Meta-Learning Operators to Optimality from Multi-Task Non-IID 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 [1], 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 [1–6] and in identification/control of linear dynamical systems [7–9]. These works study the problem of *linear representation learning*,

Preprint. Under review.

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

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

with $i = 1, \ldots, N$ iid data points from $h = 1, \ldots, H$ task distributions. Since the representation Φ is shared across all tasks, one may expect the generalization error of an approximate representation $\hat{\Phi}$ fit on HN data points to scale as $\frac{d_x r}{HN}$, where $d_x r$ is the number of parameters determining the representation. This is indeed the flavor of statistical guarantees from prior work [3–5, 9], 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 $x_i^{(h)}$ 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 [3, 9, 10]. 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 (main result, informal) *Let $\hat{\Phi}$ be the current estimate of the representation, and $\Phi_\star$ the optimal representation. Running one iteration of `DFW` yields the following improvement*

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

Critically, the second term of the right hand side scales jointly in the number of tasks and datapoints per task, whereas naively implementing other iterative 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 [1–5, 10], under the assumption of a shared but unknown linear feature representation. In particular, our proposed algorithm is adapted from the alternating optimization scheme in [1], 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, [5] 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. federated learning, that impose data locality or 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 [11–14]. In parallel, there is a body of theoretical work which studies the sample efficiency of representation learning for RL [15–17]. 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 [18–21], imitation learning for linear systems [9, 22], and notably linear system identification [7, 8, 23–26]. In many of these works [23–25], 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 optimality.

Federated learning and non-IID data: Shared global information and local models conditioned on global information appears in federated learning under the banner of *personalization* [1, 27–29] (see [30] for a survey). Recently, intense attention has been given to designing algorithms that generalize across heterogeneous agents as well as non-iid data [31–33] (see [34, 35] for surveys). However, these methods are either empirical or are not well-specified for our data assumptions. Like in [1], our algorithm is compatible with common data locality or privacy constraints.

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, $A \preccurlyeq B$ 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 $[n] := \{1, \ldots, n\}$. 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 be an indexed set $\{x[t]\}_{t\geq 1} \subset \mathbb{R}^{d_x}$. We denote a distribution $\mathbb{P}_x$ over covariate sequences, which we assume to have bounded second moments for all $t \geq 1$, i.e. $\mathbb{E}\left[x[t]x[t]^\top\right]$ is finite for all $t \geq 1$. Defining the filtration $\{\mathcal{F}[t]\}_{t\geq 0}$ where $\mathcal{F}[t] := \sigma(\{x[k]\}_{k=1}^{t+1}, \{w[k]\}_{k=1}^{t})$ is the σ-algebra generated by the covariates up to $t+1$ and noise up to t, we assume that $\{w[t]\}_{t\geq 1}$ is a σ_w^2-subgaussian martingale difference sequence (MDS):

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

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

$$y[t] = M_\star x[t] + w[t], \quad t \geq 1,$$

for $y[t]$ the *labels*, and $w[t]$ the *label noise*. We further define $\Sigma_{x,T} := \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[x[t]x[t]^\top]$.

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 $\{(x_i^{(h)}[t], y_i^{(h)}[t])\}_{t=1,i=1}^{T,N}$, consisting of N independent trajectories of length T, generated by $h = 1, \ldots, H$ task distributions. For simplicity, we assume that the number and length of trajectories N, T are the same across training tasks. For each task h, the observation model is

$$y_i^{(h)}[t] = M_\star^{(h)} x_i^{(h)}[t] + w_i^{(h)}[t], \tag{1}$$

where $M_\star^{(h)} = F_\star^{(h)} \Phi_\star$ admits a decomposition into a ground-truth representation $\Phi_\star \in \mathbb{R}^{r \times d_x}$ common across all tasks $h \in [H]$ and a task-specific weight matrix $F_\star^{(h)} \in \mathbb{R}^{d_y \times r}$, $r \leq \min\{d_x, d_y\}$. We denote the joint distribution over covariates and observations $\{x_i^{(h)}[t], y_i^{(h)}[t]\}_{t\geq 1}$ by $\mathbb{P}_{x,y}^{(h)}$. We assume that the representation $\Phi_\star$ is normalized to have orthonormal rows to prevent boundedness issues, since $F_\star^{(h)\prime} = F_\star^{(h)} Q^{-1}$, $\Phi_\star' = Q\Phi_\star$ 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 1 (Subspace Distance [1, 36]) *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 $\begin{bmatrix}\Phi_\star^\top & \Phi_{\star,\perp}^\top\end{bmatrix}$ is an orthogonal*

matrix. Define the distance between the subspaces spanned by the rows of Φ and $\Phi_\star$ by

$$\mathrm{dist}(\Phi, \Phi_\star) := \left\|\Phi\Phi_{\star,\perp}^\top\right\|_2 \tag{2}$$

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

$$X^{(h)} = \begin{bmatrix} x_1^{(h)}[0] & \cdots & x_1^{(h)}[T] & \cdots & x_i^{(h)}[t] & \cdots & x_N^{(h)}[T] \end{bmatrix}^\top \in \mathbb{R}^{NT \times d_x}.$$

The goal of multi-task operator recovery is to estimate $\{F_\star^{(h)}\}_{h=1}^H$ and $\Phi_\star$ from data collected across multiple tasks $\{(x_i^{(h)}[t], y_i^{(h)}[t])\}_{t=1,i=1}^{T,N}$, $h = 1, \ldots, H$. Some prior works [3, 9, 10] assume access to an empirical risk minimization oracle, i.e. access to

$$\{\hat{F}^{(h)}\}_{h=1}^H, \hat{\Phi} \in \operatorname*{argmin}_{\{F^{(h)}\},\Phi} \sum_{h=1}^H \sum_{i=1}^N \sum_{t=1}^T \left\| y_i^{(h)}[t] - F^{(h)}\Phi x_i^{(h)}[t] \right\|^2,$$

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 $\{F_\star^{(h)}\}_{h=1}^H$ and $\Phi_\star$. To address this problem in the linear regression setting, Collins et al. [1] propose `FedRep`, an alternating minimization-descent scheme, where on a fresh data batch, the weights $\{\hat{F}^{(h)}\}$ are computed on local data via least-squares. 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. [1]. 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 H. 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` [1]. During each optimization round, a new data batch is sampled for each task: $\{(x_i^{(h)}[t], y_i^{(h)}[t])\}_{t=1,i=1}^{T,N}$, $h \in [H]$. We then compute task-specific weights $\hat{F}^{(h)}$ on the corresponding dataset, keeping the current representation estimate $\hat{\Phi}$ fixed. For example, $\hat{F}^{(h)}$ may be the least-squares weights conditioned on $\hat{\Phi}$ [1]. Define $z_i^{(h)}[t] := \hat{\Phi} x_i^{(h)}[t]$, and the empirical covariance matrices $\hat{\Sigma}_{x,NT}^{(h)} := \frac{1}{NT} X^{(h)\top} X^{(h)}$, $\hat{\Sigma}_{z,NT}^{(h)} := \frac{1}{NT} Z^{(h)\top} Z^{(h)}$. The least squares solution $\hat{F}^{(h)}$ is given by the convex quadratic minimization

$$\begin{aligned} \hat{F}^{(h)} &= \operatorname*{argmin}_F \sum_{i=1}^N \sum_{t=1}^T \left\| y_i^{(h)}[t] - F z_i^{(h)}[t] \right\|^2 \\ &= F_\star^{(h)} \Phi_\star X^{(h)\top} Z^{(h)} (\hat{\Sigma}_{z,NT}^{(h)})^{-1} + W^{(h)\top} Z^{(h)} (\hat{\Sigma}_{z,NT}^{(h)})^{-1}, \end{aligned} \tag{3}$$

where we derive (3) through standard matrix calculus [37] and expanding (1). For each task, we then fix the weight matrix $\hat{F}^{(h)}$ 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

$$\overline{\Phi}_+^{(h)} = \hat{\Phi} - \frac{\eta}{2NT} \nabla_\Phi \left\| \sum_{i=1}^N \sum_{t=1}^T y_i^{(h)}[t] - \hat{F}^{(h)} \hat{\Phi} x_i^{(h)}[t] \right\|^2, \quad \overline{\Phi}_+ = \frac{1}{H} \sum_{h=1}^H \overline{\Phi}_+^{(h)} \tag{4}$$

where $\eta > 0$ is a given step size. We normalize $\overline{\Phi}_+$ to have orthonormal rows, e.g. by (thin/reduced) QR decomposition [38], to produce the final output $\hat{\Phi}_+$, i.e. $\overline{\Phi}_+ = R\hat{\Phi}_+$, $R \in \mathbb{R}^{r\times r}$, leading to

$$R\hat{\Phi}_+ = \hat{\Phi} - \frac{\eta}{H}\sum_{h=1}^{H} \hat{F}^{(h)\top}\left(\hat{F}^{(h)}\hat{\Phi} - F_\star^{(h)}\Phi_\star\right)\hat{\Sigma}_{x,NT}^{(h)} - \frac{\eta}{HNT}\sum_{h=1}^{H}\hat{F}^{(h)\top}W^{(h)\top}X^{(h)}. \tag{5}$$

As in [1], we right-multiply both sides of (5) by $\Phi_{\star,\perp}^\top$, recalling $\|\Phi\Phi_{\star,\perp}^\top\|_2 =: \mathrm{dist}(\Phi, \Phi_\star)$. Crucially, Collins et al. [1] assume $x_i^{(h)}[t]$ has mean 0 and identity covariance, and $w_i^{(h)}[t] \equiv 0$ across t, i, h. Therefore, the label noise terms $\hat{F}^{(h)\top}W^{(h)\top}X^{(h)}$ disappear, and the sample covariance for each task $\hat{\Sigma}_{x,NT}^{(h)}$ concentrates to identity. Under these assumptions, we get

$$\begin{aligned}\left\|R\hat{\Phi}_+\Phi_{\star,\perp}^\top\right\| &= \left\|\hat{\Phi}\Phi_{\star,\perp}^\top - \frac{\eta}{H}\sum_{h=1}^{H}\hat{F}^{(h)\top}\left(\hat{F}^{(h)}\hat{\Phi} - F_\star^{(h)}\Phi_\star\right)\hat{\Sigma}_{x,NT}^{(h)}\Phi_{\star,\perp}^\top\right\| \\ &\lesssim \underbrace{\left\|I - \frac{\eta}{H}\sum_{h=1}^{H}\hat{F}^{(h)\top}\hat{F}^{(h)}\right\|}_{\text{Contraction term}}\mathrm{dist}\left(\hat{\Phi}, \Phi_\star\right) + \underbrace{\mathcal{O}\left(\frac{1}{H}\sum_{h=1}^{H}\left\|\hat{\Sigma}_{x,NT}^{(h)} - I_{d_x}\right\|\right)}_{\text{Covariance concentration term}}.\end{aligned}$$

where we note $\Phi_\star\Phi_{\star,\perp}^\top = 0$. Therefore, under appropriate choice of η 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 $w_i^{(h)}[t]$ is introduced and when $x_i^{(h)}[t]$ has non-identity covariance.

1. When *label noise* $w_i^{(h)}[t]$ is present, since $\hat{F}^{(h)}$ is computed on $Y^{(h)}, X^{(h)}$, the gradient noise term is generally biased: $\frac{1}{NT}\mathbb{E}[\hat{F}^{(h)}W^{(h)\top}X^{(h)}] \neq 0$. Even in the simple case that all task distributions $\mathbb{P}_{x,y}^{(h)}$ are identical, $\frac{\eta}{HNT}\sum_{h=1}^{H}\hat{F}^{(h)\top}W^{(h)\top}X^{(h)}$ concentrates to its bias, and thus for large H the size of noise term is bottlenecked at $\frac{\eta}{NT}\mathbb{E}\left[\left\|\hat{F}^{(h)\top}W^{(h)\top}X^{(h)}\right\|\right]$. This critically causes the noise term to lose scaling in the *number of tasks* H, even when the tasks are identical.
2. When $x_i^{(h)}[t]$ has *non-identity covariance*, the decomposition into a contraction and covariance concentration term no longer holds, since generally $\Phi_*\mathbb{E}[\hat{\Sigma}_{x,NT}^{(h)}]\Phi_{\star,\perp}^\top \neq 0$. This causes a term whose norm otherwise concentrates around 0 in the isotropic case to scale with $\lambda_{\max}(\hat{\Sigma}_{x,NT}^{(h)}) - \lambda_{\min}(\hat{\Sigma}_{x,NT}^{(h)})$ in the worst case. Unlike prior work that assumes identical distribution of covariates $x_i^{(h)}[t]$ across tasks, this issue cannot be circumvented by whitening the covariates $x_i^{(h)}[t]$, as shifting the covariance factor to the operator $F_\star^{(h)}\Phi_\star{\Sigma_{x,T}^{(h)}}^{1/2}$ in general ruins the shared representation spanned by $\Phi_\star$.

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 $x_i^{(h)}[t]$. Toward addressing the first issue, we introduce a "de-biasing" step, where each agent computes the least squares weights $\hat{F}^{(h)}$ 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 [39], 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

$$R\hat{\Phi}_+ = \hat{\Phi} - \frac{\eta}{H}\sum_{h=1}^{H}\hat{F}^{(h)\top}\left(\hat{F}^{(h)}\hat{\Phi} - F_\star^{(h)}\Phi_\star\right) - \frac{\eta}{H}\sum_{h=1}^{H}\hat{F}^{(h)\top}W^{(h)\top}X^{(h)}\left(\hat{\Sigma}_{x,NT}^{(h)}\right)^{-1}, \tag{6}$$

where we assume $\{\hat{F}^{(h)}\}$ are computed on independent data using the aforementioned batching strategy. When $x_i^{(h)}[t], w_i^{(h)}[t]$, $t = 1, \ldots, T$, are all mutually independent, then the first two terms of

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

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

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 [40, 41]. This culminates in convergence rates that scale favorably with the number of tasks (§3.3).

To operationalize our proposed adaptations, let $D^{(h)} = \{(x_i^{(h)}[t], y_i^{(h)}[t])\}_{t=1,i=1}^{T,N}$, $h = 1, \dots, H$, be a dataset available to each agent. For simplicity, we consider partitions of multiple trajectories, but one can also subsample along single trajectories under appropriate mixing assumptions (§3.3). For the weights de-biasing step, we sub-sample trajectories $\mathcal{N}_1 \subset [N]$. For each agent, we then compute least-squares weights from $\mathcal{N}_1$:

$$\hat{F}^{(h)} = \operatorname*{argmin}_{F} \sum_{i\in\mathcal{N}_1} \sum_{t=1}^{T} \left\| y_i^{(h)}[t] - F z_i^{(h)}[t] \right\|^2. \tag{7}$$

We then sub-sample trajectories $\mathcal{N}_2 \subset [N] \setminus \mathcal{N}_1$, and compute the task-conditioned representation gradients from $\mathcal{N}_2$:

$$\hat{\mathcal{G}}_{\mathcal{N}_2}^{(h)} = \nabla_{\Phi} \frac{1}{2} \left\| \sum_{i\in\mathcal{N}_2} \sum_{t=1}^{T} y_i^{(h)}[t] - \hat{F}^{(h)} \hat{\Phi} x_i^{(h)}[t] \right\|^2. \tag{8}$$

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

$$\bar{\Phi}_+^{(h)} := \hat{\Phi} - \eta \hat{\mathcal{G}}_{\mathcal{N}_2}^{(h)} \left(\hat{\Sigma}_{x,\mathcal{N}_2 T}^{(h)}\right)^{-1}, \quad R\hat{\Phi}_+ = \frac{1}{H}\sum_{h=1}^{H} \bar{\Phi}_+^{(h)}, \tag{9}$$

We present 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):

$$\mathbb{E}\left[\operatorname{dist}(\hat{\Phi}_+, \Phi_\star)\right] = \mathbb{E}\left[\left\| R^{-1}\left(I_{d_x} - \frac{\eta}{H}\sum_{h=1}^{H} \hat{F}^{(h)\top}\hat{F}^{(h)}\right)\right\|\right] \operatorname{dist}(\hat{\Phi}, \Phi_\star), \tag{10}$$

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

Remark 1 (Choice of weights $\hat{F}^{(h)}$ vs. descent rate) *By observing the contraction expression (10), the contraction rate is seemingly solely controlled by the (average) conditioning of the weight matrices $\hat{F}^{(h)}$. Since the choice of algorithm for computing $\hat{F}^{(h)}$ is user-determined, this motivates choosing well-conditioned $\hat{F}^{(h)}$. However, the hidden trade-off lies in the orthonormalization factor R; arbitrary $\hat{F}^{(h)}$ may lead to R that undoes progress. As in [1], we analyze $\hat{F}^{(h)}$ generated by representation-conditioned least squares (7), but an optimal balance between conditioning of $\hat{F}^{(h)}$ and R can be struck by ℓ^2-regularized least squares weights $\hat{F}^{(h)}(\lambda)$ (see, e.g. [42]).*

3.3 Algorithm Guarantees

We present our main result in the form of convergence guarantees for Algorithm 1. To instantiate our bounds, we make the following assumption on the covariates.

Assumption 1 (Subgaussian covariates) *We assume the marginal distributions of $x_i^{(h)}[t]$, for all t, i, h, to be zero-mean and γ^2-subgaussian:*

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

Our final convergence rates depend on a notion of task-coverage, which we quantify as the following.

Definition 2 (Task diversity) *We define the quantities*

$$\tilde{\lambda}_{\min}^F := \lambda_{\min}\left(\frac{1}{H}\sum_{h=1}^H F_\star^{(h)\top} F_\star^{(h)}\right), \quad \tilde{\lambda}_{\max}^F := \lambda_{\max}\left(\frac{1}{H}\sum_{h=1}^H F_\star^{(h)\top} F_\star^{(h)}\right). \tag{11}$$

We further denote $\tilde{\kappa}^F := \tilde{\lambda}_{\max}^F / \tilde{\lambda}_{\min}^F$.

We now present our main result for the iid setting.

Theorem 2 (Main result, iid setting) *Let $x_i^{(h)}[t], w_i^{(h)}[t]$ be independent for all t, i, h and identically sampled for $t = 1, \ldots, T$. Let Assumption 1 hold with constant γ. Assume the current representation iterate $\hat{\Phi}$ satisfies $\mathrm{dist}(\hat{\Phi}, \Phi_\star) < \nu < 1$, and $NT \geq \tilde{\Omega}\left(\mathrm{poly}\left(\tilde{\kappa}^F, \max_h \kappa(\Sigma_x^{(h)}), \gamma, 1/\nu\right)(\max\{d_y, r\} + \log(1/\delta))\right)$. Then there exist universal constants $c_1, c_2 > 0$ such that the following is true: given $\eta \leq \frac{1}{c_1 \tilde{\lambda}_{\max}^F}$, with probability greater than $1 - \delta$, the next iterate outputted by `DFW` (Algorithm 1) satisfies*

$$\mathrm{dist}\left(\hat{\Phi}_+, \Phi_\star\right) \leq \left(1 - \eta c_2 \tilde{\lambda}_{\min}^F\right)^{1/2} \mathrm{dist}\left(\hat{\Phi}, \Phi_\star\right) + \frac{\eta \max_h \|F_\star^{(h)}\| \sigma_w^{(h)} \sqrt{d_x r}}{\sqrt{HNT}}, \tag{12}$$

where $\eta c_2 \tilde{\lambda}_{\min}^F < 1$.

We note that the first term on the RHS of (12) induces a contraction, while the second term corresponding to the variance of the gradient estimate, has the "correct" scaling: noise multiplied by # parameters of the representation, divided by the *total* amount of data H, N, T.

Remark 2 (Initialization) *We note that Theorem 2 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. [1], Tripuraneni et al. [4], Thekumparampil et al. [5] for initialization schemes in the iid linear regression setting. Our experiments suggest that a good initialization is unnecessary, which mirrors the experimental findings in Thekumparampil et al. [5, Sec. 6]. We leave constructing an initialization scheme for our general setting, or proving it is unnecessary, to future work.*

A key benefit of having (12) scale properly in H, 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 1 (Approximate ERM, iid setting) *Let the assumptions of Theorem 2 hold. Let $\hat{\Phi}_0$ be an initial representation satisfying $\mathrm{dist}(\hat{\Phi}_0, \Phi_\star) < \nu$, and define $\rho := \eta c_2 \tilde{\lambda}_{\min}^F$. Let $\boldsymbol{D} := \{\{(x_i^{(h)}[t], y_i^{(h)}[t])\}_{t=1,i=1}^{T,N}\}_{h \in [H]}$ be a given dataset. There exists a partition of $\boldsymbol{D}$ into independent batches $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_K$, such that iterating `DFW` on $\boldsymbol{B}_k$, $k \in [K]$ yields with probability greater than $1 - \delta$:*

$$\mathrm{dist}(\hat{\Phi}_K, \Phi_\star)^2 \leq \tilde{\mathcal{O}}\left(C(\rho) \frac{\max_h \sigma_w^{(h)2}(d_x r + \log(1/\delta))}{HNT}\right), \tag{13}$$

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

In particular, given a fine-tuning dataset of size $N'T'$ sampled from a task $H+1$ that shares the representation $\Phi_\star$, computing the least squares weights $\hat{F}^{(H+1)}$ conditioned on $\hat{\Phi}_K$ yields a high-probability bound on the parameter error

$$\left\|\hat{F}^{(H+1)}\hat{\Phi}_K - M_\star^{(H+1)}\right\|_F^2 \lesssim \operatorname{dist}(\hat{\Phi}_K, \Phi_\star)^2 + {\sigma_w^{(H+1)}}^2 \frac{d_y r + \log(1/\delta)}{N'T'}$$
$$\lesssim C(\rho)\frac{\max_h {\sigma_w^{(h)}}^2 (d_x r + \log(1/\delta))}{HNT} + \frac{{\sigma_w^{(H+1)}}^2 (d_y r + \log(1/\delta))}{N'T'},$$

where we omit task-related quantities for clarity. We note that the above parameter recovery bound mirrors ERM risk bounds [3, 4, 9], where we note the latter term scales with $d_y r$ (the number of parameters in $F^{(H+1)}$) as opposed to r in the linear regression setting ($d_y = 1$).

We note that the results in this section also hold for sequentially dependent data, i.e. $\{x^{(h)}[t]\}_{t\geq 1}$ are β-mixing [43–45], which we describe in detail in Appendix A.3. In particular, this opens up the applicability of `DFW` to settings where the data is fundamentally non-IID and non-isotropic. We instantiate our results to the fundamental setting of linear dynamical systems [46] in Appendix B, as well as numerically in the sequel.

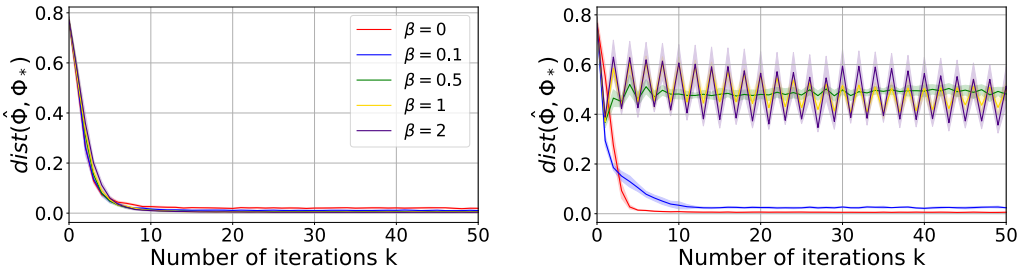
4 Numerical Results

We present numerical experiments to demonstrate the effectiveness of Algorithm 1. We consider two scenarios: 1) linear regression with IID, non-isotropic data, and 2) linear system identification The linear regression experiments highlight the necessity of our proposed `De-biasing & Feature-whitening` steps, and our system identification experiments demonstrate the ability of our algorithm to extend to sequential *non-i.i.d.* and non-isotropic data. We show that `DFW` allows us to harness the full dataset by its improved performance and variance compared to its single-task variant and `FedRep`. Full experimental details and additional experiments 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 set the number of samples $N = 100$ across all $H = 10$ tasks. We generate the H operators using the following steps: 1) the ground truth representation $\Phi_\star \in \mathbb{R}^{r\times d_x}$ is randomly generated through the `thin_svd` operation applied to a random matrix with values drawn iid from $\mathcal{N}(0,1)$, 2) a nominal task weight matrix $F_0 \in \mathbb{R}^{d_y \times r}$ is randomly generated, with its elements drawn from $\mathcal{N}(0,1)$, 3) to generate the task-specific weights $F_\star^{(h)} \in \mathbb{R}^{d_y\times r}$, $h \in [H]$, random rotations around the identity are applied to F_0. The covariates are drawn iid from a Gaussian distribution $x_i^{(h)} \sim \mathcal{N}(0, \Sigma_x)$, where the covariance matrix is expressed $\Sigma_x = \texttt{diag}([1, 1+\beta, \ldots, 1+(d_x-1)\beta])$, such that the value of β controls the non-isotropy of the covariates $x_i^{(h)}$. Note that $\beta = 0$ recovers isotropy. Figures (1a-1b) illustrate the influence of varying the degree of non-isotropy β: we examine its impact on the performance of `DFW` (Algorithm 1) and compare it with the vanilla alternating minimization-descent algorithm `FedRep` [1], which does not incorporate de-biasing and feature-whitening. The $\beta = 0$ case corresponds to an identity covariance matrix, and our results show both algorithms are effective at recovering the representation as expected. Consistent with our theoretical guarantees, we find that `DFW` remains effective when the non-isotropy measure β is increased; in contrast, Figure 1b shows how `FedRep` fails when dealing with even mildly non-isotropic data. This aligns with our findings in §3.1 and highlights the limitations of the vanilla representation gradient in handling non-isotropic feature data.

We now compare the benefit of multi-task setting over single-task. The problem dimensions d_x, d_y, r, N are carried over from the previous experiment. We now introduce a different covariance matrix parameterization: $\Sigma_x = \frac{d_x \tilde{\Sigma}_x}{\operatorname{Tr}(\tilde{\Sigma}_x)}$, where $\tilde{\Sigma}_x = \frac{1}{2}(U + U^\top)$, $U = 5 \cdot I_{d_x} + V$, with V being a random matrix whose values are drawn from $\operatorname{Unif}(0,1)$. Figure 2a compares the performance of Algorithm 1 for the single-task ($H = 1$) and multi-task settings ($H = 25$), as well as `FedRep` ($H = 25$). The figure highlights that `DFW` is able to make use of all source data to decrease variance and learn an accurate low-rank representation. As anticipated by our theoretical guarantees, Figure 2a demonstrates a significant reduction in the subspace distance as the number of tasks increases.



(a) `DFW` alt. minimization-descent (Algorithm 1) **(b)** Vanilla alt. minimization-descent (`FedRep` [1])

Figure 1: We compare subspace distance between the current and ground truth representation with respect to the number of iterations, where we vary non-isotropy through β. `DFW` suffers no degradation in performance when non-isotropy of the covariates increases, while `FedRep` fails catastrophically even under mild non-isotropy.

Furthermore, the figure reveals that the vanilla alternating descent algorithm is not able to improve beyond a certain point, as predicted in §3.1.

4.2 System Identification

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

$$x[t+1] = Ax[t] + Bu[t] + w[t], \;\; t = 0, \ldots, T-1,$$

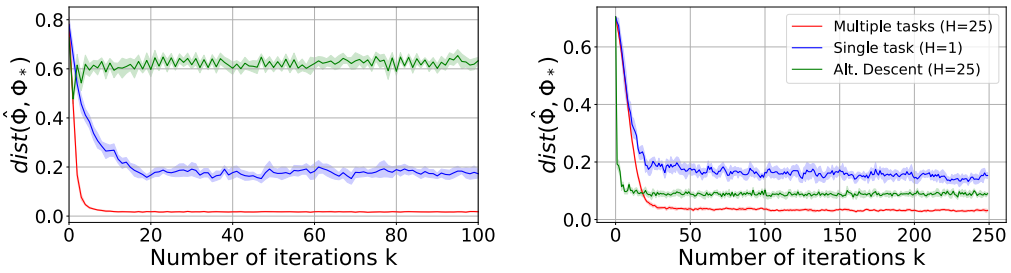
where $x[t]$ is the state of the system and $u[t]$ is the control input. In contrast to the previous example, the covariates are now additionally non-iid due to correlation over time. In particular, we can instantiate multi-task linear sysID in the form of (1),

$$x^{(h)}[t+1] = M_\star^{(h)} z^{(h)}[t] + w^{(h)}[t], \;\; t = 0, \ldots, T-1$$

where $M_\star^{(h)} := [A^{(h)} \;\; B^{(h)}] = F^{(h)}\Phi_\star \in \mathbb{R}^{d_x \times d_z}$. The state-action pair at time instant t for all tasks $h \in [H]$ is embedded as $z^{(h)}[t] = [x^{(h)}[t]^\top \;\; u^{(h)}[t]^\top]^\top$. The process noise $w^{(i)}[t]$ and control action $u^{(h)}[t]$ 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 $T = 100$, and input variance $\sigma_u^2 = 1$. The generation process of the ground truth system matrices $M_\star^{(h)}$ follows a similar approach as described in the linear regression problem, with the addition of normalization step of the nominal weight matrix F_0 to ensure system stability for all tasks $h \in [H]$. Furthermore, the process noise covariance Σ_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 $x^{(h)}[0]$ is drawn iid across tasks from the system's stationary distribution $\mathcal{N}(0, \Sigma_x^{(h)})$, which is determined by the solution to the discrete Lyapunov equation $\Sigma_x^{(h)} = A^{(h)}\Sigma_x^{(h)}(A^{(h)})^\top + \sigma_u^2 B^{(h)}(B^{(h)})^\top + \Sigma_w$. We note this implies the covariates $x_i^{(h)}[t]$ are inherently non-isotropic. Figure 2b again demonstrates the advantage of leveraging multi-task data to reduce the error in computing a shared representation across the system matrices $M_\star^{(h)}$. In line with our theoretical findings, `DFW` continues to benefit from multiple tasks, even when the data is non-iid. 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 ℓ^2-regularized weights $\hat{F}^{(h)}$.



(a) IID linear regression with random covariance

(b) Linear system identification

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`. We observe performance improvement and variance reduction for multi-task `DFW` as predicted.

Resolving these questions has important implications for the natural extension of our framework: as emphasized in [1], 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 [47] and kernel ridge regression [48]. 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

[1] 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.

[2] 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.

[3] 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.

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

[5] Kiran Koshy Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Sample efficient linear meta-learning by alternating minimization, 2021.

[6] Nikunj Saunshi, Arushi Gupta, and Wei Hu. A representation learning perspective on the importance of train-validation splitting in meta-learning, 2021.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] Matteo Hessel, Hubert Soyer, Lasse Espeholt, Wojciech Czarnecki, Simon Schmitt, and Hado van Hasselt. Multi-task deep reinforcement learning with popart, 2018.

[13] 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.

[14] 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.

[15] Rui Lu, Gao Huang, and Simon S. Du. On the power of multitask representation learning in linear mdp, 2021.

[16] Yuan Cheng, Songtao Feng, Jing Yang, Hong Zhang, and Yingbin Liang. Provable benefit of multitask representation learning in reinforcement learning, 2022.

[17] Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning, 2015.

[18] 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.

[19] 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.

[20] 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.

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

[22] Taosha Guo, Abed AlRahman Al Makdah, Vishaal Krishnan, and Fabio Pasqualetti. Imitation and transfer learning for lqg control. *arXiv preprint arXiv:2303.09002*, 2023.

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

[24] Han Wang, Leonardo F Toso, and James Anderson. Fedsysid: A federated approach to sample-efficient system identification. *arXiv preprint arXiv:2211.14393*, 2022.

[25] Lei Xin, Lintao Ye, George Chiu, and Shreyas Sundaram. Learning dynamical systems by leveraging data from similar systems. *arXiv preprint arXiv:2302.04344*, 2023.

[26] Mohamad Kazem Shirani Faradonbeh and Aditya Modi. Joint learning-based stabilization of multiple unknown linear systems. *IFAC-PapersOnLine*, 55(12):723–728, 2022.

[27] Aritra Mitra, Rayana Jaafar, George J. Pappas, and Hamed Hassani. Linear convergence in federated learning: Tackling client heterogeneity and sparse gradients, 2021.

[28] Prateek Jain, John Rush, Adam Smith, Shuang Song, and Abhradeep Guha Thakurta. Differentially private model personalization. *Advances in Neural Information Processing Systems*, 34: 29723–29735, 2021.

[29] Alberto Bietti, Chen-Yu Wei, Miroslav Dudik, John Langford, and Steven Wu. Personalization improves privacy-accuracy tradeoffs in federated learning. In *International Conference on Machine Learning*, pages 1945–1962. PMLR, 2022.

[30] Viraj Kulkarni, Milind Kulkarni, and Aniruddha Pant. Survey of personalization techniques for federated learning. In *2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*, pages 794–797. IEEE, 2020.

[31] Mi Luo, Fei Chen, Dapeng Hu, Yifan Zhang, Jian Liang, and Jiashi Feng. No fear of heterogeneity: Classifier calibration for federated learning with non-iid data, 2021.

[32] Lei Yang, Jiaming Huang, Wanyu Lin, and Jiannong Cao. Personalized federated learning on non-iid data via group-based meta-learning. *ACM Trans. Knowl. Discov. Data*, 17(4), mar 2023. ISSN 1556-4681. doi: 10.1145/3558005. URL https://doi.org/10.1145/3558005.

[33] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data, 2023.

[34] Hangyu Zhu, Jinjin Xu, Shiqing Liu, and Yaochu Jin. Federated learning on non-iid data: A survey, 2021.

[35] Xiaodong Ma, Jia Zhu, Zhihao Lin, Shanxuan Chen, and Yangjie Qin. A state-of-the-art survey on solving non-iid data in federated learning. *Future Generation Computer Systems*, 135: 244–258, 2022.

[36] Gilbert W Stewart and Ji-guang Sun. *Matrix perturbation theory*. Academic press, 1990.

[37] Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.

[38] Lloyd N Trefethen and David Bau. *Numerical linear algebra*, volume 181. Siam, 2022.

[39] 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.

[40] 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.

[41] Yasin Abbasi-Yadkori. Online learning for linearly parametrized control problems. 2013.

[42] Daniel Hsu, Sham M Kakade, and Tong Zhang. Random design analysis of ridge regression. In *Conference on learning theory*, pages 9–1. JMLR Workshop and Conference Proceedings, 2012.

[43] Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, pages 94–116, 1994.

[44] Mehryar Mohri and Afshin Rostamizadeh. Rademacher complexity bounds for non-iid processes. *Advances in Neural Information Processing Systems*, 21, 2008.

[45] Vitaly Kuznetsov and Mehryar Mohri. Generalization bounds for non-stationary mixing processes. *Machine Learning*, 106(1):93–117, 2017.

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

[47] 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.

[48] Dimitri Meunier, Zhu Li, Arthur Gretton, and Samory Kpotufe. Nonlinear meta-learning can guarantee faster rates, 2023.

[49] Tuhin Sarkar, Alexander Rakhlin, and Munther A Dahleh. Finite time lti system identification. *The Journal of Machine Learning Research*, 22(1):1186–1246, 2021.

[50] 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.

[51] 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.

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

[53] 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.

[54] 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.

[55] 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.

[56] Benjamin Recht. A tour of reinforcement learning: The view from continuous control. *Annual Review of Control, Robotics, and Autonomous Systems*, 2:253–279, 2019.

[57] 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.

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

[59] 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.

[60] 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.

[61] Naman Agarwal, Elad Hazan, and Karan Singh. Logarithmic regret for online control. *Advances in Neural Information Processing Systems*, 32, 2019.

[62] 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, 2019.

[63] 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.

[64] Ingvar Ziemann and Henrik Sandberg. Regret lower bounds for learning linear quadratic gaussian systems. *arXiv preprint arXiv:2201.01680*, 2022.

[65] 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.

[66] 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.

[67] 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.

[68] 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.

[69] Stephen Tu, Roy Frostig, and Mahdi Soltanolkotabi. Learning from many trajectories. *arXiv preprint arXiv:2203.17193*, 2022.

[70] Alexander Goldenshluger and Assaf Zeevi. Nonasymptotic bounds for autoregressive time series modeling. *The Annals of Statistics*, 29(2):417–444, 2001.

Contents

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 [40, 41].

Lemma 1 (cf. Zhang et al. [9, Lemma B.3]) *Let $\{x_t\}_{t\geq 1}$ be a $\mathbb{R}^d$-valued process adapted to a filtration $\{\mathcal{F}_t\}_{t\geq 1}$. Let $\{\eta_t\}_{t\geq 1}$ be a $\mathbb{R}^m$-valued process adapted to $\{\mathcal{F}_t\}_{t\geq 2}$. Suppose that $\{\eta_t\}_{t\geq 1}$ is a σ^2-subgaussian martingale difference sequence, i.e.,:*

$$\mathbb{E}[\eta_t \mid \mathcal{F}_t] = 0, \tag{14}$$

$$\mathbb{E}[\exp(\lambda v^\top \eta_t) \mid \mathcal{F}_t] \leq \exp\left(\frac{\lambda^2\sigma^2 \|v\|^2}{2}\right) \ \forall \mathcal{F}_t\text{-measurable } \lambda \in \mathbb{R}, v \in \mathbb{R}^m. \tag{15}$$

For $\Lambda \in \mathbb{R}^{m\times d}$, let $\{M_t(\Lambda)\}_{t\geq 1}$ be the $\mathbb{R}$-valued process:

$$M_t(\Lambda) = \exp\left(\frac{1}{\sigma}\sum_{i=1}^{t}\langle \Lambda x_i, \eta_i\rangle - \frac{1}{2}\sum_{i=1}^{t}\|\Lambda x_i\|^2\right). \tag{16}$$

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

In particular, this implies the following self-normalized martingale inequality that handles multiple matrix-valued self-normalized martingales. This can be seen as an instantiation of the Hilbert space variant from [41].

Proposition 1 (cf. Zhang et al. [9, Prop. B.1]) *Fix $H \in \mathbb{N}_+$. For $h \in [H]$, let $\{x_t^h, \eta_t^h\}_{t\geq 1}$ be a $\mathbb{R}^d \times \mathbb{R}^m$-valued process and $\{\mathcal{F}_t^h\}_{t\geq 1}$ be a filtration such that $\{x_t^h\}_{t\geq 1}$ is adapted to $\{\mathcal{F}_t^h\}_{t\geq 1}$, $\{\eta_t^h\}_{t\geq 1}$ is adapted to $\{\mathcal{F}_t^h\}_{t\geq 2}$, and $\{\eta_t^h\}_{t\geq 1}$ is a σ^2-subgaussian martingale difference sequence. Suppose that for all $h_1 \neq h_2$, the process $\{x_t^{h_1}, \eta_t^{h_1}\}$ is independent of $\{x_t^{h_2}, \eta_t^{h_2}\}$. Fix (non-random) positive definite matrices $\{V^h\}_{h=1}^H$. For $t \geq 1$ and $h \in [H]$, define:*

$$\bar{V}_t^h := V^h + V_t^h, \ \ V_t^h := \sum_{i=1}^{t} x_i x_i^\top, \ \ S_t^h := \sum_{i=1}^{t} x_i \eta_i^\top. \tag{17}$$

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

$$\sum_{h=1}^{H}\left\|(\bar{V}_T^h)^{-1/2} S_T^h\right\|_F^2 \leq 2\sigma^2\left[\sum_{h=1}^{H}\log\det((\bar{V}_T^h)^{m/2}(V^h)^{-m/2}) + \log(1/\delta)\right]. \tag{18}$$

We also consider the spectral norm variant of the self-normalized martingale bound.

Proposition 2 (Sarkar et al. [49, Proposition 8.2]) *Let $\{x_t\}_{t\geq 1}$ be an $\mathbb{R}^d$-valued stochastic process adapted to filtration $\{\mathcal{F}_t\}_{t\geq 1}$ and $\{\eta_t\}_{t\geq 1} \subset \mathbb{R}^m$ be a σ^2-subgaussian martingale difference sequence adapted to $\{\mathcal{F}_t\}_{t\geq 2}$ as defined in Lemma 1. Fix a $\delta \in (0,1)$ and a non-random positive definite matrix $V \in \mathbb{R}^{d\times d}$. For $t \geq 1$, define $\bar{V}_t := V + V_t$, $V_t := \sum_{i=1}^t x_i x_i^\top$, $S_t := \sum_{i=1}^t x_i \eta_i^\top$. Then with probability at least $1-\delta$,*

$$\left\|(V_T)^{-1/2} S_T\right\|^2 \leq 8\sigma^2\left[\log\left(5^m \det(\bar{V}_T)^{1/2}\det(V)^{-1/2}\right) + \log(1/\delta)\right]. \tag{19}$$

We introduce the following useful two-sided concentration inequality for the sample covariance of iid subgaussian covariates.

Lemma 2 (Du et al. [3, Lemma A.6]) *Let $x_1, \ldots, x_T \in \mathbb{R}^d$ be iid random vectors that satisfy $\mathbb{E}[x_t] = 0$, $\mathbb{E}\left[x_t x_t^\top\right] = I_d$, and x_t is γ^2-subgaussian. Fix $\delta \in (0,1)$. Suppose $T \gtrsim \gamma^2(d + \log(1/\delta))$. Then with probability at least $1-\delta$, the following holds*

$$0.9 I_d \preceq \frac{1}{T}\sum_{i=1}^{T} x_t x_t^\top \preceq 1.1 I_d. \tag{20}$$

In order to instantiate our bounds for non-iid covariates, we introduce the notions of β-mixing stationary processes [45, 50].

Definition 3 (β-mixing) *Let $\{x_t\}_{t\geq 1}$ be a $\mathbb{R}^d$-valued discrete-time stochastic process adapted to filtration $\{\mathcal{F}_t\}_{t=1}^{\infty}$. We denote the stationary distribution ν_∞. We define the β-mixing coefficient*

$$\beta(k) := \sup_{t\geq 1} \mathbb{E}_{\{x_\ell\}_{\ell=1}^t} \left[\left\| \mathbb{P}_{x_{t+k}}(\cdot \mid \mathcal{F}_t) - \nu_\infty \right\|_{\mathrm{tv}} \right], \tag{21}$$

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

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

$$X_{(j)}^T := \{x_t : 1 \leq t \leq T,\ (t-1 \bmod a) = j-1\},\ \ j = 1, \ldots, a. \tag{22}$$

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

Lemma 3 (Kuznetsov and Mohri [45, 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 $j = 1, \ldots, a$.*

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

In our analysis, we often instantiate Lemma 3 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 4 (Matrix Hoeffding [51]) *Let $\{X_h\}_{h=1}^H \subset \mathbb{R}^{d\times d}$ be a sequence of independent, random symmetric matrices, and let $\{B_h\}_{h=1}^H$ be a sequence of fixed symmetric matrices. Assume each random matrix satisfies*

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

Then for all $t \geq 0$,

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

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

$$\mathbb{E}[M_h] = 0, \quad X_h^2 \preceq B_h^2 \text{ almost surely.}$$

Then for all $t \geq 0$,

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

As hinted by the indexing of the matrices, by leveraging the independence of processes across tasks h, Lemma 4 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 $\{(x_i^{(h)}[t], y_i^{(h)}[t])\}_{t=1,i=1}^{T,N}$, $h = 1, \ldots, H$, and DFW trajectory partitions $\mathcal{N}_1, \mathcal{N}_2$, the least squares weights $\hat{F}^{(h)}$ can be written as

$$\begin{aligned}\hat{F}^{(h)} &= \operatorname*{argmin}_F \sum_{i\in\mathcal{N}_1}\sum_{t=1}^{T}\left\|y_i^{(h)}[t] - Fz_i^{(h)}[t]\right\|^2 \\ &= F_\star^{(h)}\Phi_\star X_{\mathcal{N}_1}^{(h)\top} Z_{\mathcal{N}_1}^{(h)}\left(\hat{\Sigma}_{z,\mathcal{N}_1 T}^{(h)}\right)^{-1} + W_{\mathcal{N}_1}^{(h)\top} Z_{\mathcal{N}_1}^{(h)}\left(\hat{\Sigma}_{z,\mathcal{N}_1 T}^{(h)}\right)^{-1} \\ &= F_\star^{(h)}\Phi_\star\hat{\Phi}^\top + F_\star^{(h)}\Phi_\star\left(I_{d_x} - \hat{\Phi}^\top\hat{\Phi}\right) X_{\mathcal{N}_1}^{(h)\top} Z_{\mathcal{N}_1}^{(h)}\left(\hat{\Sigma}_{z,\mathcal{N}_1 T}^{(h)}\right)^{-1} + W_{\mathcal{N}_1}^{(h)\top} Z_{\mathcal{N}_1}^{(h)}\left(\hat{\Sigma}_{z,\mathcal{N}_1 T}^{(h)}\right)^{-1}.\end{aligned} \tag{23}$$

Now recalling the DFW representation update in the iid setting (6), we have

$$R\hat{\Phi}_+ = \hat{\Phi} - \frac{\eta}{H}\sum_{h=1}^{H}\hat{F}^{(h)\top}\left(\hat{F}^{(h)}\hat{\Phi} - F_\star^{(h)}\Phi_\star\right) - \frac{\eta}{H}\sum_{h=1}^{H}\hat{F}^{(h)\top} W_{\mathcal{N}_2}^{(h)\top} X_{\mathcal{N}_2}^{(h)}\left(\hat{\Sigma}_{x,NT}^{(h)}\right)^{-1}. \tag{24}$$

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

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

This naturally decomposes into a contraction term and a noise term. We start with an analysis of the noise term. We observe that since $\hat{F}^{(h)}$ is by construction independent of $W_{\mathcal{N}_2}^{(h)}, X_{\mathcal{N}_2}^{(h)}$, by the independence of $x_i^{(h)}[t]$ across t and i, and the noise independence $w_i^{(h)}[t] \perp x_i^{(h)}[t]$, we find

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

Therefore, we set up for an application of Lemma 4. Toward doing so, we prove the following two ingredients: 1. a high probability bound on $\left\|\hat{F}^{(h)}\right\|$, 2. a high probability bound on the least-squares noise-esque term $\left\|\hat{F}^{(h)\top} W_{\mathcal{N}_2}^{(h)\top} X_{\mathcal{N}_2}^{(h)}\left(\hat{\Sigma}_{x,\mathcal{N}_2 T}^{(h)}\right)^{-1}\right\|$. We then condition on these two high-probability events to instantiate the almost-sure boundedness in Lemma 4. We start with the analysis of $\hat{F}^{(h)}$. We observe trivially that

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

Lemma 5 *Let $|\mathcal{N}_1|T \gtrsim \gamma^2\left(\max\{d_y, r\} + \log(1/\delta)\right)$. Then, with probability greater than $1-\delta$, we have*

$$\left\|X_{\mathcal{N}_1}^{(h)\top} Z_{\mathcal{N}_1}^{(h)}\left(\hat{\Sigma}_{z,\mathcal{N}_1 T}^{(h)}\right)^{-1}\right\| \le \frac{2}{3}\left(\kappa\left(\Sigma_x^{(h)}\right) + 1\right), \tag{25}$$

$$\left\|W_{\mathcal{N}_1}^{(h)\top} Z_{\mathcal{N}_1}^{(h)}\left(\hat{\Sigma}_{z,\mathcal{N}_1 T}^{(h)}\right)^{-1}\right\| \le 6\sigma_w^{(h)}\sqrt{\frac{d_y + \log(1/\delta)}{\lambda_{\min}(\Sigma_x^{(h)})|\mathcal{N}_1|T}}. \tag{26}$$

Therefore, setting $|\mathcal{N}_1|T \gtrsim \max\left\{\gamma^2, \frac{\sigma_w^{(h)2}}{\lambda_{\min}(\Sigma_x^{(h)})}\right\}(d_y + \log(1/\delta))$, under the assumption that $\left\|\Phi_\star\mathcal{P}_{\hat{\Phi}}^\perp\right\| \le \frac{3}{\kappa\left(\Sigma_x^{(h)}\right)+1}$, we get a high-probability bound on $\hat{F}^{(h)}$.

Lemma 6 *Assume* $|\mathcal{N}_1|T \gtrsim \max\left\{\gamma^2, \frac{\sigma_w^{(h)2}}{\lambda_{\min}(\Sigma_x^{(h)})}\right\}(\max\{d_y, r\} + \log(1/\delta))$ *and* $\left\|\Phi_\star \mathcal{P}_{\hat{\Phi}}^{\perp}\right\| \leq \frac{3}{\kappa\left(\Sigma_x^{(h)}\right)+1}$. *Then with probability at least* $1-\delta$

$$\left\|\hat{F}^{(h)}\right\| \leq 2\left\|F_\star^{(h)}\right\|. \tag{27}$$

Denote the event on which Lemma 6 holds with probability at least $1-\delta$ as $\mathcal{E}_{\hat{F}^{(h)}}(\delta)$. Then, conditioned on $\mathcal{E}_{\hat{F}^{(h)}}(\delta)$, we observe that $\hat{F}^{(h)} W_{\mathcal{N}_2}^{(h)}$ is a $4\sigma_w^{(h)2}\|F_\star^{(h)}\|^2$-subgaussian MDS supported on $\mathbb{R}^r$. Therefore, bounding

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

we invoke Proposition 2 and Lemma 2, we get the following bound

Lemma 7 *Let the conditions of Lemma 6 hold. Then, conditioned on the success event* $\mathcal{E}_{\hat{F}^{(h)}}(\delta)$, *with probability at least* $1-\delta$,

$$\left\|\hat{F}^{(h)\top} W_{\mathcal{N}_2}^{(h)\top} X_{\mathcal{N}_2}^{(h)} \left(\hat{\Sigma}_{x,\mathcal{N}_2 T}^{(h)}\right)^{-1}\right\| \leq 12\sigma_w^{(h)}\|F_\star^{(h)}\|\sqrt{\frac{d_x + \log(1/\delta)}{\lambda_{\min}(\Sigma_x^{(h)})|\mathcal{N}_2|T}}. \tag{28}$$

Therefore, we have provided a high-probability bound on the task-wise noise term. Define the matrices

$$B^{(h)} := 12\frac{\sigma_w^{(h)}\|F_\star^{(h)}\|}{H}\sqrt{\frac{d_x + \log(H/\delta)}{\lambda_{\min}(\Sigma_x^{(h)})|\mathcal{N}_2|T}} \cdot I_{d_y+r}.$$

Union-bounding over the high probability bound events of Lemma 7 with probability $1-\delta/H$ each and setting $B^{(h)}$ as above, we may apply Lemma 4 with $\sigma^2 := \sum_h {B^{(h)}}^2$ to yield the following bound.

Lemma 8 *Conditioning on the bounds of Lemma 7 holding with probability at least* $1-\delta/H$ *for each* $h \in [H]$, *we have with probability at least* $1-\delta$,

$$\left\|\frac{\eta}{H}\sum_{h=1}^{H} \hat{F}^{(h)\top} W_{\mathcal{N}_2}^{(h)\top} X_{\mathcal{N}_2}^{(h)} \left(\hat{\Sigma}_{x,\mathcal{N}_2 T}^{(h)}\right)^{-1}\right\| \lesssim \max_h \frac{\eta\sigma_w^{(h)}\|F_\star^{(h)}\|}{\sqrt{H|\mathcal{N}_2|T}}\sqrt{\frac{d_x + \log(H/\delta)}{\lambda_{\min}(\Sigma_x^{(h)})}\log\left(\frac{d_y+r}{\delta}\right)}.$$

Importantly, setting $|\mathcal{N}_2| = \Theta(N)$, we note that this establishes the desired $\tilde{O}\left(1/\sqrt{HNT}\right)$ scaling of the noise term. We note that our application of Matrix Hoeffding is rather crude, and the above bound can likely be improved in terms of polylog$(1/\delta)$ factors.

We now move on to bounding the contraction term. Defining

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

we write (23) as $\hat{F}^{(h)} = F_\star^{(h)}\Phi_\star\hat{\Phi}^\top + \Delta^{(h)} + E^{(h)}$. Expanding

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

where $\mathrm{Sym}(A) := A + A^\top$. From Lemma 6, we immediately get an upper bound on $\lambda_{\max}(\hat{F}^{(h)\top}\hat{F}^{(h)}) \leq 4\left\|F_\star^{(h)}\right\|^2$. To lower bound $\lambda_{\min}(\hat{F}^{(h)\top}\hat{F}^{(h)})$, we observe that the diagonal terms $\Delta^{(h)\top}\Delta^{(h)}$, $E^{(h)\top}E^{(h)}$ are pd, and thus can be ignored. We then observe that by Weyl's inequality [52], we have

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

We further observe that

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

such that the cross terms may be bounded as

$$\begin{aligned}&\lambda_{\max}\left(\mathrm{Sym}(\Delta^{(h)\top} F_\star^{(h)} \Phi_\star \hat{\Phi}^\top) + \mathrm{Sym}(E^{(h)\top} F_\star^{(h)} \Phi_\star \hat{\Phi}^\top) + \mathrm{Sym}(\Delta^{(h)\top} E^{(h)})\right) \\ &\leq 2\left\|\Delta^{(h)\top} F_\star^{(h)} \Phi_\star \hat{\Phi}^\top\right\| + 2\left\|E^{(h)\top} F_\star^{(h)} \Phi_\star \hat{\Phi}^\top\right\| + 2\left\|\Delta^{(h)\top} E^{(h)}\right\| \\ &= 2\left\|\Delta^{(h)}\right\|\left\|F_\star^{(h)}\right\| + 2\left\|E^{(h)}\right\|\left\|F_\star^{(h)}\right\| + 2\left\|\Delta^{(h)}\right\|\left\|E^{(h)}\right\|.\end{aligned}$$

Applying a similar analysis as in Lemma 5, we derive the following.

Lemma 9 *Assume* $|\mathcal{N}_1|T \gtrsim \max\left\{\gamma^2, \frac{\sigma_w^{(h)2}}{\lambda_{\min}(\Sigma_x^{(h)})} \frac{\|F_\star\|^2}{\sigma_{\min}(F_\star^{(h)})^4}\right\}(\max\{d_y, r\} + \log(1/\delta))$, *and* $\left\|\Phi_\star \mathcal{P}_{\hat{\Phi}}^\perp\right\| \leq \frac{3}{2} \frac{c}{\kappa\left(\Sigma_x^{(h)}\right)+1} \frac{\sigma_{\min}(F_\star^{(h)})^2}{\|F_\star\|}$, *where* $c > 0$ *is a sufficiently small, fixed numerical constant. Then with probability at least* $1 - \delta$,

$$\lambda_{\max}\left(\mathrm{Sym}(\Delta^{(h)\top} F_\star^{(h)} \Phi_\star \hat{\Phi}^\top) + \mathrm{Sym}(E^{(h)\top} F_\star^{(h)} \Phi_\star \hat{\Phi}^\top) + \mathrm{Sym}(\Delta^{(h)\top} E^{(h)})\right) \leq 3c\sigma_{\min}(F_\star^{(h)})^2.$$

Now, using the fact that

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

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

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

Under the assumption that $\left\|\Phi_{\star,\perp}\hat{\Phi}^\top\right\| = \left\|\Phi_\star \mathcal{P}_{\hat{\Phi}}^\perp\right\|$ is sufficiently small, e.g., $\left\|\Phi_{\star,\perp}\hat{\Phi}^\top\right\| \leq 1/2$ such that $1 - \left\|\Phi_{\star,\perp}\hat{\Phi}^\top\right\|^2 \geq 3/4$, then we have the following lower bound on $\lambda_{\min}(\hat{F}^{(h)\top}\hat{F}^{(h)})$: with probability at least $1 - \delta$

$$\lambda_{\min}(\hat{F}^{(h)\top}\hat{F}^{(h)}) \geq \left(\frac{3}{4} - 3c\right)\lambda_{\min}(F_\star^{(h)\top} F_\star^{(h)}).$$

As such, for $\eta \leq \min_h \frac{1}{4\|F_\star^{(h)}\|^2}$ we may bound the contraction factor by

$$\left\|I_{d_x} - \eta \frac{1}{H}\sum_{h=1}^H \hat{F}^{(h)\top}\hat{F}^{(h)}\right\| \leq \left(1 - \eta\left(\frac{3}{4} - 3c\right)\frac{1}{H}\sum_{h=1}^H \sigma_{\min}(F_\star^{(h)})^2\right).$$

Piecing this together, we have the following

Lemma 10 *Assume* $|\mathcal{N}_1|T \gtrsim \max\left\{\gamma^2, \frac{\sigma_w^{(h)2}}{\lambda_{\min}(\Sigma_x^{(h)})}\frac{\|F_\star\|^2}{\sigma_{\min}(F_\star^{(h)})^4}\right\}(\max\{d_y, r\} + \log(H/\delta))$, *and* $\left\|\Phi_\star \mathcal{P}_{\hat{\Phi}}^{\perp}\right\| \le \min\left\{\frac{3}{16}\frac{1}{\kappa\left(\Sigma_x^{(h)}\right)+1}\frac{\sigma_{\min}(F_\star^{(h)})^2}{\|F_\star\|}, 1/2\right\}$. *Let* $\eta \le \min_h 1/4\|F_\star^{(h)}\|^2$. *There exists numerical constant* $C > 0$ *such that the following holds: with probability at least* $1-\delta$,

$$\left\|R\hat{\Phi}_+\Phi_{\star,\perp}^\top\right\| \le \left(1 - \frac{3\eta}{8}\frac{1}{H}\sum_{h=1}^H \sigma_{\min}(F_\star^{(h)})^2\right)\left\|\hat{\Phi}\Phi_{\star,\perp}^\top\right\| + \max_h C\frac{\eta\sigma_w^{(h)}\|F_\star^{(h)}\|}{\sqrt{H|\mathcal{N}_2|T}}\sqrt{\frac{d_x+\log(H/\delta)}{\lambda_{\min}(\Sigma_x^{(h)})}\log\left(\frac{d_y+r}{\delta}\right)},$$

The last remaining step is to bound the effect of the orthogonalization factor R. We want to upper bound $\|R^{-1}\| = 1/\sigma_{\min}(R)$, and thus it suffices to lower bound $\sigma_{\min}(R)$. By definition, we have

$$\begin{aligned}
&RR^\top = (R\hat{\Phi}_+)(R\hat{\Phi}_+)^\top \\
&= \left(\hat{\Phi} - \frac{\eta}{H}\sum_{h=1}^H \hat{F}^{(h)\top}\left(\hat{F}^{(h)}\hat{\Phi} - F_\star^{(h)}\Phi_\star\right) - \frac{\eta}{H}\sum_{h=1}^H \hat{F}^{(h)\top}W^{(h)\top}X^{(h)}\left(\hat{\Sigma}_{x,NT}^{(h)}\right)^{-1}\right) \\
&\quad \left(\hat{\Phi} - \frac{\eta}{H}\sum_{h=1}^H \hat{F}^{(h)\top}\left(\hat{F}^{(h)}\hat{\Phi} - F_\star^{(h)}\Phi_\star\right) - \frac{\eta}{H}\sum_{h=1}^H \hat{F}^{(h)\top}W^{(h)\top}X^{(h)}\left(\hat{\Sigma}_{x,NT}^{(h)}\right)^{-1}\right)^\top \\
&\succeq I_r - \mathrm{Sym}\left(\frac{\eta}{H}\sum_{h=1}^H \hat{F}^{(h)\top}\left(\hat{F}^{(h)}\hat{\Phi} - F_\star^{(h)}\Phi_\star\right)\hat{\Phi}^\top\right) - \mathrm{Sym}\left(\frac{\eta}{H}\sum_{h=1}^H \hat{F}^{(h)\top}W^{(h)\top}X^{(h)}\left(\hat{\Sigma}_{x,NT}^{(h)}\right)^{-1}\hat{\Phi}^\top\right) \\
&\quad + \mathrm{Sym}\left(\frac{\eta^2}{H^2}\hat{F}^{(h)\top}\left(\hat{F}^{(h)}\hat{\Phi} - F_\star^{(h)}\Phi_\star\right)^\top\left(\sum_{h=1}^H \hat{F}^{(h)\top}W^{(h)\top}X^{(h)}\left(\hat{\Sigma}_{x,NT}^{(h)}\right)^{-1}\right)^\top\right),
\end{aligned}$$

where we discarded the pd diagonal terms of the expansion. Focusing on the first cross term, we have

$$\begin{aligned}
&\mathrm{Sym}\left(\frac{\eta}{H}\sum_{h=1}^H \hat{F}^{(h)\top}\left(\hat{F}^{(h)}\hat{\Phi} - F_\star^{(h)}\Phi_\star\right)\hat{\Phi}^\top\right) \\
&= \frac{\eta}{H}\sum_{h=1}^H \mathrm{Sym}\left(\hat{F}^{(h)\top}\hat{F}^{(h)} - \hat{F}^{(h)\top}F_\star^{(h)}\Phi_\star\hat{\Phi}^\top\right) \\
&= \frac{\eta}{H}\sum_{h} \mathrm{Sym}\left(\hat{F}^{(h)\top}\left(F_\star^{(h)}\Phi_\star\left(I_{d_x} - \hat{\Phi}^\top\hat{\Phi}\right)X_{\mathcal{N}_1}^{(h)\top}Z_{\mathcal{N}_1}^{(h)}\left(\hat{\Sigma}_{z,\mathcal{N}_1T}^{(h)}\right)^{-1} + W_{\mathcal{N}_1}^{(h)\top}Z_{\mathcal{N}_1}^{(h)}\left(\hat{\Sigma}_{z,\mathcal{N}_1T}^{(h)}\right)^{-1}\right)\right).
\end{aligned}$$

Conditioning on the high probability bound events of Lemma 5 and Lemma 6, we repeat a similar analysis as Lemma 9 to yield

$$\sigma_{\min}(R)^2 \ge 1 - \frac{4c\eta}{H}\sum_h \left\|F_\star^{(h)}\right\|^2.$$

Therefore, for appropriately chosen numerical constant $c > 0$, we have

$$\sigma_{\min}(R)\left\|I_{d_x} - \frac{\eta}{H}\sum_{h=1}^H \hat{F}^{(h)\top}\hat{F}^{(h)}\right\| \le \left(1 - \frac{3\eta}{16}\frac{1}{H}\sum_{h=1}^H \sigma_{\min}(F_\star^{(h)})^2\right)^{1/2}.$$

Combining the above lemmas culminates in the final descent guarantee.

Theorem 3 (Full version of Theorem 2) *Assume* $|\mathcal{N}_1|T \gtrsim \max\left\{\gamma^2, \frac{\sigma_w^{(h)2}}{\lambda_{\min}(\Sigma_x^{(h)})}\frac{\|F_\star\|^2}{\sigma_{\min}(F_\star^{(h)})^4}\right\}(\max\{d_y, r\} + \log(H/\delta))$, *and* $\left\|\hat{\Phi}\Phi_{\star,\perp}^\top\right\| \le \min\left\{\frac{3}{16}\frac{1}{\kappa\left(\Sigma_x^{(h)}\right)+1}\frac{\sigma_{\min}(F_\star^{(h)})^2}{\|F_\star\|}, 1/2\right\}$. *Let* $\eta \le \min_h 1/4\|F_\star^{(h)}\|^2$. *There exists numerical constant* $C > 0$ *such that the following holds: with probability at least* $1-\delta$,

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

A.3 The Non-IID Setting

To extend our analysis to the non-iid setting, we first instantiate our covariates as β-mixing stationary processes [43, 45].

Assumption 2 (Geometric mixing) *For each h, assume the process $\{x_i^{(h)}[t]\}_{t\geq 1}$ is a mean-zero stationary β-mixing process, with stationary covariance $\Sigma_x^{(h)}$ 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 set of N independent trajectories, where each trajectory is subsampled into a trajectories of length m (we assume $T = ma$ for notational convenience). We may then apply the analysis of the iid setting on a deflated dataset of HNm data points *drawn from the respective stationary distributions* to yield:

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

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

Now applying Lemma 3, setting $g(\cdot)$ as the indicator function for the burn-in requirement and the final descent bound, we have for all $j = 1, \ldots, a$.

$$\left|\mathbb{E}\left[g\left(\left\{X_\infty^{(h),Nm}\right\}_{h\in[H]}\right)\right] - \mathbb{E}\left[g\left\{X_{(j)}^{(h),NT}\right\}_{h\in[H]}\right]\right| \leq m\beta(a) \leq \delta'$$

Setting $\delta' = \delta/a$ and union bounding over each $j = 1, \ldots, a$, we may invert $T\beta(a) = \delta$ to find $a := \log\left(\frac{\Gamma T}{\delta}\right)/\log\left(\frac{1}{\mu}\right)$. This yields the final descent guarantee.

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

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

A.4 Converting to Sample Complexity Bounds

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

Lemma 11 *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

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

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 11 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 $\frac{C\mathrm{polylog}(M_k)}{M_k}$. However, these will only lead to inflating d_K by a $\mathrm{polylog}(M)$ factor.

In essence, Lemma 11 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 ρ. Instantiating Lemma 11 with the problem parameters of Theorem 2 yields Corollary 1.

A.4.1 Near-ERM Transfer Learning

An important consequence of Lemma 11 (thus Corollary 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 $\hat{F}^{(h)}$ computed with respect to some independent dataset of size NT,

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

where the last line follows from applying Lemma 5. We observe that the parameter error nicely decomposes into a term quadratic in $\mathrm{dist}(\hat{\Phi},\Phi_\star)$ and least squares fine-tuning error scaling with $\frac{1}{NT}$. For a fixed dataset of size HNT, one can crudely set aside $\Theta(NT)$ samples for each task, and use the rest of the $\Theta(HNT)$ samples to compute $\hat{\Phi}$. Invoking Corollary 1 and using the set-aside $\Theta(NT)$

samples to compute $\hat{F}^{(h)}$ conditioned on $\hat{\Phi}$, we recover the near-ERM high probability generalization bound on the parameter error

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

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., [50, 53–57]), as well as (online) statistical learning problems with temporally dependent covariates [40, 41, 58–66] (see [67] 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* [66, 68, 69], 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

$$\begin{aligned} &s[t+1] = A^{(h)} s[t] + B^{(h)} u[t] + w[t] \\ &w[t] \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_w^{(h)}), \quad s[0] \sim \mathcal{N}(0, \Sigma_0^{(h)}), \end{aligned} \tag{29}$$

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:

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

It is customary to consider exploratory signals that are iid zero-mean Gaussian random vectors $u[t] \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_u^{(h)})$ [58, 66, 67]. 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 $u[t] \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_u^{(h)})$ by plugging in (29) into the following equation

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

Therefore, evaluating the stationary state covariance $\Sigma_s^{(h)} := \mathbb{E}\left[s[\infty]s[\infty]^\top\right]$ 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 $y[t] \leftarrow s[t+1]$, $x[t] \leftarrow \begin{bmatrix} s[t] \\ u[t] \end{bmatrix}$, $M_\star^{(h)} \leftarrow \begin{bmatrix} A^{(h)} & B^{(h)} \end{bmatrix}$, we may instantiate multi-task linear system identification as a non-iid, non-isotropic linear operator recovery problem.

Definition 4 *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*

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

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 3 *We assume that for any task h the following hold:*

1. *The operators share a rowspace $M_\star^{(h)} := \begin{bmatrix} A^{(h)} & B^{(h)} \end{bmatrix} = F_\star^{(h)} \Phi_\star$, $F_\star^{(h)} \in \mathbb{R}^{d_s \times r}$, $\Phi_\star \in \mathbb{R}^{r \times (d_s + d_u)}$.*
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*

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

The existence of a uniform Γ' is guaranteed by Gelfand's Formula [52], and quantitative bounds may be found in, e.g., [50, 70].

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 [50].

Proposition 4 (Adapted from Tu and Recht [50, Prop. 3.1]) *For each h, let the dynamics for the linear system evolve as described in (29). Let Assumption 3 hold with constants Γ', ρ. Define $\mathbb{P}_{s[k] \sim \nu_k}[\,\cdot \mid s_0 = s]$ 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 ν_0,*

$$\mathbb{E}_{s \sim \nu_0}\left[\left\|\mathbb{P}_{s[k]}[\,\cdot \mid s_0 = s] - \mathbb{P}_{s[k]}\right\|_{\mathrm{tv}}\right] \leq \frac{\Gamma'}{2}\sqrt{\mathbb{E}_{\nu_0}[\|s[0]\|^2] + \frac{\|\Sigma^{-1}\|_*}{1-\mu^2}} \cdot \mu^k, \tag{30}$$

where $\|\cdot\|_*$ indicates the nuclear norm [52], 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 $s[t]$.

$$\left\|\mathbb{P}_{s[k]}[\,\cdot \mid s_0 = s] - \mathbb{P}_{s[k]}\right\|_{\mathrm{tv}} = \left\|\mathbb{P}_{x[k]}[\,\cdot \mid s_0 = s] - \mathbb{P}_{x[k]}\right\|_{\mathrm{tv}}$$

Since by construction the marginal distribution of states is identically $\mathcal{N}(0, \Sigma_s^{(h)})$, applying Proposition 4 to $s[t], s[t+k]$ for any t, k, we get the following quantitative bound on the mixing-time of the covariates $x[t] = \begin{bmatrix} s[t]^\top & u[t]^\top \end{bmatrix}^\top$.

Lemma 12 *Following Definition 4 and Assumption 3, the covariate process $\left\{x^{(h)}[t]\right\}_{t \geq 0}$ is a mean-zero, stationary, geometrically β-mixing process with covariance $\Sigma_x^{(h)} = \begin{bmatrix} \Sigma_s^{(h)} & 0 \\ 0 & \Sigma_u^{(h)} \end{bmatrix}$, where $\Sigma_s^{(h)} = \mathrm{dlyap}(A^{(h)}, B^{(h)} \Sigma_u^{(h)} B^{(h)} + \Sigma_w^{(h)})$, and mixing-time bounded by*

$$\begin{aligned} &\beta(k) = \Gamma \mu^k, \quad \textit{where} \\ &\Gamma := \frac{\Gamma'}{2}\sqrt{\mathrm{Tr}\left(\Sigma_s^{(h)}\right) + \frac{\|\Sigma^{-1}\|_*}{1-\mu^2}}, \ \Sigma := B^{(h)} \Sigma_u^{(h)} {B^{(h)}}^\top + \Sigma_w^{(h)}. \end{aligned} \tag{31}$$

Thus, instantiating Lemma 12 in Theorem 4 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

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

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 (29) yields that the states/covariates evolve as

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

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:

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

Analogously to linear sysID, we make the following assumptions.

Assumption 4 *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*
$$\mathbb{E}\left[x[t]x[t]^\top\right] = \Sigma_s^{(h)} =: \Sigma_x^{(h)}, \text{ for all } t \geq 0.$$
2. *The controllers share a rowspace $M_\star^{(h)} \equiv K_\star^{(h)} = F_\star^{(h)} \Phi_\star$, $F_\star^{(h)} \in \mathbb{R}^{d_u \times r}$, $\Phi_\star \in \mathbb{R}^{r \times d_s}$.*
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 Γ' is guaranteed by Gelfand's Formula [52].

By using a result almost identical to Proposition 4, we yield the following quantitative bound on the mixing time of covariates generated by stabilizing expert controllers.

Lemma 13 *Following Assumption 4, the covariate process $\left\{x^{(h)}[t]\right\}_{t \geq 0}$ is a mean-zero, stationary, geometrically β-mixing process with covariance $\Sigma_x^{(h)} = \Sigma_s^{(h)}$, where $\Sigma_s^{(h)} = \operatorname{dlyap}\left(A^{(h)} + B^{(h)} K_\star^{(h)}, B^{(h)} \Sigma_z^{(h)} {B^{(h)}}^\top + \Sigma_w^{(h)} \right)$, and mixing-time bounded by*

$$\begin{aligned} &\beta(k) = \Gamma \mu^k, \quad \textit{where} \\ &\Gamma := \frac{\Gamma'}{2} \sqrt{\operatorname{Tr}\left(\Sigma_s^{(h)}\right) + \frac{\|\Sigma^{-1}\|_*}{1 - \mu^2}}, \ \Sigma := B^{(h)} \Sigma_z^{(h)} {B^{(h)}}^\top + \Sigma_w^{(h)}. \end{aligned} \tag{32}$$

Thus, instantiating Lemma 12 in Theorem 4 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 γ 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 η 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` [1]. For example, to obtain the results depicted in Figure 1a, 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 1b. In Figure 2a, both the single-task and multi-task implementations of Algorithm 1 adopt $\eta = 7,5 \times 10^{-3}$, whereas the vanilla alternating minimization-descent approach uses $\eta = 7.5 \times 10^{-3}$ for a fair comparison. Similarly, in Figure 2b, 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 2b, we present the results for an extended range of tasks using Algorithm 1 and the alternating minimization-descent approach (`FedRep` [1]). 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 3 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 H 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 ($H = 5$), 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 operators different from those utilized to obtain the results illustrated in Figure 2b. 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 η is set to 1×10^{-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 η to 2×10^{-3}.

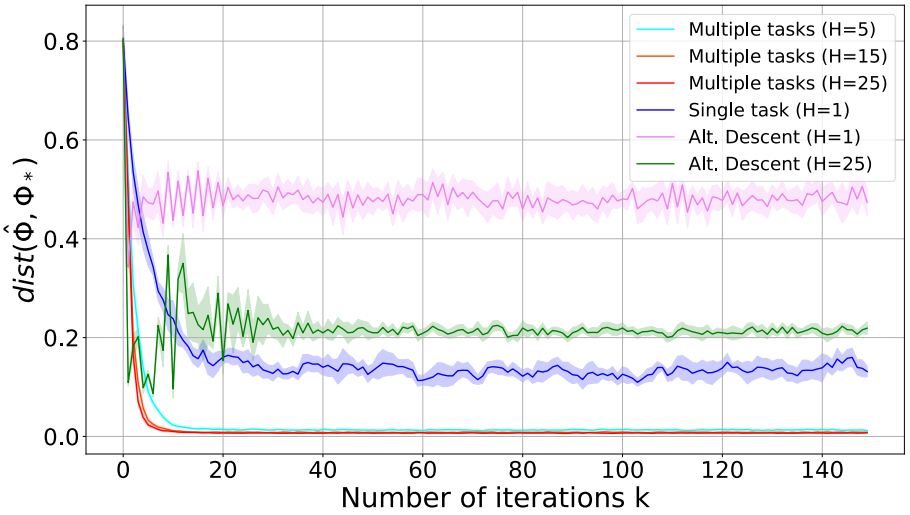


Figure 3: We plot the subspace distance between the current and ground truth representation with respect to the number of iterations, comparing between the single and multiple-task settings of Algorithm 1 and the multi-task `FedRep` for the IID linear regression with random covariance. We observe performance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

In alignment with our main theoretical findings, Figure 4 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 3 for the linear regression problem, Figure 4 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 $K^{(H+1)} = F_\star^{(H+1)}\Phi_\star$, by imitating the behavior of H expert controllers $K^{(1)}, K^{(2)}, \ldots, K^{(H)}$. These controllers share a common low-rank representation and can be decomposed into the form $K^{(h)} = F_\star^{(h)}\Phi_\star$, where $F_\star^{(h)}$ represents the task-specific weight and $\Phi_\star$ 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:

$$x^{(h)}[t+1] = Ax^{(h)}[t] + Bu^{(h)}[t], \quad t = 0, 1, \ldots, T-1,$$

with $n_x = 4$ states and $n_u = 4$ inputs, for all $h \in [H+1]$, where $u^{(h)}[t] = K^{(h)}x^{(h)}[t] + z^{(h)}[t]$, with $z^{(h)}[t] \sim \mathcal{N}(0, I_{n_u})$ 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 $y^{(h)}[t] = Gx^{(h)}[t] + w^{(h)}[t]$, 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.

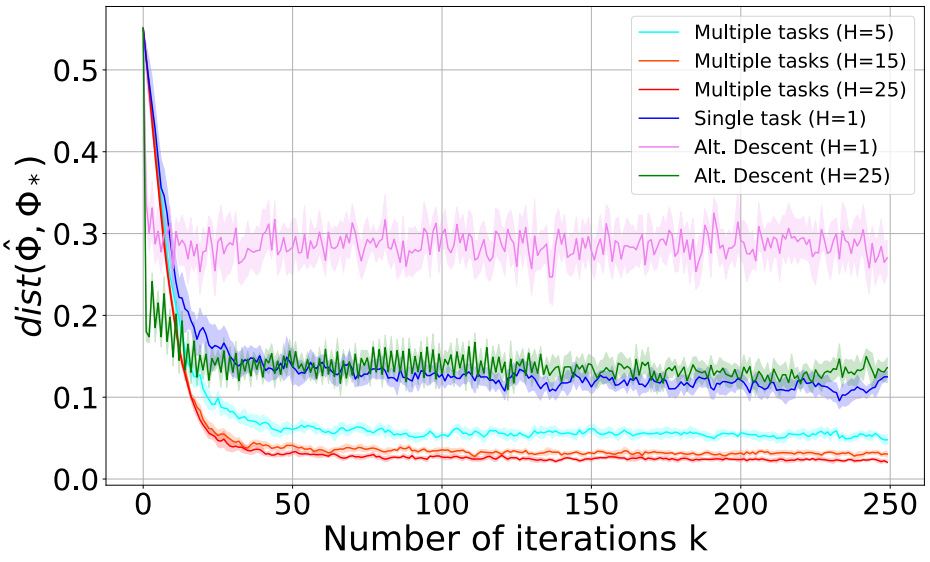


Figure 4: We plot the subspace distance between the current and ground truth representation with respect to the number of iterations, comparing between the single and multiple-task settings of Algorithm 1 multi-task `FedRep` for the linear system identification with random covariance. We observe performance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

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

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

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

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

Figure 5 presents a comparative analysis between Algorithm 1 and the vanilla alternating minimization-descent approach (`FedRep` in [1]) for computing a shared representation across linear quadratic regulators. This shared representation is then utilized to derive the learned controller $K^{(H+1)}$ in a few-shot learning manner. Consistent with our theoretical findings and in alignment with the trends observed in Figures 3-4, Figure 5 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.

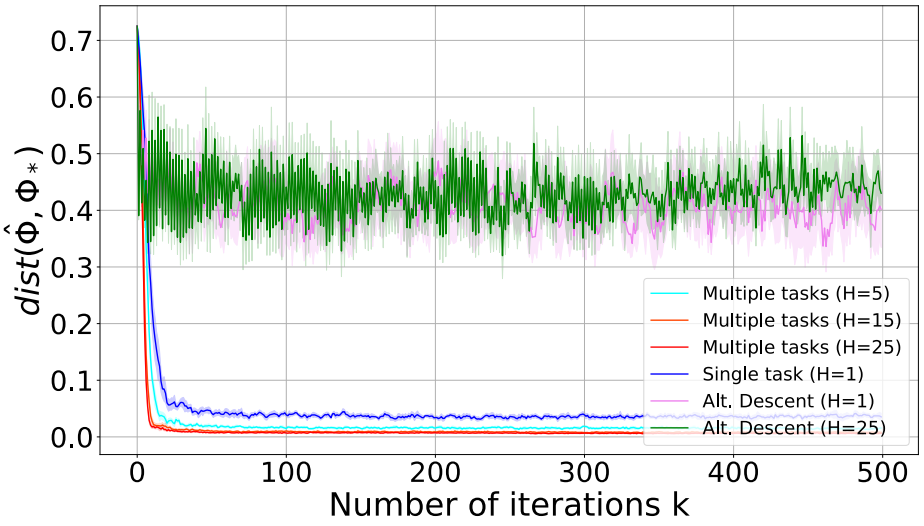


Figure 5: We plot the subspace distance between the current and ground truth representation with respect to the number of iterations, comparing between the single and multiple-task settings of Algorithm 1 and the multi-task `FedRep` for the imitation learning with random covariance. We observe performance improvement and variance reduction for multi-task `DFW` as predicted.