Guarantees for Nonlinear Representation Learning: Non-identical Covariates, Dependent Data, Fewer Samples

Thomas T. Zhang [1] **Bruce D. Lee** [1] **Ingvar Ziemann** [1] **George J. Pappas** [1] **Nikolai Matni** [1]

Abstract

A driving force behind the diverse applicability of modern machine learning is the ability to extract meaningful features across many sources. However, many practical domains involve data that are non-identically distributed across sources, and possibly statistically dependent within its source, violating vital assumptions in existing theoretical studies of representation learning. Toward addressing these issues, we establish statistical guarantees for learning general *nonlinear* representations from multiple data sources that admit different input distributions and possibly dependent data. Specifically, we study the sample-complexity of learning $T+1$ functions $f_\star^{(t)} \circ g_\star$ from a function class $\mathcal{F} \times \mathcal{G}$, where $f_\star^{(t)}$ are task specific linear functions and $g_\star$ is a shared nonlinear representation. An approximate representation $\hat{g}$ is estimated using N samples from each of T source tasks, and a fine-tuning function $\hat{f}^{(0)}$ is fit using N' samples from a target task passed through $\hat{g}$. Our results show that the excess risk of the estimate $\hat{f}^{(0)} \circ \hat{g}$ on the target task decays as $\tilde{\mathcal{O}}\left(\frac{\mathrm{C}(\mathcal{G})}{NT} + \frac{\dim(\mathcal{F})}{N'}\right)$, where $\mathrm{C}(\mathcal{G})$ denotes the complexity of $\mathcal{G}$. Notably, our rates match that of the iid setting, while requiring fewer samples per task than prior analysis and admitting *no dependence on the mixing time*. We support our analysis with numerical experiments performing imitation learning over non-linear dynamical systems.

1. Introduction

Transfer learning, in which a model is pre-trained on a large dataset, and then finetuned for a specific application, has shown great success in various fields of machine learning including computer vision (Dosovitskiy et al., 2021) and natural language processing (Devlin et al., 2019). The principle enabling the success of these approaches is the use of a large dataset to extract compressed features which are broadly useful for downstream tasks. The extraction of such generally useful features from data is referred to as *representation learning* (Bengio et al., 2013). Despite its critical role in the success of deep learning, statistical guarantees remain somewhat limited.

Only recently have studies formalized multi-task representation learning in a way that illustrates how generalization improves when data is aggregated across many tasks (Du et al., 2020; Tripuraneni et al., 2020). These regression settings consider learning $T+1$ functions $f_\star^{(t)} \circ g_\star$ in a function class $\mathcal{F} \times \mathcal{G}$ from covariate-observation pairs $\{(x_i^{(t)}, y_i^{(t)})\}$, where $f_\star^{(t)}$ are task-specific functions, and $g_\star$ is a shared representation. The tasks for $t = 1, \ldots, T$ are denoted source (training) tasks, while $t = 0$ is the target (test) task. A basic model of transfer learning can be expressed as a two-step procedure in which an estimate $\hat{g}$ for the representation is determined by solving a least squares problem using N data samples from each of the source tasks with measurements corrupted by zero-mean noise. This representation is then used to determine an estimate $\hat{f}^{(0)}$ by solving a least squares problem using N' samples from the target task, also with measurements corrupted by zero-mean noise. Du et al. (2020); Tripuraneni et al. (2020) show generalization bounds on the learned predictor in which the excess risk scales as $\tilde{\mathcal{O}}\left(\frac{\mathrm{C}(\mathcal{G})}{NT} + \frac{\mathrm{C}(\mathcal{F})}{N'}\right)$, where C quantifies the complexity of a function class. These rates capture the desirable behavior where the error from fitting the shared representation decays with the *total* amount of data aggregated across the T source tasks.

While a rather complete picture can be stitched for linear settings, for such rates to hold in settings where the representation class $\mathcal{G}$ is nonlinear, prior work crucially relies upon the assumption that covariates are independent and identically distributed (iid) across all tasks, such that the only source of variation comes from the task-specific $f_\star^{(t)}$. Such assumptions are fundamentally incompatible with many potential use cases of multi-task representation learning, such as in domain generalization and sequential decision-making.

[1]Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, USA. Correspondence to: Thomas Zhang <ttz2@seas.upenn.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

A key goal of this work is to remedy this issue and achieve multi-task rates in the absence of assumptions requiring identical covariate distributions across tasks, and independent data within tasks.

1.1. Related Work

Multi-task linear regression: Beginning with Du et al. (2020), a fairly complete picture has emerged in the setting of multi-task linear regression in which distinct tasks share a low dimensional representation, i.e. $f_\star^{(t)}(z) = Fz$ and $g(x) = \Phi x$ for some matrices $F \in \mathbb{R}^{d_{\mathsf{Y}} \times r}$ and $\Phi \in \mathbb{R}^{r \times d_{\mathsf{X}}}$ with $r \leq d_{\mathsf{X}}$[1]. In this setting, the authors demonstrate that the excess risk achieved by the empirical risk minimizer (ERM) achieves rates $\tilde{\mathcal{O}}\left(\frac{d_{\mathsf{X}} r}{NT} + \frac{d_{\mathsf{Y}} r}{N}\right)$. An active learning setting is considered by Chen et al. (2022); Wang et al. (2023), in which the assumption of uniform sampling from each task is replaced with an adaptive sampling algorithm. Chua et al. (2021) considers a setting in which the representation is fine-tuned for each task, thereby allowing the assumption of a shared Φ to be relaxed. Crucially, all of the aforementioned bounds hold only if the minimum amount of data ("burn-in time") per task exceeds a quantity proportional to d_{X}. This is counterintuitive, as a goal of aggregating data across tasks is to remove the necessity for many samples per task. Furthermore, solving the ERM is nominally a non-convex bilinear problem. Efficient algorithms to bypass ERM have been proposed to explicitly address these issues (Tripuraneni et al., 2021; Collins et al., 2021; Thekumparampil et al., 2021), while attaining same rates order-wise. The resulting analysis alleviates the dependence of the burn-in time on d_{X}, but iid covariates across tasks, such that their estimators are consistent without requiring standardizing data per-task which would otherwise reintroduce a $\approx d_{\mathsf{X}}$ burn-in per task. This is partially resolved by an algorithm proposed in Zhang et al. (2023b), which handles tasks with non-identical covariate distributions; however the burn-in remains proportional to d_{X}. These results beg the question: is the d_{X} per-task burn-in for ERM fundamental or a technical byproduct? A d_{X} burn-in is unintuitive, since given an optimal representation $\Phi_\star$, solving each task is precisely standard linear regression over r-dimensional covariates $z \triangleq \Phi_\star x$, for which the burn-in is much more lenient $\approx r$ (Wainwright, 2019).

Non-linear multi-task learning: Early works consider statistical guarantees for multi-task learning over general nonlinear function classes (Baxter, 2000; Ben-David & Borbely, 2008; Maurer et al., 2016; Hanneke & Kpotufe, 2022); however, they do not obtain rates scaling jointly in N, T due to the data model or assuming agnostic settings. (Du et al., 2020; Tripuraneni et al., 2020; Watkins et al., 2024) provide excess risk bounds in which the error benefits jointly in N and T by assuming a *shared representation*. Du et al. (2020) considers a setting in which $\mathcal{F}$ is a linear function class, and $\mathcal{G}$ is a nonlinear function class. Tripuraneni et al. (2020); Watkins et al. (2024) consider nonlinear $\mathcal{F}$ and $\mathcal{G}$; however, the resulting generalization bounds scale with diameter of covariate distributions rather than with noise-level (Du et al., 2020). These works all assume marginal covariate distributions are identical across all tasks, and the final bounds involve data-dependent complexity terms. When instantiated in linear settings, their guarantees recover suboptimal burn-ins *at least* order-d_{X} samples per task. The aforementioned results study the ERM solution rather than feasible algorithms. Meunier et al. (2023) is a notable exception, providing a feasible algorithm in the setting of Reproducing Kernel Hilbert Spaces (RKHS), in which tasks share a RKHS subspace projection.

Multi-task sequential learning: Multi-task learning has been applied to many dynamical systems settings, such as robotic manipulation (Brohan et al., 2022; Shridhar et al., 2023), agile flight (O'Connell et al., 2022), robotic locomotion . Despite its effectiveness in practice, the existing theoretical guarantees for representation learning do not apply to these settings due to their assumption that covariates are iid across tasks. Notably, when the predictor is either the dynamics function or a closed-loop control policy, the covariate distribution is inextricably linked with the predictor itself. Consider, for example, a stable autonomous system driven by white noise, $y_t \triangleq x_{t+1} = Ax_t + w_t$ with $x_0, w_t \sim \mathcal{N}(0, I_{d_{\mathsf{X}}})$. The stationary covariate distribution is inextricably linked to the "predictor" A, as demonstrated by solving the Lyapunov equation

$$\begin{aligned} \Sigma_x &\triangleq \mathbb{E}[x_{t+1} x_{t+1}^\top] = A \Sigma_x A^\top + I_{d_{\mathsf{X}}} \\ \implies \Sigma_x &= \sum_{k \geq 0} A^k (A^k)^\top. \end{aligned}$$

Therefore, in multi-task settings where multiple distinct predictors are involved, the covariate distributions will be non-identical between tasks. Furthermore, covariates generated by dynamical systems are correlated across time. These issues have been remedied in the linear setting by Modi et al. (2021); Zhang et al. (2023a) applied to system identification and imitation learning by extending the analysis of Du et al. (2020) to consider covariates generated by linear systems. For the single-task non-linear regression setting, Ziemann & Tu (2022); Ziemann et al. (2023a) demonstrate that learning in the presence of correlated covariates achieves the same rate as learning in the absence of correlation, only inflating the burn-in time by the correlation level. However, we are unaware of extensions of these results to multi-task representation learning.

In parallel, various works have considered extensions of the aforementioned multi-task linear regression works to linear bandit settings (Yang et al., 2022; 2020; Du et al.,

[1] Du et al. (2020) consider a scalar setting $d_{\mathsf{Y}} = 1$. Their analysis is extended to vector-valued settings in Zhang et al. (2023a).

2023; Mukherjee et al., 2023). We also note works studying reinforcement learning (RL) with feature approximation (or low-rank Markov decision processes (MDPs) (Agarwal et al., 2020; Jin et al., 2020; Uehara et al., 2021; Efroni et al., 2022; Du et al., 2019) which attain sample complexity gains from dimensionality reduction, but do not consider aggregating data across multiple tasks. Analysis of multi-task representation learning for MDPs has been studied by Arora et al. (2020); Lu et al. (2021); however these works assume generative models, thereby sidestepping the issues of independent data and non-identical covariates.

1.2. Contributions

In this work, we analyze the transfer learning problem in a setting where $\mathcal{F}$ is a class of linear functions mapping $\mathbb{R}^r$ to $\mathbb{R}^{d_{\mathsf{Y}}}$, and $\mathcal{G}$ is a class of nonlinear representations, as in (Du et al., 2020)[2]. In this setting, we remove assumptions of both identical covariate distributions and independent covariates within tasks, and we additionally improve the per-task burn-in requirement. We list our specific contributions:

- We derive generalization bounds that hold for non-identical covariate distributions and vector-valued measurements. In particular, we present an updated "task-diversity" measure, which takes into account overlap of non-identical covariate distributions, in addition to the similarity of linear heads $f_\star^{(t)}$ (e.g. Du et al. (2020); Zhang et al. (2023a)).
- We show our proposed bounds on ERM scale multiplicatively with noise level and jointly with number of tasks and per-task samples as in Du et al. (2020), while requiring only $\Omega(d_{\mathsf{Y}} r)$ samples per task (noting $d_{\mathsf{Y}} = 1$ in most prior work, e.g. Du et al. (2020); Tripuraneni et al. (2020)), as opposed to $\Omega(d_{\mathsf{X}})$.
- We extend our bounds to (within-task) dependent data. Adapting ideas from recent work (Ziemann & Tu, 2022; Ziemann et al., 2023b), we demonstrate that when task covariates are ϕ-mixing, our generalization bounds scale with the independent-data rate. In particular, we avoid the effective sample-size deflation incurred by standard blocking techniques (Yu, 1994; Kuznetsov & Mohri, 2017), relegating the effect of mixing to a mild increase of burn-in.

Notably, via our contributions, the guarantees in this work can be lifted from offline regression to various sequential decision-making settings, such as nonlinear system identification (Mania et al., 2022; Wagenmaker et al., 2023) and stochastic contextual bandits (Foster & Rakhlin, 2020; Simchi-Levi & Xu, 2022). Stating our main theoretical result informally:

Theorem 1.1 (Main result, informal)**.** *Assume $N \geq C_{\mathrm{mix}} \Omega\left(d_{\mathsf{Y}} r + \mathrm{C}(\mathcal{G})/T\right)$, where C_{mix} characterizes the dependency of the covariates of each task. Then the excess transfer risk of ERM is bounded with high-probability:*

$$\mathrm{ER}(\hat{f}^{(0)}, \widehat{g}) \lesssim C_{\text{task div}} \sigma^2 \left(\frac{\mathrm{C}(\mathcal{G})}{NT} + \frac{d_{\mathsf{Y}} r}{N} \right),$$

where $C_{\text{task div}}$ characterizes the relatedness between the source tasks and the target task and σ^2 characterizes the level of the noise corrupting the measurements.

Notation Expectation (resp. probability) with respect to the underlying probability space is denoted by E (resp. P). For two probability measures P and Q defined on the same probability space, their total variation is denoted $\|\mathsf{P} - \mathsf{Q}\|_{\mathsf{TV}}$. For an integer $n \in \mathbb{N}$, we also define the shorthand $[n] \triangleq \{1, \ldots, n\}$. The Euclidean norm on $\mathbb{R}^d$ is denoted $\|\cdot\|_2$, and the unit sphere in $\mathbb{R}^d$ is denoted $\mathbb{S}^{d-1}$. We also write $\|M\|_2$ for the spectral norm. For two symmetric matrices M, N, we write $M \succ N$ ($M \succeq N$) if $M - N$ is positive (semi-)definite. We use $\lesssim, \gtrsim$ to omit universal numerical factors, and $\tilde{\mathcal{O}}(\cdot)$, $\tilde{\Theta}(\cdot)$, $\tilde{\Omega}(\cdot)$ to omit polylog factors.

Samples In general, we index tasks by superscript while *within-task* samples are indexed by subscript, e.g. $x_i^{(t)}$ for task t and sample i. Let $\mathsf{P}_i^{(t)}, t \in [T]$ be probability measures over a fixed sample space S. We are given N samples from each "training task" t: $s_i^{(t)} \sim \mathsf{P}_i^{(t)}, t \in [T], i \in [N]$. For convenience, we overload notation and understand $\mathsf{P}^{(t)}$ alternatively refers to the stationary distribution when $s_i^{(t)}$ are identically distributed or to a joint *trajectory* distribution $\{s_i^{(t)}\}_{i=1}^N \sim \mathsf{P}^{(t)}$ otherwise. We use superscript $^{1:T}$ to denote a uniform mixture, e.g. $\mathsf{P}^{1:T} \triangleq \frac{1}{T}\sum_{t=1}^T \mathsf{P}^{(t)}$. We consider supervised learning: the sample space decomposes into an input (covariate) space X and and output (label) space Y: $\mathsf{S} = \mathsf{X} \times \mathsf{Y}$ and we write $s_i^{(t)} = (x_i^{(t)}, y_i^{(t)})$. Moreover, we are given N' samples from a target task distributed according to a probability measure $\mathsf{P}^{(0)}$ over Z: $(x_i^{(0)}, y_i^{(0)}) \sim \mathsf{P}_i^{(0)}, i \in [N']$. It will also be convenient to introduce empirical counterparts $\widehat{\mathsf{P}}_N^{(t)}, \widehat{\mathsf{E}}_N^{(t)}$, such that e.g. $\widehat{\mathsf{E}}_N^{(t)}[f(X)] = \frac{1}{N}\sum_{i=1}^N f(x_i^{(t)})$. We generally denote covariance matrices[3] by Σ, e.g. $\Sigma_x^{(t)} \triangleq \mathsf{E}^{(t)}[XX^\top]$.

1.3. Problem Formulation

Given the above definitions of training and test/transfer distributions, we consider a prototypical regression problem, where the goal of the learner is to perform well on the target task in terms of square loss over a fixed hypothesis class $\mathcal{H}$. To enable transfer and characterize the benefits of representation learning, we assume that the hypothesis class $\mathcal{H}$ under consideration splits into $\mathcal{H} = \mathcal{F} \times \mathcal{G}$. We define the optimal training-task predictors:

[2] This is a prototypical predictor model found across many domains, e.g. RL with feature approximation, nonlinear least squares.

[3] To be precise, second moment matrices.

$$(\{f_\star^{(t)}\}_{t=1}^T, g_\star) \in \operatorname*{argmin}_{\substack{(\{f^{(t)}\},g)\\ \in \mathcal{F}^{\otimes T}\times\mathcal{G}}} \sum_{t=1}^T \mathsf{E}^{(t)}\|f^{(t)}\circ g(X) - Y\|_2^2.$$

Hence, to each task $t \in [T]$ we associate a task-specific "head" $f_\star^{(t)} \in \mathcal{F}$, while enforcing a shared "representation" $g_\star \in \mathcal{G}$. We further denote the optimal target-task head: $f_\star^{(0)} \in \operatorname{argmin}_{f\in\mathcal{F}} \mathsf{E}^{(0)}\|f\circ g_\star(X) - Y\|_2^2$. Using our samples from both target and training tasks, we seek to find an element $(f, g) \in \mathcal{F}\times\mathcal{G}$ that renders the excess risk on the target distribution as small as possible:

$$\begin{aligned}&\mathrm{ER}(f, g) \qquad (1)\\ &\triangleq \mathsf{E}^{(0)}\|f\circ g(X) - Y\|_2^2 - \mathsf{E}^{(0)}\|f_\star^{(0)}\circ g_\star(X) - Y\|_2^2.\end{aligned}$$

In particular, we study the excess risk of a standard two-stage empirical risk minimization scheme (Du et al., 2020; Tripuraneni et al., 2020), where a representation $\widehat{g} \in \mathcal{G}$ is fit on data from the T training tasks, and a target-task head $\hat{f}^{(0)}$ is fit on the target task data, *passed through* $\widehat{g}$:

$$\begin{aligned}(\{\hat{f}^{(t)}\}_{t=1}^T, \hat{g}) &\in \operatorname*{argmin}_{\mathcal{F}^{\otimes T}\times\mathcal{G}} \sum_{t=1}^T\sum_{i=1}^N \|f^{(t)}\circ g(x_i^{(t)}) - y_i^{(t)}\|_2^2\\ \hat{f}^{(0)} &\in \operatorname*{argmin}_{f\in\mathcal{F}} \sum_{i=1}^{N'} \|f\circ\widehat{g}(x_i^{(0)}) - y_i^{(0)}\|_2^2. \qquad (2)\end{aligned}$$

Though our work mostly concerns the statistical properties of ERM, we note that in many practical settings with expressive $\mathcal{G}$, the empirical loss on a given dataset can be effectively optimized, and the error incurred by an algorithm enters as an additive factor in the generalization bounds (Vaskevicius et al., 2020). Toward characterizing bounds on the above excess risk, we consider vector-valued inputs and outputs $\mathsf{X}\times\mathsf{Y} \subseteq \mathbb{R}^{d_\mathsf{X}}\times\mathbb{R}^{d_\mathsf{Y}}$. Following prior work, we consider the *realizable* setting, i.e. there exist $(\{f_\star^{(t)}\}_{t=0}^T, g_\star)$ such that the noise term $W^{(t)} \triangleq Y^{(t)} - f_\star^{(t)}\circ g_\star(X^{(t)})$ is a (conditional) zero-mean process for every task.

Assumption 1.2. Given a filtration $\{\mathscr{F}_i^{(t)}\}_{i\geq 1}$ to which $\{x_{i-1}^{(t)}\}_{i\geq 1}$ is adapted, i.e. $x_i^{(t)}$ is predictable with respect to $\mathscr{F}_i^{(t)}$, for each $t \in [T]$, the noise sequence $\{w_i^{(t)}\}_{i\geq 1}$ is a σ_W^2*-conditionally subgaussian martingale difference sequence*:

a) $\mathsf{E}^{(t)}[w_i^{(t)}|\mathscr{F}_{i-1}^{(t)}] = 0$.

b) $\mathsf{E}^{(t)}[\exp(\lambda\langle w_i^{(t)}, v\rangle) \mid \mathscr{F}_{i-1}^{(t)}] \leq \exp(\lambda^2\sigma_W^2/2)$, for all $\lambda\in\mathbb{R}$, $v\in\mathbb{S}^{d_\mathsf{Y}-1}$, $i\geq 1$.

Assumption 1.2 simply asserts the noise is zero-mean subgaussian in the independent setting, with the additional formalism necessary when extending to sequentially dependent settings.

2. Main Results

In this section, we present our main results and the key steps in the proof. Firstly, we present the main definitions and assumptions in Section 2.1, and convert the target-task excess risk to quantities defined over the training tasks. We then instantiate in Section 2.2 a basic setting where $\mathcal{G}$ is finite and within-task samples are iid, but task-wise covariate distributions may be non-identical, $\mathsf{P}_\mathsf{X}^{(t)} \neq \mathsf{P}_\mathsf{X}^{(t')}$, in order to highlight the benefits brought by our analysis. In Section 2.3, we lift our results to general representations $\mathcal{G}$ and settings where within-task samples may be sequentially dependent. In particular, we leverage recent literature to shift the effect of dependency to the burn-in, resulting in rates analogous to the independent data setting.

2.1. Task Diversity and A Canonical Decomposition

A non-vacuous bound on the excess transfer risk is only possible if the source tasks are somehow informative for the target task. Therefore, a pervasive step in establishing a bound lies in relating the risk on the target task to the average risk over the training tasks, where the quality of this relation is determined by a "task-diversity" condition. To make this concrete, we adapt such a condition from (Tripuraneni et al., 2020).

Definition 2.1 (Task-Diversity (Tripuraneni et al., 2020))**.** The training tasks satisfy a task-diversity condition at level $\nu > 0$, if for any $g \in \mathcal{G}$ the following holds:

$$\begin{aligned}&\inf_{f\in\mathcal{F}} \mathrm{ER}(f, g)\\ &\leq \frac{\nu^{-1}}{T}\sum_{t=1}^T \inf_{f^{(t)}} \mathsf{E}^{(t)}\|f^{(t)}\circ g(X) - f_\star^{(t)}\circ g_\star(X)\|_2^2 \qquad (\mathtt{TD})\end{aligned}$$

We then consider a trivial canonical risk decomposition:

$$\begin{aligned}&\mathrm{ER}(f, g)\\ &= \mathsf{E}^{(0)}\|f\circ g(X) - Y\|_2^2 - \inf_{f'\in\mathcal{F}} \mathsf{E}^{(0)}\|f'\circ g(X) - Y\|_2^2\\ &+ \inf_{f'\in\mathcal{F}} \mathsf{E}^{(0)}\|f'\circ g(X) - Y\|_2^2 - \mathsf{E}^{(0)}\|f_\star^{(0)}\circ g_\star(X) - Y\|_2^2.\end{aligned}$$

Applying the task-diversity condition (2.1) to the last line, and observing any plug-in f' upper bounds the infimum yields the following result.

Lemma 2.2. *Let $(\{\widehat{f}^{(t)}\}_{t=0}^T, \hat{g})$ be the output of the two-stage ERM (2). Assuming the task-diversity condition (2.1) holds at level $\nu > 0$, then*

$$\begin{aligned}&\mathrm{ER}(\hat{f}^{(0)}, \widehat{g}) \leq\\ &\underbrace{\mathsf{E}^{(0)}\|\hat{f}^{(0)}\circ\hat{g}(X) - Y\|_2^2 - \inf_{f'\in\mathcal{F}} \mathsf{E}^{(0)}\|f'\circ\hat{g}(X) - Y\|_2^2}_{\text{(Risk of non-realizable regression on targets } \hat{g}(X))} \qquad (3)\\ &+ \underbrace{\frac{\nu^{-1}}{T}\sum_{t=1}^T \mathsf{E}^{(t)}\|\hat{f}^{(t)}\circ\hat{g}(X) - f_\star^{(t)}\circ g_\star(X)\|_2^2}_{\text{(Task-average estimation error of training task predictors)}}. \qquad (4)\end{aligned}$$

As outlined above, the risk of the transfer task predictor can be bounded by two main terms. The former term is

precisely the excess risk of target task head $\hat{f}^{(0)}$ over the r-dimensional inputs $\hat{g}(X)$; since $\hat{g}$ via the two-stage ERM is statistically independent of $\mathsf{P}^{(0)}$, it may be treated as fixed. Since generally $\hat{g} \neq g_\star$, regressing Y against $\hat{g}(X)$ is *non-realizable*. In particular, this breaks the conditional independence between the error $u_i = y_i - F\hat{g}(x_i)$ and covariate x_i. The latter term (4) is the population task-averaged estimation error of the predictors $\widehat{f}^{(t)} \circ \hat{g}$ with respect to optimal $f_\star^{(t)} \circ g_\star$, adjusted by task diversity ν.

Lemma 2.2 and definitions therein thus far hold for general composite classes $\mathcal{F} \times \mathcal{G}$. Toward substantiating bounds on (3) and (4), we consider a prevalent model of non-linear representation learning, where $\mathcal{G}$ is an arbitrary function class that embeds the inputs into a low-dimensional latent space in $\mathbb{R}^r$, and the task-specific heads act linearly on $\mathbb{R}^r$ (Du et al., 2020; Meunier et al., 2023; Collins et al., 2023). Besides being an established theoretical model, we note that last-layer finetuning (alternatively known as linear probing) is known empirically to benefit multi-task / out-of-distribution transfer compared to fine-tuning the full model (Kumar et al., 2022; Lee et al., 2023).

Assumption 2.3 (Low-dim. representations)**.** The representation class $\mathcal{G}$ embeds inputs to $\mathbb{R}^r$, $g : \mathbb{R}^{d_\mathsf{X}} \to \mathbb{R}^r \ \forall g \in \mathcal{G}$, where $r \leq d_\mathsf{X}$. The task-specific head class $\mathcal{F}$ is linear: $\mathcal{F} = \{f : f(z) = Fz,\ F \in \mathbb{R}^{d_\mathsf{Y} \times r}\}$.

In particular, Assumption 2.3 turns bounding (3) into bounding the excess risk of a non-realizable least squares problem (Ziemann et al., 2023b). We now discuss sufficient conditions to quantify the task-diversity parameter ν. We define the following "task-coverage" condition.

Definition 2.4. We say that the *task-coverage condition* holds if there exists a constant $\mu > 0$ such that: defining $h_\star^{(0)} \triangleq f_\star^{(0)} \circ g_\star$, for any $h = f \circ g \in \mathcal{F} \times \mathcal{G}$:

$$\begin{aligned} &\mathsf{E}^{(0)} \|h(X) - h_\star^{(0)}(X)\|_2^2 \\ &\leq \frac{\mu}{T} \sum_{t=1}^{T} \mathsf{E}^{(t)} \|h(X) - h_\star^{(0)}(X)\|_2^2. \end{aligned} \tag{TC}$$

Intuitively, Definition 2.4 quantifies the degree to which the mixture distribution of training task covariates covers the target covariate distribution.

Remark 2.5. When covariates are identically distributed for all tasks $\mathsf{P}_\mathsf{X}^{(t)} = \mathsf{P}_\mathsf{X}^{(t')}$, $t, t' \in \{0, \ldots, T\}$, then μ-(TC) holds with equality and $\mu = 1$. If the target distribution is absolutely continuous w.r.t. the mixture training distribution $\mathsf{P}_\mathsf{X}^{(0)} \ll \mathsf{P}_\mathsf{X}^{1:T}$, then μ is trivially bounded by the Radon-Nikodym derivative $\mu \leq \left\| \frac{d\mathsf{P}_\mathsf{X}^{(0)}}{d\mathsf{P}_\mathsf{X}^{1:T}} \right\|_\infty$. When $\mathcal{F}, \mathcal{G}$ are both linear classes, then μ-(TC) holds for any μ such that $\Sigma_\mathsf{X}^{(0)} \preceq \frac{\mu}{T} \sum_{t=1}^{T} \Sigma_\mathsf{X}^{(t)}$. Notably, this relaxes the "c" parameter in Du et al. (2020, Assumption 4.2) $\Sigma_\mathsf{X}^{(0)} \preceq c\Sigma_\mathsf{X}^{(t)}\ \forall t \in [T]$, as it only requires target task coverage on average over source tasks, rather than on each source task.

It turns out the notion of task-coverage implies a bound on the task-diversity parameter, captured in the following.

Proposition 2.6 ((TC) $\implies$ (TD))**.** *Let Assumption 2.3 hold. Define* $\mathbf{F}_\star^{(0)} \triangleq F_\star^{(0)\top} F_\star^{(0)}$ *and* $\mathbf{F}_\star^{1:T} \triangleq \frac{1}{T} \sum_{t=1}^{T} F_\star^{(t)\top} F_\star^{(t)} \in \mathbb{R}^{r \times r}$*, and suppose* $\mathrm{range}(\mathbf{F}_\star^{(0)}) \subseteq \mathrm{range}(\mathbf{F}_\star^{1:T})$*. Then any model satisfying* μ-(TC) *also satisfies* ν-(TD) *with* $\nu^{-1} = \mu \|\mathbf{F}_\star^{(0)} (\mathbf{F}_\star^{1:T})^\dagger\|_2$.

The proof is found in Appendix A. It may be helpful to consider scalar outputs $d_\mathsf{Y} = 1$, where the range requirement of Proposition 2.6 is equivalent to $F_\star^{(0)} \in \mathrm{span}(F_\star^{(1)}, \ldots, F_\star^{(T)})$. If this is not satisfied, then in the worst case $\mathsf{P}_\mathsf{X}^{(0)}$ may only excite the orthogonal component of $F_\star^{(0)}$, for which the training data is uninformative. We also note that $\|\mathbf{F}_\star^{(0)} (\mathbf{F}_\star^{1:T})^\dagger\|_2$ is precisely the generalization proposed in (Zhang et al., 2023a) of the "task-diversity parameter" (Du et al., 2020; Tripuraneni et al., 2021)[4]. However, we suggest that task diversity should actually be measured by the joint quantity $\mu \|\mathbf{F}_\star^{(0)} (\mathbf{F}_\star^{1:T})^\dagger\|_2$. Whereas $\|\mathbf{F}_\star^{(0)} (\mathbf{F}_\star^{1:T})^\dagger\|_2$ is precisely ν^{-1} when task covariates are identical (Remark 2.5), in general the alignment of the train and target task heads and of the covariate distributions both contribute to task diversity. Pathologically, if the train and target task covariates have disjoint supports, even if the heads are identical $F_\star^{(0)} = F_\star^{(t)},\ \forall t \in [T]$ ($\|\mathbf{F}_\star^{(0)} (\mathbf{F}_\star^{1:T})^\dagger\| = 1$), the error induced by a given (F, g) on the train distributions is in general uninformative to that on the target distribution. Similarly, non-trivial transfer risk is generally impossible when $\mathrm{range}(\mathbf{F}_\star^{(0)}) \not\subseteq \mathrm{range}(\mathbf{F}_\star^{1:T})$, even when $\mathsf{P}_\mathsf{X}^{(0)} = \mathsf{P}_\mathsf{X}^{(t)}, \forall t \in [T]$.

2.2. Warm-Up: Independent Covariates and Finite $\mathcal{G}$

In this section, we consider a basic setting where covariates are iid within-task (possibly non-identical between tasks) and where the representation class $\mathcal{G}$ is finite for simplicity. We now identify how the ideas introduced in the prequel lead to sample-efficient guarantees for representation learning. As previewed earlier, the target excess risk induced by the ERM $(\widehat{F}^{(0)}, \widehat{g})$ amounts to bounding two separate terms–the excess risk of a non-realizable least-squares regression, and the task-average estimation error of the ERM training predictors $(\{\widehat{F}^{(t)}\}_{t=1}^T, \widehat{g})$. We make the following boundedness assumptions to simplify ensuing expressions.

Assumption 2.7. Let $\mathcal{F} = \{F \in \mathbb{R}^{d_\mathsf{Y} \times r} : \|F\|_F \leq B_\mathcal{F}\}$,

[4]Therein, task-diversity is imposed by directly assuming normalization and well-conditioning $\lambda_i(\mathbf{F}_\star^{1:T}) = \Theta(T/r)$, $i = 1, \ldots, r$. Generality aside, this can also accrue an additional factor of r in the final rates when the eigenvalues of $\mathbf{F}_\star^{(0)}$ are non-uniform (see Du et al. (2020, Remark 4.2)).

and $\sup_{g\in\mathcal{G}}\sup_{x\in\mathcal{X}}\|g(x)\|_2 \le B_{\mathcal{G}}$.

Lastly, defining the centered function class $\overline{\mathcal{H}} \triangleq \mathcal{F}^{\otimes T}\times G - \{F_\star^{(t)}\}_{t=1}^T \times \{g_\star\}$, the following is an adaptation of a standard assumption (Oliveira, 2016; Koltchinskii & Mendelson, 2015; Ziemann & Tu, 2022).

Assumption 2.8 (Hypercontractivity)**.** We assume $(\overline{\mathcal{H}}, \mathsf{P}^{1:T})$ and $(\overline{\mathcal{H}}, \mathsf{P}^{(0)})$ satisfy (4-2) hypercontractivity: for each $\overline{h}\in\overline{\mathcal{H}}$,

$$\mathsf{E}^{1:T}\|\overline{h}(X)\|_2^4 \le C_{4\to 2}^{1:T}\left(\mathsf{E}^{1:T}\|\overline{h}(X)\|_2^2\right)^2, \tag{5}$$

$$\mathsf{E}^{(0)}\|\overline{h}(X)\|_2^4 \le C_{4\to 2}^{(0)}\left(\mathsf{E}^{(0)}\|\overline{h}(X)\|_2^2\right)^2. \tag{6}$$

Examples of hypercontractivity can be found in, e.g. (Ziemann & Tu, 2022).

Bounding Nonrealizable Least-Squares Error Let us define the random variable $Z \triangleq \hat{g}(X)\in\mathbb{R}^r$, and a best-in-class (misspecified) linear head on Z as

$$\widehat{F}_\star^{(0)} \triangleq \operatorname*{argmin}_{F\in\mathbb{R}^{d_{\mathcal{Y}}\times r}} \mathsf{E}^{(0)}\|Y - FZ\|_2^2$$

Since $\hat{g}$ is fixed with respect to $\mathsf{P}^{(0)}$, we may re-write (3) as

$$\begin{aligned}&\mathsf{E}^{(0)}\|\widehat{F}^{(0)}Z - Y\|_2^2 - \mathsf{E}^{(0)}\|\widehat{F}_\star^{(0)} - Y\|_2^2 \\ &= \|(\widehat{F}^{(0)} - \widehat{F}_\star^{(0)})\sqrt{\Sigma_Z^{(0)}}\|_F^2, \quad \Sigma_Z^{(0)} \triangleq \mathsf{E}^{(0)}[ZZ^\top].\end{aligned} \tag{7}$$

Define the (possibly biased) noise variable $U \triangleq Y - \widehat{F}_\star^{(0)}Z$. By the two-stage ERM, $\widehat{F}^{(0)}$ is precisely the least-squares solution on datapoints $\{(z_i^{(0)}, y_i^{(0)})\}_{i=1}^{N'}$. Therefore, we may adapt results from (Oliveira, 2016) and (Ziemann et al., 2023b) to bound the excess risk (7).

Proposition 2.9. *Fix $\delta\in(0,1)$. Define the noise-class interaction term $V \triangleq UZ^\top\Sigma_Z^{(0)-1/2}$, define $\sigma_U^2 \triangleq \sqrt{\mathsf{E}^{(0)}\left[\|U\|_2^4\right]}$, $\sigma_V^2 \triangleq \mathsf{E}^{(0)}\left[\|V\|_F^2\right]$ and $C_Z \triangleq \sup_{v\in\mathbb{S}^{d_{\mathcal{Y}}-1}}\sqrt{\mathsf{E}^{(0)}\langle v, \Sigma_Z^{(0)-1/2}Z\rangle^4}$. Then with probability at least $1-\delta$ we have*

$$\begin{aligned}\|(\widehat{F}^{(0)} - \widehat{F}_\star^{(0)})\sqrt{\Sigma_Z^{(0)}}\|_F^2 &\lesssim \frac{\sigma_V^2\log(1/\delta)}{N'} \\ &\lesssim \frac{C_Z\sigma_U^2 r\log(1/\delta)}{N'},\end{aligned}$$

as long as the burn-in $N' \gtrsim r + B_{\mathcal{G}}^2\log(1/\delta)$ is satisfied.

In Proposition 2.9, we express the excess risk of the nonrealizable least squares in terms of the variance proxy σ_U^2. We shall now relate σ_U^2 to σ_W^2, the "noise-level" of the underlying data-generating process. To reason about the magnitude of this quantity, we may re-arrange ν-(TD) to yield the following lemma.

Lemma 2.10. *Let σ_U^2 be as in Proposition 2.9. Then:*

$$\begin{aligned}\sigma_U^2 &\lesssim d_{\mathcal{Y}}\sigma_W^2 \\ &\quad + \frac{\sqrt{C_{4\to 2}^{(0)}}\nu^{-1}}{T}\sum_{t=1}^T \mathsf{E}^{(t)}\|\widehat{F}^{(t)}\hat{g}(X) - F_\star^{(t)}g_\star(X)\|_2^2.\end{aligned}$$

In other words, the noise level of the misspecified model is no more than the optimal noise level plus the familiar task-averaged estimation error (4), which we note is additionally divided by N' in Proposition 2.9. Therefore, we have isolated the task-averaged estimation error (4) as the sole remaining quantity to control.

Bounding Task-Averaged Estimation Error The goal now is to control the task-averaged estimation error. As previously discussed, the key observation is to quantify a lower isometry, such that

$$\mathsf{E}^{1:T}\left\|\overline{h}\right\|_2^2 \lesssim \widehat{\mathsf{E}}^{1:T}\left\|\overline{h}\right\|_2^2, \text{ for all } \overline{h}\in\overline{\mathcal{H}}.$$

Toward this end, we show that hypercontractivity (Assumption 2.8) leads to a lower estimate for any *given* $\overline{h}\in\overline{\mathcal{H}}$ (Proposition A.1). By an application of the *offset basic inequality* (Rakhlin & Sridharan, 2014; Liang et al., 2015), an empirical estimation error can be bounded by

$$\begin{aligned}&\frac{1}{NT}\sum_{t=1}^T\sum_{i=1}^N\|h(x_i^{(t)})\|_2^2 \\ &\le \sup_{h\in\mathcal{H}}\frac{1}{NT}\sum_{t=1}^T\sum_{i=1}^N 4\left\langle w_i^{(t)}, h(x_i^{(t)})\right\rangle - \|h(x_i^{(t)})\|_2^2 \\ &\triangleq \mathsf{M}_{NT}(\mathcal{H}),\end{aligned}$$

where $\mathsf{M}_{NT}(\mathcal{H})$ is denoted the (empirical) *martingale offset complexity* (Liang et al., 2015; Ziemann & Tu, 2022), which serves as the capacity measure of a hypothesis class $\mathcal{H}$. Notably, $\mathsf{M}_{NT}(\mathcal{H})$ scales with the *noise-level* σ_W^2, rather than the diameter of $\mathcal{H}$. Via a high-probability chaining bound (Lemma A.2), we demonstrate $\mathsf{M}_{NT}(\mathcal{H})$ is controlled by a log-covering number of $\mathcal{H}$ at a resolution γ of our choice. As a result, there is a regime of γ such that with probability at least $1-\delta$,

$$\mathsf{M}_{NT}(\mathcal{H}) \lesssim \frac{\sigma_W^2}{NT}\left(\log\mathsf{N}_\infty(\mathcal{H},\gamma) + \log(1/\delta)\right),$$

where $\mathsf{N}_\infty(\mathcal{H},\gamma)$ is the covering number of $\mathcal{H}$ in the supremum metric: $\rho(h_1,h_2) = \sup_{x\in\mathsf{X}}\|h_1(x) - h_2(x)\|_2$. For salient choices of γ, we want $\mathsf{M}_{NT}(\mathcal{H})$ to be the dominant scaling in the estimation error bound. We then proceed to a localization argument, where we can define disjoint events over elements of $\overline{\mathcal{H}}$:[5] either: 1. the population estimation

[5]Technically, over a class that subsumes $\overline{\mathcal{H}}$.

error is within an τ^2 radius around zero, or 2. the estimation error exceeds τ^2 but is dominated by the empirical error, which is bounded by the martingale offset complexity. In particular, the probability of neither event holding can be controlled by union bounding over a finite $\mathcal{O}(\tau)$-cover of $\overline{\mathcal{H}}$, such that we have with probability at least $1 - p(\tau, N, T)$, for all $\overline{h} \in \overline{\mathcal{H}}$:

$$\mathsf{E}^{1:T}\|\overline{h}\|_2^2 \lesssim \max\{\mathsf{M}_{NT}(\overline{\mathcal{H}}), \tau^2\}. \tag{8}$$

Thus, this informs choosing γ, τ such that the two terms meet at the desired rate. The failure probability $p(\tau, N, T)$ turns into a burn-in condition on N, T when inverted for δ. As the last step before bounding the estimation error, we note that $\mathcal{F}^{\otimes T}$ can be identified with a bounded set in $\mathbb{R}^{Td_{\mathsf{Y}}r}$, and therefore we get the following bound on the covering number

$$\mathsf{N}_\infty(\overline{\mathcal{H}}, \varepsilon) \le Td_{\mathsf{Y}}r \log\left(1 + \frac{2TB_{\mathcal{F}}B_{\mathcal{G}}}{\varepsilon}\right) + \log|\mathcal{G}|.$$

The aforementioned steps and proofs are found in Lemma A.2, Proposition A.3, and Lemma A.4. Optimizing resolutions γ and τ yields a bound the task-averaged estimation error.

Proposition 2.11. *Let Assumption 2.7 hold. Then, with probability at least $1 - \delta$, the estimation error of ERM predictors $\{\widehat{F}^{(t)}\}_{t=1}^T, \hat{g}$ is bounded by*

$$\frac{1}{T}\sum_{t=1}^{T}\mathsf{E}^{(t)}\|\widehat{F}^{(t)}\hat{g}(X) - F_\star^{(t)}g_\star(X)\|_2^2 \le \sigma_W^2 \cdot \tilde{\mathcal{O}}\left(\frac{d_{\mathsf{Y}}r}{N} + \frac{\log|\mathcal{G}| + \log(1/\delta)}{NT}\right),$$

as long as $N \ge C_{4\to 2}^{1:T} \cdot \tilde{\Omega}\left(d_{\mathsf{Y}}r + \log|\mathcal{G}|/T + \log(1/\delta)\right)$.

We omit logarithmic dependencies on problem-dependent constants and sample sizes for clarity. Combining Proposition 2.6, Proposition 2.9, and Proposition 2.11 yields the final bound on the excess transfer risk.

Theorem 2.12 (Transfer risk bound)**.** *Assume $\mathsf{P}^{0:T}$ satisfy μ-(TC), and let Assumption 2.7 hold. With probability at least $1 - \delta$, the target excess risk of the two-stage ERM predictor $(\widehat{F}^{(0)}, \hat{g})$ is bounded by*

$$\mathrm{ER}(\widehat{F}^{(0)}, \hat{g}) \le \frac{\sigma_W^2 C_Z d_{\mathsf{Y}} r \log(1/\delta)}{N'} + \sigma_W^2 \mu \|\mathbf{F}_\star^{(0)}(\mathbf{F}_\star^{1:T})^\dagger\|_2 \cdot \tilde{\mathcal{O}}\left(\frac{d_{\mathsf{Y}}r}{N} + \frac{\log|\mathcal{G}| + \log(1/\delta)}{NT}\right),$$

as long as $N' \gtrsim C_Z\sqrt{C_{4\to 2}^{(0)}}r + B_{\mathcal{G}}^2\log(1/\delta)$ and $N \gtrsim C_{4\to 2}^{1:T} \cdot \tilde{\Omega}\left(d_{\mathsf{Y}}r + \log|\mathcal{G}|/T + \log(1/\delta)\right)$.

The proofs of Proposition 2.11 and Theorem 2.12 can be found in Appendix A.3. We observe the following: 1. the rates are qualitatively correct, where the noise-level hits $\dim(\mathcal{F})/\{N, N'\}$ for the complexity of fitting the linear heads and $\log|\mathcal{G}|$ for the shared representation, 2. the burn-in for N' is proportional to r, which is the number of samples necessary for $\widehat{F}^{(0)}$ to be well-posed, 3. in the burn-in for N, $\log|G|$ is additionally divided by T. Therefore, for large T, the dominant term is $d_{\mathsf{Y}}r$. Compared to prior nonlinear representation learning work (Du et al., 2019; Tripuraneni et al., 2020) (where $d_{\mathsf{Y}} = 1$), this is a dramatic improvement from at least $\mathcal{O}(d_{\mathsf{X}})$ to r.

2.3. Representation Learning with Little Mixing

In this section, we extend our results to full generality, allowing possibly dependent within-task data within-task and general representation classes $\mathcal{G}$, subsuming various settings of interest, such as identification of nonlinear dynamical systems. Beyond finiteness, we instead characterize the complexity of a function class by its log-covering number $\log \mathsf{N}_\infty(\mathcal{G}, \gamma)$ in the supremum metric $\rho(g_1, g_2) = \sup_{x\in\mathsf{X}}\|g_1(x) - g_2(x)\|_2$. Besides finite classes, this subsumes various standard classes of interest, such as (Lipschitz) parametric function classes.

Example 1 (Lipschitz parametric function class)**.** *A function class $\mathcal{G}$ is called $(B_\theta, L_\theta, d_\theta)$-*Lipschitz parametric *if $\mathcal{G} = \{g_\theta(\cdot) \mid \theta \in \Theta\}$ with $\Theta \subset \mathbb{R}^{d_\theta}$, and satisfies*

$$\sup_{\theta\in\Theta}\|\theta\| \le B_\theta, \tag{9}$$

$$\sup_{x\in\mathsf{X}}\sup_{\substack{\theta_1,\theta_2\in\Theta\\ \theta_1\neq\theta_2}}\frac{\|g_{\theta_1}(x) - g_{\theta_2}(x)\|_2}{\|\theta_1 - \theta_2\|} \le L_\theta. \tag{10}$$

By a standard volumetric argument (Wainwright, 2019), it can be shown that a $(B_\theta, L_\theta, d_\theta)$-Lipschitz parametric class satisfies

$$\log \mathsf{N}_\infty(\mathcal{G}, \gamma) \le d_\theta \log\left(1 + \frac{2B_\theta L_\theta}{\gamma}\right).$$

Parametric function classes include various models of interest, such as (generalized) linear models and neural networks with smooth activations. Notably, instantiating $\mathcal{G}$ as a linear class, by identifying it with $r \times d_{\mathsf{X}}$ (orthonormal) matrices (Du et al., 2020), we may replace $\log|\mathcal{G}| \mapsto \tilde{\mathcal{O}}(rd_{\mathsf{X}})$, immediately recovering the rates from prior work on multi-task linear regression, along with the reduced burn-in and refined task diversity estimate. We note that our results are not limited to "parametric-type" covering number estimates, and can handle various non-parametric classes. We refer to (Ziemann & Tu, 2022) for various worked examples. In particular, by associating the complexity of $\mathcal{G}$ to a well-studied measure in the log-covering number, rate-optimal multi-task bounds can be painlessly extended from many existing single-task settings, avoiding the need for

custom complexity measures that may be hard to instantiate or suboptimal.

To quantify dependency, we consider ϕ-mixing (Kuznetsov & Mohri, 2017) covariates. For a sequence of random variables $\{S_i\}_{i=1}^n$ the ϕ-mixing coefficients are given by $\phi_S(i) = \sup_{t\in[n]:t+i\le n} \sup_s \|\mathsf{P}_{S_{i+t}}(s \mid S_{1:t}) - \mathsf{P}_{S_{i+t}}\|_{\mathsf{TV}}, i \in [n]$, where $\|\cdot\|_{\mathsf{TV}}$ denotes the total variation distance. With this preliminary notion in place we now state the analogue of Proposition 2.9.

Proposition 2.13. *Suppose that $\mathsf{P}^{(0)}$ is stationary and ϕ-mixing and fix $\delta \in (0,1)$. Fix a block length k dividing $N'/2$. Define the blocked noise-class interaction term $V \triangleq \frac{1}{k}\sum_{i=1}^k U_i Z_i^\top \Sigma_Z^{(0)-1/2}$ and $\sigma_V^2 \triangleq \mathsf{E}^{(0)}\left[\|V\|_F^2\right]$. Define σ_U^2 and C_Z as in Proposition 2.9. Then, defining $C^{(0)} \triangleq C_Z\sqrt{C_{4\to 2}^{(0)}}$, with probability at least $1-\delta$ we have*

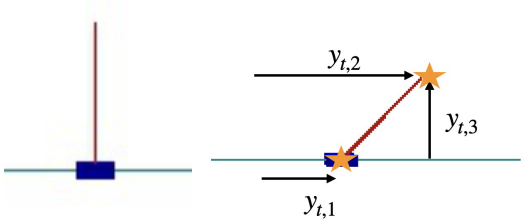
$$\|(\widehat{F}^{(0)} - \widehat{F}_\star^{(0)})\sqrt{\Sigma_Z^{(0)}}\|_F^2 \lesssim \frac{\sigma_V^2\log(1/\delta)}{N'} \lesssim \frac{C_Z\sigma_U^2 r\log(1/\delta)}{N'} + \frac{C^{(0)}k\nu^{-1}\log(1/\delta)}{N'T}\sum_{t=1}^T \mathsf{E}^{(t)}\|\widehat{F}^{(t)}\hat{g}(X) - F_\star^{(t)}g_\star(X)\|_2^2,$$

as long as the burn-in $N' \gtrsim k\Big(r + B_{\mathcal{G}}^2\log(1/\delta)\Big)$ is satisfied and the block length is sufficiently long: $k \ge N'\phi(k)\delta^{-1}$.

The proof is analogous to Proposition 2.9 but with an extra term once again including the task-average estimation error due to mixing. This term arises due to a slightly different computation of σ_V wherein we need to account for cross-terms due to dependency. Crucially, we point out that this term is of higher order in the sample size than typical occurrences of the task-average estimation error and can be rendered negligible after a burn in $N' \gtrsim k$. In other words, we recover Proposition 2.9 and are able to shift the effect of mixing to the burn-in. Additionally, as we noted earlier, a key benefit of the martingale offset complexity is that it does not depend on the data distribution beyond the conditional noise-level. Therefore, with minimal modifications there, we state the main theorem for the mixing case.

Theorem 2.14 (Transfer risk bound, mixing)**.** *Suppose that $P^{0:T}$ are each stationary and ϕ-mixing and fix $\delta \in (0,1)$. Assume that k is fixed and divides $N'/2$ and $N/2$. Define the quantity $\Phi \triangleq (\sum_{i=1}^\infty \sqrt{\phi(i)})^2$. Assume $\mathsf{P}^{0:T}$ satisfy μ-(TC), and let Assumption 2.7 hold. Define $\mathrm{C}(\mathcal{G}) \triangleq \log \mathsf{N}_\infty\left(\mathcal{G}, \frac{B_{\mathcal{F}}NT\sqrt{d_{\mathcal{Y}}}}{\sigma_W}\right)$. With probability at least $1-\delta$, the target excess risk is bounded by*

$$\mathrm{ER}(\widehat{F}^{(0)}, \hat{g}) \le \frac{\sigma_W^2 C_Z d_{\mathcal{Y}} r\log(1/\delta)}{N'} + \sigma_W^2\mu\|\mathbf{F}_\star^{(0)}(\mathbf{F}_\star^{1:T})^\dagger\|_2 \cdot \tilde{\mathcal{O}}\left(\frac{d_{\mathcal{Y}}r}{N} + \frac{\mathrm{C}(\mathcal{G}) + \log(1/\delta)}{NT}\right),$$



(a) Example observation (b) ideal keypoint extraction

Figure 1. Figure 1(a) shows an example camera observation of the pybullet simulated cartpole environment. In this image, the cartpole is at the state $x = \begin{bmatrix}0 & 0 & 0 & 0\end{bmatrix}^\top$. Figure 1(b) illustrates the ideal keypoints extracted from a cartpole image.

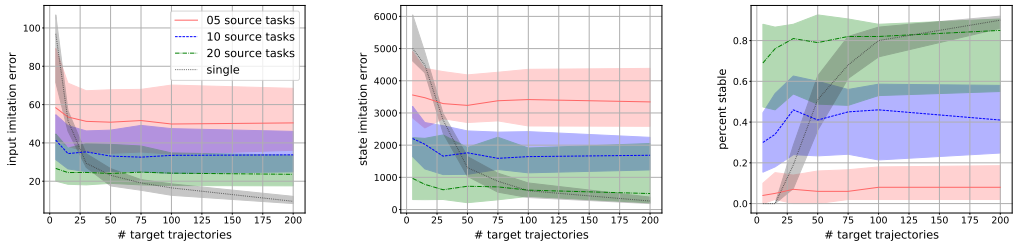
as long as $N' \gtrsim k\left(C_Z\sqrt{C_{4\to 2}^{(0)}}r + B_{\mathcal{G}}^2\log(1/\delta)\right)$ and $N \gtrsim C_{4\to 2}^{1:T}\Phi\cdot\tilde{\Omega}\left(d_{\mathcal{Y}}r + \mathrm{C}(\mathcal{G})/T + \log(1/\delta)\right)$.

To understand how mixing affects our bounds, we consider geometric ϕ-mixing, i.e. $\phi(k) \le \Gamma\rho^k$ for some $\Gamma > 0$, $\rho \in (0,1)$. Then we can find a valid block length $k \approx \log(N')$ and $\Phi = \frac{\Gamma}{(1-\sqrt{\rho})^2}$, thereby inflating the burn-in requirement on N' by a factor of $\log(N')$ and N by a constant factor. Notably, the rate remains the same as the iid setting. With ϕ-mixing, we are able to port to broader sequential settings, such as Markov Chains (Samson, 2000) and parametrized dynamical systems (Tu et al., 2022; Ziemann & Tu, 2022).

3. Numerical Validation

To validate our theoretical observations, we consider a non-trivial regression task over dynamical systems: balancing a pole atop a cart from visual observations, as pictured in Figure 1(a). A collection of systems is obtained by randomly sampling different values for the cart mass, pole mass, and pole length parameters. The regression task is to imitate expert policies controlling each collection of systems from (control input, observation) pairs. We design expert policies as linear controllers of the underlying state[6] to balance the pole in the upright position. The expert estimates the state of the system from the camera observations by first applying a keypoint extractor to the camera observations to get noisy estimates of two keypoints (visualized in Figure 1(b)), and then passing these noisy estimates into a Kalman filter. A common keypoint extractor is shared across the experts, but the linear controllers and filters are system-specific. Actuation noise is added to the expert input when it is applied to the system. We use demonstrations from the aforementioned expert policies to train imitation learning policies to replicate the experts. The policies are parameterized with convolutional neural networks. They take as input a history of 8 images, and output the control action to be applied to

[6] This consists of the cart position and velocity, and pole angular position and angular velocity.



(a) Input imitation error (b) State imitation gap (c) Percent stable

Figure 2. Three evaluation metrics comparing the performance of multi-task versus single-task imitation learning: the MSE between the input of the learned and expert controllers when evaluated on the expert trajectory, the deviation between the state trajectories generated by the learned and expert controllers, and the %trials that the learned controller keeps the pole balanced for all 500 timesteps (the dynamics are discretized to $\Delta t = 0.02$ seconds). Three curves are shown for multi-task imitation learning, generated by pre-training with a different number of source tasks. In all metrics, multi-task learning improves over single task when few target trajectories are available.

the system. The policies are trained by solving a supervised learning problem using the expert demonstrations.

Our theoretical analysis predicts that multi-task learning helps substantially in this setting, due to the shared keypoint extractor across all policies. The part that varies between expert policies is the controller and filter, which are linear maps from the keypoints to the control action to be consistent with our linear $\mathcal{F}$ nonlinear $\mathcal{G}$ model.

The experimental results in Figure 2 compare multi-task learning with single-task learning. We consider multi-task learning using a varying number of source tasks, each consisting of 10 expert demonstrations. The x-axis denotes the number of demonstrations available from the target task. For single task learning, these trajectories are used to train the entire network, while for multi-task learning they are used to fit only the final layer, keeping the representation fixed from pre-training on the source tasks. Three evaluation metrics are plotted: the MSE of the learned controller inputs, the MSE between the learned and expert trajectories, and the %trials where the controller is stabilizing. Each metric is averaged over 50 evaluation rollouts for each controller. We plot the median and shade 30%-70% quantiles for these evaluation metrics over 5 random seeds for pretraining the representation, and 10 realizations of target tasks. In all metrics, multi-task learning improves over single task learning in the low data regime as predicted, but saturates quickly when the number of target trajectories exceeds the number of per-task training trajectories, which our theory predicts is the limiting rate when T is large. Full experimental details are contained in Appendix C.

4. Discussion

We provided new guarantees for nonlinear representation learning that: 1. agree with prior work rate-wise, 2. apply to non-identical covariates and/or sequentially dependent (ϕ-mixing) covariates, 3. improve the per-task sample requirement and refine the task-diversity measure. We did not address pathologies that can arise in multi-task learning, such as class (source data) imbalance and low task diversity. Indeed, addressing these pathologies is what motivates ongoing work in active learning (Wang et al., 2023) and alignment (Wu et al., 2020), which are important directions to fully realize the benefit of learning over multiple tasks.

We conclude with some remaining technical open questions, contained in Appendix D.

Acknowledgements

Ingvar Ziemann is supported by a Swedish Research Council international postdoc grant. George J. Pappas is supported in part by NSF Award SLES-2331880. Nikolai Matni is supported in part by NSF Award SLES-2331880 and NSF CAREER award ECCS-2045834.

Impact Statement

This paper presents work whose primary purpose is to further understanding of the theoretical properties of a well-known problem. Accordingly, the results derived in this paper are descriptive by nature, and do not present any immediate societal consequences that the existing literature and implementations therein have not already presented.

References

Agarwal, A., Kakade, S., Krishnamurthy, A., and Sun, W. Flambe: Structural complexity and representation learning of low rank mdps. *Advances in neural information processing systems*, 33:20095–20107, 2020.

Arora, S., Du, S., Kakade, S., Luo, Y., and Saunshi, N. Provable representation learning for imitation learning via bi-level optimization. In *International Conference on Machine Learning*, pp. 367–376. PMLR, 2020.

Baxter, J. A model of inductive bias learning. *Journal of artificial intelligence research*, 12:149–198, 2000.

Ben-David, S. and Borbely, R. S. A notion of task relatedness yielding provable multiple-task learning guarantees. *Machine learning*, 73:273–287, 2008.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828, 2013.

Boyd, S. and Vandenberghe, L. *Convex optimization*. Cambridge university press, 2004.

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

Chen, Y., Jamieson, K., and Du, S. Active multi-task representation learning. In *International Conference on Machine Learning*, pp. 3271–3298. PMLR, 2022.

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

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

Collins, L., Hassani, H., Soltanolkotabi, M., Mokhtari, A., and Shakkottai, S. Provable multi-task representation learning by two-layer relu neural networks. *arXiv preprint arXiv:2307.06887*, 2023.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T. (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/V1/N19-1423. URL https://doi.org/10.18653/v1/n19-1423.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Du, S., Krishnamurthy, A., Jiang, N., Agarwal, A., Dudik, M., and Langford, J. Provably efficient rl with rich observations via latent state decoding. In *International Conference on Machine Learning*, pp. 1665–1674. PMLR, 2019.

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

Du, Y., Huang, L., and Sun, W. Multi-task representation learning for pure exploration in linear bandits. In *International Conference on Machine Learning*, pp. 8511–8564. PMLR, 2023.

Efroni, Y., Foster, D. J., Misra, D., Krishnamurthy, A., and Langford, J. Sample-efficient reinforcement learning in the presence of exogenous information. In *Conference on Learning Theory*, pp. 5062–5127. PMLR, 2022.

Foster, D. and Rakhlin, A. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pp. 3199–3210. PMLR, 2020.

Hanneke, S. and Kpotufe, S. A no-free-lunch theorem for multitask learning. *The Annals of Statistics*, 50(6):3119–3143, 2022.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143. PMLR, 2020.

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

Kumar, A., Raghunathan, A., Jones, R. M., Ma, T., and Liang, P. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022.

Kuznetsov, V. and Mohri, M. Generalization bounds for non-stationary mixing processes. *Machine Learning*, 106 (1):93–117, 2017.

Lee, Y., Chen, A. S., Tajwar, F., Kumar, A., Yao, H., Liang, P., and Finn, C. Surgical fine-tuning improves adaptation to distribution shifts. In *The Eleventh International Conference on Learning Representations*, 2023.

Liang, T., Rakhlin, A., and Sridharan, K. Learning with square loss: Localization through offset rademacher complexity. In *Conference on Learning Theory*, pp. 1260–1285. PMLR, 2015.

Lu, R., Huang, G., and Du, S. S. On the power of multitask representation learning in linear mdp. *arXiv preprint arXiv:2106.08053*, 2021.

Mania, H., Jordan, M. I., and Recht, B. Active learning for nonlinear system identification with guarantees. *J. Mach. Learn. Res.*, 23:32–1, 2022.

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

Mendelson, S. Improving the sample complexity using global data. *IEEE Transactions on Information Theory*, 48(7):1977–1991, 2002.

Meunier, D., Li, Z., Gretton, A., and Kpotufe, S. Nonlinear meta-learning can guarantee faster rates. *arXiv preprint arXiv:2307.10870*, 2023.

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

Mukherjee, S., Xie, Q., Hanna, J., and Nowak, R. Multi-task representation learning for pure exploration in bilinear bandits. *Advances in Neural Information Processing Systems*, 36, 2023.

Oliveira, R. I. The lower tail of random quadratic forms with applications to ordinary least squares. *Probability Theory and Related Fields*, 166(3):1175–1194, 2016.

O'Connell, M., Shi, G., Shi, X., Azizzadenesheli, K., Anandkumar, A., Yue, Y., and Chung, S.-J. Neural-fly enables rapid learning for agile flight in strong winds. *Science Robotics*, 7(66):eabm6597, 2022.

Rakhlin, A. and Sridharan, K. Online non-parametric regression. In *Conference on Learning Theory*, pp. 1232–1264. PMLR, 2014.

Samson, P.-M. Concentration of measure inequalities for markov chains and ϕ-mixing processes. *The Annals of Probability*, 28(1):416–461, 2000.

Shridhar, M., Manuelli, L., and Fox, D. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pp. 785–799. PMLR, 2023.

Simchi-Levi, D. and Xu, Y. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Mathematics of Operations Research*, 47(3):1904–1931, 2022.

Thekumparampil, K. K., Jain, P., Netrapalli, P., and Oh, S. Sample efficient linear meta-learning by alternating minimization. *arXiv preprint arXiv:2105.08306*, 2021.

Tripuraneni, N., Jordan, M., and Jin, C. On the theory of transfer learning: The importance of task diversity. *Advances in neural information processing systems*, 33: 7852–7862, 2020.

Tripuraneni, N., Jin, C., and Jordan, M. Provable meta-learning of linear representations. In *International Conference on Machine Learning*, pp. 10434–10443. PMLR, 2021.

Tu, S., Frostig, R., and Soltanolkotabi, M. Learning from many trajectories. *arXiv preprint arXiv:2203.17193*, 2022.

Uehara, M., Zhang, X., and Sun, W. Representation learning for online and offline rl in low-rank mdps. In *International Conference on Learning Representations*, 2021.

Vaskevicius, T., Kanade, V., and Rebeschini, P. The statistical complexity of early-stopped mirror descent. *Advances in Neural Information Processing Systems*, 33:253–264, 2020.

Wagenmaker, A., Shi, G., and Jamieson, K. Optimal exploration for model-based rl in nonlinear systems. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Wang, Y., Chen, Y., Jamieson, K., and Du, S. S. Improved active multi-task representation learning via lasso. In *International Conference on Machine Learning*, pp. 35548–35578. PMLR, 2023.

Watkins, A., Ullah, E., Nguyen-Tang, T., and Arora, R. Optimistic rates for multi-task representation learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Wu, S., Zhang, H. R., and Ré, C. Understanding and improving information transfer in multi-task learning. In *International Conference on Learning Representations*, 2020.

Yang, J., Hu, W., Lee, J. D., and Du, S. S. Impact of representation learning in linear bandits. In *International Conference on Learning Representations*, 2020.

Yang, J., Lei, Q., Lee, J. D., and Du, S. S. Nearly minimax algorithms for linear bandits with shared representation. *arXiv preprint arXiv:2203.15664*, 2022.

Yu, B. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, 22(1):94–116, 1994.

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

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

Ziemann, I. *Statistical Learning, Dynamics and Control: Fast Rates and Fundamental Limits for Square Loss*. PhD thesis, KTH Royal Institute of Technology, 2022.

Ziemann, I. and Tu, S. Learning with little mixing. *arXiv preprint arXiv:2206.08269. NeurIPS'22*, 2022.

Ziemann, I., Tsiamis, A., Lee, B., Jedra, Y., Matni, N., and Pappas, G. J. A tutorial on the non-asymptotic theory of system identification, 2023a.

Ziemann, I., Tu, S., Pappas, G. J., and Matni, N. The noise level in linear regression with dependent data. *arXiv preprint arXiv:2305.11165*, 2023b.

A. Proofs and Additional Information for Section 2

A.1. Section 2.1

Proposition 2.6 ((TC) $\Longrightarrow$ (TD))**.** *Let Assumption 2.3 hold. Define* $\mathbf{F}_\star^{(0)} \triangleq F_\star^{(0)\top} F_\star^{(0)}$ *and* $\mathbf{F}_\star^{1:T} \triangleq \frac{1}{T}\sum_{t=1}^T F_\star^{(t)\top} F_\star^{(t)} \in \mathbb{R}^{r\times r}$*, and suppose* $\operatorname{range}(\mathbf{F}_\star^{(0)}) \subseteq \operatorname{range}(\mathbf{F}_\star^{1:T})$*. Then any model satisfying* μ*-*(TC) *also satisfies* ν*-*(TD) *with* $\nu^{-1} = \mu\|\mathbf{F}_\star^{(0)}(\mathbf{F}_\star^{1:T})^\dagger\|_2$*.*

Proof. Let g be fixed. Then, writing out the left-hand side of (TD), we have:

$$\begin{aligned}
&\inf_{F\in\mathcal{F}} \mathsf{E}^{(0)}\|(F,g)(X)-Y\|_2^2 - \mathsf{E}^{(0)}\|F_\star^{(0)}g_\star(X)-Y\|_2^2 \\
&\quad\le \inf_{F\in\mathcal{F}} \mathsf{E}^{(0)}\|(f,g)(X)-F_\star^{(0)}g_\star(X)\|_2^2 \qquad (F_\star^{(0)} \text{ is } \mathsf{E}^{(0)}-\text{optimal}) \\
&\quad\le \inf_{F\in\mathcal{F}} \frac{\mu}{T}\sum_{t=1}^T \mathsf{P}_t\|(f,g)(X)-(f_{0,\star},g_\star)(X)\|_2^2. \quad \text{invoking (TC)}
\end{aligned} \tag{11}$$

Recall $\mathsf{P}^{1:T}$ is the uniform mixture over $\mathsf{P}^{(t)}, t\in[T]$. Define also $\Sigma \triangleq \mathsf{E}^{1:T}\begin{bmatrix} g(X)g(X)^\top & g(X)g_\star(X)^\top \\ g_\star(X)g(X)^\top & g_\star(X)g_\star(X)^\top\end{bmatrix}$ and its Schur complement

$$\overline{\Sigma} \triangleq \mathsf{E}^{1:T}[g(X)g(X)^\top] - \mathsf{E}^{1:T}[g_\star(X)g(X)^\top]\mathsf{E}^{1:T}[g_\star(X)g_\star^\top(X)]^\dagger\mathsf{E}^{1:T}[g(X)g_\star(X)^\top].$$

We now rewrite the last line of (11) as a single integral:

$$\begin{aligned}
&\inf_{F\in\mathcal{F}} \frac{\mu}{T}\sum_{t=1}^T \mathsf{E}^{(t)}\|Fg(X)-F_\star^{(0)}g(X)\|_2^2 \\
&\quad= \inf_{F\in\mathcal{F}} \mu\,\mathsf{E}^{1:T}\|Fg(X)-F_\star^{(0)}g_\star(X)\|_2^2 \\
&\quad= \inf_{F\in\mathcal{F}} \mu\operatorname{Tr}\left(\begin{bmatrix}F^\top \\ -F_\star^{(0)\top}\end{bmatrix}^\top \Sigma \begin{bmatrix}F^\top \\ -F_\star^{(0)\top}\end{bmatrix}\right) \quad \text{(linearity)} \\
&\quad= \mu\operatorname{Tr}\left(F_\star^{(0)}\overline{\Sigma}F_\star^{(0)\top}\right) \qquad \text{(optimize over } f) \\
&\quad\triangleq \mu\operatorname{Tr}\left(\overline{\Sigma}\mathbf{F}_\star^{(0)}\right).
\end{aligned} \tag{12}$$

The optimization step is a standard calculation about partial minima of quadratic forms (see e.g. Boyd & Vandenberghe, 2004, Example 3.15, Appendix A.5.4).

A similar calculation on the RHS of (TD) yields

$$\frac{1}{T}\sum_{t=1}^T \inf_{F\in\mathcal{F}} \mathsf{E}^{(t)}\|Fg(X)-F_\star^{(t)}g_\star(X)\|_2^2 = \frac{1}{T}\sum_{t=1}^T \operatorname{Tr}\left(\overline{\Sigma}F_\star^{(t)\top}F_\star^{(t)}\right) \tag{13}$$

$$\triangleq \operatorname{Tr}\left(\overline{\Sigma}\mathbf{F}_\star^{1:T}\right). \tag{14}$$

Clearly $\mu\operatorname{Tr}(\overline{\Sigma}\mathbf{F}_\star^{(0)}) \le \mu\|\mathbf{F}_\star^{(0)}\mathbf{F}_\star^{1:T\dagger}\|_2 \operatorname{Tr}\left(\overline{\Sigma}\mathbf{F}_\star^{1:T}\right)$ and so by combining (11), (12) and (13), we get

$$\begin{aligned}
\inf_{F\in\mathcal{F}} \mathrm{ER}(F,g) &\triangleq \mathsf{E}^{(0)}\|(F,g)(X)-Y\|_2^2 - \mathsf{E}^{(0)}\|F_\star^{(0)}g_\star(X)-Y\|_2^2 \\
&\le \inf_{F\in\mathcal{F}} \frac{\mu}{T}\sum_{t=1}^T \mathsf{P}_t\|(f,g)(X)-(f_{0,\star},g_\star)(X)\|_2^2 \\
&\le \frac{\mu\|\mathbf{F}_\star^{(0)}\mathbf{F}_\star^{1:T\dagger}\|_2}{T}\sum_{t=1}^T \inf_{F\in\mathcal{F}} \mathsf{E}^{(t)}\|Fg(X)-F_\star^{(t)}g_\star(X)\|_2^2.
\end{aligned}$$

By setting $\nu^{-1} = \mu\|\mathbf{F}_\star^{(0)}\mathbf{F}_\star^{1:T\dagger}\|_2$, this verifies ν-(TD).

□

A.2. Non-realizable Least Squares

Proposition 2.9. *Fix $\delta \in (0,1)$. Define the noise-class interaction term $V \triangleq UZ^\top \Sigma_Z^{(0)-1/2}$, define $\sigma_U^2 \triangleq \sqrt{\mathsf{E}^{(0)}\left[\|U\|_2^4\right]}$, $\sigma_V^2 \triangleq \mathsf{E}^{(0)}\left[\|V\|_F^2\right]$ and $C_Z \triangleq \sup_{v\in\mathbb{S}^{d_Y-1}} \sqrt{\mathsf{E}^{(0)}\langle v, \Sigma_Z^{(0)-1/2} Z\rangle^4}$. Then with probability at least $1-\delta$ we have*

$$\begin{aligned}\|(\widehat{F}^{(0)} - \widehat{F}_\star^{(0)})\sqrt{\Sigma_Z^{(0)}}\|_F^2 &\lesssim \frac{\sigma_V^2 \log(1/\delta)}{N'} \\ &\lesssim \frac{C_Z \sigma_U^2 r \log(1/\delta)}{N'},\end{aligned}$$

as long as the burn-in $N' \gtrsim r + B_{\mathcal{G}}^2 \log(1/\delta)$ is satisfied.

Proof. The result in terms of σ_V is immediate by Ziemann et al. (2023b, Theorem 3.1) and so it remains to compute the noise term σ_V for the second inequality. Namely, we have that:

$$\begin{aligned}\sigma_V^2 &= \mathsf{E}\|UZ^\top \Sigma_Z^{-1/2}\|_F^2 \\ &= \mathsf{E}\|U\|_2^2 \|\Sigma_Z^{-1/2} Z\|_2^2 \qquad \text{(rewrite rank-1 objects)} \\ &\le \sqrt{\mathsf{E}\|U\|_2^4 \mathsf{E}\|\Sigma_Z^{-1/2} Z\|_2^4}. \quad \text{(Cauchy-Schwarz)}\end{aligned} \tag{15}$$

The result follows since $\mathsf{E}\|\Sigma_Z^{-1/2} Z\|_2^4 \le C_Z^2 r^2$. □

Lemma 2.10. *Let σ_U^2 be as in Proposition 2.9. Then:*

$$\begin{aligned}\sigma_U^2 &\lesssim d_{\mathsf{Y}} \sigma_W^2 \\ &\quad + \frac{\sqrt{C_{4\to 2}^{(0)}}\nu^{-1}}{T} \sum_{t=1}^T \mathsf{E}^{(t)} \|\widehat{F}^{(t)} \hat{g}(X) - F_\star^{(t)} g_\star(X)\|_2^2.\end{aligned}$$

Proof. Recall that $U = Y - \widehat{F}_\star^{(0)} \hat{g}(X) = W + F_\star^{(0)} g_\star(X) - \widehat{F}_\star^{(0)} \hat{g}(X)$. By orthogonality (in L^2) of W_i to $\widehat{F}_\star^{(0)} \hat{g}(X)$ thus have that:

$$\begin{aligned}\sigma_U^2 &= \sqrt{\mathsf{E}\|U\|^4} \\ &= \sqrt{\mathsf{E}\|W\|^4 + \mathsf{E}\|F_\star^{(0)} g_\star(X) - \widehat{F}_\star^{(0)} \hat{g}(X)\|^4} \\ &\le \sqrt{\mathsf{E}\|W\|^4} + \sqrt{\mathsf{E}\|F_\star^{(0)} g_\star(X) - \widehat{F}_\star^{(0)} \hat{g}(X)\|^4} \qquad \text{(Triangle inequality)} \\ &\lesssim d_y \sigma_W^2 + \sqrt{C_{4\to 2}^{(0)}} \mathsf{E}\|F_\star^{(0)} g_\star(X) - \widehat{F}_\star^{(0)} \hat{g}(X)\|^2 \qquad \text{(hypercontractive estimate and sub-Gaussianity)} \\ &\lesssim d_y \sigma_W^2 + \frac{\sqrt{C_{4\to 2}^{(0)}}\nu^{-1}}{T} \sum_{t=1}^T \mathsf{E}^{(t)} \|\widehat{F}^{(t)} \hat{g}(X) - F_\star^{(t)} g_\star(X)\|_2^2 \qquad \text{(task-diversity assumption, Definition 2.1)}\end{aligned} \tag{16}$$

□

Proposition 2.13. *Suppose that $\mathsf{P}^{(0)}$ is stationary and ϕ-mixing and fix $\delta \in (0,1)$. Fix a block length k dividing $N'/2$. Define the blocked noise-class interaction term $V \triangleq \frac{1}{k}\sum_{i=1}^k U_i Z_i^\top \Sigma_Z^{(0)-1/2}$ and $\sigma_V^2 \triangleq \mathsf{E}^{(0)}\left[\|V\|_F^2\right]$. Define σ_U^2 and C_Z as in Proposition 2.9. Then, defining $C^{(0)} \triangleq C_Z\sqrt{C_{4\to 2}^{(0)}}$, with probability at least $1-\delta$ we have*

$$\begin{aligned}&\|(\widehat{F}^{(0)} - \widehat{F}_\star^{(0)})\sqrt{\Sigma_Z^{(0)}}\|_F^2 \\ &\lesssim \frac{\sigma_V^2 \log(1/\delta)}{N'} \lesssim \frac{C_Z \sigma_U^2 r \log(1/\delta)}{N'} + \\ &\frac{C^{(0)} k \nu^{-1} \log(1/\delta)}{N'T} \sum_{t=1}^T \mathsf{E}^{(t)} \|\widehat{F}^{(t)} \hat{g}(X) - F_\star^{(t)} g_\star(X)\|_2^2,\end{aligned}$$

as long as the burn-in $N' \gtrsim k\Big(r + B_{\mathcal{G}}^2 \log(1/\delta)\Big)$ *is satisfied and the block length is sufficiently long:* $k \geq N'\phi(k)\delta^{-1}$.

Proof. As noted the argument is identical to that presented in Proposition 2.9. The only difference is calculation of σ_V, detailed below.

By definition we have that (with summation ranging from 1 to k):

$$\begin{aligned}
\sigma_V^2 &= \frac{1}{k}\mathbf{tr}\left(\Sigma_Z^{-1/2}\mathsf{E}\left(\sum_{i,j} U_i^\top U_j Z_i Z_j^\top\right)\Sigma_Z^{-1/2}\right) \\
&\lesssim \frac{1}{k}\mathbf{tr}\left(\Sigma_Z^{-1/2}\mathsf{E}\left(\sum_{i} U_i^\top U_i Z_i Z_i^\top\right)\Sigma_Z^{-1/2}\right) + \frac{1}{k}\mathbf{tr}\left(\Sigma_Z^{-1/2}\mathsf{E}\left(\sum_{i\neq j} U_i^\top U_j Z_i Z_j^\top\right)\Sigma_Z^{-1/2}\right) \\
&\lesssim C_Z\sigma_U^2 + \frac{1}{k}\left(\left(\sum_{i\neq j}\mathsf{E}U_i^\top U_j\right)\right).
\end{aligned} \tag{17}$$

We now compute $\mathsf{E}U_i^\top U_j$. Notice that $U_i = Y_i - \widehat{F}_\star^{(0)}\hat{g}(X_i) = W_i + F_\star^{(0)}g_\star(X_i) - \widehat{F}_\star^{(0)}\hat{g}(X_i)$. Hence since the W_i, W_j for $i \neq j$ are orthogonal we have that:

$$\begin{aligned}
\mathsf{E}U_i^\top U_j &= \mathbf{E}\langle F_\star^{(0)}g_\star(X_i) - \widehat{F}_\star^{(0)}\hat{g}(X_i), F_\star^{(0)}g_\star(X_j) - \widehat{F}_\star^{(0)}\hat{g}(X_j)\rangle \\
&\lesssim \mathbf{E}\|F_\star^{(0)}g_\star(X_i) - \widehat{F}_\star^{(0)}\hat{g}(X_i)\|_2^2 + \mathbf{E}\|F_\star^{(0)}g_\star(X_j) - \widehat{F}_\star^{(0)}\hat{g}(X_j)\|_2^2.
\end{aligned} \tag{18}$$

The result follows by summation and using the task diversity condition, see Definition 2.1. □

A.3. Bounding the Estimation Error

The goal is to control the task-averaged estimation error. As previously discussed, the key observation is to quantify a lower isometry, such that

$$\frac{1}{T}\sum_{t=1}^{T}\mathsf{E}^{(t)}\|f^{(t)}\circ g(X) - f_\star^{(t)}\circ g_\star(X)\|_2^2 \lesssim \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\|f^{(t)}\circ g(x_i^{(t)}) - f_\star^{(t)}\circ g_\star(x_i^{(t)})\|_2^2.$$

By hypercontractivity, we have an anti-concentration result:

Proposition A.1 (Samson (2000, Theorem 2), Ziemann & Tu (2022, Prop. 5.1))**.** *Fix* $C > 0$. *Let* $g : \mathsf{X} \to \mathbb{R}$ *be a non-negative function satisfying*

$$\mathsf{E}[g(X)^2] \leq C\mathsf{E}[g(X)]^2.$$

Then we have:

$$\mathsf{P}\left[\frac{1}{m}\sum_{i=1}^{m} g(x_i) \leq \frac{1}{2}\mathsf{E}[g(X)]\right] \leq \exp\left(\frac{-m}{8C}\right).$$

Setting $g(X) \triangleq \|\overline{h}(X)\|_2^2$, Proposition A.1 yields a tail bound on the lower-isometry event for a *given* $\overline{h}$. By an application of the basic inequality (Rakhlin & Sridharan, 2014; Liang et al., 2015), an empirical estimation error can be bounded by

$$\begin{aligned}
\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\left\|h(x_i^{(t)})\right\|_2^2 &\leq \sup_{h\in\mathcal{H}}\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N} 4\left\langle w_i^{(t)}, h(x_i^{(t)})\right\rangle - \|h(x_i^{(t)})\|_2^2 &(19)\\
&\triangleq \mathsf{M}_{NT}(\mathcal{H}), &(20)
\end{aligned}$$

where $\mathsf{M}_{NT}(\mathcal{H})$ is denoted the (empirical) *martingale offset complexity* (Liang et al., 2015; Ziemann & Tu, 2022), which serves as the capacity measure of hypothesis class $\mathcal{H}$. Notably, $\mathsf{M}_{NT}(\mathcal{H})$ scales with the *noise-level* σ_w^2, rather than the diameter of $\mathcal{H}$. We control $\mathsf{M}_{NT}(\mathcal{H})$ via a high-probability chaining bound, derived from Ziemann (2022, Theorem 4.2.2):

Lemma A.2 (High-probability chaining bound)**.** *Let Assumption 1.2 hold, and fix $\delta \in (0,1)$. There exists a universal constant $c > 0$ such that given a function class $\mathcal{H}$, with probability at least $1-\delta$,*

$$\mathsf{M}_{NT}(\mathcal{H}) \le c \cdot \inf_{\gamma > 0} \left\{ \sigma_w \sqrt{d_y} \gamma \log(1/\delta) + \frac{\sigma_w^2}{NT} \log(1/\delta) + \frac{\sigma_w^2 \log \mathsf{N}_\infty(\mathcal{H}, \gamma)}{NT} + \frac{\sigma_w \gamma \sqrt{\log(1/\delta)}}{\sqrt{NT}} + \gamma^2 \right\}, \tag{21}$$

where $\mathsf{N}_\infty(\mathcal{H}, \gamma)$ is the covering number of $\mathcal{H}$ at resolution γ under the metric $\rho(h_1, h_2) = \sup_{x \in \mathsf{X}} \|h_1(x) - h_2(x)\|$.

In particular, Lemma A.2 suggests that the martingale complexity is solely a function of the class $\mathcal{H}$, and not the statistics of the data. Roughly speaking, we can choose γ to be whatever is required such that the log-covering number term is dominant, as γ manifests only logarithmically there. To determine what $\mathcal{H}$ to cover, we use the following localization result from Ziemann & Tu (2022, Theorem 5.1).

Proposition A.3. *Let Assumption 2.8 hold with $C_{4\to 2}^{1:T}$. Defining $\overline{\mathcal{H}}_\star$ as the star-hull[7] of $\overline{\mathcal{H}}$, $B(r) \triangleq \{h \in \overline{\mathcal{H}}_\star \big| \mathsf{E}^{1:T}\|h\|_2^2 \le r^2\}$, and $\partial B(r)$ the boundary of $B(r)$. Then, there exists a $r/\sqrt{8}$-net $\overline{\mathcal{H}}_\star(r)$ in the $\|\cdot\|_\infty$ of $\partial B(r)$ such that*

$$\mathsf{P}\left[\inf_{\overline{h} \in \overline{\mathcal{H}}_\star \setminus B(r)} \left\{ \widehat{\mathsf{E}}_N^{1:T} \|\overline{h}\|_2^2 - \frac{1}{8} \mathsf{E}^{1:T} \|\overline{h}\|_2^2 \right\} \le 0 \right] \le \left|\overline{\mathcal{H}}_\star(r)\right| \exp\left(\frac{-NT}{8 C_{4\to 2}^{1:T}} \right). \tag{22}$$

We note that the star-hull subsumes the original class, and thus for a given r, we have for any $\overline{h} \in \overline{\mathcal{H}}$, with probability at least $1 - \left|\overline{\mathcal{H}}_\star(r)\right| \exp\left(-NT/8C_{4\to 2}^{1:T}\right)$:

$$\mathsf{E}^{1:T}\|\overline{h}\|_2^2 \le \max\{8\mathsf{M}_{NT}(\overline{\mathcal{H}}_\star), r^2\}. \tag{23}$$

Thus, we should choose r such that the two terms meet at the desired rate, order-wise. The union bound over $\overline{\mathcal{H}}_\star(r)$ in the failure probability turns into a burn-in condition on NT. As the last step before the final bound, we derive the following covering bound:

Lemma A.4. *Under Assumption 2.7, and recalling $\overline{\mathcal{H}}_\star$ is the star-hull of $\overline{\mathcal{H}}$, we have*

$$\log \mathsf{N}_\infty(\overline{\mathcal{H}}_\star, \varepsilon) \le T d_y r \log\left(1 + \frac{4TB_{\mathcal{F}}B_{\mathcal{G}}}{\varepsilon}\right) + \log\left(1 + \frac{2B_{\mathcal{F}}B_{\mathcal{G}}}{\varepsilon}\right) + \log \mathsf{N}_\infty\left(\mathcal{G}, \frac{\varepsilon}{4B_{\mathcal{F}}}\right).$$

Proof. Firstly, noting that $\overline{\mathcal{H}}_\star$ is trivially $2B_{\mathcal{F}}B_{\mathcal{G}}$-bounded by Assumption 2.7, we invoke Mendelson (2002, Lemma 4.5) to show the covering number of star-hull of $\overline{\mathcal{H}}$ incurs only a negligible additive factor to the log-covering number.

$$\log \mathsf{N}_\infty(\overline{\mathcal{H}}_\star, \varepsilon) \le \log \mathsf{N}_\infty(\overline{\mathcal{H}}, \varepsilon/2) + \log\left(1 + \frac{2B_{\mathcal{F}}B_{\mathcal{G}}}{\varepsilon}\right).$$

It remains to demonstrate how a covering of $\mathcal{F}^{\otimes T}$ and $\mathcal{G}$ witnesses an ε-covering of $\overline{\mathcal{H}}$. Given $h_1, h_2 \in \overline{\mathcal{H}}$, define the we start with the $\|\cdot\|_\infty$ norm:

$$\begin{aligned}
\|h_1 - h_2\|_\infty &\triangleq \sup_{t\in[T]} \sup_{x \in \mathsf{X}} \left\| F_1^{(t)} g_1(x) - F_2^{(t)} g_2(x) \right\|_2 \\
&\le \sup_{t\in[T]} \sup_{x \in \mathsf{X}} \left\| (F_1^{(t)} - F_2^{(t)}) g_1(x) \right\|_2 + \left\| F_2^{(t)} (g_1 - g_2) \right\|_\infty && \text{(add and subtract, triangle ineq.)} \\
&\le \sup_{t\in[T]} B_{\mathcal{G}} \left\| F_1^{(t)} - F_2^{(t)} \right\|_2 + B_{\mathcal{F}} \|g_1 - g_2\|_\infty && \text{(Cauchy-Schwarz, boundedness)} \\
&\le B_{\mathcal{G}} \sqrt{T} \|\mathbf{F}_1 - \mathbf{F}_2\|_F + B_{\mathcal{F}} \|g_1 - g_2\|_\infty,
\end{aligned}$$

where we define the task-stacked matrix $\mathbf{F} \triangleq \begin{bmatrix} F^{(1)} & \cdots & F^{(T)} \end{bmatrix} \in \mathbb{R}^{d_y \times Tr}$. Therefore, to witness a ε-covering of $\overline{\mathcal{H}}$, it suffices to cover $\mathbb{R}^{d_y \times Tr} \sim \mathbb{R}^{T d_y r}$ at resolution $\frac{\varepsilon}{2B_{\mathcal{G}}\sqrt{T}}$ in the Euclidean norm $\|\cdot\|_2$, and $\mathcal{G}$ at resolution $\frac{\varepsilon}{2B_{\mathcal{F}}}$ in the

[7] $\mathcal{H}_\star \triangleq \mathrm{StarHull}(\mathcal{H}) = \{\alpha h,\ h \in \mathcal{H}, \alpha \in [0,1]\}$.

sup-norm $\|\cdot\|_\infty$. We recall by Assumption 2.7 that $\mathcal{F}$ is identified by an ℓ^2-ball of radius $B_\mathcal{F}$, and thus by standard volumetric arguments (e.g. Wainwright (2019, Example 5.8)), we combine bounds and recover

$$\log \mathsf{N}_\infty(\overline{\mathcal{H}}_\star, \varepsilon) \le T d_y r \log\left(1 + \frac{4TB_\mathcal{F}B_\mathcal{G}}{\varepsilon}\right) + \log\left(1 + \frac{2B_\mathcal{F}B_\mathcal{G}}{\varepsilon}\right) + \log \mathsf{N}_\infty\left(\mathcal{G}, \frac{\varepsilon}{4B_\mathcal{F}}\right).$$

□

Recalling that $\mathcal{G}$ is finite, and balancing choices of γ and r yields a bound the task-averaged estimation error.

Proposition 2.11. *Let Assumption 2.7 hold. Then, with probability at least $1-\delta$, the estimation error of ERM predictors $\{\widehat{F}^{(t)}\}_{t=1}^T, \hat{g}$ is bounded by*

$$\frac{1}{T}\sum_{t=1}^T \mathsf{E}^{(t)} \|\widehat{F}^{(t)}\hat{g}(X) - F_\star^{(t)} g_\star(X)\|_2^2 \le \sigma_W^2 \cdot \tilde{\mathcal{O}}\left(\frac{d_\mathsf{Y} r}{N} + \frac{\log|\mathcal{G}| + \log(1/\delta)}{NT}\right),$$

as long as $N \ge C_{4\to 2}^{1:T} \cdot \tilde{\Omega}\left(d_\mathsf{Y} r + \log|\mathcal{G}|/T + \log(1/\delta)\right)$.

Proof. Observing the chaining bound from Lemma A.2, we may choose $\gamma \asymp \frac{\sigma_w}{NT\sqrt{d_y}}$, and apply to $\overline{\mathcal{H}}_\star$ such that

$$\begin{aligned}\mathsf{M}_{NT}(\overline{\mathcal{H}}_\star) &\lesssim \frac{\sigma_w^2}{NT}\log(1/\delta) + \frac{\sigma_w^2 \log \mathsf{N}_\infty\left(\overline{\mathcal{H}}_\star, \frac{\sigma_w}{NT\sqrt{d_y}}\right)}{NT} + \frac{\sigma_w^2\sqrt{\log(1/\delta)}}{(NT)^{3/2}\sqrt{d_y}} + \frac{\sigma_w^2}{(NT)^2 d_y} \\ &\lesssim \frac{\sigma_w^2}{NT}\left(\log \mathsf{N}_\infty\left(\overline{\mathcal{H}}_\star, \frac{\sigma_w}{NT\sqrt{d_y}}\right) + \log(1/\delta)\right).\end{aligned}$$

Now applying the covering number bound Lemma A.4, we get

$$\log \mathsf{N}_\infty\left(\overline{\mathcal{H}}_\star, \frac{\sigma_w}{NT\sqrt{d_y}}\right) \le T d_y r \log\left(1 + \frac{4T^2 N\sqrt{d_y} B_\mathcal{F} B_\mathcal{G}}{\sigma_w}\right) + \log\left(1 + \frac{2TN\sqrt{d_y} B_\mathcal{F} B_\mathcal{G}}{\sigma_w}\right) + \log|\mathcal{G}|,$$

thus our desired rate $r^2 \asymp \frac{\sigma_w^2}{NT}\log \mathsf{N}_\infty\left(\overline{\mathcal{H}}_\star, \frac{\sigma_w}{NT\sqrt{d_y}}\right)$, yielding with probability at least $1 - \delta - \left|\overline{\mathcal{H}}_\star(r/\sqrt{8})\right| \exp\left(-NT/8C_{4\to 2}^{1:T}\right)$,

$$\frac{1}{T}\sum_{t=1}^T \mathsf{E}^{(t)} \|\widehat{F}^{(t)}\hat{g}(X) - F_\star^{(t)} g_\star(X)\|_2^2 \le \sigma_W^2 \cdot \tilde{\mathcal{O}}\left(\frac{d_y r}{N} + \frac{\log|\mathcal{G}| + \log(1/\delta)}{NT}\right).$$

Now inverting $\left|\overline{\mathcal{H}}_\star(r/\sqrt{8})\right| \exp\left(-NT/8C_{4\to 2}^{1:T}\right) \le \delta$ yields the burn-in requirement

$$N \ge C_{4\to 2}^{1:T} \cdot \tilde{\Omega}\left(d_y r + \log|\mathcal{G}|/T + \log(1/\delta)\right).$$

□

By Lemma 2.2, we simply sum up the bounds from Proposition 2.9 and Proposition 2.11 and apply Proposition 2.6 to specify ν^{-1}, which yields

Theorem 2.12 (Transfer risk bound)**.** *Assume $\mathsf{P}^{0:T}$ satisfy μ-(TC), and let Assumption 2.7 hold. With probability at least $1-\delta$, the target excess risk of the two-stage ERM predictor $(\widehat{F}^{(0)}, \hat{g})$ is bounded by*

$$\begin{aligned}\mathrm{ER}(\widehat{F}^{(0)}, \hat{g}) &\le \frac{\sigma_W^2 C_Z d_\mathsf{Y} r \log(1/\delta)}{N'} \\ &+ \sigma_W^2 \mu \|\mathbf{F}_\star^{(0)}(\mathbf{F}_\star^{1:T})^\dagger\|_2 \cdot \tilde{\mathcal{O}}\left(\frac{d_\mathsf{Y} r}{N} + \frac{\log|\mathcal{G}| + \log(1/\delta)}{NT}\right),\end{aligned}$$

as long as $N' \gtrsim C_Z\sqrt{C_{4\to 2}^{(0)}} r + B_\mathcal{G}^2 \log(1/\delta)$ and $N \gtrsim C_{4\to 2}^{1:T} \cdot \tilde{\Omega}\left(d_\mathsf{Y} r + \log|\mathcal{G}|/T + \log(1/\delta)\right)$.

The modified burn-in on N' comes from Lemma 2.10, where the additive error from misspecification in σ_U^2 when expanded is proportional to

$$\frac{rC_Z\sqrt{C_{4\to 2}^{(0)}}}{N'}\frac{\nu^{-1}}{T}\sum_{t=1}^{T}\mathsf{E}^{(t)}\|\widehat{F}^{(t)}\hat{g}(X)-F_\star^{(t)}g_\star(X)\|_2^2.$$

Therefore, it suffices to inflate the existing burn-in on N' by an additive $\approx C_z\sqrt{C_{4\to 2}^{(0)}}r$ factor so that the estimation error terms merge.

To extend bounds to the ϕ-mixing, beyond the legwork done Appendix A.2, very little changes for the estimation error bounds, apart from the sole modification in Samson's Theorem:

Proposition A.5 (Samson (2000, Theorem 2), Ziemann & Tu (2022, Prop. 5.1))**.** *Fix $C>0$. Assume $\{X\}_{i\geq 1}\sim \mathsf{P}$ is ϕ-mixing and admits dependency matrix $\Gamma_{\mathrm{dep}}(\mathsf{P})$. Let $g:\mathsf{X}\to\mathbb{R}$ be a non-negative function satisfying*

$$\mathsf{E}[g(X)^2]\leq C\mathsf{E}[g(X)]^2.$$

Then we have:

$$\mathsf{P}\left[\frac{1}{m}\sum_{i=1}^{m}g(x_i)\leq\frac{1}{2}\mathsf{E}[g(X)]\right]\leq\exp\left(\frac{-m}{8C\,\|\Gamma_{\mathrm{dep}}(\mathsf{P})\|_2^2}\right).$$

Using the bound following Definition B.2, defining $\Phi\triangleq\left(\sum_{i=1}^{\infty}\sqrt{\phi_X(i)}\right)^2$, we can follow the exact same steps above for the iid case to yield:

Theorem 2.14 (Transfer risk bound, mixing)**.** *Suppose that $P^{0:T}$ are each stationary and ϕ-mixing and fix $\delta\in(0,1)$. Assume that k is fixed and divides $N'/2$ and $N/2$. Define the quantity $\Phi\triangleq(\sum_{i=1}^{\infty}\sqrt{\phi(i)})^2$. Assume $\mathsf{P}^{0:T}$ satisfy μ-(TC), and let Assumption 2.7 hold. Define $\mathrm{C}(\mathcal{G})\triangleq\log\mathsf{N}_\infty\left(\mathcal{G},\frac{B_{\mathcal{F}}NT\sqrt{d_{\mathsf{Y}}}}{\sigma_W}\right)$. With probability at least $1-\delta$, the target excess risk is bounded by*

$$\begin{aligned}\mathrm{ER}(\widehat{F}^{(0)},\hat{g})&\leq\frac{\sigma_W^2C_Zd_{\mathsf{Y}}r\log(1/\delta)}{N'}\\&+\sigma_W^2\mu\|\mathbf{F}_\star^{(0)}(\mathbf{F}_\star^{1:T})^\dagger\|_2\cdot\tilde{\mathcal{O}}\left(\frac{d_{\mathsf{Y}}r}{N}+\frac{\mathrm{C}(\mathcal{G})+\log(1/\delta)}{NT}\right),\end{aligned}$$

as long as $N'\gtrsim k\left(C_Z\sqrt{C_{4\to 2}^{(0)}}r+B_{\mathcal{G}}^2\log(1/\delta)\right)$ and $N\gtrsim C_{4\to 2}^{1:T}\Phi\cdot\tilde{\Omega}\left(d_{\mathsf{Y}}r+\mathrm{C}(\mathcal{G})/T+\log(1/\delta)\right)$.

Note that the burn-in for N now has an additional factor of Φ.

B. Properties of Mixing Sequences of Random Variables

In Section 2.3 we extend our analysis to mixing random variables. This requires some additional machinery. Namely, for a sequence of random variables $Z_{1:n}$ we partition $[n]$ into $2m$ consecutive intervals, denoted a_j for $j\in[2m]$, so that $\sum_{j=1}^{2m}|a_j|=n$. Denote further by O (resp. by E) the union of the oddly (resp. evenly) indexed subsets of $[n]$. We further abuse notation by writing $\phi_Z(a_i)=\phi_Z(|a_i|)$ in the sequel. We will typically instantiate the below machinery with all partitions of equal length k, but for now describe the general setup.

We split the process $Z_{1:n}$ as:

$$Z_{1:|O|}^o\triangleq(Z_{a_1},\ldots,Z_{a_{2m-1}}),\quad Z_{1:|E|}^e\triangleq(Z_{a_2},\ldots,Z_{a_{2m}}).\tag{24}$$

Let $\tilde{Z}_{1:|O|}^o$ and $\tilde{Z}_{1:|E|}^e$ be blockwise decoupled versions of (24). That is we posit that $\tilde{Z}_{1:|O|}^o\sim\mathsf{P}_{\tilde{Z}_{1:|O|}^o}$ and $\tilde{Z}_{1:|E|}^e\sim\mathsf{P}_{\tilde{Z}_{1:|E|}^e}$, where:

$$\mathsf{P}_{\tilde{Z}_{1:|O|}^o}\triangleq\mathsf{P}_{Z_{a_1}}\otimes\mathsf{P}_{Z_{a_3}}\otimes\cdots\otimes\mathsf{P}_{Z_{a_{2m-1}}}\quad\text{and}\quad\mathsf{P}_{\tilde{Z}_{1:|E|}^e}\triangleq\mathsf{P}_{Z_{a_2}}\otimes\mathsf{P}_{Z_{a_4}}\otimes\cdots\otimes\mathsf{P}_{Z_{a_{2m}}}.\tag{25}$$

The process $\tilde{Z}_{1:n}$ with the same marginals as $\tilde{Z}^o_{1:|O|}$ and $\tilde{Z}^e_{1:|E|}$ is said to be the decoupled version of $Z_{1:n}$. To be clear: $\mathsf{P}_{\tilde{Z}_{1:n}} \triangleq \mathsf{P}_{Z_{a_1}} \otimes \mathsf{P}_{Z_{a_2}} \otimes \cdots \otimes \mathsf{P}_{Z_{a_{2m}}}$, so that $\tilde{Z}^o_{1:|O|}$ and $\tilde{Z}^e_{1:|E|}$ are alternatingly embedded in $\tilde{Z}_{1:n}$. The following result is key—by skipping every other block, $\tilde{Z}_{1:n}$ may be used in place of $Z_{1:n}$ for evaluating scalar functions at the cost of an additive mixing-time-related term.

Proposition B.1 (Lemma 2.6 in (Yu, 1994) instantiated to ϕ-mixing processes)**.** *Fix a ϕ-mixing process $Z_{1:n}$ and let $\tilde{Z}_{1:n}$ be its decoupled version. For any measurable function f of $Z^o_{1:|O|}$ (resp. g of $Z^e_{1:|E|}$) with joint range $[0,1]$ we have that:*

$$\begin{aligned} |\mathbf{E}(f(Z^o_{1:|O|})) - \mathbf{E}(f(\tilde{Z}^o_{1:|O|}))| &\leq \sum_{i \in E \setminus \{2m\}} \phi_Z(a_i), \\ |\mathbf{E}(g(Z^e_{1:|E|})) - \mathbf{E}(g(\tilde{Z}^e_{1:|E|}))| &\leq \sum_{i \in O \setminus \{1\}} \phi_Z(a_i). \end{aligned} \tag{26}$$

The above proposition is originally stated for β-mixing random variables in Yu (1994), but these coefficients always dominate the ϕ-mixing coefficients and so the result remains true in our setting.

We will also require a second notion of dependency.

Definition B.2 (Dependency matrix, Samson (2000, Section 2))**.** The *dependency matrix* of a process $Z_{1:n}$ with distribution P_Z is the (upper-triangular) matrix $\Gamma_{\mathsf{dep}}(\mathsf{P}_Z) = \{\Gamma_{ij}\}_{i,j=0}^{T-1} \in \mathbb{R}^{n \times n}$ defined as follows. Let $\mathcal{Z}_{1:i+1}$ denote the σ-algebra generated by $Z_{1:i+1}$. For indices $i < j$, let

$$\Gamma_{ij} = \sqrt{2 \sup_{A \in \mathcal{Z}_{1:i+1}} \|\mathsf{P}_{Z_{j+1:n}}(\cdot \mid A) - \mathsf{P}_{Z_{j+1:n}}\|_{\mathsf{TV}}}. \tag{27}$$

For the remaining indices $i \geq j$, let $\Gamma_{ii} = 1$ and $\Gamma_{ij} = 0$ when $i > j$ (below the diagonal).

It is straightforward to verify—and we will use—that

$$\|\Gamma_{\mathsf{dep}}(\mathsf{P}_Z)\| \leq \sum_{i=1}^{\infty} \sqrt{\phi_Z(i)}. \tag{28}$$

C. Additional Numerical Details

We consider the simulation task of balancing a pole atop a cart from visual observations, as pictured in Figure 1(a). This experimental setup is used to demonstrate the benefit of multi-task imitation learning (compared to single task imitation learning) for a visuomotor control task. We first describe the system, and how expert policies are generated. We then provide details about the imitation learning and evaluation process.

System Description: The pole is balanced by applying a force to the cart along a track. Denoting the position of the cart by p and the angle of the pole by θ, the system evolves according to the following dynamics:

$$\begin{aligned} u &= (M+m)(\ddot{p} + d_p \dot{p}) + m\ell((\ddot{\theta} + d_\theta \dot{\theta})\cos\theta - \dot{\theta}^2 \sin\theta), \\ 0 &= m((\ddot{p} + d_p \dot{p})\cos\theta + \ell(\ddot{\theta} + d_\theta \dot{\theta}) - g\sin\theta). \end{aligned}$$

Here, M is the mass of the cart, m is the mass of the pole, ℓ is the length of the pole, g is the acceleration due to gravity, k_p is the damping coefficient for cart on the track, and k_θ is the damping coefficient for the joint of the pole with the cart. The state of this system at time t is denoted $x_t = \begin{bmatrix} p_t & \dot{p}_t & \theta_t & \dot{\theta}_t \end{bmatrix}^\top$. These dynamics are discretized via an euler approximation with stepsize $dt = 0.02$. The discrete time dynamics will be written $x_{t+1} = f(x_t, u_t)$. We further suppose that we have a camera setup next to the track, directed towards the track and centered at the zero position of the cart. This camera gives us a partial observation of the state at any time: $o_t = \mathrm{camera}(x_t)$. Figure 1(a) is one such observation generated by the PyBullet simulator when the system is at the origin. We consider a collection of instances of this system by uniformly randomly sampling $M, m, \ell \in [0.5, 3.0] \times [0.05, 0.2] \times [1.0, 2.5]$, and setting $g = 9.8$, $k_p = k_\theta = 0.4$.

Expert Policy Description: The expert has access to a (noisy) key-point extractor that maps the image observations from the camera to a vector containing the position of the cart-pole joint along the track, the position of the pole tip along the track, and the height of the pole tip above the track. This provides the two keypoints illustrated in Figure 1(b)[8]. We denote this noisy observation as $\mathrm{keypoint}(o_t)$. A single keypoint extractor is used by all experts (across the parameter variations of the system), and is trained from labeled data across a variety of parameter settings. After applying the keypoint extractor to the images, the ideal measurements become a simple function of p and θ: they may be written $\begin{bmatrix} p_t & p_t + \sin(\theta_t)\ell & \cos(\theta_t)\ell \end{bmatrix}^\top$. As such, we can construct expert controllers using the dynamics of the system by synthesizing LQG controllers[9] for the system linearized about the upright equilibrium point. In particular, for some particular parameter realization, indexed by h, the corresponding expert controller generates the force $u_t^\star$ applied to the cart at time t as

$$\begin{aligned} \xi_{t+1} &= A_K^{(h)} \xi_t + B_K^{(h)} \mathrm{keypoint}\left(\mathrm{camera}(x_t)\right) \\ u_t^\star &= C_K^{(h)} \xi_t + D_K^{(h)} \mathrm{keypoint}\left(\mathrm{camera}(x_t)\right), \end{aligned}$$

where $(A_K^{(h)}, B_K^{(h)}, C_K^{(h)}, D_K^{(h)})$ are constructed from two Riccati equation solutions involving the linearized system, and ξ_t is a four dimensional latent state. We assume that when the input applied is applied to the system, there is an unobserved actuation noise added. Therefore, the input applied to the system at time t by the expert controller will be $u_t = u_t^\star + \eta_t$, where $\eta_t \sim \mathcal{N}(0, 0.5)$.

Imitation Learning Policy Description: We consider imitation learning agents that operate a short history of camera observations[10]. In particular, the learning agent selects inputs as

$$\hat{u}_t = K_\theta \left(\begin{bmatrix} \mathrm{camera}(x_t) \\ \vdots \\ \mathrm{camera}(x_{t-\mathrm{hist}}) \end{bmatrix} \right).$$

Here K_θ is a convolutional neural network with parameters θ. In the single task setting, the parameters are specific to the parameter realization for the task at hand. In the multi-task setting, the network parameters are partitioned into a shared component θ_{shared} and a task specific component for the final layer, θ_h.

First Stage: The shared parameters in the multi-task setting are jointly trained on a collection of H source tasks.[11]. The dataset therefore consists of demonstrations from rollouts of the expert controllers generated for H systems with different parameter realizations. Expert demonstrations are obtained from 10 independent realizations of the actuation noise sequence for each system. The length of the rollout trajectory is 500 steps (recalling the discretization timestep of 0.02.)

The multi-task network is jointly trained on the entire collection of source data to minimize the loss

$$\sum_{h=1}^{H} \sum_{i=1}^{10} \sum_{t=1}^{500} \left\| u_t^{(h)}[i] - K_{\theta_{\mathrm{shared}}, \theta_h} \left(\begin{bmatrix} \mathrm{camera}(x_t^{(h)}[i]) \\ \vdots \\ \mathrm{camera}(x_{t-\mathrm{hist}}^{(h)}[i]) \end{bmatrix} \right) \right\|^2$$

over the network parameters $\theta_{\mathrm{shared}}, \theta_1, \ldots, \theta_H$. The superscript h on the inputs and states denotes the system index that they came from, while the argument in the brackets enumerates the 10 expert trajectories collected from each system. To obtain an approximate minimizer to the above problem, we employ the adam optimizer using a batch size of 32, weight decay of $1e^{-3}$, and learning rate of $1e^{-3}$ with a decay factor of 0.5 every 10 epochs for a total of 100 epochs.[12]

Second Stage: The second stage consists of 10 target tasks, defined by new parameter realizations for the cartpole system. We compare:

[8] In our experiments, the keypoint extractor is a convolutional neural network trained on a 50000 cartpole images from instances drawn uniformly at random with states having position $p \in [-3, 3]$, $\theta \in [-\pi/3, \pi/3]$, and pole lengths $\ell \in [1, 2.5]$.

[9] We use $Q = R = \Sigma_w = \Sigma_v = I$.

[10] We use a history of 8.

[11] We consider three values of H: $H = 5, 10, 20$.

[12] Tasks are mixed together in the each batch.

1. Training a convolutional neural network for each of these tasks from scratch using the data available for the task (this is single task imitation learning).

2. Re-using the representation trained for the collection of source tasks along with a head that is fit to the target task. The head is obtained by solving a least squares problem by computing the shared representation for the history of camera observations in the expert demonstrations and solving a regression problem to match the expert inputs.

For each target task, we again collect expert demonstrations. Here, we consider a variable number of trajectories, N_{target}. Each trajectory is again obtained by rolling out the corresponding expert controller for 500 steps under new, independent realizations of the actuation noise. These expert trajectories are used to fit a linear head for the corresponding target tasks for the multi-task setting, and to train a behavior cloning agent from scratch for the single task setting.

Evaluation Results: Once the target controllers are trained, we evaluate them by rolling them out on the cartpole system with the parameters for which they were designed. These evaluation rollouts occur by rolling out out the single-task learned, multi-task learned, and expert controller under new realizations of the actuation noise. We track the input imitation error over the entire trajectory, which is the MSE of the gap between the inputs applied by the expert, and the inputs a learned controller $\hat{K}$ would apply when faced with the same observations:

$$\sum_{t=1}^{500} \left\| u_t^\star - \hat{K} \left(\begin{bmatrix} \text{camera}(x_t^\star) \\ \vdots \\ \text{camera}(x_{t-\text{hist}}^\star) \end{bmatrix} \right) \right\|^2.$$

We additionally track the state imitation error between the states $\hat{x}_t$ from rolling out the learned controller and the states $x_t^\star$ from rolling out the expert controller:

$$\sum_{t=1}^{500} \|x_t^\star - \hat{x}_t\|^2.$$

We also track whether the controller lasts 500 steps without allowing the pole to fall past an angle of $\pi/2$ in either direction. We plot the results for representation learning with 5, 10, or 20 source tasks, in addition to single task learning. The evaluation metrics are averaged across 50 evaluation rollouts for each target controller. In Figure 2, the median is plotted, with the 30%-70% quantiles are shaded. The median and quantiles are over 10 random seeds for the target tasks and 5 random seeds for the parameters of the source task instances. In the low data regime, multi-task learning excels in all metrics, with increasing benefit as more source tasks are available. In the high data regime, the single task controller eventually beats out the multi-task controllers for all metrics.

In Figure 3, we plot the input imitation error versus the number of source tasks available for pre-training on a $\log - \log$ scale with the number of target trajectories fixed at 100. Neglecting the component of the error that decays with the number of target trajectories, our theoretical results predict a decay in the error of $\frac{1}{H}$, or a slope of -1 on a $\log \log$ plot. In Figure 3, we observe a slope of approximately -0.8. The discrepancy may arise for several reasons. Firstly, the empirical risk minimizer is approximated using SGD. Secondly, the number of target trajectories used for fitting the final layer of the network is not infinite, meaning that we occur some additional error in training the final layer.

D. Open Questions

We list some open questions remaining following our analysis.

Can $(\text{poly})\log(\#\text{Tasks})$ be removed from the burn-in? To the best of our knowledge, all theoretical representation learning works that attain multi-task rates–in either the linear or non-linear representation setting–inevitably contain a term scaling with $\text{polylog}(T)$, ours no exception. In certain cases, e.g. (Du et al., 2020; Collins et al., 2021; Zhang et al., 2023b), this inevitably arises due to a union bound over a desirable event per-task. In our analysis, it arises as a byproduct of covering over T balls in $\mathbb{R}^{d_y \times r}$. A $\text{polylog}(T)$ term in the burn-in is unintuitive, as it suggests for a fixed per-task sample size N, the benefit of multi-task learning only provably extends to an upper bound on #tasks. For $d_y = 1$, and r constant (Thekumparampil et al., 2021), this equates to an upper bound of $T \approx \exp(r)$ tasks.

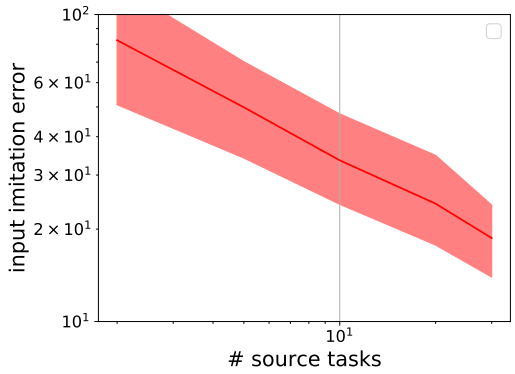


Figure 3. Input imitation error of the policies trained with a shared representation plotted against the number of source tasks used to train the representation on a log log scale. The number of target trajectories used for finetuning is fixed at 100.

For $\mathcal{F} \subseteq \mathbb{R}^{d_y \times r}$, can the burn-in be stated independently of d_x, d_y? Following our chaining bound, we are able to yield a burn-in with dominant term $d_y r$. While this is a vast improvement in previously considered scalar settings $d_y = 1$, it is not intuitively the correct order. For a fixed g, by linearity of $\mathcal{F}$, only r samples are required per task for the task-specific heads to be well-posed. Furthermore, it is known that chaining-type analysis can lead to suboptimal burn-ins for certain settings (Ziemann & Tu, 2022).