Proceedings of Machine Learning Research vol 313:1–44, 2026 37th International Conference on Algorithmic Learning Theory

Bridging Lifelong and Multi-Task Representation Learning: An Algorithm and a Complexity Measure

Zhi Wang ZHI.WANG@WISC.EDU
University of Wisconsin–Madison

Chicheng Zhang CHICHENGZ@ARIZONA.EDU
University of Arizona

Ramya Korlakai Vinayak RAMYA@ECE.WISC.EDU
University of Wisconsin–Madison

Editors: Matus Telgarsky and Jonathan Ullman

Abstract

In lifelong learning, a learner faces a sequence of tasks with shared structure and aims to identify and leverage it to accelerate learning. We study the setting where such structure is captured by a common representation of data. Unlike multi-task learning or learning-to-learn, where tasks are available upfront to learn the representation, lifelong learning requires the learner to make use of its existing knowledge while continually gathering partial information in an *online* fashion. In this paper, we consider a generalized framework of lifelong representation learning. We propose a simple algorithm that uses multi-task empirical risk minimization as a subroutine and establish a sample complexity bound based on a new notion we introduce—the *task-eluder dimension*. Our result applies to a wide range of learning problems involving general function classes. As concrete examples, we instantiate our result on classification and regression tasks under noise.

Keywords: Multi-task learning, representation learning, lifelong learning, online learning

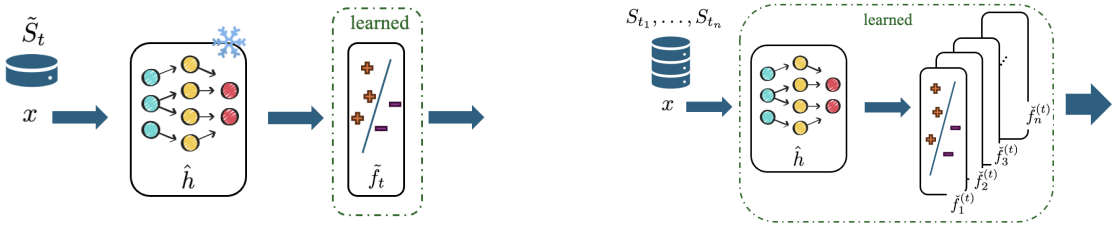
1. Introduction

In many real-world settings, learning naturally involves a collection of related tasks (Caruana, 1997). The ability to identify and leverage shared structure among tasks allows a learner to transfer knowledge and accelerate learning. One common form of structure lies in a shared *representation* of data such that simple functions operating on it can support effective and efficient learning across tasks. For example, adapting linear classifiers over a pre-trained set of deep neural network features has demonstrated state-of-the-art performance in computer vision (Donahue et al., 2014).

The benefit of representation transfer is a central topic in the study of multi-task learning (MTL) and learning to learn (LTL) (Baxter, 2000; Maurer et al., 2016; Tripuraneni et al., 2020; Aliakbarpour et al., 2024). In MTL, the learner is given a fixed set of tasks and aims to jointly learn a shared representation and task-specific prediction layers. In LTL, also known as meta-learning, tasks are drawn from an unknown distribution: the learner is first trained on a collection of tasks and then evaluated on a new task sampled from the same distribution. In both cases, tasks are available upfront, and LTL often relies on having enough *diversity* in seen tasks for the learner to fully identify the representation before applying it in an unseen task (Tripuraneni et al., 2020; Du et al., 2021).

However, learning may unfold over time, with tasks arriving in a sequence, as is the case in how humans learn. On the one hand, the learner *should* be able to begin making use of the shared structure before it is fully uncovered. On the other hand, the learner can always continually gather

© 2026 Z. Wang, C. Zhang & R.K. Vinayak.



(a) Few-shot property test (b) Multi-task ERM

Figure 1: Our algorithm maintains a representation $\hat{h}$. When a new task arrives, the algorithm first runs a few-shot property test to check whether $\hat{h}$ admits a prediction layer with low excess risk. If not, it performs MTL on data from a subset of past tasks and updates $\hat{h}$.

partial information about the structure and refine its internal representation. This motivates the study of *lifelong representation learning*, where the learner aims to identify and utilize a common representation in an online manner to reduce the sample complexity over the sequence of tasks.

While MTL and LTL with a shared representation have been extensively studied, theoretical understanding of the lifelong/online regime remains relatively limited (see Section 6 for related work). For binary classification under the noiseless, realizable assumption, Balcan et al. (2015) study lifelong learning of linear classifiers that lie in a low-dimensional linear subspace, while Pentina and Urner (2016) analyze the setting where new tasks may be represented by weighted majority votes over previously learned hypotheses from any ground class. More recently, Cao et al. (2022) consider shared one-hidden-layer representations, again under the noiseless, realizable assumption. In this work, our goal is to develop an algorithm and a theory for composite predictors over general function classes in noisy but well-specified settings (Section 2). Our main contributions are:

1. We study a simple, provably efficient algorithm (Section 3) for lifelong representation learning with composite predictors $f_t \circ h$, where $h \in \mathcal{H}$ is a shared representation (e.g., a deep neural network) and $f_t \in \mathcal{F}$ is a prediction layer specialized to task t (e.g., a linear function). Our algorithm is composed of two modular subroutines, few-shot property test and multi-task empirical risk minimization (ERM); see Figure 1. This design points toward a connection between the theory of lifelong learning and more practical algorithmic frameworks in real-world MTL.

2. We provide theoretical guarantees on the sample and space complexities of our algorithm (Section 4.3) based on a new complexity measure, the *task-eluder dimension* (Section 3.2). Together, our algorithm and complexity measure bridge lifelong learning and MTL: multi-task ERM serves as a mechanism for refining the learner's representation in lifelong learning, and the task-eluder dimension bounds how many times it needs to be invoked. When the task-eluder dimension is small, the benefit of representation transfer in lifelong learning becomes more evident.

3. To demonstrate the applicability of our findings, we provide examples of how our result can be instantiated in noisy regression and classification tasks (Section 5), and we empirically validate our theoretical results on synthetic and semi-synthetic data (Section 7).

2. Preliminaries

Notation. Throughout, we denote by $[n] := \{1, \ldots, n\}$. We use $(f \circ h)(x) = f(h(x))$ to denote the composition of functions f and h. We write $\lesssim$ and $\gtrsim$ for inequality up to a universal constant

factor. For matrices A and B, we use $A \precsim B$ to denote that there exists a universal constant $c > 0$ such that $A \preceq cB$. $\tilde{\mathcal{O}}$ hides logarithmic factors.

2.1. Problem formulation

We now present the lifelong representation learning problem studied in this work, which generalizes the settings in (e.g., Balcan et al., 2015; Cao et al., 2022). Consider a sequence of T tasks arriving over time. They share common input and output spaces, denoted by $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} \subset \mathbb{R}$. Each task $t \in [T]$ is associated with an underlying data-generating distribution $\mathcal{P}_t$ over $\mathcal{X} \times \mathcal{Y}$.[1] We assume that $\mathcal{P}_1, \ldots, \mathcal{P}_T$ share a common marginal distribution[2] over $\mathcal{X}$, denoted by P_X, as is standard in the literature for studying representation transfer (Tripuraneni et al., 2020; Xu and Tewari, 2021).

Let $\mathcal{H} \subseteq \{h : \mathcal{X} \to \mathcal{Z}\}$ be a class of representations and $\mathcal{F} \subseteq \{f : \mathcal{Z} \to \mathbb{R}\}$ be a class of task-specific prediction layers that operate on these representations. A predictor for a task is then given by $f \circ h$. As an example, $\mathcal{H}$ may represent a family of deep neural networks that map $\mathcal{X}$ to $\mathcal{Z} = \mathbb{R}^k$, while $\mathcal{F}$ is a class of linear functions in $\mathbb{R}^k$. Let $\mathcal{F} \circ \mathcal{H} := \{f \circ h : f \in \mathcal{F}, h \in \mathcal{H}\}$; we assume that $\mathcal{F} \circ \mathcal{H}$ is permissible (see (Pollard, 1984) and (Baxter, 2000, Appendix D therein)).

Let $\ell : \mathbb{R} \times \mathcal{Y} \to [0, 1]$ be a loss function. For any distribution P over $\mathcal{X} \times \mathcal{Y}$, denote by

$$\mathcal{L}_P(f \circ h) := \mathbb{E}_{(x,y)\sim P}\left[\ell\left((f \circ h)(x), y\right)\right]$$

the risk of $f \circ h$, for any $f \in \mathcal{F}$ and $h \in \mathcal{H}$.

In this paper, we consider a family of distributions parameterized by $g : \mathcal{X} \to \mathbb{R}$ of the form

$$P_g(x, y) = P_X(x)P_{Y|X}(y|x; g). \tag{1}$$

We assume that the loss ℓ is calibrated with respect to P_g, i.e., g is Bayes optimal with respect to P_g and ℓ—formally,

$$g = \operatorname*{argmin}_{g' : \mathcal{X} \to \mathbb{R}} \mathcal{L}_{P_g}(g').$$

The specific form of $P_{Y|X}(y|x; g)$ may depend on the exact learning problem at hand. For example, in classification with 0-1 loss under noiseless realizability, labels are given deterministically as $y = g(x)$. In regression where the response variable follows a distribution from an exponential family, such as logistic regression or Poisson regression, $P_{Y|X}(y \mid x; g) = \lambda(y)\exp(g(x)y - b(g(x)))$, where λ is a base measure and $b(\theta) = \ln \int_{\mathcal{Y}} e^{\theta y}\lambda(y)dy$ is the associated log partition function. We refer to $\mathbb{P} := \{P_{f\circ h} : f \in \mathcal{F}, h \in \mathcal{H}\}$ as the probabilistic model of the problem.

Throughout, we assume the tasks admit a *shared* representation under which they are well-specified:

Assumption 1 (Well-specified model with shared representation)
There exist $h^* \in \mathcal{H}$ and $f_1^*, \ldots, f_T^* \in \mathcal{F}$ such that, for each task $t \in [T]$, $\mathcal{P}_t = P_{f_t^* \circ h^*}$.

Lifelong learning. For each task $t \in [T]$, the learner can request i.i.d. samples from $\mathcal{P}_t$, possibly over more than one rounds. Using these samples, the learner produces a predictor $f_t \circ h_t$. Once completed, the learner proceeds to the next task and may never revisit a previous one. However, we assume that the learner may have access to a memory buffer of $o(T)$ size to store some data. Let

1. Following prior work, we assume that when a task is completed, it does not come back. For example, even when task 10 has the same distribution as task 1, we treat it as a new task and may produce a different model. Extending our setting to continual learning (Kirkpatrick et al., 2017) where old tasks can reappear is an interesting future direction.
2. With some care, our results can be extended to handle heterogeneous marginal distributions that are well-conditioned.

$\delta, \epsilon \in (0,1)$. The learner's goal is to return predictors $f_t \circ h_t$'s such that

$$\Pr\left(\forall t \in [T],\ \underbrace{\mathcal{L}_{\mathcal{P}_t}(f_t \circ h_t) - \mathcal{L}_{\mathcal{P}_t}(f_t^* \circ h^*)}_{\text{excess risk for task } t} \leq \epsilon\right) \geq 1 - \delta.$$

The performance of the learner is measured by its *sample complexity*, i.e., the total number of samples it acquires across all T tasks. We note that the lifelong learning objective here is akin to that of the KWIK (knows what it knows) framework studied in reinforcement learning and active learning (Li et al., 2008). Therein, for each task, the learner can always choose to acquire more data (equivalent to saying "I don't know") if they are uncertain about their current prediction. Since the task sequence may be chosen *adversarially* (so long as Assumption 1 is satisfied), the total number of samples naturally accounts for variation in task difficulty, and an important challenge in lifelong representation learning lies in carefully managing when (or in which tasks) to request more data.

Remark 2 (Comparison with prior work) *Under our formulation, the setting in (Balcan et al., 2015) corresponds to choosing $\mathcal{H}$ as a class of low-dimensional linear representations and $\mathcal{F}$ as task-specific linear layers, while Cao et al. (2022) consider a richer $\mathcal{H}$ where each h corresponds to a linear map followed by a nonlinear activation. In contrast, we allow general choices of $\mathcal{H}$ and $\mathcal{F}$.*

Pentina and Urner (2016) study a related model with a general ground class, where new tasks may be solved by combining a growing dictionary of previously learned hypotheses. In our framework, we assume a fixed representation class $\mathcal{H}$ and focus on composite predictors $f \circ h$'s.

We also note that Balcan et al. (2015); Pentina and Urner (2016); Cao et al. (2022) all assume noiseless realizability. In contrast, we consider a well-specified model in which $f_t^* \circ h^*$ may have non-zero risk due to noise (Assumption 1); we defer the fully agnostic setting to future work.

Lastly, while prior work focuses on specific learning problems (e.g., binary classification (Balcan et al., 2015; Pentina and Urner, 2016)), our formulation—via suitable choices of $\mathcal{X}$, $\mathcal{Y}$, and ℓ—offers a unified framework for lifelong representation learning across a range of learning problems.

Representation approximation error. Given a representation $h : \mathcal{X} \to \mathcal{Z}$, a class of prediction layers $\mathcal{F}$, and a distribution P_g in the form of (1); we denote by h's *representation approximation error* over P_g to be

$$\mathcal{E}(h, P_g) := \min_{f \in \mathcal{F}} \mathcal{L}_{P_g}(f \circ h) - \mathcal{L}_{P_g}(g).$$

In words, this is the best excess risk achieved by any prediction layer in $\mathcal{F}$ under the given representation h. It can be easily seen from the definition that when $g = f \circ h$ for some $f \in \mathcal{F}$, $\mathcal{E}(h, P_g) = 0$.

2.2. Background: empirical risk and multi-task ERM

Before presenting our algorithm and main results, we provide background on multi-task ERM which we use as a subroutine. Let P be the data-generating distribution over $\mathcal{X} \times \mathcal{Y}$ for a task, and $S = \{(x_j, y_j)\}_{j=1}^m$ be drawn i.i.d. from P^m. For any $h \in \mathcal{H}$ and $f \in \mathcal{F}$, define

$$\widehat{\mathcal{L}}_S(f \circ h) := \frac{1}{m} \sum_{j=1}^{m} \ell\left((f \circ h)(x_j), y_j\right)$$

to be the empirical risk of $f \circ h$ over S.

Multi-task ERM for representation transfer has been widely studied in the literature (e.g., Baxter, 2000; Maurer et al., 2016; Tripuraneni et al., 2020; Aliakbarpour et al., 2024). Let $P_1, \ldots, P_n$

Algorithm 1: Lifelong representation learning with multi-task ERM as a subroutine

Input: $\mathcal{H}$, $\mathcal{F}$, target error ϵ, confidence δ, number of tasks T, noise levels $(\kappa_t)_t$;

Initialize memory $\mathcal{M} \leftarrow \emptyset$;

Initialize $N \leftarrow 1$ (unless $\mathrm{TEdim}(\mathcal{H}, \mathcal{F}, \epsilon)$ is known (see Definition 8), in which case set N to it);

for *task* $t = 1$ **do**

  Draw a sample S_1 of size m_N from $\mathcal{P}_1^{m_N}$, apply ERM to learn $\hat{h}$ and $\hat{f}_1$ and output $\hat{f}_1 \circ \hat{h}$;

  Set $n \leftarrow 1$ and $t_n \leftarrow t$, and update the memory $\mathcal{M} \leftarrow \mathcal{M} \cup \{S_{t_n}\}$;

for *tasks* $t = 2, \ldots, T$ **do**

  // Few-shot property test: check if $\hat{h}$ admits a hypothesis for current task t with risk at most ϵ

  Draw a sample $\tilde{S}_t$ of size $\tilde{m}$ from $\mathcal{P}_t^{\tilde{m}}$, and apply ERM with current $\hat{h}$ to learn $\tilde{f}_t$;

  if $\widehat{\mathcal{L}}_{\tilde{S}_t}(\tilde{f}_t \circ \hat{h}) \leq \kappa_t + \frac{3}{4}\epsilon$ **then**

    Output $\tilde{f}_t \circ \hat{h}$;

  else

    if $n = N$ **then**

      Set $n \leftarrow 1$, $N \leftarrow 2N$, and clear the memory $\mathcal{M} \leftarrow \emptyset$;

    else

      Set $n \leftarrow n + 1$;

    Draw a sample S_t of size m_N from $\mathcal{P}_t^{m_N}$;

    Set $t_n \leftarrow t$, and update the memory $\mathcal{M} \leftarrow \mathcal{M} \cup \{S_{t_n}\}$;

    Apply ERM over the samples stored in the memory to learn

$$\hat{h}, \check{f}_1^{(t)}, \ldots, \check{f}_n^{(t)} \leftarrow \operatorname*{argmin}_{\substack{h \in \mathcal{H} \\ f_1, \ldots, f_n \in \mathcal{F}}} \frac{1}{n} \sum_{i=1}^{n} \widehat{\mathcal{L}}_{S_{t_i}}(f_i, h);$$

    Set $\hat{f}_t \leftarrow \check{f}_n^{(t)}$, update $\hat{h}$, and output $\hat{f}_t \circ \hat{h}$.

be the data-generating distribution of n tasks. Suppose for each task $i \in [n]$, we draw an i.i.d. sample of size m, $S_i \sim P_i^m$, then multi-task ERM jointly finds a shared representation and task-specific prediction layers by solving the following optimization problem:

$$\operatorname*{argmin}_{\substack{h \in \mathcal{H} \\ f_1, \ldots, f_n \in \mathcal{F}}} \frac{1}{n} \sum_{i=1}^{n} \widehat{\mathcal{L}}_{S_i}(f_i \circ h). \tag{2}$$

When $n = 1$, we refer to this approach as single-task ERM, or simply ERM.

3. Algorithm and complexity measure for lifelong representation learning

3.1. Algorithm

Our algorithm builds upon the methods of Balcan et al. (2015) and Cao et al. (2022), which are tailored to specific function classes under the noiseless realizability assumption. In contrast, our algorithm is amenable to general function classes and noise models. In addition, our algorithm is designed with practical considerations in mind, leveraging MTL as a subroutine, which has been extensively studied and aligns with real-world implementations. Algorithm 1 contains the pseudocode.

High-level overview. In our algorithm, the learner maintains a representation $\hat{h} \in \mathcal{H}$, which is initially learned from the first task. For each subsequent task, the learner performs a property test to determine if there exists a prediction layer $f \in \mathcal{F}$ such that the predictor $f \circ \hat{h}$ has excess risk at most ϵ, i.e., if $\hat{h}$'s representation approximation error is at most ϵ. Specifically, given the restricted class of functions, $\{f \circ \hat{h} : f \in \mathcal{F}\} \subset \mathcal{F} \circ \mathcal{H}$, the learner checks whether the best predictor in this restricted class has risk ϵ-close to that of the Bayes optimal predictor. This is done via few-shot learning the prediction layer with $\hat{h}$ frozen (see also Figure 1(*a*)), and requires a small number of samples depending only on the complexity of $\mathcal{F}$. We call this the *few-shot property test*.

If the test succeeds, the learner simply retains $\hat{h}$ and safely outputs the predictor (lines 8 to 9). Otherwise, the learner updates its maintained representation $\hat{h}$ by performing multi-task ERM on a subset of tasks where few-shot property test previously failed—whose samples are stored in memory (lines 15 to 18, see also Figure 1(*b*)). Applying multi-task ERM ensures that the new $\hat{h}$ is an effective representation for this subset of tasks. We note that only samples from tasks where the property test failed are added to the memory buffer. To facilitate the test, we make an additional assumption:

Assumption 3 (Known noise levels) *Let $\kappa_t := \mathcal{L}_{\mathcal{P}_t}(f_t^* \circ h^*)$ denote the Bayes-optimal risk for each task $t \in [T]$. The learner knows the values $(\kappa_t)_{t=1}^T$.*

Remark 4 *While not explicitly stated, the noiseless realizability assumption in (Balcan et al., 2015; Cao et al., 2022) entails that the learner knows $\kappa_t = 0$ for all $t \in [T]$. We introduce Assumption 3 because the excess risk measured within $\{f \circ \hat{h} : f \in \mathcal{F}\} \subset \mathcal{F} \circ \mathcal{H}$ can be misleading in estimating the excess risk with respect to the Bayes-optimal predictor, if only a limited number of examples are sampled for task t, as the risk of the best predictor in this class may itself be far from $\mathcal{L}_{\mathcal{P}_t}(f_t^* \circ h^*)$. Proposition 5 below highlights the hardness in a simple linear setting; its formal statement and proof (based on a reduction from (Kong and Valiant, 2018, Proposition 2)) are deferred to Appendix A.1. We conjecture that it may be impossible to design an algorithm whose sample complexity adapts to the quality of representation of $\hat{h}$ without the knowledge of κ_t.*

That said, Assumption 3 can be relaxed to the knowledge of a sequence of upper bounds $\{\nu_t\}$ on the Bayes optimal risks. In this case, the objective for each task t would then be to learn a hypothesis with risk at most $\nu_t + \epsilon$. Algorithm 1 can be extended to work in this setting with an an analogous sample complexity guarantee to our main result (Theorem 14 below).

Proposition 5 (informal) *Suppose we observe n examples $\{(x_i, y_i)\}_{i=1}^n \sim P^n$, where P denotes some noisy linear regression model in $\mathbb{R}^d$. Let $\mathcal{G}$ be a class of linear predictors in $\mathbb{R}^d$ and $\mathcal{G}_0 \subset \mathcal{G}$ be restricted to a fixed subspace of dimension $r \leq \frac{d}{2}$; that is, $g \in \mathcal{G}_0$ uses a given linear representation. For any P and $\mathcal{G}' \subseteq \mathcal{G}$, let $\kappa_P(\mathcal{G}') := \inf_{g \in \mathcal{G}'} \mathbb{E}_{(x,y)\sim P}\left[(g(x) - y)^2\right]$. Consider two hypotheses:*

$$H_0 = \{P : \kappa_P(\mathcal{G}_0) = \kappa_P(\mathcal{G})\} \quad \textit{and} \quad H_1 = \{P : \kappa_P(\mathcal{G}_0) > \kappa_P(\mathcal{G}) + 0.9\}.$$

There exists some constant c such that no test can successfully distinguish between H_0 and H_1 with probability $\frac{2}{3}$ using fewer than $c\sqrt{d}$ examples.

Remark 6 *The use of a memory buffer in lifelong learning is not specific to our algorithm (e.g. Isele and Cosgun, 2018). Cao et al. (2022) utilize past features stored in a memory buffer to perform representation refinement in their main algorithm, `LLL-RR`, and they also suggest a heuristic algorithm, `H-LLL`, that stores and reuses training datasets.*

It remains to understand when property tests in Algorithm 1 stop failing—that is, when multi-task ERM is no longer needed and how large the memory buffer must be. We now introduce a complexity measure that characterizes the sample and memory requirements of our algorithm, before returning to examine our algorithm in greater detail.

3.2. The task-eluder dimension

The eluder dimension is a well-established complexity measure studied in sequential decision-making (Russo and Van Roy, 2013; Foster et al., 2020; Li et al., 2022a; Hanneke, 2024). Its essence is captured by the canonical illustrative example (Russo and Van Roy, 2013): how many times can a politician "elude" by answering questions without revealing their true position? Herein, we adapt and extend the eluder dimension to characterize the complexity of lifelong representation learning.

Definition 7 (ϵ-independence) *Let $\mathcal{H} \subset \{h : \mathcal{X} \to \mathcal{Z}\}$ be a class of representations, and $\mathcal{F} \subset \{f : \mathcal{Z} \to \mathbb{R}\}$ be a class of prediction layers that operate on these representations. Let $\mathbb{P} = \{P_{f \circ h} : f \in \mathcal{F}, h \in \mathcal{H}\}$ be the probabilistic model and $\ell : \mathbb{R} \times \mathcal{Y} \to [0, 1]$ be a loss function.*

*For any representation $h \in \mathcal{H}$, we say f_n is ϵ-*independent *of $\{f_1, \ldots, f_{n-1}\}$ with respect to $(\mathcal{H}, \mathcal{F})$ and h, if there exists $h' \in \mathcal{H}$ such that*

$$\sum_{i=1}^{n-1} \mathcal{E}(h', P_{f_i \circ h}) \leq \epsilon, \tag{3}$$

but $\mathcal{E}(h', P_{f_n \circ h}) > \frac{\epsilon}{2}$.

Definition 8 (Task-eluder dimension) *Given the setting of Definition 7, for any representation $h \in \mathcal{H}$, we denote by $\mathrm{TEdim}_h(\mathcal{H}, \mathcal{F}, \epsilon)$ the length of the longest sequence $\{f_i\}_i$, such that each f_i is ϵ-independent of its predecessors with respect to $(\mathcal{H}, \mathcal{F})$ and h.*

Then, the ϵ-task-eluder dimension of $(\mathcal{H}, \mathcal{F})$ under probabilistic model $\mathbb{P}$ and loss function ℓ is $\mathrm{TEdim}_{\mathbb{P},\ell}(\mathcal{H}, \mathcal{F}, \epsilon) := \sup_{h \in \mathcal{H}} \mathrm{TEdim}_h(\mathcal{H}, \mathcal{F}, \epsilon)$. To avoid overloading the notation, we omit $\mathbb{P}$ and ℓ and write $\mathrm{TEdim}(\mathcal{H}, \mathcal{F}, \epsilon)$ when they are clear from context.

Intuitively, fix any representation h; a task f_n is independent of its predecessors if there is another representation $h' \in \mathcal{H}$ that is indistinguishable from h on earlier tasks—because good prediction layers also exist under h'—yet task f_n provides new information that helps disambiguate between h and h', as there is no prediction layer for h' that can match (h, f_n). Here, "equivalence" between predictors is measured in terms of excess risk at scale $\mathcal{O}(\epsilon)$ under the distributions induced by $(f_i \circ h)$'s.[3]

In other words, the task-eluder dimension quantifies how long the learner can continue picking up *at least some new information* from a possibly adversarial task sequence without pinning down a near-optimal representation. As a sanity check, we prove the following simple bound on the task eluder dimension when either $\mathcal{H}$ or $\mathcal{F}$ is finite, similar to the basic bounds of eluder dimension in terms of the size and domain size of the function class (Osband and Van Roy, 2014, Proposition 1):

Proposition 9 *For any $\epsilon \geq 0$,*

$$\mathrm{TEdim}(\mathcal{H}, \mathcal{F}, \epsilon) \leq 2 \min\left(|\mathcal{H}|, |\mathcal{F}|\right).$$

3. In Section 8, we examine an alternative notion of ϵ-independence (Definition 7) and show, via a linear example, that it can fail to yield meaningful bounds.

The proof of Proposition 9 is deferred to Appendix A.2. In Section 5, we provide examples of upper bounds on the task-eluder dimension for common classes.

Remark 10 *Our definition of the task eluder dimension parallels the eluder dimension defined in (Li et al., 2022a), where a central representation h is chosen when defining independence, analogous to the central predictor f^* in (Li et al., 2022a); the prediction layers f_i are analogous to the examples x_i, and the excess losses $\mathcal{E}(h', P_{f_i \circ h})$ are analogous to square errors $(f'(x_i) - f^*(x_i))^2$.*

However, this analogy is not perfect due to our lack of exact knowledge of f_i in lifelong representation learning. In fact, if we were to draw exact parallel, we may define ϵ-independence to be that: there exists $h' \in \mathcal{H}$ such that

$$\sum_{i=1}^{n-1} \mathcal{L}_{\mathcal{P}_{f_i \circ h}}(f_i \circ h') - \mathcal{L}_{\mathcal{P}_{f_i \circ h}}(f_i \circ h) \leq \epsilon, \quad \textit{but } \mathcal{L}_{\mathcal{P}_{f_n \circ h}}(f_n \circ h') - \mathcal{L}_{\mathcal{P}_{f_n \circ h}}(f_n \circ h) > \frac{\epsilon}{2}.$$

Although we are able to show bounds on the maximum length of independent sequences under this notion of independence in examples in Section 5, we were unable to bound Algorithm 1's sample complexity using this quantity.

In appendix A.4, we provide an alternative definition of task eluder dimension without a central representation, similar to the original eluder dimension definition in (Russo and Van Roy, 2013). The "centered" version of the task eluder dimension is never larger than its uncentered counterparts, and is sufficient for our applications.

Remark 11 *For noiseless, realizable binary classification, Balcan et al. (2015) and Pentina and Urner (2016) introduce notions of γ-effective dimension: the length of the longest sequence of tasks that cannot be handled by predictors built from earlier tasks, either via linear combinations (Balcan et al., 2015) or weighted majority votes (Pentina and Urner, 2016). These notions play a similar role to the task-eluder dimension, which is defined in terms of representation approximation error in the composite model and extends naturally to general losses and well-specified noise.*

3.3. A closer look at Algorithm 1

Now that we are equipped with the notion of task-eluder dimension, we revisit Algorithm 1 to analyze its sample and space complexities. For now, assume that the learner knows the task-eluder dimension $\mathrm{TEdim}(\mathcal{H}, \mathcal{F}, \epsilon) \leq \Xi$ in advance and initializes $N = \Xi$ (line 2); let us also disregard the steps in red (lines 11 to 12).[4] At the beginning of any task $t \geq 2$, suppose the counter $n = n_0$.

1. By choosing m_N appropriately (as stated in Theorem 14) and leveraging data in the memory buffer for which the initial task and tasks where property test failed, $\mathcal{M} = \left\{S_{t_1}, \ldots, S_{t_{n_0}}\right\}$, the learner ensures from multi-task ERM (line 17) that the currently maintained $\hat{h}$ satisfies, with high probability,
$$\sum_{i=1}^{n_0} \left(\mathcal{L}_{\mathcal{P}_{t_i}}(\check{f}_i^{(t_{n_0})} \circ \hat{h}) - \mathcal{L}_{\mathcal{P}_{t_i}}(f_{t_i}^* \circ h^*) \right) \leq \epsilon.$$
2. Meanwhile, with $\tilde{m}$ chosen appropriately (as stated in Theorem 14), if the few-shot property test fails, then with high probability,
$$\mathcal{E}(\hat{h}, \mathcal{P}_t) = \min_{f \in \mathcal{F}} \mathcal{L}_{\mathcal{P}_t}(f \circ \hat{h}) - \mathcal{L}_{\mathcal{P}_t}(f_t^* \circ h^*) > \frac{\epsilon}{2},$$

4. See Appendix A.5 for a clean version of the algorithm under knowledge of $\mathrm{TEdim}(\mathcal{H}, \mathcal{F}, \epsilon)$.

in other words, f_t^* is ϵ-independent of $\{f_{t_i}^*\}_{i=1}^{n_0}$ with respect to $(\mathcal{H}, \mathcal{F})$ and h^*.

Since $\mathrm{TEdim}_{h^*}(\mathcal{H}, \mathcal{F}, \epsilon) \leq \mathrm{TEdim}(\mathcal{H}, \mathcal{F}, \epsilon)$, the task-eluder dimension bounds the number of times this can happen. In other words, the task-eluder dimension characterizes the number of tasks for which the learner needs to acquire *additional* data beyond what is required for property testing, which then determines the sample and space complexities of the algorithm.

Doubling trick. In practice, the learner may not know $\mathrm{TEdim}(\mathcal{H}, \mathcal{F}, \epsilon)$ in advance. To address this, we utilize the doubling trick from the online learning literature (e.g., Shalev-Shwartz et al., 2012). Specifically, we let N be a running estimate of Ξ. We begin with a small estimate, $N = 1$ (line 2). Each time it proves insufficient—i.e., for the current estimate N, property tests fail more than N times—we double the estimate, clear the memory, and restart the process (lines 11 to 12). This ensures that, without prior knowledge of $\mathrm{TEdim}(\mathcal{H}, \mathcal{F}, \epsilon)$, the number of times that the few-shot property test fails can still be bounded by $\mathcal{O}\left(\mathrm{TEdim}(\mathcal{H}, \mathcal{F}, \epsilon)\right)$.

4. Theoretical guarantees

In this section, we first provide additional background and introduce a few technical tools for our theoretical analysis (Section 4.1). We then introduce two benchmark algorithms (Section 4.2) before presenting our main result (Section 4.3).

4.1. Background: sample complexity of multi-task ERM

Capacities of $\mathcal{H}$ and $\mathcal{F}$ based on covering numbers. As is standard in statistical learning theory, sample complexity for generalization depends on the capacity of the learner's model class, often captured by notions such as VC dimension and Rademacher complexity (e.g., Shalev-Shwartz and Ben-David, 2014). Following (Baxter, 2000), we use *covering numbers* to characterize the capacities of $\mathcal{H}$ and $\mathcal{F}$ and later analyze sample complexity. We note that our findings in this work are not tied to this specific choice and can extend to other suitable complexity measures.

For any $f \in \mathcal{F}$, let $f_\ell(z, y) := \ell(f(z), y)$ for any low-dimensional representation $z \in \mathcal{Z}$ and target $y \in \mathcal{Y}$, and let $\mathcal{F}_\ell := \{f_\ell : f \in \mathcal{F}\}$. For any distribution Q on $\mathcal{Z} \times \mathcal{Y}$, let $d_Q(f_\ell, f'_\ell) := \int_{\mathcal{Z}\times\mathcal{Y}} |f_\ell(z, y) - f'_\ell(z, y)| \, dQ(z, y)$ be the $L^1(Q)$ pseudo-metric on $\mathcal{F}_\ell$. Then, for any $\epsilon_0 > 0$, we define the capacity of $\mathcal{F}$ at scale ϵ_0 to be

$$\mathcal{C}(\mathcal{F}_\ell, \epsilon_0) := \sup_Q N(\mathcal{F}_\ell, \epsilon_0, d_Q),$$

where $N(\mathcal{F}_\ell, \epsilon_0, d_Q)$ denotes the ϵ_0-covering number of $(\mathcal{F}_\ell, d_Q)$. For ease of notation, we often use $\mathcal{C}(\mathcal{F}, \epsilon_0)$ in place of $\mathcal{C}(\mathcal{F}_\ell, \epsilon_0)$.

For measure P on $\mathcal{X}\times\mathcal{Y}$, let $d_{P,\mathcal{F}_\ell}(h, h') := \int_{\mathcal{X}\times\mathcal{Y}} \sup_{f_\ell \in \mathcal{F}_\ell} |f_\ell(h(x), y) - f_\ell(h'(x), y)| \, dP(x, y)$ be a pseudo-metric on $\mathcal{H}$. Then, for any $\epsilon_0 > 0$, we define the capacity of $\mathcal{H}$ at scale ϵ_0 to be

$$\mathcal{C}_{\mathcal{F}_\ell}(\mathcal{H}, \epsilon_0) := \sup_P N(\mathcal{H}, \epsilon_0, d_{P,\mathcal{F}_\ell}),$$

where $N(\mathcal{H}, \epsilon_0, d_{P,\mathcal{F}_\ell})$ denotes the ϵ_0-covering number of $(\mathcal{H}, d_{P,\mathcal{F}_\ell})$. When the context is clear, we write $\mathcal{C}(\mathcal{H}, \epsilon_0)$ for $\mathcal{C}_{\mathcal{F}_\ell}(\mathcal{H}, \epsilon_0)$ to avoid clutter.

We note that $\log \mathcal{C}(\mathcal{F}, \epsilon_0)$ and $\log \mathcal{C}(\mathcal{H}, \epsilon_0)$ are often referred to as the metric entropy of $\mathcal{F}$ and $\mathcal{H}$, respectively (Haussler, 1992). Section 5 provides examples of capacity bounds for $\mathcal{H}$ and $\mathcal{F}$.

Sample complexity of multi-task ERM. Baxter (2000) established the following uniform bound for multi-task ERM using capacities based on covering numbers:

Theorem 12 (Baxter, 2000, Theorem 4 and Theorem 6 thereof) *Let $P_1, \ldots, P_n$ be the data-generating distributions of n tasks. Let $\mathcal{H}_0$ be a class of representations and $\mathcal{F}_0$ a class of prediction layers. Suppose for each task i, an i.i.d. sample S_i of size m is drawn from P_i^m, where*

$$m \geq \mathcal{O}\left(\frac{1}{n\epsilon_0^2}\left(\log \mathcal{C}\big(\mathcal{H}_0, \frac{\epsilon_0}{32}\big) + n \log \mathcal{C}\big(\mathcal{F}_0, \frac{\epsilon_0}{32}\big) + \log \frac{1}{\delta_0}\right)\right),$$

then with probability at least $1 - \delta_0$, for any $(h, f_1, \ldots, f_n)$,

$$\left|\frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_{P_i}(f_i \circ h) - \frac{1}{n}\sum_{i=1}^{n} \widehat{\mathcal{L}}_{S_i}(f_i \circ h)\right| \leq \epsilon_0.$$

Corollary 13 *It follows immediately that if the number of samples per task exceeds*

$$\mathcal{O}\left(\frac{1}{n\epsilon_0^2}\left(\log \mathcal{C}\big(\mathcal{H}_0, \frac{\epsilon_0}{64}\big) + n \log \mathcal{C}\big(\mathcal{F}_0, \frac{\epsilon_0}{64}\big) + \log \frac{1}{\delta_0}\right)\right),$$

then with probability at least $1 - \delta_0$,

$$\frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_{P_i}(\hat{f}_i \circ \hat{h}) \leq \min_{h, f_1, \ldots, f_n} \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_{P_i}(f_i \circ h) + \epsilon_0,$$

where $(\hat{h}, \hat{f}_1, \ldots, \hat{f}_n)$ is the solution to multi-task ERM (Eq. (2)).

This result highlights the benefit of MTL. With only one task, the learner bears the full burden of learning both the representation and the prediction layer. Since we are mostly interested in the regime where $\log \mathcal{C}\big(\mathcal{H}_0, \frac{\epsilon_0}{64}\big) \gg \log \mathcal{C}\big(\mathcal{F}_0, \frac{\epsilon_0}{64}\big)$, learning the representation individually for each task can be costly. With MTL, while the learner still has to learn $f_1, \ldots, f_n$, the cost of learning the shared representation is amortized over the tasks.

4.2. Warm-up: two benchmarks

To better highlight the performance of our algorithm, we introduce two benchmark methods for the lifelong representation learning problem. The first is a naive *baseline* algorithm in which the learner simply ignores any shared structure and solves each task independently using single-task ERM, which we refer to as `independent-ERM`. Applying Corollary 13 (with $n = 1$) and the union bound, we obtain the following sample complexity bound of `independent-ERM`,

$$\tilde{\mathcal{O}}\left(\frac{T}{\epsilon^2}\left(\log \mathcal{C}\big(\mathcal{H}, \frac{\epsilon}{64}\big) + \log \mathcal{C}\big(\mathcal{F}, \frac{\epsilon}{64}\big)\right)\right). \tag{4}$$

Alternatively, had the learner known h^* beforehand, it suffices to only learn the prediction layer for each task by solving single-task ERM with a singleton representation class $\{h^*\}$ (using a much smaller sample). We call this *skyline* algorithm `oracle`. Formally, it follows from Corollary 13 and the union bound that `oracle` has a much lower sample complexity of

$$\tilde{\mathcal{O}}\left(\frac{T}{\epsilon^2} \log \mathcal{C}\big(\mathcal{F}, \frac{\epsilon}{64}\big)\right). \tag{5}$$

4.3. Main result

We now present our main theorem. In the interest of space, its proof is deferred to Appendix B.1.

Theorem 14 *Let $\Xi = \mathrm{TEdim}(\mathcal{H}, \mathcal{F}, \epsilon) < \infty$. Suppose $\mathcal{H}$ and $\mathcal{F}$ have finite capacities; that is, $\mathcal{C}\left(\mathcal{H}, \frac{\epsilon}{256\Xi}\right) < \infty$ and $\mathcal{C}\left(\mathcal{F}, \frac{\epsilon}{256\Xi}\right) < \infty$. In Algorithm 1, set*

$$m_N = \tilde{\mathcal{O}}\left(\frac{N}{\epsilon^2}\left[\log \mathcal{C}\left(\mathcal{H}, \frac{\epsilon}{64N}\right) + N \log \mathcal{C}\left(\mathcal{F}, \frac{\epsilon}{64N}\right)\right]\right)$$

for each N, and set

$$\tilde{m} = \tilde{\mathcal{O}}\left(\frac{1}{\epsilon^2} \log \mathcal{C}\left(\mathcal{F}, \frac{\epsilon}{128}\right)\right),$$

where $\tilde{\mathcal{O}}$ hides logarithmic factors in T and $\frac{1}{\delta}$. Then, with probability at least $1-\delta$,

- *For every task, Algorithm 1 outputs a predictor with excess risk at most ϵ;*
- *Algorithm 1 performs multi-task ERM at most $\mathcal{O}(\Xi)$ times;*
- *The sample complexity is bounded by*

$$\tilde{\mathcal{O}}\Bigg(\underbrace{\frac{T}{\epsilon^2} \log \mathcal{C}\left(\mathcal{F}, \frac{\epsilon}{128}\right)}_{\text{cost of few-shot tests}} + \underbrace{\frac{\Xi^2}{\epsilon^2}\left[\log \mathcal{C}\left(\mathcal{H}, \frac{\epsilon}{128\Xi}\right) + \Xi \log \mathcal{C}\left(\mathcal{F}, \frac{\epsilon}{128\Xi}\right)\right]}_{\text{overhead of lifelong representation learning}}\Bigg). \tag{6}$$

In addition, the size of the memory buffer it requires is at most

$$\tilde{\mathcal{O}}\left(\frac{\Xi^2}{\epsilon^2}\left[\log \mathcal{C}\left(\mathcal{H}, \frac{\epsilon}{128\Xi}\right) + \Xi \log \mathcal{C}\left(\mathcal{F}, \frac{\epsilon}{128\Xi}\right)\right]\right).$$

See Theorem 24 in Appendix B.1 for a restatement of Theorem 14 with all constants specified. To interpret the sample complexity bound in Eq. (6), observe that the first term accounts for the sample complexity from property testing. It is necessary even with a priori knowledge of h^* albeit a constant factor in the scale of the capacity (cf. the guarantee of `oracle` in Eq. (5)). The second term reflects the overhead of learning the unknown representation. It is governed by the capacities of $\mathcal{H}$ and $\mathcal{F}$, as well as the task-eluder dimension $\Xi = \mathrm{TEdim}(\mathcal{H}, \mathcal{F}, \epsilon)$. Since $\mathrm{TEdim}(\mathcal{H}, \mathcal{F}, \epsilon)$ measures the complexity of the model class under $\mathbb{P}$ and ℓ, it does not grow with T. This highlights the benefit of representation transfer when compared to the baseline in Eq. (4).

In particular, as T approaches infinity in a truly *lifelong* setting, the cost of representation learning becomes negligible, and the sample complexity nearly matches that of `oracle`. Similarly, the space complexity (size of the memory buffer) also remains bounded as T grows.

5. Examples

We now provide two concrete examples of how our results can be specialized in regression and classification tasks. Let $\mathcal{X} = \left\{x \in \mathbb{R}^d : \|x\| \le \sqrt{d}\right\}$. We focus on a class of low-dimensional linear representations, $\mathcal{H} = \left\{x \mapsto B^\top x : B \in \mathbb{R}^{d\times k}, B^\top B = I_k\right\}$, where $k \ll d$. This class has been widely studied to demonstrate the benefit of representation transfer (e.g., Balcan et al., 2015; Hu et al., 2021; Tripuraneni et al., 2021; Du et al., 2021). In the following, we sometimes abuse notation and identify a function (either representation or prediction layer) with its parameter. Due to space constraints, our proofs are deferred to Appendix C; see also Appendix C.1 for a high level description of the proof idea.

Linear regression with noise. Let $\mathcal{F}^{\text{lin}} := \left\{ z \mapsto w^\top z : w \in \mathbb{R}^k, \|w\| \leq \frac{1}{2\sqrt{d}} \right\}$ be a class of linear functions, and $\mathcal{Y} = [-1, 1]$. Consider the following probabilistic model $\mathbb{P}$: P_X over $\mathcal{X}$ satisfies $I \precsim \mathbb{E}_{x \sim P_X}\left[xx^\top\right] \precsim I$. For any h and f, given an input x, $y = (f \circ h)(x) + \eta$, where η is independently drawn from a shared noise distribution with support $[-\frac{1}{2}, \frac{1}{2}]$, mean zero, and variance κ. The T tasks are well-specified with ground truth representation B^* and prediction layers $w_1^*, \ldots, w_T^*$. Let $\ell(y', y) := \frac{1}{4}(y' - y)^2$ which has range $[0, 1]$.

Proposition 15 *Let $\epsilon \in (0, 1)$. We have*

$$\log \mathcal{C}(\mathcal{F}_\ell^{\text{lin}}, \epsilon) \leq \mathcal{O}\Big(k \log \frac{1}{\epsilon}\Big),\ \log \mathcal{C}_{\mathcal{F}_\ell^{\text{lin}}}(\mathcal{H}, \epsilon) \leq \mathcal{O}\Big(dk \log \frac{1}{\epsilon}\Big),\ \text{TEdim}_{\mathbb{P},\ell}(\mathcal{H}, \mathcal{F}^{\text{lin}}, \epsilon) \leq \mathcal{O}\big(k \log \frac{1}{\epsilon}\big).$$

Classification with logistic regression. Let $\mathcal{Y} = \{0, 1\}$. Denote by $\sigma(v) = \frac{1}{1+e^{-v}}$ the logistic sigmoid function and let $\mathcal{F}_{\log} := \left\{ z \mapsto \sigma(w^\top z) : w \in \mathbb{R}^k, \|w\| \leq \frac{1}{4\sqrt{d}} \right\}$. Data-generating distributions in $\mathbb{P}$ are defined as follows: P_X satisfies $I \precsim \mathbb{E}[xx^T] \precsim I$, and for each f and h, $P(y = 1 \mid x; f \circ h) = (f \circ h)(x)$. The T tasks are well specified by B^* and $w_1^*, \ldots, w_T^*$. Let $\ell(y', y) = -y \log y' - (1 - y) \log(1 - y')$.

Proposition 16 *Let $\epsilon \in (0, 1)$. We have*

$$\log \mathcal{C}(\mathcal{F}_\ell^{\log}, \epsilon) \leq \mathcal{O}\Big(k \log \frac{1}{\epsilon}\Big),\ \log \mathcal{C}_{\mathcal{F}_\ell^{\log}}(\mathcal{H}, \epsilon) \leq \mathcal{O}\Big(dk \log \frac{1}{\epsilon}\Big),\ \text{TEdim}_{\mathbb{P},\ell}(\mathcal{H}, \mathcal{F}^{\log}, \epsilon) \leq \mathcal{O}\big(k \log \frac{1}{\epsilon}\big).$$

In these examples, it follows straightforwardly from Theorem 14 that the sample complexity of Algorithm 1 is $\tilde{\mathcal{O}}\left(\left(kT + dk^3\right)/\epsilon^2\right)$. When T is large, this bound is dominated by $\tilde{\mathcal{O}}\left(kT/\epsilon^2\right)$ which only depends on $k \ll d$. This highlights the benefit of learning and leveraging the shared structure. We note that our goal here is to showcase the applicability of Theorem 14, rather than deriving the sharpest bound compared to what could be achieved with specialized techniques. In Appendix C, we also discuss how our analysis can apply to classification with random classification noise and the 0-1 loss, via a similar argument to Theorem 14 using multi-task ERM guarantees based on VC dimension.

6. Related work

Multi-task learning (Caruana, 1997) has been extensively studied in the literature; see (Zhang and Yang, 2021) for a survey. Lifelong learning can be traced back to (Thrun and Mitchell, 1995). Since then, it has seen applications in a wide range of domains, such as robotics (Lowrey et al., 2018), computer vision (Rebuffi et al., 2017), and natural language processing (de Masson D'Autume et al., 2019). See (Sodhani et al., 2022) for an introduction to various approaches to lifelong learning. Techniques from MTL and LTL have also been extended to lifelong learning (Finn et al., 2019). Much of the literature focuses on mitigating catastrophic forgetting (McCloskey and Cohen, 1989), whereas our main objective is to provide a theoretical study on the benefit of representation transfer.

In particular, our work builds upon a line of research that studies sample complexity guarantees of MTL and LTL where tasks share a common representation. Baxter (2000) study a general framework for inductive bias learning and establish guarantees based on covering numbers of the hypothesis space family. Maurer et al. (2016) use Gaussian complexities with a chain rule (Maurer, 2016) to derive data-dependent bounds. Tripuraneni et al. (2020) introduce a notion of task diversity which characterizes when transfer learning of representations can be achieved in LTL. Xu and Tewari (2021) study the setting where source and target tasks may use different classes of prediction

layers. Watkins et al. (2023) establish optimistic rates that adapt to the difficulty of a target task. These papers all consider the composite model $f \circ h$ that we study in this work.

MTL and LTL with shared linear representations have been studied (Tripuraneni et al., 2021; Du et al., 2021; Chen et al., 2022; Aliakbarpour et al., 2024). PAC-Bayesian bounds for LTL, with instantiations for parameter transfer and linear representation transfer, are developed by Pentina and Lampert (2014). Pentina and Ben-David (2015) study MTL and LTL of kernels, and Pentina and Ben-David (2018) approach MTL using probabilistic Lipschitzness as a quality measure of kernels.Sequential and parallel representation transfer have also been explored for linear bandits and RL (Hu et al., 2021; Yang et al., 2021; Qin et al., 2022; Duong et al., 2024).

Our work is most directly related to (Balcan et al., 2015; Pentina and Urner, 2016; Cao et al., 2022). These papers focus on lifelong learning for binary classification in the noiseless, realizable setting. Balcan et al. (2015) study low-dimensional linear representations: each task's predictor is a linear combination of k shared basis vectors. Pentina and Urner (2016) use weighted majority votes over a growing set of previously learned base hypotheses from a ground class. Cao et al. (2022) consider representations in the form of one hidden layer feature maps (a linear map followed by a nonlinear activation). We note that the guarantee for nonlinear representations in (Cao et al., 2022) relies on an assumption (Assumption 1 therein) that for two maps u and v, $d(u, v) \lesssim \Pr_{P_X}(\mathrm{sign}(u(x)) \neq \mathrm{sign}(v(x))) \lesssim d(u, v)$, where $d(\cdot, \cdot)$ denotes the angle. While this is true for linear maps when P_X is isotropic, log-concave, the first inequality may not hold in general for one-hidden-layer neural networks under the same assumption[5]. Algorithmically, these papers can be viewed as dynamically expanding their representation, with (Cao et al., 2022) additionally performing refinement. In contrast, we work with composite predictors $f \circ h$ that predict through a fixed $\mathcal{H}$, and we use multi-task ERM to relearn our representation when it becomes insufficient.

Li et al. (2022b) propose an architecture-based algorithm for continual representation learning. They provide sample complexity bounds under an assumption of *sequential task diversity*: earlier tasks are diverse enough to ensure small representational mismatch for new tasks. Similar assumptions have been made in (Tripuraneni et al., 2020) for LTL and (Qin et al., 2022) for sequential transfer in linear bandits. In contrast, we consider an online setting where the task sequence may be chosen by an adversary, and the learner needs to carefully manage when to request more data.

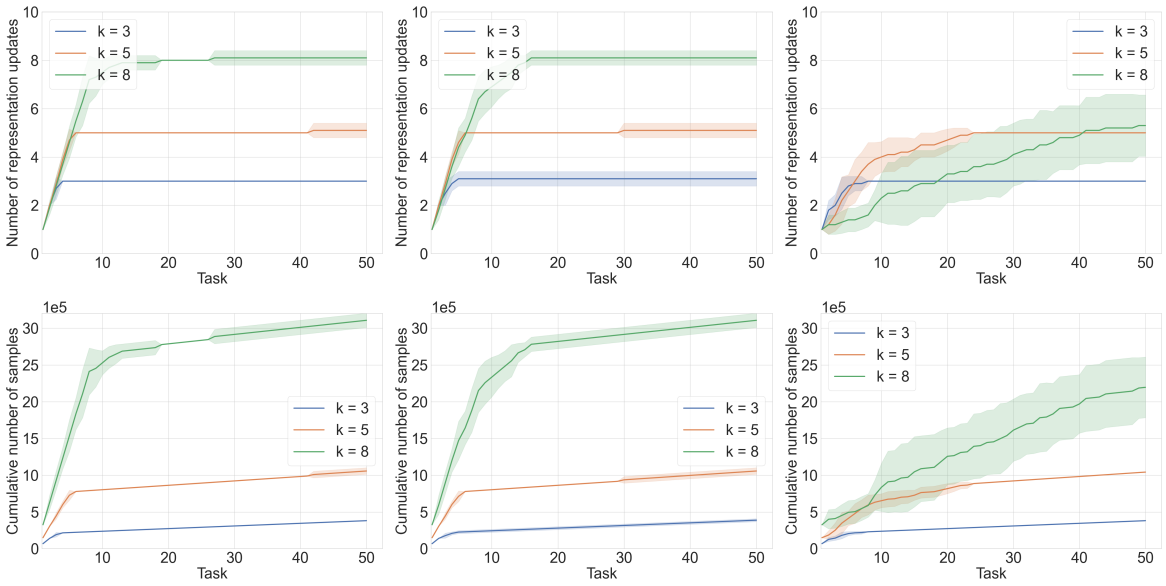
Alquier et al. (2017) study lifelong representation learning in an *online-within-online* setting: tasks arrive sequentially, and within each task, data points are also revealed sequentially with the learner predicting each instance. They also consider a *batch-within-online* setting, where data from each task is made available all at once. Alquier et al. (2017) study and establish compound regret bounds of the learner. In contrast, we allow the learner to request samples over multiple rounds, and require that, with high probability, it outputs for every task a predictor with excess risk at most ϵ.

7. Empirical validation

In this section, we empirically validate our theoretical results using synthetic and semi-synthetic data. In particular, we focus on the following questions:

1. Is the number of updates to the internal representation maintained by Algorithm 1 indeed bounded by the task-eluder dimension, $\mathrm{TEdim}(\mathcal{H}, \mathcal{F}, \epsilon)$?

5. In the nonlinear case, one may view u and v as $u(x)_{x \in \mathcal{X}}$ and $v(x)_{x \in \mathcal{X}}$ and calculate the angle in the $\mathcal{X}$-dimensional vector space.



(a) Low noise ($\beta = 8$) (b) Medium noise ($\beta = 4$) (c) High noise ($\beta = 1$)

Figure 2: Results under different noise levels. For each noise level, the top plot shows how the cumulative number of representation updates evolves over the tasks, and the bottom plot shows the average cumulative number of samples used over 50 tasks for each value of k. Shaded regions denote one standard deviation.

Table 1: Number of representation updates performed by Algorithm 1 at each noise level.

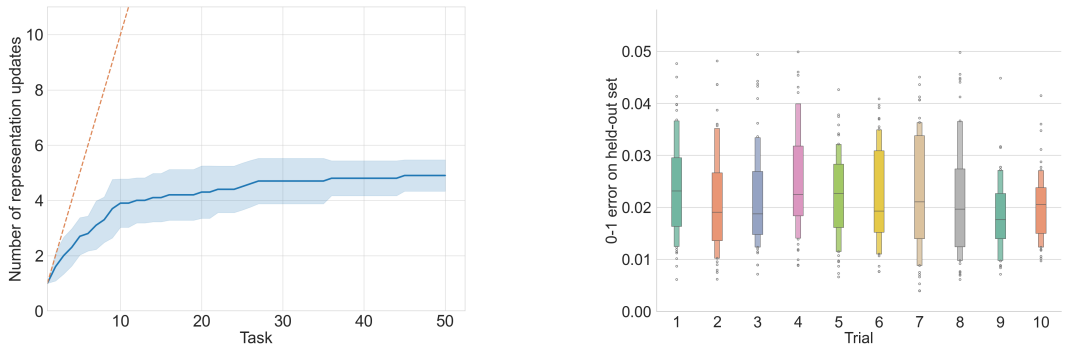
	Low noise ($\beta = 8$)	Medium noise ($\beta = 4$)	High noise ($\beta = 1$)
$k = 3$	3 ± 0	3.1 ± 0.3	3 ± 0
$k = 5$	5.1 ± 0.3	5.1 ± 0.3	5 ± 0
$k = 8$	8.1 ± 0.3	8.1 ± 0.3	5.3 ± 1.27

2. In practice, does Algorithm 1 require only a limited number of multi-task ERM calls (representation updates) for more expressive function classes beyond the examples given in Section 5?

To address these questions, we performed experiments in three settings. Implementation details are provided in Appendix D.

7.1. Synthetic linear and semi-synthetic MNIST experiments

Synthetic logistic regression tasks. We first consider a binary logistic regression setting with synthetic data. Let $\theta_1^*, \ldots, \theta_T^* \in \mathbb{R}^d$ denote the parameters associated with T tasks, and let $\beta > 0$ be a parameter that governs the noise level of the tasks. There exists a k-dimensional shared representation, given by a semi-orthogonal matrix $B^* \in \mathbb{R}^{d \times k}$, such that for each $t \in T$, $\theta_t^* = B^* w_t^*$ for some $w_t^* \in \mathbb{R}^k$, where $\|w_t^*\| = \beta$. For each task t, the covariates $x \in \mathbb{R}^d$ are drawn from $\mathcal{N}(0, I)$, and the labels are generated such that $\Pr(y = 1|x) = \sigma(x^\top \theta_t^*)$, where σ denotes the logistic sigmoid function. We set ℓ to be the binary cross-entropy loss.



(*a*) (*b*)

Figure 3: Performance on semi-synthetic experiments with MNIST digits. (a) The solid curve shows, on average, how the number of representation updates increases over 50 binary digit classification tasks, with the shaded area showing one standard deviation. The dashed line represents linear growth. (b) Each box plot shows the distribution of 0-1 errors of the 50 produced predictors when evaluated on held-out data from the MNIST test set in one of the 10 independent trials.

We set the parameters to $d = 10$, $T = 50$, and target excess risk $\epsilon = 0.05$. We varied $k \in \{3, 5, 8\}$ and $\beta \in \{1, 4, 8\}$, and for each configuration we ran 10 trials, with B^* and w_t^*'s generated randomly. We implemented Algorithm 1 with N initialized as $k \log \frac{1}{\epsilon}$, based on Proposition 16. We set

$$m_N = \frac{1}{\epsilon^2}(dk + kN) \log \frac{1}{\epsilon}, \qquad \text{and} \qquad \tilde{m} = \frac{k}{\epsilon^2} \log \frac{1}{\epsilon}.$$

Note that the choice of m_N is smaller than what Theorem 14 requires, $\tilde{O}\left(\frac{k}{\epsilon^2}(dk + kN)\right)$, which highlights the practical efficiency of our algorithm.

To certify correctness, we evaluated each predictor output on a held-out dataset of size $\frac{32}{\epsilon^2}$ to verify that its excess risk is at most ϵ. Note that this step is used only for evaluation and is not part of the algorithm.

Table 1 reports the total number of representation updates performed by our algorithm (including task 1), and the top row of Figure 2 shows how the number of updates increases over tasks for each configuration. Across all regimes, the number of updates closely tracked the dimension of the representation k. This is consistent with Theorem 14. An interesting observation occurs in the high-noise regime, where the Bayes optimal risk is larger and so the absolute target risk is higher. The learner appeared to have not needed to fully uncover the underlying representation within $T = 50$ tasks when $k = 8$, as is seen in Figure 2 which shows the cumulative number of samples and number of representation updates over the tasks for each configuration.

Binary classification with MNIST digits. We also evaluated our algorithm in a semi-synthetic setting based on the MNIST handwritten digit database (Deng, 2012). Each task is defined as a binary classification problem, formed by randomly selecting one digit as the positive class and five other digits as the negative class. We flatten each 28×28 digit image so that $d = 784$, and we consider shared representations given by one-hidden layer neural networks,

$$h(x) = \mathrm{ReLU}(B^\top x), \qquad B \in \mathbb{R}^{d \times k}, \qquad k = 128.$$

For each task t, prediction is made using a linear prediction layer $w_t \in \mathbb{R}^k$ as $\mathbb{1}\{\langle w_t, h(x)\rangle \geq 0\}$.

We set $T = 50$ tasks and evaluated the performance of the learner under the 0-1 loss with a target absolute error of $\epsilon = 0.05$. The learner uses only the MNIST training set for multi-task ERM and few-shot property tests[6]. For few-shot property tests, 800 samples are drawn from each of the positive and negative classes. If the empirical risk of the learned predictor is below $\frac{2}{3}\epsilon$, the test succeeds and the learner moves on; otherwise, the representation is updated via multi-task ERM. For each representation update (including the first task), the learner draws m samples from the positive class and m samples from the negative class, where m is the number of available training images for the positive class (around 6000), and applies multi-task ERM using these samples together with those stored in the memory.

Across 10 trials, the learner updated its representation for an average of 4.9 times, with a standard deviation of 0.54. Figure 3(*a*) illustrates the average cumulative number of updates over 50 tasks. Observe that the number of updates grew sublinearly with the number of tasks. This suggests that the learner quickly learned a representation that enabled accurate linear predictors to be learned from only a small number of examples. To verify the correctness of our algorithm, we evaluated the predictors produced by the algorithm on data from the MNIST test set[7]. Figure 3(*b*) shows that the predictors consistently had 0-1 risks below $\epsilon = 0.05$.

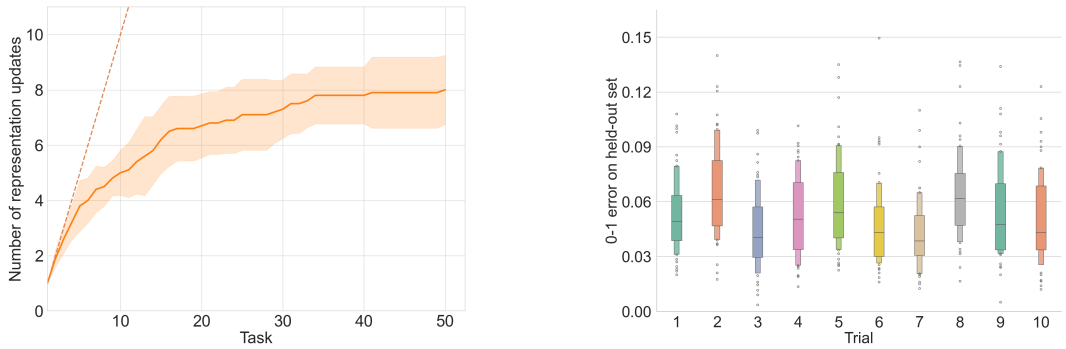
7.2. Semi-synthetic experiments with CIFAR-10 Images

So far, we have considered relatively simple settings involving linear synthetic data and the MNIST image dataset. To further address the second question (whether the learner requires only a small number of representation updates in more complex settings) and to assess the practicality of our algorithm, we now turn to a more realistic setting with CIFAR-10 images and deep convolutional representations.

Binary classification with CIFAR-10 images: experimental setup. We evaluated the algorithm on images from CIFAR-10 (Krizhevsky et al., 2009). As in the MNIST experiment, each task is formulated as a binary classification problem, with one CIFAR-10 class chosen as positive and five others as negative. We consider representations $h(x)$ given by a slightly modified ResNet-18 architecture with over 11 million parameters, which outputs a 256-dimensional feature vector (see Appendix D for details). For each task t, we again consider linear prediction layers parameterized by $w_t \in \mathbb{R}^k$ such that predictions are in the form of $\mathbb{1}\{\langle w_t, h(x)\rangle \geq 0\}$, with $k = 256$ in this setting.

We consider the 0-1 loss with a target absolute error of $\epsilon = 0.15$. The learner has access to the CIFAR-10 training set and uses binary cross-entropy as a surrogate loss during training. To make the experimental setup more realistic, we introduce minor practical modifications to the algorithm. We reserve a validation split from the CIFAR-10 training set consisting of 100 images per CIFAR-10 class. For few-shot property tests, the learner draws 800 samples from each of the positive and negative classes to train a linear prediction layer and evaluates its error on the validation split (100 validation images from each of the positive and negative classes). The test succeeds if the validation error is at most $\frac{2}{3}\epsilon$. If the test fails, the learner then draws 4900 samples per class from the training split and reuses stored data in the memory to update its representation via multi-task ERM.

6. During training, the binary cross-entropy loss is used as a surrogate loss function.
7. Again, this evaluation step is not part of the algorithm.



(*a*) (*b*)

Figure 4: Performance on semi-synthetic experiments with CIFAR-10 images. (a) The solid curve shows, on average, how the number of representation updates increases over 50 image classification tasks, with the shaded area showing one standard deviation. The dashed line represents linear growth. (b) Each box plot shows the distribution of 0-1 errors of the produced predictors when evaluated on held-out data from the CIFAR-10 test set in one of the 10 independent trials.

Results. We ran 10 trials with $T = 50$ tasks. On average, the learner updated it representation 8.0 times with a standard deviation of 1.18, and Figure 4(*a*) shows that the number of updates grew sublinearly with the number of tasks. We also evaluated the produced predictors on data from the CIFAR-10 test set. Figure 4(*b*) shows that the errors across tasks and trials were consistently below $\epsilon = 0.15$. These results confirm that even for more expressive function classes based on deep convolutional architectures, our algorithm can perform effective lifelong representation learning while using a modest number of representation updates via multi-task ERM. In addition, these results highlight that our algorithm can be implemented and run in a practical setup using standard architectures and training procedures.

8. Discussion and future work

In Section 3.2, we introduced the notion of ϵ-independence (Definition 7). We now examine an alternative definition of ϵ-independence.

Definition 17 (ϵ-pointwise-independence) *Given the setting of Definition 7, for any representation $h \in \mathcal{H}$, we say that f_n is* ϵ-pointwise *independent of $\{f_1, \ldots, f_{n-1}\}$ with respect to $(\mathcal{H}, \mathcal{F})$ and h, if there exist $h' \in \mathcal{H}$ and $f'_1, \ldots, f'_{n-1} \in \mathcal{F}$ such that*

$$\max_{i \in [n-1]} \mathcal{E}(h', P_{f_i \circ h}) \leq \epsilon, \tag{7}$$

but $\mathcal{E}(h', P_{f_n \circ h}) > \frac{\epsilon}{2}$.

At first glance, this seems to be a more natural definition. Eq. (7) requires that h' incurs low representation error on *each* of the earlier tasks induced by (h, f_i)'s, which seems well aligned with the

lifelong learning objective. In addition, this condition is weaker than Eq. (3) in Definition 7, i.e.,

$$\sum_{i=1}^{n-1} \mathcal{E}(h', P_{f_i \circ h}) \leq \epsilon,$$

and so, following a similar analysis, this definition would suggest an improved sample complexity bound.

However, we now provide a negative result which shows that the condition in Eq. (7) can, in fact, be insufficient. Consider noiseless, binary linear classification tasks with a low-dimensional linear representation. Let $\mathcal{H} = \{x \mapsto B^\top x : B \in \mathbb{R}^{d \times k}, B^\top B = I\}, \mathcal{F} = \left\{z \mapsto w^\top z : w \in \mathbb{R}^k, \|w\| = 1\right\}$, and ℓ be the 0-1 loss. Suppose $P_X = \mathcal{N}(0, I)$. Then, for unit vectors $u, v \in \mathbb{R}^d$,

$$\Pr_{x \sim P_X}\left[\text{sign}(u^\top x) \neq \text{sign}(v^\top x)\right] = \frac{\theta(u, v)}{\pi},$$

where $\theta(\cdot, \cdot)$ denotes the angle between two vectors (e.g., Long, 1995).

Proposition 18 *In the noiseless linear classification setting described above, suppose $d > k$. There exists an infinite sequence of ϵ-pointwise independent tasks under Definition 17.*

See Appendix A.3 for its proof. Despite this negative result, it may be possible to formulate intermediate notions of agreement on prior tasks that lie between the aggregate condition in Eq. (3) and the pointwise condition in Eq. (7), which could lead to improved sample complexity bounds. Understanding how to ensure that such conditions hold, both in terms of the number of samples needed and algorithmic techniques, would be an interesting direction for future study, with potential connections to collaborative PAC learning (Blum et al., 2017). At the same time, it is important to also understand whether such definitions admit meaningful bounds on their induced notions of task-eluder dimension for natural function classes.

Beyond this, several other open questions remain. In this work, we assumed a well-specified model (Assumpton 1). While this relaxes the noiseless, realizable assumption studied in prior work, it remains to be seen if it can be further weakened. Other future directions include studying whether the learner can infer noise levels from other tasks under additional structural assumptions, and bounding the task-eluder dimension for more general function classes.

Acknowledgments

This work was partially supported by NSF CAREER Award CCF 2238876. CZ acknowledges funding support by NSF CAREER Award IIS-2440266.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24.

Aliakbarpour, M., Bairaktari, K., Brown, G., Smith, A., Srebro, N., and Ullman, J. (2024). Metalearning with very few samples per task. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 46–93. PMLR.

Alquier, P., Pontil, M., et al. (2017). Regret bounds for lifelong learning. In *Artificial intelligence and statistics*, pages 261–269. PMLR.

Balcan, M.-F., Blum, A., and Vempala, S. (2015). Efficient representations for lifelong learning and autoencoding. In *Conference on Learning Theory*, pages 191–210. PMLR.

Balcan, M.-F. and Haghtalab, N. (2020). Noise in classification. *arXiv preprint arXiv:2010.05080*.

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

Blum, A., Haghtalab, N., Procaccia, A. D., and Qiao, M. (2017). Collaborative pac learning. *Advances in Neural Information Processing Systems*, 30.

Cao, X., Liu, W., and Vempala, S. (2022). Provable lifelong learning of representations. In *International Conference on Artificial Intelligence and Statistics*, pages 6334–6356. PMLR.

Caruana, R. (1997). Multitask learning. *Machine learning*, 28:41–75.

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

de Masson D'Autume, C., Ruder, S., Kong, L., and Yogatama, D. (2019). Episodic memory in lifelong language learning. *Advances in Neural Information Processing Systems*, 32.

Deng, L. (2012). The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. (2014). Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655. PMLR.

Du, S. S., Hu, W., Kakade, S. M., Lee, J. D., and Lei, Q. (2021). Few-shot learning via learning the representation, provably. In *International Conference on Learning Representations*.

Duong, T., Wang, Z., and Zhang, C. (2024). Beyond task diversity: provable representation transfer for sequential multitask linear bandits. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Finn, C., Rajeswaran, A., Kakade, S., and Levine, S. (2019). Online meta-learning. In *International conference on machine learning*, pages 1920–1930. PMLR.

Foster, D. J., Rakhlin, A., Simchi-Levi, D., and Xu, Y. (2020). Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. *arXiv preprint arXiv:2010.03104*.

Hanneke, S. (2024). The star number and eluder dimension: Elementary observations about the dimensions of disagreement. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 2308–2359. PMLR.

Haussler, D. (1992). Decision theoretic generalizations of the pac model for neural net and other learning applications. *Information and computation*, 100(1):78–150.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Hu, J., Chen, X., Jin, C., Li, L., and Wang, L. (2021). Near-optimal representation learning for linear bandits and linear rl. In *International Conference on Machine Learning*, pages 4349–4358. PMLR.

Isele, D. and Cosgun, A. (2018). Selective experience replay for lifelong learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Kearns, M. J. and Vazirani, U. (1994). *An introduction to computational learning theory*. MIT press.

Kingma, D. P. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Kong, W. and Valiant, G. (2018). Estimating learnability in the sublinear data regime. *Advances in Neural Information Processing Systems*, 31.

Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images. Technical report, University of Toronto.

Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

Li, G., Kamath, P., Foster, D. J., and Srebro, N. (2022a). Understanding the eluder dimension. *Advances in Neural Information Processing Systems*, 35:23737–23750.

Li, L., Littman, M. L., and Walsh, T. J. (2008). Knows what it knows: a framework for self-aware learning. In *Proceedings of the 25th international conference on Machine learning*, pages 568–575.

Li, Y., Li, M., Asif, M. S., and Oymak, S. (2022b). Provable and efficient continual representation learning. *arXiv preprint arXiv:2203.02026*.

Long, P. (1995). On the sample complexity of pac learning half-spaces against the uniform distribution. *IEEE Transactions on Neural Networks*, 6(6):1556–1559.

Lowrey, K., Rajeswaran, A., Kakade, S., Todorov, E., and Mordatch, I. (2018). Plan online, learn offline: Efficient learning and exploration via model-based control. *arXiv preprint arXiv:1811.01848*.

Maurer, A. (2016). A chain rule for the expected suprema of gaussian processes. *Theoretical Computer Science*, 650:109–122.

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

McCloskey, M. and Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

Osband, I. and Van Roy, B. (2014). Model-based reinforcement learning and the eluder dimension. *Advances in neural information processing systems*, 27.

Pentina, A. and Ben-David, S. (2015). Multi-task and lifelong learning of kernels. In *Algorithmic Learning Theory: 26th International Conference, ALT 2015, Banff, AB, Canada, October 4-6, 2015, Proceedings 26*, pages 194–208. Springer.

Pentina, A. and Ben-David, S. (2018). Multi-task kernel learning based on probabilistic lipschitzness. In *Algorithmic Learning Theory*, pages 682–701. PMLR.

Pentina, A. and Lampert, C. (2014). A pac-bayesian bound for lifelong learning. In *International Conference on Machine Learning*, pages 991–999. PMLR.

Pentina, A. and Urner, R. (2016). Lifelong learning with weighted majority votes. *Advances in Neural Information Processing Systems*, 29.

Pollard, D. (1984). *Convergence of Stochastic Processes*. Clinical Perspectives in Obstetrics and Gynecology. Springer New York.

Qin, Y., Menara, T., Oymak, S., Ching, S., and Pasqualetti, F. (2022). Non-stationary representation learning in sequential linear bandits. *IEEE Open Journal of Control Systems*, 1:41–56.

Rebuffi, S.-A., Kolesnikov, A., Sperl, G., and Lampert, C. H. (2017). icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010.

Russo, D. and Van Roy, B. (2013). Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Shalev-Shwartz, S. et al. (2012). Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194.

Sodhani, S., Faramarzi, M., Mehta, S. V., Malviya, P., Abdelsalam, M., Janarthanan, J., and Chandar, S. (2022). An introduction to lifelong supervised learning. *arXiv preprint arXiv:2207.04354*.

Thrun, S. and Mitchell, T. M. (1995). Lifelong robot learning. *Robotics and autonomous systems*, 15(1-2):25–46.

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

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

Verzelen, N. and Villers, F. (2010). Goodness-of-fit tests for high-dimensional gaussian linear models. *The Annals of Statistics*, 38(2):704–752.

Watkins, A., Ullah, E., Nguyen-Tang, T., and Arora, R. (2023). Optimistic rates for multi-task representation learning. *Advances in Neural Information Processing Systems*, 36:2207–2251.

Xu, Z. and Tewari, A. (2021). Representation learning beyond linear prediction functions. *Advances in Neural Information Processing Systems*, 34:4792–4804.

Yang, J., Hu, W., Lee, J. D., and Du, S. S. (2021). Impact of representation learning in linear bandits. In *International Conference on Learning Representations*.

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

Zhang, Y. and Yang, Q. (2021). A survey on multi-task learning. *IEEE transactions on knowledge and data engineering*, 34(12):5586–5609.

Outline of the Appendix

A Supplementary material for Section 3 **24**
A.1 Hardness of few-shot property testing 24
A.2 Proof of Proposition 9 25
A.3 Proof of Proposition 18 27
A.4 An uncentered version of task eluder dimension 28
A.5 Pseudocode for lifelong representation learning with known task-eluder dimension 29

B Supplementary material for Section 4 **30**
B.1 Proof of Theorem 14 30
B.2 Auxiliary lemma 33

C Supplementary material for Section 5 **35**
C.1 Key lemma for bounding the task-eluder dimension 35
C.2 Noisy linear regression 37
C.3 Classification with logistic regression 39
C.4 Classification with random classification noise and the 0-1 loss 41
C.5 Auxiliary lemmas 41

D Supplementary material for Section 7 **44**
D.1 Implementation details 44

Appendix A. Supplementary material for Section 3

A.1. Hardness of few-shot property testing

Proposition 19 (Formal version of Proposition 5) *Fix a subspace U of $\mathbb{R}^d$ with dimension $r \leq d/2$. Consider two classes of functions, $\mathcal{G} := \{x \mapsto \langle x, w\rangle : \|w\| \leq 1\}$ and $\mathcal{G}_0 := \{x \mapsto \langle x, w\rangle : \|w\| \leq 1, w \in U\} \subset \mathcal{G}$, where the latter is $\mathcal{G}$ restricted to the subspace U (equiv. to having a low-dimensional linear representation fixed). For any distribution P and $\mathcal{G}'$, let $\kappa_P(\mathcal{G}') := \inf_{g\in\mathcal{G}'} \mathbb{E}_{(x,y)\sim P}\left[(g(x)-y)^2\right]$ denote the risk of the best function in $\mathcal{G}'$.*

Consider the following two hypotheses:

$$H_0 = \left\{P \in \Delta(\mathbb{R}^d \times \mathbb{R}) : \kappa_P(\mathcal{G}_0) = \kappa_P(\mathcal{G})\right\}, \text{ and } H_1 = \left\{P \in \Delta(\mathbb{R}^d \times \mathbb{R}) : \kappa_P(\mathcal{G}_0) > \kappa_P(\mathcal{G}) + 0.9\right\}.$$

Then, there exists some constant c such that, with $n \leq c\sqrt{d}$ examples, for any hypothesis tester $\psi : (\mathbb{R}^d \times \mathbb{R})^n \to \{0,1\}$,

$$\sup_{b\in\{0,1\}} \sup_{P\in H_b} \Pr_{(x_i,y_i)_{i=1}^n \sim P^n} (\psi((x_i,y_i)_{i=1}^n) \neq b) \geq \frac{1}{3}.$$

Proof Let $U^\perp$ denote the orthogonal complement of U, and $P_{U^\perp}$ denote the orthogonal projection matrix onto $U^\perp$. The high level idea of our proof is that, with insufficient number of examples, no algorithm can reliably tell if the features in the $U^\perp$ subspace have some nontrivial correlation with the labels. Consider two distributions, Q_0 and Q_1, over $(\mathbb{R}^d \times \mathbb{R})^n$:

- Q_0: each example is drawn i.i.d. from P_0. Here, we define P_0 such that $(x, y) \sim P_0$ iff $x \sim \mathcal{N}(0, I_d)$, $y \sim \mathcal{N}(0, 1)$, and x, y are independent. We have $\kappa_{P_0}(\mathcal{G}) = \kappa_{P_0}(\mathcal{G}_0) = \sigma^2 = 1$; that is, $P_0 \in H_0$.
- Q_1: this is a mixture of distributions $P_1(w_*)^n$, where w_* is drawn from the uniform distribution over $\left\{z \in U^\perp : \|z\| = \sqrt{0.99}\right\}$. Herein, $P_1(w_*)$ is defined such that $x \sim N(0, I_d)$, $y = \langle w_*, x\rangle + \eta$, $\eta \sim \mathcal{N}(0, 0.01)$, and x, η are independent. Then, for any w_*, we have $\kappa_{P_1(w_*)}(\mathcal{G}) = 0.01$ and $\kappa_{P_1(w_*)}(\mathcal{G}_0) = 0.01 + \|P_{U^\perp} w_*\|^2 = 1$. Therefore, for any w_*, $\kappa_{P_1(w_*)}(\mathcal{G}_0) > \kappa_{P_1(w_*)}(\mathcal{G}) + 0.9$, and $P_1(w_*) \in H_1$.

To prove Proposition 19, it suffices to show that for any hypothesis tester $\psi : (\mathbb{R}^d \times \mathbb{R})^n \to \{0,1\}$,

$$\sup_{b\in\{0,1\}} \Pr_{(x_i,y_i)_{i=1}^n \sim Q_b} (\psi((x_i,y_i)_{i=1}^n) \neq b) \geq \frac{1}{3}. \tag{8}$$

To see this, we note that for any $b \in \{0,1\}$,

$$\sup_{P\in H_b} \Pr_{(x_i,y_i)_{i=1}^n \sim P^n} (\psi((x_i,y_i)_{i=1}^n) \neq b) \geq \Pr_{(x_i,y_i)_{i=1}^n \sim Q_b} (\psi((x_i,y_i)_{i=1}^n) \neq b),$$

since Q_b are mixtures of distributions of P^n, for $P \in H_b$.

To establish this, we reduce our hypothesis testing problem from a known hypothesis testing problem due to Verzelen and Villers (2010); Kong and Valiant (2018), which is known to require a large number of samples:

Proposition 20 (Kong and Valiant 2018, Proposition 2; Verzelen and Villers 2010, Proposition 4.2) *Consider two probability distributions in $\Delta((\mathbb{R}^{d'} \times \mathbb{R})^n)$ over random variables $(x'_i, y'_i)_{i=1}^n$:*

1. *Pure noise Q'_0: each (x'_i, y'_i) is drawn independently such that $x'_i \sim \mathcal{N}(0, I_{d'})$ and $y'_i \sim \mathcal{N}(0,1)$, with x'_i independent of y'_i;*
2. *Little noise Q'_1: let w'_* be a random vector drawn uniformly from the sphere in $\mathbb{R}^{d'-1}$ with radius $\sqrt{0.99}$. Conditioned on w'_*, each (x'_i, y'_i) is drawn independently such that $x'_i \sim \mathcal{N}(0, I_{d'})$ and $y'_i = \langle w'_*, x'_i \rangle + \eta_i$, where $\eta_i \sim \mathcal{N}(0, 0.01)$.*

There is a universal constant c such that when $n \leq c\sqrt{d'}$, such that for any hypothesis tester (that may use extra independent randomness in E) $f : (\mathbb{R}^{d'} \times \mathbb{R})^n \times E \to \{0,1\}$ and $\mathcal{E}$ being a distribution over random seeds e:

$$\sup_{b \in \{0,1\}} \Pr_{(x'_i, y'_i)_{i=1}^n \sim Q'_b, e \sim \mathcal{E}} \left[f((x'_i, y'_i)_{i=1}^n, e) \neq b \right] \geq \frac{1}{3}. \tag{9}$$

We now provide the reduction from the above hypothesis testing task to ours. Pick an orthogonal matrix $B = [B_\| \; B_\perp] \in R^{d \times d}$ whose first r columns span U and last d' columns span $U^\perp$. For any (x'_i, y'_i), consider

$$x_i = B_\perp x'_i + B_\| z_i,$$

where $z_i \sim \mathcal{N}(0, I_r)$ is independent of (x'_i, y'_i), and $y_i = y'_i$. It is easy to verify that $x_i \sim \mathcal{N}(0, I_d)$. Consider two cases:

- Under Q'_0, $y_i = y'_i \sim \mathcal{N}(0,1)$ and is independent of x_i. So $(x_i, y_i)_{i=1}^n$ follows the distribution Q_0.
- Under Q'_1, $y'_i = \langle x'_i, w'_* \rangle + \eta_i$, where w'_* is from the uniform distribution over $\mathbb{S}^{d'}(0, \sqrt{0.99})$. Set $w_* = B_\perp w'_*$. Then, it is easy to verify that $y = \langle x_i, w_* \rangle + \eta_i$, and w_* is drawn from the uniform distribution over $\left\{ z \in U^\perp : \|z\| = \sqrt{0.99} \right\}$. In other words, $(x_i, y_i)_{i=1}^n$ follows the same distribution as Q_1.

Given any test $\psi : (\mathbb{R}^d \times \mathbb{R})^n \to \{0,1\}$ that tries to distinguish Q_0 and Q_1, we now construct a randomized test $\psi' : (\mathbb{R}^{d'} \times \mathbb{R})^n \times \mathbb{R}^n \to \{0,1\}$ (with space of random seeds $E = \mathbb{R}^n$) that distinguishes Q'_0 and Q'_1 with the same worst-case error rate. Specifically, define $\psi'((x'_i, y'_i)_{i=1}^n, (z_i)_{i=1}^n) = \psi((x_i, y_i)_{i=1}^n)$, and $\mathcal{E} = N(0, I_r)^n$.

To see why they have the same worst-case error rates, we note that for any $b \in \{0,1\}$,

$$\Pr_{(x'_i, y'_i)_{i=1}^n \sim Q'_b, (z_i)_{i=1}^n \sim N(0, I_r)^n} \left[\psi'((x'_i, y'_i)_{i=1}^n, (z_i)_{i=1}^n) \neq b \right] = \Pr_{(x_i, y_i)_{i=1}^n \sim Q_b} \left[\psi((x_i, y_i)_{i=1}^n) \neq b \right],$$

due to the construction of $(x_i, y_i)_{i=1}^n$ based on $(x'_i, y'_i)_{i=1}^n$ and $(z_i)_{i=1}^n$ above. Thus, Eq. (8) follows from Eq. (9). ■

A.2. Proof of Proposition 9

Proposition 9. *For any $\epsilon \geq 0$,*

$$\mathrm{TEdim}(\mathcal{H}, \mathcal{F}, \epsilon) \leq 2 \min\left(|\mathcal{H}|, |\mathcal{F}| \right).$$

Proof It suffices to show $\mathrm{TEdim}_h(\mathcal{H}, \mathcal{F}, \epsilon) \leq 2 \min\left(|\mathcal{H}|, |\mathcal{F}| \right)$ for any $h \in \mathcal{H}$.

We first show that $\mathrm{TEdim}_h(\mathcal{H}, \mathcal{F}, \epsilon) \leq 2|\mathcal{H}|$. Let $\{f_i\}_{i=1}^n$ be any sequence of ϵ-independent tasks with respect to $(\mathcal{H}, \mathcal{F})$ and h. For each $j \leq n$, let $\mathcal{H}_j \subset \mathcal{H}$ denote the set of possible representations that witness the independence of task f_j from tasks $\{f_1, \ldots, f_{j-1}\}$. That is, $h' \in \mathcal{H}_j$

if

$$\sum_{i=1}^{j-1} \mathcal{E}(h', P_{f_i \circ h}) \leq \epsilon, \tag{10}$$

but $\mathcal{E}(h', P_{f_j \circ h}) > \frac{\epsilon}{2}$. For any $h' \in \mathcal{H}$ and j, let

$$m_j(h') = \left|\left\{i : i \leq j, h' \in \mathcal{H}_i\right\}\right|$$

denote the number of times h' appears in the set of witness representations up until task j. It suffices to show that for any h', $m_n(h') \leq 2$, i.e., each h' can witness at most twice, as this implies that any sequence of ϵ-independent tasks will have length at most $2|\mathcal{H}|$.

Suppose, for contradiction, $i_1 < i_2 < i_3$ are three indices in $[n]$ such that $h' \in \mathcal{H}_{i_1} \cap \mathcal{H}_{i_2} \cap \mathcal{H}_{i_3}$. Since $h' \in \mathcal{H}_{i_1} \cap \mathcal{H}_{i_2}$,

$$\mathcal{E}(h', P_{f_{i_1} \circ h}) > \frac{\epsilon}{2} \quad \text{and} \quad \mathcal{E}(h', P_{f_{i_2} \circ h}) > \frac{\epsilon}{2}; \tag{11}$$

Furthermore, since $h' \in \mathcal{H}_{i_3}$,

$$\sum_{i=1}^{i_3-1} \mathcal{E}(h', P_{f_i \circ h}) \leq \epsilon. \tag{12}$$

This is in contradiction with Eq. (11), which implies that $\sum_{i=1}^{i_3-1} \mathcal{E}(h', P_{f_i \circ h}) > \epsilon$.

We now turn to show $\mathrm{TEdim}(\mathcal{H}, \mathcal{F}, \epsilon) \leq 2|\mathcal{F}|$. To this end, it suffices to show that for any sequence of ϵ-independent tasks $f_1, \ldots, f_n$ with respect to $(\mathcal{H}, \mathcal{F})$ and h, the number of appearances of any $f \in \mathcal{F}$ in $\{f_1, \ldots, f_n\}$ can be at most 2. This will conclude the proof since this implies that any sequence of ϵ-independent tasks will have length at most $2|\mathcal{F}|$.

Fix one such ϵ-independent sequence and any $f \in \mathcal{F}$. Assume towards contradiction that i_1, i_2, i_3 are three indices in $[n]$ such that $f_{i_1} = f_{i_2} = f_{i_3}$. Since f_{i_3} is ϵ-independent of $f_1, \ldots, f_{i_3-1}$, we have that there exists $h' \in \mathcal{H}$ such that

$$\sum_{i=1}^{i_3-1} \mathcal{E}(h', P_{f_i \circ h}) \leq \epsilon, \tag{13}$$

but

$$\mathcal{E}(h', P_{f_{i_3} \circ h}) > \epsilon/2, \tag{14}$$

Since $\mathcal{E}(h', P_{f_{i_1} \circ h}) = \mathcal{E}(h', P_{f_{i_2} \circ h}) = \mathcal{E}(h', P_{f_{i_3} \circ h})$, Eq. (13) implies that $\mathcal{E}(h', P_{f_{i_3} \circ h}) \leq \frac{\epsilon}{2}$, which contradicts with Eq. (14). ■

Remark 21 *The constant* 2 *in front of* $2\min(|\mathcal{H}|, |\mathcal{F}|)$ *is due to out-of-task excess risk threshold* $\frac{\epsilon}{2}$ *in the definition of* ϵ*-independence of tasks (Definition 7). If that threshold were* $\frac{\epsilon}{C}$*, we will obtain* $\mathrm{TEdim}(\mathcal{H}, \mathcal{F}, \epsilon) \leq C\min(|\mathcal{H}|, |\mathcal{F}|)$ *here. In summary, our definition of task-eluder dimension is robust to the choice of the constant in the out-of-task excess risk threshold.*

A.3. Proof of Proposition 18

For completeness, we restate below the alternative definition of ϵ-independence (ϵ-pointwise-independence) in Section 8, the setting of the negative example, and Proposition 18.

Definition 17. *Given the setting of Definition 7, for any representation $h \in \mathcal{H}$, we say that f_n is* ϵ-pointwise *independent of $\{f_1, \ldots, f_{n-1}\}$ with respect to $(\mathcal{H}, \mathcal{F})$ and h if there exist $h' \in \mathcal{H}$ such that*

$$\max_{i \in [n-1]} \mathcal{E}(h', P_{f_i \circ h}) \leq \epsilon,$$

but $\mathcal{E}(h', P_{f_n \circ h}) > \frac{\epsilon}{2}$.

Setting. We consider noiseless, binary linear classification tasks under a shared low-dimensional linear representation. Let $\mathcal{H} = \left\{x \mapsto B^\top x : B \in \mathbb{R}^{d \times k}, B^\top B = I\right\}, \mathcal{F} = \left\{z \mapsto w^\top z : w \in \mathbb{R}^k, \|w\| = 1\right\}$, and let ℓ denote the 0-1 loss. Suppose $P_X = \mathcal{N}(0, I)$. Then, for unit vectors $u, v \in \mathbb{R}^d$,

$$\Pr_{x \sim P_X}\left[\operatorname{sign}(u^\top x) \neq \operatorname{sign}(v^\top x)\right] = \frac{\theta(u, v)}{\pi},$$

where $\theta(\cdot, \cdot)$ denotes the angle between two vectors.

Proposition 18. *In the noiseless linear classification setting described above, suppose $d > k$. There exists an infinite sequence of ϵ-pointwise independent tasks under Definition 17.*

Proof [Proof of Proposition 18] Fix $B = [e_1\ e_2\ \ldots\ e_k] \in \mathbb{R}^{d \times k}$, where e_i denotes the i-th standard basis vector in $\mathbb{R}^d$. Let $v = [1\ 0\ \ldots\ 0] \in \mathbb{R}^k$. Consider a sequence, $\{(B, w_n)\}_n$, where $w_n \equiv v$; that is, the same task is seen at every step. It suffices to show that for every n, (B, w_n) is ϵ-pointwise independent of $\{(B, w_1), \ldots, (B, w_{n-1})\}$. To this end, consider $B' = [s\ e_3\ \ldots\ e_k\ e_{k+1}] \in \mathbb{R}^{d \times k}$, where

$$s = e_1 \cos\lambda + e_2 \sin\lambda, \qquad \lambda \in \left(\frac{\pi\epsilon}{2}, \pi\epsilon\right].$$

Let $U = \operatorname{span}(B')$. Observe that

1. for each $i < n$, by choosing f_i' to be v,

$$\begin{aligned}
&\mathbb{E}_{P_{f_i \circ h}}\left[\ell((f_i' \circ h')(x), y) - \ell((f_i \circ h)(x), y)\right] \\
&= \Pr_{x \sim P_X}\left[\operatorname{sign}(x^\top B' v) \neq \operatorname{sign}(x^\top B v)\right] \\
&= \frac{\theta(Bv, B'v)}{\pi} = \frac{\lambda}{\pi} \leq \epsilon;
\end{aligned}$$

   and thus, $\mathcal{E}(h', P_{f_i \circ h}) \leq \epsilon$.

2. for any f_n',

$$\begin{aligned}
&\mathbb{E}_{P_{f_n \circ h}}\left[\ell((f_n' \circ h')(x), y) - \ell((f_n \circ h)(x), y)\right] \\
&\geq \frac{\theta(Bv, U)}{\pi} \\
&= \frac{\theta(Bv, s)}{\pi} = \frac{\lambda}{\pi} > \frac{\epsilon}{2},
\end{aligned}$$

   where $\theta(Bv, U)$ denotes the angle between Bv and the subspace U. Thus, $\mathcal{E}(h', P_{f_n \circ h}) > \frac{\epsilon}{2}$.

Therefore, by Definition 17, (B, v_n) is ϵ-pointwise independent of its predecessors at every step n, which yields an infinite sequence. ■

A.4. An uncentered version of task eluder dimension

Definition 22 (ϵ-uncentered independence) *Let $\mathcal{H} \subset \{h : \mathcal{X} \to \mathcal{Z}\}$ be a class of representations, and $\mathcal{F} \subset \{f : \mathcal{Z} \to \mathbb{R}\}$ be a class of prediction layers that operate on these representations. Let $\mathbb{P} = \{P_{f \circ h} : f \in \mathcal{F}, h \in \mathcal{H}\}$ be the probabilistic model and $\ell : \mathbb{R} \times \mathcal{Y} \to [0, 1]$ be a loss function. We say f_n is* ϵ-uncentered-independent *of $\{f_1, \ldots, f_{n-1}\}$ with respect to $(\mathcal{H}, \mathcal{F})$, if there exists $h, h' \in \mathcal{H}$ such that*

$$\sum_{i=1}^{n-1} \mathcal{E}(h', P_{f_i \circ h}) \leq \epsilon,$$

but $\mathcal{E}(h', P_{f_n \circ h}) > \frac{\epsilon}{2}$.

Definition 23 (Uncentered Task-eluder dimension) *Given the setting of Definition 22, we denote by $\mathrm{UTEdim}\,(\mathcal{H}, \mathcal{F}, \epsilon)$ the length of the longest sequence $\{f_i\}_i$, such that each tuple is ϵ-independent of its predecessors.*

A.5. Pseudocode for lifelong representation learning with known task-eluder dimension

Algorithm 2 provides pseudocode of our lifelong representation learning algorithm in the setting where the task-eluder dimension is known in advance.

Algorithm 2: Lifelong representation learning (known task-eluder dimension)

Input: $\mathcal{H}$, $\mathcal{F}$, target error ϵ, $\dim(\mathcal{H}, \mathcal{F}, \epsilon)$, confidence δ, number of tasks T, noise levels $(\kappa_t)_t$;

Initialize memory $\mathcal{M} \leftarrow \emptyset$ and $N \leftarrow \dim(\mathcal{H}, \mathcal{F}, \epsilon)$;

for *task* $t = 1$ **do**

  Draw a sample S_1 of size m_N from $\mathcal{P}_1^{m_N}$, apply ERM to learn $\hat{h}$ and $\hat{f}_1$, and output $\hat{f}_1 \circ \hat{h}$;

  Set $n \leftarrow 1$ and $t_n \leftarrow t$, and update the memory $\mathcal{M} \leftarrow \mathcal{M} \cup \{S_{t_n}\}$;

for *tasks* $t = 2, \ldots, T$ **do**

  // Few-shot property test: check if $\hat{h}$ admits a hypothesis for current task t with risk at most ϵ

  Draw a sample $\tilde{S}_t$ of size $\tilde{m}$ from $\mathcal{P}_t^{\tilde{m}}$, and apply ERM with current $\hat{h}$ to learn $\tilde{f}_t$;

  if $\widehat{\mathcal{L}}_{\tilde{S}_t}(\tilde{f}_t \circ \hat{h}) \leq \kappa_t + \frac{3}{4}\epsilon$ **then**

    Output $\tilde{f}_t \circ \hat{h}$;

  else

    // Multi-task ERM on a subset of tasks where few-shot property test failed (plus task 1)

    Set $n \leftarrow n + 1$;

    Draw a sample S_t of size m_N from $\mathcal{P}_t^{m_N}$;

    Set $t_n \leftarrow t$, and update the memory $\mathcal{M} \leftarrow \mathcal{M} \cup \{S_{t_n}\}$;

    Apply ERM over the samples stored in the memory to learn

$$\hat{h}, \check{f}_1^{(t)}, \ldots, \check{f}_n^{(t)} \leftarrow \operatorname*{argmin}_{\substack{h \in \mathcal{H} \\ f_1, \ldots, f_n \in \mathcal{F}}} \frac{1}{n} \sum_{i=1}^{n} \widehat{\mathcal{L}}_{S_{t_i}}(f_i, h)$$

    Set $\hat{f}_t \leftarrow \check{f}_n^{(t)}$, update $\hat{h}$, and output $\hat{f}_t \circ \hat{h}$.

Appendix B. Supplementary material for Section 4

B.1. Proof of Theorem 14

We first state a more precise version of Theorem 14 with explicit constants.

Theorem 24 (Restatement of Theorem 14) *Let* $\Xi = \mathrm{TEdim}(\mathcal{H}, \mathcal{F}, \epsilon) < \infty$. *Suppose* $\mathcal{C}\left(\mathcal{H}, \frac{\epsilon}{256\Xi}\right) < \infty$ *and* $\mathcal{C}\left(\mathcal{F}, \frac{\epsilon}{256\Xi}\right) < \infty$. *In Algorithm 1, for each* N, *set*

$$m_N = \frac{256N}{\epsilon^2}\left[\log \mathcal{C}\left(\mathcal{H}, \frac{\epsilon}{64N}\right) + N \log \mathcal{C}\left(\mathcal{F}, \frac{\epsilon}{64N}\right) + \log \frac{16 \log T \cdot \sum_{i=0}^{\log N} \binom{T}{2^i}}{\delta}\right] + \frac{64}{\epsilon^2},$$

and set

$$\tilde{m} = \frac{1024}{\epsilon^2}\left[\log \mathcal{C}\left(\mathcal{F}, \frac{\epsilon}{128}\right) + \log \frac{8T}{\delta}\right] + \frac{256}{\epsilon^2}.$$

Then, with probability at least $1 - \delta$,

- *For every task, Algorithm 1 outputs a predictor with excess risk at most* ϵ;
- *Algorithm 1 performs multi-task ERM at most* 2Ξ *times;*
- *The sample complexity of Algorithm 1 is upper bounded by*

$$\tilde{\mathcal{O}}\left(\frac{T}{\epsilon^2} \log \mathcal{C}\left(\mathcal{F}, \frac{\epsilon}{128}\right) + \frac{\Xi^2}{\epsilon^2}\left[\log \mathcal{C}\left(\mathcal{H}, \frac{\epsilon}{128\Xi}\right) + \Xi \log \mathcal{C}\left(\mathcal{F}, \frac{\epsilon}{128\Xi}\right)\right]\right).$$

In addition, the size of the memory buffer it requires is at most

$$\tilde{\mathcal{O}}\left(\frac{\Xi^2}{\epsilon^2}\left[\log \mathcal{C}\left(\mathcal{H}, \frac{\epsilon}{128\Xi}\right) + \Xi \log \mathcal{C}\left(\mathcal{F}, \frac{\epsilon}{128\Xi}\right)\right]\right).$$

Proof We begin by introducing some additional notation. For each task t, let $\hat{h}_t$ denote the maintained representation at the beginning of task t. Let n_t and N_t denote the values of n and N at the beginning of task t, respectively. Denote by N_{T+1} the value of N at the end of task T.

We consider a "sample tape" model (similar to the "reward stack model" in Lattimore and Szepesvári (2020, Section 4.6)), where for each $f \circ h$, there is a stack of i.i.d. samples drawn from $P_{f\circ h}$ before the first task begins. When the learner draws a sample of size m from $P_{f\circ h}$, it receives the next m entries from the corresponding tape, denoted by $S_{f\circ h}(m)$. To avoid clutter, we use $S_t(m)$ to refer to a sample of size m from $\mathcal{P}_t = P_{f_t^* \circ h^*}$.

Clean event. Consider the following events. Let

$$\mathcal{A} := \left\{\forall t \in \{2, \ldots, T\},\ \left|\mathcal{L}_{\mathcal{P}_t}(f \circ \hat{h}_t) - \widehat{\mathcal{L}}_{S_t(\tilde{m})}(f \circ \hat{h}_t)\right| \le \frac{\epsilon}{4},\ \forall f \in \mathcal{F}\right\},$$

and for each $N \in \left\{2^0, 2^1, 2^2, \ldots, 2^{\lceil \log T \rceil}\right\}$, let

$$\mathcal{B}_N := \left\{\forall n \in [N], \forall (t_1, \ldots, t_n) \in \binom{[T]}{n},\right.$$
$$\left.\left|\sum_{i=1}^{n} \mathcal{L}_{\mathcal{P}_{t_i}}(f_i \circ h) - \sum_{i=1}^{n} \widehat{\mathcal{L}}_{S_{t_i}(m_N)}(f_i \circ h)\right| \le \epsilon,\ \forall (h, f_1, \ldots, f_n)\right\}.$$

We now define the following notion of a *clean event*,

$$\mathcal{E} = \left(\bigcap_{i=0}^{\lceil \log T \rceil} \mathcal{B}_{2^i} \right) \bigcap \mathcal{A}.$$

Intuitively, $\mathcal{A}$ is the event that all samples used for few-shot property tests are $\frac{\epsilon}{4}$-representative. $\mathcal{B}_N$ is the event that every subset of tasks of size at most N satisfies the condition of a "good event" as described in (Baxter, 2000) for multi-task ERM (see Lemma 25).

We claim that the clean event happens with high probability, i.e., $\Pr(\mathcal{E}) \geq 1 - \delta$. To see this, we first examine $\Pr(\mathcal{A})$. For any $t = 2, \ldots, T$, applying Lemma 25 with one task and a singleton representation class $\{\hat{h}_t\}$—that is, $\log \mathcal{C}(\{\hat{h}_t\}, \frac{\epsilon_0}{32}) = 0$—we have

$$\Pr\left(\forall f \in \mathcal{F},\ \left| \mathcal{L}_{\mathcal{P}_t}(f \circ \hat{h}_t) - \widehat{\mathcal{L}}_{\tilde{S}_t}(f \circ \hat{h}_t) \right| \leq \frac{\epsilon}{4} \right) \geq 1 - \frac{\delta}{2T}.$$

It then follows that $\Pr(\mathcal{A}) \geq 1 - \frac{\delta}{2}$ by the union bound. Now, for each $\mathcal{B}_N$, fix any $n \leq N$ and $(t_1, \ldots, t_n)$ of $\binom{[T]}{n}$. By Lemma 25 and the definition of m_N in Theorem 24,

$$\Pr\left(\exists (h, f_1, \ldots, f_n),\ \left| \sum_{i=1}^{n} \mathcal{L}_{P_{t_i}}(f_i \circ h) - \sum_{i=1}^{n} \widehat{\mathcal{L}}_{S_{t_i}(m_N)}(f_i \circ h) \right| > \frac{\epsilon}{2} \right) \leq \frac{\delta}{4 \log T \cdot \sum_{j=1}^{\log N} \binom{T}{2^j}}.$$

It then follows by the union bound that $\Pr(\mathcal{B}_N) \geq 1 - \frac{\delta}{4 \log T}$ for every N and again by the union bound that

$$\Pr\left(\bigcap_{i=0}^{\lceil \log T \rceil} \mathcal{B}_{2^i} \right) \geq 1 - \frac{\delta}{2}.$$

Correctness of Algorithm 1. We now show that, under the clean event $\mathcal{E}$, Algorithm 1 outputs a predictor with excess risk at most ϵ for every task.

For task 1, when $\mathcal{B}_1$ happens, for any $h \in \mathcal{H}$ and $f \in \mathcal{F}$, $\left| \mathcal{L}_{\mathcal{P}_1}(f \circ h) - \widehat{\mathcal{L}}_{S_1}(f \circ h) \right| \leq \frac{\epsilon}{2}$, and so the solution to single-task ERM has excess risk $\leq \epsilon$.

For each subsequent task $t = 2, \ldots, T$, the algorithm draws an i.i.d. sample $\tilde{S}_t$ of size $\tilde{m}$ to perform a few-shot property test. When $\mathcal{A}$ happens, for any $f \in \mathcal{F}$, $\left| \mathcal{L}_{\mathcal{P}_t}(f \circ \hat{h}_t) - \widehat{\mathcal{L}}_{\tilde{S}_t}(f \circ \hat{h}_t) \right| \leq \frac{\epsilon}{4}$. Let $\tilde{f}_t$ denote the ERM solution. Then, if $\widehat{\mathcal{L}}_{\tilde{S}_t}(\tilde{f}_t \circ \hat{h}_t) \leq \kappa_t + \frac{3\epsilon}{4}$, the true risk $\mathcal{L}_{\mathcal{P}_t}(\tilde{f}_t \circ \hat{h}_t) \leq \kappa_t + \epsilon$, where we recall that κ_t is the risk of $f_t^* \circ h^*$. The algorithm can safely move on to the next task.

Otherwise, the algorithm performs multi-task ERM. Consider two cases:

1. ($n_t < N_t$) Let $t_1, \ldots, t_{n_t}$ denote the past tasks for which data are *currently* stored in memory, and let $t_{n_t+1} = t$. Given i.i.d. samples of size m_{N_t} from each of the $(n_t + 1)$ tasks, when $\mathcal{B}_{N_t}$ happens, for any $(h, f_1, \ldots, f_n)$,

$$\left| \sum_{i=1}^{n_t+1} \mathcal{L}_{P_{t_i}}(f_i \circ h) - \sum_{i=1}^{n_t+1} \widehat{\mathcal{L}}_{S_{t_i}}(f_i \circ h) \right| \leq \frac{\epsilon}{2}.$$

Let $(\hat{h}, \check{f}_1^{(t)}, \ldots, \check{f}_{n_t+1}^{(t)})$ denote the solution to multi-task ERM. It follows that

$$\begin{aligned}
&\sum_{i=1}^{n_t+1} \mathcal{L}_{P_{t_i}}(\check{f}_i^{(t)} \circ \hat{h}) - \sum_{i=1}^{n_t+1} \mathcal{L}_{P_{t_i}}(f_{t_i}^* \circ h^*) \\
&\leq \left(\sum_{i=1}^{n_t+1} \mathcal{L}_{P_{t_i}}(\check{f}_i^{(t)} \circ \hat{h}) - \sum_{i=1}^{n_t+1} \widehat{\mathcal{L}}_{S_{t_i}}(\check{f}_i^{(t)} \circ \hat{h}) \right) + \underbrace{\left(\sum_{i=1}^{n_t+1} \widehat{\mathcal{L}}_{S_{t_i}}(\check{f}_i^{(t)} \circ \hat{h}) - \sum_{i=1}^{n_t+1} \widehat{\mathcal{L}}_{S_{t_i}}(f_{t_i}^* \circ h^*) \right)}_{\leq 0} \\
&\quad + \left(\sum_{i=1}^{n_t+1} \widehat{\mathcal{L}}_{S_{t_i}}(f_{t_i}^* \circ h^*) - \sum_{i=1}^{n_t+1} \mathcal{L}_{P_{t_i}}(f_{t_i}^* \circ h^*) \right) \\
&\leq \epsilon.
\end{aligned}$$

Since each summand in $\sum_{i=1}^{n_t+1} \left(\mathcal{L}_{P_{t_i}}(\check{f}_i^{(t)} \circ \hat{h}) - \mathcal{L}_{P_{t_i}}(f_{t_i}^* \circ h^*) \right)$ is nonnegative, we have

$$\mathcal{L}_{P_t}(\check{f}_{n_t+1}^{(t)} \circ \hat{h}) - \mathcal{L}_{P_t}(f_t^* \circ h^*) \leq \epsilon.$$

2. ($n_t = N_t$) In this case, we have $n_{t+1} = 1$ and $N_{t+1} = 2N_t$. The memory is cleared. The algorithm then performs single-task ERM with an i.i.d. sample of size $m_{N_{t+1}}$. Under the event $\mathcal{B}_{N_{t+1}}$, the ERM solution has excess risk at most ϵ.

Bounding the number of times Algorithm 1 performs multi-task ERM. We now show that when $\mathcal{E}$ happens, $N_{T+1} < 2\Xi$.

Assume towards contradiction that $N_{T+1} \geq 2\Xi$. Now consider the first task t such that $N_{t+1} \geq 2\Xi$. At task t, line 11 must be reached, and we have $n_t = N_t \geq \Xi$ at the beginning of task t. For this task, we have

$$\min_{f \in \mathcal{F}} \widehat{\mathcal{L}}_{\tilde{S}_t}(f \circ \hat{h}_t) > \kappa_t + \frac{3}{4}\epsilon \implies \mathcal{E}(\hat{h}_t, \mathcal{P}_t) = \min_{f \in \mathcal{F}} \mathcal{L}_{\mathcal{P}_t}(f \circ \hat{h}_t) - \mathcal{L}_{\mathcal{P}_t}(f_t^* \circ h^*) > \frac{1}{2}\epsilon.$$

Abbreviate n_t as n. In addition, we have that for the tasks $t_1, \ldots, t_n$ in the current memory, $(\hat{h}_t, \check{f}_1^{(t_n)}, \ldots, \check{f}_n^{(t_n)})$ satisfies

$$\sum_{i=1}^{n} \mathcal{L}_{\mathcal{P}_{t_i}}(\check{f}_i^{(t_n)} \circ \hat{h}_t) - \sum_{i=1}^{n} \mathcal{L}_{\mathcal{P}_{t_i}}(f_{t_i}^* \circ h^*) \leq \epsilon,$$

which in turn implies that

$$\sum_{i=1}^{n} \mathcal{E}(\hat{h}_t, \mathcal{P}_{t_i}) \leq \epsilon.$$

This implies that f_t^* is ϵ-independent of $\{f_{t_i}^*\}_{i=1}^n$ with respect to $(\mathcal{H}, \mathcal{F})$ and h^*. This argument can be extended for each task in the current memory—each $f_{t_i}^*$ must be ϵ-independent of its predecessors. Since the task-eluder dimension is bounded by Ξ, this implies that $n_t + 1 \leq \Xi$, which contradicts with the assumption that $n_t \geq \Xi$. Hence, we have $N_{T+1} < 2\Xi$.

Sample and space complexity. We have shown that $N_{T+1} < 2\Xi$. Let $b = \log_2 N_{T+1}$. The total number of samples used by Algorithm 1 is upper bounded by

$$(\star) := \tilde{m}T + \sum_{i=0}^{b} 2^i \cdot m_{2^i},$$

where the first term is from few-shot property testing, and the second term is from samples collected and saved in memory for multi-task ERM. For each value of $N = 2^i$, $i = 0, 1, \ldots, b$, samples are drawn from at most N tasks with m_N samples per task. We focus on the second term:

$$
\begin{aligned}
&\sum_{i=0}^{b} 2^i \cdot m_{2^i} \\
&= \sum_{i=0}^{b} 2^i \left(\frac{256 \cdot 2^i}{\epsilon^2} \left[\log \mathcal{C}\left(\mathcal{H}, \frac{\epsilon}{64 \cdot 2^i}\right) + 2^i \log \mathcal{C}\left(\mathcal{F}, \frac{\epsilon}{64 \cdot 2^i}\right) + \log \frac{16 \log T \cdot \sum_{j=0}^{i} \binom{T}{2^j}}{\delta} \right] + \frac{64}{\epsilon^2} \right) \\
&\overset{(a)}{\leq} \sum_{i=0}^{b} 2^i \left(\frac{256 \cdot 2^i}{\epsilon^2} \left[\log \mathcal{C}\left(\mathcal{H}, \frac{\epsilon}{64 \cdot 2^i}\right) + 2^i \left[\log \mathcal{C}\left(\mathcal{F}, \frac{\epsilon}{64 \cdot 2^i}\right) + \log(eT) \right] + \log \frac{16 \log T}{\delta} \right] + \frac{64}{\epsilon^2} \right) \\
&\overset{(b)}{\leq} \sum_{i=0}^{b} 4^i \cdot \frac{256}{\epsilon^2} \left[\log \mathcal{C}\left(\mathcal{H}, \frac{\epsilon}{64 N_{T+1}}\right) + \log \frac{16 \log T}{\delta} \right] \\
&\qquad\qquad + \sum_{i=0}^{b} 8^i \cdot \frac{256}{\epsilon^2} \left[\log \mathcal{C}\left(\mathcal{F}, \frac{\epsilon}{64 N_{T+1}}\right) + \log(eT) \right] + \sum_{i=0}^{b} 2^i \cdot \frac{64}{\epsilon^2} \\
&\overset{(c)}{\lesssim} \frac{N_{T+1}^2}{\epsilon^2} \left[\log \mathcal{C}\left(\mathcal{H}, \frac{\epsilon}{64 N_{T+1}}\right) + N_{T+1} \left[\log \mathcal{C}\left(\mathcal{F}, \frac{\epsilon}{64 N_{T+1}}\right) + \log(eT) \right] + \log \frac{16 \log T}{\delta} \right] + \frac{N_{T+1}}{\epsilon^2},
\end{aligned}
$$

where (a) follows because

$$
\log \left(\sum_{j=0}^{i} \binom{T}{2^j} \right) \leq 2^i \log \left(\frac{eT}{2^i} \right) \leq 2^i \log (eT) ;
$$

(b) follows by algebra and the observation that covering numbers increase as scale decreases; and (c) uses the following inequality for bounding the sum of a finite geometric series: $\sum_{i=0}^{b} r^i \leq r^{b+1}$ for $r \geq 2$.

Since $N_{T+1} < 2\Xi$, we have

$$
(\star) \leq \tilde{\mathcal{O}} \left(\frac{T}{\epsilon^2} \log \mathcal{C}\left(\mathcal{F}, \frac{\epsilon}{128}\right) + \frac{\Xi^2}{\epsilon^2} \left[\log \mathcal{C}\left(\mathcal{H}, \frac{\epsilon}{128\Xi}\right) + \Xi \log \mathcal{C}\left(\mathcal{F}, \frac{\epsilon}{128\Xi}\right) \right] \right).
$$

Similarly, the size of the memory needed is non-decreasing. Since $N_{T+1} < 2\Xi$, the space complexity is upper bounded by

$$
N_{T+1} \cdot m_{N_{T+1}} \leq \tilde{\mathcal{O}} \left(\frac{\Xi^2}{\epsilon^2} \left[\log \mathcal{C}\left(\mathcal{H}, \frac{\epsilon}{128\Xi}\right) + \Xi \log \mathcal{C}\left(\mathcal{F}, \frac{\epsilon}{128\Xi}\right) \right] \right).
$$

■

B.2. Auxiliary lemma

Here, we provide a more precise version of the guarantee for multi-task ERM from (Baxter, 2000) with explicit constants. Lemma 25 follows from Corollary 19 and Theorem 6 thereof.

Lemma 25 (Baxter, 2000) *Let $P_1, \ldots, P_n$ be the data distributions for n tasks. Let $\mathcal{H}_0$ be a class of representations and $\mathcal{F}_0$ be a class of prediction layers. Let $\epsilon_0, \delta_0 \in (0, 1)$. Suppose for each task,*

an i.i.d. sample S_i of size m is drawn from P_i^m, where

$$m \geq \frac{64}{n\epsilon_0^2}\left[\log \mathcal{C}(\mathcal{H}_0, \frac{\epsilon_0}{32}) + n \log \mathcal{C}(\mathcal{F}_0, \frac{\epsilon_0}{32}) + \log \frac{4}{\delta_0}\right] + \frac{16}{\epsilon_0^2}.$$

Then, with probability at least $1 - \delta_0$, for any $(h, f_1, \ldots, f_n)$,

$$\left|\frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_{P_i}(f_i \circ h) - \frac{1}{n}\sum_{i=1}^{n} \widehat{\mathcal{L}}_{S_i}(f_i \circ h)\right| \leq \epsilon_0.$$

Appendix C. Supplementary material for Section 5

We first present a key lemma for bounding the task-eluder dimension and then provide proofs for the examples in Section 5. Auxiliary lemmas are deferred to Appendix C.5. Throughout this section, unless specified otherwise, all norms are ℓ_2 norms.

C.1. Key lemma for bounding the task-eluder dimension

Lemma 26 *Let $\mathcal{X} \subset \mathbb{R}^d$ and P_X be a distribution over $\mathcal{X}$. Let $\mathcal{Y} \subseteq \mathbb{R}$. Consider $\mathcal{H} = \{x \mapsto B^\top x : B \in \mathbb{R}^{d\times k}, B^\top B = I_k\}$ and $\mathcal{F} = \left\{z \mapsto g(w^\top z) : w \in \mathbb{R}^k, \underline{b} \le \|w\| \le \bar{b}\right\}$, where $0 \le \underline{b} \le \bar{b} \le 1$ and g is a (possibly nonlinear) map from $\mathbb{R}$ to $\mathbb{R}$.*

For any $h \in \mathcal{H}$ and $f \in \mathcal{F}$, let $P_{f\circ h}(x, y) = P_X(x)P_{Y|X}(y|x; f\circ h)$. Let $\mathbb{P} = \{P_{f\circ h} : h \in \mathcal{H}, f \in \mathcal{F}\}$, and $\ell : \mathbb{R} \times \mathcal{Y} \to [0, 1]$ be a loss function.

Fix $p \in [1, 2]$. For any representations $h, h' \in \mathcal{H}$, let $B, B' \in \mathbb{R}^{d\times k}$ be the corresponding matrices. Similarly, for any $f, f' \in \mathcal{F}$, let $w, w' \in \mathbb{R}^k$ be the corresponding vectors. Suppose

$$\left\|B'w' - Bw\right\|^p \lesssim \mathbb{E}_{P_{f\circ h}}\left[\ell((f' \circ h')(x), y) - \ell((f \circ h)(x), y)\right] \lesssim \left\|B'w' - Bw\right\|^p. \tag{15}$$

Then, for any $\epsilon \in (0, 1)$,

$$\mathrm{TEdim}_{\mathbb{P},\ell}(\mathcal{H}, \mathcal{F}, \epsilon) \lesssim k \log \frac{1}{\epsilon}.$$

Proof idea. At a high-level, our proof of Lemma 26 involves three steps. Consider any ϵ-independent sequence of tasks $\{(B, w_i)\}$:

1. First, we convert task independence to geometric independence of the task parameters, i.e., there exists B' such that $\sum_i \|P^\perp_{B'} B w_i\|$ is small but $\|P^\perp_{B'} B w_n\|$ is large.
2. Second, we use a useful identity in linear regression (Eq. (16)) to convert geometric independence to linear independence of the task prediction layers, i.e., $\|w_n\|^{-1}_{V_{n-1}} \gtrsim 1$, where $V_n = \sum_{i=1}^n w_i w_i^\top + \epsilon I$.
3. Finally, we apply the elliptical potential lemma (e.g., (Abbasi-Yadkori et al., 2011)) to argue that any sequence of independent task prediction layers can not be too long.

We note that our technique is similar in spirit to an argument in the manuscript (Yang et al., 2022) for a lifelong linear bandit algorithm. Our result is algorithm-independent and proved in a different framework.

Proof [Proof of Lemma 26] Let $\{f_i\}_{i=1}^\tau$ be any sequence of tasks such that, for each $n \le \tau$, f_n is ϵ-independent of $\{f_i\}_{i=1}^{n-1}$ with respect to $(\mathcal{H}, \mathcal{F})$ and h. When it is clear from the context, we will omit the dependence on $(\mathcal{H}, \mathcal{F})$ and h hereafter.

It suffices to show that $\tau \lesssim k \log \frac{1}{\epsilon}$. To this end, let $B \in \mathbb{R}^{d\times k}$ be the semi-orthogonal matrix associated with h, and for each $i \in [\tau]$, let $w_i \in \mathbb{R}^k$ be the vector associated with f_i. For $n \in [\tau]$, let $W_n \in \mathbb{R}^{k\times n}$ denote the matrix whose columns are $w_1, \ldots, w_n$, and let $V_n := W_n W_n^\top + \epsilon^{\frac{2}{p}} I$.

To conclude the proof, it suffices to show that for each $n \in \{2, \ldots, \tau\}$, $\|w_n\|^2_{V^{-1}_{n-1}} \gtrsim \frac{1}{4}$. Indeed, by the elliptical potential lemma (Abbasi-Yadkori et al., 2011, see also Lemma 29),

$$\sum_{n=1}^{\tau} \min\left\{1, \|w_n\|^2_{V^{-1}_{n-1}}\right\} \le 2k \log\left(1 + \frac{\tau}{k\epsilon^2}\right),$$

and it follows from Lemma 30 (see also Lattimore and Szepesvári, 2020) that $\tau \lesssim k \log \frac{1}{\epsilon}$.

For $n \in \{2, \ldots, \tau\}$, we now show $\|w_n\|^2_{V_{n-1}^{-1}} \gtrsim \frac{1}{4}$. Since f_n is ϵ-independent of $\{f_i\}_{i=1}^{n-1}$, there exist h' such that

$$\sum_{i=1}^{n-1} \mathcal{E}(h', P_{f_i \circ h}) \leq \epsilon,$$

but $\mathcal{E}(h', P_{f_n \circ h}) > \frac{\epsilon}{2}$. Fix any such h', and let B' denote the matrix and vectors associated with it.

Let $\alpha^\star := \operatorname{argmin}_{\alpha \in \mathbb{R}^{n-1}} \epsilon^{-\frac{2}{p}} \|w_n - W_{n-1}\alpha\|_2^2 + \|\alpha\|_2^2$. By Lemma 31, we have

$$w_n^\top V_{n-1}^{-1} w_n = \epsilon^{-\frac{2}{p}} \|w_n - W_{n-1}\alpha^\star\|_2^2 + \|\alpha^\star\|_2^2 . \tag{16}$$

Consider the decomposition $w_n = W_{n-1}\alpha^\star + z$. We have

$$\epsilon^{\frac{1}{p}} \overset{\text{(a)}}{\lesssim} \left\|P_{B'}^\perp B w_n\right\| \overset{\text{(b)}}{\leq} \left\|P_{B'}^\perp B W_{n-1}\alpha^\star\right\| + \left\|P_{B'}^\perp B z\right\| \overset{\text{(c)}}{\lesssim} \epsilon^{\frac{1}{p}} \|\alpha^\star\| + \|z\| , \tag{17}$$

where $P_{B'}^\perp$ denotes the orthogonal projection onto the orthogonal complement of $\operatorname{span}(B')$, and the inequalities are justified shortly. Dividing $\epsilon^{\frac{1}{p}} > 0$ on both sides, it follows that either $\|\alpha^\star\| \gtrsim \frac{1}{2}$ or $\epsilon^{-\frac{1}{p}} \|z\| \gtrsim \frac{1}{2}$. Therefore, by Eq. (16),

$$\|w_n\|^2_{V_{n-1}^{-1}} = \epsilon^{-\frac{2}{p}} \|z\|^2 + \|\alpha^\star\|^2 \gtrsim \frac{1}{4}.$$

To complete the proof, we justify the inequalities in Eq. (17) above:

(a) By Eq. (15) and that $\mathcal{E}(h', P_{f_n \circ h}) > \frac{\epsilon}{2}$, for any f_n' with corresponding vector w_n', we have

$$\left\|B' w_n' - B w_n\right\|^p \gtrsim \mathbb{E}_{(x,y)\sim P_{f_n \circ h}} \left[\ell((f_n' \circ h')(x), y) - \ell((f_n \circ h)(x), y)\right] > \frac{\epsilon}{2};$$

that is,

$$\min_{w_n' : \|w_n'\| \in [\underline{b}, \overline{b}]} \left\|B' w_n' - B w_n\right\| \gtrsim \epsilon^{\frac{1}{p}}.$$

It then follows from Lemma 32 that

$$\left\|P_{B'}^\perp B w_n\right\| \gtrsim \min_{w_n' : \|w_n'\| \in [\underline{b}, \overline{b}]} \left\|B' w_n' - B w_n\right\| \gtrsim \epsilon^{\frac{1}{p}}.$$

(b) uses the triangle inequality.

(c) Since $\sum_{i=1}^{n-1} \mathcal{E}(h', P_{f_i \circ h}) \leq \epsilon$, there exist $f_1', \ldots, f_n' \in \mathcal{F}$ with associated parameters $w_1', \ldots, w_n'$, such that

$$\sum_{i=1}^{n-1} \mathbb{E}_{(x,y)\sim P_{f_i \circ h}} \left[\ell\left((f_i' \circ h')(x), y\right) - \ell\left((f_i \circ h)(x), y\right)\right] \leq \epsilon.$$

Therefore, by Eq. (15), there exist parameters $w_1', \ldots, w_n'$

$$\sum_{i=1}^{n-1} \left\|B' w_i' - B w_i\right\|^p \lesssim \epsilon.$$

By the fact that $\left\|P_{B'}^\perp B w_i\right\| = \min_{w' \in \mathbb{R}^k} \|B' w' - B w_i\|$, we have $\sum_{i=1}^{n-1} \left\|P_{B'}^\perp B w_i\right\|^p \lesssim \epsilon$.

Using Hölder's inequality we have

$$\begin{aligned}\left\|P_{B'}^{\perp}BW_{n-1}\alpha^{\star}\right\| &\leq \sum_{i=1}^{n-1}|\alpha_i^{\star}|\left\|P_{B'}^{\perp}Bw_i\right\| \\ &\leq \left(\sum_{i=1}^{n-1}\left\|P_{B'}^{\perp}Bw_i\right\|^p\right)^{1/p}\left(\sum_{i=1}^{n-1}|\alpha^{\star}|^q\right)^{1/q} \\ &\lesssim \epsilon^{\frac{1}{p}}\left\|\alpha^{\star}\right\|_q \leq \epsilon^{\frac{1}{p}}\left\|\alpha^{\star}\right\|_2,\end{aligned}$$

where $q \in [2,\infty)$ satisfies $\frac{1}{p}+\frac{1}{q}=1$.

■

C.2. Noisy linear regression

We first revisit the setting and restate Proposition 15.

Setting. Recall that $\mathcal{X} = \left\{x \in \mathbb{R}^d : \|x\| \leq \sqrt{d}\right\}$ and $\mathcal{Y} = [-1,1]$. We consider $\mathcal{H} = \{x \mapsto B^\top x : B \in \mathbb{R}^{d\times k}, B^\top B = I\}$ and $\mathcal{F}^{\text{lin}} = \left\{z \mapsto z^\top w : w \in \mathbb{R}^k, \|w\| \leq \frac{1}{2\sqrt{d}}\right\}$. The probabilistic model $\mathbb{P} = \{P_{f\circ h} : h \in \mathcal{H}, f \in \mathcal{F}^{\text{lin}}\}$ is defined as follows: P_X satisfies $I \precsim \mathbb{E}_{x\sim P_X}[xx^\top] \precsim I$. For each h and f, given an input $x \sim P_X$, $y = (f\circ h)(x) + \eta$, where η is independent noise from a distribution that has zero mean, support $[-\frac{1}{2},\frac{1}{2}]$, and variance κ. The noise distribution is common to all $f \circ h$'s. There exist h^* and $f_1^*,\ldots,f_T^*$ such that $\mathcal{P}_t = P_{f_t^*\circ h^*}$, and let B^* and $w_1^*,\ldots,w_T^*$ be the matrix and vectors associated with these functions. Let $\ell(y',y) = \frac{1}{4}(y'-y)^2$ be the loss function.

Proposition 15. *Let $\epsilon \in (0,1)$. We have*

$$\log \mathcal{C}(\mathcal{F}_\ell^{\text{lin}},\epsilon) \leq \mathcal{O}\Big(k\log\frac{1}{\epsilon}\Big),\ \log \mathcal{C}_{\mathcal{F}_\ell^{\text{lin}}}(\mathcal{H},\epsilon) \leq \mathcal{O}\Big(dk\log\frac{1}{\epsilon}\Big),\ \text{TEdim}_{\mathbb{P},\ell}(\mathcal{H},\mathcal{F}^{\text{lin}},\epsilon) \leq \mathcal{O}\big(k\log\frac{1}{\epsilon}\big).$$

Proof [Proof of Proposition 15] We prove the three statements separately.

1. **(Capacity of $\mathcal{F}_\ell^{\text{lin}}$)** Note that each representation h maps $\mathcal{X}$ to $\mathcal{Z} := \left\{z \in \mathbb{R}^k : \|z\| \leq \sqrt{d}\right\}$, and each prediction layer $f \in \mathcal{F}^{\text{lin}}$ maps $\mathcal{Z}$ to $[-\frac{1}{2},\frac{1}{2}]$. Recall the definitions from Section 4.1. We have

$$\mathcal{C}(\mathcal{F}_\ell^{\text{lin}},\epsilon) = \sup_Q N(\mathcal{F}^{\text{lin}},\epsilon,d_Q),$$

where $d_Q(f_\ell,f_\ell') = \int_{\mathcal{Z}\times\mathcal{Y}}|\ell(f(z),y)-\ell(f'(z),y)|\,dQ(z,y)$ for any measure Q on $\mathcal{Z}\times\mathcal{Y}$.

We follow the technique from (Haussler, 1992; Baxter, 2000). For any Q,

$$\begin{aligned}d_Q(f_\ell,f_\ell') &= \frac{1}{4}\int_{\mathcal{Z}\times\mathcal{Y}}\left|(f(z)-y)^2-(f'(z)-y)^2\right|dQ(z,y) \\ &= \frac{1}{4}\int_{\mathcal{Z}\times\mathcal{Y}}\left|(f(z)-f'(z))(f(z)+f'(z)-2y)\right|dQ(z,y) \\ &\leq \int_{\mathcal{Z}}\left|(f(z)-f'(z))\right|dQ_Z(z) =: L^1(Q),\end{aligned}$$

where Q_Z is the marginal distribution derived from Q and the inequality uses the observation that $|f(z) + f'(z) - 2y| \leq 3$. It follows that

$$\sup_Q N(\mathcal{F}_\ell^{\text{lin}}, \epsilon, d_Q) \leq \sup_Q N(\mathcal{F}^{\text{lin}}, \epsilon, L^1(Q)) \leq \left(\frac{2e}{\epsilon}\right)^{2k},$$

where the second inequality uses (Haussler, 1992, Theorem 11). We then have $\log \mathcal{C}(\mathcal{F}_\ell^{\text{lin}}, \epsilon) \leq \mathcal{O}\left(k \log \frac{1}{\epsilon}\right)$.

2. **(Capacity of $\mathcal{H}$)** For any measure P on $\mathcal{X} \times \mathcal{Y}$, recall that

$$d_{P,\mathcal{F}_\ell^{\text{lin}}}(h, h') = \int_{\mathcal{X}\times\mathcal{Y}} \sup_{f_\ell \in \mathcal{F}_\ell^{\text{lin}}} \left|f_\ell(h(x)), y) - f_\ell(h'(x)), y)\right| dP(x, y).$$

For any $f \in \mathcal{F}^{\text{lin}}$, let w be the vector associated with f. We have

$$\begin{aligned}
\left|f_\ell(h(x)), y) - f_\ell(h'(x)), y)\right| &= \frac{1}{4}\left|(f(h(x)) - y)^2 - (f(h'(x)) - y)^2\right| \\
&= \frac{1}{4}\left|(f(h(x)) - f(h'(x)))(f(h(x)) + f(h'(x)) - 2y)\right| \\
&\leq \frac{3}{4}\left|\langle w, h(x) - h'(x)\rangle\right| \\
&\leq \left\|h(x) - h'(x)\right\|.
\end{aligned}$$

Therefore,

$$d_{P,\mathcal{F}_\ell^{\text{lin}}}(h, h') \leq \int_{\mathcal{X}} \left\|h(x) - h'(x)\right\| dP_X(x) =: L^1(P),$$

where P_X is the marginal distribution derived from P. By (Haussler, 1992, Theorem 11 therein), we then have

$$\mathcal{C}_{\mathcal{F}_\ell^{\text{lin}}}(\mathcal{H}, \epsilon) := \sup_P N(\mathcal{H}, \epsilon, d_{P,\mathcal{F}_\ell^{\text{lin}}}) \leq \sup_P N(\mathcal{H}, \epsilon, L^1(P)) \leq \left(\frac{2e}{\epsilon}\right)^{2dk},$$

which completes the proof.

3. **(Task-eluder dimension)** Observe that for any $h, h' \in \mathcal{H}$ and $f, f' \in \mathcal{F}^{\text{lin}}$ with corresponding parameters B, B' and w, w',

$$\begin{aligned}
&\mathbb{E}_{P_{f\circ h}}\left[\ell((f' \circ h')(x), y) - \ell((f \circ h)(x), y)\right] \\
&\overset{\text{(a)}}{=} \mathbb{E}_{x,\eta}\left[\left(x^\top B' w' - x^\top B w - \eta\right)^2 - \eta^2\right] \\
&\overset{\text{(b)}}{=} \left(B'w' - Bw\right) \mathbb{E}_x[xx^\top] \left(B'w' - Bw\right),
\end{aligned}$$

where (a) follows because under $P_{f\circ h}$, $y = (f \circ h)(x) + \eta = x^\top B w + \eta$, and (b) follows because η is independent of x and $\mathbb{E}[\eta] = 0$. Since $I \precsim \mathbb{E}_{x\sim P_X}[xx^\top] \precsim I$,

$$\left\|B'w' - Bw\right\|^2 \lesssim \mathbb{E}_{P_{f\circ h}}\left[\ell((f' \circ h')(x), y) - \ell((f \circ h)(x), y)\right] \lesssim \left\|B'w' - Bw\right\|^2.$$

Applying Lemma 26 with $g(v) = v$ and $p = 2$, we have

$$\text{TEdim}_{\mathbb{P},\ell}(\mathcal{H}, \mathcal{F}^{\text{lin}}, \epsilon) \lesssim k \log \frac{1}{\epsilon}.$$

■

C.3. Classification with logistic regression

We first restate the setting and Proposition 16.

Setting. Recall that $\mathcal{X} = \left\{x \in \mathbb{R}^d : \|x\| \leq \sqrt{d}\right\}$ and $\mathcal{Y} = \{0, 1\}$. We consider $\mathcal{H} = \{x \mapsto B^\top x : B \in \mathbb{R}^{d\times k}, B^\top B = I\}$ and $\mathcal{F}^{\log} = \left\{z \mapsto \sigma(z^\top w) : w \in \mathbb{R}^k, \|w\| \leq \frac{1}{4\sqrt{d}}\right\}$, where $\sigma(v) = 1/\left(1 + e^{-v}\right)$ is the logistic sigmoid function. The data distributions in $\mathbb{P} = \left\{P_{f\circ h} : h \in \mathcal{H}, f \in \mathcal{F}^{\log}\right\}$ are defined as follows: P_X satisfies $I \precsim \mathbb{E}_{x\sim P_X}[xx^\top] \precsim I$, and for each h and f, $\Pr(y = 1|x; f \circ h) = (f \circ h)(x)$ and $\Pr(y = 0|x; f \circ h) = 1 - (f \circ h)(x)$. There exist h^* and $f_1^*, \ldots, f_T^*$ such that $\mathcal{P}_t = P_{f_t^*\circ h^*}$, and let B^* and $w_1^*, \ldots, w_T^*$ be the associated parameters. Let $\ell(y', y) = -y \log y' - (1 - y) \log(1 - y')$ be the loss function.

Proposition 16. *Let $\epsilon \in (0, 1)$. We have*

$$\log \mathcal{C}(\mathcal{F}_\ell^{\log}, \epsilon) \leq \mathcal{O}\Big(k \log \frac{1}{\epsilon}\Big),\ \log \mathcal{C}_{\mathcal{F}_\ell^{\log}}(\mathcal{H}, \epsilon) \leq \mathcal{O}\Big(dk \log \frac{1}{\epsilon}\Big),\ \mathrm{TEdim}_{\mathbb{P},\ell}(\mathcal{H}, \mathcal{F}^{\log}, \epsilon) \leq \mathcal{O}\big(k \log \frac{1}{\epsilon}\big).$$

Proof [Proof of Proposition 16] Again, we prove the three statements separately.

1. **(Capacity of $\mathcal{F}^{\log}$)** As in the proof of Proposition 15, we use the technique from (Haussler, 1992; Baxter, 2000). First observe that each h is a mapping from $\mathcal{X}$ to $\mathcal{Z} := \left\{z \in \mathbb{R}^k : \|z\| \leq \sqrt{d}\right\}$, and each $f \in \mathcal{F}^{\log}$ is a mapping from $\mathcal{Z}$ to $[\sigma(-\frac{1}{4}), \sigma(\frac{1}{4})]$. For any probability measure Q on $\mathcal{Z} \times \mathcal{Y}$, we have

$$\begin{aligned} d_Q(f_\ell, f'_\ell) &= \int_{\mathcal{Z}\times\mathcal{Y}} \left|\ell(f(z), y) - \ell(f'(z), y)\right| dQ(z, y) \\ &\leq 3 \underbrace{\int_{\mathcal{Z}} \left|f(z) - f'(z)\right| dQ_Z(z)}_{=:L^1(Q)}, \end{aligned}$$

   where Q_Z is the marginal distribution derived from Q, and the inequality follows because ℓ is 3-Lipschitz continuous with respect to its first argument over the domain. It follows that

$$N(\mathcal{F}_\ell^{\log}, \epsilon, d_Q) \leq N(\mathcal{F}^{\log}, \frac{\epsilon}{3}, L^1(Q)),$$

   and therefore,

$$\mathcal{C}(\mathcal{F}_\ell^{\log}, \epsilon) := \sup_Q N(\mathcal{F}_\ell^{\log}, \epsilon, d_Q) \leq \sup_Q N(\mathcal{F}^{\log}, \frac{\epsilon}{3}, L^1(Q)) \leq \left(\frac{6e}{\epsilon}\right)^{2k},$$

   where the last inequality uses (Haussler, 1992, Theorem 11).

2. **(Capacity of $\mathcal{H}$)** For any probability measure P on $\mathcal{X} \times \mathcal{Y}$, consider

$$d_{P,\mathcal{F}_\ell^{\log}}(h, h') = \int_{\mathcal{X}\times\mathcal{Y}} \sup_{f_\ell \in \mathcal{F}_\ell^{\log}} \left|\ell((f \circ h)(x), y) - \ell((f \circ h')(x), y)\right| dP(x, y).$$

For any $f \in \mathcal{F}^{\log}$, $h, h' \in \mathcal{H}$, $x \in \mathcal{X}$, and $y \in \mathcal{Y}$, since ℓ is 3-Lipschitz continuous with respect to the first argument over the domain,

$$\left|\ell((f \circ h)(x), y) - \ell((f \circ h')(x), y)\right| \leq 3 \left|(f \circ h)(x) - (f \circ h')(x)\right|.$$

Let w, B, and B' denote the parameters associated with f, h, and h', respectively. It follows that

$$\begin{aligned}\left|(f \circ h)(x) - (f \circ h')(x)\right| &= \left|\sigma\left(w^\top h(x)\right) - \sigma\left(w^\top h'(x)\right)\right| \\ &\leq \frac{1}{4}\left|w^\top (h(x) - h'(x))\right| \\ &\leq \frac{1}{16}\left\|h(x) - h'(x)\right\|,\end{aligned}$$

where the second inequality uses the observation that $\sigma(\cdot)$ is $\frac{1}{4}$-Lipschitz continuous, and the last inequality follows because $\|w\| \leq \frac{1}{4}$. Therefore,

$$d_{P,\mathcal{F}_\ell^{\log}}(h, h') \leq \int_{\mathcal{X}} \left\|h(x) - h'(x)\right\| dP_X(x) =: L^1(P),$$

where P_X is the marginal distribution derived from P. It then follows from (Haussler, 1992, Theorem 11) that

$$\mathcal{C}_{\mathcal{F}_\ell^{\log}}(\mathcal{H}, \epsilon) := \sup_P N(\mathcal{H}, \epsilon, d_{P,\mathcal{F}_\ell^{\log}}) \leq \sup_P N(\mathcal{H}, \epsilon, L^1(P)) \leq \left(\frac{2e}{\epsilon}\right)^{2dk},$$

which completes the proof.

3. **(Task-eluder dimension)** Fix any $h \in \mathcal{H}$ and $f \in \mathcal{F}^{\log}$. Let B and w denote the associated matrix and vector, respectively.

   For any $\theta \in \mathbb{R}^d$ such that $\|\theta\| \leq \frac{1}{4\sqrt{d}}$, consider

   $$L(\theta) := \mathbb{E}_{(x,y)\sim P_{f\circ h}}[\ell(\sigma(x^\top \theta), y)].$$

   It is easy to verify that

   $$\nabla_\theta^2 L(\theta) = \mathbb{E}_{x\sim P_X}[\sigma'(x^\top \theta) x x^\top].$$

   For the domain $\left|x^\top \theta\right| \leq \frac{1}{4}$, $1 \lesssim \sigma'(x^\top \theta) \lesssim 1$. Since $I \precsim \mathbb{E}_{x\sim P_X}[xx^\top] \precsim I$,

   $$I \precsim \nabla_\theta^2 L(\theta) \precsim I;$$

   that is, $L(\theta)$ is *locally* strongly convex and smooth. Therefore, for any h' and f' with corresponding parameters B' and w',

   $$\left\|B'w' - Bw\right\|^2 \lesssim \underbrace{L(B'w') - L(Bw)}_{=\mathbb{E}_{P_{f\circ h}}[\ell((f'\circ h')(x),y) - \ell((f\circ h)(x),y)]} \lesssim \left\|B'w' - Bw\right\|^2.$$

   Applying Lemma 26 with $g = \sigma$ and $p = 2$, we have

   $$\mathrm{TEdim}_{\mathbb{P},\ell}\left(\mathcal{H}, \mathcal{F}^{\log}, \epsilon\right) \lesssim k \log \frac{1}{\epsilon}.$$

■

C.4. Classification with random classification noise and the 0-1 loss

We now discuss how our results may also be applied to binary classification with random classification noise under the 0-1 loss.

Setting. Let $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \{-1, 1\}$. We again consider low-dimensional linear representations, $\mathcal{H} = \left\{x \mapsto B^\top x, B \in \mathbb{R}^{d\times k}, B^\top B = I\right\}$. Let $\mathcal{F}^{\text{cls}} = \{z \mapsto \text{sign}\left(\langle w, z\rangle\right) : w \in \mathbb{R}^k, \|w\| = 1\}$ be a class of linear threshold functions. Consider the probabilistic model $\mathbb{P}$: let P_X be isotropic log-concave (e.g., normal distribution and uniform distribution over $\mathbb{S}^d(0, \sqrt{d})$). We consider random classification noise (e.g., Kearns and Vazirani, 1994). For any h and f, given an input $x \sim P_X$, $y = \text{sign}((f \circ h)(x))$ with probability $1 - \eta$, and $y = -\text{sign}((f \circ h)(x))$ with probability η, where $\eta \in [0, \frac{4}{10}]$ is the noise rate. There exist B^* and $w_1^*, \ldots, w_T^*$ such that each task t is well specified by B^* and w_t^*. Let $\ell(y', y) = \mathbb{1}\{y' \neq y\}$ denote the 0-1 loss.

Proposition 27 *Let $\epsilon \in (0, 1)$. We have*

$$\text{TEdim}_{\mathbb{P},\ell}(\mathcal{H}, \mathcal{F}^{\text{cls}}, \epsilon) \leq \mathcal{O}\big(k \log \frac{1}{\epsilon}\big).$$

Proof [Proof of Proposition 27] Since P_X is isotropic and log-concave in $\mathcal{X}$, it follows from (Balcan et al., 2015, Lemma 1 therein) that for any two unit vectors u and v,

$$\|u - v\| \lesssim \Pr_{x\sim P}\left(\text{sign}\left(u^\top x\right) \neq \text{sign}\left(v^\top x\right)\right) \lesssim \|u - v\|,$$

where we also use the fact that the Euclidean distance between two unit vectors is equivalent to the angle between them up to constant factors.

In addition, under random classification noise, for any h, h' and f, f',

$$\begin{aligned}(1 - 2\eta)\Pr((f' \circ h')(x) \neq (f \circ h)(x)) &\leq \mathbb{E}_{P_{f\circ h}}\left[\ell((f' \circ h')(x), y) - \ell((f \circ h)(x), y)\right] \\ &\leq \Pr((f' \circ h')(x) \neq (f \circ h)(x)).\end{aligned}$$

See (Balcan and Haghtalab, 2020, Section 5.1) for a reference.

For any $h, h' \in \mathcal{H}$ and $f, f' \in \mathcal{F}^{\text{cls}}$, let B, B' and w, w' denote the corresponding parameters, respectively. Since $\eta \in [0, \frac{4}{10}]$, it follows that

$$\left\|B'w' - Bw\right\| \lesssim \mathbb{E}_{P_{f\circ h}}\left[\ell((f' \circ h')(x), y) - \ell((f \circ h)(x), y)\right] \lesssim \left\|B'w' - Bw\right\|.$$

The proof is completed by applying Lemma 26 with $g(v) = \text{sign}(v)$, $\underline{b} = \overline{b} = 1$, and $p = 1$. ■

Remark 28 *Based on Proposition 27, one may derive a sample complexity guarantee for the classification example above. We note that Theorem 14 leverages sample complexity guarantees for multi-task ERM based on covering numbers (Baxter, 2000), which may no longer be suitable under the discrete 0-1 loss. Nevertheless, recent work by Aliakbarpour et al. (2024) establishes multi-task ERM guarantees based on the VC dimension, offering a path towards proving a result analogous to Theorem 14. We defer a detailed exploration of this direction to future work.*

C.5. Auxiliary lemmas

We now present the lemmas used above in the proof of Proposition 27.

Lemma 29 (Abbasi-Yadkori et al., 2011, Lemma 11) *Let $\{x_t\}_t$ be a sequence of vectors in $\mathbb{R}^d$ such that $\|x_t\| \le L$ for all t. Let $V_t = \lambda I + \sum_{s=1}^t x_s x_s^\top$. Then,*

$$\sum_{t=1}^T \min\left\{1, \|x_t\|^2_{V_{t-1}^{-1}}\right\} \le 2d \log\left(1 + \frac{L^2 T}{d\lambda}\right).$$

Lemma 30 (Lattimore and Szepesvári, 2020, Exercise 19.3) *Let $\epsilon \in (0,1)$. Suppose $z \le ck \log(1 + \frac{z}{k\epsilon^2})$ for some constant $c > 1$. Then,*

$$z \le 12c \cdot k \log \frac{1}{\epsilon}.$$

Proof This proof is due to (Lattimore and Szepesvári, 2020) for a closely related result. We provide it here with elaborated details for clarity and completeness.

Let $x = ck > 0$ and $y = \frac{1}{k\epsilon^2} > 0$. In addition, let $g(z) = z - x\log(1 + yz)$, and $z_0 = 3x\log(1 + xy)$. It suffices to show that $g(z) \le 0$ implies $z \le z_0$. The rest follows by algebra: $\log\left(1 + \frac{c}{\epsilon^2}\right) \le 4 \log \frac{1}{\epsilon}$ for $\epsilon \in (0,1)$ and $c > 1$.

To this end, we show that: (1) $g(z_0) \ge 0$ and (2) for $z > z_0$, $g(z)$ is increasing. For (1), observe that

$$\begin{aligned} x\log(1 + yz_0) &= x\log(1 + y\,(3x\log(1 + xy))) \\ &\overset{(a)}{\le} x\log(1 + 3x^2y^2) \\ &\overset{(b)}{\le} x\log(1 + xy)^3 = z_0, \end{aligned}$$

where (a) follows because $\log(1 + a) \le a$ for $a > -1$, and (b) follows because $(1 + a)^3 = 1 + 3a + 3a^2 + a^3 \ge 1 + 3a^2$ for $a > 0$. For (2), first note that

$$\frac{dg(z)}{dz} = 1 - \frac{xy}{1 + yz}.$$

Therefore, $g(z)$ is strictly increasing when $z > x - \frac{1}{y}$, and so it suffices to show that $z_0 > x - \frac{1}{y}$. To do so, let $b = xy$. We have

$$3b\log(1 + b) \overset{(c)}{\ge} \frac{3b^2}{1 + b} \overset{(d)}{>} b - 1,$$

where (c) uses the fact that $\log(1 + a) \ge \frac{a}{1+a}$ for $a > -1$, and (d) follows because $3b^2 - (1 + b)(b - 1) = 2b^2 + 1 > 0$. It then follows that

$$z_0 = 3x\log(1 + xy) > x - \frac{1}{y},$$

which completes the proof. ■

Lemma 31 *For any $x \in \mathbb{R}^d$, $U \in \mathbb{R}^{d\times n}$, and $\lambda > 0$,*

$$x^\top \left(UU^\top + \lambda I\right)^{-1} x = \min_{z\in\mathbb{R}^n} \frac{1}{\lambda}\|x - Uz\|_2^2 + \|z\|_2^2.$$

Proof Let $g(z) := \frac{1}{\lambda}\|x - Uz\|_2^2 + \|z\|_2^2$. By a little algebra, we have

$$g(z) = \frac{1}{\lambda} z^\top \left(\lambda I + U^\top U\right) z - \frac{2}{\lambda} x^\top U z + \frac{1}{\lambda} x^\top x$$

It is easy to verify that $g(z)$ is convex and minimized at $z_* = \left(\lambda I + U^\top U\right)^{-1} U^\top x$.

It then follows that

$$\begin{aligned} g(z_*) &= \frac{1}{\lambda} x^\top U \left(\lambda I + U^\top U\right)^{-1} U^\top x - \frac{2}{\lambda} x^\top U \left(\lambda I + U^\top U\right)^{-1} U^\top x + \frac{1}{\lambda} x^\top x \\ &= \frac{1}{\lambda} x^\top x - \frac{1}{\lambda} x^\top U \left(\lambda I + U^\top U\right)^{-1} U^\top x \\ &\overset{\text{(a)}}{=} x^\top \left(\lambda I + U U^\top\right)^{-1} x, \end{aligned}$$

where (a) uses the Woodbury matrix identity:

$$\left(\lambda I + U U^\top\right)^{-1} = \frac{1}{\lambda} I - \frac{1}{\lambda} U \left(\lambda I + U^\top U\right)^{-1} U^\top.$$

■

Lemma 32 *Let $B \in \mathbb{R}^{d \times k}$ be an orthonormal basis of a k-dimensional subspace of $\mathbb{R}^d$. Let u be a vector in $\mathbb{R}^d$ such that $0 \le \underline{b} \le \|u\| \le \overline{b} \le 1$. Then,*

$$\min_{w \in \mathbb{R}^k : \|w\| \in [\underline{b}, \overline{b}]} \|Bw - u\| \le 2 \left\| P_B^\perp u \right\|.$$

Proof Let $\beta = \|u\| \in [\underline{b}, \overline{b}]$, and let $z = B^\top u$. It follows that $\|z\| \in [0, \beta]$. We consider three cases:

1. $\|z\| \ge \underline{b}$. In this case we have $\|z\| \in [\underline{b}, \overline{b}]$, and so

$$\min_{w \in \mathbb{R}^k : \|w\| \in [\underline{b}, \overline{b}]} \|Bw - u\| \le \|Bz - u\| = \left\| B B^\top u - u \right\| = \left\| P_B^\perp u \right\|.$$

2. $\|z\| = 0$. It follows that

$$\left\| P_B^\perp u \right\| = \|u\| = \beta,$$

and for any $w' \in \mathbb{R}^k$ such that $\|w'\| = \underline{b}$, by the triangle inequality,

$$\min_{w \in \mathbb{R}^k : \|w\| \in [\underline{b}, \overline{b}]} \|Bw - u\| \le \left\| Bw' - u \right\| \le \left\| Bw' \right\| + \|u\| \le \underline{b} + \beta \le 2 \left\| P_B^\perp u \right\|.$$

3. $0 < \|z\| < \underline{b}$. Consider $\hat{z} = \underline{b} \frac{z}{\|z\|}$. We have $\|\hat{z}\| = \underline{b}$. Then,

$$\begin{aligned} \min_{w \in \mathbb{R}^k : \|w\| \in [\underline{b}, \overline{b}]} \|Bw - u\| &\le \|B\hat{z} - u\| \\ &\overset{\text{(a)}}{\le} \|B\hat{z} - Bz\| + \|Bz - u\| \\ &\overset{\text{(b)}}{\le} \beta - \|z\| + \left\| P_B^\perp u \right\| \\ &\overset{\text{(c)}}{\le} 2 \left\| P_B^\perp u \right\|, \end{aligned}$$

where (a) uses the triangle inequality; (b) follows because

$$\|B\hat{z} - Bz\| = \|\hat{z} - z\| = \left\| \left(\frac{\underline{b}}{\|z\|} - 1 \right) z \right\| = \left(\frac{\underline{b}}{\|z\|} - 1 \right) \|z\| = \underline{b} - \|z\| \le \beta - \|z\|;$$

and (c) again uses the triangle inequality:

$$\beta = \|u\| \leq \|P_B u\| + \left\|P_B^{\perp} u\right\| = \|z\| + \left\|P_B^{\perp} u\right\|.$$

■

Appendix D. Supplementary material for Section 7

D.1. Implementation details

All implementations were developed in PyTorch. Most experiments were conducted on machines equipped with NVIDIA GeForce RTX 4090 GPUs.

Synthetic linear experiments. For data generation, in each trial, B^* was obtained from the QR decomposition of a random $d \times k$ matrix with standard normal entries, and w_t^*'s were drawn from the uniform distribution over the origin-centered sphere with radius β. For both multi-task ERM and few-shot property tests, we used batch gradient descent with the Adam optimizer (Kingma, 2014) (learning rate 10^{-3}), for a maximum of 10^4 epochs with early stopping after 20 epochs without improvement. The Bayes-optimal risks were estimated via Monte-Carlo simulations from 10^6 samples.

MNIST experiments. Both multi-task ERM and few-shot property tests used batch gradient descent with the Adam optimizer (learning rate 10^{-3}), for up to 2000 epochs. For multi-task ERM, early stopping was applied after 10 epochs without improvement.

CIFAR-10 experiments. We consider representations given by a modified ResNet-18 architecture (He et al., 2016). The initial convolutional layer is replaced by a 3×3 kernel with stride 1, with the remaining convolutional blocks kept unchanged. After the convolutional blocks, we apply global average pooling, followed by a fully-connected linear layer, batch normalization, and ReLU activation. The resulting representation is of dimension $k = 256$. All representation networks were trained from scratch without any pretrained weights.

For multi-task ERM, we used mini-batch stochastic gradient descent with the Adam optimizer (learning rate 10^{-3}), a batch size of 256, and trained for up to 300 epochs with early stopping after 20 epochs without improvement. For few-shot property tests, we used batch gradient descent with the Adam optimizer (learning rate 10^{-2}) for a maximum of 200 epochs.