On the Theory of Transfer Learning: The Importance of Task Diversity

Nilesh Tripuraneni
University of California, Berkeley
nilesh_tripuraneni@berkeley.edu

Michael I. Jordan
University of California, Berkeley
jordan@cs.berkeley.edu

Chi Jin
Princeton University
chij@princeton.edu

Abstract

We provide new statistical guarantees for transfer learning via representation learning–when transfer is achieved by learning a feature representation shared across different tasks. This enables learning on new tasks using far less data than is required to learn them in isolation. Formally, we consider $t+1$ tasks parameterized by functions of the form $f_j \circ h$ in a general function class $\mathcal{F} \circ \mathcal{H}$, where each f_j is a task-specific function in $\mathcal{F}$ and h is the shared representation in $\mathcal{H}$. Letting $C(\cdot)$ denote the complexity measure of the function class, we show that for diverse training tasks (1) the sample complexity needed to learn the shared representation across the first t training tasks scales as $C(\mathcal{H}) + tC(\mathcal{F})$, despite no explicit access to a signal from the feature representation and (2) with an accurate estimate of the representation, the sample complexity needed to learn a new task scales only with $C(\mathcal{F})$. Our results depend upon a new general notion of task diversity–applicable to models with general tasks, features, and losses–as well as a novel chain rule for Gaussian complexities. Finally, we exhibit the utility of our general framework in several models of importance in the literature.

1 Introduction

Transfer learning is quickly becoming an essential tool to address learning problems in settings with *small* data. One of the most promising methods for multitask and transfer learning is founded on the belief that multiple, differing tasks are distinguished by a small number of task-specific parameters, but often share a common low-dimensional representation. Undoubtedly, one of the most striking successes of this idea has been to only re-train the final layers of a neural network on new task data, after initializing its earlier layers with hierarchical representations/features from ImageNet (i.e., ImageNet pretraining) [Donahue et al., 2014, Gulshan et al., 2016]. However, the practical purview of transfer learning has extended far beyond the scope of computer vision and classical ML application domains such as deep reinforcement learning [Baevski et al., 2019], to problems such as protein engineering and design [Elnaggar et al., 2020].

In this paper, we formally study the composite learning model in which there are $t + 1$ tasks whose responses are generated noisily from the function $f_j^\star \circ \mathbf{h}^\star$, where $f_j^\star$ are task-specific parameters in a function class $\mathcal{F}$ and $\mathbf{h}^\star$ an underlying shared representation in a function class $\mathcal{H}$. A large empirical literature has documented the performance gains that can be obtained by transferring a jointly learned representation $\mathbf{h}$ to new tasks in this model [Yosinski et al., 2014, Raghu et al., 2019, Lee et al., 2019].

There is also a theoretical literature that dates back at least as far as [Baxter, 2000]. However, this progress belies a lack of understanding of the basic statistical principles underlying transfer learning[1]:

How many samples do we need to learn a feature representation shared across tasks and use it to improve prediction on a new task?

In this paper we study a simple two-stage empirical risk minimization procedure to learn a new, $j = 0$th task which shares a common representation with t different training tasks. This procedure first learns a representation $\hat{\mathbf{h}} \approx \mathbf{h}^\star$ given n samples from each of t different training tasks, and then uses $\hat{\mathbf{h}}$ alongside m fresh samples from this new task to learn $\hat{f}_0 \circ \hat{\mathbf{h}} \approx f_0^\star \circ \mathbf{h}^\star$. Informally, our main result provides an answer to our sampling-complexity question by showing that the excess risk of prediction of this two-stage procedure scales (on the new task) as[2],

$$\tilde{O}\left(\frac{1}{\nu} \left(\sqrt{\frac{C(\mathcal{H}) + tC(\mathcal{F})}{nt}} \right) + \sqrt{\frac{C(\mathcal{F})}{m}} \right),$$

where $C(\mathcal{H})$ captures the complexity of the shared representation, $C(\mathcal{F})$ captures the complexity of the task-specific maps, and ν encodes a problem-agnostic notion of task diversity. The latter is a key contribution of the current paper. It represents the extent to which the t training tasks $f_j^\star$ cover the space of the features $\mathbf{h}^\star$. In the limit that $n, t \to \infty$ (i.e., training task data is abundant), to achieve a fixed level of constant prediction error on the new task only requires the number of fresh samples to be $m \approx C(\mathcal{F})$. Learning the task in isolation suffers the burden of learning both $\mathcal{F}$ and $\mathcal{H}$—requiring $m \approx C(\mathcal{F} \circ \mathcal{H})$—which can be significantly greater than the transfer learning sample complexity.

Maurer et al. [2016] present a general, uniform-convergence based framework for obtaining generalization bounds for transfer learning that scale as $O(1/\sqrt{t}) + O(1/\sqrt{m})$ (for clarity we have suppressed complexity factors in the numerator). Perhaps surprisingly, the leading term capturing the complexity of learning $\mathbf{h}^\star$ decays only in t but not in n. This suggests that increasing the number of samples per training task cannot improve generalization on new tasks. Given that most transfer learning applications in the literature collect information from only a few training tasks (i.e., $n \gg t$), this result does not provide a fully satisfactory explanation for the practical efficacy of transfer learning methods.

Our principal contributions in this paper are as follows:

- We introduce a problem-agnostic definition of task diversity which can be integrated into a uniform convergence framework to provide generalization bounds for transfer learning problems with general losses, tasks, and features. Our framework puts this notion of diversity together with a common-design assumption across tasks to provide guarantees of a fast convergence rate, decaying with *all of the samples* for the transfer learning problem.

- We provide general-purpose bounds which decouple the complexity of learning the task-specific structure from the complexity of learning the shared feature representation. Our results repose on a novel user-friendly chain rule for Gaussian processes which may be of independent interest (see Theorem 7). Crucially, this chain rule implies a form of modularity that allows us to exploit a plethora of existing results from the statistics and machine learning literatures to individually bound the sample complexity of learning task and feature functions.

- We highlight the utility of our framework for obtaining end-to-end transfer learning guarantees for several different multi-task learning models including (1) logistic regression, (2) deep neural network regression, and (3) robust regression for single-index models.

1.1 Related Work

The utility of multitask learning methods was observed at least as far back as Caruana [1997]. In recent years, representation learning, transfer learning, and meta-learning have been the subject of extensive empirical investigation in the machine learning literature (see [Bengio et al., 2013],

---

[1] A problem which is also often referred to as learning-to-learn (LTL).

[2] See Theorem 3 and discussion for a formal statement. Note our guarantees also hold for nonparametric function classes, but the scaling with n, t, m may in general be different.

[Hospedales et al., 2020] for surveys in these directions). However, theoretical work on transfer learning—particularly via representation learning—has been much more limited.

A line of work closely related to transfer learning is gradient-based meta-learning (MAML) [Finn et al., 2017]. These methods have been analyzed using techniques from online convex optimization, using a (potentially data-dependent) notion of task similarity which assumes that tasks are close to a global task parameter [Finn et al., 2019, Khodak et al., 2019a, Denevi et al., 2019a,b, Khodak et al., 2019b]. Additionally, Ben-David and Borbely [2008] define a different notion of distributional task similarity they use to show generalization bounds. However, these works do not study the question of transferring a common representation in the generic composite learning model that is our focus.

In settings restricted to linear task mappings and linear features, Lounici et al. [2011], Pontil and Maurer [2013], and Cavallanti et al. [2010] have provided sample complexity bounds for the problem of transfer learning via representation learning. Lounici et al. [2011] and Obozinski et al. [2011] also address sparsity-related issues that can arise in linear feature learning.

To our knowledge, Baxter [2000] is the first theoretical work to provide generalization bounds for transfer learning via representation learning in a general setting. The formulation of Baxter [2000] assumes a generative model over tasks which share common features; in our setting, this task generative model is replaced by the assumption that training tasks are diverse (as in Definition 3) and that there is a common covariate distribution across different tasks. In follow-up work, Maurer et al. [2016] propose a general, uniform-convergence-based framework for obtaining transfer learning guarantees which scale as $O(1/\sqrt{t}) + O(1/\sqrt{m})$ [Maurer et al., 2016, Theorem 5]. The second term represents the sample complexity of learning in a lower-dimensional space given the common representation. The first term is the bias contribution from transferring the representation—learned from an aggregate of nt samples across different training tasks—to a new task. Note this leading term decays only in t and not in n: implying that increasing the number of samples per training task cannot improve generalization on new tasks. Unfortunately, under the framework studied in that paper, this $\Omega(1/\sqrt{t})$ cannot be improved Maurer et al. [2016].

Recent work in Tripuraneni et al. [2020] and Du et al. [2020] has shown that in specific settings leveraging (1) common design assumptions across tasks and (2) a particular notion of task diversity, can break this barrier and yield rates for the leading term which decay as $O(\text{poly}(1/(nt)))$. However, the results and techniques used in both of these works are limited to the squared loss and linear task maps. Moreover, the notion of diversity in both cases arises purely from the linear-algebraic conditioning of the set of linear task maps. It is not clear from these works how to extend these ideas/techniques beyond the case-specific analyses therein.

2 Preliminaries

Notation: We use bold lower-case letters (e.g., $\mathbf{x}$) to refer to vectors and bold upper-case letters (e.g., $\mathbf{X}$) to refer to matrices. The norm $\| \cdot \|$ appearing on a vector or matrix refers to its ℓ_2 norm or spectral norm respectively. We use the bracketed notation $[n] = \{1, \ldots, n\}$ as shorthand for integer sets. Generically, we will use "hatted" vectors and matrices (e.g, $\hat{\boldsymbol{\alpha}}$ and $\hat{\mathbf{B}}$) to refer to (random) estimators of their underlying population quantities. $\sigma_1(\mathbf{A}), \ldots, \sigma_r(\mathbf{A})$ will denote the sorted singular values (in decreasing magnitude) of a rank r matrix $\mathbf{A}$. Throughout we will use $\mathcal{F}$ to refer to a function class of tasks mapping $\mathbb{R}^r \to \mathbb{R}$ and $\mathcal{H}$ to be a function class of features mapping $\mathbb{R}^d \to \mathbb{R}^r$. For the function class $\mathcal{F}$, we use $\mathcal{F}^{\otimes t}$ to refer its t-fold Cartesian product, i.e., $\mathcal{F}^{\otimes t} = \{\mathbf{f} \equiv (f_1, \ldots, f_t) \mid f_j \in \mathcal{F} \text{ for any } j \in [t]\}$. We use $\tilde{O}$ to denote an expression that hides polylogarithmic factors in all problem parameters.

2.1 Transfer learning with a shared representation

In our treatment of transfer learning, we assume that there exists a generic nonlinear feature representation that is shared across all tasks. Since this feature representation is shared, it can be utilized to transfer knowledge from existing tasks to new tasks. Formally, we assume that for a particular task j, we observe multiple data pairs $\{(\mathbf{x}_{ji}, y_{ji})\}$ (indexed over i) that are sampled i.i.d from an *unknown* distribution $\mathbb{P}_j$, supported over $\mathcal{X} \times \mathcal{Y}$ and defined as follows:

$$\mathbb{P}_j(\mathbf{x}, y) = \mathbb{P}_{f_j^\star \circ \mathbf{h}^\star}(\mathbf{x}, y) = \mathbb{P}_\mathbf{x}(\mathbf{x})\mathbb{P}_{y|\mathbf{x}}(y|f_j^\star \circ \mathbf{h}^\star(\mathbf{x})). \tag{1}$$

Here, $\mathbf{h}^\star : \mathbb{R}^d \to \mathbb{R}^r$ is the shared feature representation, and $f_j^\star : \mathbb{R}^r \to \mathbb{R}$ is a task-specific mapping. Note that we assume that the marginal distribution over $\mathcal{X}$—$\mathbb{P}_{\mathbf{x}}$—is common amongst all the tasks.

We consider transfer learning methods consisting of two phases. In the first phase (the training phase), t tasks with n samples per task are available for learning. Our objective in this phase is to learn the shared feature representation using the entire set of nt samples from the first $j \in [t]$ tasks. In the second phase (the test phase), we are presented with m fresh samples from a new task that we denote as the 0th task. Our objective in the test phase is to learn this new task based on both the fresh samples and the representation learned in the first phase.

Formally, we consider a two-stage Empirical Risk Minimization (ERM) procedure for transfer learning. Consider a function class $\mathcal{F}$ containing task-specific functions, and a function class $\mathcal{H}$ containing feature maps/representations. In the training phase, the empirical risk for t training tasks is:

$$\hat{R}_{\text{train}}(\mathbf{f}, \mathbf{h}) := \frac{1}{nt} \sum_{j=1}^{t} \sum_{i=1}^{n} \ell(f_j \circ \mathbf{h}(\mathbf{x}_{ji}), y_{ji}), \qquad (2)$$

where $\ell(\cdot, \cdot)$ is the loss function and $\mathbf{f} := (f_1, \dots, f_t) \in \mathcal{F}^{\otimes t}$. Our estimator $\hat{\mathbf{h}}(\cdot)$ for the shared data representation is given by $\hat{\mathbf{h}} = \operatorname{argmin}_{\mathbf{h} \in \mathcal{H}} \min_{\mathbf{f} \in \mathcal{F}^{\otimes t}} \hat{R}_{\text{train}}(\mathbf{f}, \mathbf{h})$.

For the second stage, the empirical risk for learning the new task is defined as:

$$\hat{R}_{\text{test}}(f, \mathbf{h}) := \frac{1}{m} \sum_{i=1}^{m} \ell(f \circ \mathbf{h}(\mathbf{x}_{0i}), y_{0i}). \qquad (3)$$

We estimate the underlying function $f_0^\star$ for task 0 by computing the ERM based on the feature representation learned in the first phase. That is, $\hat{f}_0 = \operatorname{argmin}_{f \in \mathcal{F}} \hat{R}_{\text{test}}(f, \hat{\mathbf{h}})$. We gauge the efficacy of the estimator $(\hat{f}_0, \hat{\mathbf{h}})$ by its excess risk on the new task, which we refer to as the *transfer learning risk*:

$$\text{Transfer Learning Risk} = R_{\text{test}}(\hat{f}_0, \hat{\mathbf{h}}) - R_{\text{test}}(f_0^\star, \mathbf{h}^\star). \qquad (4)$$

Here, $R_{\text{test}}(\cdot, \cdot) = \mathbb{E}[\hat{R}_{\text{test}}(\cdot, \cdot)]$ is the population risk for the new task and the population risk over the t training tasks is similarly defined as $R_{\text{train}}(\cdot, \cdot) = \mathbb{E}[\hat{R}_{\text{train}}(\cdot, \cdot)]$; both expectations are taken over the randomness in the training and test phase datasets respectively. The transfer learning risk measures the expected prediction risk of the function $(\hat{f}_0, \hat{\mathbf{h}})$ on a new datapoint for the 0th task, relative to the best prediction rule from which the data was generated—$f_0^\star \circ \mathbf{h}^\star$.

2.2 Model complexity

A well-known measure for the complexity of a function class is its Gaussian complexity. For a generic vector-valued function class $\mathcal{Q}$ containing functions $\mathbf{q}(\cdot) : \mathbb{R}^d \to \mathbb{R}^r$, and N data points, $\bar{\mathbf{X}} = (\mathbf{x}_1, \dots, \mathbf{x}_N)^\top$, the empirical Gaussian complexity is defined as

$$\hat{\mathfrak{S}}_{\bar{\mathbf{X}}}(\mathcal{Q}) = \mathbb{E}_{\mathbf{g}}[\sup_{\mathbf{q} \in \mathcal{Q}} \frac{1}{N} \sum_{k=1}^{r} \sum_{i=1}^{N} g_{ki} q_k(\mathbf{x}_i)], \qquad g_{ki} \sim \mathcal{N}(0,1) \ i.i.d.,$$

where $\mathbf{g} = \{g_{ki}\}_{k \in [r], i \in [N]}$, and $q_k(\cdot)$ is the k-th coordinate of the vector-valued function $\mathbf{q}(\cdot)$. We define the corresponding population Gaussian complexity as $\mathfrak{S}_N(\mathcal{Q}) = \mathbb{E}_{\bar{\mathbf{X}}}[\hat{\mathfrak{S}}_{\bar{\mathbf{X}}}(\mathcal{Q})]$, where the expectation is taken over the distribution of data samples $\bar{\mathbf{X}}$. Intuitively, $\mathfrak{S}_N(\mathcal{Q})$ measures the complexity of $\mathcal{Q}$ by the extent to which functions in the class $\mathcal{Q}$ can correlate with random noise g_{ki}.

3 Main Results

We now present our central theoretical results for the transfer learning problem. We first present statistical guarantees for the training phase and test phase separately. Then, we present a problem-agnostic definition of task diversity, followed by our generic end-to-end transfer learning guarantee. Throughout this section, we make the following standard, mild regularity assumptions on the loss function $\ell(\cdot, \cdot)$, the function class of tasks $\mathcal{F}$, and the function class of shared representations $\mathcal{H}$.

Assumption 1 (Regularity conditions)**.** *The following regularity conditions hold:*

- *The loss function $\ell(\cdot, \cdot)$ is B-bounded, and $\ell(\cdot, y)$ is L-Lipschitz for all $y \in \mathcal{Y}$.*

- *The function f is $L(\mathcal{F})$-Lipschitz with respect to the ℓ_2 distance, for any $f \in \mathcal{F}$.*

- *The composed function $f \circ \mathbf{h}$ is bounded: $\sup_{\mathbf{x} \in \mathcal{X}} |f \circ \mathbf{h}(\mathbf{x})| \leq D_{\mathcal{X}}$, for any $f \in \mathcal{F}, \mathbf{h} \in \mathcal{H}$.*

We also make the following realizability assumptions, which state that the true underlying task functions and the true representation are contained in the function classes $\mathcal{F}, \mathcal{H}$ over which the two-stage ERM oracle optimizes in (2) and (3).

Assumption 2 (Realizability)**.** *The true representation $\mathbf{h}^{\star}$ is contained in $\mathcal{H}$. Additionally, the true task specific functions $f_j^{\star}$ are contained in $\mathcal{F}$ for both the training tasks and new test task (i.e., for any $j \in [t] \cup \{0\}$).*

3.1 Learning shared representations

In order to measure "closeness" between the learned representation and true underlying feature representation, we need to define an appropriate distance measure between arbitrary representations. To this end, we begin by introducing the *task-averaged representation difference*, which captures the extent two representations $\mathbf{h}$ and $\mathbf{h}'$ differ in aggregate over the t training tasks measured by the population train loss.

Definition 1. For a function class $\mathcal{F}$, t functions $\mathbf{f} = (f_1, \ldots, f_t)$, and data $(\mathbf{x}_j, y_j) \sim \mathbb{P}_{f_j \circ \mathbf{h}}$ as in (1) for any $j \in [t]$, the **task-averaged representation difference** between representations $\mathbf{h}, \mathbf{h}' \in \mathcal{H}$ is:

$$\bar{d}_{\mathcal{F}, \mathbf{f}}(\mathbf{h}'; \mathbf{h}) = \frac{1}{t} \sum_{j=1}^{t} \inf_{f' \in \mathcal{F}} \mathbb{E}_{\mathbf{x}_j, y_j} \left\{ \ell(f' \circ \mathbf{h}'(\mathbf{x}_j), y_j) - \ell(f_j \circ \mathbf{h}(\mathbf{x}_j), y_j) \right\}.$$

Under this metric, we can show that the distance between a learned representation and the true underlying representation is controlled in the training phase. Our following guarantees also feature the *worst-case Gaussian complexity* over the function class $\mathcal{F}$, which is defined as:[3]

$$\bar{\mathfrak{G}}_n(\mathcal{F}) = \max_{\mathbf{Z} \in \mathcal{Z}} \hat{\mathfrak{G}}_{\mathbf{Z}}(\mathcal{F}), \text{ where } \mathcal{Z} = \{(\mathbf{h}(\mathbf{x}_1), \cdots, \mathbf{h}(\mathbf{x}_n)) \mid \mathbf{h} \in \mathcal{H}, \mathbf{x}_i \in \mathcal{X} \text{ for all } i \in [n]\}. \quad (5)$$

where $\mathcal{Z}$ is the domain induced by any set of n samples in $\mathcal{X}$ and any representation $\mathbf{h} \in \mathcal{H}$. Moreover, we will always use the subscript nt, on $\mathfrak{G}_{nt}(\mathcal{Q}) = \mathbb{E}_{\mathbf{X}}[\hat{\mathfrak{G}}_{\mathbf{X}}(\mathcal{Q})]$, to refer to the population Gaussian complexity computed with respect to the data matrix $\mathbf{X}$ formed from the concatentation of the nt training datapoints $\{\mathbf{x}_{ji}\}_{j=1, i=1}^{t, n}$. We can now present our training phase guarantee.

Theorem 1. *Let $\hat{\mathbf{h}}$ be an empirical risk minimizer of $\hat{R}_{train}(\cdot, \cdot)$ in (2). Then, if Assumptions 1 and 2 hold, with probability at least $1 - \delta$:*

$$\bar{d}_{\mathcal{F}, \mathbf{f}^{\star}}(\hat{\mathbf{h}}; \mathbf{h}^{\star}) \leq 16 L \mathfrak{G}_{nt}(\mathcal{F}^{\otimes t} \circ \mathcal{H}) + 8B\sqrt{\frac{\log(2/\delta)}{nt}}$$

$$\leq 4096 L \left[\frac{D_{\mathcal{X}}}{(nt)^2} + \log(nt) \cdot [L(\mathcal{F}) \cdot \mathfrak{G}_{nt}(\mathcal{H}) + \bar{\mathfrak{G}}_n(\mathcal{F})] \right] + 8B\sqrt{\frac{\log(2/\delta)}{nt}}.$$

Theorem 1 asserts that the *task-averaged representation difference* (Definition 1) between our learned representation and the true representation is upper bounded by the population Gaussian complexity of the vector-valued function class $\mathcal{F}^{\otimes t} \circ \mathcal{H} = \{(f_1 \circ \mathbf{h}, \ldots, f_t \circ \mathbf{h}) : (f_1, \ldots, f_t) \in \mathcal{F}^{\otimes t}, \mathbf{h} \in \mathcal{H}\}$, plus a lower-order noise term. Up to logarithmic factors and lower-order terms, this Gaussian complexity can be further decomposed into the complexity of learning a representation in $\mathcal{H}$ with nt samples—$L(\mathcal{F}) \cdot \mathfrak{G}_{nt}(\mathcal{H})$—and the complexity of learning a task-specific function in $\mathcal{F}$ using n

---

[3]Note that a stronger version of our results hold with a sharper, data-dependent version of the worst-case Gaussian complexity that eschews the absolute maxima over $\mathbf{x}_i$. See Corollary 1 and Theorem 7 for the formal statements.

samples per task—$\bar{\mathfrak{G}}_n(\mathcal{F})$. For the majority of parametric function classes used in machine learning applications, $\mathfrak{G}_{nt}(\mathcal{H}) \sim \sqrt{C(\mathcal{H})/nt}$ and $\bar{\mathfrak{G}}_n(\mathcal{F}) \sim \sqrt{C(\mathcal{F})/n}$, where the function $C(\cdot)$ measures the intrinsic complexity of the function class (e.g., VC dimension, absolute dimension, or parameter norm [Wainwright, 2019]).

We now make several remarks on this result. First, Theorem 1 differs from standard supervised learning generalization bounds. Theorem 1 provides a bound on the distance between two representations as opposed to the empirical or population training risk, despite the lack of access to a direct signal from the underlying feature representation. Second, the decomposition of $\mathfrak{G}_{nt}(\mathcal{F}^{\otimes t} \circ \mathcal{H})$ into the individual Gaussian complexities of $\mathcal{H}$ and $\mathcal{F}$, leverages a novel chain rule for Gaussian complexities (see Theorem 7), which may be of independent interest. This chain rule (Theorem 7) can be viewed as a generalization of classical Gaussian comparison inequalities and results such as the Ledoux-Talagrand contraction principle [Ledoux and Talagrand, 2013]. Further details and comparisons to the literature for this chain rule can be found in Appendix B.2 (this result also avoids an absolute maxima over $\mathbf{x}_i \in \mathcal{X}$).

3.2 Transferring to new tasks

In addition to the *task-averaged representation difference*, we also introduce the *worst-case representation difference*, which captures the distance between two representations $\mathbf{h}'$, $\mathbf{h}$ in the context of an arbitrary worst-case task-specific function $f_0 \in \mathcal{F}_0$.

Definition 2. For function classes $\mathcal{F}$ and $\mathcal{F}_0$ such that $f_0 \in \mathcal{F}_0$, and data $(\mathbf{x}, y) \sim \mathbb{P}_{f_0 \circ \mathbf{h}}$ as in (1), the **worst-case representation difference** between representations $\mathbf{h}, \mathbf{h}' \in \mathcal{H}$ is:

$$d_{\mathcal{F},\mathcal{F}_0}(\mathbf{h}'; \mathbf{h}) = \sup_{f_0 \in \mathcal{F}_0} \inf_{f' \in \mathcal{F}} \mathbb{E}_{\mathbf{x},y} \Big\{ \ell(f' \circ \mathbf{h}'(\mathbf{x}), y) - \ell(f_0 \circ \mathbf{h}(\mathbf{x}), y) \Big\}.$$

For flexibility we allow $\mathcal{F}_0$ to be distinct from $\mathcal{F}$ (although in most cases, we choose $\mathcal{F}_0 \subset \mathcal{F}$). The function class $\mathcal{F}_0$ is the set of new tasks on which we hope to generalize. The generalization guarantee for the test phase ERM estimator follows.

Theorem 2. *Let $\hat{f}_0$ be an empirical risk minimizer of $\hat{R}_{test}(\cdot, \hat{\mathbf{h}})$ in (3) for any feature representation $\hat{\mathbf{h}}$. Then if Assumptions 1 and 2 hold, and $f_0^\star \in \mathcal{F}_0$ for an unknown class $\mathcal{F}_0$, with probability at least $1 - \delta$:*

$$R_{test}(\hat{f}_0, \hat{\mathbf{h}}) - R_{test}(f_0^\star, \mathbf{h}^\star) \leq d_{\mathcal{F},\mathcal{F}_0}(\hat{\mathbf{h}}; \mathbf{h}^\star) + 16L \cdot \bar{\mathfrak{G}}_m(\mathcal{F}) + 8B\sqrt{\frac{\log(2/\delta)}{m}}$$

Here $\bar{\mathfrak{G}}_m(\mathcal{F})$ is again the worst-case Gaussian complexity[4] as defined in (5). Theorem 2 provides an excess risk bound for prediction on a new task in the test phase with two dominant terms. The first is the worst-case representation difference $d_{\mathcal{F},\mathcal{F}_0}(\hat{\mathbf{h}}; \mathbf{h}^\star)$, which accounts for the error of using a biased feature representation $\hat{\mathbf{h}} \neq \mathbf{h}^\star$ in the test ERM procedure. The second is the difficulty of learning $f_0^\star$ with m samples, which is encapsulated in $\bar{\mathfrak{G}}_m(\mathcal{F})$.

3.3 Task diversity and end-to-end transfer learning guarantees

We now introduce the key notion of task diversity. Since the learner does not have direct access to a signal from the representation, they can only observe partial information about the representation channeled through the composite functions $f_j^\star \circ \mathbf{h}^\star$. If a particular direction/component in $\mathbf{h}^\star$ is not seen by a corresponding task $f_j^\star$ in the training phase, that component of the representation $\mathbf{h}^\star$ cannot be distinguished from a corresponding one in a spurious $\mathbf{h}'$. When this component is needed to predict on a new task corresponding to $f_0^\star$ which lies along that particular direction, transfer learning will not be possible. Accordingly, Definition 1 defines a notion of representation distance in terms of information channeled through the training tasks, while Definition 2 defines it in terms of an arbitrary new test task. Task diversity essentially encodes the ratio of these two quantities (i.e. how well the training tasks can cover the space captured by the representation $\mathbf{h}^\star$ needed to predict on new tasks). Intuitively, if all the task-specific functions were quite similar, then we would only expect the training

---

[4]As before, a stronger version of this result holds with a sharper data-dependent version of the Gaussian complexity in lieu of $\bar{\mathfrak{G}}_m(\mathcal{F})$ (see Corollary 2).

stage to learn about a narrow slice of the representation—making transferring to a generic new task difficult.

Definition 3. *For a function class $\mathcal{F}$, we say t functions $\mathbf{f} = (f_1, \ldots, f_t)$ are (ν, ϵ)-**diverse** over $\mathcal{F}_0$ for a representation $\mathbf{h}$, if uniformly for all $\mathbf{h}' \in \mathcal{H}$,*

$$d_{\mathcal{F}, \mathcal{F}_0}(\mathbf{h}'; \mathbf{h}) \leq \bar{d}_{\mathcal{F}, \mathbf{f}}(\mathbf{h}'; \mathbf{h})/\nu + \epsilon.$$

Up to a small additive error ϵ, diverse tasks ensure that the worst-case representation difference for the function class $\mathcal{F}_0$ is controlled when the task-averaged representation difference for a sequence of t tasks $\mathbf{f}$ is small. Despite the abstraction in this definition of task diversity, it *exactly* recovers the notion of task diversity in Tripuraneni et al. [2020] and Du et al. [2020], where it is restricted to the special case of linear functions and quadratic loss. Our general notion allows us to move far beyond the linear-quadratic setting as we show in Section 4 and Appendix A.1.

We now utilize the definition of task diversity to merge our training phase and test phase results into an end-to-end transfer learning guarantee for generalization to the unseen task $f_0^\star \circ \mathbf{h}^\star$.

Theorem 3. *Let $(\cdot, \hat{\mathbf{h}})$ be an empirical risk minimizer of $\hat{R}_{train}(\cdot, \cdot)$ in (2), and $\hat{f}_0$ be an empirical risk minimizer of $\hat{R}_{test}(\cdot, \hat{\mathbf{h}})$ in (3) for the learned feature representation $\hat{\mathbf{h}}$. Then if Assumptions 1 and 2 hold, and the training tasks are (ν, ϵ)-diverse, with probability at least $1 - 2\delta$, the transfer learning risk in (4) is upper-bounded by:*

$$O\left(L \log(nt) \cdot \left[\frac{L(\mathcal{F}) \cdot \mathfrak{G}_{nt}(\mathcal{H}) + \bar{\mathfrak{G}}_n(\mathcal{F})}{\nu} \right] + L\bar{\mathfrak{G}}_m(\mathcal{F}) + \frac{LD_{\mathcal{X}}}{\nu(nt)^2} + B\left[\frac{1}{\nu} \cdot \sqrt{\frac{\log(2/\delta)}{nt}} + \sqrt{\frac{\log(2/\delta)}{m}} \right] + \epsilon \right).$$

Theorem 3 gives an upper bound on the transfer learning risk. The dominant terms in the bound are the three Gaussian complexity terms. For parametric function classes we expect $\mathfrak{G}_{nt}(\mathcal{H}) \sim \sqrt{C(\mathcal{H})/(nt)}$ and $\bar{\mathfrak{G}}_N(\mathcal{F}) \sim \sqrt{C(\mathcal{F})/N}$, where $C(\mathcal{H})$ and $C(\mathcal{F})$ capture the dimension-dependent size of the function classes. Therefore, when L and $L(\mathcal{F})$ are constants, the leading-order terms for the transfer learning risk scale as $\tilde{O}(\sqrt{(C(\mathcal{H}) + t \cdot C(\mathcal{F}))/(nt)} + \sqrt{C(\mathcal{F})/m})$. A naive algorithm which simply learns the new task in isolation, ignoring the training tasks, has an excess risk scaling as $\tilde{O}(\sqrt{C(\mathcal{F} \circ \mathcal{H})/m}) \approx \tilde{O}(\sqrt{(C(\mathcal{H}) + C(\mathcal{F}))/m})$. Therefore, when n and t are sufficiently large, but m is relatively small (i.e., the setting of few-shot learning), the performance of transfer learning is significantly better than the baseline of learning in isolation.

4 Applications

We now consider a varied set of applications to instantiate our general transfer learning framework. In each application, we first specify the function classes and data distributions we are considering as well as our assumptions. We then state the task diversity and the Gaussian complexities of the function classes, which together furnish the bounds on the *transfer learning risk*–from (4)–in Theorem 3.

4.1 Multitask Logistic Regression

We first instantiate our framework for one of the most frequently used classification methods—logistic regression. Consider the setting where the task-specific functions are linear maps, and the underlying representation is a projection onto a low-dimensional subspace. Formally, let $d \geq r$, and let the function classes $\mathcal{F}$ and $\mathcal{H}$ be:

$$\mathcal{F} = \{ f \mid f(\mathbf{z}) = \boldsymbol{\alpha}^\top \mathbf{z}, \ \boldsymbol{\alpha} \in \mathbb{R}^r, \ \|\boldsymbol{\alpha}\| \leq c_1 \}, \tag{6}$$

$$\mathcal{H} = \{ \mathbf{h} \mid \mathbf{h}(\mathbf{x}) = \mathbf{B}^\top \mathbf{x}, \ \mathbf{B} \in \mathbb{R}^{d \times r}, \ \mathbf{B} \text{ is a matrix with orthonormal columns} \}.$$

Here $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{0, 1\}$, and the measure $\mathbb{P}_{\mathbf{x}}$ is $\boldsymbol{\Sigma}$-sub-gaussian (see Definition 4) and D-bounded (i.e., $\|\mathbf{x}\| \leq D$ with probability one). We let the conditional distribution in (1) satisfy:

$$\mathbb{P}_{y|\mathbf{x}}(y = 1 | f \circ \mathbf{h}(\mathbf{x})) = \sigma(\boldsymbol{\alpha}^\top \mathbf{B}^\top \mathbf{x}),$$

where $\sigma(\cdot)$ is the sigmoid function with $\sigma(z) = 1/(1 + \exp(-z))$. We use the logistic loss $\ell(z, y) = -y \log(\sigma(z)) - (1 - y) \log(1 - \sigma(z))$. The true training tasks take the form $f_j^\star(\mathbf{z}) = (\boldsymbol{\alpha}_j^\star)^\top \mathbf{z}$ for all $j \in [t]$, and we let $\mathbf{A} = (\boldsymbol{\alpha}_1^\star, \ldots, \boldsymbol{\alpha}_t^\star)^\top \in \mathbb{R}^{t \times r}$. We make the following assumption on the training tasks being "diverse" and both the training and new task being normalized.

Assumption 3. $\sigma_r(\mathbf{A}^\top \mathbf{A}/t) = \tilde{\nu} > 0$ and $\|\boldsymbol{\alpha}_j^\star\| \le O(1)$ for $j \in [t] \cup \{0\}$.

In this case where the $\mathcal{F}$ contains underlying linear task functions $\boldsymbol{\alpha}_j^\star \in \mathbb{R}^r$ (as in our examples in Section 4), our task diversity definition reduces to ensuring these task vectors span the entire r-dimensional space containing the output of the representation $\mathbf{h}(\cdot) \in \mathbb{R}^r$. This is quantitatively captured by the conditioning parameter $\tilde{\nu} = \sigma_r(\mathbf{A})$ which represents how spread out these vectors are in $\mathbb{R}^r$. The training tasks will be well-conditioned in the sense that $\sigma_1(\mathbf{A}^\top \mathbf{A}/t)/\sigma_r(\mathbf{A}^\top \mathbf{A}/t) \le O(1)$ (w.h.p.) for example, if each $\boldsymbol{\alpha}_t \sim \mathcal{N}(0, \frac{1}{\sqrt{r}}\boldsymbol{\Sigma})$ i.i.d. with $\sigma_1(\boldsymbol{\Sigma})/\sigma_r(\boldsymbol{\Sigma}) \le O(1)$.

Assumption 3 with natural choices of $\mathcal{F}_0$ and $\mathcal{F}$ establishes $(\Omega(\tilde{\nu}), 0)$-diversity as defined in Definition 3 (see Lemma 1). Finally, by standard arguments, we can bound the Gaussian complexity of $\mathcal{H}$ in this setting by $\mathfrak{G}_N(\mathcal{H}) \le \tilde{O}(\sqrt{dr^2/N})$. We can also show that a finer notion of the Gaussian complexity for $\mathcal{F}$, serving as the analog of $\bar{\mathfrak{G}}_N(\mathcal{F})$, is upper bounded by $\tilde{O}(\sqrt{r/N})$. This is used to sharply bound the complexity of learning $\mathcal{F}$ in the training and test phases (see proof of Theorem 4 for more details). Together, these give the following guarantee.

Theorem 4. If Assumption 3 holds, $\mathbf{h}^\star(\cdot) \in \mathcal{H}$, and $\mathcal{F}_0 = \{ f \mid f(\mathbf{x}) = \boldsymbol{\alpha}^\top \mathbf{z}, \ \boldsymbol{\alpha} \in \mathbb{R}^r, \ \|\boldsymbol{\alpha}\| \le c_2 \}$, then there exist constants c_1, c_2 such that the training tasks $f_j^\star$ are $(\Omega(\tilde{\nu}), 0)$-diverse over $\mathcal{F}_0$. Furthermore, if for a sufficiently large constant c_3, $n \ge c_3(d + \log t)$, $m \ge c_3 r$, and $D \le c_3(\min(\sqrt{dr^2}, \sqrt{rm}))$, then with probability at least $1 - 2\delta$:

$$\text{Transfer Learning Risk} \le \tilde{O}\left(\frac{1}{\tilde{\nu}} \left(\sqrt{\frac{dr^2}{nt}} + \sqrt{\frac{r}{n}} \right) + \sqrt{\frac{r}{m}} \right).$$

A naive bound for logistic regression ignoring the training task data would have a guarantee $O(\sqrt{d/m})$. For n and t sufficiently large, the bound in Theorem 4 scales as $\tilde{O}(\sqrt{r/m})$, which is a significant improvement over $O(\sqrt{d/m})$ when $r \ll d$. Note that our result in fact holds with the empirical data-dependent quantities $\text{tr}(\boldsymbol{\Sigma}_{\mathbf{X}})$ and $\sum_{i=1}^r \sigma_i(\boldsymbol{\Sigma}_{\mathbf{X}_j})$ which can be much smaller then their counterparts d, r in Theorem 4, if the data lies on/or close to a low-dimensional subspace[5].

4.2 Multitask Deep Neural Network Regression

We now consider the setting of real-valued neural network regression. Here the task-specific functions are linear maps as before, but the underlying representation is specified by a depth-K vector-valued neural network:

$$\mathbf{h}(\mathbf{x}) = \mathbf{W}_K \sigma_{K-1}(\mathbf{W}_{K-1}(\sigma_{K-2}(\ldots \sigma(\mathbf{W}_1 \mathbf{x})))). \tag{7}$$

Each $\mathbf{W}_k$ is a parameter matrix, and each σ_k is a $\tanh$ activation function. We let $\|\mathbf{W}\|_{1,\infty} = \max_j(\sum_k |\mathbf{W}_{j,k}|)$ and $\|\mathbf{W}\|_{\infty \to 2}$ be the induced ∞-to-2 operator norm. Formally, $\mathcal{F}$ and $\mathcal{H}$ are[6]

$$\mathcal{F} = \{ f \mid f(\mathbf{z}) = \boldsymbol{\alpha}^\top \mathbf{z}, \ \boldsymbol{\alpha} \in \mathbb{R}^r, \ \|\boldsymbol{\alpha}\| \le c_1 M(K)^2 \}, \tag{8}$$
$$\mathcal{H} = \{ \mathbf{h}(\cdot) \in \mathbb{R}^r \text{ in (7) for } \mathbf{W}_k : \|\mathbf{W}_k\|_{1,\infty} \le M(k) \text{ for } k \in [K-1],$$
$$\max(\|\mathbf{W}_K\|_{1,\infty}, \|\mathbf{W}_K\|_{\infty \to 2}) \le M(K), \text{ such that } \sigma_r\left(\mathbb{E}_\mathbf{x}[\mathbf{h}(\mathbf{x})\mathbf{h}(\mathbf{x})^\top]\right) > \Omega(1) \}.$$

We consider the setting where $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}$, and the measure $\mathbb{P}_\mathbf{x}$ is D-bounded. We also let the conditional distribution in (1) be induced by:

$$y = \boldsymbol{\alpha}^\top \mathbf{h}(\mathbf{x}) + \eta \text{ for } \boldsymbol{\alpha}, \mathbf{h} \text{ as in (8)}, \tag{9}$$

with additive noise η bounded almost surely by $O(1)$ and independent of $\mathbf{x}$. We use the standard squared loss $\ell(\boldsymbol{\alpha}^\top \mathbf{h}(\mathbf{x}), y) = (y - \boldsymbol{\alpha}^\top \mathbf{h}(\mathbf{x}))^2$, and let the true training tasks take the form $f_j^\star(\mathbf{z}) = (\boldsymbol{\alpha}_j^\star)^\top \mathbf{z}$ for all $j \in [t]$, and set $\mathbf{A} = (\boldsymbol{\alpha}_1^\star, \ldots, \boldsymbol{\alpha}_t^\star)^\top \in \mathbb{R}^{t \times r}$ as in the previous example. Here we use exactly the same diversity/normalization assumption on the task-specific maps—Assumption 3—as in our logistic regression example.

---

[5] Here $\boldsymbol{\Sigma}_{\bar{\mathbf{X}}}$ denotes the empirical covariance of the data matrix $\bar{\mathbf{X}}$. See Corollary 3 for the formal statement of this sharper, more general result.

[6] For the following we make the standard assumption each parameter matrix $\mathbf{W}_k$ satisfies $\|\mathbf{W}_k\|_{1,\infty} \le M(k)$ for each j in the depth-K network [Golowich et al., 2017], and that the feature map is well-conditioned.

Choosing $\mathcal{F}_0$ and $\mathcal{F}$ appropriately establishes a $(\Omega(\tilde{\nu}), 0)$-diversity as defined in Definition 3 (see Lemma 6). Standard arguments as well as results in Golowich et al. [2017] allow us to bound the Gaussian complexity terms as follows (see the proof of Theorem 5 for details):

$$\mathfrak{G}_N(\mathcal{H}) \leq \tilde{O}\left(\frac{rM(K) \cdot D\sqrt{K} \cdot \Pi_{k=1}^{K-1}M(k)}{\sqrt{N}}\right); \quad \bar{\mathfrak{G}}_N(\mathcal{F}) \leq \tilde{O}\left(\frac{M(K)^3}{\sqrt{N}}\right).$$

Combining these results yields the following end-to-end transfer learning guarantee.

Theorem 5. *If Assumption 3 holds, $\mathbf{h}^\star(\cdot) \in \mathcal{H}$, and $\mathcal{F}_0 = \{\, f \mid f(\mathbf{z}) = \boldsymbol{\alpha}^\top \mathbf{z}, \ \boldsymbol{\alpha} \in \mathbb{R}^r, \ \|\boldsymbol{\alpha}\| \leq c_2\}$, then there exist constants c_1, c_2 such that the training tasks $f_j^\star$ are $(\Omega(\tilde{\nu}), 0)$-diverse over $\mathcal{F}_0$. Further, if $M(K) \geq c_3$ for a universal constant c_3, then with probability at least $1 - 2\delta$:*

$$\textit{Transfer Learning Risk} \leq \tilde{O}\left(\frac{rM(K)^6 \cdot D\sqrt{K} \cdot \Pi_{k=1}^{K-1}M(k)}{\tilde{\nu}\sqrt{nt}} + \frac{M(K)^6}{\tilde{\nu}\sqrt{n}} + \frac{M(K)^6}{\sqrt{m}}\right).$$

The $\mathrm{poly}(M(K))$ dependence of the guarantee on the final-layer weights can likely be improved, but is dominated by the overhead of learning the complex feature map $\mathbf{h}^\star(\cdot)$ which has complexity $\mathrm{poly}(M(K)) \cdot D\sqrt{K} \cdot \Pi_{k=1}^{K-1}M(k)$. By contrast a naive algorithm which does not leverage the training samples would have a sample complexity of $\tilde{O}\left(\mathrm{poly}(M(K)) \cdot D\sqrt{K} \cdot \Pi_{k=1}^{K-1}M(k)/\sqrt{m}\right)$ via a similar analysis. Such a rate can be much larger than the bound in Theorem 5 when $nt \gg m$: exactly the setting relevant to that of few-shot learning for which ImageNet pretraining is often used.

4.3 Multitask Index Models

To illustrate the flexibility of our framework, in our final example, we consider a classical statistical model: the index model, which is often studied from the perspective of semiparametric estimation [Bickel et al., 1993]. As flexible tools for general-purpose, non-linear dimensionality reduction, index models have found broad applications in economics, finance, biology and the social sciences [Bickel et al., 1993, Li and Racine, 2007, Otwinowski et al., 2018]. This class of models has a different flavor then previously considered: the task-specific functions are nonparametric "link" functions, while the underlying representation is a one-dimensional projection. The formal set-up and full transfer generalization guarantees are deferred to Appendix A.1.

5 Conclusion

We present a framework for understanding the generalization abilities of generic models which share a common, underlying representation. In particular, our framework introduces a novel notion of task diversity through which we provide guarantees of a fast convergence rate, decaying with *all of the samples* for the transfer learning problem. One interesting direction for future consideration is investigating the effects of relaxing the common design and realizability assumptions on the results presented here. We also believe extending the results herein to accommodate "fine-tuning" of learned representations – that is, mildly adapting the learned representation extracted from training tasks to new, related tasks – is an important direction for future work.

Broader Impact

As a theoretical paper we do not foresee our work directly having any societal consequences. However, transfer learning is a tool increasingly used in practical machine learning applications. Theoretical explorations related to transfer learning may help provide frameworks through which to reason about, and design, safer and more reliable algorithms.

Acknowledgments and Disclosure of Funding

References

Alexei Baevski, Sergey Edunov, Yinhan Liu, Luke Zettlemoyer, and Michael Auli. Cloze-driven pretraining of self-attention networks. *arXiv preprint arXiv:1903.07785*, 2019.

Peter L Bartlett, Olivier Bousquet, Shahar Mendelson, et al. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

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

Shai Ben-David and Reba Schuller Borbely. A notion of task relatedness yielding provable multiple-task learning guarantees. *Machine learning*, 73(3):273–287, 2008.

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

Peter J Bickel, Chris AJ Klaassen, Peter J Bickel, Ya'acov Ritov, J Klaassen, Jon A Wellner, and YA'Acov Ritov. *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press Baltimore, 1993.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

Giovanni Cavallanti, Nicolo Cesa-Bianchi, and Claudio Gentile. Linear algorithms for online multitask classification. *Journal of Machine Learning Research*, 11(Oct):2901–2934, 2010.

Giulia Denevi, Carlo Ciliberto, Riccardo Grazzi, and Massimiliano Pontil. Learning-to-learn stochastic gradient descent with biased regularization. *arXiv preprint arXiv:1903.10399*, 2019a.

Giulia Denevi, Dimitris Stamos, Carlo Ciliberto, and Massimiliano Pontil. Online-within-online meta-learning. In *Advances in Neural Information Processing Systems*, pages 13089–13099, 2019b.

Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655, 2014.

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

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, and Burkhard Rost. End-to-end multitask learning, from protein language to protein features without alignments. *bioRxiv*, 2020. doi: 10.1101/864405.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org, 2017.

Chelsea Finn, Aravind Rajeswaran, Sham Kakade, and Sergey Levine. Online meta-learning. *arXiv preprint arXiv:1902.08438*, 2019.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. *arXiv preprint arXiv:1712.06541*, 2017.

Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410, 2016.

Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *arXiv preprint arXiv:2004.05439*, 2020.

Sham M Kakade, Varun Kanade, Ohad Shamir, and Adam Kalai. Efficient learning of generalized linear and single index models with isotonic regression. In *Advances in Neural Information Processing Systems*, pages 927–935, 2011.

Mikhail Khodak, Maria-Florina Balcan, and Ameet Talwalkar. Provable guarantees for gradient-based meta-learning. *arXiv preprint arXiv:1902.10644*, 2019a.

Mikhail Khodak, Maria-Florina F Balcan, and Ameet S Talwalkar. Adaptive gradient-based meta-learning methods. In *Advances in Neural Information Processing Systems*, pages 5915–5926, 2019b.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10657–10665, 2019.

Qi Li and Jeffrey Scott Racine. *Nonparametric econometrics: theory and practice*. Princeton University Press, 2007.

Karim Lounici, Massimiliano Pontil, Sara Van De Geer, Alexandre B Tsybakov, et al. Oracle inequalities and optimal inference under group sparsity. *The annals of statistics*, 39(4):2164–2204, 2011.

Pascal Massart et al. About the constants in talagrand's concentration inequalities for empirical processes. *The Annals of Probability*, 28(2):863–884, 2000.

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

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *The Journal of Machine Learning Research*, 17(1):2853–2884, 2016.

Guillaume Obozinski, Martin J Wainwright, Michael I Jordan, et al. Support union recovery in high-dimensional multivariate regression. *The Annals of Statistics*, 39(1):1–47, 2011.

Jakub Otwinowski, David M McCandlish, and Joshua B Plotkin. Inferring the shape of global epistasis. *Proceedings of the National Academy of Sciences*, 115(32):E7550–E7558, 2018.

Massimiliano Pontil and Andreas Maurer. Excess risk bounds for multitask learning with trace norm regularization. In *Conference on Learning Theory*, pages 55–76, 2013.

Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. Rapid learning or feature reuse? towards understanding the effectiveness of maml. *arXiv preprint arXiv:1909.09157*, 2019.

Nilesh Tripuraneni, Chi Jin, and Michael I Jordan. Provable meta-learning of linear representations. *arXiv preprint arXiv:2002.11684*, 2020.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

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

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014.

Appendices

Notation: Here we introduce several additional pieces of notation we will use throughout.

We use $\mathbb{E}_{\mathbf{x}}[\cdot]$ to refer to the expectation operator taken over the randomness in the vector $\mathbf{x}$ sampled from a distribution $\mathbb{P}_{\mathbf{x}}$. Throughout we will use $\mathcal{F}$ to refer exclusively to a scalar-valued function class of tasks and $\mathcal{H}$ to a vector-valued function class of features. For $\mathcal{F}$, we use $\mathcal{F}^{\otimes t}$ to refer its t-fold Cartesian product such that $(f_1, \ldots, f_t) \equiv \mathbf{f} \in \mathcal{F}^{\otimes t}$ for $f_j \in \mathcal{F}, j \in [t]$. We use $f(\mathbf{h})$ as shorthand for the function composition, $f \circ \mathbf{h}$. Similarly, we define the composed function class $\mathcal{F}(\mathcal{H}) = \{f(\mathbf{h}) : f \in \mathcal{F}, \mathbf{h} \in \mathcal{H}\}$ and its vector-valued version $\mathcal{F}^{\otimes t}(\mathcal{H}) = \{(f_1(\mathbf{h}), \ldots, f_t(\mathbf{h})) : f_j \in \mathcal{F}, j \in [t], \mathbf{h} \in \mathcal{H}\}$ with this shorthand. We will use $\gtrsim, \lesssim$, and $\asymp$ to denote greater than, less than, and equal to up to a universal constant and use $\tilde{O}$ to denote an expression that hides polylogarithmic factors in all problem parameters.

In the context of the two-stage ERM procedure introduced in Section 2 we let the design matrix and responses y_{ji} for the jth task be $\mathbf{X}_j$ and $\mathbf{y}_j$ for $j \in [t] \cup \{0\}$, and the entire design matrix and responses concatenated over all $j \in [t]$ tasks as $\mathbf{X}$ and $\mathbf{y}$ respectively. Given a design matrix $\bar{\mathbf{X}} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^\top$ (comprised of mean-zero random vectors) we will let $\boldsymbol{\Sigma}_{\bar{\mathbf{X}}} = \frac{1}{N}\bar{\mathbf{X}}^\top\bar{\mathbf{X}}$ denote its corresponding empirical covariance.

Recall we define the notions of the empirical and population Gaussian complexity for a generic vector-valued function class $\mathcal{Q}$ containing functions $\mathbf{q}(\cdot) : \mathbb{R}^d \to \mathbb{R}^r$, and data matrix $\mathbf{X}$ with N datapoints as,

$$\hat{\mathfrak{G}}_{\mathbf{X}}(\mathcal{Q}) = \mathbb{E}_{\mathbf{g}}[\sup_{\mathbf{q} \in \mathcal{Q}} \frac{1}{N}\sum_{k=1}^{r}\sum_{i=1}^{N} g_{ki}q_k(\mathbf{x}_i)], \qquad \mathfrak{G}_N(\mathcal{Q}) = \mathbb{E}_{\mathbf{X}}[\hat{\mathfrak{G}}_{\mathbf{X}}(\mathcal{Q})] \qquad g_{ki} \sim \mathcal{N}(0,1) \ i.i.d.,$$

where for the latter population Gaussian complexity each its N datapoints are drawn from the $\mathbb{P}_{\mathbf{x}}(\cdot)$ design distribution. Analogously to the above we can define the empirical and population Rademacher complexities for generic vector-valued functions as,

$$\hat{\mathfrak{R}}_{\mathbf{X}}(\mathcal{Q}) = \mathbb{E}_{\boldsymbol{\epsilon}}[\sup_{\mathbf{q} \in \mathcal{Q}} \frac{1}{N}\sum_{k=1}^{r}\sum_{i=1}^{N} \epsilon_{ki}q_k(\mathbf{x}_i)], \qquad \mathfrak{R}_N(\mathcal{Q}) = \mathbb{E}_{\mathbf{X}}[\hat{\mathfrak{R}}_{\mathbf{X}}(\mathcal{Q})] \qquad \epsilon_{ki} \sim \text{Rad}(\frac{1}{2}) \ i.i.d.$$

A Further Examples

Here we present further examples to exhibit the generality of our framework.

A.1 Multitask Index Models

To illustrate the flexibility of our framework, in our final example, we consider a classical statistical model: the index model, which is often studied from the perspective of semiparametric estimation [Bickel et al., 1993]. As flexible tools for general-purpose, non-linear dimensionality reduction, index models have found broad applications in economics, finance, biology and the social sciences [Bickel et al., 1993, Li and Racine, 2007, Otwinowski et al., 2018]. This class of models has a different flavor then previously considered: the task-specific functions are nonparametric "link" functions, while the underlying representation is a one-dimensional projection. Formally, let the function classes $\mathcal{F}$ and $\mathcal{H}$ be:

$$\mathcal{F} = \{ f \mid f(\mathbf{z}) \text{ is a 1-Lipschitz, monotonic function bounded in } [0,1]\}, \tag{10}$$
$$\mathcal{H} = \{ \mathbf{h} \mid \mathbf{h}(\mathbf{x}) = \mathbf{b}^\top\mathbf{x}, \ \mathbf{b} \in \mathbb{R}^d, \ \|\mathbf{b}\| \le W\}.$$

We consider the setting where $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}$, the measure $\mathbb{P}_{\mathbf{x}}$ is D-bounded, and $DW \ge 1$. This matches the setting in Kakade et al. [2011]. The conditional distribution in (1) is induced by:

$$y = f(\mathbf{b}^\top\mathbf{x}) + \eta \ \text{ for } f, \mathbf{b} \text{ as in } (7),$$

with additive noise η bounded almost surely by $O(1)$ and independent of $\mathbf{x}$. We use the robust ℓ_1 loss, $\ell(f(\mathbf{b}^\top\mathbf{x}), y) = |y - f(\mathbf{b}^\top\mathbf{x})|$, in this example. Now, define $\mathcal{F}_t = \text{conv}\{f_1^\star, \ldots, f_t^\star\}$ as the

convex hull of the training task-specific functions $f_j^\star$. Given this, we define the $\tilde{\epsilon}$-enlargement of $\mathcal{F}_t$ by $\mathcal{F}_{t,\tilde{\epsilon}} = \{f : \exists \tilde{f} \in \mathcal{F}_t \text{ such that } \sup_z |f(z) - \tilde{f}(z)| \leq \tilde{\epsilon}\}$.

We prove a transfer generalization bound for $\mathcal{F}_0 = \mathcal{F}_{t,\tilde{\epsilon}}$, for which we can establish $(\tilde{\nu}, \tilde{\epsilon})$-diversity with $\tilde{\nu} \geq \frac{1}{t}$ as defined in Definition 3 (see Lemma 7). Standard arguments once again show that $\mathfrak{G}_N(\mathcal{H}) \leq O\left(\sqrt{(W^2 \mathbb{E}_{\mathbf{X}}[\mathrm{tr}(\mathbf{\Sigma}_{\mathbf{X}})]/N)}\right)$ and $\bar{\mathfrak{G}}_N(\mathcal{F}) \leq O\left(\sqrt{WD/N}\right)$ (see the proof of Theorem 6 for details). Together these give the following guarantee.

Theorem 6. *If $f_j^\star \in \mathcal{F}$ for $j \in [t]$, $\mathbf{h}^\star(\cdot) \in \mathcal{H}$, and $f_0^\star \in \mathcal{F}_0 = \mathcal{F}_{t,\tilde{\epsilon}}$, then the training tasks are $(\tilde{\nu}, \tilde{\epsilon})$-diverse over $\mathcal{F}_0$ where $\tilde{\nu} \geq \frac{1}{t}$. Further, with probability at least $1 - 2\delta$:*

$$\textit{Transfer Learning Risk} \leq \tilde{O}\left(\frac{1}{\tilde{\nu}} \cdot \left(\sqrt{\frac{W^2 \mathbb{E}_{\mathbf{X}}[\mathrm{tr}(\mathbf{\Sigma}_{\mathbf{X}})]}{nt}} + \sqrt{\frac{WD}{n}}\right) + \sqrt{\frac{WD}{m}}\right) + \tilde{\epsilon}.$$

As before, the complexity of learning the feature representation decays as $n \to \infty$. Hence if $\mathbb{E}[\mathrm{tr}(\mathbf{\Sigma}_{\mathbf{X}})]$ is large, the aforementioned bound will provide significant savings over the bound which ignores the training phase samples of $O\left(\sqrt{(W^2 \mathbb{E}_{\mathbf{X}}[\mathrm{tr}(\mathbf{\Sigma}_{\mathbf{X}})])/m}\right) + O(\sqrt{WD/m})$. In this example, the problem-dependent parameter $\tilde{\nu}$ does not have a simple linear-algebraic interpretation. Indeed, in the worst-case it may seem the aforementioned bound degrades with t[7]. However, note that $\mathcal{F}_0 = \mathcal{F}_{t,\tilde{\epsilon}}$, so those unseen tasks which we hope to transfer to itself *grows* with t unlike in the previous examples. The difficulty of the transfer learning problem also increases as t increases. Finally, this example utilizes the full power of (ν, ϵ)-diversity by permitting robust generalization to tasks outside $\mathcal{F}_t$, at the cost of a bias term $\tilde{\epsilon}$ in the generalization guarantee.

B   Proofs in Section 3

Here we include the proofs of central generalization guarantees and the Gaussian process chain rule used in its proof.

B.1   Training Phase/Test Phase Proofs

In all the following definitions $(\mathbf{x}_j, y_j)$ refer to datapoint drawn from the jth component of the model in (1). We first include the proof of Theorem 1 which shows that minimizing the training phase ERM objective controls the task-average distance between the underlying feature representation $\mathbf{h}$ and learned feature representation $\hat{\mathbf{h}}$.

Proof of Theorem 1. For fixed $\mathbf{f}', \mathbf{h}'$, define the centered training risk as,

$$L(\mathbf{f}', \mathbf{h}', \mathbf{f}^\star, \mathbf{h}^\star) = \frac{1}{t}\sum_{j=1}^{t} \mathbb{E}_{\mathbf{x}_j, y_j}\left\{\ell(f_j' \circ \mathbf{h}'(\mathbf{x}_j), y_j) - \ell(f_j^\star \circ \mathbf{h}^\star(\mathbf{x}_j), y_j)\right\}.$$

and its empirical counterpart,

$$\hat{L}(\mathbf{f}', \mathbf{h}', \mathbf{f}^\star, \mathbf{h}^\star) = \frac{1}{t}\sum_{j=1}^{t}\sum_{i=1}^{n}\left\{\ell(f_j' \circ \mathbf{h}'(\mathbf{x}_{ji}), y_{ji}) - \mathbb{E}_{\mathbf{x},y}[\ell(f_j^\star \circ \mathbf{h}^\star(\mathbf{x}), y)]\right\}$$

Now if $\tilde{\mathbf{f}}$ denotes a minimizer of the former expression for fixed $\hat{\mathbf{h}}$, in the sense that $\tilde{\mathbf{f}} = \frac{1}{t}\sum_{j=1}^{t} \arg\inf_{f_j' \in \mathcal{F}} \mathbb{E}_{\mathbf{x}_j, y_j}\left\{\ell(f_j' \circ \hat{\mathbf{h}}(\mathbf{x}_j), y_j) - \ell(f_j^\star \circ \mathbf{h}^\star(\mathbf{x}_j), y_j)\right\}$, then by definition, we have that $\bar{d}_{\mathcal{F}, \mathbf{f}^\star}(\hat{\mathbf{h}}; \mathbf{h}^\star)$ equals the former expression. We first decompose the average distance using the pair $(\hat{\mathbf{f}}, \hat{\mathbf{h}})$. Recall the pair $(\hat{\mathbf{f}}, \hat{\mathbf{h}})$ refers to the empirical risk minimizer in (2).

$$L(\tilde{\mathbf{f}}, \hat{\mathbf{h}}, \mathbf{f}^\star, \mathbf{h}^\star) - L(\mathbf{f}^\star, \mathbf{h}^\star, \mathbf{f}^\star, \mathbf{h}^\star) = \underbrace{L(\tilde{\mathbf{f}}, \hat{\mathbf{h}}, \mathbf{f}^\star, \mathbf{h}^\star) - L(\hat{\mathbf{f}}, \hat{\mathbf{h}}, \mathbf{f}^\star, \mathbf{h}^\star)}_{a} + L(\hat{\mathbf{f}}, \hat{\mathbf{h}}, \mathbf{f}^\star, \mathbf{h}^\star) - L(\mathbf{f}^\star, \mathbf{h}^\star, \mathbf{f}^\star, \mathbf{h}^\star)$$

---

[7]Note as $\tilde{\nu}$ is problem-dependent, for a given underlying $\mathbf{f}^\star, \mathbf{h}^\star, \mathcal{F}_0$ problem instance, $\tilde{\nu}$ may be significantly greater than $\frac{1}{t}$. See the proof of Lemma 7 for details.

Note that by definition of the $\tilde{\mathbf{f}}$, $a \leq 0$. The second pair can be controlled via the canonical risk decomposition,

$$L(\hat{\mathbf{f}}, \hat{\mathbf{h}}, \mathbf{f}^\star, \mathbf{h}^\star) - L(\mathbf{f}^\star, \mathbf{h}^\star, \mathbf{f}^\star, \mathbf{h}^\star) = \underbrace{L(\hat{\mathbf{f}}, \hat{\mathbf{h}}, \mathbf{f}^\star, \mathbf{h}^\star) - \hat{L}(\hat{\mathbf{f}}, \hat{\mathbf{h}}, \mathbf{f}^\star, \mathbf{h}^\star)}_{b} + \underbrace{\hat{L}(\hat{\mathbf{f}}, \hat{\mathbf{h}}, \mathbf{f}^\star, \mathbf{h}^\star) - \hat{L}(\mathbf{f}^\star, \mathbf{h}^\star, \mathbf{f}^\star, \mathbf{h}^\star)}_{c} +$$

$$\underbrace{\hat{L}(\mathbf{f}^\star, \mathbf{h}^\star, \mathbf{f}^\star, \mathbf{h}^\star) - L(\mathbf{f}^\star, \mathbf{h}^\star, \mathbf{f}^\star, \mathbf{h}^\star)}_{d}.$$

By definition $c \leq 0$ (note this inequality uses the realizability in Assumption 2) and $b, d \leq \sup_{\mathbf{f} \in \mathcal{F}^{\otimes t}, \mathbf{h} \in \mathcal{H}} |R_{\text{train}}(\mathbf{f}, \mathbf{h}) - \hat{R}_{\text{train}}(\mathbf{f}, \mathbf{h})|$. By an application of the bounded differences inequality and a standard symmetrization argument (see for example Wainwright [2019, Theorem 4.10] we have that,

$$\sup_{\mathbf{f} \in \mathcal{F}^{\otimes t}, \mathbf{h} \in \mathcal{H}} |R_{\text{train}}(\mathbf{f}, \mathbf{h}) - \hat{R}_{\text{train}}(\mathbf{f}, \mathbf{h})| \leq 2\mathfrak{R}_{nt}(\ell(\mathcal{F}^{\otimes t}(\mathcal{H}))) + 2B\sqrt{\frac{\log(1/\delta)}{nt}}$$

with probability at least $1 - 2\delta$.

It remains to decompose the leading Rademacher complexity term. First we center the functions to $\ell_{ji}(f_j \circ \mathbf{h}(\mathbf{x}_{ji}), y_{ji}) = \ell(f_j \circ \mathbf{h}(\mathbf{x}_{ji}), y_{ji}) - \ell(0, y_{ji})$. Then noting $|\ell_{ji}(0, y_{ji})| \leq B$, the constant-shift property of Rademacher averages Wainwright [2019, Exercise 4.7c] gives,

$$\mathbb{E}_{\boldsymbol{\epsilon}}\Big[\sup_{\mathbf{f} \in \mathcal{F}^{\otimes t}, \mathbf{h} \in \mathcal{H}} \frac{1}{nt} \sum_{j=1}^{t} \sum_{i=1}^{n} \epsilon_{ij} \ell(f_j \circ \mathbf{h}(\mathbf{x}_{ji}), y_{ji})\Big] \leq \mathbb{E}_{\boldsymbol{\epsilon}}\Big[\sup_{\mathbf{f} \in \mathcal{F}, \mathbf{h} \in \mathcal{H}} \frac{1}{nt} \sum_{j=1}^{t} \sum_{i=1}^{n} \epsilon_{ij} \ell_{ij}(f_j \circ \mathbf{h}(\mathbf{x}_{ji}), y_{ji})\Big] + \frac{B}{\sqrt{nt}}$$

Now note each $\ell_{ij}(\cdot, \cdot)$ is L-Lipschitz in its first coordinate uniformly for every choice of the second coordinate (and by construction centered in its first coordinate). So, defining the set $S = \{(f_1 \circ \mathbf{h}(\mathbf{x}_{1i}), \ldots, f_j \circ \mathbf{h}(\mathbf{x}_{ji}), \ldots, f_t \circ \mathbf{h}(\mathbf{x}_{ti}))) : j \in [t], f_j \in \mathcal{F}, \mathbf{h} \in \mathcal{H}\} \subseteq \mathbb{R}^{tn}$, and applying the contraction principle Ledoux and Talagrand [2013, Theorem 4.12] over this set shows,

$$\mathbb{E}_{\boldsymbol{\epsilon}}\Big[\sup_{\mathbf{f} \in \mathcal{F}^{\otimes t}, \mathbf{h} \in \mathcal{H}} \frac{1}{nt} \sum_{j=1}^{t} \sum_{i=1}^{n} \epsilon_{ij} \ell_{ij}(f_j \circ \mathbf{h}(\mathbf{x}_{ji}), y_{ji})\Big] \leq 2L \cdot \mathfrak{R}_{nt}(\mathcal{F}^{\otimes t}(\mathcal{H})). \tag{11}$$

Combining gives,

$$\sup_{\mathbf{f} \in \mathcal{F}^{\otimes t}, \mathbf{h} \in \mathcal{H}} |R_{\text{train}}(\mathbf{f}, \mathbf{h}) - \hat{R}_{\text{train}}(\mathbf{f}, \mathbf{h})| \leq 4L \cdot \mathfrak{R}_{nt}(\mathcal{F}^{\otimes t}(\mathcal{H})) + \frac{4B\sqrt{\log(1/\delta)}}{\sqrt{nt}}$$

with probability $1 - 2\delta$. Now note by [Ledoux and Talagrand, 2013, p.97] empirical Rademacher complexity is upper bounded by empirical Gaussian complexity: $\hat{\mathfrak{R}}_{\mathbf{X}}(\mathcal{F}^{\otimes t}(\mathcal{H})) \leq \sqrt{\frac{\pi}{2}} \hat{\mathfrak{G}}_{\mathbf{X}}(\mathcal{F}^{\otimes t}(\mathcal{H}))$. Taking expectations of this and combining with the previous display yields the first inequality in the theorem statement.

The last remaining step hinges on Theorem 7 to decompose the Gaussian complexity over $\mathcal{F}$ and $\mathcal{H}$. A direct application of Theorem 7 gives the conclusion that,

$$\hat{\mathfrak{G}}_{\mathbf{X}}(\mathcal{F}^{\otimes t}(\mathcal{H})) \leq 128\left(\frac{D_{\mathbf{X}}}{(nt)^2} + C(\mathcal{F}^{\otimes t}(\mathcal{H})) \cdot \log(nt)\right)$$

where $C(\mathcal{F}^{\otimes t}(\mathcal{H}); \mathbf{X}) = L(\mathcal{F}) \cdot \hat{\mathfrak{G}}_{\mathbf{X}}(\mathcal{H}) + \max_{\mathbf{Z} \in \mathcal{Z}} \hat{\mathfrak{G}}_{\mathbf{Z}}(\mathcal{F})$ where $\mathcal{Z} = \{\mathbf{h}(\bar{\mathbf{X}}) : \mathbf{h} \in \mathcal{H}, \bar{\mathbf{X}} \in \cup_{j=1}^{t}\{\mathbf{X}_j\}\}$. By definition of $D_{\mathbf{X}}$ we have $D_{\mathbf{X}} \leq 2D_{\mathcal{X}}$ and similarly that $\max_{\mathbf{Z} \in \mathcal{Z}} \hat{\mathfrak{G}}_{\mathbf{Z}}(\mathcal{F}) \leq \max_{\mathbf{Z} \in \mathcal{Z}_1} \hat{\mathfrak{G}}_{\mathbf{Z}}(\mathcal{F})$ for $\mathcal{Z}_1 = \{(\mathbf{h}(\mathbf{x}_1), \cdots, \mathbf{h}(\mathbf{x}_n)) \mid \mathbf{h} \in \mathcal{H}, \mathbf{x}_i \in \mathcal{X} \text{ for all } i \in [n]\}$. Taking expectations over $\mathbf{X}$ in this series of relations and assembling the previous bounds gives the conclusion after rescaling δ. $\qquad\square$

An analogous statement holds both in terms of a sharper notion of the worst-case Gaussian complexity and in terms of empirical Gaussian complexities.

14

Corollary 1. *In the setting of Theorem 1,*

$$\bar{d}_{\mathcal{F},\mathbf{f}^\star}(\hat{\mathbf{h}};\mathbf{h}^\star) \le 4096L\left[\frac{D_\mathcal{X}}{(nt)^2} + \log(nt)\cdot[L(\mathcal{F})\cdot\mathfrak{G}_\mathbf{X}(\mathcal{H}) + \mathbb{E}_\mathbf{X}[\max_{\mathbf{Z}\in\mathcal{Z}}\hat{\mathfrak{G}}_\mathbf{Z}(\mathcal{F})]\right] + 8B\sqrt{\frac{\log(1/\delta)}{n}}$$

with probability $1 - 2\delta$ *for* $\mathcal{Z} = \{\mathbf{h}(\bar{\mathbf{X}}) : \mathbf{h}\in\mathcal{H}, \bar{\mathbf{X}}\in\cup_{j=1}^{t}\{\mathbf{X}_j\}\}$. *Furthermore,*

$$\bar{d}_{\mathcal{F},\mathbf{f}^\star}(\hat{\mathbf{h}};\mathbf{h}^\star) \le 16\hat{\mathfrak{G}}_\mathbf{X}(\mathcal{F}^{\otimes t}(\mathcal{H})) + 16B\sqrt{\frac{\log(1/\delta)}{n}} \le$$

$$4096L\left[\frac{D_\mathcal{X}}{(nt)^2} + \log(nt)\cdot[L(\mathcal{F})\cdot\hat{\mathfrak{G}}_\mathbf{X}(\mathcal{H}) + \max_{\mathbf{Z}\in\mathcal{Z}}\hat{\mathfrak{G}}_\mathbf{Z}(\mathcal{F})]\right] + 16B\sqrt{\frac{\log(1/\delta)}{n}}$$

with probability at least $1 - 4\delta$.

Proof. The argument follows analogously to the proof of Theorem 1. The first statement follows identically by avoiding the relaxation–$\max_{\mathbf{Z}\in\mathcal{Z}}\hat{\mathfrak{G}}_\mathbf{Z}(\mathcal{F}) \le \max_{\mathbf{Z}\in\mathcal{Z}_1}\hat{\mathfrak{G}}_\mathbf{Z}(\mathcal{F})$ for $\mathcal{Z}_1 = \{(\mathbf{h}(\mathbf{x}_1),\cdots,\mathbf{h}(\mathbf{x}_n)) \mid \mathbf{h}\in\mathcal{H}, \mathbf{x}_i\in\mathcal{X} \text{ for all } i\in[n]\}$–after applying Theorem 7 in the proof of Theorem 1.

The second statement also follows by a direct modification of the proof of Theorem 1. In the proof another application of the bounded differences inequality would show that $|\mathfrak{R}_{nt}(\mathcal{F}^{\otimes t}(\mathcal{H})) - \hat{\mathfrak{R}}_\mathbf{X}((\mathcal{F}^{\otimes t}(\mathcal{H}))| \le 4B\sqrt{\frac{\log(1/\delta)}{nt}}$ with probability $1 - 2\delta$. Applying this inequality after (11) and union bounding over this event and the event in the theorem, followed by the steps in Theorem 1, gives the result after an application of Theorem 7. $\square$

We now show how the definition of task diversity in Definition 3 and minimizing the training phase ERM objective allows us to transfer a fixed feature representation $\hat{\mathbf{h}}$ and generalize to a new task-specific mapping f_0.

Proof of Theorem 2. Note $\tilde{f}_0 = \operatorname{argmin}_{f\in\mathcal{F}} R_{\text{test}}(f,\hat{\mathbf{h}})$–it is a minimizer of the population test risk loaded with the fixed feature representation $\hat{\mathbf{h}}$. The approach to controlling this term uses the canonical risk decomposition,

$$R_{\text{test}}(\hat{f}_0,\hat{\mathbf{h}}) - R_{\text{test}}(\tilde{f}_0,\hat{\mathbf{h}}) = \underbrace{R_{\text{test}}(\hat{f}_0,\hat{\mathbf{h}}) - \hat{R}_{\text{test}}(\hat{f}_0,\hat{\mathbf{h}})}_{a} + \underbrace{\hat{R}_{\text{test}}(\hat{f}_0,\hat{\mathbf{h}}) - \hat{R}_{\text{test}}(\tilde{f}_0,\hat{\mathbf{h}})}_{b} + \underbrace{\hat{R}_{\text{test}}(\tilde{f}_0,\hat{\mathbf{h}}) - R_{\text{test}}(\tilde{f}_0,\hat{\mathbf{h}})}_{c}$$

First by definition, $b \le 0$. Now a standard uniform convergence/symmetrization argument which also follows the same steps as in the proof of Theorem 1,

$$a + c \le 16L\cdot\mathbb{E}_{\mathbf{X}_0}[\hat{\mathfrak{G}}_{\mathbf{Z}_{\hat{\mathbf{h}}}}(\mathcal{F})] + 8B\sqrt{\frac{\log(1/\delta)}{m}} \le 16L\max_{\hat{\mathbf{h}}\in\mathcal{H}}\mathbb{E}_{\mathbf{X}_0}[\hat{\mathfrak{G}}_{\mathbf{Z}_{\hat{\mathbf{h}}}}(\mathcal{F})] + 8B\sqrt{\frac{\log(1/\delta)}{m}}$$

for $\mathbf{Z}_{\hat{\mathbf{h}}} = \hat{\mathbf{h}}(\mathbf{X}_0)$, with probability at least $1 - 2\delta$. The second inequality simply uses the fact that the map $\hat{\mathbf{h}}$ is fixed, and independent of the randomness in the test data. The bias from using an imperfect feature representation $\hat{\mathbf{h}}$ in lieu of $\mathbf{h}$ arises in $R_{\text{test}}(\tilde{f}_0,\hat{\mathbf{h}})$. For this term,

$$R_{\text{test}}(\tilde{f}_0,\hat{\mathbf{h}}) - R_{\text{test}}(f_0,\mathbf{h}^\star) = \inf_{\tilde{f}_0\in\mathcal{F}}\{R_{\text{test}}(\tilde{f}_0,\hat{\mathbf{h}}) - R_{\text{test}}(f_0,\mathbf{h}^\star)\} \le \sup_{f_0\in\mathcal{F}_0}\inf_{\tilde{f}_0\in\mathcal{F}}\{L(\tilde{f}_0,\hat{\mathbf{h}}) - L(f_0,\mathbf{h}^\star)\} =$$

$$d_{\mathcal{F},\mathcal{F}_0}(\mathbf{h};\hat{\mathbf{h}})$$

To obtain the final theorem statement we use an additional relaxation on the Gaussian complexity term for ease of presentation,

$$\max_{\hat{\mathbf{h}}\in\mathcal{H}}\mathbb{E}_{\mathbf{X}_0}[\hat{\mathfrak{G}}_{\mathbf{Z}_{\hat{\mathbf{h}}}}(\mathcal{F})] \le \bar{\mathfrak{G}}_m(\mathcal{F}).$$

Combining terms gives the conclusion. $\square$

We also present a version of Theorem 2 which can possess better dependence on the boundedness parameter in the noise terms and has data-dependence in the Gaussian complexities. As before our guarantees can be stated both in terms of population or empirical quantities. The result appeals to the functional Bernstein inequality instead of the bounded differences inequality in the concentration step. Although we only state (and use) this guarantee for the test phase generalization an analogous statement can be shown to hold for Theorem 1. Throughout the following, we use $(\mathbf{x}_i, y_i) \sim \mathbb{P}_{f_0 \circ \mathbf{h}}$ for $i \in [m]$ for ease of notation.

Corollary 2. *In the setting of Theorem 2, assuming the loss function ℓ satisfies the centering $\ell(0, y) = 0$ for all $y \in \mathcal{Y}$,*

$$R_{test}(\hat{f}_0, \hat{\mathbf{h}}) - R_{test}(f_0^\star, \mathbf{h}^\star) \le d_{\mathcal{F}, \mathcal{F}_0}(\hat{\mathbf{h}}; \mathbf{h}^\star) + 16L \cdot \mathbb{E}_{\mathbf{X}_0}[\hat{\mathfrak{G}}_{\mathbf{Z}_{\hat{\mathbf{h}}}}(\mathcal{F})] + 4\sigma\sqrt{\frac{\log(2/\delta)}{m}} + 50B\frac{\log(2/\delta)}{m}$$

for $\mathbf{Z}_{\hat{\mathbf{h}}} = \hat{\mathbf{h}}(\mathbf{X}_0)$, with probability at least $1 - \delta$. Here the maximal variance $\sigma^2 = \frac{1}{m}\sup_{f \in \mathcal{F}}\sum_{i=1}^m Var(\ell(f \circ \hat{\mathbf{h}}(\mathbf{x}_i), y_i))$. Similarly we have that,

$$R_{test}(\hat{f}_0, \hat{\mathbf{h}}) - R_{test}(f_0^\star, \mathbf{h}^\star) \le d_{\mathcal{F}, \mathcal{F}_0}(\hat{\mathbf{h}}; \mathbf{h}^\star) + 32L \cdot \hat{\mathfrak{G}}_{\mathbf{Z}_{\hat{\mathbf{h}}}}(\mathcal{F}) + 8\sigma\sqrt{\frac{\log(2/\delta)}{m}} + 100B\frac{\log(2/\delta)}{m}$$

with probability at least $1 - 2\delta$.

Proof of Corollary 2. The proof is identical to the proof of Theorem 2 save in how the concentration argument is performed. Namely in the notation of Theorem 2, we upper bound,

$$a + c \le 2\sup_{f \in \mathcal{F}}|\hat{R}_{\text{test}}(f, \hat{\mathbf{h}}) - R_{\text{test}}(f, \hat{\mathbf{h}})| = 2Z$$

Note by definition $\mathbb{E}_{\mathbf{X}_0, \mathbf{y}_0}[\hat{R}_{\text{test}}(f, \hat{\mathbf{h}})] = R_{\text{test}}(f, \hat{\mathbf{h}})$, where $\hat{R}_{\text{test}}(f, \hat{\mathbf{h}}) = \frac{1}{m}\sum_{i=1}^m \ell(f \circ \hat{\mathbf{h}}(\mathbf{x}_i), y_i)$, and the expectation is taken over the test-phase data. Instead of applying the bounded differences inequality to control the fluctuations of this term we apply a powerful form of the functional Bernstein inequality due to Massart et al. [2000]. Applying Massart et al. [2000, Theorem 3] therein, we can conclude,

$$Z \le (1 + \epsilon)\mathbb{E}[Z] + \frac{\sigma}{\sqrt{n}}\sqrt{2\kappa \log(\frac{1}{\delta})} + \kappa(\epsilon)\frac{B}{m}\log(\frac{1}{\delta})$$

for $\kappa = 2$, $\kappa(\epsilon) = 2.5 + \frac{32}{\epsilon}$ and $\sigma^2 = \frac{1}{m}\sup_{f \in \mathcal{F}}\sum_{i=1}^m Var(\ell(f \circ \hat{\mathbf{h}}(\mathbf{x}_i), y_i))$. We simply take $\epsilon = 1$ for our purposes, which gives the bound,

$$Z \le 2\mathbb{E}[Z] + 4\frac{\sigma}{\sqrt{m}}\sqrt{\log(\frac{1}{\delta})} + 35\frac{B}{m}\log(\frac{1}{\delta})$$

Next note a standard symmetrization argument shows that $\mathbb{E}[Z] \le 2\mathbb{E}_{\mathbf{X}_0, \mathbf{y}_0}[\hat{\mathfrak{R}}_{\mathbf{Z}_{\hat{\mathbf{h}}}}(\ell \circ \mathcal{F})]$ for $\mathbf{Z}_{\hat{\mathbf{h}}} = \hat{\mathbf{h}}(\mathbf{X}_0)$. Following the proof of Theorem 2 but eschewing the unnecessary centering step in the application of the contraction principle shows that, $\hat{\mathfrak{R}}_{\mathbf{Z}_{\hat{\mathbf{h}}}}(\ell \circ \mathcal{F}) \le 2L \cdot \hat{\mathfrak{R}}_{\mathbf{Z}_{\hat{\mathbf{h}}}}(\mathcal{F})$. Upper bounding empirical Rademacher complexity by Gaussian complexity and following the steps of Theorem 2 gives the first statement.

The second statement in terms of empirical quantities follows similarly. First the population Rademacher complexity can be converted into an empirical Rademacher complexity using a similar concentration inequality based result which appears in a convenient form in Bartlett et al. [2005, Lemma A.4 (i)]. Directly applying this result (with $\alpha = \frac{1}{2}$) shows that,

$$\mathbb{E}_{\mathbf{X}_0, \mathbf{y}_0}[\hat{\mathfrak{R}}_{\mathbf{Z}_{\hat{\mathbf{h}}}}(\ell \circ \mathcal{F})] \le 2\hat{\mathfrak{R}}_{\mathbf{Z}_{\hat{\mathbf{h}}}}(\ell \circ \mathcal{F}) + \frac{8B\log(\frac{1}{\delta})}{m}$$

with probability at least $1 - \delta$. The remainder of the argument follows exactly as before and as in the proof of Theorem 2 along with another union bound. $\square$

The proof of Theorem 3 is almost immediate.

Proof of Theorem 3. The result follows immediately by combining Theorem 1, Theorem 2, and the definition of task diversity along with a union bound over the two events on which Theorems 1 and 2 hold. $\square$

B.2 A User-Friendly Chain Rule for Gaussian Complexity

We provide the formal statement and the proof of the chain rule for Gaussian complexity that is used in the main text to decouple the complexity of learning the class $\mathcal{F}^{\otimes t}(\mathcal{H})$ into the complexity of learning each individual class. We believe this result may be a technical tool that is of more general interest for a variety of learning problems where compositions of function classes naturally arise.

Intuitively, the chain rule (Theorem 7) can be viewed as a generalization of the Ledoux-Talagrand contraction principle which shows that for a *fixed*, centered L-Lipschitz function ϕ, $\hat{\mathfrak{G}}_{\mathbf{X}}(\phi(\mathcal{F})) \leq 2L\hat{\mathfrak{G}}_{\mathbf{X}}(\mathcal{F})$. However, as we are learning *both* $\mathbf{f} \in \mathcal{F}^{\otimes t}$ (which is not fixed) and $\mathbf{h} \in \mathcal{H}$, $\hat{\mathfrak{G}}_{\mathbf{X}}(\mathcal{F}^{\otimes t} \circ \mathcal{H})$ features a suprema over both $\mathcal{F}^{\otimes t}$ and $\mathcal{H}$.

A comparable result for Gaussian processes to our Theorem 7 is used in Maurer et al. [2016] for multi-task learning applications, drawing on the chain rule of Maurer [2016]. Although their result is tighter with respect to logarithmic factors, it cannot be written purely in terms of Gaussian complexities. Rather, it includes a worst-case "Gaussian-like" average (Maurer et al. [2016, Eq. 4]) in lieu of $\hat{\mathfrak{G}}_{\mathcal{Z}}(\mathcal{F})$ in Theorem 7. In general, it is not clear how to sharply bound this term beyond the using existing tools in the learning theory literature. The terms appearing in Theorem 7 can be bounded, in a direct and modular fashion, using the wealth of existing results and tools in the learning theory literature.

Our proof technique and that of Maurer [2016] both hinge on several properties of Gaussian processes. Maurer [2016] uses a powerful generalization of the Talagrand majorizing measure theorem to obtain their chain rule. We take a different path. First we use the entropy integral to pass to the space of covering numbers–where the metric properties of the distance are used to decouple the features and tasks. Finally an appeal to Gaussian process lower bounds are used to come back to expression that involves only Gaussian complexities.

We will use the machinery of empirical process theory throughout this section so we introduce several useful definitions we will need. We define the empirical ℓ_2-norm as, $d_{2,\mathbf{X}}^2(\mathbf{f}(\mathbf{h}), \mathbf{f}'(\mathbf{h}')) = \frac{1}{t \cdot n} \sum_{j=1}^{t} \sum_{i=1}^{n} (f_j(\mathbf{h}(\mathbf{x}_{ji})) - f_j'(\mathbf{h}'(\mathbf{x}_{ji})))^2$, and the corresponding u-covering number as $N_{2,\mathbf{X}}(u; d_{2,\mathbf{X}}, \mathcal{F}^{\otimes t}(\mathcal{H}))$. Further, we can define the *worst-case ℓ_2-covering number* as $N_2(u; \mathcal{F}^{\otimes t}(\mathcal{H})) = \max_{\mathbf{X}} N_{2,\mathbf{X}}(u; d_{2,\mathbf{X}}, \mathcal{F}^{\otimes t}(\mathcal{H}))$. For a vector-valued function class we define the empirical ℓ_2-norm similarly as $d_{2,\mathbf{X}}^2(\mathbf{h}, \mathbf{h}') = \frac{1}{t \cdot n} \sum_{k=1}^{r} \sum_{j=1}^{t} \sum_{i=1}^{n} (\mathbf{h}_k(\mathbf{x}_{ji}) - \mathbf{h}_k'(\mathbf{x}_{ji}))^2$.

Our goal is to bound the empirical Gaussian complexity of the set $S = \{(f_1(\mathbf{h}(\mathbf{x}_{1i})), \ldots, f_j(\mathbf{h}(\mathbf{x}_{ji})), \ldots, f_t(\mathbf{h}(\mathbf{x}_{ti}))) : j \in [t], f_j \in \mathcal{F}, \mathbf{h} \in \mathcal{H}\} \subseteq \mathbb{R}^{tn}$ or function class,

$$\hat{\mathfrak{G}}_{nt}(S) = \hat{\mathfrak{G}}_{\mathbf{X}}(\mathcal{F}^{\otimes t}(\mathcal{H})) = \frac{1}{nt}\mathbb{E}[\sup_{\mathbf{f} \in \mathcal{F}^{\otimes t}, \mathbf{h} \in \mathcal{H}} \sum_{j=1}^{t} \sum_{i=1}^{n} g_{ji} f_j(\mathbf{h}(\mathbf{x}_{ji}))]; \quad g_{ji} \sim \mathcal{N}(0,1)$$

in a manner that allows for easy application in several problems of interest. To be explicit, we also recall that,

$$\hat{\mathfrak{G}}_{\mathbf{X}}(\mathcal{H}) = \frac{1}{nt}\mathbb{E}_{\mathbf{g}}[\sup_{\mathbf{h} \in \mathcal{H}} \sum_{k=1}^{r} \sum_{j=1}^{t} \sum_{i=1}^{n} g_{kji} \mathbf{h}_k(\mathbf{x}_{ji})]; \quad g_{kji} \sim \mathcal{N}(0,1)$$

We now state the decomposition theorem for Gaussian complexity.

Theorem 7. *Let the function class $\mathcal{F}$ consist of functions that are ℓ_2-Lipschitz with constant $L(\mathcal{F})$, and have boundedness parameter $D_{\mathbf{X}} = \sup_{\mathbf{f},\mathbf{f}',\mathbf{h},\mathbf{h}'} d_{2,\mathbf{X}}(\mathbf{f}(\mathbf{h}), \mathbf{f}'(\mathbf{h}'))$. Further, define $\mathcal{Z} = \{\mathbf{h}(\bar{\mathbf{X}}) : \mathbf{h} \in \mathcal{H}, \bar{\mathbf{X}} \in \cup_{j=1}^{t}\{\mathbf{X}_j\}\}$. Then the (empirical) Gaussian complexity of the function class $\mathcal{F}^{\otimes t}(\mathcal{H})$ satisfies,*

$$\hat{\mathfrak{G}}_{\mathbf{X}}(\mathcal{F}^{\otimes t}(\mathcal{H})) \leq \inf_{D_{\mathbf{X}} \geq \delta > 0} \left\{ 4\delta + 64C(\mathcal{F}^{\otimes t}(\mathcal{H})) \cdot \log\left(\frac{D_{\mathbf{X}}}{\delta}\right) \right\} \leq \frac{4D_{\mathbf{X}}}{(nt)^2} + 128C(\mathcal{F}^{\otimes t}(\mathcal{H})) \cdot \log(nt)$$

where $C(\mathcal{F}^{\otimes t}(\mathcal{H})) = L(\mathcal{F}) \cdot \hat{\mathfrak{G}}_{\mathbf{X}}(\mathcal{H}) + \max_{\mathbf{Z} \in \mathcal{Z}} \hat{\mathfrak{G}}_{\mathbf{Z}}(\mathcal{F})$. Further, if $C(\mathcal{F}^{\otimes t}(\mathcal{H})) \leq D_{\mathbf{X}}$ then by computing the exact infima of the expression,

$$\hat{\mathfrak{G}}_{\mathbf{X}}(\mathcal{F}^{\otimes t}(\mathcal{H})) \leq 64 \left(C(\mathcal{F}^{\otimes t}(\mathcal{H})) + C(\mathcal{F}^{\otimes t}(\mathcal{H})) \cdot \log\left(\frac{D_{\mathbf{X}}}{C(\mathcal{F}^{\otimes t}(\mathcal{H}))}\right) \right)$$

Proof. For ease of notation we define $N = nt$ in the following. We can rewrite the Gaussian complexity of the function class $\mathcal{F}^{\otimes t}(\mathcal{H})$ as,

$$\hat{\mathfrak{G}}_{\mathbf{X}}(\mathcal{F}^{\otimes t}(\mathcal{H})) = \mathbb{E}[\frac{1}{nt} \sup_{\mathbf{f}(\mathbf{h}) \in \mathcal{F}^{\otimes t}(\mathcal{H})} \sum_{j=1}^{t} \sum_{i=1}^{n} g_{ji} f_j(\mathbf{h}(\mathbf{x}_{ji}))] = \mathbb{E}[\frac{1}{\sqrt{N}} \cdot \sup_{\mathbf{f}(\mathbf{h}) \in \mathcal{F}^{\otimes t}(\mathcal{H})} Z_{\mathbf{f}(\mathbf{h})}]$$

from which we define the mean-zero stochastic process $Z_{\mathbf{f}(\mathbf{h})} = \frac{1}{\sqrt{N}} \sum_{j=1}^{t} \sum_{i=1}^{n} g_{ji} f_j(\mathbf{h}(\mathbf{x}_{ji}))$ for a fixed sequence of design points $\mathbf{x}_{ji}$, indexed by elements $\{\mathbf{f}(\mathbf{h}) \in \mathcal{F}^{\otimes t}(\mathcal{H})\}$, and for a sequence of independent Gaussian random variables g_{ji}. Note the process $Z_{\mathbf{f}(\mathbf{h})}$ has sub-gaussian increments, in the sense that, $Z_{\mathbf{f}(\mathbf{h})} - Z_{\mathbf{f}'(\mathbf{h}')}$ is a sub-gaussian random variable with parameter $d_{2,\mathbf{X}}^2(\mathbf{f}(\mathbf{h}), \mathbf{f}'(\mathbf{h}')) = \frac{1}{N} \sum_{j=1}^{t} \sum_{i=1}^{n} (f_j(\mathbf{h}(\mathbf{x}_{ji})) - f_j'(\mathbf{h}'(\mathbf{x}_{ji}))^2$. Since $Z_{\mathbf{f}(\mathbf{h})}$ is a mean-zero stochastic process we have that, $\mathbb{E}[\sup_{\mathbf{f}(\mathbf{h}) \in \mathcal{F}^{\otimes t}(\mathcal{H})} Z_{\mathbf{f}(\mathbf{h})}] = \mathbb{E}[\sup_{\mathbf{f}(\mathbf{h}) \in \mathcal{F}^{\otimes t}(\mathcal{H})} Z_{\mathbf{f}(\mathbf{h})} - Z_{\mathbf{f}'(\mathbf{h}')}] \leq \mathbb{E}[\sup_{\mathbf{f}(\mathbf{h}), \mathbf{f}'(\mathbf{h}') \in \mathcal{F}^{\otimes t}(\mathcal{H})} Z_{\mathbf{f}(\mathbf{h})} - Z_{\mathbf{f}'(\mathbf{h}')}]$. Now an appeal to the Dudley entropy integral bound, Wainwright [2019, Theorem 5.22] shows that,

$$\mathbb{E}[\sup_{\mathbf{f}(\mathbf{h}), \mathbf{f}'(\mathbf{h}') \in \mathcal{F}^{\otimes t}(\mathbf{h})} Z_{\mathbf{f}(\mathbf{h})} - Z_{\mathbf{f}(\mathbf{h}')}] \leq 4\mathbb{E}[\sup_{d_{2,\mathbf{X}}(\mathbf{f}(\mathbf{h}), \mathbf{f}(\mathbf{h}')) \leq \delta} Z_{\mathbf{f}(\mathbf{h})} - Z_{\mathbf{f}(\mathbf{h}')}] + 32 \int_{\delta}^{D} \sqrt{\log N_{\mathbf{X}}(u; d_{2,\mathbf{X}}, \mathcal{F}^{\otimes t}(\mathcal{H}))} du.$$

We now turn to bounding each of the above terms. Parametrizing the sequence of i.i.d. gaussian variables as $\mathbf{g}$, it follows that $\sup_{d_{2,\mathbf{X}}(\mathbf{f}(\mathbf{h}), \mathbf{f}(\mathbf{h}')) \leq \delta} Z_{\mathbf{f}(\mathbf{h})} - Z_{\mathbf{f}(\mathbf{h}')} \leq \sup_{\mathbf{v}:\|\mathbf{v}\|_2 \leq \delta} \mathbf{g} \cdot \mathbf{v} \leq \|\mathbf{g}\|\delta$. The corresponding expectation bound, after an application of Jensen's inequality to the $\sqrt{\cdot}$ function gives $\mathbb{E}[\sup_{d_{2,\mathbf{X}}(\mathbf{f}(\mathbf{h}), \mathbf{f}(\mathbf{h}')) \leq \delta} Z_{\mathbf{f}(\mathbf{h})} - Z_{\mathbf{f}(\mathbf{h}')}] \leq \mathbb{E}[\|\mathbf{g}\|_2 \delta] \leq \sqrt{N}\delta$.

We now turn to bounding the second term by decomposing the distance metric $d_{2,\mathbf{X}}$ into a distance over $\mathcal{F}^{\otimes t}$ and a distance over $\mathcal{H}$. We then use a covering argument on each of the spaces $\mathcal{F}^{\otimes t}$ and $\mathcal{H}$ to witness a covering of the composed space $\mathcal{F}^{\otimes t}(\mathcal{H})$. Recall we refer to the entire dataset concatenated over the t tasks as $\mathbf{X} \equiv \{\mathbf{x}_{ji}\}_{j=1, i=1}^{t,n}$. First, let $C_{\mathcal{H}_{\mathbf{X}}}$ be a covering of the of function space $\mathcal{H}$ in the empirical ℓ_2-norm with respect to the inputs $\mathbf{X}$ at scale ϵ_1. Then for each $\mathbf{h} \in C_{\mathcal{H}_{\mathbf{X}}}$, construct an ϵ_2-covering, $C_{\mathcal{F}^{\otimes t}_{\mathbf{h}(\mathbf{X})}}$, of the function space $\mathcal{F}^{\otimes t}$ in the empirical ℓ_2-norm with respect to the inputs $\mathbf{h}(\mathbf{X})$ at scale ϵ_2. We then claim that set $C_{\mathcal{F}^{\otimes t}(\mathcal{H})} = \cup_{\mathbf{h} \in C_{\mathcal{H}_{\mathbf{X}}}} (C_{\mathcal{F}^{\otimes t}_{\mathbf{h}(\mathbf{X})}})$ is an $\epsilon_1 \cdot L(\mathcal{F}) + \epsilon_2$-cover for the function space $\mathcal{F}^{\otimes t}(\mathcal{H})$ in the empirical ℓ_2-norm over the inputs $\mathbf{X}$. To see this, let $\mathbf{h} \in \mathcal{H}$ and $\mathbf{f} \in \mathcal{F}^{\otimes t}$ be arbitrary. Now let $\mathbf{h}' \in C_{\mathcal{H}_{\mathbf{X}}}$ be ϵ_1-close to $\mathbf{h}$. Given this $\mathbf{h}'$, there exists $\mathbf{f}' \in C_{\mathcal{F}^{\otimes t}_{\mathbf{h}'(\mathbf{X})}}$ such that $\mathbf{f}'$ is ϵ_2-close to $\mathbf{f}$ with respect to inputs $\mathbf{h}'(\mathbf{X})$. By construction $(\mathbf{h}', \mathbf{f}') \in C_{\mathcal{F}^{\otimes t}(\mathcal{H})}$. Finally, using the triangle inequality, we have that,

$$d_{2,\mathbf{X}}(\mathbf{f}(\mathbf{h}), \mathbf{f}'(\mathbf{h}')) \leq d_{2,\mathbf{X}}(\mathbf{f}(\mathbf{h}), \mathbf{f}(\mathbf{h}')) + d_{2,\mathbf{X}}(\mathbf{f}(\mathbf{h}'), \mathbf{f}'(\mathbf{h}')) =$$

$$\sqrt{\frac{1}{N} \sum_{j=1}^{t} \sum_{i=1}^{n} (f_j(\mathbf{h}(\mathbf{x}_{ji})) - f_j(\mathbf{h}'(\mathbf{x}_{ji})))^2} + \sqrt{\frac{1}{N} \sum_{j=1}^{t} \sum_{i=1}^{n} (f_j(\mathbf{h}'(\mathbf{x}_{ji})) - f_j'(\mathbf{h}'(\mathbf{x}_{ji})))^2} \leq$$

$$L(\mathcal{F}) \sqrt{\frac{1}{N} \sum_{k=1}^{r} \sum_{j=1}^{t} \sum_{i=1}^{n} (\mathbf{h}_k(\mathbf{x}_{ji}) - \mathbf{h}_k'(\mathbf{x}_{ji}))^2} + \sqrt{\frac{1}{N} \sum_{j=1}^{t} \sum_{i=1}^{n} (f_j(\mathbf{h}'(\mathbf{x}_{ji})) - f_j'(\mathbf{h}'(\mathbf{x}_{ji})))^2} =$$

$$L(\mathcal{F}) \cdot d_{2,\mathbf{X}}(\mathbf{h}, \mathbf{h}') + d_{2,\mathbf{h}'(\mathbf{X})}(\mathbf{f}, \mathbf{f}') \leq \epsilon_1 \cdot L(\mathcal{F}) + \epsilon_2$$

appealing to the uniform Lipschitz property of the function class $\mathcal{F}$ in moving from the second to third line, which establishes the claim.

We now bound the cardinality of the covering $C_{\mathcal{F}^{\otimes t}(\mathcal{H})}$. First, note $|C_{\mathcal{F}^{\otimes t}(\mathcal{H})}| = \sum_{\mathbf{h} \in C_{\mathcal{H}_{\mathbf{X}}}} |C_{\mathcal{F}^{\otimes t}_{\mathbf{h}(\mathbf{X})}}| \leq |C_{\mathcal{H}_{\mathbf{X}}}| \cdot \max_{\mathbf{h} \in \mathcal{H}_{\mathbf{X}}} |C_{\mathcal{F}^{\otimes t}_{\mathbf{h}(\mathbf{X})}}|$. To control $\max_{\mathbf{h} \in \mathcal{H}_{\mathbf{X}}} |C_{\mathcal{F}^{\otimes t}_{\mathbf{h}(\mathbf{X})}}|$, note an ϵ-cover of $\mathcal{F}^{\otimes t}_{\mathbf{h}(\mathbf{X})}$ in the empirical ℓ_2-norm with respect to $\mathbf{h}(\mathbf{X})$ can be obtained from the cover $C_{\mathcal{F}_{\mathbf{h}(\mathbf{X}_1)}} \times \ldots \times C_{\mathcal{F}_{\mathbf{h}(\mathbf{X}_t)}}$ where $C_{\mathcal{F}_{\mathbf{h}(\mathbf{X}_i)}}$ denotes a ϵ-cover of $\mathcal{F}$ in the empirical ℓ_2-norm with respect to $\mathbf{h}(\mathbf{X}_i)$. Hence $\max_{\mathbf{h} \in \mathcal{H}_{\mathbf{X}}} |C_{\mathcal{F}^{\otimes t}_{\mathbf{h}(\mathbf{X})}}| \leq |C_{\mathcal{F}_{\mathbf{h}(\mathbf{X}_1)}} \times \ldots \times C_{\mathcal{F}_{\mathbf{h}(\mathbf{X}_t)}}| \leq \underbrace{|\max_{\mathbf{z} \in \mathcal{Z}} C_{\mathcal{F}_{\mathbf{z}}} \times \ldots \times \max_{\mathbf{z} \in \mathcal{Z}} C_{\mathcal{F}_{\mathbf{z}}}|}_{t \text{ times}} \leq$

$|\max_{\mathbf{z} \in \mathcal{Z}} C_{\mathcal{F}_{\mathbf{z}}}|^t$. Combining these facts provides a bound on the metric entropy of,

$$\log N_{2,\mathbf{X}}(\epsilon_1 \cdot L(\mathcal{F}) + \epsilon_2, d_{2,\mathbf{X}}, \mathcal{F}^{\otimes t}(\mathcal{H})) \leq \log N_{2,\mathbf{X}}(\epsilon_1, d_{2,\mathbf{X}}, \mathcal{H}) + t \cdot \max_{\mathbf{Z} \in \mathcal{Z}} \log N_{2,\mathbf{Z}}(\epsilon_2, d_{2,\mathbf{Z}}, \mathcal{F}).$$

Using the covering number upper bound with $\epsilon_1 = \frac{\epsilon}{2 \cdot L(\mathcal{F})}$, $\epsilon_2 = \frac{\epsilon}{2}$ and sub-additivity of the $\sqrt{\cdot}$ function then gives a bound on the entropy integral of,

$$\int_\delta^D \sqrt{\log N_2(\epsilon, d_{2,\mathbf{X}}, \mathcal{F}^{\otimes t}(\mathcal{H}))}\, d\epsilon \leq \int_\delta^D \sqrt{\log N_{2,\mathbf{X}}(\epsilon/(2L(\mathcal{F})), d_{2,\mathbf{X}}, \mathcal{H})}\, d\epsilon + \sqrt{t}\int_\delta^D \max_{\mathbf{Z} \in \mathcal{Z}} \sqrt{\log N_{2,\mathbf{z}}(\frac{\epsilon}{2}, d_{2,\mathbf{z}}, \mathcal{F})}\, d\epsilon$$

From the Sudakov minoration theorem Wainwright [2019][Theorem 5.30] for Gaussian processes and the fact packing numbers at scale u upper bounds the covering number at scale u we find:

$$\log N_{2,\mathbf{X}}(u; d_{2,\mathbf{X}}, \mathcal{H}) \leq 4\left(\frac{\sqrt{nt}\hat{\mathfrak{G}}_{\mathbf{X}}(\mathcal{H})}{u}\right)^2 \quad \forall u > 0 \quad \text{and} \quad \log N_{2,\mathbf{Z}}(u; d_{2,\mathbf{Z}}, \mathcal{F}) \leq 4\left(\frac{\sqrt{n}\hat{\mathfrak{G}}_{\mathbf{Z}}(\mathcal{F})}{u}\right)^2 \quad \forall u > 0.$$

For the $\mathcal{H}$ term we apply the result to the mean-zero Gaussian process $Z_{\mathbf{h}} = \frac{1}{\sqrt{nt}}\sum_{k=1}^r \sum_{j=1}^t \sum_{i=1}^n g_{kji} h_k(\mathbf{x}_{ji})$, for $g_{kji} \sim \mathcal{N}(0,1)$ i.i.d. and $\mathbf{h} \in \mathcal{H}$. Combining all of the aforementioned upper bounds, shows that

$$\hat{\mathfrak{G}}_{\mathbf{X}}(\mathcal{F}^{\otimes t}(\mathcal{H})) \leq \frac{1}{\sqrt{nt}}\left(4\delta\sqrt{nt} + 64 L(\mathcal{F})\cdot\hat{\mathfrak{G}}_{\mathbf{X}}(\mathcal{H})\cdot\sqrt{nt}\int_\delta^{D_{\mathbf{X}}} \frac{1}{u} du + 64\sqrt{nt}\cdot\max_{\mathbf{Z}\in\mathcal{Z}}\hat{\mathfrak{G}}_{\mathbf{Z}}(\mathcal{F})\int_\delta^{D_{\mathbf{X}}} \frac{1}{u} du\right) \leq$$

$$4\delta + 64(L(\mathcal{F})\cdot\hat{\mathfrak{G}}_{\mathbf{X}}(\mathcal{H}) + \max_{\mathbf{Z}\in\mathcal{Z}}\hat{\mathfrak{G}}_{\mathbf{Z}}(\mathcal{F}))\cdot\log\left(\frac{D_{\mathbf{X}}}{\delta}\right) = \delta + C(\mathcal{F}^{\otimes t}(\mathcal{H}))\cdot\log\left(\frac{D_{\mathbf{X}}}{\delta}\right)$$

defining $C(\mathcal{F}^{\otimes t}(\mathcal{H})) = L(\mathcal{F})\cdot\hat{\mathfrak{G}}_{\mathbf{X}}(\mathcal{H}) + \max_{\mathbf{Z}\in\mathcal{Z}}\hat{\mathfrak{G}}_{\mathbf{Z}}(\mathcal{F})$. Choosing $\delta = D_{\mathbf{X}}/(nt)^2$ gives the first inequality. Balancing the first and second term gives the optimal choice $\delta = \frac{1}{C(\mathcal{F}^{\otimes t}(\mathcal{H}))}$ for the second inequality under the stated conditions. $\qquad\square$

C   Proofs in Section 4

In this section we instantiate our general framework in several concrete examples. This consists of two steps: first verifying a task diversity lower bound for the function classes and losses and then bounding the various complexity terms appearing in the end-to-end LTL guarantee in Theorem 3 or its variants.

C.1   Logistic Regression

Here we include the proofs of the results which both bound the complexities of the function classes $\mathcal{F}$ and $\mathcal{H}$ in the logistic regression example as well establish the task diversity lower bound in this setting. In this section we use the following definition,

Definition 4. We say the covariate distribution $\mathbb{P}_{\mathbf{x}}(\cdot)$ is $\mathbf{\Sigma}$-sub-gaussian if for all $\mathbf{v} \in \mathbb{R}^d$, $\mathbb{E}[\exp(\mathbf{v}^\top \mathbf{x}_i)] \leq \exp\left(\frac{\|\mathbf{\Sigma}^{1/2}\mathbf{v}\|^2}{2}\right)$ where the covariance $\mathbf{\Sigma}$ further satisfies $\sigma_{\max}(\mathbf{\Sigma}) \leq C$ and $\sigma_{\min}(\mathbf{\Sigma}) \geq c > 0$ for universal constants c, C.

We begin by presenting the proof of the Theorem 4 which essentially relies on instantiating a variant of Theorem 3. In order to obtain a sharper dependence in the noise terms in the test learning stage we actually directly combine Theorem 1 and Corollary 2.

Since we are also interested in stating data-dependent guarantees in this section we use the notation $\mathbf{\Sigma}_{\mathbf{X}} = \frac{1}{nt}\sum_{j=1}^t\sum_{i=1}^n \mathbf{x}_{ji}\mathbf{x}_{ji}^\top$ to refer to the empirical covariance across the the training phase samples and $\mathbf{\Sigma}_{\mathbf{X}_j}$ for corresponding empirical covariances across the per-task samples. Immediately following this result we present the statement of sharp data-dependent guarantee which depends on these empirical quantities for completeness.

Proof of Theorem 4. First note due to the task normalization conditions we can choose c_1, c_2 sufficiently large so that the realizability assumption in Assumption 2 is satisfied–in particular, we can assume that c_2 is chosen large enough to contain all the parameters $\boldsymbol{\alpha}_j^\star$ for $j \in [t] \cup \{0\}$ and $c_1 \geq \frac{C}{c}c_2$. Next note that under the conditions of the result we can use Lemma 1 to verify the task diversity condition is satisfied with parameters $(\tilde{\nu}, 0)$ with $\nu = \sigma_r(\mathbf{A}^\top\mathbf{A}/t) > 0$ with this choice of constants.

Finally, in order to combine Theorem 1 and Corollary 2 we begin by bounding each of the complexity terms in the expression. First,

- For the feature learning complexity in the training phase we obtain,

$$\hat{\mathfrak{S}}_{\mathbf{X}}(\mathcal{H}) = \frac{1}{nt}\mathbb{E}[\sup_{\mathbf{B}\in\mathcal{H}}\sum_{k=1}^{r}\sum_{j=1}^{t}\sum_{i=1}^{n}g_{kji}\mathbf{b}_k^\top\mathbf{x}_{ji}] = \frac{1}{nt}\mathbb{E}[\sup_{(\mathbf{b}_1,\ldots,\mathbf{b}_r)\in\mathcal{H}}\sum_{k=1}^{r}\mathbf{b}_k^\top(\sum_{j=1}^{t}\sum_{i=1}^{n}g_{kji}\mathbf{x}_{ji})] \le$$

$$\frac{1}{nt}\sum_{k=1}^{r}\mathbb{E}[\|\sum_{j=1}^{t}\sum_{i=1}^{n}g_{kji}\mathbf{x}_{ji}\|] \le \frac{1}{nt}\sum_{k=1}^{r}\sqrt{\mathbb{E}[\|\sum_{j=1}^{t}\sum_{i=1}^{n}g_{kji}\mathbf{x}_{ji}\|^2]} \le \frac{1}{nt}\sum_{k=1}^{r}\sqrt{\sum_{j=1}^{t}\sum_{i=1}^{n}\|\mathbf{x}_{ji}\|^2}$$

$$= \frac{r}{\sqrt{nt}}\sqrt{\mathrm{tr}(\boldsymbol{\Sigma}_{\mathbf{X}})}.$$

  Further by definition the class $\mathcal{F}$ as linear maps with parameters $\|\boldsymbol{\alpha}\|_2 \le O(1)$ we obtain that $L(\mathcal{F}) = O(1)$. We now proceed to convert this to a population quantity by noting that $\mathbb{E}[\sqrt{\mathrm{tr}(\boldsymbol{\Sigma}_{\mathbf{X}})}] \le \sqrt{d \cdot \mathbb{E}[\|\boldsymbol{\Sigma}_{\mathbf{X}}\|]} \le O(\sqrt{d})$ for $nt \gtrsim d$ by Lemma 4.

- For the complexity of learning $\mathcal{F}$ in the training phase we obtain,

$$\bar{\mathfrak{S}}_n(\mathcal{F}) = \frac{1}{n}\mathbb{E}[\sup_{\|\boldsymbol{\alpha}\|\le c_1}\sum_{i=1}^{n}g_i\boldsymbol{\alpha}^\top\mathbf{B}^\top\mathbf{x}_{ji}] = \frac{c_1}{n}\mathbb{E}[\|\sum_{i=1}^{n}g_i\mathbf{B}^\top\mathbf{x}_{ji}\|] \le \frac{c_1}{n}\sqrt{\sum_{i=1}^{n}\|\mathbf{B}^\top\mathbf{x}_{ji}\|^2} =$$

$$\frac{c_1}{\sqrt{n}}\sqrt{\mathrm{tr}(\mathbf{B}\mathbf{B}^\top\boldsymbol{\Sigma}_{\mathbf{X}_j})} = \frac{c_1}{\sqrt{n}}\sqrt{\mathrm{tr}(\mathbf{B}^\top\boldsymbol{\Sigma}_{\mathbf{X}_j}\mathbf{B})}.$$

  Now by the variational characterization of singular values it follows that $\max_{\mathbf{B}\in\mathcal{H}}\frac{c_1}{\sqrt{n}}\sqrt{\mathrm{tr}(\mathbf{B}^\top\boldsymbol{\Sigma}_{\mathbf{X}_j}\mathbf{B})} \le \frac{c_1}{n}\sqrt{\sum_{i=1}^{r}\sigma_i(\boldsymbol{\Sigma}_{\mathbf{X}_j})}$ Thus it immediately follows that,

$$\max_{\mathbf{Z}\in\mathcal{Z}}\frac{c_1}{\sqrt{n}}\sqrt{\mathrm{tr}(\boldsymbol{\Sigma}_{\mathbf{X}_j})} = \max_{\mathbf{X}_j}\max_{\mathbf{B}\in\mathcal{H}}\frac{c_1}{\sqrt{n}}\sqrt{\mathrm{tr}(\mathbf{B}^\top\boldsymbol{\Sigma}_{\mathbf{X}_j}\mathbf{B})} \le \max_{\mathbf{X}_j}\frac{c_1}{\sqrt{n}}\sqrt{\sum_{i=1}^{r}\sigma_i(\boldsymbol{\Sigma}_{\mathbf{X}_j})}.$$

  for $j \in [t]$. We can convert this to a population quantity again by applying Lemma 4 which shows $\mathbb{E}[\sqrt{\sum_{i=1}^{r}\sigma_i(\boldsymbol{\Sigma}_{\mathbf{X}_j})}] \le O(\sqrt{r})$ for $n \gtrsim d + \log t$.

- A nearly identical argument shows the complexity of learning $\mathcal{F}$ in the testing phase is,

$$\hat{\mathfrak{S}}_{\mathbf{Y}_{\hat{\mathbf{h}}}}(\mathcal{F}) = \frac{1}{m}\mathbb{E}[\sup_{\|\boldsymbol{\alpha}\|\le c_1}\sum_{i=1}^{m}\epsilon_i\boldsymbol{\alpha}^\top\hat{\mathbf{B}}^\top\mathbf{x}_{(0)i}] \le \frac{c_1}{\sqrt{m}}\sqrt{\sum_{i=1}^{r}\sigma_i(\hat{\mathbf{B}}^\top\boldsymbol{\Sigma}_{\mathbf{X}_0}\hat{\mathbf{B}})}$$

  Crucially, here we can apply the first result in Corollary 2 which allows us to take the expectation over $\mathbf{X}_0$ before maximizing over $\mathbf{B}$. Thus applying Lemma 4 as before gives the result, $\mathbb{E}[\sqrt{\sum_{i=1}^{r}\sigma_i(\mathbf{B}^\top\boldsymbol{\Sigma}_{\mathbf{X}_0}\mathbf{B})}] \le O(\sqrt{r})$ for $m \gtrsim r$.

This gives the first series of claims.

Finally we verify that Assumption 1 holds so as to use Theorem 1 and Corollary 2 to instantiate the end-to-end guarantee. First the boundedness parameter becomes,

$$D_{\mathcal{X}} = \sup_{\boldsymbol{\alpha},\mathbf{B}}(\mathbf{x}^\top\mathbf{B}\boldsymbol{\alpha}) \le O(D)$$

using the assumptions that $\|\mathbf{x}\|_2 \le D$, $\|\boldsymbol{\alpha}\|_2 \le O(1)$, $\|\mathbf{B}\|_2 = 1$. For the logistic loss bounds, $\ell(\eta;y) = y\eta - \log(1 + \exp(\eta))$. Since $|\nabla_\eta\ell(\eta;y)| = |y - \frac{\exp(\eta)}{1+\exp(\eta)}| \le 1$ it is $O(1)$-Lipschitz in its first coordinate uniformly over its second so $L = O(1)$. Moreover as, $|\ell(\eta;y)| \le O(\eta)$ where $\eta = \mathbf{x}^\top\mathbf{B}\boldsymbol{\alpha} \le \|\mathbf{x}\| \le D$ it follows the loss is uniformly bounded with parameter $O(D)$ so $B = O(D)$.

Lastly, to use Corollary 2 to bound the test phase error we need to compute the maximal variance term $\sigma^2 = \frac{1}{m} \sup_{f \in \mathcal{F}} \sum_{i=1}^{m} \mathrm{Var}(\ell(f \circ \hat{\mathbf{h}}(\mathbf{x}_i), y_i))$. Since for the logistic loss $\ell(\cdot, \cdot)$ satisfies the 1-Lipschitz property uniformly we have that, $\mathrm{Var}(\ell(f \circ \hat{\mathbf{h}}(\mathbf{x}_i), y_i)) \leq \mathrm{Var}(f \circ \hat{\mathbf{h}}(\mathbf{x}_i))$ for each $i \in [m]$. Collapsing the variance we have that,

$$\frac{1}{m} \sup_{\boldsymbol{\alpha}:\|\boldsymbol{\alpha}\|_2 \leq O(1)} \sum_{i=1}^{m} \mathrm{Var}(\mathbf{x}_i^\top \hat{\mathbf{B}} \boldsymbol{\alpha}) \leq \frac{1}{m} \sup_{\boldsymbol{\alpha}:\|\boldsymbol{\alpha}\|_2 \leq O(1)} \sum_{i=1}^{m} (\boldsymbol{\alpha}\hat{\mathbf{B}})^\top \boldsymbol{\Sigma} \hat{\mathbf{B}} \boldsymbol{\alpha} \leq O(\|\hat{\mathbf{B}}\boldsymbol{\Sigma}\hat{\mathbf{B}}\|_2) \leq$$
$$O(\|\boldsymbol{\Sigma}\|) \leq O(C) = O(1)$$

under our assumptions, which implies that $\sigma \leq O(1)$. Assembling the previous bounds shows the transfer learning risk is bounded by,

$$\lesssim \frac{1}{\tilde{\nu}} \cdot \left(\log(nt) \cdot \left[\sqrt{\frac{dr^2}{nt}} + \sqrt{\frac{r}{n}} \right] \right) + \sqrt{\frac{r}{m}}$$
$$+ \left(\frac{D}{\tilde{\nu}} \cdot \max\left(\frac{1}{(nt)^2}, \sqrt{\frac{\log(2/\delta)}{nt}} \right) + \sqrt{\frac{\log(2/\delta)}{m}} + D\frac{\log(2/\delta)}{m} \right).$$

with probability at least $1 - 2\delta$. Suppressing all logarithmic factors and using the additional condition $D \lesssim \min(dr^2, \sqrt{rm})$ guarantees the noise terms are higher-order. $\qquad \square$

Recall, in the context of the two-stage ERM procedure introduced in Section 2 we let the design matrix and responses y_{ji} for the jth task be $\mathbf{X}_j$ and $\mathbf{y}_j$ for $j \in [t] \cup \{0\}$, and the entire design matrix and responses concatenated over all $j \in [t]$ tasks as $\bar{\mathbf{X}}$ and $\mathbf{y}$ respectively. $\boldsymbol{\Sigma}_{\bar{\mathbf{X}}}$ denotes the empirical covariance of the design matrix $\bar{\mathbf{X}}$. We now state a sharp, data-dependent guarantee for logistic regression.

Corollary 3. *If Assumption 3 holds, $\mathbf{h}^\star(\cdot) \in \mathcal{H}$, and $\mathcal{F}_0 = \{ f \mid f(\mathbf{x}) = \boldsymbol{\alpha}^\top \mathbf{z}, \boldsymbol{\alpha} \in \mathbb{R}^r, \|\boldsymbol{\alpha}\| \leq c_2 \}$, then there exist constants c_1, c_2 such that the training tasks $f_j^\star$ are $(\Omega(\tilde{\nu}), 0)$-diverse over $\mathcal{F}_0$. Then with probability at least $1 - 2\delta$:*

Transfer Learning Risk $\leq$
$$O\left(\frac{1}{\tilde{\nu}} \cdot \left(\log(nt) \cdot \left[\sqrt{\frac{\mathrm{tr}(\boldsymbol{\Sigma}_{\mathbf{X}})r^2}{nt}} + \max_{j \in [t]} \sqrt{\frac{\sum_{i=1}^{r} \sigma_i(\mathbf{X}_j)}{n}} \right] \right) + \sqrt{\frac{\sum_{i=1}^{r} \sigma_i(\mathbf{X}_0)}{m}} \right)$$
$$+ O\left(\frac{D}{\tilde{\nu}} \cdot \max\left(\frac{1}{(nt)^2}, \sqrt{\frac{\log(4/\delta)}{nt}} \right) + \sqrt{\frac{\log(4/\delta)}{m}} + D\frac{\log(4/\delta)}{m} \right).$$

Proof of Corollary 3. This follows immediately from the proof of Theorem 4 and applying Corollary 1, Corollary 2. Merging terms and applying a union bound gives the result. $\qquad \square$

The principal remaining challenge is to obtain a sharp lower bound on the task diversity.

Lemma 1. *Let Assumption 3 hold in the setting of Theorem 4. Then there exists c_2 such that if $c_1 \geq \frac{C}{c}c_2$ the problem is task-diverse with parameter $(\Omega(\tilde{\nu}), 0)$ in the sense of Definition 3 where $\tilde{\nu} = \sigma_r(\mathbf{A}^\top \mathbf{A}/t)$.*

Proof. Our first observation specializes Lemma 2 to the case of logistic regression where $\Phi(\eta) = \log(1 + \exp(\eta))$, $s(\sigma) = 1$ with $\mathbf{h}(\mathbf{x}) = \mathbf{Bx}$ parametrized with $\mathbf{B} \in \mathbb{R}^{d \times r}$ having orthonormal columns and $\mathbf{f} \equiv \boldsymbol{\alpha}$. Throughout we also assume that c_2 is chosen large enough to contain all the parameters $\boldsymbol{\alpha}_j^\star$ for $j \in [t] \cup \{0\}$ and $c_1 \geq \frac{C}{c}c_2$. These conditions immediately imply the realizability conditions.

This lemma essentially allows us to use smoothness and strong convexity to bound the task-averaged representation distance and worst-case representation difference. By appealing to Lemma 2 and Lemma 3 we have that,

$$\frac{1}{8}\mathbb{E}_{\mathbf{x}_i}[\exp(-\max(|\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}}|, |\mathbf{h}(\mathbf{x}_i)^\top \boldsymbol{\alpha}|)) \cdot (\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}} - \mathbf{h}(\mathbf{x}_i)^\top \boldsymbol{\alpha})^2] \leq$$
$$\mathbb{E}_{\mathbf{x}_i, y}[\ell(\hat{f} \circ \hat{\mathbf{h}}(\mathbf{x}_i), y_i) - \ell(f \circ \mathbf{h}(\mathbf{x}_i), y_i)] \leq \frac{1}{8}\mathbb{E}_{\mathbf{x}_i}[(\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}} - \mathbf{h}(\mathbf{x}_i)^\top \boldsymbol{\alpha})^2]$$

21

We now bound each term in the task diversity,

- We first bound the representation difference where $\mathbf{x}_i, y_i \sim (\mathbb{P}_\mathbf{x}(\cdot), \mathbb{P}_{y|\mathbf{x}}(\cdot | f_0^\star \circ \mathbf{h}^\star(\mathbf{x})))$,

$$d_{\mathcal{F}, \mathcal{F}_0}(\hat{\mathbf{h}}; \mathbf{h}^\star) = \sup_{\boldsymbol{\alpha}: \|\boldsymbol{\alpha}\|_2 \leq c_2} \inf_{\hat{\boldsymbol{\alpha}}: \|\hat{\boldsymbol{\alpha}}\| \leq c_1} \mathbb{E}_{\mathbf{x},y}[\ell(\hat{f} \circ \hat{\mathbf{h}}, \mathbf{x}, y) - \ell(f_0^\star \circ \mathbf{h}^\star(\mathbf{x}), y)]] \leq$$

$$\sup_{\boldsymbol{\alpha}: \|\boldsymbol{\alpha}\|_2 \leq c_2} \inf_{\hat{\boldsymbol{\alpha}}: \|\hat{\boldsymbol{\alpha}}\| \leq c_1} \frac{1}{8} \mathbb{E}_\mathbf{x}[(\hat{\mathbf{h}}(\mathbf{x})^\top \hat{\boldsymbol{\alpha}} - \mathbf{h}^\star(\mathbf{x})^\top \boldsymbol{\alpha})^2].$$

Now for sufficiently large c_1, by Lagrangian duality the unconstrained minimizer of the inner optimization problem is equivalent to the constrained minimizer. In particular first note that under the assumptions of the problem there is unique unconstrained minimizer given by $\inf_{\hat{\boldsymbol{\alpha}}} \frac{1}{8} \mathbb{E}_{\mathbf{x}_i}[(\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}} - \mathbf{h}^\star(\mathbf{x}_i)^\top \boldsymbol{\alpha})^2] \rightarrow \hat{\boldsymbol{\alpha}}_{unconstrained} = -\mathbf{F}_{\hat{\mathbf{h}}\hat{\mathbf{h}}} \mathbf{F}_{\hat{\mathbf{h}}\mathbf{h}} \boldsymbol{\alpha} = (\hat{\mathbf{B}}^\top \boldsymbol{\Sigma} \hat{\mathbf{B}})^{-1} (\hat{\mathbf{B}}^\top \boldsymbol{\Sigma} \hat{\mathbf{B}}) \boldsymbol{\alpha}$ from the proof and preamble of Lemma 6. Note that since $\hat{\mathbf{B}}$ and $\mathbf{B}$ have orthonormal columns it follows that $\|\hat{\boldsymbol{\alpha}}\| \leq \frac{C}{c} c_2$ since $\hat{\mathbf{B}}^\top \boldsymbol{\Sigma} \hat{\mathbf{B}}$ is invertible. Thus if $c_1 \geq \frac{C}{c} c_2$, by appealing to Lagrangian duality for this convex quadratic objective with convex quadratic constraint, the unconstrained minimizer is equivalent to the constrained minimizer (since the unconstrained minimizer is contained in the constraint set). Hence leveraging the proof and result of Lemma 6 we obtain $\sup_{\boldsymbol{\alpha}: \|\boldsymbol{\alpha}\|_2 \leq c_2} \inf_{\hat{\boldsymbol{\alpha}}: \|\hat{\boldsymbol{\alpha}}\| \leq c_1} \frac{1}{8} \mathbb{E}_{\mathbf{x}_i}[(\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}} - \mathbf{h}^\star(\mathbf{x}_i)^\top \boldsymbol{\alpha})^2] \leq \frac{c_2}{8} \sigma_{\max}(\Lambda_{sc}(\mathbf{h}, \hat{\mathbf{h}}))$.

- We now turn out attention to controlling the average distance which we must lower bound. Here $\mathbf{x}_i, y_i \sim (\mathbb{P}_\mathbf{x}(\cdot), \mathbb{P}_{y|\mathbf{x}}(\cdot | f_j^\star \circ \mathbf{h}^\star(\mathbf{x})))$

$$\bar{d}_{\mathcal{F}, \mathbf{f}^\star}(\mathbf{h}; \hat{\mathbf{h}}) = \frac{1}{t} \sum_{j=1}^t \inf_{\|\hat{\boldsymbol{\alpha}}\| \leq c_1} \mathbb{E}_{\mathbf{x},y}[\ell(\hat{f} \circ \hat{\mathbf{h}}(\mathbf{x}_i), y) - \ell(f_j^\star \circ \mathbf{h}^\star(\mathbf{x}_i), y)]] \geq$$

$$\frac{1}{8t} \sum_{j=1}^t \mathbb{E}_\mathbf{x}[\exp(-\max(|\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}}|, |\mathbf{h}^\star(\mathbf{x}_i)^\top \boldsymbol{\alpha}_j^\star|)) \cdot (\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}} - \mathbf{h}^\star(\mathbf{x}_i)^\top \boldsymbol{\alpha}_j^\star)^2]$$

We will use the fact that in logistic regression $\mathbf{h}(\mathbf{x}_i) = \mathbf{B}\mathbf{x}_i$; in this case if $\mathbf{x}_i$ is C-subgaussian random vector in d dimensions, then $\mathbf{B}\mathbf{x}_i$ is C-subgaussian random vector in r dimensions. We lower bound each term in the sum over j identically and suppress the j for ease of notation in the following. For fixed j, note the random variables $Z_1 = (\boldsymbol{\alpha}_j^\star)^\top \mathbf{B}\mathbf{x}_i$ and $Z_2 = \hat{\boldsymbol{\alpha}}^\top \hat{\mathbf{B}}\mathbf{x}_i$ are subgaussian with variance parameter at most $\|\boldsymbol{\alpha}_j^\star\|_2^2 C^2$ and $\|\hat{\boldsymbol{\alpha}}\|_2^2 C^2$ respectively. Define the event $\mathbb{1}[E] = \mathbb{1}[|Z_1| \leq Ck\|\boldsymbol{\alpha}_j^\star\| \cap \mathbb{1}\{|Z_2| \leq Ck\|\hat{\boldsymbol{\alpha}}\|]$ for k to be chosen later. We use this event to lower bound the averaged task diversity since it is a non-negative random variable,

$$\mathbb{E}_\mathbf{x}[\exp(-\max(|\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}}|, |\mathbf{h}^\star(\mathbf{x}_i)^\top \boldsymbol{\alpha}_j^\star|)) \cdot (\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}} - \mathbf{h}^\star(\mathbf{x}_i)^\top \boldsymbol{\alpha}_j^\star)^2] \geq$$

$$\mathbb{E}_\mathbf{x}[\mathbb{1}[E] \exp(-\max(|\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}}|, |\mathbf{h}^\star(\mathbf{x}_i)^\top \boldsymbol{\alpha}_j^\star|)) \cdot (\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}} - \mathbf{h}^\star(\mathbf{x}_i)^\top \boldsymbol{\alpha}_j^\star)^2] \geq$$

$$\exp(-Ck \max(c_1, c_2)) \cdot \mathbb{E}_\mathbf{x}[\mathbb{1}[E](\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}} - \mathbf{h}^\star(\mathbf{x}_i)^\top \boldsymbol{\alpha}_j^\star)^2]$$

We now show that for appropriate choice of k, $\mathbb{E}_\mathbf{x}[\mathbb{1}[E](\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}} - \mathbf{h}^\star(\mathbf{x}_i)^\top \boldsymbol{\alpha}_j^\star)^2]$ is lower bounded by $\mathbb{E}_\mathbf{x}[(\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}} - \mathbf{h}^\star(\mathbf{x}_i)^\top \boldsymbol{\alpha}_j^\star)^2]$ modulo a constant factor. First write $\mathbb{E}_\mathbf{x}[\mathbb{1}[E](\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}} - \mathbf{h}^\star(\mathbf{x}_i)^\top \boldsymbol{\alpha}_j^\star)^2] = \mathbb{E}_\mathbf{x}[(\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}} - \mathbf{h}^\star(\mathbf{x}_i)^\top \boldsymbol{\alpha}_j^\star)^2] - \mathbb{E}_\mathbf{x}[\mathbb{1}[E^c](\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}} - \mathbf{h}^\star(\mathbf{x}_i)^\top \boldsymbol{\alpha}_j^\star)^2]$.

We upper bound the second term first using Cauchy-Schwarz,

$$\mathbb{E}_\mathbf{x}[\mathbb{1}[E^c](\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}} - \mathbf{h}^\star(\mathbf{x}_i)^\top \boldsymbol{\alpha}_j^\star)^2] \leq \sqrt{\mathbb{P}[E^c]} \sqrt{\mathbb{E}_\mathbf{x}(\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}} - \mathbf{h}^\star(\mathbf{x}_i)^\top \boldsymbol{\alpha}_j^\star)^4}$$

Define $Z_3 = \mathbf{x}_i^\top ((\mathbf{B}^\star)^\top \boldsymbol{\alpha}_j^\star - \hat{\mathbf{B}}^\top \hat{\boldsymbol{\alpha}})$ which by definition is subgaussian with parameter at most $((\mathbf{B}^\star)^\top \boldsymbol{\alpha}_j^\star - \hat{\mathbf{B}}^\top \hat{\boldsymbol{\alpha}}) \boldsymbol{\Sigma} ((\mathbf{B}^\star)^\top \boldsymbol{\alpha}_j^\star - \hat{\mathbf{B}}^\top \hat{\boldsymbol{\alpha}}) = \sigma^2$; since this condition implies L4-L2 hypercontractivity (see for example Wainwright [2019, Theorem 2.6]) we can also conclude that,

$$\sqrt{\mathbb{E}_\mathbf{x}(\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}} - \mathbf{h}^\star(\mathbf{x}_i)^\top \boldsymbol{\alpha}_j^\star)^4} \leq 10\sigma^2 = 10 \cdot \mathbb{E}_\mathbf{x}(\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}} - \mathbf{h}^\star(\mathbf{x}_i)^\top \boldsymbol{\alpha}_j^\star)^2.$$

Recalling the subgaussianity of Z_1 and Z_2, from an application of Markov and Jensen's inequality,

$$\mathbb{P}[|Z_1| \geq k \cdot C \|\boldsymbol{\alpha}_j^\star\|_2] \leq \frac{\mathbb{E}[Z^2]}{k^2 \cdot C^2 \|\boldsymbol{\alpha}_j^\star\|_2} \leq \frac{1}{k^2}$$

with an identical statement true for Z_2. Finally by using a union bound we have that $\sqrt{\mathbb{P}[E^c]} \leq \frac{\sqrt{2}}{k}$ using these probability bounds. Hence by taking $k = 30$ we can ensure that $\mathbb{E}_{\mathbf{x}}[\mathbb{1}[E](\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}} - \mathbf{h}^\star(\mathbf{x}_i)^\top \boldsymbol{\alpha}_j^\star)^2] \geq \frac{1}{2}\mathbb{E}_{\mathbf{x}}[(\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}} - \mathbf{h}^\star(\mathbf{x}_i)^\top \boldsymbol{\alpha}_j^\star)^2]$ by assembling the previous bounds. Finally since c_1, c_2, C, k are universal constants, by definition the conclusion that,

$$\mathbb{E}_{\mathbf{x}}[\exp(-\max(|\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}}|, |\mathbf{h}^\star(\mathbf{x}_i)^\top \boldsymbol{\alpha}_j^\star|)) \cdot (\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}} - \mathbf{h}^\star(\mathbf{x}_i)^\top \boldsymbol{\alpha}_j^\star)^2] \geq$$
$$\Omega(\mathbb{E}_{\mathbf{x}}(\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}} - \mathbf{h}^\star(\mathbf{x}_i)^\top \boldsymbol{\alpha}_j^\star)^2)$$

follows for each j. Hence the average over the t tasks is identically lower bounded as,

$$\Omega\left(\frac{1}{t}\sum_{j=1}^t \mathbb{E}_{\mathbf{x}}(\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}} - \mathbf{h}^\star(\mathbf{x}_i)^\top \boldsymbol{\alpha}_j^\star)^2\right)$$

Now using the argument from the upper bound to compute the infima since all the $\|\boldsymbol{\alpha}_j^\star\| \leq c_2$ (and hence the constrained minimizers identical to the unconstrained minimizers for each of the j terms for $c_1 \geq \frac{C}{c}c_2$) and using the proof of Lemma 6 we conclude that,

$$\bar{d}_{\mathcal{F},\mathbf{f}^\star}(\hat{\mathbf{h}}; \mathbf{h}^\star) \geq \Omega(\text{tr}(\Lambda_{sc}(\mathbf{h}^\star, \hat{\mathbf{h}})\mathbf{C})).$$

Combining these upper and lower bounds and concluding as in the proof of Lemma 6 shows

$$d_{\mathcal{F},\mathcal{F}_0}(\hat{\mathbf{h}}; \mathbf{h}^\star) \leq \frac{1}{\Omega(\tilde{\nu})}\bar{d}_{\mathcal{F},\mathbf{f}^\star}(\hat{\mathbf{h}}; \mathbf{h}^\star)$$

$\square$

Before showing the convexity-based lemmas used to control the representation differences in the loss we make a brief remark to interpret the logistic loss in the well-specified model.

Remark 1. If the data generating model satisfies the logistic model conditional likelihood as in Section 4.1, for the logistic loss ℓ we have that,

$$\mathbb{E}_{y \sim f \circ \mathbf{h}(\mathbf{x})}[\ell(\hat{f} \circ \hat{\mathbf{h}}(\mathbf{x}), y) - \ell(f \circ \mathbf{h}(\mathbf{x}), y)]] = \mathbb{E}_{\mathbf{x}}[\text{KL}[\text{Bern}(\sigma(f \circ \mathbf{h}(\mathbf{x})) \mid \text{Bern}(\sigma(\hat{f} \circ \hat{\mathbf{h}}(\mathbf{x}))]].$$

simply using the fact the data is generated from the model $y \sim \mathbb{P}_{y|\mathbf{x}}(\cdot|f \circ \mathbf{h}(\mathbf{x}))$.

To bound the task diversity we show a convexity-based lemma for general GLM/nonlinear models,

Lemma 2. *Consider the generalized linear model for which the $\mathbb{P}_{y|\mathbf{x}}(\cdot)$ distribution is,*

$$\mathbb{P}_{y|\mathbf{x}}(y_i|\boldsymbol{\alpha}^\top \mathbf{h}(\mathbf{x}_i)) = b(y_i)\exp\left(\frac{y_i\boldsymbol{\alpha}^\top \mathbf{h}(\mathbf{x}_i) - \Phi(\boldsymbol{\alpha}^\top \mathbf{h}(\mathbf{x}_i))}{s(\sigma)}\right).$$

Then if $\sup_{c \in S} \Phi''(c) = L(\mathbf{x}_i)$ and $\inf_{c \in S} \Phi''(c) = \mu(\mathbf{x}_i)$ where $c \in S = [\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}}, \mathbf{h}(\mathbf{x}_i)^\top \boldsymbol{\alpha}]$,

$$\frac{\mu(\mathbf{x}_i)}{2s(\sigma)}(\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}} - \mathbf{h}(\mathbf{x}_i)^\top \boldsymbol{\alpha})^2 \leq KL[\mathbb{P}_{y|\mathbf{x}}(\cdot|\boldsymbol{\alpha}^\top \mathbf{h}(\mathbf{x}_i)), \mathbb{P}_{y|\mathbf{x}}(\cdot|\hat{\boldsymbol{\alpha}}^\top \mathbf{h}^\star(\mathbf{x}_i))] \leq \frac{L(\mathbf{x}_i)}{2s(\sigma)}(\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}} - \mathbf{h}(\mathbf{x}_i)^\top \boldsymbol{\alpha})^2$$

where the KL is taken with respect to a fixed design point $\mathbf{x}_i$, and fixed feature functions $\mathbf{h}$, and $\hat{\mathbf{h}}$.

Proof.

$$\text{KL}[\mathbb{P}_{y|\mathbf{x}}(\cdot|\boldsymbol{\alpha}^\top \mathbf{h}(\mathbf{x}_i)), \mathbb{P}_{y|\mathbf{x}}(\cdot|\hat{\boldsymbol{\alpha}}^\top \hat{\mathbf{h}}(\mathbf{x}_i))] =$$
$$\int dy_i\, \mathbb{P}_{y|\mathbf{x}}(y_i|\boldsymbol{\alpha}^\top \mathbf{h}(\mathbf{x}_i))\left(\frac{y_i(\mathbf{h}(\mathbf{x}_i)^\top \boldsymbol{\alpha} - \hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}})}{s(\sigma)} + \frac{-\Phi(\mathbf{h}(\mathbf{x}_i)^\top \boldsymbol{\alpha}) + \Phi(\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}}))}{s(\sigma)}\right) =$$
$$\frac{1}{s(\sigma)}\left[\Phi'(\mathbf{h}(\mathbf{x}_i)^\top \boldsymbol{\alpha})(\mathbf{h}(\mathbf{x}_i)^\top \boldsymbol{\alpha} - \hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}}) - \Phi(\mathbf{h}(\mathbf{x}_i)^\top \boldsymbol{\alpha}) + \Phi(\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}})\right]$$

since we have that $\frac{\Phi(\mathbf{h}(\mathbf{x}_i)^\top \boldsymbol{\alpha})}{s(\sigma)} = \log \int dy_i\, b(y_i) \exp\left(\frac{y_i \mathbf{h}(\mathbf{x}_i)^\top \boldsymbol{\alpha}}{s(\sigma)}\right) \implies \frac{\Phi'(\mathbf{h}(\mathbf{x}_i)^\top \boldsymbol{\alpha})}{s(\sigma)} = \frac{\int dy_i\, \mathbb{P}_{y|\mathbf{x}}(y_i|\boldsymbol{\alpha}^\top \mathbf{h}(\mathbf{x}_i)) y_i}{s(\sigma)}$ as it is the log-normalizer. Using Taylor's theorem we have that

$$\Phi(\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}}) = \Phi\left(\mathbf{h}(\mathbf{x}_i)^\top \boldsymbol{\alpha}\right) + \Phi'(\mathbf{h}(\mathbf{x}_i)^\top \boldsymbol{\alpha})(\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}} - \mathbf{h}(\mathbf{x}_i)^\top \boldsymbol{\alpha}) + \frac{\Phi''(c)}{2}(\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}} - \mathbf{h}(\mathbf{x}_i)^\top \boldsymbol{\alpha})^2$$

for some intermediate $c \in [\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}}, \mathbf{h}(\mathbf{x}_i)^\top \boldsymbol{\alpha}]$. Combining the previous displays we obtain that:

$$\mathrm{KL}[\mathbb{P}_{y|\mathbf{x}}(\cdot|\boldsymbol{\alpha}^\top \mathbf{h}(\mathbf{x}_i)), \mathbb{P}_{y|\mathbf{x}}(\cdot|\hat{\boldsymbol{\alpha}}^\top \mathbf{h}(\mathbf{x}_i))] = \frac{1}{2s(\sigma)}\left[\Phi''(c)(\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}} - \mathbf{h}(\mathbf{x}_i)^\top \boldsymbol{\alpha})^2\right]$$

Now using the assumptions on the second derivative Φ'' gives,

$$\frac{\mu}{2s(\sigma)}(\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}} - \mathbf{h}(\mathbf{x}_i)^\top \boldsymbol{\alpha})^2 \leq \frac{1}{2s(\sigma)}\left[\Phi''(c)(\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}} - \mathbf{h}(\mathbf{x}_i)^\top \boldsymbol{\alpha})^2\right] \leq \frac{L}{2s(\sigma)}(\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}} - \mathbf{h}(\mathbf{x}_i)^\top \boldsymbol{\alpha})^2$$

$\square$

We now instantiate the aforementioned lemma in the setting of logistic regression.

Lemma 3. *Consider the $\mathbb{P}_{y|\mathbf{x}}(\cdot)$ logistic generative model defined in Section 4.1 for a general feature map $\mathbf{h}(\mathbf{x})$. Then for this conditional generative model in the setting of Lemma 2, where $\Phi(\eta) = \log(1 + \exp(\eta))$, $s(\sigma) = 1$, $b(y_i) = 1$,*

$$\sup_{c(\mathbf{x}) \in S(\mathbf{x})} \Phi''(c(\mathbf{x})) \leq \frac{1}{4}$$

and

$$\inf_{c(\mathbf{x}) \in S(\mathbf{x})} \Phi''(c(\mathbf{x})) \geq \frac{1}{4} \exp(-\max(|\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}}|, |\mathbf{h}(\mathbf{x}_i)^\top \boldsymbol{\alpha}|)).$$

for fixed $\mathbf{x}$.

Proof. A short computation shows $\Phi''(t) = \frac{e^t}{(e^t+1)^2}$. Note that the maxima of $\Phi''(t)$ over all $\mathbb{R}$ occurs at $t = 0$. Hence we have that, $\mathbb{E}_{\mathbf{x}}[\sup_{c \in S} \Phi''(c)] \leq \frac{1}{4}$ using a uniform upper bound. The lower bound follows by noting that

$$\inf_{c \in S} \Phi''(t) = \min(\Phi''(|\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}}|), \Phi''(|\mathbf{h}(\mathbf{x}_i)^\top \boldsymbol{\alpha})|)).$$

For the lower bound note that for $t > 0$ that $e^{2t} \geq e^t \geq 1$ implies that $\frac{e^t}{(1+e^t)^2} \geq \frac{1}{4}e^{-t}$. Since $\Phi''(t) = \Phi''(-t)$ it follows that $\Phi''(t) \geq \frac{1}{4}e^{-|t|}$ for all $t \in \mathbb{R}$. $\square$

Finally we include a simple auxiliary lemma to help upper bound the averages in our data-dependent bounds which relies on a simple tail bound for covariance matrices drawn from sub-gaussian ensembles (Vershynin [2010, Theorem 4.7.3, Exercise 4.7.1] or Wainwright [2019, Theorem 6.5]). Further recall that in Definition 4 our covariate distribution is $O(1)$-sub-gaussian.

Lemma 4. *Let the common covariate distribution $\mathbb{P}_{\mathbf{x}}(\cdot)$ satisfy Definition 4. Then if $nt \gtrsim d$,*

$$\mathbb{E}[\|\boldsymbol{\Sigma}_{\mathbf{X}}\|] \leq O(1),$$

if $n \gtrsim d + \log t$,

$$\mathbb{E}[\max_{j \in [t]} \|\boldsymbol{\Sigma}_{\mathbf{X}_j}\|] \leq O(1),$$

and if $m \gtrsim r$,

$$\max_{\mathbf{B} \in \mathcal{H}} \mathbb{E}[\|\mathbf{B}^\top \boldsymbol{\Sigma}_{\mathbf{X}_0} \mathbf{B}\|] \leq O(1),$$

where $\mathcal{H}$ is the set of $d \times r$ orthonormal matrices.

24

Proof. All of these statement essentially following by integrating a tail bound and applying the triangle inequality. For the first statement since $\mathbb{E}[\|\hat{\Sigma}_{\mathbf{X}}\|] = \mathbb{E}[\|\hat{\Sigma}_{\mathbf{X}} - \Sigma\|] + \|\Sigma\| \leq O(1)$ under the conditions $nt \gtrsim d$ of the result directly by Vershynin [2010, Theorem 4.7.3].

For the second by Wainwright [2019, Theorem 6.5], $\mathbb{E}[\exp(\lambda\|\hat{\Sigma} - \Sigma\|)] \leq \exp(c_0(\lambda^2/N) + 4d)]$ for all $|\lambda| \leq \frac{N}{c_2}$, for a sample covariance averaged over N datapoints. So using a union bound alongside a tail integration since the data is i.i.d. across tasks we have that,

$$\mathbb{E}[\max_{j\in[t]}\|\hat{\Sigma}_{\mathbf{X}_j} - \Sigma\|] \leq \int_0^\infty \min(1, t\mathbb{P}[\|\hat{\Sigma}_{\mathbf{X}_1} - \Sigma\| > \delta])d\delta \leq \int_0^\infty \min(1, \exp(c_0(\lambda^2/n) + 4d + \log t - \lambda\delta)] \leq$$

$$\int_0^\infty \min(1, \exp(4d + \log t) \cdot \exp(-c_1 \cdot n \min(\delta^2, \delta))d\delta \leq O\left(\sqrt{\frac{d + \log t}{n}} + \frac{d + \log t}{n}\right) \lesssim O(1)$$

via a Chernoff argument. The final inequality follows by bounding the tail integral and using the precondition $n \gtrsim d + \log t$. Centering the expectation and using the triangle inequality gives the conclusion.

For the last statement the crucial observation that allows the condition $m \gtrsim r$, is that $\mathbf{B}^\top\mathbf{x}_{0i}$, for all $i \in [m]$, is by definition an r-dimensional $O(1)$-sub-Gaussian random vector since $\mathbf{B}$ is an orthonormal projection matrix. Thus an identical argument to the first statement gives the result. $\square$

C.2 Deep Neural Network Regression

We first begin by assembling the results necessary to bound the Gaussian complexity of our deep neural network example. To begin we introduce a representative result from the which bounds the empirical Rademacher complexity of a deep neural network.

Theorem 8 (Theorem 2 adapted from Golowich et al. [2017]). *Let σ be a 1-Lipschitz activation function with $\sigma(0) = 0$, applied element-wise. Let $\mathcal{N}$ be the class of real-valued networks of depth K over the domain $\mathcal{X}$ with bounded data $\|\mathbf{x}_{ji}\| \leq D$, where $\|\mathbf{W}_k\|_{1,\infty} \leq M(k)$ for all $k \in [K]$. Then,*

$$\mathfrak{R}_n(\mathcal{N}; \mathbf{X}) \leq \left(\frac{2}{n}\Pi_{k=1}^K M(k)\right)\sqrt{(K + 1 + \log d) \cdot \max_{j\in[d]}\sum_{i=1}^n \mathbf{x}_{i,j}^2} \leq \frac{2D\sqrt{K + 1 + \log d} \cdot \Pi_{k=1}^K M(k)}{\sqrt{n}}.$$

where $\mathbf{x}_{i,j}$ denotes the j-th coordinate of the vector $\mathbf{x}_i$ and $\mathbf{X}$ is an $n \times d$ design matrix (with n datapoints).

With this result in hand we proceed to bound the Gaussian complexities for our deep neural network and prove Theorem 5. Note that we make use of the result $\hat{\mathfrak{R}}_{\mathbf{X}}(\mathcal{N}) \leq \sqrt{\frac{\pi}{2}} \cdot \hat{\mathfrak{G}}_{\mathbf{X}}(\mathcal{N})$ and that $\hat{\mathfrak{G}}_{\mathbf{X}}(\mathcal{N}) \leq 2\sqrt{\log N} \cdot \hat{\mathfrak{R}}_{\mathbf{X}}(\mathcal{N})$ for any function class $\mathcal{N}$ when $\mathbf{X}$ has N datapoints [Ledoux and Talagrand, 2013, p. 97].

Proof of Theorem 5. First note due to the task normalization conditions we can choose c_1, c_2 sufficiently large so that the realizability assumption in Assumption 2 is satisfied–in particular, we can assume that c_2 is chosen large enough to contain all the parameters $\boldsymbol{\alpha}_0^\star$ and c_1 large enough so that $c_1(1 \vee M(K)^2)$ is larger then the norms of the parameters $\boldsymbol{\alpha}_j^\star$ for $j \in [t]$.

Next recall that under the conditions of the result we can use Lemma 6 to verify the task diversity condition is satisfied with parameters $(\tilde{\nu}, 0)$ with $\tilde{\nu} = \sigma_r(\mathbf{A}^\top\mathbf{A}/t) > 0$. In particular under the conditions of the theorem we can verify the well-conditioning of the feature representation with $c = \Omega(1)$ which follows by definition of the set $\mathcal{H}$ and we can see that $\|\mathbb{E}_{\mathbf{x}}[\hat{\mathbf{h}}(\mathbf{x})\mathbf{h}^\star(\mathbf{x})^\top]\|_2 \leq \mathbb{E}_{\mathbf{x}}[\|\hat{\mathbf{h}}(\mathbf{x})\|\|\mathbf{h}^\star(\mathbf{x})\|] \leq O(M(K)^2)$ using the norm bound from Lemma 5. Hence under this setting we can choose c_1 sufficiently large so that $c_1 M(K)^2 \gtrsim O(M(K)^2)$. The condition $M(K) \gtrsim 1$ in the theorem statement is simply used to clean up the final bound.

In order to instantiate Theorem 3 we begin by bounding each of the complexity terms in the expression. First,

- For the feature learning complexity in the training phase we leverage Theorem 8 from Golowich et al. [2017] (which holds for scalar-valued outputs) in a modular fashion. For convenience let

nn $= \frac{2D\sqrt{K+1+\log d}\cdot \Pi_{k=1}^{K} M(k)}{\sqrt{nt}}$. To bound this term we simply pull the summation over the rank r outside the complexity and apply Theorem 8, so

$$\hat{\mathfrak{G}}_{\mathbf{X}}(\mathcal{H}) = \frac{1}{nt}\mathbb{E}[\sup_{\mathcal{W}_K} \sum_{l=1}^{r}\sum_{j=1}^{t}\sum_{i=1}^{n} g_{kji}\mathbf{h}_k(\mathbf{x}_{ji})] \leq \sum_{k=1}^{r} \hat{\mathfrak{G}}_{\mathbf{X}}(\mathbf{h}_k(\mathbf{x}_{ji})) \leq \log(nt) \cdot \sum_{k=1}^{r} \hat{\mathfrak{R}}_{\mathbf{X}}(h_k(\mathbf{x}_{ji})) \leq$$

$\log(nt) \cdot r \cdot$ nn

since under the weight norm constraints (i.e. the max ℓ_1 row norms are bounded) each component of the feature can be identically bounded. This immediately implies the population Gaussian complexity bound as the expectation over $\mathbf{X}$ is trivial. Further by definition the class $\mathcal{F}$ as linear maps with parameters $\|\boldsymbol{\alpha}\|_2 \leq M(K)^2$ we obtain that $L(\mathcal{F}) = O(M(K)^2)$.

- For the complexity of learning $\mathcal{F}$ we use the fact that for in the training phase we obtain,

$$\hat{\mathfrak{G}}_{\mathbf{X}_j}(\mathcal{F}) = \frac{1}{n}\mathbb{E}_{\mathbf{g}}[\sup_{\boldsymbol{\alpha}\in\mathcal{F}} \sum_{i=1}^{n} g_{ji}\boldsymbol{\alpha}^{\top}\mathbf{h}(\mathbf{x}_{ji})] = O\left(\frac{M(K)^2}{n}\mathbb{E}_{\mathbf{g}}[\|\sum_{i=1}^{n} g_{ji}\mathbf{h}(\mathbf{x}_{ji})\|]\right)$$

$$\leq O\left(\frac{M(K)^2}{n}\sqrt{\sum_{i=1}^{n} \|\mathbf{h}(\mathbf{x}_{ji})\|^2}\right) \leq O\left(\frac{M(K)^2}{\sqrt{n}} \max_i \|\mathbf{h}(\mathbf{x}_{ji})\|\right).$$

Now by appealing to the norm bounds on the feature map from Lemma 5 we have that $\max_{\mathbf{h}\in\mathcal{H}} \max_{\mathbf{X}_j} \max_i \|\mathbf{h}(\mathbf{x}_{ji})\| \lesssim M(K)$. Hence in conclusion we obtain the bound,

$$\bar{\mathfrak{G}}_n(\mathcal{F}) \leq O\left(\frac{M(K)^3}{\sqrt{n}}\right)$$

since the expectation is once again trivial.

- A nearly identical argument shows the complexity of learning $\mathcal{F}$ in the testing phase is,

$$\hat{\mathfrak{G}}_{\mathbf{X}_0}(\mathcal{F}) = \frac{1}{m}\mathbb{E}_{\mathbf{g}}\left[\sup_{\boldsymbol{\alpha}:\|\boldsymbol{\alpha}\|\leq c_1} \sum_{i=1}^{m} g_i\boldsymbol{\alpha}^{\top}\mathbf{h}(\mathbf{x}_{(0)i})\right] \leq \frac{c_1 M(K)^3}{\sqrt{m}}$$

from which the conclusion follows.

Finally we verify that Assumption 1 holds so as to use Theorem 3 to instantiate the end-to-end guarantee. The boundedness parameter is,

$$D_{\mathcal{X}} \leq M(K)^3$$

by Lemma 5 since it must be instantiated with $\boldsymbol{\alpha} \in \mathcal{F}$. For the ℓ_2 loss bounds, $\ell(\eta; y) = (y - \eta)^2$. Since $\nabla_\eta \ell(\eta; y) = 2(y - \eta) \leq O(N + |\eta|) = O(M(K)^3)$ where $|\eta| \leq |\boldsymbol{\alpha}^{\top}\mathbf{h}(\mathbf{x})| \leq O(M(K)^3)$ for $\boldsymbol{\alpha} \in \mathcal{F}, \mathbf{h} \in \mathcal{H}$ and $N = O(1)$ by Lemma 5. So it follows the loss is Lipschitz with $L = O(M(K)^3)$. Moreover by an analogous argument, $|\ell(\eta; y)| \leq O(M(K)^6)$ so it follows the loss is uniformly bounded with parameter $B = O(M(K)^6)$.

Assembling the previous bounds shows the transfer learning risk is bounded by.

$$\lesssim \frac{L}{\tilde{\nu}} \cdot \left(\log(nt) \cdot \left[\log(nt) \cdot r \cdot M(K)^2 \cdot nn + \frac{M(K)^3}{\sqrt{n}}\right]\right) + \frac{LM(K)^3}{\sqrt{m}}$$

$$+ \left(\frac{1}{\tilde{\nu}} \cdot \max\left(L \cdot \frac{M(K)^3}{(nt)^2}, B\sqrt{\frac{\log(1/\delta)}{nt}}\right) + B\sqrt{\frac{\log(1/\delta)}{m}}\right).$$

where nn $= \frac{2D\sqrt{K+1+\log d}\cdot \Pi_{k=1}^{K} M(k)}{\sqrt{nt}}$. Under the conditions of the result, the risk simplifies as in the theorem statement. □

We now state a simple result which allows us to bound the suprema of the empirical ℓ_2 norm (i.e. the $D_{\bar{\mathbf{X}}}$ parameter in Theorem 1) and activation outputs for various neural networks.

Lemma 5. *Let $\hat{\mathbf{h}}(\mathbf{x})$ be a vector-valued neural network of depth K taking the form in (7) with each $f_j \equiv \boldsymbol{\alpha}_j$ satisfying $\|\boldsymbol{\alpha}_j\| \le A$ with bounded data $\|\mathbf{x}\| \le D$. Then the boundedness parameter in the setting of Theorem 1 satisfies,*

$$D_{\mathcal{X}} \lesssim AD \cdot \Pi_{k=1}^K \|\mathbf{W}_k\|_2$$

If we further assume that $\sigma(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$ which is centered and 1-Lipschitz (i.e. the tanh activation function), then we obtain the further bounds that,

$$\|\mathbf{h}(\mathbf{x})\| \le \|\mathbf{W}_K\|_{\infty \to 2}$$

and

$$D_{\mathcal{X}} \lesssim A \cdot \|\mathbf{W}_K\|_{\infty \to 2}$$

which holds without requiring boundedness of $\mathbf{x}$. Note $\|\mathbf{W}_K\|_{\infty \to 2}$ is the induced ∞ to 2 operator norm.

Proof. For the purposes of induction let $\mathbf{r}_k(\cdot)$ denote the vector-valued output of the kth layer for $k \in [K]$. First note that the bound

$$D_{\mathcal{X}} \lesssim \sup_{\boldsymbol{\alpha}, \mathbf{h}, \mathbf{x}} (\boldsymbol{\alpha}^\top \mathbf{h}(\mathbf{x}))^2 \le \sup_{\mathbf{W}_k, \mathbf{x}} A^2 \|\mathbf{r}_K\|^2$$

Now, for the inductive step, $\|\mathbf{r}_K\|^2 = \|\mathbf{W}_K \sigma(\mathbf{W}_{K-1}\mathbf{r}_{K-1})\|^2 \le \|\mathbf{W}_K\|_2^2 \|\sigma(\mathbf{W}_{K-1}\mathbf{r}_{K-1})\|^2 \le \|\mathbf{W}_K\|_2^2 \|\mathbf{W}_{K-1}\mathbf{r}_{K-1}\|^2 \le \|\mathbf{W}_K\|_2^2 \|\mathbf{W}_{K-1}\|_2^2 \|\mathbf{r}_{K-1}\|^2$ where the first inequality follows because $\sigma(\cdot)$ is element-wise Lipschitz and zero-centered. Recursively applying this inequality to the base case where $\mathbf{r}_0 = \mathbf{x}$ gives the conclusion after taking square roots.

If we further assume that $\sigma(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$ which is centered and 1-Lipschitz (i.e. the tanh activation function) then we can obtain the following result by simply bounding the last layer by noting that $\|\mathbf{r}_{K-1}\|_\infty \le 1$. Then,

$$\|\mathbf{h}(\mathbf{x})\|^2 = \|\mathbf{r}_K\|_2^2 = \|\mathbf{W}_K \mathbf{r}_{K-1}\|_2^2 \le \|\mathbf{W}_K\|_{\infty \to 2}^2$$

where $\|\mathbf{W}_K\|_{\infty \to 2}$ is the induced ∞ to 2 operator norm $\qquad\square$

We now turn to proving a task diversity lower bound applicable to general ℓ_2 regression with general feature maps $\mathbf{h}(\cdot)$ under the assumptions of the $\mathbb{P}_{y|\mathbf{x}}$ of the generative model specified in (8). As our result holds only requiring $f_j^\star \equiv \boldsymbol{\alpha}_j^\star$ and applies to more then neural network features we define some generic notation.

We assume the data generating model takes the form,

$$y_{ji} = (\boldsymbol{\alpha}_j^\star)^\top \mathbf{h}^\star(\mathbf{x}_{ji}) + \eta_{ji} \text{ for } j \in \{1, \ldots, t\}, i \in \{1, \ldots, n\} \tag{12}$$

for η_{ji} with bounded second moments and independent of $\mathbf{x}_{ji}$. Here the shared feature representation $\mathbf{h}^\star(\cdot) \in \mathbb{R}^r$ is given by a generic function. In our generic framework we can identify $f_j^\star \equiv \boldsymbol{\alpha}_j^\star$ for $j \in \{1, \ldots, t\}$. As before we define the population task diversity matrix as $\mathbf{A} = (\boldsymbol{\alpha}_1^\star, \ldots, \boldsymbol{\alpha}_t^\star)^\top \in \mathbb{R}^{t \times r}$, $\mathbf{C} = \mathbf{A}^\top \mathbf{A}/t$ and $\tilde{\nu} = \sigma_r(\frac{\mathbf{A}^\top \mathbf{A}}{t})$. Given two feature representations $\hat{\mathbf{h}}(\cdot)$ and $\mathbf{h}^\star(\cdot)$, we can define their population covariance as,

$$\Lambda(\hat{\mathbf{h}}, \mathbf{h}^\star) = \begin{bmatrix} \mathbb{E}_{\mathbf{x}}[\hat{\mathbf{h}}(\mathbf{x})\hat{\mathbf{h}}(\mathbf{x})^\top] & \mathbb{E}_{\mathbf{x}}[\hat{\mathbf{h}}(\mathbf{x})\mathbf{h}^\star(\mathbf{x})^\top] \\ \mathbb{E}_{\mathbf{x}}[\mathbf{h}^\star(\mathbf{x})\hat{\mathbf{h}}(\mathbf{x})^\top] & \mathbb{E}_{\mathbf{x}}[\mathbf{h}^\star(\mathbf{x})\mathbf{h}^\star(\mathbf{x})^\top] \end{bmatrix} \equiv \begin{bmatrix} \mathbf{F}_{\hat{\mathbf{h}}\hat{\mathbf{h}}} & \mathbf{F}_{\hat{\mathbf{h}}\mathbf{h}^\star} \\ \mathbf{F}_{\mathbf{h}^\star \hat{\mathbf{h}}} & \mathbf{F}_{\mathbf{h}^\star \mathbf{h}^\star} \end{bmatrix} \succeq 0$$

and the generalized Schur complement of the representation of $\mathbf{h}^\star$ with respect to $\hat{\mathbf{h}}$ as,

$$\Lambda_{Sc}(\hat{\mathbf{h}}, \mathbf{h}^\star) = \mathbf{F}_{\mathbf{h}^\star \mathbf{h}^\star} - \mathbf{F}_{\mathbf{h}^\star \hat{\mathbf{h}}}(\mathbf{F}_{\hat{\mathbf{h}}\hat{\mathbf{h}}})^\dagger \mathbf{F}_{\hat{\mathbf{h}}\mathbf{h}^\star} \succeq 0.$$

We now instantiate the definition of task diversity in this setting. We assume that the universal constants c_2 and c_1 are large-enough to contain the true parameters $\boldsymbol{\alpha}_0^\star$ and $\boldsymbol{\alpha}_j^\star$ respectively for the following.

Lemma 6. *Consider the $\mathbb{P}_{y|\mathbf{x}}(\cdot)$ regression model defined in (12) and the loss function $\ell(\cdot, \cdot)$ taken as the squared ℓ_2 loss. Then for this conditional generative model with $\mathcal{F} = \{\boldsymbol{\alpha} : \boldsymbol{\alpha} \in \mathbb{R}^r\}$ and $\mathcal{F}_0 = \{\boldsymbol{\alpha} : \|\boldsymbol{\alpha}\|_2 \leq c_2\}$ the model is $(\frac{\check{\nu}}{c_2}, 0)$ diverse in the sense of Definition 3 and,*

$$d_{\mathcal{F}, \mathcal{F}_0}(\hat{\mathbf{h}}; \mathbf{h}^\star) = c_2 \cdot \sigma_{\max}(\Lambda_{sc}(\hat{\mathbf{h}}, \mathbf{h}^\star)); \quad \bar{d}_{\mathcal{F}, \mathbf{f}^\star}(\hat{\mathbf{h}}; \mathbf{h}^\star) = \text{tr}(\Lambda_{sc}(\hat{\mathbf{h}}, \mathbf{h}^\star)\mathbf{C}).$$

Moreover, if we assume the set of feature representations $\hat{\mathbf{h}} \in \mathcal{H}$ in the infima over $\hat{\mathbf{h}}$ are well-conditioned in the sense that $\sigma_r(\mathbb{E}_{\mathbf{x}}[\hat{\mathbf{h}}(\mathbf{x})\hat{\mathbf{h}}(\mathbf{x})^\top]) \geq c > 0$ and $\|\mathbb{E}_{\mathbf{x}}[\hat{\mathbf{h}}(\mathbf{x})\mathbf{h}^\star(\mathbf{x})^\top]\|_2 \leq C$, then if $\mathcal{F} = \{\boldsymbol{\alpha} : \|\boldsymbol{\alpha}\| \leq c_1\}$, $\mathcal{F}_0 = \{\boldsymbol{\alpha} : \|\boldsymbol{\alpha}\|_2 \leq c_2\}$ and $c_1 \geq \frac{C}{c}c_2$ the same conclusions hold.

Proof. We first bound the worst-case representation difference and then the task-averaged representation metric. For convenience we let $\mathbf{v}(\hat{\boldsymbol{\alpha}}, \boldsymbol{\alpha}) = \begin{bmatrix} \hat{\boldsymbol{\alpha}} \\ \boldsymbol{\alpha} \end{bmatrix}$ in the following. First, note that under the regression model defined with the squared ℓ_2 loss we have that,

$$\mathbb{E}_{\mathbf{x}, y \sim f \circ \mathbf{h}(\mathbf{x})}\left\{\ell(\hat{f} \circ \hat{\mathbf{h}}(\mathbf{x}), y) - \ell(f \circ \mathbf{h}(\mathbf{x}), y)\right\} = \mathbb{E}_{\mathbf{x}}[|\hat{\boldsymbol{\alpha}}^\top \hat{\mathbf{h}}(\mathbf{x}) - \boldsymbol{\alpha}^\top \mathbf{h}(\mathbf{x})|^2]$$

- the worst-case representation difference between two distinct feature representations $\mathbf{h}$ and $\mathbf{h}'$ becomes

$$d_{\mathcal{F}, \mathcal{F}_0}(\hat{\mathbf{h}}; \mathbf{h}^\star) = \sup_{\boldsymbol{\alpha}:\|\boldsymbol{\alpha}\|_2 \leq c_2} \inf_{\hat{\boldsymbol{\alpha}}} \mathbb{E}_{\mathbf{x}}|\hat{\mathbf{h}}(\mathbf{x})^\top \hat{\boldsymbol{\alpha}} - \mathbf{h}^\star(\mathbf{x})^\top \boldsymbol{\alpha}_0|^2 =$$

$$\sup_{\boldsymbol{\alpha}_0:\|\boldsymbol{\alpha}_0\|_2 \leq c_2} \inf_{\hat{\boldsymbol{\alpha}}}\{\mathbf{v}(\hat{\boldsymbol{\alpha}}, -\boldsymbol{\alpha})^\top \Lambda(\hat{\mathbf{h}}, \mathbf{h}^\star)\mathbf{v}(\hat{\boldsymbol{\alpha}}, -\boldsymbol{\alpha})\} = \sup_{\boldsymbol{\alpha}_0:\|\boldsymbol{\alpha}_0\|_2 \leq c_2} \inf_{\hat{\boldsymbol{\alpha}}}\{\mathbf{v}(\hat{\boldsymbol{\alpha}}, \boldsymbol{\alpha}_0)^\top \Lambda(\hat{\mathbf{h}}, \mathbf{h}^\star)\mathbf{v}(\hat{\boldsymbol{\alpha}}, \boldsymbol{\alpha}_0)\}.$$

Recognizing the inner infima as the partial minimization of a convex quadratic form (see for example Boyd and Vandenberghe [2004, Example 3.15, Appendix A.5.4]), we find that,

$$\inf_{\hat{\boldsymbol{\alpha}}}\{\mathbf{v}(\hat{\boldsymbol{\alpha}}, \boldsymbol{\alpha}_0)^\top \Lambda(\hat{\mathbf{h}}, \mathbf{h}^\star)\mathbf{v}(\hat{\boldsymbol{\alpha}}, \boldsymbol{\alpha}_0)\} = \boldsymbol{\alpha}_0^\top \Lambda_{sc}(\hat{\mathbf{h}}, \mathbf{h}^\star)\boldsymbol{\alpha}_0$$

Note that in order for the minimization be finite we require $\mathbf{F}_{\hat{\mathbf{h}}\hat{\mathbf{h}}} \succeq 0$ and that $\mathbf{F}_{\hat{\mathbf{h}}\mathbf{h}^\star}\boldsymbol{\alpha} \in \text{range}(\mathbf{F}_{\hat{\mathbf{h}}\hat{\mathbf{h}}})$ – which are both satisfied in our since they are constructed as expectations over appropriate rank-one operators. In this case, a sufficient condition for $\hat{\boldsymbol{\alpha}}$ to be an minimizer is that $\hat{\boldsymbol{\alpha}} = -\mathbf{F}_{\hat{\mathbf{h}}\hat{\mathbf{h}}}^\dagger \mathbf{F}_{\hat{\mathbf{h}}\mathbf{h}^\star}\boldsymbol{\alpha}$. Finally the suprema over $\boldsymbol{\alpha}$ can be computed using the variational characterization of the singular values.

$$\sup_{\boldsymbol{\alpha}_0:\|\boldsymbol{\alpha}_0\|_2 \leq c_2} \boldsymbol{\alpha}_0^\top \Lambda_{sc}(\hat{\mathbf{h}}, \mathbf{h}^\star)\boldsymbol{\alpha}_0 = c_2 \cdot \sigma_{\max}(\Lambda_{sc}(\hat{\mathbf{h}}, \mathbf{h}^\star))$$

- The task-averaged representation difference can be computed by similar means

$$\bar{d}_{\mathcal{F}, \mathbf{f}^\star}(\hat{\mathbf{h}}; \mathbf{h}^\star) = \frac{1}{t}\sum_{j=1}^{t} \inf_{\hat{\boldsymbol{\alpha}}} \mathbb{E}_{\mathbf{x}}|\hat{\mathbf{h}}(\mathbf{x})^\top \hat{\boldsymbol{\alpha}} - \mathbf{h}^\star(\mathbf{x})^\top \boldsymbol{\alpha}_j^\star|^2 = \frac{1}{t}\sum_{j=1}^{t} (\boldsymbol{\alpha}_j^\star)^\top \Lambda_{sc}(\hat{\mathbf{h}}, \mathbf{h}^\star)\boldsymbol{\alpha}_j^\star$$

$$= \text{tr}(\Lambda_{sc}(\hat{\mathbf{h}}, \mathbf{h}^\star)\mathbf{C})$$

Note that since $\Lambda_{sc}(\hat{\mathbf{h}}, \mathbf{h}^\star) \succeq 0$, and $\mathbf{C} \succeq 0$, by a corollary of the Von-Neumann trace inequality, we have that $\text{tr}(\Lambda_{sc}(\hat{\mathbf{h}}, \mathbf{h}^\star)\mathbf{C}) \geq \sum_{i=1}^{r} \sigma_i(\Lambda_{sc}(\hat{\mathbf{h}}, \mathbf{h}^\star))\sigma_{r-i+1}(\mathbf{C}) \geq \text{tr}(\Lambda_{sc}(\hat{\mathbf{h}}, \mathbf{h}^\star))\sigma_r(\mathbf{C}) \geq \sigma_{\max}(\Lambda_{sc}(\hat{\mathbf{h}}, \mathbf{h}^\star))\sigma_r(\mathbf{C}).$

Combining the above two results we can immediately conclude that,

$$d_{\mathcal{F}, \mathcal{F}_0}(\hat{\mathbf{h}}; \mathbf{h}^\star) = c_2\sigma_{\max}(\Lambda_{sc}(\hat{\mathbf{h}}, \mathbf{h}^\star)) \leq \frac{1}{\check{\nu}/c_2}\bar{d}_{\mathcal{F}, \mathbf{f}^\star}(\hat{\mathbf{h}}; \mathbf{h}^\star)$$

The second conclusion uses Lagrangian duality for the infima in both optimization problems for the worst-case and task-averaged representation differences. In particular, since the $\inf_{\hat{\boldsymbol{\alpha}}} \mathbb{E}_{\mathbf{x}}|\hat{\mathbf{h}}(\mathbf{x})^\top \hat{\boldsymbol{\alpha}} - \mathbf{h}^\star(\mathbf{x})^\top \boldsymbol{\alpha}|^2$ is a strongly-convex under the well-conditioned assumption, we have its unique minimizer is given by $\hat{\boldsymbol{\alpha}} = -(\mathbf{F}_{\hat{\mathbf{h}}\hat{\mathbf{h}}})^{-1}\mathbf{F}_{\hat{\mathbf{h}}\mathbf{h}^\star}\boldsymbol{\alpha}$; hence $\|\hat{\boldsymbol{\alpha}}\| \leq$

$\frac{C}{c}\|\boldsymbol{\alpha}\|$. Thus, if we consider the convex quadratically-constrained quadratic optimization problem $\inf_{\hat{\boldsymbol{\alpha}}:\|\hat{\boldsymbol{\alpha}}\|_2 \leq c_1} \mathbb{E}_{\mathbf{x}}|\hat{\mathbf{h}}(\mathbf{x})^\top \hat{\boldsymbol{\alpha}} - \mathbf{h}^\star(\mathbf{x})^\top \boldsymbol{\alpha}|^2$ and $c_1 \geq \frac{C}{c}\|\boldsymbol{\alpha}\|$ the constraint is inactive, and the constrained optimization problem is equivalent to the unconstrained optimization problem. Hence for the choice of $\mathcal{F} = \{\boldsymbol{\alpha} : \|\boldsymbol{\alpha}\| \leq c_1\}$ the infima in both the computation of the task-averaged distance and worst-case representation difference can be taken to be unconstrained. The second conclusion follows. $\qquad\square$

C.3 Index Models

We prove the general result which provides the end-to-end learning guarantee. Recall that we will use $\boldsymbol{\Sigma}_{\mathbf{X}}$ to refer the sample covariance over the the training phase data.

Proof of Theorem 6. First by definition of the sets $\mathcal{F}_0$ and $\mathcal{F}$ the realizability assumption holds true. Next recall that under the conditions of the result we can use Lemma 7 to verify the task diversity condition is satisfied with parameters $(\tilde{\nu}, \epsilon)$ with $\tilde{\nu} \geq \frac{1}{t}$. Note in fact we have the stronger guarantee $\tilde{\nu} \geq \frac{\|\mathbf{v}\|_1}{\|\mathbf{v}\|_\infty} \frac{1}{t}$ for $\mathbf{v}_j = \inf_{\hat{f}\in\mathcal{F}} \mathbb{E}_{\mathbf{x},\eta}[L(f_j^\star(\mathbf{b}^\star(\mathbf{x})) - \hat{f}(\hat{\mathbf{b}}(\mathbf{x})) + \eta)]$. So if $\mathbf{v}$ is well spread-out given a particular learned representation $\hat{\mathbf{b}}$, the quantity $\tilde{\nu}$ could be much larger in practice and the transfer more sample-efficient then the worst-case bound suggests.

In order to instantiate Theorem 3 we begin by bounding each of the complexity terms in the expression. First,

- For the feature learning complexity in the training phase standard manipulations give,

$$\hat{\mathfrak{G}}_{\mathbf{X}}(\mathcal{H}) \leq \frac{1}{nt}\mathbb{E}\left[\sup_{\mathbf{b}:\|\mathbf{b}\|_2 \leq W} \sum_{j=1}^{t}\sum_{i=1}^{n} g_{ji}\mathbf{b}^\top \mathbf{x}_{ji}\right] \leq \frac{W}{nt}\sqrt{\mathbb{E}[\|\sum_{j=1}^{t}\sum_{i=1}^{n} g_{ji}\mathbf{x}_{ji}\|_2^2]}$$

$$\leq \frac{W}{nt}\sqrt{\sum_{j=1}^{t}\sum_{i=1}^{n}\|\mathbf{x}_{ji}\|^2} = \sqrt{\frac{W^2\mathrm{tr}(\boldsymbol{\Sigma}_{\mathbf{X}})}{nt}}$$

  Further by definition the class $\mathcal{F}$ is 1-Lipschitz so $L(\mathcal{F}) = 1$. Taking expectations and using concavity of the $\sqrt{\cdot}$ yields the first term.

- For the complexity of learning $\mathcal{F}$ in the training phase we appeal to the Dudley entropy integral (see [Wainwright, 2019, Theorem 5.22]) and the metric entropy estimate from Kakade et al. [2011, Lemma 6(i)]. First note that $N(\mathcal{F}, d_{2,\mathbf{b}\mathbf{X}_j}, \epsilon) \leq N(\mathcal{F}, \|\cdot\|_\infty, \epsilon)$. By Kakade et al. [2011, Lemma 6(i)] $N(\mathcal{F}, \|\cdot\|_\infty, \epsilon) \leq \frac{1}{\epsilon}2^{2DW/\epsilon}$. So for all $0 \leq \epsilon \leq 1$,

$$\hat{\mathfrak{G}}_{\mathbf{Y}}(\mathcal{F}) \lesssim 4\epsilon + \frac{32}{\sqrt{n}}\int_{\epsilon/4}^{1}\sqrt{\log N(\mathcal{F}, \|\cdot\|_\infty, u)}du \lesssim \epsilon + \frac{1}{\sqrt{n}}\int_{\epsilon/4}^{1}\sqrt{\log\left(\frac{1}{u}\right) + \frac{2WD}{u}}du$$

$$\lesssim \epsilon + \frac{\sqrt{WD}}{\sqrt{n}}\int_{\epsilon/4}^{1}\frac{1}{u^{1/2}}du \lesssim \epsilon + \sqrt{\frac{WD}{n}}\cdot(2-\epsilon) \leq O\left(\sqrt{\frac{WD}{n}}\right)$$

  using the inequality that $\log(\frac{1}{u}) \leq 2\frac{WD}{u}$ and taking $\epsilon = 0$. This expression has no dependence on the input data or feature map so it immediately follows that,

$$\bar{\mathfrak{G}}_n(\mathcal{F}) \leq O\left(\sqrt{\frac{WD}{n}}\right)$$

- A nearly identical argument shows the complexity of learning $\mathcal{F}$ in the testing phase is,

$$\bar{\mathfrak{G}}_m(\mathcal{F}) \leq O\left(\sqrt{\frac{WD}{m}}\right)$$

Finally we verify that Assumption 1 holds so as to use Theorem 3 to instantiate the end-to-end guarantee. First the boundedness parameter becomes,

$$D_{\mathcal{X}} = 1$$

by definition since all the functions f are bounded between $[0, 1]$. Again, simply by definition the ℓ_1 norm is 1-Lipschitz in its first coordinate uniformly over the choice of its second coordinate. Moreover as the noise $\eta_{ij} = O(1)$, the loss is uniformly bounded by $O(1)$ so $B = O(1)$. Assembling the previous bounds and simplifying shows the transfer learning risk is bounded by,

$$\lesssim \frac{\log(nt)}{\tilde{\nu}} \cdot \left(\sqrt{\frac{W^2 \mathbb{E}_{\mathbf{X}}[\text{tr}(\boldsymbol{\Sigma}_{\mathbf{X}})]}{nt}} + \sqrt{\frac{WD}{n}} \right) + \sqrt{\frac{WD}{m}} + \frac{1}{(nt)^2} + \frac{1}{\tilde{\nu}} \sqrt{\frac{\log(1/\delta)}{nt}} + \sqrt{\frac{\log(1/\delta)}{m}} + \epsilon$$

If we hide all logarithmic factors, we can verify the noise-terms are all higher-order to get the simplified statement in the lemma. $\qquad\square$

We now introduce a generic bound to control the task diversity in a general setting. In the following recall $\mathcal{F}_t = \text{conv}\{f_1, \ldots, f_t\}$ where $f_j \in \mathcal{F}$ for $j \in [t]$ where $\mathcal{F}$ is a convex function class. Further, we define the ϵ-enlargement of $\mathcal{F}_t$ with respect to the sup-norm by $\mathcal{F}_{t,\epsilon} = \{f : \exists \tilde{f} \in \mathcal{F}_t \text{ such that } \sup_z |f(z) - f'(z)| \leq \epsilon\}$. We also assume the loss function $\ell(a, b) = L(a - b)$ for a positive, increasing function obeying a triangle inequality (i.e. a norm) for the following.

Our next results is generic and holds for all regression models of the form,

$$y = f(\mathbf{h}(\mathbf{x})) + \eta. \tag{13}$$

which encompasses the class of multi-index models.

Lemma 7. *In the aforementioned setting and consider the $\mathbb{P}_{y|\mathbf{x}}(\cdot)$ regression model defined in (13). If $\mathcal{F} = \mathcal{F}$ for a base, convex function class $\mathcal{F}$, and $\mathcal{F}_0 = \mathcal{F}_{t,\tilde{\epsilon}}$ the model is $(\tilde{\nu}, \tilde{\epsilon})$ diverse in the sense of Definition 3 for $\tilde{\nu} \geq \frac{1}{t}$.*

Proof. This result follows quickly from several properties of convex functions. First the mapping

$$(f, \hat{f}) \to \mathbb{E}_{\mathbf{x} \sim \mathbb{P}_{\mathbf{x}}(\cdot), y \sim \mathbb{P}_{y|\mathbf{x}}(f \circ \mathbf{h}(\mathbf{x}))} \left[\ell(\hat{f} \circ \hat{\mathbf{h}}(\mathbf{x}), y) - \ell(f \circ \mathbf{h}(\mathbf{x}), y) \right] =$$
$$\mathbb{E}_{\mathbf{x},n}[L(f(\mathbf{h}(\mathbf{x})) - \hat{f}(\hat{\mathbf{h}}(\mathbf{x})) + \eta)] - \mathbb{E}_n[L(n)]$$

is a jointly convex function of $(f, \hat{f})$. This follows since as an affine precomposition of a convex function, $L(f(\mathbf{h}(\mathbf{x})) - \hat{f}(\hat{\mathbf{h}}(\mathbf{x})) + \eta)$ is convex for all $\mathbf{x}, \eta$ and the expectation operator preserves convexity. Now by definition of $\mathcal{F}_{t,\tilde{\epsilon}}$, for all $f \in \mathcal{F}_{t,\tilde{\epsilon}}$ there exists $\tilde{f} \in \mathcal{F}_t$ such $\sup_z |f(z) - \tilde{f}(z)| \leq \tilde{\epsilon}$. Thus for all f we have that,

$$\mathbb{E}_{\mathbf{x},\eta}[L(f(\mathbf{h}(\mathbf{x})) - \hat{f}(\hat{\mathbf{h}}(\mathbf{x})) + \eta)] - \mathbb{E}_\eta[L(\eta)] \leq \mathbb{E}_{\mathbf{x},\eta}[L(\tilde{f}(\mathbf{h}(\mathbf{x})) - \hat{f}(\hat{\mathbf{h}}(\mathbf{x})) + \eta)] - \mathbb{E}_\eta[L(\eta)] + \tilde{\epsilon}$$

for $\tilde{f} \in \mathcal{F}_t$. Then since partial minimization of $\hat{f}$ over the convex set $\mathcal{F}$ of this jointly convex upper bound preserves convexity, we have that the mapping from f to $\inf_{\hat{f} \in \mathcal{F}} \mathbb{E}_{\mathbf{x},n}[L(f(\mathbf{h}(\mathbf{x})) - \hat{f}(\hat{\mathbf{h}}(\mathbf{x})) + \epsilon)] - \mathbb{E}_\eta[L(\eta)]$ is a convex function of f. Thus,

$$\inf_{\hat{f} \in \mathcal{F}} \mathbb{E}_{\mathbf{x},\eta}[L(f(\mathbf{h}(\mathbf{x})) - \hat{f}(\hat{\mathbf{h}}(\mathbf{x})) + \eta)] - \mathbb{E}_\eta[L(\eta)] \leq \inf_{\hat{f} \in \mathcal{F}} \mathbb{E}_{\mathbf{x},\eta}[L(\tilde{f}(\mathbf{h}(\mathbf{x})) - \hat{f}(\hat{\mathbf{h}}(\mathbf{x})) + \eta)] - \mathbb{E}_\eta[L(\eta)] + \tilde{\epsilon}$$

Now taking the suprema over $f \in \mathcal{F}_{t,\tilde{\epsilon}}$ gives,

$$\sup_{f \in \mathcal{F}_{t,\tilde{\epsilon}}} \inf_{\hat{f} \in \mathcal{F}} \mathbb{E}_{\mathbf{x},\eta}[L(f(\mathbf{h}(\mathbf{x})) - \hat{f}(\hat{\mathbf{h}}(\mathbf{x})) + \eta)] - \mathbb{E}_\eta[L(\eta)] \leq$$
$$\sup_{\tilde{f} \in \mathcal{F}_t} \inf_{\hat{f} \in \mathcal{F}} \mathbb{E}_{\mathbf{x},\eta}[L(\tilde{f}(\mathbf{h}(\mathbf{x})) - \hat{f}(\hat{\mathbf{h}}(\mathbf{x})) + \eta)] - \mathbb{E}_\eta[L(\eta)] + \tilde{\epsilon}$$

Finally, since the suprema of a a convex function over a convex hull generated by a finite set of points can be taken to occur at the generating set,

$$\sup_{\tilde{f} \in \mathcal{F}_t} \inf_{\hat{f} \in \mathcal{F}} \mathbb{E}_{\mathbf{x},\eta}[L(\tilde{f}(\mathbf{h}(\mathbf{x})) - \hat{f}(\hat{\mathbf{h}}(\mathbf{x})) + \eta)] = \max_{j \in [t]} \inf_{\hat{f} \in \mathcal{F}} \mathbb{E}_{\mathbf{x},\eta}[L(f_j(\mathbf{h}(\mathbf{x})) - \hat{f}(\hat{\mathbf{h}}(\mathbf{x})) + \eta)]$$

Finally, for a t-dimensional vector $\mathbf{v}$, $\|\mathbf{v}\|_\infty \leq \|\mathbf{v}\|_1$. Instantiating this with the vector with components $\mathbf{v}_j = \inf_{\hat{f} \in \mathcal{F}} \mathbb{E}_{\mathbf{x}, \eta}[L(f_j^\star(\mathbf{h}^\star(\mathbf{x})) - \hat{f}(\hat{\mathbf{h}}(\mathbf{x})) + \eta)]$ and combining with the above shows that[8],

$$d_{\mathcal{F}, \mathcal{F}_0}(\hat{\mathbf{h}}; \mathbf{h}^\star) \leq \bar{d}_{\mathcal{F}, \mathbf{f}^\star}(\hat{\mathbf{h}}; \mathbf{h}^\star) \cdot \frac{1}{\tilde{\nu}} + \epsilon$$

where $\tilde{\nu} \geq \frac{1}{t}$ (but might potentially be larger). Explicitly $\tilde{\nu} \geq \frac{1}{t} \frac{\|\mathbf{v}\|_1}{\|\mathbf{v}\|_\infty}$. In the case the vector $\mathbf{v}$ is well-spread out over its coordinates we expect the bound $\|\mathbf{v}\|_1 \geq \|\mathbf{v}\|_\infty$ to be quite loose and $\tilde{\nu}$ could be potentially much greater. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad\square$

Note if $\mathbf{v}$ is well-spread out – intuitively the problem possesses a problem-dependent "uniformity" and the bound $\tilde{\nu} \geq \frac{1}{t}$ is likely pessimistic. However, formalizing this notion in a clean way for nonparametric function classes considered herein seems quite difficult.

Also note the diversity bound of Lemma 7 is valid for *generic* functions and representations in addition to applying to a wide class of regression losses. In particular, all p-norms such $L(a, b) = \|a - b\|_p$. Further only mild moments boundedness conditions are required on ϵ to ensure finiteness of the objective.

---

[8]note the $\mathbb{E}_\eta[L(\eta)]$ terms cancel in the expressions for $d_{\mathcal{F}, \mathcal{F}_0}(\hat{\mathbf{h}}; \mathbf{h}^\star)$ and $\bar{d}_{\mathcal{F}, \mathbf{f}^\star}(\hat{\mathbf{h}}; \mathbf{h}^\star)$.