Title
Learning Beyond the Standard Model (of Data)

Author
Tripuraneni, Nilesh

Publication Date
2022

Peer reviewed|Thesis/dissertation

Learning Beyond the Standard Model (of Data)

by

Nilesh Tripuraneni

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Michael I. Jordan, Chair
Professor Prasad Raghavendra
Professor Jacob Steinhardt

Summer 2022

Learning Beyond the Standard Model (of Data)

Abstract

Learning Beyond the Standard Model (of Data)

by

Nilesh Tripuraneni

Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor Michael I. Jordan, Chair

Classically, most machine learning (ML) methodology has made an innocuous modeling assumption: data drawn from both the training/test sets has been independently sampled from a pair of identical distributions with nice properties. Yet, in the situations modern ML methods must confront, deviations from this idealized setting are quickly becoming the norm–not the exception. In this thesis, we address the challenges arising in understanding the often unexpected phenomenology in these settings by developing theory in two areas of interest: transfer learning and robust learning. In particular, we focus on identifying what structural conditions/techniques are needed to permit sample-efficient learning in these new settings, in order to answer questions such as why pretraining is so effective and what the limits of learning are for extremely heavy-tailed distributions.

To my family

Contents

List of Figures

Acknowledgments

My path to the Berkeley EECS Ph.D program was a winding one. The path through graduate school is no less winding, but I was fortunate to be surrounded by people whose mentorship and friendship made the path easier to walk.

Firstly, I would like to thank my advisor Mike for his support, optimism, and wisdom. Although in many instances, our research interests didn't align while I was in graduate school, Mike provided me unilateral freedom and universal support for any ideas (or internships) I pursued whilst here. I hope to emulate his big-picture thinking and (enviable) skills as both a writer and orator in some small measure as I go forward.

Mike also welcomed me to his group (SAIL) upon first arriving in Berkeley. Being a member of SAIL broadened the horizons of my knowledge far beyond what I had originally considered "machine learning". The breadth and sum of knowledge across the members of SAIL is staggering. Within SAIL, I also found an incredible set of collaborators (and mentors). I owe special thanks to several individuals I had the fortune of being seated next to in my early years in the RISE lab. Chi Jin (whom I collaborated with for much of the work in this thesis) served the dual roles of both a friend and mentor. From him I acquired not only a powerful set of technical skills but also a taste for "clean" algorithms and elegant arguments. In these early years I was also fortunate to collaborate with Nicolas Flammarion, who deepened my knowledge of optimization with his algorithmic expertise, and Jeffrey Regier who reminded me I was once (and a perhaps future) Bayesian. I benefitted from many other collaborations in the broader Berkeley community. The theory group (Prasad Raghavendra & Sam Hopkins in particular) showed me the sum-of-squares hierarchy needn't be so scary. My collaboration with Jeffrey Chan and Aldo Pacchiano on "bio-bandits" showed me you could combine bandits with just about anything. Meanwhile the only student "junior" to me I truly collaborated with, Yeshwanth Cherapanamjeri, taught me far more then I taught him about the nuances of heavy-tails. Jacob Steinhardt, Peter Bartlett, and Prasad Raghavendra have also provided insightful advice on both my work and career as members of my quals/thesis committee.

My first (& perhaps my last) teaching experiences occurred at Berkeley in my 3rd year of graduate school. Despite having multiple semesters of optimization and statistics under my belt, teaching EE 127 and Stats 210b under Laurent El Ghaoui, Alexandre Bayen and Mike's guidance was an incredible learning experience. Truly, *docendo discimus*.

I also owe thanks to the Cambridge Computational and Biological Learning Group as well as my M.Phil supervisor Zoubin Ghahramani for accepting me into their group when I knew a bit about physics but little about machine learning. They set me on my path.

During my Ph.D I was lucky to have the opportunity to intern several times in industry. Lester Mackey, who mentored me at Microsoft Research during our work on transductive prediction, continues to serve as a role model to which I aspire both for his technical brilliance and kindness. He remains the only person in machine learning I've met with the interest and ability to both write Annals of Probability papers and win weather forecasting competitions. At Google Brain, Jeffrey Pennington (& Ben Adlam) introduced me to the wonderful world

of random matrix theory. They taught me how to strike the right balance between intuition and rigor as well as the value of thinking like a physicist. Their patience and mathematical expertise throughout our remote collaboration led to a very cool paper about covariate shift. Last but not least, the forecasting team at Amazon SCOT – Dominique, Dhruv, and Dean in particular – gave me my first taste of real industry problems. Their intuitions for art of applied machine learning provided much-needed balance to my mostly theoretical education.

My years at Berkeley would've been far less enjoyable without the company of many friends. Yeshwanth Cherapanamjeri and Armin Askari (the Trailer South Park Boys list), have been great friends and foodies in addition to dutifully upvoting every meme I have ever sent. Aldo Pacchiano, Eric Mazumdar and the Ghass have similarly been a source of ghreat entertainment and culture in my years in Berkeley. Jeffrey Chan has always been a great lunch-time conversationalist and indulged my biocuriousity. Juanky Perdomo reminded me what it is to be young again. I also thank him for honorarily including me into his motley crew of community-college educated friends–which brought the blocking group experience across the coast too Berkeley. Some of my roommates over the years–namely Kush Bhatia and Vaishaal Shankar–have also been great friends, sources of advice and conversation (especially in pandemic times). I also owe thanks to several friends not in Berkeley who have always been around. I fondly recall all the Stanford Thanksgivings, Ani and (other) Armin hosted me for as well as the levity they have added to my life since college. Similarly, Diana has always been on GChat to commiserate about graduate school in machine learning. I also thank the various officemates, labmates and dumplings (who have doubled as friends) I've had the opportunity to interact with in the past years amongst many others: Robert Nishihara, Phil Moritz, Richard Liaw, Daniel Rothchild, Paras Jain, Ahmed El Alaoui, Ashia Wilson, Lydia Liu, Niladri Chatterji, Geoff Negiar, Romain Lopez, Colorado Reed, Melih Elibol, Esther Rolf, Xiang Cheng, Jianbo Chen, Lihua Lei, Horia Mania, Aadi Ramdas, Max Rabinovich, Max Simchowitz, Mitchell Stern, Karl Krauth, Akosua Busia, Clara Wong-Fannjiang, Chloe Hsu, Serena Wang, Neha Wadia, Tianyi Lin, Feng Ruan, Stephen Bates, Meena Jagadeesan, John Miller, Frances Ding, Alex Wei, Tijana Zrnic, and Paula Gradu. The former two whom I acknowledge as "my bros" at the latters behest. My various BAIR buddies: Allan Jabri, Ashish Kumar, Vitchyr Pong, Yu Sun, and Kelvin Xu (who's been a great source of job market advice) have also provided great company as well as answered many a question I've had about deep learning.

Last but not least, I am grateful for my family for their steadfast support and for always being there for me–no matter what path I've taken.

Chapter 1

Introduction

Machine learning (ML) has seen dramatic, empirical progress in recent years; with particularly notable successes in areas such as as image understanding and statistical machine translation [65]. Many of these banner successes have occurred in the regime of *static prediction*–that is, on problems possessing "nice", homogeneous data. Several key ingredients enabling success in this setting have been:

1. Access to large volumes of "nice" data (i.e. with a high signal-to-noise ratio).
2. A nearly independent, identically distributed (i.i.d.) data source (i.e. problems where the training and test data have come from essentially the same environment).

Yet, as the domains ML is applied become increasingly rich, a wealth of problems arise that move beyond this setting we refer to as the "Standard Model of Data". Beyond this setting, obstacles to learning, opportunities to accelerate learning, as well as rich class of phenomenon and methods that call for greater development and understanding emerge [100, 101, 86, 50, 23, 93, 62, 43]. The focus of this thesis is to mathematically investigate such phenomenon in two settings of interest: transfer learning and robust learning. Several questions have guided this work. In particular, we ask what (novel) structural conditions are needed to permit sample-efficient learning in settings that move beyond "nice" i.i.d. data? Moreover, what corresponding techniques/analysis must be developed to establish the algorithmic/statistical efficiency results that reflect the corresponding empirical phenomenology?

In order to precisely frame these questions we consider the fundamental data model $p(\mathbf{x}, y) = p(\mathbf{x})p(y|\mathbf{x})$–which breaks apart data into an input distribution over covariates ($p(\mathbf{x})$) and conditional distribution over labels ($p(y|\mathbf{x})$)–and lies at the core of modeling and analysis in ML. The "Standard Model" as we call it, operates in the setting that both these distributions are often Gaussian-like and identical across training and test tasks. In this thesis, we investigate the phenomenology that arises when ML deviates from these assumptions in two settings:

1. **Transfer learning** – where data from multiple tasks with varying (but related) $p(y|\mathbf{x})$ – are leveraged to improve the performance of prediction across each other (i.e. via procedures

like ImageNet pretraining). In this setting we have identified key structural conditions, such as diversity over multiple tasks, which allow the efficient transfer of information from data-rich source tasks to data-scarce target tasks. Perhaps surprisingly, our generalization bounds are the first to *decay with all of the training data samples*–which is necessary to reflect the practical efficacy of transfer learning across domains such as computer vision, natural language processing, and biology.

2. **Robust Learning** – where $p(\mathbf{x})$ and $p(y|\mathbf{x})$ may be heavy-tailed – so extreme events become commonplace. Here we have designed statistically optimal and computationally efficient methods for the simplest ML problem – mean estimation. We have established the statistical landscape of estimation for such problems is more nuanced than in the case of Gaussian-like data: since it is *not possible* to obtain Gaussian-like confidence intervals with widths $O(\sqrt{\frac{\text{DIMENSIONS}}{\text{NUMBER OF SAMPLES}}})$ for very heavy-tailed data. Far from being a theoretical curiosity, these results have algorithmic and statistical consequences in areas such as financial forecasting, AB testing of user data and even reinforcement learning [73, 81, 45] where such heavy-tails arise.

A central theme of this research has been to provide statistical understanding of the techniques and problem structure needed to make sense of the unexpected phenomenology and ML methodology in settings that move beyond "nice" i.i.d. data. Ultimately, this is not only a question of mathematics but of modeling. The results we have developed also lean on testable, data-driven conditions we believe can also be used to design more stable and efficient machine learning methods. In the following I outline these directions in greater detail and also suggest directions for further research.

Meta/Transfer Learning (Chapters 2 & 3):

In the setting of meta/transfer learning, data from multiple tasks with varying (but related) $p(y|\mathbf{x})$ is leveraged to improve the performance of prediction across each other. In the works [104, 105] we studied the paradigm of transfer learning achieved via a common, shared representation. This serves as a simple model to capture one of the most commonly used procedures in machine learning – ImageNet pretraining – where only the final layers of a neural network are retrained new task data, after initializing its earlier layers with hierarchical representations/features from ImageNet [38, 48]. This paradigm has also found widespread use in other areas such as deep reinforcement learning [4], and even protein engineering and design [42]. The work [104] studies this question in the setting of linear regression where computationally and statistically efficient algorithms for representation learning and transfer are provided. [105] generalizes these results to a generic setting with arbitrary tasks, features, and losses assuming access to an empirical risk minimization (ERM) oracle.

One principal contribution of this line of work is to introduce a problem-agnostic definition of task diversity which can be used provide generalization bounds for transfer learning problems with general losses, tasks, and features. Our framework uses this to provide guarantees of

a fast convergence rate, decaying with *all of the samples* for the transfer learning problem. Previous work in this vein proved bounds suggesting that increasing the number of samples per training task could *not* improve generalization on new tasks, which did not provide a fully satisfactory explanation for the widespread practical efficacy of transfer learning methods across numerous application domains such as computer vision, natural language processing, and biology. In addition, in [104] we shows this diversity condition is necessary for recovery of the underlying representation in the case of linear regression. From the technical stand point, our work [104] also provides provably, computationally efficient algorithms for the non-convex representation recovery task in the case of linear regression. While [105] also develops a novel chain rule for bounding the generalization error in learning problems with composite structure (i.e. where predictors are composed of a common representation and task-specific map). There exist numerous directions for continuing this line of work given that multi-task/transfer learning are heavily used across a variety of domains. One exciting direction is pursuing a theoretical understanding of the behavior of transfer learning in sequential learning settings (i.e. bandits/reinforcement learning). In this case, agents must additionally balance cooperative exploration (in order to share and efficiently transfer information) along with individual exploitation (optimizing their own reward) whilst handling the standard exploration/exploitation tradeoff. Similarly, exploring the utility of the task diversity concept developed in these works as an empirical metric to apriori gauge the efficacy of multi-task/transfer learning in could be useful for practical data analysis. Understanding the interaction of transfer learning/fine-tuning methods in problems with distributional shift is also exciting, as these algorithmic techniques inspired by transfer learning have shown promising practical utility for mitigating performance degradations that accompany covariate shift between training and test distributions [115].

Robust Learning (Chapter 4):

Even basic questions such as how to obtain (optimal) statistically and computationally efficient algorithms for amongst the simplest estimation problems in ML[1] have remained unanswered when the underlying distributions $p(\mathbf{x})$, $p(y|\mathbf{x})$ are heavy-tailed. The regime where $p(\mathbf{x})$, $p(y|\mathbf{x})$ are heavy tailed occurs routinely in applications such as financial forecasting, user data and even reinforcement learning [73, 81, 45]. The simplest instantiation of such a problem is estimating the mean of a distribution in d dimensions (which possesses a variance) with a failure probability less than δ. Remarkably, only recently were estimators proposed in [79, 51, 20], which lead to the optimal statistical rate of $O\left(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log 1/\delta}{n}}\right)$[2] in the high-probability regime. In [22], we have designed computationally-efficient algorithms (which are polynomial-time) to obtain optimal statistical rates for problems such as mean estimation in the absence

[1] Namely mean and covariance estimation along with linear regression.

[2] Note, the original estimator of [79] is not computationally-efficient. The works [51] reduced the runtime to $O(d^{28})$, while [20] further improved the runtime to $O(d^{3.5})$.

of a variance[3]. One theme of this work has been understanding when Gaussian-like rates of the form $O\left(\sqrt{\frac{d}{n}}\right)$ can be achieved–which we show is *not possible* for problems which move beyond mean estimation with a variance. In the future, extending the study of our methods to settings with heavy-tailed and correlated data is a promising direction for investigation. Similarly, understanding the behavior of more complex predictive models routinely used in machine learning (i.e. neural networks) under these extreme statistical conditions is important to designing robust, reliable methods.

[3]Data with such heavy-tails often occurs in user data [73] in large-scale technology/logistics companies.

Chapter 2

Meta-Learning Linear Representations

2.1 Introduction

The ability of a learner to transfer knowledge between tasks is crucial for robust, sample-efficient inference and prediction. One of the most well-known examples of such *transfer learning* has been in few-shot image classification where the idea is to initialize neural network weights in early layers using ImageNet pre-training/features, and subsequently re-train the final layers on a new task [38, 112]. However, the need for methods that can learn data representations that generalize to multiple, unseen tasks has also become vital in other applications, ranging from deep reinforcement learning [4] to natural language processing [3, 72]. Accordingly, researchers have begun to highlight the need to develop (and understand) generic algorithms for transfer (or meta) learning applicable in diverse domains [43]. Surprisingly, however, despite a long line of work on transfer learning, there is limited theoretical characterization of the underlying problem. Indeed, there are few efficient algorithms for feature learning that *provably* generalize to new, unseen tasks. Sharp guarantees are even lacking in the *linear* setting.

In the first chapter of this thesis, we study the problem of meta-learning of features in a linear model in which multiple tasks share a common set of low-dimensional features. Our aim is twofold. First, we ask: given a set of diverse samples from t different tasks how we can efficiently (and optimally) learn a common feature representation? Second, having learned a common feature representation, how can we use this representation to improve sample efficiency in a new $(t+1)$st task where data may be scarce?[1]

Formally, given an (unobserved) linear feature matrix $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_r) \in \mathbb{R}^{d \times r}$ with orthonormal columns, our statistical model for data pairs $(\mathbf{x}_i, y_i)$ is:

$$y_i = \mathbf{x}_i^\top \mathbf{B}\boldsymbol{\alpha}_{t(i)} + \epsilon_i \quad ; \quad \boldsymbol{\beta}_{t(i)} = \mathbf{B}\boldsymbol{\alpha}_{t(i)}, \tag{2.1}$$

where there are t (unobserved) underlying task parameters $\boldsymbol{\alpha}_j$ for $j \in \{1, \ldots, t\}$. Here $t(i) \in \{1, \ldots, t\}$ is the index of the task associated with the ith datapoint, $\mathbf{x}_i \in \mathbb{R}^d$ is a

[1] This problem is sometimes referred to as learning-to-learn (LTL).

random covariate, and ϵ_i is additive noise. We assume the sequence $\{\boldsymbol{\alpha}_{t(i)}\}_{i=1}^{\infty}$ is independent of all other randomness in the problem. In this framework, the aforementioned questions reduce to recovering $\mathbf{B}$ from data from the first $\{1, \ldots, t\}$ tasks, and using this feature representation to recover a better estimate of a new task parameter, $\boldsymbol{\beta}_{t+1} = \mathbf{B}\boldsymbol{\alpha}_{t+1}$, where $\boldsymbol{\alpha}_{t+1}$ is also unobserved.

Our main result targets the problem of learning-to-learn (LTL), and shows how a feature representation $\hat{\mathbf{B}}$ learned from t diverse tasks can improve learning on an unseen $(t+1)$st task which shares the same underlying linear representation. We informally state this result below.[2]

Theorem 2.1 (Informal)**.** *Suppose we are given n_1 total samples from t diverse and normalized tasks which are used in Algorithm 1 to learn a feature representation $\hat{\mathbf{B}}$, and n_2 samples from a new $(t+1)$st task which are used along with $\hat{\mathbf{B}}$ and Algorithm 2 to learn the parameters $\hat{\boldsymbol{\alpha}}$ of this new $(t+1)$st task. Then, the parameter $\hat{\mathbf{B}}\hat{\boldsymbol{\alpha}}$ has the following excess prediction error on a new test point $\mathbf{x}_\star$ drawn from the training data covariate distribution:*

$$\mathbb{E}_{\mathbf{x}_\star}[\langle \mathbf{x}_\star, \hat{\mathbf{B}}\hat{\boldsymbol{\alpha}} - \mathbf{B}\boldsymbol{\alpha}_{t+1}\rangle^2] \leqslant \tilde{O}\left(\frac{dr^2}{n_1} + \frac{r}{n_2}\right), \tag{2.2}$$

with high probability over the training data.

The naive complexity of linear regression which ignores the information from the previous t tasks has complexity $O(\frac{d}{n_2})$. Theorem 2.1 suggests that "positive" transfer from the first $\{1, \ldots, t\}$ tasks to the final $(t+1)$st task can dramatically reduce the sample complexity of learning when $r \ll d$ and $\frac{n_1}{n_2} \gg r^2$; that is, when (1) the complexity of the shared representation is much smaller than the dimension of the underlying space and (2) when the ratio of the number of samples used for feature learning to the number of samples present for a new unseen task exceeds the complexity of the shared representation. We believe that the LTL bound in Theorem 2.1 is the first bound, even in the *linear* setting, to sharply exhibit this phenomenon (see Section 2.1 for a detailed comparison to existing results). Prior work provides rates for which the leading term in (2.2) decays as $\sim \frac{1}{\sqrt{t}}$, not as $\sim \frac{1}{n_1}$. We identify structural conditions on the design of the covariates and diversity of the tasks that allow our algorithms to take full advantage of *all* samples available when learning the shared features. Our primary contributions in this paper are to:

- Establish that all local minimizers of the (regularized) empirical risk induced by (2.1) are close to the true linear representation up to a small, statistical error. This provides strong evidence that first-order algorithms, such as gradient descent [58], can efficiently recover good feature representations (see Section 2.3).
- Provide a method-of-moments estimator which can efficiently aggregate information across multiple differing tasks to estimate $\mathbf{B}$—even when it may be information-theoretically impossible to learn the parameters of any given task (see Section 2.3).

[2] Theorem 2.1 follows immediately from combining Theorems 2.3 and 2.4; see Theorem 2.6 for a formal statement.

- Demonstrate the benefits and pitfalls of transferring learned representations to new, unseen tasks by analyzing the bias-variance trade-offs of the linear regression estimator based on a biased, feature estimate (see Section 2.4).
- Develop an information-theoretic lower bound for the problem of feature learning, demonstrating that the aforementioned estimator is a close-to-optimal estimator of $\mathbf{B}$, up to logarithmic and conditioning/eigenvalue factors in the matrix of task parameters (see Assumption 2.2). To our knowledge, this is the first information-theoretic lower bound for representation learning in the multi-task setting (see Section 2.5).

Related Work

While there is a vast literature on papers proposing multi-task and transfer learning methods, the number of theoretical investigations is much smaller. An important early contribution is due to [6], who studied a model where tasks with shared representations are sampled from the same underlying environment. [96] and [84], using tools from empirical process theory, developed a generic and powerful framework to prove generalization bounds in multi-task and learning-to-learn settings that are related to ours. Indeed, the closest guarantee to that in our Theorem 2.1 that we are aware of is [84, Theorem 5]. Instantiated in our setting, [84, Theorem 5] provides an LTL guarantee showing that the excess risk of the loss function with learned representation on a new datapoint is bounded by $\tilde{O}(\frac{r\sqrt{d}}{\sqrt{t}} + \sqrt{\frac{r}{n_2}})$, with high probability. There are several principal differences between our work and results of this kind. First, we address the algorithmic component (or computational aspect) of meta-learning while the previous theoretical literature generally assumes access to a global empirical risk minimizer (ERM). Computing the ERM in these settings requires solving a *nonconvex* optimization problem that is in general NP hard. Second, in contrast to [84], we also provide guarantees for feature recovery in terms of the parameter estimation error—measured directly in the distance in the feature space.

Third, and most importantly, in [84], the leading term capturing the complexity of learning the feature representation decays *only in* t *but not in* n_1 (which is typically much larger than t). Although, as they remark, the $1/\sqrt{t}$ scaling they obtain is in general unimprovable in their setting, our results leverage assumptions on the distributional similarity between the underlying covariates $\mathbf{x}$ and the potential diversity of tasks to improve this scaling to $1/n_1$. That is, our algorithms make benefit of *all* the samples in the feature learning phase. We believe that for many settings (including the linear model that is our focus) such assumptions are natural and that our rates reflect the practical efficacy of meta-learning techniques. Indeed, transfer learning is often successful even when we are presented with only a few training tasks but with each having a significant number of samples per task (e.g., $n_1 \gg t$).[3]

There has also been a line of recent work providing guarantees for gradient-based meta-learning (MAML) [43]. [44, 60, 61], and [24] work in the framework of online convex

[3]See Fig. 2.3 for a numerical simulation relevant to this setting.

optimization (OCO) and use a notion of (a potentially data-dependent) task similarity that assumes closeness of all tasks to a single fixed point in parameter space to provide guarantees. In contrast to this work, we focus on the setting of learning a *representation* common to all tasks in a generative model. The task model parameters need not be close together in our setting.

In concurrent work, [40] obtain results similar to ours for multi-task linear regression and provide comparable guarantees for a two-layer ReLU network using a notion of training task diversity akin to ours. Their generalization bound for the two-layer ReLU network uses a distributional assumption over meta-test tasks, but they provide bounds for linear regression holding for both random and fixed meta-test tasks[4]. They provide purely statistical guarantees—assuming access to an ERM oracle for nonconvex optimization problems. Our focus is on providing sharp statistical rates for efficient algorithmic procedures (i.e., the method-of-moments and local minima reachable by gradient descent). Finally, we also show a (minimax)-lower bound for the problem of feature recovery (i.e., recovering $\mathbf{B}$).

2.2 Preliminaries

Throughout, we will use bold lower-case letters (e.g., $\mathbf{x}$) to refer to vectors and bold upper-case letters to refer to matrices (e.g., $\mathbf{X}$). We exclusively use $\mathbf{B} \in \mathbb{R}^{d\times r}$ to refer to a matrix with orthonormal columns spanning an r-dimensional feature space, and $\mathbf{B}_\perp$ to refer a matrix with orthonormal columns spanning the orthogonal subspace of this feature space. The norm $\|\cdot\|$ appearing on a vector or matrix refers to its ℓ_2 norm or spectral norm respectively. The notation $\|\cdot\|_{\mathrm{F}}$ refers to a Frobenius norm. $\langle \mathbf{x}, \mathbf{y}\rangle$ is the Euclidean inner product, while $\langle \mathbf{M}, \mathbf{N}\rangle = \mathrm{tr}(\mathbf{M}\mathbf{N}^\top)$ is the inner product between matrices. 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. 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. Our use of O, Ω, and Θ is otherwise standard.

Formally, an orthonormal feature matrix $\mathbf{B}$ is an element of an equivalence class (under right rotation) of a representative lying in $\mathrm{Gr}_{r,d}(\mathbb{R})$—the Grassmann manifold [41]. The Grassmann manifold, which we denote as $\mathrm{Gr}_{r,d}(\mathbb{R})$, consists of the set of r-dimensional subspaces within an underlying d-dimensional space. To define distance in $\mathrm{Gr}_{r,d}(\mathbb{R})$ we define the notion of a principal angle between two subspaces p and q. If $\mathbf{E}$ is an orthonormal matrix whose columns form an orthonormal basis of p and $\mathbf{F}$ is an orthonormal matrix whose columns form an orthonormal basis of q, then a singular value decomposition of $\mathbf{E}^\top\mathbf{F} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$ defines

[4] In a setting matching Theorem 2.1, they provide a guarantee of $\tilde{O}\left(dr^2/n_1 + tr^2/n_1 + r/n_2\right)$ for the ERM when $n_1 \gtrsim dr$ under sub-Gaussian covariate/Gaussian additive noise assumptions. Theorem 2.1 holds for the method-of-moments/linear regression estimator when $n_1 \gtrsim dr^2$ using a Gaussian covariate/sub-Gaussian additive noise assumption; the bound is free of the additional $\tilde{O}(tr^2/n_1)$ term which does not vanish as $t \to \infty$ for fixed t/n_1.

the principal angles as:

$$\mathbf{D} = \text{diag}(\cos\theta_1, \cos\theta_2, \ldots, \cos\theta_k),$$

where $0 \leqslant \theta_k \leqslant \ldots \leqslant \theta_1 \leqslant \frac{\pi}{2}$. The distance of interest for us will be the subspace angle distance $\sin\theta_1$, and for convenience we will use the shorthand $\sin\theta(\mathbf{E}, \mathbf{F})$ to refer to it. With some abuse of notation we will use $\mathbf{B}$ to refer to an explicit orthonormal feature matrix and the subspace in $\text{Gr}_{r,d}(\mathbb{R})$ it represents. We now detail several assumptions we use in our analysis.

Assumption 2.1 (Sub-Gaussian Design and Noise)**.** *The i.i.d. design vectors* $\mathbf{x}_i$ *are zero mean with covariance* $\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \mathbf{I}_d$ *and are* $\mathbf{I}_d$*-sub-gaussian, in the sense that* $\mathbb{E}[\exp(\mathbf{v}^\top \mathbf{x}_i)] \leqslant \exp\left(\frac{\|\mathbf{v}\|^2}{2}\right)$ *for all* $\mathbf{v}$*. Moreover, the additive noise variables* ϵ_i *are i.i.d. sub-gaussian with variance parameter* 1 *and are independent of* $\mathbf{x}_i$*.*

Throughout, we work in the setting of random design linear regression, and in this context Assumption 2.1 is standard. Our results do not critically rely on the identity covariance assumption although its use simplifies several technical arguments. In the following we define the population task diversity matrix as $\mathbf{A} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_t)^\top \in \mathbb{R}^{t \times r}$, $\nu = \sigma_r(\frac{\mathbf{A}^\top \mathbf{A}}{t})$, the average condition number as $\bar{\kappa} = \frac{\text{tr}(\frac{\mathbf{A}^\top \mathbf{A}}{t})}{r\nu}$, and the worst-case condition number as $\kappa = \sigma_1(\frac{\mathbf{A}^\top \mathbf{A}}{t})/\nu$.

Assumption 2.2 (Task Diversity and Normalization)**.** *The* t *underlying task parameters* $\boldsymbol{\alpha}_j$ *satisfy* $\|\boldsymbol{\alpha}_j\| = \Theta(1)$ *for all* $j \in \{1, \ldots, t\}$*. Moreover, we assume* $\nu > 0$*.*

Recovering the feature matrix $\mathbf{B}$ is impossible without structural conditions on $\mathbf{A}$. Consider the extreme case in which $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_t$ are restricted to span only the first $r-1$ columns of the column space of the feature matrix $\mathbf{B}$. None of the data points $(\mathbf{x}_i, y_i)$ contain any information about the rth column-feature which can be any arbitrary vector in the complementary $d-r-1$ subspace. In this case recovering $\mathbf{B}$ accurately is information-theoretically impossible. The parameters ν, $\bar{\kappa}$, and κ capture how "spread out" the tasks $\boldsymbol{\alpha}_j$ are in the column space of $\mathbf{B}$. The condition $\|\boldsymbol{\alpha}_j\| = \Theta(1)$ is also standard in the statistical literature and is equivalent to normalizing the signal-to-noise (snr) ratio to be $\Theta(1)$[5]. In linear models, the snr is defined as the square of the ℓ_2 norm of the underlying parameter divided by the variance of the additive noise.

Our overall approach to meta-learning of representations consists of two phases that we term "meta-train" and "meta-test". First, in the meta-train phase (see Section 2.3), we provide algorithms to learn the underlying linear representation from a set of diverse tasks. Second, in the meta-test phase (see Section 2.4) we show how to transfer these learned features to a new, unseen task to improve the sample complexity of learning. Detailed proofs of our main results can be found in the Appendix.

[5]Note that for a well-conditioned population task diversity matrix where $\bar{\kappa} \leqslant \kappa \leqslant O(1)$, our snr normalization enforces that $\text{tr}(\mathbf{A}^\top \mathbf{A}/t) = \Theta(1)$ and $\nu \geqslant \Omega(\frac{1}{r})$.

2.3 Meta-Train: Learning Linear Features

Here we address both the algorithmic and statistical challenges of provably learning the linear feature representation $\mathbf{B}$.

Local Minimizers of the Empirical Risk

The remarkable, practical success of first-order methods for training nonconvex optimization problems (including meta/multi-task learning objectives) motivates us to study the optimization landscape of the empirical risk induced by the model in (2.1). We show in this section that *all local minimizers* of a regularized version of empirical risk recover the true linear representation up to a small statistical error.

Jointly learning the population parameters $\mathbf{B}$ and $(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_t)^\top$ defined by (2.1) is reminiscent of a matrix sensing/completion problem. We leverage this connection for our analysis, building in particular on results from [46]. Throughout this section we assume that we are in a uniform task sampling model—at each iteration the task $t(i)$ for the ith datapoint is uniformly sampled from the t underlying tasks. We first recast our problem in the language of matrices, by defining the matrix we hope to recover as $\mathbf{M}_\star = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_t)^\top \mathbf{B}^\top \in \mathbb{R}^{t \times d}$. Since $\mathrm{rank}(\mathbf{M}_\star) = r$, we let $\mathbf{X}^\star \mathbf{D}^\star (\mathbf{Y}^\star)^\top = \mathrm{SVD}(\mathbf{M}_\star)$, and denote $\mathbf{U}^\star = \mathbf{X}^\star (\mathbf{D}^\star)^{1/2} \in \mathbb{R}^{t \times r}$, $\mathbf{V}^\star = (\mathbf{D}^\star)^{1/2} \mathbf{Y}^\star \in \mathbb{R}^{d \times r}$. In this notation, the responses of the regression model are written as follows:

$$y_i = \langle \mathbf{e}_{t(i)} \mathbf{x}_i^\top, \mathbf{M}_\star \rangle + \epsilon_i. \tag{2.3}$$

To frame recovery as an optimization problem we consider the Burer-Monteiro factorization of the parameter $\mathbf{M} = \mathbf{U}\mathbf{V}^\top$ where $\mathbf{U} \in \mathbb{R}^{t \times r}$ and $\mathbf{V} \in \mathbb{R}^{d \times r}$. This motivates the following objective:

$$\min_{\mathbf{U} \in \mathbb{R}^{t \times r}, \mathbf{V} \in \mathbb{R}^{d \times r}} f(\mathbf{U}, \mathbf{V}) = \frac{2t}{n} \sum_{i=1}^{n} (y_i - \langle \mathbf{e}_{t(i)} \mathbf{x}_i^\top, \mathbf{U}\mathbf{V}^\top \rangle)^2 + \frac{1}{2} \|\mathbf{U}^\top \mathbf{U} - \mathbf{V}^\top \mathbf{V}\|_{\mathrm{F}}^2. \tag{2.4}$$

The second term in (2.4) functions as a regularization to prevent solutions which send $\|\mathbf{U}\|_{\mathrm{F}} \to 0$ while $\|\mathbf{V}\|_{\mathrm{F}} \to \infty$ or vice versa. If the value of this objective (2.4) is small we might hope that an estimate of $\mathbf{B}$ can be extracted from the column space of the parameter $\mathbf{V}$, since the column space of $\mathbf{V}^\star$ spans the same subspace as $\mathbf{B}$. Informally, our main result states that all local minima of the regularized *empirical* risk are in the neighborhood of the optimal $\mathbf{V}^\star$, and have subspaces that approximate $\mathbf{B}$ well. Before stating our result we define the constraint set, which contains incoherent matrices with reasonable scales, as follows:

$$\mathcal{W} = \{ \ (\mathbf{U}, \mathbf{V}) \mid \max_{i \in [t]} \|\mathbf{e}_i^\top \mathbf{U}\|^2 \leqslant \frac{C_0 \bar{\kappa} r \sqrt{\kappa \nu}}{\sqrt{t}}, \\ \|\mathbf{U}\|^2 \leqslant C_0 \sqrt{t \kappa \nu}, \quad \|\mathbf{V}\|^2 \leqslant C_0 \sqrt{t \kappa \nu} \ \},$$

for some large constant C_0. Under Assumption 2.2, this set contains the optimal parameters. Note that $\mathbf{U}^\star$ and $\mathbf{V}^\star$ satisfy the final two constraints by definition and Lemma 2.16 can be used to show that Assumption 2.2 actually implies that $\mathbf{U}^\star$ is incoherent, which satisfies the first constraint. Our main result follows.

Theorem 2.2. *Let Assumptions 2.1 and 2.2 hold in the uniform task sampling model. If the number of samples n_1 satisfies $n_1 \gtrsim \mathrm{polylog}(n_1, d, t)(\kappa r)^4 \max\{t, d\}$, then, with probability at least $1 - 1/\mathrm{poly}(d)$, we have that given any local minimum $(\mathbf{U}, \mathbf{V}) \in int(\mathcal{W})$ of the objective (2.4), the column space of $\mathbf{V}$, spanned by the orthonormal feature matrix $\hat{\mathbf{B}}$, satisfies:*

$$\sin\theta(\hat{\mathbf{B}}, \mathbf{B}) \leqslant O\left(\frac{1}{\sqrt{\nu}}\sqrt{\frac{\max\{t, d\} r \log n_1}{n_1}}\right).$$

We make several comments on this result:

- The guarantee in Theorem 2.2 suggests that all local minimizers of the regularized empirical risk (2.4) will produce a linear representation at a distance at most $\tilde{O}(\sqrt{\max\{t, d\} r / n_1})$ from the true underlying representation. Theorem 2.5 guarantees that any estimator (including the empirical risk minimizer) must incur error $\gtrsim \sqrt{dr/n_1}$. Therefore, in the regime $t \leqslant O(d)$, all local minimizers are statistically close-to-optimal, up to logarithmic factors and conditioning/eigenvalue factors in the task diversity matrix.

- Combined with a recent line of results showing that (noisy) gradient descent can efficiently escape strict saddle points to find local minima [58], Theorem 2.2 provides strong evidence that first-order methods can efficiently meta-learn linear features.[6]

The proof of Theorem 2.2 is technical so we only sketch the high-level ideas. The overall strategy is to analyze the Hessian of the objective (2.4) at a stationary point $(\mathbf{U}, \mathbf{V})$ in $\mathrm{int}(\mathcal{W})$ to exhibit a direction Δ of negative curvature which can serve as a direction of local improvement pointing towards $\mathbf{M}^\star$ (and hence show $(\mathbf{U}, \mathbf{V})$ is not a local minimum). Implementing this idea requires surmounting several technical hurdles including (1) establishing various concentration of measure results (e.g., RIP-like conditions) for the sensing matrices $\mathbf{e}_{t(i)}\mathbf{x}_i^\top$ unique to our setting and (2) handling the interaction of the optimization analysis with the regularizer and noise terms. Performing this analysis establishes that under the aforementioned conditions all local minima in $\mathrm{int}(\mathcal{W})$ satisfy $\|\mathbf{U}\mathbf{V}^\top - \mathbf{M}^\star\|_F \leqslant O(\sqrt{t\frac{\max\{t,d\} r \log n_1}{n_1}})$ (see Theorem 2.8). Guaranteeing that this loss is small is not sufficient to ensure recovery of the underlying features. Transferring this guarantee in the Frobenius norm to a result on the subspace angle critically uses the task diversity assumption (see Lemma 2.15) to give the final result.

[6] To formally establish computational efficiency, we need to further verify the smoothness and the strict-saddle properties of the objective function (2.4) (see, e.g., [58]).

Algorithm 1 MoM Estimator for Learning Linear Features

Input: $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n_1}$.
$\hat{\mathbf{B}}\mathbf{D}_1\hat{\mathbf{B}}^\top \leftarrow$ top-r SVD of $\frac{1}{n_1} \cdot \sum_{i=1}^{n_1} y_i^2 \mathbf{x}_i \mathbf{x}_i^\top$
return $\hat{\mathbf{B}}$

Method-of-Moments Estimator

Next, we present a method-of-moments algorithm to recover the feature matrix $\mathbf{B}$ with sharper statistical guarantees. An alternative to optimization-based approaches such as maximum likelihood estimation, the method-of-moments is among the oldest statistical techniques [95] and has recently been used to estimate parameters in latent variable models [2].

As we will see, the technique is well-suited to our formulation of multi-task feature learning. We present our estimator in Algorithm 1, which simply computes the top-r eigenvectors of the matrix $(1/n_1)\sum_{i=1}^{n_1} y_i^2 \mathbf{x}_i \mathbf{x}_i^\top$. Before presenting our result, we define the averaged empirical task matrix as $\bar{\mathbf{\Lambda}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\alpha}_{t(i)} \boldsymbol{\alpha}_{t(i)}^\top$ where $\tilde{\nu} = \sigma_r(\bar{\mathbf{\Lambda}})$, and $\tilde{\kappa} = \mathrm{tr}(\bar{\mathbf{\Lambda}})/(r\tilde{\nu})$ in analogy with Assumption 2.2.

Theorem 2.3. *Suppose the n_1 data samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n_1}$ are generated from the model in (2.1) and that Assumptions 2.1 and 2.2 hold, but additionally that $\mathbf{x}_i \sim \mathcal{N}(0, \mathbf{I}_d)$. Then, if $n_1 \gtrsim \mathrm{polylog}(d, n_1) r d \tilde{\kappa}/\tilde{\nu}$, the output $\hat{\mathbf{B}}$ of Algorithm 1 satisfies*

$$\sin\theta(\hat{\mathbf{B}}, \mathbf{B}) \leqslant \tilde{O}\left(\sqrt{\frac{\tilde{\kappa}}{\tilde{\nu}}\frac{dr}{n_1}}\right),$$

with probability at least $1 - O(n_1^{-100})$. Moreover, if the number of samples generated from each task are equal (i.e., $\bar{\mathbf{\Lambda}} = \frac{1}{t}\mathbf{A}^\top\mathbf{A}$), then the aforementioned guarantee holds with $\tilde{\kappa} = \bar{\kappa}$ and $\tilde{\nu} = \nu$.

We first make several remarks regarding this result.

- Theorem 2.3 is flexible—the only dependence of the estimator on the distribution of samples across the various tasks is factored into the *empirical* task diversity parameters $\tilde{\nu}$ and $\tilde{\kappa}$. Under a uniform observation model the guarantee also immediately translates into an analogous statement which holds with the population task diversity parameters ν and $\bar{\kappa}$.

- Theorem 2.3 provides a non-trivial guarantee even in the setting where we only have $\Theta(1)$ samples from each task, but $t = \tilde{\Theta}(dr)$. In this setting, recovering the parameters of any given task is information-theoretically impossible. However, the method-of-moments estimator can efficiently aggregate information *across* the tasks to learn $\mathbf{B}$.

- The estimator does rely on the moment structure implicit in the Gaussian design to extract $\mathbf{B}$. However, Theorem 2.3 has no explicit dependence on t and is close-to-optimal in the constant-snr regime; see Theorem 2.5 for our lower bound.

Algorithm 2 Linear Regression for Learning a New Task with a Feature Estimate

Input: $\hat{\mathbf{B}}, \{(\mathbf{x}_i, y_i)\}_{i=1}^{n_2}$.
$\hat{\boldsymbol{\alpha}} \leftarrow (\sum_{i=1}^{n_2} \hat{\mathbf{B}}\mathbf{x}_i\mathbf{x}_i^\top\hat{\mathbf{B}}^\top)^\dagger\hat{\mathbf{B}}^\top \sum_{i=1}^{n_2} \mathbf{x}_i y_i$
return $\hat{\boldsymbol{\alpha}}$

We now provide a summary of the proof. Under oracle access to the population mean, $\mathbb{E}[\frac{1}{n}\sum_i y_i^2\mathbf{x}_i\mathbf{x}_i^\top] = (2\bar{\boldsymbol{\Gamma}} + (1+\mathrm{tr}(\bar{\boldsymbol{\Gamma}}))\mathbf{I}_d)$, where $\bar{\boldsymbol{\Gamma}} = \frac{1}{n}\sum_{i=1}^n \mathbf{B}\boldsymbol{\alpha}_{t(i)}\boldsymbol{\alpha}_{t(i)}^\top\mathbf{B}^\top$ (see Lemma 2.1), we can extract the features $\mathbf{B}$ by directly applying PCA to this matrix, under the condition that $\tilde{\kappa} > 0$, to extract its column space. In practice, we only have access to the samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$. Algorithm 1 uses the empirical moments $\frac{1}{n}\sum_i y_i^2\mathbf{x}_i\mathbf{x}_i^\top$ in lieu of the population mean. Thus, to show the result, we argue that $\frac{1}{n}\sum_{i=1}^n y_i^2\mathbf{x}_i\mathbf{x}_i^\top = \mathbb{E}[\frac{1}{n}\sum_{i=1}^n y_i^2\mathbf{x}_i\mathbf{x}_i^\top] + \mathbf{E}$ where $\|\mathbf{E}\|$ is a small, stochastic error (see Theorem 2.7). If this holds, the Davis-Kahan $\sin\theta$ theorem [9] shows that PCA applied to the empirical moments provides an accurate estimate of $\mathbf{B}$ under perturbation by a sufficiently small $\mathbf{E}$. The key technical step in this argument is to show sharp concentration (in spectral norm) of the matrix-valued noise terms contained in $\mathbf{E}$ which are neither bounded (in spectral norm) nor sub-gaussian/sub-exponential-like; we refer the reader to the Appendix for further details on this argument.

2.4 Meta-Test: Transfer of Features to New Tasks

Having estimated a linear feature representation $\hat{\mathbf{B}}$ shared across related tasks, our second goal is to transfer this representation to a new, unseen task—the $(t+1)$st task—to improve learning. In the context of the model in (2.1), the approach taken in Algorithm 2 uses $\hat{\mathbf{B}}$ as a plug-in surrogate for the unknown $\mathbf{B}$, and attempts to estimate $\boldsymbol{\alpha}_{t+1} \in \mathbb{R}^r$. Formally we define our estimator $\boldsymbol{\alpha}$ as follows:

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \|\mathbf{y} - \mathbf{X}\hat{\mathbf{B}}\boldsymbol{\alpha}\|^2, \tag{2.5}$$

where n_2 samples $(\mathbf{X}, \mathbf{y})$ are generated from the model in (2.1) from the $(t+1)$st task. Effectively, the feature representation $\hat{\mathbf{B}}$ performs dimension reduction on the input covariates $\mathbf{X}$, allowing us to learn in a lower-dimensional space. Our focus is on understanding the generalization properties of the estimator in Algorithm 2, since (2.5) is an ordinary least-squares objective which can be analytically solved.

Assuming we have produced an estimate $\hat{\mathbf{B}}$ of the true feature matrix $\mathbf{B}$, we can present our main result on the sample complexity of meta-learned linear regression.

Theorem 2.4. *Suppose n_2 data points, $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n_2}$, are generated from the model in (2.1), where Assumption 2.1 holds, from a single task satisfying $\|\boldsymbol{\alpha}_{t+1}\|^2 \leqslant O(1)$. Then, if $\sin\theta(\hat{\mathbf{B}}, \mathbf{B}) \leqslant \delta$ and $n_2 \gtrsim r\log n_2$, the output $\hat{\boldsymbol{\alpha}}$ from Algorithm 2 satisfies*

$$\|\hat{\mathbf{B}}\hat{\boldsymbol{\alpha}} - \mathbf{B}\boldsymbol{\alpha}_{t+1}\|^2 \leqslant \tilde{O}\left(\delta^2 + \frac{r}{n_2}\right), \tag{2.6}$$

with probability at least $1 - O(n_2^{-100})$.

Note that $\mathbf{B}\boldsymbol{\alpha}_{t+1}$ is simply the underlying parameter in the regression model in (2.1). We make several remarks about this result:

- Theorem 2.4 decomposes the error of transfer learning into two components. The first term, $\tilde{O}(\delta^2)$, arises from the bias of using an imperfect feature estimate $\hat{\mathbf{B}}$ to transfer knowledge across tasks. The second term, $\tilde{O}(\frac{r}{n_2})$, arises from the variance of learning in a space of reduced dimensionality.

- Standard generalization guarantees for random design linear regression ensure that the parameter recovery error is bounded by $O(\frac{d}{n_2})$ w.h.p. under the same assumptions [55]. Meta-learning of the linear representation $\hat{\mathbf{B}}$ can provide a significant reduction in the sample complexity of learning when $\delta^2 \ll \frac{d}{n_2}$ and $r \ll d$.

- Conversely, if $\delta^2 \gg \frac{d}{n_2}$ the bounds in (2.6) imply that the overhead of learning the feature representation may overwhelm the potential benefits of transfer learning (with respect to baseline of learning the $(t+1)$st task in isolation). This accords with the well-documented empirical phenomena of "negative" transfer observed in large-scale deep learning problems where meta/transfer-learning techniques actually result in a degradation in performance on new tasks [114]. For diverse tasks (i.e. $\kappa \leqslant O(1)$), using Algorithm 1 to estimate $\hat{\mathbf{B}}$ suggests that ensuring $\delta^2 \ll \frac{d}{n_2}$, where $\delta^2 = \tilde{O}(\frac{dr}{\nu n_1})$, requires $\frac{n_1}{n_2} \gg r/\nu$. That is, the ratio of the number of samples used for feature learning to the number of samples used for learning the new task should exceed the complexity of the feature representation to achieve "positive" transfer.

In order to obtain the rate in Theorem 2.4 we use a bias-variance analysis of the estimator error $\hat{\mathbf{B}}\hat{\boldsymbol{\alpha}} - \mathbf{B}\boldsymbol{\alpha}_{t+1}$ (and do not appeal to uniform convergence arguments). Using the definition of $\mathbf{y}$ we can write the error as,

$$\begin{aligned}\hat{\mathbf{B}}\hat{\boldsymbol{\alpha}} - \mathbf{B}\boldsymbol{\alpha}_0 &= \hat{\mathbf{B}}(\hat{\mathbf{B}}^\top \mathbf{X}^\top \mathbf{X}\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}\mathbf{X}^\top \mathbf{X}\mathbf{B}\boldsymbol{\alpha}_0 \\ &\quad - \mathbf{B}\boldsymbol{\alpha}_0 + \hat{\mathbf{B}}(\hat{\mathbf{B}}^\top \mathbf{X}^\top \mathbf{X}\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}^\top \mathbf{X}^\top \boldsymbol{\epsilon}.\end{aligned}$$

The first term contributes the bias term to (2.6) while the second contributes the variance term. Analyzing the fluctuations of the (mean-zero) variance term can be done by controlling the norm of its square, $\boldsymbol{\epsilon}^\top \mathbf{A}\boldsymbol{\epsilon}$, where $\mathbf{A} = \mathbf{X}\hat{\mathbf{B}}(\hat{\mathbf{B}}^\top \mathbf{X}^\top \mathbf{X}\hat{\mathbf{B}})^{-2}\hat{\mathbf{B}}^\top \mathbf{X}^\top$. We can bound this (random) quadratic form by first appealing to the Hanson-Wright inequality to show w.h.p. that $\boldsymbol{\epsilon}^\top \mathbf{A}\boldsymbol{\epsilon} \lesssim \mathrm{tr}(\mathbf{A}) + \tilde{O}(\|\mathbf{A}\|_{\mathrm{F}} + \|\mathbf{A}\|)$. The remaining randomness in $\mathbf{A}$ can be controlled using matrix concentration/perturbation arguments (see Lemma 2.17).

With access to the true feature matrix $\hat{\mathbf{B}}$ (i.e., setting $\hat{\mathbf{B}} = \mathbf{B}$) the term $\hat{\mathbf{B}}(\mathbf{B}^\top \mathbf{X}^\top \mathbf{X}\mathbf{B})^{-1}\mathbf{B}\mathbf{X}^\top \mathbf{X}\mathbf{B}\boldsymbol{\alpha}_0 - \mathbf{B}\boldsymbol{\alpha}_0 = 0$, due to the cancellation in the empirical covariance matrices, $(\mathbf{B}^\top \mathbf{X}^\top \mathbf{X}\mathbf{B})^{-1}\mathbf{B}\mathbf{X}^\top \mathbf{X}\mathbf{B} = \mathbf{I}_r$. This cancellation of the empirical covariance is essential to obtaining a tight analysis of the least-squares estimator. We cannot rely on this

effect in full since $\hat{\mathbf{B}} \neq \mathbf{B}$. However, a naive analysis which splits these terms, $(\hat{\mathbf{B}}^\top \mathbf{X}^\top \mathbf{X}\hat{\mathbf{B}})^{-1}$ and $\hat{\mathbf{B}}\mathbf{X}^\top \mathbf{X}\mathbf{B}$ can lead to a large increase in the variance in the bound. To exploit the fact $\hat{\mathbf{B}} \approx \mathbf{B}$, we project the matrix $\mathbf{B}$ in the leading $\mathbf{XB}$ term onto the column space of $\hat{\mathbf{B}}$ and its complement—which allows a partial cancellation of the empirical covariances in the subspace spanned by $\hat{\mathbf{B}}$. The remaining variance can be controlled as in the previous term (see Lemma 2.18).

2.5 Lower Bounds for Feature Learning

To complement the upper bounds provided in the previous section, in this section we derive information-theoretic limits for feature learning in the model (2.1). To our knowledge, these results provide the first sample-complexity lower bounds for feature learning, with regards to subspace recovery, in the multi-task setting. While there is existing literature on (minimax)-optimal estimation of low-rank matrices (see, for example, [102]), that work focuses on the (high-dimensional) estimation of matrices, whose only constraint is to be low rank. Moreover, error is measured in the additive prediction norm. In our setting, we must handle the additional difficulties arising from the fact that we are interested in (1) learning a column space (i.e., an element in the $\mathrm{Gr}_{r,d}(\mathbb{R})$) and (2) the error between such representatives is measured in the subspace angle distance. We begin by presenting our lower bound for feature recovery.

Theorem 2.5. *Suppose a total of n data points are generated from the model in (2.1) satisfying Assumption 2.1 with $\mathbf{x}_i \sim \mathcal{N}(0, \mathbf{I}_d)$, $\epsilon_i \sim \mathcal{N}(0,1)$, with an equal number from each task, and that Assumption 2.2 holds with $\boldsymbol{\alpha}_j$ for each task normalized to $\|\boldsymbol{\alpha}_j\| = \frac{1}{2}$. Then, there are $\boldsymbol{\alpha}_j$ for $r \leqslant \frac{d}{2}$ and $n \geqslant \max\left(\frac{1}{8\nu}, r(d-r)\right)$ so that:*

$$\inf_{\hat{\mathbf{B}}} \sup_{\mathbf{B} \in Gr_{r,d}(\mathbb{R})} \sin\theta(\hat{\mathbf{B}}, \mathbf{B}) \geqslant \Omega\left(\max\left(\sqrt{\frac{1}{\nu}}\sqrt{\frac{1}{n}}, \sqrt{\frac{dr}{n}}\right)\right),$$

with probability at least $\frac{1}{4}$, where the infimum is taken over the class of estimators that are functions of the n data points.

Again we make several comments on the result.

- The result of Theorem 2.5 shows that the estimator in Algorithm 1 provides a close-to-optimal estimator of the feature representation parameterized by $\mathbf{B}$–up to logarithmic and conditioning factors (i.e. κ, ν)[7] in the task diversity matrix–that is independent of the task number t. Note that under the normalization for $\boldsymbol{\alpha}_i$, as $\kappa \to \infty$ (i.e. the task matrix $\mathbf{A}$ becomes ill-conditioned) we have that $\nu \to 0$. So the first term in Theorem 2.5 establishes that task diversity is necessary for recovery of the subspace $\mathbf{B}$.

[7] Note in the setting that $\kappa \leqslant O(1)$, $\nu \sim \frac{1}{r}$.

- The dimension of $\mathrm{Gr}_{r,d}(\mathbb{R})$ (i.e., the number of free parameters needed to specify a feature set) is $r(d-r) \geqslant \Omega(dr)$ for $d/2 \geqslant r$; hence the second term in Theorem 2.5 matches the scaling that we intuit from parameter counting.

- Obtaining tight dependence of our subspace recovery bounds on conditioning factors in the task diversity matrix (i.e. κ, ν) is an important and challenging research question. We believe the gap between in conditioning/eigenvalue factors between Theorem 2.3 and Theorem 2.5 on the $\sqrt{dr/n}$ term is related to a problem that persists for classical estimators in linear regression (i.e. for the Lasso estimator in sparse linear regression). Even in this setting, a gap remains with respect to condition number/eigenvalue factors of the data design matrix $\mathbf{X}$, between existing upper and lower bounds (see [18, Section 7], [98, Theorem 1, Theorem 2] and [117] for example). In our setting the task diversity matrix $\mathbf{A}$ enters into the problem in a similar fashion to the data design matrix $\mathbf{X}$ in these aforementioned settings.

The dependency on the task diversity parameter $\frac{1}{\nu}$ (the first term in Theorem 2.5) is achieved by constructing a pair of feature matrices and an ill-conditioned task matrix $\mathbf{A}$ that cannot discern the direction along which they defer. The proof strategy to capture the second term uses a f-divergence based minimax technique from [49] (restated in Lemma 2.20 in the Appendix), similar in spirit to the global Fano (or Yang-Barron).

There are two key ingredients to using Lemma 2.20 and obtaining a tight lower bound. First, we must exhibit a large family of distinct, well-separated feature matrices $\{\mathbf{B}_i\}_{i=1}^M$ (i.e., a packing at scale η). Second, we must argue this set of feature matrices induces a family of distributions over data $\{(\mathbf{x}_i, y_i)\}_{B_i}$ which are statistically "close" and fundamentally difficult to distinguish amongst. This is captured by the fact the ϵ-covering number, measured in the space of *distributions* with divergence measure $D_f(\cdot,\cdot)$, is small. The standard (global) Fano method, or Yang-Barron method (see [113, Ch. 15]), which uses the KL divergence to measure distance in the space of measures, is known to provide rate-suboptimal lower bounds for parametric estimation problems.[8] Our case is no exception. To circumvent this difficulty we use the framework of [49], instantiated with the f-divergence chosen as the χ^2-divergence, to obtain a tight lower bound.

The argument proceeds in two steps. First, although the geometry of $\mathrm{Gr}_{r,d}(\mathbb{R})$ is complex, we can adapt results from [94] to provide sharp upper/lower bounds on the metric entropy (or global entropy) of the Grassmann manifold (see Proposition 2.9). The second technical step of the argument hinges on the ability to cover the space of distributions parametrized by $\mathbf{B}$ in the space of measures $\{\mathbf{Pr}_{\mathbf{B}} : \mathbf{B} \in \mathrm{Gr}_{r,d}(\mathbb{R})\}$—with distance measured by an appropriate f-divergence. In order to establish a covering in the space of measures parametrized by $\mathbf{B}$, the key step is to bound the distance $\chi^2(\mathbf{Pr}_{\mathbf{B}^1}, \mathbf{Pr}_{\mathbf{B}^2})$ for two different measures over data generated from the model (2.1) with two different feature matrices $\mathbf{B}^1$ and $\mathbf{B}^2$ (see Lemma 2.21). This control can be achieved in our random design setting by exploiting the

[8] Even for the simple problems of Gaussian mean estimation the classical Yang-Barron method is suboptimal; see [49] for more details.

Gaussianity of the marginals over data $\mathbf{X}$ and the Gaussianity of the conditionals of $\mathbf{y}|\mathbf{X}, \mathbf{B}$, to ultimately be expressed as a function of the angular distance between $\mathbf{B}^1$ and $\mathbf{B}^2$.

2.6 Simulations

We complement our theoretical analysis with a series of numerical experiments highlighting the benefits (and limits) of meta-learning[9]. For the purposes of feature learning we compare the performance of the method-of-moments estimator in Algorithm 1 vs. directly optimizing the objective in (2.4). Additional details on our set-up are provided in Section 2.14. We construct problem instances by generating Gaussian covariates and noise as $\mathbf{x}_i \sim \mathcal{N}(0, \mathbf{I}_d)$, $\epsilon_i \sim \mathcal{N}(0, 1)$, and the tasks and features used for the first-stage feature estimation as $\boldsymbol{\alpha}_i \sim \frac{1}{\sqrt{r}} \cdot \mathcal{N}(0, \mathbf{I}_r)$, with $\mathbf{B}$ generated as a (uniform) random r-dimensional subspace of $\mathbb{R}^d$. In all our experiments we generate an equal number of samples n_t for each of the t tasks, so $n_1 = t \cdot n_t$. In the second stage we generate a new, $(t+1)$st task instance using the same feature estimate $\mathbf{B}$ used in the first stage and otherwise generate n_2 samples, with the covariates, noise and $\boldsymbol{\alpha}_{t+1}$ constructed as before. Throughout this section we refer to features learned via a first-order gradient method as LF-FO and the corresponding meta-learned regression parameter on a new task by meta-LR-FO. We use LF-MoM and meta-LR-MoM to refer to the same quantities save with the feature estimate learned via the method-of-moments estimator. We also use LR to refer to the baseline linear regression estimator on a new task which only uses data generated from that task.

We begin by considering a challenging setting for feature learning where $d = 100$, $r = 5$, but $n_t = 5$ for varying numbers of tasks t. As Fig. 2.1 demonstrates, the method-of-moments estimator is able to aggregate information across the tasks as t increases to slowly improve its feature estimate, even though $n_t \ll d$. The loss-based approach struggles to improve its estimate of the feature matrix $\mathbf{B}$ in this regime. This accords with the extra t dependence in Theorem 2.2 relative to Theorem 2.3. In this setting, we also generated a $(t+1)$st test task with $d \ll n_2 = 2500$, to test the effect of meta-learning the linear representation on generalization in a new, unseen task against a baseline which simply performs a regression on this new task in isolation. Fig. 2.1 also shows that meta-learned regressions perform significantly worse than simply ignoring first t tasks. Theorem 2.4 indicates the bias from the inability to learn an accurate feature estimate of $\mathbf{B}$ overwhelms the benefits of transfer learning. In this regime $n_2 \gg d$ so the new task can be efficiently learned in isolation. We believe this simulation represents a simple instance of the empirically observed phenomena of "negative" transfer [114].

We now turn to the more interesting use cases where meta-learning is a powerful tool. We consider a setting where $d = 100$, $r = 5$, and $n_t = 25$ for varying numbers of tasks t. However, now we consider a new, unseen task where data is scarce: $n_2 = 25 < d$. As Fig. 2.2 shows, in this regime both the method-of-moments estimator and the loss-based approach

[9] An open-source Python implementation to reproduce our experiments can be found at https://github.com/nileshtrip/MTL.

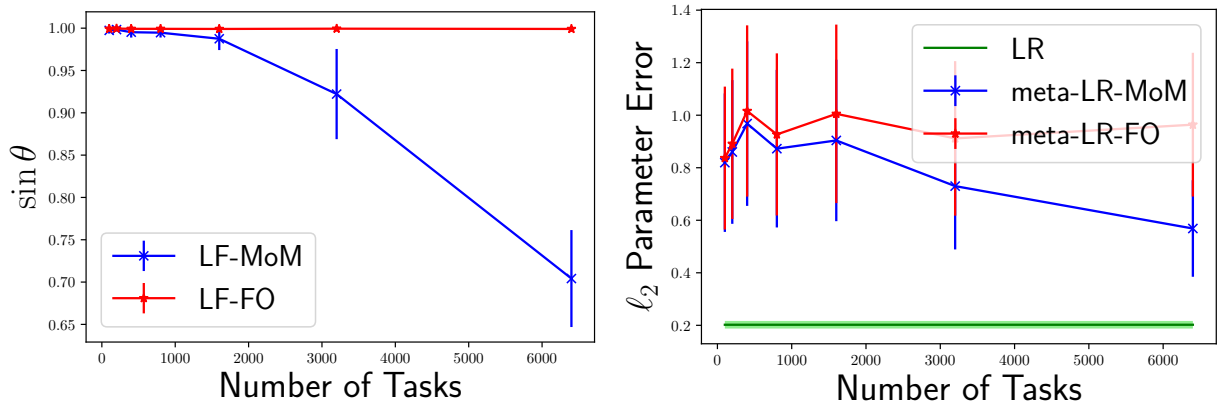

Figure 2.1: Left: LF-FO vs. LF-MoM estimator with error measured in the subspace angle distance $\sin\theta(\hat{\mathbf{B}}, \mathbf{B})$. Right: meta-LR-FO and meta-LR-MoM vs. LR on new task with error measured on new task parameter. Here $d = 100$, $r = 5$, and $n_t = 5$ while $n_2 = 2500$ as the number of tasks is varied.

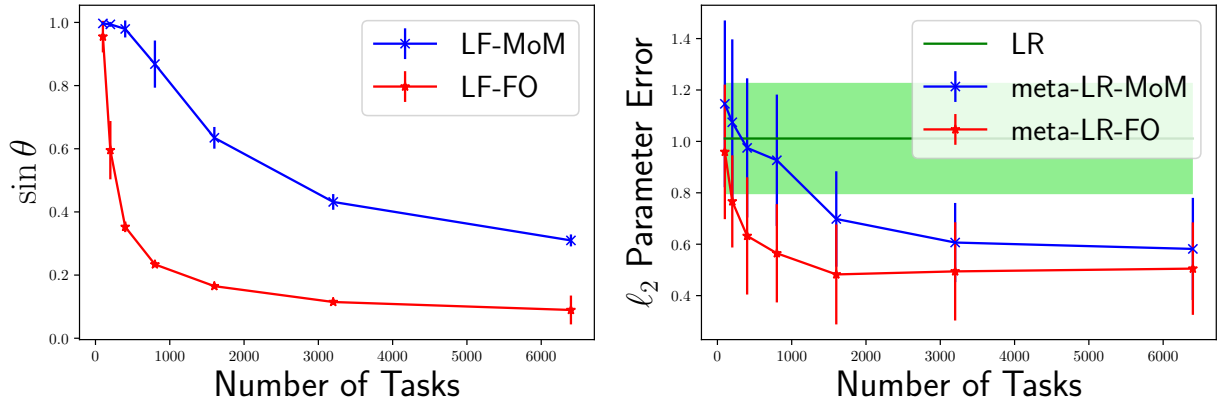

Figure 2.2: Left: LF-FO vs. LF-MoM estimator with error measured in the subspace angle distance $\sin\theta(\hat{\mathbf{B}}, \mathbf{B})$. Right: meta-LR-FO and meta-LR-MoM vs. LR on new task with error measured on new task parameter. Here $d = 100$, $r = 5$, $n_t = 25$ while $n_2 = 25$ while the number of tasks is varied.

can learn a non-trivial estimate of the feature representation. The benefits of transferring this representation are also evident in the improved generalization performance seen by the meta-regression procedures on the new task. Interestingly, the loss-based approach learns an accurate feature representation $\hat{\mathbf{B}}$ with significantly fewer samples then the method-of-moments estimator, in contrast to the previous experiment. Finally, we consider an instance where $d = 100$, $r = 5$, $t = 20$, and $n_2 = 50$ with varying numbers of training points n_t per

task. We see in Fig. 2.3 that meta-learning of representations provides significant value in a new task. Note that these numerical experiments show that as the number of tasks is fixed, but n_t increases, the generalization ability of the meta-learned regressions significantly improves as reflected in the bound (2.2).

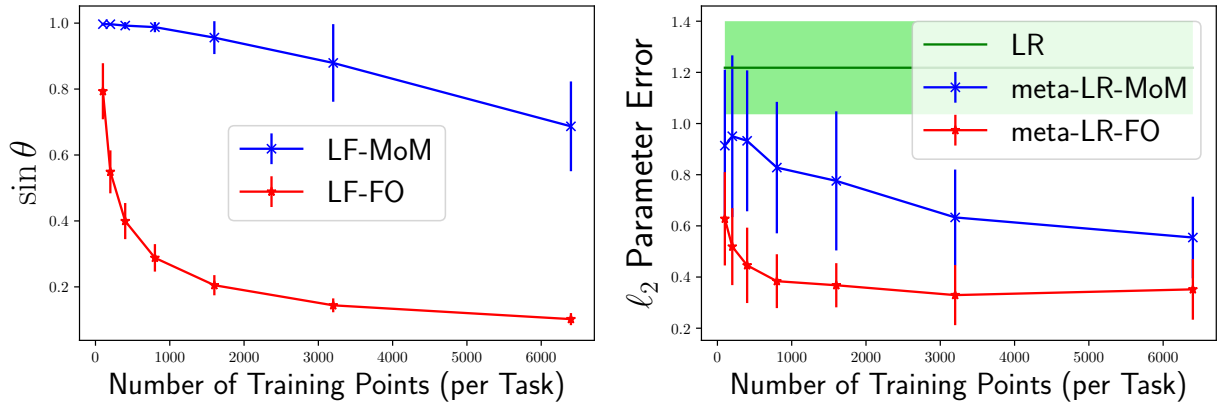

Figure 2.3: Left: LF-FO vs. LF-MoM estimator with error measured in the subspace angle distance $\sin\theta(\hat{\mathbf{B}}, \mathbf{B})$. Right: meta-LR-FO and meta-LR-MoM vs. LR on new task with error measured on new task parameter. Here $d = 100$, $r = 5$, $t = 20$, and $n_2 = 50$ while the number of training points per task (n_t) is varied.

2.7 Conclusions

In this paper we show how a shared linear representation may be efficiently learned and transferred between multiple linear regression tasks. We provide both upper and lower bounds on the sample complexity of learning this representation and for the problem of learning-to-learn. We believe our bounds capture important qualitative phenomena observed in real meta-learning applications absent from previous theoretical treatments.

Appendix

Notation and Set-up We first establish several useful pieces of notation used throughout the Appendices. We will say that a mean-zero random variable x is sub-gaussian, $x \sim \mathrm{sG}(\kappa)$, if $\mathbb{E}[\exp(\lambda x))] \leqslant \exp(\frac{\kappa^2\lambda^2}{2})$ for all λ. We will say that a mean-zero random variable x is sub-exponential, $x \sim \mathrm{sE}(\nu, \alpha)$, if $\mathbb{E}[\exp(\lambda x)] \leqslant \exp(\frac{\nu^2\lambda^2}{2})$ for all $|\lambda| \leqslant \frac{1}{\alpha}$. We will say that a mean-zero random vector is sub-gaussian, $\mathbf{x} \sim \mathrm{sG}(\kappa)$, if $\forall \mathbf{v} \in \mathbb{R}^p$, $\mathbb{E}[\exp(\mathbf{v}^\top \mathbf{x})] \leqslant \exp(\frac{\kappa^2\|\mathbf{v}\|_2^2}{2})$. A standard Chernoff argument shows that if $x \sim \mathrm{sE}(\nu, \alpha)$ then $\mathbf{Pr}[|x| \geqslant t] \leqslant 2\exp(-\frac{1}{2}\min(\frac{t^2}{\nu^2}, \frac{t}{\alpha}))$. Throughout we will use c, C to refer to universal constants that may change from line to line.

2.8 Proofs for Section 2.1

Here we provide a formal statement of Theorem 2.1.

Theorem 2.6 (Formal statement of Theorem 2.1)**.** *Suppose we are first given n_1 total samples from (2.1) which satisfy Assumption 2.1 and $\mathbf{x}_i \sim \mathcal{N}(0, \mathbf{I}_d)$, with an equal number of samples from each task, which collectively satisfy Assumption 2.2. Then, we are presented n_2 samples also from (2.1), satisfying Assumption 2.1, but from a $t+1$st task which satisfies $\|\boldsymbol{\alpha}_{t+1}\|^2 \leqslant O(1)$. If the n_1 samples are used in Algorithm 1 to learn a feature representation $\hat{\mathbf{B}}$, which is used in Algorithm 2 along with the n_2 samples to learn $\hat{\boldsymbol{\alpha}}$, and $n_1 \gtrsim \mathrm{polylog}(d, n_1)\frac{\bar{\kappa}dr}{\nu}$, $n_2 \gtrsim r \log n_2$, the excess prediction error on a new datapoint drawn from the covariate distribution, is,*

$$\mathbb{E}_{\mathbf{x}_\star}[\langle \mathbf{x}_\star, \hat{\mathbf{B}}\hat{\boldsymbol{\alpha}} - \mathbf{B}\boldsymbol{\alpha}_{t+1}\rangle^2] \leqslant \tilde{O}\left(\frac{\bar{\kappa}dr}{\nu n_1} + \frac{r}{n_2}\right),$$

with probability at least $1 - O(n_1^{-100} + n_2^{-100})$.

Proof. Note that $\mathbb{E}_{\mathbf{x}_\star}[\langle \mathbf{x}_\star, \hat{\mathbf{B}}\hat{\boldsymbol{\alpha}} - \mathbf{B}\boldsymbol{\alpha}_{t+1}\rangle^2] = \|\hat{\mathbf{B}}\hat{\boldsymbol{\alpha}} - \mathbf{B}\boldsymbol{\alpha}_{t+1}\|^2$. Combining Theorem 2.3, Theorem 2.4 and applying a union bound then gives the result. □

Note that in order to achieve the formulation in Theorem 2.1, we make the simplifying assumption that the training tasks are well-conditioned in the sense that $\bar{\kappa} \leqslant \kappa \leqslant O(1)$ and $\nu \geqslant \Omega(\frac{1}{r})$—which is consistent with the normalization in Assumption 2.2. Such a setting is for example achieved (w.h.p.) if each $\boldsymbol{\alpha}_t \sim \mathcal{N}(0, \frac{1}{\sqrt{r}}\boldsymbol{\Sigma})$ where $\sigma_1(\boldsymbol{\Sigma})/\sigma_r(\boldsymbol{\Sigma}) \leqslant O(1)$.

2.9 Proofs for Section 2.3

Analyzing the performance of the method-of-moments estimator requires two steps. First, we show that the estimator $(1/n) \cdot \sum_{i=1}^n y_i^2 \mathbf{x}_i \mathbf{x}_i^\top$ converges to its mean in spectral norm,

up to error fluctuations $\tilde{O}(\sqrt{\frac{dr}{n}})$. Showing this requires adapting tools from the theory of matrix concentration. Second, a standard application of the Davis-Kahan $\sin\theta$ theorem shows that top-r PCA applied to this noisy matrix, $(1/n)\cdot\sum_{i=1}^n y_i^2\mathbf{x}_i\mathbf{x}_i^\top$, can extract a subspace $\hat{\mathbf{B}}$ close to the true column space of $\mathbf{B}$ up to a small error. Throughout this section we let $\bar{\mathbf{\Gamma}} = \frac{1}{n}\sum_{i=1}^n \bar{\mathbf{\Gamma}}_i$ with $\bar{\mathbf{\Gamma}}_i = \mathbf{B}\boldsymbol{\alpha}_{t(i)}\boldsymbol{\alpha}_{t(i)}^\top\mathbf{B}^\top$. We also let $\bar{\mathbf{\Lambda}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\alpha}_{t(i)}\boldsymbol{\alpha}_{t(i)}^\top$ be the empirically observed task matrix. Note that under Assumption 2.2, we have that $\bar{\mathbf{\Gamma}}$ and $\bar{\mathbf{\Lambda}}$ behave identically since $\mathbf{B}$ has orthonormal columns we have that $\mathrm{tr}(\bar{\mathbf{\Gamma}}) = \mathrm{tr}(\bar{\mathbf{\Lambda}})$ and $\sigma_r(\bar{\mathbf{\Gamma}}) = \sigma_r(\bar{\mathbf{\Lambda}})$. Furthermore throughout this section we use $\tilde{\kappa}$ and $\tilde{\nu}$ to refer to the average condition number and r-th singular value of the empirically observed task matrix $\bar{\mathbf{\Lambda}}$ – since all our results hold in generality for this matrix. Note that under the uniform task observation model the task parameters of $\bar{\mathbf{\Lambda}}$ and the population task matrix $\frac{\mathbf{A}^\top\mathbf{A}}{t}$ are equal.

We first present our main theorem which shows our method-of-moments estimator can recover the true subspace $\mathbf{B}$ up to small error.

Proof of Theorem 2.3. The proof follows by combining the Davis-Kahan $\sin\theta$ theorem with our main concentration result for the matrix $\frac{1}{n}\sum_{i=1}^n y_i^2\mathbf{x}_i^\top\mathbf{x}_i - \mathbb{E}[\frac{1}{n}\sum_{i=1}^n y_i^2\mathbf{x}_i^\top\mathbf{x}_i]$. First note that $\mathbb{E}[\frac{1}{n}\sum_{i=1}^n y_i^2\mathbf{x}_i^\top\mathbf{x}_i] = (2\bar{\mathbf{\Gamma}} + (1+\mathrm{tr}(\bar{\mathbf{\Gamma}}))\mathbf{I}_d$ by Lemma 2.1 and define $\frac{1}{n}\sum_{i=1}^n y_i^2\mathbf{x}_i^\top\mathbf{x}_i - \mathbb{E}[\frac{1}{n}\sum_{i=1}^n y_i^2\mathbf{x}_i^\top\mathbf{x}_i] = \mathbf{E}$. Note that under the conditions of the result for $n \geqslant cd$ we have that $\|\mathbf{E}\| \leqslant \tilde{O}(\sqrt{\frac{d\tilde{\kappa}r\tilde{\nu}}{n}})$ by Theorem 2.7 for large-enough c due to the SNR normalization; so again by taking sufficiently large c such that $n \geqslant c\cdot\mathrm{polylog}(d,n)\tilde{\kappa}rd/\tilde{\nu}$ we can ensure that $\|\mathbf{E}\| \leqslant \delta \leqslant 2\tilde{\nu}$ for as small δ as we choose with the requisite probability. Since $\|\mathbf{E}\| \leqslant \delta$ we have that $\sigma_{r+1}(y_i^2\mathbf{x}_i\mathbf{x}_i^\top) - \sigma_{r+1}((2\bar{\mathbf{\Gamma}} + (1+\mathrm{tr}(\bar{\mathbf{\Gamma}}))\mathbf{I}_d) \leqslant \delta$ and since $\bar{\mathbf{\Gamma}}$ is rank r, $\sigma_r((2\bar{\mathbf{\Gamma}}+(1+\mathrm{tr}(\bar{\mathbf{\Gamma}}))\mathbf{I}_d) - \sigma_{r+1}((2\bar{\mathbf{\Gamma}}+(1+\mathrm{tr}(\bar{\mathbf{\Gamma}}))\mathbf{I}_d) = 2\sigma_r(\bar{\mathbf{\Gamma}})$. Hence, applying the Davis-Kahan $\sin\theta$ theorem shows that,

$$\|\hat{\mathbf{B}}_\perp^\top\mathbf{B}\| \leqslant \frac{\|\hat{\mathbf{B}}_\perp^\top\mathbf{E}\mathbf{B}\|}{2\sigma_r(\bar{\mathbf{\Lambda}}) - \delta} \leqslant \frac{\|\mathbf{E}\|}{2\sigma_r(\bar{\mathbf{\Lambda}}) - \delta} \leqslant \frac{\|\mathbf{E}\|}{\tilde{\nu}} \leqslant \tilde{O}\left(\sqrt{\frac{1}{\tilde{\nu}}\frac{d\tilde{\kappa}r}{n}}\right),$$

where the final inequalities follows by taking c large enough to ensure $\delta \leqslant \tilde{\nu}$ and Theorem 2.7. □

We now present our main result which proves the concentration of the estimator,

Theorem 2.7. *Suppose the n data samples $(\mathbf{x}_i, y_i)$ are generated from the model in (2.11) and that Assumptions 2.1 and 2.2 hold with $\mathbf{x}_i \sim \mathcal{N}(0,1)$ i.i.d. Then if $n \gtrsim c$ for sufficiently large c,*

$$\|\frac{1}{n}\sum_{i=1}^n y_i^2\mathbf{x}_i\mathbf{x}_i^\top - (2\bar{\mathbf{\Gamma}} + (1+\mathrm{tr}(\bar{\mathbf{\Gamma}}))\mathbf{I}_d))\| \leqslant$$
$$\log^3 n\cdot\log^3 d\cdot O\left(\sqrt{\frac{d\tilde{\kappa}r\tilde{\nu}}{n}} + \frac{d}{n}\right),$$

with probability at least $1 - O(n^{-100})$.

Proof. Note that the mean of $\frac{1}{n}\sum_{i=1}^n y_i^2 \mathbf{x}_i\mathbf{x}_i^\top$ is $2\bar{\mathbf{\Gamma}} + (1+\mathrm{tr}(\bar{\mathbf{\Gamma}}))\mathbf{I}_d)$ by Lemma 2.1. Then using the fact that $y_i = \mathbf{x}_i^\top \mathbf{B}\boldsymbol{\alpha}_{t(i)} + \epsilon_i$, we can write down the error decomposition for the estimator into signal and noise terms,

$$\frac{1}{n}\sum_{i=1}^n y_i^2\mathbf{x}_i\mathbf{x}_i^\top - (2\bar{\mathbf{\Gamma}} + (1+\mathrm{tr}(\bar{\mathbf{\Gamma}}))\mathbf{I}_d = \frac{1}{n}\sum_{i=1}^n (\mathbf{x}_i\mathbf{B}\boldsymbol{\alpha}_{t(i)})^2\mathbf{x}_i\mathbf{x}_i^\top - (2\bar{\mathbf{\Gamma}} + \mathrm{tr}(\bar{\mathbf{\Gamma}}))\mathbf{I}_d +$$
$$\frac{1}{n}\sum_{i=1}^n 2\epsilon_i\mathbf{x}_i\mathbf{B}\boldsymbol{\alpha}_{t(i)}\mathbf{x}_i\mathbf{x}_i^\top + \frac{1}{n}\sum_{i=1}^n \epsilon_i^2\mathbf{x}_i\mathbf{x}_i^\top - \mathbf{I}_d.$$

We proceed to control the fluctuations of each term in spectral norm individually using tools from matrix concentration. Applying Lemma 2.2, Lemma 2.3, Lemma 2.4, the triangle inequality and a union bound shows the desired quantity is upper bounded as,

$$\log^3 n \cdot \log^3 d \cdot$$
$$O\left(\sqrt{\frac{d\max(1, \mathrm{tr}(\bar{\mathbf{\Gamma}}), \mathrm{tr}(\bar{\mathbf{\Gamma}})\max_i \|\boldsymbol{\beta}_i\|^2)}{n}} + \frac{d\max(1, \max_i\|\boldsymbol{\beta}_i\|, \max_i\|\boldsymbol{\beta}_i\|^2)}{n}\right).$$

Finally, using Assumption 2.2 and the fact that $\mathrm{tr}\bar{\mathbf{\Gamma}} = \mathrm{tr}\bar{\mathbf{\Lambda}}$ and the fact $\|\boldsymbol{\beta}_i\| = \|\boldsymbol{\alpha}_i\|$, simplifies the result to the theorem statement. Note that since $\|\boldsymbol{\alpha}_i\| = \Theta(1)$ for all i we have that, $\mathrm{tr}(\bar{\mathbf{\Gamma}}) = \Theta(1)$ so the SNR normalization guarantees the leading noise term satisfies $1 \leqslant O(\mathrm{tr}(\bar{\mathbf{\Gamma}}))$. □

We begin by computing the mean of the estimator.

Lemma 2.1. *Suppose the n data samples $(\mathbf{x}_i, y_i)$ are generated from the model in (2.1) and that Assumptions 2.1 and 2.2 hold. Then,*

$$\mathbb{E}[\frac{1}{n}\sum_{i=1}^n y_i^2\mathbf{x}_i\mathbf{x}_i^\top] = 2\bar{\mathbf{\Gamma}} + (1+\mathrm{tr}(\bar{\mathbf{\Gamma}}))\mathbf{I}_d$$

where $\bar{\mathbf{\Gamma}} = \frac{1}{n}\sum_{i=1}^n \bar{\mathbf{\Gamma}}_i$ with $\mathbf{\Gamma}_i = \mathbf{B}\boldsymbol{\alpha}_{t(i)}\boldsymbol{\alpha}_{t(i)}^\top\mathbf{B}^\top$.

Proof. Since ϵ_i is mean-zero, using the definition of y_i we immediately obtain,

$$\mathbb{E}[\frac{1}{n}\sum_{i=1}^n y_i^2\mathbf{x}_i\mathbf{x}_i^\top] = \mathbf{I}_d + \mathbb{E}[\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i^\top\mathbf{\Gamma}_i\mathbf{x}_i\mathbf{x}_i\mathbf{x}_i^\top] = \mathbf{I}_d + \mathbb{E}[\mathbf{x}^\top\bar{\mathbf{\Gamma}}\mathbf{x}\mathbf{x}\mathbf{x}^\top],$$

for $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_d)$. Using the eigendecomposition of $\bar{\mathbf{\Gamma}}$ we have that $\mathbb{E}[\mathbf{x}^\top\bar{\mathbf{\Lambda}}\mathbf{x}\mathbf{x}\mathbf{x}^\top] = \sum_{i=1}^r \sigma_i\mathbb{E}[(\mathbf{x}^\top\mathbf{v}_i)^2\mathbf{x}\mathbf{x}^\top]$. Due to the isotropy of the Gaussian distribution, it suffices to compute $\mathbb{E}[(\mathbf{x}^\top\mathbf{e}_1)^2\mathbf{x}\mathbf{x}^\top]$ and rotate the result back to $\mathbf{v}_i$. In particular we have that,

$$(\mathbb{E}[(\mathbf{x}^\top\mathbf{e}_1)^2\mathbf{x}\mathbf{x}^\top])_{ij} = \begin{cases} 0 & i \neq j \\ 1 & i = j \neq 1 \\ 3 & i = j = 1 \end{cases} \implies \mathbb{E}[(\mathbf{x}^\top\mathbf{e}_1)^2\mathbf{x}\mathbf{x}^\top] = 2\mathbf{e}_1\mathbf{e}_1 + \mathbf{I}_d \implies$$

$$\implies \mathbb{E}[(\mathbf{x}^\top \mathbf{v}_i)^2 \mathbf{x}\mathbf{x}^\top] = 2\mathbf{v}_i\mathbf{v}_i + \mathbf{I}_d \implies \mathbf{E}[\mathbf{x}^\top \bar{\mathbf{\Gamma}} \mathbf{x}\mathbf{x}\mathbf{x}^\top] = 2\bar{\mathbf{\Gamma}} + \mathrm{tr}(\bar{\mathbf{\Gamma}})\mathbf{I}_d,$$

from which the conclusion follows. □

We start by controlling the fluctuations of the final noise term (which has identity mean).

Lemma 2.2. *Suppose the n data samples $(\mathbf{x}_i, y_i)$ are generated from the model in (2.1) and that Assumptions 2.1 and 2.2 hold. Then for $n \geqslant c$ for sufficiently large c,*

$$\|\frac{1}{n}\sum_{i=1}^n \epsilon_i^2 \mathbf{x}_i\mathbf{x}_i^\top - \mathbf{I}_d\| \leqslant O\left(\log^2 n \left(\sqrt{\frac{d}{n}} + \frac{d}{n}\right)\right).$$

with probability at least $1 - O(n^{-100})$.

Proof. We first decompose the expression as,

$$\|\frac{1}{n}\sum_{i=1}^n \epsilon_i^2 \mathbf{x}_i\mathbf{x}_i^\top - \mathbf{I}_d\| \leqslant \|\frac{1}{n}\sum_{i=1}^n \epsilon_i^2 \mathbf{x}_i\mathbf{x}_i^\top - \frac{1}{n}\sum_{i=1}^n \epsilon_i^2 \mathbf{I}_d\| + \|\frac{1}{n}\sum_{i=1}^n \epsilon_i^2 \mathbf{I}_d - \mathbf{I}_d\|$$

We begin by controlling the second term. By a sub-exponential tail bound we have that $\mathbf{Pr}[|\frac{1}{n}\sum_{i=1}^n \epsilon_i^2 - 1| \geqslant t] \leqslant 2\exp(-Cn\min(t^2/8^2, t/8))$, since ϵ_i^2 is $\mathrm{sE}(8,8)$ by Lemma 2.25. Letting $t = c\sqrt{\log(\frac{1}{\delta})/n}$ for sufficiently large c, and assuming $n \gtrsim \log(\frac{1}{\delta})$, implies $|\frac{1}{n}\sum_{i=1}^n \epsilon_i^2 - 1| \leqslant O(\sqrt{\frac{\log(1/\delta)}{n}})$ with probability at least $1 - 2\delta$. Hence $\|\frac{1}{n}\sum_{i=1}^n \epsilon_i^2 \mathbf{I}_d - \mathbf{I}_d\| \leqslant O(\sqrt{\frac{\log(1/\delta)}{n}})$ on this event.

Now we apply Lemma 2.26 with $a_i = \epsilon_i$ to control the first term. Using the properties of sub-Gaussian maxima we can conclude that $\mathbf{Pr}[\max_i |\epsilon|_i \geqslant t] \leqslant 2n\exp(-t^2/2)$; taking $t = 4\sqrt{\log n} + c\sqrt{\log(1/\delta)}$ for sufficiently large c implies that $\max_i |\epsilon_i| \leqslant O(\sqrt{\log n}) + O(\sqrt{\log(1/\delta)})$ with probability at least $1 - \delta$. In the setting of Lemma 2.26, conditionally on ϵ_i, $K = \max_i |\epsilon_i|$ and $\mathbf{\Sigma} = \mathbf{I}_d$ so taking $t = c\sqrt{\log(1/\delta)}$ for sufficiently large c implies that $\|\frac{1}{n}\sum_{i=1}^n \epsilon_i^2 \mathbf{x}_i\mathbf{x}_i^\top - \frac{1}{n}\sum_{i=1}^n \epsilon_i^2 \mathbf{I}_d\| \leqslant K \cdot O(\sqrt{d/n} + \sqrt{\log(1/\delta)/n} + \frac{d}{n} + \frac{\log(1/\delta)}{n})$ with probability at least $1 - 2\delta$ conditionally on ϵ_i. Conditioning on the event that $\max_i |\epsilon_i| \leqslant O(\sqrt{\log n}) + O(\sqrt{\log(1/\delta)})$ to conclude the argument finally shows that,

$$\|\frac{1}{n}\sum_{i=1}^n \epsilon_i^2 \mathbf{x}_i\mathbf{x}_i^\top - \mathbf{I}_d\| \leqslant$$
$$O(\log n + \log(1/\delta)) \cdot O(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} + \frac{d}{n} + \frac{\log(1/\delta)}{n}) + O(\sqrt{\frac{\log(1/\delta)}{n}}),$$

with probability at least $1 - 5\delta$. Selecting $\delta = n^{-100}$ implies that $\|\frac{1}{n}\sum_{i=1}^n \epsilon_i^2 \mathbf{x}_i\mathbf{x}_i^\top - \mathbf{I}_d\| \leqslant O\left(\log^2 n \left(\sqrt{\frac{d}{n}} + \frac{d}{n}\right)\right)$, with probability at least $1 - O(n^{-100})$. □

We now proceed to controlling the fluctuations of the second noise term (which is mean-zero). Our main technical tool is Lemma 2.30.

Lemma 2.3. *Suppose the n data samples $(\mathbf{x}_i, y_i)$ are generated from the model in (2.1) and that Assumptions 2.1 and 2.2 hold. Then,*

$$\|\sum_{i=1}^{n} 2\epsilon_i \mathbf{x}_i^\top \mathbf{B}\boldsymbol{\alpha}_{t(i)}\mathbf{x}_i\mathbf{x}_i^\top\| \leqslant O(\log n + \log d)\left(\sqrt{\frac{d\mathrm{tr}(\bar{\boldsymbol{\Gamma}})}{n}} + \frac{d\max_i \|\boldsymbol{\beta}_i\|(\log^2(n)+\log^2(d))}{n}\right).$$

with probability at least $1 - O((nd)^{-100})$.

Proof. To apply the truncated version of the matrix Bernstein inequality (in the form of Lemma 2.30) we need to set an appropriate truncation level R, for which need control on the norms of $Z_i = \|2\epsilon_i \mathbf{x}_i^\top \mathbf{B}\boldsymbol{\alpha}_{t(i)}\mathbf{x}_i\mathbf{x}_i^\top\| = 2|\epsilon_i|\|\mathbf{x}_i\|^2|\mathbf{x}_i^\top \mathbf{B}\boldsymbol{\alpha}_{t(i)}|$. Using sub-gaussian, and sub-exponential tail bounds we have that $|\epsilon_i| \leqslant O(1+\sqrt{\log(1/\delta)})$, $\|\mathbf{x}_i\|^2 \leqslant O(d + \max(\sqrt{d\log(1/\delta)}, \log(1/\delta))) = O(d+\sqrt{d}\log(1/\delta))$, and $|\mathbf{x}_i^\top \mathbf{B}\boldsymbol{\alpha}_{t(i)}| \leqslant O(\|\boldsymbol{\beta}_i\|(1+\sqrt{\log(1/\delta)})$ each with probability at least $1-\delta$. Accordingly with probability at least $1-3\delta$ we have that $\|Z_i\| \leqslant O(\|\boldsymbol{\beta}_i\| d(1+\log^2(1/\delta)))$. We can rearrange this statement to conclude that $\mathbf{Pr}[\|Z_i\| \geqslant c_1\|\boldsymbol{\beta}_i\|d + t] \leqslant 3\exp(-c_2(\frac{t}{\|\boldsymbol{\beta}_i\|d})^{1/2})$ for some c_1, c_2. Define a truncation level $R = c_1 \max_i \|\boldsymbol{\beta}_i\|d + K\max_i\|\boldsymbol{\beta}_i\|d$ for some K to be chosen later. We can also use the aforementioned tail bound to control $\|\mathbb{E}[Z_i] - \mathbb{E}[Z_i']\| \leqslant \mathbb{E}[Z_i \mathbb{1}[\|Z_i\| \geqslant \alpha]] \leqslant \int_{K\|\boldsymbol{\beta}_i\|d}^{\infty} 3\exp(-c_2(\frac{t}{\|\boldsymbol{\beta}_i\|d})^{1/2}) \leqslant O((1+\sqrt{K})\exp(-c\sqrt{K})\max_i\|\boldsymbol{\beta}_i\|d) = \Delta$.

Next we must compute an upper bound for the matrix variance term

$$\|\sum_{i=1}^{n}\mathbb{E}[\epsilon_i^2(\boldsymbol{\alpha}_i^\top\mathbf{B}^\top\mathbf{x}_i)^2\|\mathbf{x}_i\|^2\mathbf{x}_i\mathbf{x}_i^\top]\| = \|\sum_{i=1}^{n}\mathbb{E}[(\boldsymbol{\alpha}_i^\top\mathbf{B}^\top\mathbf{x})^2\|\mathbf{x}\|^2\mathbf{x}\mathbf{x}^\top]\|$$
$$= n\|\mathbb{E}[\mathbf{x}^\top\bar{\boldsymbol{\Gamma}}\mathbf{x}\|\mathbf{x}\|^2\mathbf{x}\mathbf{x}^\top]\| = n\|\sum_{i=1}^{r}\sigma_i\mathbb{E}[(\mathbf{v}_i^\top\mathbf{x})^2\|\mathbf{x}\|^2\mathbf{x}\mathbf{x}^\top]\|$$

, taking an expectation over ϵ_i in the first equality, and diagonalizing $\bar{\boldsymbol{\Gamma}}$. As before due to isotropy of the Gaussian it suffices to compute the expectation with $\mathbf{v}_i = \mathbf{e}_1$ and rotate the result back to $\mathbf{v}_i$. Before computing the term we note that for a standard normal gaussian random variable $g \sim \mathcal{N}(0,1)$ we have that $\mathbb{E}[g^6] = 15$, $\mathbb{E}[g^4] = 3$, $\mathbb{E}[g^2] = 1$. Then by simple combinatorics we find that,

$$(\mathbb{E}[x_1^2(\sum_{a=1}^{n}x_a^2)\mathbf{x}\mathbf{x}^\top])_{ij} = \begin{cases} 0 & i \neq j \neq 1 \\ 0 & i = 1 \neq j \\ 2\cdot 3\cdot 1 + (d-2)\cdot 1 & i = j \neq 1 \\ 15 + 3(d-1) & i = j = 1 \end{cases} \implies$$

$$\mathbb{E}[(\mathbf{x}^\top\mathbf{e}_1)^2\|\mathbf{x}\|^2\mathbf{x}\mathbf{x}^\top] = (2d+8)\mathbf{e}_1\mathbf{e}_1^\top + (d+4)\mathbf{I}_d$$

$$\implies \|\sum_{i=1}^{n} \mathbb{E}[\epsilon_i^2(\boldsymbol{\alpha}_i^\top \mathbf{B}^\top \mathbf{x}_i)^2 \|\mathbf{x}_i\|^2 \mathbf{x}_i \mathbf{x}_i^\top]\| \leqslant n(d+4)\|\bar{\boldsymbol{\Gamma}} + \mathrm{tr}(\bar{\boldsymbol{\Gamma}})\mathbf{I}_d\| \leqslant 10nd(\mathrm{tr}(\bar{\boldsymbol{\Gamma}})) = \sigma^2.$$

Finally, we can assemble the previous two computations to conclude the result with appropriate choices of R (parametrized through K) and t by combining with Lemma 2.30. Before beginning recall by definition we have that $\mathrm{tr}(\bar{\boldsymbol{\Gamma}}) = \frac{1}{n}\max_i \|\boldsymbol{\beta}_i\|^2$. Let us choose $\sqrt{K} = \frac{c_3}{c}(\log(n) + \log(d))$ for some sufficiently large c_3. In this case, we can choose c_3 such that $\Delta \leqslant O((\log n + \log d)\frac{\max_i \|\boldsymbol{\beta}_i\| d}{n^{10}d^{10}}) \leqslant O(\frac{\sqrt{\mathrm{tr}(\bar{\boldsymbol{\Gamma}})}}{n^8 d^8})$, since $\sqrt{\mathrm{tr}(\bar{\boldsymbol{\Gamma}})} \geqslant \frac{1}{\sqrt{n}}\max_i \|\boldsymbol{\beta}_i\|$. Similarly, our choice of truncation level becomes $R = O((\log^2(n) + \log^2(d))\max_i \|\boldsymbol{\beta}_i\| d)$. At this point we now choose $t = c_4(\log n + \log d)\max(\sigma/\sqrt{n}, R/n)$ for sufficiently large c_4. For large enough c_4 we can guarantee that $t \geqslant 2\Delta \implies t - \Delta \geqslant \frac{t}{2}$.

Hence combining these results together and applying Lemma 2.30 we can provide the following upper bound on the desired quantity,

$$\mathbf{Pr}[\|\frac{1}{n}\sum_{i=1}^{n} 2\epsilon_i \mathbf{x}_i^\top \mathbf{B}\boldsymbol{\alpha}_{t(i)}\mathbf{x}_i\mathbf{x}_i^\top\| \geqslant t] \leqslant$$
$$O(d\exp(-c \cdot c_4(\log n + \log d)) + O(n\sqrt{K}\exp(-c_3(\log n + \log d)) \leqslant O((nd)^{-100}),$$

by taking c_3 and c_4 sufficiently large, with

$$t = O((\log n + \log d)\left(\sqrt{\frac{d\mathrm{tr}(\bar{\boldsymbol{\Gamma}})}{n}} + \frac{d\max_i \|\boldsymbol{\beta}_i\|(\log^2(n) + \log^2(d))}{n}\right).$$

□

Finally we turn to controlling the fluctuations of the primary signal term around its mean using a similar argument to the previous term.

Lemma 2.4. *Suppose the n data samples $(\mathbf{x}_i, y_i)$ are generated from the model in (2.1) and that Assumptions 2.1 and 2.2 hold. Then*

$$\|\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i\mathbf{B}\boldsymbol{\alpha}_{t(i)})^2\mathbf{x}_i\mathbf{x}_i^\top - (2\bar{\boldsymbol{\Gamma}} + \mathrm{tr}(\bar{\boldsymbol{\Gamma}})\mathbf{I}_d)\| \leqslant$$
$$O((\log n + \log d)\left(\sqrt{\frac{d\mathrm{tr}(\bar{\boldsymbol{\Gamma}})\max_i \|\boldsymbol{\beta}_i\|^2}{n}} + \frac{d\max_i \|\boldsymbol{\beta}_i\|^2(\log^2(n) + \log^2(d))}{n}\right),$$

with probability at least $1 - O((nd)^{-100})$.

Proof. The proof is similar to the proof of Lemma 2.3 and uses Lemma 2.30. We begin by controlling the norms of $Z_i = (\mathbf{x}_i^\top \mathbf{B}\boldsymbol{\alpha}_{t(i)})^2\mathbf{x}_i\mathbf{x}_i^\top$. $\|Z_i\| = \|\mathbf{x}_i\|^2(\mathbf{x}_i^\top \mathbf{B}\boldsymbol{\alpha}_{t(i)})^2$. Using Gaussian and sub-exponential tail bounds we have that, $\|\mathbf{x}_i\|^2 \leqslant O(d + \sqrt{d}\log(1/\delta))$ and $(\mathbf{x}_i^\top \mathbf{B}\boldsymbol{\alpha}_{t(i)})^2 \leqslant$

$O(\|\boldsymbol{\beta}_i\|^2(1+\log(1/\delta)))$ each with probability at least $1-\delta$. Hence with probability at least $1-2\delta$ we find that $\|Z_i\| \leqslant O(d\|\boldsymbol{\beta}_i\|^2(1+\log^2(1/\delta))$.

We can rearrange this statement to conclude that, $\mathbf{Pr}[\|Z_i\| \geqslant c_1\|\boldsymbol{\beta}_i\|^2 d + t] \leqslant 2\exp(-c_2(\frac{t}{\|\boldsymbol{\beta}_i\|^2 d})^{1/2})$ for some c_1, c_2. Define a truncation level $R = c_1 \max_i \|\boldsymbol{\beta}_i\|^2 d + K \max_i \|\boldsymbol{\beta}_i\|^2 d$ for some K to be chosen later. We can use the aforementioned tail bound to control $\|\mathbb{E}[Z_i] - \mathbb{E}[Z_i']\| \leqslant \mathbb{E}[Z_i \mathbb{1}[\|Z_i\| \geqslant \alpha]] \leqslant \int_{K\|\boldsymbol{\beta}_i\|^2 d}^{\infty} 2\exp(-c_2(\frac{t}{\|\boldsymbol{\beta}_i\|^2 d})^{1/2}) \leqslant O((1+\sqrt{K})\exp(-c\sqrt{K})\max_i \|\boldsymbol{\beta}_i\|^2 d) = \Delta$.

Next we must compute an upper bound the matrix variance term

$$\|\sum_{i=1}^n \mathbb{E}[(\boldsymbol{\alpha}_i^\top \mathbf{B}^\top \mathbf{x}_i)^4 \|\mathbf{x}_i\|^2 \mathbf{x}_i \mathbf{x}_i^\top]\| = \|\sum_{i=1}^n \mathbb{E}[(\boldsymbol{\beta}_i^\top \mathbf{x})^4 \|\mathbf{x}\|^2 \mathbf{x}\mathbf{x}^\top]\|.$$

As before due to isotropy of the Gaussian it suffices to compute each expectation assuming $\boldsymbol{\beta}_i \propto \mathbf{e}_1$ and rotate the result back to $\boldsymbol{\beta}_i$. Before computing the term we note that for a standard normal gaussian random variable $g \sim \mathcal{N}(0,1)$ we have that $\mathbb{E}[g^8] = 105$, $\mathbb{E}[g^6] = 15$, $\mathbb{E}[g^4] = 3$, $\mathbb{E}[g^2] = 1$. Then by simple combinatorics we find that,

$$(\mathbb{E}[(\mathbf{x}^\top \mathbf{e}_1)^4 (\sum_{a=1}^n x_a^2)\mathbf{x}\mathbf{x}^\top])_{ij} = \begin{cases} 0 & i \neq j \neq 1 \\ 0 & i = 1 \neq j \\ 15 + 3\cdot 3 + (d-2)\cdot 3 & i = j \neq 1 \\ 105 + 15(d-1) & i = j = 1 \end{cases} \implies$$

$$\mathbb{E}[(\mathbf{x}^\top \boldsymbol{\beta}_i)^4 \|\mathbf{x}\|^2 \mathbf{x}\mathbf{x}^\top] = (2d+75)\mathbf{e}_1\mathbf{e}_1^\top + (3d+15)\mathbf{I}_d$$

$$\implies \|\sum_{i=1}^n \mathbb{E}[(\boldsymbol{\alpha}_i^\top \mathbf{B}^\top \mathbf{x}_i)^4 \|\mathbf{x}_i\|^2 \mathbf{x}_i \mathbf{x}_i^\top]\| \leqslant O(d)\|\sum_i \|\boldsymbol{\beta}_i\|^4 (\boldsymbol{\beta}_i\boldsymbol{\beta}_i^\top + \mathbf{I}_d)\| \leqslant O(d\sum_{i=1}^n \|\boldsymbol{\beta}_i\|_2^4) \leqslant$$

$$O(dn \max_i \|\boldsymbol{\beta}_i\|^2 \mathrm{tr}(\bar{\boldsymbol{\Gamma}})) = \sigma^2.$$

Finally, we can assemble the previous two computations to conclude the result with appropriate choices of R (parametrized through K) and t by combining with Lemma 2.30. Before beginning recall by definition we have that $\mathrm{tr}\bar{\boldsymbol{\Gamma}} \geqslant \frac{1}{n}\max_i \|\boldsymbol{\beta}_i\|^2$. Let us choose $\sqrt{K} = \frac{c_3}{c}(\log(n) + \log(d))$ for some sufficiently large c_3. In this case, we can choose c_3 such that $\Delta \leqslant O((\log n + \log d)\frac{\max_i \|\boldsymbol{\beta}_i\|^2 d}{n^{10}d^{10}}) \leqslant O(\frac{\sqrt{\mathrm{tr}(\bar{\boldsymbol{\Gamma}})}\max_i \|\boldsymbol{\beta}_i\|}{n^7 d^7})$, since $\sqrt{\mathrm{tr}(\bar{\boldsymbol{\Gamma}})} \geqslant \frac{1}{\sqrt{n}}\max_i \|\boldsymbol{\beta}_i\|$. Similarly, our choice of truncation level becomes $R = O((\log^2(n) + \log^2(d))\max_i \|\boldsymbol{\beta}_i\|^2 d)$. At this point we now choose $t = c_4(\log n + \log d)\max(\sigma/\sqrt{n}, R/n)$ for sufficiently large c_4. For large enough c_4 we can guarantee that $t \geqslant 2\Delta \implies t - \Delta \geqslant \frac{t}{2}$.

Hence combining these results together and applying Lemma 2.30 we can provide the following upper bound on the desired quantity:

$$\mathbf{Pr}[\|\frac{1}{n}\sum_{i=1}^n (\mathbf{x}_i \mathbf{B}\boldsymbol{\alpha}_{t(i)})^2 \mathbf{x}_i\mathbf{x}_i^\top - (2\bar{\boldsymbol{\Gamma}} + \mathrm{tr}(\bar{\boldsymbol{\Gamma}})\mathbf{I}_d)\| \geqslant t]$$

$$\leqslant O(d\exp(-c\cdot c_4(\log n+\log d))+O(n\sqrt{K}\exp(-c_3(\log n+\log d))$$
$$\leqslant O((nd)^{-100}),$$

by taking c_3 and c_4 sufficiently large, with

$$t=O((\log n+\log d)\left(\sqrt{\frac{d\mathrm{tr}(\bar{\mathbf{\Gamma}})\max_i\|\boldsymbol{\beta}_i\|^2}{n}}+\frac{d\max_i\|\boldsymbol{\beta}_i\|^2(\log^2(n)+\log^2(d))}{n}\right).$$

. □

2.10 Proofs for Section 2.3

In our landscape analysis we consider a setting with t tasks and we observe a datapoint from each of the t tasks uniformly at random at each iteration. Formally, we define the matrix we are trying to recover as

$$\mathbf{M}_\star=(\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_t)^\top\mathbf{B}^\top\underbrace{=}_{\text{SVD}}\mathbf{X}^\star\mathbf{D}^\star(\mathbf{Y}\star)^\top\in\mathbb{R}^{t\times d},\tag{2.7}$$

with $\mathbf{U}^\star=\mathbf{X}^\star(\mathbf{D}^\star)^{1/2}$, and $(\mathbf{D}^\star)^{1/2}(\mathbf{Y}^\star)^\top=(\mathbf{V}^\star)^\top$, from which we obtain the observations:

$$y_i=\langle\mathbf{e}_{t(i)}\mathbf{x}_i^\top,\mathbf{M}_\star\rangle+\sigma\cdot\epsilon_i,\tag{2.8}$$

where we sample tasks uniformly $t(i)\sim\{1,\ldots,t\}$ and $\mathbf{x}_i$ is a sub-gaussian random vector. Note that $\mathbf{M}^\star$ is a rank-r matrix, $\mathbf{U}^\star\in\mathbb{R}^{t\times r}$, and $\mathbf{V}^\star\in\mathbb{R}^{d\times r}$. In this section, we denote $\tilde{d}=\max\{t,d\}$ and let $\sigma_1^\star,\sigma_r^\star$ be the 1-st and r-th eigenvalue of matrix $\mathbf{M}^\star$. We denote $\kappa^\star=\sigma_1^\star/\sigma_r^\star$ as its condition number. Note that as $\mathbf{B}$ is an orthonormal matrix we have that $\mathbf{M}^\star(\mathbf{M}^\star)^\top=t\cdot\mathbf{A}^\top\mathbf{A}/t$ from which it follows that $(\sigma_1^\star)^2=t\cdot\sigma_1(\mathbf{A}^\top\mathbf{A}/t)\leqslant t\bar{\kappa}\nu\leqslant O(t)$ by the normalization on $\|\boldsymbol{\alpha}_i\|$. Similarly $(\sigma_r^\star)^2=t\sigma_r(\mathbf{A}^\top\mathbf{A}/t)\geqslant t\nu$. So it follows that $\sigma_1^\star\leqslant\sqrt{t\kappa\nu}$, $\sigma_r^\star\geqslant\sqrt{t\nu}$ and $\kappa^\star\leqslant\sqrt{\kappa}$. We use this to simplify the preconditions and the statement of the incoherence ball in the main although we work in full generality throughout the Appendix.

We now present the proof of our main result.

Proof of Theorem 2.2. Under the conditions of the theorem note that by Theorem 2.8 we have that,

$$\|\mathbf{U}\mathbf{V}^\top-\mathbf{M}^\star\|_F\leqslant O\left(\sigma\sqrt{t\frac{\max\{t,d\}r\log n}{n}}\right),$$

for $n\geqslant\mathrm{polylog}(n,d,t)C\mu^2r^4\max\{t,d\}(\kappa^\star)^4$. First recall by Lemma 2.16 the incoherence parameter can in fact be shown to be $\mu=O(\bar{\kappa})$ under our assumptions which gives the precondition on the sample complexity due to the task diversity assumption and normalization. To finally convert this bound to a guarantee on the subspace angle we directly apply Lemma 2.15 once again noting the task diversity assumption. Lastly note that as $\mathbf{B}$ is orthonormal we have that $\sigma_1^\star\leqslant\sqrt{t\kappa\nu}$, $\sigma_r^\star\geqslant\sqrt{t\nu}$ and $\kappa^\star\leqslant\sqrt{\kappa}$ as previously argued and $\sigma=1$ under the conditions of the result. □

Geometric Arguments for Landscape Analysis

Our arguments here are generally applicable to various matrix sensing/completion problems so we define some generic notation:

$$f(\mathbf{U},\mathbf{V}) = \frac{2}{n}\sum_{i=1}^{n}(\langle \mathbf{A}_i, \mathbf{U}\mathbf{V}^\top\rangle - \sqrt{t}y_i)^2 + \frac{1}{2}\|\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\|_{\mathrm{F}}^2, \tag{2.9}$$

where $\mathbf{A}_i = \sqrt{t}\mathbf{e}_{t(i)}\mathbf{x}_i^\top$. We work under the following constraint set for large constant C_0:

$$\mathcal{W} = \{\ (\mathbf{U},\mathbf{V}) \mid \max_{i\in[t]}\|\mathbf{e}_i^\top\mathbf{U}\|^2 \leqslant \frac{C_0\mu r\sigma_1^\star}{t},\quad \|\mathbf{U}\|^2 \leqslant C_0\sigma_1^\star,\quad \|\mathbf{V}\|^2 \leqslant C_0\sigma_1^\star\ \}. \tag{2.10}$$

We renormalize the statistical model for convenience simply for the purposes of the proof throughout Section 2.10 the remainder of as:

$$y_i = \langle \mathbf{A}_i, \mathbf{M}^\star\rangle + n_i, \tag{2.11}$$

where $n_i \sim \sqrt{t}\sigma\cdot\epsilon_i$ and where ϵ_i is a sub-gaussian random vector with parameter 1 (note this is because we have scaled $\mathbf{A}_i$ up by a factor of $\sqrt{t}$). $\mathbf{M}^\star$ is rank r, and we let $\mathbf{X}$ be the left singular vector of $\mathbf{M}^\star$, and assume $\mathbf{X}$ is μ-incoherent;[10] i.e., $\max_i\|\mathbf{e}_i^\top\mathbf{X}\|^2 \leqslant \mu r/t$.

We now reformulate the objective (denoting $\mathbf{M} = \mathbf{U}\mathbf{V}^\top$) as

$$f(\mathbf{U},\mathbf{V}) = 2(\mathbf{M}-\mathbf{M}^\star):\mathbf{H}_0:(\mathbf{M}-\mathbf{M}^\star) + \frac{1}{2}\|\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\|_{\mathrm{F}}^2 + Q(\mathbf{U}), \tag{2.12}$$

where $\mathbf{M}:\mathbf{H}_0:\mathbf{M} = \frac{1}{n}\sum_{i=1}^n\langle\mathbf{A}_i,\mathbf{M}\rangle^2$ and $\mathbf{E}[\mathbf{M}:\mathbf{H}_0:\mathbf{M}] = \|\mathbf{M}\|_{\mathrm{F}}^2$ and Q is a regularization term:

$$Q(\mathbf{U},\mathbf{V}) = \frac{2}{n}\sum_{i=1}^{n}[(\langle\mathbf{M}-\mathbf{M}^\star,\mathbf{A}_i\rangle - n_i)^2 - (\langle\mathbf{M}-\mathbf{M}^\star,\mathbf{A}_i\rangle)^2]. \tag{2.13}$$

In this section, we denote $\tilde{d} = \max\{t,d\}$ and let $\sigma_1^\star,\sigma_r^\star$ be the 1-st and r-th eigenvalue of matrix $\mathbf{M}^\star$. We denote $\kappa^\star = \sigma_1^\star/\sigma_r^\star$ as its condition number.

The high-level idea of the analysis uses ideas from [46]. The overall strategy is to argue that if we are currently not located at local minimum in the landscape we can certify this by inspecting the gradient or Hessian of $f(\mathbf{U},\mathbf{V})$ to exhibit a direction of local improvement Δ to decrease the function value of f. Intuitively this direction brings us close to the true underlying $(\mathbf{U}^\star,\mathbf{V}^\star)$.

We now establish some useful definitions and notation for the following analysis

[10] Note that for our particular problem this is not an additional assumption since by Lemma 2.16 our task assumptions imply this.

Definitions and Notation

Definition 2.1. Suppose $\mathbf{M}^\star$ is the optimal solution with SVD is $\mathbf{X}^\star\mathbf{D}^\star\mathbf{Y}^{\star\top}$. Let $\mathbf{U}^\star = \mathbf{X}^\star(\mathbf{D}^\star)^{\frac{1}{2}}$, $\mathbf{V}^\star = \mathbf{Y}^\star(\mathbf{D}^\star)^{\frac{1}{2}}$. Let $\mathbf{M} = \mathbf{U}\mathbf{V}^\top$ be the current point in the landscape. We reduce the problem of studying an asymmetric matrix objective to the symmetric case using the following notational transformations:

$$\mathbf{W} = \begin{pmatrix}\mathbf{U}\\ \mathbf{V}\end{pmatrix}, \mathbf{W}^\star = \begin{pmatrix}\mathbf{U}^\star\\ \mathbf{V}^\star\end{pmatrix}, \mathbf{N} = \mathbf{W}\mathbf{W}^\top, \mathbf{N}^\star = \mathbf{W}^\star\mathbf{W}^{\star\top} \tag{2.14}$$

We will also transform the Hessian operators to operate on $(t+d)\times r$ matrices. In particular, define the Hessians $\mathbf{H}_1, \mathcal{G}$ such that for all $\mathbf{W}$ we have:

$$\begin{aligned}\mathbf{N}:\mathbf{H}_1:\mathbf{N} &= \mathbf{M}:\mathbf{H}_0:\mathbf{M}\\ \mathbf{N}:\mathcal{G}:\mathbf{N} &= \|\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\|_F^2.\end{aligned}$$

Now, letting $Q(\mathbf{W}) = Q(\mathbf{U}, \mathbf{V})$, we can rewrite the objective function $f(\mathbf{W})$ as

$$\frac{1}{2}\left[(\mathbf{N}-\mathbf{N}^\star):4\mathbf{H}_1:(\mathbf{N}-\mathbf{N}^\star) + \mathbf{N}:\mathcal{G}:\mathbf{N}\right] + Q(\mathbf{W}). \tag{2.15}$$

We now introduce the definition of local alignment of two matrices.

Definition 2.2. Given matrices $\mathbf{W}, \mathbf{W}^\star \in \mathbb{R}^{d\times r}$, define their difference $\Delta = \mathbf{W} - \mathbf{W}^\star\mathbf{R}^\star$, where $\mathbf{R}^\star \in \mathbb{R}^{r\times r}$ is chosen as $\mathbf{R}^\star = \arg\min_{\mathbf{Z}^\top\mathbf{Z}=\mathbf{Z}\mathbf{Z}^\top=\mathbf{I}} \|\mathbf{W} - \mathbf{W}^\star\mathbf{Z}\|_F^2$.

Note that this definition tries to "align" $\mathbf{U}$ and $\mathbf{U}^\star$ before taking their difference, and therefore is invariant under rotations. In particular, this definition has the nice property that as long as $\mathbf{N} = \mathbf{W}\mathbf{W}^\top$ is close to $\mathbf{N}^\star = \mathbf{W}^\star(\mathbf{W}^\star)^\top$ in Frobenius norm, the corresponding Δ between them is also small (see Lemma 2.7).

Proofs for Landscape Analysis

With these definitions in hand we can now proceed to the heart of the landscape analysis. Since $\mathbf{W}^\star$ has rotation invariance, in the following section we always choose $\mathbf{W}^\star$ so that it aligns with the corresponding $\mathbf{W}$ according to Definition 2.2.

We first restate a useful result from [46],

Lemma 2.5 ([46, Lemma 16])**.** *For the objective (2.15), let $\Delta, \mathbf{N}, \mathbf{N}^\star$ be defined as in Definition 2.1, Definition 2.2. Then, for any $\mathbf{W} \in \mathbf{R}^{(t+d)\times r}$, we have*

$$\begin{aligned}\Delta:\nabla^2 f(\mathbf{W}):\Delta \leqslant &\Delta\Delta^\top:\mathbf{H}:\Delta\Delta^\top - 3(\mathbf{N}-\mathbf{N}^\star):\mathbf{H}:(\mathbf{N}-\mathbf{N}^\star)\\ &+ 4\langle\nabla f(\mathbf{W}), \Delta\rangle + [\Delta:\nabla^2 Q(\mathbf{W}):\Delta - 4\langle\nabla Q(\mathbf{W}), \Delta\rangle],\end{aligned} \tag{2.16}$$

where $\mathbf{H} = 4\mathbf{H}_1 + \mathcal{G}$. Further, if $\mathbf{H}_0$ satisfies $\mathbf{M}:\mathbf{H}_0:\mathbf{M} \in (1\pm\delta)\|\mathbf{M}\|_F^2$ for some matrix $\mathbf{M} = \mathbf{U}\mathbf{V}^\top$, let $\mathbf{W}$ and $\mathbf{N}$ be defined as in (2.14), then $\mathbf{N}:\mathbf{H}:\mathbf{N} \in (1\pm 2\delta)\|\mathbf{N}\|_F^2$.

With this result we show a key result which shows that with enough samples all stationary points in the incoherence ball $\mathcal{W}$ that are not close to $\mathbf{W}^\star$ have a direction of negative curvature.

Lemma 2.6. *If Assumption 2.1 holds, then when the number of samples, $n \geqslant C\text{polylog}(d,n,t)\mu^2 r^4 \max\{t,d\}(\kappa^\star)^4$ for a sufficiently large constant C, with probability at least $1 - 1/\text{poly}(d)$, all stationary points $\mathbf{W} \in int(\mathcal{W})$ satisfy:*

$$\Delta\Delta^\top : \mathbf{H} : \Delta\Delta^\top - 3(\mathbf{N} - \mathbf{N}^\star) : \mathbf{H} : (\mathbf{N} - \mathbf{N}^\star) \leqslant -0.1\|\mathbf{N} - \mathbf{N}^\star\|_F^2.$$

Proof. We divide the proof into two cases according to the norm of Δ and use different concentration inequalities in each case. In this proof, we denote $\Delta = (\Delta_{\mathbf{U}}^\top, \Delta_{\mathbf{V}}^\top)^\top$, clearly, we have $\|\Delta_{\mathbf{U}}\|_{\text{F}} \leqslant \|\Delta\|_{\text{F}}$ and $\|\Delta_{\mathbf{V}}\|_{\text{F}} \leqslant \|\Delta\|_{\text{F}}$.
Case 1: $\|\Delta\|_{\text{F}}^2 \leqslant \sigma_r^\star/1000$. In this case, $\|\Delta_{\mathbf{U}}\|_{\text{F}}^2 \leqslant \|\Delta\|_{\text{F}}^2 \leqslant \sigma_r^\star/1000$ and $\|\Delta_{\mathbf{V}}\|_{\text{F}}^2 \leqslant \|\Delta\|_{\text{F}}^2 \leqslant \sigma_r^\star/1000$. By (2.18), we have

$$\Delta\Delta^\top : \mathbf{H} : \Delta\Delta^\top \leqslant \|\Delta\Delta^\top\|_{\text{F}}^2 + 0.004\sigma_r^\star\|\Delta_{\mathbf{V}}\|_{\text{F}}^2 \leqslant 0.005\sigma_r^\star\|\Delta\|_{\text{F}}^2$$

On the other hand, denote $\mathbf{S} = \mathbf{W}^\star\Delta^\top + \Delta(\mathbf{W}^\star)^\top$, by (2.17) and Lemma 2.5, we know:

$$\mathbf{S} : \mathbf{H} : \mathbf{S} \geqslant 0.999\|\mathbf{S}\|_{\text{F}}^2.$$

Since we choose $\mathbf{W}^\star$ to align with the corresponding $\mathbf{W}$ according to Definition 2.2, by Lemma 2.7.

$$\|\mathbf{S}\|_{\text{F}}^2 = 2(\|\Delta^\top\mathbf{W}^\star\|_{\text{F}}^2 + \|\Delta(\mathbf{W}^\star)^\top\|_{\text{F}}^2) \geqslant 2\|\Delta(\mathbf{W}^\star)^\top\|_{\text{F}}^2 \geqslant 2\sigma_r^\star\|\Delta\|_{\text{F}}^2.$$

This gives:

$$\begin{aligned}
&\Delta\Delta^\top : \mathbf{H} : \Delta\Delta^\top - 3(\mathbf{N} - \mathbf{N}^\star) : \mathbf{H} : (\mathbf{N} - \mathbf{N}^\star) \\
&= \Delta\Delta^\top : \mathbf{H} : \Delta\Delta^\top - 3(\mathbf{S} + \Delta\Delta^\top) : \mathbf{H} : (\mathbf{S} + \Delta\Delta^\top) \\
&\leqslant -6\mathbf{S} : \mathbf{H} : \Delta\Delta^\top - 3\mathbf{S} : \mathbf{H} : \mathbf{S} \\
&\leqslant -\mathbf{S} : \mathbf{H} : \mathbf{S} - 2\sqrt{\mathbf{S} : \mathbf{H} : \mathbf{S}}(\sqrt{\mathbf{S} : \mathbf{H} : \mathbf{S}} - 3\sqrt{\Delta\Delta^\top : \mathbf{H} : \Delta\Delta^\top}) \\
&\leqslant -0.999\|\mathbf{S}\|_{\text{F}}^2 - 2\sqrt{\mathbf{S} : \mathbf{H} : \mathbf{S}} \cdot \sqrt{\sigma_r^\star} \cdot (\|\Delta\|_{\text{F}} - 0.3\|\Delta\|_{\text{F}}) \leqslant -0.999\|\mathbf{S}\|_{\text{F}}^2.
\end{aligned}$$

Finally, we know $\mathbf{N} - \mathbf{N}^\star = \mathbf{S} + \Delta\Delta^\top$, and $\|\mathbf{S}\|_{\text{F}}^2 \geqslant 2\sigma_r^\star\|\Delta\|_{\text{F}}^2 \geqslant 500\|\Delta\|_{\text{F}}^4 = 500\|\Delta\Delta^\top\|_{\text{F}}^2$. Therefore:

$$\|\mathbf{N} - \mathbf{N}^\star\|_{\text{F}} \leqslant \|\mathbf{S}\|_{\text{F}} + \|\Delta\Delta^\top\|_{\text{F}} \leqslant 2\|\mathbf{S}\|_{\text{F}}.$$

This gives:

$$\Delta\Delta^\top : \mathbf{H} : \Delta\Delta^\top - 3(\mathbf{N} - \mathbf{N}^\star) : \mathbf{H} : (\mathbf{N} - \mathbf{N}^\star) \leqslant -0.999\|\mathbf{S}\|_{\text{F}}^2 \leqslant -0.1\|\mathbf{N} - \mathbf{N}^\star\|_{\text{F}}^2.$$

Case 2: $\|\Delta\|_{\mathrm{F}}^2 \geqslant \sigma_r^\star/1000$, by (2.19), we have:

$$\frac{1}{n}\sum_{i=1}^{n}\langle \mathbf{A}_i, \mathbf{M}-\mathbf{M}^\star\rangle^2 \geqslant \|\mathbf{M}-\mathbf{M}^\star\|_{\mathrm{F}}^2 - (\sigma_r^\star)^2/10^6 \geqslant \|\mathbf{M}-\mathbf{M}^\star\|_{\mathrm{F}}^2 - 0.001\sigma_r^\star\|\Delta\|_{\mathrm{F}}^2.$$

This implies:

$$(\mathbf{N}-\mathbf{N}^\star):\mathbf{H}:(\mathbf{N}-\mathbf{N}^\star) \geqslant \|\mathbf{N}-\mathbf{N}^\star\|_{\mathrm{F}}^2 - 0.004\sigma_r^\star\|\Delta\|_{\mathrm{F}}^2.$$

Then by (2.18), we have:

$$\begin{aligned}
&\Delta\Delta^\top:\mathbf{H}:\Delta\Delta^\top - 3(\mathbf{N}-\mathbf{N}^\star):\mathbf{H}:(\mathbf{N}-\mathbf{N}^\star)\\
\leqslant&\|\Delta\Delta^\top\|_{\mathrm{F}}^2 + 0.004\sigma_r^\star\|\Delta\|_{\mathrm{F}}^2 - 3(\|\mathbf{N}-\mathbf{N}^\star\|_{\mathrm{F}}^2 - 0.004\sigma_r^\star\|\Delta\|_{\mathrm{F}}^2)\\
\leqslant&-\|\mathbf{N}-\mathbf{N}^\star\|_{\mathrm{F}}^2 + 0.016\sigma_r^\star\|\Delta\|_{\mathrm{F}}^2 \leqslant -0.1\|\mathbf{N}-\mathbf{N}^\star\|_{\mathrm{F}}^2,
\end{aligned}$$

where the last step follows by applying Lemma 2.7. This finishes the proof. □

With this key structural lemma in hand, we now present the main technical result for the section which characterizes the effect of the additive noise n_i on the landscape.

Theorem 2.8. *If Assumption 2.1 holds, when the number of samples, $n \geqslant C\mathrm{polylog}(n,d,t)\mu^2 r^4 \max\{t,d\}(\kappa^\star)^4$ for sufficiently large constant C, with probability at least $1-1/\mathrm{poly}(d)$, we have that any local minimum $(\mathbf{U},\mathbf{V}) \in int(\mathcal{W})$ of the objective (2.9) satisfies:*

$$\|\mathbf{U}\mathbf{V}^\top - \mathbf{M}^\star\|_F \leqslant O\left(\sigma\sqrt{\frac{t\max\{t,d\}r\log n}{n}}\right).$$

Proof. By Lemma 2.6, we know

$$\Delta\Delta^\top:\mathbf{H}:\Delta\Delta^\top - 3(\mathbf{N}-\mathbf{N}^\star):\mathbf{H}:(\mathbf{N}-\mathbf{N}^\star) \leqslant -0.1\|\mathbf{N}-\mathbf{N}^\star\|_F^2.$$

In order to use Lemma 2.5, we bound the contribution from the noise term Q. Recall (2.13):

$$\begin{aligned}
Q(\mathbf{W}) &= -\frac{4}{n}\sum_{i=1}^{n}(\langle \mathbf{M}-\mathbf{M}^\star, \mathbf{A}_i\rangle n_i) + \frac{2}{n}\sum_{i=1}^{n}(n_i)^2\\
\langle \nabla Q(\mathbf{W}), \Delta\rangle &= -\frac{4}{n}\sum_{i=1}^{n}(\langle \mathbf{U}\Delta_{\mathbf{V}}^\top + \Delta_{\mathbf{U}}\mathbf{V}^\top, \mathbf{A}_i\rangle n_i)\\
\Delta:\nabla^2 Q(\mathbf{W}):\Delta &= -\frac{8}{n}\sum_{i=1}^{n}(\langle \Delta_{\mathbf{U}}\Delta_{\mathbf{V}}^\top, \mathbf{A}_i\rangle n_i).
\end{aligned}$$

Let $\mathbf{B}_i$ be the $(d_1+d_2)\times(d_1+d_2)$ matrix whose diagonal blocks are 0, and off diagonal blocks are equal to $\mathbf{A}_i$ and $\mathbf{A}_i^\top$ respectively. Then we have

$$[\Delta:\nabla^2 Q(\mathbf{W}):\Delta - 4\langle \nabla Q(\mathbf{W}), \Delta\rangle]$$

$$= -\frac{8}{n}\sum_{i=1}^{n}(\langle \Delta_{\mathbf{U}}\Delta_{\mathbf{V}}^\top, \mathbf{A}_i\rangle n_i) + \frac{16}{n}\sum_{i=1}^{n}(\langle \mathbf{U}\Delta_{\mathbf{V}}^\top + \Delta_{\mathbf{U}}\mathbf{V}^\top, \mathbf{A}_i\rangle n_i)$$
$$= \frac{24}{n}\sum_{i=1}^{n}(\langle \Delta_{\mathbf{U}}\Delta_{\mathbf{V}}^\top, \mathbf{A}_i\rangle n_i) + \frac{16}{n}\sum_{i=1}^{n}(\langle \mathbf{U}^\star\Delta_{\mathbf{V}}^\top + \Delta_{\mathbf{U}}(\mathbf{V}^\star)^\top, \mathbf{A}_i\rangle n_i)$$

Now we can use Lemma 2.14 again to bound the noise terms:

$$|\frac{24}{n}\sum_{i=1}^{n}(\langle \Delta_{\mathbf{U}}\Delta_{\mathbf{V}}^\top, \mathbf{A}_i\rangle n_i)| \leqslant$$
$$O\left(\sigma\sqrt{\frac{t\max\{t,d\}r\log n}{n}}\right)\sqrt{\|\Delta_{\mathbf{U}}\Delta_{\mathbf{V}}^\top\|_{\mathrm{F}}^2 + 0.001\sigma_r^\star\|\Delta_{\mathbf{V}}\|_{\mathrm{F}}^2}$$
$$|\frac{16}{n}\sum_{i=1}^{n}(\langle \mathbf{U}^\star\Delta_{\mathbf{V}}^\top + \Delta_{\mathbf{U}}(\mathbf{V}^\star)^\top, \mathbf{A}_i\rangle n_i)| \leqslant O\left(\sigma\sqrt{\frac{t\max\{t,d\}r\log n}{n}}\right)\|\mathbf{U}^\star\Delta_{\mathbf{V}}^\top + \Delta_{\mathbf{U}}(\mathbf{V}^\star)^\top\|_F.$$

On the one hand, by Lemma 2.7, we have:

$$\|\Delta_{\mathbf{U}}\Delta_{\mathbf{V}}^\top\|_{\mathrm{F}}^2 + 0.001\sigma_r^\star\|\Delta_{\mathbf{V}}\|_{\mathrm{F}}^2 \leqslant \|\Delta\Delta^\top\|_{\mathrm{F}}^2 + 0.001\sigma_r^\star\|\Delta\|_{\mathrm{F}}^2 \leqslant 3\|\mathbf{N}-\mathbf{N}^\star\|_{\mathrm{F}}^2.$$

On the other hand, again by Lemma 2.7, we have:

$$\|\mathbf{U}^\star\Delta_{\mathbf{V}}^\top + \Delta_{\mathbf{U}}(\mathbf{V}^\star)^\top\|_F^2 \leqslant \|\mathbf{W}^\star\Delta^\top + \Delta(\mathbf{W}^\star)^\top\|_F^2 = 2[\|\mathbf{W}^\star\Delta^\top\|_{\mathrm{F}}^2 + \|\Delta^\top\mathbf{W}^\star\|_{\mathrm{F}}^2] \leqslant 10\|\mathbf{N}-\mathbf{N}^\star\|_{\mathrm{F}}^2.$$

In sum, we have:

$$[\Delta : \nabla^2 Q(\mathbf{W}) : \Delta - 4\langle \nabla Q(\mathbf{W}), \Delta\rangle] \leqslant O\left(\sigma\sqrt{\frac{t\max\{t,d\}r\log n}{n}}\right)\|\mathbf{N}-\mathbf{N}^\star\|_{\mathrm{F}}.$$

Therefore, by Lemma 2.5, the Hessian at Δ direction is equal to:

$$\Delta : \nabla^2 f(\mathbf{W}) : \Delta \leqslant -0.1\|\mathbf{N}-\mathbf{N}^\star\|_F^2 + O(\sigma\sqrt{\frac{t\max\{t,d\}r\log n}{n}})\|\mathbf{N}-\mathbf{N}^\star\|_{\mathrm{F}}.$$

When the point further satisfies the second-order optimality condition we have

$$\|\mathbf{N}-\mathbf{N}^\star\|_F \leqslant O\left(\sigma\sqrt{\frac{t\max\{t,d\}r\log n}{n}}\right).$$

In particular, $\mathbf{M}-\mathbf{M}^\star$ is a submatrix of $\mathbf{N}-\mathbf{N}^\star$, so $\|\mathbf{M}-\mathbf{M}^\star\|_F \leqslant O(\sigma\sqrt{\frac{t\max\{t,d\}r\log n}{n}})$. □

Linear Algebra Lemmas

We collect together several useful linear algebra lemmas.

Lemma 2.7. *Given matrices* $\mathbf{W}, \mathbf{W}^\star \in \mathbb{R}^{d\times r}$*, let* $\mathbf{N} = \mathbf{W}\mathbf{W}^\top$ *and* $\mathbf{N}^\star = \mathbf{W}^\star(\mathbf{W}^\star)^\top$*, and let* $\Delta, \mathbf{R}^\star$ *be defined as in Definition 2.2, and let* $\tilde{\mathbf{W}}^\star = \mathbf{W}^\star\mathbf{R}^\star$*then we have the followings properties:*

1. $\mathbf{W}(\tilde{\mathbf{W}}^\star)^\top$ *is a symmetric p.s.d. matrix;*
2. $\|\Delta\Delta^\top\|_F^2 \leqslant 2\|\mathbf{N} - \mathbf{N}^\star\|_F^2$*;*
3. $\sigma_r^\star\|\Delta\|_F^2 \leqslant \|\Delta(\tilde{\mathbf{W}}^\star)^\top\|_F^2 \leqslant \frac{1}{2(\sqrt{2}-1)}\|\mathbf{N} - \mathbf{N}^\star\|_F^2$.
4. $\|\Delta^\top\tilde{\mathbf{W}}^\star\|_F^2 \leqslant \|\mathbf{N} - \mathbf{N}^\star\|_F^2$

Proof. Statement 1 is in the proof of [46, Lemma 6]. Statement 2 is by [46, Lemma 6]. Statement 3 & 4 follow by Lemma 2.8. □

Lemma 2.8. *Let* $\mathbf{U}$ *and* $\mathbf{Y}$ *be* $d \times r$ *matrices such that* $\mathbf{U}^\top\mathbf{Y} = \mathbf{Y}^\top\mathbf{U}$ *is a p.s.d. matrix. Then,*

$$\sigma_{\min}(\mathbf{U}^\top\mathbf{U})\|\mathbf{U} - \mathbf{Y}\|_F^2 \leqslant \|(\mathbf{U} - \mathbf{Y})\mathbf{U}^\top\|_F^2 \leqslant \frac{1}{2(\sqrt{2}-1)}\|\mathbf{U}\mathbf{U}^\top - \mathbf{Y}\mathbf{Y}^\top\|_F^2$$
$$\|(\mathbf{U} - \mathbf{Y})^\top\mathbf{U}\|_F^2 \leqslant \|\mathbf{U}\mathbf{U}^\top - \mathbf{Y}\mathbf{Y}^\top\|_F.$$

Proof. For the first statement, the left inequality is immediate, so we only need to prove right inequality. To prove this, we let $\Delta = \mathbf{U} - \mathbf{Y}$, and expand:

$$\begin{aligned}
&\|\mathbf{U}\mathbf{U}^\top - \mathbf{Y}\mathbf{Y}^\top\|_F^2 = \|\mathbf{U}\Delta^\top + \Delta\mathbf{U}^\top - \Delta\Delta^\top\|_F^2 \\
&= \mathrm{tr}(2\mathbf{U}^\top\mathbf{U}\Delta^\top\Delta + (\Delta^\top\Delta)^2 + 2(\mathbf{U}^\top\Delta)^2 - 4\mathbf{U}^\top\Delta\Delta^\top\Delta) \\
&= \mathrm{tr}((4-2\sqrt{2})\mathbf{U}^\top(\mathbf{U}-\Delta)\Delta^\top\Delta + (\Delta^\top\Delta - \sqrt{2}\mathbf{U}^\top\Delta)^2 + 2(\sqrt{2}-1)\mathbf{U}^\top\mathbf{U}\Delta^\top\Delta) \\
&\geqslant \mathrm{tr}((4-2\sqrt{2})\mathbf{U}^\top\mathbf{Y}\Delta^\top\Delta + 2(\sqrt{2}-1)\mathbf{U}^\top\mathbf{U}\Delta^\top\Delta) \geqslant 2(\sqrt{2}-1)\|\mathbf{U}\Delta^\top\|_{\mathrm{F}}^2.
\end{aligned}$$

The last inequality follows since $\mathbf{U}^\top\mathbf{Y}$ is a p.s.d. matrix. For the second statement, again, we have:

$$\begin{aligned}
\|\mathbf{U}\mathbf{U}^\top - \mathbf{Y}\mathbf{Y}^\top\|_F^2 &= \|\mathbf{U}\Delta^\top + \Delta\mathbf{U}^\top - \Delta\Delta^\top\|_F^2 \\
&= \mathrm{tr}(2\mathbf{U}^\top\mathbf{U}\Delta^\top\Delta + (\Delta^\top\Delta)^2 + 2(\mathbf{U}^\top\Delta)^2 - 4\mathbf{U}^\top\Delta\Delta^\top\Delta) \\
&= \mathrm{tr}(2\mathbf{U}^\top(\mathbf{U}-\Delta)\Delta^\top\Delta + (\Delta^\top\Delta - \mathbf{U}^\top\Delta)^2 + (\mathbf{U}^\top\Delta)^2) \\
&\geqslant \mathrm{tr}(2\mathbf{U}^\top\mathbf{Y}\Delta^\top\Delta + (\mathbf{U}^\top\Delta)^2) \geqslant \|\mathbf{U}^\top\Delta\|_{\mathrm{F}}^2,
\end{aligned}$$

where the last inequality follows since $\mathbf{U}^\top\Delta = \Delta^\top\mathbf{U}$. □

Concentration Lemmas

We need to show three concentration-style results for the landscape analysis. The first is an RIP condition for over matrices in the linear space $\mathcal{T} = \{\mathbf{U}_\star \mathbf{X}^\top + \mathbf{Y}\mathbf{V}_\star^\top | \mathbf{X} \in \mathbb{R}^{t\times r}, \mathbf{Y} \in \mathbb{R}^{d\times r}\}$ using matrix concentration. The second and third are coarse concentration results that exploit the rank r structure of the underlying matrix $\mathbf{M}$ and are used in the two distinct regimes where the distance to optimality can be small or large. Also note that throughout we can assume a left-sided incoherence condition on the underlying matrix of the form $\max_{i\in[t]} \|\mathbf{e}_i^\top \mathbf{U}_\star\|^2 \leqslant \frac{\mu r}{t}$ due to Assumption 2.2.

We first present the RIP-style matrix concentration result which rests on an application of the matrix Bernstein inequality over a projected space. The proof has a similar flavor to results in [99]. First we define a projection operator on the space of matrices as $P_{\mathcal{T}}\mathbf{Z} = P_{\mathbf{U}}\mathbf{Z} + \mathbf{Z}P_{\mathbf{V}} - P_{\mathbf{U}}\mathbf{Z}P_{\mathbf{V}}$ where $P_{\mathbf{U}}$ and $P_{\mathbf{V}}$ are orthogonal projections onto the subspaces spanned by U and V. While $P_{\mathbf{U}}$ and $P_{\mathbf{V}}$ are matrices, $P_{\mathcal{T}}$ is a linear operator mapping matrices to matrices. Intuitively we wish to show that for all $\mathbf{W} \in \mathbb{R}^{t\times d}$, that the observations matrices are approximately an isometry over the space of projected matrices w.h.p: $\frac{1}{n}\sum_{i=1}^n t(\langle \mathbf{e}_{t(i)}\mathbf{x}_i^\top, P_{\mathcal{T}}\mathbf{W}\rangle)^2 \approx \|P_{\mathcal{T}}\mathbf{W}\|_{\mathrm{F}}^2 = \|\mathbf{W}\|_{\mathrm{F}}^2$. Explicitly, we define the action of the operator $\mathbf{C}_i = \mathbf{A}_i\mathbf{A}_i^\top$ where $\mathbf{A}_i = \sqrt{t}\mathbf{x}_i\mathbf{e}_j^\top$ as $\mathbf{C}_i(\mathbf{M}) = t\mathbf{x}_i\mathbf{e}_j^\top\langle \mathbf{e}_j\mathbf{x}_i^\top, \mathbf{M}\rangle$.

We record a useful fact we will use in the sequel:

$$\sqrt{t}P_{\mathcal{T}}(\mathbf{x}_i\mathbf{e}_j^\top) = P_{\mathbf{U}}\mathbf{e}_i\mathbf{x}_j^\top + \mathbf{e}_i(P_{\mathbf{V}}\mathbf{x}_j)^\top - (P_{\mathbf{U}}\mathbf{e}_i)(P_{\mathbf{V}}\mathbf{x}_j)^\top \implies$$
$$\|P_{\mathcal{T}}(\mathbf{x}_i\mathbf{e}_j^\top\|_{\mathrm{F}}^2 = \langle P_{\mathcal{T}}(\mathbf{x}_i\mathbf{e}_j^\top), \mathbf{x}_i\mathbf{e}_j\rangle = \|P_{\mathbf{U}}\mathbf{e}_i\|^2\|\mathbf{x}_i\|^2 + \|\mathbf{e}_i\|^2\|P_{\mathbf{V}}\mathbf{x}_i\|^2 - \|P_{\mathbf{U}}\mathbf{e}_i\|^2\|P_{\mathbf{V}}\mathbf{x}_j\|^2 \leqslant$$
$$\|P_{\mathbf{U}}\mathbf{e}_i\|^2\|\mathbf{x}_i\|^2 + \|P_{\mathbf{V}}\mathbf{x}_i\|^2,$$

where the last inequality holds almost surely.

We now present the proof of the RIP-style concentration result.

Lemma 2.9. *Under Assumptions 2.1 and 2.2 and the uniform task sampling model above,*

$$\|\frac{1}{n}\sum_{i=1}^n P_{\mathcal{T}}\mathbf{A}_i\mathbf{A}_i^\top P_{\mathcal{T}} - P_{\mathcal{T}}\| \leqslant (\log(ndt)) \cdot O\left(\sqrt{\frac{\mu dr^2 + tr^2}{n}} + \frac{(\mu dr + rt)\log(tdn)}{n}\right),$$

with probability at least $1 - O(n^{-100})$*, where* $\mu = O(\bar{\kappa})$*.*

Proof. Note that under the task assumption, Lemma 2.16 diversity implies incoherence of the matrix $\mathbf{U}^\star$ with incoherence parameter $\mu = O(\bar{\kappa})$. First, note $\mathbb{E}[\mathbf{C}_i(\mathbf{M})] = \mathbf{M}$ so $\mathbb{E}[\frac{1}{n}\sum_{i=1}^n P_{\mathcal{T}}\mathbf{A}_iP_{\mathcal{T}} - P_{\mathcal{T}}] = 0$. To apply the truncated version of the matrix Bernstein inequality from Lemma 2.30 we first compute a bound on the norms of each $\mathbf{C}_i$ to set the truncation level R. Note that $\|P_{\mathcal{T}}\mathbf{A}_i\mathbf{A}_iP_{\mathcal{T}}\| = \|P_{\mathcal{T}}(\mathbf{x}_i\mathbf{e}_j^\top)\|_{\mathrm{F}}^2 \leqslant t \cdot O((\frac{\mu r}{t}\|\mathbf{x}_i\|^2 + \|P_{\mathbf{V}}\mathbf{x}_i\|^2))$ using the fact the operator $\mathbf{A}_i$ is rank-one along with the Lemma 2.16 which shows task diversity implies incoherence with incoherence parameter $\bar{\kappa}$. Now exploiting Lemma 2.29 we have that $\|\mathbf{x}_i\|^2 \leqslant O(d + \max(\sqrt{d\log(1/\delta)}, \log(1/\delta))) = O(d + \sqrt{d}\log(1/\delta))$ and $\|P_V\mathbf{x}_i\|^2 \leqslant O(r + \sqrt{r}\log(1/\delta))$ with probability at least $1 - 2\delta$ using sub-exponential tail bounds and

a union bound[11]. Hence $\|P_{\mathcal{T}}\mathbf{A}_i P_{\mathcal{T}}\| \leqslant O(\mu r d + \mu r\sqrt{d}\log(1/\delta)) + tr + t\sqrt{r}\log(1/\delta)) = O(\mu r d + tr + (\mu r\sqrt{d} + t\sqrt{r})\log(1/\delta))$.

We can rearrange this statement to conclude that $\mathbf{Pr}[\|P_{\mathcal{T}}\mathbf{A}_i\mathbf{A}_i^\top P_{\mathcal{T}}\| \geqslant c_1(\mu r d + tr) + x] \leqslant \exp(-c_2(\frac{x}{\mu r\sqrt{d}+t\sqrt{r}}))$ for some c_1, c_2. Define a truncation level $R = c_1(\mu r d + rt) + K(\mu r\sqrt{d} + t\sqrt{r})$ for some K to be chosen later. We can use the aforementioned tail bound to control $\|\mathbb{E}[Z_i] - \mathbb{E}[Z_i']\| \leqslant \mathbb{E}[Z_i \mathbb{1}[\|Z_i\| \geqslant R]] \leqslant \int_{K(\mu r\sqrt{d}+t\sqrt{r})}^{\infty} \exp(-c_2(\frac{x}{\mu r\sqrt{d}+t\sqrt{r}})) \leqslant O(\exp(-cK)(\mu r\sqrt{d} + t\sqrt{r})) = \Delta$.

Now we consider the task of bounding the matrix variance term. The calculation is somewhat tedious but straightforward under our assumptions. We make use of the standard result that for two matrices $\mathbf{X}$ and $\mathbf{Y}$ that $\|\mathbf{X} - \mathbf{Y}\| \leqslant \max(\|\mathbf{X}\|, \|\mathbf{Y}\|)$.

It suffices to bound the operator norm $\|\mathbf{E}[\|P_{\mathcal{T}}\mathbf{A}_i\|_{\mathrm{F}}^2 P_{\mathcal{T}}\mathbf{A}_i(P_{\mathcal{T}}\mathbf{A}_i)^\top\|$. Using the calculation from the prequel and carefully cancelling terms we can see that,

$$\begin{aligned}
&\|\mathbb{E}[\|P_{\mathcal{T}}\mathbf{A}_i\|_{\mathrm{F}}^2 P_{\mathcal{T}}\mathbf{A}_i(P_{\mathcal{T}}\mathbf{A}_i)^\top]\| \leqslant t^2\|\mathbb{E}[(\|P_{\mathbf{U}}\mathbf{e}_i\|^2\|\mathbf{x}_i\|^2 + \|P_{\mathbf{V}}\mathbf{x}_i\|^2 - \|P_{\mathbf{U}}\mathbf{e}_i\|^2\|P_{\mathbf{V}}\mathbf{x}_j\|^2)\cdot \\
&(\|P_{\mathbf{U}}\mathbf{e}_i\|^2\mathbf{x}_i\mathbf{x}_i^\top + \|P_{\mathbf{V}}\mathbf{x}\|^2\mathbf{e}_i\mathbf{e}_i^\top - \|P_{\mathbf{V}}\mathbf{x}_i\|^2 P_{\mathbf{U}}\mathbf{e}_i(P_{\mathbf{U}}\mathbf{e}_i)^\top]\| \\
&\leqslant t^2 O(\|\mathbb{E}[(\|P_{\mathbf{U}}\mathbf{e}_i\|^2\|\mathbf{x}_i\|^2\|P_{\mathbf{U}}\mathbf{e}_i\|^2\mathbf{x}_i\mathbf{x}_i^\top + (\|P_{\mathbf{U}}\mathbf{e}_i\|^2\|\mathbf{x}_i\|^2\|P_{\mathbf{V}}\mathbf{x}_i\|^2\mathbf{e}_i\mathbf{e}_i^\top\|] + \|P_{\mathbf{V}}\mathbf{x}_i\|^4\mathbf{e}_i\mathbf{e}_i^\top).
\end{aligned}$$

We show how to calculate these leading terms as the subleading terms can be shown to be lower-order by identical calculations. First note using the fact that $\mathbb{E}[\|P_{\mathbf{U}}\mathbf{e}_i\|^2] \leqslant \frac{r}{t} \leqslant 1$, since $t \geqslant r$ by the task diversity assumption. Then $t^2 \cdot \|\mathbb{E}[(\|P_{\mathbf{U}}\mathbf{e}_i\|^2\|\mathbf{x}_i\|^2\|P_{\mathbf{U}}\mathbf{e}_i\|^2\mathbf{x}_i\mathbf{x}_i^\top\| \leqslant \|\mu r t\mathbb{E}[\|\mathbf{x}_i\|^2\mathbf{x}_i\mathbf{x}_i^\top]\mathbb{E}[\|P_{\mathbf{U}}\mathbf{e}_i\|^2]\| \leqslant O(\mu r^2 d)$ appealing to the fact $\mathbb{E}[\|\mathbf{x}\|^2\mathbf{x}\mathbf{x}^\top] \preceq O(\mathbf{I}_d)$ by Lemma 2.27.

Similarly, we have that, $t^2 \cdot \|\mathbb{E}[\|P_{\mathbf{U}}\mathbf{e}_i\|^2\|\mathbf{x}_i\|^2\|P_{\mathbf{V}}\mathbf{x}_i\|^2\mathbf{e}_i\mathbf{e}_i^\top]\| \leqslant \mu r\mathbb{E}[\|\mathbf{x}_i\|^2\|P_{\mathbf{V}}\mathbf{x}_i\|^2] \leqslant \mu r^2 d$ using incoherence and by Lemma 2.27. Finally, we have that $t^2 \cdot O(\|\mathbb{E}[\|P_{\mathbf{V}}\mathbf{x}_i\|^4\mathbf{e}_i\mathbf{e}_i^\top\|) \leqslant O(tr^2)$. Hence we have that $\sigma^2 = n \cdot O(\mu r^2 d + tr^2)$.

Finally, we can assemble the previous two computations to conclude the result with appropriate choices of R (parametrized through K) and x by combining with Lemma 2.30. Let us choose $K = \frac{c_3}{c}(\log(n) + \log(d) + \log(t))$ for some sufficiently large c_3. In this case, we can choose c_3 such that $\Delta \leqslant O(\frac{\mu r\sqrt{d}+t\sqrt{r}}{n^{10}d^{10}t^{10}}) \leqslant O(\frac{\mu}{n^{10}d^8})$. Similarly, our choice of truncation level becomes $R = O(\mu r d + tr + (\log n + \log d + \log t)(\mu r\sqrt{d} + t\sqrt{r})$. At this point we now choose $x = c_4(\log n + \log d + \log t)\max(\sigma/\sqrt{n}, R/n)$ for sufficiently large c_4. For large enough c_4 we can guarantee that $x \geqslant 2\Delta \implies x - \Delta \geqslant \frac{x}{2}$.

Hence combining these results together and applying Lemma 2.30 we can provide the following upper bound on the desired quantity:

$$\begin{aligned}
&\mathbf{Pr}[\|\frac{1}{n}\sum_{i=1}^n P_{\mathcal{T}}\mathbf{A}_i P_{\mathcal{T}} - P_{\mathcal{T}}\| \geqslant x] \leqslant \\
&O(d\exp(-c \cdot c_4(\log n + \log d + \log t)) +
\end{aligned}$$

[11]Note that by definition the orthogonal projection of a d-dimensional subgaussian random vector onto an r-dimensional subspace is an r-dimensional subgaussian random vector.

$$O(n\sqrt{K}\exp(-c_3(\log n + \log d + \log t)) \leqslant O((ndt)^{-100})$$

by taking c_3 and c_4 sufficiently large, with

$$x = O((\log(ndt))\left(\sqrt{\frac{\mu dr^2 + tr^2}{n}} + \frac{(\mu rd + tr) + (\mu r\sqrt{d} + t\sqrt{r})(\log(ndt)))}{n}\right)$$

. □

Lemma 2.10. *Let the covariates $\mathbf{x}_i$ satisfy the design conditions in Assumption 2.1 in the uniform task sampling model. Then for all matrices $\mathbf{M}$ matrices that are of rank $2r$, we have uniformly that,*

$$|\frac{1}{n}\sum_{i=1}^{n}\langle \mathbf{A}_i, \mathbf{M}\rangle^2 - \|\mathbf{M}\|_F^2| \leqslant$$
$$O\left(\sqrt{\frac{\max(t,d)r}{n}} \cdot \sqrt{t}\max_i \|\mathbf{e}_i^\top \mathbf{M}\|\|\mathbf{M}\|_F + \frac{\max(t,d)r}{n} \cdot t\max_i \|\mathbf{e}_i^\top \mathbf{M}\|^2\right).$$

with probability at least $1 - (3000r)^{-10\max(t,d)r}$.

Proof. Note that by rescaling it suffices to restrict attention to matrices $\mathbf{M}$ that are of rank $2r$ and have Frobenius norm 1 (a set which we denote Γ). Applying Lemma 2.11, we have that,

$$|\frac{1}{n}\sum_{i=1}^{n} t(\mathbf{e}_{t(i)}^\top \mathbf{M}\mathbf{x}_i)^2 - \|\mathbf{M}\|_F^2| \leqslant O\left(\frac{1}{\sqrt{n}}\sqrt{\log(\frac{1}{\delta})} + \frac{1}{n}\log(\frac{1}{\delta})\right),$$

for any fixed $\mathbf{M} \in \Gamma$ with probability at least $1 - \delta$. Now using 2.12 with $\epsilon = \frac{1}{1000}$ have that the set Γ admits a cover K of size at most $|K| = (3000r)^{(t+d+1)r}$. Now by choosing $\delta = (3000)^{-c(t+d+1)r}$ for a sufficiently large constant c we can ensure that,

$$|\frac{1}{n}\sum_{i=1}^{n} t(\mathbf{e}_{t(i)}^\top \mathbf{M}_j\mathbf{x}_i)^2 - \|\mathbf{M}_j\|_F^2| \leqslant O\left(\frac{1}{\sqrt{n}}\sqrt{(\max(t,d)r} + \frac{1}{n}\max(t,d)r\right) \quad \forall \mathbf{M}_j \in K,$$

with probability at least $1 - (3000r)^{-10\max(t,d)r}$ using a union bound. Now a straightforward Lipschitz continuity argument shows that since any $\mathbf{M} \in \Gamma$ can be written as $\mathbf{M} = \mathbf{M}_i + \epsilon a_i$ for $\mathbf{M}_i \in K$ and another $a_i \in \Gamma$, then

$$\sup_{\mathbf{M}\in\Gamma} |\frac{1}{n}\sum_{i=1}^{n} t(\mathbf{e}_{t(i)}^\top \mathbf{M}\mathbf{x}_i)^2 - \|\mathbf{M}\|_F^2| \leqslant 2(\sup_{\mathbf{M}_j\in K} |\frac{1}{n}\sum_{i=1}^{n} t(\mathbf{e}_{t(i)}^\top \mathbf{M}_j\mathbf{x}_i)^2 - \|\mathbf{M}_j\|_F^2|),$$

and hence the conclusion follows. Rescaling the result by $\|\mathbf{M}\|_{\mathrm{F}}^2$ finishes the result. □

Lemma 2.11. *Let the covariates $\mathbf{x}_i$ satisfy the design condition in Assumption 2.1 in the uniform task sampling model. Then if $Y_i = t(\mathbf{e}_{t(i)}^\top \mathbf{A}\mathbf{x}_i)^2 - \|\mathbf{A}\|_F^2$, Y_i is a sub-exponential random variable, and*

$$|\frac{1}{n}\sum_{i=1}^{n} t(\mathbf{e}_{t(i)}^\top \mathbf{M}\mathbf{x}_i)^2 - \|\mathbf{M}\|_F^2| \leqslant$$
$$O\left(\frac{\sqrt{t}\max_i \|\mathbf{e}_i^\top \mathbf{M}\|_2 \|\mathbf{M}\|_F}{\sqrt{n}}\sqrt{\log(\frac{1}{\delta})} + \frac{t\max_i \|\mathbf{e}_i^\top \mathbf{M}\|^2}{n}\log(\frac{1}{\delta})\right),$$

for any fixed $\mathbf{M}$ with probability at least $1-\delta$.

Proof. First note that under our assumptions Y_i, $\mathbb{E}[t(\mathbf{e}_j^\top \mathbf{A}\mathbf{x}_i)^2] = \|\mathbf{A}\|_F^2$. To establish the result, we show the Bernstein condition holds with appropriate parameters. To do so, we bound for $k \geqslant 1$,

$$|\mathbb{E}[Y_i^k]| \leqslant t^k 2^{2k}\mathbb{E}[(\mathbf{e}_j^\top \mathbf{M}_j \mathbf{x}_i)^{2k}] = t^k 2^{2k}\cdot \mathbb{E}[\|\mathbf{e}_j^\top \mathbf{M}_j\|^{2k}]C^{2k}k! \leqslant$$
$$(C')^{4k}k!\cdot \mathbb{E}[(t\|\mathbf{e}_j^\top \mathbf{M}\|^2)^{k-1}\cdot (t\|\mathbf{e}_j^\top \mathbf{M}\|^2)] \leqslant (C'')^k k!(t\nu^2)^{k-2}(t\nu^2\|\mathbf{A}\|_{\mathrm{F}}^2) =$$
$$\frac{1}{2}k!(\underbrace{C''t\nu^2}_{b})^{k-2}\cdot(\underbrace{C''^2 t\nu^2\|\mathbf{M}\|_{\mathrm{F}}^2}_{\sigma^2}),$$

by introducing an independent copy of Y, using Jensen's inequality, and the inequality $(\frac{a+b}{2})^k \leqslant 2^{k-1}(a^k+b^k)$ in the first inequality, and the sub-gaussian moment bound $\mathbb{E}[Z^{2k}] \leqslant 2k\Gamma(k)C^{2k} \leqslant k!C^{2k}$ for universal constant C which holds under our design assumptions. Hence directly applying the Bernstein inequality (see [113, Proposition 2.9] shows that,

$$\mathbb{E}[e^{\lambda\cdot Y_i}] \leqslant e^{\lambda^2(\sqrt{2}\sigma)^2/2} \quad \forall|\lambda| \leqslant \frac{1}{2b}.$$

Hence, using a standard sub-exponential tail bound we conclude that,

$$|\frac{1}{n}\sum_{i=1}^{n} t(\mathbf{e}_{t(i)}^\top \mathbf{M}\mathbf{x}_i)^2 - \|\mathbf{M}\|_F^2| \leqslant O\left(\frac{\sigma}{\sqrt{n}}\sqrt{\log(\frac{1}{\delta})} + \frac{b}{n}\log(\frac{1}{\delta})\right) =$$
$$O\left(\frac{\sqrt{t}\max_i \|\mathbf{e}_i^\top \mathbf{M}\|_2 \|\mathbf{M}\|_{\mathrm{F}}}{\sqrt{n}}\sqrt{\log(\frac{1}{\delta})} + \frac{t\max_i \|\mathbf{e}_i^\top \mathbf{M}\|^2}{n}\log(\frac{1}{\delta})\right),$$

for any fixed $\mathbf{A} \in \Gamma$ with probability at least $1-\delta$. □

We now restate a simple covering lemma for rank-$O(r)$ matrices from [13].

Lemma 2.12 (Lemma 3.1 from [13])**.** *Let Γ be the set of matrices $\mathbf{M} \in \mathbb{R}^{t\times d}$ that are of rank at most r and have Frobenius norm equal to 1. Then for any $\epsilon < 1$, there exists an ϵ-net covering of Γ in the Frobenius norm, S, which has cardinality at most $(\frac{9}{\epsilon})^{(t+d+1)r}$.*

We now state a central lemma which combines the previous concentration arguments into a single condition we use in the landscape analysis.

Lemma 2.13. *Let Assumptions 2.1 and 2.2 hold in the uniform task sampling model. When number of samples is greater than* $n \geqslant C\text{polylog}(d,n,t)\mu^2 r^4 \max\{t,d\}(\kappa^\star)^4$ *with large-enough constant* C*, with at least* $1-1/\text{poly}(d)$ *probability, we have following holds for all* $(\mathbf{U},\mathbf{V}) \in \mathcal{W}$ *simultaneously:*

$$\frac{1}{n}\sum_{i=1}^{n}\langle \mathbf{U}^\star\Delta_{\mathbf{V}}^\top + \Delta_{\mathbf{U}}(\mathbf{V}^\star)^\top, \mathbf{A}_i\rangle^2 \in (1\pm 0.001)\|\mathbf{U}^\star\Delta_{\mathbf{V}}^\top + \Delta_{\mathbf{U}}(\mathbf{V}^\star)^\top\|_F^2 \tag{2.17}$$

$$\frac{1}{n}\sum_{i=1}^{n}\langle \mathbf{A}_i, \Delta_{\mathbf{U}}\Delta_{\mathbf{V}}^\top\rangle^2 \leqslant \|\Delta_{\mathbf{U}}\Delta_{\mathbf{V}}^\top\|_F^2 + 0.001\sigma_r^\star\|\Delta_{\mathbf{V}}\|_F^2 \tag{2.18}$$

$$\frac{1}{n}\sum_{i=1}^{n}\langle \mathbf{A}_i, \mathbf{M}-\mathbf{M}^\star\rangle^2 \geqslant \|\mathbf{M}-\mathbf{M}^\star\|_F^2 - (\sigma_r^\star)^2/10^6, \tag{2.19}$$

where $\mathbf{M} = \mathbf{U}\mathbf{V}^\top$ *and* $\Delta_{\mathbf{U}}, \Delta_{\mathbf{V}}$ *are defined as in Definition 2.2. Here* $\mu = O(\bar{\kappa})$.

Proof. This result follows immediately by applying Lemma 2.9 to the first statement and Lemma 2.10 to the following two statements using the definition of the incoherence ball $\mathcal{W}$. □

Lemma 2.14. *Suppose the set of matrices* $\mathbf{A}_1, \mathbf{A}_2, ..., \mathbf{A}_n$ *satisfy the event in Lemma 2.13, let* $n_1, n_2, ..., n_m$ *be i.i.d. sub-gaussian random variables with variance parameter* $t\sigma^2$*, then with high probability for any* $(\mathbf{U},\mathbf{V}) \in \mathcal{W}$*, we have*

$$|\frac{1}{n}\sum_{i=1}^{n}(\langle \Delta_{\mathbf{U}}\Delta_{\mathbf{V}}^\top, \mathbf{A}_i\rangle n_i)| \leqslant O(\sigma\sqrt{\frac{t\max\{t,d\}r\log n}{n}})\sqrt{\|\Delta_{\mathbf{U}}\Delta_{\mathbf{V}}^\top\|_F^2 + 0.001\sigma_r^\star\|\Delta_{\mathbf{V}}\|_F^2}$$

$$|\frac{1}{n}\sum_{i=1}^{n}(\langle \mathbf{U}^\star\Delta_{\mathbf{V}}^\top + \Delta_{\mathbf{U}}(\mathbf{V}^\star)^\top, \mathbf{A}_i\rangle n_i)| \leqslant O(\sigma\sqrt{\frac{t\max\{t,d\}r\log n}{n}})\|\mathbf{U}^\star\Delta_{\mathbf{V}}^\top + \Delta_{\mathbf{U}}(\mathbf{V}^\star)^\top\|_F$$

for $n \gtrsim \text{polylog}(d)$.

Proof. Note since the left hand side of the expressions are linear in the matrices we can normalize to those of Frobenius norm 1. The proof of both statements is identical so we simply prove the second.

Define $\delta = \|\mathbf{U}^\star\Delta_{\mathbf{V}}^\top + \Delta_{\mathbf{U}}(\mathbf{V}^\star)^\top\|_{\text{F}}$ and $\mathbf{M} = \mathbf{U}^\star\Delta_{\mathbf{V}}^\top + \Delta_{\mathbf{U}}(\mathbf{V}^\star)^\top$ for convenience, which can be thought of as arbitrary rank-r matrices. Then let S be an ϵ-net for all rank-r matrices with Frobenius norm 1; by Lemma 2.12 we have that $\log|S| \leqslant O(\max(t,d)r\log(\frac{1}{\epsilon}))$. We set $\epsilon = \frac{1}{n^3}$ so $\log(\frac{1}{\epsilon}) = O(\log n)$. Now for any matrix $\mathbf{M} \in S$ we have that $\frac{1}{n}\langle \mathbf{A}_i, \mathbf{M}\rangle$ is a sub-gaussian random variable with variance parameter at most $t\sigma^2\frac{\delta^2}{n}$. Thus, using a sub-gaussian tail

bound along with a union bound over the net shows that uniformly over the $\mathbf{M} \in S$,

$$|\frac{1}{n}\sum_{i=1}^{n}\langle \mathbf{M}, \mathbf{A}_i\rangle n_i)| \leqslant O\left(\sigma\delta\sqrt{\frac{t\max\{t,d\}r\log n}{n}}\right),$$

with probability at least $1 - \frac{1}{\text{poly}(d)}$. We now show how to lift to the set of all $\mathbf{M}$. Note that with probability at least $1 - e^{-\Omega(n)}$ that $\|\mathbf{n}\| = O(\sqrt{t}\sigma\sqrt{n})$ by a sub-gaussian tail bound (see for example Lemma 2.29). Let $\mathbf{M}$ be an arbitrary element, and $\mathbf{M}'$ its closest element in the cover; then we have that $\mathbf{z}_i = \langle \mathbf{A}_i, \mathbf{M} - \mathbf{M}'\rangle \leqslant \frac{\delta}{n^2}$ using the precondition on $\mathbf{A}_i$. Combining and using a union bound then shows that,

$$|\frac{1}{n}\sum_{i=1}^{n} n_i\langle \mathbf{A}_i, \mathbf{M}\rangle| \leqslant |\frac{1}{n}\sum_{i=1}^{n} n_i\langle \mathbf{A}_i, \mathbf{M}'\rangle| + |\frac{1}{n}\sum_{i=1}^{n} n_i\langle \mathbf{A}_i, \mathbf{M} - \mathbf{M}'\rangle| \leqslant$$
$$O\left(\sigma\delta\sqrt{\frac{t\max\{t,d\}r\log n}{n}}\right) + \frac{\sqrt{t}\sigma\delta}{\sqrt{n}} \leqslant O\left(\sigma\delta\sqrt{\frac{t\max(t,d)r\log n}{n}}\right).$$

Rescaling and recalling the definition of δ gives the result. □

Task Diversity for the Landscape Analysis

Here we collect several useful results for interpreting the results of the landscape analysis. Throughout this section we use the notation $\mathbf{U} \in \mathbb{R}^{t\times r}$ and $\mathbf{V} \in \mathbb{R}^{d\times r}$.

The first result allow us to convert a guarantee on error in Frobenius norm to a guarantee in angular distance, assuming an appropriate diversity condition on $\mathbf{U}$.

Lemma 2.15. *Suppose $\mathbf{V}$ and $\hat{\mathbf{V}}$ are orthonormal projection matrices, that is $\mathbf{V}^\top\mathbf{V} = \mathbf{I}_r$, and $\hat{\mathbf{V}}^\top\hat{\mathbf{V}} = \mathbf{I}_r$. Then, for any $\epsilon > 0$, if $\|\hat{\mathbf{U}}\hat{\mathbf{V}}^\top - \mathbf{U}\mathbf{V}^\top\|_F^2 \leqslant \epsilon$ for some $\hat{\mathbf{U}}$ and $\mathbf{U}$, then:*

$$\text{dist}^2(\mathbf{V}, \hat{\mathbf{V}}) \leqslant \frac{\epsilon}{\nu t},$$

where $\nu = \sigma_r(\mathbf{U}^\top\mathbf{U})/t$.

Here the distance function is the sine function of the principal angle; i.e.

$$\text{dist}(\mathbf{V}, \hat{\mathbf{V}}) := \|\mathbf{V}^\top\hat{\mathbf{V}}_\perp\|,$$

and $\nu = \sigma_r(\mathbf{U}^\top\mathbf{U})/t$ represents an analog of the task diversity matrix.

Proof. Define the function $f(\tilde{\mathbf{U}}) = \|\tilde{\mathbf{U}}\hat{\mathbf{V}}^\top - \mathbf{U}\mathbf{V}^\top\|_{\text{F}}^2$. The precondition of the theorem states that there exists $\hat{\mathbf{U}}$ so that $\|\hat{\mathbf{U}}\hat{\mathbf{V}}^\top - \mathbf{U}\mathbf{V}^\top\|_{\text{F}}^2 \leqslant \epsilon$. This clearly implies the following:

$$\min_{\tilde{\mathbf{U}}} f(\tilde{\mathbf{U}}) \leqslant \epsilon. \tag{2.20}$$

Setting the gradient $\mathrm{d}f/\mathrm{d}\tilde{\mathbf{U}} = 0$, we have the minimizer $\tilde{\mathbf{U}}^\star$ satisfies:

$$(\tilde{\mathbf{U}}^\star\hat{\mathbf{V}}^\top - \mathbf{U}\mathbf{V}^\top)\hat{\mathbf{V}} = 0,$$

which gives:

$$\tilde{\mathbf{U}}^\star = \mathbf{U}\mathbf{V}^\top\hat{\mathbf{V}}.$$

Plugging this back to Eq. (2.20) gives:

$$\|\mathbf{U}\mathbf{V}^\top(\hat{\mathbf{V}}\hat{\mathbf{V}}^\top - \mathbf{I})\|_{\mathrm{F}}^2 \leqslant \epsilon.$$

Finally, we have:

$$\begin{aligned}\|\mathbf{U}\mathbf{V}^\top(\hat{\mathbf{V}}\hat{\mathbf{V}}^\top - \mathbf{I})\|_{\mathrm{F}}^2 =& \|\mathbf{U}\mathbf{V}^\top\hat{\mathbf{V}}_\perp\hat{\mathbf{V}}_\perp^\top\|_{\mathrm{F}}^2 = \|\mathbf{U}\mathbf{V}^\top\hat{\mathbf{V}}_\perp\|_{\mathrm{F}}^2 = \mathrm{tr}(\mathbf{U}^\top\mathbf{U}\mathbf{V}^\top\hat{\mathbf{V}}_\perp\hat{\mathbf{V}}_\perp^\top\mathbf{V}) \\ \geqslant& \sigma_r(\mathbf{U}^\top\mathbf{U})\|\mathbf{V}^\top\hat{\mathbf{V}}_\perp\|_{\mathrm{F}}^2 \geqslant \sigma_r(\mathbf{U}^\top\mathbf{U})\|\mathbf{V}^\top\hat{\mathbf{V}}_\perp\|^2.\end{aligned}$$

The second last inequality follow since for any p.s.d. matrices $\mathbf{A}$ and $\mathbf{B}$, we have $\mathrm{tr}(\mathbf{A}\mathbf{B}) \geqslant \sigma_{\min}(\mathbf{A})\mathrm{tr}(\mathbf{B})$. This concludes the proof. □

For the following let $\mathbf{A} = (\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_t)^\top \in \mathbb{R}^{t\times r}$ and denote the SVD of $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$. Next we remark that our assumptions on task diversity and normalization implicit in the matrix $\mathbf{A}$ are sufficient to actually imply an incoherence condition on $\mathbf{U}$ (which is used in the matrix sensing/completion style analysis).

Lemma 2.16. *If* $\mu = \frac{1}{r\sigma_r(\mathbf{A}^\top\mathbf{A}/t)}$ *and* $\max_{i\in[t]}\|\boldsymbol{\alpha}_i\|^2 \leqslant C$*, then we have:*

$$\max_{i\in[t]}\|\mathbf{e}_i^\top\mathbf{U}\|^2 \leqslant \frac{C\mu r}{t}.$$

Proof. Since $\max_{i\in[t]}\|\boldsymbol{\alpha}_i\|^2 \leqslant C$, we have, for any $i \in [t]$

$$C \geqslant \|\boldsymbol{\alpha}_i\|^2 = \|\mathbf{e}_i^\top\mathbf{A}\|^2 = \|\mathbf{e}_i^\top\mathbf{U}\boldsymbol{\Sigma}\|^2 \geqslant \|\mathbf{e}_i^\top\mathbf{U}\|^2\sigma_{\min}^2(\boldsymbol{\Sigma}) = \|\mathbf{e}_i^\top\mathbf{U}\|^2\sigma_r(\mathbf{A}^\top\mathbf{A}) = (t/\mu r)\|\mathbf{e}_i^\top\mathbf{U}\|^2,$$

which finishes the proof. □

Note in the context of Assumption 2.2 the incoherence parameter corresponds to the parameter $\bar{\kappa} \leqslant \kappa$ since under our normalization $\mathrm{tr}(\mathbf{A}^\top\mathbf{A}/t) = \Theta(1)$. Further to quickly verify the incoherence ball contains the true parameters it is important to recall the scale difference $\mathbf{M}^\star$ and $\mathbf{A}^\top\mathbf{A}/t$ by a factor of $\sqrt{t}$.

2.11 Proofs for Section 2.4

Assuming we have obtained an estimate of the column space or feature set $\hat{\mathbf{B}}$ for the initial set of tasks, such that $\|\hat{\mathbf{B}}^\top \mathbf{B}\| \leqslant \delta$, we now analyze the performance of the plug-in estimator (which explicitly uses the estimate $\hat{\mathbf{B}}$ in lieu of the unknown $\mathbf{B}$) on a new task. Recall we define the estimator for the new tasks by a projected linear regression estimator: $\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \|\mathbf{y} - \mathbf{X}\hat{\mathbf{B}}\boldsymbol{\alpha}\|^2 \implies \hat{\boldsymbol{\alpha}} = (\hat{\mathbf{B}}^\top \mathbf{X}^\top \mathbf{X}\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}^\top \mathbf{X}^\top \mathbf{y}$.

Analyzing the performance of this estimator requires first showing that the low-dimensional empirical covariance and empirical correlation concentrate in $\tilde{O}(r)$ samples and performing an error decomposition to compute the bias resulting from using $\hat{\mathbf{B}}$ in lieu of $\mathbf{B}$ as the feature representation. We measure the performance the estimator with respect to its estimation error with respect to the underlying parameter $\mathbf{B}\boldsymbol{\alpha}_0$; in particular, we use $\|\hat{\mathbf{B}}\hat{\boldsymbol{\alpha}} - \mathbf{B}\boldsymbol{\alpha}_0\|^2$. Note that our analysis can accommodate covariates $\mathbf{x}_i$ generated from non-isotropic *non-Gaussian* distributions. In fact the only condition we require on the design is that the covariates are sub-gaussian random vectors in the following sense.

Assumption 2.3. *Each covariate vector $\mathbf{x}_i$ is mean-zero, satisfies $\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}$-sub-gaussian, in the sense that $\mathbb{E}[\exp(\mathbf{v}^\top \mathbf{x}_i)] \leqslant \exp\left(\frac{\|\boldsymbol{\Sigma}^{1/2}\mathbf{v}\|^2}{2}\right)$. Moreover, the additive noise ϵ_i is i.i.d. sub-gaussian with variance parameter 1 and is independent of $\mathbf{x}_i$.*

Note that Assumption 2.1 immediately implies Assumption 2.3.

Throughout this section we will let $\hat{\mathbf{B}}$ and $\hat{\mathbf{B}}_\perp$ be orthonormal projection matrices spanning orthogonal subspaces which are rank r and rank $d - r$ respectively—so that $\text{range}(\hat{\mathbf{B}}) \oplus \text{range}(\hat{\mathbf{B}}_\perp) = \mathbb{R}^d$.

Proof of Theorem 2.4. To begin we use the definition of

$$\hat{\boldsymbol{\alpha}} = (\hat{\mathbf{B}}^\top \mathbf{X}^\top \mathbf{X}\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}\mathbf{X}^\top \mathbf{y} = (\hat{\mathbf{B}}^\top \mathbf{X}^\top \mathbf{X}\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}\mathbf{X}^\top \mathbf{X}\mathbf{B}\boldsymbol{\alpha}_0 + (\hat{\mathbf{B}}^\top \mathbf{X}^\top \mathbf{X}\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}\mathbf{X}^\top \boldsymbol{\epsilon}$$

to decompose the error as,

$$(\hat{\mathbf{B}}\hat{\boldsymbol{\alpha}} - \mathbf{B}\boldsymbol{\alpha}_0) = \hat{\mathbf{B}}(\hat{\mathbf{B}}^\top \mathbf{X}^\top \mathbf{X}\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}\mathbf{X}^\top \mathbf{X}\mathbf{B}\boldsymbol{\alpha}_0 - \mathbf{B}\boldsymbol{\alpha}_0 + \hat{\mathbf{B}}(\hat{\mathbf{B}}^\top \mathbf{X}^\top \mathbf{X}\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}^\top \mathbf{X}^\top \boldsymbol{\epsilon}.$$

Now squaring both sides of the equation gives, so

$$\|\hat{\mathbf{B}}\hat{\boldsymbol{\alpha}} - \mathbf{B}\boldsymbol{\alpha}_0\|^2 \leqslant 2(\|\hat{\mathbf{B}}(\hat{\mathbf{B}}^\top \mathbf{X}^\top \mathbf{X}\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}\mathbf{X}^\top \mathbf{X}\mathbf{B}\boldsymbol{\alpha}_0 - \mathbf{B}\boldsymbol{\alpha}_0\|^2 + \|\hat{\mathbf{B}}(\hat{\mathbf{B}}^\top \mathbf{X}^\top \mathbf{X}\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}^\top \mathbf{X}^\top \boldsymbol{\epsilon}\|^2).$$

The first bias term can be bounded by Lemma 2.17, while the the variance term can be bounded by Lemma 2.18 . Combining the results and using a union bound gives the result. □

We now present the lemmas which allow us to bound the variance terms in the aforementioned error decomposition. For the following two results we also track the conditioning dependence with respect $C_{\min}$ and $C_{\max}$. We first control the term arising from the projection of the additive noise onto the empirical covariance matrix.

Lemma 2.17. *Let the sequence of n i.i.d. covariates $\mathbf{x}_i$ and n i.i.d. additive noise variables ϵ_i satisfy Assumption 2.3. Then if $n \gtrsim C_{cond}^2 r \log n$,*

$$\|\hat{\mathbf{B}}(\hat{\mathbf{B}}^\top \mathbf{X}^\top \mathbf{X}\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}^\top \mathbf{X}^\top \boldsymbol{\epsilon}\|^2 \leqslant O\left(\frac{r\log n}{C_{\min} n}\right),$$

with probability at least $1 - O(n^{-100})$.

Proof. Since $\|\hat{\mathbf{B}}(\hat{\mathbf{B}}^\top \mathbf{X}^\top \mathbf{X}\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}^\top \mathbf{X}^\top \boldsymbol{\epsilon}\|^2 \leqslant \|(\hat{\mathbf{B}}^\top \mathbf{X}^\top \mathbf{X}\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}^\top \mathbf{X}^\top \boldsymbol{\epsilon}\|^2$, it suffices to bound the latter term. Consider $\boldsymbol{\epsilon}^\top \underbrace{\frac{1}{n}\frac{\mathbf{X}\hat{\mathbf{B}}}{\sqrt{n}}(\hat{\mathbf{B}}^\top \frac{\mathbf{X}^\top \mathbf{X}}{n}\hat{\mathbf{B}})^{-2}\frac{\hat{\mathbf{B}}^\top \mathbf{X}^\top}{\sqrt{n}}}_{\mathbf{A}} \boldsymbol{\epsilon}$. So applying the Hanson-Wright inequality [110, Theorem 6.2.1] (conditionally on $\mathbf{X}$) to conclude that $\mathbf{Pr}[|\boldsymbol{\epsilon}^\top \mathbf{A}\boldsymbol{\epsilon} - \mathbf{E}[\boldsymbol{\epsilon}^\top \mathbf{A}\boldsymbol{\epsilon}]| \geqslant t] \leqslant 2\exp(-c\min(\frac{t^2}{\|\mathbf{A}\|_F^2}, \frac{t}{\|\mathbf{A}\|}))$. Hence

$$\boldsymbol{\epsilon}^\top \mathbf{A}\boldsymbol{\epsilon} \leqslant \mathbf{E}[\boldsymbol{\epsilon}^\top \mathbf{A}\boldsymbol{\epsilon}] + O(\|\mathbf{A}\|_F\sqrt{\log(2/\delta_1)}) + O(\|\mathbf{A}\|\log(2/\delta_1))$$

with probability at least $1 - \delta_1$.

Now using cyclicity of the trace we have that $\mathbb{E}[\boldsymbol{\epsilon}^\top \mathbf{A}\boldsymbol{\epsilon}] = \frac{1}{n}\mathrm{tr}[(\hat{\mathbf{B}}^\top \frac{\mathbf{X}^\top \mathbf{X}}{n}\hat{\mathbf{B}})^{-1}]$. Similarly $\|\mathbf{A}\| = \frac{1}{n}\|(\hat{\mathbf{B}}^\top \frac{\mathbf{X}^\top \mathbf{X}}{n}\hat{\mathbf{B}})^{-1}\| = \frac{1}{n}\|(\mathbf{E} + \hat{\mathbf{B}}^\top \boldsymbol{\Sigma}\hat{\mathbf{B}})^{-1}\|$. Applying Lemma 2.19 to the matrix $\mathbf{E}$ with $\delta = n^{-200}$ and assuming $n \gtrsim C_{\mathrm{cond}}^2 r\log(1/\delta) \gtrsim C_{\mathrm{cond}}^2 r\log n$ shows that $\|(\hat{\mathbf{B}}^\top \boldsymbol{\Sigma}\hat{\mathbf{B}})^{-1}\mathbf{E}\| \leqslant \frac{1}{4}$. Also note that on this event and this regime of sufficiently large n, this concentration result shows that $\sigma_{\min}(\hat{\mathbf{B}}^\top \frac{\mathbf{X}^\top \mathbf{X}}{n}\hat{\mathbf{B}}) > C_{\min}/2$, so the matrix is invertible. Hence an application of Lemma 2.24 shows that $\|\mathbf{A}\| \leqslant \frac{1}{n}(\frac{1}{C_{\min}} \cdot (1 + C_{\mathrm{cond}}\sqrt{\frac{r\log n}{n}})) \leqslant O(\frac{1}{C_{\min} n})$. Similarly since $\frac{\mathbf{X}\hat{\mathbf{B}}}{\sqrt{n}}$ is rank r and invertible on this event, it follows $\|\mathbf{A}\|_{\mathrm{F}} \leqslant \sqrt{r}\|\mathbf{A}\| \leqslant O(\frac{\sqrt{r}}{C_{\min} n})$ and that $\frac{1}{n}\mathrm{tr}[(\hat{\mathbf{B}}^\top \frac{\mathbf{X}^\top \mathbf{X}}{n}\hat{\mathbf{B}})^{-1}] \leqslant \frac{r}{C_{\min} n}$.

Hence taking $\delta_1 = n^{-200}$, and using the union bound, we conclude that $\boldsymbol{\epsilon}^\top \mathbf{A}\boldsymbol{\epsilon} \leqslant \frac{1}{C_{\min}} \cdot O(\frac{r}{n}) + O(\frac{\sqrt{r\log n}}{n}) + O(\frac{\log n}{n}) \leqslant O(\frac{r\log n}{C_{\min} n})$ with probability at least $1 - O(n^{-100})$. □

We now control the error term which arises both from the variance in the random design matrix $\mathbf{X}$ and the bias due to mismatch between $\hat{\mathbf{B}}$ and $\mathbf{B}$.

Lemma 2.18. *Let the sequence of n i.i.d. covariates $\mathbf{x}_i$ satisfy the design assumptions in Assumption 2.3, and assume $\sin(\hat{\mathbf{B}}, \mathbf{B}) \leqslant \delta \leqslant 1$. Then if $n \gtrsim C_{cond}^2 r\log n$,*

$$\|\hat{\mathbf{B}}(\hat{\mathbf{B}}^\top \mathbf{X}^\top \mathbf{X}\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}\mathbf{X}^\top \mathbf{X}\mathbf{B}\boldsymbol{\alpha}_0 - \mathbf{B}\boldsymbol{\alpha}_0\|^2 \leqslant O(\|\boldsymbol{\alpha}_0\|^2 C_{cond}^2 \delta^2),$$

with probability at least $1 - O(n^{-100})$.

Proof. To control this term we first insert a copy of the identity $\mathbf{I}_d = \hat{\mathbf{B}}\hat{\mathbf{B}}^\top + \hat{\mathbf{B}}_\perp \hat{\mathbf{B}}_\perp^\top$ to allow the variance term in the design cancel appropriately in the span of $\hat{\mathbf{B}}$; formally,

$$\hat{\mathbf{B}}(\hat{\mathbf{B}}^\top \mathbf{X}^\top \mathbf{X}\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}\mathbf{X}^\top \mathbf{X}\mathbf{B}\boldsymbol{\alpha}_0 - \mathbf{B}\boldsymbol{\alpha}_0 =$$

$$\hat{\mathbf{B}}(\hat{\mathbf{B}}^\top \mathbf{X}^\top \mathbf{X}\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}\mathbf{X}^\top \mathbf{X}(\hat{\mathbf{B}}\hat{\mathbf{B}}^\top + \hat{\mathbf{B}}_\perp \hat{\mathbf{B}}_\perp^\top)\mathbf{B}\boldsymbol{\alpha}_0 - \mathbf{B}\boldsymbol{\alpha}_0 =$$
$$(\hat{\mathbf{B}}\hat{\mathbf{B}}^\top - \mathbf{I})\mathbf{B}\boldsymbol{\alpha}_0 + \hat{\mathbf{B}}(\hat{\mathbf{B}}^\top \mathbf{X}^\top \mathbf{X}\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}\mathbf{X}^\top \mathbf{X}\hat{\mathbf{B}}_\perp \hat{\mathbf{B}}_\perp^\top \mathbf{B}\boldsymbol{\alpha}_0 =$$
$$\hat{\mathbf{B}}_\perp \hat{\mathbf{B}}_\perp^\top \mathbf{B}\boldsymbol{\alpha}_0 + \hat{\mathbf{B}}(\hat{\mathbf{B}}^\top \mathbf{X}^\top \mathbf{X}\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}\mathbf{X}^\top \mathbf{X}\hat{\mathbf{B}}_\perp \hat{\mathbf{B}}_\perp^\top \mathbf{B}\boldsymbol{\alpha}_0 \implies$$
$$\|\hat{\mathbf{B}}(\hat{\mathbf{B}}^\top \mathbf{X}^\top \mathbf{X}\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}\mathbf{X}^\top \mathbf{X}\mathbf{B}\boldsymbol{\alpha}_0 - \mathbf{B}\boldsymbol{\alpha}_0\|^2 \leqslant$$
$$2(\|\boldsymbol{\alpha}_0\|^2\delta^2 + \|(\hat{\mathbf{B}}^\top \mathbf{X}^\top \mathbf{X}\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}\mathbf{X}^\top \mathbf{X}\hat{\mathbf{B}}_\perp\|^2\delta^2\|\boldsymbol{\alpha}_0\|^2).$$

We now turn to bounding the second error term, $\|(\hat{\mathbf{B}}^\top \mathbf{X}^\top \mathbf{X}\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}^\top \mathbf{X}^\top \mathbf{X}\mathbf{B}\|^2$. Let $\mathbf{E}_1 = \hat{\mathbf{B}}^\top \frac{\mathbf{X}^\top \mathbf{X}}{n}\hat{\mathbf{B}} - \hat{\mathbf{B}}^\top \boldsymbol{\Sigma}\hat{\mathbf{B}}$ and $\mathbf{E}_2 = \hat{\mathbf{B}}^\top \frac{\mathbf{X}^\top \mathbf{X}}{n}\mathbf{B} - \hat{\mathbf{B}}\boldsymbol{\Sigma}\mathbf{B}$. Applying Lemma 2.19 to the matrix $\mathbf{E}_1$ with $\delta = n^{-200}$ and assuming $n \gtrsim C_{\text{cond}}^2 r\log(1/\delta) \gtrsim C_{\text{cond}}^2 r\log n$ shows that $\|(\hat{\mathbf{B}}^\top \boldsymbol{\Sigma}\hat{\mathbf{B}})^{-1}\mathbf{E}_1\| \leqslant \frac{1}{4}$ and $\|\mathbf{E}_1\| \leqslant O(C_{\max}\sqrt{\frac{r\log n}{n}})$ with probability at least $1 - O(n^{-100})$. A further application of Lemma 2.24 shows that $(\hat{\mathbf{B}}^\top \frac{\mathbf{X}^\top \mathbf{X}}{n}\hat{\mathbf{B}})^{-1} = (\mathbf{E}_1 + \hat{\mathbf{B}}^\top \boldsymbol{\Sigma}\hat{\mathbf{B}})^{-1} = (\hat{\mathbf{B}}^\top \boldsymbol{\Sigma}\hat{\mathbf{B}})^{-1} + \mathbf{F}$, where $\|\mathbf{F}\| \leqslant \frac{4}{3}\|(\hat{\mathbf{B}}^\top \boldsymbol{\Sigma}\hat{\mathbf{B}})^{-1}\|\|\mathbf{E}_1(\hat{\mathbf{B}}^\top \boldsymbol{\Sigma}\hat{\mathbf{B}})^{-1}\|$ on this event. Similarly, defining $\hat{\mathbf{B}}^\top \frac{\mathbf{X}^\top \mathbf{X}}{n}\mathbf{B} = \mathbf{E}_2 + \hat{\mathbf{B}}^\top \boldsymbol{\Sigma}\mathbf{B}$ and applying Lemma 2.19 again but to the matrix $\mathbf{E}_2$ with $\delta = n^{-200}$ and assuming $n \gtrsim C_{\text{cond}}^2 r\log(1/\delta) \gtrsim C_{\text{cond}}^2 r\log n$, guarantees that $\|\mathbf{E}_2\| \leqslant O(C_{\max}(\sqrt{\frac{r\log n}{n}})$ with probability at least $1 - O(n^{-100})$.

Hence on the intersection of these two events,

$$\|(\hat{\mathbf{B}}^\top \frac{\mathbf{X}^\top \mathbf{X}}{n}\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}^\top \frac{\mathbf{X}^\top \mathbf{X}}{n}\mathbf{B}\| = \|((\hat{\mathbf{B}}^\top \boldsymbol{\Sigma}\hat{\mathbf{B}})^{-1} + \mathbf{F})(\hat{\mathbf{B}}^\top \boldsymbol{\Sigma}\mathbf{B} + \mathbf{E}_2)\| \leqslant$$
$$\|(\hat{\mathbf{B}}^\top \boldsymbol{\Sigma}\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}^\top \boldsymbol{\Sigma}\mathbf{B}\| + \|(\hat{\mathbf{B}}^\top \boldsymbol{\Sigma}\hat{\mathbf{B}})^{-1}\mathbf{E}_2\| + \|\mathbf{F}\hat{\mathbf{B}}^\top \boldsymbol{\Sigma}\mathbf{B}\| + \|\mathbf{E}_2\mathbf{F}\| \leqslant$$
$$C_{\text{cond}} + O(C_{\text{cond}}\sqrt{\frac{r\log n}{n}}) + O(C_{\text{cond}}^2\sqrt{\frac{r\log n}{n}}) + O(C_{\text{cond}}^2\frac{r\log n}{n}) \leqslant$$
$$C_{\text{cond}} + O(C_{\text{cond}}^2\sqrt{\frac{r\log n}{n}}) = O(C_{\text{cond}}),$$

under the condition $n \gtrsim C_{\text{cond}}^2 \frac{r\log n}{n}$. Taking a union bound over the aforementioned events and combining terms gives the result. □

Finally we present a concentration result for random matrices showing concentration when the matrices are projected along two (potentially different) subspaces.

Lemma 2.19. *Suppose a sequence of i.i.d. covariates $\mathbf{x}_i$ satisfy the design assumptions in Assumption 2.3. Then, if $\mathbf{A}$ and $\mathbf{B}$ are both rank r orthonormal projection matrices,*

$$\|(\mathbf{A}^\top \frac{\mathbf{X}^\top \mathbf{X}}{n}\mathbf{B}) - \mathbf{A}^\top \boldsymbol{\Sigma}\mathbf{B})\| \leqslant O(C_{\max}(\sqrt{\frac{r}{n}} + \frac{r}{n} + \sqrt{\frac{\log(1/\delta)}{n}} + \frac{\log(1/\delta)}{n})),$$

with probability at least $1 - \delta$.

Proof. The result follows by a standard sub-exponential tail bound and covering argument. First note that for any fixed $\mathbf{u}, \mathbf{v} \in \mathbb{S}^{r-1}$, we have that $\mathbf{u}^\top \mathbf{A}^\top \mathbf{x}_i$ and $\mathbf{v}^\top \mathbf{B}^\top \mathbf{x}_i$ are both $\mathrm{sG}(\sqrt{C_{\max}})$. Hence for any fixed $\mathbf{u}, \mathbf{v} \in \mathbb{S}^{r-1}$, $\mathbf{u}^\top \mathbf{A}^\top \mathbf{x}_i \mathbf{x}_i^\top \mathbf{B}\mathbf{v} - \mathbf{u}^\top \mathbf{A}^\top \mathbf{\Sigma} \mathbf{B}\mathbf{v}$ is $\mathrm{sE}(8C_{\max}, 8C_{\max})$.

Now, let S denote a ϵ-cover of $\mathbb{S}^{r-1}$ which has cardinality at most $(\frac{3}{\epsilon})^r$ by a volume-covering argument. Hence for $\epsilon = \frac{1}{5}$,

$$\mathbf{Pr}[\sup_{\mathbf{u} \in S^{r-1}, \mathbf{v} \in S^{r-1}} \mathbf{u}^\top \mathbf{A}^\top \mathbf{x}_i \mathbf{x}_i^\top \mathbf{B}\mathbf{v} - \mathbf{u}^\top \mathbf{A}^\top \mathbf{\Sigma} \mathbf{B}\mathbf{v} \geqslant t] \leqslant 225^r \exp(-cn \min(t^2/C_{\max}^2, t/C_{\max}))),$$

using a union bound over the covers and a sub-exponential tail bound. Taking $t = C \cdot C_{\max}(\sqrt{\frac{r}{n}} + \frac{r}{n} + \sqrt{\frac{\log(1/\delta)}{n}} + \frac{\log(1/\delta)}{n})$ for sufficiently large C, shows that

$$225^r \exp(-cn \min(t^2/C_{\max}^2, t/C_{\max}))) \leqslant \delta$$

. Finally a standard Lipschitz continuity argument yields

$$\begin{aligned} &\sup_{\mathbf{u} \in \mathbb{S}^{r-1}, \mathbf{v} \in \mathbb{S}^{r-1}} \mathbf{u}^\top \frac{1}{n} \sum_{i=1}^n \mathbf{u}^\top \mathbf{A}^\top \mathbf{x}_i \mathbf{x}_i^\top \mathbf{B}\mathbf{v} - \mathbf{u}^\top \mathbf{A}^\top \mathbf{\Sigma} \mathbf{B}\mathbf{v} \\ &\leqslant \frac{1}{1-3\epsilon} \sup_{\mathbf{u} \in S^{r-1}, \mathbf{v} \in S^{r-1}} \mathbf{u}^\top \mathbf{A}^\top \mathbf{x}_i \mathbf{x}_i^\top \mathbf{B}\mathbf{v} - \mathbf{u}^\top \mathbf{A}^\top \mathbf{\Sigma} \mathbf{B}\mathbf{v}, \end{aligned}$$

which gives the result. □

2.12 Proofs for Section 2.5

We begin by presenting the proof of the main statistical lower bound for recovering the feature matrix $\mathbf{B}$ and relevant auxiliary results. Following this we provide relevant background on Grassmann manifolds.

As mentioned in the main text our main tool is to use is a non-standard variant of the Fano method, along with suitable bounds on the cardinality of the packing number and the distributional covering number, to obtain minimax lower bound on the difficulty of estimating $\mathbf{B}$. We instantiate the f-divergence based lower bound below (which we instantiate with χ^2-divergence). We restate this result for convenience.

Lemma 2.20. *[49, Theorem 4.1] For any increasing function $\ell : [0, \infty) \to [0, \infty)$,*

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbf{Pr}_\theta[\ell(\rho(\hat{\theta}, \theta)) \geqslant \ell(\eta/2)] \geqslant \sup_{\eta > 0, \epsilon > 0} \left\{ 1 - \left(\frac{1}{N(\eta)} + \sqrt{\frac{(1+\epsilon^2) M_C(\epsilon, \Theta)}{N(\eta)}} \right) \right\}.$$

In the context of the previous result $N(\eta)$ denotes a lower bound on the η-packing number of the metric space (Θ, ρ). Moreover, $M_C(\epsilon, \Theta)$ is a positive real number for which there

exists a set G with cardinality $\leqslant M_C(\epsilon, S)$ and probability measures Q_α, $\alpha \in G$ such that $\sup_{\theta\in S} \min_{\alpha\in G} \chi^2(\mathbf{Pr}_\theta, Q_\alpha) \leqslant \epsilon^2$, where χ^2 denotes the chi-squared divergence. In words, $M_C(\epsilon, S)$ is an upper bound on the ϵ-covering on the space $\{\mathbf{Pr}_\theta : \theta \in S\}$ when distances are measured by the square root of the χ^2-divergence.

We obtain the other term capturing the difficulty of estimating $\hat{\mathbf{B}}$ with respect to the task diversity minimum eigenvalue by constructing a pair of feature matrices which are hard to distinguish for a particular ill-conditioned task matrix $\mathbf{A}$. Using these two results we provide the proof of our main lower bound.

Proof of Theorem 2.5. For our present purposes we simply take $\ell(\cdot)$ to be the identity function in our application of Lemma 2.20 as we obtain the second term in the lower bound. Then by the duality between packing and covering numbers we have that $\log N \geqslant \log M$ at the same scale (see for example [113, Lemma 5.5]), so once again by Proposition 2.9 we have that $\log N(\eta) \geqslant r(d-r)\log(\frac{c_1}{\eta})$. Then applying Lemma 2.21 we have that $M_C(\epsilon, \Theta') \leqslant (\frac{c_2 n}{\log(1+\epsilon^2)})^{r(d-r)/2}$. For convenience we set $k = r(d-r)$ in the following. We now choose the pair η and ϵ appropriately in Lemma 2.20. The lower bound writes as,

$$1 - \left(\frac{1}{N(\eta)} + \sqrt{\frac{(1+\epsilon^2)M_C(\epsilon,\Theta)}{N(\eta)}}\right) \geqslant 1 - \left((\frac{\eta}{c_2})^k + (\frac{\eta}{c_2})^{k/2} \cdot (c_1 n)^{k/4} \frac{(1+\epsilon^2)^{1/2}}{\log(1+\epsilon^2)^{k/4}}\right),$$

with the implicit constraint that $\epsilon' = \sqrt{\frac{2}{n}\log(1+\epsilon^2)} < 1$. A simple calculus argument shows that $\epsilon \to \sqrt{1+\epsilon^2}/(\log(1+\epsilon^2))^{k/4}$ is minimized when $1+\epsilon^2 = e^{k/2}$ (subject to $\sqrt{\frac{1}{2n}\log(1+\epsilon^2)} < 1$). This constraint can always be ensured by taking $n > \frac{k}{4}$. We then have that the lower bound becomes

$$1 - \left((\frac{\eta}{c_2})^k + (2e\frac{c_1}{c_2^2}\eta^2 n/k)^{k/4}\right). \tag{2.21}$$

We now take $\eta = C\sqrt{k/n}$ so the bound simplifies to $1 - \left((\frac{\eta}{c_2})^k + (2e\frac{c_1}{c_2^2}\eta^2 n/k)^{k/4}\right) = 1 - \left((\frac{C}{c_2}\sqrt{k/n})^k + (2e\frac{c_1}{c_2^2}C^2)^{k/4}\right)$. By choosing C to be sufficiently small and taking $n > k$ we ensure that $1 - \left((\frac{C}{c_2}\sqrt{k/n})^k + (2e\frac{c_1}{c_2^2}C^2)^{k/4}\right) \geqslant \frac{99}{100}$. Finally, under the condition $r \leqslant \frac{d}{2}$ we have that $k \geqslant \frac{dr}{2}$. Combining with Lemma 2.20 gives the result for the second term.

The first term is lower bounded using the LeCam two-point method in an indepedent fashion as a consequence of Lemma 2.23. A union bound over the events on which the lower bounds hold give the result. Note that a single choice of $\mathbf{A}$ matrix can in fact be used for both lower bounds by simply opting for the choice of $\mathbf{A}$ used in Lemma 2.23.

□

To implement the lower bound we require an upper bound on the covering number in the space of distributions of $\mathbf{Pr}_\mathbf{B}$. Throughout we use standard properties of the χ^2-divergence which can be found in [108, Section 2.4].

Lemma 2.21. *Suppose n data points, $(\mathbf{x}_i, y_i)$, are generated from the model in (2.1) with i.i.d. covariates $\mathbf{x}_i \sim \mathcal{N}(0, \mathbf{I}_d)$ and independent i.i.d. noise $\epsilon_i \sim \mathcal{N}(0,1)$. Further, assume the task parameters satisfy Assumption 2.2 with each task normalized to $\|\boldsymbol{\alpha}_i\| = \frac{1}{2}$. Then if $r \leqslant \frac{d}{2}$,*

$$M_C(\epsilon, \Theta') \leqslant \left(\frac{cn}{\log(1+\epsilon^2)}\right)^{r(d-r)/2},$$

whenever $\sqrt{\frac{1}{2n}\log(1+\epsilon^2)} < 1$.

Proof. We first upper bound the χ^2 divergence between two data distributions for two distinct $\mathbf{B}^i$ and $\mathbf{B}^j$. Now the joint distribution over the observations for each measure $\mathbf{Pr}_{\mathbf{B}^i}$ can be written as $\mathbf{Pr}_{\mathbf{B}^i} \equiv \Pi_{k=1}^t p(\mathbf{X}_k)p(\mathbf{y}|\mathbf{X}_k, \mathbf{B}^i, \boldsymbol{\alpha}_j)$, where $p(\mathbf{X}_k)$ corresponds to the density of the Gaussian design matrix, and $p(\mathbf{y}|\mathbf{X}_k, \mathbf{B}^i, \boldsymbol{\alpha}_k)$ the Gaussian conditionals of the observations $\mathbf{y}$. So using standard properties of the χ^2-divergence we find that,

$$\chi^2(\mathbf{Pr}_{\mathbf{B}^i}, \mathbf{Pr}_{\mathbf{B}^j}) = \Pi_{k=1}^t(1 + \mathbb{E}_{\mathbf{X}_k}[\chi^2(p(\mathbf{y}|\mathbf{X}_k, \mathbf{B}^i), p(\mathbf{y}|\mathbf{X}_k, \mathbf{B}^j))]) - 1 = \Pi_{k=1}^t \mathbb{E}_{\mathbf{X}_k}\left[\exp\left(\|\mathbf{X}_k(\mathbf{B}^i\boldsymbol{\alpha}_k - \mathbf{B}^j\boldsymbol{\alpha}_k)\|^2\right)\right] - 1.$$

Now note that $\|(\mathbf{B}^i\boldsymbol{\alpha}_k - \mathbf{B}^j\boldsymbol{\alpha}_k)\|^2 \leqslant 2\|\boldsymbol{\alpha}_k\|^2\sigma_1(\mathbf{I}_r - (\mathbf{B}^i)^\top\mathbf{B}^j)$. Recognizing $\sigma_r((\mathbf{B}^i)^\top\mathbf{B}^j) = \cos\theta_1(\mathbf{B}^i, \mathbf{B}^j)$ and $\|(\mathbf{B}^i_\perp)^\top\mathbf{B}^j\| = \sin\theta_1$, where θ_1 is largest principal angle between the subspaces, we have that $1 - \sigma_r((\mathbf{B}^i)^\top\mathbf{B}^j) = 1 - \sqrt{1 - \|(\mathbf{B}^i_\perp)^\top\mathbf{B}^j\|^2} \leqslant \|(\mathbf{B}^i_\perp)^\top\mathbf{B}^j\|^2 \leqslant 1$, using the elementary inequality $1 - \sqrt{1-x^2} \leqslant x^2$ for $0 \leqslant x \leqslant 1$. Thus, $\|(\mathbf{B}^i\boldsymbol{\alpha}_t - \mathbf{B}^j\boldsymbol{\alpha}_t)\|^2 \leqslant \frac{1}{2}$.

Now we use the identity that for $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I}_d)$, and $\|\mathbf{v}\| \leqslant \frac{1}{2}$ that $\mathbb{E}_{\mathbf{x}}\exp((\mathbf{v}^\top\mathbf{x})^2) = \mathbb{E}_{\mathbf{x}}[\exp(\|\mathbf{v}\|^2((\frac{\mathbf{v}^\top}{\|\mathbf{v}\|}\mathbf{x})^2 - 1)]\exp(\|\mathbf{v}\|^2) = \frac{\exp(-\|\mathbf{v}\|^2))}{\sqrt{1-2\|\mathbf{v}\|^2}} \cdot \exp(\|\mathbf{v}\|^2) \leqslant \exp(2\|\mathbf{v}\|^4) \cdot \exp(\|\mathbf{v}\|^2) \leqslant \exp(2\|\mathbf{v}\|^2)$. Hence combining the above two facts we obtain,

$$\chi^2(\mathbf{Pr}_{\mathbf{B}^i}, \mathbf{Pr}_{\mathbf{B}^j}) \leqslant \exp\left(2\sum_{k=1}^t n_t\|\mathbf{B}^i\boldsymbol{\alpha}_k - \mathbf{B}^j\boldsymbol{\alpha}_k\|^2\right) - 1 \leqslant \exp(2n \cdot \sin^2\theta(\mathbf{B}^i, \mathbf{B}^j)) - 1$$

applying Lemma 2.22 in the inequality with a rescaling. Hence to ensure that $\chi^2(\mathbf{Pr}_{\mathbf{B}^i,\mathbf{B}^j}) \leqslant \epsilon^2$ we take $\mathbf{B}^i$ to be the closest element in a ϵ' cover, S, of $\mathrm{Gr}_{r,d}(\mathbb{R})$ to $\mathbf{B}^j$. Since further $\chi^2(\mathbf{Pr}_{\mathbf{B}^i}, \mathbf{Pr}_{\mathbf{B}^j}) \leqslant \exp(2n(\epsilon')^2) - 1$ this is satisfied by taking $\epsilon' = \sqrt{\frac{1}{2n}\log(1+\epsilon^2)}$ (with the constraint we have $\epsilon' < 1$). Using Proposition 2.9, we then obtain that,

$$M_C(\epsilon, \Theta') \leqslant \left(\frac{cn}{\log(1+\epsilon^2)}\right)^{r(d-r)/2},$$

for a universal constant c. □

Lemma 2.22. *Let the task parameters $\boldsymbol{\alpha}_i$ each satisfy $\|\boldsymbol{\alpha}_i\| = 1$ with parameter $\nu = \sigma_r(\frac{\mathbf{A}^\top \mathbf{A}}{t}) > 0$, and let $\mathbf{B}^i$ and $\mathbf{B}^j$ be distinct, rank-r orthonormal feature matrices. Then,*

$$\sum_{k=1}^{t} n_t \|\mathbf{B}^i \boldsymbol{\alpha}_k - \mathbf{B}^j \boldsymbol{\alpha}_k\|^2 \leqslant n \sin^2 \theta(\mathbf{B}^i, \mathbf{B}^j).$$

where $n = t \cdot n_t$.

Proof. We can simplify the expression as follows,

$$\sum_{k=1}^{t} n_t \|\mathbf{B}^i \boldsymbol{\alpha}_k - \mathbf{B}^j \boldsymbol{\alpha}_k\|^2 = n \cdot \frac{1}{t} \sum_{j=1}^{t} 2 \cdot (\boldsymbol{\alpha}_k^\top \boldsymbol{\alpha}_k - \boldsymbol{\alpha}_k^\top (\mathbf{B}^i)^\top \mathbf{B}^j \boldsymbol{\alpha}_k) =$$
$$n \cdot \operatorname{tr}\left(\left(\mathbf{I}_r - (\mathbf{B}^i)^\top \mathbf{B}^j \right) \frac{\mathbf{A}^\top \mathbf{A}}{t} \right)$$

using the fact that $\mathbf{B}^\top \mathbf{B} = \mathbf{I}_r$ for an orthonormal feature matrix, the normalization of $\|\boldsymbol{\alpha}_k\| = 1$, and the cyclic property of the trace. Now use the fact that $\operatorname{tr}\left(\left(\mathbf{I}_r - (\mathbf{B}^i)^\top \mathbf{B}^j \right) \frac{\mathbf{A}^\top \mathbf{A}}{t} \right)) \leqslant \sigma_{\max}(\mathbf{I}_r - (\mathbf{B}^i)^\top \mathbf{B}^j)) \cdot \operatorname{tr}(\frac{\mathbf{A}^\top \mathbf{A}}{t}) \leqslant \frac{1}{4}(1 - \cos\theta_1) = \frac{1}{4}(1 - \sqrt{1 - \sin^2\theta_1}) \leqslant \frac{1}{4}\sin^2\theta_1$, using the elementary inequality $1 - \sqrt{1 - x^2} \leqslant x^2$ for $0 \leqslant x \leqslant 1$. Note that $\operatorname{tr}(\mathbf{A}^\top \mathbf{A}/t) \leqslant \frac{1}{4}$ follows from the normalization of the $\boldsymbol{\alpha}_j$. □

We now provide brief background on $\mathrm{Gr}_{r,d}(\mathbb{R})$ and establish several pieces of notation relevant to the discussion. We denote the Grassmann manifold, which consists of the the set of r-dimensional subspaces within the underlying d-dimensional space, as $\mathrm{Gr}_{r,d}(\mathbb{R})$. Another way to define it is as the homogeneous space of the orthogonal group $O(d)$ in the sense that,

$$\mathrm{Gr}_{r,d}(\mathbb{R}) \cong O(d)/(O(r) \times O(d-r)),$$

which defines its geometric structure. The underlying measure on the manifold $G_{r,n}(\mathbb{R})$ is the associated, normalized invariant (or Haar) measure.

Note that each orthonormal feature matrix $\mathbf{B}$, is contained in an equivalence class (under orthogonal rotation) of an element in $\mathrm{Gr}_{r,d}(\mathbb{R})$. To define distance in $\mathrm{Gr}_{r,d}(\mathbb{R})$ we define the notion of a principal angle between two subspaces p and q. If $\mathbf{C}$ is an orthonormal matrix whose columns form an orthonormal basis of p and $\mathbf{D}$ is an orthonormal matrix whose columns form an orthonormal basis of q, then the singular values of the decomposition of $\mathbf{C}^\top \mathbf{D} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$ defines the principal angles as follows:

$$\mathbf{D} = \operatorname{diag}(\cos\theta_1, \cos\theta_2, \ldots, \cos\theta_k),$$

where $0 \leqslant \theta_k \leqslant \ldots \leqslant \theta_1 \leqslant \frac{\pi}{2}$. As shorthand we let $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$, and let sin and cos act element-wise on its components. The subspace angle distance which is induced by ℓ_∞ norms on the vector $\sin\theta$. We refer the reader to [94] for geometric background on coding and

packing/covering bounds in the context of Grassmann manifolds relevant to our discussion here.

In the following we let $M(\mathrm{Gr}_{r,d}(\mathbb{R}), \sin\theta_1, \eta)$ denote the η-covering number of $\mathrm{Gr}_{r,d}(\mathbb{R})$ in the subspace angle distance.

Proposition 2.9. *[94, Adapted from Proposition 8] For any integers $1 \leqslant r \leqslant \frac{d}{2}$ and every $\epsilon > 0$, we have that,*

$$r(d-r)\log(\frac{c_1}{\eta}) \leqslant \log M(Gr_{r,d}(\mathbb{R}), \sin\theta_1, \eta) \leqslant r(d-r)\log(\frac{c_2}{\eta}),$$

for universal constants $c_1, c_2 > 0$.

Proof. Define for a linear operator T, $\sigma_q(T) = (\sum_{i\geqslant 1} |s_i(T)|^p)^{1/p}$ for all $1 \leqslant q \leqslant \infty$ where $s_i(T)$ denotes its ith singular value. Note that Proposition 8 in [94] states the result in the distance metric $d(E,F) = \sigma_q(P_E - P_F)$ where P_E and P_F denotes the projection operator onto the subspace E and F respectively, and $\sigma_q(T) = (\sum_{i\geqslant 1} |s_i(T)|^p)^{1/p}$. However, as the computation in Proposition 6 of [94] establishes, we have that $\sigma_q(P_E - P_F) = (2\sum_{i=1}^{r}(1-\cos^2\theta_i)^{q/2})^{1/q}$; taking $q \to \infty$ implies $\sigma_q(P_E - P_F) = \sin\theta_1$, and hence directly translating the result gives the statement of the proposition. □

Finally, we include the proof of the lower bound which captures the dependence on the task diversity parameter. The proof uses the LeCam two-point method between two problem instances which are difficult to distinguish for a particular, ill-conditioned task matrix.

Lemma 2.23. *Under the conditions of Theorem 2.5, for $n \geqslant \frac{1}{8\nu}$,*

$$\inf_{\hat{\mathbf{B}}} \sup_{\mathbf{B}\in Gr_{r,d}(\mathbb{R})} \sin\theta(\hat{\mathbf{B}}, \mathbf{B}) \geqslant \Omega\left(\sqrt{\frac{1}{\nu}}\sqrt{\frac{1}{n}}\right)$$

with probability at least $\frac{3}{10}$.

Proof. First consider an ill-conditioned task matrix where the sequence of $\boldsymbol{\alpha}_i = \frac{1}{2}\mathbf{e}_i$ for $i \in [r-1]$ but then $\boldsymbol{\alpha}_r = \frac{1}{2}(\sqrt{1-b^2}\mathbf{e}_{r-1} + b\mathbf{e}_r)$ for $0 < b < 1$ where $\mathbf{e}_i$ are the standard basis in $\mathbb{R}^r$. Now, consider two task models for two different subspaces $\mathbf{B}_1$ and $\mathbf{B}_2$ which are distinct in a single direction. Namely we take $\mathbf{B}_1 = [\mathbf{e}_1, \ldots, \mathbf{e}_r]$ and $\mathbf{B}_2 = [\mathbf{e}_1, \ldots, \mathbf{e}_{r-1}, \sqrt{1-a^2}\mathbf{e}_r + a\mathbf{e}_{r+1}]$, for $0 < a < 1$, where $\mathbf{e}_i$ refer to the standard basis in $\mathbb{R}^d$. Here we have $\cos\theta_1 = \|\mathbf{B}_2^\top \mathbf{B}_1\| = \sqrt{1-a^2} \implies \sin\theta_1 = a$, where θ_1 refers to the largest principle angle between the two subspaces.

Data is generated from the two linear models as,

$$y_i = \mathbf{x}_i^\top \mathbf{B}_1 \boldsymbol{\alpha}_j + \epsilon_i \; i = 1, \ldots, n$$
$$y_i = \mathbf{x}_i^\top \mathbf{B}_2 \boldsymbol{\alpha}_j + \epsilon_i \; i = 1, \ldots, n$$

with n total samples generated evenly from each of j in $[t]$ tasks (n_t from each task) inducing two measures $\mathbf{Pr}_1$ and $\mathbf{Pr}_2$ over their respective data. The LeCam two-point method (see [113, Ch. 15] for example) shows that,

$$\inf_{\hat{\mathbf{B}}} \sup_{\mathbf{B}\in \mathrm{Gr}_{r,d}(\mathbb{R})} \mathbf{Pr}_{\mathbf{B}}[\sin\theta(\hat{\mathbf{B}}, \mathbf{B}) \geqslant a] \geqslant \frac{1}{2}(1 - \|\mathbf{Pr}_1 - \mathbf{Pr}_2\|_{\mathrm{TV}}) \tag{2.22}$$

for the $\mathbf{B}_1$ and $\mathbf{B}_2$ of our choosing as above.

We can now upper bound the total variation distance (via the Pinsker inequality) similar to as in Lemma 2.21,

$$\|\mathbf{Pr}_1 - \mathbf{Pr}_2\|_{\mathrm{TV}}^2 \leqslant \frac{1}{2}\mathrm{KL}(\mathbf{Pr}_1|\mathbf{Pr}_2) = \frac{1}{4}\sum_{j=1}^{t}\sum_{i=1}^{n_t} \|\mathbf{B}_1\boldsymbol{\alpha}_j - \mathbf{B}_2\boldsymbol{\alpha}_j\|_2^2 = \tag{2.23}$$

$$\frac{1}{2}n_t\sum_{j=1}^{t}(\|\boldsymbol{\alpha}_j\|^2 - \boldsymbol{\alpha}_j^\top\mathbf{B}_1^\top\mathbf{B}_2\boldsymbol{\alpha}_j) = \frac{n}{2}\cdot(\frac{1}{4} - \mathrm{tr}(\mathbf{B}_1^\top\mathbf{B}_2\mathbf{C})) \tag{2.24}$$

using cyclicity of the trace. Straightforward calculations show that given the $\mathbf{A}$ matrix, $\mathbf{C} = \frac{1}{4r}\begin{bmatrix}\mathbf{I}_{r-2} & \mathbf{0}\\ \mathbf{0} & \mathbf{M}_1\end{bmatrix}$, where $\mathbf{M}_1 = \begin{bmatrix}2-b^2 & b\sqrt{1-b^2}\\ b\sqrt{1-b^2} & b^2\end{bmatrix}$. Similarly $\mathbf{B}_1^\top\mathbf{B}_2 = \begin{bmatrix}\mathbf{I}_{r-2} & \mathbf{0}\\ \mathbf{0} & \mathbf{M}_2\end{bmatrix}$ where $\mathbf{M}_2 = \begin{bmatrix}1 & 0\\ 0 & \sqrt{1-a^2}\end{bmatrix}$. Computing the trace term,

$$\frac{1}{4} - \mathrm{tr}((\mathbf{B}_1^\top\mathbf{B}_2)\mathbf{C}) = \frac{1}{4} - \frac{1}{4}\left(\frac{r-2}{r} + \frac{1}{r}\cdot\mathrm{tr}\Big(\begin{bmatrix}1 & 0\\ 0 & \sqrt{1-a^2}\end{bmatrix}\begin{bmatrix}2-b^2 & b\sqrt{1-b^2}\\ b\sqrt{1-b^2} & b^2\end{bmatrix}\Big)\right) = \tag{2.25}$$

$$\frac{1}{4r}b^2(1-\sqrt{1-a^2}). \tag{2.26}$$

Hence, $\|\mathbf{Pr}_1 - \mathbf{Pr}_2\|_{\mathrm{TV}}^2 \leqslant \frac{n}{2r}b^2(1-\sqrt{1-a^2}) \leqslant \frac{n}{2r}b^2a^2$. Combining with the LeCam two-point lemma shows that,

$$\inf_{\hat{\mathbf{B}}} \sup_{\mathbf{B}\in \mathrm{Gr}_{r,d}(\mathbb{R})} \mathbf{Pr}_{\mathbf{B}}[\sin\theta(\hat{\mathbf{B}}, \mathbf{B}) \geqslant a] \geqslant \frac{1}{2}(1 - \|\mathbf{Pr}_1 - \mathbf{Pr}_2\|_{\mathrm{TV}}) \geqslant \frac{1}{2}(1 - \sqrt{\frac{n}{2r}}ba). \tag{2.27}$$

Taking $a = \frac{1}{2}\sqrt{\frac{r}{n}}\frac{1}{b} < 1$ suffices to ensure the lower bound with probability at least $\frac{3}{10}$. This induces the constraint $\frac{1}{2}\sqrt{\frac{r}{n}}\frac{1}{b} < 1 \implies n > \frac{r}{4b^2}$. As a last remark note that the $\mathbf{C}$ matrix has maximum and minimum eigenvalues $\frac{1}{4r}(1+\sqrt{1-b^2})$ and $\frac{1}{4r}(1-\sqrt{1-b^2})$. So $\nu = \frac{1}{4r}(1-\sqrt{1-b^2}) \implies \sqrt{2}\sqrt{r\nu} \leqslant b \leqslant 2\sqrt{2}\sqrt{r\nu}$ for $0 < b < 1$ using the inequality $x^2/2 \leqslant 1-\sqrt{1-x^2} \leqslant x^2$. Hence it follows $a \geqslant \frac{1}{8}\frac{1}{\sqrt{\nu}}\sqrt{\frac{1}{n}}$ as well. Similarly the constraint can reduce too $n > \frac{1}{8\nu}$. □

2.13 Auxiliary Results

Here we collect several auxiliary results. We begin by stating a simple matrix perturbation result.

Lemma 2.24. *Let* $\mathbf{A}$ *be a positive-definite matrix and* $\mathbf{E}$ *another matrix which satisfies* $\|\mathbf{E}\mathbf{A}^{-1}\| \leqslant \frac{1}{4}$*, then,*

$$(\mathbf{A}+\mathbf{E})^{-1} = \mathbf{A}^{-1} + \mathbf{F},$$

where $\|\mathbf{F}\| \leqslant \frac{4}{3}\|\mathbf{A}^{-1}\|\|\mathbf{E}\mathbf{A}^{-1}\|$.

Proof.

$$(\mathbf{A}+\mathbf{E})^{-1} = \mathbf{A}^{-1}(\mathbf{I}+\mathbf{E}\mathbf{A}^{-1})^{-1}.$$

Under the condition, $\|\mathbf{E}\mathbf{A}^{-1}\| \leqslant \frac{1}{4}$, $\mathbf{I}+\mathbf{E}\mathbf{A}^{-1}$ is invertible and has a convergent power series expansion so

$$(\mathbf{A}+\mathbf{E})^{-1} = \mathbf{A}^{-1}(\mathbf{I}-\mathbf{E}\mathbf{A}^{-1}+(\mathbf{E}\mathbf{A}^{-1})^2+\ldots) = \mathbf{A}^{-1}+\mathbf{F},$$

where $\mathbf{F} = \mathbf{A}^{-1}\cdot(-\mathbf{E}\mathbf{A}^{-1}+(\mathbf{E}\mathbf{A}^{-1})^2+\ldots)$. Moreover,

$$\|\mathbf{F}\| \leqslant \|\mathbf{A}^{-1}\|(\|\mathbf{E}\mathbf{A}^{-1}\|+\|\mathbf{E}\mathbf{A}^{-1}\|^2+\ldots) \leqslant \frac{\|\mathbf{A}^{-1}\|\|\mathbf{E}\mathbf{A}^{-1}\|}{1-\|\mathbf{E}\mathbf{A}^{-1}\|} \leqslant \frac{4}{3}\|\mathbf{A}^{-1}\|\|\mathbf{E}\mathbf{A}^{-1}\|.$$

□

We now present several results related to the concentration of measure.

Lemma 2.25. *Let* x, y *be mean-zero random variables that are both sub-gaussian with parameters* κ_1 *and* κ_2 *respectively. Then* $z = xy - \mathbb{E}[xy] \sim sE(8\kappa_1\kappa_2, 8\kappa_1\kappa_2)$.

The proof is a standard argument and omitted. Next we prove a matrix concentration result for the individually rescaled covariance matrices of i.i.d. random variables. The proof uses a standard covering argument.

Lemma 2.26. *Let* $\mathbf{X} \in \mathbb{R}^{n\times d}$ *be a random matrix with rows* $a_i\mathbf{x}_i$*, where* $\mathbf{x}_i$ *are i.i.d. random vectors satisfying the design conditions in Assumption 2.1. Then,*

$$\|\frac{1}{n}\mathbf{X}^\top\mathbf{X} - \frac{1}{n}\sum_{i=1}^n a_i^2\mathbf{\Sigma}\| \leqslant \|\mathbf{\Sigma}\|K^2\max(\delta,\delta^2) \quad \text{for } \delta = C(\sqrt{d/n}+t/\sqrt{n}),$$

with probability at least $1-2\exp(-t^2)$*. Here* C *denotes a universal constant and* $K = \max_i|a_i|$.

Proof. First note that we bring all the vectors to isotropic position by rotating so that $\|\frac{1}{n}\mathbf{X}^\top\mathbf{X} - \frac{1}{n}\sum_{i=1}^n a_i^2\mathbf{\Sigma}\| \leqslant \|\mathbf{\Sigma}\|\|\frac{1}{n}\sum_{i=1}^n a_i^2(\mathbf{\Sigma}^{-1/2}\mathbf{x}_i)(\mathbf{\Sigma}^{-1/2}\mathbf{x}_i)^\top - \frac{1}{n}\sum_{i=1}^n a_i^2\mathbf{I}_d\|$. Now by definition for any fixed $\mathbf{v} \in \mathbb{R}^d : \|\mathbf{v}\| = 1$, each $\mathbf{v}^\top\mathbf{\Sigma}^{-1/2}\mathbf{x}_i$ is sG(1) and hence $a_i^2(\mathbf{v}^\top\mathbf{\Sigma}^{-1/2}\mathbf{x}_i)(\mathbf{\Sigma}^{-1/2}\mathbf{x}_i)^\top\mathbf{v}$ is sE$(8a_i^2, 8a_i^2)$ by Lemma 2.25. For the latter quantity [110, Theorem 4.6.1, Eq. 4.22] proves the result when $a_i = 1$ using a standard covering argument along with a sub-exponential tail bound. A close inspection of the proof of [110, Theorem 4.6.1, Eq. 4.22] shows that the aforementioned analogous statement holds when the sequence of random vectors is scaled by a_i. □

We now include two useful results on operator norm bounds of higher-order matrices. The results only require the condition of $O(1)$-L4-L2 hypercontractivity (which is directly implied by Assumption 2.1—see for example the sub-gaussian moment bounds in [113, Theorem 2.6]). Formally, we say a random vector $\mathbf{x}$ is L-L4-L2 hypercontractive if $\mathbb{E}[\langle\mathbf{v}, \mathbf{x}\rangle^4] \leqslant L^2(\mathbb{E}[\langle\mathbf{v}, \mathbf{x}\rangle^2])^2$ for all unit vectors $\mathbf{v}$. Also note that if $\mathbf{x}$ is hypercontractive this immediately implies that $P_\mathbf{V}\mathbf{x}$ is also hypercontractive with the same constant where $P_\mathbf{V}$ is an orthogonal projection operator.

Lemma 2.27. *Let $\mathbf{x}$ be a mean-zero random vector from a distribution that is L-L4-L2 hypercontractive with covariance $\mathbf{\Sigma}$ and let $P_\mathbf{V}$ be an orthogonal projection operator onto a rank-r subspace. Then*

$$\|\mathbb{E}[\|\mathbf{x}\|^2\mathbf{x}\mathbf{x}^\top]\| \leqslant L\text{tr}(\mathbf{\Sigma})\|\mathbf{\Sigma}\|; \quad \|\mathbb{E}[\|P_\mathbf{V}\mathbf{x}\|^2\mathbf{x}\mathbf{x}^\top]\| \leqslant Lr\|\mathbf{\Sigma}\|^2;$$
$$\|\mathbb{E}[\|\mathbf{x}\|^2P_\mathbf{V}\mathbf{x}(P_\mathbf{V}\mathbf{x})^\top\| \leqslant L\text{tr}(\mathbf{\Sigma})\|\mathbf{\Sigma}\|.$$

Proof of Lemma 2.27. We introduce a vector $\mathbf{v}$ with $\|\mathbf{v}\| \leqslant 1$. Then,

$$\mathbb{E}[\langle\mathbf{v}, \|\mathbf{x}\|^2\mathbf{x}\mathbf{x}^\top\mathbf{v}\rangle] = \mathbb{E}[\|\mathbf{x}\|^2\langle\mathbf{v}, \mathbf{x}\rangle^2] \leqslant (\mathbb{E}[\|\mathbf{x}\|^4])^{1/2}(\mathbb{E}[\langle\mathbf{v}, \mathbf{x}\rangle^4])^{1/2}, \tag{2.28}$$

by the Cauchy-Schwarz inequality. For the first term we have $(\mathbb{E}[\|\mathbf{x}\|^4])^{1/2} \leqslant \sqrt{L}\text{tr}(\mathbf{\Sigma})$ by Lemma 2.28. For the second term once again using L-L4-L2 hypercontractivity we have $(\mathbb{E}[\langle\mathbf{v}, \mathbf{x}\rangle^4])^{1/2} \leqslant \sqrt{L}\mathbb{E}[\langle\mathbf{v}, \mathbf{x}\rangle]^2 \leqslant \sqrt{L}\|\mathbf{\Sigma}\|$. Maximizing over $\mathbf{v}$ gives the result. The remaining statements follow using an identical calculation and appealing to Lemma 2.28. □

Lemma 2.28. *Let $\mathbf{x}$ be a mean-zero random vector from a distribution that is L-L4-L2 hypercontractive with covariance $\mathbf{\Sigma}$ and let $P_\mathbf{V}$ be an orthogonal projection operator onto a rank-r subspace. Then*

$$\mathbb{E}[\|\mathbf{x}\|^4] \leqslant L(\text{tr}\mathbf{\Sigma})^2; \quad \mathbb{E}[\|P_\mathbf{V}\mathbf{x}\|^2] \leqslant r\|\mathbf{\Sigma}\|; \quad \mathbb{E}[\|P_\mathbf{V}\mathbf{x}\|^2\|\mathbf{x}\|^2] \leqslant Lr\|\mathbf{\Sigma}\|(\text{tr}\mathbf{\Sigma});$$
$$\|\mathbb{E}[\|P_\mathbf{V}\mathbf{x}\|^4]\| \leqslant L\|\mathbf{\Sigma}\|^2r^2.$$

Proof of Lemma 2.28. A short computation using the Cauchy-Schwarz inequality and L4-L2 equivalence shows that,

$$\mathbb{E}[\|\mathbf{x}\|^4] = \mathbb{E}[(\sum_{i=1}^d\langle\mathbf{x}, \mathbf{e}_i\rangle^2)^2] = \mathbb{E}[\sum_{a,b}\langle\mathbf{x}, \mathbf{e}_a\rangle^2\langle\mathbf{x}, \mathbf{e}_b\rangle^2] \leqslant \sum_{a,b}(\mathbb{E}[\langle\mathbf{x}, \mathbf{e}_a\rangle^4]\mathbb{E}[\langle\mathbf{x}, \mathbf{e}_b\rangle^4])^{1/2} \leqslant$$

$$L\sum_{a,b}\mathbb{E}[\langle \mathbf{x}, \mathbf{e}_a\rangle^2]\mathbb{E}[\langle \mathbf{x}, \mathbf{e}_b\rangle^2] \leqslant L(\mathrm{tr}(\mathbf{\Sigma}))^2.$$

The second statement follows since $\mathbb{E}[\|P_{\mathbf{V}}\mathbf{x}\|^2] = \mathbb{E}[\mathrm{tr}(P_{\mathbf{V}}\mathbf{x}\mathbf{x}^\top] = \mathrm{tr}(P_{\mathbf{V}}\mathbf{\Sigma}) \leqslant r\|\mathbf{\Sigma}\|$ where the last line follows by the von Neumann trace inequality and the fact $P_{\mathbf{V}}$ is a projection operator. The final statements follow by combining the previous calculations. □

Lemma 2.29. *Let* $\mathbf{x}$ *be a random vector in* d *dimensions from a distribution satisfying Assumption 2.3. Then, we have:*

$$\|\mathbf{x}\| \leqslant O\left(\sqrt{C_{\max}}(\sqrt{d} + \sqrt{\log 1/\delta})\right),$$

with probability at least $1 - \delta$.

Proof of Lemma 2.29. Note that by rotating the vectors into isotropic position $\mathbf{y} = \mathbf{\Sigma}^{-1/2}\mathbf{x}$ is $\mathbf{I}_d$-subgaussian in the sense of Assumption 2.3. Since $\|\mathbf{x}\| \leqslant \sqrt{C_{\max}}\|\mathbf{\Sigma}^{-1/2}\mathbf{x}\|$ it suffices to bound the norm of $\|\mathbf{y}\|$. First note that $\mathbb{E}[\|\mathbf{y}\|] \leqslant \sqrt{\mathbb{E}[\|\mathbf{y}\|^2]} = \sqrt{d}$. Now, take an $1/2$-net over the unit sphere, G; by a standard covering argument the number of elements in G can be upper bounded by 6^d [113, Chapter 5]. By definition of $\mathbf{I}_d$-subgaussianity, for any $\mathbf{v} \in G$, we have that,

$$\mathbf{Pr}[|\langle \mathbf{v}, \mathbf{y}\rangle| \geqslant t] \leqslant 2\exp\left(-\frac{t^2}{2}\right)$$

By a standard continuity argument it follows that $\max_{\mathbf{v}\in S^{d-1}}|\langle \mathbf{v}, \mathbf{y}\rangle| \leqslant 3\max_{\mathbf{v}\in G}|\langle \mathbf{v}, \mathbf{y}\rangle|$. So by a union bound, $\mathbf{Pr}[\|\mathbf{y}\| \geqslant t] \leqslant (12)^d\exp(-t^2/20)$. Therefore, by taking $t = C(\sqrt{d} + \sqrt{\log 1/\delta})$ for large-enough constant C we can ensure that $(12)^d\exp(-t^2/20) \leqslant \delta$. Rearranging gives the conclusion. □

Finally, we prove a truncated version of the matrix Bernstein inequality we can apply to matrices that are unbounded in spectral norm. This is our primary technical tool used to show concentration of the higher-order moments used in the algorithm to recover the feature matrix $\mathbf{B}$.

Lemma 2.30. *Consider a truncation level* $R > 0$. *If* Z_i *is a sequence of symmetric independent random matrices and if* $Z_i' = Z_i\mathbb{1}[\|Z_i\| \leqslant R]$, *then*

$$\mathbf{Pr}[\|\frac{1}{n}\sum_{i=1}^n Z_i - \mathbb{E}[Z_i]\| \geqslant t] \leqslant \mathbf{Pr}[\|\frac{1}{n}\sum_{i=1}^n Z_i' - \mathbb{E}[Z_i']\| \geqslant t - \Delta] + n\mathbf{Pr}[\|Z_i\| \geqslant R],$$

where $\Delta \geqslant \|\mathbb{E}[Z_i] - \mathbb{E}[Z_i']\|$. *Further, for* $t \geqslant \Delta$, *we have that,*

$$\mathbf{Pr}[\|\frac{1}{n}\sum_{i=1}^n Z_i' - \mathbb{E}[Z_i']\| \geqslant t - \Delta] \leqslant 2d\exp\left(\frac{n^2(t-\Delta)^2}{\sigma^2 + 2Rn(t-\Delta)/3}\right),$$

where $\sigma^2 = \|\sum_{i=1}^n \mathbb{E}[(Z_i' - \mathbb{E}[Z_i'])^2]\| \leqslant \|\sum_{i=1}^n \mathbb{E}[Z_i^2]\|$.

Proof. The first statement follows by splitting on the event $\{\|Z_i\| \leqslant R : \forall i \in [n]\}$, along with a union bound, and an application of the triangle inequality to the first term. The second is simply a restatement of the matrix Bernstein inequality in [107] along with the almost sure upper bound $\|Z_i' - \mathbb{E}[Z_i']\| \leqslant \|Z_i'\| + \|\mathbb{E}[Z_i']\| \leqslant R + \mathbb{E}[\|Z_i'\|] \leqslant 2R$. The final bound on the matrix variance follows from the facts that for the p.s.d. matrix $\sum_{i=1}^n \mathbb{E}[(Z_i' - \mathbb{E}[Z_i'])^2] \preceq \sum_{i=1}^n \mathbb{E}[(Z_i')^2]$ and that $(Z_i')^2 \preceq Z_i^2$. □

2.14 Experimental Details

In our experiments we did find that gradient descent was able to decrease the loss in (2.4), but the algorithm was slow to converge. In practice, we found using the L-BFGS algorithm required no tuning and optimized the loss in (2.4) to high-precision in far fewer iterations [71]. Hence we used this first-order method throughout our experiments as our optimization routine. Our implementation is in Python, and we leveraged the `autograd` package to compute derivatives of the objective in (2.4), and the package `Ray` to parallelize our experiments [80, 87]. Each experiment is averaged over 30 repetitions with error bars representing ± 1 standard deviation over the repetitions. All the experiments herein were run on computer with 48 cores and 256 GB of RAM.

Note that after optimizing (2.4) directly using a first-order method in the variable $(\mathbf{U}, \mathbf{V})$, we can extract an estimate $\widehat{\mathbf{B}}$ of $\mathbf{B}$ by computing the column space of $\mathbf{V}$ (for example using the SVD of $\mathbf{V}$ or applying the Gram-Schmidt algorithm).

Chapter 3

On the Theory of Transfer Learning

3.1 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) [38, 48]. 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 [4], to problems such as protein engineering and design [42].

In this chapter, 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 [116, 97, 68]. There is also a theoretical literature that dates back at least as far as [6]. 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 $\hat{f}_0 \circ \hat{\mathbf{h}} \approx f_0^\star \circ \mathbf{h}^\star$.

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

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.

[84] 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 3.7). 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.

[2]See Theorem 3.3 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.

Related Work

The utility of multitask learning methods was observed at least as far back as [14]. In recent years, representation learning, transfer learning, and meta-learning have been the subject of extensive empirical investigation in the machine learning literature (see [8], [54] 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) [43]. 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 [44, 60, 24, 25, 61]. Additionally, [7] 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, [74], [96], and [16] have provided sample complexity bounds for the problem of transfer learning via representation learning. [74] and [91] also address sparsity-related issues that can arise in linear feature learning.

To our knowledge, [6] is the first theoretical work to provide generalization bounds for transfer learning via representation learning in a general setting. The formulation of [6] 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 3.3) and that there is a common covariate distribution across different tasks. In follow-up work, [84] propose a general, uniform-convergence-based framework for obtaining transfer learning guarantees which scale as $O(1/\sqrt{t}) + O(1/\sqrt{m})$ [84, 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 [84].

Recent work in [104] and [40] 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.

3.2 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.

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:

$$\mathbb{P}_j(\mathbf{x}, y) = \mathbb{P}_{f_j^\star \circ \mathbf{h}^\star}(\mathbf{x}, y) = \mathbb{P}_{\mathbf{x}}(\mathbf{x}) \mathbb{P}_{y|\mathbf{x}}(y | f_j^\star \circ \mathbf{h}^\star(\mathbf{x})). \tag{3.1}$$

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:

$$\hat{R}_{\text{train}}(\mathbf{f}, \mathbf{h}) := \frac{1}{nt} \sum_{j=1}^{t} \sum_{i=1}^{n} \ell(f_j \circ \mathbf{h}(\mathbf{x}_{ji}), y_{ji}), \tag{3.2}$$

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

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

$$\hat{R}_{\text{test}}(f, \mathbf{h}) := \frac{1}{m} \sum_{i=1}^{m} \ell(f \circ \mathbf{h}(\mathbf{x}_{0i}), y_{0i}). \tag{3.3}$$

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, $\hat{f}_0 = \arg\min_{f \in \mathcal{F}} \hat{R}_{\text{test}}(f, \hat{\mathbf{h}})$. 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*:

$$\text{Transfer Learning Risk} = R_{\text{test}}(\hat{f}_0, \hat{\mathbf{h}}) - R_{\text{test}}(f_0^\star, \mathbf{h}^\star). \tag{3.4}$$

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—$f_0^\star \circ \mathbf{h}^\star$.

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, $\bar{\mathbf{X}} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^\top$, the empirical Gaussian complexity is defined as

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

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 $\mathfrak{G}_N(\mathcal{Q}) = \mathbb{E}_{\bar{\mathbf{X}}}[\hat{\mathfrak{G}}_{\bar{\mathbf{X}}}(\mathcal{Q})]$, where the expectation is taken over the distribution of data samples $\bar{\mathbf{X}}$. Intuitively, $\mathfrak{G}_N(\mathcal{Q})$ measures the complexity of $\mathcal{Q}$ by the extent to which functions in the class $\mathcal{Q}$ can correlate with random noise g_{ki}.

3.3 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 3.1 (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: $\sup_{\mathbf{x}\in\mathcal{X}} |f \circ \mathbf{h}(\mathbf{x})| \leqslant D_{\mathcal{X}}$, 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 (3.2) and (3.3).

Assumption 3.2 (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 $j \in [t] \cup \{0\}$).*

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 3.1. 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 (3.1) for any $j \in [t]$, the **task-averaged representation difference** between representations $\mathbf{h}, \mathbf{h}' \in \mathcal{H}$ is:

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

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]

$$\bar{\mathfrak{G}}_n(\mathcal{F}) = \max_{\mathbf{Z}\in\mathcal{Z}} \hat{\mathfrak{G}}_{\mathbf{Z}}(\mathcal{F}), \text{ where } \mathcal{Z} = \{(\mathbf{h}(\mathbf{x}_1), \cdots, \mathbf{h}(\mathbf{x}_n)) \mid \mathbf{h} \in \mathcal{H}, \mathbf{x}_i \in \mathcal{X} \text{ for all } i \in [n]\}. \tag{3.5}$$

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

[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 3.1 and Theorem 3.7 for the formal statements.

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 3.1. *Let $\hat{\mathbf{h}}$ be an empirical risk minimizer of $\hat{R}_{train}(\cdot,\cdot)$ in (3.2). Then, if Assumptions 3.1 and 3.2 hold, with probability at least $1-\delta$:*

$$\bar{d}_{\mathcal{F},\mathbf{f}^\star}(\hat{\mathbf{h}};\mathbf{h}^\star) \leqslant 16L\mathfrak{G}_{nt}(\mathcal{F}^{\otimes t}\circ\mathcal{H}) + 8B\sqrt{\frac{\log(2/\delta)}{nt}}$$
$$\leqslant 4096L\left[\frac{D_{\mathcal{X}}}{(nt)^2} + \log(nt)\cdot[L(\mathcal{F})\cdot\mathfrak{G}_{nt}(\mathcal{H}) + \bar{\mathfrak{G}}_n(\mathcal{F})]\right] + 8B\sqrt{\frac{\log(2/\delta)}{nt}}.$$

Theorem 3.1 asserts that the *task-averaged representation difference* (Definition 3.1) 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 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 [113]).

We now make several remarks on this result. First, Theorem 3.1 differs from standard supervised learning generalization bounds. Theorem 3.1 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 3.7), which may be of independent interest. This chain rule (Theorem 3.7) can be viewed as a generalization of classical Gaussian comparison inequalities and results such as the Ledoux-Talagrand contraction principle [67]. Further details and comparisons to the literature for this chain rule can be found in Section 3.6 (this result also avoids an absolute maxima over $\mathbf{x}_i\in\mathcal{X}$).

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 3.2. 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 (3.1), 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 3.2. *Let $\hat{f}_0$ be an empirical risk minimizer of $\hat{R}_{test}(\cdot, \hat{\mathbf{h}})$ in (3.3) for any feature representation $\hat{\mathbf{h}}$. Then if Assumptions 3.1 and 3.2 hold, and $f_0^\star \in \mathcal{F}_0$ for an unknown class $\mathcal{F}_0$, with probability at least $1-\delta$:*

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

Here $\bar{\mathfrak{G}}_m(\mathcal{F})$ is again the worst-case Gaussian complexity[4] as defined in (3.5). Theorem 3.2 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 $\bar{\mathfrak{G}}_m(\mathcal{F})$.

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 3.1 defines a notion of representation distance in terms of information channeled through the training tasks, while Definition 3.2 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 of 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 stage to learn about a narrow slice of the representation—making transferring to a generic new task difficult.

Definition 3.3. 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 $\mathbf{h}' \in \mathcal{H}$,

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

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

[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 3.2).

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 [104] and [40], 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 3.4 and Section 3.4.

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 3.3. *Let $(\cdot, \hat{\mathbf{h}})$ be an empirical risk minimizer of $\hat{R}_{train}(\cdot,\cdot)$ in (3.2), and $\hat{f}_0$ be an empirical risk minimizer of $\hat{R}_{test}(\cdot, \hat{\mathbf{h}})$ in (3.3) for the learned feature representation $\hat{\mathbf{h}}$. Then if Assumptions 3.1 and 3.2 hold, and the training tasks are (ν, ϵ)-diverse, with probability at least $1 - 2\delta$, the transfer learning risk in (3.4) is upper-bounded by:*

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

Theorem 3.3 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.

3.4 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 (3.4)–in Theorem 3.3.

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 $d \geqslant r$, and let the function classes $\mathcal{F}$ and $\mathcal{H}$ be:

$$\begin{aligned}\mathcal{F} =& \{\ f \mid f(\mathbf{z}) = \boldsymbol{\alpha}^\top \mathbf{z},\ \boldsymbol{\alpha} \in \mathbb{R}^r,\ \|\boldsymbol{\alpha}\| \leqslant c_1\}, \\ \mathcal{H} =& \{\ \mathbf{h} \mid \mathbf{h}(\mathbf{x}) \ =\ \mathbf{B}^\top \mathbf{x},\ \mathbf{B} \in \mathbb{R}^{d\times r},\ \mathbf{B} \text{ is a matrix with orthonormal columns}\}.\end{aligned} \tag{3.6}$$

Here $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{0, 1\}$, and the measure $\mathbb{P}_{\mathbf{x}}$ is $\boldsymbol{\Sigma}$-sub-gaussian (see Definition 3.4) and D-bounded (i.e., $\|\mathbf{x}\| \leqslant D$ with probability one). We let the conditional distribution in (3.1) 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 3.3. *$\sigma_r(\mathbf{A}^\top \mathbf{A}/t) = \tilde{\nu} > 0$ and $\|\boldsymbol{\alpha}_j^\star\| \leqslant O(1)$ 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 $\sigma_1(\mathbf{A}^\top \mathbf{A}/t)/\sigma_r(\mathbf{A}^\top \mathbf{A}/t) \leqslant O(1)$ (w.h.p.) for example, if each $\boldsymbol{\alpha}_t \sim \mathcal{N}(0, \frac{1}{\sqrt{r}}\boldsymbol{\Sigma})$ i.i.d. where $\sigma_1(\boldsymbol{\Sigma})/\sigma_r(\boldsymbol{\Sigma}) \leqslant O(1)$.

Assumption 3.3 with natural choices of $\mathcal{F}_0$ and $\mathcal{F}$ establishes $(\Omega(\tilde{\nu}), 0)$-diversity as defined in Definition 3.3 (see Lemma 3.1). Finally, by standard arguments, we can bound the Gaussian complexity of $\mathcal{H}$ in this setting by $\mathfrak{G}_N(\mathcal{H}) \leqslant \tilde{O}(\sqrt{dr^2/N})$. 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 3.4 for more details). Together, these give the following guarantee.

Theorem 3.4. *If Assumption 3.3 holds, $\mathbf{h}^\star(\cdot) \in \mathcal{H}$, and $\mathcal{F}_0 = \{\ f \mid f(\mathbf{x}) = \boldsymbol{\alpha}^\top \mathbf{z},\ \boldsymbol{\alpha} \in \mathbb{R}^r,\ \|\boldsymbol{\alpha}\| \leqslant c_2\}$, 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, $n \geqslant c_3(d + \log t)$, $m \geqslant c_3 r$, and $D \leqslant c_3(\min(\sqrt{dr^2}, \sqrt{rm}))$, then with probability at least $1 - 2\delta$:*

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

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 3.4 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 $\mathrm{tr}(\mathbf{\Sigma}_{\mathbf{X}})$ and $\sum_{i=1}^{r} \sigma_i(\mathbf{\Sigma}_{\mathbf{X}_j})$ which can be much smaller then their counterparts d, r in Theorem 3.4, if the data lies on/or close to a low-dimensional subspace[5].

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:

$$\mathbf{h}(\mathbf{x}) = \mathbf{W}_K \sigma_{K-1}(\mathbf{W}_{K-1}(\sigma_{K-2}(\dots \sigma(\mathbf{W}_1 \mathbf{x})))). \tag{3.7}$$

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]

$$\begin{aligned}
\mathcal{F} =& \{\ f \mid f(\mathbf{z}) = \boldsymbol{\alpha}^\top \mathbf{z},\ \boldsymbol{\alpha} \in \mathbb{R}^r,\ \|\boldsymbol{\alpha}\| \leqslant c_1 M(K)^2\}, \\
\mathcal{H} =& \{\mathbf{h}(\cdot) \in \mathbb{R}^r \text{ in (3.7) for } \mathbf{W}_k : \|\mathbf{W}_k\|_{1,\infty} \leqslant M(k) \text{ for } k \in [K-1], \\
& \max(\|\mathbf{W}_K\|_{1,\infty}, \|\mathbf{W}_K\|_{\infty\to 2}) \leqslant M(K), \text{ such that } \sigma_r\left(\mathbb{E}_{\mathbf{x}}[\mathbf{h}(\mathbf{x})\mathbf{h}(\mathbf{x})^\top]\right) > \Omega(1)\}.
\end{aligned} \tag{3.8}$$

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 (3.1) be induced by:

$$y = \boldsymbol{\alpha}^\top \mathbf{h}(\mathbf{x}) + \eta \quad \text{for} \quad \boldsymbol{\alpha}, \mathbf{h} \ \text{ as in (3.8)}, \tag{3.9}$$

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 $\mathbf{A} = (\boldsymbol{\alpha}_1^\star, \dots, \boldsymbol{\alpha}_t^\star)^\top \in \mathbb{R}^{t \times r}$ as in the previous example. Here we use exactly the same diversity/normalization assumption on the task-specific maps—Assumption 3.3—as in our logistic regression example.

Choosing $\mathcal{F}_0$ and $\mathcal{F}$ appropriately establishes a $(\Omega(\tilde{\nu}), 0)$-diversity as defined in Definition 3.3 (see Lemma 3.6). Standard arguments as well as results in [47] allow us to bound the Gaussian complexity terms as follows (see the proof of Theorem 3.5 for details):

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

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

[5]Here $\mathbf{\Sigma}_{\bar{\mathbf{X}}}$ denotes the empirical covariance of the data matrix $\bar{\mathbf{X}}$. See Corollary 3.3 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 $\|\mathbf{W}_k\|_{1,\infty} \leqslant M(k)$ for each j in the depth-K network [47], and that the feature map is well-conditioned.

Theorem 3.5. *If Assumption 3.3 holds,* $\mathbf{h}^\star(\cdot) \in \mathcal{H}$, *and* $\mathcal{F}_0 = \{\ f \mid f(\mathbf{z}) = \boldsymbol{\alpha}^\top \mathbf{z},\ \boldsymbol{\alpha} \in \mathbb{R}^r,\ \|\boldsymbol{\alpha}\| \leqslant c_2\}$, *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* $M(K) \geqslant c_3$ *for a universal constant* c_3, *then with probability at least* $1 - 2\delta$*:*

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

The $\text{poly}(M(K))$ 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 $\text{poly}(M(K)) \cdot D\sqrt{K} \cdot \Pi_{k=1}^{K-1} M(k)$. By contrast a naive algorithm which does not leverage the training samples would have a sample complexity of $\tilde{O}\left(\text{poly}(M(K)) \cdot D\sqrt{K} \cdot \Pi_{k=1}^{K-1} M(k)/\sqrt{m}\right)$ via a similar analysis. Such a rate can be much larger than the bound in Theorem 3.5 when $nt \gg m$: exactly the setting relevant to that of few-shot learning for which ImageNet pretraining is often used.

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 [10]. As flexible tools for general-purpose, non-linear dimensionality reduction, index models have found broad applications in economics, finance, biology and the social sciences [10, 70, 92]. 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:

$$\begin{aligned} \mathcal{F} &= \{\ f \mid f(\mathbf{z}) \text{ is a 1-Lipschitz, monotonic function bounded in } [0,1]\}, \\ \mathcal{H} &= \{\ \mathbf{h} \mid \mathbf{h}(\mathbf{x}) \ =\ \mathbf{b}^\top \mathbf{x},\ \mathbf{b} \in \mathbb{R}^d,\ \|\mathbf{b}\| \leqslant W\}. \end{aligned} \tag{3.10}$$

We consider the setting where $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}$, the measure $\mathbb{P}_\mathbf{x}$ is D-bounded, and $DW \geqslant 1$. This matches the setting in [59]. The conditional distribution in (3.1) is induced by:

$$y = f(\mathbf{b}^\top \mathbf{x}) + \eta \quad \text{for } f, \mathbf{b} \text{ as in (3.7)},$$

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 $\mathcal{F}_{t,\tilde{\epsilon}} = \{f : \exists \tilde{f} \in \mathcal{F}_t \text{ such that } \sup_z |f(z) - \tilde{f}(z)| \leqslant \tilde{\epsilon}\}$.

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 $\tilde{\nu} \geqslant \frac{1}{t}$ as defined in Definition 3.3 (see Lemma 3.7). Standard arguments once again show that $\mathfrak{G}_N(\mathcal{H}) \leqslant O\left(\sqrt{(W^2 \mathbb{E}_\mathbf{X}[\text{tr}(\boldsymbol{\Sigma}_\mathbf{X}])/N)}\right)$ and $\bar{\mathfrak{G}}_N(\mathcal{F}) \leqslant O\left(\sqrt{WD/N}\right)$ (see the proof of Theorem 3.6 for details). Together these give the following guarantee.

Theorem 3.6. *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 $\tilde{\nu} \geqslant \frac{1}{t}$. Further, with probability at least $1 - 2\delta$:*

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

As before, the complexity of learning the feature representation decays as $n \to \infty$. Hence if $\mathbb{E}[\mathrm{tr}(\mathbf{\Sigma_X})]$ is large, the aforementioned bound will provide significant savings over the bound which ignores the training phase samples of $O\left(\sqrt{(W^2 \mathbb{E}_{\mathbf{X}}[\mathrm{tr}(\mathbf{\Sigma_X})])/m}\right) + O(\sqrt{WD/m})$. 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.

3.5 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.

[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 3.7 for details.

Appendix

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 $\widetilde{O}$ to denote an expression that hides polylogarithmic factors in all problem parameters.

In the context of the two-stage ERM procedure introduced in Section 3.2 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 $\mathbf{\Sigma}_{\bar{\mathbf{X}}} = \frac{1}{N}\bar{\mathbf{X}}^\top\bar{\mathbf{X}}$ 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,

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

3.6 Proofs in Section 3.3

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

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 (3.1). We first include the proof of Theorem 3.1 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 3.1. For fixed $\mathbf{f}', \mathbf{h}'$, define the centered training risk as,

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

and its empirical counterpart,

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

Now if $\tilde{\mathbf{f}}$ denotes a minimizer of the former expression for fixed $\hat{\mathbf{h}}$, in the sense that $\tilde{\mathbf{f}} = \frac{1}{t}\sum_{j=1}^{t} \arg\inf_{f_j' \in \mathcal{F}} \mathbb{E}_{\mathbf{x}_j, y_j}\Big\{\ell(f_j' \circ \hat{\mathbf{h}}(\mathbf{x}_j), y_j) - \ell(f_j^\star \circ \mathbf{h}^\star(\mathbf{x}_j), y_j)\Big\}$, 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 (3.2).

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

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

$$\begin{aligned} &L(\hat{\mathbf{f}}, \hat{\mathbf{h}}, \mathbf{f}^\star, \mathbf{h}^\star) - L(\mathbf{f}^\star, \mathbf{h}^\star, \mathbf{f}^\star, \mathbf{h}^\star) = \\ &\underbrace{L(\hat{\mathbf{f}}, \hat{\mathbf{h}}, \mathbf{f}^\star, \mathbf{h}^\star) - \hat{L}(\hat{\mathbf{f}}, \hat{\mathbf{h}}, \mathbf{f}^\star, \mathbf{h}^\star)}_{b} + \underbrace{\hat{L}(\hat{\mathbf{f}}, \hat{\mathbf{h}}, \mathbf{f}^\star, \mathbf{h}^\star) - \hat{L}(\mathbf{f}^\star, \mathbf{h}^\star, \mathbf{f}^\star, \mathbf{h}^\star)}_{c} + \\ &\underbrace{\hat{L}(\mathbf{f}^\star, \mathbf{h}^\star, \mathbf{f}^\star, \mathbf{h}^\star) - L(\mathbf{f}^\star, \mathbf{h}^\star, \mathbf{f}^\star, \mathbf{h}^\star)}_{d}. \end{aligned}$$

By definition $c \leqslant 0$ (note this inequality uses the realizability in Assumption 3.2) and $b, d \leqslant \sup_{\mathbf{f} \in \mathcal{F}^{\otimes t}, \mathbf{h} \in \mathcal{H}} |R_{\text{train}}(\mathbf{f}, \mathbf{h}) - \hat{R}_{\text{train}}(\mathbf{f}, \mathbf{h})|$. By an application of the bounded differences inequality and a standard symmetrization argument (see for example [113, Theorem 4.10] we have that,

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

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 $|\ell_{ji}(0, y_{ji})| \leqslant B$, the constant-shift property of Rademacher averages [113, Exercise 4.7c] gives,

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

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

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 [67, Theorem 4.12] over this set shows,

$$\mathbb{E}_{\boldsymbol{\epsilon}}[\sup_{\mathbf{f}\in\mathcal{F}^{\otimes t},\mathbf{h}\in\mathcal{H}}\frac{1}{nt}\sum_{j=1}^{t}\sum_{i=1}^{n}\epsilon_{ij}\ell_{ij}(f_j\circ\mathbf{h}(\mathbf{x}_{ji}),y_{ji})]\leqslant 2L\cdot\mathfrak{R}_{nt}(\mathcal{F}^{\otimes t}(\mathcal{H})). \tag{3.11}$$

Combining gives,

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

with probability $1-2\delta$. Now note by [67, p.97] empirical Rademacher complexity is upper bounded by empirical Gaussian complexity: $\hat{\mathfrak{R}}_{\mathbf{X}}(\mathcal{F}^{\otimes t}(\mathcal{H}))\leqslant\sqrt{\frac{\pi}{2}}\hat{\mathfrak{G}}_{\mathbf{X}}(\mathcal{F}^{\otimes t}(\mathcal{H}))$. 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 3.7 to decompose the Gaussian complexity over $\mathcal{F}$ and $\mathcal{H}$. A direct application of Theorem 3.7 gives the conclusion that,

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

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 $D_{\mathbf{X}}\leqslant 2D_{\mathcal{X}}$ and similarly that $\max_{\mathbf{Z}\in\mathcal{Z}}\hat{\mathfrak{G}}_{\mathbf{Z}}(\mathcal{F})\leqslant\max_{\mathbf{Z}\in\mathcal{Z}_1}\hat{\mathfrak{G}}_{\mathbf{Z}}(\mathcal{F})$ 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. □

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

Corollary 3.1. *In the setting of Theorem 3.1,*

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

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,*

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

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

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

Proof. The argument follows analogously to the proof of Theorem 3.1. The first statement follows identically by avoiding the relaxation–$\max_{\mathbf{Z}\in\mathcal{Z}}\hat{\mathfrak{G}}_{\mathbf{Z}}(\mathcal{F})\leqslant\max_{\mathbf{Z}\in\mathcal{Z}_1}\hat{\mathfrak{G}}_{\mathbf{Z}}(\mathcal{F})$ 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 3.7 in the proof of Theorem 3.1.

The second statement also follows by a direct modification of the proof of Theorem 3.1. In the proof another application of the bounded differences inequality would show that $|\mathfrak{R}_{nt}(\mathcal{F}^{\otimes t}(\mathcal{H}))-\hat{\mathfrak{R}}_{\mathbf{X}}((\mathcal{F}^{\otimes t}(\mathcal{H}))|\leqslant 4B\sqrt{\frac{\log(1/\delta)}{nt}}$ with probability $1-2\delta$. Applying this inequality after (3.11) and union bounding over this event and the event in the theorem, followed by the steps in Theorem 3.1, gives the result after an application of Theorem 3.7. □

We now show how the definition of task diversity in Definition 3.3 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 3.2. Note $\tilde{f}_0=\arg\min_{f\in\mathcal{F}}R_{\text{test}}(f,\hat{\mathbf{h}})$–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, $b\leqslant 0$. Now a standard uniform convergence/symmetrization argument which also follows the same steps as in the proof of Theorem 3.1,

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

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,

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

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

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

Combining terms gives the conclusion. □

We also present a version of Theorem 3.2 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 3.1. 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 3.2. *In the setting of Theorem 3.2, assuming the loss function ℓ satisfies the centering $\ell(0, y) = 0$ for all $y \in \mathcal{Y}$,*

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

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,

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

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

Proof of Corollary 3.2. The proof is identical to the proof of Theorem 3.2 save in how the concentration argument is performed. Namely in the notation of Theorem 3.2, we upper bound,

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

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 [82]. Applying [82, Theorem 3] therein, we can conclude,

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

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

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

Next note a standard symmetrization argument shows that $\mathbb{E}[Z] \leqslant 2\mathbb{E}_{\mathbf{X}_0,\mathbf{y}_0}[\hat{\mathfrak{R}}_{\mathbf{Z}_{\hat{\mathbf{h}}}}(\ell \circ \mathcal{F})]$ for $\mathbf{Z}_{\hat{\mathbf{h}}} = \hat{\mathbf{h}}(\mathbf{X}_0)$. Following the proof of Theorem 3.2 but eschewing the unnecessary centering step in the application of the contraction principle shows that, $\hat{\mathfrak{R}}_{\mathbf{Z}_{\hat{\mathbf{h}}}}(\ell \circ \mathcal{F}) \leqslant 2L \cdot \hat{\mathfrak{R}}_{\mathbf{Z}_{\hat{\mathbf{h}}}}(\mathcal{F})$. Upper bounding empirical Rademacher complexity by Gaussian complexity and following the steps of Theorem 3.2 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 [5, Lemma A.4 (i)]. Directly applying this result (with $\alpha = \frac{1}{2}$) shows that,

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

with probability at least $1-\delta$. The remainder of the argument follows exactly as before and as in the proof of Theorem 3.2 along with another union bound. □

The proof of Theorem 3.3 is almost immediate.

Proof of Theorem 3.3. The result follows immediately by combining Theorem 3.1, Theorem 3.2, and the definition of task diversity along with a union bound over the two events on which Theorems 3.1 and 3.2 hold. □

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 3.7) can be viewed as a generalization of the Ledoux-Talagrand contraction principle which shows that for a *fixed*, centered L-Lipschitz function ϕ, $\hat{\mathfrak{G}}_{\mathbf{X}}(\phi(\mathcal{F})) \leqslant 2L\hat{\mathfrak{G}}_{\mathbf{X}}(\mathcal{F})$. 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 3.7 is used in [84] for multi-task learning applications, drawing on the chain rule of [83]. 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 ([84, Eq. 4]) in lieu of $\hat{\mathfrak{G}}_{\mathcal{Z}}(\mathcal{F})$ in Theorem 3.7. 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 3.7 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 [83] both hinge on several properties of Gaussian processes. [83] 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, $d^2_{2,\mathbf{X}}(\mathbf{f}(\mathbf{h}), \mathbf{f}'(\mathbf{h}')) = \frac{1}{t\cdot n}\sum_{j=1}^{t}\sum_{i=1}^{n}(f_j(\mathbf{h}(\mathbf{x}_{ji})) - f'_j(\mathbf{h}'(\mathbf{x}_{ji}))^2$, 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 $d^2_{2,\mathbf{X}}(\mathbf{h}, \mathbf{h}') = \frac{1}{t\cdot n}\sum_{k=1}^{r}\sum_{j=1}^{t}\sum_{i=1}^{n}(\mathbf{h}_k(\mathbf{x}_{ji}) - \mathbf{h}'_k(\mathbf{x}_{ji}))^2$.

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 3.7. *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,*

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

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 $C(\mathcal{F}^{\otimes t}(\mathcal{H})) \leqslant D_{\mathbf{X}}$ then by computing the exact infima of the expression,

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

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 $d^2_{2,\mathbf{X}}(\mathbf{f}(\mathbf{h}), \mathbf{f}'(\mathbf{h}')) = \frac{1}{N}\sum_{j=1}^{t}\sum_{i=1}^{n}(f_j(\mathbf{h}(\mathbf{x}_{ji})) - f'_j(\mathbf{h}'(\mathbf{x}_{ji}))^2$. Since $Z_{\mathbf{f}(\mathbf{h})}$ is a mean-zero stochastic process we have that, $\mathbb{E}[\sup_{\mathbf{f}(\mathbf{h})\in\mathcal{F}^{\otimes t}(\mathcal{H})} Z_{\mathbf{f}(\mathbf{h})}] = \mathbb{E}[\sup_{\mathbf{f}(\mathbf{h})\in\mathcal{F}^{\otimes t}(\mathcal{H})} Z_{\mathbf{f}(\mathbf{h})} - Z_{\mathbf{f}'(\mathbf{h}')}] \leqslant \mathbb{E}[\sup_{\mathbf{f}(\mathbf{h}),\mathbf{f}'(\mathbf{h}')\in\mathcal{F}^{\otimes t}(\mathcal{H})} Z_{\mathbf{f}(\mathbf{h})} - Z_{\mathbf{f}'(\mathbf{h}')}]$. Now an appeal to the Dudley entropy integral bound, [113, Theorem 5.22] shows that,

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

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 $\sup_{d_{2,\mathbf{X}}(\mathbf{f}(\mathbf{h}),\mathbf{f}(\mathbf{h}'))\leqslant\delta} Z_{\mathbf{f}(\mathbf{h})} - Z_{\mathbf{f}(\mathbf{h}')} \leqslant \sup_{\mathbf{v}:\|\mathbf{v}\|_2\leqslant\delta} \mathbf{g}\cdot\mathbf{v} \leqslant \|\mathbf{g}\|\delta$. The corresponding expectation bound, after an application of Jensen's inequality to the $\sqrt{\cdot}$ function gives $\mathbb{E}[\sup_{d_{2,\mathbf{X}}(\mathbf{f}(\mathbf{h}),\mathbf{f}(\mathbf{h}'))\leqslant\delta} Z_{\mathbf{f}(\mathbf{h})} - Z_{\mathbf{f}(\mathbf{h}')}] \leqslant \mathbb{E}[\|\mathbf{g}\|_2\delta] \leqslant \sqrt{N}\delta$.

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,

$$d_{2,\mathbf{X}}(\mathbf{f}(\mathbf{h}), \mathbf{f}'(\mathbf{h}')) \leqslant d_{2,\mathbf{X}}(\mathbf{f}(\mathbf{h}), \mathbf{f}(\mathbf{h}')) + d_{2,\mathbf{X}}(\mathbf{f}(\mathbf{h}'), \mathbf{f}'(\mathbf{h}')) =$$
$$\sqrt{\frac{1}{N}\sum_{j=1}^{t}\sum_{i=1}^{n}(f_j(\mathbf{h}(\mathbf{x}_{ji})) - f_j(\mathbf{h}'(\mathbf{x}_{ji})))^2} + \sqrt{\frac{1}{N}\sum_{j=1}^{t}\sum_{i=1}^{n}(f_j(\mathbf{h}'(\mathbf{x}_{ji})) - f'_j(\mathbf{h}'(\mathbf{x}_{ji})))^2} \leqslant$$

$$L(\mathcal{F})\sqrt{\frac{1}{N}\sum_{k=1}^{r}\sum_{j=1}^{t}\sum_{i=1}^{n}(\mathbf{h}_k(\mathbf{x}_{ji})-\mathbf{h}'_k(\mathbf{x}_{ji}))^2}+\sqrt{\frac{1}{N}\sum_{j=1}^{t}\sum_{i=1}^{n}(f_j(\mathbf{h}'(\mathbf{x}_{ji}))-f'_j(\mathbf{h}'(\mathbf{x}_{ji})))^2}=$$
$$L(\mathcal{F})\cdot d_{2,\mathbf{X}}(\mathbf{h},\mathbf{h}')+d_{2,\mathbf{h}'(\mathbf{X})}(\mathbf{f},\mathbf{f}')\leqslant\epsilon_1\cdot L(\mathcal{F})+\epsilon_2$$

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 $|C_{\mathcal{F}^{\otimes t}(\mathcal{H})}|=\sum_{\mathbf{h}\in C_{\mathcal{H}_\mathbf{X}}}|C_{\mathcal{F}^{\otimes t}_{\mathbf{h}(\mathbf{X})}}|\leqslant|C_{\mathcal{H}_\mathbf{X}}|\cdot\max_{\mathbf{h}\in\mathcal{H}_\mathbf{X}}|C_{\mathcal{F}^{\otimes t}_{\mathbf{h}(\mathbf{X})}}|$. 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 $\max_{\mathbf{h}\in\mathcal{H}_\mathbf{X}}|C_{\mathcal{F}^{\otimes t}_{\mathbf{h}(\mathbf{X})}}|\leqslant|C_{\mathcal{F}_{\mathbf{h}(\mathbf{X}_1)}}\times\ldots\times C_{\mathcal{F}_{\mathbf{h}(\mathbf{X}_t)}}|\leqslant|\underbrace{\max_{\mathbf{z}\in\mathcal{Z}}C_{\mathcal{F}_\mathbf{z}}\times\ldots\times\max_{\mathbf{z}\in\mathcal{Z}}C_{\mathcal{F}_\mathbf{z}}}_{t\text{ times}}|\leqslant|\max_{\mathbf{z}\in\mathcal{Z}}C_{\mathcal{F}_\mathbf{z}}|^t$. Combining these facts provides a bound on the metric entropy of,

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

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,

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

From the Sudakov minoration theorem [113][Theorem 5.30] for Gaussian processes and the fact packing numbers at scale u upper bounds the covering number at scale u we find:

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

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

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

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

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. □

3.7 Proofs in Section 3.4

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 3.3 or its variants.

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

We begin by presenting the proof of the Theorem 3.4 which essentially relies on instantiating a variant of Theorem 3.3. In order to obtain a sharper dependence in the noise terms in the test learning stage we actually directly combine Corollaries 3.1 and 3.2.

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 3.4. First note due to the task normalization conditions we can choose c_1, c_2 sufficiently large so that the realizability assumption in Assumption 3.2 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 $c_1 \geqslant \frac{C}{c}c_2$. Next note that under the conditions of the result we can use Lemma 3.1 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 Corollaries 3.1 and 3.2 we begin by bounding each of the complexity terms in the expression. First,

- In the following we use $\mathbf{b}_k$ for $k \in [r]$ to index the orthonormal columns of $\mathbf{B}$. For the feature learning complexity in the training phase we obtain,

$$\hat{\mathfrak{G}}_{\mathbf{X}}(\mathcal{H}) = \frac{1}{nt}\mathbb{E}[\sup_{\mathbf{B}\in\mathcal{H}} \sum_{k=1}^{r}\sum_{j=1}^{t}\sum_{i=1}^{n} g_{kji}\mathbf{b}_k^\top \mathbf{x}_{ji}] = \frac{1}{nt}\mathbb{E}[\sup_{(\mathbf{b}_1,\ldots,\mathbf{b}_r)\in\mathcal{H}} \sum_{k=1}^{r}\mathbf{b}_k^\top(\sum_{j=1}^{t}\sum_{i=1}^{n} g_{kji}\mathbf{x}_{ji})] \leqslant$$
$$\frac{1}{nt}\sum_{k=1}^{r}\mathbb{E}[\|\sum_{j=1}^{t}\sum_{i=1}^{n} g_{kji}\mathbf{x}_{ji}\|] \leqslant \frac{1}{nt}\sum_{k=1}^{r}\sqrt{\mathbb{E}[\|\sum_{j=1}^{t}\sum_{i=1}^{n} g_{kji}\mathbf{x}_{ji}\|^2]} \leqslant \frac{1}{nt}\sum_{k=1}^{r}\sqrt{\sum_{j=1}^{t}\sum_{i=1}^{n}\|\mathbf{x}_{ji}\|^2}$$
$$= \frac{r}{\sqrt{nt}}\sqrt{\mathrm{tr}(\mathbf{\Sigma}_{\mathbf{X}})}.$$

  Further by definition the class $\mathcal{F}$ as linear maps with parameters $\|\boldsymbol{\alpha}\|_2 \leqslant O(1)$ we obtain that $L(\mathcal{F}) = O(1)$. We now proceed to convert this to a population quantity by noting that $\mathbb{E}[\sqrt{\mathrm{tr}(\mathbf{\Sigma}_{\mathbf{X}})}] \leqslant \sqrt{d\cdot\mathbb{E}[\|\mathbf{\Sigma}_{\mathbf{X}}\|]} \leqslant O(\sqrt{d})$ for $nt \gtrsim d$ by Lemma 3.4.

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

$$\hat{\mathfrak{G}}_{\mathbf{h}(\mathbf{X})}(\mathcal{F}) = \frac{1}{n}\mathbb{E}[\sup_{\|\boldsymbol{\alpha}\|\leqslant c_1}\sum_{i=1}^{n} g_i\boldsymbol{\alpha}^\top\mathbf{B}^\top\mathbf{x}_{ji}] = \frac{c_1}{n}\mathbb{E}[\|\sum_{i=1}^{n} g_i\mathbf{B}^\top\mathbf{x}_{ji}\|] \leqslant \frac{c_1}{n}\sqrt{\sum_{i=1}^{n}\|\mathbf{B}^\top\mathbf{x}_{ji}\|^2} =$$
$$\frac{c_1}{\sqrt{n}}\sqrt{\mathrm{tr}(\mathbf{B}\mathbf{B}^\top\mathbf{\Sigma}_{\mathbf{X}_j})} = \frac{c_1}{\sqrt{n}}\sqrt{\mathrm{tr}(\mathbf{B}^\top\mathbf{\Sigma}_{\mathbf{X}_j}\mathbf{B})}.$$

  Now by the variational characterization of singular values it follows that

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

  Thus it immediately follows that,

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

  for $j \in [t]$. We can convert this to a population quantity again by applying Lemma 3.4 which shows $\mathbb{E}[\sqrt{\sum_{i=1}^{r}\sigma_i(\mathbf{\Sigma}_{\mathbf{X}_j})}] \leqslant O(\sqrt{r})$ for $n \gtrsim d + \log t$. Hence $\bar{\mathfrak{G}}_n(\mathcal{F}) \leqslant O(\sqrt{\frac{r}{n}})$.

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

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

  Crucially, here we can apply the first result in Corollary 3.2 which allows us to take the expectation over $\mathbf{X}_0$ before maximizing over $\mathbf{B}$. Thus applying Lemma 3.4 as before gives the result, $\mathbb{E}[\sqrt{\sum_{i=1}^{r}\sigma_i(\mathbf{B}^\top\mathbf{\Sigma}_{\mathbf{X}_0}\mathbf{B})}] \leqslant O(\sqrt{r})$ for $m \gtrsim r$. Hence $\bar{\mathfrak{G}}_m(\mathcal{F}) \leqslant O(\sqrt{\frac{r}{m}})$.

This gives the first series of claims.

Finally we verify that Assumption 3.1 holds so as to use Theorem 3.1 and Corollary 3.2 to instantiate the end-to-end guarantee. First the boundedness parameter becomes,

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

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

Lastly, to use Corollary 3.2 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 the logistic loss $\ell(\cdot,\cdot)$ satisfies the 1-Lipschitz property uniformly we have that, $\mathrm{Var}(\ell(f \circ \hat{\mathbf{h}}(\mathbf{x}_i), y_i)) \leqslant \mathrm{Var}(f \circ \hat{\mathbf{h}}(\mathbf{x}_i))$ for each $i \in [m]$. Collapsing the variance we have that,

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

under our assumptions which implies that $\sigma \leqslant O(1)$. 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. □

Recall, in the context of the two-stage ERM procedure introduced in Section 3.2 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.

We now state a sharp, data-dependent guarantee for logistic regression.

Corollary 3.3. *If Assumption 3.3 holds, $\mathbf{h}^\star(\cdot) \in \mathcal{H}$, and $\mathcal{F}_0 = \{\ f \mid f(\mathbf{x}) = \boldsymbol{\alpha}^\top \mathbf{z},\ \boldsymbol{\alpha} \in \mathbb{R}^r,\ \|\boldsymbol{\alpha}\| \leqslant c_2\}$, 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 $\leqslant$

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

Proof of Corollary 3.3. This follows immediately from the proof of Theorem 3.4 and applying Corollaries 3.1 and 3.2. Merging terms and applying a union bound gives the result. □

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

Lemma 3.1. *Let Assumption 3.3 hold in the setting of Theorem 3.4. Then there exists c_2 such that if $c_1 \geqslant \frac{C}{c}c_2$ the problem is task-diverse with parameter $(\Omega(\tilde{\nu}), 0)$ in the sense of Definition 3.3 where $\tilde{\nu} = \sigma_r(\mathbf{A}^\top\mathbf{A}/t)$.*

Proof. Our first observation specializes Lemma 3.2 to the case of logistic regression where $\Phi(\eta) = \log(1 + \exp(\eta))$, $s(\sigma) = 1$ with $\mathbf{h}(\mathbf{x}) = \mathbf{B}\mathbf{x}$ 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 $c_1 \geqslant \frac{C}{c}c_2$. These conditions are consistent with the realizability assumption.

This lemma uses smoothness and (local) strong convexity to bound the task-averaged representation distance and worst-case representation difference by relating it to a result for the squared loss established in Lemma 3.6. By appealing to Lemma 3.2 and Lemma 3.3 we have that,

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

for $\mathbf{x}_j, y_j \sim (\mathbb{P}_{\mathbf{x}}(\cdot), \mathbb{P}_{y|\mathbf{x}}(\cdot|f_j \circ \mathbf{h}(\mathbf{x}))$ We now bound each term in the task diversity,

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

$$d_{\mathcal{F},\mathcal{F}_0}(\hat{\mathbf{h}}; \mathbf{h}^\star) = \sup_{f_0\in\mathcal{F}_0} \inf_{\hat{f}\in\mathcal{F}} \mathbb{E}_{\mathbf{x},y}[\ell(\hat{f} \circ \hat{\mathbf{h}}(\mathbf{x}), y) - \ell(f_0 \circ \mathbf{h}^\star(\mathbf{x}), y)]] \leqslant$$
$$\sup_{\boldsymbol{\alpha}:\|\boldsymbol{\alpha}\|_2\leqslant c_2} \inf_{\hat{\boldsymbol{\alpha}}:\|\hat{\boldsymbol{\alpha}}\|\leqslant c_1} \frac{1}{8}\mathbb{E}_{\mathbf{x}}[(\hat{\mathbf{h}}(\mathbf{x})^\top\hat{\boldsymbol{\alpha}} - \mathbf{h}^\star(\mathbf{x})^\top\boldsymbol{\alpha})^2].$$

  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 $\inf_{\hat{\boldsymbol{\alpha}}} \frac{1}{8}\mathbb{E}_{\mathbf{x}_i}[(\hat{\mathbf{h}}(\mathbf{x}_i)^\top\hat{\boldsymbol{\alpha}} - \mathbf{h}^\star(\mathbf{x}_i)^\top\boldsymbol{\alpha})^2] \to \hat{\boldsymbol{\alpha}}_{unconstrained} = -\mathbf{F}_{\hat{\mathbf{h}}\hat{\mathbf{h}}}\mathbf{F}_{\hat{\mathbf{h}}\mathbf{h}}\boldsymbol{\alpha} = (\hat{\mathbf{B}}^\top\mathbf{\Sigma}\hat{\mathbf{B}})^{-1}(\hat{\mathbf{B}}^\top\mathbf{\Sigma}\hat{\mathbf{B}})\boldsymbol{\alpha}$ from the proof and preamble of Lemma 3.6. Note that since $\hat{\mathbf{B}}$ and $\mathbf{B}$ have orthonormal

columns it follows that $\|\hat{\boldsymbol{\alpha}}\| \leqslant \frac{C}{c}c_2$ since $\hat{\mathbf{B}}^\top \boldsymbol{\Sigma}\hat{\mathbf{B}}$ is invertible. Thus if $c_1 \geqslant \frac{C}{c}c_2$, 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 3.6 we obtain $\sup_{\boldsymbol{\alpha}:\|\boldsymbol{\alpha}\|_2 \leqslant c_2} \inf_{\hat{\boldsymbol{\alpha}}:\|\hat{\boldsymbol{\alpha}}\| \leqslant c_1} \frac{1}{8}\mathbb{E}_{\mathbf{x}_i}[(\hat{\mathbf{h}}(\mathbf{x}_i)^\top \hat{\boldsymbol{\alpha}} - \mathbf{h}^\star(\mathbf{x}_i)^\top \boldsymbol{\alpha})^2] \leqslant \frac{c_2}{8}\sigma_1(\Lambda_{sc}(\mathbf{h}, \hat{\mathbf{h}}))$.

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

$$\bar{d}_{\mathcal{F},\mathbf{f}^\star}(\mathbf{h}; \hat{\mathbf{h}}) = \frac{1}{t}\sum_{j=1}^{t} \inf_{\hat{f}\in\mathcal{F}} \mathbb{E}_{\mathbf{x}_j, y_j}[\ell(\hat{f} \circ \hat{\mathbf{h}}(\mathbf{x}_j), y_j) - \ell(f_j^\star \circ \mathbf{h}^\star(\mathbf{x}_j), y_j)]] \geqslant$$

$$\frac{1}{8t}\sum_{j=1}^{t} \inf_{\|\hat{\boldsymbol{\alpha}}\| \leqslant c_1} \mathbb{E}_{\mathbf{x}_j}[\exp(-\max(|\hat{\mathbf{h}}(\mathbf{x}_j)^\top \hat{\boldsymbol{\alpha}}|, |\mathbf{h}^\star(\mathbf{x}_j)^\top \boldsymbol{\alpha}_j^\star|)) \cdot (\hat{\mathbf{h}}(\mathbf{x}_j)^\top \hat{\boldsymbol{\alpha}} - \mathbf{h}^\star(\mathbf{x}_j)^\top \boldsymbol{\alpha}_j^\star)^2]$$

We will use the fact that in our logistic regression example $\mathbf{h}(\mathbf{x}_j) = \mathbf{B}\mathbf{x}_j$; in this case if $\mathbf{x}_j$ 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. For fixed j, note the random variables $Z_1 = (\boldsymbol{\alpha}_j^\star)^\top \mathbf{B}\mathbf{x}_j$ and $Z_2 = \hat{\boldsymbol{\alpha}}^\top \hat{\mathbf{B}}\mathbf{x}_j$ are sub-gaussian 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 $\mathbb{1}[E] = \mathbb{1}[|Z_1| \leqslant Ck\|\boldsymbol{\alpha}_j^\star\| \cap \mathbb{1}\{|Z_2| \leqslant Ck\|\hat{\boldsymbol{\alpha}}\|]$ 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,

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

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

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x}}[\mathbb{1}[E](\hat{\mathbf{h}}(\mathbf{x})^\top \hat{\boldsymbol{\alpha}} - \mathbf{h}^\star(\mathbf{x})^\top \boldsymbol{\alpha}_j^\star)^2] = \\
&\mathbb{E}_{\mathbf{x}}[(\hat{\mathbf{h}}(\mathbf{x})^\top \hat{\boldsymbol{\alpha}} - \mathbf{h}^\star(\mathbf{x})^\top \boldsymbol{\alpha}_j^\star)^2] - \mathbb{E}_{\mathbf{x}}[\mathbb{1}[E^c](\hat{\mathbf{h}}(\mathbf{x})^\top \hat{\boldsymbol{\alpha}} - \mathbf{h}^\star(\mathbf{x})^\top \boldsymbol{\alpha}_j^\star)^2]
\end{aligned}$$

.

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

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

Define $Z_3 = \mathbf{x}^\top((\mathbf{B}^\star)^\top\boldsymbol{\alpha}_j^\star - \hat{\mathbf{B}}^\top\hat{\boldsymbol{\alpha}})$ which by definition is sub-gaussian 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 [113, Theorem 2.6]) we can also conclude that,

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

Recalling the sub-gaussianity of Z_1 and Z_2, from an application of Markov and Jensen's inequality,

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

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

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

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

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

Now using the argument from the upper bound to compute the infima since all the $\|\boldsymbol{\alpha}_j^\star\| \leqslant c_2$ (and hence the constrained minimizers identical to the unconstrained minimizers for each of the j terms for $c_1 \geqslant \frac{C}{c}c_2$) and using the proof of Lemma 3.6 we conclude that,

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

Combining these upper and lower bounds and concluding as in the proof of Lemma 3.6 shows

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

□

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 3.1. If the data generating model satisfies the logistic model conditional likelihood as in Section 3.4, for the logistic loss ℓ we have that,

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

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

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

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

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

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

Proof.

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

since we have that

$$\frac{\Phi(\mathbf{h}(\mathbf{x})^\top\boldsymbol{\alpha})}{s(\sigma)}=\log\int dy\ b(y)\exp(\frac{y\mathbf{h}(\mathbf{x})^\top\boldsymbol{\alpha}\rangle}{s(\sigma)})\implies\frac{\Phi'(\mathbf{h}(\mathbf{x})^\top\boldsymbol{\alpha})}{s(\sigma)}=\frac{\int dy\ \mathbb{P}_{y|\mathbf{x}}(y|\boldsymbol{\alpha}^\top\mathbf{h}(\mathbf{x}))y}{s(\sigma)}$$

as it is the log-normalizer. Using Taylor's theorem we have that

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

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

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

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

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

□

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

Lemma 3.3. *Consider the $\mathbb{P}_{y|\mathbf{x}}(\cdot)$ logistic generative model defined in Section 3.4 for a general feature map $\mathbf{h}(\mathbf{x})$. Then for this conditional generative model in the setting of Lemma 3.2, where $\Phi(\eta) = \log(1 + \exp(\eta))$, $s(\sigma) = 1$, $b(y) = 1$,*

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

and

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

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, $\mathbb{E}_{\mathbf{x}}[\sup_{p(\mathbf{x}) \in S(\mathbf{x})} \Phi''(p(\mathbf{x}))] \leqslant \frac{1}{4}$ using a uniform upper bound. The lower bound follows by noting that

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

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

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 ([111, Theorem 4.7.3, Exercise 4.7.1] or [113, Theorem 6.5]). Further recall that in Definition 3.4 our covariate distribution is $O(1)$-sub-gaussian.

Lemma 3.4. *Let the common covariate distribution $\mathbb{P}_{\mathbf{x}}(\cdot)$ satisfy Definition 3.4. Then if $nt \gtrsim d$,*

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

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

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

and if $m \gtrsim r$,

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

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

Proof. All of these statements essentially follow by integrating a tail bound and applying the triangle inequality. For the first statement since $\mathbb{E}[\|\mathbf{\Sigma}_{\mathbf{X}}\|] = \mathbb{E}[\|\mathbf{\Sigma}_{\mathbf{X}} - \mathbf{\Sigma}\|] + \|\mathbf{\Sigma}\| \leqslant O(1)$, under the conditions $nt \gtrsim d$, the result follows directly by [111, Theorem 4.7.3].

For the second by [113, Theorem 6.5], $\mathbb{E}[\exp(\lambda\|\mathbf{\Sigma} - \mathbf{\Sigma}\|) \leqslant \exp(c_0(\lambda^2/N) + 4d)]$ for all $|\lambda| \leqslant \frac{N}{c_1}$, given a sample covariance averaged over N datapoints, and universal constants c_0, c_1. So using a union bound alongside a tail integration since the data is i.i.d. across tasks,

$$\begin{aligned}
&\mathbb{E}[\max_{j\in[t]} \|\mathbf{\Sigma}_{\mathbf{X}_j} - \mathbf{\Sigma}\|] \leqslant \int_0^\infty \min(1, t\mathbb{P}[\|\mathbf{\Sigma}_{\mathbf{X}_1} - \mathbf{\Sigma}\| > \delta])d\delta \leqslant \\
&\int_0^\infty \min(1, \exp(c_0(\lambda^2/n) + 4d + \log t - \lambda\delta)] \leqslant \\
&\int_0^\infty \min(1, \exp(4d + \log t) \cdot \exp(-c_2 \cdot n\min(\delta^2, \delta))d\delta \leqslant \\
&O\left(\sqrt{\frac{d + \log t}{n}} + \frac{d + \log t}{n}\right) \leqslant O(1),
\end{aligned}$$

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. □

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 which bounds the empirical Rademacher complexity of a deep neural network.

Theorem 3.8 (Theorem 2 adapted from [47]). *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 $\|\mathbf{x}_i\| \leqslant D$ for $i \in [n]$, where $\|\mathbf{W}_k\|_{1,\infty} \leqslant M(k)$ for all $k \in [K]$. Then,*

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

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 3.5. Note that we make use of the result $\hat{\mathfrak{R}}_{\mathbf{X}}(\mathcal{N}) \leqslant \sqrt{\frac{\pi}{2}} \cdot \hat{\mathfrak{G}}_{\mathbf{X}}(\mathcal{N})$ and that $\hat{\mathfrak{G}}_{\mathbf{X}}(\mathcal{N}) \leqslant 2\sqrt{\log N} \cdot \hat{\mathfrak{R}}_{\mathbf{X}}(\mathcal{N})$ for any function class $\mathcal{N}$ when $\mathbf{X}$ has N datapoints [67, p. 97].

Proof of Theorem 3.5. First note due to the task normalization conditions we can choose c_1, c_2 sufficiently large so that the realizability assumption in Assumption 3.2 is satisfied–in particular, we can assume that c_2 is chosen large enough to contain parameter $\boldsymbol{\alpha}_0^\star$ and c_1 large enough so that $c_1 M(K)^2 \geqslant c_1 c_3^2$ 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 3.6 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 $\|\mathbb{E}_{\mathbf{x}}[\hat{\mathbf{h}}(\mathbf{x})\mathbf{h}^\star(\mathbf{x})^\top]\|_2 \leqslant \mathbb{E}_{\mathbf{x}}[\|\hat{\mathbf{h}}(\mathbf{x})\|\|\mathbf{h}^\star(\mathbf{x})\|] \leqslant O(M(K)^2)$ using the norm bound from Lemma 3.5. Hence under this setting we can choose c_1 sufficiently large so that $c_1 M(K)^2 \gtrsim \frac{M(K)^2}{c} c_2$. The condition $M(K) \gtrsim 1$ in the theorem statement is simply used to clean up the final bound.

In order to instantiate Theorem 3.3 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 3.8 from [47] (which holds for scalar-valued outputs). For convenience let $\mathrm{nn} = \frac{2D\sqrt{K+1+\log d}\cdot\Pi_{k=1}^{K}M(k)}{\sqrt{nt}}$. To bound this term we simply pull the summation over the rank r outside the complexity and apply Theorem 3.8, so

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

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 $\|\boldsymbol{\alpha}\|_2 \leqslant M(K)^2$ we obtain that $L(\mathcal{F}) = O(M(K)^2)$.

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

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

  Now by appealing to the norm bounds on the feature map from Lemma 3.5 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,

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

  since the expectation is once again trivial.

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

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

  from which the conclusion follows.

Finally we verify that Assumption 3.1 holds so as to use Theorem 3.3 to instantiate the end-to-end guarantee. The boundedness parameter is,

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

by Lemma 3.5 since it must be instantiated with $\boldsymbol{\alpha} \in \mathcal{F}$. For the ℓ_2 loss bounds, $\ell(\eta; y) = (y-\eta)^2$. Since $\nabla_\eta \ell(\eta; y) = 2(y-\eta) \leqslant O(N+|\eta|) = O(M(K)^3)$ where $|\eta| \leqslant |\boldsymbol{\alpha}^\top\mathbf{h}(\mathbf{x})| \leqslant O(M(K)^3)$ for $\boldsymbol{\alpha} \in \mathcal{F}$, $\mathbf{h} \in \mathcal{H}$ by Lemma 3.5 and $N = O(1)$. So it follows the loss is Lipschitz with $L = O(M(K)^3)$. Moreover by an analogous argument, $|\ell(\eta; y)| \leqslant O(M(K)^6)$ 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.

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

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

where $\mathrm{nn}=\frac{2D\sqrt{K+1+\log d}\cdot\Pi_{k=1}^{K}M(k)}{\sqrt{nt}}$. 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 3.1) and activation outputs for various neural networks.

Lemma 3.5. *Let $\hat{\mathbf{h}}(\mathbf{x})$ be a vector-valued neural network of depth K taking the form in (3.7) with each $f_j\equiv\boldsymbol{\alpha}_j$ satisfying $\|\boldsymbol{\alpha}_j\|\leqslant A$ with bounded data $\|\mathbf{x}\|\leqslant D$. Then the boundedness parameter in the setting of Theorem 3.1 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,

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

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

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

Now, for the inductive step, $\|\mathbf{r}_K\|^2=\|\mathbf{W}_K\sigma(\mathbf{W}_{k-1}\mathbf{r}_{K-1})\|^2\leqslant\|\mathbf{W}_K\|_2^2\|\sigma(\mathbf{W}_{K-1}\mathbf{r}_{K-1})\|^2\leqslant\|\mathbf{W}_K\|_2^2\|\mathbf{W}_{K-1}\mathbf{r}_{K-1}\|^2\leqslant\|\mathbf{W}_K\|_2^2\|\mathbf{W}_{K-1}\|_2^2\|\mathbf{r}_{K-1}\|^2$ where the first inequality follows because $\sigma(\cdot)$ is element-wise 1-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) then we can obtain the following result by simply bounding the last layer by noting that $\|\mathbf{r}_{K-1}\|_\infty\leqslant 1$. Then,

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

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

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 (3.8). 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,

$$y_{ji} = (\boldsymbol{\alpha}_j^\star)^\top \mathbf{h}^\star(\mathbf{x}_{ji}) + \eta_{ji} \text{ for } j \in \{1, \dots, t\}, i \in \{1, \dots, n\} \tag{3.12}$$

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 $j \in \{1, \dots, t\}$. As before we define the population task diversity matrix as $\mathbf{A} = (\boldsymbol{\alpha}_1^\star, \dots, \boldsymbol{\alpha}_t^\star)^\top \in \mathbb{R}^{t \times r}$, $\mathbf{C} = \mathbf{A}^\top \mathbf{A}/t$ 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 such that $\mathcal{F}$ and $\mathcal{F}_0$ contain the true parameters $\boldsymbol{\alpha}_0^\star$ and $\boldsymbol{\alpha}_j^\star$ respectively for the following.

Lemma 3.6. *Consider the $\mathbb{P}_{y|\mathbf{x}}(\cdot)$ regression model defined in (3.12) with the loss function $\ell(\cdot, \cdot)$ taken as the squared ℓ_2 loss and let Assumption 3.3 hold. Then for this conditional generative model with $\mathcal{F} = \{\boldsymbol{\alpha} : \boldsymbol{\alpha} \in \mathbb{R}^r\}$ and $\mathcal{F}_0 = \{\boldsymbol{\alpha} : \|\boldsymbol{\alpha}\|_2 \leqslant c_2\}$ the model is $(\frac{\tilde{\nu}}{c_2}, 0)$ diverse in the sense of Definition 3.3 and,*

$$d_{\mathcal{F},\mathcal{F}_0}(\hat{\mathbf{h}}; \mathbf{h}^\star) = c_2 \cdot \sigma_1(\Lambda_{sc}(\hat{\mathbf{h}}, \mathbf{h}^\star)); \quad \bar{d}_{\mathcal{F},\mathbf{f}^\star}(\hat{\mathbf{h}}; \mathbf{h}^\star) = \operatorname{tr}(\Lambda_{sc}(\hat{\mathbf{h}}, \mathbf{h}^\star)\mathbf{C}).$$

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 $\sigma_r(\mathbb{E}_{\mathbf{x}}[\hat{\mathbf{h}}(\mathbf{x})\hat{\mathbf{h}}(\mathbf{x})^\top]) \geqslant c > 0$ and $\|\mathbb{E}_{\mathbf{x}}[\hat{\mathbf{h}}(\mathbf{x})\mathbf{h}^\star(\mathbf{x})^\top]\|_2 \leqslant C$, then if $\mathcal{F} = \{\boldsymbol{\alpha} : \|\boldsymbol{\alpha}\| \leqslant c_1\}$, $\mathcal{F}_0 = \{\boldsymbol{\alpha} : \|\boldsymbol{\alpha}\|_2 \leqslant c_2\}$ and $c_1 \geqslant \frac{C}{c} c_2$, the same conclusions hold for sufficiently large constants c_1, c_2.

Proof. We first bound the worst-case representation difference and then the task-averaged representation difference. 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,

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

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

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

  Recognizing the inner infima as the partial minimization of a convex quadratic form (see for example [12, 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 here 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 $\hat{\boldsymbol{\alpha}} = -\mathbf{F}^\dagger_{\hat{\mathbf{h}}\hat{\mathbf{h}}}\mathbf{F}_{\hat{\mathbf{h}}\mathbf{h}^\star}\boldsymbol{\alpha}$. Finally the suprema over $\boldsymbol{\alpha}$ can be computed using the variational characterization of the singular values.

$$\sup_{\boldsymbol{\alpha}_0:\|\boldsymbol{\alpha}_0\|_2\leqslant c_2} \boldsymbol{\alpha}_0^\top\Lambda_{sc}(\hat{\mathbf{h}},\mathbf{h}^\star)\boldsymbol{\alpha}_0 = c_2\cdot\sigma_1(\Lambda_{sc}(\hat{\mathbf{h}},\mathbf{h}^\star))$$

- 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 $\text{tr}(\Lambda_{sc}(\hat{\mathbf{h}},\mathbf{h}^\star)\mathbf{C}) \geqslant \sum_{i=1}^{r}\sigma_i(\Lambda_{sc}(\hat{\mathbf{h}},\mathbf{h}^\star))\sigma_{r-i+1}(\mathbf{C}) \geqslant \text{tr}(\Lambda_{sc}(\hat{\mathbf{h}},\mathbf{h}^\star))\sigma_r(\mathbf{C}) \geqslant \sigma_1(\Lambda_{sc}(\hat{\mathbf{h}},\mathbf{h}^\star))\cdot\sigma_r(\mathbf{C})$.

Combining the above two results we can immediately conclude that,

$$d_{\mathcal{F},\mathcal{F}_0}(\hat{\mathbf{h}};\mathbf{h}^\star) = c_2\sigma_1(\Lambda_{sc}(\hat{\mathbf{h}},\mathbf{h}^\star)) \leqslant \frac{1}{\tilde{\nu}/c_2}\bar{d}_{\mathcal{F},\mathbf{f}^\star}(\hat{\mathbf{h}};\mathbf{h}^\star)$$

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 $\|\hat{\boldsymbol{\alpha}}\| \leqslant \frac{C}{c}\|\boldsymbol{\alpha}\|$. Thus, if we consider the convex quadratically-constrained quadratic optimization problem

$\inf_{\hat{\boldsymbol{\alpha}}:\|\hat{\boldsymbol{\alpha}}\|_2 \leqslant c_0} \mathbb{E}_{\mathbf{x}}|\hat{\mathbf{h}}(\mathbf{x})^\top \hat{\boldsymbol{\alpha}} - \mathbf{h}^\star(\mathbf{x})^\top \boldsymbol{\alpha}|^2$ and $c_0 \geqslant \frac{C}{c}\|\boldsymbol{\alpha}\|$ the constraint is inactive, and the constrained optimization problem is equivalent to the unconstrained optimization problem. Hence for the choices of $\mathcal{F} = \{\boldsymbol{\alpha} : \|\boldsymbol{\alpha}\| \leqslant c_1\}$ and $\mathcal{F}_0 = \{\boldsymbol{\alpha} : \|\boldsymbol{\alpha}\| \leqslant c_2\}$, since all the $\|\boldsymbol{\alpha}_j^\star\| \leqslant O(1)$ for $j \in [t] \cup \{0\}$, the infima in both the computation of the task-averaged distance and worst-case representation difference can be taken to be unconstrained for sufficiently large c_1, c_2. The second conclusion follows. □

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 3.6. 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 3.7 to verify the task diversity condition is satisfied with parameters $(\tilde{\nu}, \tilde{\epsilon})$ for $\tilde{\nu} \geqslant \frac{1}{t}$. Note in fact we have the stronger guarantee $\tilde{\nu} \geqslant \frac{\|\mathbf{v}\|_1}{\|\mathbf{v}\|_\infty}\frac{1}{t}$ 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 3.3 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,

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

  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 [113, Theorem 5.22]) and the metric entropy estimate from [59, Lemma 6(i)]. First note that $N_{2,\mathbf{b}\mathbf{X}_j}(\mathcal{F}, d_{2,\mathbf{b}\mathbf{X}_j}, \epsilon) \leqslant N(\mathcal{F}, \|\cdot\|_\infty, \epsilon)$, where the latter term refers to the covering number in the absolute sup-norm. By [59, Lemma 6(i)], $N(\mathcal{F}, \|\cdot\|_\infty, \epsilon) \leqslant \frac{1}{\epsilon}2^{2DW/\epsilon}$. So for all $0 \leqslant \epsilon \leqslant 1$,

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

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

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

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

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

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

Finally we verify that Assumption 3.1 holds so as to use Theorem 3.3 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,

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

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

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

$$y = f(\mathbf{h}(\mathbf{x})) + \eta. \tag{3.13}$$

which encompasses the class of multi-index models.

Lemma 3.7. *In the aforementioned setting and consider the $\mathbb{P}_{y|\mathbf{x}}(\cdot)$ regression model defined in (3.13). If $\mathcal{F}$ is a convex function class, and $\mathcal{F}_0 = \mathcal{F}_{t,\tilde{\epsilon}}$ the model is $(\tilde{\nu}, \tilde{\epsilon})$ diverse in the sense of Definition 3.3 for $\tilde{\nu} \geqslant \frac{1}{t}$.*

Proof. This result follows quickly from several properties of convex functions. We will use the pair $(\mathbf{x}, y)$ to refer to samples drawn from the generative model in (3.13); that is $\mathbf{x} \sim \mathbb{P}_{\mathbf{x}}(\cdot), y \sim \mathbb{P}_{y|\mathbf{x}}(f \circ \mathbf{h}(\mathbf{x}))$. First the mapping

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

is a jointly convex function of $(f, \hat{f})$. This follows since first 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 second 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 $\sup_z |f(z) - \tilde{f}(z)| \leqslant \tilde{\epsilon}$. Thus for all f we have that,

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

for some $\tilde{f} \in \mathcal{F}_t$. 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,

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

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

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

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,

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

To relate the worst-case and task-averaged representation differences, recall for a t-dimensional vector $\mathbf{v}$, $\|\mathbf{v}\|_\infty \leqslant \|\mathbf{v}\|_1$. 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],

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

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

Note if $\mathbf{v}$ is well-spread out – intuitively the problem possesses a problem-dependent "uniformity" and the bound $\tilde{\nu}\geqslant\frac{1}{t}$ 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 3.7 is valid for *generic* functions and representations in addition to applying to a wide class of regression losses. In particular, all p-norms such $L(a,b)=\|a-b\|_p$ satisfy the requisite conditions. 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 $\bar{d}_{\mathcal{F},\mathbf{f}^\star}(\hat{\mathbf{h}};\mathbf{h}^\star)$.

Chapter 4

Optimal Mean Estimation without Variance

4.1 Introduction

In this chapter, we aim to solve a fundamental problem in statistical inference—mean estimation—under minimal assumptions. Formally, we seek the tightest confidence interval (up to constants) achievable for the problem of mean estimation, equipped solely with a weak moment assumption on the X_i (say, when X_i are drawn from a multivariate t-distribution):

Problem 4.1. Consider a sequence of n i.i.d. vectors from a distribution $\mathcal{D}$ over $\mathbb{R}^d$ with mean μ satisfying the following weak moment condition for some $0 \leqslant \alpha \leqslant 1$:

$$\forall v \in \mathbb{R}^d, \|v\| = 1 : \ \mathbb{E}\left[|\langle v, X_i - \mu\rangle|^{1+\alpha}\right] \leqslant 1, \quad \forall i \in \{1, \ldots, n\}. \tag{MC}$$

Given a confidence level $\delta > 0$, output an estimate $\hat{\mu}$ with the smallest radius r_δ satisfying:

$$\mathbb{P}\left\{\|\hat{\mu} - \mu\| > r_\delta\right\} \leqslant \delta.$$

At first blush, such a question might seem only a theoretical curiosity. However, distributions lacking a variance (i.e. those with $\alpha < 1$) routinely arise in settings involving AB testing of user data and even reinforcement learning [106, 73, 45]. In these applications, the statistical consequences of Problem 4.1 are important to basic questions that often arise–such as how to effectively compute a treatment effect (the difference-in-means between a response variable in a heavy-tailed treatment group vs. a heavy-tailed control group) or a policy gradient (the mean across heavy-tailed stochastic gradients from a reinforcement learning simulator).

For the case of $\alpha = 1$, Problem 4.1 amounts to an assumption on the spectral norm of the covariance matrix of the distribution $\mathcal{D}$. Even in this special case, estimators with optimal confidence interval, r_δ, were only recently discovered ([75, 15]), building upon the

one-dimensional median-of-means (MoM) framework introduced in [89]. These estimators achieve the following rate:

$$r_\delta = O\left(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log 1/\delta}{n}}\right).$$

Unfortunately, however, there is no known polynomial-time algorithm to compute the estimators proposed in these papers. Computationally-efficient estimators achieving the optimal confidence interval were first proposed in [51] based on the sum-of-squares family of semidefinite relaxations of the estimator from [75]. By combining these ideas with a algorithm based on gradient descent, faster mean estimators were subsequently developed in [20]. Perhaps surprisingly, this line of work shows as long as the variance of the random vector exists, neither statistical nor computational efficiency is necessarily sacrificed when estimating μ. In particular, the dependence on d, n, δ of the confidence interval for MoM estimators, when the samples have bounded second moments, exactly matches the optimal dependence on d, n, δ, *when the samples X_i are Gaussian.*

When $\alpha < 1$, the situation is markedly different. In the one-dimensional case, the (optimal) achievable radius satisfies [29]:

$$r_\delta = O\left(\frac{\log 1/\delta}{n}\right)^{\frac{\alpha}{1+\alpha}},$$

which can be achieved by a univariate MoM-style estimator. Even in one dimension, the lack of a second moment degrades the information-theoretic bound with respect to both n and δ. Unlike the case $\alpha \geqslant 1$, Gaussian-like confidence intervals cannot be obtained in this regime. Moreover, in d dimensions, there is very little known about the optimal achievable radius save for the fact that one can obtain the following trivial rate by applying the univariate estimator coordinate-wise:[1]

$$r_\delta = \widetilde{O}\left(\sqrt{d}\left(\frac{\log 1/\delta}{n}\right)^{\frac{\alpha}{1+\alpha}}\right).$$

Two natural questions thus present themselves. First, in the regime where $\alpha < 1$, what is minimax-optimal rate for mean estimation in higher dimensions? Second, can this (hitherto unknown) rate also be achieved in polynomial time?

The primary contribution of the current paper is to present to sharp answers to both of these questions. These answers are contained in the following two theorems, the first of which presents a rate that is achievable by a polynomial-time algorithm and the second of which establishes the optimality of this rate.

Theorem 4.2. *Let $\boldsymbol{X} = X_1, \ldots, X_n$ be iid random vectors with mean μ, satisfying the weak moment assumption (MC) for some known $\alpha > 0$. There is a polynomial-time algorithm*

[1] This follows by the triangle inequality and union bound.

which, when given inputs $\boldsymbol{X}$ *and a target confidence* $\delta > 2^{-\frac{n}{16000}}$*, returns a point* x^* *satisfying:*

$$\|x^* - \mu\| \leqslant 10^8 \left(\sqrt{\frac{d}{n}} + \left(\frac{d}{n}\right)^{\frac{\alpha}{1+\alpha}} + \left(\frac{\log 1/\delta}{n}\right)^{\frac{\alpha}{1+\alpha}} \right),$$

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

Theorem 4.2 is based on a two-stage estimation procedure. In the first step, the set of inputs $\boldsymbol{X}$ is truncated to discard samples that are too far from the empirical centroid of $\boldsymbol{X}$. The second step uses an estimation-to-testing framework for heavy-tailed estimation [20] by first setting up a testing problem which decides if a candidate mean is close to the true mean μ and, subsequently, using this procedure to improve the estimate. By iterating this procedure, we eventually converge to a good approximation of the true mean.

Our second main result is a matching minimax lower bound establishing the optimality of the rate in Theorem 4.2.

Theorem 4.3. *Let* $n > 0$ *and let* $\delta \in \left(e^{-\frac{n}{4}}, \frac{1}{4}\right)$*. Then there exists a set of distributions* $\mathcal{F}$ *over* $\mathbb{R}^d$ *such that each* $\mathcal{D} \in \mathcal{F}$ *obeys the weak moment condition (MC) for some* $\alpha > 0$*, and any estimator* $\hat{\mu}$ *satisfies:*

$$\mathbb{P}_{\mathcal{D}\in\mathcal{F}} \left\{ \|\hat{\mu}(\boldsymbol{X}) - \mu(\mathcal{D})\| \geqslant \frac{1}{24} \cdot \max\left(\left(\frac{d}{n}\right)^{\frac{\alpha}{1+\alpha}}, \sqrt{\frac{d}{n}}, \left(\frac{\log 2/\delta}{n}\right)^{\frac{\alpha}{1+\alpha}} \right) \right\} \geqslant \delta,$$

where the data $\boldsymbol{X}$ *are generated iid from* $\mathcal{D}$*.*

The main challenge in proving Theorem 4.3 lies in obtaining tight dependence on dimension. The proof begins by using a standard reduction from estimation to testing for proving minimax rates [see, for example, 113, Chapter 15]. We then avoid traditional Fano-style information-theoretic approaches, however, in establishing difficulty of the testing problem. We instead take a Bayesian approach and use a carefully chosen set of discrete distributions to instantiate the testing problem, allowing us to establish a sharp dependence on dimension in our lower bound, once various technical challenges are surmounted.

Together, Theorem 4.2 and Theorem 4.3 have the following implications for the problem of mean estimation without a variance:

- In the case in which the failure probability δ is a constant, our upper and lower bounds simplify to $O(\sqrt{d/n} + (d/n)^{\alpha/(1+\alpha)})$. Interestingly, Theorem 4.2 and Theorem 4.3 reveal the existence of a phase transition in the estimation rate when $n \asymp d$—the estimation rate is dominated by $\sqrt{\frac{d}{n}}$ when $n \lesssim d$ and $(\frac{d}{n})^{\alpha/(1+\alpha)}$ when $n \gtrsim d$.
- While it is established in [29] that it is impossible to obtain subgaussian rates in this setting even in one dimension, our results reveal a decoupling between the terms depending on the failure probability and the dimension that parallels the finite-variance case (where $\alpha = 1$).

- Finally, our results also extend to the more general problem of mean estimation under adversarial corruption, which has received much recent attention in both the computer science and statistics communities. In this setting, an adversary is allowed to inspect the data points and *arbitrarily* corrupt a fraction η of them. We recover the mean up to an error of $O(\eta^{\alpha/(1+\alpha)})$ which is information-theoretically optimal (Theorem 4.8). Furthermore, as a consequence of Theorem 4.3, our sample complexity of $O(d/\eta)$ is also optimal.

The main technical challenge in establishing our upper bound is the analysis of the estimation-to-testing framework of [20] in the weak-moment setting. The analysis in [20] makes critical use of the decomposition of the variance of sums of independent random variables which does not hold in our setting. This allows tight control of the *second* moments of $\sum_{i=1}^{m} X_i$ and $\|X - \mu\|$, which are crucial to that analysis. Despite the lack of such decompositions for weak moments, we establish tight control over the appropriate quantities allowing us to establish our optimal recovery guarantees.

Similarly, the presence of weak moments also complicates the task of establishing a matching lower bound with tight dependence on the dimension d. The main difficulty is in proving the optimality of the dimension-dependent term, $(d/n)^{\alpha/(1+\alpha)}$. For the specific case where $\alpha = 1$, the lower bound may be proved within the estimation-to-testing framework by utilizing a distribution over a well-separated collection of Gaussian distributions. However, this approach fails for the weak-moment mean estimation problem; indeed, hypercontractivity properties of Gaussian distributions ensure a bounded variance leading to a lower bound that scales as $1/\sqrt{n}$ as opposed to the slower rate $n^{-\alpha/(1+\alpha)}$. To prove our lower bound, we instead use a collection of carefully chosen distributions with discrete supports whose means are separated by $O((d/n)^{\alpha/(1+\alpha)})$. Further challenges arise at this point—if we follow the standard path of bounding the complexity of the testing problem in terms of pairwise f-divergences between distributions in the hypothesis set, we obtain vacuous bounds. We instead directly analyze the posterior distribution obtained from the framework and show that random independent samples from the posterior tend to be well separated, yielding our tight lower bound.

4.2 Related Work

There has been much interest in designing information-theoretically optimal estimators for fundamental inferential tasks under minimal assumptions on the distributions generating the data [89, 57, 1, 75, 78, 51, 20, 21, 76, 28, 69, 26, 27, 77]. In the one-dimensional setting, estimators achieving the information theoretically-optimal *subgaussian* rate were obtained in the seminal work of [1, 57, 89]. In recent years, focus has shifted towards the high-dimensional setting where one aims for optimal recovery error in terms of the number of samples n, the dimension d, and the failure probability δ, without making strong distributional assumptions such as Gaussianity. As a consequence of this effort, information-theoretically optimal

estimators have been developed for mean estimation [75], linear regression [78] and covariance estimation [85]. However, the estimators proposed in these works lack computationally efficient algorithms to compute them. The first computationally efficient estimator was proposed by [51] and its runtime and analysis were subsequently improved in [20, 28, 69]. Subsequently, improved algorithms have been devised for linear regression and covariance estimation [21]. Most recently, [26] extended the approach of [69] to linear regression, improving on [21] in settings where the covariance matrix of the data-generating distribution is known. We direct the interested reader to the survey [77] and the references therein for more detailed discussion of this line of research. Note that the optimal recovery guarantee obtainable in all these settings is the subgaussian rate. This is provably not possible in the weak-moment scenario even in the one-dimensional case as evidenced by our lower bound.

Another approach towards achieving distributional robustness which has received much attention in the computer science community is robust estimation under a contamination model. In this setting, an adversary is allowed to inspect a set of data points generated from a well-behaved distribution and can arbitrarily corrupt a fraction η of them according to their choosing. Broadly speaking, the primary goal of this field is to obtain optimal recovery of the underlying parameter as a function of the corruption factor η as opposed to achieving optimal dependence on n, d, δ. Starting with the foundational works of [56, 109], which obtain information-theoretically optimal (albeit computationally intractable) estimators, computationally efficient estimators are now known for a range of statistical estimation problems in various settings [64, 35, 17, 103, 34, 32, 63, 36, 39, 19]. Since the literature of this field is vast, we restrict ourselves to the specific setting of mean estimation and direct the reader to [30] for more context on these developments. For the mean estimation problem under adversarial corruption, this line of work has resulted in estimators which succeed with constant probability, say $2/3$, and achieve the optimal recovery error of $\sqrt{\eta}$ assuming the data is drawn from a distribution with finite covariance [39, 19, 28]. In addition, [76, 28] achieve this rate along with the optimal dependence on δ in the recovery guarantees. Finally, recent work [31, 53] has led to a formal unification of algorithmic approaches towards each of these settings along with extensions of these approaches to the differentially private setting as well [52]. A corollary of our work extends these results to the setting where the covariance matrix is not defined.

4.3 Algorithm

In this section, we describe our algorithm for mean estimation problem in the setting of the weak moment condition (MC). We build on the approach of [20] which operates in the setting where the covariance matrix of the distribution generating the data is defined. However, the absence of second moments complicates the design of our algorithm leading to a more intricate procedure. Concretely, our algorithm is comprised of the following three broad stages:

1. Data Pruning: First, we compute an initial crude estimate of the mean which is within

$O(\sqrt{d})$ of μ and then proceed to use this estimate to filter out data points which are far from our estimate. This truncation must be chosen carefully to ensure that the mean of the truncated data points still approximates the true mean well. As a consequence, this truncation is a function of n, d and α. Algorithms 4 and 5 describe this crude estimation procedure and truncation step in greater detail. Intriguingly, this additional thresholding step (which is not necessary in the $\alpha = 1$ case) is critical to achieve the statistically optimal confidence intervals in this scenario.

2. Data Batching: In this stage, the data points that survive the truncation procedure in Algorithm 5 are then divided into k bins and mean estimates are computed by averaging the set of points in each bin. The number of bins is chosen depending on the desired failure probability, δ. The precise setting of parameters is described in Algorithm 3.

3. Median Computation: Finally, the bucket estimates, Z_i, produced in the previous stage are aggregated to produce our final estimate of the mean. The procedure to do this follows along the testing-to-estimation framework for robust estimation explored in [20]. Here, one first designs a procedure to *test* whether a given candidate, x, is close to μ. Subsequently, a solution to the testing program is then used to improve the estimate. In this setting, one shows that the testing program can be used to estimate both $\|x - \mu\|$ and an approximation to $\Delta = \frac{\mu - x}{\|x - \mu\|}$ which may be used to improve our estimate. Algorithms 6 and 7 and Algorithm 2 display the estimation and tuning components of this stage. The testing program is defined in **MT** and is discussed in more detail subsequently.

Algorithm 1 Mean Estimation

Input: Data Points $\boldsymbol{X} \in \mathbb{R}^{n \times d}$, Target Confidence δ
$x^{\dagger} \leftarrow$ Initial Mean Estimate($\{X_1, \ldots, X_{n/2}\}$)
$\boldsymbol{Z} \leftarrow$ Produce Bucket Estimates($\left\{X_{n/2+1}, \ldots, X_n\right\}, x^{\dagger}, \delta$)
$T \leftarrow 10^6 \log dn$
$x^* =$ Gradient Descent($\boldsymbol{Z}, x^{\dagger}, T$)
Return: x^*

Algorithm 2 Gradient Descent

Input: Bucket Means $\boldsymbol{Z} \in \mathbb{R}^{k \times d}$, Initialization $x^\dagger$, Number of Iterations T
$x^*, x_0 \leftarrow x^\dagger$ and $d^*, d_0 \leftarrow \infty$
for $t = 0 : T$ **do**
  $d_t \leftarrow$ Distance Estimation$(\boldsymbol{Z}, x_t)$
  $g_t \leftarrow$ Gradient Estimation$(\boldsymbol{Z}, x_t)$
  if $d_t < d^*$ **then**
    $x^* \leftarrow x_t$
    $d^* \leftarrow d_t$
  $x_{t+1} \leftarrow x_t + \frac{1}{20} d_t g_t$
Return: x^*

Algorithm 3 Produce Bucket Estimates

Input: Data Points $\boldsymbol{X} \in \mathbb{R}^{n \times d}$, Mean Estimate $x^\dagger$, Target Confidence δ
$\boldsymbol{Y} \leftarrow$ Prune Data$(\boldsymbol{X}, x^\dagger)$
$m \leftarrow |\boldsymbol{Y}|$
$k \leftarrow 4000 \log 1/\delta$
Split data points into k buckets with bucket $\mathcal{B}_i$ consisting of the points $X_{(i-1)\frac{m}{k}+1}, \ldots, X_{i\frac{m}{k}}$
$Z_i \leftarrow \text{Mean}(\mathcal{B}_i)\ \forall\ i \in [k]$ and $\boldsymbol{Z} \leftarrow (Z_1, \ldots, Z_k)$
Return: $\boldsymbol{Z}$

Algorithm 4 Initial Mean Estimate

Input: Set of data points $\boldsymbol{X} = \{X_i\}_{i=1}^n$
$\hat{\mu} \leftarrow \arg\min_{X_i \in \boldsymbol{X}} \min \left\{ r > 0 : \sum_{j=1}^n \mathbf{1}\left\{\|X_j - X_i\| \leqslant r\right\} \geqslant 0.6n \right\}$
Return: $\hat{\mu}$

Algorithm 5 Prune Data

Input: Set of data points $\boldsymbol{X} = \{X_i\}_{i=1}^n$, Mean Estimate $x^\dagger$
$\tau \leftarrow \max\left(100 n^{\frac{1}{1+\alpha}} d^{-\frac{(1-\alpha)}{2(1+\alpha)}}, 100\sqrt{d}\right)$
$\mathcal{C} \leftarrow \{X_i : \|X_i - x^\dagger\| \leqslant \tau\}$
Return: $\mathcal{C}$

Algorithm 6 Distance Estimation

Input: Data Points $\boldsymbol{Z} \in \mathbb{R}^{k \times d}$, Current point x
$d^* = \arg\max_{r>0} \mathbf{MT}(x, r, \boldsymbol{Z}) \geqslant 0.9k$
Return: d^*

Algorithm 7 Gradient Estimation

Input: Data Points $\boldsymbol{Z} \in \mathbb{R}^{k \times d}$, Current point x
$d^* =$ Distance Estimation$(\boldsymbol{Z}, x)$
$(v, X) = \mathbf{MT}(x, d^*, \boldsymbol{Z})$
$g \leftarrow$ Top Singular Vector(X_v)
Return: g

As in [51, 20], the following polynomial optimization problem and its semidefinite relaxation play a key role in our subsequent analysis. Intuitively, given a test point, x, the program searches for a direction (denoted by v) such that a large fraction of the bucket estimates, Z_i, are far away from x along v. Formally, the polynomial optimization problem, parameterized by x, r and $\boldsymbol{Z}$, is defined below:

$$\begin{aligned} \max \sum_{i=1}^{k} b_i \\ \text{s.t } b_i^2 = b_i \\ \|v\|^2 = 1 \\ b_i(\langle v, Z_i - x\rangle - r) \geqslant 0 \ \forall i \in [k]. \end{aligned} \tag{MTE}$$

The binary variables b_i indicates whether the i^{th} bucket mean Z_i is far away along v. Unfortunately, the binary constraints on b_i, the restriction of v and the final constraint make this problem nonconvex and there are no efficient algorithms known to compute it. Accordingly, we work with the semidefinite relaxation defined as follows:

$$\begin{aligned} \max \sum_{i=1}^{k} X_{1,b_i} \\ X_{1,b_i} = X_{b_i,b_i} \\ \sum_{j=1}^{d} X_{v_j,v_j} = 1 \\ \langle v_{b_i}, Z_i - x\rangle \geqslant X_{b_i,b_i} r \ \forall i \in [k] \\ X_{1,1} = 1 \\ X \succcurlyeq 0, \end{aligned} \tag{MT}$$

where $v_{b_i} = [X_{b_i,v_1}, \ldots, X_{b_i,v_d}]^\top$. The matrix $X \in \mathbb{S}_+^{(k+d+1)}$ is symbolically indexed by 1 and the variables $b_1, \ldots, b_k$ and $v_1, \ldots, v_d$. We will restrict ourselves to analyzing **MT** and will refer to the program initialized with x, r and $\boldsymbol{Z}$ as $\mathbf{MT}(x, r, \boldsymbol{Z})$. We will use $(v, X) = \mathbf{MT}(x, r, \boldsymbol{Z})$ to denote the optimal value, v, and solution, X, of **MT** initialized with x, r and $\boldsymbol{Z}$. For the

sake of clarity, $\mathbf{MT}(x, r, \boldsymbol{Z})$ in the absence of any specified output will refer to the optimal value of $\mathbf{MT}(x, r, \boldsymbol{Z})$.

4.4 Proof Overview

In this section, we outline the main steps in proving Theorem 4.2, providing full details in the Appendix. For ease of exposition, we restrict most of our attention to those steps of our proof which are complicated by the weaker assumptions used in our work. From Section 4.3, we see that our estimation procedure is divided into three stages:

1. We obtain an initial coarse estimate, $\hat{\mu}$, of μ (Algorithm 4);
2. We then use $\hat{\mu}$ to prune data points far away from μ (Algorithm 5);
3. Finally, the remaining data points are incorporated into a gradient-descent algorithm to obtain our final estimate (Algorithm 1).

To obtain our tight rates, we crucially require the following correctness guarantees on these three steps, each with high probability: our initial estimate $\hat{\mu}$ is within a radius of $O(\sqrt{d})$ of μ, a large fraction of data points pass the pruning steps in Algorithm 5, and finally, a tight analysis on the error of the gradient descent procedure in Algorithm 1. The first two steps are novel to the weak-moment setting and the third step, while explored previously for the case $\alpha = 1$, is complicated here due to the lack of strong decomposition structure in the weak moments.

To deal with these difficulties, we establish two crucial structural lemmas, proved in Section 4.6, on distributions satisfying weak-moment conditions. The first is a bound on the $1 + \alpha$ moments of the lengths of such random vectors:

Lemma 4.1. *Let X be a zero-mean random vector satisfying the weak-moment assumption for some $0 \leqslant \alpha \leqslant 1$. We have the following bound:*

$$\mathbb{E}[\|X\|^{1+\alpha}] \leqslant \frac{\pi}{2} \cdot d^{\frac{1+\alpha}{2}}.$$

As we will see, this lemma is crucial in all three steps of our analysis. Note that the upper bound obtained by the lemma is tight up to a small constant factor. (A standard Gaussian random vector yields an upper bound of $d^{(1+\alpha)/\alpha}$). The proof follows by first considering independent random Gaussian projections of the random vector along with an application of Jensen's inequality.

The second key lemma is a bound on the $1 + \alpha$ moments of the sum of random variables satisfying weak-moment assumptions. This lemma plays a key role in obtaining tight bounds on the accuracy of the gradient-descent procedure in Algorithm 1:

Lemma 4.2. *Let $X_1, \ldots, X_n$ be n mean-zero i.i.d. random variables satisfying the following bound, for some $0 \leqslant \alpha \leqslant 1$:*

$$\mathbb{E}[|X_i|^{1+\alpha}] \leqslant 1. \tag{4.1}$$

We have:

$$\mathbb{E}\left[|\sum_{i=1}^{n} X_i|^{1+\alpha}\right] \leqslant 2n. \tag{4.2}$$

The proof of this lemma uses techniques employed previously for establishing the Nemirovskii inequalities for Banach spaces but crucially hold even when the covariance of the matrices are not defined [88]. We now sketch the argument establishing guarantees on the first two steps of the algorithm. Firstly, from Lemma 4.1, we have that most of the sample sample data points are within a radius of $O(\sqrt{d})$ of μ with high probability. Therefore, the value of the minimizer in Line 2 of Algorithm 4 is $O(\sqrt{d})$ (by simply picking any of the data points close to μ) and, furthermore, at least one of the points within $O(\sqrt{d})$ of $\hat{\mu}$ must be at a distance at most $O(\sqrt{d})$ from μ as most data points are close to μ. This establishes the required correctness guarantees on Algorithm 4 which is formalized in Lemma 4.11. For the second step, we condition on the success of the first step and note that our threshold, τ, is chosen such that all points within $O(\sqrt{d})$ of μ are within τ of $\hat{\mu}$. Another application of Lemma 4.1 now ensures that most data points pass this threshold, establishing correctness for the second step of our procedure. This is outlined in Lemma 4.15. We devote the following subsection to the final and most technical step in our analysis.

Gradient Descent Analysis

In this section we sketch the main steps in the analysis of the gradient-descent procedure used in Algorithm 1. Throughout this subsection, we assume that the previous two steps of the procedure are successful; that is, $\hat{\mu}$ constructed as part of Algorithm 5 is within $O(\sqrt{d})$ of μ and as a consequence at least $\Omega(n)$ points are used to construct the bucket estimates. Now, let $\{Z_i\}_{i=1}^{k}$ denote the bucket estimates produced as part of Algorithm 3 and let $\widetilde{\mu} = \mathbb{E}[Z_i]$. From prior work [20], the estimate returned by Algorithm 1 is within a radius of $O(r^*)$ of $\widetilde{\mu}$, where:

$$r^* \coloneqq \min\left\{r > 0 : \mathbf{MT}(\widetilde{\mu}, r, \boldsymbol{Z}) \leqslant 0.05k\right\}.$$

Therefore, the error of our estimate may be upper bounded by the sum of two terms: the first is the degree to which $\widetilde{\mu}$ approximates μ and the second is an upper bound on r^*. We will see that there is an inherent tradeoff between these two terms—by picking the threshold τ in Algorithm 5 to be extremely large, $\widetilde{\mu}$ may be an arbitrarily good approximation of μ but our bound on r^* may be poor. We now state a structural lemma capturing the tradeoff between r^* and τ.

Lemma 4.3. *Let $\boldsymbol{Z} = \{Z_i\}_{i=1}^k$ be k iid random vectors with mean $\widetilde{\mu}$ and covariance matrix Λ. Suppose that $\mathbb{E}\left[|\langle v, Z_i - \widetilde{\mu}\rangle|^{1+\alpha}\right] \leqslant \beta$ for all $\|v\| = 1$. We have:*

$$r^* \leqslant 1000\left(\sqrt{\frac{\operatorname{Tr}\Lambda}{k}} + \beta^{1/(1+\alpha)}\right) \text{ where } r^* \coloneqq \min\{r > 0 : \boldsymbol{MT}(\widetilde{\mu}, r, \boldsymbol{Z}) \leqslant 0.05k\},$$

with probability at least $1 - e^{-k/800}$.

Given this lemma, the main remaining difficulty is in obtaining bounds on $\mathbb{E}\left[|\langle Z_i, v\rangle|^{1+\alpha}\right]$, $\operatorname{Tr}(\Lambda)$, and the deviation of $\widetilde{\mu}$ from μ. Note that from the definition of Algorithm 3, the Z_i are means of truncated data points and hence their covariance matrix is well defined.

To obtain bounds on $\|\widetilde{\mu} - \mu\|$, an application of Markov's inequality and Lemma 4.1 establishes that the probability of X_i being truncated in Algorithm 5 is at most $O(d/n)$. Then, a standard variational argument along with our weak-moment assumption allows us to bound $\|\widetilde{\mu} - \mu\|$ (see Lemma 4.9 for more details).

We obtain a bound on $\operatorname{Tr}\Lambda$ by first observing that each X_i used to compute one of the bucket estimates, Z_j, is truncated with respect to its distance from $\hat{\mu}$. By an application of the triangle inequality, we infer that all of the X_i used in the computation satisfy $\|X_i - \hat{\mu}\| \leqslant \tau + O(\sqrt{d})$. Therefore, to bound $\operatorname{Tr}\Lambda$, all we need is a bound on $\mathbb{E}[\|X_i - \widetilde{\mu}\|^2 \mathbf{1}\left\{\|X_i - \hat{\mu}\| \leqslant \tau + O(\sqrt{d})\right\}]$. Another appeal to Lemma 4.1 and a straightforward truncation argument establishes a bound of $O(\sqrt{d/n} + (d/n)^{\alpha/(1+\alpha)})$ (see Lemmas 4.10 and 4.15).

For the final term, observe that the Z_i are averages of truncated versions of X_i. A simple argument shows that truncated X_i satisfy a weak-moment bound (Lemma 4.9). Therefore, a direct application of Lemma 4.2 yields a bound of $O(k/n)$ on β. Incorporating these bounds into Lemma 4.3 and the gradient-descent framework for heavy-tailed estimation from [20] concludes the proof of Theorem 4.2.

4.5 Lower Bound

In this section, we present a lower bound for heavy-tailed regression which shows that the bound obtained in Theorem 4.2 is tight.

For a given dimension d, and sample size n, we will consider a family of distributions parameterized by size $d/2$ subsets of $[d]$. That is, we will consider a family of distributions $\mathcal{F} = \{\mathcal{D}_S : S \subset [d] \text{ and } |S| = d/2\}$. Now, for each particular distribution $\mathcal{D}_S$, we have $X \sim \mathcal{D}_S$ as follows:

$$X = \begin{cases} 0, & \text{with probability } 1 - \frac{d}{8n} \\ n^{\frac{1}{1+\alpha}} \cdot d^{-\frac{(1-\alpha)}{2(1+\alpha)}} \cdot \boldsymbol{e_i}, & \text{for } i \in S \text{ with probability } \frac{1}{4n}. \end{cases}$$

We will first show that the distribution $\mathcal{D}_S$ satisfies the $1 + \alpha$ moment condition.

Lemma 4.4. *Let $X \sim \mathcal{D}_S$ for some $S \subset [d]$ such that $|S| = d/2$. Then, X satisfies the following:*

$$\forall v : \|v\| = 1 : \mathbb{E}\left[|\langle v, X - \mu_S\rangle|^{1+\alpha}\right] \leqslant \frac{1}{2}.$$

Proof. We first note that:

$$(\mu_S)_i = \begin{cases} 0, & \text{for } i \notin S \\ \frac{n^{-\frac{\alpha}{1+\alpha}} \cdot d^{-\frac{(1-\alpha)}{2(1+\alpha)}}}{4}, & \text{otherwise.} \end{cases}$$

Let v be such that $\|v\| = 1$. We have:

$$\begin{aligned}\mathbb{E}\left[|\langle v, X - \mu_S\rangle|^{1+\alpha}\right] &= \sum_{i\in S} \frac{1}{4n} \cdot |v_i \left(n^{\frac{1}{1+\alpha}} \cdot d^{-\frac{(1-\alpha)}{2(1+\alpha)}} - \frac{n^{-\frac{\alpha}{1+\alpha}} \cdot d^{-\frac{(1-\alpha)}{2(1+\alpha)}}}{4}\right)|^{1+\alpha} \\ &\quad + \sum_{i\in S} \left(1 - \frac{1}{4n}\right) \cdot |v_i \left(\frac{n^{-\frac{\alpha}{1+\alpha}} \cdot d^{-\frac{(1-\alpha)}{2(1+\alpha)}}}{4}\right)|^{1+\alpha}.\end{aligned}$$

For the first term in this sum, we have:

$$\begin{aligned}\sum_{i\in S} \frac{1}{4n} \cdot |v_i \left(n^{\frac{1}{1+\alpha}} \cdot d^{-\frac{(1-\alpha)}{2(1+\alpha)}} - \frac{n^{-\frac{\alpha}{1+\alpha}} \cdot d^{-\frac{(1-\alpha)}{2(1+\alpha)}}}{4}\right)|^{1+\alpha} &\leqslant \sum_{i\in S} \frac{1}{4n} \cdot |v_i n^{\frac{1}{1+\alpha}} \cdot d^{-\frac{(1-\alpha)}{2(1+\alpha)}}|^{1+\alpha} \\ = \frac{1}{4}\sum_{i\in S} |v_i|^{1+\alpha} \cdot d^{-\frac{(1-\alpha)}{2}} &\leqslant \frac{1}{4}\left(\sum_{i\in S} v_i^2\right)^{\frac{1+\alpha}{2}} \cdot \left(\sum_{i\in S} d^{-1}\right)^{\frac{1-\alpha}{2}} \leqslant \frac{1}{4},\end{aligned}$$

where the second inequality follows from Hölder's inequality. For the second term, we have:

$$\begin{aligned}\sum_{i\in S} \left(1 - \frac{1}{4n}\right) \cdot |v_i \left(\frac{n^{-\frac{\alpha}{1+\alpha}} \cdot d^{-\frac{(1-\alpha)}{2(1+\alpha)}}}{4}\right)|^{1+\alpha} &\leqslant \frac{1}{4}\sum_{i\in S} |v_i|^{1+\alpha} n^{-\alpha} d^{-\frac{(1-\alpha)}{2}} \\ &\leqslant \frac{1}{4}\sum_{i\in S} |v_i|^{1+\alpha} d^{-\frac{(1-\alpha)}{2}} \leqslant \frac{1}{4},\end{aligned}$$

where the last inequality again follows from Hölder's inequality. Putting the two bounds together, we obtain:

$$\forall v : \|v\| = 1 : \mathbb{E}\left[|\langle v, X - \mu_S\rangle|^{1+\alpha}\right] \leqslant \frac{1}{2}.$$

□

We now prove a lemma that establishes the optimality of Theorem 4.2 in the regime of constant failure probability. We use the following generative process for the data $\boldsymbol{X} = X_1, \ldots, X_n$:

1. Randomly pick a subset S uniformly from the set $\{T \subset [d] : |T| = d/2\}$.
2. Generate $X_1, \ldots, X_n$ iid from the distribution, $\mathcal{D}_S$.

Lemma 4.5. *Let $(S, \boldsymbol{X})$ be generated according to the above process. We have, for any estimator $\hat{\mu}(\boldsymbol{X})$,*

$$\mathbb{P}_{S,\boldsymbol{X}}\left\{\|\hat{\mu}(\boldsymbol{X}) - \mu_S\| \geqslant \frac{1}{24} \cdot \left(\frac{d}{n}\right)^{\frac{\alpha}{1+\alpha}}\right\} \geqslant \frac{1}{4}.$$

Proof. We first define the random variable $Y \coloneqq \sum_{i=1}^{n} \mathbf{1}\{X_i \neq 0\}$. From the definition of the distributions $\mathcal{D}_S$ we have:

$$\mathbb{E}[Y] = \frac{d}{8}$$

Therefore, we have that $Y \leqslant d/4$ with probability at least $1/2$, by Markov's inequality. We now define the following random set: $T \coloneqq \{i \in [d] : \exists j \in [n] \text{ such that } (X_j)_i \neq 0\}$. We see from the definition of T and Y that $|T| \leqslant Y$. We have with probability at least $1/2$ that $|T| \leqslant d/4$. Let $\boldsymbol{X}$ be an outcome for which $|T| = k \leqslant d/4$. We have by the symmetry of the distribution that:

$$\mathbb{P}\{S|\boldsymbol{X}\} = \begin{cases} \frac{1}{\binom{d-k}{d/2-k}}, & \text{if } T \subset S \text{ and } |S| = d/2 \\ 0, & \text{otherwise.} \end{cases}$$

For given $\boldsymbol{X}$, define $Z_i = \mathbf{1}\{i \in S\}$ for $i \notin T$ (For $i \in T$, Z_i is 1). We have for Z_i and Z_j for distinct $i, j \notin T$:

$$\mathbb{E}[Z_i|\boldsymbol{X}] = \mathbb{E}[Z_j|\boldsymbol{X}] = \frac{d-2k}{2(d-k)}.$$

Furthermore, we have:

$$\begin{aligned} \operatorname{Cov}(Z_i, Z_j|\boldsymbol{X}) &= \frac{(d-2k)(d-2k-2)(d-k) - (d-2k)^2(d-k-1)}{4(d-k)^2(d-k-1)} \\ &= \frac{(d-2k)((d-2k)(d-k) - 2(d-k) - (d-2k)(d-k) + (d-2k))}{4(d-k)^2(d-k-1)} \\ &= \frac{-d(d-2k)}{4(d-k)^2(d-k-1)} < 0. \end{aligned}$$

Now, consider some $R \subset [d]$ such that $|R| = d/2$ and $T \subset R$. Let $Q = R \setminus T$. For Q, we have $|Q| = d/2 - k$. We have for S:

$$|S \cap R| = k + \sum_{i \in Q} Z_i.$$

This means that:

$$\operatorname{Var}(|S \cap R| \mid \boldsymbol{X}) = \operatorname{Var}\left(\sum_{i \in Q} Z_i \mid \boldsymbol{X}\right) \leqslant \sum_{i \in Q}\left(\frac{d-2k}{2(d-k)}\right)^2 \leqslant \frac{d/2-k}{4} = \frac{d}{8}.$$

Furthermore, we have that:

$$\mathbb{E}\left(|S \cap R| \mid \boldsymbol{X}\right) = k + \left(\frac{d}{2} - k\right) \cdot \frac{(d-2k)}{2(d-k)} \leqslant \frac{d}{4} + \frac{d}{4} \cdot \frac{d}{4(3d/4)} = \frac{d}{4} + \frac{d}{12} = \frac{d}{3}.$$

Therefore, we have by Chebyshev's inequality that:

$$\mathbb{P}\left\{|S \cap R| \geqslant \frac{5d}{12}\right\} \leqslant \frac{1}{2}.$$

Note that for any S_1, S_2 such that $|S_i| = \frac{d}{2}$ and $|S_1 \cap S_2| \leqslant \frac{5d}{12}$, we have:

$$\|\mu_{S_1} - \mu_{S_2}\| \geqslant \sqrt{2 \cdot \frac{d}{12} \cdot \left(\frac{n^{-\frac{\alpha}{1+\alpha}} \cdot d^{-\frac{1-\alpha}{2(1+\alpha)}}}{4}\right)^2} \geqslant \frac{1}{12} \cdot \left(\frac{d}{n}\right)^{\frac{\alpha}{1+\alpha}}.$$

Consider any estimator $\hat{\mu}$. Suppose that there exists R such that $T \subset R$, $|R| = d/2$ and $\|\hat{\mu}(\boldsymbol{X}) - \mu_R\| \leqslant \frac{1}{24} \cdot \left(\frac{d}{n}\right)^{\frac{\alpha}{1+\alpha}}$. Then, we have that by the triangle inequality that:

$$\mathbb{P}\left\{\|\hat{\mu}(\boldsymbol{X}) - \mu_S\| \geqslant \frac{1}{24} \cdot \left(\frac{d}{n}\right)^{\frac{\alpha}{1+\alpha}} \mid \boldsymbol{X}\right\} \geqslant \frac{1}{2}.$$

In the alternate case where $\|\hat{\mu}(\boldsymbol{X}) - \mu_R\| \geqslant \frac{1}{24} \cdot \left(\frac{d}{n}\right)^{\frac{\alpha}{1+\alpha}}$ for all such R, the same conclusion holds true trivially. From these two cases, we obtain:

$$\mathbb{P}\left\{\|\hat{\mu}(\boldsymbol{X}) - \mu_S\| \geqslant \frac{1}{24} \cdot \left(\frac{d}{n}\right)^{\frac{\alpha}{1+\alpha}} \mid \boldsymbol{X}\right\} \geqslant \frac{1}{2}.$$

Since such an $\boldsymbol{X}$ occurs with probability at least $1/2$, we arrive at our result:

$$\mathbb{P}\left\{\|\hat{\mu}(\boldsymbol{X}) - \mu_S\| \geqslant \frac{1}{24} \cdot \left(\frac{d}{n}\right)^{\frac{\alpha}{1+\alpha}}\right\} \geqslant \frac{1}{4}.$$

□

As part of our proof, we use the following one-dimensional lower bound from [29].

Theorem 4.4. *For any n, $\delta \in \left(2^{-\frac{n}{4}}, \frac{1}{2}\right)$, there exists a set of distributions $\mathcal{G}$ such that any $\mathcal{D} \in \mathcal{G}$ satisfies the weak-moment condition for some $\alpha > 0$ such that for any estimator $\hat{\mu}$:*

$$\mathbb{P}_{\mathcal{D} \in \mathcal{G}}\left\{|\hat{\mu}(\boldsymbol{X}) - \mu(\mathcal{D})| \geqslant \left(\frac{\log 2/\delta}{n}\right)^{\frac{\alpha}{1+\alpha}}\right\} \geqslant \delta$$

where $\boldsymbol{X}$ are drawn iid from $\mathcal{D}$.

Finally, we have the main theorem of the section:

Theorem 4.5. *Let $n > 0$ and $\delta \in \left(e^{-\frac{n}{4}}, \frac{1}{4}\right)$. Then, there exists a set of distributions over $\mathbb{R}^d$, $\mathcal{F}$ such that each $\mathcal{D} \in \mathcal{F}$ satisfies the weak-moment condition for some $\alpha > 0$ and the following holds, for any estimator $\hat{\mu}$:*

$$\mathbb{P}_{\mathcal{D}\in\mathcal{F}}\left\{\|\hat{\mu}(\boldsymbol{X}) - \mu(\mathcal{D})\| \geqslant \frac{1}{24}\cdot\max\left(\left(\frac{d}{n}\right)^{\frac{\alpha}{1+\alpha}}, \sqrt{\frac{d}{n}}, \left(\frac{\log 2/\delta}{n}\right)^{\frac{\alpha}{1+\alpha}}\right)\right\} \geqslant \delta,$$

where $\boldsymbol{X}$ are generated iid from $\mathcal{D}$.

Proof. When $n > d$, the bound follows from applications of Lemma 4.5 and Theorem 4.4. When $n \leqslant d$, the bound follows from known results for the bounded-covariance ($\alpha = 1$) setting [77]; we include a proof for completeness (see Lemma 4.22). □

Appendix

4.6 Auxiliary Results

Lemma 4.6. *For any $\boldsymbol{Z} \in \mathbb{R}^{k\times d}$ and $x \in \mathbb{R}^d$, the optimal value of $MT(x, r, \boldsymbol{Z})$ is monotonically nonincreasing in r.*

Proof. The lemma follows trivially from the fact that a feasible solution X of $MT(x, r, \boldsymbol{Z})$ is also a feasible solution for $MT(x, r', \boldsymbol{Z})$ for $r' \leqslant r$. □

Lemma 4.7. *For $X \sim \mathcal{N}(0, 1)$, $\mathbb{E}[|X|] = \sqrt{\frac{2}{\pi}}$.*

Proof.

$$\begin{aligned}\mathbb{E}[|X|] &= \int_{-\infty}^{\infty} |x| \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} dx = 2\int_0^{\infty} x \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\} dx \\ &= \sqrt{\frac{2}{\pi}} \int_0^{\infty} \exp\{-t\} dt = \sqrt{\frac{2}{\pi}}.\end{aligned}$$

□

Proof of Lemma 4.1. The argument hinges on a Gaussian projection trick which introduces $g \sim \mathcal{N}(0, I)$ to rewrite the norm. From Lemma 4.7 and the convexity of the function $f(x) = |x|^{1+\alpha}$, we have:

$$\begin{aligned}\mathbb{E}[\|X\|^{1+\alpha}] &= \mathbb{E}_X \left[\left(\sqrt{\frac{\pi}{2}} \mathbb{E}_g |\langle X, g\rangle|\right)^{1+\alpha}\right] \leqslant \frac{\pi}{2} \mathbb{E}_X \mathbb{E}_g \left[|\langle X, g\rangle|^{1+\alpha}\right] \\ &= \frac{\pi}{2} \mathbb{E}_g \|g\|^{1+\alpha} \mathbb{E}_X \left[|\left\langle X, \frac{g}{\|g\|}\right\rangle|^{1+\alpha}\right] \leqslant \frac{\pi}{2} \mathbb{E}_g[\|g\|^{1+\alpha}] \leqslant \frac{\pi}{2} \cdot d^{\frac{1+\alpha}{2}}.\end{aligned}$$

□

The following result derives an analogue of the Chebyshev inequality that applies under the weak-moment assumption. The primary technical difficulty to showing concentrations of sums of such random variables is that we cannot exploit orthogonality of independent random variables in L_2 by "expanding" out the square—since the requisite second moments do not necessarily exist.

Proof of Lemma 4.2. The case where $\alpha = 0$ is trivial. When $\alpha > 0$, we start by defining:

$$S_i = \sum_{j=1}^{i} X_j, \qquad S_0 = 0, \qquad f(x) = |x|^{1+\alpha}.$$

Therefore, we have from an application of Lemma 4.8:

$$\begin{aligned}
\mathbb{E}\left[f(S_n)\right] &= \mathbb{E}\left[\sum_{i=1}^{n} f(S_i) - f(S_{i-1})\right] = \sum_{i=1}^{n} \mathbb{E}\left[f(S_i) - f(S_{i-1})\right] \\
&= \sum_{i=1}^{n} \mathbb{E}\left[\int_{S_{i-1}}^{S_i} f'(x)dx\right] = \sum_{i=1}^{n} \mathbb{E}\left[X_i f'(S_{i-1}) + \int_{S_{i-1}}^{S_i} f'(x) - f'(S_{i-1})dx\right] \\
&= \sum_{i=1}^{n} \mathbb{E}\left[\int_{S_{i-1}}^{S_i} f'(x) - f'(S_{i-1})dx\right] \leqslant 2^{1-\alpha} \sum_{i=1}^{n} \mathbb{E}\left[\int_{0}^{|X_i|} f'\left(\frac{t}{2}\right) dt\right] \\
&= 2^{1-\alpha} \sum_{i=1}^{n} \mathbb{E}\left[\int_{0}^{|X_i|/2} 2f'\left(s\right) ds\right] = 2^{2-\alpha} \sum_{i=1}^{n} \mathbb{E}\left[f\left(\frac{|X_i|}{2}\right)\right] \leqslant 2n.
\end{aligned}$$

□

Lemma 4.8. *Let $g(x) = \operatorname{sgn}(x)|x|^\alpha$ for some $0 < \alpha \leqslant 1$. Then we have for any $h \geqslant 0$:*

$$\max_x g(x+h) - g(x) = 2^{1-\alpha} h^\alpha.$$

Proof. Consider the function $l(x) = g(x+h) - g(x)$. We see that l is differentiable everywhere except at $x = 0$ and $x = -h$. As long as $x \neq 0, -h$, we have:

$$l'(x) = g'(x+h) - g'(x) = \alpha(|x+h|^{\alpha-1} - |x|^{\alpha-1})$$

Since, we have $\alpha \leqslant 1$, $x = -\frac{h}{2}$ is a local maxima for $l(x)$. Furthermore, note that $l'(x) \geqslant 0$ for $x \in (-\infty, -\frac{h}{2}) \setminus \{-h\}$ and $l'(x) \leqslant 0$ for $x \in (-\frac{h}{2}, \infty) \setminus \{0\}$. Therefore, we get from the continuity of l that $x = -\frac{h}{2}$ is a global maxima for $l(x)$. □

We now provide an auxiliary result which will be useful to controlling the moments of the thresholded versions of the vectors X_i.

Lemma 4.9. *Let ν be a mean-zero distribution over $\mathbb{R}^d$ such that $X \sim \nu$ satisfies the weak-moment condition for some $\alpha > 0$. Furthermore, let $A \subset \mathbb{R}^d$ be such that $\nu(A) = \delta \leqslant \frac{1}{2}$. Let $\nu_S()$ be the conditional distribution of ν conditioned on the event $\{X \in S\}$ for any $X \subset \mathbb{R}^d$. Then we have for $Y \sim \nu(A^c)$:*

$$\textit{Claim 1: } \|\mu(\nu_{A^c})\| \leqslant 2\delta^{\frac{\alpha}{1+\alpha}}, \qquad \textit{Claim 2: } \forall v \in \mathcal{S}^{d-1}, \quad \mathbb{E}\left[|\langle v, Y - \mu(\nu_{A^c})\rangle|^{1+\alpha}\right] \leqslant 20.$$

Proof. Letting $p_A = \mathbb{P}\{X \in A\}$, we have $\nu = p_A \nu_A + p_{A^c} \nu_{A^c}$. Then,

$$\|\mu(\nu) - \mu(\nu_{A^c})\| = \max_{v \in \mathcal{S}^{d-1}} \langle v, \mu(\nu) - \mu(\nu_{A^c})\rangle.$$

So for any $v \in \mathcal{S}^{d-1}$:

$$\langle v, \mu(\nu) - \mu(\nu_{A^c})\rangle = \langle v, p_A \mu(\nu_A) + p_{A^c}\mu(\nu_{A^c}) - \mu(\nu_{A^c})\rangle$$

$$= \langle v, p_A \mu(\nu_A) - p_A \mu(\nu_{A^c}) \rangle = p_A \langle v, \mu(\nu_A) - \mu(\nu_{A^c}) \rangle.$$

Since $\mu(\nu) = 0$, we have $p_A \mu(\nu_A) = -p_{A^c} \mu(\nu_{A^c})$. We now get:

$$p_A \langle v, \mu(\nu_A) - \mu(\nu_{A^c}) \rangle = p_A \left\langle v, \mu(\nu_A) + \frac{p_A}{p_{A^c}} \mu(\nu_A) \right\rangle = \left(1 + \frac{p_A}{p_{A^c}}\right) \langle v, p_A \mu(\nu_A) \rangle.$$

Finally,

$$\begin{aligned} \langle v, p_A \mu(\nu_A) \rangle &= \mathbb{E}_{X \sim \mu} \left[\mathbf{1}\left\{X \in A\right\} \langle X, v \rangle \right] \\ &\leqslant \left(\mathbb{E}\left[\left(\mathbf{1}\left\{X \in A\right\} \right)^{\frac{1+\alpha}{\alpha}} \right] \right)^{\frac{\alpha}{1+\alpha}} \cdot \left(\mathbb{E}\left[|\langle X, v \rangle|^{1+\alpha} \right] \right)^{\frac{1}{1+\alpha}} = p_A^{\frac{\alpha}{1+\alpha}} \end{aligned}$$

where the inequality follows by an application of Hölder's inequality. We get the first claim as:

$$\max_{v \in \mathcal{S}^{d-1}} \langle v, \mu(\nu) - \mu(\nu_{A^c}) \rangle = \left(1 + \frac{p_A}{p_{A^c}}\right) \langle v, p_A \mu(\nu_{A^c}) \rangle \leqslant \left(1 + \frac{p_A}{p_{A^c}}\right) p_A^{\frac{\alpha}{1+\alpha}} \leqslant 2\delta^{\frac{\alpha}{1+\alpha}},$$

where the final inequality follows from the fact that $p_{A^c} \geqslant p_A$.

For the second claim, let $Y \sim \nu_{A^c}$ and $\mu_Y = \mathbb{E}[Y]$. We decompose the required term as follows:

$$\mathbb{E}\left[|\langle Y - \mu_Y, v \rangle|^{1+\alpha} \right] \leqslant 2^{1+\alpha} \cdot \mathbb{E}\left[|\langle \mu_Y, v \rangle|^{1+\alpha} + |\langle Y, v \rangle|^{1+\alpha} \right].$$

For the first term, we have with $Z \sim \nu_A$:

$$\mathbb{E}\left[|\langle Y, v \rangle|^{1+\alpha} \right] = p_{A^c}^{-1} \left(\mathbb{E}\left[|\langle X, v \rangle|^{1+\alpha} \right] - p_A \mathbb{E}\left[|\langle Z, v \rangle|^{1+\alpha} \right] \right) \leqslant 2.$$

Therefore, we finally have:

$$\mathbb{E}\left[|\langle Y - \mu_Y, v \rangle|^{1+\alpha} \right] \leqslant 8 + 2^{1+\alpha} \cdot 2^{1+\alpha} \cdot \delta^{\alpha} \leqslant 16,$$

which proves the second claim of the lemma. □

Lemma 4.10. *Let $X \sim \nu$ be a mean-zero random vector satisfying the weak-moment condition for some $0 \leqslant \alpha \leqslant 1$. Then, we have for any $\tau > 0$:*

$$\mathbb{E}\left[\|X\|^2 \cdot \mathbf{1}\left\{\|X\| \leqslant \tau\right\} \right] \leqslant \frac{\pi}{2} d^{\frac{1+\alpha}{2}} \tau^{1-\alpha}.$$

Proof. The proof of the lemma proceeds as follows:

$$\mathbb{E}\left[\|X\|^2 \cdot \mathbf{1}\left\{\|X\| \leqslant \tau\right\} \right] \leqslant \tau^{1-\alpha} \mathbb{E}\left[\|X\|^{1+\alpha} \mathbf{1}\left\{\|X\| \leqslant \tau\right\} \right] \leqslant \tau^{1-\alpha} \mathbb{E}\left[\|X\|^{1+\alpha} \right] \leqslant \frac{\pi}{2} d^{\frac{1+\alpha}{2}} \tau^{1-\alpha},$$

where the last inequality follows from Lemma 4.1.

□

4.7 Initial Estimate

In this subsection, we analyze the initial estimate used in the thresholding step. We will show that the estimate is within $O(\sqrt{d})$ of the true mean with high probability.

Lemma 4.11. *Let $X_1, \ldots, X_n$ be i.i.d. random vectors with mean μ, satisfying the weak-moment condition for some $\alpha > 0$. Then the mean estimate, $\hat{\mu}$, provided by Algorithm 4 satisfies:*

$$\|\mu - \hat{\mu}\| \leqslant 24\sqrt{d},$$

with probability at least $1 - e^{-\frac{n}{50}}$.

Proof. Since our algorithm is translation invariant, we may assume without loss of generality that $\mu = \mathbf{0}$. Therefore, it suffices to prove that with probability at least $2^{-\Omega(n)}$:

$$\|\hat{\mu}\| \leqslant 16\sqrt{d}.$$

We have from Lemma 4.1 that $\mathbb{E}[\|X\|] \leqslant \frac{\pi}{2} \cdot \sqrt{d}$. Applying Markov's inequality:

$$\mathbb{P}\left\{\|X\| \leqslant 8\sqrt{d}\right\} \geqslant \frac{3}{4}.$$

Combining with Hoeffding's inequality we conclude that:

$$\mathbb{P}\left\{\sum_{i=1}^{n} \mathbf{1}\left\{\|X_i\| \leqslant 8\sqrt{d}\right\} \leqslant 0.6n\right\} \leqslant \exp\left\{-\frac{n}{50}\right\}.$$

Since Algorithm 4 returns as an estimate one of the data points, let $\hat{\mu} = X_i$ for some i and let $r_j = \min\{r > 0 : \sum_{k=1}^{n} \mathbf{1}\{\|X_j - X_k\| \leqslant r\} \geqslant 0.6n\}$ for any $j \in [n]$. We now condition on the following event:

$$\sum_{i=1}^{n} \mathbf{1}\left\{\|X_i\| \leqslant 8\sqrt{d}\right\} > 0.6n.$$

Let $\mathcal{S} = \left\{j : \|X_j\| \leqslant 8\sqrt{d}\right\}$. By the triangle inequality, for any $j \in \mathcal{S}$ we have:

$$\sum_{k=1}^{n} \mathbf{1}\left\{\|X_k - X_j\| \leqslant 16\sqrt{d}\right\} \geqslant 0.6n.$$

Therefore, by the definition of $\hat{\mu}$ we infer that $r_i \leqslant 16\sqrt{d}$. Now, let $\mathcal{T} = \{k : \|X_k - X_i\| \leqslant r_i\}$. We have by the definition of r_i that $|\mathcal{T}| \geqslant 0.6n$. By the pigeonhole principle, we have that $\mathcal{T} \cap \mathcal{S} \neq \phi$. Let $j \in \mathcal{T} \cap \mathcal{S}$. By the triangle inequality we obtain:

$$\|X_i\| \leqslant \|X_i - X_j\| + \|X_j\| \leqslant 16\sqrt{d} + 8\sqrt{d} = 24\sqrt{d}.$$

Since the event being conditioned on occurs with probability at least $1 - e^{-\frac{n}{50}}$, this concludes the proof of the lemma. □

4.8 Analyzing Relaxation

We first show that the optimal value of the semidefinite program **MT** satisfies a bounded-difference condition with respect to the Z_i's.

Lemma 4.12. *Let $\boldsymbol{Y} = (Y_1, \ldots, Y_k)$ be any set of k vectors in $\mathbb{R}^d$. Now, let $\boldsymbol{Y}' = (Y_1, \ldots, Y_i', \ldots, Y_k)$ be the same set of k vectors with the i^{th} vector replaced by $Y_i' \in \mathbb{R}^d$. If m and m' are the optimal values of $MT(x, r, \boldsymbol{Y})$ and $MT(x, r, \boldsymbol{Y}')$, we have:*

$$|m - m'| \leqslant 1.$$

Proof. First, assume that X is a feasible solution to $MT(x, r, \boldsymbol{Y})$. Let us define X' as:

$$X'_{i,j} = \begin{cases} X_{i,j} & \text{if } i, j \neq b_i \\ 0 & \text{otherwise.} \end{cases}$$

That is, X' is equal to X except with the row and column corresponding to b_i being set to 0. We see that X' forms a feasible solution to $MT(x, r, \boldsymbol{Y}')$. Therefore, we have that:

$$\sum_{j=1}^{k} X_{b_j,b_j} = \sum_{j=1, j\neq i}^{k} X'_{b_j,b_j} + X_{b_i,b_i} \leqslant \sum_{j=1, j\neq i}^{k} X'_{b_j,b_j} + 1 \leqslant m' + 1,$$

where the bound $X_{b_i,b_i} \leqslant 1$ follows from the fact that the 2×2 submatrix of X formed by the rows and columns indexed by 1 and b_i is positive semidefinite and from the constraint that $X_{b_i,b_i} = X_{1,b_i}$. Since the series of equalities holds for all feasible solutions X of $MT(x, r, \boldsymbol{Y})$, we get:

$$m \leqslant m' + 1.$$

Through a similar argument, we also conclude that $m' \leqslant m + 1$. Putting the two inequalities together, we obtain the required conclusion. □

For the next few lemmas, we are concerned with the case where $x = \mu$. Since we already know that the optimal SDP value satisfies the bounded differences condition, we need to verify that the expectation is small. As a first step towards this, we define the 2-to-1 norm of a matrix M.

Definition 4.1. The 2-to-1 norm of $M \in \mathbb{R}^{n\times d}$ is defined as

$$\|M\|_{2\to 1} = \max_{\substack{\|v\|=1 \\ \sigma_i \in \{\pm 1\}}} \sigma^\top M v = \max_{\|v\|=1} \|Mv\|_1.$$

We consider the classical semidefinite programming relaxation of the 2-to-1 norm. To start with, we will define a matrix $X \in \mathbb{R}^{(n+d+1)\times(n+d+1)}$ with the rows and columns indexed

by 1 and the elements σ_i and v_j. The semidefinite programming relaxation is defined as follows:

$$\max \sum_{i,j} M_{i,j} X_{\sigma_i, v_j}$$
$$\text{s.t } X_{1,1} = 1,\ \sum_{j=1}^{d} X_{v_j,v_j} = 1,\ X_{\sigma_i,\sigma_i} = 1,\ X \succcurlyeq 0. \tag{TOR}$$

We now state a theorem of Nesterov as stated in [51]:

Theorem 4.6. *([90]) There is a constant $K_{2\to 1} = \sqrt{\pi/2} \leqslant 2$ such that the optimal value, m, of the semidefinite programming relaxation TOR satisfies:*

$$m \leqslant K_{2\to 1} \|M\|_{2\to 1}.$$

In the next step, we bound the expected 2-to-1 norm of the random matrix Z. To do this, we begin by recalling the Ledoux-Talagrand Contraction Theorem [66].

Theorem 4.7. *Let $X_1, \ldots, X_n \in \mathbb{R}^d$ be i.i.d. random vectors, $\mathcal{F}$ be a class of real-valued functions on $\mathbb{R}^d$ and $\sigma_i, \ldots, \sigma_n$ be independent Rademacher random variables. If $\phi : \mathbb{R} \to \mathbb{R}$ is an L-Lipschitz function with $\phi(0) = 0$, then:*

$$\mathbb{E} \sup_{f\in\mathcal{F}} \sum_{i=1}^{n} \sigma_i \phi(f(X_i)) \leqslant L \cdot \mathbb{E} \sup_{f\in\mathcal{F}} \sum_{i=1}^{n} \sigma_i f(X_i).$$

We are now ready to bound the expected 2-to-1 norm of the random matrix Z.

Lemma 4.13. *Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n) \in \mathbb{R}^{n\times d}$ be a set of n i.i.d. random vectors such that $\mathbb{E}[Y_i] = 0$ and $\mathbb{E}[Y_i Y_i^\top] = \Lambda$ and assume that:*

$$\max_{v\in\mathcal{S}^{d-1}} \mathbb{E}\left[|\langle v, Y\rangle|^{1+\alpha}\right] \leqslant \beta.$$

Then we have:

$$\mathbb{E}\|\boldsymbol{Y}\|_{2\to 1} \leqslant 2\sqrt{n \operatorname{Tr}\Lambda} + n\beta^{\frac{1}{1+\alpha}}.$$

Proof. Denoting by Y and Y_i' random vectors that are independently and identically distributed as Y_i and by σ_i independent Rademacher random variables, we have:

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{Y}\|_{2\to 1}] &= \mathbb{E}\left[\max_{\|v\|=1} \sum_{i=1}^{n} |\langle Y_i, v\rangle|\right] = \mathbb{E}\left[\max_{\|v\|=1} \sum_{i=1}^{n} |\langle Y_i, v\rangle| + \mathbb{E}|\langle v, Y_i\rangle| - \mathbb{E}|\langle v, Y_i\rangle|\right] \\ &\leqslant \mathbb{E}\left[\max_{\|v\|=1} \sum_{i=1}^{n} |\langle Y_i, v\rangle| - \mathbb{E}|\langle Y_i', v\rangle|\right] + n \max_{\|v\|=1} \mathbb{E}[|\langle v, Y\rangle|]\end{aligned}$$

$$\leqslant \mathbb{E}\left[\max_{\|v\|=1}\sum_{i=1}^{n}\sigma_i(|\langle Y_i, v\rangle| - |\langle Y_i', v\rangle|)\right] + n\max_{\|v\|=1}\mathbb{E}\left[|\langle v, Y\rangle|\right].$$

Now, we have for the second term:

$$\max_{\|v\|=1}\mathbb{E}[|\langle v, Y\rangle|] \leqslant \max_{\|v\|=1}\left(\mathbb{E}\langle v, Y\rangle^{1+\alpha}\right)^{\frac{1}{1+\alpha}} \leqslant \beta^{\frac{1}{1+\alpha}}.$$

For the first term, we employ a standard symmetrization argument:

$$\begin{aligned}
\mathbb{E}\left[\max_{\|v\|=1}\sum_{i=1}^{n}\sigma_i(|\langle Y_i, v\rangle| - |\langle Y_i', v\rangle|)\right] &\leqslant \mathbb{E}\left[\max_{\|v\|=1}\sum_{i=1}^{n}\sigma_i|\langle Y_i, v\rangle|\right] + \mathbb{E}\left[\max_{\|v\|=1}\sum_{i=1}^{n}-\sigma_i|\langle Y_i', v\rangle|\right] \\
&= 2\mathbb{E}\left[\max_{\|v\|=1}\sum_{i=1}^{n}\sigma_i|\langle v, Y_i\rangle|\right] \leqslant 2\mathbb{E}\left[\max_{\|v\|=1}\sum_{i=1}^{n}\sigma_i\langle v, Y_i\rangle\right] \\
&= 2\mathbb{E}\left[\|\sum_{i=1}^{n}\sigma_i Y_i\|\right] \leqslant 2\left(\mathbb{E}\left[\|\sum_{i=1}^{n}\sigma_i Y_i\|^2\right]\right)^{1/2} \\
&= 2\left(\mathbb{E}\sum_{1\leqslant i,j\leqslant n}\sigma_i\sigma_j\langle Y_i, Y_j\rangle\right)^{1/2} = 2\sqrt{n\operatorname{Tr}\Lambda},
\end{aligned}$$

where the second inequality follows from the Ledoux-Talagrand Contraction Principle (Theorem 4.7). By putting these two bounds together, we prove the lemma. □

We now bound the expected value of $MT(\mu, r, \boldsymbol{Y})$ by relating it to $\|\boldsymbol{Y}\|_{2\to 1}$.

Lemma 4.14. *Let $\boldsymbol{Y} = (Y_1, \ldots, Y_k) \in \mathbb{R}^{k\times d}$ be a collection of k i.i.d. random vectors with mean μ and covariance Λ and assume that:*

$$\max_{v\in\mathcal{S}^{d-1}}\mathbb{E}\left[|\langle v, Y\rangle|^{1+\alpha}\right] \leqslant \beta.$$

Denoting by $\mathcal{S}$ the set of feasible solutions for $MT(\mu, r, \boldsymbol{Y})$, we have:

$$\mathbb{E}\max_{x\in\mathcal{S}}\sum_{i=1}^{k}X_{1,b_i} \leqslant \frac{1}{2r}\left(5\sqrt{k\operatorname{Tr}\Lambda} + 2k\beta^{\frac{1}{1+\alpha}}\right).$$

Proof. First, let X be a feasible solution for $MT(\mu, r, \boldsymbol{Y})$. We construct a new matrix W which is indexed by σ_i and v_j as opposed to b_i and v_j for X:

$$\begin{aligned}
W_{\sigma_i,\sigma_j} = 4X_{b_i,b_j} - 2X_{1,b_i} - 2X_{1,b_j} + 1, \quad W_{v_i,v_j} = X_{v_i,v_j}, \quad W_{1,1} = 1, \\
W_{1,v_i} = X_{1,v_i}, \quad W_{1,\sigma_i} = 2X_{1,b_i} - 1, \quad W_{v_i,\sigma_j} = 2X_{v_i,b_j} - X_{1,v_i}.
\end{aligned}$$

We prove that W is a feasible solution to the SDP relaxation TOR of $\boldsymbol{Y} - \mu$. We see that:

$$W_{\sigma_i,\sigma_i} = 1 \text{ and } \sum_{i=1}^{d} W_{v_i,v_i} = 1.$$

Thus, we simply need to verify that W is positive semidefinite. Let $w \in \mathbb{R}^{k+d+1}$ be indexed by 1, σ_i and v_j. We construct from w a new vector w', indexed by 1, b_i and v_j and defined as follows:

$$w'_1 = w_1 - \sum_{i=1}^{k} w_{\sigma_i}, \quad w'_{b_i} = 2w_{\sigma_i}, \quad w'_{v_j} = w_{v_j}.$$

With w' defined in this way, we have the following equality:

$$w^\top W w = (w')^\top X w' \geqslant 0.$$

Since the condition holds for all $w \in \mathbb{R}^{k+d+1}$, we get that $W \succcurlyeq 0$. Therefore, we conclude that W is a feasible solution to the SDP relaxation TOR of $\boldsymbol{Y} - \mu$.

We bound the expected value of $MT(\mu, r, \boldsymbol{Y})$ as follows, denoting by v_{b_i} the vector $(X_{b_i,v_1}, \ldots, X_{b_i,v_d})$ and by v the vector $(X_{1,v_1}, \ldots, X_{1,v_d})$:

$$\begin{aligned}
\mathbb{E} \max_{X \in \mathcal{S}} \sum_{i=1}^{k} X_{1,b_i} = \mathbb{E} \max_{X \in \mathcal{S}} \sum_{i=1}^{k} X_{b_i,b_i} &\leqslant \frac{1}{r} \mathbb{E} \max_{X \in \mathcal{S}} \sum_{i=1}^{k} \langle v_{b_i}, Y_i - \mu \rangle \\
&= \frac{1}{2r} \mathbb{E} \max_{X \in S} \Big[\sum_{i=1}^{k} \langle 2v_{b_i} - v, Y_i - \mu \rangle + \sum_{i=1}^{k} \langle v, Y_i - \mu \rangle \Big] \\
&\leqslant \frac{1}{2r} \left(\mathbb{E} \max_{X \in S} \sum_{i=1}^{k} \langle 2v_{b_i} - v, Y_i - \mu \rangle + \mathbb{E} \max_{X \in \mathcal{S}} \sum_{i=1}^{k} \langle v, Y_i - \mu \rangle \right).
\end{aligned}$$

From the fact that X is positive semidefinite, and from the fact that the 2×2 submatrix indexed by v_i and b_j is positive semidefinite, we obtain:

$$X_{v_i,b_j}^2 \leqslant X_{v_i,v_i} X_{b_j,b_j} \leqslant X_{v_i,v_i} \implies \|v_{b_j}\|^2 = \sum_{i=1}^{d} X_{v_i,b_j}^2 \leqslant \sum_{i=1}^{d} X_{v_i,v_i} = 1.$$

Therefore, we get for the second term in the above equation:

$$\mathbb{E} \max_{X \in \mathcal{S}} \sum_{i=1}^{k} \langle v, Y_i - \mu \rangle \leqslant \mathbb{E} \| \sum_{i=1}^{k} Y_i - \mu \| \leqslant \left(\mathbb{E} \| \sum_{i=1}^{k} Y_i - \mu \|^2 \right)^{1/2} = (k \operatorname{Tr} \Lambda)^{1/2}.$$

We bound the first term using the following series of inequalities where W is constructed from X as described above:

$$\mathbb{E} \max_{x \in \mathcal{S}} \sum_{i=1}^{k} \langle 2v_{b_i} - v, Y_i - \mu \rangle = \mathbb{E} \max_{x \in \mathcal{S}} \sum_{i=1}^{k} \sum_{j=1}^{d} (Y_i - \mu)_j W_{\sigma_i,v_j} = \mathbb{E} \max_{x \in \mathcal{S}} \sum_{i=1}^{k} \sum_{j=1}^{d} (\boldsymbol{Y}_{i,j} - \mu_j) W_{\sigma_i,v_j}$$

$$\leqslant 2\mathbb{E}\|\boldsymbol{Y} - \mathbf{1}\mu^\top\|_{2\to 1} \leqslant 4\sqrt{k\,\mathrm{Tr}\,\Lambda} + 2k\beta^{\frac{1}{1+\alpha}},$$

where the first inequality follows from Theorem 4.6 and the second inequality follows from Lemma 4.13. By combining these three inequalities, we get:

$$\mathbb{E}\max_{x\in\mathcal{S}}\sum_{i=1}^{k} X_{1,b_i} \leqslant \frac{1}{2r}\left(5\sqrt{k\,\mathrm{Tr}\,\Lambda} + 2k\beta^{\frac{1}{1+\alpha}}\right).$$

□

Finally, we establish the main technical result of this section, Lemma 4.3.

Proof of Lemma 4.3. From Lemma 4.14, we see that:

$$\mathbb{E}\max_{X\in\mathcal{S}}\sum_{i=1}^{k} X_{b_i,b_i} \leqslant \frac{k}{40}.$$

Now from Lemma 4.12 and an application of the bounded difference inequality (see, for example, Theorem 6.2 in [11]), with probability at least $1 - e^{k/800}$:

$$\max_{X\in\mathcal{S}}\sum_{i=1}^{k} X_{b_i,b_i} \leqslant \frac{k}{20}.$$

□

In the following lemma, we analyze the set of random vectors returned by Algorithm 5. It will be useful to condition on the conclusion of Lemma 4.11.

Lemma 4.15. *Let $\boldsymbol{X} = \{X_i\}_{i=1}^n \sim \nu$ be iid zero-mean random vectors, satisfying the weak-moment condition for some $\alpha > 0$. Furthermore, suppose $x^\dagger$ satisfies $\|x^\dagger\| \leqslant 24\sqrt{d}$. Then, the set of vectors $\boldsymbol{Y}$ returned by Algorithm 5 with input $\boldsymbol{X}$ and $x^\dagger$ are iid random vectors vectors with mean $\widetilde{\mu}$ and covariance $\widetilde{\Sigma}$ and satisfy:*

$$\textit{Claim 1: } \mathbb{P}\left\{|\boldsymbol{Y}| \geqslant \frac{3n}{4}\right\} \geqslant 1 - e^{-\frac{n}{50}},$$

$$\textit{Claim 2: } \|\widetilde{\mu}\| \leqslant 2\left(\frac{d}{n}\right)^{\frac{\alpha}{1+\alpha}}$$

$$\textit{Claim 3: } \forall\|v\| = 1 : \mathbb{E}\left[|\langle Y_i - \widetilde{\mu}, v\rangle|^{1+\alpha}\right] \leqslant 20,$$

$$\textit{Claim 4: } \mathrm{tr}(\widetilde{\Sigma}) \leqslant 750\max\left(n^{\frac{1-\alpha}{1+\alpha}}d^{\frac{2\alpha}{(1+\alpha)}}, d\right).$$

Proof. First, consider the set $A = \{x : \|x - x^\dagger\| \leqslant \tau\}$ as defined in Algorithm 5. Note from the definition of the set A that $\{x : \|x\| \leqslant 0.75\tau\} \subseteq A$. We have from Markov's inequality and Lemma 4.1:

$$\mathbb{P}\{X_i \in A\} \geqslant 1 - \min\left(\frac{d}{n}, \frac{1}{25}\right)$$

Therefore, by an application of Hoeffding's inequality, using the definition of the set of points $Y_1, \ldots, Y_m$, we have that, with probability at least $1 - e^{-\frac{n}{50}}$:

$$|\boldsymbol{Y}| \geqslant \frac{3n}{4}$$

This proves the first claim of the lemma. For the next two claims, note that conditioned on the random variable $\hat{\mu}$, each of the Y_i are independent and identically distributed according to ν_A. Again, we get from the bound on $\mathbb{P}\{X_i \in A\}$ by an application of Lemma 4.9, the next two claims of the lemma:

$$\text{Claim 2: } \|\widetilde{\mu}\| \leqslant 2\left(\frac{d}{n}\right)^{\frac{\alpha}{1+\alpha}}, \qquad \text{Claim 3: } \forall \|v\| = 1 : \mathbb{E}\left[|\langle Y_i - \widetilde{\mu}, v\rangle|^{1+\alpha}\right] \leqslant 20.$$

For the final claim, note that as $\|x^\dagger\| \leqslant 24\sqrt{d}$, we have $A \subseteq B := \{x : \|x\| \leqslant 1.25\tau\}$. Therefore, we have by the property of the mean that:

$$\begin{aligned}\mathrm{tr}\widetilde{\Sigma} = \mathbb{E}\left[\|Y_i - \widetilde{\mu}\|^2\right] &\leqslant \mathbb{E}\left[\|Y_i\|^2\right] = \frac{1}{\nu(A)}\mathbb{E}\left[\|X_j\|^2 \mathbf{1}\{X_j \in A\}\right] \\ &\leqslant 2\mathbb{E}\left[\|X_j\|^2 \mathbf{1}\{X_j \in B\}\right] \leqslant 750 \max\left(n^{\frac{1-\alpha}{1+\alpha}} d^{\frac{2\alpha}{(1+\alpha)}}, d\right),\end{aligned}$$

where the final inequality follows from Lemma 4.10 and the definition of τ. □

The main result of this section is the following high probability guarantee on the set of points output by Algorithm 3.

Lemma 4.16. *Let $\delta > e^{n/8000}$ and $\boldsymbol{X} = \{X_i\}_{i=1}^n$ be iid random vectors with mean μ, satisfying the weak-moment condition for some known $\alpha > 0$. Furthermore, suppose that $x^\dagger$ satisfies $\|x^\dagger - \mu\| \leqslant 24\sqrt{d}$. Let $\boldsymbol{Z} = \{Z_i\}_{i=1}^k$ denote the set of vectors output by Algorithm 3 run with inputs $\boldsymbol{X}$, $x^\dagger$ and δ. Then, there exists a point $\widetilde{\mu}$ such that for all $r \geqslant 10^6\left(\sqrt{\frac{d}{n}} + \left(\frac{d}{n}\right)^{\frac{\alpha}{1+\alpha}} + \left(\frac{\log 1/\delta}{n}\right)^{\frac{\alpha}{1+\alpha}}\right)$:*

$$\|\widetilde{\mu} - \mu\| \leqslant 2\left(\frac{d}{n}\right)^{\frac{\alpha}{1+\alpha}} \text{ and } \max_{X \in \mathcal{S}} \sum_{i=1}^k X_{b_i, b_i} \leqslant \frac{k}{20},$$

with probability at least $1 - \delta/2$ where $\mathcal{S}$ denotes the set of feasible solutions of $\boldsymbol{MT}(\widetilde{\mu}, r, \boldsymbol{Z})$.

Proof. Note that it is sufficient to prove the lemma in the specific case where $\mu = \mathbf{0}$. We may now assume each of the Y_i are iid random variables satisfying the conclusions of Lemma 4.15. Therefore, Z_i are iid random vectors with mean $\widetilde{\mu}$ and covariance $\widetilde{\Sigma}$ satisfying:

$$\|\widetilde{\mu}\| \leqslant 2\left(\frac{d}{n}\right)^{\frac{\alpha}{1+\alpha}} \qquad \operatorname{tr}(\widetilde{\Sigma}) \leqslant \frac{750k \max\left(n^{\frac{1-\alpha}{1+\alpha}} d^{\frac{2\alpha}{(1+\alpha)}}, d\right)}{n}.$$

Furthermore, we have by an application of Lemma 4.2 that:

$$\forall \|v\| = 1 : \mathbb{E}\left[|\langle v, Z_i - \widetilde{\mu}\rangle|^{1+\alpha}\right] \leqslant 80\left(\frac{k}{n}\right)^{\alpha}.$$

Finally, the conclusion of the lemma follows by an application of Lemma 4.3 and the bound on the probabilities follows from the bounds in Lemmas 4.3 and 4.15. □

4.9 Gradient Descent Step

In this section, we analyze the gradient descent step in Algorithm 1. This part of our proof is essentially identical to prior work for the finite covariance setting and we repeat it here for the sake of completeness [20]. Throughout the section, we will analyze the convergence of gradient descent to an *arbitrary* point $\widetilde{\mu}$. However, in the final application, we will pick $\widetilde{\mu}$ to be close to μ. As discussed previously, the recovery guarantees of the gradient descent procedure are determined by the parameter r^* defined below:

Definition 4.2. For the bucket means, $\boldsymbol{Z} = (Z_1, \ldots, Z_k)$, and point $\widetilde{\mu}$, let r^* be defined as follows:

$$r^* := \min\left\{r > 0 : \mathbf{MT}(\widetilde{\mu}, r, \boldsymbol{Z}) \leqslant \frac{k}{20}\right\}.$$

We also make use of the following remark implied by Definition 4.2 (the implication follows by picking integral solutions for X_{b_i,b_i} and setting the submatrix of X corresponding to v to be rank one in the semidefinite program **MT**:

Remark 4.1. For the bucket means, $\boldsymbol{Z} = (Z_1, \ldots, Z_k)$, we have:

$$\forall v \in \mathbb{R}^d, \|v\| = 1 \;\Rightarrow |\{i : \langle Z_i - \widetilde{\mu}, v\rangle \geqslant r^*\}| \leqslant 0.05k$$

Distance Estimation Step

In this subsection, we analyze the distance estimation step from Algorithm 6. We show that an accurate estimate of the distance of the current point from $\widetilde{\mu}$ can be found. We begin by stating a lemma that shows that a feasible solution for $\mathbf{MT}(x, r, \boldsymbol{Z})$ can be converted to a feasible solution for $\mathbf{MT}(\widetilde{\mu}, r^*, \boldsymbol{Z})$ with a reduction in optimal value.

Lemma 4.17. *Let $X \in \mathbb{R}^{(k+d+1)\times(k+d+1)}$ be a positive semidefinite matrix, symbolically indexed by 1 and the variables b_i and v_j. Moreover, suppose that X satisfies:*

$$X_{1,1} = 1, \quad X_{b_i,b_i} = X_{1,b_i}, \quad \sum_{j=1}^{d} X_{v_j,v_j} = 1, \quad \sum_{i=1}^{k} X_{b_i,b_i} \geqslant 0.9k.$$

Then, there is a set of at least $0.85k$ indices $\mathcal{T}$ such that for all $i \in \mathcal{T}$:

$$\langle Z_i - \widetilde{\mu}, v_{b_i} \rangle < X_{b_i,b_i} r^*,$$

and a set of at least $k/3$ indices $\mathcal{R}$ such that for all $j \in \mathcal{R}$, we have $X_{b_j,b_j} \geqslant 0.85$.

Proof. We prove the lemma by contradiction. Firstly, note that X is infeasible for $\mathbf{MT}(\widetilde{\mu}, r^*, \boldsymbol{Z})$ as the optimal value for $\mathbf{MT}(\widetilde{\mu}, r^*, \boldsymbol{Z})$ is less than $k/20$ (Definition 4.2). Note that the only constraints of $\mathbf{MT}(\widetilde{\mu}, r^*, \boldsymbol{Z})$ that are violated by X are constraints of the form:

$$\langle Z_i - \widetilde{\mu}, v_{b_i} \rangle < X_{b_i,b_i} r^*.$$

Now, let $\mathcal{T}$ denote the set of indices for which the above inequality is violated. We can convert X to a feasible solution for $\mathbf{MT}(\widetilde{\mu}, r^*, \boldsymbol{Z})$ by setting to zero the rows and columns corresponding to the indices in $\mathcal{T}$. Let X' be the matrix obtained by the above operation. We have from Definition 4.2:

$$0.05k \geqslant \sum_{i=1}^{k} X'_{b_i,b_i} = \sum_{i=1}^{k} X_{b_i,b_i} - \sum_{i\in\mathcal{T}} X_{b_i,b_i} \geqslant 0.9k - |\mathcal{T}|,$$

where the last inequality follows from the fact that $X_{b_i,b_i} \leqslant 1$. By rearranging the above inequality, we get the first claim of the lemma.

For the second claim, let $\mathcal{R}$ denote the set of indices j satisfying $X_{b_j,b_j} \geqslant 0.85$. We have:

$$0.9k \leqslant \sum_{j=1}^{k} X_{b_j,b_j} = \sum_{j\in\mathcal{R}} X_{b_j,b_j} + \sum_{j\notin\mathcal{R}} X_{b_j,b_j} \leqslant |\mathcal{R}| + 0.85k - 0.85|\mathcal{R}| \implies \frac{k}{3} \leqslant |\mathcal{R}|.$$

This establishes the second claim of the lemma. □

The following lemma shows that if the distance between $\widetilde{\mu}$ and a point x is small then the estimate returned by Algorithm 6 is also small.

Lemma 4.18. *Suppose a point $x \in \mathbb{R}^d$ satisfies $\|x - \widetilde{\mu}\| \leqslant 20r^*$. Then, Algorithm 6 returns a value d' satisfying*

$$d' \leqslant 25r^*.$$

Proof. Let $r' = 25r^*$. Suppose that the optimal value of $\mathbf{MT}(x, r', \boldsymbol{Z})$ is greater than $0.9k$ and let its optimal solution be X. Let $\mathcal{R}$ and $\mathcal{T}$ denote the two sets whose existence is guaranteed by Lemma 4.17. From, the cardinalities of $\mathcal{R}$ and $\mathcal{T}$, we see that their intersection is not empty. For $j \in \mathcal{R} \cap \mathcal{T}$, we have:

$$0.85r' \leqslant \langle Z_j - x, v_{b_j} \rangle = \langle Z_j - \widetilde{\mu}, v_{b_j} \rangle + \langle \widetilde{\mu} - x, v_{b_j} \rangle < r^* + \|x - \widetilde{\mu}\|,$$

where the first inequality follows from the fact that $j \in \mathcal{R}$ and the fact that X is feasible for $\mathbf{MT}(x, r', \boldsymbol{Z})$ and the last inequality follows from the inclusion of j in $\mathcal{T}$ and Cauchy-Schwarz.

By plugging in the bounds on r' and r, we get:

$$\|x - \widetilde{\mu}\| > 20.25r^*.$$

This contradicts the assumption on $\|x - \widetilde{\mu}\|$ and concludes the proof of the lemma. □

The next lemma shows that the distance between $\widetilde{\mu}$ and a point x can be accurately estimated as long as x is sufficiently far from $\widetilde{\mu}$.

Lemma 4.19. *Suppose a point x satisfies $\tilde{d} = \|x - \widetilde{\mu}\| \geqslant 20r^*$. Then, Algorithm 6 returns a value d' satisfying:*

$$0.95\tilde{d} \leqslant d' \leqslant 1.25\tilde{d}.$$

Proof. Let us define the direction Δ to be the unit vector in the direction of $x - \widetilde{\mu}$. From Remark 4.1, the number of Z_i satisfying $\langle Z_i - \widetilde{\mu}, \Delta \rangle \geqslant r^*$ is less than $k/20$. Therefore, we have that for at least $0.95k$ points:

$$\langle Z_i - x, -\Delta \rangle = \langle x - \widetilde{\mu} + \widetilde{\mu} - Z_i, \Delta \rangle = \|x - \widetilde{\mu}\| - r^* \geqslant 0.95\tilde{d}.$$

Along with the monotonicity of $\mathbf{MT}(x, r, \boldsymbol{Z})$ in r (Lemma 4.6), this implies the lower bound.

For the upper bound, we show that the optimal value of $\mathbf{MT}(x, 1.25\tilde{d}, \boldsymbol{Z})$ is less than $0.9k$. For the sake of contradiction, suppose that this optimal value is greater than $0.9k$. Let X be a feasible solution of $\mathbf{MT}(x, 1.25\tilde{d}, \boldsymbol{Z})$ that achieves $0.9k$. Let $\mathcal{R}$ and $\mathcal{T}$ be the two sets whose existence is guaranteed by Lemma 4.17 and j be an element in their intersection. We have for j:

$$\begin{aligned} 0.85(1.25\tilde{d}) \leqslant X_{b_j,b_j} 1.25\tilde{d} \leqslant \langle Z_j - x, v_{b_j} \rangle &= \langle Z_j - \widetilde{\mu}, v_{b_j} \rangle + \langle \widetilde{\mu} - x, v_{b_j} \rangle \\ &< X_{b_j,b_j} r^* + \|\widetilde{\mu} - x\| = X_{b_j,b_j} r^* + \tilde{d}, \end{aligned}$$

where the first inequality follows from the inclusion of j in $\mathcal{R}$ and the last inequality follows from the inclusion of j in $\mathcal{T}$ and Cauchy-Schwarz. By rearranging the above inequality, we get:

$$X_{b_j,b_j} > (1.0625\tilde{d} - \tilde{d})(r^*)^{-1} > 1,$$

which is a contradiction. Therefore, we get from the monotonicity of $\mathbf{MT}(x, r, \boldsymbol{Z})$ (see Lemma 4.6), that $d' \leqslant 1.25\tilde{d}$ and this concludes the proof of the lemma. □

Gradient Estimation Step

In this section, we analyze the gradient estimation step of the algorithm. We show that an approximate gradient can be found as long as the current point x is not too close to $\widetilde{\mu}$. The following lemma shows that Algorithm 7 produces a nontrivial estimate of the gradient.

Lemma 4.20. *Suppose a point x satisfies $\|x - \widetilde{\mu}\| \geqslant 20r^*$ and let Δ be the unit vector along $\widetilde{\mu} - x$. Then, Algorithm 7 returns a vector g satisfying:*

$$\langle g, \Delta \rangle \geqslant \frac{1}{15}.$$

Proof. In the running of Algorithm 7, let X denote the solution of $\mathbf{MT}(x, d^*, \boldsymbol{Z})$. We begin by factorizing the solution X as $UU^\top$, with the rows of U denoted by u_1, $u_{b_1}, \ldots, u_{b_k}$ and $u_{v_1}, \ldots, u_{v_d}$. We also define the matrix $U_v = (u_{v_1}, \ldots, u_{v_d})$ in $\mathbb{R}^{(k+d+1)\times d}$. From the constraints in $\mathbf{MT}$, we have:

$$X_{b_i,b_i} = \|u_{b_i}\|^2 \leqslant 1 \implies \|u_{b_i}\| \leqslant 1, \quad \sum_{j=1}^{d} X_{v_j,v_j} = \sum_{j=1}^{d} \|u_{v_j}\|^2 = \|U_v\|_F^2 = 1 \implies \|U_v\|_F = 1.$$

Let $\mathcal{R}$ and $\mathcal{T}$ denote the sets defined in Lemma 4.17. Let $j \in \mathcal{T} \cap \mathcal{R}$. By noting that $v_{b_j} = u_{b_j}^\top U_v$, we have:

$$0.85d^* \leqslant \langle Z_j - \widetilde{\mu}, v_{b_j} \rangle + \langle \widetilde{\mu} - x, v_{b_j} \rangle \leqslant X_{b_j,b_j} r^* + u_{b_j}^\top U_v (\widetilde{\mu} - x),$$

where the first inequality follows from the inclusion of j in $\mathcal{R}$ and the second from its inclusion in $\mathcal{T}$. By rearranging this equation and using our bound on d^* from Lemma 4.19, we obtain:

$$0.80\|\widetilde{\mu} - x\| \leqslant 0.85d^* \leqslant X_{b_j,b_j} r^* + u_{b_j}^\top U_v (\widetilde{\mu} - x). \tag{4.3}$$

By rearranging (4.3), using Cauchy-Schwarz, $\|u_{b_i}\| \leqslant 1$ and the assumption on $\|x - \widetilde{\mu}\|$:

$$\|U_v(\widetilde{\mu} - x)\| \geqslant u_{b_j}^\top U_v (\widetilde{\mu} - x) \geqslant 0.75\|\widetilde{\mu} - x\|,$$

which yields:

$$\|U_v \Delta\| \geqslant 0.75.$$

Now, we have:

$$1 = \|U_v\|_F^2 = \|U_v \mathcal{P}_\Delta\|_F^2 + \|U_v \mathcal{P}_\Delta^\perp\|_F^2 \geqslant \|U_v \mathcal{P}_\Delta^\perp\|_F^2 + (0.75)^2 \implies \|U_v \mathcal{P}_\Delta^\perp\|_F \leqslant 0.67.$$

Let y be the top singular vector of X_v. Note that $X_v = U_v^\top U_v$ and y is also the top right singular vector of U_v. We have that:

$$0.75 \leqslant \|U_v y\| \leqslant \|U_v \mathcal{P}_\Delta y\| + \|U_v \mathcal{P}_\Delta^\perp y\| \leqslant \|\mathcal{P}_\Delta y\| + \|U_v \mathcal{P}_\Delta^\perp\|_F \leqslant \|\mathcal{P}_\Delta y\| + 0.67.$$

This means that we have:

$$|\langle y, \Delta \rangle| \geqslant \frac{1}{15}.$$

Note that the algorithm returns either y or $-y$. Consider the case where $\langle y, \Delta \rangle > 0$. From Remark 4.1 (implied by Definition 4.2), we have for at least $0.95k$ points:

$$\langle Z_i - \widetilde{\mu}, y \rangle \leqslant r^*.$$

Therefore, we have for $0.95k$ points:

$$\langle Z_i - x, y \rangle = \langle Z_i - \widetilde{\mu}, y \rangle + \langle \widetilde{\mu} - x, y \rangle \geqslant -r^* + \frac{20r^*}{15} > 0.$$

This means that in the case where $\langle y, \Delta \rangle > 0$, we return y which satisfies $\langle \widetilde{\mu} - x, y \rangle > 0$. This implies the lemma in this case. The case where $\langle y, \Delta \rangle < 0$ is similar, with $-y$ used instead of y. This concludes the proof of the lemma. □

We now prove a lemma regarding the output of Algorithm 2.

Lemma 4.21. *Let $\boldsymbol{Z} = \{Z_i\}_{i=1}^k$ be k points in $\mathbb{R}^d$, $\widetilde{\mu} \in \mathbb{R}^d$ and r^* be as in Definition 4.2. Then, Algorithm 2, with input $\boldsymbol{Z}$, initialization $x^\dagger$, and number of iterations $T \geqslant 10^6 \log \frac{\|\widetilde{\mu} - x^\dagger\|}{\epsilon}$ satisfies:*

$$\|x^* - \widetilde{\mu}\| \leqslant \max\{30r^*, \epsilon\}.$$

Proof. First, let $\mathcal{G} = \{x : \|x - \widetilde{\mu}\| \leqslant 20r^*\}$. We prove the lemma in two cases:

Case 1: None of the iterates x_t lie in $\mathcal{G}$. In this case, we have from Lemma 4.19:

$$0.95\|x_t - \widetilde{\mu}\| \leqslant d_t \leqslant 1.25\|x_t - \widetilde{\mu}\|. \tag{4.4}$$

This yields:

$$\begin{aligned}\|x_{t+1} - \widetilde{\mu}\|^2 &= \|x_t - \widetilde{\mu}\|^2 - 2\frac{d_t}{20}\langle g_t, \widetilde{\mu} - x_t \rangle + \frac{d_t^2}{400} \leqslant \|x_t - \widetilde{\mu}\|^2 - \frac{d_t\|\widetilde{\mu} - x_t\|}{150} + \frac{d_t^2}{400} \\ &\leqslant \|x_t - \widetilde{\mu}\|^2 - d_t\left(\frac{\|\mu - x_t\|}{150} - \frac{d_t}{400}\right) \leqslant \left(1 - \frac{1}{500}\right)\|x_t - \widetilde{\mu}\|^2,\end{aligned}$$

where the first inequality follows from Lemma 4.20 and the last inequality follows by substituting the lower bound on d_t in the first term and the upper bound on d_t in the second term ((4.4)). By an iterated application of the above inequality, we have:

$$\|x^* - \widetilde{\mu}\| \leqslant \frac{1}{0.95} \cdot d^* \leqslant \frac{1}{0.95} \cdot d_{T+1} \leqslant \frac{\epsilon}{10},$$

which concludes the proof of the lemma in this case.

Case 2: One of the iterates x_t falls into the set $\mathcal{G}$. If the algorithm returns an element from $\mathcal{G}$, the lemma is true from the definition of $\mathcal{G}$. Otherwise, from Lemma 4.18, we have for $x_t \in \mathcal{G}$ that $d_t \leqslant 25r^*$. Therefore, we have at the conclusion of the algorithm a value $d^* \leqslant 25r^*$ along with a returned vector x^* lying outside $\mathcal{G}$. Thus, we have from Lemma 4.19:

$$0.95\|x^* - \widetilde{\mu}\| \leqslant 25r^* \implies \|x^* - \widetilde{\mu}\| \leqslant 30r^*.$$

This concludes the proof of the lemma. □

4.10 Proof of Theorem 4.2

We assemble the results established in other sections to prove Theorem 4.2. Let $x^\dagger$ denote the output of Algorithm 4 in the running of Algorithm 1. Note that this is passed as input to Algorithms 3 and 5. We now define the event $\mathcal{E}$:

$$\mathcal{E} := \{\|x^\dagger - \mu\| \leqslant 24\sqrt{d}\}.$$

From Lemmas 4.11 and 4.16, we may assume the conclusions of Lemma 4.16, an event which occurs with probability at least $1 - \delta$. Since $\|x^\dagger - \mu\| \leqslant 24\sqrt{d}$, the proof of the theorem follows from Lemma 4.21, our bound on the number of iterations T and the final desired accuracy.

□

4.11 Lower Bound for Robust Estimation under Weak Moments

In this section, we establish a lower bound for robust mean estimation under weak moments. The lower bound will be a consequence of the following theorem:

Theorem 4.8. *Given $\eta, \alpha \in (0, 1)$, there exist two distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ over $\mathbb{R}$ with means μ_1 and μ_2, respectively, satisfying:*

1. $d_{TV}(\mathcal{D}_1, \mathcal{D}_2) \leqslant \frac{\eta}{4}$
2. $|\mu_1 - \mu_2| \geqslant \frac{1}{4} \cdot \eta^{\alpha/(1+\alpha)}$
3. $\mathbb{E}_{X\sim\mathcal{D}_1}[|X - \mu_1|^{1+\alpha}], \mathbb{E}_{X\sim\mathcal{D}_2}[|X - \mu_2|^{1+\alpha}] \leqslant 1.$

Proof. We prove the theorem by explicit construction. Let $\mathcal{D}_1$ be a δ-distribution on 0: $\mathbb{P}_{X\sim\mathcal{D}_1}(X = 0) = 1$. We have $\mu_1 = 0$ and the weak moment condition holds trivially for $\mathcal{D}_1$. Now, for $\mathcal{D}_2$, we have:

$$\mathbb{P}_{X\sim\mathcal{D}_2}(X=x)=\begin{cases}1-\frac{\eta}{4}, & \text{when } x=0\\ \frac{\eta}{4}, & \text{when } x=\left(\frac{1}{\eta}\right)^{1/(1+\alpha)}\\ 0, & \text{otherwise.}\end{cases}.$$

From the definitions of $\mathcal{D}_1$ and $\mathcal{D}_2$, we obtain the first conclusion of the lemma. By direct computation, we have $\mu_1 = 0$ and $\mu_2 = \frac{1}{4}\cdot\eta^{\alpha/(1+\alpha)}$. Finally, we verify the weak moment condition on $\mathcal{D}_2$ using the convexity of the function $f(x)=|x|^{1+\alpha}$:

$$\mathbb{E}_{X\sim\mathcal{D}_2}[|X-\mu_2|^{1+\alpha}]\leqslant 2^\alpha\cdot\mathbb{E}[|X|^{1+\alpha}+|\mu_2|^{1+\alpha}]\leqslant 2^\alpha\left(\frac{1}{4}+\frac{\eta^\alpha}{4^{(1+\alpha)}}\right)\leqslant 1.$$

This concludes the proof of the theorem. □

4.12 Lower Bound for the Bounded Covariance Setting

In this section, we a present a proof of the lower bound for high-dimensional mean estimation in the bounded covariance setting (with $\alpha = 1$) in the constant probability regime; i.e, a lower bound on the best attainable confidence interval for any estimation algorithm which succeeds with probability at least $1/2$. The main result of the section is the following lemma used to prove Theorem 4.5.

Lemma 4.22. *Let $n > 0, d \geqslant 2$. Then, under the following Bayesian generative process:*

1. *Draw μ from $\mathcal{N}(0, I)$*
2. *Draw $\boldsymbol{X} \coloneqq X_1, \ldots, X_n$ iid from $\mathcal{N}(\mu, I)$,*

we have for any estimator $\hat{\mu}(\boldsymbol{X})$:

$$\mathbb{P}\left\{\|\mu-\hat{\mu}(\boldsymbol{X})\|\geqslant\frac{\sqrt{d-2}-2\sqrt{\log 2}}{\sqrt{n+1}}\right\}\geqslant\frac{1}{2}.$$

Proof. Note that the posterior distribution over μ upon observation of the samples $\boldsymbol{X}$ remains a Gaussian; indeed, we have:

$$\begin{aligned}f(\mu\mid\boldsymbol{X})&\propto\exp\left\{-\frac{\|\mu\|^2+\sum_{i=1}^n\|X_i-\mu\|^2}{2}\right\}\propto\exp\left\{-\frac{(n+1)\|\mu\|^2-2\langle\sum_{i=1}^n X_i,\mu\rangle}{2}\right\}\\&\propto\exp\left\{-\frac{\|\mu\|^2-2\langle\sum_{i=1}^n X_i/(n+1),\mu\rangle}{2\cdot(n+1)}\right\}\propto\exp\left\{-\frac{\|\mu-\sum_{i=1}^n X_i/(n+1)\|^2}{2\cdot(n+1)}\right\}\end{aligned}$$

Hence, the posterior distribution over μ is a Gaussian with covariance $I/(n+1)$. Letting $\widehat{\mu}(\boldsymbol{X})$ be any estimator of μ, we have from Lemma 4.23:

$$\forall r>0:\mathbb{P}\left\{\|\mu-\hat{\mu}(\boldsymbol{X})\|\leqslant r\mid\boldsymbol{X}\right\}\leqslant\mathbb{P}\left\{\|\mu-\frac{\sum_{i=1}^n X_i}{(n+1)}\|\leqslant r\mid\boldsymbol{X}\right\}. \tag{4.5}$$

Hence, we have from Lemma 4.24, the fact that $\|g\|$ is a Lipschitz function of g:

$$\mathbb{P}\left\{\|\mu - \frac{\sum_{i=1}^n X_i}{(n+1)}\| \geqslant \frac{\sqrt{d-2} - 2\sqrt{\log 2}}{\sqrt{n+1}} \mid \boldsymbol{X}\right\} \geqslant \frac{1}{2}.$$

With (4.5), taking expectation over $\boldsymbol{X}$ in the above display yields:

$$\mathbb{P}\left\{\|\mu - \hat{\mu}(\boldsymbol{X})\| \geqslant \frac{\sqrt{d-2} - 2\sqrt{\log 2}}{\sqrt{n+1}}\right\} \geqslant \frac{1}{2}.$$

□

The following two lemmas feature in the proof of Lemma 4.22.

Lemma 4.23. *For $g \sim \mathcal{N}(0, I)$, we have:*

$$\forall r > 0 : \arg\max_c \mathbb{P}\left\{\|g - c\| \leqslant r\right\} = 0.$$

Proof. Fix $r > 0$ and consider the functions:

$$\forall \|v\| = 1, t \geqslant 0 : f_v(t) \coloneqq \log \int_{\|z\| \leqslant r} \exp\left\{-\frac{\|z + tv\|^2}{2}\right\} dz.$$

We will prove that $f_v(t)$ is maximized at $t = 0$ for all $\|v\| = 1$ which will imply the lemma. To do so, fix $\|v\| = 1$ and observe that the derivative of f_v, f_v' has the following expression:

$$f_v'(t) = -\frac{1}{\int_{\|z\| \leqslant r} \exp\left\{-\frac{\|z+tv\|^2}{2}\right\} dz} \cdot \int_{\|z\| \leqslant r} \exp\left\{-\frac{\|z + tv\|^2}{2}\right\} (t + \langle z, v\rangle) dz$$

Note that when $t \geqslant r$, the above expression is trivially negative as it corresponds to a negative conditional expectation of the (non-negative) quantity $(t + \langle z, x\rangle)$ which is positive on a set of non-zero measure. When $t < r$, we have:

$$\begin{aligned}
&\int_{\|z\| \leqslant r} \exp\left\{-\frac{\|z + tv\|^2}{2}\right\} (t + \langle z, v\rangle) dz \\
&= \int_{w=t-r}^{t+r} w e^{-w^2/2} \int_{\|y\| \leqslant \sqrt{r^2 - (w-t)^2}} e^{-\|y\|^2/2} dy dw \geqslant \\
&\int_{w=t-r}^{r-t} w e^{-w^2/2} \int_{\|y\| \leqslant \sqrt{r^2 - (w-t)^2}} e^{-\|y\|^2/2} dy dw \\
&\geqslant \int_{w=0}^{r-t} w e^{-w^2/2} \left[\int_{\|y\| \leqslant \sqrt{r^2-(w-t)^2}} e^{-\|y\|^2/2} dy - \int_{\|y\| \leqslant \sqrt{r^2-(w+t)^2}} e^{-\|y\|^2/2} dy\right] dw > 0
\end{aligned}$$

where the last inequality follows from the fact that the inner difference of integrals is positive when $w \in (0, r - t)$. Hence, $f_v'(t) < 0$ for all $t > 0$. Noting that $f_v(\cdot)$ is continuous, we get that $f_v(t)$ is *uniquely* maximized at $t = 0$. Now, the lemma follows from fact that for any $c \neq 0$, we have $c = tv$ for some $\|v\| = 1$ and $t > 0$ and therefore, $f_v(t) < f_v(0)$. □

Lemma 4.24. *We have for all $d \geqslant 2$ and $g \sim \mathcal{N}(0, I)$*

$$\mathbb{E}[\|g\|] \geqslant \sqrt{d-2}.$$

Proof. We have $\mathbb{E}[\|g\|^2] = d$. Furthermore, note that $\|g\|$ is a Lipschitz function of g and hence, we have [11, Theorem 5.6]:

$$\mathbb{P}\left\{|\|g\| - \mathbb{E}[\|g\|]| \geqslant t\right\} \leqslant 2\exp\left\{-\frac{t^2}{2}\right\}.$$

By integrating the tails we have:

$$\mathbb{E}[(\|g\| - \mathbb{E}[\|g\|])^2] \leqslant 2 \implies \mathbb{E}[\|g\|] = \sqrt{\mathbb{E}[\|g\|^2] - \mathbb{E}[(\|g\| - \mathbb{E}[\|g\|])^2]} \geqslant \sqrt{d-2}.$$

□

Bibliography

[1] N. Alon, Y. Matias, and M. Szegedy. "The space complexity of approximating the frequency moments". In: *Journal of Computer and System Sciences* 58.1 (1999), pp. 137–147.

[2] Animashree Anandkumar, Daniel Hsu, and Sham M Kakade. "A method of moments for mixture models and hidden Markov models". In: *Conference on Learning Theory*. 2012, pp. 33–1.

[3] Rie Kubota Ando and Tong Zhang. "A framework for learning predictive structures from multiple tasks and unlabeled data". In: *Journal of Machine Learning Research* 6.Nov (2005), pp. 1817–1853.

[4] Alexei Baevski et al. "Cloze-driven pretraining of self-attention networks". In: *arXiv preprint arXiv:1903.07785* (2019).

[5] Peter L Bartlett, Olivier Bousquet, Shahar Mendelson, et al. "Local rademacher complexities". In: *The Annals of Statistics* 33.4 (2005), pp. 1497–1537.

[6] Jonathan Baxter. "A model of inductive bias learning". In: *Journal of artificial intelligence research* 12 (2000), pp. 149–198.

[7] Shai Ben-David and Reba Schuller Borbely. "A notion of task relatedness yielding provable multiple-task learning guarantees". In: *Machine learning* 73.3 (2008), pp. 273–287.

[8] Yoshua Bengio, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives". In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), pp. 1798–1828.

[9] Rajendra Bhatia. *Matrix analysis*. Vol. 169. Springer Science & Business Media, 2013.

[10] Peter J Bickel et al. *Efficient and adaptive estimation for semiparametric models*. Vol. 4. Johns Hopkins University Press Baltimore, 1993.

[11] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

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

[13] EJ Candes and Y Plan. "Tight oracle bounds for low-rank matrix recovery from a mininal number of noisy random measurements". In: *arXiv preprint arXiv:1001.0339* (2010).

[14] Rich Caruana. "Multitask learning". In: *Machine learning* 28.1 (1997), pp. 41–75.

[15] O. Catoni and I. Giulini. "Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression". In: *arXiv preprint arXiv:1712.02747* (2017).

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

[17] Moses Charikar, Jacob Steinhardt, and Gregory Valiant. "Learning from untrusted data". In: *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017, Montreal, QC, Canada, June 19-23, 2017*. Ed. by Hamed Hatami, Pierre McKenzie, and Valerie King. ACM, 2017, pp. 47–60. ISBN: 978-1-4503-4528-6. DOI: 10.1145/3055399.3055491. URL: https://doi.org/10.1145/3055399.3055491.

[18] Xi Chen, Adityanand Guntuboyina, and Yuchen Zhang. "On Bayes risk lower bounds". In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 7687–7744.

[19] Yu Cheng, Ilias Diakonikolas, and Rong Ge. "High-dimensional robust mean estimation in nearly-linear time". In: *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019*. Ed. by Timothy M. Chan. SIAM, 2019, pp. 2755–2771. ISBN: 978-1-61197-548-2. DOI: 10.1137/1.9781611975482.171. URL: https://doi.org/10.1137/1.9781611975482.171.

[20] Yeshwanth Cherapanamjeri, Nicolas Flammarion, and Peter L. Bartlett. "Fast Mean Estimation with Sub-Gaussian Rates". In: *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*. 2019, pp. 786–806.

[21] Yeshwanth Cherapanamjeri et al. "Algorithms for heavy-tailed statistics: regression, covariance estimation, and beyond". In: *Proccedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC 2020, Chicago, IL, USA, June 22-26, 2020*. Ed. by Konstantin Makarychev et al. ACM, 2020, pp. 601–609. ISBN: 978-1-4503-6979-4. DOI: 10.1145/3357713.3384329. URL: https://doi.org/10.1145/3357713.3384329.

[22] Yeshwanth Cherapanamjeri et al. *Optimal Mean Estimation without a Variance*. 2020. arXiv: 2011.12433 [math.ST].

[23] Alexander D'Amour et al. "Underspecification presents challenges for credibility in modern machine learning". In: *Journal of Machine Learning Research* (2020). forthcoming.

[24] Giulia Denevi et al. "Learning-to-learn stochastic gradient descent with biased regularization". In: *arXiv preprint arXiv:1903.10399* (2019).

[25] Giulia Denevi et al. "Online-Within-Online Meta-Learning". In: *Advances in Neural Information Processing Systems*. 2019, pp. 13089–13099.

[26] Jules Depersin. *A spectral algorithm for robust regression with subgaussian rates*. 2020. arXiv: 2007.06072 [stat.ML].

[27] Jules Depersin. *Robust subgaussian estimation with VC-dimension*. 2020. arXiv: 2004.11734 [stat.ML].

[28] Jules Depersin and Guillaume Lecué. *Robust subgaussian estimation of a mean vector in nearly linear time*. 2019. arXiv: 1906.03058 [math.ST].

[29] L. Devroye et al. "Sub-Gaussian mean estimators". In: *Ann. Statist.* 44.6 (2016), pp. 2695–2725.

[30] I. Diakonikolas and D. M Kane. "Recent advances in algorithmic high-dimensional robust statistics". In: *arXiv preprint arXiv:1911.05911* (2019).

[31] Ilias Diakonikolas, Daniel M. Kane, and Ankit Pensia. "Outlier Robust Mean Estimation with Subgaussian Rates via Stability". In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. Ed. by Hugo Larochelle et al. 2020.

[32] Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. "List-decodable robust mean estimation and learning mixtures of spherical gaussians". In: *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018, Los Angeles, CA, USA, June 25-29, 2018*. Ed. by Ilias Diakonikolas, David Kempe, and Monika Henzinger. ACM, 2018, pp. 1047–1060. DOI: 10.1145/3188745.3188758. URL: https://doi.org/10.1145/3188745.3188758.

[33] Ilias Diakonikolas, David Kempe, and Monika Henzinger, eds. *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018, Los Angeles, CA, USA, June 25-29, 2018*. ACM, 2018. URL: http://dl.acm.org/citation.cfm?id=3188745.

[34] Ilias Diakonikolas et al. "Being robust (in high dimensions) can be practical". In: *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, 2017, pp. 999–1008. URL: http://proceedings.mlr.press/v70/diakonikolas17a.html.

[35] Ilias Diakonikolas et al. "Robust estimators in high dimensions without the computational intractability". In: *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 9-11 October 2016, Hyatt Regency, New Brunswick, New Jersey, USA*. Ed. by Irit Dinur. IEEE Computer Society, 2016, pp. 655–664. ISBN: 978-1-5090-3933-3. DOI: 10.1109/FOCS.2016.85. URL: https://doi.org/10.1109/FOCS.2016.85.

[36] Ilias Diakonikolas et al. "Robustly learning a gaussian: getting optimal error, efficiently". In: *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2018, New Orleans, LA, USA, January 7-10, 2018*. Ed. by Artur Czumaj. SIAM, 2018, pp. 2683–2702. ISBN: 978-1-61197-503-1. DOI: 10.1137/1.9781611975031.171. URL: https://doi.org/10.1137/1.9781611975031.171.

[37] Irit Dinur, ed. *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 9-11 October 2016, Hyatt Regency, New Brunswick, New Jersey, USA*. IEEE Computer Society, 2016. ISBN: 978-1-5090-3933-3. URL: https://ieeexplore.ieee.org/xpl/conhome/7781469/proceeding.

[38] Jeff Donahue et al. "Decaf: A deep convolutional activation feature for generic visual recognition". In: *International conference on machine learning*. 2014, pp. 647–655.

[39] Yihe Dong, Samuel B. Hopkins, and Jerry Li. "Quantum entropy scoring for fast robust mean estimation and improved outlier detection". In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*. Ed. by Hanna M. Wallach et al. 2019, pp. 6065–6075. URL: http://papers.nips.cc/paper/8839-quantum-entropy-scoring-for-fast-robust-mean-estimation-and-improved-outlier-detection.

[40] Simon S Du et al. "Few-shot learning via learning the representation, provably". In: *arXiv preprint arXiv:2002.09434* (2020).

[41] Alan Edelman, Tomás A Arias, and Steven T Smith. "The geometry of algorithms with orthogonality constraints". In: *SIAM journal on Matrix Analysis and Applications* 20.2 (1998), pp. 303–353.

[42] Ahmed Elnaggar et al. "End-to-end multitask learning, from protein language to protein features without alignments". In: *bioRxiv* (2020). DOI: 10.1101/864405.

[43] 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*. JMLR. org. 2017, pp. 1126–1135.

[44] Chelsea Finn et al. "Online meta-learning". In: *arXiv preprint arXiv:1902.08438* (2019).

[45] Saurabh Garg et al. "On Proximal Policy Optimization's Heavy-tailed Gradients". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 3610–3619.

[46] Rong Ge, Chi Jin, and Yi Zheng. "No Spurious Local Minima in Nonconvex Low Rank Problems: A Unified Geometric Analysis". In: *arXiv preprint arXiv:1704.00708* (2017).

[47] Noah Golowich, Alexander Rakhlin, and Ohad Shamir. "Size-independent sample complexity of neural networks". In: *arXiv preprint arXiv:1712.06541* (2017).

[48] Varun Gulshan et al. "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs". In: *Jama* 316.22 (2016), pp. 2402–2410.

[49] Adityanand Guntuboyina. "Lower bounds for the minimax risk using f-divergences, and applications". In: *IEEE Transactions on Information Theory* 57.4 (2011), pp. 2386–2399.

[50] Dan Hendrycks et al. "The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2021, pp. 8340–8349.

[51] Samuel B. Hopkins. "Mean estimation with sub-Gaussian rates in polynomial time". In: *Ann. Statist.* 48.2 (2020), pp. 1193–1213. ISSN: 0090-5364. DOI: 10.1214/19-AOS1843. URL: https://doi.org/10.1214/19-AOS1843.

[52] Samuel B. Hopkins, Gautam Kamath, and Mahbod Majid. "Efficient Mean Estimation with Pure Differential Privacy via a Sum-of-Squares Exponential Mechanism". In: *CoRR* abs/2111.12981 (2021). arXiv: 2111.12981. URL: https://arxiv.org/abs/2111.12981.

[53] Samuel B. Hopkins, Jerry Li, and Fred Zhang. "Robust and Heavy-Tailed Mean Estimation Made Simple, via Regret Minimization". In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. Ed. by Hugo Larochelle et al. 2020. URL: https://proceedings.neurips.cc/paper/2020/hash/8a1276c25f5efe85f0fc4020fbf5b4f8-Abstract.html.

[54] Timothy Hospedales et al. "Meta-learning in neural networks: A survey". In: *arXiv preprint arXiv:2004.05439* (2020).

[55] Daniel Hsu, Sham M Kakade, and Tong Zhang. "Random design analysis of ridge regression". In: *Conference on learning theory*. 2012, pp. 9–1.

[56] Peter J. Huber. "Robust estimation of a location parameter". In: *Ann. Math. Statist.* 35 (1964), pp. 73–101. ISSN: 0003-4851. DOI: 10.1214/aoms/1177703732. URL: https://doi.org/10.1214/aoms/1177703732.

[57] Mark Jerrum, Leslie G. Valiant, and Vijay V. Vazirani. "Random Generation of Combinatorial Structures from a Uniform Distribution". In: *Theor. Comput. Sci.* 43 (1986), pp. 169–188. DOI: 10.1016/0304-3975(86)90174-X. URL: https://doi.org/10.1016/0304-3975(86)90174-X.

[58] Chi Jin et al. "How to escape saddle points efficiently". In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org. 2017, pp. 1724–1732.

[59] Sham M Kakade et al. "Efficient learning of generalized linear and single index models with isotonic regression". In: *Advances in Neural Information Processing Systems*. 2011, pp. 927–935.

[60] Mikhail Khodak, Maria-Florina Balcan, and Ameet Talwalkar. "Provable guarantees for gradient-based meta-learning". In: *arXiv preprint arXiv:1902.10644* (2019).

[61] Mikhail Khodak, Maria-Florina F Balcan, and Ameet S Talwalkar. "Adaptive gradient-based meta-learning methods". In: *Advances in Neural Information Processing Systems*. 2019, pp. 5915–5926.

[62] Pang Wei Koh et al. "WILDS: A Benchmark of in-the-Wild Distribution Shifts". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, July 2021, pp. 5637–5664. URL: https://proceedings.mlr.press/v139/koh21a.html.

[63] Pravesh K. Kothari, Jacob Steinhardt, and David Steurer. "Robust moment estimation and improved clustering via sum of squares". In: *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018, Los Angeles, CA, USA, June 25-29, 2018*. Ed. by Ilias Diakonikolas, David Kempe, and Monika Henzinger. ACM, 2018, pp. 1035–1046. DOI: 10.1145/3188745.3188970. URL: https://doi.org/10.1145/3188745.3188970.

[64] Kevin A. Lai, Anup B. Rao, and Santosh S. Vempala. "Agnostic estimation of mean and covariance". In: *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 9-11 October 2016, Hyatt Regency, New Brunswick, New Jersey, USA*. Ed. by Irit Dinur. IEEE Computer Society, 2016, pp. 665–674. ISBN: 978-1-5090-3933-3. DOI: 10.1109/FOCS.2016.76. URL: https://doi.org/10.1109/FOCS.2016.76.

[65] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *nature* 521.7553 (2015), pp. 436–444.

[66] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Vol. 23. Springer Science & Business Media, 1991.

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

[68] Kwonjoon Lee et al. "Meta-learning with differentiable convex optimization". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 10657–10665.

[69] Zhixian Lei et al. "A fast spectral algorithm for mean estimation with sub-Gaussian rates". In: *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*. Ed. by Jacob D. Abernethy and Shivani Agarwal. Vol. 125. Proceedings of Machine Learning Research. PMLR, 2020, pp. 2598–2612. URL: http://proceedings.mlr.press/v125/lei20a.html.

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

[71] Dong C Liu and Jorge Nocedal. "On the limited memory BFGS method for large scale optimization". In: *Mathematical programming* 45.1-3 (1989), pp. 503–528.

[72] Xiaodong Liu et al. "Multi-task deep neural networks for natural language understanding". In: *arXiv preprint arXiv:1901.11504* (2019).

[73] Hedibert Freitas Lopes, Matthew Taddy, and Matthew Gardner. "Scalable Semiparametric Inference for the Means of Heavy-tailed Distributions". In: *Topics in Identification, Limited Dependent Variables, Partial Observability, Experimentation, and Flexible Modeling: Part B*. Emerald Publishing Limited, 2019.

[74] Karim Lounici et al. "Oracle inequalities and optimal inference under group sparsity". In: *The annals of statistics* 39.4 (2011), pp. 2164–2204.

[75] G. Lugosi and S. Mendelson. "Sub-Gaussian estimators of the mean of a random vector". In: *Ann. Statist.* 47.2 (Apr. 2019), pp. 783–794.

[76] Gabor Lugosi and Shahar Mendelson. *Robust multivariate mean estimation: the optimality of trimmed mean*. 2019. arXiv: 1907.11391 [math.ST].

[77] Gábor Lugosi and Shahar Mendelson. "Mean estimation and regression under heavy-tailed distributions: A survey". In: *Found. Comput. Math.* 19.5 (2019), pp. 1145–1190. DOI: 10.1007/s10208-019-09427-x. URL: https://doi.org/10.1007/s10208-019-09427-x.

[78] Gábor Lugosi and Shahar Mendelson. "Risk minimization by median-of-means tournaments". In: *J. Eur. Math. Soc. (JEMS)* 22.3 (2020), pp. 925–965. ISSN: 1435-9855. DOI: 10.4171/jems/937. URL: https://doi.org/10.4171/jems/937.

[79] Gábor Lugosi and Shahar Mendelson. "Sub-Gaussian estimators of the mean of a random vector". In: *The Annals of Statistics* 47.2 (2019), pp. 783–794.

[80] Dougal Maclaurin, David Duvenaud, and Ryan P Adams. "Autograd: Effortless gradients in numpy". In: *ICML 2015 AutoML Workshop*. Vol. 238. 2015.

[81] Benoit B Mandelbrot. "The variation of certain speculative prices". In: *Fractals and scaling in finance*. Springer, 1997, pp. 371–418.

[82] Pascal Massart et al. "About the constants in Talagrand's concentration inequalities for empirical processes". In: *The Annals of Probability* 28.2 (2000), pp. 863–884.

[83] Andreas Maurer. "A chain rule for the expected suprema of Gaussian processes". In: *Theoretical Computer Science* 650 (2016), pp. 109–122.

[84] Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. "The benefit of multitask representation learning". In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 2853–2884.

[85] Shahar Mendelson and Nikita Zhivotovskiy. "Robust covariance estimation under $L_4 - L_2$ norm equivalence". In: *Ann. Statist.* 48.3 (2020), pp. 1648–1664. ISSN: 0090-5364. DOI: 10.1214/19-AOS1862. URL: https://doi.org/10.1214/19-AOS1862.

[86] John Miller et al. "The Effect of Natural Distribution Shift on Question Answering Models". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, July 2020, pp. 6905–6916. URL: http://proceedings.mlr.press/v119/miller20a.html.

[87] Philipp Moritz et al. "Ray: A distributed framework for emerging {AI} applications". In: *13th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 18)*. 2018, pp. 561–577.

[88] Arkadi Nemirovski. "Topics in non-parametric statistics". In: *Lectures on Probability Theory and Statistics (Saint-Flour, 1998)*. Vol. 1738. Lecture Notes in Math. Springer, Berlin, 2000, pp. 85–277.

[89] A. S. Nemirovsky and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience Series in Discrete Mathematics. John Wiley & Sons, 1983.

[90] Y. Nesterov. "Semidefinite relaxation and nonconvex quadratic optimization". In: *Optimization Methods and Software* 9.1-3 (1998), pp. 141–160.

[91] Guillaume Obozinski, Martin J Wainwright, Michael I Jordan, et al. "Support union recovery in high-dimensional multivariate regression". In: *The Annals of Statistics* 39.1 (2011), pp. 1–47.

[92] Jakub Otwinowski, David M McCandlish, and Joshua B Plotkin. "Inferring the shape of global epistasis". In: *Proceedings of the National Academy of Sciences* 115.32 (2018), E7550–E7558.

[93] Yaniv Ovadia et al. "Can you trust your models uncertainty? Evaluating predictive uncertainty under dataset shift". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/file/8558cb408c1d76621371888657d2eb1d-Paper.pdf.

[94] Alain Pajor. "Metric entropy of the Grassmann manifold". In: *Convex Geometric Analysis* 34 (1998), pp. 181–188.

[95] Karl Pearson. "Contributions to the mathematical theory of evolution". In: *Philosophical Transactions of the Royal Society of London. A* 185 (1894), pp. 71–110.

[96] Massimiliano Pontil and Andreas Maurer. "Excess risk bounds for multitask learning with trace norm regularization". In: *Conference on Learning Theory*. 2013, pp. 55–76.

[97] Aniruddh Raghu et al. "Rapid Learning or Feature Reuse? Towards Understanding the Effectiveness of MAML". In: *International Conference on Learning Representations*. 2020. URL: https://openreview.net/forum?id=rkgMkCEtPB.

[98] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. "Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls". In: *IEEE transactions on information theory* 57.10 (2011), pp. 6976–6994.

[99] Benjamin Recht. "A simpler approach to matrix completion". In: *Journal of Machine Learning Research* 12.Dec (2011), pp. 3413–3430.

[100] Benjamin Recht et al. "Do cifar-10 classifiers generalize to cifar-10?" In: *arXiv preprint arXiv:1806.00451* (2018).

[101] Benjamin Recht et al. "Do ImageNet Classifiers Generalize to ImageNet?" In: *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 5389–5400. URL: http://proceedings.mlr.press/v97/recht19a.html.

[102] Angelika Rohde, Alexandre B Tsybakov, et al. "Estimation of high-dimensional low-rank matrices". In: *The Annals of Statistics* 39.2 (2011), pp. 887–930.

[103] Jacob Steinhardt, Moses Charikar, and Gregory Valiant. "Resilience: A criterion for learning in the presence of arbitrary outliers". In: *9th Innovations in Theoretical Computer Science Conference, ITCS 2018, January 11-14, 2018, Cambridge, MA, USA*. Ed. by Anna R. Karlin. Vol. 94. LIPIcs. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2018, 45:1–45:21. ISBN: 978-3-95977-060-6. DOI: 10.4230/LIPIcs.ITCS.2018.45. URL: https://doi.org/10.4230/LIPIcs.ITCS.2018.45.

[104] Nilesh Tripuraneni, Chi Jin, and Michael Jordan. "Provable Meta-Learning of Linear Representations". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, July 2021, pp. 10434–10443. URL: https://proceedings.mlr.press/v139/tripuraneni21a.html.

[105] Nilesh Tripuraneni, Michael Jordan, and Chi Jin. "On the Theory of Transfer Learning: The Importance of Task Diversity". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 7852–7862. URL: https://proceedings.neurips.cc/paper/2020/file/59587bffec1c7846f3e34230141556ae-Paper.pdf.

[106] Nilesh Tripuraneni et al. "Assessment of Treatment Effect Estimators for Heavy-Tailed Data". In: *arXiv preprint arXiv:2112.07602* (2021).

[107] Joel A Tropp. "User-friendly tail bounds for sums of random matrices". In: *Foundations of computational mathematics* 12.4 (2012), pp. 389–434.

[108] Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.

[109] J. W. Tukey. "A survey of sampling from contaminated distributions". In: *Contributions to Probability and Statistics* (1960), pp. 448–485.

[110] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge University Press, 2018.

[111] Roman Vershynin. "Introduction to the non-asymptotic analysis of random matrices". In: *arXiv preprint arXiv:1011.3027* (2010).

[112] Oriol Vinyals et al. "Matching networks for one shot learning". In: *Advances in neural information processing systems*. 2016, pp. 3630–3638.

[113] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, Cambridge, 2019. DOI: 10.1017/9781108627771. URL: https://doi.org/10.1017/9781108627771.

[114] Zirui Wang et al. "Characterizing and avoiding negative transfer". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 11293–11302.

[115] Mitchell Wortsman et al. *Robust fine-tuning of zero-shot models*. 2021. arXiv: 2109.01903 [cs.CV].

[116] Jason Yosinski et al. "How transferable are features in deep neural networks?" In: *Advances in neural information processing systems*. 2014, pp. 3320–3328.

[117] Yuchen Zhang, Martin J Wainwright, and Michael I Jordan. "Lower bounds on the performance of polynomial-time algorithms for sparse linear regression". In: *Conference on Learning Theory*. 2014, pp. 921–948.