Provable Meta-Learning of Linear Representations

Nilesh Tripuraneni [1] **Chi Jin** [2] **Michael I. Jordan** [1]

Abstract

Meta-learning, or learning-to-learn, seeks to design algorithms that can utilize previous experience to rapidly learn new skills or adapt to new environments. Representation learning—a key tool for performing meta-learning—learns a data representation that can transfer knowledge across multiple tasks, which is essential in regimes where data is scarce. Despite a recent surge of interest in the practice of meta-learning, the theoretical underpinnings of meta-learning algorithms are lacking, especially in the context of learning transferable representations. In this paper, we focus on the problem of multi-task linear regression—in which multiple linear regression models share a common, low-dimensional linear representation. Here, we provide provably fast, sample-efficient algorithms to address the dual challenges of (1) learning a common set of features from multiple, related tasks, and (2) transferring this knowledge to new, unseen tasks. Both are central to the general problem of meta-learning. Finally, we complement these results by providing information-theoretic lower bounds on the sample complexity of learning these linear features.

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 (Donahue et al., 2014; Vinyals et al., 2016). 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 (Baevski et al., 2019) to natural language processing (Ando & Zhang, 2005; Liu et al., 2019). Accordingly, researchers have begun to highlight the need to develop (and understand) generic algorithms for transfer (or meta) learning applicable in diverse domains (Finn et al., 2017). 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 this paper, 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:

$$y_i = \mathbf{x}_i^\top \mathbf{B} \boldsymbol{\alpha}_{t(i)} + \epsilon_i \quad ; \quad \boldsymbol{\beta}_{t(i)} = \mathbf{B}\boldsymbol{\alpha}_{t(i)}, \tag{1}$$

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 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]

[1]Department of EECS, University of California, Berkeley [2]Department of Electrical Engineering, Princeton University. Correspondence to: Nilesh Tripuraneni <nilesh_tripuraneni@berkeley.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

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

[2]Theorem 1 follows immediately from combining Theorems 3 and 4; see Theorem 6 for a formal statement.

Theorem 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:*

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

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 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 1 is the first bound, even in the *linear* setting, to sharply exhibit this phenomenon (see Section 1.1 for a detailed comparison to existing results). Prior work provides rates for which the leading term in (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 (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 (Jin et al., 2017), can efficiently recover good feature representations (see Section 3.1).
- 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 3.2).
- 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 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). To our knowledge, this is the first information-theoretic lower bound for representation learning in the multi-task setting (see Section 5).

1.1. 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 Baxter (2000), who studied a model where tasks with shared representations are sampled from the same underlying environment. Pontil & Maurer (2013) and Maurer et al. (2016), 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 1 that we are aware of is Maurer et al. (2016, Theorem 5). Instantiated in our setting, Maurer et al. (2016, 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 Maurer et al. (2016), 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 Maurer et al. (2016), 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 guaran-

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

tees for gradient-based meta-learning (MAML) (Finn et al., 2017). Finn et al. (2019); Khodak et al. (2019a;b), and Denevi et al. (2019) work in the framework of online convex 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, Du et al. (2020) 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 bounds use a distributional assumption over meta-test tasks, while our bounds for linear regression are sharp for fixed meta-test tasks. Moreover, their focus is on purely statistical guarantees—they assume access to an ERM oracle for non-convex optimization problems. Our focus is on providing 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. 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 (Edelman et al., 1998). 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 the principal angles as:

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

where $0 \le \theta_k \le \ldots \le \theta_1 \le \frac{\pi}{2}$. 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 $\mathrm{Gr}_{r,d}(\mathbb{R})$ it represents. We now detail several assumptions we use in our analysis.

Assumption 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 $\mathbb{E}[\exp(\mathbf{v}^\top\mathbf{x}_i)] \le \exp\left(\frac{\|\mathbf{v}\|^2}{2}\right)$ 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 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 $\bar{\kappa} = \frac{\mathrm{tr}(\frac{\mathbf{A}^\top\mathbf{A}}{t})}{r\nu}$, and the worst-case condition number as $\kappa = \sigma_1(\frac{\mathbf{A}^\top\mathbf{A}}{t})/\nu$.

Assumption 2 (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 $\Theta(1)$[4]. 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 3), we

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

provide algorithms to learn the underlying linear representation from a set of diverse tasks. Second, in the meta-test phase (see Section 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.

3. Meta-Train: Learning Linear Features

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

3.1. 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 (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 (1) is reminiscent of a matrix sensing/completion problem. We leverage this connection for our analysis, building in particular on results from Ge et al. (2017). 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:

$$y_i = \langle \mathbf{e}_{t(i)} \mathbf{x}_i^\top, \mathbf{M}_\star \rangle + \epsilon_i. \tag{3}$$

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:

$$\min_{\mathbf{U} \in \mathbb{R}^{t \times r}, \mathbf{V} \in \mathbb{R}^{d \times r}} f(\mathbf{U}, \mathbf{V}) = \frac{2t}{n} \sum_{i=1}^{n} (y_i - \langle \mathbf{e}_{t(i)} \mathbf{x}_i^\top, \mathbf{U}\mathbf{V}^\top \rangle)^2 + \frac{1}{2} \|\mathbf{U}^\top \mathbf{U} - \mathbf{V}^\top \mathbf{V}\|_{\mathrm{F}}^2. \tag{4}$$

The second term in (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 (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:

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

for some large constant C_0. Under Assumption 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 16 can be used to show that Assumption 2 actually implies that $\mathbf{U}^\star$ is incoherent, which satisfies the first constraint. Our main result follows.

Theorem 2. *Let Assumptions 1 and 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 (4), the column space of $\mathbf{V}$, spanned by the orthonormal feature matrix $\hat{\mathbf{B}}$, satisfies:*

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

We make several comments on this result:

- The guarantee in Theorem 2 suggests that all local minimizers of the regularized empirical risk (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 5 guarantees that any estimator (including the empirical risk minimizer) must incur error $\gtrsim \sqrt{dr/n_1}$. Therefore, in the regime $t \le O(d)$, 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 (Jin et al., 2017), Theorem 2 provides strong evidence that first-order methods can efficiently meta-learn linear features.[5]

The proof of Theorem 2 is technical so we only sketch the high-level ideas. The overall strategy is to analyze the Hessian of the objective (4) at a stationary point $(\mathbf{U}, \mathbf{V})$ in int$(\mathcal{W})$ to exhibit a direction Δ of negative curvature which

[5] To formally establish computational efficiency, we need to further verify the smoothness and the strict-saddle properties of the objective function (4) (see, e.g., (Jin et al., 2017)).

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 $\hat{\mathbf{B}}$

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 $\hat{\boldsymbol{\alpha}}$

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 $\text{int}(\mathcal{W})$ satisfy $\|\mathbf{U}\mathbf{V}^\top - \mathbf{M}^\star\|_F \le O(\sqrt{t\frac{\max\{t,d\}r\log n_1}{n_1}})$ (see Theorem 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 15) to give the final result.

3.2. 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 (Pearson, 1894) and has recently been used to estimate parameters in latent variable models (Anandkumar et al., 2012).

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 $\bar{\boldsymbol{\Lambda}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\alpha}_{t(i)}\boldsymbol{\alpha}_{t(i)}^\top$ where $\tilde{\nu} = \sigma_r(\bar{\boldsymbol{\Lambda}})$, and $\tilde{\kappa} = \text{tr}(\bar{\boldsymbol{\Lambda}})/(r\tilde{\nu})$ in analogy with Assumption 2.

Theorem 3. *Suppose the n_1 data samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n_1}$ are generated from the model in (1) and that Assumptions 1 and 2 hold, but additionally that $\mathbf{x}_i \sim \mathcal{N}(0, \mathbf{I}_d)$. Then, if $n_1 \gtrsim \text{polylog}(d, n_1)rd\tilde{\kappa}/\tilde{\nu}$, the output $\hat{\mathbf{B}}$ of Algorithm 1 satisfies*

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

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

We first make several remarks regarding this result.

- Theorem 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 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 3 has no explicit dependence on t and is close-to-optimal in the constant-snr regime; see Theorem 5 for our lower bound.

We now provide a summary of the proof. Under oracle access to the population mean $\mathbb{E}[\frac{1}{n}\sum_i y_i^2\mathbf{x}_i\mathbf{x}_i^\top] = (2\bar{\boldsymbol{\Gamma}} + (1+\text{tr}(\bar{\boldsymbol{\Gamma}}))\mathbf{I}_d)$, 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 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 7). If this holds, the Davis-Kahan $\sin\theta$ theorem (Bhatia, 2013) 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.

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 (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:

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \|\mathbf{y} - \mathbf{X}\hat{\mathbf{B}}\boldsymbol{\alpha}\|^2, \tag{5}$$

where n_2 samples $(\mathbf{X}, \mathbf{y})$ are generated from the model in (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 (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 4. *Suppose n_2 data points, $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n_2}$, are generated from the model in (1), where Assumption 1 holds, from a single task satisfying $\|\boldsymbol{\alpha}_{t+1}\|^2 \leq O(1)$. Then, if $\sin\theta(\hat{\mathbf{B}}, \mathbf{B}) \leq \delta$ and $n_2 \gtrsim r \log n_2$, the output $\hat{\boldsymbol{\alpha}}$ from Algorithm 2 satisfies*

$$\|\hat{\mathbf{B}}\hat{\boldsymbol{\alpha}} - \mathbf{B}\boldsymbol{\alpha}_{t+1}\|^2 \leq \tilde{O}\left(\delta^2 + \frac{r}{n_2}\right), \tag{6}$$

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 (1). We make several remarks about this result:

- Theorem 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 (Hsu et al., 2012). 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 (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 (Wang et al., 2019). For diverse tasks (i.e. $\kappa \leq O(1)$), 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 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,

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

The first term contributes the bias term to (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 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 $\mathbf{X}\mathbf{B}$ 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 18).

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 (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, Rohde et al. (2011)), 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 5. *Suppose a total of n data points are generated from the model in (1) satisfying Assumption 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 holds with $\boldsymbol{\alpha}_j$ for each task normalized to $\|\boldsymbol{\alpha}_j\| = \frac{1}{2}$. Then, there are $\boldsymbol{\alpha}_j$ for $r \leq \frac{d}{2}$ and $n \geq \max\left(\frac{1}{8\nu}, r(d-r)\right)$ so that:*

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

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 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)[6] 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 5 establishes that task diversity is necessary for recovery of the subspace $\mathbf{B}$.
- The dimension of $\mathrm{Gr}_{r,d}(\mathbb{R})$ (i.e., the number of free parameters needed to specify a feature set) is $r(d-r) \geq \Omega(dr)$ for $d/2 \geq r$; hence the second term in Theorem 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 3 and Theorem 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 Chen et al. (2016, Section 7), Raskutti et al. (2011, Theorem 1, Theorem 2) and Zhang et al. (2014) 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.

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

The dependency on the task diversity parameter $\frac{1}{\nu}$ (the first term in Theorem 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 Guntuboyina (2011) (restated in Lemma 20 in the Appendix), similar in spirit to the global Fano (or Yang-Barron).

There are two key ingredients to using Lemma 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 Wainwright (2019, 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.[7] Our case is no exception. To circumvent this difficulty we use the framework of Guntuboyina (2011), 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 Pajor (1998) to provide sharp upper/lower bounds on the metric entropy (or global entropy) of the Grassmann manifold (see Proposition 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 $\{\mathbb{P}_{\mathbf{B}} : \mathbf{B} \in \mathrm{Gr}_{r,d}(\mathbb{R})\}$—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 $\chi^2(\mathbb{P}_{\mathbf{B}^1}, \mathbb{P}_{\mathbf{B}^2})$ for two different measures over data generated from the model (1) with two different feature matrices $\mathbf{B}^1$ and $\mathbf{B}^2$ (see Lemma 21). This control can be achieved in our random design setting by exploiting the 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 $\mathbf{B}^2$.

6. Simulations

We complement our theoretical analysis with a series of numerical experiments highlighting the benefits (and lim-

[7] Even for the simple problems of Gaussian mean estimation the classical Yang-Barron method is suboptimal; see Guntuboyina (2011) for more details.

its) of meta-learning[8]. 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 (4). Additional details on our set-up are provided in Appendix G. 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. 1 demonstrates, the method-of-

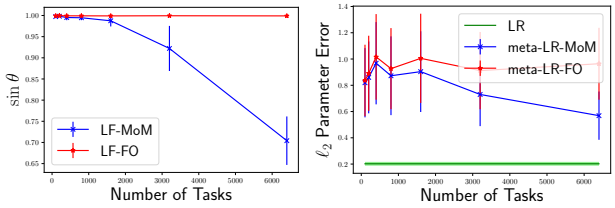

Figure 1. Left: LF-FO vs. LF-MoM estimator with error measured in the subspace angle distance $\sin\theta(\hat{\mathbf{B}}, \mathbf{B})$. Right: meta-LR-FO and meta-LR-MoM vs. LR on new task with error measured on new task parameter. Here $d = 100$, $r = 5$, and $n_t = 5$ while $n_2 = 2500$ as the number of tasks is varied.

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 relative to Theorem 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. 1 also shows that meta-learned regressions perform significantly worse than simply ignoring first t tasks. Theorem 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 (Wang et al., 2019).

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 shows, in

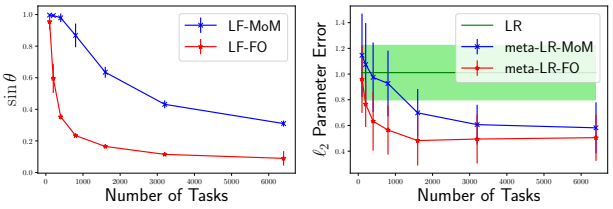

Figure 2. Left: LF-FO vs. LF-MoM estimator with error measured in the subspace angle distance $\sin\theta(\hat{\mathbf{B}}, \mathbf{B})$. Right: meta-LR-FO and meta-LR-MoM vs. LR on new task with error measured on new task parameter. Here $d = 100$, $r = 5$, $n_t = 25$ while $n_2 = 25$ while the number of tasks is varied.

this regime both the method-of-moments estimator and the loss-based approach 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. 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).

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.

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

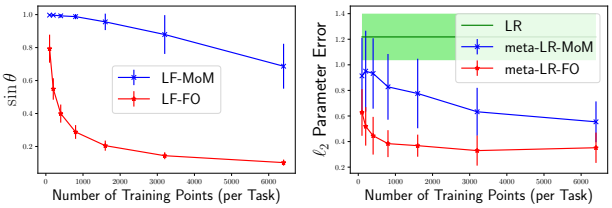

Figure 3. Left: LF-FO vs. LF-MoM estimator with error measured in the subspace angle distance $\sin\theta(\hat{\mathbf{B}}, \mathbf{B})$. Right: meta-LR-FO and meta-LR-MoM vs. LR on new task with error measured on new task parameter. Here $d = 100$, $r = 5$, $t = 20$, and $n_2 = 50$ while the number of training points per task (n_t) is varied.

References

Anandkumar, A., Hsu, D., and Kakade, S. M. A method of moments for mixture models and hidden Markov models. In *Conference on Learning Theory*, pp. 33–1, 2012.

Ando, R. K. and Zhang, T. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6(Nov):1817–1853, 2005.

Baevski, A., Edunov, S., Liu, Y., Zettlemoyer, L., and Auli, M. Cloze-driven pretraining of self-attention networks. *arXiv preprint arXiv:1903.07785*, 2019.

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

Bhatia, R. *Matrix Analysis*, volume 169. Springer Science & Business Media, 2013.

Candes, E. and Plan, Y. Tight oracle bounds for low-rank matrix recovery from a minimal number of noisy random measurements. *arXiv preprint arXiv:1001.0339*, 2010.

Chen, X., Guntuboyina, A., and Zhang, Y. On bayes risk lower bounds. *The Journal of Machine Learning Research*, 17(1):7687–7744, 2016.

Denevi, G., Ciliberto, C., Grazzi, R., and Pontil, M. Learning-to-learn stochastic gradient descent with biased regularization. *arXiv preprint arXiv:1903.10399*, 2019.

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

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

Edelman, A., Arias, T. A., and Smith, S. T. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

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

Finn, C., Rajeswaran, A., Kakade, S., and Levine, S. Online meta-learning. *arXiv preprint arXiv:1902.08438*, 2019.

Ge, R., Jin, C., and Zheng, Y. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. *arXiv preprint arXiv:1704.00708*, 2017.

Guntuboyina, A. Lower bounds for the minimax risk using f-divergences, and applications. *IEEE Transactions on Information Theory*, 57(4):2386–2399, 2011.

Hsu, D., Kakade, S. M., and Zhang, T. Random design analysis of ridge regression. In *Conference on learning theory*, pp. 9–1, 2012.

Jin, C., Ge, R., Netrapalli, P., Kakade, S. M., and Jordan, M. I. How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1724–1732. JMLR. org, 2017.

Khodak, M., Balcan, M.-F., and Talwalkar, A. Provable guarantees for gradient-based meta-learning. *arXiv preprint arXiv:1902.10644*, 2019a.

Khodak, M., Balcan, M.-F. F., and Talwalkar, A. S. Adaptive gradient-based meta-learning methods. In *Advances in Neural Information Processing Systems*, pp. 5915–5926, 2019b.

Liu, D. C. and Nocedal, J. On the limited memory BFGS method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

Liu, X., He, P., Chen, W., and Gao, J. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*, 2019.

Maclaurin, D., Duvenaud, D., and Adams, R. P. Autograd: Effortless gradients in numpy. In *ICML 2015 AutoML Workshop*, volume 238, 2015.

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

Moritz, P., Nishihara, R., Wang, S., Tumanov, A., Liaw, R., Liang, E., Elibol, M., Yang, Z., Paul, W., Jordan, M. I., et al. Ray: A distributed framework for emerging {AI} applications. In *13th {USENIX} Symposium on*

Operating Systems Design and Implementation ({OSDI} 18), pp. 561–577, 2018.

Pajor, A. Metric entropy of the Grassmann manifold. *Convex Geometric Analysis*, 34:181–188, 1998.

Pearson, K. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

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

Raskutti, G., Wainwright, M. J., and Yu, B. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE transactions on information theory*, 57 (10):6976–6994, 2011.

Recht, B. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(Dec):3413–3430, 2011.

Rohde, A., Tsybakov, A. B., et al. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930, 2011.

Tropp, J. A. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12 (4):389–434, 2012.

Tsybakov, A. B. *Introduction to Nonparametric Estimation*. Springer Science & Business Media, 2008.

Vershynin, R. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge University Press, 2018.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pp. 3630–3638, 2016.

Wainwright, M. J. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019.

Wang, Z., Dai, Z., Póczos, B., and Carbonell, J. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11293–11302, 2019.

Zhang, Y., Wainwright, M. J., and Jordan, M. I. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *Conference on Learning Theory*, pp. 921–948, 2014.