Provable Pathways: Learning Multiple Tasks over Multiple Paths

Yingcong Li,[1] **Samet Oymak**[1,2]

[1] University of California, Riverside,
[2] University of Michigan, Ann Arbor
{yli692@, oymak@ece.}ucr.edu

Abstract

Constructing useful representations across a large number of tasks is a key requirement for sample-efficient intelligent systems. A traditional idea in multitask learning (MTL) is building a shared representation across tasks which can then be adapted to new tasks by tuning last layers. A desirable refinement of using a shared one-fits-all representation is to construct task-specific representations. To this end, recent PathNet/muNet architectures represent individual tasks as pathways within a larger supernet. The subnetworks induced by pathways can be viewed as task-specific representations that are composition of modules within supernet's computation graph. This work explores the pathways proposal from the lens of statistical learning: We first develop novel generalization bounds for empirical risk minimization problems learning multiple tasks over multiple paths (Multipath MTL). In conjunction, we formalize the benefits of resulting multipath representation when adapting to new downstream tasks. Our bounds are expressed in terms of Gaussian complexity, lead to tangible guarantees for the class of linear representations, and provide novel insights into the quality and benefits of a multipath representation. When computation graph is a tree, Multipath MTL hierarchically clusters the tasks and builds cluster-specific representations. We provide further discussion and experiments for hierarchical MTL and rigorously identify the conditions under which Multipath MTL is provably superior to traditional MTL approaches with shallow supernets.

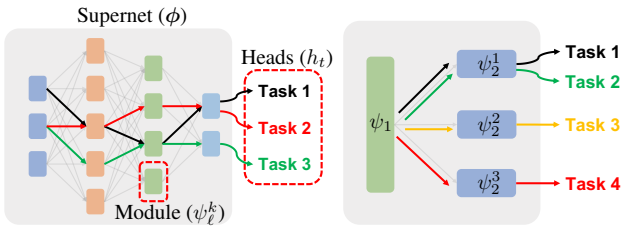
1 Introduction

Multitask learning (MTL) promises to deliver significant accuracy improvements by leveraging similarities across many tasks through shared representations. The potential of MTL has been recognized since 1990s (Caruana 1997) however its impact has grown over time thanks to more recent machine learning applications arising in computer vision and NLP that involve large datasets with thousands of classes/tasks. Representation learning techniques (e.g. MTL and self-supervision) are also central to the success of deep learning as large pretrained models enable data-efficient learning for downstream transfer learning tasks (Deng et al. 2009; Brown et al. 2020).

As we move from tens of tasks trained with small models to thousands of tasks trained with large models, new statistical and computational challenges arise: First, not all tasks will be closely related to each other, for instance, tasks might admit a natural clustering into groups. This is also connected to heterogeneity challenge in federated learning where clients have distinct distributions and benefit from personalization. To address this challenge, rather than a single task-agnostic representation, it might be preferable to use a task-specific representation. Secondly, pretrained language and vision models achieve better accuracy with larger sizes which creates computational challenges as they push towards trillion parameters. This motivated new architectural proposals such as Pathways/PathNet (Fernando et al. 2017; Dean 2021; Gesmundo and Dean 2022b) where tasks can be computed over compute-efficient subnetworks. At a high-level, each subnetwork is created by a composition of modules within a larger supernet which induces a pathway as depicted in Figure 1. Inspired from these challenges, we ask

Q: What are the statistical benefits of learning task-specific representations along supernet pathways?

Our primary contribution is formalizing the Multipath MTL problem depicted in Figure 1 and developing associated statistical learning guarantees that shed light on its benefits. Our formulation captures important aspects of the problem including learning compositional MTL representations, multilayer



(a) General computation graph (b) Hierarchical MTL

Figure 1: In Multipath MTL, each task selects a pathway within a supernet graph. The composition of the modules along the pathway forms the task-specific representation. Fig. 1a depicts a general supernet graph (highlighted in gray block), and the pathways for different tasks are shown in colored arrows. Fig. 1b is a special instance where related tasks are hierarchically clustered: For instance, Tasks 1 and 2 are assinged the same representation $\psi_2^1 \circ \psi_1$.

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

nature of supernet, assigning optimal pathways to individual tasks, and transferring learned representations to novel downstream tasks. Our specific contributions are as follows.

• Suppose we have N samples per task and T tasks in total. Denote the hypothesis sets for multipath representation by Φ, task specific heads by $\mathcal{H}$ and potential pathway choices by $\mathcal{A}$. Our main result bounds the task-averaged risk of MTL as

$$\sqrt{\frac{\text{DoF}(\Phi_{\text{used}})}{NT}} + \sqrt{\frac{\text{DoF}(\mathcal{H}) + \text{DoF}(\mathcal{A})}{N}}. \tag{1}$$

Here, $\text{DoF}(\cdot)$ returns the *degrees of freedom* of a hypothesis set (i.e. number of parameters). More generally, Theorem 1 states our guarantees in terms of Gaussian complexity. $\Phi_{\text{used}} \subseteq \Phi$ is the supernet spanned by the pathways of the empirical solution and $1/NT$ dependence implies that cost of representation learning is shared across tasks. We also show a *no-harm* result (Lemma 1): If the supernet is sufficiently expressive to achieve zero empirical risk, then, the excess risk of individual tasks will not be harmed by the other tasks. Theorem 2 develops guarantees for transferring the resulting MTL representation to a new task in terms of representation bias of the empirical MTL supernet.

• When the supernet has a single module, the problem boils down to (vanilla) MTL with single shared representation and our bounds recover the results by (Maurer, Pontil, and Romera-Paredes 2016; Tripuraneni, Jin, and Jordan 2021). When the supernet graph is hierarchical (as in Figure 1b), our bounds provide insights for the benefits of clustering tasks into similar groups and superiority of multilayer Multipath MTL over using single-layer shallow supernets (Section 5).

• We develop stronger results for linear representations over a supernet and obtain novel MTL and transfer learning bounds (Sec. 4 and Theorem 4). These are accomplished by developing new task-diversity criteria to account for the task-specific (thus heterogeneous) nature of multipath representations. Numerical experiments support our theory and verify the benefits of multipath representations. Finally, we also highlight multiple future directions.

2 Setup and Problem Formulations

Notation. Let $\|\cdot\|$ denote the ℓ_2-norm of a vector and operator norm of a matrix. $|\cdot|$ denotes the absolute value for scalars and cardinality for discrete sets. We use $[K]$ to denote the set $\{1, 2, \ldots, K\}$ and $\lesssim, \gtrsim$ for inequalities that hold up to constant/logarithmic factors. $\mathcal{Q}^K$ denotes K-times Cartesian product of a set $\mathcal{Q}$ with itself. $\circ$ denotes functional composition, i.e., $f \circ g(x) = f(g(x))$.

Setup. Suppose we have T tasks each following data distribution $\{\mathcal{D}_t\}_{t=1}^T$. During MTL phase, we are given T training datasets $\{\mathcal{S}_t\}_{t=1}^T$ each drawn i.i.d. from its corresponding distribution $\mathcal{D}_t$. Let $\mathcal{S}_t = \{(\boldsymbol{x}_{ti}, y_{ti})\}_{i=1}^N$, where $(\boldsymbol{x}_{ti}, y_{ti}) \in (\mathcal{X}, \mathbb{R})$ is an input-label pair and $\mathcal{X}$ is the input space, and $|\mathcal{S}_t| = N$ is the number of samples per task. We assume the same N for all tasks for cleaner exposition. Define the union of the datasets by $\mathcal{S}_{\text{all}} = \bigcup_{t=1}^T \mathcal{S}_t$ (with $|\mathcal{S}_{\text{all}}| = NT$), and the set of distributions by $\bar{\mathcal{D}} = \{\mathcal{D}_t\}_{t=1}^T$.

Following the setting of related works (Tripuraneni, Jin, and Jordan 2021), we will consider two problems: **(1) MTL problem** will use these T datasets to learn a supernet and establish guarantees for representation learning. **(2) Transfer learning problem** will use the resulting representation for a downstream task in a sample efficient fashion.

Problem (1): Multipath Multitask Learning (M^{2}TL). We consider a supernet with L layers where layer ℓ has K_ℓ modules for $\ell \in [L]$. As depicted in Figure 1, each task will compose a task-specific representation by choosing one module from each layer. We refer to each sequence of L modules as a *pathway*. Let $\mathcal{A} = [K_1] \times \cdots \times [K_L]$ be the set of all pathway choices obeying $|\mathcal{A}| = \prod_{\ell=1}^L K_\ell$. Let $\alpha_t \in \mathcal{A}$ denote the pathway associated with task $t \in [T]$ where $\alpha_t[\ell] \in [K_\ell]$ denotes the selected module index from layer ℓ. We remark that results can be extended to more general pathway sets as discussed in Section 3.1.

As depicted in Figure 1, let Ψ_ℓ be the hypothesis set of modules in ℓ_{th} layer and $\psi_\ell^k \in \Psi_\ell$ denote the k_{th} module function in the ℓ_{th} layer, referred to as (ℓ, k)'th module. Let $h_t \in \mathcal{H}$ be the prediction head of task t where all tasks use the same hypothesis set $\mathcal{H}$ for prediction. Let us denote the combined hypothesis

$$\begin{aligned}
&\boldsymbol{h} = [h_1, \ldots, h_T] \in \mathcal{H}^T,\\
&\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_T] \in \mathcal{A}^T,\\
&\boldsymbol{\psi}_\ell = [\psi_\ell^1, \ldots, \psi_\ell^{K_\ell}] \in \Psi_\ell^{K_\ell},\ \forall \ell \in [L],\\
&\boldsymbol{\phi} := [\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_L] \in \Phi
\end{aligned}$$

where $\Phi = \Psi_1^{K_1} \times \cdots \times \Psi_L^{K_L}$ is the supernet hypothesis class containing all modules/layers. Given a supernet $\boldsymbol{\phi} \in \Phi$ and pathway α, $\boldsymbol{\phi}_\alpha = \psi_L^\alpha \circ \cdots \circ \psi_1^\alpha$ denotes the representation induced by pathway α where we use the convention $\psi_\ell^\alpha := \psi_\ell^{\alpha[\ell]}$. Hence, $\boldsymbol{\phi}_{\alpha_t}$ is the representation of task t. We would like to solve for supernet weights $\boldsymbol{\phi}$, pathways $\boldsymbol{\alpha}$, and heads $\boldsymbol{h}$. Thus, given a loss function $\ell(\hat{y}, y)$, Multipath MTL (M^{2}TL) solves the following empirical risk minimization problem over $\mathcal{S}_{\text{all}}$ to optimize the combined hypothesis $\boldsymbol{f} = (\boldsymbol{h}, \boldsymbol{\alpha}, \boldsymbol{\phi})$:

$$\hat{\boldsymbol{f}} = \operatorname*{arg\,min}_{\boldsymbol{f} \in \mathcal{F}} \widehat{\mathcal{L}}_{\mathcal{S}_{\text{all}}}(\boldsymbol{f}) := \frac{1}{T} \sum_{t=1}^T \widehat{\mathcal{L}}_t(h_t \circ \boldsymbol{\phi}_{\alpha_t}) \quad \text{(M}^2\text{TL)}$$

$$\text{where} \quad \widehat{\mathcal{L}}_t(f) = \frac{1}{N} \sum_{i=1}^N \ell(f(\boldsymbol{x}_{ti}), y_{ti})$$

$$\mathcal{F} := \mathcal{H}^T \times \mathcal{A}^T \times \Phi.$$

Here $\widehat{\mathcal{L}}_t$ and $\widehat{\mathcal{L}}_{\mathcal{S}_{\text{all}}}$ are task-conditional and task-averaged empirical risks. We are primarily interested in controlling the task-averaged test risk $\mathcal{L}_{\bar{\mathcal{D}}}(\boldsymbol{f}) = \mathbb{E}[\widehat{\mathcal{L}}_{\mathcal{S}_{\text{all}}}(\boldsymbol{f})]$. Let $\mathcal{L}_{\bar{\mathcal{D}}}^\star := \min_{\boldsymbol{f} \in \mathcal{F}} \mathcal{L}_{\bar{\mathcal{D}}}(\boldsymbol{f})$, then the *excess MTL risk* is defined as

$$\mathcal{R}_{\text{M}^2\text{TL}}(\hat{\boldsymbol{f}}) = \mathcal{L}_{\bar{\mathcal{D}}}(\hat{\boldsymbol{f}}) - \mathcal{L}_{\bar{\mathcal{D}}}^\star. \tag{2}$$

Problem (2): Transfer Learning with Optimal Pathway (TLOP). Suppose we have a novel target task with i.i.d. training dataset $\mathcal{S}_{\mathcal{T}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^M$ with M samples drawn from

distribution $\mathcal{D}_\mathcal{T}$. Given a pretrained supernet $\boldsymbol{\phi}$ (e.g., following (M^{2}TL)), we can search for a pathway α so that $\boldsymbol{\phi}_\alpha$ becomes a suitable representation for $\mathcal{D}_\mathcal{T}$. Thus, for this new task, we only need to optimize the path $\alpha \in \mathcal{A}$ and the prediction head $h \in \mathcal{H}_\mathcal{T}$ while reusing weights of $\boldsymbol{\phi}$. This leads to the following problem:

$$\hat{f}_{\boldsymbol{\phi}} = \underset{h \in \mathcal{H}_\mathcal{T}, \alpha \in \mathcal{A}}{\arg\min}\ \widehat{\mathcal{L}}_\mathcal{T}(f) \quad \text{where} \quad f = h \circ \boldsymbol{\phi}_\alpha \qquad \text{(TLOP)}$$

$$\text{and} \quad \widehat{\mathcal{L}}_\mathcal{T}(f) = \frac{1}{M}\sum_{i=1}^{M} \ell(f(\boldsymbol{x}_i), y_i).$$

Here, $\hat{f}_{\boldsymbol{\phi}}$ reflects the fact that solution depends on the suitability of pretrained supernet $\boldsymbol{\phi}$. Let $f^\star_{\boldsymbol{\phi}}$ be a population minima of (TLOP) given supernet $\boldsymbol{\phi}$ (as $M \to \infty$) and define the population risk $\mathcal{L}_\mathcal{T}(f) = \mathbb{E}[\widehat{\mathcal{L}}_\mathcal{T}(f)]$. (TLOP) will be evaluated against the hindsight knowledge of optimal supernet for target: Define the optimal target risk $\mathcal{L}^\star_\mathcal{T} := \min_{h \in \mathcal{H}_\mathcal{T}, \boldsymbol{\phi} \in \Phi} \mathcal{L}_\mathcal{T}(h \circ \boldsymbol{\phi}_\alpha)$ which optimizes $h, \boldsymbol{\phi}$ for the target task along the fixed pathway $\alpha = [1, \ldots, 1]$. Here we can fix α since all pathways result in the same search space. We define the *excess transfer learning risk* to be

$$\begin{aligned}\mathcal{R}_{\text{TLOP}}(\hat{f}_{\boldsymbol{\phi}}) &= \mathcal{L}_\mathcal{T}(\hat{f}_{\boldsymbol{\phi}}) - \mathcal{L}^\star_\mathcal{T} \qquad (3)\\ &= \underbrace{\mathcal{L}_\mathcal{T}(\hat{f}_{\boldsymbol{\phi}}) - \mathcal{L}_\mathcal{T}(f^\star_{\boldsymbol{\phi}})}_{\text{variance}} + \underbrace{\mathcal{L}_\mathcal{T}(f^\star_{\boldsymbol{\phi}}) - \mathcal{L}^\star_\mathcal{T}}_{\text{supernet bias}}.\end{aligned}$$

The final line decomposes the overall risk into a *variance* term and *supernet bias* . The former arises from the fact that we solve the problem with finite training samples. This term will vanish as $M \to \infty$. The latter term quantifies the bias induced by the fact that (TLOP) uses the representation $\boldsymbol{\phi}$ rather than the optimal representation. Finally, while supernet $\boldsymbol{\phi}$ in (TLOP) is arbitrary, for end-to-end guarantees we will set it to the solution $\hat{\boldsymbol{\phi}}$ of (M^{2}TL). In this scenario, we will refer to $\{\mathcal{D}_t\}_{t=1}^T$ as source tasks.

3 Main Results

We are ready to present our results that establish generalization guarantees for multitask and transfer learning problems over supernet pathways. Our results will be stated in terms of Gaussian complexity which is introduced below.

Definition 1 (Gaussian Complexity) *Let $\mathcal{Q}$ be a set of hypotheses that map $\mathcal{Z}$ to $\mathbb{R}^r$. Let $(\boldsymbol{g}_i)_{i=1}^n$ ($\boldsymbol{g}_i \in \mathbb{R}^r$) be n independent vectors each distributed as $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_r)$ and let $\boldsymbol{Z} = (\boldsymbol{z}_i)_{i=1}^n \in \mathcal{Z}^n$ be a dataset of input features. Then, the empirical Gaussian complexity is defined as*

$$\widehat{\mathcal{G}}_{\boldsymbol{Z}}(\mathcal{Q}) = \mathbb{E}_{\boldsymbol{g}_i}\left[\sup_{q \in \mathcal{Q}} \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}_i^\top q(\boldsymbol{z}_i)\right].$$

The worst-case Gaussian complexity is obtained by considering the supremum over $\boldsymbol{Z} \in \mathcal{Z}^n$ as follows

$$\widetilde{\mathcal{G}}^{\mathcal{Z}}_n(\mathcal{Q}) = \sup_{\boldsymbol{Z} \in \mathcal{Z}^n} [\widehat{\mathcal{G}}_{\boldsymbol{Z}}(\mathcal{Q})].$$

For cleaner notation, we drop the superscript $\mathcal{Z}$ from the worst-case Gaussian complexity (using $\widetilde{\mathcal{G}}_n(\mathcal{Q})$) as its input space will be clear from context. When $\boldsymbol{Z} = (\boldsymbol{z}_i)_{i=1}^n$ are drawn i.i.d. from $\mathcal{D}$, the (usual) Gaussian complexity is defined by $\mathcal{G}_n(\mathcal{Q}) = \mathbb{E}_{\boldsymbol{Z} \sim \mathcal{D}^n}[\widehat{\mathcal{G}}_{\boldsymbol{Z}}(\mathcal{Q})]$. Note that, we always have $\mathcal{G}_n(\mathcal{Q}) \le \widetilde{\mathcal{G}}_n(\mathcal{Q})$ assuming $\mathcal{D}$ is supported on $\mathcal{Z}$. In our setting, keeping track of distributions along exponentially many pathways proves challenging, and we opt to use $\widetilde{\mathcal{G}}_n(\mathcal{Q})$ which leads to clean upper bounds. The supplementary material[1] also derives tighter but more convoluted bounds in terms of empirical complexity. Finally, it is well-known that Gaussian/Rademacher complexities scale as $\sqrt{\text{comp}(\mathcal{Q})/n}$ where $\text{comp}(\mathcal{Q})$ is a set complexity such as VC-dimension, which links to our informal statement (1).

We will first present our generalization bounds for the Multipath MTL problem using empirical process theory arguments. Our bounds will lead to meaningful guarantees for specific MTL settings, including vanilla MTL where all tasks share a single representation, as well as hierarchical MTL depicted in Fig. 1b. We will next derive transfer learning guarantees in terms of supernet bias, which quantifies the performance difference of a supernet from its optimum for a target. To state our results, we introduce two standard assumptions.

Assumption 1 *Elements of hypothesis sets $\mathcal{H}$ and $(\Psi_\ell)_{\ell=1}^L$ are Γ-Lipschitz functions with respect to Euclidean norm.*

Assumption 2 *Loss function $\ell(\cdot, y) : \mathbb{R} \times \mathbb{R} \to [0, 1]$ is Γ-Lipschitz with respect to Euclidean norm.*

3.1 Results for Multipath Multitask Learning

This section presents our task-averaged generalization bound for Multipath MTL problem. Recall that $\hat{\boldsymbol{f}} = (\hat{\boldsymbol{h}}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\phi}})$ is the outcome of the ERM problem (M^{2}TL). Observe that, if we were solving the problem with only one task, the generalization bound would depend on only one module per layer rather than the overall size of the supernet. This is because each task gets to select a single module through their pathway. In light of this, we can quantify the utilization of supernet layers as follows: Let $\hat{K}_\ell$ be the number of modules utilized by the empirical solution $\hat{\boldsymbol{f}}$. Formally, $\hat{K}_\ell = |\{\hat{\alpha}_t[\ell] \ \text{ for } \ t \in [T]\}|$. The following theorem provides our guarantee in terms of Gaussian complexities of individual modules.

Theorem 1 *Suppose Assumptions 1&2 hold. Let $\hat{\boldsymbol{f}}$ be the empirical solution of* (M^{2}TL)*. Then, with probability at least $1 - \delta$, the excess test risk in* (2) *obeys $\mathcal{R}_{M^2TL}(\hat{\boldsymbol{f}})$*

$$\lesssim \widetilde{\mathcal{G}}_N(\mathcal{H}) + \sum_{\ell=1}^{L} \sqrt{\hat{K}_\ell}\widetilde{\mathcal{G}}_{NT}(\Psi_\ell) + \sqrt{\frac{\log|\mathcal{A}|}{N} + \frac{\log(2/\delta)}{NT}}.$$

Here, the input spaces for $\mathcal{H}$ and Ψ_ℓ are $\mathcal{X}_\mathcal{H} = \Psi_L \circ \ldots \Psi_1 \circ \mathcal{X}$, $\mathcal{X}_{\Psi_\ell} = \Psi_{\ell-1} \circ \ldots \Psi_1 \circ \mathcal{X}$ for $\ell > 1$, and $\mathcal{X}_{\Psi_1} = \mathcal{X}$.

In Theorem 1, $\sqrt{\frac{\log|\mathcal{A}|}{N}}$ quantifies the cost of learning the pathway and $\widetilde{\mathcal{G}}_N(\mathcal{H})$ quantifies the cost of learning the

[1] All proofs and additional details are provided in the extended technical report (Li and Oymak 2023)

prediction head for each task $t \in [T]$. $\log|\mathcal{A}|$ dependence is standard for the discrete search space $|\mathcal{A}|$. The $\widetilde{\mathcal{G}}_{NT}(\Psi_\ell)$ terms are more interesting and reflect the benefits of MTL. The reason is that, these modules are essentially learned with NT samples rather than N samples, thus cost of representation learning is shared across tasks. The $\sqrt{\hat{K}_\ell}$ multiplier highlights the fact that, we only need to worry about the used modules rather than all possible K_ℓ modules we could have used. In essence, $\sum_{\ell=1}^{L}\sqrt{\hat{K}_\ell}\widetilde{\mathcal{G}}_{NT}(\Psi_\ell)$ summarizes the Gaussian complexity of $\widetilde{\mathcal{G}}(\Phi_{\text{used}})$ where Φ_{used} is the subnetwork of the supernet utilized by the ERM solution $\hat{\boldsymbol{f}}$. By definition $\widetilde{\mathcal{G}}(\Phi_{\text{used}}) \le \widetilde{\mathcal{G}}(\Phi)$. With all these in mind, Theorem 1 formalizes our earlier statement (1).

A key challenge we address in Theorem 1 is decomposing the complexity of the combined hypothesis class $\mathcal{F}$ in (M^2TL) into its building blocks $\mathcal{A}, \mathcal{H}, (\Psi_\ell)_{\ell=1}^{L}$. This is accomplished by developing Gaussian complexity chain rules inspired from the influential work of (Tripuraneni, Jordan, and Jin 2020; Maurer 2016). While this work focuses on two layer composition (prediction heads composed with a shared representation), we develop bounds to control arbitrarily long compositions of hypotheses. Accomplishing this in our multipath setting presents additional technical challenges because each task gets to choose a unique pathway. Thus, tasks don't have to contribute to the learning process of each module unlike the vanilla MTL with shared representation. Consequently, ERM solution is highly heterogeneous and some modules and tasks will be learned better than the others. Worst-case Gaussian complexity plays an important role to establish clean upper bounds in the face of this heterogeneity. In fact, in supplementary material, we provide tighter bounds in terms of empirical Gaussian complexity $\widehat{\mathcal{G}}$, however, they necessitate more convoluted definitions that involve the number of tasks that choose a particular module.

Finally, we note that our bound has a natural interpretation for parametric classes whose $\log(\varepsilon$-covering number) (i.e. metric entropy) grows with degrees of freedom as $\text{DoF}\cdot\log(1/\varepsilon)$. Then, Theorem 1 implies a risk bound proportional to $\sqrt{\frac{T\cdot(\text{DoF}(\mathcal{H})+\log|\mathcal{A}|)+\sum_{\ell=1}^{L}\hat{K}_\ell\cdot\text{DoF}(\Psi_\ell)}{NT}}$. For a neural net implementation, this means small risk as soon as total sample size NT exceeds total number of weights.

We have a few more remarks in place, discussed below.

• **Dependencies.** In Theorem 1, $\lesssim$ suppresses dependencies on $\log(NT)$ and Γ^L. The latter term arises from the exponentially growing Lipschitz constant as we compose more/deeper modules, however, it can be treated as a constant for fixed depth L. We note that such exponential depth dependence is frequent in existing generalization guarantees in deep learning literature (Golowich, Rakhlin, and Shamir 2018; Bartlett, Foster, and Telgarsky 2017; Neyshabur et al. 2018, 2017). In supplementary material, we prove that the exponential dependence can be replaced with a much stronger bound of $\sqrt{L}$ by assuming parameterized hypothesis classes.

• **Implications for Vanilla MTL.** Observe that Vanilla MTL with single shared representation corresponds to the setting $L = 1$ and $K_1 = 1$. Also supernet is simply $\Phi = \Psi_1$ and $\log|\mathcal{A}| = 0$. Applying Theorem 1 to this setting with T tasks each with N samples, we obtain an excess risk upper bound of $\widetilde{\mathcal{O}}\left(\widetilde{\mathcal{G}}_{NT}(\Phi) + \widetilde{\mathcal{G}}_{N}(\mathcal{H})\right)$, where representation Φ is trained with NT samples with input space $\mathcal{X}$, and task-specific heads $h_t \in \mathcal{H}$ are trained with N samples with input space $\Phi \circ \mathcal{X}$. This bound recovers earlier guarantees by (Maurer, Pontil, and Romera-Paredes 2016; Tripuraneni, Jordan, and Jin 2020).

• **Unselected modules do not hurt performance.** A useful feature of our bound is its dependence on Φ_{used} (spanned by empirical pathways) rather than full hypothesis class Φ. This feature arises from a uniform concentration argument where we uniformly control the excess MTL risk over all potential Φ_{used} choices. This uniform control ensures $\widetilde{\mathcal{G}}_{NT}(\Phi_{\text{used}})$ cost for the actual solution $\hat{\boldsymbol{f}}$ and it only comes at the cost of an additional $\sqrt{\frac{\log|\mathcal{A}|}{N}}$ term which is free (up to constant)!

• **Continuous pathways.** This work focuses on relatively simple pathways where tasks choose one module from each layer. The results can be extended to other choices of pathway sets $\mathcal{A}$. First, note that, as long as $\mathcal{A}$ is a discrete set, we will naturally end up with the excess risk dependence of $\sqrt{\frac{\log|\mathcal{A}|}{N}}$. However, one can also consider continuous α, for instance, due to relaxation of the discrete set with a simplex constraint. Such approaches are common in differentiable architecture search methods (Liu, Simonyan, and Yang 2019). In this case, each entry $\alpha[\ell]$ can be treated as a K_ℓ dimensional vector that chooses a continuous superposition of ℓ'th layer modules. Thus, the overall $\alpha \in \mathcal{A}$ parameter would have $\text{comp}(\mathcal{A}) = \sum_{\ell=1}^{L} K_\ell$ resulting in an excess risk term of $\sqrt{\sum_{\ell=1}^{L} K_\ell/N}$. Note that, these are high-level insights based on classical generalization arguments. In practice, performance can be much better than these uniform concentration based upper bounds.

• **No harm under overparameteration.** A drawback of Theorem 1 is that, it is an average-risk guarantee over T tasks. In practice, it is possible that some tasks are hurt during MTL because they are isolated or dissimilar to others (see supplementary for examples). Below, we show that, if the supernet achieves zero empirical risk, then, no task will be worse than the scenario where they are individually trained with N samples, i.e. Multipath MTL does not hurt any task.

Lemma 1 *Recall $\hat{\boldsymbol{f}}$ is the solution of* (M^2TL) *and $\hat{f}_t = \hat{h}_t \circ \hat{\boldsymbol{\phi}}_{\hat{\alpha}_t}$ is the associated task-t hypothesis. Define the excess risk of task t as $\mathcal{R}_t(\hat{f}_t) = \mathcal{L}_t(\hat{f}_t) - \mathcal{L}_t^\star$ where $\mathcal{L}_t(f) = \mathbb{E}_{\mathcal{D}_t}[\widehat{\mathcal{L}}_t(f)]$ is the population risk of task t and $\mathcal{L}_t^\star$ is the optimal achievable test risk for task t over $\mathcal{F}$. With probability at least $1 - \delta - \mathbb{P}(\widehat{\mathcal{L}}_{\mathcal{S}_{all}}(\hat{\boldsymbol{f}}) \neq 0)$, for all tasks $t \in [T]$,*

$$\mathcal{R}_t(\hat{f}_t) \lesssim \widetilde{\mathcal{G}}_N(\mathcal{H}) + \sum_{\ell=1}^{L}\widetilde{\mathcal{G}}_N(\Psi_\ell) + \sqrt{\frac{\log(2T/\delta)}{N}}.$$

Here, $\mathbb{P}(\widehat{\mathcal{L}}_{\mathcal{S}_{\text{all}}}(\hat{\boldsymbol{f}}) = 0)$ is the event of interpolation (zero empirical risk) under which the guarantee holds. We call this

no harm because the bound is same as what one would get by applying union bound over T empirical risk minimizations where each task is optimized individually.

3.2 Transfer Learning with Optimal Pathway

Following Multipath MTL problem, in this section, we discuss guarantees for transfer learning on a supernet. Recall that $\mathcal{A}$ is the set of pathways and our goal in (TLOP) is finding the optimal pathway $\alpha \in \mathcal{A}$ and prediction head $h \in \mathcal{H}_{\mathcal{T}}$ to achieve small target risk. In order to quantify the bias arising from the Multipath MTL phase, we introduce the following definition.

Definition 2 (Supernet Bias) *Recall the definitions $\mathcal{D}_{\mathcal{T}}$, $\mathcal{H}_{\mathcal{T}}$, and $\mathcal{L}^{\star}_{\mathcal{T}}$ stated in Section 2. Given a supernet $\boldsymbol{\phi}$, we define the supernet/representation bias of $\boldsymbol{\phi}$ for a target $\mathcal{T}$ as*

$$\textit{Bias}_{\mathcal{T}}(\boldsymbol{\phi}) = \min_{h \in \mathcal{H}_{\mathcal{T}}, \alpha \in \mathcal{A}} \mathcal{L}_{\mathcal{T}}(h \circ \boldsymbol{\phi}_{\alpha}) - \mathcal{L}^{\star}_{\mathcal{T}}.$$

Definition 2 is a restatement of the supernet bias term in (3). Importantly, it ensures that the optimal pathway-representation over $\boldsymbol{\phi}$ can not be worse than the optimal performance by $\text{Bias}_{\mathcal{T}}(\boldsymbol{\phi})$. Following this, we can state a generalization guarantee for transfer learning problem (TLOP).

Theorem 2 *Suppose Assumptions 1&2 hold. Let supernet $\hat{\boldsymbol{\phi}}$ be the solution of* (M^2TL) *and $\hat{f}_{\hat{\boldsymbol{\phi}}}$ be the empirical minima of* (TLOP) *with respect to supernet $\hat{\boldsymbol{\phi}}$. Then with probability at least $1-\delta$,*

$$\mathcal{R}_{\textit{TLOP}}(\hat{f}_{\hat{\boldsymbol{\phi}}}) \lesssim \textit{Bias}_{\mathcal{T}}(\hat{\boldsymbol{\phi}}) + \sqrt{\frac{\log(2|\mathcal{A}|/\delta)}{M}} + \widetilde{\mathcal{G}}_M(\mathcal{H}_{\mathcal{T}}),$$

where input space of $\widetilde{\mathcal{G}}_M(\mathcal{H}_{\mathcal{T}})$ is given by $\{\hat{\boldsymbol{\phi}}_{\alpha} \circ \mathcal{X} \mid \alpha \in \mathcal{A}\}$.

Theorem 2 highlights the sample efficiency of transfer learning with optimal pathway. While the derivation is straightforward relative to Theorem 1, the key consideration is the supernet bias $\text{Bias}_{\mathcal{T}}(\hat{\boldsymbol{\phi}})$. This term captures the excess risk in (TLOP) introduced by using $\hat{\boldsymbol{\phi}}$. Let $\boldsymbol{\phi}^{\star}$ be the population minima of (M^2TL). Then we can define the *supernet distance* of $\hat{\boldsymbol{\phi}}$ and $\boldsymbol{\phi}^{\star}$ by $d_{\mathcal{T}}(\hat{\boldsymbol{\phi}}; \boldsymbol{\phi}^{\star}) = \text{Bias}_{\mathcal{T}}(\hat{\boldsymbol{\phi}}) - \text{Bias}_{\mathcal{T}}(\boldsymbol{\phi}^{\star})$. The distance measures how well the finite sample solution $\hat{\boldsymbol{\phi}}$ from (M^2TL) performs compared to the optimal MTL solution $\boldsymbol{\phi}^{\star}$. A plausible assumption is so-called *task diversity* proposed by Chen et al. (2021); Tripuraneni, Jordan, and Jin (2020); Xu and Tewari (2021). Here, the idea (or assumption) is that, if a target task is similar to the source tasks, the distance term for target can be controlled in terms of the excess MTL risk $\mathcal{R}_{\text{M}^2\text{TL}}(\hat{\boldsymbol{f}})$ (e.g. by assuming $d_{\mathcal{T}}(\hat{\boldsymbol{\phi}}; \boldsymbol{\phi}^{\star}) \lesssim \mathcal{R}_{\text{M}^2\text{TL}}(\hat{\boldsymbol{f}}) + \varepsilon$). Plugging in this assumption would lead to end-to-end transfer guarantees by integrating Theorems 1 and 2, and we extend the formal analysis to appendix. However, as discussed in Theorem 4, in multipath setting, the problem is a lot more intricate because source tasks can choose totally different task-specific representations making such assumptions unrealistic. In contrast, Theorem 4 establishes concrete guarantees by probabilistically relating target and source distributions. Finally, $\text{Bias}_{\mathcal{T}}(\boldsymbol{\phi}^{\star})$ term is unavoidable, however, similar to $d_{\mathcal{T}}(\hat{\boldsymbol{\phi}}; \boldsymbol{\phi}^{\star})$, it will be small as long as source and target tasks benefit from a shared supernet at the population level.

4 Guarantees for Linear Representations

As a concrete instantiation of Multipath MTL, consider a linear representation learning problem where each module ψ^k_{ℓ} applies matrix multiplications parameterized by $\boldsymbol{B}^k_{\ell}$ with dimensions $p_{\ell} \times p_{\ell-1}$: $\psi^k_{\ell}(\boldsymbol{x}) = \boldsymbol{B}^k_{\ell}\boldsymbol{x}$. Here p_{ℓ} are module dimensions with input dimension $p_0 = p$ and output dimension p_L. Given a path α, we obtain the linear representation $\boldsymbol{B}_{\alpha} = \Pi_{\ell=1}^{L} \boldsymbol{B}_{\ell}^{\alpha[\ell]} \in \mathbb{R}^{p_L \times p}$ where p_L is the number of rows of the final module $\boldsymbol{B}_L^{\alpha[L]}$. When $p_L \ll p$, $\boldsymbol{B}_{\alpha}$ is a fat matrix that projects $\boldsymbol{x} \in \mathbb{R}^p$ onto a lower dimensional subspace. This way, during few-shot adaptation, we only need to train $p_L \ll p$ parameters with features $\boldsymbol{B}_{\alpha}\boldsymbol{x}$. This is also the central idea in several works on linear meta-learning (Kong et al. 2020a; Sun et al. 2021; Bouniot et al. 2020; Tripuraneni, Jin, and Jordan 2021) which focus on a single linear representation. Our discussion within this section extends these results to the Multipath MTL setting.

Denote $\boldsymbol{f} = \{((\boldsymbol{B}^k_{\ell})_{k=1}^{K_{\ell}})_{\ell=1}^{L}, (\boldsymbol{h}_t, \alpha_t)_{t=1}^{T}\}$ where $\boldsymbol{h}_t \in \mathbb{R}^{p_L}$ are linear prediction heads. Let $\mathcal{F}$ be the search space associated with $\boldsymbol{f}$. Follow the similar setting as in Section 2 and let $\mathcal{X} \subset \mathbb{R}^p$. Given dataset $\mathcal{S}_{\text{all}} = (\mathcal{S}_t)_{t=1}^{T}$, we study

$$\hat{\boldsymbol{f}} = \min_{\boldsymbol{f} \in \mathcal{F}} \widehat{\mathcal{L}}_{\mathcal{S}_{\text{all}}}(\boldsymbol{f}) := \frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} (y_{ti} - \boldsymbol{h}_t^{\top} \boldsymbol{B}_{\alpha_t} \boldsymbol{x}_{ti})^2. \tag{4}$$

Let $\mathcal{B}^p(r) \subset \mathbb{R}^p$ be the Euclidean ball of radius r. To proceed, we make the following assumption for a constant $C \geq 1$.

Assumption 3 *For all $\ell \in [L]$, Ψ_{ℓ} is the set of matrices with operator norm bounded by C and $\mathcal{H} = \mathcal{B}^{p_L}(C)$.*

The result below is a variation of Theorem 1 where the bound is refined for linear representations (with finite parameters).

Theorem 3 *Suppose Assumptions 2&3 hold, and input set $\mathcal{X} \subset \mathcal{B}^p(R)$ for a constant $R > 0$. Then, with probability at least $1-\delta$,*

$$\mathcal{R}_{M^2TL}(\hat{\boldsymbol{f}}) \lesssim \sqrt{\frac{L \cdot \textit{DoF}(\mathcal{F})}{NT}} + \sqrt{\frac{\log|\mathcal{A}|}{N} + \frac{\log(2/\delta)}{NT}},$$

where $\textit{DoF}(\mathcal{F}) = T \cdot p_L + \sum_{\ell=1}^{L} K_{\ell} \cdot p_{\ell} \cdot p_{\ell-1}$ is the total number of trainable parameters in $\mathcal{F}$.

We note that Theorem 3 can be stated more generally for neural nets by placing ReLU activations between layers. Here $\lesssim$ subsumes the logarithmic dependencies, and the sample complexity has linear dependence on L (rather than exponential dependence as in Thm 1). In essence, it implies small task-averaged excess risk as soon as total sample size $\gtrsim$ total number of weights.

While flexible, this result does not guarantee that $\hat{\boldsymbol{f}}$ can benefit transfer learning for a new task. To proceed, we introduce additional assumptions under which we can guarantee the success of (TLOP). The first assumption is a realizability condition that guarantees tasks share same supernet representation (so that supernet bias is small).

Assumption 4 ***(A)*** *Task datasets are generated from a planted model $(\boldsymbol{x}_t, y_t) \sim \mathcal{D}_t$ where $y_t = \boldsymbol{x}_t^{\top}\boldsymbol{\theta}^{\star}_t + z_t$ where*

*$\boldsymbol{x}_t, z_t$ are zero mean, $\mathcal{O}(1)$-subgaussian and $\mathbb{E}[\boldsymbol{x}_t\boldsymbol{x}_t^\top] = \boldsymbol{I}_p$. **(B)** Task vectors are generated according to ground-truth supernet $\boldsymbol{f}^\star = \{((\bar{\boldsymbol{B}}_\ell^k)_{k=1}^{K_\ell})_{\ell=1}^L, (\bar{\boldsymbol{h}}_t, \bar{\alpha}_t)_{t=1}^T\}$ so that $\boldsymbol{\theta}_t^\star = \bar{\boldsymbol{B}}_{\bar{\alpha}_t}^\top \bar{\boldsymbol{h}}_t$. $\boldsymbol{f}^\star$ is normalized so that $\|\bar{\boldsymbol{B}}_\ell^k\| = \|\bar{\boldsymbol{h}}_t\| = 1$.*

Our second assumption is a task diversity condition adapted from (Tripuraneni, Jin, and Jordan 2021; Kong et al. 2020b) that facilitates the identifiability of the ground truth supernet.

Assumption 5 (Diversity during MTL) *Cluster the tasks by their pathways via $\boldsymbol{H}_\alpha = \{\boldsymbol{h}_t \mid \bar{\alpha}_t = \alpha\}$. Define cluster population $\gamma_\alpha = |\boldsymbol{H}_\alpha|/p_L$ and covariance $\boldsymbol{\Sigma}_\alpha = \gamma_\alpha^{-1}\sum_{\boldsymbol{h}\in\boldsymbol{H}_\alpha}\boldsymbol{h}\boldsymbol{h}^\top$. For a proper constant $c > 0$ and for all pathways α we have $\boldsymbol{\Sigma}_\alpha \succeq c\boldsymbol{I}_{p_L}$.*

Verbally, this condition requires that, if a pathway is chosen by a source task, that pathway should contain diverse tasks so that (M^{2}TL) phase can learn a good representation that can benefit transfer learning. However, this definition is flexible in the sense that pathways can still have sophisticated interactions/intersections and we don't assume anything for the pathways that are not chosen by source. We also have the challenge that, some pathways can be a lot more populated than others and target task might suffer from poor MTL representation quality over less populated pathways. The following assumption is key to overcoming this issue by enforcing a distributional prior on the target task pathway so that *its pathway is similar to the source tasks in average*.

Assumption 6 (Distribution of target task) *Draw $\alpha_\mathcal{T}$ uniformly at random from source pathways $(\bar{\alpha}_t)_{t=1}^T$. Target task is distributed as in Assumption 4(A) with pathway $\alpha_\mathcal{T}$ and $\boldsymbol{\theta}_\mathcal{T}^\star = \bar{\boldsymbol{B}}_{\alpha_\mathcal{T}}^\top \boldsymbol{h}_\mathcal{T}$ with $\|\boldsymbol{h}_\mathcal{T}\| = 1$.*

With these assumptions, we have the following result that guarantees end-to-end multipath learning ((M^{2}TL) phase followed by (TLOP) using MTL representation).

Theorem 4 *Suppose Assumptions 3–6 hold and $\ell(\hat{y}, y) = (y-\hat{y})^2$. Additionally assume input set $\mathcal{X} \subset \mathcal{B}^p(R)$ for a constant $R > 0$ and $\mathcal{H}_\mathcal{T} \subset \mathbb{R}^{p_L}$. Solve MTL problem* (M^{2}TL) *with the knowledge of ground-truth pathways $(\bar{\alpha}_t)_{t=1}^T$ to obtain a supernet $\hat{\boldsymbol{\phi}}$ and $NT \gtrsim DoF(\mathcal{F})\log(NT)$. Solve transfer learning problem* (TLOP) *with $\hat{\boldsymbol{\phi}}$ to obtain a target hypothesis $\hat{f}_{\hat{\boldsymbol{\phi}}}$. Then, with probability at least $1 - 3e^{-cM} - \delta$, path-averaged excess target risk* (3) *obeys $\mathbb{E}_{\alpha_\mathcal{T}}[\mathcal{R}_{TLOP}(\hat{f}_{\hat{\boldsymbol{\phi}}})]$*

$$\lesssim p_L\sqrt{\frac{L\cdot DoF(\mathcal{F}) + \log(8/\delta)}{NT}} + \frac{p_L}{M} + \sqrt{\frac{\log(8|\mathcal{A}|/\delta)}{M}}.$$

Here $DoF(\mathcal{F}) = T\cdot p_L + \sum_{\ell=1}^L K_\ell \cdot p_\ell \cdot p_{\ell-1}$, and $\mathbb{E}_{\alpha_\mathcal{T}}$ denotes the expectation over the random target pathways.

In words, this result controls the target risk in terms of the sample size of the target task and sample size during multitask representation learning, and provides a concrete instantiation of discussion following Theorem 2. In appendix, we provide a tighter bound for expected transfer risk when linear head $\boldsymbol{h}_\mathcal{T}$ is uniformly drawn from the unit sphere. The primary challenge in our work compared to related vanilla MTL results by (Tripuraneni, Jin, and Jordan 2021; Du et al. 2020; Kong et al. 2020b) is the fact that, we deal with exponentially many pathway representations many of which may be low quality. Assumption 6 allows us to convert task-averaged MTL risk into a transfer learning guarantee over a *random pathway*. Finally, Theorem 4 assumes that source pathways are known during MTL phase. In appendix, we show that this assumption is indeed necessary: Otherwise, one can construct scenarios where (M^{2}TL) problem admits an alternative solution $\tilde{\boldsymbol{f}}$ with optimal MTL risk but the resulting supernet $\tilde{\boldsymbol{\phi}}$ achieves poor target risk. Supplementary material discusses this challenge and identifies additional conditions that make ground-truth pathways uniquely identifiable when we solve (M^{2}TL).

5 Insights from Hierarchical Representations

We now discuss the special two-layer supernet structure depicted in Figure 1b. This setting groups tasks into $K := K_2$ clusters and first layer module is shared across all tasks ($K_1 = 1$). Ignoring first layer, pathway $\alpha_t \in [K]$ becomes the clustering assignment for task t. Applying Theorem 1, we obtain a generalization bound of

$$\mathcal{R}_{\text{M}^2\text{TL}}(\hat{\boldsymbol{f}}) \lesssim \widetilde{\mathcal{G}}_{NT}(\Psi_1) + \sqrt{K}\widetilde{\mathcal{G}}_{NT}(\Psi_2) + \widetilde{\mathcal{G}}_N(\mathcal{H}) + \sqrt{\frac{\log K}{N}}.$$

Here, $\psi_1 \in \Psi_1$ is the shared first layer module, $\psi_2^k \in \Psi_2$ is the module assigned to cluster $k \in [K]$ that personalizes its representation, and we have $|\mathcal{A}| = K$. To provide further insights, let us focus on linear representations with the notation of Section 4: $\psi_1(\boldsymbol{x}) = \boldsymbol{B}_1\boldsymbol{x}$, $\psi_2^k(\boldsymbol{x}') = \boldsymbol{B}_2^k\boldsymbol{x}'$, and $h_t(\boldsymbol{x}'') = \boldsymbol{h}_t^\top\boldsymbol{x}''$ with dimensions $\boldsymbol{B}_1 \in \mathbb{R}^{R\times p}$, $\boldsymbol{B}_2^k \in \mathbb{R}^{r\times R}$, $\boldsymbol{h}_t \in \mathbb{R}^r$ and $r \le R \le p$. Our bound now takes the form

$$\mathcal{R}_{\text{M}^2\text{TL}}(\hat{\boldsymbol{f}}) \lesssim \sqrt{\frac{Rp + KrR + T(r+\log K)}{NT}},$$

where Rp and KrR are the number of parameters in supernet layers 1 and 2, and $(r+\log K)/N$ is the cost of learning pathway and prediction head per task. Let us contrast this to the shallow MTL approaches with 1-layer supernets.

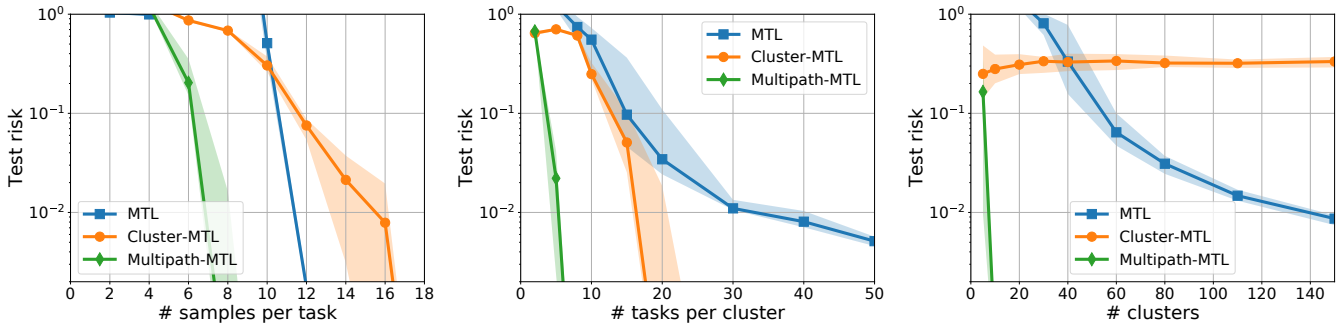
• **Vanilla MTL:** Learn $\boldsymbol{B}_1 \in \mathbb{R}^{R\times p}$ and learn larger prediction heads $\boldsymbol{h}_t^V \in \mathbb{R}^R$ (no clustering needed).

• **Cluster MTL:** Learn larger cluster modules $\boldsymbol{B}_2^{C,k} \in \mathbb{R}^{r\times p}$, and learn pathway α_t and head $\boldsymbol{h}_t \in \mathbb{R}^r$ (no $\boldsymbol{B}_1$ needed).

Experimental Insights. Before providing a theoretical comparison, let us discuss the experimental results where we compare these three approaches in a realizable dataset generated according to Figure 1b. Specifically, we generate $\bar{\boldsymbol{B}}_1$ and $\{\bar{\boldsymbol{B}}_2^k\}_{k=1}^K$ with orthonormal rows uniformly at random independently. We also generate $\bar{\boldsymbol{h}}_t$ uniformly at random over the unit sphere independently. Let $\bar{\alpha}_t$ be the cluster assignment of task t where each cluster has same size/number of tasks with $\bar{T} = T/K$ tasks. The distribution $\mathcal{D}_t$ associated with task t is generated as

$$y = \boldsymbol{x}^\top\boldsymbol{\theta}_t^\star \quad \text{where} \quad \boldsymbol{\theta}_t^\star = (\bar{\boldsymbol{h}}_t^\top\bar{\boldsymbol{B}}_2^{\alpha_t}\bar{\boldsymbol{B}}_1)^\top,\ \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p),$$

without label noise. We evaluate and present results from two scenarios where cluster assignment of each task $\bar{\alpha}_t$ is known (Figure 2) or not (Figure 3). MTL, Cluster-MTL and



(a) Varying n with $\bar{T} = 10, K = 40$ (b) Varying $\bar{T}$ with $N = 10, K = 40$ (c) Varying K with $N = 10, \bar{T} = 10$

Figure 2: We compare the sample complexity of MTL, Cluster-MTL and Multipath-MTL in a noiseless linear regression setting. For each figure, we fix two of the configurations and vary the other one. We find that Multipath-MTL is superior to both baselines of MTL and Cluster-MTL as predicted by our theory. The solid curves are the median risk and the shaded regions highlight the first and third quantile risks. Each marker is obtained by averaging 20 independent realizations.

Multipath-MTL labels corresponds to our single representation, clustering and hierarchical MTL strategies respectively, in the figures.

In Figure 2, we solve MTL problems with the knowledge of clustering $\bar{\alpha}_t$. We set ambient dimension $p = 32$, shared embedding $R = 8$, and cluster embeddings $r = 2$. We consider a base configuration of $K = 40$ clusters, $\bar{T} = T/K = 10$ tasks per cluster and $N = 10$ samples per task (see supplementary material for further details). Figure 2 compares the performance of three approaches for the task-averaged MTL test risk and demonstrates consistent benefits of Multipath MTL for varying $K, \bar{T}, N$.

We also consider the setting where $\bar{\alpha}_t$, $t \in [T]$ are unknown during training. Set $p = 128$, $R = 32$ and $r = 2$, and fix number of clusters $K = 50$ and cluster size $\bar{T} = 10$. In this experiment, instead of using the ground truth clustering $\bar{\alpha}_t$, we also learn the clustering assignment $\hat{\alpha}_t$ for each task. As we discussed and visualized in supplementary material, it is not easy to cluster random tasks even with the hindsight knowledge of task vectors $\boldsymbol{\theta}_t^\star$. To overcome this issue, we add correlation between tasks in the same cluster. Specifically, generate the prediction head by $\bar{\boldsymbol{h}}_t' = \gamma \bar{\boldsymbol{h}}^k + (1-\gamma)\bar{\boldsymbol{h}}_t$ where $\bar{\boldsymbol{h}}^k, \bar{\boldsymbol{h}}_t$ are random unit vectors corresponding to the cluster k and task t (assuming $\bar{\alpha}_t = k$). To cluster tasks, we first run vanilla MTL and learn the shared representation $\hat{\boldsymbol{B}}_1$ and heads $(\hat{\boldsymbol{h}}_t^V)_{t=1}^T$. Next build task vector estimates by $\hat{\boldsymbol{\theta}}_t := \hat{\boldsymbol{B}}_i^\top \hat{\boldsymbol{h}}_t^V$, and get $T \times T$ task similarity matrix using Euclidean distance metric. Applying standard K-means clustering to it provides a clustering assignment $\hat{\alpha}_t$. In the experiment, we set $\gamma = 0.6$ to make sure hindsight knowledge of $\boldsymbol{\theta}_t^\star$ is sufficient to correctly cluster all tasks. Results are presented in Figure 3, where solid curves are solving MTL with ground truth $\bar{\alpha}_t$ while dashed curves are using $\hat{\alpha}_t$. We observe that when given enough samples ($N \geq 60$), all tasks are grouped correctly even if the MTL risk is not zero. More importantly, Multipath MTL does outperform both vanilla MTL and cluster MTL even when the clustering is not fully correct.

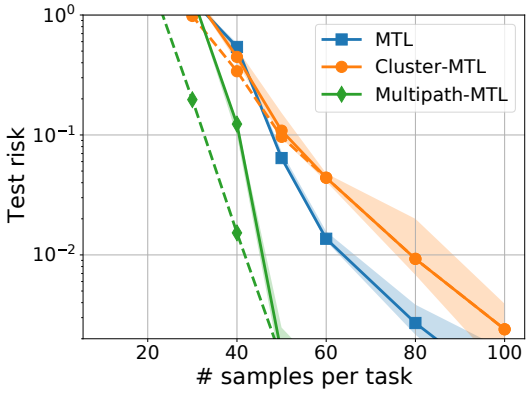


Figure 3: We group the $T = 500$ tasks into $K = 50$ clusters and compare the sample complexity of different MTL strategies. Given different sample size, we cluster tasks based on the trained MTL model and solve Cluster-/Multipath-MTL based on the assigned clusters. Solid curves are results using ground truth cluster knowledge $\bar{\alpha}_t$ and dashed are using the learned clustering $\hat{\alpha}_t$. Experimental setting follows the same setting as in Figure 2.

Understanding the benefits of Multipath MTL. Naturally, superior numerical performance of Multipath MTL in Figure 2&3 partly stems from the hierarchical dataset model we study. This model will also shed light on shortcomings of 1-layer supernets drawing from our theoretical predictions. First, observe that all three baselines are exactly specified: We use the smallest model sizes that capture the ground-truth model so that they can achieve zero test risk as N, K, T grows. For instance, Vanilla MTL achieves zero risk by setting $\boldsymbol{B}_1 = \bar{\boldsymbol{B}}_1, \boldsymbol{h}_t^V = (\bar{\boldsymbol{B}}_2^{\alpha_t})^\top \bar{\boldsymbol{h}}_t$ and cluster MTL achieves zero risk by setting $\boldsymbol{B}_2^{C,k} = \bar{\boldsymbol{B}}_2^k \bar{\boldsymbol{B}}_1, \boldsymbol{h}_t = \bar{\boldsymbol{h}}_t$. Thus, the benefit of Multipath MTL arises from stronger weight sharing across tasks that reduces test risk. In light of Sec. 4, the generalization risks of these approaches can be bounded as

$\sqrt{\text{DoF}(\mathcal{F})/NT}$ where Number-of-Parameters compare as **Vanilla:** $Rp + TR$, **Cluster:** $Krp + Tr$, **Multipath:** $Rp + KrR + Tr$. From this, it can be seen that Multipath is never worse than the others as long as $Kr \geq R$ and $\bar{T} = T/K \geq r$. These conditions hold under the assumption that multipath model is of minimal size: Otherwise, there would be a strictly smaller zero-risk model by setting $R \leftarrow Kr$ and $r \leftarrow \bar{T}$.

Conversely, Multipath shines in the regime $Kr \gg R$ or $\bar{T} \gg r$. As $\frac{Kr}{R}, \frac{p}{R} \to \infty$, Multipath strictly outperforms Cluster MTL. This arises from a *cluster diversity* phenomenon that connects to the *task diversity* notions of prior art. In essence, since r-dimensional clusters lie on a shared R dimensional space, as we add more clusters beyond $Kr \geq R$, they will collaboratively estimate the shared subspace which in turn helps estimating their local subspaces by projecting them onto the shared one. As $\frac{\bar{T}}{r}, \frac{R}{r} \to \infty$, Multipath strictly outperforms Vanilla MTL. $\frac{\bar{T}}{r}$ is needed to ensure that there is enough task diversity within each cluster to estimate its local subspace. Finally, $\frac{R}{r}$ ratio is the few-shot learning benefit of clustering over Vanilla MTL. The prediction heads of vanilla MTL is larger which necessitates a larger N, at the minimum $N \geq R$. Whereas Multipath works with as little as $N \geq r$. The same argument also implies that clustering/hierarchy would also enable better transfer learning.

6 Related Work

Our work is related to a large body of literature spanning efficient architectures and statistical guarantees for MTL, representation learning, task similarity, and subspace clustering.

• **Multitask Representation Learning.** While MTL problems admit multiple approaches, an important idea is building shared representations to embed tasks in a low-dimensional space (Thrun and Pratt 2012; Baxter 2000). After identifying this low-dimensional representation, new tasks can be learned in a sample efficient fashion inline with the benefits of deep representations in modern ML applications. While most earlier works focus on linear models, (Maurer, Pontil, and Romera-Paredes 2016) provides guarantees for general hypothesis classes through empirical process theory improving over (Baxter 2000). More recently, there is a growing line of work on multitask representations that spans tighter sample complexity analysis (Garg and Liang 2020; Hanneke and Kpotufe 2020; Du et al. 2020; Kong et al. 2020b; Xu and Tewari 2021; Lu, Huang, and Du 2021), convergence guarantees (Collins et al. 2022; Ji et al. 2020; Collins et al. 2021; Wu, Zhang, and Ré 2020), lifelong learning (Xu and Tewari 2022; Li et al. 2022), and decision making problems (Yang et al. 2020; Qin et al. 2022; Sodhani, Zhang, and Pineau 2021). Closest to our work is (Tripuraneni, Jin, and Jordan 2021) which provides tighter sample complexity guarantees compared to (Maurer, Pontil, and Romera-Paredes 2016). Our problem formulation generalizes prior work (that is mostly limited to single shared representation) by allowing deep compositional representations computed along supernet pathways. To overcome the associated technical challenges, we develop multilayer chain rules for Gaussian Complexity, introduce new notions to assess the quality of supernet representations, and develop new theory for linear representations.

• **ML Architectures and Systems.** While traditional ML models tend to be good at a handful of tasks, next-generation of neural architectures are expected to excel at a diverse range of tasks while allowing for multiple input modalities. To this aim, task-specific representations can help address both computational and data efficiency challenges. Recent works (Shu et al. 2021; Fifty et al. 2021; Yao et al. 2019; Vuorio et al. 2019; Mansour et al. 2020; Tan et al. 2022; Ghosh et al. 2020; Collins et al. 2021) propose hierarchical/clustering approaches to group tasks in terms of their similarities, (Qin et al. 2020; Ye, Zha, and Ren 2022; Gupta et al. 2022; Asai et al. 2022; He et al. 2022) focus on training mixture-of-experts (MoE) models, and similar to the pathways (Strezoski, Noord, and Worring 2019; Rosenbaum, Klinger, and Riemer 2017; Chen, Gu, and Fu 2021; Ma et al. 2019) study on task routing. In the context of lifelong learning, PathNet, PackNet (Fernando et al. 2017; Mallya and Lazebnik 2018) and many other existing methods (Parisi et al. 2019; Mallya, Davis, and Lazebnik 2018; Hung et al. 2019; Wortsman et al. 2020; Cheung et al. 2019) propose to embed many tasks into the same network to facilitate sample/compute efficiency. PathNet as well as SNR (Ma et al. 2019) propose methods to identify pathways/routes for individual tasks and efficiently compute them over the conditional subnetwork. With the advent of large language models, conditional computation paradigm is witnessing a growing interest with architectural innovations such as muNet, GShard, Pathways, and PaLM (Gesmundo and Dean 2022a,b; Barham et al. 2022; Dean 2021; Lepikhin et al. 2020; Chowdhery et al. 2022; Driess et al. 2023) and provide a strong motivation for theoretically-grounded Multipath MTL methods.

7 Discussion

This work explored novel multitask learning problems which allow for task-specific representations that are computed along pathways of a large supernet. We established generalization bounds under a general setting which proved insightful when specialized to linear or hierarchical representations. We believe there are multiple exciting directions to explore. First, it is desirable to develop a stronger control over the generalization risk of specific groups of tasks. Our Lemma 1 is a step in this direction. Second, what are risk upper/lower bounds for Multipath MTL as we vary the depth and width of the supernet graph? Discussion in Section 5 falls under this question where we demonstrate the sample complexity benefits of Multipath MTL over traditional MTL approaches. Finally, following experiments in Section 5, can we establish similar provable guarantees for computationally-efficient algorithms (e.g. method of moments, gradient descent)?

Acknowledgements

Authors would like to thank Zhe Zhao for helpful discussions and pointing out related works. This work was supported in part by the NSF grants CCF-2046816 and CCF-2212426, Google Research Scholar award, and Army Research Office grant W911NF2110312.

References

Asai, A.; Salehi, M.; Peters, M. E.; and Hajishirzi, H. 2022. Attentional Mixtures of Soft Prompt Tuning for Parameter-efficient Multi-task Knowledge Sharing. *arXiv:2205.11961*.

Barham, P.; Chowdhery, A.; Dean, J.; Ghemawat, S.; Hand, S.; Hurt, D.; Isard, M.; Lim, H.; Pang, R.; Roy, S.; et al. 2022. Pathways: Asynchronous distributed dataflow for ML. *Proceedings of Machine Learning and Systems*, 4: 430–449.

Bartlett, P. L.; Foster, D. J.; and Telgarsky, M. J. 2017. Spectrally-normalized margin bounds for neural networks. In *Neural Information Processing Systems*, 6241–6250.

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

Bouniot, Q.; Redko, I.; Audigier, R.; Loesch, A.; Zotkin, Y.; and Habrard, A. 2020. Towards better understanding meta-learning methods through multi-task representation learning theory. *arXiv preprint arXiv:2010.01992*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Neural information processing systems*, 33: 1877–1901.

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

Chen, S.; Crammer, K.; He, H.; Roth, D.; and Su, W. J. 2021. Weighted Training for Cross-Task Learning. *arXiv preprint arXiv:2105.14095*.

Chen, X.; Gu, X.; and Fu, L. 2021. Boosting share routing for multi-task learning. In *Companion Proceedings of the Web Conference 2021*, 372–379.

Cheung, B.; Terekhov, A.; Chen, Y.; Agrawal, P.; and Olshausen, B. 2019. Superposition of many models into one. *Advances in neural information processing systems*, 32.

Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H. W.; Sutton, C.; Gehrmann, S.; et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

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

Collins, L.; Mokhtari, A.; Oh, S.; and Shakkottai, S. 2022. MAML and ANIL provably learn representations. *arXiv preprint arXiv:2202.03483*.

Dean, J. 2021. Introducing Pathways: A next-generation AI architecture. https://blog.google/technology/ai/introducing-pathways-next-generation-ai-architecture/, Google AI Blog.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Driess, D.; Xia, F.; Sajjadi, M. S. M.; Lynch, C.; Chowdhery, A.; Ichter, B.; Wahid, A.; Tompson, J.; Vuong, Q.; Yu, T.; Huang, W.; Chebotar, Y.; Sermanet, P.; Duckworth, D.; Levine, S.; Vanhoucke, V.; Hausman, K.; Toussaint, M.; Greff, K.; Zeng, A.; Mordatch, I.; and Florence, P. 2023. PaLM-E: An Embodied Multimodal Language Model. In *arXiv preprint arXiv:2303.03378*.

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

Fernando, C.; Banarse, D.; Blundell, C.; Zwols, Y.; Ha, D.; Rusu, A. A.; Pritzel, A.; and Wierstra, D. 2017. Pathnet: Evolution channels gradient descent in super neural networks. *arXiv preprint arXiv:1701.08734*.

Fifty, C.; Amid, E.; Zhao, Z.; Yu, T.; Anil, R.; and Finn, C. 2021. Efficiently identifying task groupings for multi-task learning. *Advances in Neural Information Processing Systems*, 34: 27503–27516.

Garg, S.; and Liang, Y. 2020. Functional regularization for representation learning: A unified theoretical perspective. *Neural Information Processing Systems*, 33: 17187–17199.

Gesmundo, A.; and Dean, J. 2022a. An Evolutionary Approach to Dynamic Introduction of Tasks in Large-scale Multitask Learning Systems. *arXiv preprint arXiv:2205.12755*.

Gesmundo, A.; and Dean, J. 2022b. muNet: Evolving Pretrained Deep Neural Networks into Scalable Auto-tuning Multitask Systems. *arXiv preprint arXiv:2205.10937*.

Ghosh, A.; Chung, J.; Yin, D.; and Ramchandran, K. 2020. An efficient framework for clustered federated learning. *Neural Information Processing Systems*, 33: 19586–19597.

Golowich, N.; Rakhlin, A.; and Shamir, O. 2018. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, 297–299. PMLR.

Gupta, S.; Mukherjee, S.; Subudhi, K.; Gonzalez, E.; Jose, D.; Awadallah, A. H.; and Gao, J. 2022. Sparsely activated mixture-of-experts are robust multi-task learners. *arXiv preprint arXiv:2204.07689*.

Hanneke, S.; and Kpotufe, S. 2020. A no-free-lunch theorem for multitask learning. *arXiv preprint arXiv:2006.15785*.

He, C.; Zheng, S.; Zhang, A.; Karypis, G.; Chilimbi, T.; Soltanolkotabi, M.; and Avestimehr, S. 2022. SMILE: Scaling Mixture-of-Experts with Efficient Bi-level Routing. *arXiv preprint arXiv:2212.05191*.

Hung, C.-Y.; Tu, C.-H.; Wu, C.-E.; Chen, C.-H.; Chan, Y.-M.; and Chen, C.-S. 2019. Compacting, picking and growing for unforgetting continual learning. *Advances in Neural Information Processing Systems*, 32.

Ji, K.; Lee, J. D.; Liang, Y.; and Poor, H. V. 2020. Convergence of meta-learning with task-specific adaptation over partial parameters. *Advances in Neural Information Processing Systems*, 33: 11490–11500.

Kong, W.; Somani, R.; Kakade, S.; and Oh, S. 2020a. Robust meta-learning for mixed linear regression with small batches. *Neural information processing systems*, 33: 4683–4696.

Kong, W.; Somani, R.; Song, Z.; Kakade, S.; and Oh, S. 2020b. Meta-learning for mixed linear regression. In *International Conf on Machine Learning*, 5394–5404. PMLR.

Lepikhin, D.; Lee, H.; Xu, Y.; Chen, D.; Firat, O.; Huang, Y.; Krikun, M.; Shazeer, N.; and Chen, Z. 2020. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*.

Li, Y.; Li, M.; Asif, M. S.; and Oymak, S. 2022. Provable and Efficient Continual Representation Learning. *arXiv preprint arXiv:2203.02026*.

Li, Y.; and Oymak, S. 2023. Provable Pathways: Learning Multiple Tasks over Multiple Paths. *arXiv preprint arXiv:2303.04338*.

Liu, H.; Simonyan, K.; and Yang, Y. 2019. Darts: Differentiable architecture search. *ICLR*.

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

Ma, J.; Zhao, Z.; Chen, J.; Li, A.; Hong, L.; and Chi, E. H. 2019. Snr: Sub-network routing for flexible parameter sharing in multi-task learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 216–223.

Mallya, A.; Davis, D.; and Lazebnik, S. 2018. Piggyback: Adapting a single network to multiple tasks by learning to mask weights. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 67–82.

Mallya, A.; and Lazebnik, S. 2018. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7765–7773.

Mansour, Y.; Mohri, M.; Ro, J.; and Suresh, A. T. 2020. Three approaches for personalization with applications to federated learning. *arXiv preprint arXiv:2002.10619*.

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

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

Neyshabur, B.; Bhojanapalli, S.; McAllester, D.; and Srebro, N. 2017. Exploring generalization in deep learning. *Advances in neural information processing systems*, 30.

Neyshabur, B.; Li, Z.; Bhojanapalli, S.; LeCun, Y.; and Srebro, N. 2018. Towards understanding the role of over-parametrization in generalization of neural networks. *arXiv preprint arXiv:1805.12076*.

Parisi, G. I.; Kemker, R.; Part, J. L.; Kanan, C.; and Wermter, S. 2019. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113: 54–71.

Qin, Y.; Menara, T.; Oymak, S.; Ching, S.; and Pasqualetti, F. 2022. Non-Stationary Representation Learning in Sequential Linear Bandits. *IEEE Open Journal of Control Systems*.

Qin, Z.; Cheng, Y.; Zhao, Z.; Chen, Z.; Metzler, D.; and Qin, J. 2020. Multitask mixture of sequential experts for user activity streams. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3083–3091.

Rosenbaum, C.; Klinger, T.; and Riemer, M. 2017. Routing networks: Adaptive selection of non-linear functions for multi-task learning. *arXiv preprint arXiv:1711.01239*.

Shu, Y.; Kou, Z.; Cao, Z.; Wang, J.; and Long, M. 2021. Zoo-tuning: Adaptive transfer from a zoo of models. In *International Conference on Machine Learning*, 9626–9637. PMLR.

Sodhani, S.; Zhang, A.; and Pineau, J. 2021. Multi-task reinforcement learning with context-based representations. In *International Conference on Machine Learning*, 9767–9779. PMLR.

Strezoski, G.; Noord, N. v.; and Worring, M. 2019. Many task learning with task routing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1375–1384.

Sun, Y.; Narang, A.; Gulluk, I.; Oymak, S.; and Fazel, M. 2021. Towards sample-efficient overparameterized meta-learning. *Advances in Neural Information Processing Systems*, 34: 28156–28168.

Tan, A. Z.; Yu, H.; Cui, L.; and Yang, Q. 2022. Towards personalized federated learning. *IEEE Transactions on Neural Networks and Learning Systems*.

Thrun, S.; and Pratt, L. 2012. *Learning to learn*. Springer Science & Business Media.

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

Tripuraneni, N.; Jordan, M.; and Jin, C. 2020. On the theory of transfer learning: The importance of task diversity. *Advances in Neural Information Processing Systems*, 33: 7852–7862.

Vuorio, R.; Sun, S.-H.; Hu, H.; and Lim, J. J. 2019. Multimodal model-agnostic meta-learning via task-aware modulation. *Advances in Neural Information Processing Systems*, 32.

Wortsman, M.; Ramanujan, V.; Liu, R.; Kembhavi, A.; Rastegari, M.; Yosinski, J.; and Farhadi, A. 2020. Supermasks in superposition. *Advances in Neural Information Processing Systems*, 33: 15173–15184.

Wu, S.; Zhang, H. R.; and Ré, C. 2020. Understanding and improving information transfer in multi-task learning. *arXiv preprint arXiv:2005.00944*.

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

Xu, Z.; and Tewari, A. 2022. On the statistical benefits of curriculum learning. In *International Conference on Machine Learning*, 24663–24682. PMLR.

Yang, J.; Hu, W.; Lee, J. D.; and Du, S. S. 2020. Impact of representation learning in linear bandits. *arXiv preprint arXiv:2010.06531*.

Yao, H.; Wei, Y.; Huang, J.; and Li, Z. 2019. Hierarchically structured meta-learning. In *International Conference on Machine Learning*, 7045–7054. PMLR.

Ye, Q.; Zha, J.; and Ren, X. 2022. Eliciting Transferability in Multi-task Learning with Task-level Mixture-of-Experts. *arXiv preprint arXiv:2205.12701*.