Provable Pathways: Learning Multiple Tasks over Multiple Paths

Yingcong Li[*] **Samet Oymak**[*†]

[*] University of California, Riverside
[†] University of Michigan, Ann Arbor

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.

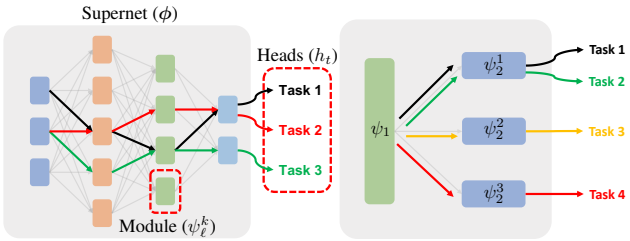
1 Introduction

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 $\psi_2^1 \circ \psi_1$.

[*]Emails: {yli692@,oymak@ece.}ucr.edu

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

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

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

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

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

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

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 $\mathcal{G}_n(\mathcal{Q}) \leq \widetilde{\mathcal{G}}_n(\mathcal{Q})$ 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 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 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 $\widetilde{\mathcal{G}}(\Phi_{\text{used}}) \leq \widetilde{\mathcal{G}}(\Phi)$. 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 $\log(\varepsilon\text{-covering number})$ (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*

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

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 $1-\delta$,*

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

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

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 $1-\delta$,*

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

where $\mathit{DoF}(\mathcal{F}) = T \cdot p_L + \sum_{\ell=1}^{L} K_{\ell} \cdot p_{\ell} \cdot p_{\ell-1}$ 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 Theorem 9 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 E, 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.

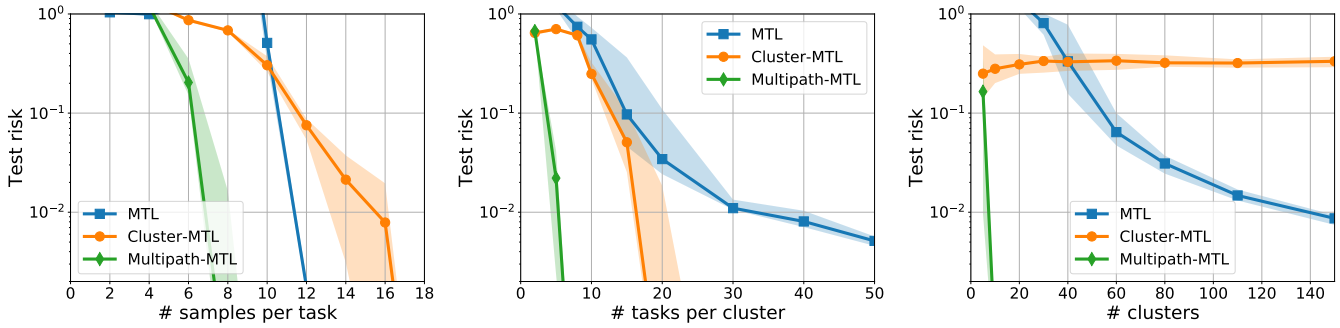
• **Vanilla MTL:** Learn $\boldsymbol{B}_1 \in \mathbb{R}^{R\times p}$ and learn larger prediction heads $\boldsymbol{h}_t^V \in \mathbb{R}^R$ (no clustering needed).

• **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.

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* phe-

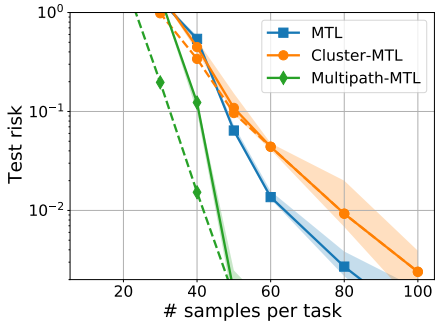


Figure 3: We group the $T = 500$ tasks into $K = 50$ clusters and compare the sample complexity of different MTL strategies. Given different sample size, we cluster tasks based on the trained MTL model and solve Cluster-/Multipath-MTL based on the assigned clusters. Solid curves are results using ground 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.

nomenon 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 (Zhang and Yang 2021; Thrun and Pratt 2012; Wang, Kolar, and Srebro 2016; 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 (Balcan, Khodak, and Talwalkar 2019; Khodak, Balcan, and Talwalkar 2019; 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; Cheng 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.

• **Quantifying Task Similarity and Clustering.** We note that task similarity and clustering has been studied by (Shui et al. 2019; Nguyen, Do, and Carneiro 2021; Zhou et al. 2020; Fifty et al. 2021; Kumar and Daume III 2012; Kang, Grauman, and Sha 2011; Aribandi et al. 2021; Zamir et al. 2018) however these works do not come with comparable statistical guarantees. Leveraging relations between tasks are explored even more broadly (Zhuang et al. 2020; Achille et al. 2021). Our experiments on linear Multipath MTL connects well with the broader subspace clustering literature (Vidal 2011; Parsons, Haque, and Liu 2004; Elhamifar and Vidal 2013). Specifically, each learning task $\boldsymbol{\theta}_t$ can be viewed as a point on a high-dimensional subspace. Multipath MTL aims to cluster these points into smaller subspaces that correspond to task-specific representations. Our challenge is that we only get to see the points through the associated datasets.

• **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 (Ramesh and Chaudhari 2021a; Shu et al. 2021; Ramesh and Chaudhari 2021b; 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

Achille, A.; Paolini, G.; Mbeng, G.; and Soatto, S. 2021. The information complexity of learning tasks, their structure and their distance. *Information and Inference: A Journal of the IMA*, 10(1): 51–72.

Aribandi, V.; Tay, Y.; Schuster, T.; Rao, J.; Zheng, H. S.; Mehta, S. V.; Zhuang, H.; Tran, V. Q.; Bahri, D.; Ni, J.; et al. 2021. Ext5: Towards extreme multi-task scaling for transfer learning. *arXiv preprint arXiv:2111.10952*.

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 preprint arXiv:2205.11961*.

Balcan, M.-F.; Khodak, M.; and Talwalkar, A. 2019. Provable guarantees for gradient-based meta-learning. In *International Conference on Machine Learning*, 424–433. PMLR.

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 *Advances 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. *Advances in 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.

Cheng, Y.; Feng, S.; Yang, J.; Zhang, H.; and Liang, Y. 2022. Provable benefit of multitask representation learning in reinforcement learning. *arXiv preprint arXiv:2206.05900*.

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*.

Elhamifar, E.; and Vidal, R. 2013. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE transactions on pattern analysis and machine intelligence*, 35(11): 2765–2781.

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. *Advances in 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. *Advances in 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.

Ji, Z.; and Telgarsky, M. 2018. Gradient descent aligns the layers of deep linear networks. *arXiv preprint arXiv:1810.02032*.

Kang, Z.; Grauman, K.; and Sha, F. 2011. Learning with whom to share in multi-task feature learning. In *ICML*.

Khodak, M.; Balcan, M.-F. F.; and Talwalkar, A. S. 2019. Adaptive gradient-based meta-learning methods. *Advances in Neural Information Processing Systems*, 32.

Kong, W.; Somani, R.; Kakade, S.; and Oh, S. 2020a. Robust meta-learning for mixed linear regression with small batches. *Advances in 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 Conference on Machine Learning*, 5394–5404. PMLR.

Kumar, A.; and Daume III, H. 2012. Learning task grouping and overlap in multi-task learning. *arXiv preprint arXiv:1206.6417*.

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*.

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 preprint 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.

Mohri, M.; Rostamizadeh, A.; and Talwalkar, A. 2018. *Foundations of machine learning*. MIT press.

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*.

Nguyen, C.; Do, T.-T.; and Carneiro, G. 2021. Similarity of classification tasks. *arXiv preprint arXiv:2101.11201*.

Oymak, S. 2018. Learning compact neural networks with regularization. In *International Conference on Machine Learning*, 3966–3975. PMLR.

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.

Parsons, L.; Haque, E.; and Liu, H. 2004. Subspace clustering for high dimensional data: a review. *Acm sigkdd explorations newsletter*, 6(1): 90–105.

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.

Ramesh, R.; and Chaudhari, P. 2021a. Boosting a model zoo for multi-task and continual learning. *arXiv preprint arXiv:2106.03027*.

Ramesh, R.; and Chaudhari, P. 2021b. Model Zoo: A Growing Brain That Learns Continually. In *International Conference on Learning Representations*.

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.

Shui, C.; Abbasi, M.; Robitaille, L.-É.; Wang, B.; and Gagné, C. 2019. A principled approach for learning task similarity in multitask learning. *arXiv preprint arXiv:1903.09109*.

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.

Talagrand, M. 2006. *The generic chaining: upper and lower bounds of stochastic processes*. Springer Science & Business Media.

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.

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

Vershynin, R. 2018. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Vidal, R. 2011. Subspace clustering. *IEEE Signal Processing Magazine*, 28(2): 52–68.

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.

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

Wang, J.; Kolar, M.; and Srebro, N. 2016. Distributed multi-task learning with shared representation. *arXiv preprint arXiv:1603.02185*.

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*.

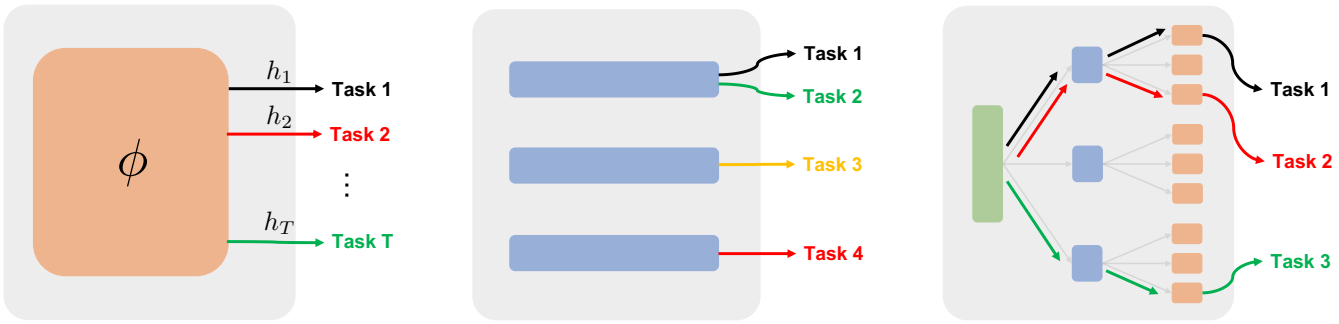
Yu, Y.; Wang, T.; and Samworth, R. J. 2015. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2): 315–323.

Zamir, A. R.; Sax, A.; Shen, W.; Guibas, L. J.; Malik, J.; and Savarese, S. 2018. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3712–3722.

Zhang, Y.; and Yang, Q. 2021. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*.

Zhou, F.; Shui, C.; Abbasi, M.; Robitaille, L.-É.; Wang, B.; and Gagné, C. 2020. Task similarity estimation through adversarial multitask neural network. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2): 466–480.

Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; and He, Q. 2020. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1): 43–76.



(a) Vanilla MTL (b) Cluster MTL (c) Hierarchical MTL

Figure 4: Three specific MTL settings: Vanilla MTL, Cluster MTL and Hierarchical MTL. In vanilla MTL, all the tasks share the same representation $\phi \in \Phi$, and each task learns its specific head $h_t \in \mathcal{H}$. It corresponds to the setting that $|\mathcal{A}| = 1$, $L = 1$ and $K_1 = 1$. In Cluster MTL, tasks are clustered into groups and different groups are assigned with different and uncorrelated representations. If we assume there are K clusters, then $|\mathcal{A}| = K$, $L = 1$ and $K_1 = K$. While, Fig. 1b shows the Hierarchical MTL with only two layers, here we present the more general Hierarchical MTL setting. Assume the degree of a hierarchical supernet is K (In Fig. 4c, $K = 3$), then $|\mathcal{A}| = K^{L-1}$ where L is the number of layers in supernet, and $K_\ell = K^{\ell-1}$.

Organization of the Supplementary Material

The supplementary material (SM) is organized as follows.

1. In Appendix A we introduce additional notions used throughout the supplementary material.
2. Appendix B provides our main proofs in Section 3 and introduces two direct corollaries of Theorem 1. We also provide a data-dependent bound in terms of empirical Gaussian complexity (rather than worst-case). In Appendix B.5 we also provide end-to-end transfer learning bound by introducing a proper notion of task diversity.
3. Appendix C provides additional guarantees (Thm 8) for parametric classes via non-data-dependent covering argument. The advantages of Theorem 8 are: (1) Sample complexity has linear dependence on supernet depth L (rather than exponential), (2) It applies to unbounded loss functions, (3) It is also a supporting result for the proof of Theorem 3&4.
4. Appendix D provides our proofs in Section 4. We also introduce Corollary 4, which is a direct application of Theorem 1. Lemma 7 proves the necessity of our Assumption 5.
5. In Appendix E, we include a short discussion on the challenges of transfer learning: Specifically, we provide a lemma/example that shows that, under the assumptions of Theorem 4, if ground-truth MTL pathways are not known, there are MTL settings for which transfer learning can provably fail. This construction highlights the (combinatorial) challenge of finding the right task clusterings during MTL phase that are actually useful for transfer phase.
6. Appendix F provides further details, algorithms, and results on numerical experiments in Section 5.

A Useful Definitions

We will start with some useful notions. Let $\|\cdot\|$ denote the ℓ_2-norm of a vector, and $[L]$ denote the set $\{1, 2, \ldots, L\}$. We denote the K times Cartesian product of a hypothesis set $\mathcal{Q}$ with itself by $\mathcal{Q}^K$. Now assume we have a hypothesis set $\mathcal{Q} : \mathcal{X} \to \mathbb{R}^r$ and an input dataset of size n, defined by $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, where $\boldsymbol{x}_i \in \mathcal{X}$. Let $\{\sigma_{ij}\}_{i\in[n], j\in[r]}$ denote Rademacher variables uniformly and independently taking values in $\{-1, 1\}$ and $\{g_{ij}\}_{i\in[n], j\in[r]}$ denote i.i.d. standard random Gaussian variables. Then we can define the empirical and population Rademacher/Gaussian complexities of a hypothesis set $\mathcal{Q}$ over inputs $\boldsymbol{X}$ and data space $\mathcal{X}$ with sample size n as

$$\text{Empirical/Population Rademacher complexities:} \quad \widehat{\mathcal{R}}_{\boldsymbol{X}}(\mathcal{Q}) = \mathbb{E}_{\sigma_{ij}}\left[\sup_{q\in\mathcal{Q}} \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{r} \sigma_{ij} q_j(\boldsymbol{x}_i)\right], \quad \mathcal{R}_n(\mathcal{Q}) = \mathbb{E}_{\boldsymbol{X}}\left[\widehat{\mathcal{R}}_{\boldsymbol{X}}(\mathcal{Q})\right],$$

$$\text{Empirical/Population Gaussian complexities:} \quad \widehat{\mathcal{G}}_{\boldsymbol{X}}(\mathcal{Q}) = \mathbb{E}_{g_{ij}}\left[\sup_{q\in\mathcal{Q}} \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{r} g_{ij} q_j(\boldsymbol{x}_i)\right], \quad \mathcal{G}_n(\mathcal{Q}) = \mathbb{E}_{\boldsymbol{X}}\left[\widehat{\mathcal{G}}_{\boldsymbol{X}}(\mathcal{Q})\right],$$

where we have $q \in \mathcal{Q}$ and $q(\boldsymbol{x}) = [q_1(\boldsymbol{x}), \ldots, q_r(\boldsymbol{x})]^\top$. Note that in vector notation one can also write $\widehat{\mathcal{R}}_{\boldsymbol{X}}(\mathcal{Q}) = \mathbb{E}_{\boldsymbol{\sigma}_i}\left[\sup_{q\in\mathcal{Q}} \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\sigma}_i^\top q(\boldsymbol{x}_i)\right]$ and $\widehat{\mathcal{G}}_{\boldsymbol{X}}(\mathcal{Q}) = \mathbb{E}_{\boldsymbol{g}}\left[\sup_{q\in\mathcal{Q}} \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}_i^\top q(\boldsymbol{x}_i)\right]$, where $\boldsymbol{\sigma}_i$ and $\boldsymbol{g}_i$ are r-dimensional with independent Rademacher/Gaussian variables in each entry. Also recall that worst-case versions $\widetilde{\mathcal{R}}_n, \widetilde{\mathcal{G}}_n$ are obtained by taking supremum over the input space.

B Proofs in Section 3

We first introduce some lemmas used throughout this section, then provide the proofs of our mean results.

B.1 Supporting Lemmas

The following is a seminal contraction lemma due to Talagrand (Talagrand 2006).

Lemma 2 (Talagrand's Contraction inequality) *Let $\boldsymbol{\varepsilon} = (\varepsilon_i)_{i=1}^n$ be i.i.d. random variables with symmetric sign (e.g. Rademacher, standard normal). Let $(\phi_i)_{i=1}^n$ be L-Lipschitz functions and $\mathcal{F}$ be a hypothesis set. We have that*

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^n \varepsilon_i\phi_i(f(\boldsymbol{x}_i))\right] \le L\,\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^n \varepsilon_i f(\boldsymbol{x}_i)\right].$$

As a corollary of this, we can deduce that adjusted empirical Gaussian complexity $n\widehat{\mathcal{G}}_{\boldsymbol{X}}(\mathcal{F})$ is non-decreasing in sample size n.

Corollary 1 *Let $\mathcal{X}$ be a bounded input space and $\mathcal{F}:\mathcal{X}\to\mathbb{R}$ be a hypothesis set. Let $\boldsymbol{X}_m$ be a dataset of size m and $\boldsymbol{X}_n=(\boldsymbol{x}_i)_{i=1}^n$ be a dataset of size n that contains $\boldsymbol{X}_m$. We have that*

$$m\widehat{\mathcal{G}}_{\boldsymbol{X}_m}(\mathcal{F}) \le n\widehat{\mathcal{G}}_{\boldsymbol{X}_n}(\mathcal{F}).$$

We note that, when $\mathcal{F}:\mathcal{X}\to\mathbb{R}^p$ is vector valued and we apply $p\times n$ L-Lipschitz functions ϕ_{ij}, the identical results (Lemmas 2 and Corollary 1) follow from Sudakov-Fernique inequality under Gaussian $\boldsymbol{\varepsilon}\in\mathbb{R}^{n\times p}$ (e.g. Exercise 7.2.13 of (Vershynin 2018)).

This also implies usual (distributional) and worst-case Gaussian complexities are also non-decreasing.

Proof Let $(\phi_i)_{i=1}^n$ be functions that are identity for $i\le m$ and zero for $i>m$. Observe that

$$m\widehat{\mathcal{G}}_{\boldsymbol{X}_m} = \mathbb{E}_{\boldsymbol{\varepsilon}}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^m \varepsilon_i f(\boldsymbol{x}_i)\right] = \mathbb{E}_{\boldsymbol{\varepsilon}}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^n \varepsilon_i\phi_i(f(\boldsymbol{x}_i))\right] \le n\widehat{\mathcal{G}}_{\boldsymbol{X}_n}.$$

■

The following lemma shows that adjusted worst-case Gaussian complexity $\sqrt{n}\widetilde{\mathcal{G}}_{\boldsymbol{X}}(\mathcal{F})$ is essentially non-decreasing in sample size n.

Lemma 3 (Worst-case Gaussian Complexity over Input Space and Sample Size) *For any bounded input space $\mathcal{X}$ and hypothesis set $\mathcal{F}$, we have that*

$$\sup_{1\le m\le n}\sqrt{m}\widetilde{\mathcal{G}}_m(\mathcal{F}) \le \sqrt{2n}\widetilde{\mathcal{G}}_n(\mathcal{F}).$$

Proof First suppose $n/2\le m\le n$. In this case, from Corollary 1, we know that $m\widetilde{\mathcal{G}}_m(\mathcal{F})\le n\widetilde{\mathcal{G}}_n(\mathcal{F}) \implies \sqrt{m}\widetilde{\mathcal{G}}_m(\mathcal{F}) \le \frac{\sqrt{2}m}{\sqrt{n}}\widetilde{\mathcal{G}}_m(\mathcal{F}) \le \sqrt{2n}\widetilde{\mathcal{G}}_n(\mathcal{F})$. What remains is the scenario $m<n/2$. To do this, we will show monotonicity under doubling $\sqrt{m}\widetilde{\mathcal{G}}_m(\mathcal{F})\le\sqrt{2m}\widetilde{\mathcal{G}}_{2m}(\mathcal{F})$. If this holds, then you can double m until a point $n/2\le m\le n$ and apply the first bound.

Consider worst-case dataset for $\widetilde{\mathcal{G}}_m$ defined as

$$\boldsymbol{Y} = \arg\max_{\boldsymbol{X}\in\mathcal{X}^m}\widehat{\mathcal{G}}_{\boldsymbol{X}}(\mathcal{F}).$$

Let $\boldsymbol{Y}'$ be a dataset of size $2m$ that repeats the elements of $\boldsymbol{Y}$ twice so that $\boldsymbol{y}'_{m+i}=\boldsymbol{y}'_i=\boldsymbol{y}_i$. Here, we consider hypothesis set $\mathcal{F}:\mathcal{X}\to\mathbb{R}^p$, and then $f(\boldsymbol{y}_i)=[f_1(\boldsymbol{y}_i),\cdots,f_p(\boldsymbol{y}_i)]^\top$. Also let $\boldsymbol{\varepsilon}\in\mathbb{R}^{m\times p}, \boldsymbol{\varepsilon}'\in\mathbb{R}^{2m\times p}$ where $\boldsymbol{\varepsilon}'_i\sim\mathcal{N}(0,I_p), i\in[2m]$ and $\boldsymbol{\varepsilon}_i=\frac{\boldsymbol{\varepsilon}'_i+\boldsymbol{\varepsilon}'_{m+i}}{\sqrt{2}}\sim\mathcal{N}(0,I_p)$. We have that

$$\begin{aligned}
2m\widetilde{\mathcal{G}}_{2m}(\mathcal{F}) &\ge \mathbb{E}_{\boldsymbol{\varepsilon}}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{2m}\sum_{j=1}^p \varepsilon'_{ij}f_j(\boldsymbol{y}'_i)\right]\\
&\ge \mathbb{E}_{\boldsymbol{\varepsilon}}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sum_{j=1}^p \varepsilon'_{i,j}f_j(\boldsymbol{y}'_i)+\varepsilon'_{(m+i),j}f_j(\boldsymbol{y}'_{m+i})\right]\\
&\ge \sqrt{2}\,\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sum_{j=1}^p \varepsilon_{ij}f_j(\boldsymbol{y}_i)\right]\\
&= \sqrt{2}m\widetilde{\mathcal{G}}_m.
\end{aligned}$$

Dividing both sides by $\sqrt{2m}$, we conclude with the claim $\sqrt{m}\widetilde{\mathcal{G}}_m(\mathcal{F})\le\sqrt{2m}\widetilde{\mathcal{G}}_{2m}(\mathcal{F})$. ■

The following is a model selection argument shows that $\widetilde{\mathcal{G}}(\Phi)$ can be replaced with $\widetilde{\mathcal{G}}(\Phi_{\text{used}})$.

Lemma 4 (Only utilized supernet matters) *Observe that T tasks can choose from up to $|\mathcal{A}|^T$ supernets in total. Let $\mathbf{\Phi}_{all} = (\Phi_i)_{i=1}^H$ with $H \leq |\mathcal{A}|^T$ be the set of unique supernets (since two supernets that choose same number of modules per layer are identical architectures). Suppose the outcome of empirical risk minimization* (M^{2}TL) *obeys $\hat{\phi} \in \Phi_{used} \in \mathbf{\Phi}_{all}$. Let $\hat{K}_\ell$ be the number of (used) modules in Φ_{used}. With probability $1-\delta$, we have that*

$$\mathcal{L}_{\bar{\mathcal{D}}}(\hat{\boldsymbol{f}}) - \widehat{\mathcal{L}}_{\mathcal{S}_{all}}(\hat{\boldsymbol{f}}) \lesssim \widetilde{\mathcal{G}}_N(\mathcal{H}) + \sum_{\ell=1}^{L} \sqrt{\hat{K}_\ell}\widetilde{\mathcal{G}}_{NT}(\Psi_\ell) + \sqrt{\frac{\log|\mathcal{A}|}{N} + \frac{\log(2/\delta)}{NT}}, \tag{5}$$

$$\mathcal{R}_{M^2TL}(\hat{\boldsymbol{f}}) := \mathcal{L}_{\bar{\mathcal{D}}}(\hat{\boldsymbol{f}}) - \mathcal{L}^\star_{\bar{\mathcal{D}}} \lesssim \widetilde{\mathcal{G}}_N(\mathcal{H}) + \sum_{\ell=1}^{L} \sqrt{\hat{K}_\ell}\widetilde{\mathcal{G}}_{NT}(\Psi_\ell) + \sqrt{\frac{\log|\mathcal{A}|}{N} + \frac{\log(2/\delta)}{NT}}. \tag{6}$$

Proof Let $\mathcal{L}_{\Phi'}, \widehat{\mathcal{L}}_{\Phi'}$ be the population and empirical risks we achieve when we run the (M^{2}TL) problem over $\Phi' \in \mathbf{\Phi}_{\text{all}}$ rather than Φ. Additionally, let $K_\ell(\Phi')$ denote the number of modules in the ℓth layer of the architecture Φ'. Given Φ', also define $\mathcal{C}_N(\Phi', \delta)$ to be the excess risk bound one obtains via (9) ((9) in Theorem 5 is obtained without using Lemma 4), that is,

$$\mathcal{C}_N(\Phi', \delta) = \widetilde{\mathcal{G}}_N(\mathcal{H}) + \sum_{\ell=1}^{L} \sqrt{K_\ell(\Phi')}\widetilde{\mathcal{G}}_{NT}(\Psi_\ell) + \sqrt{\frac{\log|\mathcal{A}|}{N} + \frac{\log(2/\delta)}{NT}}.$$

To proceed, applying (9) over $\Phi' \in \mathbf{\Phi}_{\text{all}}$ and union bounding over all $H \leq |\mathcal{A}|^T$, with probability at least $1-\delta$, we find that, all $\Phi' \in \mathbf{\Phi}_{\text{all}}$ obeys

$$|\widehat{\mathcal{L}}_{\Phi'}(\hat{\boldsymbol{f}}) - \mathcal{L}_{\Phi'}(\hat{\boldsymbol{f}})| \lesssim \mathcal{C}_N(\Phi', \delta/H).$$

Fortunately, $\mathcal{C}_N(\Phi', \delta/H) \lesssim \mathcal{C}_N(\Phi', \delta)$ since the latter already includes a $\sqrt{\frac{\log|\mathcal{A}|}{N}}$ term. Using this union bound, optimality of $\hat{\phi} \in \Phi_{\text{used}}$ (and that of the associated $\hat{\boldsymbol{f}} \in \mathcal{F}_{\text{used}}$), and using $\widehat{\mathcal{L}}_{\Phi_{\text{used}}}(\hat{\boldsymbol{f}}) = \widehat{\mathcal{L}}_{\Phi}(\hat{\boldsymbol{f}}) = \widehat{\mathcal{L}}_{\mathcal{S}_{\text{all}}}(\hat{\boldsymbol{f}})$, we find that

$$\mathcal{L}_{\Phi_{\text{used}}}(\hat{\boldsymbol{f}}) \leq \widehat{\mathcal{L}}_{\Phi_{\text{used}}}(\hat{\boldsymbol{f}}) + \mathcal{O}\left(\mathcal{C}_N(\Phi_{\text{used}}, \delta)\right) \tag{7}$$

$$\leq \widehat{\mathcal{L}}_{\mathcal{S}_{\text{all}}}(\hat{\boldsymbol{f}}) + \mathcal{O}\left(\mathcal{C}_N(\Phi_{\text{used}}, \delta)\right). \tag{8}$$

The last line establishes Inequality (5). To conclude with the second inequality, we control the excess risk error by observing test risk upper bounds the training risk. Namely, let $\boldsymbol{f}_\star \in \mathcal{F}$ be the population minima. First, with $1-\delta$ probability, for this singleton hypothesis, we have that

$$|\mathcal{L}_{\bar{\mathcal{D}}}(\boldsymbol{f}_\star) - \widehat{\mathcal{L}}_{\mathcal{S}_{\text{all}}}(\boldsymbol{f}_\star)| \leq \sqrt{\frac{\log(2/\delta)}{NT}}.$$

Second, we can write

$$\widehat{\mathcal{L}}_{\mathcal{S}_{\text{all}}}(\hat{\boldsymbol{f}}) \leq \widehat{\mathcal{L}}_{\mathcal{S}_{\text{all}}}(\boldsymbol{f}_\star) \leq \mathcal{L}_{\bar{\mathcal{D}}}(\boldsymbol{f}_\star) + \sqrt{\frac{\log(2/\delta)}{NT}}.$$

Combining this with (8), we establish the guarantee against the ground-truth optima $\boldsymbol{f}_\star$

$$\mathcal{L}_{\Phi_{\text{used}}}(\hat{\boldsymbol{f}}) - \left[\mathcal{L}_{\bar{\mathcal{D}}}(\boldsymbol{f}_\star) + \sqrt{\frac{\log(2/\delta)}{NT}}\right] \leq \mathcal{L}_{\Phi_{\text{used}}}(\hat{\boldsymbol{f}}) - \widehat{\mathcal{L}}_{\mathcal{S}_{\text{all}}}(\hat{\boldsymbol{f}}) \leq \mathcal{O}\left(\mathcal{C}_N(\Phi_{\text{used}}, \delta)\right),$$

which establishes the claim (6) after subsuming $\sqrt{\frac{\log(2/\delta)}{NT}}$ within $\mathcal{C}_N(\Phi_{\text{used}}, \delta)$. ■

B.2 Proof of Theorem 1

Let us define the covering number of a hypothesis as well as natural data-dependent Euclidean distance for ease of reference in the subsequent discussion (see (Wainwright 2019)).

Definition 3 (Covering number) *Let $\mathcal{Q} : \mathcal{X} \to \mathbb{R}^r$ be a family of functions. Given $q, q' \in \mathcal{Q}$, and a distance metric $d(q, q') \geq 0$, an ε-cover of set $\mathcal{Q}$ with respect to $d(\cdot,\cdot)$ is a set $\{q^1, q^2, \ldots, q^N\} \subset \mathcal{Q}$ such that for any $q \in \mathcal{Q}$, there exists some $i \in [N]$ such that $d(q, q^i) \leq \varepsilon$. The ε-covering number $\mathcal{N}(\varepsilon; \mathcal{Q}, d)$ is defined to be the cardinality of the smallest ε-cover.*

Definition 4 (Data-dependent distance metric ρ) *Let $\mathcal{Q} : \mathcal{X} \to \mathbb{R}^r$ be a family of functions. Given $q, q' \in \mathcal{Q}$ and an input dataset $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ with $\boldsymbol{x}_i \in \mathcal{X}$, we define the dataset-dependent Euclidean distance by $\rho_{\boldsymbol{X}}(q, q') := \sqrt{\frac{1}{n}\sum_{i\in[n], j\in[r]}(q_j(\boldsymbol{x}_i) - q'_j(\boldsymbol{x}_i))^2} = \sqrt{\frac{1}{n}\sum_{i\in[n]}\|q(\boldsymbol{x}_i) - q'(\boldsymbol{x}_i)\|^2}$, where $q(\boldsymbol{x}) = [q_1(\boldsymbol{x}), \ldots, q_r(\boldsymbol{x})]^\top$.*

Now we are ready to prove our main theorem which incorporates additional dependencies that were omitted from the original statement.

Theorem 5 (Theorem 1 restated) *Suppose Assumptions 1&2 hold. Let $\hat{\boldsymbol{f}}$ be the empirical solution of* (M^2TL). *Let $D_{\mathcal{X}} = \sup_{\boldsymbol{x}\in\mathcal{X},h\in\mathcal{H},\boldsymbol{\phi}\in\Phi,\alpha\in\mathcal{A}} |h\circ\boldsymbol{\phi}_\alpha(\boldsymbol{x})| < \infty$, and set $\Gamma^\dagger = \sum_{\ell=0}^{L}\Gamma^\ell$. Then, with probability at least $1-\delta$, the excess test risk in* (2) *obeys*

$$\mathcal{R}_{M^2TL}(\hat{\boldsymbol{f}}) \leq 768\Gamma\left(\frac{D_{\mathcal{X}}}{NT} + D_{\mathcal{X}}\sqrt{\frac{\log|\mathcal{A}|}{N}} + \Gamma^\dagger \log NT\left(\widetilde{\mathcal{G}}_N(\mathcal{H}) + \sum_{\ell=1}^{L}\sqrt{K_\ell}\widetilde{\mathcal{G}}_{NT}(\Psi_\ell)\right)\right) + 2\sqrt{\frac{\log\frac{2}{\delta}}{NT}}. \tag{9}$$

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 $\mathcal{X}_{\Psi_1}=\mathcal{X}$. The above is our general results, which we do not focus on the actual modules used in $\hat{\boldsymbol{f}}$. Now let $\hat{K}$ be the number of modules utilized by $\hat{\boldsymbol{f}}$, then with probability at least $1-\delta$, we can obtain

$$\mathcal{R}_{M^2TL}(\hat{\boldsymbol{f}}) \lesssim \widetilde{\mathcal{G}}_N(\mathcal{H}) + \sum_{\ell=1}^{L}\sqrt{\hat{K}_\ell}\widetilde{\mathcal{G}}_{NT}(\Psi_\ell) + \sqrt{\frac{\log|\mathcal{A}|}{N} + \frac{\log(2/\delta)}{NT}}. \tag{10}$$

Here, $\lesssim$ suppresses dependencies on $\log NT$, $\Gamma^\dagger$ and $D_{\mathcal{X}}$.

Remark. While this result is stated with worst-case Gaussian complexity, the line (20) states our result in terms of empirical Gaussian complexity which is always a lower bound and is in terms of the training dataset. However, (20) is more convoluted and involves worst-case hypothesis being applied to the training data. The latter arises from the fact that it is difficult to track the evolution of features across arbitrary pathways and hierarchical layers.

Proof To start with, let us recap some notations. Assume we have T tasks each with N training samples i.i.d. drawn from $(\mathcal{D}_t)_{t=1}^T$ respectively, and let $\bar{\mathcal{D}} = \{\mathcal{D}_t\}_{t=1}^T$. Denote the training dataset and inputs of t_{th} task by $\mathcal{S}_t = \{(\boldsymbol{x}_{ti}, y_{ti})\}_{i=1}^N$ and $\boldsymbol{X}_t = \{\boldsymbol{x}_{ti}\}_{i=1}^N$, and define the union by $\mathcal{S}_{\text{all}} = \bigcup_{t=1}^T \mathcal{S}_t$ and $\boldsymbol{X} = \bigcup_{t=1}^T \boldsymbol{X}_t$. Let $\boldsymbol{h} = [h_1,\ldots,h_T]\in\mathcal{H}^T$, $\boldsymbol{\alpha} = [\alpha_1,\ldots,\alpha_T]\in\mathcal{A}^T$, $\boldsymbol{\psi}_\ell = [\psi_\ell^1,\ldots,\psi_\ell^{K_\ell}]\in\Psi_\ell^{K_\ell}, \ell\in[L]$, and $\boldsymbol{\phi} = [\boldsymbol{\psi}_1,\ldots,\boldsymbol{\psi}_L]\in\Phi = \Psi_1^{K_1}\times\ldots\Psi_L^{K_L}$. $\hat{\boldsymbol{f}} := (\hat{\boldsymbol{h}},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\phi}})$ is the empirical solution of (M^2TL) and $\boldsymbol{f}^\star := (\boldsymbol{h}^\star,\boldsymbol{\alpha}^\star,\boldsymbol{\phi}^\star)$ is the population solution of (M^2TL) when each task has infinite i.i.d training samples ($N=\infty$). Let $\mathcal{F}$ denote the hypothesis set of functions $\boldsymbol{f}$. Since multitask problem is task-aware, that is, the task identification of each data is given during training and test, we can rewrite samples in $\mathcal{S}_t$ as $\{(\boldsymbol{x}_i, y_i, t_i\equiv t)\}_{i=1+(t-1)N}^{tN}$ and the overall multitask training dataset can be seen as $\mathcal{S}_{\text{all}} = \{(\boldsymbol{x}_i,y_i,t_i)\}_{i=1}^{NT}$. Letting $\boldsymbol{f}(\boldsymbol{x},t) = f_t(\boldsymbol{x}) = h_t\circ\boldsymbol{\phi}_{\alpha_t}(\boldsymbol{x})$, the loss functions can be rewritten by $\widehat{\mathcal{L}}_{\mathcal{S}_{\text{all}}}(\boldsymbol{f}) = \frac{1}{NT}\sum_{i=1}^{NT}\ell(\boldsymbol{f}(\boldsymbol{x}_i,t_i),y_i)$ and $\mathcal{L}_{\bar{\mathcal{D}}}(\boldsymbol{f}) = \mathbb{E}[\widehat{\mathcal{L}}_{\mathcal{S}_{\text{all}}}(\boldsymbol{f})]$. In the following, we drop the subscript $\bar{\mathcal{D}}$ and $\mathcal{S}_{\text{all}}$ for cleaner notations. Then we have

$$\underbrace{\mathcal{L}(\hat{\boldsymbol{f}}) - \mathcal{L}(\boldsymbol{f}^\star)}_{\mathcal{R}_{\mathrm{M}^2\mathrm{TL}}(\hat{\boldsymbol{f}})} = \underbrace{\mathcal{L}(\hat{\boldsymbol{f}}) - \widehat{\mathcal{L}}(\hat{\boldsymbol{f}})}_{a} + \underbrace{\widehat{\mathcal{L}}(\hat{\boldsymbol{f}}) - \widehat{\mathcal{L}}(\boldsymbol{f}^\star)}_{b} + \underbrace{\widehat{\mathcal{L}}(\boldsymbol{f}^\star) - \mathcal{L}(\boldsymbol{f}^\star)}_{c}, \tag{11}$$

where $b\leq 0$ because of the fact that $\hat{\boldsymbol{f}}$ is the empirical risk minimizer of $\widehat{\mathcal{L}}(\boldsymbol{f})$. Then, following the proof of Theorem 3.3 of (Mohri, Rostamizadeh, and Talwalkar 2018), we make two observations: 1) Their Equation (3.8) in the proof still holds when we restrict N i.i.d samples in each task instead of NT i.i.d. samples over distribution $\bar{\mathcal{D}}$. Therefore, the symmetrization augment does not change, and this theorem holds under our setting. 2) The identical results hold for any function set mapping to $[-1,1]$. In this work, based on these two observations, following Assumption 2 and Theorem 11.3 in (Mohri, Rostamizadeh, and Talwalkar 2018), we have that with probability at least $1-\delta/2$, $a,c\leq 2\Gamma\mathcal{R}_{NT}(\mathcal{F}) + \sqrt{\frac{\log(2/\delta)}{2NT}}$. Therefore, we can conclude that with probability at least $1-\delta$,

$$\mathcal{R}_{\mathrm{M}^2\mathrm{TL}}(\hat{\boldsymbol{f}}) \leq 4\Gamma\mathcal{R}_{NT}(\mathcal{F}) + \sqrt{\frac{2\log\frac{2}{\delta}}{NT}}, \tag{12}$$

$$\text{and similarly,}\quad \mathcal{R}_{\mathrm{M}^2\mathrm{TL}}(\hat{\boldsymbol{f}}) \leq 4\Gamma\widehat{\mathcal{R}}_{\boldsymbol{X}}(\mathcal{F}) + 3\sqrt{\frac{2\log\frac{4}{\delta}}{NT}}, \tag{13}$$

where $\widehat{\mathcal{R}}_{\boldsymbol{X}}(\mathcal{F})$ is the empirical complexity with respect to the inputs $\boldsymbol{X}$ and $\mathcal{R}_{NT}(\mathcal{F})$ is the Rademacher complexity with respect to the sample size NT. Exercise 5.5 in (Wainwright 2019) shows that Rademacher complexity can be bounded in terms of Gaussian complexity, that is $\widehat{\mathcal{R}}_{\boldsymbol{X}}(\mathcal{F})\leq\sqrt{\frac{\pi}{2}}\widehat{\mathcal{G}}_{\boldsymbol{X}}(\mathcal{F})$ and $\mathcal{R}_{NT}(\mathcal{F})\leq\sqrt{\frac{\pi}{2}}\mathcal{G}_{NT}(\mathcal{F})$. Combining them together, we have that with probability at least $1-\delta$,

$$\mathcal{R}_{\mathrm{M}^2\mathrm{TL}}(\hat{\boldsymbol{f}}) \leq 6\Gamma\mathcal{G}_{NT}(\mathcal{F}) + 2\sqrt{\frac{\log\frac{2}{\delta}}{NT}},\ \text{ and }\ \mathcal{R}_{\mathrm{M}^2\mathrm{TL}} \leq 6\Gamma\widehat{\mathcal{G}}_{\boldsymbol{X}}(\mathcal{F}) + 6\sqrt{\frac{\log\frac{4}{\delta}}{NT}}. \tag{14}$$

In what follows, we will move to Gaussian complexity instead. Now, it remains to decompose the Gaussian complexity of a set of composition functions $\mathcal{F}$ into basic function sets $\mathcal{H}$, $\mathcal{A}$ and $\{\Psi_\ell\}_{\ell=1}^L$. We will first bound the empirical Gaussian complexity with respect to any training inputs $\boldsymbol{X}$, which turns to be worst-case Gaussian complexity defined in Definition 1. Then, population complexity is simply bounded by the worst-case Gaussian complexity.

Inspired by (Tripuraneni, Jordan, and Jin 2020), we use the Dudley's entropy integral bound showed in (Wainwright 2019) (Theorem 5.22) to derive the upper bound. Define $Z_{\boldsymbol{f}} := \frac{1}{\sqrt{NT}} \sum_{i=1}^{NT} g_i \boldsymbol{f}(\boldsymbol{x}_i, t_i)$ where $\boldsymbol{f} \in \mathcal{F}$ and g_is are standard random Gaussian variables. Sine $Z_{\boldsymbol{f}}$ has zero-mean, we have $\widehat{\mathcal{G}}_{\boldsymbol{X}}(\mathcal{F}) = \frac{1}{\sqrt{NT}} \mathbb{E}_{\boldsymbol{g}}[\sup_{\boldsymbol{f}\in\mathcal{F}} Z_{\boldsymbol{f}}] \leq \frac{1}{\sqrt{NT}} \mathbb{E}_{\boldsymbol{g}}[\sup_{\boldsymbol{f},\boldsymbol{f}'\in\mathcal{F}}(Z_{\boldsymbol{f}} - Z_{\boldsymbol{f}'})]$. Following Definition 4, let $\rho_{\boldsymbol{X}}(\boldsymbol{f}, \boldsymbol{f}') = \sqrt{\frac{1}{NT}\sum_{i=1}^{NT}(\boldsymbol{f}(\boldsymbol{x}_i, t_i) - \boldsymbol{f}'(\boldsymbol{x}_i, t_i))^2}$. Define $D_{\boldsymbol{X}} = \sup_{\boldsymbol{f},\boldsymbol{f}'\in\mathcal{F}} \rho_{\boldsymbol{X}}(\boldsymbol{f}, \boldsymbol{f}') \leq 2D_{\mathcal{X}}$. Following Theorem 5.22 in (Wainwright 2019), we have that for any $\varepsilon \in [0, D_{\boldsymbol{X}}]$,

$$\mathbb{E}_{\boldsymbol{g}}\left[\sup_{\boldsymbol{f},\boldsymbol{f}'\in\mathcal{F}}(Z_{\boldsymbol{f}} - Z_{\boldsymbol{f}'})\right] \leq 2\,\mathbb{E}_{\boldsymbol{g}}\left[\sup_{\substack{\boldsymbol{f},\boldsymbol{f}'\in\mathcal{F}\\ \rho_{\boldsymbol{X}}(\boldsymbol{f},\boldsymbol{f}')\leq\varepsilon}}(Z_{\boldsymbol{f}} - Z_{\boldsymbol{f}'})\right] + 32\int_{\varepsilon/4}^{D_{\boldsymbol{X}}} \sqrt{\log \mathcal{N}(u; \mathcal{F}, \rho_{\boldsymbol{X}})}du, \tag{15}$$

where $\mathcal{N}(u; \mathcal{F}, \rho_{\boldsymbol{X}})$ is the u-covering number of function set $\mathcal{F}$ with respect to metric $\rho_{\boldsymbol{X}}(\cdot, \cdot)$ following Definition 3.

The first term in the right hand side above is easy to bound. As shown in proof of Theorem 7 in (Tripuraneni, Jordan, and Jin 2020), we have $\mathbb{E}_{\boldsymbol{g}}[\sup_{\rho_{\boldsymbol{X}}(\boldsymbol{f},\boldsymbol{f}')\leq\varepsilon}(Z_{\boldsymbol{f}} - Z_{\boldsymbol{f}'})] \leq \mathbb{E}_{\boldsymbol{g}}[\sup_{\|\boldsymbol{v}\|_2\leq\varepsilon} \boldsymbol{g}^\top \boldsymbol{v}] \leq \mathbb{E}_{\boldsymbol{g}}[\sup_{\|\boldsymbol{v}\|_2\leq\varepsilon} \|\boldsymbol{g}\|_2\|\boldsymbol{v}\|_2] = \sqrt{NT}\varepsilon$. Next, it remains to bound the integral term. Here, since $\boldsymbol{f} \in \mathcal{F}$ is a sophisticated function composed with $\psi_\ell^k \in \Psi_\ell, \alpha_t \in \mathcal{A}$ and $h_t \in \mathcal{H}$, its covering number is not well-defined. Hence, instead, we relate the cover of $\mathcal{F}$ to the covers of basic function sets, Ψ_ℓ, $\mathcal{A}$ and $\mathcal{H}$. To this end, we need to decompose the distance metric $\rho_{\boldsymbol{X}}$ into distances over basic sets. Since $\mathcal{A}$ is a discrete set with cardinality $|\mathcal{A}|$. Let $\mathcal{F}^{\boldsymbol{\alpha}} \subset \mathcal{F}$ be the function set given pathways of all tasks $\boldsymbol{\alpha}$. Then we have $\log \mathcal{N}(u; \mathcal{F}, \rho_{\boldsymbol{X}}) \leq T \log |\mathcal{A}| + \max_{\boldsymbol{\alpha}\in\mathcal{A}^T} \log \mathcal{N}(u; \mathcal{F}^{\boldsymbol{\alpha}}, \rho_{\boldsymbol{X}})$. For any $\boldsymbol{f}, \boldsymbol{f}' \in \mathcal{F}^{\boldsymbol{\alpha}}$, we have

$$\begin{aligned}
\rho_{\boldsymbol{X}}(\boldsymbol{f}, \boldsymbol{f}') &= \sqrt{\frac{1}{NT}\sum_{i=1}^{NT}(\boldsymbol{f}(\boldsymbol{x}_i, t_i) - \boldsymbol{f}'(\boldsymbol{x}_i, t_i))^2} = \sqrt{\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\left(h_t \circ \boldsymbol{\phi}_{\alpha_t}(\boldsymbol{x}_{ti}) - h'_t \circ \boldsymbol{\phi}'_{\alpha_t}(\boldsymbol{x}_{ti})\right)^2} \\
&\leq \underbrace{\sqrt{\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\left(h_t \circ \boldsymbol{\phi}_{\alpha_t}(\boldsymbol{x}_{ti}) - h'_t \circ \boldsymbol{\phi}_{\alpha_t}(\boldsymbol{x}_{ti})\right)^2}}_{d} + \underbrace{\sqrt{\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\left(h'_t \circ \boldsymbol{\phi}_{\alpha_t}(\boldsymbol{x}_{ti}) - h'_t \circ \boldsymbol{\phi}'_{\alpha_t}(\boldsymbol{x}_{ti})\right)^2}}_{e}.
\end{aligned}$$

To proceed, let us introduce some notations. For any function ϕ with inputs $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, define output set w.r.t. the inputs $\boldsymbol{X}$ by $\phi(\boldsymbol{X}) = \{\phi(\boldsymbol{x}_1), \ldots, \phi(\boldsymbol{x}_n)\}$. In the multipath setting, since different tasks have different pathways, different modules are chosen by different set of tasks. Given $\boldsymbol{\alpha}$, the task clustering methods in different layers are determined. Let $\mathcal{I}_\ell^k$ denote the union of task IDs who select (ℓ, k)'th module, and $\mathcal{I}_\ell^k, \ell \in [K_\ell]$ are disjoint sets satisfying $\bigcup_{k=1}^{K_\ell} \mathcal{I}_\ell^k = [T]$. What's more, let $\boldsymbol{Z}_\ell^k$ denote the latent inputs of (ℓ, k)'th module, where we have

$$\boldsymbol{Z}_\ell^k = \bigcup_{t\in\mathcal{I}_\ell^k} \psi_{\ell-1}^{\alpha_t} \cdots \circ \psi_1^{\alpha_t}(\boldsymbol{X}_t), \quad 1 < \ell \leq L, \tag{16}$$

and $\boldsymbol{Z}_1^k = \bigcup_{t\in\mathcal{I}_1^k} \boldsymbol{X}_t$. In short, (ℓ, k)'th module (whose function is ψ_ℓ^k) is utilized by tasks $\mathcal{I}_\ell^k$ with latent inputs $\boldsymbol{Z}_\ell^k$. The inputs of heads are

$$\boldsymbol{Z}_{\mathcal{H}}^t = \psi_L^{\alpha_t} \cdots \circ \psi_1^{\alpha_t}(\boldsymbol{X}_t) = \boldsymbol{\phi}_{\alpha_t}(\boldsymbol{X}_t), \quad \forall\, t \in [T].$$

Then we can obtain that

$$(d) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{1}{N}\sum_{i=1}^{N}\left(h_t \circ \boldsymbol{\phi}_{\alpha_t}(\boldsymbol{x}_{ti}) - h'_t \circ \boldsymbol{\phi}_{\alpha_t}(\boldsymbol{x}_{ti})\right)^2} \leq \sqrt{\frac{1}{T}\sum_{t=1}^{T}\rho^2_{\boldsymbol{Z}^t_{\mathcal{H}}}(h_t, h'_t)},$$

$$\begin{aligned}(e) &\leq \Gamma\sqrt{\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\left\|\boldsymbol{\phi}_{\alpha_t}(\boldsymbol{x}_{ti}) - \boldsymbol{\phi}'_{\alpha_t}(\boldsymbol{x}_{ti})\right\|^2} \\ &\leq \Gamma\sum_{\ell=1}^{L}\Gamma^{L-\ell}\sqrt{\frac{1}{K_\ell}\sum_{k=1}^{K_\ell}\frac{1}{|\boldsymbol{Z}^k_\ell|}\sum_{\boldsymbol{z}_i\in\boldsymbol{Z}^k_\ell}\left\|\psi^k_\ell(\boldsymbol{z}_i) - \psi'^k_\ell(\boldsymbol{z}_i)\right\|^2} \\ &\leq \sum_{\ell=1}^{L}\Gamma^{L-\ell+1}\sqrt{\frac{1}{K_\ell}\sum_{k=1}^{K_\ell}\rho^2_{\boldsymbol{Z}^k_\ell}(\psi^k_\ell, \psi'^k_\ell)}.\end{aligned}$$

Here $|\boldsymbol{Z}^k_\ell| = |\mathcal{I}^k_\ell|N$ is the number of samples used in training (ℓ, k)'th module. The result follows the fact that all functions $h \in \mathcal{H}, \psi^k_\ell \in \Psi_\ell, \ell \in [L], k \in [K_\ell]$ are Γ-Lipschitz, and it also applies an implicit chain rule for composition Lipschitz functions. Now, we decompose distance (d) into distances of each head function $h_t, t \in [T]$, with inputs $\boldsymbol{Z}^t_{\mathcal{H}}$, and decompose distance (e), which captures the distance of composition functions $\boldsymbol{\phi}$ and $\boldsymbol{\phi}'$, into distances of module functions $\psi^k_\ell, \psi'^k_\ell, \ell \in [L], k \in [K_\ell]$, w.r.t. inputs of ψ^k_ℓ, $\boldsymbol{Z}^k_\ell$. Combining them together and assuming $\rho_{\boldsymbol{Z}^t_{\mathcal{H}}}(h_t, h'_t) \leq \varepsilon'$ and $\rho_{\boldsymbol{Z}^k_\ell}(\psi^k_\ell, \psi'^k_\ell) \leq \varepsilon'$ for all $t \in [T], \ell \in [L]$ and $k \in [K^\ell]$, we can obtain

$$\rho_{\boldsymbol{X}}(\boldsymbol{f}, \boldsymbol{f}') \leq \sqrt{\frac{1}{T}\sum_{t=1}^{T}\rho^2_{\boldsymbol{Z}^t_{\mathcal{H}}}(h_t, h'_t)} + \sum_{\ell=1}^{L}\Gamma^{L-\ell+1}\sqrt{\frac{1}{K_\ell}\sum_{k=1}^{K_\ell}\rho^2_{\boldsymbol{Z}^k_\ell}(\psi^k_\ell, \psi'^k_\ell)} \leq \left(1 + \sum_{\ell=1}^{L}\Gamma^{L-\ell+1}\right)\varepsilon' := \Gamma^\dagger\varepsilon'.$$

It shows that given pathway assignments $\boldsymbol{\alpha}$ and inputs $\boldsymbol{X}$, ε'-covers of all heads and modules result in $(\Gamma^\dagger\varepsilon)$-cover of $\mathcal{F}^{\boldsymbol{\alpha}}$. Recalling that $\log\mathcal{N}(u; \mathcal{F}, \rho_{\boldsymbol{X}}) \leq T\log|\mathcal{A}| + \max_{\boldsymbol{\alpha}\in\mathcal{A}^T}\log\mathcal{N}(u; \mathcal{F}^{\boldsymbol{\alpha}}, \rho_{\boldsymbol{X}})$, we have

$$\log\mathcal{N}\left(\Gamma^\dagger\varepsilon'; \mathcal{F}, \rho_{\boldsymbol{X}}\right) \leq T\log|\mathcal{A}| + \max_{\boldsymbol{\alpha}\in\mathcal{A}^T}\log\mathcal{N}\left(\Gamma^\dagger\varepsilon'; \mathcal{F}^{\boldsymbol{\alpha}}, \rho_{\boldsymbol{X}}\right) \tag{17}$$

$$\leq T\log|\mathcal{A}| + \max_{\boldsymbol{\alpha}\in\mathcal{A}^T}\left(\sum_{t=1}^{T}\log\mathcal{N}\left(\varepsilon'; \mathcal{H}, \rho_{\boldsymbol{Z}^t_{\mathcal{H}}}\right) + \sum_{\ell=1}^{L}\sum_{k=1}^{K_\ell}\log\mathcal{N}\left(\varepsilon'; \Psi_\ell, \rho_{\boldsymbol{Z}^k_\ell}\right)\right). \tag{18}$$

Till now, we have decomposed the covering number of $\mathcal{F}^{\boldsymbol{\alpha}}$ into product of covering numbers of all basic function sets $\mathcal{H}, \Psi_\ell, \ell \in [L]$. Next, following (Tripuraneni, Jordan, and Jin 2020), and the Sudakov minoration theorem (Theorem 5.30) and Lemma 5.5 in (Wainwright 2019), and recalling Definition 1, we have that for any $\varepsilon' > 0$,

$$\max_{\boldsymbol{\alpha}\in\mathcal{A}^T}\sum_{t=1}^{T}\log\mathcal{N}\left(\varepsilon'; \mathcal{H}, \rho_{\boldsymbol{Z}^t_{\mathcal{H}}}\right) \leq \max_{\boldsymbol{\alpha}\in\mathcal{A}^T}\sum_{t=1}^{T}\left(\frac{2\sqrt{N}}{\varepsilon'}\widehat{\mathcal{G}}_{\boldsymbol{Z}^t_{\mathcal{H}}}(\mathcal{H})\right)^2 \leq T\left(\frac{2\sqrt{N}}{\varepsilon'}\widetilde{\mathcal{G}}^{\mathcal{X}_{\mathcal{H}}}_N(\mathcal{H})\right)^2,$$

$$\begin{aligned}\max_{\boldsymbol{\alpha}\in\mathcal{A}^T}\sum_{k=1}^{K_\ell}\log\mathcal{N}\left(\varepsilon'; \Psi_\ell, \rho_{\boldsymbol{Z}^k_\ell}\right) &\leq \max_{\boldsymbol{\alpha}\in\mathcal{A}^T}\sum_{k=1}^{K_\ell}\left(\frac{2\sqrt{|\boldsymbol{Z}^k_\ell|}}{\varepsilon'}\widehat{\mathcal{G}}_{\boldsymbol{Z}^k_\ell}(\Psi_\ell)\right)^2 \leq \max_{\boldsymbol{\alpha}\in\mathcal{A}^T}\sum_{k=1}^{K_\ell}\left(\frac{2\sqrt{|\boldsymbol{Z}^k_\ell|}}{\varepsilon'}\widetilde{\mathcal{G}}^{\mathcal{X}_{\Psi_\ell}}_{|\boldsymbol{Z}^k_\ell|}(\Psi_\ell)\right)^2 \\ &\leq K_\ell\left(\frac{2\sqrt{2NT}}{\varepsilon'}\widetilde{\mathcal{G}}^{\mathcal{X}_{\Psi_\ell}}_{NT}(\Psi_\ell)\right)^2,\end{aligned}$$

where 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 $\mathcal{X}_{\Psi_1} = \mathcal{X}$. The last inequality is drawn from Lemma 3, which shows $\sqrt{|\boldsymbol{Z}^k_\ell|}\widetilde{\mathcal{G}}^{\mathcal{X}_{\Psi_\ell}}_{|\boldsymbol{Z}^k_\ell|}(\Psi_\ell) \leq \sqrt{2NT}\mathcal{G}^{\mathcal{X}_{\Psi_\ell}}_{NT}(\Psi_\ell)$. Since Definition 1 eliminates the input($\boldsymbol{X}$)-dependency, the inequalities hold for any valid inputs $\boldsymbol{X}$. In what follows, we drop the superscripts from the worst-case Gaussian complexities for cleaner exposition as they are clear from context. Then, setting $\varepsilon' = \frac{u}{\Gamma^\dagger}$, applying triangle inequality, we can obtain that for any $\boldsymbol{X}$,

$$\sqrt{\log\mathcal{N}(u; \mathcal{F}, \rho_{\boldsymbol{X}})} \leq \sqrt{T\log|\mathcal{A}|} + \frac{2\Gamma^\dagger\sqrt{NT}}{u}\widetilde{\mathcal{G}}_N(\mathcal{H}) + \sum_{\ell=1}^{L}\frac{2\Gamma^\dagger\sqrt{2K_\ell NT}}{u}\widetilde{\mathcal{G}}_{NT}(\Psi_\ell). \tag{19}$$

Now it is time to combine everything together! Recall (14), (15) and (19). Since, $D_{\boldsymbol{X}} \leq 2D_{\mathcal{X}}$ for any inputs $\boldsymbol{X}$, choosing $\varepsilon = \frac{8D_{\mathcal{X}}}{NT}$, we can obtain that with probability at least $1-\delta$,

$$\begin{aligned}
\mathcal{R}_{\text{M}^2\text{TL}}(\hat{\boldsymbol{f}}) &\leq 6\Gamma\mathcal{G}_{NT}(\mathcal{F}) + 2\sqrt{\frac{\log\frac{2}{\delta}}{NT}} \\
&\leq 12\Gamma\left(\varepsilon + 32D_{\mathcal{X}}\sqrt{\frac{\log|\mathcal{A}|}{N}} + 32\Gamma^{\dagger}\left(\widetilde{\mathcal{G}}_N(\mathcal{H}) + \sum_{\ell=1}^{L}\sqrt{2K_\ell}\widetilde{\mathcal{G}}_{NT}(\Psi_\ell)\right)\int_{\varepsilon/4}^{2D_{\mathcal{X}}}\frac{1}{u}du\right) + 2\sqrt{\frac{\log\frac{2}{\delta}}{NT}} \\
&\leq 768\Gamma\left(\frac{D_{\mathcal{X}}}{NT} + D_{\mathcal{X}}\sqrt{\frac{\log|\mathcal{A}|}{N}} + \Gamma^{\dagger}\log NT\left(\widetilde{\mathcal{G}}_N(\mathcal{H}) + \sum_{\ell=1}^{L}\sqrt{K_\ell}\widetilde{\mathcal{G}}_{NT}(\Psi_\ell)\right)\right) + 2\sqrt{\frac{\log\frac{2}{\delta}}{NT}}.
\end{aligned}$$

Till now, we have obtained the result for general $\hat{\boldsymbol{f}}$. Finally, consider the case that $\hat{\boldsymbol{f}}$ might not utilize all the modules in the supernet. Let $\hat{K}_\ell \leq K_\ell$ be the number of modules used by the empirical solution $\hat{\boldsymbol{f}}$. Applying Lemma 4, we can now replace Φ with Φ_{used} which replaces K_ℓ with $\hat{K}_\ell$ for $\ell \in [L]$, which concludes our final result. ■

• **Developing an input-dependent bound.** In Theorem 1, we present the bound of Multipath MTL problem based on the worst-case Gaussian complexity. However, as shown in Definition 1, it computes the complexity of a function set by searching for the worst-case latent inputs, which ignores the data distribution and how the data collected as tasks. In the following argument, we present an input-based guarantee that bounds the excess risk of Multipath MTL problem tightly. To begin with, recall that $\boldsymbol{X} = \{\boldsymbol{X}_t\}_{t=1}^T$ and $\boldsymbol{X}_t = \{\boldsymbol{x}_{ti}\}_{i=1}^N$ denote the actual raw feature sets. Given inputs in T tasks, we can define the *empirical* worst-case Gaussian complexities of $\mathcal{H}$ and $\Psi_\ell, \ell \in [L]$ as follows.

$$C_{\boldsymbol{X}}^{\mathcal{H}} = \max_{t\in[T]}\sup_{\boldsymbol{Z}\in\mathcal{Z}_t}\widehat{\mathcal{G}}_{\boldsymbol{Z}}(\mathcal{H}), \quad \text{where } \mathcal{Z}_t = \Psi_L \circ \ldots \Psi_1(\boldsymbol{X}_t),$$

$$C_{\boldsymbol{X}}^{\Psi_\ell} = \max_{\mathcal{I}\subset[T]}\sup_{\boldsymbol{Z}\in\mathcal{Z}_{\mathcal{I}}}\sqrt{\frac{|\mathcal{I}|}{T}}\widehat{\mathcal{G}}_{\boldsymbol{Z}}(\Psi_\ell), \quad \text{where } \mathcal{Z}_{\mathcal{I}} = \bigcup_{t\in\mathcal{I}}\Psi_{\ell-1}\circ\ldots\Psi_1(\boldsymbol{X}_t),$$

where $\widehat{\mathcal{G}}_{\boldsymbol{Z}}(\mathcal{H})$ and $\widehat{\mathcal{G}}_{\boldsymbol{Z}}(\Psi_\ell)$ are empirical Gaussian complexities and input spaces of $\mathcal{H}$ and Ψ_ℓ are corresponding to the raw input $\boldsymbol{X}$. Then, such statement provide another method to bound (18). That is, we have for any $\varepsilon' > 0$,

$$\begin{aligned}
\max_{\boldsymbol{\alpha}\in\mathcal{A}^T}\sum_{t=1}^{T}\log\mathcal{N}\left(\varepsilon';\mathcal{H},\rho_{\boldsymbol{Z}_{\mathcal{H}}^t}\right) &\leq \sum_{t=1}^{T}\left(\frac{2\sqrt{N}}{\varepsilon'}\max_{\boldsymbol{\alpha}\in\mathcal{A}^T}\widehat{\mathcal{G}}_{\boldsymbol{Z}_{\mathcal{H}}^t}(\mathcal{H})\right)^2 \leq T\left(\frac{2\sqrt{N}}{\varepsilon'}C_{\boldsymbol{X}}^{\mathcal{H}}\right)^2, \\
\max_{\boldsymbol{\alpha}\in\mathcal{A}^T}\sum_{k=1}^{K_\ell}\log\mathcal{N}\left(\varepsilon';\Psi_\ell,\rho_{\boldsymbol{Z}_\ell^k}\right) &\leq \sum_{k=1}^{K_\ell}\left(\frac{2\sqrt{NT}}{\varepsilon'}\max_{\boldsymbol{\alpha}\in\mathcal{A}^T}\sqrt{\frac{|\boldsymbol{Z}_\ell^k|}{NT}}\widehat{\mathcal{G}}_{\boldsymbol{Z}_\ell^k}(\Psi_\ell)\right)^2 \leq K_\ell\left(\frac{2\sqrt{NT}}{\varepsilon'}C_{\boldsymbol{X}}^{\Psi_\ell}\right)^2.
\end{aligned}$$

The statements provided to prove Theorem 1 utilize the worst-case Gaussian complexity, and it bounds both empirical and population Gaussian complexities. Here, $C_X^{\mathcal{H}}$ and $C_{\boldsymbol{X}}^{\Psi_\ell}$ depend on the input $\boldsymbol{X}$, and by construction, they are larger than their corresponding empirical complexities, however there is no guarantee that they will be larger than the corresponding population Gaussian complexities. Combining the result with (14), we can obtain that with probability at least $1-\delta$,

$$\mathcal{R}_{\text{M}^2\text{TL}}(\hat{\boldsymbol{f}}) \leq 384\Gamma\left(\frac{D_{\boldsymbol{X}}}{NT} + D_{\boldsymbol{X}}\sqrt{\frac{\log|\mathcal{A}|}{N}} + \Gamma^{\dagger}\log NT\left(C_{\boldsymbol{X}}^{\mathcal{H}} + \sum_{\ell=1}^{L}\sqrt{K_\ell}C_{\boldsymbol{X}}^{\Psi_\ell}\right)\right) + 6\sqrt{\frac{\log\frac{4}{\delta}}{NT}}, \tag{20}$$

where $D_{\boldsymbol{X}} = \sup_{\boldsymbol{f},\boldsymbol{f}'\in\mathcal{F}}\rho_{\boldsymbol{X}}(\boldsymbol{f},\boldsymbol{f}')$. Here we consider complexity of each task-specific head separately and bound it using the task with the largest head complexity ($C_{\boldsymbol{X}}^{\mathcal{H}}$). As for the complexity of each layer, in the general case (as shown in Theorem 1), all the modules in the same layer share the same input space $\mathcal{X}_{\Psi_\ell}$ by assuming raw input space $\mathcal{X}$, and because of Lemma 3, the sample complexity of ℓ_{th} layer is bounded by $\mathcal{O}\left(\sqrt{K_\ell}\widetilde{\mathcal{G}}_{NT}(\Psi_\ell)\right)$. When given actual training data $\boldsymbol{X}$, we need to search to find the worst-case cluster method of ℓ_{th} layer, which results in $C_{\boldsymbol{X}}^{\Psi_\ell}$.

Below, we extend our theoretical result of Multipath MTL to two specific settings, vanilla MTL and hierarchical MTL.

Corollary 2 (Vanilla MTL) *Given the same data setting described in Section 2, consider a vanilla MTL problem as depicted in Figure 4a, which can be formulated as follows.*

$$\{\hat{h}_t\}_{t=1}^T, \hat{\phi} = \operatorname*{arg\,min}_{h_t\in\mathcal{H},\phi\in\Phi}\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\ell(h_t\circ\phi(\boldsymbol{x}_{ti}), y_{ti}).$$

Suppose $\mathcal{H}$, Φ are sets of Γ-Lipschitz functions with respect to Euclidean norm, and $\ell(\cdot, y) : \mathbb{R} \times \mathbb{R} \to [0, 1]$ is also Γ-Lipschitz with respect to Euclidean norm. Define $\mathcal{D}_{\mathcal{X}} = \sup_{\boldsymbol{x}\in\mathcal{X}, h\in\mathcal{H}, \phi\in\Phi} |h \circ \phi(\boldsymbol{x})| < \infty$. Let $\mathcal{L}(\{h_t\}_{t=1}^T, \phi) = \mathbb{E}_{\bar{\mathcal{D}}}[\ell(h_t \circ \phi(\boldsymbol{x}), y)]$ and $\mathcal{L}^\star = \min_{h_t\in\mathcal{H}, \phi\in\Phi} \mathbb{E}_{\bar{\mathcal{D}}}[\ell(h_t \circ \phi(\boldsymbol{x}), y)]$. Then we have that with probability at least $1 - \delta$,

$$\mathcal{L}(\{\hat{h}_t\}_{t=1}^T, \hat{\phi}) - \mathcal{L}^\star \leq 384\Gamma \left(\frac{D_{\mathcal{X}}}{NT} + (\Gamma + 1) \log NT \left(\widetilde{\mathcal{G}}_N(\mathcal{H}) + \mathcal{G}_{NT}(\Phi) \right) \right) + 2\sqrt{\frac{\log \frac{2}{\delta}}{NT}}.$$

Here, the input space for $\mathcal{H}$ is $\Psi \times \mathcal{X}$.

This corollary is consistent with (Tripuraneni, Jordan, and Jin 2020), and it can be simply deduced following the statement of Theorem 5, by setting $L = 1$, $K_1 = 1$. Since there is only one pathway selection, $|\mathcal{A}| = 1$ and $\log |\mathcal{A}| = 0$. Here, the input space for representation Φ is $\mathcal{X}$, and its complexity is shown in Gaussian complexity fashion.

Corollary 3 (Hierarchical MTL) *Consider the hierarchical MTL problem depicted in Fig. 4c and consider a hierarchical supernet with degree K. Follow the same settings in Section 2. Suppose Assumptions 1&2 hold. Let $\hat{\boldsymbol{f}}$ be the empirical solution of* (M^{2}TL)*. Let $D_{\mathcal{X}} = \sup_{\boldsymbol{x}\in\mathcal{X}, h\in\mathcal{H}, \boldsymbol{\phi}\in\Phi, \alpha\in\mathcal{A}} |h \circ \boldsymbol{\phi}_\alpha(\boldsymbol{x})| < \infty$ and $\Gamma^\dagger = \sum_{\ell=0}^L \Gamma^\ell$. Then, with probability at least $1 - \delta$, the excess test risk in* (2) *obeys*

$$\mathcal{R}_{M^2TL}(\hat{\boldsymbol{f}}) \leq 768\Gamma \left(\frac{D_{\mathcal{X}}}{NT} + D_{\mathcal{X}} \sqrt{\frac{(L-1)\log K}{N}} + \Gamma^\dagger \log NT \left(\widetilde{\mathcal{G}}_N(\mathcal{H}) + \sum_{\ell=1}^L K^{\frac{\ell-1}{2}} \widetilde{\mathcal{G}}_{NT}(\Psi_\ell) \right) \right) + 2\sqrt{\frac{\log \frac{2}{\delta}}{NT}}.$$

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 $\mathcal{X}_{\Psi_1} = \mathcal{X}$. Now if we consider a two-layer hierarchical representations as depicted in Fig. 1b, we can immediately obtain the result by setting $L = 2$ ($\Gamma^\dagger = 1 + \Gamma + \Gamma^2$). Then with probability at least $1 - \delta$,

$$\mathcal{R}_{M^2TL}(\hat{\boldsymbol{f}}) \leq 768\Gamma \left(\frac{D_{\mathcal{X}}}{NT} + D_{\mathcal{X}} \sqrt{\frac{\log K}{N}} + \Gamma^\dagger \log NT \left(\widetilde{\mathcal{G}}_N(\mathcal{H}) + \mathcal{G}_{NT}(\Psi_1) + \sqrt{K} \widetilde{\mathcal{G}}_{NT}(\Psi_2) \right) \right) + 2\sqrt{\frac{\log \frac{2}{\delta}}{NT}}.$$

The result is consistent with Section 5, and proof can be immediately done by setting $|\mathcal{A}| = K^{L-1}$ and $K_\ell = K^{\ell-1}$ in Theorem 5. Here we observe that if the complexity of Ψ_ℓ decreasing exponentially as $\text{comp}(\Psi_\ell) \propto K^{-\frac{\ell}{2}}$, then each layer has a constant complexity. We believe this and similar bounds can potentially provide guidelines on how we should design hierarchical supernets.

B.3 Proof of Lemma 1

Lemma 5 (Lemma 1 restated) *Recall $\hat{\boldsymbol{f}}$ is the solution of* (M^{2}TL) *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 $t \in [T]$,*

$$\mathcal{R}_t(\hat{f}_t) \lesssim \widetilde{\mathcal{G}}_N(\mathcal{H}) + \sum_{\ell=1}^L \widetilde{\mathcal{G}}_N(\Psi_\ell) + \sqrt{\frac{\log(2T/\delta)}{N}}. \tag{21}$$

Proof Let $\mathcal{F}_{\text{IND}}$ be the hypothesis class of a single task induced by a pathway in the supernet. Since modules are same, $\mathcal{F}_{\text{IND}}$ is same regardless of pathway. First, applying our main theorem (Thm 1) for a single supernet with $K_\ell = 1$ (i.e. on $\mathcal{F}_{\text{IND}}$), for a single task t, we end up with the uniform concentration guarantee, for all $f \in \mathcal{F}_{\text{IND}}$, with probability at least $1 - \delta$,

$$|\widehat{\mathcal{L}}_{\mathcal{S}_t}(f) - \mathcal{L}_t(f)| \lesssim \widetilde{\mathcal{G}}_N(\mathcal{H}) + \sum_{\ell=1}^L \widetilde{\mathcal{G}}_N(\Psi_\ell) + \sqrt{\frac{\log(2/\delta)}{N}}.$$

Union bounding, for all $f_t \in \mathcal{F}_{\text{IND}}$, $t \in [T]$, with probability at least $1 - \delta$, we obtain

$$|\widehat{\mathcal{L}}_{\mathcal{S}_t}(f_t) - \mathcal{L}_t(f_t)| \lesssim \widetilde{\mathcal{G}}_N(\mathcal{H}) + \sum_{\ell=1}^L \widetilde{\mathcal{G}}_N(\Psi_\ell) + \sqrt{\frac{\log(2T/\delta)}{N}}. \tag{22}$$

Let us call this intersection event $\mathcal{E}_{\text{all}}$. Intersecting this with the events $\min_{f_t\in\mathcal{F}_{\text{IND}}} \widehat{\mathcal{L}}_{\mathcal{S}_t}(f_t) = 0$ for $t \in [T]$, we exactly end up with (21). Thus, the statement is indeed what one would obtain by union bounding individualized training.

To proceed, we argue that same bound holds when solving (M^{2}TL). We know (22) holds for all f_t chosen from $\mathcal{F}_{\text{IND}}$, therefore it holds for $\hat{f}_t$, $t \in [T]$. Consider its intersection with the event $\mathbb{P}(\widehat{\mathcal{L}}_{\mathcal{S}_{\text{all}}}(\hat{\boldsymbol{f}}) \neq 0)$. Given that $\widehat{\mathcal{L}}_{\mathcal{S}_t}(\hat{f}_t) = 0$, we obtain $\mathcal{R}_t(\hat{f}_t) \leq \mathcal{L}_t(\hat{f}_t)$ upper bounded by the RHS of (22).

■

B.4 Proof of Theorem 2

Theorem 6 (Theorem 2 restated) *Suppose Assumptions 1&2 hold. Let supernet $\hat{\boldsymbol{\phi}}$ be the solution of* (M^{2}TL) *and $\hat{f}_{\hat{\boldsymbol{\phi}}}$ be the empirical minima of* (TLOP) *with respect to supernet $\hat{\boldsymbol{\phi}}$. Let $D_{\mathcal{X}} = \sup_{\boldsymbol{x}\in\mathcal{X},\alpha\in\mathcal{A},h\in\mathcal{H}_{\mathcal{T}}} |h\circ\hat{\boldsymbol{\phi}}_\alpha(\boldsymbol{x})| < \infty$. Then with probability at least $1-\delta$,*

$$\mathcal{R}_{TLOP}(\hat{f}_{\hat{\boldsymbol{\phi}}}) \leq Bias_{\mathcal{T}}(\hat{\boldsymbol{\phi}}) + 768\Gamma\left(\frac{D_{\mathcal{X}}}{M} + D_{\mathcal{X}}\sqrt{\frac{\log|\mathcal{A}|}{M}} + \log M\cdot\widetilde{\mathcal{G}}_M(\mathcal{H}_{\mathcal{T}})\right) + 2\sqrt{\frac{\log\frac{2}{\delta}}{M}},$$

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

Proof For short notation, let $\mathcal{H} := \mathcal{H}_{\mathcal{T}}$. We consider the transfer learning problem over a target task, with distribution $\mathcal{D}_{\mathcal{T}}$ and training dataset $\mathcal{S}_{\mathcal{T}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^M$ with M samples i.i.d. drawn from $\mathcal{D}_{\mathcal{T}}$. Let $\hat{\boldsymbol{\phi}}$ and $\boldsymbol{\phi}^\star$ denote the empirical and population solution of (M^{2}TL). Then, we can recap the excess transfer learning risk

$$\mathcal{R}_{\text{TLOP}}(\hat{f}_{\hat{\boldsymbol{\phi}}}) = \mathcal{L}_{\mathcal{T}}(\hat{f}_{\hat{\boldsymbol{\phi}}}) - \mathcal{L}_{\mathcal{T}}^\star = \underbrace{\mathcal{L}_{\mathcal{T}}(\hat{f}_{\hat{\boldsymbol{\phi}}}) - \mathcal{L}_{\mathcal{T}}(f^\star_{\hat{\boldsymbol{\phi}}})}_{\text{variance}(a)} + \underbrace{\mathcal{L}_{\mathcal{T}}(f^\star_{\hat{\boldsymbol{\phi}}}) - \mathcal{L}_{\mathcal{T}}^\star}_{\text{supernet bias }(b)}.$$

Following Definition 2, $b = \text{Bias}_{\mathcal{T}}(\hat{\boldsymbol{\phi}})$, and it remains to bound variance (a). Let $\hat{f}_{\hat{\boldsymbol{\phi}}} := (\hat{h}_{\hat{\boldsymbol{\phi}}}, \hat{\alpha}_{\hat{\boldsymbol{\phi}}})$ and $f^\star_{\hat{\boldsymbol{\phi}}} := (h^\star_{\hat{\boldsymbol{\phi}}}, \alpha^\star_{\hat{\boldsymbol{\phi}}})$. For short notations, we remove the subscript $\hat{\boldsymbol{\phi}}$, and we assume supernet $\hat{\boldsymbol{\phi}}$ is implied. Following the similar statements in Appendix B.2, we can decompose variance as follows.

$$a = \mathcal{L}_{\mathcal{T}}(\hat{f}) - \mathcal{L}_{\mathcal{T}}(f^\star) = \underbrace{\mathcal{L}_{\mathcal{T}}(\hat{f}) - \widehat{\mathcal{L}}_{\mathcal{T}}(\hat{f})}_{c} + \underbrace{\widehat{\mathcal{L}}_{\mathcal{T}}(\hat{f}) - \widehat{\mathcal{L}}_{\mathcal{T}}(f^\star)}_{d} + \underbrace{\widehat{\mathcal{L}}_{\mathcal{T}}(f^\star) - \mathcal{L}_{\mathcal{T}}(f^\star)}_{e}$$

where $\mathcal{L}_{\mathcal{T}}(f) = \mathbb{E}_{\mathcal{D}_{\mathcal{T}}}[\ell(h\circ\hat{\boldsymbol{\phi}}_\alpha(\boldsymbol{x}), y)]$ and $\widehat{\mathcal{L}}_{\mathcal{T}}(f) = \frac{1}{M}\sum_{i=1}^M \ell(h\circ\hat{\boldsymbol{\phi}}_\alpha(\boldsymbol{x}_i), y_i)$ where $f = (h,\alpha)$ and $(\boldsymbol{x}_i, y_i)\in\mathcal{S}_{\mathcal{T}}$. Since $\hat{f}$ minimizes the training loss given $\hat{\boldsymbol{\phi}}$, $d\leq 0$. Let $\boldsymbol{X}$ denote the input dataset, that is, $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^M$. Same as Inequality (14) in Appendix B.2, we derive the similar result that with probability at least $1-\delta$,

$$\mathcal{R}_{\text{TLOP}}(\hat{f}) \leq \text{Bias}_{\mathcal{T}}(\hat{\boldsymbol{\phi}}) + 6\Gamma\mathcal{G}_M(\mathcal{H}\circ\hat{\boldsymbol{\phi}}(\mathcal{A})) + 2\sqrt{\frac{\log\frac{2}{\delta}}{M}},$$

$$\text{and}\ \ \mathcal{R}_{\text{TLOP}}(\hat{f}) \leq \text{Bias}_{\mathcal{T}}(\hat{\boldsymbol{\phi}}) + 6\Gamma\widehat{\mathcal{G}}_{\boldsymbol{X}}(\mathcal{H}\circ\hat{\boldsymbol{\phi}}(\mathcal{A})) + 6\sqrt{\frac{\log\frac{2}{\delta}}{M}},$$

where $\widehat{\mathcal{G}}_{\boldsymbol{X}}(\mathcal{H}\circ\hat{\boldsymbol{\phi}}(\mathcal{A})) = \mathbb{E}_{\boldsymbol{g}}\left[\sup_{h\in\mathcal{H},\alpha\in\mathcal{A}}\frac{1}{M}\sum_{i=1}^M g_i h\circ\hat{\boldsymbol{\phi}}_\alpha(\boldsymbol{x}_i)\right]$ and $\mathcal{G}_M(\mathcal{H}\circ\hat{\boldsymbol{\phi}}(\mathcal{A})) = \mathbb{E}_{\mathcal{D}_{\mathcal{T}}}\left[\widehat{\mathcal{G}}_{\boldsymbol{X}}(\mathcal{H}\circ\hat{\boldsymbol{\phi}}(\mathcal{A}))\right]$. Following the Definition 4, let $D = \sup_{h,h'\in\mathcal{H},\alpha,\alpha'\in\mathcal{A}}\rho_{\boldsymbol{X}}(h\circ\hat{\boldsymbol{\phi}}_\alpha, h'\circ\hat{\boldsymbol{\phi}}_{\alpha'}) \leq 2D_{\mathcal{X}}$. By applying the Dudley's theorem, and following the same statements in Appendix B.2, we obtain that given any $\varepsilon\in[0, D]$

$$\widehat{\mathcal{G}}_{\boldsymbol{X}}(\mathcal{H}\circ\hat{\boldsymbol{\phi}}(\mathcal{A})) \leq 2\varepsilon + \frac{32}{\sqrt{M}}\int_{\varepsilon/4}^D\sqrt{\log\mathcal{N}\left(u;\mathcal{H}\circ\hat{\boldsymbol{\phi}}(\mathcal{A}),\rho_{\boldsymbol{X}}\right)}du.$$

Now we need to decompose the covering number of $\mathcal{H}\circ\hat{\boldsymbol{\phi}}(\mathcal{A})$ into the covering numbers of separate hypothesis sets $\mathcal{H}$ and $\mathcal{A}$. For short notations, let $\mathcal{H}(\mathcal{A}) := \mathcal{H}\circ\hat{\boldsymbol{\phi}}(\mathcal{A})$ and $\mathcal{H}(\alpha) := \mathcal{H}\circ\hat{\boldsymbol{\phi}}_\alpha$, and we omit the subscript $\boldsymbol{X}$ from ρ. Since pathway set $\mathcal{A}$ is discrete with cardinality $|\mathcal{A}|$, the covering number of $\mathcal{H}(\mathcal{A})$ is the product of covering number of $\mathcal{H}(\alpha)$ for all $\alpha\in\mathcal{A}$, and can be bounded by the $|\mathcal{A}|$ times product of the worst-case covering number of $\mathcal{H}(\alpha)$, that is $\mathcal{N}(u;\mathcal{H}(\mathcal{A}),\rho) = \Pi_{\alpha\in\mathcal{A}}\mathcal{N}(u;\mathcal{H}(\alpha),\rho) \leq \max_{\alpha\in\mathcal{A}}\mathcal{N}^{|\mathcal{A}|}(u;\mathcal{H}(\alpha),\rho)$. Logarithm of it results in $\log\mathcal{N}(u;\mathcal{H}(\mathcal{A}),\rho) \leq \log|\mathcal{A}| + \max_{\alpha\in\mathcal{A}}\mathcal{N}(u;\mathcal{H}(\alpha),\rho)$. Now let $\boldsymbol{Z}_\alpha = \hat{\boldsymbol{\phi}}_\alpha(\boldsymbol{X}) = \{\hat{\boldsymbol{\phi}}_\alpha(\boldsymbol{x}_i) : \boldsymbol{x}_i\in\boldsymbol{X}\}$, which is the set of latent inputs of prediction head. Then for any given $\alpha\in\mathcal{A}$,

$$\rho_{\boldsymbol{X}}(h\circ\hat{\boldsymbol{\phi}}_\alpha, h'\circ\hat{\boldsymbol{\phi}}_\alpha) = \sqrt{\frac{1}{M}\sum_{i=1}^M\left(h\circ\hat{\boldsymbol{\phi}}_\alpha(\boldsymbol{x}_i) - h'\circ\hat{\boldsymbol{\phi}}_\alpha(\boldsymbol{x}_i)\right)^2} = \sqrt{\frac{1}{M}\sum_{i=1}^M(h(\boldsymbol{z}_i) - h'(\boldsymbol{z}_i))^2} = \rho_{\boldsymbol{Z}_\alpha}(h,h'),$$

where $\boldsymbol{z}_i = \hat{\boldsymbol{\phi}}_\alpha(\boldsymbol{x}_i)$ and then $\boldsymbol{Z}_\alpha = \{\boldsymbol{z}_1,\ldots,\boldsymbol{z}_M\}$. Such equality states that if pathway α is fixed, u-cover of head $\mathcal{H}$ results in u-cover of the prediction function, and simply, $\mathcal{N}(u;\mathcal{H}(\alpha),\rho_{\boldsymbol{X}}) = \mathcal{N}(u;\mathcal{H},\rho_{\boldsymbol{Z}_\alpha})$. Next, following the same statements

in Appendix B.2, if we utilize the Sudakov minoration theorem in (Wainwright 2019), we obtain $\sqrt{\log \mathcal{N}(u; \mathcal{H}, \rho_{\boldsymbol{Z}_\alpha})} \leq \frac{2\sqrt{M}}{u} \widehat{\mathcal{G}}_{\boldsymbol{Z}_\alpha}(\mathcal{H})$. Finally, combining all we have together obtains

$$\begin{aligned}\widehat{\mathcal{G}}_{\boldsymbol{X}}(\mathcal{H}(\mathcal{A})) &\leq 2\varepsilon + \frac{32}{\sqrt{M}} \int_{\varepsilon/4}^{D} \sqrt{\log \mathcal{N}\left(u; \mathcal{H}(\mathcal{A}), \rho_{\boldsymbol{X}}\right)} du \leq 2\varepsilon + 32D\sqrt{\frac{\log|\mathcal{A}|}{M}} + 64 \max_{\alpha \in \mathcal{A}} \widehat{\mathcal{G}}_{\boldsymbol{Z}_\alpha}(\mathcal{H}) \int_{\varepsilon/4}^{D} \frac{1}{u} du \\ &\leq 2\varepsilon + 32D\sqrt{\frac{\log|\mathcal{A}|}{M}} + 64 \log \frac{4D}{\varepsilon} \max_{\alpha \in \mathcal{A}} \widehat{\mathcal{G}}_{\boldsymbol{Z}'_\alpha}(\mathcal{H}) \leq 64 \left(\frac{D}{M} + D\sqrt{\frac{\log|\mathcal{A}|}{M}} + \log M \max_{\alpha \in \mathcal{A}} \widehat{\mathcal{G}}_{\boldsymbol{Z}_\alpha}(\mathcal{H}) \right),\end{aligned}$$

by choosing $\varepsilon = \frac{4D}{M}$.

• **Input-dependent bound.** If we define the worst case *empirical* Gaussian complexity based on the raw input data $\boldsymbol{X}$ and given supernet $\hat{\boldsymbol{\phi}}$, that is $C_{\boldsymbol{X}}^{\mathcal{H}} := \max_{\alpha \in \mathcal{A}} \widehat{\mathcal{G}}_{\boldsymbol{Z}_\alpha}(\mathcal{H})$, where $\boldsymbol{Z}_\alpha$ shows as above with respect to $\hat{\boldsymbol{\phi}}$ and α, we have that with probability at least $1 - \delta$,

$$\mathcal{R}_{\text{TLOP}}(\hat{f}_{\hat{\boldsymbol{\phi}}}) \leq \text{Bias}_{\mathcal{T}}(\hat{\boldsymbol{\phi}}) + 384\Gamma \left(\frac{D}{M} + D\sqrt{\frac{\log|\mathcal{A}|}{M}} + \log M \cdot C_{\boldsymbol{X}}^{\mathcal{H}} \right) + 6\sqrt{\frac{\log \frac{4}{\delta}}{M}}.$$

Furthermore, let input space be $\mathcal{X}$. If we define the worst case Gaussian complexity independent to the specific training dataset and supernet, that is, $\widetilde{\mathcal{G}}_M^{\mathcal{X}_\mathcal{H}}(\mathcal{H}) := \sup_{\boldsymbol{Z} \in \mathcal{X}_\mathcal{H}^M} \widehat{\mathcal{G}}_{\boldsymbol{X}}(\mathcal{H})$, where $\mathcal{X}_\mathcal{H} = \{\hat{\boldsymbol{\phi}}_\alpha \circ \mathcal{X} | \alpha \in \mathcal{A}\}$, then we have that

$$\mathcal{G}_M(\mathcal{H}(\mathcal{A})) \leq 64 \left(\frac{D_\mathcal{X}}{M} + D_\mathcal{X}\sqrt{\frac{\log|\mathcal{A}|}{M}} + \log M \cdot \widetilde{\mathcal{G}}_M^{\mathcal{X}_\mathcal{H}}(\mathcal{H}) \right),$$

which leads to the result that with probability at least $1 - \delta$,

$$\mathcal{R}_{\text{TLOP}}(\hat{f}_{\hat{\boldsymbol{\phi}}}) \leq \text{Bias}_{\mathcal{T}}(\hat{\boldsymbol{\phi}}) + 384\Gamma \left(\frac{D_\mathcal{X}}{M} + D_\mathcal{X}\sqrt{\frac{\log|\mathcal{A}|}{M}} + \log M \cdot \widetilde{\mathcal{G}}_M(\mathcal{H}) \right) + 2\sqrt{\frac{\log \frac{2}{\delta}}{M}}.$$

Here input space of $\mathcal{H}$ is given by $\mathcal{X}_\mathcal{H} = \{\hat{\boldsymbol{\phi}}_\alpha \circ \mathcal{X} | \alpha \in \mathcal{A}\}$. ■

B.5 End-to-End Transfer Learning

In this section, we present an end-to-end transfer learning guarantee based on task diversity. We start with two useful definitions: supernet distance and task diversity. Here, supernet distance has been mentioned in Section 3.2 and following provides the intact definition. It measures the performance gap of two supernets. Similar to the previous work (Chen et al. 2021; Tripuraneni, Jordan, and Jin 2020; Xu and Tewari 2021), we define task diversity in Definition 6. It captures the similarity of target task to source tasks over a supernet by comparing their representation distance over it. Finally, using the task diversity argument, we can immediately obtain the theoretical guarantee for transfer learning risk.

Definition 5 (Supernet Distance) *Consider a transfer learning with optimal pathway* (TLOP) *problem. Recall the definitions $\mathcal{D}_\mathcal{T}$ and $\mathcal{H}_\mathcal{T}$ stated in Section 2. Given two supernets $\boldsymbol{\phi}$ and $\boldsymbol{\phi}'$, define the supernet/representation distance of $\boldsymbol{\phi}$ from $\boldsymbol{\phi}'$ for a target $\mathcal{T}$ as*

$$\textit{Dist}_\mathcal{T}(\boldsymbol{\phi}; \boldsymbol{\phi}') = \textit{Bias}_\mathcal{T}(\boldsymbol{\phi}) - \textit{Bias}_\mathcal{T}(\boldsymbol{\phi}') = \min_{h \in \mathcal{H}_\mathcal{T}, \alpha \in \mathcal{A}} \mathcal{L}_\mathcal{T}(h \circ \boldsymbol{\phi}_\alpha) - \min_{h \in \mathcal{H}_\mathcal{T}, \alpha \in \mathcal{A}} \mathcal{L}_\mathcal{T}(h \circ \boldsymbol{\phi}'_\alpha).$$

Here, we do not restrict the supernet distance to target task $\mathcal{T}$ only. Given source task $t \in [T]$, we can still define the corresponding supernet distance of $\boldsymbol{\phi}$ from $\boldsymbol{\phi}'$ as

$$\text{Dist}_t(\boldsymbol{\phi}; \boldsymbol{\phi}') = \min_{h \in \mathcal{H}, \alpha \in \mathcal{A}} \mathcal{L}_t(h \circ \boldsymbol{\phi}_\alpha) - \min_{h \in \mathcal{H}, \alpha \in \mathcal{A}} \mathcal{L}_t(h \circ \boldsymbol{\phi}'_\alpha), \tag{23}$$

and the hypothesis set for head is $\mathcal{H}$ instead.

Definition 6 (Task Diversity) *For any supernets $\boldsymbol{\phi}$ and $\boldsymbol{\phi}'$, given T source tasks with distribution $(\mathcal{D}_t)_{t=1}^T$ and a target task with distribution $\mathcal{D}_\mathcal{T}$, we say that the source tasks are (ν, ϵ)-diverse over the target task for a supernet $\boldsymbol{\phi}'$ if for any $\boldsymbol{\phi} \in \Phi$,*

$$\textit{Dist}_\mathcal{T}(\boldsymbol{\phi}; \boldsymbol{\phi}') \leq \left(\frac{1}{T} \sum_{t=1}^{T} \textit{Dist}_t(\boldsymbol{\phi}; \boldsymbol{\phi}') \right) / \nu + \epsilon,$$

where we assume that head hypothesis sets $\mathcal{H}$, $\mathcal{H}_\mathcal{T}$ are implied for source and target distances.

Theorem 7 (End-to-end transfer learning) *Suppose Assumption 1&2 hold. Let supernet $\hat{\boldsymbol{\phi}}$ and $\boldsymbol{\phi}^\star$ be the empirical and population solutions of* (M^2TL) *and $\hat{f}_{\hat{\boldsymbol{\phi}}}$ be the empirical minima of* (TLOP) *with respect to supernet $\hat{\boldsymbol{\phi}}$. Assume the source tasks used in Multipath MTL phase are (ν, ϵ)-diverse over target task $\mathcal{T}$ for the optimal supernet $\boldsymbol{\phi}^\star$. Then with probability at least $1 - 2\delta$,*

$$\mathcal{R}_{TLOP}(\hat{f}_{\hat{\boldsymbol{\phi}}}) \lesssim Bias_{\mathcal{T}}(\boldsymbol{\phi}^\star) + \frac{1}{\nu}\left(\widetilde{\mathcal{G}}_N(\mathcal{H}) + \sum_{\ell=1}^{L}\sqrt{\hat{K}_\ell}\widetilde{\mathcal{G}}_{NT}(\Psi_\ell) + \sqrt{\frac{\log|\mathcal{A}|}{N}}\right) + \sqrt{\frac{\log|\mathcal{A}|}{M}} + \widetilde{\mathcal{G}}_M(\mathcal{H}_{\mathcal{T}}) + \frac{1}{\nu}\sqrt{\frac{\log\frac{2}{\delta}}{NT}} + \sqrt{\frac{\log\frac{2}{\delta}}{M}} + \epsilon.$$

Here, the input spaces for $\mathcal{H}$, Ψ_ℓ and $\mathcal{H}_{\mathcal{T}}$ are same to the statements in Theorem 1 and Theorem 2.

Proof Recall Theorem 6. To state end-to-end transfer learning risk, we need to bound supernet bias $\mathrm{Bias}_{\mathcal{T}}(\hat{\boldsymbol{\phi}})$. Following Definition 5, we have that $\mathrm{Bias}_{\mathcal{T}}(\hat{\boldsymbol{\phi}}) = \mathrm{Dist}_{\mathcal{T}}(\hat{\boldsymbol{\phi}}; \boldsymbol{\phi}^\star) + \mathrm{Bias}_{\mathcal{T}}(\boldsymbol{\phi}^\star)$. Next, from Definition 6, since we assume source tasks are (ν, ϵ)-diverse over target task $\mathcal{T}$ for the supernet $\boldsymbol{\phi}^\star$, we can obtain $\mathrm{Dist}_{\mathcal{T}}(\hat{\boldsymbol{\phi}}; \boldsymbol{\phi}^\star) \leq \left(\frac{1}{T}\sum_{t=1}^{T}\mathrm{Dist}_t(\hat{\boldsymbol{\phi}}; \boldsymbol{\phi}^\star)\right)/\nu + \epsilon$. To process, following (23), we have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathrm{Dist}_t(\hat{\boldsymbol{\phi}}; \boldsymbol{\phi}^\star) &= \frac{1}{T}\sum_{t=1}^{T}\left(\min_{h\in\mathcal{H},\alpha\in\mathcal{A}}\mathcal{L}_t(h\circ\hat{\boldsymbol{\phi}}_\alpha) - \min_{h\in\mathcal{H},\alpha\in\mathcal{A}}\mathcal{L}_t(h\circ\boldsymbol{\phi}^\star_\alpha)\right)\\ &\leq \frac{1}{T}\sum_{t=1}^{T}\left(\mathcal{L}_t(\hat{h}_t\circ\hat{\boldsymbol{\phi}}_{\hat{\alpha}_t}) - \mathcal{L}_t(h_t^\star\circ\boldsymbol{\phi}^\star_{\alpha_t^\star})\right) = \mathcal{L}_{\bar{\mathcal{D}}}(\hat{\boldsymbol{f}}) - \mathcal{L}^\star_{\bar{\mathcal{D}}} = \mathcal{R}_{\mathrm{M}^2\mathrm{TL}}(\hat{\boldsymbol{f}}).\end{aligned}$$

Here, $(\{\hat{h}_t, \hat{\alpha}_t\}_{t=1}^T, \hat{\boldsymbol{\phi}})$ and $(\{h_t^\star, \alpha_t^\star\}_{t=1}^T, \boldsymbol{\phi}^\star)$ are the empirical and population solutions of (M^2TL), and we set $\hat{\boldsymbol{f}} := (\{\hat{h}_t, \hat{\alpha}_t\}_{t=1}^T, \hat{\boldsymbol{\phi}})$. The inequality term holds from the fact that: 1) $\min_{h\in\mathcal{H},\alpha\in\mathcal{A}}\mathcal{L}_t(h\circ\hat{\boldsymbol{\phi}}_\alpha) \leq \mathcal{L}_t(\hat{h}_t\circ\hat{\boldsymbol{\phi}}_{\hat{\alpha}_t})$, and 2) $\min_{h\in\mathcal{H},\alpha\in\mathcal{A}}\mathcal{L}_t(h\circ\boldsymbol{\phi}^\star_\alpha) = \mathcal{L}_t(h_t^\star\circ\boldsymbol{\phi}^\star_{\alpha_t^\star})$ since $h_t^\star$ and $\alpha_t^\star$ can be seen as the optimal solutions given supernet $\boldsymbol{\phi}^\star$. Combining them together with Theorem 1 and Theorem 2 completes the proof. ■

C Multipath MTL under Subexponential Loss Functions

The goal of this section is proving an MTL result under unbounded loss functions (e.g. least-squares). The high-level proof strategy is essentially a simplified version of proof of Theorem 1, where we use a more naive covering argument for parametric classes that have $\mathcal{O}(\log(1/\varepsilon))$ covering numbers. For this reason, we will make some simplifications in the proof to avoid repetitions. Instead, we will highlight key differences such as how the concentration argument changes due to unbounded losses. We first make the following assumptions.

Assumption 7 *For any task distribution $(\boldsymbol{x}, y) \sim \mathcal{D}_t$ and for any task hypothesis $f_t \in \mathcal{F}_t$ (induced by $\alpha_t, h_t, \boldsymbol{\phi}$), we have that $\ell(y, f_t(\boldsymbol{x}))$ is a Ξ subexponential random variable for some $\Xi > 0$. Additionally, assume loss is a $\Gamma > 0$ Lipschitz function of $\hat{y} = f_t(\boldsymbol{x})$.*

We also assume a standard covering assumption. Note that, unlike the proof of Theorem 1, we focus on parametric classes and use data-agnostic covers.

Assumption 8 *All modules $\psi_\ell^k \in \Psi_\ell$ are Γ Lipschitz and map $\psi_\ell^k : \mathbb{R}^{p_{\ell-1}} \to \mathbb{R}^{p_\ell}$. For the sake of simplicity assume $\psi_\ell^k(0) = 0$ (e.g. neural net layer with ReLU activation). Additionally, for any Euclidean ball of radius R, the covering dimension of Ψ_ℓ follows the parametric classes, namely,*

$$\mathcal{N}(\varepsilon; \Psi_\ell, R) \leq d_\ell \log(\frac{3R}{\varepsilon}),$$

where d_ℓ is the covering dimension of Ψ_ℓ. Verbally, there exists a cover Ψ_ℓ^ε, $|\Psi_\ell^\varepsilon| \leq \mathcal{N}(\varepsilon; \Psi_\ell, R)$, such that for any $\|\boldsymbol{x}\| \leq R$ and for any $\psi \in \Psi_\ell$, there exists $\psi' \in \Psi_\ell^\varepsilon$ such that $\|\psi'(\boldsymbol{x}) - \psi(\boldsymbol{x})\| \leq \varepsilon$. Additionally, let head $\mathcal{H}$ be Γ Lipschitz and $d_{\mathcal{H}}$ be the covering dimension for $\mathcal{H}$.

Theorem 8 *Suppose $\mathcal{X} \subset \mathcal{B}^p(R)$ and Assumptions 7 and 8 hold. Suppose we have a* (M^2TL) *problem with N_{tot} samples in total where all training samples are independent, however, task sample sizes are arbitrary*[1]*. Note that, in the specific setting of Theorem 1, we have $N_{tot} = NT$ with identical sample sizes. Assume that $N_{tot} \gtrsim DoF(\mathcal{F})\log(N_{tot}) + T\log|\mathcal{A}|$. Declare population risk $\mathcal{L}_{\bar{\mathcal{D}}}(\boldsymbol{f}) = \mathbb{E}[\widehat{\mathcal{L}}_{\mathcal{S}_{all}}(\boldsymbol{f})]$. We have that with probability at least $1 - \delta$, for all Multipath hypothesis $\boldsymbol{f} \in \mathcal{F}$*

$$|\widehat{\mathcal{L}}_{\mathcal{S}_{all}}(\boldsymbol{f}) - \mathcal{L}_{\bar{\mathcal{D}}}(\boldsymbol{f})| \lesssim \Xi\sqrt{\frac{L\cdot DoF(\mathcal{F}) + T\log|\mathcal{A}| + \log(2/\delta)}{N_{tot}}}. \tag{24}$$

[1] In words, tasks don't have to have N samples each. We will simply control the gap between empirical and population. If tasks have different sizes, then their population weights will similarly change.

The right hand side bounds similarly hold for the excess risk $\mathcal{R}_{M^2TL}(\hat{\boldsymbol{f}})$ where $\hat{\boldsymbol{f}}$ is the ERM solution. $\lesssim$ subsumes the logarithmic dependence on R, Γ, L, N_{tot}. The exact bound is below (39)*. Finally, if we solve* (M^2TL) *with fixed pathway choices (rather than searching over $\mathcal{A}$), with same probability and assuming $N_{tot} \gtrsim DoF(\mathcal{F}) \log(N_{tot})$ we have the simplified bound*

$$|\widehat{\mathcal{L}}_{\mathcal{S}_{all}}(\boldsymbol{f}) - \mathcal{L}_{\bar{\mathcal{D}}}(\boldsymbol{f})| \lesssim \Xi \sqrt{\frac{L \cdot DoF(\mathcal{F}) + \log(2/\delta)}{N_{tot}}}. \tag{25}$$

The theorem is automatically applicable to loss functions bounded by $\Xi > 0$. Also note that, this theorem avoids exponential depth dependence compared to Theorem 1. This is primarily because of the strong coverability of parametric classes which (essentially) applies a log operation to the Lipschitz constant of $\mathcal{F}$.

Proof Let $R_0 = R$ and note that, at the ℓth layer, the input(output) space has radius $R_{\ell-1}(R_\ell)$ where $R_\ell = \Gamma^\ell R$. Let Ψ_ℓ denote the hypothesis set of the modules of ℓth layer. Fix an ε cover $\mathcal{F}_\varepsilon$ for the sets $(\Psi_\ell^{K_\ell})_{\ell=1}^L, \mathcal{H}^T$ and $\mathcal{A}^T$, such that Ψ_ℓ is covered according to its input space radius $R_{\ell-1}$ with resolution $\varepsilon_\ell = \frac{\varepsilon}{\Gamma^{L-\ell+1}}$, where ℓ is layer depth and prediction head is layer $L+1$. This implies that

$$\log|\mathcal{F}_\varepsilon| \le T d_{\mathcal{H}} \log \frac{3R_L}{\varepsilon} + T \log|\mathcal{A}| + \sum_{\ell=1}^{L} K_\ell d_\ell \log \frac{3R_{\ell-1}}{\varepsilon_\ell} \tag{26}$$

$$= (T d_{\mathcal{H}} + \sum_{\ell=1}^{L} K_\ell d_\ell) \log \frac{3R_L}{\varepsilon} + T \log|\mathcal{A}| \tag{27}$$

$$\le \mathrm{DoF}(\mathcal{F}) \log \frac{3R_L}{\varepsilon} + T \log|\mathcal{A}| \tag{28}$$

$$= \mathrm{DoF}(\mathcal{F}) \left(\log \frac{3R}{\varepsilon} + L \log \Gamma \right) + T \log|\mathcal{A}|. \tag{29}$$

• **Step 1: Union bound over the cover.** We now show a uniform concentration argument over this cover. Since each sample is independent of others and each loss is Ξ subexponential, using subexponential Bernstein inequality (e.g. Prop 5.16 of (Vershynin 2010)), we have that

$$\mathbb{P}\left(|\widehat{\mathcal{L}}_{\mathcal{S}_{\text{all}}}(\boldsymbol{f}) - \mathcal{L}_{\bar{\mathcal{D}}}(\boldsymbol{f})| \ge \frac{t}{\sqrt{N_{\text{tot}}}} \right) \le 2 \exp\left(-c \min\left\{ \frac{t^2}{\Xi^2}, \frac{t\sqrt{N_{\text{tot}}}}{\Xi} \right\} \right)$$

Let $\varepsilon = \frac{1}{\Gamma(L+1)N_{\text{tot}}}$, and recall that we assumed $N_{\text{tot}} \gtrsim \log|\mathcal{F}_\varepsilon| + \tau$. Now, setting $t \propto \sqrt{\log|\mathcal{F}_\varepsilon| + \tau}$ and union bounding over all $\boldsymbol{f} \in \mathcal{F}_\varepsilon$, we find that, uniformly over $\mathcal{F}_\varepsilon$,

$$\mathbb{P}\left(|\widehat{\mathcal{L}}_{\mathcal{S}_{\text{all}}}(\boldsymbol{f}) - \mathcal{L}_{\bar{\mathcal{D}}}(\boldsymbol{f})| \ge \Xi \sqrt{\frac{\log|\mathcal{F}_\varepsilon| + \tau}{N_{\text{tot}}}} \right) \le 2e^{-\tau}. \tag{30}$$

• **Step 2: Perturbation analysis.** Now that covering analysis is done, we proceed with controlling the perturbation. Let $\boldsymbol{f} \in \mathcal{F}$ be a Multipath MTL hypothesis. We choose $\boldsymbol{f}' \in \mathcal{F}_\varepsilon$ such that:

- $\boldsymbol{f}'$ chooses the same pathways.
- $\boldsymbol{f}'$ chooses heads $(h'_t)_{t=1}^T$ and modules $((\psi'^k_\ell)_{k=1}^{K_\ell})_{\ell=1}^L$ such that these hypotheses are ε close over their respective input spaces to the hypotheses of $\boldsymbol{f}$ denoted by $(h_t)_{t=1}^T$ and modules $((\psi_\ell^k)_{k=1}^{K_\ell})_{\ell=1}^L$.

Fix an arbitrary $\boldsymbol{x} \in \mathcal{X}$ and task $t \in [T]$. Set the short-hand notation $\tilde{\psi}_\ell = \psi_\ell^{\alpha_t}$ and $\tilde{\psi}'_\ell = \psi'^{\alpha_t}_\ell$. Along the pathway α_t, define the functions

$$f_t^\ell(\boldsymbol{x}) = \begin{cases} f_t(\boldsymbol{x}) & \text{if} \quad \ell = L+1, \\ f'_t(\boldsymbol{x}) & \text{if} \quad \ell = 0, \\ h'_t \circ \tilde{\psi}'_L \circ \ldots \tilde{\psi}'_{\ell+1} \circ \tilde{\psi}_\ell \circ \cdots \circ \tilde{\psi}_1(\boldsymbol{x}) & \text{if} \quad 1 \le \ell \le L. \end{cases}$$

Let $\boldsymbol{x}_\ell = \tilde{\psi}_\ell \circ \cdots \circ \tilde{\psi}_1(\boldsymbol{x})$. Recall that, Ψ_ℓ is covered with resolution ε_ℓ. Now, through a standard perturbation decomposition,

we find that

$$|f_t(\boldsymbol{x}) - f'_t(\boldsymbol{x})| \le \sum_{\ell=0}^{L} |f_t^{\ell+1}(\boldsymbol{x}) - f_t^{\ell}(\boldsymbol{x})| \tag{31}$$

$$\le \sum_{\ell=0}^{L} |h'_t \circ \tilde{\psi}'_L \circ \ldots \tilde{\psi}_{\ell+1}(\boldsymbol{x}_\ell) - h'_t \circ \tilde{\psi}'_L \circ \ldots \tilde{\psi}'_{\ell+1}(\boldsymbol{x}_\ell)| \tag{32}$$

$$\le \sum_{\ell=0}^{L} \Gamma^{L-\ell} \varepsilon_{\ell+1} \tag{33}$$

$$= (L+1)\varepsilon. \tag{34}$$

This establishes that if tasks choose the same pathways, proposed ε cover ensures that for all $\boldsymbol{x} \in \mathcal{X}$ and task t, $\mathcal{F}_\varepsilon$ is an $(L+1)\varepsilon$ cover of $\mathcal{F}$. To conclude, using Γ Lipschitzness of the loss function, we obtain

$$\mathcal{L}_{\bar{\mathcal{D}}}(\boldsymbol{f}) - \mathcal{L}_{\bar{\mathcal{D}}}(\boldsymbol{f}') \le \sup_{\boldsymbol{x},t} |\ell(y, f_t(\boldsymbol{x})) - \ell(y, f'_t(\boldsymbol{x}))| \le \sup_{\boldsymbol{x},t} \Gamma |f_t(\boldsymbol{x}) - f'_t(\boldsymbol{x})| \le \Gamma(L+1)\varepsilon \tag{35}$$

$$\widehat{\mathcal{L}}_{\mathcal{S}_{\text{all}}}(\boldsymbol{f}) - \widehat{\mathcal{L}}_{\mathcal{S}_{\text{all}}}(\boldsymbol{f}') \le \sup_{\boldsymbol{x},t} |\ell(y, f_t(\boldsymbol{x})) - \ell(y, f'_t(\boldsymbol{x}))| \le \sup_{\boldsymbol{x},t} \Gamma |f_t(\boldsymbol{x}) - f'_t(\boldsymbol{x})| \le \Gamma(L+1)\varepsilon. \tag{36}$$

Combining with uniform concentration, we found that, for all $\boldsymbol{f} \in \mathcal{F}$, with probability $1-\delta$,

$$|\widehat{\mathcal{L}}_{\mathcal{S}_{\text{all}}}(\boldsymbol{f}) - \mathcal{L}_{\bar{\mathcal{D}}}(\boldsymbol{f})| \lesssim \Xi \sqrt{\frac{\log|\mathcal{F}_\varepsilon| + \log(2/\delta)}{N_{\text{tot}}}} + \Gamma(L+1)\varepsilon \tag{37}$$

$$\lesssim \Xi \sqrt{\frac{\text{DoF}(\mathcal{F})\left(\log\frac{3R}{\varepsilon} + L\log\Gamma\right) + T\log|\mathcal{A}| + \log(2/\delta)}{N_{\text{tot}}}} + \Gamma(L+1)\varepsilon. \tag{38}$$

Recall that $\varepsilon = \frac{1}{\Gamma(L+1)N_{\text{tot}}}$, then we obtain the advertised uniform concentration guarantee

$$|\widehat{\mathcal{L}}_{\mathcal{S}_{\text{all}}}(\boldsymbol{f}) - \mathcal{L}_{\bar{\mathcal{D}}}(\boldsymbol{f})| \lesssim \Xi \sqrt{\frac{\text{DoF}(\mathcal{F})\left(\log(3R\Gamma(L+1)N_{\text{tot}}) + L\log\Gamma\right) + T\log|\mathcal{A}| + \log(2/\delta)}{N_{\text{tot}}}}. \tag{39}$$

We get the simplified statement (24) after ignoring the log factors. Finally, (25) arises by repeating above argument step-by-step while ignoring $|\mathcal{A}|$ term in (26). ■

D Proofs in Section 4

D.1 A direct corollary of Theorem 1 to linear representations

We start with a lemma that controls the worst-case Gaussian complexity of linear models. The proof is standard and stated for completeness.

Lemma 6 (Linear models) *Let $\mathcal{B} \subset \mathbb{R}^{d\times p}$ be a set of matrices with operator norm bounded by a constant $C > 0$ and let $\mathcal{X} \subset \mathcal{B}^p(R)$ (subset of ℓ_2 ball of radius R) . Then*

$$\widetilde{\mathcal{G}}_n^{\mathcal{X}}(\mathcal{B}) \le CR\sqrt{\frac{dp}{n}}.$$

Proof Set $\boldsymbol{X}_\varepsilon = \sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{g}_i^\top = \boldsymbol{X}^\top \boldsymbol{G}$ where $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is dataset and $\boldsymbol{G} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1) \in \mathbb{R}^{n\times d}$. Applying Cauchy-Schwarz, we write

$$\widetilde{\mathcal{G}}_n^{\mathcal{X}}(\mathcal{B}) = \frac{1}{n} \sup_{\boldsymbol{X}\in\mathcal{X}^n} \mathbb{E}\left[\sup_{\boldsymbol{B}\in\mathcal{B}} \sum_{i=1}^n \boldsymbol{g}_i^\top \boldsymbol{B}\boldsymbol{x}_i\right] = \frac{1}{n} \sup_{\boldsymbol{X}\in\mathcal{X}^n} \mathbb{E}\left[\sup_{\boldsymbol{B}\in\mathcal{B}} \text{trace}(\boldsymbol{X}_\varepsilon \boldsymbol{B})\right] \tag{40}$$

$$\le C\frac{\sqrt{p}}{n} \sup_{\boldsymbol{X}\in\mathcal{X}^n} \mathbb{E}[\|\boldsymbol{X}_\varepsilon\|_F] \le C\frac{\sqrt{p}}{n} \sup_{\boldsymbol{X}\in\mathcal{X}^n} \sqrt{\mathbb{E}[\|\boldsymbol{X}^\top\boldsymbol{G}\|_F^2]} \le CR\sqrt{\frac{dp}{n}}. \tag{41}$$

■

Corollary 4 *Suppose Assumptions 2&3 hold and input set $\mathcal{X} \subset \mathcal{B}^p(c\sqrt{p})$[2] for a constant $c > 0$. Let $\hat{\boldsymbol{f}}$ be empirical solution of* (4)*. Then, with probability at least $1 - \delta$,*

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

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

Proof This proof is immediately done by following Theorem 1 and Lemma 6. Since $\mathcal{X} \subset \mathcal{B}^p(c\sqrt{p})$, and we assume $\Psi_\ell, \ell \in [L]$ have bounded operator norm C, for each layer, the input space $\mathcal{X}_{\Psi_\ell} \subset \mathcal{B}^{p_{\ell-1}}(C^{\ell-1}c\sqrt{p})$ and then following Lemma 6, $\widetilde{\mathcal{G}}_{NT}(\Psi_\ell) \leq C^\ell c\sqrt{p}\sqrt{\frac{p_\ell p_{\ell-1}}{NT}}$, $\ell \in [L]$. Since $\mathcal{H} = \mathcal{B}^{p_L}(C)$ and $\mathcal{X}_{\mathcal{H}} \subset \mathcal{B}^{p_L}(C^L c\sqrt{p})$, we have $\widetilde{\mathcal{G}}_N(\mathcal{H}) \leq C^{L+1}c\sqrt{p}\sqrt{\frac{p_L}{N}}$. Then we obtain

$$\widetilde{\mathcal{G}}_N(\mathcal{H}) + \sum_{\ell=1}^{L} \sqrt{K_\ell}\widetilde{\mathcal{G}}_{NT}(\Psi_\ell) \lesssim c\sqrt{p} \cdot \sqrt{\frac{C^{L+1} \cdot T \cdot p_L + \sum_{\ell=1}^{L} C^\ell \cdot K_\ell \cdot p_\ell \cdot p_{\ell-1}}{NT}}.$$

Combining it with Theorem 1 finishes the proof. ■

D.2 Proof of Theorem 3

Corollary 4 directly follows by applying Theorem 1 to the linear representation setting, and therefore $\lesssim$ subsumes dependencies on $\log NT$ and Γ^L. Instead in Theorem 3 we establish a tighter bound for parametric hypothesis classes and the sample complexity is only logarithmic in the input space radius R ($R = c\sqrt{p}$ in Corollary 4) and linearly dependent on the number of layers L.

Proof The theorem is a direct application of Theorem 8 after verifying the assumptions. First, bounded loss function $\ell : \mathbb{R} \times \mathbb{R} \to [0, 1]$ implies it is sub-exponential, which verifies Assumption 7. Second, all module/head functions have bounded spectral/Euclidean norms, which verifies Assumption 8. One remark (compared to Theorem 8) is that, since the loss function is bounded, by applying Hoeffding's inequality, (30) holds without enforcing a lower bound constraint on N_{tot}. ■

D.3 The Need for Well-Populated Source Tasks in Assumption 5

Lemma 7 *Consider a weaker version of Assumption 5 where we enforce $\boldsymbol{\Sigma}_\alpha \succeq c\boldsymbol{I}_{p_L}$ over all clusters with two or more tasks*[3] *(i.e. only when $\gamma_\alpha \geq 2/p_L$). Then, there exists a (*(M^2TL)*,* (TLOP)*) problem pair such that the excess transfer learning risk obeys $\mathcal{R}_{TLOP}(\hat{f}_{\hat{\phi}}) \geq 1$ as soon as $N \geq p$.*

Proof The idea is packing supernet with isolated MTL tasks that are uncorrelated with target while achieving zero MTL risk. We consider a simple supernet construction where T tasks will be processed in parallel and all layers have exactly $K_\ell = T$ modules. Specifically, task t will use the pathway $\alpha_t = [t, t, \ldots, t]$ by selecting tth module from each layer. This way each task will use a unique pathway and supernet will be fully occupied. Set noise level $\sigma = 0$. Observe that as soon as $N \geq p$, $\boldsymbol{\theta}_t^\star$ minimizes both empirical and population risks. Consequently, for any $\|\bar{\boldsymbol{h}}_t\| = 1$, $\boldsymbol{B}_{\alpha_t} = \boldsymbol{h}_t(\boldsymbol{\theta}_t^\star)^\top$ is a valid (and minimum norm) minimizer of empirical and population risks. Here, we highlight the minimum norm aspect because this solution is what gradient descent would converge during MTL phase (while we acknowledge the existence of infinitely-many solutions) (Ji and Telgarsky 2018). To wrap up the proof, suppose transfer task is orthogonal to all source tasks and observe that, regardless of the transfer prediction head $\hat{\boldsymbol{h}}_{\mathcal{T}}$ and pathway choice t, we have

$$\begin{aligned}\mathcal{R}_{\text{TLOP}}(\hat{f}_{\hat{\phi}}) &= \mathbb{E}\left[(y - \hat{f}_{\hat{\phi}}(\boldsymbol{x}))^2\right] = \mathbb{E}\left[(\boldsymbol{\theta}_{\mathcal{T}}^\top \boldsymbol{x} - \hat{\boldsymbol{h}}_{\mathcal{T}}^\top \boldsymbol{h}_t(\boldsymbol{\theta}_t^\star)^\top \boldsymbol{x})^2\right] \\ &\geq \|\boldsymbol{\theta}_{\mathcal{T}} - (\hat{\boldsymbol{h}}_{\mathcal{T}}^\top \boldsymbol{h}_t)\boldsymbol{\theta}_t^\star\|^2 \geq \|\boldsymbol{\theta}_{\mathcal{T}}\|^2 = 1.\end{aligned}$$

This concludes the proof. We note that, if $\sigma \neq 0$ same argument would work as $N \to \infty$. Additionally, through same argument with $\sigma = 0$, it can be observed that, a more general lower bound on excess transfer risk is $\min_{t\in[T]} \|\boldsymbol{\theta}_{\mathcal{T}} - \boldsymbol{\theta}_t^\star\|^2/2$. ■

[2] Observe that, this input space is rich enough to capture a random vector with $\mathcal{O}(1)$ subgaussian norm. For instance, a standard normal vector would fall into this set with exponentially high probability as soon as $c > 1$.

[3] The relaxation is not enforcing anything on pathways containing a single task.

D.4 Proof of Theorem 4 and Supporting Results

We start with a useful lemma to show excess risk of linear least squares problem with dependent noise.

Lemma 8 (Linear least squares risk with dependent noise) *Let $\mathcal{S} = (\boldsymbol{x}_i, y_i)_{i=1}^n \overset{i.i.d.}{\sim} \mathcal{D}$ where $y = \boldsymbol{\theta}^\top \boldsymbol{x} + z$ where $\boldsymbol{x}$ is $\mathcal{O}(1)$ subgaussian vector with isotropic covariance and z is $\mathcal{O}(\sigma)$ subgaussian noise. Here, we assume that $\boldsymbol{x}\&z$ can be dependent, however, orthogonal (i.e. $\mathbb{E}[\boldsymbol{x}z] = 0$). Let $\boldsymbol{X} = [\boldsymbol{x}_1 \ \cdots \ \boldsymbol{x}_n]^\top \in \mathbb{R}^{n\times p}$ and $\boldsymbol{X}^\dagger$ be the Moore-Penrose pseudoinverse of $\boldsymbol{X}$. Let $\wedge$ be the minimum symbol. For $n \geq Cp$ for a sufficiently large constant $C \geq 1$, the excess least squares risk and population-empirical risk gap of $\hat{\boldsymbol{\theta}} = \boldsymbol{X}^\dagger \boldsymbol{y}$ is given by*

$$\mathcal{L}_{\mathcal{D}}(\hat{\boldsymbol{\theta}}) - \sigma^2 \leq C\sigma^2 \frac{p+t}{n} \qquad \text{with probability at least } 1 - e^{-cn} - 2e^{-\sqrt{tn}\wedge t} \tag{42}$$

$$\mathcal{L}_{\mathcal{D}}(\hat{\boldsymbol{\theta}}) - \widehat{\mathcal{L}}_{\mathcal{S}}(\hat{\boldsymbol{\theta}}) \leq C\sigma^2 \left(\frac{p}{n} + \sqrt{\frac{t}{n}}\right) \qquad \text{with probability at least } 1 - 2e^{-cn} - 4e^{-\sqrt{tn}\wedge t}. \tag{43}$$

Proof Let $\boldsymbol{z} = [z_1 \ \ldots \ z_n]^\top$ and $\sigma_{\min}(\cdot), \sigma_{\max}(\cdot)$ return the smallest and biggest singular value of a matrix. We can write

$$\mathcal{L}_{\mathcal{D}}(\hat{\boldsymbol{\theta}}) - \mathbb{E}[z^2] = \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|^2 = \|(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{z}\|^2 \leq \frac{\|\boldsymbol{X}^\top \boldsymbol{z}\|^2}{\sigma_{\min}(\boldsymbol{X})^4}.$$

Following (Vershynin 2010), we have $\sqrt{2n} \geq \sigma_{\max}(\boldsymbol{X}), \sigma_{\min}(\boldsymbol{X}) \geq \sqrt{n/2}$ each with probability at least $1 - e^{-cn}$. The crucial term of interest is $\|\boldsymbol{X}^\top \boldsymbol{z}\|$. To control this, observe that $\boldsymbol{X}^\top \boldsymbol{z} = \sum_{i=1}^n z_i \boldsymbol{x}_i$. Since $z_i \boldsymbol{x}_i$ is $\mathcal{O}(\sigma)$-subexponential (multiplication of two subgaussians), the summand $\boldsymbol{X}^\top \boldsymbol{z}$ has a mixed subgaussian/subexponential tail. Specifically, it obeys (Oymak 2018, Lemma D.7)

$$\mathbb{P}\left(\|\boldsymbol{X}^\top \boldsymbol{z}\|^2 \gtrsim \sigma^2(p+t)n\right) \leq 2e^{-\sqrt{tn}\wedge t}.$$

Combining both, with advertised probability we establish the first claim.

$$\frac{\|\boldsymbol{X}^\top \boldsymbol{z}\|^2}{\sigma_{\min}(\boldsymbol{X})^4} \lesssim \sigma^2 \frac{p+t}{n}.$$

For the second claim, observe that

$$\widehat{\mathcal{L}}_{\mathcal{S}}(\hat{\boldsymbol{\theta}}) - \frac{1}{n}\|\boldsymbol{z}\|^2 = \frac{1}{n}\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2 - \frac{1}{n}\|\boldsymbol{z}\|^2 = \frac{1}{n}[\|(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{X}^\dagger)\boldsymbol{z}\|^2 - \|\boldsymbol{z}\|^2] \tag{44}$$

$$= \frac{1}{n}\|\boldsymbol{X}\boldsymbol{X}^\dagger \boldsymbol{z}\|^2 \leq \frac{\sigma_{\max}(\boldsymbol{X})^2}{n}\|(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{z}\|^2 \tag{45}$$

$$\leq 2\|(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{z}\|^2 \lesssim \sigma^2 \frac{p+t}{n}. \tag{46}$$

Here, the first and second inequalities of last line hold with respective probabilities at least $1 - e^{-cn}$ and $1 - e^{-cn} - 2e^{-\sqrt{tn}\wedge t}$. Additionally, since z^2 is $\mathcal{O}\left(\sigma^2\right)$-subexponential, $|\frac{1}{n}\|\boldsymbol{z}\|^2 - \mathbb{E}[z^2]| \lesssim \sigma^2\sqrt{t/n}$ with probability at least $1 - 2e^{-\sqrt{tn}\wedge t}$. Combining all provides the final equation bounding the gap between empirical and population risks. ■

Then we present the following lemma that converts an MTL guarantee into a transfer learning guarantee on a single subspace.

Lemma 9 *Let $\boldsymbol{B} \in \mathbb{R}^{r\times p}$ be a matrix with orthonormal rows and fix $\{\boldsymbol{h}_t\}_{t=1}^T \in \mathbb{R}^r$ with unit covariance and declare distributions $(\boldsymbol{x}, y) \sim \mathcal{D}_t$ obeying $y = \boldsymbol{h}_t^\top \boldsymbol{B}\boldsymbol{x} + z$ with $\mathbb{E}[z^2] = \sigma^2$ and $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] = \boldsymbol{I}_p$. Form $\boldsymbol{H} = [\boldsymbol{h}_1 \ \ldots \ \boldsymbol{h}_t]^\top$ and assume $C\frac{1}{r}\boldsymbol{I}_r \succeq \frac{1}{T}\boldsymbol{H}^\top \boldsymbol{H} \succeq c\frac{1}{r}\boldsymbol{I}_r$. Now, for some $\varepsilon > 0$, suppose that $\hat{\boldsymbol{f}} = (\hat{\boldsymbol{B}}, \{\hat{\boldsymbol{h}}_t\}_{t=1}^T)$ with orthonormal $\hat{\boldsymbol{B}}$ achieves small population risk in average that is*

$$\mathcal{L}_{\bar{\mathcal{D}}}(\hat{\boldsymbol{f}}) - \mathcal{L}_{\bar{\mathcal{D}}}(\boldsymbol{f}_\star) = \frac{1}{T}\sum_{t=1}^T \mathbb{E}_{\mathcal{D}_t}[(\boldsymbol{h}_t^\top \boldsymbol{B}\boldsymbol{x} - \hat{\boldsymbol{h}}_t^\top \hat{\boldsymbol{B}}\boldsymbol{x})^2] \leq \varepsilon$$

where $\mathcal{L}_{\bar{\mathcal{D}}}(\boldsymbol{f}_\star) = \sigma^2$ is the optimal risk achieved by $\boldsymbol{f}_\star = (\boldsymbol{B}, \{\boldsymbol{h}_t\}_{t=1}^T)$. Let $\mathcal{D}_{\mathcal{T}}$ be a new distribution with $y = \boldsymbol{h}_{\mathcal{T}}^\top \boldsymbol{B}\boldsymbol{x} + z$ where $\boldsymbol{x}, z$ are independent $\mathcal{O}(1), \mathcal{O}(\sigma)$ subgaussian respectively and $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] = \boldsymbol{I}_p$. With probability at least $1 - e^{-cM} - 2e^{-\sqrt{tM}\wedge t}$, the transfer learning risk on $\hat{\boldsymbol{B}}$ with M samples is bounded as

$$\mathcal{L}_{\mathcal{T}}(\hat{f}) - \sigma^2 \lesssim r\varepsilon \wedge 1 + C\frac{r+t}{M}.$$

where $\wedge$ is the minimum symbol. Additionally, if target task vector $\boldsymbol{h}_{\mathcal{T}}$ is uniformly drawn from unit Euclidean sphere, in expectation over $\boldsymbol{h}_{\mathcal{T}}$ and in probability over target training datasets (with probability at least $1-e^{-cM}-2e^{-\sqrt{tM}\wedge t}$), we have the tighter bound

$$\mathbb{E}_{\boldsymbol{h}_{\mathcal{T}}}[\mathcal{L}_{\mathcal{T}}(\hat{f})]-\sigma^2 \lesssim \varepsilon + C\frac{r+t}{M}.$$

Finally, in both cases, population-empirical transfer gaps $|\mathcal{L}_{\mathcal{T}}(\hat{f})-\widehat{\mathcal{L}}_{\mathcal{S}_{\mathcal{T}}}(\hat{f})|$, $\mathbb{E}_{\boldsymbol{h}_{\mathcal{T}}}[|\mathcal{L}_{\mathcal{T}}(\hat{f})-\widehat{\mathcal{L}}_{\mathcal{S}_{\mathcal{T}}}(\hat{f})|]$ are bounded by $\mathcal{O}\left(\frac{r+t}{M}\right)$ with same probability.

Proof Let $\boldsymbol{\theta}_t = \boldsymbol{B}^\top \boldsymbol{h}_t$ and $\hat{\boldsymbol{\theta}}_t = \hat{\boldsymbol{B}}^\top \hat{\boldsymbol{h}}_t$. We first observe that task t risk is simply

$$\mathcal{L}_t(\hat{\boldsymbol{\theta}}_t) = \mathbb{E}_{\mathcal{D}_t}[(y-\hat{\boldsymbol{h}}_t^\top \hat{\boldsymbol{B}}\boldsymbol{x})^2] = \sigma^2 + \mathbb{E}_{\mathcal{D}_t}[(\boldsymbol{\theta}_t\boldsymbol{x}-\hat{\boldsymbol{\theta}}_t\boldsymbol{x})^2] = \sigma^2 + \|\hat{\boldsymbol{\theta}}_t-\boldsymbol{\theta}_t\|^2.$$

Thus, the excess MTL risk is simply

$$\mathcal{L}_{\bar{\mathcal{D}}}(\hat{\boldsymbol{f}}) - \mathcal{L}_{\bar{\mathcal{D}}}(\boldsymbol{f}_\star) = \frac{1}{T}\|\boldsymbol{\Theta}-\hat{\boldsymbol{\Theta}}\|_F^2 \le \varepsilon,$$

where $\boldsymbol{\Theta}, \hat{\boldsymbol{\Theta}} \in \mathbb{R}^{T\times p}$ are the concatenated task vectors.

Now, we aim to obtain the transfer learning risk over $\hat{\boldsymbol{B}}$. We first write the target regression task $(y,\boldsymbol{x}) \sim \mathcal{D}_{\mathcal{T}}$ with $y = \boldsymbol{x}^\top\boldsymbol{\theta}_{\mathcal{T}} + z$ (for some $\boldsymbol{h}$) as

$$y = \boldsymbol{x}^\top \hat{\boldsymbol{B}}^\top \boldsymbol{h} + z + \boldsymbol{x}^\top \Pi_{\hat{\boldsymbol{B}}^\perp}(\boldsymbol{\theta}_{\mathcal{T}}). \tag{47}$$

Here set $\boldsymbol{x}' = \hat{\boldsymbol{B}}\boldsymbol{x}$ and $z' = \boldsymbol{x}^\top(\boldsymbol{I}-\hat{\boldsymbol{B}}^\top\hat{\boldsymbol{B}})\boldsymbol{\theta}_{\mathcal{T}}$, and then $y = \boldsymbol{h}^\top\boldsymbol{x}' + z + z'$. Note that

$$\mathbb{E}[\boldsymbol{x}'z'] = \mathbb{E}[\hat{\boldsymbol{B}}\boldsymbol{x}\boldsymbol{x}^\top(\boldsymbol{I}-\hat{\boldsymbol{B}}^\top\hat{\boldsymbol{B}})\boldsymbol{\theta}_{\mathcal{T}}] = 0,$$

verifying that we can treat the representation mismatch as a dependent but orthogonal subgaussian noise. Combined with Lemma 8, with probability $1-e^{-cM}-2e^{-\sqrt{tM}\wedge t}$ (conditioned on $\hat{\boldsymbol{B}}$) this leads to a transfer learning risk of

$$\mathcal{L}_{\mathcal{T}}(\hat{f}) - \sigma^2 \le \|\Pi_{\hat{\boldsymbol{B}}^\perp}(\boldsymbol{\theta}_{\mathcal{T}})\|^2 + C\frac{r+t}{M}.$$

Following the proof of Lemma 8, the σ^2 term on the right hand side of (42) is related to the inputs ($\boldsymbol{x}'$) and noise ($z+z'$) levels, which are $\mathcal{O}(1)$.

To proceed, observe that $\mathbb{E}[z'^2] = \|\Pi_{\hat{\boldsymbol{B}}^\perp}(\boldsymbol{\theta}_{\mathcal{T}})\|^2 = \|\boldsymbol{B}\boldsymbol{\theta}_{\mathcal{T}}\|^2 - \|\hat{\boldsymbol{B}}\boldsymbol{\theta}_{\mathcal{T}}\|^2 = \boldsymbol{\theta}_{\mathcal{T}}^\top(\boldsymbol{B}^\top\boldsymbol{B}-\hat{\boldsymbol{B}}^\top\hat{\boldsymbol{B}})\boldsymbol{\theta}_{\mathcal{T}}$. In the worst case, this risk is equal to

$$\sup_{\|\boldsymbol{\theta}\|=1, \boldsymbol{B}^\top\boldsymbol{B}\boldsymbol{\theta}=\boldsymbol{\theta}} \|\Pi_{\hat{\boldsymbol{B}}^\perp}(\boldsymbol{\theta})\|^2 = \|\boldsymbol{B}^\top\boldsymbol{B}-\hat{\boldsymbol{B}}^\top\hat{\boldsymbol{B}}\|.$$

Recall that we are given $\frac{1}{T}\|\boldsymbol{\Theta}-\hat{\boldsymbol{\Theta}}\|^2 \le \frac{1}{T}\|\boldsymbol{\Theta}-\hat{\boldsymbol{\Theta}}\|_F^2 \le \varepsilon$. Additionally, $\boldsymbol{\Theta}^\top\boldsymbol{\Theta}/T$ is a well-conditioned matrix over the subspace Range($\boldsymbol{B}$) with minimum nonzero eigenvalue at least $c/r > 0$ and condition number upper bounded by C/c (equal to that of $\boldsymbol{H}$). If $\varepsilon \le c/2r$, this also implies $\lambda_{\min}(\hat{\boldsymbol{\Theta}}^\top\hat{\boldsymbol{\Theta}}/T) \ge c/2r$ and condition number at most $3C/c$. Consequently, applying Davis-Kahan theorem (Yu, Wang, and Samworth 2015) on $\boldsymbol{\Theta}, \hat{\boldsymbol{\Theta}}$ pair implies that the eigenspaces $\boldsymbol{B}, \hat{\boldsymbol{B}}$ of $\boldsymbol{\Theta}, \hat{\boldsymbol{\Theta}}$ obey

$$\|\boldsymbol{B}^\top\boldsymbol{B}-\hat{\boldsymbol{B}}^\top\hat{\boldsymbol{B}}\| \le \mathcal{O}\left(\frac{\varepsilon}{c/r}\right) = \mathcal{O}(r\varepsilon).$$

If $\varepsilon r \ge c/2$, we can simply use the tighter estimate $\|\boldsymbol{B}^\top\boldsymbol{B}-\hat{\boldsymbol{B}}^\top\hat{\boldsymbol{B}}\| \le 1$, which completes the proof of first part of the lemma.

Secondly, consider the average case scenario where $\boldsymbol{h}_{\mathcal{T}} \sim$ unif_over_sphere. In this case, we observe that, the target-averaged transfer risk follows

$$\mathbb{E}_{\boldsymbol{\theta}_{\mathcal{T}}}[\min_{\boldsymbol{h}} \|\boldsymbol{\theta}_{\mathcal{T}}-\hat{\boldsymbol{B}}^\top\boldsymbol{h}\|^2] = \mathbb{E}[\|(\boldsymbol{I}-\hat{\boldsymbol{B}}^\top\hat{\boldsymbol{B}})\boldsymbol{\theta}_{\mathcal{T}}\|^2] = \frac{1}{2r}\|\boldsymbol{B}^\top\boldsymbol{B}-\hat{\boldsymbol{B}}^\top\hat{\boldsymbol{B}}\|_F^2. \tag{48}$$

This time, Davis-Kahan theorem yields the tighter estimate (for our purposes) $\frac{1}{2r}\|\boldsymbol{B}^\top\boldsymbol{B}-\hat{\boldsymbol{B}}^\top\hat{\boldsymbol{B}}\|_F^2 \lesssim \frac{1}{T}\|\boldsymbol{\Theta}-\hat{\boldsymbol{\Theta}}\|_F^2 \le \varepsilon$. To proceed, we find that, the expected transfer learning risk over task distribution obey the tighter guarantee (with probability at least $1-e^{-cM}-2e^{-\sqrt{tM}\wedge t}$)

$$\mathbb{E}_{\boldsymbol{h}_{\mathcal{T}}}[\mathcal{L}_{\mathcal{T}}(\hat{f})] - \sigma^2 \lesssim \varepsilon + C\frac{r+t}{M}.$$

The final claim arises as a direct result of our application of Lemma 8 in (47). ■

The following corollary is a Multipath MTL guarantee for least-squares regression obtained by specializing the more general Theorem 8.

Corollary 5 *Suppose $\mathcal{X} \subset \mathcal{B}^p(R)$, $\ell(\hat{y}, y)$ is quadratic, and Assumptions 3&4 hold. Solving* (M^2TL) *with the fixed choice of ground-truth pathways and $NT \gtrsim DoF(\mathcal{F}) \log(NT)$, with probability at least $1-\delta$, we have that*

$$\mathcal{R}_{M^2TL}(\hat{\boldsymbol{f}}) \lesssim \sqrt{\frac{L \cdot DoF(\mathcal{F}) + \log(2/\delta)}{NT}}. \tag{49}$$

Proof We need to verify the assumptions of Theorem 8. Observe that $\boldsymbol{x}, z$ are subgaussian and ground-truth model $\|\boldsymbol{\theta}_t^\star\| \le 1$ and all feasible task hypothesis $\boldsymbol{\theta}$ obeys $\|\boldsymbol{\theta}\| \le C^{L+1}$ which we treat as a constant (i.e. fixed depth L). Consequently, subexponential norm obeys $\|(y - \boldsymbol{\theta}^\top \boldsymbol{x})^2\|_{\psi_1} = \|(z + \boldsymbol{x}^\top(\boldsymbol{\theta} - \boldsymbol{\theta}_\star))^2\|_{\psi_1} \le \mathcal{O}\left(C^{2(L+1)}\right)$ which verifies $\mathcal{O}(1)$ subexponential condition. Similarly, loss function is Lipschitz with $\Gamma = \sup_{y,\boldsymbol{x}} |y - f_t(\boldsymbol{x})| \le 2C^{L+1}R$. Together, these verify Assumption 7. Note that, Theorem 8 has logarithmic dependence on Γ which is subsumed within $\lesssim$. Finally, each module is C Lipschitz (due to spectral norm bounds) and log-covering number of $d \times p$ matrices with C-bounded spectral norm obeys $dp \log(3CR/\varepsilon)$. These two verify Assumption 8. ∎

Finalizing the Proof of Theorem 4 Following the discussion above, we provide a proof of Theorem 4. The result below is a formal restatement of the theorem with a few caveats. First, we state two closely-related guarantees. First guarantee is when target head $\boldsymbol{h}_\mathcal{T}$ is arbitrary (worst-case) and second one is for when it is uniformly distributed over unit sphere (average case). The latter shaves a factor of p_L in the MTL risk term. Second, the probability term in Theorem 4 is chosen to be approximate for notational simplicity. Namely, we ignored the $\log(1/\delta)/NT$ term and second order effects. We state the full dependence here which is a bit more convoluted.

Theorem 9 *Suppose Assumptions 3–6 hold and $\ell(\hat{y}, y) = (y - \hat{y})^2$. Additionally assume input space is $\mathcal{B}^p(c\sqrt{p})$ and $\mathcal{H}_\mathcal{T} = \mathbb{R}^{p_L}$[4]. Solve MTL problem* (M^2TL) *with the knowledge of ground-truth pathways $(\bar{\alpha}_t)_{t=1}^T$ to obtain a supernet $\hat{\boldsymbol{\phi}}$ and assume $NT \gtrsim DoF(\mathcal{F}) \log(NT)$. Solve transfer learning problem* (TLOP) *with $\hat{\boldsymbol{\phi}}$ to obtain a target hypothesis $\hat{f}_{\hat{\boldsymbol{\phi}}}$. Then, with probability at least $1 - 3e^{-cM} - 4\delta$, excess target risk* (3) *of TLOP obeys*

$$\mathbb{E}_{\alpha_\mathcal{T}}[\mathcal{R}_{TLOP}(\hat{f}_{\hat{\boldsymbol{\phi}}})] \lesssim \frac{p_L}{M} + p_L\sqrt{\frac{L \cdot DoF(\mathcal{F}) + \log(2/\delta)}{NT}} + \left[\sqrt{\frac{\log(2|\mathcal{A}|/\delta)}{M}}\right]_+. \tag{50}$$

Here, the probability is over the source datasets and the (input, noise) pairs of the target dataset i.e. $(\boldsymbol{x}_i^\mathcal{T}, z_i^\mathcal{T})_{i=1}^M$, and we used the short hand $[x]_+ = x + x^2$. Additionally, if target distribution follows the same generative model with prediction head $\boldsymbol{h}_\mathcal{T}$ drawn uniformly at random over the unit sphere, we obtain the tighter bound

$$\mathbb{E}_{\alpha_\mathcal{T}, \boldsymbol{h}_\mathcal{T}}[\mathcal{R}_{TLOP}(\hat{f}_{\hat{\boldsymbol{\phi}}})] \lesssim \frac{p_L}{M} + \sqrt{\frac{L \cdot DoF(\mathcal{F}) + \log(2/\delta)}{NT}} + \left[\sqrt{\frac{\log(2|\mathcal{A}|/\delta)}{M}}\right]_+. \tag{51}$$

Remark. Note that, above, we split probability space into three independent variables. Source datasets $\mathcal{S}_{\text{all}}$, (input, noise) pairs of the target dataset i.e. $(\boldsymbol{x}_i^\mathcal{T}, z_i^\mathcal{T})_{i=1}^M$, and finally target path $\alpha_\mathcal{T}$. The result is with high probability over the former two and expectation over the latter.

Proof In this proof, we aim to reduce the Multipath MTL guarantee to a Vanilla MTL scenario so that we can utilize Lemma 9. Assumptions 5 and 6 will be critical towards this goal. Recall that, we have the Multipath MTL guarantee from Corollary 5 so that, with probability $1-\delta$,

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

Let us call this event $\mathcal{E}_1$. Here, we omitted the $\log|\mathcal{A}|/N$ term because our transfer guarantee will require the knowledge of ground-truth pathways for sources (even if it is not required for the target). The main idea is to show that small $\mathcal{R}_{\text{M}^2\text{TL}}(\hat{\boldsymbol{f}})$ implies that target will fall on a pathway with small source-averaged risk. This way, we can apply Lemma 9 to provide a guarantee for the target. To proceed, we gather all unique ground-truth pathways via $\Gamma = \{\gamma_i\}_{i=1}^S$. Additionally, let $C(\gamma)$ be the number of tasks that chooses pathway γ.

Finally let $\mathcal{L}_i'(\boldsymbol{f})$ be the excess task-averaged population risk over γ_i, that is $\mathcal{L}_i'(\boldsymbol{f}) = \frac{1}{C(\gamma_i)} \sum_{\bar{\alpha}_t = \gamma_i} \{\mathcal{L}_t(f_t) - \sigma^2\}$. With this definition, we can write MTL excess risk as

$$\frac{1}{T}\sum_{i=1}^S C(\gamma_i)\mathcal{L}_i'(\boldsymbol{f}) \lesssim \sqrt{\frac{L \cdot \text{DoF}(\mathcal{F})}{NT}} + \sqrt{\frac{\log(2/\delta)}{NT}}.$$

[4] We make this assumption (no norm constraint unlike MTL phase) since during transfer learning, we simply solve least-squares. Thanks to this, we achieve faster rates.

To proceed, we will view each $\mathcal{L}_i'$ as a vanilla MTL problem over pathway γ_i. Following Assumption 6, we draw the random pathway $\alpha_{\mathcal{T}}$ of the target task and it is equal to $\alpha_{\mathcal{T}} = \gamma_i \in \Gamma$. Note that this event happens with probability $\mathbb{P}(\alpha_{\mathcal{T}} = \gamma_i) = C(\gamma_i)/T$. Conditioned on this, let us control the transfer risk.

Note that during TLOP we will search over all pathways $\alpha \in \mathcal{A}$. Denote $\bar{\boldsymbol{B}}_\alpha, \hat{\boldsymbol{B}}_\alpha \in \mathbb{R}^{p_L \times p}$ are the ground-truth and empirical weights of the linear model induced by α. Denote the transfer learning model over $\hat{\boldsymbol{B}}_\alpha$ via $\hat{f}_\alpha$. For any choice of α, applying Lemma 8, we know that empirical-population transfer gap $\mathcal{L}_{\mathcal{D}_{\mathcal{T}}}(\hat{f}_\alpha) - \widehat{\mathcal{L}}_{\mathcal{S}_{\mathcal{T}}}(\hat{f}_\alpha)$ is bounded by $\mathcal{O}\left(\frac{p_L}{M} + \left[\sqrt{\frac{\log(2/\delta)}{M}}\right]_+\right)$ with probability at least $1 - 2e^{-cM} - 2\delta$. This is over the input/noise distribution of target samples (arbitrary $\alpha_{\mathcal{T}} = \gamma_i$ and associated ground-truth $\boldsymbol{\theta}_{\mathcal{T}}$). Union bounding over all potential pathways target task may use, we obtain that,

$$\sup_{\alpha \in \mathcal{A}} |\mathcal{L}_{\mathcal{D}_{\mathcal{T}}}(\hat{f}_\alpha) - \widehat{\mathcal{L}}_{\mathcal{S}_{\mathcal{T}}}(\hat{f}_\alpha)| \leq \mathcal{O}\left(\frac{p_L}{M} + \left[\sqrt{\frac{\log(2|\mathcal{A}|/\delta)}{M}}\right]_+\right).$$

Consequently, empirical risk minimization over all pathways will choose a target model $\hat{f}_{\hat{\boldsymbol{\phi}}}$ guaranteeing with probability at least $1 - 2e^{-cM} - 2\delta$

$$\mathcal{R}_{\text{TLOP}}(\hat{f}_{\hat{\boldsymbol{\phi}}}) \leq \min_\alpha \mathcal{R}_{\text{TLOP}}(\hat{f}_\alpha) + \mathcal{O}\left(\frac{p_L}{M} + \left[\sqrt{\frac{\log(2|\mathcal{A}|/\delta)}{M}}\right]_+\right) \tag{52}$$

$$\leq \mathcal{R}_{\text{TLOP}}(\hat{f}_{\gamma_i}) + \mathcal{O}\left(\frac{p_L}{M} + \left[\sqrt{\frac{\log(2|\mathcal{A}|/\delta)}{M}}\right]_+\right). \tag{53}$$

The latter line is reasonable because we know that ground-truth pathway $\alpha_{\mathcal{T}} = \gamma_i$ is a great candidate for being population minima. Applying Lemma 9 again over the path γ_i, with probability $1 - e^{-cM} - \delta$, we obtain

$$\mathcal{R}_{\text{TLOP}}(\hat{f}_{\gamma_i}) \leq p_L \mathcal{L}_i'(\boldsymbol{f}) \wedge 1 + C\frac{p_L}{M} + \left[\frac{\log(2/\delta)}{M}\right]_+.$$

Combining with above, with probability $1 - 3e^{-cM} - 3\delta$, the ERM solution over all pathways obeys

$$\mathcal{R}_{\text{TLOP}}(\hat{f}_{\hat{\boldsymbol{\phi}}}) \leq p_L \mathcal{L}_i'(\boldsymbol{f}) \wedge 1 + \mathcal{O}\left(\frac{p_L}{M} + \left[\sqrt{\frac{\log(2|\mathcal{A}|/\delta)}{M}}\right]_+\right).$$

Note that above holds for worst-case prediction head $\boldsymbol{h}_{\mathcal{T}}$. Additionally, applying Lemma 9 again and assuming $\boldsymbol{h}_{\mathcal{T}}$ is generated uniformly over the unit sphere, on the same event, we find

$$\mathbb{E}_{\boldsymbol{h}_{\mathcal{T}}}[\mathcal{R}_{\text{TLOP}}(\hat{f}_{\hat{\boldsymbol{\phi}}})] \leq \mathcal{L}_i'(\boldsymbol{f}) + \mathcal{O}\left(\frac{p_L}{M} + \left[\sqrt{\frac{\log(|\mathcal{A}|/\delta)}{M}}\right]_+\right). \tag{54}$$

Now, for fixed MTL dataset, taking expectation over $\alpha_{\mathcal{T}}$, with same probability over the input/noise distribution

$$\mathbb{E}_{\alpha_{\mathcal{T}}}[\mathcal{R}_{\text{TLOP}}(\hat{f}_{\hat{\boldsymbol{\phi}}})] \leq \sum_{i=1}^{S} \frac{C(\gamma_i)}{T} p_L \mathcal{L}_i'(\boldsymbol{f}) \wedge 1 + \mathcal{O}\left(\frac{p_L}{M} + \left[\sqrt{\frac{\log(2|\mathcal{A}|/\delta)}{M}}\right]_+\right) \tag{55}$$

$$\leq p_L \mathcal{R}_{\text{M}^2\text{TL}}(\hat{\boldsymbol{f}}) \wedge 1 + \mathcal{O}\left(\frac{p_L}{M} + \left[\sqrt{\frac{\log(2|\mathcal{A}|/\delta)}{M}}\right]_+\right). \tag{56}$$

Let us call this event $\mathcal{E}_2$ which is independent of $\alpha_{\mathcal{T}}$. Union bounding $\mathcal{E}_1$ and $\mathcal{E}_2$ (both independent of $\alpha_{\mathcal{T}}$), we obtain the advertised bound (50). We obtain (51) through same argument following the average-case control (54). ■

E Not All Optimal MTL Pathways are Good for Transfer Learning

Ideally we would like to prove Theorem 4 without assuming that MTL phase is solved with the knowledge of ground-truth pathways. While we believe this may be possible under stronger assumptions, here, we discuss why this problem is pretty challenging with a simple example on linear representations.

Setting: Suppose we have a single layer linear supernet with $K = 2$ modules each with size $2R \times p$. This corresponds to the Cluster MTL model where we simply wish to group the tasks into two clusters and train vanilla MTL over individual clusters. This simple setting will already highlight the issue.

• **Source tasks:** Consider four groups of tasks $(\boldsymbol{\Theta}_i)_{i=1}^4$ where $\boldsymbol{\Theta}_i = (\boldsymbol{\theta}_{ij})_{j=1}^{T/4}$. We assume that $T/4$ task vectors from $\boldsymbol{\Theta}_i$ perfectly span an R dimensional subspace S_i (at least $T \geq 4R$). Additionally, set $(S_i)_{i=1}^4$ to be perfectly orthogonal over $\mathbb{R}^p$. Also assume that the tasks are linear and noiseless i.e. $y_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\theta}_{ij}$.

Lemma 10 *Suppose representation modules $\boldsymbol{B}_1, \boldsymbol{B}_2 \in \mathbb{R}^{2R\times p}$ are constrained to have orthonormal rows. Define the ground-truth pathways where $\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2$ are on pathway 1 and $\boldsymbol{\Theta}_3, \boldsymbol{\Theta}_4$ are on pathway 2. Now, assume that transfer learning task $\boldsymbol{\theta}_\mathcal{T}$ is drawn uniformly at random from the $2R$ dimensional subspace of one of these pathways. Assume target is linear & isotropic: $(\boldsymbol{x}, y) \sim \mathcal{D}_\mathcal{T}$ obeys $y = \boldsymbol{x}^\top \boldsymbol{\theta}_\mathcal{T} + z$ where $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] = \boldsymbol{I}_p$ and $\boldsymbol{x}, z$ are orthogonal. Then, regardless of the source sample/task sizes N, T and target sample size M, there exists an MTL solution such that, excess transfer risk of final target hypothesis $\hat{f}_\mathcal{T}$ obeys*

$$\mathbb{E}_{\boldsymbol{\theta}_\mathcal{T}}[\mathcal{R}_{TLOP}(\hat{f}_\mathcal{T})] \geq c.$$

for some absolute constant $c > 0$. Additionally, $\mathcal{R}_{TLOP}(\hat{f}_\mathcal{T}) \geq 0.5$ almost surely as $R \to \infty$.

Proof As the reader might have noticed, the argument is straightforward. Create the following MTL solution: Let $\boldsymbol{B}_1$ be an orthonormal basis for $\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_3$ and let $\boldsymbol{B}_2$ be an orthonormal basis for $\boldsymbol{\Theta}_2, \boldsymbol{\Theta}_4$. Without losing generality, for $\boldsymbol{B}_1$, let us set it so that first R rows are assigned to $\boldsymbol{\Theta}_1$ and last R assigned to $\boldsymbol{\Theta}_3$ (same for $\boldsymbol{B}_2$). Note that, we simply swapped $\boldsymbol{\Theta}_2$ with $\boldsymbol{\Theta}_3$ in pathway assignments.

Observe that $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ achieves zero MTL risk because they contain all task vectors $\boldsymbol{\Theta}_i = (\boldsymbol{\theta}_{ij})_{j=1}^{T/4}$ in their range and problems are noiseless. What remains to show is that $\boldsymbol{B}_1, \boldsymbol{B}_2$ assignments are poor choices for the target task drawn from either $\boldsymbol{B}_1^\star$ induced by $\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2$ or $\boldsymbol{B}_2^\star$ induced by $\boldsymbol{\Theta}_3, \boldsymbol{\Theta}_4$. Without losing generality, suppose $\boldsymbol{\theta}_\mathcal{T}$ is drawn from $\boldsymbol{B}_1^\star$. Observing $\boldsymbol{\theta}_\mathcal{T}$ lies on the combined range of $\boldsymbol{B}_1, \boldsymbol{B}_2$, and using properties of linear regression with isotropic features, we bound the target transfer risk via

$$\begin{aligned}
\mathcal{L}_\mathcal{T}(\hat{f}_\mathcal{T}) - \mathbb{E}[z^2] &= \min_{i\in\{1,2\}} \mathcal{L}_\mathcal{T}(\boldsymbol{B}_i^\top \hat{\boldsymbol{h}}_\mathcal{T}) - \mathbb{E}[z^2] \\
&= \min_{i\in\{1,2\}} \|\boldsymbol{B}_i^\top \hat{\boldsymbol{h}}_\mathcal{T} - \boldsymbol{\theta}_\mathcal{T}\|^2 \\
&\geq \min_{i\in\{1,2\}} \min_{\boldsymbol{h}} \|\boldsymbol{B}_i^\top \boldsymbol{h} - \boldsymbol{\theta}_\mathcal{T}\|^2 \\
&= \min_{i\in\{1,2\}} \|\boldsymbol{B}_{3-i}\boldsymbol{\theta}_\mathcal{T}\|^2 = \min_{i\in\{1,2\}} \|\boldsymbol{B}_i\boldsymbol{\theta}_\mathcal{T}\|^2 \\
&= \min_{i\in\{1,2\}} \|\text{Proj}_{\mathcal{S}_i}(\boldsymbol{\theta}_\mathcal{T})\|^2.
\end{aligned}$$

The last line highlights the fact that S_i lies on $\boldsymbol{B}_i$ and projection of $\boldsymbol{\theta}_\mathcal{T}$ on $\boldsymbol{B}_i$ is exactly equal to its projection on S_i by pathway assignments. Since $\boldsymbol{\theta}_\mathcal{T}$ is uniformly drawn, the last line is equivalent to $X(\boldsymbol{g}, \boldsymbol{g}') = \frac{\|\boldsymbol{g}\|^2}{\|\boldsymbol{g}\|^2+\|\boldsymbol{g}'\|^2} \wedge \frac{\|\boldsymbol{g}'\|^2}{\|\boldsymbol{g}\|^2+\|\boldsymbol{g}'\|^2}$ for $\boldsymbol{g}, \boldsymbol{g}' \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \boldsymbol{I}_R)$. Observing $\|\boldsymbol{g}\|^2, \|\boldsymbol{g}'\|^2$ are Chi-squared, it is clear that, for all R and for some constant $c_0 > 0$, we have $\mathbb{P}(0.5 \leq \frac{\|\boldsymbol{g}\|^2}{R} \leq 1.5) \geq c_0$. On these events on $\boldsymbol{g}, \boldsymbol{g}'$, we have $X(\boldsymbol{g}, \boldsymbol{g}') \geq 1/4$ and $\mathbb{E}[X(\boldsymbol{g}, \boldsymbol{g}')] \geq c = c_0^2/4$. Finally, as dimension $R \to \infty$, we have $\|\boldsymbol{g}\|^2/\|\boldsymbol{g}'\|^2 \to 1$ almost surely, which similarly implies $X(\boldsymbol{g}, \boldsymbol{g}') \to 0.5$. ■

F Experimental Details on Section 5

We provide further details on the experiments in Section 5 as well as incorporate additional experiments.

F.1 Algorithms for Vanilla MTL, Cluster MTL, and Multipath MTL

To facilitate faster and more stable convergence of all three algorithms, we used a conventional approach from nonconvex optimization literature which has also been proposed in the context of linear representation learning (Kong et al. 2020a; Sun et al. 2021; Bouniot et al. 2020; Tripuraneni, Jin, and Jordan 2021). Specifically, linear representation learning with Vanilla MTL has a bilinear form similar to matrix factorization. Thus, first-order method to solve Vanilla MTL benefits from proper initialization of the representation. In our experiments, we use such a two-step procedure:

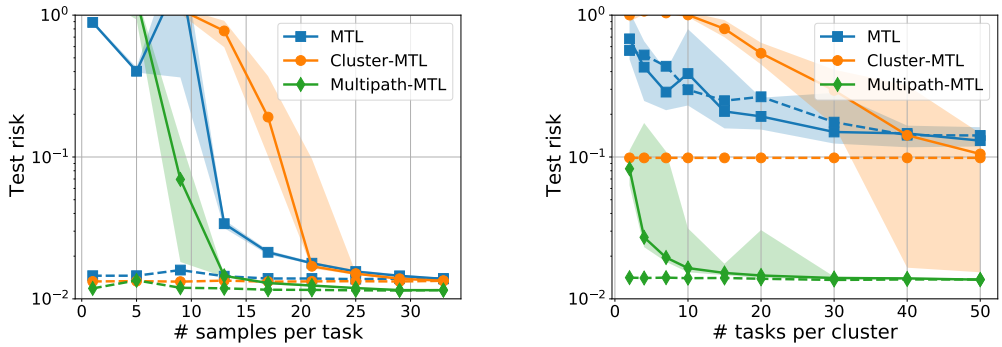
- **Initialization:** At the start of MTL, build an initialization for the representation.
- **Alternating least-squares (ALS):** Train prediction heads and representation layers through alternating least-squares.

Here, we note that ALS is same as alternating gradient descent (AGD) however we are essentially running infinitely many gradient iterations before alternating. The reason we use this procedure for all three algorithms is to provide a fair comparison without the worry of tuning learning rates for each algorithm individually. Initialization plays a useful role in further stabilizing ALS.

While prior works provide initialization methods for MTL, we will also develop a novel initialization algorithm for Multipath MTL. We believe this may be an interesting future direction for providing provable computational guarantees for Multipath MTL.

Initialization procedures: We first revise the procedure for Vanilla MTL. Suppose we are given T tasks with dataset $\mathcal{S}_{\text{all}}$ where input features have isotropic covariance. We will use the procedure discussed in (Sun et al. 2021) where the authors claim improvement over (Tripuraneni, Jin, and Jordan 2021; Kong et al. 2020b).

- **Vanilla MTL:** initialization is a method-of-moments procedure as follows:



(a) Varying N_1 with $\bar{T} = 10, K = 20, N_2 = 33$ (b) Varying $\bar{T}_1$ with $N = 10, K = 20, \bar{T}_2 = 50$

Figure 5: We evaluate (Vanilla) MTL, Cluster-MTL and Multipath-MTL using imbalanced data in a linear regression setting (half tasks have more, half have less data resources). In both experiments, there are $K = 20$ clusters in total. In Fig. 5a, all clusters have $\bar{T} = 10$ tasks. Then, we fix the sample size of the tasks in 10 of the clusters ($N_2 = 33$), and change the sample size in the other 10 clusters (N_1) from 1 to 33. Then we solve the three MTL (Vanilla, Cluster, Multipath) problems. Solid/Dashed curves show the test risk of the tasks who have fewer/more samples. In Fig. 5b, instead we fix the sample size of each task ($N = 10$). In 10 of the clusters, we set the number of tasks $\bar{T}_2 = 50$. While in the other clusters, the number of tasks is varied from 2 to 50. Again, we run experiments under all the three settings and plot the test risk of fewer/more tasks in solid/dashed curves. The curves in both figures show the median risks and the shaded regions highlight the first and third quantile risks. Each marker is average of 20 independent runs.

1. Form the $\hat{\boldsymbol{\theta}}_t$ estimates via $\hat{\boldsymbol{\theta}}_t = \frac{1}{N}\sum_{i=1}^{N} y_{ti}\boldsymbol{x}_{ti}$.
2. Form the moment matrix $\boldsymbol{M} = \sum_{t=1}^{T} \hat{\boldsymbol{\theta}}_t \hat{\boldsymbol{\theta}}_t^\top$.
3. Set $\hat{\boldsymbol{B}}_0 \in \mathbb{R}^{R\times p}$ to be the top R eigenvectors of $\boldsymbol{M}$.

At this point, we can start running our favorite choice of first order method starting from the initialization $\hat{\boldsymbol{B}}_0$. In our implementation, we run ALS where we estimate $\{\hat{\boldsymbol{h}}_t\}_{t=1}^T$ (by fitting LS given $\hat{\boldsymbol{B}}$), then re-estimate $\hat{\boldsymbol{B}}$ and keep going.

• **Cluster MTL:** In our experiments, we assumed clusters (i.e. pathways) are known. This is in order to decouple the challenge of task-clustering from the comparisons in Figure 2. We note that task clustering has been studied by (Fifty et al. 2021; Kumar and Daume III 2012; Kang, Grauman, and Sha 2011) (Leveraging relations between tasks are explored even more broadly (Zhuang et al. 2020).) however these works don't come with comparable statistical guarantees. In our setup, Cluster MTL simply runs K Vanilla MTL algorithms in parallel over individual clusters using ground-truth pathways.

• **Multipath MTL:** We propose an initialization algorithm which is inspired from the Vanilla MTL algorithm as follows. Again, we assume knowledge of clustering/pathways.

1. Estimate shared first layer $\hat{\boldsymbol{B}}_1 \in \mathbb{R}^{R\times p}$ via **Vanilla MTL** initialization using all data.
2. Estimate cluster-specific representations $(\tilde{\hat{\boldsymbol{B}}}_2^k)_{k=1}^K \in \mathbb{R}^{r\times p}$ via **Vanilla MTL** initialization over each cluster data.
3. Estimate the second layer $(\hat{\boldsymbol{B}}_2^k)_{k=1}^K \in \mathbb{R}^{r\times R}$ by projecting $\tilde{\hat{\boldsymbol{B}}}_2^k$ onto the R-dimensional first layer as follows

$$\hat{\boldsymbol{B}}_2^k = \tilde{\hat{\boldsymbol{B}}}_2^k \hat{\boldsymbol{B}}_1^\top.$$

We then run ALS where we go in the order: Prediction heads, second layers, first layer (repeat).

Remark on unknown clusters: We note that a simple approach to identifying clusters when they are unknown is by solving Vanilla MTL and then clustering the resulting weight vectors $\{\hat{\boldsymbol{\theta}}_t\}_{t=1}^T$ of the Vanilla MTL solution (e.g. via K-means). The reason is that, the ground-truth weights $\{\boldsymbol{\theta}_t^\star\}_{t=1}^T$ are simply points that lie on r-dimensional latent cluster-subspaces that we would like to recover. Naturally, the (random) points on the same subspace will have higher correlation. This viewpoint (restricted to linear setting) also connects well with the broader subspace clustering literature where each learning task is a point on a high-dimensional subspace (Vidal 2011; Parsons, Haque, and Liu 2004; Elhamifar and Vidal 2013). The challenge in our setting is we only get to see the points through the associated datasets. Figure 3 shows our results assuming unknown source pathways.

In the next section, we discuss a few more experiments comparing these three approaches.

F.2 Additional Numerical Experiments

In Figure 5, we conduct more experiments to see how tasks with less data resources perform in MTL when trained together with other tasks which have more resources. Here, by resources we either mean a task having more samples N or a task having other

(related) tasks along its pathway/cluster. Thus, our experiments involve imbalanced training data. We consider two experimental settings to show how Multipath MTL benefits accuracy compared to the other two MTL models: Vanilla MTL and Cluster MTL.
Experimental settings: Consider the same Vanilla MTL, Cluster MTL and Multipath MTL problems in linear regression regime as discussed in Section 5 and follow the same algorithm in Section F.1. In the experiments, same as Section 5, we set $p = 32$, $R = 8$, and $r = 2$. We consider MTL problem with $K = 20$ clusters. Here, data is noisy. In Fig. 5a, there are 10 tasks in each cluster. In half of the clusters, each task has fixed sample size, $N_2 = 33$ (more resource); while in the remaining 10 clusters, the sample size (N_1) varies from 1 to 33 (less resource). Solid curves display the test risk of the tasks with N_1 samples and dashed curves present the test risk of tasks with N_2 samples. Rather than changing number of samples, in experiments shown in Figure 5b, we create another scenario where number of tasks per cluster is varied (as a measure of data resource). Here, instead all tasks contain $N = 10$ samples. For 10 of the total clusters, there are fixed $\bar{T}_2 = 50$ tasks in each cluster. However, the other 10 contain only $\bar{T}_1$ tasks in each cluster, and we compare the performance with different $\bar{T}_1$ selections. We change $\bar{T}_1$ from 2 to 50 and results are displayed in Fig 5b. Similar, solid curves present the results of the clusters who contain fewer tasks (less resource), to the contrary, dashed curves present the test risk of clusters with fixed $\bar{T}_2 = 50$ tasks (more resource).

In both figures, Multipath-MTL performs better than the other two models, which again shows that the sample complexity of hierarchical model is smaller than the vanilla and clustering models. When there are fewer samples or fewer tasks, all the three methods fail at learning a good representation. The three dashed curves in Fig. 5a behave in line with expectations: They follow from the fact that tasks with more samples can learn decent representations by themselves. The solid curve of Cluster MTL decreases slower, and it is because other than the other two methods where clusters are correlated and representations are shared, in Cluster MTL setting (as depicted in Fig. 4b), clusters are separately trained. Therefore, there is no benefit across the clusters. In Fig. 5b, firstly, the evidence that orange and blue dashed curves are above the green one again shows the sample efficiency of Multipath MTL. Here, when there are only 2 tasks for the 10 resource-poor clusters, the Cluster MTL has the worst performance because there is no representation sharing across clusters. Test risk of Vanilla MTL does not change too much even the task number increases. It is because MTL representation of vanilla model is larger and tasks don't have enough samples to train their prediction heads. For instance, blue solid curve hits blue dashed curve at very beginning, which shows that the model is already trained well and adding more tasks cannot help too much (both more resource tasks and less resource tasks are doing similar).