We thank all the reviewers for their helpful corrections and feedback. We will incorporate them into the revised version.

R1: We thank R1 for pointing out the missing eigenvalue notation. With regards to comments in weakness section,

1. In the fully agnostic setting (where the model is completely misspecified) we don't believe our results apply because the identity of the underlying shared representation becomes ambiguous. But for small degrees of model misspecification, we believe our results can be extended there at the price of additional error terms in our guarantees.

2. Our current results assume homogeneous sample size across different tasks for a clean presentation. We believe our techniques/results can be easily extended to non-homogeneous sample sizes across different tasks (with additional notational modifications in our results).

3. The "classifier head" of Example 3 (in Appendix A) is actually a nonlinear function. We agree that further investigating the task diversity definition for different examples is an important direction for future research.

4. We believe the neural network example will generalize to the logistic loss (by combining with the contents in Section 4.1). We opted for our current choice of different losses/functions to exhibit the utility of our general framework.

5. We believe that other bounds on the Gaussian/Rademacher complexity can directly be applied in our framework with little modification (such as [1, Theorem 5] for example). However, our framework does not directly accommodate margin bounds–such as in [2]–and extending our results to include these is an interesting future direction for research.

6: We agree that investigating the experimental implications of our bounds is an important direction for future work.

R2: We thank R2 for pointing out the typos, and suggestions for clarification. With regards to the typos in the "Additional Feedback" Section. For 1: $\hat{\mathbf{f}}$ refers to an ERM solution in the task variables of the training risk in Eq. 2–the pair $(\hat{\mathbf{f}}, \hat{\mathbf{h}})$ refers an entire ERM solution in Eq.2; we will define/clarify this. For 2: The union should be written $\mathbf{h} \in C_{\mathcal{H}_X}$ (referring to the $\mathcal{H}$ covering with respect to inputs $\mathbf{X}$). For 3: Yes, both terms should be primed (and the statement should be more clearly written as, "given this $\mathbf{h}'$, $\exists \mathbf{f}' \in C_{\mathcal{F}^{\otimes t}_{\mathbf{h}'(\mathbf{X})}}$ that is ϵ_2-close to $\mathbf{f}$ with respect to inputs $\mathbf{h}'(\mathbf{X})$. By construction, $\mathbf{h}' \in C_{\mathcal{H}_X}$ and $\mathbf{f}' \in C_{\mathcal{F}^{\otimes t}(\mathcal{H})}$."). We will correct/clarify the notation here.

With regards to Gaussian complexity vs Rademacher complexity, we chose to include only Gaussian complexities in the main paper to simplify the presentation, since the chain rule is most naturally stated in terms of Gaussian complexities. Since Gaussian/Rademacher complexities are equivalent up to logarithmic factors [3, p.97] we could also rewrite all our results in terms of Rademacher complexities at the cost of only logarithmic factors.

With regards to the comments in the weakness section, we believe our notion of task diversity can be understood in a natural way, and will provide further discussion and intuition to interpret it—see the following. Our framework/arguments (which hold for general $\mathcal{F}$, $\mathcal{H}$ and ℓ) do require abstraction, but we also believe this abstraction is a strength that allows our guarantees to be applied to a wide class of problems.

Definition 1 and Definition 2 seek to define two notions of distance between two representations $\mathbf{h}, \mathbf{h}'$. In our framework, information about the representations is only observed through the composite functions $f \circ \mathbf{h}$. For any direction/component in $\mathbf{h}$ that is not seen by a corresponding task f, that component of the representation $\mathbf{h}$ cannot be distinguished from a corresponding one in a spurious $\mathbf{h}'$. When this component is needed to predict on a new task f_0 which lies along that direction, transfer learning will not be possible. Therefore, Definition 1 defines a notion of representation distance in terms of information channeled through the training tasks while Definition 2 defines it in terms of an arbitrary new test task. Task diversity essentially encodes the ratio of these two quantities (i.e. how well the training tasks can observe relevant parts of the representations useful for the new task). In the case where the $\mathcal{F}$ contains underlying linear task functions $\boldsymbol{\alpha}_j^\star \in \mathbb{R}^r$ (as in our examples in Section 4), our task diversity definition reduces to ensuring these task vectors span the entire r-dimensional space containing the output of the representation $\mathbf{h}(\cdot) \in \mathbb{R}^r$. This is quantitatively captured by the conditioning parameter $\tilde{\nu} = \sigma_r(\mathbf{A})$ for $\mathbf{A} = (\boldsymbol{\alpha}_1^\star, \ldots, \boldsymbol{\alpha}_t^\star)^\top \in \mathbb{R}^{t \times r}$ which represents how correlated these vectors are in $\mathbb{R}^r$. Appendix A gives a further task diversity example when $\mathcal{F}$ contains nonparametric functions. We will provide further explanation of these definitions and their relationship to each specific example in the final version.

R3: We thank R3 for their comments and agree studying transfer learning in frameworks other then those using Gaussian complexities (i.e. with more refined data-dependent bounds as R1 mentions), is an interesting future direction.

[1] Golowich, Noah, Alexander Rakhlin, and Ohad Shamir. "Size-independent sample complexity of neural networks." Conference On Learning Theory. 2018.

[2] Wei, Colin, and Tengyu Ma. "Improved sample complexities for deep neural networks and robust classification via an all-layer margin." International Conference on Learning Representations. 2019.

[3] Ledoux, Michel, and Talagrand, Michel. Probability in Banach Spaces.