ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10095627

ON THE FAIRNESS OF MULTITASK REPRESENTATION LEARNING

Yingcong Li *Samet Oymak*

Department of Electrical and Computer Engineering
University of California, Riverside

ABSTRACT

In the context of multitask learning (MTL), representation learning is often accomplished through a feature-extractor ϕ that is shared across all tasks. This way, intuitively, the statistical cost of learning ϕ is collaboratively split across all tasks which enables sample efficiency. In this work, we consider a novel fairness scenario where T tasks can be split into majority and minority groups of sizes T_{maj} and T_{min} respectively: The group assignments are unknown during MTL and $T_{\text{min}}/T_{\text{maj}}$ ratio corresponds to the imbalance level of the problem. We further assume that these groups admit r_0, r_1-dimensional linear representations which are orthogonal to each other, thus, they would not benefit each other during MTL. Our main finding is that misspecification disproportionately hurts the minority tasks and over-parameterization is key to ensuring fairness of MTL representations. Specifically, we prove that, when we fit a $R = r_0$ dimensional misspecified representation, MTL model achieves small task-averaged risk however it has vanishing explanatory power on minority tasks. Conversely, when we fit a $R = r_0 + r_1$ dimensional well-specified representation, MTL model achieves small risks on both majority and minority tasks which are on par with the oracle baseline of training each group individually with the hindsight knowledge of assignments. Finally, we provide experimental results which are consistent with our theoretical findings.

***Index Terms*—** multitask learning, fairness, representation learning, imbalanced data, upper/lower bounds

1. INTRODUCTION

Multitiask learning (MTL) aims to learn a broad representation for numerous tasks by leveraging the useful information shared with different tasks. Many empirical and theoretical evidences have shown that MTL can significantly improve the task performance ([1, 2]). Here, a key consensus is that *highly similar tasks get more benefits from MTL than dissimilar tasks*. Based on this hypothesis, many data-dependent MTL methods are presented, and show considerable improvement compared to the vanilla MTL ([3, 4, 5]).

This work was supported by the NSF grants CCF-2046816 and CCF-2212426, Google Research Scholar award, and ARO grant W911NF2110312.

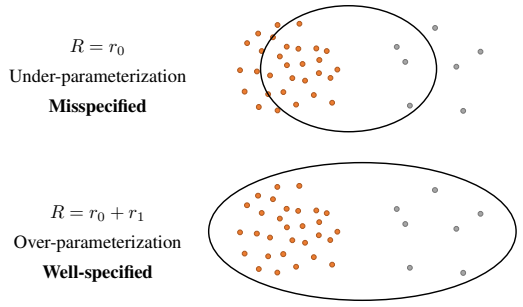

Fig. 1. Illustration of MTL with imbalanced tasks. Here, orange and gray dots are majority and minority tasks, which lie in r_0 and r_1-dimensional uncorrelated subspaces; and the black ellipses are R-dimensional subspaces the MTL models span. **Top**: Consider the under-parameterized setting where $R = r_0$. The representation is misspecified and can not fit all majority and minority well. **Bottom**: In the over-parameterization scenario where $R = r_0 + r_1$, MTL representation is sufficient to cover all the tasks and each task can be predicted well if trained properly.

However, most of the existing work focuses on the average performance over all tasks without considering the single task's performance. As larger models arise, thousands of tasks are trained together in MTL manner, while the minority/isolated tasks might not get benefits from MTL due to their inability in dominating the average performance. As depicted in Figure 1, minority tasks (gray dots) are separable from majority tasks (orange dots) and they do not share representations. This raises a question: *What is the performance of isolated minority tasks when trained jointly with majority tasks?* To answer it, in this work, we establish a MTL scenario with $T = T_{\text{maj}} + T_{\text{min}}$ tasks in total, where T_{maj}, T_{min} are the sizes of *majority* and *minority* groups ($T_{\text{maj}} \geq T_{\text{min}}$), and they correspond to irrelevant linear representations. We analyze the fairness in linear multitask representation by assuming majority and minority tasks are in two orthogonal subspaces and provide theoretical bounds for each group when MTL representation (with dimension R) is misspecified or well-specified. Our specific contributions are as follows.

- The misspecified representation learned with the under-parameterized model ($R = r_0$) can significantly hurt the minority tasks while performing well at average. In

fact, explanatory power of the MTL model over the minority tasks is proportional to $T_{\text{min}}/T_{\text{maj}}$ and vanishes as $T_{\text{maj}} \to \infty$.

- In the overparameterized setting ($R = r_0 + r_1$), minority tasks are not (significantly) harmed by the majority when trained jointly, and their performance is comparable to training minority tasks via a separate MTL.

1.1. Prior Art

Our work is most related to the literature on multitask and fairness representation learning.

Multitask representation learning. Generally, the goal of MTL is to train a task-shared feature extractor ϕ which maps the high-dimensional inputs to lower-dimensional features, and tasks are learned in a sample efficient fashion ([6, 7, 8, 9, 10, 11]) by utilizing the latent features. Moreover, task similarity also plays a role, since intuitively, highly correlated tasks benefit more from jointly training and representation sharing, and many existing methods ([3, 4, 5, 12, 13, 14]) significantly improve MTL performance by taking data into consideration. To the best of our knowledge, though evidences are shown that task relations are important in MTL, little focus on the theoretical analysis of isolated and minority tasks' performance. In this work, we address this challenge in the linear representation setting by setting ϕ to be a linear projection, and training with imbalanced and uncorrelated data (majority/minority).

Fairness and Imbalanced Data. As machine learning is increasingly used in a wide range of applications, fair learning has witnessed growing interest due to potential biases in the data ([15, 16, 17]). Prior works ([18, 19, 20]) have studied the trade-offs between accuracy and fairness in MTL. Another fairness-related literature is imbalanced classification problems ([21, 22]). However, most of the existing work focuses on the class-imbalance and provides methods that improve performance under imbalanced setting. Importantly, these don't discuss the impact of representation learning, where groups of tasks admit different optimal representations resulting in fairness challenges.

2. PROBLEM SETUP

Notation. Let $[n]$ denote set $\{1, \cdots, n\}$, $\|\mathbf{x}\|$ denote the ℓ_2-norm of a vector $\mathbf{x}$, and $\boldsymbol{I}_d$ denote the identity $d \times d$ matrix. Let $\lambda_{\max}(\boldsymbol{A})$, $\lambda_{\min}(\boldsymbol{A})$ return the maximal and minimal eigenvalues of positive semi-definite matrix $\boldsymbol{A}$. We use $\mathcal{O}(\cdot)$ to denote a equality up to a constant and $\widetilde{\mathcal{O}}(\cdot)$ hides constant and logarithmic factors.

Consider a linear multitask learning problem with T tasks and each task has N samples, denoted by $\{(\mathbf{x}_{ti}, y_{ti})\}_{i=1}^N \in \mathbb{R}^d \times \mathbb{R}$, $t \in [T]$. Assume tasks are partitioned into two groups, majority and minority, each with T_{maj} and T_{min} tasks respectively, where $T = T_{\text{maj}} + T_{\text{min}}$, $T_{\text{maj}} \geq T_{\text{min}}$. Let us assume that the groups lie in two orthogonal subspaces, so that their representations are helpless to each other. Different to standard MTL analysis where task-averaged performance is evaluated, in this work, we aim to quantify the performance bounds for each group. To this end, we first formulate a subspace-based MTL problem via introducing linear representation matrices. Let $\boldsymbol{B}^\star_{\text{maj}} \in \mathbb{R}^{r_0 \times d}$, $\boldsymbol{B}^\star_{\text{min}} \in \mathbb{R}^{r_1 \times d}$ denote the two representations corresponding to the two groups where we assume $\boldsymbol{B}^\star_{\text{maj}} \boldsymbol{B}^{\star\top}_{\text{maj}} = \boldsymbol{I}_{r_0}$, $\boldsymbol{B}^\star_{\text{min}} \boldsymbol{B}^{\star\top}_{\text{min}} = \boldsymbol{I}_{r_1}$ and $\boldsymbol{B}^\star_{\text{maj}} \boldsymbol{B}^{\star\top}_{\text{min}} = \boldsymbol{0}$. Here, rows of $\boldsymbol{B}^\star_{\text{maj}}$ span the r_0-dimensional subspace of majority groups, and similar for $\boldsymbol{B}^\star_{\text{min}}$. To clean notations, let $\mathcal{T}_{\text{maj}}$ and $\mathcal{T}_{\text{min}}$ be the sets of majority and minority task identifiers, where $|\mathcal{T}_{\text{maj}}| = T_{\text{maj}}$, $|\mathcal{T}_{\text{min}}| = T_{\text{min}}$ and $\mathcal{T}_{\text{maj}} \bigcup \mathcal{T}_{\text{min}} = [T]$. We assume linear labeling function. Specifically, for $t \in \mathcal{T}_{\text{maj}}$, data is generated by $y_{ti} = \boldsymbol{h}_t^\star \boldsymbol{B}^\star_{\text{maj}} \mathbf{x}_{ti} + z_{ti}$ where $(\boldsymbol{h}_t^\star)_{t \in \mathcal{T}_{\text{maj}}} \in \mathbb{R}^{r_0}$; whereas for $t \in \mathcal{T}_{\text{min}}$, $y_{ti} = \boldsymbol{h}_t^\star \boldsymbol{B}^\star_{\text{min}} \mathbf{x}_{ti} + z_{ti}$ where $(\boldsymbol{h}_t^\star)_{t \in \mathcal{T}_{\text{min}}} \in \mathbb{R}^{r_1}$. Here we assume $\|\boldsymbol{h}_t^\star\| \leq C$ for some constant $C \geq 1$, and inputs $\mathbf{x}_{ti} \in \mathbb{R}^d$ and noise $z_{ti} \in \mathbb{R}$ are zero-mean and independent with $\mathcal{O}(1)$ and $\mathcal{O}(\sigma)$ sub-Gaussian norm. In this work, we assume inputs have identity covariance where $\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \boldsymbol{I}_d$, and noise level $\mathbb{E}[z^2] = \sigma^2$.

Next, let us consider training a linear model which maps inputs to R-dimensional features using representation $\boldsymbol{B} \in \mathbb{R}^{R \times d}$, and each task has its specific-head $\boldsymbol{h}_t \in \mathbb{R}^R$. Then after applying quadratic loss function, we can define the empirical risk minimization problem as follows.

$$\hat{\boldsymbol{f}} = \underset{\boldsymbol{f} \in \mathcal{F}}{\arg\min} \ \widehat{\mathcal{L}}_{\text{avg}}(\boldsymbol{f}) \tag{1}$$

$$\text{where } \widehat{\mathcal{L}}_{\text{avg}}(\boldsymbol{f}) = \frac{1}{TN} \sum_{t=1}^{T} \sum_{i=1}^{N} \left(y_{ti} - \boldsymbol{h}_t^\top \boldsymbol{B} \mathbf{x}_{ti}\right)^2.$$

Here, $\boldsymbol{f} := ((\boldsymbol{h}_t)_{t=1}^T, \boldsymbol{B})$ and we define hypothesis set of $\boldsymbol{f}$ by $\mathcal{F} = \{((\boldsymbol{h}_t)_{t=1}^T, \boldsymbol{B}) \big| \boldsymbol{B}\boldsymbol{B}^\top = \boldsymbol{I}_R, \ \|\boldsymbol{h}_t\| \leq C\}$. Given finite samples, $\widehat{\mathcal{L}}_{\text{avg}}(\boldsymbol{f})$ defines the task-averaged training risk. Let $\mathcal{L}_{\text{avg}}(\boldsymbol{f}) := \mathbb{E}[\widehat{\mathcal{L}}_{\text{avg}}(\boldsymbol{f})]$ be the population risk, and define the *task-averaged excess test risk*: $\mathcal{R}_{\text{avg}}(\boldsymbol{f}) = \mathcal{L}_{\text{avg}}(\boldsymbol{f}) - \sigma^2$. While instead of focusing on the average risk where how majority and minority contribute is unclear, we study the fairness of MTL representation over each group. To this goal, define the population excess risks of majority and minority tasks as:

$$\mathcal{R}_{\text{maj}}(\boldsymbol{f}) = \frac{1}{T_{\text{maj}}} \sum_{t \in \mathcal{T}_{\text{maj}}} \mathbb{E}\left[(y_{ti} - \boldsymbol{h}_t \boldsymbol{B} \mathbf{x}_{ti})^2\right] - \sigma^2,$$

$$\mathcal{R}_{\text{min}}(\boldsymbol{f}) = \frac{1}{T_{\text{min}}} \sum_{t \in \mathcal{T}_{\text{min}}} \mathbb{E}\left[(y_{ti} - \boldsymbol{h}_t \boldsymbol{B} \mathbf{x}_{ti})^2\right] - \sigma^2.$$

Intuitively, we have $\mathcal{R}_{\text{avg}}(\boldsymbol{f}) = \frac{T_{\text{maj}}}{T} \mathcal{R}_{\text{maj}}(\boldsymbol{f}) + \frac{T_{\text{min}}}{T} \mathcal{R}_{\text{min}}(\boldsymbol{f})$.

In the following, we will present our main theoretical results of fairness analysis in linear MTL representation for both underparameterized and overparameterized settings in Section 3, and Section 4 presents our experimental evaluations.

Here, minimal population risk obeys $\arg\min_{\boldsymbol{f} \in \mathcal{F}} \mathcal{L}_{\text{avg}}(\boldsymbol{f}) \geq \sigma^2$ and equality holds when $R \geq r_0 + r_1$.

3. MAIN RESULTS

In this section, we consider both underparameterized (misspecified) and overparameterized (well-specified) scenarios and establish generalization bounds for excess test risks of majority and minority tasks.

3.1. Underparameterized MTL Representations

Recall that all majority tasks lie in a r_0-dimensional subspace denoted by $\boldsymbol{B}^\star_{\text{maj}} \in \mathbb{R}^{r_0 \times d}$ whereas minority lie in a r_1-dimensional subspace (orthogonal to $\boldsymbol{B}^\star_{\text{maj}}$) denoted by $\boldsymbol{B}^\star_{\text{min}} \in \mathbb{R}^{r_1 \times d}$. In this subsection, we consider the underparameterized setting where $R = r_0$, and present generalization risk bounds for majority and minority by introducing group covariances generating from the ground-truth heads.

Theorem 1 *Consider the MTL problem in* (1) *(without the $\|\boldsymbol{h}_t^\star\| \leq C$ constraints). For majority and minority tasks, define group covariances based on the heads*

$$\boldsymbol{H}_{maj} = \frac{1}{T_{maj}} \sum_{t \in \mathcal{T}_{maj}} \boldsymbol{h}_t^\star \boldsymbol{h}_t^{\star\top}, \ \boldsymbol{H}_{min} = \frac{1}{T_{min}} \sum_{t \in \mathcal{T}_{min}} \boldsymbol{h}_t^\star \boldsymbol{h}_t^{\star\top}.$$

Assume $\lambda_{min}(\boldsymbol{H}_{maj}) > 0$ and define $\kappa_0 = \frac{\lambda_{max}(\boldsymbol{H}_{maj})}{\lambda_{min}(\boldsymbol{H}_{maj})}$ and $\kappa_1 = \frac{\lambda_{max}(\boldsymbol{H}_{min})}{\lambda_{min}(\boldsymbol{H}_{maj})}$. Also assume prediction heads are normalized to be unit norm, that is, $\|\boldsymbol{h}_t^\star\| = 1$ for all t. Let $\bar{\boldsymbol{f}} = \{\bar{\boldsymbol{B}}, (\bar{\boldsymbol{h}}_t)_{t=1}^T\}$ be the population minima obtained as $N \to \infty$. Then, the majority/minority excess risks obey

$$\textit{Majority: } \mathcal{R}_{maj}(\bar{\boldsymbol{f}}) \leq \kappa_0 \frac{T_{min}}{T_{maj}},$$
$$\textit{Minority: } \mathcal{R}_{min}(\bar{\boldsymbol{f}}) \geq 1 - \kappa_1 \frac{T_{min}}{T_{maj}}.$$

Here in Theorem 1, we obtain upper and lower excess risk bounds for majority and minority tasks respectively. We observe that both bounds are corresponding to the imbalance level of the data, denoted as $T_{\text{min}}/T_{\text{maj}}$. Specifically, as the ratio decreases, we prove that optimal representation is less and less aligned with minority subspace. The observation can be easily interpreted as follows: If the model is given more majority tasks (or less minority tasks), the majority tasks then dominate the average performance ($\widehat{\mathcal{L}}_{\text{avg}}$), and learning majority benefits more in reducing the training risk compared to the minority. Therefore, the underdetermined model tends to learn the representation that aligned with majority tasks. Since majority and minority are uncorrelated, it in turn hurts minority.

The following corollary draws bounds considering the special case where $T_{\text{min}}/T_{\text{maj}} \to 0$.

Corollary 1 *Consider the setting of Theorem 1. We have that*

$$\lim_{T_{min}/T_{maj} \to 0} \mathcal{R}_{maj}(\bar{\boldsymbol{f}}) = 0, \quad \lim_{r_0 T_{min}/r_1 T_{maj} \to 0} \mathcal{R}_{min}(\bar{\boldsymbol{f}}) \geq 1.$$

This corollary states that, in the proper limit, minority tasks achieve the trivial risk $\mathcal{R}_{\text{min}}(\bar{\boldsymbol{f}}) = 1$ that corresponds to making zero prediction $\hat{y} = 0$. The limit condition $r_0 T_{\text{min}}/r_1 T_{\text{maj}} \to 0$ can be interpreted as follows: Majority tasks have T_{maj}/r_0 label energy per subspace dimension. In contrast, minority tasks have T_{min}/r_1 label energy per dimension. When majority energy-per-dimension dominates minority, all subspace dimensions of the representations are assigned to majority to minimize task-averaged risk.

3.2. Overparameterized MTL Representations

Different to the underdertermined problem where optimal solution ($\mathcal{R}_{\text{avg}} = 0$) is not feasible, in this subsection we consider the overparameterized setting where $R = r_0 + r_1$. Now if consider the case where $N \to \infty$, the population solution of MTL problem (1) $\boldsymbol{f}^\star$ satisfies $\mathcal{R}_{\text{avg}}(\boldsymbol{f}^\star) = 0$ which concludes $\mathcal{R}_{\text{maj}}(\boldsymbol{f}^\star) = \mathcal{R}_{\text{min}}(\boldsymbol{f}^\star) = 0$. To formalize this under the finite-sample setting, we provide a generalization bound which will help us accurately control the excess risk on minority tasks.

Theorem 2 *Let $\hat{\boldsymbol{f}} = ((\hat{\boldsymbol{h}}_t)_{t=1}^T, \hat{\boldsymbol{B}})$ be the empirical solution of problem* (1) *with $R = r_0 + r_1$. Assume per-task sample size obeys $N \gtrsim d + \log(T/\delta)$. Then with probability at least $1 - \delta$, the task-averaged test excess risk obeys*

$$\mathcal{R}_{avg}(\hat{\boldsymbol{f}}) \lesssim \sigma^2 \frac{dR + TR + \log(1/\delta)}{NT}.$$

Here $\lesssim$ subsumes constant and logarithmic factors. This result is obtained as a variation of Theorem 4.1 in [8]. In our technical report [23], we also provide an additional theorem that can circumvent the per-task sample size requirement $N \gtrsim d$. Combining Theorem 2 with the fact $\mathcal{R}_{\text{min}}(\boldsymbol{f}) \leq \frac{T}{T_{\text{min}}} \mathcal{R}_{\text{avg}}$ directly obtains a generalization bound for minority tasks as follows.

Corollary 2 *Consider the setting of Theorem 2. We have that with probability at least $1 - \delta$, the excess minority risk obeys*

$$\mathcal{R}_{min}(\hat{\boldsymbol{f}}) \lesssim \sigma^2 \frac{dR + TR + \log(1/\delta)}{NT_{min}}.$$

This corollary demonstrates fairness benefits of the overparameterized setting as the excess risk on minority tasks is at most $\frac{(T+d)R}{(T_{\text{min}}+d)r_1}$ times larger than training the minority group individually: Following Theorem 2, excess risk of individually training minority tasks is bounded by $\widetilde{\mathcal{O}}\left(dr_1 + T_{\text{min}} r_1 + \log(1/\delta)\right)/NT_{\text{min}}$. Hence, once $dr_1 + T_{\text{min}} r_1$ is proportional to $dR + TR$, well-specified joint MTL training is as good as individual MTL training up to a constant factor. Fortunately, this holds under mild conditions, namely, when $R \lesssim r_1$ and $T \lesssim \min(T_{\text{min}}, d)$.

4. SIMULATIONS

In this section, we discuss our experiments for both underparameterized and overparameterized settings, and results are

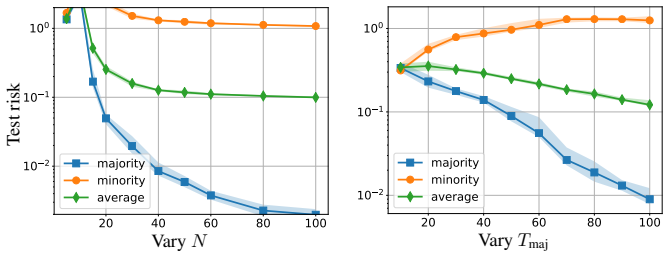

Fig. 2. We evaluate the performance of majority/minority tasks in MTL manner using underparameterized linear model ($R = r_0 = 8$). Here, blue, orange and green curves are test risks of majority, minority and all T tasks (average) respectively. **Left**: Fix $T_{\text{maj}} = 100$ and $T_{\text{min}} = 10$, while change the per-task sample size from 5 to 100. **Right**: Fix $N = 40$ and $T_{\text{min}} = 10$, while change T_{maj} from 10 to 100. Each marker is obtained by averaging 20 independent realizations.

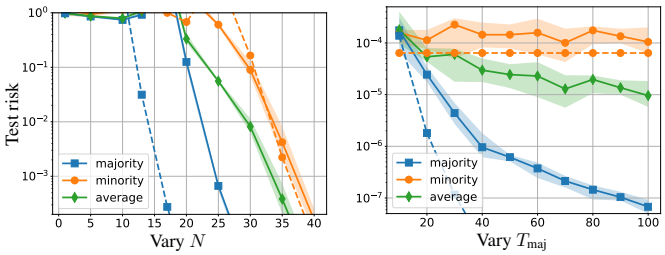

Fig. 3. We compare the performances of majority/minority groups, jointly/separately training in the overparameterized setting. Here, same as Fig. 2, blue, orange and green solid curves are corresponded test risks trained with overparameterized model ($R = r_0 + r_1 = 16$). Dashed curves present the results of training majority/minority tasks separately in MTL manner with $R = 8$. **Left/Right**: Follow the same settings in Fig. 2. The sample size varies form 1 to 40 on the left.

displayed in Fig. 2 and Fig. 3. We begin with our data generation model and hyperparameter selections for both settings.

Data generation. Following Section 2, we generate $\boldsymbol{B}^\star_{\text{maj}} \in \mathbb{R}^{r_0 \times d}$ and $\boldsymbol{B}^\star_{\text{min}} \in \mathbb{R}^{r_1 \times d}$ with orthonormal rows and $\boldsymbol{B}^\star_{\text{maj}} \boldsymbol{B}^{\star\top}_{\text{min}} = \boldsymbol{0}$ (which implies $d \geq (r_1 + r_2)$). Specifically, we first generate $r_0 + r_1$ d-dimensional orthonormal vectors uniformly at random independently, and then without losing generality and randomness, build $\boldsymbol{B}^\star_{\text{maj}}$ with the first r_0 vectors and $\boldsymbol{B}^\star_{\text{min}}$ with the latter. We also generate $\boldsymbol{h}^\star_t, t \in [T]$ uniformly at random over the unit sphere independently with proper r_0 and r_1 dimensions. The task t is generated by

$$y = \boldsymbol{h}^{\star\top}_t \tilde{\boldsymbol{B}} \mathbf{x} \quad \text{where} \quad \mathbf{x} \in \mathcal{N}(0, \boldsymbol{I}_d),$$

without label noise. Here, $\tilde{\boldsymbol{B}} = \boldsymbol{B}^\star_{\text{maj}}$ for $t \in \mathcal{T}_{\text{maj}}$, and $\tilde{\boldsymbol{B}} = \boldsymbol{B}^\star_{\text{min}}$ for $t \in \mathcal{T}_{\text{min}}$. In all experiments, we set ambient dimension $d = 32$ and local representation dimensions $r_0 = r_1 = 8$. Both under/over-parameterized settings are evaluated with two experiments showing in Fig. 2&3. On the left, we fix the numbers of majority and minority tasks, where $T_{\text{maj}} = 100, T_{\text{min}} = 10$, while change sample size of each task. Whereas on the right, sample size and minority size are fixed to be $N = 40, T_{\text{min}} = 10$, and T_{maj} varies from 10 to 100, and we can observe that at beginning points where $T_{\text{maj}} = T_{\text{min}} = 10$, majority and minority groups have similar performance. Blue, orange, and green solid curves display the test risks of majority, minority, and all T tasks respectively, and the dashed curves in Fig. 3 presents the corresponded individual results of training single model for tasks in the majority/minority group only.

In Fig. 2, we set $R = 8$ and evaluate the underdetermined model where zero average risk is not achievable even with noiseless labels. We observe that on the left, given sufficiently small imbalance ratio ($T_{\text{min}}/T_{\text{maj}} = 0.1$), the test risk of minority is strictly bigger than one even more training samples are added, which shows that representation of minority tasks is never learned when they are trained jointly with majority tasks, and the R-dimensional representation tends to align with the subspace majority tasks span. Here $\mathcal{L}_{\text{avg}} \approx T_{\text{min}} \mathcal{L}_{\text{min}}/T \approx 0.1$. Similar phenomenon appears on the right. On the leftmost where $T_{\text{maj}} = T_{\text{min}}$, the learned representation is shared equally to both majority and minority tasks and therefore, $\mathcal{L}_{\text{maj}} \approx \mathcal{L}_{\text{min}} \approx \mathcal{L}_{\text{avg}} < 1$. However, as T_{maj} grows, same as the left, the misspecified representation tends to fit the majority subspace, which in turn hurts minority tasks. We can observe that for $T_{\text{min}}/T_{\text{maj}} < 0.2$, $\mathcal{L}_{\text{min}} \geq 1$.

We consider overparameterized setting in Fig. 3 by setting $R = 16$, where instead zero risk is feasible under noiseless assumption (e.g., $\boldsymbol{B} = [\boldsymbol{B}^{\star\top}_{\text{maj}} \ \boldsymbol{B}^{\star\top}_{\text{min}}]^\top$). Therefore on the left, performances of both majority and minority tasks get improved and are approaching zero risks when training with more and more samples. While different to the underparameterized setting where enlarging size of majority group hurts minority tasks (Orange curve on the right of Fig. 2 increases.), ignoring some perturbations from randomness, orange solid curve on the right of Fig. 3 stays at the same level. It shows that once the representation is sufficient, the minority can never be harmed by the majority tasks. The decreasing of the blue curve is from the fact that the size of majority group is increased and more samples in training results in better performance. In these experiments, we also provide individually training results where majority/minority groups are trained in separate MTL with representation dimension $R = 8$ and results are displayed in dashed curves. Both sub-figures in Fig. 3 show that though trained with majority tasks, minority performs as good as individually training.

5. DISCUSSION

During recent years, there has been growing research on identifying and understanding the benefits of over-parameterization. Most of these research focus on either optimization benefits or statistical benefits through the lens of linearized models such as random feature regression. In this work, we identified and rigorously characterized a novel benefit of over-parameterization for representational fairness. As future directions, it would be of interest to empirically verify our theory through experiments on real datasets and also extending our theory to more realistic nonlinear settings.

6. REFERENCES

[1] Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes, "The benefit of multitask representation learning," *Journal of Machine Learning Research*, vol. 17, no. 81, pp. 1–32, 2016.

[2] Yu Zhang and Qiang Yang, "A survey on multi-task learning," *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[3] Rahul Ramesh and Pratik Chaudhari, "Model zoo: A growing brain that learns continually," in *International Conference on Learning Representations*, 2021.

[4] Shuxiao Chen, Koby Crammer, Hangfeng He, Dan Roth, and Weijie J Su, "Weighted training for cross-task learning," *arXiv preprint arXiv:2105.14095*, 2021.

[5] Chris Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn, "Efficiently identifying task groupings for multi-task learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 27503–27516, 2021.

[6] Sebastian Thrun and Lorien Pratt, *Learning to learn*, Springer Science & Business Media, 2012.

[7] Jonathan Baxter, "A model of inductive bias learning," *Journal of artificial intelligence research*, vol. 12, pp. 149–198, 2000.

[8] Simon S Du, Wei Hu, Sham M Kakade, Jason D Lee, and Qi Lei, "Few-shot learning via learning the representation, provably," *arXiv preprint arXiv:2002.09434*, 2020.

[9] Weihao Kong, Raghav Somani, Zhao Song, Sham Kakade, and Sewoong Oh, "Meta-learning for mixed linear regression," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5394–5404.

[10] Ziping Xu and Ambuj Tewari, "Representation learning beyond linear prediction functions," *Advances in Neural Information Processing Systems*, vol. 34, pp. 4792–4804, 2021.

[11] Rui Lu, Gao Huang, and Simon S Du, "On the power of multitask representation learning in linear mdp," *arXiv preprint arXiv:2106.08053*, 2021.

[12] Zhuoliang Kang, Kristen Grauman, and Fei Sha, "Learning with whom to share in multi-task feature learning," in *ICML*, 2011.

[13] Abhishek Kumar and Hal Daume III, "Learning task grouping and overlap in multi-task learning," *arXiv preprint arXiv:1206.6417*, 2012.

[14] Yingcong Li and Samet Oymak, "Provable pathways: Learning multiple tasks over multiple paths," *In Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[15] Luca Oneto and Silvia Chiappa, "Fairness in machine learning," in *Recent Trends in Learning From Data*, pp. 155–196. Springer, 2020.

[16] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[17] Simon Caton and Christian Haas, "Fairness in machine learning: A survey," *arXiv preprint arXiv:2010.04053*, 2020.

[18] Yuyan Wang, Xuezhi Wang, Alex Beutel, Flavien Prost, Jilin Chen, and Ed H Chi, "Understanding and improving fairness-accuracy trade-offs in multi-task learning," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 1748–1757.

[19] Luca Oneto, Michele Doninini, Amon Elders, and Massimiliano Pontil, "Taking advantage of multitask learning for fair classification," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 227–237.

[20] Chen Zhao and Feng Chen, "Rank-based multi-task learning for fair regression," in *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2019, pp. 916–925.

[21] Alexander Spangher, Jonathan May, Sz-Rung Shiang, and Lingjia Deng, "Multitask semi-supervised learning for class-imbalanced discourse classification," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 498–517.

[22] Ganesh Ramachandra Kini, Orestis Paraskevas, Samet Oymak, and Christos Thrampoulidis, "Label-imbalanced and group-sensitive classification under overparameterization," *arXiv:2103.01550, in submission to NeurIPS*, 2021.

[23] Yingcong Li and Samet Oymak, "On the fairness of multitask representation learning," `https://intra.ece.ucr.edu/~oymak/fairness.pdf`, 2023.