Abstract

In order to create machine learning systems that serve a variety of users well, it is vital to not only achieve high average performance but also ensure equitable outcomes across diverse groups. However, most machine learning methods are designed to improve a model’s average performance on a chosen end task without consideration for their impact on worst group error. Multitask learning (MTL) is one such widely used technique. In this paper, we seek not only to understand the impact of MTL on worst-group accuracy but also to explore its potential as a tool to address the challenge of group-wise fairness. We primarily consider the standard setting of fine-tuning a pre-trained model, where, following recent work (Gururangan et al., 2020; Dery et al., 2023), we multitask the end task with the pre-training objective constructed from the end task data itself. In settings with few or no group annotations, we find that multitasking often, but not consistently, achieves better worst-group accuracy than Just-Train-Twice (JTT; Liu et al., 2021) – a representative distributionally robust optimization (DRO) method. Leveraging insights from synthetic data experiments, we propose to modify standard MTL by regularizing the joint multitask representation space. We run a large number of fine-tuning experiments across computer vision and natural language processing datasets and find that our regularized MTL approach consistently outperforms JTT on both average and worst-group outcomes. Our official code can be found here: https://github.com/atharvajk98/MTL-group-robustness

*Equal contribution.
1 Introduction

As machine learning systems exert ever-increasing influence on the real world, it is paramount that they not only perform well on aggregate but also exhibit equitable outcomes across diverse subgroups characterized by attributes like race (Buolamwini & Gebru, 2018; Liang et al., 2021), gender (Buolamwini & Gebru, 2018; Srinivasan & Bisk, 2022) and geographic location (Jurgens et al., 2017; De Vries et al., 2019; Ayush et al., 2021). Therefore, it is important to understand the impact of widely-used machine learning techniques with respect to these desiderata. Multitask learning (MTL) (Caruana, 1997; Baxter, 2000; Ruder et al., 2019; Dery et al., 2021a) is one example of such a technique that features prominently in machine learning practitioners’ toolbox for improving a model’s aggregate performance. However, the effect of multitasking on worst group outcomes is underexplored. In this paper, we both study the impact of MTL, as is, on worst group error and also consider whether modifications can be made to improve its effect on worst group outcomes.

Traditionally, the problem of worst-group generalization has been tackled explicitly via methods such as distributionally robust optimization (DRO) (Ben-Tal et al., 2013; Duchi & Namkoong, 2018; Hashimoto et al., 2018; Sagawa et al., 2020a). In contrast to traditional empirical risk minimization, DRO aims to minimize the worst-case risk over a predefined set of distributions (the uncertainty set). Defining the uncertainty set usually (but not always) requires access to group annotations. Since average performance based approaches like MTL are typically designed without consideration of group annotations, our focus will be on settings with limited-to-no group annotations. In these settings, there exist a number of generalized Reweighting (GRW) algorithms for distributional robustness (Nam et al., 2020; Liu et al., 2021; Zhang et al., 2022; Nam et al., 2022; Qiu et al., 2023; Zhai et al., 2023; Izmailov et al., 2022). These approaches minimize the weighted average risk based on the weight assigned to each example. One such widely used method is the Just-Train-Twice (JTT) algorithm (Liu et al., 2021), which performs two model training runs: one to identify poorly performing examples and another run that upweights these examples. For our empirical explorations, we take JTT as a representative DRO method and use it to provide reference performance to situate our study of multitask learning.

We focus our investigations of multitask learning on the ubiquitous setting of fine-tuning a pre-trained model. Here, a common way to improve end task average performance is to multitask the end task with the pre-training objective constructed over the task data itself (Gururangan et al., 2020; Dery et al., 2021b). We intuit that this multitasking approach could improve robustness to worst group outcomes since previous work like Hendrycks et al. (2019); 2020; Mao et al. (2020) has established a favorable connection between pre-training and robustness (both adversarial and out-of-distribution). We test our intuition by conducting preliminary experiments across a pair of computer vision and natural language tasks in two settings: one with limited group annotations and the other with none. Initial results (Table 1) reveal that multitasking shows promise in that it can improve worst group outcomes over ERM and JTT. However, these improvements are not consistent. Therefore, we are spurred to consider modifications to make it a more competitive tool against worst group outcomes.

In order to build intuition about how to adapt MTL to target the worst group error, we conduct controlled experiments on two-layer linear models trained from scratch on synthetic data. We borrow the synthetic data setup introduced by Sagawa et al. (2020b) in which training data consists of two majority groups, where spurious features (features that are not required to robustly solve the end task) are predictive of the end task output, and two minority groups, where spurious features are uncorrelated with the output. Sagawa et al. (2020b) demonstrated that under certain conditions on the generative distribution of the input features, linear models trained on such data would probably rely on spurious features and thus, have poor worst group error. Working with this simplified setup where worst-group outcomes are easily inducible allows us to study MTL’s effects more incisively. We instantiate reconstruction from noised input as our auxiliary

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>No Group Labels</th>
<th>Val Group Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Waterbirds</td>
<td>ERM</td>
<td>80.1(\pm)1</td>
<td>85.4(\pm)1</td>
</tr>
<tr>
<td></td>
<td>JTT</td>
<td>82.1(\pm)1</td>
<td>85.9(\pm)1</td>
</tr>
<tr>
<td></td>
<td>(MTL) ERM+MM</td>
<td>80.1(\pm)1</td>
<td>85.4(\pm)1</td>
</tr>
<tr>
<td>Civil-Small</td>
<td>ERM</td>
<td>51.6(\pm)1</td>
<td>67.4(\pm)1</td>
</tr>
<tr>
<td></td>
<td>JTT</td>
<td>52.5(\pm)1</td>
<td>68.0(\pm)1</td>
</tr>
<tr>
<td></td>
<td>(MTL) ERM+MLM</td>
<td>58.3(\pm)1</td>
<td>68.5(\pm)1</td>
</tr>
</tbody>
</table>

Table 1: Standard multitasking improves worst group outcomes over ERM and JTT but not consistently. Experimental details can be found in Section A.1.
task to perform multitask learning in the synthetic setup. This choice is partially informed by the fact that many pre-training objectives like masked language modeling (MLM) (Devlin et al., 2018) and masked image modeling (He et al., 2022; Tong et al., 2022) are based on input reconstruction. When training solely on this auxiliary task, we uncover that regularizing the pre-output layer of the model is critical for ensuring that the model upweights the core features (features required to robustly solve the end task) over the spurious ones. This leads us to the following recipe for improving worst-group error: regularized multitask learning of the end task with the (appropriately chosen) auxiliary objective.

Through a battery of experiments across natural language processing (NLP) and computer vision (CV) datasets, we demonstrate that multitasking the end task with the pre-training objective along with \( \ell_1 \) regularization on the shared, pre-prediction layer activations is competitive when pitted against state-of-the-art DRO approaches like JTT (Liu et al., 2021) and Bitrate-Constrained DRO (Sethur et al., 2023). Specifically, in settings where only validation group annotations are available, regularized MTL outperforms JTT and BR-DRO on 3/3 and 2/3 datasets, respectively. Our approach improves worst-group accuracy over ERM (by as much as \( \sim 4\% \)) and JTT (by \( \sim 1\% \)) in settings where group annotations are completely unavailable. Moreover, regularized MTL consistently outperforms both ERM and JTT on average performance, regardless of whether group annotations are available or not. Thus, within the prevailing framework of utilizing pre-trained models for downstream fine-tuning, our results demonstrate that regularized multitask learning can be a simple yet versatile and robust tool for improving both average and worst-group outcomes.

2 Informal motivation for our regularized MTL method

Problem Setup / Preliminaries: Let each input example \( x \in X \) have a classification label \( y \in Y \) and a demographic attribute \( s \in S \). Each group \( g = (s, y) \in G \) is defined by the label \( y \) and the attribute \( s \), such that \( s \) spuriously correlates with \( y \). Thus, the sample space of \( G \) is the Cartesian product of \( Y \) and \( S \). Our goal is to learn a model that minimizes the worst-group error. We evaluate the models based on their worst group accuracy (WGA), i.e., the minimum predictive accuracies of our model across all groups. We are interested in the setting where the spurious attribute \( s \), and consequently, the group identity \( g \) are unavailable (or available to a very limited degree) at training time, as annotating spurious attributes is typically expensive.

Why would we expect multitask learning to help mitigate worst group outcomes? It would be naive to assume that multitasking the end task with any auxiliary task would prevent poor group outcomes. In order to better understand intuitively which auxiliary tasks may be helpful, we first provide an example using the data generation process and linear model setup presented in Sagawa et al. (2020b)’s work on the effect of spurious and core features on worst-group accuracy.

When do models incur high worst group error? Sagawa et al. (2020b) describe a simple data-generating distribution that defines, for each example, a label \( y \in \{-1, 1\} \), a spurious attribute \( s \in \{-1, 1\} \), and features \( x \). The features are described as either core features \( x_{\text{core}} \) if they are associated with the label \( y \), or spurious features \( x_{\text{spur}} \) if they are associated with the spurious attribute \( s \):

\[
\begin{align*}
  x_{\text{core}} | y & \sim \mathcal{N}(y\mathbf{1}, \sigma_{\text{core}}^2 I_d) \\
  x_{\text{spur}} | s & \sim \mathcal{N}(s\mathbf{1}, \sigma_{\text{spur}}^2 I_d)
\end{align*}
\]

\( x = [x_{\text{core}}; x_{\text{spur}}] \in \mathbb{R}^d \) and \( d = (d_c + d_s) \)

We can then define a linear model, parameterized by \( \hat{w} \), that predicts the label given the features.

\( \hat{y}^{(i)} = \hat{w} \cdot x^{(i)} \).

Note that because the core features \( x_{\text{core}} \) are the ones associated with the label to be predicted, they are the ones that the model should use in order to attain high predictive accuracy.

*We will delve deeper into other motivations in Section 3.*
The cross-product of the space of possible labels \( y = \pm 1 \) and spurious attributes \( s = \pm 1 \) divides the samples generated from this distribution into four groups. When some groups are more frequent than others in the training data, a correlation between \( \{ y, s \} \) is created. Further still, in the presence of the above correlation, if the spurious features have lower variance with respect to the data generating process (Equation 1), i.e., \( \sigma_{\text{spur}}^2 \leq \sigma_{\text{core}}^2 \), linear models will tend to rely more on (assign a higher weight to) the spurious features over the core ones (Sagawa et al., 2020b). This learned reliance on the spurious features – instead of the core features that are truly predictive of the label – results in poor worst-group error.

Why does reconstruction help? Considering the above, for an auxiliary task to be helpful, it should discourage the model from using spurious features by showing a stronger preference for core features in exactly the case when \( \sigma_{\text{spur}}^2 \leq \sigma_{\text{core}}^2 \). In this paper, we argue that one class of tasks that fulfills this criterion are reconstruction tasks, where we predict original input features from noised versions.

For instance, in the example above, if we add noise with a constant variance of \( \sigma_{\text{noise}}^2 \) over each dimension, it results in noised inputs that have variances \( \tilde{\sigma}_{\text{spur}}^2 = (\sigma_{\text{spur}}^2 + \sigma_{\text{noise}}^2) \leq \tilde{\sigma}_{\text{core}}^2 = (\sigma_{\text{core}}^2 + \sigma_{\text{noise}}^2) \) per spurious and core feature dimension, respectively. Under the assumption that both true labels \( y = 1 \) and \( y = -1 \) are equally probable, and in the simplest case where we are reconstructing features independently of each other with a linear predictor \( \hat{x}_i = w_i \tilde{x}_i \) (where \( \tilde{x} \) is the noised input), the Bayes optimal weight on a feature \( i \) would be (see Appendix A.3 for the full proof):

\[
\begin{align*}
\hat{w}^\text{bayes}_i &= \frac{\sigma_i^2 + 0.5 (\mu_i^2|y=1 + \mu_i^2|y=-1)}{\sigma_i^2 + 0.5 (\mu_i^2|y=1 + \mu_i^2|y=-1) + \sigma_{\text{noise}}} \\
&= \frac{\sigma_i^2 + 0.5 (\mu_i^2|y=1 + \mu_i^2|y=-1)}{\sigma_i^2 + 0.5 (\mu_i^2|y=1 + \mu_i^2|y=-1) + \sigma_{\text{noise}}}
\end{align*}
\]

(3)

where \( \mu_i^2|y=\pm 1 \) are the per-dimension means from Eqn 1

Note that \( \hat{w}^\text{bayes}_i \) is larger for dimensions with higher variances \( \sigma_i^2 \), assuming \( \mu_i^2|y=\pm 1 \) are symmetric across core and spurious features (i.e all \( i \)). Thus, this reconstruction task places more weight on the core features in exactly the setting where a linear predictor for the end task would prefer to use the spurious features. Note that for this preference of the core features to be effectively realized, the auxiliary task needs to be sufficiently up-weighted, but not so much so that the end task is not learned at all.

Why is regularization necessary? Given an auxiliary task with the above property of preferring core to spurious features (under \( \sigma_{\text{spur}}^2 \leq \sigma_{\text{core}}^2 \)), a model with sufficient capacity can still rely on spurious features for solving the end task. We can incentivize the model to mostly use core features by applying sufficient regularization to the parts of the model that are shared between the two tasks (such as shared feature extractors). The restricted capacity encourages the model to rely on features that would cause it to do well on both tasks, which would be the core features.

Based on the intuition established in this section, we propose a simple yet effective method for improving worst-group outcomes: Multitasking the end task with the pre-training objective – which tends to be reconstruction tasks – while regularizing the shared (pre-prediction) layer. In Section 3, we will test this intuition through synthetic data experiments, and in Sections 4 and 5, we demonstrate its empirical efficacy through natural data experiments.

### 3 Synthetic Data Experiments

We initiate our investigation with an empirical study in a simplified context, involving the training of a two-layer linear model on synthetic data. This exploration is designed to substantiate the informal intuition introduced in Section 2 through concrete empirical findings.

**3.1 Data Generating Distribution**

We base our experiment on the data generation distribution from Equation 1, where features are divided into core and spurious ones.
As an instantiation, we consider end the task data defined by \( T_{\text{end}} = \{(x_i, y_i)\}_{i \in \mathbb{N}} \) where we have \( N \) total samples. Here \( d_c = 1; \ d_s = 1 \implies d = 2 \). The data is dominated by samples where \( \{s = y\} \) and thus, we have two majority groups \( G_{s=y=1} \) and \( G_{s=y=-1} \), each with \( \frac{n_{\text{max}}}{2} \) samples. The two minority groups are when \( \{s = -y\} \) each with \( \frac{n_{\text{min}}}{2} \): \( G_{s=-y=1}, G_{s=-y=-1} \). Due to the fact that \( n_{\text{max}} > n_{\text{min}} \), the attribute \( s \) is highly correlated with the label \( y \) in the training data and thus, is a spurious feature when considering the true data generation distribution. The end task is to predict the true label \( y_i \) from the given input data \( x_i \).

Figure 1 shows data sampled from the generative process described in Equation 1. We produce 1000 samples in \( \mathbb{R}^2 \) with \( \sigma_{\text{core}} = 0.6 \) and \( \sigma_{\text{spur}} = 0.1 \). \( n_{\text{min}} = 100 \) and \( n_{\text{max}} = 900 \), making the spurious feature highly correlated with the true label.

### 3.2 Training on the end task only

We train on \( T_{\text{end}} \) only to confirm that the resulting model has poor worst-group outcomes. Since we will eventually be do multitasking, we use a two layer linear model where the first layer is a linear featurizer – that will eventually be shared between all tasks being multitasked – and the second layer is a prediction head dedicated to the end task. This shared featurizer but separate head architecture is common in modern multitask learning (Yu et al., 2020; Michel et al., 2021; Dery et al., 2021a).

For simplicity, the featurizer layer \( f(\cdot) \) is a diagonal linear function parameterized by \( a^T \):

\[
f : \mathbb{R}^d \to \mathbb{R}^d \mid f(\alpha)(x) = (\text{diag}(\alpha)) x \tag{4}
\]

And the final output prediction layer is given by

\[
y_{\text{pred}}^{\text{end}} = (w_{\text{end}}^T f(x)) = (w_{\text{end}}^T \text{diag}(\alpha)) x = (y_{\text{end}}^T)^T x
\]

Note that we have effectively parameterized a linear model with decomposed formulation which will be helpful once we proceed to multitasking. The end task loss is binary cross-entropy with \( \ell_2 \) regularization on

*the diagonal parameterization allows us to easily read off how much weight is assigned to core features versus spurious ones*
As motivated in Section 2, we proceed to introduce a reconstruction based auxiliary task. The auxiliary task data is defined by \( T_{\text{aux}} = \{ (x_i, x_i) \}_{i \in M} \) where we have \( M \) total samples. \( M \) unlabelled points (with respect to the end task) are taken from the distribution described by Equation 4. Noise of the form
\[
\epsilon_{\text{noise}} \sim \mathcal{N}(0, \sigma_{\text{noise}}^2 I_d) \quad | \quad \tilde{x} = x + \varepsilon
\]
is applied to each point. The task is to reconstruct \( x_i \) from \( \tilde{x}_i \). Reusing the featurizer from Equation 4, we define the following prediction model:
\[
x_{\text{pred}}^{\text{aux}} = (W^{\text{aux}})^T f(a)(\tilde{x})
\]
\( W^{\text{aux}} \) parameterizes the auxiliary prediction head which we regularize to \( \{ W^{\text{aux}} \in \mathbb{R}^{d \times d} \mid ||W^{\text{aux}}||_F^2 = 1 \} \).

Finally, our reconstruction loss is given by:
\[
L_{\text{recon}}(W^{\text{aux}}, a) = \frac{1}{2M} \sum_{(x_i, x_i) \in T_{\text{aux}}} ||x_i - (W^{\text{aux}})^T f(\tilde{x})||^2
\]

Using the synthetic data instantiation in Figure 1 we apply noise from \( \mathcal{N}(0, I_2) \), i.e., \( \sigma_{\text{noise}}^2 = 1 \) on each of the 1000 training points to get the training data for the auxiliary task. We fit the model solely to the auxiliary task by running batched stochastic gradient descent on \( L_{\text{recon}} \). We use learning rates in the set \( \{10^{-2}, 10^{-3}\} \) and batch sizes in the set \( \{64, 256\} \).

We consider two cases when the intermediate layer \( a \) has low versus high capacity, as reflected in \( \ell_1 \)-norm. Low \( \ell_1 \)-norm \( \|a\|_1 = |a_{\text{spur}}| + |a_{\text{core}}| = 0.1 \) means restricted capacity since this constraint (along with \( ||W^{\text{aux}}||_F^2 = 1 \)) results in models that cannot fit the training data perfectly. High \( \ell_1 \)-norm \( (\|a\|_1 = 10) \)
means that the model is expressive enough to perfectly fit the training data. Figure 3 provides insight into the learned intermediate layer in either case. When the model has enough capacity, there is no competition between the core and spurious features. This means solutions where the spurious feature is weighted more than the core feature are feasible as long as the core feature weight is enough to reconstruct the noised core features well.

However, under restricted capacity where the learned weight of the core and spurious features are in direct competition, the model has to put more weight on the core features in order to achieve a lower auxiliary loss (as motivated in Section 2). This can be seen in Figure 3. Thus, for the auxiliary task to be effective at forcing a model to use core features over spurious ones, the model’s capacity must be reasonably restricted.

3.4 Multitasking with regularization

Given the findings from Section 3.3, we proceed to multitask \( \mathcal{L}_{\text{end}} \) and \( \mathcal{L}_{\text{recon}} \) along with regularization on the shared layer \( \mathbf{a} \). Let \( \mathbf{A}(\tau) = \{ \mathbf{a} \in \mathbb{R}^d \mid ||\mathbf{a}||_1 = \tau \} \) be a set of \( \ell_1 \)-norm constrained vectors, we solve the following multitask optimization problem:

\[
\bar{\mathbf{W}}_{\text{aux}}, \bar{\mathbf{w}}_{\text{end}}, \bar{\mathbf{a}} = \arg\min_{\mathbf{W}_{\text{aux}} \in \mathbb{R}^{d \times d}, \mathbf{a} \in \mathbf{A}(\tau)} \|\mathbf{W}_{\text{aux}}\|_F = 1, \alpha \cdot \mathcal{L}_{\text{recon}}(\mathbf{W}_{\text{aux}}, \mathbf{a}) + \mathcal{L}_{\text{end}}(\mathbf{w}_{\text{end}}, \mathbf{a})
\]  

(8)

We implement the multitask model illustrated in Figure 4. We use the same set of hyper-parameters as used in Section 3.3 and perform joint stochastic gradient descent on both \( \mathbf{T}_{\text{end}} \) and \( \mathbf{T}_{\text{aux}} \). When \( \tau \) is chosen to be small enough, model capacity is restricted, and the model is forced to rely chiefly on the core features to do well on both the end and auxiliary tasks. Figure 5 evinces this. When the norm of the shared layer is high \( ||\mathbf{a}||_1 = 10 \), the end task can still predominantly rely on the spurious features leading to poor worst group error. On the other hand, when model capacity is reasonably restricted by setting \( ||\mathbf{a}||_1 = 0.1 \), we see from Figure 5 (left) that we can achieve improved worst group accuracy. Thus, we can effectively leverage the reconstruction auxiliary task in this simplified setting by applying sufficient regularization to ensure improved worst-group outcomes.

Table 2: Summary of results from Sections 3.3 and 3.4. Regularized multitasking leads to improved worst-group outcomes.

| Method       | \( ||\mathbf{a}||_1 = 0.1 \) | \( ||\mathbf{a}||_1 = 10 \) |
|--------------|-------------------------------|-------------------------------|
| End task only| 64.15                         | 48.30                         |
| Regularized MTL | 94.02%                       | 0.00%                         |

Table 2: Summary of results from Sections 3.3 and 3.4. Regularized multitasking leads to improved worst-group outcomes.

Figure 5: Depicted are the learned half-spaces for the multitask model under \( \tau = \{0.1, 10\} \) and \( \alpha = 10 \). Restricting the capacity of the shared feature space is critical for multitasking to be effective for improving worst group error. Examples visualized are 1000 balanced test examples sampled from Equation 1.
4 Details For Natural Data Experiments

We have made a case for using regularized multitask learning to combat poor worst-group performance through empirical explorations in a simplified, synthetic setting. In this section, we review the experimental details for our investigations of tasks of more practical interest.

4.1 Datasets

We conduct experiments across three datasets. To relieve the burden of compute, we introduce a fourth dataset, a smaller, sub-sampled version of one of the original datasets for ablations.

1. **Waterbirds**: This image classification dataset was introduced by Sagawa et al. (2020a). The task is to distinguish between species of land and water birds. It consists of bird images sourced from the CUB dataset Wah et al. (2011) and superimposed on land or water backgrounds from the Places dataset Zhou et al. (2018). The label (type of bird) is spuriously correlated with the background, resulting in 4 groups. Since this is a small dataset (4795 train examples), we also use it for ablations.

2. **MultiNLI**: This is a natural language inference dataset. The task is to classify whether the second sentence is entailed by, contradicts, or is neutral with respect to the first sentence (Williams et al., 2018). Following Sagawa et al. (2020a), we utilize the presence of negation words as a spurious attribute, leading to the creation of a total of 6 groups.

3. **Civilcomments**: The Civilcomments dataset is a toxicity classification dataset that contains comments from online forums Borkan et al. (2019); Koh et al. (2021). Along with the toxicity label, each text is annotated with additional overlapping sub-group labels of 8 demographic identities: male, female, LGBTQ, Christian, Muslim, other religions, Black, and White. As per Koh et al. (2021) and Sagawa et al. (2020a), we define 16 overlapping groups by taking the Cartesian product of the binary toxicity label and each of the above 8 demographic identities.

4. **Civilcomments-small**: As Civilcomments is a large dataset of about 448000 datapoints, we create a sub-group stratified subset of 5% for conducting ablations and other detailed experiments. Our subset contains 13770, 2039, and 4866 datapoints in our train, validation, and test split, respectively.

4.2 Multitask Model and Training Details

We follow the parameter sharing paradigm (Ruder, 2017; Sener & Koltun, 2018) where both \( \mathbf{T}_{\text{end}} \) and \( \mathbf{T}_{\text{aux}} \) share the same model body, parameterized by \( \theta_{\text{base}} \). We instantiate task-specific heads, parameterized by \( \theta_{\text{end}} \) and \( \theta_{\text{aux}} \), respectively. We introduce \( \ell_1 \) regularization to the last layer activations immediately before the per-task prediction heads. Specifically, let \( h_{\text{end}}, h_{\text{aux}} \in \mathbb{R}^d \) be the output representations generated by the base model, which are fed into their respective task-specific heads. Our final multitask learning objective is expressed as follows:

\[
\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{end}} + \alpha_{\text{aux}} \cdot \mathcal{L}_{\text{aux}} + \alpha_{\text{reg}} \left( \|h_{\text{end}}\|_1 + \|h_{\text{aux}}\|_1 \right)
\]

We cross-validate optimizing \( \mathcal{L}_{\text{final}} \) with different weighting schemes. We choose \( \alpha_{\text{aux}} \) and \( \alpha_{\text{reg}} \) from the set \( \{e^{-1}, e^0, e^1\} \). Note that whilst we optimize \( \mathcal{L}_{\text{final}} \) we care only about improving worst-group error on \( \mathbf{T}_{\text{aux}} \). We use the pretrained BERT\textsubscript{base} (Devlin et al., 2018) and ViT\textsubscript{base} (Dosovitskiy et al., 2020) as the shared base models for NLP and CV tasks, respectively. We leverage the base models’ self-supervised pre-training objectives, namely, masked language modeling (MLM) and masked image modeling (MIM) for our auxiliary transfer task \( \mathbf{T}_{\text{aux}} \) as in Dery et al. (2021a). These auxiliary objectives are based on end-task data itself (unless specified otherwise). We do this to maintain an apples-to-apples comparison with our chosen baselines, which do not use external data. As Section 5 will show, we obtain performance improvements even in this setting. In Section 5.4 we show that our improvements when using task-only data are predicated on sufficient prior pre-training. More details on the multitask model and batching scheme are presented in Appendix A.

*In contrast to synthetic data experiments, where the norm constraint is applied to pre-prediction layer weights, in this section we directly apply the norm constraint to the features, indirectly constraining model weights. In the synthetic experiment, core and spurious features have a one-to-one mapping to model weights, enabling direct regularization of the features, but that is no longer possible in more complex models.*
For training, we vary the fine-tuning learning rate within the ranges of \(10^{-3}, 10^{-4}\) for Waterbirds, and \(10^{-4}, 10^{-5}\) for the text datasets. We experiment with batch sizes in the set \{4, 8, 16, 32\}. We use the same batch sizes for \(T_{end}\) and \(T_{aux}\). We train for 50 epochs for the NLP datasets and 200 epochs for Waterbirds, with an early stopping patience of 10, as per the check-pointing scheme explained in section 4.2. We use the Adam optimizer for NLP datasets with decoupled weight decay regularization of \(10^{-2}\) (Loshchilov & Hutter, 2017). Consistent with the recent studies on ViT (Dosovitskiy et al., 2020; Steiner et al., 2022), we use SGD with a momentum of 0.9 (Sutskever et al., 2013) to fine-tune Waterbirds. We run each hyperparameter configuration across 5 seeds and report the averaged results. We report the ERM, JTT, and groupDRO results for Civilcomments and MultiNLI from Idrissi et al. (2022) as the authors conducted extensive hyperparameter tuning across all these methods. However, since Idrissi et al. (2022) report results on Waterbirds using a ResNet-50 model (He et al., 2016) and our experiments employ ViT, we re-run all baselines using ViT with a consistent set of hyperparameters, as mentioned above.

**Evaluation Details** We assess all methods and datasets using two model selection strategies:

1. **Val-GP**: This strategy requires group annotations in the validation data during training. Here, we select the model based on the maximum worst-group accuracy on the validation data.
2. **No-GP**: This strategy requires no access to any group annotations during training. We select the model based on the average validation accuracy.

**Baseline Methods** Since we evaluate our method based on its ability to generalize to worst performing groups, we benchmark it against three popular methods found in group generalization literature. These methods either directly or indirectly optimize for worst-group improvements.

1. **Empirical Risk Minimization (ERM)**: This is the standard approach of minimizing the average loss over all the training data. No group information is used during training except when the Val-GP strategy is used for model selection.
2. **Just Train Twice (JTT)**: It presents a two-step approach for worst group generalization (Liu et al., 2021). JTT first trains a standard ERM model for \(T\) epochs to identify misclassified datapoints. Then, a second model is trained on a reweighted dataset constructed by upweighting the misclassified examples by \(\alpha_{up}\). It does not use group information during training except for the Val-GP strategy.
3. **Bit-rate Constrained DRO (BR-DRO)**: Traditionally, in the two-player formulation of DRO, the adversary can use complex reweighting functions, resulting in overly pessimistic solutions. In contrast, BR-DRO (Setlur et al., 2023) constrains the adversary’s complexity based on information theory under a data-independent prior. While BR-DRO offers weaker robustness without performance guarantees for arbitrary reweighting, it is less pessimistic and suitable for simpler distribution shifts, characterized by a reweighting function contained in a simpler complexity class. BR-DRO does not use group information during training except for the Val-GP setting.
4. **Group-DRO**: Group distributionally robust optimization minimizes the maximum loss across all the sub-groups (Sagawa et al., 2020a). This optimization method incorporates group annotations during training. Similar to prior works (Liu et al., 2021; Idrissi et al., 2022; Setlur et al., 2023), we treat it as an oracle, as this is the only method that uses group annotations.

## 5 Results and Discussion

In this section, we provide empirical evidence demonstrating the effectiveness of our regularized MTL approach in mitigating worst-group error while maintaining average performance across different scenarios.

### 5.1 Multitasking is Competitive with Bespoke DRO Methods

We first compare our approach with previously proposed methods for tackling worst-group accuracy. Table 3 details the performance of various methods across the tasks of interest for the Val-GP setting. As expected, groupDRO yields the highest worst-group accuracy as it directly optimizes for it. Our MTL approach outperforms JTT and BR-DRO on two datasets (MNLI and Waterbirds) while performing comparatively
Table 3: Mean and standard deviations of the test worst-group accuracies across all the methods under consideration. Regularized MTL consistently reduces the gap between ERM and groupDRO when considering worst-group accuracy.

<table>
<thead>
<tr>
<th>Method</th>
<th>Group Labels</th>
<th>Civilcomments</th>
<th>MNLI</th>
<th>Waterbirds</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERM</td>
<td>Val Only</td>
<td>61.3 ± 2.0</td>
<td>67.6 ± 1.2</td>
<td>85.4 ± 4.4</td>
</tr>
<tr>
<td>JTT</td>
<td>Val Only</td>
<td>67.8 ± 1.6</td>
<td>67.5 ± 1.9</td>
<td>85.9 ± 2.5</td>
</tr>
<tr>
<td>BR-DRO</td>
<td>Val Only</td>
<td>68.9 ± 0.7</td>
<td>68.5 ± 0.8</td>
<td>86.7 ± 1.3</td>
</tr>
<tr>
<td>ERM + MT + L1</td>
<td>Val Only</td>
<td>68.2 ± 1.2</td>
<td>69.7 ± 1.5</td>
<td>87.5 ± 2.7</td>
</tr>
<tr>
<td>groupDRO (Upper Bound)</td>
<td>Train and Val</td>
<td>69.9 ± 1.2</td>
<td>78.0 ± 0.7</td>
<td>93.9 ± 0.7</td>
</tr>
</tbody>
</table>

with BR-DRO on the CivilComments dataset. Given the competitive results in Table 3, we argue that our regularized MTL formulation is an attractive option over JTT and BR-DRO. Multitasking already features prominently in many ML code bases. Thus, introducing our simple regularization modification to existing MTL implementations represents a smaller technical overhead compared to introducing JTT or BR-DRO to target worst-group error. Also, as we will see in Section 5.2 below, regularized MTL is a single approach capable of improving both worst-group and average accuracy.

5.2 Multitasking Improves both Average and worst-group Performance Even in the Absence Group Annotations

Figure 6: Comparison of the performance of different approaches with respect to average and worst-group accuracy on the Waterbirds dataset under val-GP and no-GP settings. Regularized MTL improves both average and worst-group accuracy even without group annotations. All methods enjoy lift in when validation group annotations are available.

Though previous works typically assume that practitioners have access to the group annotations on the validation set [Liu et al., 2021; Kirichenko et al., 2022], we are interested in settings where no such annotations are available. This covers many tasks of practical interest since, in some cases, it may be prohibitively cost-intensive (financially and in terms of human labor) to acquire group annotations even for the smaller validation set [Paranjape et al., 2023]. Consequently, we present a comparative performance analysis in Figure 6 encompassing settings with and without access to group annotations. With respect to worst-group accuracy, our regularized MTL approach outperforms JTT and achieves ≈ 2% lift over ERM when group
5.3 Are both Regularization and Multitasking Jointly necessary?

Table 4: Disentangling the impact of L1 regularization and SSL objective on worst-group accuracy. We see that regularized multitasking is necessary for gains in both average and worst-group performance.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>No Group Annotations</th>
<th>Val Group Annotations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Avg Acc</td>
<td>WG Acc</td>
</tr>
<tr>
<td>Waterbirds</td>
<td>JTT</td>
<td>95.6_{0.3}</td>
<td>82.1_{1.2}</td>
</tr>
<tr>
<td></td>
<td>ERM</td>
<td>95.5_{0.2}</td>
<td>80.1_{4.6}</td>
</tr>
<tr>
<td></td>
<td>+ L1</td>
<td>95.6_{0.3}</td>
<td>82.0_{5.4}</td>
</tr>
<tr>
<td></td>
<td>+ MIM</td>
<td>95.3_{0.4}</td>
<td>80.1_{4.6}</td>
</tr>
<tr>
<td></td>
<td>+ MIM + L1</td>
<td>95.8_{0.3}</td>
<td>83.3_{3.4}</td>
</tr>
<tr>
<td>Civilcomments-Small</td>
<td>JTT</td>
<td>83.3_{0.3}</td>
<td>52.5_{2}</td>
</tr>
<tr>
<td></td>
<td>ERM</td>
<td>83.9_{0.4}</td>
<td>51.6_{5.6}</td>
</tr>
<tr>
<td></td>
<td>+ L1</td>
<td>83.7_{0.4}</td>
<td>51.6_{4.0}</td>
</tr>
<tr>
<td></td>
<td>+ MLM</td>
<td>83.9_{0.3}</td>
<td>58.3_{5.6}</td>
</tr>
<tr>
<td></td>
<td>+ MLM + L1</td>
<td>84.4_{0.4}</td>
<td>53.7_{4.3}</td>
</tr>
</tbody>
</table>

In this section, we conduct an ablation to verify if both multitask learning and regularizing the final layer joint embedding space are necessary to improve average and worst-group performance. Our results are captured in Table 4. When assessing the worst-group accuracy, we find that regularizing the final embedding space during ERM can, at times, result in worse performance compared to training via standard ERM (66.3_{1.6} vs 67.4_{2.1} on CivilComments-small with validation group labels). On the other hand, multitasking without regularization can fail to improve over ERM, as evinced by the lack of improvement on Waterbirds. The regularized MTL approach is the only setting consistently improving on both datasets with and without validation group annotations. In line with these findings, we observe that joint L1 regularization and multitask learning setup yields the highest average accuracy in most cases.

5.4 Impact of Pre-Training

Table 5: Waterbirds: Impact of pre-training on average and worst-group accuracy.

<table>
<thead>
<tr>
<th>Pretrained</th>
<th>Method</th>
<th>No Group Annotations</th>
<th>Val Group Annotations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Avg Acc</td>
<td>WG Acc</td>
</tr>
<tr>
<td>no</td>
<td>ERM</td>
<td>65.1_{0.5}</td>
<td>4.5_{1.6}</td>
</tr>
<tr>
<td></td>
<td>JTT</td>
<td>67.0_{3.3}</td>
<td>10.8_{1.7}</td>
</tr>
<tr>
<td></td>
<td>ERM + MIM + L1</td>
<td>67.0_{1.3}</td>
<td>1.65_{0.7}</td>
</tr>
<tr>
<td>yes</td>
<td>ERM</td>
<td>95.5_{0.2}</td>
<td>80.1_{4.6}</td>
</tr>
<tr>
<td></td>
<td>JTT</td>
<td>95.6_{0.3}</td>
<td>82.1_{1.2}</td>
</tr>
<tr>
<td></td>
<td>ERM + MIM + L1</td>
<td>95.8_{0.3}</td>
<td>83.3_{3.4}</td>
</tr>
</tbody>
</table>

Fine-tuning pre-trained models is arguably the de-facto paradigm in machine learning (Devlin et al., 2018; Dosovitskiy et al., 2020; Dery et al., 2021). Consequently, our experiments so far have exclusively focused on pre-trained models. In this section, we wish to understand the effect of deviating from this paradigm on our MTL approach. Thus, we compare against JTT and ERM when the model is trained from scratch instead of starting with a pre-trained model. Tables 4 and 5 depict our results on Waterbirds and Civilcomments-small, respectively. Our results show that pre-training is critical for setting up regularized MTL as a viable
remedy against poor worst-group outcomes. We posit the following explanation for this outcome. Note that our informal motivation in Section 2 presupposes an ability to solve the auxiliary task to a reasonable degree. Solving the MLM and MIM tasks effectively from scratch with only the inputs of the relatively small supervised dataset is difficult. This poor performance on the auxiliary task translates to an inability to constrain the use of the spurious features on the end-task. Consistent with prior works (Tu et al., 2020; Wiles et al., 2022), our recommendation to practitioners is to use our approach during the fine-tuning of pre-trained models to be maximally effective.

Another consequence of our findings is that caution is warranted in interpreting the results of previous work on DRO in light of the new paradigm of mostly using pre-trained models. Most previous results on DRO have examined the setting of training from scratch, and as Tables 6 and 5, DRO methods significantly outperform competitors in that setting. However, the originally outsized gains in worst-group error significantly shrink when we move to pre-trained models whilst our method shows superior performance.

6 Related Work

- **Multitask Learning.** Multitask learning is a common stratergy for ML practitioners to improve the average performance of their models (Ruder, 2017; Ruder et al., 2019; Liu et al., 2019). While work like Hendrycks et al. (2019, 2020); Mao et al. (2020) have shown that multitasking can improve the adversarial and out-of-distribution robustness of models, the impact of multitasking on worst-group outcomes has been relatively unexplored. Makino et al. (2022) propose generative multitask learning (GMLT), a method that bolsters robustness to target shift by conditioning the input on all available targets, thereby addressing challenges associated with target-causing confounders and spurious dependencies between input and targets. However, it is important to note that their approach necessitates all target annotations during training, a requirement we do not assume in our scenario. Our work is inspired by Gururangan et al. (2020), who introduce constructing auxiliary objectives directly from end-task data for continued pre-training (they dub this Task Adaptive Pre-training – TAPT). Following Dery et al. (2021b, 2023), we multitask this auxiliary task with the end-task. However, unlike these studies, our focus is on improving the worst-case group accuracy of the final model. The work by Michel et al. (2021) explores the balancing of worst and average performance in multitask learning. In their study, they focus on a set of equally important end tasks, striving for proficient model performance across all of them. In contrast, our work delves into the asymmetrical multitask setting, where the presence of the auxiliary task is determined by its contribution to enhancing our target metric in the end task.

- **Robustness using group demographics.** Our multitask learning approach is primarily designed for settings with limited-to-no group annotations. However, many DRO approaches assume the presence of group annotations for all training points. Among the approaches that leverage group information, *Group Distributionally Robust Optimization* (Sagawa et al., 2020a) is the most popular technique that tries to minimize the maximum loss over the sub-groups. Goel et al. (2021) presented *Model Patching*, a data augmentation method designed to enhance the representation of minority groups. *FISH*, proposed by Shi et al. (2022), focuses on domain generalization via inter-domain gradient matching. In the settings where group annotations are expensive (financially or in terms of human resources) to procure, these methods are not viable options.

<table>
<thead>
<tr>
<th>Pretrained</th>
<th>Method</th>
<th>No Group Annotations</th>
<th>Val Group Annotations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Avg Acc</td>
<td>WG Acc</td>
</tr>
<tr>
<td>No</td>
<td>ERM</td>
<td>80.70.8</td>
<td>31.17.2</td>
</tr>
<tr>
<td></td>
<td>JTT</td>
<td>79.60.6</td>
<td>34.90.9</td>
</tr>
<tr>
<td></td>
<td>ERM+MLM+L1</td>
<td>80.70.6</td>
<td>31.30.6</td>
</tr>
<tr>
<td>Yes</td>
<td>ERM</td>
<td>83.90.4</td>
<td>51.65.6</td>
</tr>
<tr>
<td></td>
<td>JTT</td>
<td>83.30.2</td>
<td>52.50.2</td>
</tr>
<tr>
<td></td>
<td>ERM+MLM+L1</td>
<td>84.40.4</td>
<td>53.74.3</td>
</tr>
</tbody>
</table>

Table 6: Civilcomments-small : Impact of pre-training on average and worst-group accuracy.
• **Robustness without group demographics.** Extensive research has been dedicated to addressing the challenges of worst-group generalization in the more realistic scenario where access to group annotations during training is unavailable. *GEORGE* [Sohoni et al. (2020)](#) adopts a clustering-based methodology to unveil latent groups within the dataset and subsequently employs groupDRO for improved robustness. *Learning from Failure (LfF)* [Nam et al. (2020)](#) introduces a two-stage strategy. In the first stage, an intentionally biased model aims to identify minority instances where spurious correlations do not apply. In the second stage, the identified examples are given increased weight during the training of a second model. *Just Train Twice (JTT)* method [Liu et al. (2021)](#) follows a similar principle by training a model that minimizes loss over a reweighted dataset. This dataset is constructed by up-weighting training examples misclassified during the initial few epochs. Our regularized MTL approach has several advantages over these methods, even though they are all deployed in the same limited-to-no group annotations settings. As we have demonstrated, our approach can improve both worst-group and average performance, unlike the other approaches targeted against worst-group error only. Secondly, due to the widespread usage of multitask learning by many ML practitioners, implementing our modification represents minimal overhead instead of introducing one of the above bespoke approaches.

### 7 Conclusion

In this work, we presented an empirical investigation of the impact of multitasking on worst-group outcomes. We found that deploying multitasking, as is, does not consistently improve upon worst performing groups. We have shown that while DRO methods, like JTT, display superior performance when models are trained from scratch, this is not the case in the currently more widespread setting of fine-tuning a pre-trained model. Specifically, when fine-tuning, our method – regularized multitasking of the end-task with the pre-training objective constructed over end-task data – leads to improvements in worst-case group accuracy over JTT. Our work has demonstrated that it is possible to design a single, simple method that improves both worst-case group accuracy average accuracy regardless of the availability of group annotations. Since multitask learning is already a standard part of many practitioner’s toolbox, and our modification to adapt it against worst-group accuracy is simple, our approach requires minimal overhead to integrate into existing systems compared to bespoke DRO approaches. We thus encourage practitioners to introduce our modification to their MTL pipelines as an essentially free way of improving worst-group performance without sacrificing gains in average performance.

In order to keep an apples-to-apples comparison with DRO approaches, we have primarily focused on multitasking with auxiliary objectives based on end-task data only. For future work, it would be interesting to explore the impact of multitasking with auxiliary objectives based on external data more deeply. It would also be interesting to leverage meta-learning to dynamically adapt the auxiliary tasks towards improving worst-case group outcomes [Dery et al. (2021b), (2023)]. Additionally, we leave the study on the generalizability of our method to adversarial robustness, domain shift, and label shift to future work.

### 8 Acknowledgements

This work was also supported in part by the Tang AI Innovation Fund, Defence Science and Technology Agency Singapore, National Science Foundation grants IIS1705121, IIS1838017, IIS2046613, IIS2112471, and funding from Meta, Morgan Stanley, Amazon, Google, Schmidt Early Career Fellowship and Apple. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of any of these funding agencies.
References


Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 16000–16009, 2022.


A Appendix

A.1 Training Details

For $T_{\text{prim}}$, we employ a task-specific classification head of a single-layer multi-layer perceptron (MLP). For $T_{\text{aux}}$, we leverage the pre-trained MLM and MIM heads from BERT and ViT, respectively. We utilize the embedding of the [CLS] token from the base model through this MLP for classification. To facilitate effective multitask training, we adopt a task-heterogeneous batching scheme [Aghajanyan et al. 2021]. This facilitates the accumulation of gradients across tasks prior to each parameter update, contributing to improved training efficiency and convergence. Lastly, to ensure proper scaling, the L1 loss is normalized by the number of parameters in the shared representation.

A.2 Going Beyond the Pre-training Objective

Table 7: Impact of different pre-training objectives on average and worst-group accuracy. Not all auxiliary tasks can help improve worst-group performance.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>No Group Annotations</th>
<th>Val Group Annotations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Avg Acc</td>
<td>WG Acc</td>
</tr>
<tr>
<td>Waterbirds</td>
<td>ERM</td>
<td>95.5_{0.2}</td>
<td>80.1_{4.6}</td>
</tr>
<tr>
<td></td>
<td>+ MIM + L1</td>
<td>95.8_{0.3}</td>
<td>83.3_{3.4}</td>
</tr>
<tr>
<td></td>
<td>+ SimCLR + L1</td>
<td>96.1_{4.3}</td>
<td>84.0_{2.4}</td>
</tr>
<tr>
<td>Civilcomments-Small</td>
<td>ERM</td>
<td>83.9_{0.4}</td>
<td>51.6_{5.6}</td>
</tr>
<tr>
<td></td>
<td>+ MLM + L1</td>
<td>84.4_{0.4}</td>
<td>53.7_{1.3}</td>
</tr>
<tr>
<td></td>
<td>+ CLM + L1</td>
<td>83.9_{0.7}</td>
<td>50.9_{4.9}</td>
</tr>
</tbody>
</table>

Previous works in multitasking with self-supervised objectives suggest that different auxiliary objectives have disparate impacts on end task performance [Dery et al. 2023]. Out of curiosity about the impact of the choice of auxiliary objective, we explore the impact of going beyond the model’s original pre-training objective. For the Waterbirds dataset, we experiment with SimCLR – a constrastive prediction task based on determining whether two distinct augmented images originate from the same base image [Chen et al. 2020]. For BERT experiments on Civilcomments-small, we substitute the standard masked language modeling (MLM) task with causal language modeling (CLM) as the auxiliary task. From the results in Table 7, we observe that SimCLR’s performance closely resembles that of the MIM pre-training objective, whereas CLM shows relatively inferior results compared to MLM. We hypothesize that BERT’s intrinsic bidirectional attention mechanism and non-autoregressive nature are not ideally suited for causal language modeling [Song et al. 2019], resulting in the model underperforming in our multitask setup. Given the model performance’s sensitivity to the replacement objective’s choice, we proffer a practical recommendation to practitioners: use the pre-training objective as the auxiliary task. This aligns with recent work on best practices for fine-tuning pre-trained models [Goyal et al. 2023].

A.3 Bayes optimal model for dimension-independent reconstruction under noised inputs

$$\ell(w_i) = \frac{1}{2} \mathbb{E} \left[ (x_i - w_i \hat{x}_i)^2 \right]$$

$$= \frac{1}{2} \mathbb{E} \left[ (x_i)^2 - 2w_i x_i \hat{x}_i + (w_i \hat{x}_i)^2 \right]$$

$$\frac{\partial \ell(w_i)}{\partial w_i} = -\mathbb{E} [x_i \hat{x}_i] + w_i \mathbb{E} \left[ (\hat{x}_i)^2 \right]$$
The optimal weighting $w^*_i$ is achieved when $\frac{\partial (w_i)}{\partial w_i} = 0$.

$$
 w^*_i = \frac{E[x_i\tilde{E}]}{E[(\tilde{x}_i)^2]}
 = \frac{E[x_i(x_i + \epsilon_i)]}{E[(x_i + \epsilon_i)^2]}
 = \frac{E[(x_i)^2] + E[x_i]E[\epsilon_i]}{E[(x_i)^2] + 2E[x_i]E[\epsilon_i] + E[\epsilon_i]^2}
$$

note $E[\epsilon_i] = 0$, $E[(x_i)^2] = \sigma^2_i + 0.5 (\mu_{i|y=1}^2 + \mu_{i|y=-1}^2)$

$$
\text{E} = \frac{\sigma^2_i + 0.5 (\mu_{i|y=1}^2 + \mu_{i|y=-1}^2)}{\sigma^2_i + 0.5 (\mu_{i|y=1}^2 + \mu_{i|y=-1}^2) + \sigma^2_{\text{noise}}}
$$

### B Broader Impact Statement

In terms of broader impact, our work has beneficial implications for group fairness and mitigating demographic-based bias in machine learning systems. Specifically, we present a simple approach to improving worst-case group error. This means that our work contributes positively to certain groups that would otherwise be impacted by the poor performance of ML models. We have also provided a new lens from which to view the problem of the impacts of multitasking, resulting in showing that it is tangible against poor group-based outcomes. However, our method introduces compute overhead due to the optimization of multiple objectives. This results in increased power consumption and, thus, greenhouse emissions during the training of models. Nevertheless, introducing our simple regularization modification to existing MTL implementations represents a negligible technical overhead than introducing JTT or BR-DRO to target worst-group error.

Table 8: Sensitivity analysis of all the methods to different hyperparameters wrt worst-group accuracy.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>Seed</th>
<th>Learning Rate</th>
<th>Batch Size</th>
<th>Up</th>
<th>T</th>
<th>$\lambda_{aux}$</th>
<th>$\lambda_{reg}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Waterbirds</td>
<td>ERM</td>
<td>0.0734</td>
<td>0.5717</td>
<td>-0.4384</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>JTT</td>
<td>0.0573</td>
<td>-</td>
<td>-0.380</td>
<td>0.0246</td>
<td>0.1862</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>ERM+MT+L1</td>
<td>-0.0314</td>
<td>-</td>
<td>0.0142</td>
<td>-</td>
<td>-</td>
<td>0.0943</td>
<td>0.4114</td>
</tr>
<tr>
<td></td>
<td>groupDRO</td>
<td>-1.688</td>
<td>0.196</td>
<td>-0.1384</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Civilcomments-Small</td>
<td>ERM</td>
<td>0.0617</td>
<td>-0.80</td>
<td>0.1029</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>JTT</td>
<td>-0.0315</td>
<td>-</td>
<td>0.2104</td>
<td>-0.0047</td>
<td>0.1111</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>ERM+MT+L1</td>
<td>0.0677</td>
<td>-</td>
<td>0.0453</td>
<td>-</td>
<td>-</td>
<td>0.0601</td>
<td>-0.2955</td>
</tr>
<tr>
<td></td>
<td>groupDRO</td>
<td>-0.0392</td>
<td>-0.7352</td>
<td>-0.1726</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**Sensitivity to Hyperparameters:** For evaluating the sensitivity of worst group accuracy to various hyperparameters across different methods, we employ Kendall’s rank coefficient $\tau$ [Kendall 1938]. To adhere to our computation budget, we opt for the optimal learning rate utilized in ERM for both JTT and ERM+MT+L1. Our sensitivity analysis from Table 8 reveals distinct preferences in learning rates. The ViT model applied to waterbirds, demonstrating a preference for higher learning rates. Conversely, when trained on the civilcomments-small dataset, BERT exhibits a preference for lower learning rates. The epoch at upsampling is also a sensitive hyperparameter for JTT. The sensitivity to different seed values is negligible across all the methods. Notably, our method displays the most minor sensitivity to changes in batch size. However, it is noteworthy that our method is most responsive to variations in the regularization parameter ($\lambda_{reg}$), with lower values resulting in higher worst group accuracy.