A. Image Classification on CIFAR10-Skewed

The UTKFace dataset is an excellent choice for our main paper as it contains valuable metadata and has been extensively studied in trustworthy ML literature. However, it does come with certain limitations, such as its focus on binary classification tasks, and being confined to a highly specialized domain, i.e., facial images. To address these limitations, we will conduct a second case study using a skewed version of the CIFAR-10 dataset (more details below).

A.1. Experiment Setup

Dataset We will adopt the CIFAR10-Skewed setup from Wang *et al*. [40] for our case study. In this setup, the 10 object classes of CIFAR10 [20] are divided into two groups, i.e., colour majority and grayscale majority. The first 5 classes (*airplane, automobile, bird, cat, deer*) are marked as the colour majority, i.e., 95% of images from these classes are left as is, while the other 5% are converted into grayscale images. Conversely, the last 5 classes (*dog, frog, horse, ship, truck*). are marked as the grayscale majority, i.e., 95% of images from these classes are converted into grayscale images, while the other 5% are left as is.

Training Details The training details are the same as the UTKFace Setup, except that we train the models for only 20 epochs on CIFAR10-Skewed.

Axes of Multiplicity The axes of multiplicity are the same as the UTKFace Setup, except for the model architecture. We will only use a modified ResNet18 model (adapted for CIFAR10 images of size 32x32) and we will not study multiplicity across changing architecture in this case study.

A.2. Accuracy

The multiplicity sheet for accuracy on the CIFAR10-Skewed dataset is created in Fig. 10. Note that the test dataset for CIFAR10-Skewed, i.e., the dataset on which we measure this accuracy, is also skewed and has the same formulation as the training dataset defined above. The trends of accuracy multiplicity are quite similar to that of UTKFace, i.e., no significant accuracy variance is present across any hyperparameter choice or random seeds.

A.3. Group Fairness

For the CIFAR10-Skewed dataset, grayscale images in the 'colour majority' object classes and colour images in the 'grayscale majority' object classes are both minority groups. For this particular case study, we will measure the group accuracy of the GS Minority, i.e., the grayscale minority in the colour majority object classes, as the fairness score under intervention. The results are collected in Fig. 11.

As previously noted, the multiplicity sheet in Fig. 11 highlights the importance of random seeds in fairness. While other hyperparameter choices do have a noticeable impact (in order - training data augmentation, batch size, learning rate, and the optimization algorithm), clearly the most consistently dominant source of fairness multiplicity is the randomness in model training. Moreover, the overall fairness multiplicity ($\Delta^{all}_{max} = 11.16$) is almost seven times higher than the accuracy multiplicity ($\Delta^{all}_{max} = 1.63$), further highlighting the severe impact of multiplicity on trustworthy metrics.

A.4. Out-of-Distribution Robustness

We will use accuracy on a grayscale version of the CIFAR10 dataset [20] as the measure of our OOD robustness multiplicity. In Fig. 12, we present the multiplicity sheet for Accuracy on the CIFAR10-GS dataset. The results show similar trends to robustness multiplicity for UTKFace, with both hyperparameter choices and the random seed being equally important in affecting the model's robustness. The range of overall robustness multiplicity ($\Delta^{all}_{max} = 4.01$) is a little more than two times higher than accuracy multiplicity, which is unsurprising since despite being grayscale, the test images still belong to CIFAR10. A more severe robustness check on a dataset that is quite different from CIFAR10 might introduce and even higher OOD robustness multiplicity.

A.5. Differential Privacy

We use the same perturbation and trade-off setup for privacy multiplicity as done for UTKFace, i.e., we will measure the accuracy of the model under output perturbations from an exponential distribution with a fixed *rate parameter* λ. We present the multiplicity sheet for privacy by recording the Perturbation Accuracy with $\lambda = 5$ in Fig. 13. The same trends as Fig. 5 are noticed, i.e., the random seed has minimal impact and it's the hyperparameters that dramatically influence the privacy multiplicity, in line with the existing literature on practical tips for privacy [28]. The overall privacy multiplicity range ($\Delta^{all}_{max} = 10.28$) is also almost six times larger than the accuracy multiplicity but would depend on the rate parameter λ.

A.6. Security against Adversarial Attacks

We will use the same setup for security multiplicity as UTKFace, i.e., we use projected gradient descent (PGD) [23] to progressively move out of the local minima until we reach the given distance budget, and then measure the accuracy of the model under these perturbations. We present the multiplicity sheet for PGD Accuracy with $\delta = 0.005$ in Fig. 14. The trends are again similar to the ones seen in the main paper, i.e., no single factor dominates the multiplicity. The overall multiplicity range ($\Delta^{all}_{max} = 4.51$) for security is almost five times larger than the accuracy multiplicity and clearly depends on the adversarial distance budget δ.

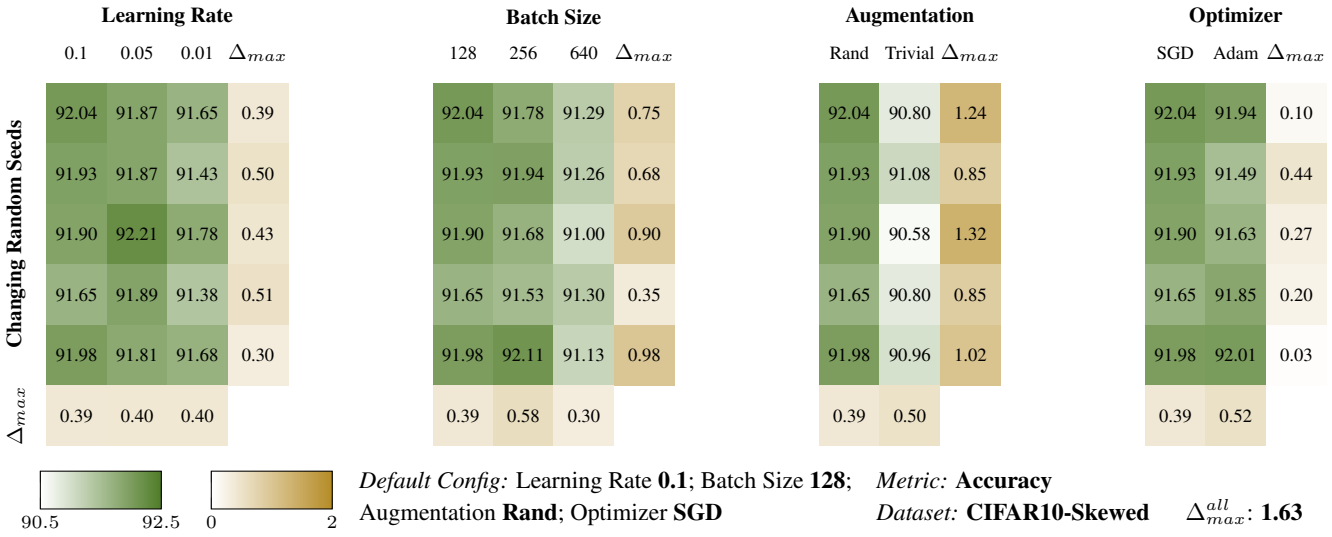


Figure 10. *Multiplicity sheet* for **Accuracy** on **CIFAR10-Skewed** dataset.

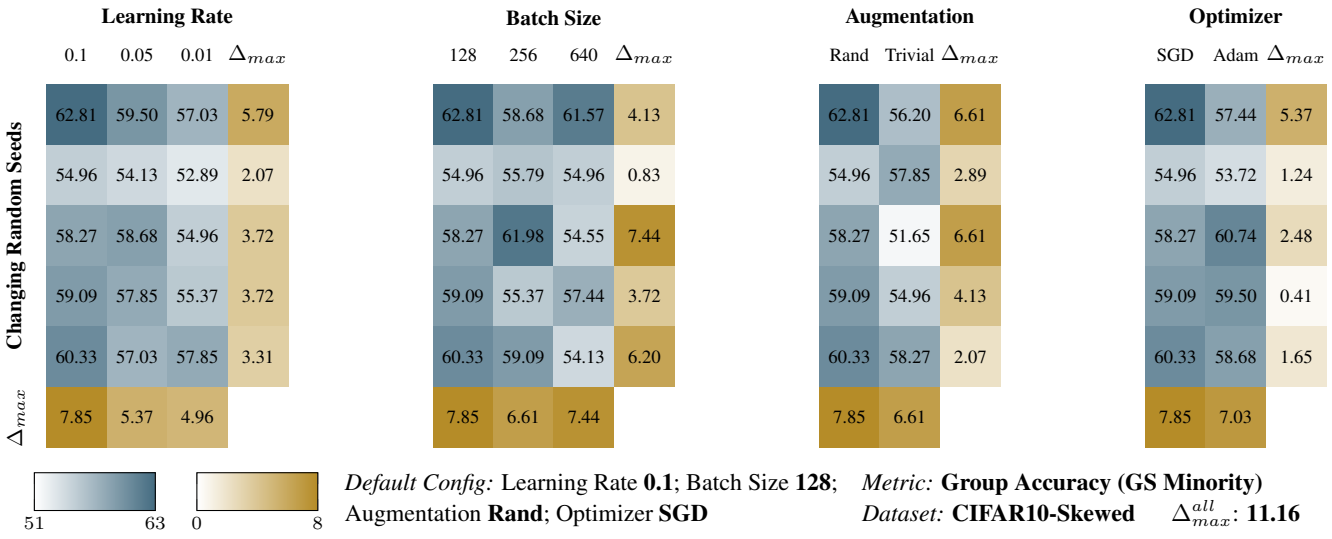
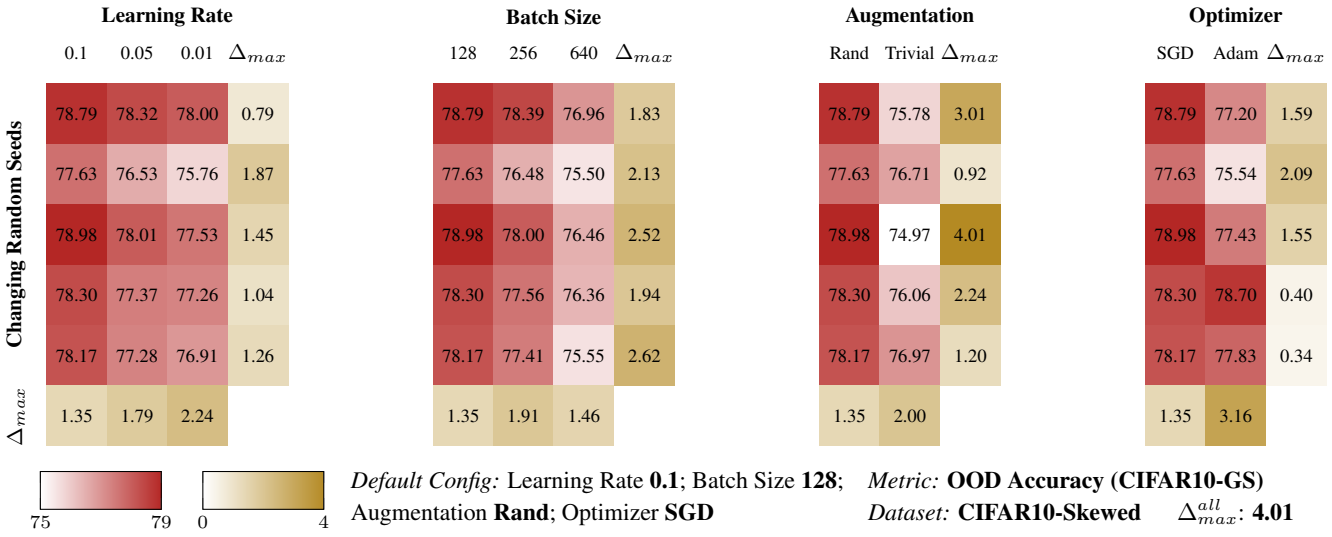


Figure 11. *Multiplicity sheet* for **Group Accuracy (GS Minority)** on **CIFAR10-Skewed** dataset.

A.7. Discussion

We have provided an additional case study on the CIFAR10-Skewed dataset as a companion to our main case study on the UTKFace dataset. These results help us cement certain trends, for example, the impact of random seeds on fairness, the impact of hyperparameter choices on privacy-utility trade-off, etc., all of which are unsurprising as these trends have been noted previously in the literature (albeit in isolated settings). We believe these experiments will serve as a useful companion to our main paper, and help establish the importance of multiplicity sheets in image classification.

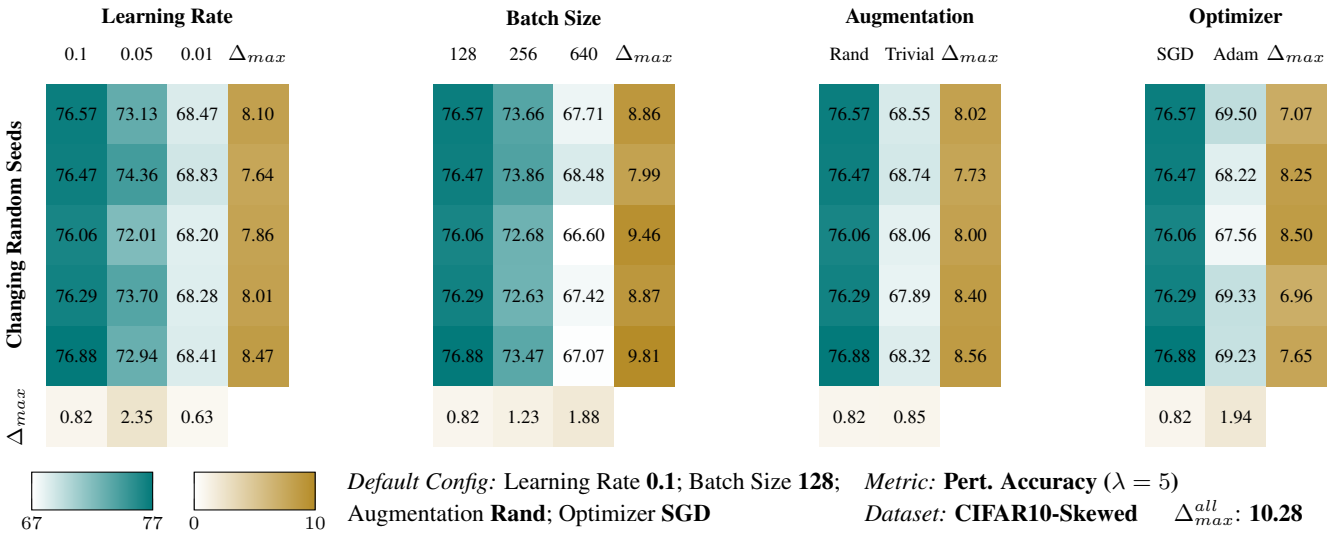


Default Config: Learning Rate **0.1**; Batch Size **128**; Augmentation **Rand**; Optimizer **SGD**

Metric: **OOD Accuracy (CIFAR10-GS)**

Dataset: **CIFAR10-Skewed** Δ^{all}_{max}: **4.01**

Figure 12. *Multiplicity sheet* for **OOD Accuracy (CIFAR10-GS)** on **CIFAR10-Skewed** dataset.



Default Config: Learning Rate **0.1**; Batch Size **128**; Augmentation **Rand**; Optimizer **SGD**

Metric: **Pert. Accuracy** ($\lambda = 5$)

Dataset: **CIFAR10-Skewed** Δ^{all}_{max}: **10.28**

Figure 13. *Multiplicity sheet* for **Perturbation Accuracy** ($\lambda = 5$) on **CIFAR10-Skewed** dataset.

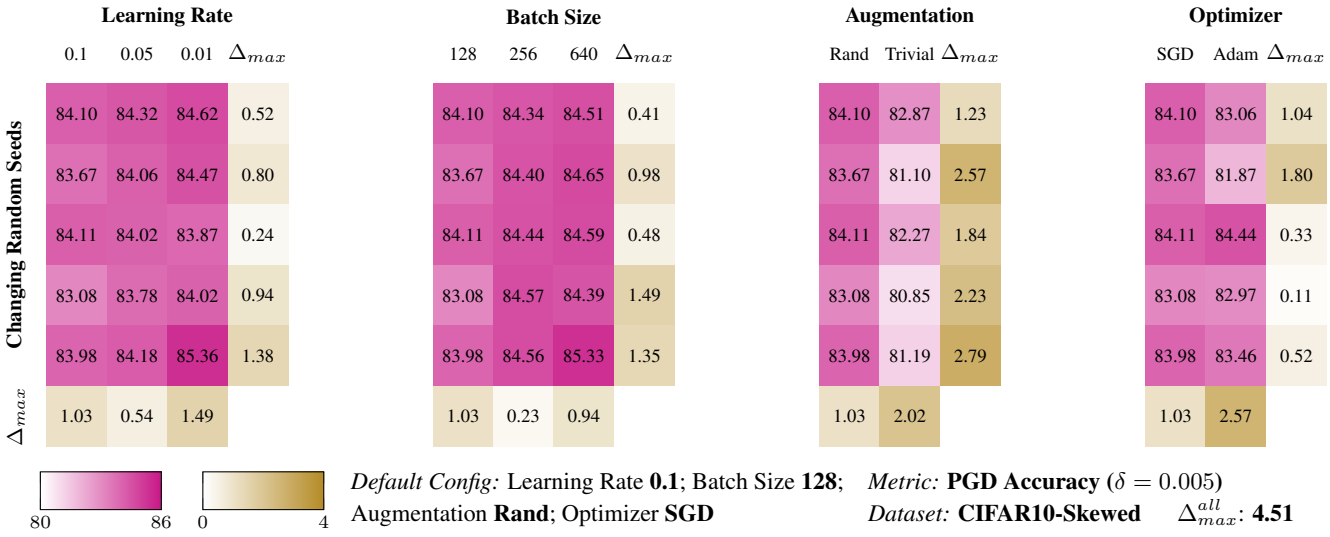


Figure 14. *Multiplicity sheet* for **PGD Accuracy** ($\delta = 0.005$) on **CIFAR10-Skewed** dataset.