# A ADDITIONAL EXPERIMENTS

We present additional experimental results that explore a few key hypotheses:

- how sensitive are the results to inexact subgroup specification?
- how does the precise choice of target label and spurious attribute affect results?
- how generalisable are the results to multi-class settings?

## A.1 INEXACT SUBGROUP SPECIFICATION

The post-hoc modification techniques in the body crucially rely on knowledge of the precise subgroup specification of each example. This is unrealistic in practice, where the subgroups may be latent or inexactly specified. Following Sagawa et al. (2020a)[Appendix B], we simulate a setting of inexact specification of the subgroups on `celebA`. Here, we use as spurious attributes $\mathcal{A}' = \texttt{WearingLipstick} \times \texttt{Eyeglasses} \times \texttt{Smiling} \times \texttt{DoubleChin} \times \texttt{OvalFace}$, comprising 32 distinct values. We then learn using the subgroups $\mathcal{Y} \times \mathcal{A}'$ as input, and then measure performance with respect to the *original* subgroups $\mathcal{Y} \times \mathcal{A}'$.

As noted in the body, the threshold adjustment technique (THR) is challenging to apply as-is in this setting, as it requires setting 32 distinct thresholds. As suggested in §4.3, we thus apply a simple heuristic of tying the thresholds to the subgroup frequencies, i.e., $t_{a(x)} = \log \mathbb{P}(y = +1 \mid a(x)) - \log \mathbb{P}(y = -1 \mid a(x))$. This has the effect of implicitly requiring higher model confidences to classify examples into a dominant subgroup. This technique is seen to have a worst-subgroup error of $16.67\%$, which is a modest increase compared to the $12.10\%$ obtained when using the exact subgroups $\mathcal{Y} \times \mathcal{A}'$. This illustrates that one can still make useful predictions given imperfect subgroup information.

## A.2 CHOICE OF TARGET LABEL AND SPURIOUS ATTRIBUTE

The results in the body involve data with one or more rare subgroups. Is this rarity the primary factor that influences performance, or does the definition of the subgroups themselves matter? To test this, we consider a variant of the `celebA` dataset in the body where the target label and sensitive attribute are swapped. In this variant of the dataset, we have $\mathcal{Y} = \{\texttt{male}, \texttt{female}\}$ and $\mathcal{A} = \{\texttt{blond}, \texttt{dark}\}$. This exactly preserves the subgroup definitions and their rarity, but fundamentally changes the target label and feature used in training.

Interestingly, this simple modification dramatically improves performance of the baseline: on the rarest subgroup, the error is $13.67\%$, which is a significant reduction over the $56.94\%$ for the original dataset. This indicates that the precise choice of subgroup definition can play a non-trivial role in final performance. Intuitively, performance can be hampered when the target variable is spuriously correlated with many features in the training set.

Nonetheless, even with this improved model, we find that threshold adjustment (THR) can further improve performance to $9.11\%$. The average subgroup errors of both techniques are similar, being $1.29\%$ and $1.28\%$ respectively.

## A.3 MULTI-CLASS SETTINGS

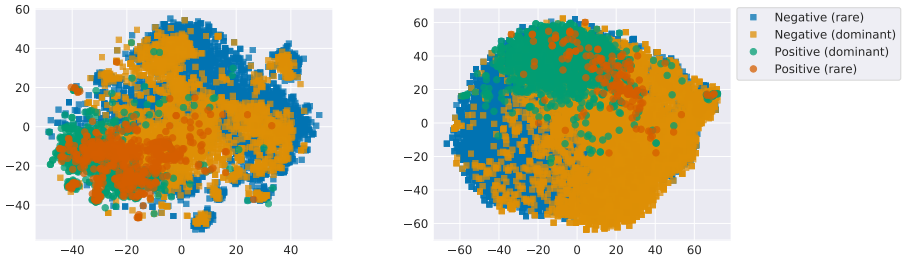
The results in the body involve problems with binary labels. To assess the effect of working with multi-class labels, we employ a modified version of MNIST based on Goel et al. (2020). Here, one mixes the standard MNIST dataset with samples from a corrupted version of MNIST comprising zig-zag images. The zig-zag images are made to be strongly correlated with the digit parity, so that most odd digits are zig-zagged. We then consider subgroups defined by $\mathcal{Y} \times \mathcal{A}$, where $\mathcal{A} = \{\texttt{normal}, \texttt{zig} - \texttt{zag}\}$. Note that we consider $\mathcal{Y} = \{0, 1, \ldots, 9\}$ to illustrate performance in a multi-class setting, unlike Goel et al. (2020) who consider $\mathcal{Y} = \{0, 1\}$ to be the digit parity.

We train a LeNet-5 for $100$ epochs using a learning rate of $0.0001$, momentum $0.9$, weight decay $0.05$, and batch size $100$. Here, ERM achieves a worst-subgroup error of $67.11\%$. Classifier retraining (cRT) based on subsampling all non-minority samples improves this to $74.52\%$. Similarly, threshold adjustment (THR) based on the heuristic of tying the thresholds to the subgroup frequencies (as

described in the previous section) achieves $78.95\%$. This illustrates the potential for post-hoc techniques to also be useful in scenarios other than binary classification.

## B VISUALISATION OF EMBEDDINGS UNDER GROUP-BASED DRO

Figure 6 shows a tSNE visualisation of the embeddings learned by a model trained to minimise the group-based DRO objective of Sagawa et al. (2020a). Similar to the results of ERM, there is generally a notable separation of samples from the four subgroups.




(a) `waterbirds` (b) `celebA`

Figure 6: Two-dimensional tSNE visualisation of test embeddings as produced by an overparameterised model trained to minimise group-based DRO.

## C ADDITIONAL EXPERIMENTAL ABLATIONS

We present additional experimental results, highlighting several key points:

- the separation of scores between rare and dominant subgroups is consistent across all datasets considered in the paper; however, the score distributions are markedly different on train and test sets, owing to models being overly confident on training samples.
- increasing model complexity systematically exacerbates the distribution shift in decision scores.
- early stopping has limited effect on the score distributions; even after a single of epoch of training, there may be a distinction between rare and dominant subgroups' scores.
- increased $\ell_2$ regularisation strength has a favourable effect on the score distributions, encouraging samples from both rare and dominant subgroups to be correctly classified.
- subsampling (per Sagawa et al. (2020a)) has a positive effect on the score distributions, making them almost perfectly align across subgroups.
- increasing the fraction of majority samples has a deleterious effect on overall performance; however, even at extreme levels of imbalance, the score distribution for rare samples may be shifted to correct for bias.

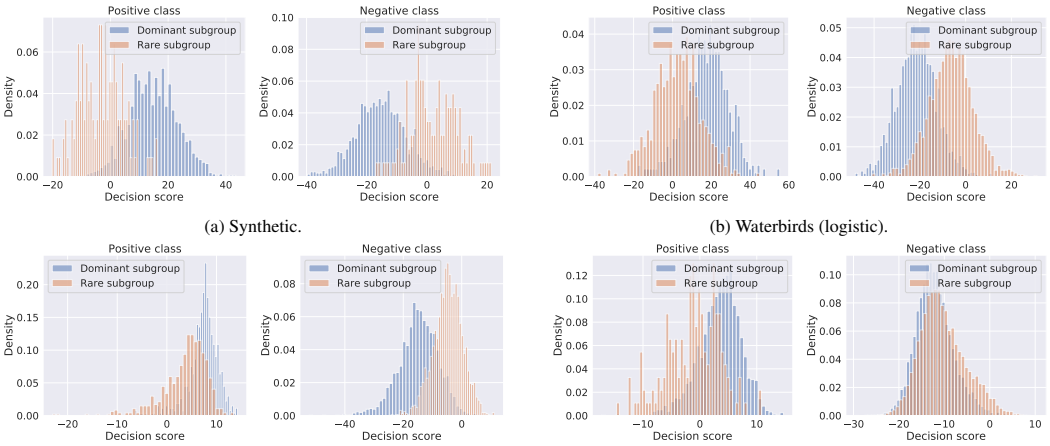
### C.1 HISTOGRAM OF TRAIN AND TEST SCORES

Figure 7 plots histograms of test scores for all datasets considered in this paper. We consistently find that there is a separation between the scores for rare and dominant subgroups.

We see similar behaviour on training scores in Figure 8. However, note the vastly different scale, owing to the model being more confident in its predictions for these samples. In general, while there are differences in the distributions for the rare and dominant subgroup scores, nearly all such scores are on the correct side of the decision boundary. This is expected, since overparameterised models perfectly fit the training data, and thus correctly classify all samples. The ability of these models to nonetheless produce meaningful results on test samples is owing to their inductive bias.

### C.2 IMPACT OF MODEL COMPLEXITY ON SCORES

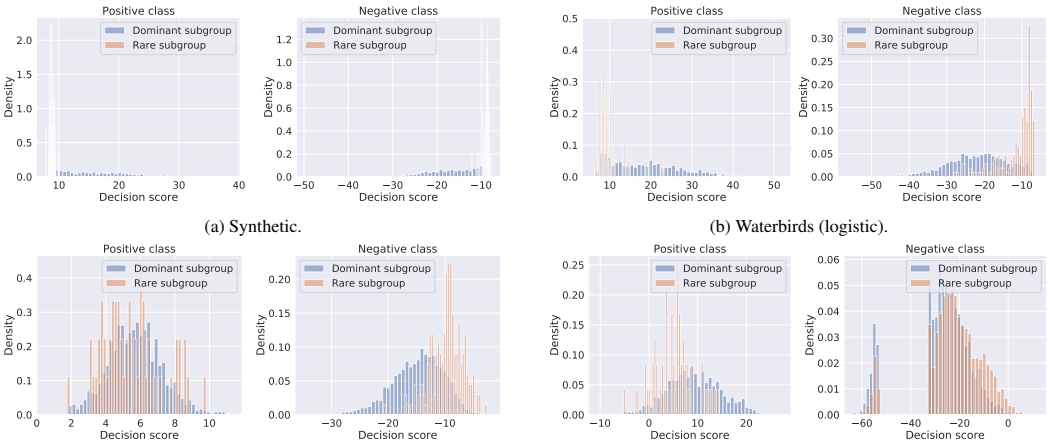
Figure 9 shows model scores on test samples on `synth` as number of projection features $m$ is varied. We see that as the model complexity increases, there is a steady increase in the separation




(a) Synthetic. (b) Waterbirds (logistic).

(c) Waterbirds (ResNet). (d) CelebA.

Figure 7: Histograms of model scores on test samples on various datasets, comprising two labels with two sub-groups each. In general, there are differences in the distributions for the rare and dominant subgroup scores, with the former often lying on the incorrect side of the decision boundary.




(a) Synthetic. (b) Waterbirds (logistic).

(c) Waterbirds (ResNet). (d) CelebA.

Figure 8: Histograms of model scores on train samples on various datasets, comprising two labels with two sub-groups each. In general, while there are differences in the distributions for the rare and dominant subgroup scores, nearly all such scores are on the correct side of the decision boundary.

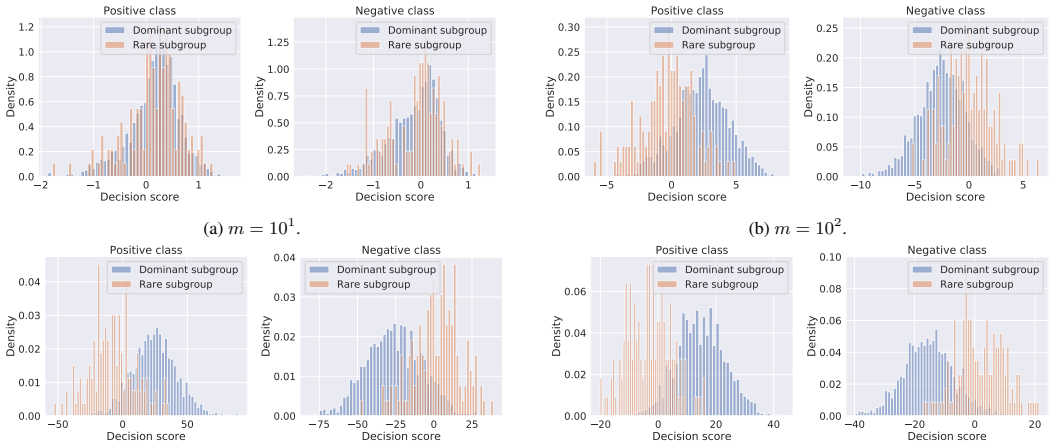
of decision scores between the rare and dominant subgroups for a label. This is in keeping with overparameterisation exacerbating worst-subgroup error: as the decision scores have more pronounced separation, using a default classification threshold will lead to significantly worse performance.

### C.3 Impact of early stopping on scores

Figures 10 and 11 shows the evolution of model scores on test samples on the CelebA and Waterbirds datasets. Here, the distinction between the scores amongst subgroups of the positive class is visible even after early stopping. With increased training epochs, there is a systematic shift of the negative scores, as the network becomes increasingly confident on them.

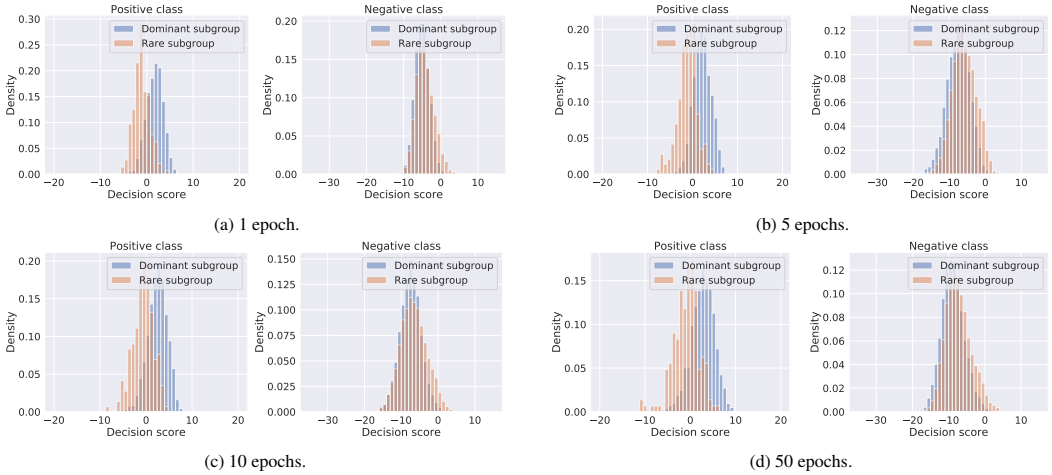
### C.4 Impact of regularisation on scores

Figures 12 and 13 shows how model scores on test samples vary as we modify the strength of regularisation. Increasing the strength is seen to favourably impact the scores on the negative class,




(a) $m = 10^1$. (b) $m = 10^2$. (c) $m = 10^3$. (d) $m = 10^4$.

Figure 9: Histograms of model scores on test samples on `synth` as number of projection features $m$ is varied. We see that as the model complexity increases, there is a steady increase in the separation of decision scores between the rare and dominant subgroups for a label. This is in keeping with overparameterisation exacerbating worst-subgroup error.



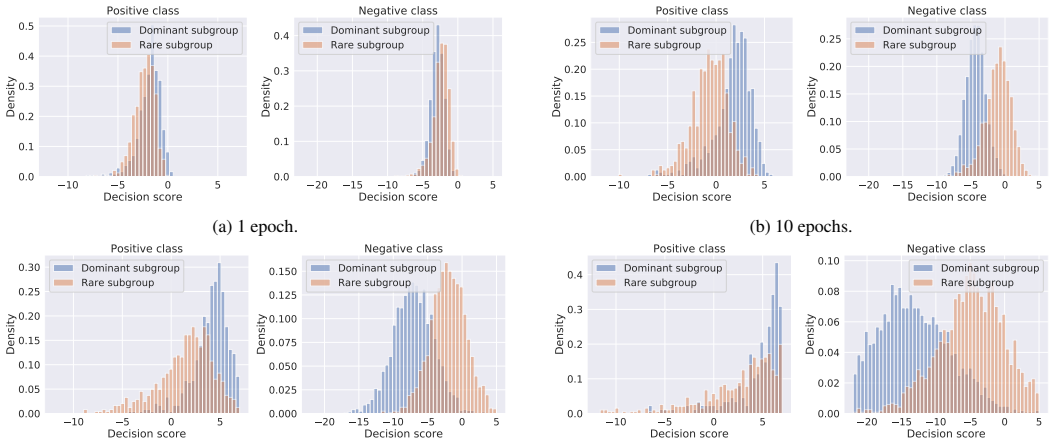

(a) 1 epoch. (b) 5 epochs. (c) 10 epochs. (d) 50 epochs.

Figure 10: Evolution of histograms of model scores on test samples on `celebA`. The distinction between the scores amongst subgroups of the positive class is visible even after early stopping. With increased training epochs, there is a systematic shift of the negative scores, as the network becomes increasingly confident on them.

for both the dominant and rare subgroups. This provides another perspective on why regularisation can be somewhat effective at improving worst-subgroup performance.

### C.5 Impact of subsampling on scores

Figure 14 shows histograms of model scores on test samples on synthetic dataset, with and without subsampling per Sagawa et al. (2020a). Subsampling is seen to make the scores equitable across the subgroups, which provides another perspective on why this technique can effectively mitigate a bias against rare subgroups.



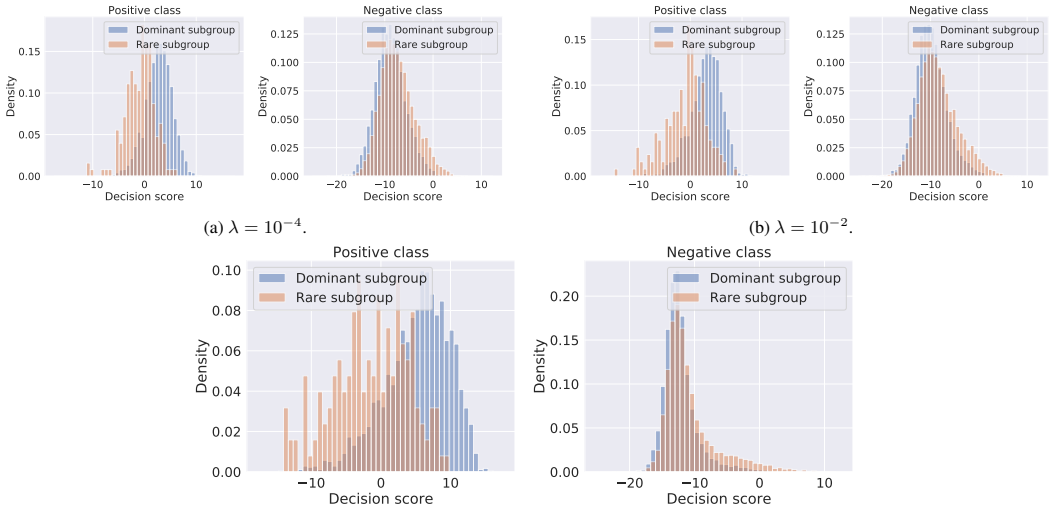

(a) 1 epoch.

(b) 10 epochs.

(c) 100 epochs.

(d) 1000 epochs.

Figure 11: Evolution of histograms of model scores on test samples on `waterbirds`.




(a) $\lambda = 10^{-4}$.

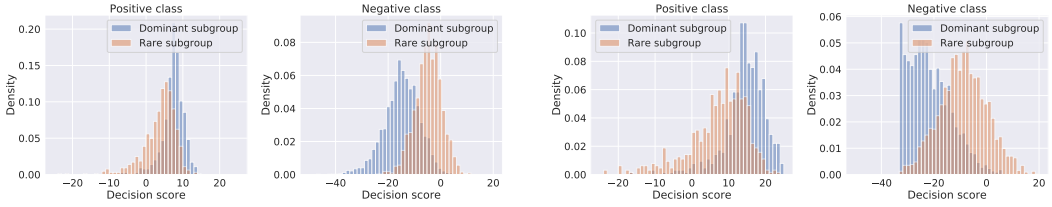
(b) $\lambda = 10^{-2}$.

(c) $\lambda = 10^{0}$.

Figure 12: Histograms of model scores on test samples on `celebA` with various strenghts of regularisation. Increasing the strength is seen to favourably impact the scores on the negative class, for both the dominant and rare subgroups.

### C.6 Impact of fraction of rare subgroups on scores

The `synth` dataset involves a parameter $p_{\text{dom}}$in its construction, which controls the relative number of samples belonging to the dominant class. By default, following Sagawa et al. (2020b), we use $p_{\text{dom}}= 0.90$. Figure 15 shows how the tradeoff is affected by changing $p_{\text{dom}}$. As $p_{\text{dom}}$increases, as expected, it is more challenging to minimise the worst-subgroup error at large $m$. Figure 16 further shows how the test scores for each subgroup are affected by the choice of $p_{\text{dom}}$. As $p_{\text{dom}}$increases, the rare subgroup scores are seen to significantly diverge from the dominant one.



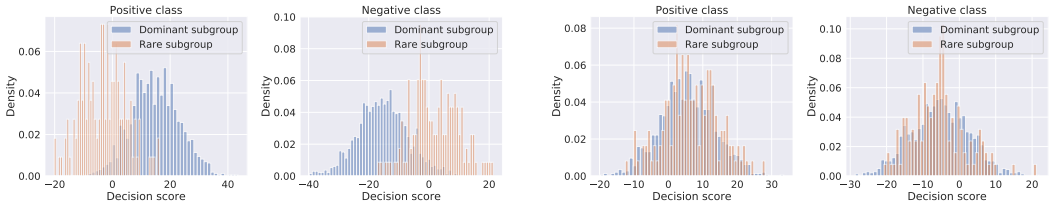

(a) $\lambda = 10^{-4}$.




(b) $\lambda = 10^{-2}$.


(c) $\lambda = 10^{0}$.

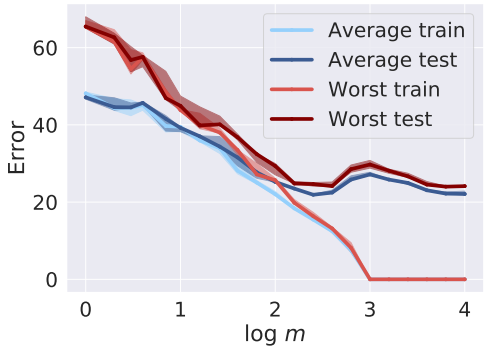
Figure 13: Histograms of model scores on test samples on `waterbirds` with various strengths of regularisation. Increasing the strength is seen to favourably impact the scores on the negative class, for both the dominant and rare subgroups.



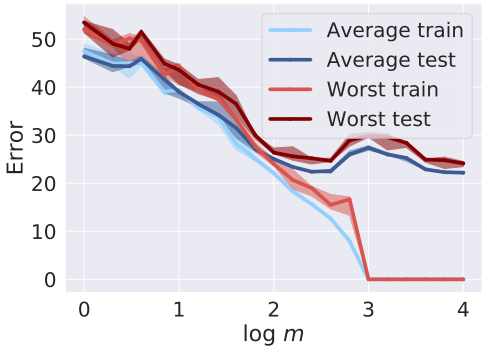

(a) Baseline.

(b) Baseline + subsampling.

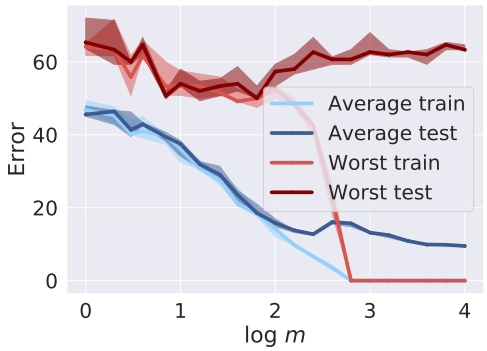
Figure 14: Histograms of model scores on test samples on `synth`, with and without subsampling per Sagawa et al. (2020a). Subsampling is seen to make the scores equitable across the subgroups.



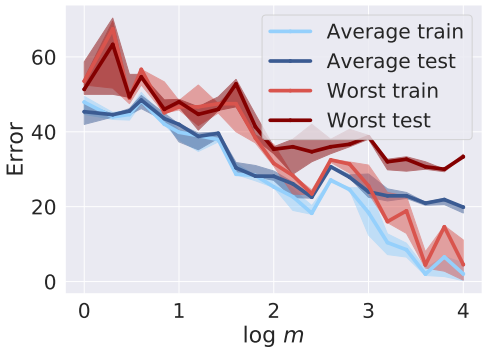

(a) Baseline $p_{\mathrm{maj}} = 0.5$.



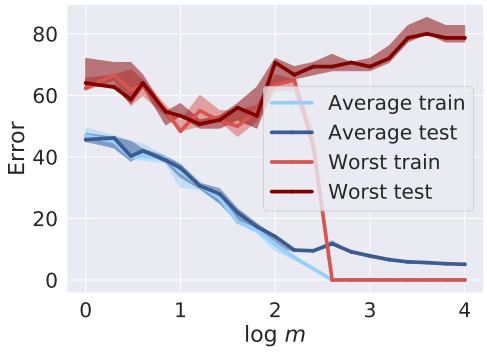

(b) Threshold correction $p_{\mathrm{maj}} = 0.5$.



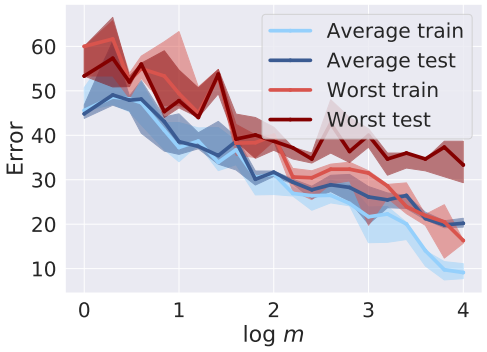

(c) Baseline $p_{\mathrm{maj}} = 0.9$.




(d) Threshold correction $p_{\mathrm{maj}} = 0.9$.

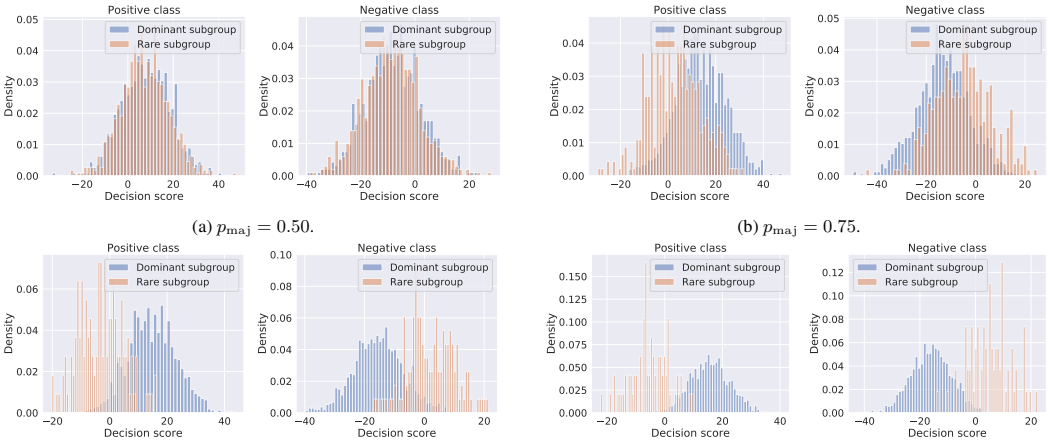



(e) Baseline $p_{\mathrm{maj}} = 0.95$.




(f) Threshold correction $p_{\mathrm{maj}} = 0.95$.

Figure 15: Performance of baseline and threshold correction on `synth`. We vary the fraction of majority samples $p_{\mathrm{maj}}$. This is seen to adversely affect the performance of ERM. However, threshold correction can effectively reduce this error, albeit at the expense of higher variance.




(a) $p_{\mathrm{maj}} = 0.50$.

(b) $p_{\mathrm{maj}} = 0.75$.

(c) $p_{\mathrm{maj}} = 0.90$.

(d) $p_{\mathrm{maj}} = 0.95$.

Figure 16: Histograms of model scores on test samples on `synth`, for various values of fraction of majority samples $p_{\mathrm{dom}}$. As this fraction becomes larger, the rare subgroups see progressive shift in their scores compared to the dominant ones.