

## A APPENDIX

### A.1 THEORETICAL RESULTS FOR CMNIST+

*Fitting Color for Classification.* In expectation, the ERM model would learn color as the feature representation for classification. Here, we show the theoretical results for such cases. When the model learns color  $C$  as the feature representation  $F(X)$ , we have:

$$\begin{aligned} P(Y|C) &= \sum_E P(Y|C, E)P(E|C) \\ &= \sum_E \frac{P(C|Y, E)P(Y|E)}{P(C|E)}P(E|C) \\ &= \sum_E P(C|Y, E)P(Y|E) \frac{P(E)}{P(C)} \end{aligned} \quad (9)$$

The second equality is by Bayes' rule. We know that  $P(C) = \sum_Y \sum_E P(C|Y, E)P(Y|E)P(E)$ . Then, given  $P(C|Y, E)$ ,  $P(Y|E)$ ,  $P(C)$  and Eq. 9, we can obtain  $P(Y|C)$  as shown in Table 4. These results imply that the deterministic classifier  $\hat{P}(Y = 1|C = G) = 1$ ,  $\hat{P}(Y = 1|C = B) = 1$ ,  $\hat{P}(Y = 1|C = R) = 0$  for all  $\rho \in [0.55, 0.9]$  as shown in Table 4. So, the accuracy of the deterministic classifier  $\hat{P}(Y|C)$  on the test set is 0.2. Its accuracy on the training set is  $\sum_C P(\hat{Y} = Y|C)P(C)$  where  $P(C) = \sum_E \sum_Y P(C|Y, E)P(Y|E)P(E)$ .

*Fitting Domain Label for Classification.* Since the spurious correlation between the domain label and the class label is strong, and IRM cannot penalize models fitting the domain label,  $P(Y|E)$  can help us understand the expected behavior of the IRM model. Note that the test data is from an unseen domain. So, we analyze the model that first predicts domain by color, then predicts class label by domain. First, we analyze  $P(E|C)$  as below:

$$\begin{aligned} P(E|C) &= \sum_Y P(E|C, Y)P(Y|C) \\ &= \sum_Y \frac{P(C|Y, E)P(E|Y)P(Y|C)}{P(C|Y)} \\ &= \sum_Y \frac{P(C|Y, E)P(E|Y)P(Y)}{P(C)} \\ &= \sum_Y \frac{P(C|Y, E)P(Y|E)P(E)}{P(C)}, \end{aligned} \quad (10)$$

where the second and forth qualities are by Bayes' rule. With Eq. 10, we can list the values of  $P(E|C)$  in Table 6 for  $\rho \in [0.55, 0.9]$ . Thus, we know the deterministic domain prediction results would be  $\hat{P}(E = 1|C = G) = 1$ ,  $\hat{P}(E = 1|C = B) = 0$ , and  $\hat{P}(E = 1|C = R) = 0$ . Since  $P(Y|E)$  is given in Table 1, we can obtain the deterministic classifier's predictions as:

$$\hat{P}(Y = 1|C = G) = 1, \hat{P}(Y = 1|C = B) = 0, \hat{P}(Y = 1|C = R) = 0. \quad (11)$$

So, the expected test accuracy of the model would be 0.35. Recall that the model first predicts domain label by color and then predict class label by the predicted domain. By doing this, it would have  $F(X) \approx E$ . This makes it approximately satisfy the IRM constraint  $Y \perp\!\!\!\perp E|F(X)$  since  $Y \perp\!\!\!\perp E|E$ . This implies that the model's performance can be treated as the expected performance of the IRM model in CMNIST+. In terms of the performance of  $\hat{P}(Y|\hat{E})$  on the training set, we can get the results using the same prediction rules as in Eq. 11.

One may argue that the IRM model can perform well if we balance the two classes in each domain. Here, we theoretically show this is not the case. By setting  $P(Y|E) = 0.5$ , we can obtain the values of  $P(E|C)$  in Table 7. Note that practically this can be done by oversampling the minority class of each domain in each mini-batch. However, since  $P(Y|E) = 0.5$ , the predictions made by the deterministic classifier  $\hat{P}(Y|E)$  would be just random guess, leading to a test accuracy of 0.5.

*Fitting both Domain and Color for Classification.* Here, we consider the model that first predicts the domain label by color and then predicts the class label by both the color and the predicted domain label. The first step is the same as the model fitting the domain label. For the second step, we analyze  $P(Y|C, E)$  as below:

$$P(Y|C, E) = \frac{P(C|Y, E)P(Y|E)}{P(C|E)}. \quad (12)$$

With  $P(C|E) = \sum_Y P(C|Y, E)P(Y|E)$  and Eq. 12, we obtain values of  $P(Y|C, E)$  as shown in Table 8. So, given the predicted domains,  $\hat{P}(E = 1|C = G) = 1$ ,  $\hat{P}(E = 1|C = B) = 0$ , and  $\hat{P}(E = 1|C = R) = 0$ , the predictions on the class label are  $\hat{P}(Y = 1|C = G, E = 1) = 1$ ,  $\hat{P}(Y = 1|C = B, E = 2) = 1$ ,  $\rho > 0.8$ ,  $\hat{P}(Y = 1|C = B, E = 2) = 0$ ,  $\rho \leq 0.8$  and  $\hat{P}(Y = 1|C = R, E = 2) = 0$ . So, the test accuracy is 0.35 when  $\rho \leq 0.8$  and 0.2 when  $\rho > 0.8$ .

Similarly, when we make  $P(Y|E) = 0.5$  by oversampling the minority class in each domain, we can obtain the predictions shown in Table 9. So, the deterministic model would make predictions as  $\hat{P}(Y = 1|C = G, E = 1) = 1$ ,  $\hat{P}(Y = 1|C = B, E = 2) = 1$ , and  $\hat{P}(Y = 1|C = R, E) = 0$ . This would lead to a test accuracy of 0.2. These results reflect the reasons why the IRM model trained with the balanced classes in each domain ( $P(Y|E) = 0.5$ ) has worse performance compared to its counterpart trained with the original CMNIST+ data.

Table 4: Theoretical analysis results: fitting color

| $\rho$ | $P(Y = 1 C = G)$ | $P(Y = 1 C = B)$ | $P(Y = 1 C = R)$ |
|--------|------------------|------------------|------------------|
| 0.55   | 0.697            | 0.534            | 0.29             |
| 0.6    | 0.737            | 0.545            | 0.25             |
| 0.65   | 0.775            | 0.56             | 0.212            |
| 0.7    | 0.811            | 0.577            | 0.176            |
| 0.8    | 0.88             | 0.63             | 0.111            |
| 0.85   | 0.912            | 0.67             | 0.081            |
| 0.9    | 0.942            | 0.73             | 0.053            |

Table 5: Theoretical analysis results: fitting color, when  $P(Y|E) = 0.5$ .

| $\rho$ | $P(Y = 1 C = G)$ | $P(Y = 1 C = B)$ | $P(Y = 1 C = R)$ |
|--------|------------------|------------------|------------------|
| 0.55   | 0.633            | 0.633            | 0.29             |
| 0.6    | 0.667            | 0.667            | 0.25             |
| 0.65   | 0.702            | 0.702            | 0.212            |
| 0.7    | 0.739            | 0.739            | 0.176            |
| 0.8    | 0.818            | 0.818            | 0.111            |
| 0.85   | 0.86             | 0.86             | 0.081            |
| 0.9    | 0.905            | 0.905            | 0.053            |

Table 6: Theoretical analysis results: predicting domain by color.

| $\rho$ | $P(E = 1 C = G)$ | $P(E = 1 C = B)$ | $P(E = 1 C = R)$ |
|--------|------------------|------------------|------------------|
| 0.55   | 0.697            | 0.466            | 0.332            |
| 0.6    | 0.737            | 0.455            | 0.3              |
| 0.65   | 0.775            | 0.44             | 0.27             |
| 0.7    | 0.811            | 0.423            | 0.241            |
| 0.8    | 0.88             | 0.37             | 0.189            |
| 0.85   | 0.912            | 0.33             | 0.165            |
| 0.9    | 0.942            | 0.27             | 0.142            |

## B EXPERIMENTAL SETUP AND DATASETS

Here, we include more details on experimental setup and datasets.

Table 7: Theoretical analysis results: predicting domain by color, when  $P(Y|E) = 0.5$ .

| $\rho$ | $P(E = 1 C = G)$ | $P(E = 1 C = B)$ | $P(E = 1 C = R)$ |
|--------|------------------|------------------|------------------|
| 0.55   | 0.633            | 0.367            | 0.5              |
| 0.6    | 0.667            | 0.333            | 0.5              |
| 0.65   | 0.702            | 0.298            | 0.5              |
| 0.7    | 0.739            | 0.261            | 0.5              |
| 0.8    | 0.818            | 0.182            | 0.5              |
| 0.85   | 0.86             | 0.14             | 0.5              |
| 0.9    | 0.905            | 0.095            | 0.5              |

Table 8: Predicting by both color and domain

| $\rho$ | $P(Y = 1 G, 1)$ | $P(Y = 1 G, 2)$ | $P(Y = 1 B, 1)$ | $P(Y = 1 B, 2)$ | $P(Y = 1 R, 1)$ | $P(Y = 1 R, 2)$ |
|--------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| 0.55   | 0.957           | 0.1             | 0.9             | 0.214           | 0.786           | 0.043           |
| 0.6    | 0.964           | 0.1             | 0.9             | 0.25            | 0.75            | 0.036           |
| 0.65   | 0.971           | 0.1             | 0.9             | 0.292           | 0.708           | 0.029           |
| 0.7    | 0.977           | 0.1             | 0.9             | 0.341           | 0.659           | 0.023           |
| 0.8    | 0.986           | 0.1             | 0.9             | 0.471           | 0.529           | 0.014           |
| 0.85   | 0.99            | 0.1             | 0.9             | 0.557           | 0.443           | 0.01            |
| 0.9    | 0.994           | 0.1             | 0.9             | 0.667           | 0.333           | 0.006           |

Table 9: Predicting by both color and domain, class balanced

| $\rho$ | $P(Y = 1 G, 1)$ | $P(Y = 1 G, 2)$ | $P(Y = 1 B, 1)$ | $P(Y = 1 B, 2)$ | $P(Y = 1 R, 1)$ | $P(Y = 1 R, 2)$ |
|--------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| 0.55   | 0.71            | 0.5             | 0.5             | 0.71            | 0.29            | 0.29            |
| 0.6    | 0.75            | 0.5             | 0.5             | 0.75            | 0.25            | 0.25            |
| 0.65   | 0.788           | 0.5             | 0.5             | 0.788           | 0.212           | 0.212           |
| 0.7    | 0.824           | 0.5             | 0.5             | 0.824           | 0.176           | 0.176           |
| 0.8    | 0.889           | 0.5             | 0.5             | 0.889           | 0.111           | 0.111           |
| 0.85   | 0.919           | 0.5             | 0.5             | 0.919           | 0.081           | 0.081           |
| 0.9    | 0.947           | 0.5             | 0.5             | 0.947           | 0.053           | 0.053           |

## B.1 EXPERIMENTAL SETUP

Here, we provide more details on experimental setup. **Grid Search.** We perform grid search for hyperparameter tuning. For IRM, IRM-MMD and IRM-ACDM, we search the iteration number to plug in the IRM penalty term ( $K_{IRM}$ ) in 200, 400, 600 and IRM penalty weight  $\alpha$  in  $\{1, 10, \dots, 10^8\}$ . For MMD, ACDM, IRM-MMD, IRM-ACDM, we search CDM penalty weight  $\beta$  in  $\{1, 10, \dots, 10^5\}$ . For ACDM and IRM-ACDM, we set the number of steps we train the discriminator  $D$  in each iteration to 10.

## B.2 DATASETS

Here, we present more details about the datasets.

**CMNIST.** CMNIST is introduced by (Arjovsky et al., 2019). We can see the two training domains of CMNIST are similar to each other in terms of both  $P(C|Y, E)$  and  $P(Y|E)$  in Table 10. This means CMNIST does not cover the case of strong  $\Lambda$  spurious, since the spurious correlations, color–domain and domain–class, are not strong.

Table 10: Description of CMNIST

| $E$     | $P(Y = 1 E)$ | $Y$     | $P(C = G Y, E)$ | $P(C = B Y, E)$ | $P(C = R Y, E)$ |
|---------|--------------|---------|-----------------|-----------------|-----------------|
| $E = 1$ | 0.5          | $Y = 1$ | 0.9             | 0.0             | 0.1             |
|         |              | $Y = 0$ | 0.1             | 0.0             | 0.9             |
| $E = 2$ | 0.5          | $Y = 1$ | 0.8             | 0.0             | 0.2             |
|         |              | $Y = 0$ | 0.2             | 0.0             | 0.8             |
| $E = 3$ | 0.5          | $Y = 1$ | 0.1             | 0.0             | 0.9             |
|         |              | $Y = 0$ | 0.9             | 0.0             | 0.1             |

**CMNIST+.** We also visualize the CMNIST+ dataset with different values of  $\rho$  in Fig. 5.

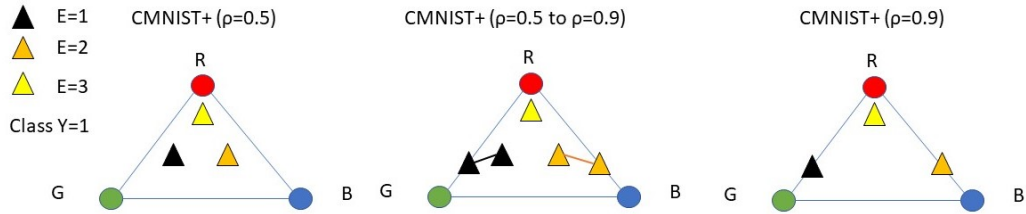


Figure 5: We visualize CMNIST+ with  $\rho \in [0.5, 0.9]$  in terms of  $P(C|Y=1, E)$ . Each large triangle represents the space of  $P(C|Y, E)$ . Each small triangle shows the values of  $P(C=c|Y=1, E=e)$ ,  $c \in \{R, G, B\}$ ,  $e \in \{1, 2, 3\}$ .