## A. New Experiments on the Adult Data Set

**Adult Income Prediction**. In this task we use the Adult data set[1] to predict if an individual's income exceeds \$50K/yr based on census data. We split the data set into four subgroups regarded as separated environments according to $race \in \{\text{Black}, \text{Non-Black}\}$ and $sex \in \{\text{Male}, \text{Female}\}$. We randomly choose two thirds of data from the subgroups Black Males and Non-Black Females for training, and then verify models across all four subgroups with the rest data. Six integral variables including *Age*, *FNLWGT*, *Eduction-Number*, *Capital-Gain*, *Capital-Loss*, and *Hours-Per-Week* are fed into ZIN and Minimax-TV-$\ell_1$ for environment inference. Ground-truth environment indicators are provided for groupDRO, IRM and IRM-TV-$\ell_1$.

Categorical variables except race and sex are encoded by one-hot coding, followed by a principal component analysis (PCA) transform, keeping an over 99% cumulative explained variance ratio. The transformed features are combined with the integral variables, yielding 59-dimensional features in total, which are subsequently normalized to have zero mean and unit variance for invariant feature learning. The prediction results are shown in Table 1.

The architectures of the invariant feature mapping $\Phi$ and the environment inference $\rho$ in Minimax-TV-$\ell_1$ instantiated in the Adult data set are similar to those in CelebA, shown in Table 2. All the compared methods except TIVA share the same $\Phi$, while TIVA adopts a linear model for $\Phi$. Minimax-TV-$\ell_1$ and ZIN share the same $\rho$. TIVA takes the normalized features for invariant feature learning as well as environment inference. The cross entropy is used as the loss function of prediction. The training process for the Adult data set is similar to that for CelebA.

Table 1: Mean accuracy (%) of competing methods on adult income prediction task with 10 repetitions.

| | | MEAN | | | STD | | |
|---|---|---|---|---|---|---|---|
| ENV PARTITION | METHODS | TRAIN | TEST | WORST | TRAIN | TEST | WORST |
| FALSE | ERM | 93.34 | 82.16 | 79.55 | 0.31 | 0.33 | 0.37 |
| | EIIL | 79.97 | 72.77 | 70.94 | 0.56 | 0.61 | 0.73 |
| | LFF | 82.03 | 75.04 | 73 | 5.54 | 3.01 | 2.45 |
| | TIVA | 91.45 | 81.95 | 79.28 | 0.12 | 0.39 | 0.46 |
| | ZIN | 93.16 | 82.26 | 79.67 | 0.17 | 0.27 | 0.29 |
| | **MINMAX-TV-$\ell_1$** | 92.40 | **83.33** | **80.95** | 0.11 | 0.14 | 0.16 |
| TRUE | GROUPDRO | 87.51 | 76.42 | 73.07 | 0.59 | 1.29 | 1.54 |
| | IRM | 93.19 | 82.32 | 79.76 | 0.28 | 0.23 | 0.29 |
| | **IRM-TV-$\ell_1$** | 92.42 | **83.31** | **80.93** | 0.19 | 0.18 | 0.19 |

Table 2: Pytorch-style architectures of the invariant feature mapping $\Phi$ and the environment inference $\rho$.

| DATA SET | $\Phi$ | $\rho$ |
|---|---|---|
| ADULT | LINEAR(59, 16)→RELU()→LINEAR(16, 1) | LINEAR(6, 16)→RELU()→LINEAR(16, 4)→SOFMAX() |

[1] https://archive.ics.uci.edu/dataset/2/adult