A Appendix – Framework and Algorithm

A.1 Lambda Optimization for Demographic Parity

Continuing from Sec. 3.3, we state FairBatch's optimization for demographic parity disparity:

$$\min_{\boldsymbol{\lambda}} \max\{|\frac{|S_{\boldsymbol{\lambda}(y,z')}|}{|S_{\boldsymbol{\lambda}(z')}|} L_{(y,z')} - \frac{|S_{\boldsymbol{\lambda}(y,z)}|}{|S_{\boldsymbol{\lambda}(z)}|} L_{(y,z)}|\},\ z \neq z',\ z, z' \in \mathbb{Z}, y \in \mathbb{Z} \tag{9}$$

where $L_{(y,z)} = 1/|S_{\boldsymbol{\lambda}(y,z)}| \sum_{S_{\boldsymbol{\lambda}(y,z)}} \ell_\theta(s_i)$, and $S_{\boldsymbol{\lambda}(y,z)}$ is a subset of $S_{\boldsymbol{\lambda}}$ from Eq. 4 for the (y $=$ y, z $=$ z) set. More details are described in Roh et al. [2021].

A.2 Fairness Constraints in the Multidimensional Knapsack Problem

Continuing from Sec. 4, we describe how we rearrange the fairness constraints so that the right-hand side expressions of Eq. 2 become constants instead of containing the variable S_y. We first express Eq. 2 as a summation using the indicator function $\mathbf{1}_D(\cdot)$ and then move the right-hand side expression to the left-hand side:

$$\begin{aligned}
&\sum_{j \in \mathbb{I}_{(y,z)}} p_j \leq \lambda_{(y,z)} |S_y| \\
\iff\ &\sum_{i=1}^{n} \mathbf{1}_{D_{(y,z)}}(d_i)\, p_i \leq \lambda_{(y,z)} |S_y| \\
&\qquad\qquad\qquad\qquad = \lambda_{(y,z)} \sum_{i=1}^{n} \mathbf{1}_{D_y}(d_i)\, p_i \\
\iff\ &\sum_{i=1}^{n} \mathbf{1}_{D_{(y,z)}}(d_i)\, p_i - \lambda_{(y,z)} \sum_{i=1}^{n} \mathbf{1}_{D_y}(d_i)\, p_i \leq 0
\end{aligned}$$

where p_i indicates whether the data sample d_i is selected or not, $\mathbb{I}_{(y,z)}$ is an index set of the (y, z) class, $D_{(y,z)}$ is a subset for the (y, z) class, $\mathbf{1}_{D_{(y,z)}}(d_i)$ is an indicator function that returns 1 if $d_i \in D_{(y,z)}$ and 0 otherwise, and S_y is the selected samples for y $= y$.

By considering each case formulated by the indicator functions, we can rewrite the inequality with example weights v_i:

$$\sum_{i=1}^{n} v_i p_i \leq 0, \text{where } v_i = \begin{cases} 0 & \text{if } d_i \notin D_y \\ -\lambda_{(y,z)} & \text{if } d_i \in D_y \text{ and } d_i \notin D_z \\ 1 - \lambda_{(y,z)} & \text{if } d_i \in D_{(y,z)} \end{cases}.$$

Finally, we add 1 for the above weights v_i to make the new weights w_i that are always positive:

$$\sum_{i=1}^{n} w_i p_i \leq \tau n, \text{where } w_i = \begin{cases} 1 & \text{if } d_i \notin D_y \\ 1 - \lambda_{(y,z)} & \text{if } d_i \in D_y \text{ and } d_i \notin D_z \\ 2 - \lambda_{(y,z)} & \text{if } d_i \in D_{(y,z)} \end{cases}.$$

A.3 Convergence of the Algorithm

Continuing from Sec. 4, we discuss the convergence of our algorithm. Currently, our algorithm does not have theoretical guarantees for convergence. However, both ITLM and FairBatch do have convergence guarantees under some assumptions. Hence, we suspect that our algorithm will converge under reasonable circumstances as well. Indeed in our experiments, we did not run into convergence issues so far. In more general applications, averaging the model predictions over the last few epochs can be a reasonable choice.

B Appendix – Experiments

B.1 Other Experimental Settings

Continuing from Sec. 5, we provide more details of the experimental settings. The batch sizes of the synthetic, COMPAS, and AdultCensus datasets are 100, 200, and 2000, respectively. For the

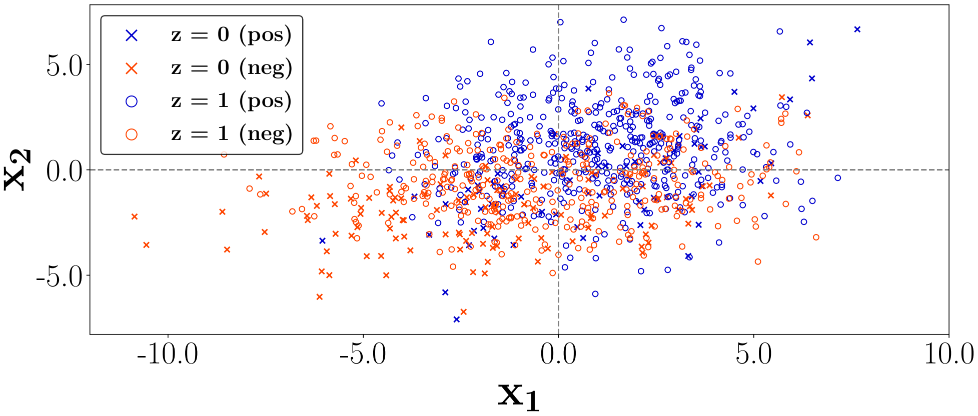


Figure 4: The synthetic dataset.

synthetic dataset, we split the data into 2000, 1000, and 200 samples for the training, test, and validation sets, respectively. For the real datasets, we use 20% of the entire data as a test set and split the remaining data into 10:1 for the training and validation sets. Note that the validation set is only used in FR-Train. We choose the learning rate from the candidate set {0.0001, 0.0005} using cross-validation. For ITLM-related algorithms (i.e., ITLM, ITLM→FB, ITLM→Penalty, and Ours), we utilize warm-starting in training, where we train the first 100 epochs without fair or robust training.

B.2 Synthetic Data

Continuing from Sec. 5, we visualize the synthetic dataset. Figure 4 shows the synthetic dataset we utilized in Sec. 5. As we explained in the experimental setting, the synthetic dataset has 3,200 samples and consists of two non-sensitive features $(\mathrm{x}_1, \mathrm{x}_2)$, one binary sensitive feature z, and one binary label class y.

B.3 Other Fair and Robust Baselines

Continuing from Sec. 5, we compare two more fair and robust baselines that are variants of ITLM→FB and ITLM→Penalty. First, (ITLM+Penalty)→FB is similar to ITLM->FB except that we improve ITLM using fairness penalty terms. In particular, we add the following covariance term to the optimization: $\min_{S:|S|=\lfloor \tau N \rfloor} \sum_{s_i \in S} [l_\theta(s_i) + \mu \; |Cov(z_i, \hat{y_i})|]$ where z_i and $\hat{y_i}$ are the sensitive group and the predicted label of the sample s_i, respectively. After selecting samples via the optimization in ITLM+Penalty, we run FairBatch on the selected data. The second method (ITLM+Penalty)->Penalty is identical except that it runs Penalty as its second step instead of FairBatch.

Table 6 shows the performances of the algorithms when using label flipping and group-targeted label flipping. For both types of label flipping, the new fair and robust baselines (i.e., (ITLM+Penalty)→FB and (ITLM+Penalty)→Penalty) perform better than LR, but worse than our algorithm in terms of accuracy and fairness. These results are similar to those of ITLM->FB and ITLM->Penalty.

B.4 A Trade-off Curve Comparison of Our Algorithm and FR-Train

Continuing from Sec. 5.1, we draw accuracy-fairness disparity trade-off curves of our algorithm and FR-Train. Figure 5 shows results using the synthetic dataset w.r.t. equalized odds (EO) disparity where the experimental settings are identical to Table 1. The trends are consistent with the other results in Table 1, where our algorithm usually shows better fairness (i.e., lower disparity) than FR-Train when the accuracy is similar. Another observation is that FR-Train's trade-off curve is noisy due to its adversarial training.

Table 6: Performances on the *synthetic* test set w.r.t. equalized odds disparity (EO Disp.) and demographic parity disparity (DP Disp.). We compare our algorithm with LR and the two-step fair and robust baselines: ITLM→FB [Roh et al., 2021], ITLM→Penalty [Zafar et al., 2017a,b], (ITLM+Penalty)→FB [Roh et al., 2021], and (ITLM+Penalty)→Penalty [Zafar et al., 2017a,b]. We flip 10% of labels in the training data. Experiments are repeated 5 times.

	Label Flipping				Group-Targeted Label Flipping			
Method	Acc.	EO Disp.	Acc.	DP Disp.	Acc.	EO Disp.	Acc.	DP Disp.
LR	.665±.003	.557±.015	.665±.003	.400±.010	.600±.002	.405±.008	.600±.002	.300±.006
ITLM→FB	.718±.003	.199±.020	.725±.002	.089±.032	.707±.001	.108±.030	.704±.003	.067±.027
ITLM→Penalty	.651±.051	.172±.046	.674±.012	.068±.014	.706±.001	.080±.004	.688±.004	.044±.004
(ITLM+Penalty)→FB	.668±.015	.183±.036	.714±.002	.045±.021	.702±.004	.125±.035	.688±.008	.049±.024
(ITLM+Penalty)→Penalty	.685±.030	.213±.030	.694±.017	.069±.012	.700±.013	.182±.062	.695±.003	.058±.002
Ours	.727±.005	.064±.005	.720±.001	.006±.001	.726±.001	.040±.002	.720±.001	.039±.007

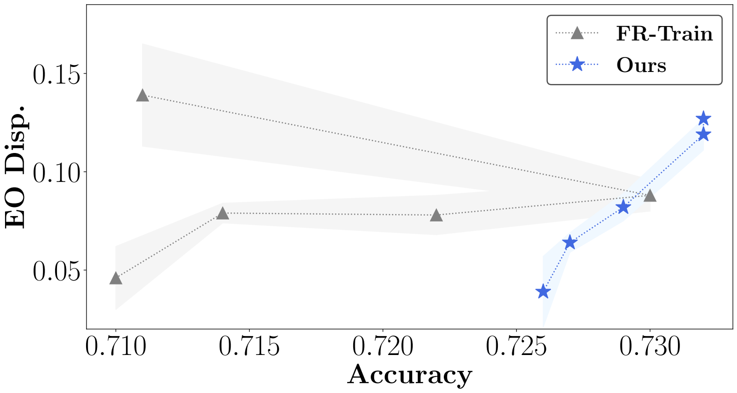


Figure 5: Accuracy-EO disp. trade-off curves of our algorithm and FR-Train on the synthetic dataset.

B.5 Accuracy and Fairness – AdultCensus

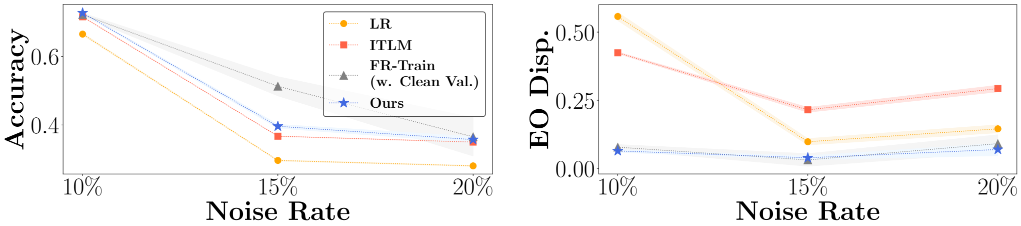
Continuing from Sec. 5.1, we compare our algorithm with other methods on the AdultCensus dataset. The overall results are similar to those for the COMPAS dataset (Table 2), where our algorithm shows the best or second-best accuracy and fairness performances among the fair and robust training algorithms. Compared to FR-Train, our algorithm has similar accuracy and better fairness.

B.6 Varying the Noise Rate – Equalized Odds

Continuing from Sec. 5.3, we observe the accuracy and fairness w.r.t. equalized odds of the algorithms when varying the noise rate (i.e., flipping different amounts of labels) in the training data. Figure 6 shows the performances of logistic regression (LR), ITLM, FR-Train, and our algorithm for different noise rates. Similar to the demographic parity disparity results (Figure 3), our algorithm outperforms LR and ITLM while having worse accuracy, but better fairness compared to FR-Train.

Table 7: Performances on the *AdultCensus* test set w.r.t. equalized odds disparity (EO Disp.) and demographic parity disparity (DP Disp.). We compare our algorithm with four types of baselines: (1) vanilla training: LR; (2) robust training: ITLM [Shen and Sanghavi, 2019]; (3) fair training: FB [Roh et al., 2021]; and (4) fair and robust training: ITLM→FB, ITLM→Penalty [Zafar et al., 2017a,b], and FR-Train [Roh et al., 2020]. We flip 10% of labels in the training data. Experiments are repeated 5 times. We highlight the best and second-best performances among the fair and robust algorithms.

	Label Flipping				Group-Targeted Label Flipping			
Method	Acc.	EO Disp.	Acc.	DP Disp.	Acc.	EO Disp.	Acc.	DP Disp.
LR	.746±.003	.070±.002	.746±.003	.073±.001	.748±.006	.095±.002	.748±.006	.127±.006
ITLM	.785±.018	.087±.031	.785±.018	.092±.016	.785±.011	.144±.023	.785±.011	.105±.006
FB	.748±.002	.022±.002	.758±.004	.046±.007	.739±.014	.086±.037	.693±.002	.015±.005
ITLM→FB	.772±.024	**.047±.008**	.776±.023	.073±.010	**.773±.013**	**.047±.006**	**.769±.014**	.053±.005
ITLM→Penalty	**.776±.023**	.082±.015	.774±.024	**.054±.018**	.755±.026	.161±.018	.757±.003	**.013±.003**
Ours	.771±.015	**.029±.005**	**.782±.015**	**.049±.012**	.761±.006	**.047±.018**	.760±.007	**.034±.016**
FR-Train	**.779±.007**	.061±.009	**.782±.007**	.089±.005	**.782±.008**	.075±.018	**.773±.012**	.049±.010



(a) Accuracy.

(b) Equalized odds disparity.

Figure 6: Performances of LR, ITLM, FR-Train, and our algorithm (Ours) on the synthetic data while varying the noise rate using label flipping [Paudice et al., 2018].

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] We believe our abstract and introduction (Section 1) are accurate.
   (b) Did you describe the limitations of your work? [Yes] We described the limitation of our work in the last part of Section 7.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] We discussed the potential societal impacts of our work in the last part of Section 7.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] We carefully read the guidelines.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] We state the conditions for the theoretical results in Sections 3 and 4 (e.g., the range of the variables in the optimization).
   (b) Did you include complete proofs of all theoretical results? [Yes] We include all details for theoretical results in Section 3, Section 4, and Section A.2 (e.g., observing that the proposed optimization is strongly NP-hard).
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We include the code, data, and instructions to reproduce the main results in the supplementary.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We specify the training details in Section 5 and Section B.1.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] All experiments are repeated with 5 different random seeds. We report the error ranges in Section 5.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We specify this information in the experimental settings (Section 5).
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We cite the references of the public datasets COMPAS and AdultCensus in Section 5.
   (b) Did you mention the license of the assets? [N/A] The public datasets we used are not licensed.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We include our code in the supplementary.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] We use publicly-shared benchmark datasets and cite the references in Section 5.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] We explain in Section 5 that we do not use any direct personal identifier (e.g., name and date of birth).
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]