

399 A Proof of Lemma 1

400 *Proof.* By Eq.(1) we have

$$\hat{\text{TPR}}_a = (1 - \eta_{1a})\text{TPR}_a + \eta_{1a}\text{TPR}_{a'},$$

401

$$\hat{\text{TNR}}_a = (1 - \eta_{0a})\text{TNR}_a + \eta_{0a}\text{TNR}_{a'}.$$

402 And we have EÔd as

$$\text{EÔd} = |\hat{\text{TPR}}_0 - \hat{\text{TPR}}_1| + |\hat{\text{TNR}}_0 - \hat{\text{TNR}}_1| = (1 - \eta_{10} - \eta_{11})|\text{TPR}_0 - \text{TPR}_1| + (1 - \eta_{00} - \eta_{01})|\text{TNR}_0 - \text{TNR}_1|.$$

403 For $\hat{\text{DI}}$, we have the positive prediction rate under noisy data as follows:

$$\hat{\text{PR}}_a = \hat{\alpha}_a \hat{\text{TPR}}_a + (1 - \hat{\alpha}_a) \hat{\text{FPR}}_a,$$

404 where $\hat{\alpha}_a = \frac{|\{i|\hat{a}_i=a, y_i=1\}|}{|\{i|\hat{a}_i=a\}|}$ is the noisy base rate of group a . Correspondingly, we have $\hat{\text{DI}}$ as

$$\begin{aligned} \hat{\text{DI}} &= |\hat{\text{PR}}_0 - \hat{\text{PR}}_1| \\ &= |(1 - \hat{\alpha}_0)[(1 - \eta_{00})\text{FPR}_0 + \eta_{00}\text{FPR}_1] - (1 - \hat{\alpha}_1)[(1 - \eta_{01})\text{FPR}_1 + \eta_{01}\text{FPR}_0] \\ &\quad + \hat{\alpha}_0[(1 - \eta_{10})\text{TPR}_0 + \eta_{10}\text{TPR}_1] - \hat{\alpha}_1[(1 - \eta_{11})\text{TPR}_1 + \eta_{11}\text{TPR}_0]| \\ &= |[1 - (\hat{\alpha}_0 + \eta_{00}) + \hat{\alpha}_0\eta_{00} - \eta_{01} + \hat{\alpha}_1\eta_{01}]\text{FPR}_0 - [1 - (\hat{\alpha}_1 + \eta_{01}) + \hat{\alpha}_1\eta_{01} - \eta_{00} + \hat{\alpha}_0\eta_{00}]\text{FPR}_1 \\ &\quad + (\hat{\alpha}_0 - \hat{\alpha}_0\eta_{10} - \hat{\alpha}_1\eta_{11})\text{TPR}_0 - (\hat{\alpha}_1 - \hat{\alpha}_0\eta_{10} - \hat{\alpha}_1\eta_{11})\text{TPR}_1|. \end{aligned}$$

405

□

406 **Remark 1.** *There is no deterministic relationship between $\hat{\text{DI}}$ and DI under group- and label-*
407 *dependent flip. Specially, under group-dependent flip, i.e., $\eta_{ya} = \eta_{y'a} = \eta_a$, we have*

$$\begin{aligned} \hat{\text{DI}} &= |[1 - (\hat{\alpha}_0 + \eta_0) + \hat{\alpha}_0\eta_0 - \eta_1 + \hat{\alpha}_1\eta_1]\text{FPR}_0 - [1 - (\hat{\alpha}_1 + \eta_1) + \hat{\alpha}_1\eta_1 - \eta_0 + \hat{\alpha}_0\eta_0]\text{FPR}_1 \\ &\quad + (\hat{\alpha}_0 - \hat{\alpha}_0\eta_0 - \hat{\alpha}_1\eta_1)\text{TPR}_0 - (\hat{\alpha}_1 - \hat{\alpha}_0\eta_0 - \hat{\alpha}_1\eta_1)\text{TPR}_1| \\ &= (1 - \eta_0 - \eta_1)\text{DI}, \end{aligned}$$

408 where the last equality is obtained by substituting $\hat{\alpha}_a$ with $\hat{\alpha}_a = (1 - \eta_a)\alpha_a + \eta_a\alpha_{a'}$.

409 B Proof of Theorem 1

410 *Proof.* Since $0 \leq (1 - \eta_{ya} - \eta_{ya'}) \leq 1$, we have from Lemma 1 that $\text{EÔd} \leq |\text{TPR}_0 - \text{TPR}_1| +$
411 $|\text{TNR}_0 - \text{TNR}_1| = \text{EOd}$. For the upper-bound, we have the following relationship regarding Q_{ya}
412 and \hat{Q}_{ya} :

$$Q_{ya} = \frac{1 - \eta_{ya'}}{1 - \eta_{ya} - \eta_{ya'}} \hat{Q}_{ya} + \frac{\eta_{ya}}{1 - \eta_{ya} - \eta_{ya'}} \hat{Q}_{ya'},$$

413 where $a' = |1 - a|$. Therefore we have the total variation distance between Q_{ya} and \hat{Q}_{ya} as follows:

$$\begin{aligned} \text{TV}(Q_{ya}, \hat{Q}_{ya}) &= \frac{1}{2} \int |Q_{ya} - \hat{Q}_{ya}| \\ &= \frac{1}{2} \int \left| \frac{1 - \eta_{ya'}}{1 - \eta_{ya} - \eta_{ya'}} \hat{Q}_{ya} + \frac{\eta_{ya}}{1 - \eta_{ya} - \eta_{ya'}} \hat{Q}_{ya'} - \hat{Q}_{ya} \right| \\ &= \frac{\eta_{ya}}{1 - \eta_{ya} - \eta_{ya'}} \text{TV}(\hat{Q}_{ya}, \hat{Q}_{ya'}). \end{aligned}$$

414 And we have the EOd under clean distribution as follows:

E_{Od}

$$\begin{aligned}
&= |\text{TPR}_0 - \text{TPR}_1| + |\text{TNR}_0 - \text{TNR}_1| \\
&= \left| \int_{\frac{1}{2}}^1 Q_{10} - Q_{11} \right| + \left| \int_0^{\frac{1}{2}} Q_{00} - Q_{01} \right| \\
&= \left| \int_{\frac{1}{2}}^1 Q_{10} - \hat{Q}_{10} + \hat{Q}_{10} - Q_{11} + \hat{Q}_{11} - \hat{Q}_{11} \right| + \left| \int_0^{\frac{1}{2}} Q_{00} - \hat{Q}_{00} + \hat{Q}_{00} - Q_{01} - \hat{Q}_{01} + \hat{Q}_{01} \right| \\
&\leq \left| \int_{\frac{1}{2}}^1 \hat{Q}_{10} - \hat{Q}_{11} \right| + \left| \int_{\frac{1}{2}}^1 Q_{10} - \hat{Q}_{10} \right| + \left| \int_{\frac{1}{2}}^1 \hat{Q}_{11} - Q_{11} \right| + \left| \int_0^{\frac{1}{2}} \hat{Q}_{01} - Q_{01} \right| + \left| \int_0^{\frac{1}{2}} \hat{Q}_{00} - \hat{Q}_{01} \right| + \left| \int_0^{\frac{1}{2}} Q_{00} - \hat{Q}_{00} \right| \\
&\leq \text{E}\hat{\text{O}}\text{d} + \frac{\eta_{10} + \eta_{11}}{1 - \eta_{10} - \eta_{11}} 2TV(\hat{Q}_{10}, \hat{Q}_{11}) + \frac{\eta_{00} + \eta_{01}}{1 - \eta_{00} - \eta_{01}} 2TV(\hat{Q}_{00}, \hat{Q}_{01}) \\
&\leq \text{E}\hat{\text{O}}\text{d} + \frac{\eta_{10} + \eta_{11}}{1 - \eta_{10} - \eta_{11}} \sqrt{D_{KL}(\hat{Q}_{10}, \hat{Q}_{11})} + \frac{\eta_{00} + \eta_{01}}{1 - \eta_{00} - \eta_{01}} \sqrt{D_{KL}(\hat{Q}_{00}, \hat{Q}_{01})},
\end{aligned}$$

415 where the last inequality is due to $TV(P, Q) \leq \sqrt{\frac{1}{2} D_{KL}(P \| Q)}$. Since $D_{KL}(\hat{P}_{ya}, \hat{P}_{ya'}) \leq \epsilon_y$, we
416 have the following upper- and lower-bound regarding E_{Od} under clean distribution and E_{Od} under
417 noisy distribution:

$$\text{E}\hat{\text{O}}\text{d} \leq \text{E}\hat{\text{O}}\text{d} + \frac{\eta_{00} + \eta_{01}}{1 - \eta_{00} - \eta_{01}} \sqrt{\epsilon_0} + \frac{\eta_{10} + \eta_{11}}{1 - \eta_{10} - \eta_{11}} \sqrt{\epsilon_1}.$$

418

□

419 C Proof of Lemma 2

420 *Proof.* By Eq. (8) we have

$$\begin{aligned}
\widetilde{\text{TPR}}_a &= (1 - \eta_{1a})\text{TPR}_a + \eta_{1a}\text{FPR}_a, \\
\widetilde{\text{TNR}}_a &= (1 - \eta_{0a})\text{TNR}_a + \eta_{0a}\text{FNR}_a.
\end{aligned}$$

422 And we have the following relationship regarding $\widetilde{\text{E}}\text{Od}$:

$$\begin{aligned}
\widetilde{\text{E}}\text{Od} &= |(1 - \beta_{10})\text{TPR}_0 + \beta_{10}\text{FPR}_0 - (1 - \beta_{11})\text{TPR}_1 + \beta_{11}\text{FPR}_1| \\
&\quad + |(1 - \beta_{00})\text{TNR}_0 + \beta_{00}\text{FNR}_0 - (1 - \beta_{01})\text{TNR}_1 + \beta_{01}\text{FNR}_1| \\
&= |\text{TPR}_0 - \text{TPR}_1 + \beta_{10}(\text{FPR}_0 - \text{TPR}_0) - \beta_{11}(\text{FPR}_1 - \text{TPR}_1)| \\
&\quad + |\text{TNR}_0 - \text{TNR}_1 + \beta_{00}(\text{FNR}_0 - \text{TNR}_0) - \beta_{01}(\text{FNR}_1 - \text{TNR}_1)|
\end{aligned}$$

423 For $\widetilde{\text{DI}}$, we have

$$\begin{aligned}
\widetilde{\text{DI}} &= \left| \frac{(1 - \beta_{10})\text{TP}_0 + \beta_{00}\text{FP}_0 + (1 - \beta_{00})\text{FP}_0 + \beta_{10}\text{TP}_0}{\text{TP}_0 + \text{TN}_0 + \text{FP}_0 + \text{FN}_0} - \frac{(1 - \beta_{11})\text{TP}_1 + \beta_{01}\text{FP}_1 + (1 - \beta_{01})\text{FP}_1 + \beta_{11}\text{TP}_1}{\text{TP}_1 + \text{TN}_1 + \text{FP}_1 + \text{FN}_1} \right| \\
&= \left| \frac{\text{TP}_0 + \text{FP}_0}{\text{TP}_0 + \text{TN}_0 + \text{FP}_0 + \text{FN}_0} - \frac{\text{TP}_1 + \text{FP}_1}{\text{TP}_1 + \text{TN}_1 + \text{FP}_1 + \text{FN}_1} \right| = \text{DI}.
\end{aligned}$$

424

□

425 D Proof of Lemma 3

426 *Proof.* Let Q_{ya} be the distribution of predicted soft labels in the clean subgroup $\{i | y_i = y, a_i = a\}$
427 and \tilde{Q}_{ya} be the corresponding distribution in the noisy group, let β_{ya} be the label noise rate of
428 subgroup $\{i | \tilde{y}_i = y, a_i = a\}$, we have

$$Q_{ya} = \frac{1 - \beta_{y'a}}{(1 - \beta_{ya} - \beta_{y'a})} \tilde{Q}_{ya} - \frac{\beta_{ya}}{(1 - \beta_{ya} - \beta_{y'a})} \tilde{Q}_{y'a}. \quad (9)$$

429 And we have EO_p under clean data as follows:

$$\begin{aligned}
\text{EO}_p &= \left| \int_{\frac{1}{2}}^1 Q_{10} - \int_{\frac{1}{2}}^1 Q_{11} \right| \\
&= \left| \int_{\frac{1}{2}}^1 \frac{(1-\beta_{00})}{(1-\beta_{10}-\beta_{00})} \tilde{Q}_{10} - \frac{(1-\beta_{01})}{(1-\beta_{11}-\beta_{01})} \tilde{Q}_{11} + \frac{\beta_{11}}{(1-\beta_{11}-\beta_{01})} \tilde{Q}_{01} - \frac{\beta_{10}}{(1-\beta_{00}-\beta_{10})} \tilde{Q}_{00} \right| \\
&\leq \left| \frac{(1-\beta_{00})}{(1-\beta_{10}-\beta_{00})} \widetilde{\text{TPR}}_0 - \frac{(1-\beta_{01})}{(1-\beta_{11}-\beta_{01})} \widetilde{\text{TPR}}_1 \right| + \left| \frac{\beta_{11}}{(1-\beta_{11}-\beta_{01})} \widetilde{\text{FPR}}_1 - \frac{\beta_{10}}{(1-\beta_{00}-\beta_{10})} \widetilde{\text{FPR}}_0 \right| \\
&\leq \min \left\{ \left| \frac{(1-\beta_{00})}{(1-\beta_{10}-\beta_{00})}, \frac{(1-\beta_{01})}{(1-\beta_{11}-\beta_{01})} \right| \right\} \widetilde{\text{EO}}_p + \min \left\{ \left| \frac{\beta_{11}}{(1-\beta_{11}-\beta_{01})}, \frac{\beta_{10}}{(1-\beta_{00}-\beta_{10})} \right| \right\} \widetilde{\text{DFPR}} \\
&\quad + \left| \frac{(1-\beta_{00})}{(1-\beta_{10}-\beta_{00})} - \frac{(1-\beta_{01})}{(1-\beta_{11}-\beta_{01})} \right| (\widetilde{\text{TPR}}_a + \widetilde{\text{FPR}}_{\bar{a}}),
\end{aligned} \tag{10}$$

430 where $a = 0$ if $\frac{(1-\beta_{00})}{(1-\beta_{10}-\beta_{00})} \geq \frac{(1-\beta_{01})}{(1-\beta_{11}-\beta_{01})}$ and $a = 1$ otherwise, and $\bar{a} = |1-a|$. And we have

431 DTNR as follows:

$$\begin{aligned}
\text{DTNR} &= \left| \int_0^{\frac{1}{2}} p_{00} - \int_0^{\frac{1}{2}} p_{01} \right| \\
&= \left| \int_0^{\frac{1}{2}} \frac{(1-\beta_{10})}{(1-\beta_{10}-\beta_{00})} \tilde{P}_{00} - \frac{(1-\beta_{11})}{(1-\beta_{11}-\beta_{01})} \tilde{P}_{01} + \frac{\beta_{01}}{(1-\beta_{11}-\beta_{01})} \tilde{P}_{11} - \frac{\beta_{00}}{(1-\beta_{00}-\beta_{10})} \tilde{P}_{10} \right| \\
&\leq \min \left\{ \left| \frac{(1-\beta_{10})}{(1-\beta_{10}-\beta_{00})}, \frac{(1-\beta_{11})}{(1-\beta_{11}-\beta_{01})} \right| \right\} \widetilde{\text{DTNR}} + \min \left\{ \left| \frac{\beta_{01}}{(1-\beta_{11}-\beta_{01})}, \frac{\beta_{00}}{(1-\beta_{00}-\beta_{10})} \right| \right\} \widetilde{\text{DFNR}} \\
&\quad + \left| \frac{(\beta_{00})}{(1-\beta_{10}-\beta_{00})} - \frac{(\beta_{01})}{(1-\beta_{11}-\beta_{01})} \right| (\widetilde{\text{TNR}}_{a'} + \widetilde{\text{FNR}}_{\bar{a}'}).
\end{aligned} \tag{11}$$

432 where $a' = 0$ if $\frac{(1-\beta_{10})}{(1-\beta_{10}-\beta_{00})} \geq \frac{(1-\beta_{11})}{(1-\beta_{11}-\beta_{01})}$ and 0 otherwise, and $\bar{a}' = |1-a'|$. Therefore we have

433 clean EO_d as follows:

$$\begin{aligned}
\text{EO}_d &\leq \min \left\{ \frac{1}{1-\beta_{00}-\beta_{10}}, \frac{1}{1-\beta_{01}-\beta_{11}} \right\} \widetilde{\text{EO}}_d + \left| \frac{(1-\beta_{00})}{(1-\beta_{10}-\beta_{00})} - \frac{(1-\beta_{01})}{(1-\beta_{11}-\beta_{01})} \right| (\widetilde{\text{TPR}}_a + \widetilde{\text{FPR}}_{\bar{a}}) \\
&\quad + \left| \frac{(\beta_{00})}{(1-\beta_{10}-\beta_{00})} - \frac{(\beta_{01})}{(1-\beta_{11}-\beta_{01})} \right| (\widetilde{\text{TNR}}_{a'} + \widetilde{\text{FNR}}_{\bar{a}'}) \\
&\leq \min \left\{ \frac{1}{1-\beta_{00}-\beta_{10}}, \frac{1}{1-\beta_{01}-\beta_{11}} \right\} \widetilde{\text{EO}}_d \\
&\quad + (2 + \widetilde{\text{EO}}_d) \max \left\{ \left| \frac{\beta_{00}}{1-\beta_{00}-\beta_{10}} - \frac{\beta_{01}}{1-\beta_{01}-\beta_{11}} \right|, \left| \frac{1-\beta_{00}}{1-\beta_{10}-\beta_{00}} - \frac{1-\beta_{01}}{1-\beta_{11}-\beta_{01}} \right| \right\} \\
&= \min \left\{ \frac{1}{1-\beta_{00}-\beta_{10}} + \beta, \frac{1}{1-\beta_{01}-\beta_{11}} + \beta \right\} \widetilde{\text{EO}}_d + 2\beta.
\end{aligned} \tag{12}$$

434 \square

435 E Results under different sensitive attributes

436 We also include results under different sensitive attributes. Specifically, we choose race as sensitive
437 information for Adult dataset and sex as sensitive information for COMPAS dataset. Results are
438 shown in Table 7-8. As shown in the tables, our method performs better or comparably than other
439 methods under varied sensitive attributes, which validates the effectiveness of our method.

Method	Accuracy	Disparate Impact	EOD
Baseline	66.80 \pm 0.34%	19.64 \pm 1.79%	21.24 \pm 2.17%
Inprocessing (Wang et al., 2022)	63.14 \pm 0.49%	12.25 \pm 1.14%	14.35 \pm 1.22%
DLR (Celis et al., 2021)	63.34 \pm 0.54%	9.47 \pm 1.24%	8.58 \pm 1.63%
FairExpec (Mehrotra and Celis, 2021)	63.26 \pm 1.45%	9.83 \pm 1.55%	10.17 \pm 1.29%
CorScale (Lamy et al., 2019)	62.37 \pm 0.77%	11.23 \pm 1.25%	11.19 \pm 1.75%
Ours	63.45 \pm 0.42%	7.57 \pm 1.22%	6.54 \pm 1.38%

Table 7: Experimental results on COMPAS dataset under sensitive attribute noise with *sex* as sensitive attribute. The noise rates are set as $\eta_{00} = 0.2$, $\eta_{01} = 0.1$, $\eta_{10} = 0.3$, $\eta_{11} = 0.2$.

Method	Accuracy	Disparate Impact	EOD
Baseline	84.16 \pm 0.45%	13.31 \pm 1.47%	15.13 \pm 1.24%
Inprocessing (Wang et al., 2022)	82.54 \pm 0.73%	10.26 \pm 1.58%	11.27 \pm 1.64%
DLR (Celis et al., 2021)	81.57 \pm 0.63%	7.42 \pm 1.36%	7.47 \pm 1.18%
FairExpec (Mehrotra and Celis, 2021)	82.35 \pm 0.49%	8.63 \pm 1.21%	8.15 \pm 1.46%
CorScale (Lamy et al., 2019)	80.84 \pm 0.68%	8.45 \pm 1.16%	9.23 \pm 1.58%
Ours	82.14 \pm 0.71%	6.56 \pm 1.14%	5.84 \pm 1.27%

Table 8: Experimental results on Adult dataset under sensitive attribute noise with *race* as sensitive attribute. The noise rates are set as $\eta_{00} = 0.15$, $\eta_{01} = 0.1$, $\eta_{10} = 0.1$, $\eta_{11} = 0.3$.