A. Proof of Theorem 2

We first show that the accuracy $A_F(\tau)$ of a binary selective classification task is an increasing function if the confidence κ is constructed from a calibrated score function $R = s(x)$. And the monotonicity of the precision $PPV_F(\tau)$ can be proved similarly.

The following lemma from (Jones et al., 2020) characterizes the condition for monotonicity of selective accuracy.

Lemma 4. *$A_F(\tau)$ is monotone increasing in τ if and only if*

$$\frac{f_M(\tau)}{f_M(-\tau)} \leq \frac{1 - F_M(\tau)}{F_M(-\tau)}, \tag{21}$$

for all $\tau \geq 0$.

Our proof also relies on the following lemma.

Lemma 5. *Suppose the score function R is calibrated by group, and $\hat{Y} = \arg\max_{y\in\{0,1\}} P(Y = y|R = r)$. Denote the maximum a posteriori probability $S = \max\{R, 1 - R\}$, then*

$$P(Y = \hat{Y}|S = s, D = d) = s, \tag{22}$$

for all $d \in \mathcal{D}$.

Proof. Since R is calibrated by group, then $\forall a, b \in \mathcal{D}$,

$$P(Y = 1|R = r, D = a) = P(Y = 1|R = r, D = b) = r,$$

where $r \in [0, 1]$. Thus, for any $s \in [0.5, 1]$ and $d \in \mathcal{D}$, we have

$$\begin{aligned}
&P(Y = \hat{Y}|S = s, D = d)\\
&= P(Y = \hat{Y}|R \in \{s, 1 - s\}, D = d)\\
&= P(Y = 1, \hat{Y} = 1|R \in \{s, 1 - s\}, D = d)\\
&\quad + P(Y = 0, \hat{Y} = 0|R \in \{s, 1 - s\}, D = d)\\
&= \frac{P(Y = 1, \hat{Y} = 1, R \in \{s, 1 - s\}, D = d)}{P(R \in \{s, 1 - s\}, D = d)}\\
&\quad + \frac{P(Y = 0, \hat{Y} = 0, R \in \{s, 1 - s\}, D = d)}{P(R \in \{s, 1 - s\}, D = d)}\\
&\overset{(a)}{=} \frac{P(Y = 1, \hat{Y} = 1, R = s, D = d)}{P(R \in \{s, 1 - s\}, D = d)}\\
&\quad + \frac{P(Y = 0, \hat{Y} = 0, R = 1 - s, D = d)}{P(R \in \{s, 1 - s\}, D = d)}\\
&= \frac{P(Y = 1|R = s, D = d)P(R = s, D = d)}{P(R = s, D = d) + P(R = 1 - s, D = d)}\\
&\quad + \frac{P(Y = 0|R = 1 - s, D = d)P(R = 1 - s, D = d)}{P(R = s, D = d) + P(R = 1 - s, D = d)}\\
&\overset{(b)}{=} s,
\end{aligned}$$

where (a) follows from the fact that $\hat{Y} = 1$ iff $R \geq 0.5$, and $\hat{Y} = 0$ iff $R < 0.5$, and (b) is due to the calibration by group assumption $P(Y = 0|R = 1 - s, D = d) = s$. □

By Lemma 5, the accuracy $P(Y = \hat{Y}|S = s, D = d)$ is independent of the group D given S and we can drop the conditioning of the group in the following proof.

In the selective classification problem, we convert the maximum a posteriori probability s into confidence κ using

$$\kappa(s) = \frac{1}{2} \log\left(\frac{s}{1 - s}\right), \tag{23}$$

which maps $[0.5, 1]$ to $[0, \infty]$. So for any sample with confidence $z \in \mathbb{R}^+$,

$$\begin{aligned}
P(Y = \hat{Y}|\kappa = z) &= P(Y = \hat{Y}|S = \kappa^{-1}(z))\\
&= \kappa^{-1}(z),
\end{aligned} \tag{24}$$

where $\kappa^{-1}(\cdot)$ is the inverse function of $\kappa(\cdot)$. We use $f_\kappa(z)$ to denote the pdf of the confidence score κ for $z \in \mathbb{R}^+$, then the pdf of the margin $f_M(t)$ can be written as,

$$f_M(t) = \begin{cases} P(Y = \hat{Y}|\kappa = t)f_\kappa(t), & \text{for } t \geq 0\\ P(Y \neq \hat{Y}|\kappa = -t)f_\kappa(-t), & \text{for } t < 0, \end{cases}$$

or equivalently,

$$f_M(t) = \begin{cases} \kappa^{-1}(t)f_\kappa(t), & \text{for } t \geq 0\\ (1 - \kappa^{-1}(-t))f_\kappa(-t), & \text{for } t < 0. \end{cases} \tag{25}$$

It can be verified that $\kappa^{-1}(z)$ is a increasing function for $z \in \mathbb{R}^+$, and $\kappa^{-1}(0) = \frac{1}{2}$. Thus,

$$\frac{f_M(z)}{f_M(-z)} = \frac{\kappa^{-1}(z)f_\kappa(z)}{(1 - \kappa^{-1}(z))f_\kappa(z)} = \frac{\kappa^{-1}(z)}{(1 - \kappa^{-1}(z))} \geq 1. \tag{26}$$

We can conclude that the cdf of the margin $F_M(t)$ satisfies

$$F_M(0) = \int_{-\infty}^{0} f_M(t)dt < \frac{1}{2}, \tag{27}$$

which implies that $A_F(0) > 0.5$.

To show that $A_F(\tau)$ is monotonically increasing with the threshold τ, we need to verify the condition in Lemma 4. Note that

$$\begin{aligned}
\frac{1 - F_M(\tau)}{F_M(-\tau)} &= \frac{\int_\tau^\infty f_M(t)dt}{\int_{-\infty}^{-\tau} f_M(t)dt}\\
&= \frac{\int_\tau^\infty \kappa^{-1}(t)f_\kappa(t)dt}{\int_\tau^\infty (1 - \kappa^{-1}(t))f_\kappa(t)dt}\\
&\geq \frac{\kappa^{-1}(\tau)\int_\tau^\infty f_\kappa(t)dt}{(1 - \kappa^{-1}(\tau))\int_\tau^\infty f_\kappa(t)dt}\\
&= \frac{f_M(\tau)}{f_M(-\tau)},
\end{aligned}$$

which completes the proof for the selective accuracy.

By replacing the margin distribution $f_M(t)$ with the margin distribution condition on $\hat{Y} = 1$, i.e., $f_{M|\hat{Y}=1}(t)$, the monotonicity of the precision $PPV_F(\tau)$ can be obtained following similar steps.

Note that the condition for monotonicity of the precision is given by

$$\frac{f_{M|\hat{Y}=1}(\tau)}{f_{M|\hat{Y}=1}(-\tau)} \leq \frac{1 - F_{M|\hat{Y}=1}(\tau)}{F_{M|\hat{Y}=1}(-\tau)}, \tag{28}$$

and Lemma 5 is replaced by the following simple fact due to calibration by group

$$\begin{aligned} &P(Y = 1|\hat{Y} = 1, S = s) \\ &= P(Y = 1|R = s) \\ &= s. \end{aligned} \tag{29}$$

In our proof, it only requires that the confidence function κ is a increasing function that maps $[0.5, 1]$ to $[0, \infty]$, so that $\kappa^{-1}(\cdot)$ is a increasing function and $\kappa^{-1}(0) = \frac{1}{2}$. Thus, Theorem 2 also holds for confidence functions satisfying these conditions, which is not limited to the function in (3).

B. Additional Experimental Results

B.1. Varying λ on Adult Dataset

To illustrate the relative insensitivity of our choice in λ (the regularizer weight for the conditional mutual information penalty term), we plot the area under the accuracy-coverage curve and (b) area between the precision-coverage curves against the value of λ for the Adult dataset in Figure 8, with 95% confidence intervals calculated from five trials per value of λ.

We see that for a wide range of λ, the performance of our method remains very stable, with the performance falling off only as λ grows close to zero.

B.2. Additional Baselines for Adult Dataset

In addition to the previous methods, we also implement three additional baselines on the Adult dataset. The first is the Chi-squared fairness regularizer of (Mary et al., 2019), which is designed for enforcing Equality of Opportunity in the full-coverage case, similarly to DRO. The second is logistic regression, which has been shown previously to provide well-calibrated scores at the cost of poorer performance on the dataset itself (Liu et al., 2019).

The final baseline is an oracle baseline whereby we use Platt Scaling (Platt et al., 1999) on each group individually to produce two different calibrated classifiers, then assume that we have the group label in the test set and apply the appropriate classifier. While this does represent a different setup than the other methods (which do not require the group label at test time), it does give us a good bound for the possible performance that might be expected. The precision-coverage curves can be found in 9, and the area under curve results can be found in Table 3.

Table 3. Area under curve results for all datasets.

Dataset	Method	Area under accuracy curve	Area between precision curves
Adult	Baseline	0.931	0.220
	DRO	0.911	0.116
	Chi-Sq.	0.920	0.140
	LogReg	0.620	0.231
	Ours	0.887	0.021
	Oracle	0.880	0.007

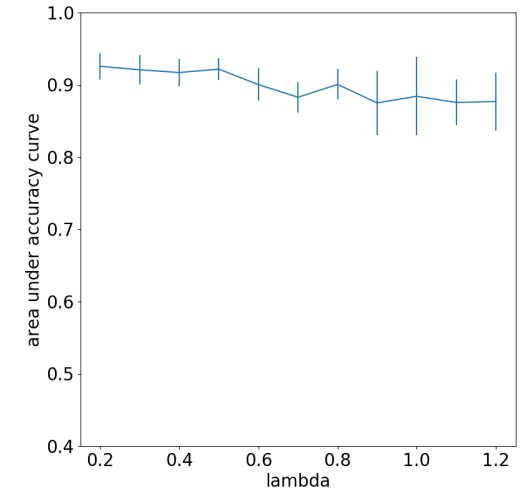
We can see that our method provides a lower area between precision curves than the Chi-squared regularizer, and that while logistic regression provides very low disparities across all coverages, it does so at the cost of incredibly poor average overall performance across all coverages.

B.3. CelebA and Civil Comments Datasets

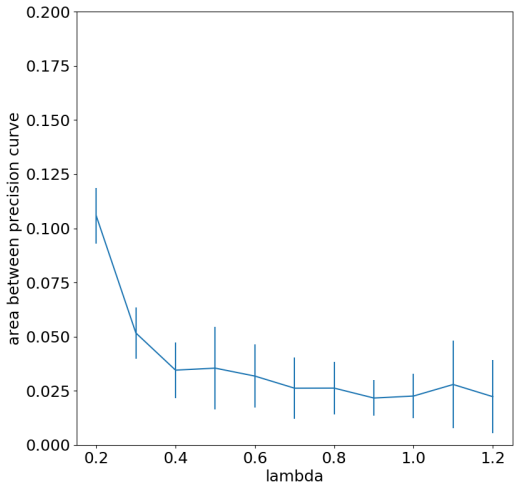
Figures 10 and 11 show the group-specific precision-coverage curves for the CelebA and Civil Comments datasets, with the margin distributions in Figures 12 and 13.

For the CelebA dataset, we note that while all methods converge towards perfect precision, the baseline has a significant period in which there is a large difference in precision between the two groups. This period is much smaller in the case of DRO and our sufficiency-based method, and ours ultimately converges the fastest.

The baseline method in the Civil Comments dataset shows the magnified disparities phenomenon in the precisions as coverage decreases, which is mitigated in the case of both DRO and our method. Our method also shows faster convergence of the two precision-coverage curves.

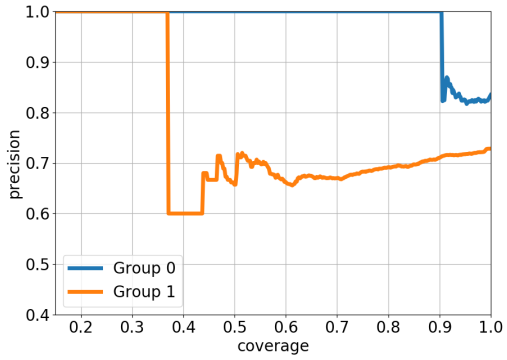


(a)

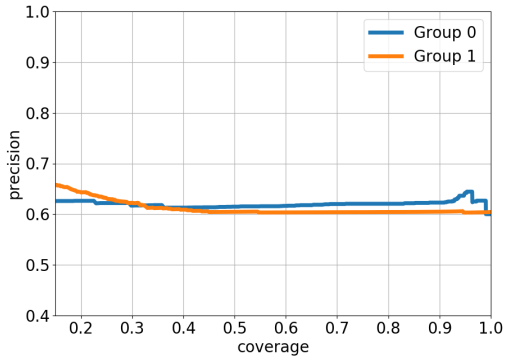


(b)

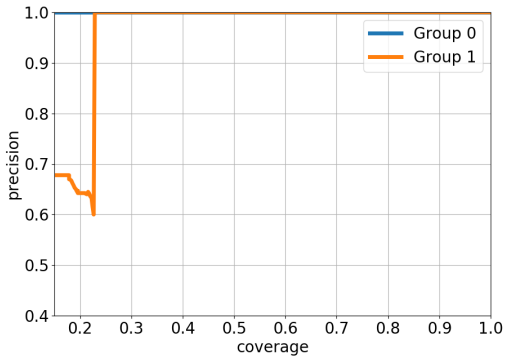
Figure 8. (a) Area under the accuracy-coverage curve and (b) area between the precision-coverage curves for different values of λ on the Adult dataset using our method.



(a) Chi-squared regularizer

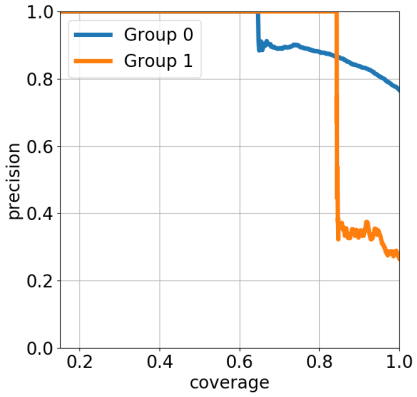


(b) Logistic Regression

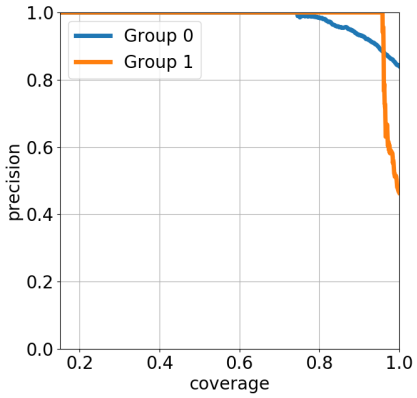


(c) Platt Scaling Oracle

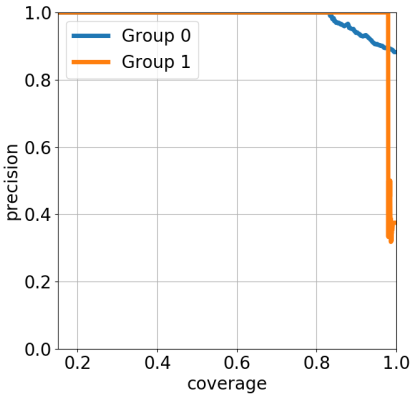
Figure 9. Group-specific precision-coverage curves for Adult dataset for additional methods.



(a) Baseline

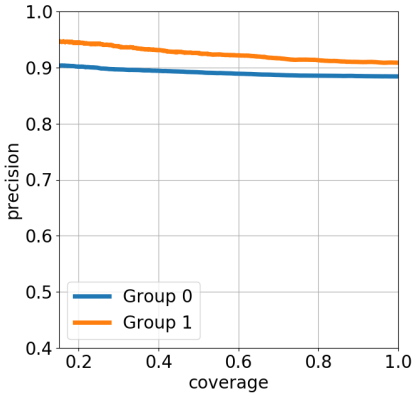


(b) DRO

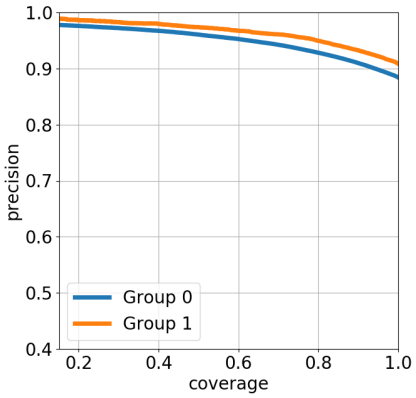


(c) Sufficiency-regularized

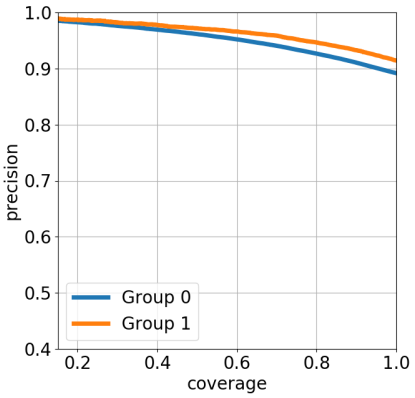
Figure 10. Group-specific precision-coverage curves for CelebA dataset for the three methods.



(a) Baseline

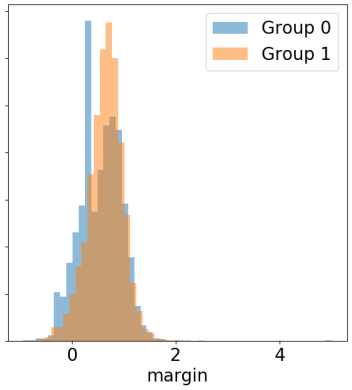


(b) DRO

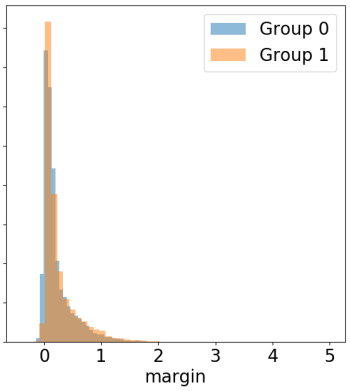


(c) Sufficiency-regularized

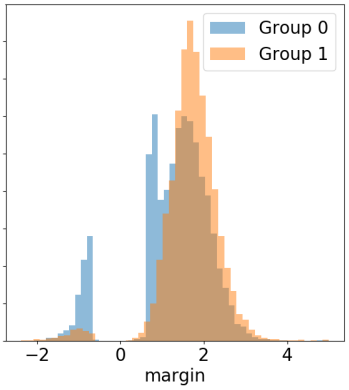
Figure 11. Group-specific precision-coverage curves for Civil Comments dataset for the three methods.



(a) Baseline

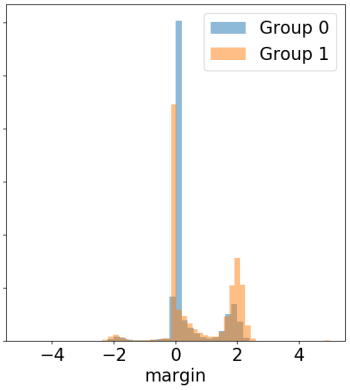


(b) DRO

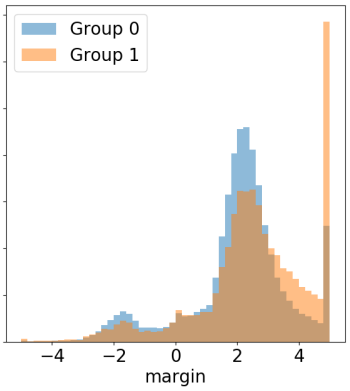


(c) Sufficiency-regularized

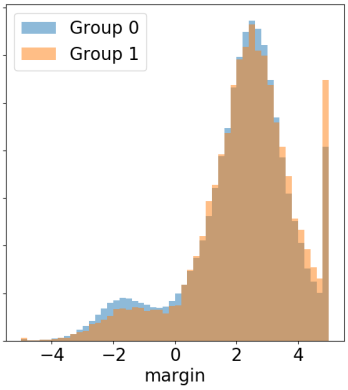
Figure 12. Margin distributions for CelebA dataset for the three methods.



(a) Baseline



(b) DRO



(c) Sufficiency-regularized

Figure 13. Margin distributions for Civil Comments dataset for the three methods.