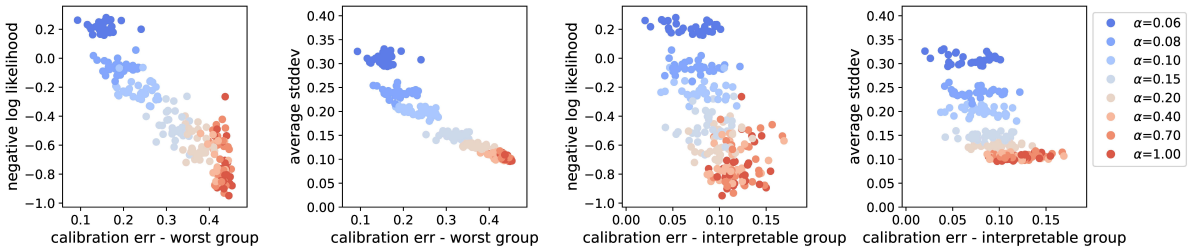

Figure 6: The same plot as Figure 3, but with recalibration by isotonic regression. The results are qualitatively the same with and without recalibration.

A. Experiments Details and Additional Results

A.1. Additional Theoretic Results

A.1.1. Relationship between PAIC and ECE

Given a forecaster $\mathbf{H}$ we can define expected calibration error (ECE) as

$$\mathrm{ECE}(\mathbf{H}) = \int_{c=0}^{1} |\Pr[\mathbf{H}[\mathbf{X}](\mathbf{Y}) \le c] - c| \, dc$$

Proposition 3. $\mathrm{ECE}(\mathbf{H}) = d_{W1}(\mathbb{F}_{\mathbf{H}[\mathbf{X}](\mathbf{Y})}, \mathbb{F}_{\mathbf{U}})$.

Intuitively, both $d_{W1}(\mathbb{F}_{\mathbf{H}[\mathbf{X}](\mathbf{Y})}, \mathbb{F}_{\mathbf{U}})$ try to integrate the difference between the curve $c \mapsto \Pr[\mathbf{H}[\mathbf{X}](\mathbf{Y}) \le c]$ and the curve $c \mapsto c$. The difference is that they integrate the difference in different ways (similar to the difference between Riemann and Lebesgue integral).

A.1.2. Trivial Construction of mPAIC forecaster

We construct a trivial forecaster that is always mPAIC. Let Φ be the standard Gaussian CDF, In particular for some $c > 0$, choose

$$\bar{h}[x, r](y) = \Phi(y/c - \Phi^{-1}(r))$$

Then when $c \to \infty$, we have $\bar{h}[x, r](y) = \Phi(\Phi^{-1}(r)) = r$. In other words, for any ϵ, δ, $\bar{h}$ is (ϵ, δ)-mPAIC for sufficiently large c. However, this forecaster is certainly not useful in practice because it outputs a distribution with variance $\to \infty$.

A.2. Fairness Experiment Details

We use the UCI crime and communities dataset (Dua & Graff, 2017) and we predict the crime rate based on features about the neighborhood (such as racial composition). The prediction model is a fully connected deep network, where the additional input r is concatenated into each hidden layer (except the last one). Other than this difference, all other setups are standard — with dropout and early stopping on validation data to prevent over-fitting. For details please refer to the code included with this paper.

During evaluation of calibration error for interpretable groups, we only consider groups with at least 150 samples to avoid excessive estimation error.

A.3. Additional Plots and Comparisons for Fairness Experiments

In Figure 6 we plot the same experimental results in Figure 3, where the only difference is we apply post-training re-calibration (Kuleshov et al., 2018). There is no qualitative difference between Figure 3 and Figure **??** because (average) calibration does not improve calibration for the worst group.

A.4. Experiment Details for Credit Approval

Dataset We will use the "Give Me Some Credit" dataset on Kaggle. Because it is a binary classification dataset (credit delinquency vs. no delinquency), we first train a classifier to predict the Bernoulli probability, and use the probability (plus

	No Recalibration	With Recalibration
$\alpha = 0.1$ (PAIC)	population density (+) pct drug officer (-)	pct immigrant 8yr (-) vac house boarded (-)
$\alpha = 1.0$ (NLL)	pct black (+) pct < 3 bedroom (+)	pct immigrant (+) pct dense house (+)

Table 1: Least calibrated group for each setup. A + sign indicates this feature is above the median and a - sign indicates the feature is below the median. These are indeed groups where fairness can be a consideration (e.g. immigrants, race or economic condition).

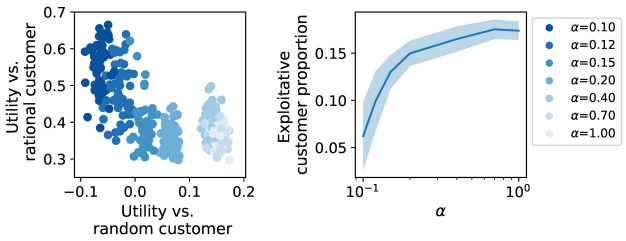

Figure 7: The experiment in Figure 5 without post training recalibration.

a small Gaussian noise) as the label. We synthesize a training set and a validation set, where the validation set is very large to simulate a stream of non-repeating customers. We train the bank's forecaster $\mathbf{H}$ on the training set, and apply it to interacting with customers sampled from the validation set.

Customer Model The customer utility we use is

	$y \geq y_0$	$y < y_0$
'yes'	0.2	1.0
'no'	-0.5	-0.5

We assume the customers knows their own credit worthiness y. Based on previous customers x, y and the actual utility from playing the game, we learn a function $\psi(x, y) \to \mathbb{R}$ by gradient descent to predict the customer's utility. The prediction function ψ is also a fully connected deep neural network. Each new customer $(x_{\text{new}}, y_{\text{new}}) \sim \mathbb{F}_{\mathbf{XY}}$ will only apply if $\psi(x_{\text{new}}, y_{\text{new}}) \geq 0$.

Decision Rule The "Bayesian" decision rule in Eq.(6) can be written as

$$\phi_{\mathbf{H}}(x) = \begin{cases} \text{'yes'} & \mathbf{H}[x](y_0) \leq 1/4 \\ \text{'no'} & \text{otherwise} \end{cases} \tag{7}$$

Recalibration Since post training recalibration (Kuleshov et al., 2018; Malik et al., 2019) is usually beneficial, we will report both results with and without recalibration by isotonic regression. The results with recalibration is in Figure 5 and the results without recalibration is in Figure 7.

A.5. Additional Plots for Credit Approval

In Figure 7 we plot the results without post training recalibration. They are qualitatively similar to Figure 5. Post training recalibration has little effect on calibration of the worst sub-group, and therefore do not improve performance in this experiment.

A.6. Additional Discussion

Stochastic vs. Deterministic Distributions Our results are most useful when $\mathbf{Y} \mid \mathbf{X}$ is almost deterministic, i.e. the uncertainty comes from model ignorance instead of the environment. When $\mathbf{Y} \mid \mathbf{X}$ is highly stochastic, Definition 5 can still be achieved, but with a significant sacrifice to sharpness. This is because $\bar{h}[x, r]$ must be a high variance distribution

to satisfy $\bar{h}[x, r](\mathbf{Y}) \approx r$ for all likely values for $\mathbf{Y} \mid x$. Even though individual calibration can still be (approximately) achieved, the sharpness may be prohibitively poor in practice.

B. Proofs

B.1. Proofs for Section 3

Proposition 1. *For any distribution $\mathbb{F}$ on $\mathcal{X} \times \mathcal{Y}$ such that $\mathbb{F}_{\mathbf{X}}$ assigns zero measure to individual points $\{x \in \mathcal{X}\}$ sample $\mathcal{D} = \{(x_1, y_1), \cdots, (x_n, y_n)\} \overset{i.i.d.}{\sim} \mathbb{F}$. For any deterministic forecaster h and any function $T(\mathcal{D}, h) \to \{\text{yes}, \text{no}\}$ such that*

$$\Pr_{\mathcal{D}\sim\mathbb{F}}[T(\mathcal{D}, h) = \text{yes}] = \kappa > 0\,,$$

there exists a distribution $\mathbb{F}'$ such that (a) h is not (ϵ, δ)-PAIC w.r.t $\mathbb{F}'$ for any $\epsilon < 1/4$ and $\delta < 1$, and (b)

$$\Pr_{\mathcal{D}\sim\mathbb{F}'}[T(\mathcal{D}, h) = \text{yes}] \geq \kappa$$

Proof of Proposition 1. Given a distribution $\mathbb{F}$ and forecaster h such that $\Pr_{\mathcal{D}\sim\mathbb{F}}[T(\mathcal{D}, h) = \text{yes}] = \kappa > 0$, we will construct an alternative distribution $\mathbb{F}'$ by choosing some function $g : \mathcal{X} \to \mathcal{Y}$ (defined later), and define a new distribution $\mathbf{X}', \mathbf{Y}' \sim \mathbb{F}'_g$ by: $\mathbf{X}' \sim \mathbb{F}_{\mathbf{X}}$ and $\mathbf{Y}' \mid x$ is the delta distribution on $g(x)$. Then by Definition 1, $\forall x \in \mathcal{X}$

$$d_{\text{W1}}(\mathbb{F}_{h[x](\mathbf{Y}')}, \mathbb{F}_{\mathbf{U}}) \geq 1/4 \tag{8}$$

In words, the above expression is because for any distribution $h[x]$ outputs, we can never rule out a possible ground truth distribution ($\mathbb{F}'_g$) that is deterministic. Under a deterministic distribution $\mathbb{F}_{\mathbf{Y}'|x}$, it must be that Eq.(9) holds. (An alternative construction can strengthen the theorem by choosing $\mathbf{Y}' \mid x$ to be a distribution with sufficiently small non-zero variance. It can become clear that Eq.(9) is not an artifact of our requirement that h must output a continuous CDF, but rather the the variance of the ground truth distribution cannot be known).

What remains to show is that there must exist a g such that

$$\Pr_{\mathcal{D}\sim\mathbb{F}'_g}[T(\mathcal{D}, h) = \text{yes}] \geq \kappa$$

We will do this with the probabilistic method. For convenience we will represent the value 'yes' by 1 and the value 'no' by 0. We can use the notation

$$\Pr_{\mathcal{D}\sim\mathbb{F}}[T(\mathcal{D}, h) = \text{yes}] := \mathbb{E}_{\mathcal{D}\sim\mathbb{F}}[T(\mathcal{D}, h)]$$

Because $\mathbb{F}_{\mathbf{X}}$ assigns zero measure to individual points, for any finite set of $x_1, \cdots, x_n \overset{\text{i.i.d.}}{\sim} \mathbb{F}_{\mathbf{X}}$, all the x_i are distinct (i.e. $x_i \neq x_j, \forall i \neq j$) almost surely. Suppose $\mathbf{G}$ is a random function on $\{g : \mathcal{X} \to \mathcal{Y}\}$ defined by $\mathbf{G}(x) \sim \mathbb{F}_{\mathbf{Y}|x}$, then random variables $\mathcal{D} = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ defined by the following two sampling procedures are identically distributed (i.e. in the sense that they belong to any measurable subset of $(\mathcal{X} \times \mathcal{Y})^n$ with the same probability)

$$\begin{aligned} &x_1, \cdots, x_n \overset{\text{i.i.d.}}{\sim} \mathbb{F}_{\mathbf{X}}, \qquad y_i \sim \mathbb{F}_{\mathbf{Y}|x_i} \\ &x_1, \cdots, x_n \overset{\text{i.i.d.}}{\sim} \mathbb{F}_{\mathbf{X}}, \qquad g \sim \mathbf{G}, \qquad y_i = g(x_i) \end{aligned}$$

In words, we could either 1. directly sample a dataset $\mathcal{D}$ from $\mathbb{F}$, or 2. we could first sample a value $g(x) \sim \mathbb{F}_{\mathbf{Y}|x}$ for each x, then sample $x_1, \cdots, x_n \sim \mathbb{F}_{\mathbf{X}}$ and directly evaluate $y_1 = g(x_1), \cdots, y_n = g(x_n)$.

Therefore any bounded random variable must have identical expectation under the probability law defined by the two sampling procedures

$$\kappa = \mathbb{E}_{\mathcal{D}\sim\mathbb{F}}[T(\mathcal{D}, h)] = \mathbb{E}_{g\sim\mathbf{G}}\mathbb{E}_{\mathcal{D}\sim\mathbb{F}'_g}[T(\mathcal{D}, h)]$$

But this must imply there exists g such that

$$\mathbb{E}_{\mathcal{D}\sim\mathbb{F}'_g}[T(\mathcal{D}, h)] \geq \kappa$$

because a random variable must be able to take a value that is at least its expectation

For any other divergence we can get similar results by replacing Eq.(9). For example, for total variation distance we have

$$d_{\mathrm{TV}}(\mathbb{F}_{h[x](\mathbf{Y}')}, \mathbb{F}_{\mathbf{U}}) = 1 \tag{9}$$

□

Proposition 4. *For any distribution $\mathbb{F}$ on $\mathcal{X} \times \mathcal{Y}$ such that $\mathbb{F}_{\mathbf{X}}$ assigns zero measure to individual points $\{x \in \mathcal{X}\}$ sample $\mathcal{D} = \{(x_1, y_1), \cdots, (x_n, y_n)\} \overset{i.i.d.}{\sim} \mathbb{F}$. For any deterministic forecaster h and any function $T(\mathcal{D}, h) \to \{\mathrm{yes}, \mathrm{no}\}$ such that*

$$\Pr_{\mathcal{D}\sim\mathbb{F}}[T(\mathcal{D}, h) = \mathrm{yes}] = \kappa > 0\,,$$

then there exists a distribution $\mathbb{F}'$, h is not (ϵ, δ)-adversarial group calibrated with respect to $\mathbb{F}'$ for any $\epsilon < 1/8$ and $\delta < 1/2$, and

$$\Pr_{\mathcal{D}\sim\mathbb{F}'}[T(\mathcal{D}, h) = \mathrm{yes}] \geq \kappa$$

Proof of Proposition 4. The proof is almost identical to the proof of Proposition 1. We construct a $g : \mathcal{X} \to \mathcal{Y}$ such that

$$\Pr_{\mathcal{D}\sim\mathbb{F}'}[T(\mathcal{D}, h) = \mathrm{yes}] \geq \kappa$$

We can pick the subgroup $\mathcal{S}_1, \mathcal{S}_2 \subset \mathcal{X}$ defined by

$$\mathcal{S}_1 = \{x, h[x](g(x)) \geq 1/2\}, \mathcal{S}_2 = \{x, h[x](g(x)) < 1/2\}$$

Because $\mathcal{S}_1 \cup \mathcal{S}_2 = \mathcal{X}$, so at least one of $\mathcal{S}_1, \mathcal{S}_2$ must have probability measure at least $1/2$ under $\mathbb{F}_X$. Without loss of generality assume it's $\mathcal{S}_1$. Then for $\tilde{X} = X \mid \mathcal{S}_1$ we have $h[\tilde{X}](g(x)) \geq 1/2$ almost surely, $\mathbb{F}_{h[\tilde{X}](g(x))}(r) = 0, \forall r \in [0, 1/2]$, which implies

$$d_{W1}(\mathbb{F}_{h[\tilde{X}](g(x))}, \mathbb{F}_{\mathbf{U}}) \geq 1/8$$

Therefore h cannot be $(1/8, 1/2)$-adversarial group calibrated. □

B.2. Proofs for Section 4.2

Theorem 1. *If $\bar{h}$ is (ϵ, δ)-mPAIC, then for any $\epsilon' > \epsilon$ it is $(\epsilon', \delta(1-\epsilon)/(\epsilon'-\epsilon))$-PAIC with respect to the 1-Wasserstein distance.*

Proof of Theorem 1. Recall the convention that $\mathbf{Y}$ is the random variable that always has the conditional distribution $\mathbb{F}_{\mathbf{Y}|x}$, and $\mathbf{R}$ is uniformly distributed in $[0, 1]$. Denote

$$\mathrm{err}(x, y) := d_{W_1}\left(\mathbb{F}_{\bar{h}[x,\mathbf{R}](y)}, \mathbb{F}_{\mathbf{U}}\right), \qquad \mathrm{err}(x) := d_{W_1}\left(\mathbb{F}_{\bar{h}[x,\mathbf{R}](\mathbf{Y})}, \mathbb{F}_{\mathbf{U}}\right)$$

Suppose $\bar{h}[x, \cdot](y)$ is a monotonically non-decreasing function for all x, y, then

$$\mathrm{err}(x, y) = d_{W_1}\left(\mathbb{F}_{\bar{h}[x,\mathbf{R}](y)}, \mathbb{F}_{\mathbf{U}}\right) = \int_{r=0}^{1} |\bar{h}(x, r)(y) - r| dr = \mathbb{E}\left[|\bar{h}(x, \mathbf{R})(y) - \mathbf{R}|\right]$$

So in general for arbitrary $\bar{h}$ we have

$$\mathrm{err}(x, y) = d_{W_1}\left(\mathbb{F}_{\bar{h}[x,\mathbf{R}](y)}, \mathbb{F}_{\mathbf{U}}\right) \leq \int_{r=0}^{1} |\bar{h}(x, r)(y) - r| dr = \mathbb{E}\left[|\bar{h}(x, \mathbf{R})(y) - \mathbf{R}|\right] \tag{10}$$

In addition by Jensen's inequality we have $\mathrm{err}(x) \leq \mathbb{E}[\mathrm{err}(x, \mathbf{Y})]$ so

$$\mathrm{err}(x) \leq \mathbb{E}\left[|\bar{h}(x, \mathbf{R})(\mathbf{Y}) - \mathbf{R}|\right] \tag{11}$$

Suppose $\bar{h}$ is not (ϵ', δ')-PAIC, by definition we have

$$\Pr[\mathrm{err}(\mathbf{X}) \geq \epsilon'] > \delta'$$

Define the notation

$$\mathcal{S}_b := \{x \in \mathcal{X}, \mathbb{E}\left[|\bar{h}(x, \mathbf{R})(\mathbf{Y}) - \mathbf{R}|\right] \geq \epsilon'\}$$

by Eq.(11) we know that whenever $\mathrm{err}(x) \geq \epsilon'$ we have $x \in \mathcal{S}_b$, so we can conclude

$$\Pr[\mathbf{X} \in \mathcal{S}_b] > \delta' \tag{12}$$

Whenever $x \in \mathcal{S}_b$, for any $\epsilon < \epsilon'$, we have

$$\begin{aligned}
\epsilon' &\leq \mathbb{E}[|\bar{h}(x, \mathbf{R})(\mathbf{Y}) - \mathbf{R}|] \\
&\leq \epsilon \Pr[|\bar{h}(x, \mathbf{R})(\mathbf{Y}) - \mathbf{R}| < \epsilon] + \Pr[|\bar{h}(x, \mathbf{R})(\mathbf{Y}) - \mathbf{R}| \geq \epsilon] \\
&= \epsilon(1 - \Pr[|\bar{h}(x, \mathbf{R})(\mathbf{Y}) - \mathbf{R}| \geq \epsilon]) + \Pr[|\bar{h}(x, \mathbf{R})(\mathbf{Y}) - \mathbf{R}| \geq \epsilon]
\end{aligned}$$

where the second inequality is because $|\bar{h}(x, \mathbf{R})(\mathbf{Y}) - \mathbf{R}|$ is bounded in $[0, 1]$. By simple algebra we get

$$\Pr[|\bar{h}(x, \mathbf{R})(\mathbf{Y}) - \mathbf{R}| \geq \epsilon] \geq \frac{\epsilon' - \epsilon}{1 - \epsilon} \tag{13}$$

We can combine Eq.(12) and Eq.(13) to get

$$\begin{aligned}
&\Pr[|\bar{h}(\mathbf{X}, \mathbf{R})(\mathbf{Y}) - \mathbf{R}| \geq \epsilon'] \\
&= \Pr[|\bar{h}(\mathbf{X}, \mathbf{R})(\mathbf{Y}) - \mathbf{R}| \geq \epsilon \mid \mathbf{X} \in \mathcal{S}_b] \Pr[\mathbf{X} \in \mathcal{S}_b] + \Pr[|\bar{h}(\mathbf{X}, \mathbf{R})(\mathbf{Y}) - \mathbf{R}| \geq \epsilon \mid \mathbf{X} \notin \mathcal{S}_b] \Pr[\mathbf{X} \notin \mathcal{S}_b] \\
&> \frac{\epsilon' - \epsilon}{1 - \epsilon} \delta'
\end{aligned}$$

Therefore, $\bar{h}$ is not $(\epsilon, \frac{\epsilon' - \epsilon}{1 - \epsilon}\delta')$-mPAIC. To summarize, we have concluded that whenever $\bar{h}$ is not (ϵ', δ')-PAIC, for any $\epsilon < \epsilon'$, it is not $(\epsilon, \frac{\epsilon' - \epsilon}{1 - \epsilon}\delta')$-mPAIC. This is equivalent to the statement: suppose $\bar{h}$ is (ϵ, δ)-mPAIC, then $\bar{h}$ is $(\epsilon', \delta\frac{1-\epsilon}{\epsilon' - \epsilon})$-PAIC. □

Proposition 2. *[Concentration] Let $\bar{h}$ be any (ϵ, δ)-mPAIC forecaster, and $(x_1, y_1), \cdots, (x_n, y_n) \overset{i.i.d.}{\sim} \mathbb{F}_{\mathbf{XY}}$, $r_1, \cdots, r_n \overset{i.i.d.}{\sim} \mathbb{F}_{\mathbf{U}}$, then with probability $1 - \gamma$*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(|\bar{h}[x_i, r_i](y_i) - r_i| \geq \epsilon) \leq \delta + \sqrt{\frac{-\log \gamma}{2n}}$$

Proof of Proposition 2. Consider the sequence of Bernoulli random variables $b_i = \mathbb{I}(\left|\bar{h}(x_i, r_i)(y_i) - r_i\right| \geq \epsilon)$. Suppose $\mathbb{E}[b_i] = \delta$, then by Hoeffding inequality

$$Pr\left[\frac{1}{n}\sum_i b_i \geq \delta + \epsilon\right] \leq e^{-2\epsilon^2 n}$$

Plugging in $e^{-2\epsilon^2 T}$ as γ we have $\epsilon = \sqrt{\frac{-\log \gamma}{2n}}$ □

B.3. Proofs for Section 5

Theorem 2. *If a forecaster is (ϵ, δ)-PAIC with respect to distance metric $\mathcal{W}_p$, then $\forall \delta' \in [0, 1]$, $\delta' > \delta$, it is $(\epsilon + \delta/\delta', \delta')$-adversarial group calibrated with respect to $\mathcal{W}_p$.*

Proof of Theorem 2. Given a forecaster $\mathbf{H}$ if for some x, y we have $d_{\mathcal{W}_p}(\mathbb{F}_{\mathbf{H}[x](y)}, \mathbb{F}_{\mathbf{U}}) < \epsilon$ then by definition of the Wasserstein distance we have

$$\int_{r=0}^{1} |\mathbb{F}_{\mathbf{H}[x](y)}(r) - r|^p \leq \epsilon^p$$

If $\mathbf{H}$ is (ϵ, δ)-mPAIC with respect to $\mathcal{W}_p$. Consider a partition of $\mathcal{X}$ into two sets: $\mathcal{X}_g$ where $\forall x \in \mathcal{X}_g$ we have

$$\int_r \left|\mathbb{F}_{\mathbf{H}[x](y)}(r) - r\right|^p dr \leq \epsilon^p$$

and $\mathcal{X}_b$ where the above property fails. We know $\Pr[\mathcal{X}_b] \leq \delta$. In general for any $x \in \mathcal{X}$, because $\mathbb{F}_{\mathbf{H}[x](y)}$ is a monotonically increasing function of c bounded in $[0,1]$, we have

$$\int_r \left|\mathbb{F}_{\mathbf{H}[x](y)}(r) - r\right|^p dc \leq \int_r r^p dr = \frac{1}{p+1}$$

Another useful identity we will use is

$$\mathbb{F}_{\mathbf{H}[\mathbf{X}](\mathbf{Y})}(r) = \Pr[\mathbf{H}[\mathbf{X}][\mathbf{Y}] \leq r] = \mathbb{E}_{x\sim\mathbb{F}_{\mathbf{X}}}[\mathbb{E}[\mathbb{I}(\mathbf{H}[x][\mathbf{Y}] \leq r)]] = \mathbb{E}_{x\sim\mathbb{F}_{\mathbf{X}}}[\mathbb{F}_{\mathbf{H}[x](\mathbf{Y})}(r)] \tag{14}$$

Combining the above results we have for any $\tilde{X} = X \mid \mathcal{S}$

$$\begin{aligned}
d_{\mathcal{W}_p}(\mathbb{F}_{\mathbf{H}[\tilde{X}](Y)}, \mathbb{F}_{\mathbf{U}}) &= \left(\int_{r=0}^1 |\mathbb{F}_{\mathbf{H}[\tilde{X}](Y)}(r) - r|^p dr\right)^{1/p} \\
&= \left(\int_{r=0}^1 \left|\mathbb{E}_{x\sim\tilde{X}}[\mathbb{F}_{\mathbf{H}[x](\mathbf{Y})}(r) - r]\right|^p dr\right)^{1/p} && \text{(Eq.14)} \\
&\leq \mathbb{E}_{x\sim\tilde{X}}\left[\left(\int_{r=0}^1 |\mathbb{F}_{\mathbf{H}[x](\mathbf{Y})}(r) - r|^p dr\right)^{1/p}\right] && \text{(Jensen)} \\
&= \mathbb{E}_{x\sim\tilde{X}}\left[\left(\int_{r=0}^1 |\mathbb{F}_{\mathbf{H}[x](\mathbf{Y})}(r) - r|^p dr\right)^{1/p} \mid x \in \mathcal{X}_g\right] \Pr[\tilde{X} \in \mathcal{X}_g] + \\
&\quad \mathbb{E}_{x\sim\tilde{X}}\left[\left(\int_{r=0}^1 |\mathbb{F}_{\mathbf{H}[x](\mathbf{Y})}(r) - r|^p dr\right)^{1/p} \mid x \in \mathcal{X}_b\right] \Pr[\tilde{X} \in \mathcal{X}_b] && \text{(Conditional Expectation)} \\
&\leq \epsilon \Pr[\tilde{X} \in \mathcal{X}_g] + (p+1)^{-1/p} \Pr[\tilde{X} \in \mathcal{X}_b] \\
&\leq \epsilon \frac{\delta' - \delta}{\delta'} + (p+1)^{-1/p} \frac{\delta}{\delta'} && (\epsilon \leq (p+1)^{-1/p})
\end{aligned}$$

If we don't care about constants too much, we can further simplify above by

$$\epsilon \frac{\delta' - \delta}{\delta'} + (p+1)^{-1/p} \frac{\delta}{\delta'} \leq \epsilon + \delta/\delta'$$

□

B.4. Proofs for Section 6

Theorem 3. *Suppose $l : \mathcal{X} \times \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$ is a monotonic non-negative loss, let $\phi_{\mathbf{H}}$ and $l_{\mathbf{H}}$ be defined as in Eq.(6)*

1. If $\mathbf{H}$ is 0-average calibrated, then $\forall k > 0$

$$\Pr[l(\mathbf{X}, \mathbf{Y}, \phi_{\mathbf{H}}(\mathbf{X})) \geq k l_{\mathbf{H}}(\mathbf{X})] \leq 2/k$$

2. If $\mathbf{H}$ is $(0,0)$-PAIC, then $\forall x \in \mathcal{X}, k > 0$

$$\Pr[l(x, \mathbf{Y}, \phi_{\mathbf{H}}(x)) \geq k l_{\mathbf{H}}(x)] \leq 1/k$$

Proof of Theorem 3. Choose any $x \in \mathcal{X}$, $h \in \mathcal{H}$ and $r \in (0,1)$. For some action a assume $l(x, \cdot, a)$ is monotonically non-decreasing. We consider the situation where $y < h[x]^{-1}(1-r)$, or equivalently $h[x](y) < 1 - r$, then

$$l_h(x) = \int_{y' \in \mathcal{Y}} l(x, y', a) dh[x](y') \geq \int_{y' \geq y} l(x, y', a) dh[x](y') \geq l(x, y, a) \int_{y' \geq y} dh[x](y') \geq r l(x, y, a)$$

because the above is true for any $a \in \mathcal{A}$, it must also be true for the action $\phi_h(x)$. On the other hand, assume $l(x, \cdot, a)$ is monotonically non-increasing, then by a similar argument we get whenever $h[x](y) > r$ we have

$$l_h(x) \geq r l(x, y, a)$$

Consider the set $\mathcal{S}_r, \mathcal{M}_r, \bar{\mathcal{M}}_r \subset \mathcal{X} \times \mathcal{Y} \times \mathcal{H}$, defined by

$$\mathcal{S}_r = \{x, y, h \mid l_h(x) \leq rl(x, y, a)\}, \quad \mathcal{M}_r = \{x, y, h \mid h[x](y) \leq r\}, \quad \bar{\mathcal{M}}_r = \{x, y, h \mid h[x](y) \geq 1 - r\}$$

The above results would imply $\mathcal{S}_r \subset \mathcal{M}_r \cup \hat{\mathcal{M}}_r$. But we know that

$$\Pr[\mathbf{X}, \mathbf{Y}, \mathbf{H} \in \mathcal{S}_r] \leq \Pr[\mathbf{X}, \mathbf{Y}, \mathbf{H} \in \mathcal{M}_r \cup \hat{\mathcal{M}}_r] \leq 2r$$

taking $k = 1/r$ gives us the desired statement.

If $\mathbf{H}$ is individually calibrated, then it is also adversarial group calibrated by Theorem 2. Define a function $\zeta : \mathcal{X} \times \mathcal{H} \to \{0, 1\}$ that represents whether l is monotonically non-decreasing or non-increasing in $\mathcal{Y}$. Then

$$\begin{aligned}
&\Pr[\mathbf{X}, \mathbf{Y}, \mathbf{H} \in \mathcal{S}_r] \\
&\leq \Pr[\mathbf{X}, \mathbf{Y}, \mathbf{H} \in \mathcal{M}_r \mid \zeta(\mathbf{X}, \mathbf{H}) = 0] \Pr[\zeta(\mathbf{X}, \mathbf{H}) = 0] + \Pr[\mathbf{X}, \mathbf{Y}, \mathbf{H} \in \mathcal{M}_r \mid \zeta(\mathbf{X}, \mathbf{H}) = 1] \Pr[\zeta(\mathbf{X}, \mathbf{H}) = 1] \\
&\leq r
\end{aligned}$$

□

B.5. Proofs for Appendix

Proposition 3. $\mathrm{ECE}(\mathbf{H}) = d_{W1}(\mathbb{F}_{\mathbf{H}[\mathbf{X}](\mathbf{Y})}, \mathbb{F}_{\mathbf{U}})$.

Proof of Proposition 3. This proposition depends on the following Lemma.

Lemma 1. *Let $\phi : [0, 1] \to [0, 1]$ be a monotonic differentiable function such that $\phi(0) = 0$ and $\phi(1) = 1$. Let $\psi(x) = x$, then for any $1 \leq s \leq +\infty$ we have*

$$\int_{c=0}^{1} \left|\phi^{-1}(c) - \psi(c)\right|^s dc = \int_{r=0}^{1} |\phi(r) - \psi(r)|^s dr$$

First observe that by the monotonicity of $\mathbb{F}_{\mathbf{H}[\mathbf{X}](\mathbf{Y})}$ we have

$$\Pr[\mathbf{H}[\mathbf{X}](\mathbf{Y}) \leq c] = \int_{r=0}^{1} \mathbb{I}(\mathbb{F}_{\mathbf{H}[\mathbf{X}](\mathbf{Y})} \leq c) dr = {\mathbb{F}_{\mathbf{H}[\mathbf{X}](\mathbf{Y})}}^{-1}(c)$$

We get

$$\begin{aligned}
\mathrm{ECE}(\mathbf{H}) &= \int_{c=0}^{1} |\Pr[\mathbf{H}[\mathbf{X}](\mathbf{Y}) \leq c] - c| \, dc = \int_{c=0}^{1} \left|{\mathbb{F}_{\mathbf{H}[\mathbf{X}](\mathbf{Y})}}^{-1}(c) - c\right| dc \\
&= \int_{r=0}^{1} \left|\mathbb{F}_{\mathbf{H}[\mathbf{X}](\mathbf{Y})}(r) - r\right| dr = d_{W1}(\mathbb{F}_{\mathbf{H}[\mathbf{X}](\mathbf{Y})}, \mathbb{F}_{\mathbf{U}})
\end{aligned}$$

Finally we prove Lemma 1.

Proof of Lemma 1. Let $[a, b]$ be an interval where $\phi(x) - x$ does not change sign, and $f(a) = a$, $f(b) = b$. Without loss of generality, assume it is positive. Then

$$\begin{aligned}
\int_{x=a}^{b} |\phi(x) - x|^s dx - \int_{y=a}^{b} (f^{-1}(y) - y)^s dy &= \int_{x=a}^{b} (\phi(x) - x)^s dx - \int_{x=a}^{b} (-x + \phi(x))^s f'(x) dx \\
&= \int_{x=a}^{b} (\phi(x) - x)^s (f'(x) - 1) dx = \frac{(\phi(x) - x)^{s+1}}{s+1} |_a^b = 0
\end{aligned}$$

Let $0 = a_1 < a_2 < \cdots < a_n = 1$ be a set of points where $f(a_i) = a_i$, and f does not change sign between $[a_i, a_{i+1}]$. Then we have

$$\int_{x=0}^{1} |\phi(x) - x|^s dx - \int_{y=0}^{1} |f^{-1}(y) - y|^s dy$$
$$= \sum_i \left(\int_{x=a_i}^{a_{i+1}} |\phi(x) - x|^s dx - \int_{y=a_i}^{a_{i+1}} |f^{-1}(y) - y|^s dy \right) = 0$$

□

□