A $L_{\mathcal{D}}(h) = \texttt{FPR}_{\mathcal{D}}(h)$ doesn't have the uniform convergence property

Recall that the false positive rate of a binary classifier can be written as

$$\texttt{FPR}_{\mathcal{D}}(h) \equiv \mathbf{Pr}\left[h(x) = 1 | y = 0\right]$$

For intuition, the fact that we condition on the event $y = 0$ means that for distributions in which the probability of this event is small, a good estimate of the true loss will require more samples. Intuitively, this contradicts the uniform convergence requirement that there is a single number of samples that "works" for every distribution $\mathcal{D}$.

Proof. Suppose $\mathcal{X}$ is finite and $|\mathcal{X}| = n$. Fix some element $x' \in \mathcal{X}$, and consider the distribution $\mathcal{D}$ on $\mathcal{X} \times Y$ obtained by taking a uniform distribution over $\mathcal{X}$ and labeling elements via the deterministic labeling function

$$y(x) = \begin{cases} 0 & x \neq x' \\ 1 & x = x' \end{cases}$$

Consider the classifier h that labels everyone as 1: $h(x) \equiv 1$. Then, $\texttt{FPR}_{\mathcal{D}}(h) = 1$ (since there is a single negative example, which is incorrectly labeled as a positive). On the other hand, the empirical estimate w.r.t any sample $S \subseteq \mathcal{X} - \{x'\}$ is zero, so for such a sample, the difference between $L_S(h)$ and $L_D(h)$ is at a maximal value of 1. Recall that uniform convergence requires us to estimate this difference to arbitrary precision with high probability; therefore, the "bad event" in which $x \notin S$ must happen with probability at most δ. Equivalently,

$$\left(\frac{n-1}{n}\right)^m \leq \delta$$

This require taking m large enough to guarantee that $m \geq \log \frac{1}{\delta} \cdot \frac{1}{\log \frac{n}{n-1}}$. However, there is no function $f : (0,1) \to \mathbb{N}$ that guarantees that $m \geq f(\delta)$ satisfies this condition *for every n*, because as n approaches infinity, $\log \frac{n}{n-1} \to 0$, so the expression is unbounded.

□

B Exist $L \in \mathcal{L}$ that are not multi-group compatible

Proof. For the counter-example we focus on binary classification and individual (metric) fairness w.r.t a binary metric (i.e., that specifies for every two individuals, whether they are identical or completely different). Fixing a metric $d : \mathcal{X} \times \mathcal{X} \to \{0,1\}$, the loss function L is a combination of accuracy and individual (metric) fairness:

$$\begin{aligned} L_{\mathcal{D}}(h) &= a \cdot L^{IF}_{\mathcal{D}}(h) + b \cdot L^{0-1}_{\mathcal{D}}(h) \\ &\equiv a \cdot \mathop{\mathbf{Pr}}_{x,x' \sim \mathcal{D}_X}\left[h(x) \neq h(x') \wedge d(x,x') = 0\right] + b \cdot \mathop{\mathbf{Pr}}_{x,y \sim \mathcal{D}}\left[h(x) \neq y\right] \end{aligned}$$

Let us now construct the problem instance in question. Let the domain $\mathcal{X}$ be $\mathcal{X} = \{x_S, x_T, x_{ST}\}$, with $\mathcal{D}_X$ denoting the marginal distribution on $\mathcal{X}$ in which x_S has probability 0.8, and x_T, x_{ST} each have probability of 0.1. A distribution $\mathcal{D}$ is obtained as the product of $\mathcal{D}_X$ and $\mathcal{D}_{Y|X=x}$, where the latter assigns deterministic labels: $y(x_S) = 0$ and $y(x_{ST}) = y(x_T) = 1$. The class $\mathcal{H}$ consists of the constant classifiers, h^0 and h^1, and the collection of groups is $\mathcal{G} = \{S, T\}$ where $S = \{x_S, x_{ST}\}$ and $T = \{x_T, x_{ST}\}$. Finally, the metric specifies that x_S and x_{ST} are identical, and the rest are different:

$$d(x_S, x_{ST}) = 0, \quad , d(x_T, x_{ST}) = 1, \quad d(x_S, x_T) = 1$$

We argue that there is no classifier satisfying the multi-PAC requirement w.r.t $\mathcal{H}$ and $\mathcal{G}$. To see this, we first note that

$$L_{\mathcal{D}_S}(\mathcal{H}) = b/9, \; L_{\mathcal{D}_T}(\mathcal{H}) = 0$$

This is because the optimal classifier for T is h^1, which is perfect for both the accuracy and IF losses; whereas for S, the best classifier is h^0, which is perfect in terms of IF and has a 0-1 loss of $1/9$.

Assume for contradiction that for every $\varepsilon > 0$, there is a classifier that satisfies the multi-PAC requirement. From the perspective of T, the next-best classifier has a loss of $1/2$. So, for multi-PAC with $\varepsilon < 1/2$, it must be the case that $h(x_{ST}) = 1$. On the other hand, from the perspective of S, when $a \gg b$ the next-best classifier has loss $8b/9$. So, for multi-PAC with $\varepsilon < 7b/9$, it must be the case that $h(x_{ST}) = 0$. This means that for this problem instance and for $\varepsilon < \min\{7b/9, 1/2\}$, there is no classifier satisfying the ε-multi-PAC requirement.

□

C Proof of Lemma 4.3 (compatibility $\to$ f-proper)

Let L be any unambiguous and compatible loss. First, we note that by unambiguity, for any singleton distribution $\mathcal{D}$, the loss minimizer is unique. We can therefore denote it by $h^\star_{\mathcal{D}}$.

Next, we make an observation that we will use in the proof: that for any distribution $\mathcal{D}$, the classifier $h : \mathcal{X} \to [0, 1]$ defined by

$$h(x) = \begin{cases} h^\star_{\mathcal{D}_x}(x), & x \in \text{supp}(\mathcal{D}) \\ 0, & \text{otherwise} \end{cases}$$

minimizes the loss $L_{\mathcal{D}}(\cdot)$. That is, we are forming a new classifier h by predicting on an input $x \in \text{supp}(\mathcal{D})$ using the prediction of the classifier that minimizes the loss on the distribution $\mathcal{D}$ restricted to x. The claim is that this classifier is competitive with the best possible loss on the original distribution $\mathcal{D}$.

We claim that the observation follows as a corollary from the compatibility assumption. Note that this is trivially the case for any singleton distribution $\mathcal{D}$ (by unambiguity), so assume for contradiction that there is a non-singleton distribution $\mathcal{D}$ for which the observation does not hold. We define a multi-PAC problem instance as follows. For $x \in \mathcal{X}$, let $g_x = \{x\}$ denote the singleton group that consists only of x. Define $\mathcal{G}_{singletons} = \{g_x : \; x \in \text{supp}(\mathcal{D})\}$ and $\mathcal{H}_{singletons} = \left\{h^\star_{\mathcal{D}_x} : \; x \in \text{supp}(\mathcal{D})\right\}$. Additionally, let $h^\star$ denote some classifier in $\arg\min_h L_{\mathcal{D}}(h)$. Finally,

$$\mathcal{G} = \mathcal{G}_{singletons} \cup \{\text{supp}(\mathcal{D})\}$$
$$\mathcal{H} = \mathcal{H}_{singletons} \cup \{h^\star\}$$

Note that by definition, for every group g in $\mathcal{G}$, $L_{\mathcal{D}_g}(\mathcal{H}) = 0$ (because we specifically included an optimal classifier for every group in $\mathcal{H}$). Multi-PAC for the singleton groups $\mathcal{G}_{singeltons}$ with an arbitrarily small precision ε therefore requires that we predict $h^\star_{\mathcal{D}_x}(x)$ for $x \in \text{supp}(\mathcal{D})$. But by the assumption, the resulting classifier is not optimal for the group $\{\text{supp}(\mathcal{D})\}$, in violation to multi-PAC w.r.t that group. This completes the proof of the observation.

We can now use the observation to directly prove niceness. We will do this by constructing a specific function f and showing that the classifier defined by $h_{\mathcal{D}}(x) = f(x, \mathbf{E}_{\mathcal{D}}[y|x])$ always minimizes the loss $L_{\mathcal{D}}$. Consider

$$f(x, v) = h^\star_{\mathcal{D}_{x,v}}(x)$$

where $\mathcal{D}_{x,v}$ is the singleton distribution supported on x that predicts a label of 1 w.p v, and $h^\star_{\mathcal{D}_{x,v}}$ is the loss minimizer for this distribution (which, by unambiguity, is indeed unique).

We need to prove that f satisfies the requirement in the definition of $f-$proper. Fix some distribution $\mathcal{D}$; we need to prove that

$$h_{\mathcal{D}} \in \arg\min_h L_{\mathcal{D}}(h)$$

where $h_{\mathcal{D}}(x) = f(x, \mathbf{E}_{\mathcal{D}}[y|x])$.

By the observation, the classifier that predicts for $x \in \text{supp}(\mathcal{D})$ using the optimal classifier for $\mathcal{D}_x$ is itself optimal for $\mathcal{D}$. But by construction, $\mathcal{D}_x \equiv \mathcal{D}_{x,\mathbf{E}_{\mathcal{D}}[y|x]}$, so we get:

$$h^\star_{\mathcal{D}_x}(x) = h^\star_{\mathcal{D}_{x,\mathbf{E}_{\mathcal{D}}[y|x]}} = f(x, \mathop{\mathbf{E}}_{\mathcal{D}}[y|x]) = h_{\mathcal{D}}(x)$$

In other words, the classifier that predicts for $x \in \text{supp}(\mathcal{D})$ using the optimal classifier for $\mathcal{D}_x$ is precisely $h_{\mathcal{D}}$. The observation therefore proves $h_{\mathcal{D}} \in \arg\min_h L_{\mathcal{D}}(h)$, as required.

D Proof of Lemma 4.4 (f-proper $\rightarrow$ learnability)

To prove the lemma, we construct a learning algorithm, $\texttt{MultiGroup}_{\texttt{L}}$, and prove that when L is compatible and has the uniform convergence property, the output of this algorithm satisfies the requirements in the definition of multi-group learnability.

The definition of $\texttt{MultiGroup}_{\texttt{L}}$ is given in Algorithm 3. At a high-level, $\texttt{MultiGroup}_{\texttt{L}}$ accepts a class $\mathcal{H}$, collection of subgroups $\mathcal{G}$ and parameters $\varepsilon, \delta, \gamma$, and returns a classifier by invoking a learning algorithm for OI w.r.t an appropriate distinguisher class $\mathcal{A}$. The definition of each distinguisher $A \in \mathcal{A}$ is given separately – see Algorithm 4.

We begin by proving that if L is f-proper, then $h \leftarrow \texttt{MultiGroup}_{\texttt{L,f}}(\varepsilon, \delta, \mathcal{H}, \mathcal{G})$ satisfies the (ε, δ)-multi-group requirement w.r.t $\mathcal{H}$ and $\mathcal{G}$.

Algorithm 3 $\texttt{MultiGroup}_{\texttt{L,f}}(\epsilon, \delta, \gamma, \mathcal{H}, \mathcal{G})$

1: **Parameters:** loss function L, function $f : \mathcal{X} \times [0,1] \to [0,1]$

2: **Input:** accuracy parameter $\epsilon \in (0,1)$, failure probability $\delta \in (0,1)$, minimal subgroups size parameter $\gamma \in (0,1)$, hypothesis class $\mathcal{H}$, collection of subgroups $\mathcal{G}$.

3: **Output:** A classifier h satisfying the (ε, δ)-multi-group guarantee w.r.t $\mathcal{H}$ and $\mathcal{G}$

4: Set $\varepsilon' = \alpha = \varepsilon/4$ and $\delta' = \eta = \tau = \delta/4$.

5: Set $k_{\mathcal{G}} = m_L^{UC}(\varepsilon', \delta', |\mathcal{H}| + 1)$.

6: Set $k = 10 \cdot \frac{1}{\gamma} \cdot \log \frac{1}{\delta'} \cdot k_{\mathcal{G}}$.

7: Let $\mathcal{A} = \left\{ A^{L,f,k}_{g,h,\alpha} \mid g \in \mathcal{G}, h \in \mathcal{H} \right\}$ be a collection of distinguishers, as defined in Algorithm 4.

8: Invoke OI as a sub-routine to learn $\tilde{p} \leftarrow \texttt{OI}(\tau, \eta, \mathcal{A})$.

9: **return** $f(\tilde{p})$

Lemma D.1. *Suppose L is f-proper. Fix a distribution $\mathcal{D}$ over $\mathcal{X} \times Y$, a finite class $\mathcal{H}$, a finite collection of subgroups $\mathcal{G}$ and parameters $\delta, \varepsilon, \gamma \in (0,1)$. Then, w.p at least $1 - \delta$, the predictor $h \leftarrow$ $\texttt{MultiGroup}_{\texttt{L,f}}(\epsilon, \delta, \gamma, \mathcal{H}, \mathcal{G})$ satisfies*

$$\forall g \in \mathcal{G}_\gamma : \quad L_{\mathcal{D}_g}(h) \leq L_{\mathcal{D}_g}(\mathcal{H}) + \varepsilon$$

Proof. We begin by lower-bounding the acceptance probability of each distinguisher $A \in \mathcal{A}$ when it receives samples from the modeled distribution $\tilde{\mathcal{D}}$. Recall that this is the distribution in which outcomes y_i are sampled according to $\text{Ber}(\tilde{p}(x_i))$, where $\tilde{p}$ is the predictor returned by OI.

Claim D.2. *The probability that each $A \triangleq A_{g,h} \in \mathcal{A}$ accepts when given samples from the modeled distribution $\tilde{\mathcal{D}}$ is at least $1 - 2\delta'$:*

$$\Pr_{\{(x_i, y_i)\}_{i=1}^k \sim \tilde{D}^k} [A(\{(x_i, y_i, \tilde{p}(x_i)\}_{i=1}^k) = 1] \geq 1 - 2\delta'$$

To see why this is true, we first note that by construction, the predictor $f(\tilde{p})$ (where $f(\tilde{p})(x) = f(x, \tilde{p}(x))$) coincides with the predictor $h_{\tilde{\mathcal{D}}}$ from the definition of $f-$proper. Thus, invoking the assumption that L is f-proper for the distribution $\mathcal{D}_g$ guarantees that

$$L_{\tilde{\mathcal{D}}_g} f(\tilde{p}) \leq L_{\tilde{\mathcal{D}}_g}(h) \tag{4}$$

To relate this fact to the acceptance criteria of A, which is in terms of a sample S_g from $\tilde{\mathcal{D}}_g$, we need to use the uniform convergence property of L. Recall that the distinguisher operates on $k = 10 \cdot \frac{1}{\gamma} \cdot \log \frac{1}{\delta'} \cdot k_{\mathcal{G}}$ samples from $\tilde{\mathcal{D}}$; this was chosen precisely to guarantee that w.p at least $1 - \delta'$, we have at least $k_{\mathcal{G}}$ samples from $\tilde{\mathcal{D}}_g$ for every group g whose mass is at least γ. Since $k_{\mathcal{G}} = m_L^{UC}(\varepsilon', \delta', |\mathcal{H}| + 1)$, we have a uniform convergence guarantee for the class that includes $\mathcal{H}$ and $\tilde{p}$. That is, we know that w.p at least $1 - \delta'$

$$\left| L_{S_g}(h) - L_{\tilde{\mathcal{D}}_g}(h) \right| \leq \varepsilon', \qquad \left| L_{S_g}(f(\tilde{p})) - L_{\tilde{\mathcal{D}}_g}(f(\tilde{p})) \right| \leq \varepsilon' \tag{5}$$

Algorithm 4 $A^{L,f,k}_{g,h,\alpha}$ (multi-sample Sample-Access OI distinguisher)

Parameters: number of samples $k \in \mathbb{N}$, group $g \subseteq \mathcal{X}$, classifier $h : \mathcal{X} \to [0,1]$, loss function L, function $f : \mathcal{X} \times [0,1] \to [0,1]$
Input: $\{(x_i, y_i, p_i)\}_{i=1}^k$, where $x_i \in \mathcal{X}$, $y_i \in \{0,1\}$ and $p_i \in [0,1]$
Output: A binary output denoting Accept/Reject
$I_g = \{i : x_i \in g\}$
$S_g = \{(x_i, y_i)\}_{i \in I_g}$
Define a predictor f_g as

$$f_g(x) = \begin{cases} f(x_i, p_i) & \exists i \in [k] \text{ such that } x = x_i \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

if $L_{S_g}(f_g) < L_{S_g}(h) + 2\alpha$ **then**
    return 1
end if
return 0

Combining Equations (4) and (5), we have that with probability at least $1 - 2\delta'$ (obtained by union bounding with respect to the two δ' failure probabilities we used above),

$$L_{S_g}(f(\tilde{p})) \leq L_{S_g}(h) + 2\epsilon'$$

Finally, we note that w.r.t S_g, the predictor f_g defined in Equation (3) of Algorithm 4 is the same as $f(\tilde{p})$ – so the above is precisely the acceptance criterion for A in this case. We thus conclude that the acceptance probability of A when it receives samples from the modeled distribution is at least $1 - 2\delta'$, which concludes the proof of the claim.

Next, we argue that a direct corollary is a related lower bound on the acceptance probability of A when it receives samples from the true distribution $\mathcal{D}$.

Claim D.3. *The probability that each $A \triangleq A_{g,h} \in \mathcal{A}$ accepts when given samples from the true distribution $\mathcal{D}$ is at least $1 - (2\delta' + \tau + \eta)$:*

$$\mathop{\mathbf{Pr}}_{\{(x_i,y_i)\}_{i=1}^k \sim \mathcal{D}^k}[A(\{(x_i, y_i, \tilde{p}(x_i)\}_{i=1}^k) = 1] \geq 1 - (2\delta' + \tau + \eta)$$

The claim follows as a direct corollary from the previous claim. By definition, since `OI` is a learning algorithm for OI predictors, $\tilde{p}$ is $(\tau, \mathcal{A})$-OI w.p at least $1 - \eta$. Recall that if $\tilde{p}$ is $(\tau, \mathcal{A})$-OI, then we are guaranteed that the probabilities of each $A \in \mathcal{A}$ accepting on samples from $\tilde{\mathcal{D}}$ and A accepting on samples from $\mathcal{D}$ differ by at most τ:

$$\left| \mathop{\mathbf{Pr}}_{\{(x_i,y_i)\}_{i=1}^k \sim \mathcal{D}^k}[A(\{(x_i, y_i, \tilde{p}(x_i)\}_{i=1}^k) = 1] - \mathop{\mathbf{Pr}}_{\{(x_i,y_i)\}_{i=1}^k \sim \tilde{\mathcal{D}}^k}[A(\{(x_i, y_i, \tilde{p}(x_i)\}_{i=1}^k) = 1] \right| \leq \tau \tag{6}$$

in other words, w.p at least $1 - \eta$ we are guaranteed that $\mathbf{Pr}_{\{(x_i,y_i)\}_{i=1}^k \sim \mathcal{D}^k}[A(\{(x_i, y_i, \tilde{p}(x_i)\}_{i=1}^k) =$

$1] \geq 102\delta' - \tau$. This implies that a lower bound on the acceptance probability in this case is exactly $1 - (2\delta' + \tau + \eta)$, completing the proof of the claim.

Next, we recall that by the definition of the acceptance condition for A, the condition in Equation (6) is the same as saying that w.p at least $1 - (2\delta' + \tau + \eta)$ over the choice of $S_g \sim \mathcal{D}_g$,

$$L_{S_g}(f(\tilde{p})) \leq L_{S_g}(h) + 2\varepsilon'$$

Again using the uniform convergence guarantee from Equation (5), this implies

$$L_{\mathcal{D}_g}(f(\tilde{p})) \leq L_{\mathcal{D}_g}(h) + 4\varepsilon'$$

Plugging in $\varepsilon' = \varepsilon/4$ and $\delta' = \tau = \eta = \delta/4$, we conclude that w.p at least $1 - \delta$,

$$L_{\mathcal{D}_g}(f(\tilde{p})) \leq L_{\mathcal{D}_g}(h) + \varepsilon$$

which is the required. This completes the proof of Lemma D.1. □

To prove multi-group learnability, it remains to bound the sample complexity of Algorithm 3, which we do in the following claim.

Claim D.4. *The sample complexity of Algorithm 3 is*

$$m_L^{gPAC}(\varepsilon, \delta, \gamma, \mathcal{H}, \mathcal{G}) = O\left(\frac{m_{\mathcal{H}}(\varepsilon, \delta) \cdot \log\left(\frac{|\mathcal{H}| \cdot |\mathcal{G}|}{\varepsilon}\right)}{\delta^4 \cdot \gamma}\right)$$

Proof. By the definition of Algorithm 3, the number of samples required is the number of samples required to obtain OI w.r.t $(\tau, \eta, \mathcal{A})$, where $\mathcal{A}$ is a collection of $|\mathcal{H}| \cdot |\mathcal{G}|$ $k-$sample OI distinguishers. By Theorem 2.8, this requires an order of $O(\frac{k \cdot \log(|A|/\eta)}{\tau^4})$ samples. Ignoring constant factors and plugging in the settings of k, η and τ used in Algorithm 3,

$$\begin{aligned} \eta &= O(\delta) \\ \tau &= O(\varepsilon) \\ k &= O\left(\frac{1}{\gamma} \cdot \log \frac{1}{\delta} \cdot m_L^{UC}(\varepsilon, \delta, |\mathcal{H}|)\right) = O\left(\frac{1}{\gamma} \cdot \log \frac{1}{\delta} \cdot m_{\mathcal{H}}(\varepsilon, \delta\right) \end{aligned}$$

we obtain the stated bound. □

Note that when L has the uniform convergence property, this entire expression is indeed polynomial in $1/\varepsilon, 1/\delta, 1/\gamma$ and $\log(|\mathcal{H}|), \log(|\mathcal{G}|)$, as required. Together with the previous lemma, this concludes the proof of Lemma 4.4.