A. Proof of Theorem 2

In what follows, we present proofs of Theorem 2. We start a simple sufficient condition to ensure that a group prefers classifier h to another classifier h'. We will make use of this result to prove Theorem 2, and to design the score function for our decoupling procedure in Appendix B.

Lemma 3 (Generalization of Preferences) *Consider evaluating the true risk of two classifiers h and h' over group z. Given classifiers satisfy $\hat{\Delta}_z(h,h') > 0$, then $\Delta_z(h,h') > 0$ with probability at least $1-\delta$ for any $\delta \in (0,1]$ if*

$$4\mathfrak{R}(\mathcal{H}) + \sqrt{\frac{2\ln\frac{2}{\delta}}{n_z}} \leq \hat{\Delta}_z(h,h'), \tag{5}$$

where $\mathfrak{R}(\mathcal{H})$ is the Rademacher complexity of the hypothesis class $\mathcal{H}$.

Proof 1 *For any group $z \in Z$ and any classifier $h \in \mathcal{H}$ with probability at least $1-\delta/2$, we have that*

$$\left|\hat{R}_z(h) - R_z(h)\right| \leq 2\mathfrak{R}(\mathcal{H}) + \sqrt{\frac{\ln\frac{2}{\delta}}{2n_z}}. \tag{6}$$

The bound in (6) holds for both h and h' with probability at least $1-\delta$. Thus, we know that:

$$\begin{aligned}
R_z(h') - R_z(h) =& (R_z(h') - \hat{R}_z(h')) + (\hat{R}_z(h)) - R_z(h)) + \hat{R}_z(h') - \hat{R}_z(h) \\
\geq& -\left(2\mathfrak{R}(\mathcal{H}) + \sqrt{\frac{\ln\frac{2}{\delta}}{2n_z}}\right) - \left(2\mathfrak{R}(\mathcal{H}) + \sqrt{\frac{\ln\frac{2}{\delta}}{2n_z}}\right) + \hat{\Delta}_z(h,h') \\
=& -\left(4\mathfrak{R}(\mathcal{H}) + \sqrt{\frac{2\ln\frac{2}{\delta}}{n_z}}\right) + \hat{\Delta}_z(h,h') \\
\geq& 0,
\end{aligned}$$

if the condition specified in (5) holds.

We can make use of Lemma 3 to produce the following bounds on the generalization of rationality and envy-freeness. [6]

Corollary 4 (Generalization of Rationality) *Given a set of decoupled classifiers $H_Z = \{\hat{h}_z\}_{z\in Z}$ such that*

$$\hat{\Delta}_z(\hat{h}_z, \hat{h}_0) > 0 \quad \textit{for all} \quad z \in Z,$$

H_Z satisfies rationality with respect the pooled classifier $\hat{h}_0$ with probability at least $1-\delta$, if for all groups $z \in Z$:

$$4\mathfrak{R}(\mathcal{H}) + \sqrt{\frac{2}{n_z}\ln\left(\frac{2|Z|}{\delta}\right)} \leq \hat{\Delta}_z(\hat{h}_z, \hat{h}_0),$$

Corollary 5 (Generalization of Envy-freeness) *Given a set of decoupled classifiers $H_Z = \{\hat{h}_z\}_{z\in Z}$ such that*

$$\hat{\Delta}_z(\hat{h}_z, \hat{h}_{z'}) > 0 \quad \textit{for all} \quad z, z' \in Z,$$

H_Z satisfies envy-freeness with probability at least $1-\delta$ if, for all pairs of groups $z, z' \in Z$:

$$4\mathfrak{R}(\mathcal{H}) + \sqrt{\frac{2}{n_z}\ln\left(\frac{|Z|^2}{\delta}\right)} \leq \hat{\Delta}_z(\hat{h}_z, \hat{h}_{z'}).$$

[6] For the sake of clarity, we will consider a setting where each group is assigned its own classifier so that $a(z) = z$ for each $z \neq z'$. Similar results can be derived for a setting where a single classifier can be assigned to multiple groups (see e.g., Appendix B).

Both results follow from repeated applications of Lemma 2. Specifically:

- Rationality requires that the pairwise preferences in Lemma 2 hold for all groups $z \in Z$. This involves preference conditions for $|Z|$ pairs of classifiers – i.e., one for each distinct pair $\hat{h}_z, \hat{h}_0$ where $z \in Z$. Thus, we can ensure that rationality holds with probability at least $1-\delta$ by applying Lemma 2 with probability at least $1-\frac{\delta}{|Z|}$.
- Envy-freeness requires that the pairwise preferences in Lemma 2 hold for all pairs of groups $z, z' \in Z$. This involves preference conditions on $|Z|(|Z|-1)/2$ pairs of classifiers – i.e., one for each distinct pair $\hat{h}_z, \hat{h}_{z'}$ where $z, z' \in Z$. Since there are $|Z|(|Z|-1)/2$ pairs, and that $|Z|(|Z|-1)/2 \leq |Z|^2/2$, we can ensure that envy-freeness hold with probability at least $1-\delta$ by applying Lemma 2 with probability at least $\frac{\delta}{|Z|^2/2}$.

We are now ready to prove Theorem 2.

Proof 2 (Theorem 2) *Using Massart's Lemma, we have that:*

$$\mathfrak{R}(\mathcal{H}) \leq \sqrt{\frac{2\log|\mathcal{H}|}{n_z}} \tag{7}$$

Combining the bound on $\mathfrak{R}(\mathcal{H})$ in (7) with the bound in Corollary 4, we have that H_Z satisfies rationality with probability at least $1-\delta$, if for all $z \in Z$,

$$n_z \geq \frac{64\ln|\mathcal{H}| + 4\ln\left(\frac{2|Z|}{\delta}\right)}{\hat{\Delta}_z(\hat{h}_z, \hat{h}_0)^2} \tag{8}$$

Likewise, combining the bound on $\mathfrak{R}(\mathcal{H})$ in (7) with the bound in Corollary 5, we have that H_Z satisfies envy-freeness with probability at least $1-\delta$ if for all $z \in Z$,

$$n_z \geq \frac{64\ln|\mathcal{H}| + 4\ln\left(\frac{|Z|^2}{\delta}\right)}{\hat{\Delta}_z(\hat{h}_z, \hat{h}_{z'})^2}. \tag{9}$$

Given the bounds in (8) and (9), we can see that H_Z satisfies both rationality and envy-freeness with probability at least $1-\delta$ if for all $z \in Z$,

$$n_z \geq \max\left\{\frac{64\ln|\mathcal{H}| + 4\ln\left(\frac{2|Z|}{\delta}\right)}{\hat{\Delta}_z(\hat{h}_z, \hat{h}_0)^2}, \frac{64\ln|\mathcal{H}| + 4\ln\left(\frac{|Z|^2}{\delta}\right)}{\hat{\Delta}_z(\hat{h}_z, \hat{h}_{z'})^2}\right\} \tag{10}$$

Thus, the bound in Theorem 2 holds so long as we can show that:

$$\frac{64\ln|\mathcal{H}| + 4\ln(\frac{2|Z|^2}{\delta})}{\hat{\epsilon}_z^2} \geq \max\left\{\frac{64\ln|\mathcal{H}| + 4\ln\left(\frac{2|Z|}{\delta}\right)}{\hat{\Delta}_z(\hat{h}_z, \hat{h}_0)^2}, \frac{64\ln|\mathcal{H}| + 4\ln\left(\frac{|Z|^2}{\delta}\right)}{\hat{\Delta}_z(\hat{h}_z, \hat{h}_{z'})^2}\right\} \tag{11}$$

This follows given that we have defined $\hat{\epsilon}_z = \min\left(\hat{\Delta}_z(\hat{h}_z, \hat{h}_0), \min_{z' \in Z/\{z\}} \hat{\Delta}_z(\hat{h}_z, \hat{h}_{z'})\right)$, and that the inequality $4\ln\left(\frac{|Z|^2}{\delta}\right) \geq 4\ln\left(\frac{2|Z|}{\delta}\right)$ holds whenever $|Z| \geq 2$.

B. Score Function

In what follows, we formally derive the score function that we present in Section 4. The score function ensures that our procedure grows a tree in a way that is aligned with the goal of minimizing the risk of a preference violation.

We wish to bound the probability that H_T violates rationality or envy-freeness as follows:

$$\mathbb{P}\left(\begin{array}{c}H_T \text{ violates} \\ \text{rationality or} \\ \text{envy-freeness}\end{array}\right) \leq \mathsf{ViolationScore}(T) = \sum_{z \in Z} 4\exp\left(-\tfrac{n_z}{2} \cdot \hat{\Delta}_z(\hat{h}_z, \hat{h}_0)^2\right) + \sum_{z \in Z} \sum_{\substack{z' \in Z \\ a(z') \neq a(z)}} 4\exp\left(-\tfrac{n_z}{2} \cdot \hat{\Delta}_z(\hat{h}_z, \hat{h}_{z'})^2\right)$$

We restrict our attention to cases where $\hat{\Delta}_z(z, z') > 0$ since our training procedure ensures that $\hat{\Delta}_z(z, z') \geq 0$, and since $\hat{\Delta}_z(z, z') = 0$ implies indifference (i.e., it does not imply a preference violation).

Given a pair groups $z, z' \in Z$ such that $a(z) \neq a(z')$, we denote an event where group z prefers the classifier assigned to group z' as $\mathcal{E}_{z \to z'}$.. We will bound the probability of $\mathcal{E}_{z \to z'}$ in terms of the following event:

$$\mathcal{E}_{z,z'} = \left\{ |R_z(\hat{h}_z) - \hat{R}_z(\hat{h}_z)| \geq \frac{\hat{\Delta}_z(\hat{h}_z, \hat{h}_{z'})}{2} \right\} \cup \left\{ |R_z(\hat{h}_{z'}) - \hat{R}_z(\hat{h}_{z'})| \geq \frac{\hat{\Delta}_z(\hat{h}_z, \hat{h}_{z'})}{2} \right\}$$

We observe that $\mathcal{E}_{z \to z'} \subseteq \mathcal{E}_{z,z'}$. We proceed to present a proof by contradiction. Suppose that $\mathcal{E}_{z \to z'} \not\subseteq \mathcal{E}_{z,z'}$, this means that there must exist an event $\omega \in \mathcal{E}_{z \to z'}$ such that $\omega \notin \mathcal{E}_{z,z'}$. The fact that $\omega \notin \mathcal{E}_{z,z'}$ implies that both of the following inequalities must hold:

$$|R_z(\hat{h}_z) - \hat{R}_z(\hat{h}_z)| < \frac{\hat{\Delta}_z(\hat{h}_z, \hat{h}_{z'})}{2}$$
$$|R_z(\hat{h}_{z'}) - \hat{R}_z(\hat{h}_{z'})| < \frac{\hat{\Delta}_z(\hat{h}_z, \hat{h}_{z'})}{2}$$

This implies:

$$\begin{aligned} R_z(\hat{h}_z) - R_z(\hat{h}_{z'}) &= (R_z(\hat{h}_z) - \hat{R}_z(\hat{h}_z)) + (\hat{R}_z(\hat{h}_z) - \hat{R}_z(\hat{h}_{z'})) + (\hat{R}_z(\hat{h}_{z'}) - R_z(\hat{h}_{z'})) \\ &< \frac{\hat{\Delta}_z(\hat{h}_z, \hat{h}_{z'})}{2} - \hat{\Delta}_z(\hat{h}_z, \hat{h}_{z'}) + \frac{\hat{\Delta}_z(\hat{h}_z, \hat{h}_{z'})}{2} \\ &= 0. \end{aligned}$$

Thus, we have shown that z does not envy z', which contradicts the fact that $\omega \in \mathcal{E}_{z \to z'}$.

Having shown that $\mathcal{E}_{z \to z'} \subseteq \mathcal{E}_{z,z'}$, we can bound the probability of an envy-freeness violation as follows:

$$\begin{aligned} \mathbb{P}\left(\cup_{z,z'} \mathcal{E}_{z \to z'}\right) &\leq \mathbb{P}\left(\cup_{z,z'} \mathcal{E}_{z,z'}\right) && (12) \\ &\leq \sum_{\substack{z,z' \in Z \\ a(z) \neq a(z')}} \mathbb{P}\left(\mathcal{E}_{z,z'}\right) && (13) \\ &\leq \sum_{\substack{z,z' \in Z \\ a(z) \neq a(z')}} \mathbb{P}\left(|R_z(\hat{h}_z) - \hat{R}_z(\hat{h}_z)| \geq \tfrac{\hat{\Delta}_z(\hat{h}_z, \hat{h}_{z'})}{2}\right) + \mathbb{P}\left(|R_z(\hat{h}_{z'}) - \hat{R}_z(\hat{h}_{z'})| \geq \tfrac{\hat{\Delta}_z(\hat{h}_z, \hat{h}_{z'})}{2}\right) && (14) \\ &\leq \sum_{\substack{z,z' \in Z \\ a(z) \neq a(z')}} 2\exp\left(-2n_z \left(\frac{\hat{\Delta}_z(\hat{h}_z, \hat{h}_{z'})}{2}\right)^2\right) + 2\exp\left(-2n_z \left(\frac{\hat{\Delta}_z(\hat{h}_z, \hat{h}_{z'})}{2}\right)^2\right) && (15) \\ &= \sum_{\substack{z,z' \in Z \\ a(z) \neq a(z')}} 4\exp\left(-\frac{n_z}{2} \cdot \hat{\Delta}_z(\hat{h}_z, \hat{h}_{z'})^2\right) && (16) \end{aligned}$$

Here: (12) follows from the fact that $\mathcal{E}_{z \to z'} \subseteq \mathcal{E}_{z,z'}$; (13) and (14) follow from the union bound; and (15) follows from inverting the bound.

We bound the probability of a rationality violation in a similar manner. We first define the following event for each $z \in Z$:

$$\mathcal{E}_{z,0} = \left\{ |R_z(\hat{h}_z) - \hat{R}_z(\hat{h}_z)| \geq \frac{\hat{\Delta}_z(\hat{h}_z, \hat{h}_0)}{2} \right\} \cup \left\{ |R_z(\hat{h}_0) - \hat{R}_z(\hat{h}_0)| \geq \frac{\hat{\Delta}_z(\hat{h}_z, \hat{h}_0)}{2} \right\}$$

We note that $\mathcal{E}_{z \to 0} \subseteq \mathcal{E}_{z,0}$, which can be shown by deriving an analogous contradiction to the one derived for envy-freeness. With this result, we can bound the probability of an rationality violation as follows:

$$\begin{aligned}
\mathbb{P}\left(\cup_{z \in Z} \mathcal{E}_{z \to 0}\right) &\leq \mathbb{P}\left(\cup_z \mathcal{E}_{z,0}\right) && (17)\\
&\leq \sum_{z \in Z} \mathbb{P}\left(\mathcal{E}_{z,0}\right) && (18)\\
&\leq \sum_{z \in Z} \mathbb{P}\left((|R_z(\hat{h}_z) - \hat{R}_z(\hat{h}_z)| \geq \tfrac{\hat{\Delta}_z(\hat{h}_z, \hat{h}_0)}{2} \right) + \mathbb{P}\left(|R_z(\hat{h}_0) - \hat{R}_z(\hat{h}_0)| \geq \tfrac{\hat{\Delta}_z(\hat{h}_z, \hat{h}_0)}{2} \right) && (19)\\
&\leq \sum_{z \in Z} 2\exp\left(-2n_z \left(\frac{\hat{\Delta}_z(\hat{h}_z, \hat{h}_0)}{2} \right)^2 \right) + 2\exp\left(-2n_z \left(\frac{\hat{\Delta}_z(\hat{h}_z, \hat{h}_0)}{2} \right)^2 \right) && (20)\\
&= \sum_{z \in Z} 4\exp\left(-\frac{n_z}{2} \cdot \hat{\Delta}_z(\hat{h}_z, \hat{h}_0)^2 \right) && (21)
\end{aligned}$$

Here: (17) follows from the fact that $\mathcal{E}_{z \to 0} \subseteq \mathcal{E}_{z,0}$; (18) and (19) follow from the union bound; and (20) follows from inverting the bound. Our final expression for the score function is obtained by combining the terms in (16) and (21).