

Fair Wrapping for Black-box Predictions

Alexander Soen¹ Ibrahim Alabdulmohsin² Sanmi Koyejo³ Yishay Mansour^{2,4} Nyalleng Moorosi²
 Richard Nock^{2,1} Ke Sun^{5,1} Lexing Xie¹

Abstract

We introduce a new family of techniques to post-process (“wrap”) a black-box classifier in order to reduce its bias. Our technique builds on the recent analysis of improper loss functions whose optimisation can correct any *twist* in prediction, unfairness being treated as a twist. In the post-processing, we learn a wrapper function which we define as an α -tree, which modifies the prediction. We provide two generic boosting algorithms to learn α -trees. We show that our modification has appealing properties in terms of composition of α -trees, generalization, interpretability, and KL divergence between modified and original predictions. We exemplify the use of our technique in three fairness notions: conditional value at risk, equality of opportunity, and statistical parity; and provide experiments on several readily available datasets.

1. Introduction

Machine Learning has seen a dramatic increase of its impact over the past decade – enough that it has become a priority to control not just the accuracy, but also the bias of models’ outputs (Alabdulmohsin & Lucic, 2021; Hardt et al., 2016; Zafar et al., 2019). If we take into account the numerous fairness targets and models that have been defined and / or refined (Mehrabi et al., 2022) – sometimes excluding each other or in tension with accuracy, and factor in the energy and CO2 footprint of the domain (Martineau, 2020; Strubell et al., 2019), then the combinatorics of training accurate *and* fair models look non trivial. A suitable trend in the field tries to “decouple” both constraints as it seeks to post-process the outputs of *pretrained* (accurate) models to achieve a more fair output (Zafar et al., 2019). Post-processing may be the only option if e.g. we have no access to the model’s training

¹Australian National University ²Google Research
³University of Illinois Urbana-Champaign ⁴Tel Aviv University
⁵Data61/CSIRO. Correspondence to: Alexander Soen <alexander.soen@anu.edu.au>.

data / algorithm / hyperparameters (etc.).

In this cluster, three subtrends emerge: learning a new fair model close to the black-box, tweaking the output subject to fairness constraints, and exploiting sets of classifiers (see Section 2). When the task is class probability estimation (Reid & Williamson, 2011), the estimated black-box is an accurate but potentially unfair posterior $\eta_u : \mathcal{X} \rightarrow [0, 1]$ which is neither opened nor trained further. The goal is then to learn a fair posterior η_f from it. A number of relevant *desiderata* can be considered for post-processing, including: **(a)** flexibility of training to substantially different fairness metrics, **(b)** guarantees in terms of proximity of η_f to η_u if fairness requirements are lightweight, **(c)** strong algorithmic guarantees to obtain η_f , **(d)** explanability properties in the mapping from η_f to η_u , **(e)** composability properties if e.g. η_f was later treated as a black-box to be post-processed using a different fairness notion, **(f)** generalisation properties (η_u).

Our contribution explores a new solution to the post-processing problem, borne out of the analysis of loss functions for class probability estimation that are *improper* – thus for which Bayes rule, eventually unfair, is not a minimizer. Such methods are formally able to correct *any* twist in prediction (Nock et al., 2021), unfairness being treated as one. We use the α -loss (Arimoto, 1971; Liao et al., 2018), known to have such a property (Nock et al., 2021). The correction is then a function $\alpha : \mathcal{X} \rightarrow \mathbb{R}$ to be learned. The approach also addresses the goals **(a-f)** from three standpoints: analytical, representation, and algorithmic. From an analytical standpoint, we show that the correction yields convenient divergence bounds between η_f and η_u , a convenient form for the Rademacher complexity of the class of η_f , and a straightforward composability property. Representation-wise, the corrections we learn are easy-to-interpret tree-shaped functions that we define as α -*trees*. Algorithmically speaking, we provide two formal boosting algorithms to learn α -trees, building upon a seminal result on boosting decision trees (Kearns & Mansour, 1996). We exemplify the algorithm on three fairness metrics: conditional value at risk, equality of opportunity, and statistical parity. Experiments are provided against various baselines on readily available datasets.

2. Related work

Post-processing models to achieve fairness is one out of three different categories of approaches to tackle the ML + fairness challenge (Zafar et al., 2019, Section 6.2). We can segment this cluster further in three subsets: (I) approaches learning a new model with two constraints: being close to the pretrained model and being fair (Kim et al., 2019; Petersen et al., 2021; Wei et al., 2020; Yang et al., 2020); (II) approaches biasing the output of the pretrained model at classification time, modifying observations to receive a more fair outcome (Alabdulmohsin & Lucic, 2021; Hardt et al., 2016; Lohia et al., 2019; Menon & Williamson, 2018; Woodworth et al., 2017; Yang et al., 2020); and a last one (III) consisting of exploiting sets of models to achieve fairness (Dwork et al., 2018). None of those approaches formulates substantial guarantees on all of points (a-f) in the introduction. Some bring contributions applicable to more than two fairness notions (Corbett-Davies et al., 2017; Wei et al., 2020; Dwork et al., 2018; Yang et al., 2020) (a), two of which provide the convenience of analytic conditions on new fairness notions to fit in the approach (Wei et al., 2020; Dwork et al., 2018), but for all of them the algorithmic price-tag is unclear (Corbett-Davies et al., 2017; Dwork et al., 2018) or heavily depends on convex optimisation routines (Wei et al., 2020). Alabdulmohsin & Lucic (2021); Yang et al. (2020) provide strong guarantees regarding (b), in terms of *consistency and generalization*. To our knowledge, no previous approach has exploited the α -loss function (an improper loss) and its properties to correct prediction unfairness.

3. Losses for class probability estimation

Binary experiments and measures Let \mathcal{X} be a domain of observations, $\mathcal{Y} \doteq \{-1, 1\}$ labels and S is a sensitive attribute in \mathcal{X} . We assume that the modalities of S induce a partition of \mathcal{X} . (\mathcal{X}, P) and (\mathcal{X}, N) are measure spaces for “positive” and “negative” observations respectively (leaving implicit the σ -algebra, assumed to be the same everywhere). $(\mathcal{X} \times \{-1, 1\}, D)$ is the group’s product measure space of labeled examples following the (group’s supervised) *binary task* (π, P, N) (Reid & Williamson, 2011, Section 4), $\pi \doteq \mathbb{P}[Y = 1]$ being the prior. (\mathcal{X}, M) is a *mixture* measure space defined by $M \doteq \pi \cdot P + (1 - \pi) \cdot N$. As is often assumed in ML, sampling is i.i.d.; we make no notational distinction between empirical and true measure to simplify exposure as most of our results would apply for both. Distinction shall be made when discussing generalisation. Finally, $\eta \in [0, 1]^{\mathcal{X}}$ denotes a posterior that computes (an estimate of) $\mathbb{P}[Y = 1 | X]$. In this paper, blue-boxed text is used to single out algorithmic nuggets with lightweight description, e.g.,

given a mixture M and posterior η , we sample according to the product measure on $\mathcal{X} \times \{-1, 1\}$ by sampling an observation (mixture) and then the class (posterior).

Bayes posterior admits the expression $\eta^* = \pi \cdot dP/dM$ (Reid & Williamson, 2011), and is optimal for *proper* losses.

Losses for class-probability estimation a *loss for class probability estimation*, $\ell : \mathcal{Y} \times [0, 1] \rightarrow \mathbb{R}$, is expressed as

$$\ell(y, u) \doteq \llbracket y = 1 \rrbracket \cdot \ell_1(u) + \llbracket y = -1 \rrbracket \cdot \ell_{-1}(u), \quad (1)$$

where $\llbracket \cdot \rrbracket$ is Iverson’s bracket (Knuth, 1992). Functions ℓ_1, ℓ_{-1} are called *partial losses*. A loss is *symmetric* when $\ell_1(u) = \ell_{-1}(1 - u), \forall u \in [0, 1]$ (Nock & Nielsen, 2008) and *differentiable* when both partial losses are differentiable. A loss is *fair*¹ when $\ell_1(1) = \ell_{-1}(0) = 0$ and $0 = \min \ell_1 = \min \ell_{-1}$ (Reid & Williamson, 2011). The α -loss is a differentiable, symmetric and fair loss defined by the partial losses (Liao et al., 2018):

$$\ell_1^{(\alpha)}(u) \doteq \frac{\alpha \cdot (1 - u)^{\frac{\alpha-1}{\alpha}}}{\alpha - 1}, \quad \ell_{-1}^{(\alpha)}(u) \doteq \ell_1^{(\alpha)}(1 - u), \quad (2)$$

for $\alpha \geq 0$ and $\ell_1^{(\alpha)}(u) \doteq \ell_{-1}^{(-\alpha)}(u) = \ell_1^{(-\alpha)}(1 - u)$ for $\alpha < 0$. As $\alpha \rightarrow 1$, the α -loss converges to log-loss ($\ell_1^{\text{LOG}}(u) \doteq -\log(u)$) and as $\alpha \rightarrow \infty$, the α -loss converges to the 0/1-loss ($\ell_1^{0/1}(u) \doteq \llbracket u < 1/2 \rrbracket$). The pointwise conditional risk of estimator $\hat{\eta} \in [0, 1]$ with respect to ground (unknown) truth $\eta \in [0, 1]$ is $L(\hat{\eta}, \eta) \doteq \mathbb{E}_{Y \sim B(\eta)} [\ell(Y, \hat{\eta})]$, i.e.:

$$L(\hat{\eta}, \eta) = \eta \cdot \ell_1(\hat{\eta}) + (1 - \eta) \cdot \ell_{-1}(\hat{\eta}), \quad (3)$$

where $B(\cdot)$ denotes a Bernoulli for picking label $Y = 1$.

Properness and the Bayes tilted estimate The Bayes tilted estimate of loss ℓ (Nock et al., 2021),

$$t_{\ell}(\eta) \doteq \arg \inf_{u \in [0, 1]} L(u, \eta), \quad (4)$$

is the pointwise minimizer(s) of (3). When ℓ is **proper**, $\eta \in t_{\ell}(\eta)$ and when **strictly proper**, $\{\eta\} = t_{\ell}(\eta)$. $\alpha \in \{1, \infty\}$ -loss is proper and $\alpha = 1$ -loss is strictly proper. The Bayes tilted estimate of the ($\alpha \geq 0$)-loss is (Nock et al., 2021):

$$t_{\ell}(\eta) = \begin{cases} [0, 1] & \text{if } (\alpha = 0) \vee (\alpha = \infty \wedge \eta = \frac{1}{2}) \\ \left\{ \frac{\eta^{\alpha}}{\eta^{\alpha} + (1 - \eta)^{\alpha}} \right\} & \text{otherwise (taking limit if } \alpha = \infty) \end{cases} \quad (5)$$

Notably, for example when $\alpha = 1$, $t_{\ell}(\eta) = \{\eta\}$.

Population loss A model that fits a posterior η is trained to minimize a population version of (3), called *risk*, which integrates the Bayes tilted estimate, as:

$$L(\eta; M, \eta^*) \doteq \mathbb{E}_{X \sim M} [L(t_{\ell}(\eta(X)), \eta^*(X))]. \quad (6)$$

If the loss is proper, such as for the log- or square-losses, we retrieve the classical expression $L(\eta; M, \eta^*) = \mathbb{E}_{X \sim M} [L(\eta(X), \eta^*(X))]$. The tilted population loss (6) is the key to our approach to fairness correction.

¹“fair” as defined is related but distinct from the algorithmic fairness goals and metrics in this work.

4. Making black-boxes fair with guarantees

The overall recipe We have a black-box posterior η_u , accurate but eventually not fair. We wish to learn a fair posterior, η_f , which is a function of η_u , but we cannot “open” nor train further the black-box. Our task is thus to design a mapping

$$\eta_u \mapsto \eta_f \quad (7)$$

with desirable analytical, representation, and algorithmic properties as summarized in constraints **(a-f)** (Section 1). η_f integrates components that need to be learned from data to achieve fairness, as such, we need an algorithm, say \mathcal{A} :

\mathcal{A} learns η_f by minimizing a risk as:

$$\eta_f \stackrel{\mathcal{A}}{\leftarrow} \min L(\eta_f; M, \eta_t), \quad (8)$$

where the loss ℓ , the mixture M , and “target” posterior η_t are *designed* to achieve the fairness guarantee.

To constraints **(a-f)** we add a last invertibility condition, **(g)**, which states that it has to be simple to retrieve η_u from η_f used as a black-box and components learned to create η_f .

Our implementation of mapping (7) A simple choice for η_f consists in picking the Bayes tilted estimate of a twist-proper loss. We choose the α -loss so (5) gives our (7):

$$\eta_f(\mathbf{x}) \doteq \frac{\eta_u(\mathbf{x})^{\alpha(\mathbf{x})}}{\eta_u(\mathbf{x})^{\alpha(\mathbf{x})} + (1 - \eta_u(\mathbf{x}))^{\alpha(\mathbf{x})}} \in [0, 1] \quad (9)$$

where $\alpha \in \mathbb{R}_*^\mathcal{X}$ thus defines the components that need to be learned to *wrap* η_u . Because it is an α -loss and is also strictly proper, we pick the log-loss ($\alpha = 1$) as the loss of choice for (8): it follows that minimizing (8) yields some form of convergence (to be made precise later) for η_f towards η_t , with the desirable property that the log-loss being strictly proper, as $\sup |\alpha - 1| \rightarrow 0$, we get $\eta_f \rightarrow \eta_u$. We can make precise this latter convergence in our case, in the context of **(b)** above. Since posteriors η_u, η_f have the same support, a good divergence measure is an f -divergence, and we pick the KL divergence for its prominence in information theory and geometry (Amari & Nagaoka, 2000):

$$KL(\eta_u, \eta_f; M) = \mathbb{E}_{(\mathbf{X}, \mathbf{Y}) \sim D_u} \left[\log \left(\frac{dD_u((\mathbf{X}, \mathbf{Y}))}{dD_f((\mathbf{X}, \mathbf{Y}))} \right) \right],$$

where D_u, D_f are the product measures defined from M and their respective posteriors (Section 3).

Theorem 1. For any function $\alpha : \mathcal{X} \rightarrow \mathbb{R}$, any black-box posterior η_u , and any integer $K \geq 2$, using (9) yields the following bound on the KL divergence:

$$\begin{aligned} & KL(\eta_u, \eta_f; M) \\ & \leq \mathbb{E}_{\mathbf{X} \sim M} \left[\sum_{k=2}^K \frac{\eta_u(\mathbf{X})(1 - \eta_u(\mathbf{X}))f^k(\alpha(\mathbf{X}), \eta_u(\mathbf{X}))}{k(k-1)} \right] \\ & \quad + o(\mathbb{E}_{\mathbf{X} \sim M} [(\alpha(\mathbf{X}) - 1)^K]), \end{aligned} \quad (10)$$

where we have used function $f : \mathbb{R} \times [0, 1] \rightarrow \mathbb{R}$ defined as:

$$f(z, u) \doteq |\log((1-u)/u) \cdot (z-1)|. \quad (11)$$

Proof in SI, Section I. Here are two examples of concrete upperbounds on $KL(\eta_u, \eta_f; M)$. In setting **(S1)**, correction is all the smaller as the black-box posterior is far from 1/2:

(S1) $f(\alpha(\mathbf{x}), \eta_u(\mathbf{x})) \leq 1$ (a.s.), f being in (11).

To present the second setting, we need to introduce an Assumption that will be important to analyse our algorithms.

Assumption 1. The black-box prediction is bounded away from the extremes: there exists $B > 0$ such that

$$Im(\eta_u) \subseteq \mathbb{I} \doteq \left[\frac{1}{1 + \exp(B)}, \frac{1}{1 + \exp(-B)} \right] \text{ (a.s.)}. \quad (12)$$

Compliance with Assumption 1 can be done by clipping the black-box’ output with user-fixed B or making sure it is calibrated and then finding B . We now present setting **(S2)**.

(S2) Assumption 1 holds for some $0 < B \leq 3$ and function α satisfies $|\alpha(\mathbf{x}) - 1| \leq 1/B$ (a.s.).

Corollary 2. Under setting **(S2)**, we have the upperbound

$$KL(\eta_u, \eta_f; M) \leq \frac{\pi^2}{6(2 + \exp(B) + \exp(-B))}, \quad (13)$$

and under settings **(S1)**, we have the weaker guarantee $KL(\eta_u, \eta_f; M) \leq \pi^2/24 \approx 0.41$.

The proof of the Corollary is in SI, Section II and includes a graphical view of the domain of f complying with **(S1)**. To get a glimpse into the quality of the bounds, fix $B = 3$ for **(S2)**. In this case, we want $\alpha(\cdot) \in [2/3, 4/3]$ (a.s.), which is a reasonable sized interval centered at 1, the clamped black-box posterior’s interval is approximately $[0.04, 0.96]$, which is quite flexible, and the distortion to the black-box caused by α is upperbounded as $KL(\eta_u, \eta_f; M) \leq 7.5E-2$.

Overview of (8) To make the high-level process precise, we thus look after the minimisation of

$$L(\eta_f; M, \eta_t) \doteq \mathbb{E}_{\mathbf{X} \sim M} \left[\frac{\eta_t(\mathbf{X}) \cdot -\log \eta_f(\mathbf{X})}{+(1 - \eta_t(\mathbf{X})) \cdot -\log(1 - \eta_f(\mathbf{X}))} \right], \quad (14)$$

with η_f in (9). (14) has a simple and popular alternative expression: plugging (9) in (14) and simplifying using the corresponding product measure $(\mathcal{X} \times \mathcal{Y}, D_t)$ (we use M, η_t to craft D_t), yields the expression based on the logistic loss:

$$\begin{aligned} & L(\eta_f; M, \eta_t) \doteq \\ & \mathbb{E}_{(\mathbf{X}, \mathbf{Y}) \sim D_t} \left[\log \left(1 + \exp \left(-\mathbf{Y} \alpha(\mathbf{X}) \log \left(\frac{\eta_u(\mathbf{X})}{1 - \eta_u(\mathbf{X})} \right) \right) \right) \right]. \end{aligned}$$

Remarks We can make two key remarks related to points **(e,f)**. The logistic loss being Lipschitz, a relevant capacity

notion to assess the uniform convergence of this risk for the whole wrapped model is the Rademacher complexity of the following set of functions (Bartlett & Mendelson, 2002):

$$\mathcal{H}_t \doteq \left\{ \alpha(\mathbf{x}) \cdot \log \left(\frac{\eta_u(\mathbf{x})}{1 - \eta_u(\mathbf{x})} \right), \forall (\alpha, \eta_u) \right\}, \quad (15)$$

where we assume known the set of functions from which η_u was trained. The analytical form in (9) also brings the following easy-to-check composability property.

Lemma 1. *The composition of any two wrapping transformations $\eta_u \xrightarrow{\alpha} \eta_f \xrightarrow{\alpha'} \eta'_f$ following (9) is equivalent to the single transformation $\eta_u \xrightarrow{\alpha \cdot \alpha'} \eta'_f$.*

This also gives us compliance with the invertibility condition (**g**) by wrapping η_f using $\alpha' = 1/\alpha$. In the following Section, we investigate the functions we consider for α and how to train them with boosting-compliant convergence.

5. Alpha-trees and how to grow them

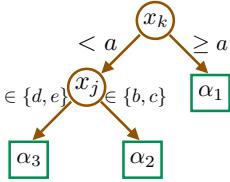


Figure 1. An example of an α -tree. Just like in a decision tree, variables of many different types can be used for the splits.

We now focus on three main components with key focus on constraint (**c**) and additional leverage on (**d,f**): the models we use for function α , a convenient upperbound on (14), and finally a fast, boosting-compliant algorithm to minimise this upperbound when learning our models. At this stage, both the mixture M and η_t remain unspecified as they will depend on the fairness objective tackled.

Alpha-trees We first define the functions we use for α .

Definition 1. *An α -tree is a rooted, directed binary tree, with internal nodes labeled with observation variables. Outgoing arcs are labeled with tests over the nodes' variable. Leaves are real valued. $\Lambda(\Upsilon)$ is the leafset of α -tree Υ .*

Figure 1 presents an example of α -tree. Just like a decision tree, an α -tree recursively splits the whole domain \mathcal{X} , the key difference being that leaf predictions are correction to the unfair posterior, not labels.

General induction of an α -tree Assumption 1 is instrumental for this part. Denote $\iota(u) \doteq \log(u/(1-u))$ the logit of $u \in [0, 1]$ and $\tilde{\iota}(u) \doteq \iota(u)/B$ a normalization which satisfies $\tilde{\iota}(\mathbb{I}) = [-1, 1]$ (12). Also, we define the edge of the normalized logit given mixture M and target posterior η_t ,

$$e(M, \eta_t) \doteq \mathbb{E}_{(\mathbf{X}, \mathbf{Y}) \sim D_t} [\mathbf{Y} \tilde{\iota}(\eta_u(\mathbf{X}))], \quad (16)$$

Algorithm 1 TOPDOWN (M, η_t, Υ_0, B)

Input mixture M , posterior η_t , α -tree Υ_0 , $B \in \mathbb{R}_{+*}$;

Step 1: $\Upsilon \leftarrow \Upsilon_0$;

Step 2 : **while** stopping condition not met **do**

Step 2.1 : pick leaf $\lambda^* \in \Lambda(\Upsilon)$

Step 2.2 : $h^* \leftarrow \arg \min_{h \in \mathcal{H}} H(\Upsilon(\lambda^*, h); M, \eta_t)$;

Step 2.3 : $\Upsilon \leftarrow \Upsilon(\lambda^*, h^*)$; // split using h^* at λ^*

Step 3 : label leaves:

$$\Upsilon(\lambda) \doteq \tilde{\iota} \left(\frac{1 + e(M_\lambda, \eta_t)}{2} \right), \forall \lambda \in \Lambda(\Upsilon), \quad (18)$$

Output Υ ;

which satisfies $e(M, \eta_t) \in [-1, 1]$ when Assumption 1 is satisfied. The blueprint of our algorithm, TOPDOWN, is given in Algorithm 1, where \mathcal{H} denote a function set for splits, each element of which is a function from \mathcal{X} to $\{-1, 1\}$, +1 indicating the observation follows the right arc at the split. TOPDOWN is similar at a high level to classical top-down decision trees induction algorithms (Kearns & Mansour, 1996, Figure 1). A notable low-level difference is the initial α -tree provided, Υ_0 ; using the decision tree induction blueprint would require Υ_0 to be a 1-node tree. Another difference is the loss used for selecting splits. We now present this criterion, letting $H(q) \doteq -q \log(q) - (1-q) \log(1-q)$.

Definition 2. *Given α -tree Υ with leafset Λ , when Assumption 1 is satisfied, the **entropy** of Υ is denoted*

$$H(\Upsilon; M, \eta_t) \doteq \mathbb{E}_{\lambda \sim M_{\Lambda(\Upsilon)}} [H_1(\lambda; M, \eta_t)], \quad (17)$$

where $H_1(\lambda; M, \eta_t) \doteq H((1 + e(M_\lambda, \eta_t))/2)$, M_λ is M conditioned to leaf $\lambda \in \Lambda$ and $M_{\Lambda(\Upsilon)}$ is measure induced on $\Lambda(\Upsilon)$ by the leaves' weights on M .

TOPDOWN is a boosting algorithm To show that TOPDOWN is a boosting algorithm, we need a *Weak Hypothesis Assumption*, which postulates informally that each chosen split brings a small edge over random splits for a tailored distribution that *locally* makes the problem “harder”.

Definition 3. *Let $\lambda \in \Lambda(\Upsilon)$ and $D_{t,\lambda}$ be the product measure on $\mathcal{X} \times \mathcal{Y}$ conditioned on λ . The **balanced** product measure $D'_{t,\lambda}$ at leaf λ is defined as ($z \doteq (\mathbf{x}, \mathbf{y})$ for short):*

$$D'_{t,\lambda}(z) \doteq \frac{1 - e(M_\lambda, \eta_t) \cdot y \tilde{\iota}(\eta_u(\mathbf{x}))}{1 - e(M_\lambda, \eta_t)^2} \cdot D_{t,\lambda}(z). \quad (19)$$

We check that $\int_{\lambda} dD'_{t,\lambda} = 1$ because of the definition of $e(M_\lambda, \eta_t)$ (16). Our balanced distribution was named after Kearns & Mansour (1996)'s: ours indeed generalises theirs. Consider the “fairness-free case” as the replacement of $\tilde{\iota}(.)$ by constant 1 in (16) and replacing η_t by η_u . This yields $e(M_\lambda, \eta_u) = \mathbb{E}_{(\mathbf{X}, \mathbf{Y}) \sim D_\lambda} [\mathbf{Y}] = 2q_\lambda - 1$, with q_λ the local proportion of positive examples in λ . The denominator

of (19) becomes $4q_\lambda(1 - q_\lambda)$, which after simplification with the numerator, depending on y , yields a factor on the right hand side of $1/(2q_\lambda)$ for positive examples and $1/(2(1-q_\lambda))$ for negative examples and brings the balanced distribution in Kearns & Mansour (1996). We now state our Weak Hypothesis Assumption (WHA).

Assumption 2. Let $h : \mathcal{X} \rightarrow \{-1, 1\}$ be the function splitting leaf λ , and let $\gamma > 0$. We say that h γ -witnesses the Weak Hypothesis Assumption (WHA) at λ iff

$$(i) \quad \left| \mathbb{E}_{(X,Y) \sim D'_{t,\lambda}} [Y\tilde{\iota}(\eta_u(X)) \cdot h(X)] \right| \geq \gamma, \\ (ii) \quad e(M_\lambda, \eta_t) \cdot \mathbb{E}_{(X,Y) \sim D_{t,\lambda}} [(1 - \tilde{\iota}^2(\eta_u(X))) \cdot h(X)] \leq 0.$$

An important remark is in place: if $\tilde{\iota}(\eta_u(.)) \in \{-1, 1\}$, the second part (ii) vanishes and our WHA looks a lot more like the conventional one (Kearns & Mansour, 1996); in fact, in the fairness-free case (see above), (i) reduces to the weak hypothesis assumption of Kearns & Mansour (1996). In the most general case, our WHA defines (i) first-order and (ii) second-order conditions on the local edges $y\tilde{\iota}(\eta_u(x))$, the second-order condition being, in the boosting jargon, a condition on *confidences* ($|\tilde{\iota}|$, Schapire & Singer (1999)). Since it is more involved than classical boosting's, let us exemplify how our WHA works if we have a leaf λ where local “treatments due to the black-box” are bad ($y\tilde{\iota}(\eta_u(x)) < 0$ often). In such a case, $e(M_\lambda, \eta_t) < 0$ so the balanced distribution (Definition 3) reweights higher examples whose treatment is better than average, *i.e.* the local minority. Suppose (i) holds as is without the $|.|$. In such a case, the split “aligns” the treatment quality with h , so $h = +1$ for a substantial part of this minority. (ii) imposes $\mathbb{E}_{(X,Y) \sim D_{t,\lambda}} [h(X)] \geq \mathbb{E}_{(X,Y) \sim D_{t,\lambda}} [\tilde{\iota}^2(\eta_u(X)) \cdot h(X)]$: $h = -1$ for a substantial part of large confidence treatment. The split thus tends to separate mostly large confidence but bad treatments (left) and mostly good treatments (right). Before the split, the value $\Upsilon(\lambda)$ would be negative (18) and thus reverse the polarity of the black-box, which would be good for badly treated examples but catastrophic for the local minority of adequately treated examples. After the split however, we still have the left ($h = -1$) leaf where this would eventually happen, but the minority at λ would have disproportionately ended in the right ($h = +1$) leaf, where it would be likely that $\Upsilon(.)$ would this time be *positive* and thus preserve the polarity of the treatment of the black-box. We now state our boosting compliance for TOPDOWN.

Theorem 3. Suppose (a) Assumption 1 holds, (b) we pick the heaviest leaf to split at each iteration in Step 2.1 of TOPDOWN and (c) $\exists \gamma > 0$ such that each split h^* (Step 2.2) in Υ γ -witnesses the WHA. Then there exists a constant $c > 0$ such that $\forall \varepsilon > 0$, if the number of leaves of Υ satisfies $|\Lambda(\Upsilon)| \geq (1/\varepsilon)^c \log(\frac{1}{\varepsilon})/\gamma^2$, then the posterior η_f crafted from (9) using TOPDOWN's Υ achieves $L(\eta_f; M, \eta_t) \leq \varepsilon$.

The proof of Theorem 3 is in SI, Section III. it proceeds in

two stages, the first being the proof that

$$L(\eta_f; M, \eta_t) \leq H(\Upsilon; M, \eta_t), \quad (20)$$

with the scoring in (18), the second being the boosting results focused on the entropy H of the α -tree.

An audacious scoring scheme for α -trees Let us call *conservative* the scoring scheme in (18). There is an alternative scoring scheme, which can lead to substantially larger corrections in absolute values, hence the naming, and yields better entropic bounds for the α -tree.

Definition 4. For any mixture M and posteriors η_u, η_t , let

$$e^+(M, \eta_t) \doteq \mathbb{E}_{(X,Y) \sim D_t} [\max\{0, Y\tilde{\iota}(\eta_u(X))\}], \quad (21)$$

$$e^-(M, \eta_t) \doteq -\mathbb{E}_{(X,Y) \sim D_t} [\min\{0, Y\tilde{\iota}(\eta_u(X))\}] \quad (22)$$

The *audacious* scoring schemes at the leaves of the α -tree replaces (18) in Step 3 by:

$$\Upsilon(\lambda) \doteq \tilde{\iota} \left(\frac{e^+(M_\lambda, \eta_t)}{e^+(M_\lambda, \eta_t) + e^-(M_\lambda, \eta_t)} \right), \forall \lambda \in \Lambda(\Upsilon).$$

Theorem 4. Suppose Assumption 1 holds and let $H_2(q) \doteq H(q)/\log 2$ ($\in [0, 1]$), H being defined in Definition 2. For any leaf $\lambda \in \Lambda(\Upsilon)$, denote for short:

$$H_2(\lambda; M, \eta_t) \doteq \log(2) \cdot \left(1 + (e_\lambda^+ + e_\lambda^-) \cdot \left(H_2 \left(\frac{e_\lambda^+}{e_\lambda^+ + e_\lambda^-} \right) - 1 \right) \right),$$

where we used shorthands $e_\lambda^b \doteq e^b(M_\lambda, \eta_t)$, $\forall b \in \{+, -\}$. Using the audacious scoring scheme, we get instead of (20):

$$L(\eta_f; M, \eta_t) \leq \mathbb{E}_{\lambda \sim M_{\Lambda(\Upsilon)}} [H_2(\lambda; M, \eta_t)]. \quad (23)$$

(proof in SI, Section IV) At first glance, the upperbounds in (20) and (23) may look non comparable, but it takes a simple argument to show that (23) is never worse and can me much tighter.

Lemma 2. $\forall \alpha$ -tree Υ , $\mathbb{E}_{\lambda \sim M_{\Lambda(\Upsilon)}} [H_2(\lambda; M, \eta_t)] \leq H(\Upsilon; M, \eta_t)$.

(proof in SI, Section V) It thus comes at no surprise that using the audacious scoring also results in a boosting result for TOPDOWN guaranteeing the same rates as in Theorem 3. It also takes a simple picture to show that the per-leaf slack in Lemma 2 can be substantial, a slack which can be represented using a simple picture, see Figure 2 (left), following from the use of Jensen's inequality in the Lemma's proof.

Conservative vs audacious corrections If we were to just care about accuracy, we would barely have any reason to use the conservative correction. Even thinking about generalisation, the Rademacher complexity of decision trees is a function of their depth so the faster the convergence, the better (Bartlett & Mendelson, 2002, Section 4.1) (see also Section 7). Adding fairness substantially changes the picture: some constraints, like equality of opportunity (Section 6) can antagonise accuracy to some extent. In such a case, using the conservative correction can keep posteriors

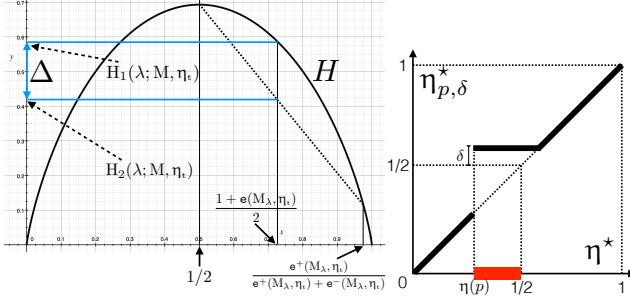


Figure 2. Left: Difference between the per-leaf bounds on risk $L(\eta_f; M, \eta_t)$ using (17) and (20) (conservative scoring) and (23) (audacious scoring). Details in the proof of Lemma 2. Right: A representation of the (p, δ) -pushup of η^* , where $\eta(p) \doteq \inf \eta^*(\mathcal{X}_p) < 1/2$ (Definition 5). All posteriors in $[\eta(p), 1/2 + \delta]$ are mapped to $1/2 + \delta$; others do not change. The new posterior $\eta_{p,\delta}^*$ eventually reduces the accuracy of classification for observations whose posterior lands in the thick red interval (x -axis).

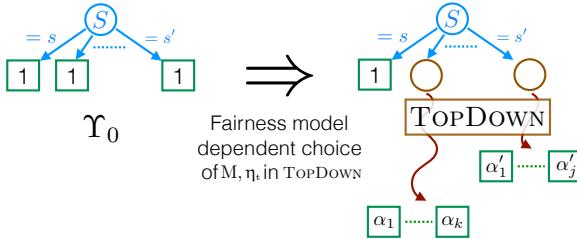


Figure 3. Picking Υ_0 a stump on the fairness attribute allows to finely tune growths of sub- α -trees to the fairness criterion at hand.

η_u and η_t close enough (Theorem 1) so that fairness can be achieved without substantial sacrifice on accuracy.

A convenient initial alpha-tree Since the fairness attribute partitions the dataset, there is a simple and convenient choice for Υ_0 in TOPDOWN, the stump whose test is on the fairness attribute (thus, not necessarily binary), resulting in separate sub- α -trees for each modality. We then run TOPDOWN with a specific choice of mixture and target posterior to accomodate the fairness model at hand, see Figure 3.

6. Handling fairness notions

To summarise, we have presented so far a general loss function (14) with plug-ins mixture M and target posterior η_t and a fast algorithm to minimise it by training interpretable models (α -trees) used to then skew the black-box prediction η_u via (9), achieving a closer guess to the target and resulting in a more fair prediction, provided M, η_t are chosen so as to tackle the fairness objective. The choices made for our three fairness notions are not meant to be optimal as other choices could provide substantial leverage; however, they provide illustrative choices of simple implementations: for η_t for example, we treat conditional value at risk with the most straightforward choice to give to each relevant x its actual posterior; for statistical parity, we rely on the simplest choice to give to all relevant x s a group's *average posterior*

as target; the most "convoluted" choice, for equality of opportunity, increases the posterior above $1/2$ to get the target posterior, for a subset of relevant x s. We now detail the example of the conditional value at risk; due to the lack of space, we defer to SI (pg 26) the case of statistical parity. **Conditional value at risk** CVAR was introduced in optimisation / finance (Rockafellar & Uryasev, 2000) and its use in fairness for ML was introduced in Williamson & Menon (2019). The criterion to minimise is:

$$\text{CVAR}_\beta(\eta_f) \doteq \mathbb{E}_{S \sim M_S} [L(\eta_f; M_S, \eta^*) | L(\eta_f; M_S, \eta^*) \geq L_\beta],$$

L_β being the risk value for the β quantile among groups, which is user defined; also, M_S is the measure induced on S by the groups' weights and M_s is the mixture conditioned on $S = s$. CVAR focuses optimisation on the worst treated groups and if we denote \mathcal{S}_β the subset of modalities used in CVAR, then a simple way to optimise CVAR is to repeatedly grow the subtree of the α -tree that makes the correction for one of those groups. Put simply, we iterate

TOPDOWN with $M \leftarrow M_s$ ($s \in \mathcal{S}_\beta$) and $\eta_t \leftarrow \eta^*$,

and we repeat until $\text{CVAR}_\beta(\eta_f)$ gets below a threshold or (more specifically) its worst treated group gets a risk below a threshold (this can be used as stopping criterion). This imposes to update \mathcal{S}_β to keep the set accurate between runs of TOPDOWN. The number of iteration to get CVAR_β below a threshold ε in our boosting framework is thus no more than the number of modalities of S times the $|\Lambda(\Upsilon)|$ bound in Theorem 3. Details are in the experimental Section.

Equality of opportunity (EOO) requires to smooth discrimination within an "advantaged" group, modeled by the label $y = 1$ (Hardt et al., 2016). We say that η_f achieves ε -equality of opportunity iff a mapping h_f of η_f to \mathcal{Y} (e.g. using the sign of its logit) satisfies

$$\max_{s \in S} \mathbb{P}_{X \sim P_s} [h_f(X) = 1] - \min_{s \in S} \mathbb{P}_{X \sim P_s} [h_f(X) = 1] \leq \varepsilon, \quad (24)$$

where P_s is the positive observations' measure conditioned to value $S = s$ for the sensitive attribute. EOO can be antagonistic with the fitting of η_f to η^* : if that latter one is close to zero in a subgroup and close to one in another one, then better fittings on η_f can arbitrarily increase the LHS in (24). To cope with this issue, we do not pick $\eta_t \leftarrow \eta^*$ as in CVAR, but rather skew the posterior for a subset of observations. Fix some $s^\circ \in \arg \min_{s \in S} \mathbb{P}_{X \sim P_s} [h_f(X) = 1]$. Our strategy consists in skewing the target posterior for $S = s^\circ$ so that for a subset of the subgroup, it becomes bigger than $1/2$. A convenient use of TOPDOWN then guarantees more positive classifications for $S = s^\circ$ – thus a more fair outcome – and thus a reduction of LHS in (24) until (24) is satisfied². To achieve this, we create a (p, δ) -pushup of η^* .

²A symmetric strategy holds if one instead wants to reduce

Definition 5. Fix $p \in [0, 1]$ and let \mathcal{X}_p be a subset of \mathcal{X} such that (i) $\inf \eta^*(\mathcal{X}_p) \geq \sup \eta^*(\mathcal{X} \setminus \mathcal{X}_p)$ and (ii) $\int_{\mathcal{X}_p} dM = p$. For any $\delta \geq 0$, the (p, δ) -pushup of η^* , $\eta_{p, \delta}^*$, is the posterior defined as $\eta_{p, \delta}^*(x) = \eta^*(x)$ if $(x \notin \mathcal{X}_p) \vee (\eta^*(x) \geq \frac{1}{2} + \delta)$ and $\eta_{p, \delta}^*(x) = \frac{1}{2} + \delta$ otherwise:

$$\eta_{p, \delta}^*(x) = \begin{cases} \eta^*(x) & \text{if } (x \notin \mathcal{X}_p) \vee (\eta^*(x) \geq \frac{1}{2} + \delta) \\ \frac{1}{2} + \delta & \text{otherwise.} \end{cases}$$

Figure 2 (right) presents an example of mapping. Notice that the transformation can introduce classification mistakes with respect to η^* , but only examples with (i) small “edge” $|1/2 - \eta^*|$ and (ii) labeled as negative on η^* are susceptible to get positive label on $\eta_{p, \delta}^*$. Notice the tradeoff achieved: (ii) is consistent with the fairness objective while (i) limits the degradation in accuracy. We then run TOPDOWN using as mixture the *positive* measure conditioned to $S = s^\circ$ and $p = \mathbb{P}_{X \sim P_{s^*}} [h_f(X) = 1] + \varepsilon/(K - 1)$, $\delta = K\varepsilon/(K - 1)$, where $K > 1$ is any user-fixed constant. In summary, we do

TOPDOWN with $M \leftarrow P_{s^\circ}$ and $\eta_t \leftarrow \eta_{p, \delta}^*$,

and we have the following guarantee:

Theorem 5. If TOPDOWN is run until $L(\eta_f; M, \eta_t) \leq (\varepsilon^4/2) + \mathbb{E}_{X \sim M} [H(\eta_t(X))]$, then after the run we observe $\mathbb{P}_{X \sim P_{s^*}} [h_f(X) = 1] - \mathbb{P}_{X \sim P_{s^\circ}} [h_f(X) = 1] \leq \varepsilon$.

The proof of Theorem 5 is in SI, Section VI. For the optimisation to be carried out properly in the full context of EOO, we should not wait to get the bound on $L(\eta_f; M, \eta_t)$. Rather, we should make sure (a) we update $\arg \min_{s \in S} \mathbb{P}_{X \sim P_s} [h_f(X) = 1]$ (and thus s°) after each split in the α -tree and (b) we keep $\arg \max_{s \in S} \mathbb{P}_{X \sim P_s} [h_f(X) = 1]$ as is, to prevent switching targets and eventually composing pushup transformations for the same $S = s^\circ$, which would not necessarily comply with our theory. One should note that the guarantee presented in Theorem 5 and Section 6 depends on the mapping h_f and not the direct posterior η_f as typically considered (Hardt et al., 2016). When taking the mapping as a threshold of the posterior (sign of the logit), h_f can be interpreted as forcing the original posterior to be extreme values of 0 or 1. If one wants to consider the typical EOO definitions depending on posterior values, the statistical parity approach can be adapted (by replacing the measure M with the measure of the positive examples P).

7. Discussion

Generalisation Moving forward with the remarks before Lemma 1, we now assume we have a m -training sample

$\mathbb{P}_{X \sim P_{s^*}} [h_f(X) = 1]$ ($s^* \in \arg \max_{s \in S} \mathbb{P}_{X \sim P_s} [h_f(X) = 1]$). Choosing one strategy depends on the application: if positive class implies money spending (e.g., for loan prediction), then our strategy implies spending more money to achieve fairness, while the latter one reduces the amount of money lent to achieve fairness.

$\mathcal{S} \doteq \{(\mathbf{x}_i, y_i) \sim D\}_{i=1}^m$. The empirical Rademacher complexity of a set of functions \mathcal{H} from \mathcal{X} to \mathbb{R} , $\mathfrak{R}_{\mathcal{S}}(\mathcal{H}) \doteq \mathbb{E}_{\sigma} \sup_{h \in \mathcal{H}} \mathbb{E}_i [\sigma_i h(\mathbf{x}_i)]$ (sampling uniform with $\sigma_i \in \{-1, 1\}$), is a capacity parameter that yields efficient control of uniform convergence when the loss used is Lipschitz (Bartlett & Mendelson, 2002, Theorem 7), which is the case of the logistic loss (Section 4). To see how the α -tree affects the Rademacher complexity of classification using η_f instead of η_u , suppose real-valued prediction based on η_u is achieved via logit mapping, $\iota \circ \eta_u$ (15). Such mappings are common for decision trees (Schapire & Singer, 1998).

Lemma 3. Suppose $\{\eta_u\}$ is the set of decision trees of depth $\leq d$ and denote $\mathfrak{R}_{\mathcal{S}}(\text{DT}(d))$ the empirical Rademacher complexity of decision trees of depth $\leq d$ (Bartlett & Mendelson, 2002) and d' the maximum depth allowed for α -trees. Then we have for \mathcal{H}_f in (15): $\mathfrak{R}_{\mathcal{S}}(\mathcal{H}_f) \leq \mathfrak{R}_{\mathcal{S}}(\text{DT}(d + d'))$.

The proof is straightforward once we remark that elements in \mathcal{H}_f can be represented as decision trees, where we plug at each leaf of η_u a copy of the α -tree Υ .

Sensitive feature use vs proxy-based prediction Post-processing methods have been flagged in the context of fair classification for the fact that they require explicit access to the sensitive feature at classification time (Zafar et al., 2019, Section 6.2.3). Our basic approach to the induction of α -trees falls in the category (Figure 3), but there is a simple way to *mask* the use of the sensitive attribute and the polarity of disparate treatment it induces: it consists in first inducing a decision tree to *predict* the sensitive feature based on the other features and use this decision tree as Υ_0 in TOPDOWN. We thus also *redefine* sensitive groups based on this decision tree – thus alleviating the need to use the sensitive attribute in the α -tree. The use of *proxy sensitive attributes* in a similar manner has seen ample use in a various domain such as health care (Bureau, 2014; Brown et al., 2016) and finance (Fremont et al., 2005). Despite the adaptation of proxy sensitive attributes, we note that its application in post-process and α -trees may not be appropriate across all domains (Datta et al., 2017).

8. Experiments

To evaluate TOPDOWN, we consider the American Community Survey (ACS) dataset preprocessed by Folktables³ (Ding et al., 2021) where we evaluate TOPDOWN’s application to various fairness models (as per Section 6 and SI pg 26). In particular, we consider the ACS dataset for income prediction in the state of CA. For these experiments, we consider *age* as the sensitive attribute in a binary and trinary modality, where it is binned with splits at 25 and 25, 50, respectively. For the black-box classifier, we consider a clipped (Assumption 1 with $B = 1$) random forest (RF) from scikit-learn calibrated using Platt’s method

³Public at: github.com/zykls/folktables

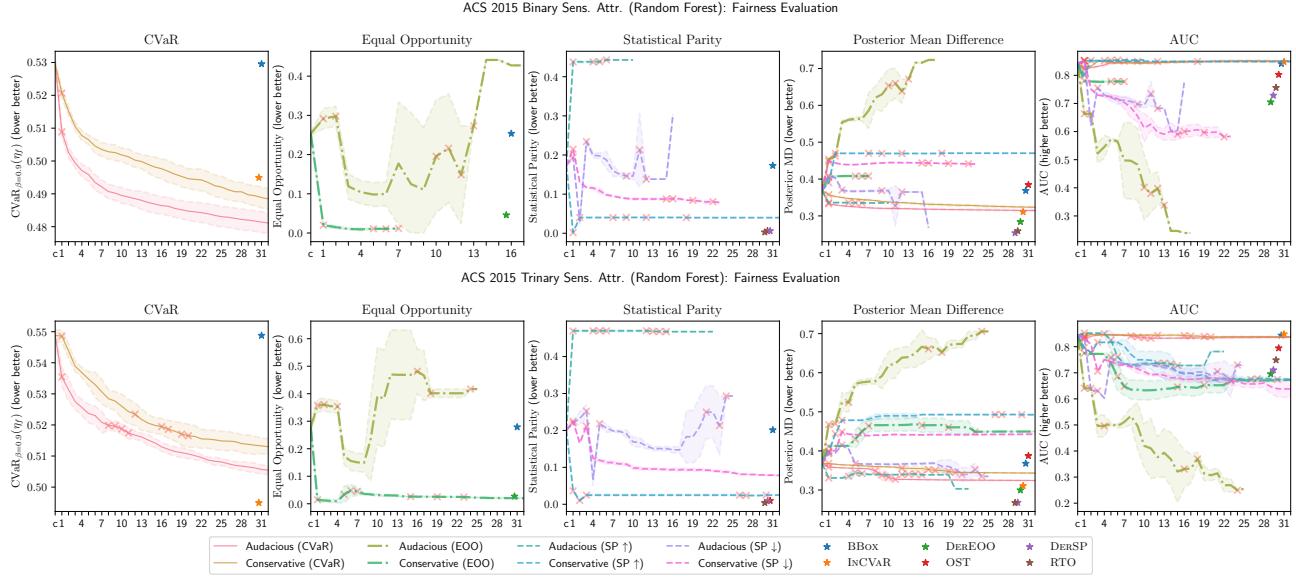


Figure 4. TOPDOWN optimized over boosting iterations for different fairness models evaluated on ACS 2015 with binary (up) and trinary (down) sensitive attributes. “c” on the x-axis denotes the clipped black-box. Crosses denote when a subgroup’s α -tree is initiated (over any fold). The shade depicts \pm a standard deviation from the mean. However, this disappears in the case where other folds stop early.

(Platt et al., 1999). The RF consists of an ensemble of 50 decision trees with a maximum depth of 4 and a random selection of 10% of the training samples per decision tree. Data is split into 3 subsets for black-box training, post-processing training, and testing; consisting of 40:40:20 splits in 5 fold cross validation. SI (pg 27) presents additional experiments on additional datasets – including considerations on proxy sensitive attributes, distribution shift, and interpretability.

Multiple fairness notions We evaluate TOPDOWN for CVAR, equality of opportunity EOO, and statistical parity SP, as per Section 6 and SI. Statistical parity aims to make subgroup’s expected posteriors similar and is popular in a various post-processing methods (Wei et al., 2020; Al-abdulmohsin & Lucic, 2021). The definition can be found in SI (pg 26) along with the strategy used in TOPDOWN. For SP, we consider two flavours: one as described directly in SI (SP \uparrow); and the symmetric strategy where the target posterior is the smallest expected subgroup posterior (SP \downarrow). Conservative and audacious updates rules are also tested. For each of these TOPDOWN configurations, we boost for 32 iterations. The initial α -tree is initialized as per Fig. 3.

To evaluate TOPDOWN, we compare against 5 baseline approaches. For CVAR we consider the in-processing approach (INCVAR) presented in Williamson & Menon (2019). For EOO, we consider a derived predictor (DEREOO) (Hardt et al., 2016). For SP, we consider an optimized score transformation approach (OST) (Wei et al., 2020); a derived predictor modified for SP (DERSP) (Hardt et al., 2016); and a randomized threshold optimizer approach (RTO) (Alabdulmohsin & Lucic, 2021). The clipped black-box is also displayed for clarity (BBox). The experiments

are summarized in Fig. 4. For clarity we only plot the baselines and wrappers which are directly associated to each fairness criterian. In addition, we also plot the posterior mean difference MD (0/1 loss) and the AUC to examine the effects on accuracy.

In the optimization of CVAR, both TOPDOWN approaches cause a decrease in CVAR. The conservative update causes a smaller decrease than the audacious approach; however as a slight trade-off the AUC of the conservative update is higher. Interesting, the MD of the audacious approach is better in both binary and ternary settings. This further demonstrates that the audacious update is more desirable when optimizing CVAR in TOPDOWN. Another observation is that in the binary case, only one sensitive attribute subgroup’s α -tree is optimized. This indicates that after 32 iterations the worse case subgroup does not change in the binary case. In comparison to the baseline approach INCVAR, both fair wrappers are capable of beating the baseline in the binary case – good news since INCVAR directly optimizes CVAR –, but are unable to do so in the trinary case.

For EOO, there is a huge difference between conservative and audacious updates as the former gets to the most fair outcomes of all baselines. Even if we used early stopping or pruning of the α -tree, audacious update would fail at producing outcomes as fair. This rejoins our remark on the interest of having a conservative update in Section 5. When compared to DEREOO, we find that the conservative TOPDOWN approach produces lower EOO for both binary and trinary cases. However, DEREOO tend to have better accuracy scores in at least one of MD and AUC (which shows interest in early stopping/pruning the α -tree).

The case of SP follows the same pattern for our technique for both targeting the largest (\uparrow) or smallest (\downarrow) expected subgroup posterior, with superior results for the conservative update vs the audacious counterpart for both binary and trinary datasets. In addition, the conservative SP \uparrow reports better SP scores and AUC scores than the conservative SP \downarrow . Comparing the best SP TOPDOWN (SP \uparrow) to the baselines, discounting OST we find that TOPDOWN only is superior in AUC; where DERSP and RTO result in lower SP and MD. This is unsurprising: our TOPDOWN treatment SP can result in harsh updates; in SI (pg 26), we discuss an alternative approach using ties with optimal transport.

References

- Alabdulmohsin, I. and Lucic, M. A near-optimal algorithm for debiasing trained machine learning models. In *NeurIPS*34*, 2021.
- Amari, S.-I. and Nagaoka, H. *Methods of Information Geometry*. Oxford University Press, 2000.
- Arimoto, S. Information-theoretical considerations on estimation problems. *Information and control*, 19:181–194, 1971.
- Bartlett, P.-L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *JMLR*, 3:463–482, 2002.
- Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., and Zhang, Y. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *IBM J. Res. Dev.*, 63:4:1–4:15, 2019.
- Brown, D. P., Knapp, C., Baker, K., and Kaufmann, M. Using bayesian imputation to assess racial and ethnic disparities in pediatric performance measures. *Health services research*, 51(3):1095–1108, 2016.
- Bureau, C. F. P. Using publicly available information to proxy for unidentified race and ethnicity: A methodology and assessment. *Washington, DC: CFPB, Summer*, 2014.
- Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. Algorithmic decision making and the cost of fairness. *CoRR*, abs/1701.08230, 2017.
- Datta, A., Fredrikson, M., Ko, G., Mardziel, P., and Sen, S. Use privacy in data-driven systems: Theory and experiments with machine learnt programs. In *24th ACM SIGSAC*, pp. 1193–1210, 2017.
- Ding, F., Hardt, M., Miller, J., and Schmidt, L. Retiring adult: New datasets for fair machine learning. *NeurIPS*34*, 2021.
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R.-S. Fairness through awareness. In *ITCS’12*, pp. 214–226, 2012.
- Dwork, C., Immorlica, N., Kalai, A.-T., and Leiserson, M.-D.-M. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on Fairness, Accountability and Transparency*, volume 81, pp. 119–133. PMLR, 2018.
- Fremont, A. M., Bierman, A., Wickstrom, S. L., Bird, C. E., Shah, M., Escarce, J. J., Horstman, T., and Rector, T. Use of geocoding in managed care settings to identify quality disparities. *Health Affairs*, 24(2):516–526, 2005.
- Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *NeurIPS’16*, pp. 3315–3323, 2016.
- Kearns, M. and Mansour, Y. On the boosting ability of top-down decision tree learning algorithms. In *Proc. of the 28th ACM STOC*, pp. 459–468, 1996.
- Kim, M.-P., Ghorbani, A., and Zou, J.-Y. Multiaccuracy: Black-box post-processing for fairness in classification. In *AAAI/ACM Conference on AI, Ethics, and Society*, pp. 247–254. ACM, 2019.
- Knuth, D.-E. Two notes on notation. *The American Mathematical Monthly*, 99(5):403–422, 1992.
- Liao, J., Kosut, O., Sankar, L., and du Pin Calmon, F. A tunable measure for information leakage. In *2018 IEEE International Symposium on Information Theory, ISIT 2018*, pp. 701–705. IEEE, 2018.
- Lohia, P.-K., Ramamurthy, K.-N., Bhide, M., Saha, D., Varshney, K.-R., and Puri, R. Bias mitigation post-processing for individual and group fairness. In *ICASSP’19*, pp. 2847–2851. IEEE, 2019.
- Martineau, K. Shrinking deep learning’s carbon footprint. <https://news.mit.edu/2020/shrinking-deep-learning-carbon-footprint-0807>, 2020.
- Mehrabi, N., Morstatter, F., Saxena, N., and Galstyan, A. A survey on bias and fairness in machine learning. *ACM CSUR*, 54:1–35, 2022.
- Menon, A.-K. and Williamson, R.-C. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*, volume 81, pp. 107–118. PMLR, 2018.

- Nock, R. and Nielsen, F. On the efficient minimization of classification-calibrated surrogates. In *NIPS*21*, pp. 1201–1208, 2008.
- Nock, R., Sypherd, T., and Sankar, L. Being properly improper. *CoRR*, abs/2106.09920, 2021.
- Petersen, F., Mukherjee, D., Sun, Y., and Yurochkin, M. Post-processing for individual fairness. *CoRR*, abs/2110.13796, 2021.
- Platt, J. et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.
- Reid, M.-D. and Williamson, R.-C. Information, divergence and risk for binary experiments. *JMLR*, 12:731–817, 2011.
- Rockafellar, R.-T. and Uryasev, S. Optimisation of conditional value-at-risk. *Journal of Risk*, 2:21–41, 2000.
- Schapire, R. E. and Singer, Y. Improved boosting algorithms using confidence-rated predictions. In *9th COLT*, pp. 80–91, 1998.
- Schapire, R. E. and Singer, Y. Improved boosting algorithms using confidence-rated predictions. *MLJ*, 37:297–336, 1999.
- Strubell, E., Ganesh, A., and McCallum, A. Energy and policy considerations for deep learning in NLP. In *ACL’19*, pp. 3645–3650, 2019.
- van Erven, T. and Harremoës, P. Rényi divergence and kullback-leibler divergence. *IEEE Trans. IT*, 60:3797–3820, 2014.
- Wei, D., Ramamurthy, K.-N., and du Pin Calmon, F. Optimized score transformation for fair classification. In *AISTATS’20*, volume 108, pp. 1673–1683, 2020.
- Williamson, R. C. and Menon, A. K. Fairness risk measures. In *International Conference on Machine Learning*, pp. 6786–6797, 2019.
- Woodworth, B.-E., Gunasekar, S., Ohannessian, M.-I., and Srebro, N. Learning non-discriminatory predictors. In *COLT’17*, volume 65, pp. 1920–1953. PMLR, 2017.
- Yang, F., Cisse, M., and Koyejo, O. O. Fairness with overlapping groups; a probabilistic perspective. *Advances in Neural Information Processing Systems*, 33, 2020.
- Zafar, M.-B., Valera, I., Gomez-Rodriguez, M., and Gummadi, K.-P. Fairness constraints: A flexible approach for fair classification. *JMLR*, 20:75:1–75:42, 2019.

Supplementary Material

Abstract

This is the Supplementary Material to Paper "Fair Wrapping for Black-box Predictions". To differentiate with the numberings in the main file, the numbering of Theorems is letter-based (A, B, ...).

Table of contents

Supplementary material on proofs and fairness models

→ Proof of Theorem 1	Pg 12
→ Proof of Corollary 2	Pg 15
→ Proof of Theorem 3	Pg 17
→ Proof of Theorem 4	Pg 21
→ Proof of Lemma 2	Pg 23
→ Proof of Theorem 5	Pg 24
→ Handling Statistical parity	Pg 26

Supplementary material on experiments

→ SI Experiment Settings	Pg 27
→ Additional Main Text Experiments	Pg 28
→ Neural Network Experiments	Pg 29
→ Proxy Sensitive Attributes	Pg 30
→ Distribution Shift	Pg 33
→ High Clip Value	Pg 35
→ Example Alpha-Tree	Pg 36

I. Proof of Theorem 1

We first show two technical Lemmata.

Lemma D. *For any $a \geq 0$, let*

$$h(z) \doteq \log \left(\frac{1}{1 + a^{1+z}} \right). \quad (25)$$

We have

$$h^{(k)}(z) = -\frac{\log^k(a) \cdot a^{1+z}}{(1 + a^{1+z})^k} \cdot P_{k-1}(a^{1+z}), \quad (26)$$

where $P_k(x)$ is a degree- $k-1$ polynomial. Letting $c_{k,j}$ the constant factor of monomial x^j in $P_k(x)$, for $j \leq k-1$, we have the following recursive definitions: $c_{1,0} = 1$ ($k=1$) and

$$c_{k+1,k} = (-1)^k, \quad (27)$$

$$c_{k+1,j} = (j+1) \cdot c_{k,j} - (k+1-j) \cdot c_{k,j-1}, \forall 0 < j < k, \quad (28)$$

$$c_{k+1,0} = 1. \quad (29)$$

Hence, we have for example $P_1(x) = 1$, $P_2(x) = -x + 1$, $P_3(x) = x^2 - 4x + 1$, $P_4(x) = -x^3 + 11x^2 - 11x + 1$, ...

Proof: We let

$$f(z) \doteq \frac{a^{1+z}}{1 + a^{1+z}}, \quad (30)$$

so that $h'(z) = -\log(a) \cdot g(z)$ and we show

$$f^{(k)}(z) = \frac{\log^k(a) \cdot a^{1+z}}{(1 + a^{1+z})^{k+1}} \cdot P_k(a^{1+z}). \quad (31)$$

We first check

$$f'(z) = \frac{\log(a) \cdot a^{1+z}}{(1 + a^{1+z})^2}, \quad (32)$$

which shows $P_1(x) = 1$. We then note that for any $k \in \mathbb{N}_*$,

$$\frac{d}{dz} \frac{a^{1+z}}{(1 + a^{1+z})^k} = \frac{\log(a) \cdot a^{1+z}}{(1 + a^{1+z})^{k+1}} \cdot ((-k)a^{1+z} + 1), \quad (33)$$

so the induction case yields $f^{(k+1)}(z) \doteq f^{(k)'}(z)$, that is:

$$\begin{aligned} f^{(k+1)}(z) &= \log^k(a) \cdot \frac{d}{dz} \left(\frac{a^{1+z}}{(1 + a^{1+z})^{k+1}} \cdot P_k(a^{1+z}) \right) \\ &= \log^k(a) \cdot \left(\frac{\log(a) \cdot a^{1+z}}{(1 + a^{1+z})^{k+2}} \cdot ((-ka^{1+z} + 1) \cdot P_k(a^{1+z}) + \frac{a^{1+z} \cdot \log(a)}{(1 + a^{1+z})^{k+1}} \cdot a^{1+z} \cdot \frac{dP_k(x)}{dx} \Big|_{x=a^{1+z}}) \right) \\ &= \frac{\log^{k+1}(a) \cdot a^{1+z}}{(1 + a^{1+z})^{k+2}} \cdot \underbrace{\left(((-ka^{1+z} + 1) \cdot P_k(a^{1+z}) + a^{1+z}(1 + a^{1+z}) \cdot \frac{dP_k(x)}{dx} \Big|_{x=a^{1+z}}) \right)}_{\doteq P_{k+1}(a^{1+z})}, \end{aligned} \quad (34)$$

from which we check that P_{k+1} is indeed a polynomial and its coefficients are obtained via identification from P_k , which establishes (31) and yields to the statement of the Lemma. \square

Lemma E. Coefficient $c_{k,j}$ admits the following bound, for any $0 \leq j \leq k$:

$$|c_{k,j}| \leq (k-1)! \binom{k-1}{j}. \quad (35)$$

Proof: First, we have the following recursive definition for the absolute value of the leveraging coefficients in $c_{\cdot,\cdot}$, (we call them $a_{\cdot,\cdot}$ for short): $|c_{\cdot,\cdot}| = a_{\cdot,\cdot}$, with

$$a_{k+1,k} = 1, \quad (36)$$

$$a_{k+1,j} = (j+1) \cdot a_{k,j} + (k+1-j) \cdot a_{k,j-1}, \forall 0 < j < k, \quad (37)$$

$$a_{k+1,0} = 1. \quad (38)$$

We now show by induction that $a_{k+1,j} \leq k! \binom{k}{j} \doteq b_{k+1,j}$. For $j = 0$, $b_{k+1,0} = k! \geq a_{k+1,0}$ ($k \geq 2$) and for $j = k$, $b_{k+1,k} = k! \geq a_{k+1,0}$ as well. We now check, assuming the property holds at all ranks k , that for ranks $k+1$, we have

$$\begin{aligned} a_{k+1,j} &= (j+1) \cdot a_{k,j} + (k+1-j) \cdot a_{k,j-1} \\ &\leq (j+1)(k-1)! \binom{k-1}{j} + (k+1-j)(k-1)! \binom{k-1}{j-1}, \end{aligned} \quad (39)$$

and we want to check that the RHS is $\leq k! \binom{k}{j}$ for any $0 < j < k$. Simplifying yields the equivalent inequality

$$(j+1)(k-j) + (k+1-j)j \leq k^2. \quad (40)$$

finding the worst case bound for j yields $j = k/2$ (we disregard the fact that j is an integer) and plugging in the bound yields the constraint on k : $k \geq 2$, which indeed holds. \square

We also check that h in Lemma D is infinitely differentiable. As a consequence, we get from Lemma D the Taylor expansion around $g = 1$ (for any $a \geq 0$) at any order $K \geq 2$,

$$\begin{aligned} &\log \left(\frac{1}{1+a^g} \right) \\ &= \log \left(\frac{1}{1+a} \right) - \frac{a \log a}{1+a} \cdot (g-1) - \underbrace{\sum_{k=2}^K \frac{a \log^k(a) P_{k-1}(a)}{k!(1+a)^k} \cdot (g-1)^k}_{\doteq R_{K,a}(g)} + o((g-1)^K). \end{aligned} \quad (41)$$

The choice to start the summation at $k = 2$ is done for technical simplifications to come. We thus have

$$\begin{aligned} \log \eta_f(\mathbf{x}) &= \log \left(\frac{1}{1 + \left(\frac{1 - \eta_u(\mathbf{x})}{\eta_u(\mathbf{x})} \right)^{\alpha(\mathbf{x})}} \right) \\ &= \log \eta_u(\mathbf{x}) - (1 - \eta_u(\mathbf{x})) \log \left(\frac{1 - \eta_u(\mathbf{x})}{\eta_u(\mathbf{x})} \right) \cdot (\alpha(\mathbf{x}) - 1) - R_{\frac{1 - \eta_u(\mathbf{x})}{\eta_u(\mathbf{x})}, K}(\alpha(\mathbf{x})) \\ &\quad + o((\alpha(\mathbf{x}) - 1)^K), \\ \log(1 - \eta_f(\mathbf{x})) &= \log(1 - \eta_u(\mathbf{x})) - \eta_u(\mathbf{x}) \log \left(\frac{\eta_u(\mathbf{x})}{1 - \eta_u(\mathbf{x})} \right) \cdot (\alpha(\mathbf{x}) - 1) - R_{\frac{\eta_u(\mathbf{x})}{1 - \eta_u(\mathbf{x})}, K}(\alpha(\mathbf{x})) \\ &\quad + o((\alpha(\mathbf{x}) - 1)^K). \end{aligned}$$

Define for short $\Delta_u(\mathbf{x}) \doteq \eta_u(\mathbf{x}) \cdot -\log \eta_f(\mathbf{x}) + (1 - \eta_u(\mathbf{x})) \cdot -\log(1 - \eta_f(\mathbf{x})) - (\eta_u(\mathbf{x}) \cdot -\log \eta_u(\mathbf{x}) + (1 - \eta_u(\mathbf{x})) \cdot -\log(1 - \eta_u(\mathbf{x})))$, so that $\text{KL}(\eta_u, \eta_f; M) = \mathbb{E}_{X \sim M} [\Delta_u(X)]$. The Taylor expansion (41) unveils an interesting simplification:

$$\begin{aligned}\Delta_u(\mathbf{x}) &= -\eta_u(\mathbf{x}) \log \eta_u(\mathbf{x}) + \eta_u(\mathbf{x})(1 - \eta_u(\mathbf{x})) \log \left(\frac{1 - \eta_u(\mathbf{x})}{\eta_u(\mathbf{x})} \right) \cdot (\alpha(\mathbf{x}) - 1) \\ &\quad + \eta_u(\mathbf{x}) \cdot R_{\frac{1 - \eta_u(\mathbf{x})}{\eta_u(\mathbf{x})}, K}(\alpha(\mathbf{x})) \\ &\quad - (1 - \eta_u(\mathbf{x})) \log(1 - \eta_u(\mathbf{x})) + (1 - \eta_u(\mathbf{x})) \eta_u(\mathbf{x}) \log \left(\frac{\eta_u(\mathbf{x})}{1 - \eta_u(\mathbf{x})} \right) \cdot (\alpha(\mathbf{x}) - 1) \\ &\quad + (1 - \eta_u(\mathbf{x})) \cdot R_{\frac{\eta_u(\mathbf{x})}{1 - \eta_u(\mathbf{x})}, K}(\alpha(\mathbf{x})) \\ &\quad - (\eta_u(\mathbf{x}) \cdot -\log \eta_u(\mathbf{x}) + (1 - \eta_u(\mathbf{x})) \cdot -\log(1 - \eta_u(\mathbf{x}))) + o((\alpha(\mathbf{x}) - 1)^K) \\ &= \eta_u(\mathbf{x}) \cdot R_{\frac{1 - \eta_u(\mathbf{x})}{\eta_u(\mathbf{x})}}(\alpha(\mathbf{x})) + (1 - \eta_u(\mathbf{x})) \cdot R_{\frac{\eta_u(\mathbf{x})}{1 - \eta_u(\mathbf{x})}, K}(\alpha(\mathbf{x})) + o((\alpha(\mathbf{x}) - 1)^K), \forall \mathbf{x} \in \mathcal{X},\end{aligned}$$

so the divergence to the black-box prediction simplifies as well, this time using Lemma E:

$$\begin{aligned}\text{KL}(\eta_u, \eta_f; M) &= \mathbb{E}_{X \sim M} \left[\eta_u(X) \cdot R_{\frac{1 - \eta_u(X)}{\eta_u(X)}, K}(\alpha(X)) + (1 - \eta_u(X)) \cdot R_{\frac{\eta_u(X)}{1 - \eta_u(X)}, K}(\alpha(X)) \right] \\ &\quad + o(\mathbb{E}_{X \sim M} [(\alpha(X) - 1)^K]).\end{aligned}\tag{42}$$

Not touching the little-oh term, we simplify further and bound the term in the expectation: for any $\mathbf{x} \in \mathcal{X}$,

$$\begin{aligned}&\eta_u(\mathbf{x}) \cdot R_{\frac{1 - \eta_u(\mathbf{x})}{\eta_u(\mathbf{x})}, K}(\alpha(\mathbf{x})) + (1 - \eta_u(\mathbf{x})) \cdot R_{\frac{\eta_u(\mathbf{x})}{1 - \eta_u(\mathbf{x})}, K}(\alpha(\mathbf{x})) \\ &= \eta_u(\mathbf{x}) \cdot \sum_{k=2}^K \frac{\frac{1 - \eta_u(\mathbf{x})}{\eta_u(\mathbf{x})} \cdot \log^k \left(\frac{1 - \eta_u(\mathbf{x})}{\eta_u(\mathbf{x})} \right) P_{k-1} \left(\frac{1 - \eta_u(\mathbf{x})}{\eta_u(\mathbf{x})} \right)}{k! \left(1 + \frac{1 - \eta_u(\mathbf{x})}{\eta_u(\mathbf{x})} \right)^k} \cdot (\alpha(\mathbf{x}) - 1)^k \\ &\quad + (1 - \eta_u(\mathbf{x})) \cdot \sum_{k=2}^K \frac{\frac{\eta_u(\mathbf{x})}{1 - \eta_u(\mathbf{x})} \cdot \log^k \left(\frac{\eta_u(\mathbf{x})}{1 - \eta_u(\mathbf{x})} \right) P_{k-1} \left(\frac{\eta_u(\mathbf{x})}{1 - \eta_u(\mathbf{x})} \right)}{k! \left(1 + \frac{\eta_u(\mathbf{x})}{1 - \eta_u(\mathbf{x})} \right)^k} \cdot (\alpha(\mathbf{x}) - 1)^k \\ &= \eta_u(\mathbf{x}) \cdot \sum_{k=2}^K \frac{\frac{1 - \eta_u(\mathbf{x})}{\eta_u(\mathbf{x})} \cdot \log^k \left(\frac{1 - \eta_u(\mathbf{x})}{\eta_u(\mathbf{x})} \right) \cdot \sum_{j=0}^{k-2} c_{k-1,j} \left(\frac{1 - \eta_u(\mathbf{x})}{\eta_u(\mathbf{x})} \right)^j}{k! \left(1 + \frac{1 - \eta_u(\mathbf{x})}{\eta_u(\mathbf{x})} \right)^k} \cdot (\alpha(\mathbf{x}) - 1)^k \\ &\quad + (1 - \eta_u(\mathbf{x})) \cdot \sum_{k=2}^K \frac{\frac{\eta_u(\mathbf{x})}{1 - \eta_u(\mathbf{x})} \cdot \log^k \left(\frac{\eta_u(\mathbf{x})}{1 - \eta_u(\mathbf{x})} \right) \cdot \sum_{j=0}^{k-2} c_{k-1,j} \left(\frac{\eta_u(\mathbf{x})}{1 - \eta_u(\mathbf{x})} \right)^j}{k! \left(1 + \frac{\eta_u(\mathbf{x})}{1 - \eta_u(\mathbf{x})} \right)^k} \cdot (\alpha(\mathbf{x}) - 1)^k \\ &= \sum_{k=2}^K \frac{\log^k \left(\frac{1 - \eta_u(\mathbf{x})}{\eta_u(\mathbf{x})} \right) \cdot \sum_{j=0}^{k-2} c_{k-1,j} \cdot \eta_u^{k-j}(\mathbf{x}) (1 - \eta_u(\mathbf{x}))^{j+1}}{k!} \cdot (\alpha(\mathbf{x}) - 1)^k \\ &\quad + \sum_{k=2}^K \frac{\log^k \left(\frac{\eta_u(\mathbf{x})}{1 - \eta_u(\mathbf{x})} \right) \cdot \sum_{j=0}^{k-2} c_{k-1,j} \cdot (1 - \eta_u(\mathbf{x}))^{k-j} \eta_u^{j+1}(\mathbf{x})}{k!} \cdot (\alpha(\mathbf{x}) - 1)^k\end{aligned}\tag{43}$$

We now note, using Lemma E that for any $\mathbf{x} \in \mathcal{X}$,

$$\begin{aligned}
 & \sum_{j=0}^{k-2} |c_{k-1,j}| \cdot \eta_u^{k-j}(\mathbf{x}) (1 - \eta_u(\mathbf{x}))^{j+1} \\
 = & \eta_u^2(\mathbf{x}) (1 - \eta_u(\mathbf{x})) \cdot \sum_{j=0}^{k-2} |c_{k-1,j}| \cdot \eta_u^{k-2-j}(\mathbf{x}) (1 - \eta_u(\mathbf{x}))^j \\
 \leq & \eta_u^2(\mathbf{x}) (1 - \eta_u(\mathbf{x})) \cdot \sum_{j=0}^{k-2} (k-2)! \binom{k-2}{j} \eta_u^{k-2-j}(\mathbf{x}) (1 - \eta_u(\mathbf{x}))^j \\
 = & (k-2)! \cdot \eta_u^2(\mathbf{x}) (1 - \eta_u(\mathbf{x})) \cdot \underbrace{\sum_{j=0}^{k-2} \binom{k-2}{j} \eta_u^{k-2-j}(\mathbf{x}) (1 - \eta_u(\mathbf{x}))^j}_{=(1-\eta_u(\mathbf{x})+\eta_u(\mathbf{x}))^{k-2}=1} \\
 = & (k-2)! \cdot \eta_u^2(\mathbf{x}) (1 - \eta_u(\mathbf{x})),
 \end{aligned}$$

and similarly

$$\sum_{j=0}^{k-2} |c_{k-1,j}| \cdot (1 - \eta_u(\mathbf{x}))^{k-j} \eta_u^{j+1}(\mathbf{x}) \leq (k-2)! \cdot \eta_u(\mathbf{x}) (1 - \eta_u(\mathbf{x}))^2,$$

so plugging the two last bounds on (43) yields the bound on $\text{KL}(\eta_u, \eta_f; M)$ from (42):

$$\begin{aligned}
 \text{KL}(\eta_u, \eta_f; M) &\leq \mathbb{E}_{\mathbf{X} \sim M} \left[\sum_{k=2}^K \frac{(\eta_u^2(\mathbf{X})(1 - \eta_u(\mathbf{X})) + \eta_u(\mathbf{X})(1 - \eta_u(\mathbf{X}))^2) \left| \log \left(\frac{1 - \eta_u(\mathbf{X})}{\eta_u(\mathbf{X})} \right) \right|^k}{k(k-1)} \cdot |\alpha(\mathbf{X}) - 1|^k \right] \\
 &\quad + o(\mathbb{E}_{\mathbf{X} \sim M} [(\alpha(\mathbf{X}) - 1)^K]) \\
 &= \mathbb{E}_{\mathbf{X} \sim M} \left[\sum_{k=2}^K \frac{\eta_u(\mathbf{X})(1 - \eta_u(\mathbf{X})) \left| \log \left(\frac{1 - \eta_u(\mathbf{X})}{\eta_u(\mathbf{X})} \right) \right|^k}{k(k-1)} \cdot |\alpha(\mathbf{X}) - 1|^k \right] \\
 &\quad + o(\mathbb{E}_{\mathbf{X} \sim M} [(\alpha(\mathbf{X}) - 1)^K]),
 \end{aligned} \tag{44}$$

which yields the statement of Theorem 1.

II. Proof of Corollary 2

We start by (S2). We study function

$$f_k(u) \doteq u(1-u) \left| \log \left(\frac{1-u}{u} \right) \right|^k, \forall u \in \left[\frac{1}{1+\exp(B)}, \frac{1}{1+\exp(-B)} \right]. \tag{45}$$

f_k being symmetric around $u = 1/2$ and zeroing in $1/2$, we consider wlog $u < 1/2$ to find its maximum, so we can drop the absolute value. We have

$$f'_k(u) = \log^{k-1} \left(\frac{1-u}{u} \right) \cdot \left((1-2u) \cdot \log \left(\frac{1-u}{u} \right) - k \right). \tag{46}$$

Function $u \mapsto (1-2u) \cdot \log \left(\frac{1-u}{u} \right)$ is strictly decreasing on $(0, 1/2)$ and has limit $+\infty$ on 0^+ , so the unique maximum of f on $[0, 1/2]$ (we close by continuity the interval in 0 since $\lim_{0^+} f = 0$) is attained at the only solution u_k of

$$(1-2u_k) \cdot \log \left(\frac{1-u_k}{u_k} \right) = k, \tag{47}$$

and such a solution always exist for any $k \ll \infty$. It also follows $u_{k+1} < u_k$, so if we denote as k^* the smallest k such that

$$u_{k^*} \leq \frac{1}{1 + \exp(B)}, \quad (48)$$

then we will have the upperbound:

$$\begin{aligned} f_k(u) &\leq \frac{1}{1 + \exp(B)} \cdot \frac{1}{1 + \exp(-B)} \cdot B^k \\ &= \frac{B^k}{2 + \exp(B) + \exp(-B)}, \forall k \geq k^*. \end{aligned} \quad (49)$$

We can also compute k^* exactly as it boils down to taking the integer part of the solution of (47) where u_k is picked as in (48):

$$k^* = \left\lfloor \frac{\exp(B) - 1}{\exp(B) + 1} \cdot B \right\rfloor, \quad (50)$$

to get $k^* = 2$, it is sufficient that $B \leq 3$, which thus gives:

$$\text{KL}(\eta_u, \eta_f; M) \leq \sum_{k=2}^K \frac{\mathbb{E}_{X \sim M} [(B \cdot |\alpha(X) - 1|)^k]}{(2 + \exp(B) + \exp(-B))k(k-1)} + G, \quad (51)$$

and if $|\alpha(x) - 1| \leq 1/B = 1/3, \forall x \in \mathcal{X}$, then we can include all terms for all $k \geq 2$ in the upperbound, which makes the little-oh remainder vanish and we get:

$$\text{KL}(\eta_u, \eta_f; M) \leq \lim_{K \rightarrow +\infty} \frac{1}{2 + \exp(B) + \exp(-B)} \cdot \sum_{k=2}^K \frac{1}{k(k-1)} \quad (52)$$

$$\leq \frac{1}{2 + \exp(B) + \exp(-B)} \cdot \sum_{k \geq 1} \frac{1}{k^2} \quad (53)$$

$$= \frac{\pi^2}{6(2 + \exp(B) + \exp(-B))}, \quad (54)$$

which is (13) and proves the Corollary for setting **(S2)**. The proof for setting **(S1)** is direct as in this case we get:

$$\begin{aligned} \text{KL}(\eta_u, \eta_f; M) &\leq \lim_{K \rightarrow +\infty} \mathbb{E}_{X \sim M} \left[\sum_{k=2}^K \frac{\eta_u(X)(1 - \eta_u(X))f^k(\alpha(X), \eta_u(X))}{k(k-1)} \right] \\ &= \mathbb{E}_{X \sim M} \left[\sum_{k=2}^K \frac{\eta_u(X)(1 - \eta_u(X))f^k(\alpha(X), \eta_u(X))}{k(k-1)} \right] \\ &\leq \mathbb{E}_{X \sim M} \left[\sum_{k=2}^K \frac{\eta_u(X)(1 - \eta_u(X))}{k(k-1)} \right] \end{aligned} \quad (55)$$

$$\leq \frac{1}{4} \cdot \sum_{k=2}^K \frac{1}{k(k-1)} \quad (56)$$

$$\leq \frac{1}{4} \cdot \sum_{k \geq 1} \frac{1}{k^2} \quad (57)$$

$$= \frac{\pi^2}{24}, \quad (58)$$

as claimed.

Figure 5 provides an idea of the set of *admissible* couples (correction, black-box posterior) that comply with **(S1)**, from which we see that the range of admissible corrections is quite flexible, even when η_u comes quite close to $\{0, 1\}$.

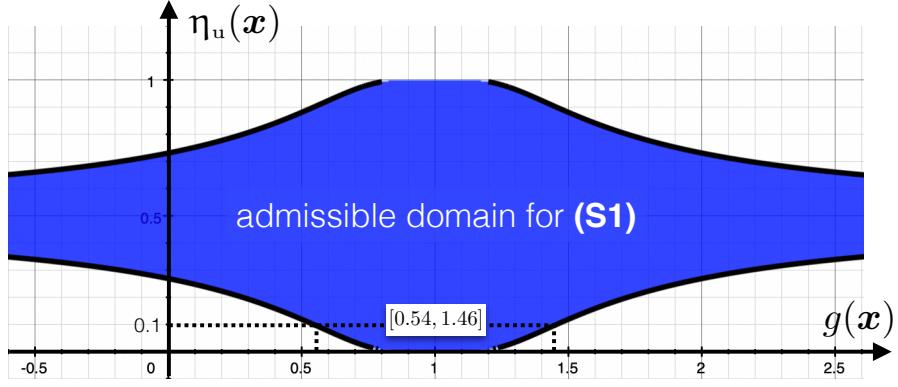


Figure 5. Admissible couples of values (g, η_u) (in blue) complying with setting (S1). For example, any couple $(g, 0.1)$ with $g \in [0.54, 1.46]$ is admissible.

III. Proof of Theorem 3

We proceed in two steps, first showing that the loss we care about for fairness (14) (main file) is upperbounded by the entropy of the α -tree Υ , then developing the boosting result from the minimisation of the entropy itself. We thus start with the following Theorem.

Theorem F. Suppose Assumption 1 holds and the outputs of Υ are:

$$\Upsilon(\mathbf{x}) \doteq \tilde{\iota}\left(\frac{1 + e(M_{\lambda(\mathbf{x})}, \eta_t)}{2}\right), \forall \mathbf{x} \in \mathcal{X}, \quad (59)$$

where $\lambda(\mathbf{x})$ is the leaf reached by \mathbf{x} in Υ . Then the following bound holds for the risk (14):

$$L(\eta_f; M, \eta_t) \leq H(\Upsilon; M, \eta_t). \quad (60)$$

Proof: We need a simple Lemma, see e.g. Nock et al. (2021).

Lemma F. $\forall \kappa \in \mathbb{R}, \forall B \geq 0, \forall |z| \leq B$,

$$\log(1 + \exp(\kappa z)) \leq \log(1 + \exp(\kappa B)) - \kappa \cdot \frac{B - z}{2}. \quad (61)$$

We then note, using $z \doteq \log\left(\frac{1-\eta_u}{\eta_u}\right)$ (stripping variables for readability) and Assumption 1,

$$\begin{aligned} -\log \eta_f &= -\log\left(\frac{\eta_u^\Upsilon}{\eta_u^\Upsilon + (1 - \eta_u)^\Upsilon}\right) \\ &= -\log\left(\frac{1}{1 + \left(\frac{1-\eta_u}{\eta_u}\right)^\Upsilon}\right) \\ &= \log\left(1 + \left(\frac{1-\eta_u}{\eta_u}\right)^\Upsilon\right) \\ &= \log\left(1 + \exp\left(\Upsilon \log\left(\frac{1-\eta_u}{\eta_u}\right)\right)\right) \\ &\leq \log(1 + \exp(\Upsilon B)) - \Upsilon \cdot \frac{B - \log\left(\frac{1-\eta_u}{\eta_u}\right)}{2} \end{aligned} \quad (62)$$

$$= \log(1 + \exp(\Upsilon B)) - \Upsilon \cdot \frac{B + \iota(\eta_u)}{2}, \quad (63)$$

where in (62) we have used (61) with $\kappa \doteq \Upsilon$, using Assumption 1 guaranteeing $|\iota(\eta_u)| \leq B$. We also get, using this time $\kappa \doteq -\Upsilon$,

$$\begin{aligned}
 -\log(1 - \eta_f) &= \log \left(1 + \exp \left(-\Upsilon \log \left(\frac{1 - \eta_u}{\eta_u} \right) \right) \right) \\
 &\leq \log(1 + \exp(-\Upsilon B)) + \Upsilon \cdot \frac{B + \iota(\eta_u)}{2} \\
 &= \log(1 + \exp(\Upsilon B)) - \Upsilon B + \Upsilon \cdot \frac{B + \iota(\eta_u)}{2} \\
 &= \log(1 + \exp(\Upsilon B)) - \Upsilon \cdot \frac{B - \iota(\eta_u)}{2}.
 \end{aligned} \tag{64}$$

Assembling (63) and (64) for an upperbound to $L(\eta_f; M, \eta_t)$, we get, using the fact that an α -tree partitions \mathcal{X} into regions with constant predictions,

$$\begin{aligned}
 L(\eta_f; M, \eta_t) &\doteq \mathbb{E}_{X \sim M} [\eta_t(X) \cdot -\log \eta_f(X) + (1 - \eta_t(X)) \cdot -\log(1 - \eta_f(X))] \\
 &\leq \mathbb{E}_{X \sim M} \left[\eta_t(X) \cdot \left(\log(1 + \exp(\Upsilon(X)B)) - \Upsilon(X) \cdot \frac{B + \iota(\eta_u(X))}{2} \right) \right. \\
 &\quad \left. + (1 - \eta_t(X)) \cdot \left(\log(1 + \exp(\Upsilon(X)B)) - \Upsilon(X) \cdot \frac{B - \iota(\eta_u(X))}{2} \right) \right] \\
 &= \mathbb{E}_{X \sim M} \left[\log(1 + \exp(\Upsilon(X)B)) - \Upsilon(X) \cdot \left(\frac{\eta_t(X) \cdot \frac{B + \iota(\eta_u(X))}{2}}{(1 - \eta_t(X)) \cdot \frac{B - \iota(\eta_u(X))}{2}} \right) \right] \\
 &= \mathbb{E}_{\lambda \sim M_{\Lambda(\Upsilon)}} \left[\log(1 + \exp(\Upsilon(\lambda)B)) - \Upsilon(\lambda) \cdot \mathbb{E}_{X \sim M_\lambda} \left[\left(\frac{\eta_t(X) \cdot \frac{B + \iota(\eta_u(X))}{2}}{(1 - \eta_t(X)) \cdot \frac{B - \iota(\eta_u(X))}{2}} \right) \right] \right] \\
 &= \mathbb{E}_{\lambda \sim M_{\Lambda(\Upsilon)}} \left[\log(1 + \exp(\Upsilon(\lambda)B)) - \Upsilon(\lambda) \cdot \mathbb{E}_{(X, Y) \sim D_{t\lambda}} \left[\frac{B + \Upsilon \cdot \iota(\eta_u(X))}{2} \right] \right] \\
 &= \mathbb{E}_{\lambda \sim M_{\Lambda(\Upsilon)}} \left[\log(1 + \exp(\Upsilon(\lambda)B)) - \Upsilon(\lambda) \cdot \frac{B + \mathbb{E}_{(X, Y) \sim D_{t\lambda}} [\Upsilon \cdot \iota(\eta_u(X))]}{2} \right] \\
 &= \mathbb{E}_{\lambda \sim M_{\Lambda(\Upsilon)}} \left[\log(1 + \exp(\Upsilon(\lambda)B)) - \Upsilon(\lambda)B \cdot \frac{1 + e(M_\lambda, \eta_t)}{2} \right],
 \end{aligned} \tag{65}$$

where we have used index notation for leaves introduced in the Theorem's statement, used the definition of $e(M_\lambda, \eta_t)$ and let $\Upsilon(\lambda)$ denote λ 's leaf value in Υ . Looking at (65), we see that we can design the leaf values to minimize each contribution to the expectation (noting the convexity of the relevant functions in $\Upsilon(\lambda)$), which for any $\lambda \in \Lambda(\Upsilon)$ we define with a slight abuse of notations as:

$$L(\Upsilon(\lambda)) \doteq \log(1 + \exp(\Upsilon(\lambda)B)) - \Upsilon(\lambda)B \cdot \frac{1 + e(M_\lambda, \eta_t)}{2}. \tag{66}$$

We note

$$L'(\Upsilon(\lambda)) = B \cdot \left(\frac{\exp(\Upsilon(\lambda)B)}{1 + \exp(\Upsilon(\lambda)B)} - \frac{1 + e(M_\lambda, \eta_t)}{2} \right),$$

which zeroes for

$$\Upsilon(\lambda) = \frac{1}{B} \cdot \log \left(\frac{1 + e(M_\lambda, \eta_t)}{1 - e(M_\lambda, \eta_t)} \right) = \tilde{\iota} \left(\frac{1 + e(M_\lambda, \eta_t)}{2} \right),$$

yielding the bound (we use $e(\lambda)$ as a shorthand for $e(M_\lambda, \eta_t)$):

$$\begin{aligned}
 L(\eta_f; M, \eta_t) &\leq \mathbb{E}_{\lambda \sim M_{\Lambda(\Upsilon)}} \left[\log \left(1 + \frac{1 + e(\lambda)}{1 - e(\lambda)} \right) - \log \left(\frac{1 + e(\lambda)}{1 - e(\lambda)} \right) \cdot \frac{1 + e(\lambda)}{2} \right] \\
 &= \mathbb{E}_{\lambda \sim M_{\Lambda(\Upsilon)}} \left[-\log \left(\frac{1 - e(\lambda)}{2} \right) - \log \left(\frac{1 + e(\lambda)}{1 - e(\lambda)} \right) \cdot \frac{1 + e(\lambda)}{2} \right] \\
 &= \mathbb{E}_{\lambda \sim M_{\Lambda(\Upsilon)}} \left[-\log \left(\frac{1 - e(\lambda)}{2} \right) + \frac{1 + e(\lambda)}{2} \cdot \log \left(\frac{1 - e(\lambda)}{2} \right) - \frac{1 + e(\lambda)}{2} \cdot \log \left(\frac{1 + e(\lambda)}{2} \right) \right] \\
 &= \mathbb{E}_{\lambda \sim M_{\Lambda(\Upsilon)}} \left[-\frac{1 - e(\lambda)}{2} \cdot \log \left(\frac{1 - e(\lambda)}{2} \right) - \frac{1 + e(\lambda)}{2} \cdot \log \left(\frac{1 + e(\lambda)}{2} \right) \right] \\
 &= H(\Upsilon; M, \eta_t),
 \end{aligned} \tag{67}$$

which is the statement of Theorem F. \square

Armed with Theorem F, what we now show is the boosting compliant convergence on the entropy of the α -tree. For the informed reader, the proof of our result relies on a generalisation of Kearns & Mansour (1996, Lemma 2), then branching on the proofs of Kearns & Mansour (1996, Lemma 6, Theorem 9) to complete our result. For this objective, we first introduce notations, summarized in Figure 6, for the split of a leaf λ_q in a subtree with two new leaves λ_p, λ_r . Here, we make use of simplified notation

$$e_p \doteq e(M_{\lambda_p}, \eta_t), \tag{68}$$

and similarly for e_q and e_r . Quantities $p, q, r \in [0, 1]^4$ are computed from the corresponding e . τ is the probability, measured from $D_{t\lambda_q}$, that an example has $h(.) = +1$, where h is the split function at λ_q . We state and prove our

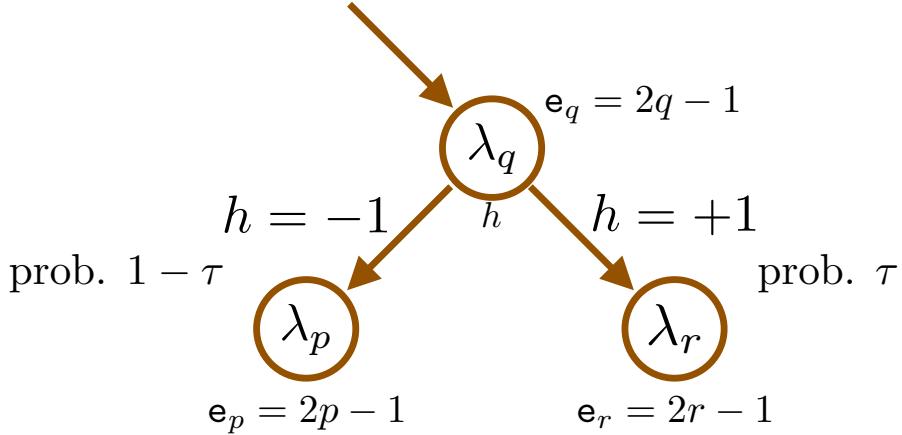


Figure 6. Main notations used in the proof of Theorem 3, closely following some notations of Kearns & Mansour (1996, Fig. 4).

generalisation to Kearns & Mansour (1996, Lemma 2).

Lemma G. Assuming notations in Figure 6 for the split h investigated at a leaf λ_q , and letting $\delta \doteq r - p$, if for some $\gamma > 0$ the split h γ -witnesses the WHA at λ , then $\tau(1 - \tau)\delta \geq \gamma \cdot q(1 - q)$.

⁴Under Assumption 1.

Proof: Using the definition of the rebalanced distribution, we have:

$$\begin{aligned}
 & \mathbb{E}_{(X,Y) \sim D'_{t\lambda_q}} [Y\tilde{\iota}(\eta_u(X))h(X)] \\
 &= \mathbb{E}_{(X,Y) \sim D_{t\lambda_q}} \left[\frac{1 - e_q \cdot Y \cdot \tilde{\iota}(\eta_u(X))}{1 - e_q^2} \cdot Yh(X)\tilde{\iota}(\eta_u(X)) \right] \\
 &= \frac{\mathbb{E}_{(X,Y) \sim D_{t\lambda_q}} [Yh(X)\tilde{\iota}(\eta_u(X))] - e_q \cdot \mathbb{E}_{(X,Y) \sim D_{t\lambda_q}} [\tilde{\iota}^2(\eta_u(X))h(X)]}{1 - e_q^2}, \tag{69}
 \end{aligned}$$

since $y^2 = 1, \forall y \in \mathcal{Y}$. We also have, by definition of the partition induced by h and the definition of τ ,

$$\begin{aligned}
 \tau e_r - (1 - \tau)e_p &= \tau \cdot \mathbb{E}_{(X,Y) \sim D_{t\lambda_r}} [Y \cdot \tilde{\iota}(\eta_u(X))] - (1 - \tau) \cdot \mathbb{E}_{(X,Y) \sim D_{t\lambda_p}} [Y \cdot \tilde{\iota}(\eta_u(X))] \\
 &= \mathbb{E}_{(X,Y) \sim D_{t\lambda_q}} [Yh(X)\tilde{\iota}(\eta_u(X))]. \tag{70}
 \end{aligned}$$

We can thus write:

$$\begin{aligned}
 & \mathbb{E}_{(X,Y) \sim D'_{t\lambda_q}} [Y\tilde{\iota}(\eta_u(X))h(X)] \\
 &= \frac{\tau e_r - (1 - \tau)e_p - e_q \cdot \mathbb{E}_{(X,Y) \sim D_{t\lambda_q}} [\tilde{\iota}^2(\eta_u(X))h(X)]}{1 - e_q^2} \tag{71}
 \end{aligned}$$

$$= \frac{2\tau e_r - e_q \cdot \left(1 + \mathbb{E}_{(X,Y) \sim D_{t\lambda_q}} [\tilde{\iota}^2(\eta_u(X))h(X)] \right)}{1 - e_q^2} \tag{72}$$

$$= \frac{2\tau e_r - 2\tau e_q - e_q \cdot \left(1 - 2\tau + \mathbb{E}_{(X,Y) \sim D_{t\lambda_q}} [\tilde{\iota}^2(\eta_u(X))h(X)] \right)}{1 - e_q^2} \tag{73}$$

$$= \frac{2\tau e_r - 2\tau e_q + e_q \cdot \left(2\tau - 1 - \mathbb{E}_{(X,Y) \sim D_{t\lambda_q}} [\tilde{\iota}^2(\eta_u(X))h(X)] \right)}{1 - e_q^2} \tag{74}$$

$$= \frac{2\tau e_r - 2\tau e_q + \frac{e_q \cdot \mathbb{E}_{(X,Y) \sim D_{t\lambda_q}} [(1 - \tilde{\iota}^2(\eta_u(X))) \cdot h(X)]}{1 - e_q^2}}{1 - e_q^2}. \tag{75}$$

Here, (71) follows from (69) and (70), (72) uses the fact that $e_q = (1 - \tau)e_p + \tau e_r$, (73) and (74) are convenient reformulations after adding $2\tau e_q - 2\tau e_q$ and (75) follows from $\mathbb{E}_{(X,Y) \sim D_{t\lambda_q}} [h(X)] = 2\tau - 1$ by definition of τ and $h \in \{-1, 1\}$. Let

$$\Delta(h) \doteq e_q \cdot \mathbb{E}_{(X,Y) \sim D_{t\lambda_q}} [(1 - \tilde{\iota}^2(\eta_u(X))) \cdot h(X)]. \tag{76}$$

We have $p = (1 + e_p)/2$ (and similarly for $q = (1 + e_q)/2$ and $r = (1 + e_r)/2$), so we reformulate (74) as:

$$\begin{aligned}
 \mathbb{E}_{(X,Y) \sim D'_{t\lambda_q}} [Y\tilde{\iota}(\eta_u(X))h(X)] &= \frac{2\tau(2r - 2q)}{4q(1 - q)} + \frac{\Delta(h)}{4q(1 - q)} \\
 &= \frac{\tau(r - q)}{q(1 - q)} + \frac{\Delta(h)}{4q(1 - q)} \\
 &= \frac{\tau(1 - \tau)\delta}{q(1 - q)} + \frac{\Delta(h)}{4q(1 - q)}, \tag{77}
 \end{aligned}$$

where the last identity comes from the fact that $r = q + (1 - \tau)\delta$. We now have two cases depending on what removing the absolute value in the WHA leads to:

Case 1 (i) is $\mathbb{E}_{(X,Y) \sim D'_{t\lambda_q}} [Y\tilde{\iota}(\eta_u(X))h(X)] \geq \gamma$. We get from (77):

$$\tau(1 - \tau)\delta \geq \gamma \cdot q(1 - q) - \frac{\Delta(h)}{4}, \tag{78}$$

and since (ii) brings $\Delta(h) \leq 0$, we obtain $\tau(1 - \tau)\delta \geq \gamma \cdot q(1 - q)$, as claimed.

Case 2 (i) is $\mathbb{E}_{(X,Y) \sim D'_{t,\lambda_q}} [Y\tilde{\iota}(\eta_u(X))h(X)] \leq -\gamma$. Since \mathcal{H} is closed by negation we replace h by $h' \doteq -h$, which satisfies $\mathbb{E}_{(X,Y) \sim D'_{t,\lambda_q}} [Y\tilde{\iota}(\eta_u(X))h'(X)] = -\mathbb{E}_{(X,Y) \sim D'_{t,\lambda_q}} [Y\tilde{\iota}(\eta_u(X))h(X)]$. The change switches the sign of δ by its definition and also $\Delta(h') = -\Delta(h)$ so (78) becomes $-\tau(1 - \tau)\delta \leq -\gamma \cdot q(1 - q) + \Delta(h')/4$, i.e.

$$\tau(1 - \tau)\delta \geq \gamma \cdot q(1 - q) - \frac{\Delta(h')}{4}, \quad (79)$$

which brings us back to Case 1 with the switch $h \leftrightarrow h'$ as h' satisfies $\mathbb{E}_{(X,Y) \sim D'_{t,\lambda_q}} [Y\tilde{\iota}(\eta_u(X))h'(X)] \geq \gamma$. This ends the proof of Lemma G. \square

Branching Lemma G to the proof of Theorem 3 via the results of (Kearns & Mansour, 1996) is simple as all major parameters p, q, r, δ, τ are either the same or satisfy the same key relationships (linked to the linearity of the expectation). This is why, if we compute the decrease $H(\Upsilon; M, \eta_t) - H(\Upsilon(\lambda, h); M, \eta_t)$, $\Upsilon(\lambda, h)$ being the α -tree Υ with the split in Figure 6 performed with h at λ , then we immediately get

$$H(\Upsilon; M, \eta_t) - H(\Upsilon(\lambda, h); M, \eta_t) \geq \gamma^2 q(1 - q), \quad (80)$$

which comes from Kearns & Mansour (1996, Lemma 6), and (80) can be directly used in the proof of Kearns & Mansour (1996, Theorem 9) – which unravels the local decrease of $H(\cdot; M, \eta_t)$ to get to the global decrease of the criterion for the whole of Υ 's induction –, and to get $H(\Upsilon; M, \eta_t) \leq \varepsilon$, it is sufficient that

$$|\Lambda(g)| \geq \left(\frac{1}{\varepsilon}\right)^{\frac{c \log(\frac{1}{\varepsilon})}{\gamma^2}}, \quad (81)$$

as claimed, for $c > 0$ a constant. This ends the proof of Theorem 3.

Remark 1. Lemma F reveals an interesting property: instead of requesting $\Pi_{S',\lambda}(h) \leq 0$ in split-fair-compliance, suppose we strengthen the assumption, requesting for some $\beta > 0$ that

$$\Pi_{S',\lambda}(h) \leq -\beta \cdot (1 - e_q^2), \quad (82)$$

then the "advantage" γ becomes an advantage $\gamma + \beta$ in (81). Since we have $\Pi_{S',\lambda}(h) = e_q \cdot \mathbb{E}_{(X,Y) \sim D_{S',\lambda_q}} [(1 - \tilde{\iota}^2(\eta_u(X))) \cdot h(X)]$, constraint (82) quickly vanishes as $|e_q| \rightarrow 1$, i.e. as the black-box gets very good –or– very bad (in this last case, we remark that $1 - \eta_u$ becomes very good, so this is not a surprise). For example, if $e_q \geq 1 - \varepsilon'$ for small ε' , then we just need

$$\mathbb{E}_{(X,Y) \sim D_{S',\lambda_q}} [(1 - \tilde{\iota}^2(\eta_u(X))) \cdot h(X)] \leq -\varepsilon' \beta \cdot \frac{2 - \varepsilon'}{1 - \varepsilon'}. \quad (83)$$

IV. Proof of Theorem 4

The proof is obtained via a generalisation of Lemma F.

Lemma H. Fix any $B > 0$. For any $\alpha \in \mathbb{R}$, any $\theta, z \in [-B, B]$, if we let

$$\vartheta(z) \doteq (z - \theta) \cdot \begin{cases} \frac{1}{B+\theta} & \text{if } z < \theta, \\ 0 & \text{if } z = \theta, \\ \frac{1}{B-\theta} & \text{if } z > \theta. \end{cases}, \quad (84)$$

then we have

$$\log(1 + \exp(\alpha z)) \leq \log\left(\frac{1 + \exp(B\alpha)}{1 + \exp(\theta\alpha)}\right) \cdot |\vartheta(z)| - B\alpha \max\{0, -\vartheta(z)\} + \log(1 + \exp(\theta\alpha)).$$

Remark: Lemma F is obtained for the choices $\theta = \pm B$.

Proof: We fix any $\theta' \in [-1, 1]$ and let

$$1 \doteq (-1, \log(1 + \exp(-\alpha))), \quad (85)$$

$$c \doteq (\theta', \log(1 + \exp(\alpha\theta'))), \quad (86)$$

$$r \doteq (1, \log(1 + \exp(\alpha))). \quad (87)$$

The equation of the line passing through l, c is

$$f_l(z) = \frac{\log\left(\frac{1+\exp(\theta'\alpha)}{1+\exp(-\alpha)}\right)}{1+\theta'} \cdot z + \frac{\log\left(\frac{1+\exp(\theta'\alpha)}{1+\exp(-\alpha)}\right)}{1+\theta'} + \log(1 + \exp(-\alpha)) \quad (88)$$

$$= -\frac{\log\left(\frac{1+\exp(\alpha)}{1+\exp(\theta'\alpha)}\right)}{1+\theta'} \cdot z + \frac{\alpha z}{1+\theta'} - \frac{\log\left(\frac{1+\exp(\alpha)}{1+\exp(\theta'\alpha)}\right)}{1+\theta'} + \frac{\alpha}{1+\theta'} + \log(1 + \exp(-\alpha)) \quad (89)$$

$$= \frac{\log\left(\frac{1+\exp(\alpha)}{1+\exp(\theta'\alpha)}\right)}{1+\theta'} \cdot (\theta' - z) + \frac{\alpha(z - \theta')}{1+\theta'} - \log\left(\frac{1+\exp(\alpha)}{1+\exp(\theta'\alpha)}\right) + \log(1 + \exp(\alpha)) \quad (90)$$

$$= \frac{\log\left(\frac{1+\exp(\alpha)}{1+\exp(\theta'\alpha)}\right)}{1+\theta'} \cdot (\theta' - z) + \frac{\alpha(z - \theta')}{1+\theta'} + \log(1 + \exp(\theta'\alpha)) \quad (91)$$

and the equation of the line passing through c, r is

$$f_r(z) = \frac{\log\left(\frac{1+\exp(\alpha)}{1+\exp(\theta'\alpha)}\right)}{1-\theta'} \cdot z - \frac{\log\left(\frac{1+\exp(\alpha)}{1+\exp(\theta'\alpha)}\right)}{1-\theta'} + \log(1 + \exp(\alpha)) \quad (92)$$

$$= \frac{\log\left(\frac{1+\exp(\alpha)}{1+\exp(\theta'\alpha)}\right)}{1-\theta'} \cdot (z - \theta') - \log\left(\frac{1+\exp(\alpha)}{1+\exp(\theta'\alpha)}\right) + \log(1 + \exp(\alpha)) \quad (93)$$

$$= \frac{\log\left(\frac{1+\exp(\alpha)}{1+\exp(\theta'\alpha)}\right)}{1-\theta'} \cdot (z - \theta') + \log(1 + \exp(\theta'\alpha)). \quad (94)$$

For any $z \in [-1, 1]$, define $\vartheta'(z) \in [-1, 1]$ to be:

$$\vartheta'(z) \doteq (z - \theta') \cdot \begin{cases} \frac{1}{1+\theta'} & \text{if } z < \theta', \\ 0 & \text{if } z = \theta', \\ \frac{1}{1-\theta'} & \text{if } z > \theta'. \end{cases} \quad (95)$$

Function $z \mapsto \log(1 + \exp(\alpha z))$ being convex, we thus get the secant upperbound:

$$\log(1 + \exp(\alpha z)) \leq \log\left(\frac{1 + \exp(\alpha)}{1 + \exp(\theta'\alpha)}\right) \cdot |\vartheta'(z)| + \alpha \min\{0, \vartheta'(z)\} + \log(1 + \exp(\theta'\alpha)), \quad (96)$$

and this holds for $z \in [-1, 1]$. If instead $z \in [-B, B]$, then letting $\theta \doteq B\theta' \in [-B, B]$, we note:

$$\begin{aligned} \log(1 + \exp(\alpha z)) &= \log(1 + \exp(\alpha B \cdot (z/B))) \\ &\leq \log\left(\frac{1 + \exp(\alpha B)}{1 + \exp(\theta'\alpha B)}\right) \cdot |\vartheta'(z/B)| + \alpha B \min\{0, \vartheta'(z/B)\} + \log(1 + \exp(\theta'\alpha B)), \end{aligned} \quad (97)$$

where this time,

$$\begin{aligned} \vartheta'\left(\frac{z}{B}\right) &\doteq \left(\frac{z}{B} - \theta'\right) \cdot \begin{cases} \frac{1}{1+\theta'} & \text{if } z < B\theta', \\ 0 & \text{if } z = B\theta', \\ \frac{1}{1-\theta'} & \text{if } z > B\theta'. \end{cases} \\ &= (z - \theta) \cdot \begin{cases} \frac{1}{B+\theta} & \text{if } z < \theta, \\ 0 & \text{if } z = \theta, \\ \frac{1}{B-\theta} & \text{if } z > \theta. \end{cases} \doteq \vartheta(z). \end{aligned} \quad (98)$$

We thus get

$$\log(1 + \exp(\alpha z)) \leq \log\left(\frac{1 + \exp(B\alpha)}{1 + \exp(\theta\alpha)}\right) \cdot |\vartheta(z)| + B\alpha \min\{0, \vartheta(z)\} + \log(1 + \exp(\theta\alpha)), \quad (99)$$

and since $\min\{0, z\} = -\max\{0, -z\}$, we get the statement of the Lemma. \square

We use Lemma H with $\theta = 0$, which yields $\vartheta(z) = z/B$; using notations from the proof of Theorem F, we thus get (using the same notations as in the proof of Theorem 3),

$$\begin{aligned} -\log \eta_f &= \log(1 + \exp(\Upsilon \cdot -\iota(\eta_u))) \\ &\leq \frac{1}{B} \cdot \log\left(\frac{1 + \exp(B\Upsilon)}{2}\right) \cdot |\iota(\eta_u(X))| + \Upsilon \min\{0, -\iota(\eta_u(X))\} + \log(2) \\ &= \frac{1}{B} \cdot \log\left(\frac{1 + \exp(B\Upsilon)}{2}\right) \cdot |\iota(\eta_u(X))| - \Upsilon \max\{0, \iota(\eta_u(X))\} + \log(2) \end{aligned} \quad (100)$$

$$\begin{aligned} -\log(1 - \eta_f) &= \log(1 + \exp(\Upsilon \cdot \iota(\eta_u(X)))) \\ &\leq \frac{1}{B} \cdot \log\left(\frac{1 + \exp(B\Upsilon)}{2}\right) \cdot |\iota(\eta_u(X))| + \Upsilon \min\{0, \iota(\eta_u(X))\} + \log(2) \\ &= \frac{1}{B} \cdot \log\left(\frac{1 + \exp(B\Upsilon)}{2}\right) \cdot |\iota(\eta_u(X))| - \Upsilon \max\{0, -\iota(\eta_u(X))\} + \log(2). \end{aligned} \quad (101)$$

We get that the inequality in (65) now reads (for *any* values $\{\Upsilon(\lambda), \lambda \in \Lambda(\Upsilon)\}$) $L(\eta_f; M, \eta_t) = \mathbb{E}_{\lambda \sim M_{\Lambda(\Upsilon)}} [J(\lambda)]$ with $J(\lambda)$ satisfying:

$$\begin{aligned} J(\lambda) &\leq \mathbb{E}_{X \sim M_\lambda} \left[\eta_t(X) \cdot \left(\frac{1}{B} \cdot \log\left(\frac{1 + \exp(B\Upsilon(\lambda))}{2}\right) \cdot |\iota(\eta_u(X))| - \Upsilon(\lambda) \max\{0, \iota(\eta_u(X))\} + \log(2) \right) \right. \\ &\quad \left. + (1 - \eta_t(X)) \cdot \left(\frac{1}{B} \cdot \log\left(\frac{1 + \exp(B\Upsilon(\lambda))}{2}\right) \cdot |\iota(\eta_u(X))| - \Upsilon(\lambda) \max\{0, -\iota(\eta_u(X))\} + \log(2) \right) \right] \\ &= \log(2) - B\Upsilon(\lambda) \cdot e^+(M_\lambda, \eta_t) + \log\left(\frac{1 + \exp(B\Upsilon(\lambda))}{2}\right) \cdot (e^+(M_\lambda, \eta_t) + e^-(M_\lambda, \eta_t)), \end{aligned} \quad (102)$$

and the bound takes its minimum on $\Upsilon(\lambda)$ for

$$\Upsilon(\lambda) = \frac{1}{B} \cdot \log\left(\frac{e^+(M_\lambda, \eta_t)}{e^-(M_\lambda, \eta_t)}\right) = \iota\left(\frac{e^+(M_\lambda, \eta_t)}{e^+(M_\lambda, \eta_t) + e^-(M_\lambda, \eta_t)}\right), \quad (103)$$

yielding (using notations from Theorem 4),

$$\begin{aligned} J(\lambda) &\leq \log(2) \cdot (1 - e^-_\lambda - e^+_\lambda) - e^+_\lambda \cdot \log\left(\frac{e^+_\lambda}{e^-_\lambda}\right) + \log\left(\frac{e^-_\lambda + e^+_\lambda}{e^-_\lambda}\right) \cdot (e^-_\lambda + e^+_\lambda) \\ &= \log(2) \cdot \left(1 + (e^-_\lambda + e^+_\lambda) \cdot \left(H_2\left(\frac{e^+_\lambda}{e^+_\lambda + e^-_\lambda}\right) - 1\right)\right), \end{aligned} \quad (104)$$

and brings the statement of Theorem 4 after plugging the bound in the expectation.

V. Proof of Lemma 2

We note that $H_2(1/2) = 1$, so we can reformulate:

$$\frac{H_2(\lambda; M, \eta_t)}{\log 2} = (1 - (e^+_\lambda + e^-_\lambda)) \cdot H_2\left(\frac{1}{2}\right) + (e^+_\lambda + e^-_\lambda) \cdot H_2\left(\frac{e^+_\lambda}{e^+_\lambda + e^-_\lambda}\right), \quad (105)$$

and we also have $e^+_\lambda \leq 0, e^-_\lambda \geq 0, e^+_\lambda + e^-_\lambda \leq 1$, plus

$$(1 - (e^+_\lambda + e^-_\lambda)) \cdot \left(\frac{1}{2}\right) + (e^+_\lambda + e^-_\lambda) \cdot \left(\frac{e^+_\lambda}{e^+_\lambda + e^-_\lambda}\right) = \frac{1 + e^+_\lambda - e^-_\lambda}{2} = \frac{1 + e(M_\lambda, \eta_t)}{2}, \quad (106)$$

as indeed $e(M_\lambda, \eta_t) = e_\lambda^+ - e_\lambda^-$ from its definition. Thus, by Jensen's inequality, since H is concave,

$$\begin{aligned}
 & \log(2) \cdot \left(1 + (e_\lambda^+ + e_\lambda^-) \cdot \left(H_2 \left(\frac{e_\lambda^+}{e_\lambda^+ + e_\lambda^-} \right) - 1 \right) \right) \\
 &= \log(2) \cdot \left((1 - (e_\lambda^+ + e_\lambda^-)) \cdot H_2 \left(\frac{1}{2} \right) + (e_{\rho,\lambda}^- + e_{\rho,\lambda}^+) \cdot H_2 \left(\frac{e_\lambda^+}{e_\lambda^+ + e_\lambda^-} \right) \right) \\
 &\leq \log(2) \cdot H_2 \left((1 - (e_\lambda^+ + e_\lambda^-)) \cdot \frac{1}{2} + (e_\lambda^+ + e_\lambda^-) \cdot \frac{e_\lambda^+}{e_\lambda^+ + e_\lambda^-} \right) \\
 &= \log(2) \cdot H_2 \left(\frac{1 + e(M_\lambda, \eta_t)}{2} \right) \\
 &= H \left(\frac{1 + e(M_\lambda, \eta_t)}{2} \right),
 \end{aligned}$$

which, after plugging in expectations and simplifying, yields the statement of Lemma 2.

VI. Proof of Theorem 5

We remind that we craft product measures using a mixture and a posterior that shall be implicit from context: we thus note that the KL divergence

$$KL(\eta_t, \eta_f; M) \doteq \mathbb{E}_{(X,Y) \sim D_t} \left[\log \left(\frac{dD_t((X,Y))}{dD_f((X,Y))} \right) \right] \quad (107)$$

$$= \mathbb{E}_{X \sim M} \left[\eta_t(X) \cdot -\log \left(\frac{\eta_f(X)}{\eta_t(X)} \right) + (1 - \eta_t(X)) \cdot -\log \left(\frac{1 - \eta_f(X)}{1 - \eta_t(X)} \right) \right] \quad (108)$$

$$= L(\eta_f; M, \eta_t) - \mathbb{E}_{X \sim M} [H(\eta_t(X))], \quad (109)$$

where D_t (resp. D_f) is obtained from couple (M, η_t) (resp. (M, η_f)). Denote

$$s^\circ \doteq \arg \min_s \mathbb{P}_{X \sim P_s} [h_f(X) = 1], \quad (110)$$

where h_f is the $+1 / -1$ prediction obtained from the posterior η_f using *e.g.* the sign of its logit. We define the total variation divergence:

$$TV(\eta_t, \eta_f; M) \doteq \int_{\mathcal{X} \times \mathcal{Y}} |dD_t((X,Y)) - dD_f((X,Y))|, \quad (111)$$

which, because of the definition of the product measures, is also equal to:

$$TV(\eta_t, \eta_f; M) = \int_{\mathcal{X}} |\eta_t(X)dM(X) - \eta_f(X)dM(X)| \quad (112)$$

$$+ \int_{\mathcal{X}} |(1 - \eta_t(X))dM(X) - (1 - \eta_f(X))dM(X)| \quad (113)$$

$$= 2 \int_{\mathcal{X}} |\eta_t(X) - \eta_f(X)|dM(X). \quad (114)$$

We have Pinsker's inequality, $TV(\eta_t, \eta_f; M) \leq \sqrt{2KL(\eta_t, \eta_f; M)}$ (see *e.g.* (van Erven & Harremoës, 2014)), so if we run `TOPDOWN` until

$$L(\eta_f; M, \eta_t) \leq \frac{\tau^2}{2} + \mathbb{E}_{X \sim M} [H(\eta_t(X))], \quad (115)$$

then because of (109) and (114),

$$\int_{\mathcal{X}} |\eta_t(X) - \eta_f(X)|dM(X) \leq \tau. \quad (116)$$

Denote subgroups $s^* \doteq \arg \max_s \mathbb{P}_{X \sim P_s} [h_f(X) = 1]$ and $s^\circ \doteq \arg \min_s \mathbb{P}_{X \sim P_s} [h_f(X) = 1]$. We pick

$$M \leftarrow P_{s^\circ} \quad (117)$$

for TOPDOWN and the (p, δ) -push up posterior η_t , with

$$p \doteq \mathbb{P}_{X \sim P_{s^*}} [h_f(X) = 1] + \frac{\delta}{2}, \quad (118)$$

assuming the RHS is ≤ 1 .

Denote \mathcal{X}_{p, s° the subset of the support of P_{s° such that $\eta_t(X) \geq (1/2) + \delta$. Notice that by definition,

$$\int_{\mathcal{X}_{p, s^\circ}} dP_{s^\circ}(X) = p. \quad (119)$$

We have two possible outcomes for η_f of relevance on \mathcal{X}_{p, s° : (i) $\eta_f(X) \leq 1/2$ and (ii) $\eta_f(X) > 1/2$. Notice that in this latter case, we are guaranteed that $h_f(X) = 1$, which counts towards bringing closer $\mathbb{P}_{X \sim P_{s^\circ}} [h_f(X) = 1]$ to $\mathbb{P}_{X \sim P_{s^*}} [h_f(X) = 1]$, so we have to make sure that (i) occurs with sufficiently small probability, and this is achieved via guarantee (116).

If the total weight on \mathcal{X}_{p, s° of the event (i) $\eta_f(X) \leq 1/2$ is more than δ , then

$$\begin{aligned} \int_{\mathcal{X}} |\eta_t(X) - \eta_f(X)| dP_{s^\circ}(X) &\geq \int_{\mathcal{X}_{p, s^\circ}} |\eta_t(X) - \eta_f(X)| dP_{s^\circ}(X) \\ &\geq \left| \frac{1}{2} + \delta - \frac{1}{2} \right| \cdot \int_{\mathcal{X}_{p, s^\circ}} [\eta_f(X) \leq 1/2] dP_{s^\circ}(X) \\ &> \left| \frac{1}{2} + \delta - \frac{1}{2} \right| \cdot \delta \\ &= \delta^2. \end{aligned} \quad (120)$$

If we have the relationship $\delta = \sqrt{\tau}$, then we get a contradiction with (116). In conclusion, if (128) holds, then

$$\int_{\mathcal{X}_{p, s^\circ}} [\eta_f(X) \leq 1/2] dP_{s^\circ}(X) \leq \delta. \quad (121)$$

In summary, for any $\tau > 0$, if we run TOPDOWN with the choices $M \leftarrow P_{s^\circ}$ (which corresponds to the "worst treated" subgroup with respect to EOO) and craft the (p, δ) -push up posterior η_t with p as in (118), then

$$\mathbb{P}_{X \sim P_{s^\circ}} [h_f(X) = 1] \geq \int_{\mathcal{X}_{p, s^\circ}} [\eta_f(X) > 1/2] dP_{s^\circ}(X) \quad (122)$$

$$= \int_{\mathcal{X}_{p, s^\circ}} (1 - [\eta_f(X) \leq 1/2]) dP_{s^\circ}(X) \quad (123)$$

$$= \int_{\mathcal{X}_{p, s^\circ}} dP_{s^\circ}(X) - \int_{\mathcal{X}_{p, s^\circ}} [\eta_f(X) \leq 1/2] dP_{s^\circ}(X) \quad (124)$$

$$\geq p - \delta \quad (125)$$

$$= \mathbb{P}_{X \sim P_{s^*}} [h_f(X) = 1] - \frac{\delta}{2}, \quad (126)$$

where (125) makes use of (119) and (121). Fixing $\delta \doteq 2\varepsilon$, ε being used in (24) (main file), we obtain

$$\mathbb{P}_{X \sim P_{s^*}} [h_f(X) = 1] - \mathbb{P}_{X \sim P_{s^\circ}} [h_f(X) = 1] \leq \varepsilon, \quad (127)$$

and via relationship $\delta = \sqrt{\tau}$, we check that (128) becomes the following function of ε :

$$L(\eta_f; M, \eta_t) \leq 8\varepsilon^4 + \mathbb{E}_{X \sim M} [H(\eta_t(X))], \quad (128)$$

and we get the statement of the Theorem for the choice (118), which corresponds to $K = 2$ and reads

$$p \doteq \mathbb{P}_{X \sim P_{s^*}} [h_f(X) = 1] + \varepsilon. \quad (129)$$

If the RHS in (129) is not ≤ 1 , we can opt for an alternative with one more free variable, $K \geq 1$,

$$p \doteq \mathbb{P}_{X \sim P_{s^*}} [h_f(X) = 1] + \frac{\delta}{K}, \quad (130)$$

where K is large enough for the constraint to hold. In this case, to keep (127) we must have $\delta(K - 1)/K = \varepsilon$, which elicitates

$$\delta = \frac{K\varepsilon}{K - 1} \quad (131)$$

instead of $\delta \doteq 2\varepsilon$, bringing

$$p \doteq \mathbb{P}_{X \sim P_{s^*}} [h_f(X) = 1] + \frac{\varepsilon}{K - 1}, \quad (132)$$

and a desired approximation guarantee for TOPDOWN of:

$$L(\eta_f; M, \eta_t) \leq \frac{K^4}{2(K - 1)^4} \cdot \varepsilon^4 + \mathbb{E}_{X \sim M} [H(\eta_t(X))]. \quad (133)$$

Since $K > 1$, $K^4/(K - 1)^4 \geq 1$, so we are guaranteed that (133) holds if we ask for

$$L(\eta_f; M, \eta_t) \leq \frac{\varepsilon^4}{2} + \mathbb{E}_{X \sim M} [H(\eta_t(X))], \quad (134)$$

VII. Handling Statistical parity

Statistical parity (SP) is a group fairness notion (Dwork et al., 2012), implemented recently in a context similar to ours (Alabdulmohsin & Lucic, 2021) as the constraint that per-group expected treatments must not be too far from each other. We say that η_f achieves ε -statistical parity (across all groups induced by sensitive attribute S) iff

$$\max_{s \in S} \mathbb{E}_{X \sim M_s} [\eta_f(X)] - \min_{s \in S} \mathbb{E}_{X \sim M_s} [\eta_f(X)] \leq \varepsilon. \quad (135)$$

Denote $s^\circ \doteq \arg \min_{s \in S} \mathbb{E}_{X \sim M_s} [\eta_f(X)]$, $s^* \doteq \arg \max_{s \in S} \mathbb{E}_{X \sim M_s} [\eta_f(X)]$. Since the risk we minimise in (14) involves a proper loss, the most straightforward use of TOPDOWN is to train the sub- α -tree for one of these two groups, giving as target posterior the *expected* posterior of the other group, *i.e.* we use $\eta_t(x) = \mathbb{E}_{X \sim M_{s^*}} [\eta_u(X)] \doteq \bar{\eta}_u$ if we grow the α -tree of s° and thus iterate

TOPDOWN with $M \leftarrow M_{s^\circ}$ and $\eta_t \leftarrow \bar{\eta}_u$,

and we repeat until s° does not achieve anymore the smallest expected posterior. We then update the group and repeat the procedure until a given slack ε is achieved between the extremes in (135). More sophisticated / gentle approaches are possible, including using the links between statistical parity and optimal transport (OT, Dwork et al. (2012, Section 3.2)), suggesting to use as target posterior the expected posterior obtained from an OT plan between groups s° and s^* .

VIII. SI Experiment Settings

In this SI section, we briefly discuss the additional datasets⁵ and experimental settings included in the subsequent sections. In particular, we highlight the datasets used, the black-boxes post-processed, and specifics of the TOPDOWN algorithm.

Datasets

- **German Credit.** In the SI, we additionally consider the German Credit dataset, preprocessed by AIF360 (Bellamy et al., 2019). The dataset consists of only 1000 examples, which is the smallest of the 3 datasets considered. On the other hand, the dataset provided by AIF360 contains 57 features, primarily from one-hot encoding.
- **Bank.** Another dataset we consider in the SI is the Bank dataset, preprocessed by AIF360 (Bellamy et al., 2019). The dataset consists 30488 examples, above the German Credit dataset but below the ACS datasets. The dataset also has 57 features which is largely from one-hot encoding.
- **ACS.** The American Community Survey dataset is the dataset we present in the main text. More specifically, we consider the income prediction task (as depicted in the Folktables Python package (Ding et al., 2021)) over 1-year survey periods in the state of CA. Our of the 3 datasets, ACS provides the largest dataset, with 187475 examples for the 2015 sample of the dataset. Despite this, Folktables only provides 10 features for its prediction task. Through one-hot encoding, this is extended to 29 features.

Additional Z -score normalization was used where appropriate. Sensitive attributes are binned into binary and trinary modalities, as specified in the main text (and one-hot encoded for the trinary case).

Each experiment / dataset is used with 5-fold cross-validation and further split such that there are subset partitions for: (1) training the black-box; (2) training a post-processing method; and (3) testing and evaluation. In particular, we utilize standard cross-validation to split the data into a 80:20 training testing split. The training split is then split randomly equally for separate training of the black-box and post-processing method. The final data splits result in 40:40:20 partitions.

black-boxes

- **Random Forest.** As per the main text, we primarily consider a calibrated random forest classifier provided by the scikit-learn Python package. The un-calibrated random forest classifier consists of 50 decision trees in an ensemble. Each decision tree has a maximum depth of 4 and is trained on a 10% subset of the black-box training data. In calibration, 5 cross validation folds are used for Platt scaling.
- **Neural Network.** Additionally to random forests, we consider a calibrated neural network in the SI, also provided by scikit-learn. The un-calibrated neural network is trained using mostly default parameters provided by scikit-learn. The exception to this is the specification of 300 training iterations and the specification of 10% of the training set to be used for early stopping.

The black-boxes are additionally clipped to adhere to Assumption 1 with $B = 1$ for all sections except for Appendix XIII.

TOPDOWN Specifics

The α -trees learnt by TOPDOWN are initialized as per Fig. 3. That is, we initialize sub- α -trees with $\alpha = 1$ for each of the modalities of the sensitive attribute. In addition, each split of the α -tree consists of projects to a specific feature / attribute. The split is either a modality of the discrete feature or a single linear threshold of a continuous feature. In addition, to avoid over-fitting we restrict splits to only those which result in children node that have at least 10% of the parent node's examples; and at a minimum have at least 30 examples for each child node.

⁵Public at: github.com/Trusted-AI/AIF360

IX. Additional Main Text Experiments

In this section, we report the experiments of those presented in the main text for the additional German Credit and Bank datasets. We additionally present any missing sensitive attribute modalities missing. Figs. 7 and 8 presents equivalent plots for Fig. 4 in the main text for the German Credit and Bank datasets.

Fairness Models

In comparison to ACS, Fig. 8 for the Bank dataset performs similarly to the main text figure. There are only slight deviations in the ordering of which TOPDOWN settings perform best. For example, the CVAR optimization of audacious and conservative updates are a lot closer in the Bank dataset than that of the ACS 2015 dataset.

In comparison, the result's of TOPDOWN on the German Credit largely deviate from that of the other experiments. This can be clearly seen in the number of boosting iteration TOPDOWN completes being significantly lower before the entropy stops being decreased (and thus terminating the algorithm). Another major deviation is that CVAR fails to get lowered for both binary and trinary sensitive attribute modalities in the German Credit dataset. Despite this, EOO and SP both have slight improvements for the best corresponding TOPDOWN setting (conservative EOO and conservative SP ↑), which is consistent with other datasets. This is despite the original classifier's EOO and SP being significantly lower than the ACS dataset. However, there is a major cost in the case of EOO, where the accuracy (both for MD and AUC) is harmed significantly.

A reason for the significantly worse performance, predominantly in CVAR optimization, of TOPDOWN for the German Credit is likely the significantly smaller number of example available in the dataset. Given that there are only 1000 examples and 57 features variables, the 40:40:20 split of the dataset results in the subsets to not be representative of the entire dataset's support. Additionally, CVAR is strongly tied to the cross-entropy loss function and empirical risk minimization (Williamson & Menon, 2019; Rockafellar & Uryasev, 2000). As such, given the nonrepresentative subsets of the dataset used for training TOPDOWN, minimizing the CVAR for low sample inputs is difficult.

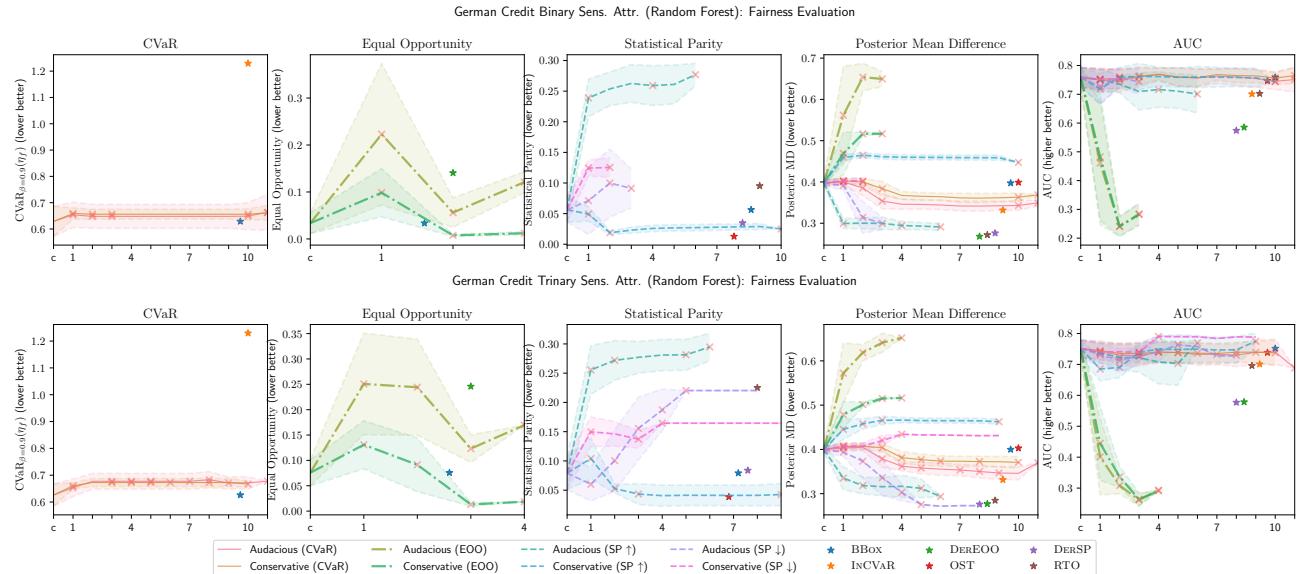


Figure 7. TOPDOWN optimized for different fairness models evaluated on German Credit with binary (up) and trinary (down) sensitive attributes. Crosses denote when a subgroup's α -tree is initiated (over any fold). The shade depicts \pm a standard deviation from the mean. However, this disappears in the case where other folds stop early.

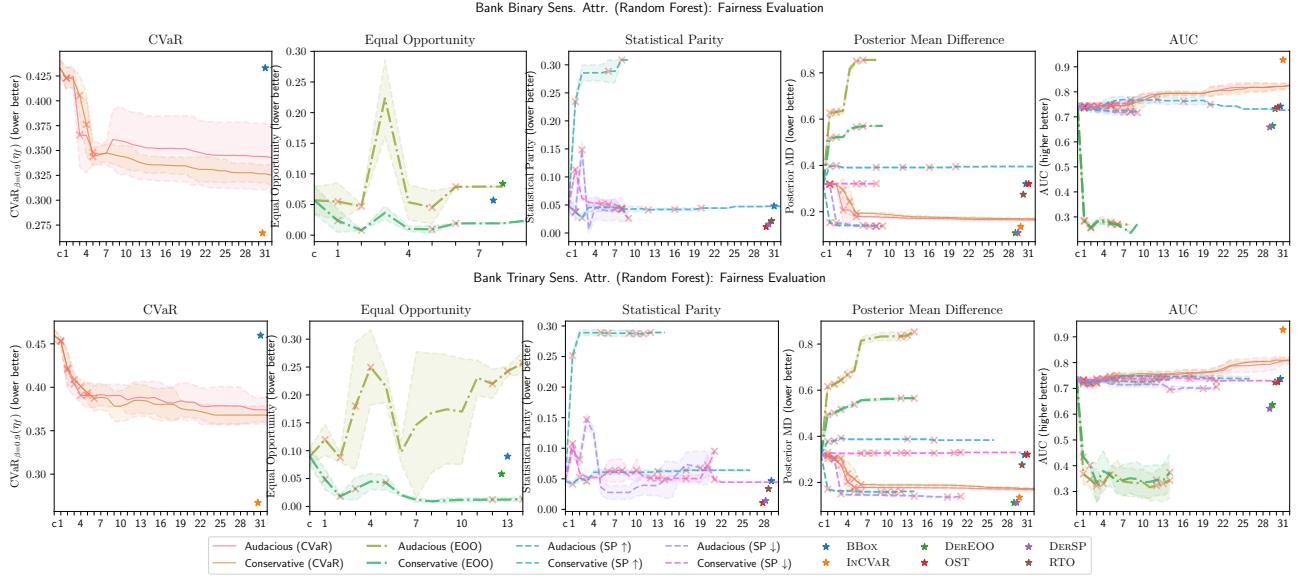


Figure 8. TOPDOWN optimized for different fairness models evaluated on Bank with binary (up) and trinary (down) sensitive attributes. Crosses denote when a subgroup’s α -tree is initiated (over any fold). The shade depicts \pm a standard deviation from the mean. However, this disappears in the case where other folds stop early.

X. Neural Network Experiments

In this SI section, we repeat all experiments evaluating different fairness models and proxy sensitive attributes using the neural network (NN) black-box. Figs. 9 to 11 presents neural network equivalent plots for all datasets to that of Fig. 4 as presented in the main text. When comparing the NN experiments to the experiments corresponding to that of the random forest (RF) black-box experiments, only minor deviation can be seen with most trends staying the same. One consistent deviation is that the CVAR criterion and accuracy measures (MD and AUC) are frequently smaller at the initial and final point of boosting. This comes from the strong representation power of the NN black-box being translated from the initial black-box to the final wrapper classifier. In this regard, switching to a NN did not help the optimization of CVAR for the German Credit dataset, see Fig. 9.

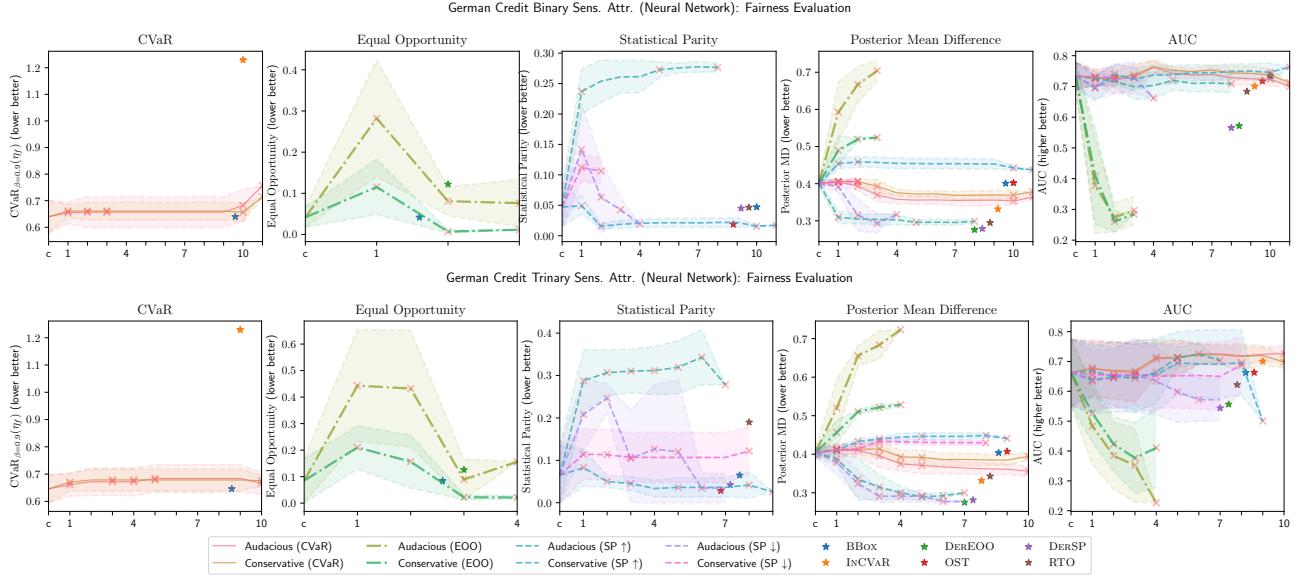


Figure 9. TOPDOWN optimized for different fairness models evaluated on German Credit with binary (up) and trinary (down) sensitive attributes. Crosses denote when a subgroup’s α -tree is initiated (over any fold). The shade depicts \pm a standard deviation from the mean. However, this disappears in the case where other folds stop early.

XI. Proxy Sensitive Attributes

We examine the use of sensitive attribute proxies to remove sensitive attribute requirements at test time. In particular, we use a decision tree with a maximum depth of 8 to predict sensitive attributes (from other features) as a proxy to the true sensitive attribute.

Fig. 14 presents the RF TOPDOWN proxy sensitive attribute experiments results of the ACS 2015 dataset not present in the main text. We focus on the binary case (left). Unsurprisingly, the proxy increases the variance of CVAR and AUC whilst also being worse than their non-proxy counterparts; but still manages to improve CVAR and AUC at the end (with an initial dip quickly erased for the later criterion). Remark the non-trivial nature of the proxy approach, as growing the α -tree is based on groups learned at the decision tree leaves *but* the CVAR computation still relies on the *original* sensitive grouping.

Figs. 12 and 13 presents the RF TOPDOWN proxy sensitive attribute results of the German Credit and Bank datasets. The ACS and Bank experiments presented here are similar to that presented in the main text. For German Credit, similar degradation in CVAR in the non-proxy case can be seen for TOPDOWN results using proxy attributes.

When comparing to the MLP variants (Figs. 15 to 17), results are quite similar with slight increases in CVAR from the change in black-box. One notable difference can be seen in Fig. 17. In particular, the proxy and regular curves do not “cross”. This indicates that (given that the sensitive attribute proxy used is the same as RF) the black-box being post-processed is an important consideration in the use of proxies. In particular, as RF has a higher / worse initial CVAR, which is highly tied to the loss / cross entropy of the black-box, the robustness of the black-box needs to be considered.

Fair Wrapping for Black-box Predictions

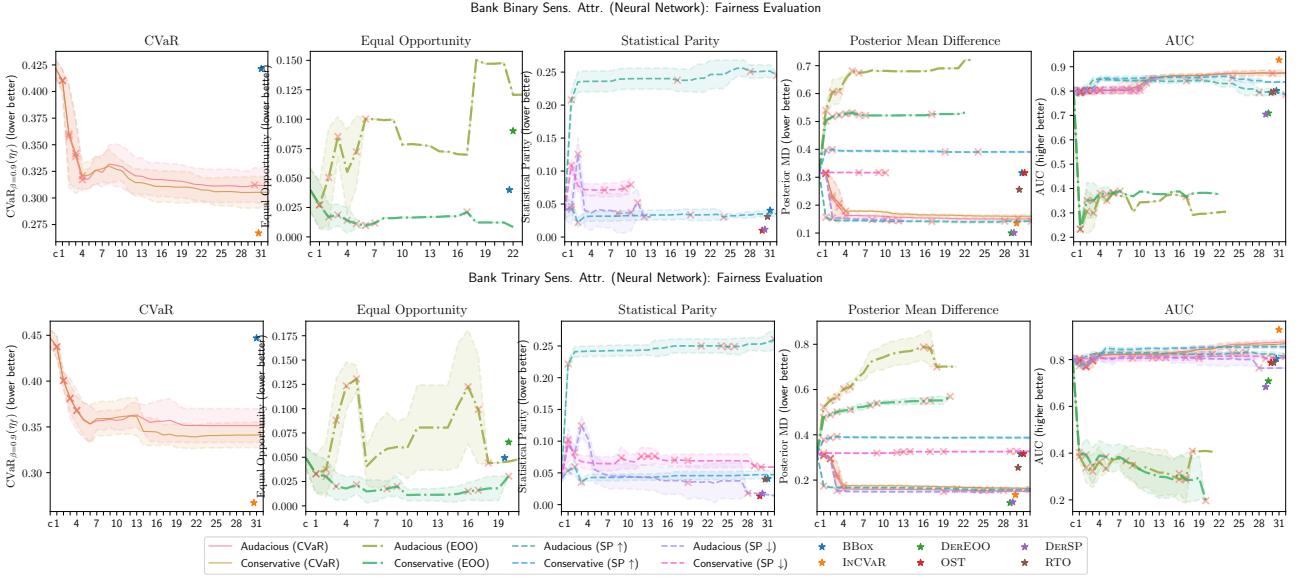


Figure 10. TOPDOWN optimized for different fairness models evaluated on Bank with binary (up) and trinary (down) sensitive attributes. Crosses denote when a subgroup’s α -tree is initiated (over any fold). The shade depicts \pm a standard deviation from the mean. However, this disappears in the case where other folds stop early.

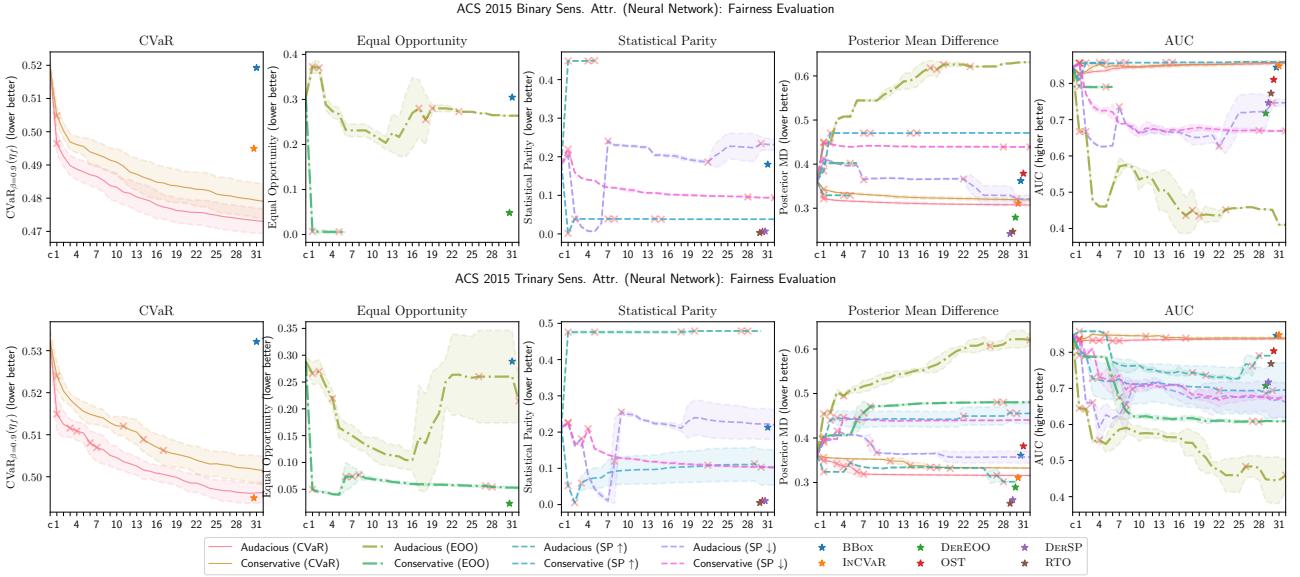


Figure 11. TOPDOWN optimized for different fairness models evaluated on ACS 2015 with binary (up) and trinary (down) sensitive attributes. Crosses denote when a subgroup’s α -tree is initiated (over any fold). The shade depicts \pm a standard deviation from the mean. However, this disappears in the case where other folds stop early.

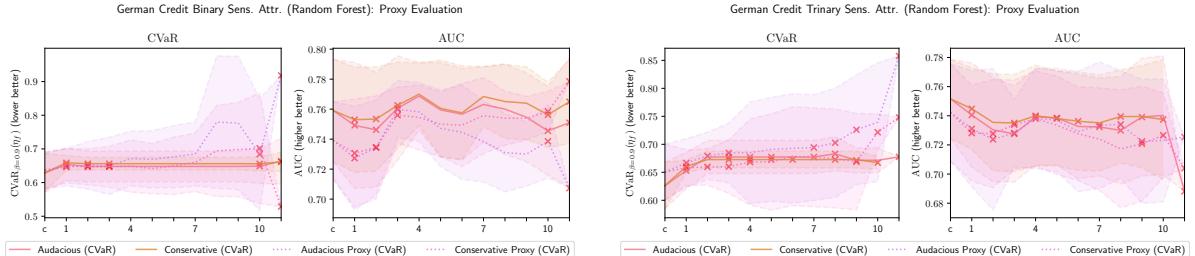


Figure 12. RF evaluation of replacing sensitive attributes with a proxy decision tree on the German Credit datasets.

Fair Wrapping for Black-box Predictions

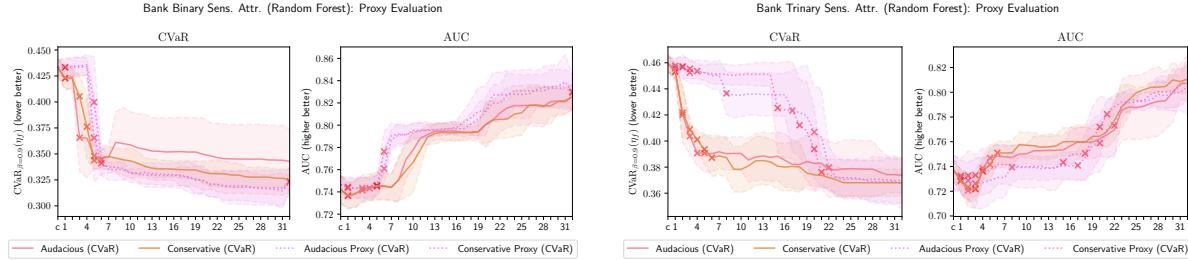


Figure 13. RF evaluation of replacing sensitive attributes with a proxy decision tree on the Bank datasets.

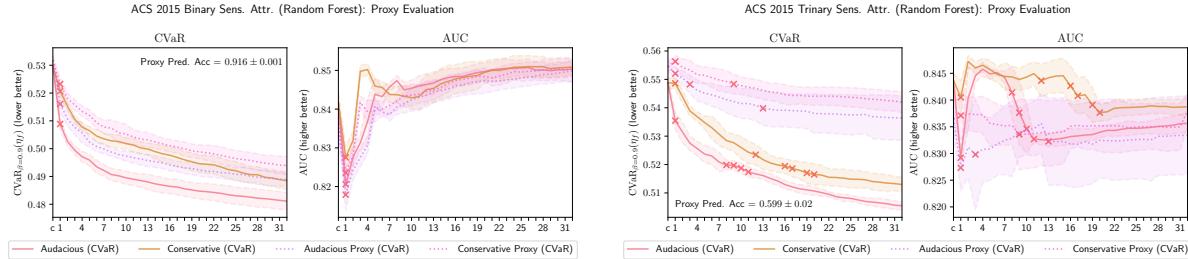


Figure 14. RF replacing sensitive attributes with a proxy decision tree on the ACS 2015 dataset (see text).

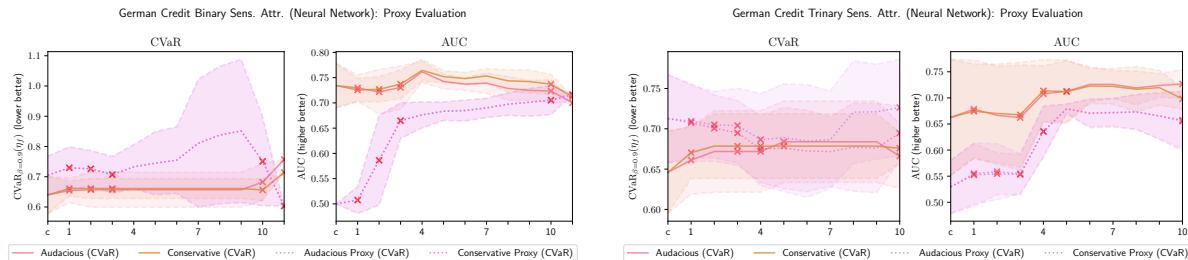


Figure 15. MLP evaluation of replacing sensitive attributes with a proxy decision tree on the German Credit datasets.

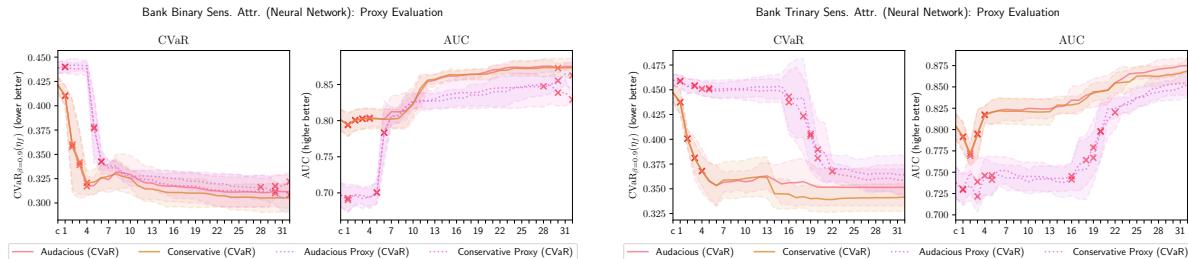


Figure 16. MLP evaluation of replacing sensitive attributes with a proxy decision tree on the Bank datasets.

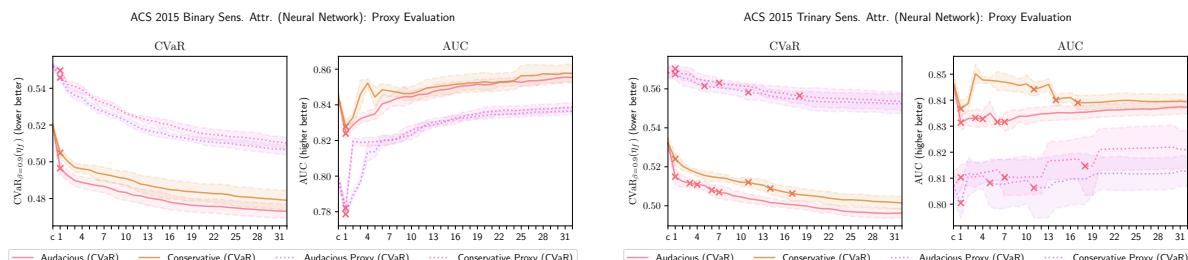


Figure 17. MLP evaluation of replacing sensitive attributes with a proxy decision tree on the ACS datasets.

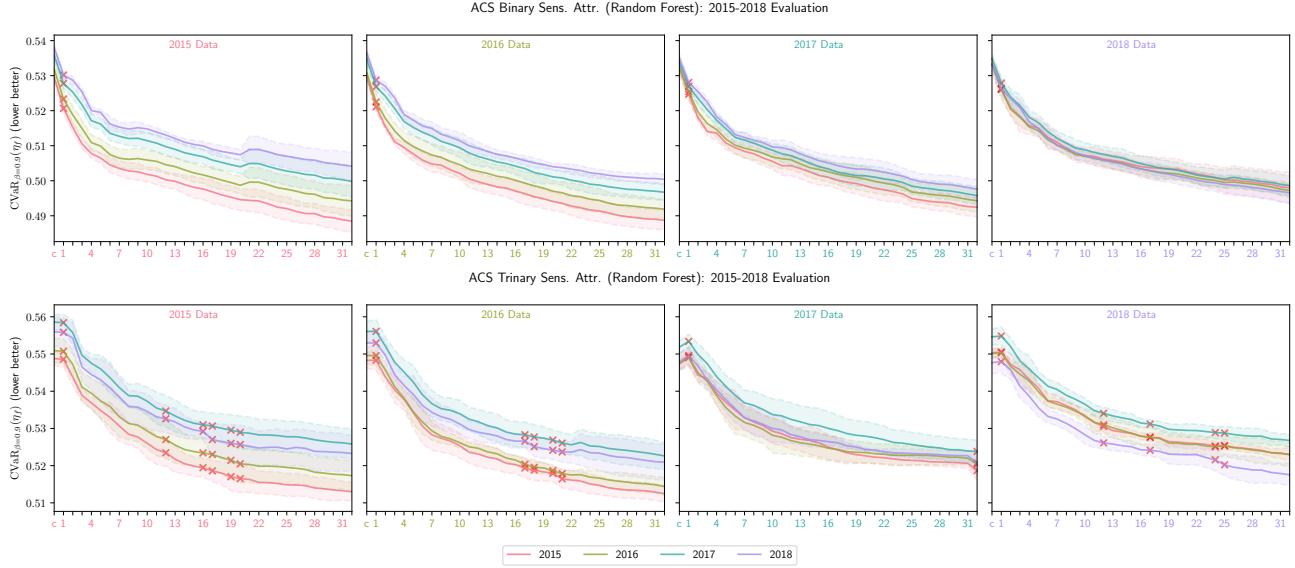


Figure 18. Random forest black-box conservative CVAR wrapper trained for ACS 2015 to 2018 datasets. Each plot is trained on a different dataset year. Each curve colour, indicates the data being used to evaluate the wrapper.

XII. Distribution shift

To examine how TOPDOWN is effected by distribution shift, we train various wrappers over multiple years of the ACS dataset. In particular, we train and evaluate CVAR wrappers over the ACS dataset from years 2015 to 2018. Figs. 18 and 19 report the CVAR values over the multiple years for the random forest (RF) black-box. Figs. 20 and 21 likewise reports corresponding results for neural network (NN) black-boxes.

As the ACS dataset consists of census data, one could expect that prior years of the data will be (somewhat) represented in subsequent years of the data. This is further emphasised in the plots, where curves become more closely group together as the training year used to train TOPDOWN increases, *i.e.*, 2018 containing enough example which are indicative of prior years’ distributions. Unsurprisingly, we can see that most circumstances the largest decrease in CVAR (mostly) comes from instances where the data matches the evaluation. *i.e.*, the 2015 curve in (top) Fig. 18. Nevertheless, we can see that despite the training data, all evaluation curves decrease from their initial values in all plots; where a slight ‘break’ in ‘monotonicity’ occurs in some instances of miss-matching data – most prominently in (top) Fig. 18 for the 2015 plot around 21 boosting iterations. We also remark, perhaps surprisingly, that there is no crossing between curves (*e.g.* as could be expected for the test-2015 and test-2016 curves on training from 2016’s data in Figure 18), but if test-2015 remains best, we also remark that it does become slightly worse for train-2016 while test-2016 expectedly improves with train-2016 compared to train-2015. Ultimately, all test-* curves converge to a ‘midway baseline’ on train-2018.

In general, there is little change when comparing the two different black-boxes. The only consist pattern in comparison is that the NN approaches start and end with a smaller CVAR value than their RF counter parts. When comparing binary versus trinary results, there is a distinct larger spread between evaluation curves (between each year within a plot) for the trinary counterparts. This is expected as in the trinary sensitive attribute modality, CVAR is sensitive to additional partitions of the dataset. The spread is further strengthened as the final α -tree in TOPDOWN often does not provide an α -correction for all subgroups, *i.e.*, at least one subgroup is not changed by the α -tree with $\alpha = 1$. When comparing conservative versus aggressive approaches, it can also be seen that there is a larger spread between evaluation curves for the aggressive variant.

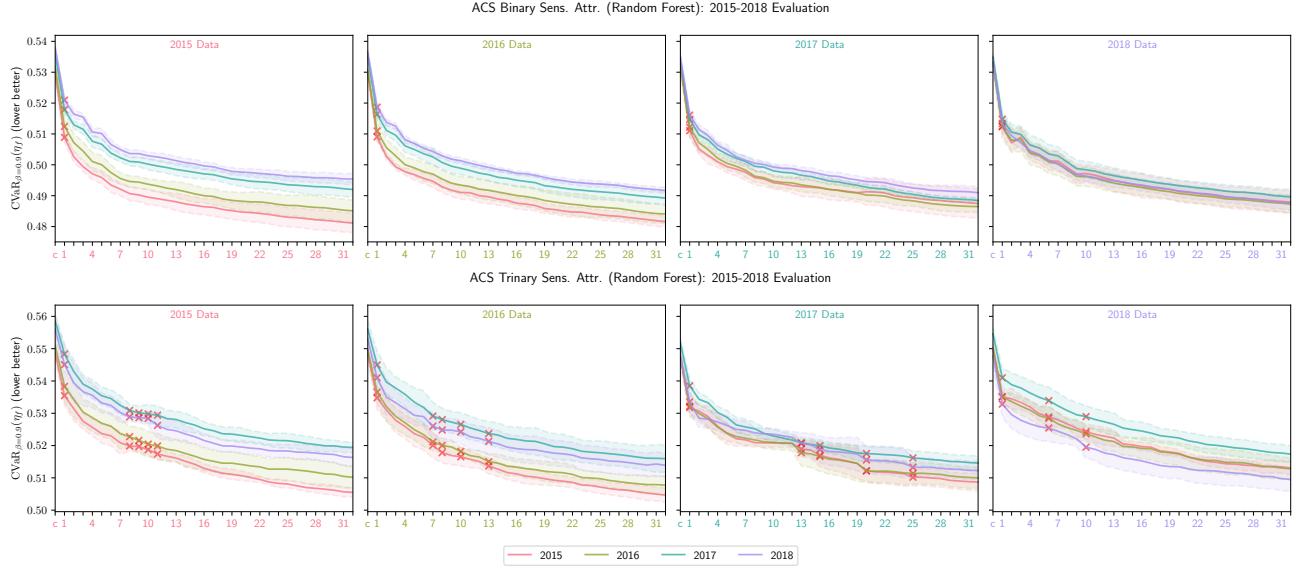


Figure 19. Random forest black-box aggressive CVAR wrapper trained for ACS 2015 to 2018 datasets Each plot is trained on a different dataset year. Each curve colour, indicates the data being used to evaluate the wrapper.

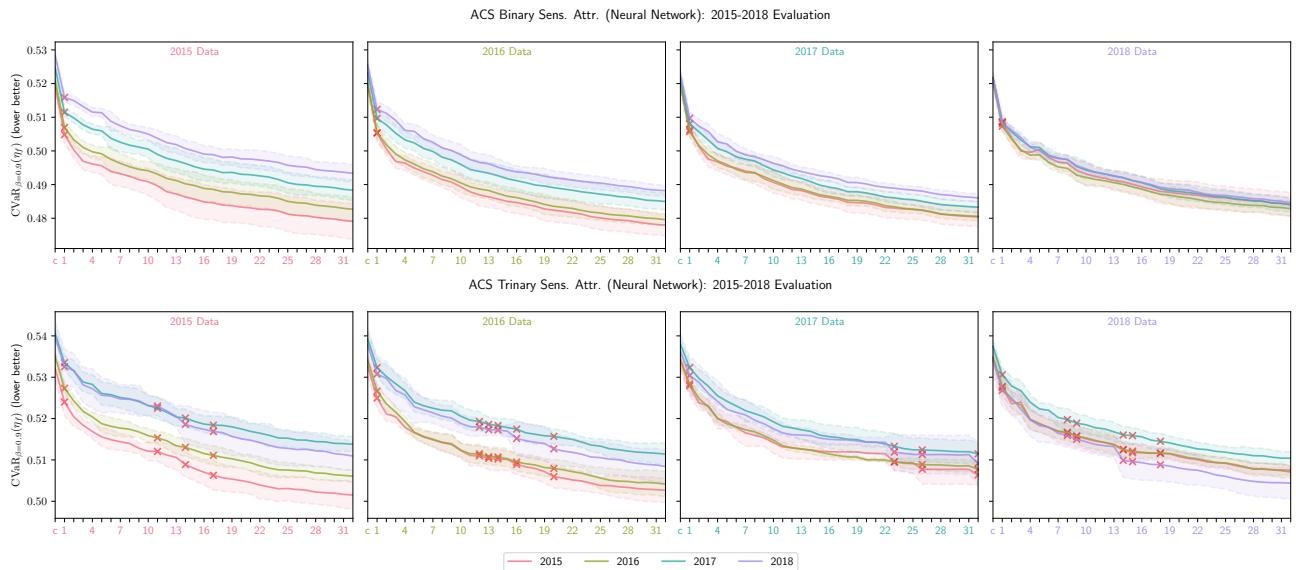


Figure 20. Neural Network black-box conservative CVAR wrapper trained for ACS 2015 to 2018 datasets Each plot is trained on a different dataset year. Each curve colour, indicates the data being used to evaluate the wrapper.

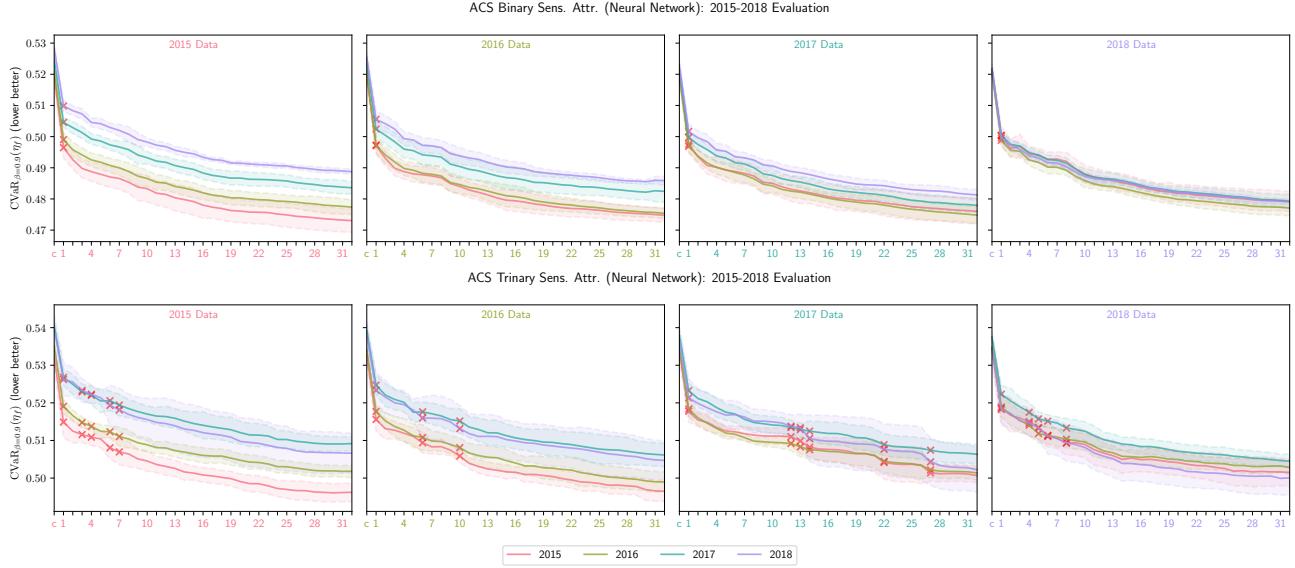


Figure 21. Neural Network black-box aggressive CVAR wrapper trained for ACS 2015 to 2018 datasets. Each plot is trained on a different dataset year. Each curve colour, indicates the data being used to evaluate the wrapper.

XIII. High Clip Value

In this section, we consider a higher clipping value than that used in other experiments. In other sections, we consider a $B = 1$ clipping value which results in posterior restricted between roughly $[0.27, 0.73]$. Although this clipping seems harsh, from the prior experiments one can see that TOPDOWN provides a lot of improvement across all fairness criterion (and we will see $B = 1$ allows TOPDOWN to improve beyond optimization for a large clip value).

We will now consider TOPDOWN experiments which correspond to evaluation over CVAR, EOO, and SP criterion with clipping $B = 3$ (as discussed in theory sections of the main text). This restricts the posterior to be between roughly $[0.05, 0.95]$. Figs. 22 to 24 presents RF plots over German, Bank, and ACS datasets; and Figs. 25 to 27 presents equivalent MLP plots. In general, there is only a slight difference between the RF and MLP plots in this clipping setting.

We focus on the RF ACS plot of the higher clipping value, Fig. 24. The most striking issue is that the minimization of CVAR is a lot worse than when using clipping $B = 1$. In particular, BBOX (which in Fig. 24 has $B = 3$) is not beaten by the final wrapped classifier produced by either update of TOPDOWN. However, for EOO and SP there is still a reduction in criterion, although a lower reduction for some cases, *i.e.*, conservative EOO. It is unsurprising that CVAR is more difficult to optimize in this case as the black-box would be closer to an optimal accuracy / cross-entropy value without larger clipping. As a result, CVAR would be more difficult to improve on as it depends on subgroup / partition cross-entropy. In particular, the large spike in the first iteration of boosting is striking. This comes from the fact that we are no directly minimizing a partition's cross-entropy directly, but an upper-bound, where the theory specifies that the upper-bound requires that the original black-box is already an α -tree with correct corrections. However, as the the original black-box is not an α -tree with correction specified by the update, the initial update can cause an increase in the CVAR (which appears to be more common with higher clipping values).

Despite the initial “jump” and in-ability to recover, let us compare the $B = 3$ plot to the original $B = 1$ RF TOPDOWN plot given in the main text, Fig. 4. From comparing the results, one can see that the final boosting iteration for the $B = 1$ aggressive updates beats the $B = 3$ black-box classifiers. Thus, even when comparing against CVAR which is highly influenced by accuracy (thus a higher clipping value is desired), a smaller clipping value resulting in a more clipped black-box posterior is potentially more useful in CVAR TOPDOWN. If one looks at the conservative curves in Fig. 4, these do not beat the $B = 3$ black-box. This further strengthens the argument that the aggressive update is preferred in CVAR TOPDOWN; and is further emphasized by the increase cap between curves with $B = 3$ black-boxes, as shown in Fig. 24.

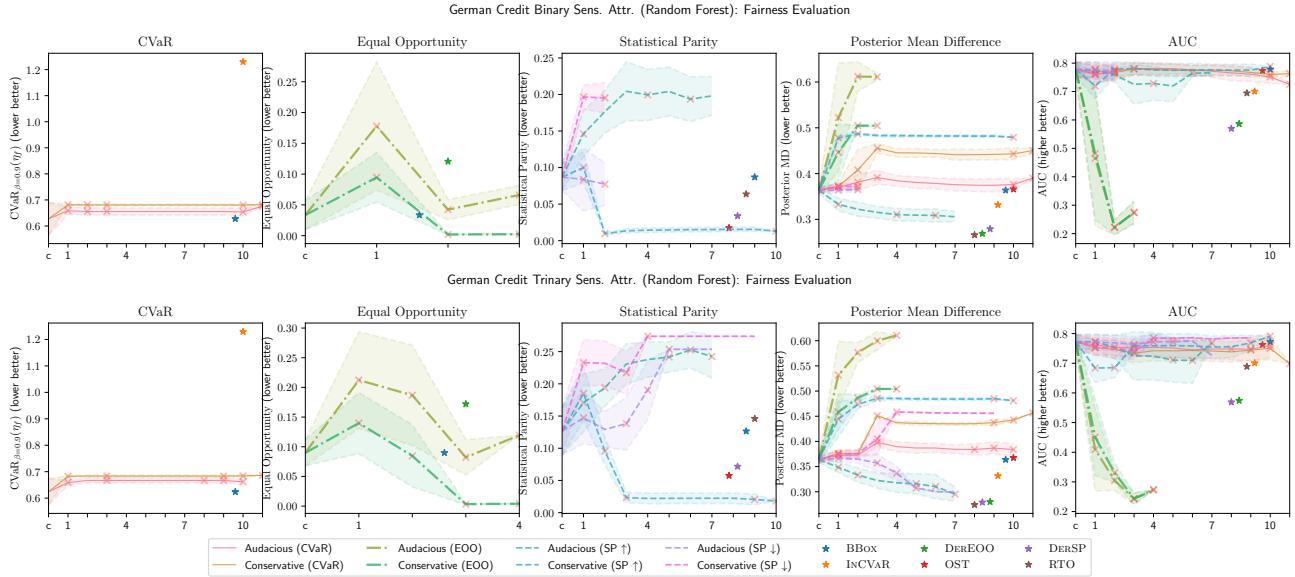


Figure 22. RF with $B = 3$ TOPDOWN optimized for different fairness models evaluated on German Credit with binary (up) and trinary (down) sensitive attributes. Crosses denote when a subgroup’s α -tree is initiated (over any fold). The shade depicts \pm a standard deviation from the mean. However, this disappears in the case where other folds stop early.

XIV. Example Alpha-Tree

In this section, we provide an example of an α -tree generated using TOPDOWN. In particular, we look at one example from training CVAR TOPDOWN on the Bank dataset with binary sensitive attributes. Fig. 28 presents the example α -tree. The tree contains information about the attributes in which splits are made and the α -correction made at leaf nodes (and their induced partition). In the example, could note that the α trees for modalities of the age sensitive attribute are imbalanced. The right tree is significantly smaller than the left. One could also note the high reliance on “education” based attributes for determining partitions. These factors could be used to scrutinise the original blackbox; and eventually, even provide constraints on the growth of an α -tree which would aim to avoid certain combinations of attribute. We leave these factors for future work.

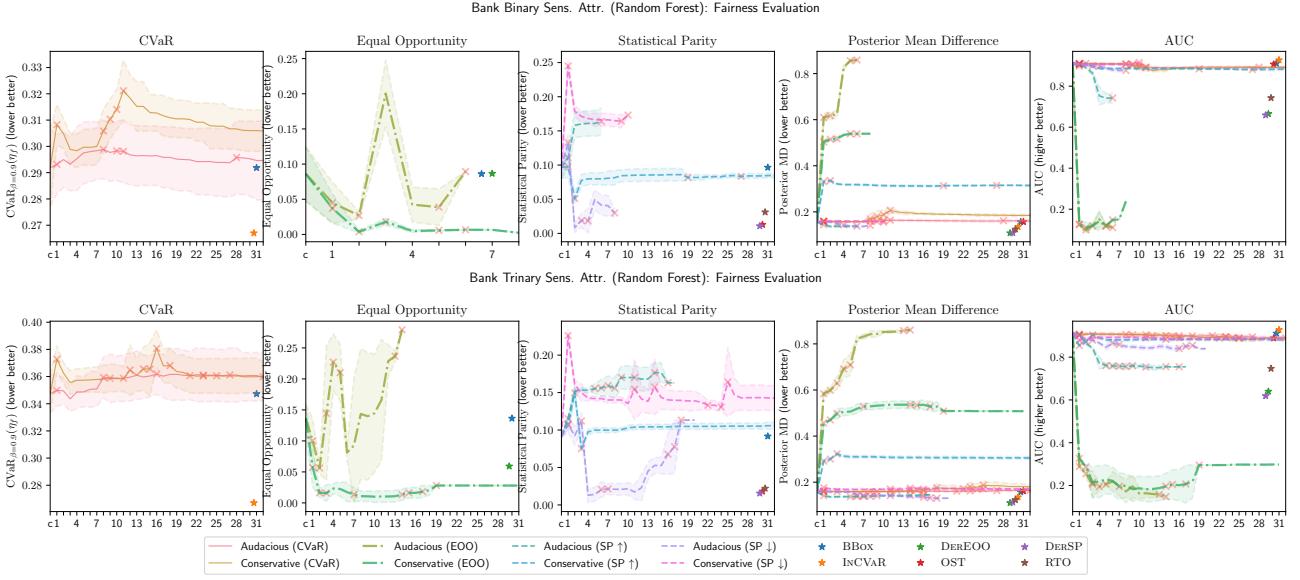


Figure 23. RF with $B = 3$ TOPDOWN optimized for different fairness models evaluated on Bank with binary (up) and trinary (down) sensitive attributes. Crosses denote when a subgroup’s α -tree is initiated (over any fold). The shade depicts \pm a standard deviation from the mean. However, this disappears in the case where other folds stop early.

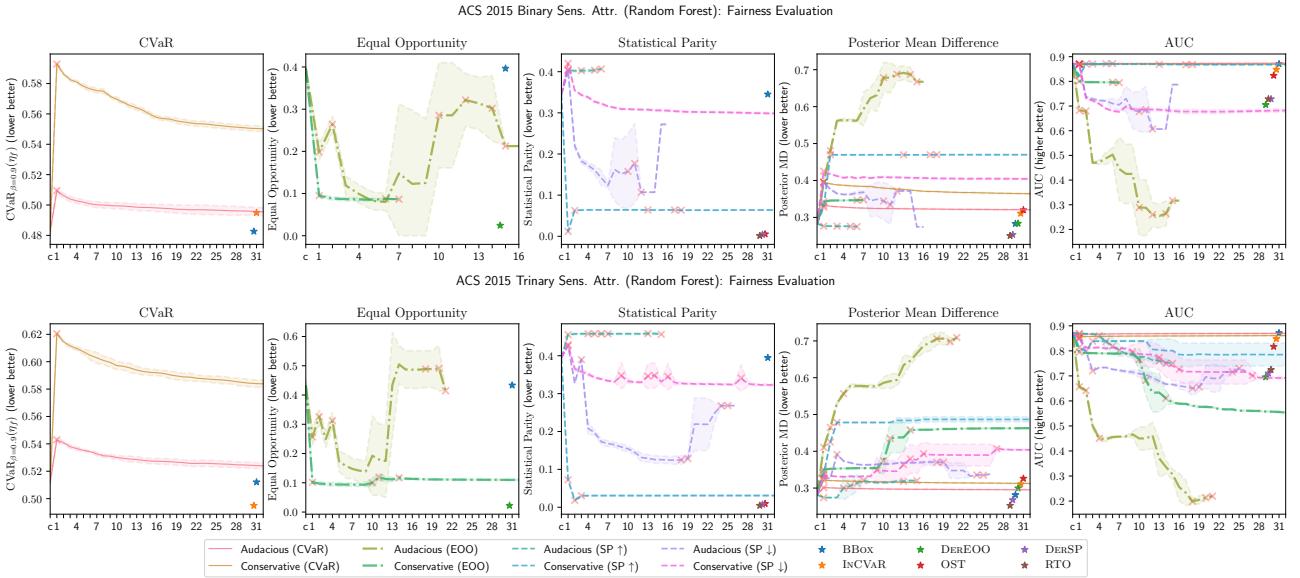


Figure 24. RF with $B = 3$ TOPDOWN optimized for different fairness models evaluated on Bank with binary (up) and trinary (down) sensitive attributes. Crosses denote when a subgroup’s α -tree is initiated (over any fold). The shade depicts \pm a standard deviation from the mean. However, this disappears in the case where other folds stop early.

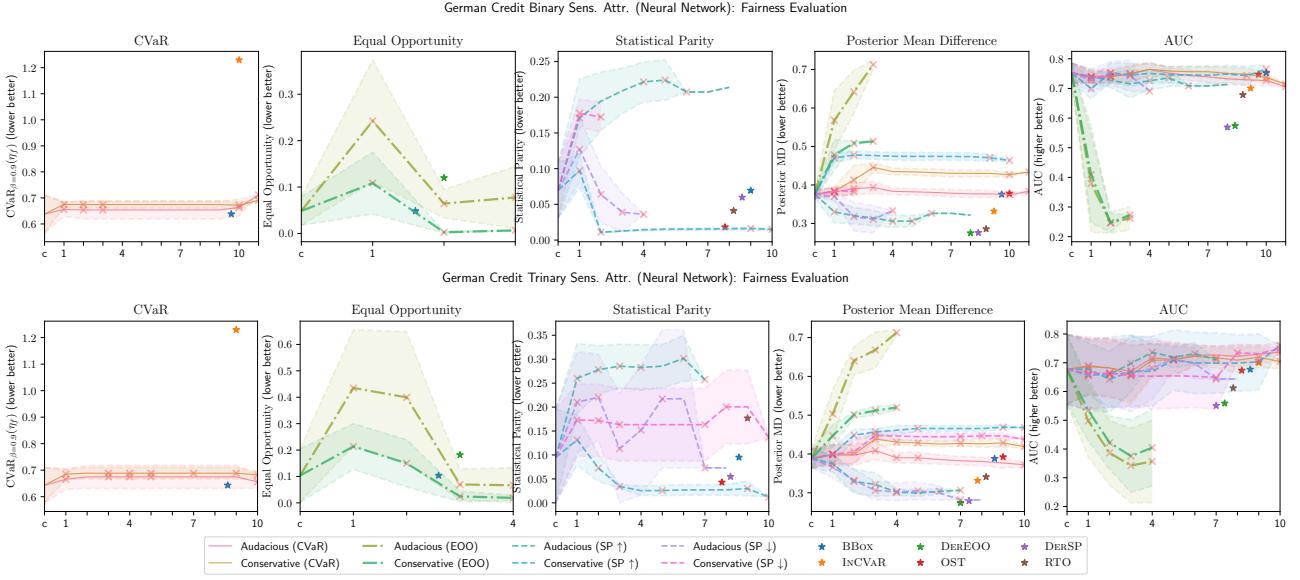


Figure 25. MLP with $B = 3$ TOPDOWN optimized for different fairness models evaluated on German Credit with binary (up) and trinary (down) sensitive attributes. Crosses denote when a subgroup’s α -tree is initiated (over any fold). The shade depicts \pm a standard deviation from the mean. However, this disappears in the case where other folds stop early.

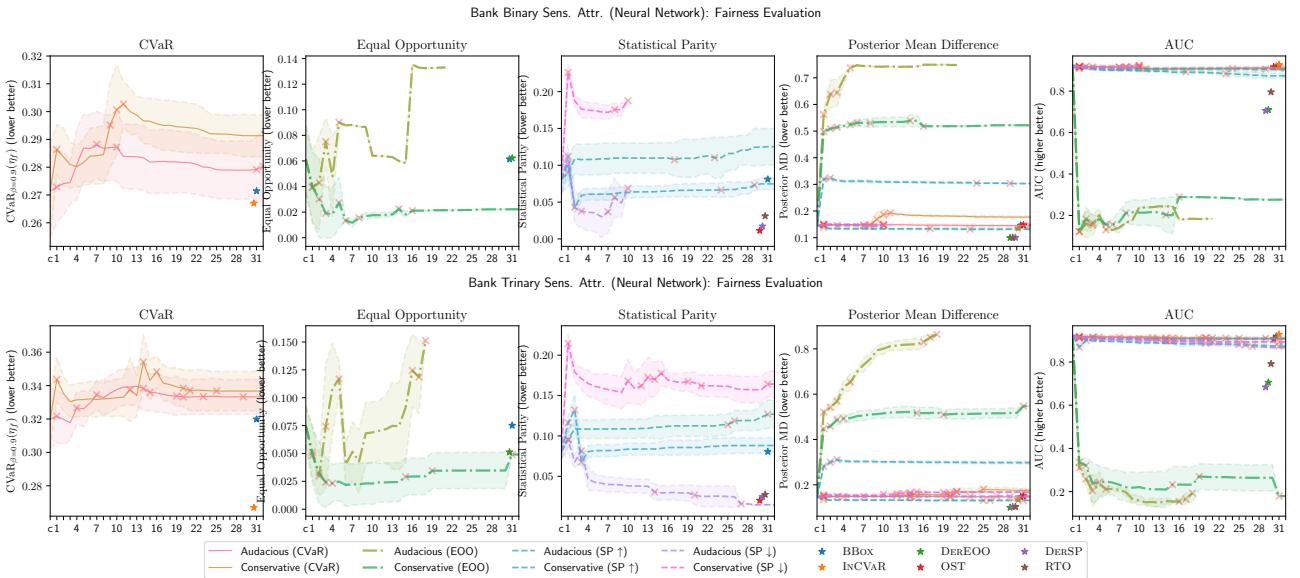


Figure 26. MLP with $B = 3$ TOPDOWN optimized for different fairness models evaluated on Bank with binary (up) and trinary (down) sensitive attributes. Crosses denote when a subgroup’s α -tree is initiated (over any fold). The shade depicts \pm a standard deviation from the mean. However, this disappears in the case where other folds stop early.

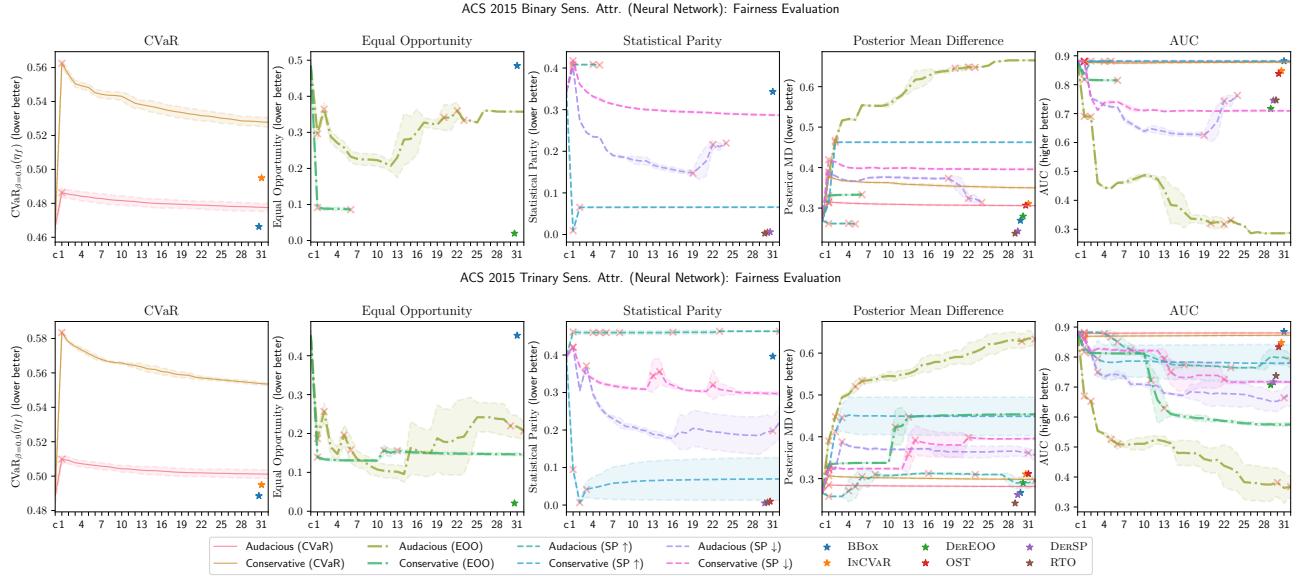


Figure 27. MLP with $B = 3$ TOPDOWN optimized for different fairness models evaluated on Bank with binary (up) and trinary (down) sensitive attributes. Crosses denote when a subgroup’s α -tree is initiated (over any fold). The shade depicts \pm a standard deviation from the mean. However, this disappears in the case where other folds stop early.

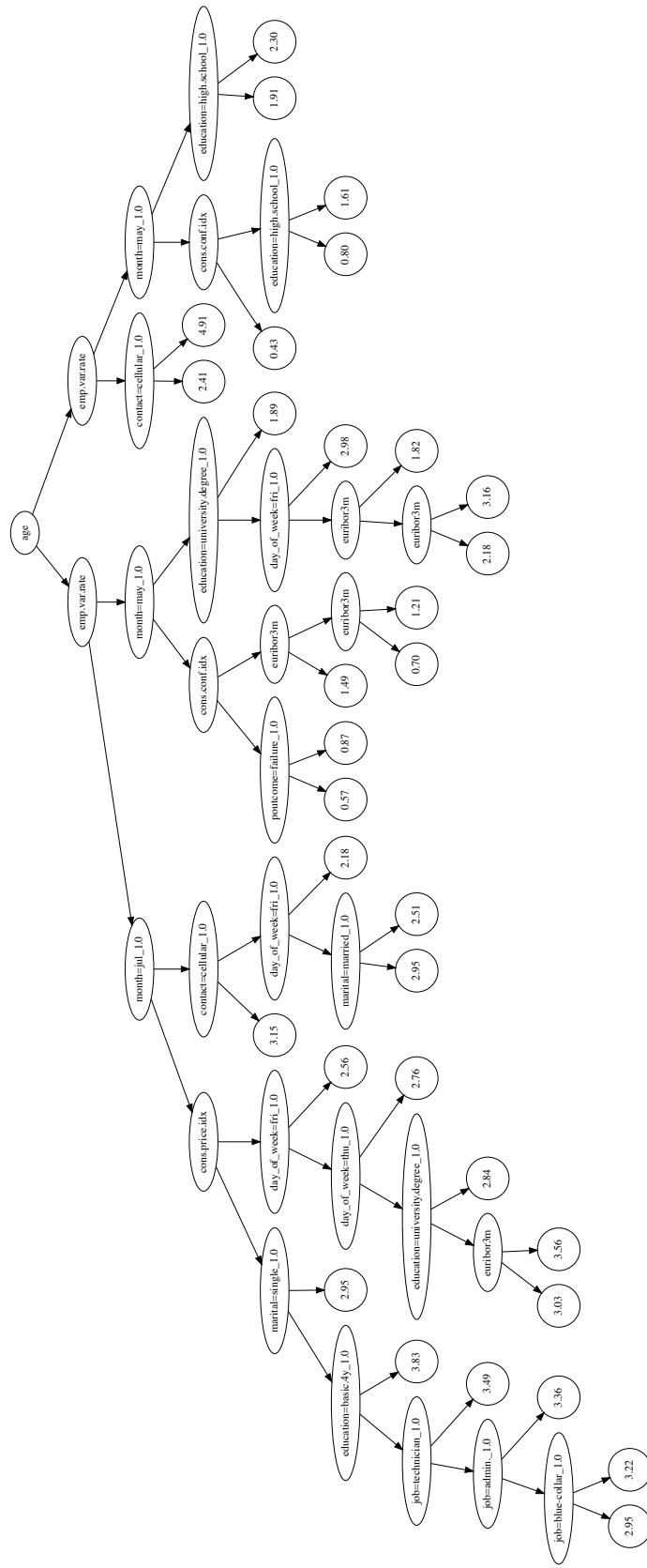


Figure 28. Example tree generated in the optimization of TopDown for CVAR in the Bank dataset with binary sensitive attributes.