

# Fair Wrapping for Black-box Predictions

Alexander Soen<sup>†</sup> Ibrahim Alabdulmohsin<sup>‡</sup> Sanmi Koyejo\*, ‡  
 Yishay Mansour<sup>‡, ◇</sup> Nyalleng Moorosi<sup>‡</sup> Richard Nock<sup>‡, †</sup>  
 Ke Sun<sup>†, ◇</sup> Lexing Xie<sup>†</sup>

Australian National University<sup>†</sup> Google Research<sup>‡</sup>  
 Stanford University<sup>\*</sup> Tel Aviv University<sup>◇</sup>  
 Data61/CSIRO<sup>◇</sup>

## 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 optimization 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  $\alpha$ -tree, which modifies the prediction. We provide two generic boosting algorithms to learn  $\alpha$ -trees. We show that our modification has appealing properties in terms of composition of  $\alpha$ -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

The social impact of Machine Learning (ML) has seen a dramatic increase over the past decade – enough so that the bias of model outputs must be accounted for alongside accuracy (Alabdulmohsin & Lucic, 2021; Hardt et al., 2016; Zafar et al., 2019). Considering the various number of fairness targets (Mehrabi et al., 2022) and the energy and CO2 footprint of ML (Martineau, 2020; Strubell et al., 2019), the combinatorics of training accurate *and* fair models is non-trivial. This is especially so given the inherit incompatibilities of fairness constraints (Kleinberg et al., 2017) and the underlying tension of satisfying fairness whilst maintaining accuracy. One trend in the field ”decouples” the two constraints by post-processing *pretrained* (accurate) models to achieve fairer outputs (Zafar et al., 2019). Post-processing may be the only option if we have no access to the model’s training data / algorithm / hyperparameters (etc.).

Within the post-processing approach, three trends have emerged: learning a new fair model close to the black-box, tweaking the output subject to fairness constraints, and exploiting sets of classifiers. If the task is class probability estimation (Reid & Williamson, 2011), the

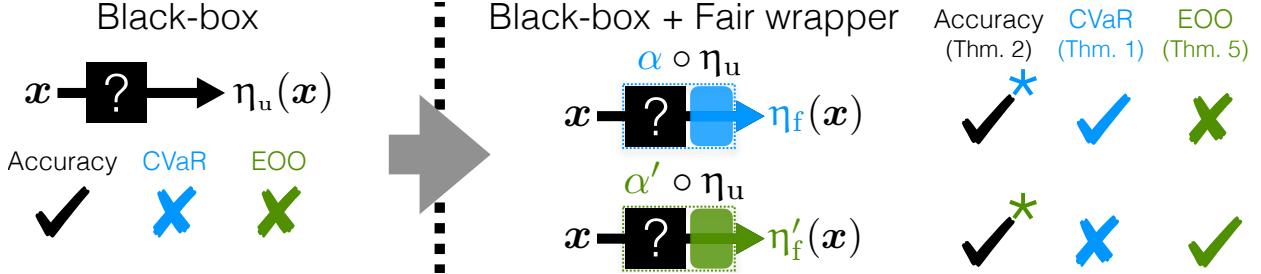


Figure 1: Summary of using different  $\alpha$ -correction wrappers to obtain different fairness criteria guarantees. See Sections 5 and 6 for full details on guarantees.

estimated black-box is an accurate but potentially unfair posterior  $\eta_u : \mathcal{X} \rightarrow [0, 1]$  which neither can be opened nor trained further. The goal is then to learn a fair posterior  $\eta_f$  from it. In addition to the black-box constraint, a number of *desiderata* can be considered for post-processing. Ideally in correcting a black-box, we would want the approach to have (**flexibility**) in satisfying substantially different fairness criteria, (**proximity**) of the learnt  $\eta_f$  to the original  $\eta_u$ , and meaningful (**composability**) properties if, *e.g.*,  $\eta_f$  was later treated as a new black-box to be post-processed. To facilitate a specific style of correction, we may also want the representation of the correction to facilitate (**explainability**) for auditing the post-processing procedure and bounds on the increased model (**complexity**) of the final classifier  $\eta_f$ . In training the correction, algorithmically we would also want guarantees for (**convergence**).

**Our contribution** satisfies the aforementioned desiderata in its correction, representation, and algorithmic guarantees. By leveraging recent theory in *improper* loss functions, we utilize a *universal* correction of black-box posteriors defined by the  $\alpha$ -loss. This allows for a flexible correction which yields convenient divergence bounds between  $\eta_f$  and  $\eta_u$ , a convenient form for the Rademacher complexity of the class of  $\eta_f$ , and a simple composability property. Representation-wise, the corrections we learn are easy-to-interpret tree-shaped functions that we define as  $\alpha$ -*trees*. Algorithmically speaking, we provide two formal boosting algorithms to learn  $\alpha$ -trees building upon seminal results (Kearns & Mansour, 1996). We demonstrate our algorithm for conditional value-at-risk (CVAR) (Williamson & Menon, 2019), equality of opportunity (EOO), and statistical parity; as depicted in Fig. 1. Experiments are provided against five baselines on readily available datasets. All proofs and more experiments are in an Appendix denoted as SI.

## 2 Related Work

Post-processing models to achieve fairness is one of three different categories in tackling the ML + fairness challenge (Zafar et al., 2019, Section 6.2). Although other notions exist, *e.g.* individual fairness (Dwork et al., 2012a), we limit our analysis to group fairness, which concerns itself with ensuring that statistics of sub-populations are similar. We further segment this cluster into 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 for fairer outcomes (Alabdulmohsin &

Lucic, 2021; Hardt et al., 2016; Lohia et al., 2019; Menon & Williamson, 2018; Woodworth et al., 2017; Yang et al., 2020); and (III) techniques consisting of exploiting sets of models to achieve fairness (Dwork et al., 2018). None of these approaches formulates substantial guarantees on all of the desiderata in the introduction. Some bring contributions with the **(flexibility)** of being applicable to more than two fairness notions (Corbett-Davies et al., 2017; Wei et al., 2020; Dwork et al., 2018; Yang et al., 2020). 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). However, for all of them, the algorithmic price-tag is unclear (Corbett-Davies et al., 2017; Dwork et al., 2018) or heavily depends on convex optimization routines (Wei et al., 2020). Alabdulmohsin & Lucic (2021); Yang et al. (2020) provide strong guarantees regarding **(proximity)**, w.r.t. *consistency and generalization*. To our knowledge, our approach of correcting prediction unfairness through improper losses (Sypherd et al., 2022) is new.

### 3 Setting and Motivating Example

Let  $\mathcal{X}$  be a domain of observations,  $\mathcal{Y} \doteq \{-1, 1\}$  be labels and  $S$  is a sensitive attribute in  $\mathcal{X}$ . We assume that the modalities of  $S$  induce a partition of  $\mathcal{X}$ . We further let  $D$  denote the joint measure over  $\mathcal{X} \times \mathcal{Y}$ ,  $M$  denote the marginal measure over  $\mathcal{X}$ , and  $\pi \doteq \mathbb{P}[\mathbf{Y} = 1]$  being the prior. We denote conditioning of  $M$  through a subscript, e.g.,  $M_s$  for  $s \in S$  denotes  $M$  conditioned on a sensitive attribute subgroup  $S = s$ . We leave the  $\sigma$ -algebra to be implicit (which is assumed to be the same everywhere). As is often assumed in ML, sampling is i.i.d.; we make no notational distinction between empirical and true measures to simplify exposition – most of our results apply for both.

Consider the task of learning a function  $\eta \in [0, 1]^{\mathcal{X}}$  to estimate the true posterior  $\eta^*(\mathbf{x}) = \mathbb{P}[\mathbf{Y} = 1 | \mathbf{X} = \mathbf{x}]$  in binary classification. For instance, we may want to predict the probability of hiring an applicant for a company. Given the pointwise loss  $L(\eta(\mathbf{X}), \eta^*(\mathbf{X}))$ , which determines the loss of a single example, the *(total) risk* is defined as

$$L(\eta; M, \eta^*) \doteq \mathbb{E}_{\mathbf{X} \sim M} [L(\eta(\mathbf{X}), \eta^*(\mathbf{X}))] \quad (1)$$

(with slight abuse of notation). A low risk corresponds to good classification performance. In this paper, we consider the risk determined by the *log-loss*:

$$L(\eta; M, \eta^*) \doteq \mathbb{E}_{\mathbf{X} \sim M} [\eta^*(\mathbf{X}) \cdot -\log \eta(\mathbf{X}) + (1 - \eta^*(\mathbf{X})) \cdot -\log(1 - \eta(\mathbf{X}))]. \quad (2)$$

We now consider a simple fairness problem, centered around the example of Fig. 2. Suppose that we are given a black-box  $\eta_u$  which predicts hiring probabilities without considering fairness. Although the minimized total risk of (1) might be small, there can be discrepancies in the performance between different subgroups. Instead of considering the total risk, the predictive performance of specific subgroups can be examined through the *subgroup risk*  $L(\eta_u; M_s, \eta^*)$ , for a subgroup  $s \in S$ . For instance, we might want to examine the discrepancy of subgroup risks among the age of applications. A natural fairness task would be to improve the worst performing subgroup, say  $s_w \in S$ .

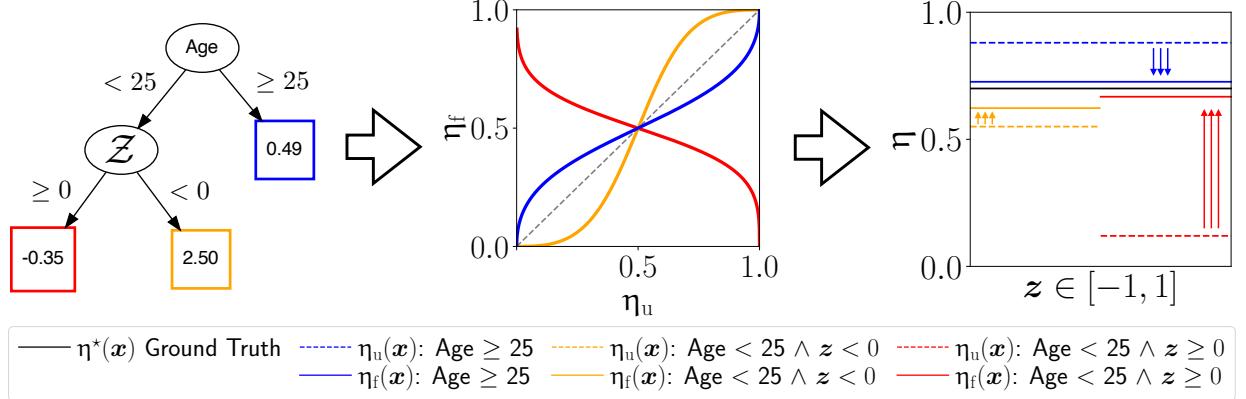


Figure 2: Improving CVAR for a toy hiring task with  $\alpha$ -trees. An  $\alpha$ -tree (left) transforms the posterior via (3) (middle); resulting in an input-dependent fairness correction of the posterior  $\eta_u$  (right).

To post-process unfairness, we want to learn a function  $\alpha : \mathcal{X} \rightarrow \mathbb{R}$  which “wraps”  $\eta_u$  and *lowers* the worst subgroup risk. We propose the following “wrapping”, inspired by improper loss functions (Sypherd et al., 2022)

$$\eta_f(\mathbf{x}) \doteq \frac{\eta_u(\mathbf{x})^{\alpha(\mathbf{x})}}{\eta_u(\mathbf{x})^{\alpha(\mathbf{x})} + (1 - \eta_u(\mathbf{x}))^{\alpha(\mathbf{x})}} \in [0, 1]. \quad (3)$$

Notice when  $\alpha(\mathbf{x}) = 1$  the resulting posterior is the original  $\eta_u(\mathbf{x})$ . Importantly, (3) is flexible enough to transform any input black-box  $\eta_u$  to any needed  $\eta_f$ . Looking at Fig. 2 (left and middle), intuitively by setting different  $\alpha(\mathbf{x})$  values, (3) “sharpens” (yellow,  $\alpha > 1$ ), “dampens” (blue,  $0 < \alpha < 1$ ), or “polarity reverses” (red,  $\alpha < 0$ ) the original posterior  $\eta_u$ . To improve fairness, we need a combination of these corrections to accommodate different subsets of the input domain (thus learning  $\alpha(\cdot)$  as a function). We specifically learn  $\alpha(\cdot)$  to be a tree structure, which allows an interpretable correction alongside other formal properties (Section 5).

**Definition 1.** An  $\alpha$ -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  $\alpha$ -tree  $\Upsilon$ . An  $\alpha$ -tree induces a correction over posteriors as per (3) with  $\alpha = \Upsilon$ .

Our fairness problem is now learning an  $\alpha$ -tree  $\eta$  which provides a corresponding correction that improves the worst subgroup risk, *i.e.*,  $L(\eta_u; M_{sw}, \eta^*) > L(\eta_f; M_{sw}, \eta^*)$ . The entirety of Fig. 2 presents such a process. In this hiring task example, the ground truth hiring rate is constant *w.r.t.* the inputs,  $\eta^*(\mathbf{x}) = 0.7$ . Despite this, the black-box  $\eta_u(\cdot)$  unfairly depends on the age of applicants and incorrectly depends on a noise feature  $Z$ . By choosing  $\alpha(\mathbf{x})$  as per Fig. 2 (left), the correction changes  $\eta_u$  to be closer to  $\eta^*$ , improving the risk of the worst subgroup (alongside the other subgroup in this example): the worse-case loss improves from 1.09 to 0.62.

Although the fairness criteria discussed might be considered simple, the procedure of iteratively minimizing the worse performing subgroup can be used to improve the *conditional*

---

**Algorithm 1** TOPDOWN ( $M_t, \eta_t, \Upsilon_0, B$ )

---

**Input** mixture  $M_t$ , posterior  $\eta_t$ ,  $\alpha$ -tree  $\Upsilon_0$ ,  $B \in \mathbb{R}_{+*}$ ;

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

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

    Step 2.1 : pick leaf  $\lambda^* \in \Lambda(\Upsilon)$ ; // i.e. heaviest leaf

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

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

Step 3 : label leaves  $\forall \lambda \in \Lambda(\Upsilon)$ :

$$\Upsilon(\lambda) \doteq \tilde{\iota} \left( \frac{1 + e(M_\lambda, \eta_t)}{2} \right), \quad // \alpha\text{-value}(5)$$

**Output**  $\Upsilon$ ;

---

*value-at-risk* (lower is better) fairness criteria (Williamson & Menon, 2019):

$$CVAR_\beta(\eta_f) \doteq \mathbb{E}_S [L(\eta_f; M_S, \eta^*) \mid L(\eta_f; M_S, \eta^*) \geq L_\beta], \quad (4)$$

where  $L_\beta$  is the risk value for the  $\beta$  quantile among subgroups, which is user-defined. The difference is that  $CVAR_\beta$  not only considers the worse case subgroup, but all subgroups above the  $L_\beta$  risk value (let these subgroups be  $S_\beta$ ). One can simply iteratively improve all  $s \in S_\beta$ , as done in the example of Fig. 2. Indeed, in the example,  $CVAR_\beta$  with  $\beta = 0.9$  improves from 1.09 to 0.62 (which is equivalent to worse-case loss in this case).

## 4 Growing Alpha-Trees

We now introduce the procedure to grow an  $\alpha$ -tree via a boosting algorithm TOPDOWN, Algorithm 1. TOPDOWN can be thought of as a generalization of the standard decision tree induction used for classification (Kearns & Mansour, 1996). We first introduce relevant concepts from decision tree induction to explain TOPDOWN. We contextualize TOPDOWN through its application in improving the CVAR criteria.

We first introduce a technical assumption for the black-box  $\eta_u$  to be post-processed:

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

$$Im(\eta_u) \subseteq \mathbb{I} \doteq [(1 + \exp(B))^{-1}, (1 + \exp(-B))^{-1}] \quad (a.s.). \quad (6)$$

Compliance with Assumption 1 can be done by clipping the black-box's output with a user-fixed  $B$  or making sure it is calibrated and then finding  $B$ .

**Entropy-based updates for fairness:** An important component in standard decision tree induction is the *edge* function, which measures the *label purity* (proportion of positive examples) of a decision tree node. We introduce a generalization which considers the *alignment purity* of a black-box.

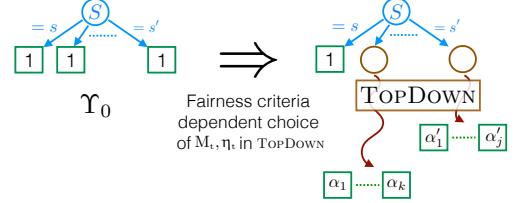


Figure 3: Picking  $\Upsilon_0$  a stump on the fairness attribute allows to finely tune growths of sub- $\alpha$ -trees to the fairness criterion at hand.

**Definition 2.** Let  $\iota(u) \doteq \log(u/(1-u))$  the logit of  $u \in (0, 1)$  and  $\tilde{\iota}(u) \doteq \iota(u)/B$  a normalization which satisfies  $\tilde{\iota}(\mathbb{I}) = [-1, 1]$ . The **alignment edge** of  $M_t$  and  $\eta_t$  is defined as,

$$e(M_t, \eta_t) \doteq \mathbb{E}_{(X,Y) \sim D_t} [Y\tilde{\iota}(\eta_u(X))], \quad (7)$$

where  $D_t$  is the joint measure induced by  $M_t$  and  $\eta_t$ . With Assumption 1,  $e(M_t, \eta_t) \in [-1, 1]$ .

By replacing the normalized logit  $\tilde{\iota}$  with a constant 1, (7) reduces to a measure of label purity used in regular classification. In our case, (7) measures how well the black-box  $\eta_u$  “aligns” with the true labels  $Y$  through the logit. This also takes into account the “confidence” of the black-box’s predictions: for a high alignment purity, predictions not only need be correct but also to be highly confident ( $\eta_u$  close to the endpoints of  $\mathbb{I}$ ). Similar to how the splits of a decision tree classifier are determined by the entropy of a tree’s label purity, an  $\alpha$ -tree splits based on its alignment entropy.

**Definition 3.** Given an  $\alpha$ -tree  $\Upsilon$  with leafset  $\Lambda$ , when Assumption 1 is satisfied, the **entropy of  $\Upsilon$**  is:

$$H(\Upsilon; M_t, \eta_t) \doteq \mathbb{E}_{\lambda \sim M_{\Lambda(\Upsilon)}} [H_1(\lambda; M_t, \eta_t)], \quad (8)$$

where  $H(q) \doteq -q \log(q) - (1-q) \log(1-q)$ ,  $H_1(\lambda; M_t, \eta_t) \doteq H((1+e(M_\lambda, \eta_t))/2)$ ,  $M_\lambda$  is  $M_t$  conditioned to leaf  $\lambda \in \Lambda(\Upsilon)$ , and  $M_{\Lambda(\Upsilon)}$  is a measure induced on  $\Lambda(\Upsilon)$  by leaf weights on  $M_t$ .

**Theorem 1.** For any  $M_t, \eta_t$ , let  $\Upsilon$ ’s leaves follow (5). Then  $L(\eta_t; M_t, \eta_t) \leq H(\Upsilon; M_t, \eta_t)$ .

Algorithm 1 can now be explained by repeatedly leveraging Theorem 1. Suppose that we have a hypothesis set of possible splits  $\mathcal{H}$  to grow our  $\alpha$ -tree. Denote  $\Upsilon(\lambda, h)$  as the  $\alpha$ -tree  $\Upsilon$  split at leaf  $\lambda \in \Lambda(\Upsilon)$  using test  $h$ . The inner loop within Step 2 is the process of finding the best possible leaf splits which helps to minimize the  $\alpha$ -tree’s entropy and to reduce the risk as per Theorem 1. The  $\alpha$ -values of (5) calculated in step 3 are those used to ensure Theorem 1 holds. By setting  $M_t \leftarrow M_{s_w}$  and  $\eta_t \leftarrow \eta^*$ , Algorithm 1 improves CVAR by iteratively improving the  $\alpha$ -tree’s worst subgroup entropy  $H(\Upsilon; M_{s_w}, \eta^*)$ , which as a surrogate improves the worst subgroup risk  $L(\eta_u; M_{s_w}, \eta^*)$ . To accommodate for different  $\beta$  quantiles values for CVAR, TOPDOWN can be run repeatedly (replacing the initial input tree  $\Upsilon_0$ ) to progressively improve all  $s \in \mathcal{S}_\beta$ . Hence, to reduce CVAR, we basically

$$\text{use TOPDOWN with } M_t \leftarrow M_s \ (s \in \mathcal{S}_\beta) \text{ and } \eta_t \leftarrow \eta^*. \quad (\text{CVAR})$$

As alluded to by the notation used to instantiate TOPDOWN, the inputs of the algorithm can be instantiated to optimize for fairness criteria beyond CVAR. This is discussed in Section 6. In the usual ML setting,  $M_t, \eta_t$  can be *estimated* from a training sample (see Section 7).

**Initialization:** In the procedure of improving CVAR, the worst subgroups can be iteratively improved. However, we also need to make sure that improvement of a subgroup does not adversely affect another subgroup (which could potentially harm CVAR instead). As such, we introduce an additional structure to the  $\alpha$ -tree  $\Upsilon$  by tweaking the initial tree structure  $\Upsilon_0$  used in TOPDOWN. Since the fairness attribute  $S$  partitions the dataset, a convenient choice of initializing the  $\alpha$ -tree is to split by the subgroup modalities, as depicted in Fig. 3. As such, we grow separate sub- $\alpha$ -trees for each of the sensitive modalities. For CVAR, this allows the subgroup risk of individual subgroups to be tweaked without adversely affecting other subgroups.

## 5 Formal Properties

We move to the formal properties of our approach. We first detail the background of improper loss functions which motivates our correction given in Section 3. We then present the formal properties of this correction. The useful properties of having  $\alpha(\cdot)$  represented by a tree structure is then discussed. Finally, we present a convergence analysis of Algorithm 1 and an alternative boosting scheme.

**Can  $\eta_f$  as per (3) correct (any) potential unfairness? Yes.** In short, this comes from recent theory in *improper loss functions* for class probability estimation (CPE) (Sypherd et al., 2022). We are interested in the pointwise minimizer (eventually set-valued) of:

$$t_\ell(\eta) \doteq \arg \inf_{u \in [0,1]} L(u, \eta). \quad (9)$$

Dubbed as the *Bayes tilted estimate* of a loss  $\ell$  (Sypherd et al., 2022),  $t_\ell(\eta)$  is the set of optimal ‘‘responses’’ given a ground truth (pointwise) posterior  $\eta$ . Common loss functions are *proper*: the ground truth value  $\eta \in t_\ell(\eta)$  is an optimal response. However, in the case where  $\eta$  cannot be trusted (for instance when it is *unfair*), we may not want to default to imitating  $\eta$ . In addition we also want to make sure that the Bayes tilted estimate can fit to any desired (in our context, *fair*) target. The so-called  $\alpha$ -loss  $\ell^\alpha$ , which generalizes the (proper) log-loss, is a good candidate parameterized by a variable  $\alpha$ . Its Bayes tilted estimate is the pointwise version of (3), for  $\alpha \notin \{0, \infty\}$  and  $\eta \neq 1/2$ :

$$t_{\ell^\alpha}(\eta) = \{\eta^\alpha / (\eta^\alpha + (1 - \eta)^\alpha)\}. \quad (10)$$

Importantly for  $\alpha$ -losses, for any  $\eta \notin \{0, 1/2, 1\}$  and any  $\eta' \in (0, 1)$ , there exists  $\alpha \in \mathbb{R}$  in (10) such that  $t_{\ell^\alpha}(\eta) = \{\eta'\}$ . This property, called *twist-properness* (Sypherd et al., 2022), allows for any pointwise correction. By extending  $\alpha$  to a function (of  $x \in \mathcal{X}$ , as per (3)), twist-properness ensures that given an initial unfair posterior an appropriately learned  $\alpha(\cdot)$  can correct any unfairness. This allows for (**flexible**) fairness post-processing – different  $\alpha$  functions can be learned for different criteria (*i.e.*, Fig. 1).

**Why use the Log-Loss?** As per Section 3, we minimize the log-loss. We choose the log-loss for two reasons: ① it is strictly proper and so minimizing  $L(\eta_f; M_t, \eta_t)$  (*i.e.*, via TOPDOWN) ‘‘pushes’’  $\eta_f$  towards target  $\eta_t$ ; and ② it is the  $\alpha$ -loss for  $\alpha = 1$ , so we are guaranteed that for the minimizer  $\eta_f \rightarrow \eta_u \iff \alpha \rightarrow 1$ . With alternative (*i.e.* non-strictly proper) loss, we might have only ‘‘ $\Rightarrow$ ’’.

**Do we have guarantees on some proximity of  $\eta_f$  with respect to  $\eta_u$ ? Yes, with light assumptions.** We examine the (**proximity**) of black-box and post-processed posteriors with the KL divergence (Amari & Nagaoka, 2000):

$$\text{KL}(\eta_u, \eta_f; M) \doteq \mathbb{E}_{(X, Y) \sim D_u} [\log (dD_u((X, Y)) / dD_f((X, Y)))], \quad (11)$$

where  $D_u, D_f$  are the product measures defined from  $M$  and their respective posteriors. To bound the proximity (11), we present setting **(S1)**.

- (S1)** Assumption 1 holds for some  $0 < B \leq 3$  and function  $\alpha$  satisfies  $|\alpha(x) - 1| \leq 1/B$  (a.s.).

This setting lead to the following *data independent* proximity bound.

**Theorem 2.** *For any  $M$ , (S1) implies  $\text{KL}(\eta_u, \eta_f; M) \leq \pi^2 / (6 \cdot (2 + \exp(B) + \exp(-B)))$ .*

As an example, fix  $B = 3$  for (S1). 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 is upperbounded as  $\text{KL}(\eta_u, \eta_f; M) \leq 7.5E - 2$ .

**Is the composition of transformations meaningful? Yes.** The analytical form in (3) 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 (3) is equivalent to the single transformation  $\eta_u \xrightarrow{\alpha \cdot \alpha'} \eta'_f$ .*

This gives an *invertibility* condition – wrapping  $\eta_f$  with  $\alpha' = 1/\alpha$  recovers the original black-box  $\eta_u$ .

**Given some capacity parameter for  $\eta_u$ , can we easily compute that of  $\eta_f$ ? Yes, e.g., for decision trees.** Such a question is particularly relevant for generalization. As we are using the log-loss (2), a relevant capacity notion to assess the uniform convergence of risk minimization for the whole wrapped model is the Rademacher (**complexity**) (Bartlett & Mendelson, 2002). We examine the following set of functions:

$$\mathcal{H}_f \doteq \{\eta_f : \eta_f(\mathbf{x}) \text{ given by (3) with } \alpha, \eta_u; \forall (\alpha, \eta_u)\}, \quad (12)$$

where we assume known the set of functions from which  $\eta_u$  was trained. We now assume we have a  $m$ -training sample  $\mathcal{S} \doteq \{(\mathbf{x}_i, y_i) \sim D\}_{i=1}^m$ . The empirical Rademacher complexity of a set of functions  $\mathcal{H}$  from  $\mathcal{X}$  to  $\mathbb{R}$ ,  $\mathfrak{R}_s(\mathcal{H}) = \mathbb{E}_\sigma \sup_{h \in \mathcal{H}} \mathbb{E}_i[\sigma_i h(\mathbf{x}_i)]$  (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 log-loss. To see how the  $\alpha$ -tree affects the Rademacher complexity of classification using  $\eta_f$  instead of  $\eta_u$ , suppose real-valued prediction based on  $\eta_u$  is achieved via logit mapping,  $\iota \circ \eta_u$  (12). Such mappings are common for decision trees (Schapire & Singer, 1998).

**Lemma 2.** *Suppose  $\{\eta_u\}$  is the set of decision trees of depth  $\leq d$  and denote  $\mathfrak{R}_s(\text{DT}(d))$  the empirical Rademacher complexity of decision trees of depth  $\leq d$  (Bartlett & Mendelson, 2002) and  $d'$  the maximum depth allowed for  $\alpha$ -trees. Then we have for  $\mathcal{H}_f$  in (12):  $\mathfrak{R}_s(\mathcal{H}_f) \leq \mathfrak{R}_s(\text{DT}(d + d'))$ .*

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  $\eta_u$  a copy of the  $\alpha$ -tree  $\Upsilon$ .

**Does Algorithm 1 have any convergence properties? Yes, it is a boosting algorithm.** Following a similar blueprint to classical decision tree induction, it comes with no surprise that TOPDOWN can achieve boosting compliant convergence. To show that TOPDOWN is a boosting algorithm, we need a *Weak Hypothesis Assumption (WHA)*, which postulates informally that each chosen split brings a small edge over random splits for a tailored distribution.

**Definition 4.** Let  $\lambda \in \Lambda(\Upsilon)$  and  $D_{t\lambda}$  be the product measure on  $\mathcal{X} \times \mathcal{Y}$  conditioned on  $\lambda$ . The **balanced** product measure  $D'_{t\lambda}$  at leaf  $\lambda$  is defined as ( $z \doteq (\mathbf{x}, y)$  for short):

$$D'_{t\lambda}(z) \doteq \frac{1 - e(M_\lambda, \eta_t) \cdot y\tilde{\iota}(\eta_u(\mathbf{x}))}{1 - e(M_\lambda, \eta_t)^2} \cdot D_{t\lambda}(z). \quad (13)$$

We check that  $\int_\lambda dD'_{t\lambda} = 1$  because of Def. 2. Our balanced distribution is named after Kearns & Mansour (1996)'s: ours indeed generalizes theirs. The key difference comes from the change in setting, where we consider the alignment purity of a leaf and not its label purity. The “fairness-free case” where  $\tilde{\iota}(\cdot)$  is replaced by constant 1 yields the original balanced distribution (Kearns & Mansour, 1996). We now state our WHA.

**Assumption 2.** Let  $h : \mathcal{X} \rightarrow \mathcal{Y}$  be the function splitting leaf  $\lambda$ . For  $\gamma > 0$ , then  $h$   $\gamma$ -**witnesses** the Weak Hypothesis Assumption (**WHA**) at  $\lambda$  iff

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

Intuitively, (i) gives a condition on the split  $h$ 's correlation with unfair posterior  $\eta_u$  agreement with labels and (ii) does the same for the confidence of  $\eta_u$  predictions. Similarly to the balanced distribution, in the fairness-free case where we only care about the label purity of splits, the WHA simplified to that of Kearns & Mansour (1996)'s – *i.e.*, only (i) matters. A further discussion on our balanced distribution and WHA is in the SI, Section III. We now state TOPDOWN's boosting compliant (**convergence**).

**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  $\Upsilon$   $\gamma$ -witnesses the WHA. Then there exists a constant  $c > 0$  such that  $\forall \varepsilon > 0$ , if the number of leaves of  $\Upsilon$  satisfies  $|\Lambda(\Upsilon)| \geq (1/\varepsilon)^{c \log(1/\varepsilon)/\gamma^2}$ , then  $\eta_f$  crafted from (3) using TOPDOWN's  $\Upsilon$  achieves  $L(\eta_f; M_t, \eta_t) \leq \varepsilon$ .

**Are there alternative ways of growing  $\alpha$ -trees?** Yes. Let us call *conservative* the scoring scheme in (5). 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  $\alpha$ -tree.

**Definition 5.** For any mixture  $M_t$  and posteriors  $\eta_u, \eta_t$ , let  $e^+(M_t, \eta_t)$  and  $e^-(M_t, \eta_t)$  be defined by

$$e^\pm(M_t, \eta_t) \doteq \mathbb{E}_{(\mathbf{X}, Y) \sim D_t} [\max\{0, \pm Y\tilde{\iota}(\eta_u(\mathbf{X}))\}]. \quad (14)$$

The **audacious** scoring schemes at the leaves of the  $\alpha$ -tree replaces (5) in Step 3 by:

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

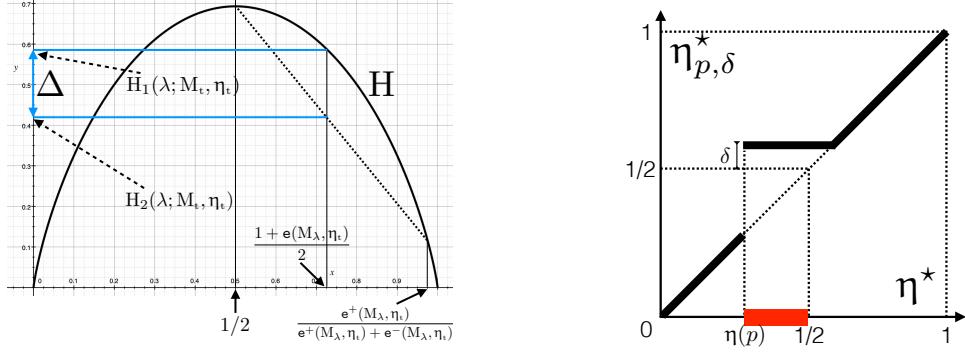


Figure 4: *Left:* Difference between the per-leaf bounds on risk  $L(\eta_u; M_t, \eta_t)$  using (8) and Theorem 1 (conservative scoring) and (15) (audacious scoring). Details in the proof of Lemma 3. *Right:* A representation of the  $(p, \delta)$ -pushup of  $\eta^*$ , where  $\eta(p) \doteq \inf \eta^*(\mathcal{X}_p) < 1/2$  (Def. 6). All posteriors in  $[\eta(p), 1/2 + \delta]$  are mapped to  $1/2 + \delta$ ; others do not change. New posterior  $\eta_{p,\delta}^*$  eventually reduces the accuracy of classification for observations whose posterior lands in the thick red interval ( $x$ -axis).

**Theorem 4.** Suppose Assumption 1 holds and let  $H_2(q) \doteq H(q)/\log 2$  ( $\in [0, 1]$ ),  $H$  being defined in Definition 3. For any leaf  $\lambda \in \Lambda(\Upsilon)$ , denote for short:

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

where let  $e_\lambda^b \doteq e^b(M_\lambda, \eta_t)$ ,  $\forall b \in \{+, -\}$ . Using audacious scoring, we get instead of Theorem 1:

$$L(\eta_f; M_t, \eta_t) \leq \mathbb{E}_{\lambda \sim M_{\Lambda(\Upsilon)}} [H_2(\lambda; M_t, \eta_t)]. \quad (15)$$

While upperbounds in Theorem 1 and Theorem 4 may look incomparable, it takes a simple argument to show that (15) is never worse and can be much tighter.

**Lemma 3.**  $\forall \alpha$ -tree  $\Upsilon$ ,  $\mathbb{E}_{\lambda \sim M_{\Lambda(\Upsilon)}} [H_2(\lambda; M_t, \eta_t)] \leq H(\Upsilon; M_t, \eta_t)$ .

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 3 can be substantial, a slack which can be represented using a simple picture, see Figure 4 (left), following from the use of Jensen's inequality in the Lemma's proof.

**As audacious scoring is better boosting-wise, is conservative scoring useful? Yes.** If we only cared about accuracy, we would barely have any reason to use the conservative correction. Even thinking about generalization, the Rademacher complexity of decision trees is a function of their depth so faster the convergence, the better (Bartlett & Mendelson, 2002, Section 4.1). Adding fairness substantially changes the picture: some constraints, like equality of opportunity (Section 6) can antagonize accuracy to some extent. In such a case, using the conservative correction can keep posteriors  $\eta_u$  and  $\eta_t$  close enough (Theorem 2) so that fairness can be achieved without substantial sacrifice on accuracy.

## 6 Fairness and Societal Considerations

In this section, we present the fairness guarantees TOPDOWN can achieve. In particular, we provide a discussion about how Theorem 3 can guarantee minimization of the CVAR criteria. Furthermore, we provide alternative inputs to TOPDOWN which allows for EOO to be targeted as a fairness criteria. In the SI Section II, we further present a treatment of statistical parity. Lastly, we discuss how using  $\alpha$ -trees provides explainable corrections and how utilization of the sensitive attribute (as per Fig. 3) can be circumvented.

**Guarantees on CVaR:** As discussed in previous sections, one way to improve the CVAR fairness criteria (as per (4)) is to focus optimization on the worst treated subgroups. Given a specified quantile group  $\beta$  and the set of worse subgroups  $\mathcal{S}_\beta$ , we can repeat (CVAR) until  $\text{CVAR}_\beta(\eta_f)$  gets below a threshold or (more specifically) its worst tread group gets a risk below a threshold (*i.e.*, a stopping criterion). Importantly, Theorem 3 provides a guarantee: to ensure  $\text{CVAR}_\beta$  is below  $\varepsilon$ , we simply need to boost for  $|S|$  times the tree size bound  $|\Lambda(\Upsilon)|$  given in Theorem 3.

**Guarantees on EOO:** EOO requires to smooth discrimination within an “advantaged” group, modeled by the label  $Y = 1$  (Hardt et al., 2016). We say that  $\eta_f$  achieves  $\varepsilon$ -equality of opportunity iff a mapping  $h_f$  of  $\eta_f$  to  $Y$  (*e.g.* using the sign of its logit  $\iota$ ) satisfies

$$\max_{s \in S} \mathbb{P}_{X \sim P_s} [h_f(X) = 1] - \min_{s \in S} \mathbb{P}_{X \sim P_s} [h_f(X) = 1] \leq \varepsilon, \quad (16)$$

where  $P_s$  is the positive observations’ measure conditioned to value  $S = s$  for the sensitive attribute. It is clear that EOO can be antagonistic to accuracy: the rate of advantage in the data  $D$  may not be equal among the subgroups. As such, unlike CVAR, we do not want to target the Bayes posterior  $\eta_t \not\leftarrow \eta^*$  for EOO. Instead, we target a skewed posterior which aims to improve the least advantaged subgroup, *i.e.*, increasing  $s^\circ \in \arg \min_{s \in S} \mathbb{P}_{X \sim P_s} [h_f(X) = 1]$ . Our strategy consists of picking a target posterior which skews part of the original  $\eta^*$  to be more advantaged, thus reducing the LHS of (16) until (16) is satisfied<sup>1</sup>. For this, we create a  $(p, \delta)$ -pushup of  $\eta^*$ , defined in SI Appendix I.

Fig. 4 (right) presents an example of a pushup map. Notice that the pushup only changes the predicted probability of example which do not have a “confident prediction” (the interval  $[\eta(p), 1/2 + \delta]$ ). Intuitively,  $p$  controls how many examples are corrected and  $\delta$  controls how much the correction “pushes up” advantage, further discussion in SI Appendix I. We then run TOPDOWN using as mixture the *positive* measure conditioned to  $S = s^\circ$  and  $p \doteq \mathbb{P}_{X \sim P_{s^*}} [h_f(X) = 1] + \varepsilon/(K - 1)$ ,  $\delta \doteq K\varepsilon/(K - 1)$ , with  $K > 1$  user-fixed. Thus, we do

Use TOPDOWN with  $M_t \leftarrow P_{s^\circ}$  and  $\eta_t \leftarrow \eta_{p, \delta}^*$ . (EOO)

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

In the full context of EOO, in the optimization we should not wait to get the bound on  $L(\eta_f; M_t, \eta_t)$ . Rather, we should make sure (a) we update  $\arg \min_{s \in S} \mathbb{P}_{X \sim P_s} [h_f(X) = 1]$

---

<sup>1</sup>If we instead reduce  $\arg \max_{s \in S} \mathbb{P}_{X \sim P_s} [h_f(X) = 1]$  we get a symmetric strategy. The application informs which to use: if positive class implies money spending (*e.g.* loan prediction), then our strategy implies spending more money; while the latter aims to reduce money lent to achieve fairness.

(and thus  $s^\circ$ ) after each split in the  $\alpha$ -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. Notably, the guarantee presented in Theorem 5 depends on the mapping  $h_f$  and not the direct posterior  $\eta_f$ , as typically considered (Hardt et al., 2016). When taking a threshold (sign of the logit),  $h_f$  can be interpreted as forcing the original posterior to be extreme values of 0 or 1.

Unlike the CVAR case, EOO (as per (17), SI) requires an explicit approximation of  $\eta^*$ . In practice, we find that taking a simple approximation of  $\eta^*$  still can yield fairness gains. However, if one does not want to make such an approximation, one can adapt the statistical parity approach (detailed in SI, Section II). Similarly, if wants to consider the typical EOO definitions depending on posterior values, the target measure can be replaced (*i.e.*, swapping measure  $M_t$  with the positive examples  $P$ ).

**Explainability:**  $\alpha$ -trees using the initialization proposed in Section 4 (and Fig. 3) allows for **(explainability)** properties similar to that of decision tree classifiers. Fixing a sensitive attribute  $S = s$ , the corresponding sub- $\alpha$ -tree  $\Upsilon_s$  can be examined to scrutinize the correction done for the corresponding subgroup. If the splits of the  $\alpha$ -tree are simple, similarly to standard decision tree classifiers, corresponding partitions of the input domain can be examined. Furthermore, the type of corrections can also be examined, as discussed in Section 3; where corrections can be classified as “sharpening”, “dampening”, or “polarity flipping” depending on the leaves’  $\alpha$ -values.

**Usage of sensitive attribute:** 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, § 6.2.3). Our basic approach to the induction of  $\alpha$ -trees falls in this category (Fig. 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 an alternative initialization to naively splitting on subgroups. We thus also *redefine* sensitive groups based on this decision tree – thus alleviating the need to use the sensitive attribute in the  $\alpha$ -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). We however note that its application in post-process and  $\alpha$ -trees may not be appropriate across all domains (Datta et al., 2017).

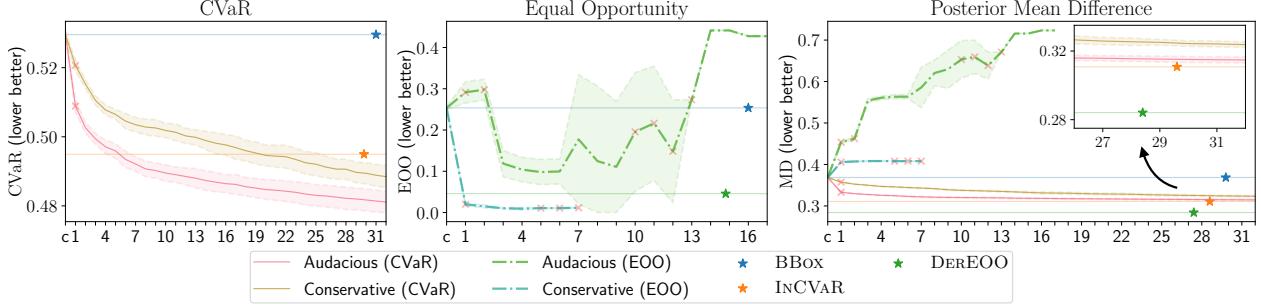
## 7 Experiments

To evaluate TOPDOWN<sup>2</sup>, we consider three datasets presenting a range of different size / feature types, Bank and German Credit (preprocessed by AIF360 (Bellamy et al., 2019)) and the American Community Survey (ACS) dataset preprocessed by Folktables<sup>3</sup> (Ding et al., 2021). The SI (pg 41) presents all results at length (including considerations on proxy sensitive attributes, distribution shift, and interpretability), along with the different black-boxes considered (random forests and neural nets). We concentrate in the Section on the ACS dataset for income prediction in the state of CA and evaluate TOPDOWN’s

---

<sup>2</sup>Implementation public at: <https://github.com/alexandersoen/alpha-tree-fair-wrappers>

<sup>3</sup>Public at: <https://github.com/zykls/folktables>



**Figure 5: ACS 2015 with Binary Sensitive Attribute and Random Forest Black-box:** Evaluation of TOPDOWN over boosting iterations (x-axis) for different fairness criteria. ‘c’ on the x-axis denotes the clipped black-box. ‘×’ denote when a subgroup’s  $\alpha$ -tree is initiated (over any fold). The shade denotes  $\pm$  a standard deviation from the mean, this disappears when folds have early stopping.

application to various fairness criteria (as per Section 6 and SI pg 19) with Random Forests (RF). For these experiments, we consider *age* as a binary sensitive attribute with a bin split at 25 (a trinary modality is deferred to the SI). For the black-box, we consider a clipped (Assumption 1 with  $B = 1$ ) random forest (RF) from `scikit-learn` calibrated using Platt’s method (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. For EOO, we utilize an out-of-the-box Gaussian Naive Bayes classifier from `scikit-learn` to approximate  $\eta^*$ .

**Multiple fairness criteria** We evaluate TOPDOWN for CVAR, equality of opportunity EOO, and statistical parity SP. The complete treatment of SP is pushed to SI (Sections II, XII). SP aims to make subgroup’s expected posteriors similar and is popular in a various post-processing methods (Wei et al., 2020; Alabdulmohsin & Lucic, 2021). The definition can be found in SI (pg 19) along with the strategy used in TOPDOWN. Conservative and audacious updates rules are also tested. For each of these TOPDOWN configurations, we boost for 32 iterations. The initial  $\alpha$ -tree is initialized as in Fig. 3.

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 (DEREEO) (Hardt et al., 2016). Our SP baselines include 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). We denote the clipped black-box as BBox. The experiments for CVAR and EOO are summarized in Fig. 5; the full plot with SP is presented at SI Fig. 10. For clarity we only plot the baselines and wrappers which are directly associated to each fairness criteria. We also plot the posterior mean difference between the data and debiased posteriors MD (0/1 loss) to examine the effects on accuracy.

**For CVaR**, both conservative and audacious approaches decreases CVAR, which results in better CVAR values than both the original BBox and in-processing baseline INCVAR – which is good news since INCVAR directly optimizes CVAR. We note that there are cases in which the in-processing approach is better than ours (trinary sensitive attributes in SI),

but this is expected given INCVAR’s optimization goal. Interestingly, the audacious update is superior in both CVAR and MD than the conservative update. This is also consistent for trinary sensitive attributes. Thus, the audacious update is desirable when optimizing CVAR. Another observation is that only one sensitive attribute subgroup’s  $\alpha$ -tree is initialized (only one ‘ $\times$ ’). This indicates that after 32 iterations the worse case subgroup does not change in the binary 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  $\alpha$ -tree (taking an earlier iterations) the audacious update would fail at producing outcomes as fair as its conservative counterpart. Furthermore, the audacious update comes with a significant degradation of accuracy MD. Furthermore, by looking at the iterations in which subgroup  $\alpha$ -trees are initialized, the audacious update causes large (primarily bad) jumps in performance. This rejoins our remark on the interest of having a conservative update in Section 5. When compared to DEREEOO, we find that the conservative TOPDOWN approach produces lower EOO. However, DEREEOO tend to have better accuracy scores in MD. These observations are consistent with the trinary sensitive attribute (SI).

**For SP**, we can observe fairness results that can be on par with contenders for the conservative update, but observe a substantial degradation of MD. This, we believe, follows from a simple plug-in instantiation of  $M, \eta_t$  for the fairness notion in SI Section II, resulting in potentially harsh updates. In SI (pg 19), we discuss an alternative approach using ties with optimal transport.

## 8 Limitations and Conclusion

Given the context of fairness, it is important to highlight possible limitations of our approach and the potential social harm from such misuse. We highlight two of these for our TOPDOWN approach. Firstly, our approach has shown to have failure cases in *small data cases*. This can be seen in the experiments on German Credit dataset (SI, Section XIII-XV). This, we believe, has a formal basis in our approach. For instances, EOO requires accurate data posterior estimation which may be difficult in small data regimes. Secondly, TOPDOWN is of course not unilaterally better than all other post-processing fairness approaches – there is *No Free Lunch*. As such, we do not claim that our instantiation of TOPDOWN is optimal in CVAR, EOO, or SP. However, considering that such different fairness models instantiated in the *same* algorithm can lead to competitive results with the respective state of the art, the avenue for improved instantiations or accurate extensions to new fairness constraints appears promising. We leave these for future work.

## Acknowledgments

YM received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (grant agreement No. 882396), the Israel Science Foundation(grant number 993/17), Tel Aviv University Center for AI and Data Science (TAD), and the Yandex Initiative for Machine Learning at Tel Aviv University.

AS and LX thank members of the ANU Humanising Machine Intelligence program for discussions on fairness and ethical concerns in AI, and the NeCTAR Research Cloud for providing computational resources, an Australian research platform supported by the National Collaborative Research Infrastructure Strategy.

## 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.
- 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 *24<sup>th</sup> 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. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012a.
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R.-S. Fairness through awareness. In *ITCS’12*, pp. 214–226, 2012b.

- 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 28<sup>th</sup> 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.
- Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.
- 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.
- 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 *9<sup>th</sup> 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.
- Sypherd, T., Nock, R., and Sankar, L. Being properly improper. In *ICML’22*, 2022.
- 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

→ Additional Equality of Opportunity Details	Pg 19
→ Handling Statistical parity	Pg 19
→ Weak Hypothesis Assumption Discussion	Pg 19
→ General KL Distortion IV	Pg 21
→ Proof of Theorem F	Pg 22
→ Proof of Theorem 2 and G	Pg 26
→ Proof of Theorem 1 and 3	Pg 28
→ Proof of Theorem 4	Pg 34
→ Proof of Lemma 3	Pg 37
→ Proof of Theorem 5	Pg 37

### Supplementary material on experiments

→ SI Experiment Settings	Pg 41
→ Experiments on Statistical Parity	Pg 43
→ Additional Main Text Experiments	Pg 44
→ Neural Network Experiments	Pg 48
→ Proxy Sensitive Attributes	Pg 52
→ Distribution Shift	Pg 56
→ High Clip Value	Pg 61
→ Example Alpha-Tree	Pg 69

# I Additional Equality of Opportunity Details

In this section, we present additional details for the equality of opportunity (EOO) strategy presented in Section 6. In particular, we present the full definition of the  $(p, \delta)$ -pushup of  $\eta^*$ .

**Definition 6.** 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  $(p, \delta)$ -pushup of  $\eta^*$ ,  $\eta_{p,\delta}^*$ , is the posterior defined as  $\eta_{p,\delta}^* = \eta^*$  if  $\inf \eta^*(\mathcal{X}_p) \geq 1/2$  and otherwise:

$$\eta_{p,\delta}^*(\mathbf{x}) \doteq \begin{cases} \frac{1}{2} + \delta & \text{if } \eta^*(\mathbf{x}) \in [\inf \eta^*(\mathcal{X}_p), 1/2 + \delta] \\ \eta^*(\mathbf{x}) & \text{otherwise.} \end{cases} \quad (17)$$

Following Fig. 4 (right) and discussion in the main-text, we make a couple more observation. From Def. 6, the selection of the set  $\mathcal{X}_p$  (*i.e.* choice of  $p$ ) presents a tradeoff between accuracy and the fairness objective. As  $p$  increases, the size of  $\mathcal{X}_p$  necessarily increases. For instance, taking  $\mathcal{X}_p = \{\mathbf{x} : \eta^*(\mathbf{x}) > l\}$ , the larger  $\mathcal{X}_p$  is the more negative prediction are flipped. Thus intuitively,  $p$  measures the size of correct posterior values and  $\delta$  defines the size of the correction.

# II Handling Statistical parity

**Statistical parity** (SP) is a group fairness notion (Dwork et al., 2012b), 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  $\eta_f$  achieves  $\varepsilon$ -statistical parity (across all groups induced by sensitive attribute  $S$ ) iff

$$\max_{s \in S} \mathbb{E}_{\mathbf{X} \sim M_s} [\eta_f(\mathbf{X})] - \min_{s \in S} \mathbb{E}_{\mathbf{X} \sim M_s} [\eta_f(\mathbf{X})] \leq \varepsilon. \quad (18)$$

Denote  $s^\circ \doteq \arg \min_{s \in S} \mathbb{E}_{\mathbf{X} \sim M_s} [\eta_f(\mathbf{X})]$ ,  $s^* \doteq \arg \max_{s \in S} \mathbb{E}_{\mathbf{X} \sim M_s} [\eta_f(\mathbf{X})]$ . Since the risk we minimize in (2) involves a proper loss, the most straightforward use of TOPDOWN is to train the sub- $\alpha$ -tree for one of these two groups, giving as target posterior the *expected* posterior of the other group, *i.e.* we use  $\eta_t(\mathbf{x}) = \mathbb{E}_{\mathbf{X} \sim M_{s^*}} [\eta_u(\mathbf{X})] \doteq \bar{\eta}_{u,s^*}$  if we grow the  $\alpha$ -tree of  $s^\circ$  and thus iterate

$\text{TOPDOWN with } M_t \leftarrow M_{s^\circ} \text{ and } \eta_t \leftarrow \bar{\eta}_{u,s^*},$

and we repeat until  $s^\circ$  does not achieve anymore the smallest expected posterior. We then update the group and repeat the procedure until a given slack  $\varepsilon$  is achieved between the extremes in (18). More sophisticated / gentle approaches are possible, including using the links between statistical parity and optimal transport (OT, Dwork et al. (2012b, Section 3.2)), suggesting to use as target posterior the expected posterior obtained from an OT plan between groups  $s^\circ$  and  $s^*$ .

# III Weak Hypothesis Assumption Discussion

In this Section, we discuss the balanced distribution introduced in Definition 4 and the Weak Hypothesis Assumption (WHA) introduced in Assumption 2. In particular, we note that

these definitions are a generalization of those introduced in “classical” decision tree boosting (Kearns & Mansour, 1996). The primary difference in our case and the “classical” case, is that we must consider the posterior values of a base black-box  $\eta_u$ , whereas the usual top-down tree induction only cares about minimizing its loss function to produce an accurate  $\eta_t$ . In-fact, we can even interpret our variants as a case where we replace the labels in a classification task  $Y$  with “labels” (in  $[0, 1]$ ) of how well another classifier does in classification  $Y\tilde{\iota}(\eta_u(X))$ . We will explore various cases to show that our definitions reduce to the classical case; and further provide settings to intuitive describe our more complicated assumption.

**The balanced distribution** in Definition 4 acts as a re-weighting mechanism for particular samples of the original distribution  $D_{t\lambda}(z)$ . Let us first consider the reduction to the balanced distribution introduced by Kearns & Mansour (1996). This comes from the fairness-free case, where we ignore the confidence of the black-box  $\eta_u$  and set  $\tilde{\iota}(\eta_u) \leftarrow 1$ . This provides two simplifications. Firstly, the edge value becomes  $e(M_\lambda, \eta_t) = \mathbb{E}_{(X, Y) \sim D_{t\lambda}}[Y] = 2q_\lambda - 1$ , where  $q_\lambda$  the local proportion of positive examples in  $\lambda$ . That is, we have  $q_\lambda = \mathbb{P}(Y = +1 | X \text{ at leaf } \lambda)$ . The second simplification is in the numerator of Equation 13. In summary, we get:

$$D'_{t\lambda}(z) \leftarrow \frac{1 - e(M_\lambda, \eta_t) \cdot y}{1 - e(M_\lambda, \eta_t)^2} \cdot D_{t\lambda}(z) = \frac{1 - y \cdot (2q_\lambda - 1)}{4q_\lambda(1 - q_\lambda)} \cdot D_{t\lambda}(z) = D_{t\lambda}(z) \cdot \begin{cases} \frac{1}{2q_\lambda} & \text{if } y = +1 \\ \frac{1}{2(1-q_\lambda)} & \text{if } y = -1 \end{cases}, \quad (19)$$

which is indeed the balanced distribution of Kearns & Mansour (1996). Intuitively, this balanced distribution simply ensures that regardless of  $x$ , it is equally likely for  $y$  be either  $-1$  or  $+1$  (hence balanced in its prediction value  $Y$ ). This can be seen by calculating the edge with the new distribution  $\mathbb{E}_{(X, Y) \sim D'_{t\lambda}}[Y] = 0$ .

To consider our balanced distribution, let us start with (19). In particular, we will replace the labels  $Y \in \{-1, +1\}$  by the predictive ability of black-box  $\eta_u$ . This predictive ability is summarized by  $Y\tilde{\iota}(\eta_u(X)) \in [-1, 1]$ , where larger is better. In particular, taking from the boosting jargon, the *confidences* ( $|\tilde{\iota}|$ , Schapire & Singer (1999)) of the black-box  $\eta_u$  are considered. For instance, a highly confident  $|\tilde{\iota}(\eta_u(X))| \rightarrow 1$  and correct  $\text{sign}(Y) = \text{sign}(\tilde{\iota}(\eta_u(X)))$  prediction will lead to a value close to  $+1$  (positive label). If the prediction is confident but incorrect  $\text{sign}(Y) \neq \text{sign}(\tilde{\iota}(\eta_u(X)))$ , then the value will be close to  $-1$  (negative label). If the prediction is not confident  $|\tilde{\iota}(\eta_u(X))| \rightarrow 0$ , then we get a neutral response 0.

With this in mind, the balanced distribution in Definition 4 is “balanced” in the prediction score  $Y\tilde{\iota}(\eta_u(X))$ . Of course, the distribution  $D'_{t\lambda}(z)$  is with respect to  $X \times Y$  and not  $X \times [-1, 1]$ , and thus we are not re-balancing for each possible  $Y\tilde{\iota}(\eta_u(X)) \in [-1, 1]$ . The balance of the distribution takes into account the confidence of the predictions, this can be seen in examination of the balanced edge in this setting:

$$\mathbb{E}_{(X, Y) \sim D'_{t\lambda}}[Y \cdot \tilde{\iota}(\eta_u(X))] = \frac{1}{1 - e(M_\lambda, \eta_t)^2} \cdot e(M_\lambda, \eta_t) \cdot \mathbb{E}_{(X, Y) \sim D_{t\lambda}}[1 - \tilde{\iota}^2(\eta_u(X))]. \quad (20)$$

(20) is in general non-zero. However, in the case that the  $\eta_u$  is confident such that  $\tilde{\iota}(\eta_u(X)) \in \{-1, +1\}$ , then the new edge goes to 0. In other words, when we have a maximally confident black-box, the balanced edge will be 0; otherwise, the edge is scaled with respect to a function of the confidence  $\mathbb{E}_{(X, Y) \sim D_{t\lambda}}[1 - \tilde{\iota}^2(\eta_u(X))]$ .

**Our Weak Learning Assumption (WHA)** in Assumption 2 has a similar analysis as that of the balanced distribution. Condition (i) of the assumption is exactly the same  $\gamma$ -witness

condition in Kearns & Mansour (1996). Condition (ii), similar to the analysis of (20), vanishes when we have a maximally confident black-box  $\eta_u$ . In general, condition (ii) ensures that one side of the split induced by  $h$  are skewed to one side.

Let us exemplify how our WHA works if we have a leaf  $\lambda$  where local “treatments due to the black-box” are bad ( $y\tilde{\iota}(\eta_u(\mathbf{x})) < 0$  often). In such a case,  $e(M_\lambda, \eta_t) < 0$  so the balanced distribution (Definition 4) 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  $\mathbb{E}_{(\mathbf{X}, Y) \sim D_{t,\lambda}} [h(\mathbf{X})] \geq \mathbb{E}_{(\mathbf{X}, Y) \sim D_{t,\lambda}} [\tilde{\iota}^2(\eta_u(\mathbf{X})) \cdot h(\mathbf{X})]$ :  $h = -1$  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 (5) 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  $\lambda$  would have disproportionately ended in the right ( $h = +1$ ) leaf, where it would be likely that  $\Upsilon(\cdot)$  would this time be *positive* and thus preserve the polarity of the treatment of the black-box.

**Why are the assumptions states using the balanced distribution?** A point worth clarifying is the “why’s” of consider a WHA in the way it is formulated and the focus on the balanced distribution. In typical boosting algorithms, a weak learning hypothesis is needed: the assumption that a classifier / split / weak learner exists for *any* distribution which is better than random. In top-down tree induction, it is in-fact sufficient to only have this assumption hold for one type of distribution — the balanced distribution (Kearns & Mansour, 1996). As such the WHA presented, is actually a weaker assumption than the “distributionally global” weak learning assumption used in other boosting algorithms.

## IV General KL Distortion

We introduce a general Theorem for upperbounding the distortion an  $\alpha$ -correction can make, as promised in Section 5. Notably, the following result does not rely on any setting.

**Theorem F.** *For any function  $\alpha : \mathcal{X} \rightarrow \mathbb{R}$ , any black-box posterior  $\eta_u$ , and any integer  $K \geq 2$ , using (3) yields the following bound on the KL divergence:*

$$\text{KL}(\eta_u, \eta_f; M) \leq \mathbb{E}_{\mathbf{X} \sim M} \left[ \sum_{k=2}^K \frac{\eta_u(\mathbf{X})(1 - \eta_u(\mathbf{X}))f^k(\alpha(\mathbf{X}), \eta_u(\mathbf{X}))}{k(k-1)} \right] + o\left(\mathbb{E}_{\mathbf{X} \sim M} [(\alpha(\mathbf{X}) - 1)^K]\right), \quad (21)$$

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

$$f(z, u) \doteq |\log(u/(1-u)) \cdot (1-z)| = |\iota(u) \cdot (1-z)|. \quad (22)$$

The proof is presented in Section V. In general, one can see that if  $\alpha(\cdot)$  does not differ from 1 too much, the KL divergence will be small. This intuitively makes sense as  $\alpha = 1$  causes  $\eta_f = \eta_u$  — there is no untwisting. Theorem F can be further weakened to provide easier to understand bounds, as per Corollary 2.

We further present an alternative setting to that in the main-text, which provides another bound on distortion. In this setting **(S2)**, we ensure that when the predictions of  $\eta_u$  are confident, then the corresponding  $\alpha$  correction are small (close to  $\alpha = 1$ ).

**(S2)**  $f(\alpha(\mathbf{x}), \eta_u(\mathbf{x})) = |\iota(\eta_u(\mathbf{x})) \cdot (1 - \alpha(\mathbf{x}))| \leq 1$  (a.s.) ,  $f$  being in (22).

This provides the alternative upperbound on the distortion.

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

$$\text{KL}(\eta_u, \eta_f; M) \leq \pi^2/24 \approx 0.41. \quad (23)$$

The proof of the Corollary is in SI, Section VI and includes a graphical view of the values of  $\eta_u(\cdot)$  and  $\alpha(\cdot)$  complying with **(S2)**.

## V Proof of Theorem F

We first show two technical Lemmata.

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

$$h(z) \doteq \log \left( \frac{1}{1 + a^{1+z}} \right). \quad (24)$$

We have

$$h^{(k)}(z) = -\frac{\log^k(a) \cdot a^{1+z}}{(1 + a^{1+z})^k} \cdot P_{k-1}(a^{1+z}), \quad (25)$$

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

$$c_{k+1,k} = (-1)^k, \quad (26)$$

$$c_{k+1,j} = (j+1) \cdot c_{k,j} - (k+1-j) \cdot c_{k,j-1}, \forall 0 < j < k, \quad (27)$$

$$c_{k+1,0} = 1. \quad (28)$$

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

**Proof:** We let

$$f(z) \doteq \frac{a^{1+z}}{1 + a^{1+z}}, \quad (29)$$

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

$$f^{(k)}(z) = \frac{\log^k(a) \cdot a^{1+z}}{(1 + a^{1+z})^{k+1}} \cdot P_k(a^{1+z}). \quad (30)$$

We first check

$$f'(z) = \frac{\log(a) \cdot a^{1+z}}{(1 + a^{1+z})^2}, \quad (31)$$

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

$$\frac{d}{dz} \frac{a^{1+z}}{(1 + a^{1+z})^k} = \frac{\log(a) \cdot a^{1+z}}{(1 + a^{1+z})^{k+1}} \cdot (-(k-1)a^{1+z} + 1), \quad (32)$$

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

$$\begin{aligned} f^{(k+1)}(z) &= \log^k(a) \cdot \frac{d}{dz} \left( \frac{a^{1+z}}{(1 + a^{1+z})^{k+1}} \cdot P_k(a^{1+z}) \right) \\ &= \log^k(a) \cdot \left( \frac{\log(a) \cdot a^{1+z}}{(1 + a^{1+z})^{k+2}} \cdot (-ka^{1+z} + 1) \cdot P_k(a^{1+z}) + \frac{a^{1+z} \cdot \log(a)}{(1 + a^{1+z})^{k+1}} \cdot a^{1+z} \cdot \frac{dP_k(x)}{dx} \Big|_{x=a^{1+z}} \right) \\ &= \frac{\log^{k+1}(a) \cdot a^{1+z}}{(1 + a^{1+z})^{k+2}} \cdot \underbrace{\left( (-ka^{1+z} + 1) \cdot P_k(a^{1+z}) + a^{1+z}(1 + a^{1+z}) \cdot \frac{dP_k(x)}{dx} \Big|_{x=a^{1+z}} \right)}_{\doteq P_{k+1}(a^{1+z})}, \end{aligned} \quad (33)$$

from which we check that  $P_{k+1}$  is indeed a polynomial and its coefficients are obtained via identification from  $P_k$ , which establishes (30) and yields to the statement of the Lemma.  $\square$

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

$$|c_{k,j}| \leq (k-1)! \binom{k-1}{j}. \quad (34)$$

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

$$a_{k+1,k} = 1, \quad (35)$$

$$a_{k+1,j} = (j+1) \cdot a_{k,j} + (k+1-j) \cdot a_{k,j-1}, \forall 0 < j < k, \quad (36)$$

$$a_{k+1,0} = 1. \quad (37)$$

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  $j = k$ ,  $b_{k+1,k} = k! \geq a_{k+1,k}$  as well. We now check, assuming the property holds at all ranks  $k$ , that for ranks  $k+1$ , we have

$$\begin{aligned} a_{k+1,j} &= (j+1) \cdot a_{k,j} + (k+1-j) \cdot a_{k,j-1} \\ &\leq (j+1)(k-1)! \binom{k-1}{j} + (k+1-j)(k-1)! \binom{k-1}{j-1}, \end{aligned} \quad (38)$$

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

$$(j+1)(k-j) + (k+1-j)j \leq k^2. \quad (39)$$

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  $K \geq 2$ ,

$$\begin{aligned} & \log\left(\frac{1}{1+a^g}\right) \\ &= \log\left(\frac{1}{1+a}\right) - \frac{a \log a}{1+a} \cdot (g-1) - \underbrace{\sum_{k=2}^K \frac{a \log^k(a) P_{k-1}(a)}{k!(1+a)^k} \cdot (g-1)^k}_{\doteq R_{K,a}(g)} + o((g-1)^K). \end{aligned} \quad (40)$$

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 (40) unveils an interesting simplification:

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

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

$$\begin{aligned} \text{KL}(\eta_u, \eta_f; M) &= \mathbb{E}_{X \sim M} \left[ \eta_u(X) \cdot R_{\frac{1-\eta_u(X)}{\eta_u(X)}, K}(\alpha(X)) + (1 - \eta_u(X)) \cdot R_{\frac{\eta_u(X)}{1-\eta_u(X)}, K}(\alpha(X)) \right] \\ &\quad + o(\mathbb{E}_{X \sim M} [(\alpha(X) - 1)^K]). \end{aligned} \quad (41)$$

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 \\
& + (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 \\
& + (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 \\
& + \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 \tag{42}
\end{aligned}$$

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 (42) yields the bound on  $\text{KL}(\eta_u, \eta_f; M)$  from (41):

$$\begin{aligned}
\text{KL}(\eta_u, \eta_f; M) &\leq \mathbb{E}_{X \sim M} \left[ \sum_{k=2}^K \frac{(\eta_u^2(X)(1 - \eta_u(X)) + \eta_u(X)(1 - \eta_u(X))^2) \left| \log \left( \frac{1 - \eta_u(X)}{\eta_u(X)} \right) \right|^k}{k(k-1)} \cdot |\alpha(X) - 1|^k \right] \\
&\quad + o(\mathbb{E}_{X \sim M} [(\alpha(X) - 1)^K]) \\
&= \mathbb{E}_{X \sim M} \left[ \sum_{k=2}^K \frac{\eta_u(X)(1 - \eta_u(X)) \left| \log \left( \frac{1 - \eta_u(X)}{\eta_u(X)} \right) \right|^k}{k(k-1)} \cdot |\alpha(X) - 1|^k \right] \\
&\quad + o(\mathbb{E}_{X \sim M} [(\alpha(X) - 1)^K]),
\end{aligned} \tag{43}$$

which yields the statement of Theorem F.

## VI Proof of Theorem 2 and G

We start by (S1). We study function

$$f_k(u) \doteq u(1-u) \left| \log \left( \frac{1-u}{u} \right) \right|^k, \forall u \in \left[ \frac{1}{1+\exp(B)}, \frac{1}{1+\exp(-B)} \right]. \tag{44}$$

$f_k$  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

$$f'_k(u) = \log^{k-1} \left( \frac{1-u}{u} \right) \cdot \left( (1-2u) \cdot \log \left( \frac{1-u}{u} \right) - k \right). \tag{45}$$

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

$$(1-2u_k) \cdot \log \left( \frac{1-u_k}{u_k} \right) = k, \tag{46}$$

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

$$u_{k^*} \leq \frac{1}{1+\exp(B)}, \tag{47}$$

then we will have the upperbound:

$$\begin{aligned}
f_k(u) &\leq \frac{1}{1+\exp(B)} \cdot \frac{1}{1+\exp(-B)} \cdot B^k \\
&= \frac{B^k}{2 + \exp(B) + \exp(-B)}, \forall k \geq k^*.
\end{aligned} \tag{48}$$

We can also compute  $k^*$  exactly as it boils down to taking the integer part of the solution of (46) where  $u_k$  is picked as in (47):

$$k^* = \left\lfloor \frac{\exp(B) - 1}{\exp(B) + 1} \cdot B \right\rfloor, \quad (49)$$

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

$$\text{KL}(\eta_u, \eta_f; M) \leq \sum_{k=2}^K \frac{\mathbb{E}_{X \sim M} [(B \cdot |\alpha(X) - 1|)^k]}{(2 + \exp(B) + \exp(-B))k(k-1)} + G, \quad (50)$$

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:

$$\text{KL}(\eta_u, \eta_f; M) \leq \lim_{K \rightarrow +\infty} \frac{1}{2 + \exp(B) + \exp(-B)} \cdot \sum_{k=2}^K \frac{1}{k(k-1)} \quad (51)$$

$$\leq \frac{1}{2 + \exp(B) + \exp(-B)} \cdot \sum_{k \geq 1} \frac{1}{k^2} \quad (52)$$

$$= \frac{\pi^2}{6(2 + \exp(B) + \exp(-B))}, \quad (53)$$

which is (21) and proves the Corollary for setting **(S1)**. The proof for setting **(S2)** is direct as in this case we get:

$$\begin{aligned} \text{KL}(\eta_u, \eta_f; M) &\leq \lim_{K \rightarrow +\infty} \mathbb{E}_{X \sim M} \left[ \sum_{k=2}^K \frac{\eta_u(X)(1 - \eta_u(X))f^k(\alpha(X), \eta_u(X))}{k(k-1)} \right] \\ &= \mathbb{E}_{X \sim M} \left[ \sum_{k=2}^K \frac{\eta_u(X)(1 - \eta_u(X))f^k(\alpha(X), \eta_u(X))}{k(k-1)} \right] \\ &\leq \mathbb{E}_{X \sim M} \left[ \sum_{k=2}^K \frac{\eta_u(X)(1 - \eta_u(X))}{k(k-1)} \right] \end{aligned} \quad (54)$$

$$\leq \frac{1}{4} \cdot \sum_{k=2}^K \frac{1}{k(k-1)} \quad (55)$$

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

$$= \frac{\pi^2}{24}, \quad (57)$$

as claimed.

Figure 6 provides an idea of the set of *admissible* couples (correction, black-box posterior) that comply with **(S2)**, from which we see that the range of admissible corrections is quite flexible, even when  $\eta_u$  comes quite close to  $\{0, 1\}$ .

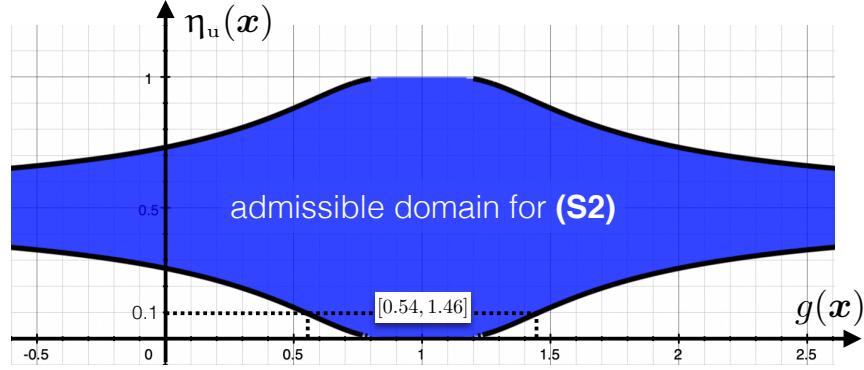


Figure 6: Admissible couples of values  $(g, \eta_u)$  (in blue) complying with setting **(S2)**. For example, any couple  $(g, 0.1)$  with  $g \in [0.54, 1.46]$  is admissible.

## VII Proof of Theorem 1 and 3

We proceed in two steps, first showing that the loss we care about for fairness (2) (main file) is upperbounded by the entropy of the  $\alpha$ -tree  $\Upsilon$ , then developing the boosting result from the minimization of the entropy itself. We thus start with the following Theorem.

**Theorem H** (Theorem 1 in Main Paper). *Suppose Assumption 1 holds and the outputs of  $\Upsilon$  are:*

$$\Upsilon(\mathbf{x}) \doteq \tilde{\iota} \left( \frac{1 + e(M_{\lambda(\mathbf{x})}, \eta_t)}{2} \right), \forall \mathbf{x} \in \mathcal{X}, \quad (58)$$

where  $\lambda(\mathbf{x})$  is the leaf reached by  $\mathbf{x}$  in  $\Upsilon$ . Then the following bound holds for the risk (2):

$$L(\eta_t; M_t, \eta_t) \leq H(\Upsilon; M_t, \eta_t). \quad (59)$$

**Proof:** We need a simple Lemma, see *e.g.* Sypherd et al. (2022).

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

$$\log(1 + \exp(\kappa z)) \leq \log(1 + \exp(\kappa B)) - \kappa \cdot \frac{B - z}{2}. \quad (60)$$

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

1,

$$\begin{aligned}
-\log \eta_f &= -\log \left( \frac{\eta_u^\gamma}{\eta_u^\gamma + (1 - \eta_u)^\gamma} \right) \\
&= -\log \left( \frac{1}{1 + \left( \frac{1 - \eta_u}{\eta_u} \right)^\gamma} \right) \\
&= \log \left( 1 + \left( \frac{1 - \eta_u}{\eta_u} \right)^\gamma \right) \\
&= \log \left( 1 + \exp \left( \gamma \log \left( \frac{1 - \eta_u}{\eta_u} \right) \right) \right) \\
&\leq \log(1 + \exp(\gamma B)) - \gamma \cdot \frac{B - \log \left( \frac{1 - \eta_u}{\eta_u} \right)}{2} \tag{61} \\
&= \log(1 + \exp(\gamma B)) - \gamma \cdot \frac{B + \iota(\eta_u)}{2}, \tag{62}
\end{aligned}$$

where in (61) we have used (60) with  $\kappa \doteq \gamma$ , using Assumption 1 guaranteeing  $|\iota(\eta_u)| \leq B$ . We also get, using this time  $\kappa \doteq -\gamma$ ,

$$\begin{aligned}
-\log(1 - \eta_f) &= \log \left( 1 + \exp \left( -\gamma \log \left( \frac{1 - \eta_u}{\eta_u} \right) \right) \right) \\
&\leq \log(1 + \exp(-\gamma B)) + \gamma \cdot \frac{B + \iota(\eta_u)}{2} \\
&= \log(1 + \exp(\gamma B)) - \gamma B + \gamma \cdot \frac{B + \iota(\eta_u)}{2} \\
&= \log(1 + \exp(\gamma B)) - \gamma \cdot \frac{B - \iota(\eta_u)}{2}. \tag{63}
\end{aligned}$$

Assembling (62) and (63) for an upperbound to  $L(\eta_f; M_t, \eta_v)$ , we get, using the fact that an

$\alpha$ -tree partitions  $\mathcal{X}$  into regions with constant predictions,

$$\begin{aligned}
L(\eta_f; M, \eta_t) &\doteq \mathbb{E}_{X \sim M} [\eta_t(X) \cdot -\log \eta_f(X) + (1 - \eta_t(X)) \cdot -\log(1 - \eta_f(X))] \\
&\leq \mathbb{E}_{X \sim M_t} \left[ \eta_t(X) \cdot \left( \log(1 + \exp(\Upsilon(X)B)) - \Upsilon(X) \cdot \frac{B + \iota(\eta_u(X))}{2} \right) \right. \\
&\quad \left. + (1 - \eta_t(X)) \cdot \left( \log(1 + \exp(\Upsilon(X)B)) - \Upsilon(X) \cdot \frac{B - \iota(\eta_u(X))}{2} \right) \right] \\
&= \mathbb{E}_{X \sim M_t} \left[ \log(1 + \exp(\Upsilon(X)B)) - \Upsilon(X) \cdot \left( \frac{\eta_t(X) \cdot \frac{B + \iota(\eta_u(X))}{2}}{(1 - \eta_t(X)) \cdot \frac{B - \iota(\eta_u(X))}{2}} \right) \right] \\
&= \mathbb{E}_{\lambda \sim M_{\Lambda(\Upsilon)}} \left[ \log(1 + \exp(\Upsilon(\lambda)B)) - \Upsilon(\lambda) \cdot \mathbb{E}_{X \sim M_\lambda} \left[ \left( \frac{\eta_t(X) \cdot \frac{B + \iota(\eta_u(X))}{2}}{(1 - \eta_t(X)) \cdot \frac{B - \iota(\eta_u(X))}{2}} \right) \right] \right] \\
&= \mathbb{E}_{\lambda \sim M_{\Lambda(\Upsilon)}} \left[ \log(1 + \exp(\Upsilon(\lambda)B)) - \Upsilon(\lambda) \cdot \mathbb{E}_{(X, Y) \sim D_{t\lambda}} \left[ \frac{B + Y \cdot \iota(\eta_u(X))}{2} \right] \right] \\
&= \mathbb{E}_{\lambda \sim M_{\Lambda(\Upsilon)}} \left[ \log(1 + \exp(\Upsilon(\lambda)B)) - \Upsilon(\lambda) \cdot \frac{B + \mathbb{E}_{(X, Y) \sim D_{t\lambda}} [Y \cdot \iota(\eta_u(X))]}{2} \right] \\
&= \mathbb{E}_{\lambda \sim M_{\Lambda(\Upsilon)}} \left[ \log(1 + \exp(\Upsilon(\lambda)B)) - \Upsilon(\lambda)B \cdot \frac{1 + e(M_\lambda, \eta_t)}{2} \right], \tag{64}
\end{aligned}$$

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  $\lambda$ 's leaf value in  $\Upsilon$ . Looking at (64), 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:

$$L(\Upsilon(\lambda)) \doteq \log(1 + \exp(\Upsilon(\lambda)B)) - \Upsilon(\lambda)B \cdot \frac{1 + e(M_\lambda, \eta_t)}{2}. \tag{65}$$

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  $e(M_\lambda, \eta_t)$ ):

$$\begin{aligned}
L(\eta_f; M_t, \eta_t) &\leq \mathbb{E}_{\lambda \sim M_{\Lambda(\Upsilon)}} \left[ \log \left( 1 + \frac{1 + e(\lambda)}{1 - e(\lambda)} \right) - \log \left( \frac{1 + e(\lambda)}{1 - e(\lambda)} \right) \cdot \frac{1 + e(\lambda)}{2} \right] \\
&= \mathbb{E}_{\lambda \sim M_{\Lambda(\Upsilon)}} \left[ -\log \left( \frac{1 - e(\lambda)}{2} \right) - \log \left( \frac{1 + e(\lambda)}{1 - e(\lambda)} \right) \cdot \frac{1 + e(\lambda)}{2} \right] \\
&= \mathbb{E}_{\lambda \sim M_{\Lambda(\Upsilon)}} \left[ -\log \left( \frac{1 - e(\lambda)}{2} \right) + \frac{1 + e(\lambda)}{2} \cdot \log \left( \frac{1 - e(\lambda)}{2} \right) - \frac{1 + e(\lambda)}{2} \cdot \log \left( \frac{1 + e(\lambda)}{2} \right) \right] \\
&= \mathbb{E}_{\lambda \sim M_{\Lambda(\Upsilon)}} \left[ -\frac{1 - e(\lambda)}{2} \cdot \log \left( \frac{1 - e(\lambda)}{2} \right) - \frac{1 + e(\lambda)}{2} \cdot \log \left( \frac{1 + e(\lambda)}{2} \right) \right] \\
&= H(\Upsilon; M_t, \eta_t),
\end{aligned} \tag{66}$$

which is the statement of Theorem H.  $\square$

Armed with Theorem H, what we now show is the boosting compliant convergence on the entropy of the  $\alpha$ -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 7, for the split of a leaf  $\lambda_q$  in a subtree with two new leaves  $\lambda_p, \lambda_r$ . Here, we make use of simplified notation

$$e_p \doteq e(M_{\lambda_p}, \eta_t), \tag{67}$$

and similarly for  $e_q$  and  $e_r$ . Quantities  $p, q, r \in [0, 1]^4$  are computed from the corresponding  $e$ .  $\tau$  is the probability, measured from  $D_{t\lambda_q}$ , that an example has  $h(\cdot) = +1$ , where  $h$  is the split function at  $\lambda_q$ . We state and prove our generalisation to Kearns & Mansour (1996,

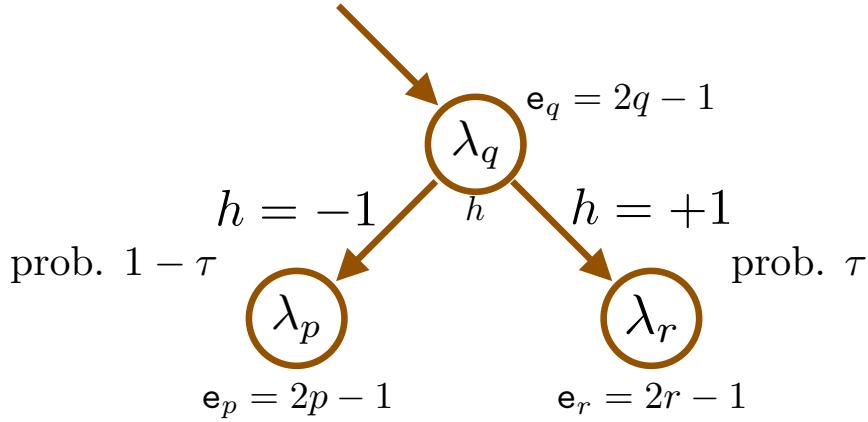


Figure 7: Main notations used in the proof of Theorem 3, closely following some notations of Kearns & Mansour (1996, Fig. 4).

Lemma 2).

---

<sup>4</sup>Under Assumption 1.

**Lemma G.** Assuming notations in Figure 7 for the split  $h$  investigated at a leaf  $\lambda_q$ , and letting  $\delta \doteq r - p$ , if for some  $\gamma > 0$  the split  $h$   $\gamma$ -witnesses the WHA at  $\lambda$ , then  $\tau(1-\tau)\delta \geq \gamma \cdot q(1-q)$ .

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

$$\begin{aligned} & \mathbb{E}_{(\mathbf{X}, \mathbf{Y}) \sim D'_{t\lambda_q}} [Y \tilde{\iota}(\eta_u(\mathbf{X})) h(\mathbf{X})] \\ &= \mathbb{E}_{(\mathbf{X}, \mathbf{Y}) \sim D_{t\lambda_q}} \left[ \frac{1 - e_q \cdot Y \cdot \tilde{\iota}(\eta_u(\mathbf{X}))}{1 - e_q^2} \cdot Y h(\mathbf{X}) \tilde{\iota}(\eta_u(\mathbf{X})) \right] \\ &= \frac{\mathbb{E}_{(\mathbf{X}, \mathbf{Y}) \sim D_{t\lambda_q}} [Y h(\mathbf{X}) \tilde{\iota}(\eta_u(\mathbf{X}))] - e_q \cdot \mathbb{E}_{(\mathbf{X}, \mathbf{Y}) \sim D_{t\lambda_q}} [\tilde{\iota}^2(\eta_u(\mathbf{X})) h(\mathbf{X})]}{1 - e_q^2}, \end{aligned} \quad (68)$$

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

$$\begin{aligned} \tau e_r - (1 - \tau) e_p &= \tau \cdot \mathbb{E}_{(\mathbf{X}, \mathbf{Y}) \sim D_{t\lambda_r}} [Y \cdot \tilde{\iota}(\eta_u(\mathbf{X}))] - (1 - \tau) \cdot \mathbb{E}_{(\mathbf{X}, \mathbf{Y}) \sim D_{t\lambda_p}} [Y \cdot \tilde{\iota}(\eta_u(\mathbf{X}))] \\ &= \mathbb{E}_{(\mathbf{X}, \mathbf{Y}) \sim D_{t\lambda_q}} [Y h(\mathbf{X}) \tilde{\iota}(\eta_u(\mathbf{X}))]. \end{aligned} \quad (69)$$

We can thus write:

$$\begin{aligned} & \mathbb{E}_{(\mathbf{X}, \mathbf{Y}) \sim D'_{t\lambda_q}} [Y \tilde{\iota}(\eta_u(\mathbf{X})) h(\mathbf{X})] \\ &= \frac{\tau e_r - (1 - \tau) e_p - e_q \cdot \mathbb{E}_{(\mathbf{X}, \mathbf{Y}) \sim D_{t\lambda_q}} [\tilde{\iota}^2(\eta_u(\mathbf{X})) h(\mathbf{X})]}{1 - e_q^2} \end{aligned} \quad (70)$$

$$= \frac{2\tau e_r - e_q \cdot \left( 1 + \mathbb{E}_{(\mathbf{X}, \mathbf{Y}) \sim D_{t\lambda_q}} [\tilde{\iota}^2(\eta_u(\mathbf{X})) h(\mathbf{X})] \right)}{1 - e_q^2} \quad (71)$$

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

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

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

Here, (70) follows from (68) and (69), (71) uses the fact that  $e_q = (1 - \tau)e_p + \tau e_r$ , (72) and (73) are convenient reformulations after adding  $2\tau e_q - 2\tau e_q$  and (74) follows from  $\mathbb{E}_{(\mathbf{X}, \mathbf{Y}) \sim D_{t\lambda_q}} [h(\mathbf{X})] = 2\tau - 1$  by definition of  $\tau$  and  $h \in \{-1, 1\}$ . Let

$$\Delta(h) \doteq e_q \cdot \mathbb{E}_{(\mathbf{X}, \mathbf{Y}) \sim D_{t\lambda_q}} [(1 - \tilde{\iota}^2(\eta_u(\mathbf{X}))) \cdot h(\mathbf{X})]. \quad (75)$$

We have  $p = (1 + \epsilon_p)/2$  (and similarly for  $q = (1 + \epsilon_q)/2$  and  $r = (1 + \epsilon_r)/2$ ), so we reformulate (73) as:

$$\begin{aligned}\mathbb{E}_{(\mathbf{X}, \mathbf{Y}) \sim D'_{t, \lambda_q}} [\mathbf{Y} \tilde{\iota}(\eta_u(\mathbf{X})) h(\mathbf{X})] &= \frac{2\tau(2r - 2q)}{4q(1 - q)} + \frac{\Delta(h)}{4q(1 - q)} \\ &= \frac{\tau(r - q)}{q(1 - q)} + \frac{\Delta(h)}{4q(1 - q)} \\ &= \frac{\tau(1 - \tau)\delta}{q(1 - q)} + \frac{\Delta(h)}{4q(1 - q)},\end{aligned}\quad (76)$$

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  $\mathbb{E}_{(\mathbf{X}, \mathbf{Y}) \sim D'_{t, \lambda_q}} [\mathbf{Y} \tilde{\iota}(\eta_u(\mathbf{X})) h(\mathbf{X})] \geq \gamma$ . We get from (76):

$$\tau(1 - \tau)\delta \geq \gamma \cdot q(1 - q) - \frac{\Delta(h)}{4}, \quad (77)$$

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  $\mathbb{E}_{(\mathbf{X}, \mathbf{Y}) \sim D'_{t, \lambda_q}} [\mathbf{Y} \tilde{\iota}(\eta_u(\mathbf{X})) h(\mathbf{X})] \leq -\gamma$ . Since  $\mathcal{H}$  is closed by negation we replace  $h$  by  $h' \doteq -h$ , which satisfies  $\mathbb{E}_{(\mathbf{X}, \mathbf{Y}) \sim D'_{t, \lambda_q}} [\mathbf{Y} \tilde{\iota}(\eta_u(\mathbf{X})) h'(\mathbf{X})] = -\mathbb{E}_{(\mathbf{X}, \mathbf{Y}) \sim D'_{t, \lambda_q}} [\mathbf{Y} \tilde{\iota}(\eta_u(\mathbf{X})) h(\mathbf{X})]$ . The change switches the sign of  $\delta$  by its definition and also  $\Delta(h') = -\Delta(h)$  so (77) becomes  $-\tau(1 - \tau)\delta \leq -\gamma \cdot q(1 - q) + \Delta(h')/4$ , i.e.

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

which brings us back to Case 1 with the switch  $h \leftrightarrow h'$  as  $h'$  satisfies  $\mathbb{E}_{(\mathbf{X}, \mathbf{Y}) \sim D'_{t, \lambda_q}} [\mathbf{Y} \tilde{\iota}(\eta_u(\mathbf{X})) h'(\mathbf{X})] \geq \gamma$ . 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  $p, q, r, \delta, \tau$  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  $H(\Upsilon; M_t, \eta_t) - H(\Upsilon(\lambda, h); M_t, \eta_t)$ ,  $\Upsilon(\lambda, h)$  being the  $\alpha$ -tree  $\Upsilon$  with the split in Figure 7 performed with  $h$  at  $\lambda$ , then we immediately get

$$H(\Upsilon; M_t, \eta_t) - H(\Upsilon(\lambda, h); M_t, \eta_t) \geq \gamma^2 q(1 - q), \quad (79)$$

which comes from Kearns & Mansour (1996, Lemma 6), and (79) can be directly used in the proof of Kearns & Mansour (1996, Theorem 9) – which unravels the local decrease of  $H(\cdot; M_t, \eta_t)$  to get to the global decrease of the criterion for the whole of  $\Upsilon$ 's induction –, and to get  $H(\Upsilon; M_t, \eta_t) \leq \varepsilon$ , it is sufficient that

$$|\Lambda(g)| \geq \left(\frac{1}{\varepsilon}\right)^{\frac{c \log(\frac{1}{\varepsilon})}{\gamma^2}}, \quad (80)$$

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

$$\Pi_{S',\lambda}(h) \leq -\beta \cdot (1 - e_q^2), \quad (81)$$

then the "advantage"  $\gamma$  becomes an advantage  $\gamma + \beta$  in (80). Since we have  $\Pi_{S',\lambda}(h) = e_q \cdot \mathbb{E}_{(\mathbf{X}, \mathbf{Y}) \sim D_{S',\lambda_q}} [(1 - \tilde{\iota}^2(\eta_u(\mathbf{X}))) \cdot h(\mathbf{X})]$ , constraint (81) quickly vanishes as  $|e_q| \rightarrow 1$ , i.e. as the black-box *ges*t 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  $\varepsilon'$ , then we just need

$$\mathbb{E}_{(\mathbf{X}, \mathbf{Y}) \sim D_{S',\lambda_q}} [(1 - \tilde{\iota}^2(\eta_u(\mathbf{X}))) \cdot h(\mathbf{X})] \leq -\varepsilon' \beta \cdot \frac{2 - \varepsilon'}{1 - \varepsilon'}. \quad (82)$$

## VIII 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

$$\vartheta(z) \doteq (z - \theta) \cdot \begin{cases} \frac{1}{B+\theta} & \text{if } z < \theta, \\ 0 & \text{if } z = \theta, \\ \frac{1}{B-\theta} & \text{if } z > \theta. \end{cases}, \quad (83)$$

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

$$l \doteq (-1, \log(1 + \exp(-\alpha))), \quad (84)$$

$$c \doteq (\theta', \log(1 + \exp(\alpha\theta'))), \quad (85)$$

$$r \doteq (1, \log(1 + \exp(\alpha))). \quad (86)$$

The equation of the line passing through  $l, c$  is

$$f_l(z) = \frac{\log\left(\frac{1+\exp(\theta'\alpha)}{1+\exp(-\alpha)}\right)}{1+\theta'} \cdot z + \frac{\log\left(\frac{1+\exp(\theta'\alpha)}{1+\exp(-\alpha)}\right)}{1+\theta'} + \log(1 + \exp(-\alpha)) \quad (87)$$

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

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

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

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

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

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

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

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

$$\vartheta'(z) \doteq (z - \theta') \cdot \begin{cases} \frac{1}{1+\theta'} & \text{if } z < \theta', \\ 0 & \text{if } z = \theta', \\ \frac{1}{1-\theta'} & \text{if } z > \theta'. \end{cases} \quad (94)$$

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

$$\begin{aligned} & \log(1 + \exp(\alpha z)) \\ & \leq \log\left(\frac{1 + \exp(\alpha)}{1 + \exp(\theta'\alpha)}\right) \cdot |\vartheta'(z)| + \alpha \min\{0, \vartheta'(z)\} + \log(1 + \exp(\theta'\alpha)), \end{aligned} \quad (95)$$

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:

$$\begin{aligned} & \log(1 + \exp(\alpha z)) \\ & = \log(1 + \exp(\alpha B \cdot (z/B))) \\ & \leq \log\left(\frac{1 + \exp(\alpha B)}{1 + \exp(\theta'\alpha B)}\right) \cdot |\vartheta'(z/B)| + \alpha B \min\{0, \vartheta'(z/B)\} + \log(1 + \exp(\theta'\alpha B)), \end{aligned} \quad (96)$$

where this time,

$$\begin{aligned} \vartheta'\left(\frac{z}{B}\right) & \doteq \left(\frac{z}{B} - \theta'\right) \cdot \begin{cases} \frac{1}{1+\theta'} & \text{if } z < B\theta', \\ 0 & \text{if } z = B\theta', \\ \frac{1}{1-\theta'} & \text{if } z > B\theta'. \end{cases} \\ & = (z - \theta) \cdot \begin{cases} \frac{1}{B+\theta} & \text{if } z < \theta, \\ 0 & \text{if } z = \theta, \\ \frac{1}{B-\theta} & \text{if } z > \theta. \end{cases} \doteq \vartheta(z). \end{aligned} \quad (97)$$

We thus get

$$\begin{aligned} & \log(1 + \exp(\alpha z)) \\ & \leq \log\left(\frac{1 + \exp(B\alpha)}{1 + \exp(\theta\alpha)}\right) \cdot |\vartheta(z)| + B\alpha \min\{0, \vartheta(z)\} + \log(1 + \exp(\theta\alpha)), \end{aligned} \quad (98)$$

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 H, we thus get (using the same notations as in the proof of Theorem 3),

$$\begin{aligned} & -\log \eta_f \\ &= \log(1 + \exp(\Upsilon \cdot -\iota(\eta_u))) \\ &\leq \frac{1}{B} \cdot \log\left(\frac{1 + \exp(B\Upsilon)}{2}\right) \cdot |\iota(\eta_u(X))| + \Upsilon \min\{0, -\iota(\eta_u(X))\} + \log(2) \\ &= \frac{1}{B} \cdot \log\left(\frac{1 + \exp(B\Upsilon)}{2}\right) \cdot |\iota(\eta_u(X))| - \Upsilon \max\{0, \iota(\eta_u(X))\} + \log(2) \end{aligned} \quad (99)$$

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

We get that the inequality in (64) now reads (for *any* values  $\{\Upsilon(\lambda), \lambda \in \Lambda(\Upsilon)\}$ )  $L(\eta_f; M_t, \eta_t) = \mathbb{E}_{\lambda \sim M_{\Lambda(\Upsilon)}} [J(\lambda)]$  with  $J(\lambda)$  satisfying:

$$\begin{aligned} & J(\lambda) \\ &\leq \mathbb{E}_{X \sim M_\lambda} \left[ \eta_t(X) \cdot \left( \frac{1}{B} \cdot \log\left(\frac{1 + \exp(B\Upsilon(\lambda))}{2}\right) \cdot |\iota(\eta_u(X))| - \Upsilon(\lambda) \max\{0, \iota(\eta_u(X))\} + \log(2) \right) \right. \\ &\quad \left. + (1 - \eta_t(X)) \cdot \left( \frac{1}{B} \cdot \log\left(\frac{1 + \exp(B\Upsilon(\lambda))}{2}\right) \cdot |\iota(\eta_u(X))| - \Upsilon(\lambda) \max\{0, -\iota(\eta_u(X))\} + \log(2) \right) \right] \\ &= \log(2) - B\Upsilon(\lambda) \cdot e^+(M_\lambda, \eta_t) \\ &\quad + \log\left(\frac{1 + \exp(B\Upsilon(\lambda))}{2}\right) \cdot (e^+(M_\lambda, \eta_t) + e^-(M_\lambda, \eta_t)), \end{aligned} \quad (101)$$

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

$$\Upsilon(\lambda) = \frac{1}{B} \cdot \log\left(\frac{e^+(M_\lambda, \eta_t)}{e^-(M_\lambda, \eta_t)}\right) = \tilde{\iota}\left(\frac{e^+(M_\lambda, \eta_t)}{e^+(M_\lambda, \eta_t) + e^-(M_\lambda, \eta_t)}\right), \quad (102)$$

yielding (using notations from Theorem 4),

$$\begin{aligned} J(\lambda) &\leq \log(2) \cdot (1 - e_\lambda^- - e_\lambda^+) - e_\lambda^+ \cdot \log\left(\frac{e_\lambda^+}{e_\lambda^-}\right) + \log\left(\frac{e_\lambda^- + e_\lambda^+}{e_\lambda^-}\right) \cdot (e_\lambda^- + e_\lambda^+) \\ &= \log(2) \cdot \left(1 + (e_\lambda^- + e_\lambda^+) \cdot \left(H_2\left(\frac{e_\lambda^+}{e_\lambda^+ + e_\lambda^-}\right) - 1\right)\right), \end{aligned} \quad (103)$$

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

## IX Proof of Lemma 3

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

$$\frac{H_2(\lambda; M, \eta_t)}{\log 2} = (1 - (e_\lambda^+ + e_\lambda^-)) \cdot H_2\left(\frac{1}{2}\right) + (e_\lambda^+ + e_\lambda^-) \cdot H_2\left(\frac{e_\lambda^+}{e_\lambda^+ + e_\lambda^-}\right), \quad (104)$$

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

$$(1 - (e_\lambda^+ + e_\lambda^-)) \cdot \left(\frac{1}{2}\right) + (e_\lambda^+ + e_\lambda^-) \cdot \left(\frac{e_\lambda^+}{e_\lambda^+ + e_\lambda^-}\right) = \frac{1 + e_\lambda^+ - e_\lambda^-}{2} = \frac{1 + e(M_\lambda, \eta_t)}{2}, \quad (105)$$

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 3.

## X 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

$$\text{KL}(\eta_t, \eta_f; M_t) \doteq \mathbb{E}_{(X,Y) \sim D_t} \left[ \log \left( \frac{dD_t((X, Y))}{dD_f((X, Y))} \right) \right] \quad (106)$$

$$= -\mathbb{E}_{X \sim M_t} \left[ \eta_t(X) \cdot \log \left( \frac{\eta_f(X)}{\eta_t(X)} \right) + (1 - \eta_t(X)) \cdot \log \left( \frac{1 - \eta_f(X)}{1 - \eta_t(X)} \right) \right] \quad (107)$$

$$= L(\eta_f; M_t, \eta_t) - \mathbb{E}_{X \sim M_t} [H(\eta_t(X))], \quad (108)$$

where  $D_t$  (resp.  $D_f$ ) is obtained from couple  $(M_t, \eta_t)$  (resp.  $(M_t, \eta_f)$ ). Denote

$$s^\circ \doteq \arg \min_s \mathbb{P}_{X \sim P_s} [h_f(X) = 1], \quad (109)$$

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

$$\text{TV}(\eta_t, \eta_f; M_t) \doteq \int_{\mathcal{X} \times \mathcal{Y}} |dD_t((X, Y)) - dD_f((X, Y))|, \quad (110)$$

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

$$\text{TV}(\eta_t, \eta_f; M_t) = \int_{\mathcal{X}} |\eta_t(\mathbf{X}) dM_t(\mathbf{X}) - \eta_f(\mathbf{X}) dM_t(\mathbf{X})| \quad (111)$$

$$+ \int_{\mathcal{X}} |(1 - \eta_t(\mathbf{X})) dM_t(\mathbf{X}) - (1 - \eta_f(\mathbf{X})) dM_t(\mathbf{X})| \quad (112)$$

$$= 2 \int_{\mathcal{X}} |\eta_t(\mathbf{X}) - \eta_f(\mathbf{X})| dM_t(\mathbf{X}). \quad (113)$$

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

$$L(\eta_f; M_t, \eta_t) \leq \frac{\tau^2}{2} + \mathbb{E}_{\mathbf{X} \sim M_t} [H(\eta_t(\mathbf{X}))], \quad (114)$$

then because of (108) and (113),

$$\int_{\mathcal{X}} |\eta_t(\mathbf{X}) - \eta_f(\mathbf{X})| dM_t(\mathbf{X}) \leq \tau. \quad (115)$$

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

$$M_t \leftarrow P_{s^\circ} \quad (116)$$

for TOPDOWN and the  $(p, \delta)$ -push up posterior  $\eta_t$ , with

$$p \doteq \mathbb{P}_{\mathbf{X} \sim P_{s^*}} [h_f(\mathbf{X}) = 1] + \frac{\delta}{2}, \quad (117)$$

assuming the RHS is  $\leq 1$ .

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

$$\int_{\mathcal{X}_{p, s^\circ}} dP_{s^\circ}(\mathbf{X}) = p. \quad (118)$$

We have two possible outcomes for  $\eta_f$  of relevance on  $\mathcal{X}_{p, s^\circ}$ : (i)  $\eta_f(\mathbf{X}) \leq 1/2$  and (ii)  $\eta_f(\mathbf{X}) > 1/2$ . Notice that in this latter case, we are guaranteed that  $h_f(\mathbf{X}) = 1$ , which counts towards bringing closer  $\mathbb{P}_{\mathbf{X} \sim P_{s^\circ}} [h_f(\mathbf{X}) = 1]$  to  $\mathbb{P}_{\mathbf{X} \sim P_{s^*}} [h_f(\mathbf{X}) = 1]$ , so we have to make sure that (i) occurs with sufficiently small probability, and this is achieved via guarantee (115).

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

$$\begin{aligned} \int_{\mathcal{X}} |\eta_t(\mathbf{X}) - \eta_f(\mathbf{X})| dP_{s^\circ}(\mathbf{X}) &\geq \int_{\mathcal{X}_{p, s^\circ}} |\eta_t(\mathbf{X}) - \eta_f(\mathbf{X})| dP_{s^\circ}(\mathbf{X}) \\ &\geq \left| \frac{1}{2} + \delta - \frac{1}{2} \right| \cdot \int_{\mathcal{X}_{p, s^\circ}} [\eta_f(\mathbf{X}) \leq 1/2] dP_{s^\circ}(\mathbf{X}) \\ &> \left| \frac{1}{2} + \delta - \frac{1}{2} \right| \cdot \delta \\ &= \delta^2. \end{aligned} \quad (119)$$

If we have the relationship  $\delta = \sqrt{\tau}$ , then we get a contradiction with (115). In conclusion, if (114) holds, then

$$\int_{\mathcal{X}_{p,s^\circ}} [\eta_f(\mathbf{X}) \leq 1/2] dP_{s^\circ}(\mathbf{X}) \leq \delta. \quad (120)$$

In summary, for any  $\tau > 0$ , if we run TOPDOWN with the choices  $M_t \leftarrow P_{s^\circ}$  (which corresponds to the "worst treated" subgroup with respect to EOO) and craft the  $(p, \delta)$ -push up posterior  $\eta_t$  with  $p$  as in (117), then

$$\mathbb{P}_{\mathbf{X} \sim P_{s^\circ}} [h_f(\mathbf{X}) = 1] \geq \int_{\mathcal{X}_{p,s^\circ}} [\eta_f(\mathbf{X}) > 1/2] dP_{s^\circ}(\mathbf{X}) \quad (121)$$

$$= \int_{\mathcal{X}_{p,s^\circ}} (1 - [\eta_f(\mathbf{X}) \leq 1/2]) dP_{s^\circ}(\mathbf{X}) \quad (122)$$

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

$$\geq p - \delta \quad (124)$$

$$= \mathbb{P}_{\mathbf{X} \sim P_{s^*}} [h_f(\mathbf{X}) = 1] - \frac{\delta}{2}, \quad (125)$$

where (124) makes use of (118) and (120). Fixing  $\delta \doteq 2\varepsilon$ ,  $\varepsilon$  being used in (16) (main file), we obtain

$$\mathbb{P}_{\mathbf{X} \sim P_{s^*}} [h_f(\mathbf{X}) = 1] - \mathbb{P}_{\mathbf{X} \sim P_{s^\circ}} [h_f(\mathbf{X}) = 1] \leq \varepsilon, \quad (126)$$

and via relationship  $\delta = \sqrt{\tau}$ , we check that (114) becomes the following function of  $\varepsilon$ :

$$L(\eta_f; M_t, \eta_t) \leq 8\varepsilon^4 + \mathbb{E}_{\mathbf{X} \sim M_t} [H(\eta_t(\mathbf{X}))], \quad (127)$$

and we get the statement of the Theorem for the choice (117), which corresponds to  $K = 2$  and reads

$$p \doteq \mathbb{P}_{\mathbf{X} \sim P_{s^*}} [h_f(\mathbf{X}) = 1] + \varepsilon. \quad (128)$$

If the RHS in (128) is not  $\leq 1$ , we can opt for an alternative with one more free variable,  $K \geq 1$ ,

$$p \doteq \mathbb{P}_{\mathbf{X} \sim P_{s^*}} [h_f(\mathbf{X}) = 1] + \frac{\delta}{K}, \quad (129)$$

where  $K$  is large enough for the constraint to hold. In this case, to keep (126) we must have  $\delta(K-1)/K = \varepsilon$ , which elicitates

$$\delta = \frac{K\varepsilon}{K-1} \quad (130)$$

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

$$p \doteq \mathbb{P}_{\mathbf{X} \sim P_{s^*}} [h_f(\mathbf{X}) = 1] + \frac{\varepsilon}{K-1}, \quad (131)$$

and a desired approximation guarantee for TOPDOWN of:

$$L(\eta_f; M_t, \eta_t) \leq \frac{K^4}{2(K-1)^4} \cdot \varepsilon^4 + \mathbb{E}_{X \sim M_t} [H(\eta_t(X))]. \quad (132)$$

Since  $K > 1$ ,  $K^4/(K-1)^4 \geq 1$ , so we are guaranteed that (132) holds if we ask for

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

## XI SI Experiment Settings

In this SI section, we briefly discuss the additional datasets<sup>5</sup> 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. German Credit and Bank are standard public benchmark datasets in the literature. ACS is a newer dataset with curation details listed in (Ding et al., 2021, Section 3).

### 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. Out 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.

`AIF360` uses a Apache License 2.0 and `Folktables` uses a MIT License.

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.

---

<sup>5</sup>Public at: <https://github.com/Trusted-AI/AIF360>

- **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 XVII.

For the criteria we evaluate TOPDOWN and baselines to in the SI, we consider those introduced in the main-text alongside AUC as an additional metric for accuracy.

## Compute

Compute was done with no GPUs. Virtual machines were used with RAM 16GB VCPUs 8 VCPU Disk 30GB from [a local HPC facility] (to be named after publication).

## Code

The code used in this submission is attached in the supplementary material, which will be released upon publication.

## TopDown Specifics

The  $\alpha$ -trees learnt by TOPDOWN are initialized as per Fig. 3. That is, we initialize sub- $\alpha$ -trees with  $\alpha = 1$  for each of the modalities of the sensitive attribute. In addition, each split of the  $\alpha$ -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.

For EOO we utilize a Gaussian Naive Bayes classifier from `scikit-learn` with default parameters to fit  $\eta^*$  from data. We note that no fine-tuning was done for this classifier.

In the SI, we consider all variants TOPDOWN examined in the main-text. Additionally we consider the symmetric variant of the SP TOPDOWN approach. We reiterate the two ways of enforcing SP: as per Section 6, there are two symmetric strategies for SP. In particular, we aim to match the either the maximum or minimum subgroup to the opposite extreme (conditionally) expected posterior. As such, we can either match to the largest posterior, which we denote as  $(\uparrow)$ , or we can match to the smallest posterior  $(\downarrow)$ . We already present SP  $\uparrow$  in the main-text and additionally present evaluation for SP  $\downarrow$  here.

## XII Experiments on Statistical Parity

We refer to SI, Section II for the formal aspects of handling statistical parity. **For SP**, a similar pattern to that of EOO follows, except the differences are more extreme. In particular, the audacious approach fails to optimize for SP and instead harms it significantly, but does slightly improve MD. The audacious update is problematic here as the target  $\eta_t$  in the SP strategy is a constant (and does not take into consideration of the subgroup being updated). As such the audacious approach should not be used for SP as it will optimize to match the constant  $\eta_t$  more harshly. On the other hand, the conservative update variant provides an improvement to SP whilst antagonizing MD accuracy. Notably, the “best” iteration for SP and MD occurs at its first iteration (which shows interest in early stopping / pruning the  $\alpha$ -tree). This is expected, as there is large shift when changing from an  $\alpha = 1$  to the initial rooted value (*e.g.*, (5)). The pattern of conservative updates being superior to the audacious counterparts is consistent in other datasets and sensitive attribute modalities. Comparing the conservative SP TOPDOWN to the baselines, discounting OST for MD, we find that DERSP and RTO result in lower SP and MD. This is unsurprising: our TOPDOWN treatment SP can result in harsh updates; in SI (pg 19), we discuss an alternative approach using ties with optimal transport.

## XIII Additional Main Text Experiments

In this section, we report the experiments identical to that presented in the main-text, including missing criteria, settings, and the additional German Credit and Bank datasets. Each plot we present provides the binary and trinary sensitive attribute settings over all criteria discussed in the previous setting.

In particular:

- Fig. 8 presents the evaluation using a RF black-box with  $B = 1$  clipping on the German Credit dataset.
- Fig. 9 presents the evaluation using a RF black-box with  $B = 1$  clipping on the Bank dataset.
- Fig. 10 presents the evaluation using a RF black-box with  $B = 1$  clipping on the ACS dataset.

## Fairness Models

In comparison to ACS, Fig. 9 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  $\uparrow$ ), 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. Thus for such a failure cases, one should confirm that TOPDOWN is not decreasing fairness to prevent social harms.

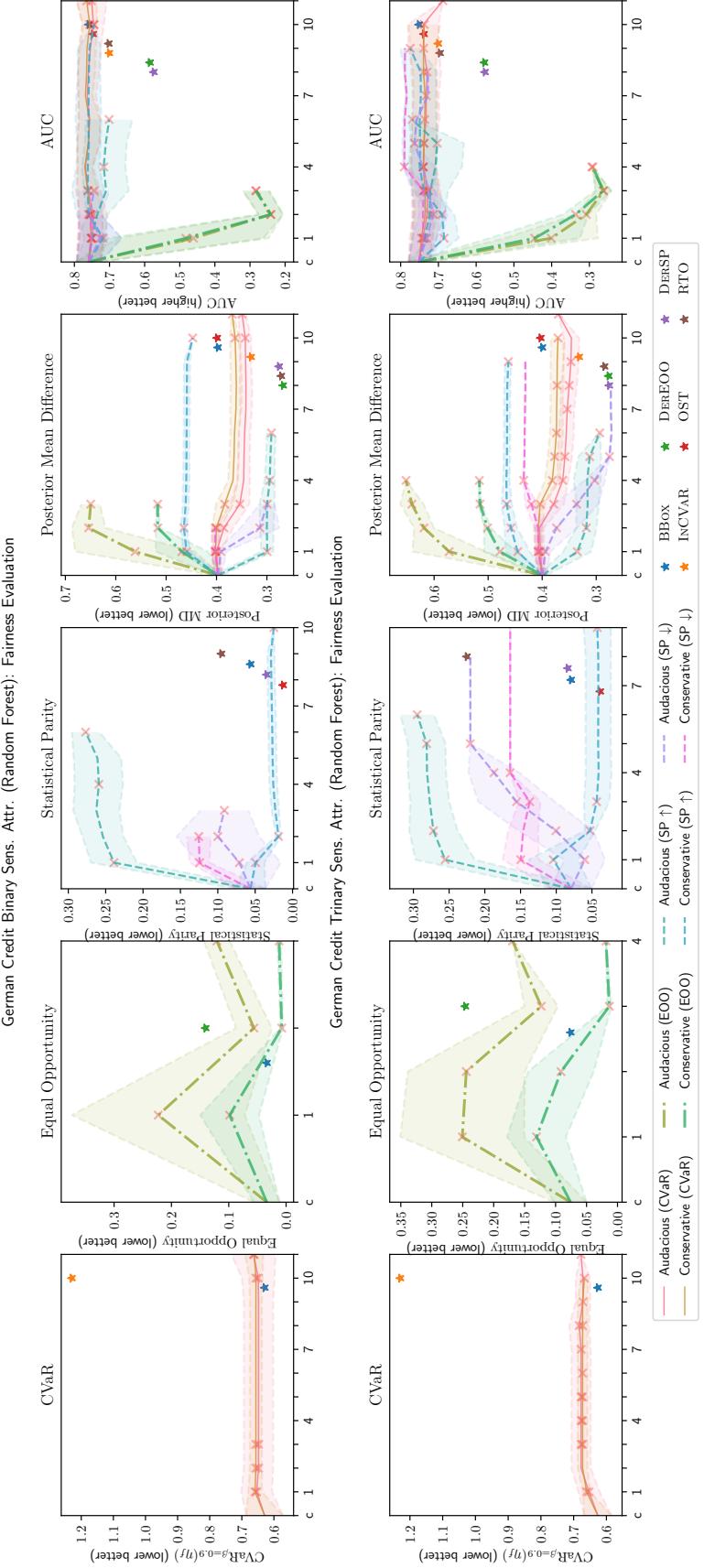


Figure 8: TOPDOWN optimized for different fairness models evaluated on German Credit with binary (up) and trinary (down) sensitive attributes. Crosses denote when a subgroup's  $\alpha$ -tree is initiated (over any fold). The shade depicts  $\pm$  a standard deviation from the mean. However, this disappears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

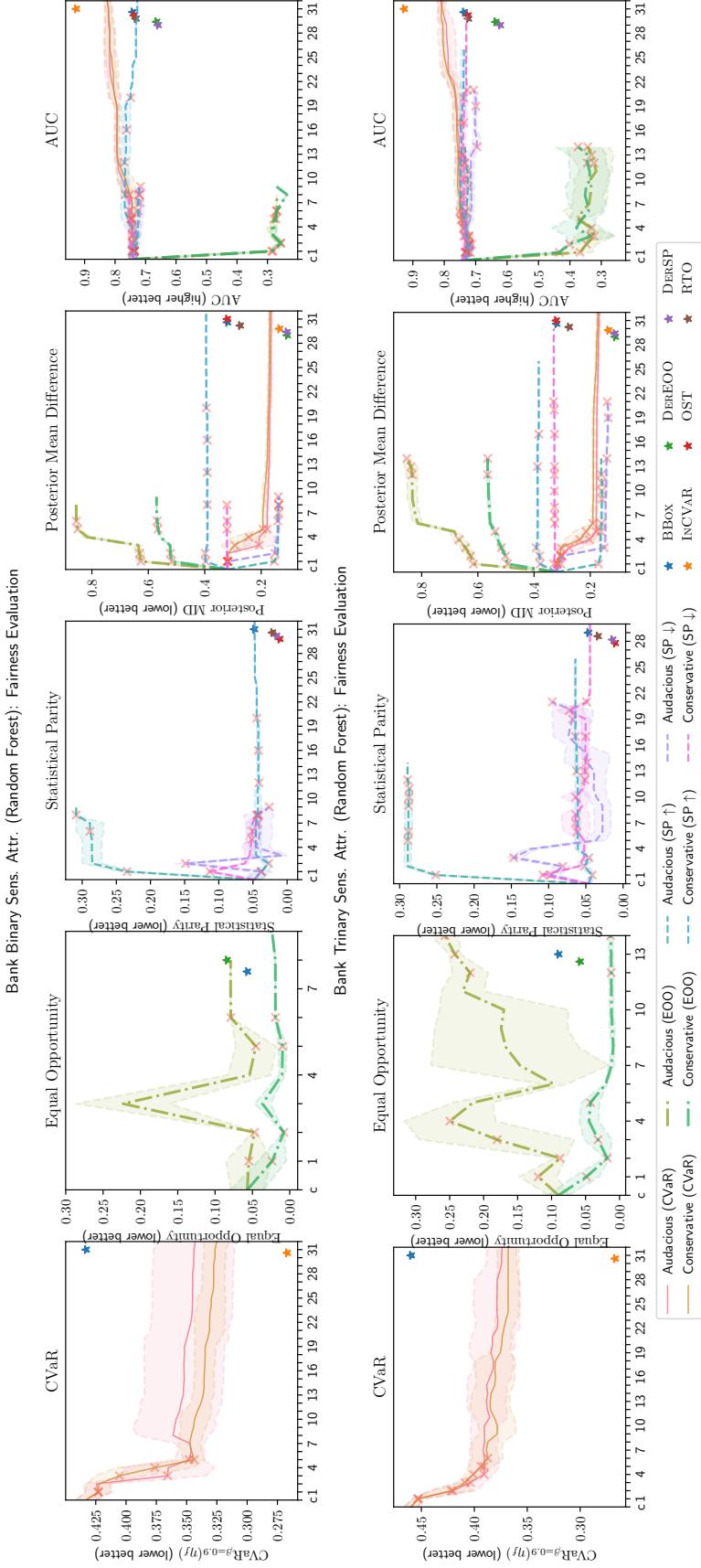


Figure 9: TOPDOWN optimized for different fairness models evaluated on Bank with binary (up) and trinary (down) sensitive attributes. Crosses denote when a subgroup's  $\alpha$ -tree is initiated (over any fold). The shade depicts  $\pm$  a standard deviation from the mean. However, this disappears in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

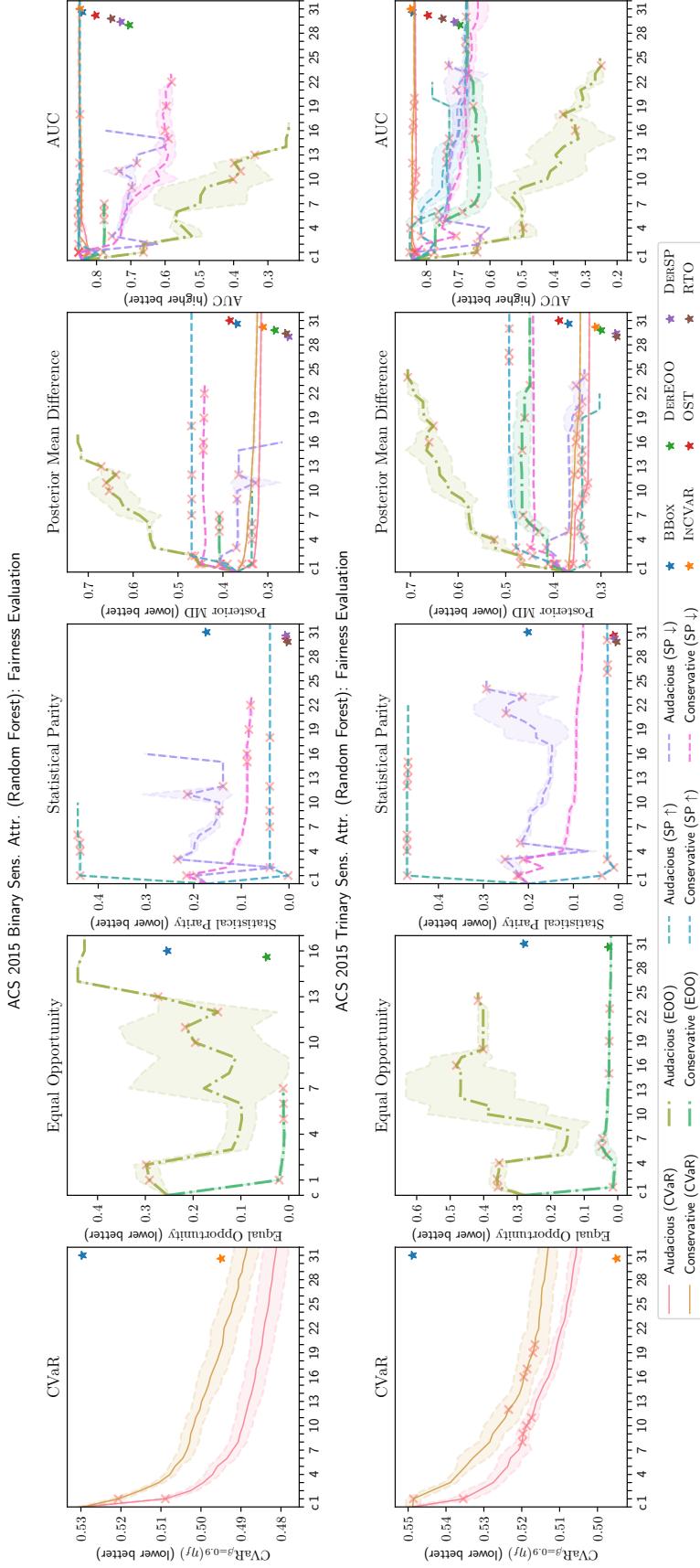


Figure 10: TOPDOWN optimized over boosting iterations for different fairness models evaluated on ACS 2015 with binary (up) and trinary (down) sensitive attributes. “c” on the x-axis denotes the clipped black-box. Crosses denote when a subgroup’s  $\alpha$ -tree is initiated (over any fold). The shade depicts  $\pm$  a standard deviation from the mean. However, this disappears in the case where other folds stop early.

## XIV 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. 11 to 13 presents neural network equivalent plots to that of Figs. 8 to 10.

In particular:

- Fig. 11 presents the evaluation using a NN black-box with  $B = 1$  clipping on the German Credit dataset.
- Fig. 12 presents the evaluation using a NN black-box with  $B = 1$  clipping on the Bank dataset.
- Fig. 13 presents the evaluation using a NN black-box with  $B = 1$  clipping on the ACS dataset.

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. 11.

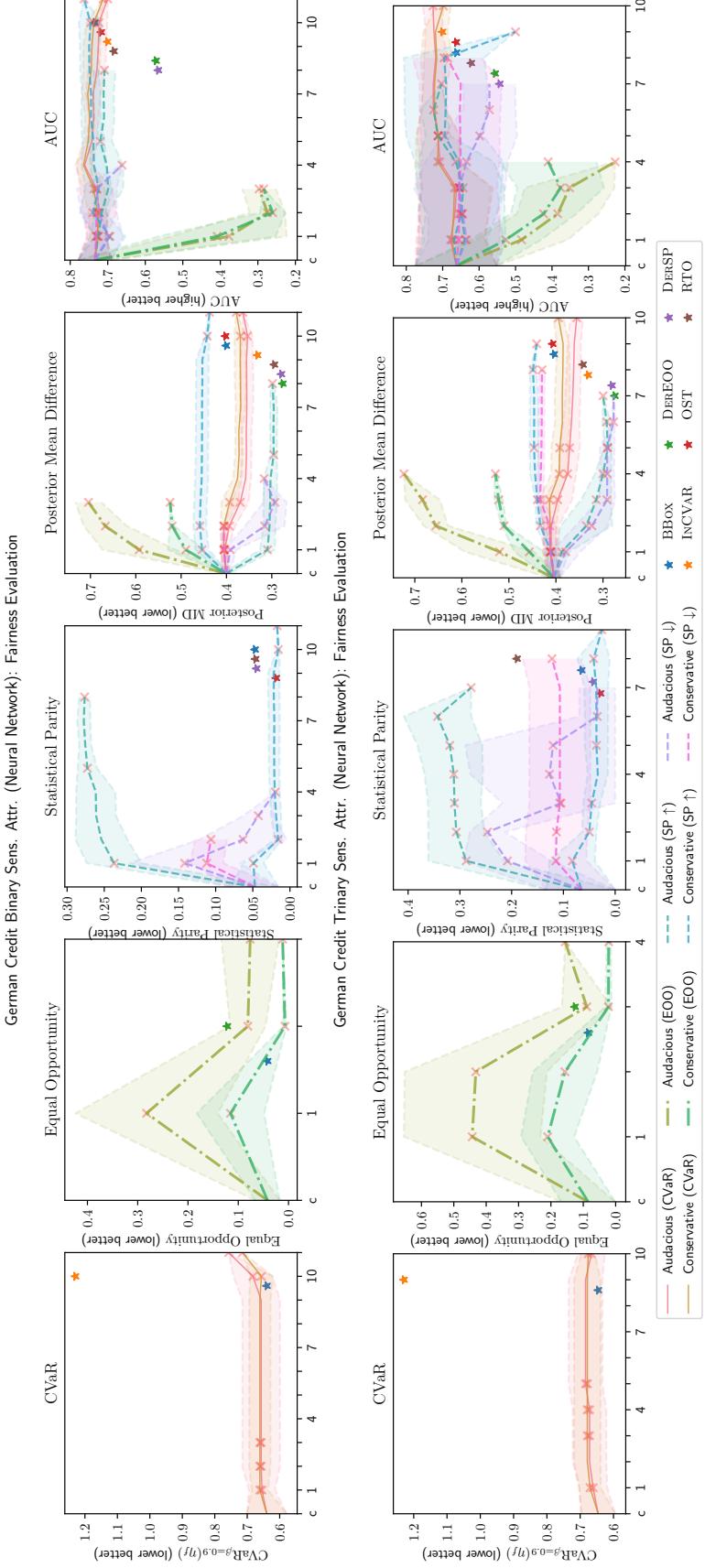


Figure 11: TOPDOWN optimized for different fairness models evaluated on German Credit with binary (up) and trinary (down) sensitive attributes. Crosses denote when a subgroup's  $\alpha$ -tree is initiated (over any fold). The shade depicts  $\pm$  a standard deviation from the mean. However, this disappears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

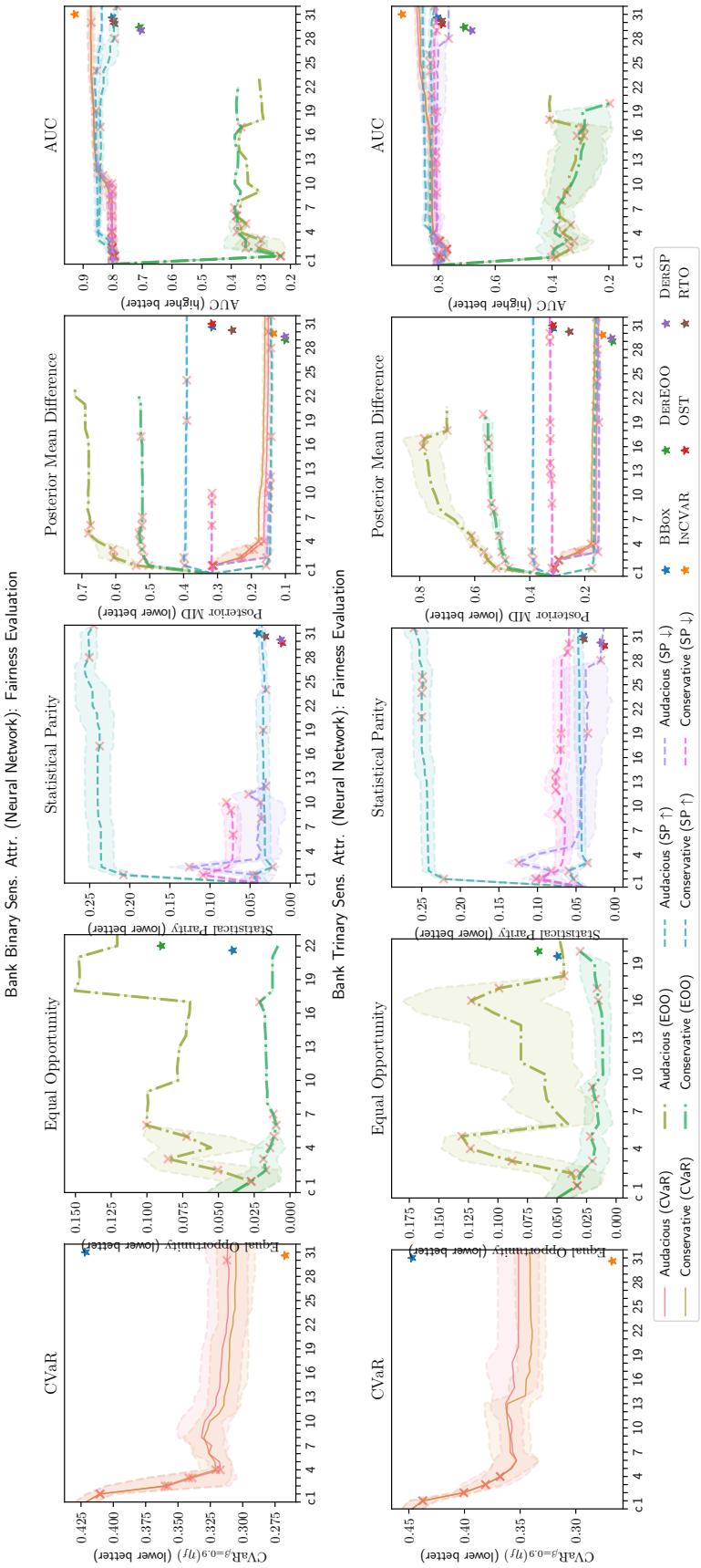


Figure 12: **TOPDOWN** optimized for different fairness models evaluated on Bank with binary (up) and trinary (down) sensitive attributes. Crosses denote when a subgroup's  $\alpha$ -tree is initiated (over any fold). The shade depicts  $\pm$  a standard deviation from the mean. However, this disappears in the case where other folds stop early.

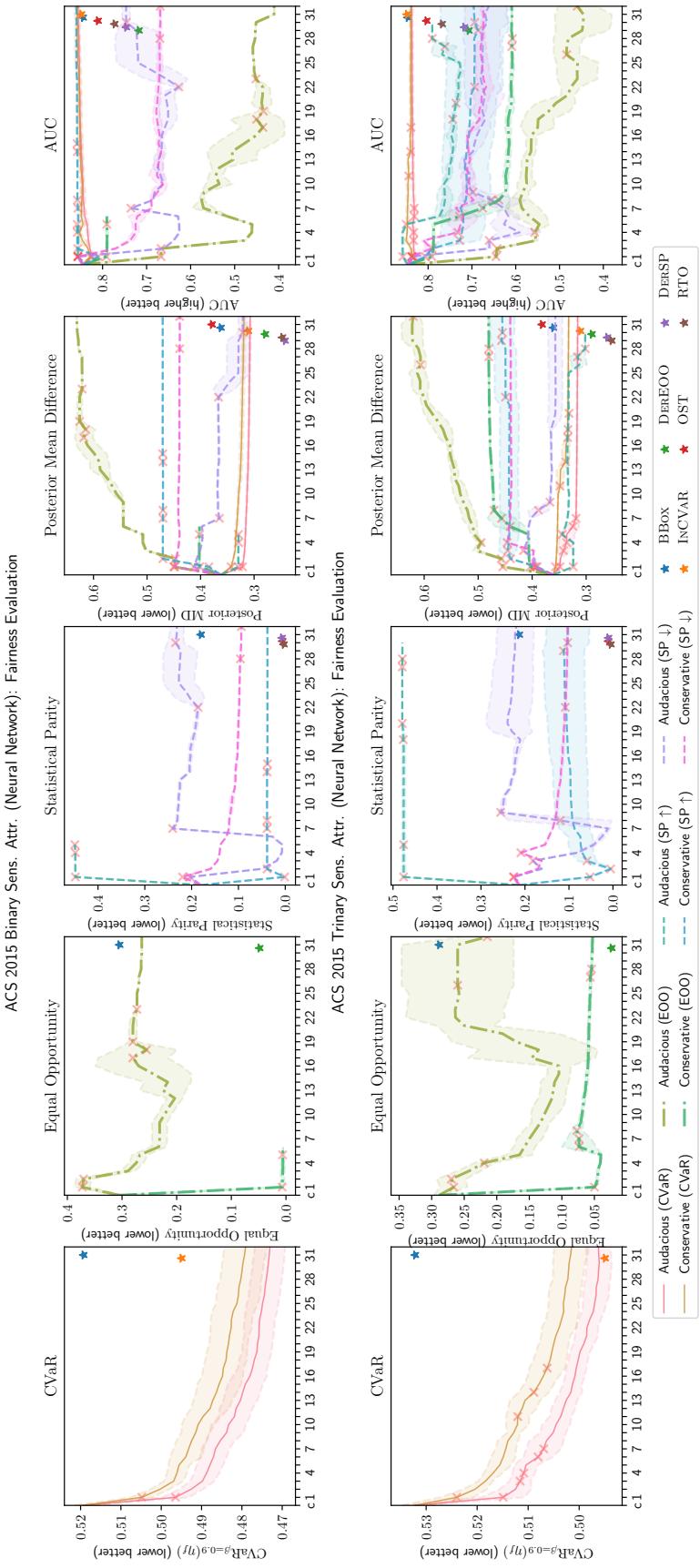


Figure 13: **TOPDOWN** optimized for different fairness models evaluated on Bank with binary (up) and trinary (down) sensitive attributes. Crosses denote when a subgroup's  $\alpha$ -tree is initiated (over any fold). The shade depicts  $\pm$  a standard deviation from the mean. However, this disappears in the case where other folds stop early.

## XV 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. We present results for both RF and NN black-boxes.

In particular:

- Fig. 14 presents the proxy sensitive attribute evaluation using a RF and NN black-box with  $B = 1$  clipping on the German Credit dataset.
- Fig. 15 presents the proxy sensitive attribute evaluation using a RF and NN black-box with  $B = 1$  clipping on the Bank dataset.
- Fig. 16 presents the proxy sensitive attribute evaluation using a RF and NN black-box with  $B = 1$  clipping on the ACS dataset.

Fig. 16 (top) 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 latter criterion). Remark the non-trivial nature of the proxy approach, as growing the  $\alpha$ -tree is based on groups learned at the decision tree leaves *but* the CVAR computation still relies on the *original* sensitive grouping.

Figs. 14 and 15 (top) 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. 14 to 16 bottom), results are quite similar with slight increases in CVAR from the change in black-box. One notable difference can be seen in Fig. 16. 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.

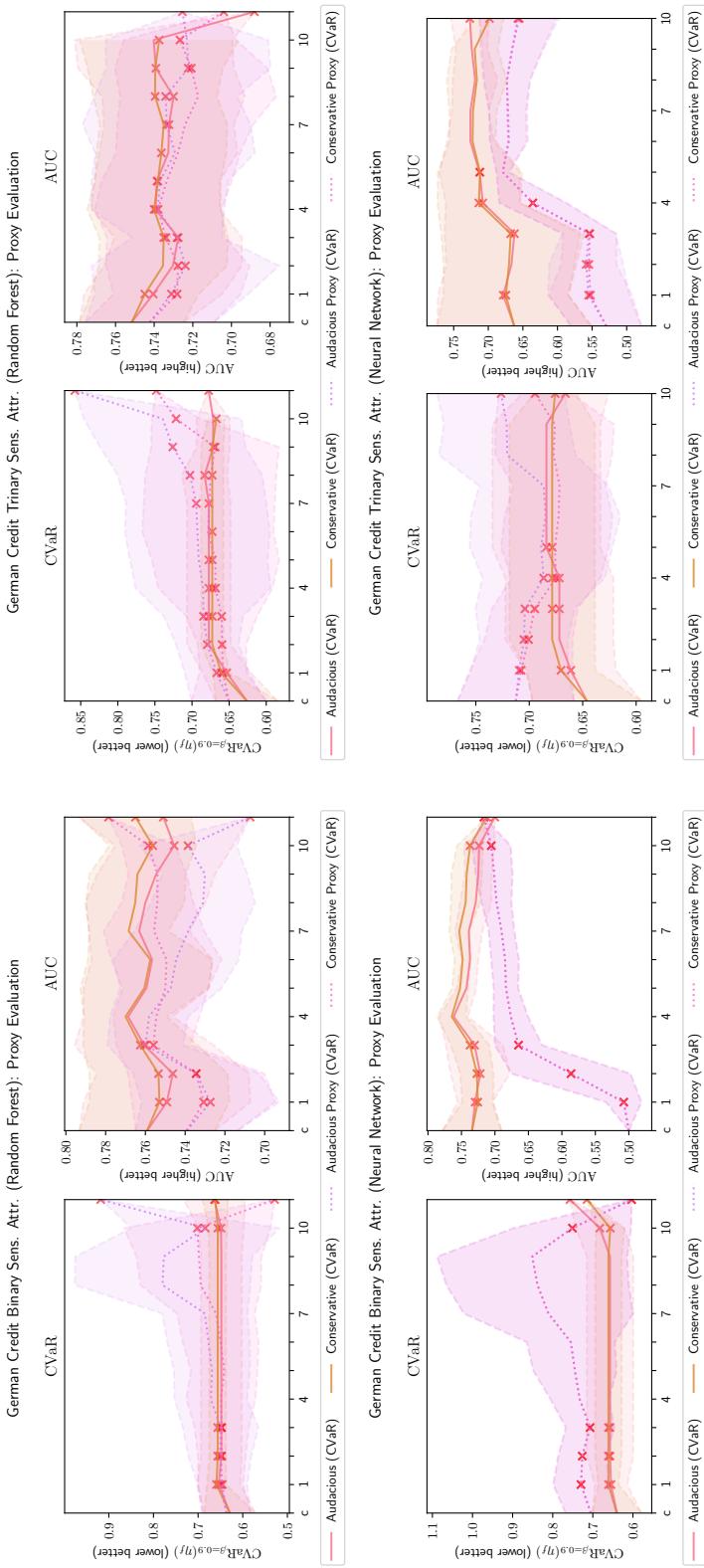


Figure 14: RF (top) and MLP (bottom) evaluation of replacing sensitive attributes with a proxy decision tree on the German Credit datasets.

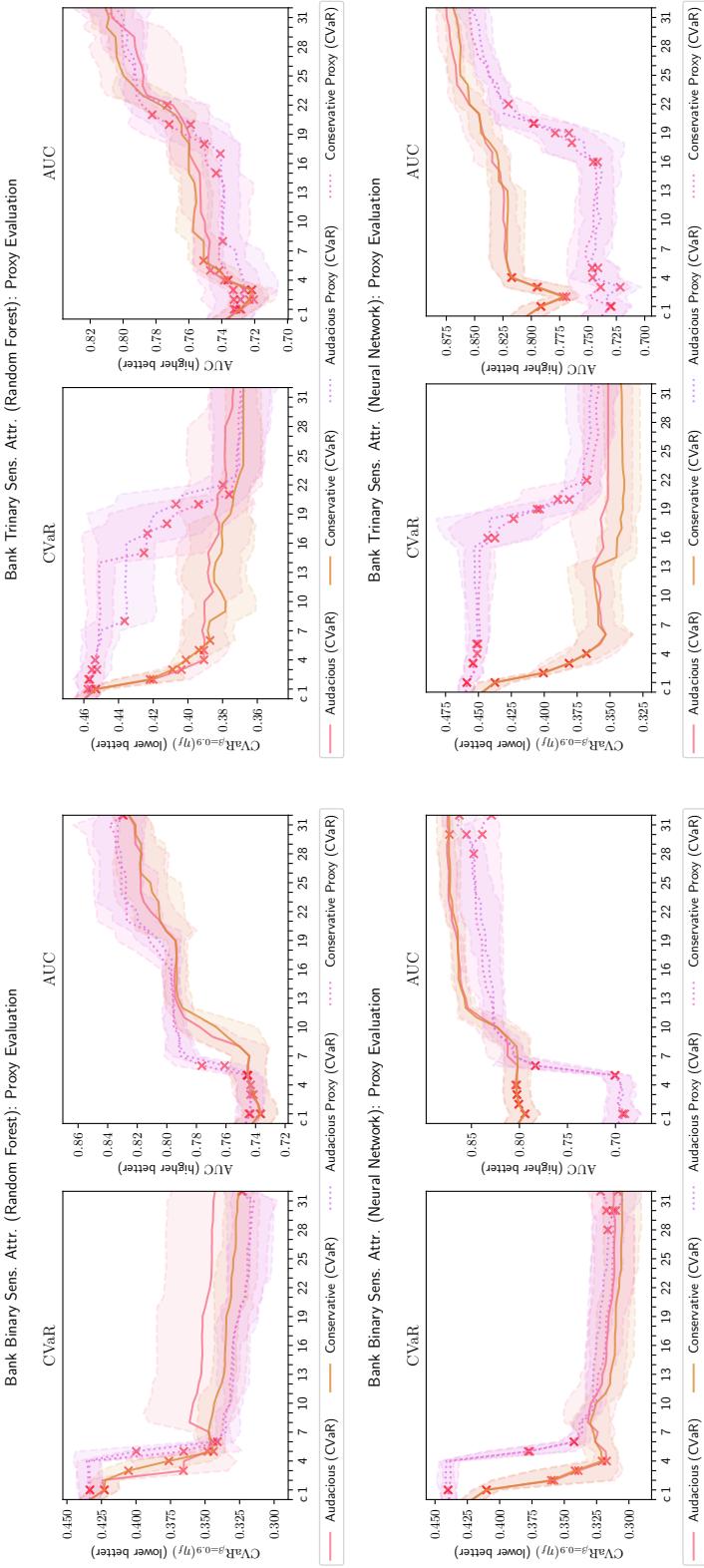


Figure 15: RF (top) and MLP (bottom) evaluation of replacing sensitive attributes with a proxy decision tree on the Bank datasets.

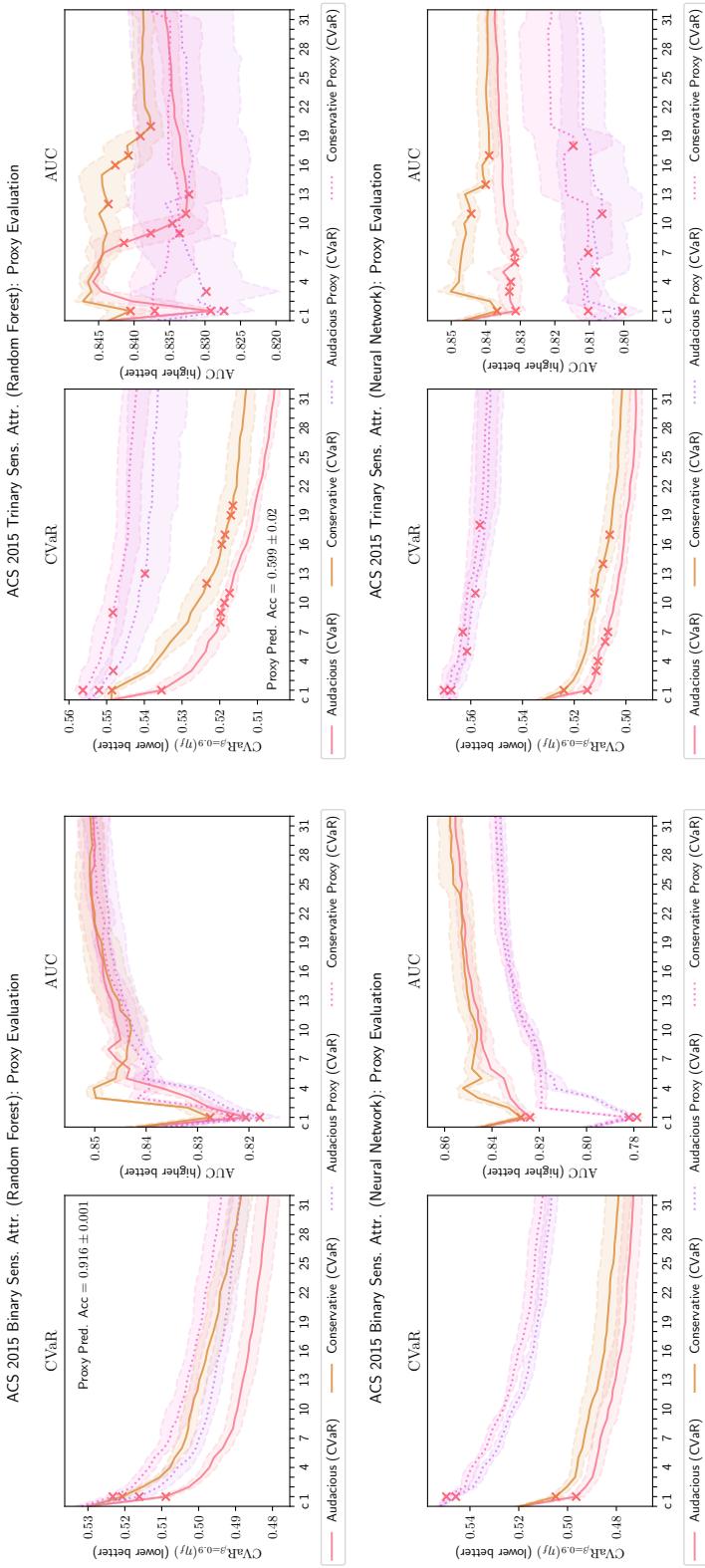


Figure 16: RF (top) and MLP (bottom) evaluation of replacing sensitive attributes with a proxy decision tree on the ACS 2015 datasets.

## XVI 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. 17 and 18 report the CVAR values over the multiple years for the random forest (RF) black-box. Figs. 19 and 20 likewise reports corresponding results for neural network (NN) black-boxes.

In particular:

- Fig. 17 presents the conservative update distribution shift evaluation using a RF black-box with  $B = 1$  clipping on the ACS dataset over years 2015 to 2018.
- Fig. 18 presents the aggressive update distribution shift evaluation using a RF black-box with  $B = 1$  clipping on the ACS dataset over years 2015 to 2018.
- Fig. 19 presents the conservative update distribution shift evaluation using a MLP black-box with  $B = 1$  clipping on the ACS dataset over years 2015 to 2018.
- Fig. 20 presents the aggressive update distribution shift evaluation using a MLP black-box with  $B = 1$  clipping on the ACS dataset over years 2015 to 2018.

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. 17. 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. 17 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 17), 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  $\alpha$ -tree in TOPDOWN often does not provide an  $\alpha$ -correction for all subgroups, *i.e.*, at least one subgroup is not changed by the  $\alpha$ -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.

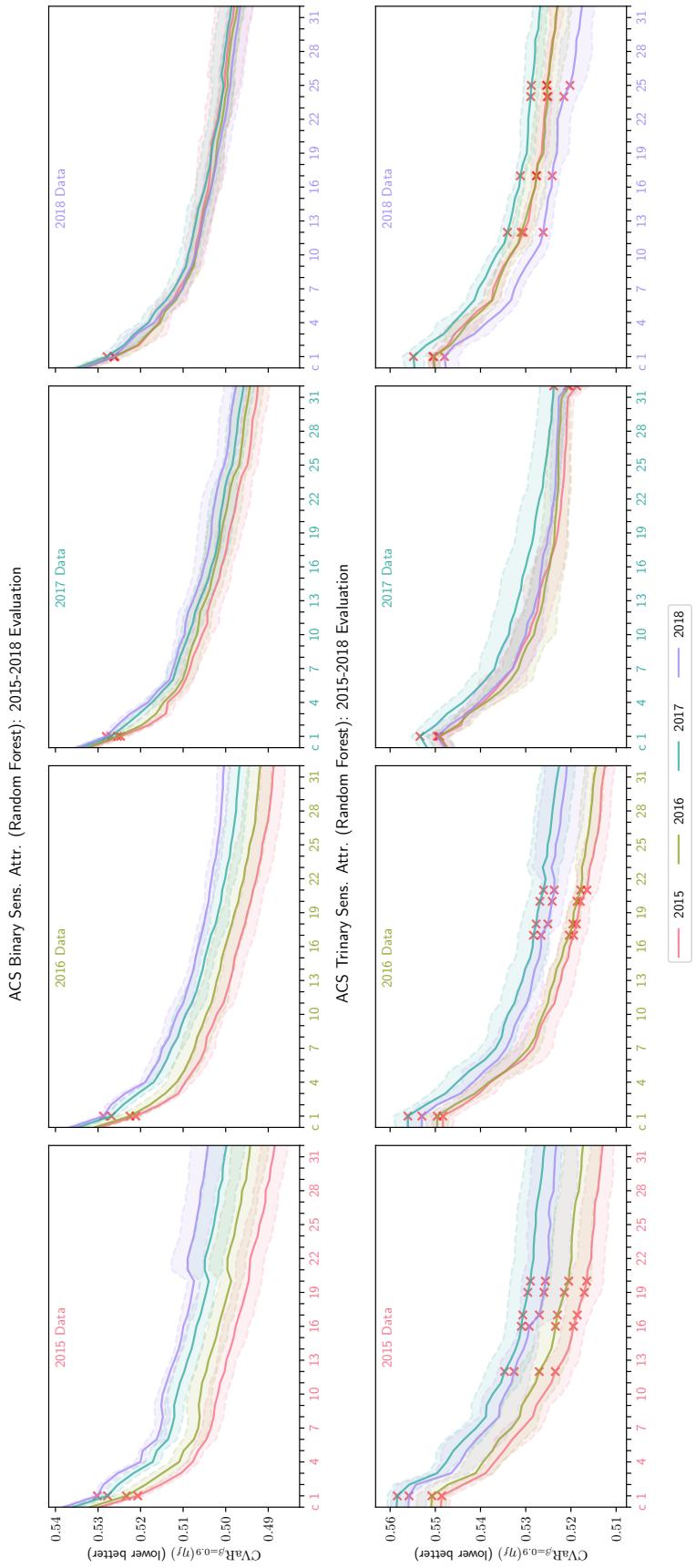


Figure 17: Random forest black-box conservative CVAR wrapper trained for ACS 2015 to 2018 datasets. Each plot is trained on a different dataset year. Each curve colour, indicates the data being used to evaluate the wrapper.

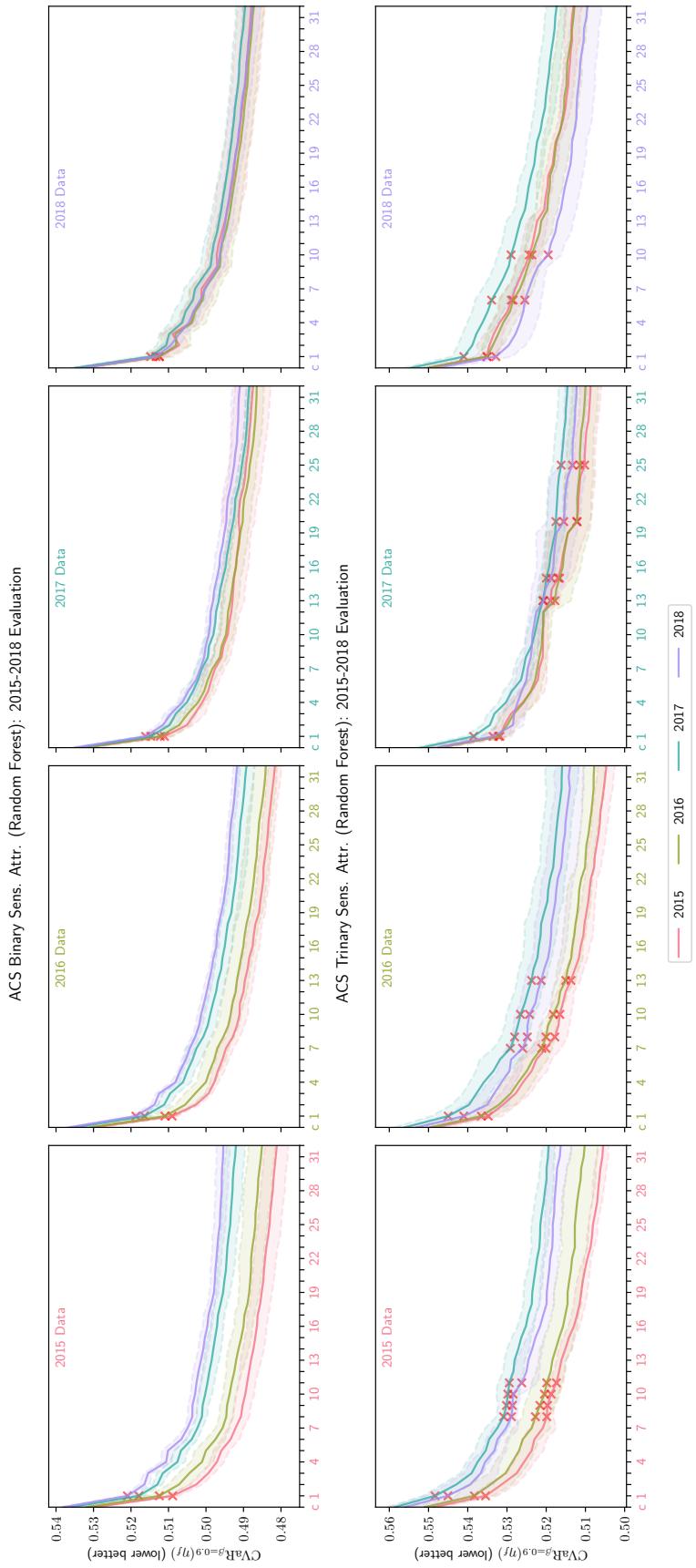


Figure 18: Random forest black-box aggressive CVaR wrapper trained for ACS 2015 to 2018 datasets. Each plot is trained on a different dataset year. Each curve colour, indicates the data being used to evaluate the wrapper.

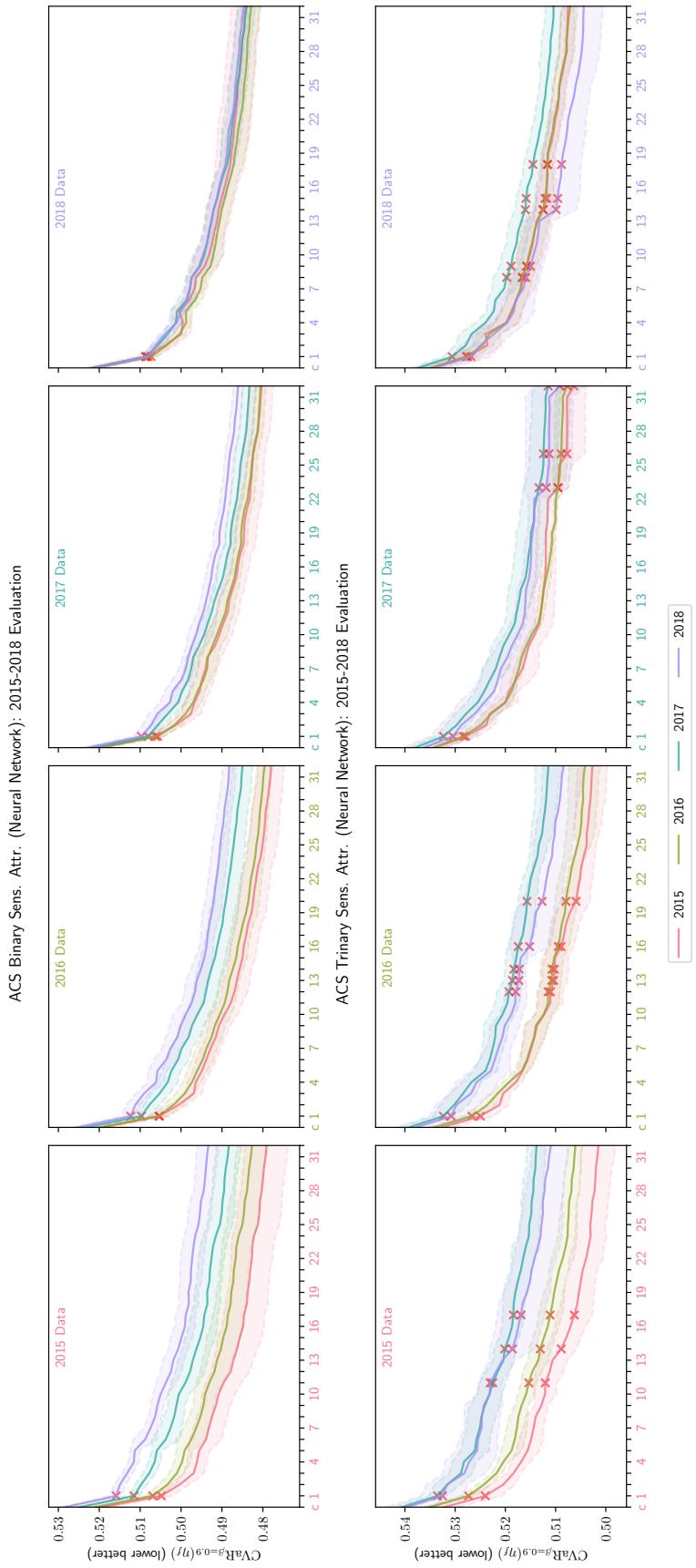


Figure 19: Neural Network black-box conservative CVaR wrapper trained for ACS 2015 to 2018 datasets. Each plot is trained on a different dataset year. Each curve colour, indicates the data being used to evaluate the wrapper.

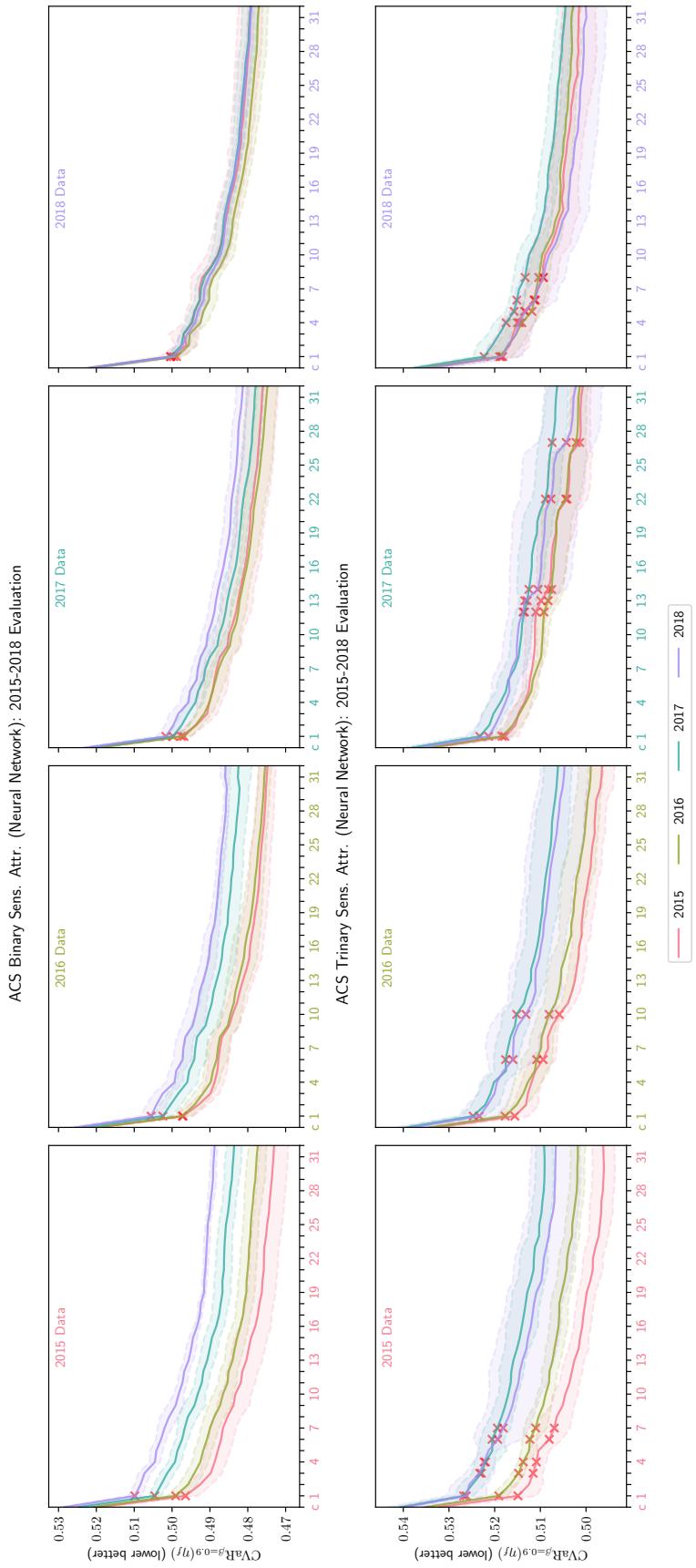


Figure 20: Neural Network black-box aggressive CVaR wrapper trained for ACS 2015 to 2018 datasets. Each plot is trained on a different dataset year. Each curve colour, indicates the data being used to evaluate the wrapper.

## XVII 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. 21 to 23 presents RF plots over German, Bank, and ACS datasets; and Figs. 24 to 26 presents equivalent MLP plots.

In particular:

- Fig. 21 presents the evaluation using a RF black-box with  $B = 3$  clipping on the German Credit dataset.
- Fig. 22 presents the evaluation using a RF black-box with  $B = 3$  clipping on the Bank dataset.
- Fig. 23 presents the evaluation using a RF black-box with  $B = 3$  clipping on the ACS dataset.
- Fig. 24 presents the evaluation using a NN black-box with  $B = 3$  clipping on the German Credit dataset.
- Fig. 25 presents the evaluation using a NN black-box with  $B = 3$  clipping on the Bank dataset.
- Fig. 26 presents the evaluation using a NN black-box with  $B = 3$  clipping on the ACS dataset.

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. 23. 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. 23 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  $\alpha$ -tree with correct corrections. However, as the the original black-box is not an  $\alpha$ -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 Fig. 10. 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. 10, 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. 23.

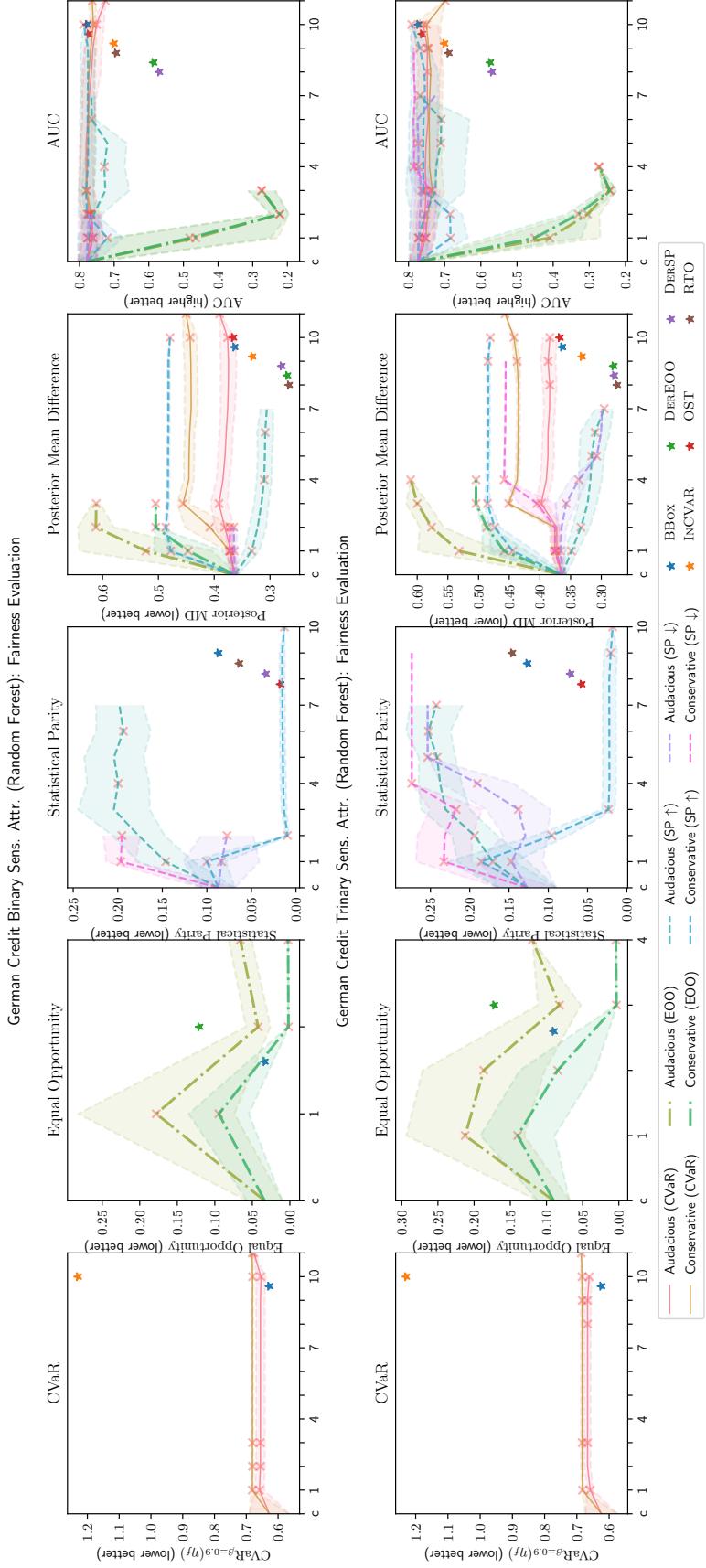


Figure 21: RF with  $B = 3$  TOPDOWN optimized for different fairness models evaluated on German Credit with binary (up) and trinary (down) sensitive attributes. Crosses denote when a subgroup's  $\alpha$ -tree is initiated (over any fold). The shade depicts  $\pm$  a standard deviation from the mean. However, this disappears in the case where other folds stop early.

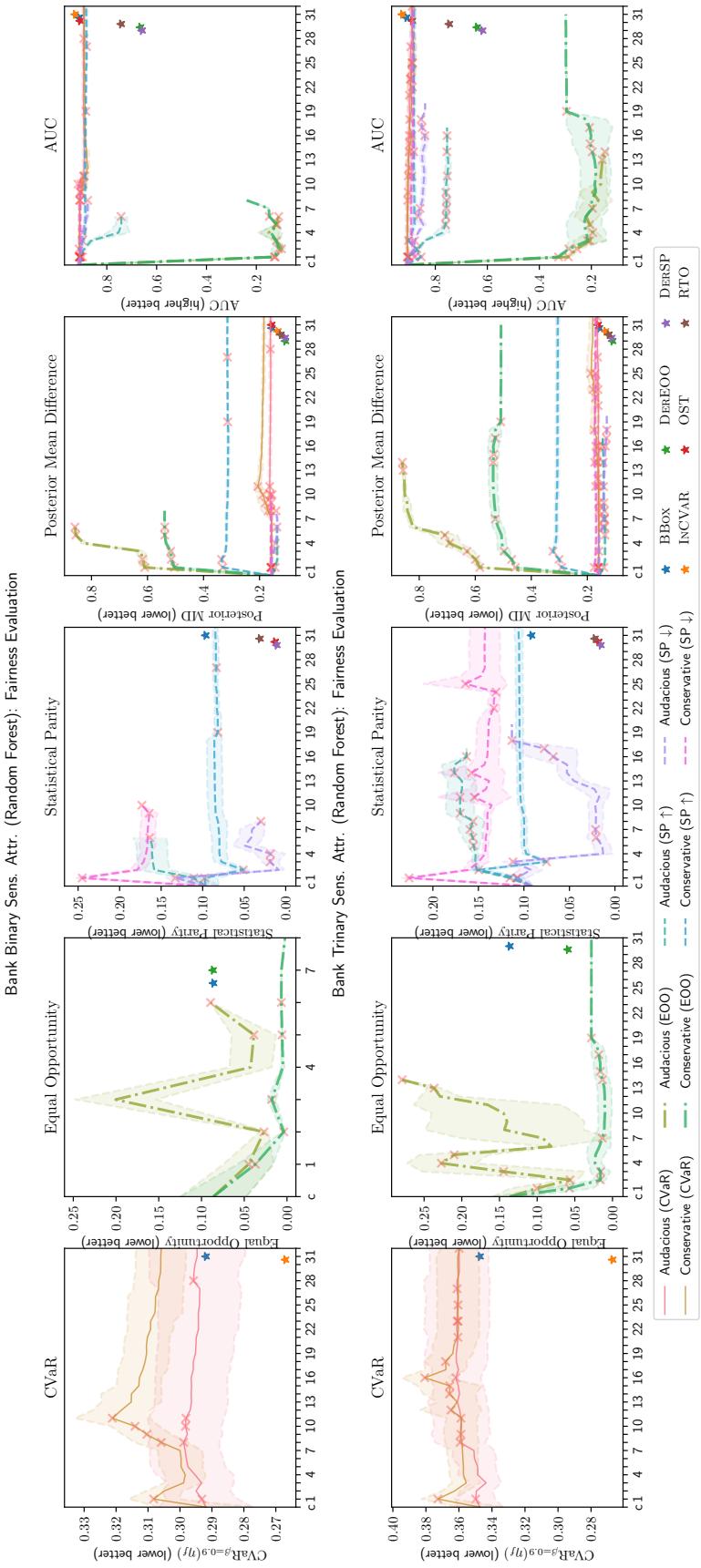


Figure 22: RF with  $B = 3$  TopDown optimized for different fairness models evaluated on Bank with binary (up) and trinary (down) sensitive attributes. Crosses denote when a subgroup's  $\alpha$ -tree is initiated (over any fold). The shade depicts  $\pm$  a standard deviation from the mean. However,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

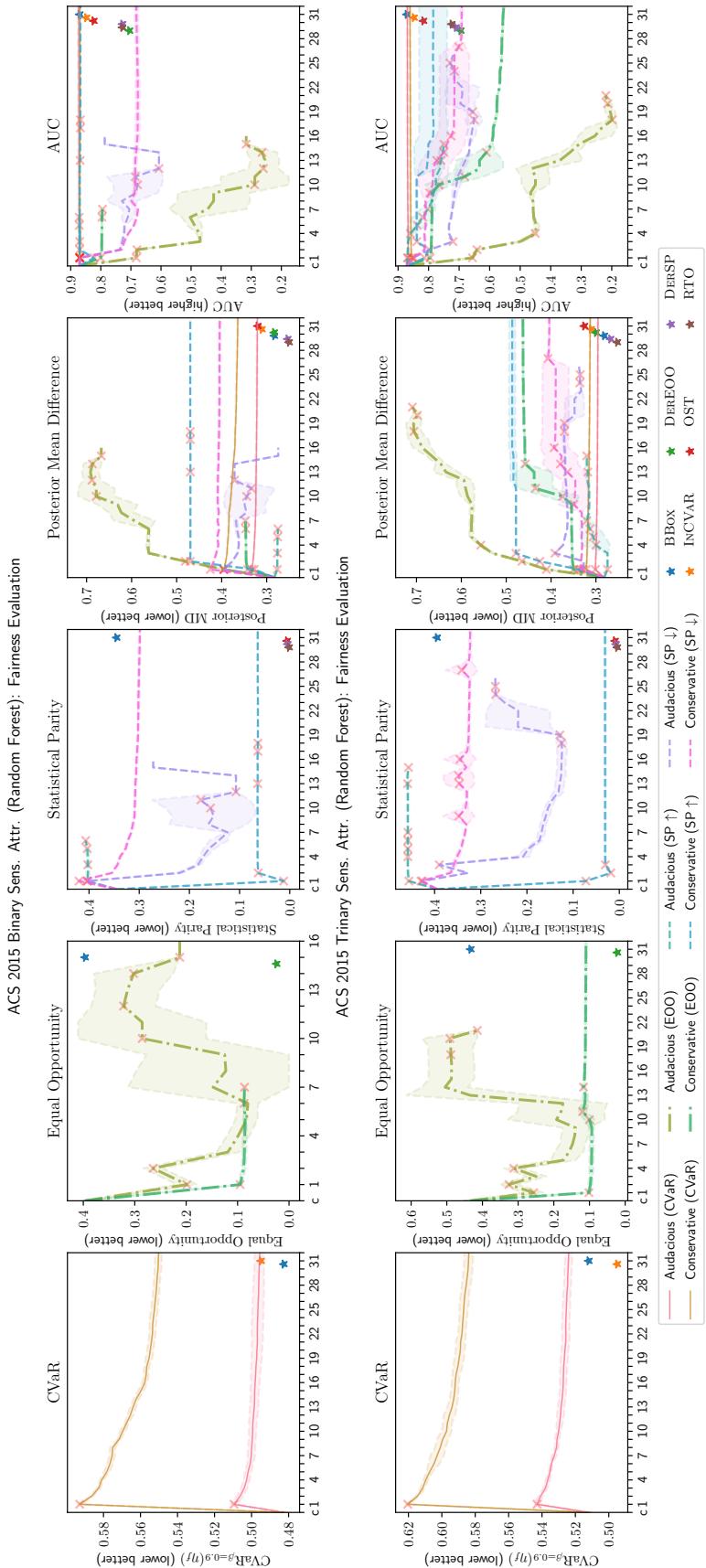


Figure 23: RF with  $B = 3$  TopDown optimized for different fairness models evaluated on Bank with binary (up) and trinary (down) sensitive attributes. Crosses denote when a subgroup’s  $\alpha$ -tree is initiated (over any fold). The shade depicts  $\pm$  a standard deviation from the mean. However,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

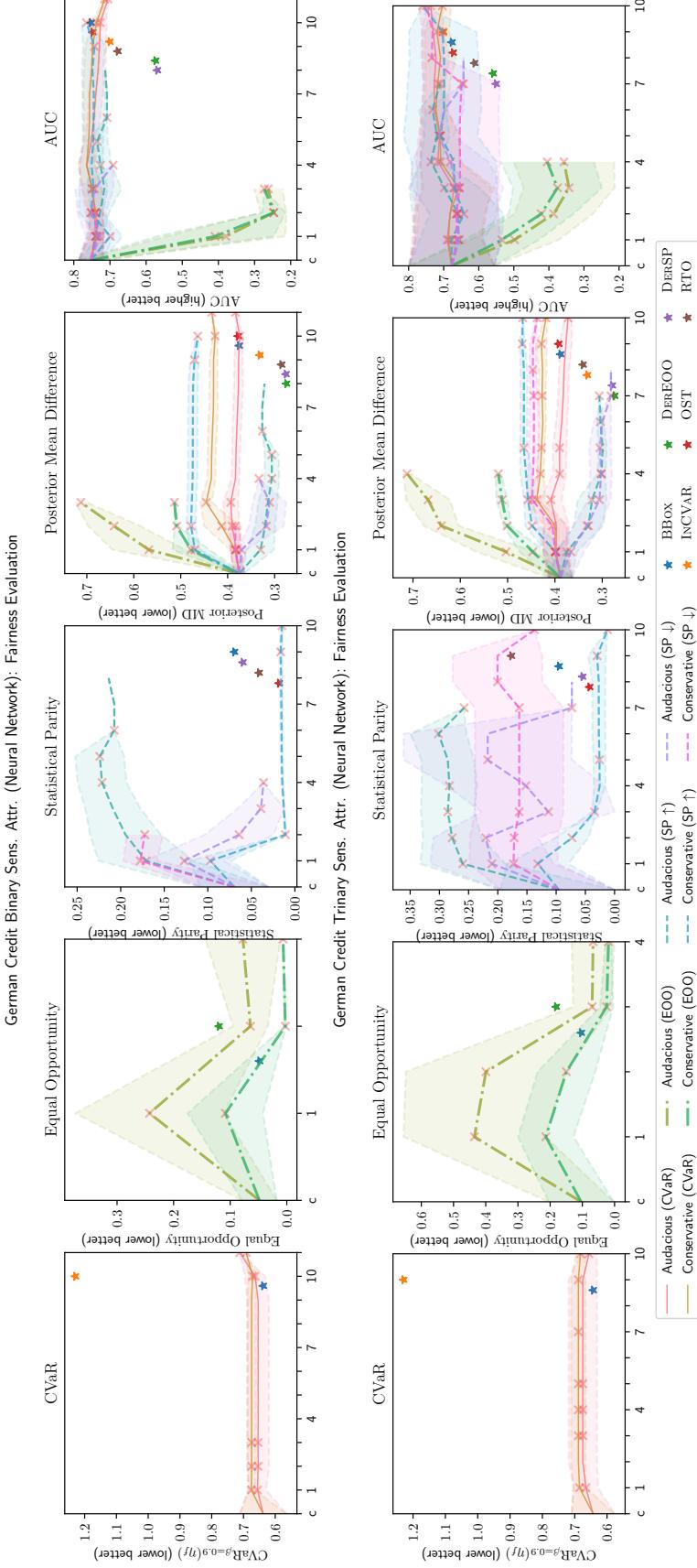


Figure 24: MLP with  $B = 3$  TopDown optimized for different fairness models evaluated on German Credit with binary (up) and trinary (down) sensitive attributes. Crosses denote when a subgroup's  $\alpha$ -tree is initiated (over any fold). The shade depicts  $\pm$  a standard deviation from the mean.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

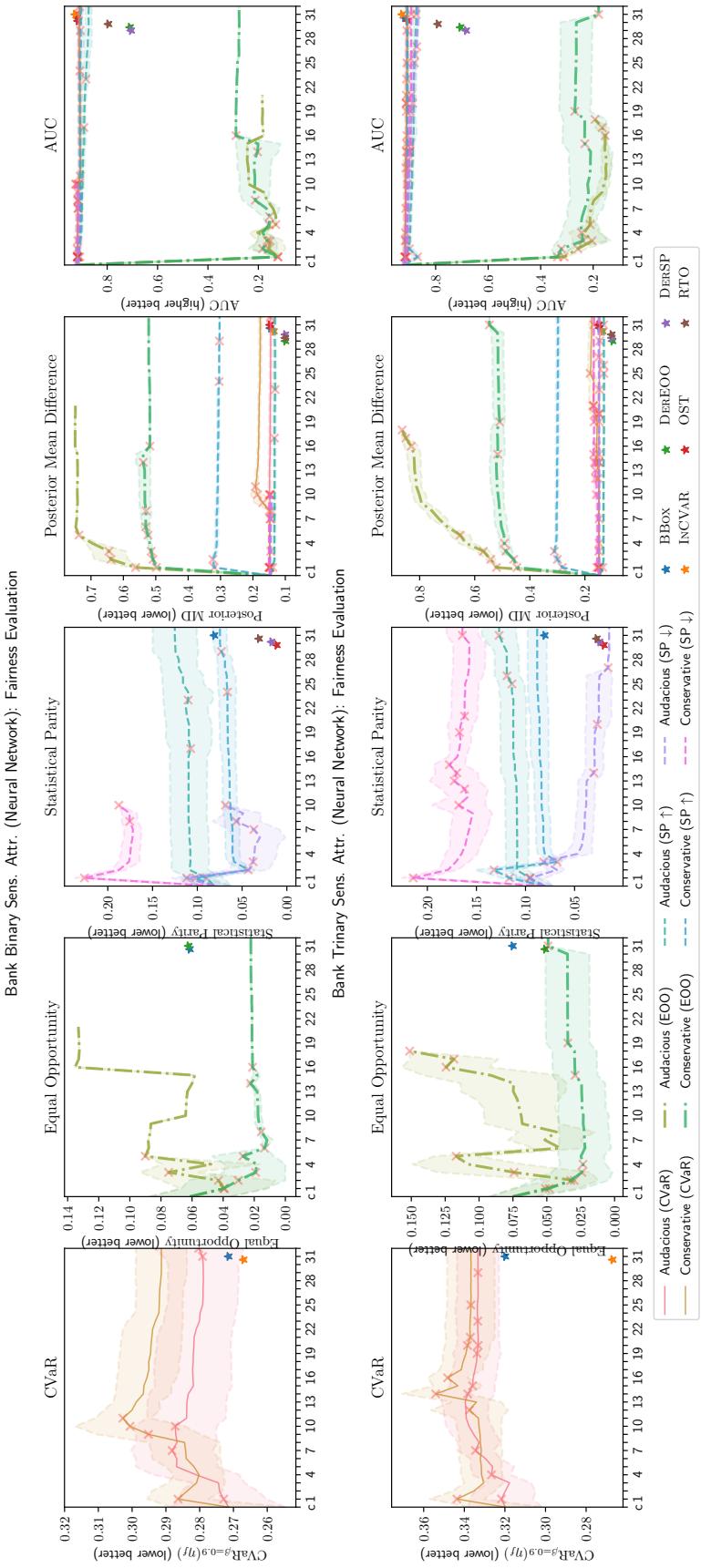


Figure 25: MLP with  $B = 3$  TOPDOWN optimized for different fairness models evaluated on Bank with binary (up) and trinary (down) sensitive attributes. Crosses denote when a subgroup's  $\alpha$ -tree is initiated (over any fold). The shade depicts  $\pm$  a standard deviation from the mean. However,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

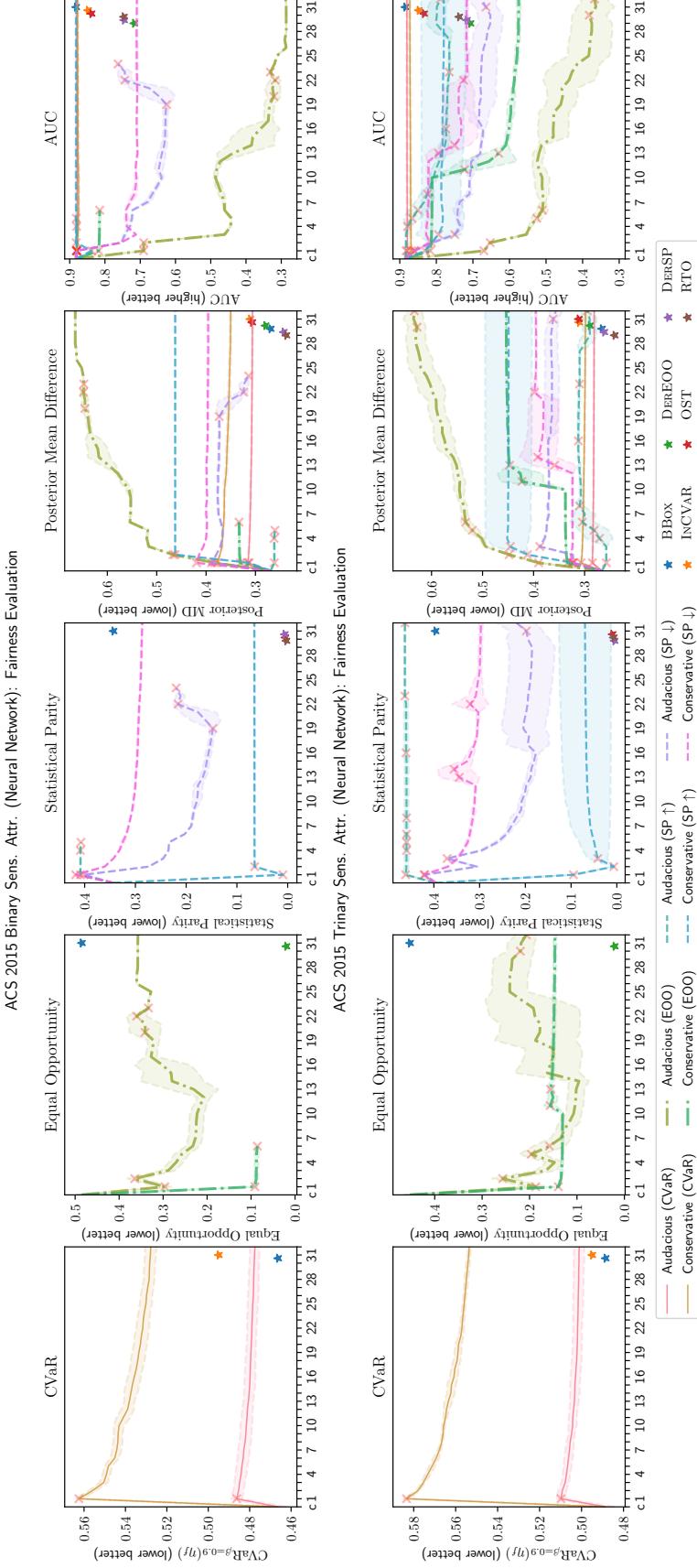


Figure 26: MLP with  $B = 3$  TOPDOWN optimized for different fairness models evaluated on Bank with binary (up) and trinary (down) sensitive attributes. Crosses denote when a subgroup’s  $\alpha$ -tree is initiated (over any fold). The shade depicts  $\pm$  a standard deviation from the mean. However,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

## XVIII Example Alpha-Tree

In this section, we provide an example of an  $\alpha$ -tree generated using TOPDOWN. In particular, we look at one example from training CVAR TOPDOWN on the Bank dataset with binary sensitive attributes. Fig. 27 presents the example  $\alpha$ -tree. The tree contains information about the attributes in which splits are made and the  $\alpha$ -correction made at leaf nodes (and their induced partition). In the example, could note that the  $\alpha$  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  $\alpha$ -tree which would aim to avoid certain combinations of attribute. We leave these factors for future work.

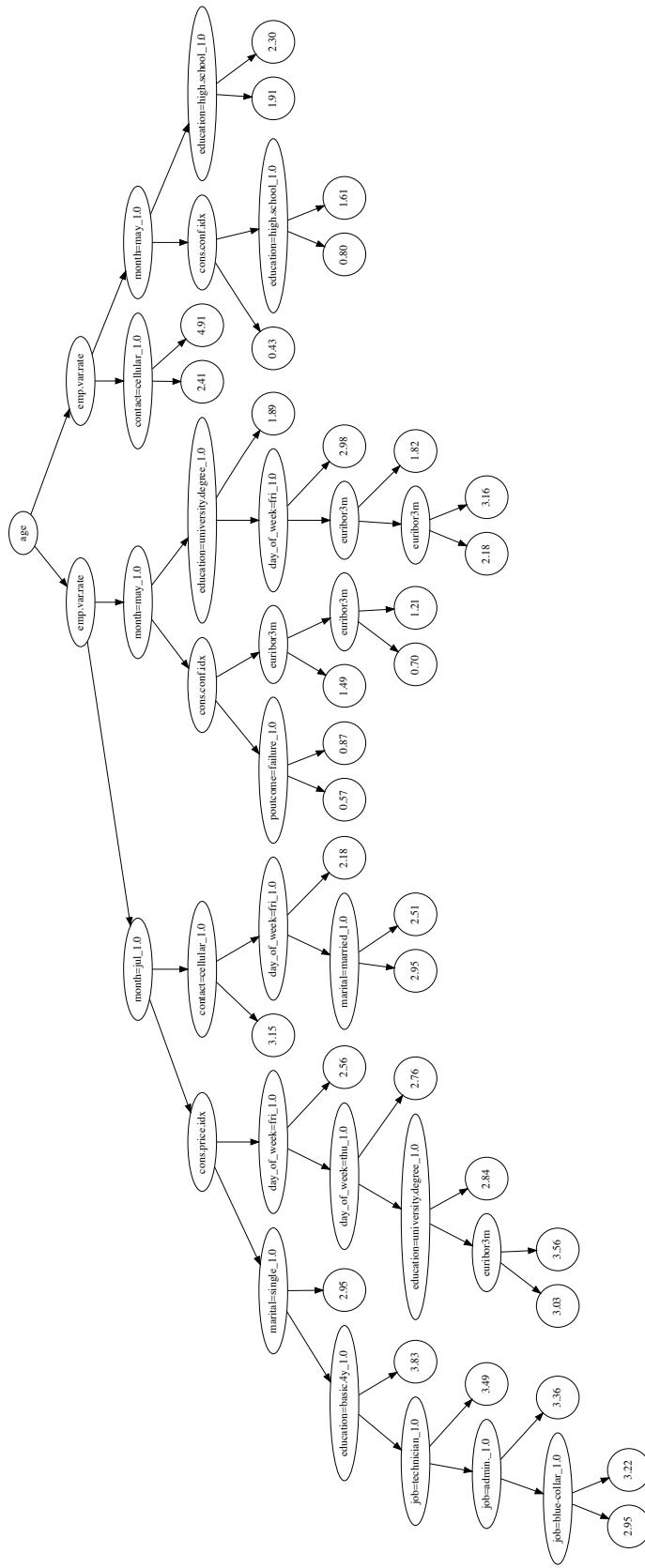


Figure 27: Example tree generated in the optimization of TopDown for CVaR in the Bank dataset with binary sensitive attributes.