Fair Densities via Boosting the Sufficient Statistics of Exponential Families

Alexander Soen [* 1 2] **Hisham Husain** [* 2] **Richard Nock** [† 3 1]

Abstract

We introduce a boosting algorithm to pre-process data for fairness. Starting from an initial fair but inaccurate distribution, our approach shifts towards better data fitting while still ensuring a minimal fairness guarantee. To do so, it learns the sufficient statistics of an exponential family with boosting-compliant convergence. Importantly, we are able to theoretically prove that the learned distribution will have a *representation rate* and *statistical rate* data fairness guarantee. Unlike recent optimization based pre-processing methods, our approach can be easily adapted for continuous domain features. Furthermore, when the weak learners are specified to be decision trees, the sufficient statistics of the learned distribution can be examined to provide clues on sources of (un)fairness. Empirical results are present to display the quality of result on real-world data.

1. Introduction

It is hard to exaggerate the importance that fairness has now taken within Machine Learning (ML) (Calmon et al., 2017; Celis et al., 2020; Williamson & Menon, 2019) (and references therein). ML being a data processing field, a common upstream source of biases leading to downstream discrimination is the data itself (Calders & Žliobaitė, 2013; Kay et al., 2015; De-Arteaga et al., 2019). Targeting data is especially important when the downstream use-case is unknown. As such, pre-processing algorithm have been introduced to debias data and mitigate unfairness in potential downstream tasks. These approaches commonly target two sources of (un)fairness (Celis et al., 2020): ① the balance of different social groups across the dataset; and ② the balance of positive outcomes between different social groups.

---
[*]Work complete prior to joining Amazon [†]Work complete prior to joining Google Research [1]Australian National University [2]Amazon [3]Google Research. Correspondence to: Alexander Soen <alexander.soen@anu.edu.au>.
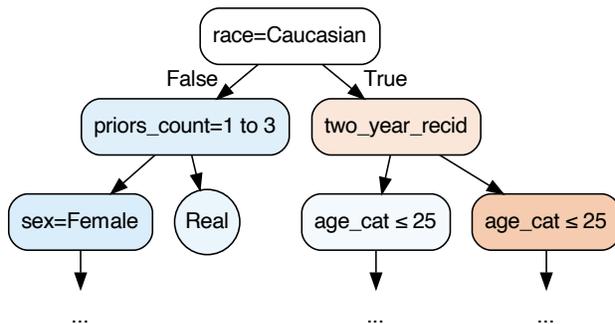
Figure 1. Our approach learns weak learners at each boosting iteration of the algorithm. The weak learner attempts to differentiate between the 'generated' current distribution (more blue) versus the 'real' data (more red) — whose predicted 'realness' are used to update the current distribution. The following shows, the first 3 levels of a decision tree weak learner produced in the first round of boosting M-E-1.0 (with the COMPAS dataset and sensitive attribute = race). The decision tree unveils a race and sex-dependent segregation. The circular node depicts a leaf node.

Recently, optimization based approaches have been proposed to directly learn debias distributions (Calmon et al., 2017; Celis et al., 2020). Such approaches share two drawbacks: tuning parameters to guarantee realizable solutions or fairness guarantees can be tedious, and the distribution learned has the same *discrete* support as the training sample. This latter drawback inherently prevents extending the distribution (support) and its guarantees (fairness) "beyond the training sample's". On the other hand, Generative Adversarial Networks (GANs) based approaches have been leveraged to learn fair distributions (Choi et al., 2017; Xu et al., 2018; Rajabi & Garibay, 2022). These can naturally be adapted to continuous spaces. Unfortunately, in discrete spaces rounding or domain transformations need to be used. Furthermore, these approaches often lack any theoretical fairness guarantees; or requires casual structures to be known (van Breugel et al., 2021). In addition, the fairness of GAN based approaches are difficult to estimate a prior to training, where fairness is achieved via a regularization penalty or separate training rounds. A shared downside of optimization and GAN approaches have is the lack of interpretability of how the fair distributions are learned.

Our Contributions We propose a novel boosting algorithm to pre-process data: Fair Boosted Density Estimation

(FBDE). Intuitively, we start with a 'fair' initial distribution that is updated to move closer to the true data with each boosting step. The learned distribution consists of an exponential family distribution with sufficient statistics given by the weak learners (WLs). Notably, the computational complexity of our approach is proportional to the number of boosting steps and, thus, is clearly readable. Additionally, our approach can naturally be used in both discrete and continuous spaces by choosing an appropriate initial distribution and WLs. In the case when we are examining a discrete domain (a common setting for social data benchmarks (Fabris et al., 2022)), interpretable WLs can be used to examine how FBDE is learning its distribution, see Fig. 1. Theoretically, we provide two styles of fairness guarantees (Theorems 3.2 and 3.3) which balance between fairness and 'aggressiveness' of boosting updates. In addition, we provide a boosting-style analysis of convergence of FBDE (Theorems 3.5 and 3.6). We also extend the theory of *mollifiers* previously used in privacy (Husain et al., 2020) to characterize sets of fair distributions we are boosting within (Lemmas 4.2 and 4.3). Finally, we empirically evaluate our approach with respect to data, prediction, and clustering fairness; and further present an empirical test of FBDE used for continuous domains.

Related Work Pre-processing provides a flexible approach to algorithmic fairness in ML. Although there other points of intervention in the fairness pipeline (*e.g.*, in-processing or post-processing fairness (Zafar et al., 2019, Section 6.2)) and entirely different notation of fairness (*e.g.*, individual fairness (Dwork et al., 2012)), we limit our work to subgroup fairness for pre-processing data. That is, instead of approaches which targets fairness at a model prediction level, we instead target the data itself. We also note that boosting methods have seen success in post-processing methods (Kim et al., 2019; Soen et al., 2022).

Other pre-processing methods not already discussed include re-labeling or re-weighting (King & Zeng, 2001; Calders et al., 2009; Kamiran et al., 2012; Kamiran & Calders, 2012). Although many of these approaches are computationally efficient, they often do not consider elements in the domain which do not appear in the dataset and lack fairness guarantees. Similarly, repair methods aim to change the input data to break the dependence on sensitive attributes for trained classifiers (Johndrow & Lum, 2019; Feldman et al., 2015). To achieve distributional repair, one can employ frameworks of optimal transport (Gordaliza et al., 2019) or counterfactual distributions (Wang et al., 2019).

Limitations of pre-processing methods in the presence of distribution shift has also been examined: under all possible distribution shifts, there does not exist a generic non-trivial fair representation (Lechner et al., 2021). However, we note that fairness guarantees breaking due distribution shift is not a problem exclusive to pre-processing algorithms (Schrouff et al., 2022). Furthermore, this limitation is pessimistic and some data shifts may not significantly harm fairness. Nevertheless, from a practical standpoint one should look out for fairness degradation from distribution shift.

2. Setting and Motivation

Let $\mathfrak{X}$ be a domain of inputs, $\mathcal{Y}$ be labels, and $\mathcal{S}$ be a set of sensitive attributes (separate from $\mathfrak{X}$). We assume that both $\mathcal{S}, \mathcal{Y}$ are finite. Unlike prior works (*i.e.*, Calmon et al. (2017); Celis et al. (2020)) we *do not* assume that $\mathfrak{X}$ is discrete. Denote $\mathcal{Z} = \mathfrak{X} \times \mathcal{Y} \times \mathcal{S}$. We short hand the tuple (x, y, s) as z, *i.e.*, for a function f we have $f(z) = f(x, y, s)$. We further let $\mathcal{D}(\mathcal{Z})$ denote the set of distributions with common support $\mathcal{Z}$. With slight abuse of notation, we distinguish between marginals, conditional, and joint distributions of $P \in \mathcal{D}(\mathcal{Z})$ by its arguments, *i.e.*, $P(x \mid y, s)$ vs $P(y, s)$. All proofs are deferred to the Appendix.

The goal of pre-processing is to correct an input distribution $P \in \mathcal{D}(\mathcal{Z})$ to an output distribution $Q \in \mathcal{D}(\mathcal{Z})$ which adheres to or improves a fairness criteria on the distribution. We focus on two common data fairness criteria (Celis et al., 2020): representation rate and statistical rate.

The first of these criteria simply measures the balance of representation of the different sensitive subgroups.

Definition 2.1. A density $P \in \mathcal{D}(\mathcal{Z})$ has ρ-**representation rate** if

$$\mathrm{RR}(P, s, s') \doteq \frac{P(s)}{P(s')} \geq \rho, \quad \forall s, s' \in \mathcal{S} \qquad (1)$$

and define the **representation rate of a distribution** P as $\mathrm{RR}(P) \doteq \min_{s,s' \in \mathcal{S}} \mathrm{RR}(P, s, s') \in [0, 1]$.

The second of these criteria measure the balance of a prediction outcome across different subgroups.

Definition 2.2. A density $P \in \mathcal{D}(\mathcal{Z})$ has τ-**statistical rate** (w.r.t. fixed label $y \in \mathcal{Y}$) if

$$\mathrm{SR}(P, s, s'; y) \doteq \frac{P(y \mid s)}{P(y \mid s')} \geq \tau, \quad \forall s, s' \in \mathcal{S} \qquad (2)$$

and define the **statistical rate of a distribution** P (w.r.t. y) as $\mathrm{SR}(P) \doteq \min_{s,s' \in \mathcal{S}} \mathrm{SR}(P, s, s'; y) \in [0, 1]$.

The selection of label $y \in \mathcal{Y}$ used to specify statistical rate typically corresponds to that corresponding to a positive outcome. As such, the balance being measured corresponds to the rate of positive outcomes across different subgroups. We take this convention and represent the advantaged group as $\mathsf{Y} = +1$. It should be noted that statistical rate can also be interpreted as a form of *discrimination control*, inspired by the "80% rule" (Calmon et al., 2017).

Intuitively for both notions of fairness, a distribution is maximally fair when the fairness parameter is equal to 1 (*i.e.*, $\tau = 1$ or $\rho = 1$). With such a rate requirement, the constraint requires probability of representation or the positive label rate to be equal for all subgroups.

We present an algorithm which provides an estimate of an input distribution P which satisfies a statistical rate τ. Although we are targeting statistical rate, the learned distribution also has strong guarantees for representation rate.

2.1. Why Data Fairness?

A reasonable question one might make is why target the data and not just make models 'fair' directly. We highlight a few examples of why one might want, or even require, fairness within the data itself.

Firstly, one might want to make data fair to allow for models trained further down in the ML pipeline to be fair. A tantalizing question data fairness faces is how it influences model fairness downstream in ML pipelines. It has previously been shown that downstream classifier compound representation injustices within the data (De-Arteaga et al., 2019, Theorem 1). Furthermore, it has been previously shown experimentally that pre-processing approaches can improve prediction fairness for various metrics (Calmon et al., 2017; Celis et al., 2020). As such, providing 'fair' representations can be a critical components of a ML pipeline.

Also, providing data fairness guarantees can also potentially provide improvement for non-classification notions of fairness, *i.e.*, clustering, see experiments in Section 5.

Lastly, we would also like to highlight that finding fair distributions has recently become of independent interest to certify the fairness of models (Kang et al., 2022). The goal is to provide a certificate of a models performance under distribution shift restricted to distribution which are statistical rate fair. As a side-effect, the evaluated model can be proven to be 'fair' (Kang et al., 2022, Proposition 1).

3. Fair Boosted Density Estimator

We propose *Fair Boosted Density Estimation* (FBDE), a boosting algorithm which iteratively degrades a fair but 'inaccurate' initial distribution Q_0 to become closer to an input data distribution, we denote as P. In particular, the user specifies a statistical rate target / budget $\tau \in (0, 1]$ which controls the fairness degradation size of each iterative boosting update. Pseudo-code is given in Algorithm 1.

3.1. Distribution Estimator

The learned fair distribution Q_T is an exponential family distribution constructed by iteratively aggregating classifiers. The distribution consists of two main components: ① the

initially fair distribution; and ② the boosting updates.

For the initial distribution we require it to be more statistical rate fair than the target τ. We also require the initial distribution to be representation rate fair to get a guarantee for representation rate. As such, we define an initial distribution Q_{INIT} hierarchically. First, we specify the label-sensitive marginal given an initial budget $\tau < \text{SR}_0 \leq 1$:

$$Q_{\text{INIT}}(s) \doteq 1/|\mathcal{S}|; \tag{3}$$
$$Q_{\text{INIT}}(Y = 1|s) \doteq \max\{P(Y = 1|s), \text{SR}_0 \cdot p_{\max}\}, \tag{4}$$

where $p_{\max} = \max_{s \in \mathcal{S}} P(Y = 1|s)$[1]. Finally, we specify the conditional $Q_{\text{INIT}}(x|y, s)$ for each $(y, s) \in \mathcal{Y} \times \mathcal{S}$ by fitting an empirical distribution if X is discrete, or a normal distribution if X is continuous. One can verify that $\text{SR}(Q_{\text{INIT}}) \geq \text{SR}_0$ and $\text{RR}(Q_{\text{INIT}}) = 1$. Alternatively, Eq. (3) can also be altered to satisfy a weaker representation rate budget RR_0. As suggested by the notation, we denote SR_0 and RR_0 as the statistical rate and representation of the initial distribution, respectively.

One additional consideration we should make is in the circumstance when finding an adequate Q_{INIT} is difficult. For instance, in low data regimes, approximating $Q_{\text{INIT}}(x|y, s)$ for each (y, s) could be difficult (especially for under represented demographic subgroups). This can be particularly catastrophic when we are estimating the conditional distribution in a way such that by having no examples of inputs occurring in the training data, the estimated distribution has zero support for those inputs, *i.e.*, an empirical distribution. A remedy for such as situation comes from taking a mixture distribution of our proposed initial distribution (as per Eqs. (3) and (4)) and a 'prior' distribution. Celis et al. (2020, Section 3.1 "Prior distributions") discusses a similar approach for interpolating between distributions. Appendix K presents additional discussion and an experimental example using real world datasets.

In the boosting step of our algorithm, binary classifiers $c_t : \mathcal{Z} \rightarrow \mathbb{R}$ are used to make an exponential reweighting of the previous iteration's distribution Q_{t-1}. In particular, the classifiers c_t acts as a *discriminator* to distinguish between real P and fake samples Q_{t-1} (via the $\text{sign}(c_t(z))$). We assume that more positive outputs of c_t indicate greater 'realness'. We take the common technical assumption that each c_t have bounded output: $c_t(z) \in [-C, C]$ for some $C > 0$ (Schapire & Singer, 1999).

We assume we have a weak learner $\text{WL}(P, Q_{t-1})$ which is used to produce such discriminators c_t at each step of the algorithm. Thus, after t iterations the learned distribution

---

[1]The initialization specified increases fairness by increasing positive outcomes. Depending on the application, one may want to take the opposite strategy: replacing Eq. (4) by $\min\{\text{SR}_0 \cdot P(Y = 1|s), p_{\min}\}$ where $p_{\min} = \min_{s \in \mathcal{S}} P(Y = 1|s)$.

Algorithm 1 FBDE (WL, $T, \tau, \mathrm{Q}_{\mathrm{INIT}}, \vartheta_t$)

1: **input**: Weak learner WL, # iter. T, SR τ,
    init. $\mathrm{Q}_{\mathrm{INIT}}$, input dist. P, leverage ϑ_t;
2: $\mathrm{Q}_0 \leftarrow \mathrm{Q}_{\mathrm{INIT}}$ (with $\mathrm{SR}_0 > \tau$)
3: **for** $t = 1, \ldots, T$ **do**
4:    $c_t \leftarrow \mathrm{WL}(\mathrm{P}, \mathrm{Q}_{t-1})$
5:    $\mathrm{Q}_t \propto \mathrm{Q}_{t-1} \cdot \exp(\vartheta_t c_t)$
6: **end for**
7: **return**: Q_T (for fairness $\vartheta_t \in \{\vartheta_t^{\mathrm{E}}, \vartheta_t^{\mathrm{R}}\}$)

Table 1. Summary of of different leveraging schemes ϑ_t in FBDE.

FAIR	ϑ_t	$\mathrm{SR}(\mathrm{Q}_t)$	SIZE ε_t ($\mathrm{SR}_0 = 1$)
EXACT	$O(2^{-t})$	τ	$-\log\tau$
REL.	$O(t^{-1})$	$\tau^{\Omega(\log t)}$	$-(1 + \log t)\log\tau$

have the following exponential family functional form:

$$\mathrm{Q}_t(x, y, s) = \frac{1}{Z_t} \exp(\vartheta_t c_t(x, y, s))\mathrm{Q}_{t-1}(x, y, s), \quad (5)$$

where $\vartheta_t \geq 0$ are *leveraging coefficients* and Z_t is the normalizer of $\mathrm{Q}_t(x, y, s)$. The sufficient statistics of this exponential family distribution is exactly the learned $c_1 \ldots c_t$. Intuitively, the less (more) 'real' a c_t deems an input to be, the lower (higher) the reweighting $\exp(\vartheta_t c_t(\cdot))$ makes.

3.2. Fairness Guarantees

To establish fairness guarantees for statistical rate fairness, we require multiplicative parity of positive subgroup rates $\mathrm{Q}_t(\mathrm{Y} = 1|s)$. To calculate these rates, we define the following normalization terms:

$$Z_t(y, s) = \int_{\mathcal{X}} \exp(\vartheta_t c_t(z)) \, \mathrm{dQ}_{t-1}(x|y, s);$$

$$Z_t(s) = \int_{\mathcal{X} \times \mathcal{Y}} \exp(\vartheta_t c_t(z)) \, \mathrm{dQ}_{t-1}(x, y|s).$$

We verify that $\mathrm{Q}_t(y, s) = \mathrm{Q}_{t-1}(y, s) \cdot (Z_t(y, s)/Z_t)$ and $\mathrm{Q}_t(s) = \mathrm{Q}_{t-1}(s) \cdot (Z_t(s)/Z_t)$.

This recursive definition of these marginal distributions provides a convenient lower bound of the statistical rate.

Lemma 3.1. *Suppose that Q_0 has statistical rate SR_0. Then for all $s, s' \in \mathcal{S}$ we have*

$$\mathrm{SR}(\mathrm{Q}_t, s, s'; y) \geq \mathrm{SR}_0 \cdot \prod_{i=1}^{t} \left(\frac{Z_i(s')}{Z_i(s)} \cdot \frac{Z_i(y, s)}{Z_i(y, s')} \right). \quad (6)$$

It should be noted that for each of the normalizers there is a hidden dependence on Q_0 and ϑ_i's. As such, the pairwise statistical rate of the boosted distributions are determined by two factors: ① the initial SR_0; and ② the leveraging coefficients ϑ_i. We note that similar argumentation can be made for representation rate fairness. In our approach, the coefficients ϑ_i are taken to be a function of the iteration number i, initial fairness SR_0, and fairness budget τ.

Thus given the initial distribution specified above, we propose two leveraging schemes ϑ_t which can be used to accommodate different fairness guarantees.

Exact fairness guarantees that fairness holds irrespective of other parameters of Q_T (*i.e.* boosting steps T). This type of fairness guarantee can be established by setting the leveraging coefficient as $\vartheta_t^{\mathrm{E}} := -(C2^{t+1})^{-1}\log(\tau/\mathrm{SR}_0)$.

Theorem 3.2. *Suppose that $\vartheta_t = \vartheta_t^{\mathrm{E}}$, then $\mathrm{SR}(\mathrm{Q}_T) > \tau$ and $\mathrm{RR}(\mathrm{Q}_T) > \mathrm{RR}_0 \sqrt{\tau/\mathrm{SR}_0}$ for $T \geq 1$.*

This setting is not just appealing for its absolute fairness it provides, but also the exponentially decreasing leverage ϑ_t^{E} — which fits the setting where only a few classifiers are required to provide a good estimate of P (or are enough to break the WLA, Definition 3.4). For representation rate, the $\sqrt{\tau/\mathrm{SR}_0}$ term only degrades the initial representation rate RR_0 slightly when the statistical rate budget τ is high (which is assumed as we want to achieve a fair Q_T). For instance, when $\tau = 0.8$ and we use the initial distribution proposed ($\mathrm{RR}_0 = 1$), then $\mathrm{RR}(\mathrm{Q}_T) > \sqrt{\tau} \approx 0.894$.

Given that the exact fairness guarantees hold regardless of T, one could theoretically keep adding classifiers c_t forever. In practice this never happens, and thus we explore a notion of 'relative' fairness. Instead of exact fairness, the guarantee on fairness gradually becomes weaker 'relative' to an initial fairness constraint over update iterations.

Relative fairness proves a fairness guarantee which degrades gracefully with the number of boosting iterations. To do so, we define $\vartheta_t^{\mathrm{R}} := -(4Ct)^{-1}\log(\tau/\mathrm{SR}_0)$.

Theorem 3.3. *Suppose that $\vartheta_t = \vartheta_t^{\mathrm{R}}$, then $\mathrm{SR}(\mathrm{Q}_T) > \tau^{1+\log T}$ and $\mathrm{RR}(\mathrm{Q}_T) > \mathrm{RR}_0(\sqrt{\tau/\mathrm{SR}_0})^{1+\log T}$ for $T \geq 1$.*

Notice the key boosting difference with Theorem 3.2: the sum of the series of leveraging coefficients diverges, so relative fairness accommodates for more aggressive boosting schemes. Table 1 summarizes the implications of the two theorems. It is not surprising that both leveraging schemes display $\vartheta_t \to 0$ as $\tau/\mathrm{SR}_0 \to 1$, as maximal fairness forces the distribution to stick to Q_0 and is therefore data oblivious. Differences are apparent when we consider the number of boosting iterations T: should we boost for $T = 5$, we still get $\mathrm{SR}(\mathrm{Q}_T) > \tau^{2.3}$ and $\mathrm{RR}(\mathrm{Q}_T) > \tau^{1.15}$ with relative fairness (taking $\mathrm{SR}_0 = \mathrm{RR}_0 = 1$), which can still be reasonable depending on the problem. In real-world data scenarios (see Section 5), we find that large numbers of boosting iterations can be taken without major degradation of fairness.

3.3. Sampling for Q_T

A desirable property for the learned Q_T would be the ability to sample efficiently from it. In practice, to utilize the weak learner WL to create classifiers c_t, we require samples from Q_{t-1}. In the case when X consists of a discrete finite domain, one can simply enumerate the possible inputs and sample from a large multinomial distribution. However, when X consists of continuous random variables, we cannot use the same strategy. Instead, we utilize Langevin Monte Carlo (LMC) and a sampling trick to efficiently calculate samples. It should be noted that LMC can be directly applied to sampling Q_t if Y, S are approximated to be continuous, *i.e.* via Grathwohl et al. (2021) or Choi et al. (2017); Xu et al. (2018); Rajabi & Garibay (2022).

We first note that the conditional distribution $Q_t(x|y,s)$ is of a similar functional form given by Eq. (5). As such, a natural way to sample from these conditionals is via LMC sampling algorithms using the distribution's *score function*:

$$\nabla \log Q_t(x|y,s) = \sum_{i=1}^{t} \vartheta_i \nabla c_i(z) + \nabla \log Q_0(x|y,s), \quad (7)$$

where $z = (x,y,s)$, as noted in Section 2. Of course, we require the classifiers c_t and log-likelihood of the initial distribution Q_0 to be differentiable. The former can be achieved by taking c_t's to be simple neural networks. The latter can be achieved using our proposed initial distribution (taking $Q_0(x|y,s)$ to be normal distributions).

Secondly, the marginal distribution $Q_t(y,s)$ can be calculated by *sampling from* Q_0:

$$Q_t(y,s) \propto Q_0(y,s) \underset{x|y,s \sim Q_0}{\mathbb{E}} \left[\exp\left(\sum_{k=1}^{t} \vartheta_k c_k(z) \right) \right], \quad (8)$$

where the expectation can be approximated by sampling from $Q_0(x|y,s)$.

With these ingredients, we can now state our sampling routine. For any t, we first calculate the current marginal $Q_t(y,s)$ via Eq. (8). We can now hierarchically sample the joint distribution by first sampling from $Q_t(y,s)$, and then sampling from $Q_t(x|y,s)$ using LMC.

FBDE as Boosting the Score An interesting perspective of FBDE is to examine the overdamped Langevin diffusion process which can be used to sample Q_t. Assuming the necessary assumptions of continuity and differentiability (including Y and S to simplify the narrative), the (joint) diffusion process of Q_t, as per (7), is given by

$$\dot{Z}_t = \sum_{i=1}^{t} \vartheta_i \nabla c_i(Z) + \nabla \log Q_0(Z) + \sqrt{2}\dot{W}, \quad (9)$$

where W is standard Brownian motion and $Z = (X, Y, S)$.

In the limit (w.r.t. the time derivative), the distribution of Z approaches Q_t (Roberts & Tweedie, 1996).

Removing the initial summation, Eq. (9) simplifies to the Langevin diffusion process for the initial diffusion process — the score function of Q_0 'pushes' the process to areas of high likelihood. Reintroducing the first summation term, as the goal of $c_i(.)$'s is to predict the 'realness' of samples, the term defines a vector fields which 'points' towards the most 'real' direction in the distribution. Thus, the WLs $c_i(.)$ can be interpreted as correction terms which 'pushes' the diffusion process closer to P.

3.4. Convergence

To discuss properties of convergence FBDE has, we first introduce a variant of *weak learning assumption* of boosting (Husain et al., 2020; Cranko & Nock, 2019).

Definition 3.4 (WLA). A learner $\mathrm{WL}(\cdot, \cdot)$ satisfies the **weak learning assumption** (WLA) for $\gamma_P, \gamma_Q \in (0,1]$ iff for all $P, Q \in \mathcal{D}(\mathcal{Z})$, $\mathrm{WL}(P,Q)$ produces a *discriminator* $c : \mathcal{Z} \to \mathbb{R}$ satisfying $\mathbb{E}_P[c] \geq C \cdot \gamma_P$ and $\mathbb{E}_Q[-c] \geq C \cdot \gamma_Q$.

Intuitively, the WLA constants γ_P, γ_Q measures a degree of separability between the two distributions P and Q — with $\gamma_P = \gamma_Q = 1$ giving maximal separability. As such, in our algorithm, as $Q_t \to P$ finding higher boosting constants γ^t becomes harder. Although our WLA may appear dissimilar to the typical WLAs which rely on a single inequality (Schapire & Singer, 1999), it has been shown that Definition 3.4 is equivalent (Cranko & Nock, 2019, Appendix Lemma 7). When applying the WLA to FBDE, we refer to constants γ_P^t, γ_Q^t when referring to the WL learned via $\mathrm{WL}(P, Q_{t-1})$, *i.e.*, learning the tth classifier c_t.

Using the WLA, we examine the progress we can make per boosting step. In particular, we examine the drop in KL between input P and successive boosted densities Q_{t-1}, Q_t.

Theorem 3.5. *Let ϑ_t be any leverage such that $\vartheta_t \leq 1$ for all t, $\tau/\mathrm{SR}_0 > \exp(-4C)$, and $C > 0$. If WL satisfies WLA (for P, Q_{t-1}) with γ_P^t, γ_Q^t, then:*

$$\mathrm{KL}(P, Q_{t-1}) - \mathrm{KL}(P, Q_t) \geq \vartheta_t \cdot \Lambda_t, \quad (10)$$

where $\Lambda_t = \gamma_P^t \cdot C + \Gamma(\mathbb{V}_{Q_{t-1}}[c_t], \gamma_Q^t)$; $\mathbb{V}_{Q_t}[c_t]$ denotes the variance of c_t; and $\Gamma(.,.)$ is decreasing w.r.t. the first argument and increasing w.r.t. the second argument.

The full definition of $\Gamma(.,.)$ can be found in the Appendix. Intuitively, more accurate WLs (which gives higher γ's) leads to a higher KL drop. The above is a strict improvement upon Husain et al. (2020, Theorem 5) and, therefore, extends the theoretical guarantees for boosted density estimation (not just FBDE) at large. In particular, we find that a constraint on the variance of the classifier allows us to remove the two boosting regimes analysis, prevalent in

Husain et al. (2020) — the conditions for > 0 KL drop only depends on the variance and WLA constants now.

Specifically, Theorem 3.5 allows for smaller values of γ_{Q}^t to result in a > 0 KL drop than Husain et al. (2020, Theorem 5) (which requires $\gamma_{\mathrm{Q}}^t > 1/3$ for $C = \log 2$). In particular, taking $C = \log 2$ then having $\mathbb{V}_{\mathrm{Q}_{t-1}}[c_t] < C^2 \cdot 2/3 \approx 0.32$ allows us to have a > 0 KL drop with smaller γ_{Q}^t constants (with an extended discussion in Appendix I). Notably, given the bounded nature of the classifier, the variance is already bounded with $\mathbb{V}_{\mathrm{Q}_{t-1}}[c_t] \leq C^2 \approx 0.48$ for $C = \log 2$ — thus the condition itself is reasonable. In addition, we should note that convergence, unlike in the privacy case (Husain et al., 2020), depends on how the fairness τ parameter interacts with the update's leveraging coefficient ϑ_t.

In addition to consider the per iteration KL drop, we consider "how far from Q_0" we progressively get, in an information-theoretic sense. For this, we define $\Delta(\mathrm{Q}) \doteq \mathrm{KL}(\mathrm{P}, \mathrm{Q}_0) - \mathrm{KL}(\mathrm{P}, \mathrm{Q})$. To simplify the analysis, we assume that constants $\gamma_{\mathrm{P}}, \gamma_{\mathrm{Q}}$ are fixed throughout the boosting process (or simply taking worse-case constants).

Theorem 3.6. *Suppose that $\lambda = -\log(\tau/\mathrm{SR}_0)$, $\alpha(\gamma) = \min_t \Gamma(\mathbb{V}_{\mathrm{Q}_{t-1}}[c_t], \gamma)/(\gamma C)$, and $C, T > 0$.*
• *If $\vartheta_t := \vartheta_t^{\mathrm{E}}$, then*

$$\Delta(\mathrm{Q}_T) \in \lambda \left(1 - 2^{-T}\right) \left[\frac{\gamma_{\mathrm{P}} + \gamma_{\mathrm{Q}} \cdot \alpha(\gamma_{\mathrm{Q}})}{2}, 1\right];$$

• *If $\vartheta_t := \vartheta_t^{\mathrm{R}}$, then*

$$\Delta(\mathrm{Q}_T) \in \frac{\lambda}{2} \left[\frac{\gamma_{\mathrm{P}} + \gamma_{\mathrm{Q}} \cdot \alpha(\gamma_{\mathrm{Q}})}{2} \left(\frac{1}{T} + \log T\right), (1 + \log T)\right],$$

where $a[b, c] = [ab, ac]$.

In addition to the worse-case convergence rates (lower bounds) typically analyzed in boosting algorithms, Theorem 3.6 also characterizes the best-case scenario (upper bounds) achievable by our algorithm. Notably, the gap, computed as the ratio best-to-worst, solely depends on the parameters of the WLA, and thus is guaranteed to be reduced as the WL becomes better. As $T \to \infty$, then with the exact leverage ϑ_t^{E} we have that $\Delta(\mathrm{Q}_T) = \Theta(-\log(\tau/\mathrm{SR}_0))$. Additionally assuming $\gamma_{\mathrm{P}}, \gamma_{\mathrm{Q}} \to 1$ and setting $\mathbb{V}_{\mathrm{Q}_{t-1}}[c_t] \to 0$ uniformly, the upper and lower bound exactly matches at $-\log(\tau/\mathrm{SR}_0)$ — our bound is tight. Despite this, with the relative leverage ϑ_t^{R}, there will always be a $\Omega(-\log(\tau/\mathrm{SR}_0))$ gap between the upper and lower bound of $\Delta(\mathrm{Q}_T)$. Furthermore, in contrast to the geometric convergence of the bounds in the exact fairness case, in relative fairness the bounds grow logarithmically with T. However, we do note that for relative leverage we can achieve a tight bound if we do not attempt to simplify the series $\sum_k^T \vartheta_k$, see Theorem E.3.

On particularly nice interpretation of Theorem 3.6 is to utilize the lower bounds of $\Delta(\mathrm{Q}_T)$ to determine sufficient conditions for small KL. In particular, we find the sufficient number of boosting steps T to make $\mathrm{KL}(\mathrm{P}, \mathrm{Q}_T)$ small.

Corollary 3.7. *Suppose that the same condition in Theorem 3.6 hold and $\varepsilon < \mathrm{KL}(\mathrm{P}, \mathrm{Q}_0)$. Then $\mathrm{KL}(\mathrm{P}, \mathrm{Q}_T) < \varepsilon$, for:*
• $\vartheta_t := \vartheta_t^{\mathrm{E}}$ *if*

$$T > \frac{1}{\log 2} \cdot \log \left(\left(1 - 2 \cdot \frac{\mathrm{KL}(\mathrm{P}, \mathrm{Q}_0) - \varepsilon}{\lambda \cdot (\gamma_{\mathrm{P}} + \gamma_{\mathrm{Q}} \cdot \alpha(\gamma_{\mathrm{Q}}))}\right)^{-1}\right).$$

• $\vartheta_t := \vartheta_t^{\mathrm{R}}$ *if*

$$T > \exp \left(4 \cdot \frac{\mathrm{KL}(\mathrm{P}, \mathrm{Q}_0) - \varepsilon}{\lambda \cdot (\gamma_{\mathrm{P}} + \gamma_{\mathrm{Q}} \cdot \alpha(\gamma_{\mathrm{Q}}))}\right).$$

One should note, that when comparing the exact leverage versus the relative versus in Corollary 3.7, we are comparing the growth of two functions: $x \mapsto \log((1 - 2x)^{-1})$ for exact leverage; and $x \mapsto \exp(4x)$ for relative leverage. The former dominates asymptotic dominates the latter. That is, using the exact leverage will require more boosting updates to achieve same drop in KL; which follows our intuition that relative leveraging has better boosting convergence at the cost of decaying fairness.

4. Mollifier Interpretation

Previously it has been shown that boosted density estimation for privacy can be interpreted as boosting within a set of a set of private densities — dubbed a mollifier (Husain et al., 2020). To study the set of distributions FBDE boosts within, we adapt the *relative mollifier* construction for fairness.

Definition 4.1. Let Q_0 be a reference distribution. Then the ε-fair **relative mollifier** $\mathcal{M}_{\varepsilon, \mathrm{Q}_0}$ is the set of distributions Q satisfying $\forall s, s' \in \mathcal{S}$

$$\max \left\{\frac{\mathrm{SR}(\mathrm{Q}, s, s'; y)}{\mathrm{SR}(\mathrm{Q}_0, s, s'; y)}, \frac{\mathrm{SR}(\mathrm{Q}_0, s, s'; y)}{\mathrm{SR}(\mathrm{Q}, s, s'; y)}\right\} \leq e^{\varepsilon}. \quad (11)$$

The reference distribution Q_0 can be interpreted as the initial distribution chosen in FBDE. Notably, one can change the constraint in Eq. (11) to accommodate for different notions of data fairness, *i.e.*, replacing "SR" to "RR". The following Lemmas 4.2 and 4.3 holds identically for representation rate style mollifiers — with the first lemma providing a fairness guarantees for *all* elements in $\mathcal{M}_{\varepsilon, \mathrm{Q}_0}$.

Lemma 4.2. *If $\mathrm{Q} \in \mathcal{M}_{\varepsilon, \mathrm{Q}_0}$, then $\mathrm{SR}(\mathrm{Q}) \geq \mathrm{SR}_0 \cdot \exp(-\varepsilon)$.*

Lemma 4.2 shows a significant difference from the privacy case: the distribution within a carefully constructed mollifier are fair; not just the sampler (Husain et al., 2020). This difference is significant in fairness as knowledge of selected elements in the mollifier can elucidate sources of (un)fairness
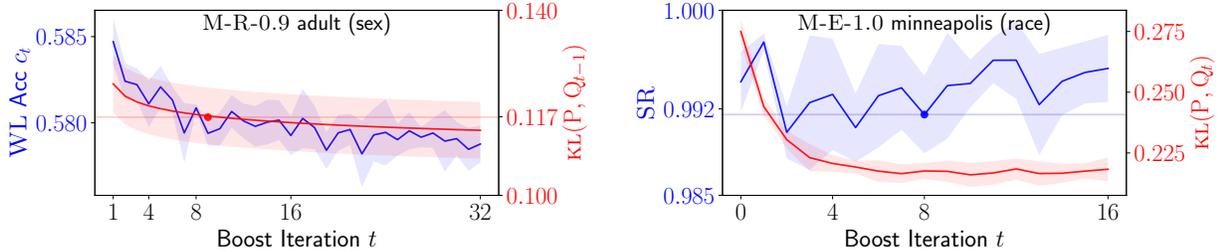
Figure 2. Evaluations of FBDE over boosting iterations. Left: WL accuracy vs KL of M-R-0.9 over boosting iterations on the ADULT dataset. Right: SR vs KL of M-E-1.0 over boosting iterations on the MINNEAPOLIS dataset — original data's SR is 0.684 ± 0.005. The shaded region depicts the 1 std. range. The horizontal line / point indicates the $T = 8$ (left) $\mathrm{KL}(\mathrm{P}, \mathrm{Q}_8)$ value and (right) $\mathrm{SR}(\mathrm{Q}_8)$ value.

learned by a model. Furthermore, this detail is crucial in applications where we need a set of fair distributions (Kang et al., 2022). Given Lemma 4.2, a natural question to ask is, *what kinds of fair distributions are included in the relative mollifier?* The following lemma answers this question.

Lemma 4.3. *Suppose that $\mathcal{M}_{\varepsilon, \mathrm{Q}_0}$ is a relative mollifier and $\mathrm{Q} \in \mathcal{D}(\mathcal{Z})$ has $\mathrm{SR}(\mathrm{Q}) \geq \exp(-\varepsilon)$. If $\forall s, s' \in \mathcal{S}$ we have either $\mathrm{SR}(\mathrm{Q}, s, s'; y) > 1$ or $\mathrm{SR}(\mathrm{Q}, s, s'; y) \leq \mathrm{SR}(\mathrm{Q}_0, s, s'; y) \leq 1$, then $\mathrm{Q} \in \mathcal{M}_{\varepsilon, \mathrm{Q}_0}$.*

Intuitively, the condition of membership in Lemma 4.3 can be interpreted as Q and Q_0 having a shared 'type' of (un)fairness: letting $\mathcal{S} = \{\text{male}, \text{female}\}$, if the probability of positive outcomes in both Q, Q_0 is higher given $\mathsf{S} = $ male than $\mathsf{S} = $ female but Q_0 is more fair, then $\mathrm{Q} \in \mathcal{M}_{\varepsilon, \mathrm{Q}_0}$. A significant instantiation of Lemma 4.3 is when $\mathrm{SR}_0 = 1$, which gives us that $\mathrm{SR}(\mathrm{Q}) \geq \exp(-\varepsilon) \implies \mathrm{Q} \in \mathcal{M}_{\varepsilon, \mathrm{Q}_0}$. In other-words, we have a *complete* mollifier: by combining Lemma 4.2, $\mathrm{Q} \in \mathcal{M}_{\varepsilon, \mathrm{Q}_0} \iff \mathrm{SR}(\mathrm{Q}) \geq \exp(\varepsilon)$ which notably including all perfectly fair distributions.

We can now express FBDE as the process of mollification $\operatorname{argmin}_{\mathrm{Q} \in \mathcal{M}} \mathrm{KL}(\mathrm{P}, \mathrm{Q})$ (Husain et al., 2020), *i.e.*, finding the closest element of a mollifier $\mathcal{M}_{\varepsilon, \mathrm{Q}_0}$ w.r.t. to an input P. In particular, by taking $\varepsilon = -\log(\tau / \mathrm{SR}_0)$ the statistical rate conditions in Lemmas 4.2 and 4.3 correspond exactly to the fairness budget τ of FBDE. Taking $\mathrm{SR}_0 = 1$, the minimization of KL (as per Theorem 3.5) can be interpreted as an iterative boosting procedure for mollification. Even in the case when $\mathcal{M}_{\varepsilon, \mathrm{Q}_0}$ is *incomplete* ($\mathrm{SR} \neq 1$), it is still useful as careful selection of Q_0 can practically yield all distributions we are interested in, *i.e.*, those close to P. Furthermore, it is fine to further focus on a subset of a mollifier. Especially so in our case, where we only consider a 'boost'-able exponential family subset of $\mathcal{M}_{\varepsilon, \mathrm{Q}_0}$; we note the strong approximation capabilities of exponential families (Nielsen & Garcia, 2009, Section 1.5.4).

One interesting perspective of FBDE and its different fairness settings is to examine the mollifiers they are boosting within. In particular, we can consider the ε_t parameter in Definition 4.1, or the '*size*', of the relative mollifiers

w.r.t. iteration value t — summarized in Table 1. Assuming $\mathrm{SR}_0 = 1$, using the exact leverage ϑ_t^{E} implies that we are boosting within a constant sized mollifier as per Theorem 3.2, with $\varepsilon_t = -\log(\tau / \mathrm{SR}_0)$. For the relative leverage ϑ_t^{R}, as the fairness degrades in this case, the corresponding mollifier grows, for which $\varepsilon_t = -(1 + \log t) \log(\tau / \mathrm{SR}_0)$. Notice that size of the mollifier also coincides (up to constant multiplication) with the upper bounds of $\Delta(\mathrm{Q}_T)$ as per Theorem 3.6.

5. Experiments

In this section, we (1) verify that FBDE debiased densities and ad-hears to specified data fairness measures; (2) inspect the implications of utilizing samples produced by a FBDE debiased density in downstream tasks, specifically, prediction and clustering tasks; (3) explore the interpretability of FBDE when utilizing decision tree (DT) weak learners (WLs); and (4) present an experimental on a dataset with continuous $\mathcal{X}$.

To analyze these points, we evaluate FBDE over preprocessed COMPAS (binary $\mathcal{S} = $ race) and ADULT (binary $\mathcal{S} = $ sex) datasets provided by AIF360[2] (Bellamy et al., 2019). This consists of a discrete binary domain.

We consider 4 configurations of FBDE, boosted for $T = 32$ iterations. We consider a fixed fairness budget $\tau = 0.8$ throughout and take a combination of exact vs relative leverage; and $\mathrm{SR}_0 = 1$ vs $\mathrm{SR}_0 = 0.9$. We designate each configuration by M-X-Y, where X encodes the leverage and Y encodes the base rate, *i.e.*, M-E-1.0 uses exact leverage with $\mathrm{SR}_0 = 1$. DT WLs are calibrated using Platt's method (Platt et al., 1999). For baselines, we consider two different data pre-processing approaches. Firstly, we consider the max entropy approach proposed by Celis et al. (2020) (MAXENT) with default parameters. Secondly, we compare against TabFairGAN proposed by Rajabi & Garibay (2022) (TABFAIR), which includes a separate training phase for fairness. In clustering, we consider a Rényi Fair K-means

---

[2]Public at: www.github.com/Trusted-AI/AIF360

7

Table 2. FBDE ($T = 32$) and baselines evaluated for COMPAS and ADULT datasets. The table reports the mean and s.t.d.

			DATA	M-E-1.0	M-E-0.9	M-R-1.0	M-R-0.9	MAXENT	TABFAIR	FAIRK
COMPAS (S = RACE)	DATA	RR	.662±.007	.966±.008	.977±.008	.944±.010	.964±.009	.992±.004	.632±.052	–
		SR	.747±.013	.988±.006	.899±.011	.978±.011	.896±.011	.992±.006	.727±.100	–
		KL	–	.135±.020	.129±.020	.132±.020	.127±.020	.164±.018	2.71±.735	–
	PRED	SR$_c$	.747±.020	.959±.025	.875±.025	.945±.027	.872±.024	.939±.018	.802±.074	–
		EO	.781±.034	.960±.026	.900±.041	.950±.028	.895±.039	.944±.020	.815±.079	–
		ACC	.660±.004	.641±.014	.653±.012	.642±.010	.656±.012	.648±.015	.591±.043	–
	CLUS	δPR	.334±.019	.332±.022	.334±.019	.334±.019	.319±.036	.333±.020	.286±.027	.259±.013
		δSR	.288±.055	.235±.082	.284±.055	.284±.055	.256±.084	.288±.059	.268±.063	.284±.036
		DIST	.065±.000	.078±.000	.078±.000	.078±.000	.078±.000	.078±.000	.078±.001	.069±.001
ADULT (S = SEX)	DATA	RR	.496±.002	.958±.003	.979±.003	.919±.004	.957±.003	.995±.002	.516±.023	–
		SR	.360±.005	.961±.005	.883±.006	.924±.006	.865±.006	.979±.004	.862±.101	–
		KL	–	.122±.006	.119±.006	.113±.006	.114±.006	.182±.005	1.68±.538	–
	PRED	SR$_c$	.360±.003	.818±.010	.766±.010	.793±.011	.753±.008	.919±.011	.823±.118	–
		EO	.471±.008	.959±.016	.908±.018	.935±.016	.895±.016	.981±.010	.867±.105	–
		ACC	.803±.003	.785±.002	.788±.002	.787±.002	.788±.002	.773±.005	.781±.006	–
	CLUS	δPR	.125±.067	.093±.019	.101±.025	.116±.027	.111±.033	.130±.061	.137±.038	.117±.039
		δSR	.426±.232	.252±.082	.302±.076	.190±.061	.254±.086	.468±.237	.437±.179	.296±.151
		DIST	.034±.000	.037±.000	.037±.000	.037±.000	.037±.000	.037±.000	.037±.000	.035±.000

($K = 4$) approach as per Baharlouei et al. (2019) (FAIRK).

In evaluating all approaches, we utilize 5-fold cross validation and evaluate all measurements using the test set (whenever appropriate). All training was on a MacBook Pro (16 GB memory, M1, 2020).

Additional experiments and discussion is presented in the Appendix, including, additional dataset comparisons, inprocessing algorithm comparison, and the runtime of approaches. Code for FBDE is available at `www.github.com/alexandersoen/fbde`.

Data Fairness Table 2 summarizes all results for COMPAS and ADULT. To evaluate the data fairness of our approach and baselines MAXENT and TABFAIR we compare the representation rate RR and statistical rate SR. We first note that TABFAIR does not target RR directly. To evaluate the information maintained after debiasing, we measure the KL between the original dataset and debiased samples.

All approaches apart from TABFAIR provides an increase in fairness (for both RR and SR). Across FBDE variants, we notice a trade-off between fairness and KL, where exact leveraging provides stronger fairness than relative leveraging at the cost of higher KL. FBDE performs better for both fairness and KL when compared to TABFAIR. We notice that TABFAIR struggles in the smaller COMPAS dataset and can even harm fairness (notice the s.t.d.). Notably, the KL is significantly worse than other approaches as a result of miss-matching of input distribution P support (with small $\delta = 10^{-9}$ added for unsupported regions) — possibly as a result of the transformation from discrete to continuous domains (Rajabi & Garibay, 2022); in addition to possible mode collapse in training (Thanh-Tung et al., 2019). MAXENT has the best SR, although it comes with a slight cost in KL when compared to FBDE.

It should be noted that FBDE's fairness target / budget is set to $\tau = 0.8$. So if a higher SR is required, τ can be increased. Although we take $T = 32$, as the leveraging coefficients decrease rapidly, we would expect a majority of the learning to occur in the initial iterations of the algorithm. This is indeed the case, as shown in Fig. 2 (left). We also note that the decrease in KL is proportional to the accuracy of WLs. This follows the intuition of Theorem 3.5 — more accurate weak learners implies larger boosting constants $\gamma_.$, which leads to larger drops in KL.

Prediction Fairness To evaluate the prediction fairness, we evaluate a decision tree classifier (CLF) (from `scikit-learn` with max depth of 32) trained on debiased samples. CLF is evaluated on CLF's statistical rate (with $\hat{Y} = 1$) (SR$_c$) and equality of opportunity ratio / true positive rate ratio (EO). Accuracy (ACC) is evaluated to measure the degradation from utilizing debiased samples.

When comparing for prediction, all approaches compared to data provide an increase for both SR$_c$ and EO with a slight decrease in ACC. For COMPAS, the FBDE variants and M-E-1.0 are comparable across fairness and utility; with TABFAIR slightly lower in fairness and accuracy scores. In ADULT, the FBDE variants and TABFAIR are similar in performance, with MAXENT having the best SR$_c$ and EO at the cost of the worst ACC.

Clustering Fairness To evaluate the clustering fairness, a $K(= 4)$-Means classifier (from `scikit-learn`) is trained

using the debiased samples. The fairness considerations in clustering is two fold. First we measure the difference in fairness across clusters (lower is better): we take the difference between the min and max ratio of privileged data points ($S = 1$) in each cluster (as per Chierichetti et al. (2017)) (δPR); and between ratios of statistical rate (δSR). Secondly, to evaluate the quality, we measure the average Manholobis distance to designated cluster centers (DIST).

FAIRK has the best utility DIST across approaches. Interestingly, pre-processing approaches can still be competitive to FAIRK across the fairness measures. In particular, δSR of the pre-processing approaches, particularly FBDE algorithms, can actually beat FAIRK— which is perhaps expected as data SR is specifically targeted. In ADULT, δPR was also significantly improved using pre-processing.

Interpretability Uniquely, FBDE allows for the learned distribution to be examined by analyzing the WLs learned. For instance, Fig. 1 depicts a DT WL learned for COMPAS by M-E-1.0 in one of its folds. Importantly, the 'realness' and 'fakeness' equate to modifications of the initial perfectly fair distribution Q_0 to be come closer to P: by Eq. (5) real predictions are up-weighted while fake predictions are down-weighted. This can be used to examine the underlying bias within input P. For example, taking 'True' (twice) on the "race"-"two_year_recid" path ends on a majority "Fake" node, which indicates that to get closer to P, the probability of non-Caucasians re-offenders was increased.

From a proactive point of view, one can instead restrict the hypothesis class in which the WL outputs. For instance, one can restrict the input domain to not include sensitive attributes — however, this may a cause harm in KL utility, and may still potentially learn spurious correlations in the form of proxies (Datta et al., 2017). Alternatively, the interpretability of the WLs allows for a human-in-the-loop algorithm: an auditor can become the stopping criteria of FBDE by comparing the gain in utility of an additional WL against the possible unfairness its reweighting could causes.

Continuous Domain In addition to the discrete datasets of COMPAS and ADULT, we consider the MINNEAPO-LIS police stop dataset[3] (binary S = race) which contains numeric / continuous features. In particular, we consider only a subset of features, taking position, race, and 'person searched'. We evaluate M-E-1.0 with 1 hidden layer (20 neuron) neural networks as WLs, with $T = 16$ iterations.

Fig. 2 (right) plots the change in SR versus KL across boosted updates. Here, a small $\delta = 10^{-9}$ probability value is added to a binned domain to estimate the KL. We verify that the SR stays above the budget in this continuous domain setting, where $SR(P) = 0.684 \pm 0.005$. Similarly,

---

[3]Public at: www.opendata.minneapolismn.gov/datasets/police-stop-data

the RR is improved (see Appendix). However, we note that the learned RR is lower than what theory may suggest, *i.e.*, M-E-1.0 gives $RR(Q_T) = 0.847 \pm 0.002$ which is slightly lower than the expected $\sqrt{\tau} \approx 0.894$ given by Theorem 3.2 (original $RR(P) = 0.183 \pm 0.001$). We attribute this miss-match in theory due to numerical approximation error in the estimation of samples via LMC and the MCMC estimation of Eq. (8), where better estimates could improve performance. For instance, in our experiments we only use the Unadjusted Langevin algorithm for LMC; more sophisticated methods can be used for better samples (Roberts & Tweedie, 1996). One differing aspect readers may notice is the non-monotonicity of the KL curve. In addition to numerical approximation error, the non-monotonicity may be occurring due to the binned (δ) estimation of the KL.

6. Limitations and Conclusion

In this paper, we introduce a new boosting algorithm, FBDE, which learns exponential family distributions which are both representation rate and statistical rate fair. To conclude, we highlight a few limitations of FBDE.

Firstly, FBDE is a boosting algorithm and thus relies on the WLA holding, *i.e.*, our fairness and convergence guarantees rely on this. Thus the performance of the WLs should be interrogated to ensure that FBDE does not harm fairness and cause subsequent social harm, *e.g.*, Fig. 2 (left).

Secondly, we reiterate that FBDE is not an unilateral replacement of other (types of) fairness algorithms, where specialized algorithms can provide stronger guarantees for specific criteria. Instead, FBDE targets an upstream source of unfairness in the ML pipeline, which can eventually allow for other forms of fairness. Nevertheless, when attempting to achieve these downstream fairness metrics one should be careful by examining FBDE's performance per boosting iterations. Indeed, some data settings can cause FBDE to have critical failure in, *e.g.*, SR_C, see Appendix O.

Lastly, we note that the fairness guarantees (Theorems 3.2 and 3.3) can break in practice as a result of numerical approximation error — especially in continuous domain datasets. We leave improvements of, *e.g.*, sampling for FBDE in continuous domains for future work.

Acknowledgements

References

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H.-M. A reductions approach to fair classification. In *ICML'18*, pp. 60–69, 2018.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine bias: There's software used across the country to predict future criminals. *And it's biased against blacks. ProPublica*, 23, 2016.

Baharlouei, S., Nouiehed, M., Beirami, A., and Razaviyayn, M. Rényi fair inference. In *International Conference on Learning Representations*, 2019.

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.

Calders, T. and Žliobaitė, I. Why unbiased computational processes can lead to discriminative decision procedures. In *Discrimination and privacy in the information society*, pp. 43–57. Springer, 2013.

Calders, T., Kamiran, F., and Pechenizkiy, M. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pp. 13–18, 2009.

Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K. N., and Varshney, K. R. Optimized pre-processing for discrimination prevention. In *NeurIPS'17*, pp. 3992–4001, 2017.

Celis, L. E., Keswani, V., and Vishnoi, N. Data preprocessing to mitigate bias: A maximum entropy based approach. In *International Conference on Machine Learning*, pp. 1349–1359, 2020.

Chierichetti, F., Kumar, R., Lattanzi, S., and Vassilvitskii, S. Fair clustering through fairlets. In *NeurIPS'17*, pp. 5036–5044, 2017.

Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. F., and Sun, J. Generating multi-label discrete patient records using generative adversarial networks. In *Machine learning for healthcare conference*, pp. 286–305. PMLR, 2017.

Cranko, Z. and Nock, R. Boosted density estimation remastered. In *International Conference on Machine Learning*, pp. 1416–1425. PMLR, 2019.

Datta, A., Fredrikson, M., Ko, G., Mardziel, P., and Sen, S. Use privacy in data-driven systems: Theory and experiments with machine learnt programs. In *24th ACM SIGSAC*, pp. 1193–1210, 2017.

De-Arteaga, M., Romanov, A., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., Geyik, S., Kenthapadi, K., and Kalai, A. T. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 120–128, 2019.

Do, H., Putzel, P., Martin, A. S., Smyth, P., and Zhong, J. Fair generalized linear models with a convex penalty. In *International Conference on Machine Learning*, pp. 5286–5308. PMLR, 2022.

Dua, D. and Karra Taniskidou, E. Uci machine learning repository [http://archive. ics. uci. edu/ml]. irvine, ca: University of california. *School of Information and Computer Science*, 2017.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R.-S. Fairness through awareness. In *ITCS'12*, pp. 214–226, 2012.

Fabris, A., Messina, S., Silvello, G., and Susto, G. A. Algorithmic fairness datasets: the story so far. *arXiv preprint arXiv:2202.01711*, 2022.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. In *21st KDD*, pp. 259–268, 2015.

Gordaliza, P., Del Barrio, E., Fabrice, G., and Loubes, J.-M. Obtaining fairness using optimal transport theory. In *International Conference on Machine Learning*, pp. 2357–2365. PMLR, 2019.

Grathwohl, W., Swersky, K., Hashemi, M., Duvenaud, D., and Maddison, C. Oops i took a gradient: Scalable sampling for discrete distributions. In *International Conference on Machine Learning*, pp. 3831–3841. PMLR, 2021.

Husain, H., Balle, B., Cranko, Z., and Nock, R. Local differential privacy for sampling. In *International Conference on Artificial Intelligence and Statistics*, pp. 3404–3413. PMLR, 2020.

Johndrow, J. E. and Lum, K. An algorithm for removing sensitive information: application to race-independent recidivism prediction. *Annals of Applied Statistics*, 13(1): 189–220, 2019.

Kamiran, F. and Calders, T. Classifying without discriminating. In *2009 2nd international conference on computer, control and communication*, pp. 1–6. IEEE, 2009.

Kamiran, F. and Calders, T. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

Kamiran, F., Karim, A., and Zhang, X. Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining*, pp. 924–929, 2012.

Kang, M., Li, L., Weber, M., Liu, Y., Zhang, C., and Li, B. Certifying some distributional fairness with sub-population decomposition. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=6mej19W1ppP.

Kasiviswanathan, S. P., Lee, H. K., Nissim, K., Raskhodnikova, S., and Smith, A. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

Kay, M., Matuszek, C., and Munson, S. A. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd annual acm conference on human factors in computing systems*, pp. 3819–3828, 2015.

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.

King, G. and Zeng, L. Logistic regression in rare events data. *Political analysis*, 9(2):137–163, 2001.

Lechner, T., Ben-David, S., Agarwal, S., and Ananthakrishnan, N. Impossibility results for fair representations. *arXiv preprint arXiv:2107.03483*, 2021.

Nielsen, F. and Garcia, V. Statistical exponential families: A digest with flash cards. *arXiv preprint arXiv:0911.4863*, 2009.

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.

Rajabi, A. and Garibay, O. O. Tabfairgan: Fair tabular data generation with generative adversarial networks. *Machine Learning and Knowledge Extraction*, 4(2):488–501, 2022.

Roberts, G. O. and Tweedie, R. L. Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, pp. 341–363, 1996.

Schapire, R. E. and Singer, Y. Improved boosting algorithms using confidence-rated predictions. *MLJ*, 37:297–336, 1999.

Schrouff, J., Harris, N., Koyejo, S., Alabdulmohsin, I. M., Schnider, E., Opsahl-Ong, K., Brown, A., Roy, S., Mincu, D., Chen, C., et al. Diagnosing failures of fairness transfer across distribution shift in real-world medical settings. *Advances in Neural Information Processing Systems*, 35: 19304–19318, 2022.

Soen, A., Alabdulmohsin, I., Koyejo, O. O., Mansour, Y., Moorosi, N., Nock, R., Sun, K., and Xie, L. Fair wrapping for black-box predictions. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=rxrLt7rTlAr.

Thanh-Tung, H., Tran, T., and Venkatesh, S. Improving generalization and stability of generative adversarial networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=ByxPYjC5KQ.

van Breugel, B., Kyono, T., Berrevoets, J., and van der Schaar, M. Decaf: Generating fair synthetic data using causally-aware generative networks. *Advances in Neural Information Processing Systems*, 34:22221–22233, 2021.

Van der Laan, P. The 2001 census in the netherlands. In *conference The Census of Population*, 2000.

Wang, H., Ustun, B., and Calmon, F. Repairing without retraining: Avoiding disparate impact with counterfactual distributions. In *International Conference on Machine Learning*, pp. 6618–6627. PMLR, 2019.

Williamson, R. C. and Menon, A. K. Fairness risk measures. In *International Conference on Machine Learning*, pp. 6786–6797, 2019.

Xu, D., Yuan, S., Zhang, L., and Wu, X. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pp. 570–575. IEEE, 2018.

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.

Appendix

Abstract

This is the Appendix to Paper "Fair Densities via Boosting the Sufficient Statistics of Exponential Families". To differentiate with the numbering in the main file, the sectioning is letter-based (A, B, ..., *etc*.). Thus, Theorems in the Appendix will be displayed as (A.1, B.3, ..., *etc*.). Additionally, figure and tables in the Appendix are numbered with Roman numerals (I, II, ..., *etc*.).

Table of contents

A. Proof of Lemma 3.1

Proof: To lower bound the pairwise statistical rate, we utilize the 'unrolled' computation of the required marginal measures. For arbitrary y, s we have

$$Q_t(y|s) = \frac{Q_t(y,s)}{Q_t(s)} = \frac{Q_{t-1}(y,s)}{Q_{t-1}(s)} \cdot \frac{Z_t(y,s)}{Z_t(s)} \cdots = \frac{Q_0(y,s)}{Q_0(s)} \cdot \prod_{i=1}^t \frac{Z_i(y,s)}{Z_i(s)} = Q_0(y|s) \cdot \prod_{i=1}^t \frac{Z_i(y,s)}{Z_i(s)}.$$

Thus, the pairwise statistical rate for s, s' is recovered immediately by taking a lower bound on the statistical rate of Q_0. □

B. Proof of Theorem 3.2

To prove Theorem 3.2, we split the proof into the statistical rate component of the theorem and the representation rate component of the theorem.

B.I. Statistical Rate

To prove the exact fairness guarantee for statistical rate, we first use the following lemmas.

Lemma B.1. *For $T \geq 1$, we have*

$$\mathrm{SR}(Q_T) \geq \mathrm{SR}_0 \cdot \exp\left(-4C \sum_{i=1}^T \vartheta_t\right). \tag{12}$$

Proof: We first consider bounds on the difference between log-normalizer (log-partition) functions. First by taking the smallest and largest values of the classifier $c_t(.)$ (which assumed to be bounded in $[-C, C]$), we have

$$-C \cdot \vartheta_t \leq \vartheta_t c_t(x, y, s) \leq C \cdot \vartheta_t$$

Then by taking the exponential, integrand (w.r.t., measure $dQ_{t-1}(x|y, s)$), and logarithm, we get

$$\begin{aligned}-C\vartheta_t &\leq & \log\left(\int_{\mathcal{X}} \exp(\vartheta_t c_t(x, y, s)) dQ_{t-1}(x|y, s)\right) &\leq C\vartheta_t \\ -C\vartheta_t &\leq & \log Z_t(y, s) &\leq C\vartheta_t.\end{aligned} \tag{13}$$

Then by taking the largest difference between $\log Z_t(y, s)$ and $\log Z_t(y, s')$, we have that for any s, s'

$$-2C\vartheta_t \leq \log Z_t(y, s) - \log Z_t(y, s') \leq 2C\vartheta_t. \tag{14}$$

Identically, we can bound the difference for $\log Z_t(s)$ by replacing the measure in the integrand step with $dQ_{t-1}(x, y|s)$, giving:

$$-2C\vartheta_t \leq \log Z_t(s) - \log Z_t(s') \leq 2C\vartheta_t. \tag{15}$$

Remark B.2. We can also bound $\log Z_t$ (*i.e.*, Appendix B.I) using the same method by, again, replacing the measure to be dQ_{t-1}

Together, (14) and (15) allows us to simplify Lemma 3.1:

$$\mathrm{SR}(Q_t, s, s'; y) \geq \mathrm{SR}_0 \cdot \prod_{i=1}^t \left(\frac{Z_i(s')}{Z_i(s)} \cdot \frac{Z_i(y, s)}{Z_i(y, s')}\right) = \mathrm{SR}_0 \cdot \exp\left(\sum_{i=1}^t \log\left(\frac{Z_i(s')}{Z_i(s)} \cdot \frac{Z_i(y, s)}{Z_i(y, s')}\right)\right)$$

$$\geq \mathrm{SR}_0 \cdot \exp\left(-4C \sum_{i=1}^t \vartheta_i\right).$$

As this holds for any t, s, s', the bound holds for $\mathrm{SR}(Q_T)$ as required. □

We can now prove the statistical rate part of Theorem 3.2.

Proof: Taking $\vartheta_t = -(C2^{t+1})^{-1}\log(\tau/\text{SR}_0)$, from Lemma B.1 we get:

$$\text{SR}(Q_t) \geq \text{SR}_0 \cdot \exp\left(-4C\sum_{i=1}^{t}\vartheta_i\right) = \text{SR}_0 \cdot \exp\left(2\log\left(\frac{\tau}{\text{SR}_0}\right)\sum_{i=1}^{t}\frac{1}{2^i}\right) = \text{SR}_0 \cdot \exp\left(2\log\left(\frac{\tau}{\text{SR}_0}\right)\left(1-\frac{1}{2^t}\right)\right)$$

$$\geq \text{SR}_0 \cdot \exp\left(\log\left(\frac{\tau}{\text{SR}_0}\right)\right) = \tau.$$

The last inequality follows that we are only considering $t \geq 1$. Thus $\text{SR}(Q_T) > \tau$ as required. □

We now move to proving the representation rate component of Theorem 3.2.

B.II. Representation Rate

To lower bound the corresponding representation rate, we present a similar lemma to Lemma B.1.

Lemma B.3. *For $T \geq 1$, we have*

$$\text{RR}(Q_T) \geq \text{RR}_0 \cdot \exp\left(-2C\sum_{i=1}^{T}\vartheta_t\right). \tag{16}$$

Proof: Proving Lemma B.3 is similar to the proof of Lemma 3.1. Specifically, we use Eq. (15) to bound the representation rate.

First, we calculate the marginal distribution via a recursive relation similar to Lemma 3.1:

$$Q_t(s) = Q_{t-1}(s) \cdot \frac{Z_t(s)}{Z_t} = \ldots = Q_0(s) \cdot \prod_{i=1}^{t}\frac{Z_i(s)}{Z_t}.$$

Now, we Eq. (15) to bound the resulting representation rate:

$$\text{RR}(Q_t, s, s') \geq \text{RR}_0 \cdot \prod_{i=1}^{t}\frac{Z_i(s)}{Z_i(s')} = \text{RR}_0 \cdot \exp\left(\sum_{i=1}^{t}\log\left(\frac{Z_i(s)}{Z_i(s')}\right)\right)$$

$$\geq \text{RR}_0 \cdot \exp\left(-2C\sum_{i=1}^{t}\vartheta_t\right).$$

As the bound holds for all s, s', the bound holds for $\text{RR}(Q_t)$ required. □

Remark B.4. Notice that Lemma B.3 holds even when we restrict Q_t to only have non-sensitive features x and sensitive features s (*i.e.*, no labels being separately considered). This allows for leveraging schemes ϑ_t to be designed for representation rate specifically. For instance, for a target representation rate τ_{RR} we can establish exact and relative representation rate fairness guarantees by replacing "τ/SR_0" to "$\tau_{\text{RR}}/\text{RR}_0$".

We can now prove the representation rate component of Theorem 3.2.

Proof: Taking $\vartheta_t = -(C2^{t+1})^{-1}\log(\tau/\text{SR}_0)$, from Lemma B.3 we get:

$$\text{RR}(Q_t) \geq \text{RR}_0 \cdot \exp\left(-2C\sum_{i=1}^{t}\vartheta_t\right) = \text{RR}_0 \cdot \exp\left(\log\left(\frac{\tau}{\text{SR}_0}\right)\sum_{i=1}^{t}\frac{1}{2^i}\right) = \text{RR}_0 \cdot \exp\left(\log\left(\frac{\tau}{\text{SR}_0}\right)\left(1-\frac{1}{2^t}\right)\right)$$

$$\leq \text{RR}_0 \cdot \exp\left(\frac{1}{2}\log\left(\frac{\tau}{\text{SR}_0}\right)\right) = \text{RR}_0 \cdot \sqrt{\frac{\tau}{\text{SR}_0}}.$$

As required. □

C. Proof of Theorem 3.3

The relative fairness guarantee is proven similarly to the exact fairness guarantee in Appendix B. We also split the proof into statistical rate and representation rate components of the theorem.

C.I. Statistical Rate

Proof: Taking $\vartheta_t = -(4Ct)^{-1}\log(\tau/\mathrm{SR}_0)$, from Lemma B.1 we get:

$$
\begin{aligned}
\mathrm{SR}(\mathrm{Q}_t) &\geq \mathrm{SR}_0 \cdot \exp\left(-4C\sum_{i=1}^{t}\vartheta_i\right) \\
&= \mathrm{SR}_0 \cdot \exp\left(\log\left(\frac{\tau}{\mathrm{SR}_0}\right)\sum_{i=1}^{t}\frac{1}{i}\right) \\
&> \mathrm{SR}_0 \cdot \exp\left(\log\left(\frac{\tau}{\mathrm{SR}_0}\right)\cdot\left(1+\int_1^t\frac{1}{i}\mathrm{d}i\right)\right) \\
&= \mathrm{SR}_0 \cdot \exp\left(\log\left(\frac{\tau}{\mathrm{SR}_0}\right)\cdot(1+\log t)\right) \\
&= \mathrm{SR}_0 \cdot \exp\left(\log\left(\frac{\tau}{\mathrm{SR}_0}\right)\right)\cdot\exp\left(\log\left(\frac{\tau}{\mathrm{SR}_0}\right)\cdot\log t\right) \\
&= \tau \cdot \left(\frac{\tau}{\mathrm{SR}_0}\right)^{\log t} \\
&= \mathrm{SR}_0^{-\log t}\cdot\tau^{1+\log t}.
\end{aligned}
$$

Here we note that $\log(\tau/\mathrm{SR}_0) < 0$ as $\tau < \mathrm{SR}_0$. Furthermore, note that $\mathrm{SR}_0^{-\log t}$ is an increasing function of $t \geq 1$, Thus taking $t = 1$ for this term:

$$
\mathrm{SR}(\mathrm{Q}_t) > \mathrm{SR}_0^{-\log 1}\cdot\tau^{1+\log t} = \tau^{1+\log t}.
$$

Thus $\mathrm{SR}(\mathrm{Q}_T) > \tau^{1+\log T}$ as required. $\qquad\square$

C.II. Representation Rate

Using Lemma B.3, we prove the representation rate component of Theorem 3.3.

Proof: Taking $\vartheta_t = -(4Ct)^{-1}\log(\tau/\mathrm{SR}_0)$, from Lemma B.3 we get:

$$
\begin{aligned}
\mathrm{RR}(\mathrm{Q}_t) &\geq \mathrm{RR}_0 \cdot \exp\left(-2C\sum_{i=1}^{t}\vartheta_t\right) \\
&= \mathrm{RR}_0 \cdot \exp\left(\frac{1}{2}\cdot\log\left(\frac{\tau}{\mathrm{SR}_0}\right)\sum_{i=1}^{t}\frac{1}{i}\right) \\
&> \mathrm{RR}_0 \cdot \exp\left(\frac{1}{2}\cdot\log\left(\frac{\tau}{\mathrm{SR}_0}\right)\left(1+\int_1^t\frac{1}{i}\mathrm{d}i\right)\right) \\
&= \mathrm{RR}_0 \cdot \exp\left(\frac{1}{2}\cdot\log\left(\frac{\tau}{\mathrm{SR}_0}\right)(1+\log t)\right) \\
&= \mathrm{RR}_0 \cdot \left(\sqrt{\frac{\tau}{\mathrm{SR}_0}}\right)^{1+\log t}.
\end{aligned}
$$

As required. $\qquad\square$

D. Proof of Theorem 3.5

Lemma D.1. *The* KL*-drop is given by:*

$$
\mathrm{KL}(\mathrm{P},\mathrm{Q}_{t-1}) - \mathrm{KL}(\mathrm{P},\mathrm{Q}_t) = \vartheta_t\cdot\mathbb{E}_{\mathrm{P}}[c_t] - \log\mathbb{E}_{\mathrm{Q}_{t-1}}[\exp(\vartheta_t c_t)]. \tag{17}
$$

Proof: The drop is given by the following:

$$
\begin{aligned}
\mathrm{KL}(\mathrm{P}, \mathrm{Q}_{t-1}) - \mathrm{KL}(\mathrm{P}, \mathrm{Q}_t) &= \int \log\left(\frac{\mathrm{P}}{\mathrm{Q}_{t-1}}\right) d\mathrm{P} - \int \log\left(\frac{\mathrm{P}}{\mathrm{Q}_t}\right) d\mathrm{P} \\
&= \int \log\left(\frac{\mathrm{P}}{\mathrm{Q}_{t-1}}\right) d\mathrm{P} - \int \log\left(\frac{\mathrm{P}}{\exp\left(\vartheta_t c_t - \log Z_t\right) \cdot \mathrm{Q}_{t-1}}\right) d\mathrm{P} \\
&= \int_{\mathcal{X} \times \mathcal{Y} \times \mathcal{S}} \log\left(\frac{\exp\left(\vartheta_t c_t(x,y,s) - \log Z_t\right) \cdot \mathrm{Q}_{t-1}(x,y,s)}{\mathrm{Q}_{t-1}(x,y,s)}\right) d\mathrm{P}(x,y,s) \\
&= \int_{\mathcal{X} \times \mathcal{Y} \times \mathcal{S}} \left(\vartheta_t c_t(x,y,s) - \log Z_t\right) d\mathrm{P}(x,y,s) \\
&= \vartheta_t \cdot \mathbb{E}_{\mathrm{P}}[c_t] - \log Z_t \\
&= \vartheta_t \cdot \mathbb{E}_{\mathrm{P}}[c_t] - \log \mathbb{E}_{\mathrm{Q}_{t-1}}[\exp(\vartheta_t c_t)],
\end{aligned}
$$

where the last line follows from the definition of the normalizing term (noting it can be expressed as an expectation). ∎

Lemma D.2. *Given the WLA and WLs with bounding constant $C > 0$, given coefficients:*

$$
A = \frac{1}{4C^2}\left[\exp(C) - \exp(-C)\right]; \qquad B = \frac{1}{2C}\left[\exp(C) - \exp(-C)\right]; \qquad K = \frac{\exp(C) + \exp(-C)}{2} - AC^2,
$$

then for WL $c_t(.)$,

$$
\mathbb{E}_{\mathrm{Q}_{t-1}}[\exp(c_t)] < \exp(-\Gamma(\mathbb{V}_{\mathrm{Q}_{t-1}}[c_t], \gamma_{\mathrm{Q}}^t)) \tag{18}
$$

where

$$
\Gamma(\mathbb{V}_{\mathrm{Q}_{t-1}}[c_t], \gamma_{\mathrm{Q}}^t) = \log\left((A \cdot \mathbb{V}_{\mathrm{Q}_{t-1}}[c_t] - C^2 A \cdot \gamma_{\mathrm{Q}}^t + K)^{-1}\right). \tag{19}
$$

Proof: First notice that $f(x) = Ax^2 + Bx + K$ is an upper bound of $\exp(x)$ on $[-C, C]$, *i.e.*, $\exp(x) \le f(x)$. The coefficients are derived from fitting $f(x)$ to satisfy points $(-C, \exp(-C))$ and $(C, \exp(C))$; and ensuring monotonicity on the interval. Fig. I depicts the bound (and a comparison to a prior bound used in Husain et al. (2020)).

From this, we have:

$$
\begin{aligned}
\mathbb{E}_{\mathrm{Q}_{t-1}}[\exp(c_t)] &\le \mathbb{E}_{\mathrm{Q}_{t-1}}[Ac_t^2 + Bc_t + K] \\
&= A \cdot \mathbb{E}_{\mathrm{Q}_{t-1}}[c_t^2] + B \cdot \mathbb{E}_{\mathrm{Q}_{t-1}}[c_t] + K \\
&= A \cdot (\mathbb{V}_{\mathrm{Q}_{t-1}}[c_t] + \mathbb{E}_{\mathrm{Q}_{t-1}}[c_t]^2) + B \cdot \mathbb{E}_{\mathrm{Q}_{t-1}}[c_t] + K \\
&= A \cdot \mathbb{V}_{\mathrm{Q}_{t-1}}[c_t] + \mathbb{E}_{\mathrm{Q}_{t-1}}[c_t] \cdot (A \cdot \mathbb{E}_{\mathrm{Q}_{t-1}}[c_t] + B) + K.
\end{aligned}
$$

Note from the WLA we have $\mathbb{E}_{\mathrm{Q}_{t-1}}[-c_t] \ge C \cdot \gamma_{\mathrm{Q}}^t \iff \mathbb{E}_{\mathrm{Q}_{t-1}}[c_t] \le -C \cdot \gamma_{\mathrm{Q}}^t < 0$. Furthermore, notice that $Az + B > 0$ and increasing for $z \in [-C, C]$:

$$
Az + B = \frac{\exp(C) - \exp(-C)}{2C}\left[\frac{z}{2C} + 1\right] > 0 \qquad \text{(for } z \in [-C, C]\text{)}.
$$

Thus,

$$
\begin{aligned}
\mathbb{E}_{\mathrm{Q}_{t-1}}[\exp(c_t)] &\le A \cdot \mathbb{V}_{\mathrm{Q}_{t-1}}[c_t] + \mathbb{E}_{\mathrm{Q}_{t-1}}[c_t] \cdot (A \cdot \mathbb{E}_{\mathrm{Q}_{t-1}}[c_t] + B) + K \\
&\le A \cdot \mathbb{V}_{\mathrm{Q}_{t-1}}[c_t] - \gamma_{\mathrm{Q}} \cdot C \cdot (A \cdot \mathbb{E}_{\mathrm{Q}_{t-1}}[c_t] + B) + K \\
&\le A \cdot \mathbb{V}_{\mathrm{Q}_{t-1}}[c_t] - \gamma_{\mathrm{Q}} \cdot C \cdot (B - C \cdot A) + K.
\end{aligned}
$$

Which gives the Lemma, as per:

$$
\begin{aligned}
\mathbb{E}_{\mathrm{Q}_{t-1}}[\exp(c_t)] &\le A \cdot \mathbb{V}_{\mathrm{Q}_{t-1}}[c_t] - \gamma_{\mathrm{Q}} \cdot C \cdot (B - C \cdot A) + K \\
&= \exp\left(\log\left(A \cdot \mathbb{V}_{\mathrm{Q}_{t-1}}[c_t] - \gamma_{\mathrm{Q}} \cdot C \cdot (B - C \cdot A) + K\right)\right) \\
&= \exp\left((-1) \cdot -\log\left(A \cdot \mathbb{V}_{\mathrm{Q}_{t-1}}[c_t] - \gamma_{\mathrm{Q}} \cdot C \cdot (B - C \cdot A) + K\right)\right) \\
&= \exp\left(-\log\left(A \cdot \mathbb{V}_{\mathrm{Q}_{t-1}}[c_t] - \gamma_{\mathrm{Q}} \cdot C \cdot (B - C \cdot A) + K\right)^{-1}\right).
\end{aligned}
$$

Finally, we conclude the proof by noting that $2CA = B$. ∎

Remark D.3. We note that the quadratic approximation $f(x) = Ax^2 + Bx + K$ gets worse as C increases. Intuitively, as the interval approximation increases, the quadratic finds it more difficult to "catch-up" to the exponential growth of $x \mapsto \exp(x)$. See Fig. I for a pictorial view of this.
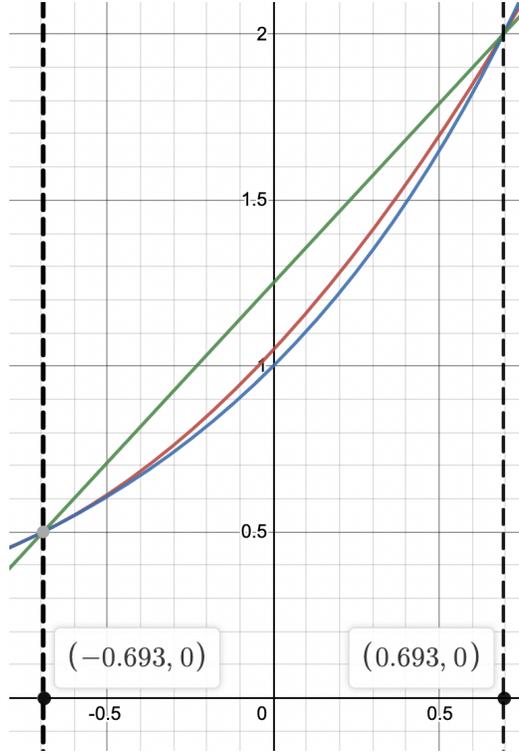

Figure I. Upper bounds for $\exp(x)$. The curves are: original $\exp(x)$ function (blue); our quadratic function (red); and linear function used in Husain et al. (2020) (green).

Now we can prove Theorem 3.5.

Proof: Given that $\tau/\text{SR}_0 > \exp(-4C)$, we have that for each leveraging scheme in Theorems 3.2 and 3.3 the leverage is bounded by $\vartheta_t \leq 1$ for all t.

Thus with Lemmas D.1 and D.2, we bound the KL:

$$
\begin{aligned}
\text{KL}(P, Q_{t-1}) - \text{KL}(P, Q_t) &= \vartheta_t \cdot \mathbb{E}_P[c_t] - \log \mathbb{E}_{Q_{t-1}}[\exp(\vartheta_t c_t)] && \text{(Lemma D.1)} \\
&\geq \vartheta_t \cdot \mathbb{E}_P[c_t] - \vartheta_t \cdot \log \mathbb{E}_{Q_{t-1}}[\exp(c_t)] && \text{(a)} \\
&= \vartheta_t \cdot (\mathbb{E}_P[c_t] - \log \mathbb{E}_{Q_{t-1}}[\exp(c_t)]) \\
&\geq \vartheta_t \cdot (\mathbb{E}_P[c_t] - \log \exp(-\Gamma(\mathbb{V}_{Q_{t-1}}[c_t], \gamma_Q^t))) && \text{(Lemma D.2)} \\
&= \vartheta_t \cdot (\mathbb{E}_P[c_t] + \Gamma(\mathbb{V}_{Q_{t-1}}[c_t], \gamma_Q^t) \\
&\geq \vartheta_t \cdot (C \cdot \gamma_Q^t + \Gamma(\mathbb{V}_{Q_{t-1}}[c_t], \gamma_Q^t), && \text{(WLA cond. on P)}
\end{aligned}
$$

where (a) is given by Jensen's inequality and noting that $x \mapsto x^{1/\vartheta_t}$ is convex as $\vartheta_t \leq 1$. $\qquad \square$

E. Proof of Theorem 3.6

To prove the theorem, we breakdown the proof into the corresponding upper and lower bounds. In particular, we first prove general Lemmas first and then specialize for the specific leveraging coefficients ϑ_t to give the Theorem.

E.I. Upper Bound

Lemma E.1. *Suppose $C > 0$, then*

$$\text{KL}(\text{P}, \text{Q}_0) - \text{KL}(\text{P}, \text{Q}_T) \leq 2C \cdot \sum_{k=1}^{T} \vartheta_k. \tag{20}$$

Proof: We first note that by unrolling the definition of Q_T (Eq. (5)), the density can be stated directly in terms of the initial distribution:

$$\text{Q}_T = \text{Q}_0 \exp(\langle \boldsymbol{\vartheta}, \boldsymbol{c} \rangle - \varphi(\boldsymbol{\vartheta})), \tag{21}$$

where $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_T)$, $\boldsymbol{c} = (c_1, \ldots, c_T)$, and $\varphi(\boldsymbol{\vartheta}) = \sum_{i=1}^{T} \log Z_i$.

Thus similarly to Lemma D.1, calculate the total drop:

$$
\begin{aligned}
\text{KL}(\text{P}, \text{Q}_0) - \text{KL}(\text{P}, \text{Q}_T) &= \int \log\left(\frac{\text{P}}{\text{Q}_0}\right) d\text{P} - \int \log\left(\frac{\text{P}}{\text{Q}_T}\right) d\text{P} \\
&= \int \log\left(\frac{\text{P}}{\text{Q}_0}\right) d\text{P} - \int \log\left(\frac{\text{P}}{\text{Q}_0 \exp(\langle \boldsymbol{\vartheta}, \boldsymbol{c} \rangle - \varphi(\boldsymbol{\vartheta}))}\right) d\text{P} \\
&= \int \log\left(\frac{\text{Q}_0 \exp\left(\langle \boldsymbol{\vartheta}, \boldsymbol{c} \rangle - \varphi(\boldsymbol{\vartheta})\right)}{\text{Q}_0}\right) d\text{P} \\
&= \int \left(\langle \boldsymbol{\vartheta}, \boldsymbol{c} \rangle - \varphi(\boldsymbol{\vartheta})\right) d\text{P} \\
&= \left\langle \boldsymbol{\vartheta}, \int \boldsymbol{c} \, d\text{P} \right\rangle - \varphi(\boldsymbol{\vartheta}).
\end{aligned}
$$

We now upper bound each of these terms. Firstly, we note that as each weak learner c_t is upper bounded by C, we can easily bound the first term:

$$\left\langle \boldsymbol{\vartheta}, \int \boldsymbol{c} \, d\text{P} \right\rangle \leq \left\langle \boldsymbol{\vartheta}, \int C \cdot \mathbf{1} \, d\text{P} \right\rangle = C \cdot \sum_{k=1}^{T} \vartheta_k.$$

Similarly, we can bound the second term by noting Remark B.2. Thus using an identical bound to Appendix B.I we get:

$$\varphi(\boldsymbol{\vartheta}) = \sum_{i=1}^{T} \log Z_i \geq \sum_{k=1}^{T} (-C\vartheta_k) = -C \cdot \sum_{k=1}^{T} \vartheta_k$$

Thus it follows that

$$\text{KL}(\text{P}, \text{Q}_0) - \text{KL}(\text{P}, \text{Q}_T) \leq 2C \cdot \sum_{k=1}^{T} \vartheta_k.$$

As required. $\qquad\square$

E.II. Lower Bound

Lemma E.2. *Suppose $\alpha(\gamma) \doteq \min_t \Gamma(\mathbb{V}_{\text{Q}_{t-1}}[c_t], \gamma)/(\gamma C)$, $C > 0$, $\vartheta \leq 1$ for all t, and $T > 0$, then:*

$$\text{KL}(\text{P}, \text{Q}_0) - \text{KL}(\text{P}, \text{Q}_T) \geq (\gamma_\text{P} + \gamma_\text{Q} \cdot \alpha(\gamma_\text{Q})) \cdot C \cdot \sum_{k=1}^{T} \vartheta_k.$$

Proof: To establish the lower bound, we repeatedly apply Theorem 3.5:

$$\text{KL}(\text{P}, \text{Q}_T) \leq \text{KL}(\text{P}, \text{Q}_{T-1}) - \vartheta_T \cdot \Lambda_T \leq \quad \ldots \quad \leq \text{KL}(\text{P}, \text{Q}_0) - \sum_{k=1}^{T} \vartheta_k \cdot \Lambda_k.$$

Let $\alpha(\gamma) \doteq \min_t \Gamma(\mathbb{V}_{Q_{t-1}}[c_t], \gamma)/(\gamma C)$. Note that the variance is bounded by $0 \leq \mathbb{V}_{Q_{t-1}}[c_t] \leq \mathbb{E}_Q[c_t^2] \leq C^2$ and that $\Gamma(\cdot, \gamma)$ is a decreasing function for all $\gamma \in [0, 1]$. Thus we have

$$
\begin{aligned}
\text{KL}(P, Q_0) - \text{KL}(P, Q_T) &\geq \sum_{k=1}^{T} \vartheta_k \cdot \Lambda_k \\
&= \sum_{k=1}^{T} \vartheta_k \cdot \left(\gamma_P^t \cdot C + \Gamma(\mathbb{V}_{Q_{t-1}}[c_t], \gamma_Q^t) \right) \\
&= \sum_{k=1}^{T} \vartheta_k \cdot \left(\gamma_P \cdot C + \Gamma(\mathbb{V}_{Q_{t-1}}[c_t], \gamma_Q) \right) \quad &\text{(Fixed const. assum.)} \\
&\geq \sum_{k=1}^{T} \vartheta_k \cdot \left(\gamma_P \cdot C + \gamma_Q \cdot C \cdot \alpha(\gamma_Q) \right) \quad &\text{(Definition of } \alpha) \\
&= (\gamma_P + \gamma_Q \cdot \alpha(\gamma_Q)) \cdot C \cdot \sum_{k=1}^{T} \vartheta_k.
\end{aligned}
$$

As required. $\qquad\square$

E.III. For Fairness Leverage Coefficients

From Lemmas E.1 and E.2 we have that:

Theorem E.3. *Suppose* $\alpha(\gamma) \doteq \min_t \Gamma(\mathbb{V}_{Q_{t-1}}[c_t], \gamma)/(\gamma C)$, $C > 0$, $\vartheta \leq 1$ *for all t, and* $T > 0$, *then:*

$$
(\gamma_P + \gamma_Q \cdot \alpha(\gamma_Q)) \cdot C \cdot \sum_{k=1}^{T} \vartheta_k \leq \text{KL}(P, Q_0) - \text{KL}(P, Q_T) \leq 2C \cdot \sum_{k=1}^{T} \vartheta_k.
$$

To upper and lower bound the statistical difference $\Delta(Q_T)$, we upper and lower bound the sum of leveraging coefficients ϑ_t for each of the fairness guarantees.

E.III.1. EXACT FAIRNESS LEVERAGE

Proof: Taking $\vartheta_t = -(C2^{t+1})^{-1} \log(\tau/\text{SR}_0)$ from geometric series, we have

$$
\sum_{k=1}^{T} \vartheta_k = -\frac{\log(\tau/\text{SR}_0)}{2C} \sum_{k=1}^{T} \frac{1}{2^k} = -\frac{\log(\tau/\text{SR}_0)}{2C} \left(1 - \frac{1}{2^T} \right)
$$

Thus we have

$$
-\log(\tau/\text{SR}_0) \cdot \frac{\gamma_P + \gamma_Q \cdot \alpha(\gamma_Q)}{2} \cdot \left(1 - \frac{1}{2^T} \right) \leq \text{KL}(P, Q_0) - \text{KL}(P, Q_T) \leq -\log(\tau/\text{SR}_0) \left(1 - \frac{1}{2^T} \right)
$$

As required. $\qquad\square$

E.III.2. RELATIVE FAIRNESS LEVERAGE

Proof: From harmonic series, we have that

$$
\frac{1}{T} + \log T < \sum_{k=1}^{T} \frac{1}{k} < 1 + \log T.
$$

Thus taking $\vartheta_t = -(4Ct)^{-1} \log(\tau/\text{SR}_0)$ we have that

$$
-\log(\tau/\text{SR}_0) \cdot \frac{1}{4C} \cdot \left(\frac{1}{T} + \log T \right) < \sum_{k=1}^{T} \vartheta_t < -\log(\tau/\text{SR}_0) \cdot \frac{1}{4C} \cdot (1 + \log T).
$$

This gives the following bounds:

$$-\log(\tau/\text{SR}_0) \cdot \frac{\gamma_P + \gamma_Q \cdot \alpha(\gamma_Q)}{4} \cdot \left(\frac{1}{T} + \log T\right) \leq \text{KL}(P, Q_0) - \text{KL}(P, Q_T) \leq -\log(\tau/\text{SR}_0) \cdot \frac{1}{2} \cdot (1 + \log T).$$

As required. $\qquad\qquad\square$

F. Proof of Corollary 3.7

Proof: The statement follows directly from utilizing Theorem E.3 and utilizing the series lower bounds used in the proof of Theorem 3.6.

Indeed, we have

$$\text{KL}(P, Q_0) - \text{KL}(P, Q_T) \geq (\gamma_P + \gamma_Q \cdot \alpha(\gamma_Q)) \cdot C \cdot \sum_{k=1}^{T} \vartheta_k$$

$$\iff \text{KL}(P, Q_T) \leq \text{KL}(P, Q_0) - (\gamma_P + \gamma_Q \cdot \alpha(\gamma_Q)) \cdot C \cdot \sum_{k=1}^{T} \vartheta_k.$$

Thus, a sufficient conditions for $\text{KL}(P, Q_T) < \varepsilon$ is that the RHS is $< \varepsilon$.

Thus we are looking to find conditions where the following holds:

$$\frac{\text{KL}(P, Q_0) - \varepsilon}{(\gamma_P + \gamma_Q \cdot \alpha(\gamma_Q)) \cdot C} < \sum_{k=1}^{T} \vartheta_k. \tag{22}$$

We will take $\lambda = -\log(\tau/\text{SR}_0)$.

F.I. Exact Leverage

We can directly rewrite the condition as

$$\frac{\text{KL}(P, Q_0) - \varepsilon}{(\gamma_P + \gamma_Q \cdot \alpha(\gamma_Q)) \cdot C} < \sum_{k=1}^{T} \vartheta_k = \frac{\lambda}{2C}\left(1 - \frac{1}{2^T}\right)$$

$$2^T > \left(1 - 2 \cdot \frac{\text{KL}(P, Q_0) - \varepsilon}{\lambda \cdot (\gamma_P + \gamma_Q \cdot \alpha(\gamma_Q))}\right)$$

$$T > \frac{1}{\log 2} \log \cdot \left(1 - 2 \cdot \frac{\text{KL}(P, Q_0) - \varepsilon}{\lambda \cdot (\gamma_P + \gamma_Q \cdot \alpha(\gamma_Q))}\right).$$

As required.

F.II. Relative Leverage

For this leverage, we need a weaker bound. We use the fact that

$$\frac{\text{KL}(P, Q_0) - \varepsilon}{(\gamma_P + \gamma_Q \cdot \alpha(\gamma_Q)) \cdot C} < \sum_{k=1}^{T} \impliedby \frac{\text{KL}(P, Q_0) - \varepsilon}{(\gamma_P + \gamma_Q \cdot \alpha(\gamma_Q)) \cdot C} < \frac{\lambda}{4C} \cdot \log T.$$

Thus, we immediately have that

$$T > \exp\left(4 \cdot \frac{\text{KL}(P, Q_0) - \varepsilon}{\lambda \cdot (\gamma_P + \gamma_Q \cdot \alpha(\gamma_Q))}\right)$$

implies the statement holds. $\qquad\qquad\square$

Remark F.1. Tighter bounds can be made for relative leverage w.r.t. the "omega function". More generally, by leaving the leverage series as is (without simplification) gives sufficient condition for a bounded KL. However, the condition may not be w.r.t. T being large.

G. Proof of Lemma 4.2

Proof: We prove the following by contradiction. Suppose that there exists distribution $Q \in \mathcal{M}_{\varepsilon, Q_0}$ with $SR(Q) < SR_0 \cdot \exp(-\varepsilon)$. Thus for there exists a $s, s' \in \mathcal{S}$ such that

$$SR(Q, s', s; y) < SR_0 \cdot \exp(-\varepsilon) \implies SR(Q, s, s'; y) > \frac{1}{SR_0} \cdot \exp(\varepsilon).$$

But given that $SR(Q_0) = SR_0$, by definition, we have that

$$\forall a, b \in \mathcal{S} : \quad SR(Q_0, a, b; y) \geq SR_0$$

Thus we have that,

$$
\begin{aligned}
\max \left\{ \frac{SR(Q, s, s'; y)}{SR(Q_0, s, s'; y)}, \frac{SR(Q_0, s, s'; y)}{SR(Q, s, s'; y)} \right\} &\geq \frac{SR(Q, s, s'; y)}{SR(Q_0, s, s'; y)} \\
&\geq SR(Q, s, s'; y) \cdot SR(Q_0, s', s; y) \\
&\geq SR(Q, s, s'; y) \cdot sr_0 \\
&> \frac{1}{SR_0} \cdot \exp(\varepsilon) \cdot SR_0 = \exp(\varepsilon).
\end{aligned}
$$

However, $Q \in \mathcal{M}_{\varepsilon, Q_0}$ and thus by Definition 4.1 we have a contradiction.

Thus $Q \in \mathcal{M}_{\varepsilon, Q_0} \implies SR(Q) \geq SR_0 \cdot \exp(-\varepsilon)$.

□

H. Proof of Lemma 4.3

Proof: We first note that the condition of "$\forall s, s' \in \mathcal{S}$ we have either $SR(Q, s, s'; y) > 1$ or $SR(Q, s, s'; y) \leq SR(Q_0, s, s'; y) \leq 1$" is equivalent to "$\forall s, s' \in \mathcal{S}$ we have $SR(Q, s, s'; y) \leq 1$ implies $SR(Q, s, s'; y) \leq SR(Q_0, s, s'; y) \leq 1$".

We suppose that Q is an arbitrary distribution that satisfies this condition. Let $s, s' \in \mathcal{S}$ be arbitrary such that $SR(Q, s, s'; y) \leq 1$, and thus $SR(Q, s, s'; y) \leq SR(Q_0, s, s'; y) \leq 1$.

We have that

$$\frac{SR(Q, s, s'; y)}{SR(Q_0, s, s'; y)} \geq SR(Q, s, s'; y)$$

and

$$\frac{SR(Q, s, s'; y)}{SR(Q_0, s, s'; y)} \leq 1 \leq \frac{SR(Q_0, s, s'; y)}{SR(Q, s, s'; y)}.$$

As $SR(Q) \geq \exp(-\varepsilon)$, we have that

$$
\begin{aligned}
SR(Q, s, s'; y) \geq \exp(-\varepsilon) &\implies \frac{SR(Q, s, s'; y)}{SR(Q_0, s, s'; y)} \geq \exp(-\varepsilon) \\
&\implies \min \left\{ \frac{SR(Q, s, s'; y)}{SR(Q_0, s, s'; y)}, \frac{SR(Q_0, s, s'; y)}{SR(Q, s, s'; y)} \right\} \geq \exp(-\varepsilon).
\end{aligned}
$$

Furthermore, if $SR(Q, s, s'; y) > 1$, then necessarily $SR(Q, s', s; y) \leq 1$. However, we also have

$$\min \left\{ \frac{SR(Q, s, s'; y)}{SR(Q_0, s, s'; y)}, \frac{SR(Q_0, s, s'; y)}{SR(Q, s, s'; y)} \right\} = \min \left\{ \frac{SR(Q_0, s', s; y)}{SR(Q, s', s; y)}, \frac{SR(Q, s', s; y)}{SR(Q_0, s', s; y)} \right\}.$$

Thus for all $s, s' \in \mathcal{S}$

$$\min \left\{ \frac{SR(Q, s, s'; y)}{SR(Q_0, s, s'; y)}, \frac{SR(Q_0, s, s'; y)}{SR(Q, s, s'; y)} \right\} \geq \exp(-\varepsilon) \iff \max \left\{ \frac{SR(Q, s, s'; y)}{SR(Q_0, s, s'; y)}, \frac{SR(Q_0, s, s'; y)}{SR(Q, s, s'; y)} \right\} \leq \exp(\varepsilon).$$

That is $Q \in \mathcal{M}_{\varepsilon, Q_0}$.

□

I. Analyzing KL Drop Bound

In the following section, we analysis the bound given in Theorem 3.5. In particular, we examine the non-negativity of the bound and compare it against Husain et al. (2020, Theorem 5).

Lemma I.1. *The following holds for* $\Gamma(\cdot, \cdot)$:

$$\Gamma(C^2, \gamma) = \log\left(\frac{4}{(2 + \gamma)\exp(-C)) + (2 - \gamma)\exp(C)}\right);$$

$$\Gamma(0, \gamma) = \log\left(\frac{4}{(3 - \gamma)\exp(-C)) + (1 - \gamma)\exp(C)}\right);$$

$$\Gamma(v, 1) = \log\left(\frac{4C^2}{v(\exp(C) - \exp(-C) + 4C^2\exp(-C)}\right);$$

$$\Gamma(0, 1) = C.$$

Proof: Follows directly from the definition of Γ, as per Lemma D.2. $\square$

To compare against Husain et al. (2020, Theorem 5), we consider the maximum variance which is required to make $\Gamma(\cdot, \cdot) > 0$.

Theorem I.2. *For the function* $\Gamma(v, \gamma)$*, we have that*

$$\Gamma(v, \gamma) > 0 \iff \gamma > \frac{v}{C^2} + 1 - \frac{4}{\exp(C) + 1}; \tag{23}$$

$$\Gamma(v, \gamma) > 0 \iff v < C^2 \cdot \left(\gamma + \frac{4}{\exp(C) + 1} - 1\right). \tag{24}$$

Proof: We consider the conditions for $\Gamma(v, \gamma) > 0$. This occurs when:

$$\Gamma(v, \gamma) > 0$$
$$\iff \log\left((A \cdot v - C^2 A \cdot \gamma + K)^{-1}\right) > 0$$
$$\iff A \cdot v - C^2 A \cdot \gamma + K < 1$$
$$\iff \gamma > \frac{Av + K - 1}{C^2 A}$$
$$\iff \gamma > \frac{v}{C^2} + \frac{K - 1}{C^2 A}.$$

Through computation, one can note that $(K - 1)/(C^2 A) = 1 - 4/(\exp(C) + 1)$. Thus we have the following:

$$\Gamma(v, \gamma) > 0 \iff \gamma > \frac{v}{C^2} + 1 - \frac{4}{\exp(C) + 1};$$

$$\Gamma(v, \gamma) > 0 \iff v < C^2 \cdot \left(\gamma + \frac{4}{\exp(C) + 1} - 1\right).$$

As required. $\square$

Corollary I.3 (Comparing to Husain et al. (2020)). *Let* $C = \log 2$. *If* $\mathbb{V}_{Q_{t-1}}[c_t] < C^2 \cdot \frac{2}{3}$, *when Husain et al. (2020, Theorem 5)'s KL drop is positive, so is ours.*

Furthermore, lower $\mathbb{V}_{Q_{t-1}}[c_t]$ *values allows for positive KL drops even when Husain et al. (2020, Theorem 5) has a non-positive drop.*

Proof: To guarantee a positive drop, Husain et al. (2020, Theorem 5) requires the WLA to be within a "high boosting regime". This is equivalent to having $\gamma_Q^t > 1/3$.

It follows from Theorem I.2 that we get a positive drop when:

$$\gamma > \frac{2}{3} + 1 - \frac{4}{3} = \frac{1}{3}. \tag{25}$$

As required.

The last part of the corollary follows from noting that Eq. (24) is linearly increasing function of v. Thus lower v will relax the constraint in Eq. (25). $\qquad\blacksquare$

J. General Fair Mollifiers

The relative mollifier construction present in Definition 4.1 was not the original definition of a mollifier, but only a convenient method for construction. Originally in Husain et al. (2020), elements of private mollifiers satisfy a pairwise privacy constraint. We introduce the *fair mollifier* which is equivalent to this original definition.

Definition J.1. Let $\mathcal{M} \subset \mathcal{D}(\mathcal{Z})$ and $\varepsilon > 0$. $\mathcal{M}$ is an ε-**fair mollifier** iff[4] $\forall Q, Q' \in \mathcal{M}$,

$$\mathrm{SR}(Q, s, s'; y) \le e^{\varepsilon} \cdot \mathrm{SR}(Q', s, s'; y), \quad \forall s, s' \in \mathcal{S}. \tag{26}$$

Definition J.1 is inspired by the local notion of differential privacy (Kasiviswanathan et al., 2011). Likewise to the main-text, all statistical rate definitions can be interchanged with representation rate.

Lemma J.2. *Suppose that $\mathcal{M}$ is an ε-fair mollifier. If there exists a $\mathrm{F} \in \mathcal{M}$ with $\mathrm{SR}(\mathrm{F}) = \mathrm{SR}_f$, then all $Q \in \mathcal{M}$ have $\mathrm{SR}(Q) \ge \mathrm{SR}_f \cdot \exp(-\varepsilon)$.*

Proof: As $\mathcal{M}$ is a ε-fair mollifier, the following constraint holds for all $Q \in \mathcal{M}$ and for all $s, s' \in \mathcal{S}$ with the distribution $\mathrm{F} \in \mathcal{M}$, where $\mathrm{SR}(\mathrm{F}) = \mathrm{SR}_f$:

$$\mathrm{SR}(\mathrm{F}, s, s'; y) \le \exp(\varepsilon) \cdot \mathrm{SR}(Q, s, s'; y) \implies \mathrm{SR}_0 \le \exp(\varepsilon) \cdot \mathrm{SR}(Q, s, s'; y)$$
$$\implies \mathrm{SR}_0 \cdot \exp(-\varepsilon) \le \mathrm{SR}(Q, s, s'; y).$$

Thus all $Q \in \mathcal{M}$ has statistical rate $\mathrm{SR}(Q) \ge \mathrm{SR}_f \cdot \exp(-\varepsilon)$. $\qquad\blacksquare$

It should be noted that Lemma J.2 is not tight, a mollifier might have a strong fairness guarantee.

Lemma J.3. $\mathcal{M}_{\varepsilon, Q_0}$ *is a 2ε-mollifier.*

Proof: Follows directly from Eq. (11) applies to two distributions. $\qquad\blacksquare$

K. Improvement in the Low / Sparse Data Regime via Prior Mixing

When utilizing the FBDE algorithm, picking a good initial distribution can be pivotal to ensure that our final distribution is not only fair, but also close to the true data distribution. However, when we only have access to a small volume of data; or when a dataset has a large number of features causing sparsity over input combinations, it may be difficult to find good approximations of the input distribution. These scenarios can be further problematic when we utilize an empirical distribution to create our initial distribution as per Eqs. (3) and (4): inputs in the training set may not appear in the test set. In this case, the final distribution will have zero support for inputs which occur in the test set.

To remedy this issue, we introduce the *mixed* initial distribution. The idea is to take a mixture distribution between the initial distribution introduced Eqs. (3) and (4) with a *prior distribution*. That is, we an initial distribution as per,

$$Q_{\mathrm{INIT}} \leftarrow (1 - \alpha) \cdot Q_{\mathrm{INIT}} + \alpha \cdot Q_{\mathrm{PRIOR}}, \tag{27}$$

where $\alpha \in [0, 1]$ and $Q_{\mathrm{PRIOR}} \in \mathcal{D}(\mathcal{Z})$ is a 'prior' distribution.

By taking $\alpha \ne 1$ and a Q_{PRIOR} with non-zero support over all of $\mathcal{Z}$, we can ensure that the final boosted distribution also has non-zero support regardless of the initial distribution.

The one caveat to utilizing this mixture is that, of course, we still want the final Q_{INIT} to satisfy the initial fairness constraints, *i.e.*, the dependence on SR_0 and RR_0 in Theorems 3.2 and 3.3.

Nevertheless, as long as the final Q_{INIT} satisfies the initial fairness constraints, the mixing of distribution provides a convenient mechanism for adding prior knowledge into FBDE. Indeed, beyond dealing with miss-matching distribution support issues, Q_{PRIOR} can be used to add any user prior knowledge to FBDE.

We present experiment where mixing priors are not used for DUTCH and GERMAN in Appendix P.

---

[4]Absolute continuity also required for all $\{P\} \cup \mathcal{M}$ w.r.t. all $\mathcal{M}$.

L. Additional Dataset Descriptions

We present the following additional dataset descriptions for COMPAS and ADULT not included in the main-text. We note that COMPAS and ADULT are standard public benchmark datasets; and AIF360 uses a Apache License 2.0.

(a) **COMPAS (Angwin et al., 2016)** The pre-processed dataset contains data where each defendant has the sex, race, age, number of priors, charge degree, and recidivism within two years. These attributes are given as binary features, where counts and continuous values are discretized into bins. The resulting dataset has a domain size of 144 and with 5,278 data points. We consider sensitive attributes of sex (only in appendix) and race.

(b) **ADULT (Dua & Karra Taniskidou, 2017)** The pre-processed dataset contains data where for each person the dataset contains the race, sex, age, years of educations, and binary label whether they earn more than 50K a year. The counts and continuous values are discretized similarly to COMPAS. The pre-processed dataset has a domain size of 504 and with 48,842 data points. We consider sex and race (only in appendix) as the sensitive attribute.

For the MINNEAPOLIS dataset, we also include additional dataset descriptions. We note that it uses a CC0 1.0 License.

(c) **MINNEAPOLIS** The dataset consists of stop data for the Minneapolis Police Department. We only consider the features consisting of latitude, longitude, race, and 'search' features, where the last feature is a binary variable that reports an individual was searched at a stop. We remove any rows with missing values. We remove outliers (those consisting of longitude values less than -13). Furthermore, we binarize the race features to 'Black' vs all other categories. In total, the resulting dataset has 121,965 data points.

In addition the datasets examined in the main-text, we also include the DUTCH and GERMAN standard datasets. The primary results for these extra datasets are presented in Appendix O. We also include an additional discussion in Appendix K, where we use a "trick" to improve the performance of mollifiers in GERMAN. We further note that these are standard public benchmark datasets.

(d) **DUTCH (Van der Laan, 2000)** The Dutch Census dataset is from a 2001 Netherlands census, where data represents aggregated groups of people. The prediction task is to determine whether an individual has a prestigious job or not. The dataset consists of 60,420 examples with 12 different attributes (many with large number of categorical features). The pre-processed dataset has a domain size of 5,971,968, thus, the dataset is extremely sparse. We consider sex and age as sensitive attributes.

(e) **GERMAN (Kamiran & Calders, 2009)** The German Credit dataset consists of individual bank holders, with the prediction task being to determine whether or not the grant credit to someone. We utilize the pre-processed version provided by AIF360. Notably, the dataset is small with only 1,000 samples with a domain size of 216.

M. Extended Discrete Experiments

We present additional experiments and figures corresponding to the discrete datasets of COMPAS and ADULT. In particular, we include all combinations of 'sex' and 'race' sensitive attributes settings for the datasets. An extended Table 2 is presented in Table I. Furthermore, in Fig. I we plot the complete set of weak learner accuracy vs KL over boosting iterations plots that was given by Fig. 2 (left). We further include similar plots for both representation rate and statistical rate in Figs. II and III, respectively. We also present plots corresponding to SR vs SR_c in Fig. IV, which suggests that the classifier statistical rate SR_c tends to follow a similar trend to that of the data statistical rate SR.

M.I. Runtime Comparison

In addition, we report the runtime of each approach in Table II. We notice that in general, M-E-1.0 is the quickest approach (by a significant margin) taking only a few seconds on both datasets. By strictly looking at the table, TABFAIR is the next quickest; and lastly FBDE and FAIRK (with FAIRK scaling significantly worse on the larger dataset). With this strict analysis, one may come to a conclusion that FBDE has a large computational trade-off when compared to M-E-1.0 (and even TABFAIR). However, if one examines Fig. I they will notice that the change in the distribution is quite small after $T = 8$ iterations. In particular, the KL will still be superior than M-E-1.0 if we early stopped at $T = 8$. Furthermore, we would have a fairer distribution as well (we would have moved from Q_0 less). As the boosting algorithms time complexity is linear w.r.t. T (*i.e.*, the number of WLs learned) we could reduce the boosting iterations / runtime by 4x; whilst still maintaining superior KL and also having strong fairness. Indeed, the fairness and KL can be verified by examining Figs. II to IV.

Although, this would still result in a runtime many times larger than M-E-1.0, we believe that the possible interpretability / human-in-the-loop capabilities of FBDE are out-weight the higher runtime — we can interpret this comparison as a trade-off between lower runtime vs interpretability properties.

M.II. A Note about MAXENT

When using MAXENT, we utilize take the following parameters: statistical rate setting $\rho = 1$, prior interpolation parameter $C = 0.5$, and marginal vector constrained by a weighted mean (from Celis et al. (2020, Algorithm 1)).

M.III. A Note About TABFAIR

We note the extremely poor performance of TABFAIR in the ADULT dataset with 'race' sensitive attribute. We believe that the reason for this performance is the selection of suboptimal (hyper)parameters. Unfortunately, the model parameters presented in Rajabi & Garibay (2022, Table A1 in Appendix) are not applicable to our experiments as they utilize a different pre-processing of the COMPAS and ADULT datasets. For COMPAS, we use a total of 128 epochs with 40 of those being 'fair epochs' (Rajabi & Garibay (2022)'s fairness training phase); in addition the batch sizes are specified to be 128 and the fairness regularizer parameter $\lambda_f = 2.5$. For ADULT, we have 128 epochs with 30 of those being 'fair epochs'; in addition the batch sizes are specified to be 128 and the fairness regularizer parameter $\lambda_f = 2$. Although we did not make an effort to fine-tune these (hyper)parameters for these extra settings in the Appendix, the lack of performance transfer when only the sensitive attribute is switched highlights a significant downside when compared to FBDE and other optimization approaches.

M.IV. Full Weak Learner Plot

We present the full plot of Fig. 1 in Fig. V. We leave the careful examination of such weak learners to future work and rather present these plots as a proof of concept.

N. Additional Continuous Experiments

We present additional experiments and figures for the MINNEAPOLIS dataset. We note that all computation was done without a GPU. In addition to M-E-1.0 explored in the main-text, we present results for M-E-0.9, M-R-1.0, and M-R-0.9 similar to the discrete dataset experiments. In Table I, we present a summary of evaluations used in the discrete experiments. In Fig. I we present RR vs KL, SR vs KL, and SR vs SR$_c$ plots (one each row), as per Fig. 2.

O. Extra Discrete Experiments

We present the exact same plots and tables for DUTCH and GERMAN as we did for DUTCH and GERMAN. All settings for FBDE are identical except for the initial distribution utilized. Here we incorporate a "prior" into the initial distribution, similar to Celis et al. (2020). An extended discussion of the trick we use is presented in Appendix K. We plot weak learner accuracy vs KL, SR vs KL, RR vs KL, and SR vs SR$_c$ over boosting iterations in Figs. I to IV, respectively.

We would like to note one particularly poor showing in Table I: M-R-0.9 on DUTCH with $\mathcal{S} = $ sex. In this case, M-R-0.9 has a critical failure in SR$_C$. However, one should note that in Fig. IV, there actually include boosting iterations which M-R-0.9 significantly performs better in SR$_C$. Despite this, we should still note that the overall average SR$_C$ is still quite poor. This is similar to M-E-0.9 in this dataset setting. This indicates that having a weaker (in terms of fairness) initial distribution might not work for sparse datasets. Nevertheless, one wants to make sure that FBDE is performing correctly in these downstream fairness considerations by examining performance per boosting iterations.

P. DUTCH and GERMAN Without Mixing Priors

Experimentally, in Appendix O we utilize a initial distribution mixing with Q_{PRIOR} as a uniform distribution and $\alpha = 1/2$ (as per Appendix K). In this case, our initial distribution is very similar to the prior distribution utilized in Celis et al. (2020, Section 3.1 "Prior distributions"). This seems to explain why MAXENT performs so well in KL for DUTCH and GERMAN datasets, where TABFAIR does not.

Indeed, Table I presents a repeat of those experiments without prior mixing. In this case, the KL divergence is significantly

worse. Despite this, it should also be noted that the empirical distribution's KL (comparing the training set vs the test set) is also quite large.

Q. In-Processing Experiments

To evaluate FBDE learned predictions against in-processing algorithms, we consider the following in-processing approaches:

- FAIRCONS: from Zafar et al. (2019);

- REDUCT: from Agarwal et al. (2018);

- FAIRGLM: from Do et al. (2022).

We utilize Do et al. (2022)'s implementation of these in-processing algorithms, provided in `www.github.com/hyungrok-do/fair-glm-cvx`. Table I provides a summary of these comparisons.

Table I. FBDE ($T = 32$) and baselines evaluated for COMPAS and ADULT datasets. The table reports the mean and s.t.d.

			DATA	M-E-1.0	M-E-0.9	M-R-1.0	M-R-0.9	MAXENT	TABFAIR	FAIRK
COMPAS (δ = SEX)	DATA	RR	.443±.004	.916±.011	.944±.007	.852±.010	.907±.007	.987±.007	.231±.044	–
		SR	.728±.044	.993±.003	.904±.008	.986±.010	.898±.008	.989±.011	.823±.138	–
		KL	–	.288±.020	.295±.022	.268±.020	.283±.021	.335±.021	3.14±.831	–
	PRED	SR$_C$	.726±.025	.952±.006	.874±.009	.938±.019	.874±.016	.966±.012	.844±.113	–
		EO	.764±.031	.966±.021	.905±.027	.963±.027	.907±.030	.978±.015	.797±.098	–
		ACC	.660±.004	.651±.008	.654±.005	.657±.009	.655±.008	.650±.007	.595±.036	–
	CLUS	δPR	.122±.014	.122±.014	.122±.014	.122±.014	.119±.017	.119±.017	.106±.032	.143±.019
		δSR	.380±.123	.380±.123	.380±.123	.380±.123	.438±.097	.385±.122	.364±.065	.295±.089
		DIST	.064±.000	.078±.000	.078±.000	.078±.000	.078±.000	.078±.000	.078±.000	.069±.001
COMPAS (δ = RACE)	DATA	RR	.662±.007	.966±.008	.977±.008	.944±.010	.964±.009	.992±.004	.632±.052	–
		SR	.747±.013	.988±.006	.899±.011	.978±.011	.896±.011	.992±.006	.727±.100	–
		KL	–	.135±.020	.129±.020	.132±.020	.127±.020	.164±.018	2.71±.735	–
	PRED	SR$_C$	.747±.020	.959±.025	.875±.025	.945±.027	.872±.044	.939±.018	.802±.074	–
		EO	.781±.034	.960±.026	.900±.041	.950±.028	.895±.039	.944±.020	.815±.079	–
		ACC	.660±.004	.641±.014	.653±.012	.642±.010	.656±.012	.648±.015	.591±.043	–
	CLUS	δPR	.334±.019	.332±.022	.334±.019	.334±.019	.319±.036	.333±.020	.286±.027	.259±.013
		δSR	.288±.055	.235±.082	.284±.055	.284±.055	.256±.084	.288±.059	.268±.063	.284±.036
		DIST	.065±.000	.078±.000	.078±.000	.078±.000	.078±.000	.078±.000	.078±.001	.069±.001
ADULT (δ = SEX)	DATA	RR	.496±.002	.958±.003	.979±.003	.919±.004	.957±.003	.995±.002	.516±.023	–
		SR	.360±.005	.961±.005	.883±.006	.944±.006	.865±.006	.979±.004	.862±.101	–
		KL	–	.122±.006	.119±.006	.113±.006	.114±.006	.182±.005	1.68±.538	–
	PRED	SR$_C$	.360±.003	.818±.010	.766±.010	.793±.011	.753±.008	.919±.011	.823±.118	–
		EO	.471±.008	.959±.016	.908±.018	.935±.016	.895±.016	.981±.010	.867±.105	–
		ACC	.803±.003	.785±.002	.788±.002	.787±.002	.788±.002	.773±.005	.781±.006	–
	CLUS	δPR	.125±.067	.093±.019	.101±.025	.116±.027	.111±.033	.130±.061	.137±.038	.117±.039
		δSR	.426±.232	.252±.082	.302±.076	.190±.061	.254±.086	.468±.237	.437±.179	.296±.151
		DIST	.034±.000	.037±.000	.037±.000	.037±.000	.037±.000	.037±.000	.037±.000	.035±.000
ADULT (δ = RACE)	DATA	RR	.170±.001	.914±.002	.955±.003	.837±.002	.910±.002	.998±.003	.089±.019	–
		SR	.601±.011	.985±.006	.892±.006	.974±.007	.888±.006	.986±.006	.284±.032	–
		KL	–	.292±.001	.305±.001	.264±.001	.289±.001	.382±.005	1.98±.173	–
	PRED	SR$_C$	.600±.031	.867±.035	.808±.031	.860±.034	.803±.035	.935±.019	.480±.114	–
		EO	.787±.053	.933±.019	.960±.037	.935±.017	.957±.034	.929±.038	.588±.103	–
		ACC	.803±.003	.800±.002	.801±.003	.800±.002	.801±.003	.794±.004	.745±.019	–
	CLUS	δPR	.042±.019	.051±.038	.061±.032	.051±.029	.043±.012	.047±.021	.071±.039	.052±.044
		δSR	.313±.096	.422±.111	.366±.212	.303±.102	.327±.115	.290±.096	.359±.225	.400±.218
		DIST	.034±.000	.037±.000	.037±.000	.037±.000	.037±.000	.037±.000	.037±.000	.035±.000

Table II. Runtime (s) of FBDE ($T = 32$) and baselines evaluated for COMPAS and ADULT datasets. The table reports the mean and s.t.d.

	M-E-1.0	M-E-0.9	M-R-1.0	M-R-0.9	MAXENT	TABFAIR	FAIRK
COMPAS (SEX)	88.467 ± 0.193	90.362 ± 0.296	89.070 ± 0.443	89.123 ± 0.210	0.740 ± 1.127	25.234 ± 0.131	62.365 ± 0.251
COMPAS (RACE)	88.026 ± 0.122	88.568 ± 0.220	89.621 ± 0.106	87.935 ± 0.233	0.629 ± 0.894	25.091 ± 0.310	61.705 ± 0.386
ADULT (SEX)	311.700 ± 2.941	308.990 ± 0.904	310.606 ± 1.777	309.986 ± 1.433	2.068 ± 3.039	169.611 ± 3.659	5446.451 ± 54.161
ADULT (SEX)	313.745 ± 1.462	311.175 ± 1.319	313.996 ± 0.755	315.955 ± 6.093	1.735 ± 2.094	171.614 ± 3.297	5149.282 ± 43.155

Figure I. All WL accuracy *vs* KL over boosting iterations plots for COMPAS and ADULT. Horizontal line depicts the KL(P, Q_8) value.

Figure II. All RR vs KL over boosting iterations plots for COMPAS and ADULT. Horizontal line depicts the $T = 8$ RR value.
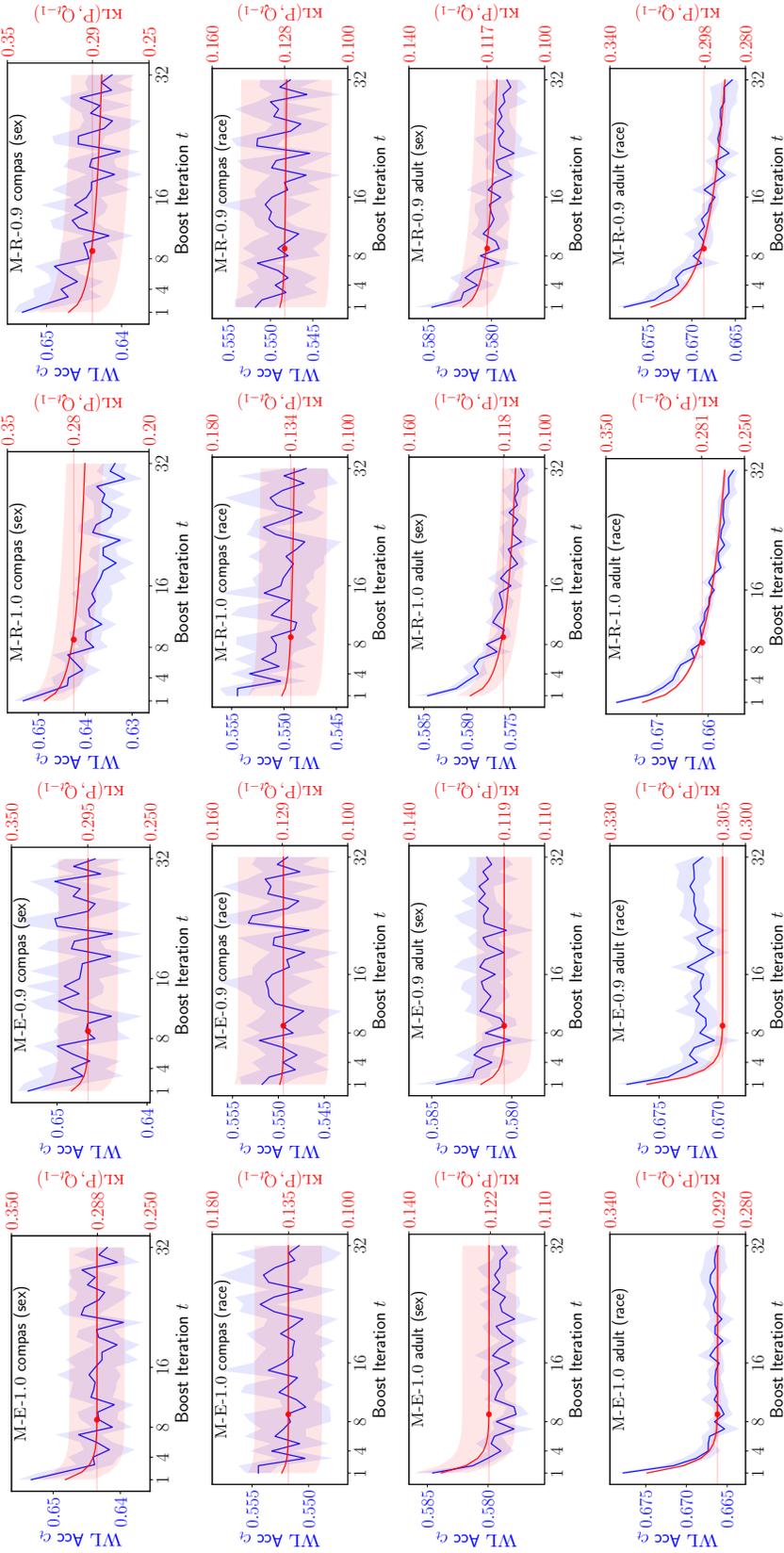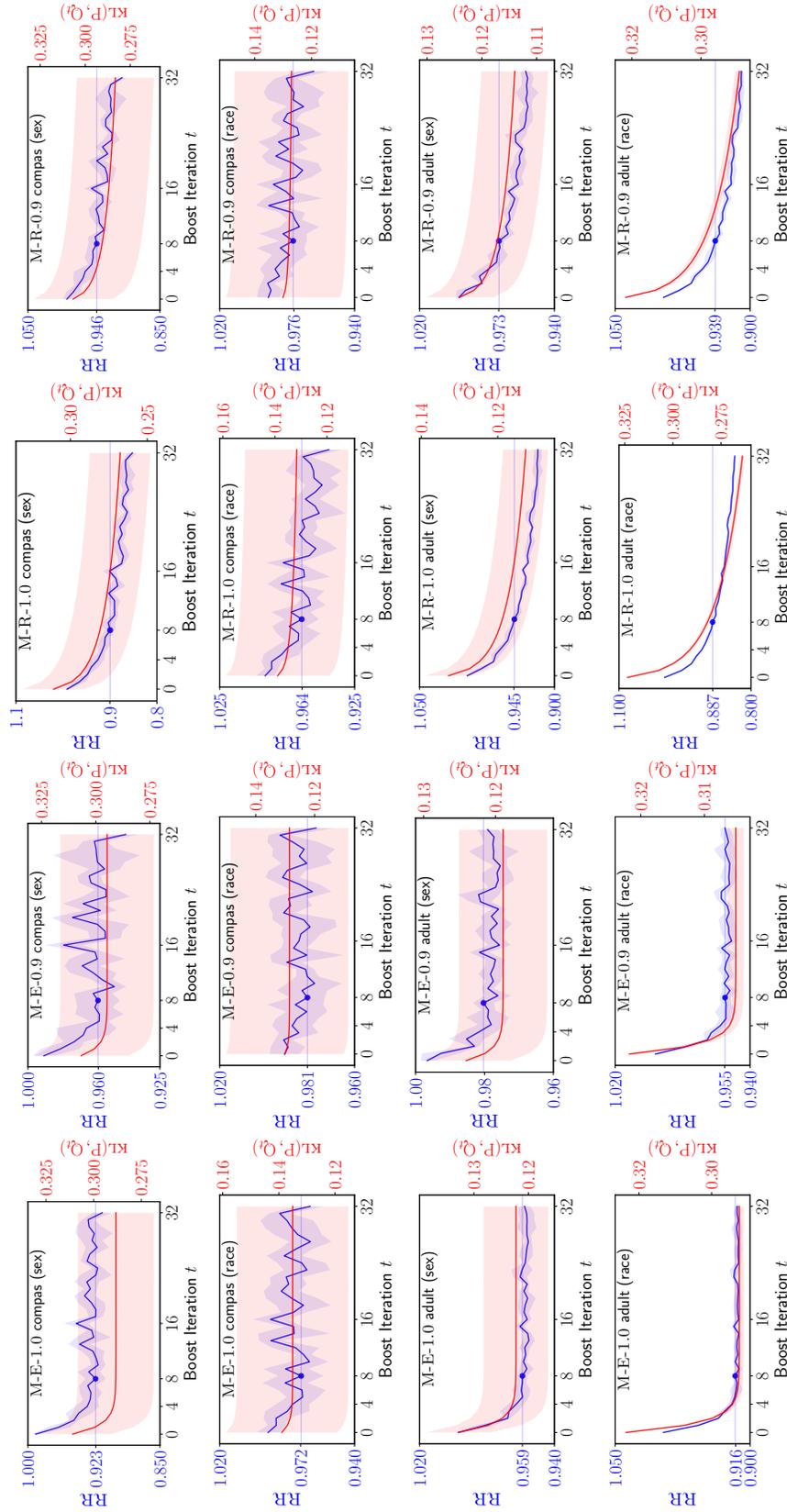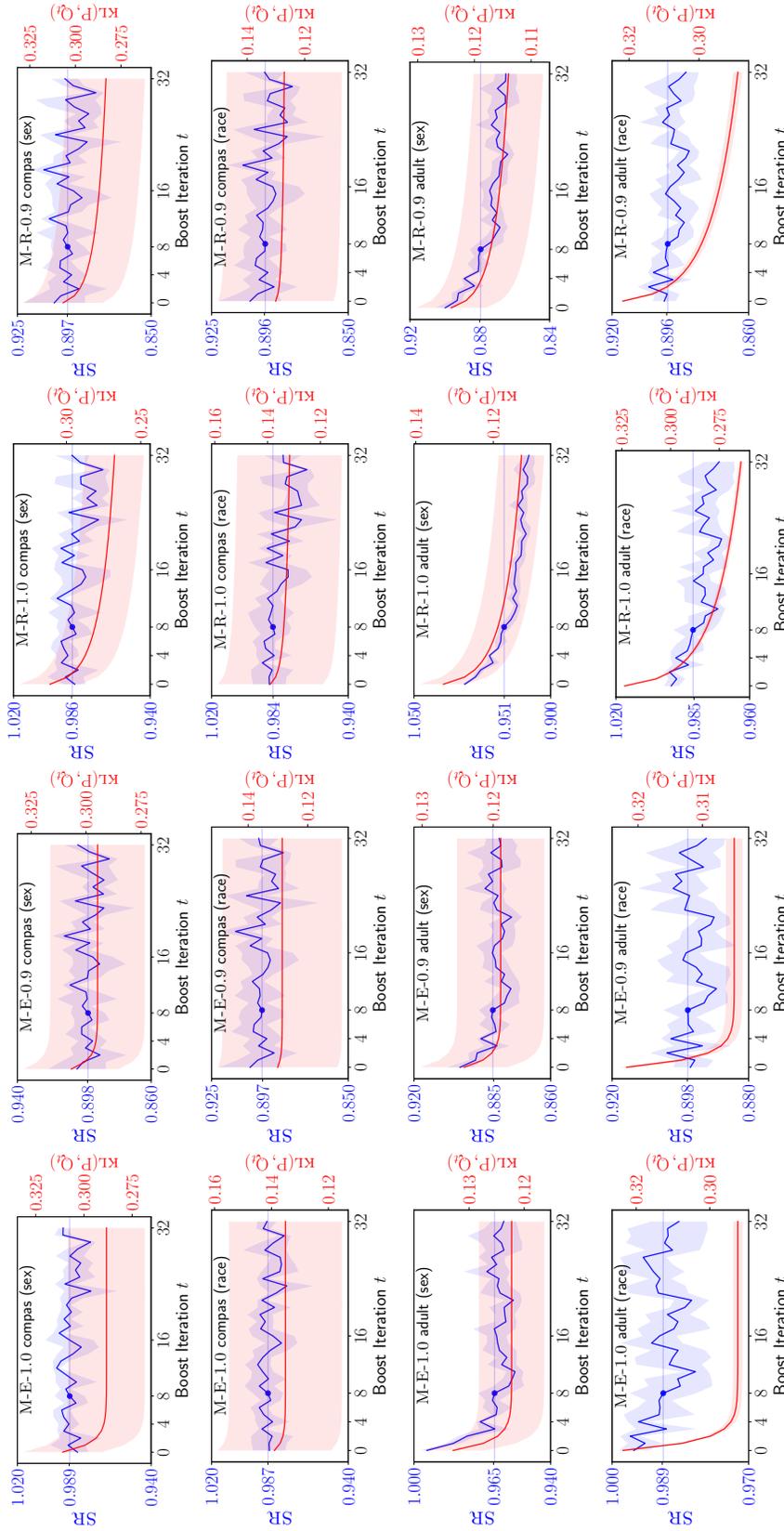
Figure III. All SR vs KL over boosting iterations plots for COMPAS and ADULT. Horizontal line depicts the $T = 8$ SR value.
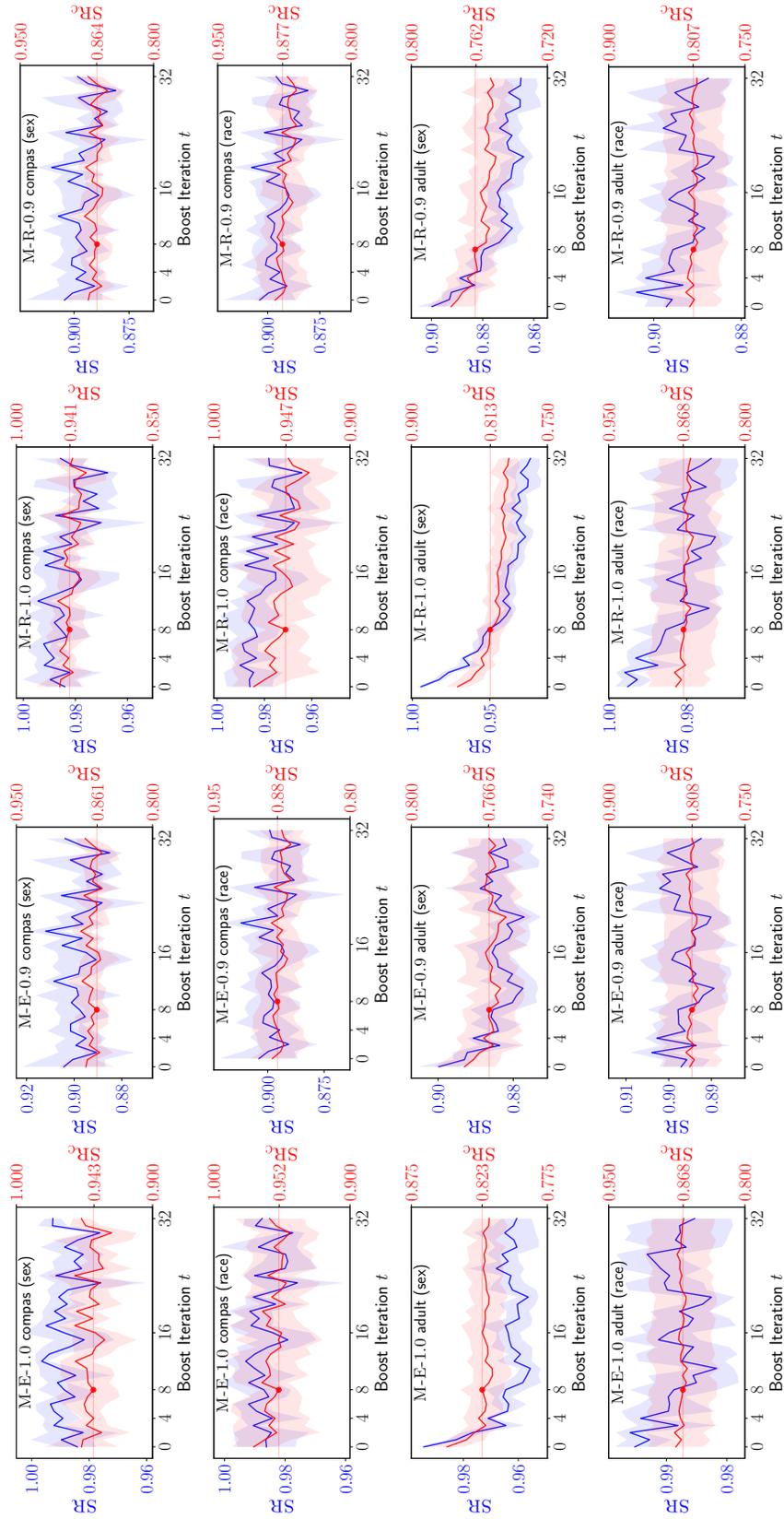
Figure IV. All SR *vs* Classification SR over boosting iterations plots for COMPAS and ADULT. Horizontal line depicts the $T = 8$ SR$_c$ value.
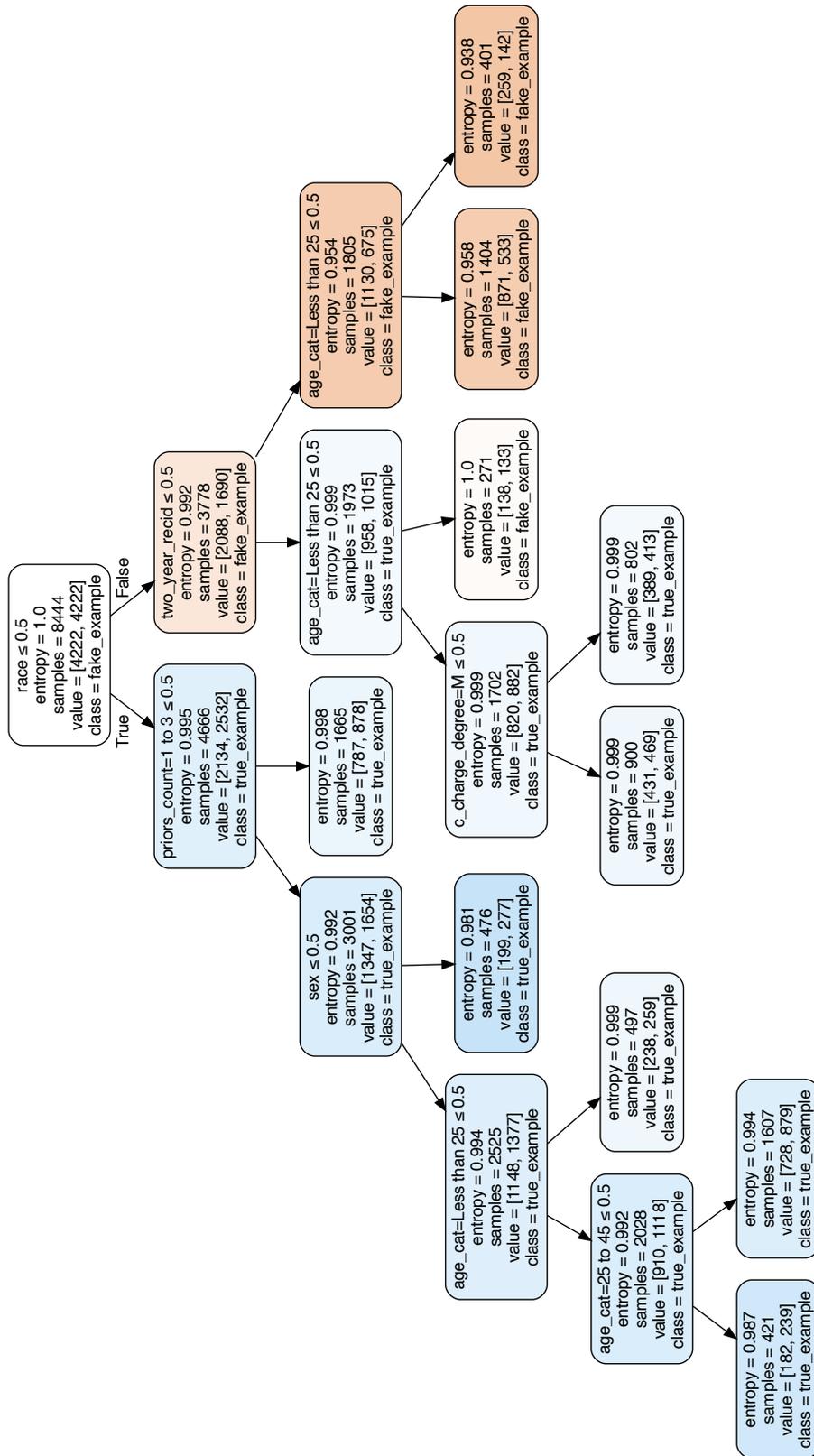
Figure V. Full WL example of Fig. 1: M-E-1.0 with COMPAS (S = race).

Table I. FBDE ($T = 16$) evaluated for the MINNEAPOLIS dataset. The table reports the mean and s.t.d.

			DATA	M-E-1.0	M-E-0.9	M-R-1.0	M-R-0.9
MINNEAPOLIS ($\mathcal{S}$ = RACE)	DATA	RR	$.183 \pm .001$	$.846 \pm .004$	$.918 \pm .000$	$.756 \pm .003$	$.860 \pm .004$
		SR	$.684 \pm .005$	$.995 \pm .003$	$.901 \pm .000$	$.991 \pm .004$	$.902 \pm .004$
		KL	$-$	$.218 \pm .005$	$.236 \pm .000$	$.186 \pm .004$	$.219 \pm .005$
	PRED	SR_C	$.685 \pm .012$	$.930 \pm .018$	$.754 \pm .000$	$.929 \pm .038$	$.835 \pm .038$
		EO	$.778 \pm .028$	$.974 \pm .018$	$.856 \pm .000$	$.943 \pm .031$	$.929 \pm .047$
		ACC	$.649 \pm .004$	$.594 \pm .016$	$.625 \pm .000$	$.600 \pm .015$	$.601 \pm .015$
	CLUS	δPR	$.134 \pm .013$	$.137 \pm .011$	$.155 \pm .000$	$.138 \pm .012$	$.153 \pm .023$
		δSR	$.282 \pm .072$	$.235 \pm .033$	$.310 \pm .000$	$.443 \pm .057$	$.262 \pm .044$
		DIST	$.013 \pm .000$	$.016 \pm .000$	$.016 \pm .000$	$.016 \pm .000$	$.016 \pm .000$
		TIME(S)	$-$	1540.731 ± 2.615	$1542.730 \pm .000$	1369.129 ± 55.183	1410.402 ± 45.461
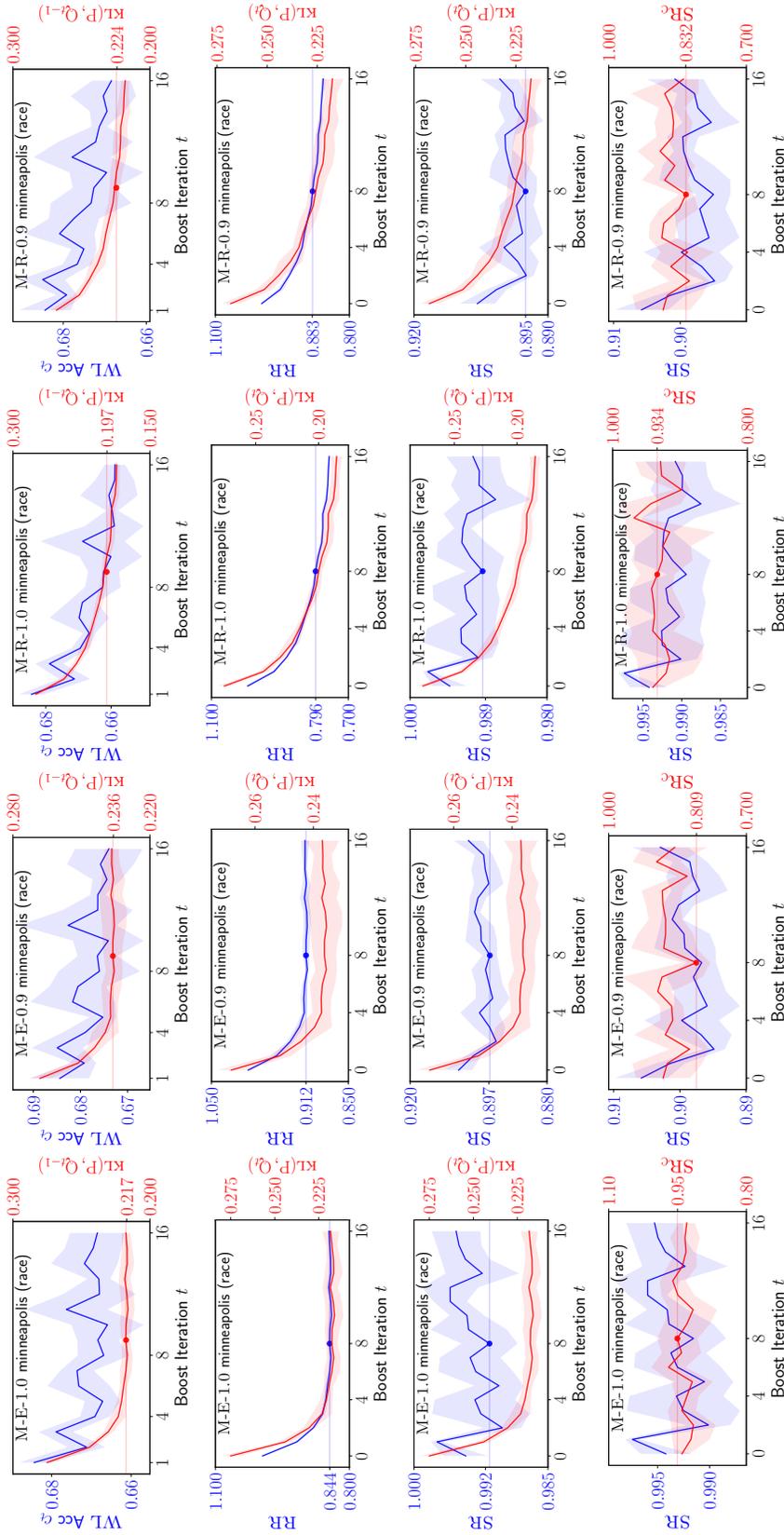
Figure I. All boosting iterations plots for MINNEAPOLIS. Horizontal line depicts the $\mathrm{KL}(\mathrm{P}, \mathrm{Q}_8)$ value.

Table I. FBDE ($T = 32$) and baselines evaluated for DUTCH and GERMAN datasets. The table reports the mean and s.t.d.

			DATA	M-E-1.0	M-E-0.9	M-R-1.0	M-R-0.9	MAXENT	TABFAIR	FAIRK
DUTCH (S = SEX)	DATA	RR	.996±.002	.987±.001	.993±.000	.980±.001	.992±.000	.997±.002	.411±.260	–
		SR	.523±.002	.988±.001	.894±.000	.971±.001	.887±.000	.961±.004	.344±.338	–
		KL	–	.939±.004	.916±.004	.920±.004	.906±.004	1.637±.017	4.367±.668	–
	PRED	SR$_c$	.464±.022	.696±.193	.518±.255	.628±.139	.275±.068	.496±.020	.628±.223	–
		EO	.890±.006	.973±.005	.986±.008	.980±.007	.987±.007	.967±.007	.585±.203	–
		ACC	.827±.001	.763±.003	.778±.004	.770±.003	.781±.001	.814±.003	.667±.068	–
	CLUS	δPR	.080±.029	.070±.021	.136±.053	.113±.017	.069±.023	.075±.006	.094±.056	.203±.103
		δSR	.341±.126	.273±.066	.254±.053	.449±.052	.254±.058	.204±.038	.291±.121	.428±.195
		DIST	.051±.000	.056±.000	.056±.000	.056±.000	.056±.000	.056±.000	.056±.000	.053±.000
DUTCH (S = AGE)	DATA	RR	.621±.001	.980±.001	.990±.000	.964±.001	.980±.000	.996±.004	.227±.026	–
		SR	.610±.003	.993±.001	.895±.000	.984±.001	.893±.000	.963±.005	.430±.011	–
		KL	–	.935±.004	.929±.005	.916±.004	.919±.005	1.656±.015	4.132±.165	–
	PRED	SR$_c$	.695±.030	.816±.094	.676±.257	.747±.084	.709±.180	.722±.023	.271±.046	–
		EO	.892±.009	.993±.005	.960±.007	.987±.007	.961±.009	.991±.003	.742±.018	–
		ACC	.827±.001	.781±.004	.788±.004	.785±.005	.790±.005	.820±.003	.661±.004	–
	CLUS	δPR	.699±.080	.454±.212	.660±.096	.675±.070	.718±.056	.440±.175	.436±.190	.373±.051
		δSR	.525±.079	.421±.138	.489±.103	.492±.098	.542±.036	.340±.058	.376±.068	.443±.126
		DIST	.052±.000	.056±.000	.056±.000	.056±.000	.056±.000	.056±.000	.056±.000	.053±.000
GERMAN (S = SEX)	DATA	RR	.449±.014	.934±.033	.960±.027	.903±.038	.946±.037	.980±.013	.575±.154	–
		SR	.897±.018	.982±.018	.917±.044	.983±.017	.916±.044	.985±.011	.985±.010	–
		KL	–	.606±.027	.612±.026	.588±.026	.602±.025	.531±.035	9.02±1.44	–
	PRED	SR$_c$	.903±.055	.971±.014	.887±.049	.970±.019	.886±.050	.951±.030	.837±.120	–
		EO	.928±.055	.980±.021	.909±.046	.969±.023	.907±.051	.954±.031	.819±.120	–
		ACC	.690±.022	.712±.011	.705±.011	.714±.019	.706±.007	.671±.044	.509±.170	–
	CLUS	δPR	.234±.103	.229±.144	.244±.104	.197±.082	.210±.055	.180±.075	–	.168±.070
		δSR	.268±.100	.359±.310	.327±.094	.422±.265	.393±.298	.441±.144	.396±.315	.376±.299
		DIST	.158±.001	.197±.001	.197±.001	.197±.000	.197±.001	.195±.001	.196±.011	.157±.001
GERMAN (S = AGE)	DATA	RR	.235±.008	.911±.037	.951±.036	.847±.033	.915±.033	.971±.011	.419±.083	–
		SR	.794±.029	.982±.020	.911±.019	.982±.019	.908±.022	.980±.013	.968±.014	–
		KL	–	.717±.035	.725±.034	.687±.034	.708±.034	.691±.043	9.812±1.357	–
	PRED	SR$_c$	.802±.067	.947±.027	.881±.045	.950±.044	.890±.048	.969±.016	.780±.212	–
		EO	.823±.078	.940±.028	.894±.059	.943±.027	.914±.058	.951±.034	.744±.193	–
		ACC	.690±.022	.700±.016	.693±.020	.697±.013	.707±.011	.671±.023	.486±.128	–
	CLUS	δPR	.162±.039	.183±.042	.210±.080	.183±.097	.174±.052	.200±.060	.228±.025	.208±.102
		δSR	.387±.333	.494±.231	.784±.250	.729±.270	.665±.441	.601±.226	.499±.298	.406±.236
		DIST	.161±.001	.198±.001	.198±.000	.198±.001	.198±.001	.196±.001	.198±.011	.158±.001

Figure I. All WL accuracy *vs* KL over boosting iterations plots for DUTCH and GERMAN. horizontal line depicts the $\mathrm{KL}(\mathrm{p}, \mathrm{q}_8)$ value.

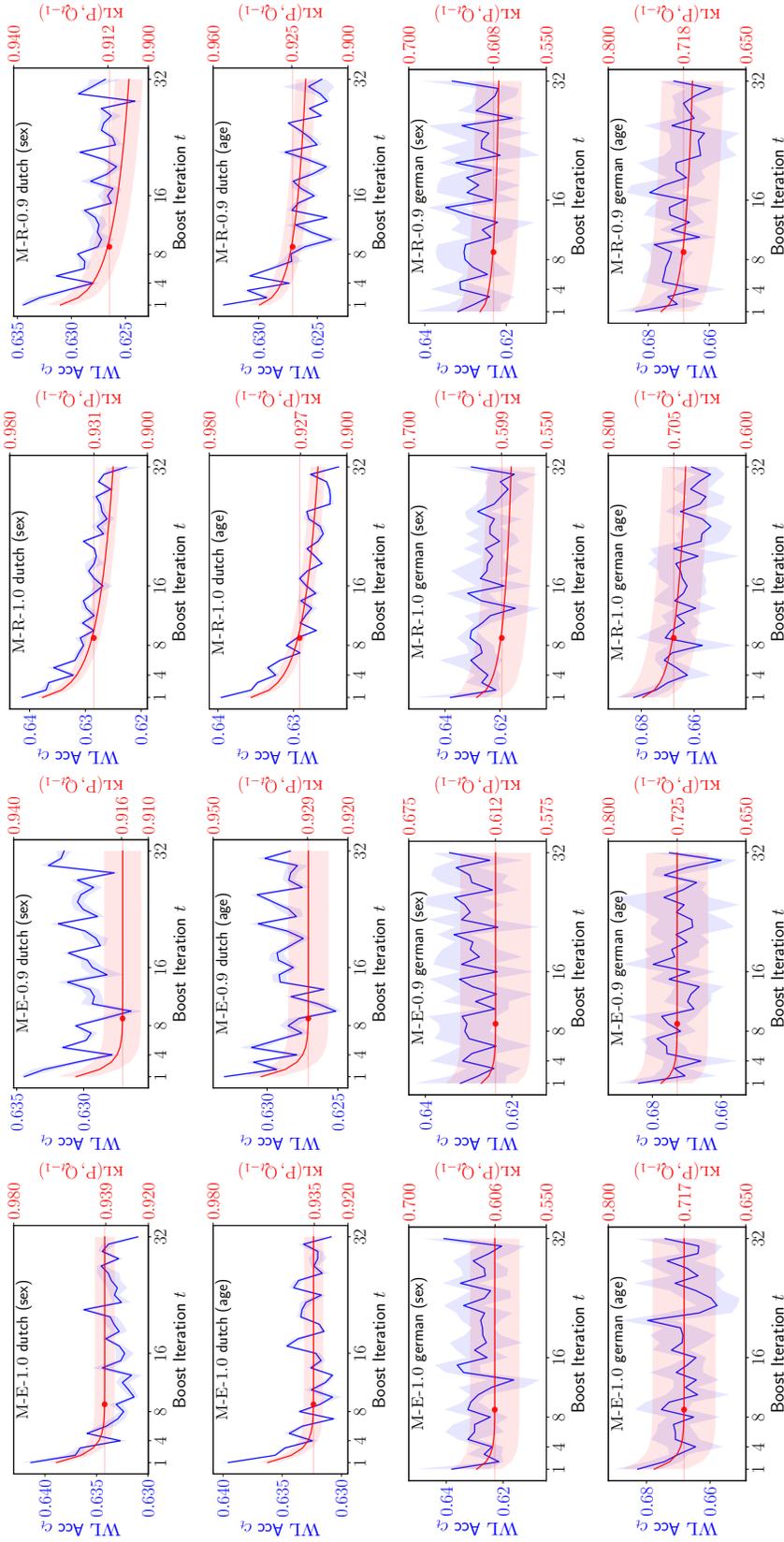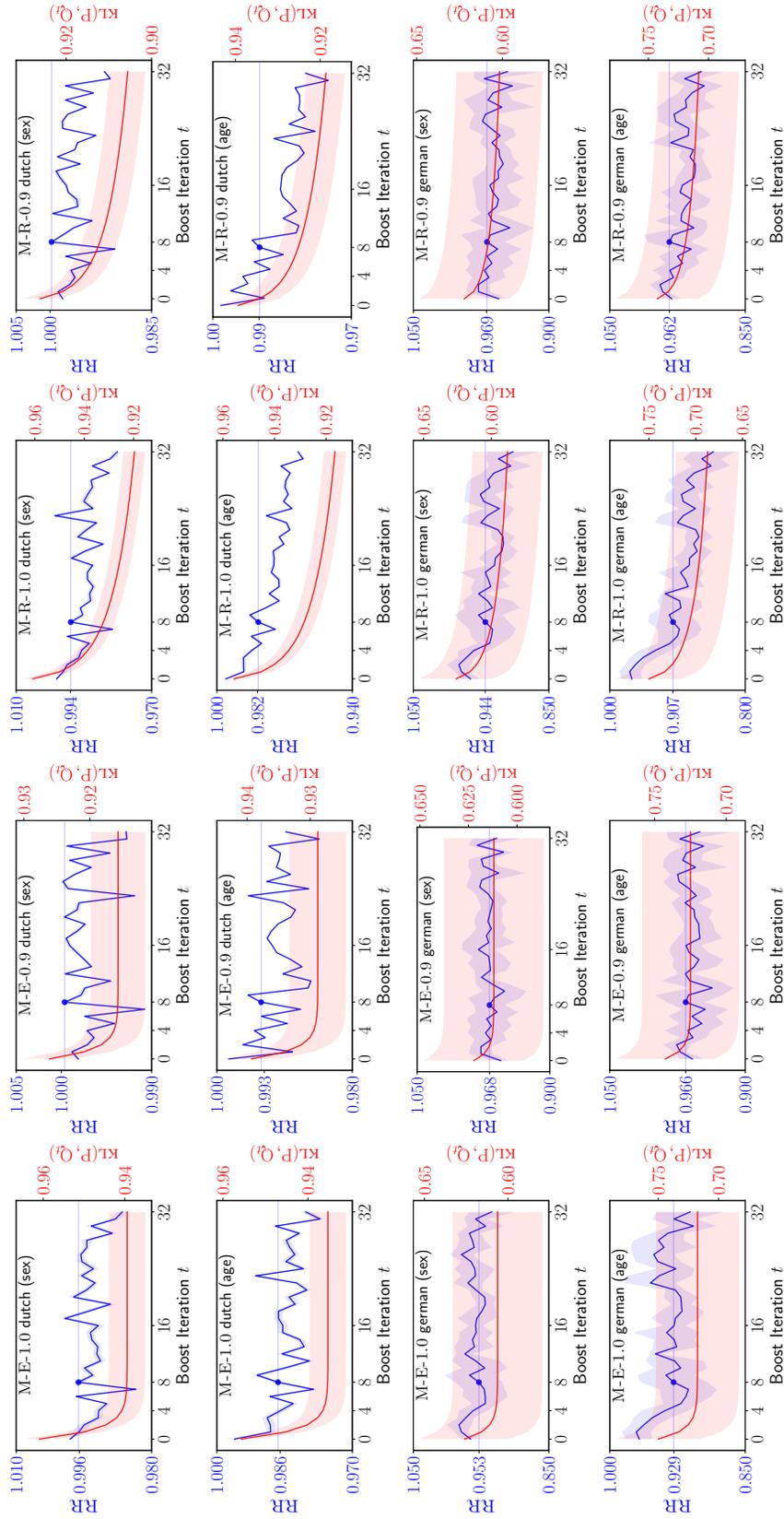*Figure II*. All RR vs KL over boosting iterations plots for DUTCH and GERMAN. horizontal line depicts the $t = 8$ rr value.
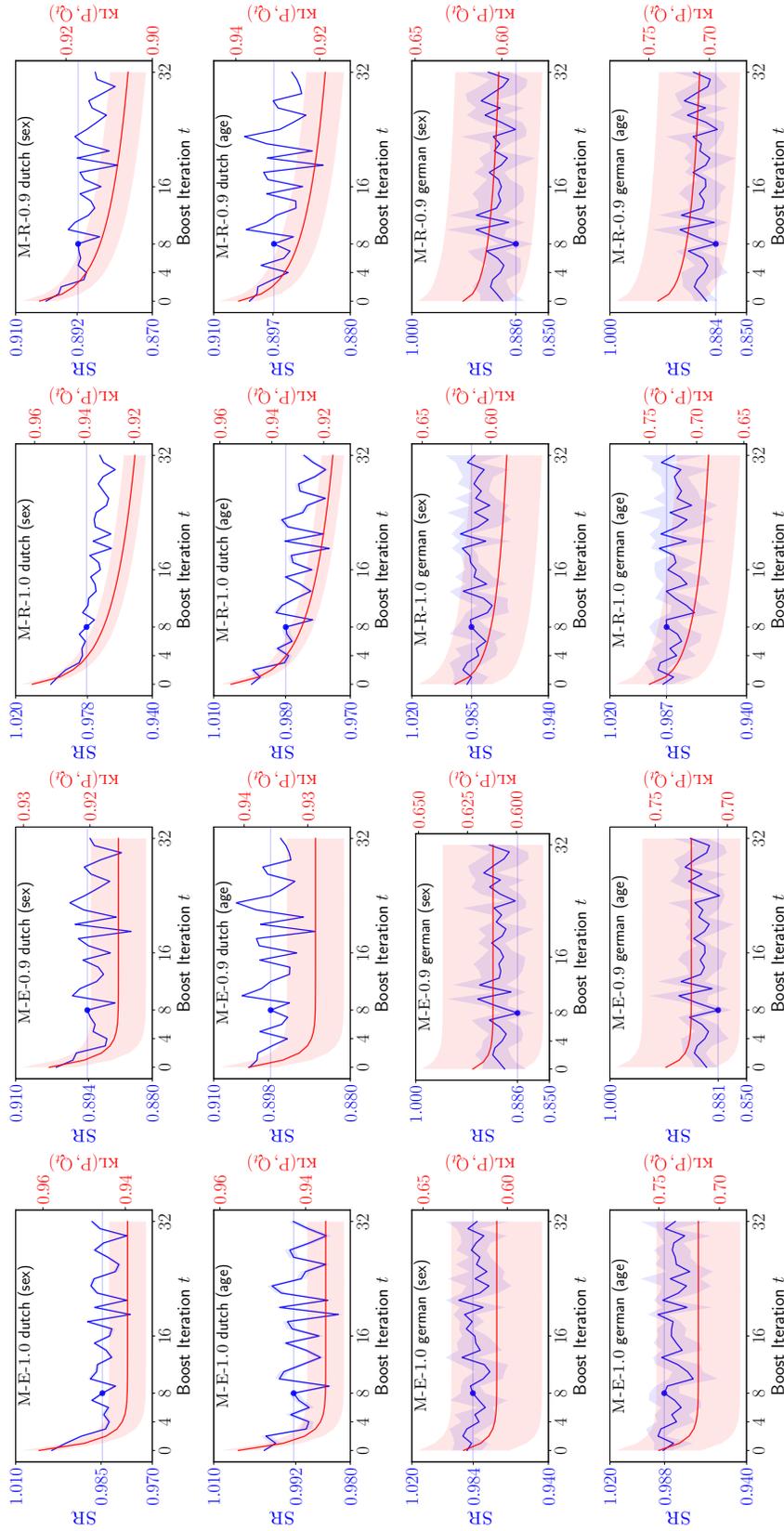
Figure III. All SR vs KL over boosting iterations plots for DUTCH and GERMAN. horizontal line depicts the $t = 8$ sr value.
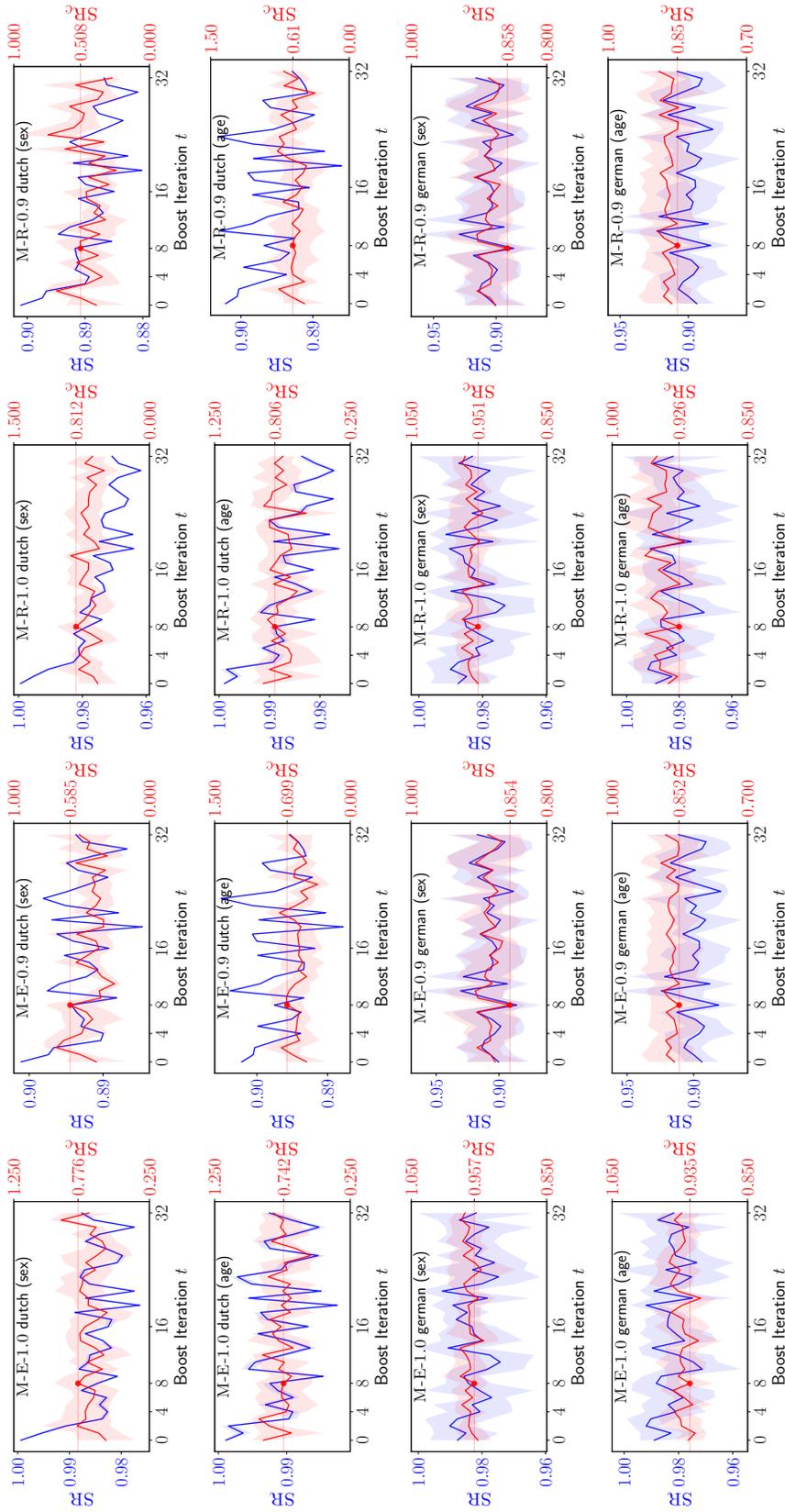
Figure IV. All SR vs Classification SR over boosting iterations plots for DUTCH and GERMAN. Horizontal line depicts the $T = 8$ SR_c value.

Table I. FBDE ($T = 32$) for DUTCH and GERMAN datasets without mixing. The table reports the mean and s.t.d.

			DATA	M-E-1.0	M-E-0.9	M-R-1.0	M-R-0.9
DUTCH	SEX	RR	$.996 \pm .002$	$.984 \pm .000$	$.992 \pm .000$	$.977 \pm .000$	$.990 \pm .000$
		SR	$.523 \pm .002$	$.978 \pm .000$	$.889 \pm .000$	$.952 \pm .000$	$.878 \pm .000$
		KL	$1.715 \pm .021$	$1.797 \pm .022$	$1.768 \pm .022$	$1.788 \pm .022$	$1.765 \pm .022$
	AGE	RR	$.621 \pm .001$	$.966 \pm .000$	$.985 \pm .000$	$.938 \pm .000$	$.969 \pm .000$
		SR	$.610 \pm .003$	$.979 \pm .000$	$.889 \pm .000$	$.958 \pm .000$	$.880 \pm .000$
		KL	$1.716 \pm .021$	$1.773 \pm .022$	$1.761 \pm .022$	$1.767 \pm .022$	$1.758 \pm .022$
GERMAN	SEX	RR	$.449 \pm .014$	$.933 \pm .036$	$.959 \pm .027$	$.896 \pm .039$	$.943 \pm .039$
		SR	$.897 \pm .018$	$.983 \pm .018$	$.919 \pm .024$	$.985 \pm .014$	$.917 \pm .024$
		KL	$1.178 \pm .278$	$1.245 \pm .286$	$1.246 \pm .286$	$1.238 \pm .285$	$1.242 \pm .286$
	AGE	RR	$.235 \pm .008$	$.906 \pm .037$	$.950 \pm .037$	$.838 \pm .032$	$.908 \pm .033$
		SR	$.794 \pm .029$	$.984 \pm .018$	$.910 \pm .020$	$.983 \pm .015$	$.905 \pm .022$
		KL	$1.192 \pm .294$	$1.382 \pm .264$	$1.386 \pm .266$	$1.360 \pm .265$	$1.374 \pm .266$

Table I. FBDE ($T = 32$) for COMPAS, ADULT, DUTCH, and GERMAN datasets compared against in-processing algorithms. The table reports the mean and s.t.d.

			DATA	M-E-1.0	M-E-0.9	M-R-1.0	M-R-0.9	FAIRCONS	REDUCT	FAIRGLM
COMPAS	SEX	SR$_c$	$.726 \pm .025$	$.952 \pm .006$	$.874 \pm .009$	$.938 \pm .019$	$.874 \pm .016$	$.980 \pm .012$	$.874 \pm .021$	$.874 \pm .016$
		EO	$.764 \pm .031$	$.966 \pm .021$	$.905 \pm .027$	$.963 \pm .027$	$.907 \pm .030$	$.981 \pm .012$	$.899 \pm .028$	$.899 \pm .027$
		ACC	$.660 \pm .004$	$.651 \pm .008$	$.654 \pm .005$	$.657 \pm .009$	$.655 \pm .008$	$.573 \pm .027$	$.665 \pm .009$	$.667 \pm .007$
	RACE	SR$_c$	$.747 \pm .020$	$.959 \pm .025$	$.875 \pm .025$	$.945 \pm .027$	$.872 \pm .024$	$.997 \pm .001$	$.822 \pm .043$	$.797 \pm .008$
		EO	$.781 \pm .034$	$.960 \pm .026$	$.900 \pm .041$	$.950 \pm .028$	$.895 \pm .039$	$.987 \pm .007$	$.858 \pm .033$	$.838 \pm .033$
		ACC	$.660 \pm .004$	$.641 \pm .014$	$.653 \pm .012$	$.642 \pm .010$	$.656 \pm .012$	$.515 \pm .016$	$.664 \pm .009$	$.664 \pm .009$
ADULT	SEX	SR$_c$	$.360 \pm .003$	$.818 \pm .010$	$.766 \pm .010$	$.793 \pm .011$	$.753 \pm .008$	$.920 \pm .013$	$.814 \pm .011$	$.857 \pm .009$
		EO	$.471 \pm .008$	$.959 \pm .016$	$.908 \pm .018$	$.935 \pm .016$	$.895 \pm .016$	$.983 \pm .009$	$.964 \pm .023$	$.979 \pm .018$
		ACC	$.803 \pm .003$	$.785 \pm .002$	$.788 \pm .002$	$.787 \pm .002$	$.788 \pm .002$	$.786 \pm .003$	$.789 \pm .002$	$.787 \pm .003$
	RACE	SR$_c$	$.600 \pm .031$	$.867 \pm .035$	$.808 \pm .031$	$.860 \pm .034$	$.803 \pm .035$	$.806 \pm .020$	$.817 \pm .018$	$.843 \pm .015$
		EO	$.787 \pm .053$	$.933 \pm .019$	$.960 \pm .037$	$.935 \pm .017$	$.957 \pm .034$	$.964 \pm .036$	$.969 \pm .027$	$.970 \pm .029$
		ACC	$.803 \pm .003$	$.800 \pm .002$	$.801 \pm .003$	$.800 \pm .002$	$.801 \pm .003$	$.803 \pm .002$	$.803 \pm .003$	$.803 \pm .002$
DUTCH	SEX	SR$_c$	$.464 \pm .022$	$.696 \pm .193$	$.518 \pm .255$	$.628 \pm .139$	$.275 \pm .068$	$1.00 \pm .000$	$1.00 \pm .000$	$1.00 \pm .000$
		EO	$.890 \pm .006$	$.973 \pm .005$	$.986 \pm .008$	$.980 \pm .007$	$.987 \pm .007$	$.821 \pm .011$	$.958 \pm .012$	$.938 \pm .008$
		ACC	$.827 \pm .001$	$.763 \pm .003$	$.778 \pm .004$	$.770 \pm .003$	$.781 \pm .001$	$.771 \pm .004$	$.818 \pm .003$	$.816 \pm .003$
	AGE	SR$_c$	$.695 \pm .030$	$.816 \pm .094$	$.676 \pm .257$	$.747 \pm .084$	$.709 \pm .180$	$1.00 \pm .000$	$1.00 \pm .000$	$1.00 \pm .000$
		EO	$.892 \pm .009$	$.993 \pm .005$	$.960 \pm .007$	$.987 \pm .007$	$.961 \pm .009$	$.929 \pm .005$	$.912 \pm .011$	$.912 \pm .010$
		ACC	$.827 \pm .001$	$.781 \pm .004$	$.788 \pm .004$	$.785 \pm .005$	$.790 \pm .005$	$.797 \pm .003$	$.835 \pm .003$	$.827 \pm .002$
GERMAN	SEX	SR$_c$	$.903 \pm .055$	$.971 \pm .014$	$.887 \pm .049$	$.970 \pm .019$	$.886 \pm .050$	$.997 \pm .001$	$.939 \pm .004$	$.936 \pm .024$
		EO	$.928 \pm .055$	$.980 \pm .021$	$.909 \pm .046$	$.969 \pm .023$	$.907 \pm .051$	$.991 \pm .008$	$.929 \pm .002$	$.923 \pm .019$
		ACC	$.690 \pm .022$	$.712 \pm .011$	$.705 \pm .011$	$.714 \pm .019$	$.706 \pm .007$	$.712 \pm .008$	$.700 \pm .015$	$.700 \pm .015$
	AGE	SR$_c$	$.802 \pm .067$	$.947 \pm .027$	$.881 \pm .045$	$.950 \pm .024$	$.890 \pm .048$	$.978 \pm .018$	$.872 \pm .057$	$.910 \pm .028$
		EO	$.823 \pm .078$	$.940 \pm .028$	$.894 \pm .059$	$.943 \pm .027$	$.914 \pm .058$	$.963 \pm .018$	$.879 \pm .077$	$.916 \pm .047$
		ACC	$.690 \pm .022$	$.700 \pm .016$	$.693 \pm .020$	$.697 \pm .013$	$.707 \pm .011$	$.698 \pm .015$	$.690 \pm .015$	$.698 \pm .017$