Pairwise Fairness for Ordinal Regression

Matthäus Kleindessner
Amazon

Samira Samadi
MPI-IS Tübingen

Muhammad Bilal Zafar
Amazon

Krishnaram Kenthapadi*
Fiddler AI

Chris Russell
Amazon

Abstract

We initiate the study of fairness for ordinal regression. We adapt two fairness notions previously considered in fair ranking and propose a strategy for training a predictor that is approximately fair according to either notion. Our predictor has the form of a threshold model, composed of a scoring function and a set of thresholds, and our strategy is based on a reduction to fair binary classification for learning the scoring function and local search for choosing the thresholds. We provide generalization guarantees on the error and fairness violation of our predictor, and we illustrate the effectiveness of our approach in extensive experiments.

1 INTRODUCTION

As machine learning (ML) algorithms have become an integral part of numerous human-centric domains, they have shown a wide range of concerning behaviors: criminal recidivism tools mislabeling black low-risk defendants as high-risk (Angwin et al., 2016); word2vec embeddings encoding stereotypes such as "a father is to a doctor as a mother is to a nurse" (Bolukbasi et al., 2016); and facial recognition systems having lower accuracy on darker-skinned or female faces (Buolamwini and Gebru, 2018), to name just the most prominent examples. These observations have led to the study of fairness in ML (Barocas et al., 2019), and in the past years numerous ML tasks have been studied from a fairness perspective. While most works consider (binary) classification (e.g., Hardt et al., 2016), fair algorithms have also been developed for regression (e.g., Berk et al., 2017; Agarwal et al., 2019) and several unsupervised learning tasks (e.g., Chierichetti et al., 2017; Celis et al., 2018a,b; Ekstrand et al., 2018; Samadi et al., 2018; Kleindessner et al., 2019a,b; Ghadiri et al., 2021).

While ordinal regression is a widespread task in ML and data science, this is the first work to address the problem of fair ordinal regression. Ordinal regression (aka ordinal classification—see Sec. 2 for a formal description; Gutiérrez et al., 2016) refers to multiclass classification over an ordered label set. Consider a hiring scenario, where given a job applicant's features, such as education, we want to predict a label in {bad, okay, good, excellent}. Clearly, it is less critical to misclassify an excellent applicant as good than misclassifying an excellent one as bad. Algorithms for ordinal regression take such order information into account and typically assign different costs to different misclassifications. However, the order information not only entails that different kinds of misclassifications should be weighted differently; it also carries fairness implications. For example, an applicant that should be scored as okay would feel treated unfairly if misclassified as bad, but probably not if they were misclassified as good. Or it might be acceptable to misclassify all excellent applicants from a minority group as good as long as the excellent applicants from other (majority) groups are misclassified in the same way.

In this paper, we initiate the study of fairness for ordinal regression by making the following **contributions**:

- We propose two pairwise fairness notions for ordinal regression, which we adapt from the literature on fair ranking (Sec. 2.1; see Sec. 4 for related work). We also sketch other possible fairness notions (Sec. 6).
- Focusing on the two pairwise notions, we propose

*Work done at Amazon.

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

a strategy to learn a predictor that is accurate and approximately fair according to either notion (Sec. 3.1 & Sec. 3.2). Our approach is based on a reduction to fair binary classification and some local search procedure. It allows us to control the trade-off between accuracy and fairness that typically exists via a parameter.

- We provide generalization bounds on the error and the fairness violation of our predictor (Sec. 3.3).
- We shed light on computational aspects of our strategy (Sec. 3.2), discuss its limitations (Sec. 3.4), and prove that some simpler alternatives can perform arbitrarily worse (Sec. 3.5).
- We perform extensive experiments and compare to "unfair" state-of-the-art methods in oder to illustrate the effectiveness of our approach (Sec. 5 & App. B).

2 SETUP & FAIRNESS NOTIONS

Setup in Ordinal Regression (Without Fairness) Given an input point $x \in \mathcal{X}$, we want to accurately predict its label $y \in \mathcal{Y}$, where the label set $\mathcal{Y}$ is totally ordered (such as bad $\prec$ okay $\prec$ good $\prec$ excellent in the example of Sec. 1). W.l.o.g., we identify $\mathcal{Y}$ with $[k] := \{1, \ldots, k\}$, that is $\mathcal{Y} = [k]$. Accuracy is measured according to a cost matrix $C \in \mathbb{R}_{\geq 0}^{k \times k}$, where $C_{i,j}$ is the cost that we incur when misclassifying a point with true label i as having label j. The order on $\mathcal{Y}$ entails information about the proximity of labels, and we assume that the misclassification cost $C_{i,j}$ can only increase as j moves away from i: formally, we assume C to have V-shaped rows, that is $C_{i,j-1} \geq C_{i,j}$ for $2 \leq j \leq i$ and $C_{i,j} \leq C_{i,j+1}$ for $i \leq j \leq k-1$, with $C_{i,i} = 0$, $i \in [k]$. One choice for C is the binary cost matrix $C_{i,j} = \mathbb{1}\{i \neq j\}$, used in standard multiclass classification, but not taking label order into account. A popular choice in ordinal regression is the absolute cost matrix $C_{i,j} = |i - j|$.

Datapoints (x, y) are assumed to be drawn i.i.d. from a joint distribution $\mathbb{P}$ on $\mathcal{X} \times [k]$. Given a set of training points $((x_i, y_i))_{i=1}^n$, our goal is to learn a predictor $f : \mathcal{X} \to [k]$ with small expected cost $\mathbb{E}_{(x,y)\sim\mathbb{P}}\, C_{y,f(x)}$. If C is the absolute cost matrix, we refer to the expected cost as mean absolute error (MAE).

2.1 Pairwise Fairness Notions

When concerned about fairness, we assume that datapoints come with a protected attribute $a \in \mathcal{A}$ and consider $(x, y, a) \sim \mathbb{P}$, where $\mathbb{P}$ is now a distribution on $\mathcal{X} \times [k] \times \mathcal{A}$. The attribute a encodes sensitive information such as gender or race, and we assume $\mathcal{A}$ is finite.[1] As before, we want to learn a predictor $f : \mathcal{X} \to [k]$ with small expected cost $\mathbb{E}_{(x,y,a)\sim\mathbb{P}}\, C_{y,f(x)}$, but now f should also be fair with respect to the protected attribute a and avoid discrimination against any protected group (i.e., all individuals that share a value of a). Although satisfying our proposed fairness notions would be easier if we granted f access to a, we assume that f does not use a.[2] This is a common assumption in the literature on fair ML, because otherwise we would commit disparate treatment (Barocas et al., 2019) or the attribute a might not even be available at test time. At training time or when assessing the fairness of a predictor, we assume access to a.

There are several plausible ways to define what it means that f is fair. In this paper we focus on two pairwise fairness notions as we formalize them in the following. We sketch other options in Section 6.

Pairwise demographic parity (DP) The predictor f satisfies pairwise DP if for all $\tilde{a}, \hat{a} \in \mathcal{A}$

$$\begin{aligned} &\mathbb{P}[f(x_1) > f(x_2) | a_1 = \tilde{a}, a_2 = \hat{a}] = \\ &\qquad \mathbb{P}[f(x_1) < f(x_2) | a_1 = \tilde{a}, a_2 = \hat{a}], \end{aligned} \tag{1}$$

where the probability is over (x_1, y_1, a_1) and (x_2, y_2, a_2) being independent samples from $\mathbb{P}$ and potentially the randomness of the predictor f.

Pairwise equal opportunity (EO) Similarly, we say that f satisfies pairwise EO if for all $\tilde{a}, \hat{a} \in \mathcal{A}$

$$\begin{aligned} &\mathbb{P}[f(x_1) > f(x_2) | a_1 = \tilde{a}, a_2 = \hat{a}, y_1 > y_2] = \\ &\qquad \mathbb{P}[f(x_1) < f(x_2) | a_1 = \tilde{a}, a_2 = \hat{a}, y_1 < y_2]. \end{aligned} \tag{2}$$

To motivate these two definitions, assume that for a point x to be classified (e.g., a job applicant) some predictions $f(x)$ are preferable to others and that the order of preference coincides with the order on $\mathcal{Y}$ (e.g., a prediction "good" is preferred to "okay"). Pairwise DP asks that it is as likely for a point sampled from one protected group to be preferred over a point sampled from a second group as it is for the converse to happen (e.g., a female applicant being considered better than a male applicant happens just as likely as a male applicant being considered better than a female one). This is a pairwise analogue of the standard fairness notion of DP (Kamiran and Calders, 2011), which requires the prediction to be independent of the protected attribute, that is $\mathbb{P}[f(x) = \tilde{y} | a = \tilde{a}] = \mathbb{P}[f(x) = \tilde{y} | a = \hat{a}]$ for all $\tilde{y} \in \mathcal{Y}$ and $\tilde{a}, \hat{a} \in \mathcal{A}$, in the sense that pairwise

[1] Our notions fall into the category of group fairness as opposed to individual fairness (Friedler et al., 2016).

[2] Technically, our formulation allows f to use a by encoding a as part of x. Using a would make the problem easier since it would allow us to learn group-dependent thresholds (cf. Section 3).

DP requires the order of the predicted labels for an input pair to be independent of whether the first point is from group $\tilde{a}$ and the second one from group $\hat{a}$, or the other way round. Pairwise EO asks for the same condition as pairwise DP, but conditioned on the order of the ground-truth labels (e.g., a female applicant that is de facto better than a male competitor being considered better happens just as likely as a male, de facto better applicant being considered better than a female competitor). This is a pairwise analogue of the standard notion of EO (Hardt et al., 2016), which considers $y = 1$ to be the preferred outcome and requires $\mathbb{P}[f(x) = 1|a = \tilde{a}, y = 1] = \mathbb{P}[f(x) = 1|a = \hat{a}, y = 1]$ for all $\tilde{a}, \hat{a} \in \mathcal{A}$. The analogue of $y = 1$ being the preferred outcome is having a higher label in our case.

One might wonder whether there is also a pairwise analogue of the notion of equalized odds (Hardt et al., 2016), which is a stricter notion than standard EO and requires that $\mathbb{P}[f(x) = \hat{y}|a = \tilde{a}, y = \tilde{y}] = \mathbb{P}[f(x) = \hat{y}|a = \hat{a}, y = \tilde{y}]$, $\tilde{a}, \hat{a} \in \mathcal{A}$, $\tilde{y}, \hat{y} \in \mathcal{Y}$. A pairwise analogue yields only a slightly stronger notion than pairwise EO (see App. A.1), and we do not consider it here.

Fairness notions similar to pairwise DP / EO as defined in (1) or (2) were recently introduced in the context of ranking (where $y \in [n]$ and $f : \mathcal{X} \to [n]$ if the dataset to be ranked comprises n elements; Beutel et al., 2019), bipartite ranking ($y \in \{0, 1\}$ and $f : \mathcal{X} \to \mathbb{R}$; Kallus and Zhou, 2019) and standard regression ($y \in \mathbb{R}$ and $f : \mathcal{X} \to \mathbb{R}$; Narasimhan et al., 2020). However, they have not been studied in the context of ordinal regression. We discuss related work in Section 4.

We make some remarks on pairwise DP and EO as defined in (1) and (2) (proofs can be found in App. A.2):

- Any constant predictor $f(x) = i$, for some $i \in [k]$, satisfies both fairness notions. The perfect predictor $f(x) = y$ satisfies pairwise EO, but not necessarily pairwise DP.
- When $k = 2$, pairwise DP and standard DP are equivalent. For general k, standard DP implies pairwise DP, but not the other way round.
- Standard EO is neither sufficient nor necessary for pairwise EO, even for $k = 2$. The notion of equalized odds implies pairwise EO for $k = 2$, but not for $k > 2$.

Consequently, pairwise DP is a less restrictive fairness notion than standard DP, and pairwise EO and standard EO / equalized odds are incomparable. We believe that each of these notions can be the most appropriate one in a given scenario (cf. Sec. 6). For example, assume we predict the quality of cars offered on a marketplace on a scale from one to ten, and we want to be fair with respect to different vendors. Vendors would care more that their cars do not unjustifiably loose the comparisons with the cars offered by their competitors (thus requesting pairwise EO), rather than caring whether the probability of a car with true quality "3" being predicted a quality of "7" is the same for all vendors (corresponding to equalized odds). Standard EO, which is primarily designed for binary classification, is not appropriate in this scenario since it requires declaring a single preferred outcome and treats all other outcomes as equal, and both standard and pairwise DP do not take the different ground-truth qualities into account. In Appendix A.3 we present a similar example in more detail.

In the next section we propose a strategy to learn a predictor that is accurate and approximately satisfies either pairwise DP or EO. We measure the amount by which a predictor f violates the two notions by DP-viol or EO-viol, defined as the maximum absolute difference between the left and right sides of (1) or (2); e.g.,

$$\begin{aligned}\text{DP-viol}(f; \mathbb{P}) = \max_{\tilde{a},\hat{a} \in \mathcal{A}} |\mathbb{P}[f(x_1) > f(x_2)|a_1 = \tilde{a}, a_2 = \hat{a}] \\ - \mathbb{P}[f(x_1) < f(x_2)|a_1 = \tilde{a}, a_2 = \hat{a}]|.\end{aligned}$$

On a dataset $\mathcal{D}$ of size n we can evaluate DP-viol$(f; \mathcal{D})$ in time $\mathcal{O}(n + k|\mathcal{A}|^2)$ and EO-viol$(f; \mathcal{D})$ in time $\mathcal{O}(n + k^2|\mathcal{A}|^2)$, assuming evaluating f on x takes time $\mathcal{O}(1)$. If $\mathcal{D}$ is an i.i.d. sample from $\mathbb{P}$, for a given f, we have $|\,\text{Fair-viol}(f; \mathbb{P}) - \text{Fair-viol}(f; \mathcal{D})| \leq M\sqrt{(\log \frac{|\mathcal{A}|^2}{\delta})/n}$ with probability $1 - \delta$ over the sample $\mathcal{D}$, where M is some constant (see App. A.4).

3 LEARNING A FAIR MODEL

Threshold models are a common approach to ordinal regression (Gutiérrez et al., 2016). They consist of a scoring function $s : x \mapsto s(x) \in \mathbb{R}$ and $k - 1$ thresholds $\theta_1 \leq \ldots \leq \theta_{k-1}$ and predict label i for input point x if $s(x) \in (\theta_{i-1}, \theta_i]$ (with $\theta_0 = -\infty$ and $\theta_k = +\infty$). Our proposed strategy consists of learning an accurate and approximately fair threshold model in a two-step approach: first, we learn a scoring function s that approximately satisfies pairwise DP or EO (i.e., s satisfies (1) or (2) with f replaced by s) via a reduction to fair binary classification. Next, we choose thresholds that result in an approximately fair predictor f. Although our definitions of pairwise DP and EO allow for a randomized f, we aim to learn a deterministic f since in decisions strongly affecting humans' lives (such as hiring) randomization is often seen as problematic (Cotter et al., 2019b). From now on, we assume that $\mathcal{X} \subseteq \mathbb{R}^d$. The proofs of all statements are in Appendix A.5-A.10.

3.1 Learning a Fair Scoring Function

Ideally, s satisfies $s(x_1) < s(x_2) \Leftrightarrow y_1 < y_2$. Considering a linear scoring function $s(x) = w \cdot x$ for some $w \in \mathbb{R}^d$, we have $s(x_1) < s(x_2) \Leftrightarrow \mathrm{sgn}(w \cdot (x_1 - x_2)) = -1$. Given training data $\mathcal{D} = ((x_i, y_i))_{i=1}^n$, the well-known approach of Herbrich et al. (2000) to ordinal regression exploits this equivalence and learns a linear scoring function, parameterized by w, by learning a linear classifier, also parameterized by w, that aims to solve the binary classification problem with training dataset $\mathcal{D}' = \{(x', y') = (x_i - x_j, \mathrm{sgn}(y_i - y_j)) : i, j \in [n], y_i \neq y_j\}$. Concretely, the SVM-based algorithm of Herbrich et al. chooses w to minimize the hinge loss on $\mathcal{D}'$. We adapt their approach by learning a linear classifier on $\mathcal{D}'$ that approximately satisfies a fairness constraint closely related to standard DP or EO (cf. Sec. 2.1) on $\mathcal{D}'$ with respect to some attribute a'. As we prove, this implies that the scoring function approximately satisfies pairwise DP or EO on $\mathcal{D} = ((x_i, y_i, a_i))_{i=1}^n$:

Proposition 1 (Reduction to fair binary classification)**.** *Let* $\mathcal{D} = ((x_i, y_i, a_i))_{i=1}^n \subseteq \mathbb{R}^d \times [k] \times \mathcal{A}$ *and* $\mathcal{D}' = \{(x', y', a') = (x_i - x_j, \mathrm{sgn}(y_i - y_j), (a_i, a_j)) : i, j \in [n], y_i \neq y_j\} \subseteq \mathbb{R}^d \times \{-1, 1\} \times \mathcal{A}^2$*. For* $w \in \mathbb{R}^d$*, let* c_w *be the binary classifier* $c_w(x) = \mathrm{sgn}(w \cdot x)$ *and* s_w *be the scoring function* $s_w(x) = w \cdot x$*. We have, for arbitrary* ε*,*

$$\begin{aligned}\max_{\tilde{a}, \hat{a} \in \mathcal{A}} |\mathbb{P}_{(x',y',a') \sim \mathcal{D}'}[c_w(x') = 1 | a' = (\tilde{a}, \hat{a})] - \\ \mathbb{P}_{(x',y',a') \sim \mathcal{D}'}[c_w(x') = 1 | a' = (\hat{a}, \tilde{a})]| \leq \varepsilon\end{aligned} \tag{3}$$

if and only if s_w *satisfies* $\text{DP-viol}(s_w; \mathcal{D}) \leq \varepsilon$ *conditioned on* $y_1 \neq y_2$*, and we have* (3) *with the probabilities conditioned on* $y' = 1$ *if and only if* s_w *satisfies* $\text{EO-viol}(s_w; \mathcal{D}) \leq \varepsilon$*.*

When aiming for pairwise DP, having $\text{DP-viol}(s_w) \leq \varepsilon$ conditioned on $y_1 \neq y_2$ is sufficient for our purposes since we can hope for $s_w(x_1) \approx s_w(x_2)$ for $y_1 = y_2$. This allows us to construct a predictor f with $f(x_1) = f(x_2)$ for $y_1 = y_2$ and $\text{DP-viol}(f) \leq \varepsilon$. We note that the fairness constraint (3) is easier to satisfy than standard DP since it only compares probabilities conditioned on $a' = (\tilde{a}, \hat{a})$ and $\bar{a}' = (\hat{a}, \tilde{a})$, respectively, but not arbitrary $a', \bar{a}' \in \mathcal{A}^2$. For learning c_w that satisfies (3) (or the related constraint in case of EO) on $\mathcal{D}'$, we can utilize the existing approaches for fair binary linear classification from the literature, such as the reduction approach of Agarwal et al. (2018) or the various relaxation-based approaches (e.g., Donini et al., 2018; Wu et al., 2019; Zafar et al., 2019). These methods allow us to choose any of the standard convex loss functions for performing constrained regularized empirical risk minimization (ERM) to learn c_w, and they also allow us to control the trade-off between optimizing for accuracy and satisfying the fairness constraint (i.e., controlling ε in (3)).

3.2 Learning Fair Thresholds

Once s is learned, we require thresholds $\theta_1 \leq \ldots \leq \theta_{k-1}$. We propose to choose thresholds that minimize a weighted combination of the misclassification cost and the fairness violation of the resulting predictor on the training data $\mathcal{D} = ((x_i, y_i, a_i))_{i=1}^n$: if $f(\cdot\,; s, \boldsymbol{\theta})$ denotes the threshold model-predictor with scoring function s and thresholds $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{k-1})$, we solve

$$\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} C_{y_i, f(x_i; s, \boldsymbol{\theta})} + \lambda \cdot \text{Fair-viol}(f(\cdot\,; s, \boldsymbol{\theta}); \mathcal{D}), \tag{4}$$

where Fair-viol is a generic notation for DP-viol or EO-viol as defined in Section 2.1 and $\lambda \geq 0$ controls the extent to which we optimize for accuracy (i.e., small cost) and fairness, respectively. In our experiments, we choose λ aligned with ε from the previous section.

Perhaps surprisingly, the problem (4) can be solved in polynomial time using dynamic programming:

Proposition 2 ((4) can be solved in poly-time)**.** *If* $|\mathcal{A}| = 2$*, then for any* $\lambda \geq 0$ *and* $\text{Fair-viol} \in \{\text{DP-viol}, \text{EO-viol}\}$*, we can find optimal thresholds solving* (4) *in time* $\mathcal{O}(n^6 k^3)$*.*

The dynamic programming approach of Proposition 2 can be generalized to $|\mathcal{A}| > 2$, but then has running time exponential in $|\mathcal{A}|$. Even when $|\mathcal{A}| = 2$, the running time is prohibitively high in practice. Instead, we have to make use of heuristics to efficiently find local minima of (4). We propose to perform a local search, moving one threshold θ_i at a time. We can implement such local search with running time $\mathcal{O}(n|\mathcal{A}|^2)$ per iteration in case of pairwise DP or $\mathcal{O}(nk|\mathcal{A}|^2)$ in case of pairwise EO (see App. A.10 for details).

3.3 Generalization Guarantees

We provide generalization bounds on the MAE and the fairness violation of the predictor learned by our strategy ($B_R(0)$ denotes the ball $\{x \in \mathbb{R}^d : \|x\|_2 \leq R\}$):

Theorem 1 (Generalization bounds)**.** *Assume that the training data* $\mathcal{D}$ *is an i.i.d. sample from a distribution* $\mathbb{P}$ *on* $B_R(0) \times [k] \times \mathcal{A}$ *with* $\mathbb{P}[a = \tilde{a}] \geq \beta > 0$ *for all* $\tilde{a} \in \mathcal{A}$*. Assume that we learn a scoring function* s *as outlined in Section 3.1 by means of constrained regularized ERM with Ivanov* l_2*-regularization* $\|w\|_2^2 \leq \nu$*, and thresholds as outlined in Section 3.2. There exists a constant* M *such that for any* $\gamma > 0$ *and* $0 < \delta < 1$*, our learned predictor* $f = f(\cdot\,; s, \boldsymbol{\theta})$ *satisfies with probability at least* $1 - \delta$ *over the training sample* $\mathcal{D}$ *of size* n*, for* n

sufficiently large,

$$\mathrm{MAE}(f;\mathbb{P}) \leq \widehat{L}^{\gamma}_{\mathcal{D}}(s,\boldsymbol{\theta}) + (k-1)\cdot \sqrt{\frac{M}{n}\left(\frac{R^2\nu}{\gamma^2}\ln(n) + \ln\left(2\frac{(k-1)R\sqrt{\nu}}{\gamma\delta}\right)\right)}$$

and

$$\text{DP-viol}(f;\mathbb{P}) \leq \text{DP-viol}(f;\mathcal{D}) + Mk^2\cdot \sqrt{\left(d+\log\frac{4|\mathcal{A}|}{\delta}\right)\cdot\left[\left(1-\sqrt{\frac{2\log\frac{4|\mathcal{A}|}{\delta}}{n\beta}}\right)n\beta\right]^{-1}}.$$

$\widehat{L}^{\gamma}_{\mathcal{D}}(s,\boldsymbol{\theta})$ *is the empirical γ-margin loss (see App. A.7).*

The statement for EO-viol is provided in Appendix A.7.

3.4 Limitations of our Approach

We briefly discuss limitations of our approach and potential remedies.

- Jointly optimizing the objective function in (4) over the scoring function s and thresholds $\boldsymbol{\theta}$ could potentially yield better results than our two-step approach. This is an interesting direction for future work.
- Our approach learns a linear scoring function, which might be too restrictive in some settings. If the method for learning the fair binary classifier c_w can be kernelized, then our approach can be kernelized similarly to the unfair approach of Herbrich et al. (2000). Alternatively, random features (Rahimi and Recht, 2007) could be used to increase expressiveness.
- Starting with a training set $\mathcal{D}$ of size n our approach constructs a training set $\mathcal{D}'$ of size n^2, which seems natural in light of the pairwise nature of our fairness notions, but might be infeasible for large $\mathcal{D}$. If so, we subsample a training set for learning the classifier c_w.

3.5 Is Fairness Needed in Both Steps?

One might wonder whether we must enforce fairness on the scoring function s, or whether it would suffice to enforce fairness only when choosing the thresholds. Indeed, since any constant predictor is perfectly fair (cf. Sec. 2.1), no matter what s is, we can always choose thresholds so as to achieve any desired level of fairness. However, if s is fair to some extent, there tend to be more choices of thresholds that yield a fair predictor (see App. A.12 for some simulations) and hence we can get a more accurate predictor. Similarly, one can wonder whether it suffices to enforce fairness only when learning the scoring function. The next lemma shows that an approach in which we enforce fairness in only one of the two learning steps can perform arbitrarily worse compared to our two-step approach:

Lemma 1 (Enforcing fairness in both steps can be necessary)**.** *There exist datasets in* $\mathbb{R}^2\times[4]\times\{0,1\}$ *and a class S of scoring functions* $s:\mathbb{R}^2\rightarrow\mathbb{R}$ *that show:*

1. *Choosing a scoring function s that minimizes the number of label flips, but is not required to satisfy pairwise DP (EO), and subsequently thresholds that minimize the* MAE *under the constraint that the resulting predictor satisfies pairwise DP (EO) can result in a* MAE *that is arbitrarily high compared to enforcing pairwise DP (EO) both when choosing s and the thresholds.*
2. *Choosing a scoring function s to minimize the number of label flips under the constraint that s satisfies pairwise DP (EO) and subsequently thresholds that minimize the* MAE*, but are not required to yield a predictor satisfying pairwise DP (EO), can result in a predictor with* DP-viol $=0.5$ *(*EO-viol $=1$*). In contrast, thresholds required to yield a predictor satisfying pairwise DP (EO), will always give* DP-viol $=0$ *(*EO-viol $=0$*).*

The examples used to prove Lemma 1 are worst-case examples that might not occur in practice. However, in Section 5 we also present experiments on real-world datasets where our two-step approach clearly outperforms an approach that is fair in only one step. The second statement of Lemma 1 is slightly different for pairwise DP and EO (DP-viol $=0.5$ vs EO-viol $=1$). The next lemma shows that this is not an artefact of our proof: for pairwise DP, the fairness violation of the predictor can be upper bounded by the fairness violation of the scoring function for any choice of thresholds, such that DP-viol $=0.5$ is the worst we can get using a perfectly fair scoring function.

Lemma 2 (Fairness of scoring function vs fairness of predictor for DP)**.** *Let* $s:\mathcal{X}\rightarrow\mathbb{R}$ *be a scoring function and* $f:\mathcal{X}\rightarrow[k]$ *be a predictor that is obtained from thresholding s. No matter what the thresholds are, we have* DP-viol$(f)\leq 1/2+$DP-viol$(s)/2$.

4 RELATED WORK

Ordinal Regression A review of classical techniques, such as the proportional odds model (POM; McCullagh, 1980), is provided by O'Connell (2006). Gutiérrez et al. (2016) provide a more recent survey. They categorize ordinal regression methods into naive approaches, binary decomposition, and threshold models. Our approach falls into the latter category. It is based on the SVM pairwise approach of Herbrich et al. (2000), which has been refined by SVM pointwise approaches (Shashua and Levin, 2002; Chu and Keerthi, 2007). The pointwise approaches avoid the issue of transforming a training set of size n to one of size n^2. However, it is unclear how to incorporate fairness into these approaches. Some newer methods are the deep learning-based ones by Niu et al. (2016), Liu et al. (2017), Polania et al. (2019) and Cao et al. (2020), which are designed for image data and age or body mass index estimation. None of the existing methods for ordinal regression takes fairness into account.

Related Fairness Notions Beutel et al. (2019) introduce a fairness notion similar to pairwise EO as defined in (2) for a ranking function used in a recommendation system. They train a fair model by adding a correlation penalty term to the loss function. Kuhlman et al. (2019) propose fairness criteria equivalent to pairwise DP or EO for general rankings, and an auditing mechanism that evaluates the criteria on several subparts of a ranking. Kallus and Zhou (2019) study a fairness notion similar to pairwise EO for bipartite ranking. They are mainly concerned with evaluating their notion, but also propose a simple post-processing mechanism to obtain a fair group-dependent scoring function from an unfair one. The notion of Kallus and Zhou and some closely related notions by Borkan et al. (2019) have been generalized by Vogel et al. (2021), who provide generalization guarantees for a scoring function learned under their fairness constraints and train a fair model by adding a highly non-convex surrogate of the fairness violation to the loss function. Pairwise DP and EO as well as related variants, including the case of a continuous protected attribute, are also introduced by Narasimhan et al. (2020) for a ranking or regression function $f : \mathcal{X} \to \mathbb{R}$. They train a fair model using the algorithm of Cotter et al. (2019a) for optimization problems with non-differentiable constraints. When Narasimhan et al. consider ranking, their approach is similar to the first step in our two-step approach. Note that there is no obvious way how to formulate our problem of fair ordinal regression as a fair ranking / bipartite ranking / regression problem. For example, when trying to formulate it as a fair regression problem and use the techniques of Narasimhan et al., we would have to carefully choose the encoding of a training point's label $y \in \mathcal{Y}$ as a label $y \in \mathbb{R}$. An arbitrary monotone encoding does not work since Narasimhan et al. only consider the squared loss on $\mathbb{R}$. Furthermore, we would have to decide how to map a prediction $f(x) \in \mathbb{R}$ to one in $\mathcal{Y}$.

Other existing, non-pairwise fairness notions for ranking (e.g., Zehlike et al., 2017; Biega et al., 2018; Celis et al., 2018b; Singh and Joachims, 2018) or regression (e.g., Johnson et al., 2016; Berk et al., 2017; Pérez-Suay et al., 2017; Agarwal et al., 2019) are very different from our proposed notions.

5 EXPERIMENTS

To illustrate the relevance of our approach, we first consider two real-world datasets for which fairness is of concern. We then investigate the performance of our approach by comparing it to "unfair" state-of-the-art methods for ordinal regression on numerous benchmark datasets. These benchmark datasets do not come with a meaningful protected attribute, and we instead treat other features as "protected". Although not ideal, such evaluation is widely common in the literature on fair ML. See Appendix B.2 for details about all datasets.

We implemented our proposed approach in Python.[3] In the first step of our approach, that is when learning the scoring function via the reduction to fair binary classification as described in Section 3.1, we built on the approach of Agarwal et al. (2018), whose implementation is available as part of FAIRLEARN (https://fairlearn.github.io/). We used the GRIDSEARCH-method (corresponding to Sec. 3.4 of Agarwal et al.), which returns a deterministic rather than a randomized classifier. We provide some details in Appendix B.1. For solving the second problem in our approach, that is choosing the thresholds, we ran the local search as explained in Section 3.2 with random initialization for ten times and kept the solution with the lowest objective value (evaluated on training data).

The two main parameters in our strategy correspond to ε in (3) and λ in (4) and govern the accuracy-vs-fairness trade-off when learning the scoring function and thresholds, respectively. The GRIDSEARCH-method is not parameterized in ε, but in $\mu \in [0, 1]$ such that the method aims to minimize $(1 - \mu) \cdot \text{error}(c_w) + \mu \cdot \text{Fair-viol}(c_w)$. We reparameterized (4) using $\lambda' = \frac{\lambda}{k+\lambda} \in [0, 1)$ so that μ and λ' have the same interpretation, and in the following we study our approach for various choices of $\mu = \lambda'$ (without systematically searching for an optimal correspondence between μ and λ', we found that choosing $\mu = \lambda'$ generally works well). Throughout, we identify labels in $\mathcal{Y}$ with $1, \ldots, k$, set C to the absolute cost matrix and aim for a small MAE as defined in Section 2 (in App. B.4 we present an experiment with an asymmetric cost matrix).

5.1 Experiments on Real-World Datasets

We applied our strategy to the Drug Consumption dataset (referred to as DC dataset; Fehrman et al., 2015) and the Communities and Crime dataset (C&C

[3]Code available on https://github.com/amazon-research/fair-ordinal-regression.

dataset; Redmond and Baveja, 2002). On the DC dataset, we predicted when an individual has last consumed cannabis (never or more than a decade ago - last decade - last year - last month - last day). The input features are a person's age, education and seven features measuring personality traits, and we used a person's gender (f or m) as protected attribute. In such a setting, fairness can be of central concern, for example, when a security firm makes this kind of predictions for its job applicants. On the C&C dataset, we predicted the total number of crimes per 10^5 inhabitants (discretized to 8 classes) for communities represented by 95 input features such as the median income. As protected attribute $a \in \{\text{white}, \text{diverse}\}$, we used whether a community is predominantly (i.e., $> 80\%$) inhabited by Caucasians (in App. B.3 we consider $a \in \{\text{white}, \text{African-American}, \text{Hispanic or Asian}\}$). Also in this scenario, fairness is highly relevant (Lum and Isaac, 2016). The datasets comprise 1885 and 1994 records, respectively, which we randomly split into a training set of size 1500 and a test set of size 385 or 494. We report results on the test sets, averaged over 20 random splits.

The plots in the top row of Figure 1 show the results for the DC dataset and the plots in the bottom row for the C&C dataset. They show on the x-axis the fairness violation DP-viol (left) or EO-viol (right) and on the y-axis the MAE of various predictors: **(i)** the predictors that we obtained from running our approach for $\mu = \lambda' \in \{0, 0.1, 0.2, 0.3, 0.33, 0.36, 0.4, 0.5, \ldots, 0.9\}$. For the left plots we enforce approximate pairwise DP and for the right plots pairwise EO. **(ii)** the predictor obtained from fitting a POM model (McCullagh, 1980), a classical threshold model technique that does not require any hyperparameter tuning and is among the most widely used ordinal regression methods in practice (Gutiérrez et al., 2016, Sec. 4.4). This is a simple unfair baseline. **(iii)** the best constant predictor; since we want to minimize the MAE, this is $f(x) = \text{median}(y_1, \ldots, y_n)$ for training data $\mathcal{D} = ((x_i, y_i, a_i))_{i=1}^n$. This is the most simple fair baseline. **(iv)** randomized predictors that predict $\text{median}(y_1, \ldots, y_n)$ with probability p and the prediction of the POM-predictor with probability $1 - p$, for $p = \frac{j}{50}$, $j \in [49]$. For these randomized predictors, the plots show the performance obtained from averaging over 100 times of applying such a predictor to the test set. This is a simple approximately fair baseline. **(v)** the predictors that we obtained by choosing approximately fair thresholds for the scoring function of the POM model based on (4). In doing so, we used the same values of λ' and ran the local search strategy in the same way as for our approach. This is supposed to show the need of incorporating fairness when learning the scoring function (cf. Sec. 3.5).

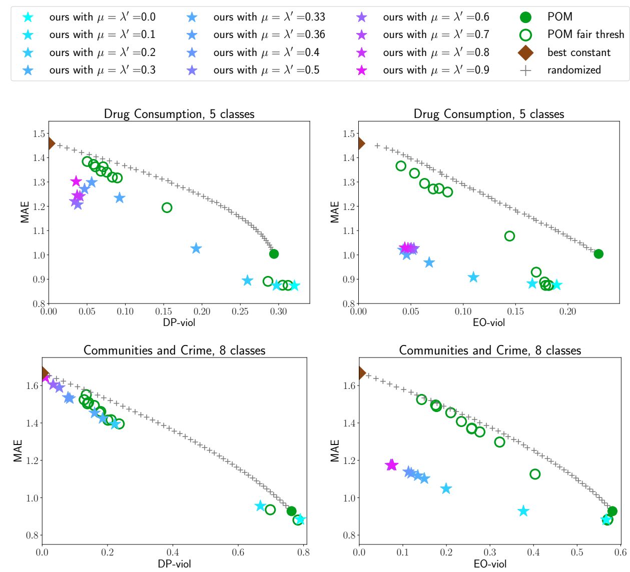

Figure 1: MAE vs DP-viol **(left)** and MAE vs EO-viol **(right)** for various predictors: the predictors produced by our approach (cyan to purple stars), by the POM algorithm (green filled circle), by solving (4) for the scoring function of the POM predictor for various values of λ (green circles), and by randomly mixing the POM predictor with the best constant one (grey crosses); the best constant predictor is represented by the brown diamond. See Appendix B.3 for a version of the bottom left plot where we fill the empty area by considering additional values of $\mu = \lambda'$, and another version where we consider three protected groups.

We make the following observations: **(i)** the POM model can be highly unfair with DP-viol $= 0.29$ / EO-viol $= 0.23$ on the DC dataset and DP-viol $= 0.76$ / EO-viol $= 0.58$ on the C&C dataset. This shows that we cannot expect a standard method for ordinal regression to be fair and that there is a need for an approach that explicitly takes fairness into account. **(ii)** The predictors produced by our approach, in general, nicely explore the trade-off between accuracy and fairness. As expected, the larger the value of $\mu = \lambda'$, the more fair and the less accurate is the predictor. However, on the C&C dataset and when aiming for pairwise DP (bottom left), there is an area in the MAE-vs-DP-viol plot (in the range of DP-viol $\in [0.25, 0.65]$) that is unexplored. One can try to fill that area by running our approach with additional values of $\mu = \lambda' \in [0.1, 0.2]$, and indeed by doing so we succeed. The corresponding plot is shown in Appendix B.3. We deliberately chose not to show that plot instead of the one shown in Figure 1 to illustrate that we have to choose the parameter $\mu = \lambda'$ in an adaptive way if we want to fully explore the accuracy-vs-fairness trade-off that is achievable by our method on a given dataset. **(iii)** For small $\mu = \lambda'$,

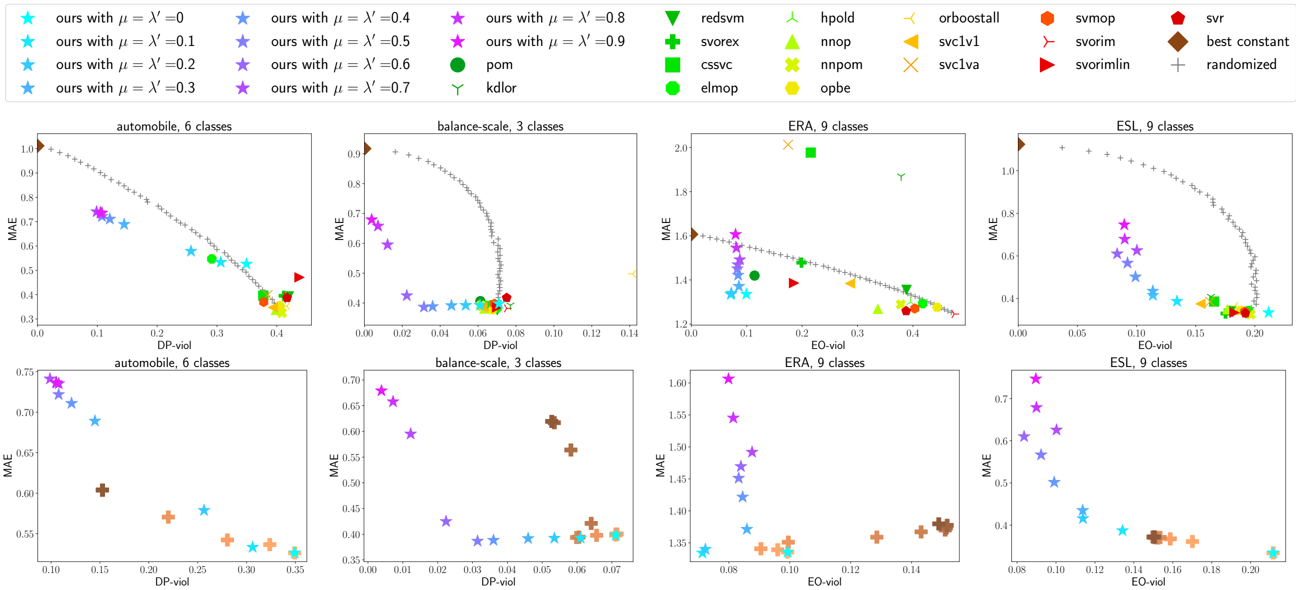

Figure 2: **Top row:** MAE vs DP-viol (1st + 2nd plot) or MAE vs EO-viol (3rd + 4th plot) for the predictors produced by our approach in comparison to the algorithms in the ORCA toolbox for four of the datasets. The plots for the other datasets are in Appendix B.5. **Bottom row:** MAE vs DP-viol or MAE vs EO-viol for the predictors produced by our approach with $\mu = \lambda'$ and with the same values of μ, but $\lambda' = 0$ (light to dark brown crosses).

our predictors have even smaller MAE than the POM predictor. Since $|\mathbb{P}[y_1 > y_2 | a_1 = \mathrm{f}, a_2 = \mathrm{m}] - \mathbb{P}[y_1 < y_2 | a_1 = \mathrm{f}, a_2 = \mathrm{m}]| = 0.35$ on the DC dataset and $|\mathbb{P}[y_1 > y_2 | a_1 = \text{white}, a_2 = \text{diverse}] - \mathbb{P}[y_1 < y_2 | a_1 = \text{white}, a_2 = \text{diverse}]| = 0.72$ on the C&C dataset, it is not surprising that when aiming for pairwise DP and for larger values of $\mu = \lambda'$, our predictors have higher MAE. When aiming for pairwise EO, even for large values of $\mu = \lambda'$ the MAE of our predictors is comparable to the MAE of the POM predictor, but our predictors are significantly more fair. **(iv)** Our predictors, which are deterministic, clearly outperform the randomized predictors. Note that the performance of the randomized predictors interpolates between the performance of the POM predictor and the performance of the best constant predictor in a non-linear way due to the pairwise nature of Fair-viol. Once more, we emphasize that applying a randomized predictor to human subjects might be problematic (cf. Sec. 3). **(v)** Choosing thresholds for the scoring function of the POM model based on (4) improves the MAE of the POM predictor for small values of λ. For larger values of λ, it improves the fairness of the POM predictor, but also increases its MAE. In particular when aiming for pairwise EO, its MAE is much larger than the MAE of our predictors. This confirms our findings of Section 3.5, and in particular the first claim of Lemma 1, where we showed that solving (4) for an unfair scoring function can result in a MAE that is arbitrarily high compared to when using a fair scoring function.

5.2 Comparison to Unfair State-Of-The-Art Methods on Benchmark Datasets

Gutiérrez et al. (2016) performed an extensive evaluation of 16 algorithms for ordinal regression, including classical methods (such as the POM model used in the previous section) as well as state-of-the-art methods (such as the SVM-based algorithms of Chu and Keerthi, 2007, or the neural network approach of Cheng et al., 2008). They applied these algorithms to 41 benchmark datasets, 17 of which come with an actual ordinal regression task (referred to as real ordinal regression datasets) and 24 of which originally come with a standard regression task, that is $y \in \mathbb{R}$, and for which the label y has been discretized to either five or ten classes (referred to as discretized regression datasets). The algorithms and the datasets are publicly available as part of the ORCA toolbox for MATLAB (Sánchez-Monedero et al., 2019), which also comprises the hierarchical model of Sánchez-Monedero et al. (2018); hence, there are 17 algorithms in total.

We applied the 17 ORCA algorithms and our strategy for $\mu = \lambda' \in \{0, 0.1, 0.2, \ldots, 0.9\}$ to the 33 datasets containing at least 200 datapoints. The datasets do not have a meaningful protected attribute, and we treat an arbitrary binary feature as the protected attribute $a \in \{0, 1\}$, or chose some real-valued feature x_r and set $a = \mathbb{1}\{x_r \geq \text{median}(x_r)\}$ for those datasets that do not have any binary features. We ran all ORCA algorithms in the same way as Gutiérrez et al. did; in particular, we performed 5-fold cross validation over the same sets of hyperparameters with the goal of minimizing the MAE. However, no algorithm observes the protected attribute a (or the feature x_r if applicable) as part of the input x. Hence, our experiments do not exactly replicate the ones by Gutiérrez et al.. We also used the same splits into 30 or 20 training and test sets and report all results on the test sets, averaged over the various splits. In Appendix B.5, we show results that include the standard deviation over the splits.

The plots in the top row of Figure 2 show the results for the first four of the real ordinal regression datasets and when aiming for pairwise DP or EO. The results for all datasets, both for pairwise DP and EO, are in Appendix B.5. The plots in the bottom row show the performance of predictors obtained from our approach with $\mu = \lambda'$ in comparison with the performance of predictors obtained from our approach with $\lambda' = 0$. In the latter case, we enforce fairness only when learning the scoring function. The main findings are similar to the ones from Figure 1 (we provide some more interpretation in App. B.5): **(i)** the state-of-the-art methods can be highly unfair. **(ii)** Our predictors explore the accuracy-vs-fairness trade-off, but sometimes there are unexplored areas and we would need to adaptively choose additional values of $\mu = \lambda'$ that we run our strategy with. **(iii)** In particular when aiming for pairwise EO, our predictors are often significantly more fair, but only slightly less accurate than the competitors. **(iv)** We outperform the predictors that we obtain by randomly mixing the predictor with the smallest MAE with the best constant one (similarly as we did in the previous section; shown by the grey crosses). **(v)** Our predictors can potentially produce a much better accuracy-vs-fairness trade-off than the predictors obtained by running our approach with $\lambda' = 0$. For example, on the balance-scale dataset shown in the second plot of the bottom row of Figure 2, the predictor corresponding to $\mu = \lambda' = 0.5$ has MAE $= 0.39$ and DP-viol $= 0.03$, while the best predictor that corresponds to some $\mu \geq 0$ and $\lambda' = 0$ and has the same value of MAE has twice as high a value of DP-viol. This supports our findings of Section 3.5 and in particular the second claim of Lemma 1.

6 DISCUSSION & FUTURE WORK

This paper initiates the study of fair ordinal regression. We adapted two pairwise fairness notions from the literature on fair ranking and proposed a two-step strategy for training a threshold model that allows us to control the trade-off between accuracy and fairness.

There are two main directions for future work: we designed an algorithm for training an accurate and approximately fair predictor. It would be interesting to understand how we can do so in an optimal way (cf. Sec. 3.4, first item) and what is the best accuracy-vs-fairness trade-off that we can achieve, both in principle and for a given model class. The latter question is still receiving considerable attention even in the binary classification setting (e.g., Dutta et al., 2020; Kim et al., 2020). On the other hand, it would be interesting to study other fairness notions for ordinal regression. Two examples of other possible notions are to require that, for all $\tilde{a} \in \mathcal{A}$, $\mathbb{E}[C_{y,f(x)}|f(x) < y, a = \tilde{a}] = \mathbb{E}[C_{y,f(x)}|f(x) < y]$ or $\mathbb{E}[C_{y,f(x)}\mathbb{1}\{f(x) < y\}|a = \tilde{a}] = \mathbb{E}[C_{y,f(x)}\mathbb{1}\{f(x) < y\}]$. The motivation behind these two fairness notions is that an input point x suffers some harm whenever its predicted label $f(x)$ is less preferable than its actual label y. One could also consider a notion of individual fairness (Dwork et al., 2012) for ordinal regression, requiring that similar datapoints are treated similarly. In contrast to standard multiclass classification, in case of ordinal regression, the order on the label set $\mathcal{Y}$ already provides some similarity information about $\mathcal{Y}$. Of course, once we have several fairness notions for ordinal regression at our disposal, it is a non-trivial question to choose the most appropriate one for a given task, just as it is the case in standard binary classification (Makhlouf et al., 2020).

References

A. Agarwal, A. Beygelzimer, M. Dudik, J. Langford, and H. Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning (ICML)*, 2018.

A. Agarwal, M. Dudik, and Z. Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning (ICML)*, 2019.

S. Agarwal. Generalization bounds for some ordinal regression algorithms. In *International Conference on Algorithmic Learning Theory (ALT)*, 2008.

H. Altay Guvenir and I. Uysal. Bilkent University Function Approximation Repository, 2000. http://funapp.cs.bilkent.edu.tr.

J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Propublica—machine bias, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

R. Berk, H. Heidari, S. Jabbari, M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth. A convex framework for fair regression. In *Fairness, Ac-*

countability, and Transparency in Machine Learning (FATML), 2017.

A. Beutel, J. Chen, T. Doshi, H. Qian, L. Wei, Y. Wu, L. Heldt, Z. Zhao, L. Hong, E. Chi, and C. Goodrow. Fairness in recommendation ranking through pairwise comparisons. In *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, 2019.

A. Biega, K. Gummadi, and G. Weikum. Equity of attention: Amortizing individual fairness in rankings. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, 2018.

A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth. Support vector ordinal regression. *Journal of the ACM*, 36(4):929–965, 1989.

T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Neural Information Processing Systems (NeurIPS)*, 2016.

D. Borkan, L. Dixon, J. Sorensen, N. Thain, and L. Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *The World Wide Web Conference (WWW)*, 2019.

J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability, and Transparency (ACM FAT)*, 2018.

W. Cao, V. Mirjalili, and S. Raschka. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140:325–331, 2020.

L. E. Celis, V. Keswani, D. Straszak, A. Deshpande, T. Kathuria, and N. K. Vishnoi. Fair and diverse DPP-based data summarization. In *International Conference on Machine Learning (ICML)*, 2018a.

L. E. Celis, D. Straszak, and N. K. Vishnoi. Ranking with fairness constraints. In *International Colloquium on Automata, Languages and Programming (ICALP)*, 2018b.

J. Cheng, Z. Wang, and G. Pollastri. A neural network approach to ordinal regression. In *International Joint Conference on Neural Networks (IJCNN)*, 2008.

F. Chierichetti, R. Kumar, S. Lattanzi, and S. Vassilvitskii. Fair clustering through fairlets. In *Neural Information Processing Systems (NeurIPS)*, 2017.

W. Chu and Z. Ghahramani. Gaussian processes for ordinal regression. *Journal of Machine Learning Research (JMLR)*, 6:1019–1041, 2005. Code and datasets available on http://www.gatsby.ucl.ac.uk/~chuwei/ordinalregression.html.

W. Chu and S. Keerthi. Support vector ordinal regression. *Neural Computation*, 19(3):792–815, 2007.

P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, 2009. Dataset available on https://archive.ics.uci.edu/ml/datasets/wine+quality.

A. Cotter, H. Jiang, M. Gupta, S. Wang, T. Narayan, S. You, and K. Sridharan. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *Journal of Machine Learning Research (JMLR)*, 20(172):1–59, 2019a.

A. Cotter, H. Narasimhan, and M. Gupta. On making stochastic classifiers deterministic. In *Neural Information Processing Systems (NeurIPS)*, 2019b.

M. Csikos, A. Kupavskii, and N. Mustafa. Tight lower bounds on the VC-dimension of geometric set systems. *Journal of Machine Learning Research (JMLR)*, 20(81):1–8, 2019.

M. Donini, L. Oneto, S. Ben-David, J. Shawe-Taylor, and M. Pontil. Empirical risk minimization under fairness constraints. In *Neural Information Processing Systems (NeurIPS)*, 2018.

D. Dua and C. Graff. UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences, 2019. http://archive.ics.uci.edu/ml.

S. Dutta, D. Wei, H. Yueksel, P.-Y. Chen, S. Liu, and K. Varshney. Is there a trade-off between fairness and accuracy? A perspective using mismatched hypothesis testing. In *International Conference on Machine Learning (ICML)*, 2020.

C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science Conference (ITCS)*, 2012.

M. Ekstrand, M. Tian, I. M. Azpiazu, J. Ekstrand, O. Anuyah, D. McNeill, and M. S. Pera. All the cool kids, how do they fit in? Popularity and demographic biases in recommender evaluation and effectiveness. In *Conference on Fairness, Accountability, and Transparency (ACM FAT)*, 2018.

E. Fehrman, A. K. Muhammad, E. M. Mirkes, V. Egan, and A. N. Gorban. The five factor model of personality and evaluation of drug consumption risk. arXiv:1506.06297 [stat.AP], 2015. Dataset available on https://archive.ics.uci.edu/ml/datasets/Drug+consumption+(quantified).

S. Friedler, C. Scheidegger, and S. Venkatasubramanian. On the (im)possibility of fairness. arXiv:1609.07236 [cs.CY], 2016.

D. Gavinsky, S. Lovett, M. Saks, and S. Srinivasan. A tail bound for read-k families of functions. *Random Structures and Algorithms*, 47(1):99–108, 2014.

M. Ghadiri, S. Samadi, and S. Vempala. Socially fair k-means clustering. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021.

P. Gutiérrez, M. Pérez-Ortiz, J. Sánchez-Monedero, Francisco Fernández-Navarro, and C. Hervás-Martínez. Ordinal regression methods: Survey and experimental study. *IEEE Transactions on Knowledge and Data Engineering*, 28(1):127–146, 2016. Supplementary material and datasets available on http://www.uco.es/grupos/ayrna/orreview.

M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Neural Information Processing Systems (NeurIPS)*, 2016.

R. Herbrich, T. Graepel, and K. Obermayer. Large margin rank boundaries for ordinal regression. In *Advances in Large Margin Classifiers*, pages 115–132. MIT Press, 2000.

K. Johnson, D. Foster, and R. Stine. Impartial predictive modeling: Ensuring group fairness in arbitrary models. arXiv:1608.00528 [stat.ME], 2016.

N. Kallus and A. Zhou. The fairness of risk scores beyond classification: Bipartite ranking and the xAUC metric. In *Neural Information Processing Systems (NeurIPS)*, 2019.

F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33:1–33, 2011.

J. Kim, J. Chen, and A. Talwalkar. FACT: A diagnostic for group fairness trade-offs. In *International Conference on Machine Learning (ICML)*, 2020.

M. Kleindessner, P. Awasthi, and J. Morgenstern. Fair k-center clustering for data summarization. In *International Conference on Machine Learning (ICML)*, 2019a.

M. Kleindessner, S. Samadi, P. Awasthi, and J. Morgenstern. Guarantees for spectral clustering with fairness constraints. In *International Conference on Machine Learning (ICML)*, 2019b.

C. Kuhlman, M. VanValkenburg, and E. Rundensteiner. FARE: diagnostics for fair ranking using pairwise error metrics. In *International Conference on World Wide Web (WWW)*, 2019.

Y. Liu, A. Kong, and C. Goh. Deep ordinal regression based on data relationship for small datasets. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.

M. Lohaus, M. Perrot, and U. von Luxburg. Too relaxed to be fair. In *International Conference on Machine Learning (ICML)*, 2020.

K. Lum and W. Isaac. To predict and serve? *Significance*, 13(5):14–19, 2016.

K. Makhlouf, S. Zhioua, and C. Palamidessi. On the applicability of ML fairness notions. arXiv:2006.16745 [cs.LG], 2020.

P. McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(2):109–142, 1980.

H. Narasimhan, A. Cotter, M. Gupta, and S. Wang. Pairwise fairness for ranking and regression. In *AAAI Conference on Artificial Intelligence*, 2020.

Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua. Ordinal regression with multiple output CNN for age estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

A. O'Connell. *Logistic regression models for ordinal response variables*, volume 146. Sage Publications, 2006.

A. Pérez-Suay, V. Laparra, G. Mateo-García, J. Muñoz-Marí, L. Gómez-Chova, and G. Camps-Valls. Equity of attention: Amortizing individual fairness in rankings. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, 2017.

J. Pinto da Costa, H. Alonso, and J. Cardoso. The unimodal model for the classification of ordinal data. *Neural Networks*, 21(1):78–91, 2008.

L. Polania, D. Wang, and G. Fung. Ordinal regression using noisy pairwise comparisons for body mass index range estimation. In *IEEE Winter Conference on Applications of Computer Vision*, 2019.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Neural Information Processing Systems (NeurIPS)*, 2007.

M. Redmond and A. Baveja. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141(3):660–678, 2002. Dataset available on https://archive.ics.uci.edu/ml/datasets/communities+and+crime.

S. Samadi, U. Tantipongpipat, J. Morgenstern, M. Singh, and S. Vempala. The price of fair PCA: One extra dimension. In *Neural Information Processing Systems (NeurIPS)*, 2018.

J. Sánchez-Monedero, M. Pérez-Ortiz, P. Sáez, A. Gutiérrez, and C. Hervás-Martínez. Partial order label decomposition approaches for melanoma diagnosis. *Applied Soft Computing*, 64:341–355, 2018.

J. Sánchez-Monedero, P. Gutiérrez, and M. Pérez-Ortiz. ORCA: A Matlab/Octave toolbox for ordinal regression. *Journal of Machine Learning Research (JMLR)*, 20(125):1–5, 2019. https://github.com/ayrna/orca. Datasets available on http://www.uco.es/grupos/ayrna/orreview.

S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

A. Shashua and A. Levin. Ranking with large margin principle: Two approaches. In *Neural Information Processing Systems (NeurIPS)*, 2002.

A. Singh and T. Joachims. Fairness of exposure in rankings. In *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, 2018.

U.S. Department of Commerce, Bureau of the Census. Census of population and housing 1990 United States: Summary tape file 1a & 3a (computer files).

U.S. Department Of Commerce, Bureau of the Census Producer, Washington, DC and Inter-university Consortium for Political and Social Research Ann Arbor, Michigan, 1992.

U.S. Department of Justice, Bureau of Justice Statistics. Law enforcement management and administrative statistics (computer file).

U.S. Department of Justice, Federal Bureau of Investigation. Crime in the United States (computer file), 1995.

J. Vanschoren, J. van Rijn, B. Bischl, and L. Torgo. Openml: Networked science in machine learning. *SIGKDD Explorations*, 15(2):49–60, 2013. https://www.openml.org/.

R. Vogel, A. Bellet, and S. Clémencon. Learning fair scoring functions: Bipartite ranking under ROC-based fairness constraints. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

Y. Wu, L. Zhang, and X. Wu. On convexity and bounds of fairness-aware classification. In *The World Wide Web Conference (WWW)*, 2019.

M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research (JMLR)*, 20(75):1–42, 2019.

M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. Baeza-Yates. FA*IR: A fair top-k ranking algorithm. In *ACM Conference on Information and Knowledge Management (CIKM)*, 2017.

T. Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research (JMLR)*, 2:527–550, 2002.

APPENDIX

A PROOFS AND DETAILED EXPLANATIONS

A.1 Pairwise Equalized Odds

One might wonder whether there is also a pairwise analogue of the standard fairness notion of equalized odds (Hardt et al., 2016), which is a stricter notion than standard EO and requires that $\mathbb{P}[f(x)=\hat{y}|a=\tilde{a},y=\tilde{y}]=\mathbb{P}[f(x)=\hat{y}|a=\hat{a},y=\tilde{y}]$ for all $\tilde{a},\hat{a}\in\mathcal{A}$ and $\tilde{y},\hat{y}\in\mathcal{Y}$. When interpreting pairwise EO as a pairwise analogue of standard EO (cf. Section 2.1), we considered having a higher label, in a pair of points, as an analogue of $y=1$ being the preferred outcome in standard EO. To derive a pairwise analogue of standard equalized odds, we can consider having a non-higher label (i.e., a smaller or equal label), in a pair of points, as an analogue of $y=0$ being the unpreferred outcome in standard equalized odds for binary classification (with $y\in\{0,1\}$). We then require

$$\mathbb{P}[f(x_1)\leq f(x_2)|a_1=\tilde{a},a_2=\hat{a},y_1\leq y_2]=\mathbb{P}[f(x_1)\geq f(x_2)|a_1=\tilde{a},a_2=\hat{a},y_1\geq y_2] \tag{5}$$

in addition to (2). But (5) is the same as (2), except that the strict inequalities are replaced by non-strict ones.

A.2 Proofs of the Remarks Made in Section 2.1

- When $f(x)=i$, then $\mathbb{P}[f(x_1)>f(x_2)|a_1=\tilde{a},a_2=\hat{a}]=0$ and $\mathbb{P}[f(x_1)>f(x_2)|a_1=\tilde{a},a_2=\hat{a},y_1>y_2]=0$ for all $\tilde{a},\hat{a}\in\mathcal{A}$. It is clear that the perfect predictor $f(x)=y$ satisfies pairwise EO, but not necessarily pairwise DP.

- If $\mathbb{P}[f(x)=\tilde{y}|a=\tilde{a}]=\mathbb{P}[f(x)=\tilde{y}|a=\hat{a}]$ for all $\tilde{y}\in\mathcal{Y}$ and $\tilde{a},\hat{a}\in\mathcal{A}$, then

$$\begin{aligned}
\mathbb{P}[f(x_1)>f(x_2)|a_1=\tilde{a},a_2=\hat{a}] &= \sum_{i=1}^{k-1}\sum_{j=i+1}^{k}\mathbb{P}[f(x_1)=j,f(x_2)=i|a_1=\tilde{a},a_2=\hat{a}]\\
&= \sum_{i=1}^{k-1}\sum_{j=i+1}^{k}\mathbb{P}[f(x_1)=j|a_1=\tilde{a}]\mathbb{P}[f(x_2)=i|a_2=\hat{a}]\\
&= \sum_{i=1}^{k-1}\sum_{j=i+1}^{k}\mathbb{P}[f(x_1)=j|a_1=\hat{a}]\mathbb{P}[f(x_2)=i|a_2=\tilde{a}]\\
&= \sum_{i=1}^{k-1}\sum_{j=i+1}^{k}\mathbb{P}[f(x_1)=j,f(x_2)=i|a_1=\hat{a},a_2=\tilde{a}]\\
&= \mathbb{P}[f(x_1)>f(x_2)|a_1=\hat{a},a_2=\tilde{a}]\\
&= \mathbb{P}[f(x_1)<f(x_2)|a_1=\tilde{a},a_2=\hat{a}],
\end{aligned}$$

  which shows that standard DP implies pairwise DP.

  When $k=2$ and $\mathbb{P}[f(x_1)>f(x_2)|a_1=\tilde{a},a_2=\hat{a}]=\mathbb{P}[f(x_1)<f(x_2)|a_1=\tilde{a},a_2=\hat{a}]$, then $\mathbb{P}[f(x_1)=2,f(x_2)=1|a_1=\tilde{a},a_2=\hat{a}]=\mathbb{P}[f(x_1)=1,f(x_2)=2|a_1=\tilde{a},a_2=\hat{a}]$ and

$$\begin{aligned}
\mathbb{P}[f(x_1)=2|a_1=\tilde{a}]\cdot\mathbb{P}[f(x_2)=1|a_2=\hat{a}]=\mathbb{P}[f(x_1)=1|a_1=\tilde{a}]\cdot\mathbb{P}[f(x_2)=2|a_2=\hat{a}] \quad &\Leftrightarrow\\
\mathbb{P}[f(x_1)=2|a_1=\tilde{a}]\cdot(1-\mathbb{P}[f(x_2)=2|a_2=\hat{a}])=(1-\mathbb{P}[f(x_1)=2|a_1=\tilde{a}])\cdot\mathbb{P}[f(x_2)=2|a_2=\hat{a}] \quad &\Leftrightarrow\\
\mathbb{P}[f(x_1)=2|a_1=\tilde{a}]=\mathbb{P}[f(x_2)=2|a_2=\hat{a}], &
\end{aligned}$$

  which shows that for $k=2$ pairwise DP implies standard DP.

  Let $k=3$, $\mathcal{D}=\{(x_1,y_1,0),(x_2,y_2,1),(x_3,y_3,0)\}\subseteq\mathcal{X}\times[3]\times\{0,1\}$, and $f(x_1)=1$, $f(x_2)=2$, $f(x_3)=3$. Then f satisfies pairwise DP on $\mathcal{D}$, but not standard DP, which shows that for general k the two fairness notions are not equivalent.

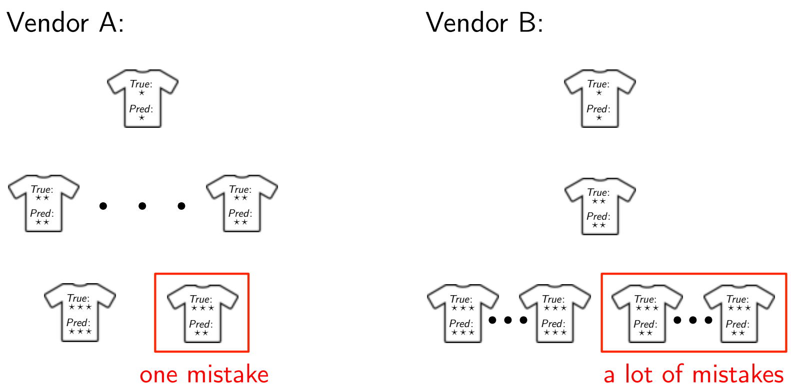

Figure 3: Example of a classification scenario in which we deem the predictions unfair to Vendor B. While the predictions satisfy the fairness notion of equalized odds (and hence also standard EO), our notion of pairwise EO is heavily violated.

- Let $k = 2$ and $\mathcal{D} = \{(x_1, 1, 0), (x_2, 1, 0), (x_3, 1, 1), (x_4, 2, 0), (x_5, 2, 1), (x_6, 2, 0)\} \subseteq \mathcal{X} \times [2] \times \{0, 1\}$. If $f(x_1) = 2$, $f(x_2) = 1$, $f(x_3) = 1$, $f(x_4) = 2$, $f(x_5) = 2$, $f(x_6) = 1$, then f satisfies pairwise EO, but not standard EO on $\mathcal{D}$. If $f(x_1) = 1$ instead of $f(x_1) = 2$, then f satisfies standard EO, but not pairwise EO on $\mathcal{D}$.

  If $k = 2$ and $\mathbb{P}[f(x) = 1 | a = \tilde{a}, y = \tilde{y}] = \mathbb{P}[f(x) = 1 | a = \hat{a}, y = \tilde{y}]$ for all $\tilde{a}, \hat{a} \in \mathcal{A}$ and $\tilde{y} \in \mathcal{Y} = \{0, 1\}$, then

$$\begin{aligned}
\mathbb{P}[f(x_1) > f(x_2) | a_1 = \tilde{a}, a_2 = \hat{a}, y_1 > y_2] &= \mathbb{P}[f(x_1) = 1, f(x_2) = 0 | a_1 = \tilde{a}, a_2 = \hat{a}, y_1 = 1, y_2 = 0] \\
&= \mathbb{P}[f(x_1) = 1 | a_1 = \tilde{a}, y_1 = 1] \cdot \mathbb{P}[f(x_2) = 0 | a_2 = \hat{a}, y_2 = 0] \\
&= \mathbb{P}[f(x_1) = 1 | a_1 = \hat{a}, y_1 = 1] \cdot \mathbb{P}[f(x_2) = 0 | a_2 = \tilde{a}, y_2 = 0] \\
&= \mathbb{P}[f(x_1) = 1, f(x_2) = 0 | a_1 = \hat{a}, a_2 = \tilde{a}, y_1 = 1, y_2 = 0] \\
&= \mathbb{P}[f(x_1) > f(x_2) | a_1 = \hat{a}, a_2 = \tilde{a}, y_1 > y_2] \\
&= \mathbb{P}[f(x_1) < f(x_2) | a_1 = \tilde{a}, a_2 = \hat{a}, y_1 < y_2],
\end{aligned}$$

  which shows that for $k = 2$ equalized odds implies pairwise EO.

  Let $k = 3$, $\mathcal{D} = \{(x_1, 1, 0), (x_2, 2, 0), (x_3, 3, 0), (x_4, 3, 0), (x_5, 1, 1), (x_6, 2, 1), (x_7, 2, 1), (x_8, 3, 1), (x_9, 3, 1)\} \subseteq \mathcal{X} \times [3] \times \{0, 1\}$, and $f(x_1) = 1$, $f(x_2) = 2$, $f(x_3) = 3$, $f(x_4) = 2$, $f(x_5) = 1$, $f(x_6) = 2$, $f(x_7) = 2$, $f(x_8) = 3$, $f(x_9) = 2$. Then f satisfies equalized odds on $\mathcal{D}$, but not pairwise EO. □

A.3 Detailed Example Comparing Standard Fairness Notions and our Pairwise Fairness Notions

Consider the scenario that there are two vendors selling T-shirts with ground-truth (quality) labels in $\{\star, \star\star, \star\star\star\}$, which we want to predict. Assume that Vendor A sells one T-shirt with a ground-truth label of one star, n T-shirts with a ground-truth label of two stars, and two T-shirts with a ground-truth label of three stars. Assume that Vendor B sells one T-shirt with a ground-truth label of one star, one T-shirt with a ground-truth label of two stars, and $2n$ T-shirts with a ground-truth label of three stars. Finally, assume that all our predictions are correct, except that for half of the 3-star T-shirts of each vendor we only predict two stars. A sketch of the scenario is provided in Figure 3. Our predictions satisfy equalized odds (and hence also standard EO), but our predictions appear to be very unfair to vendor B since they incorrectly downgrade a lot of Vendor B's high-quality T-shirts. Other than equalized odds, our notion of pairwise EO is heavily violated by our predictions, thus identifying them as unfair:

$$\mathbb{P}[\text{prediction for T-shirt } i > \text{prediction for T-shirt } j \mid \text{T-shirt } i \text{ from Vendor A, T-shirt } j \text{ from Vendor B,}$$
$$\text{ground-truth label of T-shirt } i > \text{ground-truth label of T-shirt } j] = \frac{n+3}{n+4} \to 1 \text{ (as } n \to \infty),$$
$$\mathbb{P}[\text{prediction for T-shirt } i < \text{prediction for T-shirt } j \mid \text{T-shirt } i \text{ from Vendor A, T-shirt } j \text{ from Vendor B,}$$
$$\text{ground-truth label of T-shirt } i < \text{ground-truth label of T-shirt } j] = \frac{1+2n+n^2}{1+2n+2n^2} \to \frac{1}{2} \text{ (as } n \to \infty),$$
$$\text{and EO-viol} \to \frac{1}{2}.$$

Our predictions neither satisfy standard DP nor pairwise DP. However, both of these fairness notions do not take the ground-truth labels and the fact that most of Vendor B's T-shirts are of higher quality than the ones of Vendor A into account, and hence they are not desirable in this scenario.

A.4 Convergence of $\text{Fair-viol}(f;\mathcal{D})$ to $\text{Fair-viol}(f;\mathbb{P})$

We only consider the case that Fair-viol = DP-viol. The case Fair-viol = EO-viol can be treated in an analogous way.

It is not hard to see that

$$|\,\text{DP-viol}(f;\mathbb{P}) - \text{DP-viol}(f;\mathcal{D})| \leq 2 \max_{\tilde{a},\hat{a}\in\mathcal{A}} |\mathbb{P}[f(x_1) > f(x_2)|a_1 = \tilde{a}, a_2 = \hat{a}] - \widehat{\mathbb{P}}_n[f(x_1) > f(x_2)|a_1 = \tilde{a}, a_2 = \hat{a}]|,$$

where $\widehat{\mathbb{P}}_n$ is the empirical distribution on $\mathcal{D}$, which is an i.i.d. sample from $\mathbb{P}$, and (x_1, y_1, a_1), (x_2, y_2, a_2) are independent samples from $\mathbb{P}$ and $\widehat{\mathbb{P}}_n$, respectively. Let $\mathcal{D} = ((x_i, y_i, a_i))_{i=1}^n$ and fix $\tilde{a}, \hat{a}$. Then

$$\widehat{\mathbb{P}}_n[f(x_1) > f(x_2)|a_1 = \tilde{a}, a_2 = \hat{a}] = \frac{\sum_{i,j\in[n]} \mathbb{1}[f(x_i) > f(x_j), a_i = \tilde{a}, a_j = \hat{a}]}{\sum_{i,j\in[n]} \mathbb{1}[a_i = \tilde{a}, a_j = \hat{a}]}.$$

The random variables $\mathbb{1}[f(x_i) > f(x_j), a_i = \tilde{a}, a_j = \hat{a}]_{i,j\in[n]}$ are a read-n family (Gavinsky et al., 2014, Definition 1). It is $\mathbb{P}[\mathbb{1}[f(x_i) > f(x_j), a_i = \tilde{a}, a_j = \hat{a}] = 1] = \mathbb{P}[f(x_i) > f(x_j), a_i = \tilde{a}, a_j = \hat{a}]$ for $i \neq j$ and $\mathbb{P}[\mathbb{1}[f(x_i) > f(x_j), a_i = \tilde{a}, a_j = \hat{a}] = 1] = 0$ for $i = j$. It follows from (Gavinsky et al., 2014, Theorem 1.1) that for any $\varepsilon > 0$, with probability $1 - 2e^{-2\varepsilon^2 n}$ over the sample $\mathcal{D}$ we have

$$\frac{\sum_{i,j\in[n]} \mathbb{1}[f(x_i) > f(x_j), a_i = \tilde{a}, a_j = \hat{a}]}{n^2} \leq \mathbb{P}[f(x_1) > f(x_2), a_1 = \tilde{a}, a_2 = \hat{a}] + \varepsilon$$

as well as

$$\frac{\sum_{i,j\in[n]} \mathbb{1}[f(x_i) > f(x_j), a_i = \tilde{a}, a_j = \hat{a}]}{n^2} \geq \mathbb{P}[f(x_1) > f(x_2), a_1 = \tilde{a}, a_2 = \hat{a}] - \frac{1}{n} - \varepsilon.$$

Similarly, we have with probability $1 - 2e^{-2\varepsilon^2 n}$ over the sample $\mathcal{D}$ that

$$\mathbb{P}[a_1 = \tilde{a}, a_2 = \hat{a}] - \varepsilon - \frac{1}{n} \leq \frac{\sum_{i,j\in[n]} \mathbb{1}[a_i = \tilde{a}, a_j = \hat{a}]}{n^2} \leq \mathbb{P}[a_1 = \tilde{a}, a_2 = \hat{a}] + \varepsilon.$$

It follows that if $\mathbb{P}[a_i = \tilde{a}, a_j = \hat{a}]$ is lower bounded by some positive constant, for any $\delta \in (0, 1)$, with probability $1 - \delta$ over the sample $\mathcal{D}$ we have

$$|\mathbb{P}[f(x_1) > f(x_2)|a_1 = \tilde{a}, a_2 = \hat{a}] - \widehat{\mathbb{P}}_n[f(x_1) > f(x_2)|a_1 = \tilde{a}, a_2 = \hat{a}]| \leq M\sqrt{\frac{\log\frac{1}{\delta}}{n}}.$$

for some constant M. This implies that with probability $1 - \delta$ over the sample $\mathcal{D}$ we have

$$\max_{\tilde{a},\hat{a}\in\mathcal{A}} |\mathbb{P}[f(x_1) > f(x_2)|a_1 = \tilde{a}, a_2 = \hat{a}] - \widehat{\mathbb{P}}_n[f(x_1) > f(x_2)|a_1 = \tilde{a}, a_2 = \hat{a}]| \leq M\sqrt{\frac{\log\frac{|\mathcal{A}|^2}{\delta}}{n}}.$$

A.5 Proof of Proposition 1

It is for all $\tilde{a}, \hat{a} \in \mathcal{A}$

$$\mathbb{P}_{(x',y',a')\sim\mathcal{D}'}[c_w(x') = 1 | a' = (\tilde{a}, \hat{a})] =$$

$$\frac{1}{|\{(x', y', a') \in \mathcal{D}' : a' = (\tilde{a}, \hat{a})\}|} \sum_{(x',y',a')\in\mathcal{D}':a'=(\tilde{a},\hat{a})} \mathbb{1}\{\operatorname{sgn}(w \cdot x') = 1\} =$$

$$\frac{\sum_{((x_i,y_i,a_i),(x_j,y_j,a_j))\in\mathcal{D}\times\mathcal{D}:\, y_i\neq y_j, a_i=\tilde{a}, a_j=\hat{a}} \mathbb{1}\{w \cdot (x_i - x_j) > 0\}}{|\{((x_i, y_i, a_i), (x_j, y_j, a_j)) \in \mathcal{D} \times \mathcal{D} :\; y_i \neq y_j, a_i = \tilde{a}, a_j = \hat{a}\}|} =$$

$$\mathbb{P}_{((x_i,y_i,a_i),(x_j,y_j,a_j))\sim\mathcal{D}^2}[s_w(x_i) > s_w(x_j) | a_i = \tilde{a}, a_j = \hat{a}, y_i \neq y_j],$$

which implies the statement for DP.

Similarly, for all $\tilde{a}, \hat{a} \in \mathcal{A}$ we have

$$\mathbb{P}_{(x',y',a')\sim\mathcal{D}'}[c_w(x') = 1 | a' = (\tilde{a}, \hat{a}), y' = 1] =$$

$$\frac{1}{|\{(x', y', a') \in \mathcal{D}' : a' = (\tilde{a}, \hat{a}), y' = 1\}|} \sum_{(x',y',a')\in\mathcal{D}':a'=(\tilde{a},\hat{a}),y'=1} \mathbb{1}\{\operatorname{sgn}(w \cdot x') = 1\} =$$

$$\frac{\sum_{((x_i,y_i,a_i),(x_j,y_j,a_j))\in\mathcal{D}\times\mathcal{D}:\, y_i> y_j, a_i=\tilde{a}, a_j=\hat{a}} \mathbb{1}\{w \cdot (x_i - x_j) > 0\}}{|\{((x_i, y_i, a_i), (x_j, y_j, a_j)) \in \mathcal{D} \times \mathcal{D} :\; y_i > y_j, a_i = \tilde{a}, a_j = \hat{a}\}|} =$$

$$\mathbb{P}_{((x_i,y_i,a_i),(x_j,y_j,a_j))\sim\mathcal{D}^2}[s_w(x_i) > s_w(x_j) | a_i = \tilde{a}, a_j = \hat{a}, y_i > y_j],$$

which implies the statement for EO. □

A.6 Proof of Proposition 2

We write $s_i = s(x_i)$, $i \in [n]$, for the values of the scoring function s on the training data $\mathcal{D} = ((x_i, y_i, a_i))_{i=1}^n$ and assume the values to be sorted, that is $s_i \leq s_{i+1}$, $i \in [n-1]$, and given, that is we do not take the time for evaluating s into account. Here we only consider the case of a binary protected attribute: let $\mathcal{A} = \{\tilde{a}, \hat{a}\}$, and let $\tilde{G} = \{i \in [n] : a_i = \tilde{a}\}$ and $\hat{G} = \{i \in [n] : a_i = \hat{a}\}$.

<u>When aiming for pairwise DP:</u>

We build a table $T \in (\mathbb{R}_{\geq 0} \cup \{\infty\})^{(n+1)\times k\times k\times(2\lfloor \frac{n^2}{4}\rfloor+1)}$ with

$$T(i, f, l, v) = \min_{\hat{y}\in\mathcal{H}_{i,f,l,v}} \sum_{j=1}^{i} C_{y_j,\hat{y}_j}$$

for $i \in \{0\} \cup [n]$, $f, l \in [k]$ and $v \in \{-\lfloor \frac{n^2}{4} \rfloor, \ldots, \lfloor \frac{n^2}{4} \rfloor\}$, where

$$\mathcal{H}_{i,f,l,v} = \Big\{\hat{y} = (\hat{y}_1, \ldots, \hat{y}_i) \in [k]^i : \hat{y} \text{ are predictions for } x_1, \ldots, x_i \text{ that are sorted, that is } y_r \leq y_{r+1} \text{ for } r \in [i-1],$$
$$\text{with } \hat{y}_i = l, \text{ take at most } f \text{ different values and satisfy } \hat{y}_r = \hat{y}_{r'} \text{ for } s_r = s_{r'}, \text{ and for which}$$
$$\sum_{1\leq r,r'\leq i:\, r\in\tilde{G}, r'\in\hat{G}} \left[\mathbb{1}\{\hat{y}_r > \hat{y}_{r'}\} - \mathbb{1}\{\hat{y}_r < \hat{y}_{r'}\}\right] = v\Big\}$$

and $T(i, f, l, v) = \infty$ if $\mathcal{H}_{i,f,l,v} = \emptyset$. Note that $\sum_{1\leq r,r'\leq i:\, r\in\tilde{G}, r'\in\hat{G}} 1 \leq |\tilde{G}| \cdot |\hat{G}| \leq \lfloor \frac{n^2}{4} \rfloor$. Given the table T, we can compute the optimal value of (4) as

$$\min_{l\in[k],v\in\{-\lfloor\frac{n^2}{4}\rfloor,\ldots,\lfloor\frac{n^2}{4}\rfloor\}} \frac{1}{n}\cdot T(n,k,l,v)+\frac{\lambda}{|\tilde{G}|\cdot|\hat{G}|}\cdot|v|.$$

It is

$$\begin{aligned}
&T(0,f,l,0)=0,\quad f,l\in[k],\qquad T(0,f,l,v)=\infty\quad f,l\in[k],v\neq 0,\\
&T(1,f,l,0)=C_{y_1,l}\quad f,l\in[k],\qquad T(1,f,l,v)=\infty\quad f,l\in[k],v\neq 0,\\
&T(i,1,l,0)=\sum_{j=1}^{i} C_{y_j,l}\quad i\in[n],l\in[k],\qquad T(i,1,l,v)=\infty\quad i\in[n],l\in[k],v\neq 0,\\
&T(i,f,1,0)=\sum_{j=1}^{i} C_{y_j,1}\quad i\in[n],f\in[k],\qquad T(i,f,1,v)=\infty\quad i\in[n],f\in[k],v\neq 0,
\end{aligned}$$

and

$$T(i,f,l,v)=\min_{i'<i,l'<l}\left[T(i',f-1,l',v^*)+\infty\cdot\mathbb{1}\{i'>0\wedge(s_{i'}=s_{i'+1})\}+\sum_{r=i'+1}^{i}C_{y_r,l}\right],$$

where $v^*=v-|\{i'+1,\ldots,i\}\cap\tilde{G}|\cdot|[i']\cap\hat{G}|+|\{i'+1,\ldots,i\}\cap\hat{G}|\cdot|[i']\cap\tilde{G}|$.

We can build T in time $\mathcal{O}(n^4k^3)$. If we store the minimizing (i',l') with $T(i,f,l,v)$, we can easily construct an optimal solution from T.

<u>When aiming for pairwise EO:</u>

We build a table $T\in(\mathbb{R}_{\geq 0}\cup\{\infty\})^{(n+1)\times k\times k\times(\lfloor\frac{n^2}{4}\rfloor+1)\times(\lfloor\frac{n^2}{4}\rfloor+1)}$ with

$$T(i,f,l,v_1,v_2)=\min_{\hat{y}\in\mathcal{H}_{i,f,l,v_1,v_2}}\sum_{j=1}^{i}C_{y_j,\hat{y}_j}$$

for $i\in\{0\}\cup[n]$, $f,l\in[k]$ and $v_1,v_2\in\{0\}\cup[\lfloor\frac{n^2}{4}\rfloor]$, where

$$\begin{aligned}
\mathcal{H}_{i,f,l,v_1,v_2}=\Big\{&\hat{y}=(\hat{y}_1,\ldots,\hat{y}_i)\in[k]^i:\hat{y}\text{ are predictions for }x_1,\ldots,x_i\text{ that are sorted, that is }y_r\leq y_{r+1}\text{ for}\\
&r\in[i-1]\text{, with }\hat{y}_i=l\text{, take at most }f\text{ different values and satisfy }\hat{y}_r=\hat{y}_{r'}\text{ for }s_r=s_{r'}\text{, and for which}\\
&\sum_{1\leq r,r'\leq i:\,r\in\tilde{G},r'\in\hat{G}}\mathbb{1}\{\hat{y}_r\geq\hat{y}_{r'}\wedge y_r<y_{r'}\}=v_1\text{ and }\sum_{1\leq r,r'\leq i:\,r\in\tilde{G},r'\in\hat{G}}\mathbb{1}\{\hat{y}_r\leq\hat{y}_{r'}\wedge y_r>y_{r'}\}=v_2\Big\}
\end{aligned}$$

and $T(i,f,l,v)=\infty$ if $\mathcal{H}_{i,f,l,v_1,v_2}=\emptyset$. Note that, just as before, $\sum_{1\leq r,r'\leq i:\,r\in\tilde{G},r'\in\hat{G}}1\leq\lfloor\frac{n^2}{4}\rfloor$. Given the table T, we can compute the optimal value of (4) as

$$\min_{l\in[k],v_1,v_2\in\{0\}\cup[\lfloor\frac{n^2}{4}\rfloor]}\frac{1}{n}\cdot T(n,k,l,v_1,v_2)+\lambda\cdot\left|\frac{v_1}{|\{(r,r')\in\tilde{G}\times\hat{G}:y_r<y_{r'}\}|}-\frac{v_2}{|\{(r,r')\in\tilde{G}\times\hat{G}:y_r>y_{r'}\}|}\right|.$$

It is for all $i \in [n]$ and $f, l \in [k]$

$$
\begin{aligned}
&T(0,f,l,0,0) = 0, \qquad T(0,f,l,v_1,v_2) = \infty, \quad (v_1,v_2) \neq (0,0),\\
&T(1,f,l,0,0) = C_{y_1,l}, \qquad T(1,f,l,v_1,v_2) = \infty, \quad (v_1,v_2) \neq (0,0),
\end{aligned}
$$

$$
T(i,1,l,v_1,v_2) = \begin{cases} \sum_{j=1}^{i} C_{y_j,l} & (v_1,v_2) = \\ & \quad \left(\sum_{1\le r,r'\le i:\, r\in\tilde{G}, r'\in\hat{G}} \mathbb{1}\{y_r < y_{r'}\}, \sum_{1\le r,r'\le i:\, r\in\tilde{G}, r'\in\hat{G}} \mathbb{1}\{\hat{y}_r \ge \hat{y}_{r'} \wedge y_r > y_{r'}\}\right), \\ \infty & \text{else} \end{cases}
$$

$$
T(i,f,1,v_1,v_2) = \begin{cases} \sum_{j=1}^{i} C_{y_j,1} & (v_1,v_2) = \\ & \quad \left(\sum_{1\le r,r'\le i:\, r\in\tilde{G}, r'\in\hat{G}} \mathbb{1}\{y_r < y_{r'}\}, \sum_{1\le r,r'\le i:\, r\in\tilde{G}, r'\in\hat{G}} \mathbb{1}\{\hat{y}_r \ge \hat{y}_{r'} \wedge y_r > y_{r'}\}\right) \\ \infty & \text{else} \end{cases}
$$

and

$$
T(i,f,l,v_1,v_2) = \min_{i'<i, l'<l} \left[T(i', f-1, l', v_1^*, v_2^*) + \infty \cdot \mathbb{1}\{i' > 0 \wedge (s_{i'} = s_{i'+1})\} + \sum_{r=i'+1}^{i} C_{y_r,l} \right],
$$

where

$$
\begin{aligned}
v_1^* &= v_1 - |\{(r,r') \in (\tilde{G} \cap \{i'+1,\dots,i\}) \times (\hat{G} \cap [i]) : y_a < y_b\}|,\\
v_2^* &= v_2 - |\{(r,r') \in (\tilde{G} \cap [i]) \times (\hat{G} \cap \{i'+1,\dots,i\}) : y_a > y_b\}|.
\end{aligned}
$$

Using two helper tables $H_1, H_2 \in (\mathbb{N} \cup \{0\})^{n\times n}$ with

$$
\begin{aligned}
H_1(i,j) &= |\{(r,r') \in (\tilde{G} \cap [i]) \times (\hat{G} \cap [j]) : y_r < y_{r'}\}|,\\
H_2(i,j) &= |\{(r,r') \in (\tilde{G} \cap [i]) \times (\hat{G} \cap [j]) : y_r > y_{r'}\}|,
\end{aligned}
$$

which we can certainly build in time $\mathcal{O}(n^6)$, we can build T in time $\mathcal{O}(n^6k^3)$. If we store the minimizing (i', l') with $T(i,f,l,v)$, we can easily construct an optimal solution from T. □

A.7 Addendum to Section 3.3

We first provide the statement for EO-viol.

Under the same assumptions as in Theorem 1, but instead of $\mathbb{P}[a = \tilde{a}] \ge \beta > 0$ for all $\tilde{a} \in \mathcal{A}$ requiring that $\mathbb{P}[a = \tilde{a}, y = j'] \ge \beta > 0$ for all $\tilde{a} \in \mathcal{A}, j' \in [k]$, our learned predictor $f = f(\cdot; s, \boldsymbol{\theta})$ satisfies with probability at least $1-\delta$ over the training sample $\mathcal{D}$ of size n, for n sufficiently large, the bound on $\mathrm{MAE}(f;\mathbb{P})$ from Theorem 1 and

$$
|\,\mathrm{EO\text{-}viol}(f;\mathbb{P}) - \mathrm{EO\text{-}viol}(f;\mathcal{D})| \le Mk^4 \sqrt{\frac{d + \log\frac{M|\mathcal{A}|^2k}{\delta}}{\left(1 - \sqrt{\frac{2\log\frac{M|\mathcal{A}|^2k}{\delta}}{n\beta}}\right) n\beta}} + \frac{M}{\beta^2}\sqrt{\frac{\log\frac{M|\mathcal{A}|^2k}{\delta}}{2n}}.
$$

We prove the bounds on MAE, DP-viol and EO-viol separately. Theorem 1 and the statement above then follow from a simple union bound.

- **Bound on** $\mathrm{MAE}(f;\mathbb{P})$

  According to Theorem 4 of Zhang (2002), the covering number $\mathcal{N}_\infty(\gamma, \mathcal{F}, m)$, where $\mathcal{F}$ is the class of linear scoring functions $s_w(x) = w \cdot x$ with $\|w\|_2 \le \sqrt{\nu}$ and domain $B_R(0)$, satisfies

$$
\log_2 \mathcal{N}_\infty(\gamma, \mathcal{F}, m) \le 36\frac{R^2\nu}{\gamma^2} \log_2\left(2\left\lceil \frac{4R\sqrt{\nu}}{\gamma} + 2\right\rceil m + 1\right).
$$

Furthermore, since $|s_w(x)| \leq \|w\|_2 \cdot \|x\|_2$, our strategy learns thresholds in $[-R\sqrt{\nu}, +R\sqrt{\nu}]$. The bound on $\mathrm{MAE}(f;\mathbb{P})$ is now a simple corollary of Theorem 7 of Agarwal (2008).

The empirical γ-margin loss $\widehat{L}^{\gamma}_{\mathcal{D}}(s,\boldsymbol{\theta})$, with $\mathcal{D} = ((x_i, y_i, a_i))_{i=1}^n$, is defined as

$$\widehat{L}^{\gamma}_{\mathcal{D}}(s,\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} l_\gamma(s,\boldsymbol{\theta},(x_i,y_i))$$

with

$$l_\gamma(s,\boldsymbol{\theta},(x,y)) = \sum_{j=1}^{k-1} \mathbb{1}\{y^j(s(x)-\theta_j) \leq \gamma\},$$

where

$$y^j = \begin{cases} +1 & \text{if } j \in \{1,\dots,y-1\} \\ -1 & \text{if } j \in \{y,\dots,k-1\} \end{cases}.$$

- **Bound on** $\text{DP-viol}(f;\mathbb{P})$

  It is not hard to see that

$$|\text{DP-viol}(f;\mathbb{P}) - \text{DP-viol}(f;\mathcal{D})| \leq \\ 2\max_{\tilde{a},\hat{a}\in\mathcal{A}} |\mathbb{P}[f(x_1) > f(x_2)|a_1=\tilde{a}, a_2=\hat{a}] - \widehat{\mathbb{P}}_n[f(x_1) > f(x_2)|a_1=\tilde{a}, a_2=\hat{a}]|,$$

where $\widehat{\mathbb{P}}_n$ is the empirical distribution on $\mathcal{D}$. For now, let us consider fixed $\tilde{a}, \hat{a}$. We have

$$\mathbb{P}[f(x_1) > f(x_2)|a_1=\tilde{a}, a_2=\hat{a}] = \sum_{i=1}^{k-1}\sum_{j=i+1}^{k} \mathbb{P}[f(x_1)=j|a_1=\tilde{a}]\cdot\mathbb{P}[f(x_2)=i|a_2=\hat{a}]$$

and

$$\widehat{\mathbb{P}}_n[f(x_1) > f(x_2)|a_1=\tilde{a}, a_2=\hat{a}] = \sum_{i=1}^{k-1}\sum_{j=i+1}^{k}\left[\frac{1}{\tilde{n}}\sum_{s=1}^{\tilde{n}}\mathbb{1}[f(\tilde{x}_s)=j]\cdot\frac{1}{\hat{n}}\sum_{t=1}^{\hat{n}}\mathbb{1}[f(\hat{x}_t)=i]\right],$$

where we write $(\tilde{x}_i, \tilde{y}_i, \tilde{a}_i)_{i=1}^{\tilde{n}}$ for those training sample points with $a_i = \tilde{a}$ and $(\hat{x}_i, \hat{y}_i, \hat{a}_i)_{i=1}^{\hat{n}}$ for those with $a_i = \hat{a}$.

We assume f to be a threshold model with a linear scoring function $s(x) = w\cdot x$, that is

$$f(x) = j \quad\Leftrightarrow\quad \theta_{j-1} < w\cdot x \leq \theta_j.$$

Let $\mathbb{P}^{\tilde{mod}}$ be a distribution on $\mathcal{X}\times\{0,1\}$, corresponding to random variables x and y, with $\mathbb{P}^{\tilde{mod}}[y=0]=1$ and $\mathbb{P}^{\tilde{mod}}[x] = \mathbb{P}[x|a=\tilde{a}]$. Note that $(\tilde{x}_i, 0)_{i=1}^{\tilde{n}}$ is a sample from $\mathbb{P}^{\tilde{mod}}$. We have

$$\mathbb{P}[f(x)=j|a=\tilde{a}] = \mathbb{P}^{\tilde{mod}}[\mathbb{1}[\theta_{j-1} < w\cdot x \leq \theta_j]=1] = \mathbb{E}_{(x,y)\sim\mathbb{P}^{\tilde{mod}}}[\text{01-loss}(y, \mathbb{1}[\theta_{j-1} < w\cdot x \leq \theta_j])]$$

and

$$\frac{1}{\tilde{n}}\sum_{s=1}^{\tilde{n}}\mathbb{1}[f(\tilde{x}_s)=j] = \frac{1}{\tilde{n}}\sum_{s=1}^{\tilde{n}}\mathbb{1}[\theta_{j-1} < w\cdot\tilde{x}_s \leq \theta_j] = \frac{1}{\tilde{n}}\sum_{s=1}^{\tilde{n}}\text{01-loss}(0, \mathbb{1}[\theta_{j-1} < w\cdot\tilde{x}_s \leq \theta_j]).$$

We assume $x \in \mathbb{R}^d$. The class of halfspace functions $\{x \mapsto \mathbb{1}[w\cdot x \geq b] : w\in\mathbb{R}^d, b\in\mathbb{R}\}$ has VC-dimension $d+1$ (e.g., Shalev-Shwartz and Ben-David, 2014, Theorem 9.3), and it follows from a theorem in Blumer et al. (1989) (also stated as Theorem A in Csikos et al., 2019) that there exists a constant M such that the class of functions $\{x \mapsto \mathbb{1}[\theta_{j-1} < w \cdot x \leq \theta_j] : w \in \mathbb{R}^d, \theta_{j-1}, \theta_j \in \mathbb{R}\}$ has VC-dimension at most Md (in the following, we write M for an absolute constant that may change from line to line). According to the fundamental theorem of statistical learning (e.g., Shalev-Shwartz and Ben-David, 2014, Theorem 6.8), for fixed $\delta \in (0,1)$, we have with probability of at least $1-\delta$ over the sample $(\tilde{x}_i, \tilde{y}_i, \tilde{a}_i)_{i=1}^{\tilde{n}}$ (with $\tilde{n}$ fixed)

$$\sup_{w\in\mathbb{R}^d,\theta_{j-1},\theta_j\in\mathbb{R}} \left| \mathbb{P}[f(x) = j | a = \tilde{a}] - \frac{1}{\tilde{n}} \sum_{s=1}^{\tilde{n}} \mathbb{1}[f(\tilde{x}_s) = j] \right| \leq M \sqrt{\frac{d + \log\frac{1}{\delta}}{\tilde{n}}}.$$

Assuming that $\mathbb{P}[a = \tilde{a}] \geq \beta$ it follows from Chernoff's bound and a simple union bound that for any $\delta \in (0,1)$ with probability at least $1-2\delta$ over the sample $(x_i, y_i, a_i)_{i=1}^{n}$ we have

$$\tilde{n} \geq \left(1 - \sqrt{\frac{2\log\frac{1}{\delta}}{n\beta}}\right) n\beta \qquad \text{and}$$

$$\forall j: \sup_{w\in\mathbb{R}^d,\theta_{j-1},\theta_j\in\mathbb{R}} \left| \mathbb{P}[f(x) = j | a = \tilde{a}] - \frac{1}{\tilde{n}} \sum_{s=1}^{\tilde{n}} \mathbb{1}[f(\tilde{x}_s) = j] \right| \leq M \sqrt{\frac{d + \log\frac{1}{\delta}}{\left(1 - \sqrt{\frac{2\log\frac{1}{\delta}}{n\beta}}\right) n\beta}}.$$

Assuming that $\mathbb{P}[a = \tilde{a}] \geq \beta$ for all $\tilde{a} \in \mathcal{A}$, it follows that with probability $1 - 2|\mathcal{A}|\delta$ over the sample $(x_i, y_i, a_i)_{i=1}^{n}$ we have

$$\forall \tilde{a} \in \mathcal{A}, \forall j: \sup_{w\in\mathbb{R}^d,\theta_{j-1},\theta_j\in\mathbb{R}} \left| \mathbb{P}[f(x) = j | a = \tilde{a}] - \frac{1}{\tilde{n}} \sum_{s=1}^{\tilde{n}} \mathbb{1}[f(\tilde{x}_s) = j] \right| \leq M \sqrt{\frac{d + \log\frac{1}{\delta}}{\left(1 - \sqrt{\frac{2\log\frac{1}{\delta}}{n\beta}}\right) n\beta}}.$$

Since $|uv - \hat{u}\hat{v}| \leq 3\max\{|u - \hat{u}|, |v - \hat{v}|\}$ for $u, v, \hat{u}, \hat{v} \in [0,1]$, it follows that with probability $1 - 2|\mathcal{A}|\delta$ we have

$$\max_{\tilde{a},\hat{a}\in\mathcal{A}} \sup_f |\mathbb{P}[f(x_1) > f(x_2) | a_1 = \tilde{a}, a_2 = \hat{a}] - \widehat{\mathbb{P}}_n[f(x_1) > f(x_2) | a_1 = \tilde{a}, a_2 = \hat{a}]| \leq Mk^2 \sqrt{\frac{d + \log\frac{1}{\delta}}{\left(1 - \sqrt{\frac{2\log\frac{1}{\delta}}{n\beta}}\right) n\beta}}.$$

This implies that with probability $1-\delta$ over the sample $(x_i, y_i, a_i)_{i=1}^{n}$ we have

$$\sup_f |\operatorname{DP-viol}(f;\mathbb{P}) - \operatorname{DP-viol}(f;\mathcal{D})| \leq Mk^2 \sqrt{\left(d + \log\frac{2|\mathcal{A}|}{\delta}\right) \cdot \left[\left(1 - \sqrt{\frac{2\log\frac{2|\mathcal{A}|}{\delta}}{n\beta}}\right) n\beta\right]^{-1}}.$$

In particular, this provides a bound on DP-viol$(f;\mathbb{P})$ for the threshold model f learned by our strategy.

- **Bound on** EO-viol$(f;\mathbb{P})$

  The bound on EO-viol$(f;\mathbb{P})$ can be derived in a similar way as the bound on DP-viol$(f;\mathbb{P})$, and we refrain from presenting the details here.

A.8 Proof of Lemma 1

We choose S to consist only of the two projections onto the first and second coordinate, respectively, that is $S = \{p_1 : \mathbb{R}^2 \to \mathbb{R} \text{ with } p_1(u_1, u_2) = u_1, p_2 : \mathbb{R}^2 \to \mathbb{R} \text{ with } p_2(u_1, u_2) = u_2\}$.

1.

- Aiming for pairwise DP:

Consider $n = 2m+1$ many datapoints $(x_1, y_1, a_1) \ldots, (x_n, y_n, a_n) \in \mathbb{R}^2 \times [4] \times \{0,1\}$ with

$$x_i = \begin{cases} (1,3) & i = 1, \\ (i,2) & 2 \leq i \leq m+1 \\ (i,4) & m+2 \leq i \leq 2m+1 \end{cases}, \qquad y_i = \begin{cases} 1 & i = 1, \\ 2 & 2 \leq i \leq m+1, \\ 4 & m+2 \leq i \leq 2m+1 \end{cases}, \qquad a_i = \begin{cases} 0 & i = 1, \\ 1 & 2 \leq i \leq 2m+1 \end{cases}.$$

The scoring function p_1 does not suffer from any label flips whereas the scoring function p_2 incurs label flips for m many datapoint pairs. However, p_2 satisfies pairwise DP whereas p_1 does not. When choosing thresholds that minimize the MAE under the constraint that the resulting predictor f has to satisfy pairwise DP, when using p_1 as scoring function, we must choose thresholds such that $f(x_i) = 2$ for all $i \in \mathbb{N}$, and if we use p_1, we obtain MAE $= 1$. If we use p_2 as scoring function, we can choose thresholds $\theta_1 = 0, \theta_2 = 2.5, \theta_3 = 3.5$ and obtain a predictor f with MAE $= \frac{2}{2m+1}$.

- Aiming for pairwise EO:
  Consider $n = 2m+5$ many datapoints $(x_1, y_1, a_1) \ldots, (x_n, y_n, a_n) \in \mathbb{R}^2 \times [4] \times \{0,1\}$ with

$$\begin{aligned} &(x_1, y_1, a_1) = ((1,1),1,0), \qquad (x_2, y_2, a_2) = ((2,3),1,1), \qquad (x_3, y_3, a_3) = ((3,2),2,0), \\ &(x_4, y_4, a_4) = ((4,4),2,1), \qquad (x_5, y_5, a_5) = ((5,n+1),4,1) \\ &(x_i, a_i, y_i) = ((i,i),3,1), \quad 6 \leq i \leq m+5, \qquad (x_i, a_i, y_i) = ((i,i),4,1), \quad m+6 \leq i \leq 2m+5. \end{aligned}$$

  The scoring function p_1 incurs label flips for m many datapoint pairs whereas the scoring function p_2 incurs label flips for only one datapoint pair. However, p_1 satisfies pairwise EO whereas p_2 does not. When choosing thresholds that minimize the MAE under the constraint that the resulting predictor f has to satisfy pairwise EO, when using p_2 as scoring function, we have to choose thresholds such that $f(x_i) = 3$ for all $i \in \mathbb{N}$, and if we use p_2, we obtain MAE $= \frac{6+m}{2m+5} > \frac{1}{2}$. If we use p_1 as scoring function, we can choose thresholds $\theta_1 = 2.5, \theta_2 = 4.5, \theta_3 = m+5.5$ and obtain a predictor f with MAE $= \frac{1}{n}$.

2.

- Aiming for pairwise DP:
  Consider six datapoints $(x_1, y_1, a_1) \ldots, (x_6, y_6, a_6) \in \mathbb{R}^2 \times [4] \times \{0,1\}$ with

$$\begin{aligned} &(x_1, y_1, a_1) = ((1,0),1,0), \qquad (x_2, y_2, a_2) = ((2,0),1,0), \qquad (x_3, y_3, a_3) = ((3,1),1,1), \\ &(x_4, y_4, a_4) = ((4,1),1,1), \qquad (x_5, y_5, a_5) = ((5,0),2,0), \qquad (x_6, y_6, a_6) = ((6,0),2,0). \end{aligned}$$

  The scoring function p_1 satisfies pairwise DP whereas p_2 does not. Using p_1, we can choose thresholds $\theta_1 = 4.5, \theta_2 = \theta_3 = 7$ in order to obtain a predictor with MAE $= 0$. However, for this predictor we have DP-viol $= 0.5$.

- Aiming for pairwise EO:
  Consider six datapoints $(x_1, y_1, a_1) \ldots, (x_6, y_6, a_6) \in \mathbb{R}^2 \times [4] \times \{0,1\}$ with

$$\begin{aligned} &(x_1, y_1, a_1) = ((1,0),1,0), \qquad (x_2, y_2, a_2) = ((2,1),2,1), \qquad (x_3, y_3, a_3) = ((3,1),1,1), \\ &(x_4, y_4, a_4) = ((4,1),1,1), \qquad (x_5, y_5, a_5) = ((5,0),2,0), \qquad (x_6, y_6, a_6) = ((6,0),2,0). \end{aligned}$$

  The scoring function p_1 satisfies pairwise EO whereas p_2 does not. Using p_1, if we choose thresholds that minimize the MAE, we can choose $\theta_1 = 4.5, \theta_2 = \theta_3 = 7$ and obtain a predictor with MAE $= \frac{1}{6}$. However, for this predictor we have EO-viol $= 1$.

$\square$

A.9 Proof of Lemma 2

Let

$$\begin{aligned}
A &= \mathbb{P}[s(x_1) > s(x_2) | a_1 = \tilde{a}, a_2 = \hat{a}],\\
B &= \mathbb{P}[s(x_1) < s(x_2) | a_1 = \tilde{a}, a_2 = \hat{a}],\\
C &= \mathbb{P}[s(x_1) = s(x_2) | a_1 = \tilde{a}, a_2 = \hat{a}],\\
D &= \mathbb{P}[f(x_1) > f(x_2) | a_1 = \tilde{a}, a_2 = \hat{a}],\\
E &= \mathbb{P}[f(x_1) < f(x_2) | a_1 = \tilde{a}, a_2 = \hat{a}].
\end{aligned}$$

It is $A + B + C = 1$ and $|A - B| \leq \text{DP-viol}(s)$. It follows that

$$A, B \in \left[\frac{1 - C - \text{DP-viol}(s)}{2}, \frac{1 - C + \text{DP-viol}(s)}{2}\right].$$

Since $f(x_1) > f(x_2)$ implies that $s(x_1) > s(x_2)$ and $f(x_1) < f(x_2)$ implies that $s(x_1) < s(x_2)$, we have $D \leq A$ and $E \leq B$. Hence,

$$D, E \in \left[0, \frac{1 - C + \text{DP-viol}(s)}{2}\right]$$

and

$$|D - E| \leq \frac{1 - C + \text{DP-viol}(s)}{2} \leq \frac{1}{2} + \frac{\text{DP-viol}(s)}{2}.$$

□

A.10 Performing Local Search to Minimize (4)

We first sort the datapoints such that $s(x_j) \leq s(x_{j+1})$, $j \in [n-1]$. Including the time it takes to evaluate s, this can be done in time $\mathcal{O}(nd + n \log n)$. We can then perform a local search in order to find a local minimum of (4) as follows: given thresholds $\theta_1 \leq \ldots \leq \theta_{k-1}$, in each round, we move one threshold θ_i to the left or to the right (within $[\theta_{i-1}, \theta_{i+1}]$ and thus maintaining the order of the thresholds) such that we decrease the value of (4) as much as possible. Since the value of (4) does not depend on the exact location of θ_i within $[s(x_j), s(x_{j+1}))$, it is enough to consider moving θ_i to a position in $[\theta_{i-1}, \theta_i) \cap \{s(x_j) : j \in [n]\}$ or $(\theta_i, \theta_{i+1}] \cap \{s(x_j) : j \in [n]\}$.

<u>When aiming for pairwise DP:</u>

For thresholds $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{k-1})$, set $\theta_0 = -\infty$ and $\theta_k = +\infty$, and let for $i \in [k]$ and $\tilde{a}, \hat{a} \in \mathcal{A}$ with $\tilde{a} \neq \hat{a}$

$$\begin{aligned}
T(i, \tilde{a}; \boldsymbol{\theta}) &= |\{j \in [n] : a_j = \tilde{a}, s(x_j) \in (\theta_{i-1}, \theta_i]\}|,\\
P(\tilde{a}, \hat{a}; \boldsymbol{\theta}) &= |\{(j, l) \in [n]^2 : a_j = \tilde{a}, a_l = \hat{a}, f(x_j; s, \boldsymbol{\theta}) > f(x_l; s, \boldsymbol{\theta})\}|,\\
Q(\tilde{a}, \hat{a}; \boldsymbol{\theta}) &= |\{(j, l) \in [n]^2 : a_j = \tilde{a}, a_l = \hat{a}, f(x_j; s, \boldsymbol{\theta}) < f(x_l; s, \boldsymbol{\theta})\}|,
\end{aligned}$$

where $f(\cdot\,; s, \boldsymbol{\theta})$ denotes the threshold model-predictor with scoring function s and thresholds $\boldsymbol{\theta}$. The value of the objective function (4) using thresholds $\boldsymbol{\theta}$ is

$$\begin{aligned}
\text{Obj}(\boldsymbol{\theta}) &= \text{Cost}(\boldsymbol{\theta}) + \lambda \cdot \text{DP-viol}(f(\cdot\,; s, \boldsymbol{\theta}); \mathcal{D})\\
&= \frac{1}{n} \sum_{i=1}^{n} C_{y_i, f(x_i; s, \boldsymbol{\theta})} + \lambda \cdot \max_{\tilde{a} \neq \hat{a}} \frac{|P(\tilde{a}, \hat{a}; \boldsymbol{\theta}) - Q(\tilde{a}, \hat{a}; \boldsymbol{\theta})|}{|\{j \in [n] : a_j = \tilde{a}\}| \cdot |\{j \in [n] : a_j = \hat{a}\}|}.
\end{aligned}$$

Given some initial thresholds $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{k-1})$, we can compute $T(i, \tilde{a}; \boldsymbol{\theta})$, $P(\tilde{a}, \hat{a}; \boldsymbol{\theta})$, $Q(\tilde{a}, \hat{a}; \boldsymbol{\theta})$ and $\text{Obj}(\boldsymbol{\theta})$ in time $\mathcal{O}(n + k|\mathcal{A}|^2)$ (we assume $k \leq n$ and $|\mathcal{A}| \leq n$). We can then compute the best possible right-move in time $\mathcal{O}(n|\mathcal{A}|^2)$ by going through the sorted array of scores $(s(x_j))_{j=1}^n$ once and using the following: if θ_{i_0} (for some

$i_0 \in [k-1]$) is located at $s(x_j)$ and we move it to $s(x_{j+1})$, thus yielding new thresholds $\boldsymbol{\theta}'$, we have

$$\begin{aligned}
T(i_0, a_{j+1}; \boldsymbol{\theta}') &= T(i_0, a_{j+1}; \boldsymbol{\theta}) + 1, \\
T(i_0+1, a_{j+1}; \boldsymbol{\theta}') &= T(i_0+1, a_{j+1}; \boldsymbol{\theta}) - 1, \\
T(i, \tilde{a}; \boldsymbol{\theta}') &= T(i, \tilde{a}; \boldsymbol{\theta}) \quad \text{for} \quad (i, \tilde{a}) \notin \{(i_0, a_{j+1}), (i_0+1, a_{j+1})\}, \\
P(\tilde{a}, \hat{a}; \boldsymbol{\theta}') &= \begin{cases} P(\tilde{a}, \hat{a}; \boldsymbol{\theta}) - T(i_0, \hat{a}; \boldsymbol{\theta}) & \text{if } a_{j+1} = \tilde{a} \\ P(\tilde{a}, \hat{a}; \boldsymbol{\theta}) + T(i_0+1, \tilde{a}; \boldsymbol{\theta}) & \text{if } a_{j+1} = \hat{a} \\ P(\tilde{a}, \hat{a}; \boldsymbol{\theta}) & \text{else} \end{cases}, \\
Q(\tilde{a}, \hat{a}; \boldsymbol{\theta}') &= \begin{cases} Q(\tilde{a}, \hat{a}; \boldsymbol{\theta}) + T(i_0+1, \hat{a}; \boldsymbol{\theta}) & \text{if } a_{j+1} = \tilde{a} \\ Q(\tilde{a}, \hat{a}; \boldsymbol{\theta}) - T(i_0, \tilde{a}; \boldsymbol{\theta}) & \text{if } a_{j+1} = \hat{a} \\ Q(\tilde{a}, \hat{a}; \boldsymbol{\theta}) & \text{else} \end{cases}, \\
\mathrm{Obj}(\boldsymbol{\theta}') &= \mathrm{Cost}(\boldsymbol{\theta}) - \frac{1}{n} \cdot C_{y_{j+1}, i+1} + \frac{1}{n} \cdot C_{y_{j+1}, i} + \lambda \cdot \max_{\tilde{a} \neq \hat{a}} \frac{|P(\tilde{a}, \hat{a}; \boldsymbol{\theta}') - Q(\tilde{a}, \hat{a}; \boldsymbol{\theta}')|}{|\{j \in [n] : a_j = \tilde{a}\}| \cdot |\{j \in [n] : a_j = \hat{a}\}|}.
\end{aligned}$$

Similarly, we can compute the best possible left-move in time $\mathcal{O}(n|\mathcal{A}|^2)$. Some care has to be taken when there are datapoints x_j, x_{j+1} with $s(x_j) = s(x_{j+1})$ since we cannot move a threshold to $s(x_j)$ then. Clearly, we have $\mathcal{O}(n + k|\mathcal{A}|^2) \subseteq \mathcal{O}(n|\mathcal{A}|^2)$.

When aiming for pairwise EO:

We can proceed similarly as in case of pairwise DP. For thresholds $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{k-1})$, set $\theta_0 = -\infty$ and $\theta_k = +\infty$, and let for $i, i' \in [k]$ and $\tilde{a}, \hat{a} \in \mathcal{A}$ with $\tilde{a} \neq \hat{a}$

$$\begin{aligned}
T(i, i', \tilde{a}; \boldsymbol{\theta}) &= |\{j \in [n] : a_j = \tilde{a}, y_j = i', s(x_j) \in (\theta_{i-1}, \theta_i]\}|, \\
P(\tilde{a}, \hat{a}; \boldsymbol{\theta}) &= |\{(j, l) \in [n]^2 : a_j = \tilde{a}, a_l = \hat{a}, y_j > y_l, f(x_j; s, \boldsymbol{\theta}) > f(x_l; s, \boldsymbol{\theta})\}|, \\
Q(\tilde{a}, \hat{a}; \boldsymbol{\theta}) &= |\{(j, l) \in [n]^2 : a_j = \tilde{a}, a_l = \hat{a}, y_j < y_l, f(x_j; s, \boldsymbol{\theta}) < f(x_l; s, \boldsymbol{\theta})\}|.
\end{aligned}$$

It is

$$\begin{aligned}
\mathrm{Obj}(\boldsymbol{\theta}) &= \mathrm{Cost}(\boldsymbol{\theta}) + \lambda \cdot \text{EO-viol}(f(\cdot\,; s, \boldsymbol{\theta}); \mathcal{D}) \\
&= \frac{1}{n} \sum_{i=1}^{n} C_{y_i, f(x_i; s, \boldsymbol{\theta})} + \\
&\qquad \lambda \cdot \max_{\tilde{a} \neq \hat{a}} \left| \frac{P(\tilde{a}, \hat{a}; \boldsymbol{\theta})}{|\{(j, l) \in [n]^2 : a_j = \tilde{a}, a_l = \hat{a}, y_j > y_l\}|} - \frac{Q(\tilde{a}, \hat{a}; \boldsymbol{\theta})}{|\{(j, l) \in [n]^2 : a_j = \tilde{a}, a_l = \hat{a}, y_j < y_l\}|} \right|.
\end{aligned}$$

Given some initial thresholds $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{k-1})$, we can compute $T(i, i', \tilde{a}; \boldsymbol{\theta})$, $P(\tilde{a}, \hat{a}; \boldsymbol{\theta})$, $Q(\tilde{a}, \hat{a}; \boldsymbol{\theta})$ and $\mathrm{Obj}(\boldsymbol{\theta})$ in time $\mathcal{O}(n + k^2|\mathcal{A}|^2)$. We can then compute the best possible right-move in time $\mathcal{O}(nk|\mathcal{A}| + n|\mathcal{A}|^2)$ by going through the sorted array of scores $(s(x_j))_{j=1}^n$ once and using the following: if θ_{i_0} (for some $i_0 \in [k-1]$) is located at $s(x_j)$ and we move it to $s(x_{j+1})$, thus yielding new thresholds $\boldsymbol{\theta}'$, we have

$$\begin{aligned}
T(i_0, y_{j+1}, a_{j+1}; \boldsymbol{\theta}') &= T(i_0, y_{j+1}, a_{j+1}; \boldsymbol{\theta}) + 1, \\
T(i_0+1, y_{j+1}, a_{j+1}; \boldsymbol{\theta}') &= T(i_0+1, y_{j+1}, a_{j+1}; \boldsymbol{\theta}) - 1, \\
T(i, i', \tilde{a}; \boldsymbol{\theta}') &= T(i, i', \tilde{a}; \boldsymbol{\theta}) \quad \text{for} \quad (i, i', \tilde{a}) \notin \{(i_0, y_{j+1}, a_{j+1}), (i_0+1, y_{j+1}, a_{j+1})\}
\end{aligned}$$

and

$$P(\tilde{a},\hat{a};\boldsymbol{\theta}') = \begin{cases} P(\tilde{a},\hat{a};\boldsymbol{\theta}) - \sum_{i'=1}^{y_{j+1}-1} T(i_0,i',\hat{a};\boldsymbol{\theta}) & \text{if } a_{j+1}=\tilde{a} \\ P(\tilde{a},\hat{a};\boldsymbol{\theta}) + \sum_{i'=y_{j+1}+1}^{k} T(i_0+1,i',\tilde{a};\boldsymbol{\theta}) & \text{if } a_{j+1}=\hat{a} \\ P(\tilde{a},\hat{a};\boldsymbol{\theta}) & \text{else} \end{cases},$$

$$Q(\tilde{a},\hat{a};\boldsymbol{\theta}') = \begin{cases} Q(\tilde{a},\hat{a};\boldsymbol{\theta}) + \sum_{i'=y_{j+1}+1}^{k} T(i_0+1,i',\hat{a};\boldsymbol{\theta}) & \text{if } a_{j+1}=\tilde{a} \\ Q(\tilde{a},\hat{a};\boldsymbol{\theta}) - \sum_{i'=1}^{y_{j+1}-1} T(i_0,i',\tilde{a};\boldsymbol{\theta}) & \text{if } a_{j+1}=\hat{a} \\ Q(\tilde{a},\hat{a};\boldsymbol{\theta}) & \text{else} \end{cases},$$

$$\begin{aligned}\mathrm{Obj}(\boldsymbol{\theta}') = \mathrm{Cost}(\boldsymbol{\theta}) - \frac{1}{n}\cdot C_{y_{j+1},i+1} + \frac{1}{n}\cdot C_{y_{j+1},i} + \\ \lambda\cdot\max_{\tilde{a}\neq\hat{a}}\left|\frac{P(\tilde{a},\hat{a};\boldsymbol{\theta}')}{|\{(j,l)\in[n]^2: a_j=\tilde{a},a_l=\hat{a},y_j>y_l\}|} - \frac{Q(\tilde{a},\hat{a};\boldsymbol{\theta}')}{|\{(j,l)\in[n]^2: a_j=\tilde{a},a_l=\hat{a},y_j<y_l\}|}\right|.\end{aligned}$$

Similarly, we can compute the best possible left-move in time $\mathcal{O}(nk|\mathcal{A}|+n|\mathcal{A}|^2)$. Clearly, we have $\mathcal{O}(n+k^2|\mathcal{A}|^2), \mathcal{O}(nk|\mathcal{A}|+n|\mathcal{A}|^2) \subseteq \mathcal{O}(nk|\mathcal{A}|^2)$.

Initializing the local search:

To initialize the local search, we can choose $\theta_1,\ldots,\theta_{k-1}$ at random or, for small values of λ, we can choose them as a solution to (4) with $\lambda=0$, which we can compute in time $\mathcal{O}(n^2k^3)$ (see Appendix A.11).

A.11 Solving (4) When $\lambda = 0$

We write $s_i = s(x_i)$, $i\in[n]$, for the values of the scoring function s on the training data $\mathcal{D}=((x_i,y_i,a_i))_{i=1}^n$ and assume the values to be sorted, that is $s_i\leq s_{i+1}$, $i\in[n-1]$, and given, that is we do not take the time for evaluating s into account.

We build a table $T\in(\mathbb{R}_{\geq 0}\cup\{\infty\})^{(n+1)\times k\times k}$ with

$$T(i,f,l) = \min_{\hat{y}\in\mathcal{H}_{i,f,l}} \sum_{j=1}^{i} C_{y_j,\hat{y}_j}$$

for $i\in\{0\}\cup[n]$ and $f,l\in[k]$, where

$$\begin{aligned}\mathcal{H}_{i,f,l} = \{\hat{y}=(\hat{y}_1,\ldots,\hat{y}_i)\in[k]^i: &\ \hat{y} \text{ are predictions for } x_1,\ldots,x_i \text{ that are sorted, that is } y_r\leq y_{r+1} \text{ for } r\in[i-1],\\ &\text{with } \hat{y}_i=l, \text{ take at most } f \text{ different values and satisfy } \hat{y}_r=\hat{y}_{r'} \text{ for } s_r=s_{r'}\}.\end{aligned}$$

The optimal value of (4) with $\lambda=0$ is given by $(1/n)\cdot\min_{l\in[k]} T(n,k,l)$.

It is

$$\begin{aligned} T(0,f,l) &= 0, \quad f,l\in[k],\\ T(1,f,l) &= C_{y_1,l} \quad f,l\in[k],\\ T(i,1,l) &= \sum_{j=1}^{i} C_{y_j,l} \quad i\in[n], l\in[k],\\ T(i,f,1) &= \sum_{j=1}^{i} C_{y_j,1} \quad i\in[n], f\in[k], \end{aligned}$$

and

$$T(i,f,l) = \min_{i'<i,l'<l}\left[T(i',f-1,l') + \infty\cdot\mathbb{1}\{i'>0\wedge(s_{i'}=s_{i'+1})\} + \sum_{r=i'+1}^{i} C_{y_r,l}\right].$$

We can build T in time $\mathcal{O}(n^2k^3)$. If we store the minimizing (i', l') with $T(i, f, l)$, we can easily construct an optimal solution from T.

A.12 Simulations Illustrating the Observation Made in Section 3.5

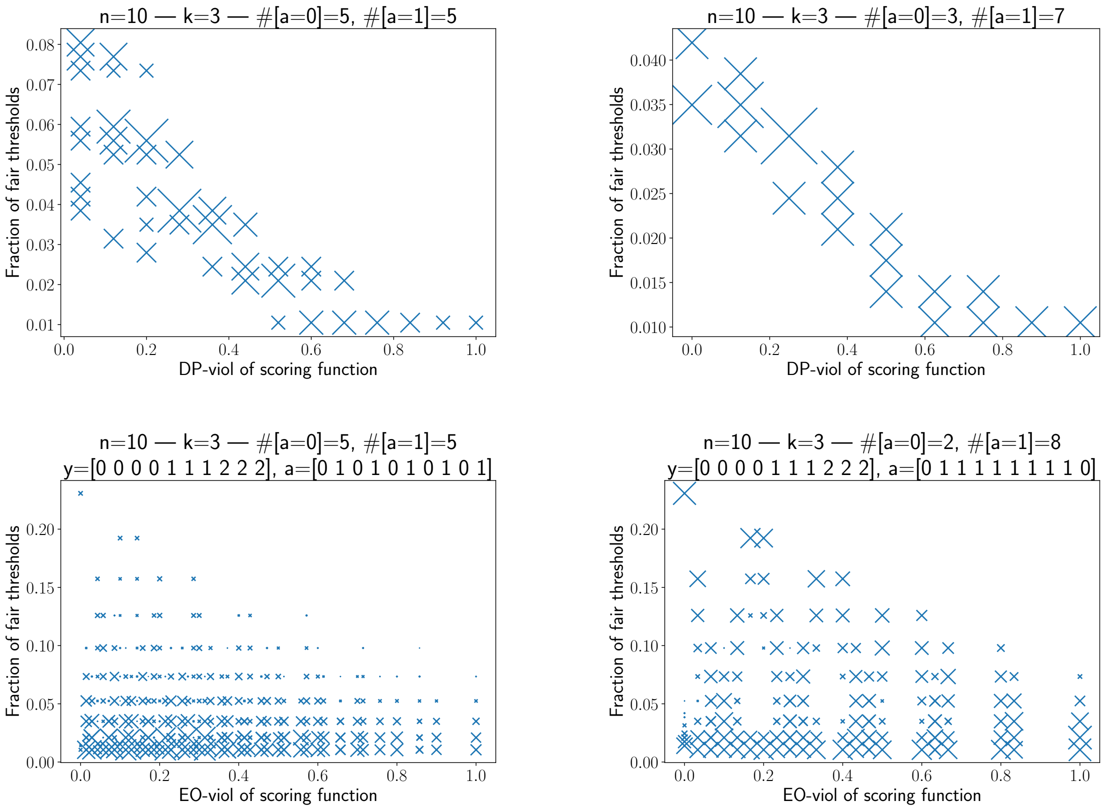

Figure 4: Simulations illustrating the observation made in Section 3.5: the more fair the scoring function, the more choices of thresholds yield a fair predictor. The size of a marker is proportional to how many times we observed a particular outcome.

Figure 4 shows some simulations illustrating the observation that we made in Section 3.5: if the scoring function is fair to some extent, then there exist more choices of thresholds that yield a fair predictor. For $n = 10$ many datapoints and $k = 3$ we considered all possible injective scoring functions by considering all permutations of $1, \ldots, 10$. We then computed the fairness violation Fair-viol of the scoring function and considering all possible choices of thresholds that yield a different predictor, we computed the fraction of choices of thresholds for which the resulting predictor is fair. The plots of Figure 4 show that fraction on the y-axis and the fairness violation Fair-viol of the scoring function on the x-axis. In the plots in the top row we studied pairwise DP and there Fair-viol = DP-viol; in the plots in the bottom row we studied pairwise EO and there Fair-viol = EO-viol. The size of a marker in these plots is proportional to how many times we observed a particular outcome (since there can be several different scoring functions with the same value of Fair-viol and the same fraction of fair thresholds). We can see that there is a strong correlation between the fairness of the scoring function and the number of choices of thresholds that yield a fair predictor. Note that when considering pairwise DP (top row), the results do not depend on which ground-truth labels we assign to the datapoints, but only on the number of datapoints with $a = 0$ and $a = 1$, respectively. When considering pairwise EO (bottom row), the results depend on the ground-truth labels and the protected attribute a of each datapoint. Each row shows two plots in different settings—the relevant information can be read from the titles of the plots.

B SOME DETAILS AND FURTHER EXPERIMENTS

B.1 Details About Implementation

We chose to build on the reduction approach of Agarwal et al. (2018) for solving the fair binary classification problem described in Section 3.1 since that approach is theoretically well-established. Furthermore, its implementation is available as part of the official Python package FAIRLEARN[4].

We used the GRIDSEARCH-method, which corresponds to Section 3.4 of Agarwal et al. and returns a deterministic rather than a randomized classifier as the EXPONENTIATEDGRADIENT-method does. As Agarwal et al. discuss, the GRIDSEARCH-method is only feasible when $|\mathcal{A}| \leq 3$. In fact, in the current version v0.6.1 of FAIRLEARN the GRIDSEARCH-method allows for $|\mathcal{A}| = 3$ only when aiming for demographic parity; it requires $|\mathcal{A}| = 2$ when aiming for equal opportunity. If we wanted to extend our implementation to arbitrary $|\mathcal{A}| \geq 2$ for both DP and EO, we would have to use the EXPONENTIATEDGRADIENT-method instead of the GRIDSEARCH-method. Since we aim to learn a deterministic classifier, we then would have to use the predictor with the highest weight in the mixture returned by the EXPONENTIATEDGRADIENT-method. Alternatively, we could try to use the other methods that are available for fair binary classification such as the various relaxation based approaches (e.g., Donini et al., 2018; Wu et al., 2019; Zafar et al., 2019). However, the relaxation based approaches have recently been criticized for several reasons (Lohaus et al., 2020).

In the GRIDSEARCH-method we set the parameters `grid_size` and `grid_limit` to 100 and 3, respectively. As base classification method we used logistic regression in the implementation available in Scikit-learn. Its main parameter is the parameter `C`, which is the inverse of a regularization parameter. We set it to $1/(2 \cdot \text{size of training data} \cdot \gamma)$ for some $\gamma \in \{10^{-i} : i \in [5]\}$ that we chose by means of 10-fold cross validation on the classification problem described in Section 3.1 without any fairness constraint, aiming for small 01-loss. We set all other parameters to their default values, except for `max_iter` and `fit_intercept`, which we set to 2500 and `False`, respectively.

Finally, in case the dataset $\mathcal{D}'$ comprised more than $6 \cdot 10^5$ datapoints, we subsampled $6 \cdot 10^5$ datapoints (cf. Section 3.4, third item).

For fitting a POM model, we used the implementation that comes with MATLAB.

B.2 Details About Datasets

In Section 5.1 we used the Drug Consumption dataset (Fehrman et al., 2015) and the Communities and Crime dataset (Redmond and Baveja, 2002; U.S. Department of Commerce, Bureau of the Census; U.S. Department Of Commerce, Bureau of the Census Producer, Washington, DC and Inter-university Consortium for Political and Social Research Ann Arbor, Michigan, 1992; U.S. Department of Justice, Bureau of Justice Statistics; U.S. Department of Justice, Federal Bureau of Investigation, 1995), which are both publicly available in the UCI repository (Dua and Graff, 2019).

The benchmark datasets used in Section 5.2 together with their splits into training and test sets are available on the website accompanying the survey paper of Gutiérrez et al. (2016).

They extracted the real ordinal regression datasets from the UCI repository (Dua and Graff, 2019) and OpenML (Vanschoren et al., 2013), and they generated one synthetic dataset as proposed by Pinto da Costa et al. (2008). We provide links to the data sources and references as requested by the data creator / donor if applicable:

- automobile: https://archive.ics.uci.edu/ml/datasets/automobile
- balance-scale: https://archive.ics.uci.edu/ml/datasets/balance+scale
- car: https://archive.ics.uci.edu/ml/datasets/car+evaluation
- ERA: https://www.openml.org/d/1030
- ESL: https://www.openml.org/d/1027
- eucalyptus: https://www.openml.org/d/188

[4] https://fairlearn.github.io/ (MIT License)

- LEV: https://www.openml.org/d/1029
- newthyroid: https://archive.ics.uci.edu/ml/datasets/thyroid+disease
- SWD: https://www.openml.org/d/1028
- toy: synthetic dataset generated as proposed by Pinto da Costa et al. (2008)
- winequality-red (Cortez et al., 2009): https://archive.ics.uci.edu/ml/datasets/wine+quality

Gutiérrez et al. (2016) downloaded the discretized regression datasets from the website accompanying the paper of Chu and Ghahramani (2005). Chu and Ghahramani, in turn, obtained them from the website of Luis Torgo[5], who obtained them from the UCI repository (Dua and Graff, 2019), the Delve project[6], the StatLib datasets archive[7], and the Bilkent University Function Approximation Repository (Altay Guvenir and Uysal, 2000). We provide links to the data sources and references as requested by the data creator / donor if applicable:

- abalone: http://archive.ics.uci.edu/ml/datasets/Abalone
- bank: http://www.cs.toronto.edu/~delve/data/bank/desc.html
- calhousing: http://lib.stat.cmu.edu/datasets/ as `houses.zip`; submitted by Kelley Pace
- census: http://www.cs.toronto.edu/~delve/data/census-house/desc.html
- computer: http://www.cs.toronto.edu/~delve/data/comp-activ/desc.html
- housing: http://lib.stat.cmu.edu/datasets/boston
- machine: http://archive.ics.uci.edu/ml/datasets/Computer+Hardware
- stock: http://pcaltay.cs.bilkent.edu.tr/DataSets/

We normalized every dataset such that each feature has zero mean and unit variance on the training set. Table 1 and Table 2 provide some additional statistics of the Drug Consumption dataset and the Communities and Crime dataset, respectively, which we used in Section 5.1. Table 3 and Table 4 provide some statistics of the real ordinal datasets and discretized regression datasets used in Section 5.2. Further statistics, such as the distribution per class, that is $\mathbb{P}[y = i]$, $i \in [k]$, can be found in Section 4 of Gutiérrez et al. (2016).

Table 1: Statistics of the Drug Consumption dataset used in Section 5.1.

$\mathbb{P}[a = \text{f}]$	$\mathbb{P}[y = 1]$	$\mathbb{P}[y = 2]$	$\mathbb{P}[y = 3]$	$\mathbb{P}[y = 4]$	$\mathbb{P}[y = 5]$
0.50	0.33	0.14	0.11	0.17	0.24

Table 2: Statistics of the Communities and Crime dataset used in Section 5.1.

$\mathbb{P}[a = \text{white}]$	$\mathbb{P}[y = 1]$	$\mathbb{P}[y = 2]$	$\mathbb{P}[y = 3]$	$\mathbb{P}[y = 4]$	$\mathbb{P}[y = 5]$	$\mathbb{P}[y = 6]$	$\mathbb{P}[y = 7]$	$\mathbb{P}[y = 8]$
0.57	0.20	0.18	0.22	0.13	0.08	0.05	0.05	0.09

[5] https://www.dcc.fc.up.pt/~ltorgo/Regression/DataSets.html
[6] http://www.cs.toronto.edu/~delve/
[7] http://lib.stat.cmu.edu/datasets/

Table 3: Statistics of the real ordinal regression datasets used in Section 5.2.

	# train	# test	# features	# classes	$\mathbb{P}[a=0]$	$\mathbb{P}[y_1 > y_2 \| a_1 = 0, a_2 = 1]$	$\mathbb{P}[y_1 < y_2 \| a_1 = 0, a_2 = 1]$	$\|\mathbb{P}[y_1 > y_2 \| a_1 = 0, a_2 = 1] -$ $\mathbb{P}[y_1 < y_2 \| a_1 = 0, a_2 = 1]\|$
automobile	153	52	70	6	0.53	0.62	0.20	0.42
balance-scale	468	157	3	3	0.41	0.51	0.13	0.38
car	1296	432	20	4	0.67	0.45	0.0	0.45
ERA	750	250	3	9	0.46	0.25	0.63	0.38
ESL	366	122	3	9	0.40	0.06	0.81	0.75
eucalyptus	552	184	90	5	0.59	0.38	0.41	0.06
LEV	750	250	3	5	0.44	0.17	0.57	0.41
newthyroid	161	54	4	3	0.49	0.42	0.07	0.35
SWD	750	250	9	4	0.55	0.23	0.46	0.23
toy	225	75	1	5	0.49	0.29	0.49	0.20
winequality-red	1199	400	10	6	0.48	0.27	0.38	0.11

Table 4: Statistics of the discretized regression datasets used in Section 5.2.

	# train	# test	# features	# classes	$\mathbb{P}[a=0]$	$\mathbb{P}[y_1 > y_2 \mid a_1=0, a_2=1]$	$\mathbb{P}[y_1 < y_2 \mid a_1=0, a_2=1]$	$\lvert\mathbb{P}[y_1 > y_2 \mid a_1=0, a_2=1] - \mathbb{P}[y_1 < y_2 \mid a_1=0, a_2=1]\rvert$
abalone-5	1000	3177	9	5	0.63	0.28	0.53	0.25
bank1-5	50	8142	7	5	0.50	0.38	0.42	0.05
bank2-5	75	8117	31	5	0.50	0.41	0.39	0.02
calhousing-5	150	20490	7	5	0.50	0.35	0.46	0.11
census1-5	175	22609	7	5	0.50	0.14	0.72	0.58
census2-5	200	22584	15	5	0.50	0.14	0.72	0.58
computer1-5	100	8092	11	5	0.50	0.70	0.15	0.55
computer2-5	125	8067	20	5	0.50	0.70	0.15	0.55
housing-5	300	206	12	5	0.50	0.65	0.19	0.46
machine-5	150	59	5	5	0.49	0.75	0.12	0.63
stock-5	600	350	8	5	0.50	0.18	0.69	0.51
abalone-10	1000	3177	9	10	0.63	0.32	0.58	0.26
bank1-10	50	8142	7	10	0.50	0.43	0.47	0.05
bank2-10	75	8117	31	10	0.50	0.46	0.44	0.01
calhousing-10	150	20490	7	10	0.50	0.40	0.51	0.11
census1-10	175	22609	7	10	0.50	0.17	0.76	0.59
census2-10	200	22584	15	10	0.50	0.16	0.76	0.60
computer1-10	100	8092	11	10	0.49	0.75	0.18	0.57
computer2-10	125	8067	20	10	0.50	0.75	0.18	0.57
housing-10	300	206	12	10	0.50	0.71	0.22	0.48
machine-10	150	59	5	10	0.48	0.80	0.14	0.67
stock-10	600	350	8	10	0.50	0.20	0.75	0.54

B.3 Addendum to the Experiment of Section 5.1 on the Communities and Crime Dataset

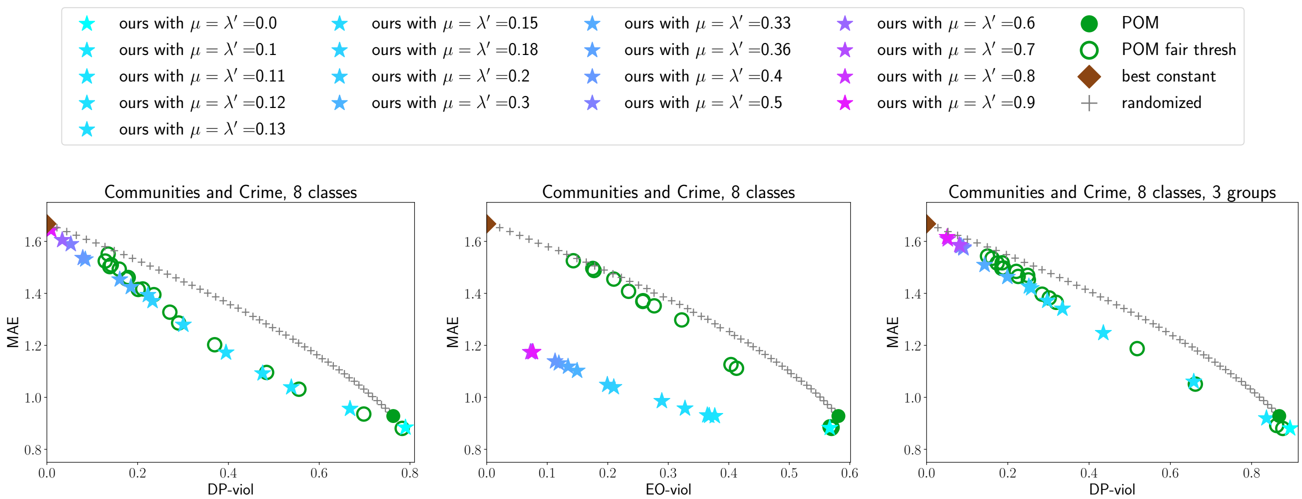

Figure 5: **Left and center:** Results of the experiment of Section 5.1 on the Communities and Crime dataset when we run our strategy for additional values of $\mu = \lambda'$. **Right:** MAE vs DP-viol for the various predictors on a version of the dataset with three protected groups.

Figure 5 (left and center) show the results of the experiment on the Communities and Crime dataset presented in Section 5.1, when we additionally run our strategy with $\mu = \lambda' \in \{0.11, 0.12, 0.13, 0.15, 0.18\}$. This results in nicely exploring the MAE-vs-DP-viol trade-off over the whole range of DP-viol $\in [0, 0.8]$.

The right plot shows the results (when aiming for pairwise DP) for a version of the dataset with three protected groups: rather than $a \in \{\text{white}, \text{diverse}\}$, we consider $a \in \{\text{white}, \text{African-American}, \text{Hispanic or Asian}\}$. We set $a = $ African-American if at least 25% of a community's population are African-Americans, $a = $ Hispanic or Asian if at least 25% are Hispanics or Asians and less than 25% are African-Americans, and $a = $ white otherwise. It is $\mathbb{P}[a = \text{white}] = 0.46$, $\mathbb{P}[a = \text{African-American}] = 0.24$ and $\mathbb{P}[a = \text{Hispanic or Asian}] = 0.3$. We can see that also in this case the predictors produced by our approach nicely explore the accuracy-vs-fairness trade-off. Note that our implementation does not allow us to aim for pairwise EO in the case of three groups due to the limitations of Fairlearn's GridSearch-method (cf. Appendix B.1).

B.4 Experiment with Asymmetric Cost Matrix

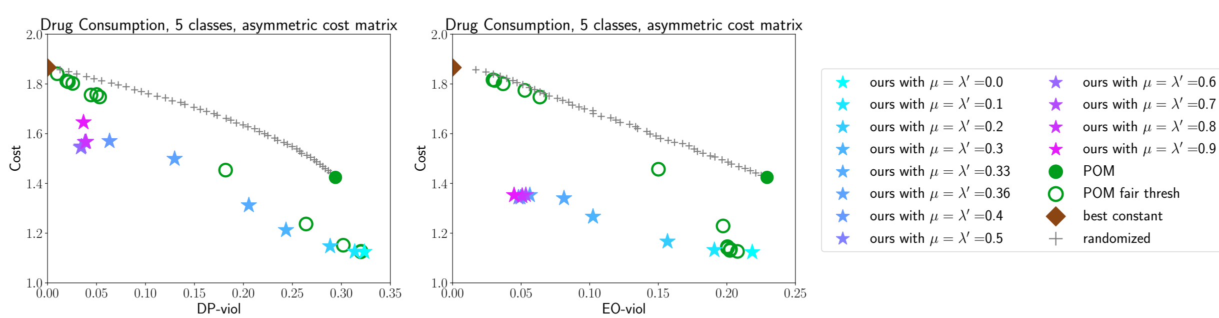

Figure 6: Same experiment on the DC dataset as shown in Figure 1, but with an asymmetric cost matrix.

We performed the same experiment on the Drug Consumption dataset as described in Section 5.1, but with the asymmetric cost matrix $C_{i,j} = |i - j| + |i - j| \cdot \mathbb{1}\{j > i\}$ instead of the absolute cost matrix $C_{i,j} = |i - j|$. All observations made in Section 5.1 for the symmetric case hold similarly in the asymmetric case.

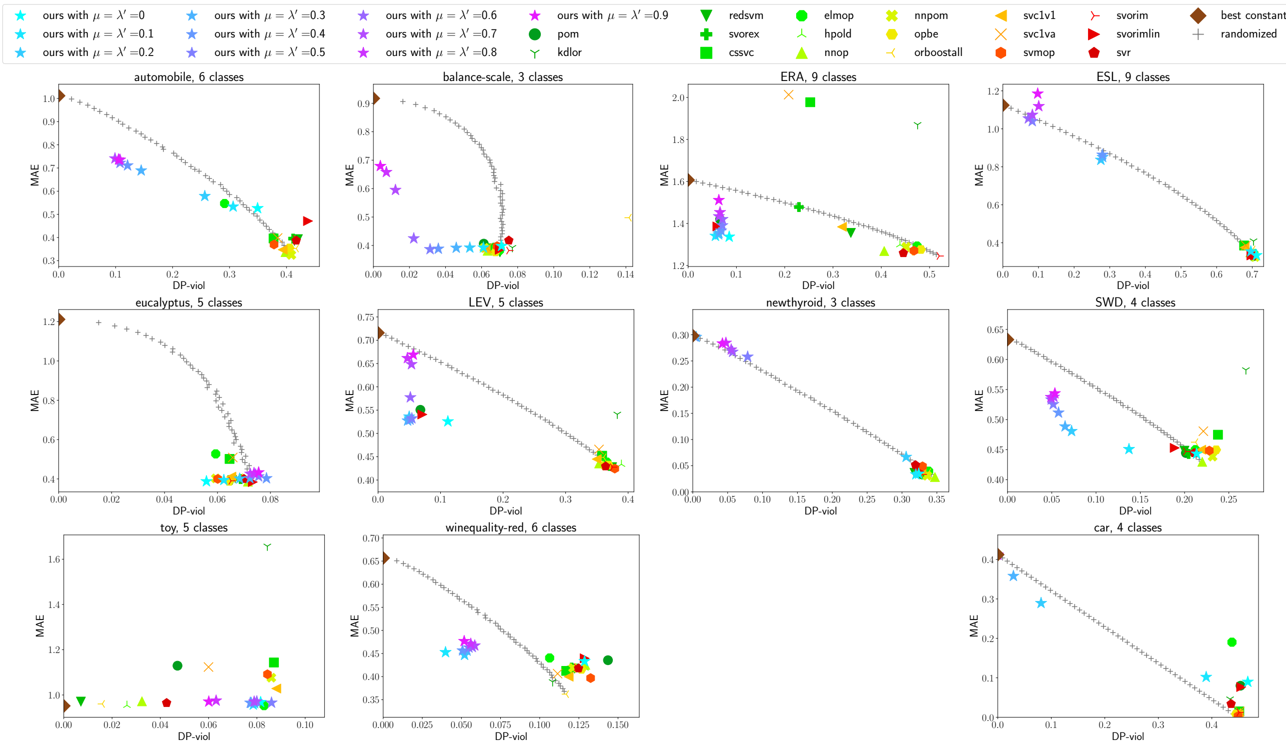

Figure 7: Experiments of Section 5.2 on the **real ordinal regression datasets** when aiming for **pairwise DP**. Note that the toy dataset has only a single feature that is provided as input to a predictor and that the best method on the toy dataset (svorex) coincides with the best constant predictor; we do not see any grey crosses corresponding to randomly mixing the best predictor with the best constant one.

B.5 Experiments of Section 5.2

Figures 7 to 12 show the results of the experiments described in Section 5.2. Our observations discussed in Section 5 generally hold across all datasets, with a few exceptions: on the eucalyptus dataset, all methods are quite fair with DP-viol ≤ 0.08 and EO-viol ≤ 0.07, and our method does not provide any significant improvement; note that our predictors are competitive with all state-of-the-art predictors in terms of the MAE. On the newthyroid dataset, we observe two interesting phenomena when aiming for pairwise EO. First, it is striking that the randomized predictors are significantly less fair than the predictor with the smallest MAE for values of p around 0.5 ("the grey curve strings a big bow"). Second, the EO-viol of our predictors increases as $\mu = \lambda'$ increases, which is in stark contrast to what we would expect. We do not have an explanation for these two phenomena, but suspect that they have a common cause. On the toy dataset, the best constant predictor outperforms all other predictors. Since the toy dataset has only a single feature that is provided as input to a predictor, these results are not too meaningful. On the bank1-5 and bank1-10 datasets (when aiming for pairwise DP) and on the bank2-5 and bank2-10 datasets (when aiming for pairwise EO) we observe the same two phenomena as on the newthyroid dataset, which we do not have an explanation for. However, note that in these cases *all* predictors satisfy Fair-viol ≤ 0.012 (!) and are almost perfectly fair.

Figures 13 to 18 show the same results, but additionally show the standard deviation of the MAE and Fair-viol over the various splits. We can see that the standard deviations for our predictors are mostly comparable with the standard deviations for the predictors produced by the ORCA algorithms, both in case of the MAE and Fair-viol.

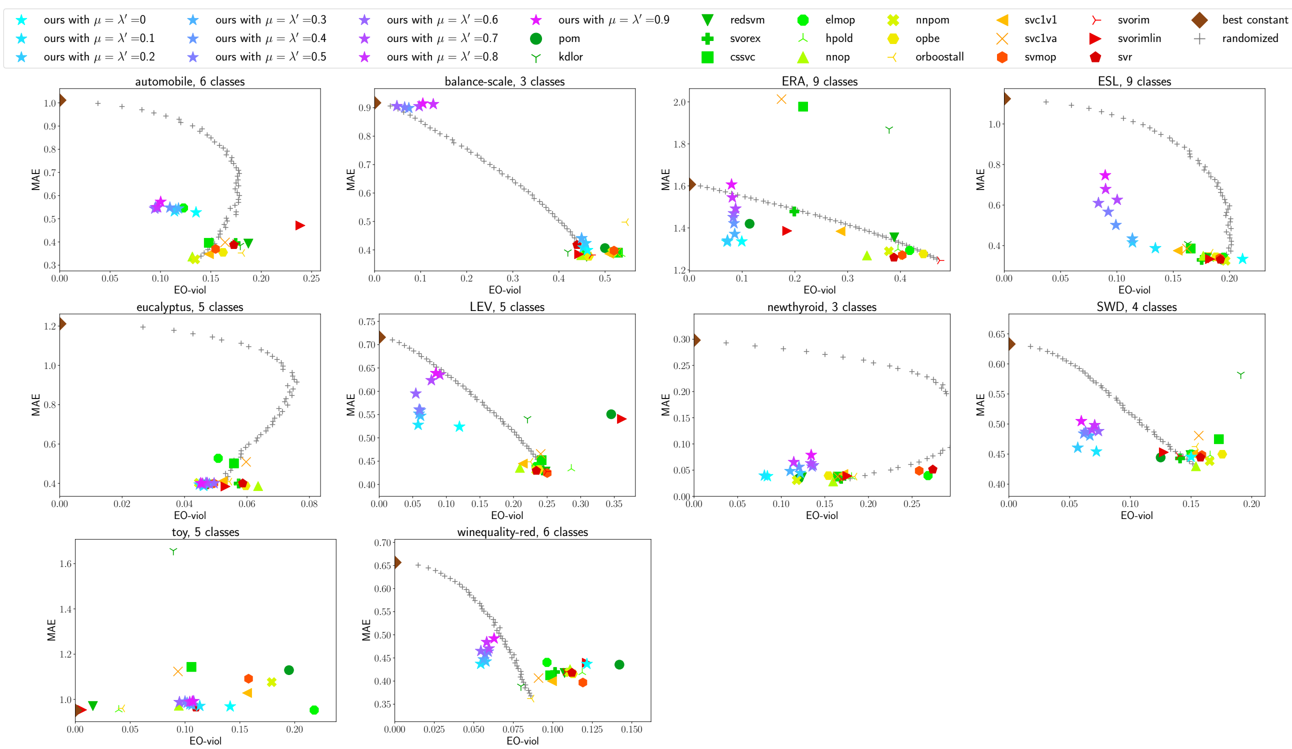

Figure 8: Experiments of Section 5.2 on the **real ordinal regression datasets** when aiming for **pairwise EO**. Note that the toy dataset has only a single feature that is provided as input to a predictor and that the best method on the toy dataset (svorex) coincides with the best constant predictor; we do not see any grey crosses corresponding to randomly mixing the best predictor with the best constant one. Also note that we do not provide a plot for the car dataset; since $\mathbb{P}[y_1 < y_2, a_1 = 0, a_2 = 1] = 0$ for the car dataset (cf. Table 3), the notion of pairwise EO is not well-defined for this dataset.

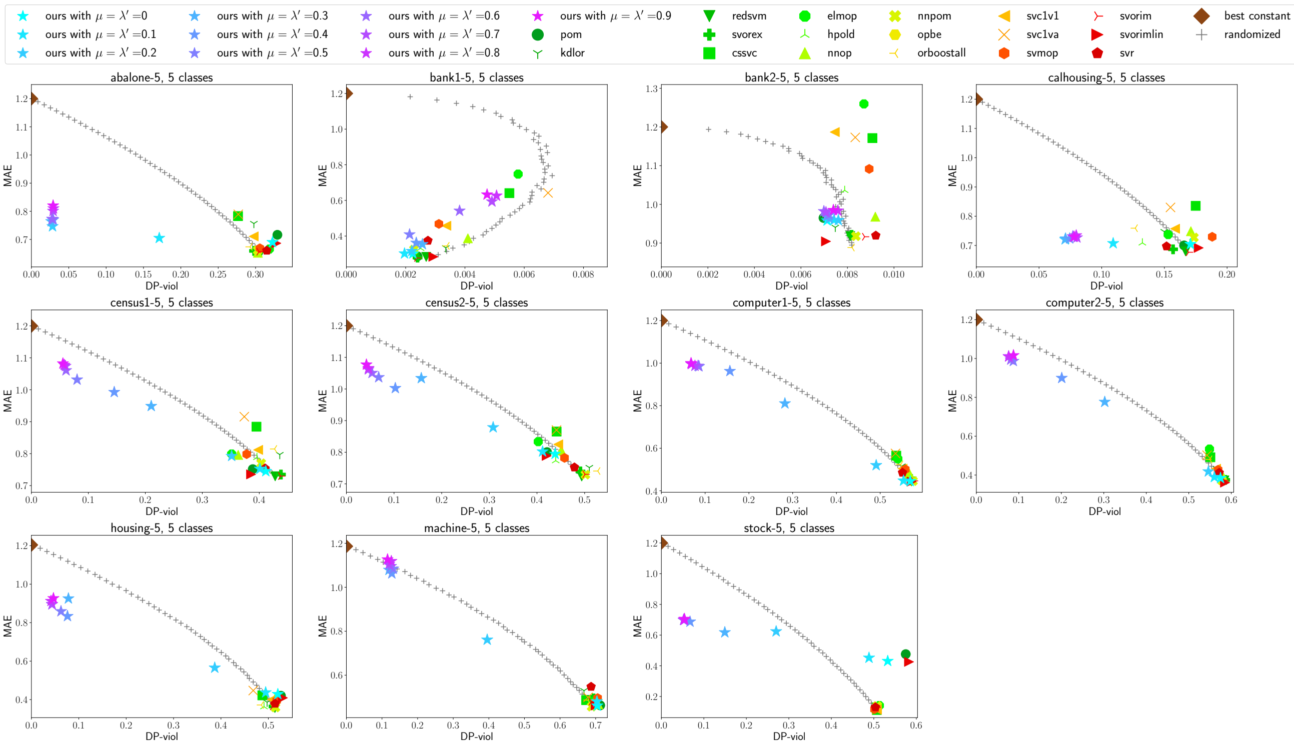

Figure 9: Experiments of Section 5.2 on the **discretized regression datasets with 5 classes** when aiming for **pairwise DP**.

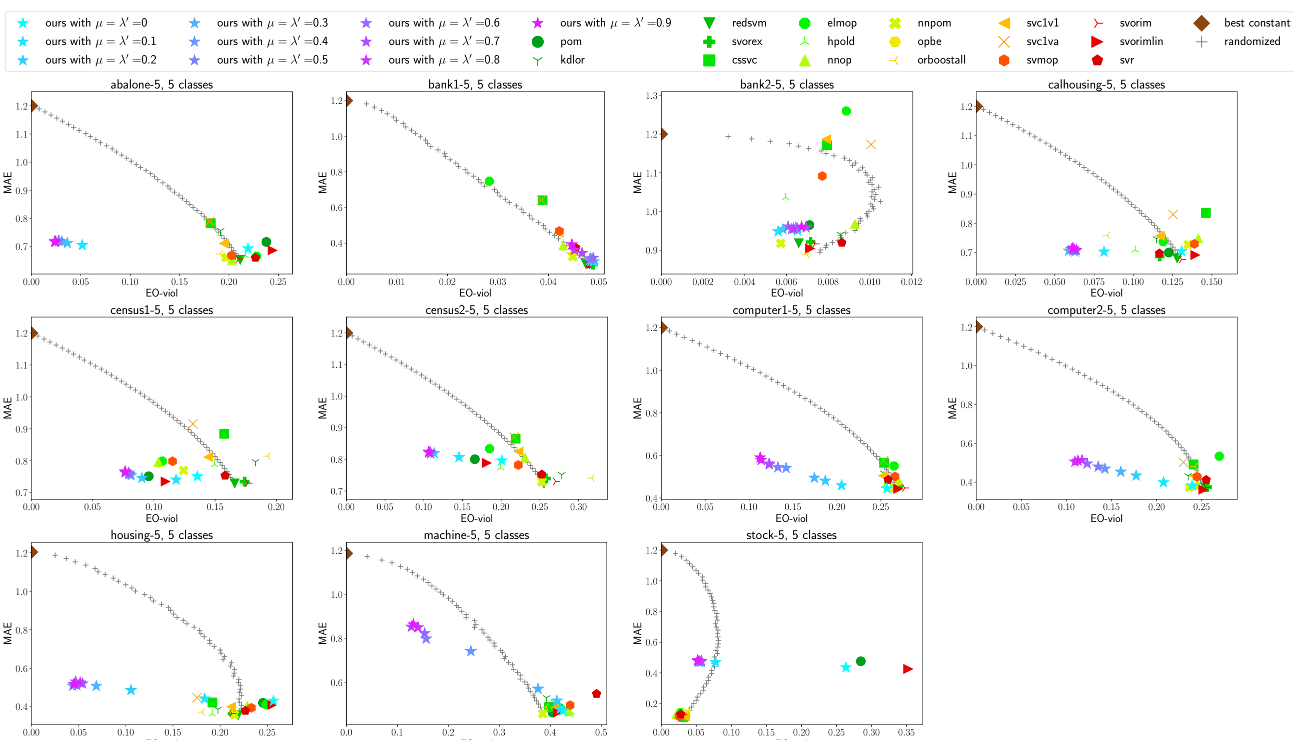

Figure 10: Experiments of Section 5.2 on the **discretized regression datasets with 5 classes** when aiming for **pairwise EO**.

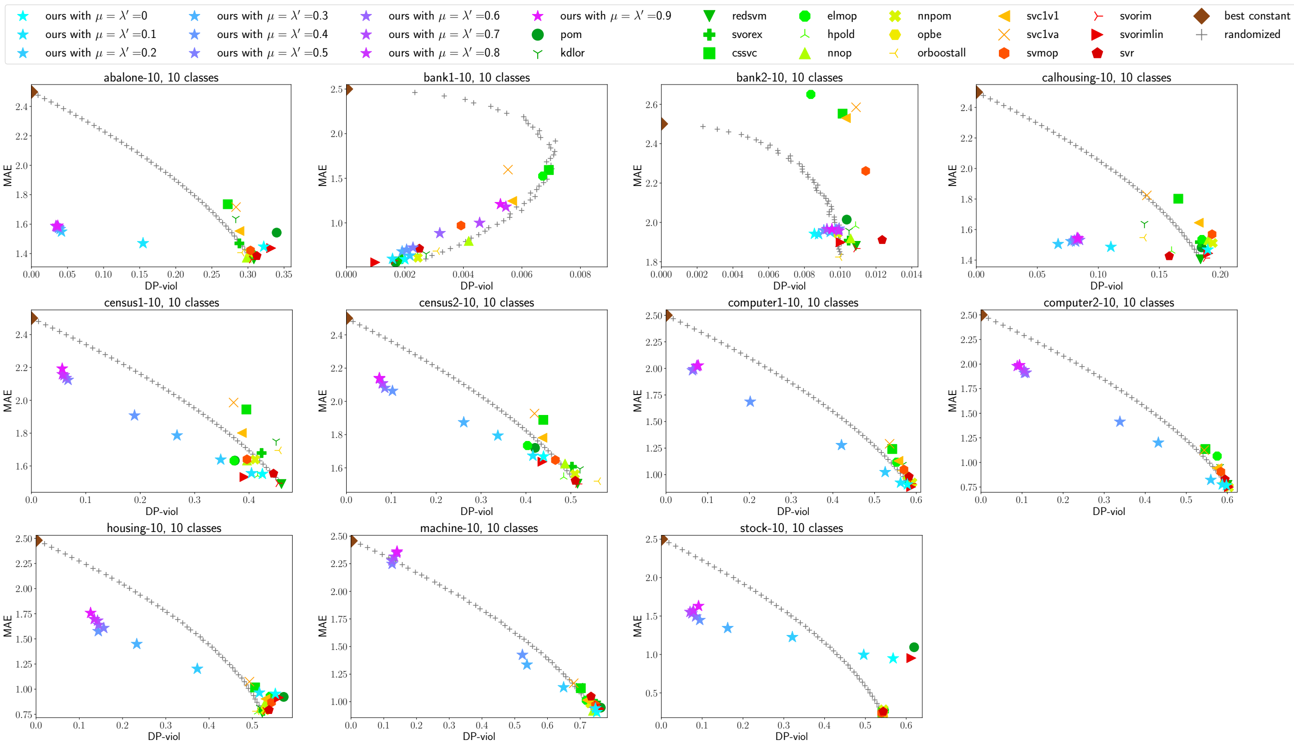

Figure 11: Experiments of Section 5.2 on the **discretized regression datasets with 10 classes** when aiming for **pairwise DP**.

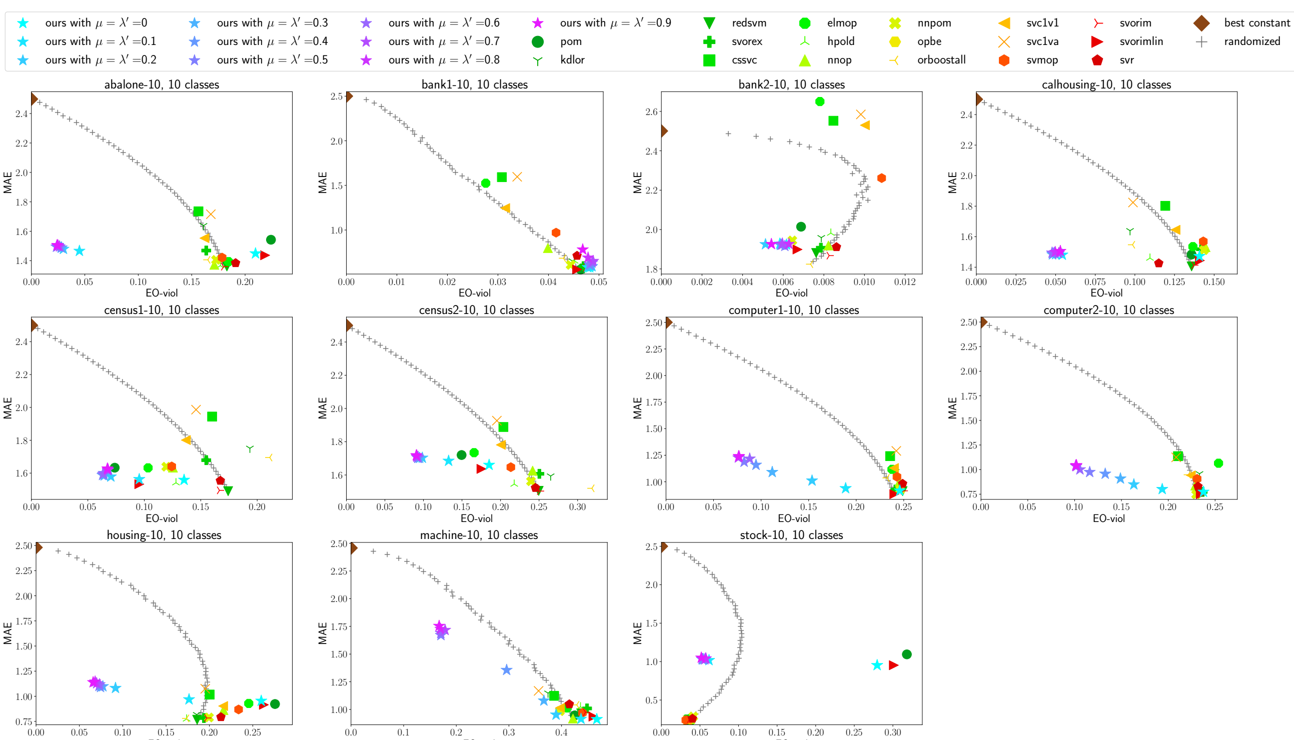

Figure 12: Experiments of Sec. 5.2 on the **discretized regression datasets with 10 classes** when aiming for **pairwise EO**.

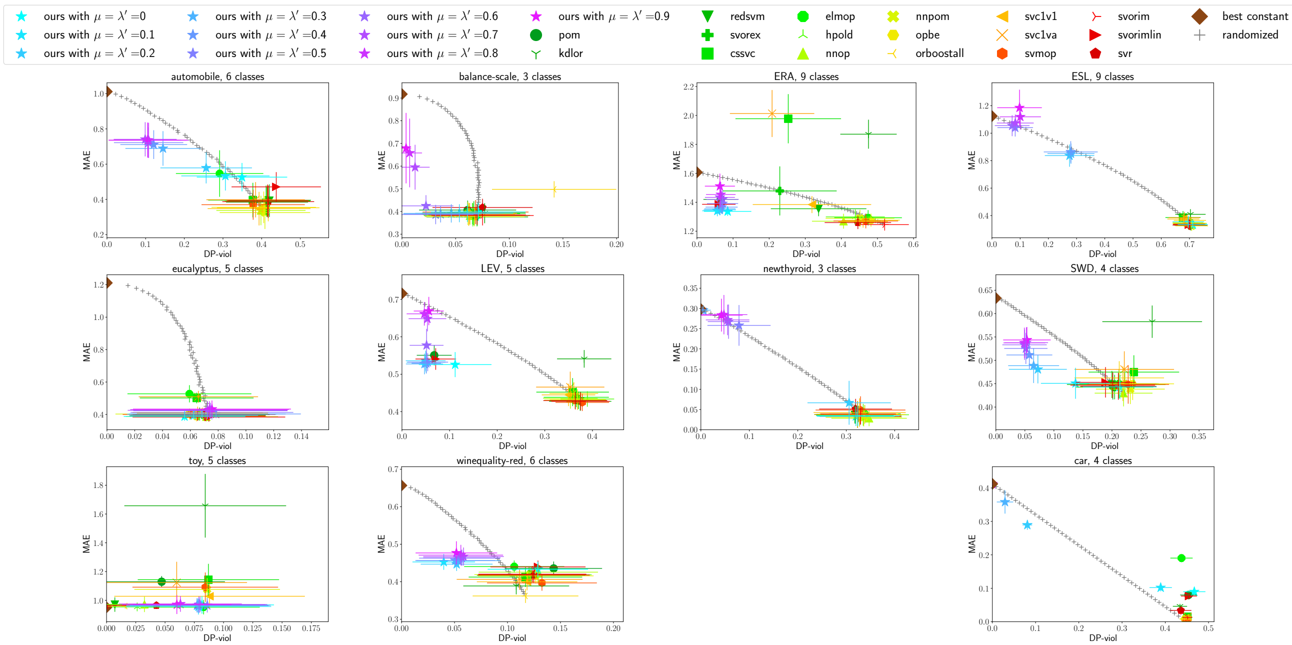

Figure 13: Experiments of Sec. 5.2 on the **real ordinal regression datasets** when aiming for **pairwise DP**. Note that the toy dataset has only a single feature that is provided as input to a predictor and that the best method on the toy dataset (svorex) coincides with the best constant predictor; we do not see any grey crosses corresponding to randomly mixing the best predictor with the best constant one. The errorbars show the standard deviation over the 30 splits into training and test sets.

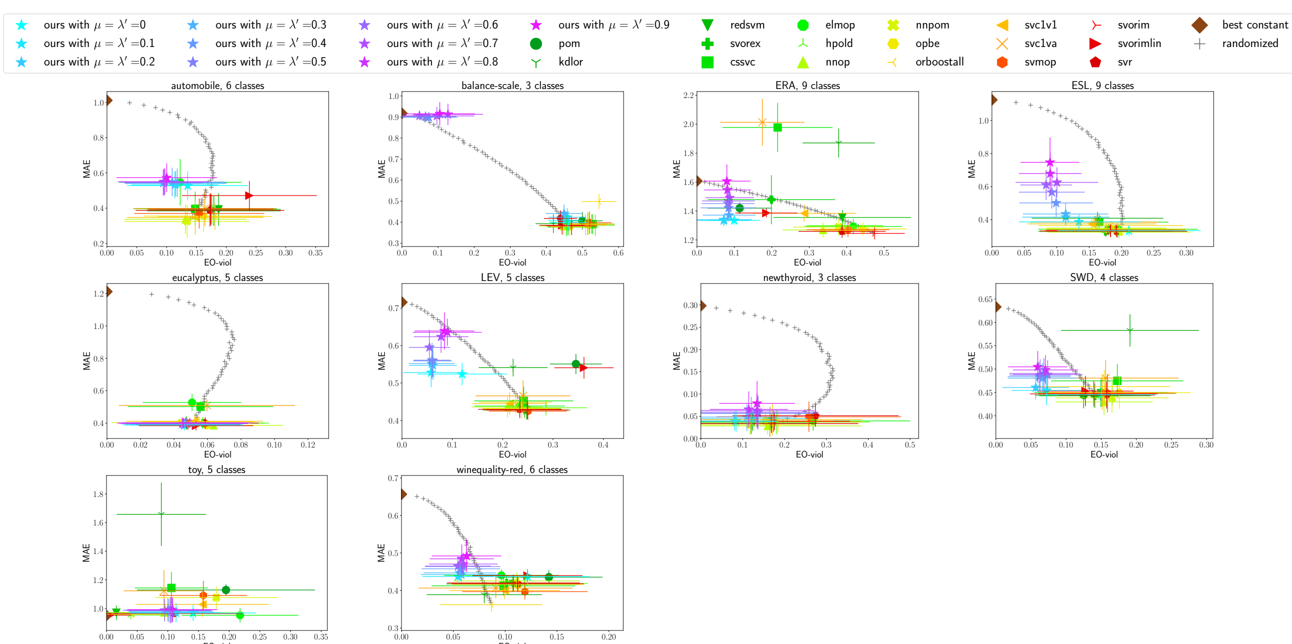

Figure 14: Experiments of Section 5.2 on the **real ordinal regression datasets** when aiming for **pairwise EO**. Note that the toy dataset has only a single feature that is provided as input to a predictor and that the best method on the toy dataset (svorex) coincides with the best constant predictor; we do not see any grey crosses corresponding to randomly mixing the best predictor with the best constant one. Also note that we do not provide a plot for the car dataset; since $\mathbb{P}[y_1 < y_2, a_1 = 0, a_2 = 1] = 0$ for the car dataset (cf. Table 3), the notion of pairwise EO is not well-defined for this dataset. The errorbars show the standard deviation over the 30 splits into training and test sets.

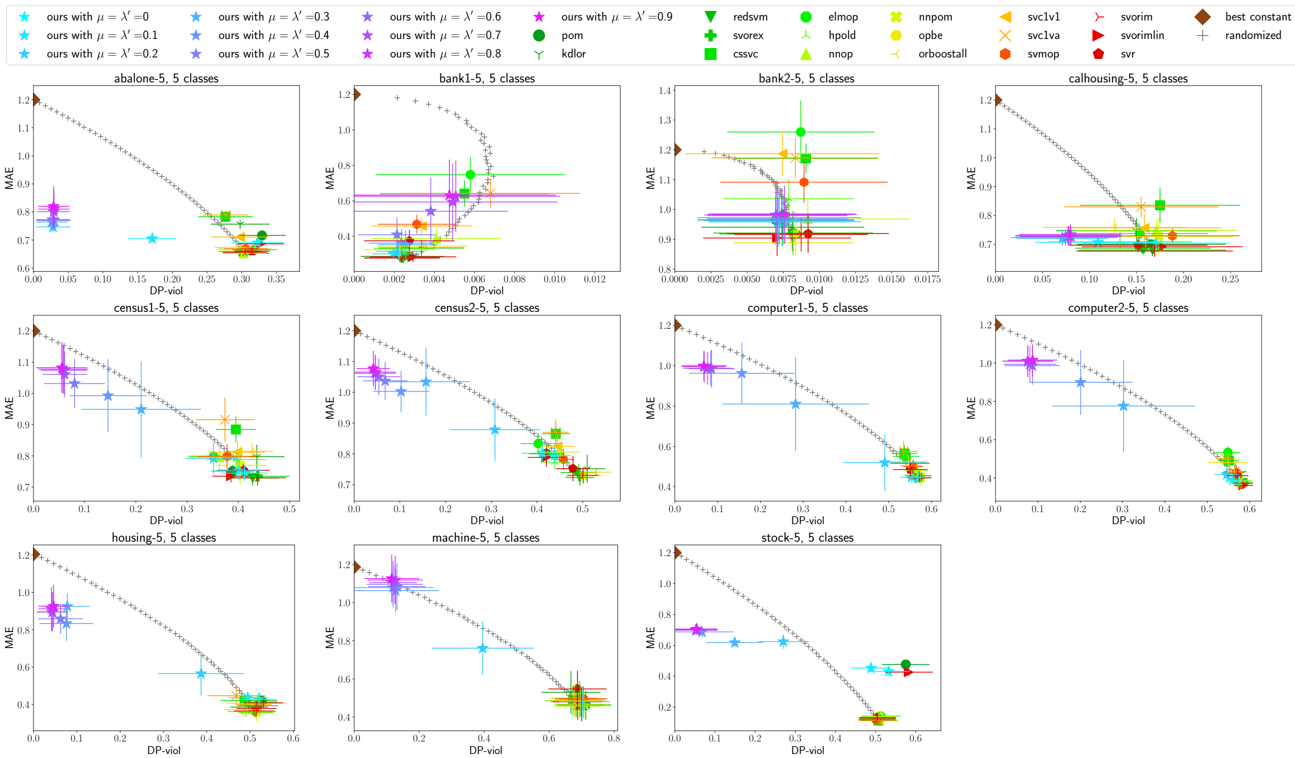

Figure 15: Experiments of Section 5.2 on the **discretized regression datasets with 5 classes** when aiming for **pairwise DP**. The errorbars show the standard deviation over the 20 splits into training and test sets.

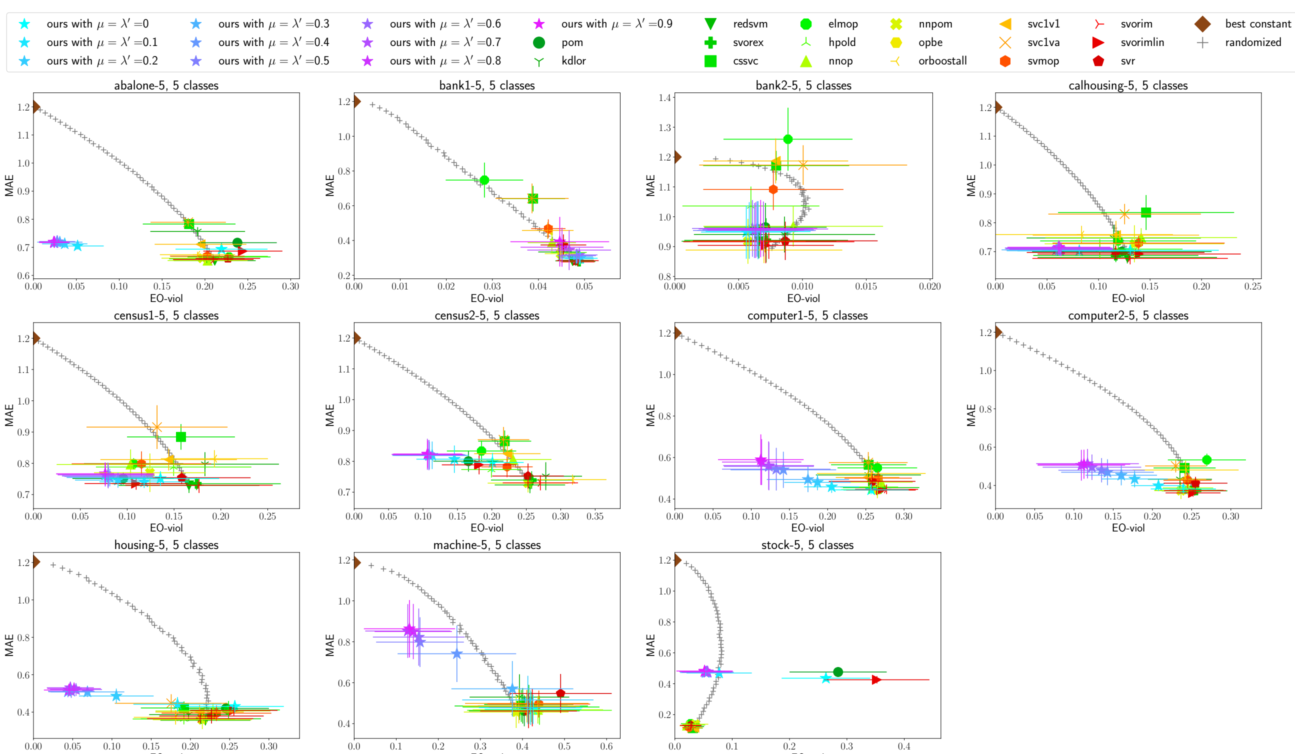

Figure 16: Experiments of Section 5.2 on the **discretized regression datasets with 5 classes** when aiming for **pairwise EO**. The errorbars show the standard deviation over the 20 splits into training and test sets.

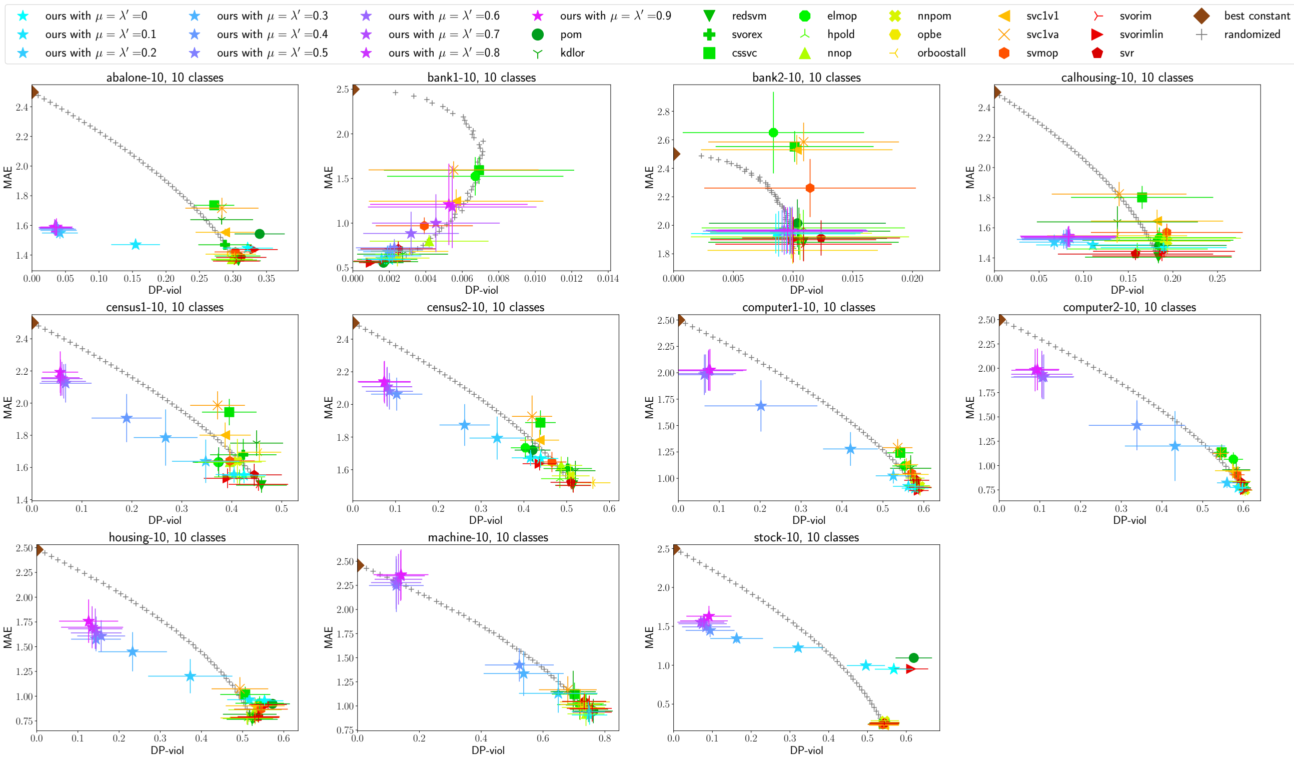

Figure 17: Experiments of Section 5.2 on the **discretized regression datasets with 10 classes** when aiming for **pairwise DP**. The errorbars show the standard deviation over the 20 splits into training and test sets.

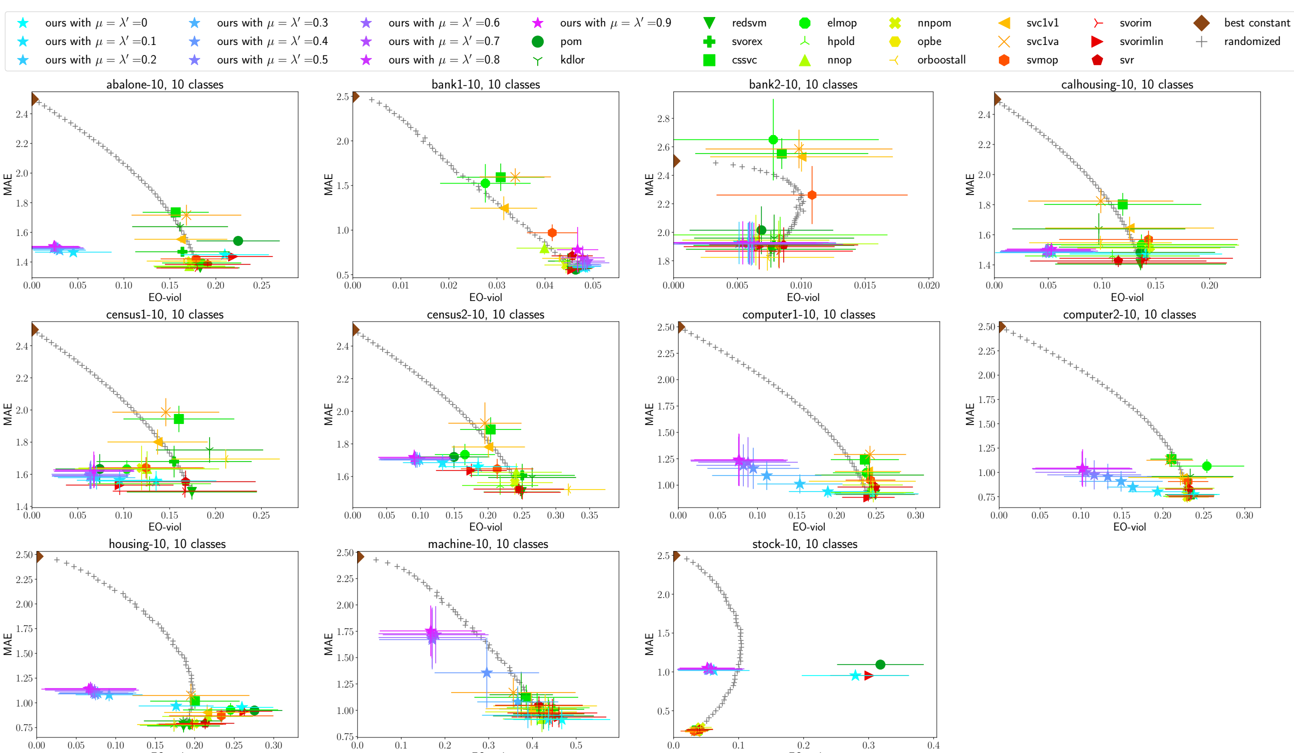

Figure 18: Experiments of Section 5.2 on the **discretized regression datasets with 10 classes** when aiming for **pairwise EO**. The errorbars show the standard deviation over the 20 splits into training and test sets.