Fair Regression via Plug-In Estimator and Recalibration

Evgenii Chzhen*
LMO, Université Paris-Saclay
CNRS, Inria

Christophe Denis
LAMA, Université Gustave Eiffel
MIA-Paris, AgroParisTech
INRAE, Université Paris-Saclay

Mohamed Hebiri
LAMA, Université Gustave Eiffel
CREST, ENSAE, IP Paris

Luca Oneto
DIBRIS, University of Genoa

Massimiliano Pontil
Istituto Italiano di Tecnologia
University College London

Abstract

We study the problem of learning an optimal regression function subject to a fairness constraint. It requires that, conditionally on the sensitive feature, the distribution of the function output remains the same. This constraint naturally extends the notion of demographic parity, often used in classification, to the regression setting. We tackle this problem by leveraging on a proxy-discretized version, for which we derive an explicit expression of the optimal fair predictor. This result naturally suggests a two stage approach, in which we first estimate the (unconstrained) regression function from a set of labeled data and then we recalibrate it with another set of unlabeled data. The recalibration step can be efficiently performed via a smooth optimization. We derive rates of convergence of the proposed estimator to the optimal fair predictor both in terms of the risk and fairness constraint. Finally, we present numerical experiments illustrating that the proposed method is often superior or competitive with state-of-the-art methods.

1 Introduction

During the recent years algorithmic fairness has emerged as a fundamental area of machine learning, due to the potential risk that standard learning algorithms, when trained on sensitive datasets, may inherit or amplify bias present in the data. This has raised the challenge to design novel algorithms that, while still optimizing prediction performance, mitigate or remove unfairness of the learned predictor, see the papers and books [4, 13, 15, 21, 24, 29, 31, 33, 34, 40, 41, 57–60] and references therein. Until very recently, most work has focused on classification problems, with regression receiving far less attention. However regression problems are equally important for algorithmic fairness. For example, both the problems of predicting students' performance without discriminating based on the gender, or predicting the crime risk of a community without discriminating based on the race, can be cast as regression.

In this paper we study the problem of designing computationally efficient and statistically principled learning methods for fair regression. We define the optimal fair regression function as the one that

*evgenii.chzhen@universite-paris-saclay.fr

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

minimizes the population squared error subject to a fairness constraint that asks that the function output is independent from the sensitive feature. This notion of fairness is referred to as demographic parity and is widely used in the literature, both in the context of classification and regression [2, 14, 28, 32, 47].

The above definition of optimal fair regression function is not well suited to design an efficient algorithm. Therefore, we first consider a proxy-discretized version of the fair regression problem, for which we derive an explicit expression of the optimal fair predictor. Importantly, we show that this discretization scheme does not alter the quality of the optimal rule: the optimal fair predictors for both problems (the discretized and the original one) have close risks, controlled by the discretization parameter. Our expression for the discretized optimal predictor naturally suggests a plug-in two stage approach, in which we first estimate the (unconstrained) regression function from a set of labeled data and then we recalibrate it with another set of unlabeled data. The latter step can be efficiently performed via a smooth optimization.

A key feature of our approach is that it can be employed alongside any off-the-shelf regression learning method and, provided this one is consistent, our recalibration step transforms in a simple way the original (unconstrained) regression estimator into a prediction function which consistently estimates the optimal fair regression function. This strategy is particularly appealing in those applications where the cost of re-training an existing learning algorithm is high. Furthermore, we derive rates of convergence of the proposed estimator to the optimal fair predictor both in terms of the risk and the fairness constraint violation.

Finally, we present numerical experiments with the proposed method on five real datasets, indicating that our method is often superior or competitive with state-of-the-art methods. In particular, when using random forest as the base regression estimator, our approach results in substantial decrease in fairness violation, at the costs of only a moderate increase in the prediction error rate.

Previous work. One of the first work on fair regression is [12], where the authors study the problem of linear regression imposing constraints on the mean outcome or residuals of the models (fairness in expectation). More recently, several authors [2, 6, 26, 36, 38, 43, 47–49] focus on the fair regression problem all employing various fairness definitions. Similarly to [12], the works [6, 36, 48] deal with the linear regression setup by refining the definition of fairness. Raff et al. [49] and Fitzsimons et al. [26] examine the incorporation of fairness in expectation constraints in tree based regression methods. Pérez-Suay et al. [48] incorporate a penalty on the dependence between the predictor and the sensitive attribute into the kernel ridge regression formulation. Unlike these contributions, we do not assume neither linear nor linear in a kernel space relationships between the input and the output.
More related to our work are the papers by Oneto et al. [47] and Agarwal et al. [2]. The former introduces a framework for fair Empirical Risk Minimization (ERM) in the context of regression, providing general bounds in the case of fair regression in RKHS, using a relaxed notion of linearized fairness. The latter paper elegantly transforms the problem of bounded fair regression to a classification problem and then employs the reduction approach of [1]. They derive ERM-type generalization guarantees which are applicable to any class of predictors with bounded pseudo-dimension. Two notions of fairness are used, closest to ours being the Kolmogorov-Smirnov (KS) distance. In contrast to the above papers, we measure unfairness by the Total Variation (TV) distance, which is a stronger notion than the KS distance. Furthermore, our guarantees do not require the optimal predictor to be in a Glivenko–Cantelli or a bounded pseudo-dimension class. Yet, the price for such a guarantee is an extra mild assumption on the distribution of the observations.
Our theoretical contribution is partly inspired by recent work of Chzhen et al. [18], where the authors study binary classification using the Equal Opportunity constraint [see 29]. While they also provide a two stage plug-in approach, the setting considered here induces a non-trivial adaptation of their method of proof, involving a discretization step to deal with the uncountable nature of the constraint. Moreover, contrary to them, we derive finite sample bounds.

2 Fair regression

In this section, we introduce the fair regression problem and describe a discretized version of it, for which we derive an explicit form of the optimal regression function.

2.1 Learning setting

We let $(X, S, Y) \in \mathbb{R}^d \times \mathcal{S} \times \mathbb{R}$ be random tuple distributed according to a Borel probability measure $\mathbb{P}$ on $\mathbb{R}^d \times \mathcal{S} \times \mathbb{R}$. Here $X \in \mathbb{R}^d$ is a feature vector, $S \in \mathcal{S} := \{-1, 1\}$ is a binary sensitive feature (*i.e.,* protected attribute), and $Y \in \mathbb{R}$ is a real valued signal to be predicted. For all $s \in \mathcal{S}$ we denote by $\mathbb{P}_{X|S=s}$ the conditional distribution of $X|S = s$, by $p_s = \mathbb{P}(S = s)$ the marginal distribution of S, and by $\eta(X, s) = \mathbb{E}[Y|X, S = s]$ the conditional expectation of Y. Throughout the paper, we denote by $\mathcal{F}$ the set of *all* Borel measurable functions $f : \mathbb{R}^d \times \mathcal{S} \to \mathbb{R}$. In this work we study predictors which include $s \in \mathcal{S}$ in their functional form.

We consider the standard mean squared risk of a predictor f defined as $\mathcal{R}(f) := \mathbb{E}(Y - f(X, S))^2$. We consider a natural extension of the Demographic Parity [11] as the notion of fairness[2], which was previously used in the context of regression by [2, 14, 32].

Definition 2.1 (Fair predictor)**.** *We say that a predictor $f \in \mathcal{F}$ is* fair *with respect to the distribution $\mathbb{P}$ on $\mathbb{R}^d \times \mathcal{S} \times \mathbb{R}$ if for all Borel sets $\mathcal{C} \subset \mathbb{R}$ it holds that*

$$\mathbb{P}\left(f(X, S) \in \mathcal{C} \,|\, S = -1\right) = \mathbb{P}\left(f(X, S) \in \mathcal{C} \,|\, S = 1\right) \quad .$$

In other words, a function f is fair if the total variation distance between the two conditional distributions of the function output associated to the two values of the sensitive feature is zero. For any Borel set $\mathcal{C} \subset \mathbb{R}$ and any predictor f, we also introduce the shorthand notation

$$\mathcal{U}(f, \mathcal{C}) := \left|\mathbb{P}\left(f(X, S) \in \mathcal{C} \,|\, S = -1\right) - \mathbb{P}\left(f(X, S) \in \mathcal{C} \,|\, S = 1\right)\right| \quad . \tag{1}$$

Finally, we define the fair optimal predictor as a minimizer of the risk under the fairness constraint, that is,

$$f^* \in \operatorname*{arg\,min}_{f \in \mathcal{F}} \left\{ \mathcal{R}(f) \;:\; \sup_{\mathcal{C} \subset \mathbb{R}} \mathcal{U}(f, \mathcal{C}) = 0 \right\} \quad . \tag{$\mathcal{P}$}$$

Notice that the feasible set of the problem ($\mathcal{P}$) is non-empty for any distribution $\mathbb{P}$ as it contains all constant predictors.

Remark 2.2. *In this work the sensitive attribute $s \in \mathcal{S}$ enters explicitly in the functional form of the predictor. However, in some applications (e.g. in the law domain) this may not be permitted. In Supplementary Material we show how to modify our methodology to address the case when the predictors take the form of $f : \mathbb{R}^d \to \mathbb{R}$.*

Let us also emphasize that, unlike previous theoretical investigations of fair regression [2, 47], we do not restrict $\mathcal{F}$. Throughout this work we pose the following boundedness assumption on the signal $Y \in \mathbb{R}$, which is also made in the above papers.

Assumption 2.3 (Bounded signal)**.** *There exists $M > 0$ such that $|Y| \leq M$ almost surely.*

The constant M or its upper bound is assumed known a-priori. This knowledge may naturally arise from the specific application at hand, *e.g.,* GPA of a student.

2.2 Reduction via finite discretization

The optimization problem ($\mathcal{P}$) is challenging, since it involves an uncountable number of constraints. To address this difficulty, a natural approach is to consider a proxy of problem ($\mathcal{P}$), based on a finite discretization step. To describe our observation, for any positive integer L, let $\mathcal{Q}_L$ be the uniform grid of $2L + 1$ points on $[-M, M]$, that is, $\mathcal{Q}_L = \{\ell M/L\}_{\ell=-L}^{L}$. Denote by $\mathcal{G}_L$ the set of measurable functions from $\mathbb{R}^d \times \mathcal{S}$ to $\mathcal{Q}_L$. The fair optimal discretized predictor $g_L^* : \mathbb{R}^d \times \mathcal{S} \to \mathcal{Q}_L$ is defined as

$$g_L^* \in \operatorname*{arg\,min}_{g \in \mathcal{G}_L} \left\{ \mathcal{R}(g) \;:\; \max_{q \in \mathcal{Q}_L} \mathcal{U}\left(g, \{q\}\right) = 0 \right\} \quad . \tag{$\mathcal{P}'_L$}$$

Note that unlike f^*, which takes values in the whole interval $[-M, M]$, the function g_L^* only takes values in the uniform grid $\mathcal{Q}_L$.

The following lemma confirms the intuition that for large values of L, the risk of g_L^* should be similar to that of f^*.

[2]For simplicity, in what follows we only consider the case of a binary sensitive feature. However, our methodology extends to non-binary case.

Lemma 2.4. *Let $\sigma^2 = \mathrm{Var}(Y)$. For every positive integer L, all solutions g_L^* of $(\mathcal{P}'_L)$ are fair in the sense of Definition 2.1. Moreover,*

$$\mathcal{R}(g_L^*) \le \mathcal{R}(f^*) + 2\sigma\frac{M}{L} + \frac{M^2}{L^2} \ .$$

Interestingly, problem $(\mathcal{P}'_L)$ can be solved analytically under the following mild assumption.

Assumption 2.5. *Assume, for all $s \in \mathcal{S}$, that the mappings $t \mapsto \mathbb{P}\left(\eta(X,s) \le t \mid S = s\right)$ are continuous.*

It is satisfied if the random variable $\eta(X,S)$ does not have atoms conditionally on $S = \pm 1$.

Proposition 2.6 (Optimal fair predictor)**.** *Under Assumption 2.5 for all positive integers L a solution g_L^* of problem $(\mathcal{P}'_L)$ is given for all $(x,s) \in \mathbb{R}^d \times \mathcal{S}$ by*

$$g_L^*(x,s) = \operatorname*{arg\,min}_{\ell \in \{-L,\dots,L\}} \left\{-s\lambda_\ell^* + p_s\left(\eta(x,s) - \tfrac{\ell M}{L}\right)^2\right\} \times \frac{M}{L} \ , \tag{2}$$

where $\lambda_{-L}^*, \dots, \lambda_L^*$ are solutions of

$$\min_{\lambda \in \mathbb{R}^{2L+1}} \sum_{s \in \mathcal{S}} \mathbb{E}_{X|S=s} \max_{\ell} \left\{s\lambda_\ell - p_s\left(\eta(X,s) - \tfrac{\ell M}{L}\right)^2\right\} \ . \tag{3}$$

The proof of this result borrows ideas from [16, 18]. In particular, we first write problem $(\mathcal{P}'_L)$ in the minmax form. It appears that its dual maxmin version can be solved analytically and Assumption 2.5 guarantees the strong duality.

The above result says that an optimal solution of the discretized fair regression problem $(\mathcal{P}'_L)$ is obtained by first computing the standard regression function η and then transforming this function via problems (2) and (3). In virtue of Lemma 2.4 a tempting approach to ultimately estimate the optimal fair regression function in problem $(\mathcal{P})$, would be to use an estimator of g_L^*, by first estimating the regression function η and then implementing an empirical version of problem (3). The next section describe in more details this estimator and, crucially, justify its choice by proving non-asymptotic error bounds for its excess risk and fairness constraint.

3 Proposed approach

In what follows we propose a data-driven procedure $\hat{g}$, which is based on *two* data samples: a labeled sample $\mathcal{D}_n{=}(X_i, S_i, Y_i)_{i=1}^n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$ of size n, and an independent unlabeled sample $\mathcal{D}'_N = (X'_i, S'_i)_{i=1}^N \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{(X,S)}$ of size N, where $\mathbb{P}_{(X,S)}$ is the marginal distribution of (X,S) induced by $\mathbb{P}$. That is, our algorithm is performed in a semi-supervised manner. The *principal goal* of this work is to construct a procedure $\hat{g}$ which meets two criteria:

$$\text{Fairness: } \mathbf{E}[\sup_{\mathcal{C} \in \mathbb{R}} \mathcal{U}(\hat{g}, \mathcal{C})] \le \delta_{n,N}, \qquad \text{Risk optimality: } \mathbf{E}[\mathcal{E}(\hat{g})] \le \delta'_{n,N},$$

where $\delta_{n,N}$ and $\delta'_{n,N}$ are two decreasing sequences of n and N, the excess risk $\mathcal{E}(f) := \mathcal{R}(f) - \mathcal{R}(f^*)$, and $\mathbf{E}$ is the expectation taken *w.r.t.* the distribution of the observations $\mathcal{D}_n, \mathcal{D}'_N$.

The proposed method is a plug-in approach which mimics the conditions imposed on g_L^* from Proposition 2.6. We require an off-the-shelf estimator $\hat{\eta}(X,S)$ of $\eta(X,S) = \mathbb{E}[Y|X,S]$ which is constructed using *only* the first *labeled* sample. This problem has been studied to a great extent and it is not of the main concern in this work. For instance such estimators include locally polynomial methods [39, 53], k-nearest neighbours [52, 20], random forests [9, 51], ridge and lasso regressions [3, 7], and many more. We also require the following, rather technical, assumption on the constructed estimator $\hat{\eta}$.

Assumption 3.1. *The mappings $t \mapsto \mathbb{P}\left(\hat{\eta}(X,s) \le t \mid S = s\right)$ are almost surely continuous.*

We refer to [17] for an in-depth discussion on this assumption and an ad-hoc method which allows to satisfy this condition for any estimator $\hat{\eta}$ and any distribution $\mathbb{P}_{X|S=s}$ which admits a density *w.r.t.* the Lebesgue measure. Yet, this assumption is of little or no concern for the practitioner as we demonstrate in our experimental study in Section 4.

To proceed with our plug-in method, we first decompose the unlabeled sample $\mathcal{D}'_N$ into three groups $\mathcal{D}'_{N_{-1}}$, $\mathcal{D}'_{N_1}$ and $\mathcal{D}^S_N$ of sizes N_{-1}, N_1, and N respectively. So that $\mathcal{D}'_{N_s} = \{X'_i \,:\, (X'_i, S'_i) \in \mathcal{D}'_N,\ S'_i = s\}$ for all $s \in \{-1, 1\}$ and $\mathcal{D}^S_N = \{S'_i \,:\, (X'_i, S'_i) \in \mathcal{D}'_N\}$.

Our next goal is to mimic the condition on $\lambda^*_{-L}, \ldots, \lambda^*_L$ imposed by Eq. (3), which requires the knowledge of η, p_s, and $\mathbb{P}_{X|S=s}$ for $s \in \mathcal{S}$ and $\ell \in \{-L, \ldots, L\}$. The estimator $\hat{p}_1$ of $p_1 = \mathbb{P}(S = 1)$ is based on the empirical frequencies on $\mathcal{D}^S_N$ and $\hat{p}_{-1} = 1 - \hat{p}_1$. For each $s \in \mathcal{S}$, the conditional expectation $\mathbb{E}_{X|S=s}$ is estimated using its empirical version on $\mathcal{D}'_{N_s}$ as $\hat{\mathbb{P}}_{X|S=s} = \frac{1}{N_s}\sum_{X' \in \mathcal{D}'_{N_s}} \delta_{X'}$. The *final estimator* $\hat{g}_L$ is then defined for all $(x, s) \in \mathbb{R}^d \times \mathcal{S}$ as

$$\hat{g}_L(x, s) = \operatorname*{arg\,min}_{\ell \in \{-L, \ldots, L\}} \left\{ -s\hat{\lambda}_\ell + \hat{p}_s \left(\hat{\eta}(x, s) - \frac{\ell M}{L} \right)^2 \right\} \times \frac{M}{L} \ , \tag{4}$$

where $\hat{\lambda}_{-L}, \ldots, \hat{\lambda}_L$ are solutions of

$$\min_{\lambda \in \mathbb{R}^{2L+1}} \sum_{s \in \mathcal{S}} \hat{\mathbb{E}}_{X|S=s} \max_{\ell} \left\{ s\lambda_\ell - \hat{p}_s \left(\hat{\eta}(X, s) - \frac{\ell M}{L} \right)^2 \right\} \ . \tag{5}$$

If the $\arg\min$ in Eq. (4) is not a singleton, we use the convention that the smallest value of ℓ is taken. Also notice that the minimization problem in Eq. (5) is convex. Therefore, it can be efficiently solved. In Section 3.2 we address this point and propose an efficient iterative algorithm based on the smoothing technique of Nesterov [45].

In summary, the proposed procedure is composed of two steps. First, we estimate the regression function η by standard methods using only labeled data, and then we estimate the thresholds $\lambda^*_{-L}, \ldots, \lambda^*_L$ using *unlabeled* data and the estimator $\hat{\eta}$ constructed on the first step. In many applications of fairness, an accurate initial estimator $\hat{\eta}$ is already available. Thus, our work suggests that in order to transform $\hat{\eta}$ into a fair predictor it is sufficient to gather only *unlabeled* data and solve the minimization problem in Eq. (5), which may be much cheaper than training a fair predictor from scratch.

3.1 Rates of convergence

In this section we present the rates of convergence of the proposed algorithm for an arbitrary value of $L \in \mathbb{N}$. These bounds demonstrate a bias-variance trade-off and a way to select L which optimizes it. We begin with bound on the violation of the fairness constraint of the proposed algorithm.

Theorem 3.2. *Under Assumption 3.1, there exists a universal constant $C > 0$ such that for each $L \in \mathbb{N}$ the proposed procedure $\hat{g}_L$ satisfies*

$$\mathbf{E}\left[\sup_{\mathcal{C} \subset \mathbb{R}} \mathcal{U}(\hat{g}_L, \mathcal{C})\right] \leq C \sum_{s \in \mathcal{S}} \sqrt{\frac{L}{p_s N}} \ .$$

Proof sketch. We first derive the first order optimality condition for the problem in Eq. (5). Since this problem is non-smooth (due to the $\max$) the optimality condition involves a sub-gradient of the objective. Using Assumption 3.1 we show that the non-smooth part of the objective has a little impact on the sub-gradient. On the final step, we show that the quantity of interest is controlled by a properly chosen empirical process plus the impact of the non-smooth part of the objective. □

The bound depends only on the size of the *unlabeled* dataset, and not on the quality of the initial estimator $\hat{\eta}$. It can be intuitively explained by the fact that the notion of fairness in Definition 2.1 depends only on the conditional distribution of X given S and not on the regression function η. A consequence of our findings is that when a large unlabeled dataset is available, achieving fairness becomes an easy task based only on the recalibration step we propose.

The next bound is on the excess-risk of the proposed algorithm. It establishes the trade-off introduced by the discretization step.

Theorem 3.3. *Let Assumptions 2.5 and 3.1 be satisfied. Then there exists a universal constant $C > 0$ such that for all $L \in \mathbb{N}$, the proposed procedure $\hat{g}_L$ satisfies*

$$\mathbf{E}[\mathcal{E}(\hat{g}_L)] \leq C M^2 \sum_{s \in \mathcal{S}} \left(\sqrt{\frac{L^2}{p_s N}} + \frac{1}{2L} \right) + 8M\mathbf{E}\left\|\eta - \hat{\eta}\right\|_1 \ .$$

Algorithm 1 Smoothed accelerated gradient descent

Input: temperature parameter β, number of iterations T
Initialize $\lambda_1 = z_1 = \tau_0 = 0$.
for $t = 1$ **to** T **do**
$\gamma_t = \frac{1-\tau_{t-1}}{\tau_t}$, $\tau_t = \frac{1+\sqrt{1+4\tau_{t-1}^2}}{2}$
$(z_{t+1})_\ell = (\lambda_t)_\ell - \frac{\beta}{2}\sum_{s\in\mathcal{S}} s\hat{\mathbb{E}}_{X|S=s}\left[\sigma_\beta\left(s(\lambda_t)_\ell - \hat{p}_s\left(\hat{\eta}(X,s) - \ell M/L\right)^2\right)\right], \quad \forall \ell$
$\lambda_{t+1} = (1-\gamma_t)z_{t+1} + \gamma_t z_t$
end for
Output: λ_T

Proof sketch. The proof of this result goes in two steps. On the first step we leverage the form of the optimal predictor g_L^* and the constructed plug-in rule $\hat{g}_L$ to show that $\mathcal{R}(\hat{g}_L) - \mathcal{R}(g_L^*)$ can be bounded by two terms. The first term involves the violation of the fairness constraints and is controlled by Theorem 3.2. The second term can be controlled by the estimation error of $\hat{\eta}$ and $\hat{p}_s$. Finally, we combine Lemma 2.4 with the bound on $\mathcal{R}(\hat{g}_L) - \mathcal{R}(g_L^*)$ to obtain the result on $\mathcal{E}(\hat{g}_L)$. □

Unlike the bound on fairness, the excess-risk bound already depends on the quality of $\hat{\eta}$. Importantly, the last term in the above bound decreases with n instead of $p_s n$, that is, this term is not affected by the unbalanced distributions. Finally, from the excess-risk bound we can observe that the parameter L should be chosen in an optimal way, performing the bias-variance trade-off. Setting $L = N^{1/4}$ in the previous results we immediately get the following corollary.

Corollary 3.4. *Let Assumptions 2.5 and 3.1 be satisfied and let $L = N^{1/4}$. Then there exists universal constants $C, C' > 0$ such that the proposed procedure $\hat{g}_L$ satisfies*

$$\mathbf{E}\left[\sup_{\mathcal{C}\subset\mathbb{R}} \mathcal{U}(\hat{g}_L, \mathcal{C})\right] \le C\sum_{s\in\mathcal{S}}(p_s^{\frac{8}{5}}N)^{-\frac{3}{8}} \quad \textit{and} \quad \mathbf{E}[\mathcal{E}(\hat{g}_L)] \le C'M^2\sum_{s\in\mathcal{S}}(p_s^2 N)^{-\frac{1}{4}} + 8M\mathbf{E}\left\|\eta - \hat{\eta}\right\|_1 \quad .$$

Note that the choice of L is independent from the size of the labeled data n and it does not affect the second term on the right hand side of the excess-risk guarantee. A careful analysis of our proof reveals that a data driven choice of L that depends on $\hat{p}_s$ would improve the above result. Namely, instead of $p_s^2 N$ we could obtain $p_s N$. However, this proof is much more technical and is thus omitted. Finally, we remark that the result of Corollary 3.4 and both Theorems 3.2–3.3 explicitly assumes that $\hat{\lambda}_{-L}, \dots, \hat{\lambda}_L$ is a global minimizer of the problem in Eq. (5). In the next section we provide an optimization algorithm which finds an ϵ solution of the problem in Eq. (5).

3.2 Optimization algorithm

Recall that the proposed estimator sets $\hat{\lambda}_{-L}, \dots, \hat{\lambda}_L$ to be a solution of the minimization problem in Eq. (5). This problem is convex but non-smooth, thus subgradient methods can be used to numerically approximate a solution. While being optimal in a black-box optimization paradigm [46], subgradient methods often can be significantly accelerated if the structure of the non-smooth problem is "simple". In our setting, we follow the smoothing technique due to Nesterov [45], which leads to Algorithm 1. The key insight in this approach is to approximate the inner maximum in the objective function of Eq. (5) by a smooth convex function with Lipschitz gradient. This results in the LogSumExp (also known as soft-max) instead of the "hard" max. Such smoothed problem is then solved using an optimal method, such as the accelerated gradient descent [44].

To understand the proposed optimization algorithm, let us introduce some notation. For any vector $\lambda \in \mathbb{R}^{2L+1}$, the soft argmax (also known as Gibbs distribution) of λ with the temperature parameter β is defined component-wise for all $\ell \in \{-L, \dots, L\}$ as $\sigma_\beta(\lambda)_\ell := \exp\left(\frac{\lambda_\ell}{\beta}\right)/\sum_{\ell=-L}^{L}\exp\left(\frac{\lambda_\ell}{\beta}\right)$. Finally, denote by $G : \mathbb{R}^{2L+1} \to \mathbb{R}$ the objective function of the minimization in Eq. (5). That is, the vector $\hat{\lambda} = (\hat{\lambda}_{-L}, \dots, \hat{\lambda}_L)^\top$ is a minimizer of $\min_{\lambda\in\mathbb{R}^{2L+1}} G(\lambda)$. To find an ε-solution of this problem we run Algorithm 1, which takes as an input two parameters $T \in \mathbb{N}$ and $\beta > 0$.

Method	CRIME MSE	CRIME KS	LAW MSE	LAW KS	NLSY MSE	NLSY KS	STUD MSE	STUD KS	UNIV MSE	UNIV KS
RLS	$.033{\pm}.003$	$.55{\pm}.06$	$.107{\pm}.010$	$.15{\pm}.02$	$.153{\pm}.016$	$.73{\pm}.07$	$4.77{\pm}.49$	$.50{\pm}.05$	$2.24{\pm}.22$	$.14{\pm}.01$
RLS+Berk	$.037{\pm}.004$	$.16{\pm}.02$	$.121{\pm}.013$	$.10{\pm}.01$	$.189{\pm}.019$	$.49{\pm}.05$	$5.28{\pm}.57$	$.32{\pm}.03$	$2.43{\pm}.23$	$.05{\pm}.01$
RLS+Oneto	$.037{\pm}.004$	$.14{\pm}.01$	$.112{\pm}.012$	$.07{\pm}.01$	$.156{\pm}.016$	$.50{\pm}.05$	$5.02{\pm}.54$	$.23{\pm}.02$	$2.44{\pm}.26$	$.05{\pm}.01$
RLS+Ours	$.035{\pm}.003$	$.22{\pm}.02$	$.117{\pm}.011$	$.04{\pm}.01$	$.177{\pm}.017$	$.16{\pm}.02$	$5.14{\pm}.46$	$.11{\pm}.01$	$2.63{\pm}.24$	$.04{\pm}.01$
KRLS	$.024{\pm}.003$	$.52{\pm}.05$	$.040{\pm}.004$	$.09{\pm}.01$	$.061{\pm}.006$	$.58{\pm}.06$	$3.85{\pm}.36$	$.47{\pm}.05$	$1.43{\pm}.15$	$.10{\pm}.01$
KRLS+Oneto	$.028{\pm}.003$	$.19{\pm}.02$	$.046{\pm}.004$	$.05{\pm}.01$	$.066{\pm}.007$	$.06{\pm}.01$	$4.07{\pm}.39$	$.18{\pm}.02$	$1.46{\pm}.13$	$.04{\pm}.01$
KRLS+Perez	$.033{\pm}.003$	$.25{\pm}.02$	$.048{\pm}.005$	$.04{\pm}.01$	$.065{\pm}.007$	$.08{\pm}.01$	$3.97{\pm}.38$	$.14{\pm}.02$	$1.50{\pm}.15$	$.06{\pm}.01$
KRLS+Ours	$.032{\pm}.003$	$.12{\pm}.01$	$.050{\pm}.005$	$.02{\pm}.01$	$.070{\pm}.007$	$.05{\pm}.01$	$4.04{\pm}.37$	$.06{\pm}.01$	$1.51{\pm}.15$	$.02{\pm}.01$
RF	$.020{\pm}.002$	$.45{\pm}.04$	$.046{\pm}.005$	$.11{\pm}.01$	$.055{\pm}.006$	$.55{\pm}.06$	$3.59{\pm}.39$	$.45{\pm}.05$	$1.31{\pm}.13$	$.10{\pm}.01$
RF+Raff	$.030{\pm}.003$	$.21{\pm}.02$	$.058{\pm}.006$	$.06{\pm}.01$	$.066{\pm}.006$	$.08{\pm}.01$	$4.28{\pm}.40$	$.09{\pm}.01$	$1.38{\pm}.12$	$.02{\pm}.01$
RF+Agar	$.029{\pm}.003$	$.13{\pm}.01$	$.050{\pm}.005$	$.04{\pm}.01$	$.065{\pm}.006$	$.07{\pm}.01$	$3.87{\pm}.41$	$.07{\pm}.01$	$1.40{\pm}.13$	$.02{\pm}.01$
RF+Ours	$.031{\pm}.003$	$.09{\pm}.01$	$.061{\pm}.006$	$.03{\pm}.01$	$.064{\pm}.007$	$.05{\pm}.01$	$3.93{\pm}.36$	$.05{\pm}.01$	$1.41{\pm}.14$	$.02{\pm}.01$

Table 1: Results for all the datasets and all the methods concerning MSE and KS when the sensitive feature is exploited in the functional form of the model.

Theorem 3.5. *For every $L > 0$ and every $\varepsilon > 0$ the output λ_T of Algorithm 1 with*

$$\beta = \frac{M^2\sqrt{2L+1}}{T\log(2L+1)} \quad \textit{and} \quad T = \left\lceil \frac{256M^2}{\varepsilon}\sqrt{(2L+1)}\log(2L+1)\right\rceil \ ,$$

satisfies $G(\lambda_T) - G(\hat{\lambda}) \le \varepsilon$.

Unlike subgradient methods that require $T = O(\varepsilon^{-2})$ iterations to achieve ε-solution, smoothing technique allows to achieve T of order ε^{-1} as stated in Theorem 3.5. More precisely, when we set $L = N^{1/4}$ as suggested by Corollary 3.4, $T = O(\varepsilon^{-1}N^{1/8}\log(N))$. Following our statistical results a reasonable choice of the optimization accuracy is $\varepsilon = O(N^{-1/4})$, implying that the total amount of iterations $T = O(N^{3/8}\log(N))$.

Remark 3.6. *We did not attempt to improve the constant* 256 *present in the choice of T, as our main interest in this result is the dependence on N and ε.*

On each iteration, Algorithm 1 computes the soft argmax function and averages it over the unlabeled dataset, which can be done in time linear in N. Note that the averaging step involves only unlabeled data and can be pre-computed before running the algorithm. Finally, to compute the estimator $\hat{g}_L(x)$ at a new point x (see Eq. (4)) we need to find the minimum over a finite set, which is performed in time linear in $L = N^{1/4}$.

4 Empirical study

In this section, we present numerical experiments[3] with the proposed fair regression estimator.

Experimental setting. In all experiments, we collect statistics on the test set $\mathcal{T} = \{(X_i, S_i, Y_i)\}_{i=1}^{|\mathcal{T}|}$. The empirical mean squared error (MSE) is defined as

$$\mathrm{MSE}\,(g) = \frac{1}{|\mathcal{T}|}\sum_{(X,S,Y)\in\mathcal{T}}(Y - g(X,S))^2 \ .$$

We also would like to measure the violation of fairness constraint imposed by Definition 2.1. Since it involves a computationally expensive TV variation distance, we replace it by the empirical Kolmogorov-Smirnov distance:

$$\mathrm{KS}\,(g) = \sup_{t\in\mathbb{R}}\left|\frac{1}{|\mathcal{T}_{-1}|}\sum_{(X,Y)\in\mathcal{T}_{-1}}\mathbf{1}_{\{g(X,-1)\le t\}} - \frac{1}{|\mathcal{T}_{+1}|}\sum_{(X,Y)\in\mathcal{T}_{+1}}\mathbf{1}_{\{g(X,+1)\le t\}}\right| \ ,$$

[3]The source of our method can be found at `https://github.com/lucaoneto/NIPS2020_Fairness`.

where for all $s\in\{-1,+1\}$ the set $\mathcal{T}_s=\{(X,S,Y)\in\mathcal{T}\,:\,S=s\}$. For all datasets we split the data in two parts (70% train and 30% test), this procedure is repeated 30 times, and we report the average performance on the test set alongside its standard deviation. We employ the 2-steps 10-fold CV procedure considered by [22] to select the best hyperparameters with the training set. In the first step, we shortlist all the hyperparameters with MSE close to the best one (in our case, the hyperparameters which lead to 10% larger MSE w.r.t. the best MSE). Then, from this list, we select the hyperparameters with the lowest KS.

Methods. We compare our method to different fair regression approaches for both linear and non-linear regression. In the case of linear models we consider the following methods: Linear RLS plus [6] (RLS+Berk), Linear RLS plus [47] (RLS+Oneto), and Linear RLS plus Our Method (RLS+Ours), where RLS is the abbreviation of Regularized Least Squares. In the case of non-linear models we compare to the following methods: Kernel RLS (KRLS), Kernel RLS plus [47] (KRLS+Oneto), Kernel RLS plus [48] (KRLS+Perez), Kernel RLS plus Our Method (KRLS+Ours), Random Forests (RF), Random Forests plus [49] (RF+Raff), Random Forests plus [2][4] (RF+Agar), and Random Forests plus Our Method (RF+Ours).

The hyperparameters of the methods are set as follows. As our theory suggests that $L = N^{1/4}$ leads to a statistically grounded approach, we choose $L \in \{6, 12, 24\}$ since the size of the considered datasets is smaller than $24^4 \approx 3\times10^5$ and $\beta \in \{0.1, 0.01\}$. For RLS we set the regularization hyperparameters $\lambda \in 10^{\{-4.5,-3.5,\cdots,3\}}$ and for KRLS we set $\lambda \in 10^{\{-4.5,-3.5,\cdots,3\}}$ and $\gamma \in 10^{\{-4.5,-3.5,\cdots,3\}}$. Finally, for RF we set to 1000 the number of trees and for the number of features to select during the tree creation we search in $\{d^{1/4}, d^{1/2}, d^{3/4}\}$.

Datasets. In order to analyze the performance of our methods and test it against the state-of-the-art alternatives, we consider five benchmark datasets, CRIME, LAW, NLSY, STUD, and UNIV, which are briefly described below:

Communities&Crime (CRIME) contains socio-economic, law enforcement, and crime data about communities in the US [50] with 1994 examples. The task is to predict the number of violent crimes per 10^5 population (normalized to $[0,1]$) with race as the protected attribute. Following [12], we made a binary sensitive attribute s as to the percentage of black population, which yielded 970 instances of s=1 with a mean crime rate 0.35 and 1024 instances of s=-1 with a mean crime rate 0.13.

Law School (LAW) refers to the Law School Admissions Councils National Longitudinal Bar Passage Study [56] and has 20649 examples. The task is to predict a students GPA (normalized to $[0,1]$) with race as the protected attribute (white versus non-white).

National Longitudinal Survey of Youth (NLSY) involves survey results by the U.S. Bureau of Labor Statistics that is intended to gather information on the labor market activities and other life events of several groups [10]. Analogously to [37] we model a virtual company's hiring decision assuming that the company does not have access to the applicants' academic scores. We set as target the person's GPA (normalized to $[0,1]$), with race as sensitive attribute.

Student Performance (STUD), approaches 649 students achievement (final grade) in secondary education of two Portuguese schools using 33 attributes [19], with gender as the protected attribute.

University (UNIV)[5] is a proprietary and highly sensitive dataset containing all the data about the past and present students enrolled at the University of *Genoa*. In this study we take into consideration students who enrolled, in the academic year 2017-2018. The dataset contains 5000 instances, each one described by 35 attributes (both numeric and categorical) about ethnicity, gender, financial status, and previous school experience. The scope is to predict the average grades at the end of the first semester, with gender as the protected attribute.

Comparison w.r.t. state-of-the-art. In Table 1, we present the performance of different methods on various datasets described in Section 4. Our findings indicate that the proposed method is generally superior or competitive with state-of-the-art methods. In particular, our method is extremely good in enforcing fairness, even though, often, this comes at the cost of a slight increase in the MSE. Overall, RF+Ours tends to be the most effective method, and the one we would recommend to use in practice.

[4]We thank the authors for sharing a prototype of their code.

[5]The data and the research are related to the project DROP@UNIGE of the University of Genoa.

5 Discussion and conclusion

We proposed a new method to fair regression, which is able to estimate the optimal fair regression function, when the demographic parity constraint is imposed. This approach is very general and can be employed on top of any standard estimator, by means of the recalibration step which only involves an additional independent unlabeled dataset. This step can be efficiently implemented by solving a small-scale convex optimization problem. We derived non-asymptotic error rates for this estimator, relative to both the squared risk and a fairness violation based on the total variation distance. Numerical experiments demonstrated that the proposed method is effective and often superior to previous fair regression methods. In future it would be valuable to study the theoretical impact of the smoothing parameter on the risk/fairness trade-off as well as to understand the optimality of our bounds. Finally, an important open problem is whether an estimator having the same guarantees as the proposed one, could be constructed on the basis of a single dataset, used both to estimate the regression function and the recalibrations step.

Broader impact

Although theoretical, our work may have at least two indirect positive future societal effects. First, the proposed algorithm is easy to implement and use, since it works in a post-processing regime. This makes it potentially attractive to practitioners who use computationally demanding methods. Second, our procedure comes with strong theoretical guarantees, making our contribution more reliable in practice.

Despite this potentially positive impact, one should be aware that the notion of demographic parity is only a way to formalize the idea of fairness in a rigorous mathematical manner and, as other similar formalizations, it might not reflect the reality. Hence, while this fact does not compromise our theoretical contribution, one has to pay extra care to the choice of the fairness notion when machine learning algorithms are to be deployed to society.

Acknowledgments and Disclosure of Funding

This work was partially supported Amazon Web Services and SAP SE.

References

[1] Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453*, 2018.

[2] Agarwal, A., Dudik, M., and Wu, Z. S. Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning*, 2019.

[3] Arlot, S. and Bach, F. Data-driven calibration of linear estimators with minimal penalties. In *Advances in Neural Information Processing Systems*, pp. 46–54, 2009.

[4] Barocas, S., Hardt, M., and Narayanan, A. *Fairness and Machine Learning*. fairmlbook.org, 2018.

[5] Beck, A. and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[6] Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., Neel, S., and Roth, A. A convex framework for fair regression. In *Fairness, Accountability, and Transparency in Machine Learning*, 2017.

[7] Bickel, P. J., Ritov, Y., and Tsybakov, A. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009.

[8] Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989.

[9] Breiman, L. Consistency for a simple model of random forests. Technical report, Statistics Department University Of California At Berkeley, 2004.

[10] Bureau of Labor Statistics. National longitudinal surveys of youth data set. `www.bls.gov/nls/`, 2019.

[11] Calders, T., Kamiran, F., and Pechenizkiy, M. Building classifiers with independency constraints. In *IEEE international conference on Data mining*, 2009.

[12] Calders, T., Karim, A., Kamiran, F., Ali, W., and Zhang, X. Controlling attribute effect in linear regression. In *IEEE International Conference on Data Mining*, 2013.

[13] Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K. N., and Varshney, K. R. Optimized pre-processing for discrimination prevention. In *Neural Information Processing Systems*, 2017.

[14] Chiappa, S., Jiang, R., Stepleton, T., Pacchiano, A., Jiang, H., and Aslanides, J. A general approach to fairness with optimal transport. In *AAAI*, 2020.

[15] Chierichetti, F., Kumar, R., Lattanzi, S., and Vassilvitskii, S. Fair clustering through fairlets. In *Neural Information Processing Systems*, 2017.

[16] Chzhen, E. *Plug-in methods in classification*. PhD thesis, Université Paris-Est, September 2019. URL `https://tel.archives-ouvertes.fr/tel-02400552`.

[17] Chzhen, E., Denis, C., and Hebiri, M. Minimax semi-supervised confidence sets for multi-class classification. *arXiv preprint arXiv:1904.12527*, 2019.

[18] Chzhen, E., Denis, C., Hebiri, M., Oneto, L., and Pontil, M. Leveraging labeled and unlabeled data for consistent fair binary classification. In *Advances in Neural Information Processing Systems 32*, pp. 12739–12750. Curran Associates, Inc., 2019.

[19] Cortez, P. and Silva, A. Using data mining to predict secondary school student performance. In *FUture BUsiness TEChnology Conference*, 2008.

[20] Devroye, L. The uniform convergence of nearest neighbor regression function estimators and their application in optimization. *IEEE Transactions on Information Theory*, 24(2):142–151, 1978.

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

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

[23] Dudley, R. The sizes of compact subsets of Hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.

[24] Dwork, C., Immorlica, N., Kalai, A. T., and Leiserson, M. D. M. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on Fairness, Accountability and Transparency*, 2018.

[25] Eisenstat, D. and Angluin, D. The VC dimension of k-fold union. *Information Processing Letters*, 101(5):181–184, 2007.

[26] Fitzsimons, J., Ali, A. A., Osborne, M., and Roberts, S. Equality constrained decision trees: For the algorithmic enforcement of group fairness. *arXiv preprint arXiv:1810.05041*, 2018.

[27] Gao, B. and Pavel, L. On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805*, 2017.

[28] Gordaliza, P., Del Barrio, E., Fabrice, G., and Loubes, J. M. Obtaining fairness using optimal transport theory. In *International Conference on Machine Learning*, 2019.

[29] Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Neural Information Processing Systems*, 2016.

[30] Haussler, D. Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69(2):217–232, 1995.

[31] Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., and Roth, A. Fair learning in markovian environments. In *Conference on Fairness, Accountability, and Transparency in Machine Learning*, 2016.

[32] Jiang, R., Pacchiano, A., Stepleton, T., Jiang, H., and Chiappa, S. Wasserstein fair classification. *arXiv preprint arXiv:1907.12059*, 2019.

[33] Joseph, M., Kearns, M., Morgenstern, J. H., and Roth, A. Fairness in learning: Classic and contextual bandits. In *Neural Information Processing Systems*, 2016.

[34] Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. Avoiding discrimination through causal reasoning. In *Neural Information Processing Systems*, 2017.

[35] Koltchinskii, V. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Springer Science & Business Media, 2011.

[36] Komiyama, J. and Shimao, H. Two-stage algorithm for fairness-aware machine learning. *arXiv preprint arXiv:1710.04924*, 2017.

[37] Komiyama, J. and Shimao, H. Comparing fairness criteria based on social outcome. *arXiv preprint arXiv:1806.05112*, 2018.

[38] Komiyama, J., Takeda, A., Honda, J., and Shimao, H. Nonconvex optimization for regression with fairness constraints. In *International Conference on Machine Learning*, 2018.

[39] Korostelëv, A. and Tsybakov, A. *Minimax theory of image reconstruction*, volume 82 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1993.

[40] Kusner, M. J., Loftus, J., Russell, C., and Silva, R. Counterfactual fairness. In *Neural Information Processing Systems*, 2017.

[41] Lum, K. and Johndrow, J. A statistical framework for fair predictive algorithms. *arXiv preprint arXiv:1610.08077*, 2016.

[42] Matoušek, J. *Lectures on discrete geometry*, volume 108. Springer, 2002.

[43] Nabi, R., Malinsky, D., and Shpitser, I. Learning optimal fair policies. *International Conference on Machine Learning*, 2019.

[44] Nesterov, Y. A method for solving the convex programming problem with convergence rate O(1/kˆ2). In *Doklady Akademii Nauk SSSR*, 1983.

[45] Nesterov, Y. Smooth minimization of non-smooth functions. *Mathematical programming*, 103 (1):127–152, 2005.

[46] Nesterov, Y. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[47] Oneto, L., Donini, M., and Pontil, M. General fair empirical risk minimization. *arXiv preprint arXiv:1901.10080*, 2019.

[48] Pérez-Suay, A., Laparra, V., Mateo-García, G., Muñoz-Marí, J., Gómez-Chova, L., and Camps-Valls, G. Fair kernel learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2017.

[49] Raff, E., Sylvester, J., and Mills, S. Fair forests: Regularized tree induction to minimize model bias. In *AAAI/ACM Conference on AI, Ethics, and Society*, 2018.

[50] Redmond, M. and Baveja, A. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141(3):660–678, 2002.

[51] Scornet, E., Biau, G., and Vert, J. Consistency of random forests. *Ann. Statist.*, 43(4):1716–1741, 08 2015.

[52] Stone, C. Consistent nonparametric regression. *Ann. Statist.*, pp. 595–620, 1977.

[53] Tsybakov, A. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009.

[54] Vapnik, V. and Chervonenkis, A. On the uniform convergence of relative frequencies of events to their probabilities. *Doklady Akademii Nauk SSSR*, 181(4):781–787, 1968.

[55] Vershynin, R. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

[56] Wightman, L. F. and Ramsey, H. *LSAC national longitudinal bar passage study*. Law School Admission Council, 1998.

[57] Yao, S. and Huang, B. Beyond parity: Fairness objectives for collaborative filtering. In *Neural Information Processing Systems*, 2017.

[58] Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *International Conference on World Wide Web*, 2017.

[59] Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In *International Conference on Machine Learning*, 2013.

[60] Zliobaite, I. On the relation between accuracy and fairness in binary classification. *arXiv preprint arXiv:1505.05723*, 2015.

Supplementary Material

Below we give an overview of the structure of the supplementary material and highlight the main novel results of this work.

- Appendix A is mainly devoted to the derivation of the expression for the optimal predictor g^*_L. The proof of Lemma 2.4 is also placed in this section.
- Appendix B states general preparation results which are used for the proof of fairness rates. Appendix B.1 is devoted to the proof of Theorem 3.2, which establishes fairness guarantees of the proposed procedure.
- Similarly, Appendix C starts by stating supporting results, whose proofs are postponed to Section C.2. Appendix C.1 is devoted to the proof of Theorem 3.3, which establishes guarantees on the excess-risk of the proposed procedure.
- Appendix D is devoted to the optimization part of our contribution and establishes guarantees on Algorithm 1.
- Appendix F shows the impact of unlabeled data on the performance of the estimator.

Let us also mention that in the supplementary material we omit the underscript L, when no confusion can rise. That is, instead of g^*_L and $\hat{g}_L$ we write g^* and $\hat{g}$ respectively. Finally, before proceeding further, let us point out one technical subtlety: in what follows it is assumed that the estimator $|\hat{\eta}(\cdot, s)| \leq M$, this assumption is never restrictive in practice as long as M is known. Indeed, if $\hat{\eta}(\cdot, s)$ take values outside of $[-M, M]$, then its truncationn on this interval is strictly better in terms of the ℓ_1 error, since the true $|\eta(\cdot, s)| \leq M$.

A Derivation of the optimal predictor and its properties

First we state a rather intuitive statement. Informally, if the signal Y is almost surely bounded on the interval $[-M, M]$, then the fair optimal predictor f^* is also bounded almost surely on the interval $[-M, M]$. This result allows to consider only those predictors f, which take values in $[-M, M]$.

Lemma A.1. *Assume that $|Y| \leq M$ almost surely, then $|f^*(X, S)| \leq M$ almost surely.*

Proof. Let f^* be the minimizer of problem $\mathcal{P}$. Denote by $f \mapsto \Pi_f$ the projection defined as

$$\Pi_f(x, s) = f(x, s)\mathbf{1}_{\{|f(x,s)| \leq M\}} + M \operatorname{sign}(f(x, s))\mathbf{1}_{\{|f(x,s)| > M\}} \enspace .$$

Now our goal is to show that Π_{f^*} is fair in the sense of Definition 2.1 and that its risk is upper bounded by the risk of f^*. This would imply that $\Pi_{f^*} = f^*$ almost surely. The fairness of Π_{f^*} follows directly from the fairness of f^*. Moreover, we can write

$$\begin{aligned}
\mathbb{E}(Y - \Pi_{f^*}(X, S))^2 &= \mathbb{E}(Y - f^*(X, S) + f^*(X, S) - \Pi_{f^*}(X, S))^2 \\
&= \mathbb{E}(Y - f^*(X, S))^2 + 2\mathbb{E}(Y - f^*(X, S))(f^*(X, S) - \Pi_{f^*}(X, S)) \\
&\quad + \mathbb{E}(f^*(X, S) - \Pi_{f^*}(X, S))^2 \enspace .
\end{aligned}$$

Let us introduce the following notation

$$Z = 2(Y - f^*(X, S))(f^*(X, S) - \Pi_{f^*}(X, S)) + (f^*(X, S) - \Pi_{f^*}(X, S))^2 \enspace .$$

Notice that

$$Z = (2Y - f^*(X, S) - \Pi_{f^*}(X, S))(f^*(X, S) - \Pi_{f^*}(X, S)) \enspace .$$

If we can show that $Z \leq 0$ almost surely, the proof is finished. To see this, we first notice that

$$\begin{aligned}
f^*(X, S) - \Pi_{f^*}(X, S) &= (|f^*(X, S)| - M)\operatorname{sign}(f^*(X, S))\mathbf{1}_{\{|f^*(X,S)| > M\}} \enspace , \\
f^*(X, S) + \Pi_{f^*}(X, S) &= 2f^*(X, S)\mathbf{1}_{\{|f^*(x,s)| \leq M\}} \\
&\quad + (M + |f^*(X, S)|)\operatorname{sign}(f^*(X, S))\mathbf{1}_{\{|f^*(X,S)| > M\}} \enspace .
\end{aligned}$$

After simple algebraic manipulations Z can be expressed as

$$\begin{aligned}
Z &= (2Y\,\mathrm{sign}(f^*(X,S)) - M - |f^*(X,S)|)\,(|f^*(X,S)| - M)\,\mathbf{1}_{\{|f^*(X,S)|>M\}}\\
&\leq 2\,(Y\,\mathrm{sign}(f^*(X,S)) - M)\,(|f^*(X,S)| - M)\,\mathbf{1}_{\{|f^*(X,S)|>M\}}\\
&\leq 2\,(|Y| - M)\,(|f^*(X,S)| - M)\,\mathbf{1}_{\{|f^*(X,S)|>M\}} \enspace .
\end{aligned}$$

Finally, since $|Y| \leq M$ we conclude. $\square$

Now, we prove Lemma 2.4, which gives a theoretical justification to the reduction scheme and the introduction of g_L^*. Let us recall the statement of this result first.

Lemma (Lemma 2.4)**.** *For every positive integer L, all solutions g_L^* of $(\mathcal{P}_L')$ are fair in the sense of Definition 2.1. Moreover*

$$\mathcal{R}(g_L^*) \leq \mathcal{R}(f^*) + 2\sigma\frac{M}{L} + \frac{M^2}{L^2} \enspace ,$$

where $\sigma^2 = \mathrm{Var}(Y)$.

Proof of Lemma 2.4. First we show that g_L^* is fair. Fix arbitrary $C \in [-M, M]$, thus for any $s \in \mathcal{S}$ we can write

$$\begin{aligned}
\mathbb{P}\left(g_L^*(X,S) \in C \,|\, S = s\right) &= \mathbb{P}\left(g_L^*(X,S) \in C \cap \mathcal{Q}_{L,M} \,|\, S = s\right)\\
&= \sum_{y \in C \cap \mathcal{Q}_{L,M}} \mathbb{P}\left(g_L^*(X,S) = y \,|\, S = s\right) \enspace .
\end{aligned}$$

Every $y \in C \cap \mathcal{Q}_{L,M}$ can be expressed as $\ell M/L$ for some $\ell \in \{-L, \ldots, L\}$ and for every $\ell \in \{-L, \ldots, L\}$

$$\mathbb{P}\left(g_L^*(X,S) = \ell M/L \,|\, S = -1\right) = \mathbb{P}\left(g_L^*(X,S) = \ell M/L \,|\, S = 1\right) \enspace ,$$

which implies that g_L^* is fair.

Finally, to demonstrate the inequality in this result we first construct an operator $T_L : \mathcal{F} \to \mathcal{G}_{L,M}$ defined point-wise for all $(x,s) \in \mathbb{R}^d \times \mathcal{S}$ as

$$(T_L(f))(x,s) = \lfloor Lf(x,s)/M \rfloor M/L \enspace ,$$

where for $x \in \mathbb{R}$, $\lfloor x \rfloor$ stands for the closest integer smaller or equal to x. Now, we show that $T_L(f^*)$ is feasible for problem $(\mathcal{P}_L')$. Indeed, for any $\ell \in \{-L, \ldots, L-1\}$ and any $(x,s) \in \mathbb{R}^d \times \mathcal{S}$, by construction of T_L, we have

$$(T_L(f^*))(x,s) = \ell M/L \qquad \Leftrightarrow \qquad f^*(x,s) \in \left[\frac{\ell M}{L}, \frac{(\ell+1)M}{L}\right) \enspace .$$

Therefore, since f^* is fair and the set $[\ell M/L, (\ell+1)M/L)$ is Borel we have for all $\ell \in \{-L, \ldots, L-1\}$

$$\mathbb{P}\left((T_L(f^*))(X,S) = \ell M/L \,|\, S = -1\right) = \mathbb{P}\left((T_L(f^*))(X,S) = \ell M/L \,|\, S = 1\right) \enspace .$$

Moreover, we also have for all $(x,s) \in \mathbb{R}^d \times \mathcal{S}$

$$T_L(f^*)(x,s) = M \qquad \Leftrightarrow \qquad f^*(x,s) = M \enspace ,$$

which implies that for $\ell = L$ we have

$$\mathbb{P}\left((T_L(f^*))(X,S) = \ell M/L \,|\, S = -1\right) = \mathbb{P}\left((T_L(f^*))(X,S) = \ell M/L \,|\, S = 1\right) \enspace .$$

Thus, $T_L(f^*)$ is feasible for problem $(\mathcal{P}_L')$ and we can write

$$\begin{aligned}
\mathbb{E}(Y - g_L^*(X,S))^2 &\leq \mathbb{E}\Big(Y - (T_L(f^*))(X,S)\Big)^2\\
&= \mathbb{E}(Y - f^*(X,S))^2 + \mathbb{E}\Big(f^*(X,S) - (T_L(f^*))(X,S)\Big)^2\\
&\quad + 2\mathbb{E}(Y - f^*(X,S))(f^*(X,S) - (T_L(f^*))(X,S)) \enspace .
\end{aligned}$$

Notice that for all (x, s) we have $|f^*(x, s) - (T_L(f^*))(x, s)| \leq M/L$, and thus using the Cauchy-Schwartz inequality we get

$$\mathbb{E}(Y - g_L^*(X, S))^2 \leq \mathbb{E}(Y - f^*(X, S))^2 + 2M\frac{\sqrt{\mathbb{E}(Y - f^*(X, S))^2}}{L} + \frac{M^2}{L^2} \enspace .$$

Finally, since $f(x, s) \equiv \mathbb{E}[Y]$ is a feasible function for problem ($\mathcal{P}$), we have

$$\mathbb{E}(Y - f^*(X, S))^2 \leq \operatorname{Var}(Y) \enspace ,$$

which concludes the proof. □

The next proof is devoted to the derivation of the optimal predictor g_L^* provided in Proposition 2.6. Below we recall the statement of Proposition 2.6.

Proposition (Proposition 2.6)**.** *Under Assumption 2.5 for all positive integers L a solution g_L^* of problem ($\mathcal{P}'_L$) is given for all $(x, s) \in \mathbb{R}^d \times \mathcal{S}$ by*

$$g_L^*(x, s) = \mathop{\arg\min}_{\ell \in \{-L, \ldots, L\}} \{-s\lambda_\ell^* + Z_\ell(x, s)\} \times \frac{M}{L} \enspace ,$$

where, for every $s \in \mathcal{S}$ and $\ell \in \{-L, \ldots, L\}$, we have defined the quantity $Z_\ell(x, s) = p_s\left(\eta(X, s) - \frac{\ell M}{L}\right)^2$ and $\lambda_{-L}^*, \ldots, \lambda_L^*$ are solutions of

$$\min_{\lambda \in \mathbb{R}^{2L+1}} \sum_{s \in \mathcal{S}} \mathbb{E}_{X|S=s} \max_{\ell} \{s\lambda_\ell - Z_\ell(X, s)\} \enspace .$$

Proof of Proposition 2.6. Our goal is to solve the following problem

$$\min_g \max_{\lambda \in \mathbb{R}^{2L+1}} \mathbb{E}(Y - g(X, S))^2 + \sum_{\ell=-L}^{L} \lambda_\ell \left(\mathbb{P}\left(g(X, -1) = \ell M/L \mid S = -1\right) - \mathbb{P}\left(g(X, 1) = \ell M/L \mid S = 1\right)\right) \enspace .$$

First of all notice that the minimization of $\mathbb{E}(Y - g(X, S))^2$ is equivalent to the minimization of $\mathbb{E}_{(X,S)}(\eta(X, S) - g(X, S))^2$, where $\eta(X, S) = \mathbb{E}[Y|X, S]$. Therefore, instead of the above saddle point problem we target a solution of

$$\min_g \max_{\lambda \in \mathbb{R}^{2L+1}} \mathbb{E}_{(X,S)}(\eta(X, S) - g(X, S))^2 + \sum_{\ell=-L}^{L} \lambda_\ell \left(\mathbb{P}\left(g(X, -1) = \ell M/L \mid S = -1\right) - \mathbb{P}\left(g(X, 1) = \ell M/L \mid S = 1\right)\right) \enspace .$$

Since $s \in \{-1, 1\}$, then the objective function of this saddle point problem can be rewritten as

$$\sum_{s \in \mathcal{S}} \left(p_s \mathbb{E}_{X|S=s}(\eta(X, s) - g(X, s))^2 - s\sum_{\ell=-L}^{L} \lambda_\ell \mathbb{P}\left(g(X, s) = \ell M/L \mid S = s\right)\right) \enspace ,$$

where $p_s = \mathbb{P}(S = s)$. Moreover, since $\sum_{\ell=-L}^{L} \mathbf{1}_{\{g(X,s)=\ell M/L\}} \equiv 1$ we can rewrite the original saddle point problem as

$$\min_g \max_{\lambda \in \mathbb{R}^{2L+1}} \sum_{s \in \mathcal{S}} \mathbb{E}_{X|S=s}\left[\sum_{\ell=-L}^{L} \left(p_s(\eta(X, s) - \ell M/L)^2 - s\lambda_\ell\right) \mathbf{1}_{\{g(X,s)=\ell M/L\}}\right] \enspace .$$

Let us first solve the dual $\max\min$ problem, that is, we would like to find a solution of

$$\max_{\lambda \in \mathbb{R}^{2L+1}} \min_g \sum_{s \in \mathcal{S}} \mathbb{E}_{X|S=s}\left[\sum_{\ell=-L}^{L} \left(p_s(\eta(X, s) - \ell M/L)^2 - s\lambda_\ell\right) \mathbf{1}_{\{g(X,s)=\ell M/L\}}\right] \enspace .$$

Clearly, for every fixed $\lambda \in \mathbb{R}^{2L+1}$ the solution of minimization problem inside is given by $\tilde{g}_\lambda$ defined point-wise as

$$\tilde{g}_\lambda(x, s) = \arg\min_{\ell} \left\{p_s(\eta(X, s) - \ell M/L)^2 - s\lambda_\ell\right\} M/L \enspace .$$

Therefore, the $\max\min$ problem boils down to

$$\max_{\lambda \in \mathbb{R}^{2L+1}} \sum_{s \in \mathcal{S}} \mathbb{E}_{X|S=s} \left[\min_{\ell} \left\{p_s(\eta(X, s) - \ell M/L)^2 - s\lambda_\ell\right\}\right] \enspace .$$

Which is equivalent to

$$-\min_{\lambda \in \mathbb{R}^{2L+1}} \sum_{s \in \mathcal{S}} \mathbb{E}_{X|S=s} \left[\max_{\ell} \left\{s\lambda_\ell - p_s(\eta(X, s) - \ell M/L)^2\right\}\right] \enspace .$$

As we are only interested in the minimizer of the above problem and not in the value of the minimum, we can write that the above problem is equivalent in this sense to

$$-\min_{\lambda \in \mathbb{R}^{2L+1}} \sum_{s \in \mathcal{S}} \mathbb{E}_{X|S=s} \left[\max_{\ell} \left\{s\lambda_\ell + 2p_s \frac{\ell M}{L} \eta(X, s) - p_s \frac{l^2 M^2}{L^2}\right\}\right] \enspace .$$

The objective function of the above minimization problem is convex and is uniformly lower-bounded. The convexity is obvious. Let us show that it is lower bounded. We have the following sequence

$$\begin{aligned}
&\sum_{s \in \mathcal{S}} \mathbb{E}_{X|S=s} \left[\max_{\ell} \left\{s\lambda_\ell + 2p_s \frac{\ell M}{L} \eta(X, s) - p_s \frac{l^2 M^2}{L^2}\right\}\right] \\
&\qquad \geq \sum_{s \in \mathcal{S}} \max_{\ell} \left\{\mathbb{E}_{X|S=s} \left[s\lambda_\ell + 2p_s \frac{\ell M}{L} \eta(X, s) - p_s \frac{l^2 M^2}{L^2}\right]\right\} \\
&\qquad = \sum_{s \in \mathcal{S}} \max_{\ell} \left\{s\lambda_\ell + 2p_s \frac{\ell M}{L} \mathbb{E}_{X|S=s}[\eta(X, s)] - p_s \frac{l^2 M^2}{L^2}\right\} \\
&\qquad \geq \max_{\ell} \left\{\sum_{s \in \mathcal{S}} \left(s\lambda_\ell + 2p_s \frac{\ell M}{L} \mathbb{E}_{X|S=s}[\eta(X, s)] - p_s \frac{l^2 M^2}{L^2}\right)\right\} \\
&\qquad = \max_{\ell} \left\{2\frac{\ell M}{L} \sum_{s \in \mathcal{S}} p_s \mathbb{E}_{X|S=s}[\eta(X, s)] - \frac{l^2 M^2}{L^2} \sum_{s \in \mathcal{S}} p_s\right\} \\
&\qquad = \max_{\ell} \left\{2\frac{\ell M}{L} \mathbb{E}[Y] - \frac{l^2 M^2}{L^2}\right\} \\
&\qquad = \max_{\ell} \left\{\left(\mathbb{E}[Y] - \frac{\ell M}{L}\right)^2\right\} - \mathbb{E}[Y]^2 \geq 0 \enspace .
\end{aligned}$$

To conclude the proof notice that under Assumption 2.5, the first order optimality condition for the minimization over λ reads for all $\ell \in \{-L, \ldots, L\}$ as

$$\sum_{s \in \mathcal{S}} s\mathbb{P}_{X|S=s}\left(\tilde{g}_{\lambda^*}(X, s) = \frac{\ell M}{L}\right) = 0 \enspace ,$$

where λ^* is a minimizer. Which implies that $\tilde{g}_{\lambda^*}$ is fair and thus is feasible for problem ($\mathcal{P}'_L$). Using this argument, it is easy to see that $\mathcal{R}(\tilde{g}_{\lambda^*}) = \mathcal{R}(g^*)$ which concludes the proof. □

The next proposition shows that the thresholds $\lambda^*_{-L}, \ldots, \lambda^*_L$ can be found in a compact region. Note that the same, line by line, proof can be applied for $\hat{\lambda}_{-L}, \ldots, \hat{\lambda}_L$, which is thus omitted.

Proposition A.2. *The minimization problem in Eq.* (3) *admits a global minimizer $\lambda^*_{-L}, \ldots, \lambda^*_L$ which satisfies*

$$\min_{\ell \in \{-L, \ldots, L\}} \{\lambda^*_\ell\} = 0, \qquad \max_{\ell \in \{-L, \ldots, L\}} \{\lambda^*_\ell\} \leq 4M^2 \enspace .$$

Proof. Before proceeding to the proof of this result let us first introduce some notation. We denote by $H(\lambda_{-L}, \ldots, \lambda_L)$ the objective function of the minimization problem in Eq. (3). That is,

$$H(\lambda_{-L}, \ldots, \lambda_L) = \sum_{s \in \mathcal{S}} \mathbb{E}_{X|S=s}\left[\max_{\ell \in \{-L,\ldots,L\}}\left\{s\lambda_\ell - p_s\left(\eta(X,s) - \frac{\ell M}{L}\right)^2\right\}\right] \ .$$

Fix any minimizing sequence $\lambda^k = (\lambda^k_{-L}, \ldots, \lambda^k_L)^\top$ of $H(\lambda_{-L}, \ldots, \lambda_L)$, then for all $\varepsilon > 0$ there exists $K \in \mathbb{N}$ such that for all $k \geq K$ it holds that

$$H(0, \ldots, 0) + \varepsilon \geq H(\lambda^k_{-L}, \ldots, \lambda^k_L) \ .$$

Furthermore, notice that for any $(\lambda_\ell)_{\ell=-L,\ldots,L}$ and any $c \in \mathbb{R}$ it holds that

$$\begin{aligned}
H(\lambda_{-L}, \ldots, \lambda_L) &= H(\lambda_{-L}, \ldots, \lambda_L) + \sum_{s \in \mathcal{S}} sc \\
&= \sum_{s \in \mathcal{S}} \mathbb{E}_{X|S=s}\left[\max_{\ell \in \{-L,\ldots,L\}}\left\{s(\lambda_\ell + c) - p_s\left(\eta(X,s) - \frac{\ell M}{L}\right)^2\right\}\right] \\
&= H(\lambda_{-L} + c, \ldots, \lambda_L + c) \ ,
\end{aligned}$$

which implies that $(\lambda_\ell)_{\ell=-L,\ldots,L} + c$ achieves the same value of the objective function. Thus, if we introduce $\tilde{\lambda}^k = (\tilde{\lambda}^k_{-L}, \ldots, \tilde{\lambda}^k_L)^\top$ such that for all $\ell \in \{-L, \ldots, L\}$ and all $k \in \mathbb{N}$ it holds that $\tilde{\lambda}^k_{-\ell} = \lambda^k_{-\ell} + c^k$ with $c^k = -\min_{\ell \in \{-L,\ldots,L\}} \lambda^k_\ell$, then $\tilde{\lambda}^k$ is also a minimizing sequence. Hence, by the definition of a minimizing sequence, for all $\varepsilon > 0$ there exists $K \in \mathbb{N}$ such that for all $k \geq K$

$$H(0, \ldots, 0) + \varepsilon \geq H(\lambda^k_{-L}, \ldots, \lambda^k_L) = H(\tilde{\lambda}^k_{-L}, \ldots, \tilde{\lambda}^k_L) \ . \tag{6}$$

Using the definition of H we can upperbound $H(0, \ldots, 0)$ as follows

$$\begin{aligned}
H(0, \ldots, 0) &= -\sum_{s \in \mathcal{S}} p_s \mathbb{E}_{X|S=s}\left[\min_{\ell \in \{-L,\ldots,L\}}\left(\eta(X,s) - \frac{\ell M}{L}\right)^2\right] \\
&= -\mathbb{E}_{(X,S)}\left[\min_{\ell \in \{-L,\ldots,L\}}\left(\eta(X,S) - \frac{\ell M}{L}\right)^2\right] \leq 0 \ .
\end{aligned} \tag{7}$$

Substituting Eq. (7) to Eq. (6) we get for all $k \geq K$ that

$$\varepsilon \geq H(\tilde{\lambda}^k_{-L}, \ldots, \tilde{\lambda}^k_L) \ .$$

Moreover, for any $\lambda_{-L}, \ldots, \lambda_L$

$$\begin{aligned}
H(\lambda_{-L}, \ldots, \lambda_L) &= \sum_{s \in \mathcal{S}} \mathbb{E}_{X|S=s}\left[\max_{\ell \in \{-L,\ldots,L\}}\left\{s\lambda_\ell - p_s\left(\eta(X,s) - \frac{\ell M}{L}\right)^2\right\}\right] \\
&\geq \max\{\lambda_\ell\} - \min\{\lambda_\ell\} - 4M^2 \ ,
\end{aligned} \tag{8}$$

where the inequality is obtained from the fact that for all $s \in \mathcal{S}$ it holds that

$$\max_{\ell \in \{-L,\ldots,L\}}\left\{s\lambda_\ell - p_s\left(\eta(X,s) - \frac{\ell M}{L}\right)^2\right\} \geq \max_{\ell \in \{-L,\ldots,L\}}\{s\lambda_\ell\} - p_s 4M^2 \ .$$

Applying Eq. (8) to $H(\tilde{\lambda}^k_{-L}, \ldots, \tilde{\lambda}^k_L)$ and using the fact that by construction of $\tilde{\lambda}^k_\ell$ we have $\min\left\{\tilde{\lambda}^k_\ell\right\} = 0$ for all $k \in \mathbb{N}$ we can derive

$$\varepsilon \geq \max_\ell\left\{\tilde{\lambda}^k_\ell\right\} - \min_\ell\left\{\tilde{\lambda}^k_\ell\right\} - 4M^2 = \max_\ell\left\{\tilde{\lambda}^k_\ell\right\} - 4M^2 \ .$$

We have shown that for all $k \geq K$ it holds that

$$\max_\ell\left\{\tilde{\lambda}^k_\ell\right\} \leq 4M^2 + \varepsilon, \quad \min_\ell\left\{\tilde{\lambda}^k_\ell\right\} = 0 \ . \tag{9}$$

Thus, for all $k \geq K$ the minimizing sequence $\tilde{\lambda}_\ell^k$ is bounded. Extracting convergent subsequence from $\tilde{\lambda}_\ell^k$, which by the abuse of notation we also denote by $\tilde{\lambda}_\ell^k$, and using the fact that $H : \mathbb{R}^{2L+1} \to \mathbb{R}$ is continuous we conclude that

$$\inf_{\lambda_{-L},\ldots,\lambda_L} H(\lambda_{-L},\ldots,\lambda_L) \overset{(A)}{=} \lim_{k\to\infty} H(\tilde{\lambda}_{-L}^k,\ldots,\tilde{\lambda}_L^k) \overset{(B)}{=} H(\lambda_{-L}^*,\ldots,\lambda_L^*) \ , \tag{10}$$

where $\lambda^* = (\lambda_{-L}^*,\ldots,\lambda_L^*)^\top$ is the limit of the sequence $\tilde{\lambda}_\ell^k$. In the above equalities (A) is due to the definition of a minimizing sequence, while (B) is thanks to the continuity of $H : \mathbb{R}^{2L+1} \to \mathbb{R}$. This implies that $\lambda_{-L}^*,\ldots,\lambda_L^*$ is a global minimizer. Lastly, taking the limit in Eq. (9) we conclude that for all $\varepsilon > 0$

$$\max_\ell \{\lambda_\ell^*\} \leq 4M^2 + \varepsilon, \quad \min_\ell \{\lambda_\ell^*\} = 0 \ ,$$

the proof is concluded by the fact that ε is arbitrary. □

B Preparation for fairness rates

Before establishing the main theoretical results of this work, let us introduce some notation, which compacts the proofs. We strongly suggest the reader to be familiar with this notation as it will greatly simplify the reading flow.

For all $x \in \mathbb{R}^d$, $s \in \mathcal{S}$ and $\ell \in \{-L,\ldots,L\}$ and all $\lambda \in \mathbb{R}$ we define

$$\hat{h}_\ell^s(x,\lambda) := s\lambda - \hat{p}_s(\hat{\eta}(x,s) - \ell M/L)^2 \ .$$

Therefore, using this notation, the proposed procedure $\hat{g}_L$ defined in Eq. (4) can be written as

$$\hat{g}_L(x,s) = \min\left\{ \mathop{\arg\min}_{\ell\in\{-L,\ldots,L\}} \left\{-\hat{h}_\ell^s(x,\hat{\lambda}_\ell)\right\}\right\} \times \frac{M}{L} \ , \tag{11}$$

where $\hat{\lambda}_{-L},\ldots,\hat{\lambda}_L$ is a solutions of Eq. (5) rewritten as

$$\min_{\lambda_{-L},\ldots,\lambda_L} \sum_{s\in\mathcal{S}} \hat{\mathbb{E}}_{X|S=s}\left[\max_{\ell\in\{-L,\ldots,L\}} \left\{\hat{h}_\ell^s(X,\lambda_\ell)\right\}\right] \ . \tag{12}$$

This notation is only going to be used in the section where we derive the fairness guarantees.

In this part we also would like to introduce several standard results from empirical process theory and establish some generic properties of the minimization algorithm for $\hat{\lambda}_{-L},\ldots,\hat{\lambda}_L$ under the continuity Assumption 3.1.

Reminder on VC theory.

Here we remind some standard definitions of VC theory [54, 42] and already classical results from the empirical process theory on VC classes [55, 35].

Definition B.1 (Projection). *Consider a set system $(\mathcal{X},\mathcal{R})$ with element set $\mathcal{X}$ and a set of subsets $\mathcal{R}$. Let $\mathcal{Y} \subset \mathcal{X}$ we define the projection of $\mathcal{R}$ onto $\mathcal{Y}$ as*

$$\mathcal{R}|_\mathcal{Y} := \{\mathcal{Y} \cap R \ : \ R \in \mathcal{R}\} \ .$$

Definition B.2 (Shattering). *Let $(\mathcal{X},\mathcal{R})$ be a set system with element set $\mathcal{X}$ and a set of subsets $\mathcal{R}$. Let $\mathcal{Y} \subset \mathcal{X}$, we say that $\mathcal{R}$ shatters $\mathcal{Y}$ if*

$$|\mathcal{R}|_\mathcal{Y}| = 2^{|\mathcal{Y}|} \ ,$$

where $|\cdot|$ stands for the cardinality when we consider sets.

Definition B.3 (VC-dimension). *Let $(\mathcal{X},\mathcal{R})$ be a set system. The VC-dimension of $\mathcal{R}$, denoted by* $\mathrm{VC}(\mathcal{R})$ *is the size of the largest subset of $\mathcal{X}$ which is shattered by $\mathcal{R}$.*

Definition B.4 (k-Unions of ranges). *Let $(\mathcal{X},\mathcal{R})$ be a set system, for any integer $k \geq 2$, define the k-fold union of $\mathcal{R}$ as the set system induced on $\mathcal{X}$ by the ranges*

$$\mathcal{R}^{k\cup} := \{R_1 \cup \ldots \cup R_k \ : \ R_1,\ldots,R_k \in \mathcal{R}\} \ .$$

Notice that the k-fold union of a range set $\mathcal{R}$ are nested, that is

$$\mathcal{R} \subset \mathcal{R}^{2\cup} \subset \ldots \subset \mathcal{R}^{k\cup} \ ,$$

in particular, for all $K > 0$ it holds that

$$\bigcup_{k=1}^{K} \mathcal{R}^{k\cup} = \mathcal{R}^{K\cup} \ .$$

The next very simple result gives a bound on the VC-dimension of k-union of a particular range set. General treatment of this type of questions can be found in [8, 25].

Lemma B.5. *Let $k \geq 2$ be a positive integer and $f : \mathbb{R}^d \to \mathbb{R}$ be a fixed function. Consider the following set system $(\mathbb{R}^d, \mathcal{R}^{k\cup})$, where $\mathcal{R}$ is defined as*

$$\mathcal{R} = \left\{R_{w_-,w_+} \,:\, w_-, w_+ \in \mathbb{R}\right\} \ ,$$

with $R_{w_-,w_+} = \left\{x \in \mathbb{R}^d \,:\, w_- > f(x) > w_+\right\}$. Then,

$$\mathrm{VC}\left(\mathcal{R}^{k\cup}\right) \leq 2k \ .$$

Proof. Let $\mathcal{Y} = \{x_1, \ldots, x_{2k}, x_{2k+1}\}$ be any subset of $\mathbb{R}^d$ of cardinality $2k+1$. *W.l.o.g* suppose that

$$f(x_1) \geq \ldots \geq f(x_{2k}) \geq f(x_{2k+1}) \ .$$

Clearly, the set $\{x_1, x_3, x_5, \ldots, x_{2k+1}\}$ cannot be obtained by intersecting $\mathcal{Y}$ with any $R \in \mathcal{R}^{k\cup}$, therefore $\mathcal{Y}$ is not shattered by $\mathcal{R}^{k\cup}$. □

The next result is classical and is typically derived using the entropy integral [23] combined with the Haussler's lemma [30].

Theorem B.6 ([55])**.** *Let $X, X_1, \ldots, X_n$ be* i.i.d. *random variables distributed according to $\mathbb{P}$ on $\mathbb{R}^d$ and $(\mathbb{R}^d, \mathcal{R})$ be a range system of VC-dimension V, then there exists a universal constant $C > 0$ such that*

$$\mathbb{E} \sup_{R \in \mathcal{R}} \left|(\mathbb{P} - \hat{\mathbb{P}})\mathbf{1}_{\{X \in R\}}\right| \leq C\sqrt{\frac{V}{n}} \ ,$$

where the expectation is taken w.r.t. *the joint distribution of $X_1, \ldots, X_n$, and $\hat{\mathbb{P}}$ is the empirical distribution on $X_1, \ldots, X_n$.*

Some properties of the minimization problem in Equation (5)**.**

Let P be a finite set of points from $\mathbb{R}^d$, we denote by $\mathrm{Co}(P)$ its convex hull. The next lemma gives the first order optimality condition for the minimization problem in Equation (5).

Lemma B.7. *Any solution $\hat{\lambda}_{-L}, \ldots, \hat{\lambda}_L$ of the minimization problem in Equation* (5) *satisfies for each $\ell \in \{-L, \ldots, L\}$*

$$\begin{aligned}0 \in &\sum_{s \in \mathcal{S}} s\hat{\mathbb{P}}_{X|S=s}\left(\forall j \neq \ell \ \hat{h}_\ell^s(X, \hat{\lambda}_\ell) > \hat{h}_j^s(X, \hat{\lambda}_j)\right)\\ &+ \sum_{s \in \mathcal{S}} \mathrm{Co}\left(\{0, s\}\right)\hat{\mathbb{P}}_{X|S=s}\left(\forall j \neq \ell \ \hat{h}_\ell^s(X, \hat{\lambda}_\ell) \geq \hat{h}_j^s(X, \hat{\lambda}_j), \ \exists j \neq \ell \ \hat{h}_\ell^{-1}(X, \hat{\lambda}_\ell) = \hat{h}_j^s(X, \hat{\lambda}_j)\right) \ .\end{aligned}$$

Proof. Fix an arbitrary $\ell \in \{-L, \ldots, L\}$. For all $j \in \{-L, \ldots, L\}$, $x \in \mathbb{R}^d$, and $s \in \mathcal{S}$ it holds that

$$\partial_{\lambda_\ell} \hat{h}_j^s(x, \lambda_j) = s\delta_{lj} \ ,$$

where δ_{lj} is the Kronecker symbol. Thus, the subdifferential of $\max_{j \in \{-L,\ldots,L\}}\left\{\hat{h}_j^s(X, \lambda_j)\right\}$ *w.r.t.* λ_ℓ is given by

$$\begin{aligned}\partial_{\lambda_\ell}\left(\max_{j \in \{-L,\ldots,L\}}\left\{\hat{h}_j^s(x, \lambda_j)\right\}\right) =& s\mathbf{1}_{\left\{\forall j \neq \ell \ \hat{h}_\ell^s(x,\lambda_\ell) > \hat{h}_j^s(x,\lambda_j)\right\}}\\ &+ \mathrm{Co}\left(\{0, s\}\right)\mathbf{1}_{\left\{\forall j \neq \ell \ \hat{h}_\ell^s(x,\hat{\lambda}_\ell) \geq \hat{h}_j^s(x,\hat{\lambda}_j), \ \exists j \neq \ell \ \hat{h}_\ell^s(x,\hat{\lambda}_\ell) = \hat{h}_j^s(x,\hat{\lambda}_j)\right\}} \ .\end{aligned}$$

We conclude the proof using the linearity of the empirical expectation and applying the first order optimality condition for convex non-differentiable problems. □

The next Lemma is used to bound the second term on the right hand side of Lemma B.7. The proof of this result heavily relies on Assumption 3.1.

Lemma B.8. *Let Assumption 3.1 be satisfies, then for all $\ell \in \{-L, \dots, L\}$, all $\lambda_{-L}, \dots, \lambda_L \in \mathbb{R}$, and all $s \in \mathcal{S}$ it holds that*

$$\hat{\mathbb{P}}_{X|S=s}\left(\exists j \neq \ell \; \hat{h}^s_\ell(X, \lambda_\ell) = \hat{h}^s_j(X, \lambda_j)\right) \leq \frac{2L}{N_s} \ ,$$

almost surely.

Proof. We provide the proof for $s = 1$ and the proof for $s = -1$ follows the same arguments line by line. Fix an arbitrary $\ell \in \{-L, \dots, L\}$ and $\lambda_{-L}, \dots, \lambda_L \in \mathbb{R}$. If $2L \geq N_1$, then the bound is trivial, thus *w.l.o.g.,* we can assume that $2L + 1 \leq N_1$. Recall that by defintion we have

$$\hat{\mathbb{P}}_{X|S=1}\left(\exists j \neq \ell \; \hat{h}^1_\ell(X, \lambda_\ell) = \hat{h}^1_j(X, \lambda_j)\right) = \frac{1}{N_1} \sum_{X \in \mathcal{D}'_{N_1}} \mathbf{1}_{\left\{\exists j \neq \ell \; \hat{h}^1_\ell(X, \lambda_\ell) = \hat{h}^1_j(X, \lambda_j)\right\}} \ .$$

The proof goes by contradiction. Assume that

$$\frac{1}{N_1} \sum_{X \in \mathcal{D}'_{N_1}} \mathbf{1}_{\left\{\exists j \neq \ell \; \hat{h}^1_\ell(X, \lambda_\ell) = \hat{h}^1_j(X, \lambda_j)\right\}} \geq \frac{2L + 1}{N_1} \ ,$$

with non-zero probability. It implies that in the sum on the left hand side there are at least $2L + 1$ terms, which are exactly equal to one, while in the set $\{-L, \dots, L\} \setminus \{\ell\}$ there are only $2L$ elements. Applying the pingeonhole principle we can conclude that there exists $j \in \{-L, \dots, L\} \setminus \{\ell\}$ and $X, X' \in \mathcal{D}'_{N_1}$ such that simultaneously

$$\begin{aligned}
\hat{h}^1_\ell(X, \lambda_\ell) &= \hat{h}^1_j(X, \lambda_j) \\
\hat{h}^1_\ell(X', \lambda_\ell) &= \hat{h}^1_j(X', \lambda_j) \ .
\end{aligned}$$

Recall that

$$\hat{h}^1_j(x, \lambda) := \lambda - \hat{p}_s(\hat{\eta}(x, 1) - jM/L)^2 \ .$$

Thus, the above two equations become:

$$\begin{aligned}
\lambda_\ell - \hat{p}_1(\hat{\eta}(X, 1) - \ell M/L)^2 &= \lambda_j - \hat{p}_1(\hat{\eta}(X, 1) - jM/L)^2 \\
\lambda_\ell - \hat{p}_1(\hat{\eta}(X', 1) - \ell M/L)^2 &= \lambda_j - \hat{p}_1(\hat{\eta}(X', 1) - jM/L)^2 \ .
\end{aligned}$$

Solving the above equalities for $\hat{\eta}(X, 1)$ and $\hat{\eta}(X', 1)$ implies that

$$\hat{\eta}(X, 1) = \hat{\eta}(X', 1) \ .$$

Since X and X' are sampled from $\mathbb{P}_{X|S=1}$, the above arguments imply that the following bound holds

$$\begin{aligned}
0 &< \mathbb{P}\left(\frac{1}{N_1} \sum_{X \in \mathcal{D}'_{N_1}} \mathbf{1}_{\left\{\exists j \neq \ell \; \hat{h}^1_\ell(X, \lambda_\ell) = \hat{h}^1_j(X, \lambda_j)\right\}} \geq \frac{2L + 1}{N_1}\right) \\
&\leq \mathbb{P}\left(\exists X, X' \in \mathcal{D}'_{N_1} \; \hat{\eta}(X, 1) = \hat{\eta}(X', 1)\right) \ .
\end{aligned}$$

Finally, notice that thanks to the continuity assumption, the random variable $\hat{\eta}(X, 1)$ almost surely does not have any atoms *w.r.t.* the measure $\mathbb{P}_{X|S=1}$, which implies that

$$\mathbb{P}\left(\exists X, X' \in \mathcal{D}'_{N_1} \; \hat{\eta}(X, 1) = \hat{\eta}(X', 1)\right) = 0 \ ,$$

and we arrive to a contradiction. □

B.1 Rates for fairness

We are now in position to prove Theorem 3.2, one of the main theoretical results of this work. Let us recall its statement in a slightly more general form.

Theorem B.9. *Under Assumption 3.1, there exists a universal constant $C > 0$ such that for each Borel set $\mathcal{C} \subset \mathbb{R}$ it holds that*

$$\mathbb{E}_{(\mathcal{D}_n, \mathcal{D}'_N)} \underbrace{\left|\mathbb{P}_{X|S=1}\left(\hat{g}(X,1) \in \mathcal{C}\right) - \mathbb{P}_{X|S=-1}\left(\hat{g}(X,-1) \in \mathcal{C}\right)\right|}_{\mathcal{U}(\hat{g},\mathcal{C})} \leq C \sum_{s \in \mathcal{S}} \left(\sqrt{\frac{|\mathcal{M}|}{p_s N}} + \frac{|\mathcal{M}|\, L}{p_s N}\right) \enspace ,$$

where $\mathcal{M} = \frac{L}{M} \times \left(\{-L, -\frac{(L-1)M}{L}, \ldots, \frac{(L-1)M}{L}, L\} \cap \mathcal{C}\right)$, Moreover, under the same assumptions there exists a universal constant C' such that

$$\mathbb{E}_{(\mathcal{D}_n, \mathcal{D}'_N)} \sup_{\mathcal{C} \subset \mathbb{R}} \underbrace{\left|\mathbb{P}_{X|S=1}\left(\hat{g}(X,1) \in \mathcal{C}\right) - \mathbb{P}_{X|S=-1}\left(\hat{g}(X,-1) \in \mathcal{C}\right)\right|}_{\mathcal{U}(\hat{g},\mathcal{C})} \leq C' \sum_{s \in \mathcal{S}} \left(\sqrt{\frac{L}{p_s N}} + \frac{L^2}{p_s N}\right) \enspace .$$

Proof of Theorem 3.2. Fix some Borel subset $\mathcal{C} \subset \mathbb{R}$. First notice that thanks to the continuity assumption 3.1 it holds for all $s \in \mathcal{S}$ and all $\ell \in \{-L, \ldots, L\}$ that

$$\mathbb{P}_{X|S=s}\left(\hat{g}(X,s) = \frac{\ell M}{L}\right) = \mathbb{P}_{X|S=s}\left(\forall j \neq \ell \; \hat{h}^s_\ell(X, \hat{\lambda}_\ell) > \hat{h}^s_j(X, \hat{\lambda}_j)\right) \enspace ,$$

almost surely. Denote by $\mathcal{M} = \frac{L}{M} \times \left(\{-M, -\frac{(L-1)M}{L}, \ldots, \frac{(L-1)M}{L}, M\} \cap \mathcal{C}\right)$, the scaling of those points in the grid $\mathcal{Q}_L = \{-M, -\frac{(L-1)M}{L}, \ldots, \frac{(L-1)M}{L}, M\}$ which end up in $\mathcal{C}$, thus we can write

$$\begin{aligned}
\mathbb{P}_{X|S=s}\left(\hat{g}(X,s) \in \mathcal{C}\right) &= \mathbb{P}_{X|S=s}\left(\bigcup_{\ell \in \mathcal{M}} \left\{\hat{g}(X,s) = \frac{\ell M}{L}\right\}\right)\\
&= \mathbb{P}_{X|S=s}\left(\bigcup_{\ell \in \mathcal{M}} \left\{\forall j \neq \ell \; \hat{h}^s_\ell(X, \hat{\lambda}_\ell) > \hat{h}^s_j(X, \hat{\lambda}_j)\right\}\right) \enspace .
\end{aligned}$$

Therefore, the unfairness $\mathcal{U}(\hat{g}, \mathcal{C})$ can be written as

$$\mathcal{U}(\hat{g}, \mathcal{C}) = \left|\sum_{s \in \mathcal{S}} s \mathbb{P}_{X|S=s}\left(\bigcup_{\ell \in \mathcal{M}} \left\{\forall j \neq \ell \; \hat{h}^s_\ell(X, \hat{\lambda}_\ell) > \hat{h}^s_j(X, \hat{\lambda}_j)\right\}\right)\right| \enspace ,$$

and first of all we are interested in a bound on $\mathcal{U}(\hat{g}, \mathcal{C})$ which holds almost surely.

Using the first order optimality condition for the problem in Eq. (5), derived in Lemma B.7, we can conclude that for each $\ell \in \{-L. \ldots, L\}$ there exists $\rho^\ell_1 \in [0,1]$ and $\rho^\ell_{-1} \in [-1, 0]$ such that

$$\begin{aligned}
0 &= \sum_{s \in \mathcal{S}} s \hat{\mathbb{P}}_{X|S=s}\left(\forall j \neq \ell \; \hat{h}^s_\ell(X, \hat{\lambda}_\ell) > \hat{h}^s_j(X, \hat{\lambda}_j)\right)\\
&\quad + \sum_{s \in \mathcal{S}} \rho^\ell_s \hat{\mathbb{P}}_{X|S=s}\left(\forall j \neq \ell \; \hat{h}^s_\ell(X, \hat{\lambda}_\ell) \geq \hat{h}^s_j(X, \hat{\lambda}_j), \; \exists j \neq \ell \; \hat{h}^s_\ell(X, \hat{\lambda}_\ell) = \hat{h}^s_j(X, \hat{\lambda}_j)\right) \enspace .
\end{aligned}$$

Note that for each $\ell \in \{-L, \ldots, L\}$ the events $\{\forall j \neq \ell \; \hat{h}^s_\ell(X, \hat{\lambda}_\ell) > \hat{h}^s_j(X, \hat{\lambda}_j)\}$ are disjoint. Therefore, summing the above equality over $\ell \in \mathcal{M}$ we conclude that

$$\begin{aligned}
0 &= \sum_{s \in \mathcal{S}} s \hat{\mathbb{P}}_{X|S=s}\left(\bigcup_{\ell \in \mathcal{M}} \left\{\forall j \neq \ell \; \hat{h}^s_\ell(X, \hat{\lambda}_\ell) > \hat{h}^s_j(X, \hat{\lambda}_j)\right\}\right)\\
&\quad + \sum_{\ell \in \mathcal{M}} \sum_{s \in \mathcal{S}} \rho^\ell_s \hat{\mathbb{P}}_{X|S=s}\left(\forall j \neq \ell \; \hat{h}^s_\ell(X, \hat{\lambda}_\ell) \geq \hat{h}^s_j(X, \hat{\lambda}_j), \; \exists j \neq \ell \; \hat{h}^s_\ell(X, \hat{\lambda}_\ell) = \hat{h}^s_j(X, \hat{\lambda}_j)\right) \enspace .
\end{aligned}$$

The later implies that $\mathcal{U}(\hat{g}, \mathcal{C})$ can be bounded as

$$\begin{aligned}\mathcal{U}(\hat{g}, \mathcal{C}) \leq & \sum_{s \in \mathcal{S}} \left| \left(\mathbb{P}_{X|S=s} - \hat{\mathbb{P}}_{X|S=s} \right) \mathbf{1}_{\left\{ \bigcup_{\ell \in \mathcal{M}} \left\{ \forall j \neq \ell \ \hat{h}^s_\ell(X, \hat{\lambda}_\ell) > \hat{h}^s_j(X, \hat{\lambda}_j) \right\} \right\}} \right| \\ & + \sum_{\ell \in \mathcal{M}} \sum_{s \in \mathcal{S}} \hat{\mathbb{P}}_{X|S=s} \left(\forall j \neq \ell \ \hat{h}^s_\ell(X, \hat{\lambda}_\ell) \geq \hat{h}^s_j(X, \hat{\lambda}_j), \ \exists j \neq \ell \ \hat{h}^s_\ell(X, \hat{\lambda}_\ell) = \hat{h}^s_j(X, \hat{\lambda}_j) \right) \\ \leq & \sum_{s \in \mathcal{S}} \left| \left(\mathbb{P}_{X|S=s} - \hat{\mathbb{P}}_{X|S=s} \right) \mathbf{1}_{\left\{ \bigcup_{\ell \in \mathcal{M}} \left\{ \forall j \neq \ell \ \hat{h}^s_\ell(X, \hat{\lambda}_\ell) > \hat{h}^s_j(X, \hat{\lambda}_j) \right\} \right\}} \right| \\ & + \sum_{\ell \in \mathcal{M}} \sum_{s \in \mathcal{S}} \hat{\mathbb{P}}_{X|S=s} \left(\exists j \neq \ell \ \hat{h}^s_\ell(X, \hat{\lambda}_\ell) = \hat{h}^s_j(X, \hat{\lambda}_j) \right) \quad .\end{aligned}$$

Lemma B.8 allows to control the second term on the r.h.s. of the above inequality. Thus, applying the result of Lemma B.8 and taking supremum over all $\lambda_{-L}, \ldots, \lambda_L$ in the first term on the r.h.s. we arrive at

$$\begin{aligned}\mathcal{U}(\hat{g}, \mathcal{C}) \leq & \sum_{s \in \mathcal{S}} \sup_{\lambda \in \mathbb{R}^{2L+1}} \left| \left(\mathbb{P}_{X|S=s} - \hat{\mathbb{P}}_{X|S=s} \right) \mathbf{1}_{\left\{ \bigcup_{\ell \in \mathcal{M}} \left\{ \forall j \neq \ell \ \hat{h}^s_\ell(X, \lambda_\ell) > \hat{h}^s_j(X, \lambda_j) \right\} \right\}} \right| \\ & + 2 \left| \mathcal{M} \right| L \left(\frac{1}{N_{-1}} + \frac{1}{N_1} \right) \quad ,\end{aligned}$$

almost surely. Thus, to bound the expected value of $\mathcal{U}(\hat{g}, \mathcal{C})$ it remains to bound the expected deviation of the empirical process above and $\mathbb{E}_{(\mathcal{D}_n, \mathcal{D}'_N)}[1/N_s]$ for all $s \in \mathcal{S}$.

We start by bounding the empirical process. As before, we focus on $s = 1$ and the proof for $s = -1$ is identical. To this end, for a fixed $\ell \in \mathcal{M}$, let us examine the event $\left\{ \forall j \neq \ell \ \hat{h}^1_\ell(X, \lambda_\ell) > \hat{h}^1_j(X, \lambda_j) \right\}$. Using the definition oh $\hat{h}^1_j$ we can write

$$\left\{ \forall j \neq \ell \ \hat{h}^1_\ell(X, \lambda_\ell) > \hat{h}^1_j(X, \lambda_j) \right\} \Leftrightarrow \left\{ \forall j \neq \ell \ \lambda_\ell - \hat{p}_1 (\hat{\eta}(X, 1) - \ell M/L)^2 > \lambda_j - \hat{p}_1 (\hat{\eta}(X, 1) - jM/L)^2 \right\} \quad .$$

Rewriting the condition on the right hand side of the equivalence above we arrive at

$$\left\{ \forall j \neq \ell \ \hat{h}^1_\ell(X, \lambda_\ell) > \hat{h}^1_j(X, \lambda_j) \right\} \Leftrightarrow \left\{ \forall j \neq \ell \ \frac{(\lambda_\ell - \lambda_j) L}{2M\hat{p}_1} - \frac{(\ell^2 - j^2)M}{2L} > \hat{\eta}(X, 1)(j - \ell) \right\} \quad .$$

Denote by $\theta^\ell_j = \theta^\ell_j(\lambda_{-L}, \ldots, \lambda_L) := \frac{(\lambda_\ell - \lambda_j)L}{2M\hat{p}_1} - \frac{(\ell^2 - j^2)M}{2L}$, thus we have

$$\begin{aligned}\left\{ \forall j \neq \ell \ \hat{h}^1_\ell(X, \lambda_\ell) > \hat{h}^1_j(X, \lambda_j) \right\} & \Leftrightarrow \left\{ \forall j \neq \ell \ \theta^\ell_j > \hat{\eta}(X, 1)(j - \ell) \right\} \\ & \Leftrightarrow \left\{ \forall j > \ell \ \frac{\theta^\ell_j}{j - \ell} > \hat{\eta}(X, 1) \right\} \cap \left\{ \forall j < \ell \ \frac{\theta^\ell_j}{j - \ell} < \hat{\eta}(X, 1) \right\} \\ & \Leftrightarrow \left\{ \min_{j > \ell} \frac{\theta^\ell_j}{j - \ell} > \hat{\eta}(X, 1) > \max_{j < \ell} \frac{\theta^\ell_j}{j - \ell} \right\} \quad .\end{aligned}$$

Denoting by $w^\ell_+ = w^\ell_+(\lambda_{-L}, \ldots, \lambda_L) = \min_{j > \ell} \frac{\theta^\ell_j}{j - \ell}$ and by $w^\ell_- = w^\ell_-(\lambda_{-L}, \ldots, \lambda_L) = \max_{j < \ell} \frac{\theta^\ell_j}{j - \ell}$ we get

$$\left\{ \forall j \neq \ell \ \hat{h}^1_\ell(X, \lambda_\ell) > \hat{h}^1_j(X, \lambda_j) \right\} \Leftrightarrow \left\{ w^\ell_+ > \hat{\eta}(X, 1) > w^\ell_- \right\} \quad .$$

Thus, we have

$$\begin{aligned} & \sup_{\lambda} \left| \left(\mathbb{P}_{X|S=1} - \hat{\mathbb{P}}_{X|S=1} \right) \mathbf{1}_{\left\{ \bigcup_{\ell \in \mathcal{M}} \left\{ \forall j \neq \ell \ \hat{h}^1_\ell(X, \lambda_\ell) > \hat{h}^1_j(X, \lambda_j) \right\} \right\}} \right| \\ & \leq \sup_{(w^{-L}_+, w^{-L}_-), \ldots, (w^L_+, w^L_-) \in \mathbb{R}^2} \left| \left(\mathbb{P}_{X|S=1} - \hat{\mathbb{P}}_{X|S=1} \right) \mathbf{1}_{\left\{ \bigcup_{\ell \in \mathcal{M}} \left\{ w^\ell_+ > \hat{\eta}(X, 1) > w^\ell_- \right\} \right\}} \right| \quad .\end{aligned}$$

This implies that for all $\mathcal{C}$ it holds that

$$\mathcal{U}(\hat{g}, \mathcal{C}) \leq \sum_{s \in \mathcal{S}} \left(\sup_{(w_+^{-L}, w_-^{-L}), \ldots, (w_+^{L}, w_-^{L}) \in \mathbb{R}^2} \left| \left(\mathbb{P}_{X|S=1} - \hat{\mathbb{P}}_{X|S=1} \right) \mathbf{1}_{\left\{ \bigcup_{\ell \in \mathcal{M}} \left\{ w_+^{\ell} > \hat{\eta}(X,1) > w_-^{\ell} \right\} \right\}} \right| + 2 \left| \mathcal{M} \right| L N_s^{-1} \right) \enspace .$$

We are ready to prove the **first claim** of the result. Combining Lemma B.5 with Lemma B.6 we conclude that there exists $C > 0$ such that

$$\mathbb{E} \left[\sup_{(w_+^{-L}, w_-^{-L}), \ldots, (w_+^{L}, w_-^{L}) \in \mathbb{R}^2} \left| \left(\mathbb{P}_{X|S=1} - \hat{\mathbb{P}}_{X|S=1} \right) \mathbf{1}_{\left\{ \bigcup_{\ell \in \mathcal{M}} \left\{ w_+^{\ell} > \hat{\eta}(X,1) > w_-^{\ell} \right\} \right\}} \right| \;\middle|\; \mathcal{D}_N^S, \mathcal{D}_n \right] \leq C \sqrt{\frac{2 \left| \mathcal{M} \right|}{N_1}} \enspace .$$

Finally, repeating the same argument for $s = -1$ we obtain for some universal $C > 0$

$$\mathbb{E}_{(\mathcal{D}_n, \mathcal{D}'_N)}[\mathcal{U}(\hat{g}, \mathcal{C})] \leq C \mathbb{E} \left(\sqrt{\frac{2 \left| \mathcal{M} \right|}{N_{-1}}} + \sqrt{\frac{2 \left| \mathcal{M} \right|}{N_1}} \right) + 2 \mathbb{E} \left(\frac{\left| \mathcal{M} \right| L}{N_{-1}} + \frac{\left| \mathcal{M} \right| L}{N_1} \right) \enspace .$$

Note that N_{-1} and N_1 are binomial random variables with parameters (p_{-1}, N) and (p_1, N) respectively. Applying the bound on the moment of binomials random variables we conclude that for some universal $C > 0$ it holds that

$$\mathbb{E}_{(\mathcal{D}_n, \mathcal{D}'_N)}[\mathcal{U}(\hat{g}, \mathcal{C})] \leq C \sum_{s \in \mathcal{S}} \left(\sqrt{\frac{\left| \mathcal{M} \right|}{p_s N}} + \frac{\left| \mathcal{M} \right| L}{p_s N} \right) \enspace .$$

In order to prove the **second claim** of the result, we first notice that following the same argument we can write

$$\sup_{\mathcal{C} \subset \mathbb{R}} \mathcal{U}(\hat{g}, \mathcal{C}) \leq \sum_{s \in \mathcal{S}} \left(\sup_{R \in \mathcal{R}_s} \left| \left(\mathbb{P}_{X|S=s} - \hat{\mathbb{P}}_{X|S=s} \right) \mathbf{1}_{\{X \in R\}} \right| + \frac{4L^2}{N_s} \right) \enspace ,$$

almost surely. Here, for all $s \in \mathcal{S}$ the range set $\mathcal{R}_s$ is defined as

$$\mathcal{R}_s = \bigcup_{\ell=1}^{2L+1} \mathcal{R}_{\hat{\eta}, s}^{\ell \cup} \enspace ,$$

where $\mathcal{R}_{\hat{\eta}, s} = \left\{ R_{a,b}^s \,:\, a, b \in \mathbb{R} \right\}$ and $R_{a,b}^s = \left\{ x \in \mathbb{R}^d \,:\, a > \hat{\eta}(x, s) > b \right\}$. In words, the ranges of $\mathcal{R}_s$ are induced by $2L + 1$-fold union of level sets of $\hat{\eta}(\cdot, s)$, with $\hat{\eta}(\cdot, s)$ being fixed conditionally on the labeled dataset. Note that again thanks to Lemma B.5 and the inclusion of k-fold unions it holds that

$$\mathrm{VC}(\mathcal{R}_s) \leq 2L + 1 \enspace ,$$

for all $s \in \mathcal{S}$. We conclude similarly to the previous case applying Lemma B.6, which formally replaces $\left| \mathcal{M} \right|$ by $2L + 1$. $\square$

C Preparation for risk rates

As in the previous part, we first present some preparation results which allow to establish the consistency of the proposed procedure in terms of the risk measure. We suggest the reader to understand the statements the following lemmas first and immediately proceed to the proof of the risk consistency result. After the proof of the main result, the interested reader could proceed to the proofs of the lemmas of this section.

The next tautology is used to simplify the presentation.

Lemma C.1. *For any g it holds that*

$$\mathcal{R}(g) = \mathbb{E}[Y^2] - \mathbb{E}[\eta^2(X, S)] + \sum_{s \in \mathcal{S}} p_s \mathbb{E}_{X|S=s} (\eta(X, s) - g(X, s))^2 \enspace .$$

Let $r(\cdot)$ be defined as

$$r(g) := \sum_{s\in\mathcal{S}} p_s \mathbb{E}_{X|S=s}(\eta(X,s) - g(X,s))^2 = \mathbb{E}_{(X,S)}(\eta(X,S) - g(X,S))^2 \enspace .$$

Notice that for any g, g' it holds that

$$\mathcal{R}(g) - \mathcal{R}(g') = r(g) - r(g') \enspace ,$$

therefore, from now on we focus on $r(\hat{g}) - r(g^*)$ instead of $\mathcal{R}(\hat{g}) - \mathcal{R}(g^*)$.

The next result provides an alternative expression for the risk of the oracle g^*.

Lemma C.2. *Let the continuity Assumption 2.5 be satisfied, then*

$$r(g^*) = \max_{\lambda\in\mathbb{R}^{2L+1}} \sum_{s\in\mathcal{S}} \mathbb{E}_{X|S=s} \min_{\ell\in\{-L,\ldots,L\}} \left\{-s\lambda_\ell + p_s\left(\eta(X,s) - \frac{\ell M}{L}\right)^2\right\} \enspace .$$

We also need a suitable upper bound on the risk of the proposed procedure $\hat{g}$, which is derived very similarly to Lemma C.2.

Lemma C.3. *The proposed estimator $\hat{g}$ satisfies almost surely*

$$\begin{aligned} r(\hat{g}) \leq &\sum_{s\in\mathcal{S}} \mathbb{E}_{X|S=s} \min_{\ell\in\{-L,\ldots,L\}} \left\{-s\hat{\lambda}_\ell + \hat{p}_s\left(\hat{\eta}(X,s) - \frac{\ell M}{L}\right)^2\right\} \\ &+ \sum_{\ell=-L}^{L} \hat{\lambda}_\ell \sum_{s\in\mathcal{S}} s\mathbb{P}_{X|S=s}\left(\hat{g}(X,s) = \frac{\ell M}{L}\right) + 4M\left\|\eta - \hat{\eta}\right\|_1 + 4M^2\sum_{s\in\mathcal{S}}|p_s - \hat{p}_s| \enspace , \end{aligned}$$

where $\left\|\eta - \hat{\eta}\right\|_1 = \mathbb{E}_{(X,S)}\left|\eta(X,S) - \hat{\eta}(X,S)\right|$.

There are four terms in the expression for $r(\hat{g})$: the first one is the risk of $\hat{g}$ if the practitioner had access to the marginal distribution of (X,S); the second term described the violation of the fairness constraints; the third is coming from the fact that we use $\hat{\eta}$ in place of η; the last term appears due to estimation of the marginal distribution of S. Equipped with the two above results we deduce the following corollary on the excess risk of the proposed procedure.

Corollary C.4. *Under Assumption 2.5 the proposed estimator $\hat{g}$ satisfies almost surely*

$$r(\hat{g}) - r(g^*) \leq 8M\left\|\eta - \hat{\eta}\right\|_1 + 8M^2\sum_{s\in\mathcal{S}}|p_s - \hat{p}_s| + \sum_{\ell=-L}^{L}\hat{\lambda}_\ell\sum_{s\in\mathcal{S}} s\mathbb{P}_{X|S=s}\left(\hat{g}(X,s) = \frac{\ell M}{L}\right) \enspace .$$

Proof. Let us introduce some short-hand notation to save space

$$\begin{aligned} \alpha &= \sum_{s\in\mathcal{S}} \mathbb{E}_{X|S=s} \min_{\ell\in\{-L,\ldots,L\}} \left\{-s\hat{\lambda}_\ell + \hat{p}_s\left(\hat{\eta}(X,s) - \frac{\ell M}{L}\right)^2\right\} \\ \beta &= \max_{\lambda} \sum_{s\in\mathcal{S}} \mathbb{E}_{X|S=s} \min_{\ell\in\{-L,\ldots,L\}} \left\{-s\lambda_\ell + p_s\left(\eta(X,s) - \frac{\ell M}{L}\right)^2\right\} \enspace . \end{aligned}$$

Using the above we can write

$$
\begin{aligned}
\alpha-\beta \leq & \sum_{s\in\mathcal{S}}\mathbb{E}_{X|S=s}\min_{\ell\in\{-L,\ldots,L\}}\left\{-s\hat{\lambda}_\ell+\hat{p}_s\left(\hat{\eta}(X,s)-\frac{\ell M}{L}\right)^2\right\}\\
& -\sum_{s\in\mathcal{S}}\mathbb{E}_{X|S=s}\min_{\ell\in\{-L,\ldots,L\}}\left\{-s\hat{\lambda}_\ell+p_s\left(\eta(X,s)-\frac{\ell M}{L}\right)^2\right\}\\
\leq & \sum_{s\in\mathcal{S}}\mathbb{E}_{X|S=s}\min_{\ell\in\{-L,\ldots,L\}}\left\{-s\hat{\lambda}_\ell+p_s\left(\hat{\eta}(X,s)-\frac{\ell M}{L}\right)^2\right\}+4M^2\sum_{s\in\mathcal{S}}|p_s-\hat{p}_s|\\
& \quad -\sum_{s\in\mathcal{S}}\mathbb{E}_{X|S=s}\min_{\ell\in\{-L,\ldots,L\}}\left\{-s\hat{\lambda}_\ell+p_s\left(\eta(X,s)-\frac{\ell M}{L}\right)^2\right\}\\
\leq & \sum_{s\in\mathcal{S}}p_s\mathbb{E}_{X|S=s}\max_{\ell}\left|\left(\hat{\eta}(X,s)-\frac{\ell M}{L}\right)^2-\left(\eta(X,s)-\frac{\ell M}{L}\right)^2\right|+4M^2\sum_{s\in\mathcal{S}}|p_s-\hat{p}_s|\\
\leq & 4M\sum_{s\in\mathcal{S}}p_s\mathbb{E}_{X|S=s}\left|\hat{\eta}(X,s)-\eta(X,s)\right|+4M^2\sum_{s\in\mathcal{S}}|p_s-\hat{p}_s|\\
= & 4M\left\|\eta-\hat{\eta}\right\|_1+4M^2\sum_{s\in\mathcal{S}}|p_s-\hat{p}_s| \quad .
\end{aligned}
$$

Finally, combining Lemma C.2 with Lemma C.3 implies the statement of the corollary. □

C.1 Rates for the excess risk

We are ready to present the proof of the rates of convergence of the excess risk of the proposed procedure stated in Theorem 3.3. Recall that Lemma 2.4 gives a way to control $\mathcal{R}(g_L^*)-\mathcal{R}(f^*)$. Thus, to control the excess risk of g_L^* it only remains to bound $\mathcal{R}(\hat{g}_L)-\mathcal{R}(g_L^*)$. From now on we again omit the index L. We also recall[6] the statement of Theorem 3.3

Theorem C.5. *Let Assumptions 2.5 and 3.1 be satisfied, then for the proposed estimator $\hat{g}$ there exists a universal constant $C>0$ such that*

$$
\mathbb{E}_{(\mathcal{D}_n,\mathcal{D}'_N)}[\mathcal{R}(\hat{g})]-\mathcal{R}(g^*)\leq 8M\mathbb{E}_{(\mathcal{D}_n,\mathcal{D}'_N)}\left\|\eta-\hat{\eta}\right\|_1+CM^2\sum_{s\in\mathcal{S}}\left(L\sqrt{\frac{1}{p_sN}}+\frac{L^2}{p_sN}\right) \quad .
$$

Proof of Theorem C.5. As already discussed we have

$$
\mathbb{E}_{(\mathcal{D}_n,\mathcal{D}'_N)}[\mathcal{R}(\hat{g})]-\mathcal{R}(g^*)=\mathbb{E}_{(\mathcal{D}_n,\mathcal{D}'_N)}[r(\hat{g})]-r(g^*) \quad . \tag{13}
$$

Thanks to Corollary C.4 we have

$$
\begin{aligned}
\mathbb{E}_{(\mathcal{D}_n,\mathcal{D}'_N)}[r(\hat{g})]-r(g^*)\leq & \sum_{\ell=-L}^{L}\mathbb{E}_{(\mathcal{D}_n,\mathcal{D}'_N)}\left[\hat{\lambda}_\ell\sum_{s\in\mathcal{S}}s\mathbb{P}_{X|S=s}\left(\hat{g}(X,s)=\frac{\ell M}{L}\right)\right]\\
& +8M\mathbb{E}_{(\mathcal{D}_n,\mathcal{D}'_N)}\left\|\eta-\hat{\eta}\right\|_1+8M^2\sum_{s\in\mathcal{S}}\mathbb{E}_{(\mathcal{D}_n,\mathcal{D}'_N)}\left|\hat{p}_s-p_s\right| \quad .
\end{aligned}
$$

Let us bound the first term on the right hand side of the above inequality. Thanks to Proposition A.2 we know that for all $\ell\in\{-L,\ldots,L\}$ it holds that $|\hat{\lambda}_\ell|\leq 4M^2$. Note that Proposition A.2 is proven for λ^*, yet an identical proof yields the same conclusion on $\hat{\lambda}$. Using this we can write introducing the notation

$$
(*)=\sum_{\ell=-L}^{L}\mathbb{E}_{(\mathcal{D}_n,\mathcal{D}'_N)}\left[\hat{\lambda}_\ell\sum_{s\in\mathcal{S}}s\mathbb{P}_{X|S=s}\left(\hat{g}(X,s)=\frac{\ell M}{L}\right)\right] \quad ,
$$

[6]Theorem 3.3 provides a bound on $\mathcal{E}(\hat{g})=\mathcal{R}(\hat{g})-\mathcal{R}(f^*)$, while Theorem C.5 is stated on $\mathcal{R}(\hat{g})-\mathcal{R}(g^*)$. The result of Theorem 3.3 is recovered immediately from Lemma 2.4 and Theorem C.5.

that

$$(*) \leq 4M^2 \sum_{\ell=-L}^{L} \mathbb{E}_{(\mathcal{D}_n, \mathcal{D}'_N)} \left| \sum_{s \in \mathcal{S}} s \mathbb{P}_{X|S=s} \left(\hat{g}(X,s) = \frac{\ell M}{L} \right) \right| \enspace .$$

For each $\ell \in \{-L, \ldots, L\}$ we can use Theorem B.9 with $|\mathcal{M}| = 1$ which implies that for some universal constant $C > 0$ we have

$$(*) \leq CM^2 \sum_{s \in \mathcal{S}} \left(L\sqrt{\frac{1}{p_s N}} + \frac{L^2}{p_s N} \right) \enspace .$$

Finally, we can write for some universal $C > 0$ that

$$\sum_{s \in \mathcal{S}} \mathbb{E}_{(\mathcal{D}_n, \mathcal{D}'_N)} \left| \hat{p}_s - p_s \right| = 2\mathbb{E}_{(\mathcal{D}_n, \mathcal{D}'_N)} \left| p_1 - \hat{p}_1 \right| \leq C\sqrt{\frac{1}{N}} \enspace .$$

Combining all of the above we conclude. □

The proof of Theorem 3.3 ends if we combine Theorem C.5 with Lemma 2.4..

C.2 Proofs of preparation results

Proof of Lemma C.2. We have the following chain of equalities

$$\begin{aligned}
r(g^*) &= \sum_{s \in \mathcal{S}} p_s \mathbb{E}_{X|S=s} (\eta(X,s) - g^*(X,s))^2 \\
&= \sum_{\ell=-L}^{L} \sum_{s \in \mathcal{S}} p_s \mathbb{E}_{X|S=s} \left(\eta(X,s) - \frac{\ell M}{L} \right)^2 \mathbf{1}_{\left\{ g^*(X,s) = \frac{\ell M}{L} \right\}} \\
&= \sum_{\ell=-L}^{L} \sum_{s \in \mathcal{S}} \mathbb{E}_{X|S=s} \left(-s\lambda^*_\ell + p_s \left(\eta(X,s) - \frac{\ell M}{L} \right)^2 \right) \mathbf{1}_{\left\{ g^*(X,s) = \frac{\ell M}{L} \right\}} \\
&\quad + \sum_{\ell=-L}^{L} \lambda^*_\ell \sum_{s \in \mathcal{S}} s \mathbb{P}_{X|S=s} \left(g^*(X,s) = \frac{\ell M}{L} \right) \enspace .
\end{aligned}$$

Since g^* is fair it holds that

$$\sum_{s \in \mathcal{S}} s \mathbb{P}_{X|S=s} \left(g^*(X,s) = \frac{\ell M}{L} \right) = 0 \enspace ,$$

for all $\ell \in \{-L, \ldots, L\}$. Thus we have

$$r(g^*) = \sum_{\ell=-L}^{L} \sum_{s \in \mathcal{S}} \mathbb{E}_{X|S=s} \left(-s\lambda^*_\ell + p_s \left(\eta(X,s) - \frac{\ell M}{L} \right)^2 \right) \mathbf{1}_{\left\{ g^*(X,s) = \frac{\ell M}{L} \right\}} \enspace .$$

Recall that for every $(x,s) \in \mathbb{R}^d \times \mathcal{S}$ the oracle g^* is defined as

$$g^*(x,s) = \arg\min_{\ell} \left\{ -s\lambda^*_\ell + p_s \left(\eta(x,s) - \frac{\ell M}{L} \right)^2 \right\} \times \frac{M}{L} \enspace ,$$

thus for $r(g^*)$ we can write

$$r(g^*) = \sum_{s \in \mathcal{S}} \mathbb{E}_{X|S=s} \min_{\ell \in \{-L, \ldots, L\}} \left\{ -s\lambda^*_\ell + p_s \left(\eta(X,s) - \frac{\ell M}{L} \right)^2 \right\} \enspace .$$

Using the definition of $\lambda^*_{-L}, \ldots, \lambda^*_L$ we have

$$r(g^*) = \max_{\lambda} \sum_{s \in \mathcal{S}} \mathbb{E}_{X|S=s} \min_{\ell \in \{-L, \ldots, L\}} \left\{ -s\lambda_\ell + p_s \left(\eta(X,s) - \frac{\ell M}{L} \right)^2 \right\} \enspace .$$

□

Proof of Lemma C.3. Conditionally on all data we can write

$$r(\hat{g}) = \mathbb{E}(\hat{\eta}(X,S) - \hat{g}(X,S))^2 + \mathbb{E}(\eta(X,S) - \hat{g}(X,S))^2 - \mathbb{E}(\hat{\eta}(X,S) - \hat{g}(X,S))^2 \enspace .$$

Note that the boundness of $Y \in \mathbb{R}$, implies the boundness of $\eta(X,S)$. Thus, we have

$$\mathbb{E}(\eta(X,S) - \hat{g}(X,S))^2 - \mathbb{E}(\hat{\eta}(X,S) - \hat{g}(X,S))^2 \leq 4M \left\|\eta - \hat{\eta}\right\|_1 \enspace .$$

So far we showed that the following bound holds almost surely

$$r(\hat{g}) \leq \mathbb{E}(\hat{\eta}(X,S) - \hat{g}(X,S))^2 + 4M \left\|\eta - \hat{\eta}\right\|_1 \enspace .$$

Now, let us work with $\mathbb{E}(\hat{\eta}(X,S) - \hat{g}(X,S))^2$. We can write

$$\begin{aligned}
\mathbb{E}(\hat{\eta}(X,S) - \hat{g}(X,S))^2 &= \sum_{s \in \mathcal{S}} p_s \mathbb{E}_{X|S=s}(\hat{\eta}(X,s) - \hat{g}(X,s))^2 \\
&= \sum_{s \in \mathcal{S}} \hat{p}_s \mathbb{E}_{X|S=s}(\hat{\eta}(X,s) - \hat{g}(X,s))^2 \\
&\quad + \sum_{s \in \mathcal{S}} (p_s - \hat{p}_s) \mathbb{E}_{X|S=s}(\hat{\eta}(X,s) - \hat{g}(X,s))^2 \\
&\leq \sum_{s \in \mathcal{S}} \hat{p}_s \mathbb{E}_{X|S=s}(\hat{\eta}(X,s) - \hat{g}(X,s))^2 + 4M^2 \sum_{s \in \mathcal{S}} |p_s - \hat{p}_s| \enspace .
\end{aligned}$$

Lastly, for the first term on the right hand side of the above inequality we can write

$$\begin{aligned}
\sum_{s \in \mathcal{S}} \hat{p}_s \mathbb{E}_{X|S=s}(\hat{\eta}(X,s) - \hat{g}(X,s))^2 &= \sum_{\ell=-L}^{L} \sum_{s \in \mathcal{S}} \mathbb{E}_{X|S=s} \hat{p}_s \left(\hat{\eta}(X,s) - \frac{\ell M}{L}\right)^2 \mathbf{1}_{\left\{\hat{g}(X,s) = \frac{\ell M}{L}\right\}} \\
&= \sum_{\ell=-L}^{L} \sum_{s \in \mathcal{S}} \mathbb{E}_{X|S=s} \left(-s\hat{\lambda}_\ell + \hat{p}_s \left(\hat{\eta}(X,s) - \frac{\ell M}{L}\right)^2\right) \mathbf{1}_{\left\{\hat{g}(X,s) = \frac{\ell M}{L}\right\}} \\
&\quad + \sum_{\ell=-L}^{L} \hat{\lambda}_\ell \sum_{s \in \mathcal{S}} s \mathbb{P}_{X|S=s}\left(\hat{g}(X,s) = \frac{\ell M}{L}\right) \enspace .
\end{aligned}$$

Recall that for each $(x,s) \in \mathbb{R}^d \times \mathcal{S}$ the estimator $\hat{g}$ is defined as

$$\hat{g}(x,s) = \min\left\{\arg\min_{\ell} \left\{-s\hat{\lambda}_\ell + \hat{p}_s \left(\hat{\eta}(x,s) - \frac{\ell M}{L}\right)^2\right\}\right\} \times \frac{M}{L} \enspace ,$$

thus we have

$$\begin{aligned}
\sum_{s \in \mathcal{S}} \hat{p}_s \mathbb{E}_{X|S=s}(\hat{\eta}(X,s) - \hat{g}(X,s))^2 &= \sum_{s \in \mathcal{S}} \mathbb{E}_{X|S=s} \min_{\ell \in \{-L,\ldots,L\}} \left\{-s\hat{\lambda}_\ell + \hat{p}_s \left(\hat{\eta}(X,s) - \frac{\ell M}{L}\right)^2\right\} \\
&\quad + \sum_{\ell=-L}^{L} \hat{\lambda}_\ell \sum_{s \in \mathcal{S}} s \mathbb{P}_{X|S=s}\left(\hat{g}(X,s) = \frac{\ell M}{L}\right) \enspace .
\end{aligned}$$

Combining all of the above concludes the proof. □

D Optimization algorithm to approximate the thresholds

The whole section is devoted to the proof of Theorem 3.5. We denote by Δ the probability simplex in $\mathbb{R}^{2L+1}$. As pointed out, in the main body of the paper, we set $\hat{\lambda}_{-L}, \ldots, \hat{\lambda}_L$ to be a solution of Eq. (5). Let us recall that the problem in Eq. (5) is an example of non-smooth convex optimization, and subgradient methods can be used to find a solution numerically. Yet, subgradient methods suffer from instability of the outcome and have slow rates of convergence. To alleviate this issue we leverage the structure of problem (5) and apply the idea of smoothing, developed in the context of optimization [45].

Thus, instead of building an iterative scheme for problem (5) we focus on its proxy-problem defined for all $\beta > 0$ as

$$\min_{\lambda_{-L},\dots,\lambda_L} \sum_{s\in\mathcal{S}} \hat{\mathbb{E}}_{X|S=s} \max_{w\in\Delta} \left\{ \sum_{\ell=-L}^{L} w_\ell \left(s\lambda_\ell - \hat{Z}_\ell(X,s) \right) - \beta \, \mathrm{KL}(w||\pi) \right\} \quad , \tag{$\mathcal{P}^{\beta}_{\hat{\lambda}}$}$$

where $\pi = (1/(2L+1), \dots, 1/(2L+1))^\top \in \mathbb{R}^{2L+1}$ and the KL-divergence is defined as

$$\mathrm{KL}(w||\pi) = \sum_{\ell=-L}^{L} w_\ell \log \frac{w_l}{\pi_l} \quad . \tag{14}$$

Denote by G and G_β the objective functions of the minimization problems in Eq. (5) and in ($\mathcal{P}^{\beta}_{\hat{\lambda}}$) respectively.

Therefore, $\hat{\lambda} = (\hat{\lambda}_{-L}, \dots, \hat{\lambda}_L)^\top$ is defined as

$$\hat{\lambda} \in \operatorname*{arg\,min}_{\lambda\in\mathbb{R}^{2L+1}} G(\lambda) \quad .$$

Also, define $\hat{\lambda}_\beta$ as

$$\hat{\lambda}_\beta \in \operatorname*{arg\,min}_{\lambda\in\mathbb{R}^{2L+1}} G_\beta(\lambda) \quad .$$

The next result tells that G_β is indeed an approximation of G as long as β is sufficiently small.

Lemma D.1. *For all $\lambda \in \mathbb{R}^{2L+1}$ it holds that*

$$G_\beta(\lambda) \le G(\lambda) \le G_\beta(\lambda) + 2\beta \log(2L+1) \quad .$$

Proof of Lemma D.1. For any probability vector w it holds that $0 \le \sum_{\ell=-L}^{L} w_\ell \log \frac{w_l}{\pi_l} \le \log(2L+1)$. Applying this fact concludes the proof. □

We also need to derive an explicit expression for G_β.

Lemma D.2. *For any $\beta > 0$ it holds that*

$$G_\beta(\lambda) = \beta \sum_{s\in\mathcal{S}} \hat{\mathbb{E}}_{X|S=s} \log \left(\sum_{\ell=-L}^{L} \exp\left(\frac{1}{\beta} s\lambda_\ell - \frac{1}{\beta} \hat{Z}_\ell(X,s) \right) \right) - 2\beta \log(2L+1) \quad .$$

Proof of Lemma D.2. For a fixed $s \in \mathcal{S}$ and a fixed $x \in \mathbb{R}^d$, let us first solve another problem, namely we would like to find a maximizer of

$$\max \left\{ \sum_{\ell=-L}^{L} w_\ell \left(s\lambda_\ell - \hat{Z}_\ell(x,s) - \beta \log \frac{w_\ell}{\pi_\ell} \right) \;:\; \sum_{\ell=-L}^{L} w_\ell = 1 \right\} \quad . \tag{15}$$

To solve this problem analytically, we construct the Lagrangian function as

$$\mathcal{L}(w,\kappa) = \sum_{\ell=-L}^{L} w_\ell \left(s\lambda_\ell - \hat{Z}_\ell(x,s) - \beta \log \frac{w_\ell}{\pi_\ell} \right) + \kappa \left(\sum_{\ell=-L}^{L} w_\ell - 1 \right) \quad .$$

The KKT conditions read as

$$\partial_{w_\ell} \mathcal{L}(w,\kappa) = 0 \quad ,$$
$$\sum_{\ell=-L}^{L} w_\ell = 1 \quad ,$$

for all $\ell \in \{-L, \dots, L\}$. Taking the partial derivatives we get

$$\partial_{w_\ell} \mathcal{L}(p, \kappa) = s\lambda_\ell - \hat{Z}_\ell(x, s) - \beta \log \frac{w_\ell}{\pi_\ell} - \beta + \kappa = 0 \ , \tag{16}$$

$$\sum_{\ell=-L}^{L} w_\ell = 1 \ . \tag{17}$$

Solving Eq. (16) for w_l we obtain

$$\begin{aligned}
&-\beta \log \frac{w_\ell}{\pi_\ell} = -s\lambda_\ell + \hat{Z}_\ell(x, s) + \beta - \kappa \ , \\
&\log \frac{w_\ell}{\pi_\ell} = \frac{1}{\beta} s\lambda_\ell - \frac{1}{\beta} \hat{Z}_\ell(x, s) - 1 + \frac{1}{\beta}\kappa \ , \\
&w_\ell = \frac{1}{2L+1} \exp\left(\frac{1}{\beta} s\lambda_\ell - \frac{1}{\beta} \hat{Z}_\ell(x, s)\right) \exp\left(-1 + \frac{1}{\beta}\kappa\right) \ .
\end{aligned}$$

Using the relation in Eq. (17), we find the value of the dual variable κ as

$$\exp\left(-1 + \frac{1}{\beta}\kappa\right) = \left(\frac{1}{2L+1} \sum_{\ell=-L}^{L} \exp\left(\frac{1}{\beta} s\lambda_\ell - \frac{1}{\beta} \hat{Z}_\ell(x, s)\right)\right)^{-1} \tag{18}$$

Plug-in the above into the expression for w_ℓ we arrive at

$$w_\ell = \frac{\exp\left(\frac{1}{\beta} s\lambda_\ell - \frac{1}{\beta} \hat{Z}_\ell(x, s)\right)}{\sum_{\ell=-L}^{L} \exp\left(\frac{1}{\beta} s\lambda_\ell - \frac{1}{\beta} \hat{Z}_\ell(x, s)\right)} \ .$$

Note that $w_\ell \in [0, 1]$ and $\sum_\ell w_\ell = 1$, therefore it is a minimizer of

$$\max_{w \in \Delta} \left\{ \sum_{\ell=-L}^{L} w_\ell \left(s\lambda_\ell - \hat{Z}_\ell(x, s) - \beta \log \frac{w_\ell}{\pi_\ell} \right) \right\} \ .$$

Plug-in the expression for w_ℓ into the above objective function we conclude that

$$\begin{aligned}
\max_{w \in \Delta} \left\{ \sum_{\ell=-L}^{L} w_\ell \left(s\lambda_\ell - \hat{Z}_\ell(x, s) - \beta \log \frac{w_\ell}{\pi_\ell} \right) \right\} =& \beta \log \left(\sum_{\ell=-L}^{L} \exp\left(\frac{1}{\beta} s\lambda_\ell - \frac{1}{\beta} \hat{Z}_\ell(x, s)\right) \right) \\
&- \beta \log(2L+1) \ .
\end{aligned}$$

Thus the minimizer of problem $(\mathcal{P}^\beta_{\hat{\lambda}})$ is also the solution of

$$\min_\lambda \left\{ \beta \sum_{s \in \mathcal{S}} \hat{\mathbb{E}}_{X|S=s} \log \left(\sum_{\ell=-L}^{L} \exp\left(\frac{1}{\beta} s\lambda_\ell - \frac{1}{\beta} \hat{Z}_\ell(X, s)\right) \right) - \beta \log(2L+1) \right\} \ .$$

Therefore,

$$G_\beta(\lambda) = \beta \sum_{s \in \mathcal{S}} \hat{\mathbb{E}}_{X|S=s} \log \left(\sum_{\ell=-L}^{L} \exp\left(\frac{1}{\beta} s\lambda_\ell - \frac{1}{\beta} \hat{Z}_\ell(X, s)\right) \right) - 2\beta \log(2L+1) \ . \tag{19}$$

$\square$

The function G_β is appealing due to the fact that it is smooth and its gradient is Lipschitz.

Lemma D.3 ([27]). *The function G_β has a continuous gradient with Lipschitz constant $2/\beta$, that is, for all λ, λ' it holds that*

$$\left\| \nabla G_\beta(\lambda) - \nabla G_\beta(\lambda') \right\|_2 \le \frac{2}{\beta} \left\| \lambda - \lambda' \right\|_2 \ .$$

Note that small values of β induce large Lipschitz constant and thus this function is harder to minimize.

Let us also derive the gradient of G_β in order to apply iterative procedures.

Lemma D.4. *For every $\lambda \in \mathbb{R}^{2L+1}$, the following expression holds for the gradient of G_β*

$$\Big(\nabla G_\beta(\lambda)\Big)_\ell = \sum_{s \in \mathcal{S}} s \hat{\mathbb{E}}_{X|S=s} \frac{\exp\left(\frac{s}{\beta}\lambda_\ell - \frac{1}{\beta}\hat{Z}_\ell(X,s)\right)}{\sum_{\ell=-L}^{L} \exp\left(\frac{s}{\beta}\lambda_\ell - \frac{1}{\beta}\hat{Z}_\ell(X,s)\right)} \ ,$$

for each $\ell \in \{-L, \dots, L\}$.

Let us recall the accelerated gradient descent for convex $(2/\beta)$ -smooth functions. The goal is to approximate

$$\min G_\beta(\lambda) \ .$$

The iterations of the accelerated gradient descent are given by

$$\begin{aligned}
&\lambda_1 = y_1 = \tau_0 = 0 \ , \\
&y_{t+1} = \lambda_t - \frac{\beta}{2}\nabla G_\beta(\lambda_t) \ , \\
&\lambda_{t+1} = (1-\gamma_t)y_{t+1} + \gamma_t y_t \ , \\
&\tau_t = \frac{1 + \sqrt{1 + 4\tau_{t-1}^2}}{2} \ , \\
&\gamma_t = \frac{1-\tau_t}{\tau_{t+1}} \ .
\end{aligned}$$

The next result is already classical in the optimization literature, its proof can be found in [44, 5].

Theorem D.5 ([44])**.** *The above iteration satisfies*

$$G_\beta(\lambda_T) - G_\beta(\hat{\lambda}_\beta) \le \frac{4\|\lambda_1 - \hat{\lambda}_\beta\|_2^2}{\beta T^2} \ .$$

Combination of Theorem D.5 with Lemma D.1 immediately yields.

Corollary D.6. *Let λ_T be the output Algorithm 1, therefore*

$$G(\lambda_T) - G(\hat{\lambda}) \le \frac{4\|\hat{\lambda}_\beta\|_2^2}{\beta T^2} + 2\beta\log(2L+1) \ .$$

Proof. Thanks to Lemma D.1 we have

$$\begin{aligned}
&G(\lambda_T) \le G_\beta(\lambda_T) + 2\beta\log(2L+1) \ , \\
&G(\hat{\lambda}) \ge G_\beta(\hat{\lambda}) \ge G_\beta(\hat{\lambda}_\beta) \ .
\end{aligned}$$

Moreover, using Theorem D.5 we get

$$G(\lambda_T) - G(\hat{\lambda}) \le \frac{4\|\hat{\lambda}_\beta\|_2^2}{\beta T^2} + 2\beta\log(2L+1) \ .$$

□

Let us understand the order of magnitude of $\|\hat{\lambda}_\beta\|_2^2$.

Lemma D.7. *For any positive β it holds that*

$$\|\hat{\lambda}_\beta\|_\infty \le 4M^2 + 2\beta\log(2L+1) \ .$$

Proof. Notice that

$$G_\beta(0) \leq G(0) \leq 0 \ .$$

Moreover, for any $\lambda \in \mathbb{R}^{2L+1}$ we have

$$\begin{aligned}
G_\beta(\lambda) &= \beta \sum_{s\in\mathcal{S}} \hat{\mathbb{E}}_{X|S=s} \log\left(\sum_{\ell=-L}^{L} \exp\left(\frac{1}{\beta}s\lambda_\ell - \frac{1}{\beta}\hat{Z}_\ell(X,s)\right)\right) - 2\beta\log(2L+1)\\
&\geq G(\lambda) - 2\beta\log(2L+1)\\
&\geq \max\{\lambda_\ell\} - \min\{\lambda_\ell\} - 4M^2 - 2\beta\log(2L+1) \ .
\end{aligned}$$

And we conclude similarly to Proposition A.2. □

Corollary D.8. *For any positive β it holds that*

$$G(\lambda_T) - G(\hat{\lambda}) \leq 128M^4\frac{2L+1}{\beta T^2} + 128\beta\log^2(2L+1) \ .$$

Proof. Recall that for any $\lambda \in \mathbb{R}^{2L+1}$ it holds

$$\|\lambda\|_2^2 \leq \|\lambda\|_\infty^2 (2L+1) \ .$$

Therefore thanks to Lemma D.7, for $\hat{\lambda}_\beta$ we have

$$\begin{aligned}
\left\|\hat{\lambda}_\beta\right\|_2^2 &\leq (2L+1)\left(4M^2 + 2\beta\log(2L+1)\right)^2\\
&\leq 32(2L+1)M^4 + 8(2L+1)\beta^2\log^2(2L+1) \ .
\end{aligned}$$

Substituting this bound into the result of Corollary D.6 we get

$$G(\lambda_T) - G(\hat{\lambda}) \leq 128M^4\frac{2L+1}{\beta T^2} + 32\beta\left(\frac{(2L+1)\log^2(2L+1)}{T^2} + 2\log(2L+1)\right) \ .$$

Finally, notice that for all positive integer $L > 0$ it holds that $\log(2L+1) \leq \log^2(2L+1)$ and if $T \geq \sqrt{2L+1}$ then we have

$$G(\lambda_T) - G(\hat{\lambda}) \leq 128M^4\frac{2L+1}{\beta T^2} + 32\beta\left(\log^2(2L+1) + 2\log^2(2L+1)\right) \ .$$

□

Finally, if we set β as

$$\beta = M^2\frac{\sqrt{2L+1}}{T\log(2L+1)} \ ,$$

the bound reads as

$$G(\lambda_T) - G(\hat{\lambda}) \leq 256M^2\frac{\sqrt{2L+1}\log(2L+1)}{T} \ .$$

Thus, in order to achieve an ε precision, we need to set T as

$$T = \frac{256M^2}{\varepsilon}\sqrt{(2L+1)}\log(2L+1) \ .$$

Our statistical analysis summarized in Theorem 3.3 suggests that $L \sim N^{1/4}$ gives the best convergence rate in terms of the excess risk. Therefore, in order to achieve and ε precision for the desired minimization it is sufficient to satisfy

$$T \sim \frac{N^{1/8}\log(N)}{\varepsilon} \ .$$

In order to match the rate of convergence for the excess risk an fairness, it is desirable to set $\varepsilon \sim N^{-1/4}$. So the final runtime of our algorithm is $O(N^{3/8}\log(N))$ + the time spent on the construction of $\hat{\eta}$.

E Algorithm for predictions without sensitive attribute

In this section we propose a modification of our methodology for the case when the predictions are defined as $f : \mathbb{R}^d \to \mathbb{R}$. That is, the fair optimal predictor $f^* : \mathbb{R}^d \to \mathbb{R}$ is defined as a solution of

$$\min_{f:\mathbb{R}^d\to\mathbb{R}} \left\{\mathbb{E}(Y - f(X))^2 \,:\, \forall \mathcal{C} \subset \mathbb{R}\ \ \mathbb{P}\left(f(X) \in \mathcal{C} \,|\, S = 1\right) = \mathbb{P}\left(f(X) \in \mathcal{C} \,|\, S = -1\right)\right\} \enspace .$$

Remark E.1. *In this part of the supplementary material we use the same notation as in the main body. This section should be seen independently from the main body. For instance, the reader should not confuse f^* defined in the main body of the paper and f^* defined above.*

Similarly to the case with the use of $S \in \mathcal{S}$, we work under the bounded signal Assumption 2.3, that is, $|Y| \leq M$. First we define the binned optimal fair predictor $g^*_L : \mathbb{R}^d \to \mathcal{Q}_L$, where $\mathcal{Q}_L$ is the uniform grid on $[-M.M]$ of $2L + 1$ points defined in the main body. The binned optimal fair predictor $g^*_L : \mathbb{R}^d \to \mathcal{Q}_L$ is a solution of

$$\min_{g:\mathbb{R}^d\to\mathcal{Q}_L} \left\{\mathbb{E}(Y - g(X))^2 \,:\, \forall q \in \mathcal{Q}_L\ \ \mathbb{P}\left(g(X) = q \,|\, S = 1\right) = \mathbb{P}\left(g(X) = q \,|\, S = -1\right)\right\} \enspace .$$

Following the proof of Lemma 2.4 line by line, it is clear that an analogous statement holds in this case. Thus, in order to extend the approach of the main body of this work to the case where the prediction function does not bring into play the sensitive feature, we need to derive the form of g^*_L for all integer $L > 0$.

Let us define[7] $\eta(X) := \mathbb{E}[Y|X]$, $\tau(X) := \mathbb{P}\left(S = 1 \,|\, X\right)$, and $p_s = \mathbb{P}(S = s)$ for all $s \in \mathcal{S}$.

Assumption E.2. *The mappings $t \mapsto \mathbb{P}_X(\eta(X) \geq t)$ and $t \mapsto \mathbb{P}_X(\tau(X) \geq t)$ are continuous.*

Theorem E.3. *For each $L > 0$ under Assumption E.2 it holds for all $x \in \mathbb{R}^d$ that*

$$g^*_L(x) = \operatorname*{arg\,min}_{\ell\in\{-L,\ldots,L\}} \left\{(\eta(x) - \ell M/L)^2 + \lambda^*_\ell\left(\frac{\tau(x)}{p_1} - 1\right)\right\} \times \frac{M}{L} \enspace ,$$

where $\lambda^* = (\lambda^*_{-L}, \ldots, \lambda^*_L)^\top$ is a solution of

$$\min_\lambda \left\{\mathbb{E}_X \max_\ell \left\{\lambda_\ell\left(1 - \frac{\tau(X)}{p_1}\right) - (\eta(X) - \ell M/L)^2\right\}\right\} \enspace .$$

Proof. Fix some integer $L > 0$. Notice that we can write for all $g : \mathbb{R}^d \to \mathcal{Q}_L$

$$\mathbb{E}(Y - g(X))^2 = \mathbb{E}_X(\eta(X) - g(X))^2 + \mathbb{E}(Y^2) - \mathbb{E}(\eta^2(X)) \enspace .$$

Thus, g^*_L can be equivalently defined as a solution of

$$\min_{g:\mathbb{R}^d\to\mathcal{Q}_L} \left\{\mathbb{E}_X(\eta(X) - g(X))^2 \,:\, \forall q \in \mathcal{Q}_L\ \ \mathbb{P}\left(g(X) = q \,|\, S = 1\right) = \mathbb{P}\left(g(X) = q \,|\, S = -1\right)\right\} \enspace .$$

For an arbitrary $q \in \mathcal{Q}_L$ and $s \in \mathcal{S}$ we can write

$$\mathbb{P}\left(g(X) = q \,|\, S = s\right) = p_s^{-1}\mathbb{P}(g(X) = q, S = s) = p_s^{-1}\mathbb{E}_X[\mathbf{1}_{\{g(X)=q\}}\mathbb{P}\left(S = s \,|\, X\right)] \enspace ,$$

therefore for $(*) = \mathbb{P}\left(g(X) = q \,|\, S = 1\right) - \mathbb{P}\left(g(X) = q \,|\, S = -1\right)$ we can write

$$\begin{aligned}
(*) &= \sum_{s\in\mathcal{S}} s p_s^{-1}\mathbb{E}_X[\mathbf{1}_{\{g(X)=q\}}\mathbb{P}\left(S = s \,|\, X\right)]\\
&= p_1^{-1}\mathbb{E}_X[\mathbf{1}_{\{g(X)=q\}}\mathbb{P}\left(S = 1 \,|\, X\right)] - p_{-1}^{-1}\mathbb{E}_X[\mathbf{1}_{\{g(X)=q\}}\mathbb{P}\left(S = -1 \,|\, X\right)]\\
&= p_1^{-1}\mathbb{E}_X[\mathbf{1}_{\{g(X)=q\}}\tau(X)] - p_{-1}^{-1}\mathbb{E}_X[\mathbf{1}_{\{g(X)=q\}}(1 - \tau(X))]\\
&= \mathbb{E}_X\left[\left(\frac{\tau(X)}{p_1 p_{-1}} - \frac{1}{p_{-1}}\right)\mathbf{1}_{\{g(X)=q\}}\right] \enspace .
\end{aligned}$$

The above implies that

$$(*) = 0 \Leftrightarrow \mathbb{E}_X\left[\left(\frac{\tau(X)}{p_1} - 1\right)\mathbf{1}_{\{g(X)=q\}}\right] = 0 \enspace .$$

Hence, g^*_L is a solution of

$$\min_{g:\mathbb{R}^d\to\mathcal{Q}_L} \left\{\mathbb{E}_X(\eta(X) - g(X))^2 \,:\, \forall q \in \mathcal{Q}_L\ \ \mathbb{E}_X\left[\left(\frac{\tau(X)}{p_1} - 1\right)\mathbf{1}_{\{g(X)=q\}}\right] = 0\right\} \enspace . \tag{20}$$

[7] The reader should not confuse η defined in the main body with η defined in this section.

Remark E.4. *Notice that if X is independent from S, then $\tau(X) \equiv p_1$ and* any *predictor is fair.*

The rest of the proof is similar to the proof of Proposition 2.6. Let us write the problem in Eq. (20) in its unconstrained form. That is, we would like to solve

$$\min_{g:\mathbb{R}^d \to \mathcal{Q}_L} \max_{\lambda} \left\{ \mathbb{E}_X(\eta(X) - g(X))^2 + \sum_{\ell=-L}^{L} \lambda_\ell \mathbb{E}_X \left[\left(\frac{\tau(X)}{p_1} - 1 \right) \mathbf{1}_{\{g(X) = \ell M/L\}} \right] \right\} .$$

The objective function of this minmax problem can be equivalently written as

$$\mathbb{E}_X \sum_{\ell=-L}^{L} \left[(\eta(X) - \ell M/L)^2 + \lambda_\ell \left(\frac{\tau(X)}{p_1} - 1 \right) \right] \mathbf{1}_{\{g(X) = \ell M/L\}} .$$

Now, as before we focus on the dual maxmin formulation of the problem

$$\max_{\lambda} \min_{g:\mathbb{R}^d \to \mathcal{Q}_L} \left\{ \mathbb{E}_X \sum_{\ell=-L}^{L} \left[(\eta(X) - \ell M/L)^2 + \lambda_\ell \left(\frac{\tau(X)}{p_1} - 1 \right) \right] \mathbf{1}_{\{g(X) = \ell M/L\}} \right\} .$$

The inner minimization problem can be solved explicitly and the solution for all $\lambda \in \mathbb{R}^{2L+1}$ is given by $\tilde{g}_\lambda$ defined for all $x \in \mathbb{R}$ as

$$\tilde{g}_\lambda(x) = \arg\min_{\ell \in \{-L, \dots, L\}} \left\{ (\eta(x) - \ell M/L)^2 + \lambda_\ell \left(\frac{\tau(x)}{p_1} - 1 \right) \right\} \times \frac{M}{L} .$$

Substituting the expression for $\tilde{g}_\lambda$ into the objective function of the maxmin formulation we get

$$\max_{\lambda} \left\{ \mathbb{E}_X \min_{\ell} \left\{ (\eta(X) - \ell M/L)^2 + \lambda_\ell \left(\frac{\tau(X)}{p_1} - 1 \right) \right\} \right\} .$$

Let λ^* be any minimizer of the above problem. To finish the proof we show that $\tilde{g}_{\lambda^*}$ is fair. It is done similarly to the proof of Proposition 2.6. That is, we first make use of Assumption E.2 to conclude that the objective function in the maximization problem for λ^* is almost surely smooth. Then, we write the first order optimality condition for smooth concave maximization problem which precisely gives the fairness of $\tilde{g}_{\lambda^*}$. Thus, $g_L^* = \tilde{g}_{\lambda^*}$ and we conclude. □

Remark E.5. *It is straightforward to construct a plug-in method once the form of the optimal predictor is established. Indeed, we only need to solve three problems:*

- *Unconstrained regression on (X, Y), to estimate $\mathbb{E}[Y|X]$.*
- *Unconstrained classification on (X, S) to estimate $\mathbb{P}(S = 1|X)$.*
- *Unconstrained minimization over $\lambda \in \mathbb{R}^{2L+1}$.*

The statistical analysis of this method is left for future research.

F The Impact of Unlabeled Data on the Performance of the Estimator

In this section, we empirically study the behavior of the proposed estimator as a function of unlabeled data sample used for recalibration. For this purpose, since the benchmark datasets considered in this paper are fully labeled, we subsample from the original dataset a smaller labeled sample $\mathcal{D}_n$ and then simulate a scenario in which the unlabeled sample $\mathcal{D}'_N$ varies. Specifically, we choose $n = 1/10$ the size of dataset used to estimate η, and $N \in \{0, 1/10, 2/10, 4/10, 8/10\}$ the size of the dataset considered to recalibrate η as a fair predictor. This data generation procedure is applied to the LAW dataset, since it is the largest dataset. We apply our method using the random forest algorithm, using the same cross-validation scheme as in Section 4. The above pipeline is repeated 30 times and the variance of the results is reported in Table 2. Notice that both MSE and DDP are improving with N, highlighting the importance of the unlabeled data. We believe that the improvement could have been more significant if the unlabeled data were provided initially.

LAW - RF+Ours	MSE	DDP
$\mathcal{D}_n$=1/10	.096±.012	.046±.005
$\mathcal{D}_n$=1/10, $\mathcal{D}'_N$=1/10	.093±.011	.044±.005
$\mathcal{D}_n$=1/10, $\mathcal{D}'_N$=2/10	.092±.010	.041±.005
$\mathcal{D}_n$=1/10, $\mathcal{D}'_N$=4/10	.090±.010	.039±.005
$\mathcal{D}_n$=1/10, $\mathcal{D}'_N$=8/10	.089±.010	.038±.004

Table 2: Impact of the size of the unlabeled dataset on MSE and DDP. The size of the labeled sample $\mathcal{D}_n$ is fixed to $1/10$ of the original dataset size. The unlabeled $\mathcal{D}_N$ is initially empty (meaning that we both estimate η and recalibrate it using the same sample $\mathcal{D}_n$, as in the previous experiments of Table 1), and then it increases from $1/10$ to $8/10$ of the original dataset.