Fair Generalized Linear Models with a Convex Penalty

Hyungrok Do [1] **Preston Putzel** [2] **Axel Martin** [1] **Padhraic Smyth** [2] **Judy Zhong** [1]

Abstract

Despite recent advances in algorithmic fairness, methodologies for achieving fairness with generalized linear models (GLMs) have yet to be explored in general, despite GLMs being widely used in practice. In this paper we introduce two fairness criteria for GLMs based on equalizing expected outcomes or log-likelihoods. We prove that for GLMs both criteria can be achieved via a convex penalty term based solely on the linear components of the GLM, thus permitting efficient optimization. We also derive theoretical properties for the resulting fair GLM estimator. To empirically demonstrate the efficacy of the proposed fair GLM, we compare it with other well-known fair prediction methods on an extensive set of benchmark datasets for binary classification and regression. In addition, we demonstrate that the fair GLM can generate fair predictions for a range of response variables, other than binary and continuous outcomes.

1. Introduction

Though machine learning is increasingly being used to support and perform crucial decision making tasks, recent research has clearly demonstrated that data-driven predictive models can often retain systematic biases that are present in the underlying data and can propagate these inequalities to their predictions. For example, large biases in prediction performance have been detected for machine learning models in areas such as criminal recidivism prediction relative to race (Angwin et al., 2016), ranking of job candidates relative to gender (Lahoti et al., 2018), face recognition relative to both race and gender (Ryu et al., 2018; Buolamwini & Gebru, 2018), and in multiple healthcare applications relative to gender, race, and insurance status (Char et al., 2018; Larrazabal et al., 2020; Seyyed-Kalantari et al., 2020).

To address these issues there has recently been a significant body of work in the machine learning community on algorithmic fairness in the context of predictive modeling, including (i) data preprocessing methods that try to reduce disparities, (ii) in-process approaches which enforce fairness during model training, and (iii) post-process approaches which adjust a model's predictions to achieve fairness after training is completed. However, the majority of this work has focused on classification problems with binary outcome variables, and to a lesser extent on regression. There has been little to no investigation of fairness in contexts that include other types of outcome variables such as multiclass or count outputs.

Generalized linear models (GLMs) provide a natural and systematic approach to handle a variety of different types of response variables Y including real-valued, binary, categorical, ordinal, and count outcomes (Nelder & Wedderburn, 1972; McCullagh & Nelder, 1989; Hilbe, 1994). More specifically, GLMs can be viewed as a generalization of standard linear regression, where the normality assumption on the conditional distribution of Y is relaxed to allow for a range of distributional forms, including binomial, multinomial, and Poisson. While GLMs have a significantly simpler functional form compared to flexible modern machine learning models (such as tree-based models and deep neural networks), they nonetheless are a frequent method of choice for building predictive models across many applications areas such as biology, medicine, social science, engineering, climate analysis, and risk analysis (Lindsey, 2000).

Thus, there is a gap between the practical use of GLMs and the development of fairness-aware methodologies for GLMs in the research literature. This paper addresses the gap by:

- Developing a new framework for GLMs to achieve fair predictions for under-represented groups;
- Providing theoretical performance properties and optimization guarantees for fair GLMs;
- Demonstrating that the proposed fair GLM can improve prediction parities for a variety of outcomes including the less-studied count and multinomial outcomes, and investigating, via a systematic empirical study across 11 datasets, the accuracy-disparity trade-

[1]Department of Population Health, NYU Grossman School of Medicine, New York, NY, USA [2]Department of Computer Science, University of California, Irvine, CA, USA. Correspondence to: Judy Zhong <judy.zhong@nyulangone.org>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

Table 1. In-process fair prediction methods and outcome types they can handle (✓: demonstrated in the paper).

Methods	Outcome Types			
	Binary	Continuous	Multiclass	Count
Fair Constraints (Zafar et al., 2017b)	✓			
Disparate Mistreatment (Zafar et al., 2017a)	✓			
Absolute/Squared Difference (Bechavod & Ligett, 2017)	✓			
Group/Individual Fairness (Berk et al., 2017)	✓	✓		
Independence measured by HSIC (Pérez-Suay et al., 2017)		✓		
Fair ERM (Donini et al., 2018)	✓			
Statistical Parity (Agarwal et al., 2018; 2019)	✓	✓		
Bounded Group Loss (Agarwal et al., 2019)	✓	✓		
General FERM (Oneto et al., 2020)		✓		
Fair GLM (Ours)	✓	✓	✓	✓

offs of our fair GLM compared with existing alternative approaches.

Full code and datasets for our experiments are available on `https://github.com/hyungrok-do/fair-glm-cvx`. The proofs for all of the theoretical results in the paper can be found in Appendix A.

2. Related Work

We focus on approaches that consider *group fairness*, which require models to have similar predictive performances across groups. Among the general class of group fairness methods, we focus on the widely-used *in-process* approach, where fairness criteria are introduced during the training process, typically by adding a fairness constraint or penalty to the formulation of their objective function. We limit our scope in this paper to methods that do not include sensitive features as inputs to the prediction model but instead use them during training as part of a penalty or constraint on disparity.

Below we discuss related work organized by the types of outcome variables that are handled by each approach, as summarized in Table 1.

For tackling fair binary classification tasks, Zafar et al. (2017b) and Zafar et al. (2017a) proposed the fair constraint (FC) and the constraint for avoiding disparate mistreatment (DM), respectively. Furthermore, Donini et al. (2018) formalized fair empirical risk minimization (FERM) as a constrained risk minimization problem and applied it to linear and nonlinear support vector machines (SVMs). Agarwal et al. (2018) proposed a reductions approach which can transform binary classification problems under statistical parity (SP) or under equalized odds constraints into unconstrained cost-sensitive classification problems.

For fair regression tasks, Pérez-Suay et al. (2017) proposed a penalty based on the Hilbert-Schmidt independence criterion (HSIC) to encourage independence between predicted values and sensitive attributes, and applied this approach to both linear and kernel regression.

There have also been attempts to develop frameworks that can be generalized to multiple types of outcomes. Berk et al. (2017) proposed individual fairness (IF) and group fairness (GF) penalties and applied them to binary logistic and linear regression models. Oneto et al. (2020) proposed a generalized FERM (GFERM) framework, extending the FERM idea to regression models. Agarwal et al. (2019) extended their reductions approach for a general class of problems defined by Lipschitz loss functions and applied the approach to regression and binary classification. They considered statistical parity (SP) and bounded group loss (BGL) as fairness criteria.

Fairness for multiclass classification has also seen relatively little investigation despite its potential utility, and in particular in-process frameworks that cover multiclass classification do not appear to have been investigated in prior work. A likely reason is that extending the fairness approaches used for other problems, such as enforcing equalized odds, is not trivial to extend to the multiclass case. Ye & Xie (2020) proposed to use one-versus-rest SVMs with a penalty on misclassification rates, while Putzel & Lee (2022) have investigated several different extensions of the demographic parity and equalized odds criteria in the multiclass setting. Denis et al. (2021) proposed a plug-in estimator that guarantees demographic parity for multiclass classification.

Thus, overall, to the best of our knowledge, there has been no prior work providing a unified framework for fairness

methods with GLMs.

3. Problem Formulation

3.1. Notation and Problem Definition

Throughout the paper, we consider the generalized linear model framework

$$\mathsf{E}[Y|\mathsf{X}] = g^{-1}(\mathsf{X}\beta) = \mu(\theta), \tag{1}$$

where X and Y are predictor and response variables distributed over R^p and $\mathcal{Y}$, respectively, $\beta \in \mathsf{R}^p$ is a regression coefficient vector, θ is a function of $\mathsf{X}\beta$, g is a link function, and $\mu = g^{-1}$ is a mean function. We provide representative examples of GLMs in Table 2. The probability density/mass function of Y given X has the following form, which is parameterized by θ and ϕ as

$$f(y|\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\},$$

where a, b, and c are functions that vary depending on the choice of link functions or distributions. In this paper, we limit our scope to canonical link functions so that $\theta = \mathsf{X}\beta$.

Suppose that we are given K groups, defined as the possible values $\mathcal{A} = \{a_1, \ldots, a_K\}$ of a sensitive attribute A such as race/ethnicity or gender. We denote the predictor and response variables of group k as X^k and Y^k and their probability distributions as $p^k_{\mathbf{X}}$ and p^k_Y, respectively. In turn, $p^k_{\mathbf{X}|Y=y}$ represents the conditional distribution of X^k given that $Y^k = y$.

Our primary goal is to build a prediction model that learns the relationship between X and Y well. A conventional GLM approach is to estimate a parameter vector β_{GLM} that maximizes the expected log-likelihood, or equivalently minimizes expected negative log-likelihood, that is,

$$\beta_{\text{GLM}} = \operatorname{argmin}_\beta \mathsf{E}[\ell(\beta; \mathsf{X}, Y)],$$

where the expectation is with respect to the joint distribution $p_{\mathbf{X}Y}$ of (X, Y). To relieve prediction disparity, a fairness penalty term $D(\beta; \mathsf{X}, Y, A)$ is included to encourage fair prediction performance between groups.

$$\beta_{\text{FGLM}} = \operatorname{argmin}_\beta \mathsf{E}[\ell(\beta; \mathsf{X}, Y)] + \lambda D(\beta; \mathsf{X}, Y, A), \tag{2}$$

where λ is a hyperparameter. As mentioned in the previous section, several versions of fairness penalties have been proposed. In this paper, we investigate a general framework for fairness criteria for GLMs, a framework that is theoretically applicable to all types of outcomes as well as being computationally efficient.

3.2. Fairness Criteria for GLMs

Definition 3.1 (Equalized Expected Outcomes)**.** A GLM, parameterized by β, satisfies the criterion of *equalized expected outcomes*, with respect to a sensitive attribute A and response variable Y, if the GLM's expected outcomes are identical for all possible outcomes for every pair of groups, that is,

$$\mathsf{E}\,\mu(\mathsf{X}^{ky}\beta) = \mathsf{E}\,\mu(\mathsf{X}^{ly}\beta), \tag{3}$$

where $\mathsf{X}^{ky} \sim p^k_{\mathbf{X}|Y=y}$ and $\mathsf{X}^{ly} \sim p^l_{\mathbf{X}|Y=y}$ for all $k, l \in \mathcal{A}$ and $y \in \mathcal{Y}$. Here, $\mathcal{Y}$ is a set of all possible outcomes of Y. In practice, for real-valued or unbounded outcomes, $\mathcal{Y}$ is discretized in the fairness penalty (as in prior work, Donini et al. (2018)), which we will discuss later in Section 5.2.

Note that the set of β satisfying the equalized expected outcome is non-empty as it can be achieved by the trivial solution $\beta = 0$. This criterion is a natural extension of *equalized odds* (Hardt et al., 2016), which requires the predicted values for each group to be the same for each true outcome $y \in \{0, 1\}$ or equivalently, requires that the predicted values and the sensitive attribute be conditionally independent given the true outcomes. However, statistical independence is not equivalent to equalized expected outcomes for types of prediction tasks other than binary classification; indeed, the equalized expected outcomes is a weaker condition than statistical independence. Even though the equalized expected outcomes does not guarantee statistical independence, producing the same expected prediction for the same true outcome is still a meaningful criterion.

Previous work in Donini et al. (2018) considered equalizing loss functions across sensitive attributes for the case of binary classification. This was further extended to the case of regression tasks in Oneto et al. (2020). Here we introduce an extension of this approach to the case of GLMs. For GLMs, negative log-likelihoods are used as loss functions, thus, we consider the conditional expected log-likelihood

$$\mathsf{E}[\ell(\beta; \mathsf{X}, y)] = \int_{\mathbf{x} \in \mathcal{X}} \left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\} dp_{\mathbf{X}|Y=y}(\mathbf{x}),$$

which measures the expected log-likelihood given y.

Definition 3.2 (Equalized Expected Log-likelihoods)**.** A GLM satisfies equalized expected log-likelihoods with respect to a sensitive attribute A and response variable Y if

$$\mathsf{E}[\ell(\beta; \mathsf{X}^{ky}, y)] = \mathsf{E}[\ell(\beta; \mathsf{X}^{ly}, y)], \tag{4}$$

for all $k, l \in \mathcal{A}$ and $y \in \mathcal{Y}$.

The trivial solution $\beta = 0$ also satisfies the criterion of equalized expected log-likelihoods, so Definition 3.2 is always achievable. However, we emphasize that Definition 3.1 and 3.2 are different. Equalizing expected outcomes

Table 2. Common GLM Distributions with Canonical Link Functions

Distribution	Link	Support	$\mu = g^{-1}(X\theta)$	ϕ	μ'	ℓ'
Bernoulli(μ)	Logit	$\{0, 1\}$	$\frac{1}{1+\exp(-X\theta)}$	1	$\mu(1-\mu)$	$y-\mu$
Multinomial(μ_i)	Logit	$\{0, 1\}$	$\frac{\exp(X\theta_i)}{1+\sum_{j\neq i}\exp(X\theta_j)}$	1	$\mu_i(1-\mu_i)$	$y_i-\mu_i$
Normal(μ, σ^2)	Identity	$\mathbb{R}$	$X\theta$	σ^2	1	$\frac{(y-\mu)}{\sigma^2}$
Poisson(μ)	Log	$\{0\} \cup \mathbb{Z}_+$	$\exp(X\theta)$	1	μ	$y-\mu$

need not result in equalizing expected log-likelihoods and vice versa.

Based on Definition 3.1 and 3.2, we define a measure of disparity between groups by summing up the squared differences of the pairwise expected outcomes or log-likelihoods for all possible true outcomes:

$$D_{\text{EO}} = \sum_{k,l \in A} \sum_{y \in Y} \left(\mathbb{E}[\mu(X^{ky}\theta)] - \mathbb{E}[\mu(X^{ly}\theta)]\right)^2,$$
$$D_{\text{ELL}} = \sum_{k,l \in A} \sum_{y \in Y} \left(\mathbb{E}[\ell(\theta; X^{ky}, y)] - \mathbb{E}[\ell(\theta; X^{ly}, y)]\right)^2.$$

Here we assume y is discretized into a finite number of regions causing the summation $\sum_{y \in Y}$ to be finite. Both disparities above can directly be plugged into (2) as a penalty term to get a fair GLM estimator. However, they are not convex in general, depending on the choice of link functions, and thus, it will be hard to attain globally optimal solutions. In the following section, we introduce a new convex penalty, an upper bound on each of D_{EO} and D_{ELL}.

3.3. A Linear Component Fairness Penalty for GLMs

Lemma 3.3. *Let $h : \mathbb{R} \to \mathbb{R}$ be a differentiable function. Then, for any random variables θ^k and θ^l distributed over $\mathbb{R}$, the following inequality holds:*

$$\left(\mathbb{E}[h(\theta^k)] - \mathbb{E}[h(\theta^l)]\right)^2 \leq \mathbb{E}[h'(\theta^m)^2]\mathbb{E}[(\theta^k - \theta^l)^2], \quad (5)$$

where $\theta^m = \alpha\theta^k + (1-\alpha)\theta^l$, for some $\alpha \in [0, 1]$.

The lemma implies that the squared difference of the expected value of the function h is bounded by the second-order moment of the difference of the θs, provided that the expectation of h' is finite. Based on this lemma, we provide two key results of our work.

Proposition 3.4. *Given $y \in Y$, let μ be an inverse link function and θ be the coefficient vector of GLM defined in (1). Let $h = \mu$, $\theta^k = X^{ky}\theta$, and $\theta^l = X^{ly}\theta$. Then,*

$$\left(\mathbb{E}[\mu(X^{ky}\theta)] - \mathbb{E}[\mu(X^{ly}\theta)]\right)^2 \leq \mathbb{E}[\mu'(X^m\theta)^2]\mathbb{E}[(X^{ky}\theta - X^{ly}\theta)^2], \quad (6)$$

where $X^m = \alpha X^{ky} + (1-\alpha)X^{ly}$, for some $\alpha \in [0, 1]$.

The implication of the proposition is that the squared difference of expected GLM outcomes of group k and l is bounded by the second-order moment of the difference of linear components of group k and l. Therefore, we can minimize the left-hand side of (6) by minimizing the second-order moment provided $\mathbb{E}[\mu'(X^m\theta)^2]$ is bounded above.

This result motivates us to use $\mathbb{E}[(X^{ky}\theta - X^{ly}\theta)^2]$ as a penalty to encourage fairness. As shown in Appendix A, $\mathbb{E}[\mu'(X^m\theta)^2]$ is bounded above for some outcome distributions including normal, binomial and multinomial, but not for all distributions. To address this issue, later in this section we introduce results that support the usage of the term as a fairness penalty for broader classes of distributions.

The next proposition states that $\mathbb{E}[(X^{ky}\theta - X^{ly}\theta)^2]$ bounds the difference of expected log-likelihoods as well.

Proposition 3.5. *Given $y \in Y$, let*

$$h(\theta) = \ell(\theta; X, y) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi),$$

where $\theta = X\theta$ and let $\theta^k = X^{ky}\theta$, and $\theta^l = X^{ly}\theta$. Then, we have

$$\left(\mathbb{E}[\ell(\theta; X^{ky}, y)] - \mathbb{E}[\ell(\theta; X^{ly}, y)]\right)^2 \leq \mathbb{E}[\ell'(\theta; X^m, y)^2]\mathbb{E}[(X^{ky}\theta - X^{ly}\theta)^2], \quad (7)$$

where $X^m = \alpha X^{ky} + (1-\alpha)X^{ly}$, for some $\alpha \in [0, 1]$.

As in the previous proposition, for some distributions, $\mathbb{E}[\ell'(\theta; X^m, y)^2]$, is bounded above but not for all. More rigorous theoretical arguments are presented Section 3.4.

The two propositions above motivate the definition of fairness using the linear components:

$$D_{\mathrm{LC}} = \sum_{k;l \in A} \sum_{y \in Y} \mathrm{E}[(\mathsf{X}^{ky} - \mathsf{X}^{ly})^2]. \quad (8)$$

We can rewrite the penalty term by applying the bias-variance trade-off as

$$\mathrm{E}[(\mathsf{X}^{ky} - \mathsf{X}^{ly})^2] = \mathrm{V}[\mathsf{X}^{ky} - \mathsf{X}^{ly}] + (\mathrm{E}[\mathsf{X}^{ky}] - \mathrm{E}[\mathsf{X}^{ly}])^2.$$

That is, our penalty term consists of the variance of the difference of linear components and the expected difference of linear components. The second term, the difference of linear components, has been considered as a fairness penalty previously in Bechavod & Ligett (2017) for binary classification and Berk et al. (2017) for both binary classification and linear regression. Our analysis shows that adding the variance term can effectively bound the difference of the expected log-likelihoods. In addition, computationally, D_{LC} is still convex in β, thus permitting efficient optimization.

3.4. The Fair Generalized Linear Model (F-GLM)

The above results provide the basis for our F-GLM estimator: $\hat{\beta}_{\mathrm{FGLM}}$ is estimated by minimizing the function

$$\mathrm{E}[\ell(\beta; \mathsf{X}, Y)] + \frac{\lambda}{\kappa} \sum_{k;l \in A} \sum_{y \in Y} \mathrm{E}[(\mathsf{X}^{ky} - \mathsf{X}^{ly})^2] \quad (9)$$

where $\lambda \geq 0$ is a tuning parameter that controls the trade-off between fairness and log-likelihood and $\kappa = |Y|K(K-1)/2$ which is the number of all possible combinations.

We now provide two theorems proving that, at the value of $\hat{\beta}_{\mathrm{FGLM}}$ which minimizes (9), the corresponding D_{EO} and D_{ELL} are bounded by D_{LC} multiplied by respective constants C and C' independent of λ. Thus, even if we impose a greater value of λ to get $\hat{\beta}_{\mathrm{FGLM}}$ with smaller D_{LC}, the constant does not change, and thus, the upper bound of D_{EO} gets smaller. That is, we can obtain $\hat{\beta}_{\mathrm{FGLM}}$ with D_{EO} as small as we want, by increasing λ. This applies to D_{ELL} equivalently, provided by Theorem 3.7.

Theorem 3.6. *Given $y \in Y$, let μ be an inverse link function and $\hat{\beta}_{\mathrm{FGLM}}$ be the minimizer of (9), there exists $C > 0$, which is independent of λ, satisfying*

$$(\mathrm{E}[\mu(\mathsf{X}^{ky}\hat{\beta}_{\mathrm{FGLM}})] - \mathrm{E}[\mu(\mathsf{X}^{ly}\hat{\beta}_{\mathrm{FGLM}})])^2 \leq C\,\mathrm{E}[(\mathsf{X}^{ky}\hat{\beta}_{\mathrm{FGLM}} - \mathsf{X}^{ly}\hat{\beta}_{\mathrm{FGLM}})^2]. \quad (10)$$

The proof makes use of the fact that μ' is either bounded or monotonically increasing for the case of canonical GLMs. Full details of the proof can be found in Appendix A.4

Theorem 3.7. *Similar to the previous theorem, we also have $C' > 0$, independent of λ, satisfying*

$$(\mathrm{E}[\ell(\hat{\beta}_{\mathrm{FGLM}}; \mathsf{X}^{ky}, y)] - \mathrm{E}[\ell(\hat{\beta}_{\mathrm{FGLM}}; \mathsf{X}^{ly}, y)])^2 \leq C'\mathrm{E}[(\mathsf{X}^{ky}\hat{\beta}_{\mathrm{FGLM}} - \mathsf{X}^{ly}\hat{\beta}_{\mathrm{FGLM}})^2]. \quad (11)$$

While we do not have theoretical results on the tightness of the bounds, our empirical results later in the paper suggest that optimizing the bounds produces models with useful prediction-fairness trade-offs.

The corresponding empirical objective function for (9), given a training dataset $\{(y_i, \mathsf{x}_i, A_i) \in Y \times \mathbb{R}^p \times A : i = 1, \ldots, n\}$ is

$$\frac{1}{n}\sum_{i=1}^{n} \ell(\beta; \mathsf{x}_i, y_i) + \frac{\lambda}{\kappa} \sum_{k;l \in A} \sum_{y \in Y} \frac{1}{n^{kly}} \sum_{(i;j) \in S^{kly}} (\mathsf{x}_i\beta - \mathsf{x}_j\beta)^2, \quad (12)$$

where

$$\ell(\beta; \mathsf{x}_i, y_i) = \frac{y_i \mathsf{x}_i\beta - b(\mathsf{x}_i\beta)}{a(\phi)} + c(y_i, \phi),$$

$S^{kly} = \{(i,j) : y_i = y_j = y, A_i = k, A_j = l\}$, and $n^{kly} = |S^{kly}|$.

Note that our fairness definitions are formulated conditional on $Y = y$ to equalize the terms (log-likelihood, expected outcomes, or linear expectations) within each group of subjects with outcomes equal to the discretized value y or included within the discretized region. For binary outcomes, this means enforcing parity separately for $y = 0$ and $y = 1$, which has previously been reported to achieve better fairness (Hardt et al., 2016) rather than encouraging demographic parity, which does not condition on y. For continuous or unbounded outcomes, define a discretization mapping that maps Y into a discretized set $\{[\delta_i, \delta_{i+1})\}$, rendering finite numbers of regions. We denote the number of $[\delta_i, \delta_{i+1})$ by $|Y|$.

We emphasize that the discretization is only applied to the fairness penalty term, but not to the log-likelihood term. Additional details about the discretization process can be found in Section 5.2. In principle, the proposed fairness definition of equalized log-likelihoods can also be formulated without conditioning on y to achieve marginal fairness between the sensitive attributes.

We further note that our penalty term is similar to the individual fairness penalty of Berk et al. (2017) defined as the sum of squared difference of linear components x_i and x_j of two individuals sampled from different groups weighted by a distance function $d(y_i, y_j)$. Berk et al. (2017) used

$d(y_i, y_j) = \mathbb{I}(y_i = y_j)$ and $d(y_i, y_j) = \exp(-(y_i - y_j)^2)$ for binary classification and regression tasks, respectively. The choice of the identity function as the distance function seems to yield a formulation which bears similarity to ours. However, two penalty terms have different denominators, and the Individual Fairness penalty is not a finite sample estimation of the expectation (8). In addition, our work is the first to provide theoretical support for learning fair GLMs; we combine these theoretical results with extensive empirical results in Section 6, considerably broadening the scope of Berk et al. (2017).

4. Consistency

Here we present the $\sqrt{n}$-consistency of the F-GLM estimator. The full proofs can be found in Appendix A.

Lemma 4.1. *As* $\min_k n^k = n \to \infty$, *we have,*

$$D \to \frac{1}{\kappa} \sum_{k,l \in A} \sum_{y \in Y} E\,(X^{ky} - X^{ly})^2 =$$

One can easily confirm that if all the pairs X^{ky} and X^{ly} are identically distributed, then D becomes 0.

Theorem 4.2 ($\sqrt{n}$-consistency)**.** *Let* β *be the true GLM coefficient. Assuming the two regularity conditions specified in Zou (2006):*

If $\sqrt{n}\lambda_n \to \lambda_0 \geq 0$ *and* $I(\beta) = \Sigma^{-1}$, *as* $\min_k n^k = n \to \infty$, *then,*

$$\sqrt{n}\,(\hat{\beta}_{\text{FGLM}} - \beta) \xrightarrow{d} -\Sigma(W + \lambda_0 D), \quad (13)$$

as $\min_k n^k = n \to \infty$, *where* $W \sim N(0, \Sigma)$ *and* $I(\beta)$ *is the Fisher information matrix. Thus,*

$$\hat{\beta}_{\text{FGLM}} \to \operatorname{argmin}_\beta \; E[\ell(\beta; X, Y)] + \lambda_0 \beta^T D\beta. \quad (14)$$

The introduction of the penalty term does not alter the asymptotic variance of the standard GLM estimator; however, it does introduce bias. The amount of bias depends on the variance and difference of expectations between the pairs X^{ky} and X^{ly}.

5. Optimization

In this section we outline our approach for optimizing (12). We first introduce the matrix form of the penalty term in (12):

$$\frac{\lambda}{\kappa} \beta^T D^{kly}$$

where

$$D^{kly} = \frac{1}{n^{kly}} \sum_{(i,j) \in S^{kly}} (x_i - x_j)^T (x_i - x_j),$$

with $x_i \in \mathbb{R}^{1 \times p}$. Therefore, we can rewrite the objective function (12) as

$$\frac{1}{n} \sum_{i=1}^{n} \ell(\beta; x_i, y_i) + \lambda \beta^T \underbrace{\frac{1}{\kappa} \sum_{y \in Y} D^{kly}}_{D},$$

since D is positive semi-definite the objective function (12) is convex; thus it can efficiently be solved with first or second-order methods.

5.1. Newton-Raphson Optimization

Below we describe a Newton-Raphson method, a widely used second-order approach, for minimizing the objective function defined in (12). Starting with an initial solution $\beta^{(0)} = 0$, the Newton-Raphson method iteratively improves the current solution with the following update rule:

$$\beta^{(t+1)} = \beta^{(t)} - [\nabla^2 F(\beta^{(t)})]^{-1} \nabla F(\beta^{(t)}), \quad (15)$$

where

$$\nabla F(\beta) = \frac{1}{n} X^T (y - \mu) + \lambda D\beta$$

and

$$\nabla^2 F(\beta) = \frac{1}{n} X^T W X + \lambda D,$$

where $\mu = g^{-1}(X\beta)$ and $W = \operatorname{diag}(\mu)$.

Convergence of the algorithm is guaranteed if line search is applied to determine step sizes, provided (12) is convex in β. It is well known that the Newton-Raphson method is less sensitive to the choice of step sizes/learning rates than first-order methods.

In addition, it has been noted that convergence of first-order methods is in general not guaranteed for maximizing the Poisson log-likelihood (He et al., 2016), due to its non-global Lipschitz continuity. Second-order methods can be more efficient and effective for finding the global optimum solution for Poisson log-likelihood loss functions which are not globally Lipschitz-continuous.

In the results reported in this paper we focus on second-order methods for the reasons above, but we emphasize that other optimization approaches (e.g., first-order stochastic gradient methods) can be used in practice.

5.2. Discretization for Continuous and Unbounded Outcomes

We consider a discretization mapping such that each segment $[\delta_i, \delta_{i+1})$ contains at least one instance from each group. In particular, for continuous outcomes we use an *equal counts* strategy where each segment contains the same number of samples. We achieve smoother approximations

Table 3. Real-world datasets categorized by their outcome types. Sensitive attribute and number of its unique categories (K), sample size (n) and number of the predictor variables (p) are those of the *preprocessed* datasets. Preprocessing details can be found in Appendix D.

Outcome Type	Dataset	Outcome	Sensitive Attribute (K)	n	p
Binary	Adult (Kohavi, 1996)	Income 50K	Gender (2)	45,222	34
	Arrhythmia (Guvenir et al., 1997)	Presence of Arrhythmia	Gender (2)	418	80
	COMPAS (Larson et al., 2016)	Recidivism in 2Y	Race (4)	6,172	11
	Drug Consumption (Fehrman et al., 2017)	Methadone usage (Y vs N)	Race (2)	1,885	25
	German Credit (Dua & Graff, 2017)	Credit (Good vs Bad)	Gender (2)	1,000	46
Continuous	Communities and Crime (Redmond & Baveja, 2002)	Violent Crimes per Capita	Race (3)	1,993	97
	Law School (LSAC) (Wightman, 1998)	GPA	Race (5)	20,715	7
	Parkinsons Telemonitoring (Tsanas et al., 2009)	UPDRS Score	Gender (2)	5,875	25
	Student Performance (Cortez & Silva, 2008)	Final Grade	Gender (2)	649	39
Count	Health & Retirement Survey (HRS) (`https://hrs.isr.umich.edu/about`)	# of dependencies in daily activities	Race (4)	12,774	23
Multiclass	Drug Consumption (Fehrman et al., 2017)	Meth usage: never used vs within 1Y vs over 1Y ago	Race (2)	1,885	25
	Obesity (Palechor & de la Hoz Manotas, 2019)	Obesity Levels (6)	Gender (2)	2,111	23

if we use more segments t. However, because of the constraint that at least one sample from each group has to be included in each segment it is not always possible to increase t as large as we would like. Instead, we start from a large desired value of t and check if the constraint is satisfied, then if not, we continually decrease t until we get a proper mapping. Empirically, we find that the performance is relatively robust as a function of discretization strategy and number of segments. For count outcomes, which are discrete but unbounded, we choose integer maximum and minimum thresholds. Then we set any values greater than the maximum threshold equal to the threshold while keeping the other values the same and vice versa. Additional details are in Appendix B.

5.3. Computational Complexity

Our estimation procedure can be divided into two stages: (i) preparing D and (ii) Newton-Raphson iterations. The complexity of computing D^{kly} is $O\left(n^{kly}p^2\right)$. As a reminder n^{kly} represents the number of pairs of individuals (i, j) with $A_i = k$, $A_j = l$, and $y_i = y_j = y$. Since n^{kly} is a subset of the total number of pairs, it is bounded by n^2. There are K^2jYj total terms D^{kly} that make up the full matrix D. Thus in total computing D takes $O(n^2K^2jYjp^2)$ The per iteration complexity of the Newton-Raphson algorithm is $O(np^2 + p^3)$ for all outcomes besides multiclass. The complexity in the case of multiclass outcomes becomes $O(mnp^2 + mp^3)$ since there are an additional $O(m)$ set of parameters . However, the complexity of computing D remains the same in all cases. The complete derivation can be found in Appendix C.

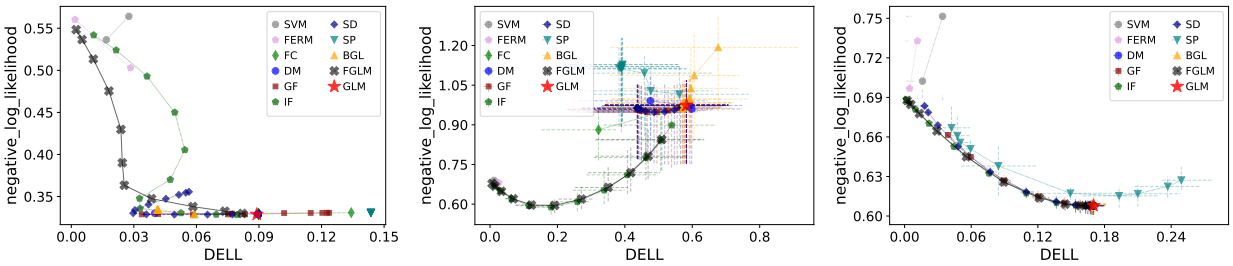
Thus, the largest contribution to the computational complexity of our approach is from pre-computing D. With very large dataset sizes (large n), this bottleneck could be mitigated by using a subsample of the dataset to compute D. For a dataset which also is high-dimensional, the Newton Raphson optimization could also be replaced by faster (per iteration) first order gradient methods. However, in practice, such speedups were not necessary for any of the real-world datasets used in this paper. For the largest dataset in our experiments, with $n = 45,222$ and $p = 14$, fitting an F-GLM took 7 seconds to compute D and an additional 9 seconds until convergence of Newton-Raphson iterations.

6. Experiments and Results

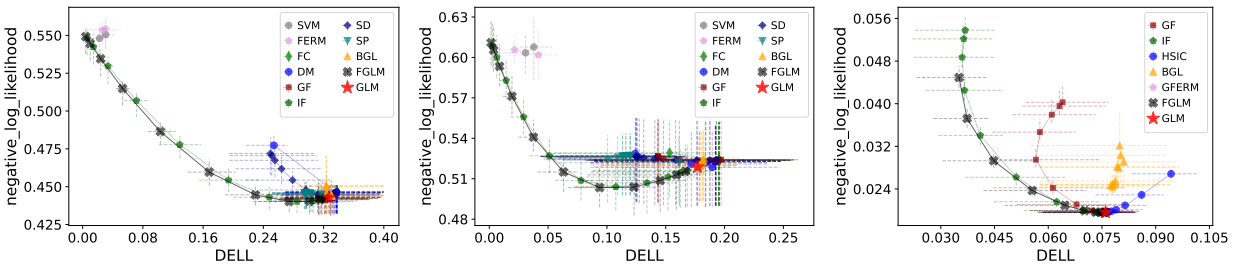
We performed experiments for a comprehensive list of benchmark datasets to evaluate the proposed F-GLM, comparing it with the naive GLM and with multiple *in-process* linear model-based fairness-aware methods.

6.1. Datasets and Fairness-Aware Methods

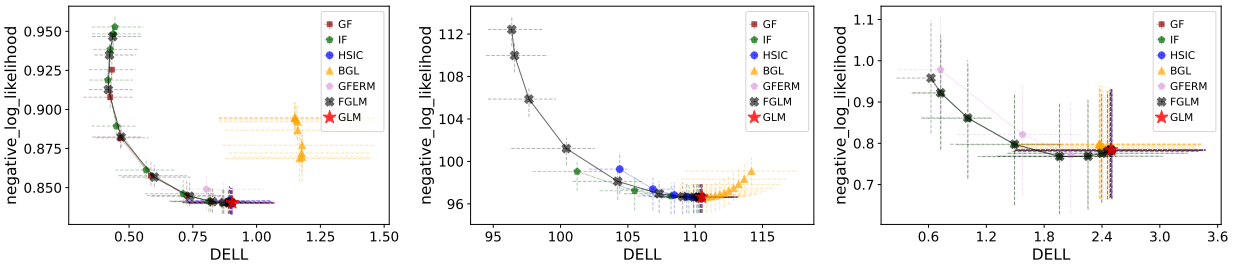
We consider four different tasks/outcome types: binary classification (5 datasets), multiclass classification (2 datasets), continuous outcomes (4 datasets), and count outcomes (1 dataset). General characteristics of the datasets are summarized in Table 3. For the binary classification and regression tasks, we evaluated the naive GLM, the proposed F-GLM, and the methods listed in Table 1. For count and multinomial outcome prediction tasks, we evaluated the naive GLM and the proposed F-GLM. Canonical link functions are used for the GLM and F-GLM: the identity function for normal continuous outcomes, logit function for binary outcomes, log function for count outcomes, and logit functions for multinomial outcomes.



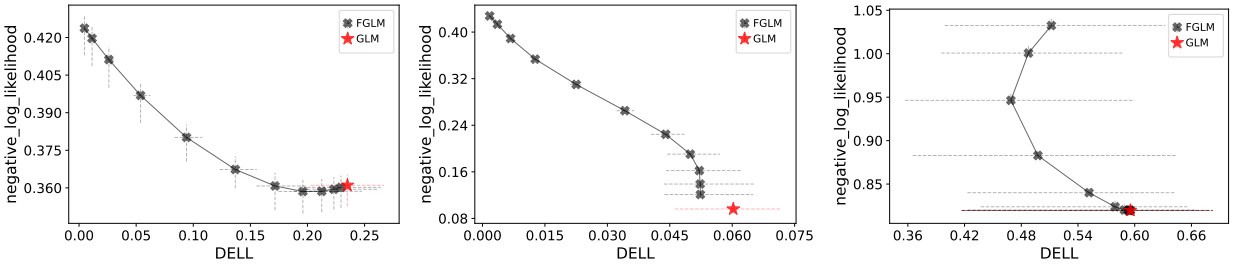
(a) Adult–Binary–Gender(2) (b) Arrhythmia–Binary–Gender(2) (c) COMPAS–Binary–Race(4)



(d) Drug–Binary–Race(2) (e) German–Binary–Race(2) (f) Crime-Cont-Race(3)



(g) LSAC–Cont–Race(5) (h) Parkinsons–Cont–Gender(2) (i) Student–Cont–Gender(2)



(j) Drug–Multi–Race(2) (k) Obesity–Multi–Gender(2) (l) HRS–Count–Race(4)

Figure 1. Experimental results *for negative log-likelihoods* and D_{ELL} for 11 real world datasets, with binary (a-e) and continuous outcomes (f-i). Each subtitle is in the form of Dataset–Outcome Type–Sensitive Attribute(K). For both binary and continuous outcomes we use a Generalized Linear Model (GLM, red star ★), Fair Generalized Linear Model (F-GLM, black X ✖), Individual Fairness penalty (IF, green pentagon ⬟), Group Fairness penalty (GF, dark red square ■), and Bounded Group Loss (BGL, orange triangle ▲). Methods for binary outcomes also include the Support Vector Machine (SVM, grey hexagon ⬢), Fair Constraints (FC, green diamond ◆), Disparate Mistreatment (DM, blue circle ●), Squared Difference Penalizer (SD, dark blue diamond ◆), Fair Empirical Risk Minimization (FERM, plum pentagon ⬟), Statistical Parity (SP, teal triangle ▼). Methods for continuous outcomes include the HSIC penalty (HSIC, blue circle ●), General Fair Empirical Risk Minimization (GFERM, plum pentagon ⬟). See Table 1 for additional information for each method. Each dot represents the mean performance across test sets for a specific hyperparameter value λ and the vertical and horizontal dotted lines reflect variation across test sets (IQRs) for each of performance and disparity.

6.2. Evaluation Metrics

Each method was evaluated in two aspects: (i) overall prediction performance measured by *log-likelihood* and (ii) *disparity of the log-likelihoods* between groups. Both metrics were computed using the test instances. Note that the disparity of the log-likelihoods is an empirical estimate of D_{ELL}. Since there is not a clear consensus on the best choice for a prediction disparity measure in the fairness literature, we also investigated model performances for other disparity measures (including D_{EO}) and included them in Appendix F.

6.3. Experimental Methods

We randomly divided each dataset into training (70%) and testing (30%) sets, except for the Adult dataset which has predefined train/test splits. Each model was trained on the training set by varying its fairness-related hyperparameter (if it exists) over a suitable range. Varying the hyperparameters in this manner produces trajectories of model performance that illustrate the trade-offs in prediction-disparity for each method. Evaluation for a range of such operating points is commonly done in the fairness literature rather than focusing on selecting a single hyperparameter value for each method. The performance and disparity measures were then estimated on the test dataset. For each value of the hyperparameter, the performance and disparity measures were estimated by averaging over 20 replicates of random splits of the training and testing sets (except for the Adult dataset).

6.4. Results

The results for binary and continuous outcomes are displayed in Figure 1, where the x axis represents disparity in the group log-likelihoods and the y axis is the overall log-likelihood, as measured on test sets. For most datasets, the naive GLM is the most unfair solution (largest value of the x axis), which is expected due to the absence of any fairness constraint. Most of the fairness-aware methods we evaluated show wide-ranging trajectories of trade-offs between overall prediction performance and disparity. The trajectories of the competitive methods seem to be broadly consistent with earlier empirical results. The proposed F-GLM is generally one of the best performers relative to competitors for most datasets, in that it can decrease disparity substantially while maintaining overall predictive accuracy.

The results for multiclass and count outcomes are also displayed in Figure 1. For the Drug and Obesity datasets, the negative log-likelihoods of the underrepresented sensitive attribute group were 19% and 43% worse than those of the majority group. For the HRS dataset, the MSEs of the non-Hispanic black and Hispanic groups are 1.75 and 1.61 times of that of the non-Hispanic white subjects, respectively, highlighting the need for fairness-aware prediction algorithms in practice. Overall the proposed F-GLM results in trajectories that flexibly trade-off overall prediction accuracy with disparity.

Results of the disparity and overall prediction performance using other metrics, including log-likelihoods and AUROCs, are presented in the supplementary materials, and showed mostly similar patterns as Figure 1.

7. Conclusions

We presented a fair generalized linear model (F-GLM), incorporating a convex penalty term based solely on the linear components of the GLM, in order to learn fair predictions. We provided statistical justification that the F-GLM achieves fairness both for the expected outcomes and log-likelihoods between groups. Thus, the framework is appealing both theoretically and computationally. Experimental results on benchmark datasets suggest that the F-GLM framework can improve prediction parity while maintaining overall accuracy for binary classification and regression, as well as for less-studied outcomes such as count and multiclass.

One limitation of the F-GLM is its linearity; even though it provides good interpretability, its predictive power can be sub-optimal if the relationship between predictor and response variables is non-linear and complex. We conjecture that the F-GLM approach proposed here could be extended beyond linear models to provide a useful fairness framework for nonlinear machine learning models such as kernel machines or neural networks, by equalizing the linear components in the final decision layer.

An interesting extension of work would be to explore the connection between our fairness penalty and variance regularization, particularly when used for robust learning of predictive models or domain generalization.

Acknowledgements

We thank the ICML reviewers for their suggestions on improving the original version of this paper. This work was supported by in part by the National Institutes of Health under awards NIH R01-LM013344, R01AG054467, R01AG065330 and R01-AG065330-02S1, by the National Science Foundation under award IIS-1900644, by the HPI Research Center in Machine Learning and Data Science at UC Irvine (PP), and by a Qualcomm Faculty award (PS). The Health and Retirement Study public use dataset is produced and distributed by the University of Michigan, Ann Arbor, MI. with funding from the National Institute on Aging (grant number NIA U01AG009740).

References

Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., and Wallach, H. A reductions approach to fair classification. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 60–69. PMLR, 10–15 Jul 2018.

Agarwal, A., Dudik, M., and Wu, Z. S. Fair regression: Quantitative definitions and reduction-based algorithms. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 120–129. PMLR, 09–15 Jun 2019.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine bias. *ProPublica*, 23, May 2016.

Bechavod, Y. and Ligett, K. Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044*, 2017.

Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., Neel, S., and Roth, A. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*, 2017.

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Friedler, S. A. and Wilson, C. (eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81, pp. 77–91, New York, NY, USA, 23–24 Feb 2018. PMLR.

Char, D. S., Shah, N. H., and Magnus, D. Implementing machine learning in health care—addressing ethical challenges. *The New England Journal of Medicine*, 378(11): 981, 2018.

Cortez, P. and Silva, A. M. G. Using data mining to predict secondary school student performance. 2008. URL http://www3.dsi.uminho.pt/pcortez/student.pdf.

Denis, C., Elie, R., Hebiri, M., and Hu, F. Fairness guarantee in multi-class classification. *arXiv preprint arXiv:2109.13642*, 2021.

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J., and Pontil, M. Empirical risk minimization under fairness constraints. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 2796–2806, Red Hook, NY, USA, 2018. Curran Associates Inc.

Dua, D. and Graff, C. UCI Machine Learning Repository, 2017. URL http://archive.ics.uci.edu/ml.

Fehrman, E., Muhammad, A. K., Mirkes, E. M., Egan, V., and Gorban, A. N. The five factor model of personality and evaluation of drug consumption risk. In *Data Science*, pp. 231–242. Springer, 2017. doi: 10.1007/978-3-319-55723-6_18.

Guvenir, H., Acar, B., Demiroz, G., and Cekin, A. A supervised machine learning algorithm for arrhythmia analysis. In *Computers in Cardiology*, pp. 433–436, 1997.

Hardt, M., Price, E., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

He, N., Harchaoui, Z., Wang, Y., and Song, L. Fast and simple optimization for Poisson likelihood models. *arXiv preprint arXiv:1608.01264*, 2016.

Hilbe, J. M. Generalized linear models. *The American Statistician*, 48(3):255–265, 1994.

Kohavi, R. Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pp. 202–207. AAAI Press, 1996.

Lahoti, P., Weikum, G., and Gummadi, K. P. ifair: Learning individually fair data representations for algorithmic decision making. *CoRR*, abs/1806.01059, 2018.

Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H., and Ferrante, E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594, 2020.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. How we analyzed the COMPAS recidivism algorithm. Technical report, ProPublica, May 2016. URL https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

Lindsey, J. K. *Applying Generalized Linear Models*. Springer Science & Business Media, 2000.

McCullagh, P. and Nelder, J. A. *Generalized Linear Models*. Chapman and Hall, 2nd edition, 1989.

Nelder, J. A. and Wedderburn, R. W. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.

Oneto, L., Donini, M., and Pontil, M. General fair empirical risk minimization. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2020.

Palechor, F. M. and de la Hoz Manotas, A. Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from colombia, peru and mexico. *Data in Brief*, 25:104344, 2019.

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*, pp. 339–355. Springer, 2017.

Putzel, P. and Lee, S. Blackbox post-processing for multiclass fairness. *arXiv preprint arXiv:2201.04461*, 2022.

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.

Ryu, H. J., Adam, H., and Mitchell, M. Inclusivefacenet: Improving face attribute detection with race and gender diversity. In *Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)*, 2018.

Seyyed-Kalantari, L., Liu, G., McDermott, M., Chen, I. Y., and Ghassemi, M. Chexclusion: Fairness gaps in deep chest x-ray classifiers. In *BIOCOMPUTING 2021: Proceedings of the Pacific Symposium*, pp. 232–243. World Scientific, 2020.

Tsanas, A., Little, M., McSharry, P., and Ramig, L. Accurate telemonitoring of Parkinson's disease progression by non-invasive speech tests. *Nature Precedings*, pp. 1–1, 2009.

Wightman, L. F. LSAC national longitudinal bar passage study. 1998.

Ye, Q. and Xie, W. Unbiased subdata selection for fair classification: A unified framework and scalable algorithms. *arXiv preprint arXiv:2012.12356*, 2020.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pp. 1171–1180, 2017a.

Zafar, M. B., Valera, I., Rogriguez, M. G., and Gummadi, K. P. Fairness Constraints: Mechanisms for Fair Classification. In Singh, A. and Zhu, J. (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pp. 962–970. PMLR, 20–22 Apr 2017b.

Zou, H. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

A. Proofs

A.1. Lemma 3.3

Since h is differentiable, by the mean value theorem, we have $h(\theta^k) \quad h(\theta^l) = h'(\theta^m)(\theta^k \quad \theta^l)$, where $\theta^m = \alpha\theta^k + (1 \quad \alpha)\theta^l$, $\alpha \; 2 \; [0,1]$. Thus, by applying Cauchy-Schwarz inequality, we obtain

$$\begin{aligned} \mathsf{E}\ h(\theta^k) \quad h(\theta^l)\ ^2 &= \mathsf{E}\ h'(\theta^m)(\theta^k \quad \theta^l)\ ^2 \\ &\quad \mathsf{E}\ h'(\theta^m)^2\ \mathsf{E}\ (\theta^k \quad \theta^l)^2\ . \end{aligned}$$

Note that it is not necessary to assume θ^k and θ^l are independent.

A.2. Proposition 3.4

The inverse of the canonical link functions μ for GLMs are monotone and differentiable (Dobson & Barnett, 2018). Let $\theta^k = \mathsf{X}^{ky}$, $\theta^l = \mathsf{X}^{ly}$, and $h = \mu$. Note that it is not necessary to assume X^{ky} and X^{ly} are independent. Applying Lemma 1 yields

$$\begin{aligned} \mathsf{E}\ \mu(\mathsf{X}^{ky}\ ) \quad \mu(\mathsf{X}^{ly}\ )\ ^2 &= \mathsf{E}\ \mu'(\mathsf{X}^{m}\ )(\mathsf{X}^{ky} \quad \mathsf{X}^{ly}\ )\ ^2 \\ &\quad \mathsf{E}\ \mu'(\mathsf{X}^{m}\ )^2\ \mathsf{E}\ (\mathsf{X}^{ky} \quad \mathsf{X}^{ly}\ )^2\ , \end{aligned}$$

where $\mathsf{X}^m = \alpha\mathsf{X}^{ky} + (1 \quad \alpha)\mathsf{X}^{ly}$, for some $\alpha \; 2 \; [0,1]$.

A.3. Proposition 3.5

Let

$$h(\theta) = \frac{y\theta \quad b(\theta)}{a(\phi)} + c(y,\phi) = \frac{y\mathsf{X} \quad b(\mathsf{X}\ )}{a(\phi)} + c(y,\phi) = \ell(\ ;\mathsf{X},y),$$

here, y is a fixed value and $\theta = \mathsf{X}$. The log-likelihood is a concave and differentiable function of θ. Thus, with $\theta^k = \mathsf{X}^{ky}$ and $\theta^l = \mathsf{X}^{ly}$, we can apply Lemma 1. That is,

$$\begin{aligned} \mathsf{E}\ \ell(\ ;\mathsf{X}^{ky},y) \quad \ell(\ ;\mathsf{X}^{ly},y)\ ^2 &= \mathsf{E}\ \ell'(\ ;\mathsf{X}^{m},y)(\mathsf{X}^{ky} \quad \mathsf{X}^{ly}\ )\ ^2 \\ &\quad \mathsf{E}\ \ell'(\ ;\mathsf{X}^{m},y)^2\ \mathsf{E}\ (\mathsf{X}^{ky} \quad \mathsf{X}^{ly}\ )^2\ , \end{aligned}$$

where $\mathsf{X}^m = \alpha\mathsf{X}^{ky} + (1 \quad \alpha)\mathsf{X}^{ly}$, for some $\alpha \; 2 \; [0,1]$.

A.4. Theorem 3.6

If μ' is bounded (e.g. Bernoulli or Multinomial, see Table 2), we can easily find $C = \sup(\mu')^2$; however, for some link functions (or outcomes) μ' is not bounded and in that case we cannot find C in the same way. Instead, for that case, we use the fact that μ' is *monotonically increasing* and nonnegative to complete the proof. We first introduce a lemma.

Lemma A.1. *Let* b_{FGLM} *be the solution for* (9) *given* $\lambda \quad 0$. *Then,*

$$k\mathsf{b}_{\text{FGLM}}k_2^2 \quad \frac{\delta_{\max}(\ )}{\delta_{\min}(\ )}k\mathsf{b}_{\text{GLM}}k_2^2, \tag{16}$$

where $\delta_{\max}(\ )$ and $\delta_{\min}(\ )$ are the largest and the smallest nonnegative eigenvalues of . We further note that, by the Perron-Frobenius Theorem, $\delta_{\max}$ and $\delta_{\min}$ are the largest and the smallest row sums of .

Proof. The lemma follows from the following chain of inequalities:

$$\delta_{\min}(\ )k\mathsf{b}_{\text{FGLM}}k_2^2 \quad \mathsf{b}_{\text{FGLM}}^T \quad \mathsf{b}_{\text{FGLM}} \quad \mathsf{b}_{\text{GLM}}^T \quad \mathsf{b}_{\text{GLM}} \quad \delta_{\max}(\ )k\mathsf{b}_{\text{GLM}}k_2^2. \tag{17}$$

The left-most inequality follows from the eigenvalue decomposition of , that is,

$$\mathsf{b}_{\text{FGLM}}^T \quad \mathsf{b}_{\text{FGLM}} = \mathsf{b}_{\text{FGLM}}^T Q^{\ 1} \quad Q\mathsf{b}_{\text{FGLM}} \quad \delta_{\min}(\ )\mathsf{b}_{\text{FGLM}}^T Q^{\ 1} Q\mathsf{b}_{\text{FGLM}} = \delta_{\min}(\ )k\mathsf{b}_{\text{FGLM}}k_2^2,$$

where is a diagonal matrix whose entries are the eigenvalues of . Since is positive semi-definite, the entries of are all non-negative. Likewise, we have

$$\mathsf{b}_{\text{GLM}}^T \quad \mathsf{b}_{\text{GLM}} = \mathsf{b}_{\text{GLM}}^T Q^{\ 1} \quad Q\mathsf{b}_{\text{GLM}} \quad \delta_{\max}(\ )\mathsf{b}_{\text{GLM}}^T Q^{\ 1} Q\mathsf{b}_{\text{GLM}} = \delta_{\max}(\ )k\mathsf{b}_{\text{GLM}}k_2^2,$$

which yields the right-most inequality. The inequality in the middle trivially holds because b_{FGLM} and b_{GLM} are optimal solutions for the F-GLM and the naive GLM problems. □

Proof of Theorem 3.6. Here we only consider monotonically increasing μ^0 because otherwise (Bernoulli, multinomial, and normal) μ^0 is bounded and thus we can easily find the quantity that bounds $\mathsf{E}[\mu^0(\mathsf{X}^m\mathsf{b}_{\text{FGLM}})^2]$. We have a chain of inequalities that follows from Lemma A.1 as well as the eigenvalue decomposition:

$$\mathsf{b}_{\text{FGLM}}(\mathsf{X}^{mT}\mathsf{X}^m)\mathsf{b}_{\text{FGLM}} \quad \delta_{\max}(\mathsf{X}^{mT}\mathsf{X}^m)k\mathsf{b}_{\text{FGLM}}k_2^2 \quad \delta_{\max}(\mathsf{X}^{mT}\mathsf{X}^m)(\delta_{\max}(\ )/\delta_{\min}(\ ))k\mathsf{b}_{\text{GLM}}k_2^2.$$

Therefore,

$$j\mathsf{X}^m\mathsf{b}_{\text{FGLM}}j = (\mathsf{b}_{\text{FGLM}}^T(\mathsf{X}^{mT}\mathsf{X}^m)\mathsf{b}_{\text{FGLM}})^{1=2} \quad (\delta_{\max}(\mathsf{X}^{mT}\mathsf{X}^m)(\delta_{\max}(\ )/\delta_{\min}(\ )))^{1=2}k\mathsf{b}_{\text{GLM}}k_2,$$

which yields,

$$\mathsf{E}[\mu^0(\mathsf{X}^m\mathsf{b}_{\text{FGLM}})^2] \quad \mathsf{E}[\mu^0(j\mathsf{X}^m\mathsf{b}_{\text{FGLM}}j)^2] \quad \mathsf{E}[\mu^0((\delta_{\max}(\mathsf{X}^{mT}\mathsf{X}^m)(\delta_{\max}(\ )/\delta_{\min}(\ )))^{1=2}k\mathsf{b}_{\text{GLM}}k_2)^2].$$

This term is independent of λ but depend on the predictors and responses. □

A.5. Theorem 3.7

Proof of Theorem 3.7. In the case of Bernoulli and multinomial, we can take $\sup_x(y \quad \mu(x))^2$ where $0 \quad \mu(x) \quad 1$; otherwise even if μ is unbounded, it is still monotonically increasing. Thus,

$$\begin{aligned}
\mathsf{E}[\ell^0(\mathsf{b}_{\text{FGLM}}; \mathsf{X}^m, y)^2] &= a(\phi)^{\ 2}\mathsf{E}[y^2 \quad 2y\mu(\mathsf{X}^m\mathsf{b}_{\text{FGLM}}) + \mu(\mathsf{X}^m\mathsf{b}_{\text{FGLM}})^2] && (18)\\
&\quad a(\phi)^{\ 2}(y^2 + 2jyj\mathsf{E}[j\mu(\mathsf{X}^m\mathsf{b}_{\text{FGLM}})j] + \mathsf{E}[\mu(\mathsf{X}^m\mathsf{b}_{\text{FGLM}})^2]) && (19)\\
&\quad a(\phi)^{\ 2}(y^2 + 2jyj\mathsf{E}[j\mu(j\mathsf{X}^m\mathsf{b}_{\text{FGLM}}j)j] + \mathsf{E}[\mu(j\mathsf{X}^m\mathsf{b}_{\text{FGLM}}j)^2]) && (20)\\
&\quad a(\phi)^{\ 2}(y^2 + 2jyj\mathsf{E}[j\mu((\delta_{\max}(\mathsf{X}^{mT}\mathsf{X}^m)(\delta_{\max}(\ )/\delta_{\min}(\ )))^{1=2}k\mathsf{b}_{\text{GLM}}k_2)j] && (21)\\
&\quad + \mathsf{E}[\mu(j(\delta_{\max}(\mathsf{X}^{mT}\mathsf{X}^m)(\delta_{\max}(\ )/\delta_{\min}(\ )))^{1=2}k\mathsf{b}_{\text{GLM}}k_2j)^2]) && (22)\\
&= a(\phi)^{\ 2}\mathsf{E}[(jyj + \mu(j(\delta_{\max}(\mathsf{X}^{mT}\mathsf{X}^m)(\delta_{\max}(\ )/\delta_{\min}(\ )))^{1=2}k\mathsf{b}_{\text{GLM}}k_2j)^2] && (23)
\end{aligned}$$

□

Moreover, we have an inequality $\delta_{\max}(\mathsf{X}^{mT}\mathsf{X}^m) \quad \delta_{\max}(\mathsf{X}^{kyT}\mathsf{X}^{ky}) + \delta_{\max}(\mathsf{X}^{lyT}\mathsf{X}^{ly}) + \delta_{\max}(\mathsf{X}^{kyT}\mathsf{X}^{ly} + \mathsf{X}^{lyT}\mathsf{X}^{ky})$ that allows us to remove X^m and α which are unknown.

A.6. Lemma 4.1

For any $k, l \in A$ and $y \in Y$, we have a chain of inequalities

$$\mathsf{D}^{kly} = \frac{1}{n^{kly}} \sum_{(i;j)2S^{kly}} (\mathsf{x}_i \quad \mathsf{x}_j)^T(\mathsf{x}_i \quad \mathsf{x}_j), \tag{24}$$

where $S^{kly} = \{(i,j) : y_i = y_j = y, A_i = k, A_j = l\}$, which is a set of samples drawn from the joint distribution of $(\mathsf{X}^{ky}, \mathsf{X}^{ly})$. Thus, as $n^k, n^l \to \infty$, $\mathsf{D}^{kly} \to \mathsf{E}[(\mathsf{X}^{ky} \quad \mathsf{X}^{ly})^2] = \mathsf{V}[\mathsf{X}^{ky} \quad \mathsf{X}^{ly}] + \mathsf{E}[\mathsf{X}^{ky} \quad \mathsf{X}^{ly}]^2$. Therefore,

$$\mathsf{D} = \frac{2\lambda}{jYjK(K \quad 1)} \sum_{k;l2A} \sum_{y2Y} \mathsf{D}^{kly} \to \frac{2\lambda}{jYjK(K \quad 1)} \sum_{k;l2A} \sum_{y2Y} \mathsf{V}[\mathsf{X}^{ky} \quad \mathsf{X}^{ly}] + \mathsf{E}[\mathsf{X}^{ky} \quad \mathsf{X}^{ly}]^2 \ , \tag{25}$$

as $\min_k n^k \to \infty$.

A.7. Theorem 4.2

For the proof, we assume the two regularity conditions given in Zou (2006):

1. The Fisher information matrix $I(\quad) = \quad^{1} = \mathsf{E}[b''(\mathsf{X}\quad)\mathsf{X}^T\mathsf{X}]$ is finite and positive definite.
2. There is a sufficiently large enough open set U that contains the true $\quad$ such that $\forall \quad \in U$,

$$|b'''(\mathsf{X}\quad)| \quad M(\mathsf{X}) < \infty$$

and

$$\mathsf{E}[M(\mathsf{X})|\mathsf{X}_j\mathsf{X}_k\mathsf{X}_l|] < \infty$$

for all $1 \quad j, k, l \quad p$.

Note that these regularity conditions are considered to be *mild* (Zou, 2006).

Now define

$$V_n(\mathsf{u}) = F \quad + \frac{\mathsf{u}}{\sqrt{n}} \quad F(\quad),$$

where F is the objective function for the F-GLM. Then $V_n(\mathsf{u})$ is minimized at $\mathsf{u} = \sqrt{n}\left(\mathsf{b}_{\text{FGLM}} \quad\right)$. Using the Taylor series expansion, we can rewrite $V_n(\mathsf{u})$ as

$$V_n(\mathsf{u}) = \sum_{i=1}^{n}\left(y_i \quad b'(\mathsf{x}_i \quad)\right)\frac{\mathsf{x}_i\mathsf{u}}{\sqrt{n}} + \sum_{i=1}^{n}\frac{1}{2}b''(\mathsf{x}_i \quad)\mathsf{u}^T\frac{\mathsf{x}_i^T\mathsf{x}_i}{n}\mathsf{u} + n^{-\frac{3}{2}}\sum_{i=1}^{n}\frac{1}{6}b'''(\mathsf{x}_i\tilde{\ }\quad)(\mathsf{x}_i\mathsf{u})^3$$
$$+ \lambda_n \quad + \frac{\mathsf{u}}{\sqrt{n}}\Big)^T\mathsf{D} \quad + \frac{\mathsf{u}}{\sqrt{n}}\Big)^T\mathsf{D} \quad ,$$

where $\tilde{\ }$ is between $\quad$ and $\quad + \frac{\mathsf{u}}{\sqrt{n}}$. Given the regularity conditions the first three terms converges to

$$\mathsf{u}^T\mathsf{W} + \frac{1}{2}\mathsf{u}^T \quad \mathsf{u}$$

in distribution, where $\mathsf{W} \quad N(0, \quad)$. On the other hand, for the last term, we have

$$\lambda_n \quad \frac{2\mathsf{u}^T\mathsf{D}}{\sqrt{n}} + \frac{\mathsf{u}^T\mathsf{D}\mathsf{u}}{n} \quad \stackrel{p}{!} \quad 2\lambda_0\mathsf{u}^T \quad + 0,$$

provided $\lambda_n/\sqrt{n} \; ! \; \lambda_0 \quad 0$ and $\mathsf{D} \; ! \quad$ as $\min_k n^k \; ! \; \infty$. Thus, we have

$$V_n(\mathsf{u}) \stackrel{d}{!} V(\mathsf{u}) = \mathsf{u}^T\mathsf{W} + \frac{1}{2}\mathsf{u}^T \quad \mathsf{u} + 2\lambda_0\mathsf{u}^T \quad .$$

Therefore,

$$\sqrt{n}\left(\mathsf{b}_{\text{FGLM}} \quad\right) = \operatorname*{argmin}_{\mathsf{u}} V_n(\mathsf{u}) \stackrel{d}{!} \operatorname*{argmin}_{\mathsf{u}} V(\mathsf{u}).$$

Note that $V(\mathsf{u})$ is minimized at $\mathsf{u} = \quad^{-1}(2\lambda_0 \quad + \mathsf{W})$.

B. Discretization

B.1. Continuous Outcomes

For continuous outcomes, i.e., the regression task, we investigated two different discretization strategies, which we refer to as *equal counts* and *equal lengths*. For the equal counts strategy, we construct segments $[\delta_j, \delta_{j+1})$ which each include the same amount of samples, regardless of their group memberships, while the length of the segments are allowed to vary. In contrast, the equal lengths strategy makes each segment be the same length while the number of samples inside each segment can differ. Both strategies do not guarantee that each segment includes at least one sample from all groups which causes the penalty term to be undefined for some segments. To avoid this, we vary the number of segments starting from a large number and check if all the segments include at least one sample from all the groups. If not, we continually decrease the number of segments until we get a set of segments with each including at least one sample from all the groups. Algorithm 1 describes this discretization procedure for continuous outcomes. We intuitively expect that a larger number of segments will provide better approximations. Thus, for the experiments, we set the max number of segments to 100. We found the equal counts based discretization results (on average across 20 different splits of the training data) in 2.75, 25, and 8 segments for crime, parkinsons, and student datasets, respectively, while the equal lengths results on average in 4.2, 7, and 5 segments, respectively.

We performed additional experiments to check if the F-GLM with continuous outcomes is sensitive to the number of segments. The results are summarized in Figure 2. We see that the choice of segments can change the performance-disparity trade-off trajectories; however, the overall patterns do not change much.

We note that discretization of y was not a primary focus of our study: further investigation is likely to be worthwhile, both from a theoretical perspective as well as investigating other algorithmic discretization strategies.

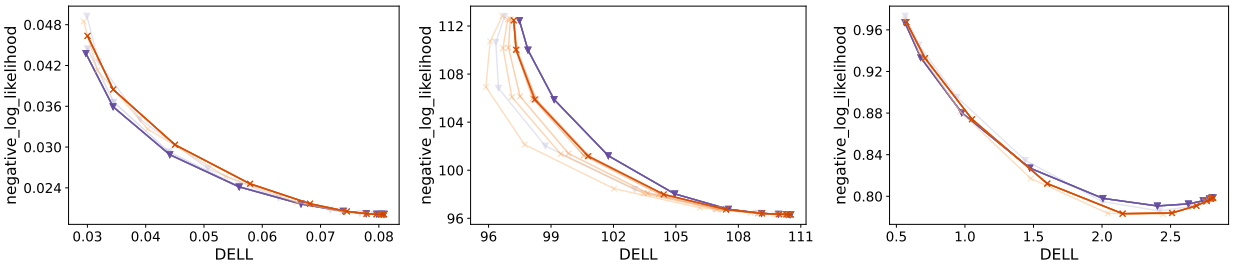
B.2. Count Outcomes

For count outcomes, i.e., the Poisson regression task, we find the smallest and the largest integers L and U satisfying $f(\mathbf{x}_i, y_i, A_i) : y_i = y, A_i = k)g \neq ?$ for all k 2 A and $L < y < U$. Then, we set $y_i = \min f y_i, Lg$ and $y_i = \max f y_i, Ug$ for all i.

C. Computational Complexity

C.1. Preparing D

Since $\mathbf{x}_i, \mathbf{x}_j$ 2 $\mathbb{R}^1$ ^{p}, the complexity of computing $(\mathbf{x}_i \quad \mathbf{x}_j)^T(\mathbf{x}_i \quad \mathbf{x}_j)$ is $O(p^2)$. Thus, the complexity to compute $\mathbf{D}^{kly}$ is $O(n^{kly}p^2)$. Moreover, we have $n^{kly} \quad n(n \quad 1)/2$. Thus, the complexity of preparing $\mathbf{D}$ is $O(n^2p^2K^2jYj)$.



(a) Crime–Continuous–Race(3) (b) Parkinsons–Continuous–Gender(2) (c) Student–Continuous–Gender(2)

Figure 2. Experimental results for equal counts (orange X markers) and equal lengths (purple triangle markers) with various numbers of segments (ranging from 1 to 3 – 25 depending on the dataset); a darker color means a greater number of segments. Our observation from the three datasets is that there is no straightforward relationship between the number of segments and better trade-off trajectories. However, the overall shape of the tradeoff trajectories between performance and disparity remains similar.

C.2. Newton-Raphson Iteration

Computing the gradient consists of two matrix multiplication operations with complexities $O(np)$ and $O(p^2)$. Moreover, the complexity of computing the Hessian is $O(np^2)$ provided W is a diagonal matrix. Also, inverting the Hessian is $O(p^3)$ since the Hessian is a dense $p \quad p$ matrix. Therefore, the per-iteration complexity of the Newton-Raphson algorithm for the F-GLM is $O(np^2 + p^3)$.

D. Datasets and Preprocessing Details

D.1. Adult Dataset

The Adult dataset contains the *income* records for 45,222 individuals from the 1994 census database, where the outcome *income* is dichotomized into a binary variable (*below $50K in income versus above $50K in income*). Each record is composed of the outcome and 14 predictors, among which 8 are categorical and 6 are continuous.

The variable *gender* was considered a sensitive attribute in this dataset (Male: 67%, Female = 33%). Other covariates of interest included age, professional occupation, education level, marital and relationship status, capital gains and losses, ethnicity, and country of origin.

Note that due to missing data the total of 48,842 instances was decreased to 45,222 after filtering out incomplete records.

D.2. Arrhythmia Dataset

The Arrhythmia dataset contains the *presence of arrhythmia status* for 418 individuals, which we treat as a binary outcome (*presence vs absent*). Each record is composed of the outcome and 80 predictors of interest.

The variable *sex* was considered a sensitive attribute in this dataset (Male: 53%, Female = 47%). Other attributes included variables such as age, height, weight, QRS duration among others.

Note that the original sample size of 452 was reduced down to 418 records due to the removal to all samples with missing data and 2 individuals with nonsensical height values.

D.3. COMPAS Dataset

The COMPAS dataset contains records for 6,172 criminal defendants across the United States of America, where we use whether each defendant became a recidivist within 2 years of the first offense as outcome (*was a recidivist within 2 years vs was not*). Each defendant has 10 predictor variables, including *gender* and *race/ethnicity*. The latter variables were considered as sensitive features and have imbalanced distributions across defendants. (Sex: Female: 19%, Male: 81%, we set Female as the baseline category. Race: Caucasian: 34%, African-American: 52%, Hispanic: 8%, Other: 6% which contain Asian and Native-American ethnicities, we set Caucasian as the baseline).

The dataset further contains 1 categorical variable, degree of the charge (F: 0.36, M: 0.64, we set F as the baseline), 5 continuous variables (age in years with mean = 34.5, number of priors counts with mean = 3.2, juvenile felony counts with mean = 0.06, juvenile misconduct counts with mean = 0.1, juvenile other category counts with mean = 0.11) and 2 time variables (time in jail (days) with mean = 15, time in custody (days) with mean = 35).

Note that the original dataset contained 7,214 records. However to ensure data quality we removed records that had a charge date of a defendant's COMPAS score crime that was not within 30 days from when the person was arrested, under the assumption that this is not the correct offense for this record. We further removed records that had missing fields for recidivism or the degree of the charge of interest, resulting in a total sample size of 6,172 with complete observation data.

D.4. Drug Consumption Dataset

The drug consumption dataset contains records for 1,885 respondents and each respondent has 12 predictor variables, including gender and race/ethnicity. Participants were questioned concerning their use of 18 legal and illegal drugs and answered with one of the following seven categories: *never used, used over a decade ago, used in the last decade, used in last year, used in last month, used in last week, and used in last day*.

Thus, we can define three different tasks based on the participant's response: binary classification of classifying *never used* versus *the others (ever used)* and both ordered and unordered multiclass classification classifying the original seven

categories of the outcomes. For the binary classification we used *methadone*. Further, we choose to use *methamphetamine* as a response variable, one of the most addictive and widely used drugs.

The dataset consists of 4 categorical and 8 continuous predictors. All the continuous predictors were standardized to have zero mean and unit variance a priori, by the data provider, so we did not apply any transformations. We applied one-hot encoding to all the categorical predictors.

There are two sensitive attributes: *gender* (Female: 50%, Male: 50%) and *race/ethnicity* (Asian: 1.38%, Black: 1.75%, Mixed-Black/Asian 0.16%, Mixed-White/Asian: 1.06%, Mixed-White/Black: 1.06%, Other 3.34%, and White: 91.25%). Because of the severe imbalance in race/ethnicity, we merged all the non-White race/ethnicity into a single non-White category.

D.5. German Credit Dataset

The German Credit dataset contains records for 1000 individuals, describing the level of risk for their *credit*, which we treated as binary (*good vs bad*). Each record is composed of the outcome and 21 predictor variables among which 14 were categorical and 7 were continuous.

The variable *sex* was considered a sensitive attribute in this dataset (Male: 69%, Female = 31%). Other covariates of interest included the status of checking account, credit history, saving in accounts and bonds, employment status, property ownership, disposable income and age among others.

There were no reduction from the original sample size due to missingness, since all records had complete information available.

D.6. Communities and Crime Dataset

The Communities and Crime dataset contains the criminal records for 1,994 communities in the United States of America from socio-economic data from 1990 US Census and law enforcement data from the 1990 US LEMAS survey. The outcome of interest is the *violent crimes per population* (continuous). Each record is composed of the outcome and 31 predictors, among which is the sensitive attribute *race* (stratified between Asian: 4%, Black: 11%, Hispanic: 6%, White: 79%). Other predictors included age, income, urbanism, police budget among others.

Note that the original sample size was significantly reduced due to high levels of data missingness across predictors. Moreover, the variables, state, county, community, community name and the fold for cross-validation were removed as they serve no purpose for prediction.

D.7. Law School Admission Council Dataset

The Law School Admission Council (LSAC) dataset contains records for 22,407 Law School students gathered by a National Longitudinal Study primarily undertaken in response to reports suggesting bar passage rates were lower among examinees of color. The outcome of interest is the *Grade Point Average* of students during Law School which is a continuous variable. We consider both the *race/ethnicity* and the *gender* of students to be sensitive factors. (Gender: Female = 44%, Male = 56%, we set Female as the baseline category. Race: White/Caucasian: 88.2%, African-American: 6%, Asian: 4%, we group all other ethnicities under Other: 1.8%. We set White/Caucasian as the baseline category). We further included two continuous variables as predictors, namely the LSAT score of each student (median 37, IQR 33-41) and the university GPA of students (median 3.2, IQR 3-3.5) and a single categorical variable specifying if students are participating in the academic program at full or part time (part time: 7.7%, with full time being the baseline category).

After removing observations containing missing data, the final dataset contained records for 22,368 students with complete information.

D.8. Parkinson's Telemonitoring Dataset

The Parkinson's Telemonitoring dataset contains the record of 42 patients with early-stage Parkinson's disease recruited through a six-month trial of telemonitoring for remote symptom progression monitoring. The dataset includes 5,875 instances of data observations across all patients with outcome *Unified Parkinson's Disease Rating Scale (UPDR) score* which is a *continuous* value evaluating various aspects of Parkinson's disease.

The sensitive attribute of the dataset was set to be the *sex* of patients (Female: 33%, Male: 67%). We further included 16 predictors. All records had complete information and thus there was no sample size reduction due to missingness.

D.9. Student Performance Dataset

The Student dataset contains records for 382 students for 2 separate classes (mathematics and Portuguese) over 3 trimesters, we use the grades students received in the third trimester (numeric score between 0 and 20). We separate this dataset into two separate datasets, with mathematics scores and Portuguese scores respectively. The sensitive variable in this dataset is the *sex* of students. (Sex: Female: 48%, Male: 52%, we set Female as the baseline). The dataset share the same set of 25 further predictors, among which 3 are continuous and 21 are categorical. Note that some categorical variables were further collapsed to eliminate smaller categories (such as the education level of the mother and father, the travel and study times, family relationship statuses and free time levels of students).

D.10. Health & Retirement Survey Dataset

The University of Michigan Health and Retirement Study (HRS) longitudinal dataset, recording survey responses on health and aging. The dataset contains 12,744 instances. The *number of dependencies in daily activities* was set as the target outcome as a *count variable*. This is encoded as the *score* in the dataset, ranging from 0 to 10.

The *ethnicity* of patients was set as the sensitive attribute of the dataset (Afro-American: 15%, Hispanic: 10%, Other: 2% and White: 73%). The 22 predictor variables included gender, marital status, age, education and net worth among others. Note that large portions of the data entries were missing and we considered only complete cases.

D.11. Obesity Dataset

The Obesity dataset contains the health records of 2,111 individuals with their assessed level of *obesity*. *Obesity* is treated a multilevel outcome with levels: *Insufficient weight* (13%), *normal weight* (13%), *overweight level I* (14%), *overweight level 2* (14%), *obesity type I* (17%), and *obesity type II/III* (29%).

The *gender* of patients was set as the sensitive attribute of the dataset (Female: 49%, Male: 51%), and we further included 14 predictor variables such as age, family history and smoking status. Note that there were no missing data in this dataset, all individuals had complete information.

E. Experimental Setting Details

E.1. Implementation of Competitive Methods

For the fair constraints (Zafar et al., 2017b) and disparate mistreatment (Zafar et al., 2017a) methods, we used the Python code provided by the authors[1]. For the squared difference penalty (Bechavod & Ligett, 2017), and the group and individual fairness convex penalty (Berk et al., 2017), we adapted our Newton-Raphson method (all three methods can be expressed in the same form as that of F-GLM using a different $\mathbf{D}$.) Note that both papers suggested using CVXPY (Diamond & Boyd, 2016; Agrawal et al., 2018), which is an off-the-shelf optimization solver, for solving their problems. We also implemented the HSIC penalty (Pérez-Suay et al., 2017) which can easily be solved because the problem has a closed form solution for the linear case. For the linear FERM method, we used its Python implementation provided by the authors[2]. We used the `fairlearn`[3] Python package for the reductions approach with statistical parity or bounded group loss (Agarwal et al., 2018; 2019). Specifically, we used grid search (`GridSearch` function) instead of the exponentiated gradient method to get a single model. For the general FERM (Oneto et al., 2020), since we did not find any available code online, we implemented the method with CVXPY. Note that the authors suggested using CPLEX[4], which is an off-the-shelf optimization solver.

[1] https://github.com/mbilalzafar/fair-classification
[2] https://github.com/jmikko/fair_ERM
[3] https://fairlearn.org/
[4] https://www.ibm.com/analytics/cplex-optimizer

E.2. Hyperparameters

We presented the range of the hyperparameters used for our experiments in Table 4. For some datasets, we used slightly different range of hyperparameters for some methods. Details can be found in our code.

Table 4. The range of the hyperparameters that control accuracy-fairness trade-off used for the experiments

Methods	hyperparameter	min value	max value
Generalized Linear Model	-	-	-
Linear SVM	-	-	-
Fair Constraints	c	10^{-3}	20
Disparate Mistreatment	c	10^{-3}	20
Squared Difference		10^{-3}	10
Group Fairness		10^{-3}	10
Individual Fairness		10^{-3}	10
HSIC Penalty		10^{-3}	5
FERM		0	0
GFERM		0	100
Statistical Parity	w	0	1
Bounded Group Loss	w	0	1
Fair GLM		10^{-3}	10

*We used $\in \{0.05, 0.01\}$ for SVM and FERM.
**The code provided by the authors allows only $= 0$.
***We varied constraint _weight parameter of GridSearch function.

F. Additional Experimental Results

We provide additional plots that summarize the experimental results here. Figure 4 shows the overall performance and disparity in mean squared error (for binary and multiclass classification) – also referred to as Brier score. The overall patterns are quite similar to those in Figure 1 in the main paper, supporting the conclusion that the F-GLM can produce favorable performance-disparity trajectories. For regression tasks, the negative log-likelihoods and mean squared errors are equivalent.

Figure 5 shows the performance and disparity measured for miscellaneous task-specific metrics; AUROC for binary classification, mean absolute error (MAE) for regression, misclassification rate for multiclass classification, and MAE for Poisson regression. We note that AUROC is the only higher-the-better metric so we plotted 1-AUROC instead to be consistent with other metrics. Also, AUROC cannot be calculated for each class because it is a concordance score, so it was not separately calculated for each class. Here the results are more mixed than in Figure 1, in Figure 3 or in Figure 4 (in terms of comparing F-GLM with other methods), but this is to be expected since performance criteria such as AUROC are not necessarily highly correlated with the other performance metrics.

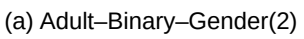
References – Supplementary

Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., and Wallach, H. A reductions approach to fair classification. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 60–69. PMLR, 10–15 Jul 2018.

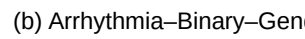
Agarwal, A., Dudik, M., and Wu, Z. S. Fair regression: Quantitative definitions and reduction-based algorithms. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 120–129. PMLR, 09–15 Jun 2019.

Agrawal, A., Verschueren, R., Diamond, S., and Boyd, S. A rewriting system for convex optimization problems. *Journal of Control and Decision*, 5(1):42–60, 2018.

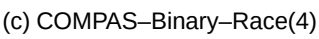
Bechavod, Y. and Ligett, K. Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044*, 2017.



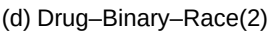
(a) Adult–Binary–Gender(2)



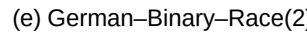
(b) Arrhythmia–Binary–Gender(2)



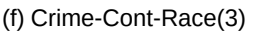
(c) COMPAS–Binary–Race(4)



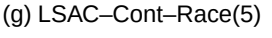
(d) Drug–Binary–Race(2)



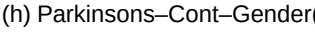
(e) German–Binary–Race(2)



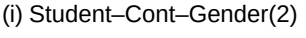
(f) Crime-Cont-Race(3)



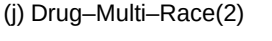
(g) LSAC–Cont–Race(5)



(h) Parkinsons–Cont–Gender(2)

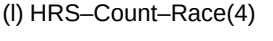


(i) Student–Cont–Gender(2)



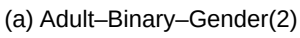
(j) Drug–Multi–Race(2)

(k) Obesity–Multi–Gender(2)

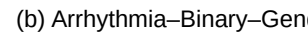


(l) HRS–Count–Race(4)

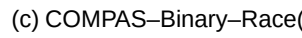
Figure 3.Experimental results in negative log-likelihoods and D_{EO} from 11 real world datasets with binary (a-e) and continuous outcomes (f-i). Each subtitle is in the form of Dataset–Outcome Type–Sensitive Attribute(K). For both binary and continuous outcomes we use a Generalized Linear Model (GLM, red star), Fair Generalized Linear Model (F-GLM, black X), Individual Fairness penalty (IF, green pentagon), Group Fairness penalty (GF, dark red square), and Bounded Group Loss (BGL, orange triangle). Methods for binary outcomes also include the Support Vector Machine (SVM, grey hexagon), Fair Constraints (FC, green diamond), Disparate Mistreatment (DM, blue circle), Squared Difference penalizer (SD, dark blue diamond), Fair Empirical Risk Minimization (FERM, plum pentagon), Statistical Parity (SP, teal triangle). Methods for continuous outcomes include the HSIC penalty (HSIC, blue circle), General Fair Empirical Risk Minimization (GFERM, plum pentagon). See Table 1 for additional information for each method. Each dot represents mean performance across test sets for a specific hyperparameter value



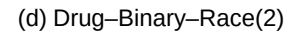
(a) Adult–Binary–Gender(2)



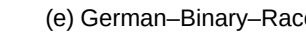
(b) Arrhythmia–Binary–Gender(2)



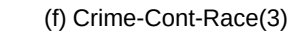
(c) COMPAS–Binary–Race(4)



(d) Drug–Binary–Race(2)

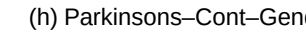


(e) German–Binary–Race(2)

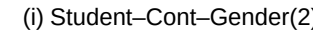


(f) Crime-Cont-Race(3)

(g) LSAC–Cont–Race(5)

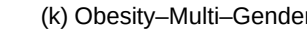


(h) Parkinsons–Cont–Gender(2)

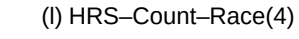


(i) Student–Cont–Gender(2)

(j) Drug–Multi–Race(2)

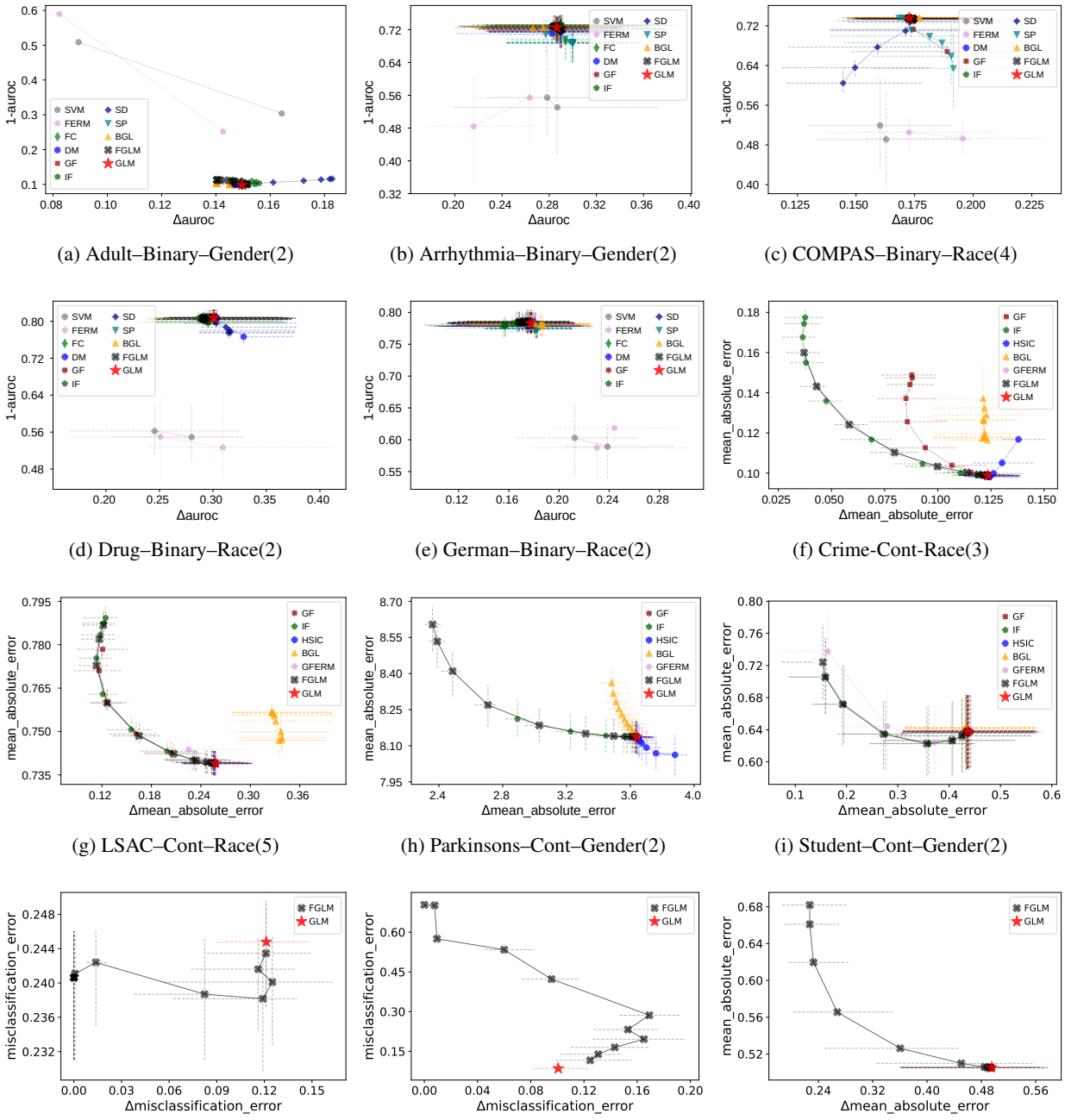


(k) Obesity–Multi–Gender(2)



(l) HRS–Count–Race(4)

Figure 4.Experimental results in mean squared errors (also referred to as brier scores for classi cation problems) from 11 real world datasets with binary (a-e) and continuous outcomes (f-k). (x-axis is disparity of mean squared errors. Each subtitle is in the form of Dataset–Outcome Type–Sensitive Attribute(k)) For both binary and continuous outcomes we use a Generalized Linear Model (GLM, red star), Fair Generalized Linear Model (F-GLM, black X), Individual Fairness penalty (IF, green pentagon), Group Fairness penalty (GF, dark red square), and Bounded Group Loss (BGL, orange triangle). Methods for binary outcomes also include the Support Vector Machine (SVM, grey hexagon), Fair Constraints (FC, green diamond), Disparate Mistreatment (DM, blue circle), Squared Difference penalizer (SD, dark blue diamond), Fair Empirical Risk Minimization (FERM, plum pentagon), Statistical Parity (SP, teal triangle). Methods for continuous outcomes include the HSIC penalty (HSIC, blue circle), General Fair Empirical Risk Minimization (GFERM, plum pentagon). See Table 1 for additional information for each method. Each dot represents mean performance across test sets for a speci c hyperparameter value



(a) Adult–Binary–Gender(2) (b) Arrhythmia–Binary–Gender(2) (c) COMPAS–Binary–Race(4)

(d) Drug–Binary–Race(2) (e) German–Binary–Race(2) (f) Crime-Cont-Race(3)

(g) LSAC–Cont–Race(5) (h) Parkinsons–Cont–Gender(2) (i) Student–Cont–Gender(2)

(j) Drug–Multi–Race(2) (k) Obesity–Multi–Gender(2) (l) HRS–Count–Race(4)

Figure 5. Experimental results *in some other metrics (AUROC/MAE/misclassification rate)* from 11 real world datasets with binary (a-e) and continuous outcomes (f-i). The x-axis is the disparity of each metric. Each subtitle is in the form of Dataset–Outcome Type–Sensitive Attribute(K). For both binary and continuous outcomes we use a Generalized Linear Model (GLM, red star ★), Fair Generalized Linear Model (F-GLM, black X ✖), Individual Fairness penalty (IF, green pentagon ⬟), Group Fairness penalty (GF, dark red square ■), and Bounded Group Loss (BGL, orange triangle ▲). Methods for binary outcomes also include the Support Vector Machine (SVM, grey hexagon ⬢), Fair Constraints (FC, green diamond ♦), Disparate Mistreatment (DM, blue circle ●), Squared Difference penalizer (SD, dark blue diamond ◆), Fair Empirical Risk Minimization (FERM, plum pentagon ⬟), Statistical Parity (SP, teal triangle ▼). Methods for continuous outcomes include the HSIC penalty (HSIC, blue circle ●), General Fair Empirical Risk Minimization (GFERM, plum pentagon ⬟). See Table 1 for additional information for each method. Each dot represents mean performance across test sets for a specific hyperparameter value λ.

Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., Neel, S., and Roth, A. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*, 2017.

Diamond, S. and Boyd, S. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

Dobson, A. J. and Barnett, A. G. *An introduction to generalized linear models*. CRC press, 2018.

Oneto, L., Donini, M., and Pontil, M. General fair empirical risk minimization. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2020.

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*, pp. 339–355. Springer, 2017.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pp. 1171–1180, Republic and Canton of Geneva, CHE, 2017a. International World Wide Web Conferences Steering Committee. ISBN 9781450349130. doi: 10.1145/3038912.3052660.

Zafar, M. B., Valera, I., Rogriguez, M. G., and Gummadi, K. P. Fairness Constraints: Mechanisms for Fair Classification. In Singh, A. and Zhu, J. (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 962–970. PMLR, 20–22 Apr 2017b.

Zou, H. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.