Sensitivity-based measures of discrimination in insurance pricing

Mathias Lindholm[*] Ronald Richman[†] Andreas Tsanakas[‡]
Mario V. Wüthrich[§]

July 17, 2024

Abstract

Different notions of fairness and discrimination have been extensively discussed in the machine learning and insurance pricing literatures. As not all fairness criteria can be concurrently satisfied, it is important to develop metrics that allow the assessment of materiality of discriminatory effects and the trade-offs between various criteria. Methods from sensitivity analysis have been deployed for the measurement of demographic unfairness, that is, the statistical dependence of risk predictions on protected attributes. We produce a sensitivity-based measure for the different phenomenon of proxy discrimination, referring to the implicit inference of protected attributes from other covariates. For this, we first define a set of admissible prices that avoid proxy discrimination. Then, the measure is defined as the normalised L^2-distance of a price from the closest element in that set. Furthermore, we consider the attribution of the proxy discrimination measure to individual (or subsets of) covariates and investigate how properties of the data generating process are reflected in those metrics. Finally, we build on the global (i.e., portfolio-wide) measures of demographic unfairness and proxy discrimination to propose local (i.e., policyholder-specific) measures, which allow a fine-grained understanding of discriminatory effects across a collective of policyholders.

Keywords: Proxy discrimination, demographic parity, global sensitivity analysis, insurance pricing, algorithmic fairness.

1 Introduction

Questions of fairness and discrimination have become central to the wider machine learning literature (Barocas & Selbst 2016, Mehrabi et al. 2021) and, more specifically, the literature on insurance pricing (Lindholm et al. 2022, Frees & Huang 2023, Charpentier 2024). The

[*] Department of Mathematics, Stockholm University.

[†] Old Mutual Insure and University of the Witwatersrand.

[‡] Bayes Business School (formerly Cass), City, University of London.

[§] RiskLab, Department of Mathematics, ETH Zurich.

salience of these questions has increased with the rise of artificial intelligence and advanced predictive analytics, as seen by the regulatory attention directed towards the topic (e.g., EIOPA 2021, MAS 2022)

A variety of criteria have been formulated, which insurance prices should satisfy in order to be considered fair (Charpentier 2024, Xin & Huang 2024). On the one hand, *group fairness* criteria interrogate the statistical relationship between claims, prices, and protected attributes such as gender or ethnicity; these criteria are typically formulated via (conditional) independence statements. On the other hand, *individual fairness* criteria focus on whether insurance policyholders with similar risk profiles are treated similarly, i.e., are quoted comparable premiums. Here different fairness criteria arise from different notions of similarity and information restrictions, e.g., not using gender or ethnicity as a rating factor. Hence, in an insurance setting, while group fairness criteria consider the outcome of a pricing strategy, individual fairness notions revolve more around the way that these prices were generated. Furthermore, as part of the rich literature on fairness and discrimination, an understanding has developed that such criteria are not necessarily consistent with each other and can even be mutually exclusive (Kleinberg et al. 2016, Lindholm et al. 2024).

Within that context, the need emerges to construct metrics for different forms of discrimination and unfairness: one does not just need to know whether such phenomena take place within a particular insurance portfolio, but also whether the effects are material enough to justify concern and eventual action. Hence, measuring the materiality of discriminatory effects is a key ingredient of eventual regulatory action, in order to move from broad principles (EIOPA 2021) to binding regulations. Furthermore, the potential incompatibility of different fairness notions means that one cannot require for all of them to hold at the same time. Regardless of which notions of fairness are prioritised, this creates the need to monitor many aspects of possible unfairness and measure the respective trade-offs, including those arising from any (mandated or otherwise) price adjustment; for a wide discussion see the sequence of white papers by the Monetary Authority of Singapore (MAS 2022).

Here we develop measures for two distinct phenomena, *demographic unfairness* and *proxy discrimination*, with a clear emphasis on the latter. Demographic unfairness relates to violations of demographic parity, that is, the requirement that prices are statistically independent of policyholders' protected attributes. While the applicability of this particular group fairness notion in insurance has been criticised (Lindholm et al. 2024), we consider it for two reasons: first, because it is easy to explain and politically salient, therefore a potential source of reputational risk for insurers (e.g., Cook et al. 2022); and second because it helps with introducing the construction of the measures we are using. The notion of proxy discrimination builds on the understanding that some policyholder attributes, like gender or ethnicity, should not be used to calculate the price for individual policyholders, as this would constitute *direct discrimination*. Based on that premise, it is additionally desirable to avoid the effective proxying of protected attributes by other variables (e.g., car engine size, postal code) that are correlated to them. A wide range of conceptualisations of proxy discrimination exist (Tschantz 2022), with the causal structure of covariates often taking centre stage (Araiza Iturria et al. 2024, Côté et al. 2024). Here we take a view of proxy discrimination as a form of omitted variable bias, which is not contingent on assumptions of causality, but focuses on the indirect inference of protected attributes from other covariates, in the sense of Lindholm et al. (2022, 2023, 2024).

The measures of discrimination we develop are based on methods from Global Sensitivity Analysis, originating in the work of Sobol' (2001) and having been deeply studied by a variety of authors (e.g., Saltelli et al. 2008, 2010, Owen 2014, Borgonovo & Plischke 2016, Owen & Prieur 2017). Sensitivity analysis is deployed to provide insight into complex computational models, evaluate the relative importance of model inputs and identify model vulnerabilities; for applications specifically to insurance risk portfolios and insurance regulation, see respectively Rabitti & Borgonovo (2020), Vallarino et al. (2024) and Borgonovo et al. (2024). The variable importance metrics used in sensitivity analysis can thus be suitable tools for evaluating the direct and indirect impact of protected attributes on insurance prices. This is already recognised in the work of Bénesse et al. (2022) who provide sensitivity-based measures for a variety of fairness criteria, though, to our knowledge, applications of sensitivity analysis to the problem of measuring proxy discrimination are currently lacking in the literature (a very brief discussion is given in Hiabu et al. (2023)).

In Section 2 we formally introduce the ideas of demographic unfairness and proxy discrimination. Specifically, in Sections 2.3 and 2.4 we define, respectively, measures of demographic unfairness and proxy discrimination and discuss their properties. The former is already found in Bénesse et al. (2022) and we only deal with it briefly. The measure of proxy discrimination is to our knowledge new and can be understood as the distance between any given price and the closest element in a set of prices that avoid proxy discrimination. This set of admissible prices arises as a convex combination of the discrimination-free prices in Lindholm et al. (2022) and constant prices that do not depend on any policyholder characteristics. This idea is operationalised through a constrained regression of the price on best-estimate prices, calculated using different scenarios regarding the value of a protected attribute. The measure of proxy discrimination takes values between 0 and 1 for an insurance portfolio and can be evaluated for any system of prices, without reference to how these were calculated. As a result it lends itself to empirical evaluation and price auditing. We note that the process of identifying the closest element in the set of proxy-discrimination-free prices generally relies of knowledge of the joint distribution of protected characteristics and other covariates. As a result, using that element as an adjusted price could be problematic, as it would violate the stringent conditions of Lindholm et al. (2024).

While a global measure of proxy discrimination for a portfolio is useful, it is also necessary to understand which covariates are the sources for such discrimination. In Section 2.5, we discuss how to attribute the measure of proxy discrimination to covariates. Given that the measure is a result of an L^2-projection, this can be achieved by a simple adaptation of the variance-based Sobol' and total sensitivity indices (Saltelli et al. 2008).

In Section 3.1, we introduce some standard properties of the underlying data generating process and explain how the sensitivity measures respond to such properties. Subsequently, in Sections 3.2 and 3.3 we introduce, respectively, local measures of demographic unfairness and proxy discrimination. These measures are evaluated for individual policies and allow a more fine-grained understanding and visualisation of the way in which discriminatory effects may arise in a portfolio. The local measures we propose are simply given by the price of a policy minus a benchmark price that is free of the particular type of unfairness considered. For the case of demographic unfairness, the benchmark price comes from an Output Optimal Transport transformation of the portfolio's prices (Lindholm et al. 2024, Charpentier 2024). For the case of proxy discrimination, the benchmark is given by the closest element in the set of prices that avoid proxy discrimination.

Finally, we conclude with Section 4, where we briefly discuss various important aspects of the problem not fully addressed by the definition of the measures of demographic unfairness and proxy discrimination, namely: the differences in the notion of proxy discrimination of Lindholm et al. (2024) with the current paper; questions of whether evaluation should be under a portfolio or market distribution; the impact of model uncertainty; and the calculation of proxy discrimination for the prices actually charged to policyholders, which are generally different to pure risk predictions.

2 Measures of proxy discrimination and demographic unfairness

2.1 Setup and notation

We work on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with $\mathbb{P}$ the real-world probability measure. On that space we consider the random vector $(Y, \boldsymbol{X}, D)$. The random variable Y represents a loss (or loss frequency), to be predicted based on covariates $(\boldsymbol{X}, D)$. Of these, $\boldsymbol{X}$ captures *non-protected covariates* (non-discriminatory characteristics), while D reflects a *protected attribute* (discriminatory or sensitive characteristic). The variability of $(\boldsymbol{X}, D)$ under $\mathbb{P}$ represents portfolio heterogeneity, while the variability of Y conditional on $\{\boldsymbol{X} = \boldsymbol{x}, D = d\}$ reflects loss uncertainty for a policyholder with known features $(\boldsymbol{x}, d)$. We will assume throughout that D is discrete and finite, taking values in the set $\mathfrak{D}$.

Throughout, for a generic random variable Z, we represent its distribution function by $\mathbb{P}(z)$; the conditional distribution of Z given $W = w$ is denoted accordingly by $\mathbb{P}(z|w)$. In the case of absolutely continuous random variables Z, probability density functions are given by $\mathrm{d}\mathbb{P}(z)/\mathrm{d}z$. In the case of discrete Z we have probability weights $\mathbb{P}(z) = \mathbb{P}(Z = z) > 0$.

2.2 Pricing functions and discriminatory effects

We define the *best-estimate price* as

$$\mu(\boldsymbol{x}, d) := \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}, D = d], \tag{1}$$

such that $\mu(\boldsymbol{x}, d)$ is the optimal (in L^2-norm) prediction of the loss Y for a policyholder with features $(\boldsymbol{x}, d)$.

Best-estimate prices have discriminatory effects because of their direct dependence on protected characteristics D. The most straightforward way of correcting for such *direct discrimination*, is to calculate insurance prices without including the information D as a covariate for prediction of Y. The resulting conditional expectation based only on non-protected characteristics $\boldsymbol{X}$ is termed the *unawareness price* and is defined by

$$\mu(\boldsymbol{x}) := \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}]. \tag{2}$$

Nonetheless, the unawareness price may still have discriminatory effects, arising from the potential dependence between the random vectors $\boldsymbol{X}$ and D. We note that such dependence need not have a causal source, but just be a feature of a particular portfolio of insurance policies, e.g., when female and male policyholders have different age distributions.

Out of many different notions of unfairness or discrimination we focus on two complementary perspectives on the impact of the dependence of protected attributes and non-protected covariates. First, given such dependence it will generally hold that $\mu(\boldsymbol{X})$ is also dependent on D. This means that insurance prices may vary across demographic groups, such that for some $d \neq d', \ d, d' \in \mathfrak{D}$, we have that $\mathbb{E}[\mu(\boldsymbol{X}) \mid D = d] \neq \mathbb{E}[\mu(\boldsymbol{X}) \mid D = d']$. For short, we call such insurance prices *demographically unfair*.

Second, removing D from the set of covariates does not imply that these are not used *indirectly* in pricing. A concern is that, if D can be partially predicted from $\boldsymbol{X}$, it is possible that insurance prices, derived with the aim of maximising predictive accuracy, implicitly use $\boldsymbol{X}$ to infer D. In fact, by observing that we can write unawareness prices as

$$\mu(\boldsymbol{x}) = \sum_{d \in \mathfrak{D}} \mu(\boldsymbol{x}, d)\, \mathbb{P}(d \mid \boldsymbol{x}), \tag{3}$$

it becomes clear that unawareness prices do indeed rely on implicit inference of D from $\boldsymbol{X}$, via the conditional probability $\mathbb{P}(d \mid \boldsymbol{x})$ used in averaging over best-estimates. We say that prices utilising such inference of protected characteristics are subject to *proxy discrimination*.

2.3 Measuring demographic unfairness

We now put the previous ideas on a more formal footing, which applies to a general pricing functional $\boldsymbol{X} \mapsto \pi(\boldsymbol{X})$. We start with the well-known idea of demographic parity.

Definition 1. *The pricing functional* $\boldsymbol{X} \mapsto \pi(\boldsymbol{X})$ *satisfies* demographic parity *with respect to* $\mathbb{P}(\boldsymbol{X}, D)$*, if the random variable* $\pi(\boldsymbol{X})$ *is independent of* D *under* $\mathbb{P}$*. If* π *violates demographic parity, we say that it is* demographically unfair*.*

Note that independence of $\boldsymbol{X}$ and D is a sufficient but not a necessary condition for demographic parity. Furthermore, establishing that a pricing functional π is demographically unfair does not necessarily mean that the resulting impact is substantial. To understand the materiality of unfairness, one can visualise the changes in (empirical estimates of) the conditional density of $(\pi(\boldsymbol{X})|D = d)$ across demographic groups $d \in \mathfrak{D}$. Sometimes though, a simple numerical metric is useful. We now present such a metric, following Bénesse et al. (2022) who apply ideas from sensitivity analysis to evaluate various concepts of algorithmic (un)fairness.

Definition 2. *The demographic unfairness metric* UF *is defined as*

$$\mathrm{UF}(\pi) = \frac{\mathrm{Var}(\mathbb{E}[\pi(\boldsymbol{X}) \mid D])}{\mathrm{Var}(\pi(\boldsymbol{X}))}, \tag{4}$$

with the convention that if $\mathrm{Var}(\pi(\boldsymbol{X})) = 0$*, then* $\mathrm{UF}(\pi) = 0$*.*

The rationale for the construction (4) is well established in the different context of Global Sensitivity Analysis and the metric is known as a *Sobol' Index* (e.g. Sobol' 2001, Saltelli et al. 2008). Its interpretation follows from the decomposition $\mathrm{Var}(\pi(\boldsymbol{X})) = \mathrm{Var}(\mathbb{E}[\pi(\boldsymbol{X}) \mid$

$D]) + \mathbb{E}[\mathrm{Var}(\pi(\boldsymbol{X}) \mid D)]$, where the term standing as numerator in (4) represents the amount of variation in $\pi(\boldsymbol{X})$ attributable to the protected variable D.

In (4), we suppress the dependence of the metric on $\mathbb{P}(\boldsymbol{X}, D)$. The unfairness metric reflects demographic disparities by representing the variability of average prices across demographic groups as a portion of the overall variance of prices. These demographic properties are reflected by $\mathbb{P}(\boldsymbol{X}, D)$, which can be a population distribution, reflecting the interdependence of covariates across a society. However, $\mathbb{P}(\boldsymbol{X}, D)$ can also be the distribution of covariates within a specific insurance portfolio; in that case it may reflect to a lesser extent some widely applicably demographic realities and more the structure of an individual portfolio, which can differ across insurers. We return to this point in Section 4.2.

The following easily derived properties are stated without proof.

Proposition 1. *The unfairness metric* UF *satisfies the following properties.*

i) $0 \leq \mathrm{UF}(\pi) \leq 1$. *Furthermore, for all* $a, b \in \mathbb{R}$ *it holds that* $\mathrm{UF}(a + b\pi) = \mathrm{UF}(\pi)$.

ii) If π *satisfies demographic parity with respect to* $\mathbb{P}(\boldsymbol{X}, D)$, *then* $\mathrm{UF}(\pi) = 0$.

iii) If $\sigma(\pi(\boldsymbol{X})) \subseteq \sigma(D)$, *i.e.,* $\pi(\boldsymbol{X})$ *is* D*-measurable, then* $\mathrm{UF}(\pi) = 1$.

Example 1. Let $D \in \{0, 1\}$, $X \sim \mathrm{U}(0, 1)$ and $\mathbb{P}(D = 1|X) = \mathbb{E}[D \mid X] = X$. This implies $\mathbb{P}(D = 1) = \frac{1}{2}$. Assume that the best-estimate price is

$$\mu(X, D) = \frac{1}{2} + X + D,$$

which includes a fixed cost. In this model there is the potential for demographic unfairness, since (X, D) are dependent. The unawareness price equals

$$\mu(X) = \mathbb{E}\left[\frac{1}{2} + X + D \;\middle|\; X\right] = \frac{1}{2} + 2X.$$

Straightforward calculations lead to

$$\begin{aligned}\mathbb{E}[\mu(X) \mid D = 0] &= \frac{1}{2} + 2\mathbb{E}[X \mid D = 0] = \frac{1}{2} + \frac{2}{3},\\ \mathbb{E}[\mu(X) \mid D = 1] &= \frac{1}{2} + 2\mathbb{E}[X \mid D = 1] = \frac{1}{2} + \frac{4}{3}.\end{aligned}$$

It then follows that $\mathrm{Var}(\mathbb{E}[X \mid D]) = 1/36$. Hence, if the unawareness price is used, the average premium for $D = 0$ is different compared to $D = 1$. Consequently demographic unfairness arises. We can quantify this effect via the UF metric:

$$\mathrm{UF} = \frac{\mathrm{Var}(\mathbb{E}[1/2 + 2X \mid D])}{\mathrm{Var}(1/2 + 2X)} = \frac{\mathrm{Var}(\mathbb{E}[X \mid D])}{\mathrm{Var}(X)} = \frac{1}{3} > 0.$$

■

2.4 Defining and quantifying proxy discrimination

We now turn our attention to the measurement of proxy discrimination. Here the situation is different, compared to unfairness. While violations of demographic parity relate to the statistical properties of the pricing functional, the issue of proxy discrimination arises from the way that the pricing functional is constructed. Recall that the best-estimate prices $\mu(\boldsymbol{X}, D)$ are not used, as they would give rise to direct discrimination and some other pricing functional $\pi(\boldsymbol{X})$ must be used instead. Effectively this corresponds to merging rating classes with the same non-protected profile $\{\boldsymbol{X} = \boldsymbol{x}\}$ but different protected characteristics $\{D = d\}$. Hence, one can calculate the price in each of those new classes as a weighted average over d of the corresponding costs $\mu(\boldsymbol{x}, d)$. Thus, pricing relates to reallocating the claims costs $\mu(\boldsymbol{x}, d)$, following the removal of the explanatory effect of D. To avoid proxy discrimination – unlike the situation of unawareness prices (3) – the weights used should not depend on $\boldsymbol{x}$.

To address this issue, Lindholm et al. (2022) suggest the pricing formula

$$h^*(\boldsymbol{X}) = \sum_{d \in \mathfrak{D}} \mu(\boldsymbol{X}, d)\mathbb{P}^*(d), \tag{5}$$

for some distribution $\mathbb{P}^*(d)$ on $\mathfrak{D}$. Here, we build on Lindholm et al. (2022), to construct an expanded set of admissible pricing functionals that we consider to be free from proxy discrimination. We argue that a constant price $\pi(\boldsymbol{X}) \equiv \pi$ that does not depend on the covariates $\boldsymbol{X}$ cannot proxy-discriminate, as it does not discriminate between policyholders in any sense. Consequently, we would not like to *ex ante* exclude convex combinations of the form

$$(1-\alpha)\pi + \alpha h^*(\boldsymbol{X}), \quad \alpha \in [0,1], \tag{6}$$

from our admissible set of prices. Comparing with equation (5), this means that we both allow for an additive constant offset and also that the weights $v_d := \alpha\mathbb{P}^*(d)$ that are applied on the best-estimate prices $\mu(\boldsymbol{X}, d)$ may sum to less than 1.

Proxy discrimination is now defined with the above arguments in mind. Define the set $\mathcal{V} := \{\boldsymbol{v} \in [0,1]^{|\mathfrak{D}|} \ : \ \sum_{d \in \mathfrak{D}} v_d \le 1\}$.

Definition 3. *The pricing functional* $\boldsymbol{X} \mapsto \pi(\boldsymbol{X})$ avoids proxy discrimination *with respect to* $\mu(\boldsymbol{X}, D)$, *if for* $\mathbb{P}$*-almost every* $\boldsymbol{X}$ *we can write*

$$\pi(\boldsymbol{X}) = c + \sum_{d \in \mathfrak{D}} \mu(\boldsymbol{X}, d) v_d, \tag{7}$$

for some $c \in \mathbb{R}$ *and* $\boldsymbol{v} \in \mathcal{V}$ *that do not depend on* $\boldsymbol{X}$. *If* π *does not have that structure, we say that it is* proxy-discriminatory.

Following Definition 3, we can now define a measure of proxy discrimination.

Definition 4. *The proxy discrimination metric* PD *is defined as*

$$\mathrm{PD}(\pi) = \frac{\min_{c \in \mathbb{R},\ \boldsymbol{v} \in \mathcal{V}} \mathbb{E}\left[\left(\pi(\boldsymbol{X}) - c - \sum_{d \in \mathfrak{D}} \mu(\boldsymbol{X}, d) v_d\right)^2\right]}{\mathrm{Var}(\pi(\boldsymbol{X}))}, \tag{8}$$

with the convention that if $\mathrm{Var}(\pi(X)) = 0$, *then* $\mathrm{PD}(\pi) = 0$.

The metric PD quantifies the extent to which a pricing functional cannot be expressed as a weighted average of best-estimate cost terms $\mu(\boldsymbol{x}, d)$, allowing also for a fixed cost term. Note that the presence of the intercept c, even with the constraints on $\boldsymbol{v}$, ensures that any solution $(c^*, \boldsymbol{v}^*)$ of the regression problem (8), satisfies $\mathbb{E}\left[\pi(\boldsymbol{X}) - c^* - \sum_{d\in\mathfrak{D}} \mu(\boldsymbol{X}, d)v_d^*\right] = 0$. (We note that even though $c^*, \boldsymbol{v}^*$ need not be unique, the quantity $c^* + \sum_{d\in\mathfrak{D}} \mu(\boldsymbol{X}, d)v_d^*$ is.) Hence, we can explicitly solve for c and express the numerator of (8) as

$$\min_{\boldsymbol{v}\in\mathcal{V}} \mathbb{E}\left[\left(\pi(\boldsymbol{X}) - \mathbb{E}[\pi(\boldsymbol{X})] - \sum_{d\in\mathfrak{D}} \big(\mu(\boldsymbol{X}, d) - \mathbb{E}[\mu(\boldsymbol{X}, d)]\big)v_d\right)^2\right].$$

Consequently we can write

$$\mathrm{PD}(\pi) = \frac{\mathrm{Var}\left(\pi(\boldsymbol{X}) - \sum_{d\in\mathfrak{D}} \mu(\boldsymbol{X}, d)v_d^*\right)}{\mathrm{Var}(\pi(\boldsymbol{X}))} \tag{9}$$

and $\mathrm{PD}(\pi)$ can be understood as one minus the coefficient of determination for the constrained regression of $\pi(\boldsymbol{X})$ on $\mu(\boldsymbol{X}, d),\ d \in \mathfrak{D}$.

The quantity $\pi^*(\boldsymbol{X}) := c^* + \sum_{d\in\mathfrak{D}} \mu(\boldsymbol{X}, d)v_d^*$ can be understood as the closest element to $\pi(\boldsymbol{X})$ in the set of prices that are free from proxy discrimination. We do *not* specifically suggest $\pi^*(\boldsymbol{X})$ as a suitable price correction for $\pi(\boldsymbol{X})$, since adopting such a practice would be inconsistent with the requirements on avoiding proxy discrimination, as formulated by Lindholm et al. (2024), given the optimal $(c^*, \boldsymbol{v}^*)$ generally will depend on the joint distribution of $(\boldsymbol{X}, D)$. We return to this point in the discussion of Section 4.1.

Again, simple properties of the metric can be stated.

Proposition 2. *The proxy discrimination metric* PD *satisfies the following properties.*

i) $0 \leq \mathrm{PD}(\pi) \leq 1$. *Furthermore, for all* $a \in \mathbb{R}, b \in \mathbb{R}_+$ *it holds that* $\mathrm{PD}(a + b\pi) = \mathrm{PD}(\pi)$.

ii) If π *avoids proxy discrimination with respect to* $\mu(\boldsymbol{X}, D)$*, then* $\mathrm{PD}(\pi) = 0$.

iii) If $\pi(\boldsymbol{X})$ *is uncorrelated with* $\mu(\boldsymbol{X}, d)$ *for all* $d \in \mathfrak{D}$*, then* $\mathrm{PD}(\mu) = 1$.

Proof. Parts i) and ii) are immediate.

For iii), uncorrelatedness implies that in the regression (8) we have $v_d^* = 0$ for all $d \in \mathfrak{D}$ and $c^* = \mathbb{E}[\pi(\boldsymbol{X})]$. Consequently

$$\min_{c\in\mathbb{R},\ \boldsymbol{v}\in\mathcal{V}} \mathbb{E}\left[\left(\pi(\boldsymbol{X}) - c - \sum_{d\in\mathfrak{D}} \mu(\boldsymbol{X}, d)v_d\right)^2\right] = \mathrm{Var}(\pi(\boldsymbol{X})).$$

□

Parts i)-ii) of Proposition 2 show that PD is an interpretable metric for proxy discrimination, while part iii) describes a situation where proxy discrimination is maximal: the insurance prices are not at all explained by claim costs and, thus, any discrimination they achieve between policyholders must be undesirable.

In the definition of the proxy discrimination metric (8) we constrained the weights $\boldsymbol{v}$ to be less than one. The reason for this is that higher weights on terms $\mu(\boldsymbol{X}, d)$ may also produce proxying effects, as the next example shows.

Example 2. We continue with the simple model of Example 1. In that model we expect the unawareness price to be subject to proxy discrimination, as (X, D) are dependent and the best-estimate prices are sensitive in D. Let us evaluate the numerator of (8), for $\pi(X) = \mu(X) = \frac{1}{2} + 2X$. We have that

$$\begin{aligned}\mu(X) - c - \sum_{d\in\{0,1\}} \mu(X,d)v_d &= \frac{1}{2} + 2X - c - \left(\frac{1}{2} + X\right) v_0 - \left(\frac{3}{2} + X\right) v_1 \\ &= (2 - v_0 - v_1)X - c - \frac{1}{2}(v_0 + 3v_1 - 1).\end{aligned}$$

Since c can be chosen to remove the bias for any choices of v_0, v_1, we have that

$$\mathbb{E}\left[\left(\mu(X) - c^* - \sum_{d\in\{0,1\}} \mu(X,d)v_d^*\right)^2\right] = (2 - v_0^* - v_1^*)^2 \operatorname{Var}(X).$$

From this it is clear that the minimum is achieved by $v_0^* + v_1^* = 1$. Hence, the proxy discrimination metric (8) becomes

$$\mathrm{PD}(\mu) = \frac{\operatorname{Var}(X)}{\operatorname{Var}(1/2 + 2X)} = \frac{1}{4}.$$

We now consider the alternative price $\pi(X) = 3X$, noting that it agrees on average with the unawareness price, i.e., $\mathbb{E}[\mu(X)] = \mathbb{E}[\pi(X)] = 3/2$. This price penalises further policyholders with X close to 1, which we know are more likely to satisfy $D = 1$. Hence, we expect the price to proxy-discriminate even more than $\mu(X)$; moreover this will be in a gratuitous way, as the increased level of proxy discrimination does not benefit prediction accuracy. Let us now calculate $\mathrm{PD}(\pi)$. Using similar arguments as above, it follows that

$$\mathbb{E}\left[\left(\pi(X) - c^* - \sum_{d\in\{0,1\}} \mu(X,d)v_d^*\right)^2\right] = \operatorname{Var}(2X).$$

Then,

$$\mathrm{PD}(\pi) = \frac{\operatorname{Var}(2X)}{\operatorname{Var}(3X)} = \frac{4}{9} > \frac{1}{4} = \mathrm{PD}(\mu),$$

such that the increase in the degree of proxy discrimination is reflected in our metric.

Finally, any price that is free of proxy discrimination according to Definition 3 will, for some $c \in \mathbb{R}$, $\boldsymbol{v} \in \mathcal{V}$, take the form

$$\begin{aligned}\mu^*(X) &= c + v_0\mu(X,0) + v_1\mu(X,1) \\ &= c + v_0\left(\frac{1}{2} + X\right) + v_1\left(\frac{3}{2} + X\right) \\ &= c + \frac{1}{2}(v_0 + 3v_1) + (v_0 + v_1)X.\end{aligned}$$

This allows for different choices of prices that avoid proxy discrimination. For example

$$\begin{aligned} \mu_1^*(X) &:= 1 + X && (v_0 + v_1 = 1), \text{ or} \\ \mu_2^*(X) &:= \frac{5}{4} + \frac{1}{2}X && (v_0 + v_1 = 1/2), \end{aligned}$$

where $\mathbb{E}[\mu_1^*(X)] = \mathbb{E}[\mu_2^*(X)] = \mathbb{E}[\mu(X)]$. For the price $\mu_2^*(X)$ by choosing $v_0 + v_1 < 1$ we have a decreased sensitivity to claim costs and we compensate by a higher flat premium part c. However, this does not manifest in a reduction of demographic unfairness, since

$$\begin{aligned} \mathrm{UF}(\mu_1^*) &= \frac{\mathrm{Var}(\mathbb{E}[1 + X \mid D])}{\mathrm{Var}(1 + X)} = \frac{1}{3}, \\ \mathrm{UF}(\mu_2^*) &= \frac{\mathrm{Var}(\mathbb{E}[1.25 + 0.5X \mid D])}{\mathrm{Var}(1.25 + 0.5X)} = \frac{1}{3}, \end{aligned}$$

such that $\mathrm{UF}(\mu_1^*) = \mathrm{UF}(\mu_2^*) = \mathrm{UF}(\mu)$. A key observation here is that $\mu_2^*(X)$ has a lower variance, as it is less able to differentiate between policyholders and therefore provides less accurate predictions of claim costs Y. However, demographic unfairness is quantified by the UF metric as a percentage of the variance of prices, i.e., with reference to a given pricing functional's potential to differentiating between risk profiles.

■

2.5 Attribution of proxy discrimination to individual covariates

Given the measurement of proxy discrimination by (8), a next question of interest is which (subsets of) covariates – elements of $\boldsymbol{X}$ – are mostly responsible. Here we draw again from literature on Global Sensitivity Analysis (e.g. Saltelli et al. 2008). Let the dimension of $\boldsymbol{X}$ be q and $\mathcal{S} \subseteq \{1, \dots, q\} =: \mathcal{Q}$ a set of indices, such that $\boldsymbol{X}_{\mathcal{S}}$ is the corresponding sub-vector of $\boldsymbol{X}$. Analogously, denote $\mathcal{S}^c = \mathcal{Q} \setminus \mathcal{S}$, and $\boldsymbol{X}_{\mathcal{S}^c}$, such that $\boldsymbol{X} = (\boldsymbol{X}_{\mathcal{S}}, \boldsymbol{X}_{\mathcal{S}^c})$.

Noting the form (9), we can attribute the variance in the numerator to the subset $\mathcal{S}$ of covariates, by conditioning on sub-vectors. Specifically, following Definition 4, denote the regression residual by

$$\Lambda(\pi, \boldsymbol{X}) := \pi(\boldsymbol{X}) - c^* - \sum_{d \in \mathfrak{D}} \mu(\boldsymbol{X}, d) v_d^*. \tag{10}$$

We now define two metrics that reflect the contribution of (a subset of) covariates to proxy discrimination.

Definition 5. *For the proxy discrimination metric* PD *of* (8) *and* $\Lambda(\pi, \boldsymbol{X})$ *as in* (10)*, we define the contribution of the sub-vector* $\boldsymbol{X}_{\mathcal{S}}$*,* $\mathcal{S} \subseteq \mathcal{Q}$ *to proxy discrimination by the two metrics,*

$$\mathrm{PD}_{\mathcal{S}}(\pi) = \frac{\mathrm{Var}(\mathbb{E}[\Lambda(\pi, \boldsymbol{X}) \mid \boldsymbol{X}_{\mathcal{S}}])}{\mathrm{Var}(\pi(\boldsymbol{X}))}, \tag{11}$$

$$\widetilde{\mathrm{PD}}_{\mathcal{S}}(\pi) = \frac{\mathrm{Var}(\Lambda(\pi, \boldsymbol{X})) - \mathrm{Var}(\mathbb{E}[\Lambda(\pi, \boldsymbol{X}) \mid \boldsymbol{X}_{\mathcal{S}^c}])}{\mathrm{Var}(\pi(\boldsymbol{X}))}. \tag{12}$$

When $\mathcal{S} = \{i\}$*, we write* $\mathrm{PD}_i(\pi)$*,* $\widetilde{\mathrm{PD}}_i(\pi)$*.*

The metric $\mathrm{PD}_{\mathcal{S}}$ is thus understood as the sensitivity of the residual $\Lambda(\pi, \boldsymbol{X})$ to the subset of covariates $\boldsymbol{X}_{\mathcal{S}}$, reflecting the amount of variability in $\Lambda(\pi, \boldsymbol{X})$ driven by $\boldsymbol{X}_{\mathcal{S}}$. The metric $\widetilde{\mathrm{PD}}_{\mathcal{S}}$ reflects the expected reduction in the variance of $\Lambda(\pi, \boldsymbol{X})$ achieved by averaging out $\boldsymbol{X}_{\mathcal{S}}$. When $\mathcal{S} = \{i\}$, then PD_i is identified by a (rescaled) Sobol' Index (or first-order sensitivity), while $\widetilde{\mathrm{PD}}_i$ is known as a *total sensitivity* (Saltelli et al. 2008). A difference to standard sensitivity measures is that here we are normalising with $\mathrm{Var}(\pi(\boldsymbol{X}))$ – rather than $\mathrm{Var}(\Lambda(\pi, \boldsymbol{X}))$ – to maintain the direct connection with the global PD metric (8).

Estimation of the metrics (11) requires evaluating conditional expectations with respect to subsets of covariates. The sensitivity analysis literature presents various methods to do this (e.g., Jansen 1999, Sobol' 2001, Saltelli et al. 2010), though generally under the assumption of independent $\boldsymbol{X}$ which is not tenable in an insurance context. Alternative approaches are based on predictive modelling (Da Veiga et al. 2009); for consistent and efficient evaluation of non-linear regressions on all $\boldsymbol{X}_{\mathcal{S}}$, $\mathcal{S} \subseteq \mathcal{Q}$, through a single model see Richman & Wüthrich (2023).

The metrics introduced in Definition 5 have the following properties, which we state without proof.

Proposition 3. *The metrics* $\mathrm{PD}_{\mathcal{S}}$, *and* $\widetilde{\mathrm{PD}}_{\mathcal{S}}$ *satisfy the following properties.*

i) $0 \leq \mathrm{PD}_{\mathcal{S}}(\pi), \widetilde{\mathrm{PD}}_{\mathcal{S}}(\pi) \leq \mathrm{PD}(\pi)$.

ii) a) If $\boldsymbol{X}_{\mathcal{S}} \perp\!\!\!\perp \Lambda(\pi, \boldsymbol{X})$, *then* $\mathrm{PD}_{\mathcal{S}}(\pi) = 0$.

b) If $\Lambda(\pi, \boldsymbol{X})$ *is* $\boldsymbol{X}_{\mathcal{S}}$*-measurable, then* $\mathrm{PD}_{\mathcal{T}}(\pi) = \mathrm{PD}(\pi)$ *for all* $\mathcal{S} \subseteq \mathcal{T} \subseteq \mathcal{Q}$.

iii) a) If $\Lambda(\pi, \boldsymbol{X})$ *is* $\boldsymbol{X}_{\mathcal{S}^c}$*-measurable, then* $\widetilde{\mathrm{PD}}_{\mathcal{T}}(\pi) = 0$, *for all* $\mathcal{T} \subseteq \mathcal{S}$.

b) If $\boldsymbol{X}_{\mathcal{S}^c} \perp\!\!\!\perp \Lambda(\pi, \boldsymbol{X})$, *then* $\widetilde{\mathrm{PD}}_{\mathcal{S}}(\pi) = \mathrm{PD}(\pi)$.

iv) If there are functions g, h *such that* $\Lambda(\pi, \boldsymbol{X}) = g(\boldsymbol{X}_{\mathcal{S}}) + h(\boldsymbol{X}_{\mathcal{S}^c})$ *and* $\boldsymbol{X}_{\mathcal{S}} \perp\!\!\!\perp \boldsymbol{X}_{\mathcal{S}^c}$, *then* $\mathrm{PD}_{\mathcal{S}}(\pi) + \mathrm{PD}_{\mathcal{S}^c}(\pi) = \mathrm{PD}(\pi)$ *and* $\widetilde{\mathrm{PD}}_{\mathcal{S}}(\pi) + \widetilde{\mathrm{PD}}_{\mathcal{S}^c}(\pi) = \mathrm{PD}(\pi)$.

Part i) of Proposition 3 is a natural condition for stating that the metrics $\mathrm{PD}_{\mathcal{S}}(\pi)$, $\widetilde{\mathrm{PD}}_{\mathcal{S}}(\pi)$ reflect contributions to the overall proxy discrimination $\mathrm{PD}(\pi)$. Parts ii) and iii) give conditions for the metrics taking their extremal values. The conditions are complementary, using different independence or measurability assumptions to reflect irrelevance or full relevance of $\boldsymbol{X}_{\mathcal{S}}$. Specifically, when $\Lambda(\pi, \boldsymbol{X})$ is $\boldsymbol{X}_{\mathcal{S}}$-measurable (case ii)b)), then $\Lambda(\pi, \boldsymbol{X})$ is fully determined by $\boldsymbol{X}_{\mathcal{S}}$ (and its super-vectors). Hence a high value of $\mathrm{PD}_{\mathcal{S}}(\pi)$ indicates that the set $\mathcal{S}$ contains variables that have high importance. When $\Lambda(\pi, \boldsymbol{X})$ is $\boldsymbol{X}_{\mathcal{S}^c}$-measurable (case iii)a)), then $\Lambda(\pi, \boldsymbol{X})$ is fully determined by $\boldsymbol{X}_{\mathcal{S}^c}$. Hence a low value of $\widetilde{\mathrm{PD}}_{\mathcal{S}}(\pi)$ indicates that the set $\mathcal{S}$ contains variables that have low importance. Finally, part iv) gives strong conditions – independence of $\boldsymbol{X}_{\mathcal{S}}$ from $\boldsymbol{X}_{\mathcal{S}^c}$ and additivity of $\Lambda(\pi, \boldsymbol{X})$ – for being able to additively decompose the proxy discrimination metric $\mathrm{PD}(\pi)$. We resume the discussion of the properties of $\mathrm{PD}_{\mathcal{S}}$, $\widetilde{\mathrm{PD}}_{\mathcal{S}}(\pi)$ in Section 3.1.

The question of additivity is important for interpreting the contributions of covariates to proxy discrimination. Specifically, it will generally be

$$\sum_{i=1}^{q} \mathrm{PD}_i(\pi) \neq \mathrm{PD}(\pi), \quad \sum_{i=1}^{q} \widetilde{\mathrm{PD}}_i(\pi) \neq \mathrm{PD}(\pi).$$

Nonetheless, an additive decomposition of $\mathrm{PD}(\pi)$ is achievable by employing the game-theoretical concept of the *Shapley value* (Shapley et al. 1953) for the value functional $\mathcal{S} \mapsto \mathrm{Var}(\mathbb{E}[\Lambda(\pi, \boldsymbol{X}) \mid \boldsymbol{X}_{\mathcal{S}}])$. While recent literature on model interpretability has focused on the use of Shapley values to derive local model explanations for given instances $\boldsymbol{X} = \boldsymbol{x}$ (Lundberg & Lee 2017, Aas et al. 2021), we use the Shapley value for a decomposition of a global sensitivity measure, following Owen (2014), Owen & Prieur (2017), Song et al. (2016). This leads to following definition.

Definition 6. *For the proxy discrimination metric* PD *of* (8) *and* $\Lambda(\pi, \boldsymbol{X})$ *as in* (10)*, denote*

$$w(\mathcal{S}) = \mathrm{Var}(\mathbb{E}[\Lambda(\pi, \boldsymbol{X}) \mid \boldsymbol{X}_{\mathcal{S}}]), \ \mathcal{S} \subseteq \mathcal{Q}.$$

Then, we define the Shapley attribution of the covariate X_i to proxy discrimination as the metric,

$$\mathrm{PD}_i^{\mathrm{sh}}(\pi) = \frac{1}{\mathrm{Var}(\pi(\boldsymbol{X}))\, q} \sum_{\mathcal{S} \subseteq \mathcal{Q} \setminus \{i\}} \binom{q-1}{|\mathcal{S}|}^{-1} \big(w(\mathcal{S} \cup \{i\}) - w(\mathcal{S})\big). \tag{13}$$

As the Shapley value is a well-known concept across literatures, we do not review its properties here. The key practical feature is that by the use of Shapley values we achieve an additive attribution.

$$\sum_{i=1}^{q} \mathrm{PD}_i^{\mathrm{sh}}(\pi) = \mathrm{PD}(\pi).$$

Furthermore, we note that while Definition 6 calculates the Shapley value with respect to $w(\mathcal{S}) = \mathrm{Var}(\mathbb{E}[\Lambda(\pi, \boldsymbol{X}) \mid \boldsymbol{X}_{\mathcal{S}}])$ and is thus based on $\mathrm{PD}_{\mathcal{S}}$, a result identical to (13) is obtained if instead we additivise the alternative metric $\widetilde{\mathrm{PD}}_{\mathcal{S}}$ (Song et al. 2016). Hence the distinction between the two metrics of Definition 6 collapses when Shapley values are employed.

3 Structural properties, price adjustments, and local measures

3.1 Structural properties

The ideas of demographic unfairness and proxy discrimination discussed in Section 2 related to the statistical properties and construction of pricing functionals π. Here we associate such properties with structural properties of the data generating process, that is, with features of the joint distribution $\mathbb{P}(\boldsymbol{X}, D, Y)$. First, we define a number of such properties.

Definition 7 (Structural properties)**.**

P1 $\boldsymbol{X} \perp\!\!\!\perp D$ *(independence)*

P2 $Y \perp\!\!\!\perp D \mid \boldsymbol{X}$ *($\boldsymbol{X}$-sufficiency)*

P3 $\mu(\boldsymbol{X}, D) = \mu(\boldsymbol{X})$ *(weak $\boldsymbol{X}$-sufficiency)*

P4 $\sigma(\boldsymbol{X}) \subseteq \sigma(D)$ *($\boldsymbol{X}$-irrelevance)*

P5 $Y \perp\!\!\!\perp \boldsymbol{X} \mid D$ *(D-sufficiency)*

P6 $\mu(\boldsymbol{X}, D) = \mu(D)$ *(weak D-sufficiency)*

The formulation of properties P1-P6 is not reliant on any assumed causal relation between Y, $\boldsymbol{X}$ and D. Nonetheless, one can formulate Directed Acyclical Graphs (DAG) representing causal relations such that, e.g. properties P2 or P5 are satisfied. We do not pursue this route here. We note however that causal inference is at the heart of many discussions of algorithmic fairness – indicatively we mention the seminal paper by Kusner et al. (2017), the overview of proxy discrimination by Tschantz (2022) and the insurance-specific investigations Araiza Iturria et al. (2024), Côté et al. (2024).

In the light of Definitions 1, 2, 3, 4, the relationships summarised in Proposition 4, below, hold. These show the implications of structural properties P1-P6 for the metrics UF and PD, for either a general price $\pi(\boldsymbol{X})$ or, more specifically, the unawareness price $\mu(\boldsymbol{X})$.

Proposition 4.

- *i) If P1 holds, any pricing functional* $\pi(\boldsymbol{X})$ *satisfies demographic parity and* $\mathrm{UF}(\pi) = 0$.
- *ii) If any of P1, P2 or P3 holds, the unawareness price* $\mu(\boldsymbol{X})$ *avoids proxy discrimination and* $\mathrm{PD}(\mu) = 0$.
- *iii) If P4 holds, any pricing functional* $\pi(\boldsymbol{X})$ *is demographically unfair and* $\mathrm{UF}(\pi) = 1$, *except if* $\pi(\boldsymbol{X}) \equiv \pi$*, a constant.*
- *iv) If any of P4, P5 or P6 holds, any pricing functional* $\pi(\boldsymbol{X})$ *is proxy discriminatory and* $\mathrm{PD}(\pi) = 1$, *except if* $\pi(\boldsymbol{X}) \equiv \pi$*, a constant.*

Proof. Part i) follows from P1 $\implies \pi(\boldsymbol{X}) \perp\!\!\!\perp D$. For part ii) note from (3), that P1 $\implies$ $\mathbb{P}(D = d|\boldsymbol{x}) = \mathbb{P}(D = d)$, from which $\mathrm{PD}(\mu) = 0$ follows. Finally, P2 $\implies$ P3. Then the regression in Definition 4 reduces to

$$\min_{c,\boldsymbol{v}} \mathbb{E}\left[\left(\mu(\boldsymbol{X}) - c - \mu(\boldsymbol{X})\sum_d v_d\right)^2\right] = 0.$$

Part iii) is immediate. For part iv) note that either of P4 or P5 implies P6. The implication from P6 is a special case of Proposition 2iii); specifically we have that

$$\min_{c,\boldsymbol{v}} \mathbb{E}\left[\left(\pi(\boldsymbol{X}) - c - \sum_d \mu(d) v_d\right)^2\right] = \mathrm{Var}(\pi(\boldsymbol{X})).$$

□

Proposition 4, parts i)-ii), give conditions for avoiding demographic unfairness or proxy discrimination; in ii) limiting to the case of unawareness prices. Demographic fairness relates to the joint law of the pricing functional π and the response Y, hence the strong requirement P1 arises as a natural sufficient condition. Property P2 means that Y depends on D only via $\boldsymbol{X}$. Hence measurement of non-protected characteristics $\boldsymbol{X}$ eliminates any benefit to predictions from collecting protected characteristics D. This in turn implies P3, which means that the best-estimate prices are insensitive in D. Hence, if we restrict to unawareness prices, these conditions guarantee the absence of proxy discrimination.

Conversely, parts iii)-iv) of Proposition 4 give conditions for maximal levels of demographic unfairness and proxy discrimination. Here, properties P4-P6, in different ways, mean that knowing $\boldsymbol{X}$ in addition to D adds no new information useful for predicting claims Y.

As the properties P1-P6 are formulated with respect to the data generating process rather than arbitrary pricing functionals, they are strong and only give rise to sufficient conditions. The following example outlines a situation where properties P1-P3 are violated, but no demographic unfairness or proxy discrimination arises, demonstrating how the properties do not generally give necessary conditions. This illustrates also that, despite the insight given by understanding sufficient structural conditions, there is a need to quantify the materiality of discriminatory effects, for example via the metrics UF and PD discussed above.

Example 3. Let $\boldsymbol{X} \in \mathcal{X}$ and $D \in \{0, 1\}$, and assume the following

$$\begin{cases} \mathbb{P}(D = 0 \mid \boldsymbol{X} = \boldsymbol{x}) = \mathbb{P}(D = 0) & \text{if } \boldsymbol{x} \in \mathcal{A} \subset \mathcal{X}, \\ \mathbb{P}(D = 0 \mid \boldsymbol{X} = \boldsymbol{x}) \neq \mathbb{P}(D = 0) & \text{if } \boldsymbol{x} \in \mathcal{X} \setminus \mathcal{A}. \end{cases}$$

That is, $\boldsymbol{X} \not\perp\!\!\!\perp D$ under $\mathbb{P}$, and P1 from Definition 7 is violated. Further, assume that the best-estimate prices satisfy, for a non-trivial $\nu(\boldsymbol{x}, d)$

$$\begin{cases} \mu(\boldsymbol{x}, d) = \nu(\boldsymbol{x}, d) & \text{if } \boldsymbol{x} \in \mathcal{A} \subset \mathcal{X}, \\ \mu(\boldsymbol{x}, d) = 0 & \text{if } \boldsymbol{x} \in \mathcal{X} \setminus \mathcal{A} \end{cases},$$

which means that $\mathbb{E}[Y \mid \boldsymbol{X}, D] \neq \mathbb{E}[Y \mid \boldsymbol{X}]$, i.e., P2-P3 from Definition 7 are violated. Consequently, we may consider that $\boldsymbol{X}$ potentially acts as a proxy of D. Nonetheless, the materiality of this proxying effect is zero, if the unawareness price is used. To see this, consider the best-estimate and unawareness prices:

$$\begin{aligned} \mu(\boldsymbol{X}, D) = \mathbb{E}[Y \mid \boldsymbol{X}, D] &= \mathbb{1}_{\{\boldsymbol{X} \in \mathcal{A}\}} \nu(\boldsymbol{X}, D), \\ \mu(\boldsymbol{X}) &= \mathbb{1}_{\{\boldsymbol{X} \in \mathcal{A}\}} \nu(\boldsymbol{X}, 0) \mathbb{P}(D = 0 \mid \boldsymbol{X}) + \mathbb{1}_{\{\boldsymbol{X} \in \mathcal{A}\}} \nu(\boldsymbol{X}, 1) \mathbb{P}(D = 1 \mid \boldsymbol{X}) \\ &= \mu(\boldsymbol{X}, 0) \mathbb{P}(D = 0) + \mu(\boldsymbol{X}, 1) \mathbb{P}(D = 1). \end{aligned}$$

Hence, by Definition 3, $\mu(\boldsymbol{X})$ avoids proxy discrimination and we have $\mathrm{PD}(\mu) = 0$. Furthermore, notwithstanding the dependence of $(\boldsymbol{X}, D)$, the unawareness price is also demographically fair, given that:

$$\mathbb{E}[\mu(\boldsymbol{X}) \mid D = d] = \int_{\boldsymbol{x} \in \mathcal{A}} \mu(\boldsymbol{x}) \mathrm{d}\mathbb{P}(\boldsymbol{x} \mid d) = \int_{\boldsymbol{x} \in \mathcal{A}} \mu(\boldsymbol{x}) \mathrm{d}\mathbb{P}(\boldsymbol{x}),$$

which does not depend on the value of d and thus $\mathrm{UF}(\mu) = 0$.

■

We now turn our attention to the way that structural properties of the data generating process impact on the sensitivity of the PD measure to covariate sub-vectors $\boldsymbol{X}_{\mathcal{S}}$. In the following result we deal with the common case of unawareness prices.

Proposition 5. *i) Assume that $\boldsymbol{X}_{\mathcal{S}^c} \perp\!\!\!\perp (Y, D) \mid \boldsymbol{X}_{\mathcal{S}}$. Then, for the unawareness price $\pi(\boldsymbol{X}) = \mu(\boldsymbol{X})$ the following hold.*

a) $\mathrm{PD}_{\mathcal{S}}(\mu) = \mathrm{PD}(\mu)$ and $\widetilde{\mathrm{PD}}_{\mathcal{S}^c}(\mu) = 0$.

b) If additionally $\boldsymbol{X}_{\mathcal{S}^c} \perp\!\!\!\perp \boldsymbol{X}_{\mathcal{S}}$, then it it also holds that $\widetilde{\mathrm{PD}}_{\mathcal{S}}(\mu) = \mathrm{PD}(\mu)$ and $\mathrm{PD}_{\mathcal{S}^c}(\mu) = 0$.

ii) Let the best-estimate price take the form $\mu(\boldsymbol{X}, D) = g(\boldsymbol{X}) + h(D)$ and assume that $\mathrm{Cov}(g(\boldsymbol{X}), h(D)) \geq 0$ and $\boldsymbol{X}_{\mathcal{S}^c} \perp\!\!\!\perp D \mid \boldsymbol{X}_{\mathcal{S}}$. Then, for the unawareness price $\pi(\boldsymbol{X}) = \mu(\boldsymbol{X})$ the following hold.

a) $\mathrm{PD}_{\mathcal{S}}(\mu) = \mathrm{PD}(\mu)$ and $\widetilde{\mathrm{PD}}_{\mathcal{S}^c}(\mu) = 0$.

b) If additionally $\boldsymbol{X}_{\mathcal{S}^c} \perp\!\!\!\perp \boldsymbol{X}_{\mathcal{S}}$, then it also holds that $\widetilde{\mathrm{PD}}_{\mathcal{S}}(\mu) = \mathrm{PD}(\mu)$ and $\mathrm{PD}_{\mathcal{S}^c}(\mu) = 0$.

Proof. i) a) By the conditional independence assumption we have that:

$$\begin{aligned}\mathbb{P}(D = d \mid \boldsymbol{X}) &= \mathbb{P}(D = d \mid \boldsymbol{X}_{\mathcal{S}}),\\ \mu(\boldsymbol{X}, D) = \mathbb{E}[Y \mid \boldsymbol{X}, D] &= \mathbb{E}[Y \mid \boldsymbol{X}_{\mathcal{S}}, D] =: \mu(\boldsymbol{X}_{\mathcal{S}}, D).\end{aligned} \tag{14}$$

Consequently, noting (3), we have that

$$\begin{aligned}\mu(\boldsymbol{X}) &= \sum_{d \in \mathfrak{D}} \mu(\boldsymbol{X}_{\mathcal{S}}, d) \mathbb{P}(D = d \mid \boldsymbol{X}_{\mathcal{S}}) \implies\\ \Lambda(\mu, \boldsymbol{X}) &= \sum_{d \in \mathfrak{D}} \mu(\boldsymbol{X}_{\mathcal{S}}, d) \left(\mathbb{P}(D = d \mid \boldsymbol{X}_{\mathcal{S}}) - v_d^*\right) - c^*.\end{aligned}$$

As the last expression is $\boldsymbol{X}_{\mathcal{S}}$-measurable, it holds that $\mathbb{E}\left[\Lambda(\mu, \boldsymbol{X}) \mid \boldsymbol{X}_{\mathcal{S}}\right] = \Lambda(\mu, \boldsymbol{X})$, from which it follows that,

$$\begin{aligned}\mathrm{PD}_{\mathcal{S}}(\mu) &= \frac{\mathrm{Var}(\Lambda(\mu, \boldsymbol{X}))}{\mathrm{Var}(\pi(\boldsymbol{X}))} = \mathrm{PD}(\mu),\\ \widetilde{\mathrm{PD}}_{\mathcal{S}^c}(\pi) &= \frac{\mathrm{Var}(\Lambda(\pi, \boldsymbol{X})) - \mathrm{Var}(\mathbb{E}[\Lambda(\pi, \boldsymbol{X}) \mid \boldsymbol{X}_{\mathcal{S}}])}{\mathrm{Var}(\pi(\boldsymbol{X}))} = 0.\end{aligned}$$

b) Here it is sufficient to show that $\mathbb{E}[\Lambda(\mu, \boldsymbol{X}) \mid \boldsymbol{X}_{\mathcal{S}^c}]$ is a constant. It has already been shown that $\Lambda(\mu, \boldsymbol{X})$ is $\boldsymbol{X}_{\mathcal{S}}$-measurable; hence the result follows from $\boldsymbol{X}_{\mathcal{S}^c} \perp\!\!\!\perp \boldsymbol{X}_{\mathcal{S}}$.

ii) a) The unawareness price will take the form

$$\mu(\boldsymbol{X}) = g(\boldsymbol{X}) + \mathbb{E}[h(D) \mid \boldsymbol{X}].$$

We consider the form of the quantity to be minimised in the numerator of the PD measure. From the specific form of the best-estimate and unawareness prices we have:

$$\begin{aligned}&\mu(\boldsymbol{X}) - \mathbb{E}[\mu(\boldsymbol{X})] - \sum_{d \in \mathfrak{D}} v_d \left(\mu(\boldsymbol{X}, d) - \mathbb{E}[\mu(\boldsymbol{X}, d)]\right)\\ &= \left(1 - \sum_{d \in \mathfrak{D}} v_d\right) (g(\boldsymbol{X}) - \mathbb{E}[g(\boldsymbol{X})]) + \mathbb{E}[h(D) \mid \boldsymbol{X}] - \mathbb{E}[h(D)].\end{aligned}$$

Consequently, we can just let $v := \sum_{d\in\mathfrak{D}} v_d \in [0,1]$ and optimise over that. By checking the Karush-Kuhn-Tucker conditions (calculations not documented here, we find that the condition $\mathrm{Cov}(g(\boldsymbol{X}), h(D)) \geq 0$ is necessary and sufficient for $v = 1$. Thus we obtain

$$\begin{aligned}\Lambda(\mu, \boldsymbol{X}) &= \mathbb{E}[h(D) \mid \boldsymbol{X}] - \mathbb{E}[h(D)] \\ &= \mathbb{E}[h(D) \mid \boldsymbol{X}_{\mathcal{S}}] - \mathbb{E}[h(D)],\end{aligned}$$

where we used the additional assumption $\boldsymbol{X}_{\mathcal{S}^c} \perp\!\!\!\perp D \mid \boldsymbol{X}_{\mathcal{S}}$ in the second equation. Furthermore,

$$\mathbb{E}[\Lambda(\mu, \boldsymbol{X}) \mid \boldsymbol{X}_{\mathcal{S}}] = \Lambda(\mu, \boldsymbol{X}),$$

from which the stated result follows.

b) Again, it is sufficient to show that $\mathbb{E}[\Lambda(\mu, \boldsymbol{X}) \mid \boldsymbol{X}_{\mathcal{S}^c}]$ is a constant, which follows from the additional independence assumption.

□

Proposition 5 provides dependence scenarios, under which the contributions of sub-vectors of $\boldsymbol{X}$ are either zero or equal to the total level of proxy discrimination $\mathrm{PD}(\mu)$. These dependence scenarios are thus different conceptualisations of full relevance of $\boldsymbol{X}_{\mathcal{S}}$ and irrelevance of $\boldsymbol{X}_{\mathcal{S}^c}$ with respect to the alternative metrics $\mathrm{PD}_{\mathcal{S}}$ and $\widetilde{\mathrm{PD}}_{\mathcal{S}}$. Specifically, the conditional independence of parts i)a) and ii)a) means that knowing $\boldsymbol{X}_{\mathcal{S}}$ makes any information on $\boldsymbol{X}_{\mathcal{S}^c}$ redundant for risk predictions. The independence statement of parts ib) and ii)b) goes further, reducing even more the relevance of $\boldsymbol{X}_{\mathcal{S}^c}$. The strength of this requirement is consistent with the interpretability of high values of PD and low values $\widetilde{\mathrm{PD}}$, rather than the converse. Finally, we note that the key difference between parts i) and ii) is that the additional structure imposed in part ii) makes the properties of the attributions of the PD metric functions only of the dependence structure for $(\boldsymbol{X}, D)$, with no reference to Y. The simplifying assumptions here are lack of interactions between $\boldsymbol{X}$ and D when predicting Y and a positive correlation of the portions of claim costs arising respectively from $\boldsymbol{X}$ and D.

The properties discussed in Definition 7 and Proposition 5 describe features of the data generating process for a particular insurance portfolio. While these are affected by marketing and underwriting decisions, they typically remain out of the control of, e.g., a pricing actuary. Hence, any adjustments carried out to avoid demographic unfairness or proxy discrimination need to focus on the design of pricing functionals, in order to comply with Definitions 1 and/or 3.

At the same time, any such price adjustment allows the quantification of the difference of a price-in-use and the adjusted price for a particular policyholder profile. Thus, one can form easily *local* (policyholder-specific) measures of demographic unfairness and proxy discrimination in contrast to the *global* (portfolio-wide) measures introduced above. We develop these ideas in the rest of this section.

3.2 Local measurement of demographic unfairness

A standard construction of demographically fair prices has been via optimal transport (OT) methods (Gordaliza et al. 2019, Chiappa et al. 2020). In an insurance context, such methods aim at inducing independence between prices $\pi(\boldsymbol{X})$ and the protected characteristics D by suitable variable transformations; for insurance-specific investigations see Lindholm et al. (2024), Charpentier et al. (2023). The broader mathematical problem of approximating a random variable with another, subject to an independence constraint, is treated by Delbaen & Majumdar (2024). There are two key approaches found in the literature. *Input OT* aims at transforming (pre-processing) the covariates $\boldsymbol{X}$ so that it is independent of D, thus satisfying property P1 from Definition 7. Then any functional of the transformed covariates will satisfy demographic parity. *Output OT* refers to the transformation (post-processing) of prices themselves in order to achieve independence from D; thus this procedure targets Definition 1 and, as it is specific to a given pricing functional, it is in a sense a weaker intervention compared to Input OT. For a comparison of the two approaches in the context of insurance pricing and a discussion of their respective interpretability, see Lindholm et al. (2024).

Here, to create a demographically fair benchmark price, we follow an Output OT approach. For the pricing functional π, denote the conditional distributions of the corresponding prices by

$$G_d(m) = \mathbb{P}(\pi(\boldsymbol{X}) \leq m \mid D = d), \quad d \in \mathfrak{D},$$

and assume for simplicity that they are continuous. Then, for any continuous distribution G, we may construct the prices

$$\tilde{\pi}(\boldsymbol{X}, D) = \sum_{d \in \mathfrak{D}} \mathbb{1}_{\{D=d\}} G^{-1} \circ G_d\big(\pi(\boldsymbol{X})\big). \tag{15}$$

The price $\tilde{\pi}(\boldsymbol{X}, D)$ satisfies

$$\mathbb{P}(\tilde{\pi}(\boldsymbol{X}, D) \leq m) = G(m),$$
$$\mathbb{P}(\tilde{\pi}(\boldsymbol{X}, D) \leq m \mid D = d) = G(m), \quad \text{for all } d \in \mathfrak{D},$$

such that $D \perp\!\!\!\perp \tilde{\pi}(\boldsymbol{X}, D) \sim G$; the construction (15) works by making the conditional distribution of prices the same on each demographic subgroup $D = d$. Note that the transformed price $\tilde{\pi}(\boldsymbol{X}, D)$ explicitly depends on D – even as $\pi(\boldsymbol{X})$ does not. This is a form of direct discrimination arising in the process of engineering demographic parity; see Lindholm et al. (2024) for more discussion of this point.

Finally, to construct a demographically fair benchmark, we need to select the target distribution G. A standard choice is given by Chzhen et al. (2020)

$$G^{-1}(u) = \sum_{d' \in \mathfrak{D}} \mathbb{P}\left(D = d'\right) G_{d'}^{-1}(u), \tag{16}$$

From now on we will consistently refer to the *Output OT price* as the construction $\tilde{\pi}(\boldsymbol{X}, D)$ from (15) and (16).

We can now proceed with the definition of a local measure of demographic unfairness.

Definition 8. *Consider a pricing functional π and the Output OT price $\tilde{\pi}(\boldsymbol{X}, D)$. Then, for the policyholder with profile $\boldsymbol{X} = \boldsymbol{x}$, $D = d$, the local measure of demographic unfairness is defined as:*

$$\delta_{\mathrm{UF}}(\boldsymbol{x}, d; \pi) = \pi(\boldsymbol{x}) - \tilde{\pi}(\boldsymbol{x}, d). \tag{17}$$

If $\pi(\boldsymbol{X}) = \mu(\boldsymbol{X})$ is the unawareness price, we just write $\delta_{\mathrm{UF}}(\boldsymbol{x}, d)$.

A value of $\delta_{\mathrm{UF}}(\boldsymbol{x}, d; \pi) > 0$ implies that policyholders with attributes $\boldsymbol{X} = \boldsymbol{x}$, $D = d$ suffer from demographic unfairness in the sense that the price they are charged is higher than the corresponding benchmark demographically fair price. Clearly, if $\boldsymbol{X}$ and D are already independent, the prices $\pi(\boldsymbol{X})$ and $\tilde{\pi}(\boldsymbol{X}, D)$ coincide such that the measure becomes zero. This is stated formally below.

Proposition 6. *If P1 in Definition 7 holds, then* $\delta_{\mathrm{UF}}(\boldsymbol{x}, d) = 0$.

Furthermore, it is of interest to establish conditions for the sign of the metric $\delta_{\mathrm{UF}}(\boldsymbol{x}, d; \pi)$. In the simple but common case of a binary D, this is straightforward. Denote by $\preceq_{\mathrm{st}}$ precedence in the usual stochastic order, such that for two distributions F, G, we have that $F \preceq_{\mathrm{st}} G \iff F(x) \geq G(x)$ for all x; this is a strong condition not allowing the crossing of distributions.

Proposition 7. *Let* $\mathfrak{D} = \{0, 1\}$ *and* $G_0 \preceq_{\mathrm{st}} G_1$.

i) $\delta_{\mathrm{UF}}(\boldsymbol{x}, 0; \pi) \leq 0$ *and* $\delta_{\mathrm{UF}}(\boldsymbol{x}, 1; \pi) \geq 0$.

ii) For $\boldsymbol{x}$ *such that* $\mathbb{P}(D = 0 \mid \boldsymbol{X} = \boldsymbol{x}) > 0$, *it holds that:*

$$\mathbb{E}[\delta_{\mathrm{UF}}(\boldsymbol{X}, D; \pi) \mid \boldsymbol{X} = \boldsymbol{x}] > 0 \iff$$
$$\frac{\mathbb{P}(D = 1 \mid \boldsymbol{X} = \boldsymbol{x})}{\mathbb{P}(D = 0 \mid \boldsymbol{X} = \boldsymbol{x})} > \frac{-\delta_{\mathrm{UF}}(\boldsymbol{x}, 0; \pi)}{\delta_{\mathrm{UF}}(\boldsymbol{x}, 1; \pi)} = \frac{G^{-1} \circ G_0(\pi(\boldsymbol{x})) - \pi(\boldsymbol{x})}{\pi(\boldsymbol{x}) - G^{-1} \circ G_1(\pi(\boldsymbol{x}))} \geq 0,$$

where G *is given by* (16).

Proof.

i) The statement follows by noting that construction (16) implies $G_0 \preceq_{\mathrm{st}} G \preceq_{\mathrm{st}} G_1$ and consequently

$$\delta_{\mathrm{UF}}(\boldsymbol{x}, 0; \pi) = \pi(\boldsymbol{x}) - G^{-1} \circ G_0(\pi(\boldsymbol{x})) \leq 0,$$
$$\delta_{\mathrm{UF}}(\boldsymbol{x}, 1; \pi) = \pi(\boldsymbol{x}) - G^{-1} \circ G_1(\pi(\boldsymbol{x})) \geq 0.$$

ii) We have that

$$\mathbb{E}[\delta_{\mathrm{UF}}(\boldsymbol{X}, D; \pi) \mid \boldsymbol{X} = \boldsymbol{x}] =$$
$$= \mathbb{P}(D = 0 \mid \boldsymbol{X} = \boldsymbol{x})\delta_{\mathrm{UF}}(\boldsymbol{x}, 0; \pi) + \mathbb{P}(D = 1 \mid \boldsymbol{X} = \boldsymbol{x})\delta_{\mathrm{UF}}(\boldsymbol{x}, 1; \pi),$$

with the stated result following directly from the inequalities of part i).

□

To interpret part i) of Proposition 7, first note that $G_0 \preceq_{\mathrm{st}} G_1$ means that the policyholders with protected attribute $D = 1$ tend to be considered as higher risk, according to best-estimate prices. So these are the policyholders for whom the use of the Output OT price (15) should confer a discount, compared with the unawareness price. For part ii), we consider a situation where for a policyholder with $\boldsymbol{X} = \boldsymbol{x}$ the value of D may not be known. The left-hand side of the stated condition implies that, on average, the local measure of unfairness will be positive when $\mathbb{P}(D=1 \mid \boldsymbol{X}=\boldsymbol{x})/\mathbb{P}(D=0 \mid \boldsymbol{X}=\boldsymbol{x})$ is high, such that there is a high chance that, given $\boldsymbol{X}=\boldsymbol{x}$, the policyholder belongs to the demographically disadvantaged group $D=1$. Furthermore, the condition is more likely to be satisfied when the ratio $-\delta_{\mathrm{UF}}(\boldsymbol{x},0;\pi)/\delta_{\mathrm{UF}}(\boldsymbol{x},1;\pi)$ is low. That fraction becomes small if the comparative disadvantage for group $D=1$ (denominator) becomes much higher than the comparative advantage of group $D=0$ (numerator), given the information $\boldsymbol{X}=\boldsymbol{x}$.

These ideas are illustrated in the following example.

Example 4. We continue from Example 2, where $\mu(X,D)=\frac{1}{2}+X+D$ and $X \sim \mathrm{U}(0,1)$. Note that $\mu(x,1)>\mu(x,0)$ for all x. However, we now allow a variety of positive and negative dependence relations between (X,D) by assuming that

$$\mathbb{P}(D=1 \mid X=x)=\frac{1-a}{2}+ax, \quad a \in (-1,1].$$

By setting $a=1$, we recover the exact setting of Example 2; $0<a<1$ gives a weaker positive dependence, while $-1<a<0$ gives a negative dependence. With this modification the unawareness price changes to $\mu(X)=\frac{2-a}{2}+(a+1)X$.

We now evaluate the local measure of unfairness (17) for the unawareness price $\mu(X)$ and various levels of a. The calculations are simple but tedious and are not reported here. In Figure 1 we plot the functions $\delta_{\mathrm{UF}}(x,d)$ (blue for $d=0$, red for $d=1$), as well as their conditional mean $\mathbb{E}[\delta_{\mathrm{UF}}(X,D) \mid X=x]$ (black) for $a=0.75$ (positive dependence) and $a=-0.75$ (negative dependence). While the shapes appear similar, there are two observations. First, for $a=0.75$, the red line is above zero and the blue line below, showing that policyholders with $D=1$ are adversely affected by demographic unfairness, while policyholders with $D=0$ are benefiting. This pattern is reversed when the dependence of (X,D) becomes negative ($a=-0.75$). Second, the two plots are at very different scales, with the absolute value of local demographic unfairness being an order of magnitude higher in the case of positive dependence. The reason is that, for $a=0.75$, the positive dependence of (X,D) works in the same direction as the impact of each of those two variables on claims costs. However, when dependence is negative, then the unawareness price becomes less sensitive to X (in the extreme $a \to -1$ leads to a constant $\mu(X)$). As a result, for negative dependence much smaller disparities between demographic groups emerge. This point is reinforced through Figure 2, where we plot $\mathbb{E}[\delta_{\mathrm{UF}}(X,D) \mid X=x]$ against different values of the dependence parameter a.

■

In Example 4, we had a single non-protected covariate $\boldsymbol{X}=X$, but, in general, $\delta_{\mathrm{UF}}(\boldsymbol{x},d)$ will be a multivariate function. Then the question arises as to how individual covariates contribute to this metric. This can be done by using standard local model explainability methods, e.g., by, analogously to (13), calculating Shapley values with respect to the alternative value functional $\mathcal{S} \mapsto \mathbb{E}[\delta_{\mathrm{UF}}(\boldsymbol{X},D) \mid \boldsymbol{X}_{\mathcal{S}}=\boldsymbol{x}_{\mathcal{S}}]$, where $\boldsymbol{x}_{\mathcal{S}}$ is a sub-vector of $\boldsymbol{x}$ for the specific instance of interest (Lundberg & Lee 2017, Aas et al. 2021). The same comments apply to the local metric for proxy discrimination introduced in the next section.

3.3 Local measurement of proxy discrimination

Analogously to the last section, we propose a local measure of proxy discrimination. Again, we need for that purpose a benchmark price that avoids proxy discrimination. Any such

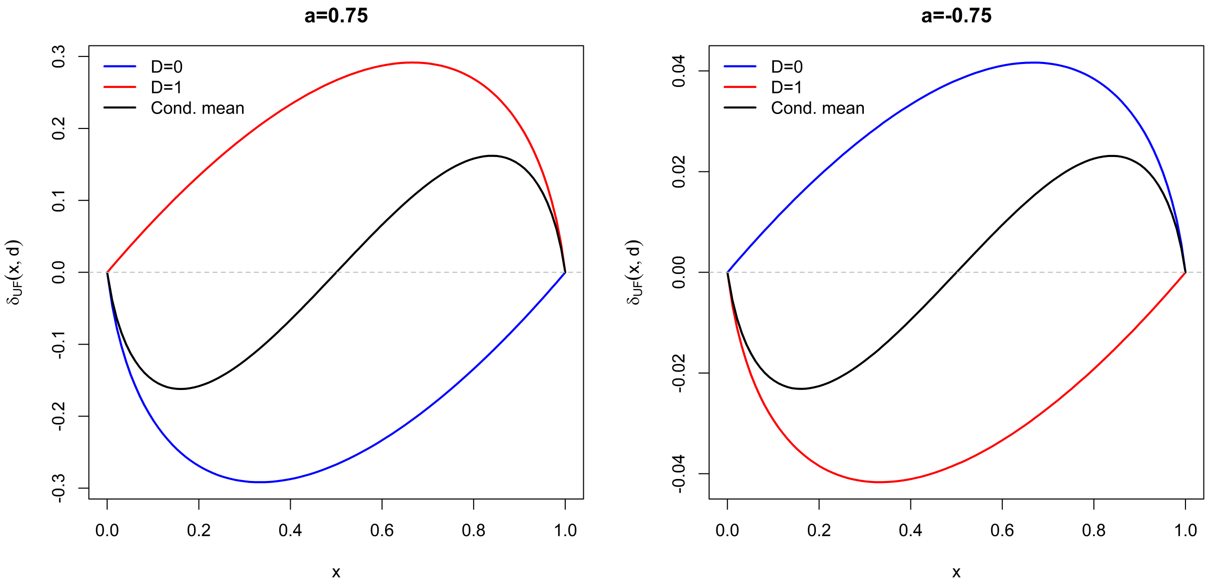

Figure 1: Local unfairness metric $\delta_{\mathrm{UF}}(x,d)$ (blue for $d=0$, red for $d=1$) and conditional mean $\mathbb{E}[\delta_{\mathrm{UF}}(X,D) \mid X=x]$ (black) for $a=0.75$ (left) and $a=-0.75$ (right).

price takes the form (7); nonetheless, to produce a benchmark one needs to choose the values of $c, v_d,\ d \in \mathfrak{D}$. Here we choose values that minimise the numerator in (8), such that the benchmark is the price closest to the original price $\pi(\boldsymbol{X})$, which is free from proxy discrimination. (Note that while the optimal values c^*, v_d^* may not be unique, the resulting approximation is.) The difference between a price and its closest proxy-discrimination-free approximation has already been defined in Section 2.5, as the regression residual $\Lambda(\boldsymbol{X},\pi)$. Hence we re-purpose this quantity as a local measure of proxy discrimination.

Definition 9. *Consider a pricing functional π and $\Lambda(\boldsymbol{X},\pi)$ in equation (10). Then, for the policyholder with profile $\boldsymbol{X}=\boldsymbol{x}$, the local measure of proxy discrimination is defined as:*

$$\delta_{\mathrm{PD}}(\boldsymbol{x};\pi) = \Lambda(\boldsymbol{x},\pi). \tag{18}$$

If $\pi(\boldsymbol{X}) = \mu(\boldsymbol{X})$ is the unawareness price, we just write $\delta_{\mathrm{PD}}(\boldsymbol{x})$.

We state Proposition 8 below without proof – for the case of each property it is easy to show that the unawareness price is already free from proxy discrimination, such that $\Lambda(\boldsymbol{X},\pi)$ is identically zero.

Proposition 8. *If any one of P1, P2 or P3 in Definition 7 holds, then $\delta_{\mathrm{PD}}(\boldsymbol{x}) = 0$.*

We conclude this section with a continuation of our running example.

Example 5. We continue from Example 4, considering the evaluation of our local measure of proxy discrimination. We make a qualitative argument, omitting a formal proof. Recall the forms of the best-estimate and unawareness prices, $\mu(X,D) = 1/2 + X + D$ and $\mu(X) =$

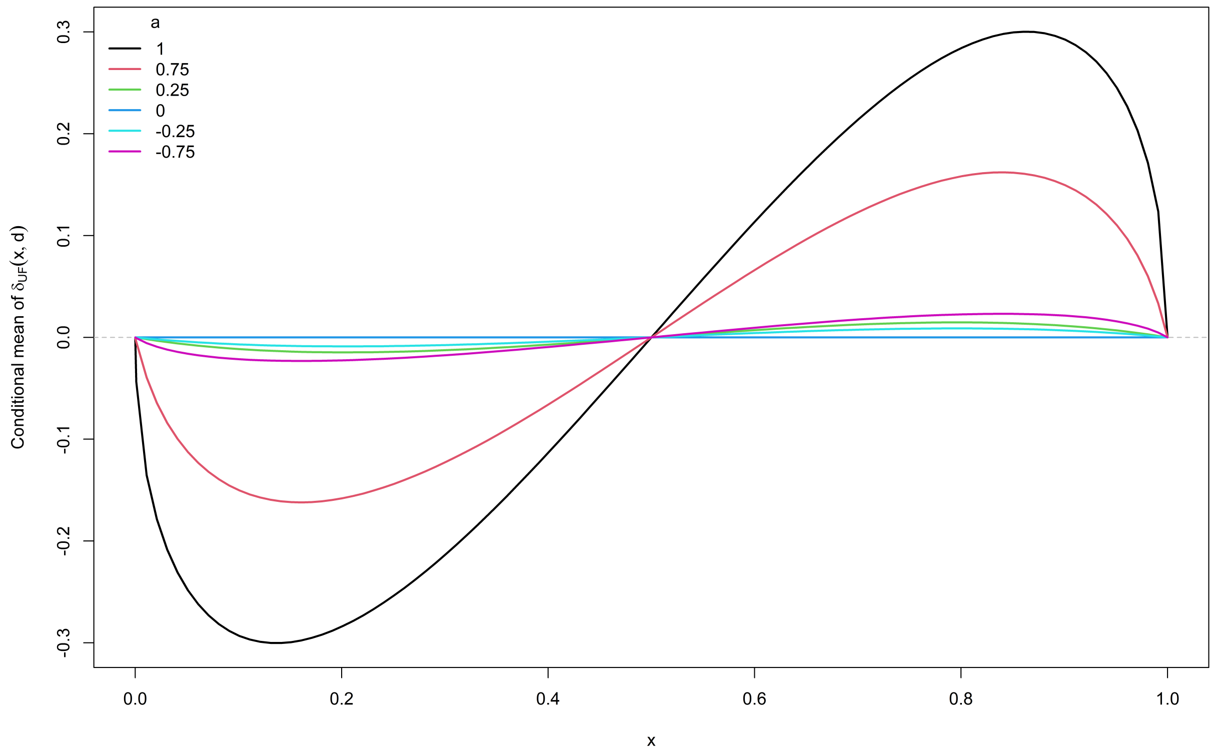

Figure 2: Conditional mean of the local unfairness metric $\mathbb{E}[\delta_{\mathrm{UF}}(X, D) \mid X = x]$ for the unawareness price and a range of dependence parameters a.

$1 - a/2 + (1 + a)X$ respectively. If $-1 < a \leq 0$, the slope of the unawareness price in X is less or equal to that of the best-estimate price. As a result $\Lambda(X, \mu) = 0$ and there is no proxy discrimination. However, when $0 < \alpha \leq 1$, proxy discrimination arises. The closest approximation to $\mu(X)$ that avoids proxy discrimination is the price $\mu^*(X) = 1 + X$, which reduces the slope in X to 1. Consequently we have

$$\delta_{\mathrm{PD}}(x) = \begin{cases} 0, & -1 \leq a \leq 0 \\ -\frac{a}{2} + ax, & 0 < a \leq 1. \end{cases}$$

In Figure 3 we show the function $\delta_{\mathrm{PD}}(x)$ for different (non-negative) values of the dependence parameter a. Because of the positive dependence, policyholders with $x > 0.5$ are implicitly inferred to have protected attribute $D = 1$ and hence are disadvantaged in the sense of proxy discrimination; the reverse happens for $x < 0.5$.

Finally, we consider the extent to which adopting a price that avoids proxy discrimination also has beneficial impacts in reducing demographic unfairness. For this, we calculate, alongside $\mathbb{E}[\delta_{\mathrm{UF}}(X, D) \mid X = x]$, the quantity $\mathbb{E}[\delta_{\mathrm{UF}}(X, D; \pi) \mid X = x]$, where $\pi(X) = \mu(X)$ for $-1 \leq a \leq 0$ and $\pi(X) = \mu^*(X)$ for $0 < a \leq 1$. We plot those metrics in Figure 4, as functions of the dependence parameter a. Naturally, for $a \leq 0$ the two plots are identical. For $a > 0$, the discrimination-free price offers a modest improvement, though demographic unfairness persists. This is expected, since the issues of demographic unfairness and proxy discrimination are quite separate and there is no simple way of addressing both at the same time; for a detailed discussion see Lindholm et al. (2024).

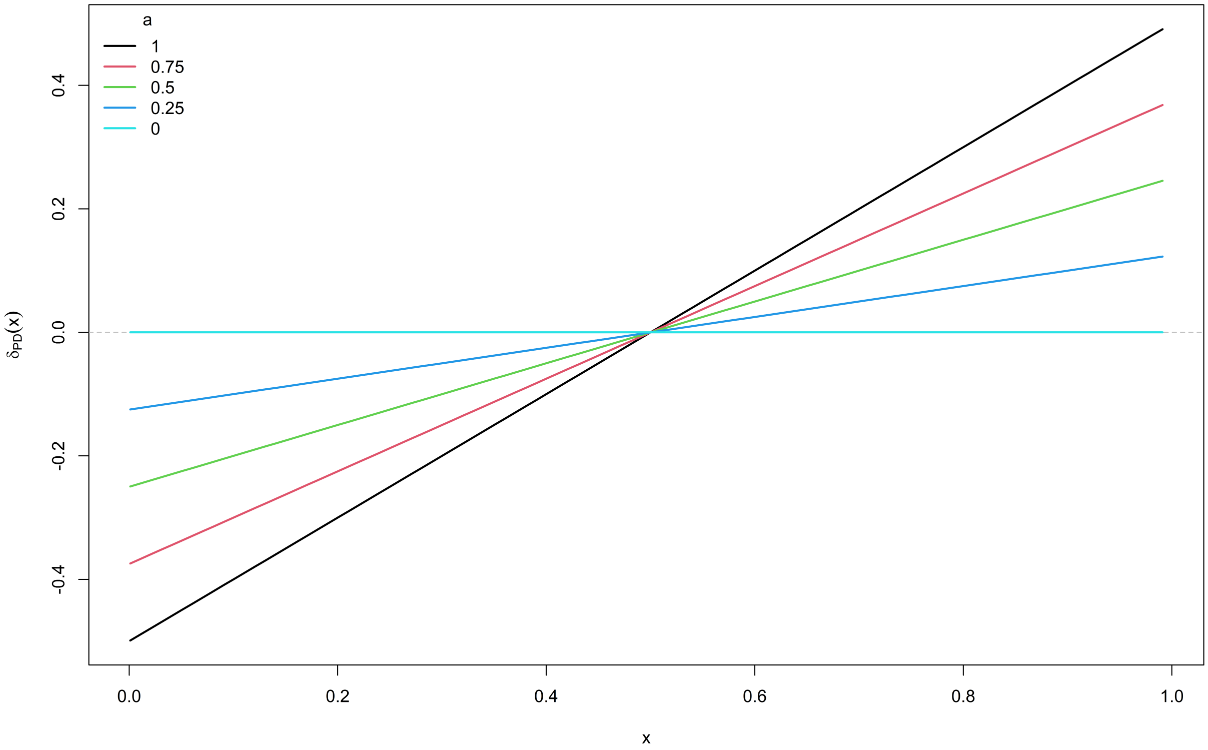

Figure 3: Local proxy discrimination metric δ_{PD} for the unawareness price and a range of dependence parameters a.

4 Discussion and extensions

We propose measures of demographic unfairness and proxy discrimination with a focus on insurance pricing. The measure of demographic unfairness is already present in the literature (Bénesse et al. 2022), while the measure of proxy discrimination is new. For that measure, we also propose methods for attributing any proxying effects to different covariates. These measures are global, in the sense that they quantify unfairness or discrimination across a portfolio. In addition to studying the properties of these measures, we develop related local measures, which allow quantification of demographic unfairness and proxy discrimination at the granular policy level.

In the rest of this section we discuss limitations of our approach, possible extensions, and wider issues that we consider relevant and can form directions for future research.

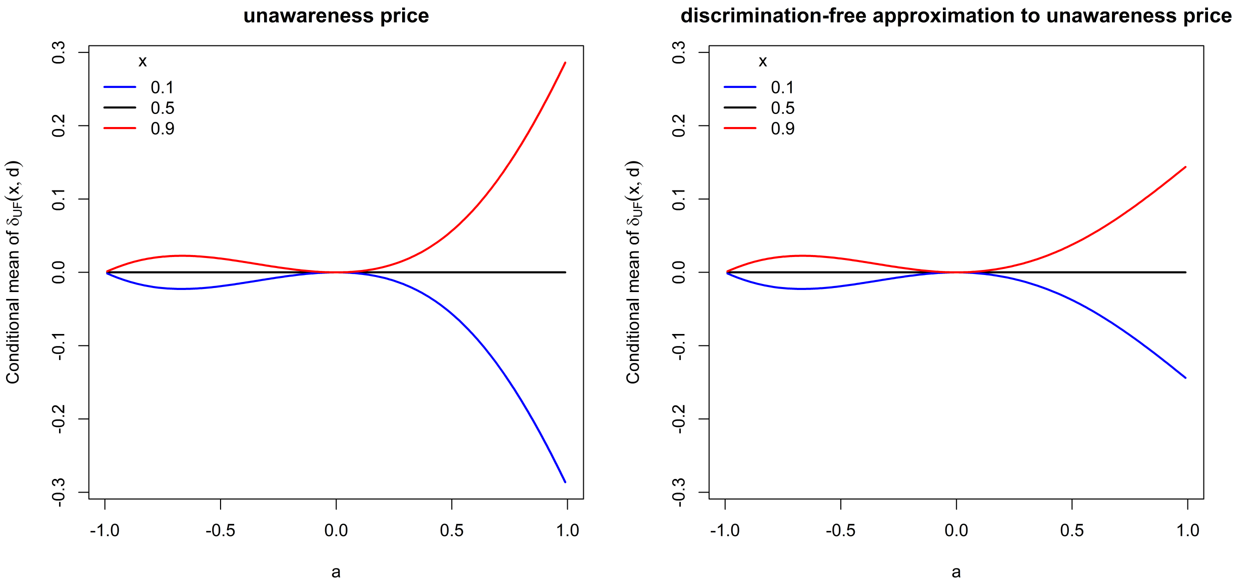

Figure 4: Conditional mean of local unfairness metric $\mathbb{E}[\delta_{\mathrm{UF}}(X, D) \mid X = x]$ for the unawareness price (left) and its closest approximation that avoids proxy discrimination (right), plotted against dependence parameter a.

4.1 Conceptualisation of proxy discrimination

Our aim in this paper was to provide a measure of proxy discrimination that is useful in practice, in that it allows quantification from a set of observable insurance prices. This is important for the empirical assessment and auditing of proxy discrimination in insurance portfolios. We note that the design of metrics that can be evaluated on the joint distribution of observed quantities, e.g., $(Y, D, \pi(\boldsymbol{X}))$, is generally well suited for group fairness notions. These notions revolve around the *outcomes* of a pricing strategy. However, as argued in Lindholm et al. (2024), proxy discrimination can be better seen as an individual fairness concern. In particular, proxy discrimination relates to the *mechanism* by which prices are calculated, rather than the outcome of the process. We try to bridge this gap by the construction (8), which makes the metric $\mathrm{PD}(\pi)$ a function of the joint distribution of $(\pi(\boldsymbol{X}), \mu(\boldsymbol{X}, d),\ d \in \mathfrak{D})$, thus considering the way that prices relate to predictions for different values of the protected attribute D.

While we believe that through this choice our primary aim has been achieved, tensions remain. First, Definition 4 used here is different to the definition of proxy discrimination in Lindholm et al. (2024). The two definitions make requirements of different type. The definition of proxy discrimination in Lindholm et al. (2024) as an individual fairness property requires that prices are calculated in a way that is unaffected by the conditional probability $\mathbb{P}(D \mid \boldsymbol{X})$, which is precisely the statistical manifestation of proxying. This definition assumes an implicit commitment of insurers on how they would be pricing the same risk within the context of alternative portfolios with different statistical dependence structures. This allows more flexibility in the choice of the pricing functional itself – but it is also harder to monitor. On the other hand, Definition 4 does not require any such commitment and is instead specific to a particular portfolio. The *quid pro quo* is then less flexibility as to how a discrimination-free pricing functional can be selected, namely, as a constrained weighted average of $\mu(\boldsymbol{X}, d),\ d \in \mathfrak{D}$, and a constant.

The differences between the notion of proxy discrimination in this paper and in Lindholm et al. (2024) are also reflected in the role of the object $\pi^*(\boldsymbol{X}) = c^* + \sum_{d\in\mathfrak{D}} \mu(\boldsymbol{X}, d)v_d^*$, the closest element to $\pi(\boldsymbol{X})$ within the set of prices that are free from proxy discrimination. Using $\pi^*(\boldsymbol{X})$ as a corrected price appears problematic, since a solution $(c^*, \boldsymbol{v}^*)$ of the regression problem (8), will generally depend on the joint distribution of $(\boldsymbol{X}, D)$, which, from the perspective of Lindholm et al. (2024), would make any price depending on those optimised coefficients not permissible. At the same time we note the following: by the existence of the offset c^* we have that $\mathbb{E}[\pi^*(\boldsymbol{X})] = \mathbb{E}[\pi(\boldsymbol{X})]$. Consequently, if the pricing functional used is unbiased (e.g., the unawareness price $\mu(\boldsymbol{X})$ is used), then it will also hold that $\mathbb{E}[\pi^*(\boldsymbol{X})] = \mathbb{E}[Y]$. Unbiasedness is fundamental property of technical insurance prices, since its violation implies that portfolios are not mean-self-financing. Note that the discrimination-free price $h^*(\boldsymbol{X})$ in (5) is generally not unbiased. Furthermore, any bias correction method for discrimination-free prices, including the methods discussed in Lindholm et al. (2022), would suffer from the same limitation of sensitivity in the joint distribution of $(\boldsymbol{X}, D)$.

Finally, we have not considered the situation where there is *direct discrimination*, i.e. where the prices are also functions of the protected characteristic D. This is because we tacitly assumed that such prices are not permitted due to legal or regulatory constraints. Nonetheless, the measures (4) and (8) can also be evaluated with respect to broader classes of prices, by substituting for $\pi(\boldsymbol{X})$ some random variable Π that represents prices that are not necessarily $\boldsymbol{X}$-measurable. For example, it is clear that the variability of such a price Π cannot be explained by regressing on $\mu(\boldsymbol{X}, d),\ d \in \mathfrak{D}$, such that in the presence of direct discrimination it will generally be $\mathrm{PD}(\Pi) > 0$. This also implies that when observing a high value of PD, the metric cannot tell us whether direct discrimination is part of the reason. In situations where direct discrimination is the focal issue, one can follow the suggestion of Bénesse et al. (2022) and use the *total sensitivity* $1 - \mathrm{Var}(\mathbb{E}[\Pi \mid \boldsymbol{X}])/\mathrm{Var}(\Pi)$; note that if Π is $\boldsymbol{X}$-measurable that metric becomes equal to zero.

4.2 Dependence under a market distribution

Demographic unfairness is easily understood and communicated to all stakeholders. At the same time, the requirement to satisfy this property is very strong and can become problematic. Part of the reason for that is that the dependence between $\pi(\boldsymbol{X})$ and D may not reflect causal relations but be an artifact of portfolio composition, which makes company data not representative of the wider population (Mehrabi et al. 2021, Côté et al. 2024). Lindholm et al. (2024) discuss how, consequently, modifying prices to avoid demographic unfairness will generally lead to different adjustments for different portfolios, creating inconsistencies and potential market distortions.

An alternative to the standard definition of demographic parity is to ask for independence of $\pi(\boldsymbol{X})$ and D across a whole market, rather than any one particular portfolio. This requirement would ensure that across the broader population of policyholders, no demographic group is systematically disadvantaged. However, at the level of an individual portfolio, it may still be the case that, on average, different prices are charged for different demographic groups.

To formalise this idea, consider $\bar{\mathbb{P}}(\boldsymbol{X}, D)$ the joint distribution of $(\boldsymbol{X}, D)$ across such a broad population, describing the statistical relationship between policyholder characteristics across an insurance market. Then we can modify Definition 2, by considering a demographic unfairness measure with respect to a population distribution $\bar{\mathbb{P}}(\boldsymbol{X}, D)$, defined as

$$\mathrm{UF}(\pi, \bar{\mathbb{P}}) := \frac{\mathrm{Var}(\mathbb{E}_{\bar{\mathbb{P}}}[\pi(\boldsymbol{X}) \mid D])}{\mathrm{Var}(\pi(\boldsymbol{X}))}. \tag{19}$$

In (19), the conditional expectation, whose variability reflects the dependence between prices and protected attributes, is calculated with respect to the population (conditional) distribution $\bar{\mathbb{P}}(\boldsymbol{X} \mid D)$. However, the two variances are calculated with respect to the portfolio-specific distribution $\mathbb{P}(\boldsymbol{X})$, since it is of interest the extent to which the chosen pricing functional performs within that portfolio context. If the portfolio and population distributions coincide, then (19) reduces to the standard version of the metric (4), $\mathrm{UF}(\pi, \bar{\mathbb{P}}) = \mathrm{UF}(\pi)$.

4.3 Model uncertainty

Both the metrics UF and PD rely on quantities that need to be estimated from data – hence their evaluation is subject to potential model error. A first observation is that data on protected attributes D is needed and this may not be available for all policies. For the discussion that follows we make the strong assumption that such data are available. Methods for dealing with only partial information on D in the context of calculating discrimination-free prices were developed by Lindholm et al. (2023).

Then, in the case of UF, model uncertainty does not present a major issue, if estimation takes place on a large enough set of insurance policies. First, as D is generally discrete and typically does not have a high number of states, the conditional expectation $\mathbb{E}[\pi(\boldsymbol{X}) \mid D = d]$ can be evaluated empirically on different subsets of the data. Similarly, the variances in the numerator and denominator of (4) can be respectively evaluated by the empirical distribution of D and $\boldsymbol{X}$. However, note that the population-based measure (19) raises the additional problem of estimating $\bar{\mathbb{P}}(\boldsymbol{X}, D)$; the challenges associated with this exercise are primarily about data sharing and privacy.

However, the estimation of PD is more susceptible to model error, since it relies on knowledge of the best-estimate prices $\mu(\boldsymbol{X}, D) = \mathbb{E}[Y \mid \boldsymbol{X}, D]$. While this can be calculated by regression methods from available policy data, the outcome will be heavily contingent on the class of models chosen (e.g., a GLM, a tree-based model or a deep neural network). This problem is fundamental to the idea of proxy discrimination, which is always understood within the context of a particular predictive model. One way to deal with this issue is to consider a discrete set of alternative plausible models for best-estimate prices $\mu_k(\boldsymbol{X}, D), k \in \mathcal{K}$. We may interpret each $\mu_k(\boldsymbol{X}, D)$ as a conditional expectation $\mathbb{E}_{\mathbb{P}_k}[Y \mid \boldsymbol{X}, D]$ under a different competing predictive model $\mathbb{P}_k$. Then, we may take some inspiration from robust decision making (Ben-Tal et al. 2009) and provide a suitable generalisation of Definition 4, by defining a version of PD, with respect to a model set $\mathcal{M} := \{\mu_k(\boldsymbol{X}, D) \mid k \in \mathcal{K}\}$:

$$\mathrm{PD}(\pi, \mathcal{M}) := \min_{c \in \mathbb{R},\ \boldsymbol{v} \in \mathcal{V}} \max_{k \in \mathcal{K}} \mathbb{E}\left[\left(\pi(\boldsymbol{X}) - c - \sum_{d \in \mathfrak{D}} \mu_k(\boldsymbol{X}, d) v_d\right)^2\right] \Bigg/ \mathrm{Var}(\pi(\boldsymbol{X})). \tag{20}$$

When there is no model uncertainty, $\mathcal{M}$ is a singleton and equation (20) simplifies to (8), i.e., $\mathrm{PD}(\pi, \{\mu(\boldsymbol{X}, D)\}) = \mathrm{PD}(\pi)$.

Finally, we note that, so far, we have not discussed the quality of the price functional π as a potential predictor of Y. This is a perspective we need to consider, since technical prices in insurance are explicitly linked to claim costs. The predictive performance of π cannot be disentangled from its discriminatory potential. In both Definitions 2 and 4, a large variance of the pricing functional will suppress the value of the metric. One may shed some light on this by focusing attention on pricing functionals that are auto-calibrated predictors of Y, that is, they satisfy $\pi(\boldsymbol{X}) = \mathbb{E}[Y \mid \pi(\boldsymbol{X})]$. For such predictors, a higher variability (in the sense of convex order) reflects a greater predictive accuracy with respect to a class of a convex loss functions; for details see, e.g., Krüger & Ziegel (2021), Denuit et al. (2021), Wüthrich (2023). Then, a high variance of the prices $\pi(\boldsymbol{X})$ can be seen as evidence of it being a good predictor, as it is more effective in differentiating between the predicted costs of policies with profile $\boldsymbol{X} = \boldsymbol{x}$. With this in mind, we can understand each of the Definitions 2 and 4 as quantifying the part of that differentiation potential that is viewed as undesirable, under the lens of either demographic unfairness or proxy discrimination.

4.4 Commercial prices and proxy discrimination

While much of the focus in the actuarial literature is on technical prices, when examining discrimination effects we need to consider the prices actually charged to policyholders and the commercial adjustments that these may include. For our discussion of demographic unfairness this creates no issues; commercial prices can (and should) be used in equation (4).

However, when considering the measurement of proxy discrimination, there are some complications. The constraints on the weights $\boldsymbol{v}$ in Definition 4 indicate that care should be taken when applying PD to commercial prices, which may include market discounts or penalties. On the one hand, the inclusion of the constant c allows for any additive bias of a commercial price $\pi(\boldsymbol{X})$ with respect to a convex combination of $\mu(\boldsymbol{X}, d)$ over d. On the other hand, multiplicative adjustments are potentially considered discriminatory. For instance, in the context of Example 2, a price of the form $\mu^*(X) = 1 + X$ would not be subject to proxy discrimination. On the other hand, any price of the form $\pi(X) = \lambda(1 + X)$, $\lambda > 1$ leads to proxy discrimination according to Definition 3, since this implies $\sum_d v_d > 1$. This can also be understood as the higher slope in X implicitly (and disproportionately) disadvantaging policyholders with $D = 1$.

Such penalisation of proportional price adjustments appears excessive. One potential way to address this is to require additional transparency, by decomposing any commercial price $\hat{\pi}(\boldsymbol{X})$ as

$$\hat{\pi}(\boldsymbol{X}) = \pi(\boldsymbol{X})\zeta(\boldsymbol{X}),$$

where $\pi(\boldsymbol{X})$ is understood as a predictor of Y and $\zeta(\boldsymbol{X})$ as a multiplicative commercial price adjustment. Then, in equation (8), the predictor $\pi(\boldsymbol{X})$ rather than the commercial price $\hat{\pi}(\boldsymbol{X})$ is used. Consequently, there remains the question of how to deal with the commercial adjustment $\zeta(\boldsymbol{X})$, which may itself introduce some discriminatory effects. Here, one could take a strong view and require that $\zeta(\boldsymbol{X})$ be independent of D. While some degree of demographic unfairness can be considered acceptable within a portfolio from a claims costing perspective, as long as proxying effects are adjusted for, the application of different average levels of commercial discounts or loadings per demographic group would be much more problematic. Independence between $\zeta(\boldsymbol{X})$ and D is most straightforwardly achieved by making $\zeta(\boldsymbol{X})$ a constant; a more advanced alternative entails pre-processing $\boldsymbol{X}$ in the spirit of the Input Optimal Transport methods discussed in, e.g., Lindholm et al. (2024).

A different solution would be to consider a 'best-estimate commercial price' $\nu(\boldsymbol{X}, D)$ that explicitly depends on both non-protected covariates and protected attributes. Of course even the existence of such a functional (let alone attempts to estimate it) would be controversial. Having said that, in the same way that $\mu(\boldsymbol{X}, D)$ is calculated by minimising the deviation of Y from a statistical predictor, one could envisage $\nu(\boldsymbol{X}, D)$ as the result of a price optimisation exercise (e.g., Guelman & Guillén 2014, Verschuren 2022) that involves all policyholders' characteristics. Of course $\nu(\boldsymbol{X}, D)$ cannot be used in pricing, as it depends on D. However it can be used to assess proxy discrimination. Consider $\nu(\boldsymbol{X})$ a commercial analogue of the unawareness price, now derived by price optimisation using data on $\boldsymbol{X}$ only. In that context, one could assess the extent to which $\nu(\boldsymbol{X})$ proxy-discriminates by the corresponding measure:

$$\mathrm{PD}\big(\nu(\boldsymbol{X}), \{\nu(\boldsymbol{X}, D)\}\big) = \frac{\min_{c \in \mathbb{R},\ \boldsymbol{v} \in \mathcal{V}} \mathbb{E}\left[\Big(\nu(\boldsymbol{X}) - c - \sum_{d \in \mathfrak{D}} \nu(\boldsymbol{X}, d) v_d\Big)^2\right]}{\mathrm{Var}(\nu(\boldsymbol{X}))}.$$

Such an approach would be contingent on the prices $\nu(\boldsymbol{X}), \nu(\boldsymbol{X}, D)$ being derived algorithmically, with estimation of the latter strictly regulated, with the sole purpose of acting as a benchmark to measure proxy discrimination. This complex situation could be nonetheless simplified, if it can be argued that an insurer's costs associated with a policy are proportional to claims costs such that $\nu(\boldsymbol{X}, D) := \lambda\,\mu(\boldsymbol{X}, D),\ \nu(\boldsymbol{X}) := \lambda\,\mu(\boldsymbol{X}),\ \lambda > 1$. With this justification, one can restore proportional loadings, without falling foul of requirements to avoid proxy discrimination.

References

Aas, K., Jullum, M. & Løland, A. (2021), 'Explaining individual predictions when features are dependent: More accurate approximations to shapley values', *Artificial Intelligence* **298**, 103502.

Araiza Iturria, C. A., Hardy, M. & Marriott, P. (2024), 'A discrimination-free premium under a causal framework', *North American Actuarial Journal* pp. 1–21.

Barocas, S. & Selbst, A. D. (2016), 'Big data's disparate impact', *California law review* pp. 671–732.

Ben-Tal, A., El Ghaoui, L. & Nemirovski, A. (2009), *Robust optimization*, Vol. 28, Princeton university press.

Bénesse, C., Gamboa, F., Loubes, J.-M. & Boissin, T. (2022), 'Fairness seen as global sensitivity analysis', *Machine Learning* pp. 1–28.

Borgonovo, E., Clemente, G. P. & Rabitti, G. (2024), 'Why insurance regulators need to require sensitivity settings of internal models for their approval', *Finance Research Letters* **60**, 104859.

Borgonovo, E. & Plischke, E. (2016), 'Sensitivity analysis: A review of recent advances', *European Journal of Operational Research* **248**(3), 869–887.

Charpentier, A. (2024), *Insurance, biases, discrimination and fairness*, Springer.

Charpentier, A., Hu, F. & Ratz, P. (2023), 'Mitigating Discrimination in Insurance with Wasserstein Barycenters', *ArXiv:2306.12912* .

Chiappa, S., Jiang, R., Stepleton, T., Pacchiano, A., Jiang, H. & Aslanides, J. (2020), A general approach to fairness with optimal transport, *in* 'Proceedings of the AAAI Conference on Artificial Intelligence', Vol. 34, pp. 3633–3640.

Chzhen, E., Denis, C., Hebiri, M., Oneto, L. & Pontil, M. (2020), 'Fair regression with Wasserstein barycenters', *Advances in Neural Information Processing Systems* **33**, 7321–7331.

Cook, T., Greenall, A. & Sheehy, E. (2022), Discriminatory pricing: Exploring the 'ethnicity penalty' in the insurance market, Technical report, Citizens Advice. https://www.citizensadvice.org.uk/policy/publications/discriminatory-pricing-exploring-the-ethnicity-penalty-in-the-insurance-market1/ [Accessed: 08 July 2024].

Côté, O., Côté, M.-P. & Charpentier, A. (2024), 'A fair price to pay: exploiting causal graphs for fairness in insurance', *Available at SSRN 4709243* .

Da Veiga, S., Wahl, F. & Gamboa, F. (2009), 'Local polynomial estimation for sensitivity analysis on models with correlated inputs', *Technometrics* **51**(4), 452–463.

Delbaen, F. & Majumdar, C. (2024), Approximation with independent variables, *in* 'Peter Carr Gedenkschrift: Research Advances in Mathematical Finance', World Scientific, pp. 311–327.

Denuit, M., Charpentier, A. & Trufin, J. (2021), 'Autocalibration and tweedie-dominance for insurance pricing with machine learning', *Insurance: Mathematics and Economics* **101**, 485–497.

EIOPA (2021), Artificial intelligence governance principles: towards ethical and trustworthy artificial intelligence in the european insurance sector. a report from eiopa´s consultative expert group on digital ethics in insurance, Technical report, European Insurance and Occupational Pensions Authority.

Frees, E. W. & Huang, F. (2023), 'The discriminating (pricing) actuary', *North American Actuarial Journal* **27**(1), 2–24.

Gordaliza, P., Del Barrio, E., Fabrice, G. & Loubes, J.-M. (2019), Obtaining fairness using optimal transport theory, *in* 'International conference on machine learning', PMLR, pp. 2357–2365.

Guelman, L. & Guillén, M. (2014), 'A causal inference approach to measure price elasticity in automobile insurance', *Expert Systems with Applications* **41**(2), 387–396.

Hiabu, M., Meyer, J. T. & Wright, M. N. (2023), Unifying local and global model explanations by functional decomposition of low dimensional structures, *in* 'International Conference on Artificial Intelligence and Statistics', PMLR, pp. 7040–7060.

Jansen, M. J. (1999), 'Analysis of variance designs for model output', *Computer Physics Communications* **117**(1-2), 35–43.

Kleinberg, J., Mullainathan, S. & Raghavan, M. (2016), 'Inherent trade-offs in the fair determination of risk scores', *arXiv preprint arXiv:1609.05807* .

Krüger, F. & Ziegel, J. F. (2021), 'Generic conditions for forecast dominance', *Journal of Business & Economic Statistics* **39**(4), 972–983.

Kusner, M. J., Loftus, J., Russell, C. & Silva, R. (2017), 'Counterfactual fairness', *Advances in neural information processing systems* **30**.

Lindholm, M., Richman, R., Tsanakas, A. & Wüthrich, M. V. (2022), 'Discrimination-free insurance pricing', *ASTIN Bulletin: The Journal of the IAA* **52**(1), 55–89.

Lindholm, M., Richman, R., Tsanakas, A. & Wüthrich, M. V. (2023), 'A multi-task network approach for calculating discrimination-free insurance prices', *European Actuarial Journal* pp. 1–41.

Lindholm, M., Richman, R., Tsanakas, A. & Wuthrich, M. V. (2024), 'What is fair? proxy discrimination vs. demographic disparities in insurance pricing', *Scandinavian Actuarial Journal* pp. 1–36.

Lundberg, S. M. & Lee, S.-I. (2017), 'A unified approach to interpreting model predictions', *Advances in neural information processing systems* **30**.

MAS (2022), MAS-led Industry Consortium Publishes Assessment Methodologies for Responsible Use of AI by Financial Institutions, Technical report, Monetary Authority of Singapore. https://www.mas.gov.sg/news/media-releases/2022/mas-led-industry-consortium-publishes-assessment-methodologies-for-responsible-use-of-ai-by-financial-institutions [Accessed: 08 July 2024].

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. (2021), 'A survey on bias and fairness in machine learning', *ACM computing surveys (CSUR)* **54**(6), 1–35.

Owen, A. B. (2014), 'Sobol'indices and shapley value', *SIAM/ASA Journal on Uncertainty Quantification* **2**(1), 245–251.

Owen, A. B. & Prieur, C. (2017), 'On shapley value for measuring importance of dependent inputs', *SIAM/ASA Journal on Uncertainty Quantification* **5**(1), 986–1002.

Rabitti, G. & Borgonovo, E. (2020), 'Is mortality or interest rate the most important risk in annuity models? a comparison of sensitivity analysis methods', *Insurance: Mathematics and Economics* **95**, 48–58.

Richman, R. & Wüthrich, M. V. (2023), 'Conditional expectation network for shap', *arXiv preprint arXiv:2307.10654* .

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M. & Tarantola, S. (2010), 'Variance based sensitivity analysis of model output. design and estimator for the total sensitivity index', *Computer physics communications* **181**(2), 259–270.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M. & Tarantola, S. (2008), *Global sensitivity analysis: the primer*, John Wiley & Sons.

Shapley, L. S. et al. (1953), 'A value for n-person games'.

Sobol', I. M. (2001), 'Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates', *Mathematics and computers in simulation* **55**(1-3), 271–280.

Song, E., Nelson, B. L. & Staum, J. (2016), 'Shapley effects for global sensitivity analysis: Theory and computation', *SIAM/ASA Journal on Uncertainty Quantification* **4**(1), 1060–1083.

Tschantz, M. C. (2022), What is proxy discrimination?, *in* 'Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency', pp. 1993–2003.

Vallarino, A., Rabitti, G. & Chokami, A. K. (2024), 'Construction of rating systems using global sensitivity analysis: A numerical investigation', *ASTIN Bulletin: The Journal of the IAA* **54**(1), 25–45.

Verschuren, R. M. (2022), 'Customer price sensitivities in competitive insurance markets', *Expert Systems with Applications* **202**, 117133.

Wüthrich, M. V. (2023), 'Model selection with gini indices under auto-calibration', *European Actuarial Journal* **13**(1), 469–477.

Xin, X. & Huang, F. (2024), 'Antidiscrimination insurance pricing: Regulations, fairness criteria, and models', *North American Actuarial Journal* **28**(2), 285–319.