# A MULTIVARIATE ENERGY-BASED FAIRNESS ADJUSTER FOR PREMIUMS

Charlotte Jamotton, Donatien Hainaut

REPRINT | 2026 / 20

UCLouvain

ISBA | LIDAM
Louvain Institute of Data Analysis and Modeling in economics and statistics

**ISBA**

Voie du Roman Pays 20 - L1.04.01

B-1348 Louvain-la-Neuve

Email : lidam-library@uclouvain.be

https://uclouvain.be/en/research-institutes/lidam/isba/publication

# A Multivariate Energy-based Fairness Adjuster for Premiums

Charlotte Jamotton*[1] and Donatien Hainaut[1]

[1]Institute of Statistics, Biostatistics and Actuarial Sciences (ISBA), Université catholique de Louvain (UCLouvain), Voie du Roman Pays 20/L1.04.01, Louvain-la-Neuve, 1348, Belgium

**Abstract**

Fairness in insurance pricing has received increasing attention, particularly in light of regulations calling for non-discriminatory premium estimation such as the EU Gender Directive (2012). This research focuses on removing structural bias in predicted premiums as a solidarity-driven fairness concept. We propose a gradient-based distributional adjustment designed to mitigate disparities across protected groups. The method relies on the Energy distance, a multivariate metric that enables the joint alignment of predicted premium distributions across multiple sensitive attributes (including non-binary ones) simultaneously. To maintain overall prediction balance, a post-hoc autocalibration step corrects biases in the total predicted number of claims. In addition, the method supports fairness adjustments for new policyholders without requiring retraining of the underlying fairness model. We evaluate the proposed methodology in a car insurance pricing setting, where demographic factors such as age and gender are commonly used for risk assessment and premium determination. Results show that the method effectively reduces group-level disparities while retaining a significant share of predictive accuracy.

**Keywords**: fairness, demographic parity, autocalibration, actuarial pricing, Energy distance

*Corresponding author: charlotte.jamotton@uclouvain.be

# 1. Introduction

A key characteristic of the insurance industry is its inverted production cycle. Ideally, premiums should be a function of the insured's risk, but this risk is unobservable when the contract is signed. Therefore, to set the premium that must be paid upfront, insurers instead rely on salient (and observable) features that correlate with the insured's risk. However, some of these rating factors, such as gender and ethnicity, are prohibited under anti-discrimination laws. While policymakers seek to ensure that insurers do not make distinctions based on these legally prohibited characteristics (Avraham, 2017, p.2), the law of insurance underwriting inherently grants insurers the authority to discriminate because it makes statistical sense to do so (Gandy, 2016). This phenomenon, often described as an injustice by generalisation (Schanze, 2013), is rooted in the assumption that individuals meet the average characteristics of the risk group to which they belong. This assumption lies at the core of the mutualization mechanism in insurance pricing, where risks are segmented into homogeneous rate-making classes or risk pools, imposing a form of solidarity among policyholders with similar risk profiles. Nonetheless, as Charpentier (2023, p.11, p.241–242) argues, individuals should not be treated differently because of their group membership, particularly when that membership is assigned at birth, as in the case of gender or ethnicity.

The seemingly incompatible notions of actuarial pricing and fairness present significant challenges for actuaries, who must comply with anti-discrimination regulations while still differentiating risk based on relevant characteristics. Moreover, the concept of discrimination itself is complex, multifaceted, and subject to interpretation (Edwards, 2007), making its identification in insurance particularly difficult. As Charpentier (2023, p.1–10) explains, achieving fairness in insurance requires integrating justice and ethics into data-driven pricing models. For a broader discussion on the legal, economic, and ethical aspects of discrimination in actuarial pricing, we refer to Frees and Huang (2021) and Charpentier (2023).

In recent years, the rise of big data and advances in machine learning techniques have further exacerbated these ongoing questions about fairness in insurance. While insurers are prohibited from directly discriminating based on specific policyholder attributes, simply removing these sensitive variables from pricing models does not entirely eliminate the risk of discrimination. This raises a critical question (Tschantz, 2022). If insurers are not allowed to directly discriminate based on gender, why are proxy variables, such as vehicle power which correlate with and indirectly infer gender, still permissible?

In the actuarial literature, recent studies by Lindholm et al. (2022) and Charpentier et al. (2023) have proposed different post-processing approaches to mitigate both direct and indirect discrimination in insurance pricing. Lindholm et al. (2022, 2024) individual fairness method aims to address proxy discrimination, commonly referred to as confounding bias in causal theory (Côté et al., 2025). Inspired by post-stratification techniques,

it consists of averaging over sensitive information and a change of measure to prevent implicit inference of protected characteristics. Conversely, group fairness methodologies, such as those using Wasserstein barycenters proposed by Charpentier et al. (2023) and extended by Hu et al. (2023a, 2023b), aim to align outcome distributions across protected groups so that no association exists between sensitive information and premiums. According to Côté et al. (2025, p.10), these so-called corrective premiums offer the most stringent mitigation of indirect discrimination. A comprehensive review and classification of existing fairness methodologies in insurance can be found in Xin and Huang (2023) and Côté et al. (2025).

In this chapter, we focus on Demographic Parity (DP), a solidarity-driven group fairness criterion, to remove structural bias in predicted premiums. Although EU regulations on gender-neutral pricing align more closely with an individual fairness perspective, we argue that DP is more consistent with the risk pooling mechanism inherent in insurance pricing. In practice, insurance contracts are typically priced using regression models (Denuit et al., 2019), such as Generalized Linear Models (GLMs), which assign a limited set of unique premiums to a much larger pool of policies. In other words, the risk assessment and premium allocation are not individualised but structured around groups with shared characteristics. By adopting DP, we ensure that the probability of receiving a given premium level becomes independent of demographic attributes such as gender, promoting fair treatment across groups in line with the risk pooling principle. We acknowledge that different theoretical and practical considerations may support alternative fairness criteria.

The main contributions of this paper are threefold. First, we propose a gradient-based fairness correction using the Energy distance to reduce disparities across multiple sensitive attributes simultaneously, including non-binary variables. The multivariate nature of the Energy distance is a key advantage of our approach. It allows for simultaneous fairness adjustments across protected variables with different numbers of sensitive classes, such as gender (binary) and discretised age (multi-class), and is therefore order-independent. Second, we introduce a post-hoc autocalibration step (Denuit et al., 2021) to correct biases in the total predicted claim count after fairness adjustments. Third, we design a simple out-of-sample mapping mechanism to apply fairness corrections to new policyholders. Note that, while several recent approaches, including individual fairness and counterfactual fairness methods (Fahrenwaldt et al., 2024; Grari et al., 2020; Kusner et al., 2017), can generalise fairness adjustments to new observations, our contribution lies in providing a straightforward mapping mechanism that enables the out-of-sample transformation of premiums without re-optimising fairness corrections for each new policyholder. In the machine-learning literature, post-hoc corrections have similarly been shown to enforce group-fairness constraints without retraining (Hardt et al., 2016; Pleiss et al., 2017).

The chapter is structured as follows. Section 2 introduces notations and defines Demographic Parity as a group fairness criterion. Section 3 presents the theoretical justifica-

tion for choosing the Energy distance and its relationship to other statistical distances. Section 4 details the gradient-based adjustment, the autocalibration procedure, and the out-of-sample mapping. Section 5 evaluates the methodology in a car insurance pricing setting, demonstrating its effectiveness in reducing group-level disparities while preserving a substantial share of predictive accuracy.

## 2. Notations and definitions

Consider an insurance portfolio represented as a tabular dataset $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \subset \mathbb{R}^m$ with $n$ policies (rows) and $m$ rating factors (columns). Each policy $i = 1, \ldots, n$ is described by a vector of characteristics $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{im}) \in \mathbb{R}^m$. We partition the $m$ rating factors into two disjoint sets. The set of allowed covariates $\mathcal{A} \subseteq \{1, \ldots, m\}$, and the set of prohibited (or discriminatory) covariates $\mathcal{D} \subseteq \{1, \ldots, m\}$, such that $\mathcal{A} \cup \mathcal{D} = \{1, \ldots, m\}$ and $\mathcal{A} \cap \mathcal{D} = \varnothing$. For each policy $i$, we denote by $\boldsymbol{A}_i = (x_{ij})_{j \in \mathcal{A}}$, and $\boldsymbol{D}_i = (x_{ij})_{j \in \mathcal{D}}$ the subvectors of allowed and prohibited rating factors, respectively.

For every policy $i$, we observe the number of claims $N_i$ (count data) reported during its contract duration (or exposure) $\nu_i$. Although we present the notations using claim counts $N$, the methodology applies to any real-valued insurance response. Throughout the analysis, we will often refer to the claim frequency per unit of exposure, defined as the ratio $\lambda_i = N_i/\nu_i$. Counts are low-support integers (typically 0–3), while $\lambda$ is continuous on $[0, N/\min(\nu)]$, more stable, and more interpretable for illustrating fairness adjustments.

In actuarial pricing, the objective is to approximate the (unknown) conditional expectation of the response variable $Y$ given the available rating information $(\boldsymbol{A}, \boldsymbol{D})$:

$$\mu(\boldsymbol{A}, \boldsymbol{D}) = \mathbb{E}[Y | (\boldsymbol{A}, \boldsymbol{D})] \,. \tag{1}$$

In the empirical analysis presented in Section 5, we use a Light Gradient Boosting machine (LightGBM) to model the premium best-estimates $\hat{\pi}(\boldsymbol{A}, \boldsymbol{D})$ for the (true) pure premium $\mu(\boldsymbol{A}, \boldsymbol{D})$ defined in Eq. (1). This best-estimate predictor uses all available features, including the protected attributes $\boldsymbol{D}$. Although it is statistically optimal (because it exploits all predictive information), it is also directly discriminatory by construction. In this chapter, we focus on Demographic Parity, a solidarity-driven group fairness criterion, to remove structural bias from predicted premiums.

*Definition* 2.1 (Demographic Parity). Demographic Parity (DP) is a group fairness criterion that requires the conditional distribution of the premiums to be identical across all protected demographic groups. For two groups $a$ and $b$ in the support of $\boldsymbol{D}$,

$$\mathbb{P}(\hat{\pi}(\boldsymbol{A}, \boldsymbol{D}) | \boldsymbol{D} = a) = \mathbb{P}(\hat{\pi}(\boldsymbol{A}, \boldsymbol{D}) | \boldsymbol{D} = b) \,. \tag{2}$$

In practice, DP ensures that no protected group receives systematically different pricing.

In a regression setting, violations of Eq. (2) can be quantified by a statistical dis-

tance between the conditional distributions $\hat{\pi}(\boldsymbol{A}, \boldsymbol{D}) \mid (\boldsymbol{D} = a)$ and $\hat{\pi}(\boldsymbol{A}, \boldsymbol{D}) \mid (\boldsymbol{D} = b)$. Many distances may be used to quantify disparities across protected groups, including divergences (e.g., Kullback-Leibler) or optimal-transport metrics (e.g., Wasserstein). As detailed in Section 4, we propose the use of the Energy distance, defined in Section 3, to align the group-conditional distributions of predicted premiums in order to satisfy the DP condition in Eq. (2).

# 3. Measures of disparities

In this section, we review several distances and divergences commonly used to quantify discrepancies between probability distributions, with a particular emphasis on their suitability for gradient-based fairness correction. We discuss the Kullback-Leibler divergence, the Wasserstein distance, the Cramér distance, and its multivariate extension i.e., the Energy distance. Their respective properties and limitations motivate our choice of the Energy distance for enforcing demographic parity in premium predictions.

## 3.1 Classical distances and divergences

In machine learning, Kullback and Leibler (1951) divergence is among the most widely used measures of dissimilarity between two distributions.

*Definition* 3.1 (Kullback-Leibler Divergence). For probability distributions $P$ and $Q$ with densities $p$ and $q$, the Kullback-Leibler (KL) divergence is defined as:

$$\mathrm{d}_{\mathrm{KL}}\left(P \,\|\, Q\right) = \int_{-\infty}^{\infty} p(u) \log\left(\frac{p(u)}{q(u)}\right) dx\,.$$

As a measure of relative entropy, the KL divergence does not account for the geometric proximity between outcomes. For example, consider an actuarial pricing context where $p(u)$ assigns a 0.01% probability to a premium level $u = 100$, while $q(u)$ assigns the same probability to $u = 101$. Despite the two premiums being very close, the KL divergence treats them as entirely different outcomes. To address this limitation, researchers (Panaretos & Zemel, 2019; Peyré & Cuturi, 2019) have turned to alternative distance metrics that consider the underlying geometry between outcomes, such as Wasserstein.

*Definition* 3.2 (Wasserstein Distance). For two probability distributions $P$ and $Q$, the Wasserstein distance of order $p$, also known as the Earth Mover distance when $p = 1$, is defined as (Kantorovich & Rubinshtein, 1958):

$$W_p(P, Q) = \left(\inf_{\pi \in \Pi(P,Q)} \int \|u - v\|^p d\pi\left(u, v\right)\right)^{1/p}, \tag{3}$$

where $\Pi(P, Q)$ denotes the set of all joint distributions for the pair $(U, V)$ with marginals $P$ and $Q$, respectively. The infimum in Eq. (3) is taken over all possible couplings

(defined below) $\pi \in \Pi(P, Q)$ ensuring that the optimal transport plan minimises the cost of transporting mass from $u$ to $v$ (Villani, 2009).

*Definition* 3.3 (Coupling). A coupling of two probability distributions $P$ and $Q$ is any joint distribution $\pi$ on $(U, V)$ such that $U \sim P$ and $V \sim Q$. The set of all such couplings is denoted $\Pi(P, Q)$ (Santambrogio, 2015; Villani, 2009).

The Wasserstein distance accounts for the underlying geometry between outcomes through the transport cost $\|u - v\|^p$. However, a key drawback for optimisation is that $W_p$ does not satisfy the unbiased sample gradients property (Bellemare et al., 2017), one of three key properties (the other two being sum invariance and scale sensitivity) in machine learning when using a distance as a loss function.

*Definition* 3.4 (Unbiased Sample Gradients). Let $\mathrm{d}(P, Q)$ be a statistical divergence. We say d has unbiased sample gradients if, for all $m$ independent samples $\boldsymbol{u}_m$ drawn from $P$, the expected gradient of the sample loss $\theta \mapsto \mathrm{d}\left(\hat{P}_m, Q_\theta\right)$ satisfies (Bellemare et al., 2017, p.3):

$$\mathbb{E}_{\boldsymbol{u}_m \sim P}\left[\nabla_\theta \, \mathrm{d}\left(\hat{P}_m, Q_\theta\right)\right] = \nabla_\theta \, \mathrm{d}\left(P, Q_\theta\right) .$$

Bellemare et al. (2017) show that the Wasserstein distance lacks unbiased sample gradients, leading to convergence issues and motivating Cramér (1928) distance as an unbiased alternative.

*Definition* 3.5 (Cramér Distance). For one-dimensional distributions $P$ and $Q$ with cumulative distribution functions (CDFs) $F_P$ and $F_Q$, Cramér (1928) distance is defined as the $\ell_2^2$-distance:

$$\ell_2^2\left(P, Q\right) = \int_{-\infty}^{\infty} \left|F_P(u) - F_Q(u)\right|^2 \, dx \, . \tag{4}$$

Its square root is a member of the $\ell_p$ family of metrics:

$$\ell_p\left(P, Q\right) = \left(\int_{-\infty}^{\infty} \left|F_P(u) - F_Q(u)\right|^p \, dx\right)^{1/p} . \tag{5}$$

Interestingly, for one-dimensional distributions $P$ and $Q$, the $p$-Wasserstein distance in Eq. (3) can be expressed in terms of the quantile functions $F_P^{-1}$ and $F_Q^{-1}$ (Panaretos & Zemel, 2019):

$$W_p(P, Q) = \left(\int_0^1 \left|F_P^{-1}(u) - F_Q^{-1}(u)\right|^p du\right)^{1/p} . \tag{6}$$

For $p = 1$, Eq. (6) simplifies to the $\ell_1$-distance, a member of the $\ell_p$ family defined in

Eq. (5) like the square-root of the Cramér distance:

$$\begin{aligned} W_1(P,Q) &= \int_0^1 \left|F_P^{-1}(u) - F_Q^{-1}(u)\right| dx \\ &= \int_{-\infty}^{+\infty} |F_P(u) - F_Q(u)|\, dx \\ &= \ell_1(P,Q)\,. \end{aligned}$$

Besides accounting for the underlying geometry between outcomes (like Wasserstein and unlike KL), the Cramér distance also possesses desirable optimisation properties, including unbiased sample gradients (unlike Wasserstein), sum invariance, and scale sensitivity, making it more suitable for gradient-based learning than the Wasserstein metric. The following property establishes that the squared $\ell_2$-distance admits unbiased sample gradients, as demonstrated in Appendix A.

*Proposition* 3.1 (Unbiased sample gradient of the squared $\ell_2$-distance). Let $P$ be a probability measure and let $\{Q_\theta\}_{\theta\in\Theta}$ denote a parameterized family of probability measures, with cumulative distribution functions $F_P(u)$ and $F_{Q_\theta}(u)$, respectively. The gradient of the squared $\ell_2$-distance between $P$ and $Q_\theta$ satisfies the following identity (Bellemare et al., 2017, p.15):

$$\nabla_\theta\, \ell_2^2\,(P, Q_\theta) = \mathbb{E}_{\boldsymbol{u}_m\sim P}\left[\nabla_\theta\, \ell_2^2\left(\hat{P}_m, Q_\theta\right)\right],$$

where $\hat{P}_m$ denotes the empirical measure based on $m$ independent samples drawn from $P$.

Note that while Bellemare et al. (2017) show that the Wasserstein distance lacks unbiased sample gradients, leading to convergence issues and motivating Cramér (1928) distance as an unbiased alternative, this problem was further confirmed in practice for minibatch optimisation by Fatras et al. (2019), and addressed by Genevay et al. (2017) via the Sinkhorn divergence.

## 3.2 Energy distance

Although the Cramér distance combines the best of KL and Wasserstein, it is limited to univariate distributions. In the context of actuarial fairness, real-world scenarios often involve multiple sensitive attributes. A key observation by Székely (2003) is that $\ell_2^2(P,Q)$ in Eq. (4) can be rewritten as:

$$\ell_2^2(P,Q) = \mathbb{E}\big[|U-V|\big] - \tfrac{1}{2}\mathbb{E}\big[|U-U'|\big] - \tfrac{1}{2}\mathbb{E}\big[|V-V'|\big]\,,$$

where $|.|$ denotes the absolute value for scalars $u, v \in \mathbb{R}$, and $U'$ and $V'$ are independent and identically distributed copies of $U$ and $V$. As shown by Székely (2003), replacing absolute values by Euclidean norms $\|\cdot\|$ generalises Eq. (4) to higher dimensions, known as the Energy distance.

*Definition* 3.6 (Energy Distance). For $d$-dimensional independent random variables $U, V$,

the Energy distance (for the Euclidean metric) is defined as:

$$\begin{aligned}\mathcal{E}(P,Q) =& 2\,\mathbb{E}_{U\sim P,V\sim Q}\big[\|U-V\|\big] \\ &- \mathbb{E}_{U,U'\sim P}\big[\|U-U'\|\big] - \mathbb{E}_{V,V'\sim Q}\big[\|V-V'\|\big]\,. \end{aligned} \tag{7}$$

In each expectation the relevant draws are independent (independent pairs for the within-distribution terms and independent cross draws for the first term).

For univariate distributions, the equivalence $\mathcal{E}(P,Q) = 2\ell_2^2(P,Q)$ is proved in Appendix A. Like Cramér distance, the $\mathcal{E}$-distance admits unbiased sample gradients, as formally demonstrated in Appendix A.

*Proposition* 3.2 (Unbiased sample gradient of the $\mathcal{E}$-distance). Let $U \sim P$, $V \sim Q_\theta$, and let $\hat{U}_m \sim \hat{P}_m$ denote $m$ independent samples drawn from $P$. The gradient $\nabla$ of the Energy distance $\mathcal{E}\left(P, Q_\theta\right)$ with respect to $\theta$ satisfies the following identity (Bellemare et al., 2017, p.15):

$$\nabla_\theta\, \mathcal{E}\left(P, Q_\theta\right) = \mathbb{E}_{\boldsymbol{u}_m\sim P}\left[\nabla_\theta\, \mathcal{E}\left(\hat{P}_m, Q_\theta\right)\right]\,,$$

showing that $\mathcal{E}$ admits unbiased sample gradients.

The $\mathcal{E}$-distance also retains a strong connection to optimal transport. The Wasserstein distance in Eq. (3) can be defined in terms of expectations over coupled random variables $U \sim P$ and $V \sim Q$:

$$\begin{aligned} W_p(P,Q) &= \left(\inf_{\pi\in\Pi(P,Q)} \int \|u-v\|^p d\pi\,(u,v)\right)^{1/p} \\ &= \inf_{\pi\in\Pi(P,Q)} \mathbb{E}_{(U,V)\sim\pi}\Big[\|U-V\|^p\Big]^{1/p}\,. \end{aligned}$$

As discussed in Sejdinović et al. (2013), the Energy distance can be equivalently represented in the integral form. The first expectation in Eq. (7) may be written as the iterated integral:

$$\mathbb{E}_{U\sim P,\,V\sim Q}\big[\|U-V\|\big] = \int_{\mathcal{U}} \left(\int_{\mathcal{V}} \|u-v\|\, Q(dy)\right) P(dx)\,.$$

Given that $U$ and $V$ are independent, the joint law of $(U,V)$ is the product measure $P\otimes Q$. By Fubini/Tonelli the iterated integral equals the integral with respect to the product measure $P\otimes Q$:

$$\int_{\mathcal{U}}\int_{\mathcal{V}} \|u-v\|\, Q(dy)\, P(dx) = \int_{\mathcal{U}\times\mathcal{V}} \|u-v\|\, (P\otimes Q)(dx\,dy) = \mathbb{E}_{(U,V)\sim P\otimes Q}\big[\|U-V\|\big]\,.$$

Consequently, Eq. (7) is equivalently written as:

$$\mathcal{E}(P,Q) = 2\,\mathbb{E}_{(U,V)\sim P\otimes Q}\big[\|U-V\|\big] - \mathbb{E}_{P\otimes P}\big[\|U-U'\|\big] - \mathbb{E}_{Q\otimes Q}\big[\|V-V'\|\big]\,.$$

The last two expectations are already written as expectations under $(U,U') \sim P\otimes P$ and $(V,V') \sim Q\otimes Q$ to emphasize that those terms depend only on the marginals $P$ and $Q$.

The product measure (joint law) $P \otimes Q$ is an element of the set of couplings $\Pi(P, Q)$. The usual Energy distance is closely related to the 1-Wasserstein distance evaluated at the independent/product coupling $\pi_\perp = P \otimes Q \in \Pi(P, Q)$:

$$\mathcal{E}(P, Q) \;=\; 2\,\mathbb{E}_{(U,V)\sim\pi_\perp}\big[\|U - V\|\big] \;-\; \mathbb{E}_{P\otimes P}\big[\|U - U'\|\big] \;-\; \mathbb{E}_{Q\otimes Q}\big[\|V - V'\|\big]\,.$$

As noted by Duong et al. (2022), both the Wasserstein and Energy distances can be interpreted as the energy required to move one probability distribution $P$ to another $Q$. The Wasserstein distance measures the cost of the optimal transport plan, whereas the Energy distance considers the cost of a random (maximum-entropy) transport plan. In the following Section 4, we introduce a gradient-based fairness correction mechanism based on the Energy distance.

# 4. Energy gradient fairness adjuster

As outlined in Section 2, our goal is to mitigate group-level disparities across $G \geq 2$ protected groups by enforcing demographic parity. More specifically, we seek to align the conditional distribution of predicted premiums $\hat{\pi}(\boldsymbol{A}, \boldsymbol{D})$ across all values of the sensitive attribute(s) $\boldsymbol{D}$. The Energy distance provides a differentiable and geometrically meaningful measure of disparity, making it particularly suitable for this type of gradient-based fairness correction.

## 4.1 Gradient-based fairness correction

Let $U, V \in \mathbb{R}^d$ be two independent random variables drawn from distributions $P$ and $Q$, respectively. Given samples $\{\boldsymbol{u}_i\}_{i=1}^{n_u}$ and $\{\boldsymbol{v}_j\}_{j=1}^{n_v}$ in $\mathbb{R}^d$, the Energy distance, defined in Eq. (7), admits the empirical estimate:

$$\begin{aligned}\mathcal{E}(P, Q) =& \frac{2}{n_u\, n_v} \sum_{i=1}^{n_u} \sum_{j=1}^{n_v} \|\boldsymbol{u}_i - \boldsymbol{v}_j\| \\ &- \frac{1}{n_u^2} \sum_{i=1}^{n_u} \sum_{j\neq i} \|\boldsymbol{u}_i - \boldsymbol{u}_j\| - \frac{1}{n_v^2} \sum_{i=1}^{n_v} \sum_{j\neq i} \|\boldsymbol{v}_i - \boldsymbol{v}_j\|\,, \end{aligned} \tag{8}$$

with the pairwise Euclidean distance defined as:

$$\|\boldsymbol{u}_i - \boldsymbol{v}_j\| = \sqrt{\sum_{k=1}^{d} (u_{ik} - v_{jk})^2}\,.$$

We propose a fairness correction based on the Energy gradient that iteratively nudges the distributions of $U$ and $V$ toward each other. For an observed pair $(\boldsymbol{u}_i, \boldsymbol{v}_j)$, the gradients

of the pairwise norm with respect to a component $u_{ik}$ or $v_{jk}$ are:

$$\frac{\partial\|\boldsymbol{u}_i-\boldsymbol{v}_j\|}{\partial u_{ik}}=\frac{u_{ik}-v_{ik}}{\|\boldsymbol{u}_i-\boldsymbol{v}_j\|}, \quad \text{and} \quad \frac{\partial\|\boldsymbol{u}_i-\boldsymbol{v}_j\|}{\partial v_{jk}}=-\frac{y_{jk}-u_{jk}}{\|\boldsymbol{v}_j-\boldsymbol{u}_i\|}.$$

Hence, the gradient of the Energy distance with respect to individual observations $\{\boldsymbol{u}_i\}_{i=1}^{n_u}$ and $\{\boldsymbol{v}_j\}_{j=1}^{n_v}$ is:

$$\begin{aligned}\nabla_{\boldsymbol{u}_i}\mathcal{E}(P,Q)=&\frac{2}{n_u n_v}\sum_{j=1}^{n_v}\frac{\boldsymbol{u}_i-\boldsymbol{v}_j}{\|\boldsymbol{u}_i-\boldsymbol{v}_j\|}-\frac{2}{n_u^2}\sum_{j=1}^{n_u}\frac{\boldsymbol{u}_i-\boldsymbol{u}_j}{\|\boldsymbol{u}_i-\boldsymbol{u}_j\|}, \quad \text{and}\\ \nabla_{\boldsymbol{v}_j}\mathcal{E}(P,Q)=&-\frac{2}{n_u n_v}\sum_{i=1}^{n_u}\frac{\boldsymbol{v}_j-\boldsymbol{u}_i}{\|\boldsymbol{v}_j-\boldsymbol{u}_i\|}+\frac{2}{n_v^2}\sum_{i=1}^{n_v}\frac{\boldsymbol{v}_j-\boldsymbol{v}_i}{\|\boldsymbol{v}_j-\boldsymbol{v}_i\|}.\end{aligned}$$

By iteratively updating $U$ and $V$ via gradient descent, we progressively align their distributions:

$$\begin{pmatrix}U^{(e+1)}\\V^{(e+1)}\end{pmatrix}=\begin{pmatrix}U^{(e)}\\V^{(e)}\end{pmatrix}-\eta\begin{pmatrix}\nabla_U\mathcal{E}(P,Q)\\\nabla_V\mathcal{E}(P,Q)\end{pmatrix}, \tag{9}$$

where $\eta$ is the learning rate and $e$ indexes the iteration (epoch).

**Example 4.1** (Two sensitive groups)**.** For $G=2$ groups (e.g., gender), let $U=\hat{\pi}_{\text{female}}$ and $V=\hat{\pi}_{\text{male}}$, with $d=1$. Eq. (9) can be rewritten as:

$$\hat{\pi}^{\text{fair}(e+1)}=\begin{pmatrix}\hat{\pi}^{(e)}_{\text{female}}\\\hat{\pi}^{(e)}_{\text{male}}\end{pmatrix}-\eta\begin{pmatrix}\nabla_{\hat{\pi}_{\text{female}}}\mathcal{E}(P,Q)\\\nabla_{\hat{\pi}_{\text{male}}}\mathcal{E}(P,Q)\end{pmatrix}. \tag{10}$$

For $G>2$ groups, the premium vector can be partitioned as:

$$\hat{\pi}=\{\hat{\pi}_g\}_{g=1}^{G}=\begin{pmatrix}\{\hat{\pi}_g\}_{g=1}^{\lceil G/2\rceil}\\\{\hat{\pi}_g\}_{g=\lfloor G/2\rfloor+1}^{G}\end{pmatrix}=\begin{pmatrix}U\\V\end{pmatrix}. \tag{11}$$

Note that the ordering of dimensions is not meaningful. Dimensions can be treated as exchangeable.

**Example 4.2** (More than two sensitive groups)**.** Consider the protected attributes $\boldsymbol{D}$ include both the policyholder's gender $D_1\in\{\text{female, male}\}$ and his age group $D_2\in\{\text{young, middle-aged, senior}\}$, leading to $G=6$ distinct groups (e.g., young female, middle-aged male, etc.). Given that $G=6$, $U$ and $V$ are both 3-dimensional variables in $\mathbb{R}^3$, with $d=3$, where $\{u_{i,1}\}_{i=1}^{n_1}$ represents the young female premiums, $\{v_{j,1}\}_{j=1}^{n_{\lfloor G/2\rfloor+1}}$ the young male premiums, $\{u_{i,2}\}_{i=1}^{n_2}$ the middle-aged female premiums, etc.

Since the vectors $\hat{\pi}_g$ sizes $n_g$ may differ across groups $g\in(1,\ldots,G)$, we account for unequal group sizes by sampling equal-sized subsets from $U$ and $V$ at each iteration (see Proposition 3.2). Moreover, to ensure pairwise comparisons are made not only within

the same age category (e.g., comparing young female with young male, and middle-aged female with middle-aged male, etc.) but across different pairs (e.g., comparing young female with senior male), a uniform random permutation (Knuth, 1981), denoted $\tau$, is applied to $U$ (or $V$ equivalently) at each iteration $e$.

Algorithm 1 below is a pseudo-algorithm outlining the procedure. The premium predictions $\hat{\pi}$ are adjusted iteratively using gradient descent (via backpropagation) over multiple epochs until the fairness loss function is minimised. The gradient $\nabla\mathcal{L}_{\text{fair}} = \nabla\mathcal{E}$ is computed via automatic differentiation (using loss.backward in PyTorch).

**Algorithm 1:** $\mathcal{E}$-based fairness adjuster

**Data:** $\hat{\pi} = \{\hat{\pi}_g\}_{g=1}^{G}$

**Initialisation:** $\hat{\pi}^{\text{fair}}(0) \leftarrow \{\hat{\pi}_g\}_{g=1}^{G}$

**for** $e = 1$ **to** $E$ **do**

Sample predictions as in Eq. (11): $U^{\tau}(e-1) \xleftarrow{\tau} \{\hat{\pi}_g^{\text{fair}}(e-1)\}_{g=1}^{\lceil G/2 \rceil} \in \mathbb{R}^{n_u \times d}$

$V(e-1) \leftarrow \{\hat{\pi}_g^{\text{fair}}(e-1)\}_{g=\lfloor G/2 \rfloor+1}^{G} \in \mathbb{R}^{n_v \times d}$

Compute pairwise distances from Eq. (8):

$\text{dist}_{UY} \leftarrow \frac{1}{n_u n_v} \sum_{i=1}^{n_u} \sum_{j=1}^{n_v} \sqrt{\sum_{k=1}^{d} \left(u_{ik}(e-1) - v_{jk}(e-1)\right)^2}$

$\text{dist}_{UU} \leftarrow \frac{1}{n_u^2} \sum_{i=1}^{n_u} \sum_{j \neq i} \sqrt{\sum_{k=1}^{d} \left(u_{ik}(e-1) - u_{jk}(e-1)\right)^2}$

$\text{dist}_{VV} \leftarrow \frac{1}{n_v^2} \sum_{i=1}^{n_v} \sum_{j \neq i} \sqrt{\sum_{k=1}^{d} \left(v_{ik}(e-1) - v_{jk}(e-1)\right)^2}$

Compute Energy distance fairness loss: $\mathcal{L}_{\text{fair}}(e) \leftarrow 2 \cdot \text{dist}_{UV} - \text{dist}_{UU} - \text{dist}_{VV}$

Compute gradient $\nabla\mathcal{L}_{\text{fair}}(e)$ w.r.t. $U(e-1), V(e-1)$ using Eq. (9) via automatic differentiation and update predictions:

$$\hat{\pi}^{\text{fair}}(e) \leftarrow \begin{pmatrix} U(e) \\ V(e) \end{pmatrix} = \begin{pmatrix} U(e-1) \\ V(e-1) \end{pmatrix} - \eta \begin{pmatrix} \nabla_{U(e-1)}\mathcal{L}_{\text{fair}}(e) \\ \nabla_{V(e-1)}\mathcal{L}_{\text{fair}}(e) \end{pmatrix}$$

**end**

**return** $\hat{\pi}^{\text{fair}}(E)$

## 4.2 New policyholders

We further propose a mechanism to adjust predictions for new policyholders, so that we do not need to retrain the entire fairness adjuster. For each policyholder in the set used to train the fairness adjuster model, an individual fairness correction factor $\gamma_i$ is defined as:

$$\gamma_i = \frac{\hat{\pi}_i^{\text{fair}}}{\hat{\pi}_i},$$

where $\hat{\pi}_i$ is the original predicted premium before fairness adjustment, and $\hat{\pi}_i^{\text{fair}}$ is the fairness-adjusted premium. This correction factor $\gamma_i$ quantifies the individual adjustment required for fairness. We train a random forest regressor $f_\theta$ to predict $\gamma_i$ based on the

policyholder's features $(\boldsymbol{A}_i, \boldsymbol{D}_i)$:

$$\hat{\gamma}_i = f_\theta\left(\boldsymbol{A}_i, \boldsymbol{D}_i\right).$$

This model learns how different features influence fairness adjustments. For a new policyholder with features $(\boldsymbol{A}_{\text{new}}, \boldsymbol{D}_{\text{new}})$, the fairness correction factor is predicted as:

$$\hat{\gamma}_{\text{new}} = f_\theta(\boldsymbol{A}_{\text{new}}, \boldsymbol{D}_{\text{new}}).$$

The initial regression model (e.g., LightGBM) predictions $\hat{\pi}_{\text{new}}$ are then adjusted using this factor:

$$\hat{\pi}_{\text{new}}^{\text{fair}} = \hat{\pi}_{\text{new}} \cdot \hat{\gamma}_{\text{new}}.$$

As showed in Section 5.3, this mechanism assigns to new policyholders premiums that align with the fairness correction learned during the fairness adjuster training.

## 4.3 Post-hoc autocalibration

Finally, note that the regression models (particularly tree-based boosting) used to predict the premium best-estimates $\hat{\pi}\left(\boldsymbol{A}, \boldsymbol{D}\right)$ often exhibit imbalances between predicted and actual claims total. Such discrepancies can adversely affect the financial stability of an insurance portfolio.

A model is globally unbiased when the expected value of the predicted premiums matches the expected value of the actual outcomes $Y$:

$$\mathbb{E}\left[\hat{\pi}\left(\boldsymbol{A}, \boldsymbol{D}\right)\right] = \mathbb{E}\left[Y\right].$$

A model is locally unbiased when, for every specific value $s$ of the predicted premiums:

$$\mathbb{E}\left[Y | \hat{\pi}\left(\boldsymbol{A}, \boldsymbol{D}\right) = s\right] = s.$$

Deviations from these conditions indicate that certain segments may be under-priced or overpriced, with substantial implications for insurance pricing strategies.

Denuit et al. (2021) address these calibration issues by proposing a local regression technique to correct model bias. The expected actual loss $\mathbb{E}\left[Y | \hat{\pi}\left(\boldsymbol{A}, \boldsymbol{D}\right)\right]$ is first estimated using local regression techniques on $\{(\hat{\pi}\left(\boldsymbol{a}_i, \boldsymbol{d}_i\right), y_i)\}$. The predictions $\hat{\pi}$ are adjusted with a multiplicative factor $\lambda_\alpha(s)$ derived from the local regression estimates:

$$\lambda_\alpha(s) = \frac{\mathbb{E}\left[Y | \hat{\pi}\left(\boldsymbol{A}, \boldsymbol{D}\right) = s\right]}{s}.$$

The balance corrected premiums $\hat{\pi}^{\text{BC}}$ are then given by:

$$\hat{\pi}^{\text{BC}} = \lambda_\alpha\left(\hat{\pi}\left(\boldsymbol{a}, \boldsymbol{d}\right)\right) \cdot \hat{\pi}\left(\boldsymbol{a}, \boldsymbol{d}\right).$$

As illustrated in Section D, the balance correction is applied after the fairness adjustment, thereby ensuring that the final adjusted premiums accurately reflect the underlying risk across different segments:

$$\hat{\pi}^{\text{fair, BC}} = \lambda_{\alpha}^{\text{fair}}\left(\hat{\pi}^{\text{fair}}\right) \cdot \hat{\pi}^{\text{fair}} .$$

# 5. Results

In this Section 5, we empirically evaluate the approach proposed in Section 4. In Section 5.1, we first introduce the datasets and their key characteristics. In Section 5.2, we model the premium best-estimates $\hat{\pi}(\boldsymbol{A}, \boldsymbol{D})$ using a tree-based boosting model (LightGBM). In Sections 5.3 and 5.4, we adjust the premium best-estimates for fairness across two and six protected demographic groups, and correct for the balance bias.

## 5.1 Datasets

To evaluate the proposed methodology, we first simulate a motor insurance portfolio $\mathcal{X}$ of $n = 100\,000$ policies. Table 1 displays a sample of five policies, where each column corresponds to a rating factor. In this case study, the policyholder's gender and age, two key risk factors in motor insurance, are considered as prohibited variables $\boldsymbol{D} = (D_1, D_2)$. For each policy, we observe the number of claims $N_i \in \{0, 1, 2, 3\}$ reported during its contract duration (or exposure) $\nu_i \in (1, 365)$ measured in days. As the majority of policies report zero claims (see Table B1 in Appendix), the count variable $N$ is highly zero-inflated. Throughout the analysis, we work with the claim frequency per unit of exposure $\lambda_i = N_i/\nu_i$. Note that we assume that the models predictions $\hat{\lambda}$ (per unit of exposure), are the so-called premium best-estimates $\hat{\pi}(\boldsymbol{A}, \boldsymbol{D})$.

To gain further insight into the dataset, Fig. B1 and B2 in Appendix illustrate the distributions of the rating factors, presented in Table 1, with respect to the variable of interest $N$. Fig. B1 reports average claim counts across various categories, while Fig. B2 shows the empirical distributions of age and income across $N \in \{0, 1, 2, 3\}$.

Note that the synthetic portfolio was generated using a combination of marginal distributions and dependency structures. For example, vehicle power class and vehicle type were generated using gender-dependent probabilities, while geographic location was generated conditionally on ethnicity. Income was constructed as a function of age, education, gender, ethnicity, location, vehicle characteristics, and their associated multiplicative effects, creating a network of dependencies among variables. Interaction effects were also deliberately introduced when generating claim frequencies. Examples include Gender $\times$ Age, Age $\times$ Vehicle Type, Vehicle Power Class $\times$ Age, Geographic Location $\times$ Vehicle Power Class, and Education $\times$ Income. These interactions create non-linear relationships between protected and non-protected variables, such that seemingly legitimate rating factors become statistically associated with protected attributes. As a result, several risk factors act as proxy variables for protected characteristics. For instance, vehicle type

and power class indirectly encode information about gender, while geographic location partially reflects ethnicity. Because claim frequencies depend on these interacting factors, discrimination can emerge indirectly even when protected attributes are excluded from the pricing model. This design provides a realistic setting for evaluating fair actuarial models, as it reproduces both direct and indirect pathways through which sensitive information can influence premium predictions.

| | Income | Age | Power | Education | Location | Type | Gender | Ethnicity |
|---|---|---|---|---|---|---|---|---|
| 1 | 37047.85 | 27.48 | Low | High School | Urban | Sports | M | American |
| 2 | 14426.15 | 24.31 | High | Bachelor | Suburban | SUV | M | Asian |
| 3 | 25834.59 | 28.24 | Medium | High School | Suburban | SUV | F | European |
| 4 | 51670.00 | 32.62 | Low | Master | Urban | Sports | M | American |
| 5 | 28547.69 | 26.45 | High | PhD | Rural | Sedan | F | African |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

Table 1: Illustrative sample of five policies from the simulated motor insurance portfolio $\mathcal{X}$. Each row represents an individual policyholder characterised by continuous covariates (policyholder's income and age) and categorical attributes (vehicle's power, policyholder's education and geographic location, vehicle's type, policyholder's gender and ethnicity).

To evaluate whether the gradient-based fairness correction maintain its effectiveness beyond controlled simulation settings, we also consider a real-world car insurance portfolio publicly available at Roy (2021). After removing observations with missing values, the working dataset consists of 8 149 complete policies. Each contract is assigned a unit exposure $\nu_i = 1$, and the observed number of past accidents is used as the claim count, denoted $N_i$. Accordingly, we define the claim frequency as $\lambda_i = N_i/\nu_i$, which we model under a Poisson specification consistent with classical actuarial frequency modelling.

Table 2 presents five representative observations to illustrate the structure of the dataset. Each policy is described by two continuous variables (credit score and annual mileage) and a set of categorical attributes capturing demographic, socio-economic, and vehicle-related characteristics. These include the policyholder's gender and race (the sensitive attributes in our analysis), age, driving experience, education level, income class, household composition, postal code, vehicle type, and traffic violations (speeding and DUI).

| | Age | Gender | Race | Driving Exp. | Education | Income | Credit Score | ⋯ |
|---|---|---|---|---|---|---|---|---|
| 1 | 65+ | female | majority | 30y+ | university | upper class | 0.592 | ⋯ |
| 2 | 40–64 | female | majority | 20–29y | high school | middle class | 0.654 | ⋯ |
| 3 | 26–39 | female | minority | 10–19y | none | poverty | 0.422 | ⋯ |
| 4 | 16–25 | female | majority | 0–9y | high school | poverty | 0.344 | ⋯ |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

Table 2: Illustrative sample of four policies from the car insurance portfolio publicly available at Roy (2021).

For each dataset, we split the observations into three disjoint sets. A GLM and a LightGBM are trained on 60% of the data (in-sample training set) to obtain baseline predictive models for discriminatory premiums. The trained GLM and LightGBM models are used to predict discriminatory premiums on a first out-of-sample test set (with 20% observations). These discriminatory premiums are then corrected for fairness and balance (post-hoc autocalibration). A Random Forest model is finally trained on the first out-of-sample set to approximate the mapping from features to corrected premiums. This allows prediction of fairness-adjusted premiums for a held-out second out-of-sample test set (also with 20% observations), representing new policyholders not used in the training or correction stages.

The analysis presented in the following sections on the simulated data is reproduced for the real-world dataset in Appendix F. As in the simulated study, enforcing demographic parity leads to a clear fairness-accuracy trade-off. While the adjustment successfully reduces dependence between premiums and protected characteristics, it deteriorates predictive performance and increases claim count overestimation. The loss in accuracy is particularly pronounced when demographic parity is enforced jointly across gender and ethnicity, as the (already limited) number of observations are further partitioned into smaller protected groups, reducing the reliability of the estimated group-specific distributions. The subsequent autocalibration step partially recovers actuarial balance and predictive accuracy without affecting demographic parity.

## 5.2 Best-estimates

We consider a Generalized Linear Model (GLM) and a boosting algorithm (LightGBM) to model the premium best-estimates $\hat{\pi}$. The boosting model is tuned via `lightgbm` and `optuna`, with optimal hyper-parameters reported in Table C1.

Fig. 5.1 compares predicted and observed claim frequencies across 10 quantile bins. Bars represent total exposure within each bin, while the solid and dashed lines show the predicted $\hat{\lambda}$ and observed $\lambda$ frequencies, respectively. A closer alignment between the two lines across bins indicates superior model performance. The GLM (left plot)

systematically under- and over-predicts frequencies in several bins, whereas LightGBM (right plot) aligns much more closely with the empirical frequencies across all bins.

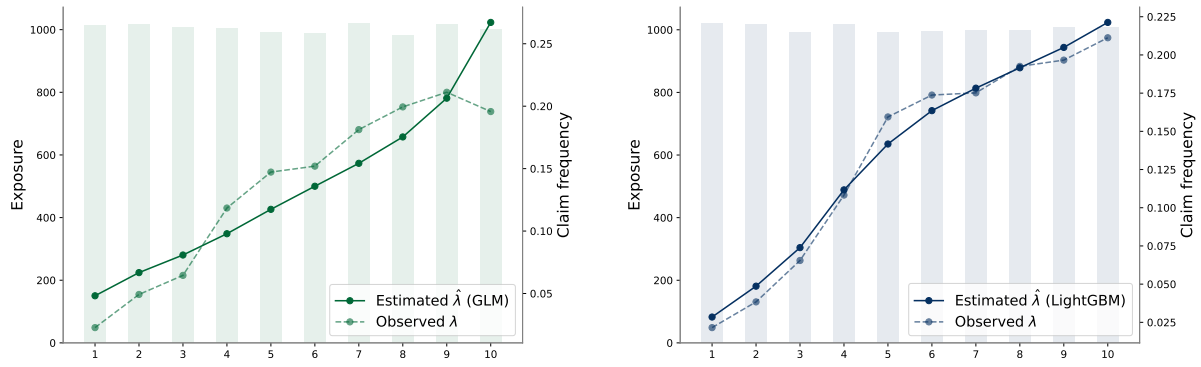


Figure 5.1: Comparison of predicted and observed claim frequencies per unit exposure across 10 quantile-based bins of predicted frequency. The solid line shows the estimated frequency $\hat{\lambda}$ from the model, while the dashed line shows the observed frequency $\lambda$ computed from the data. The bars indicate the total exposure in each bin. Left plot: GLM; right plot: LightGBM.

The Poisson deviance (a likelihood-based measure of predictive accuracy) is widely used to evaluate claim count regression models. It is defined as:

$$\mathcal{D}_{\text{Poisson}}\left(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}}\right) = 2\sum_{i=1}^{n} N_i \left(\frac{\nu_i}{N_i}\hat{\lambda}_i - \left(\ln\left(\frac{\nu_i}{N_i}\hat{\lambda}_i\right) + 1\right)\mathbb{I}_{\{N_i \geq 1\}}\right), \tag{12}$$

where $N_i$, $\nu_i$, and $\hat{\lambda}_i$ are respectively the $i$th policyholder's observed number of claims, observed exposure, and estimated claim frequency (per unit of exposure).

Table 3 confirms that LightGBM has a lower deviance than the GLM, reflecting its ability to better capture the variation in the data by modelling nonlinear relationships and interactions more effectively than GLM (which is restricted to linear relationships unless explicitly specified).

| Model | Poisson deviance |
|---|---|
| GLM | 6625.6248 |
| LightGBM | 6539.1550 |

Table 3: Comparison of Poisson deviance values on the first out-of-sample test set for the GLM and the LightGBM. Lower deviance indicates a better fit to the observed claim frequencies.

Fig. 5.2 further illustrates the distribution of predicted premiums and highlight differences in risk allocation between the two models.

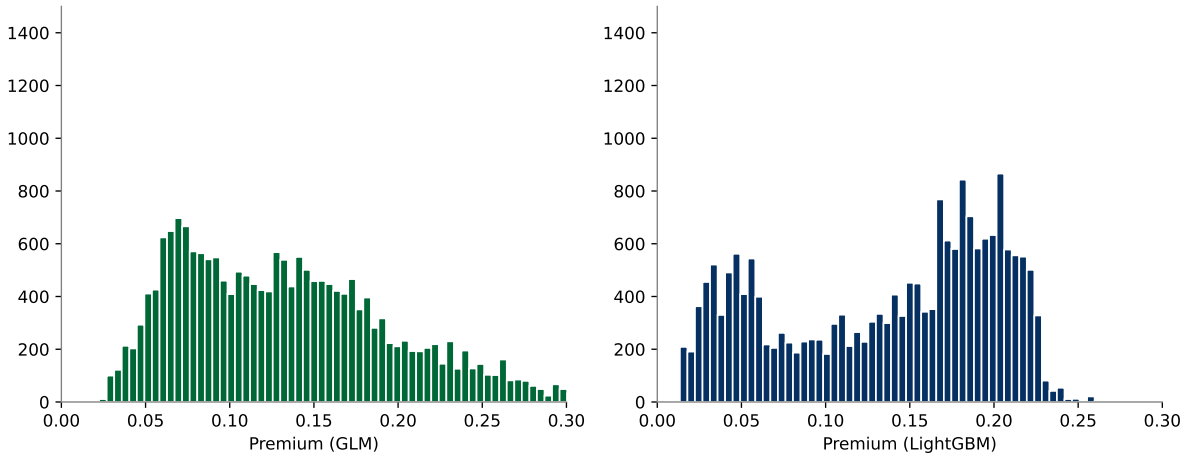


Figure 5.2: Histogram comparison of predicted premiums $\hat{\pi}$ for the GLM (left) and LightGBM (right). The plots illustrate the distribution of predicted premiums and highlight differences in risk allocation between the two models.

Given the models predictions $\hat{\boldsymbol{\lambda}}$, the probability of observing exactly $c$ claim(s) is given by Poisson probability mass function (pmf):

$$f(c, \lambda_i) = \mathbb{P}(N_i = c) = \frac{(\lambda_i)^c e^{-\lambda_i}}{c!},$$

with $\mathbb{P}(N_i > 0) = 1 - \mathbb{P}(N_i = 0) = 1 - e^{-\lambda_i}$. Summing these probabilities over all policies yields the expected total claim count for each claim level $c$. Table 4 compares observed and predicted counts. Although LightGBM better fits the quantile patterns (see Fig. 5.1) and deviance metrics (see Table 3), it overestimates by 13 the total number of policies with at least one claim $c > 0$. This overestimation bias is a well-known consequence of minimising the Poisson deviance outside the canonical GLM setting (Denuit et al., 2021).

| $c$ | Observed | GLM | LightGBM | GLM gap | LightGBM gap |
|---|---|---|---|---|---|
| 0 | 18719 | 18717.5133 | 18705.6169 | -1.4867 | -13.3831 |
| 1 | 1221 | 1213.9650 | 1224.7363 | -7.0350 | 3.7363 |
| 2 | 56 | 65.2688 | 66.5456 | 9.2688 | 10.5456 |
| 3 | 4 | 3.1122 | 2.9848 | -0.8878 | -1.0152 |
| $> 0$ | 1281 | 1282.4867 | 1294.3831 | 1.4867 | 13.3831 |

Table 4: Comparison between observed and predicted claim counts for the LightGBM and GLM. The column $c$ denotes the number of claims per policy, and the gap columns show the difference between predicted and observed counts within each group. The aggregated row ($c > 0$) summarises the total predicted versus observed claim counts for policies with at least one claim. Highlighted value indicates the total gap.

We can further check the estimation bias by running a simple Poisson regression of the

intercept ($\hat{\beta}_0 = -2.0023$) and consider its prediction per unit of exposure $\hat{\lambda} = \overline{\pi} = 0.13502$ as the true premium mean. In Table 5, we compare it with the GLM and LightGBM average premium values $\overline{\pi} = \frac{1}{n}\sum_{i=1}^{n} \hat{\lambda}_i$ (Denuit et al., 2021). While both models are close to the true premium mean $\overline{\pi} = 0.13502$, LightGBM does slightly overestimate the average premium.

| Statistic | $\text{GLM}_{\beta_0}$ | GLM | LightGBM |
|---|---|---|---|
| Mean $\overline{\pi}$ | 0.1350 | 0.1350 | 0.1363 |
| 10% Quantile | – | 0.0602 | 0.0391 |
| 90% Quantile | – | 0.2288 | 0.2115 |

Table 5: Summary statistics of the expected premiums $\hat{\pi}$ for the baseline intercept-only GLM ($\text{GLM}_{\beta_0}$), the full GLM, and the LightGBM.Missing quantiles for $\text{GLM}_{\beta_0}$ are denoted by – because it produces a constant expected premium across policies. The highlighted value indicates that the LightGBM's mean prediction diverges from the baseline mean.

The negative bias (or overestimation) observed in the LightGBM predictions can be corrected with a multiplicative factor $\mathbb{E}\left[Y|\hat{\pi}\left(\boldsymbol{X}\right) = F_{\hat{\pi}}^{-1}\left(u\right)\right] / F_{\hat{\pi}}^{-1}\left(u\right)$ proposed by Denuit et al. (2021). This balance correction, detailed in Appendix D, is as key post-hoc step of the fairness adjustment presented in the following Sections 5.3 and 5.4.

The best-estimate premiums $\hat{\pi}\left(\boldsymbol{A}, \boldsymbol{D}\right)$, presented in Section 5.2 and in Appendix D, are discriminatory with respect to sensitive attributes such as gender. Following the theoretical framework outlined in Section 4, our goal is to enforce demographic parity by aligning the distributions of predicted premiums across $G \geq 2$ demographic groups. We first illustrate this mechanism for a binary sensitive attribute in Section 5.3, and extend the approach to settings where $G > 2$ in Section 5.4.

## 5.3 Fairness adjustment G=2

In the binary setting $G = 2$, the predicted (discriminatory) premiums $\hat{\pi}$ are partitioned into two demographic groups $\{\hat{\pi}_g\}_{g=1}^{G}$ corresponding to $g = \{\text{Female}, \text{Male}\}$. These premiums are iteratively adjusted using algorithm 1 to equalise not only the means, but also the dispersion and overall shape of their distributions.

Fig. 5.3 illustrates the distributions of predicted premiums before and after fairness adjustment across the two demographic groups. Solid lines correspond to the original discriminatory premiums, while dashed lines represent the premiums after fairness correction. Prior to adjustment, the distributions clearly differ between genders, with females (in red) concentrated in the lower-premium range and males (in blue) in the higher range, reflecting gender-based pricing differences. After fairness adjustment, the distributions for both groups nearly overlap, demonstrating that our Energy-based fairness correction

effectively mitigates gender-based disparities.

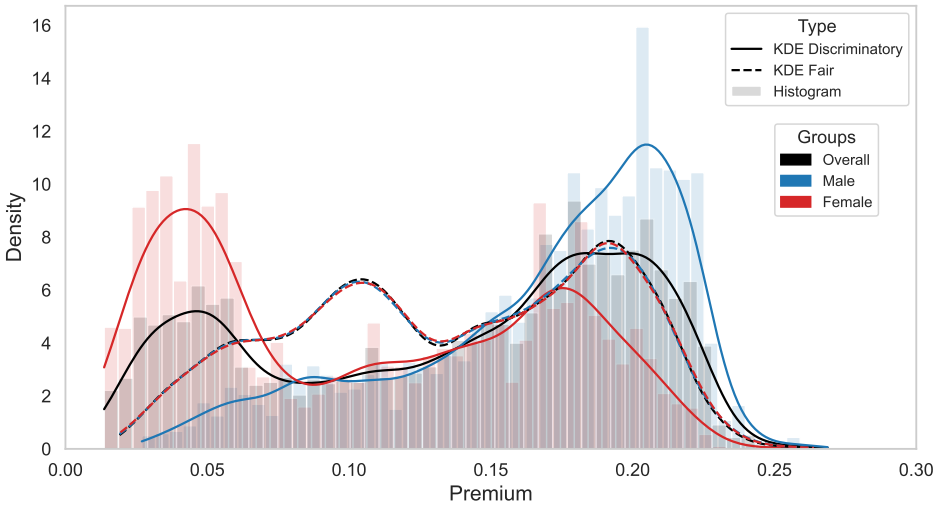


Figure 5.3: Kernel density estimates (KDEs) of predicted premiums by gender. Solid lines represent the original LightGBM discriminatory premiums, while dashed lines correspond to the fairness-adjusted premiums. The male group is shown in blue, the female group in red, and the overall portfolio in black.

Enforcing independence between premiums and gender (a key risk factor) constrains the predictions and prevents the model from fully exploiting gender-related risk information. As illusrated in Fig. 5.4, the model inevitably sacrifices some predictive accuracy to enforce demographic parity. As later detailed in Table 10, the deterioration in predictive accuracy remains however modest, with a +0.62% increase in deviance relative to the GLM benchmark.

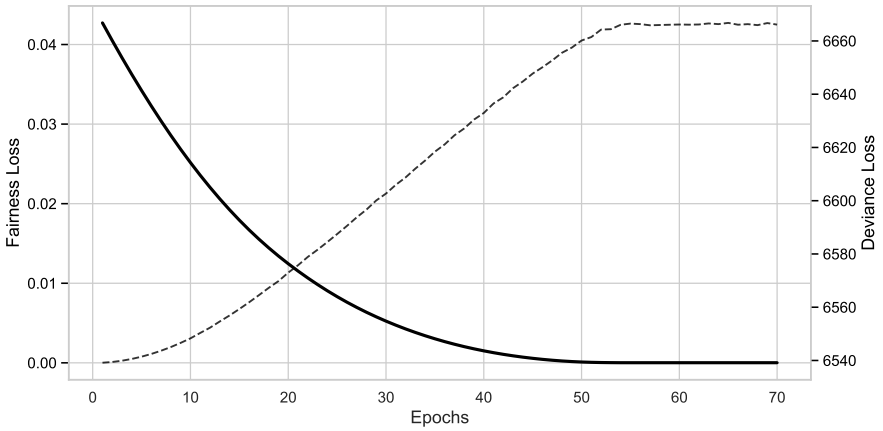


Figure 5.4: Evolution of the fairness loss (in Eq. (7)) and the Poisson deviance loss (in Eq. (12)) over training epochs.

While the fairness adjustment inevitably alters the original premium distribution, Fig. E1 in Appendix illustrates the crucial role of the intra-group terms $\mathbb{E}[\|U - U'\|]$ and $\mathbb{E}[\|V - V'\|]$ in our fairness adjustement. Omitting these terms collapses the distributional spread (see last column), and leads to highly concentrated premium with unnaturally

sharp and narrow peaks. This results in a poor model accuracy, as the fairness-adjusted premiums do not respect the natural distributional spread of the initial (discriminatory) premiums $\hat{\pi}$. Including these intra-group terms (see middle column) preserves the predictions dispersion, reflecting the model's capacity to actuarially discriminate between policyholders, thereby ensuring a meaningful demographic parity adjustment, and highlighting the suitability of the Energy distance for this type of distributional fairness correction.

Tables 7 and 6 further reveal that the fairness adjustment also leads to an increase in the overestimation of the overall claim count (from 13 to 19), and to a larger deviation from the true mean $\mu = 0.13502$.

| Statistic | $\text{GLM}_{\beta_0}$ | LightGBM | Fair LightGBM |
|---|---|---|---|
| Mean $\overline{\pi}$ | 0.1350 | 0.1363 | 0.1366 |
| 10% Quantile | – | 0.0391 | 0.0587 |
| 90% Quantile | – | 0.2115 | 0.2050 |

Table 6: Summary statistics of expected premiums $\hat{\pi}$ for the baseline intercept-only GLM ($\text{GLM}_{\beta_0}$), the standard LightGBM, and the fairness-adjusted LightGBM. Highlighted values indicate divergence from the target mean premium, indicating the effect of fairness adjustment.

| $c$ | Observed | LightGBM | Fair LightGBM | LightGBM gap | Fair LightGBM gap |
|---|---|---|---|---|---|
| 0 | 18719 | 18705.6169 | 18699.0905 | -13.3831 | -19.9095 |
| 1 | 1221 | 1224.7363 | 1234.2958 | 3.7363 | 13.2958 |
| 2 | 56 | 66.5456 | 63.7865 | 10.5456 | 7.7865 |
| 3 | 4 | 2.9848 | 2.7249 | -1.0152 | -1.2751 |
| $> 0$ | 1281 | 1294.3831 | 1300.9095 | 13.3831 | 19.9095 |

Table 7: Comparison of observed claim counts with predictions from standard LightGBM and fairness-adjusted LightGBM. The gap columns show the difference between predicted and observed counts. Highlighted values indicate the aggregated total gaps, illustrating the impact of the fairness adjustment on total balance.

In Fig. 5.5, a noticeable drop below the diagonal is observed (right plot) for higher levels of $s$, which may explain the increased negative bias detected after fairness adjustment in Table 6. A detailed view of the local bias is provided in Fig. E2 (in Appendix).

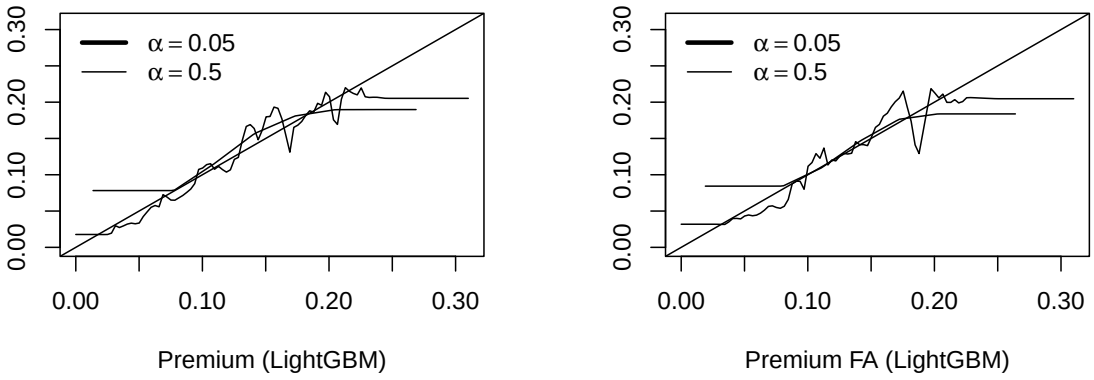


Figure 5.5: Visualisation of $s \mapsto \mathbb{E}\left[Y|\hat{\pi}\left(\boldsymbol{X}\right) = s\right]$ before and after fairness adjustment (FA) across the gender variable.

The negative bias (or overestimation) observed in the fair LightGBM predictions can be corrected using the multiplicative bias-correction factor proposed by Denuit et al. (2021) detailed in Appendix D:

$$\frac{\mathbb{E}\left[Y|\hat{\pi}\left(\boldsymbol{X}\right) = F_{\hat{\pi}}^{-1}\left(x\right)\right]}{F_{\hat{\pi}}^{-1}\left(x\right)},$$

and estimated using a local GLM smoother with bandwidth $\alpha = 0.16$.

Table F5 reports summary statistics before and after bias correction. The average premium shifts from 0.1366 to 0.1351, bringing it close to the reference mean 0.13502.

| Statistic | $\text{GLM}_{\beta_0}$ | LightGBM | Fair LightGBM | Fair BC LightGBM |
|---|---|---|---|---|
| Mean $\overline{\pi}$ | 0.1350 | 0.1363 | 0.1366 | 0.1351 |
| 10% Quantile | – | 0.0391 | 0.0587 | 0.0461 |
| 90% Quantile | – | 0.2115 | 0.2050 | 0.2035 |

Table 8: Summary statistics of expected premiums $\hat{\pi}$ for the baseline intercept-only GLM ($\text{GLM}_{\beta_0}$), standard LightGBM, fairness-adjusted LightGBM, and bias-corrected fairness-adjusted LightGBM. Highlighted values indicate divergence from the target mean premium, illustrating the effect of fairness adjustment and subsequent bias correction.

Table 9 shows the improvement in predicted claim counts. The gap in the number of policies with at least one claim drops from 19.91 to 5.55.

| $c$ | Observed | Fair LightGBM | Fair BC LightGBM | Fair LightGBM gap | Fair BC LightGBM gap |
|---|---|---|---|---|---|
| 0 | 18719 | 18699.0905 | 18713.4489 | -19.9095 | -5.5511 |
| 1 | 1221 | 1234.2958 | 1221.2914 | 13.2958 | 0.2914 |
| 2 | 56 | 63.7865 | 62.5397 | 7.7865 | 6.5397 |
| 3 | 4 | 2.7249 | 2.6244 | -1.2751 | -1.3756 |
| $> 0$ | 1281 | 1300.9095 | 1286.5511 | 19.9095 | 5.5511 |

Table 9: Comparison of observed claim counts with predicted claims from the fairness-adjusted LightGBM before and after bias correction. The gap columns show the difference between predicted and observed counts. Highlighted values indicate the aggregated total gaps, illustrating the reduction of prediction bias after applying balance correction.

Table 10 summarises the Poisson deviance values. We note that the fairness–accuracy trade-off manifests at two distinct levels. First, a slight degradation in predictive accuracy (individual deviance fit) from 6539.155 to 6666.717 immediately after the fairness adjustment. The fairness correction inevitably reduces predictive accuracy to some extent, as the optimisation process prioritizes demographic parity over likelihood-based fit. Enforcing independence between premiums and gender (a key risk factor) further prevents the model from fully exploiting gender-related risk information. Second, a minor imbalance in aggregate calibration (alignment between predicted and observed total claims shown in Table 7). To address the latter, we introduced an additional post-hoc step, which partially restores, not only the actuarial balance, but also the predictive accuracy (6654.98) that was temporarily reduced by the fairness correction, without affecting fairness, as shown in Fig. 5.6.

| Model | Poisson Deviance | $\Delta$ GLM | $\Delta$ LightGBM |
|---|---|---|---|
| GLM | 6625.6248 | | |
| LightGBM | 6539.1550 | $-1.3051\%$ | |
| LightGBM BC | 6542.6450 | | |
| LightGBM fair (gender) | 6666.7170 | $+0.6202\%$ | $+1.9507\%$ |
| LightGBM fair BC (gender) | 6654.9800 | $+0.4430\%$ | $+1.7713\%$ |

Table 10: Poisson deviance comparison at the same evaluation stage (out-of-sample 1) for the generalised linear model (GLM), bias-corrected (BC), and fairness-adjusted versions of the LightGBM model. Lower deviance indicates a better fit to the observed claim frequencies. $\Delta$ percentages compare each fairness (and balance) corrected model relative deterioration with the GLM and LightGBM.

Fig. 5.6 shows the local redistribution of density after balance correction (dotted lines).

It further confirms that, even after balance correction, fairness is maintained as the conditional premium densities for both gender groups remain well-aligned.

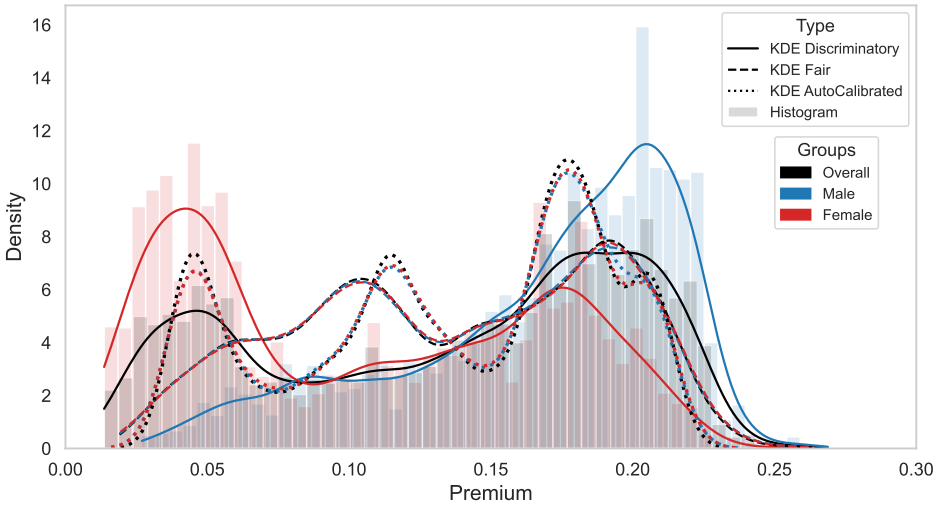


Figure 5.6: Kernel density estimates (KDEs) of predicted premiums by gender. Solid lines represent the original LightGBM discriminatory premiums, dashed lines the fairness-adjusted premiums, and dotted lines the fairness-adjusted and balance-corrected premiums. The male group is shown in blue, the female group in red, and the overall portfolio in black.

Table 11 summarises several keys fairness metrics discussed in the literature. These include the Wasserstein distance $W_p(U,V)$ extensively studied by Charpentier (2023), Charpentier et al. (2023) and Hu et al. (2023a, 2023b), and inspired by Chzhen et al. (2020), the MMD and Sinkhorn divergence $S_\epsilon(U,V)$ proposed by Oneto et al. (2020), as well as widely adopted metrics in the statistical and causal theory (see e.g., Charpentier (2023) and Deka and Sutherland (2023)) such as the Group Fairness in Expectation $\Delta_{GFE}$ and Statistical Parity $\Delta_{SP}$. We further provide the Wasserstein barycenter and MMD-based adjustments benchmarks for comparison in Appendix G. The comparison with the Wasserstein benchmark highlights important differences in how demographic parity is enforced. Although both methods reduce group disparities, the Wasserstein correction achieves parity through quantile-level redistribution, resulting in larger distortions among high-risk policyholders and weaker rank preservation. In contrast, the proposed energy-based adjustment results in larger distortions among lower-risk policyholders, better preserves the original risk ranking, and achieves stronger conditional parity for high-risk policyholders.

| Metric with $U = \hat{\pi}_{\text{Female}}$ and $V = \hat{\pi}_{\text{Male}}$ | | $\hat{\pi}$ | $\hat{\pi}^{\text{fair, BC}}$ |
|---|---|---|---|
| $\mathcal{E}(U,V) = 2\mathbb{E}\Big[\|U-V\|\Big] - \mathbb{E}\Big[\|U-U'\|\Big] - \mathbb{E}\Big[\|V-V'\|\Big]$ | | 0.042704 | 8e-06 |
| $W_p(U,V) = \left(\inf_{\gamma\in\Gamma(U,V)} \mathbb{E}_{(U,V)\sim\gamma}\|U-V\|^p\right)^{1/p}$ | $(p=1)$ | 0.064760 | 0.001529 |
| | $(p=2)$ | 0.070531 | 0.002113 |
| $\text{MMD}^2(U,V) = \mathbb{E}[k(U,U')] + \mathbb{E}[k(V,V')] - 2\mathbb{E}[k(U,V)]$ | | 0.205325 | 2e-06 |
| $S_{\epsilon=0.1}(U,V) = W_\epsilon(U,V) - \frac{1}{2}W_\epsilon(U,U) - \frac{1}{2}W_\epsilon(V,V)$ | | 0.028611 | 2.2e-05 |
| $\Delta_{GFE} = \|\mathbb{E}[U] - \mathbb{E}[V]\|$ | | 0.062959 | 0.000155 |
| $\Delta_{SP} = \sup_s \|P(U \le s) - P(V \le s)\|$ | | 0.396203 | 0.030354 |

Table 11: Comparison of key fairness metrics before and after $\mathcal{E}$-based fairness adjustment and balance correction. MMD is computed using a Gaussian RBF kernel $k(u,v) = \exp\left(-\|u-v\|^2/\left(2\sigma^2\right)\right)$ with fixed bandwidth $\sigma = 0.1$.

For new policyholders, we extend this methodology by applying an individual correction factor $\gamma_i$, as outlined in Section 4.2. A random forest regressor learns the mapping from policyholder characteristics to correction factors, generalising the fairness adjustment mechanism beyond the dataset (out-of-sample 1) used to train the adjuster.

Fig. 5.7 shows that the adjusted predictions for new policyholders (out-of-sample 2) remain consistent across sensitive groups with the results observed in Fig. 5.6, with only minor deviations.

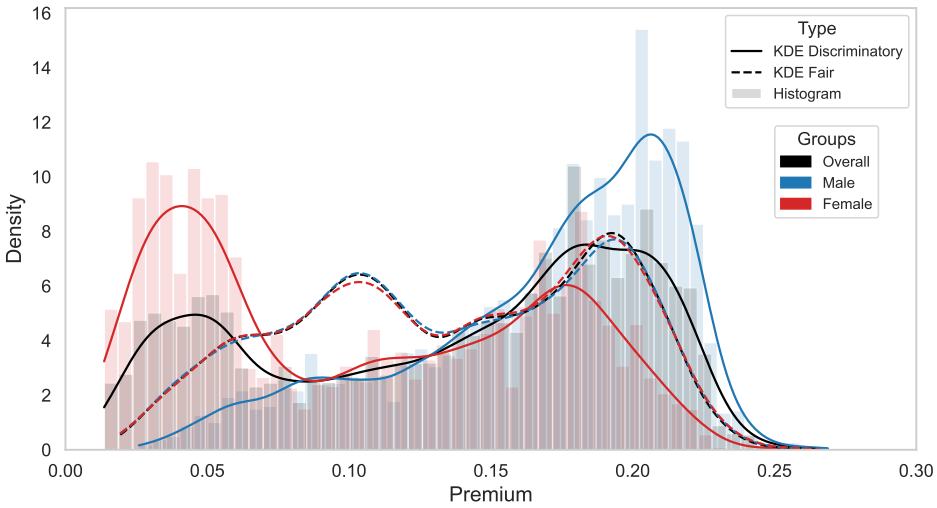


Figure 5.7: Kernel density estimates (KDEs) of predicted premiums by gender for new policyholders. Solid lines represent the LightGBM discriminatory premiums, and dashed lines the fairness-adjusted premiums. The male group is shown in blue, the female group in red, and the overall portfolio in black.

Table 12 summarises the Poisson deviance results before and after fairness correction on the new test set (out-of-sample 2) with 20 000 policies. The fair LightGBM

model exhibits slightly higher deviance than the unconstrained LightGBM on both out-of-sample sets (+1.95% and +2.60% relative to LightGBM), consistent with typical accuracy–fairness trade-offs. All models, including GLM and baseline LightGBM, deteriorate when predicting genuinely new policyholders ($\Delta$ of +3.8–4.13%), indicating that a portion of the deviance increase for the new policyholders (from +1.95% to +2.60%) stems from intrinsic out-of-sample difficulty rather than instability introduced by the fairness mechanism (Hainaut, 2025).

| | in-sample | out-of-sample 1 | out-of-sample 2 | $\Delta$ |
|---|---|---|---|---|
| GLM | 19997.9223 | 6625.6248 | 6880.0621 | +3.8102% |
| LightGBM | 19539.5379 | 6539.1550 (-1.3051%) | 6809.2512 (-1.0292%) | +4.1304% |
| LightGBM fair | – | 6666.7170 (+1.9507%) (+0.6202%) | 6986.1686 (+2.5982%) (+1.5422%) | +4.7917% |

Table 12: Poisson deviance comparison across models: (i) in-sample (training) set ($n$ = 60,000) to fit the discriminatory GLM and LightGBM models, (ii) a first out-of-sample (test) set ($n$ = 20,000) to predict discriminatory premiums and correct for fairness and autocalibration, and (iii) a second out-of-sample (test) set ($n$ = 20,000) of new policyholders, for which corrected premiums were predicted using a Random Forest trained on the first test set. Percentages in parentheses compare each model with the GLM and LightGBM at the same evaluation stage. $\Delta$ measures the relative deterioration when moving from the first to the second out-of-sample (new policyholders).

## 5.4 Fairness adjustment G>2

As discussed in Section 4, our $\mathcal{E}$-adjuster can simultaneously adjust premiums across more than two sensitive groups. For $G > 2$, the predicted premiums $\hat{\pi}$ are partitioned into $G$ vectors $\{\hat{\pi}_g\}_{g=1,\dots,G}$ of sizes $n_g$ organized in $d$-dimensional variables $U$ and $V$ as in Eq. (11). In this setting, equal-sized samples are drawn from each group in $U$, and for each group in $V$. A uniform random permutation is also applied to ensure the pairwise comparisons span all group combinations (see algorithm 1).

Fig. 5.8 illustrates the fairness adjustment of $\hat{\pi}$ across six demographic groups, determined by the policyholders' gender (female or male) and age (young, middle-aged, or senior). The solid lines correspond to the initial predicted premiums partitioned into $G$ demographic groups $\{\hat{\pi}_g\}_{g=1}^{6}$. The overlapping dashed lines correspond to the corresponding fairness adjusted premiums $\{\hat{\pi}_g^{\text{fair}}\}_{g=1}^{6}$. We observe that after adjustment, the dashed lines largely overlap within each gender-age pairing, indicating that the $\mathcal{E}$-adjuster has effectively reduced disparity across these six groups, while preserving the spread of the initial distribution (in black).

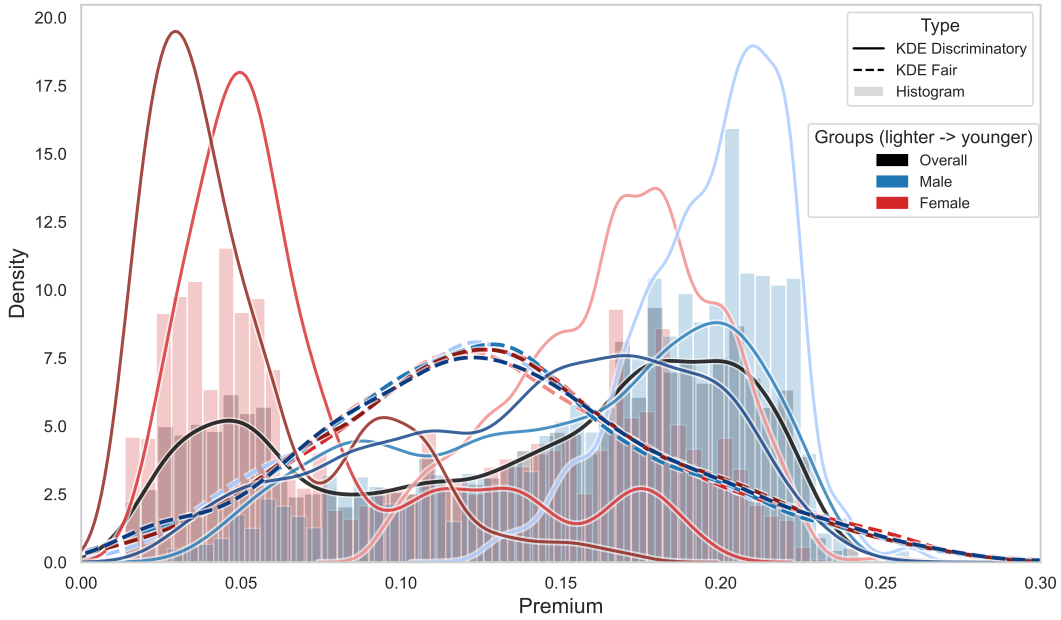


Figure 5.8: Kernel density estimates (KDEs) of predicted premiums by gender and age group. Solid lines represent the LightGBM discriminatory premiums, and dashed lines the fairness-adjusted premiums. The male groups are shown in blue, the female groups in red, and the overall portfolio in black. The color shades correspond to age groups. Histogram correspond to the overall empirical frequencies of female, male, and overall discriminatory premiums.

Tables 13 and 14 reveal that the fairness adjustment leads to a significant underestimation (around $-42$) of the overall number of claims, and a larger deviation from the mean value.

| $c$ | Observed | LightGBM | Fair LightGBM | LightGBM gap | Fair LightGBM gap |
|---|---|---|---|---|---|
| 0 | 18719.0 | 18705.6169 | 18761.1463 | -13.3831 | 42.1463 |
| 1 | 1221.0 | 1224.7363 | 1178.2408 | 3.7363 | -42.7592 |
| 2 | 56.0 | 66.5456 | 58.1117 | 10.5456 | 2.1117 |
| 3 | 4.0 | 2.9848 | 2.4107 | -1.0152 | -1.5893 |
| $> 0$ | 1281.0 | 1294.3831 | 1238.8537 | 13.3831 | -42.1463 |

Table 13: Comparison of observed claim counts with predictions from standard LightGBM and fairness-adjusted LightGBM. The gap columns show the difference between predicted and observed counts. Highlighted values indicate the aggregated total gaps, illustrating the impact of the fairness adjustment on total balance.

| Statistic | $\text{GLM}_{\beta_0}$ | LightGBM | Fair LightGBM |
|---|---|---|---|
| Mean $\overline{\pi}$ | 0.13502 | 0.136334 | 0.129824 |
| 10% Quantile | | 0.039125 | 0.061754 |
| 90% Quantile | | 0.211547 | 0.203506 |

Table 14: Summary statistics of expected premiums $\hat{\pi}$ for the baseline intercept-only GLM ($\text{GLM}_{\beta_0}$), standard LightGBM, and fairness-adjusted LightGBM. Highlighted values indicate divergence from the target mean premium, illustrating the effect of fairness adjustment.

For $\alpha = 0.02$, we manage to significantly improve the underestimation after balance correction. In Table 15 we observe a smaller divergence from the true mean $\mu = 0.13502$, from 0.1298 to 0.1341, and in Table 16 we observe a decrease in the overall overestimation from $-42$ to 6.94 claims.

| Statistic | $\text{GLM}_{\beta_0}$ | LightGBM | Fair LightGBM | Fair BC LightGBM |
|---|---|---|---|---|
| Mean $\overline{\pi}$ | 0.13502 | 0.136334 | 0.129824 | 0.134132 |
| 10% Quantile | | 0.039125 | 0.061754 | 0.107966 |
| 90% Quantile | | 0.211547 | 0.203506 | 0.160423 |

Table 15: Summary statistics of expected premiums $\hat{\pi}$ for the baseline intercept-only GLM ($\text{GLM}_{\beta_0}$), standard LightGBM, fairness-adjusted LightGBM, and bias-corrected fairness-adjusted LightGBM. Highlighted values indicate divergence from the target mean premium, illustrating the effect of fairness adjustment and subsequent bias correction.

| $c$ | Observed | Fair LightGBM | Fair BC LightGBM | Fair LightGBM gap | Fair BC LightGBM gap |
|---|---|---|---|---|---|
| 0 | 18719.0 | 18761.1463 | 18712.0574 | 42.1463 | -6.9426 |
| 1 | 1221.0 | 1178.2408 | 1230.1127 | -42.7592 | 9.1127 |
| 2 | 56.0 | 58.1117 | 55.8166 | 2.1117 | -0.1834 |
| 3 | 4.0 | 2.4107 | 1.9556 | -1.5893 | -2.0444 |
| $> 0$ | 1281.0 | 1238.8537 | 1287.9426 | -42.1463 | 6.9426 |

Table 16: Comparison of observed claim counts with predicted claims from the fairness-adjusted LightGBM before and after bias correction. The gap columns show the difference between predicted and observed counts. Highlighted values indicate the aggregated total gaps, illustrating the reduction of prediction bias after applying balance correction.

Table 17 summarises the deviance loss before and after fairness adjustment (for $G = 2$

and $G = 6$) and balance correction. As expected, the deviance deteriorates much more when adjusting across six demographic groups (gender and age jointly) compared to two (gender only). This reflects the increasing difficulty of aligning multiple conditional premium distributions simultaneously, especially when these attributes are strong predictors. The post-hoc autocalibration step does recover a non-negligible portion of predictive accuracy. For example, for the six-group setting, the deviance decreases from 7341.93 to 6891.32 after balance correction (i.e., +4% relative to the GLM benchmark), showing that the post-hoc calibration mitigates a substantial share of the initial accuracy loss.

| Model | Poisson Deviance | Δ GLM | Δ LightGBM |
|---|---|---|---|
| GLM | 6625.6248 | | |
| LightGBM | 6539.1550 | −1.3051% | |
| LightGBM BC | 6542.6450 | | |
| LightGBM fair (gender) | 6666.7170 | +0.6202% | +1.9507% |
| LightGBM fair BC (gender) | 6654.9800 | +0.4430% | +1.7713% |
| LightGBM fair (age & gender) | 7341.9258 | +10.8111% | +12.2764% |
| LightGBM fair BC (age & gender) | 6891.3200 | +4.0101% | +5.3855% |

Table 17: Poisson deviance comparison across models at the same evaluation stage (out-of-sample 1). Δ percentages compare each fairness (and balance) corrected model relative deterioration with the GLM and LightGBM.

Insurers should be aware that the fairness adjustment may disproportionately affect smaller or extreme subgroups. Table 18 presents the Poisson deviance for each gender–age subgroup, comparing the baseline LightGBM predictions with the fairness-adjusted and balance-corrected models. The relative change (%) in Poisson deviance quantifies the trade-off between enforcing demographic parity and preserving predictive accuracy across demographic group. Fairness adjustment alone deteriorates deviance for all groups, with the largest relative increases observed for young males (23.36%) and senior females (24.71%). This reflects that substantial redistribution of predicted premiums is needed in these groups with higher-frequencies to achieve fairness, which naturally affects predictive accuracy.

Applying the post-hoc balance correction substantially mitigates these increases. For example, young males' deviance increase drops from 23.36% to 2.37%, and young females' from 21.10% to 1.35%. Similarly, middle-aged and senior male groups also see their relative increases fall below 5%, indicating that the BC step effectively restores predictive accuracy (while retaining the fairness adjustment).

Interestingly, the post-hoc balance correction slightly amplifies the increase in deviance for middle-aged and senior female policyholders. This highlights the difficulty of retaining

predictive accuracy for lower claim frequency groups (see darker red KDEs in Fig. 5.8 where a substantial mass of the initial premium distributions is around 0).

Overall, the results highlight that, although fairness enforcement inevitably perturbs the original predictions, especially for groups with higher claim frequencies (i.e., young policyholders), the post-hoc autocalibration mitigates much of the associated loss.

| Gender | Age group | $n$ | LightGBM | Fair LightGBM | Fair BC LightGBM | % Fair | % Fair BC |
|---|---|---|---|---|---|---|---|
| Female | Middle-aged | 3,391 | 708.7603 | 812.7800 | 820.6330 | 14.6763% | 15.7843% |
| Female | Senior | 2,636 | 451.4856 | 563.0406 | 572.7071 | 24.7084% | 26.8495% |
| Female | Young | 4,049 | 1,631.4034 | 1,975.6220 | 1,653.4445 | 21.0995% | 1.3510% |
| Male | Middle-aged | 3,328 | 1,184.7231 | 1,384.3420 | 1,204.5646 | 16.8494% | 1.6748% |
| Male | Senior | 2,644 | 849.8874 | 974.0947 | 886.5005 | 14.6146% | 4.3080% |
| Male | Young | 3,952 | 1,712.8952 | 2,113.0673 | 1,753.4705 | 23.3623% | 2.3688% |

Table 18: Poisson deviance by demographic group for LightGBM, fairness-adjusted LightGBM, and fairness-adjusted LightGBM with balance correction (BC). Percentage increases relative to LightGBM are shown for both adjustments.

Calibration tables in Appendix E further show the impact of fairness corrections by subgroup. For senior policyholders, females (Table E3) experience a substantial degradation in calibration after fairness adjustment, with deviations of $-103.00$ ($c = 0$), whereas males (Table E6), despite being similarly represented, show negligible changes (2.78). This also reflects that substantial redistribution of predicted premiums is needed in the lower claim frequencies groups (see senior and middle-aged KDEs in Fig. 5.8) to achieve fairness, which naturally affects calibration.

# 6. Discussion

In this chapter, we introduce a novel gradient-based group fairness correction using the multivariate Energy distance to enforce demographic parity and reduce group-level disparities across multiple sensitive attributes simultaneously. The multivariate nature of the Energy distance is a key advantage of our approach. It allows for simultaneous fairness adjustments across protected variables with different numbers of sensitive classes, such as gender (binary) and discretised age (multi-class), and is therefore order-independent.

As expected, the loss in accuracy increases when the number of demographic groups grows, reflecting the greater difficulty of simultaneously aligning multiple conditional premium distributions, especially when enforcing fairness for key risk factors, such as gender and age in car pricing. However, the predictive accuracy is partially recovered through a post-hoc autocalibration step (Denuit et al., 2021) introduced to correct biases in the total predicted claim count after fairness adjustments.

We further introduce a simple out-of-sample mapping mechanism to apply fairness corrections to new policyholders. While the random forest used to map policyholder characteristics to individual correction factors does require training, it is important to distinguish this lightweight post-processing step from retraining the underlying pricing model. The LightGBM estimator and gradient-based fairness correction are not refitted once the fairness-adjusted premiums have been produced. As with any predictive model, some degradation is observed when moving to genuinely new policyholders, but our results indicate that this deterioration is comparable to the baseline in/out-of-sample gap observed for both GLM and unconstrained LightGBM models.

While gradient descent introduces an optimisation layer, the process remains transparent and explainable to regulators for several reasons.

First, the LightGBM model that provides the initial discriminatory premium predictions is itself a standard, interpretable model in actuarial practice, with feature importances and partial dependence plots readily available to explain its drivers.

Second, the fairness adjustment is applied as a post-hoc correction, rather than as an in-processing constraint embedded inside the model fitting. We believe this distinction is important for transparency. In-processing methods typically modify the loss function or introduce latent constraints during training, making each change in the predictions less interpretable. In contrast, in our post-processing setup, the correction is defined by a single, well-specified quantitative objective, that is, minimising the Energy distance between conditional premium distributions across sensitive groups. At every iteration, the magnitude and direction of the adjustment can be quantified as a deterministic transformation of the original predictions, allowing practitioners to track (and justify) fairness improvements step by step.

Moreover, this Energy distance itself has a well-established statistical interpretation as a measure of disparity between distributions. The between-group term encourages the alignment of premiums across groups (reducing disparities between $U$ and $V$), while the within-group terms preserve the internal variability of each group. This structure ensures that the correction smooths inter-group differences (improving inter-group fairness) without collapsing intra-group heterogeneity/diversity, thereby maintaining portfolio diversity and the credibility structure of the premiums.

Third, the post-hoc autocalibration step (following Denuit et al. (2021)) adds a local GLM correction to the adjusted premiums. GLMs are a well-documented and interpretable regression framework in actuarial practice, and this step systematically restores balance with actual claims while preserving actuarial soundness.

We believe that the combination of these sequential steps makes the contribution of each fairness or calibration correction transparent.

Besides this transparency concern, another practical challenge in fair insurance pricing arises when protected attributes are not available. Recent works have addressed this issue explicitly. For instance, Gabric et al. (2024) propose a variational Bayesian framework

for discrimination-free pricing in the absence of observed sensitive attributes, while Zhang et al. (2025) consider settings where sensitive attributes are available only in a noisy or privatized form. These approaches highlight an important distinction between fairness under full observability and fairness under partial access to protected information.

In the context of the proposed energy-distance-based framework, the case of noisy or privatized sensitive attributes can in principle be accommodated by replacing exact group membership with its observed corrupted counterpart, leading to fairness constraints defined over induced noisy group partitions. However, the resulting fairness penalty is then aligned with the distribution of the observed (noisy) attribute rather than the true latent protected attribute, introducing an additional layer of measurement error.

By contrast, the fully unobserved case is fundamentally more challenging. In the absence of any direct information on protected attributes, enforcing group-wise distributional alignment requires introducing additional modelling assumptions, such as latent group structure or the use of proxy variables correlated with the protected attribute. This includes approaches where sensitive attributes are inferred from observable features (e.g., name or location based inference as in the Colorado BIFSG framework), but such strategies raise well-known concerns regarding identifiability, stability, and the normative validity of the resulting fairness guarantees, since the fairness constraint becomes dependent on an estimated rather than observed protected structure, which may itself embed bias and introduce additional layers of uncertainty.

**Funding Statement** This research was supported by the project ASTeRISK under research grant ID 40007517, from the Excellence of Science (EoS) program call by FWO and FNRS.

**Competing Interests** No potential conflict of interest was reported by the authors.

**Acknowledgments** The authors are grateful to the anonymous reviewers and the Associate Editor for their constructive comments and suggestions, which substantially improved the manuscript.

# References


Avraham, R. (2017). Discrimination and insurance. *The Routledge handbook of the ethics of discrimination*, 335–347. https://doi.org/10.2139/ssrn.3089946

Bellemare, M. G., Danihelka, I., Dabney, W., Mohamed, S., Lakshminarayanan, B., Hoyer, S., & Munos, R. (2017). The cramer distance as a solution to biased wasserstein gradients. https://arxiv.org/abs/1705.10743

Charpentier, A. (2023). Insurance, biases, discrimination and fairness. https://doi.org/10.1007/978-3-031-49783-4

Charpentier, A., Flachaire, E., & Gallic, E. (2023). Optimal transport for counterfactual estimation: A method for causal inference. In *Optimal transport statistics for economics and related topics* (pp. 45–89). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-35763-3_3

Chzhen, E., Denis, C., Hebiri, M., Oneto, L., & Pontil, M. (2020). Fair regression with wasserstein barycenters. https://arxiv.org/abs/2006.07286

Côté, O., Côté, M.-P., & Charpentier, A. (2025). A fair price to pay: Exploiting causal graphs for fairness in insurance. *Journal of Risk and Insurance*, 92(1), 33–75. https://doi.org/10.1111/jori.12503

Cramér, H. (1928). On the composition of elementary errors. *Scandinavian Actuarial Journal*, 1928(1), 13–74. https://doi.org/10.1080/03461238.1928.10416862

Deka, N., & Sutherland, D. J. (2023). Mmd-b-fair: Learning fair representations with statistical testing. https://arxiv.org/abs/2211.07907

Denuit, M., Charpentier, A., & Trufin, J. (2021). Autocalibration and tweedie-dominance for insurance pricing with machine learning. *Insurance: Mathematics and Economics*, 101, 485–497. https://doi.org/10.1016/j.insmatheco.2021.09.001

Denuit, M., Hainaut, D., & Trufin, J. (2019). Insurance risk classification, 3–26. https://doi.org/10.1007/978-3-030-25820-7_1

Duong, L. R., Zhou, J., Nassar, J., Berman, J., Olieslagers, J., & Williams, A. H. (2022). Representational dissimilarity metric spaces for stochastic neural networks. https://doi.org/10.48550/ARXIV.2211.11665

Edwards, A. (2007). Wouter vandenhole, non-discrimination and equality in the view of the un human rights treaty bodies. *Human Rights Law Review*, 7(1), 267–269. https://doi.org/10.1093/hrlr/ngl032

Fahrenwaldt, M., Furrer, C., Hiabu, M. E., Huang, F., Jørgensen, F. H., Lindholm, M., Loftus, J., Steffensen, M., & Tsanakas, A. (2024). Fairness: Plurality, causality, and insurability. *European Actuarial Journal*, 14(2), 317–328. https://doi.org/10.1007/s13385-024-00387-3

Fatras, K., Zine, Y., Flamary, R., Gribonval, R., & Courty, N. (2019). Learning with minibatch wasserstein : Asymptotic and gradient properties. https://doi.org/10.48550/ARXIV.1910.04091

Frees, E. W., & Huang, F. (2021). The discriminating (pricing) actuary. *North American Actuarial Journal*, 27(1), 2–24. https://doi.org/10.1080/10920277.2021.1951296

Gabric, L., Zhou, S., & Zhou, K. Q. (2024). A bayesian approach to discrimination-free insurance pricing. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4785927

Gandy, O. H. (2016). *Coming to terms with chance: Engaging rational discrimination and cumulative disadvantage*. Routledge. https://doi.org/10.4324/9781315572758

Genevay, A., Peyré, G., & Cuturi, M. (2017). Learning generative models with sinkhorn divergences. https://doi.org/10.48550/ARXIV.1706.00292

Grari, V., Lamprier, S., & Detyniecki, M. (2020). Adversarial learning for counterfactual fairness. https://doi.org/10.48550/ARXIV.2008.13122

Hainaut, D. (2025). In-processing of actuarial and equity fairness constraints for neural networks. https://doi.org/10.2139/ssrn.5652150

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. https://arxiv.org/abs/1610.02413

Hu, F., Ratz, P., & Charpentier, A. (2023a). Fairness in multi-task learning via wasserstein barycenters. https://doi.org/10.48550/ARXIV.2306.10155

Hu, F., Ratz, P., & Charpentier, A. (2023b). A sequentially fair mechanism for multiple sensitive attributes. https://doi.org/10.48550/ARXIV.2309.06627

Kantorovich, L. V., & Rubinshtein, S. (1958). On a space of totally additive functions. *Vestnik of the St. Petersburg University: Mathematics*, 13(7), 52–59.

Knuth, D. (1981). Seminumerical algorithms. *The art of computer programming*, 2. https://doi.org/10.1002/spe.4380120909

Kullback, S., & Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1), 79–86. https://doi.org/10.1214/aoms/1177729694

Kusner, M. J., Loftus, J. R., Russell, C., & Silva, R. (2017). Counterfactual fairness. https://doi.org/10.48550/ARXIV.1703.06856

Lindholm, M., Richman, R., Tsanakas, A., & Wüthrich, M. V. (2022). A discussion of discrimination and fairness in insurance pricing. https://arxiv.org/abs/2209.00858

Lindholm, M., Richman, R., Tsanakas, A., & Wüthrich, M. V. (2024). What is fair? proxy discrimination vs. demographic disparities in insurance pricing. *Scandinavian Actuarial Journal*, 2024(9), 935–970. https://doi.org/10.1080/03461238.2024.2364741

Oneto, L., Donini, M., Luise, G., Ciliberto, C., Maurer, A., & Pontil, M. (2020). Exploiting mmd and sinkhorn divergences for fair and transferable representation learning. *Advances in Neural Information Processing Systems*, 33, 15360–15370.

Panaretos, V. M., & Zemel, Y. (2019). Statistical aspects of wasserstein distances. *Annual Review of Statistics and Its Application*, 6(Volume 6, 2019), 405–431. https://doi.org/10.1146/annurev-statistics-030718-104938

Peyré, G., & Cuturi, M. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6), 355–607. https://doi.org/10.1561/2200000073

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. https://arxiv.org/abs/1709.02012

Roy, S. (2021). Car insurance data [Kaggle Dataset]. https://www.kaggle.com/datasets/sagnik1511/car-insurance-data/data

Royden, H. L., & Fitzpatrick, P. (1988). *Real analysis* (Vol. 32). Macmillan New York.

Santambrogio, F. (2015). *Optimal transport for applied mathematicians: Calculus of variations, pdes and modeling*. Springer. https://doi.org/10.1007/978-3-319-20828-2

Schanze, E. (2013). Injustice by generalization: Notes on the test-achats decision of the european court of justice. *German Law Journal*, 14(2), 423–433. https://doi.org/10.1017/S2071832200001863

Sejdinović, D., Sriperumbudur, B. K., Gretton, A., & Fukumizu, K. (2013). Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *Annals of Statistics*, 41(5), 2263–2291. https://doi.org/10.1214/13-AOS1140

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1), 72. https://doi.org/10.2307/1412159

Székely, G. J. (2003). E-statistics: The energy of statistical samples. *Bowling Green State University, Department of Mathematics and Statistics Technical Report*, 3(05), 1–18. https://doi.org/10.13140/RG.2.1.5063.9761

Tschantz, M. C. (2022). What is proxy discrimination? https://arxiv.org/abs/2205.05265

Villani, C. (2009). *Optimal transport*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-71050-9

Xin, X., & Huang, F. (2023). Antidiscrimination insurance pricing: Regulations, fairness criteria, and models. *North American Actuarial Journal*, 28(2), 285–319. https://doi.org/10.1080/10920277.2023.2190528

Zhang, T., Liu, S., & Shi, P. (2025). Discrimination-free insurance pricing with privatized sensitive attributes. https://doi.org/10.48550/ARXIV.2504.11775

# A. Proofs

*Proof of Proposition 3.1.* The Cramér distance is unique among $\ell_p^p$ distances as its sample gradient is an unbiased estimator of the population gradient. This property, established by Bellemare et al. (2017, p. 16), is crucial in stochastic gradient-based optimisation when probability measures are used as loss functions.

Let $P$ be a probability distribution, and let $\{Q_\theta\}_{\theta\in\Theta}$ denote a differentiable parametric family of probability measures with cumulative distribution functions $F_P(u)$ and $F_{Q_\theta}(u)$. The squared $\ell_2$ distance admits the representation:

$$\ell_2^2(P, Q_\theta) = \int_{-\infty}^{+\infty} \big(F_{Q_\theta}(u) - F_P(u)\big)^2 du.$$

By definition, differentiating with respect to $\theta$ gives:

$$\nabla_\theta\, \ell_2^2\big(P, Q_\theta\big) = \nabla_\theta \int_{-\infty}^{+\infty} \Big(F_{Q_\theta}(u) - F_P(u)\Big)^2 du.$$

Under the hypothesis that the convergence of the square follows from the convergence of the ordinary values (see e.g., the Leibniz rule under dominated convergence), we may swap differentiation and integration:

$$= \int_{-\infty}^{+\infty} \nabla_\theta \Big(F_{Q_\theta}(u) - F_P(u)\Big)^2 du.$$

Application of the chain rule of differentiation then gives:

$$= \int_{-\infty}^{+\infty} 2\Big(F_{Q_\theta}(u) - F_P(u)\Big)\, \nabla_\theta F_{Q_\theta}(u)\, du.$$

By Lemma 1 in Bellemare et al. (2017, p. 14), we replace the true CDF $F_P(u)$ by its expected empirical CDF $\mathbb{E}_{\boldsymbol{U}_m\sim P}[F_{\hat{P}_m}(u)]$:

$$= \int_{-\infty}^{+\infty} 2\Big(F_{Q_\theta}(u) - \mathbb{E}_{\boldsymbol{U}_m\sim P}[F_{\hat{P}_m}(u)]\Big)\, \nabla_\theta F_{Q_\theta}(u)\, du.$$

By linearity of expectation:

$$= \int_{-\infty}^{+\infty} 2\,\mathbb{E}_{\boldsymbol{U}_m\sim P}\Big(F_{Q_\theta}(u) - F_{\hat{P}_m}(u)\Big)\, \nabla_\theta F_{Q_\theta}(u)\, du.$$

By Fubini's theorem (Royden & Fitzpatrick, 1988), we interchange the order of integration and expectation:

$$
\begin{aligned}
&= \mathbb{E}_{\boldsymbol{U}_m \sim P} \int_{-\infty}^{+\infty} 2\Big(F_{Q_\theta}(u) - F_{\hat{P}_m}(u)\Big)\, \nabla_\theta F_{Q_\theta}(u)\, du \\
&= \mathbb{E}_{\boldsymbol{U}_m \sim P}\Big[\nabla_\theta\, \ell_2^2\big(\hat{P}_m, Q_\theta\big)\Big].
\end{aligned}
$$

□

*Proof of Proposition 3.2.* Following Bellemare et al. (2017, p. 16), the Energy distance between two probability measures $P$ and $Q_\theta$ is defined as:

$$
\mathcal{E}(P, Q) = 2\mathbb{E}\|U - V\| - \mathbb{E}\|U - U'\| - \mathbb{E}\|V - V'\|,
$$

where $X \sim P$, $Y \sim Q_\theta$, and $U', V'$ are independent copies of $U$ and $V$, respectively. For the empirical measure $\hat{P}_m$, we write $\hat{U}_m \sim \hat{P}_m$. The gradient of $\mathcal{E}\,(P, Q_\theta)$ w.r.t. $\theta$ is:

$$
\begin{aligned}
\nabla_\theta\, \mathcal{E}(P, Q_\theta) &= \nabla_\theta\Big(2\mathbb{E}\|U - V\| - \mathbb{E}\|U - U'\| - \mathbb{E}\|V - V'\|\Big) \\
&= 2\nabla_\theta\, \mathbb{E}\|U - V\| - \nabla_\theta\, \mathbb{E}\|U - U'\| - \nabla_\theta\, \mathbb{E}\|V - V'\|.
\end{aligned}
$$

Since $\mathbb{E}\|U - U'\|$ does not depend on $\theta$, this simplifies to:

$$
= 2\,\nabla_\theta\, \mathbb{E}\|U - V\| - \nabla_\theta\, \mathbb{E}\|V - V'\|.
$$

Similarly, for the empirical measure $\hat{P}_m$, we have:

$$
\begin{aligned}
\nabla_\theta\, \mathcal{E}\big(\hat{P}_m, Q_\theta\big) &= \nabla_\theta\Big(2\mathbb{E}\|\hat{U}_m - V\| - \mathbb{E}\|\hat{U}_m - \hat{U}'_m\| - \mathbb{E}\|V - V'\|\Big) \\
&= 2\,\nabla_\theta\, \mathbb{E}\|\hat{U}_m - V\| - \nabla_\theta\, \mathbb{E}\|V - V'\|.
\end{aligned}
$$

All together, we have:

$$
\begin{aligned}
\nabla_\theta\, \mathcal{E}(P, Q_\theta) &= \mathbb{E}_{\boldsymbol{U}_m \sim P}\, \nabla_\theta\, \mathcal{E}\big(\hat{P}_m, Q_\theta\big) \\
2\,\nabla_\theta\, \mathbb{E}\|U - V\| - \nabla_\theta\, \mathbb{E}\|V - V'\| &= 2\,\mathbb{E}_{\boldsymbol{U}_m \sim P} \nabla_\theta\, \mathbb{E}\|\hat{U}_m - V\| - \nabla_\theta\, \mathbb{E}\|V - V'\| \\
\nabla_\theta\, \mathbb{E}\|U - V\| &= \mathbb{E}_{\boldsymbol{U}_m \sim P} \nabla_\theta\, \mathbb{E}\|\hat{U}_m - V\|.
\end{aligned}
$$

Under the assumption that $\nabla_\theta$ and the expectation over $\boldsymbol{U}_m$ commute, we have:

$$
\nabla_\theta\, \mathbb{E}\|U - V\| = \nabla_\theta \mathbb{E}_{\boldsymbol{U}_m \sim P}\, \mathbb{E}\|\hat{U}_m - V\|.
$$

Under the independence of the variables with $\hat{P}_m$ and $Q_\theta$, we have:

$$
= \nabla_\theta \mathbb{E}_{\boldsymbol{U}_m \sim P}\, \mathbb{E}\, \mathbb{E}_{u \sim \hat{P}_m}\|u - V\|.
$$

Finally, given that the expected empirical distribution is $P$, we find:

$$
\begin{aligned}
&= \nabla_\theta \mathbb{E}_{u\sim P}\, \mathbb{E}\|u - V\| \\
&= \nabla_\theta\, \mathbb{E}\|U - V\|.
\end{aligned}
$$

□

*Proof of equivalence with the Energy distance.* In the univariate case, Székely (2003) established that the Energy distance and the $\ell_2^2$ distance are linearly related:

$$\ell_2^2(P, Q) = \frac{1}{2}\,\mathcal{E}(P, Q).$$

Let $P$ and $Q$ be probability distributions on $\mathbb{R}$ with cumulative distribution functions $F_P$ and $F_Q$, respectively. In one dimension, the Euclidean norm $\|x - y\|$ reduces to the absolute value $|x-y|$ for scalars $x, y \in \mathbb{R}$. Therefore, the Energy distance can be expressed as:

$$\mathcal{E}(P, Q) = 2\,\mathbb{E}\Big[|U - V|\Big] - \mathbb{E}\Big[|U - U'|\Big] - \mathbb{E}\Big[|V - V'|\Big]\,,$$

where $U, U' \sim P$ i.i.d. and $V, V' \sim Q$ i.i.d., with all pairs independent. By definition of the expectation of functions of independent random variables:

$$
\begin{aligned}
\mathbb{E}[|U - V|] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |u - v|\, dP(u) dQ(v). \\
\mathbb{E}[|U - U'|] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |u - u'|\, dP(u)\, dP(u'). \\
\mathbb{E}[|V - V'|] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |v - v'|\, dQ(v)\, dQ(v').
\end{aligned}
$$

It is well-known that (Székely, 2003), in one dimension, the expected absolute difference can be expressed in terms of the squared difference of the CDFs:

$$
\begin{aligned}
\mathbb{E}[|U - V|] &= \int_{-\infty}^{\infty} \Big(F_P(u) + F_Q(u) - 2F_P(u)F_Q(u)\Big)\, du = \int_{-\infty}^{\infty} \Big(F_P(u) - F_Q(u)\Big)^2\, du \\
\mathbb{E}[|U - U'|] &= \int_{-\infty}^{\infty} \Big(F_P(u) + F_P(u) - 2F_P(u)^2\Big)\, du = \int_{-\infty}^{\infty} \Big(2F_P(u) - 2F_P(u)^2\Big)\, du \\
\mathbb{E}[|V - V'|] &= \int_{-\infty}^{\infty} \Big(2F_Q(u) - 2F_Q(u)^2\Big)\, du\,.
\end{aligned}
$$

Substituting these results into the definition of Energy distance gives:

$$\begin{aligned}
\mathcal{E}(P,Q) =& 2\mathbb{E}[|U-V|] - \mathbb{E}[|U-U'|] - \mathbb{E}[|V-V'|] \\
=& \int_{-\infty}^{\infty} \Big[ 2F_P(u) + 2F_Q(u) - 4F_P(u)F_Q(u) \\
& - \big(2F_P(u) - 2F_P(u)^2\big) - \big(2F_Q(u) - 2F_Q(u)^2\big) \Big] \, du \\
=& \int_{-\infty}^{\infty} \Big[ -4F_P(u)F_Q(u) + 2F_P(u)^2 + 2F_Q(u)^2 \Big] \, du \\
=& 2\int_{-\infty}^{\infty} \Big[ -2F_P(u)F_Q(u) + F_P(u)^2 + F_Q(u)^2 \Big] \, du \\
=& 2\int_{-\infty}^{\infty} \big(F_P(u) - F_Q(u)\big)^2 \, du \\
=& 2\ell_2^2(P,Q) \, .
\end{aligned}$$

□

# B. Datasets

| Number of claims | Count | Percentage |
|---|---|---|
| 0 | 93,535 | 93.53% |
| 1 | 6,109 | 6.11% |
| 2 | 335 | 0.34% |
| 3 | 21 | 0.02% |

Table B1: The table reports the number of policies and the corresponding percentage for each observed claim count $N_i \in \{0,1,2,3\}$ in the synthetic motor insurance portfolio. As shown, the majority of policies have zero claims, highlighting the rarity of the event.

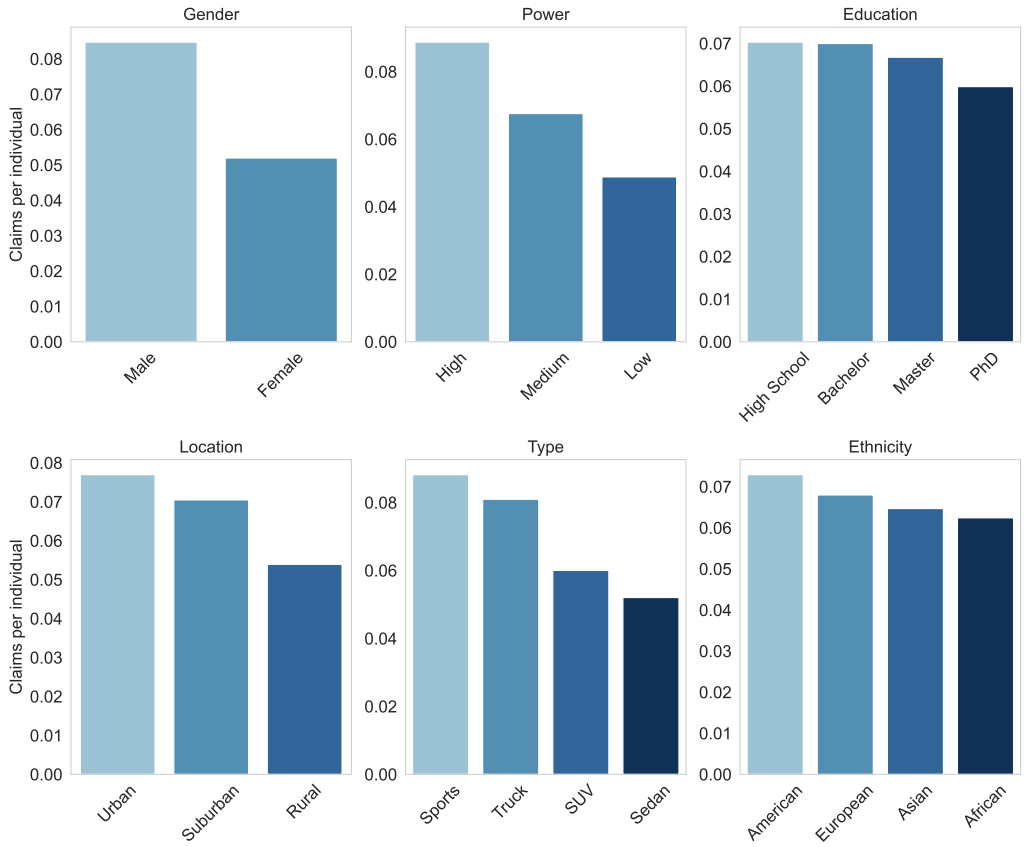


Figure B1: Average number of claims per policyholder for each category of the discrete rating factors in the simulated motor insurance portfolio. Categories within each factor are sorted in descending order of claim frequency to highlight the most at-risk segments.

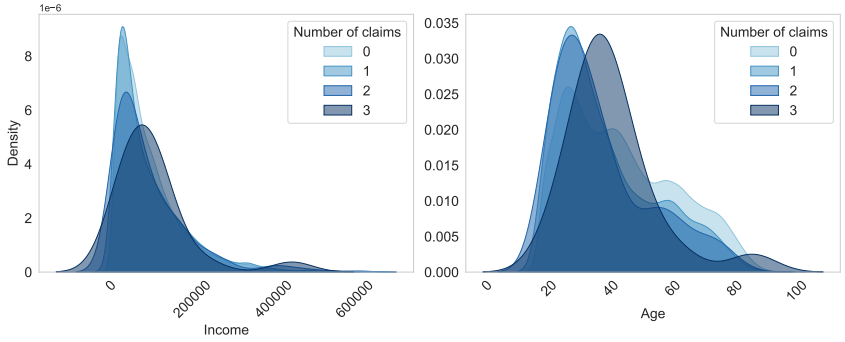


Figure B2: Density distributions of the continuous rating factors (age and income) stratified by the number of claims $N_i \in \{0, 1, 2, 3\}$ in the simulated motor insurance portfolio. These plots show how the distribution of continuous covariates varies across different claim counts.

# C. Best-estimates

| Parameter | Value | Description |
|---|---|---|
| learning_rate | 0.0099 | Step size shrinkage used in each boosting step to prevent overfitting. Smaller values make learning more conservative. |
| feature_fraction | 0.9368 | Fraction of features randomly selected for each tree (column subsampling), used to prevent overfitting. |
| bagging_fraction | 0.6151 | Fraction of data randomly selected for each bagging iteration (row subsampling), helps reduce variance. |
| bagging_freq | 4 | Frequency of bagging; LightGBM performs data subsampling every 4 boosting iterations. |
| lambda_l2 | 2.5348e-05 | L2 regularization term on weights to prevent overfitting (ridge penalty). |
| min_sum_hessian_in_leaf | 0.7570 | Minimum sum of instance Hessians (second-order gradients) in a leaf; larger values can make the model more conservative. |
| max_depth | 6 | Maximum depth of each tree; limits model complexity and helps control overfitting. |
| min_gain_to_split | 0.4139 | Minimum gain required to make a further partition on a leaf; acts as a regularization term to control splitting. |
| min_data_in_leaf | 74 | Minimum number of data points allowed in a leaf; prevents overly specific leaf nodes. |
| num_leaves | 51 | Maximum number of leaves per tree; controls model complexity and expressiveness. |
| objective | poisson | Objective function optimized by the model; here, a Poisson regression suited for count data. |

Table C1: Optimized LightGBM hyperparameters using `optuna` and their descriptions.

# D. Best-estimates autocalibration

The negative bias (or overestimation) observed in the LightGBM predictions in Table 5 can be corrected using the multiplicative bias-correction factor proposed by Denuit et al.

(2021):

$$\frac{\mathbb{E}\left[Y|\hat{\pi}\left(\boldsymbol{X}\right)=F_{\hat{\pi}}^{-1}\left(x\right)\right]}{F_{\hat{\pi}}^{-1}\left(x\right)}, \tag{D1}$$

estimated using a local smoother with bandwidth $\alpha = 0.25$. This correction reduces global bias while preserving the Poisson deviance (barely increases from 6539.15 to 6542.64).

Table D1 reports summary statistics before and after bias correction. The average premium shifts from 0.1363 to 0.1346, bringing it close to the reference mean.

| Statistic | $\text{GLM}_{\beta_0}$ | LightGBM | BC LightGBM |
|---|---|---|---|
| Mean $\overline{\pi}$ | 0.1350 | 0.1363 | 0.1346 |
| 10% Quantile | – | 0.0391 | 0.0374 |
| 90% Quantile | – | 0.2115 | 0.2003 |

Table D1: Summary statistics of the expected premiums $\hat{\pi}$ for the baseline intercept-only GLM ($\text{GLM}_{\beta_0}$), the LightGBM, and the bias-corrected LightGBM. Highlighted value indicates the shift in deviation from the baseline mean premium, illustrating the effect of bias correction.

Table D2 shows the improvement in predicted claim counts. The gap in the number of policies with at least one claim drops from 13.38 to $-1.37$, effectively removing the systematic bias.

| $c$ | Observed | LightGBM | BC LightGBM | LightGBM gap | BC LightGBM gap |
|---|---|---|---|---|---|
| 0 | 18719 | 18705.6169 | 18720.3746 | -13.3831 | 1.3746 |
| 1 | 1221 | 1224.7363 | 1211.9909 | 3.7363 | -9.0091 |
| 2 | 56 | 66.5456 | 64.7117 | 10.5456 | 8.7117 |
| 3 | 4 | 2.9848 | 2.8178 | -1.0152 | -1.1822 |
| $> 0$ | 1281 | 1294.3831 | 1279.6254 | 13.3831 | -1.3746 |

Table D2: Comparison of observed claim counts with predicted claims from LightGBM before and after bias correction. The columns gap show the difference between predicted and observed counts. Highlighted value indicates the reduction of prediction bias after correction.

Fig. D1 displays histograms of premium predictions before and after bias correction.

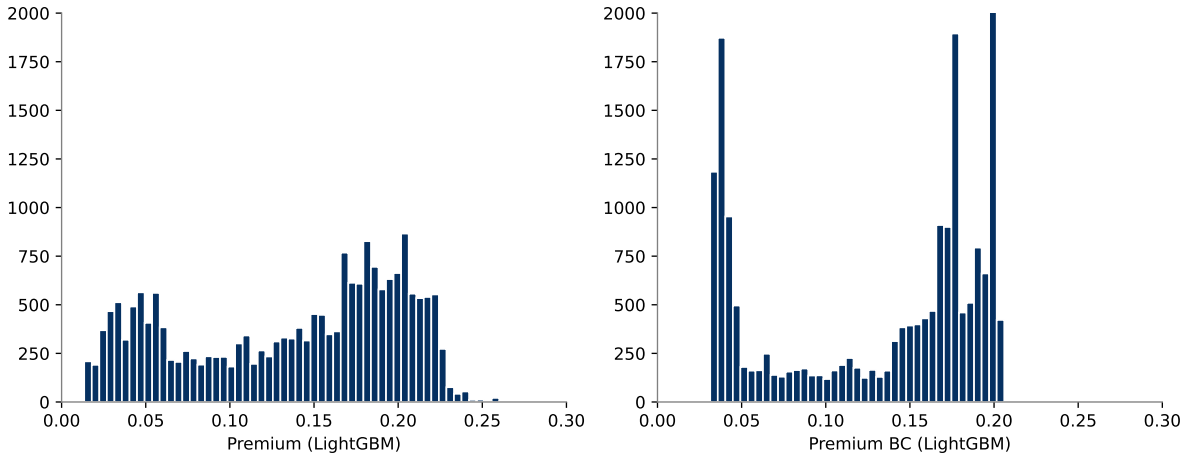


Figure D1: Histogram comparison of predicted premiums $\hat{\pi}$ for the LightGBM before (left) and after (right) balance correction.

Fig. D2 illustrates how the mapping $s \mapsto \mathbb{E}\left[Y|\hat{\pi}\left(\boldsymbol{X}\right)=s\right]$ moves closer to the first bisector (i.e., true mean) $\mu\left(\boldsymbol{x}\right)=\mathbb{E}\left[Y|\boldsymbol{X}=x\right]$ after correction.

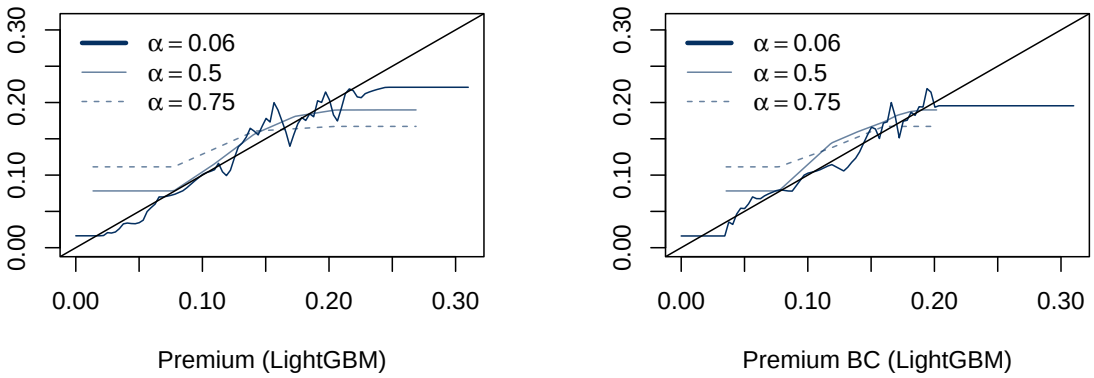


Figure D2: Visualisation of $s \mapsto \mathbb{E}\left[Y|\hat{\pi}\left(\boldsymbol{X}\right)=s\right]$ before (left plot) and after (right plot) bias correction (BC). $\alpha$ is a smoothing hyper-parameter in the `locfit` package in R, and the diagonal corresponds to $\hat{\pi}=\mu$.

Overall, the autocalibration step successfully realigns the LightGBM predictions with the true mean while maintaining predictive performance (measured by the Poisson deviance). This balance correction is a key post-hoc step of the fairness adjustment.

# E. Simulated dataset

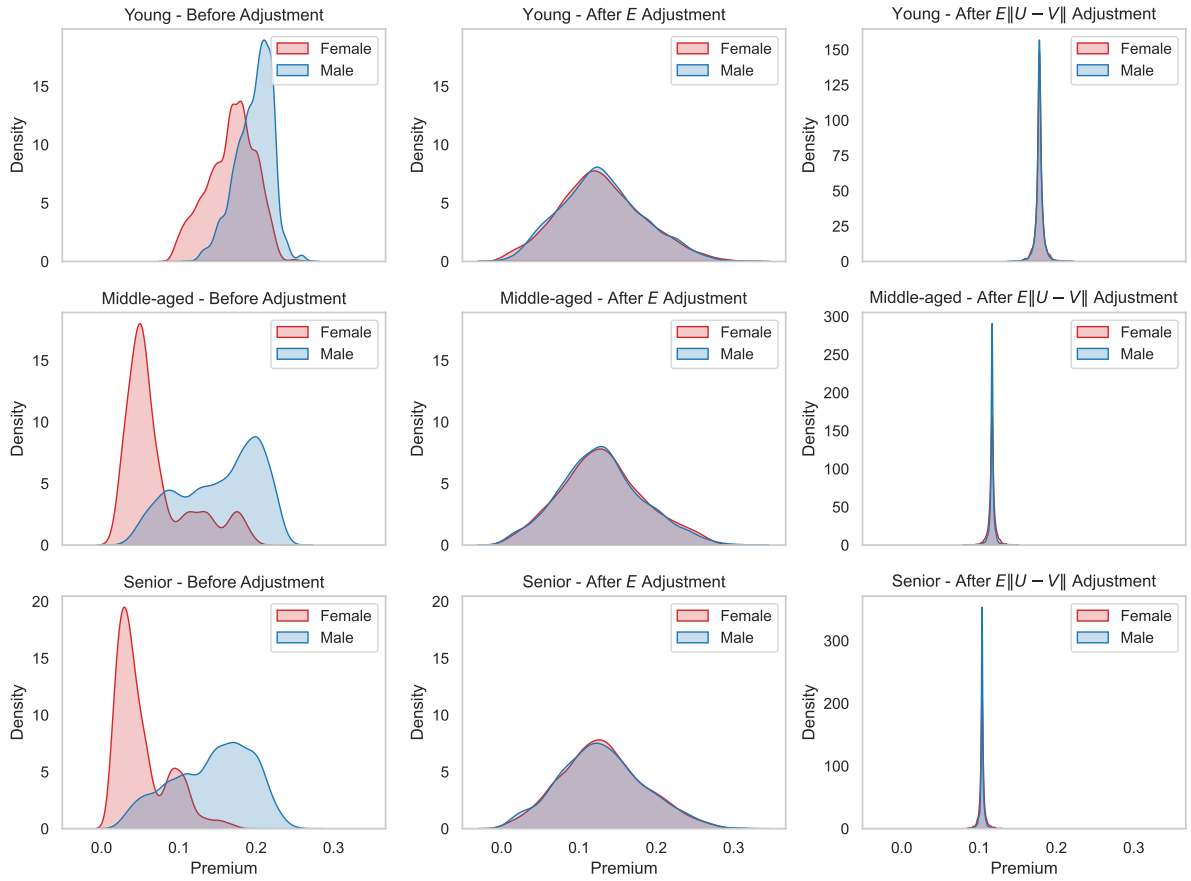


Figure E1: Visualisation of the role of the intra-group variability terms $\mathbb{E}\big[\|U-U'\|\big]$ and $\mathbb{E}\big[\|V-V'\|\big]$ in the fairness adjuster.

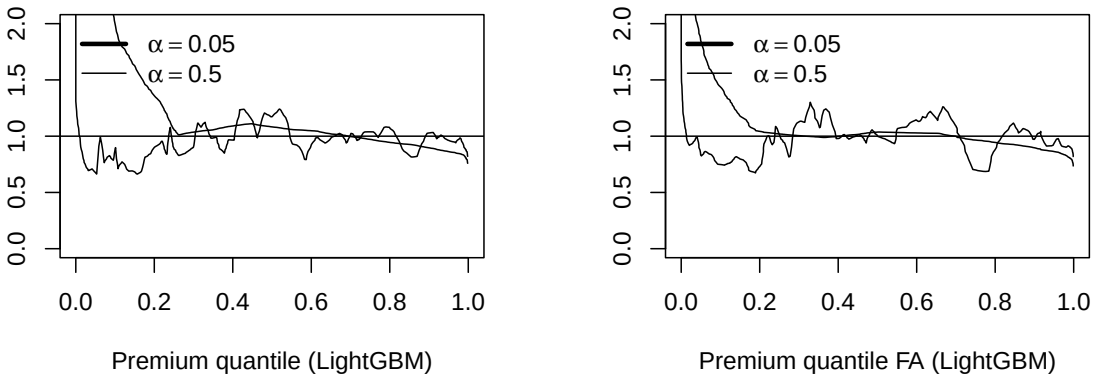


Figure E2: Visualisation of the local bias $\mathbb{E}\left[Y|\hat{\pi}\left(\boldsymbol{X}\right)=F_{\hat{\pi}}^{-1}\left(x\right)\right]/F_{\hat{\pi}}^{-1}\left(x\right)$ before and after fairness adjustment (FA).

| $c$ | Observed | LightGBM | Fair LightGBM | Fair BC LightGBM | $\Delta$ | $\Delta$ Fair | $\Delta$ Fair BC |
|---|---|---|---|---|---|---|---|
| 0 | 3711 | 3723.96 | 3795.55 | 3786.66 | 12.96 | 84.55 | 75.66 |
| 1 | 324 | 306.64 | 241.02 | 250.56 | −17.36 | −82.98 | −73.44 |
| 2 | 14 | 17.58 | 11.92 | 11.37 | 3.58 | −2.08 | −2.63 |
| 3 | 0 | 0.78 | 0.49 | 0.40 | 0.78 | 0.49 | 0.40 |

Table E1: Young-females calibration table showing the observed number of claims and the predicted counts from the LightGBM, the fairness-adjusted LightGBM, and the bias-corrected and fairness-adjusted LightGBM, along with the differences between predicted and observed counts.

| $c$ | Observed | LightGBM | Fair LightGBM | Fair BC LightGBM | $\Delta$ | $\Delta$ Fair | $\Delta$ Fair BC |
|---|---|---|---|---|---|---|---|
| 0 | 3281 | 3271.76 | 3180.63 | 3172.47 | -9.24 | -100.37 | -108.53 |
| 1 | 105 | 115.35 | 200.10 | 208.72 | 10.35 | 95.10 | 103.72 |
| 2 | 3 | 3.77 | 9.85 | 9.47 | 0.77 | 6.85 | 6.47 |
| 3 | 2 | 0.12 | 0.41 | 0.33 | -1.88 | -1.59 | -1.67 |

Table E2: Middle-aged-females calibration table showing the observed number of claims and the predicted counts from the LightGBM, the fairness-adjusted LightGBM, and the bias-corrected and fairness-adjusted LightGBM, along with the differences between predicted and observed counts.

| $c$ | Observed | LightGBM | Fair LightGBM | Fair BC LightGBM | $\Delta$ | $\Delta$ Fair | $\Delta$ Fair BC |
|---|---|---|---|---|---|---|---|
| 0 | 2576 | 2569.34 | 2473.00 | 2465.42 | -6.66 | -103.00 | -110.58 |
| 1 | 56 | 65.09 | 155.14 | 162.94 | 9.09 | 99.14 | 106.94 |
| 2 | 4 | 1.54 | 7.54 | 7.38 | -2.46 | 3.54 | 3.38 |
| 3 | 0 | 0.04 | 0.31 | 0.26 | 0.04 | 0.31 | 0.26 |

Table E3: Senior-females calibration table showing the observed number of claims and the predicted counts from the LightGBM, the fairness-adjusted LightGBM, and the bias-corrected and fairness-adjusted LightGBM, along with the differences between predicted and observed counts.

| $c$ | Observed | LightGBM | Fair LightGBM | Fair BC LightGBM | $\Delta$ | $\Delta$ Fair | $\Delta$ Fair BC |
|---|---|---|---|---|---|---|---|
| 0 | 3579 | 3584.58 | 3707.25 | 3697.33 | 5.58 | 128.25 | 118.33 |
| 1 | 355 | 343.67 | 232.74 | 243.22 | −11.33 | −122.26 | −111.78 |
| 2 | 17 | 22.57 | 11.51 | 11.05 | 5.57 | −5.49 | −5.95 |
| 3 | 1 | 1.13 | 0.48 | 0.39 | 0.13 | −0.52 | −0.61 |

Table E4: Young-males calibration table showing the observed number of claims and the predicted counts from the LightGBM, the fairness-adjusted LightGBM, and the bias-corrected and fairness-adjusted LightGBM, along with the differences between predicted and observed counts.

| $c$ | Observed | LightGBM | Fair LightGBM | Fair BC LightGBM | $\Delta$ | $\Delta$ Fair | $\Delta$ Fair BC |
|---|---|---|---|---|---|---|---|
| 0 | 3093 | 3090.95 | 3122.93 | 3116.95 | −2.05 | 29.93 | 23.95 |
| 1 | 224 | 224.27 | 194.96 | 201.64 | 0.27 | −29.04 | −22.36 |
| 2 | 10 | 12.22 | 9.69 | 9.08 | 2.22 | −0.31 | −0.92 |
| 3 | 1 | 0.54 | 0.41 | 0.32 | −0.46 | −0.59 | −0.68 |

Table E5: Middle-aged-males calibration table showing the observed number of claims and the predicted counts from the LightGBM, the fairness-adjusted LightGBM, and the bias-corrected and fairness-adjusted LightGBM, along with the differences between predicted and observed counts.

| $c$ | Observed | LightGBM | Fair LightGBM | Fair BC LightGBM | $\Delta$ | $\Delta$ Fair | $\Delta$ Fair BC |
|---|---|---|---|---|---|---|---|
| 0 | 2479 | 2465.02 | 2481.78 | 2473.23 | −13.98 | 2.78 | −5.77 |
| 1 | 157 | 169.71 | 154.28 | 163.03 | 12.71 | −2.72 | 6.03 |
| 2 | 8 | 8.88 | 7.61 | 7.47 | 0.88 | −0.39 | −0.53 |
| 3 | 0 | 0.38 | 0.32 | 0.26 | 0.38 | 0.32 | 0.26 |

Table E6: Senior-males calibration table showing the observed number of claims and the predicted counts from the LightGBM, the fairness-adjusted LightGBM, and the bias-corrected and fairness-adjusted LightGBM, along with the differences between predicted and observed counts.

# F. Real-world dataset

The analysis presented in Section 5 on simulated data is reproduced for the real-world dataset introduced in Section 5.1.

## F.1 Best-estimates

We consider a boosting algorithm (LightGBM) to model the premium best-estimates $\hat{\pi}$. The boosting model is tuned via `lightgbm` and `optuna` with optimal hyper-parameters reported in Table F1.

| Parameter | Value | Parameter | Value |
|---|---|---|---|
| learning_rate | 0.0164 | min_sum_hessian_in_leaf | 0.0002 |
| feature_fraction | 0.8795 | max_depth | 7 |
| bagging_fraction | 0.8946 | min_gain_to_split | 0.3384 |
| bagging_freq | 1 | min_data_in_leaf | 35 |
| lambda_l2 | 1.1452e-09 | num_leaves | 30 |

Table F1: Optimized LightGBM (with poisson objective) hyperparameters using `optuna`. Their descriptions are detailed in Table C1.

Fig. 5.1 compares predicted and observed claim frequencies across 10 quantile bins. Bars represent total exposure within each bin, while the solid and dashed lines show the predicted $\hat{\lambda}$ and observed $\lambda$ frequencies, respectively. A closer alignment between the two lines across bins indicates superior model performance.

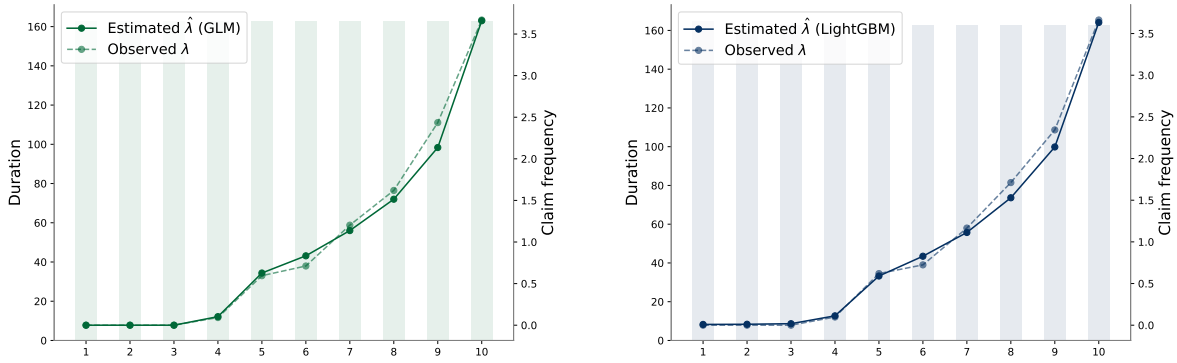


Figure F1: Comparison of predicted and observed claim frequencies per unit exposure across 10 quantile-based bins. The solid line shows the estimated frequency $\hat{\lambda}$ from the model, while the dashed line shows the observed frequency $\lambda$. The bars indicate the total exposure in each bin. Left plot: GLM; right plot: LightGBM.

As observed in Table F2, LightGBM has a slightly higher deviance than the GLM. In the analysis on simulated data, we observed the opposite.

| Model | Poisson deviance |
|---|---|
| GLM | 1436.7176 |
| LightGBM | 1476.3945 |

Table F2: Comparison of Poisson deviance values, defined in Eq. (12), on the first out-of-sample test set for the GLM and the LightGBM. Lower deviance indicates a better fit to the observed claim frequencies.

Fig. F2 illustrates the distribution of predicted premiums and highlights differences in risk allocation between the two models. Since the predicted premiums correspond to Poisson count estimates, many values (in the first histogram bin) are close to zero (but strictly positive). For readability, all visualisations are based on the transformed scale $\log(1 + \hat{\lambda}_i)$.

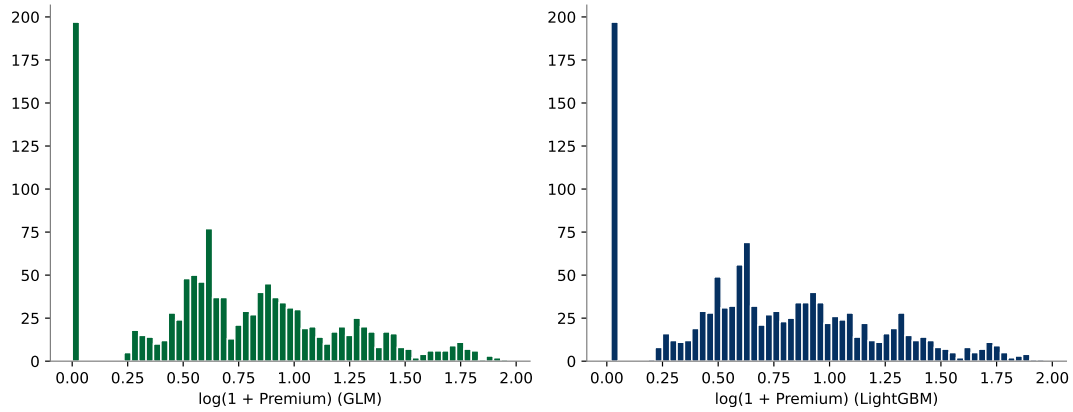


Figure F2: Histogram comparison of predicted premiums $\hat{\pi}$ for the GLM (left) and LightGBM (right). The plots illustrate the distribution of predicted premiums and highlight differences in risk allocation between the two models.

## F.2 Fairness adjustment G=2

In the binary setting $G = 2$, the predicted (discriminatory) premiums $\hat{\pi}$ are partitioned into two demographic groups $\{\hat{\pi}_g\}_{g=1}^{G}$ corresponding to $g = \{\text{female}, \text{male}\}$. These premiums are iteratively adjusted using algorithm 1 to equalise not only the means, but also the dispersion and overall shape of their distributions.

Fig. F3 presents the cumulative distributions of predicted premiums for the overall portfolio and for the male and female subgroups, before (solid lines) and after (dashed lines) the fairness adjustment. All predictions are visualised on the transformed scale $\log(1+\hat{\lambda}_i)$. While the analysis on the simulated portfolio in Section 5.3 relied on standard (non-cumulative) kernel density estimates, the cumulative KDEs in Fig. F3 allow for

a better comparison of the distributional shift across groups, particularly in the lower premium range where many values are close to zero.

Before the adjustment, the cumulative distributions differ noticeably between genders. The cumulative curve for females lies above that for males (i.e., female policyholders have lower predicted premiums on average than males). After the fairness adjustment, the cumulative distributions for both groups (dashed lines) nearly coincide, indicating that our Energy-based fairness correction effectively mitigates gender-based pricing disparities.

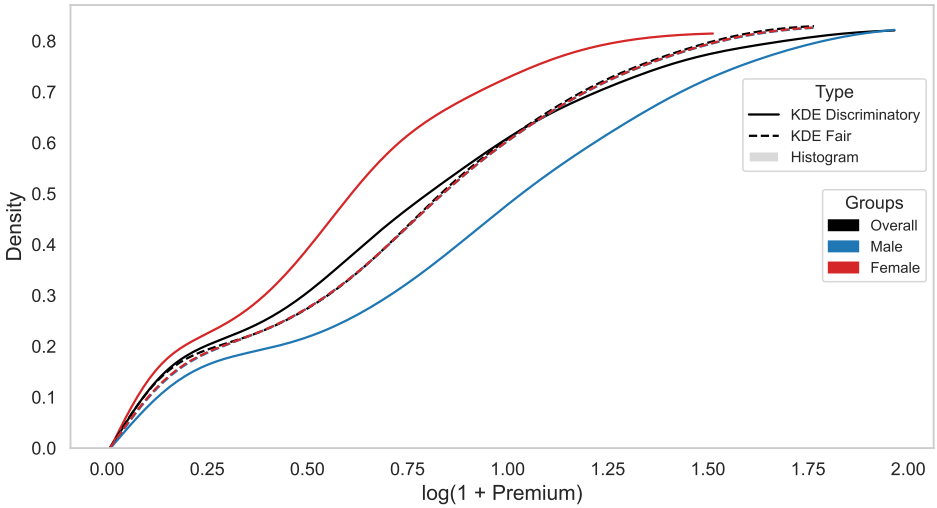


Figure F3: Cumulative kernel density estimates (KDEs) of predicted premiums (on the transformed scale $\log(1+\hat{\lambda}_i)$) by gender. Solid lines represent the original LightGBM discriminatory premiums, while dashed lines correspond to the fairness-adjusted premiums. The male group is shown in blue, the female group in red, and the overall portfolio in black.

Enforcing independence between premiums and gender (a key risk factor) constrains the predictions and prevents the model from fully exploiting gender-related risk information. As illusrated in Fig. F4, the model inevitably sacrifices some predictive accuracy to enforce demographic parity.

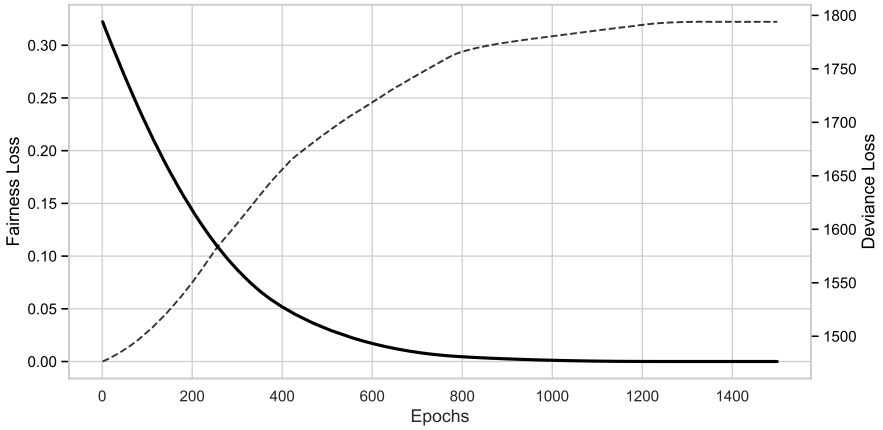


Figure F4: Evolution of the fairness loss (i.e., the Energy distance in Eq. (7)) and the Poisson deviance loss (in Eq. (12)) over training epochs.

Table F3 shows that the fairness adjustment also leads to an increase in the overestimation of the overall claim count (from 44 to 81). However, we also observe in Table F4 that the fairness adjustment leads to a smaller deviation from the true mean $\mu = 1.0119$.

| $c$ | Observed | LightGBM | $\text{LightGBM}_{\text{fair}}$ | LightGBM gap | Fair LightGBM gap |
|---|---|---|---|---|---|
| 0 | 951 | 906.9875 | 869.6073 | -44.0125 | -81.3927 |
| 1 | 272 | 304.0601 | 318.1396 | 32.0601 | 46.1396 |
| 2 | 154 | 188.3684 | 208.1326 | 34.3684 | 54.1326 |
| 3 | 101 | 106.6544 | 117.8240 | 5.6544 | 16.8240 |
| 4 | 62 | 58.4609 | 61.4248 | -3.5391 | -0.5752 |
| 5 | 36 | 31.6624 | 30.2977 | -4.3376 | -5.7023 |
| 6 | 28 | 16.8910 | 14.1690 | -11.1090 | -13.8310 |
| 7 | 12 | 8.7817 | 6.2518 | -3.2183 | -5.7482 |
| 8 | 8 | 4.3992 | 2.5920 | -3.6008 | -5.4080 |
| 9 | 4 | 2.1040 | 1.0082 | -1.8960 | -2.9918 |
| $> 0$ | 679.0 | 723.0125 | 760.3927 | 44.0125 | 81.3927 |

Table F3: Comparison of observed claim counts with predictions from standard LightGBM and fairness-adjusted LightGBM. The gap columns show the difference between predicted and observed counts. Highlighted values indicate the aggregated total gaps, illustrating the impact of the fairness adjustment on total balance.

| Statistic | $\text{GLM}_{\beta_0}$ | LightGBM | Fair LightGBM |
|---|---|---|---|
| Mean $\overline{\pi}$ | 1.0609 | 0.9983 | 1.0119 |
| 10% Quantile | – | 0.0099 | 0.0099 |
| 90% Quantile | – | 2.6672 | 2.4771 |

Table F4: Summary statistics of expected premiums $\hat{\pi}$ for the baseline intercept-only GLM ($\text{GLM}_{\beta_0}$), the standard LightGBM, and the fairness-adjusted LightGBM. Highlighted values indicate divergence from the target mean premium, indicating the effect of fairness adjustment.

The negative bias (or overestimation) observed in the fair LightGBM predictions can be corrected using the multiplicative bias-correction factor proposed by Denuit et al. (2021):

$$\frac{\mathbb{E}\left[Y|\hat{\pi}\left(\boldsymbol{X}\right) = F_{\hat{\pi}}^{-1}\left(x\right)\right]}{F_{\hat{\pi}}^{-1}\left(x\right)}, \tag{F1}$$

estimated using a local smoother with bandwidth $\alpha = 0.16$. Table F5 reports summary statistics before and after bias correction. The average premium shifts from 1.0119 to 1.0397, bringing it close to the reference mean.

| Statistic | $\text{GLM}_{\beta_0}$ | LightGBM | Fair LightGBM | Fair BC LightGBM |
|---|---|---|---|---|
| Mean $\overline{\pi}$ | 1.0609 | 0.9983 | 1.0119 | 1.0397 |
| 10% Quantile | – | 0.0099 | 0.0099 | 0.0000 |
| 90% Quantile | – | 2.6672 | 2.4771 | 2.5625 |

Table F5: Summary statistics of expected premiums $\hat{\pi}$ for the baseline intercept-only GLM ($\text{GLM}_{\beta_0}$), standard LightGBM, fairness-adjusted LightGBM, and bias-corrected fairness-adjusted LightGBM. Highlighted values indicate divergence from the target mean premium, illustrating the effect of fairness adjustment and subsequent bias correction.

Table F6 summarises the Poisson deviance values. We note that the fairness–accuracy trade-off manifests at two distinct levels. The fairness correction inevitably reduces predictive accuracy, as the optimisation process prioritises demographic parity over likelihood-based fit. The post-hoc autocalibration step partially restores, not only the actuarial balance, but also a share of the predictive accuracy (1667.6080) that was initially reduced by the fairness correction, without affecting fairness, as shown in Fig. F5.

| Model | Poisson Deviance | Δ GLM | Δ LightGBM |
|---|---|---|---|
| GLM | 1436.7176 | | |
| LightGBM | 1476.3944 | +0.0276 | |
| LightGBM fair (gender) | 1793.9142 | +0.2486 | +0.2151 |
| LightGBM fair BC (gender) | 1667.6080 | +0.1607 | +0.1295 |

Table F6: Poisson deviance comparison at the same evaluation stage (out-of-sample 1) for the generalised linear model (GLM), bias-corrected (BC), and fairness-adjusted versions of the LightGBM model. Lower deviance indicates a better fit to the observed claim frequencies. $\Delta = (a - b)/b$ values compare each fairness (and balance) corrected model relative deterioration with the GLM and LightGBM.

Fig. F5 shows the local redistribution of density after balance correction (dotted lines). Even after balance correction, fairness is maintained as the conditional premium densities for both gender groups remain well-aligned.

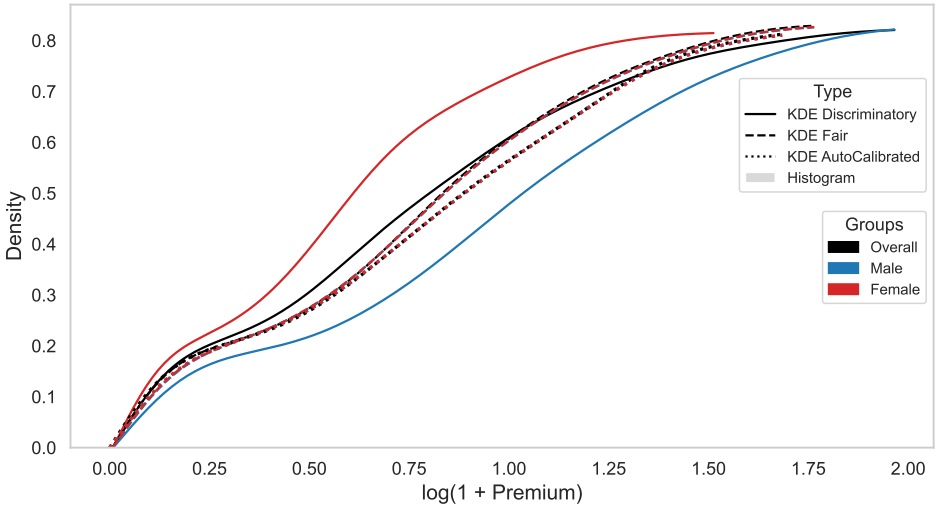


Figure F5: Cumulative kernel density estimates (KDE) of predicted premiums by gender. Solid lines represent the original LightGBM discriminatory premiums, dashed lines the fairness-adjusted premiums, and dotted lines the fairness-adjusted and balance-corrected premiums. The male group is shown in blue, the female group in red, and the overall portfolio in black.

Table F7 summarises several keys fairness metrics discussed in the literature (see dicussion related to Table 11).

| Metric with $U = \hat{\pi}_{\text{female}}$ and $V = \hat{\pi}_{\text{male}}$ | $\hat{\pi}$ | $\hat{\pi}^{\text{fair, BC}}$ |
|---|---|---|
| $W_1(U, V)$ | 0.883082 | 0.175620 |
| $W_2(U, V)$ | 1.264377 | 0.306100 |
| $\text{MMD}^2(U, V)$ | 0.037794 | $8.2 \times 10^{-5}$ |
| $\Delta_{\text{GFE}} = \lvert\mathbb{E}[U] - \mathbb{E}[V]\rvert$ | 0.760072 | 0.001787 |
| $\Delta_{\text{SP}} = \sup_s \lvert P(U \leq s) - P(V \leq s)\rvert$ | 0.335124 | 0.017543 |
| $S_{\epsilon=0.1}(U, V)$ | 0.701744 | 0.000864 |

Table F7: Comparison of key fairness metrics before and after $\mathcal{E}$-based fairness adjustment and balance correction. MMD is computed using a Gaussian RBF kernel $k(u, v) = \exp\left(-\|u - v\|^2 / (2\sigma^2)\right)$ with fixed bandwidth $\sigma = 0.1$.

For new policyholders, a random forest regressor learns the mapping from policyholder characteristics to correction factors, as outlined in Section 4.2. Fig. F6 shows that the adjusted predictions for new policyholders (out-of-sample 2) remain consistent across sensitive groups with the results observed in Fig. F5.

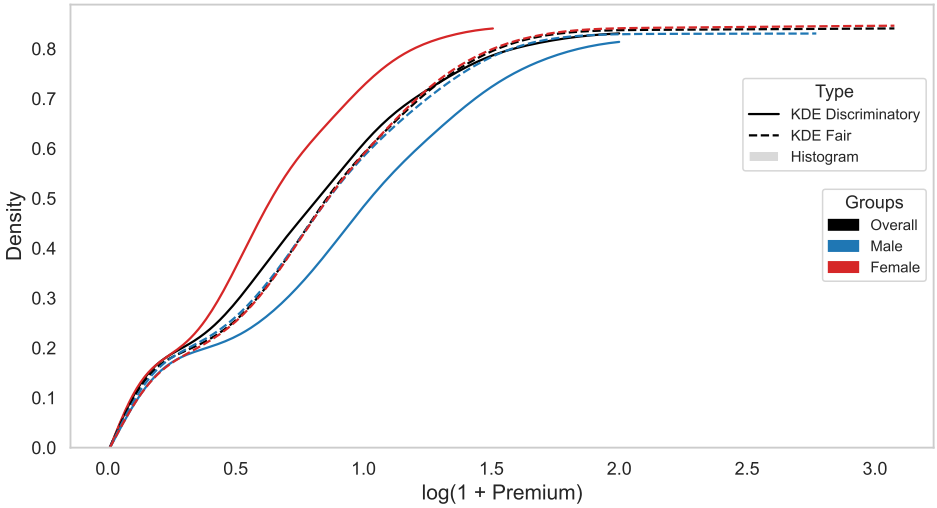


Figure F6: Cumulative kernel density estimates (KDEs) of predicted premiums by gender for new policyholders. Solid lines represent the LightGBM discriminatory premiums, dashed lines the fairness-adjusted premiums, and dotted lines the fairness-adjusted and balance-corrected premiums. The male group is shown in blue, the female group in red, and the overall portfolio in black.

Table F8 summarises the Poisson deviance results before and after fairness correction on the new test (out-of-sample 2) set with 1 630 policies. The transition from out-of-sample 1 to out-of-sample 2 reveals a nearly negligible change in loss (from 1793.91 to 1794.33). Both out-of-sample sets contain only $n = 1{,}630$ observations, which naturally induces some instability. With such small test samples, even moderate changes in the composition of policyholders between the two out-of-sample sets can produce nontrivial fluctuations in deviance after fairness correction.

| | in-sample | out-of-sample 1 | out-of-sample 2 | $\Delta$ |
|---|---|---|---|---|
| GLM | 4019.5312 | 1436.7176 | 1380.0193 | -3.9464% |
| LightGBM | 3814.3971 | 1476.3945 (+0.0276) | 1440.2732 (+0.0437) | -2.4466% |
| LightGBM fair | – | 1793.9142 (+0.2151) (+0.2486) | 1794.3334 (+0.2458) (+0.3002) | -0.0234% |

Table F8: Poisson deviance comparison across models: (i) in-sample (training) set ($n = 4{,}889$) to fit the discriminatory GLM and LightGBM models, (ii) a first out-of-sample (test) set ($n = 1{,}630$) to predict discriminatory premiums and correct for fairness and autocalibration, and (iii) a second out-of-sample (test) set ($n = 1{,}630$) of new policyholders, for which corrected premiums were predicted using a Random Forest trained on the first test set. Values in parentheses compare each model relatively (i.e., $(a - b)/b$) with the GLM and LightGBM at the same evaluation stage. $\Delta = (a - b)/b$ measures the relative deterioration when moving from the first to the second out-of-sample (new policyholders).

## F.3 Fairness adjustment G>2

For $G > 2$, the predicted premiums $\hat{\pi}$ are partitioned into $G$ vectors $\{\hat{\pi}_g\}_{g=1}^{G}$ of sizes $n_g$ organized in $d$-dimensional variables $U$ and $V$ as in Eq. (11). In this setting, equal-sized samples are drawn from each group in $U$, and for each group in $V$. A uniform random permutation is also applied to ensure the pairwise comparisons span all group combinations (see algorithm 1).

Fig. F7 illustrates the fairness adjustment of $\hat{\pi}$ across four demographic groups, determined by the policyholders' gender (female or male) and ethnicity (minority or majority). The solid lines correspond to the initial predicted premiums partitioned into $G$ demographic groups $\{\hat{\pi}_g\}_{g=1}^{4}$. The overlapping dashed lines correspond to the corresponding fairness adjusted premiums $\{\hat{\pi}_g^{\text{fair}}\}_{g=1}^{4}$. We observe that after adjustment, the dashed lines largely overlap within each gender-ethnicity pairing, indicating that the $\mathcal{E}$-adjuster has effectively reduced disparity across these four groups.

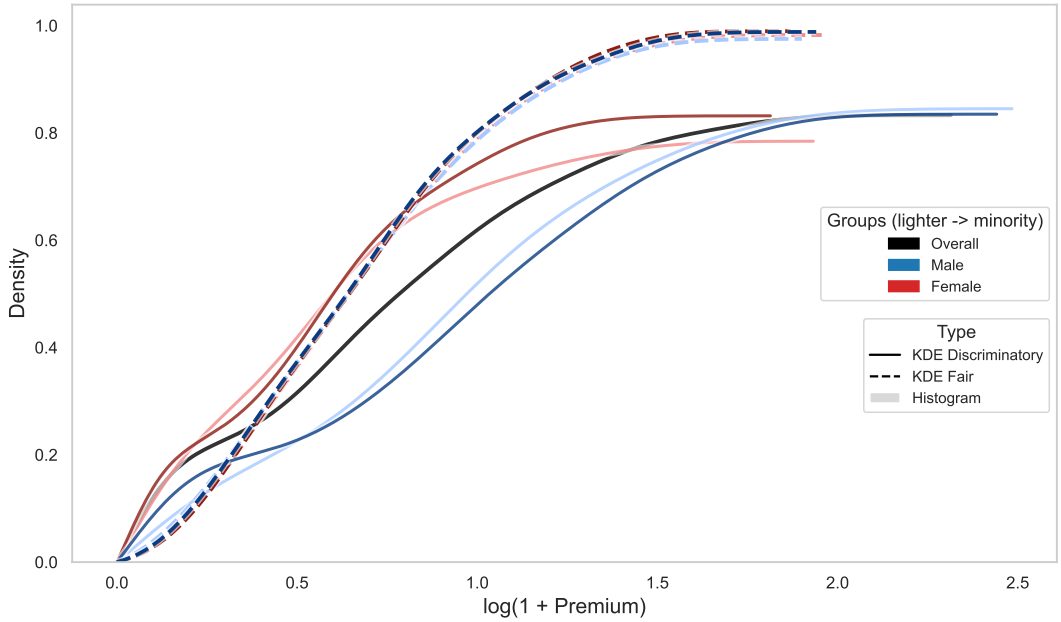


Figure F7: Kernel density estimates (KDEs) of predicted premiums by gender and age group. Solid lines represent the LightGBM discriminatory premiums, and dashed lines the fairness-adjusted premiums. The male groups are shown in blue, the female groups in red, and the overall portfolio in black. The lighter color shade corresponds to the minority group.

Table F9 reveals that the fairness adjustment actually mitigated the deviation from the mean value.

| Statistic | $\text{GLM}_{\beta_0}$ | LightGBM | Fair LightGBM |
|---|---|---|---|
| Mean $\overline{\pi}$ | 1.0609 | 1.0119 | 1.0597 |
| 10% Quantile | | 0.0099 | 0.2556 |
| 90% Quantile | | 2.4771 | 2.2170 |

Table F9: Summary statistics of expected premiums $\hat{\pi}$ for the baseline intercept-only GLM ($\text{GLM}_{\beta_0}$), standard LightGBM, and fairness-adjusted LightGBM. Highlighted values indicate divergence from the target mean premium, illustrating the effect of fairness adjustment.

Table F10 summarises the deviance loss before and after fairness adjustment (for $G = 2$ and $G = 4$). We observe overall a relatively stark deterioration in predictive accuracy following fairness adjustement, especially when adjusting across four protected groups. In the simulated study, model adjustment relied on 100 000 policies (with 60% for training and 20% for testing). Here, however, fairness adjustments are derived from only 1 630 policies in the first out-of-sample set, further split across two or four sensitive groups, leaving roughly 400 premium predictions per group. Using such small group-specific samples to estimate conditional distributions and to enforce demographic parity clearly leads to poor predictive accuracy (in terms of Poisson deviance).

| Model | Poisson Deviance | Δ GLM | Δ LightGBM |
|---|---|---|---|
| GLM | 1436.7176 | | |
| LightGBM | 1476.3944 | +0.0276 | |
| LightGBM fair (gender) | 1793.9142 | +0.2486 | +0.2151 |
| LightGBM fair BC (gender) | 1667.6080 | +0.1607 | +0.1295 |
| LightGBM fair BC (gender & ethnicity) | 4718.2180 | +2.284 | +2.1958 |

Table F10: Poisson deviance comparison at the same evaluation stage (out-of-sample 1) for the generalised linear model (GLM), bias-corrected (BC), and fairness-adjusted versions of the LightGBM model. Lower deviance indicates a better fit to the observed claim frequencies. Δ compare each fairness (and balance) corrected model relative deterioration with the GLM and LightGBM.

# G. Wasserstein and MMD benchmarks

To further benchmark the proposed gradient-based fairness correction, we compare it against two post-processing baselines. First, against a distributional alignment method based on kernel density estimation, optimal transport barycenters, and quantile remap-

ping. Second, against a maximum mean discrepancy (MMD)-based gradient adjustment applied directly on the predictions. Both approaches enforce demographic parity by reducing distributional discrepancies between the conditional predictive distributions of men and women after training (here, LightGBM predictions).

## G.1 Wasserstein (OT)

Because predictions lie in $[0, 1]$, we estimate bounded densities using a logit-transformed kernel density estimator:

$$f_G(\pi) = \exp\big(\mathrm{KDE}(\mathrm{logit}(\pi))\big)(\pi(1-\pi))^{-1}, \qquad G \in W, M.$$

From these KDEs we compute empirical PDF $f_W, f_M$ and CDF $F_W, F_M$ estimates.

Before applying optimal transport correction, a simple baseline consists of scaling each group's predictions to match the global mean:

$$\hat{\pi}_G^{\text{scaled}} = \frac{\bar{\pi}}{\bar{\pi}_G}\hat{\pi}_G.$$

This ensures equal means but leaves distributional differences largely untouched, precisely the limitation the barycenter method addresses.

Let $w_W$, $w_M$ be empirical group weights. For each group CDF $(F_G)$, define the barycentric target distribution $F^\star$ through a pair of reference-specific barycenters:

$$\begin{aligned} F_W^\star(\pi^\star) &= F_W(\pi), \qquad \pi^\star = w_W\pi + w_M F_M^{-1}(F_W(\pi)), \\ F_M^\star(\pi^\star) &= F_M(\pi), \qquad \pi^\star = w_M\pi + w_W F_W^{-1}(F_M(\pi)). \end{aligned}$$

Both constructions lead to the same Wasserstein barycenter $F^\star$, whose density $f^\star$ is obtained by differentiating the monotone transport map. This $F^\star$ acts as the demographically fair predictive distribution toward which both groups are transported.

For each individual prediction $\hat{\pi}_i$, we first compute the group-specific empirical quantile:

$$u_i = F_G(\hat{\pi}_i).$$

We map it through the opposite group's inverse CDF:

$$T_G(\hat{\pi}_i) = F_{\bar{G}}^{-1}(u_i).$$

We then construct the barycentric transport map:

$$\hat{\pi}_i^{\text{fair}} = w_G, \hat{\pi}_i + w_{\bar{G}}, T_G(\hat{\pi}_i).$$

This yields a set of adjusted predictions $\hat{\pi}^{\text{fair}}$ whose distribution matches the barycenter

$F^\star$. In practice, this approach nearly eliminates discrepancy:

$$W_1(F_W^{\text{fair}}, F_M^{\text{fair}}) \approx 0, \quad F_W^{\text{fair}} \simeq F_M^{\text{fair}} \simeq F^\star.$$

## G.2 MMD-based gradient fairness adjustment

As a second baseline, we implement a gradient-based fairness correction driven by the Maximum Mean Discrepancy (MMD) between group-conditional predictions. Both the Wasserstein-OT approach and the MMD gradient method aim at enforcing demographic parity at the distributional level, but while the former constructs an explicit transport map between empirical distributions, the latter performs iterative RKHS alignment in prediction space.

Let $M = \{\hat{\pi}_i : S_i = M\}$ and $W = \{\hat{\pi}_j : S_j = W\}$. The squared MMD between the two groups under an RBF kernel $k_\sigma$ is:

$$\text{MMD}^2(W, M) = \mathbb{E}[k_\sigma(W, W')] + \mathbb{E}[k_\sigma(M, M')] - 2\mathbb{E}[k_\sigma(W, M)]\,.$$

We minimize this quantity by directly updating predictions via gradient descent:

$$\hat{\pi} \leftarrow \hat{\pi} - \eta \nabla_{\hat{\pi}} \text{MMD}^2(W, M)\,,$$

where $\eta$ is a step size.

The corresponding gradients for individual samples in group M and W are given by:

$$\nabla_{\hat{\pi}_i \in M} = -\frac{2}{n_M^2 \sigma^2} \sum_{i'} (\hat{\pi}_i - \hat{\pi}_{i'}) k_\sigma(\hat{\pi}_i, \hat{\pi}_{i'}) + \frac{2}{n_M n_W \sigma^2} \sum_{j} (\hat{\pi}_i - \hat{\pi}_j) k_\sigma(\hat{\pi}_i, \hat{\pi}_j)\,,$$

$$\nabla_{\hat{\pi}_j \in W} = -\frac{2}{n_W^2 \sigma^2} \sum_{j'} (\hat{\pi}_j - \hat{\pi}_{j'}) k_\sigma(\hat{\pi}_j, \hat{\pi}_{j'}) + \frac{2}{n_M n_W \sigma^2} \sum_{i} (\hat{\pi}_j - \hat{\pi}_i) k_\sigma(\hat{\pi}_j, \hat{\pi}_i)\,.$$

This procedure iteratively adjusts predictions such that the RKHS embeddings of both groups become aligned, yielding:

$$\text{MMD}(\hat{\pi}_W^{\text{fair}}, \hat{\pi}_M^{\text{fair}}) \to 0\,.$$

Note that while MMD is a well-established discrepancy measure, the empirical MMD requires all pairwise kernel evaluations between samples across and within groups, resulting in $\mathcal{O}(n^2)$ complexity per iteration. Although approximations such as mini-batch MMD or random Fourier features can reduce computational cost, they introduce additional stochastic variance in the estimated discrepancy. In practice, these approaches require careful tuning of batch size, learning rate, and kernel bandwidth. These parameters are known to strongly affect both convergence stability and the quality of the resulting alignment.

Given that our goal is not to design a large-scale stochastic optimization procedure,

but rather to evaluate MMD as a deterministic distributional alignment baseline, we adopt the full-batch formulation to ensure numerical stability and reproducibility.

We further fix the kernel bandwidth to 0.0622. It is a median heuristic bandwidth choice, which is one of the most standard (and surprisingly effective) ways to set the kernel scale in MMD.

We first evaluated the MMD gradient adjustment using a learning rate $\eta$ of 0.01. This configuration required approximately 143 minutes for 500 iterations and resulted in only marginal changes in the group distributions, which exhibited only a slight convergence toward each other. Increasing the learning rate to 0.1 still did not yield meaningful overlap between the marginal distributions after 500 iterations. For comparison, the energy-based gradient method converged in fewer than 70 iterations under the same experimental setting with $\eta = 0.001$. Further increasing the learning rate to 1.0 led to faster displacement of the distributions (124 minutes for 500 iterations), with the marginals beginning to overlap. However, the resulting alignment remained insufficient to achieve demographic parity at a level comparable to the energy-gradient method. Finally, we chose to set the learning rate to 1.5, to improve distributional alignments.

In this setting, MMD-based gradient alignment exhibits slower convergence and higher sensitivity to optimization hyperparameters compared to energy-distance-based corrections.

## G.3 Results

As shown in Table G1, the discriminatory model exhibits a positive group mean difference, indicating that men tend to receive higher premium than women. The energy-based adjustment removes the mean disparity exactly, outperforming both MMD and Wasserstein. Wasserstein slightly over-corrects the gap and produces a negative mean difference ($-0.0054$), whereas MMD slightly undercorrects it ($+0.0008$). Although the Wasserstein barycenter enforces equality of marginal distributions by construction, it requires the numerical inversion of empirical CDFs. This step is inherently less stable, especially in sparse regions or in the presence of heavy tails (common in insurance data). The fact that MMD is kernel-based may also explain the residual differences, as further reductions in marginal discrepancies become increasingly difficult.

| Model | Group mean difference (men − women) |
|---|---|
| LightGBM | 0.0630 |
| Energy LightGBM | 0.0000 |
| Wasserstein LightGBM | -0.0054 |
| MMD LightGBM | 0.0008 |

Table G1: Group mean differences (men − women) for predicted claim frequencies. Positive values indicate that men have higher predicted frequencies than women.

Fig. G1 shows the different effects of the fairness adjustments per group. Both methods reduce premiums for men while increasing premiums for women. In the lower and middle quantiles, the Wasserstein-corrected premium (solid line, with crosses) are overall slightly closer to the observed discriminatory predictions (dotted line). Wasserstein adjustment, however, introduces a stronger distortion in higher frequencies bins. MMD adjustments induce similar stronger distortions than Energy ones in the upper tails. Overall, women see substantially higher premiums than originally, while men receive lower premiums. This highlights a trade-off between overall distributional parity and extreme per-group distortions, particularly for high-risk policyholders.

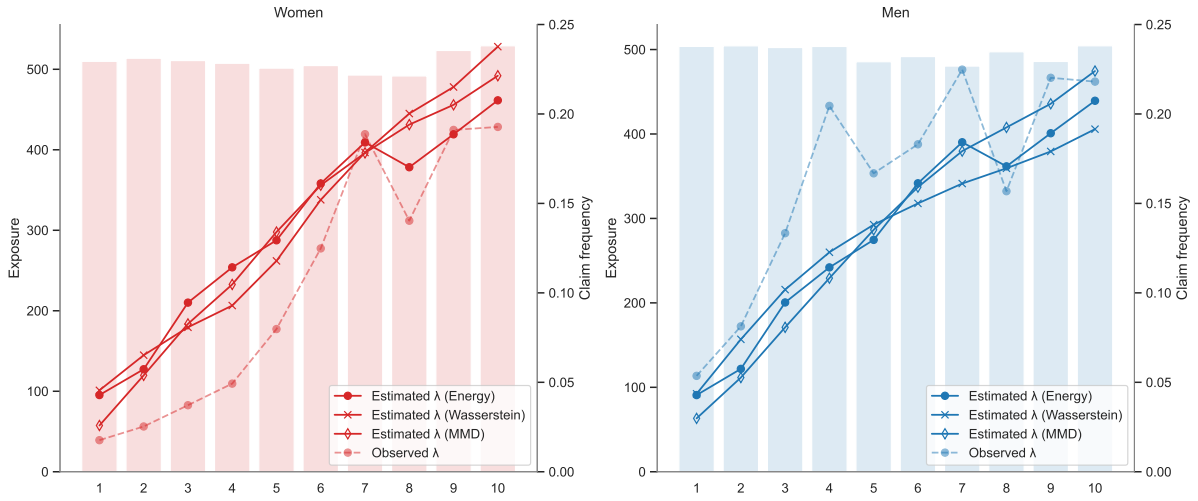


Figure G1: Comparison of predicted and observed claim frequencies by gender per unit exposure across 10 LightGBM quantile-based bins. The dashed line shows the observed frequency $\lambda$ computed from the data. The solid line shows the estimated frequency $\hat{\lambda}$. Energy-based adjustments are indicated by circles, Wasserstein adjustments by crosses, and MMD by diamonds. Colors differentiate gender.

Fig. G2 further shows a comparison of original and adjusted predicted premiums for male and female policyholders. The energy-based adjustment leads to only mild distortion in the tail, and its group-specific effects look stronger for men (further from the diagonal). MMD clearly leads to the strongest adjustments for men (further from the diagonal) in the lower and mid-tail. In contrast, the Wasserstein adjustment leads to a stronger

deformation of the upper tail, reflecting a heavier redistribution of extreme values. Unlike Wasserstein-based corrections, which explicitly couple quantiles across groups and induce visible redistribution in the tails, MMD and Energy do not enforce a structured cross-group matching of quantiles, leading to less pronounced adjustments in the extremes of the distribution.

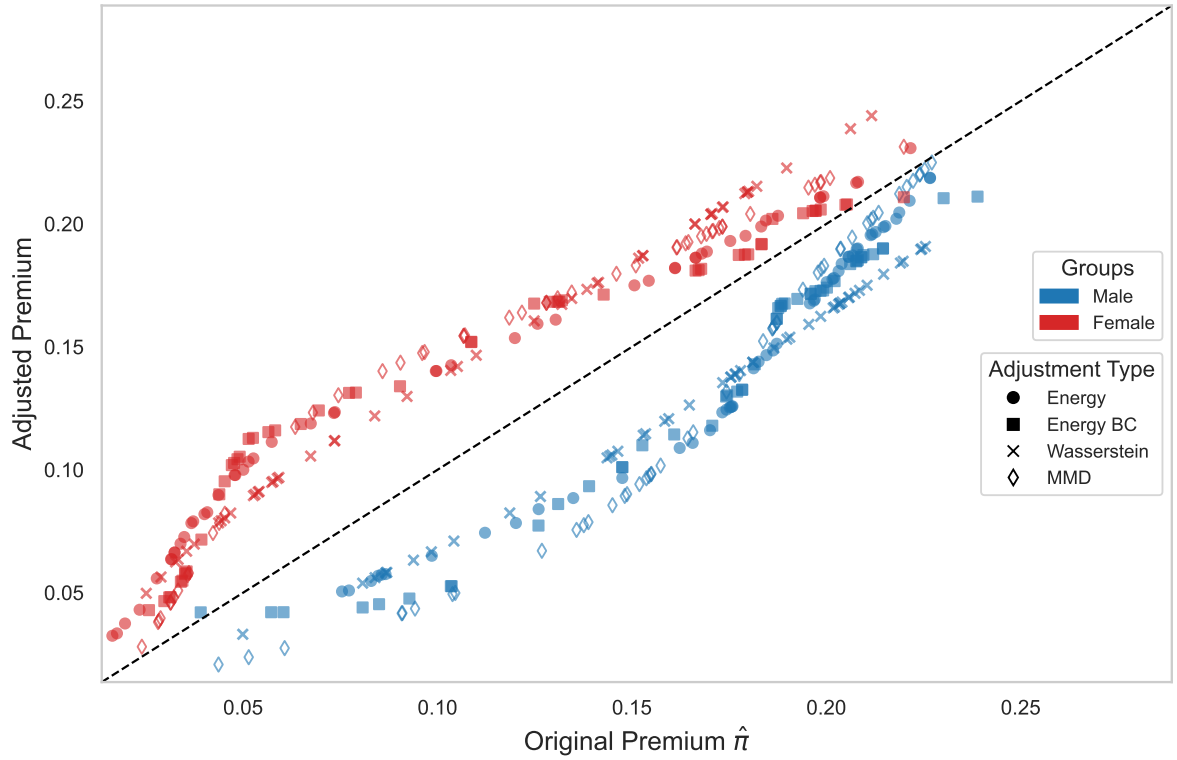


Figure G2: Scatter plot of original discriminatory premiums and adjusted premiums for male and female policyholders. Each point represents an individual prediction, with the x-axis showing the original LightGBM-predicted premium $\hat{\pi}$ and the y-axis showing the corresponding adjusted premium. Energy-based adjustments are indicated by circles, Energy-based and BC-corrected by squares, Wasserstein barycenter by crosses, and MMD by diamonds. Colors differentiate gender: blue for male and red for female. The dashed black identity line indicates perfect preservation of the original discriminatory predictions. Subsampling (10% of points) is applied for visualization clarity.

Table G2 provides a quantitative assessment of how disruptive each correction is with respect to the original risk differentiation. Fairness adjustments should mitigate unfair dependence on protected attributes without disrupting the core risk ranking. High-risk policyholders should still receive higher premiums than low-risk ones. Otherwise, this would break actuarial credibility, tariff stability, and portfolio steering, leading to anti-selection and misaligned underwriting incentives.

The Spearman (1904) rank correlation between the discriminatory and adjusted premiums in Table G2 show that all corrections affect the initial ordering of risks. The energy-based adjustment preserves it slightly more effectively ($\rho = 0.8604$) than the Wasserstein barycentric correction ($\rho = 0.8266$). Wasserstein barycenters enforces equality at each

quantile level. This can locally reorder individuals, especially in distributional tails, leading to more pronounced rank distortions.

| Model | Spearman $\rho$ |
|---|---|
| Energy LightGBM | 0.8604 |
| Wasserstein LightGBM | 0.8266 |
| MMD LightGBM | 0.8615 |

Table G2: Spearman rank correlation between discriminatory and adjusted premium.

Table G3 further reports the rank correlation by gender. Because Spearman correlation only depends on ranks, values close to 1 indicate that the transformation applied by each method is almost monotone within each group. This is exactly what happens here. All methods act essentially as group-specific monotone recalibrations, especially Wasserstein, so they do not meaningfully reorder individuals within the same gender.

| Model | Men | Women |
|---|---|---|
| Energy LightGBM | 0.9998 | 0.9998 |
| Wasserstein LightGBM | 1.0000 | 1.0000 |
| MMD LightGBM | 0.9999 | 1.0000 |

Table G3: Spearman rank correlation between discriminatory and adjusted premium, by gender.

Fig. G3 evaluates conditional demographic parity:

$$\mathbb{E}\left[\hat{\pi}_{\text{fair}}|\text{men}, q_{\hat{\pi}}\right] = \mathbb{E}\left[\hat{\pi}_{\text{fair}}|\text{women}, q_{\hat{\pi}}\right], \tag{G1}$$

where $\hat{\pi}_{\text{fair}}$ are the Energy, Wasserstein, or MMD corrected premium, and $q_{\hat{\pi}}$ are the quantiles of the original discriminatory premium. This shows local fairness, not just global parity.

The Wasserstein and Energy-corrected premium show mild differences in the middle quantiles. Wasserstein stays consistently negative across all upper quantiles, showing persistent conditional disparity across premium levels. This indicates that women tend to receive slightly higher predicted premiums than men, but the deviation is stable across the risk spectrum. Whereas Energy returns toward zero in the upper tail, improving conditional parity. Energy seems to better enforce parity for higher-risk policyholders. The MMD curve shows stronger non-monotonicity across the risk spectrum, with relatively large negative disparities in mid-quantiles (down to approximately -0.10) followed by partial recovery in the upper tail (similarly to Energy).

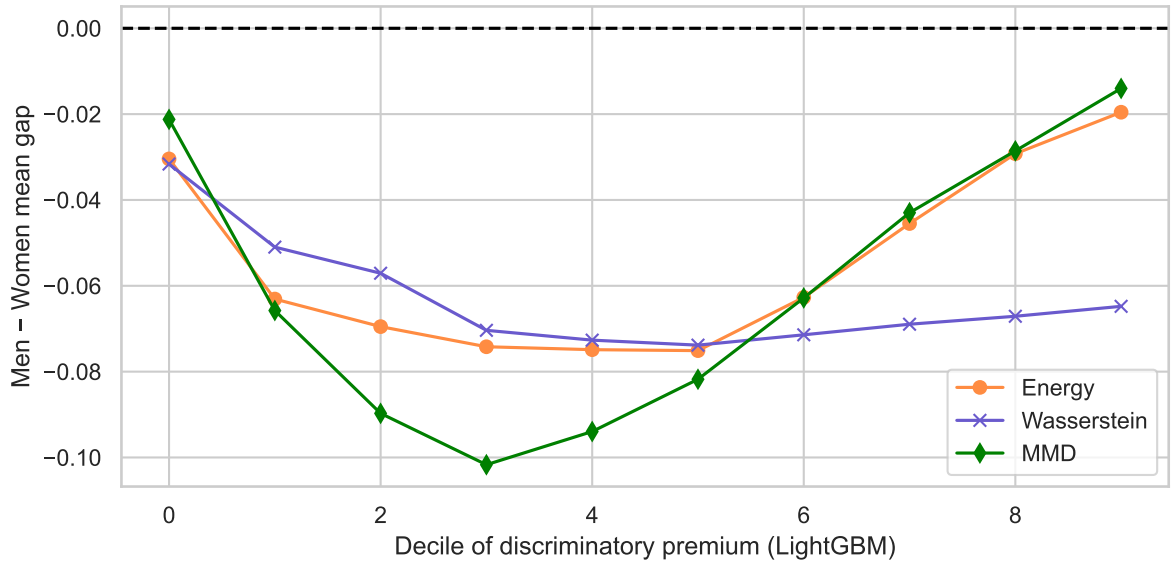


Figure G3: Conditional demographic parity gaps (men − women) by discriminatory premium quantile for each fairness method. Negative values indicate higher predicted premiums for women.

These empirical differences are consistent with the theoretical properties of the metrics.

The Energy distance acts as a global discrepancy measure. It integrates distributional differences over the entire support. Since the upper tail typically represents only a small fraction of the probability mass (5% here), its contribution to the Energy distance gradient is small, implying that the optimisation primarily aligns the bulk of the distribution.

The Wasserstein distance, in contrast, operates directly on quantile differences, giving each quantile, including extreme ones, equal weight in its objective. As a result, any misalignment in the far right tail contributes proportionally to the objective, forcing the correction to act more aggressively on large predicted values.

The Maximum Mean Discrepancy (MMD) sensitivity is strongly governed by the choice of kernel bandwidth. Small bandwidth translates to very local, almost point matching, alignments. Whereas large bandwidth translates to very global, almost mean matching, alignments. Consequently, MMD can induce stronger localised adjustments than Energy, particularly in low-density or mid-tail regions, while remaining less disruptive in the extremes than Wasserstein.

Hence, from a distributional standpoint, the Energy distance preserves the extremal structure more faithfully, the Wasserstein distance enforces stricter quantile-level parity at the cost of stronger distortions in the tail, and MMD provides a kernel-based comparison that can capture both local and global discrepancies without explicit quantile coupling.

Note that both energy-based and MMD-based gradient adjustments are unconstrained functional perturbations of the prediction surface, meaning they do not inherently enforce positivity of the output. Since premiums are updated via gradient-driven corrections, sufficiently large updates (especially in sparse regions or tails) can in principle push some corrected values below zero unless an explicit positivity constraint is imposed (e.g., floor

constraint or a log-link parameterisation in the case of count data). In our experiments, we verified that no negative premiums were produced after applying the corrections.