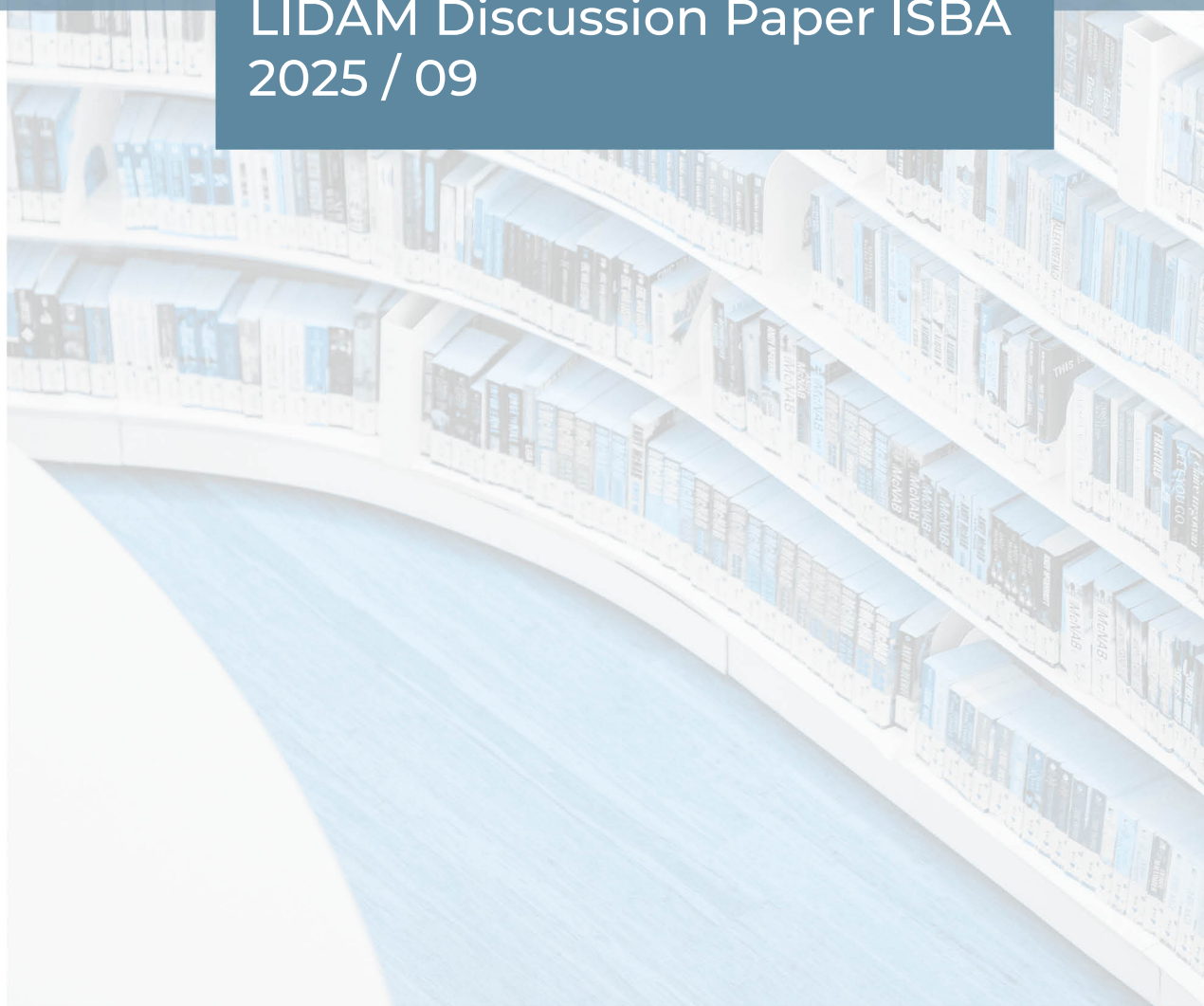
A MULTIVARIATE ENERGY-BASED FAIRNESS ADJUSTER FOR PREMIUMS

Charlotte Jamotton, Donatien Hainaut

LIDAM Discussion Paper ISBA
2025 / 09



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 has received increasing attention, particularly in light of regulations calling for non-discriminatory premium estimation, such as the EU Gender Directive (2012). This research explores group fairness mechanisms designed to mitigate group-level disparities. We propose a method based on the energy distance, a multivariate metric that enables fairness adjustments across multiple sensitive attributes simultaneously, even where they are non-binary. To maintain overall prediction balance, our approach integrates an autocalibration step to correct for biases in the total predicted number of claims. Moreover, our method supports fairness adjustments for new policyholders without requiring retraining of the fairness model. We evaluate the proposed methodology in the context of car insurance pricing, where demographic factors such as age and gender are commonly used for risk assessment and premium determination. Results show that our method effectively reduces group-level disparities while preserving the predictive accuracy.

Keywords: fairness, demographic parity, autocalibration, actuarial pricing, energy distance

*Corresponding author: charlotte.jamotton@uclouvain.be
†donatien.hainaut@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. To estimate 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 (Gandy, 2016). As Seligman (1983) questions: if young women have fewer automobile accidents than young men—and they do, on average—should women not receive better rates? Similarly, if industry data shows that women spend more time in hospitals than men, why should women not pay more for health policies?

This phenomenon, often described as an injustice by generalization (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 (Vandenhole, 2005), 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 & Huang (2023) 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 (Birnbaum, 2020): 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 mitigation in causal theory (Côté et al., 2024). Inspired by post-stratification techniques, it consists of an average over the sensitive information and a change of measure to prevent the implicit inference of protected characteristics from allowed attributes. Conversely, group fairness methodologies, such as those using Wasserstein barycenters proposed by Charpentier et al. (2023) and extended by Hu et al. (2023, 2024), recognize the need for equity across demographic groups. They emphasize the importance of aligning outcomes across diverse demographic groups such that no association between sensitive information and premiums is allowed. According to Côté et al. (2024, 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 & Huang (2024) and Côté et al. (2024).

In this article, we explore the application of Demographic Parity (DP) as a group fairness criterion in insurance pricing. While EU regulations on gender-neutral pricing align more with an individual fairness perspective, we argue that DP, as a group fairness criterion, 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 individualized but structured around groups with shared characteristics. By adopting DP, we ensure that the probability of receiving a certain premium level remains 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 article is structured as follows. In Section 2, we set the notations and define Demographic Parity (DP) as a group fairness criterion. In Section 3, we provide a theoretical justification for choosing the energy distance to quantify disparities between premium distributions across sensitive demographic groups. We also discuss its relationship with other well-known statistical distances, such as Kullback-Leibler (KL) divergence and Wasserstein distance. In Section 4, we highlight the multivariate nature of the energy distance, which allows for simultaneous fairness adjustments across multiple sensitive demographic groups. This is a key advantage of our approach. It can handle variables with different numbers of sensitive classes, such as gender (binary) and discretized age (multi-class), at once, and is therefore order-independent. In Section 4, we additionally introduce the autocalibration technique (Denuit et al., 2021) to correct biases in the total number of claims following fairness adjustments, and we propose a methodology to adjust premiums for new policyholders without retraining the fairness model. To the best of our knowledge, existing methodologies do not support the adjustment for new policyholders. Finally, in Section 5, we evaluate the proposed methodology in the context of car insurance pricing,

where demographic factors such as age and gender are commonly used for risk assessment and premium determination. Results show that our method effectively reduces group-level disparities while preserving the predictive accuracy.

2. Notations and definitions

Consider an insurance portfolio $\mathcal{X} \subset \mathbb{R}^m$ represented as a tabular dataset with n rows and m columns. Each row $i = 1, \ldots, n$ represents a policy defined by a vector of characteristics, denoted $\boldsymbol{x}_i$. Each column $j = 1 \ldots, m$ represents a rating factor. We distinguish the allowed rating factors, denoted $\boldsymbol{X} \in \mathcal{X}$, from the discriminatory rating factors, denoted $\boldsymbol{D} \in \mathcal{X}$. For each policy, we observe the number of claims N_i (count data) reported during its contract duration (or exposure) ν_i. Throughout the analysis, we will often refer to the claims frequency per unit of exposure, defined as the ratio $\lambda_i = N_i/\nu_i$.

In actuarial pricing, the goal is to approximate the (unknown) conditional expectation $\mu(\boldsymbol{X}, \boldsymbol{D}) = \mathbb{E}[Y|(\boldsymbol{X}, \boldsymbol{D})]$ of the response variable Y (here, the number of claims) given the available information $(\boldsymbol{X}, \boldsymbol{D})$. In the empirical analysis in Section 5, we use a Light Gradient Boosting machine (LGB) to model the premium best-estimates $\hat{\pi}(\boldsymbol{X}, \boldsymbol{D})$ for the (true) pure premium $\mu(\boldsymbol{X}, \boldsymbol{D})$.

Definition 2.1 (Group Fairness). Group fairness requires the distribution of the predicted outcome $\hat{\pi}(\boldsymbol{X}, \boldsymbol{D})$ to be independent of the protected attributes $\boldsymbol{D}$. Formally, group fairness is satisfied if:

$$\hat{\pi}(\boldsymbol{X}, \boldsymbol{D}) \perp \boldsymbol{D}. \tag{1}$$

Definition 2.2 (Demographic Parity (DP)). Demographic Parity (DP) is a group fairness criterion that requires the distribution of the predicted outcomes $\hat{\pi}(\boldsymbol{X}, \boldsymbol{D})$ to be identical across different demographic groups, ensuring that no group receives systematically different pricing based on protected attributes $\boldsymbol{D}$. Formally, fairness at the group level across two groups a and b is achieved if:

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

In a regression setting, statistical distances are commonly used to quantify the disparities between conditional distributions $\hat{\pi}(\boldsymbol{X}, \boldsymbol{D})|D = a$ and $\hat{\pi}(\boldsymbol{X}, \boldsymbol{D})|D = b$. As discussed in Section 3, there exists a wide range of distances/divergences to quantify the disparities between two distributions, including the Kullback-Leibler divergence and the Wasserstein distance. In Section 4, we propose the use of the energy distance, defined in Section 3, to quantify these disparities and to align premium distributions across sensitive demographic groups in order to satisfy the DP condition in Eq.(2).

3. Statistical Distances

In machine learning, Kullback & Leibler (1951) divergence is one of the most commonly used measures to quantify the dissimilarity between two probability distributions. For distributions P and Q, the KL divergence is defined as:

$$\mathrm{d}_{KL}\left(P \,\|\, Q\right) = \int_{-\infty}^{\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx, \tag{3}$$

where p and q are the probability densities of distributions P and Q, respectively.

As a measure of relative entropy, KL divergence focuses on the relative distance between distributions but does not account for the geometric proximity between outcomes. For instance, consider an actuarial context where $p(x)$ assigns a 0.01% probability to a premium level $x = 100$, while $q(x)$ assigns the same probability to $x = 101$. Despite the values being very close, the KL divergence treats them as entirely different outcomes due to its reliance on relative entropy.

To address this limitation, researchers (Panaretos & Zemel, 2019; Peyré, Cuturi et al., 2019) have turned to alternative distance metrics that consider the underlying geometry between outcomes, such as the Wasserstein probability metric.

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 & Rubinstein, 1958):

$$W_p(P, Q) = \left(\inf_{\pi \in \Pi(P,Q)} \int \|x - y\|^p d\pi\left(x, y\right)\right)^{1/p}, \tag{4}$$

where $\Pi(P, Q)$ denotes the set of all joint distributions for the pair (X, Y) with marginals P and Q, respectively. The infimum is taken over all possible couplings $\pi \in \Pi$, ensuring that the optimal transport plan minimizes the cost of transporting mass from x to y (Villani et al., 2008).

Unlike the KL divergence, the Wasserstein distance does not possess the unbiased sample gradients property, one of the three key properties desired for a divergence (the other two being sum invariance and scale sensitivity). This is an important consideration in machine learning when using a metric as a loss function. Specifically, when we minimize the loss function $\theta \mapsto \mathrm{d}\left(P, Q_\theta\right)$ to find the optimal parameters θ^* of a distribution Q_θ.

Definition 3.1 (Unbiased sample gradients property). A measure d is said to have unbiased sample gradients if, for all m independent samples $\boldsymbol{x}_m$ 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):

$$\mathbb{E}_{\boldsymbol{x}_m \sim P} \nabla_\theta \mathrm{d}\left(\hat{P}_m, Q_\theta\right) = \nabla_\theta \mathrm{d}\left(P, Q_\theta\right). \tag{5}$$

If a metric lacks unbiased sample gradients, optimizing it using gradient descent may fail or converge to an incorrect minimum.

Interestingly, the p-Wasserstein distance can be expressed in terms of the quantile functions F_P^{-1} and F_Q^{-1} for one-dimensional distributions P and Q as follows:

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

For $p = 1$, this simplifies to the l_1-distance (a member of the l_p family) between the cumulative distribution functions (CDFs) F_P and F_Q:

$$W_1(P,Q) = \int_0^1 \left| F_P^{-1}(x) - F_Q^{-1}(x) \right| dx = \int_{-\infty}^{+\infty} \left| F_P(x) - F_Q(x) \right| dx = l_1^1(P,Q). \tag{7}$$

A related measure is the Cramér (1928) distance, which is defined as the l_2^2-distance between the CDFs of the two distributions:

$$l_2^2\left(P,Q\right) = \int_{-\infty}^{\infty} \left| F_P(x) - F_Q(x) \right|^2 dx, \tag{8}$$

whose square root is a member of the l_p family of metrics:

$$l_p\left(P,Q\right) = \left(\int_{-\infty}^{\infty} \left| F_P(x) - F_Q(x) \right|^p dx \right)^{1/p}. \tag{9}$$

The Cramér distance (Cramér, 1928) combines the strengths of both the Wasserstein and KL metrics, as it considers the underlying geometry of the outcomes while respecting the unbiased sample gradients property (along with sum invariance and scale sensitivity).

Proposition 3.1 (Unbiased sample gradient for l_2^2-distance). Let P be a probability measure and Q_θ be a parameterized family of probability measures with cumulative distribution functions $F_P(x)$ and $F_{Q_\theta}(x)$, respectively. The gradient of the squared l_2-distance satisfies the following relationship (Bellemare et al., 2017, p.15):

$$\nabla_\theta \, l_2^2\left(P, Q_\theta\right) = \mathbb{E}_{\boldsymbol{x}_m \sim P}\left[\nabla_\theta \, l_2^2\left(\hat{P}_m, Q_\theta \right) \right]. \tag{10}$$

The unbiased sample gradient for the l_2^2-distance is demonstrated in Appendix 6.

Although, in theory, the Cramér distance combines the best of Kullback-Leibler and Wasserstein, it is limited to univariate distributions. In the context of actuarial fairness, real-world scenarios often involve multiple sensitive attributes. Interestingly, the Cramér distance in Eq.(8) can be expressed in the following form:

$$l_2^2\left(P,Q\right) = \mathbb{E}[|X - Y|] - \frac{1}{2}\mathbb{E}[|X - X'|] - \frac{1}{2}\mathbb{E}[|Y - Y'|], \tag{11}$$

where $|.|$ denotes the absolute value for scalars $x, y \in \mathbb{R}$, and X' and Y' are independent and identically distributed copies of X and Y.

As shown by Székely (2003), Eq.(11) allows for the generalization of the Cramér dis-

tance to higher dimensions, known as the energy distance. For d-dimensional independent random variables $X, Y \in \mathbb{R}^d$, the energy distance is defined as:

$$\mathcal{E}(P,Q) = 2\mathbb{E}[\|X-Y\|] - \mathbb{E}[\|X-X'\|] - \mathbb{E}[\|Y-Y'\|], \tag{12}$$

where $\|\cdot\|$ denotes the Euclidean norm.

The equivalence $\mathcal{E}(P,Q) = 2l_2^2(P,Q)$ for univariate distributions P and Q is demonstrated in Appendix 6.

Proposition 3.2 (Unbiased sample gradient for $\mathcal{E}$-distance). For $X \sim P$, $Y \sim Q_\theta$, and $\hat{X}_m \sim \hat{P}_m$, the gradient ∇ of $\mathcal{E}(P, Q_\theta)$ satisfies the following expression, as noted in Bellemare et al. (2017, p.15):

$$\nabla_\theta \mathcal{E}(P, Q_\theta) = \mathbb{E}_{\boldsymbol{x}_m \sim P}\left[\nabla_\theta \mathcal{E}\left(\hat{P}_m, Q_\theta\right)\right] \tag{13}$$

The unbiased sample gradient for $\mathcal{E}$-distance is demonstrated in Appendix 6.

Another interesting aspect of the $\mathcal{E}$-distance is its similarity to the Wasserstein distance. The Wasserstein distance in Eq.(4) can be defined in terms of expectations over coupled random variables $X \sim P$ and $Y \sim Q$:

$$W_p(P,Q) = \inf_{\pi \in \Pi(P,Q)} \mathbb{E}_{(X,Y)\sim\pi}\left[\|X-Y\|^p\right]^{1/p}. \tag{14}$$

Since the energy distance $\mathcal{E}(P,Q)$ involves $\mathbb{E}\|X-Y\|$, we can consider minimizing $\mathbb{E}\|X-Y\|$ over all possible couplings π of $X \sim P$ and $Y \sim Q$ such that:

$$\mathcal{E}(P,Q) = \inf_{\pi \in \Pi(P,Q)} 2\mathbb{E}_{(X,Y)\sim\pi}\left[\|X-Y\|\right] - \mathbb{E}\left[\|X-X'\|\right] - \mathbb{E}\left[\|Y-Y'\|\right]. \tag{15}$$

Since the additional terms $\mathbb{E}\|X-X'\|$ and $\mathbb{E}\|Y-Y'\|$ depend solely on the distributions P and Q, they remain constant with respect to the coupling π. Consequently, when $p=1$, the infimum over all couplings in Eq.(14) coincides with the formulation of the energy distance in Eq.(15).

4. Energy distance fairness adjuster

In this Section 4, we introduce a fairness adjustment mechanism based on the energy distance metric, denoted by $\mathcal{E}$. As outlined in Section 2, our goal is to mitigate disparities in predicted premiums across $G \geq 2$ sensitive demographic groups by enforcing demographic parity. More specifically, we seek to align the distribution of predicted premiums $\hat{\pi}(X, D)$ across all values of the sensitive attribute(s) $\boldsymbol{D}$.

Let $X, Y \in \mathbb{R}^d$ be two independent random variables drawn from distributions P and Q, respectively. The energy distance $\mathcal{E}(P,Q)$ measures the statistical difference between

these distributions and is defined as:

$$\mathcal{E}(P,Q) = 2\mathbb{E}\left[\|X-Y\|\right] - \mathbb{E}\left[\|X-X'\|\right] - \mathbb{E}\left[\|Y-Y'\|\right],$$

where $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^d$. Notably, $\mathcal{E}(P,Q) = 0$ if and only if X and Y are identically distributed.

Given samples $\{\boldsymbol{x}_i\}_{i=1}^n$ and $\{\boldsymbol{y}_j\}_{j=1}^m$ in $\mathbb{R}^d$, the energy distance can be empirically estimated as:

$$\mathcal{E}(P,Q) := \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\|\boldsymbol{x}_i-\boldsymbol{y}_j\| - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\|\boldsymbol{x}_i-\boldsymbol{x}_j\| - \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j\neq i}\|\boldsymbol{y}_i-\boldsymbol{y}_j\|, \quad (16)$$

where

$$\|\boldsymbol{x}_i-\boldsymbol{y}_j\| = \sqrt{\sum_{k=1}^{d}(x_{ik}-y_{jk})^2}. \quad (17)$$

This formulation captures both the inter-group distance (between X and Y) and the intra-group variability (within X and within Y). The intra-group terms are particularly important for preserving the structure and natural spread of each group's premium distribution, a point we explore further below.

For an observed pair $(\boldsymbol{x}_i, \boldsymbol{y}_j)$, the gradients of the pairwise norm with respect to a component x_{ik} and y_{jk} are:

$$\frac{\partial\|\boldsymbol{x}_i-\boldsymbol{y}_j\|}{\partial x_{ik}} = \frac{x_{ik}-y_{ik}}{\|\boldsymbol{x}_i-\boldsymbol{y}_j\|}, \quad \frac{\partial\|\boldsymbol{x}_i-\boldsymbol{y}_j\|}{\partial y_{jk}} = -\frac{y_{jk}-x_{jk}}{\|\boldsymbol{y}_j-\boldsymbol{x}_i\|}.$$

Thus, the gradients of the energy distance with respect to the individual observations $\{\boldsymbol{x}_i\}_{i=1}^n$ and $\{\boldsymbol{y}_j\}_{j=1}^m$ are:

$$\nabla_{\boldsymbol{x}_i}\mathcal{E}(P,Q) = \frac{2}{nm}\sum_{j=1}^{m}\frac{\boldsymbol{x}_i-\boldsymbol{y}_j}{\|\boldsymbol{x}_i-\boldsymbol{y}_j\|} - \frac{2}{n^2}\sum_{j=1}^{n}\frac{\boldsymbol{x}_i-\boldsymbol{x}_j}{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|} \quad (18)$$

$$\nabla_{\boldsymbol{y}_j}\mathcal{E}(P,Q) = -\frac{2}{nm}\sum_{i=1}^{n}\frac{\boldsymbol{y}_j-\boldsymbol{x}_i}{\|\boldsymbol{y}_j-\boldsymbol{x}_i\|} + \frac{2}{m^2}\sum_{i=1}^{m}\frac{\boldsymbol{y}_j-\boldsymbol{y}_i}{\|\boldsymbol{y}_j-\boldsymbol{y}_i\|} \quad (19)$$

By iteratively updating X and Y using gradient descent, we gradually nudge their distribution toward each other:

$$\begin{pmatrix} X(e+1) \\ Y(e+1) \end{pmatrix} = \begin{pmatrix} X(e) \\ Y(e) \end{pmatrix} - \eta \begin{pmatrix} \nabla_X\mathcal{E}(P,Q) \\ \nabla_Y\mathcal{E}(P,Q) \end{pmatrix}, \quad (20)$$

where η is the learning rate and e indexes the iteration (or epoch).

Consider random variables $X \sim P$ and $Y \sim Q$ correspond to the estimated premiums for female and male policyholders, respectively. For each demographic group $g \in \{1,\ldots,G\}$, let $\hat{\pi}_g \in \mathbb{R}^{n_g\times 1}$ denote the predicted premiums vector for group g, where n_g is

the number of observations. Since we have $G = 2$ demographic groups from a single sensitive variable D_{gender}, variables X and Y can be rewritten as $X = \hat{\pi}_{\text{female}}$ and $Y = \hat{\pi}_{\text{male}}$ with dimension $d = 1$. Therefore, Eq.(20) can be rewritten as:

$$\hat{\pi}^{\text{fair}}(e+1) = \begin{pmatrix} \hat{\pi}_{\text{female}}(e) \\ \hat{\pi}_{\text{male}}(e) \end{pmatrix} - \eta \begin{pmatrix} \nabla_{\hat{\pi}_{\text{female}}} \mathcal{E}(P,Q) \\ \nabla_{\hat{\pi}_{\text{male}}} \mathcal{E}(P,Q) \end{pmatrix}. \tag{21}$$

For $G \geq 2$, we can generalize the premium representation as follows:

$$\hat{\pi} = \{\hat{\pi}_g\}_{g=1}^{G} = \begin{pmatrix} \{\hat{\pi}_g\}_{g=1}^{\lceil G/2 \rceil} \\ \{\hat{\pi}_g\}_{g=\lfloor G/2 \rfloor + 1}^{G} \end{pmatrix} = \begin{pmatrix} X \\ Y \end{pmatrix}. \tag{22}$$

Consider the protected attributes $\boldsymbol{D}$ include both gender $D_1 \in \{\text{female, male}\}$ and age $D_2 \in \{\text{young, middle-aged, senior}\}$, leading to $G = 6$ distinct groups (e.g., young female, middle-aged male, etc.). Given that $G = 6$, X and Y are both 3-dimensional variables where $\{x_{i,1}\}_{i=1}^{n_1}$ represents the young female premiums, $\{y_{j,1}\}_{j=1}^{n_{\lfloor G/2 \rfloor + 1}}$ the young male premiums, $\{x_{i,2}\}_{i=1}^{n_2}$ the middle-aged female premiums, etc. Note that the ordering of dimensions is not meaningful. Dimensions can be treated as exchangeable.

When $G > 2$, equal-sized samples of size n for $\hat{\pi}_g \in X$ and m for $\hat{\pi}_g \in Y$ must be drawn at each iteration (see Proposition 3.2) since the vectors $\hat{\pi}_g$ sizes n_g may differ across groups $g \in (1, \ldots, G)$. Moreover, given Eq.(12), a uniform random permutation (Knuth, 1981), denoted P, is applied to Y (or X equivalently) at each iteration e 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 key insight of our $\mathcal{E}$-based fairness approach is the importance of the intra-group terms $\mathbb{E}[\|X - X'\|]$ and $\mathbb{E}[\|Y - Y'\|]$ in Eq.(12). It ensures meaningful demographic parity in premium distributions. As shown in the last column of Figure 17 in Appendix, minimizing only the inter-group term $\mathbb{E}\left[\|X - Y\|\right]$ leads to highly concentrated premium distributions with unnaturally sharp and narrow peaks, indicating that the variability within each (gender) group collapses. This results in a poor model accuracy, as it does not respect the original structure of the premiums $\hat{\pi}$ distribution. In the middle column, including the additional intra-group terms $\mathbb{E}[\|X - X'\|]$ and $\mathbb{E}[\|Y - Y'\|]$ allows to preserve the natural spread and variability of each demographic group's premium distribution.

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 minimized. 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

Input $\hat{\pi} = \{\hat{\pi}_g\}_{g=1}^{G}$ (Initial premium predictions per unit of exposure produced by a base regressor (e.g., a LightGBM) partitioned into $G \geq 2$ sensitive groups)

Output: $\hat{\pi}^{\text{fair}}$ (Fairness-adjusted predictions)

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

for e *from* 1 *to* E **do**

// Sample predictions:

$X(e-1) \leftarrow \{\hat{\pi}_g^{\text{fair}}(e-1)\}_{g=1}^{\lceil G/2 \rceil} \in \mathbb{R}^{n \times d}$;

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

// Compute pairwise distances from Eq.(16):

$\text{dist}_{XY} \leftarrow \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \sqrt{\sum_{k=1}^{d} \left(x_{ik}(e-1) - y_{jk}(e-1)\right)^2}$;

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

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

// Compute fairness loss (energy distance):

$\mathcal{L}_{\text{fair}}(e) \leftarrow 2\,\text{dist}_{XY} - \text{dist}_{XX} - \text{dist}_{YY}$

// Compute gradient $\nabla \mathcal{L}_{\text{fair}}(e)$ with respect to $X(e-1)$ and $Y(e-1)$ from Eq.(19) via automatic differentiation

// Update predictions:

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

end

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

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 training set (used to train the fairness adjuster model), an individual fairness correction factor γ_i is computed:

$$\gamma_i = \frac{\hat{\pi}_{\text{fair},i}}{\hat{\pi}_i} \tag{23}$$

where $\hat{\pi}_i$ is the original predicted premium before fairness adjustment, and $\hat{\pi}_{\text{fair},i}$ is the fairness-adjusted premium obtained through the fairness optimization process. This correction factor γ_i quantifies the individual adjustment required for fairness.

We train a random forest regressor $f_\theta(X)$ to predict γ_i based on the policyholder's features X_i:

$$\hat{\gamma}_i = f_\theta\left(X_i\right) \tag{24}$$

This model learns how different features influence fairness adjustments. For a *new* policyholder with features X_{new}, their fairness correction factor is predicted as:

$$\hat{\gamma}_{new} = f_\theta(X_{\text{new}}) \tag{25}$$

The initial regression model (e.g., Light Gradient Boosting machine) predictions $\hat{\pi}$ are then adjusted using this factor:

$$\hat{\pi}_{\text{fair,new}} = \hat{\pi}_{\text{new}} \cdot \hat{\gamma}_{\text{new}} \tag{26}$$

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

Finally, note that the regression models used to predict the premiums best-estimates $\hat{\pi}(\boldsymbol{X}, \boldsymbol{D})$, particularly tree-based boosting, 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:

$$\mathbb{E}\left[\hat{\pi}(\boldsymbol{X}, \boldsymbol{D})\right] = \mathbb{E}\left[Y\right] \tag{27}$$

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

$$\mathbb{E}\left[Y|\hat{\pi}(\boldsymbol{X}, \boldsymbol{D}) = s\right] = s \tag{28}$$

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 $\mathbb{E}\left[Y|\hat{\pi}(\boldsymbol{X}, \boldsymbol{D})\right]$ is first estimated using local regression techniques on $\{(\hat{\pi}(\boldsymbol{x}_i), y_i)\}$. The predictions $\hat{\pi}$ are adjusted with a multiplicative factor $\lambda_\alpha(s)$ derived from the local regression estimates:

$$\lambda_\alpha(s) = \frac{\mathbb{E}\left[Y|\hat{\pi}(\boldsymbol{X}, \boldsymbol{D}) = s\right]}{s} \tag{29}$$

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

$$\hat{\pi}_{BC}(\boldsymbol{x}) = \lambda_\alpha\left(\hat{\pi}(\boldsymbol{x})\right) \cdot \hat{\pi}(\boldsymbol{x}) \tag{30}$$

As illustrated in Section 5.2, the balance correction is applied after the fairness adjustment, thereby ensuring that the final adjusted premiums accurately reflect the underlying risk across different segments.

5. Results

In this Section 5, we empirically evaluate the approach proposed in Section 4. First, we introduce the dataset and its key characteristics. In Section 5.1, we model the premium

best-estimates $\hat{\pi}(\boldsymbol{X}, \boldsymbol{D})$ using a tree-based boosting model (LGB). We then compare, in Section 5.2, the overall total balance calibration of the LGB model with that of a Generalized Linear Model (GLM) with a canonical link function. Finally, in Section 5.3, we adjust the premium best-estimates for fairness, considering both two and six sensitive demographic groups.

To assess our methodology empirically, we simulate a motor insurance portfolio $\mathcal{X}$ consisting of $n = 100,000$ individual policies. Table 1 provides a preview of the first five policies, where each column represents a rating factor. In this case study, the gender (binary) and age of the policyholders, two key risk factors in motor insurance, are considered as prohibited variables $\boldsymbol{D} = (D_1, D_2)$. For each policy, we also 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 most policies report zero claims (see Table 19 in Appendix), the count variable N is considered a rare event. In the remainder of the analysis, we focus on claims frequency per unit of exposure, the ratio $\lambda_i = N_i/\nu_i$. To gain further insight

	X_{income}	D_{age}	X_{power}	$X_{\text{education}}$	X_{location}	X_{type}	D_{gender}	X_{ethnicty}
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: Preview of five policies from the simulated insurance portfolio $\mathcal{X}$.

into the dataset, Figures 13 and 14 in Appendix illustrate the distributions of the rating factors, presented in Table 1, with respect to the variable of interest N. Figure 13 shows the average number of claims per policyholder across various categories (e.g., female, male, SUV, Sudan, ...), while Figure 14 displays the density distributions of the continuous age and income variables across $N \in \{0, 1, 2, 3\}$.

5.1 Best estimates

We consider a Generalized Linear Model (GLM) and a boosting algorithm (LGB) to model the premium best-estimates $\hat{\pi}$ for each policyholder. For the boosting model, we use the `lightgbm` and `optuna` Python packages to find the optimal set of hyper-parameters, detailed in Table 20. Both models are trained on 60,000 policies. Their performance is evaluated on a disjoint test set of 20,000 policies. A second disjoint test set, with the remaining 20,000 observations, is set aside to later assess the fairness adjustment procedure for new policyholders.

Figure 5.1 compares the performance of the two models. In this figure, the date are segmented into 10 quantile bins, each containing approximately the same number of observations. The bar chart represents the sum of exposures within in each bin, while

the line plot compares the estimated claim frequency $\hat{\lambda}$ (solid line) against the observed frequency λ (dashed line). A closer alignment between the two lines across bins indicates superior model performance. The GLM (left plot) systematically under-predicts or over-predicts frequencies in certain bins, whereas LGB (right plot) provides a closer match across all bins.

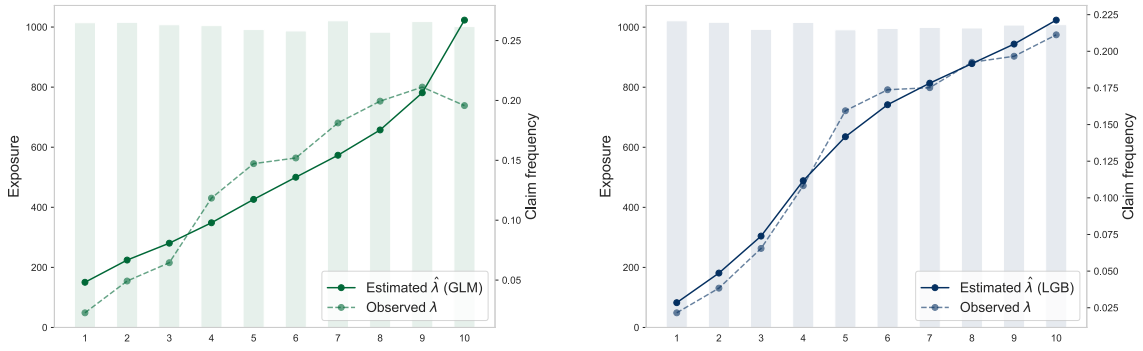


Figure 5.1: Comparison of estimated $\hat{\boldsymbol{\lambda}}$ (solid line) against observed $\boldsymbol{\lambda}$ (dashed line) across exposure bins. The data is divided into 10 equal-sized bins based on the predicted claim frequency per unit of exposure. Left plot: GLM. Right plot: LGB model.

The Poisson deviance is widely used to formally evaluate the accuracy of a regression model that estimates claim counts in insurance. It is defined as:

$$\mathcal{D}_{\text{Poisson}}\left(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}}\right) = 2\sum_{i=1}^{n} N_i \left(\frac{\nu_i}{N_i}\hat{\lambda}_i - \left(\ln\left(\frac{\nu_i}{N_i}\hat{\lambda}_i\right) + 1\right)\mathbb{I}_{\{N_i \geq 1\}}\right), \tag{31}$$

where N_i, ν_i, and $\hat{\lambda}_i$ are respectively the ith policyholder's observed number of claims, observed exposure, and estimated claim frequency (per unit of exposure). Table 2 summarizes the Poisson deviance for both models on the test set. The LGB has a slightly lower Poisson deviance than the GLM, suggesting that the LGB machine better captures the variation in the data. This can be attributed to the LGB's ability to model nonlinear relationships and interactions more effectively than GLM, which is restricted to linear relationships unless explicitly transformed.

Model	Poisson Deviance
GLM	6625.6248
LGB	6539.1550

Table 2: GLM and LGB Poisson deviance (test set).

Given the models predictions $\hat{\boldsymbol{\lambda}}$, the probability of observing exactly c claim(s) is given

by Poisson probability mass function (pmf):

$$f(c, \lambda_i) = \mathbb{P}(N_i = c) = \frac{(\lambda_i)^c e^{-\lambda_i}}{c!}, \tag{32}$$

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 across the (test) portfolio. Table 3 summarizes these values. While the LGB model has a lower deviance (see Table 2) and a closer fit in the quantile plots (see Figure 5.1), it overestimates by 13 the total number of policies with at least one claim $c > 0$ compared to the GLM. The boosting model's overestimation bias is the result of the use of the Poisson deviance as a loss function outside canonical GLMs (Denuit et al., 2021). Note that any gap in nonzero claim counts is counterbalanced by an opposite gap in zero claim counts, as required by the Poisson pmf in Eq.(32).

c	Observed	LGB	GLM	LGB gap	GLM gap
0	18719.0	18705.6169	18717.5133	-13.3831	-1.4867
1	1221.0	1224.7363	1213.9650	3.7363	-7.0350
2	56.0	66.5456	65.2688	10.5456	9.2688
3	4.0	2.9848	3.1122	-1.0152	-0.8878
> 0	1281.0	1294.3831	1282.4867	**13.3831**	**1.4867**

Table 3: Comparison of claims counts for LGB and GLM regression models.

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 4, we compare it with the GLM and LGB average premium values $\overline{\pi} = \frac{1}{n}\sum_{i=1}^{n} \hat{\lambda}_i$ (Denuit et al., 2021). Note that in the remainder of the analysis, we assume that the models predictions $\hat{\lambda}$ (per unit of exposure), are the so-called premium best-estimates $\hat{\pi}(\boldsymbol{X}, \boldsymbol{D})$, illustrated in Figure 5.2. While both models are close to the true premium mean $\overline{\pi} = 0.13502$ (see Table 4), the boosting machine does slightly overestimate the average premium.

Statistic	GLM_{β_0}	GLM	LGB
Mean $\overline{\pi}$	0.13502	0.135001	**0.136334**
10% Quantile		0.060189	0.039125
90% Quantile		0.228846	0.211547

Table 4: Summary statistics on $\hat{\pi}$.

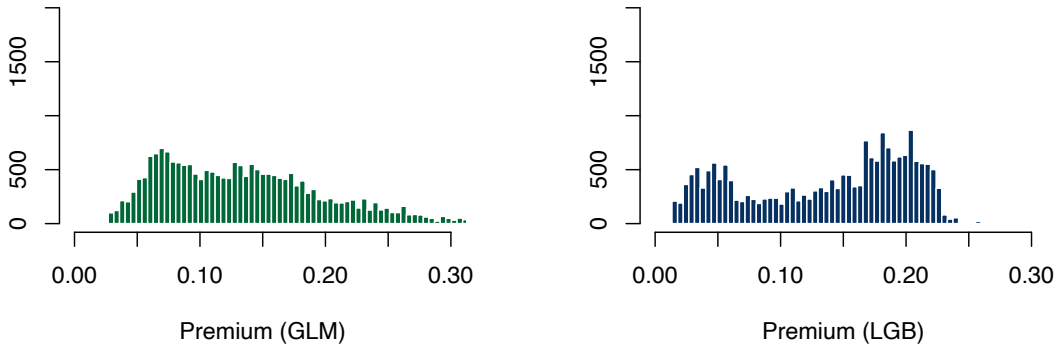


Figure 5.2: Histogram of $\hat{\pi}$ modeled by the GLM (left plot) and LGB (right plot).

5.2 Autocalibration

The negative bias (or overestimation) observed in the LGB 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). With a smoothing hyper-parameter of $\alpha = 0.25$, this correction helps reduce reduce the global bias, while preserving the Poisson deviance (barely increases from 6539.155 to 6542.645). Table 5 presents summary statistics before and after bias correction, showing that the mean premium prediction is much closer to the true value after adjustment. Table 6 details the improvement in the expected claim counts. The gap in the count of policies with at least one claim is nearly eliminated after correction. Figure 5.3 displays

Statistic	GLM_{β_0}	LGB	LGB_{BC}
Mean $\overline{\pi}$	0.13502	**0.136334**	**0.134647**
10% Quantile		0.039125	0.037364
90% Quantile		0.211547	0.200250

Table 5: Divergence from the true mean μ before and after bias correction (BC).

c	Observed	LGB	LGB_{BC}	LGB gap	LGB_{BC} gap
0	18719.0	18705.6169	18720.3746	-13.3831	1.3746
1	1221.0	1224.7363	1211.9909	3.7363	-9.0091
2	56.0	66.5456	64.7117	10.5456	8.7117
3	4.0	2.9848	2.8178	-1.0152	-1.1822
> 0	1281.0	1294.3831	1279.6254	**13.3831**	**-1.3746**

Table 6: Comparison of claims counts for LGB before and after bias correction (BC).

histograms of premium predictions before and after bias correction.

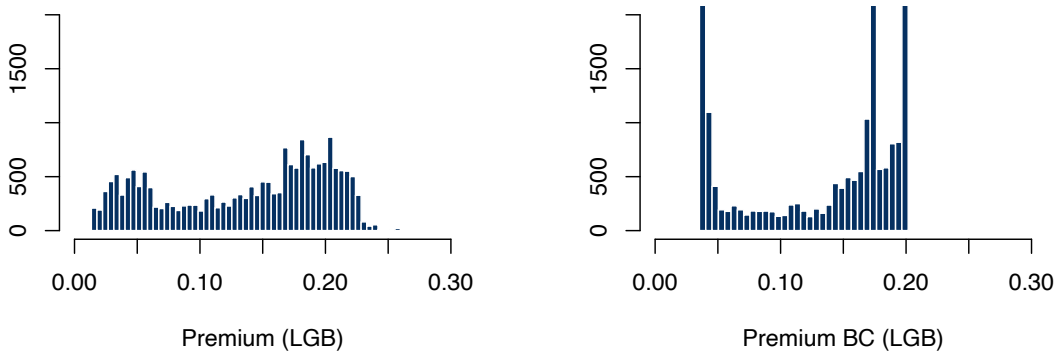


Figure 5.3: Histogram of $\hat{\pi}$ before (left) and after (right) balance correction.

Furthermore, the mapping $s \mapsto \mathbb{E}\left[Y|\hat{\pi}\left(\boldsymbol{X}\right)=s\right]$, in Figure 5.4, shows that after bias correction (right plot), the estimated premiums are much closer to the true mean $\mu\left(\boldsymbol{x}\right)=\mathbb{E}\left[Y|\boldsymbol{X}=x\right]$ (i.e., the first bisector). We also observe that $\mathbb{E}\left[Y|\hat{\pi}^{BC}\left(\boldsymbol{X}\right)=F_{\hat{\pi}}^{-1}\left(u\right)\right]/F_{\hat{\pi}}^{-1}\left(u\right)$ is now closer to 1 in Figure 15 in Appendix, indicating a significant reduction in local bias.

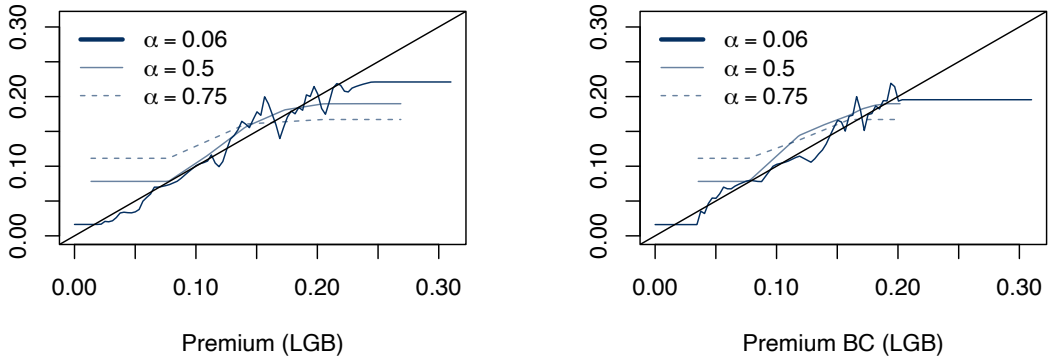


Figure 5.4: Visualization of $s \mapsto \mathbb{E}\left[Y|\hat{\pi}\left(\boldsymbol{X}\right)=s\right]$ before (left plot) and after (right plot) bias correction (BC). α is a smoothing hyper-parameter in the `locfit` package in R, and the diagonal corresponds to $\hat{\pi}=\mu$.

These results demonstrate that the bias correction procedure successfully realigns the LGB predictions with the true premium mean while maintaining model performance as measured by the Poisson deviance. This balance correction is an important component of the subsequent fairness adjustment steps.

5.3 Fairness adjustment

The best-estimates premiums $\hat{\pi}(\boldsymbol{X}, \boldsymbol{D})$ presented in Sections 5.1 and 5.2 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. To achieve this, we apply the gradient-descent optimization on the energy distance metric defined in Eq.(12). We first illustrate this mechanism for a binary sensitive attribute and then extend the approach to settings where $G > 2$.

5.3.1 G=2 demographic groups

In the binary setting $G = 2$, premiums $\hat{\pi}$ are split into two groups $\{\hat{\pi}_g\}_{g=1}^{G}$ with $g = \{\text{Female}, \text{Male}\}$. The premiums are iteratively adjusted (see algorithm 1) to equalize, not only the means, but also the dispersion and overall shape of their distributions.

Figure 5.5 shows the premium densities before and after fairness adjustment across the two demographic groups. The solid blue line represents the initial premiums $\hat{\pi}$. The dashed and dotted blue lines correspond to the initial premiums for male $\hat{\pi}_{\text{Male}}$ and female $\hat{\pi}_{\text{Female}}$ policyholders, respectively. The adjusted premiums $\hat{\pi}^{\text{fair}}$ are similarly depicted in red. The initial distributions clearly differ between males and females, with a higher concentration of low premiums for females and conversely for males, indicating gender-based pricing differences. As theorized, after adjustment the distributions for both groups (dashed and dotted red lines) nearly overlap. This indicates that our gradient-descent optimization on the energy distance has effectively reduced gender-based disparities.

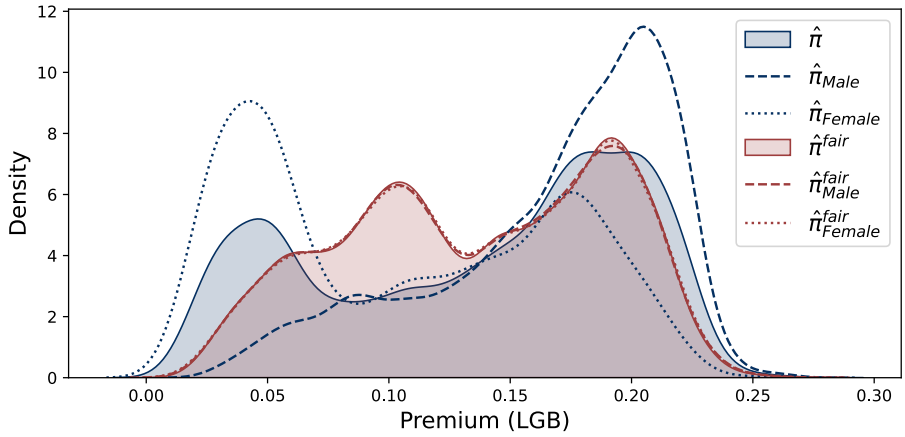


Figure 5.5: Premium densities before and after fairness adjustment across $g = \{\text{Female}, \text{Male}\}$.

While our fairness adjuster successfully enforced demographic parity, fairness should not come at the cost of the model predictive performance. As already discussed in Section 4 and illustrated in Figure 17 (in Appendix), the intra-group terms $\mathbb{E}[\|X - X'\|]$ and $\mathbb{E}[\|Y - Y'\|]$ in Eq.(12) ensure the natural spreads (intra-group variability) of the

premium distributions are preserved, and, therefore, prevent the collapse of the distributions into unrealistic concentrated peaks (which would lead to poor predictive accuracy). Figure 5.6, which compares the overall premium distribution before and after fairness adjustment, confirms the fairness adjustment results in a realistic redistribution of density in the adjusted distribution.

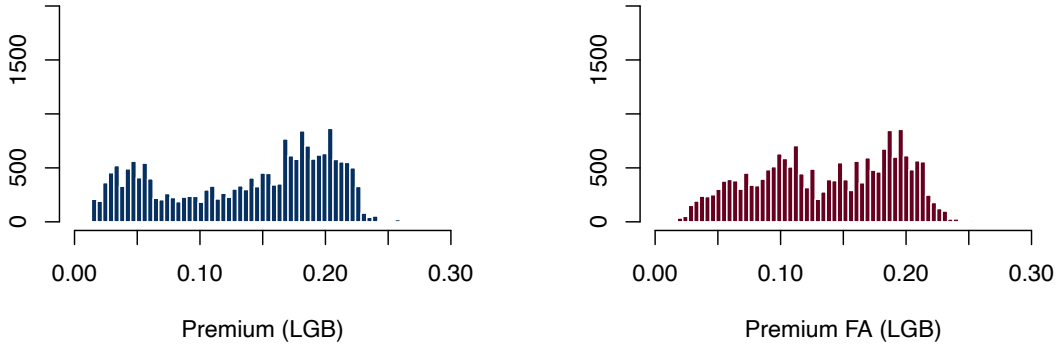


Figure 5.6: Premium distribution before and after fairness adjustment (FA) with respect to the Gender (binary) variable. Left plot: $\hat{\pi}$. Right plot: $\hat{\pi}^{\text{fair}}$

Figure 5.7 further shows the evolution of the fairness loss (i.e., the energy distance) and the deviance loss (a likelihood-based measure of predictive accuracy) over training epochs. The fairness loss, which begins at a relatively high value (reflecting significant disparities between groups), steadily decreases and approaches near zero by around 50 epochs. This decline confirms the effectiveness of the update rule in Eq.(21), which continuously nudges the group distributions closer together toward demographic parity. Conversely, the deviance loss, which reflects the model predictive performance, increases over epochs but remains acceptable (around 1.5% increase).

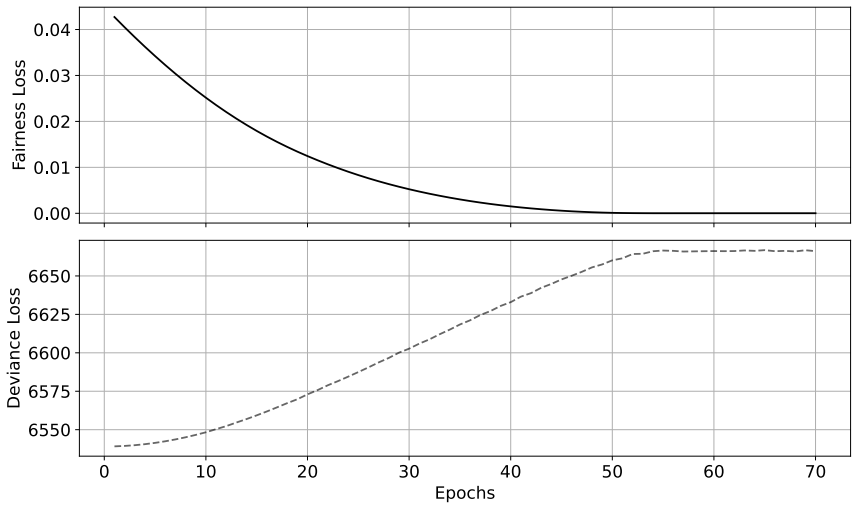


Figure 5.7: Evolution of fairness and deviance losses over epochs.

The analysis in Tables 8 and 7 reveals that, while the fairness adjustment effectively reduces gender disparities (see Figure 5.6), it also increases the overestimation of the

overall claims count (from 13.38 to 19.9) and creates a larger deviation from the true mean $\mu = 0.13502$.

Statistic	GLM_{β_0}	LGB	LGB_{fair}
Mean $\overline{\pi}$	0.13502	**0.136334**	**0.13657**
10% Quantile		0.039125	0.058666
90% Quantile		0.211547	0.204979

Table 7: Divergence from the true mean μ.

c	Observed	LGB	LGB_{fair}	LGB gap	LGB_{fair} gap
0	18719.0	18705.6169	18699.0905	-13.3831	-19.9095
1	1221.0	1224.7363	1234.2958	3.7363	13.2958
2	56.0	66.5456	63.7865	10.5456	7.7865
3	4.0	2.9848	2.7249	-1.0152	-1.2751
> 0	1281.0	1294.3831	1300.9095	**13.3831**	**19.9095**

Table 8: Comparison of claims counts before and after fairness adjustment.

In Figure 5.8, 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 7. A detailed view of the local bias is provided in Figure 16 (in Appendix).

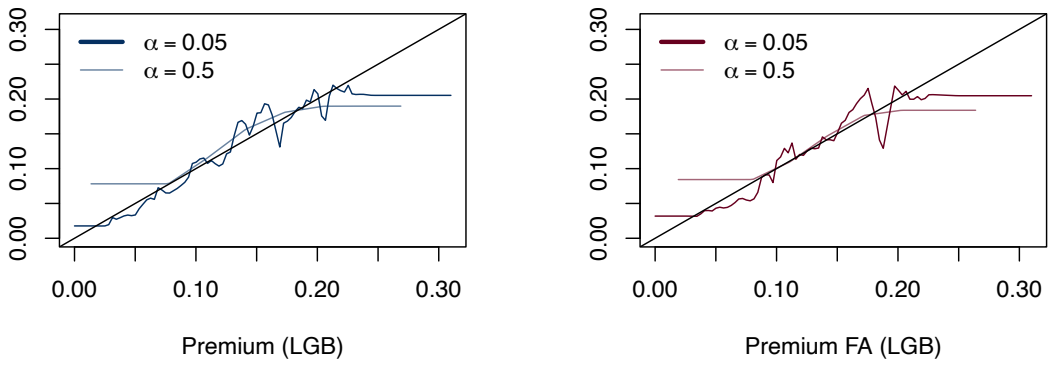


Figure 5.8: Visualization of $s \mapsto \mathbb{E}\left[Y|\hat{\pi}\left(\boldsymbol{X}\right) = s\right]$ before and after fairness adjustment across the Gender variable.

To mitigate the observed bias, we propose a secondary balance correction after the initial $\mathcal{E}$-based adjustment using the calibration method described at the end of Section 4 and illustrated in Section 5.2. For $\alpha = 0.16$, the bias correction step successfully reduces the initial overestimation. This improvement is detailed in Tables 9 and 10, with a noticeable decrease of the total claims counts from around twenty to five, and an average premium of 0.135058, that is much closer to the reference of 0.13502.

c	Observed	LGB_{fair}	$\text{LGB}_{\text{fair, BC}}$	LGB_{fair} gap	$\text{LGB}_{\text{fair, BC}}$ gap
0	18719.0	18699.0905	18713.4489	-19.9095	-5.5511
1	1221.0	1234.2958	1221.2914	13.2958	0.2914
2	56.0	63.7865	62.5397	7.7865	6.5397
3	4.0	2.7249	2.6244	-1.2751	-1.3756
> 0	1281.0	1300.9095	1286.5511	**19.9095**	**5.5511**

Table 9: Comparison of claims counts for $\hat{\pi}^{\text{fair}}$ before and after BC.

Statistic	GLM_{β_0}	LGB	LGB_{fair}	$\text{LGB}_{\text{fair, BC}}$
Mean $\overline{\pi}$	0.13502	**0.136334**	**0.13657**	**0.135058**
10% Quantile		0.039125	0.058666	0.046071
90% Quantile		0.211547	0.204979	0.203513

Table 10: Divergence from the true mean μ.

Figure 5.9 further shows the local redistribution of density after balance correction (right plot).

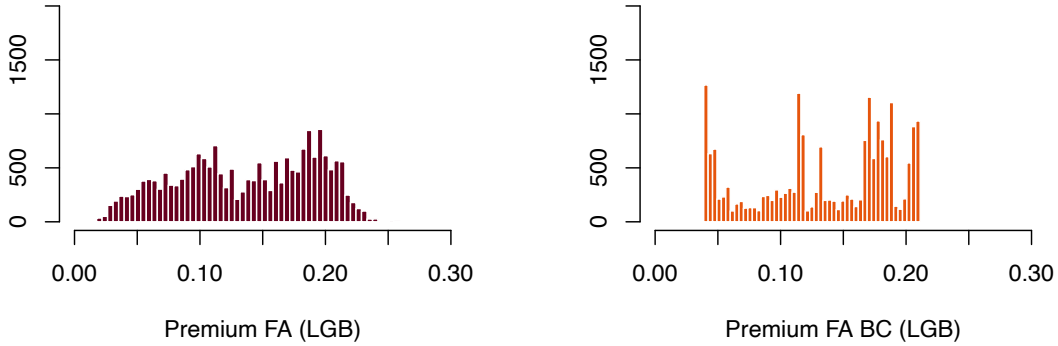


Figure 5.9: Fair premiums distributions $\hat{\pi}^{\text{fair}}$ before and after bias correction (BC).

Table 11 summarizes the Poisson deviance values. While fairness adjustments slightly increase the deviance (from 6539.155 to 6666.717), the loss of accuracy remains limited (1.5% increase), especially after balance correction (6654.98) which improves the deviance without affecting the fairness as showed in Figure 5.10.

Model	Poisson deviance
GLM	6625.6248
LGB	6539.155
LGB BC	6542.645
LGB fair	6666.717
LGB fair BC	6654.98

Table 11: Poisson deviance values for GLM and LGB models on the test set.

Figure 5.10 confirms that, even after balance correction, fairness is maintained as the conditional premium densities (in orange) for both gender groups remain well-aligned.

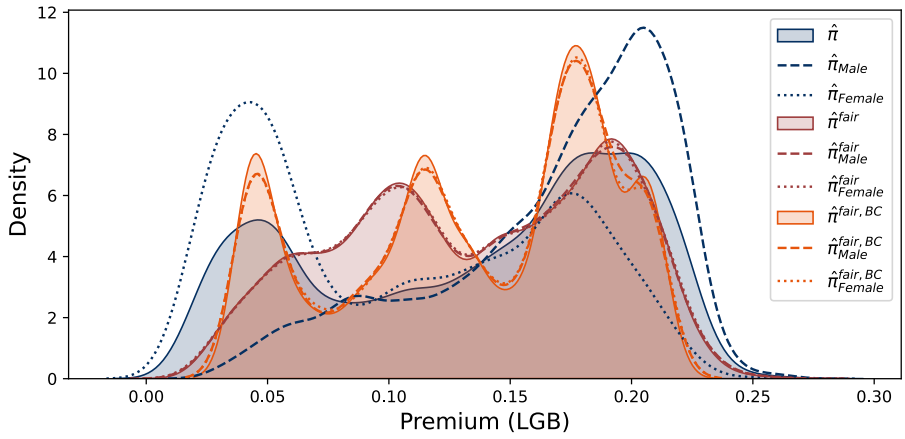


Figure 5.10: Premiums densities before (in blue) and after FA (in red) and FA & BC (in orange) across $g = \{\text{Female}, \text{Male}\}$.

Table 12 summarizes several keys fairness metrics discussed in the literature. These include the Wasserstein distance $W_p(X, Y)$ extensively studied by Charpentier (2023), Charpentier et al. (2023) and Hu et al. (2023, 2024), the MMD and Sinkhorn divergence $S_\epsilon(X, Y)$ proposed by Oneto et al. (2020) to impose demographic parity, as well as widely adopted metrics in the statistical and causal theory (see e.g., Charpentier (2023)) such as the Group Fairness in Expectation Δ_{GFE} and Statistical Parity Δ_{SP}.

Metric Formula with $X = \hat{\pi}_{\text{Female}}$ and $Y = \hat{\pi}_{\text{Male}}$	$\hat{\pi}$	$\hat{\pi}^{\text{fair, BC}}$
$\mathcal{E}(X,Y) = 2\mathbb{E}\|X-Y\| - \mathbb{E}\|X-X'\| - \mathbb{E}\|Y-Y'\|$	0.042704	8e-06
$W_p(X,Y) = \left(\inf_{\gamma\in\Gamma(X,Y)} \mathbb{E}_{(X,Y)\sim\gamma}\|X-Y\|^p\right)^{1/p}$	0.006167	0.001024
$\text{MMD}^2(X,Y) = \mathbb{E}[k(X,X')] + \mathbb{E}[k(Y,Y')] - 2\mathbb{E}[k(X,Y)]$	0.205325	2e-06
$S_\epsilon(X,Y) = W_\epsilon(X,Y) - \frac{1}{2}W_\epsilon(X,X) - \frac{1}{2}W_\epsilon(Y,Y)$	0.028611	2.2e-05
$\Delta_{GFE} = \lvert\mathbb{E}[\hat{\pi}\lvert D_{\text{Gender}} = \text{F}] - \mathbb{E}[\hat{\pi}\lvert D_{\text{Gender}} = \text{M}]\rvert$	0.062959	0.000155
$\Delta_{SP} = \sup_s \lvert P(\hat{\pi} \le s\lvert D_{\text{Gender}} = \text{F}) - P(\hat{\pi} \le s\lvert D_{\text{Gender}} = \text{M})\rvert$	0.396203	0.030354

Table 12: Comparison of key fairness metrics before and after $\mathcal{E}$-based fairness adjustment and balance correction.

For new policyholders, we extend this methodology by applying an individual correction factor γ_i, as outlined in Section 4. A random forest regressor learns the mapping from policyholder's characteristics to correction factors, generalizing the fairness adjustment mechanism beyond the dataset used to train the adjuster.

Figure 5.11 shows that the adjusted predictions for new policyholders remain consistent across sensitive groups with the results observed in Figure 5.10, with some minor deviations.

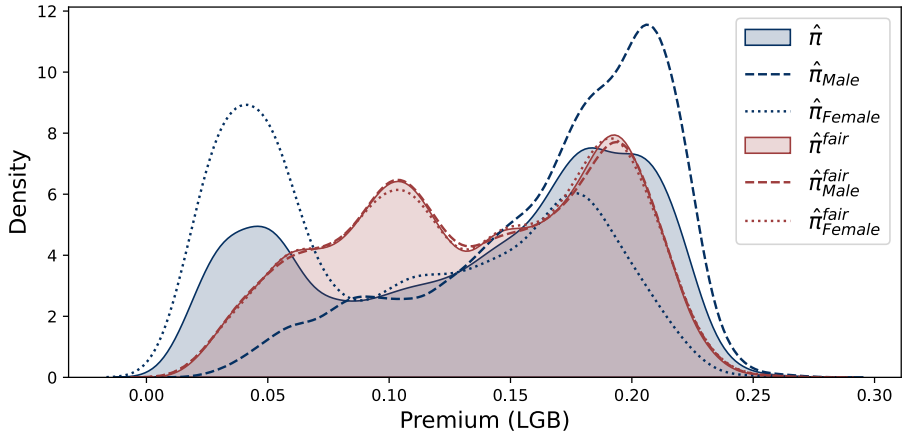


Figure 5.11: LGB premiums densities, before and after FA for new policies.

Table 13 summarizes the Poisson deviance values before and after fairness correction on the new test set with 20 000 policies. Once again, we observe a limited decrease in the model predictive performance.

Model	Poisson deviance
GLM	6880.0621
LGB	6809.2512
LGB fair	6986.1686

Table 13: Poisson deviance values on the test set.

5.3.2 G>2 demographic groups

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 $\{\hat{\pi}_g\}_{g=1}^{G}$ of sizes n_g organized in d-dimensional variables X and Y as in Eq. (22). In this setting, equal-sized samples are drawn from each group in X, and for each group in Y. A uniform random permutation is also applied to ensure the pairwise comparisons span all group combinations (see algorithm 1).

Figure 5.12 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).

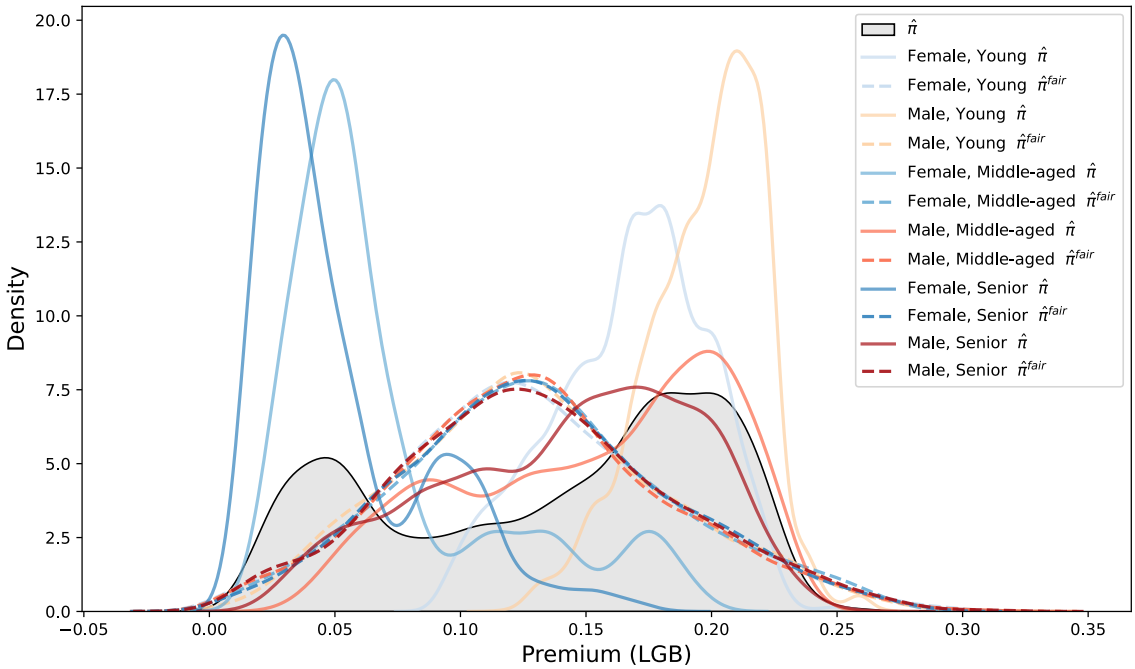


Figure 5.12: Premium densities across six demographic groups before (solid lines) and after fairness adjustment (dotted lines).

In Table 15, we observe that the fairness adjustment lead to a significant underestimation (around 42) of the overall number of claims with a decrease of the mean value in Table 14 from 0.136334 to 0.129824.

Statistic	GLM_{β_0}	LGB	LGB_{fair}
Mean $\overline{\pi}$	0.13502	**0.136334**	**0.129824**
10% Quantile		0.039125	0.061754
90% Quantile		0.211547	0.203506

Table 14: Divergence from the true mean μ before and after fairness adjustment.

c	Observed	LGB	LGB_{fair}	LGB gap	LGB_{fair} 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 15: Comparison of claims counts before and after fairness adjustment.

For $\alpha = 0.02$, we manage to significantly improve the underestimation after balance correction. In Table 16 we observe a smaller divergence from the true mean $\mu = 0.13502$, from 0.1298 to 0.1341, and in Table 17 we observe a decrease in the overall overestimation from -42 to 6.94 claims.

Statistic	GLM_{β_0}	LGB	LGB_{fair}	$\text{LGB}_{\text{fair, BC}}$
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 16: Divergence from the true mean μ before and after balance correction.

c	Observed	LGB_{fair}	$\text{LGB}_{\text{fair, BC}}$	LGB_{fair} gap	$\text{LGB}_{\text{fair, BC}}$ 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 17: Comparison of claims counts before and after balance correction.

Table 18 summarizes the deviance loss before and after fairness adjustment (for $G = 2$ and $G = 6$) and balance correction. The deviance deteriorates much more when adjusting across six demographic groups (compared to two). It was to be expected given that we used as demographic groups, both the gender and the age of the policyholders, which are known to be key risk factors in car insurance pricing. However, the overall loss in the model's accuracy after balance correction remains acceptable (6891.32 compared to the initial GLM deviance of 6625.6248).

Model	Poisson Deviance
GLM	6625.6248
LGB	6539.1550
LGB BC	6542.6450
LGB fair (gender)	6666.7170
LGB fair BC (gender)	6654.9800
LGB fair (age & gender)	7341.9258
LGB fair BC (age & gender)	6891.3200

Table 18: Poisson deviance values for GLM and LGB models (see Table 11).

6. Conclusion

In this article, we introduce a novel group fairness adjuster based on the multivariate energy distance to enforce demographic parity and address group-level disparities in insurance premium estimation. By leveraging the multivariate properties of the energy distance, our approach can simultaneously adjust premiums across more than two sensitive

demographic groups. The adjustment mechanism relies on a gradient-descent optimization of the energy distance that minimizes inter-group disparities while preserving the intra-group variability. Any biases in the total number of claims after fairness adjustment are mitigated by a post-adjustment balance correction (BC) step. It improves the final model calibration by aligning the adjusted premiums more closely with the empirical mean of the observed claims distribution.

We assess our method in the context of car insurance pricing, where demographic factors such as age and gender have a great influence in the risk assessment and premium determination. Results show that significant reductions in group-level disparities are achievable without overly compromising the model's accuracy. Furthermore, the flexibility of our approach allows for the adjustment of discriminatory premiums for new policyholders without the need for retraining.

This research contributes to the growing body of work addressing fairness in the insurance industry by highlighting the importance of group fairness criteria like demographic parity, in alignment with the risk group-based nature of insurance pricing.

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.

References

Avraham, R. (2017). Discrimination and insurance. *The Routledge handbook of the ethics of discrimination*, 335–347.

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. *arXiv preprint arXiv:1705.10743*.

Birnbaum, B. (2020). Insurance consumer protection issues resulting from, or heightened bycovid-19. *Center for Economic Justice Report*.

Charpentier, A. (2023). Insurance, biases, discrimination and fairness.

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.

Côté, O., Côté, M.-P., & Charpentier, A. (2024). A fair price to pay: Exploiting causal graphs for fairness in insurance. *Journal of Risk and Insurance*.

Cramér, H. (1928). On the composition of elementary errors: First paper: Mathematical deductions. *Scandinavian Actuarial Journal*, 1928(1), 13–74.

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

Denuit, M., Hainaut, D., Trufin, J., Denuit, M., Hainaut, D., & Trufin, J. (2019). Insurance risk classification. *Effective Statistical Learning Methods for Actuaries I: GLMs and Extensions*, 3–26.

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

Gandy, O. H. (2016). *Coming to terms with chance: Engaging rational discrimination and cumulative disadvantage*. Routledge.

Hu, F., Ratz, P., & Charpentier, A. (2023). Fairness in multi-task learning via wasserstein barycenters. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 295–312.

Hu, F., Ratz, P., & Charpentier, A. (2024). A sequentially fair mechanism for multiple sensitive attributes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(11), 12502–12510.

Kantorovich, L., & Rubinstein, G. S. (1958). On a space of totally additive functions. *Vestnik Leningrad. Univ*, 13, 52–59.

Knuth, D. (1981). Seminumerical algorithms. *The art of computer programming*, 2.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1), 79–86.

Lindholm, M., Richman, R., Tsanakas, A., & Wuthrich, M. V. (2024). What is fair? proxy discrimination vs. demographic disparities in insurance pricing. *Proxy Discrimination vs. Demographic Disparities in Insurance Pricing (February 1, 2024)*.

Lindholm, M., Richman, R., Tsanakas, A., & Wüthrich, M. V. (2022). A discussion of discrimination and fairness in insurance pricing. *arXiv preprint arXiv:2209.00858*.

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(1), 405–431.

Peyré, G., Cuturi, M., et al. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6), 355–607.

Royden, H. L. (1988). *Real analysis*. Krishna Prakashan Media.

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.

Seligman, D. (1983). Insurance and the price of sex.

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.

Vandenhole, W. (2005). *Non-discrimination and equality in the view of the un human rights treaty bodies*. Intersentia nv.

Villani, C., et al. (2008). *Optimal transport: Old and new* (Vol. 338). Springer.

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

Appendix A

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

Table 19: Claims count and percentage.

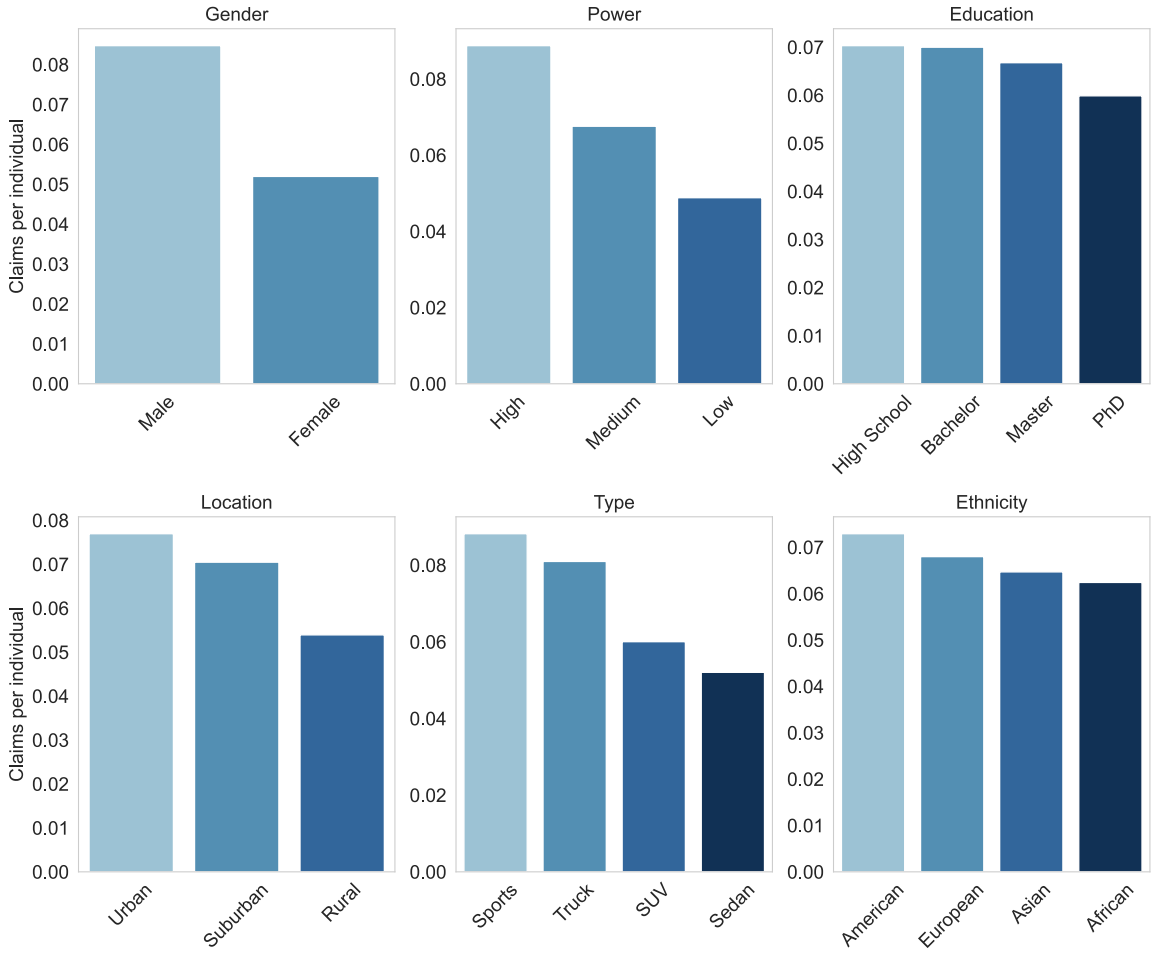


Figure 13: Each subplot represents a specific rating factor, with bars showing the average number of claims per individual within each category. Categories are sorted in descending order of claim frequency.

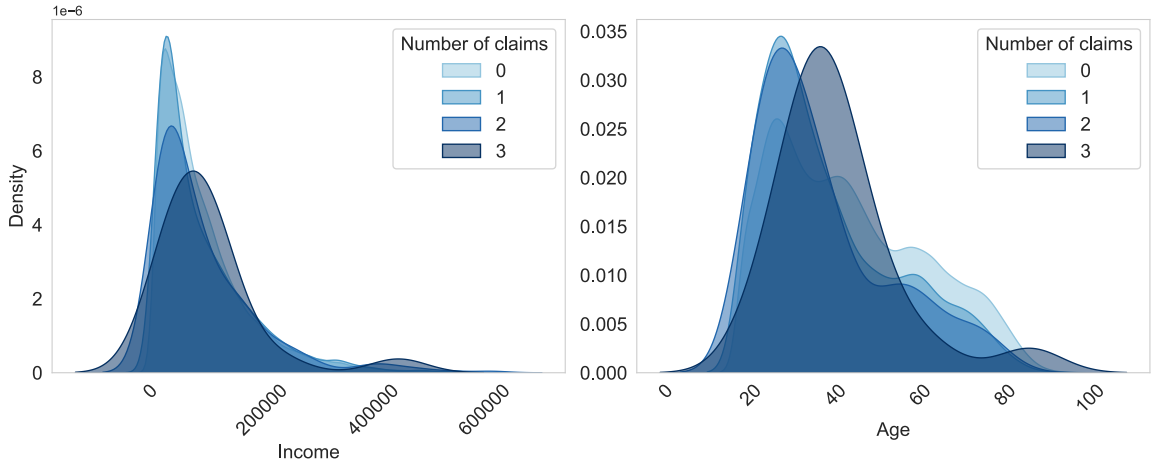


Figure 14: Density distributions of continuous variables across different number of claims.

Parameter	Value
learning_rate	0.0099
feature_fraction	0.9368
bagging_fraction	0.6151
bagging_freq	4
lambda_l2	2.5348e-05
min_sum_hessian_in_leaf	0.7570
max_depth	6
min_gain_to_split	0.4139
min_data_in_leaf	74
num_leaves	51
objective	poisson

Table 20: Best LightGBM hyperparameters values, optimized using `optuna`.

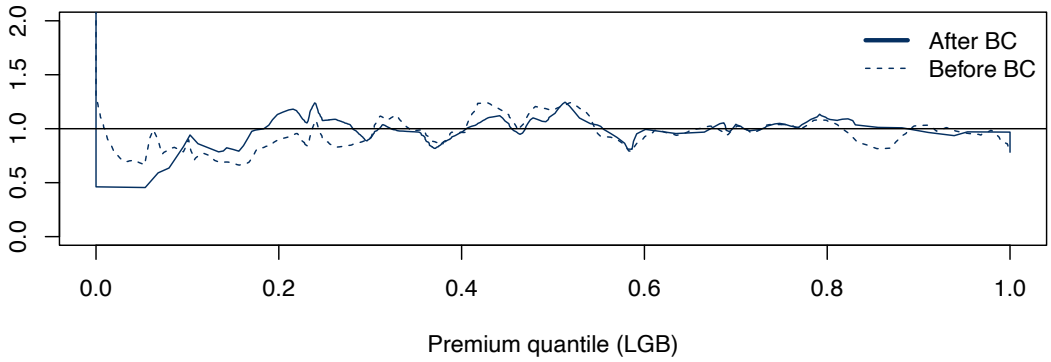


Figure 15: Local bias, i.e., $\mathbb{E}\left[Y|\pi\left(\boldsymbol{X}\right)=F_{\pi}^{-1}\left(u\right)\right]/F_{\pi}^{-1}\left(u\right)$ should be close to 1. Dotted line: before bias correction. Solid line: after bias correction.

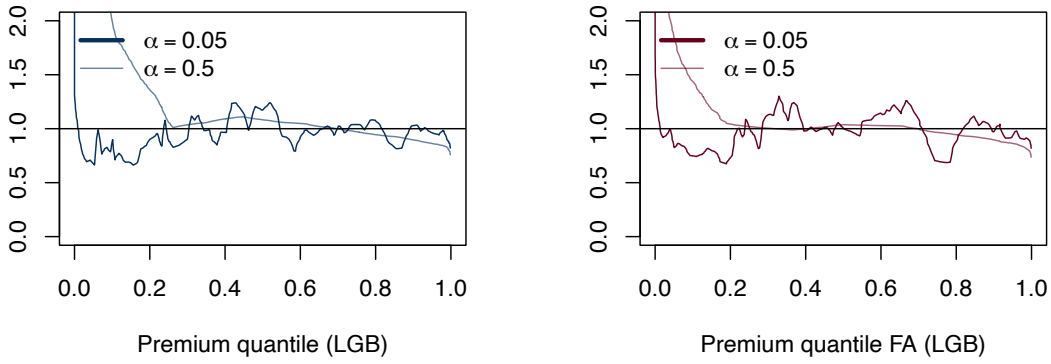


Figure 16: Visualization of the local bias $\mathbb{E}\left[Y|\hat{\pi}\left(\boldsymbol{X}\right) = F_{\hat{\pi}}^{-1}\left(u\right)\right]/F_{\hat{\pi}}^{-1}\left(u\right)$ before and after fairness adjustment (FA).

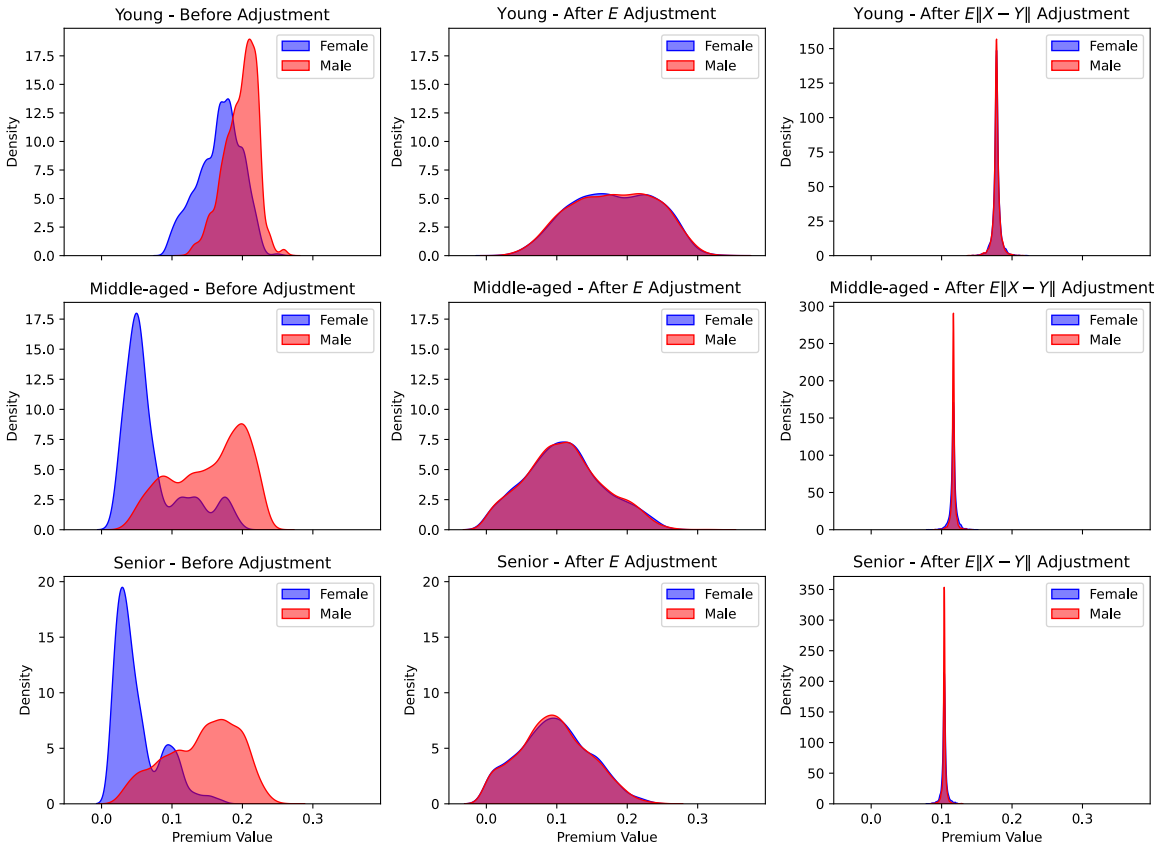


Figure 17: Visualization of the intra-group variability role in the fairness adjuster.

Appendix B

In this appendix, we present detailed proofs (Bellemare et al., 2017) for the unbiased sample gradients properties of the l_2^2 and Energy distances as well as the equivalence of these distances in the univariate case.

B.1. Proof of unbiased sample gradients for the l_2^2 distance

As demonstrated by Bellemare et al. (2017, p.15), the Cramér l_2^2 distance is the only l_p distance that possesses the unbiased sample gradients property. This is a key property in machine learning when using measures as loss functions for model training.

The gradient of the l_2^2 distance is defined as the derivative with respect to θ of the integral:

$$\nabla_\theta \, l_2^2\big(P, Q_\theta\big) = \nabla_\theta \int_{-\infty}^{+\infty} \Big(F_{Q_\theta}(x) - F_P(x)\Big)^2 \, dx.$$

Under the hypothesis that the convergence of the square follows from the convergence of the ordinary values, we swap the gradient and the integral:

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

Applying the chain rule of differentiation yields:

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

We use Lemma 1 from Bellemare et al. (2017, p. 14) to substitute the true CDF $F_P(x)$ with the expected empirical CDF $\mathbb{E}_{\boldsymbol{X}_m \sim P}[F_{\hat{P}_m}(x)]$:

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

By linearity of the expectation:

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

Finally, Fubini's theorem (Royden, 1988) allows to interchange the order of integration and expectation:

$$\begin{aligned} &= \mathbb{E}_{\boldsymbol{X}_m \sim P} \int_{-\infty}^{+\infty} 2\Big(F_{Q_\theta}(x) - F_{\hat{P}_m}(x)\Big) \, \nabla_\theta F_{Q_\theta}(x) \, dx, \\ &= \mathbb{E}_{\boldsymbol{X}_m \sim P}\Big[\nabla_\theta \, l_2^2\big(\hat{P}_m, Q_\theta\big)\Big]. \end{aligned}$$

B.2. Proof of unbiased sample gradients for the energy distance

Following Bellemare et al. (2017, p. 16), the Energy distance is defined as

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

where $X \sim P$, $Y \sim Q_\theta$, and $\hat{X}_m \sim \hat{P}_m$.

The gradient ∇ of $\mathcal{E}\left(P, Q_\theta\right)$ with respect to θ is:

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

Given that $\mathbb{E}\|X - X'\|$ does not depend on θ, we have:

$$= 2\,\nabla_\theta \,\mathbb{E}\|X - Y\| - \nabla_\theta \,\mathbb{E}\|Y - Y'\|.$$

Similarly, for the samples $\hat{X}_m \sim \hat{P}_m$, we have:

$$\begin{aligned}\nabla_\theta \,\mathcal{E}\left(\hat{P}_m, Q_\theta\right) &= \nabla_\theta \Big(2\mathbb{E}\|\hat{X}_m - Y\| - \mathbb{E}\|\hat{X}_m - \hat{X}'_m\| - \mathbb{E}\|Y - Y'\|\Big), \\ &= 2\,\nabla_\theta \,\mathbb{E}\|\hat{X}_m - Y\| - \nabla_\theta \,\mathbb{E}\|Y - Y'\|.\end{aligned}$$

All together, we have:

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

Under the assumption that ∇_θ and the expectation over $\boldsymbol{X}_m$ commute, we have:

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

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

$$= \nabla_\theta \mathbb{E}_{\boldsymbol{X}_m \sim P}\, \mathbb{E}\, \mathbb{E}_{x \sim \hat{P}_m} \|x - Y\|.$$

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

$$\begin{aligned}&= \nabla_\theta \mathbb{E}_{x \sim P}\, \mathbb{E}\|x - Y\| \\ &= \nabla_\theta \,\mathbb{E}\|X - Y\|.\end{aligned}$$

B.3. Proof of equivalence in one dimension with the Energy distance

In one dimension, Székely (2003) demonstrated that the Energy distance and the l_2^2 distance are linearly related by

$$l_2^2(P, Q) = \frac{1}{2}\,\mathcal{E}(P, Q).$$

For univariate distributions P and Q in $\mathbb{R}$, the Euclidean norm $\|\cdot\|$ simplifies to the absolute value $\|x - y\| = \sqrt{(x - y)^2} = |x - y|$ for scalars $x, y \in \mathbb{R}$. Therefore, the Energy

distance is written as:

$$\mathcal{E}(P, Q) = 2\mathbb{E}[|X - Y|] - \mathbb{E}[|X - X'|] - \mathbb{E}[|Y - Y'|].$$

Given that $X \sim P$ and $Y \sim Q$ are independent, their joint distribution is the product of their marginals $P(x)$ and $Q(y)$, and the expectation of a function of two random variables $f(X, Y)$ can be written as a double integral over their joint probability distribution:

$$\mathbb{E}[|X - Y|] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |x - y| \, dP(x) dQ(y).$$

Similarly, for $X \sim P$ and $X' \sim P$ (iid), we have:

$$\mathbb{E}[|X - X'|] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |x - x'| \, dP(x) \, dP(x').$$

For $Y \sim Q$ and $Y' \sim Q$ (iid), we have:

$$\mathbb{E}[|Y - Y'|] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |y - y'| \, dQ(y) \, dQ(y').$$

In one dimension, the absolute difference $|x-y|$ can be expressed in terms of the difference between the CDFs $P(x)$ and $Q(x)$:

$$|x - y| = \int_{x}^{y} \operatorname{sgn}(x - y) \, dx,$$

where $\operatorname{sgn}(x-y)$ is the sign function. By integrating over all pairs (x, y), the expected absolute difference can be expressed in terms of the squared difference of the CDFs (Székely, 2003):

$$\mathbb{E}[|X - Y|] = \int_{-\infty}^{\infty} (F_P(x) - F_Q(x))^2 \, dx.$$

This result follows from the decomposition of the probability measures using their CDFs and observing how discrepancies accumulate across the domain of the random variables.

Similarly, the terms $\mathbb{E}[|X - X'|]$ and $\mathbb{E}[|Y - Y'|]$ also involve the squared differences. Substituting the above into the Energy distance formula, we have:

$$\begin{aligned}\mathcal{E}(P, Q) = 2 \int_{-\infty}^{\infty} (F_P(x) - F_Q(x))^2 \, dx \\ - \int_{-\infty}^{\infty} (F_P(x) - F_P(x))^2 \, dx - \int_{-\infty}^{\infty} (F_Q(x) - F_Q(x))^2 \, dx.\end{aligned}$$

Given that the last two terms distributions are identical within their respective groups in the one-dimensional case, it simplifies to:

$$\mathcal{E}(P, Q) = 2 \int_{-\infty}^{\infty} (F_P(x) - F_Q(x))^2 \, dx = 2l_2^2(P, Q).$$