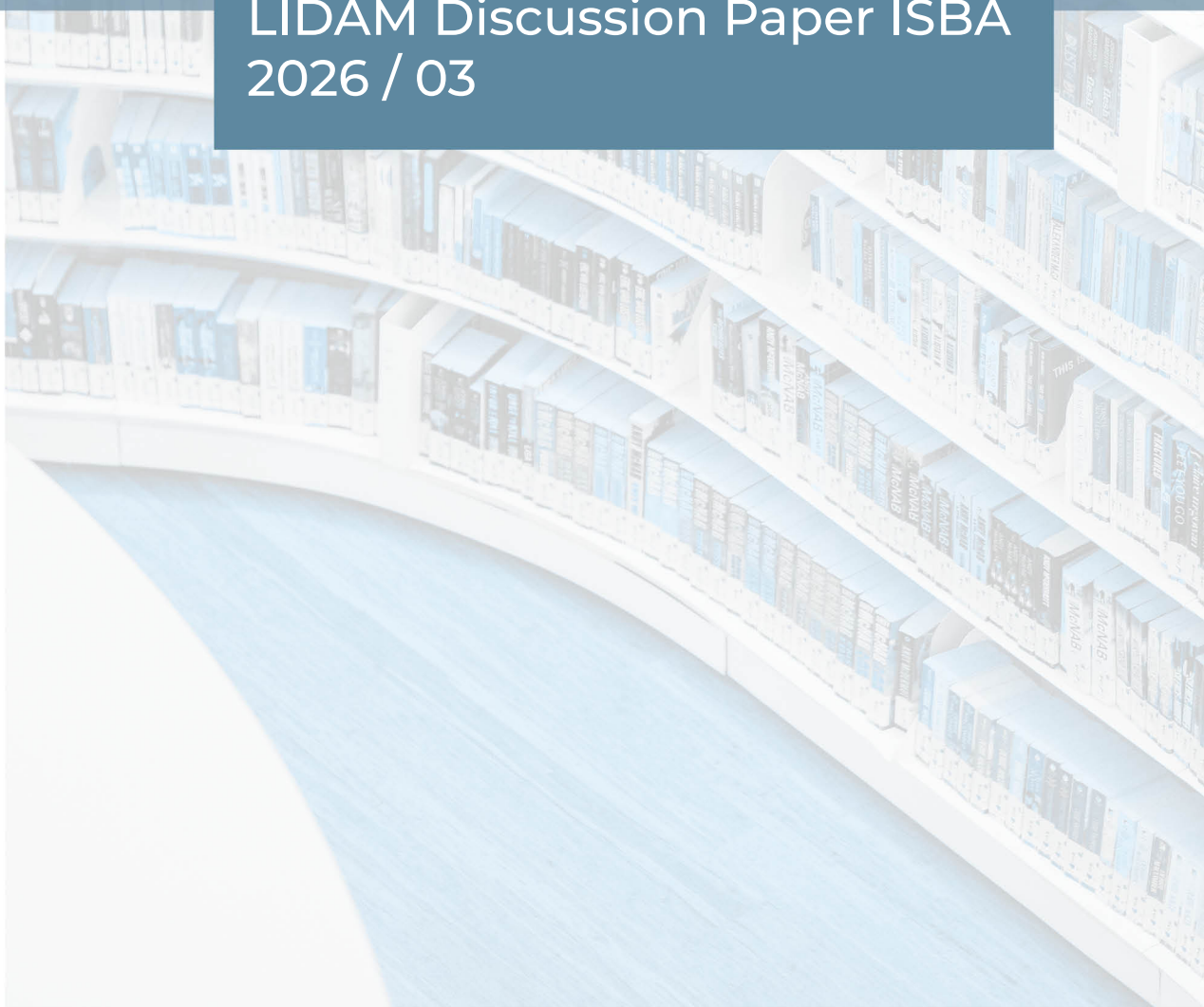
A MULTI-CRITERIA FAIR GAUSSIAN REGRESSOR FOR INSURANCE PREMIUM

Charlotte Jamotton, Donatien Hainaut

LIDAM Discussion Paper ISBA
2026 / 03



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 Multi-Criteria Fair Gaussian Regressor for Insurance Premium

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

[1]Institute of Statistics, Biostatistics and Actuarial Sciences (ISBA), Université catholique de Louvain (UCLouvain), Louvain-la-Neuve, 1348, Belgium

Abstract

This article studies how multiple notions of fairness can be incorporated into a single Bayesian non-parametric regression framework for insurance pricing, with a focus on claim frequency modeling under a log-link. We consider a Generalized Gaussian Process Regression (GGPR) model for count data with risk exposure and introduce fairness interventions in its architecture. Specifically, we address notions of individual fairness by altering the kernel structure to control the similarity between policies (e.g., to mitigate omitted variable bias). We also address group-level fairness by enforcing demographic parity through linear constraints affecting the posterior. This modified GGPR architecture allows us to jointly enforce multiple fairness definitions, spanning both group and individual-level criteria, within a single probabilistic model. We empirically explore trade-offs with actuarial fairness, and how different fairness criteria interact when combined. The results highlight the importance of adopting a multi-criteria, context-aware approach to fairness in insurance pricing.

Keywords: gaussian process regression, count data, confounding bias, actuarial fairness, omitted variable bias, indirect discrimination, direct discrimination

*Corresponding author: charlotte.jamotton@uclouvain.be

1. Introduction

Insurance pricing is fundamentally concerned with the accurate assessment of risk. Modern actuarial practice increasingly relies on big data and flexible machine learning methods (e.g., gradient boosting machine) to capture complex nonlinearities and interactions in claim data (Barry & Charpentier, 2020). While these models often improve predictive accuracy (Yang et al., 2017), they have also intensified concerns regarding discrimination, and compliance with evolving regulatory frameworks. In response, a rapidly growing actuarial literature (Fahrenwaldt et al., 2024) has emerged to question whether algorithmic pricing models amplify, perpetuate, or mitigate existing societal inequalities.

Fairness in insurance pricing remains a deeply contested concept. Competing definitions and criteria, such as individual versus group fairness, causal versus statistical notions of discrimination, proxy versus direct discrimination, are often mutually incompatible. As noted by O. Côté et al. (2024, p. 2), the insurance literature struggles with a lack of conceptual consensus, with fairness terms frequently used inconsistently across methodological, legal, and ethical perspectives (Dolman et al., 2021). Different fairness criteria and methodologies are defended as the correct solution, yet there is still no consensus on a single approach that aligns with both regulatory constraints and actuarial soundness.

This article adopts a deliberately neutral stance. Rather than advocating for a single definition of fairness, we propose a framework that embeds multiple fairness paradigms into the mathematical structure of a single regression model. Our objective is not to resolve the fairness debate normatively, but to provide a coherent experimental and analytical setting in which the practical implications and trade-offs of competing fairness notions can be studied side by side.

We focus in particular on the tension between individual-level and group-level fairness interventions. On the one hand, Lindholm et al. (2024c) and related work advocate for so-called discrimination-free prices derived via inverse probability weighting and sensitivity-based corrections, arguing that group fairness constraints such as demographic parity may induce inefficiencies or distort incentives. On the other hand, a series of papers by M.-P. Côté et al. (2024) and O. Côté et al. (2024) and Hu et al. (2023, 2024) emphasize the market-level consequences of individual corrections, showing that portfolio-specific fairness does not necessarily extend to market-wide equity and may exacerbate adverse selection. These contributions highlight fundamental disagreements not only about what constitutes fairness, but also about at which level (individual, portfolio, or market) it should be enforced.

To structure this landscape, we adopt the typology introduced by O. Côté et al. (2024), which classifies fairness interventions into five families: best-estimate, unaware, aware, hyperaware, and corrective. Additional insights are drawn from Xin and Huang (2024), who systematically contrast individual and group fairness criteria under different regulatory regimes, and from sensitivity-based analyses (Lindholm et al., 2024b) which

further sharpen the disagreement regarding proxy discrimination and causal relevance. Together, these works reveal a fragmented but important debate that strongly motivates the need for a unified modeling framework capable of accommodating multiple fairness objectives simultaneously.

To study these competing fairness notions, we require a predictive model that can capture non-linear dependencies of the response (such as claim amounts, claim counts, or claim frequencies) on multiple rating factors and that can also embed multiple fairness interventions directly into its architecture.

Traditional regression models in insurance often rely on parametric assumptions, where the actuary specifies a fixed functional form of the mapping relating the covariates to the target variable. For instance, Generalized Linear Models (GLMs) are widely used in pricing motor and health insurance, where claim frequency is modeled using a Poisson distribution with a log-link function (Denuit et al., 2019c). Similarly, Generalized Additive Models (GAMs) extend GLMs by allowing smooth, non-linear effects of predictors, such as driver age or vehicle power, using spline-based components (Denuit et al., 2019b). While these approaches are interpretable and statistically efficient, they require substantial prior knowledge about the data-generating process and involve model selection procedures to ensure the most appropriate model is retained.

When the true relationship between features and the outcome is complex or unknown, non-parametric regression offers a more flexible alternative. Methods such as splines or kernel regression can capture arbitrary non-linear patterns by defining local smoothing over the covariate space. In insurance, kernel methods have been used for tasks like claim cost reserving (Hiabu, 2016; Shim & Hwang, 2011) and mortality modeling (Risk & Ludkovski, 2024). However, such methods depend on design choices like knot placement or bandwidth selection, which are often heuristic and sensitive to input dimensionality. More recently, regression trees, random forests, and gradient boosting machines (e.g., XGBoost, LightGBM) have gained popularity in actuarial modeling due to their ability to model complex interactions without explicit specification of the functional form (Chevalier & Côté, 2025; Denuit et al., 2020). These ensemble methods have proven effective in various insurance tasks, including pricing (Henckaerts et al., 2020), mortality forecasting (Bjerre, 2022), and fraud detection (Taha & Malebary, 2020).

A third alternative is to adopt a Bayesian non-parametric strategy, and directly model the unknown underlying function. Bayesian non-parametric methods like Gaussian processes (GPs) place a prior distribution directly over the space of functions, allowing the data to inform both the shape and complexity of the model. In other words, rather than assuming a fixed functional form, we formulate the uncertainty about the mapping into a prior over mappings and then use Bayes rule to reach the posterior.

As a distribution over functions, a GP does not require an explicit parametric form and yields a posterior that reflects both prior beliefs and empirical evidence. This makes it particularly suitable for capturing complex, non-linear relationships in insurance data

where model uncertainty is a central concern. In insurance applications, GPs have been used in mortality modeling (Ludkovski et al., 2018), claim reserving (Ang & Lee, 2022), and option pricing (Hainaut, 2026; Hainaut & Vrins, 2024).

We contribute to the actuarial literature as follow. First, we extend the applicability of GPR to realistic insurance settings by considering claim count models with exposure, relying on Generalized Gaussian Process Regression (GGPR) for non-Gaussian data. In particular, we build on existing work on GGPR for Poisson and Tweedie-type likelihoods with offset terms, which has been shown to provide a flexible alternative to GLMs in frequency modeling for insurance portfolios (see, e.g., Lloyd et al. (2015), Rigby and Stasinopoulos (2005) and Wang and Shi (2014)). This extension is essential to ensure that fairness corrections remain compatible with standard actuarial practice. Second, rather than committing to a single criterion a priori, we show how a wide range of fairness criteria, spanning both individual and group-level notions, can be incorporated into a single GPR model through targeted modifications of the kernel structure and objective function. Third, we assess these fairness interventions both individually and jointly, comparing their effects side by side on standard fairness and accuracy metrics, and on the overall balance between collected premiums and incurred claim.

The quantitative experiment aims to provide clarity in the fairness versus performance debate, challenging the common binary framing that one fairness notion must be chosen over another. Instead, we advocate for a calibrated and context-aware application, acknowledging that fairness is multidimensional and sensitive to data, legal, and ethical context. To the best of our knowledge, this is the first work that integrates multiple, potentially conflicting fairness criteria within a single Bayesian non-parametric model. By doing so, we provide a framework through which fairness definitions can be compared, and evaluated according to the specific objectives and constraints faced by insurers and regulators.

The remainder of the article is organized as follows. Section 2 provides a comprehensive review of the actuarial fairness literature, highlighting key contributions, points of disagreement, and tensions between individual and group fairness paradigms. Section 3 introduces the Generalized Gaussian Process Regression framework for Poisson count data with risk exposure and sets up the notations. Section 4 presents the multi-criteria fair GGPR model and details how different fairness criteria are integrated within the Bayesian architecture. Section 5 reports empirical results, focusing on fairness-accuracy trade-offs, and actuarial fairness. Section 6 concludes by discussing implications for insurers and regulators, and by arguing for a context-dependent application of fairness criteria in insurance pricing.

2. Fairness in Insurance Pricing

This section reviews the existing literature on fairness in insurance pricing, with the objective of clarifying conceptual distinctions, methodological choices, and points of disagreement. Rather than proposing yet another definition of fairness, we emphasize how different fairness notions emerge from distinct normative, regulatory, and modeling assumptions. We show that many apparent contradictions in the literature stem from enforcing fairness at different levels (individual, portfolio, or market) and from implicitly privileging one objective over others.

Classical actuarial fairness is rooted in the principle of risk equivalence. Policyholders with identical risk profiles should, in expectation, pay identical premiums. Under standard assumptions, this principle is operationalized through best-estimate pricing, typically via GLMs for claim frequency and severity (Denuit et al., 2019a). From this perspective, fairness is closely tied to long-term financial balance, mutualization, and insurer solvency.

The rise of algorithmic pricing models has challenged this view. Machine learning methods, while often improving predictive accuracy, may rely on complex interactions and proxy variables that correlate with protected attributes, raising concerns about discrimination even when sensitive variables are excluded. This shift has brought concepts from algorithmic fairness into actuarial science, including demographic parity, individual fairness, counterfactual fairness, and omitted variable bias (often referred to as proxy discrimination in the actuarial literature).

A key difficulty is that these notions are not merely technical constraints but encode normative choices about what constitutes unjust discrimination. As a result, fairness in insurance pricing has become a multidimensional concept, spanning legal compliance, ethical acceptability, statistical properties, and market outcomes.

A prominent strand of the literature advocates for individual-level notions of fairness, focusing on the elimination of proxy discrimination. Lindholm et al. (2024c) argue that group-based constraints such as demographic parity are ill-suited to insurance pricing, as they may distort risk signals and undermine efficiency. Instead, they propose a discrimination-free price, constructed by reweighting premiums via inverse probability weighting to remove the influence of protected attributes while preserving causal risk factors.

This approach interprets fairness as a property of individual pricing decisions. Two individuals with the same non-protected characteristics should face the same expected premium, regardless of group membership. In this framework, proxy discrimination is treated as the central concern, and fairness violations arise when protected attributes affect premiums through inadmissible causal pathways. Note that causal notions of admissibility are inherently subjective and depend on unverifiable assumptions.

The sensitivity-based extensions proposed in Lindholm et al. (2024b) further refine this view by quantifying the marginal influence of protected attributes on predicted premiums.

These measures offer a continuous notion of discrimination, allowing insurers to assess how strongly pricing outcomes depend on sensitive variables, even indirectly.

While conceptually appealing, individual fairness approaches have been criticized for largely abstracting away from portfolio composition and market equilibrium. By construction, they ensure fairness at the individual level, but remain agnostic about distributional effects across groups or the dynamics induced by policyholder behavior.

In contrast, a second strand of the literature emphasizes group-level fairness criteria, such as demographic parity or conditional parity, motivated by concerns about social equity and regulatory compliance. Early actuarial contributions by Hu et al. (2023) and Hu et al. (2024) demonstrate how premium corrections enforcing demographic parity can be implemented within standard pricing models, and analyze their impact on predictive accuracy.

Xin and Huang (2024) provide a comprehensive overview of fairness criteria in insurance pricing, explicitly distinguishing between individual and group notions of fairness and mapping them to different regulatory regimes. They classify proxy discrimination as a form of individual unfairness, aligning with Lindholm et al. (2024c).

Building on these perspectives, O. Côté et al. (2024) propose a typology of fairness interventions (best-estimate, unaware, aware, hyperaware, and corrective) that explicitly distinguishes between models that ignore protected attributes and those that actively correct for group disparities.

A central insight of this literature is that fairness cannot be evaluated solely at the individual or portfolio level. M.-P. Côté et al. (2024) show that portfolio-specific fairness interventions may fail to generalize market-wide, particularly when policyholders can select into or out of contracts. In such settings, individual fairness corrections may unintentionally exacerbate adverse selection, leading to inequitable outcomes at the population level.

These results highlight a fundamental tension. Fairness interventions that preserve individual pricing accuracy may undermine collective solidarity, while group-based corrections may induce cross-subsidization and efficiency losses.

The divergence between individual and group fairness is further complicated by regulatory constraints. In many jurisdictions, insurers are prohibited from using certain sensitive attributes (e.g., gender, ethnicity), yet may still be required to demonstrate that pricing models do not induce indirect discrimination. Under the Affordable Care Act in the United States, health insurers are restricted to a small set of rating variables, prioritizing solidarity over risk-based pricing. Conversely, in jurisdictions such as Colorado, insurers are prohibited from using ethnicity data while still being required to prove the absence of discriminatory outcomes, creating an implicit demand for counterfactual or group-level fairness audits.

These regulatory asymmetries underscore the impracticality of committing to a single fairness definition. Instead, insurers must navigate a complex landscape where fairness is enforced differently depending on legal, social, and market contexts.

Despite the variety of fairness approaches proposed in the literature, implementations typically rely on a single criterion. Sensitivity-based measures provide diagnostic insight but do not resolve trade-offs between competing fairness notions. Moreover, fairness is usually enforced through post-processing adjustments rather than embedded within the pricing model, limiting transparency into how fairness constraints interact with model uncertainty, nonlinearity, and predictive structure. These limitations motivate a unified modeling framework in which multiple fairness criteria can be jointly incorporated.

3. GPR for Count Data with Risk Exposure

This section introduces the Generalized Gaussian Process Regression (GGPR) for Poisson count data with exposure, which is particularly relevant for modeling insurance claim frequencies. The GGPR was initially proposed by Chan and Dong (2011) where only similarity between risk profiles matters. GGPR applies to response obeying a distribution in the Exponential Dispersion (ED) family, exactly as with GLMs and GAMs. The GGPR was recently adapted by Hainaut and Denuit (2025) to account for policies with different exposures to risk, like the coverage duration. This is important for actuarial applications where exposures often play a major role, as they reflect the volume of the risk.

3.1 Poisson Frequency Model with Exposure

Consider an insurance portfolio $\mathcal{X} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\} \subset \mathbb{R}^m$, where each policy $i = 1, \dots, n$ is described by a vector of rating factors $\boldsymbol{x}_i = (x_{i1}, \dots, x_{im})$. These covariates are partitioned into a set of allowed covariates $\mathcal{A}$ and a set of prohibited or sensitive covariates $\mathcal{D}$, with $\mathcal{A} \cup \mathcal{D} = \{1, \dots, m\}$ and $\mathcal{A} \cap \mathcal{D} = \varnothing$. We write $\boldsymbol{a}_i = (x_{ij})_{j \in \mathcal{A}}$ and $\boldsymbol{d}_i = (x_{ij})_{j \in \mathcal{D}}$ for the corresponding subvectors.

Proposition 3.1. *Let N denote the number of claims reported by a policy over an exposure period $\nu > 0$. Conditionally on the covariates $\boldsymbol{x}$, we assume a Poisson model with exposure:*

$$N \mid \boldsymbol{x} \sim \mathcal{P}(\nu \lambda(\boldsymbol{x}))\,, \tag{1}$$

where $\lambda(\boldsymbol{x}) > 0$ is the claim intensity per unit of exposure. Define the key ratio:

$$Y := \frac{N}{\nu}, \tag{2}$$

which represents the claim frequency per unit of exposure. Since $N = \nu Y$ takes values in $\mathbb{N}$, the conditional probability mass function of Y is induced by that of N:

$$p(y \mid \boldsymbol{x}) = \mathbb{P}(N = \nu y \mid \boldsymbol{x}) = \exp(-\nu\lambda(\boldsymbol{x})) \, \frac{(\nu\lambda(\boldsymbol{x}))^{\nu y}}{(\nu y)!}, \tag{3}$$

which is a density in the ED family (Wuthrich & Merz, 2021) as shown in Proof 1 in Appendix A.

The objective is to approximate the (unknown) conditional expectation of the response $\nu Y \sim \mathcal{P}\left(\lambda\left(\boldsymbol{x}\right)\nu\right)$ given the covariates to derive the amount of pure premium:

$$\mu(\boldsymbol{x}) = \mathbb{E}[Y \mid \boldsymbol{x}] = \gamma'(\theta(\boldsymbol{x})) = \lambda(\boldsymbol{x})\,, \tag{4}$$

with $\theta(\boldsymbol{x}) = \ln \lambda(\boldsymbol{x})$ and $\gamma'(\cdot) = \exp(\cdot)$.

Remark 3.2 (Generalization to the ED family). While we present the notation using count data, the methodology can be generalized to Gamma or Inverse Gaussian distribution for claim severities.

3.2 GGPR for Log-Link Counts

To model the conditional mean $\mu(\boldsymbol{x})$, we introduce a latent function $g(\cdot)$ defined on the log scale:

$$\ln \mu(\boldsymbol{x}) = g(\boldsymbol{x}) + \varepsilon, \qquad \varepsilon \sim \mathcal{N}\big(0, \sigma_n^2\big) \ . \tag{5}$$

This formulation induces nonlinear dependence of the claim frequency on the covariates through a kernel function (similarity matrix), while retaining the interpretability of a log-link model.

Proposition 3.3. *Let* $\boldsymbol{g} = (g(\boldsymbol{x}_1), \ldots, g(\boldsymbol{x}_n))^\top$. *Under the Gaussian Process (GP) prior,* $\boldsymbol{g}$ *follows a multivariate Gaussian distribution with* n*-mean vector* $\mathbf{0}$ *and* $n \times n$ *Gram matrix:*

$$C(\boldsymbol{X}, \boldsymbol{X}) = K(\boldsymbol{X}, \boldsymbol{X}) + \sigma_n^2 \boldsymbol{I}_n, \tag{6}$$

where K *is the kernel matrix and* σ_n^2 *is a small scale variability. See Proof 2 in Appendix A.*

3.3 Posterior Inference via Laplace Approximation

Given observations (key ratios) $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ and exposures $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_n)^\top$, the conditional log-likelihood takes the Poisson form:

$$\ln p(\boldsymbol{y} \mid \boldsymbol{g}) = \sum_{i=1}^{n} \nu_i \big(y_i g_i - \exp(g_i)\big) + c(y_i, \nu_i), \tag{7}$$

where $c(\cdot)$ collects terms independent of $\boldsymbol{g}$. The posterior density of the latent vector $\boldsymbol{g}$ is then:

$$p(\boldsymbol{g} \mid \boldsymbol{y}, \boldsymbol{X}) \propto p(\boldsymbol{y} \mid \boldsymbol{g})\, p(\boldsymbol{g} \mid \boldsymbol{X}) \, . \tag{8}$$

Because the key-ratio Y follows a non-Gaussian distribution, the Poisson likelihood is not conjugate to the Gaussian prior. Therefore, the posterior distribution of the latent function $\boldsymbol{g}$ is analytically intractable (since it is no longer Gaussian) and must be approximated.

Following Williams and Rasmussen (2006), we approximate the posterior using a Laplace approximation. Specifically, the posterior is approximated by a Gaussian distribution centered at its mode:

$$q(\boldsymbol{g} \mid \boldsymbol{y}, \boldsymbol{X}) \approx \mathcal{N}(\hat{\boldsymbol{g}}, \Sigma_{\boldsymbol{g}}). \tag{9}$$

The n-dimensional mean vector $\hat{\boldsymbol{g}}$ is equal to (see Proof 6):

$$\hat{\boldsymbol{g}} = \arg\max_{\boldsymbol{g}} \{\psi(\boldsymbol{g})\} \ , \qquad \psi(\boldsymbol{g}) = \boldsymbol{\nu}^\top (\boldsymbol{y} \odot \boldsymbol{g} - \exp(\boldsymbol{g})) - \tfrac{1}{2} \boldsymbol{g}^\top C(\boldsymbol{X}, \boldsymbol{X})^{-1} \boldsymbol{g} \, , \tag{10}$$

where $\odot$ denotes the Hadamard product. The $n \times n$ covariance matrix $\Sigma_{\boldsymbol{g}}$ is given by the inverse negative Hessian of the log-posterior evaluated at $\hat{\boldsymbol{g}}$ (see Proof 7):

$$\Sigma_{\boldsymbol{g}} = \left(C(\boldsymbol{X}, \boldsymbol{X})^{-1} + \operatorname{diag}(\boldsymbol{\nu} \odot \exp(\hat{\boldsymbol{g}}))\right)^{-1}, \tag{11}$$

with the $n \times n$ diagonal matrix $H(\boldsymbol{g}) = \operatorname{diag}(\boldsymbol{\nu} \odot \exp(\hat{\boldsymbol{g}}))$.

The optimization problem defining $\hat{\boldsymbol{g}}$ in Eq. (10) is typically solved using Newton-Raphson iterations with $\boldsymbol{g}^{(0)} = \ln\left(\frac{1}{n}\sum_i y_i\right)$:

$$\hat{\boldsymbol{g}}^{(j)} = \hat{\boldsymbol{g}}^{(j-1)} - \left(\nabla^2 \psi\left(\hat{\boldsymbol{g}}^{(j-1)}\right)\right)^{-1} \nabla \psi\left(\hat{\boldsymbol{g}}^{(j-1)}\right), \tag{12}$$

where the gradient $\nabla \psi\left(\hat{\boldsymbol{g}}^{(j-1)}\right)$ takes the following simple form:

$$\nabla \psi\left(\hat{\boldsymbol{g}}^{(j-1)}\right) = \boldsymbol{\nu}\left(\boldsymbol{y} - \exp\left(\hat{\boldsymbol{g}}^{(j-1)}\right)\right) - C(\boldsymbol{X}, \boldsymbol{X})^{-1} \hat{\boldsymbol{g}}^{(j-1)}, \tag{13}$$

and the Hessian is equal to:

$$\begin{aligned} \nabla^2 \psi\left(\hat{\boldsymbol{g}}^{(j-1)}\right) &= -\operatorname{diag}\left(\boldsymbol{\nu} \odot \exp\left(\hat{\boldsymbol{g}}^{(j-1)}\right)\right) - C(\boldsymbol{X}, \boldsymbol{X})^{-1} && (14) \\ &= -H\left(\hat{\boldsymbol{g}}^{(j-1)}\right) - C(\boldsymbol{X}, \boldsymbol{X})^{-1}. && (15) \end{aligned}$$

3.4 Prediction and Premium Estimation

Proposition 3.4 (New Policy)**.** *Let $\boldsymbol{x}_*$ denote a new policy with kernel vector:*

$$k_* = k(\boldsymbol{x}_*, \boldsymbol{X}), \qquad k_{**} = k(\boldsymbol{x}_*, \boldsymbol{x}_*). \tag{16}$$

Under the Laplace approximation, the predictive distribution of the latent variable $g_* = g(\boldsymbol{x}_*)$ is approximately Gaussian,

$$g_* \mid \boldsymbol{x}_*, \boldsymbol{X}, \boldsymbol{y} \approx \mathcal{N}(m_*, s_*^2), \tag{17}$$

with

$$\begin{aligned} m_* &= k_* C(\boldsymbol{X}, \boldsymbol{X})^{-1} \hat{\boldsymbol{g}}, && (18) \\ s_*^2 &= k_{**} - k_*\left(C(\boldsymbol{X}, \boldsymbol{X}) + H(\hat{\boldsymbol{g}})^{-1}\right)^{-1} k_*^\top. && (19) \end{aligned}$$

See Proof 4 in Appendix A.

Transforming back to the original scale, the predictive mean, that is the estimate of the key ratio Y_*, satisfies:

$$\mathbb{E}[Y_* \mid \boldsymbol{x}_*, \boldsymbol{X}, \boldsymbol{y}] \approx \mathbb{E}[\exp(g_*)] = \exp\left(m_* + \tfrac{1}{2} s_*^2\right), \tag{20}$$

which corresponds to the log-normal moment correction induced by the exponential link. This quantity provides an estimate of the pure premium per unit of exposure.

The corresponding expected claim count is:

$$\mathbb{E}[N_* \mid \boldsymbol{x}_*, \boldsymbol{X}, \boldsymbol{y}] \;\approx\; \nu_* \exp\!\big(m_* + \tfrac{1}{2} {s_*}^2\big)\,. \tag{21}$$

Remark 3.5 (Use of covariates). In this section, the GGPR model is formulated using the full set of available covariates $\boldsymbol{X} = (\boldsymbol{A}, \boldsymbol{D}) \in \mathbb{R}^{n \times m}$. In the next section, we explicitly distinguish between allowed (non-sensitive) covariates $\boldsymbol{A}$ and prohibited (sensitive) attributes $\boldsymbol{D}$ in order to introduce fairness modeling constraints.

Remark 3.6 (Kernel choice). Throughout this work, we use the radial basis function (RBF), also known as the squared exponential kernel, as a baseline choice due to its smoothness and universal approximation properties. While more elaborate kernel constructions may be preferable for mixed continuous and categorical data, kernel engineering is not the focus of this study and is left for future work. The RBF is defined as:

$$k\left(\boldsymbol{x}, \boldsymbol{x'}\right) = \sigma_f^2 \exp\!\left(-\frac{1}{2\ell^2}\|\boldsymbol{x} - \boldsymbol{x'}\|^2\right), \tag{22}$$

where $\|\cdot\|^2$ denotes the squared Euclidean distance, σ_f^2 controls the marginal variance of the latent function, and ℓ represents the characteristic length scale governing smoothness. Larger values of ℓ induce smoother functions, while smaller values allow for more rapid local variation.

Algorithm 1: GGPR for Log-Link Count Data

Data: Training data X, frequencies y, exposures ν, kernel $k(\cdot, \cdot)$

Compute Gram matrix: $C = K(X, X) + \sigma_n^2 I_n$

Initialize latent function: $\hat{\boldsymbol{g}}^{(0)} = \ln\left(\frac{1}{n}\sum_i y_i\right)$

for $j = 1, 2, \ldots$ *until convergence* **do**

- Compute gradient: $\nabla\psi(\hat{\boldsymbol{g}}^{(j-1)}) = \nu \odot \left(y - \exp(\hat{\boldsymbol{g}}^{(j-1)})\right) - C^{-1}\hat{\boldsymbol{g}}^{(j-1)}$
- Compute Hessian: $H^{(j-1)} = -\mathrm{diag}(\nu \odot \exp(\hat{\boldsymbol{g}}^{(j-1)})) - C^{-1}$
- Update latent mode using Newton-Raphson:
  $\hat{\boldsymbol{g}}^{(j)} = \hat{\boldsymbol{g}}^{(j-1)} - \left(H^{(j-1)}\right)^{-1}\nabla\psi(\hat{\boldsymbol{g}}^{(j-1)})$

Compute posterior covariance: $\Sigma_g = \left(C^{-1} + \mathrm{diag}(\nu \odot \exp(\hat{\boldsymbol{g}}))\right)^{-1}$

Predict new policy x_*: $k_* = k(x_*, X), \quad k_{**} = k(x_*, x_*)$

Compute predictive mean and variance of latent function:

$$m_* = {k_*}^\top C^{-1}\hat{\boldsymbol{g}}, \qquad {s_*}^2 = k_{**} - {k_*}^\top (C + H^{-1})^{-1} k_*$$

Compute predictive premium (key ratio): $\hat{\pi}_* = \mathbb{E}[Y_*] \approx \exp\left(m_* + \frac{1}{2}s_*^2\right)$

Compute expected claim count: $\mathbb{E}[N_*] \approx \nu_*\,\hat{\pi}_*$

3.5 Actuarial Fairness

In insurance pricing, actuarial fairness (or portfolio-level calibration) requires that the total predicted number of claims across all policies matches the total observed claims. Formally, if N_i denotes the observed claims and ν_i the exposure for policy i, a model is actuarially fair if:

$$\sum_{i=1}^{n} \nu_i \hat{Y}_i = \sum_{i=1}^{n} N_i, \tag{23}$$

where $\hat{Y}_i$ is the predicted claim frequency. This property ensures that the insurer's total expected losses are consistent with the realized experience, and is critical for regulatory and solvency compliance (Lindholm & Wüthrich, 2025).

Proposition 3.7 (Actuarial Fairness in GGPR)**.** *Let N_i denote the number of claims for policy i, ν_i its exposure, and $Y_i = N_i/\nu_i$ the claim frequency. Under the GGPR for log-link counts, let $\hat{\boldsymbol{g}}$ denote the Laplace posterior mode. Define the predictive claim frequency using the log-normal correction:*

$$\hat{Y}_i := \exp\Big(m_i + \tfrac{1}{2} s_i^2\Big), \tag{24}$$

where m_i and s_i^2 are the posterior predictive mean and variance of the latent g_i. Then, the posterior mode satisfies actuarial fairness at the portfolio level:

$$\sum_{i=1}^{n} \nu_i \exp(\hat{g}_i) = \sum_{i=1}^{n} N_i. \tag{25}$$

In contrast, the prediction $\hat{Y}_i$ does not satisfy this equality:

$$\sum_{i=1}^{n} \nu_i \hat{Y}_i \neq \sum_{i=1}^{n} N_i. \tag{26}$$

Proposition 3.8 (Balance corrected GGPR at training points)**.** *Let $\hat{\boldsymbol{\pi}}(\hat{\boldsymbol{g}})$ be the predictive premiums at the training points:*

$$\hat{\boldsymbol{\pi}}(\hat{\boldsymbol{g}}) = \exp\big(\hat{\boldsymbol{g}} + \tfrac{1}{2}\operatorname{diag}(\Sigma_{\hat{\boldsymbol{g}}})\big)\,, \qquad \Sigma_{\hat{\boldsymbol{g}}} = \big(C^{-1} + \operatorname{diag}\big(\boldsymbol{\nu} \odot e^{\hat{\boldsymbol{g}}}\big)\big)^{-1}. \tag{27}$$

There exists a first-order bias-corrected mode $\hat{\boldsymbol{g}}_{\mathrm{BC}} = \hat{\boldsymbol{g}} + \Delta\boldsymbol{g}$, obtained by solving the augmented linear system:

$$\begin{bmatrix} \nabla^2\psi(\hat{\boldsymbol{g}}) & (\nu^\top J)^\top \\ \nu^\top J & 0 \end{bmatrix} \begin{bmatrix} \Delta\boldsymbol{g} \\ \lambda_{\mathrm{act}} \end{bmatrix} = \begin{bmatrix} \nabla\psi(\hat{\boldsymbol{g}}) \\ \sum_{i=1}^{n} N_i - \sum_{i=1}^{n} \nu_i \hat{\pi}_i \end{bmatrix}, \tag{28}$$

where

$$J \;\approx\; \operatorname{diag}(\hat{\boldsymbol{\pi}})(C + \Sigma_{\hat{\boldsymbol{g}}})^{-1}. \tag{29}$$

The resulting predictive premiums $\hat{\boldsymbol{\pi}}(\hat{\boldsymbol{g}}_{\mathrm{BC}})$ satisfy the portfolio-level calibration constraint:

$$\sum_{i=1}^{n} \nu_i \hat{\pi}_i(\hat{\boldsymbol{g}}_{\mathrm{BC}}) = \sum_{i=1}^{n} N_i \tag{30}$$

to first order in $\Delta \boldsymbol{g}$.

Algorithm 2: Balance corrected GGPR for Log-Link Count Data

Data: Training data X, frequencies y, exposures ν, kernel $k(\cdot, \cdot)$

Result: Predictive premium $\hat{\pi}_*$ satisfying portfolio-level calibration

Compute Gram matrix: $C = K(X, X) + \sigma_n^2 I_n$

Initialize latent function: $\hat{\boldsymbol{g}}^{(0)} = \log\big(\frac{1}{n}\sum_{i=1}^{n} y_i\big)$

for $j = 1, 2, \ldots$ *until convergence* **do**

- Compute gradient: $\nabla\psi(\hat{\boldsymbol{g}}^{(j-1)}) = \nu \odot \left(y - e^{\hat{\boldsymbol{g}}^{(j-1)}}\right) - C^{-1}\hat{\boldsymbol{g}}^{(j-1)}$
- Compute Hessian: $H^{(j-1)} = -\operatorname{diag}\left(\nu \odot e^{\hat{\boldsymbol{g}}^{(j-1)}}\right) - C^{-1}$
- Newton update: $\hat{\boldsymbol{g}}^{(j)} = \hat{\boldsymbol{g}}^{(j-1)} - \big(H^{(j-1)}\big)^{-1}\nabla\psi(\hat{\boldsymbol{g}}^{(j-1)})$

Compute Laplace posterior covariance: $\Sigma_{\hat{\boldsymbol{g}}} = \big(C^{-1} + \operatorname{diag}\left(\nu \odot e^{\hat{\boldsymbol{g}}}\right)\big)^{-1}$

if *balance correction* **then**

- **for** $k = 1, 2, \ldots$ *until balance convergence* **do**
  - Compute training-point premiums: $\hat{\boldsymbol{\pi}} = \exp\big(\hat{\boldsymbol{g}} + \frac{1}{2}\operatorname{diag}(\Sigma_{\hat{\boldsymbol{g}}})\big)$
  - Compute balance residual: $r = \sum_{i=1}^{n} N_i - \sum_{i=1}^{n} \nu_i \hat{\pi}_i$
  - Approximate Jacobian: $J \approx \operatorname{diag}(\hat{\boldsymbol{\pi}})\big(C + \Sigma_{\hat{\boldsymbol{g}}}\big)^{-1}$
  - Solve the KKT system:

$$\begin{bmatrix} \nabla^2\psi(\hat{\boldsymbol{g}}) & (\nu^\top J)^\top \\ \nu^\top J & 0 \end{bmatrix} \begin{bmatrix} \Delta\boldsymbol{g} \\ \lambda_{\mathrm{act}} \end{bmatrix} = \begin{bmatrix} \nabla\psi(\hat{\boldsymbol{g}}) \\ r \end{bmatrix}$$

  - Update latent mode: $\hat{\boldsymbol{g}} \leftarrow \hat{\boldsymbol{g}} + \Delta\boldsymbol{g}$
  - Update covariance: $\Sigma_{\hat{\boldsymbol{g}}} = \big(C^{-1} + \operatorname{diag}(\nu \odot e^{\hat{\boldsymbol{g}}})\big)^{-1}$

For a new policy x_*, compute: $k_* = k(x_*, X), \qquad k_{**} = k(x_*, x_*)$

Predictive mean and variance: $m_* = k_*^\top C^{-1}\hat{\boldsymbol{g}}, \qquad s_*^2 = k_{**} - k_*^\top (C + H^{-1})^{-1} k_*$

Predictive balance corrected premium: $\hat{\pi}_* = \exp\big(m_* + \frac{1}{2}s_*^2\big), \qquad \mathbb{E}[N_*] = \nu_* \hat{\pi}_*$

4. Fair Generalized Gaussian Process Regression

Fairness in actuarial pricing involves multiple, generally incompatible constraints without explicit trade-offs, and most existing approaches address them in isolation. From an actuarial perspective, premiums must additionally satisfy calibration and portfolio balance conditions to prevent systematic overpricing or underpricing at the portfolio level (see Section 3.5). Fairness interventions generally violate this condition unless explicit rebalancing constraints are imposed (see e.g., Jamotton and Hainaut (2025)).

In this Section, we detail the introduction of fairness notions whithin the Generalized Gaussian Process Regression presented in Section 3. Specifically, we control notions of individual fairness through the similarity matrix, whereas group-level constraints are imposed as expectations or covariance restrictions on posterior moments of g (similarily to actuarial fairness in Section 3.5).

To motivate the construction, we first review existing fairness approaches organized into five families (best-estimate, unaware, aware, hyperaware, and corrective) by O. Côté et al. (2024). This classification refines the classical pre-/in-/post-processing taxonomy of Kamishima et al. (2012). Throughout this work, fairness is enforced directly during GGPR training (i.e., in-processing).

4.1 Fairness Families

Let $Y \in \mathbb{R}_+$ denote the target loss variable (the claim frequency per unit of exposure), $A \in \mathcal{A}$ the vector of permitted rating factors, and $D \in \mathcal{D}$ the vector of protected attributes. We assume a latent risk function $g(\boldsymbol{x})$, with $\boldsymbol{x} = (\boldsymbol{a}, \boldsymbol{d})$, and a pricing rule $\pi(\cdot)$ derived from the predictive distribution of g.

Definition 4.1 (Best-estimate premium)**.** A best-estimate premium uses all available information (A, D) and is sufficient for predicting Y:

$$\pi^{BE}(A, D) = \mathbb{E}[Y \mid A, D]. \tag{31}$$

While it may be statistically optimal, the best-estimate premium explicitly encodes direct discrimination whenever D is causally or associatively linked to Y. While indirect discrimination through A becomes irrelevant once D is included, such pricing is typically prohibited by anti-discrimination regulation.

Definition 4.2 (Unaware premium)**.** An unaware premium excludes protected attributes D:

$$\pi^{U}(A) = \mathbb{E}[Y \mid A]. \tag{32}$$

Although unaware pricing appears fair by construction, it generally suffers from omitted variable bias. When $A \leftarrow D \rightarrow Y$, the variables in A act as proxies for D, inducing indirect discrimination. This phenomenon has been extensively documented in both econometrics

and algorithmic fairness (see e.g. Barocas, Solon; Selbst, Andrew D (2016) and Kusner et al. (2017)).

Definition 4.3 (Aware premium)**.** An aware premium uses only permitted features A while correcting for the confounding effect of D:

$$\pi^A(A) = \mathbb{E}_D[\mathbb{E}[Y \mid A, D]]\,. \tag{33}$$

Aware premiums eliminate direct discrimination and correct for omitted variable bias by marginalising over D. The discrimination-free pricing framework of Lindholm et al. (2022, 2024a, 2024c) provides an actuarial formulation of this approach. However, $\pi^A(A)$ may still be statistically associated with D, leading to residual disparate impact through A (O. Côté et al., 2024, p.57).

Definition 4.4 (Hyperaware premium)**.** A hyperaware premium predicts only the component of Y that is independent of D:

$$\pi^H(A) = \mathbb{E}[Y^{\perp D} \mid A], \tag{34}$$

where $Y^{\perp D}$ denotes a component of Y statistically independent of D.

Hyperaware pricing aims to remove both direct and indirect pathways from D to Y by predicting only the component of the target variable Y that is conditionally independent of the sensitive attribute D, while restricting inputs to allowed features A. Unlike the aware model, which adjusts for omitted variable bias in a predictive sense, the hyperaware model attempts to target the fair signal in Y, i.e., the variation in Y that is invariant under changes in D. Formally, we consider the decomposition:

$$Y = Y^{\perp D} + Y^{\parallel D}, \tag{35}$$

where $Y^{\perp D}$ is the component of Y orthogonal (independent) to D, and $Y^{\parallel D}$ is the component of Y explained by D.

However, since $Y^{\perp D}$ is not directly observable, this decomposition is only well-defined under additional (strong) assumptions, and fairness improvements are difficult to quantify robustly. For these reasons, we chose not to model a hyperaware premium in this work. Hyperaware models nevertheless serve as an important conceptual bridge toward corrective pricing.

Definition 4.5 (Corrective premium)**.** A corrective premium uses only the component of A independent of D:

$$\pi^C(A^{\perp D}) = \mathbb{E}[Y \mid A^{\perp D}], \qquad A^{\perp D} \perp D. \tag{36}$$

Corrective premiums explicitly enforce independence between premiums and protected

attributes, typically achieving demographic parity:

$$\pi^C(A^{\perp D}) \perp D. \tag{37}$$

This may require corrective direct discrimination, which remains controversial from a legal standpoint but may be justified when addressing structural inequalities. One can say that this corrective premium is fair because it avoids historically accurate direct discrimination and disparate impact through A. Recent approaches include distributional alignment via optimal transport and Wasserstein barycenters Hu et al. (2023, 2024).

4.2 Multi-criteria fair GGPR for Log-Link Counts

These different fairness notions are enforced by altering the kernel architecture or by introducing linear equality constraints on the predictive distribution.

Remark 4.1 (Use of Non-Autocalibrated GGPR for Presentation). For clarity and to simplify the notation, the fairness interventions are presented in this section using the non-balance corrected GGPR, but they can be directly applied to the balance-corrected GGPR as well as shown in the empirical results in Section 5.

4.2.1 Best-estimate GGPR

Proposition 4.2 (Best-estimate kernel)**.** *The best-estimate kernel is defined as:*

$$k^{BE}(\boldsymbol{x}, \boldsymbol{x'}) = \sigma_f^2 \exp\left(-\frac{1}{2l^2}\|\boldsymbol{x} - \boldsymbol{x'}\|^2\right), \tag{38}$$

where $\boldsymbol{x}$ contains all available rating factors, including protected attributes, to encode similarity between policies.

Proposition 4.3 (Best-estimate GGPR premium)**.** *Let $\boldsymbol{x}_*$ denote a new policy with best-estimate kernel vector:*

$$k_*^{BE} = k(\boldsymbol{x_*}, \boldsymbol{X}), \qquad k_{**}^{BE} = k(\boldsymbol{x_*}, \boldsymbol{x_*}). \tag{39}$$

Under the Laplace approximation (with canonical link), the best-estimate GGPR premium, defined as the conditional expectation of the claim frequency, satisfies:

$$\pi^{BE}(\boldsymbol{x}_*) := \mathbb{E}[Y_* \mid \boldsymbol{x}_*, \boldsymbol{X}, \boldsymbol{y}] \;\approx\; \exp\Big(m_*^{BE} + \tfrac{1}{2}(s_*^{BE})^2\Big), \tag{40}$$

with:

$$m_*^{BE} = k_*^{BE}\big(C^{BE}\big)^{-1}\hat{\boldsymbol{g}}^{BE}, \tag{41}$$

$$(s_*^{BE})^2 = k_{**}^{BE} - k_*^{BE}\Big(C^{BE} + H(\hat{\boldsymbol{g}}^{BE})^{-1}\Big)^{-1}k_*^{BE^\top}, \tag{42}$$

where the best-estimate Gram matrix is:

$$C^{BE} = \sigma_n^2 \boldsymbol{I}_n + k^{BE}. \tag{43}$$

See Proof 10 in Appendix A.

If policies $\boldsymbol{x}$ and $\boldsymbol{x'}$ are similar then the corresponding $\pi(\boldsymbol{x})$ and $\pi(\boldsymbol{x'})$ will be strongly similar. Since the best-estimate kernel, in Eq. (38), uses all attributes, even, small changes in $\boldsymbol{d}$ can lead to large changes in $k(\boldsymbol{x}, \boldsymbol{x'})$, explicitly discriminating.

The kernel architecture can be modified to enforce fairness, especially individual fairness, which requires that similar individuals should be treated similarly (i.e., receive similar premium). Translated into the kernel language, if individuals $\boldsymbol{x}$ and $\boldsymbol{x'}$ are similar, except in the sensitive variables space, then $k(\boldsymbol{x}, \boldsymbol{x'})$ should remain high. This leads us to the unaware and aware kernels.

4.2.2 Unaware GGPR

Proposition 4.4 (Unaware kernel)**.** *The unaware kernel is defined as:*

$$k^U(\boldsymbol{a}, \boldsymbol{a'}) = \sigma_f^2 \exp\left(-\frac{1}{2l^2}\|\boldsymbol{a} - \boldsymbol{a'}\|^2\right), \tag{44}$$

where $\boldsymbol{a}$ contains allowed rating factors.

This ensures that, different sensitive values do not affect similarity, that is, regardless of $\boldsymbol{d} \neq \boldsymbol{d'}$, $\boldsymbol{a} = \boldsymbol{a'} \Rightarrow k = 1$.

Proposition 4.5 (Unaware GGPR premium)**.** *Let $\boldsymbol{a}_*$ denote a new policy with unaware kernel vector:*

$$k_*^U = k(\boldsymbol{a_*}, \boldsymbol{A}), \qquad k_{**}^U = k(\boldsymbol{a_*}, \boldsymbol{a_*}). \tag{45}$$

Under the Laplace approximation (with canonical link), the best-estimate GGPR premium, defined as the conditional expectation of the claim frequency, satisfies:

$$\pi^U(\boldsymbol{a}_*) := \mathbb{E}[Y_* \mid \boldsymbol{a}_*, \boldsymbol{A}, \boldsymbol{y}] \;\approx\; \exp\left(m_*^U + \tfrac{1}{2}(s_*^U)^2\right), \tag{46}$$

with:

$$m_*^U = k_*^U \left(C^U\right)^{-1} \hat{\boldsymbol{g}}^U, \tag{47}$$

$$(s_*^U)^2 = k_{**}^U - k_*^U \left(C^U + H(\hat{\boldsymbol{g}}^U)^{-1}\right)^{-1} k_*^{U\top}, \tag{48}$$

where the unaware Gram matrix is:

$$C^U = \sigma_n^2 \boldsymbol{I}_n + k^U. \tag{49}$$

See Proof 11 in Appendix A.

While the unaware GGPR removes direct dependence and discrimination on sensitive attributes D (through the kernel architecture in Eq. (44)), it does not fully address discrimination. This is because ignoring sensitive attributes does not eliminate their indirect influence on non-sensitive features, a phenomenon known as proxy discrimination or omitted variable bias.

4.2.3 Aware GGPR

Proposition 4.6 (Aware kernel)**.** *Let $\boldsymbol{x} = (\boldsymbol{a}, \boldsymbol{d})$ denote the full feature vector, where $\boldsymbol{a}$ are allowed attributes and $\boldsymbol{d}$ sensitive attributes. The aware kernel is defined as the population-averaged kernel:*

$$k^A(\boldsymbol{a}, \boldsymbol{a}') \;:=\; \mathbb{E}_{\boldsymbol{d},\boldsymbol{d}'}\left[k^{BE}((\boldsymbol{a}, \boldsymbol{d}), (\boldsymbol{a}', \boldsymbol{d}'))\right], \tag{50}$$

where the expectation is taken with respect to the empirical distribution of the sensitive attributes.

In other words, the aware kernel measures similarity between two individuals by averaging the best-estimate kernel over the distribution of sensitive attributes. Importantly, the aware GGPR is strictly stronger than the unaware GGPR. While the unaware kernel removes sensitive attributes from the feature space, it does not correct for proxy effects encoded in allowed variables. The aware kernel neutralises the sensitive attributes impact through marginalisation, thereby mitigating both direct and indirect discrimination.

In practice, the expectation in Eq. (50) is approximated by Monte-Carlo sampling. Let $\{\boldsymbol{d}^{(m)}\}_{m=1}^{M}$ be independent and identically distributed samples from the empirical distribution of $\boldsymbol{D}$. The aware kernel admits the approximation:

$$k^A(\boldsymbol{a}, \boldsymbol{a}') \;\approx\; \frac{1}{M}\sum_{m=1}^{M} k^{BE}\Big((\boldsymbol{a}, \boldsymbol{d}^{(m)}), (\boldsymbol{a}', \boldsymbol{d}^{(m)})\Big), \tag{51}$$

which corresponds to kernel-level averaging over sensitive attributes.

This construction is closely related to counterfactual fairness (Plecko & Bareinboim, 2024). For a fixed individual with allowed characteristics $\boldsymbol{a}$, the predicted premium is invariant, in expectation, to counterfactual changes in the sensitive attributes $\boldsymbol{d}$. The model therefore approximates the counterfactual quantity:

$$\mathbb{E}[Y_* \mid \boldsymbol{a}_*, \mathrm{do}(\boldsymbol{D}_* = \boldsymbol{d})], \tag{52}$$

averaged over the population distribution of $\boldsymbol{D}_*$. We refer to causal theory for more details about counterfactual fairness (Cornacchia et al., 2023; Pearl, 2009; Plecko & Bareinboim, 2024).

Proposition 4.7 (Aware GGPR premium)**.** *Let $\boldsymbol{a}_*$ denote a new policy with aware kernel*

vector

$$k_*^A = k^A(\boldsymbol{a}_*, \boldsymbol{A}), \qquad k_{**}^A = k^A(\boldsymbol{a}_*, \boldsymbol{a}_*). \tag{53}$$

Under the Laplace approximation (with canonical link), the aware GGPR premium, defined as the conditional expectation of the claim frequency, satisfies

$$\pi^A(\boldsymbol{a}_*) := \mathbb{E}[Y_* \mid \boldsymbol{a}_*, \boldsymbol{A}, \boldsymbol{y}] \;\approx\; \exp\!\Big(m_*^A + \tfrac{1}{2}(s_*^A)^2\Big)\,, \tag{54}$$

with

$$m_*^A = k_*^A \big(C^A\big)^{-1} \hat{\boldsymbol{g}}^A, \tag{55}$$

$$(s_*^A)^2 = k_{**}^A - k_*^A \Big(C^A + H(\hat{\boldsymbol{g}}^A)^{-1}\Big)^{-1} {k_*^A}^\top, \tag{56}$$

where the aware Gram matrix is defined as:

$$C^A = \sigma_n^2 \boldsymbol{I}_n + k^A. \tag{57}$$

See Proof 12 in Appendix A.

The approach is inspired by, but conceptually distinct from, the post-hoc fairness correction proposed by Lindholm et al. (2022). In their framework, fairness is enforced by averaging predicted premiums over sensitive attributes,

$$\pi^{\text{fair}}(\boldsymbol{a}) = \mathbb{E}_{\boldsymbol{d}}\big[\pi^{BE}(\boldsymbol{a}, \boldsymbol{d})\big]\,, \tag{58}$$

which can be interpreted as a post-stratification step applied after model estimation. Post-hoc averaging, as considered for instance by Lindholm et al. (2022), removes explicit dependence on $\boldsymbol{d}$ at the level of point predictions (by averaging pointwise expected premiums over the sensitive attributes) but systematically understates predictive uncertainty by ignoring variability arising from marginalization over sensitive attributes.

In contrast, the aware GGPR integrates out sensitive information at the kernel level (into the similarity measure), prior to posterior inference. This affects the covariance structure of the latent function and, consequently, both posterior mean and uncertainty. As a consequence, predictive uncertainty under aware GGPR reflects both the conditional variability given sensitive attributes and the additional uncertainty arising from marginalizing over them. This distinction can be made explicit using the law of total variance. For the latent variable g,

$$\mathrm{Var}^A(g_*) = \mathbb{E}_{\boldsymbol{d}}[\mathrm{Var}(g_* \mid \boldsymbol{a}_*, \boldsymbol{d})] + \mathrm{Var}_{\boldsymbol{d}}[\mathbb{E}(g_* \mid \boldsymbol{a}_*, \boldsymbol{d})]\,. \tag{59}$$

From an actuarial perspective, this difference has important implications. Aware GGPR yields premiums that are counterfactually fair in expectation while producing wider and more realistic uncertainty bands, thereby allowing uncertainty due to sensitive attributes

to be explicitly accounted for in pricing, reserving, and risk management.

4.2.4 Corrective GGPR

Proposition 4.8 (Corrective GGPR)**.** *Under the Laplace approximation (with canonical log-link), the corrective GGPR premium, defined as the conditional expectation of the claim frequency with an empirical covariance constraint with respect to a sensitive attribute* $\boldsymbol{d}$*, satisfies:*

$$\pi^C(\boldsymbol{x}_*) := \mathbb{E}[Y_* \mid \boldsymbol{x}_*, \boldsymbol{X}, \boldsymbol{y}] \;\approx\; \exp\Big(m_*^C + \tfrac{1}{2}(s_*^C)^2\Big), \tag{60}$$

with:

$$m_*^C = k_*^\top C^{-1} \hat{\boldsymbol{g}}^C, \tag{61}$$

$$(s_*^C)^2 = k_{**} - k_*^\top \big(C + H(\hat{\boldsymbol{g}}^C)^{-1}\big)^{-1} k_*. \tag{62}$$

The corrective posterior mode $\hat{\boldsymbol{g}}^C$ *is obtained by solving the augmented KKT system:*

$$\begin{bmatrix} C^{-1} + H(\hat{\boldsymbol{g}}_0) & a \\ a^\top & 0 \end{bmatrix} \begin{bmatrix} \Delta \boldsymbol{g} \\ \lambda \end{bmatrix} = \begin{bmatrix} \nabla \psi(\hat{\boldsymbol{g}}_0) \\ c \end{bmatrix}, \tag{63}$$

with:

$$\hat{\boldsymbol{g}}^C = \hat{\boldsymbol{g}}_0 + \Delta \boldsymbol{g}, \qquad a := J^\top C_n \boldsymbol{d}, \qquad c := -\boldsymbol{d}^\top C_n \hat{\boldsymbol{\pi}}_0. \tag{64}$$

Here $C_n = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$ *is the centering matrix,* $H(\hat{\boldsymbol{g}}) = \operatorname{diag}(\boldsymbol{\nu} \odot \exp(\hat{\boldsymbol{g}}))$*,* $\hat{\boldsymbol{\pi}}_0 = \exp(m_0 + \frac{1}{2}s_0^2)$*, and*

$$J = \frac{\partial \hat{\boldsymbol{\pi}}}{\partial \boldsymbol{g}}\Big|_{\hat{\boldsymbol{g}}_0} \approx \operatorname{diag}(\hat{\boldsymbol{\pi}}_0)\big(C + H(\hat{\boldsymbol{g}}_0)^{-1}\big)^{-1}. \tag{65}$$

This construction ensures that the empirical covariance between the predictive premium and sensitive attributes:

$$\boldsymbol{d}^\top C_n \hat{\boldsymbol{\pi}}^C \approx 0, \tag{66}$$

is satisfied at the linearized level.

Algorithm 3: Corrective GGPR

Data: Training data X, frequencies y, exposures ν, kernel $k(\cdot,\cdot)$

Result: Predictive premium $\hat{\pi}^C$ satisfying $\boldsymbol{d}^\top C_n \hat{\boldsymbol{\pi}} \approx 0$

Compute Gram matrix: $C = K(X,X) + \sigma_n^2 I_n$

Initialize latent mode:

$$\hat{\boldsymbol{g}}^{(0)} = \ln\left(\frac{1}{n}\sum_i y_i\right)$$

for $j = 1, 2, \ldots$ *until convergence* **do**

Compute gradient: $\nabla\psi(\hat{\boldsymbol{g}}^{(j-1)}) = \nu \odot (y - \exp(\hat{\boldsymbol{g}}^{(j-1)})) - C^{-1}\hat{\boldsymbol{g}}^{(j-1)}$

Compute Hessian components: $H^{(j-1)} = \mathrm{diag}(\nu \odot \exp(\hat{\boldsymbol{g}}^{(j-1)}))$

Compute predictive mean and variance at training points:

$$m^{(j-1)} = C^{-1}\hat{\boldsymbol{g}}^{(j-1)}, \qquad s^{2,(j-1)} = \mathrm{diag}\big((C + H^{-1})^{-1}\big)$$

Compute linearized premium: $\hat{\boldsymbol{\pi}}^{(j-1)} = \exp\big(m^{(j-1)} + \frac{1}{2}s^{2,(j-1)}\big)$

Define constraint vector and offset:

$$a^{(j-1)} = C_n\,\mathrm{diag}(\hat{\boldsymbol{\pi}}^{(j-1)})\,\boldsymbol{d}, \qquad c^{(j-1)} = -\,\boldsymbol{d}^\top C_n \hat{\boldsymbol{\pi}}^{(j-1)}$$

Solve augmented KKT system:

$$\begin{bmatrix} C^{-1} + H^{(j-1)} & a^{(j-1)} \\ (a^{(j-1)})^\top & 0 \end{bmatrix} \begin{bmatrix} \Delta\boldsymbol{g}^{(j)} \\ \lambda \end{bmatrix} = \begin{bmatrix} \nabla\psi(\hat{\boldsymbol{g}}^{(j-1)}) \\ c^{(j-1)} \end{bmatrix}$$

Update latent mode:

$$\hat{\boldsymbol{g}}^{(j)} = \hat{\boldsymbol{g}}^{(j-1)} + \Delta\boldsymbol{g}^{(j)}$$

Compute posterior covariance: $\Sigma_g = (C^{-1} + H(\hat{\boldsymbol{g}}))^{-1}$

Predict new policy x_*: $m_* = k_*^\top C^{-1}\hat{\boldsymbol{g}}, \qquad s_*^2 = k_{**} - k_*^\top (C + H^{-1})^{-1} k_*$

Compute predictive premium: $\hat{\pi}_* = \exp\big(m_* + \frac{1}{2}s_*^2\big)$

Ensure: $\boldsymbol{d}^\top C_n \hat{\boldsymbol{\pi}} \approx 0$

5. Results

In this section, we evalute our approach based on a non-life insurance portfolio. The analysis is organized in two stages.

First, we assess the predictive performance of the Generalized Gaussian Process Regression (GGPR) model for claim frequency estimation by benchmarking it against well-established regression approaches in the actuarial literature, namely the Poisson Generalized Linear Model (GLM) and a Poisson gradient boosting model implemented with LightGBM. This comparison focuses on the goodness-of-fit and calibration of the resulting pure premiums.

Second, we analyze the impact of introducing and combining different fairness notions into the GGPR in terms of predictive accuracy, and actuarial fairness. We benchmark our FGGPR with a post-processing fairness approach where GGPR predictions are adjusted a posteriori using Wasserstein barycenters.

5.1 Regression Models

We consider a non-life insurance portfolio composed of $n = 62\,289$ individual policies. For each policy i, we observe a vector of two quantitative (policyholder's age and vehicle age) and three categorical (gender, geographic area, and vehicle power) rating characteristics $\boldsymbol{x}_i$, an exposure period $\nu_i > 0$, and an observed claim count $N_i \in \mathbb{N}$. We assume that, conditionally on $\boldsymbol{x}_i$, the claim count follows a Poisson-type structure $N_i \mid \boldsymbol{x}_i \sim \text{Poisson}(\nu_i \lambda_i)$ where λ_i denotes the claim intensity per unit of exposure.

To estimate $\hat{\lambda}_i$, we consider three regression models. A classical Generalized Linear Model (GLM), a gradient-boosted decision tree model (LightGBM), and the (autocalibrated) GGPR adapted to Poisson data with different risk exposures.

The baseline model is a Poisson GLM with log-link function. The conditional mean is specified as:

$$\mathbb{E}[N_i \mid \boldsymbol{x}_i] = \nu_i \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}), \tag{67}$$

or equivalently,

$$\ln \hat{\lambda}_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}. \tag{68}$$

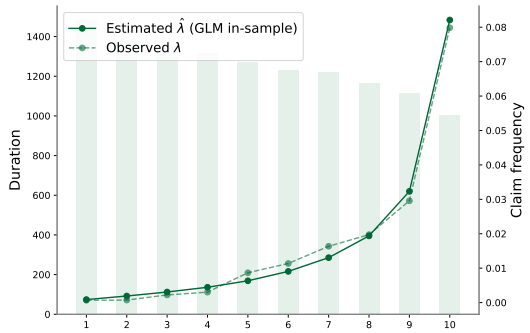
The exposure ν_i enters the model as an offset, ensuring interpretability in terms of frequency per exposure unit. The parameter vector $\boldsymbol{\beta}$ is estimated by maximum likelihood. This model serves as a transparent actuarial benchmark with well-understood statistical properties.

While a GLM is actuarially fair (i.e., the total predicted claim count equals the total observed count) by construction, it has limited flexibility. To capture nonlinearities and higher-order interactions between rating factors, we also fit a Poisson gradient boosting model using LightGBM. The model assumes $\ln \lambda_i = f(\boldsymbol{x}_i)$ where f is an additive ensemble of decision trees trained under a Poisson loss function, with exposure incorporated as

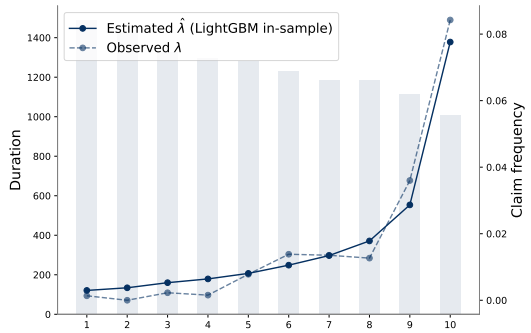
an offset. Hyperparameters (including tree depth, learning rate, number of leaves, and regularization terms) are tuned using an optuna study (see Table 13), with the Poisson deviance on a validation set as the optimization criterion.

As a flexible non-parametric alternative, we consider the GGPR model for count data with risk exposure (with and without balance correction) detailed in Section 3.

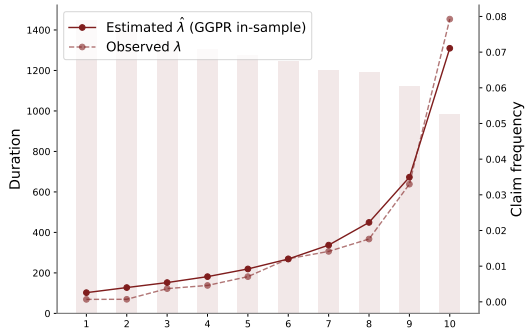
Figure 1 assesses local calibration by comparing observed and predicted claim frequencies per unit exposure across risk score bins. Policies are partitioned into ten quantile-based groups according to their predicted claim frequency $\hat{\lambda}_i$. A well-calibrated model is characterized by close alignment between predicted and empirical frequencies across all bins. Deviations can indicate local miscalibration, tail effects, systematic over or underestimation across the risk spectrum, or distributional distortions introduced by model structure.



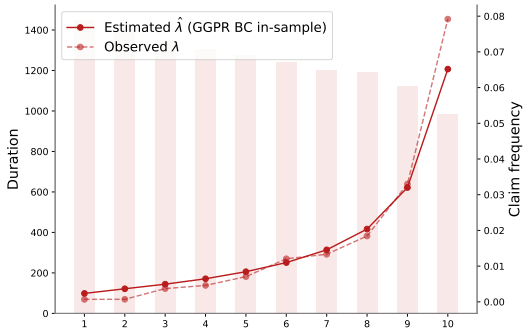
(a) GLM in-sample



(b) LightGBM in-sample



(c) GGPR in-sample



(d) GGPR BC in-sample

Figure 1: Observed versus predicted claim frequencies (per unit exposure) by risk score bins. For each quantile-bin, bars represent the total risk exposure (contract duration), while lines show the predicted claim frequency and the corresponding empirical frequency computed from observed claims.

Overall goodness-of-fit is assessed in Table 1 using the Poisson deviance:

$$D = 2 \sum_{i=1}^{n} \left[N_i \log \left(\frac{N_i}{\nu_i \hat{\lambda}_i} \right) - (N_i - \nu_i \hat{\lambda}_i) \right], \tag{69}$$

with the observed number of claims N_i, the observed risk exposure ν_i and the predicted

claim rate $\hat{\lambda}_i$ (i.e., predicted claim frequency per unit exposure). Lower deviance indicates a better fit to the observed claim counts. Overall, the deviance values are of comparable magnitude across models, indicating that all approaches achieve a similar level of global goodness-of-fit. In-sample, LightGBM attains the lowest deviance, reflecting its greater flexibility and ability to closely adapt to the training data. The GLM follows closely, while GGPR and its balance-corrected variant have a slightly higher deviance, consistent with the stronger regularisation imposed by the Gaussian process prior. Out-of-sample, the GLM achieves the lowest deviance, while LightGBM shows a small deterioration relative to its in-sample performance, suggesting mild overfitting. GGPR and balance-corrected GGPR remain competitive, with deviance values close to those of LightGBM and GLM. Importantly, the balance correction (BC) applied to GGPR results in only a slight decrease in deviance, demonstrating that improved fairness properties can be obtained without materially degrading predictive fit.

	in-sample	out-of-sample
GLM	1651.3378	1477.1372
LightGBM	1638.0187	1483.4738
GGPR	1681.9403	1488.9850
GGPR BC	1680.6705	1484.9006

Table 1: Poisson deviance values on the training set (in-sample) and test set (out-of-sample). Lower deviance indicates a better fit to the observed claim counts.

Finally, we assess total balance by comparing observed and predicted numbers of claims both in-sample in Table 2 and out-of-sample in Table 3. It highlights over or under-prediction across claim counts and overall totals. In-sample, the Poisson GLM reproduces the total number of observed claims exactly, because of maximum likelihood estimation under a log-link which enforces portfolio-level balance. The balance-calibrated GGPR exhibits the same property up to numerical precision. Despite this global constraint, both models exhibit misallocation across claim counts. All models overpredict single-claim policies ($N = 1$) and underpredict multi-claim policies ($N \geq 2$). This pattern is most pronounced for GGPR, which overestimates $N = 1$ claims by nearly 30 and underestimates $N = 2$ claims by more than 12. The balance correction substantially reduces this distortion, bringing the overprediction for $N = 1$ closer to that of the GLM. LightGBM does not enforce portfolio-level balance and consequently overpredicts total claims by approximately 5.6 in-sample. While its distribution across claim counts is closer to the GLM than uncorrected GGPR, it still amplifies the overprediction of single-claim policies. Overall, these results indicate that portfolio-level balance does not imply correct allocation across claim counts, and that model flexibility tends to shift mass toward single-claim outcomes.

N	Obs	GLM	LightGBM	GGPR	GGPR BC
1	172	183.1752 (+11.18)	189.4931 (+17.49)	201.8983 (+29.90)	185.3868 (+13.39)
2	18	6.5506 (−11.45)	5.8621 (−12.14)	5.3310 (−12.67)	4.5004 (−13.50)
3	0	0.2625 (+0.26)	0.2057 (+0.21)	0.1301 (+0.13)	0.1011 (+0.10)
Total	190	190.0000 (0.00)	195.5686 (+5.57)	207.3626 (+17.36)	189.9905 (−0.01)

Table 2: Observed and predicted (in-sample) number of claims by claim count, with deviations from the observed values shown in parentheses.

Out-of-sample, none of the models preserve total balance, as expected. All models systematically overpredict total claims, with the magnitude increasing with model flexibility. The GLM exhibits the smallest deviation (+13.7), followed by balance-corrected GGPR (+15.8), LightGBM (+19.4), and uncorrected GGPR (+32.6). The same structural pattern observed in-sample persists out-of-sample (i.e., overprediction of single-claim policies, underprediction of multiple claims, and small but non-zero mass assigned to $N = 3$ despite no observed occurrences). Uncorrected GGPR shows the largest imbalance, suggesting that smoothing and posterior shrinkage, while beneficial for calibration curves, may inflate the expected number of low-count claims when extrapolated out-of-sample. The balance correction mitigates this effect, reducing both the total overprediction and the distortion across claim counts.

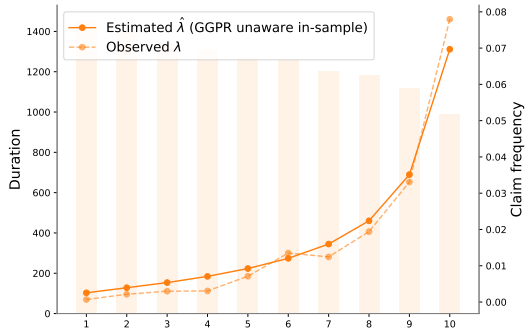
N	Obs	GLM	LightGBM	GGPR	GGPR BC
1	156	175.419 (+19.42)	181.7178 (+25.72)	195.2421 (+39.24)	179.2878 (+23.29)
2	12	6.0718 (−5.93)	5.4775 (−6.52)	5.1808 (−6.82)	4.3737 (−7.63)
3	0	0.2298 (+0.23)	0.1798 (+0.18)	0.1238 (+0.12)	0.0962 (+0.10)
Total	168	181.7304 (+13.73)	187.3814 (+19.38)	200.5495 (+32.55)	183.7596 (+15.76)

Table 3: Observed and predicted (out-of-sample) number of claims by claim count, with deviations from the observed values shown in parentheses.

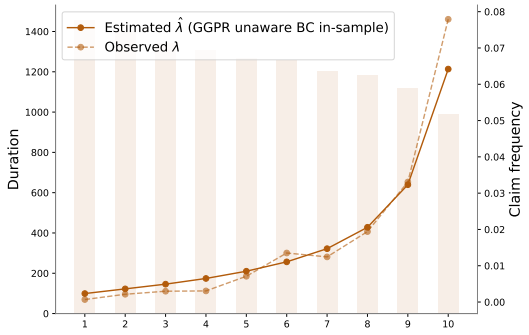
5.2 Fairness Benchmark

In this section, we focus on enforcing fairness in predicted claim rates with respect to a binary sensitive attribute (gender). Building on the GGPR model, we compare the effects of different fairness interventions (unaware, aware, corrective) on the distribution of predicted claim frequencies, individual and group-level measures of disparities, accuracy, overall balance.

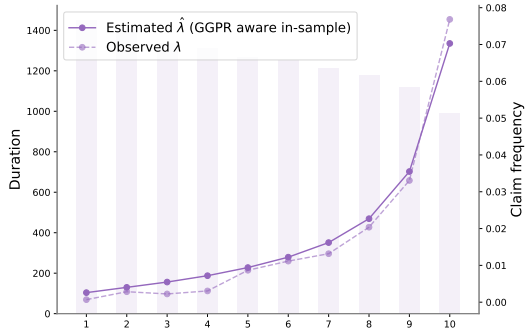
The unaware GGPR model is trained without any knowledge of the sensitive attribute. The aware GGPR model's kernel is averaged over the sensitive information. The corrective GGPR model is designed to reduce group-level disparities during training. We benchmark these fairness-aware GGP variants against a post-hoc Wasserstein barycenter correction.



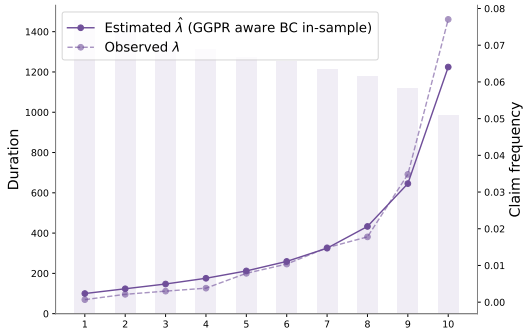
(a) Unaware GGPR in-sample



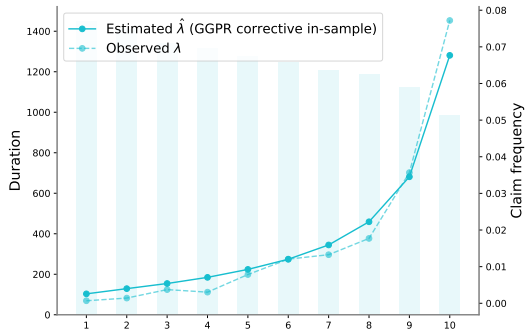
(b) Unaware GGPR BC in-sample



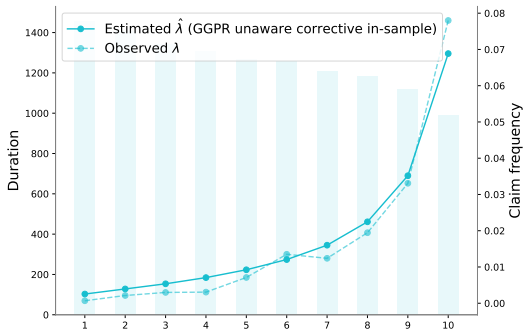
(c) Aware GGPR in-sample



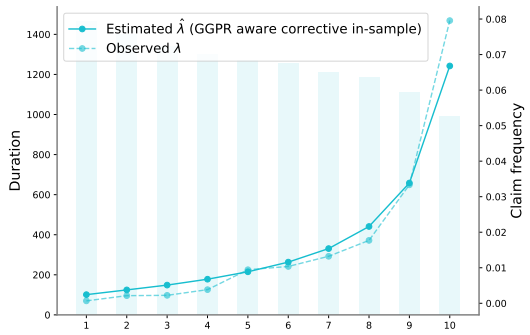
(d) Aware GGPR BC in-sample



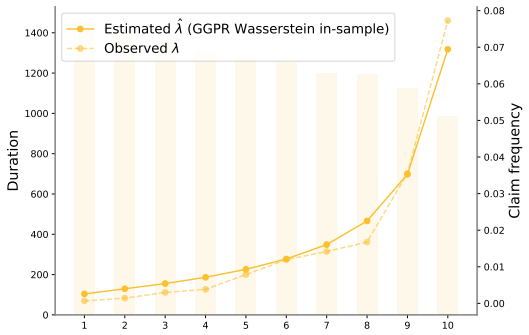
(e) Corrective GGPR in-sample



(f) Corrective unaware GGPR in-sample



(g) Corrective aware GGPR in-sample



(h) Wasserstein GGPR in-sample

Figure 2: Actual vs predicted claim frequency for unaware and aware GGPR models with and without balance correction, and for corrective GGPR models with best-estimate, unaware, or aware kernel, and post-hoc Wasserstein.

As reported in Table 4, all models have similar in-sample and out-of-sample total loss, suggesting that fairness corrections do not hurt predictive performance.

	in-sample	out-of-sample
best-estimate GGPR	1681.9403	1488.9850
best-estimate GGPR BC	1680.6705	1484.9006
unaware GGPR	1688.2966	1498.5991
unaware GGPR BC	1687.1544	1494.6337
aware GGPR	1689.7740	1499.7334
aware GGPR BC	1689.0642	1495.3120
corrective GGPR	1691.2243	1503.3934
corrective unaware GGPR	1688.5636	1498.6971
corrective aware GGPR	1680.5754	–
Wasserstein	1688.2064	1497.5140

Table 4: Poisson deviance values on the training set (in-sample) and test set (out-of-sample) across all FGGPR models.

In the in-sample setting in Table 5, the observed covariance between the predicted premiums and the sensitive attribute is 49.06, serving as a baseline. Standard predictive models, such as the GLM and LightGBM, partially reduce this covariance to 26.41 and 16.35, respectively, indicating some mitigation of the sensitive attribute's influence while leaving a significant signal. The GGPR baseline achieves a covariance of 18.74, and applying bias correction (GGPR BC) reduces it slightly further to 17.16. Both unaware and aware variants of GGPR, particularly when combined with balance correction, considerably lower the covariance. Direct fairness-corrective approaches are the most effective: the GGPR corrective model reduces covariance to 0, while the aware corrective variant effectively achieves zero covariance (-0.000001). Post-hoc adjustment using the Wasserstein distance also reduces covariance to 4.93, performing well but not as drastically as the direct corrective methods. Overall, these results highlight the effectiveness of fairness-aware strategies in controlling the influence of sensitive attributes in in-sample predictions.

Model	Empirical Covariance $dC_n\hat{\lambda}$
Observed	49.060785
GLM	26.411420
LightGBM	16.352894
GGPR	18.744371
GGPR BC	17.157110
GGPR unaware	5.243294
GGPR unaware BC	4.862886
GGPR aware	5.284778
GGPR aware BC	4.870753
GGPR corrective	0.000000
GGPR unaware corrective	0.000000
GGPR aware corrective	-0.000001
GGPR post-hoc Wasserstein	4.933428

Table 5: In-sample empirical covariance of predicted premiums across models.

Tables 8–7 report calibration by number of claims for both in-sample and out-of-sample predictions. In the training set, several GGPR variants, particularly GGPR BC and its unaware and aware counterparts, achieve near-perfect total calibration, with aggregate deviations close to zero. This is expected, as bias correction is explicitly designed to enforce calibration constraints in-sample.

Out-of-sample, however, all models exhibit systematic overestimation of the total number of claims, with deviations ranging from approximately +14 (GLM) to over +35 (GGPR aware). Fairness-aware corrective approaches do not fully eliminate this effect, although GGPR BC variants consistently show smaller deviations compared to their uncorrected counterparts.

Across both samples, models tend to overpredict policies with exactly one claim while substantially underpredicting higher claim counts. This pattern is particularly pronounced for GGPR models, reflecting regression-to-the-mean effects common in count modeling. Post-hoc Wasserstein correction improves fairness metrics but does not materially improve calibration relative to direct bias-corrected GGPR variants.

Overall, the results highlight a trade-off between fairness enforcement and calibration robustness, especially out-of-sample, underscoring the importance of jointly evaluating fairness and predictive reliability.

N	Observed	GLM	LightGBM	GGPR
1	172	183.18 (11.18)	189.49 (17.49)	201.90 (29.90)
2	18	6.55 (-11.45)	5.86 (-12.14)	5.33 (-12.67)
3	0	0.26 (0.26)	0.21 (0.21)	0.13 (0.13)
Total	190	190.00 (0.00)	195.57 (5.57)	207.36 (17.36)

Table 6: In-sample calibration by number of claims (baseline models). Values in parentheses denote deviations from the observed counts.

N	GGPR BC	GGPR unaware	GGPR unaware BC	GGPR aware	GGPR aware BC	GGPR corrective	GGPR post-hoc W.
1	185.39 (13.39)	201.11 (29.11)	185.45 (13.45)	203.64 (31.64)	185.45 (13.45)	198.30 (26.30)	201.72 (29.72)
2	4.50 (-13.50)	5.22 (-12.78)	4.44 (-13.56)	5.32 (-12.68)	4.43 (-13.57)	5.15 (-12.85)	5.20 (-12.80)
3	0.10 (0.10)	0.12 (0.12)	0.10 (0.10)	0.13 (0.13)	0.10 (0.10)	0.14 (0.14)	0.12 (0.12)
Total	189.99 (-0.01)	206.45 (16.45)	189.99 (-0.01)	209.10 (19.10)	189.98 (-0.02)	203.60 (13.60)	207.04 (17.04)

Table 7: In-sample calibration by number of claims (fairness-aware GGPR variants). Values in parentheses denote deviations from the observed counts.

N	Observed	GLM	LightGBM	GGPR
1	156	175.42 (19.42)	181.72 (25.72)	195.24 (39.24)
2	12	6.07 (-5.93)	5.48 (-6.52)	5.18 (-6.82)
3	0	0.23 (0.23)	0.18 (0.18)	0.12 (0.12)
Total	168	181.73 (13.73)	187.38 (19.38)	200.55 (32.55)

Table 8: Out-of-sample calibration by number of claims (baseline models). Values in parentheses denote deviations from the observed counts.

N	GGPR BC	GGPR unaware	GGPR unaware BC	GGPR aware	GGPR aware BC	GGPR corrective	GGPR post-hoc W.
1	179.29 (23.29)	195.20 (39.20)	180.02 (24.02)	197.67 (41.67)	180.18 (24.18)	192.73 (36.73)	195.13 (39.13)
2	4.37 (-7.63)	5.13 (-6.87)	4.37 (-7.63)	5.24 (-6.76)	4.37 (-7.63)	5.11 (-6.89)	5.08 (-6.92)
3	0.10 (0.10)	0.12 (0.12)	0.09 (0.09)	0.12 (0.12)	0.09 (0.09)	0.14 (0.14)	0.12 (0.12)
Total	183.76 (15.76)	200.45 (32.45)	184.48 (16.48)	203.03 (35.03)	184.65 (16.65)	197.98 (29.98)	200.33 (32.33)

Table 9: Out-of-sample calibration by number of claims (fairness-aware GGPR variants). Values in parentheses denote deviations from the observed counts.

Tables 10 and 11 report out-of-sample calibration results by number of claims, disaggregated by gender. Across all models, calibration errors are substantially larger for female policyholders, reflecting the smaller sample size and higher relative variability in this subgroup.

Standard predictive models (GLM, LightGBM, GGPR) consistently overpredict one-claim policies for both genders, with markedly larger deviations for females. This imbalance propagates to higher claim counts, where underprediction is systematic, particularly for male policyholders.

Bias-corrected and fairness-aware GGPR variants reduce aggregate calibration error for males but often exacerbate overprediction for females, especially in the one-claim category. Corrective approaches improve fairness metrics but do not uniformly improve calibration across genders, illustrating a tension between demographic parity constraints and distributional calibration.

Overall, these results emphasize the importance of evaluating calibration jointly with fairness metrics at the subgroup level, as aggregate calibration may conceal substantial gender-specific discrepancies.

Table 12 reports mean predicted claim frequencies by gender, both out-of-sample and in-sample:

$$\mu_0 := \mathbb{E}[\hat{Y} \mid S = 0], \qquad \mu_1 := \mathbb{E}[\hat{Y} \mid S = 1], \tag{70}$$

and their difference:

$$\Delta_\mu := \mu_1 - \mu_0. \tag{71}$$

N	Observed	GLM		LightGBM		GGPR	
		Female	Male	Female	Male	Female	Male
1	18 / 154	19.62 (1.62)	163.56 (9.56)	26.05 (8.05)	163.44 (9.44)	26.68 (8.68)	175.22 (21.22)
2	2 / 16	0.37 (-1.63)	6.18 (-9.82)	0.51 (-1.49)	5.35 (-10.65)	0.44 (-1.56)	4.90 (-11.10)
3	0	0.008 (0.008)	0.255 (0.255)	0.011 (0.011)	0.195 (0.195)	0.006 (0.006)	0.124 (0.124)
Total	20 / 170	20 (0)	170 (0)	26.58 (6.58)	168.99 (-1.01)	27.12 (7.12)	180.24 (10.24)

Table 10: In-sample calibration by number of claims and gender (baseline models). Female values are shown in blue, male values in red. Deviations from observed counts are shown in parentheses.

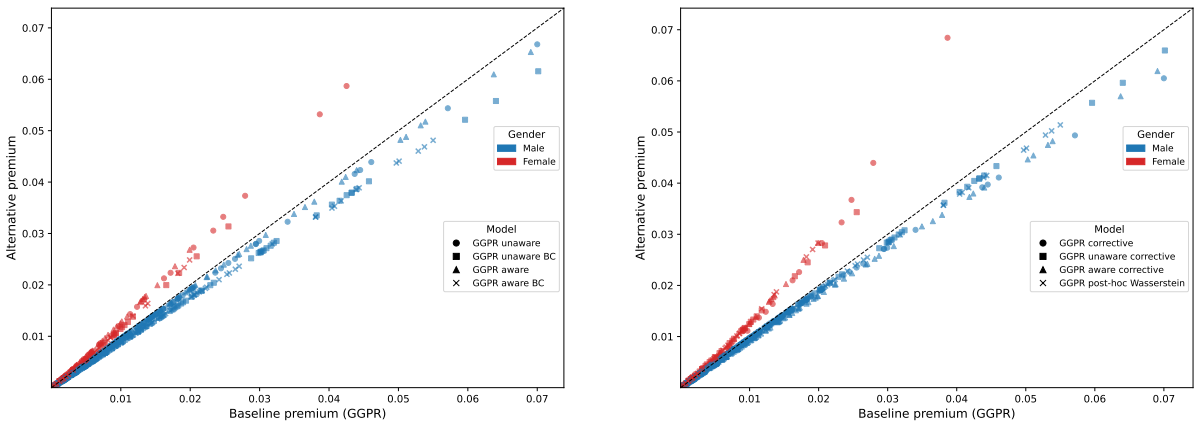
N	GGPR BC		GGPR unaware		GGPR aware		GGPR corrective	
	Female	Male	Female	Male	Female	Male	Female	Male
1	24.48 (6.48)	160.90 (6.90)	33.75 (15.75)	167.36 (13.36)	34.15 (16.15)	169.49 (15.49)	34.70 (16.70)	163.60 (9.60)
2	0.37 (-1.63)	4.13 (-11.87)	0.76 (-1.24)	4.46 (-11.54)	0.77 (-1.23)	4.55 (-11.45)	1.21 (-0.79)	3.94 (-12.06)
3	0.005 (0.005)	0.096 (0.096)	0.015 (0.015)	0.109 (0.109)	0.015 (0.015)	0.111 (0.111)	0.063 (0.063)	0.082 (0.082)
Total	24.86 (4.86)	165.13 (-4.87)	34.52 (14.52)	171.93 (1.93)	34.94 (14.94)	174.16 (4.16)	35.97 (15.97)	167.62 (-2.38)

Table 11: In-sample calibration by number of claims and gender (fairness-aware GGPR variants). Female values are shown in blue, male values in red.

A zero mean difference $\Delta_\mu = 0$ corresponds to demographic parity in expectation. Classical models such as the GLM and LightGBM exhibit substantial gender gaps, with male predictions exceeding female predictions by approximately 0.002 to 0.005 claims on average. Similar patterns are observed for the baseline GGPR, indicating that gender-related disparities persist even under flexible non-linear modeling.

Unaware and aware GGPR variants substantially reduce the male-female gap, both in-sample and out-of-sample, with differences close to zero. Bias correction further stabilizes these gaps, suggesting improved alignment with demographic parity constraints. The smallest gaps are achieved by the corrective GGPR variants, which occasionally slightly reverse the direction of the disparity, reflecting numerical enforcement of fairness constraints rather than residual sampling variation.

Overall, these results are consistent with the empirical covariance analysis: models explicitly designed to enforce fairness succeed in aligning average predictions across genders, while standard predictive models retain systematic gender differences. Importantly, these effects persist out-of-sample, indicating that fairness corrections generalize beyond the training data.



(a) Unaware and aware GGPR variants.

(b) Corrective and post-hoc GGPR variants.

Figure 3: Training-set comparison of fair premiums against the discriminatory GGPR baseline. Each point represents a policyholder, coloured by gender, with the dashed line indicating the identity. Panel (a) shows kernel fairness constraints (unaware and aware models), while panel (b) reports corrective approaches, including the post-hoc Wasserstein adjustment.

	In-sample			Out-of-sample		
Model	Female	Male	Gap	Female	Male	Gap
GLM	0.00569	0.01031	0.00463	0.00609	0.01024	0.00415
LightGBM	0.00756	0.01025	0.00270	0.00784	0.01022	0.00238
GGPR	0.00771	0.01094	0.00322	0.00815	0.01099	0.00283
GGPR BC	0.00707	0.01002	0.00295	0.00747	0.01007	0.00259
GGPR unaware	0.00982	0.01043	0.00062	0.01045	0.01048	0.00003
GGPR unaware BC	0.00903	0.00960	0.00057	0.00961	0.00965	0.00004
GGPR aware	0.00993	0.01057	0.00063	0.01061	0.01061	−0.00001
GGPR aware BC	0.00904	0.00960	0.00056	0.00964	0.00965	0.00002
GGPR corrective	0.01023	0.01017	−0.00006	0.01103	0.01020	−0.00084
GGPR unaware corrective	0.00985	0.01038	0.00054	0.01046	0.01043	−0.00003
GGPR aware corrective	0.01022	0.00986	−0.00036	–	–	–
GGPR post-hoc Wasserstein	0.01018	0.01039	0.00021	0.01034	0.01050	0.00016

Table 12: Mean predicted claim frequency by gender for all models. The gap is defined as the difference between male and female mean predictions. In-sample results correspond to the training set, while out-of-sample results are computed on the test set.

6. Discussion

This work proposes a framework for multi-criteria fair Generalized Gaussian Process Regression (GGPR) adapted to count data with heterogeneous risk exposure, as commonly encountered in insurance pricing and frequency modeling. By embedding several fairness notions directly into the probabilistic learning architecture, the proposed approach moves beyond single-criterion corrections and post-hoc adjustments.

A key contribution lies in the adaptation of fairness constraints to the generalized GP setting, where the response variable follows a Poisson distribution and the risk exposure enters through the canonical log-link. Fairness is enforced on the latent intensity function, rather than on observed counts. This modeling choice is essential. Enforcing fairness directly on observed claim counts would mechanically penalize policyholders with longer exposure periods, thereby conflating exposure heterogeneity with discriminatory effects. By imposing fairness constraints on the latent intensity function, the model preserves the proportionality between expected claim counts and exposure that is intrinsic to Poisson frequency models.

Within this framework, multiple fairness criteria are formulated as constraints on functionals of the posterior predictive distribution. This allows them to be combined within a single objective, yielding a multi-criteria fair GGPR that trades off predictive accuracy, smoothness, and several fairness objectives simultaneously. These constraints are enforced during training, rather than applied ex post, which ensures internal consistency of predictions.

The empirical results confirm several key findings. First, fairness constraints can be enforced without overly affecting predictive performance. Poisson deviance remains comparable across baseline, fairness-aware, and corrective GGPR variants both in-sample and out-of-sample. Second, unaware and aware kernels reduce the empirical covariance between predicted premiums and gender relative to standard models, while corrective GGPR variants achieve near-zero covariance by construction. These effects persist out-of-sample, indicating that fairness corrections generalize beyond the training data. Moreover, while balance-corrected GGPR variants achieve near-perfect in-sample calibration, all models exhibit systematic overprediction out-of-sample, with errors disproportionately affecting the smaller female subgroup.

From an actuarial perspective, the proposed framework offers several advantages. First, fairness corrections are interpretable as controlled deviations from the best-estimate posterior. Second, each fairness criterion corresponds to a clearly defined mathematical object (e.g., covariance with a sensitive attribute, or distributional alignment across groups). Moreover, the ability to combine multiple fairness criteria and sensitive attributes within a single probabilistic model reflects the multi-dimensional nature of discrimination in real-world pricing systems.

Overall, this article demonstrates that fairness can be embedded directly into the

probabilistic learning mechanism. By jointly accounting for count data, exposure, and multiple fairness objectives within a single model, this framework provides a flexible foundation for fair insurance pricing. Future research may extend this framework to severity modeling, and dynamic portfolio settings.

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.

References

Ang, Z. Q., & Lee, S. K. (2022). Hierarchical bayesian gaussian process regression model for loss reserving using combinations of squared exponential kernels. *Insurance: Mathematics and Economics*, 105, 54–63. https://doi.org/10.1016/j.insmatheco.2022.03.008

Barocas, Solon; Selbst, Andrew D. (2016). Big data's disparate impact. https://doi.org/10.15779/Z38BG31

Barry, L., & Charpentier, A. (2020). Personalization as a promise: Can big data change the practice of insurance? *Big Data and Society*, 7(1), 205395172093514. https://doi.org/10.1177/2053951720935143

Bjerre, D. S. (2022). Tree-based machine learning methods for modeling and forecasting mortality. *ASTIN Bulletin*, 52(3), 765–787. https://doi.org/10.1017/asb.2022.11

Chan, A. B., & Dong, D. (2011). Generalized gaussian process models. *CVPR 2011*, 2681–2688. https://doi.org/10.1109/cvpr.2011.5995688

Chevalier, D., & Côté, M.-P. (2025). From point to probabilistic gradient boosting for claim frequency and severity prediction. *European Actuarial Journal*, 15(3), 707–752. https://doi.org/10.1007/s13385-025-00428-5

Cornacchia, G., Anelli, V. W., Biancofiore, G. M., Narducci, F., Pomo, C., Ragone, A., & Di Sciascio, E. (2023). Auditing fairness under unawareness through counterfactual reasoning. *Information Processing and Management*, 60(2), 103224. https://doi.org/10.1016/j.ipm.2022.103224

Côté, M.-P., Côté, O., & Charpentier, A. (2024). Selection bias in insurance: Why portfolio-specific fairness fails to extend market-wide. https://doi.org/10.2139/ssrn.5018749

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*.

Denuit, M., Hainaut, D., & Trufin, J. (2019a). *Effective statistical learning methods for actuaries i: Glms and extensions*. Springer International Publishing. https://doi.org/10.1007/978-3-030-25820-7

Denuit, M., Hainaut, D., & Trufin, J. (2019b). Generalized additive models (gams). In *Effective statistical learning methods for actuaries i: Glms and extensions* (pp. 253–327). Springer.

Denuit, M., Hainaut, D., & Trufin, J. (2019c). Generalized linear models (glms). *Effective Statistical Learning Methods for Actuaries I: GLMs and Extensions*, 97–196.

Denuit, M., Hainaut, D., & Trufin, J. (2020). Bagging trees and random forests. In *Effective statistical learning methods for actuaries ii: Tree-based methods and extensions* (pp. 107–130). Springer.

Dolman, C., Frees, E. (, & Huang, F. (2021). Multidisciplinary collaboration on discrimination – not just "nice to have". *Annals of Actuarial Science*, 15(3), 485–487. https://doi.org/10.1017/s174849952100021x

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

Hainaut, D. (2026). American option pricing with model constrained gaussian process regressions. *Applied Mathematics and Computation*, 512, 129747. https://doi.org/10.1016/j.amc.2025.129747

Hainaut, D., & Denuit, M. (2025). *Insurance risk classification with generalized gaussian process regression models* (tech. rep.). https://dial.uclouvain.be/pr/boreal/object/boreal%3A299662/datastream/PDF_01/view

Hainaut, D., & Vrins, F. D. (2024). European option pricing with model constrained gaussian process regressions. https://doi.org/10.2139/ssrn.4982344

Henckaerts, R., Côté, M.-P., Antonio, K., & Verbelen, R. (2020). Boosting insights in insurance tariff plans with tree-based machine learning methods. *North American Actuarial Journal*, 25(2), 255–285. https://doi.org/10.1080/10920277.2020.1745656

Hiabu, M. (2016). On the relationship between classical chain ladder and granular reserving. *Scandinavian Actuarial Journal*, 2017(8), 708–729. https://doi.org/10.1080/03461238.2016.1240709

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.

Jamotton, C., & Hainaut, D. (2025). *A multivariate energy-based fairness adjuster for premiums* (tech. rep.). https://dial.uclouvain.be/pr/boreal/object/boreal:300678

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23*, 35–50.

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). Discrimination-free insurance pricing. *ASTIN Bulletin: The Journal of the IAA*, 52(1), 55–89.

Lindholm, M., Richman, R., Tsanakas, A., & Wüthrich, M. V. (2024a). A multi-task network approach for calculating discrimination-free insurance prices. *European Actuarial Journal*, 14(2), 329–369.

Lindholm, M., Richman, R., Tsanakas, A., & Wüthrich, M. V. (2024b). Sensitivity-based measures of discrimination in insurance pricing. https://ssrn.com/abstract=4897265

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

Lindholm, M., & Wüthrich, M. V. (2025). The balance property in insurance pricing. *Scandinavian Actuarial Journal*, 1–38.

Lloyd, C., Gunter, T., Osborne, M., & Roberts, S. (2015). Variational inference for gaussian process modulated poisson processes. In F. Bach & D. Blei (Eds.), *Proceedings of the 32nd international conference on machine learning* (pp. 1814–1822, Vol. 37). PMLR. https://doi.org/10.48550/ARXIV.1411.0254

Ludkovski, M., Risk, J., & Zail, H. (2018). Gaussian process models for mortality rates and improvement factors. *ASTIN Bulletin*, 48(3), 1307–1347. https://doi.org/10.1017/asb.2018.24

Pearl, J. (2009). *Causality: Models, reasoning and inference* (2nd). Cambridge University Press.

Plecko, D., & Bareinboim, E. (2024). Causal fairness analysis: A causal toolkit for fair machine learning. *Foundations and Trends in Machine Learning*, 17(3), 304–589. https://doi.org/10.1561/2200000106

Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 54(3), 507–554. https://doi.org/10.1111/j.1467-9876.2005.00510.x

Risk, J., & Ludkovski, M. (2024). Expressive mortality models through gaussian process kernels. *ASTIN Bulletin*, 54(2), 327–359. https://doi.org/10.1017/asb.2023.39

Shim, J., & Hwang, C. (2011). Kernel poisson regression machine for stochastic claims reserving. *Journal of the Korean Statistical Society*, 40(1), 1–9. https://doi.org/10.1016/j.jkss.2010.01.004

Taha, A. A., & Malebary, S. J. (2020). An intelligent approach to credit card fraud detection using an optimized light gradient boosting machine. *IEEE Access*, 8, 25579–25587. https://doi.org/10.1109/access.2020.2971354

Wang, B., & Shi, J. Q. (2014). Generalized gaussian process regression model for non-gaussian functional data. *Journal of the American Statistical Association*, 109(507), 1123–1133. https://doi.org/10.1080/01621459.2014.889021

Williams, C. K., & Rasmussen, C. E. (2006). *Gaussian processes for machine learning* (Vol. 2). MIT press Cambridge, MA.

Wuthrich, M. V., & Merz, M. (2021). Statistical foundations of actuarial learning and its applications. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3822407

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

Yang, Y., Qian, W., & Zou, H. (2017). Insurance premium prediction via gradient tree-boosted tweedie compound poisson models. *Journal of Business and Economic Statistics*, 36(3), 456–470. https://doi.org/10.1080/07350015.2016.1200981

A. Proofs

A.1 GGPR

Proof 1 (Proof of Proposition 3.1). Recall that a density in the ED family admits the representation

$$p(y \mid \boldsymbol{X}) = \exp\left(\frac{y\theta(\boldsymbol{X}) - \gamma(\theta(\boldsymbol{X}))}{\phi/\nu} + c(y, \phi, \nu)\right), \tag{72}$$

where θ is the canonical parameter, γ the cumulant function, ϕ the dispersion parameter, and c a normalizing term independent of θ.

For the Poisson model, the dispersion parameter is $\phi = 1$ and the canonical link is logarithmic. Define:

$$\theta(\boldsymbol{X}) := \ln \lambda(\boldsymbol{X}). \tag{73}$$

The cumulant function corresponding to the Poisson exponential family is:

$$\gamma(\theta) = \exp(\theta), \qquad \gamma'(\theta) = \exp(\theta), \qquad \gamma'^{-1}(\mu) = \ln \mu. \tag{74}$$

Substituting $\theta(\boldsymbol{X}) = \ln \lambda(\boldsymbol{X})$ and $\phi = 1$ into the ED form yields:

$$\frac{y\theta(\boldsymbol{X}) - \gamma(\theta(\boldsymbol{X}))}{\phi/\nu} = \nu\left(y \ln \lambda(\boldsymbol{X}) - \lambda(\boldsymbol{X})\right) \tag{75}$$

$$= \nu y \ln \lambda(\boldsymbol{X}) - \nu\lambda(\boldsymbol{X}). \tag{76}$$

Choosing the normalizing term

$$c(y, \phi, \nu) = -\ln((\nu y)!) + \nu y \ln \nu, \tag{77}$$

we obtain:

$$p(y \mid \boldsymbol{X}) = \exp(\nu y \ln \lambda(\boldsymbol{X}) - \nu\lambda(\boldsymbol{X}) - \ln((\nu y)!) + \nu y \ln \nu) \tag{78}$$

$$= \exp(-\nu\lambda(\boldsymbol{X}))\, \frac{(\nu\lambda(\boldsymbol{X}))^{\nu y}}{(\nu y)!}. \tag{79}$$

□

Proof 2 (Proof of Proposition 3.3). Let $\boldsymbol{X} \in \mathbb{R}^{n \times m}$ be the training set and let Φ be a (design) kernel matrix:

$$\Phi = \left[\phi\left(\boldsymbol{x}_1\right), \ldots, \phi\left(\boldsymbol{x}_n\right)\right]^\top \in \mathbb{R}^{n \times m}. \tag{80}$$

The Bayesian linear regression model can be expressed as:

$$\boldsymbol{g} = \Phi\boldsymbol{w} + \epsilon, \quad \epsilon \sim \mathcal{N}\left(0, \sigma_n^2 \boldsymbol{I}_n\right), \tag{81}$$

with a Gaussian prior on the weights:

$$\boldsymbol{w} \sim \mathcal{N}\left(\mathbf{0}, \Sigma_m\right), \tag{82}$$

Conditioned on $\boldsymbol{w}$, the likelihood is Gaussian,

$$\boldsymbol{g} \mid \boldsymbol{X}, \boldsymbol{w} \sim \mathcal{N}(\Phi \boldsymbol{w}, \sigma_n^2 \boldsymbol{I}_n). \tag{83}$$

The marginal mean of $\boldsymbol{g} \mid \boldsymbol{X}$ is obtained by iterated expectation:

$$\begin{aligned} \mathbb{E}[\boldsymbol{g} \mid \boldsymbol{X}] &= \mathbb{E}_{\boldsymbol{w}}[\mathbb{E}[\boldsymbol{g} \mid \boldsymbol{X}, \boldsymbol{w}]] && (84) \\ &= \mathbb{E}_{\boldsymbol{w}}[\Phi \boldsymbol{w}] = \Phi \mathbb{E}[\boldsymbol{w}] = \mathbf{0}. && (85) \end{aligned}$$

The marginal covariance follows from the law of total covariance:

$$\begin{aligned} \operatorname{Cov}[\boldsymbol{g} \mid \boldsymbol{X}] &= \operatorname{Cov}_{\boldsymbol{w}}[\mathbb{E}[\boldsymbol{g} \mid \boldsymbol{X}, \boldsymbol{w}]] + \mathbb{E}_{\boldsymbol{w}}[\operatorname{Cov}[\boldsymbol{g} \mid \boldsymbol{X}, \boldsymbol{w}]] && (86) \\ &= \operatorname{Cov}_{\boldsymbol{w}}[\Phi \boldsymbol{w}] + \mathbb{E}_{\boldsymbol{w}}[\sigma_n^2 \boldsymbol{I}_n] && (87) \\ &= \Phi \operatorname{Cov}[\boldsymbol{w}] \Phi^\top + \sigma_n^2 \boldsymbol{I}_n && (88) \\ &= \Phi \Sigma_m \Phi^\top + \sigma_n^2 \boldsymbol{I}_n && (89) \\ &= K + \sigma_n^2 \boldsymbol{I}_n && (90) \\ &= C\left(\boldsymbol{X}, \boldsymbol{X}\right). && (91) \end{aligned}$$

Since the marginal mean and covariance fully characterize a Gaussian distribution, it follows that:

$$\boldsymbol{g} \mid \boldsymbol{X} \sim \mathcal{N}\Big(\mathbf{0}, K + \sigma_n^2 \boldsymbol{I}_n\Big), \quad K := \Phi \Sigma_m \Phi^\top, \tag{92}$$

or equivalently,

$$\boldsymbol{g} \mid \boldsymbol{X} \sim \mathcal{N}(\mathbf{0},\ C\left(\boldsymbol{X}, \boldsymbol{X}\right)). \tag{93}$$

□

Proposition A.1 (Predictive distribution in kernel form)**.** *Let $k(\boldsymbol{x}_*, \boldsymbol{X}) = \phi(\boldsymbol{x}_*)^\top \Sigma_c \Phi^\top$ denote the cross-kernel between a new input $\boldsymbol{x}_*$ and the training data, and let $K = \Phi \Sigma_c \Phi^\top$. Then the predictive distribution for a new output g_* is*

$$g_* \mid \boldsymbol{x}_*, \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{g} \sim \mathcal{N}(\mu_*, \sigma_*^2), \tag{94}$$

with

$$\begin{aligned} \mu_* &= k(\boldsymbol{x}_*, \boldsymbol{X})(\sigma_n^2 \boldsymbol{I}_n + K)^{-1} \boldsymbol{g}, && (95) \\ \sigma_*^2 &= k(\boldsymbol{x}_*, \boldsymbol{x}_*) - k(\boldsymbol{x}_*, \boldsymbol{X})(\sigma_n^2 \boldsymbol{I}_n + K)^{-1} k(\boldsymbol{X}, \boldsymbol{x}_*). && (96) \end{aligned}$$

Proof 3 (Proof of Proposition A.1). Starting from the posterior predictive mean in feature

space (see Proposition A.2):

$$\phi\left(\boldsymbol{x}_*\right)^\top \bar{\boldsymbol{w}} = \sigma_n^{-2} \phi\left(\boldsymbol{x}_*\right)^\top A^{-1} \Phi^\top \boldsymbol{g} \tag{97}$$

Applying the Woodbury identity:

$$A^{-1} = \left(\sigma_n^{-2} \Phi^\top \Phi + \Sigma_m^{-1}\right)^{-1} = \Sigma_m - \Sigma_m \Phi^\top \left(\sigma_n^2 \boldsymbol{I}_n + \Phi \Sigma_m \Phi^\top\right)^{-1} \Phi \Sigma_m, \tag{98}$$

yields:

$$\begin{aligned}
\phi\left(\boldsymbol{x}_*\right)^\top \bar{\boldsymbol{w}} &= \sigma_n^{-2} \phi\left(\boldsymbol{x}_*\right)^\top A^{-1} \Phi^\top \boldsymbol{g} \\
&= \sigma_n^{-2} \phi\left(\boldsymbol{x}_*\right)^\top \left(\Sigma_m - \Sigma_m \Phi^\top \left(\sigma_n^2 \boldsymbol{I}_n + \Phi \Sigma_m \Phi^\top\right)^{-1} \Phi \Sigma_m\right) \Phi^\top \boldsymbol{g} \\
&= \sigma_n^{-2} \phi(\boldsymbol{x}_*)^\top \Sigma_m \Phi^\top \boldsymbol{g} - \sigma_n^{-2} \phi\left(\boldsymbol{x}_*\right)^\top \Sigma_m \Phi^\top \left(\sigma_n^2 \boldsymbol{I}_n + \Phi \Sigma_m \Phi^\top\right)^{-1} \Phi \Sigma_m \Phi^\top \boldsymbol{g}.
\end{aligned}$$

We introduce the $1 \times n$ cross-kernel k and $n \times n$ kernel K matrices:

$$k(\boldsymbol{x}_*, \boldsymbol{X}) = \phi(\boldsymbol{x}_*)^\top \Sigma_m \Phi^\top, \quad K = \Phi \Sigma_m \Phi^\top. \tag{99}$$

The predictive mean becomes:

$$\phi(\boldsymbol{x}_*)^\top \bar{\boldsymbol{w}} = \sigma_n^{-2} k(\boldsymbol{x}_*, \boldsymbol{X}) \boldsymbol{g} - \sigma_n^{-2} k(\boldsymbol{x}_*, \boldsymbol{X}) \left(\sigma_n^2 \boldsymbol{I}_n + K\right)^{-1} K \boldsymbol{g}.$$

We factor out the cross-kernel term $k(\boldsymbol{x}_*, \boldsymbol{X})$:

$$\phi(\boldsymbol{x}_*)^\top \bar{\boldsymbol{w}} = k(\boldsymbol{x}_*, \boldsymbol{X}) \sigma_n^{-2} \Big[\boldsymbol{I}_n - \left(\sigma_n^2 \boldsymbol{I}_n + K\right)^{-1} K\Big] \boldsymbol{g}. \tag{100}$$

Using the identity $\left(\sigma_n^2 \boldsymbol{I}_n + K\right)^{-1} \left(\sigma_n^2 \boldsymbol{I}_n + K\right) = \boldsymbol{I}_n$, we simplify the brackets:

$$\begin{aligned}
\boldsymbol{I}_n - \left(\sigma_n^2 \boldsymbol{I}_n + K\right)^{-1} K &= \left(\sigma_n^2 \boldsymbol{I}_n + K\right)^{-1} \left(\sigma_n^2 \boldsymbol{I}_n + K - K\right) && (101) \\
&= \left(\sigma_n^2 \boldsymbol{I}_n + K\right)^{-1} \sigma_n^2 \boldsymbol{I}_n && (102) \\
&= \sigma_n^2 \left(\sigma_n^2 \boldsymbol{I}_n + K\right)^{-1}. && (103)
\end{aligned}$$

We get the final predictive mean in kernel form:

$$\phi(\boldsymbol{x}_*)^\top \bar{\boldsymbol{w}} = k(\boldsymbol{x}_*, \boldsymbol{X}) \left(\sigma_n^2 \boldsymbol{I}_n + K\right)^{-1} \boldsymbol{g}. \tag{104}$$

This is the standard Bayesian linear regression predictive mean expressed in terms of kernels.

Similarly, the predictive covariance $\phi(\boldsymbol{x}_*)^\top A^{-1}\phi(\boldsymbol{x}_*)$ is equal to:

$$\begin{aligned}\phi(\boldsymbol{x}_*)^\top A^{-1}\phi(\boldsymbol{x}_*) &= \phi(\boldsymbol{x}_*)^\top \left(\Sigma_m - \Sigma_m\Phi^\top\left(\sigma_n^2\boldsymbol{I}_n + \Phi\Sigma_m\Phi^\top\right)^{-1}\Phi\Sigma_m\right)\phi(\boldsymbol{x}_*)\\ &= \phi(\boldsymbol{x}_*)^\top\Sigma_m\phi(\boldsymbol{x}_*) - \phi(\boldsymbol{x}_*)^\top\Sigma_m\Phi^\top(\sigma_n^2\boldsymbol{I}_n + \Phi\Sigma_m\Phi^\top)^{-1}\Phi\Sigma_m\phi(\boldsymbol{x}_*)\\ &= k(\boldsymbol{x}_*,\boldsymbol{x}_*) - k(\boldsymbol{x}_*,\boldsymbol{X})(\sigma_n^2\boldsymbol{I}_n + K)^{-1}k(\boldsymbol{X},\boldsymbol{x}_*),\end{aligned}$$

where $k(\boldsymbol{X},\boldsymbol{x}_*) = k(\boldsymbol{x}_*,\boldsymbol{X})^\top$. The final predictive covariance in kernel form is equal to:

$$\phi(\boldsymbol{x}_*)^\top A^{-1}\phi(\boldsymbol{x}_*) = k(\boldsymbol{x}_*,\boldsymbol{x}_*) - k(\boldsymbol{x}_*,\boldsymbol{X})(\sigma_n^2\boldsymbol{I}_n + K)^{-1}k(\boldsymbol{X},\boldsymbol{x}_*). \tag{105}$$

This is the Bayesian linear regression predictive variance in kernel form. The predictive distribution of scalar g_* for a new policyholder $\boldsymbol{x}_*$ in Gaussian Process is formally defined as:

$$p(g_* \mid \boldsymbol{x}_*,\boldsymbol{X},\boldsymbol{y},\boldsymbol{g}) = \mathcal{N}(k(\boldsymbol{x}_*,\boldsymbol{X})(\sigma_n^2\boldsymbol{I}_n + K)^{-1}\boldsymbol{g}, \tag{106}$$

$$k(\boldsymbol{x}_*,\boldsymbol{x}_*) - k(\boldsymbol{x}_*,\boldsymbol{X})(\sigma_n^2\boldsymbol{I}_n + K)^{-1}k(\boldsymbol{X},\boldsymbol{x}_*)). \tag{107}$$

□

Proof 4 (Proof of Proposition 3.4). Let $\boldsymbol{g} = (g(\boldsymbol{x}_1),\ldots,g(\boldsymbol{x}_n))^\top$ denote the latent function values at the training inputs $\boldsymbol{X} = (\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$. Under the Gaussian process prior, the joint distribution of $(\boldsymbol{g}, g_*)$ is:

$$\begin{pmatrix}\boldsymbol{g}\\ g_*\end{pmatrix} \sim \mathcal{N}\left(\boldsymbol{0}, \begin{pmatrix}C(\boldsymbol{X},\boldsymbol{X}) & k_*^\top\\ k_* & k_{**}\end{pmatrix}\right), \tag{108}$$

where $k_* = k(\boldsymbol{x}_*,\boldsymbol{X})$ and $k_{**} = k(\boldsymbol{x}_*,\boldsymbol{x}_*)$.

For non-Gaussian likelihoods, the posterior distribution of $\boldsymbol{g}$ is analytically intractable. Under the Laplace approximation, the posterior is approximated by a Gaussian:

$$q(\boldsymbol{g} \mid \boldsymbol{y},\boldsymbol{X}) \approx \mathcal{N}(\hat{\boldsymbol{g}}, \Sigma_{\boldsymbol{g}}), \tag{109}$$

where $\hat{\boldsymbol{g}}$ is the posterior mode and:

$$\Sigma_{\boldsymbol{g}} = \left(C(\boldsymbol{X},\boldsymbol{X})^{-1} + H(\hat{\boldsymbol{g}})\right)^{-1}, \tag{110}$$

with $H(\hat{\boldsymbol{g}})$ denoting the Hessian matrix of the negative log-likelihood evaluated at $\hat{\boldsymbol{g}}$. The predictive distribution of g_* is obtained by marginalising over the approximate posterior $q(\boldsymbol{g} \mid \boldsymbol{y},\boldsymbol{X})$:

$$p(g_* \mid \boldsymbol{y},\boldsymbol{X}) = \int p(g_* \mid \boldsymbol{g})\, q(\boldsymbol{g} \mid \boldsymbol{y},\boldsymbol{X})\, d\boldsymbol{g}. \tag{111}$$

From the joint Gaussian prior, the conditional distribution of g_* given $\boldsymbol{g}$ is (see Proposi-

tion A.1):

$$p(g_* \mid \boldsymbol{g}) = \mathcal{N}\big(k_* C(\boldsymbol{X}, \boldsymbol{X})^{-1} \boldsymbol{g},\ k_{**} - k_* C(\boldsymbol{X}, \boldsymbol{X})^{-1} k_*^\top\big)\,. \tag{112}$$

The predictive mean is therefore given by:

$$\begin{aligned} m_* &= \mathbb{E}_{q(\boldsymbol{g}\mid\boldsymbol{y},\boldsymbol{X})}\left[\mathbb{E}(g_* \mid \boldsymbol{g})\right] \\ &= \mathbb{E}_{q(\boldsymbol{g}\mid\boldsymbol{y},\boldsymbol{X})}\left[k_* C(\boldsymbol{X}, \boldsymbol{X})^{-1} \boldsymbol{g}\right] \\ &= k_* C(\boldsymbol{X}, \boldsymbol{X})^{-1} \mathbb{E}_{q(\boldsymbol{g}\mid\boldsymbol{y},\boldsymbol{X})}[\boldsymbol{g}] \\ &= k_* C(\boldsymbol{X}, \boldsymbol{X})^{-1} \hat{\boldsymbol{g}}. \end{aligned} \tag{113}$$

Using the law of total variance:

$$s_*^2 = \mathbb{E}_{q(\boldsymbol{g}\mid\boldsymbol{y},\boldsymbol{X})}[\mathrm{Var}(g_* \mid \boldsymbol{g})] + \mathrm{Var}_{q(\boldsymbol{g}\mid\boldsymbol{y},\boldsymbol{X})}[\mathbb{E}(g_* \mid \boldsymbol{g})]\,. \tag{114}$$

The first term is constant with respect to $\boldsymbol{g}$ and equals:

$$\mathbb{E}_{q(\boldsymbol{g}\mid\boldsymbol{y},\boldsymbol{X})}[\mathrm{Var}(g_* \mid \boldsymbol{g})] = k_{**} - k_* C(\boldsymbol{X}, \boldsymbol{X})^{-1} k_*^\top. \tag{115}$$

The second term is:

$$\begin{aligned} \mathrm{Var}_{q(\boldsymbol{g}\mid\boldsymbol{y},\boldsymbol{X})}[\mathbb{E}(g_* \mid \boldsymbol{g})] &= \mathrm{Var}_{q(\boldsymbol{g}\mid\boldsymbol{y},\boldsymbol{X})}\left[k_* C(\boldsymbol{X}, \boldsymbol{X})^{-1} \boldsymbol{g}\right] \\ &= k_* C(\boldsymbol{X}, \boldsymbol{X})^{-1} \Sigma_{\boldsymbol{g}} C(\boldsymbol{X}, \boldsymbol{X})^{-1} k_*^\top. \end{aligned} \tag{116}$$

Combining both terms yields:

$$s_*^2 = k_{**} - k_* C(\boldsymbol{X}, \boldsymbol{X})^{-1} k_*^\top + k_* C(\boldsymbol{X}, \boldsymbol{X})^{-1} \Sigma_{\boldsymbol{g}} C(\boldsymbol{X}, \boldsymbol{X})^{-1} k_*^\top. \tag{117}$$

Substituting $\Sigma_{\boldsymbol{g}} = (C^{-1} + H)^{-1}$ and applying the Woodbury identity gives:

$$C^{-1} - C^{-1} \Sigma_{\boldsymbol{g}} C^{-1} = \big(C + H^{-1}\big)^{-1}. \tag{118}$$

Therefore:

$$s_*^2 = k_{**} - k_*\big(C(\boldsymbol{X}, \boldsymbol{X}) + H(\hat{\boldsymbol{g}})^{-1}\big)^{-1} k_*^\top. \tag{119}$$

The predictive distribution of the latent function value at $\boldsymbol{X}_*$ is therefore:

$$g_* \mid \boldsymbol{x}_*, \boldsymbol{X}, \boldsymbol{y} \approx \mathcal{N}(m_*, s_*^2), \tag{120}$$

with m_* and s_*^2 as stated in Proposition 3.4. □

Proposition A.2 (Posterior distribution in feature space)**.** *Let $g(x) = \Phi(x)^\top \boldsymbol{w}$ be a finite-dimensional feature representation of the latent function, with design matrix $\Phi \in \mathbb{R}^{n \times m}$*

evaluated at the inputs $\boldsymbol{X} = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_n})$. *Assume the Gaussian prior:*

$$\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \Sigma_m), \tag{121}$$

and the Gaussian likelihood:

$$\boldsymbol{y} \mid \Phi, \boldsymbol{w} \sim \mathcal{N}\left(\Phi\boldsymbol{w}, \sigma_n^2 \boldsymbol{I}_n\right). \tag{122}$$

Then the posterior distribution of $\boldsymbol{w}$ *is Gaussian,*

$$\boldsymbol{w} \mid \Phi, \boldsymbol{y} \sim \mathcal{N}\left(\bar{\boldsymbol{w}}, A^{-1}\right), \tag{123}$$

with

$$A = \sigma_n^{-2}\Phi^\top\Phi + \Sigma_m^{-1}, \qquad \bar{\boldsymbol{w}} = A^{-1}\boldsymbol{b} = A^{-1}\sigma_n^{-2}\Phi^\top\boldsymbol{y}. \tag{124}$$

Proof 5 (Proof of Proposition A.2). We condition on the feature matrix $\Phi = \Phi(\boldsymbol{X})$ rather than directly on the inputs $\boldsymbol{X}$. Recall that the density of an m-dimensional multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma_m)$ is given by:

$$p(\boldsymbol{w}) = (2\pi)^{-m/2}|\Sigma_m|^{-1/2}\exp\Big(-\tfrac{1}{2}(\boldsymbol{w}-\boldsymbol{\mu})^\top\Sigma_m^{-1}(\boldsymbol{w}-\boldsymbol{\mu})\Big). \tag{125}$$

We assume a zero-mean Gaussian prior on the weights $\boldsymbol{w} \in \mathbb{R}^m$:

$$\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \Sigma_m), \tag{126}$$

with density:

$$\begin{aligned} p(\boldsymbol{w}) &= (2\pi)^{-m/2}|\Sigma_m|^{-1/2}\exp\big(-\tfrac{1}{2}\boldsymbol{w}^\top\Sigma_m^{-1}\boldsymbol{w}\big) \\ &\propto \exp\big(-\tfrac{1}{2}\boldsymbol{w}^\top\Sigma_m^{-1}\boldsymbol{w}\big). \end{aligned} \tag{127}$$

The likelihood is similarly assumed Gaussian:

$$\boldsymbol{y} \mid \Phi, \boldsymbol{w} \sim \mathcal{N}\left(\Phi\boldsymbol{w}, \sigma_n^2 \boldsymbol{I}_n\right), \tag{128}$$

with density:

$$p(\boldsymbol{y}|\Phi, \boldsymbol{w}) = (2\pi)^{-n/2}|\left(\sigma_n^2\boldsymbol{I}_n\right)|^{-1/2}\exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\Phi w})^\top\left(\sigma_n^2\boldsymbol{I}_n\right)^{-1}(\boldsymbol{y}-\boldsymbol{\Phi w})\right) \tag{129}$$

$$\propto \exp\left(-\frac{1}{2(\sigma_n^2\boldsymbol{I}_n)}(\boldsymbol{y}-\boldsymbol{\Phi w})^\top(\boldsymbol{y}-\boldsymbol{\Phi w})\right). \tag{130}$$

Ignoring terms not depending on $\boldsymbol{w}$, the posterior density is proportional to:

$$p(\boldsymbol{w} \mid \Phi, \boldsymbol{y}) \propto p\left(\boldsymbol{y} \mid \Phi, \boldsymbol{w}\right) p\left(\boldsymbol{w}\right) \tag{131}$$

$$\propto \exp\left(\left(-\frac{1}{2\sigma_n^2}(\boldsymbol{y}-\boldsymbol{\Phi w})^\top(\boldsymbol{y}-\boldsymbol{\Phi w}) - \frac{1}{2}\boldsymbol{w}^\top {\Sigma_m}^{-1}\boldsymbol{w}\right)\right). \tag{132}$$

Expanding the quadratic in the exponential, we get:

$$p(\boldsymbol{w} \mid \Phi, \boldsymbol{y}) \propto \exp\left(-\frac{1}{2\sigma_n^2}\left(\boldsymbol{y}^\top\boldsymbol{y} - 2\boldsymbol{w}^\top\Phi^\top\boldsymbol{y} + \boldsymbol{w}^\top\Phi^\top\Phi\boldsymbol{w}\right) - \frac{1}{2}\boldsymbol{w}^\top {\Sigma_m}^{-1}\boldsymbol{w}\right) \tag{133}$$

$$= \exp\left(-\frac{1}{2\sigma_n^2}\boldsymbol{y}^\top\boldsymbol{y} + \sigma_n^{-2}\boldsymbol{w}^\top\Phi^\top\boldsymbol{y} - \frac{1}{2}\boldsymbol{w}^\top\left(\sigma_n^{-2}\Phi^\top\Phi + {\Sigma_m}^{-1}\right)\boldsymbol{w}\right) \tag{134}$$

$$= \exp\left(\text{const}_y + b^\top\boldsymbol{w} - \frac{1}{2}\boldsymbol{w}^\top A\boldsymbol{w}\right), \tag{135}$$

where

$$A := \sigma_n^{-2}\Phi^\top\Phi + \Sigma_m^{-1} \in \mathbb{R}^{m\times m}, \qquad b := \sigma_n^{-2}\Phi^\top\boldsymbol{y} \in \mathbb{R}^m. \tag{136}$$

We drop const_y, since it collects terms not depending on w. The posterior is now proportional to:

$$p(\boldsymbol{w} \mid \Phi, \boldsymbol{y}) \propto \exp\left(-\frac{1}{2}\boldsymbol{w}^\top A\boldsymbol{w} + b^\top\boldsymbol{w}\right). \tag{137}$$

To identify the posterior distribution, we complete the square. For symmetric positive-definite A,

$$-\tfrac{1}{2}\boldsymbol{w}^\top A\boldsymbol{w} + \boldsymbol{b}^\top\boldsymbol{w} = -\tfrac{1}{2}(\boldsymbol{w}-A^{-1}\boldsymbol{b})^\top A(\boldsymbol{w}-A^{-1}\boldsymbol{b}) + \tfrac{1}{2}\boldsymbol{b}^\top A^{-1}\boldsymbol{b}. \tag{138}$$

Defining:

$$\bar{\boldsymbol{w}} := A^{-1}\boldsymbol{b} = (\sigma_n^{-2}\Phi^\top\Phi + \Sigma_m^{-1})^{-1}\sigma_n^{-2}\Phi^\top\boldsymbol{y}, \tag{139}$$

we obtain:

$$p(\boldsymbol{w} \mid \Phi, \boldsymbol{y}) \propto \exp\left(-\frac{1}{2}\boldsymbol{w}^\top A\boldsymbol{w} + b^\top\boldsymbol{w}\right) \tag{140}$$

$$\propto \exp\left(-\frac{1}{2}(\boldsymbol{w}-\bar{\boldsymbol{w}})^\top A\left(\boldsymbol{w}-\bar{\boldsymbol{w}}\right)\right). \tag{141}$$

Hence the posterior $\boldsymbol{w} \mid \boldsymbol{y}$ is Gaussian with covariance $A^{-1} \in \mathbb{R}^{m\times m}$ and mean $\bar{\boldsymbol{w}} \in \mathbb{R}^m$:

$$\boldsymbol{w} \mid \Phi, \boldsymbol{y} \sim \mathcal{N}\left(\bar{\boldsymbol{w}}, A^{-1}\right). \tag{142}$$

□

Proof 6 (Proof of Eq. (10)). Assume the latent vector satisfies $\boldsymbol{g} \mid \boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, C(\boldsymbol{X}, \boldsymbol{X}))$,

where $C(\boldsymbol{X}, \boldsymbol{X}) = K + \sigma_n^2 \boldsymbol{I}_n$. Its density is:

$$p(\boldsymbol{g} \mid \boldsymbol{X}) = \frac{1}{(2\pi)^{n/2}|C(\boldsymbol{X}, \boldsymbol{X})|^{1/2}} \exp\left(-\tfrac{1}{2}\boldsymbol{g}^\top C(\boldsymbol{X}, \boldsymbol{X})^{-1}\boldsymbol{g}\right). \tag{143}$$

Taking logarithms and dropping terms independent of $\boldsymbol{g}$ yields:

$$\ln p(\boldsymbol{g} \mid \boldsymbol{X}) = -\tfrac{1}{2}\boldsymbol{g}^\top C(\boldsymbol{X}, \boldsymbol{X})^{-1}\boldsymbol{g} + \text{const.} \tag{144}$$

By Bayes' rule, the posterior mode $\hat{\boldsymbol{g}}$ satisfies:

$$\hat{\boldsymbol{g}} = \arg\max_{\boldsymbol{g}} \left\{\ln p(\boldsymbol{y} \mid \boldsymbol{g}) + \ln p(\boldsymbol{g} \mid \boldsymbol{X})\right\} \tag{145}$$

$$= \arg\max_{\boldsymbol{g}} \left\{\ln p(\boldsymbol{y} \mid \boldsymbol{g}) - \tfrac{1}{2}\boldsymbol{g}^\top C(\boldsymbol{X}, \boldsymbol{X})^{-1}\boldsymbol{g}\right\}. \tag{146}$$

For a Poisson likelihood with exposure vector $\boldsymbol{\nu}$, the log-likelihood is:

$$\ln p(\boldsymbol{y} \mid \boldsymbol{g}) = \boldsymbol{\nu}^\top (\boldsymbol{y} \odot \boldsymbol{g} - \exp(\boldsymbol{g})) + \text{const.} \tag{147}$$

Defining: $\psi(\boldsymbol{g}) = \boldsymbol{\nu}^\top (\boldsymbol{y} \odot \boldsymbol{g} - \exp(\boldsymbol{g})) - \frac{1}{2}\boldsymbol{g}^\top C(\boldsymbol{X}, \boldsymbol{X})^{-1}\boldsymbol{g}$, we obtain:

$$\hat{\boldsymbol{g}} = \arg\max_{\boldsymbol{g}} \psi(\boldsymbol{g}). \tag{148}$$

□

Proof 7 (Proof of Eq. (11)). Under the Laplace approximation, the inverse covariance matrix is given by the negative Hessian of the log-posterior evaluated at the mode $\hat{\boldsymbol{g}}$:

$$\Sigma_{\boldsymbol{g}}^{-1} = -\nabla_{\boldsymbol{g}}^2 \ln p(\boldsymbol{g} \mid \boldsymbol{y}, \boldsymbol{X})\big|_{\boldsymbol{g}=\hat{\boldsymbol{g}}} \tag{149}$$

$$= -\nabla^2 \psi(\hat{\boldsymbol{g}}). \tag{150}$$

For a Poisson likelihood with exposure vector $\boldsymbol{\nu}$, the objective function ψ satisfies:

$$\psi(\hat{\boldsymbol{g}}) = \boldsymbol{\nu}^\top \exp(\hat{\boldsymbol{g}}) - \boldsymbol{\nu}^\top(\boldsymbol{y} \odot \hat{\boldsymbol{g}}) + \tfrac{1}{2}\hat{\boldsymbol{g}}^\top C(\boldsymbol{X}, \boldsymbol{X})^{-1}\hat{\boldsymbol{g}}. \tag{151}$$

Taking the second derivative yields:

$$-\nabla^2 \psi(\hat{\boldsymbol{g}}) = \operatorname{diag}(\boldsymbol{\nu} \odot \exp(\hat{\boldsymbol{g}})) + C(\boldsymbol{X}, \boldsymbol{X})^{-1}. \tag{152}$$

Therefore, the $n \times n$ covariance matrix is:

$$\Sigma_{\boldsymbol{g}} = \left(C(\boldsymbol{X}, \boldsymbol{X})^{-1} + \operatorname{diag}(\boldsymbol{\nu} \odot \exp(\hat{\boldsymbol{g}}))\right)^{-1}, \tag{153}$$

where $H(\hat{\boldsymbol{g}}) = \operatorname{diag}(\boldsymbol{\nu} \odot \exp(\hat{\boldsymbol{g}}))$ denotes the Hessian of the negative log-likelihood. □

Proof 8 (Proof of Proposition 3.7). The posterior mode $\hat{\boldsymbol{g}}$ maximizes the log-posterior:

$$\psi(\boldsymbol{g}) = \boldsymbol{\nu}^\top (\boldsymbol{y} \odot \boldsymbol{g} - \exp(\boldsymbol{g})) - \tfrac{1}{2}\boldsymbol{g}^\top C^{-1}\boldsymbol{g}. \tag{154}$$

The first-order condition (gradient of ψ) at the mode is:

$$\nabla\psi(\hat{\boldsymbol{g}}) = \boldsymbol{\nu} \odot (\boldsymbol{y} - \exp(\hat{\boldsymbol{g}})) - C^{-1}\hat{\boldsymbol{g}} = 0. \tag{155}$$

Summing over all policies and left-multiplying by $\mathbf{1}_n^\top$:

$$\sum_{i=1}^{n} \nu_i\big(y_i - \exp(\hat{g}_i)\big) - \mathbf{1}_n^\top C^{-1}\hat{\boldsymbol{g}} = 0. \tag{156}$$

So exact calibration at the posterior mode would follow if $C^{-1}\hat{\boldsymbol{g}}$ sums to zero (e.g., for a zero-mean GP prior with centered kernel):

$$\sum_{i=1}^{n} \nu_i \exp(\hat{g}_i) = \sum_{i=1}^{n} N_i \qquad \text{s.t.} \quad C^{-1}\hat{\boldsymbol{g}} = 0. \tag{157}$$

Even if Eq. (157) holds, the predictive frequency at each point nonetheless includes the posterior variance term:

$$\hat{Y}_i = \exp(m_i + \tfrac{1}{2}s_i^2) \geq \exp(\hat{g}_i), \tag{158}$$

where $m_i = \hat{g}_i$ and $s_i^2 = \operatorname{diag}(C^{-1} + \operatorname{diag}(\nu \odot \exp(\hat{\boldsymbol{g}})))^{-1}$ at training points. So summing over the portfolio:

$$\sum_{i=1}^{n} \nu_i \hat{Y}_i \geq \sum_{i=1}^{n} \nu_i \exp(\hat{g}_i) \approx \sum_{i=1}^{n} N_i. \tag{159}$$

This shows that the predictions break actuarial fairness. □

Proof 9 (Proof of Proposition 3.8). At the training points, the posterior predictive premiums under the Laplace approximation are given by:

$$\hat{\boldsymbol{\pi}}(\hat{\boldsymbol{g}}) = \exp\big(\hat{\boldsymbol{g}} + \tfrac{1}{2}\operatorname{diag}(\Sigma_{\hat{\boldsymbol{g}}})\big)\,, \tag{160}$$

where $\hat{\boldsymbol{g}}$ denotes the posterior mode maximizing the log-posterior:

$$\psi(\boldsymbol{g}) = \sum_{i=1}^{n}\big(N_i g_i - \nu_i e^{g_i}\big) - \tfrac{1}{2}\boldsymbol{g}^\top C^{-1}\boldsymbol{g}. \tag{161}$$

The Laplace posterior covariance at the mode is:

$$\Sigma_{\hat{\boldsymbol{g}}} = \big(C^{-1} + \operatorname{diag}\big(\nu \odot e^{\hat{\boldsymbol{g}}}\big)\big)^{-1}. \tag{162}$$

Define the portfolio-level calibration constraint:

$$\phi(\hat{\boldsymbol{g}}) := \sum_{i=1}^{n} \nu_i \hat{\pi}_i(\hat{\boldsymbol{g}}) - \sum_{i=1}^{n} N_i. \tag{163}$$

In general, the unconstrained Laplace mode $\hat{\boldsymbol{g}}$ does not satisfy $\phi(\hat{\boldsymbol{g}}) = 0$. We perform a first-order Taylor expansion of ϕ around $\hat{\boldsymbol{g}}$. For $\hat{\boldsymbol{g}}_{\mathrm{BC}} = \hat{\boldsymbol{g}} + \Delta\boldsymbol{g}$:

$$\phi(\hat{\boldsymbol{g}}_{\mathrm{BC}}) \approx \phi(\hat{\boldsymbol{g}}) + \nabla\phi(\hat{\boldsymbol{g}})^{\top} \Delta\boldsymbol{g}. \tag{164}$$

Imposing $\phi(\hat{\boldsymbol{g}}_{\mathrm{BC}}) = 0$, we get the linearized constraint:

$$\nabla\phi(\hat{\boldsymbol{g}})^{\top} \Delta\boldsymbol{g} = \sum_{i=1}^{n} N_i - \sum_{i=1}^{n} \nu_i \hat{\pi}_i(\hat{\boldsymbol{g}}). \tag{165}$$

By the chain rule:

$$\frac{\partial \hat{\pi}_i}{\partial g_j} = \hat{\pi}_i \left(\delta_{ij} + \tfrac{1}{2} \frac{\partial (\Sigma_{\boldsymbol{g}})_{ii}}{\partial g_j} \right), \tag{166}$$

where δ_{ij} denotes the Kronecker delta. Differentiating the covariance matrix gives:

$$\frac{\partial (\Sigma_{\boldsymbol{g}})_{ii}}{\partial g_j} = -(\Sigma_{\boldsymbol{g}})_{ij}^2 \, \nu_j e^{g_j}. \tag{167}$$

This expression is exact but dense and costly to evaluate for large n. To obtain a tractable constrained Newton step, we approximate the Jacobian of $\hat{\boldsymbol{\pi}}$ by:

$$J := \nabla\hat{\boldsymbol{\pi}}(\hat{\boldsymbol{g}}) \;\approx\; \mathrm{diag}(\hat{\boldsymbol{\pi}})(C + \Sigma_{\hat{\boldsymbol{g}}})^{-1}. \tag{168}$$

Substituting Eq. (168) into Eq. (165) gives:

$$\nu^{\top} J \, \Delta\boldsymbol{g} = \sum_{i=1}^{n} N_i - \sum_{i=1}^{n} \nu_i \hat{\pi}_i. \tag{169}$$

In the unconstrained case, the Newton step in Eq. (12) can be rewritten as:

$$\nabla^2\psi(\hat{\boldsymbol{g}}) \, \Delta\boldsymbol{g} = -\nabla\psi(\hat{\boldsymbol{g}}) \tag{170}$$

Enforcing the linearized calibration constraint via a Lagrange multiplier λ_{act} leads to the augmented linear system:

$$\begin{bmatrix} \nabla^2\psi(\hat{\boldsymbol{g}}) & (\nu^{\top} J)^{\top} \\ \nu^{\top} J & 0 \end{bmatrix} \begin{bmatrix} \Delta\boldsymbol{g} \\ \lambda_{\mathrm{act}} \end{bmatrix} = \begin{bmatrix} \nabla\psi(\hat{\boldsymbol{g}}) \\ \sum_{i=1}^{n} N_i - \sum_{i=1}^{n} \nu_i \hat{\pi}_i \end{bmatrix}. \tag{171}$$

Solving this system yields a balance-corrected mode $\hat{\boldsymbol{g}}_{\mathrm{BC}} = \hat{\boldsymbol{g}} + \Delta\boldsymbol{g}$ that satisfies the portfolio-level calibration constraint to first order.

A.2 FGGPR

Proof 10 (Proof of Proposition 4.3). Under the Laplace approximation, the best-estimate posterior of the latent vector $\boldsymbol{g}^{BE}$ associated with the best-estimate kernel is approximated by a Gaussian distribution (see details in Section 3):

$$\boldsymbol{g}^{BE} \mid \boldsymbol{y}, \boldsymbol{X} \;\approx\; \mathcal{N}(\hat{\boldsymbol{g}}^{BE}, \Sigma_{\boldsymbol{g}^{BE}}), \tag{172}$$

where

$$\hat{\boldsymbol{g}}^{BE} = \arg\max_{\boldsymbol{g}^{BE}} \left\{ \psi\left(\boldsymbol{g}^{BE}\right) \right\} \tag{173}$$

$$\Sigma_{\boldsymbol{g}^{BE}} = \left(\left(C^{BE}\right)^{-1} + \operatorname{diag}\left(\boldsymbol{\nu} \odot \exp(\hat{\boldsymbol{g}}^{BE})\right) \right)^{-1}, \tag{174}$$

with

$$\psi\left(\boldsymbol{g}^{BE}\right) = \boldsymbol{\nu}^{\top}\left(\boldsymbol{y} \odot \boldsymbol{g}^{BE} - \exp\left(\boldsymbol{g}^{BE}\right)\right) - \tfrac{1}{2}\boldsymbol{g}^{BE\top}\left(C^{BE}\right)^{-1}\boldsymbol{g}^{BE}, \tag{175}$$

$$C^{BE} = \sigma_n^2 \boldsymbol{I}_n + k^{BE}. \tag{176}$$

The posterior mode is obtained via Newton-Raphson iterations:

$$\hat{\boldsymbol{g}}^{BE\,(j)} = \hat{\boldsymbol{g}}^{BE\,(j-1)} - \left(\nabla^2 \psi(\hat{\boldsymbol{g}}^{BE\,(j-1)}) \right)^{-1} \nabla \psi(\hat{\boldsymbol{g}}^{BE\,(j-1)}), \tag{177}$$

with

$$\nabla \psi(\boldsymbol{g}^{BE}) = \boldsymbol{\nu} \odot \left(\boldsymbol{y} - \exp(\boldsymbol{g}^{BE})\right) - \left(C^{BE}\right)^{-1}\boldsymbol{g}^{BE}, \tag{178}$$

$$\nabla^2 \psi(\boldsymbol{g}^{BE}) = -H(\boldsymbol{g}^{BE}) - \left(C^{BE}\right)^{-1}, \tag{179}$$

where

$$H(\boldsymbol{g}^{BE}) = \operatorname{diag}\left(\boldsymbol{\nu} \odot \exp(\boldsymbol{g}^{BE})\right). \tag{180}$$

Conditionally on $\boldsymbol{g}^{BE}$, the predictive distribution of $g_* = g(\boldsymbol{x}_*)$ is Gaussian. Applying the standard Laplace predictive approximation (see Proposition 3.4) yields:

$$g_* \mid \boldsymbol{x}_*, \boldsymbol{X}, \boldsymbol{y} \;\approx\; \mathcal{N}(m_*^{BE}, (s_*^{BE})^2). \tag{181}$$

Finally, since $Y_* = \exp(g_*)$, the best-estimate GGPR premium is obtained via the log-normal moment correction:

$$\mathbb{E}^{BE}[Y_* \mid \boldsymbol{x}_*, \boldsymbol{X}, \boldsymbol{y}] \approx \exp\left(m_*^{BE} + \tfrac{1}{2}(s_*^{BE})^2 \right) \tag{182}$$

with:

$$m_*^{BE} = k_*^{BE}\big(C^{BE}\big)^{-1}\hat{\boldsymbol{g}}^{BE}, \tag{183}$$

$$\big(s_*^{BE}\big)^2 = k_{**}^{BE} - k_*^{BE}\Big(C^{BE} + H(\hat{\boldsymbol{g}}^{BE})^{-1}\Big)^{-1}k_*^{BE\top}, \tag{184}$$

where the best-estimate Gram matrix defined by the best-estimate kernel in Proposition 4.2:

$$C^{BE} = \sigma_n^2\boldsymbol{I}_n + k^{BE}. \tag{185}$$

□

Proof 11 (Proof of Proposition 4.5). Under the Laplace approximation, the unaware posterior of the latent vector $\boldsymbol{g}^U$ associated with the unaware kernel is approximated by a Gaussian distribution (see details in Section 3):

$$\boldsymbol{g}^U \mid \boldsymbol{y}, \boldsymbol{A} \;\approx\; \mathcal{N}(\hat{\boldsymbol{g}}^U, \Sigma_{\boldsymbol{g}^U}), \tag{186}$$

where

$$\hat{\boldsymbol{g}}^U = \arg\max_{\boldsymbol{g}^U}\left\{\psi\left(\boldsymbol{g}^U\right)\right\} \tag{187}$$

$$\Sigma_{\boldsymbol{g}^U} = \Big(\big(C^U\big)^{-1} + \operatorname{diag}\big(\boldsymbol{\nu}\odot\exp(\hat{\boldsymbol{g}}^U)\big)\Big)^{-1}, \tag{188}$$

with

$$\psi\left(\boldsymbol{g}^U\right) = \boldsymbol{\nu}^\top\left(\boldsymbol{y}\odot\boldsymbol{g}^U - \exp\left(\boldsymbol{g}^U\right)\right) - \tfrac{1}{2}\boldsymbol{g}^{U\top}\big(C^U\big)^{-1}\boldsymbol{g}^U, \tag{189}$$

$$C^U = \sigma_n^2\boldsymbol{I}_n + k^U. \tag{190}$$

The posterior mode is obtained via Newton-Raphson iterations:

$$\hat{\boldsymbol{g}}^{U\,(j)} = \hat{\boldsymbol{g}}^{U\,(j-1)} - \Big(\nabla^2\psi(\hat{\boldsymbol{g}}^{U\,(j-1)})\Big)^{-1}\nabla\psi(\hat{\boldsymbol{g}}^{U\,(j-1)}), \tag{191}$$

with

$$\nabla\psi(\boldsymbol{g}^U) = \boldsymbol{\nu}\odot\big(\boldsymbol{y} - \exp(\boldsymbol{g}^U)\big) - \big(C^U\big)^{-1}\boldsymbol{g}^U, \tag{192}$$

$$\nabla^2\psi(\boldsymbol{g}^U) = -H(\boldsymbol{g}^U) - \big(C^U\big)^{-1}, \tag{193}$$

where

$$H(\boldsymbol{g}^U) = \operatorname{diag}\big(\boldsymbol{\nu}\odot\exp(\boldsymbol{g}^U)\big)\,. \tag{194}$$

Conditionally on $\boldsymbol{g}^U$, the predictive distribution of $g_* = g(\boldsymbol{a}_*)$ is Gaussian. Applying the standard Laplace predictive approximation (see Proposition 3.4) yields:

$$g_* \mid \boldsymbol{a}_*, \boldsymbol{A}, \boldsymbol{y} \;\approx\; \mathcal{N}(m_*^U, \big(s_*^U\big)^2). \tag{195}$$

Finally, since $Y_* = \exp(g_*)$, the best-estimate GGPR premium is obtained via the log-normal moment correction:

$$\mathbb{E}^U[Y_* \mid \boldsymbol{a}_*, \boldsymbol{A}, \boldsymbol{y}] \approx \exp\Big(m_*^U + \tfrac{1}{2}(s_*^U)^2\Big) \tag{196}$$

with:

$$m_*^U = k_*^U \big(C^U\big)^{-1} \hat{\boldsymbol{g}}^U, \tag{197}$$

$$(s_*^U)^2 = k_{**}^U - k_*^U \Big(C^U + H(\hat{\boldsymbol{g}}^U)^{-1}\Big)^{-1} {k_*^U}^\top, \tag{198}$$

where the unaware Gram matrix is defined by the unaware kernel in Proposition 4.4:

$$C^U = \sigma_n^2 \boldsymbol{I}_n + k^U. \tag{199}$$

□

Proof 12 (Proof of Proposition 4.7). Under the Laplace approximation, the aware posterior distribution of the latent vector $\boldsymbol{g}^A$ associated with the aware kernel is approximated by a multivariate Gaussian distribution (see Section 3):

$$\boldsymbol{g}^A \mid \boldsymbol{y}, \boldsymbol{A} \;\approx\; \mathcal{N}\big(\hat{\boldsymbol{g}}^A, \Sigma_{\boldsymbol{g}^A}\big)\,, \tag{200}$$

where the posterior mode $\hat{\boldsymbol{g}}^A$ maximises the log-posterior:

$$\hat{\boldsymbol{g}}^A = \arg\max_{\boldsymbol{g}^A} \big\{\psi(\boldsymbol{g}^A)\big\}\,, \tag{201}$$

$$\psi(\boldsymbol{g}^A) = \boldsymbol{\nu}^\top \big(\boldsymbol{y} \odot \boldsymbol{g}^A - \exp(\boldsymbol{g}^A)\big) - \tfrac{1}{2}{\boldsymbol{g}^A}^\top \big(C^A\big)^{-1} \boldsymbol{g}^A, \tag{202}$$

and the covariance matrix is given by

$$\Sigma_{\boldsymbol{g}^A} = \Big(\big(C^A\big)^{-1} + \operatorname{diag}\big(\boldsymbol{\nu} \odot \exp(\hat{\boldsymbol{g}}^A)\big)\Big)^{-1}. \tag{203}$$

Here,

$$C^A = \sigma_n^2 \boldsymbol{I}_n + k^A, \tag{204}$$

where k^A denotes the Gram matrix associated with the aware kernel defined in Proposition 4.6. The posterior mode $\hat{\boldsymbol{g}}^A$ is obtained via Newton-Raphson iterations:

$$\hat{\boldsymbol{g}}^{A\,(j)} = \hat{\boldsymbol{g}}^{A\,(j-1)} - \Big(\nabla^2 \psi(\hat{\boldsymbol{g}}^{A\,(j-1)})\Big)^{-1} \nabla \psi(\hat{\boldsymbol{g}}^{A\,(j-1)}), \tag{205}$$

where the gradient and Hessian of ψ are given by:

$$\nabla \psi(\boldsymbol{g}^A) = \boldsymbol{\nu} \odot \big(\boldsymbol{y} - \exp(\boldsymbol{g}^A)\big) - \big(C^A\big)^{-1} \boldsymbol{g}^A, \tag{206}$$

$$\nabla^2 \psi(\boldsymbol{g}^A) = -H(\boldsymbol{g}^A) - \big(C^A\big)^{-1}, \tag{207}$$

with

$$H(\boldsymbol{g}^A) = \operatorname{diag}\big(\boldsymbol{\nu} \odot \exp(\boldsymbol{g}^A)\big)\,. \tag{208}$$

Conditionally on the latent vector $\boldsymbol{g}^A$, the predictive distribution of the latent risk for a new policy with allowed attributes $\boldsymbol{a}_*$ and sensitive attributes $\boldsymbol{d}_*$ is Gaussian. Applying the standard Laplace predictive approximation (see Proposition 3.4) yields:

$$g_* \mid \boldsymbol{a}_*, \boldsymbol{d}_*, \boldsymbol{A}, \boldsymbol{y} \;\approx\; \mathcal{N}\Big(m_*^A(\boldsymbol{d}_*), \big(s_*^A(\boldsymbol{d}_*)\big)^2\Big)\,, \tag{209}$$

where

$$m_*^A(\boldsymbol{d}_*) = k_*^A(\boldsymbol{d}_*)\big(C^A\big)^{-1}\hat{\boldsymbol{g}}^A, \tag{210}$$

$$\big(s_*^A(\boldsymbol{d}_*)\big)^2 = k_{**}^A(\boldsymbol{d}_*) - k_*^A(\boldsymbol{d}_*)\Big(C^A + H(\hat{\boldsymbol{g}}^A)^{-1}\Big)^{-1} k_*^A(\boldsymbol{d}_*)^\top. \tag{211}$$

The defining feature of the aware GGPR is that the dependence on sensitive attributes is handled at the kernel level. Specifically, the aware kernel k^A encodes similarity between policies by averaging counterfactual realisations of the sensitive attributes inside the covariance function, rather than by conditioning or marginalising at prediction time. As a result, once the aware Gram matrix $C^A = \sigma_n^2 \boldsymbol{I}_n + k^A$ is constructed, inference and prediction proceed exactly as in the standard GGPR with a fixed kernel.

Conditionally on the training data, the predictive distribution of the latent variable $g_* = g(\boldsymbol{a}_*)$ therefore satisfies

$$g_* \mid \boldsymbol{a}_*, \boldsymbol{A}, \boldsymbol{y} \;\approx\; \mathcal{N}\Big(m_*^A, \big(s_*^A\big)^2\Big)\,, \tag{212}$$

with predictive mean and variance given by

$$m_*^A = k_*^A\big(C^A\big)^{-1}\hat{\boldsymbol{g}}^A, \tag{213}$$

$$\big(s_*^A\big)^2 = k_{**}^A - k_*^A\Big(C^A + H(\hat{\boldsymbol{g}}^A)^{-1}\Big)^{-1} {k_*^A}^\top. \tag{214}$$

Although no explicit marginalisation over sensitive attributes appears in these expressions, the predictive variance $\big(s_*^A\big)^2$ implicitly reflects counterfactual uncertainty induced by the sensitive attributes through the structure of the aware kernel k^A. Finally, recalling that $Y_* = \exp(g_*)$ denotes the claim frequency per unit of exposure, the aware GGPR premium is obtained via the log-normal moment correction:

$$\mathbb{E}^A[Y_* \mid \boldsymbol{a}_*, \boldsymbol{A}, \boldsymbol{y}] \approx \exp\Big(m_*^A + \tfrac{1}{2}\big(s_*^A\big)^2\Big)\,. \tag{215}$$

□

Proof 13 (Proof of Proposition 4.8). The Laplace approximation yields a second-order

Taylor expansion of the log-posterior around a point $\hat{\boldsymbol{g}}_0$:

$$\psi(\boldsymbol{g}) \approx \psi(\hat{\boldsymbol{g}}_0) + \nabla\psi(\hat{\boldsymbol{g}}_0)^\top \Delta\boldsymbol{g} - \tfrac{1}{2}\Delta\boldsymbol{g}^\top \big(C^{-1} + H(\hat{\boldsymbol{g}}_0)\big)\Delta\boldsymbol{g}, \tag{216}$$

where $\Delta\boldsymbol{g} = \boldsymbol{g} - \hat{\boldsymbol{g}}_0$. Maximizing this approximation is equivalent to minimizing the quadratic objective:

$$\min_{\Delta\boldsymbol{g}} \ \tfrac{1}{2}\Delta\boldsymbol{g}^\top Q \Delta\boldsymbol{g} - \nabla\psi(\hat{\boldsymbol{g}}_0)^\top \Delta\boldsymbol{g}, \qquad Q := C^{-1} + H(\hat{\boldsymbol{g}}_0). \tag{217}$$

The predictive premium is nonlinear in $\boldsymbol{g}$:

$$\hat{\boldsymbol{\pi}} = \exp\big(m(\boldsymbol{g}) + \tfrac{1}{2}s^2(\boldsymbol{g})\big). \tag{218}$$

Linearizing around $\hat{\boldsymbol{g}}_0$ gives:

$$\hat{\boldsymbol{\pi}} \approx \hat{\boldsymbol{\pi}}_0 + J\Delta\boldsymbol{g}, \tag{219}$$

with Jacobian:

$$J \approx \operatorname{diag}(\hat{\boldsymbol{\pi}}_0)\big(C + H(\hat{\boldsymbol{g}}_0)^{-1}\big)^{-1}. \tag{220}$$

Imposing zero empirical covariance with the sensitive attribute yields the linear constraint:

$$\boldsymbol{d}^\top C_n J \Delta\boldsymbol{g} = -\boldsymbol{d}^\top C_n \hat{\boldsymbol{\pi}}_0 =: c, \tag{221}$$

which can be written as:

$$a^\top \Delta\boldsymbol{g} = c, \qquad a := J^\top C_n \boldsymbol{d}. \tag{222}$$

Introducing a Lagrange multiplier λ, the Lagrangian associated with Eq. (217) is:

$$\mathcal{L}(\Delta\boldsymbol{g}, \lambda) = \tfrac{1}{2}\Delta\boldsymbol{g}^\top Q \Delta\boldsymbol{g} - \nabla\psi(\hat{\boldsymbol{g}}_0)^\top \Delta\boldsymbol{g} + \lambda(a^\top \Delta\boldsymbol{g} - c). \tag{223}$$

The first-order optimality conditions are:

$$Q\Delta\boldsymbol{g} - \nabla\psi(\hat{\boldsymbol{g}}_0) + \lambda a = 0, \qquad a^\top \Delta\boldsymbol{g} = c, \tag{224}$$

which together form the KKT system:

$$\begin{bmatrix} C^{-1} + H(\hat{\boldsymbol{g}}_0) & a \\ a^\top & 0 \end{bmatrix} \begin{bmatrix} \Delta\boldsymbol{g} \\ \lambda \end{bmatrix} = \begin{bmatrix} \nabla\psi(\hat{\boldsymbol{g}}_0) \\ c \end{bmatrix}, \tag{225}$$

with

$$\hat{\boldsymbol{g}}^C = \hat{\boldsymbol{g}}_0 + \Delta\boldsymbol{g}, \qquad a := J^\top C_n \boldsymbol{d}, \qquad c := -\boldsymbol{d}^\top C_n \hat{\boldsymbol{\pi}}_0. \tag{226}$$

Here $C_n = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$ is the centering matrix, $H(\hat{\boldsymbol{g}}) = \operatorname{diag}(\boldsymbol{\nu} \odot \exp(\hat{\boldsymbol{g}}))$, $\hat{\boldsymbol{\pi}}_0 = \exp(m_0 + \frac{1}{2}s_0^2)$, and

$$J = \frac{\partial\hat{\boldsymbol{\pi}}}{\partial\boldsymbol{g}}\Big|_{\hat{\boldsymbol{g}}_0} \approx \operatorname{diag}(\hat{\boldsymbol{\pi}}_0)\big(C + H(\hat{\boldsymbol{g}}_0)^{-1}\big)^{-1}. \tag{227}$$

Solving for $\Delta \boldsymbol{g}$ yields the corrective posterior mode $\hat{\boldsymbol{g}}^C = \hat{\boldsymbol{g}}_0 + \Delta \boldsymbol{g}$. The predictive mean and variance then follow from standard Laplace-approximated GP prediction formulas. □

A.3 Results

Parameter	Value	Description
Best value	0.05169	Objective metric (e.g., validation loss) achieved in the best trial
Std. Dev.	0.00607	Standard deviation of the objective metric across cross-validation folds
learning_rate	0.00315	Step size applied at each boosting iteration
feature_fraction	0.8925	Fraction of features randomly selected for building each tree
bagging_fraction	0.9662	Fraction of data randomly selected for each iteration (subsampling)
bagging_freq	6	Number of iterations between each bagging operation
lambda_l2	0.00155	L2 regularization term applied to leaf weights
min_sum_hessian_in_leaf	48.31	Minimum sum of second-order gradients (Hessians) in a leaf to allow a split
max_depth	8	Maximum depth allowed for each tree
min_gain_to_split	0.296	Minimum gain required for a split to be performed
min_data_in_leaf	43	Minimum number of data points required in a leaf
num_leaves	233	Maximum number of leaves in a single tree

Table 13: Best LightGBM hyperparameter configuration and performance from the optimization study. Long descriptions are wrapped inline.