α-FAIR INSURANCE PRICING: A FAIRNESS CONTINUUM

Anonymous authors
Paper under double-blind review

ABSTRACT

Fairness in insurance pricing remains a long-standing and deeply debated puzzle. On one hand, insurers, driven by profitability considerations, set premiums that differentiate across individual risks to achieve actuarial fairness. On the other hand, insurance serves a critical societal function by pooling risks across a population, motivating cross-subsidization among groups to promote solidarity fairness. The tension between these two competing notions of fairness makes insurance pricing inherently complex, particularly in modern settings where granular data allow for increasingly fine risk differentiation and regulators face growing pressure to protect vulnerable groups. To address this challenge, we propose an α-**F**air **I**ndividual **S**olvent **P**remium (α-FISP) framework for insurance pricing that explicitly captures the trade-off between actuarial and solidarity fairness while guaranteeing solvency, a fundamental requirement in insurance operations. We formulate the pricing problem as a constrained optimization task, where actuarially fair premiums are adjusted subject to budget constraints on cross-subsidization within each risk class. This formulation naturally yields a family of solutions parameterized by α, tracing a continuum between purely actuarial and purely solidarity-based pricing and enabling decision-makers to select an operating point along this fairness spectrum. We derive various theoretical results for the proposed framework. Numerical experiments show that α-FISP is computationally tractable and aligns well with the U.S. regulatory regimes featuring heterogeneous state-level fairness requirements.

1 INTRODUCTION

Fairness has become a central theme in the era of data-driven decision-making, particularly as machine learning (ML) algorithm are increasingly deployed in high-stakes domains such as credit scoring, criminal justice, and healthcare. In this paper, we focus on fairness in insurance pricing, a setting where algorithmic decisions have widespread financial and societal impact. The insurance industry is inherently data-driven, relying heavily on statistical and predictive models to derive insights from data and set premiums [15]. With the rapid growth of InsurTech and the availability of granular data, algorithmic pricing plays a central role in modern insurance operations [32].Given the insurance market's size and economic criticality, ensuring fair pricing is a practical and societal imperative.

Insurance operates in a profit-driven environment while simultaneously providing financial protection as a societal good. This dual role makes fairness in insurance pricing a long-standing and deeply debated puzzle. On one hand, insurers—driven by profitability, competition, and solvency considerations—set premiums that differentiate across individual risks to achieve actuarial fairness. A premium is actuarially fair if it equals or closely reflects the insured's expected loss [20]. Traditionally, insurers have implemented actuarial fairness through risk classification, grouping individuals with similar attributes into homogeneous risk classes. Refining these classes improves alignment between premiums and underlying risk, which helps mitigate adverse selection and encourages loss-control behavior by making policyholders financially accountable for their risk exposure [30; 31]. On the other hand, insurance fulfills a critical social function by pooling risks across the population, which motivates cross-subsidization between groups to promote solidarity fairness. A solidarity-oriented view of fairness, often grounded in principles of social justice [22], emphasizes equality in sharing the burden of losses. Under this view, fairness may require some policyholders to pay more than their pure expected loss—or others to pay less—to maintain access and affordability for those at higher risk through no fault of their own.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The tension between these two competing notions of fairness—solidarity and actuarial—makes insurance pricing inherently complex. This debate is often conceptualized along a continuum, with a pure "moral" or solidarity-driven perspective at one end and a pure "technical" or actuarially driven perspective at the other [33]. While these notions are incompatible when taken to their extremes, empirical studies suggest that public preferences tend to occupy a nuanced middle ground, accepting some degree of cross-subsidization as long as pricing remains broadly risk-reflective [11]. This balancing act has become even more challenging in modern settings. Advances in data availability and predictive modeling have enabled insurers to differentiate risks with unprecedented granularity. While such refinement improves actuarial precision, it can also magnify premium differences between groups, raising concerns about affordability and social exclusion. In addition, because many rating variables are correlated with sensitive attributes such as gender, race, or socioeconomic status, data-driven pricing can unintentionally reproduce or reinforce existing inequities [16]. Consequently, regulators face growing pressure to protect vulnerable populations and restrict the extent of permissible risk differentiation. Recent regulatory interventions highlight this trend, including the EU Gender Directive (2004/113/EC), which prohibits the use of gender in life insurance pricing [10], and Colorado's SB 21-169 statute, which bars unfair discrimination across multiple protected characteristics [7]. Taken together, the tension between market-driven risk differentiation and regulation-mandated solidarity highlights the need for frameworks that balance these objectives, creating pricing systems that are both economically viable and socially acceptable.

Motivated by these observations, we introduces a novel insurance pricing framework, termed α-fairness, to explicitly capture the trade-off between actuarial and solidarity fairness. This framework yields premiums parameterized by $\alpha \in [0, 1]$, defining a continuum from purely actuarial to purely solidarity-based pricing and enabling decision-makers to select an operating point aligned with their strategic or regulatory priorities. First, we formally define the notion of α-fairness. A premium structure is said to be α-fair if, within any group of policyholders sharing the same non-sensitive attributes (a risk class), the ratio of the lowest to the highest premium is at least α. At the extremes, $\alpha = 0$ corresponds to fully differentiated premiums based on sensitive attributes (pure actuarial fairness), while $\alpha = 1$ enforeces uniform premiums within the group (pure solidarity fairness). Second, we formulate the computation of α-fair premiums as a constrained optimization problem that adjusts actuarially fair premiums subject to budget constraints controlling the degree of cross-subsidization within each risk class. The resulting framework is computationally tractable, statistically well-grounded, and robust. Third, the proposed α-fair premiums satisfy individual-level solvency requirements, a critical property for insurance pricing that is often overlooked in existing fair-pricing studies. More broadly, the framework is flexible and accommodates practical considerations such as portfolio-level profit targets and regulatory caps on rate changes.

2 Related Work

2.1 Fairness in Machine Learning

Fairness in machine learning has received growing attention due to concerns about algorithmic bias [26; 3]. Existing definitions broadly fall into individual fairness—treating similar individuals similarly [12]—and group fairness, which enforces parity across subpopulations via criteria such as equalized odds, demographic parity, or calibration [17; 14; 19]. These notions often conflict, highlighting trade-offs under practical constraints [6; 28; 8].

Methodologically, fairness interventions include pre-processing (e.g., reweighting, fair representations), in-processing (e.g., constrained or regularized objectives), and post-processing (e.g., adjusting outputs) [18; 37; 4; 36; 1; 17]. Recent work emphasizes fairness–efficiency trade-offs via multi-objective or constrained optimization [27; 34; 25].

Our work aligns with this optimization-based perspective but differs in focus. Rather than statistical parity, we study fairness in insurance pricing as an economic allocation problem, explicitly balancing actuarial and solidarity fairness in premium design.

2.2 Fairness in Insurance Pricing

The discussion of fairness in insurance pricing has a long history in actuarial science, well before the emergence of algorithmic fairness in machine learning. Classical insurance theory distinguishes

between actuarial fairness and solidarity fairness (or social fairness) [33; 22]. Much of the early actuarial literature treated this as a normative tension rather than a quantitative trade-off, with fairness operationalized through the design of rating structures that balance heterogeneity and pooling [5]. Recent research in actuarial science has revived this debate in light of advances in data availability and machine learning. A growing body of work has sought to formalize fairness criteria and clarify how regulatory objectives translate into modeling constraints [16; 35; 13; 9]. Collectively, these studies emphasize that fairness in insurance cannot be reduced to a single metric, but must account for causal structure, legal standards, and market functioning.

A complementary line of research focuses on computing discrimination-free premiums. For example, Lindholm et al. [23] define a framework for removing discriminatory dependence between sensitive attributes and premiums, and Lindholm et al. [24] extend this approach using multi-task networks. Araiza Iturria et al. [2] propose a causal formulation that achieves fairness through counterfactual adjustments, ensuring pricing decisions are invariant to changes in protected attributes. More recently, Zhang et al. [38] incorporate local differential privacy to handle scenarios where sensitive attributes are unobserved or privatized. These approaches primarily operationalize fairness at the prediction level, using constraints or transformations applied to the outputs of pricing models.

In contrast, we take a structural view, modeling fairness as an allocation mechanism that adjusts actuarially fair premiums under cross-subsidization. Our α-fairness framework provides a tunable trade-off between actuarial and solidarity fairness, linking classical actuarial principles with modern optimization-based approaches.

2.3 Our Contribution

This paper advances the literature on fairness in both machine learning and insurance pricing in several important ways. First, we introduce a new paradigm, the α-fairness framework, which formalizes a continuum between actuarial and solidarity fairness. Unlike conventional fairness metrics that focus on prediction errors or demographic parity, α-fairness directly regulates premium differentials within risk classes, reflecting the economic nature of fairness in insurance. Second, we formulate the computation of α-fair premiums as a constrained optimization problem that adjusts actuarially fair premiums subject to group-specific cross-subsidization constraints, while guaranteeing individual-level solvency—a property often missing in existing fairness-based pricing approaches. The resulting α-Fair Individual Solvent Premium (α-FISP) is supported by rigorous theoretical guarantees, including existence, boundedness, sample complexity preservation, and order preservation. Third, the framework yields a computationally tractable premium and is regulatorily adaptable: varying state-level fairness requirements can be implemented through the interpretable fairness budget parameter α, enabling a unified pricing pipeline aligned with U.S. insurance regulation. Simulation and empirical analyses further demonstrate that α-fair pricing aligns naturally with regulatory objectives. Finally, by framing fairness as an allocation problem rather than purely a predictive constraint, our framework bridges machine learning and actuarial perspectives, integrating ethical and regulatory considerations directly into insurance pricing.

3 Problem Formulation

3.1 Preliminaries

Let (X, D, Y) be a random triplet with unknown joint distribution $\mathbb{P}_{X,D,Y}$. We observe n samples:

$$(x_i, d_i, y_i) \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{X,D,Y}, \quad i = 1, \ldots, n.$$

In the insurance pricing context, $x_i \in \mathcal{X} \subseteq \mathbb{R}^p$ denotes a p-dimensional vector of non-sensitive rating variables used for risk classification, $d_i \in \mathcal{D}$ denotes sensitive attributes, and $y_i \in \mathcal{Y}$ represents the realized loss (or risk outcome) of a policyholder. We assume that Y is a non-negative random variable with finite second moment, i.e., $\mathbb{E}[Y^2] < \infty$, and that D takes values in a finite discrete set, i.e., $|\mathcal{D}| < \infty$.

Definition 3.1 (Actuarial Fair)**.** *The actuarial fair premium for Y w.r.t. (X, D) is defined as:*

$$\mu(X, D) := \mathbb{E}[Y|X, D].$$

This premium fully reflects differences in expected losses attributable to both non-sensitive and sensitive attributes. For notational convenience, we denote $\mu(X, D)$ by $\mu_D(X)$ throughout the paper.

Definition 3.2 (Solidarity Fair)**.** *The solidarity fair premium for Y w.r.t. X is defined as:*

$$\mu(X) := \mathbb{E}[Y|X].$$

Under this pricing rule, individuals sharing the same non-sensitive attributes are charged a common premium, thereby pooling risks across sensitive attributes.

Definition 3.3 (Individual Fairness)**.** *Let $f \in \mathcal{F}$ be a pricing function $f : \mathcal{Z} \to \mathbb{R}$ defined on an input space $\mathcal{Z}$. Equip $\mathcal{Z}$ with a metric $d_{\mathcal{Z}}$ and the output space with metric $d_{\mathbb{R}}$. For $\epsilon \geq 0$ and $\delta \geq 0$ given, f is* individually fair *if for any $z, z' \in \mathcal{Z}$ satisfying $d_{\mathcal{Z}}(z, z') \leq \epsilon$, we have $d_{\mathbb{R}}(f(z), f(z')) \leq \delta$.*

This principle requires that individuals who are similar according to the chosen input-space metric receive similar outcomes. The thresholds (ϵ, δ) allow relaxation of strict fairness requirements when exact parity is not feasible.

3.2 α-Fairness and Individual Solvency

For fair insurance pricing, we introduce two purpose-built distance functions.

Definition 3.4 (Metric on Feature Space)**.** *Let $dist_{\mathcal{X}\times\mathcal{D}} : (\mathcal{X} \times \mathcal{D}) \times (\mathcal{X} \times \mathcal{D}) \to [0, \infty]$ be defined for $(x, d), (x', d') \in \mathcal{X} \times \mathcal{D}$ as*

$$dist_{\mathcal{X}\times\mathcal{D}}\big((x, d), (x', d')\big) = \begin{cases} 0, & \text{if } x = x', \\ +\infty, & \text{if } x \neq x'. \end{cases}$$

This extended pseudometric is characterized exclusively by non-sensitive attributes. It treats policyholders with identical non-sensitive attributes ($x = x'$) as maximally similar regardless of sensitive attributes, while deeming those with differing attributes ($x \neq x'$) incomparable.

Definition 3.5 (Metric on Outcome Space)**.** *For strictly positive premiums $y, y' > 0$ assigned by a pricing function f, define*

$$dist_{\mathcal{Y}}(y, y') := 1 - \min\left(\frac{y}{y'}, \frac{y'}{y}\right).$$

This scale-invariant metric satisfies all metric axioms for $y, y' > 0$. It quantifies proportional deviation between premiums, where $dist_{\mathcal{Y}}(y, y') = 0$ if and only if $y = y'$ (exact/strict fairness constraint), and approaches 1 as the ratio between premiums becomes unbounded (no fairness constraint).

With the distance functions defined on the feature space and outcome space, we can now formalize a proportional notion of individual fairness for insurance premiums.

Definition 3.6 (α-Fairness)**.** *Let $f \in \mathcal{F}$ be a pricing function $f : \mathcal{X} \times \mathcal{D} \to \mathbb{R}_+$, and let $dist_{\mathcal{X}\times\mathcal{D}}$ and $dist_{\mathcal{Y}}$ be the feature-space and outcome-space distance functions defined in Definitions 3.4 and 3.5, respectively. Given thresholds $\epsilon \geq 0$ and $\alpha \in [0, 1]$, f is said to be* α-fair *if, for any pair of individuals $(x, d), (x', d') \in \mathcal{X} \times \mathcal{D}$ satisfying $dist_{\mathcal{X}\times\mathcal{D}}\big((x, d), (x', d')\big) \leq \epsilon$, we have $dist_{\mathcal{Y}}\big(f(x, d), f(x', d')\big) \leq 1 - \alpha$.*

When $x = x'$, α-fairness requires $\min\left(\frac{f(x,d)}{f(x,d')}, \frac{f(x,d')}{f(x,d)}\right) \geq \alpha$, ensuring that individuals with identical non-sensitive attributes receive proportionally similar premiums, with α controlling the allowable deviation. When $x \neq x'$, no fairness constraint is imposed, allowing full use of non-sensitive attributes for risk classification. The scale-invariant design of $dist_{\mathcal{Y}}$ ensures this proportional constraint holds across all premium levels.

For notational clarity, let $\{d_1, \ldots, d_{|\mathcal{D}|}\}$ denote the support of D, and let $p_{k|x} := \mathbb{P}(D = d_k|X = x)$.

Definition 3.7 (Individual Solvency)**.** *Let $f \in \mathcal{F}$ be a pricing function $f : \mathcal{X} \times \mathcal{D} \to \mathbb{R}_+$. The function f is* individually solvent *if, for every fixed $x \in \mathcal{X}$,*

$$\sum_{k=1}^{|\mathcal{D}|} \left(f(x, d_k) - \mu_k(x)\right) p_{k|x} \geq 0,$$

where $\mu_k(x) := \mathbb{E}[Y|X = x, D = d_k]$ denotes the actuarially fair premium.

Individual solvency enforces self-sufficiency by requiring premiums for identical non-sensitive attributes to cover their group's expected loss. This precludes cross-subsidization across heterogeneous groups ($x \neq x'$), confining financial liability to actuarially comparable pools.

3.3 Constrained Optimization

Our objective is to compute insurance premiums that satisfy both α-fairness and individual solvency. We formulate this task as a constrained risk minimization problem. Specifically, for a pricing function $f : \mathcal{X} \times \mathcal{D} \to \mathbb{R}_+$ and a fixed $\alpha \in (0, 1]$, we seek to minimize the mean squared error for Y:

$$\min_{f \in \mathcal{F}} \mathbb{E}_{(X,D,Y)\sim\mathbb{P}_{X,D,Y}} \left[(f(X, D) - Y)^2\right], \tag{1}$$

subject to the constraints:

$$1 - dist_{\mathcal{Y}} \left(f(x, d_k), f(x, d_{k'})\right) \geq \alpha, \quad \forall x \in \mathcal{X},\ k, k' \in [|\mathcal{D}|]. \tag{2}$$

$$\sum_{k=1}^{|\mathcal{D}|} \left(f(x, d_k) - \mu_k(x)\right) \cdot p_{k|x} \geq 0, \quad \forall x \in \mathcal{X}. \tag{3}$$

The solution to this optimization problem is denoted the α-*Fair Individual Solvent Premium* (α-FISP). To simplify, we express the nonlinear constraint (2) in the following equivalent linear inequalities:

$$f(x, d_k) - \alpha\, f(x, d_{k'}) \geq 0, \tag{4}$$

$$f(x, d_{k'}) - \alpha\, f(x, d_k) \geq 0, \tag{5}$$

for all $x \in \mathcal{X}$ and all pairs $k < k'$.

Taken together, individual solvency and α-fairness yield a two-tier equity framework. Individual solvency anchors premiums to legitimate risk differentiation based on non-sensitive attributes, ensuring inter-group equity. Conditional on risk class, α-fairness governs intra-group treatment across sensitive attributes, enforcing proportional fairness. This hierarchical separation resolves the tension between actuarial fairness and societal fairness by decoupling risk classification from fairness enforcement.

4 Properties of the Fair Pricing Optimization

In this section, we analyze theoretical properties of the constrained problem defined by (1)–(5). We establish existence and uniqueness of the optimal solution, ensuring well-posedness. We then show that imposing α-fairness and individual solvency leads to only a controlled increase in generalization error relative to the unconstrained solution. For completeness, we derive the out-of-sample constraint violation rate. Finally, we show that for any $\alpha \in (0, 1]$, the optimal pricing function is order-preserving in risk, supporting its fairness and practical relevance.

For the population optimization and its structural analysis, we take $\mathcal{F} = L^2(\mathbb{P}_{X,D})$ as the ambient space of real-valued, square-integrable functions and interpret constraints (3) and (4)–(5) $\mathbb{P}_X$-a.e. Under the assumptions of Proposition 4.1, these constraints imply $f(x, d_k) \geq \alpha\mu(x) \geq \alpha\mu_{\min} > 0$ for every k, $\mathbb{P}_X$-a.e. Thus, the formulation over $\mathcal{F}$ is equivalent to the original formulation for strictly positive premiums.

Proposition 4.1 (Existence and Uniqueness). *Fix $\alpha \in (0, 1]$, suppose there exists a constant $\mu_{\min} > 0$ s.t. $\mu_k(x) \geq \mu_{\min}$ and $p_{k|x} > 0$ for $\mathbb{P}_X$-a.e. $x \in \mathcal{X}$ and all $k \in [|\mathcal{D}|]$. Let $\mathcal{F}_{\alpha,IS} \subset \mathcal{F}$ be the feasible set of functions satisfying the individual solvency constraint* (3) *and the α-fairness constraints* (4)–(5). *Then the optimization problem* (1) *admits a unique minimizer $f^* \in \mathcal{F}_{\alpha,IS}$ up to $\mathbb{P}_{X,D}$-a.s. equality.*

For the finite-sample analysis, Let $\mathcal{H} \subset L^2(\mathbb{P}_{X,D})$ be a fixed hypothesis class. Define

$$\mathcal{H}_\alpha := \{f \in \mathcal{H} \mid f \text{ satisfies (4)–(5)}\}, \text{ and } \mathcal{H}_{\alpha,IS} := \mathcal{H} \cap \mathcal{F}_{\alpha,IS}.$$

Write $\mathcal{R}(f) := \mathbb{E}\left[(f(X, D) - Y)^2\right]$ and denote the minimizer in Proposition 4.1 by $f^*_{\alpha,IS}$. Then, the following theorem compares an empirical pricing rule with the actuarially fair premium $\mu_D(X)$. Technical assumptions and the empirical optimization problem are specified in Appendix C.2.

Theorem 4.2. *Fix $\alpha \in (0,1]$ and $\delta \in (0,1)$. Suppose Assumptions C.2.1-C.2.4 in Appendix C hold and $f^*_{\alpha,IS} \in \mathcal{H}$. Assume that $|Y| \le B$ a.s. and $|f(X,D)| \le B_f$ a.s. for every $f \in \mathcal{H}$, and set $M := B + B_f$. Suppose there exist constants $\mu_{\min}$ and $\mu_{\max}$ s.t. $0 < \mu_{\min} \le \mu_D(X) \le \mu_{\max} < +\infty$ a.s. and define the heterogeneity ratio $R_\mu := \frac{\mu_{\max}}{\mu_{\min}}$. There exists a universal constant $C_0 > 0$ s.t. when the slack calibration (S1) uses $C \ge C_0$, w.p. at least $1-\delta$, the empirical feasible set (S2) is nonempty and its measurable empirical risk minimizer $\hat{f}_{\alpha,IS}$, defined in (S3) satisfies:*

$$\mathcal{R}(\hat{f}_{\alpha,IS}) - \mathcal{R}(\mu_D) \le 8M\mathfrak{R}_n(\mathcal{H}_\alpha) + 6M^2\sqrt{\log(8/\delta)/2n} + \mu^2_{\max}\left[(R_\mu - 1/\alpha)_+\right]^2,$$

where $\mathfrak{R}_n(\mathcal{H}_\alpha)$ is the expected Rademacher complexity of $\mathcal{H}_\alpha$.

Remark 4.3. *The excess risk bound decomposes into two components. The first captures the statistical error arising from estimation and function class complexity. The second represents a fairness-induced bias that quantifies the intrinsic cost of enforcing α-fairness under heterogeneous risk profiles. The effect of individual solvency enters through the slack term η_n, which is governed by the accuracy of the plug-in estimators and does not alter the leading convergence rate. As $\alpha \to 0$, the fairness constraint becomes inactive, and the bias term vanishes. Conversely, as $\alpha \to 1$, the fairness constraint is most restrictive, and the bias term attains its maximum.*

Corollary 4.4 (Out-of-sample Constraint Violations)**.** *Under the setup of Theorem 4.2, $\hat{f}_{\alpha,IS}$ satisfies* (4)- (5) *$\mathbb{P}_X$-a.e. Moreover, w.p. at least $1-\delta$ over the training data, simultaneously for all $\tau > 0$,*

$$\mathbb{P}_X\left(\sum_{k=1}^{|\mathcal{D}|} p_{k|X}\hat{f}_{\alpha,IS}(X,d_k) < \mu(X) - \tau\right) \le \frac{8M\mathfrak{R}_n(\mathcal{H}_\alpha) + 6M^2\sqrt{\log(8/\delta)/2n}}{\tau^2},$$

where the test probability holds the fitted estimator fixed. For fixed α, δ, τ, hypothesis class, and boundedness and complexity constants, this upper bound is $O(n^{-1/2})$.

We have established that the proposed constrained optimization problem is well-posed and admits a consistent $\mathcal{O}(n^{-1/2})$ constraint violation rate. We now study the structural properties of α-FISP that align naturally with insurance operations and economic intuition.

Proposition 4.5 (Boundedness of α-FISP)**.** *Fix $\alpha \in (0,1]$. Under the assumptions of Proposition 4.1, for $k \in [|\mathcal{D}|]$ and $\mathbb{P}_X$-a.e., its population minimizer satisfies*

$$\min_k \mu_k(x) \le f^*(x,d_k) \le \max_k \mu_k(x).$$

Remark 4.6. *Proposition 4.5 ensures that α-FISP lies within the convex hull of the actuarially fair premiums for each risk class x. This property is critical both operationally and economically. Operationally, it guarantees that fairness and solvency constraints do not induce extreme pricing behavior: no premium is pushed below the minimum expected loss or inflated beyond the highest actuarially fair benchmark within the same risk class. Such bounds are consistent with underwriting practices and regulatory expectations, which prohibit loss-leading or excessively punitive pricing. Economically, the result shows that fairness-induced adjustments operate as a reallocation of premiums within an actuarially meaningful range, rather than as a distortion that breaks the link between price and risk.*

Proposition 4.7 (Order Preservation)**.** *Fix $\alpha \in (0,1]$. Under the assumptions of Proposition 4.1. Let f^* denote the population minimizer. For $\mathbb{P}_X$-a.e. x, the following hold:*

(i) For every $k \in [|\mathcal{D}|]$,

$$\mu_k(x) \ge \mu(x) \implies f^*(x,d_k) \ge \mu(x).$$

(ii) Suppose there is a partition of $\mathcal{D}$ into a low-risk cluster C_L and a high-risk cluster $C_H = \mathcal{D} \setminus C_L$, s.t. for some $M(\alpha,x) > 0$,

$$\begin{cases} f^*(x,d_k) = \alpha M(\alpha,x), & \forall k \in C_L \\ f^*(x,d_k) = M(\alpha,x), & \forall k \in C_H. \end{cases}$$

If $\sum_{k\in C_L} p_{k|x} \le \frac{1}{1+\alpha}$, then for every $k \in C_L$,

$$f^*(x,d_k) \le \mu(x).$$

(iii) For any $k, j \in [|\mathcal{D}|]$,

$$\mu_k(x) \ge \mu_j(x) \quad \implies \quad f^*(x,d_k) \ge f^*(x,d_j).$$

Remark 4.8. *Part (i) shows that if a risk class (x, d_k) is priced above the solidarity premium under actuarial fairness ($\alpha = 0$), then for any $\alpha \in (0, 1]$, the α-FISP remains lower bounded by $\mu(x)$. Part (ii) provides a complementary upper bound under a structural condition on the optimal solution and a sufficient mass of low-risk groups; see Appendix A.1 for illustration. While this condition is not universal, part (iii) establishes a robust property: α-FISP preserves the ordering of premiums across sensitive attributes within each risk class. In particular, higher-risk groups are never priced below lower-risk ones, ensuring monotonicity and interpretability.*

5 Numerical Experiments

5.1 Simulations

In this section, we conduct a series of controlled numerical experiments to illustrate the behavior of α-FISP under varying fairness budgets and to empirically validate our theoretical results.

We generate two datasets, each with $n = 10,000$ observations. In Dataset A, the response variable Y represents insurance claim costs. The non-sensitive attributes X consist of age and smoking status. The age variable X_{A} takes integer values between 18 and 80, and the smoking status variable X_{S} is binary, denoting smoker (S) or non-smoker (NS). The sensitive attribute D represents gender and is binary with levels male (M) and female (F). The data-generating process for Dataset A is:

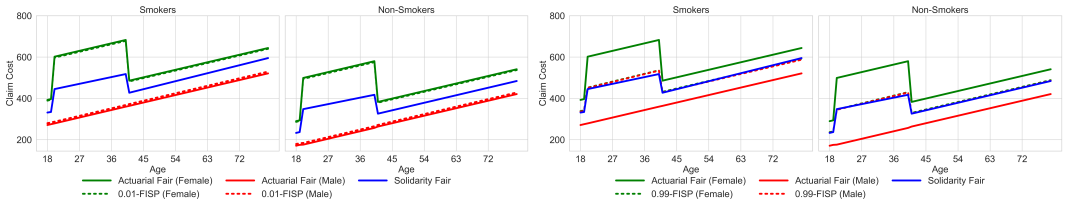
$$\begin{aligned} Y =& 100 + 4X_{\mathrm{A}} + 100 \cdot \mathbf{1}\{X_{\mathrm{S}} = \mathrm{S}\} + 120 \cdot \mathbf{1}\{D = \mathrm{F}\} \\ &+ 200 \cdot \mathbf{1}\{X_{\mathrm{A}} \in [20, 40], D = \mathrm{F}\} + \epsilon, \epsilon \sim N(0, 40^2) \end{aligned}$$

Dataset B retains the same response Y and non-sensitive attributes X, but extends the sensitive attributes to three levels (A, B, and C). Due to limited space, we defer the results in Appendix B.3.

To examine the order-preserving property in Proposition 4.7, we further construct variants A1/B1 and A2/B2 by assigning different group probabilities. Specifically, A1 and B1 satisfy the condition $\sum_{k \in [|C_L|]} p_{k|x} \leq \frac{1}{1+\alpha}$, whereas A2 and B2 violate this condition. In all simulations, we assume $X_{\mathrm{A}} \perp D$ and $X_{\mathrm{A}} \perp X_{\mathrm{S}}$. Detailed probability assignments are provided in Appendix B.1.

All models are trained using two-layer feed-forward neural networks, with hidden layer sizes of $(5, 5)$ for Dataset A and $(16, 8)$ for Dataset B. Model training is performed by optimizing the Lagrangian associated with the constrained optimization problem defined by the objective (1) and the constraints (3), (4), and (5). The final model parameters are obtained by averaging across five random splits.

Figure 1 illustrates the behavior of α-FISP using Dataset A1. Panel (a) shows that as $\alpha \to 0$, the premiums converge to the actuarially fair premiums defined in Definition 3.1, whereas panel (b) shows that as $\alpha \to 1$, they approach solidarity-fair premiums as defined in Definition 3.2.



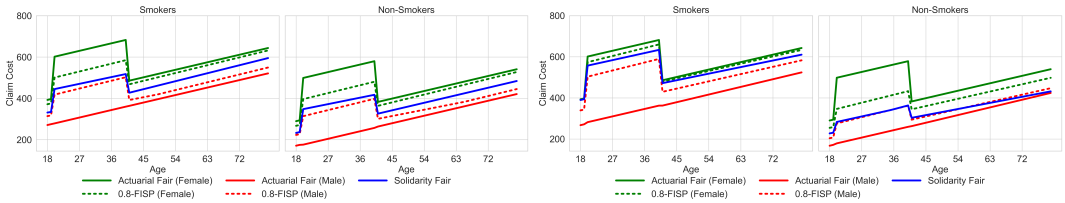
(a) Convergence to actuarially fair premiums as $\alpha \to 0$ (b) Convergence to solidarity-fair premiums as $\alpha \to 1$

Figure 1: α-fair premiums at the two fairness extremes

Figure 2 illustrates the comparison between lower-tier premiums and the solidarity benchmark in Proposition 4.7. When the sufficient mass condition is not satisfied, a lower-tier premium may exceed the solidarity benchmark; this does not contradict pairwise order preservation in Proposition 4.7.

5.2 Health Insurance

We analyze the U.S. Health Insurance dataset [21], treating sex as the sensitive attribute. Reflecting common industry practices where tail risks are ceded to reinsurance, we restrict our study to the



(a) Condition satisfied ($P_L = 0.5 \leq \frac{1}{1+\alpha}$) (b) Condition violated ($P_L = 0.7 > \frac{1}{1+\alpha}$)

Figure 2: Order preservation of α-fair premiums for $|\mathcal{D}| = 2$ ($\alpha = 0.8$)

dense region of frequent claims. Within this region, we examine observations at the 25th, 50th, and 75th percentile of health expenditures as representative risk classes. As in Section 5.1, we employ a two-layer feed-forward NN with 16 and 8 neurons in the first and second hidden layers, respectively. Figure 3 displays α-FISP for $\alpha \in \{0.25, 0.5, 0.75\}$, with actuarially fair and solidarity-fair premiums shown as benchmarks, for the risk class corresponding to the 50th percentile of health expenditures.

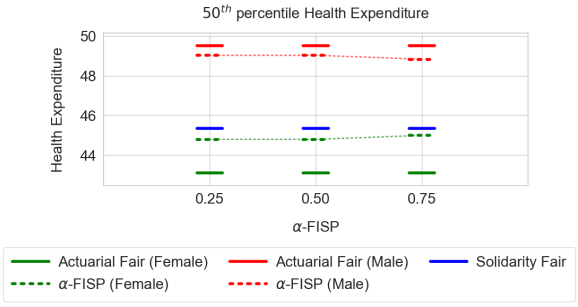


Figure 3: α-fair premiums for the median health risk class

The results illustrate that increasing α tightens the fairness constraint, progressively shifting premiums from actuarial levels toward the solidarity benchmark. This visualization highlights the cross-subsidization mechanism embedded in the α-fair framwork and empirically validates the order-preserving property established in Proposition 4.7. These findings are consistent with the simulation results in Section 5.1, and analogous patterns for other expenditure percentiles are in Appendix B.4.

5.3 Solvency Margin and Choice of α

In this section, we empirically compare α-FISP with the widely cited fair insurance pricing method (DFIP) proposed by [23]. We also provide some guidance on how to choose α in practice.

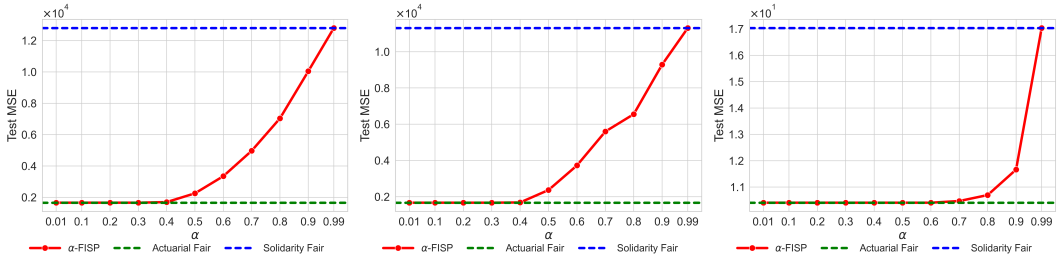
Table 1: α-FISP and DFIP Comparison on Dataset A ($|\mathcal{D}| = 2$)

Condition	**Metric**	**α**					
		0.01	**0.20**	**0.40**	**0.60**	**0.80**	**0.99**
Satisfied	α-FISP Test MSE	1658.57	1659.62	1697.04	3349.90	7039.25	12798.40
	DFIP Test MSE	13081.87	13081.87	13081.87	13081.87	13081.87	13081.87
	α-FISP Solvency Margin	**2.16**	**2.20**	**3.33**	**9.35**	**10.72**	**7.09**
	DFIP Solvency Margin	-9.87	-9.87	-9.87	-9.87	-9.87	-9.87
	α-FISP Solvency Violation Rate (%)	0.00	0.00	0.00	0.00	0.00	0.00
	α-FISP Fairness Violation Rate (%)	0.00	0.00	0.74	1.22	1.11	0.70
Violated	α-FISP Test MSE	1674.11	1674.11	1687.53	3727.72	6555.59	11302.19
	DFIP Test MSE	12551.79	12551.79	12551.79	12551.79	12551.79	12551.79
	α-FISP Solvency Margin	**2.63**	**2.63**	**3.42**	**13.99**	**9.29**	**15.45**
	DFIP Solvency Margin	-7.17	-7.17	-7.17	-7.17	-7.17	-7.17
	α-FISP Solvency Violation Rate (%)	0.00	0.00	0.00	0.00	0.00	0.00
	α-FISP Fairness Violation Rate (%)	0.00	0.00	0.73	0.88	0.57	0.00

Table 2: α-FISP and DFIP Comparison on Healthcare Insurance Dataset

Metric	$\boldsymbol{\alpha}$					
	0.01	**0.20**	**0.40**	**0.60**	**0.80**	**0.99**
α-FISP Test MSE	10.41	10.41	10.41	10.41	10.70	17.03
DFIP Test MSE	18.48	18.48	18.48	18.48	18.48	18.48
α-FISP Solvency Margin	**0.68**	**0.68**	**0.68**	**0.68**	**0.62**	**1.32**
DFIP Solvency Margin	-0.19	-0.19	-0.19	-0.19	-0.19	-0.19
α-FISP Solvency Violation Rate (%)	0.00	0.00	0.00	0.00	0.00	0.17
α-FISP Fairness Violation Rate (%)	0.00	0.00	0.00	0.00	0.00	2.93

Table 1 and 2 show that α-FISP maintains a positive solvency margin across all fairness budgets. In contrast, DFIP fails to maintain a positive solvency margin. This signifies one key merit of α-FISP: explicitly incorporating the critical solvency requirement in risk pricing, which is overlooked in the current fair insurance pricing literature. While Corollary 4.4 suggests no out-of-sample fairness violation, the positive fairness violation rates here do not invalidate our theory. Corollary 4.4 concerns an estimator belonging to the globally fair class $\mathcal{H}_\alpha$. The numerical implementation enforces constraints approximately and may not satisfy them at every test input. The reported violations therefore assess the practical implementation rather than the exact estimator covered by the corollary.



(a) Dataset A1 (b) Dataset A2 (c) Healthcare Insurance Dataset

Figure 4: MSE vs. α Plots

Figure 4a, 4b, and 4c show different behaviors of MSE as α increases. This suggests that creating similar plots may yield some insight into selecting an optimal α that best balances two competing fairness notions. For example, a smaller α may be more defensible for Dataset A, while a larger α may be more defensible for Dataset B since the MSE does not rise significantly even for $\alpha = 0.9$.

6 Conclusion

This paper introduces α-fair insurance pricing, a principled framework that formalizes fairness in insurance as a continuum between actuarial and solidarity perspectives. By directly regulating premium dispersion within risk classes, α-fairness captures the economic substance of fairness in pricing decisions, rather than relying on indirect predictive or demographic criteria. We formulate pricing as a constrained optimization problem, yielding the α-FISP as a flexible and tractable solution. We establish theoretical guarantees and show through numerical experiments that the framework is computationally efficient and compatible with heterogeneous regulatory environments. More broadly, this work reframes fairness in insurance pricing as an allocation problem, offering a transparent and interpretable mechanism for navigating regulatory and ethical trade-offs.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

During manuscript preparation, we used ChatGPT to assist with formulating and revising mathematical statements, developing and checking proof arguments, identifying potential gaps, and improving notation and internal consistency. This assistance included suggestions for alternative proof constructions and feedback on the presentation and interpretation of numerical experiments. We also used the tool to assist with bibliographic checks, language editing, and drafting manuscript passages, including these statements. The authors take full responsibility for the final manuscript, including the correctness and originality of its mathematical arguments, the accuracy of its references, and the validity of its empirical claims. We checked the retained AI-assisted mathematical arguments against the stated assumptions, verified bibliographic information against the original sources, and reviewed the revised text for accuracy.

ETHICS STATEMENT

This work aims to make the trade-off between actuarial and solidarity fairness in insurance pricing explicit and interpretable. The proposed α-fairness criterion regulates premium differences across sensitive groups within a given risk class; it does not provide a universal definition of fairness or determine which rating variables are ethically appropriate. In particular, the designation of variables as non-sensitive does not rule out their use as proxies for sensitive attributes. The choice of α and the construction of risk classes therefore require context-specific consideration of affordability, distributional effects, and stakeholder interests. Our individual-solvency condition concerns conditional expected-loss coverage, rather than guaranteed realized profitability or insurer capital adequacy. The numerical experiments serve as methodological illustrations, not as authorization for operational pricing. Any deployment should be subject to applicable restrictions on sensitive-attribute use, appropriate privacy safeguards, and evaluation of potential harms to affected policyholders.

REPRODUCIBILITY STATEMENT

Section 3 specifies the pricing objective and constraints, and Appendix C provides the proofs of the theoretical results in Section 4. In particular, Appendix C.2 details the statistical assumptions, plug-in estimation conditions, empirical optimization problem, and slack calibration underlying the finite-sample analysis. The simulation response models are presented in Section 5.1 and Appendix B.3, with group-probability assignments in Appendix B.1. Section 5 describes the experimental settings and neural-network architectures, and Appendices B.2–B.4 provide additional numerical results. The supplementary material includes anonymized code for data preparation, model training, evaluation, and figure generation, together with the software dependencies, experimental configurations, and instructions needed to reproduce the reported results.

REFERENCES

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 60–69, 2018.

[2] Carlos Andrés Araiza Iturria, Mary Hardy, and Paul Marriott. A discrimination-free premium under a causal framework. *North American Actuarial Journal*, 28(4):801–821, 2024.

[3] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT press, 2023.

[4] Flavio Calmon, Dennis Wei, Bhaskar Vinzamuri, Karthikeyan N Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, pp. 3992–4001, 2017.

[5] Arthur Charpentier. *Insurance, biases, discrimination and fairness*. Springer, 2024.

[6] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017.

[7] Colorado General Assembly. Restrict insurers' use of external consumer data. Colorado Senate Bill 21-169, 2021 Regular Session, 2021. URL `https://leg.colorado.gov/bills/sb21-169`. Signed July 6, 2021; effective September 7, 2021.

[8] Sam Corbett-Davies, Johann D. Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. The measure and mismeasure of fairness. *Journal of Machine Learning Research*, 24(312): 1–117, 2023. URL `https://jmlr.org/papers/v24/22-1511.html`.

[9] Olivier Côté, Marie-Pier Côté, and Arthur Charpentier. A fair price to pay: Exploiting causal graphs for fairness in insurance. *Journal of Risk and Insurance*, 92(1):33–75, 2025.

[10] Council of the European Union. Council directive 2004/113/ec of 13 december 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services. Official Journal of the European Union L 373 (21 December 2004), 37–43, December 2004. URL `https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32004L0113`. CELEX 32004L0113; commonly referred to as the "Gender Directive".

[11] Lloyd Dixon and James M Anderson. *Public Attitudes Toward Risk-based Pricing and Risk-sharing in Insurance*. RAND Corporation, 2024.

[12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.

[13] Matthias Fahrenwaldt, Christian Furrer, Munir Eberhardt Hiabu, Fei Huang, Frederik Hytting Jørgensen, Mathias Lindholm, Joshua Loftus, Mogens Steffensen, and Andreas Tsanakas. Fairness: plurality, causality, and insurability. *European Actuarial Journal*, 14(2):317–328, 2024.

[14] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 259–268, 2015.

[15] Edward W Frees. Analytics of insurance markets. *Annual Review of Financial Economics*, 7(1): 253–277, 2015.

[16] Edward W.(̃Jed) Frees and Fei Huang. The discriminating (pricing) actuary. *North American Actuarial Journal*, 27(1):2–24, January 2023. doi: 10.1080/10920277.2021.1951296.

[17] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS 2016)*, volume 29, 2016.

[18] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowl. Inf. Syst.*, 33(1):1–33, October 2012. ISSN 0219-1377. doi: 10.1007/s10115-011-0463-8. URL https://doi.org/10.1007/s10115-011-0463-8.

[19] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, pp. 43:1–43:23. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2017. doi: 10.4230/LIPIcs.ITCS.2017.43. URL https://doi.org/10.4230/LIPIcs.ITCS.2017.43.

[20] Xavier Landes. How fair is actuarial fairness? *Journal of Business Ethics*, 128(3):519–533, 2015.

[21] Brett Lantz. *Machine Learning with R*. Packt Publishing, 2013. ISBN 1782162143.

[22] Turo-Kimmo Lehtonen and Jyri Liukko. The forms and limits of insurance solidarity. *Journal of Business Ethics*, 103(Suppl 1):33–44, 2011.

[23] Mathias Lindholm, Ronald Richman, Andreas Tsanakas, and Mario V. Wüthrich. Discrimination-free insurance pricing. *ASTIN Bulletin: The Journal of the IAA*, 52(1), 2022. doi: 10.1017/asb.2021.23.

[24] Mathias Lindholm, Ronald Richman, Andreas Tsanakas, and Mario V Wüthrich. A multi-task network approach for calculating discrimination-free insurance prices. *European Actuarial Journal*, 14(2):329–369, 2024.

[25] Natalia Martinez, Martin Bertran, and Guillermo Sapiro. Minimax pareto fairness: a multi objective perspective. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

[26] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35, 2021.

[27] Aditya Krishna Menon and Robert C. Williamson. The cost of fairness in binary classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, pp. 107–118, 2018.

[28] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8:141–163, 2021.

[29] Weigutian Ou and Helmut Bölcskei. Covering numbers for deep ReLU networks with applications to function approximation and nonparametric regression. *Foundations of Computational Mathematics*, 2026. doi: 10.1007/s10208-026-09762-w.

[30] Michael Rothschild and Joseph Stiglitz. Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. *The Quarterly Journal of Economics*, 90(4): 629–649, 1976. doi: 10.2307/1885326.

[31] Steven Shavell. On moral hazard and insurance. *The Quarterly Journal of Economics*, 93(4): 541–562, 1979.

[32] Iván Sosa and Óscar Montes. Understanding the insurtech dynamics in the transformation of the insurance sector. *Risk Management and Insurance Review*, 25(1):35–68, 2022.

[33] Yves Thiery and Caroline Van Schoubroeck. Fairness and equality in insurance classification. *The Geneva Papers on Risk and Insurance-Issues and Practice*, 31(2):190–211, 2006.

[34] Robert C. Williamson and Aditya Krishna Menon. Fairness risk measures. *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 6786–6797, 2019.

[35] Xi Xin and Fei Huang. Antidiscrimination insurance pricing: Regulations, fairness criteria, and models. *North American Actuarial Journal*, 28(2):285–319, 2024. doi: 10.1080/10920277.2023.2190528.

[36] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pp. 1171–1180, 2017.

[37] Rich Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pp. 325–333, 2013.

[38] T. Zhang, S. Liu, and P. Shi. Discrimination-free insurance pricing with privatized sensitive attributes. *arXiv preprint arXiv:2504.11775*, 2025.

A Appendix A: Deferred Example

A.1 Proposition 4.7 (ii) Example

We present a simple numeric example in this section. Consider a simplified setting with a single risk class (ignoring non-sensitive risk class x for brevity) and a binary sensitive attribute $D \in \{L, H\}$, representing Low-risk and High-risk groups. Assume actuarial fair premiums, $\mu_L = 10$ and $\mu_H = 100$ for each class, and a fairness budget $\alpha = 0.5$. According to Proposition 4.7 (ii), we calculate the sufficient cutoff $p_L \leq \frac{1}{1+\alpha} = \frac{1}{1+0.5}$, which is approximately 0.67. We compare two scenarios varying the population mix $p_L = \mathbb{P}(D = L)$.

Case 1: Condition Satisfied ($p_L = 0.5$). We assume low-risk mass equals to 0.5, which is below the threshold 0.67. Solidarity premium, μ, is calculated through a weighted average: $0.5 \cdot 10 + (1 - 0.5) \cdot 100 = 55$. Recall optimal solution is defined as the minimizer of weighted loss function $L = p_L \cdot (f_L - \mu_L)^2 + p_H \cdot (f_H - \mu_H)^2$. Under α-FISP constraint, we have $f_L = \alpha \cdot f_H$. Then, re-formulate loss function to $L(f_H) = p_L \cdot (\alpha \cdot f_H - \mu_L)^2 + p_H \cdot (f_H - \mu_H)^2$, and take derivative respect to f_H, we have $f_H^* = \dfrac{\alpha p_L \mu_L + p_H \mu_H}{\alpha^2 p_L + p_H}$. Plug-in numerical values, we have $f_H^* = 84$, and $f_L^* = 42$. We can see solvency constraint is naturally satisfied as $0.5 \cdot 84 + 0.5 \cdot 42 = 63 \geq 55$. We also observe that $f_L^* = 42 < 55 = \mu$. The result is intuitive: the fair premium for the low-risk group is subsidized but remains below the population average.

Case 2: Condition Violated ($p_L = 0.9$). Now the low-risk group dominates the population ($0.9 > 0.67$). Use similar formulation as in previous case, μ equals to 19; f_H^* equals to approximately 44.62; f_L^* equals to approximately 22.31. Individual solvency is satisfied ($0.9 \cdot 22.31 + 0.1 \cdot 44.62 \approx 24.54 \geq 19$), but notice that $f_L^* > \mu$. $f_L^* < f_H^*$ confirms result Proposition 4.7 (iii), but low-risk mass domination, under α-fair constraint, leads to price paid by low-risk individuals exceed the solidarity benchmark.

The example distinguishes pairwise order preservation from comparison with the solidarity benchmark. In both cases, the low-risk premium remains below the high-risk premium. When the sufficient mass condition fails, however, the low-risk premium can exceed the solidarity premium.

B Appendix B: Deferred Figures

B.1 Simulation Probability Assignment

In A1 (condition satisfied): $\mathbb{P}(X_{\mathrm{S}} = \mathrm{S}) = 0.5$ and $\mathbb{P}(D = \mathrm{F}|X_{\mathrm{S}} = \mathrm{S}) = \mathbb{P}(D = \mathrm{F}|X_{\mathrm{S}} = \mathrm{NS}) = 0.5$.

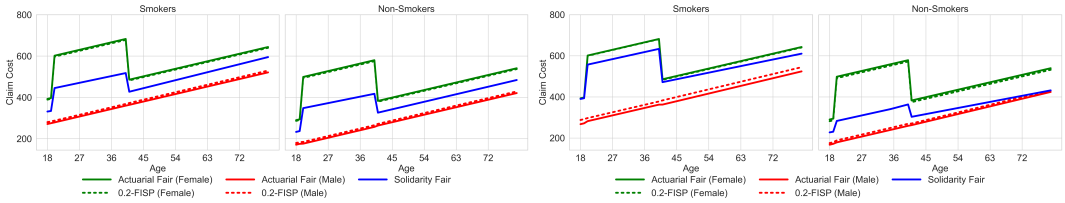
In A2 (condition violated): $\mathbb{P}(X_{\mathrm{S}} = \mathrm{S}) = 0.3$, $\mathbb{P}(D = \mathrm{F}|X_{\mathrm{S}} = \mathrm{S}) = 0.8$, and $\mathbb{P}(D = \mathrm{F}|X_{\mathrm{S}} = \mathrm{NS}) = 0.3$.

In B1 (condition satisfied): $\mathbb{P}(X_{\mathrm{S}} = \mathrm{S}) = 0.5$ and $\mathbb{P}(D = \mathrm{A}|X_{\mathrm{S}} = \mathrm{S}) = \mathbb{P}(D = \mathrm{B}|X_{\mathrm{S}} = \mathrm{S}) = \mathbb{P}(D = \mathrm{C}|X_{\mathrm{S}} = \mathrm{S}) = \frac{1}{3}$.

In B2 (condition violated): $\mathbb{P}(X_{\mathrm{S}} = \mathrm{S}) = 0.5$, $\mathbb{P}(D = \mathrm{A}|X_{\mathrm{S}} = \mathrm{S}) = 0.6$, and $\mathbb{P}(D = \mathrm{B}|X_{\mathrm{S}} = \mathrm{S}) = \mathbb{P}(D = \mathrm{C}|X_{\mathrm{S}} = \mathrm{S}) = 0.2$

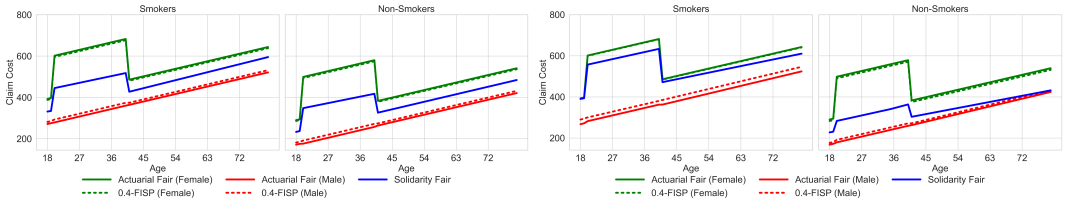
B.2 SIMULATION FIGURES (D = 2)

we present α-FISP under varying fairness budget α for datasets A1 and A2 (i.e. $|\mathcal{D}| = 2$) in the following:



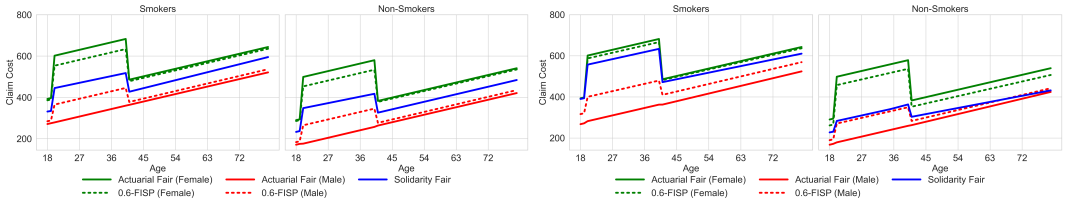
(a) Condition satisfied ($P_L = 0.5 \leq \frac{1}{1+\alpha}$) (b) Condition satisfied ($P_L = 0.7 \leq \frac{1}{1+\alpha}$)

Figure 5: Order preservation of α-fair premiums for $|\mathcal{D}| = 2$ and $\alpha = 0.2$



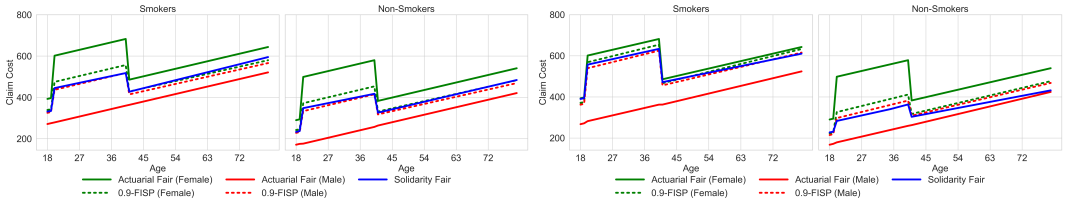
(a) Condition satisfied ($P_L = 0.5 \leq \frac{1}{1+\alpha}$) (b) Condition satisfied ($P_L = 0.7 \leq \frac{1}{1+\alpha}$)

Figure 6: Order preservation of α-fair premiums for $|\mathcal{D}| = 2$ and $\alpha = 0.4$



(a) Condition satisfied ($P_L = 0.5 \leq \frac{1}{1+\alpha}$) (b) Condition violated ($P_L = 0.7 > \frac{1}{1+\alpha}$)

Figure 7: Order preservation of α-fair premiums for $|\mathcal{D}| = 2$ and $\alpha = 0.6$



(a) Condition satisfied ($P_L = 0.5 \leq \frac{1}{1+\alpha}$) (b) Condition violated ($P_L = 0.7 > \frac{1}{1+\alpha}$)

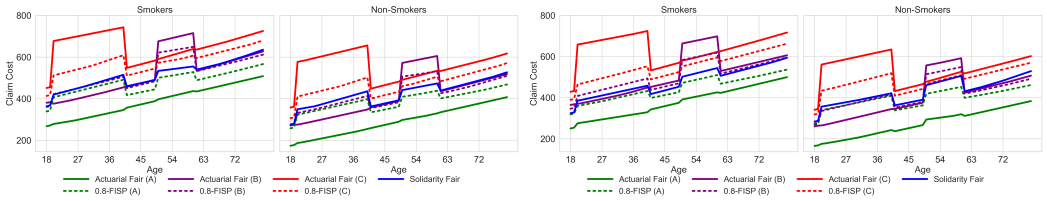
Figure 8: Order preservation of α-fair premiums for $|\mathcal{D}| = 2$ and $\alpha = 0.9$

B.3 SIMULATION FIGURES (D ¿ 2)

Dataset B retains the same response Y and non-sensitive attributes X, but extends the sensitive attributes to three levels (A, B, and C). The data-generating process is:

$$\begin{aligned} Y =& 100 + 4X_{\mathrm{A}} + 100 \cdot \mathbf{1}\{X_{\mathrm{S}} = \mathrm{S}\} + 100 \cdot \mathbf{1}\{D = \mathrm{B}\} \\ &+ 200 \cdot \mathbf{1}\{D = C\} + 200 \cdot \mathbf{1}\{X_{\mathrm{A}} \in [20, 40], D = \mathrm{B}\} \\ &+ 180 \cdot \mathbf{1}\{X_{\mathrm{A}} \in [50, 60], D = \mathbf{C}\} + \epsilon, \epsilon \sim N(0, 40^2) \end{aligned}$$

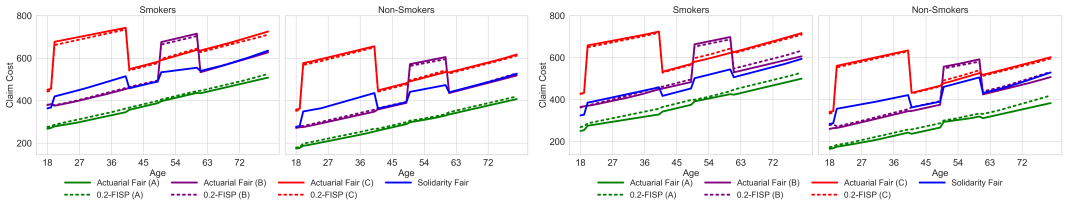
Figure 9 is analogous to Figure 2, but is obtained using Datasets B1 and B2 for panels (a) and (b), respectively, and illustrates the general case with $|\mathcal{D}| > 2$. The contrast is more pronounced in this multi-group setting: when the low-risk condition is violated, premium ordering is no longer preserved, underscoring the necessity of the condition stated in Proposition 4.7.



(a) Condition satisfied ($P_L = \frac{1}{3} \leq \frac{1}{1+\alpha}$) (b) Condition violated ($P_L = 0.6 > \frac{1}{1+\alpha}$)

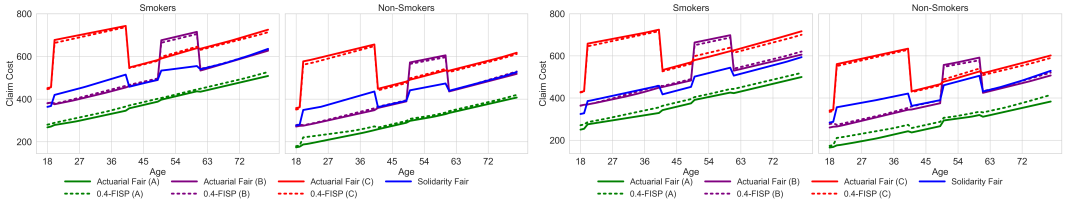
Figure 9: Order preservation of α-fair premiums for $|\mathcal{D}| > 2$ ($\alpha = 0.8$)

We then present α-FISP under varying fairness budget α for datasets B1 and B2 (i.e. $|\mathcal{D}| > 2$) in the following:



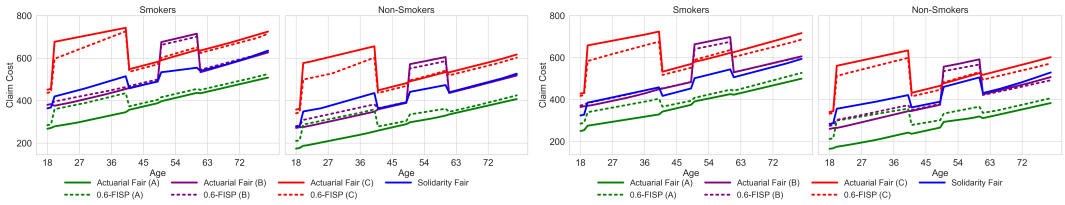
(a) Condition satisfied ($P_L = 0.5 \leq \frac{1}{1+\alpha}$) (b) Condition satisfied ($P_L = 0.6 \leq \frac{1}{1+\alpha}$)

Figure 10: Order preservation of α-fair premiums for $|\mathcal{D}| > 2$ and $\alpha = 0.2$



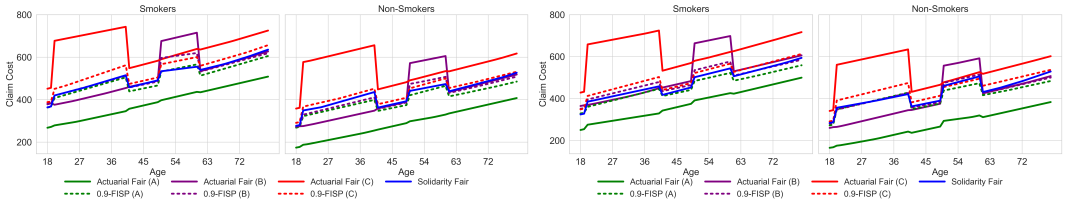
(a) Condition satisfied ($P_L = 0.5 \leq \frac{1}{1+\alpha}$) (b) Condition satisfied ($P_L = 0.6 \leq \frac{1}{1+\alpha}$)

Figure 11: Order preservation of α-fair premiums for $|\mathcal{D}| > 2$ and $\alpha = 0.4$



(a) Condition satisfied ($P_L = 0.5 \leq \frac{1}{1+\alpha}$) (b) Condition satisfied ($P_L = 0.6 \leq \frac{1}{1+\alpha}$)

Figure 12: Order preservation of α-fair premiums for $|\mathcal{D}| > 2$ and $\alpha = 0.6$



(a) Condition satisfied ($P_L = 0.5 \leq \frac{1}{1+\alpha}$) (b) Condition violated ($P_L = 0.6 > \frac{1}{1+\alpha}$)

Figure 13: Order preservation of α-fair premiums for $|\mathcal{D}| > 2$ and $\alpha = 0.9$

B.4 HEALTH INSURANCE FIGURES

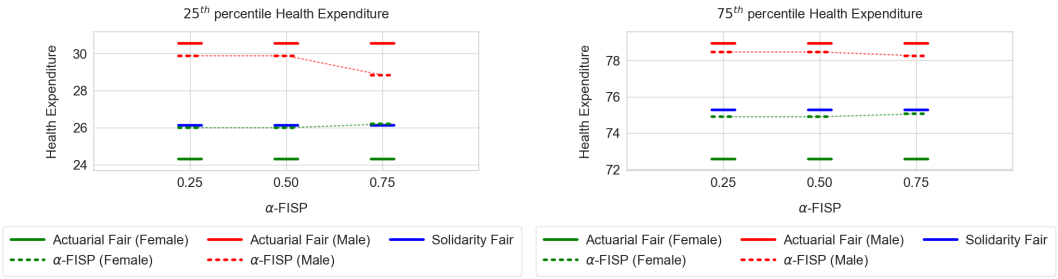


Figure 14: α-fair premiums for health expenditure risk classes at the 25th and 75th percentiles. The panels illustrate the convergence of α-FISP premiums (dashed lines) to the Solidarity Fair benchmark (blue solid line) as the fairness constraint tightens.

C Appendix C: Deferred Proofs

C.1 Proposition 4.1 Proof

Proof. We first show that the feasible set is non-empty. For $x \in \mathcal{X}$, define the candidate function:

$$f_0(x, d_k) = \mu(x),\ k = 1, \ldots, |\mathcal{D}|.$$

Since $\mu(x) = \sum_{k=1}^{|\mathcal{D}|} \mu_k(x) \cdot p_{k|x}$. Therefore, f_0 trivially satisfies IS and α-fairness constraints. Thus, f_0 is feasible. Moreover, we know $f_0 \in \mathcal{F}$, since

$$\begin{aligned}
\mathbb{E}\left[f_0(X, D)^2\right] &= \mathbb{E}\left[\mu(X)^2\right] \\
&= \mathbb{E}\left[\left(\mathbb{E}\left[Y \mid X\right]\right)^2\right] \\
&\leq \mathbb{E}\left[\mathbb{E}\left[Y^2 \mid X\right]\right] \\
&= \mathbb{E}\left[Y^2\right] \\
&< +\infty.
\end{aligned}$$

Therefore, we have $f_0 \in \mathcal{F}_{\alpha,IS}$ and $\mathcal{F}_{\alpha,IS} \neq \emptyset$.

Note that the constraints also imply strict positivity. For any real-valued function f satisfying the constraints, let $M_f(x) := \max_k f(x, d_k)$. Individual solvency gives

$$M_f(x) \geq \sum_{k=1}^{|\mathcal{D}|} p_{k|x} f(x, d_k) \geq \mu(x) \geq \mu_{\min} > 0.$$

Consequently, the α-fairness constraints imply

$$f(x, d_k) \geq \alpha M_f(x) \geq \alpha \mu_{\min} > 0$$

for every k and $\mathbb{P}_X$-almost every x.

Next, we show that the feasible set $\mathcal{F}_{\alpha,IS}$ is convex. Let $f, g \in \mathcal{F}_{\alpha,IS}$ and $t \in [0, 1]$. Define $h := tf + (1-t)g$, then we have for $a.e.\ x$:

$$\begin{aligned}
\sum_{k=1}^{|\mathcal{D}|} (h(x, d_k) - \mu_k(x)) \cdot p_{k|x} &= t \sum_{k=1}^{|\mathcal{D}|} (f(x, d_k) - \mu_k(x)) \cdot p_{k|x} + (1-t) \sum_{k=1}^{|\mathcal{D}|} (g(x, d_k) - \mu_k(x)) \cdot p_{k|x} \\
&\geq 0.
\end{aligned}$$

Also, for any $k < k'$ and $a.e.\ x$, we have:

$$\begin{aligned}
h(x, d_k) - \alpha h(x, d_{k'}) &= t\left(f(x, d_k) - \alpha f(x, d_{k'})\right) + (1-t)\left(g(x, d_k) - \alpha g(x, d_{k'})\right) \\
&\geq 0.
\end{aligned}$$

Therefore, $h \in \mathcal{F}_{\alpha,IS}$ meaning $\mathcal{F}_{\alpha,IS}$ is convex.

Then, we show that $\mathcal{F}_{\alpha,IS}$ is closed in $L^2(\mathbb{P}_{X,D})$.

Let $\{f_n\}_{n \geq 1} \subset \mathcal{F}_{\alpha,IS}$ s.t. $f_n \to f$ in $\mathcal{F} = L^2(\mathbb{P}_{X,D})$. We know that L^2 convergence implies convergence in probability. Thus, we can find a subsequence $\{f_{n_j}\}_{j \geq 1}$ s.t. $f_{n_j} \overset{a.s.}{\to} f$ w.r.t. the probability measure $\mathbb{P}_{X,D}$. That is, there is some $\Omega_0 \subset \mathcal{X} \times \mathcal{D}$ and $\mathbb{P}_{X,D}(\Omega_0) = 1$ s.t.

$$f_{n_j}(x, d) \to f(x, d), \forall (x, d) \in \Omega_0.$$

But, note that IS and α-fairness constraints are evaluated pointwise. We need to leverage the $a.s.$ convergence to ensure for $\mathbb{P}_X$-$a.e.\ x$, the limit function f also satisfies the constraints.

Now, fix $k \in [|\mathcal{D}|]$. Define

$$A_k := \left\{x \in \mathcal{X} \mid f_{n_j}(x, d_k) \to f(x, d_k)\right\}.$$

Since we know $f_{n_j}(x, d) \overset{a.s.}{\to} f(x, d)$, therefore, we have

$$\mathbb{P}_{X,D}\left(\{(x, d) \mid x \in A_k^c, d = d_k\}\right) = 0.$$

By conditioning on X, we get:

$$\mathbb{P}_{X,D}\left(\{(x,d) \mid x \in A_k^c, d = d_k\}\right) = \int_{A_k^c} p_{k|x} d\mathbb{P}_X(x) = 0.$$

But, since $p_{k|x} > 0$, then this forces

$$\mathbb{P}_X(A_k^c) = 0.$$

Therefore, for each $k \in [|\mathcal{D}|]$, we have

$$\mathbb{P}_X(A_k) = 1.$$

Now, define

$$A := \bigcap_{k=1}^{|\mathcal{D}|} A_k,$$

then, this gives

$$\mathbb{P}_X(A) = 1,$$

as A is a finite intersection of probability-1 sets.

Thus, we have shown that $\forall x \in A$,

$$f_{n_j}(x, d_k) \to f(x, d_k), \forall k \in [|\mathcal{D}|].$$

Next, fix a pair $k < k'$. Then, for each subsequence index j, define

$$B_j^{k,k'} := \left\{x \in \mathcal{X} \mid f_{n_j}(x, d_k) - \alpha f_{n_j}(x, d_{k'}) \geq 0 \text{ and } f_{n_j}(x, d_{k'}) - \alpha f_{n_j}(x, d_k) \geq 0\right\}.$$

We know f_{n_j} is feasible, therefore, we have

$$\mathbb{P}_X\left(B_j^{k,k'}\right) = 1.$$

Now, define

$$B^{k,k'} := \bigcap_{j=1}^{\infty} B_j^{k,k'},$$

then, this gives

$$\mathbb{P}_X\left(B^{k,k'}\right) = 1,$$

as $B^{k,k'}$ is a countable intersection of probability-1 sets.

Now, we further define

$$B^{\alpha-\text{fair}} := \bigcap_{k<k'} B^{k,k'},$$

then, this again gives

$$\mathbb{P}_X\left(B^{\alpha-\text{fair}}\right) = 1,$$

since $|\mathcal{D}| < \infty$, there are only finitely many pairs of (k, k').

Similarly, for each j, define

$$B_j^{IS} := \left\{x \in \mathcal{X} \mid \sum_{k=1}^{|\mathcal{D}|} \left(f_{n_j}(x, d_k) - \mu_k(x)\right) \cdot p_{k|x} \geq 0\right\},$$

since f_{n_j} is feasible, we have

$$\mathbb{P}_X\left(B_j^{IS}\right) = 1.$$

Define

$$B^{IS} := \bigcap_{j=1}^{\infty} B_j^{IS},$$

we have

$$\mathbb{P}_X\left(B^{IS}\right) = 1.$$

Finally, define

$$B := B^{\alpha-\text{fair}} \cap B^{IS},$$

we get:

$$\mathbb{P}_X(B) = 1.$$

Thus, we have shown that $\forall x \in B$, all constraints hold for each f_{n_j}.

Note that since $\mathbb{P}_X(A) = \mathbb{P}_X(B) = 1$, thus, we have $\mathbb{P}_X(A \cap B) = 1$. Now, we are ready to take the limit.

Pick any $x \in A \cap B$. For each j, we have:

$$f_{n_j}(x, d_k) - \alpha f_{n_j}(x, d_{k'}) \geq 0,$$

take $j \to \infty$, we have:

$$f(x, d_k) - \alpha f(x, d_{k'}) = \lim_{j\to\infty} \left(f_{n_j}(x, d_k) - \alpha f_{n_j}(x, d_{k'})\right) \geq 0.$$

The same argument will show that the α-fairness constraint is satisfied for the reverse way.

Pick any $x \in A \cap B$. For each j, we have:

$$\sum_{k=1}^{|\mathcal{D}|} \left(f_{n_j}(x, d_k) - \mu_k(x)\right) \cdot p_{k|x} \geq 0.$$

Since $|\mathcal{D}| < \infty$, and $f_{n_j}(x, d_k) \to f(x, d_k)$. By letting $j \to \infty$, we get:

$$\lim_{j\to\infty} \left(\sum_{k=1}^{|\mathcal{D}|} \left(f_{n_j}(x, d_k) - \mu_k(x)\right) \cdot p_{k|x} \right) = \sum_{k=1}^{|\mathcal{D}|} \left(f(x, d_k) - \mu_k(x)\right) \cdot p_{k|x} \geq 0.$$

Therefore, we have proved that $f \in \mathcal{F}_{\alpha,IS}$ and this implies that $\mathcal{F}_{\alpha,IS}$ is closed in $\mathcal{F}$. $\mathcal{F}_{\alpha,IS}$ is also weakly closed since it is also convex.

Next, define the objective

$$\begin{aligned}
\psi(f) &:= \mathbb{E}\left[\left(f(X, D) - Y\right)^2\right] \\
&\geq \frac{1}{2}\mathbb{E}\left[f(X, D)^2\right] - \mathbb{E}\left[Y^2\right] \\
&= \frac{1}{2}\|f\|_2^2 - \mathbb{E}\left[Y^2\right].
\end{aligned}$$

Since $\mathbb{E}\left[Y^2\right] < \infty$, we have $\psi(f) \to \infty$ whenever $\|f\|_2^2 \to \infty$. Thus, ψ is coercive.

To show ψ is weakly LSC, expand the square to get:

$$\begin{aligned}
\psi(f) &= \mathbb{E}\left[\left(f(X, D) - Y\right)^2\right] \\
&= \mathbb{E}\left[f(X, D)^2\right] - 2\mathbb{E}\left[f(X, D)Y\right] + \mathbb{E}\left[Y^2\right] \\
&= \mathbb{E}\left[f(X, D)^2\right] - 2\mathbb{E}\left[f(X, D)\mathbb{E}\left[Y \mid X, D\right]\right] + \mathbb{E}\left[Y^2\right] \\
&= \mathbb{E}\left[f(X, D)^2\right] - 2\mathbb{E}\left[f(X, D)\mu_D(X)\right] + \mathbb{E}\left[Y^2\right].
\end{aligned}$$

But, note that, by Jensen's inequality, we have

$$\mathbb{E}\left[\mu_D(X)^2\right] \leq \mathbb{E}\left[Y^2\right] < \infty,$$

meaning $\mu_D(X) \in L^2(\mathbb{P}_{X,D})$. Thus,

$$\psi(f) = \|f\|_2^2 - 2\langle f, \mu_D\rangle + \mathbb{E}\left[Y^2\right].$$

Since $\|\cdot\|_2^2$ is weakly LSC on $L^2(\mathbb{P}_{X,D})$ and $\langle \cdot, \mu_D\rangle$ is a bounded linear operator (inner product operator), hence weakly continuous. Therefore, ψ is also weakly LSC.

Now, let

$$m := \inf_{f \in \mathcal{F}_{\alpha,IS}} \psi(f).$$

Since $\mathcal{F}_{\alpha,IS} \neq \emptyset$ and $m < \infty$. Take $\{f_n\}_{n\geq 1} \subset \mathcal{F}_{\alpha,IS}$ with $\psi(f_n) \to m$. Then, by coercivity, $\{f_n\}_{n\geq 1}$ is bounded in $\mathcal{F}$. Therefore, we can find a weakly convergent subsequence $f_{n_j} \rightharpoonup f^*$ in $\mathcal{F}$. But, we already proved that $\mathcal{F}_{\alpha,IS}$ is weakly closed. Thus, $f^* \in \mathcal{F}_{\alpha,IS}$. By weakly LSC, we have:

$$\psi(f^*) \leq \liminf_{j\to\infty} \psi(f_{n_j}) = m.$$

Therefore $\psi(f^*) = m$, hence f^* is a minimizer.

To show f^* is the unique minimizer, let $f, g \in \mathcal{F}$ and $t \in (0,1)$, then apply the identity with $a, b \in \mathbb{R}$, $t \in (0,1)$:

$$(ta + (1-t)b)^2 = ta^2 + (1-t)b^2 - t(1-t)(a-b)^2,$$

we get:

$$\psi(tf + (1-t)g) = t\psi(f) + (1-t)\psi(g) - t(1-t)\|f-g\|_2^2.$$

If $f \neq g$ in $L^2(\mathbb{P}_{X,D})$, then $\|f-g\|_2^2 > 0$, resulting in

$$\psi(tf + (1-t)g) < t\psi(f) + (1-t)\psi(g).$$

This implies that ψ is strictly convex.

Now, let f^*, g^* be two distinct minimizers. Take $t = \frac{1}{2}$ and define

$$h := \frac{1}{2}f^* + \frac{1}{2}g^* \in \mathcal{F}_{\alpha,IS}.$$

Since we have proven that $\mathcal{F}_{\alpha,IS}$ is convex. Then, we have

$$\begin{aligned}
\psi(h) &= \frac{1}{2}\psi(f^*) + \frac{1}{2}\psi(g^*) - \frac{1}{4}\|f^* - g^*\|_2^2 \\
&= m - \frac{1}{4}\|f^* - g^*\|_2^2 \\
&< m,
\end{aligned}$$

contradicting the optimality of f^*, g^*, and the infimum m over $F_{\alpha,IS}$.

Therefore, there exists a unique minimizer f^* up to $\mathbb{P}_{X,D}$-$a.s.$ equality. □

C.2 THEOREM 4.2 PROOF

C.2.1 STATISTICAL ASSUMPTIONS

Throughout the proof, fix $\alpha \in (0, 1]$ and $\delta \in (0, 1)$, and let

$$S_n := \{(X_i, D_i, Y_i)\}_{i=1}^n.$$

The notations $\mathcal{H}_\alpha$, $\mathcal{H}_{\alpha,IS}$, and $f^*_{\alpha,IS}$ are defined in the main text.

We have $(X_i, D_i, Y_i) \overset{iid}{\sim} \mathbb{P}_{X,D,Y}$, $i = 1, \ldots, n$ and $\mathcal{D} = \{d_1, \ldots, d_{|\mathcal{D}|}\}$, $|\mathcal{D}| < +\infty$.

There exist constants B, $B_f > 0$ and $0 < \mu_{\min} \le \mu_{\max} \le B$ s.t.

$$0 \le Y \le B \text{ a.s.},$$

and

$$|f(x, d_k)| \le B_f, \text{ for every } f \in \mathcal{H}, x \in \mathcal{X}, k \in [|\mathcal{D}|],$$

and

$$\mu_{\min} \le \mu_k(x) \le \mu_{\max}, \; p_{k|x} > 0,$$

for every $k \in [|\mathcal{D}|]$ and $\mathbb{P}_X$-a.e. x.

The class $\mathcal{H}$ and $\mathcal{T} \subset \{t \mid \mathcal{X} \to [0, 1]\}$ are fixed independently of the observed sample, consist of measurable functions, and $\mathcal{T}$ is nonempty. The classes $\mathcal{H}_\alpha$ and $\mathcal{T}$ are point-wise measurable: each admits a countable subclass such that every member of the original class is the point-wise limit of a sequence from that subclass.

For empirical evaluation, hypothesis functions are represented by fixed measurable versions. Versions of μ_k and $p_{k|}$. are chosen so that their stated bounds and probability-vector properties hold on any exceptional $\mathbb{P}_X$-null set as well.

C.2.2 UNIFORM COVERING NUMBERS

There exist constants $V_F, V_T \ge 1$, $A_F, A_T \ge e$ s.t. for every $\epsilon \in (0, 1]$,

$$\sup_Q \log N\left(\epsilon B_f, \mathcal{H}_\alpha, L^2(Q)\right) \le V_F \log\left(\frac{A_F}{\epsilon}\right),$$

and

$$\sup_{Q_X} \log N\left(\epsilon, \mathcal{T}, L^2(Q_X)\right) \le V_T \log\left(\frac{A_T}{\epsilon}\right),$$

where Q ranges over finitely supported probability measures on $\mathcal{X} \times \mathcal{D}$. Q_X ranges over finitely supported probability measures on $\mathcal{X}$, and $N(\epsilon, G, L^2(Q))$ is the covering number of G w.r.t. $L^2(Q)$ metric.

Remark C.1. *Theorem 2.1 of [29] provides covering-number bounds for fully connected ReLU networks with bounded width, depth, and parameter magnitudes on a bounded input domain. In particular, its L^∞ bound yields uniform $L^2(Q)$ covering number control because*

$$\|f - g\|_{L^2(Q)} \le \|f - g\|_{L^\infty},$$

for every probability measure Q. Therefore, suitable bounded neural-network classes satisfy the hypothesis-class component of Assumption C.2.2, with adjusted constants when restricting to $\mathcal{H}_\alpha$.

C.2.3 PLUG-IN ESTIMATOR ERRORS

The estimators $\hat{\mu}_k(x)$ and $\hat{p}_{k|x}$ are jointly measurable in their data and arguments. Almost surely, for every $x \in \mathcal{X}$,

$$\hat{p}_{k|x} \ge 0, \; \sum_{k=1}^{|\mathcal{D}|} \hat{p}_{k|x} = 1.$$

There exist $\epsilon_\mu, \epsilon_p \ge 0$ s.t.

$$\mathbb{P}(\mathcal{E}_{\text{plug}}) \ge 1 - \frac{\delta}{2},$$

where

$$\mathcal{E}_{\text{plug}} := \left\{ \sup_{x\in\mathcal{X},k\in[|\mathcal{D}|]} |\hat{\mu}_k(x) - \mu_k(x)| \le \epsilon_\mu, \sup_{x\in\mathcal{X}} \sum_{k=1}^{|\mathcal{D}|} \left|\hat{p}_{k|x} - p_{k|x}\right| \le \epsilon_p \right\}$$

C.2.4 Empirical Optimization and Slack Calibration

Define the empirical squared risk

$$\mathcal{R}_n(f) := \frac{1}{n}\sum_{i=1}^{n}(f(X_i, D_i) - Y_i)^2.$$

For $f \in \mathcal{H}_\alpha$, define the population and plug-in conditional solvency residuals,

$$IS_f(x) := \sum_{k=1}^{|\mathcal{D}|} p_{k|x} \cdot (f(x, d_k) - \mu_k(x)),$$

$$\hat{IS}_f(x) := \sum_{k=1}^{|\mathcal{D}|} \hat{p}_{k|x} \cdot (f(x, d_k) - \hat{\mu}_k(x)).$$

For $t \in \mathcal{T}$, let

$$G(f,t) := \mathbb{E}\left[IS_f(X)t(X)\right],$$

$$G_n(f,t) := \frac{1}{n}\sum_{i=1}^{n} IS_f(X_i)t(X_i),$$

$$\hat{G}_n(f,t) := \frac{1}{n}\sum_{i=1}^{n} \hat{IS}_f(X_i)t(X_i).$$

Set

$$M := B + B_f,$$
$$\Gamma := V_F \log(eA_F) + V_T \log(eA_T).$$

For a sufficiently large universal constant $C > 0$, choose

$$\eta_n := M\epsilon_p + \epsilon_\mu + CM\left(\sqrt{\frac{\Gamma}{n}} + \sqrt{\frac{\log(8/\delta)}{n}}\right). \tag{S1}$$

Define the relaxed empirical feasible set

$$\hat{\mathcal{H}}_{\alpha,IS}^{\eta_n} := \left\{ f \in \mathcal{H}_\alpha \mid \hat{G}_n(f,t) \ge -\eta_n, \text{ for every } t \in \mathcal{T} \right\}. \tag{S2}$$

Whenever this set is nonempty, define

$$\hat{f}_{\alpha,IS} \in \arg\min_{f\in\hat{\mathcal{H}}_{\alpha,IS}^{\eta_n}} \mathcal{R}_n(f). \tag{S3}$$

We assume that the event

$$\hat{\mathcal{H}}_{\alpha,IS}^{(\eta_n)} \neq \emptyset$$

is measurable and that, whenever this set is nonempty, the minimum in (S3) is attained and admits a selection jointly measurable in the data and the prediction input. On the event that the empirical feasible set is empty, define $\hat{f}_{\alpha,IS}$ to equal a fixed reference function $h_{\text{ref}} \in \mathcal{H}_\alpha$. The reference function exists because we assumed that $f^*_{\alpha,IS} \in \mathcal{H}$ so $\mathcal{H}_\alpha \neq \emptyset$.

The expected Rademacher complexity is defined as:

$$\Re_n(\mathcal{H}_\alpha) := \mathbb{E}_{S_n^{X,D},\sigma}\left[\sup_{f\in\mathcal{H}_\alpha} \frac{1}{n}\sum_{i=1}^{n}\sigma_i f(X_i, D_i)\right],$$

where $S_n^{X,D} = \{(X_i, D_i)\}_{i=1}^n$ and $\sigma_1, \ldots, \sigma_n$ are independent Rademacher random variables.

C.2.5 PROOF

Proof. **Step 1:** Uniform deviation of the solvency functionals

Decompose

$$G(f,t) - \hat{G}_n(f,t) = (G(f,t) - G_n(f,t)) + (G_n(f,t) - \hat{G}_n(f,t)).$$

Since we know that $|f(x,d_k) - \mu_k(x)| \leq M$, then on the event $\mathcal{E}_{\text{plug}}$, we have:

$$\sup_{f\in\mathcal{H}_\alpha, x\in\mathcal{X}} |\hat{IS}_f(x) - IS_f(x)| \leq M\epsilon_p + \epsilon_\mu.$$

Then, we have:

$$\sup_{f\in\mathcal{H}_\alpha, t\in\mathcal{T}} |G_n(f,t) - \hat{G}_n(f,t)| \leq M\epsilon_p + \epsilon_\mu.$$

For the sampling deviation, consider the fixed class

$$\mathcal{G} := \{x \mapsto IS_f(x)t(x) \mid f \in \mathcal{H}_\alpha, t \in \mathcal{T}\}.$$

Note that the class is bounded by M.

Let Q_X be a finitely supported probability measure on $\mathcal{X}$, and define

$$Q(dx, \{d_k\}) := p_{k|x}Q_X(dx).$$

Then, for $f, g \in \mathcal{H}_\alpha$, we apply Jensen's inequality and get:

$$\begin{aligned}
\|IS_f - IS_g\|^2_{L^2(Q_X)} &= \int \left[\sum_{k=1}^{|\mathcal{D}|} p_{k|x}(f(x,d_k) - g(x,d_k))\right]^2 dQ_X(x) \\
&\leq \int \sum_{k=1}^{|\mathcal{D}|} p_{k|x}(f(x,d_k) - g(x,d_k))^2 dQ_X(x) \\
&= \|f - g\|^2_{L^2(Q)}.
\end{aligned}$$

Thus, for $t, s \in \mathcal{T}$, we have:

$$\|IS_f t - IS_g s\|_{L^2(Q_X)} \leq \|f - g\|_{L^2(Q)} + M\|t - s\|_{L^2(Q_X)}.$$

Then, by Assumption C.2.2, we get:

$$\log N(\epsilon M, G, L^2(Q_X)) \leq V_F \log\left(\frac{2A_F}{\epsilon}\right) + V_T \log\left(\frac{2A_T}{\epsilon}\right), 0 < \epsilon \leq 1.$$

Then, using the entropy-integral bound with the signed class:

$$\mathcal{G}^{\pm} := \mathcal{G} \cup (-\mathcal{G}).$$

By the entropy-integral bound, symmetrization, and bounded-difference concentration, there exists a universal constant $C_1 > 0$ s.t. w.p. at least $1 - \delta/4$:

$$\sup_{f\in\mathcal{H}_\alpha, t\in\mathcal{T}} |G(f,t) - G_n(f,t)| \leq C_1 M\sqrt{\frac{\Gamma}{n}} + M\sqrt{\frac{2\log(4/\delta)}{n}}.$$

Denote this event by $\mathcal{E}_{\text{sample}}$. On $\mathcal{E}_{\text{sample}} \cap \mathcal{E}_{\text{plug}}$, choose a sufficiently large C in S1, we get:

$$\sup_{f\in\mathcal{H}_\alpha, t\in\mathcal{T}} |G(f,t) - \hat{G}_n(f,t)| \leq \eta_n$$

Now, if $f \in \mathcal{H}_{\alpha,IS}$, then $G(f,t) \geq 0$, for every $t \in \mathcal{T}$. Hence

$$\hat{G}_n(f,t) \geq -\eta_n,$$

for every $t \in \mathcal{T}$. This implies:

$$\mathcal{H}_{\alpha,IS} \subset \hat{\mathcal{H}}^{(\eta_n)}_{\alpha,IS}.$$

Then, by the representability assumption, we have

$$f^*_{\alpha,IS} \in \mathcal{H}_{\alpha,IS},$$

which ensures that the empirical feasible set is nonempty on this event.

Step 2: Uniform risk deviation and empirical comparison

Define

$$\mathcal{L}_\alpha := \{(x,d,y) \mapsto (f(x,d)-y)^2 \mid f \in \mathcal{H}_\alpha\}.$$

Then, for $u, v \in [-B_f, B_f]$ and $y \in [0, B]$, we have:

$$\begin{aligned}|(u-y)^2 - (v-y)^2| &= |u-v||u+v-2y| \\ &\leq 2M|u-v|.\end{aligned}$$

By the contraction inequality, we get

$$\Re_n(\mathcal{L}_\alpha) \leq 2M\Re_n(\mathcal{H}_\alpha).$$

Then, applying the bounded-loss Rademacher bound and taking the union bound, we have w.p. at least $1-\delta/4$:

$$\Delta_n := \sup_{f\in\mathcal{H}_\alpha} |\mathcal{R}(f) - \mathcal{R}_n(f)| \leq 4M\Re_n(\mathcal{H}_\alpha) + 3M^2\sqrt{\frac{\log(8/\delta)}{2n}}.$$

On the event from step 1, we have

$$\mathcal{R}_n(\hat{f}_{\alpha,IS}) \leq \mathcal{R}_n(f^*_{\alpha,IS}).$$

Thus,

$$\begin{aligned}\mathcal{R}(\hat{f}_{\alpha,IS}) - \mathcal{R}(f^*_{\alpha,IS}) =& \mathcal{R}(\hat{f}_{\alpha,IS}) - \mathcal{R}_n(\hat{f}_{\alpha,IS}) \\ &+ \mathcal{R}_n(\hat{f}_{\alpha,IS}) - \mathcal{R}_n(f^*_{\alpha,IS}) \\ &+ \mathcal{R}_n(f^*_{\alpha,IS}) - \mathcal{R}(f^*_{\alpha,IS}) \\ \leq& 2\Delta_n.\end{aligned}$$

Step 3: Fairness Cost

Define:

$$\begin{aligned}m(x) &:= \min_k \mu_k(x), \\ u(x) &:= \max_k \mu_k(x),\end{aligned}$$

and

$$g_\alpha(x,d_k) := \min\left\{\mu_k(x), \frac{m(x)}{\alpha}\right\}.$$

Set

$$\begin{aligned}r_k(x) &:= \mu_k(x) - g_\alpha(x,d_k) \geq 0, \\ \lambda(x) &:= \sum_{k=1}^{|\mathcal{D}|} p_{k|x} r_k(x),\end{aligned}$$

and

$$\tilde{g}_\alpha(x,d_k) := g_\alpha(x,d_k) + \lambda(x).$$

Since

$$m(x) \leq g_\alpha(x,d_k) \leq \frac{m(x)}{\alpha}.$$

Then, by adding the nonnegative term $\lambda(x)$, we have:

$$\tilde{g}_\alpha(x,d_k) - \alpha\tilde{g}_\alpha(x,d_j) = g_\alpha(x,d_k) - \alpha g_\alpha(x,d_j) + (1-\alpha)\lambda(x) \geq 0.$$

Moreover,

$$\sum_{k=1}^{|\mathcal{D}|} p_{k|x}(\tilde{g}_\alpha(x, d_k) - \mu_k(x)) = -\sum_{k=1}^{|\mathcal{D}|} p_{k|x} r_k(x) + \lambda(x) = 0.$$

Since $\tilde{g}_\alpha$ is bounded, hence square-integrable. Then, we get

$$\tilde{g}_\alpha \in \mathcal{F}_{\alpha,IS}.$$

Now, conditional on $X = x$, we can write:

$$\begin{aligned} \tilde{g}_\alpha(x, D) - \mu_D(x) &= \lambda(x) - r_D(x), \\ \lambda(x) &= \mathbb{E}[r_D(X) \mid X = x]. \end{aligned}$$

Then, we have

$$\begin{aligned} \mathbb{E}\left[(\tilde{g}_\alpha(X, D) - \mu_D(X))^2 \mid X = x\right] &= Var(r_D(X) \mid X = x) \\ &\leq \sum_{k=1}^{|\mathcal{D}|} p_{k|x} r_k(x)^2. \end{aligned}$$

Then, for each k,

$$\begin{aligned} r_k(x) &\leq \left(u(x) - \frac{m(x)}{\alpha}\right)_+ \\ &= m(x)\left(\frac{u(x)}{m(x)} - \frac{1}{\alpha}\right)_+ \\ &\leq \mu_{\max}(R_\mu - 1/\alpha)_+. \end{aligned}$$

Therefore,

$$\begin{aligned} \mathcal{R}(f^*_{\alpha,IS}) - \mathcal{R}(\mu_D) &\leq \mathcal{R}(\tilde{g}_\alpha) - \mathcal{R}(\mu_D) \\ &= \mathbb{E}\left[(\tilde{g}_\alpha(X, D) - \mu_D(X))^2\right] \\ &\leq \mu^2_{\max}\left[\left(R_\mu - \frac{1}{\alpha}\right)_+\right]^2 \end{aligned}$$

Step 4: Combining Results

Let $\mathcal{E}_{\text{risk}}$ denote the risk-deviation event in Step 2, by a union bound, we have:

$$\mathbb{P}(\mathcal{E}_{\text{risk}} \cap \mathcal{E}_{\text{sample}} \cap \mathcal{E}_{\text{plug}}) \geq 1 - \frac{\delta}{4} - \frac{\delta}{4} - \frac{\delta}{2} = 1 - \delta.$$

Combining the results, we have:

$$\begin{aligned} \mathcal{R}(\hat{f}_{\alpha,IS}) - \mathcal{R}(\mu_D) &= \mathcal{R}(\hat{f}_{\alpha,IS}) - \mathcal{R}(f^*_{\alpha,IS}) + \mathcal{R}(f^*_{\alpha,IS}) - \mathcal{R}(\mu_D) \\ &\leq 2\Delta_n + \mu^2_{\max}\left[\left(R_\mu - \frac{1}{\alpha}\right)_+\right]^2 \\ &\leq 8M\Re_n(\mathcal{H}_\alpha) + 6M^2\sqrt{\frac{\log(8/\delta)}{2n}} + \mu^2_{\max}\left[\left(R_\mu - \frac{1}{\alpha}\right)_+\right]^2. \end{aligned}$$

□

C.3 COROLLARY 4.4 PROOF

Proof. For simplicity, write $\hat{f} = \hat{f}_{\alpha,IS}$ with all test expectations evaluated holding the fitted estimator fixed. Since $\hat{f} \in \mathcal{H}_\alpha$, including under the reference-function convention, (4)- (5) hold $\mathbb{P}_X$-a.e.

For solvency, define

$$a(x) := \left(\mu(x) - \sum_{k=1}^{|\mathcal{D}|} p_{k|x} \hat{f}(x, d_k) \right)_+,$$
$$g(x, d_k) := \hat{f}(x, d_k) + a(x).$$

Since $\hat{f}$ and a are bounded and measurable, g is measurable and square-integrable. Moreover, we have

$$\begin{aligned} g(x, d_k) - \alpha g(x, d_j) &= \hat{f}(x, d_k) - \alpha \hat{f}(x, d_j) + (1 - \alpha) a(x) \\ &\geq 0, \end{aligned}$$

and,

$$\begin{aligned} \sum_{k=1}^{|\mathcal{D}|} p_{k|x} g(x, d_k) &= \max \left\{ \sum_{k=1}^{|\mathcal{D}|} p_{k|x} \hat{f}(x, d_k), \mu(x) \right\} \\ &\geq \mu(x). \end{aligned}$$

Therefore, $g \in \mathcal{F}_{\alpha,IS}$. However, g need not belong to $\mathcal{H}$.

Now, expand the square loss and, conditioning on X, we have

$$\begin{aligned} \mathcal{R}(g) - \mathcal{R}(\hat{f}) &= 2\mathbb{E}[a(X)(\hat{f}(X, D) - Y)] + \mathbb{E}_{X \sim \mathbb{P}_X}[a(X)^2] \\ &= -\mathbb{E}_{X \sim \mathbb{P}_X}[a(X)^2]. \end{aligned}$$

Then, the population optimality gives

$$\mathbb{E}_{X \sim \mathbb{P}_X}[a(X)^2] \leq \mathcal{R}(\hat{f}) - \mathcal{R}(f^*_{\alpha,IS}).$$

The intermediate risk comparison in the proof of Theorem 4.2 establishes that, on an event w.p. at least $1 - \delta$,

$$\begin{aligned} \mathbb{E}_{X \sim \mathbb{P}_X}[a(X)^2] &\leq \mathcal{R}(\hat{f}) - \mathcal{R}(f^*_{\alpha,IS}) \\ &\leq 8M\Re_n(\mathcal{H}_\alpha) + 6M^2 \sqrt{\frac{\log(8/\delta)}{2n}}. \end{aligned}$$

On this same event, applying Markov's inequality and we get that for every $\tau > 0$,

$$\begin{aligned} \mathbb{P}_X \left(\sum_{k=1}^{|\mathcal{D}|} p_{k|X} \hat{f}(X, d_k) < \mu(X) - \tau \right) &= \mathbb{P}_X(a(X) > \tau) \\ &\leq \frac{\mathbb{E}_{X \sim \mathbb{P}_X}[a(X)^2]}{\tau^2}. \end{aligned}$$

Substituting the preceding bound proves the claim simultaneously for all $\tau > 0$. The $O(n^{-1/2})$ rate follows from the uniform covering-number assumption (Assumption C.2.2), with the stated parameters and class-complexity constants fixed. □

C.4 PROPOSITION 4.5 PROOF

Proof. Let

$$m(x) := \min_k \mu_k(x), \text{ and } M(x) := \max_k \mu_k(x).$$

We will first show that

$$\min_k \mu_k(x) \leq f^*(x, \cdot),$$

for $\mathbb{P}_X$-$a.e.$ x.

Similar to the proof of Proposition 4.7. Define

$$S_L := \{x \mid \exists k \text{ s.t. } f^*(x, d_k) < m(x)\},$$

and show that $\mathbb{P}_X(S_L) = 0$.

For each k, define

$$S_{L,k} := \{x \mid f^*(x, d_k) < m(x)\}.$$

Then, we know that

$$S_L = \bigcup_{k=1}^{|\mathcal{D}|} S_{L,k},$$

then given $|\mathcal{D}| < \infty$, if $\mathbb{P}_X(S_L) > 0$, there must be some k_0 s.t. $\mathbb{P}_X(S_{L,k_0}) > 0$.

Assume for contradiction that such k_0 exists.

Define

$$F(x) := \max_k f^*(x, d_k),$$

then, under IS, we have:

$$\sum_{k=1}^{|\mathcal{D}|} f^*(x, d_k) \cdot p_{k|x} \geq \sum_{k=1}^{|\mathcal{D}|} \mu_k(x) \cdot p_{k|x} = \mu(x) \geq m(x).$$

Thus, we have $F(x) \geq m(x)$ for $\mathbb{P}_X$-$a.e.$ x.

Now, define

$$\epsilon(x) := \mathbf{1}_{S_{L,k_0}}(x) \cdot \min\{m(x) - f^*(x, d_{k_0}), F(x) - f^*(x, d_{k_0})\}.$$

$\epsilon(x) > 0$, since we assumed that $f^*(x, d_{k_0}) < m(x)$, for $x \in S_{L,k_0}$.

We now construct $\tilde{f}$ as follows:

$$\tilde{f}(x, d_k) := \begin{cases} f^*(x, d_{k_0}) + \epsilon(x), & \text{if } k = k_0, \\ f^*(x, d_k) & \text{, if } k \neq k_0. \end{cases}$$

Then, we show that $\tilde{f} \in \mathcal{F}$ and that $\tilde{f}$ is feasible.

Since $f^* \in \mathcal{F}$ and f^* is feasible. Then, enough to bound $\epsilon(x)$.

$$\begin{aligned} \mathbb{E}\left[\left(\tilde{f}(X, D) - f^*(X, D)\right)^2\right] &= \mathbb{E}\left[\epsilon(X)^2 \cdot \mathbf{1}_{\{D=d_{k_0}\}}\right] \\ &\leq \mathbb{E}\left[(\mu_{k_0}(X) - f^*(X, d_{k_0}))^2 \cdot \mathbf{1}_{\{D=d_{k_0}\}}\right] \\ &\leq 2\mathbb{E}\left[\mu_{k_0}(X)^2 \cdot \mathbf{1}_{\{D=d_{k_0}\}}\right] + 2\mathbb{E}\left[f^*(X, d_{k_0})^2 \cdot \mathbf{1}_{\{D=d_{k_0}\}}\right] \\ &< \infty, \end{aligned}$$

since we have shown that $\mathbb{E}\left[\mu_D(X)^2\right] \leq \mathbb{E}\left[Y^2\right] < \infty$ and $f^* \in \mathcal{F}$.

Since $\epsilon(x) > 0$, IS is preserved.

For α-fairness:

If $x \notin S_{L,k_0}$, $\tilde{f}(x,\cdot) = f^*(x,\cdot)$.

If $x \in S_{L,k_0}$, then we increased the k_0 coordinate and capped it at $F(x)$. Thus, we know the maximum of $\tilde{f}(x,\cdot)$ is also $F(x)$. Since f^* is already α-fair, then we must have $f^*(x,d_k) \geq \alpha F(x)$. Thus, with $\epsilon(x) > 0$, we have for all k,

$$\alpha F(x) \leq \tilde{f}(x,d_k) \leq F(x).$$

Therefore, $\tilde{f} \in \mathcal{F}$ and is feasible.

Since f^* and $\tilde{f}$ differ only on $\{X \in S_{L,k_0}, D = d_{k_0}\}$, thus, we have:

$$\psi(\tilde{f}) - \psi(f^*) = \mathbb{E}\left[\mathbf{1}_{\{X \in S_{L,k_0}, D=d_{k_0}\}} \cdot \left(\left(\tilde{f}(X,d_{k_0}) - \mu_{k_0}(X)\right)^2 - \left(f^*(X,d_{k_0}) - \mu_{k_0}(X)\right)^2\right)\right],$$

where $\psi(f) := \mathbb{E}\left[(f(X,D) - Y)^2\right]$ (same as defined in previous proofs).

For $x \in S_{L,k_0}$, define

$$g(x) := \mu_{k_0}(x) - f^*(x,d_{k_0}) > 0.$$

Also, we have

$$\tilde{f}(x,d_{k_0}) = f^*(x,d_{k_0}) + \epsilon(x) \leq m(x) \leq \mu_{k_0}(x).$$

Thus, we get

$$0 < \epsilon(x) \leq g(x).$$

Point-wise, we have:

$$\begin{aligned}\left(\tilde{f}(x,d_{k_0}) - \mu_{k_0}(x)\right)^2 - \left(f^*(x,d_{k_0}) - \mu_{k_0}(x)\right)^2 &= (-g(x) + \epsilon(x))^2 - (-g(x))^2 \\ &= \epsilon(x)^2 - 2\epsilon(x)g(x) \\ &< 0.\end{aligned}$$

Therefore, we get:

$$\mathbb{P}(\{X \in S_{L,k_0}, D = d_{k_0}\}) = \int_{x \in S_{L,k_0}} p_{k_0|x} d\mathbb{P}_X(x) > 0.$$

But, this immediately implies that

$$\psi(\tilde{f}) < \psi(f^*),$$

contradicting the optimality of f^*, as desired.

Now, we show that

$$f^*(x,\cdot) \leq \max_k \mu_k(x),$$

for $\mathbb{P}_X$-$a.e.$ x.

Define

$$S_H := \{x \mid \exists k \text{ s.t. } f^*(x,d_k) > M(x)\},$$

similarly, we show $\mathbb{P}_X(S_H) = 0$.

Assume for contradiction that $\mathbb{P}_X(S_H) > 0$.

Define

$$0 < g(x,d_k) := \min\{f^*(x,d_k), M(x)\} \leq f^*(x,d_k).$$

Therefore

$$\mathbb{E}[g(X,D)^2] \leq \mathbb{E}[f^*(X,D)^2] < +\infty.$$

Then, we know $g \in \mathcal{F}$.

We now argue that g is α-fair. Fix x.

If $F(x) \leq M(x)$, then $g(x,\cdot) = f^*(x,\cdot)$, α-fairness is preserved.

If $F(x) > M(x)$, then at least one coordinate is capped. Thus, we have:

$$\max_k g(x,d_k) = M(x).$$

Since f^* is α-fair, thus

$$f^*(x, d_k) \geq \alpha F(x) \geq \alpha M(x).$$

If $f^*(x, d_k) > M(x)$, then $g(x, d_k) = M(x) \geq \alpha M(x)$.

If $f^*(x, d_k) \leq M(x)$, then $g(x, d_k) = f^*(x, d_k) \geq \alpha F(x) > \alpha M(x)$. Hence, $\min_k g(x, d_k) \geq \alpha M(x) = \alpha \max_k g(x, d_k)$.

Therefore, g is also α-fair.

For IS, define

$$s_g(x) := \sum_{k=1}^{|\mathcal{D}|} (g(x, d_k) - \mu_k(x)) \cdot p_{k|x}.$$

If $s_g(x) \geq 0$ for $\mathbb{P}_X$-a.e. x, then g is feasible, and it is enough to get a contradiction by showing that $\psi(g) < \psi(f^*)$.

However, recall that in the construction of g, the cap brought down the premium for some coordinates. Thus, IS is not guaranteed. Thus, $s_g < 0$ may happen on some set.

Let

$$T := \{x \in \mathcal{X} \mid s_g(x) < 0\},$$

then, on T, define the deficit

$$\text{def}(x) := -s_g(x) = \sum_{k=1}^{|\mathcal{D}|} (\mu_k(x) - g(x, d_k)) \cdot p_{k|x} > 0,$$

and the corresponding positive correction as:

$$\text{pos}(x) := \sum_{k=1}^{|\mathcal{D}|} (\mu_k(x) - g(x, d_k))_+ \cdot p_{k|x}.$$

Note that if $\text{def}(x) > 0$, then at least one of the $\mu_k(x) - g(x, d_k), k = 1, \ldots, |\mathcal{D}|$ is strictly positive. Thus, $\text{pos}(x) > 0$. Also, note that $\text{def}(x) \leq \text{pos}(x)$ by the presence of $(\cdot)_+$ in $\text{pos}(\cdot)$. Therefore, we have the ratio:

$$t(x) := \begin{cases} \frac{\text{def}(x)}{\text{pos}(x)}, x \in T, \\ 0 \qquad , x \notin T. \end{cases}$$

Now, define

$$h(x, d_k) := g(x, d_k) + t(x)(\mu_k(x) - g(x, d_k))_+.$$

We will verify that $h \in \mathcal{F}$ and is feasible.

For each (x, d_k):

if $\mu_k(x) \leq g(x, d_k)$ then $h = g$.

If $\mu_k(x) > g(x, d_k)$, then

$$h(x, d_k) = (1 - t(x))g(x, d_k) + t(x)\mu_k(x)$$

is a convex combination of $g(x, d_k)$ and $\mu_k(x)$. Hence, point-wise, we have:

$$|h(x, d_k)| \leq |g(x, d_k)| + |\mu_k(x)|.$$

Then, we have:

$$h(X, D)^2 \leq 2g(X, D)^2 + 2\mu_D(X)^2,$$

since $(a + b)^2 \leq 2a^2 + 2b^2$. But, we know that $g \in \mathcal{F}$ and $\mu_D \in \mathcal{F}$. Therefore, $h \in \mathcal{F}$.

For IS, we have:

$$\begin{aligned} \sum_{k=1}^{|\mathcal{D}|} (h(x, d_k) - \mu_k(x)) \cdot p_{k|x} &= \sum_{k=1}^{|\mathcal{D}|} (g(x, d_k) - \mu_k(x)) \cdot p_{k|x} + t(x) \sum_{k=1}^{|\mathcal{D}|} (\mu_k(x) - g(x, d_k))_+ \cdot p_{k|x} \\ &= s_g(x) + t(x)\text{pos}(x). \end{aligned}$$

If $x \notin T$, then $t(x) = 0$, $s_g(x) \geq 0$.

If $x \in T$, then

$$s_g(x) + t(x)\text{pos}(x) = -\text{def}(x) + \frac{\text{def}(x)}{\text{pos}(x)} \cdot \text{pos}(x) = 0.$$

For α-fairness:

Since $g = f^*$ on S_H^c, feasibility of f^* implies $s_g(x) \geq 0$ there. Hence $T \subseteq S_H$ up to a null set, and $h = g = f^*$ on S_H^c, so fairness is preserved on that set. On S_H, we have

$$\alpha M(x) \leq g\,(x, d_k) \leq M(x) \quad \text{for every } k.$$

Since $t(x) \in [0, 1]$, each coordinate of h is either unchanged or increased toward $\mu_k(x) \leq M(x)$. Therefore,

$$\alpha M(x) \leq g\,(x, d_k) \leq h\,(x, d_k) \leq M(x) \quad \text{for every } k.$$

Thus, h satisfies the α-fairness constraints.

Thus, $h \in \mathcal{F}$ and is feasible.

Now, we show that h strictly decreases the objective.

Fix (x, d_k):

If $g(x, d_k) \geq \mu_k(x)$, then $h(x, d_k) = g(x, d_k)$.

If $g(x, d_k) < \mu_k(x)$, then

$$h(x, d_k) = (1 - t(x))g(x, d_k) + t(x)\mu_k(x).$$

Then, we have:

$$\mu_k(x) - h(x, d_k) = (1 - t(x))(\mu_k(x) - g(x, d_k)),$$

which implies

$$\begin{aligned}(h(x, d_k) - \mu_k(x))^2 &= (1 - t(x))^2(g(x, d_k) - \mu_k(x))^2 \\ &\leq (g(x, d_k) - \mu_k(x))^2.\end{aligned}$$

Thus, we have

$$\psi(h) \leq \psi(g).$$

Then, enough to show that g strictly decreases the objective.

Define the event

$$A := \{f^*(X, D) > M(X)\}.$$

Then, for any $x \in S_H$, there exists at least one k with $f^*(x, d_k) > M(x)$. Then, we have:

$$\sum_{k=1}^{|\mathcal{D}|} \mathbf{1}_{\{f^*(x,d_k) > M(x)\}} \cdot p_{k|x} > 0,$$

since $p_{k|x} > 0$. Then

$$\begin{aligned}\mathbb{P}(A) &= \int_{x \in S_H} \sum_{k=1}^{|\mathcal{D}|} \mathbf{1}_{\{f^*(x,d_k) > M(x)\}} \cdot p_{k|x} \cdot d\mathbb{P}_X(x) \\ &> 0.\end{aligned}$$

On A, we have $g(X, D) = M(X) < f^*(X, D)$ and $\mu_D(X) \leq M(X)$. Then, point-wise we have:

$$\begin{aligned}(g(x, d_k) - \mu_k(x))^2 - (f^*(x, d_k) - \mu_k(x))^2 &= (M(x) - \mu_k(x))^2 - (f^*(x, d_k) - \mu_k(x))^2 \\ &< 0.\end{aligned}$$

This immediately gives

$$\psi(h) \leq \psi(g) < \psi(f^*),$$

contradicting the optimality of f^*, as desired.

This completes the proof. □

C.5 Proposition 4.7 Proof

C.5.1 Proposition 4.7 (i) Proof

Proof. We will prove by contradiction.

Fix a group index k and define

$$S := \{\mu_k(x) \geq \mu(x) \text{ and } f^*(x, d_k) < \mu(x)\}.$$

Then, enough to show that $\mathbb{P}_X(S) = 0$.

Assume for contradiction that $\mathbb{P}_X(S) > 0$.

Note that for $\mathbb{P}_X$-$a.e.$ x, by IS, we have:

$$\sum_{k=1}^{|\mathcal{D}|} (f^*(x, d_k) - \mu_k(x)) \cdot p_{k|x} \geq 0 \iff \sum_{k=1}^{|\mathcal{D}|} f^*(x, d_k) \cdot p_{k|x} \geq \sum_{k=1}^{|\mathcal{D}|} \mu_k(x) \cdot p_{k|x}.$$

This implies

$$\sum_{k=1}^{|\mathcal{D}|} f^*(x, d_k) \cdot p_{k|x} \geq \mu(x).$$

Let $M(x) := \max_k f^*(x, d_k)$. Then, we know that

$$\sum_{k=1}^{|\mathcal{D}|} f^*(x, d_k) \cdot p_{k|x} \leq M(x),$$

since the average is always upper-bounded by the maximum. Therefore, for any $x \in S$,

$$M(x) \geq \mu(x) > f^*(x, d_k)$$

with the strict inequalities:

$$M(x) - f^*(x, d_k) > 0,$$

and

$$\mu_k(x) - f^*(x, d_k) > 0.$$

Define

$$\Delta_M(x) := M(x) - f^*(x, d_k)$$
$$\Delta_\mu(x) := \mu_k(x) - f^*(x, d_k),$$

and the perturbation

$$\epsilon(x) := \mathbf{1}_S(x) \min\{\Delta_M(x), \Delta_\mu(x)\}.$$

Then, we construct an alternative solution $\tilde{f}$ as follows:

$$\tilde{f}(x, d_j) := \begin{cases} f^*(x, d_k) + \epsilon(x), & \text{if } j = k \\ f^*(x, d_j) & \text{, if } j \neq k. \end{cases}$$

Then, we verify that $\tilde{f} \in \mathcal{F}$.

Note that $\tilde{f}$ and f^* only differ at coordinate k. Therefore, enough to show the differed part is square-integrable. Concretely, the differed part is:

$$\mathbb{E}\left[\left(\tilde{f}(X, D) - f^*(X, D)\right)^2\right] = \mathbb{E}\left[\epsilon(X)^2 \mathbf{1}_{\{D=d_k\}}\right].$$

Note that

$$\epsilon(x) \leq \mathbf{1}_S(x)\Delta_\mu(x) = \mathbf{1}_S(x)\left(\mu_k(x) - f^*(x, d_k)\right),$$

then, we have:

$$\begin{aligned}\mathbb{E}\left[\epsilon(X)^2\mathbf{1}_{\{D=d_k\}}\right] &\leq \mathbb{E}\left[(\mu_k(X)-f^*(X,d_k))^2\cdot\mathbf{1}_{\{D=d_k\}}\right]\\ &\leq 2\mathbb{E}\left[\mu_k(X)^2\cdot\mathbf{1}_{\{D=d_k\}}\right]+2\mathbb{E}\left[f^*(X,d_k)^2\cdot\mathbf{1}_{\{D=d_k\}}\right]\\ &<\infty,\end{aligned}$$

since in Proposition 4.1, we have shown that $\mathbb{E}\left[\mu_D(X)^2\right]\leq\mathbb{E}\left[Y^2\right]<\infty$, and $f^*\in\mathcal{F}$. Therefore, we have $\tilde{f}\in\mathcal{F}$.

Next, we verify that $\tilde{f}$ is feasible.

For IS, we have:

$$\begin{aligned}\sum_{k=1}^{|\mathcal{D}|}\tilde{f}(x,d_k)\cdot p_{k|x} &= \sum_{k=1}^{|\mathcal{D}|}f^*(x,d_k)\cdot p_{k|x}+\epsilon(x)\cdot p_{k|x}\\ &\geq \mu(x)+\epsilon(x)\cdot p_{k|x}\\ &\geq \mu(x).\end{aligned}$$

For α-fairness.

If $x\notin S$, then $\tilde{f}(x,\cdot)=f^*(x,\cdot)$, the constraints are satisfied since f^* is feasible.

Now, fix $x\in S$, then,

$$\begin{aligned}\tilde{f}(x,d_k) &= f^*(x,d_k)+\epsilon(x)\\ &\leq f^*(x,d_k)+\Delta_M(x)\\ &= M(x).\end{aligned}$$

This implies that the maximum of $\tilde{f}(x,\cdot)$ is also upper-bounded by $M(x)$.

Let j^* be the index s.t. $M(x)=f^*(x,d_{j^*})$, then, for every j, we have

$$f^*(x,d_j)\geq\alpha f^*(x,d_{j^*})=\alpha M(x).$$

Therefore, for $j\neq k$, we have:

$$\tilde{f}(x,d_j)=f^*(x,d_{j^*})\geq\alpha M(x).$$

Therefore, for all j, we have:

$$\alpha M(x)\leq\tilde{f}(x,d_j)\leq M(x).$$

Thus, for any pair of j,l,

$$\tilde{f}(x,d_j)-\alpha\tilde{f}(x,d_l)\geq\alpha M(x)-\alpha M(x)=0,$$

and the same holds for the reverse way as one can easily verify.

Now, it remains to show that $\tilde{f}$ strictly decreases the objective.

Let

$$\psi(f):=\mathbb{E}\left[(f(X,D)-Y)^2\right].$$

Note that by conditioning, we have:

$$\mathbb{E}\left[(f(X,D)-Y)^2\mid X,D\right]=\left(f(X,D)-\mu_D(X)\right)^2+Var(Y\mid X,D),$$

but the variance term does not depend on f. Thus, enough to show that the squared bias term is decreased with $\tilde{f}$.

Since $\tilde{f}$ and f^* differ only on $\{D=k\}\cap\{X\in S\}$, so, we get:

$$\psi(\tilde{f})-\psi(f^*)=\mathbb{E}\left[\mathbf{1}_{\{X\in S,D=d_k\}}\left(\left(\tilde{f}(X,d_k)-\mu_k(X)\right)^2-\left(f^*(X,d_k)-\mu_k(X)\right)^2\right)\right].$$

Recall that on $\{X\in S\}$, we have $\tilde{f}(X,d_k)=f^*(X,d_k)+\epsilon(X)$.

Define

$$g(X) := \mu_k(X) - f^*(X, d_k).$$

Then, we have $g(X) > 0$ and $\epsilon(X) \leq g(X)$ on S.

Point-wise, we have:

$$\begin{aligned}\left(\tilde{f}(x, d_k) - \mu_k(x)\right)^2 - (f^*(x, d_k) - \mu_k(x))^2 &= (f^*(x, d_k) + \epsilon(x) - \mu_k(x))^2 - (f^*(x, d_k) - \mu_k(x))^2 \\ &= (-g(x) + \epsilon(x))^2 - (-g(x))^2 \\ &= \epsilon(x)^2 - 2\epsilon(x)g(x).\end{aligned}$$

But, since $0 < \epsilon(x) \leq g(x)$ for every $x \in S$. Thus, we have $\epsilon(x) - 2g(x) \leq -g(x) < 0$. Hence, for $x \in S$, we have

$$\epsilon(x)^2 - 2\epsilon(x)g(x) < 0.$$

Therefore, since we assumed $\mathbb{P}_X(S) > 0$. But, we also know that $p_{k|x} > 0$ for $\mathbb{P}_X$-$a.e.\ x$. Thus,

$$\begin{aligned}\mathbb{P}\left(\{X \in S, D = d_k\}\right) &= \int_{x \in S} p_{k|x} \cdot d\mathbb{P}_X(x) \\ &> 0.\end{aligned}$$

Then, this implies

$$\begin{aligned}\psi(\tilde{f}) - \psi(f^*) &= \int_S p_{k|x} \left(\epsilon(x)^2 - 2\epsilon(x)g(x)\right) d\mathbb{P}_X(x) \\ &< 0,\end{aligned}$$

contradicting the optimality of f^*. Therefore, $\mathbb{P}_X(S) = 0$, as desired.

This completes the proof. □

C.5.2 PROPOSITION 4.7 (II) PROOF

Proof. Define

$$Q_x(z) := \sum_{k=1}^{|\mathcal{D}|} p_{k|x} \left(z_k - \mu_k(x)\right)^2 .$$

For a.e. x, let $q(x)$ minimize Q_x over vectors satisfying

$$z_k - \alpha z_j \geq 0 \quad (k \neq j), \quad \sum_k p_{k|x} z_k \geq \mu(x).$$

This problem has a unique minimizer: its feasible set is nonempty, closed, and convex, while its objective is coercive and strictly convex because every $p_{k|x} > 0$. The minimizer depends continuously on the parameter vector $(p, \boldsymbol{\mu})$ on the domain $p_k > 0$. Indeed, along convergent parameter sequences, comparison with the feasible constant premium gives local boundedness of the minimizers. Every feasible vector z for the limiting parameters has feasible approximants

$$z^{(r)} = z + \left(p^{(r)} \cdot \mu^{(r)} - p^{(r)} \cdot z\right)_+ \mathbf{1} \longrightarrow z.$$

Passing to limits in feasibility and optimality, and using uniqueness, proves continuity. Consequently, $q(x)$ is measurable. Comparison with the feasible vector $\mu(x)\mathbf{1}$ gives

$$Q_x(q(x)) \leq \sum_k p_{k|x} \left(\mu(x) - \mu_k(x)\right)^2 \leq \sum_k p_{k|x} \mu_k(x)^2.$$

Hence

$$\sum_k p_{k|x} q_k(x)^2 \leq 2Q_x(q(x)) + 2\sum_k p_{k|x} \mu_k(x)^2 \leq 4\sum_k p_{k|x} \mu_k(x)^2.$$

Thus $q_D(X) \in L^2\left(\mathbb{P}_{X,D}\right)$. Integrating the point-wise optimality inequality and using Proposition 4.1 yields

$$f^*\left(x, d_k\right) = q_k(x) \quad \text{for every } k, \quad \mathbb{P}_X\text{-a.e.}$$

Fix $\alpha \in (0, 1]$, and let x belong to the full-measure set on which the preceding pointwise characterization holds. If $\alpha = 1$, fairness forces all premiums to equal a common value c, and solvency requires $c \geq \mu(x)$. Since

$$\sum_k p_{k|x} \left(c - \mu_k(x)\right)^2 = \left(c - \mu(x)\right)^2 + \sum_k p_{k|x} \left(\mu_k(x) - \mu(x)\right)^2 ,$$

the optimum is $c = \mu(x)$, proving the claim. Henceforth, suppose $0 < \alpha < 1$. We first consider $|\mathcal{D}| = 2$, labeling the groups so that $\mu_L(x) \leq \mu_H(x)$. If $\mu_L(x) \geq \alpha\mu_H(x)$ and $\mu_H(x) \geq \alpha\mu_L(x)$, then, we know $(f^*(x, d_L), f^*(x, d_H)) = (\mu_L(x), \mu_H(x))$ is feasible (IS hits equality by definition of $\mu(x)$) and optimal with $\mu_L(x) \leq \mu(x) \implies f^*(x, d_L) = \mu_L(x) \leq \mu(x)$, as desired.

Now suppose $\mu_L(x) < \alpha\mu_H(x)$. For this fixed x, write the conditional squared-bias objective as

$$\psi(f) := p_{L|x} \left(f\left(x, d_L\right) - \mu_L(x)\right)^2 + p_{H|x} \left(f\left(x, d_H\right) - \mu_H(x)\right)^2 .$$

Suppose, for contradiction, that

$$f^*\left(x, d_L\right) > \alpha f^*\left(x, d_H\right).$$

For $t \in (0,1)$, define

$$f_t(x, d_r) := (1-t) f^*(x, d_r) + t\mu_r(x), \quad r \in \{L, H\}.$$

Solvency is preserved because the actuarial premium vector satisfies it with equality. The inequality $f_t(x, d_L) \geq \alpha f_t(x, d_H)$ remains valid for sufficiently small $t > 0$, by strict slack at $t = 0$. The reverse fairness inequality is also preserved because

$$\mu_H(x) - \alpha\mu_L(x) > 0.$$

However,

$$\psi(f_t) = (1-t)^2 \psi(f^*) < \psi(f^*),$$

where $\psi(f^*) > 0$ because the actuarial premium vector is infeasible. This contradicts pointwise optimality. Therefore,

$$f^*(x, d_L) = \alpha f^*(x, d_H).$$

Now, let $f(x, d_L) = \alpha f(x, d_H)$, we have:

$$\psi(f) = (\alpha f(x, d_H) - \mu_L(x))^2 \cdot p_{L|x} + (f(x, d_H) - \mu_H(x))^2 \cdot p_{H|x}.$$

Fairness constraint is automatically satisfied by the setup of $f(x, d_L) = \alpha f(x, d_H)$. For IS, we have:

$$\left(p_{H|x} + \alpha p_{L|x}\right) f(x, d_H) = f(x, d_L) \cdot p_{L|x} + f(x, d_H) \cdot p_{H|x} \geq \mu(x).$$

This implies

$$f(x, d_H) \geq \frac{\mu(x)}{p_{H|x} + \alpha p_{L|x}}.$$

Now, if we minimize ψ with $f(x, d_L) = \alpha f(x, d_H)$, we would obtain

$$\frac{\partial}{\partial f(x, d_H)} \psi(f) = 2\alpha \left(\alpha f(x, d_H) - \mu_L(x)\right) \cdot p_{L|x} + 2\left(f(x, d_H) - \mu_H(x)\right) \cdot p_{H|x} = 0.$$

This gives

$$\tilde{f}(x, d_H) = \frac{\alpha\mu_L(x) \cdot p_{L|x} + \mu_H(x) \cdot p_{H|x}}{\alpha^2 p_{L|x} + p_{H|x}},$$

consequently,

$$\tilde{f}(x, d_L) = \alpha\tilde{f}(x, d_H) = \frac{\alpha^2\mu_L(x) \cdot p_{L|x} + \alpha\mu_H(x) \cdot p_{H|x}}{\alpha^2 p_{L|x} + p_{H|x}}.$$

Therefore, we have that

$$f^*(x, d_H) = \max\left\{\tilde{f}(x, d_H), \frac{\mu(x)}{p_{H|x} + \alpha p_{L|x}}\right\},$$

and

$$f^*(x, d_L) = \alpha f^*(x, d_H).$$

If the maximum is attained in order to satisfy IS, then

$$\begin{aligned} f^*(x, d_L) &= \alpha f^*(x, d_H) \\ &= \frac{\alpha\mu(x)}{p_{H|x} + \alpha p_{L|x}}. \end{aligned}$$

Then, enough to verify that if

$$\begin{aligned} \frac{\alpha\mu(x)}{p_{H|x} + \alpha p_{L|x}} \leq \mu(x) \iff \alpha &\leq p_{H|x} + \alpha p_{L|x} \\ &= 1 - p_{L|x} + \alpha p_{L|x} \\ &= \alpha + (1-\alpha)(1 - p_{L|x}), \end{aligned}$$

as desired.

Now, suppose IS is slack, that is $f^*(x, d_H) = \tilde{f}(x, d_H)$, then, we have:

$$\begin{aligned} f^*(x, d_L) &= \alpha f^*(x, d_H) \\ &= \frac{\alpha^2 \mu_L(x) \cdot p_{L|x} + \alpha \mu_H(x) \cdot p_{H|x}}{\alpha^2 p_{L|x} + p_{H|x}}. \end{aligned}$$

Then, enough to verify if

$$\frac{\alpha^2 \mu_L(x) \cdot p_{L|x} + \alpha \mu_H(x) \cdot p_{H|x}}{\alpha^2 p_{L|x} + p_{H|x}} \leq \mu(x) \iff \frac{\alpha^2 \mu_L(x) \cdot p_{L|x} + \alpha \mu_H(x) \cdot p_{H|x}}{\alpha^2 p_{L|x} + p_{H|x}} \leq \mu_L(x) \cdot p_{L|x} + \mu_H(x) \cdot p_{H|x}.$$

Then, simplify to get:

$$\alpha \tilde{f}(x, d_H) - \mu(x) = \frac{(1-\alpha)(1-p_{L|x}) \left((1+\alpha)(\mu_H(x) - \mu_L(x)) \cdot p_{L|x} - \mu_H(x)\right)}{\alpha^2 p_{L|x} + p_{H|x}},$$

Under $p_{L|x} \leq 1/(1+\alpha)$,

$$\begin{aligned} &(1+\alpha) p_{L|x} \left(\mu_H(x) - \mu_L(x)\right) - \mu_H(x) \\ &\quad = \left((1+\alpha) p_{L|x} - 1\right) \mu_H(x) - (1+\alpha) p_{L|x} \mu_L(x) \leq 0. \end{aligned}$$

Since the prefactor is nonnegative, we obtain

$$f^*\left(x, d_L\right) = \alpha \tilde{f}\left(x, d_H\right) \leq \mu(x),$$

as desired.

For $|\mathcal{D}| > 2$, use the two-tier structure assumed in part (ii). Let $\mathcal{C}_L$ and $\mathcal{C}_H$ be nonempty sets partitioning the group-index set $[|\mathcal{D}|]$, such that, for some $M(x) > 0$,

$$f^*(x, d_k) = \begin{cases} \alpha M(x), \ \forall k \in C_L \\ M(x) \quad , \ \forall k \in C_H. \end{cases}$$

Then, let

$$\begin{aligned} p_{L|x} &:= \sum_{k \in C_L} p_{k|x}, \quad \mu_L(x) := \frac{1}{p_{L|x}} \sum_{k \in C_L} \mu_k(x) \cdot p_{k|x}, \\ p_{H|x} &:= \sum_{k \in C_H} p_{k|x}, \quad \mu_H(x) := \frac{1}{p_{H|x}} \sum_{k \in C_H} \mu_k(x) \cdot p_{k|x}. \end{aligned}$$

For a candidate upper-tier premium u, the conditional squared-bias objective is

$$\begin{aligned} &\sum_{k \in C_L} p_{k|x} \left(\alpha u - \mu_k(x)\right)^2 + \sum_{k \in C_H} p_{k|x} \left(u - \mu_k(x)\right)^2 \\ &\quad = p_{L|x} \left(\alpha u - \mu_L(x)\right)^2 + p_{H|x} \left(u - \mu_H(x)\right)^2 + C_x \end{aligned}$$

where C_x is independent of u. Solvency requires

$$\left(\alpha p_{L|x} + p_{H|x}\right) u \geq \mu(x).$$

Because $f^*(x, \cdot)$ is a conditional optimum and belongs to this two-tier family, its tier level $M(x)$ minimizes this scalar objective over the displayed feasible interval. Hence the preceding scalar minimizer formula and sufficient-condition calculation apply.

Then, with the same argument used for $|\mathcal{D}| = 2$. Therefore, if

$$\sum_{k \in \mathcal{C}_L} p_{k|x} \leq \frac{1}{1+\alpha},$$

then

$$f^*(x, d_k) = \alpha M(x) \leq \mu(x) \quad \text{for every } k \in \mathcal{C}_L.$$

This completes the proof. □

C.5.3 PROPOSITION 4.7 (III) PROOF

Proof. Fix $\alpha \in (0, 1]$. By the pointwise characterization established at the beginning of Appendix C.5.2, let x belong to the full-measure set on which $f^*(x, \cdot)$ minimizes the conditional objective subject to fairness and individual solvency. If $\alpha = 1$, fairness forces all coordinates to be equal, so the ordering conclusion holds immediately. Henceforth, suppose $0 < \alpha < 1$. The conditional risk can be written as follows:

$$\begin{aligned}\mathbb{E}\left[(f(X,D)-Y)^2 \mid X=x\right] &= \sum_{k=1}^{|\mathcal{D}|} \mathbb{E}\left[(f(x,d_k)-Y)^2 \mid X=x, D=d_k\right] \cdot p_{k|x} \\ &= \sum_{k=1}^{|\mathcal{D}|} \mathbb{E}\left[\left((f(x,d_k)-\mu_k(x)) + (\mu_k(x)-Y)\right)^2 \mid X=x, D=d_k\right] \cdot p_{k|x}.\end{aligned}$$

Expand the square, we have that for each fixed k,

$$\mathbb{E}\left[(f(x,d_k)-Y)^2 \mid X=x, D=d_k\right] = (f(x,d_k)-\mu_k(x))^2 + Var(Y \mid X=x, D=d_k).$$

Then, since the variance term does not depend on f, we can drop it when optimizing over f.

Let

$$M(x) := \max_k f(x, d_k).$$

For a jointly feasible premium vector, individual solvency implies

$$M(x) \geq \sum_k p_{k|x} f\left(x, d_k\right) \geq \mu(x) > 0.$$

Then, feasibility implies that

$$\alpha M(x) \leq f(x, \cdot) \leq M(x).$$

On the other hand, if there is some $M(x)$ with

$$\alpha M(x) \leq f(x, \cdot) \leq M(x),$$

then

$$\alpha \leq \frac{f(x, d_k)}{f(x, d_j)} \leq \frac{1}{\alpha},$$

for all pairs (k, j). In other words, for fixed x, α-fairness constraints are precisely the existence of a scalar $M(x)$ satisfying the above inequalities.

Thus, for fixed x, the optimization becomes:

$$\min_{z \in \mathbb{R}^{|\mathcal{D}|}, M(x) \in \mathbb{R}} \sum_{k=1}^{|\mathcal{D}|} (z_k - \mu_k(x))^2 \cdot p_{k|x},$$

subject to

$$\begin{aligned}\sum_{k=1}^{|\mathcal{D}|} (z_k - \mu_k(x)) \cdot p_{k|x} &\geq 0 \\ \alpha M(x) \leq z_k &\leq M(x), \forall k.\end{aligned}$$

Here z_k represents a candidate premium $f\left(x, d_k\right)$, and M is an auxiliary scalar. At an optimal pair, the premium coordinates satisfy $z_k^* = f^*\left(x, d_k\right)$; denote the corresponding auxiliary scalar by $M(x)$.

The program is convex, and its objective is strictly convex in z. For $0 < \alpha < 1$, choose $c > \mu(x)$, set $z_k = c$ for every k, and take

$$M = \frac{2c}{1+\alpha}.$$

Then

$$\alpha M < c < M, \quad \sum_k p_{k|x} z_k = c > \mu(x).$$

Thus Slater's condition holds, and there exist nonnegative Lagrange multipliers satisfying the KKT conditions at the conditional optimum.

Now, define the Lagrange multipliers:

- $\lambda(x) \geq 0$, for $\mu(x) - \sum_{k=1}^{|\mathcal{D}|} f(x, d_k) \cdot p_{k|x} \leq 0$.
- $u_k(x) \geq 0$, for $f(x, d_k) - M(x) \leq 0$.
- $v_k(x) \geq 0$, for $\alpha M(x) - f(x, d_k) \leq 0$.

At the optimal pair, stationarity in the premium coordinates gives

$$2p_{k|x} \left(f^* \left(x, d_k \right) - \mu_k(x) \right) - \lambda(x) p_{k|x} + u_k(x) - v_k(x) = 0.$$

Complementary slackness gives

$$u_k(x) \left(f^* \left(x, d_k \right) - M(x) \right) = 0, \quad v_k(x) \left(\alpha M(x) - f^* \left(x, d_k \right) \right) = 0.$$

Now, consider the following cases:

1. If $\alpha M(x) < f(x, d_k) < M(x)$, then $u_k(x) = v_k(x) = 0$, we get:
$$f(x, d_k) = \mu_k(x) + \frac{\lambda(x)}{2}.$$
2. If $f(x, d_k) = M(x)$, then for $\alpha < 1$, the lower bound is slack, meaning $v_k(x) = 0$, we get:
$$u_k(x) = (\lambda(x) - 2(M(x) - \mu_k(x))) \cdot p_{k|x} \geq 0 \implies \mu_k(x) + \frac{\lambda(x)}{2} \geq M(x).$$
3. If $f(x, d_k) = \alpha M(x)$, then the upper bound is slack, meaning $u_k(x) = 0$, we get:
$$v_k(x) = (2(\alpha M(x) - \mu_k(x)) - \lambda(x)) \cdot p_{k|x} \geq 0 \implies \mu_k(x) + \frac{\lambda(x)}{2} \leq \alpha M(x).$$

Then, for every k, we have:

$$f^*(x, d_k) = \Pi_{[\alpha M(x), M(x)]} \left(\mu_k(x) + \frac{\lambda(x)}{2} \right).$$

Note that if $\mu_k(x) \geq \mu_j(x)$, then

$$\mu_k(x) + \frac{\lambda(x)}{2} \geq \mu_j(x) + \frac{\lambda(x)}{2} \implies f(x, d_k) = \Pi \left(\mu_k(x) + \frac{\lambda(x)}{2} \right) \geq \Pi \left(\mu_j(x) + \frac{\lambda(x)}{2} \right) = f(x, d_j).$$

Hence, we have

$$\mu_k(x) \geq \mu_j(x) \implies f^*(x, d_k) \geq f^*(x, d_j),$$

for $\mathbb{P}_X$-$a.e.$ x. □