Optimal insurance design under distortion risk measures with variance constraint*

Wei Wang[1], Tim J. Boonen[2], Wenjun Jiang[3], and Yiying Zhang[4]

[1] *Research Institute for Interdisciplinary Sciences, School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai 200433, P.R. China. E-mail:* wangwei1@sufe.edu.cn

[2] *Department of Statistics and Actuarial Science, School of Computing and Data Science, The University of Hong Kong, Hong Kong, P.R. China. E-mail:* tjboonen@hku.hk

[3] *Department of Mathematics and Statistics, University of Calgary, Calgary, AB, T2N 1N4, Canada. E-mail:* wenjun.jiang@ucalgary.ca

[4] *Department of Mathematics, Southern University of Science and Technology, Shenzhen 518055, P.R. China. E-mail:* zhangyy3@sustech.edu.cn

January 6, 2026

Abstract

Variance constraints on insured risk can help to reduce the risk exposure of an insurer. In this paper, we investigate the optimal insurance design under distortion risk measures, focusing on minimizing the risk measure of the insured's end-of-period exposure, while the insurer limits the variance of his risk exposure. The design adheres to the incentive compatibility condition, and premiums are charged according to the generalized distortion premium principle. The optimal policy is derived semi-analytically by using the signed Choquet integral representation of standard deviation when the variance constraint is binding, and the optimal indemnity function is of the layered form. For cases where the underlying loss is discretely distributed, the optimal policy can be determined by solving a second-order cone program. In contrast, for continuous loss distributions, we apply the classical sample average approximation scheme and establish qualitative convergence results. To demonstrate the efficiency of the proposed model, we conduct several numerical tests. These tests indicate that as the variance constraint tightens, the insured cedes fewer losses while the insurer is exposed to less tail risk.

Keywords: Risk management; optimal insurance; distortion risk measure; variance constraint; second-order cone program.

*Corresponding author: Yiying Zhang, Department of Mathematics, Southern University of Science and Technology, Shenzhen 518055, P.R. China. E-mail: zhangyy3@sustech.edu.cn

JEL Classification: C60, G22

1 Introduction

Insurance is an important mechanism for the reallocation of risk between the insurer and the insured. The study of the optimal insurance design can be traced back to the seminal work of Borch (1960), in which the focus is on variance-based objectives of the insured. Since then, numerous scholars have embarked on exploring the optimal insurance problem from various perspectives and settings/frameworks, utilizing a wide range of methodologies. Many objective functions have been studied; for example, the expected utility framework has been studied by Arrow (1963, 1974), Raviv (1979), and Chi & Wei (2020). Other objectives include distortion risk measures (Cui et al., 2013; Assa, 2015), regret-based objectives (Chi & Zhuang, 2022), rank-dependent expected utilities (Bernard et al., 2015; Xu et al., 2019; Ghossoub, 2019), maxmin expected utilities (Birghila et al., 2023), narrow framing (Zheng, 2020), and the ruin probability (Jin et al., 2024). Moreover, recently, various forms of distributionally robust objectives with distortion risk measures are considered by Boonen & Jiang (2024), Cai et al. (2024), and Fadina et al. (2025).

As the objective of the insured, our focus is on the class of distortion risk measures (DRMs), which includes the Value-at-Risk (VaR) and the Conditional Value-at-Risk (CVaR) as two prominent examples. This is primarily because DRMs possess many desirable mathematical properties for risk measures, such as translation invariance, comonotonic additivity, and positive homogeneity (Wang et al., 1997). Moreover, DRMs are related to dual utility, which is characterized by Yaari (1987) via a modification of the independence axiom in Von Neumann-Morgenstern expected utility theory.

In this paper, we extend this stream of research by studying the optimal insurance problem within a risk minimization framework, where the insured's risk preference is characterized by the DRM and the insurer imposes a variance constraint on his risk exposure. If the insurer manages a sufficiently large number of independent and homogeneous policies, his risk will, according to the law of large numbers, be sufficiently diversified and approach zero. However, this assumption regarding the number of policies may not hold in practice, especially in reinsurance. In fact, the number of insureds may be small, and policies may be heterogeneous. Therefore, the variance of the insurer's loss gains greater significance. Moreover, the variance is favored in optimal insurance design due to its simplicity, interpretability, and broad relevance. It is widely applied across various domains of insurance and finance, such as portfolio selection and premium calculation. It provides a balance between mathematical tractability, practical applicability, and economic intuition. While other risk measures, such as CVaR, focus solely on tail risk and have their merits, the variance considers the entire distribution. This offers a well-rounded approach to assessing and managing risk in insurance contracts, making it a reliable and efficient choice.

The use of variance in optimal insurance design is not new in the literature, but related studies are relatively rare. Borch (1960) is the first to study the optimal insurance design problem that aims to minimize the variance of the insurer's retained risk. He proves that, given a fixed premium in accordance with the expected-value premium principle, a stop-loss (or deductible) insurance policy is optimal. This work is later generalized by Kaluszka (2001), who incorporates a variance-related premium principle, and shows that the optimal contract to minimize the variance of the insured's payment can be stop-loss, quota share (i.e., the insurer covers a constant proportion of the loss) or change-loss (i.e., a combination of stop-loss and quota share). For further developments in this direction, we refer to Kaluszka (2005). Additionally, Vajda (1962) examines the optimal insurance problem from the insurer's perspective by minimizing the variance of indemnity, showing that a quota share policy is optimal under certain conditions.

This paper is closely related to the recent work by Chi et al. (2024), who study the design of an optimal insurance contract in which the insured maximizes her expected utility and the insurer limits the variance of his risk exposure while maintaining the principle of indemnity and charging the premium according to the expected-value premium principle. However, our work differs from Chi et al. (2024) in several key aspects. First, we employ the DRM to capture the insured's risk preference instead of the expected utility framework. Many empirical studies in finance, such as those by Allais (1953) and Ellsberg (1961), indicate that individuals often have non-additive priors. Second, while Chi et al. (2024) use the expected-value premium principle, we adopt the general distortion premium principle, which includes the expected-value premium principle as a special case. Third, Chi et al. (2024) restrict the feasible set of indemnity function to ensure adherence to the principle of indemnity and then show that the desired policies satisfy the no-sabotage condition, whereas our model focuses on the no-sabotage condition directly, which helps rule out ex post moral hazard, as discussed in Huberman et al. (1983). Finally, unlike the approach in Chi et al. (2024), we develop a novel method by utilizing the signed Choquet integral representation of the standard deviation, combined with Lagrangian duality and the minimax theorem, to derive the optimal insurance policy, when the variance constraint is binding. This technique is novel in related literature on optimal insurance under DRMs.

The main contributions of the paper can be summarized as follows.

First, we propose an insurance design model from the insured's perspective under DRMs with a variance constraint. By considering the variance constraint, the insurer can effectively manage and assess the financial stability of its portfolio by limiting the potential fluctuations in underwriting risk. Moreover, we demonstrate that the proposed model is closely related to the insurance design model under a DRM with a mean-variance premium principle. However, to the best of our knowledge, the latter has not been explored in the literature.

Second, by utilizing the signed Choquet integral representation of the standard deviation (Wang et al., 2020, Example 3), we show that the Lagrange dual function for any fixed positive Lagrange multiplier is essentially a saddle-point problem, and use the classical minimax theorem to disentangle the structure of the optimal indemnity function. The resulting optimal indemnity function is of a layered form, determined by both the optimal Lagrange multiplier and the optimal integrand in the signed Choquet integral representation of the standard deviation for the optimal policy. To determine the parameters in the optimal policy, we show that when the underlying loss is discretely distributed, these parameters can be obtained by solving a second-order cone program, which is computationally efficient. For the continuously distributed loss, we employ the classical sample average approach to determine these parameters. We also establish qualitative convergence results for both the optimal value and the set of optimal solutions.

Third, some numerical experiments are presented to demonstrate the efficiency of the proposed model and the computational scheme. Our numerical tests support certain recent empirical findings from Armantier et al. (2023), such as the observation that if the variance constraint tightens, the insured cedes fewer losses while the insurer is exposed to less tail risk. The findings of this paper are useful to both academics and practitioners, as they shed light on the optimal indemnities for the well-studied class of DRMs by incorporating the variance constraint.

The rest of the paper is organized as follows. We formulate the main problem in Section 2. In Section 3, we develop the solution approach and present the optimal insurance contracts. In Section 4, we derive the tractable reformulation for the discretely distributed random loss. In Section 5, we establish the qualitative convergence results for the sample average approximation. We provide some numerical tests to the proposed model and computational framework in Section 6. Section 7 concludes the paper, and auxiliary numerical results are provided in Appendix A.

2 Problem formulation

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a given probability space, where Ω is a sample space with sigma-algebra $\mathcal{F}$ and $\mathbb{P}$ is a reference probability measure. Consider an insurer faced with an insurable loss X, which is a non-negative, essentially bounded random variable defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with the cumulative distribution function (c.d.f.) $F_X(x) := \mathbb{P}\{X \leq x\}$ and the essential supremum $M < \infty$. The survival function of X is denoted by $S_X(x) := 1 - F_X(x)$ and the left-continuous generalized inverse of F_X is denoted by

$$F_X^{-1}(t) = \inf\{x \in \mathbb{R} : F_X(x) \geq t\}, \ t \in (0, 1].$$

For simplicity, let $F_X^{-1}(0) = \inf\{x \in \mathbb{R} : F_X(x) > 0\} \geq 0$ as X is non-negative. In other words, we have $F_X^{-1}(t) \in [0, M]$, for all $t \in [0, 1]$. Throughout this paper, we adopt the following commonly used notation. Let $x \wedge y = \min\{x, y\}$. The indicator function is denoted by $\mathbb{1}_A(s)$, so that $\mathbb{1}_A(s) = 1$ for $s \in A$, and $\mathbb{1}_A(s) = 0$ otherwise. We use boldface lowercase letters for vectors (e.g., $\boldsymbol{v}$), and calligraphic letters for sets (e.g., $\mathcal{H}$). For a vector $\boldsymbol{v} \in \mathbb{R}^n$, $\|\boldsymbol{v}\|_2$ denotes its Euclidean norm. We adopt the convention that $\sup \emptyset = -\infty$. For any two sets $A, B \subset \mathbb{R}$, we use

$$\mathbb{D}(A, B) = \sup_{x \in A} \inf_{y \in B} |x - y|$$

to denote the deviation of set A from set B. For any two sets $\mathcal{H}_1, \mathcal{H}_2$ of continuous functions on $[0, 1]$, we use

$$\mathbb{D}(\mathcal{H}_1, \mathcal{H}_2) = \sup_{h_1 \in \mathcal{H}_1} \inf_{h_2 \in \mathcal{H}_2} \|h_1 - h_2\|_\infty$$

to denote the deviation of set $\mathcal{H}_1$ from set $\mathcal{H}_2$, where $\|h\|_\infty = \sup_{t\in[0,1]} |h(t)|$.

2.1 Admissible indemnity functions

An insurance contract design problem involves partitioning X into two parts: $I(X)$ and $X - I(X)$, where $I(X)$ (the indemnity) is the portion of the loss that is ceded to the insurer ("he"), and $R_I(X) := X - I(X)$ (the retention) is the portion borne by the insured ("she"). I and R_I are also called the insured's ceded and retained loss functions, respectively.

It is natural to require that a contract satisfies the *principle of indemnity*, namely, the indemnity is non-negative and less than the amount of loss. However, it may lead to ex post moral hazard, see, e.g., Huberman et al. (1983). To address this issue, Huberman et al. (1983) propose the *incentive-compatible condition* (also known as the *no-sabotage* condition in the existing insurance literature), which states that both the insured and the insurer pay more for a larger loss as a hard constraint on admissible insurance policies, in addition to the principle of indemnity. Thus, the no-sabotage set of indemnity functions is

$$\mathcal{I} := \{I : I(0) = 0,\ 0 \le I(x_2) - I(x_1) \le x_2 - x_1 \ \text{ for all } \ 0 \le x_1 \le x_2 \le M\}. \tag{2.1}$$

If $I \in \mathcal{I}$, then $X - I(X)$ and $I(X)$ are comonotonic.[1] Note that the set $\mathcal{I}$ is quite large and includes many well-known indemnity functions, such as the quota-share, the stop-loss, and the truncated stop-loss functions. Any $I \in \mathcal{I}$ is 1-Lipschitz continuous and consequently admits the following integral representation:

$$I(x) = \int_0^x h(t)dt,\ x \in [0, \infty),$$

where h is called the *marginal indemnification function* (MIF), see Assa (2015) and Zhuang et al. (2016). Thus, the set $\mathcal{I}$ can be rewritten as

$$\mathcal{I} = \left\{ I : I(x) = \int_0^x h(t)dt,\ 0 \le h(t) \le 1, \text{ and } h \text{ is Lebesgue measurable} \right\}.$$

As the insurer covers part of the loss for the insured, she is compensated by collecting the insurance premium from the insured. In the literature, such premium is often calculated using only the expected indemnity (see, e.g., Bernard et al. (2015); Xu et al. (2019)). Throughout this paper, we assume that the insurer calculates the insurance premium using the generalized distortion premium principle, as it covers many interested principles widely used in insurance economics, which will be formally introduced in the next subsection.

2.2 Distortion risk measure and general premium principle

A distortion function $g : [0, 1] \mapsto [0, 1]$ is a non-decreasing function such that $g(0) = 0$ and $g(1) = 1$. The set of all distortion functions is denoted by $\mathcal{G}_d$, and the subset of distortion functions that are

[1] Random variables Y, Z are called comonotonic if there exists a random variable W and non-decreasing functions f and g such that $Y = f(W)$ and $Z = g(W)$.

concave on $[0,1]$ is denoted by $\mathcal{G}_{cv}$. Let

$$\mathcal{G} = \left\{g : [0,1] \mapsto \mathbb{R}_+ \mid g(0) = 0,\ g(t) \text{ is non-decreasing and bounded}\right\}. \tag{2.2}$$

Obviously, it holds that $\mathcal{G}_{cv} \subset \mathcal{G}_d \subset \mathcal{G}$.

Definition 2.1. *A DRM ρ_g of a non-negative random variable X with a distortion function $g \in \mathcal{G}_d$ is defined as*

$$\rho_g(X) = \int_0^\infty g\left(S_X(x)\right) dx \tag{2.3}$$

provided that the integral exists.

It is well known that ρ_g is a coherent risk measure[2] if and only if $g \in \mathcal{G}_{cv}$, see e.g., Acerbi (2002) and Dhaene et al. (2006). For more discussions on the properties and applications of DRMs, we refer to Wang et al. (1997). Two prominent examples of DRMs are the VaR and the CVaR.

Definition 2.2. *The* VaR *and* CVaR *of a non-negative random variable X at a confidence level $\alpha \in (0,1)$ are respectively defined as*

$$\mathrm{VaR}_\alpha(X) = \inf\{x \in \mathbb{R}_+ : \mathbb{P}\{X \le x\} \ge \alpha\},$$

and

$$\mathrm{CVaR}_\alpha(X) = \frac{1}{1-\alpha}\int_\alpha^1 \mathrm{VaR}_s(X) ds$$

provided that the integral exists.

Indeed, for $\alpha \in (0,1)$, VaR_α and CVaR_α are two DRMs, corresponding to the distortion functions $g(t) = \mathbb{1}_{(1-\alpha,1]}(t)$ and $g(t) = 1 \wedge \frac{t}{1-\alpha}$, respectively. Clearly, the distortion function of VaR_α belongs to $\mathcal{G}_d$ while that of CVaR_α belongs to $\mathcal{G}_{cv}$. More detailed discussions on their properties can be found in, e.g., Föllmer & Schied (2011).

Now, we focus on the choice of insurance premium principle. It is important to note that Wang et al. (1997) propose several natural axioms for pricing insurance contracts that characterize the premium principle of Wang (1996), which is referred to as *Wang's premium principle*. Wang's premium principle is closely related to the dual theory of choice proposed by Yaari (1987). Interested readers may refer to Goovaerts et al. (2010) for more details on the evolution of DRM and their connections with premium principles. Under these axioms, they prove that the price to insure a

[2] A risk measure ρ is called coherent if it satisfies the following four properties:

- **Monotonicity**: $\rho(Y) \le \rho(Z)$ if $\mathbb{P}\{Y \le Z\} = 1$;
- **Sub-addivity**: $\rho(Y+Z) \le \rho(Y) + \rho(Z)$ for all Y, Z;
- **Positive homogeneity**: $\rho(c \cdot Y) = c \cdot \rho(Y)$ for all $c \ge 0$ and all Y;
- **Translation invariance**: $\rho(Y+d) = \rho(Y) + d$ for all $d \in \mathbb{R}$ and all Y.

See Artzner et al. (1999) for a discussion of these four properties.

risk is given by the expectation of the risk with respect to a distorted probability that depends on the underlying risk. More precisely, Wang's premium principle is defined as

$$\pi(I(X)) = (1+\theta)\rho_g(I(X)),$$

where $g \in \mathcal{G}_{cv}$ and the constant $\theta \geq 0$ is the *safety loading factor*. As pointed out by Wang et al. (1997) and Young (1999), Wang's premium principle satisfies many convenient properties and has been widely used both in academia and the insurance industry. Interested readers can refer to Denuit et al. (2006) for more discussions on its theoretical properties and applications.

In this paper, we assume that the insurer will adopt a more general insurance premium principle (Cui et al., 2013; Jin et al., 2024). For any ceded loss function $I \in \mathcal{I}$ purchased by the insurer, the insurance premium charged is determined by the following *general premium principle*

$$\pi_g(I(X)) = \rho_g(I(X)) = \int_0^\infty g\left(S_{I(X)}(s)\right) ds = \int_0^\infty g\left(S_X(s)\right) dI(s), \tag{2.4}$$

where $g \in \mathcal{G}$ is the pricing (or premium) function employed by the insurer. It is important to note that we do not require g to be concave or even continuous, as opposed to Wang's premium principle. It is obvious that the expected-value premium principle, Wang's premium principle, the distortion premium principle, and the proportional hazard transform principle are special cases of the general premium principle (2.4). Moreover, the general premium principle (2.4) essentially coincides with the *extended distortion premium principle*, as discussed in Jin et al. (2024), by interpreting $(1+\theta_0)g(t)$, where $\theta_0 > -1$ (as opposed to $\theta_0 \geq 0$), as a new distortion function within the set $\mathcal{G}$. This extension is appropriate; however, it is essential to note that any reasonable premium principle must satisfy some desirable properties outlined in Wang et al. (1997), such as non-negative loading and non-excessive loading.

In the following, we provide some other examples of such general distortion premium principles used in the literature.

Example 2.1. *Two well-known examples of the general distortion premium principles are as follows:*

(i) The Mean-CVaR premium principle proposed by Tan et al. (2020) is defined as

$$\pi(I) := (1+\theta)\left[\frac{1}{1+\beta}\mathbb{E}[I(X)] + \frac{\beta}{1+\beta}\mathrm{CVaR}_\alpha(I(X))\right],$$

where $\theta > 0$ is a safety loading factor, $\alpha \in (0,1)$, and $\beta > 0$ captures the relative weight between the two risk measures "Mean" and "CVaR". To recover the Mean-CVaR premium principle, we set the distortion function g as

$$g(t) = \begin{cases} \frac{1-\alpha+\beta}{(1-\alpha)(1+\beta)}t, & t \in [0, 1-\alpha], \\ \frac{1}{1+\beta}t + \frac{\beta}{1+\beta}, & t \in (1-\alpha, 1]. \end{cases}$$

(ii) *The generalized percentile premium principle, see, e.g., Kaluszka (2005), is given by*

$$\pi(I) := \mathbb{E}[I(X)] + \beta \left[F^{-1}_{I(X)}(\alpha) - \mathbb{E}[I(X)] \right],$$

where $\alpha, \beta \in (0,1)$, and β captures the weight between the risk measure "Mean" and "VaR". To recover the generalized percentile premium principle, we set the distortion function g as

$$g(t) = (1-\beta)t + \beta \mathbb{1}_{(1-\alpha,1]}(t),\ t \in [0,1].$$

2.3 Problem formulation

Suppose that the insured is interested in purchasing an insurance contract $(I, \pi(I))$ to hedge the risk X he bears. After buying the insurance contract, her loss becomes $X - I(X) + \pi(I(X))$. For a fixed insurance premium principle $\pi(I)$, we assume that the insured aims to choose an insurance contract that can minimize her risk exposure, measured by a DRM. The insurer may evaluate this risk using different measures for different purposes (see, e.g., Kaye, 2005). Following the same reasoning as in Chi et al. (2024), where it is assumed that the insurer has sufficient regulatory capital and therefore focuses on the volatility of the underwriting risk, we also use variance as the measure of risk in this paper. This approach allows the insurer to manage and assess the financial stability of its portfolio by limiting the potential fluctuations in the underwriting risk.

We now present the main problem studied in this paper.

Problem 1. *The insured aims to find the optimal ceded loss function by solving the following optimization problem:*

$$\min_{I \in \mathcal{I}} \quad \rho_{g_1}\left(X - I(X) + \pi_{g_2}(I(X))\right) \tag{2.5a}$$

$$\text{s.t.} \quad \mathrm{Var}(I(X) - \pi_{g_2}(I(X))) \leq \nu, \tag{2.5b}$$

where $g_1 \in \mathcal{G}_d$, $g_2 \in \mathcal{G}$, $\mathrm{Var}(\cdot)$ is the variance functional, and $\nu > 0$ is a constant imposed by the insurer.

In Problem 1, the insured's risk attitude is captured by the DRM ρ_{g_1} (Yaari, 1987), whereas the insurer's return is reflected by the general distortion premium π_{g_2} and his risk level by the variance bound ν. Moreover, we can interpret Problem 1 from various perspectives to offer some insights into the structure and implications of the problem. One may interpret the problem as one faced by an insurer who seeks to design a contract with the best interest of a representative insured in mind, so as to remain marketable and competitive, while maintaining the desired profitability and risk control. From the insured's perspective, she aims to minimize her risk exposure while accommodating the insurer's participation constraint imposed on the variance level. Note that if the upper bound ν is set to be greater than or equal to $\mathbb{E}[X^2]$, Problem 1 reduces to the model considered by Assa (2015):

$$\min_{I \in \mathcal{I}} \rho_{g_1}\left(X - I(X) + \pi_{g_2}(I(X))\right).$$

This holds because $\mathrm{Var}(I(X) - \pi_{g_2}(I(X))) = \mathrm{Var}(I(X)) \leq \mathbb{E}[X^2]$ for all $I \in \mathcal{I}$. A more precise upper bound will be discussed in Section 3.

Remark 2.1. *Problem 1 is closely related to the following unconstrained problem*

$$\min_{I\in\mathcal{I}} \rho_{g_1}(X - I(X) + \hat{\pi}(I(X))) \tag{2.6}$$

with the mean-variance premium principle $\hat{\pi}(I(X))$. To the best of our knowledge, Problem (2.6) has not been studied in the literature. Recall that the mean-variance premium principle is given by

$$\hat{\pi}(I(X)) := (1+\theta)\mathbb{E}[I(X)] + \frac{\eta}{2}\mathrm{Var}(I(X)), \tag{2.7}$$

in which $\theta \geq 0$ and $\eta \geq 0$ are safety loading parameters. Note that the insurance premium (2.7) reduces to the standard variance premium principle if $\theta = 0$ and to the expected-value premium principle if $\eta = 0$. To understand this, by considering the translation invariance of DRM, we get

$$\begin{aligned}\rho_{g_1}(X - I(X) + \hat{\pi}(I(X))) &= \rho_{g_1}(X - I(X) + (1+\theta)\mathbb{E}[I(X)]) + \frac{\eta}{2}\mathrm{Var}(I(X)) \\ &= \rho_{g_1}(X - I(X) + \pi_{\hat{g}}(I(X))) + \frac{\eta}{2}\mathrm{Var}(I(X)) \\ &= \rho_{g_1}(X - I(X) + \pi_{\hat{g}}(I(X))) + \frac{\eta}{2}\mathrm{Var}(I(X) - \pi_{\hat{g}}(I(X))),\end{aligned}$$

where $\hat{g}(t) = (1+\theta)t \in \mathcal{G}$. If we choose the value of η as the twice of the optimal Lagrange multiplier to Problem 1 with $g_2 = \hat{g}$, then we can immediately state that the optimal indemnity I^* to Problem 1 is also optimal to Problem (2.6).

3 Optimal insurance contracts

In this section, we present our approach for solving Problem 1. We begin by addressing the case where the variance constraint is slack, i.e., when ν is large. In this scenario, the solution is straightforward by directly applying the MIF method, allowing us to determine the smallest ν for which the constraint is slack.

Because of the translation invariance and comonotonic additivity[3] of ρ_{g_1}, we have

$$\rho_{g_1}(X - I(X) + \pi_{g_2}(I(X))) = \rho_{g_1}(X) - \rho_{g_1}(I(X)) + \pi_{g_2}(I(X)),$$

and consequently,

$$\begin{aligned}\rho_{g_1}(X - I(X) + \pi_{g_2}(I(X))) &= \rho_{g_1}(X) + \int_0^M \big(g_2(S_X(x)) - g_1(S_X(x))\big)\, dI(x) \\ &= \rho_{g_1}(X) + \int_0^M \big(g_2(S_X(x)) - g_1(S_X(x))\big)\, h(x)dx,\end{aligned} \tag{3.1}$$

where $h \in \mathcal{H}$ and the last equality holds as $I \in \mathcal{I}$. Minimizing the integral in (3.1) over the set $\mathcal{I}$ is equivalent to minimizing its integrand point-wise for almost every $x \in [0, M]$. That is, $I \in \mathcal{I}$ minimizes the integral in (3.1) if and only if $I = I_u^*$, where

$$I_u^*(x) = \int_0^x \left\{\mathbb{1}_{\{\tilde{g}(S_X(t))<0\}}(t) + \xi(t)\mathbb{1}_{\{\tilde{g}(S_X(t))=0\}}(t)\right\} dt, \tag{3.2}$$

[3] A risk measure is called comonotonic additive if $\rho(Y+Z) = \rho(Y) + \rho(Z)$ for all comonotonic Y, Z.

with $\tilde{g}(t) := g_2(t) - g_1(t)$, and $\xi(t)$ is a Lebesgue-measurable function taking values in $[0, 1]$. We refer to Theorem 3.1 in Zhuang et al. (2016) for more details. Apparently, I_u^* is the optimal indemnity function for Problem 1 without considering the variance constraint. We also define $\underline{I_u^*}(x) = \int_0^x \mathbb{1}_{\{\tilde{g}(S_X(t)<0)\}}(t)dt$, which denotes the minimum needed indemnity that minimizes the insured's risk exposure in the absence of variance constraint.

Let $\tilde{\nu} = \sqrt{\nu}$. Problem 1 can then be written as:

$$\min_{I\in\mathcal{I}} \quad \rho_{g_1}(X - I(X) + \pi_{g_2}(I(X))) \tag{3.3a}$$

$$\text{s.t.} \quad \text{SD}(I(X)) \leq \tilde{\nu}, \tag{3.3b}$$

where $\text{SD}(\cdot) := \sqrt{\text{Var}(\cdot)}$ is the standard deviation functional. Note that SD is a convex deviation measure (Rockafellar et al., 2006), implying that $\text{SD}(I(X))$ is convex in I. Moreover, the objective (3.3a) is linear in I, since it can be represented by (3.1), and the set I is convex. Therefore, Problem (3.3) is an infinite-dimensional convex program. Since $M < \infty$, the ceded loss function I is defined on the compact set $[0, M]$. Applying the Arzelà-Ascoli Theorem, we conclude that the set $\mathcal{I}$ is compact. Since the convex deviation measure SD is continuous (Rockafellar et al., 2006), the feasible set of Problem (3.3), being a closed subset of I, is also compact.

Clearly, the value of $\tilde{\nu}$ (or ν) plays a critical role in the problem (3.3). The following proposition characterizes the conditions under which the constraint in (3.3) becomes binding.

Proposition 3.1. *(i). If $\tilde{\nu} \geq \text{SD}(\underline{I_u^*}(X))$, then $\underline{I_u^*}$ is a solution to the problem (3.3).*

(ii). If $\tilde{\nu} \in (0, \text{SD}(\underline{I_u^*}(X)))$ then the problem (3.3) is convex, who attains its minimum. Moreover, the optimal solution I^* satisfies $\text{SD}(I^*(X)) = \tilde{\nu}$.

Proof. The proof of (i) is straightforward. We focus on the proof of (ii). The result of the first part follows directly from the linearity of the objective function with respect to I and the convexity and compactness of the feasible set. For the second part, let I^* be the solution to the problem (3.3). If $\tilde{\nu} < \text{SD}(\underline{I_u^*}(X))$, then I^* cannot be of the form (3.2), since $\tilde{\nu} < \text{SD}(\underline{I_u^*}(X)) \leq \text{SD}(I_u^*(X))$ due to the signed Choquet integral representation of SD, as introduced later in (3.7) and (3.9). Since all solutions to the unconstrained problem (3.3a) must be of the form (3.2), the optimal value of the constrained problem (3.3) must exhibit a strict increase compared with that of the unconstrained problem (3.3a). As a result, we have

$$\rho_{g_1}(X - I^*(X) + \pi_{g_2}(I^*(X))) > \rho_{g_1}(X - \underline{I_u^*}(X) + \pi_{g_2}(\underline{I_u^*}(X))).$$

Define $\hat{I}_\kappa^* = (1-\kappa)I^* + \kappa \underline{I_u^*}$, $\kappa \in [0, 1]$. Clearly, $\hat{I}_\kappa^* \in \mathcal{I}$. Since ρ_{g_1} and π_{g_2} are comonotonic additive and $\underline{I_u^*}(X)$ and $I^*(X)$ are comontonic, it follows that

$$\begin{aligned}&\rho_{g_1}(X - \hat{I}_\kappa^*(X) + \pi_{g_2}(\hat{I}_\kappa^*(X))) \\ &= (1-\kappa)(\rho_{g_1}(X - I^*(X) + \pi_{g_2}(I^*(X)))) + \kappa(\rho_{g_1}(X - \underline{I_u^*}(X) + \pi_{g_2}(\underline{I_u^*}(X)))),\end{aligned}$$

and hence

$$\rho_{g_1}(X - \hat{I}^*_\kappa(X) + \pi_{g_2}(\hat{I}^*_\kappa(X))) < \rho_{g_1}(X - I^*(X) + \pi_{g_2}(I^*(X)))$$

for all $\kappa \in (0,1]$.

Since $\mathrm{SD}(I(X))$ is continuous in I under the L_∞ metric, if $\mathrm{SD}(I^*(X)) < \tilde{\nu}$, then there exists a $\kappa \in (0,1)$ such that $\mathrm{SD}(\hat{I}^*_\kappa(X)) \le \tilde{\nu}$ holds, contradicting that I^* is the solution to the problem (3.3). Hence, we conclude that $\mathrm{SD}(I^*(X)) = \tilde{\nu}$. □

Proposition 3.1 tells exactly what the bound should be for the variance constraint to be active. Intuitively, if $\nu = 0$, then no insurance will be bought. Therefore, to focus on the non-trivial case, in the sequel we assume that $\tilde{\nu} \in (0, \mathrm{SD}(\underline{I^*_u}(X)))$. To solve Problem (3.3), we consider its Lagrangian dual formulation. By introducing an auxiliary variable $\lambda \ge 0$, the Lagrangian function of the problem in (3.3) is given by:

$$L(I,\lambda) := \rho_{g_1}(X - I(X) + \pi_{g_2}(I(X))) + \lambda\left(\mathrm{SD}(I(X)) - \tilde{\nu}\right).$$

Consequently, the dual problem of Problem (3.3) can be written as

$$\max_{\lambda\ge 0}\min_{I\in\mathcal{I}}\ \rho_{g_1}(X - I(X) + \pi_{g_2}(I(X))) + \lambda\left(\mathrm{SD}(I(X)) - \tilde{\nu}\right). \tag{3.4}$$

Since $\tilde{\nu} > 0$, Slater's condition holds for Problem (3.3), and consequently strong duality holds. Thus, to solve Problem (3.3), we can solve (3.4) instead.

Proposition 3.2. *If $\tilde{\nu} \in (0, \mathrm{SD}(\underline{I^*_u}(X)))$, then the Lagrange multiplier λ for Problem (3.3) is strictly positive and bounded.*

Proof. The first result follows directly from the complementary slackness condition for the convex program (3.3) and the boundedness of the Lagrange multipliers set, as discussed in Pomerol (1981). □

In the following discussion, we will first focus on the inner minimization problem of (3.4). For any $\lambda > 0$, let

$$\psi(\lambda) := \min_{I\in\mathcal{I}}\ \pi_{g_2}(I(X)) - \rho_{g_1}(I(X)) + \lambda\mathrm{SD}(I(X)). \tag{3.5}$$

Then, Problem (3.4) can be reformulated as

$$\max_{\lambda>0}\{\psi(\lambda) - \lambda\tilde{\nu}\} + \rho_{g_1}(X). \tag{3.6}$$

The next proposition establishes the relationship between the parameter $\tilde{\nu}$ and the optimal Lagrange multiplier λ^* that solves the problem (3.6). In particular, it implies that if the problem (3.6) admits a unique maximizer, then the maximizer $\lambda^*(\tilde{\nu})$ is non-increasing in $\tilde{\nu}$.

Proposition 3.3. *The function $\psi(\lambda)$ is concave in λ and the optimal Lagrange multiplier $\lambda^*(\tilde{\nu}) := \inf\left\{\arg\max_{\lambda>0}\{\psi(\lambda) - \lambda\tilde{\nu}\}\right\}$ is non-increasing in $\tilde{\nu}$.*

Proof. For any $\lambda_1, \lambda_2 > 0$ and $\beta \in [0,1]$, we have

$$\begin{aligned}
&\pi_{g_2}(I(X)) - \rho_{g_1}(I(X)) + \big((1-\beta)\lambda_1 + \beta\lambda_2\big)\,\mathrm{SD}(I(X)) \\
&= (1-\beta)\left(\pi_{g_2}(I(X)) - \rho_{g_1}(I(X)) + \lambda_1 \mathrm{SD}(I(X))\right) + \beta\left(\pi_{g_2}(I(X)) - \rho_{g_1}(I(X)) + \lambda_2 \mathrm{SD}(I(X))\right) \\
&\geq (1-\beta)\psi(\lambda_1) + \beta\psi(\lambda_2),
\end{aligned}$$

which implies that $\psi((1-\beta)\lambda_1+\beta\lambda_2) \geq (1-\beta)\psi(\lambda_1)+\beta\psi(\lambda_2)$. Hence, $\psi(\lambda)$ is concave. Thus, $\psi(\lambda)$ is almost everywhere differentiable over its domain, leading to $\lambda^*(\tilde{\nu}) = \inf\left\{\lambda > 0 : \ \psi'(\lambda+) - \tilde{\nu} \leq 0\right\}$, where $\psi'(\lambda+)$ denotes the right-side derivative of ψ. Since $\psi'(\lambda+) - \tilde{\nu}$ is decreasing in both λ and $\tilde{\nu}$, it follows that $\lambda^*(\tilde{\nu}_1) \leq \lambda^*(\tilde{\nu}_2)$ when $\tilde{\nu}_1 \geq \tilde{\nu}_2$. This completes the proof. □

To determine the function $\psi(\lambda)$, we need to address the standard deviation term in the minimization problem (3.5). Specifically, we need to reformulate or approximate $\mathrm{SD}(I(X))$ to facilitate solving this minimization problem. Thanks to the signed Choquet integral representation shown in Wang et al. (2020, Example 3), the SD can be represented as

$$\mathrm{SD}(X) = \sup_{h \in \widetilde{\mathcal{H}}} \int_0^1 F_X^{-1}(1-t)dh(t), \tag{3.7}$$

where

$$\widetilde{\mathcal{H}} := \left\{ h : [0,1] \mapsto \mathbb{R}, \ h(0) = h(1) = 0, \ h \text{ is concave}, \ \int_0^1 (h'(t))^2 dt \leq 1 \right\}. \tag{3.8}$$

Note that the functions in the set $\widetilde{\mathcal{H}}$ (as shown by (3.8)) are not necessarily continuous on the interval $[0,1]$, as any concave function is continuous on the interior of its domain rather than on the entire domain. Moreover, the integral in (3.7) should be understood as the Lebesgue-Stieltjes integral, and we cannot ignore the discontinuity points of h when calculating the integral. According to Wang et al. (2020, Example 3), a worst-case h is both attainable and continuous. Therefore, throughout this paper, we restrict the elements in $\widetilde{\mathcal{H}}$ to be continuous.

Remark 3.1. *For any $h \in \widetilde{\mathcal{H}}$, $h(t) \leq 0.5$ for $t \in [0,1]$. To see this, note that by Jensen's inequality, we have*

$$1 \geq \int_0^1 (h'(z))^2 dz \geq \left(\int_0^1 |h'(z)| dz\right)^2 \implies 1 \geq \int_0^1 |h'(z)| dz = \int_0^{t_0} h'(z)dz + \int_{t_0}^1 -h'(z)dz = 2h(t_0).$$

Moreover, there exists an $h \in \widetilde{\mathcal{H}}$ such that $h(t_0) = 0.5$ for some $t_0 \in [0,1]$ and h is piecewise linear having only two pieces with $t_0 = 0.5$.

Lemma 3.1. *The set $\widetilde{\mathcal{H}}$ is compact and the problem in (3.7) attains its maximum.*

Proof. We first show that the set $\widetilde{\mathcal{H}}$ is compact. For any $h \in \widetilde{\mathcal{H}}$, let $t, s \in [0,1]$, using the fundamental theorem of calculus, we have

$$|h(t) - h(s)| = \left|\int_s^t h'(u)du\right| \leq \left|\int_s^t |h'(u)|du\right|.$$

Applying the Cauchy-Schwarz inequality to the integral, we obtain

$$|h(s)-h(t)| \leq \left(\int_s^t 1^2 du\right)^{1/2}\left(\int_s^t (h'(u))^2 du\right)^{1/2} = \sqrt{|s-t|}\left(\int_s^t (h'(u))^2 du\right)^{1/2}.$$

Since $\int_0^1 (h'(u))^2 du \leq 1$ for all $h \in \widetilde{\mathcal{H}}$, it follows that

$$|h(s)-h(t)| \leq \sqrt{|s-t|}.$$

Note that the above bound is independent of the particular function h and only depends on the distance $|s-t|$. Thus, by the definition of uniformly equicontinuity, we can state that the set $\widetilde{\mathcal{H}}$ is uniformly continuous. Moreover, since $h(0)=0$, then $|h(t)| \leq \sqrt{t}$ for $t \in [0,1]$. Note that the bound is independent of the particular function h, thus, the set $\widetilde{\mathcal{H}}$ is uniformly bounded. Therefore, by applying the Arzelà-Ascoli Theorem, the set $\widetilde{\mathcal{H}}$ is compact.

Note that the objective function in Problem (3.7) is linear in h, and of course is continuous in h. Therefore, the problem in (3.7) attains its maximum. This completes the proof. □

Note that

$$\begin{aligned}\int_0^1 F^{-1}_{I(X)}(1-u)dh(u) &= \int_0^1 I(F_X^{-1}(1-u))dh(u)\\ &= I(F_X^{-1}(1-u))h(u)\Big|_0^1 - \int_0^1 h(u)dI(F_X^{-1}(1-u))\\ &= \int_0^M h(S_X(x))dI(x).\end{aligned}$$

Then, we have

$$\mathrm{SD}(I(X)) = \sup_{h\in\widetilde{\mathcal{H}}}\int_0^1 F^{-1}_{I(X)}(1-u)dh(u) = \sup_{h\in\widetilde{\mathcal{H}}}\int_0^M h(S_X(x))dI(x). \tag{3.9}$$

Therefore, by combining (3.1) and (3.9), Problem (3.5) can be rewritten as:

$$\begin{aligned}&\min_{I\in\mathcal{I}}\left\{\int_0^M \big(g_2(S_X(x))-g_1(S_X(x))\big)\,dI(x) + \lambda\cdot\sup_{h\in\widetilde{\mathcal{H}}}\int_0^M h(S_X(x))dI(x)\right\}\\ =&\min_{I\in\mathcal{I}}\sup_{h\in\widetilde{\mathcal{H}}}\left\{\int_0^M \big(\tilde{g}(S_X(x))+\lambda h(S_X(x))\big)\,dI(x)\right\},\end{aligned} \tag{3.10}$$

where $\tilde{g} := g_2 - g_1$.

To solve the minimax problem in (3.10), we recall the well-known minimax theorem (Fan, 1953).

Theorem 3.1 (Minimax theorem)**.** *Let Ξ_1 be a non-empty compact convex Hausdorff topological vector space*[4] *and Ξ_2 be a convex set. If $\mathcal{H}$ is a real-valued function defined on $\Xi_1\times\Xi_2$ such that*

- *$\xi_1 \mapsto \mathcal{H}(\xi_1,\xi_2)$ is convex and lower semi-continuous on Ξ_1 for each $\xi_2\in\Xi_2$; and*

[4] A Hausdorff topological vector space is a topological vector space with the separation property, i.e. any two distinct points in the space can be separated by disjoint open sets.

- $\xi_2 \mapsto \mathcal{H}(\xi_1, \xi_2)$ *is concave on* Ξ_2 *for each* $\xi_1 \in \Xi_1$,

then

$$\inf_{\xi_1 \in \Xi_1} \sup_{\xi_2 \in \Xi_2} \mathcal{H}(\xi_1, \xi_2) = \sup_{\xi_2 \in \Xi_2} \inf_{\xi_1 \in \Xi_1} \mathcal{H}(\xi_1, \xi_2).$$

The minimax theorem states that, under certain conditions, the infimum of the supremum of a real-valued function defined on a product of two convex sets is equal to the supremum of the infimum of the function. Note that $\mathcal{I}$ is compact and convex, and $\tilde{\mathcal{H}}$ is convex. Furthermore, the objective function

$$\int_0^M \big(\tilde{g}(S_X(x)) + \lambda h(S_X(x))\big)\, dI(x)$$

is linear both in I and h. Thus, all the requirements of Theorem 3.1 are fulfilled, allowing us to rewrite Problem (3.5) as

$$\sup_{h \in \widetilde{\mathcal{H}}} \min_{I \in \mathcal{I}} \left\{ \int_0^M \big(\tilde{g}(S_X(x)) + \lambda h(S_X(x))\big)\, dI(x) \right\}. \tag{3.11}$$

Note that the solution to the inner problem of Problem (3.11) is standard. To see this, for any fixed $\lambda > 0$ and $h \in \widetilde{\mathcal{H}}$, by applying the element-wise minimization, we obtain the optimal indemnity function to Problem (3.11) is $I^*_{\lambda,h}(x) = \int_0^x (I^*_{\lambda,h})'(s)ds$, where

$$(I^*_{\lambda,h})'(s) = \begin{cases} 0, & \text{if } \tilde{g}(S_X(s)) + \lambda h(S_X(s)) > 0, \\ \xi(s), & \text{if } \tilde{g}(S_X(s)) + \lambda h(S_X(s)) = 0, \\ 1, & \text{if } \tilde{g}(S_X(s)) + \lambda h(S_X(s)) < 0, \end{cases} \tag{3.12}$$

where $\xi(s)$ is $[0,1]$-valued and Lebesgue measurable. With the $I^*_{\lambda,h}$ as in (3.12), Problem (3.11) reduces to

$$\begin{aligned} & \sup_{h \in \widetilde{\mathcal{H}}} \int_0^M \big\{\tilde{g}(S_X(x)) + \lambda h(S_X(x))\big\}\, dI^*_{\lambda,h}(x) \\ = & \sup_{h \in \widetilde{\mathcal{H}}} \int_0^M \big\{\tilde{g}(S_X(x)) + \lambda h(S_X(x))\big\}\, \mathbb{1}_{\{\tilde{g}(S_X(x)) + \lambda h(S_X(x)) < 0\}}(x) dx \\ = & \sup_{h \in \widetilde{\mathcal{H}}} \int_0^M \Big\{\big(\tilde{g}(S_X(x)) + \lambda h(S_X(x))\big) \wedge 0\Big\}\, dx. \end{aligned} \tag{3.13}$$

Moreover, by applying the Radon measure

$$\mu([a,b)) := -F_X^{-1}(1-b) - (-F_X^{-1}(1-a)), \tag{3.14}$$

and using change-of-variable technique, Problem (3.11) can be further rewritten as

$$\sup_{h \in \widetilde{\mathcal{H}}} \int_0^1 \Big\{\big(\tilde{g}(t) + \lambda h(t)\big) \wedge 0\Big\}\, \mu(dt), \tag{3.15}$$

where the integral should be understood as the Lebesgue-Stieltjes integral. Therefore, Problem (3.3) can be reformulated as

$$\max_{\lambda > 0} \sup_{h \in \widetilde{\mathcal{H}}} \int_0^1 \Big\{\big(\tilde{g}(t) + \lambda h(t)\big) \wedge 0\Big\}\, \mu(dt) - \lambda \tilde{\nu} + \rho_{g_1}(X). \tag{3.16}$$

Remark 3.2. *Since the objective function in (3.15) is continuous in h and $\widetilde{\mathcal{H}}$ is compact, its optimum is attainable. Moreover, Problem (3.15) is an infinite-dimensional concave program, which is generally challenging to solve due to the complexity of the feasible set $\widetilde{\mathcal{H}}$, particularly because of the concavity requirement on function h.*

The following proposition shows a seemingly trivial property of the solution to Problem (3.15).

Proposition 3.4. *If the optimal value of Problem (3.15) is strictly negative, then the optimal h^* that solves Problem (3.15) must satisfy*

$$\int_0^1 ((h^*)'(t))^2 dt = 1. \tag{3.17}$$

Proof. Let $h^* \in \widetilde{\mathcal{H}}$ be an optimal solution to Problem (3.15). Then we have $h^*(0) = 0$, $h^*(1) = 0$, h^* is concave, and $\int_0^1 ((h^*)'(t))^2 dt \leq 1$. Since the optimal value of Problem (3.15) is strictly negative, it must hold that $\tilde{g}(t) + \lambda h^*(t) < 0$ on a set $C_1 \subseteq (0,1)$ with $\mu(C_1) > 0$. Clearly, this implies $\tilde{g}(t) < 0$ when $t \in C_1$. Consequently, we assert that $h^*(t) \not\equiv 0$ on C_1. Otherwise, one could construct $\hat{h}(t) > 0$ over C_1 such that $\tilde{g}(t) < \tilde{g}(t) + \lambda \hat{h}(t)$, thereby improving the objective value of Problem (3.15)—a contradiction. This implies that $h^*(t) > 0$ for all $t \in (0,1)$, since $h^* \in \widetilde{\mathcal{H}}$ is concave and satisfies $h^*(0) = h^*(1) = 0$. Assume that $\int_0^1 ((h^*)'(t))^2 dt < 1$. We consider $h(t) = (1+\epsilon)h^*(t)$ with $\epsilon > 0$. Then $h(0) = h(1) = 0$, h is also concave, and

$$\int_0^1 (h'(t))^2 dt = (1+\epsilon)^2 \int_0^1 ((h^*)'(t))^2 dt > \int_0^1 ((h^*)'(t))^2 dt.$$

Let $\epsilon = 1/\sqrt{\int_0^1 ((h^*)'(t))^2 dt} - 1 > 0$, so we have $h \in \widetilde{\mathcal{H}}$ and $\int_0^1 (h'(t))^2 dt = 1$. Furthermore, since $h^*(t) > 0$ for $t \in (0,1)$, we have $h(t) > h^*(t) > 0$ for all $t \in (0,1)$, which results in

$$(\tilde{g}(t) + \lambda h(t)) \wedge 0 \geq (\tilde{g}(t) + \lambda h^*(t)) \wedge 0, \text{ for } t \in [0,1],$$

and

$$(\tilde{g}(t) + \lambda h(t)) \wedge 0 > \tilde{g}(t) + \lambda h^*(t) = (\tilde{g}(t) + \lambda h^*(t)) \wedge 0, \text{ for } t \in C_1,$$

which results in

$$\int_0^1 \left\{ (\tilde{g}(t) + \lambda h(t)) \wedge 0 \right\} \mu(dt) > \int_0^1 \left\{ (\tilde{g}(t) + \lambda h^*(t)) \wedge 0 \right\} \mu(dt),$$

yielding a contradiction. This completes the proof. □

Theorem 3.2. *If $\nu \in (0, \mathrm{Var}(\underline{I_u^*}(X)))$ and the optimal value of Problem (3.15) is strictly negative, then there exist $\lambda^* > 0$ and $h^* \in \widetilde{\mathcal{H}}$ satisfying (3.17) such that the optimal indemnity function to Problem 1 is of the layered form, that is,*

$$I^*_{\lambda^*,h^*}(x) = \int_0^x (I^*_{\lambda^*,h^*})'(s) ds, \tag{3.18}$$

where $(I^*_{\lambda^*,h^*})'(s)$ satisfies (3.12) for some $\xi(s)$ such that $\mathrm{Var}(I^*_{\lambda^*,h^*}(X)) = \nu$.

Proof. This result is straightforward based on the above discussions. □

From Theorem 3.2, it is clear that the derivative of the optimal indemnity function to Problem 1, denoted as $(I^*_{\lambda^*,h^*})'(x)$, must take a value of either 0 or 1, except at points x where $\tilde{g}(S_X(x)) + \lambda^* h^*(S_X(x)) = 0$. If the Lebesgue measure of the set

$$E := \{x \in [0, M] : \tilde{g}(S_X(x)) + \lambda^* h^*(S_X(x)) = 0\}$$

is zero, then the optimal indemnity function $I^*_{\lambda^*,h^*}(x)$ satisfies $\mathrm{Var}(I^*_{\lambda^*,h^*}(X)) = \nu$ for any $[0,1]$-valued, Lebesgue measurable $\xi(s)$. Otherwise, $\xi(s)$ must be carefully chosen to ensure that

$$\mathrm{Var}(I^*_{\lambda^*,h^*}(X)) = \nu.$$

Let us consider the case where the function $\tilde{g}(S_X(x)) + \lambda^* h^*(S_X(x))$ is strictly positive on the interval $[0, x_1]$ and strictly negative on the interval $[x_2, M]$, with $E = [x_1, x_2]$. Figure 1 illustrates three different selections of $\xi(s)$ on the interval E. From the figure, we can observe that when $E = [x_1, x_2]$,

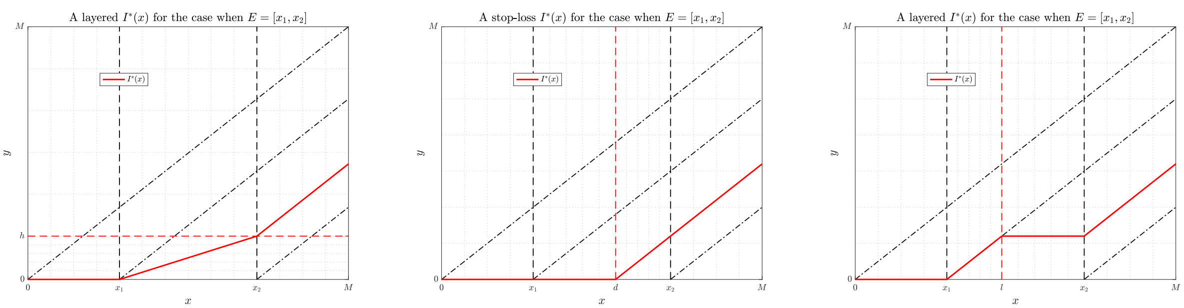

Figure 1: The illustration of $I^*_{\lambda^*,h^*}(x)$ ($I^*(x)$ for short) for different choices of $\xi(s)$ on the interval E. (Left): a linear function with a slope of $\frac{h}{x_2 - x_1}$; (Middle): a piecewise linear function with two slopes (0 on $[x_1, d]$ and 1 on $[d, x_2]$); (Right): a piecewise linear function with two slopes (1 on $[x_1, l]$ and 0 on $[l, x_2]$). In all cases, the parameters h, d, and l are determined from the condition $\mathrm{Var}(I^*(X)) = \nu$ under the assumption that $\mathrm{Var}((X - x_2)_+) < \nu < \mathrm{Var}((X - x_1)_+)$.

the selection of $\xi(s)$ within the interval E would affect the optimal $I^*(x)$ and consequently the parameters of $\xi(s)$ should be carefully chosen within the interval E. This consideration might be a fundamental difference between the models with a variance constraint and those without. It is important to note that different choices of $\xi(s)$ on the interval E result in the same indifference objective values for the insured. However, from the insurer's perspective, selecting the stop-loss indemnity as the optimal one may be more advantageous. This choice ensures that the insurer maximizes his profit by limiting his exposure to extreme losses while maintaining a predictable structure for covering smaller claims. Moreover, (3.12) provides an explicit expression of $(I^*_{\lambda^*,h^*})'(x)$, provided that the parameters λ^* and h^* are determined. However, the optimal h^* implicitly depends on the optimal ceded loss $I^*_{\lambda^*,h^*}(X)$ to the insurer, as seen in (3.9), which seems to complicate the determination of h^*. In the following section, we will discuss the approach for determining both λ^* and h^*.

4 Tractable reformulations

To solve Problem (3.15) efficiently, we first consider the case where the underlying random variable X is discrete. For a continuously distributed random variable X, we can use the *sample average approximation* (SAA) to discretize the probability distribution, thereby approximately solving Problem (3.15). For more details on this approach, see Kleywegt et al. (2002), Shapiro et al. (2021), and Wang & Xu (2023). To examine the convergence of the SAA approach, we also establish related convergence results with respect to both the optimal value and optimal solutions in Section 5.

The next theorem states that Problem (3.5) can be computed by solving a finite-dimensional quadratic program of reasonable size when X is discretely distributed with finite outcomes.

Theorem 4.1. *Let X be a non-negative finitely distributed random loss with $\mathbb{P}\{X = x_i\} = p_i$ for $i = 1, \ldots, n$, where $0 \le x_1 < x_2 < \cdots < x_n$. Then for a fixed $\lambda > 0$, the problem in (3.15) can be reformulated as the following optimization program*

$$
\begin{aligned}
\sup_{v,\eta,s} \quad & \sum_{i=0}^{n-1}(x_{i+1} - x_i)\eta_i && \text{(4.1a)}\\
\text{s.t.} \quad & \eta_i \le 0,\ i = 0, 1, \ldots, n-1, && \text{(4.1b)}\\
& \eta_i \le \tilde{g}(1-\pi_i) + \lambda v_i,\ i = 0, 1, \ldots, n-1, && \text{(4.1c)}\\
& v_i + s_i(\pi_i - \pi_{i+1}) \ge v_{i+1},\ i = 0, 1, \ldots, n-1, && \text{(4.1d)}\\
& v_{i+1} + s_{i+1}(\pi_{i+1} - \pi_i) \ge v_i,\ i = 0, 1, \ldots, n-1, && \text{(4.1e)}\\
& \sum_{i=0}^{n-1}\left(\frac{v_{i+1} - v_i}{\pi_i - \pi_{i+1}}\right)^2 (\pi_{i+1} - \pi_i) \le 1, && \text{(4.1f)}\\
& v_0 = 0,\ v_n = 0, && \text{(4.1g)}
\end{aligned}
$$

where $\pi_0 = 0$, $\pi_i = \sum_{j \le i} p_j$ for $i = 1, \ldots, n$, and $x_0 = 0$. Moreover, a worst-case concave function depending on λ (i.e., one achieving the supremum in (3.15)) is given by

$$h^*(t) = \min_{i=0,1,\ldots,n}\{s_i^*(t - (1-\pi_i)) + v_i^*\},\ t \in [0,1],$$

where v^* and s^* are optimal solutions to Problem (4.1). This is a piecewise linear concave function.

Proof. Since X is a discretely-distributed random loss, then its induced Radon measure μ defined as $\mu([a,b)) := -F_X^{-1}(1-b) - (-F_X^{-1}(1-a))$ is a discrete measure and is given by

$$\mu = \sum_{i=0}^{n-1}(x_{i+1} - x_i)\delta_{\{1-\pi_i\}},$$

where $x_0 = 0$, δ_{s_i} is a Dirac measure, and $\pi_i = \sum_{j \le i} p_j$ for $i = 1, \ldots, n$. Consequently, we have

$$\int_0^1 \left\{\left(\tilde{g}(t) + \lambda h(t)\right) \wedge 0\right\} \mu(dt) = \sum_{i=0}^{n-1}(x_{i+1} - x_i)\left\{\left(\tilde{g}(1-\pi_i) + \lambda h(1-\pi_i)\right) \wedge 0\right\}. \qquad (4.2)$$

From (4.2), it is obvious that its value only depends on the values of h at the points $1-\pi_i$ for $i=0,1,\ldots,n-1$ rather than the whole function of h. This finding helps us to derive the tractable reformulation of the problem in (3.15).

Let $\mathscr{H}$ be the set of all continuous functions defined on the interval $[0,1]$. For each fixed $v=(v_0,v_1,\ldots,v_n)\in\mathbb{R}^{n+1}$, we define the set

$$\mathcal{H}(v):=\{h\in\mathscr{H}:h(1-\pi_i)=v_i,\ i=0,1,\ldots,n-1\}.$$

Thus, Problem (3.15) is equivalent to

$$\sup_{v\in\mathbb{R}^{n+1}}\left\{\sup_{h\in\widetilde{\mathcal{H}}\cap\mathcal{H}(v)}\int_0^1\left\{\left(\tilde{g}(t)+\lambda h(t)\right)\wedge 0\right\}\mu(dt):\widetilde{\mathcal{H}}\cap\mathcal{H}(v)\neq\emptyset\right\},\tag{4.3}$$

which is a two-stage program. It first fixes the values of h at the points $1-\pi_i$ for $i=0,1,\ldots,n$, and then finds the worst-case v satisfying all the constraints.

We now establish conditions for the points $(1-\pi_i,v_i)$ for $i=0,1,\ldots,n$ to determine when $\widetilde{\mathcal{H}}\cap\mathcal{H}(v)\neq\emptyset$. For any $h\in\widetilde{\mathcal{H}}\cap\mathcal{H}(v)\neq\emptyset$, we immediately have $v_0=h(1)=0$ and $v_n=h(0)=0$.

Concavity requirement. For each i, let $l_i(t)=s_i(t-(1-\pi_i))+v_i$ be the affine function passing through $(1-\pi_i,v_i)$ with slope s_i. By Boyd & Vandenberghe (2004, Section 6.5.5), the concave function h passing through $(1-\pi_i,v_i)$ for $i=0,1,\ldots,n$ implies that they satisfy the following linear system

$$l_i(1-\pi_{i+1})=s_i(\pi_i-\pi_{i+1})+v_i\geq v_{i+1},\ i=0,1,\ldots,n-1$$

and

$$l_{i+1}(1-\pi_i)=s_{i+1}(\pi_{i+1}-\pi_i)+v_i\geq v_{i+1},\ i=0,1,\ldots,n-1.$$

Conversely, let $\hat{h}(t)=\min_{i=0,1,\ldots,n}l_i(t)$ for $t\in[0,1]$, then $\hat{h}$ is a piecewise linear concave function. By the proof of Wang & Xu (2023, Lemma 4.1), we have $\hat{h}(1-\pi_i)=v_i$ for $i=0,1,\ldots,n$. Thus, $\hat{h}\in\mathcal{H}(v)$.

Bounded variation requirement. For any $h\in\widetilde{\mathcal{H}}\cap\mathcal{H}(v)\neq\emptyset$, by the concavity of h, we know that

$$h(t)\geq\frac{v_{i+1}-v_i}{\pi_i-\pi_{i+1}}(t-(1-\pi_i))+v_i,\ t\in[1-\pi_{i+1},1-\pi_i)$$

and

$$\int_{1-\pi_{i+1}}^{1-\pi_i}h'(t)dt=h(1-\pi_i)-h(1-\pi_{i+1})=v_i-v_{i+1}=\int_{1-\pi_{i+1}}^{1-\pi_i}\frac{v_{i+1}-v_i}{\pi_i-\pi_{i+1}}dt$$

for $i=0,1,\ldots,n-1$. Consequently, we have

$$\int_{1-\pi_{i+1}}^{1-\pi_i}(h'(t))^2dt\geq\int_{1-\pi_{i+1}}^{1-\pi_i}\left(\frac{v_{i+1}-v_i}{\pi_i-\pi_{i+1}}\right)^2dt=\left(\frac{v_{i+1}-v_i}{\pi_{i+1}-\pi_i}\right)^2(\pi_{i+1}-\pi_i).$$

Therefore, the bounded variation requirement for h implies that points $(1-\pi_i,v_i)$ for $i=0,1,\ldots,n$ satisfy

$$1\geq\int_0^1(h'(t))^2dt\geq\sum_{i=0}^{n-1}\left(\frac{v_{i+1}-v_i}{\pi_{i+1}-\pi_i}\right)^2(\pi_{i+1}-\pi_i).\tag{4.4}$$

Conversely, let

$$\widetilde{h}(t) := \min_{i=0,1,\dots,n-1} \left\{ \frac{v_{i+1}-v_i}{\pi_i-\pi_{i+1}}(t-(1-\pi_i))+v_i \right\},\ t\in[0,1].$$

Then, from (4.4), we immediately have that $\int_0^1 (\widetilde{h}'(t))^2 dt \le 1$.

To sum up, Problem (4.3) can be reformulated as

$$\begin{aligned}
\sup_{v,s}\quad & \sum_{i=0}^{n-1}(x_{i+1}-x_i)\left\{\left(\tilde{g}(1-\pi_i)+\lambda v_i\right)\wedge 0\right\} && \text{(4.5a)}\\
\text{s.t.}\quad & v_i+s_i(\pi_i-\pi_{i+1})\ge v_{i+1},\ i=0,1,\dots,n-1, && \text{(4.5b)}\\
& v_{i+1}+s_{i+1}(\pi_{i+1}-\pi_i)\ge v_i,\ i=0,1,\dots,n-1, && \text{(4.5c)}\\
& \sum_{i=0}^{n-1}\left(\frac{v_{i+1}-v_i}{\pi_i-\pi_{i+1}}\right)^2(\pi_{i+1}-\pi_i)\le 1, && \text{(4.5d)}\\
& v_0=0,\ v_n=0. && \text{(4.5e)}
\end{aligned}$$

By introducing auxiliary variables η_i for $i=0,1,\dots,n-1$, the objective function (4.5a) can be rewritten as:

$$\begin{aligned}
\sup_{\eta}\quad & \sum_{i=0}^{n-1}(x_{i+1}-x_i)\eta_i && \text{(4.6a)}\\
\text{s.t.}\quad & \eta_i\le\tilde{g}(1-\pi_i)+\lambda v_i,\ \ i=0,1,\dots,n-1, && \text{(4.6b)}\\
& \eta_i\le 0,\ i=0,1,\dots,n-1. && \text{(4.6c)}
\end{aligned}$$

This is because the optimal solution to a linear program must lie on the boundary of the feasible set. Combining (4.5) and (4.6), we complete the proof. □

Note that except for the constraint (4.1f), the objective and constraints in Problem (4.1) are all linear. We now reformulate the constraint (4.1f). Note that

$$\begin{aligned}
&\sum_{i=0}^{n-1}\left(\frac{v_{i+1}-v_i}{\pi_i-\pi_{i+1}}\right)^2(\pi_{i+1}-\pi_i)=\sum_{i=0}^{n-1}\frac{1}{\pi_{i+1}-\pi_i}(v_{i+1}^2-2v_iv_{i+1}+v_i^2)\\
=&\frac{1}{\pi_1-\pi_0}v_0^2+\sum_{i=1}^{n-1}\left(\frac{1}{\pi_i-\pi_{i-1}}+\frac{1}{\pi_{i+1}-\pi_i}\right)v_i^2+\frac{1}{\pi_n-\pi_{n-1}}v_n^2-2\sum_{i=0}^{n-1}\frac{1}{\pi_{i+1}-\pi_i}v_iv_{i+1}.
\end{aligned}$$

Consequently, let $\boldsymbol{v}=(v_0,v_1,\dots,v_n)^{\mathrm{T}}\in\mathbb{R}^{n+1}$, then the constraint (4.1f) can be represented as

$$\boldsymbol{v}^{\mathrm{T}}A\boldsymbol{v}\le 1,$$

where the tridiagonal matrix A is given by

$$A:=\begin{bmatrix}
\frac{1}{\pi_1-\pi_0} & -\frac{1}{\pi_1-\pi_0} & & & \\
-\frac{1}{\pi_1-\pi_0} & \frac{1}{\pi_1-\pi_0}+\frac{1}{\pi_2-\pi_1} & -\frac{1}{\pi_2-\pi_1} & & \\
 & & \ddots & & \\
 & & & \ddots & -\frac{1}{\pi_n-\pi_{n-1}}\\
 & & & -\frac{1}{\pi_n-\pi_{n-1}} & \frac{1}{\pi_n-\pi_{n-1}}
\end{bmatrix}\in\mathbb{R}^{(n+1)\times(n+1)}.$$

Note that A is a positive semidefinite matrix. Consequently, the constraint (4.1f) can be further reformulated to the second-order cone constraint

$$\|A^{1/2}\boldsymbol{v}\|_2 \leq 1.$$

Problem (4.1) is a second-order cone program (SOCP), with $3n+2$ variables and $4n$ linear constraints and a convex quadratic constraint, which can be solved efficiently by using specialized solvers such as `coneprog` in MATLAB or MOSEK. Note that the SAA method and the SOCP approach have been employed in the discussion for optimal (re)insurance contracts by Tan & Weng (2014), Asimit et al. (2017), Asimit & Boonen (2018) and Asimit et al. (2020).

We now derive the tractable reformulation for Problem (3.4) based on the result in Theorem 4.1.

Corollary 4.1. *Let X be a non-negative finitely distributed random loss with $\mathbb{P}\{X = x_i\} = p_i$ for $i = 1, \dots, n$, where $0 \leq x_1 < x_2 < \cdots < x_n$. Then the problem in (3.4) can be reformulated as the following biconvex program*

$$\begin{aligned}
\sup_{v,\eta,s,\lambda} \quad & \sum_{i=0}^{n-1}(x_{i+1} - x_i)\eta_i - \lambda\tilde{\nu} + \rho_{g_1}(X) && (4.7a)\\
\text{s.t.} \quad & \eta_i \leq 0, \ i = 0, 1, \dots, n-1, && (4.7b)\\
& \eta_i \leq \tilde{g}(1 - \pi_i) + \lambda v_i, \ i = 0, 1, \dots, n-1, && (4.7c)\\
& v_i + s_i(\pi_i - \pi_{i+1}) \geq v_{i+1}, \ i = 0, 1, \dots, n-1, && (4.7d)\\
& v_{i+1} + s_{i+1}(\pi_{i+1} - \pi_i) \geq v_i, \ i = 0, 1, \dots, n-1, && (4.7e)\\
& \sum_{i=0}^{n-1}\left(\frac{v_{i+1} - v_i}{\pi_i - \pi_{i+1}}\right)^2(\pi_{i+1} - \pi_i) \leq 1, && (4.7f)\\
& v_0 = 0, \ v_n = 0, \ \lambda \geq 0, && (4.7g)
\end{aligned}$$

where $\pi_0 = 0$, $\pi_i = \sum_{j \leq i} p_j$ for $i = 1, \dots, n$, and $x_0 = 0$.

In general, it is challenging to solve a biconvex program. Fortunately, since the multiplier is strictly positive (see Proposition 3.2), Problem (4.7) can be further reformulated as a convex program by the change of some variables as follows:

$$\tilde{v}_i = \lambda v_i, \ \tilde{s}_i = \lambda s_i, \ i = 0, 1, \dots, n, \tag{4.8}$$

that is,

$$\begin{aligned}
\sup_{\tilde{v},\eta,\tilde{s},\lambda} \quad & \sum_{i=0}^{n-1}(x_{i+1} - x_i)\eta_i - \lambda\tilde{\nu} + \rho_{g_1}(X) && (4.9a)\\
\text{s.t.} \quad & \eta_i \leq 0, \ i = 0, 1, \dots, n-1, && (4.9b)\\
& \eta_i \leq \tilde{g}(1 - \pi_i) + \tilde{v}_i, \ i = 0, 1, \dots, n-1, && (4.9c)\\
& \tilde{v}_i + \tilde{s}_i(\pi_i - \pi_{i+1}) \geq \tilde{v}_{i+1}, \ i = 0, 1, \dots, n-1, && (4.9d)
\end{aligned}$$

$$\tilde{v}_{i+1} + \tilde{s}_{i+1}(\pi_{i+1} - \pi_i) \geq \tilde{v}_i, \ i = 0, 1, \ldots, n-1, \tag{4.9e}$$

$$\sum_{i=0}^{n-1} \left(\frac{\tilde{v}_{i+1} - \tilde{v}_i}{\pi_i - \pi_{i+1}} \right)^2 (\pi_{i+1} - \pi_i) \leq \lambda^2, \tag{4.9f}$$

$$v_0 = 0, \ v_n = 0, \ \lambda \geq 0, \tag{4.9g}$$

where $\pi_0 = 0$, $\pi_i = \sum_{j \leq i} p_j$ for $i = 1, \ldots, n$, and $x_0 = 0$. Note that the problem in (4.9) is also an SOCP with $3n + 3$ variables and $4n$ linear constraints, and a second-order cone constraint

$$\|A^{1/2} \tilde{\boldsymbol{v}}\|_2 \leq \lambda.$$

Moreover, the optimal solution to Problem (4.7) can be easily found from the optimal solution to Problem (4.9) via (4.8). This method will be used and discussed in numerical tests to find the optimal λ^* and optimal h^* in Problem (3.16).

5 Convergence analysis for the continuous case

Throughout this section, we assume that X has continuous support on $[0, M]$, so that its induced Radon measure μ as defined in (3.14) is atomless. In this section, we will establish the convergence results for the SAA counterparts of both Problem (3.15) and the outer maximization of Problem (3.16). Specifically, we will show that the functional solutions obtained from the SAA counterpart (Problem (4.1)) converge to a functional solution to the original Problem (3.15) and the optimal value of Problem (4.1) uniformly converges to the optimal value of Problem (3.15) for any λ in a certain closed bounded interval. While such convergence is intuitively expected, its formal proof requires careful extensions from known results. This is because our solution is in the space of functions, which is not a finite-dimensional space. Finally, we show that the optimal λ^* of Problem (4.9) converges to that of Problem (3.16).

Let $x_1, \ldots, x_N$, $N \in \mathbb{N}$ be a sequence of independent and identically distributed (i.i.d.) samples drawn from the distribution of a continuously distributed loss X. Without loss of generality, we assume that $x_1 < x_2 < \cdots < x_N$, and denote the approximated loss as X_N. Then the induced Radon measure μ_N defined as

$$\mu_N([a, b)) := -F_{X_N}^{-1}(1 - b) - (-F_{X_N}^{-1}(1 - a))$$

is a discrete measure and is given by

$$\mu_N = \sum_{i=0}^{N-1} (x_{i+1} - x_i) \delta_{\frac{N-i}{N}}, \ N \in \mathbb{N},$$

where $x_0 = 0$, and δ_s is the Dirac measure on $[0, 1]$.

5.1 Convergence results for Problem (3.15)

To begin with our convergence analysis, we introduce some notation. For any $\lambda > 0$, let

$$\psi_N(\lambda) := \sup_{h \in \widetilde{\mathcal{H}}} \int_0^1 \left\{ \big(\tilde{g}(t) + \lambda h(t)\big) \wedge 0 \right\} \mu_N(dt) \tag{5.1}$$

be the optimal value of the SAA counterpart of Problem (3.15), $\mathcal{H}_N^*(\lambda)$ and $\mathcal{H}^*(\lambda)$ be the optimal solution sets of problems (5.1) and (3.15), respectively.

Theorem 5.1. *For any $\lambda > 0$, we have $\psi_N(\lambda) \to \psi(\lambda)$ and $\mathbb{D}(\mathcal{H}_N^*(\lambda), \mathcal{H}^*(\lambda)) \to 0$ with probability 1 as $N \to \infty$. Moreover, for any closed bounded interval denoted by $E \subset \mathbb{R}_+$, the following holds*

$$\lim_{N\to\infty} \sup_{\lambda \in E} |\psi_N(\lambda) - \psi(\lambda)| = 0. \tag{5.2}$$

Proof. To begin with, for any fixed $\lambda > 0$, we introduce a function $f_h(t) := \big(\tilde{g}(t) + \lambda h(t)\big) \wedge 0$ for $t \in [0, 1]$.

Step 1: Point-wise convergence

Since $g_1 \in \mathcal{G}_d$ and $g_2 \in \mathcal{G}$, the function $\tilde{g}(t) = g_2(t) - g_1(t)$ is of bounded variation on the interval $[0, 1]$. Consequently, the function $f_h(t)$ is also of bounded variation for any $\lambda > 0$ and any $h \in \widetilde{\mathcal{H}}$, which implies that it can be decomposed as the difference of two non-decreasing functions:

$$f_h(t) = f_1(t) - f_2(t),\ t \in [0, 1],$$

where f_1 and f_2 are non-decreasing. Both f_1 and f_2 are of bounded variation and have at most countably many discontinuities, i.e., the set $D_{f_i} := \{t \in [0, 1] : f_i \text{ is discountinuous at } t\}$ is countable for $i = 1, 2$. Since μ is a probability measure without mass, then it follows

$$\mu(D_{f_i}) = 0,\ i = 1, 2.$$

Consequently, $\mu(D_{f_h}) = 0$. Given that f_h is bounded and measurable on $[0, 1]$ and μ_N weakly converges to μ as N increases by the Glivenko-Gantelli theorem, then by Portemanteau Theorem (Klenke, 2013, Theorem 13.16) we directly have

$$\lim_{N\to\infty} \int_0^1 f_h(t)\mu_N(dt) = \int_0^1 f_h(t)\mu(dt),\ \text{for all } h \in \widetilde{\mathcal{H}}. \tag{5.3}$$

Step 2: Uniformly convergence

Note that $f_h(t)$ is continuous in h with respect to $\|\cdot\|_\infty$-norm. Since $\widetilde{\mathcal{H}}$ is compact, there exist finitely many points, $h_1, \ldots, h_r \in \widetilde{\mathcal{H}}$, and corresponding neighborhoods $\mathcal{N}_1, \ldots, \mathcal{N}_r$ covering $\widetilde{\mathcal{H}}$ such that, for any $\epsilon > 0$ and $j = 1, \ldots, r$:

$$\sup_{h \in \mathcal{N}_j \cap \widetilde{\mathcal{H}}} \left| \int_0^1 f_h(t)\mu_N(dt) - \int_0^1 f_{h_j}(t)\mu_N(dt) \right| < \frac{\epsilon}{3},$$
$$\text{and} \sup_{h \in \mathcal{N}_j \cap \widetilde{\mathcal{H}}} \left| \int_0^1 f_h(t)\mu(dt) - \int_0^1 f_{h_j}(t)\mu(dt) \right| < \frac{\epsilon}{3}.$$

From (5.3), we know that

$$\left| \int_0^1 f_{h_j}(t)\mu_N(dt) - \int_0^1 f_{h_j}(t)\mu(dt) \right| < \frac{\epsilon}{3},\ j = 1, \ldots, r,$$

for N large enough. Without loss of generality, we assume that a given $h \in \widetilde{\mathcal{H}}$ is covered by the neighborhood $\mathcal{N}_j$ for some j. Then

$$\left|\int_0^1 f_h(t)\mu_N(dt) - \int_0^1 f_h(t)\mu(dt)\right| \leq \left|\int_0^1 f_h(t)\mu_N(dt) - \int_0^1 f_{h_j}(t)\mu_N(dt)\right| + \left|\int_0^1 f_{h_j}(t)\mu_N(dt) - \int_0^1 f_{h_j}(t)\mu(dt)\right| + \left|\int_0^1 f_{h_j}(t)\mu(dt) - \int_0^1 f_h(t)\mu(dt)\right| < \epsilon.$$

Therefore,

$$\sup_{h\in\widetilde{\mathcal{H}}}\left|\int_0^1 f_h(t)\mu_N(dt) - \int_0^1 f_h(t)\mu(dt)\right| < \epsilon,$$

which shows that $\int_0^1 f_h(t)\mu_N(dt)$ uniformly converges to $\int_0^1 f_h(t)\mu(dt)$ on $\widetilde{\mathcal{H}}$.

Step 3: $\psi_N(\lambda) \to \psi(\lambda)$ as $N \to \infty$

By the definitions of $\psi_N(\lambda)$ and $\psi(\lambda)$, we have

$$|\psi_N(\lambda) - \psi(\lambda)| \leq \sup_{h\in\widetilde{\mathcal{H}}}\left|\int_0^1 f_h(t)\mu_N(dt) - \int_0^1 f_h(t)\mu(dt)\right| < \epsilon$$

for sufficiently large N.

Step 4: $\mathbb{D}(\mathcal{H}_N^*(\lambda), \mathcal{H}^*(\lambda)) \to 0$ as $N \to \infty$

We argue now by contradiction. Suppose that $\mathbb{D}(\mathcal{H}_N^*(\lambda), \mathcal{H}^*(\lambda)) \not\to 0$. Since $\widetilde{\mathcal{H}}$ is compact, by passing to a subsequence if necessary, there exists $\hat{h}_N^* \in \mathcal{H}_N^*(\lambda)$ such that $d(\hat{h}_N^*, \mathcal{H}^*(\lambda)) \geq \epsilon$ for some $\epsilon > 0$ and that $\hat{h}_N^*$ tends to a point $\hat{h} \in \widetilde{\mathcal{H}}$. It follows that $\hat{h} \notin \mathcal{H}^*(\lambda)$ and hence

$$\int_0^1 f_{\hat{h}}(t)\mu(dt) < \psi(\lambda),$$

where the strict inequality holds since $\hat{h} \in \widetilde{\mathcal{H}} \setminus \mathcal{H}^*(\lambda)$. Moreover, $\psi_N(\lambda) = \int_0^1 f_{\hat{h}_N}(t)\mu_N(dt)$ and

$$\begin{aligned} &\left|\int_0^1 f_{\hat{h}_N}(t)\mu_N(dt) - \int_0^1 f_{\hat{h}}(t)\mu(dt)\right| \\ \leq\ &\left|\int_0^1 f_{\hat{h}_N}(t)\mu_N(dt) - \int_0^1 f_{\hat{h}_N}(t)\mu(dt)\right| + \left|\int_0^1 f_{\hat{h}_N}(t)\mu(dt) - \int_0^1 f_{\hat{h}}(t)\mu(dt)\right|. \end{aligned} \tag{5.4}$$

The first term in the right-hand side of (5.4) converges to zero due to the uniform convergence of $\int_0^1 f_h(t)\mu_N(dt)$ and the second term converges to zero due to the continuity of $\int_0^1 f_h(t)\mu(dt)$. To this end, we obtain that $\psi_N(\lambda)$ tends to $\int_0^1 f_{\hat{h}}(t)\mu(dt) < \psi(\lambda)$, which is a contradiction.

Step 5: (5.2) holds.

Since the closed bounded interval on $\mathbb{R}_+$ is compact and $\psi_N(\lambda)$ and $\psi(\lambda)$ are continuous in λ, consequently, by the Heine-Cantor theorem, $\psi_N(\lambda)$ uniformly converges to $\psi(\lambda)$. This completes the proof. □

5.2 Convergence results for the outer maximization of Problem (3.16)

We now focus on the convergence of the outer maximization of Problem (3.16). Let

$$v_N^* := \max_{\lambda \geq 0} \left\{ \psi_N(\lambda) - \lambda \tilde{\nu} \right\} + \rho_{g_1}(X) \text{ and } v^* := \max_{\lambda \geq 0} \left\{ \psi(\lambda) - \lambda \tilde{\nu} \right\} + \rho_{g_1}(X).$$

Let Λ_N^* and Λ^* denote the optimal solution sets to them, respectively.

Theorem 5.2. *$v_N^* \to v^*$ and $\mathbb{D}(\Lambda_N^*, \Lambda^*) \to 0$ with probability one as $N \to \infty$.*

Proof. From Proposition 3.2, the Lagrangian multiplier λ to Problem (3.16) is bounded, then there exists a closed bounded interval denoted by $E \subset \mathbb{R}_+$ such that

$$v_N^* := \max_{\lambda \in E} \left\{ \psi_N(\lambda) - \lambda \tilde{\nu} \right\} + \rho_{g_1}(X) \text{ and } v^* := \max_{\lambda \in E} \left\{ \psi(\lambda) - \lambda \tilde{\nu} \right\} + \rho_{g_1}(X).$$

Note that (i) the function $\psi(\lambda) - \lambda\tilde{\nu}$ is finite valued and continuous in λ on compact set E; (ii) $\psi_N(\lambda) - \lambda\tilde{\nu}$ uniformly converges to $\psi(\lambda) - \lambda\tilde{\nu}$ on E by Theorem 5.1. Therefore, by Theorem 5.3 in Shapiro et al. (2021), we complete the proof. □

6 Numerical experiments

In this section, we present some numerical tests to further analyze the effect of the variance constraint on the indemnity function and to evaluate the efficiency of the proposed computational scheme. All experiments were carried out using MATLAB 2024a installed on a Macbook Pro (i5-5257 CPU, 2.90GHz dual core processor, 8GB memory).

6.1 The effect of the variance constraint on the indemnity function

In this subsection, we investigate the effect of the variance constraint on the indemnity function using the results in Section 3. We adopt the following setup for the numerical tests:

- The underlying continuous-distributed random loss X is assumed to follow a truncated exponential distribution with a mean of 1000, i.e.,

  $$F_X(x) = \frac{1 - \exp\left(-\frac{x}{1000}\right)}{1 - \exp(-10^2)}, \quad x \in [0, 10^5],$$

  where $M = 10^5$.

- We randomly generate N samples from $F_X(x)$. In these tests, we choose $N = 1000$. The largest loss in our sample is approximately 7557, and consequently, we can update M to this value.

- The insured's concave distortion function is of the power type, also known as the proportional hazards (PH) transform (Wang, 1995), i.e.,

  $$g_1(t) = t^p, \; p \in (0, 1).$$

A lower value of p corresponds to a more concave distortion function and, thus, a higher aversion to mean-preserving spreads (Yaari, 1987). We set $p \in \{0.3, 0.5, 0.7\}$.

- We assume that the insurer calculates the premium using the expected-value principle that is widely used in insurance economics (see, e.g., Mossin, 1968; Wang et al., 1997; Xu et al., 2019), namely, the premium for making a non-negative random payment $I(X)$ is calculated as

  $$\pi(I(X)) = (1+\theta)\mathbb{E}[I(X)],$$

  where $\theta \geq 0$ is the so-called safety loading coefficient. The premium principle is also a special case of the general premium principle, corresponding to the pricing function $g_2(t) = (1+\theta)t$, so that $g_2 \in \mathcal{G}$. In our tests, we set $\theta = 0.1$.

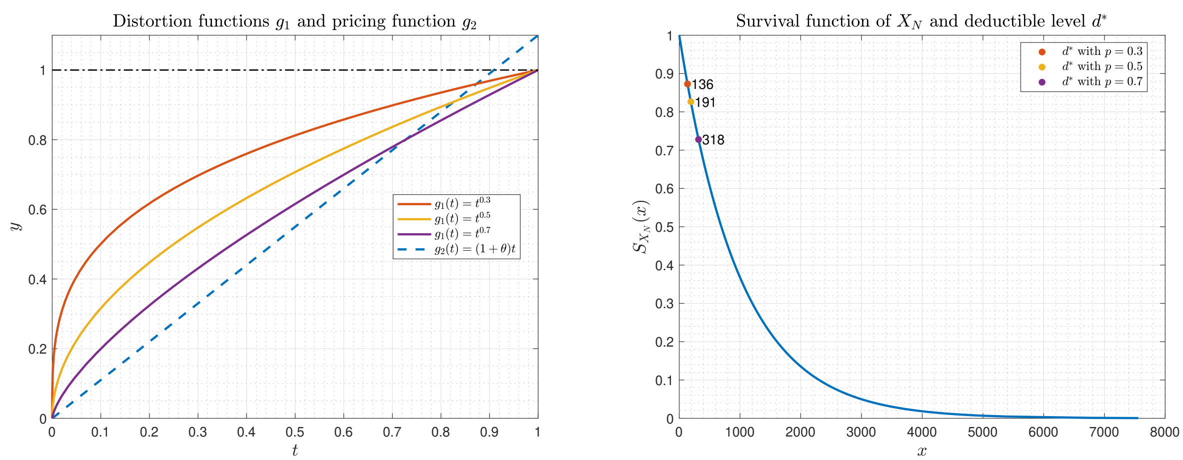

Figure 2: (Left): The distortion functions g_1 for different parameters and pricing function g_2; (Right): The survival function of X_N and the deductible level d^* without variance constraint for different parameters of g_1.

Figure 2 (Left) displays the distortion functions for different parameters and the pricing function g_2. A larger value of p corresponds to a less concave distortion function. Moreover, there is only one (non-zero) intersection point between g_2 and g_1 for all considered p, which means that the optimal indemnity function to Problem 1 without the variance constraint is a stop-loss function (see (3.2)), i.e., $I(x) = (x - d^*)_+$, where $d^* \geq 0$ is the specified deductible level. Figure 2 (Right) illustrates the deductible level d^* on the survival function of X_N for different parameters of the distortion function g_1. A smaller value of p corresponds to a smaller deductible level d^*. This is because a more risk-averse insured is likely to prefer a policy that offers greater coverage.

For a given deductible level d^*, we can easily calculate the standard deviation of $(X_N - d^*)_+$. For $p = 0.3, 0.5, 0.7$, the standard deviations are 926, 920, and 898, respectively. Therefore, we will choose the value of the standard deviation bound $\tilde{\nu}$ from the interval $(0, 800]$ to ensure that the variance constraint in Problem 1 is always binding.

Figure 3 displays the optimal h_N^* for Problem (3.16) by solving the problem in (4.9) for different values of the parameter p in g_1 when the standard deviation bounds $\tilde{\nu}$ are given by 1, 200, and 500. For the case when $\tilde{\nu} = 800$, the optimal h_N^* is shown in Figure A.1 (Left) in Appendix A. From

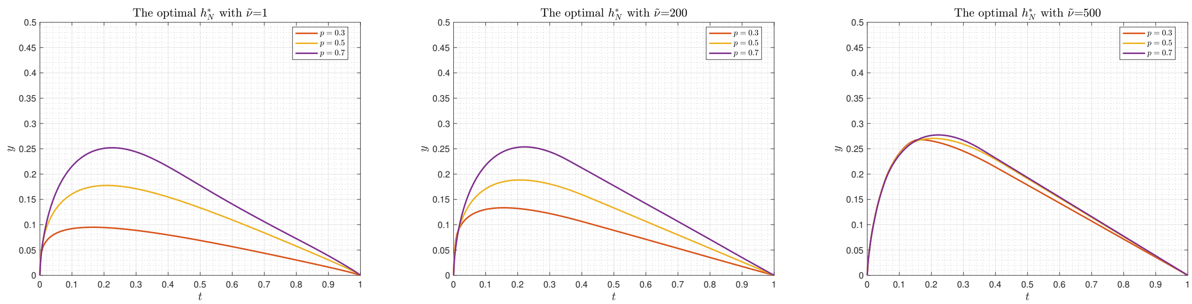

Figure 3: The optimal h_N^* for different values of the parameter p with the standard deviation bound given by $\tilde{\nu} = 1, 200$, and 500.

Figure 3, we can see that all $h_N^*(t)$ are bounded by 0.5 for different values of the parameter p and the bound $\tilde{\nu}$, which is supported by Remark 3.1. For relatively small ν (e.g., $\tilde{\nu} = 1, 200$), $h_N^*(t)$ has the largest derivative when $t \to 0_+$ for $p = 0.3$. This is because $\tilde{g}(t)$ has the smallest derivative when $t \to 0_+$ for $p = 0.3$, see Figure A.1. Moreover, for relatively large $\tilde{\nu}$ (e.g., $\tilde{\nu} = 500, 800$), the differences between $h_N^*(t)$ for different p become smaller. This might be due to the fact that $h_N^*(t)$ has a smaller impact on the objective function of Problem (3.16), whereas λ^* would have a more significant impact when $\tilde{\nu}$ is relatively large.

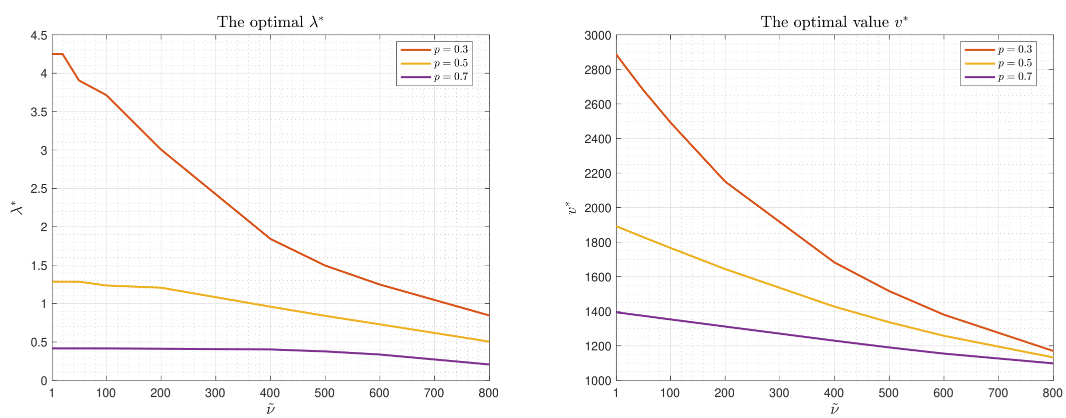

Figure 4: (Left): The optimal λ^* for different parameters of g_1 as a function of the standard deviation bound $\tilde{\nu}$; (Right): The optimal value v^* for different parameters of g_1 as a function of the standard deviation bound $\tilde{\nu}$.

Figure 4 displays the optimal solution λ^* (Left) and the optimal values v^* (Right) to Problem (3.16) for different parameters of g_1 with respect to the increase of the standard deviation bound $\tilde{\nu}$. Table A.1 in Appendix A presents the values of the optimal λ^* for different parameters of g_1 and various values of $\tilde{\nu}$. From Figure 4, we can see that the optimal solution λ^* of Problem (3.16) decreases for all considered values of p as the standard deviation bound $\tilde{\nu}$ increases. This echoes

Proposition 3.3. Moreover, for any fixed $\tilde{\nu}$, we can see that the optimal solution to Problem (3.16) is also decreasing as p increases. This is because $\tilde{g}(t)$ has the smallest value at each $t \in [0, 1]$ for $p = 0.3$, see Figure A.1 in Appendix A.

Figure 5 displays the function $\tilde{g}(S_{X_N}(x)) + \lambda^* h_N^*(S_{X_N}(x))$ for different parameters of g_1 when the standard deviation bounds are given by $\tilde{\nu} = 1, 200, 500$, and 800. For other values of $\tilde{\nu}$, see Figure A.2 in Appendix A for details. From Figures 5 and A.2, we can observe that there always exists

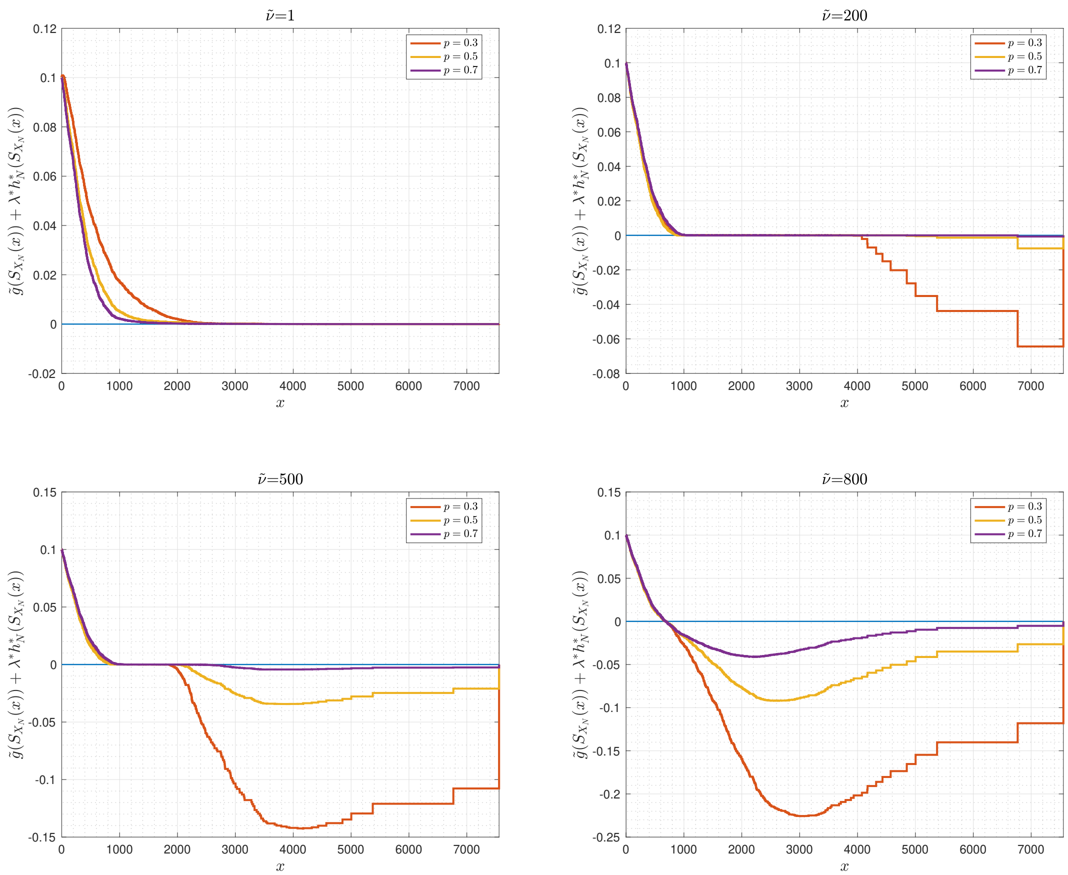

Figure 5: The function $\tilde{g}(S_{X_N}(x)) + \lambda^* h_N^*(S_{X_N}(x))$ for different parameters of g_1 when the standard deviation bounds are given by $\tilde{\nu} = 1, 200, 500$, and 800.

an interval, denoted by E_+, on which the function $\tilde{g}(S_{X_N}(x)) + \lambda^* h_N^*(S_{X_N}(x))$ remains positive in all considered cases. This is mainly because the function $\tilde{g}(t)$, as shown in Figure A.1 (Right), is positive on some interval for all p, and consequently the function $\tilde{g}(t) + \lambda^* h_N^*(t)$ is positive on some interval for all p and $\tilde{\nu}$ since $0 \neq h_N^*(t) \in \widetilde{\mathcal{H}}$ and $\lambda^* > 0$. From this observation, we can conclude that there always exists a deductible for the optimal indemnity $I^*_{\lambda^*, h_N^*}$, and the deductible is non-zero in all considered cases. Moreover, we can observe that the set $E = \{x \in [0, M) : \tilde{g}(S_{X_N}(x)) + \lambda^* h_N^*(S_{X_N}(x)) = 0\}$ is non-zero when $\tilde{\nu}$ is relatively small (e.g., $\tilde{\nu} = 1, 5, 100, 200, 400, 500, 600$) but becomes zero when $\tilde{\nu}$ is relatively large (e.g., $\tilde{\nu} = 800$). From this observation, we can infer that the optimal indemnity function is not unique for relatively small values of $\tilde{\nu}$ but is unique for relatively large values of $\tilde{\nu}$. Furthermore, based on the discussions of the optimal indemnity function shown in Figure 1, if we select $\xi(s)$ on the interval E as a piecewise linear function with two slopes, 0 and 1 (as in Figure 1 (Middle)), the optimal indemnity function I will be a stop-loss function and the parameter d is determined by solving the equation $\tilde{\nu} = \sqrt{\mathrm{Var}((X_N - d)_+)}$, as shown in Table 1 for different values of the standard deviation bound $\tilde{\nu}$. It is important to note that the deductible level

Table 1: The optimal deductible level d for different values of the standard deviation bound $\tilde{\nu}$

$\tilde{\nu}$	1	5	10	20	50	100	200	400	500	600	800
d	7525.5	7399.0	7240.9	6924.7	6115.7	4977.7	3468.9	2105.5	1687.1	1332.6	690.5

d is independent of the parameter p of g_1, which is essentially different from the models without variance constraint (see Figure 2 (Right)). Furthermore, as the standard deviation bound $\tilde{\nu}$ tightens, the optimal deductible increases for all considered cases. This is because as the constraint tightens, the insured cedes fewer losses while the insurer is exposed to less tail risk, consistent with recent empirical findings in Armantier et al. (2023).

Next, we focus on the case where $\xi(s)$ on the interval E is a linear function (as shown in Figure 1 (Left)) because it allows us to gain more insights into the impact of the parameter p of g_1 and $\tilde{\nu}$ on the optimal indemnity function. Figure 6 displays the optimal indemnity function I^* where $\xi(s)$ is linear on the interval E for different parameters of g_1 when the standard deviation bounds $\tilde{\nu}$ are given by 1, 200, 500, and 800. For other cases of $\tilde{\nu}$, see details in Figure A.3 in Appendix A. From Figures 6 and A.3, we can see that the deductible level d for stop-loss optimal indemnity function I^* always lies in the set E, denoted by $[x_1, x_2]$, which is consistent with the fact that $\mathrm{Var}((X_N - x_2)_+) < \nu = \mathrm{Var}((X_N - d)_+) < \mathrm{Var}((X_N - x_1)_+)$ for all considered cases. For a fixed p, we can see that both the length of the interval E and the value of x_2 decrease as $\tilde{\nu}$ increases. This indicates that the layered optimal indemnity function I^* increasingly resembles a stop-loss indemnity as the variance constraint becomes less restrictive. Moreover, for a fixed $\tilde{\nu}$, the parameter p of g_1 has less impact on the interval E when $\tilde{\nu}$ is relatively small (e.g., $\tilde{\nu} = 1, 5$) or $\tilde{\nu}$ is relatively large (e.g., $\tilde{\nu} = 600, 800$). In these cases, the variance constraint either severely limits or significantly relaxes the structure of the optimal indemnity function, making the influence of p on the interval E less significant. However, for intermediate values of $\tilde{\nu}$, the parameter p has a more noticeable impact on the length and positioning of the interval E, leading to greater sensitivity of the layered optimal indemnity function I^* to changes in p. To establish more concrete insights regarding the impact of the parameter p on the optimal layered indemnity function I^*, further extensive numerical tests are necessary.

The numerical examples above illustrate that the optimal ceded loss function of the insured is of the stop-loss treaty when the premium is charged via the expected-value premium principle and g_1 is a PH transform. As shown in Theorem 3.2, the solutions may exhibit multiple layers, depending on the specific shapes of the distortion functions g_1 and g_2. Figure A.4 in Appendix A shows that the optimal indemnity function becomes a limited stop-loss contract when the insured is a CVaR minimizer and g_2 is an inverse S-shaped distortion function.

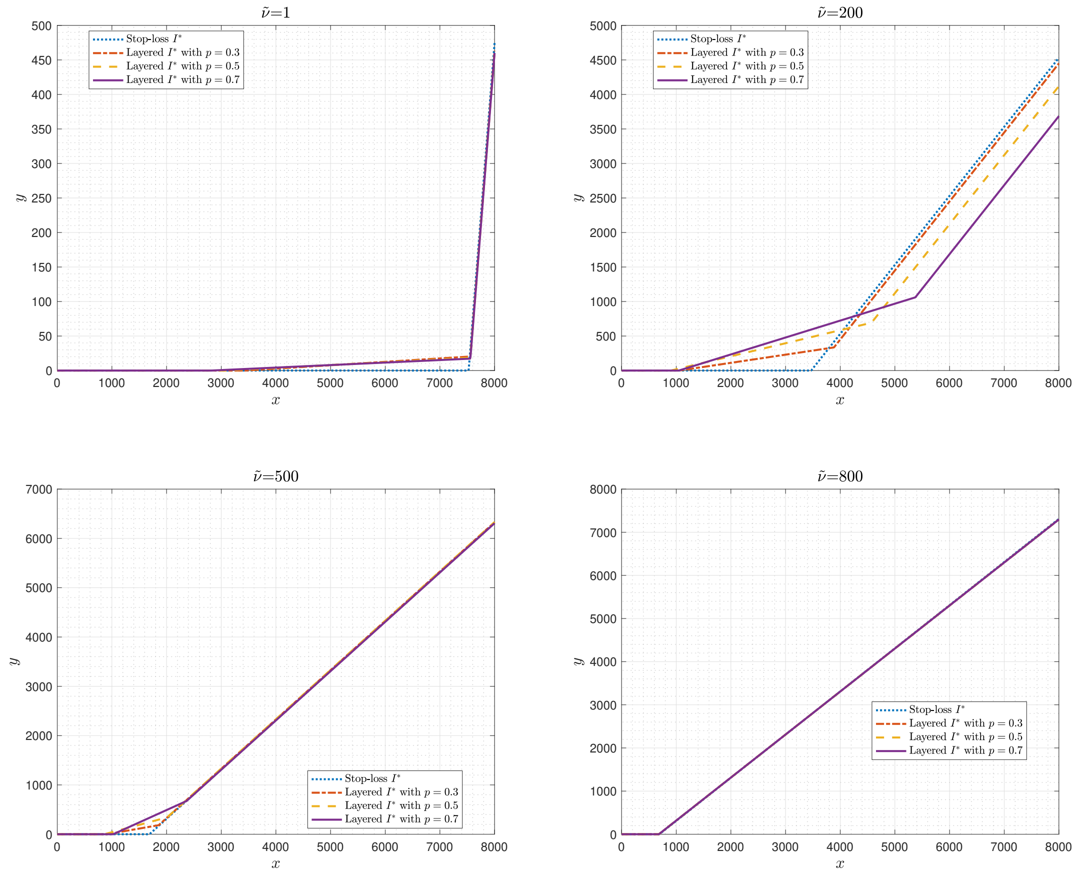

Figure 6: The optimal indemnity function I^* when $\xi(s)$ is linear on the interval E for different parameters of g_1. The standard deviation bounds are given by $\tilde{\nu} = 1, 200, 500$, and 800.

6.2 The performance of the proposed computational scheme

In this section, we conduct extensive numerical tests to evaluate the performance of the proposed computational scheme. Specifically, we will carry out the following numerical tests. (i) The convergence of the optimal values and optimal solutions of the SAA counterpart of the problem. (ii) The CPU time of solving the SAA counterpart of the problem.

(i) Convergence analysis

In this test, we randomly generate 100 data sets for each N and solve the problem in(3.16) via solving the problem in (4.9) to obtain the optimal value v_N^* and the optimal solution λ_N^* for each data set. Figure 7 depicts a boxplot of the optimal solution and the optimal value for different parameters of g as N increases when the standard deviation bound is $\tilde{\nu} = 800$. We can see that for all considered parameters of g, the optimal value v_N^* converges rapidly, while the optimal solution λ_N^* exhibits a slightly slower convergence.

(ii) CPU time

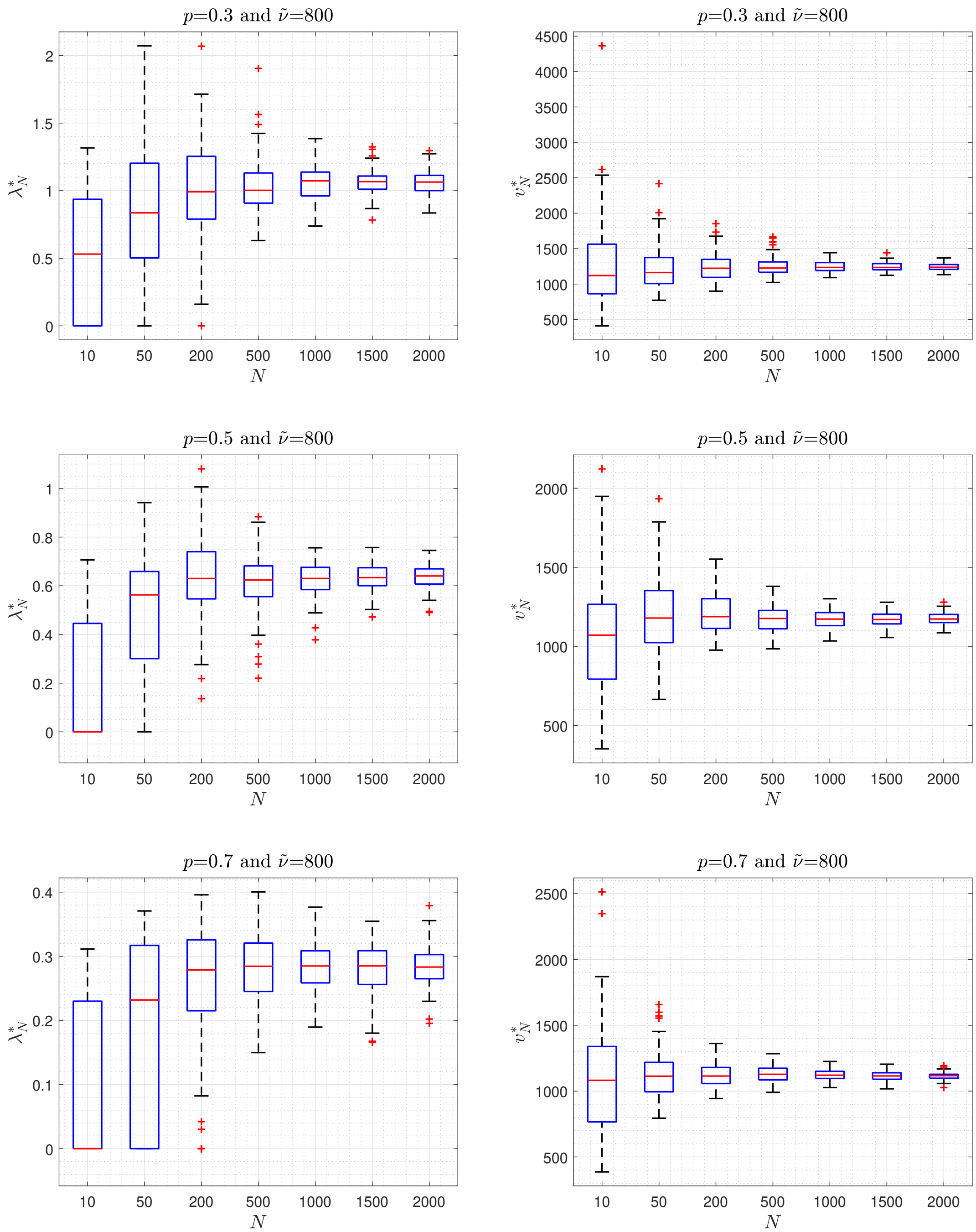

Figure 7: The boxplot for the optimal λ_N^* and the optimal value $\tilde{v}_N^*$ to Problem (3.16) for different parameters of g as a function of the sample size N, when the standard deviation bound is given by $\tilde{\nu} = 800$.

In this test, we present the solution time for Problem (3.16) as a function of the sample size N. Specifically, we solve the problem in (4.9) using the solver *coneprog* on Matlab 2024a with an interior-point algorithm. To ensure stability in the solution times, we randomly generate 100 data sets for each N and record the solution time for each data set. The average CPU time as a function of the sample size N is shown in Figure 8. As shown in Figure 8, the solution time grows

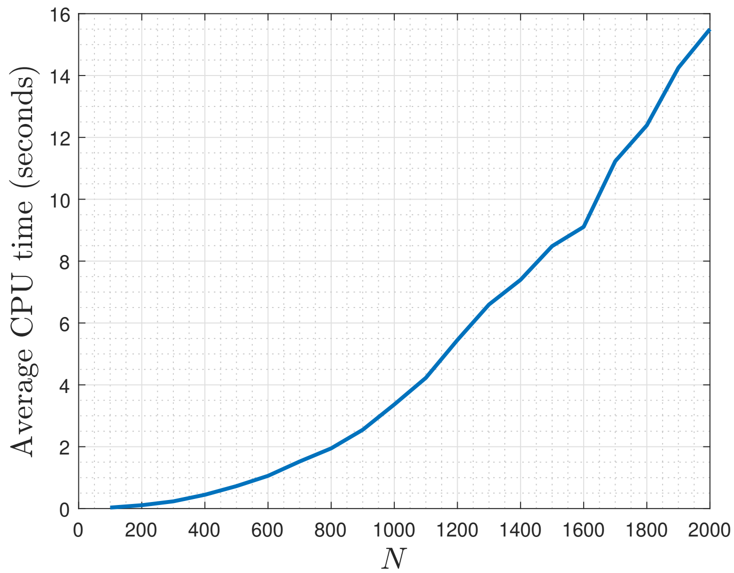

Figure 8: The average CPU time for solving the problem as a function of the sample size N.

gradually with the sample size N, which can be attributed to the fact that the proposed tractable reformulation to Problem (4.9) is a finite-dimensional second order cone program of reasonable size.

7 Concluding remarks

In this paper, we consider the optimal insurance design problem under the DRMs by introducing a variance constraint on the insurer's risk exposure, which is motivated by the insurer's need to manage underwriting risk effectively. However, while it serves to control risk, it also poses considerable technical challenges in solving the problem. We develop an approach to determine the structure of the optimal indemnity function analytically when the variance constraint is binding. To determine the parameters in the optimal contract, we reformulate the proposed problem as a second-order cone program by assuming that the underlying random loss is discrete. When the underlying loss is continuously distributed, we propose using the sample average approximation technique to approximate the parameters in the optimal contract. The convergence results are also established. We provide some numerical experiments to support our theoretical results. Although considering a variance constraint to control the insurer's risk exposure in actuarial science is rare, our results provide essential economic insights that explain some prevalent provisions in insurance policies. It is important to note that the optimal indemnity function can also be derived based on the convex representation of the variance (Liu et al., 2020, Example 2.2), following the same line of argument as in the paper.

There are several potential directions for extending this work. First, we may extend the current model to the case of general convex deviation measures (Rockafellar et al., 2006; Han et al., 2025) to control the insurer's risk exposure. Second, if the insured has a limited budget for the insurance premium, then some premium constraint might be considered. Third, it might be interesting to analyze the impact of the variance constraint on the optimal contract under a more general framework, such as rank-dependent expected utility and cumulative prospect theory.

Acknowledgements

The authors are grateful to three anonymous referees for their constructive comments and suggestions, which significantly improved the paper. The research of Wei Wang is supported by the Fundamental Research Funds for the Central Universities and the National Natural Science Foundation of China (No. 12501456). Wenjun Jiang acknowledges the financial support received from the Natural Sciences and Engineering Research Council of Canada (grant RGPIN-2020-04204), Alberta Innovates and Canadian Institute of Actuaries. Yiying Zhang acknowledges the financial support from the National Natural Science Foundation of China (No. 12571513), and Shenzhen Science and Technology Program (No. JCYJ20230807093312026, No. JCYJ20250604144210013).

Disclosure statements

No potential competing interests were reported by the authors.

References

Acerbi, C. (2002). Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking & Finance*, *26*, 1505–1518.

Allais, M. (1953). Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école américaine. *Econometrica*, *21*, 503–546.

Armantier, O., Foncel, J., & Treich, N. (2023). Insurance and portfolio decisions: Two sides of the same coin? *Journal of Financial Economics*, *148*, 201–219.

Arrow, K. J. (1963). Uncertainty and the welfare economics of medical care. *American Economic Review*, *53*, 941–973.

Arrow, K. J. (1974). Optimal insurance and generalized deductibles. *Scandinavian Actuarial Journal*, *1974*, 1–42.

Artzner, P., Delbaen, F., Eber, J.-M., & Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, *9*, 203–228.

Asimit, A. V., Bignozzi, V., Cheung, K. C., Hu, J., & Kim, E.-S. (2017). Robust and pareto optimality of insurance contracts. *European Journal of Operational Research*, *262*, 720–732.

Asimit, A. V., Cheung, K. C., Chong, W. F., & Hu, J. (2020). Pareto-optimal insurance contracts with premium budget and minimum charge constraints. *Insurance: Mathematics and Economics*, *95*, 17–27.

Asimit, V., & Boonen, T. J. (2018). Insurance with multiple insurers: A game-theoretic approach. *European Journal of Operational Research*, *267*, 778–790.

Assa, H. (2015). On optimal reinsurance policy with distortion risk measures and premiums. *Insurance: Mathematics and Economics*, *61*, 70–75.

Bernard, C., He, X., Yan, J.-A., & Zhou, X. Y. (2015). Optimal insurance design under rank-dependent expected utility. *Mathematical Finance*, *25*, 154–186.

Birghila, C., Boonen, T. J., & Ghossoub, M. (2023). Optimal insurance under maxmin expected utility. *Finance and Stochastics*, *27*, 467–501.

Boonen, T. J., & Jiang, W. (2024). Robust insurance design with distortion risk measures. *European Journal of Operational Research*, *316*, 694–706.

Borch, K. (1960). An attempt to determine the optimum amount of stop loss reinsurance. In *Transactions of the 16th International Congress of Actuaries* (pp. 597–610). Uden forlag.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Cai, J., Liu, F., & Yin, M. (2024). Worst-case risk measures of stop-loss and limited loss random variables under distribution uncertainty with applications to robust reinsurance. *European Journal of Operational Research*, *318*, 310–326.

Chi, Y., & Wei, W. (2020). Optimal insurance with background risk: An analysis of general dependence structures. *Finance and Stochastics*, *24*, 903–937.

Chi, Y., Zhou, X. Y., & Zhuang, S. C. (2024). Variance insurance contracts. *Insurance: Mathematics and Economics*, *115*, 62–82.

Chi, Y., & Zhuang, S. C. (2022). Regret-based optimal insurance design. *Insurance: Mathematics and Economics*, *102*, 22–41.

Cui, W., Yang, J., & Wu, L. (2013). Optimal reinsurance minimizing the distortion risk measure under general reinsurance premium principles. *Insurance: Mathematics and Economics*, *53*, 74–85.

Denuit, M., Dhaene, J., Goovaerts, M., & Kaas, R. (2006). *Actuarial Theory for Dependent Risks: Measures, Orders and Models*. John Wiley & Sons.

Dhaene, J., Vanduffel, S., Goovaerts, M. J., Kaas, R., Tang, Q., & Vyncke, D. (2006). Risk measures and comonotonicity: a review. *Stochastic Models*, *22*, 573–606.

Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *Quarterly Journal of Economics*, *75*, 643–669.

Fadina, T., Hu, J., Liu, P., & Xia, Y. (2025). Optimal reinsurance with multivariate risks and dependence uncertainty. *European Journal of Operational Research*, *321*, 231–242.

Fan, K. (1953). Minimax theorems. *Proceedings of the National Academy of Sciences*, *39*, 42–47.

Föllmer, H., & Schied, A. (2011). *Stochastic Finance: An Introduction in Discrete Time*. Walter de Gruyter.

Ghossoub, M. (2019). Optimal insurance under rank-dependent expected utility. *Insurance: Mathematics and Economics*, *87*, 51–66.

Goovaerts, M. J., Kaas, R., & Laeven, R. J. (2010). Decision principles derived from risk measures. *Insurance: Mathematics and Economics*, *47*, 294–302.

Han, X., Wang, R., & Wu, Q. (2025). Monotonic mean-deviation risk measures. *Finance and Stochastics*, *Forthcoming*, 1–36.

Huberman, G., Mayers, D., & Smith Jr, C. W. (1983). Optimal insurance policy indemnity schedules. *Bell Journal of Economics*, *14*, 415–426.

Jin, Z., Xu, Z. Q., & Zou, B. (2024). Optimal moral-hazard-free reinsurance under extended distortion premium principles. *SIAM Journal on Control and Optimization*, *62*, 1390–1416.

Kaluszka, M. (2001). Optimal reinsurance under mean-variance premium principles. *Insurance: Mathematics and Economics*, *28*, 61–67.

Kaluszka, M. (2005). Optimal reinsurance under convex principles of premium calculation. *Insurance: Mathematics and Economics*, *36*, 375–398.

Kaye, P. (2005). Risk measurement in insurance a guide to risk measurement, capital allocation and related decision support issues. *Casualty Actuarial Society Discussion Paper*, (pp. 1–34).

Klenke, A. (2013). *Probability theory: a comprehensive course*. Springer Science & Business Media.

Kleywegt, A. J., Shapiro, A., & Homem-de Mello, T. (2002). The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, *12*, 479–502.

Liu, F., Cai, J., Lemieux, C., & Wang, R. (2020). Convex risk functionals: Representation and applications. *Insurance: Mathematics and Economics*, *90*, 66–79.

Mossin, J. (1968). Aspects of rational insurance purchasing. *Journal of Political Economy*, *76*, 553–568.

Pomerol, J. C. (1981). The boundedness of the Lagrange multipliers set and duality in mathematical programming. *Zeitschrift für Operations Research*, *25*, 191–204.

Raviv, A. (1979). The design of an optimal insurance policy. *American Economic Review*, *69*, 84–96.

Rockafellar, R. T., Uryasev, S., & Zabarankin, M. (2006). Generalized deviations in risk analysis. *Finance and Stochastics*, *10*, 51–74.

Shapiro, A., Dentcheva, D., & Ruszczynski, A. (2021). *Lectures on Stochastic Programming: Modeling and Theory*. SIAM.

Tan, K. S., Wei, P., Wei, W., & Zhuang, S. C. (2020). Optimal dynamic reinsurance policies under a generalized Denneberg's absolute deviation principle. *European Journal of Operational Research*, *282*, 345–362.

Tan, K. S., & Weng, C. (2014). Empirical approach for optimal reinsurance design. *North American Actuarial Journal*, *18*, 315–342.

Vajda, S. (1962). Minimum variance reinsurance. *ASTIN Bulletin: The Journal of the IAA*, *2*, 257–260.

Wang, R., Wei, Y., & Willmot, G. E. (2020). Characterization, robustness, and aggregation of signed Choquet integrals. *Mathematics of Operations Research*, *45*, 993–1015.

Wang, S. S. (1995). Insurance pricing and increased limits ratemaking by proportional hazards transforms. *Insurance: Mathematics and Economics*, *17*, 43–54.

Wang, S. S. (1996). Premium calculation by transforming the layer premium density. *ASTIN Bulletin: The Journal of the IAA*, *26*, 71–92.

Wang, S. S., Young, V. R., & Panjer, H. H. (1997). Axiomatic characterization of insurance prices. *Insurance: Mathematics and Economics*, *21*, 173–183.

Wang, W., & Xu, H. (2023). Preference robust distortion risk measure and its application. *Mathematical Finance*, *33*, 389–434.

Xu, Z. Q., Zhou, X. Y., & Zhuang, S. C. (2019). Optimal insurance under rank-dependent utility and incentive compatibility. *Mathematical Finance*, *29*, 659–692.

Yaari, M. E. (1987). The dual theory of choice under risk. *Econometrica*, *55*, 95–115.

Young, V. R. (1999). Optimal insurance under Wang's premium principle. *Insurance: Mathematics and Economics*, *25*, 109–122.

Zheng, J. (2020). Optimal insurance design under narrow framing. *Journal of Economic Behavior & Organization*, *180*, 596–607.

Zhuang, S. C., Weng, C., Tan, K. S., & Assa, H. (2016). Marginal indemnification function formulation for optimal reinsurance. *Insurance: Mathematics and Economics*, *67*, 65–76.

Appendix A Additional numerical results

In this appendix, we collect some other illustrative numerical results.

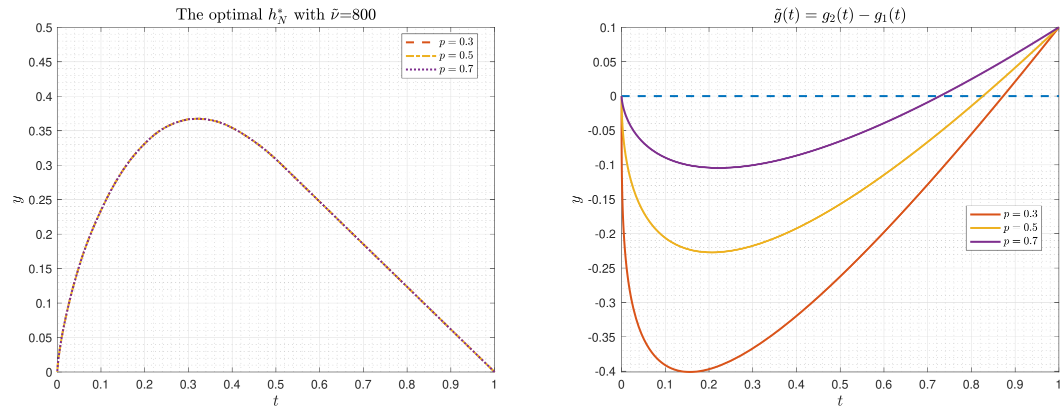

Figure A.1: (Left) The optimal h_N^* for different values of the parameter p with the standard deviation bound $\tilde{\nu} = 800$; (Right): The function $\tilde{g}(t)$ for different values of the parameter p.

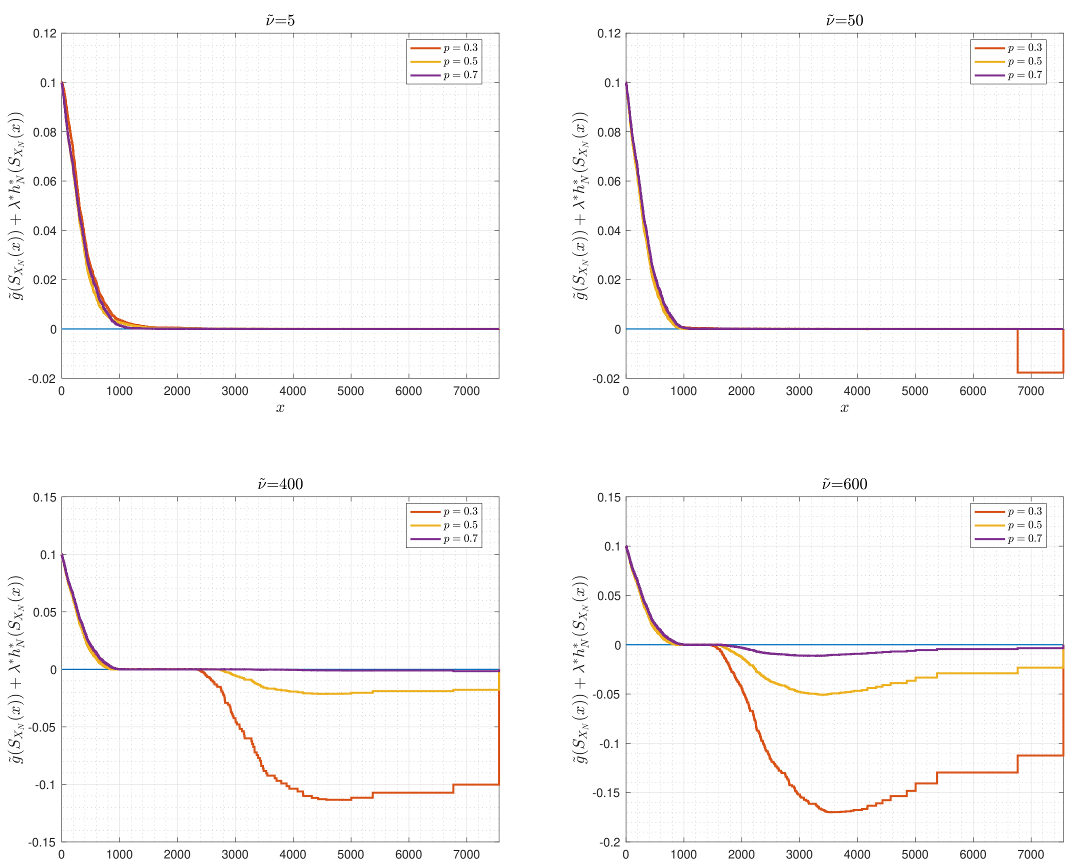

Figure A.2: The function $\tilde{g}(S_{X_N}(x)) + \lambda^* h_N^*(S_{X_N}(x))$ for different parameters of g_1, when the standard deviation bounds are given by $\tilde{\nu} = 5, 50, 400$, and 600.

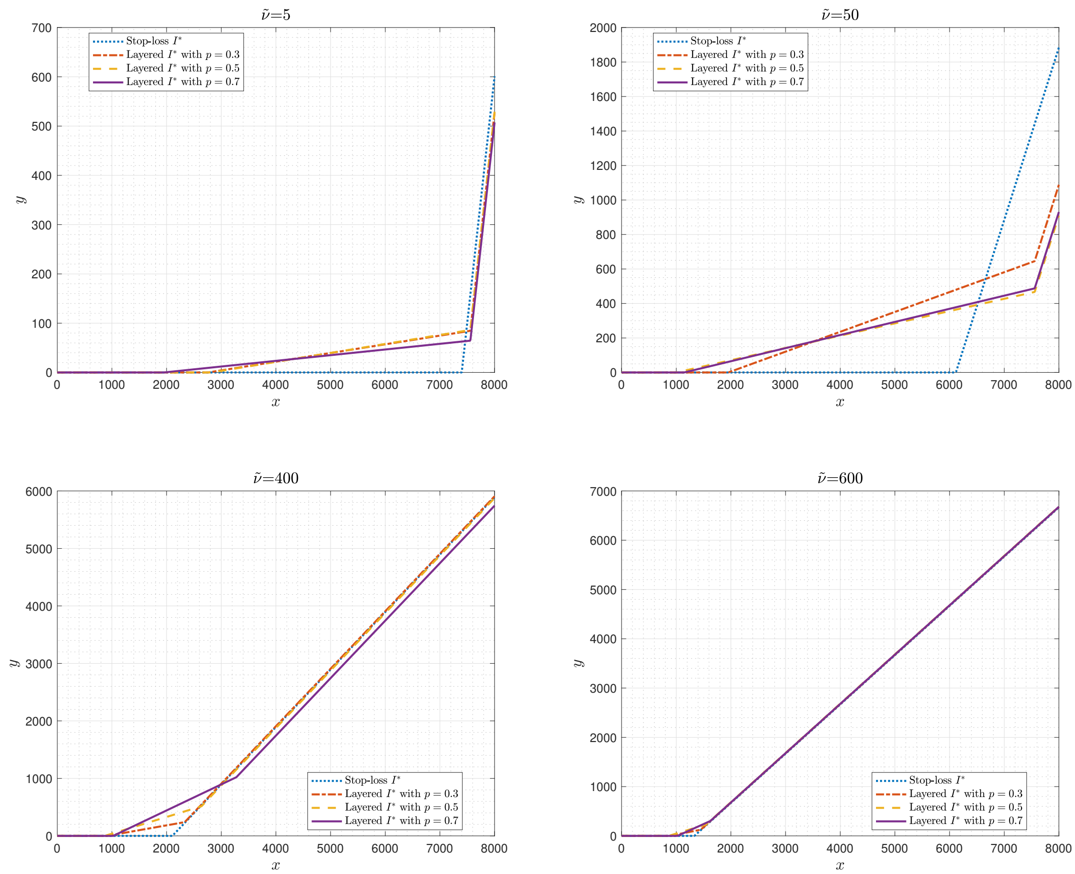

Figure A.3: The optimal indemnity function I^* where $\xi(s)$ is linear on the interval E for different parameters of g_1, when the standard deviation bounds are given by $\tilde{\nu} = 5, 50, 400,$ and 600.

Table A.1: The optimal λ^* for different values of the parameter p and the standard deviation bound $\tilde{\nu}$.

p \ $\tilde{\nu}$	1	5	10	20	50	100	200	400	500	600	800
0.3	4.2508	4.2492	4.2490	4.2488	3.9055	3.7146	3.0054	1.8415	1.4948	1.2746	0.8450
0.5	1.2863	1.2852	1.2849	1.2847	1.2845	1.2353	1.2067	0.9592	0.8402	0.7294	0.5047
0.7	0.4166	0.4160	0.4159	0.4158	0.4158	0.4157	0.4119	0.4024	0.3768	0.3364	0.2072

We also supplement another numerical example by looking at another case of g_1 and g_2. Most parts of the setup remain the same, except for the distortion functions and the safety loading factor. We remark that specific parameters are adopted for illustration purposes.

- The insured is a CVaR minimizer, whose distortion function is $g_1(t) = \frac{t}{1-\alpha} \wedge 1$, and the insurer prices the insurance using an inverse S-shaped distortion function[5], i.e. $g_2(t) = \frac{t^\gamma}{(t^\gamma + (1-t)^\gamma)^{\frac{1}{\gamma}}}$. We set $\alpha = 0.6$ and $\gamma = 0.65$.

- The safety loading factor is now set to $\theta = 0.3$.

The optimal indemnity functions under the new setting are shown in Figure A.4. Notably, the optimal indemnity function without the variance constraint is of a limited stop-loss form, which can be derived based on Figure A.4 (Left). In the presence of variance constraint, the optimal parametric form of the indemnity function remains the same, with parameters adjusted to satisfy the constraint. When the constraint becomes tighter, the insurance provides less coverage. This is consistent with our observations in the first numerical example.

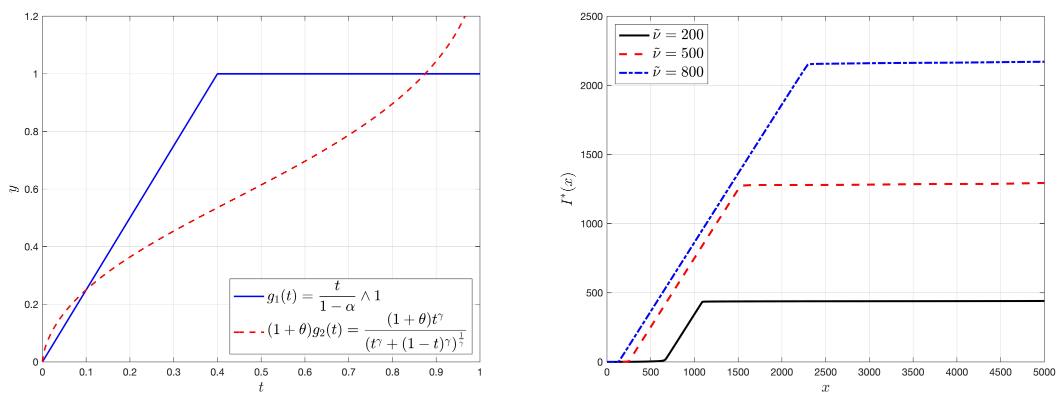

Figure A.4: (Left) The distortion functions $g_1(t)$ and $(1+\theta)g_2(t)$; (Right): The optimal indemnity functions under the new setting.

[5]The inverse S-shaped distortion function is popularized due to the development of behavioral insurance, where decision makers would exaggerate the probabilities of extremely good and extremely bad results.