

Optimal dividend and capital injection under self-exciting claims

Paulin Aubert* Etienne Chevalier† Vathana Ly Vath‡

November 26, 2025

Abstract

In this paper, we study an optimal dividend and capital-injection problem in a Cramér-Lundberg model where claim arrivals follow a Hawkes process, capturing clustering effects often observed in insurance portfolios. We establish key analytical properties of the value function and characterise the optimal capital-injection strategy through an explicit threshold. We also show that the value function is the unique viscosity solution of the associated HJB variational inequality. For numerical purposes, we first compute a benchmark solution via a monotone finite-difference scheme with Howard's policy iteration. We then develop a reinforcement learning approach based on policy-gradient and actor-critic methods. The learned strategies closely match the PDE benchmark and remain stable across initial conditions. The results highlight the relevance of policy-gradient techniques for dividend optimisation under self-exciting claim dynamics and point toward scalable methods for higher-dimensional extensions.

Keywords: Optimal dividend, Singular stochastic control, Hawkes processes, Viscosity solutions, Reinforcement learning, Policy gradient.

1 Introduction

The allocation of an insurer's surplus between solvency and shareholder remuneration is a central question in actuarial science, traditionally addressed through ruin probabilities and optimal dividend policies. In this context, the surplus process provides a framework for quantifying the trade-off between long-term financial stability and the distribution of profits. Since the seminal contribution of de Finetti [10], a vast literature has emerged at the intersection of probability theory, stochastic control, and insurance mathematics.

Classical studies build upon the Cramér–Lundberg model introduced by Lundberg [24] and Cramér [8], and further developed by Gerber [12, 11]. Over the past decades, the dividend optimisation problem has been analysed using both regular and singular control techniques in models driven by compound Poisson processes or Brownian motion. See for instance Jeanblanc and Shiryaev [17], Asmussen and Taksar [4], and Gerber and Shiu [14, 13]. Numerous extensions have since been proposed to incorporate investment risk, reinsurance, capital injections, and taxation, as documented in the works of Paulsen and Gjessing [26], Hojgaard and Taksar [16], Azcue and Muler [5], Kulenko and Schmidli [21], Lokka and Zervos [23], and Albrecher and Thonhauser [2]. Comprehensive reviews of these developments can be found in Albrecher and Thonhauser [3] and in the monograph by Schmidli [29]. A persistent assumption in the classical literature is that claim arrivals are independent and identically distributed, typically modelled by a Poisson process. Yet real insurance portfolios—particularly those exposed to catastrophic, environmental, cyber, or systemic risks—often

*Laboratoire de Mathématiques et Modélisation d'Évry, Université Évry Paris-Saclay, Exiom Partners, France, paulin.aubert@univ-evry.fr.

†Laboratoire de Mathématiques et Modélisation d'Évry, Université Évry Paris-Saclay, UMR 8071 CNRS, France, etienne.chevalier@univ-evry.fr.

‡Laboratoire de Mathématiques et Modélisation d'Évry, Université Paris-Saclay, ENSIIE, UMR 8071 CNRS, France, vathana.lyvath@ensiie.fr.

display pronounced clustering, generating temporal dependence in claim occurrences. This has motivated the use of more general point processes, including Cox and shot-noise dynamics [1], and more recently Hawkes processes, as studied by Brachetta, Callegaro, Ceci, and Sgarra [7]. Dividend optimisation has been analysed in some of these non-Poisson settings. However, the combined optimisation of dividends and capital injections in a Cramér-Lundberg model driven by Hawkes claim arrivals has not been addressed in the existing literature. The present work develops a dividend optimisation framework for a Cramér–Lundberg model with Hawkes claim arrivals, allowing for both dividend distributions and capital injections. This extends classical results obtained under compound Poisson dynamics, including those of Kulenko and Schmidli [21]. From an analytical perspective, we establish fundamental properties of the value function, including bounds, monotonicity, and local Lipschitz continuity, and characterise the optimal capital-injection strategy through an explicit threshold. We then show that the value function is the unique viscosity solution of the associated Hamilton–Jacobi–Bellman variational inequality.

Because Hawkes dynamics considerably increase the analytical complexity of the model, numerical methods are required to approximate the value function and the associated optimal policy. As a classical benchmark, we first compute a reference solution using a monotone finite-difference approximation of the HJB variational inequality combined with Howard’s policy iteration algorithm. This PDE-based approach serves to validate the structure of the optimal strategy in our setting. The main numerical contribution of the paper lies in the development of a reinforcement learning methodology tailored to this class of singular stochastic control problems. A growing body of work aims to connect stochastic control theory with reinforcement learning by developing policy-gradient and actor–critic formulations, as illustrated by the contributions of Wang et al. [31], Jia and Zhou [19, 18, 20], as well as the recent advances of Hamdouche et al. [15] and Pham and Warin [27]. Building on these developments, we examine whether parameterised stochastic policies can learn near-optimal dividend and capital-injection strategies in our setting. Our methodology is related to the framework of Hamdouche et al. [15], who study policy-gradient approaches for control problems with random exit times. The results obtained reinforce the view that policy-gradient algorithms offer a scalable alternative to PDE-based methods, and can be effectively applied to higher-dimensional or path-dependent extensions of the dividend optimisation problem where classical numerical techniques become impractical.

The remainder of the paper is structured as follows. Section 2 introduces the surplus model with capital injections and Hawkes-driven claims. Section 3 establishes key analytical properties of the value function and derives the structure of the optimal injection strategy. Section 3.5 shows that the value function is the unique viscosity solution to the associated HJB variational inequality. Section 4 presents the finite-difference framework and the corresponding numerical results, which serve as a benchmark and illustrate the economic features of the optimal policy. Finally, Section 5 develops the reinforcement learning methodology and compares the learned strategies with the PDE benchmark.

2 Modelling insurer’s portfolio and clustering effect

2.1 Uncontrolled surplus dynamics

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space on which all stochastic processes and random variables are defined and such that \mathbb{F} is complete and right-continuous. The insurer’s cash reserve is represented by a stochastic process $R = (R_t)_{t \geq 0}$, whose dynamics, in the absence of any control, follow the classical Cramér-Lundberg model:

$$R_t^x = x + ct - \sum_{k=1}^{N_t} Y_k,$$

where $x \in \mathbb{R}^+$ is the company’s initial capital, $c > 0$ is the constant premium income per unit of time, $N = (N_t)_{t \geq 0}$ is a counting process representing the number of claims occurring up to time t and $(Y_k)_{k \in \mathbb{N}}$ is a positive random variable with density f , assumed independent of the counting process.

Traditionally, N is assumed to be a homogeneous Poisson process with constant intensity $\lambda > 0$, which implies independent, exponentially distributed inter-arrival times. While analytically convenient, this framework is not designed to account for temporal dependence in claim arrivals, which motivates the use of more flexible models such as self-exciting processes.

2.2 Temporal claim dependence via a Hawkes process

In practice, claim arrivals often exhibit temporal clustering: events such as natural disasters, cyber incidents, or pandemics tend to generate multiple claims in short time intervals. This behaviour, known as the clustering effect, contradicts the memoryless nature of the Poisson process.

To model the clustering behaviour of claims, we choose to represent the arrival process N as a Hawkes process. Hawkes processes are well known for their ability to model clustering effects. In our framework, the claim arrival intensity $\lambda = (\lambda_t)_{t \geq 0}$ evolves dynamically according to the following equation:

$$\lambda_t = a(b - \lambda_t)dt + \eta dN_t,$$

where $a, b, \eta > 0$ are model parameters, and the initial condition is $\lambda_0 = y \in [b, +\infty)$. Between claim arrivals, the intensity λ_t reverts toward the long-term level b at rate a , while each claim at time t increases λ_t by η . This dynamics captures both the self-exciting nature and the memory effects in claim arrivals. We assume $\lambda_0 \geq b$ without loss of generality. Indeed, under exponential kernels and as soon as a few claims occur, the intensity will almost surely exceed b and remain above it due to the accumulation of excitation. This assumption also simplifies several technical arguments in the analysis that follows.

2.3 Controlled surplus dynamics

We assume that the company is owned by a group of shareholders whose objective is to extract value from the surplus through dividend distributions, while preserving solvency via capital injections when needed. These two financial levers modify the surplus dynamics, leading to a controlled process.

Let $\alpha = (Z_t, K_t)_{t \geq 0}$ be a control strategy, where Z is a non-decreasing, right-continuous, \mathcal{F} -adapted process representing the cumulative dividends paid out to shareholders and K is a non-decreasing, left-continuous, \mathcal{F} -adapted process representing the cumulative capital injections by shareholders. Under strategy α the controlled surplus process is given by:

$$\begin{aligned} X_t &= R_t^x - Z_t + K_t \\ &= x + ct - \sum_{k=1}^{N_t} Y_k - Z_t + K_t. \end{aligned}$$

Dividend payments reduce the reserve, while capital injections increase it. These interventions are subject to economic constraints and are only permitted within an admissible set. To ensure both economic relevance and mathematical well-posedness of the model, we restrict our attention to a class of admissible strategies defined as follows:

Definition 2.1 (Set of admissible strategies). *A strategy $\alpha_t = (Z_t, K_t)_{t \geq 0}$ is said to be admissible if:*

- Z is càd-làg, \mathbb{F} -adapted, non-decreasing and such that $Z_t - Z_{t-} \leq X_{t-} + R_t^0 - R_{t-}^0$ and $Z_{0-} = 0$,
- K is càg-làd, \mathbb{F} -adapted, non-decreasing and such that $K_{0-} = 0$.

When $(X_0, \lambda_0) = (x, y) \in \mathbb{R} \times [b, +\infty)$, the set of admissible strategies is denoted by $\mathcal{A}(x, y)$.

The condition $Z_t - Z_{t-} \leq X_{t-} + R_t^0 - R_{t-}^0$ enforces that dividends cannot be paid beyond the available reserve at any time.

2.4 Ruin and objective function

As is standard in risk theory, we assume that the company ceases operations at the time of ruin, i.e., when its reserve becomes negative. The ruin time under strategy α is defined as:

$$T^\alpha = \inf\{t \geq 0, X_{t+} < 0\}.$$

The shareholders' objective is to maximize the expected discounted net gains until ruin. The gain includes the total discounted dividends and subtracts a penalty proportional to the capital injected. Formally, for $(x, y) \in \mathbb{R} \times [b, +\infty)$, the reward associated with a strategy $\alpha \in \mathcal{A}(x, y)$ is given by:

$$J_\alpha(x, y) = \mathbb{E} \left[\int_0^{T^\alpha} e^{-\rho s} dZ_s - \delta \int_0^{T^\alpha} e^{-\rho s} dK_s \right],$$

where $\rho > 0$ is the discount rate, and $\delta > 1$ is the penalty coefficient reflecting the opportunity cost of capital injections. The optimization problem then consists in maximizing $J_\alpha(x, y)$ over all admissible strategies:

$$v(x, y) = \sup_{\alpha \in \mathcal{A}(x, y)} J_\alpha(x, y) \quad \text{on } \mathbb{R} \times [b, +\infty). \quad (2.1)$$

Remark 2.1. *The condition $\delta > 1$ is crucial to prevent excessive capital injections, which would otherwise be incentivized if $\delta \leq 1$. Similarly, the discount rate $\rho > 0$ ensures finiteness of the value function and rules out infinite accumulation of dividends over time. See [21] for a detailed discussion.*

3 Theoretical analysis

We now examine the analytical and structural properties of the value function associated with the stochastic control problem (2.1). These results provide the mathematical foundations required to characterize the value function as a viscosity solution of the Hamilton–Jacobi–Bellman (HJB) equation, a task carried out in Section 3.5.

3.1 Pre-claim intensity

A recurring element in our analysis is the conditional behaviour of the claim intensity process prior to the first jump of the counting process N . In order to simplify computations involving the law of the first claim time, we introduce the deterministic intensity process $\tilde{\lambda}$, defined on the event $\{t \leq \tau_1\}$, where τ_1 denotes the first jump time of N . On $\{t \leq \tau_1\}$, the intensity process satisfies the deterministic ordinary differential equation:

$$d\lambda_t = a(b - \lambda_t)dt, \quad \lambda_0 = y \geq b,$$

which integrates explicitly to:

$$\tilde{\lambda}_t := \lambda_t = b - (b - y)e^{-at}. \quad (3.1)$$

This expression appears frequently in computations involving expectations conditional on the absence of claims. In particular, the following expression for the survival probability will be used repeatedly in the analysis. Let $h > 0$. The probability that no claim occurs up to time h is given by:

$$\begin{aligned} \mathbb{P}(\tau_1 \geq h) &= e^{-\int_0^h \tilde{\lambda}_s ds} \\ &= e^{-bh - \frac{y-b}{a}(1-e^{-ah})} \\ &\underset{h \rightarrow 0}{=} 1 - yh + o(h). \end{aligned}$$

This approximation is especially useful when analysing the infinitesimal behaviour of the controlled process, as will be required in the rest of this section.

3.2 Dynamic programming principle

We begin by establishing the dynamic programming principle (DPP) associated with the control problem (2.1). This fundamental result expresses the value function in terms of sequentially optimal decisions over time intervals, and serves as the cornerstone for deriving the Hamilton–Jacobi–Bellman equation and for analyzing the structural properties of the value function. In our problem, the dynamic programming principle can be stated as follows:

Proposition 3.1 (Dynamic Programming Principle). *Let θ be any \mathcal{F} -stopping time and $(x, y) \in \mathbb{R} \times [b, +\infty)$, it follows from the dynamic programming principle that*

$$v(x, y) = \sup_{\alpha \in \mathcal{A}(x, y)} \mathbb{E} \left[\int_0^{T^\alpha \wedge \theta} e^{-\rho s} dZ_s - \delta \int_0^{T^\alpha \wedge \theta} e^{-\rho s} dK_s + e^{-\rho(T^\alpha \wedge \theta)} v(X_{T^\alpha \wedge \theta}, \lambda_{T^\alpha \wedge \theta}) \right]. \quad (3.2)$$

We refer to standard texts (e.g., [21]) for further details and omit the proof, which follows classical arguments.

3.3 Analytical properties of the value function

We first derive upper and lower bounds for the value function. The lower bound is immediate, as the controller can always choose to take no action. The upper bound corresponds to an idealized scenario where all available surplus is instantly paid as dividends without receiving further claims.

Proposition 3.2 (Value function boundaries). *For $x \in \mathbb{R}$ and $y \geq b$, we have*

$$x^+ \leq v(x, y) \leq x^+ + \frac{c}{\rho}.$$

Proof. The lower bound follows by considering a strategy $\hat{\alpha}$ which immediately distributes the whole cash reserve, x^+ and then does not distribute any dividends and does not inject capital. We get

$$v(x, y) \geq J_{\hat{\alpha}}(x, y) = x^+, \quad \text{for } (x, y) \in \mathbb{R} \times [b, +\infty).$$

Let $(x, y) \in \mathbb{R}^+ \times [b, +\infty)$. We know that, for any strategy $\alpha = (Z, K)$, we have $0 \leq Z_u \leq x + cu + K_u$, on before $\{u \leq T^\alpha\}$, so we deduce that:

$$J_\alpha(x, y) \leq x + \mathbb{E} \left[\int_0^{T^\alpha} e^{-\rho s} c ds + (1 - \delta) \int_0^{T^\alpha} e^{-\rho s} dK_s \right] \leq x + \int_0^{+\infty} e^{-\rho s} c ds = x + \frac{c}{\rho}.$$

If $(x, y) \in \mathbb{R}^- \times [b, +\infty)$, there are only two admissible actions at time 0: letting the firm going to bankruptcy or injecting capital up to 0. Hence we have

$$v(x, y) \leq \max(x + v(0, y); 0) \leq \frac{c}{\rho}.$$

□

We next establish monotonicity properties, reflecting the natural intuition that higher surplus enhances value, whereas higher claim intensity reduces it.

Proposition 3.3 (Monotonicity in x). *Let $0 \leq x < x'$ and $y \in [b, +\infty)$. Then:*

$$v(x', y) - v(x, y) \geq x' - x.$$

Proof. Let $\varepsilon > 0$ and α_ε be an ε -suboptimal strategy, i.e., $J_{\alpha_\varepsilon}(x, y) \geq v(x, y) - \varepsilon$. Let $0 \leq x < x'$. We consider the strategy consisting in distributing dividends up to x and then apply strategy α_ε . By the dynamic programming principle (3.2) we obtain:

$$\begin{aligned} v(x', y) &\geq x' - x + J_{\alpha_\varepsilon}(x, y) \\ &\geq x' - x + v(x, y) - \varepsilon. \end{aligned}$$

Letting $\varepsilon \rightarrow 0$ yields the desired result. □

Proposition 3.4 (Monotonicity in y). *Let $x \in \mathbb{R}^+$. The function $y \rightarrow v(x, y)$ is non-increasing on $[b, +\infty)$.*

Proof. Let $x \in \mathbb{R}$ and $b \leq y < y'$. For $\varepsilon > 0$, there exists ε -suboptimal strategy $\alpha^\varepsilon = (Z^\varepsilon, K^\varepsilon) \in \mathcal{A}(x, y')$ such that $v(x, y') \leq J^{\alpha^\varepsilon}(x, y') + \varepsilon$.

From theorem 3.2 in [9] we know that the intensity process is such that $\lambda^y \leq \lambda^{y'}$ almost surely, which, from Lemma A.2 in [1] implies that $X^{\alpha^\varepsilon, y'} \leq X^{\alpha^\varepsilon, y}$ and therefore $T^{\alpha^\varepsilon, y'} \leq T^{\alpha^\varepsilon, y}$. From the dynamic programming principle we have:

$$\begin{aligned} v(x, y) &\geq \mathbb{E} \left[\int_0^{T^{\alpha^\varepsilon, y'}} e^{-\rho s} d(Z_s^\varepsilon - \delta K_s^\varepsilon) + e^{-\rho T^{\alpha^\varepsilon, y'}} v(X_{T^{\alpha^\varepsilon, y'}}^\varepsilon, \lambda_{T^{\alpha^\varepsilon, y'}}^\varepsilon) \right], \\ &\geq \mathbb{E} \left[\int_0^{T^{\alpha^\varepsilon, y'}} e^{-\rho s} d(Z_s^\varepsilon - \delta K_s^\varepsilon) \right] \\ &\geq v(x, y') + \varepsilon. \end{aligned}$$

As this is true for every $\varepsilon > 0$ we deduce, by letting ε going to 0, that v is non increasing in y . \square

Corollary 3.1. *For $x \in \mathbb{R}^+$, we have*

$$\lim_{y \rightarrow +\infty} v(x, y) = x.$$

Proof. Let $x \in \mathbb{R}^+$. For $y > e^b$, we set $t^*(y) = \frac{1}{a} \ln \left(\frac{y-b}{\ln(y)-b} \right)$. $t^*(y)$ is then such that

$$\lambda_t^y \geq \tilde{\lambda}_t^y \geq \hat{\lambda}_{t^*(y)}^y, \quad \text{for all } t \leq t^*(y).$$

Let $\varepsilon > 0$, it follows from the dynamic programming principle (3.1), that

$$v(x, y) \leq \varepsilon + \mathbb{E} \left[\int_0^{T^\alpha \wedge t^*(y)} e^{-\rho s} d(Z_s - \delta K_s) + e^{-\rho(T^\alpha \wedge t^*(y))} v(X_{T^\alpha \wedge t^*(y)}, \lambda_{T^\alpha \wedge t^*(y)}) \right]$$

As $\lim_{y \rightarrow +\infty} t^*(y) = +\infty$ and from the monotonicity of v in intensity, for y big enough, we have

$$\mathbb{E} \left[\int_0^{T^\alpha \wedge t^*(y)} e^{-\rho s} d(Z_s - \delta K_s) \mathbf{1}_{\{T^\alpha \leq t^*(y)\}} \right] \leq v^{\ln(y)}(x),$$

where we denote by v^ζ the value function of our control problem with a constant intensity equal to ζ . We recall that $\lim_{\zeta \rightarrow +\infty} v^\zeta(x) = x$. On the other hand, from Propositions (3.2), (3.3) and (3.4), we have

$$\mathbb{E} \left[\left(\int_0^{t^*(y)} e^{-\rho s} d(Z_s - \delta K_s) + e^{-\rho t^*(y)} v(X_{t^*(y)}, \lambda_{t^*(y)}) \right) \mathbf{1}_{\{T^\alpha > t^*(y)\}} \right] \leq e^{-\rho t^*(y)} \left(x + ct^*(y) + \frac{c}{\rho} \right).$$

Hence we can conclude, thanks to Proposition (3.2), that

$$x \leq \lim_{y \rightarrow +\infty} v(x, y) \leq \varepsilon + \lim_{y \rightarrow +\infty} v^{\ln(y)}(x) + e^{-\rho t^*(y)} \left(x + ct^*(y) + \frac{c}{\rho} \right) = \varepsilon + x$$

We obtain the result by letting ε going to 0. \square

We establish the local Lipschitz continuity of the value function, which is essential for applying comparison principles in the viscosity solution analysis.

Proposition 3.5 (Local Lipschitz continuity). *The value function v is locally Lipschitz continuous on $\mathbb{R} \times (b, +\infty)$. More precisely:*

i) *For $b \leq y$ and $0 \leq x < x'$ and $y > 0$, we have:*

$$x' - x \leq v(x', y) - v(x, y) \leq \delta(x' - x).$$

If $x < x' \leq 0$, we have

$$0 \leq v(x', y) - v(x, y) = \max(\delta x' + v(0, y), 0) - \max(\delta x + v(0, y), 0) \leq \delta(x' - x).$$

ii) Let $x \in \mathbb{R}^+$ and $b < y < y'$. For $\varepsilon > 0$ such that $b + \varepsilon \leq y$, we have:

$$0 \leq v(x, y) - v(x, y') \leq (x + \frac{c}{\rho}) \frac{a}{\varepsilon} (y' - y) + o(y' - y).$$

If $x \leq 0$, we have $v(x, y) - v(x, y') \leq \max(\delta x + v(0, y), 0) - \max(\delta x + v(0, y'), 0) \leq v(0, y) - v(0, y') \leq \frac{ac}{\rho\varepsilon} (y' - y) + o(y' - y)$.

Proof. We start by showing the first point:

- i) Let $y \geq b$ and $0 \leq x < x'$, we consider the strategy, in $\mathcal{A}(x, y)$ which consists in immediately inject some capital up to the cash level x' . It follows from the dynamic programming principle that:

$$v(x, y) \geq v(x', y) - \delta(x' - x).$$

Set $t_0 = (x' - x)/c$ and

$$\begin{aligned} v(x, y) &\geq \mathbb{E} \left[e^{-\rho(\tau_1 \wedge t_0)} v(X_{\tau_1 \wedge t_0}, \lambda_{\tau_1 \wedge t_0}) \right] \\ &\geq \mathbb{E} \left[e^{-\rho t_0} v(X_{t_0}, \lambda_{t_0}) \mathbf{1}_{\{t_0 \leq \tau_1\}} \right] \\ &\geq e^{-\rho t_0} v(x', \lambda_{t_0}) \mathbb{P}(t_0 \leq \tau_1) \\ &= e^{-\rho t_0} e^{-\int_0^{t_0} \tilde{\lambda}_s ds} v(x', y), \quad \text{with } \tilde{\lambda}_s = b - (b - y)e^{-as}. \end{aligned}$$

Hence, from Proposition 3.3, we have:

$$x' - x \leq v(x', y) - v(x, y) \leq \left(e^{\rho t_0} e^{\int_0^{t_0} \tilde{\lambda}_s ds} - 1 \right) v(x, y)$$

As $z \rightarrow ze^z - e^z + 1$ takes values in \mathbb{R}^+ and that $\int_0^{t_0} \tilde{\lambda}_s ds = bt_0 + \frac{y-b}{a} (1 - e^{-at_0})$, we get^c

$$\begin{aligned} v(x', y) - v(x, y) &\leq \left[(\rho + b)t_0 + \frac{y-b}{a} (1 - e^{-at_0}) \right] e^{(\rho+y)t_0} e^{\int_0^{t_0} \tilde{\lambda}_s ds} v(x, y) \\ &\leq (\rho + y)t_0 e^{(\rho+y)t_0} v(x, y) \\ &\leq \frac{\rho + y}{c} e^{\frac{\rho+y}{c}(x'-x)} (x + \frac{c}{\rho})(x' - x). \end{aligned}$$

We conclude that v is locally Lipschitz in x and that

$$x' - x \leq v(x', y) - v(x, y) \leq (x + \frac{c}{\rho}) \left((x' - x) \frac{\rho + y}{c} + o(x' - x) \right)$$

Hence v is Lipschitz in x .

- ii) We now consider $x \in \mathbb{R}^+$, $b < y < y'$ and $\varepsilon > 0$ such that $y \geq b + \varepsilon$. Let t_0 be such that $\tilde{\lambda}_{t_0}^{y'} = y$. From the definition of $\tilde{\lambda}$ (see (3.1)), we have that:

$$t_0 = -\frac{1}{a} \ln \left(\frac{y-b}{y'-b} \right) \leq \frac{y'-y}{a(y-b)} \leq \frac{1}{a\varepsilon} (y' - y).$$

Applying the strategy $(0, 0) \in \mathcal{A}(x, y')$, the dynamic programming principle implies that:

$$\begin{aligned} v(x, y') &\geq \mathbb{E} \left[e^{-\rho(\tau_1 \wedge t_0)} v(X_{\tau_1 \wedge t_0}, \lambda_{\tau_1 \wedge t_0}) \right] \\ &\geq e^{-\rho t_0} v(x + ct_0, y) \mathbb{P}(t_0 \leq \tau_1) \\ &\geq e^{-\rho t_0} e^{-\int_0^{t_0} \tilde{\lambda}_s ds} v(x + ct_0, y). \end{aligned}$$

As $v(x + ct_0, y) \geq v(x, y)$, we deduce from Proposition 3.4 that

$$0 \leq v(x, y) - v(x, y') \leq v(x, y') \left(e^{\rho t_0} e^{\int_0^{t_0} \bar{\lambda}_s ds} - 1 \right) \leq v(x, y') (\rho + y') t_0 e^{(\rho + y') t_0},$$

v is then locally lipschitz in its second variable and

$$0 \leq v(x, y) - v(x, y') \leq (x + \frac{c}{\rho}) \frac{a}{\varepsilon} (y' - y) e^{\frac{a}{\varepsilon} (\rho + y') (y' - y)}.$$

□

3.4 Capital injection strategies

From an economic perspective, capital injections represent a costly measure that should only be used to prevent imminent ruin. Injecting capital before it is strictly necessary is therefore suboptimal, as we formally establish below.

Proposition 3.6 (Capital injection policy). *Injecting capital is only optimal when strictly necessary to prevent ruin. In particular, capital injection at time t can only be optimal if the controlled surplus satisfies $X_t < 0$.*

Proof. For $y \geq b$ and $x \in \mathbb{R}$, two scenarios must be considered:

i) If $x \geq 0$, we claim that:

$$v(x + \varepsilon, y) - \varepsilon \delta < v(x, y)$$

Let $\kappa > 0$ and $\varepsilon > 0$ and α_ε a ε -suboptimal strategy, i.e. it is such that:

$$v(x + \kappa, y) \leq J_{\alpha_\varepsilon}(x + \kappa, y) + \varepsilon.$$

We let $\hat{\alpha}$ be the strategy consisting in applying α_ε while $s < T^{\alpha_\varepsilon}$ and increasing capital if $s = T^{\alpha_\varepsilon}$. Then we have:

$$\begin{aligned} v(x, y) &\geq J_{\hat{\alpha}}(x, y) \\ &= \mathbb{E} \left[\int_0^{T^{\alpha_\varepsilon}} e^{-\rho s} (dZ_s - \delta dK_s) - e^{-\rho T^{\alpha_\varepsilon}} \kappa \delta + e^{-\rho T^{\alpha_\varepsilon}} v(X_{T^{\alpha_\varepsilon}} + \kappa, Y_{T^{\alpha_\varepsilon}}) \right] \\ &= -\delta \kappa \mathbb{E} \left[e^{-\rho T^{\alpha_\varepsilon}} \right] + \mathbb{E} \left[\int_0^{T^{\alpha_\varepsilon}} e^{-\rho s} (dZ_s - \delta dK_s) + e^{-\rho T^{\alpha_\varepsilon}} v(X_{T^{\alpha_\varepsilon}} + \kappa, Y_{T^{\alpha_\varepsilon}}) \right] \\ &\geq -\delta \kappa \mathbb{E} \left[e^{-\rho T^{\alpha_\varepsilon}} \right] + v(x + \kappa, y) - \varepsilon. \end{aligned}$$

Finally, letting $\varepsilon \rightarrow 0$ and as $\mathbb{E} [e^{-\rho T^{\alpha_\varepsilon}}] < 1$ we conclude that:

$$v(x, y) > v(x + \kappa, y) - \kappa \delta.$$

ii) If $x < 0$, capital injection of at least $|x|$ is needed to avoid ruin, incurring a cost of $\delta|x|$. Then:

- Either $v(0, y) > 0$ and we inject at least capital $|x|$ if and only if $v(0, y) + \delta x > 0$,
- Or $v(0, y) = 0$ and we have $v(0, y) + \delta x < 0$ so we let the firm go bankrupt.

□

We now provide an explicit characterization of the value function for negative surplus values. The following result introduces a threshold that determines whether capital injection is optimal or if letting the firm go bankrupt is preferable.

Proposition 3.7 (Capital injection threshold). *Let $x < 0$ and $y \in [b, +\infty)$. We define $\kappa^*(y) = -\frac{v(0,y)}{\delta}$. Then, the value function satisfies:*

$$v(x, y) = \begin{cases} 0 & \text{if } x < \kappa^*(y), \\ v(0, y) + \delta x & \text{if } \kappa^*(y) < x < 0. \end{cases}$$

Proof. Let $(x, y) \in (-\infty, 0) \times [b, +\infty)$. By the dynamic programming principle (3.2) and Proposition 3.6, it is never optimal to inject more than $|x|$ units of capital. Let us define:

$$\kappa^*(y) = \inf\{z \in \mathbb{R}^-, v(z, y) > 0\}.$$

Then, for $x < 0$, we have that:

$$v(x, y) = \max(v(0, y) + \delta x, 0).$$

Which implies that capital is injected only if $v(0, y) + \delta x > 0$, or equivalently, if $x \geq -v(0, y)/\delta$. The result follows directly. \square

Thus, the capital injection threshold $\kappa^*(y)$ clearly delineates the boundary between solvency and bankruptcy, allowing us to precisely describe the insurer's optimal behaviour in situations of financial distress.

3.5 Hamilton–Jacobi–Bellman equation

In this section, we first state the HJB equation related to our control problem and then we show that the value function is the unique locally Lipschitz viscosity solution of the HJB equation. It will allow us to build a benchmark numerical method, based on the discretization of the variational inequality satisfied by the value function v in the next section.

We set $\mathcal{D}^+ := [0, +\infty) \times (b, +\infty)$ and $\mathcal{D}^- := (-\infty, 0) \times (b, +\infty)$. The HJB equation associated with our control problem is given by the following variational inequality:

$$\left\{ \min(\varphi \mathbf{1}_{\mathcal{D}^-}, (\partial_x \varphi - 1) \mathbf{1}_{\mathcal{D}^+}, \delta - \partial_x \varphi, -\mathcal{L}\varphi \mathbf{1}_{\mathcal{D}^+}) = 0 \quad \text{on } \mathbb{R} \times (b, +\infty) = \mathcal{D}^- \cup \mathcal{D}^+ \right. \quad (3.3)$$

where, \mathcal{L} denotes the infinitesimal generator of the controlled surplus process, defined by:

$$\mathcal{L}\varphi(x, y) := -(\rho + y)\varphi + c\partial_x \varphi + a(b - y)\partial_y \varphi + y \int_0^{+\infty} \varphi(x - z, y + \eta) dF(z),$$

and F denotes the cumulative distribution function of the claim sizes.

3.6 Viscosity solution characterization

The HJB equation stated in Equation (3.3) reflects the optimal trade-off between three control actions: paying dividends, injecting capital to prevent ruin, or passively allowing the surplus to evolve under the stochastic environment driven by the claim process. Due to the complexity introduced by the two-dimensional state space, classical solutions to the HJB equation are not expected to exist. For this reason, we adopt the framework of viscosity solutions. We now define the notion of viscosity solution used throughout the paper.

Definition 3.2 (Viscosity subsolution). *A function $\underline{u} : \mathbb{R} \times [b, +\infty) \rightarrow \mathbb{R}$ is said to be a viscosity subsolution of (3.3) at point $(x, y) \in \mathbb{R} \times (b, +\infty)$ if any continuously differentiable function $\varphi : \mathbb{R} \times [b, +\infty) \rightarrow \mathbb{R}$ with $\varphi(x, y) = \underline{u}(x, y)$ such that $\underline{u} - \varphi$ reaches a local maximum, 0, at (x, y) satisfies:*

$$\min(\varphi(x, y) \mathbf{1}_{(-\infty, 0)}(x), (\partial_x \varphi(x, y) - 1) \mathbf{1}_{\mathbb{R}^+}(x), \delta - \partial_x \varphi(x, y), -\mathcal{L}\varphi(x, y) \mathbf{1}_{\mathbb{R}^+}(x)) \leq 0.$$

Definition 3.3 (Viscosity supersolution). *A function $\bar{u} : \mathbb{R} \times [b, +\infty) \rightarrow \mathbb{R}$ is said to be a viscosity supersolution of (3.3) at point $(x, y) \in \mathbb{R} \times (b, +\infty)$ if any continuously differentiable function $\varphi : \mathbb{R} \times [b, +\infty) \rightarrow \mathbb{R}$ with $\varphi(x, y) = \bar{u}(x, y)$ such that $\bar{u} - \varphi$ reaches a local minimum, 0, at (x, y) satisfies:*

$$\min(\varphi(x, y) \mathbf{1}_{(-\infty, 0)}(x), (\partial_x \varphi(x, y) - 1) \mathbf{1}_{\mathbb{R}^+}(x), \delta - \partial_x \varphi(x, y), -\mathcal{L}\varphi(x, y) \mathbf{1}_{\mathbb{R}^+}(x)) \geq 0.$$

Definition 3.4 (Viscosity solution). A function $u : \mathbb{R} \times [b, +\infty) \rightarrow \mathbb{R}$ is said to be a viscosity solution of (3.3) if it is both a viscosity subsolution and a viscosity supersolution.

We now justify the viscosity characterization of the value function v by proving that it satisfies the HJB equation (3.3) both as a subsolution and as a supersolution in the viscosity sense.

Lemma 3.1 (Value function as viscosity supersolution). The value function v is a viscosity supersolution of (3.3) at every point $(x, y) \in \mathbb{R} \times (b, +\infty)$

Proof. Let $\varphi \in \mathcal{C}^1$ be a test function such that $v - \varphi$ has a local minimum at $(x, y) \in \mathbb{R} \times (b, +\infty)$ and $v(x, y) = \varphi(x, y)$. We verify the four conditions defining the viscosity supersolution are in force:

- i) **Bankruptcy constraint:** As $\varphi(x, y) = v(x, y) \geq 0$ we have $\varphi(x, y)\mathbf{1}_{\mathcal{D}^-}(x, y) \geq 0$.
- ii) **Dividend constraint:** Assume that $x > 0$. For any $\varepsilon > 0$ small enough, it follows from the possibility to immediately pay ε in dividends, that we have:

$$v(x, y) \geq v(x - \varepsilon, y) + \varepsilon \geq \varphi(x - \varepsilon, y) + \varepsilon.$$

As at point (x, y) we have $v(x, y) = \varphi(x, y)$ and $\varphi \in \mathcal{C}^1$ we deduce:

$$\frac{\partial \varphi}{\partial x}(x, y) - 1 \geq 0.$$

For $x = 0$, we know that $\varphi(-\varepsilon, y) \leq v(-\varepsilon, y) = \max(v(0, y) - \delta\varepsilon, 0)$. Moreover, $\varphi(0, y) = v(0, y) > 0$ therefore, for ε going to 0 we get $\partial_x \varphi(0, x) \geq \delta \geq 1$.

- iii) **Capital injection constraint:** Similarly, for any $\varepsilon > 0$, the possibility to inject capital at cost δ leads to:

$$v(x, y) \geq v(x + \varepsilon, y) - \delta\varepsilon.$$

This implies:

$$\delta - \frac{\partial \varphi}{\partial x}(x, y) \geq 0. \quad (3.4)$$

- iv) **Generator inequality:** The inequality is obvious for $x < 0$, so we shall assume that $x \geq 0$. We define the stopping time θ_h as:

$$\theta_h := \inf\{u \geq 0 : (x + cu, \tilde{\lambda}_u) \notin B(x, y)\} \wedge h.$$

And recall that τ_1 is the time of arrival of the first claim. Let $d\tilde{\lambda}_t = a(b - \lambda_t)dt$. Then we have:

$$\begin{aligned} v(x, y) &\geq \mathbb{E} \left[\int_0^{\tau_1 \wedge \theta_h} e^{-\rho s} dZ_s - \delta \int_0^{\tau_1 \wedge \theta_h} e^{-\rho s} dK_s + e^{-\rho(\tau_1 \wedge \theta_h)} \varphi(X_{\tau_1 \wedge \theta_h}, \lambda_{\tau_1 \wedge \theta_h}) \right] \\ &= \mathbb{E} \left[\int_0^{\tau_1 \wedge \theta_h} e^{-\rho s} dZ_s + e^{-\rho(\tau_1 \wedge \theta_h)} \varphi(X_{\tau_1 \wedge \theta_h}, \lambda_{\tau_1 \wedge \theta_h}) \right] \\ &= \mathbb{E} \left[\int_0^{\tau_1 \wedge \theta_h} e^{-\rho u} dZ_u + e^{-\rho \theta_h} \varphi(x + c\theta_h, \tilde{\lambda}_{\theta_h}) \mathbf{1}_{\{\theta_h < \tau_1\}} + e^{-\rho \tau_1} \varphi(x + c\tau_1 - Y_1, \tilde{\lambda}_{\tau_1} + \eta) \mathbf{1}_{\{\tau_1 \leq \theta_h\}} \right] \\ &= \mathbb{E} \left[\int_0^{\tau_1 \wedge \theta_h} e^{-\rho u} dZ_u + \left(\varphi(x, y) - \rho \theta_h \varphi(x, y) + c\theta_h \frac{\partial \varphi}{\partial x} + a(b - y)\theta_h \frac{\partial \varphi}{\partial y} + o(\theta_h) \right) \mathbf{1}_{\{\theta_h < \tau_1\}} \right] \\ &\quad + \int_0^{\theta_h} \left(\int_0^{+\infty} \varphi(x + cs - u, \tilde{\lambda}_s + \eta) dF(u) \right) e^{-\rho s} p_{\tau_1}(s) ds. \end{aligned}$$

Rearranging the terms and dividing by θ_h we obtain:

$$\begin{aligned} 0 \geq \mathbb{E} & \left[\int_0^{\tau_1 \wedge \theta_h} \frac{1}{\theta_h} e^{-\rho u} dZ_u \right] + \left(-\rho \varphi(x, y) + c \frac{\partial \varphi}{\partial x} + a(b-y) \frac{\partial \varphi}{\partial y} + o(\theta_h) \right) \mathbb{E} [\mathbf{1}_{\{\theta_h < \tau_1\}}] \\ & + \int_0^{\theta_h} \left(\int_0^{+\infty} \frac{1}{\theta_h} (\varphi(x+cs-u, \tilde{\lambda}_s + \eta) - v(x, y)) dF(u) \right) e^{-\rho s} p_{\tau_1}(s) ds \\ & + \frac{v(x, y)}{\theta_h} \int_0^{\theta_h} e^{-\rho s} p_{\tau_1}(s) ds. \end{aligned}$$

But we have that:

$$\frac{v(x, y)}{\theta_h} \int_0^{\theta_h} e^{-\rho s} p_{\tau_1}(s) ds \xrightarrow[h \rightarrow 0]{} yv(x, y).$$

Because:

$$p_{\tau_1}(s) = (b + (y-b)e^{-as}) e^{-bs - \frac{y-b}{a}(1-e^{as})} \xrightarrow[s \rightarrow 0]{} y.$$

Finally, letting $h \rightarrow 0$ we obtain:

$$0 \geq \mathbb{E}[Z_{0^+} - Z_0] - (\rho + y)\varphi(x, y) + c \frac{\partial \varphi}{\partial x} + a(b-y) \frac{\partial \varphi}{\partial y} + y \int_0^{+\infty} \varphi(x-u, y+\eta) dF(u).$$

Finally, choosing a strategy Z such that $\mathbb{E}[Z_{0^+} - Z_0] = 0$ we obtain:

$$(\rho + y)\varphi(x, y) - c \frac{\partial \varphi}{\partial x} - a(b-y) \frac{\partial \varphi}{\partial y} - y \int_0^{+\infty} \varphi(x-u, y+\eta) dF(u) \geq 0.$$

□

Lemma 3.2 (Value function as viscosity subsolution). *The value function v is a viscosity subsolution of (3.3) at every point $(x, y) \in \mathbb{R} \times (b, +\infty)^2$*

Proof. This proof is inspired by [2]. Arguing by contradiction that v is not a viscosity subsolution of (3.3) at point $(x, y) \in \mathbb{R} \times (b, +\infty)$. By definition this means that one can find $\nu > 0$ and $\varphi \in \mathcal{C}^1$ such that $\varphi(x, y) = v(x, y)$ and $\varphi(x', y') \geq v(x', y')$ for $(x', y') \in \mathbb{R} \times (b, +\infty)$:

$$\min (\varphi(x, y) \mathbf{1}_{(-\infty, 0)}(x), (\partial_x \varphi(x, y) - 1) \mathbf{1}_{(0, +\infty)}(x), \delta - \partial_x \varphi(x, y), -\mathcal{L}\varphi(x, y) \mathbf{1}_{[0, +\infty)}(x)) > \nu.$$

i) If $x < 0$, this implies that

$$\min (\varphi(x, y), \delta - \partial_x \varphi(x, y)) > \nu.$$

As $v(x, y) = \varphi(x, y) > 0$, the optimal policy is to inject capital and there exists $\varepsilon > 0$ such that $v(x, y) = v(x + \varepsilon) - \delta\varepsilon$. On the other hand, φ is continuously differentiable, so for $\varepsilon > 0$ small enough $\partial_x \varphi(x', y) \leq \delta - \frac{\nu}{2}$ for $x' \in (x, x + \varepsilon)$. Integrating the last inequality between x and $x + \varepsilon$, we get

$$\varepsilon(\delta - \frac{\nu}{2}) \geq \varphi(x + \varepsilon, y) - \varphi(x, y) \geq v(x + \varepsilon, y) - v(x, y) = \varepsilon\delta.$$

That leads to a contradiction.

ii) For $x = 0$, we have

$$\min (\partial_x \varphi(0, y) - 1, \delta - \partial_x \varphi(0, y), -\mathcal{L}\varphi(0, y)) > \nu.$$

Let $h > 0$ and $\varepsilon > 0$. From the dynamic programming principle, there exists $(Z^\varepsilon, K^\varepsilon) \in \mathcal{A}(0, y)$ such that:

$$\begin{aligned} v(0, y) & \leq \mathbb{E} \left[\int_0^{\tau_1 \wedge h} e^{-\rho s} dZ_s^\varepsilon - \delta \int_0^{\tau_1 \wedge h} e^{-\rho s} dK_s^\varepsilon + e^{-\rho \tau_1 \wedge h} v(X_{\tau_1 \wedge h}^\varepsilon, \lambda_{\tau_1 \wedge h}^y) \right] + \varepsilon h \\ & \leq \mathbb{E} \left[\int_0^{\tau_1 \wedge h} e^{-\rho s} d(Z_s^\varepsilon - \delta K_s^\varepsilon) + e^{-\rho \tau_1 \wedge h} \varphi(X_{\tau_1 \wedge h}^\varepsilon, \lambda_{\tau_1 \wedge h}^y) \right] + \varepsilon h \end{aligned}$$

It follows from Proposition 3.6 that $K_u^\varepsilon = 0$ on $\{u \leq \tau_1\}$. Hence, we have $Z_u^\varepsilon \leq cu$ on $0 \leq u \leq 0\tau_1$. It follows that there exists $\hat{c} \in [0, c]$ such that, on $\{h < \tau_1\}$:

$$\int_0^h e^{-\rho s} d(Z_s^\varepsilon - \delta K_s^\varepsilon) = \int_0^h e^{-\rho s} dZ_s^\varepsilon = (c - \hat{c})h + o(h) \text{ and } X_h^\varepsilon = \hat{c}h + o(h).$$

Hence, as $\varphi(0, y) = v(0, y)$, we have:

$$\begin{aligned} \varphi(0, y) &\leq \mathbb{E} \left[((c - \hat{c})h + o(h)) + e^{-\rho h} \varphi(\hat{c}h + o(h), \tilde{\lambda}_h^y) \mathbf{1}_{\{\tau_1 > h\}} \right] \\ &\quad + \mathbb{E} \left[\left(\int_0^{\tau_1} e^{-\rho s} dZ_s^\varepsilon + e^{-\rho \tau_1} \varphi(\tau_1 c - Y_1, \tilde{\lambda}_{\tau_1}^y + \eta) \right) \mathbf{1}_{\{\tau_1 \leq h\}} \right] + \varepsilon h, \end{aligned}$$

where $\tilde{\lambda}$ is solution of the following ODE: $d\tilde{\lambda}_s = a(b - \tilde{\lambda}_s)ds$. One can easily check that:

$$\tilde{\lambda}_s^y = (y - b)e^{-as} + b; \quad \text{for } s \geq 0.$$

For s going to 0, we get:

$$\tilde{\lambda}_s^y = y - as(y - b) + o(s).$$

Then we have:

$$\begin{aligned} \varphi(0, y) &\leq e^{-\rho h} \varphi(\hat{c}h + o(h), y - a(y - b)h + o(h)) \mathbb{P}(\tau_1 > h) + (c - \hat{c})h + o(h) \\ &\quad + \int_0^h \left(\int_0^{z^*(y)} e^{-\rho s} \varphi(cs - z, y + \eta - as(y - b) + o(s)) p(z) dz \right) \tilde{\lambda}_s e^{-\int_0^s \tilde{\lambda}_u du} ds + \varepsilon h \\ &= e^{-\rho h} \left(\varphi(0, y) + h \left[\hat{c} \frac{\partial \varphi}{\partial x}(0, y) + a(b - y) \frac{\partial \varphi}{\partial y}(0, y) \right] + o(h) \right) e^{-\int_0^h \tilde{\lambda}_s ds} + (c - \hat{c})h + \varepsilon h + o(h) \\ &\quad + hy \int_0^{z^*(y)} \varphi(-z, y) p(z) dz + o(h) \end{aligned}$$

For h and then ε going to 0, we obtain that:

$$\varphi(0, y) \leq \frac{1}{\rho + y} \left[c + \hat{c} \left(\frac{\partial \varphi}{\partial x}(0, y) - 1 \right) + a(b - y) \frac{\partial \varphi}{\partial y}(0, y) + y \int_0^{z^*(y)} \varphi(-z, y) p(z) dz \right].$$

As we have $\frac{\partial \varphi}{\partial x}(0, y) - 1 \geq 0$, we get a contradiction between $\mathcal{L}\varphi(0, y) < -\nu$ and

$$\varphi(0, y) \leq \frac{1}{\rho + y} \left[c \frac{\partial \varphi}{\partial x}(0, y) + a(b - y) \frac{\partial \varphi}{\partial y}(0, y) + y \int_0^{z^*(y)} \varphi(-z, y) p(z) dz \right]. \quad (3.5)$$

iii) Assume that $x > 0$ and set $B_r(x, y) \subset (0, +\infty) \times (b, +\infty)$ be a closed ball of radius $r > 0$. We define:

$$\tau_B = \inf\{t > 0 | X_t^\alpha \notin B_r(x, y)\}.$$

We denote by $\tau^* = \tau_B \wedge T$.

Case 1: On $\{\tau^* = \tau_B\}$ two cases are possible:

– There was no jump and:

$$\begin{cases} X_{\tau^*-}^\alpha = X_{\tau^*}^\alpha = x + r \Rightarrow y - r \leq \lambda_{\tau^*-} = \lambda_{\tau^*} \leq x + r, \\ \lambda_{\tau^*-} = \lambda_{\tau^*} = y - r \Rightarrow x - r \leq X_{\tau^*}^\alpha \leq x + r. \end{cases}$$

– There has been a jump and:

$$\begin{cases} X_{\tau^*-}^\alpha \geq X_{\tau^*}^\alpha \text{ and } X_{\tau^*}^\alpha \leq x - r \text{ and } \lambda_{\tau^*-} \leq \lambda_{\tau^*}, \\ \lambda_{\tau^*-} \leq \lambda_{\tau^*} \text{ and } \lambda_{\tau^*} \geq y + r \text{ and } X_{\tau^*}^\alpha \leq x + r. \end{cases}$$

Taken together, these elements give us $X_{\tau^*}^\alpha \leq x \pm r := x'$ and $\lambda_{\tau^*-} \leq \lambda_{\tau^*}$ and as v is increasing in x and decreasing in y we have:

$$v(X_{\tau^*}^\alpha, \lambda_{\tau^*}) \leq v(x', \lambda_{\tau^*}) \leq \varphi(x', \lambda_{\tau^*}) \leq \varphi(X_{\tau^*-}^\alpha, \lambda_{\tau^*-}).$$

Case 2: On $\{\tau^* = T\}$ we have $X_{\tau^*-}^\alpha \geq X_{\tau^*}^\alpha$ and $\lambda_{\tau^*-} \leq \lambda_{\tau^*}$, then we can write:

$$v(X_{\tau^*}^\alpha, \lambda_{\tau^*}) \leq \varphi(X_{\tau^*-}^\alpha, \lambda_{\tau^*-}).$$

Then, for both cases one can write:

$$e^{-\rho\tau^*} v(X_{\tau^*}^\alpha, \lambda_{\tau^*}) \leq e^{-\rho\tau^*^-} \varphi(X_{\tau^*-}^\alpha, \lambda_{\tau^*-}).$$

Recall that:

$$\begin{cases} X_{\tau^*-}^\alpha = x + c\tau^{*-} - \sum_{k=1}^{N_{\tau^*-}} Y_k - Z_{\tau^*-} + K_{\tau^*-}, \\ d\lambda_{\tau^*-} = a(b - \lambda_{\tau^*-})d\tau^{*-} + \eta dN_{\tau^*-}. \end{cases}$$

By applying Itô's formula on $e^{-\rho\tau^*^-} \varphi(X_{\tau^*-}^\alpha, \lambda_{\tau^*-})$ we obtain:

$$\begin{aligned} e^{-\rho\tau^*^-} \varphi(X_{\tau^*-}^\alpha, \lambda_{\tau^*-}) - \varphi(x, y) &= \int_0^{\tau^*^-} e^{-\rho s} \frac{\partial \varphi}{\partial x}(X_{s-}^\alpha, \lambda_{s-}^\alpha) [cds - dZ_s + dK_s] \\ &\quad + \int_0^{\tau^*^-} e^{-\rho s} \frac{\partial \varphi}{\partial y}(X_{s-}^\alpha, \lambda_{s-}^\alpha) [a(b - y)ds] - \rho \int_0^{\tau^*^-} e^{-\rho s} \varphi(X_{s-}^\alpha, \lambda_{s-}^\alpha) ds \\ &\quad + \sum_{\substack{0 \leq s \leq \tau^*^- \\ X_s^\alpha \neq X_{s-}^\alpha}} (\varphi(X_s^\alpha, \lambda_s^\alpha) - \varphi(X_{s-}^\alpha, \lambda_{s-}^\alpha)) e^{-\rho s} \\ &\quad + \sum_{\substack{0 \leq s \leq \tau^*^- \\ X_{s+}^\alpha \neq X_s^\alpha}} (\varphi(X_{s+}^\alpha, \lambda_{s+}^\alpha) - \varphi(X_s^\alpha, \lambda_s^\alpha)) e^{-\rho s}. \end{aligned}$$

By construction before τ^* we are in $B_r(x)$, so this can not be optimal to inject capital, which leads to $K = 0$ before τ^* .

$$\begin{aligned} e^{-\rho\tau^*^-} \varphi(X_{\tau^*-}^\alpha, \lambda_{\tau^*-}) - \varphi(x, y) &= \int_0^{\tau^*^-} e^{-\rho s} \left[\frac{\partial \varphi}{\partial x}(X_{s-}^\alpha, \lambda_{s-}^\alpha) c + \frac{\partial \varphi}{\partial y}(X_{s-}^\alpha, \lambda_{s-}^\alpha) a(b - y) \right] ds \\ &\quad - \int_0^{\tau^*^-} e^{-\rho s} \frac{\partial \varphi}{\partial x}(X_{s-}^\alpha, \lambda_{s-}^\alpha) dZ_s - \rho \int_0^{\tau^*^-} e^{-\rho s} \varphi(X_{s-}^\alpha, \lambda_{s-}^\alpha) ds \\ &\quad + \sum_{\substack{0 \leq s \leq \tau^*^- \\ X_s^\alpha \neq X_{s-}^\alpha}} (\varphi(X_s^\alpha, \lambda_s^\alpha) - \varphi(X_{s-}^\alpha, \lambda_{s-}^\alpha)) e^{-\rho s} \\ &\quad + \sum_{\substack{0 \leq s \leq \tau^*^- \\ X_{s+}^\alpha \neq X_s^\alpha}} (\varphi(X_{s+}^\alpha, \lambda_{s+}^\alpha) - \varphi(X_s^\alpha, \lambda_s^\alpha)) e^{-\rho s}. \end{aligned}$$

$- X_{s+}^\alpha - X_s^\alpha \neq 0$ corresponds to the case where dividends has been distributed. So, we have:

$$X_{s+}^\alpha - X_s^\alpha = -(Z_{s+} - Z_s),$$

hence:

$$\begin{aligned} \sum_{\substack{0 \leq s \leq \tau^* \\ X_{s+}^\alpha \neq X_s^\alpha}} (\varphi(X_{s+}^\alpha, \lambda_{s+}) - \varphi(X_s^\alpha, \lambda_s)) e^{-\rho s} &\leq \sum_{\substack{0 \leq s \leq \tau^* \\ X_{s+}^\alpha \neq X_s^\alpha}} (\varphi(X_{s+}^\alpha, \lambda_s) - \varphi(X_s^\alpha, \lambda_s)) e^{-\rho s} \\ &= - \sum_{\substack{0 \leq s \leq \tau^* \\ Z_{s+} \neq Z_s^\alpha}} \left(\int_0^{Z_{s+} - Z_s} \frac{\partial \phi}{\partial x}(X_s^\alpha - u, \lambda_s) du \right). \end{aligned}$$

Using the fact that $\partial_x \varphi - 1 > \nu$ i.e. $\partial_x \varphi > 1$ we obtain:

$$\begin{aligned} &- \left(\int_0^{\tau^*} e^{-\rho s} \frac{\partial \varphi}{\partial x}(X_{s-}^\alpha, \lambda_{s-}^\alpha) dZ_s + \sum_{\substack{0 \leq s \leq \tau^* \\ Z_{s+} \neq Z_s^\alpha}} e^{-\rho s} \left(\int_0^{Z_{s+} - Z_s} \frac{\partial \phi}{\partial x}(X_s^\alpha - u, \lambda_s) du \right) \right) \leq \\ &- \left(\int_0^{\tau^*} e^{-\rho s} dZ_s + \sum_{\substack{0 \leq s \leq \tau^* \\ Z_{s+} \neq Z_s^\alpha}} e^{-\rho s} (Z_{s+} - Z_s) \right) = - \int_0^{\tau^*} e^{-\rho s} dZ_s. \end{aligned}$$

– $X_s^\alpha - X_{s-}^\alpha$ corresponds to the case where there has been a jump in the cash process (claims has arrived). As pointed out in [21] and used in [2], the following process is a martingale:

$$\begin{aligned} &\sum_{\substack{0 \leq s \leq \tau^* \\ X_s^\alpha \neq X_{s-}^\alpha}} (\varphi(X_s^\alpha, \lambda_s^\alpha) - \varphi(X_{s-}^\alpha, \lambda_{s-}^\alpha)) e^{-\rho s} \\ &- y \int_0^{\tau^*} \left(\int_0^{+\infty} (\varphi(X_{s-}^\alpha - u, \lambda_{s-} + \eta) - \varphi(X_{s-}^\alpha, \lambda_{s-})) dF(u) \right) e^{-\rho s} ds. \end{aligned}$$

We obtain:

$$\begin{aligned} &e^{-\rho \tau^*} \varphi(X_{\tau^*}^\alpha, \lambda_{\tau^*}) + \int_0^{\tau^*} e^{-\rho s} dZ_s \leq \varphi(x, y) \\ &+ \int_0^{\tau^*} e^{-\rho s} \left[c \frac{\partial \varphi}{\partial x}(X_{s-}^\alpha, \lambda_{s-}^\alpha) + a(b - y) \frac{\partial \varphi}{\partial y}(X_{s-}^\alpha, \lambda_{s-}^\alpha) - \rho \varphi(X_{s-}^\alpha, \lambda_{s-}^\alpha) \right. \\ &\left. + y \int_0^{+\infty} (\varphi(X_{s-}^\alpha - u, \lambda_{s-} + \eta) - \varphi(X_{s-}^\alpha, \lambda_{s-})) dF(u) \right] ds. \end{aligned}$$

Finally:

$$\begin{aligned} \varphi(x, y) = v(x, y) &\leq \mathbb{E} \left[\int_0^{\tau^*} e^{-\rho s} (dZ_s - \delta dK_s) + e^{\tau^*} v(X_{\tau^*}^\alpha, \lambda_{\tau^*}) \right] + \varepsilon \\ &\leq \mathbb{E} \left[\int_0^{\tau^*} e^{-\rho s} dZ_s + e^{\tau^*} v(X_{\tau^*}^\alpha, \lambda_{\tau^*}) \right] + \varepsilon \\ &\leq \varphi(x, y) + \mathbb{E} \left[\int_0^{\tau^*} e^{-\rho s} \left[c \frac{\partial \varphi}{\partial x}(X_{s-}^\alpha, \lambda_{s-}^\alpha) + a(b - y) \frac{\partial \varphi}{\partial y}(X_{s-}^\alpha, \lambda_{s-}^\alpha) \right. \right. \\ &\quad \left. \left. - \rho \varphi(X_{s-}^\alpha, \lambda_{s-}^\alpha) + y \int_0^{+\infty} (\varphi(X_{s-}^\alpha - u, \lambda_{s-} + \eta) - \varphi(X_{s-}^\alpha, \lambda_{s-})) dF(u) \right] ds \right] + \varepsilon. \end{aligned}$$

This implies the following contradiction:

$$0 \leq -\mathbb{E} \left[\int_0^{\tau^*^-} e^{-\rho s} \mathcal{L}\varphi(X_{s^-}, \lambda_{s^-}) ds \right] \leq -\nu \mathbb{E} \left[\int_0^{\tau^*^-} e^{-\rho s} ds \right] < 0.$$

□

A direct consequence of Lemmas 3.2 and 3.1 is the following result.

Theorem 3.1 (Value function as viscosity solution). *The value function v is a viscosity solution of the Hamilton–Jacobi–Bellman equation (3.3) on the domain $\mathbb{R} \times (b, +\infty)$.*

Remark 3.2. Notice that the variational inequality at points $(0, y)$ could be considered as a boundary condition because it could be written as

$$(\rho + y)\varphi(0, y) = c \frac{\partial \varphi}{\partial x}(0, y) + a(b - y) \frac{\partial \varphi}{\partial y}(0, y) + y \int_0^{\varphi(0, y+\eta)/\delta} (\varphi(0, y+\eta) - \delta z) dF(z) \quad (3.6)$$

Following the proof of Proposition 4.2 in [1], we can now give a characterization of the value function as the smallest viscosity supersolution of equation (3.3).

Theorem 3.2. *v is the smallest viscosity supersolution of (3.3) that is non-increasing in y , locally Lipschitz continuous and satisfies the growth condition established in Proposition 3.2.*

4 Finite-difference estimate

In this section, we present the classical finite-difference scheme used as a numerical benchmark for the solution of the HJB variational inequality (3.3). The method relies on a monotone discretization of the state dynamics combined with Howard’s policy iteration algorithm to obtain the stationary solution. This framework provides a consistent and interpretable reference against which the reinforcement learning approach introduced later can be compared.

4.1 Discrete HJB variational inequality

Computational grid and domain truncation

To approximate the value function numerically, we truncate the state space and construct a uniform grid over the resulting bounded domain. Let $X_{\min} < 0 < X_{\max}$ and $Y_{\max} \in (b, +\infty)$, and define $\mathcal{D} := [X_{\min}, X_{\max}] \times [b, Y_{\max}]$ as the computational domain for the surplus and intensity variables. The domain \mathcal{D} is discretized using $N_x \in \mathbb{N}$ and $N_y \in \mathbb{N}$ spatial subdivisions along the x and y directions, respectively, leading to the mesh sizes

$$\Delta x := \frac{X_{\max} - X_{\min}}{N_x}, \quad \Delta y := \frac{Y_{\max} - b}{N_y}.$$

The corresponding grid points are defined by

$$x_i := X_{\min} + i\Delta x, \quad y_j := b + j\Delta y.$$

The full grid is therefore

$$\mathcal{G} := \{(x_i, y_j) : 0 \leq i \leq N_x, 0 \leq j \leq N_y\},$$

with its interior nodes denoted by

$$\mathcal{G}^\circ := \{(x_i, y_j) \in \mathcal{G} : 1 \leq i \leq N_x - 1, 1 \leq j \leq N_y - 1\}.$$

At each grid point $(x_i, y_j) \in \mathcal{G}$, the numerical approximation of the value function is denoted by $V_{i,j} \approx v(x_i, y_j)$ and will be used consistently throughout the discrete formulation. The index corresponding to the origin $x = 0$ is denoted by i_0 , so that $x_{i_0} = 0$.

Discretization of differential operators

We approximate the differential operators in the HJB variational inequality by means of a monotone finite-difference discretization. One-sided differences are employed in each direction to preserve the directionality of the underlying drift terms. For x and y coordinates, the discrete first-order operators are defined as

$$D_x^- V_{i,j} := \frac{V_{i,j} - V_{i-1,j}}{\Delta x}, \quad D_x^+ V_{i,j} := \frac{V_{i+1,j} - V_{i,j}}{\Delta x},$$

and similarly

$$D_y^- V_{i,j} := \frac{V_{i,j} - V_{i,j-1}}{\Delta y}, \quad D_y^+ V_{i,j} := \frac{V_{i,j+1} - V_{i,j}}{\Delta y}.$$

For a generic convective term $s\partial_\xi V$ with $\xi \in \{x, y\}$, an upwind discretization is adopted:

$$s\partial_\xi V \approx \begin{cases} sD_\xi^- V, & \text{if } s \geq 0, \\ sD_\xi^+ V, & \text{if } s < 0. \end{cases}$$

In particular, since $c > 0$, we approximate $-c\partial_x V$ by $cD_x^+ V$. For the intensity dynamics, the drift satisfies $a(b - y) \leq 0$ on $[b, Y_{\max}]$, yielding

$$-a(b - y_j)\partial_y V_{i,j} \approx -a(b - y_j)D_y^- V_{i,j}.$$

Approximation of the jump integral

The discrete infinitesimal generator \mathcal{L}_h acting on the grid interior \mathcal{G}° is then defined by

$$-\mathcal{L}_h V_{i,j} := (\rho + y_j)V_{i,j} - cD_x^+ V_{i,j} - a(b - y_j)D_y^- V_{i,j} - y_j \mathcal{Q}_h[V]_{i,j}, \quad (4.1)$$

where \mathcal{Q}_h denotes the discrete approximation of the jump operator

$$\mathcal{Q}[V](x, y) := \int_0^{+\infty} V(x - z, y + \eta) dF(z).$$

To approximate this integral, we truncate the support of f at $Z_{\max} = (M + \frac{1}{2})\Delta x$ and apply a midpoint quadrature rule on $[0, Z_{\max}]$:

$$\int_0^{Z_{\max}} V(x - z, y + \eta) f(z) dz \approx \sum_{m=0}^M V(x - (m + \frac{1}{2})\Delta x, y + \eta) f((m + \frac{1}{2})\Delta x) \Delta x,$$

When $x_i - (m + \frac{1}{2})\Delta x < 0$, the capital injection condition given in Proposition 3.7 is enforced to evaluate V , while off-grid values are obtained by bilinear interpolation.

Discrete HJB variational inequality

Combining the spatial and integral approximations introduced above, the discrete counterpart of the HJB variational inequality takes the form

$$\min \left(\underbrace{D_x^- V_{i,j} - 1}_{\text{dividends}}, \underbrace{\delta - D_x^+ V_{i,j}}_{\text{capital injection}}, \underbrace{-\mathcal{L}_h V_{i,j}}_{\text{continuation}} \right) = 0, \quad (i, j) \in \mathcal{G}^\circ, \quad (4.2)$$

The resulting non-linear system is monotone and consistent with the viscosity framework, providing a robust basis for numerical resolution.

4.2 Numerical implementation

Local update rules

The discrete variational inequality (4.2) defines, at each grid node, the local optimality condition between the three possible regimes: dividend payment, capital injection, and continuation. In practice, this translates into a set of region-specific update formulas that can be used to iteratively compute the value function over the grid. The expressions below follow directly from the monotone discretization introduced in the previous subsection.

In the dividend region, the optimal action corresponds to an immediate payout, leading to the first-order condition $D_x^- V_{i,j} = 1$, which yields

$$V_{i,j} = V_{i-1,j} + \Delta x.$$

In the continuation region, the process evolves according to the controlled surplus dynamics without intervention, and the value function satisfies the discrete HJB equation obtained from (4.1):

$$V_{i,j} = \left(\rho + y_j + \frac{c}{\Delta x} + \frac{a(y_j - b)}{\Delta y} \right)^{-1} \left(\frac{c}{\Delta x} V_{i+1,j} + \frac{a(y_j - b)}{\Delta y} V_{i,j-1} + y_j Q_h[V]_{i,j} \right).$$

The capital injection region requires a specific treatment, as its behaviour is entirely characterized by Proposition 3.7. According to this result, the value function is known explicitly for $x < 0$, where injections occur whenever the surplus lies below the optimal boundary. Hence, the relation

$$V_{i,j} = \max(0, V_{i_0,j} + \delta x_i),$$

is imposed directly as a boundary condition for all grid points with $x_i < 0$, ensuring consistency with the theoretical characterization of the optimal policy.

Boundary conditions

In the negative surplus region $x < 0$, the value function is entirely determined by the theoretical characterization established in Proposition 3.7, which directly governs the capital injection mechanism. Hence, no numerical update is required in this area, and the boundary relation at $x = 0^+$ serves as the effective entry condition for the computational domain. According to the HJB equation 3.3, for $j \in \{0, \dots, N_y\}$ the value function at $x = 0$ satisfies

$$(\rho + y_j)V_{i_0,j} = cD_x^+ V_{i_0,j} + a(b - y_j)D_y^- V_{i_0,j} + y_j I_h(V_{i_0,j}^{(\eta)}/\delta),$$

where we recall that i_0 denotes the index corresponding to $x_{i_0} = 0$. Here, the term $V_{i_0,j}^{(\eta)}$ represents the numerical approximation of $v(0, y_j + \eta)$ obtained by linear interpolation, while $I_h(\cdot)$ denotes the approximation of the integral term arising from the infinitesimal generator at $x = 0$, using the injection characterization given in Proposition 3.7. In practice, I_h can be evaluated using a midpoint or trapezoidal rule depending on the discretization of F , although for many standard claim size distributions, this integral admits a closed-form expression, allowing for an exact and computationally efficient evaluation.

At the upper boundary of the intensity domain, the asymptotic behaviour derived in Corollary 3.1 implies $\lim_{y \rightarrow \infty} v(x, y) = x$ for all $x \in \mathbb{R}^+$, which translates numerically into

$$V_{i,N_y} = x_i,$$

for all $i \geq i_0$. Together, these two boundary conditions fully close the discrete problem and ensure the well-posedness of the numerical resolution of the value function.

Howard policy iteration

The non-linear discrete system (4.2) is solved using Howard's policy iteration algorithm. The method alternates between a policy evaluation step, where the value function is computed for a fixed control configuration, and a policy improvement step, where the control is updated pointwise according to the minimization operator in (4.2). Starting from an initial value function $V^{(0)}$ and an initial policy $\pi^{(0)}$, the iteration proceeds as follows:

(i) **Policy evaluation:** For a fixed policy $\pi^{(k)}$, the corresponding value function $V^{(k+1)}$ is obtained by solving the discrete HJB system (4.1) induced by this policy. This consists in applying, at each grid point, the update rule associated with the prescribed regime. The system is solved by fixed-point iteration until convergence, under the boundary conditions described above.

(ii) **Policy improvement:** The policy is then updated pointwise by selecting the locally optimal regime,

$$\pi^{(k+1)}(i, j) = \arg \min_{\pi \in \{\text{dividend, injection, continuation}\}} \mathcal{H}_h^\pi[V^{(k+1)}]_{i,j},$$

where \mathcal{H}_h^π denotes the local discrete HJB operator associated with regime π .

The algorithm iterates between these two steps until the policy stabilizes, that is, when $\pi^{(k+1)} = \pi^{(k)}$ over the grid, indicating convergence to the stationary optimal control. The convergence of the numerical scheme follows from standard arguments for monotone finite-difference approximations and policy iteration methods. Under the usual monotonicity, consistency, and stability assumptions on the discrete operator \mathcal{L}_h , the fixed-point evaluation step preserves the viscosity solution framework of the continuous HJB variational inequality [6]. Moreover, the outer Howard iteration, alternating between policy evaluation and improvement, converges to the unique stationary solution of the discrete control problem under these same structural conditions [28, 22]. Overall, the scheme is guaranteed to converge to the discrete viscosity solution, which consistently approximates the continuous value function as the mesh is refined.

4.3 Numerical results and sensitivity analysis

4.3.1 Reference configuration and qualitative analysis

Model and grid setup

We begin with a balanced baseline configuration of parameters, chosen to represent a typical regime where claim arrivals, excitation effects, and premium inflows are of comparable magnitude. In particular, claim sizes are assumed to follow an exponential distribution with parameter β . This setup serves as a reference for the numerical results presented below and will later be used to assess the sensitivity of the optimal policy to individual model parameters. The corresponding values are reported in Table 1, while the discretization settings are summarized in Table 2.

| a | b | η | ρ | c | δ | β |
|-----|-----|--------|--------|-----|----------|---------|
| 2.0 | 2.0 | 0.4 | 0.1 | 1.0 | 1.8 | 3.0 |

Table 1: Baseline configuration of model parameters.

Instead of fixing N_x and N_y directly, we define the grid resolution through the auxiliary parameters M and n_η , which determine the number of discretization steps relative to Z_{\max} and η . This construction ensures that Δy is an exact multiple of η and Δx an exact multiple of Z_{\max} , thereby avoiding interpolation errors when evaluating the jump and excitation terms. The origin $x = 0$ is explicitly enforced to belong to the grid, with minor adjustments of the bounds if necessary.

| X_{\min} | X_{\max} | Y_{\max} | n_η | M | Z_{\max} |
|------------|------------|------------|----------|-----|------------|
| -5.0 | 4.0 | 25.0 | 8 | 80 | 5.0 |

Table 2: Grid parameters used for the numerical discretization.

Value function and associated optimal policy

Figure 1a displays the estimated value function obtained from the finite-difference scheme. The surface exhibits the expected qualitative behaviour: the value increases with the surplus x and decreases with the claim intensity y , reaching its highest levels for large surpluses and low intensities. These numerical patterns

are fully consistent with the theoretical monotonicity properties established in Propositions 3.3 and 3.4, confirming the accuracy and stability of the discretization procedure. Figure 1b shows the optimal control policy under the baseline configuration. The solution exhibits a threshold structure, with two distinct regions for $x < 0$ (a ruin region and a capital-injection region) and two regions for $x \geq 0$ (a continuation region and a dividend region), yielding a clear and interpretable partition of the state space.

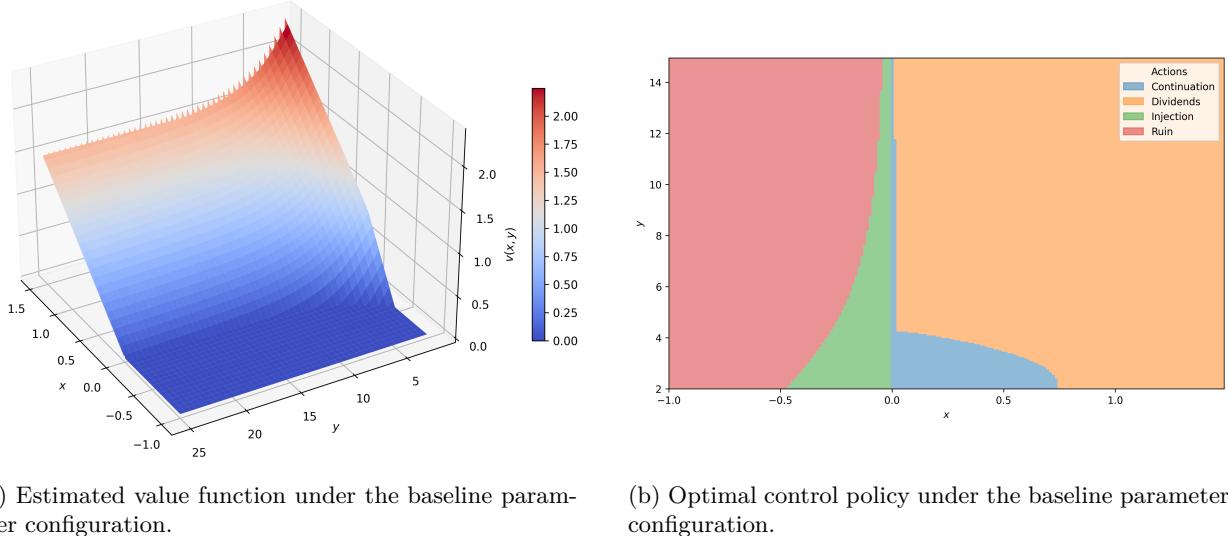


Figure 1: Estimated value function and corresponding optimal control policy under the baseline parameter configuration.

In the positive surplus region, the policy exhibits the expected two-zone structure: a continuation region and a dividend region. For sufficiently large surplus levels, it is always optimal to distribute dividends. This behaviour reflects the fact that the insurer holds enough reserves to absorb potential losses, making the immediate distribution of excess capital preferable. By contrast, the continuation region corresponds to states in which the activity remains exposed to significant risk. In this zone, it is optimal to retain earnings until the reserve reaches a safer level, at which point dividend payments resume. A critical feature emerging from the numerical solution is the existence of an intensity threshold y above which the optimal action is to liquidate the surplus down to $x = 0^+$. In this high-intensity regime, there exists an increased and persistent risk of claim occurrences, leading to a high likelihood of large loss clusters and little chance that the intensity will decline rapidly enough to restore profitability. Operating under such conditions is no longer profitable, and the optimal strategy is to distribute all available capital before the firm is driven to ruin.

In the negative surplus region, the numerical policy reproduces the expected qualitative behaviour, featuring a clear capital-injection region and a ruin region. The boundary separating these two zones coincides with the one derived in Proposition 3.7. Since $\kappa^*(y) = -v(0, y)/\delta$ and Corollary 3.1 establishes that $v(0, y) \rightarrow 0$ as $y \rightarrow \infty$, the convergence of the injection boundary toward 0 for large intensities is fully consistent with the theoretical predictions. Beyond this boundary, capital injection is no longer optimal. When incoming claims push the cash reserves past this threshold, the activity becomes too costly to refinance. Injecting capital up to $x = 0$ would not generate future earnings sufficient to offset the cost of the refinancing itself. In such circumstances, further investment is economically dominated, and the optimal decision is to let ruin occur.

4.3.2 Sensitivity of the optimal policy

We now examine how the optimal control policy reacts to changes in the model parameters. Each parameter is varied independently around its baseline value while keeping the others fixed. The resulting policy maps illustrate how the intervention thresholds adapt to the underlying economic and risk conditions. Overall, the numerical outcomes remain consistent with theoretical expectations and economic intuition.

Impact of Hawkes dynamics parameters

The parameters (a, b, η) govern the temporal behaviour of the claim intensity process. An increase in the mean-reversion rate a accelerates the return of λ_t to its baseline level b , reducing the persistence of high-intensity episodes. This results in wider continuation and injection regions, as the system spends less time in high-risk states. In contrast, a higher excitation parameter η amplifies clustering effects, making the environment significantly more risky. When the self-excitation of future claim arrivals makes the business unprofitable, the optimal strategy shifts toward full liquidation: distributing all available surplus rather than continuing operations.

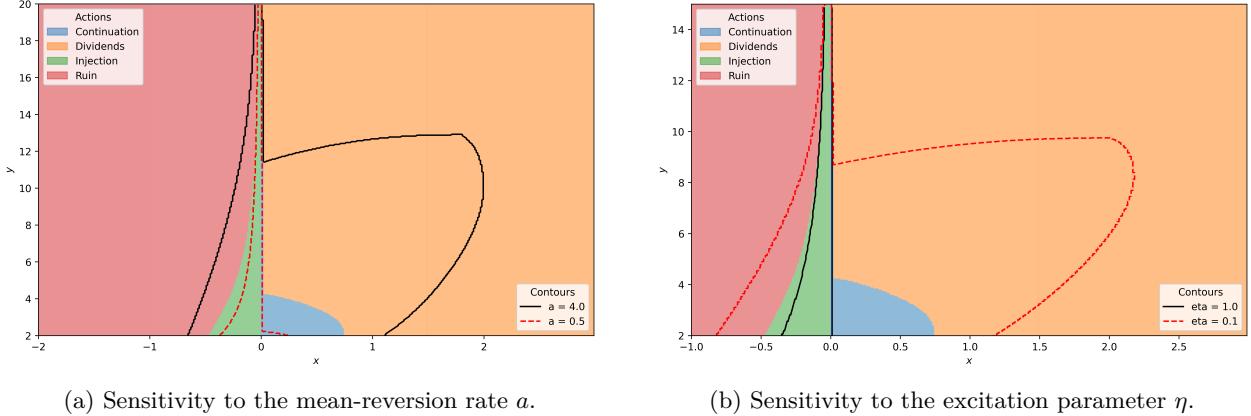


Figure 2: Sensitivity of optimal policy to Hawkes dynamics parameters.

A closer inspection of the continuation region also reveals a distinctive shape that depends sensitively on the model parameters. For a fixed but sufficiently high intensity level, the optimal policy in the positive surplus region may switch from dividend distribution to continuation and then back to dividend distribution as x increases. This non-monotone pattern appears for specific parameter configurations, such as $a = 4$ and $\eta = 0.1$, but also emerges under other parameter variations in the subsequent sensitivity analyses.

The initial dividend region observed at low surplus levels reflects situations where the intensity has risen too sharply for profitability to be restored. Such states necessarily arise from a sequence of adverse claims originating in the continuation region, which simultaneously depletes the surplus and drives the intensity upward. Under these conditions, continued operation is no longer viable, and the optimal action is to liquidate the available surplus immediately. For the same intensity level, a slightly higher surplus would allow the insurer to absorb potential short-term losses while waiting for the intensity to revert, making continuation preferable. As the surplus becomes large, the policy reverts to its usual behaviour: the company holds enough reserves to withstand adverse shocks, and distributing dividends again becomes optimal. This layered structure of the continuation region thus captures a subtle interplay between short-term risk exposure and long-term mean reversion in the intensity dynamics, and aligns with the economic interpretation of the Hawkes-driven claim environment.

Impact of the premium–claim balance

The premium rate c determines the rate of surplus accumulation, directly affecting the insurer's capacity to sustain operations. Higher values of c expand the continuation region and postpone both injections and dividend payments. In contrast, the claim size parameter β affects the expected cost of claims, with larger β (smaller expected losses) leading to higher profitability and a broader dividend region.

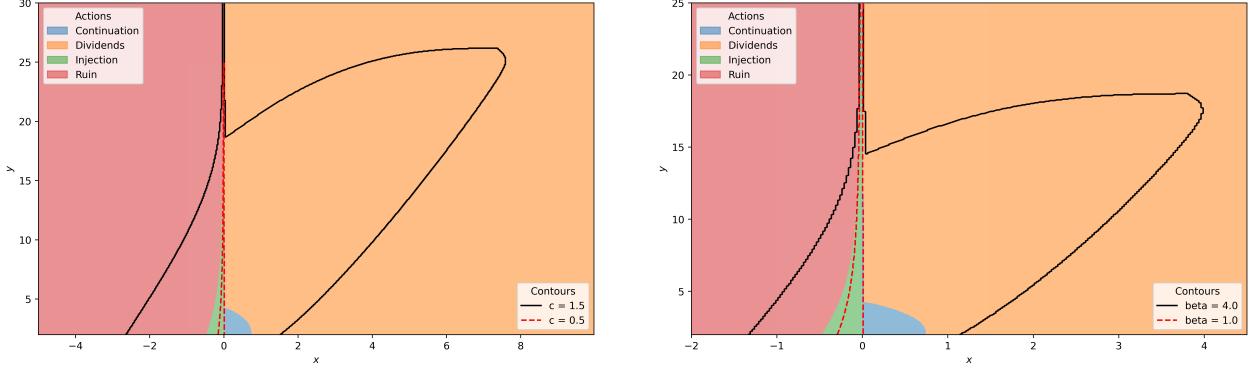


Figure 3: Sensitivity of the optimal policy to insurance parameters.

Impact of financing and valuation parameters

The discount rate ρ and the capital injection cost δ capture financial and valuation effects. A higher discount rate reduces the present value of future profits, leading the insurer to liquidate earlier rather than maintaining operations with limited expected returns. This translates into a contraction of the continuation region and an expansion of the dividend area. Conversely, increasing the injection cost δ discourages recapitalization and makes the firm more reluctant to support temporary losses, thereby enlarging the liquidation region and shrinking the domain where capital injections are optimal.

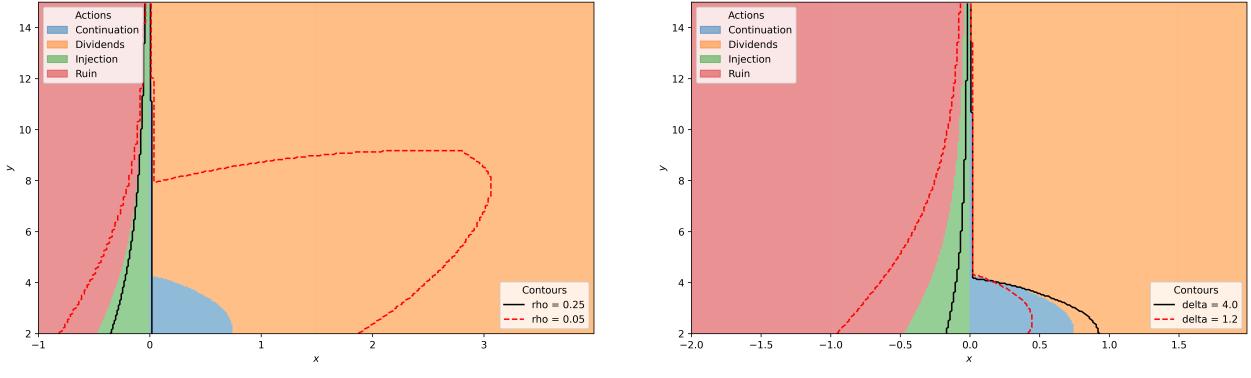


Figure 4: Sensitivity of the optimal policy to financing and valuation parameters.

5 Reinforcement learning estimate

In this section, we introduce a numerical method based on policy optimisation techniques from reinforcement learning to solve the control problem (2.1).

5.1 Discrete-time reformulation of the control problem

Formulation as a Markov Decision Process

We begin by reformulating the problem within a general MDP framework, following the approach of [15] for control problems with random exit times. Let $(S_t^\alpha)_{t \geq 0}$ denote the controlled state process taking values in a domain $\mathcal{S} \subset \mathbb{R}^d$, and let α be an admissible control with values in a subset of \mathbb{R}^m . The process evolves until the random exit time

$$T^\alpha := \inf\{t \geq 0 : S_t^\alpha \notin \mathcal{O}\},$$

where $\mathcal{O} \subset \mathcal{S}$ is an open domain.

Remark 5.3. In our setting, the state variable is $S_t = (X_t, \lambda_t)$, where X_t denotes the surplus process and λ_t the Hawkes intensity. The control is $\alpha_t = (Z_t, K_t)$, consisting of cumulative dividends and capital injections.

The performance criterion is defined in terms of a running reward function f and a terminal reward function g . Given an initial state $S_0 = s$, the expected return under a control α is

$$J_\alpha(s) = \mathbb{E} \left[\int_0^{T^\alpha} e^{-\rho t} f(S_t^\alpha, d\alpha_t) + g(S_{T^\alpha}^\alpha) \right]. \quad (5.1)$$

The associated value function is

$$v(s) = \sup_{\alpha \in \mathcal{A}(s)} J_\alpha(s).$$

Remark 5.4. In our model, there is no terminal reward, i.e., $g \equiv 0$. The running reward is

$$f(S_t^\alpha, d\alpha_t) = dZ_t - \delta dK_t,$$

reflecting dividend payments and penalised capital injections.

We allow for general controlled state dynamics, potentially involving drift, diffusion, jumps, and control actions. A typical form is

$$dS_t = \mu(S_t)dt + \sigma(S_t)dW_t + \eta(S_t)dN_t + \sum_{i=1}^m d\alpha_t^i,$$

where W_t is a Brownian motion and N_t a jump process (e.g., a Hawkes process). In our model, the state $S_t = (X_t, \lambda_t)$ evolves with deterministic drift and jump-driven increments: λ_t follows Hawkes dynamics, while X_t is affected by premium inflows, claim jumps, and the control (Z_t, K_t) .

Discretisation and finite-horizon MDP approximation

Let $\mathbb{T} = \{t_0 = 0 < t_1 < \dots < t_N\}$ be a uniform time grid with step size $h > 0$. The state space is $\mathcal{S} \subset \mathbb{R}^d$, and we denote by $s \in \mathcal{S}$ the initial state. At each state s_i , the set of admissible controls is $\mathcal{A}(s_i) \subset \mathbb{R}^m$.

We consider the discretised controlled process $(S_{t_i})_{i=0}^N$, where the transition from $S_{t_i} = s_i$ to $S_{t_{i+1}}$ under control $a \in \mathcal{A}(s_i)$ is specified by a transition kernel

$$p(\cdot | t_i, s_i, a),$$

that is, $p(\cdot | t_i, s_i, a)$ is the law of $S_{t_{i+1}}$ given (S_{t_i}, a) .

Definition 5.5 (Randomised policy). A randomised policy is a measurable transition kernel

$$\pi : (t_i, s_i) \in \mathbb{T} \times \mathcal{S} \mapsto \pi(\cdot | t_i, s_i) \in \mathcal{P}(\mathcal{A}(s_i)),$$

assigning to each state a probability distribution over admissible actions. We write $\alpha \sim \pi$ for the random control sequence generated under π .

We denote by Π_h the set of all admissible discrete-time randomised policies. Under $\pi \in \Pi_h$, the controlled state process is $(S_{t_i}^\pi)_{i=0}^N$. The discrete-time exit time is defined as

$$\tau := \inf\{t_i \in \mathbb{T} : S_{t_i}^\pi \notin \mathcal{O}\}.$$

We introduce the corresponding exit index

$$N(\tau) := \inf\{i \in \{0, \dots, N\} : S_{t_i}^\pi \notin \mathcal{O}\}.$$

To obtain a discrete-time counterpart of the objective (5.1), let $A_{t_{i+1}}^\pi := \alpha_{t_{i+1}} - \alpha_{t_i}$ denote the action increment on $[t_i, t_{i+1}]$. The expected cumulative reward under a policy π is then

$$J(\pi) = \mathbb{E}_{\alpha \sim \pi} \left[\sum_{i=0}^{N(\tau)-1} f(S_{t_i}^\pi, A_{t_{i+1}}^\pi) + g(S_\tau^\pi) \right], \quad (5.2)$$

where f is the instantaneous reward function and g the terminal reward.

5.2 Policy gradient estimators

Gradient representations for policy optimization

Policy optimisation methods developed in the reinforcement learning literature provide an alternative way to approximate the solution of stochastic control problems. In this framework, the control is modelled through a parametrised family of stochastic policies $\{\pi_\theta : \theta \in \mathbb{R}^p\}$, where each policy assigns to a state a probability distribution over actions. Such distributions are typically represented by neural networks, whose parameters depend on the current state. Sampling actions from these distributions yields unbiased gradient estimators of the expected return, giving rise to policy gradient algorithms.

We now introduce a formal definition of a parametrised stochastic policy.

Definition 5.6 (Parametrised stochastic policy). *Let $\theta \in \mathbb{R}^p$. A stochastic policy π_θ is said to be parametrised if, for each (t_i, s_i) , it admits a density with respect to a reference measure ν on $\mathcal{A}(s_i)$:*

$$\pi_\theta(da | t_i, s_i) = \rho_\theta(t_i, s_i, a)\nu(da),$$

where $\rho_\theta : \mathbb{T} \times \mathcal{S} \times \mathcal{A} \rightarrow (0, +\infty)$ is a measurable function.

We restrict attention to parametrised policies of the form π_θ , and the objective becomes to optimise the parameter $\theta \in \mathbb{R}^p$ so as to maximise the discrete-time functional (5.2).

Theorem 5.3 (Objective function gradient). *Let $\theta \in \mathbb{R}^p$ and π_θ be a randomized parametrised policy. Then, the gradient of (5.2) with respect to θ is given by:*

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\alpha \sim \pi_\theta} \left[\left(\sum_{i=0}^{N(\tau)-1} f(S_{t_i}^{\pi_\theta}, A_{t_{i+1}}^{\pi_\theta}) + g(S_\tau^{\pi_\theta}) \right) \left(\sum_{i=0}^{N(\tau)-1} \nabla_\theta \log(\rho_\theta(t_i, S_{t_i}^{\pi_\theta}, A_{t_{i+1}}^{\pi_\theta})) \right) \right], \quad (5.3)$$

where we recall that $A_{t_{i+1}}^{\pi_\theta} = \alpha_{t_{i+1}}^{\pi_\theta} - \alpha_{t_i}^{\pi_\theta}$

Proof. Recall that:

$$J(\pi_\theta) = \mathbb{E}_{\alpha \sim \pi_\theta} \left[\sum_{i=0}^{N(\tau)-1} f(S_{t_i}^{\pi_\theta}, A_{t_{i+1}}^{\pi_\theta}) + g(S_\tau^{\pi_\theta}) \right]$$

The proof relies on the arguments presented in the work by Hamdouche et al. [15]. In our setting we need to increase the dimension of the dynamics. Hence, we consider the process Y defined as follows:

$$Y_t = \int_0^t e^{-\rho s} f(S_s^\alpha, d\alpha_s) + g(S_t^\alpha), \quad \text{for } t \geq 0$$

As the process $\tilde{S}_t = (S_t, Y_t)_{t \geq 0}$ is Markovian and we can apply Theorem (2.1) of Hamdouche et al. to get that

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\alpha \sim \pi_\theta} \left[\left(\sum_{i=0}^{N(\tau)-1} f(S_{t_i}^{\pi_\theta}, A_{t_{i+1}}^{\pi_\theta}) + g(S_\tau^{\pi_\theta}) \right) \left(\sum_{i=0}^{N(\tau)-1} \nabla_\theta \log(\rho_\theta(t_i, S_{t_i}^{\pi_\theta}, A_{t_{i+1}}^{\pi_\theta})) \right) \right]$$

□

The representation of Theorem 5.3 expresses the gradient of the performance functional in terms of the cumulative realised reward and the score of the policy. An alternative and often more stable estimator can be obtained by exploiting the dynamic programming structure of the value process. To this end, following [15], we introduce a dynamic version of the performance functional under the policy π_θ .

For each index $i \in \{0, \dots, N\}$ and state $s \in \mathcal{S}$, define

$$v_i^\theta(s) := \mathbb{E}_{\alpha \sim \pi_\theta} \left[\sum_{j=i}^{N(\tau_i)-1} f(S_{t_j}^{\pi_\theta}, A_{t_{j+1}}^{\pi_\theta}) + g(S_{\tau_i}^{\pi_\theta}) \middle| S_{t_i}^{\pi_\theta} = s \right],$$

where the local exit time is

$$\tau_i := \inf\{t_j \in \mathbb{T} : t_j \geq t_i, S_{t_j}^{\pi_\theta} \notin \mathcal{O}\} \wedge t_N.$$

Clearly, $v_N^\theta(s) = g(s)$, and $v_i^\theta(s) = g(s)$ for all $i < N$ whenever $s \notin \mathcal{O}$. Moreover, by the discrete-time dynamic programming principle for $s \in \mathcal{O}, i = 0, \dots, N - 1$,

$$v_i^\theta(s) = \mathbb{E}_{\alpha \sim \pi_\theta} \left[v_{i+1}^\theta(S_{t_{i+1}}^{\pi_\theta}) \middle| S_{t_i}^{\pi_\theta} = s \right].$$

Theorem 5.4 (Martingale representation). *We have:*

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\alpha \sim \pi_\theta} \left[\sum_{i=0}^{N(\tau)-1} v_{i+1}^\theta(S_{t_{i+1}}^{\pi_\theta}) \nabla_\theta \log(\rho_\theta(t_i, S_{t_i}^{\pi_\theta}, A_{t_{i+1}}^{\pi_\theta})) \right] \quad (5.4)$$

Proof. For a trajectory controlled by π_θ , define the cumulative reward process on the discrete grid by

$$Y_{t_0} := 0, \quad Y_{t_{i+1}} := Y_{t_i} + f(S_{t_i}^{\pi_\theta}, A_{t_{i+1}}^{\pi_\theta}), \quad i = 0, \dots, N(\tau) - 1,$$

and set $Y_\tau := Y_{t_{N(\tau)}} + g(S_\tau^{\pi_\theta})$. Then $J(\pi_\theta) = \mathbb{E}_{\alpha \sim \pi_\theta}[Y_\tau]$, and the augmented process $\tilde{S}_{t_i} := (S_{t_i}^{\pi_\theta}, Y_{t_i})$ is Markovian. Applying the results in Hamdouche et al. [15] to $(\tilde{S}_{t_i})_{i \geq 0}$ yields the gradient representation stated in Theorem 5.4. \square

Remark 5.5. *An equivalent expression based on the temporal differences of the value function is given by:*

$$\nabla J(\pi_\theta) = \mathbb{E}_{\alpha \sim \pi_\theta} \left[\sum_{i=0}^{N(\tau)-1} \left(v_{i+1}^\theta(S_{t_{i+1}}^{\pi_\theta}) - v_i^\theta(S_{t_i}^{\pi_\theta}) \right) \nabla_\theta \log(\rho_\theta(t_i, S_{t_i}^{\pi_\theta}, A_{t_{i+1}}^{\pi_\theta})) \right]$$

This form is particularly relevant in actor-critic methods where v_i^θ is replaced by a learned critic.

After time discretisation with step $h > 0$, the controlled process (X_t, λ_t) induces a Markov decision process with continuous action space. Rather than discretising the actions, we restrict attention to a parametrised class of stochastic policies, typically implemented through neural networks. From a theoretical perspective, the work of Kushner and Dupuis [22] shows that, when the full admissible action space is retained, the value functions of the discrete-time control problems converge to their continuous-time counterpart as $h \rightarrow 0$. The use of parametrised stochastic policies introduces a second level of approximation: the optimisation is now restricted to a subset of all admissible randomised controls. This additional approximation does not affect the consistency of the time discretisation itself, but it may prevent the algorithm from attaining the true optimal value if the optimal policy lies outside the chosen parametrised class.

Gradient-based learning algorithm

We now leverage Theorems 5.3 and 5.4 to design learning algorithms aimed at approximating optimal policies. Our first method is a direct policy gradient algorithm based on Theorem 5.3. This approach corresponds to an extension of the well-known REINFORCE algorithm introduced by Sutton [30], and has been adapted in recent works. The algorithm proceeds as follows: the policy is initialized and used to generate a collection of sample paths. For each path, the cumulative reward and the log-probabilities of the actions taken are recorded. These are then used to compute a Monte Carlo estimate of the gradient, which serves to update the policy parameters via stochastic gradient ascent.

Algorithm 1 Policy gradient algorithm

Number of episodes E , number of Monte Carlo trajectories K and learning rate η
Initialize policy π_θ with its parameters θ
for epoch $e = 1, \dots, E$ **do**
 for trajectory $k = 1, \dots, K$ **do**
 Apply current policy up to the end of trajectory $N(\tau_k) : \{(s_i, a_i)\}_{i=0}^{\tau_k}$
 Calculate total reward and log probabilities:

$$G_k = \sum_{i=0}^{N(\tau_k)-1} f(s_{t_i}, a_{t_i}) + g(s_{\tau_k})$$

$$\Lambda_k = \sum_{i=0}^{N(\tau_k)-1} \nabla_\theta \log(\rho_\theta(t_i, s_{t_i}, a_{t_i}))$$

 end for
 Compute total loss and update policy parameters by gradient ascent:

$$\theta \leftarrow \theta + \eta \frac{1}{K} \sum_{k=1}^K G_k \Lambda_k$$

end for

While the policy gradient algorithm is straightforward to implement and only requires that the policy admit a differentiable density, it does not rely on any value function approximation. This simplicity is one of its main advantages. However, a well-known drawback of REINFORCE-type methods is the high variance of the gradient estimators, which can lead to slow and unstable convergence. To address this issue, several variance-reduction techniques have been proposed. A common strategy is to subtract a baseline from the return: typically, an estimate of the value function. It helps reduce variance without introducing bias. This idea motivates the actor-critic methods presented in the next section.

Actor-critic algorithm

The second approach is an actor-critic algorithm, based on the gradient formula provided by Theorem 5.4. This method combines elements of both value-based and policy-based methods: the actor updates the policy, while the critic estimates the value function. This dual update often results in improved sample efficiency and convergence stability. We follow the methodology introduced in [30] and adapted in [15], using two neural networks: one for the policy π_θ (the actor) and one for the value function \hat{q}_ω (the critic).

Algorithm 2 Off-line actor critic policy gradient algorithm

Number of episodes E , number of trajectories to use K and η_θ and η_ω the learning rates
Initialize policy and value function π_θ and \hat{q}_ω with their parameters θ and ω
for epoch $e = 1, \dots, E$ **do**
 for trajectory $k = 1, \dots, K$ **do**
 Apply current policy π_θ up to the end of trajectory $N(\tau_k) : \{(s_{t_i}, a_{t_i})\}_{i=0}^{N(\tau_k)}$
 Calculate total advantage and log probabilities:

$$\Phi_k = \sum_{i=0}^{N(\tau_k)-1} (\hat{q}_\omega(s_{t_{i+1}}) - \hat{q}_\omega(s_{t_i})) \nabla_\theta \log(\rho_\theta(t_i, s_{t_i}, a_{t_{i+1}}))$$

$$\Psi_k = \sum_{i=0}^{N(\tau_k)-1} (\hat{q}_\omega(s_{t_{i+1}}) - \hat{q}_\omega(s_{t_i})) \nabla_\omega \hat{q}_\omega(s_{t_i})$$

 end for
 Compute total losses and update policy and value function parameters by gradient ascent:

$$\theta \leftarrow \theta + \eta_\theta \frac{1}{K} \sum_{k=1}^K \Phi_k$$

$$\omega \leftarrow \omega + \eta_\omega \frac{1}{K} \sum_{k=1}^K \Psi_k$$

end for

5.3 Reinforcement learning setup

In our framework, the observation space consists of two state variables: the current value of the Hawkes intensity process and the insurer's available cash reserves. To simulate the stochastic dynamics of the claim process and its intensity, we rely on Ogata's thinning algorithm [25]. Notably, the evolution of the intensity process is independent of the agent's actions, and thus remains unaffected by the control policy.

On the other hand, the agent directly influences the cash reserve through its actions. It is therefore essential to clearly define how the chosen policy impacts the surplus process.

In the theoretical formulation of the problem, the exit time is random and may potentially never be reached. To address this issue in our numerical implementation, we introduce a maximum time horizon $T > 0$ and define the stopping time as:

$$\tau = T \wedge \inf\{t_i \in \mathbb{T}, X_{t_i}^\pi < 0\}.$$

Naturally, the introduction of T modifies the original problem and may introduce a bias if not handled carefully. To mitigate this, we choose T large enough so that, in the absence of any control intervention, ruin occurs before time T with high probability. Formally, we select T such that

$$\mathbb{P}(\inf\{t_i \in \mathbb{T}, X_{t_i}^\pi < 0\} \geq T) \leq \varepsilon,$$

where $\varepsilon > 0$. This ensures that the finite-horizon approximation remains faithful to the structure of the original problem.

Naïve setup

We follow the MDP framework introduced in the previous section, where \mathbb{T} denotes the discretized time grid, and S_{t_i} represents the state at time $t_i \in \mathbb{T}$. In the most basic setup, we define the observation space as the pair $S_{t_i} = (X_{t_i}, \lambda_{t_i})$, and let the agent sample an action $A_{t_i}^\pi$ from a policy π , constrained to the interval $(-\infty, X_{t_i}]$. A positive action corresponds to a dividend payment, while a negative action corresponds to a capital injection. The cash reserve then evolves according to:

$$X_{t_{i+1}}^\pi = X_{t_i}^\pi + hc - \sum_{k=1}^{N_{t_{i+1}} - N_{t_i}} Y_k - A_{t_i}^\pi.$$

The agent's expected reward under policy π is then defined by:

$$J(\pi) = \mathbb{E} \left[\sum_{j=1}^{N(\tau)} e^{-\rho t_j} \left(A_{t_j}^\pi \mathbb{1}_{\{A_{t_j}^\pi \geq 0\}} + \delta A_{t_j}^\pi \mathbb{1}_{\{A_{t_j}^\pi < 0\}} \right) \right].$$

While this approach is theoretically valid, it grants the agent considerable freedom, which can significantly slow down learning due to the difficulty of balancing exploration and exploitation. In particular, it becomes challenging for the agent to discover optimal intervention timings. For this reason, we propose a more structured approach that incorporates theoretical insights derived from the analytical study presented in the first part of the paper.

Setup based on theoretical knowledge

This second approach restricts the admissible controls by imposing a two-barrier structure. For capital injections, Proposition 3.7 provides an explicit optimal threshold κ^* . For dividend payments, guided by the numerical solution of the HJB variational inequality, we postulate the existence of a state-dependent payout threshold $x^*(y)$ for $y \in [b, +\infty)$. Such a threshold is economically natural: once the surplus becomes sufficiently large, an optimal strategy must eventually prescribe dividend distributions.

We define the observation space as $S_{t_i} = \lambda_{t_i}$ and use the policy to predict the values of the optimal boundaries $\kappa^*(y)$ and $x^*(y)$. The surplus process then evolves according to:

$$X_{t_{i+1}}^\pi = X_{t_i}^\pi + hc - \sum_{k=1}^{N_{t_{i+1}} - N_{t_i}} Y_k - (X_{t_i}^\pi - x^*(y)) \mathbb{1}_{\{X_{t_i}^\pi \geq x^*(y)\}} - X_{t_i}^\pi \mathbb{1}_{\{\kappa^*(y) \leq X_{t_i}^\pi < 0\}}. \quad (5.5)$$

In this context, the agent's expected reward is:

$$J(\pi) = \mathbb{E} \left[\sum_{j=1}^{N(\tau)} e^{-\rho t_j} \left((X_{t_j}^\pi - x^*(y)) \mathbb{1}_{\{X_{t_j}^\pi \geq x^*(y)\}} + \delta X_{t_j}^\pi \mathbb{1}_{\{\kappa(y) \leq X_{t_j}^\pi < 0\}} \right) \right]. \quad (5.6)$$

This approach reduces the complexity of the learning task by restricting the agent's output to the prediction of the two optimal boundaries, rather than a full-range action. As a result, it helps accelerate training and improves the stability of the learned policy.

5.4 Numerical results

Learning boundaries

We implement both reinforcement learning algorithms together with standard regularisation techniques—such as entropy bonuses—to stabilise training and improve convergence. For comparability with the PDE-based results, we adopt the same model parameters as those reported in Table 1.

The learning procedure proceeds as follows. Given a parameter vector $\theta \in \mathbb{R}^p$, we initialise a neural network policy π_θ that takes as input the current value of the Hawkes intensity and outputs four real numbers corresponding to the parameters used to sample the control. The network architecture consists of two hidden layers of 64 neurons with ReLU activations. Trajectory generation under the policy is carried out through the following steps:

- i) At each time step, the current intensity is observed and passed through π_θ , which returns the parameters $(\mu_1, \sigma_1, \mu_2, \sigma_2)$.
- ii) From these parameters, we construct two normal distributions $\mathcal{N}(\mu_1, \sigma_1)$ and $\mathcal{N}(\mu_2, \sigma_2)$.
- iii) One sample is drawn from each distribution, and the log-probabilities of the sampled actions are recorded.
- iv) The corresponding control boundaries are constructed and applied to the surplus process according to Equation 5.5, after which steps (i)–(iii) are repeated until the ruin time is reached.

Each simulated trajectory yields a total reward together with its associated sequence of log-probabilities. Repeating this procedure M times provides a Monte Carlo estimate of the policy gradient, using either Theorem 5.3 or Theorem 5.4. The policy parameters are then updated via stochastic gradient ascent. In the actor–critic setting, the procedure remains identical except that a second neural network, with the same architecture as the policy network, is introduced to approximate the value function and serve as a learned baseline for variance reduction.

Comparison to baseline

To assess the performance of the reinforcement learning methods, we train the agents under the benchmark parameter set reported in Table 1 and compare the learned values to the reference solution obtained from the numerical resolution of the HJB variational inequality. The time discretisation step is set to $h = 1/50$, and the time maximum horizon to $T = 50$, which corresponds to a maximum of $T/h = 2,500$ time steps, an upper limit that is never reached in practice due to earlier ruin. Each policy update relies on $M = 2048$ Monte Carlo trajectories generated in parallel, with learning rates of order 10^{-3} for both the actor and the critic. Training is performed over 200 epochs for each algorithm. We consider two initial surplus–intensity states, $(x_0, y_0) = (1, 2.8)$ and $(x_0, y_0) = (0, 2.8)$, representing respectively a comfortably capitalised position and a near-boundary initial surplus.

Figures 5a–5b display the evolution of the empirical objective J during training, together with the PDE benchmark value. In both initial configurations, the actor–critic method exhibits the fastest and most stable convergence, reaching the PDE benchmark within relatively few epochs. The REINFORCE estimator also converges toward the correct value, although with slightly higher variance, which is expected for Monte Carlo policy gradients. The variance remains moderate thanks to the inclusion of a baseline term, which

stabilises the updates without introducing bias. Overall, both algorithms succeed in learning policies whose performance matches the PDE solution, thereby validating the discrete-time formulation and the policy gradient estimators developed in this section.

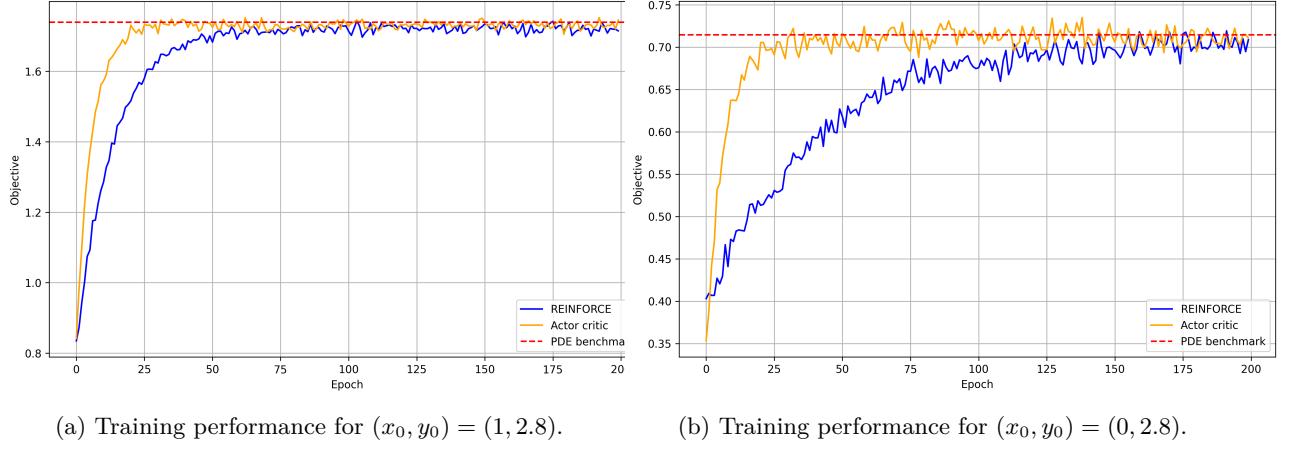


Figure 5: Convergence of the learned objective toward the PDE benchmark value.

In Figures 6a and 6b, we display the control regions learned by the reinforcement learning agent. The colour map indicates the action selected in each state: the yellow region corresponds to inaction, the purple region to capital injection, and the blue region to dividend distribution.

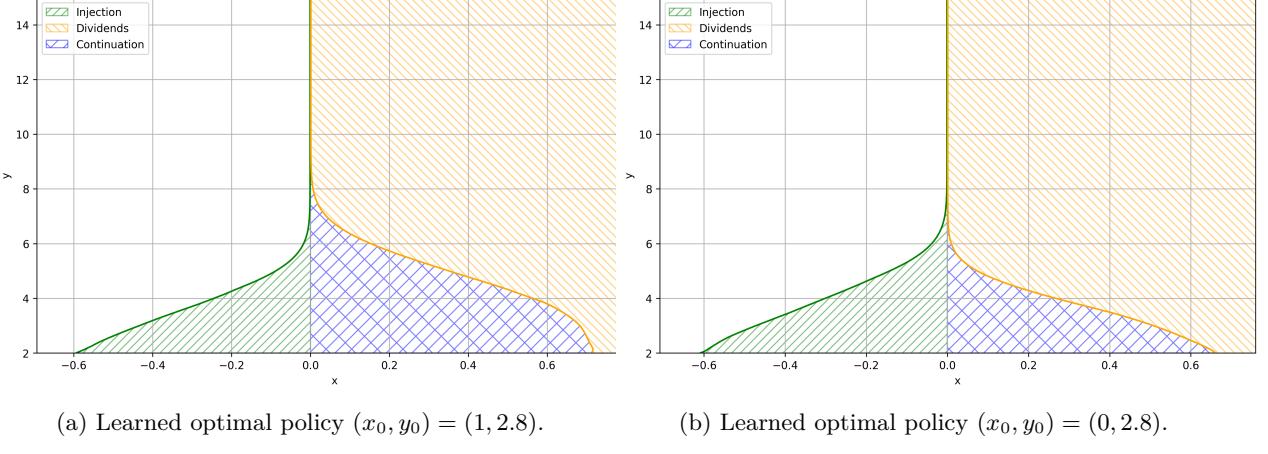


Figure 6: Learned control regions obtained by the policy gradient algorithm for two initial states.

The learned strategies display the same qualitative structure as the optimal policy obtained through the variational inequality formulation in the PDE section. This close agreement provides strong evidence for the validity of the reinforcement learning approach. Some discrepancies between the two training runs can be observed in the precise location of the control boundaries. This behaviour is expected: since policy-gradient methods optimise over a restricted class of parametrised stochastic policies, they converge to near-optimal strategies rather than an exact optimum. For such quasi-optimal policies, the control boundary is not uniquely defined. Our Monte Carlo experiments confirm that the estimated value is only slightly sensitive to variations in the dividend boundary, provided that the global structure of the optimal strategy is preserved.

| x | y | PDE | MC (Opt.) | IC95% (MC Opt.) | Rel. err. | MC (RL) | IC95% (RL) | Rel. err. |
|-----|-----|--------|-----------|------------------|-----------|---------|------------------|-----------|
| 0 | 2 | 0.8588 | 0.8414 | [0.8023, 0.8805] | -2.07% | 0.8677 | [0.8263, 0.9090] | 1.03% |
| | 3 | 0.6811 | 0.6642 | [0.6269, 0.7014] | -2.54% | 0.6840 | [0.6455, 0.7225] | 0.44% |
| | 4 | 0.5298 | 0.5181 | [0.4833, 0.5528] | -2.27% | 0.5360 | [0.5016, 0.5705] | 1.17% |
| 0.5 | 2 | 1.3874 | 1.3412 | [1.2987, 1.3838] | -3.44% | 1.3890 | [1.3467, 1.4313] | 0.12% |
| | 3 | 1.2031 | 1.1581 | [1.1166, 1.1995] | -3.89% | 1.2368 | [1.1948, 1.2788] | 2.80% |
| | 4 | 1.0360 | 0.9882 | [0.9514, 1.0249] | -4.84% | 1.0324 | [0.9920, 1.0728] | -0.35% |
| 1.0 | 2 | 1.8881 | 1.8673 | [1.8257, 1.9089] | -1.11% | 1.8872 | [1.8451, 1.9293] | -0.05% |
| | 3 | 1.7033 | 1.6886 | [1.6477, 1.7294] | -0.87% | 1.7143 | [1.6727, 1.7558] | 0.64% |
| | 4 | 1.5360 | 1.4894 | [1.4527, 1.5261] | -3.13% | 1.5312 | [1.4911, 1.5714] | -0.31% |

Table 3: Comparison of the PDE and RL estimates of the value function.

Finally, Table 3 reports three sets of values for representative state pairs. The first column (PDE) shows the benchmark value computed from the numerical solution of the HJB variational inequality. The second block provides a Monte Carlo estimate of the value obtained when applying the theoretically optimal policy to the discretised environment. The third block reports the corresponding estimate obtained using the policy learned by reinforcement learning. Both Monte Carlo values are computed from 4,096 simulated trajectories, and the reported confidence intervals are the standard asymptotic 95% confidence intervals. The relative errors reported in the table are computed by comparing respectively the Monte Carlo estimate of the value function applying theoretical optimal policy and RL Monte Carlo estimate to the PDE benchmark value. The policy learned by RL consistently outperforms the theoretically optimal continuous-time policy when both are evaluated on the discretised environment, highlighting the ability of RL to adapt favourably to numerical discretisation effects.

Beyond this qualitative agreement, the reinforcement learning framework offers two significant advantages. First, it scales naturally to higher-dimensional settings in which PDE-based methods become impractical or computationally prohibitive. Second, it provides a flexible modelling environment: changes to the claim distribution, richer dependence structures between claims and intensity, or more complex interactions in the dynamics can be incorporated with minimal modifications to the learning procedure. In this regard, reinforcement learning constitutes a powerful and adaptable tool for approximating optimal strategies in stochastic control problems with complex or high-dimensional dynamics.

Bibliography

- [1] Hansjörg Albrecher, Pablo Azcue, and Nora Muler. “Optimal dividend strategies for a catastrophe insurer”. In: *Frontiers of Mathematical Finance* 3 (Nov. 2023), pp. 304–344. DOI: 10.3934/fmf.2024008.
- [2] Hansjörg Albrecher and Stefan Thonhauser. “Optimal dividend strategies for a risk process under force of interest”. In: *Insurance: Mathematics and Economics* 43.1 (2008), pp. 134–149. ISSN: 0167-6687. DOI: 10.1016/j.insmatheco.2008.03.012.
- [3] Hansjörg Albrecher and Stefan Thonhauser. “Optimality results for dividend problems in insurance”. In: *Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales. Serie A, Matemáticas* 103 (Mar. 2009), pp. 295–320. DOI: 10.1007/BF03191909.
- [4] Soren Asmussen and Michael Taksar. “Controlled diffusion models for optimal dividend pay-out”. In: *Insurance: Mathematics and Economics* 20.1 (1997), pp. 1–15. ISSN: 0167-6687. DOI: 10.1016/S0167-6687(96)00017-0.
- [5] Pablo Azcue and Nora Muler. “Optimal reinsurance and dividend distribution policies in the Cramér-Lundberg model”. In: *Mathematical Finance* 15.2 (2005), pp. 261–308. DOI: 10.1111/j.0960-1627.2005.00220.x.

- [6] Guy Barles and Panagiotis E Souganidis. “Convergence of approximation schemes for fully nonlinear second order equations”. In: *Asymptotic analysis* 4.3 (1991), pp. 271–283. DOI: 10.3233/ASY-1991-4305.
- [7] Matteo Brachetta, Giorgia Callegaro, Claudia Ceci, and Carlo Sgarra. “Optimal reinsurance via BSDEs in a partially observable model with jump clusters”. In: *Finance and Stochastics* 28.2 (2024), pp. 453–495. DOI: 10.1007/s00780-023-00523-z.
- [8] Harald Cramér. “On the Mathematical Theory of Risk”. In: *Skandia Jubilee Volume*. Stockholm, Sweden: Skandia Insurance Company, 1930.
- [9] Donald A. Dawson and Zenghu Li. “Stochastic equations, flows and measure-valued processes”. In: *The Annals of Probability* 40.2 (2012), pp. 813–857. ISSN: 00911798. DOI: 10.1214/10-AOP629.
- [10] Bruno de Finetti. “Su un’Impostazione Alternativa Della Teoria Collettiva del Rischio”. In: *Proceedings of the Transactions of the XV International Congress of Actuaries*. New York, 1957, pp. 433–443.
- [11] Hans U. Gerber. “An extension of the renewal equation and its application in the collective theory of risk”. In: *Scandinavian Actuarial Journal* 1970.3-4 (1970), pp. 205–210. DOI: 10.1080/03461238.1970.10405664.
- [12] Hans U. Gerber. “Entscheidungskriterien für den zusammengesetzten Poisson-Prozess”. In: *Mitteilungen der Vereinigung Schweizerischer Versicherungsmathematiker* 69 (1969), pp. 185–227.
- [13] Hans U. Gerber and Elias S. W. Shiu. “On Optimal Dividend Strategies In The Compound Poisson Model”. In: *North American Actuarial Journal* 10.2 (2006), pp. 76–93. DOI: 10.1080/10920277.2006.10596249.
- [14] Hans U. Gerber and Elias S. W. Shiu. “Optimal Dividends”. In: *North American Actuarial Journal* 8.1 (2004), pp. 1–20. DOI: 10.1080/10920277.2004.10596125.
- [15] Mohamed Hamdouche, Pierre Henry-Labordere, and Huyêñ Pham. “Policy Gradient Learning Methods for Stochastic Control with Exit Time and Applications to Share Repurchase Pricing”. In: *Applied Mathematical Finance* 29 (July 2023). DOI: 10.1080/1350486X.2023.2239850.
- [16] Bjarne Hojgaard and Michael Taksar. “Optimal proportional reinsurance policies for diffusion models with transaction costs”. In: *Insurance: Mathematics and Economics* 22.1 (1998). Special issue on the interplay between insurance, finance and control, pp. 41–51. ISSN: 0167-6687. DOI: 10.1016/S0167-6687(98)00007-9.
- [17] Monique Jeanblanc-Picqué and A. N. Shiryaev. “Optimization of the flow of dividends”. In: *Russian Mathematical Surveys* 50.2 (Apr. 1995), p. 257. DOI: 10.1070/RM1995v05n02ABEH002054.
- [18] Yanwei Jia and Xun Yu Zhou. “Policy Evaluation and Temporal-Difference Learning in Continuous Time and Space: A Martingale Approach”. In: *Journal of Machine Learning Research* 23.154 (2022), pp. 1–55. URL: <https://www.jmlr.org/papers/v23/21-0947.html>.
- [19] Yanwei Jia and Xun Yu Zhou. “Policy Gradient and Actor-Critic Learning in Continuous Time and Space: Theory and Algorithms”. In: *Journal of Machine Learning Research* 23.275 (2022), pp. 1–50. URL: <https://www.jmlr.org/papers/v23/21-1387.html>.
- [20] Yanwei Jia and Xun Yu Zhou. “q-Learning in Continuous Time”. In: *Journal of Machine Learning Research* 24.161 (2023), pp. 1–61. URL: <https://www.jmlr.org/papers/v24/22-0755.html>.
- [21] Natalie Kulenko and Hanspeter Schmidli. “Optimal dividend strategies in a Cramér–Lundberg model with capital injections”. In: *Insurance: Mathematics and Economics* 43.2 (2008), pp. 270–278. ISSN: 0167-6687. DOI: 10.1016/j.insmatheco.2008.05.013.
- [22] Harold J. Kushner and Paul Dupuis. *Numerical Methods for Stochastic Control Problems in Continuous Time*. 2nd ed. Vol. 24. Stochastic Modelling and Applied Probability. 2001. DOI: 10.1007/978-1-4613-0007-6.
- [23] Arne Lokka and Mihail Zervos. “Optimal dividend and issuance of equity policies in the presence of proportional costs”. In: *Insurance: Mathematics and Economics* 42.3 (2008), pp. 954–961. ISSN: 0167-6687. DOI: 10.1016/j.insmatheco.2007.10.013.

- [24] Filip Lundberg. *1. Approximerad framställning af sannolikhetsfunktionen: 2. Återförsäkring af kollektivrisker*. Thèse de doctorat. Stockholm, 1903.
- [25] Yosihiko Ogata. “On Lewis’ Simulation Method for Point Processes”. In: *IEEE Transactions on Information Theory* 27 (Jan. 1981), pp. 23–30. DOI: 10.1109/TIT.1981.1056305.
- [26] Jostein Paulsen and Håkon K. Gjessing. “Optimal choice of dividend barriers for a risk process with stochastic return on investments”. In: *Insurance: Mathematics and Economics* 20.3 (1997), pp. 215–223. ISSN: 0167-6687. DOI: 10.1016/S0167-6687(97)00011-5.
- [27] Huyêñ Pham and Xavier Warin. “Actor-Critic Learning Algorithms for Mean-Field Control with Moment Neural Networks”. In: *Methodology and Computing in Applied Probability* 27 (2025). DOI: 10.1007/s11009-025-10142-0.
- [28] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Statistics. 1994. ISBN: 978-0-471-61977-2. DOI: 10.1002/9780470316887.
- [29] Hanspeter Schmidli. *Stochastic Control in Insurance*. 2008. DOI: 10.1007/978-1-84800-003-2.
- [30] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. Second. 2018. URL: <http://incompleteideas.net/book/the-book-2nd.html>.
- [31] Haoran Wang, Thaleia Zariphopoulou, and Xun Yu Zhou. “Reinforcement Learning in Continuous Time and Space: A Stochastic Control Approach”. In: *Journal of Machine Learning Research* 21.198 (2020), pp. 1–34. URL: <http://jmlr.org/papers/v21/19-144.html>.