Fault-Tolerant Reinforcement Learning in Continuous Time

David Mguni[1]

PROWLER.io, Cambridge, UK. davidmg@prowler.io

Abstract. Reinforcement learning (RL) is applicable in a wide number of settings such as robotics and industrial applications. Presently, continuous-time RL methods fail to produce frameworks that are reliable in instances of abrupt systemic failures, actuator faults and random stoppages. To this end, we propose a new framework that enables an RL agent to learn controls that are robust against faults and random stoppages in worst case scenarios suitable for applications with continuous state and action spaces. We use a path integral (PI) formulation that enables efficient and expedient learning, suitable for numerous practical applications. By constructing a variant of a stochastic game between a player that controls the dynamics 'controller' and a player that terminates the game 'stopper', we prove theoretical results that transform the problem into an approximation problem of a path integral with no open algorithmic parameters beyond exploration noise.

Keywords: Robust optimal stopping, reinforcement learning, stochastic differential game.

1 Introduction

Reinforcement learning (RL) seeks to address the problem of how an agent learns to maximise a cumulative sequence of rewards in unknown environments. It provides the basis for tackling a large number complex tasks in unknown environments and has been used to learn to perform a wide variety of complex tasks. Increasingly, RL methods are being deployed in a number of environments in which safe operating standards must be ensured such as healthcare, factory automation, supply chain management and autonomous helicopter control [32, 22, 18, 21].

However, at present, RL methods do not provide a satisfactory solution that can cope with random actuator faults or partial systemic failure. Such failures include randomly occurring faults and random terminal events. System failures can occur within aviation, automotive vehicles and robotics such as actuator faults [17, 9] and can severely compromise safe completion of tasks and lead to catastrophic outcomes. In such circumstances, the controller is required to act without the full availability of all actuators.

Many tasks of interest such as physical control of actuators involve continuous control. In this paper, we construct a method that enables an RL controller to determine an optimal sequence of actions that is robust against failures that

would otherwise lead to adverse outcomes in continuous environments. Our analysis produces the necessary framework that provides a fault-tolerant (FT) RL method that is suitable for financial applications, robotics and other settings involving physical control.

In order to find the optimal FT control policy, it is firstly necessary to determine a worst-case scenario stopping criterion to which the control policy is able to respond. The framework we introduce determines at which points stopping the system induces the worst outcomes and, the corresponding FT control. As we then show, the problem admits a stochastic differential game (SDG) representation between a controller and an adversarial stopper.

Under this interpretation, the outcome is determined by a controller that affects the state process by exercising a control whilst playing against an adversary that selects the time at which to stop the game. The resulting framework finds an optimal control that is robust against faults and stoppages at times that pose adverse risk. We define risk in the worst-case scenario sense — given the complete set of probability distributions, the agent considers the worst-case in assessing the expected payoff.

The framework provides a solution to optimal stopping problems (SPs) under worst-case scenarios. SPs belong to a class of optimal stochastic control (OSC) problems in which the goal is to determine a criterion for stopping the system at a time that maximises some state-dependent payoff [30]. SPs are widespread within economics and industrial applications.

The contributions of the paper are as follows: we, for the first time show that the FT RL problem in continuous-time admits an SDG representation. This allows us to use SDG theory to show that the solution to the problem can be computed by finding a saddle point equilibrium of the corresponding SDG. Second, by adapting techniques from PI control theory to a game setting, we show that the problem of finding the optimal value function (VF) is reducible to constructing a PI estimate which, in turn, can be approximated by forward sampling of a diffusion process. Thirdly, we prove an equivalence between the game of control and stopping and optimal stopping problems under worst-case scenarios. This in turn, allows us to show for the first time that SPs under worst-case scenarios can be tackled using PI approximation.

The resulting framework yields a learning framework that learns FT control. The method requires neither the transition dynamics nor the reward function to be known up-front and involves no free algorithmic parameters other than exploration noise.

The paper is organised as follows: in Sec. 2 we discuss relevant works within OSC and RL. In Sec. 3, we give a formal description of the problems and establish the connection between FT RL and SDGs. and provide a practical example drawn from robotics. In Sec. 4., we give some necessary mathematical background in SDG theory and discuss PI control. In Sec. 5, we summarise our main results; we defer the proofs the main analysis which is performed in Sec. 6. In Sec. 7, we study SP and show that our method solves SPs in worst-case scenarios. Lastly in Sec. 8, we provide some concluding remarks.

2 Related Work

The literature on FT control within RL is limited. In [25] an FT framework was considered in a discrete setting using approximate dynamic programming (APD) which suffers from the curse of dimensionality — the rate of convergence grows exponentially with the state space dimensionality. Additionally, given the discrete time, state and action setting, the method in [25] is generally unsuitable for environments that require continuous control such as finance and robotics.

Within OSC, combined control and stopping has been studied in a few cases such as [6]. Similarly, games of control and stopping have been analysed in continuous-time in specific contexts e.g. linear diffusions [20], geometric Brownian motion [4] and jump-diffusions [2, 24, 23]. These analyses however, assume that the model of the transition dynamics and reward are known up-front and are therefore unsuitable for RL settings.

There is a plethora of work on SPs [30]. [35] uses approximate dynamic programming methods to construct an iterative scheme to compute the solution of an SP. Despite the importance of risk in RL, work on SPs is limited to value function based iterative methods [25] and are typically restricted to risk-neutral settings e.g. [35] which do not permit the inclusion of a controller. Introducing a notion of risk (generated adversarially) adds considerable difficulty as the solution concept is now an SDG saddle point equilibrium, the existence of which must be established.

Our results generalise existing analyses to strategic settings with both a controller and an adversarial stopper which tackles risk within SPs.[1]

3 Problem Formulation

We now describe the problem with which we are concerned, namely FT RL. We will later prove an equivalence between the FT problem and an SP and construct a method to generate solutions for both problems.

3.1 Fault-Tolerant Reinforcement Learning

The FT control problem we are concerned with requires learning a criterion for stopping a system (sub)process at the worst possible time — that is, the point (state and time) at which terminating some system process incurs the greatest cost to the controller. Applying this stopping rule to the system subsequently induces a response by the controller that is robust against systemic faults in worst-case scenarios. Tackling this problem necessitates a formalism that combines both an SP which seeks to determine an optimal time to arrest at some worst possible state with single control RL.

In this problem the controller uses a control variate $u \in \mathcal{U}$ to modify the system dynamics. At time $s \in [0, T]$ and when the system is at state X_s, the

[1] SPs have been extended to games where the action of each player 1s restricted to one of two actions; to stop the game or continue. [8].

controller incurs a cost $f(s, X_s, u)$, where u is the magnitude of the controller's influence on the system. At any given point $\tau_S \leq T$ the system may stop and the problem terminates where $\tau_S \sim \ell([0,T])$ is a measurable, random exit time and f is some distribution on $[0,T]$. At this point, the controller incurs a cost $K(X_{\tau_S})$ and the system terminates.

The objective function J is given by the following:

$$J^{\tau_S}[x] = \mathbb{E}\left[\int_0^{\tau_S} f(s, X_s, u)ds + e^{-\gamma(\tau_S \wedge T)}K(X_{\tau_S})\right], \tag{1}$$

where $\tau_S \sim \ell([0,T]), x \in \mathcal{S}$ is the initial point of the system, $\mathcal{S} \subset \mathbb{R}^d$ is the state space. The parameter $u \in \mathcal{U}$ is the control contained within some admissible control set $\mathcal{U} \subseteq \mathbb{R}^{d\times 1}$ and the set $\mathcal{T} \subseteq [0,T]$ consists of all $\mathcal{F}-$measurable stopping times. The functions $f : [0,T] \times \mathcal{S} \to \mathbb{R}$ and $K : [0,T] \times \mathcal{S} \to \mathbb{R}$ represent the running cost or cost-to-go and the terminal cost functions (resp.) and $\gamma \in \mathbb{R}_{>0}$ is the discount factor.

The FT control problem which we tackle is one in which the controller acts with concern for stopping at states which incur high costs or potentially catastrophic outcomes. Such examples are often encountered in practical scenarios, for example aviation faults at specific times, actuator failures of medical failures and life support machines in addition to assisting device failures.

To this end, we develop a framework which produces fault-tolerant control that can cope with abrupt system or (sub)process stoppages and failures at the worst possible time. To produce such control it is firstly necessary to determine a stopping rule for the worst possible stopping times; applying this stopping rule to the system subsequently induces a response by the controller that is robust against systemic faults in worst-case scenarios. Tackling this problem necessitates a formalism that combines both an SP which seeks to determine an optimal time to arrest at some worst possible state with single control RL.

The task involves finding both a worst-case stopping time $\hat{\tau}$ and an optimal (FT) control $\hat{u}$. We are therefore concerned with problems of the following kind:

Find $(\hat{\tau}, \hat{u}) \in \mathcal{T} \times \mathcal{U}$ and $J^{\hat{\tau},\hat{u}}[s]$ s.th.

$$\sup_{u\in\mathcal{U}}\left(\inf_{\tau\in\mathcal{T}} J^{\tau,u}[x]\right) = J^{\hat{\tau},\hat{u}}[x], \tag{2}$$

where the objective function J is given by:

$$J^{\tau,u}[x] = \mathbb{E}\left[\int_0^{\tau\wedge T} f(s, X_s, u)ds + e^{-\gamma(\tau\wedge T)}K(X_{\tau\wedge T})\right], \tag{3}$$

where $a \wedge b := \min\{a, b\}$.

The transition dynamics are described by a diffusion process on X; without the inclusion of the controller's influence, the system dynamics are given by:

$$dX_s = \mu(s, X_s)ds + \sigma(s, X_s)dB_s + h(s, X_s)dP_s, \tag{4}$$

where $P \in \mathbb{R}^{m\times 1}$ is Poisson distributed and $h : [0,T] \times \mathcal{S} \to \mathbb{R}^{d\times m}$ is the Poisson process coefficient or jump amplitude. $B(s,x) : [0,T] \times \mathbb{R}^d \to \mathbb{R}^m$ is an $m-$dimensional standard Brownian motion. Both P and B are independent and are supported by the filtered probability space and $\mathcal{F}$ is the filtration of $(\Omega, \mathbb{P}, \mathcal{F} = (\mathcal{F}_s)_{s\in[0,T]})$. The functions $\mu : [0,T] \times \mathcal{S} \to \mathbb{R}$ and $\sigma : [0,T] \times \mathcal{S} \to \mathbb{R}^{d\times m}$ are the drift coefficient and the diffusion coefficient which describe the central tendency and volatility of the system respectively. We assume that σ, μ are Lipschitz continuous and satisfy a polynomial growth condition (see appendix).

Note the process (4) has the following discrete analogue:

$$X_{n+1} = X_n + \mu(s_n, X_n)\delta s + \sigma(s_n, X_n)(B_{s_{n+1}} - B_{s_n})h(s_n, X_n)(P_{s_{n+1}} - P_{s_n}),$$

When the controller acts on the state process, the process (4) evolves according to the expression:

$$dX_s = \hat{\mu}(s, X_s, u)ds + \sigma(s, X_s)dB_s + h(s, X_s)dP(s, X_s), \tag{5}$$

where $\hat{\mu} : [0,T] \times \mathcal{S} \times \mathcal{U} \to \mathbb{R}^p$ is the controlled drift coefficient.

The construction (2) can be thought of as being an SDG between a controller that is delegated the task of executing the control $\hat{u} \in \mathcal{U}$ to maximise J against a stopper that seeks to find a time $\hat{\tau} \leq T$ to stop the process at the worst possible time (i.e. that which minimises J). With this interpretation, the pair $(\hat{\tau}, \hat{u})$ constitutes an equilibrium in which each player responds optimally to their opponents' control, hence the induced control $\hat{u} \in \mathcal{U}$ is FT.

Note that the problem (2) is degenerate — if the set $\mathcal{U}$ is restricted to a singleton, then the problem collapses to a (risk-neutral) SP. Conversely if the set $\mathcal{T}$ is a singleton then the problem reduces to a risk-neutral MDP.

In this paper, we prove theoretical results that demonstrate how to solve problems of this kind. We specialise to the linear quadratic case in which the running cost for each player i takes the form:

$$f(s, x, u) \equiv f(s, x) + \frac{1}{2}u^T Ru, \tag{6}$$

where R is an invertible symmetric matrix which is the control weight. The drift coefficient $\hat{\mu}$ is now:

$$\hat{\mu}(s, x, u) \equiv \mu(s, x) + G(s, x)^T u. \tag{7}$$

where $G : [0,T] \times \mathcal{S} \to \mathbb{R} \to \mathbb{R}^{p\times d}$.

Linear quadratic models have been shown to have high performance in tasks throughout robotics and engineering [1] and have been successfully within RL settings to tackle physical control [34, 33].

To elucidate the idea, we give an example of the FT problem. As the following illustrates, the framework applies to actuator failure in RL applications. The example is adapted from the discrete setting in [25].

3.2 Example: Control with abrupt actuator failure

Consider an agent e.g. a robot that uses an actuator to perform actions. Given full operability of its actuator, the agent exercises a control $u \in \mathcal{U}$ that modifies the dynamics of a diffusive process X^u (we emphasise the control on X with the superscript). In many systems, there exists some risk of actuator failure at which point the agent thereafter can affect the state transitions without the use of its actuator. Subsequently, the agent's control is reduced to influencing the system in only a subset of the system state components. Considering the case of full actuator failure. Denoting by u_0 a control variate that does not act on the system, the agent's subsequent payoff (after failure) is determined by the uncontrolled dynamics of the system X^{u_0} in which case, its expected return as of that point is given by the VF, $\tilde{V}$.

In order to perform robustly against actuator failure, it is therefore necessary to consider a set of stopping times $\mathcal{T} \subseteq [0, T]$ after which, the robot can no longer select actions that require functionality of its actuators. In particular, in order to construct a robust policy against catastrophic outcomes, it is necessary to consider actuator failure in worst-case scenarios.

The problem involves finding a pair $(\hat{\tau}, \hat{u}) \in \mathcal{T} \times \mathcal{U}$ which consists of a stopping time and control policy s.th.

$$\sup_{u \in \mathcal{U}} \left(\inf_{\tau \in \mathcal{T}} \mathbb{E} \left[\int_t^{\tau \wedge \infty} e^{-\gamma s} f(X_s^u, u) ds + e^{-\gamma(\tau \wedge \infty)} \tilde{V}(X^u_{\tau \wedge \infty}) \right] \right),$$

where $\tilde{V}(X^u_\tau) := \mathbb{E}\left[\int_t^{\tau_S \wedge \infty} e^{-\gamma s} f(X_s^{u_0}, u_0) ds + e^{-\gamma(\tau \wedge \infty)} K\left(X^{u_0}_{\tau_S \wedge \infty}\right) \right]$.

In order to generate the control $\hat{u}$, the adversary is included which simulates actuator failure at the time which inflicts the greatest cost to the controller. The resulting control $\hat{u}$ is a best-response against actuator failures at the worst-possible times.

In order to extract the pair $(\hat{\tau}, \hat{u})$ we formulate the problem as an SDG and seek to compute the equilibrium controls of the game. To do so, we appeal to the theoretical machinery within SDG theory — to this end, we first introduce some necessary SDG concepts that shall underpin our analysis. Having formulated the problem as a game, we will then appeal to a formulation of OSC, namely path integral control in order to construct a simulation-based method that enables us to compute the value function of the game.

4 Background

In this section we give some mathematical background which underpins the main analysis in Sec. 5.

4.1 Stochastic Differential Game Theory

SGDs are strategic settings in which two or more players continuously alter the dynamics of a stochastic system by strategically selected magnitudes. The task

of each player is to alter the system dynamics so as to maximise their individual state-dependant payoff [3].

A description of a two-player SDG is as follows: consider two players, player 1 and player 2. Each player influences the diffusion process (4) using a set of admissible controls $u \in \mathcal{U}$ (resp., $v \in \mathcal{V}$) where $\mathcal{U}$ (resp., $\mathcal{V}$) is an admissible control set for player 1 (resp., player 2). Each player's control modifies the drift coefficient (which is now augmented to include the controls of two players) so that the controlled process of the SDG is given by:

$$dX_s = \hat{\mu}(s, X_s, u, v)ds + \sigma(s, X_s)dB_s + h(s, X_s)dP_s,$$

where $\hat{\mu} : [0, T] \times \mathcal{S} \times \mathcal{U} \times \mathcal{V} \to \mathbb{R}$ is the controlled drift function and σ, B, h, P are as described in Sec. 3.1.

Each player $i \in \{1, 2\}$ has a cost function $J_i : [0, T] \times \mathcal{S} \times \mathcal{U} \times \mathcal{V} \to \mathbb{R}$ which it seeks to minimise:

$$J_i^{u,v}[x] = \mathbb{E}\left[\int_0^T f_i(s, X_s, u, v)ds + e^{-\gamma T}K_i(X_T)\right],$$

where $f_i : [0, T] \times \mathbb{R}^p \times \mathcal{U} \times \mathcal{V} \to \mathbb{R}$ is the running cost, $K_i : [0, T] \times \mathbb{R}^p \to \mathbb{R}$, is the terminal cost and $\gamma \in \mathbb{R}_{>0}$ is the game discount factor.

We are interested in zero-sum games which are games in which following condition $J_1 = -J_2$ holds.

We are interested in constructing an equilibrium concept for zero-sum games. The value of the game exists if we can commute the sup and inf operators so that we have that $\sup_{u\in\mathcal{U}} \inf_{\tau\in\mathcal{T}} J^{\tau,u}[\cdot] = \inf_{\tau\in\mathcal{T}} \sup_{u\in\mathcal{U}} J^{\tau,u}[\cdot]$. We denote the value by $\hat{J}$ and denote by $(\hat{k}, \hat{u}) \in \mathcal{T} \times \mathcal{U}$ the pair that satisfies $J^{\hat{k},\hat{u}} \equiv \hat{J}$. The value, should it exist, is the minimum payoff each player can guarantee itself under the equilibrium strategy. Should the value $\hat{J}$ exist, it constitutes a saddle point equilibrium of the game in which neither player can improve their payoff by playing some other control — an analogous concept to a Nash equilibrium for the case of two-player zero-sum games.

Definition 1. The pair $(\hat{\tau}, \hat{u}) \in \mathcal{T} \times \mathcal{U}$ is a saddle point equilibrium iff $\forall x \in [0, T] \times \mathcal{S}$:

$$J^{\hat{\tau},\hat{u}}[x] = \sup_{u\in\mathcal{U}} J^{\hat{\tau},u}[x] = \inf_{\tau\in\mathcal{T}} J^{\tau,\hat{u}}[x]. \tag{8}$$

We now introduce a central concept within game theory:

A saddle point equilibrium is a strategic configuration in which each player executes it best-response (BR) strategy where a BR strategy is defined by:

Definition 2. The set of BR strategies for player 1 against the stopping time $\tau \in \mathcal{T}$ (BR strategies for player 1I against the control $u \in \mathcal{U}$) is defined by $\hat{u} \in \underset{u'\in\mathcal{U}}{\arg\sup}\, \mathbb{E}[J^{\tau,u'}[\cdot]]$ (resp., $\hat{\tau} \in \underset{\tau'\in\mathcal{T}}{\arg\inf}\, \mathbb{E}[J^{\tau',u}[\cdot]]$).

It can be shown that the above transition dynamics admit the Markov property [28]. For SDGs with Markovian transition dynamics, path dependent strategies can be disposed of without destroying the existence of an equilibrium of the game. Consequently, without loss of generality, we restrict ourselves to the class of Markov controls, that is controls that depend only on the current state and round. It is well-known that for stochastic games, an equilibrium exists in Markov strategies even when the opponent can draw from non-Markovian strategies [15].

The literature on SDGs is primarily concerned with two player settings in which both players exercise their controls to jointly modify the diffusive process [10, 12]. In order to tackle problem (2), we depart from this model and consider a game in which player 2 is entitled to choose a time to stop the process i.e. the player 2 control set is $\mathcal{T} \subseteq [0, T]$ which consists of ($\mathcal{F}-$measurable) stopping times. Thus, player 1 can manipulate the system dynamics with its control and, at any point, player 2 can decide to stop the process.

Formulating the problem as a game enables us to construct a characterisation of the optimal controls, in particular, when an equilibrium is achieved, the controls being exercised by each player will be best-response controls. However, the issue of how to compute the value function and hence, extract the optimal controls remains.

In order to compute the value of the game, we appeal to a formulation of OSC theory that uses a path integral to produce an estimate of the VF. We proceed to give a brief overview of path integral control theory.

4.2 Path Integral Control Theory

In continuous-time, the problem of finding the optimal VF in OSC problems can be reduced to solving a partial differential equation (PDE) known as the Hamilton-Jacobi-Bellman (HJB) equation. In general, the HJB equation is a non-linear second order PDE.

Path integral control (PIC) is a formulation of OSC in which the evolution of the VF is described as a functional integral over all intermediate paths satisfying some given boundary conditions [19]. Under a coupling assumption between control costs and the variance, the PIC framework enables the VF of RL problems with quadratic control costs and linear control to be computed by forward sampling of the uncontrolled diffusion process. Since estimates of the optimal control policies are now reduced to approximating a PI, computing the VF involves only the path costs of the state trajectory.[2] This has led to significant gains in computing optimal controls in continuous-time applications of RL [34, 33]. In PIC, the following coupling relation between the control cost and the diffusion coefficient is assumed: $G(s, x)^T R^{-1} G(s, x) = \lambda \sigma^T \sigma 1_p$, where 1_p is the $p-$dimensional

[2] Also, the convergence rate of PI approximation is independent of the computational explosion with the state space dimensionality [31] as encountered in other methods for solving HJB equations e.g. finite difference methods [26].

identity matrix, G is the control matrix in (7) and $\lambda \in \mathbb{R}_{>0}$. The coupling assumption implies that the control cost decreases as the system variance increases which enables the linearisation of the HJB equation. This allows for estimation of the VF by approximation of a PI (see next section). The assumption is however frequently violated in numerous settings in finance, economics and RL [7, 13] rendering PIC unsuitable for these problems.

Using the theory of backward stochastic differential equations (BSDEs), we demonstrate how the PI approximation techniques can be used without imposing the coupling assumption. This enables application in a broad range of settings within robotics economics for which the coupling assumption does not hold.

The classical Feynman-Kac lemma enables linear parabolic PDEs to be solved numerically by Monte Carlo simulations of the stochastic process [12]. Recent developments using backward stochastic differential equations (BSDEs) have led to the development of versions of the Feynman-Kac lemma that apply to semilinear second order PDEs — PDEs in which the PDE may have nonlinear of all terms up to and including the first order derivative. This nonlinear Feynman-Kac formula yields a nonlinear Monte Carlo method via the BSDE to solve the semilinear HJB equation in a numerical fashion. Full details of this approach can be found in [29, 31].

5 Summary of Main Results

We now summarise the main results of the paper the proofs of which we defer to the next section.

We define the adversary's continuation region D as:

$$D := \{X(s) \in \mathcal{S}; \psi(X(S)) > K(X(S))\}. \tag{9}$$

The main result of the paper is to characterise the set of BR controls and construct an RL method for computing the VF. In particular, we prove the following:

The FT optimal control $\hat{u} \in \mathcal{U}$ is given by:

$$\hat{u} = -\frac{1}{2} R^{-1} [G(s,x)]^T \nabla_x \psi(s,x), \tag{10}$$

and the adversary's optimal stopping time $\hat{\tau} \in \mathcal{T}$ is:

$$\hat{\tau} = \inf\{s > 0 | X(s) \notin D; s \in [0,T]\}, \tag{11}$$

where

$$\psi(x) = \int P\left(X_{\hat{\tau}} \middle| x\right) \cdot \left(\int_0^{\hat{\tau}} f(X_s) ds + K(X_{\hat{\tau}})\right) dx, \tag{12}$$

with $s \in [t, T-t)$ and for any $x(s+t) \equiv X_{s+t} \in \mathbb{R}^p$ and where P is the probability transition matrix.

We prove that the pair $(\hat{\tau}, \hat{u}) \in \mathcal{T} \times \mathcal{U}$ consists of BR strategies and constitutes a saddle point equilibrium. We lastly show that the stopping time $\hat{\tau}$ is a solution to the SP described in Sec. 6.1. The results allow the VF to be computed by forward sampling of (4).

6 Main Analysis

We now develop the main theory of the paper and prove the results of Sec. 5. In particular, we provide a full characterisation of the game and show that the VF can be approximated by estimating a PI. This allows for an RL method with no open parameters other than exploration noise. We then characterise the optimal stopping time for the stopper and the FT optimal control.

We denote by $\mathscr{C}([a,b],\mathbb{F})$ the set of continuous functions from $\mathbb{R}$ to a field $\mathbb{F}$ over the interval $[a,b]\subseteq\mathbb{R}$. The temporal derivative is denoted by ∂_s and the first, second and n^{th} spatial derivative (resp.): $\nabla_{x_i}, \nabla^2_{x_i,x_j}, \nabla^n_{x_1,\dots x_n}$. We denote by $\mathscr{C}^{1,2}([0,T],\Omega)=\{h\in\mathscr{C}(\Omega):\partial_t h, \nabla_{x_i,x_j}h\in\mathscr{C}([0,T],\Omega)\}$ and by $\mathbb{S}(p)$ is the set of invertible symmetric $p\times p$ matrices for a field $\mathbb{F}\subseteq\mathbb{R}^p$.

We begin introducing the Dynkin operator of the controlled process (5), acting on some function $\phi\in\mathscr{C}^{1,2}([0,T]\times\mathcal{S})$ and $\forall x\in[0,T]\times\mathcal{S}$ by:

$$\mathscr{L}^u\phi(x)=\sum_{i=1}^{k}\nabla_{x_i}\phi(x)^T\hat{\mu}_i(s,x,u)+\frac{1}{2}\sum_{i,j=1}^{p}(\sigma\sigma^T)_{ij}\nabla^2{}_{x_i,x_j}\phi(x)+D_{\text{jump}}\phi(x),$$

where D_{jump} is defined by: $D_{\text{jump}}\phi(s,x):=\sum_{j=1}^{p}\gamma_j[\phi(s,x+h_j(s,x))-\phi(s,x)]$, and where $\rho_j:\mathbb{R}\to\mathbb{R}^{m+1}$ is the jump-rate vector of P.

We begin with the following lemma that describes the Dynkin operator under the optimal player 1 control:

Lemma 1. The following holds $\forall x\in[0,T]\times\mathcal{S}$:

$$\sup_{u\in\mathcal{U}}\mathscr{L}^u\phi(x)=\mathscr{A}\phi(x) \tag{13}$$

where $\mathscr{L}^u$ is the Dynkin operator (c.f. (13)) and the operator $\mathscr{A}$ is given by: $\mathscr{A}:=\mu(s,x)\nabla_x\phi(x)+\sum_{i,j=1}^{d}(\sigma\sigma^T)_{ij}\nabla^2_{xx}\phi(x)+\nabla_x\phi(x)G(s,x)R^{-1}G(s,x)\nabla_x\phi(x)+D_{\text{jump}}\phi(x)$.

The following theorem fully characterises the VF:

Theorem 1. Let $\psi\in\mathscr{C}^{1,2}([t,\tau_S];\mathbb{R}^p)$ satisfy:

$$\sup_{u\in\mathcal{U}}\left[\frac{\partial\psi}{\partial s}(x)+\mathscr{A}\psi(x)+f(s,x,u)\right]\begin{cases}=0, & \forall x\in D,\\ \leq 0, & \forall x\notin D,\end{cases} \tag{14}$$

then ψ is the VF of the game, that is $\forall x\in[0,T]\times\mathcal{S}$ we have $\psi[x]=J^{\hat{\tau},\hat{u}}[x]$.

The proof relies on an application of Itō's lemma for jump diffusions [28], the mean value theorem and constructing a sequence of continuation regions in which stopping is suboptimal for the stopper. The result is a minor modification of Theorem 2.1 in [2] in which the result is derived for diffusions with jumps generated by compensated Poisson random measures.

Theorem 1 states that the solution to (14) enables us to recover the VF to the game. Crucially, this reduces the problem to finding a solution to (14). Even

though this represents some progress, obtaining a solution to (14) is generally inaccessible through analytic methods.

The following is a direct consequence of the Theorem:

Corollary 1. The pair $(\hat{\tau}, \hat{u}) \in \mathcal{T} \times \mathcal{U}$ consists of BR strategies and constitutes a saddle point equilibrium in Markov strategies.

We now state a key result:

Proposition 1. For all $x \in [t, \tau_S] \times \mathcal{S}$, the VF ψ satisfies:

$$\min\left\{-\frac{\partial \psi}{\partial s}(x) - \left(\mathcal{A}\psi(x) + f(x)\right), \psi(x) - K(x)\right\} = 0, \tag{15}$$

Prop. 1 exhibits the fact that the problem satisfies an obstacle problem. The intuition behind (1) is that whilst the game is in play, player 1 executes its BR strategy and the VF satisfies the HJB equation. However, the expected cumulative future costs for player 2 can never exceed $K(x)$ since the rational choice for player 1 is to stop the game as soon as parity of its future costs occur.

Proof of Prop. 1. We shall refer to (15) as the Hamilton-Jacobi-Bellman-Isaacs condition or HJBI condition for short.

We initiate the proof by considering a suboptimal stopping time for player 2, $\tilde{\tau} \in \mathcal{T}$, hence:

$$V(\cdot) = \sup_{u \in \mathcal{U}} \inf_{\tau \in \mathcal{T}} J^{\tau, u}[\cdot] \leq \sup_{u \in \mathcal{U}} J^{\tilde{\tau}, u}$$

We can equivalently write this as:

$$V(x) \leq \sup_{u \in \mathcal{U}} \mathbb{E}\left[\int_0^{\tilde{\tau}} f(s, X_s, u) ds + e^{-\gamma(\tilde{\tau} \wedge T)} K(X_{\tilde{\tau} \wedge T})\right],$$

Using the definition of the VF and subtracting $V(x)$ from both sides, we find::

$$0 \leq \sup_{u \in \mathcal{U}} \mathbb{E}\left[\int_0^{dt \wedge \tilde{\tau}} f(s, X_s, u) ds + e^{-\gamma(\tilde{\tau} \wedge T)} V(t + dt, X_{dt}) - V(x)\right],$$

After taking the limit as $dt \to 0$ we find:

$$0 \leq \sup_{u \in \mathcal{U}} \mathbb{E}\left[f(s, X_s, u) ds + e^{-\gamma(\tilde{\tau} \wedge T)} dV(x)\right], \tag{16}$$

By Itō's lemma for jump-diffusion processes [28], the total variation, $d\phi$ is:

$$d\phi(x) = \sum_{i=1}^{k} \nabla_{x_i} \phi(x)^T \left(\hat{\mu}_i(s, x, u) ds + \sigma_i(s, x) dB_s\right) + \frac{1}{2} \sum_{i,j=1}^{p} (\sigma\sigma^T)_{ij} \nabla^2{}_{x_i, x_j} \phi(x) + D_{\text{jump}} \phi(x) + \frac{\partial \phi}{\partial s}(x). \tag{17}$$

Further, by the standard properties of Brownian motion (and since $V \in \mathscr{C}^{1,2}([0,T],\mathbb{R}^p)$:

$$\mathbb{E}\left[\sum_{i=1}^{k} \nabla_{x_i} V(x)^T \sigma_i(s,x) dB_s\right] = 0 \tag{18}$$

Therefore, inserting (17) into (16) and using (18) we find:

$$-\frac{\partial V}{\partial s}(x) - \left(\sup_{u\in\mathcal{U}} \mathscr{L}^u V(x) + f(x)\right) \geq 0, \tag{19}$$

Using Lemma 1, we can equivalently express (19) as:

$$-\frac{\partial V}{\partial s}(x) - (\mathscr{A} V(x) + f(x)) \geq 0 \tag{20}$$

It remains only to consider the case when an immediate termination is optimal for the stopper. In this case, the following expression holds:

$$V(x) = K(x) \tag{21}$$

Hence, combining (21) with (20), yields the following:

$$\min\left\{-\frac{\partial V}{\partial s}(x) - (\mathscr{A} V(x) + f(x)), V(x) - K(x)\right\} = 0,$$

which is the required result. □

The issue of how to compute the solution to (14) remains. Moreover, the analysis has thus far proceeded as if the reward function and transition model are known up-front. We now show how an RL method can be used to overcome these problems by way of estimating a PI.

The PDE in (14) belongs to a class of PDEs known as semilinear parabolic PDEs. After further deduction, we can extract the BR controls for each player (this in proven in Theorem 2) . For the case involving a controller and no stopper, in order to linearise the operator $\mathscr{A}$, current PI methods impose the coupling assumption. As we now demonstrate, since the VF is a solution to a semilinear parabolic PDE, we can construct a PI representation of the VF without imposing the coupling assumption.

We now state the Feynman-Kac lemma (a proof for jump-diffusion processes can be found in [34]):

Lemma 2. [Generalised Feynman-Kac Lemma for Jump-Diffusion Processes] Suppose that $\psi \in \mathscr{C}^{1,2}([0,T];\mathbb{R}^p)$ and $a, \eta \in \mathscr{C}([0,T];\mathbb{R}^p)$, given the LP PDE:

$$\partial_s \psi(s,y) + \hat{H}\psi(s,y) = \eta(s,y), \tag{22}$$

where $\hat{H}$ is the Hamiltonian operator: $\hat{H}[\psi] := \alpha a(s,x)\psi(x) + \mathscr{L}\psi(x)$ and $\mathscr{L}$ is the Dynkin operator of (5) and $(s,y) \in [0,T] \times \mathcal{S}, \alpha \in \mathbb{R}$. The VF is ψ then:

$$\psi(x) = \mathbb{E}\left[\psi(X_T)\exp\left(\alpha \int_0^T a(r,x_r)dr\right) + \int_0^T \eta(x)\exp\left(-\int_0^T a(r,x_r)dr\right) ds\right],$$

where the expectation is taken w.r.t the probability measure $\mathbb{P}$ such that the forward diffusion process obeys (4).

Applying the generalised Feynman-Kac lemma yields:

$$\psi(x) = \mathbb{E}\left[K\left(X_{\hat{\tau}}|x\right) + \int_0^{\hat{\tau}} f(s, X_s, u_0)ds\right]. \tag{23}$$

where $\hat{\tau} = \inf\{s > 0|X(s) : \psi(X(S)) \leq K(X(S))\}$.

Expression (23) can be equivalently written as:

$$\psi(x) = \int P\left(X_{\hat{\tau}}\Big|x\right) \cdot \left(\int_0^{\hat{\tau}} f(s, X_{\hat{\tau},u_0})ds + K(X_{\hat{\tau}})\right) dx, \tag{24}$$

with $s \in [t, T-t)$ and for any $x(s+t) \equiv X_{s+t} \in \mathbb{R}^p$, the probability transition matrix P is given by:

$$\begin{aligned}
&P(s+t, x(s+t)|t, x(t)) \\
&= (1-\rho(t)\delta t) \cdot \left[\frac{1}{\sqrt{2\pi}\sigma\delta t}\exp\left\{-((x(s+t)-x(t)-\mu(x,t)\delta t)^2/2\sigma)\right\}\right] \\
&\quad + \rho(t)\delta t\left[\frac{1}{\sqrt{2\pi}\sigma\delta t}\exp\left\{\frac{-1}{2\sigma}((x(s+t)-x(t)-(f(x,t)+h(x,t))\delta t)^2)\right\}\right].
\end{aligned} \tag{25}$$

This underscores the following result; we defer the proof to the appendix:

Theorem 2. The optimal FT control $\hat{u} \in \mathcal{U}$ is given by:

$$\hat{u} = -\frac{1}{2}R^{-1}[G(s,x)]^T\nabla_x\psi(s,x), \tag{26}$$

and the adversary's optimal stopping time $\hat{\tau} \in \mathcal{T}$ is:

$$\hat{\tau} = \inf\{s > 0; X(s) \notin D|s \in [0,T]\}, \tag{27}$$

where$D$ is the adversary's continuation region (c.f. (9)) and ψ is given in (24).

The probabilistic interpretation of (24) yields a Monte-carlo method for estimating the VF by empirical means: $\psi(x) \approx \psi^N(x) := \frac{1}{N}\sum_{i=1}^N h(x + W^i_{T-t})$ where $h(x) := \int_0^\tau f(s,x,u_0)ds + G(x_\tau)$.

To implement the control in a number of settings, it is necessary to formulate the control in discrete-time. To this end, we deduce the following result:

Lemma 3. The optimal control policy, $\hat{u} \in \mathcal{U}$ has a discrete-time representation given by:

$$\hat{u}(s_j, x) = -\lim_{ds\to 0}R^{-1}G(s_j,x)^T\int \alpha(s_i,x)(\nabla_x S(s_j,x) + S(s_j,x)\nabla_x Z(s_j,x))d\tau_i,$$

where

$$\alpha(s_i,x) := \frac{1}{D(\tau_i)}\exp(Z(s_i,x)), \quad S(x) := \chi(T, X_T^{t,x_0,\hat{u}}) + \sum_{j=1}^{N-1} f(s_i,x)ds,$$

$$Z(x) := -\frac{1}{2}\sum_{j=1}^{N-1} \|(x_{s_{j+1}} - x_{s_j})/ds - \mu(s_j,x)\|_\sigma^2, \quad D(s_i) := (2\pi ds)^{\frac{1}{2}(N-i)l}.$$

We now describe an SP in which the goal is to find an optimal stopping criterion under worst-case scenarios.

6.1 Optimal Stopping in Worst-Case Scenarios

SPs involve finding an optimal stopping criterion in a dynamic system given some state dependent reward. SPs are ubiquitous in economics, finance and statistical hypotheses testing. In many instances with concern for safety or financial security, it is necessary to consider SPs under worst-case scenarios.

We tackle an important variant of the problem in which the agent seeks to find the optimal stopping criterion under worst-case conditions. To describe worst case scenarios, we perform a change of measure.

The problem involves an agent that seeks to finding an optimal stopping time $\hat{\tau}$ under the adverse non-linear expectation $\mathscr{E}_P[\cdot] := -|\rho|^{-1}\inf_{\mathbb{Q}\in\mathscr{M}_a} \log \mathbb{E}_{\mathbb{Q}}[\cdot]$ s.th.

$$\hat{\tau} \in \arg\sup_{\tau\in\mathscr{T}} \mathscr{E}_{\mathbb{P}}\left[Y_\tau\right] = \arg\sup_{\tau\in\mathscr{T}} \left(-|\rho|^{-1}\inf_{\mathbb{Q}\in\mathscr{M}_a} \log \mathbb{E}_{\mathbb{Q}}\left[Y_\tau\right]\right) \tag{28}$$

where $\mathscr{M}_a$ is a family of measures equivalent[3] to $\mathbb{P}$, the reference measure under (30), $\rho \in \mathbb{R}_{>0}$ and $Y_\tau = \exp\left\{-|\rho|\left(\int_t^{\tau\wedge T} f(s,u_s,x_s)ds + e^{-\gamma(\tau\wedge T)}K(X_{\tau\wedge T})\right)\right\}$.

Note that the risk sensitive minimisation (28) generalises the risk neutral minimisation since by a version of the Laplace-Varadhan lemma [11] we obtain:

$$\lim_{\rho\to 0} \mathscr{E}_{\mathbb{P}}\left[Y_\tau\right] = \inf_{\mathbb{Q}\in\mathscr{M}_a} \mathbb{E}_{\mathbb{Q}}\left[\bar{Y}_\tau\right] \tag{29}$$

where $\bar{Y}_\tau := \int_t^{\tau\wedge T} f(s,X_s,u_s)ds + e^{-\gamma(\tau\wedge T)}K(X_{\tau\wedge T})$.

The uncontrolled system dynamics are given by:

$$dX_s = \mu(s,X_s)ds + \sigma dB(s,X_s), \tag{30}$$

where the drift μ is as described previously and $\sigma \in \mathbb{R}$.

The problem describes an agent that seeks to find an optimal stopping time $\tau \in \mathscr{T}$ under a worst-case scenario. We now illustrate the SP within an example in a financial setting which is adapted from the discrete example in [25].

[3] The measure $\mathbb{P}$ is said to be equivalent (denoted by $\mathbb{Q} \ll \mathbb{P}$) to $\mathbb{Q}$ if whenever the measure is 0 on $\mathbb{Q}$ it is also 0 on $\mathbb{P}$.

6.2 Example: Optimal investing in financial markets

An investor (I) seeks to exit the market (sell all market holdings) at an optimal stopping time $\tau \in \mathcal{T}$. It is assumed that the market acts in such a way to minimise risk-free profit opportunities for the investor.[4] When I exits the market, I receives a return of $\lambda^\tau X_\tau$ where $X_t \equiv X(t,\omega) \in [0,\infty[\times\Omega$ is a Markov process that determines the asset price at time t and $\lambda \in]0,1]$ is I's discount factor. Classically, the exit time is computed as the solution to the following problem:

$$\sup_{\tau\in\mathcal{T}} \mathbb{E}_{\mathbb{P}}\left[e^{-\gamma\tau}X_\tau\right]. \tag{31}$$

In (31), the expectation is taken w.r.t. a risk-neutral measure $\mathbb{P}$ and hence, neglects the adversarial effect of the market. To accommodate market effects on investment opportunities, the objective is modified to the following:

$$\sup_{\tau\in\mathcal{T}}\left(-|\rho|\inf_{\mathbb{Q}}\log\mathbb{E}_{\mathbb{Q}}\left[\exp\{-|\rho|e^{-\gamma\tau}X_\tau\}\right]\right);\ \rho\in\mathbb{R}_{>0}. \tag{32}$$

In this problem, the agent finds an optimal time to exit a financial market under an adversarial market scenario.

6.3 Solving the SP

Having characterised the optimal controller (and adversary) behaviour for the game of control and stopping, we return to the SP under worst-case scenarios and show that the solution can be recovered by solving the game.

The following theorem shows that the solution to the SP is given by the stopping time $\hat{u}$ of the SDG (2):

Theorem 3. Let $\hat{\tau} \in \mathcal{T}$ be the equilibrium pair in Theorem 2, then $\hat{\tau} \in \mathcal{T}$ is a solution to the worst-case SP.

The theorem is proven by establishing an equivalence of the two problems. In particular, the proof of the theorem works by demonstrating that the objective function of the game of control and stopping corresponds to the objective of the SP and secondly, showing an equivalence between the (optimal) VF for the game of control and stopping and the (optimal) VF for the SP.

7 Conclusion

We constructed a novel method for generating fault-tolerant control policies. The framework produces policies that are robust against random system faults that can lead to catastrophic outcomes. We showed that the method tackles optimal stopping under worst case scenarios. The continuous (in action and state spaces and, time) yields a solution suitable for numerous problems within finance, robotics and physical control. For both problems, we showed that the each solution can be approximated by forward sampling of a diffusion process.

[4] This is the no arbitrage principle [5].

Supplementary Material

Assumptions

The results of the paper are built under the following additional assumptions:

The functions $\mu, K, \hat{\mu}, \hat{\sigma}, h$ are deterministic, measurable and Lipschitz continuous that satisfy a polynomial growth condition. These conditions ensure the existence of (5) [16].

In particular, for $R \in \{\mu, K, \hat{\mu}, \hat{\sigma}, h\}$, we assume there exist real-valued constants $c_R > 0$ s.th. $\forall s \in [t, \tau_S], \forall x, y \in \mathbb{R}^p$ we have:

$$|R(s,x) - R(s,y)| \leq c_R |x - y| .$$

We assume the functions $\mu, K, \hat{\mu}, \hat{\sigma}, h$ satisfy a polynomial growth condition, that is for $R \in \{\mu, K, \hat{\mu}, \hat{\sigma}, h\}$, we assume that there exist real-valued constants $d_R > 0$ s.th. $\forall (s,x) \in [t, \tau_S] \times \mathbb{R}^p$ we have:

$$|R(s,x)| \leq d_R \left(|1 + |x|^{\rho}\right)$$

Poisson Stochastic Calculus

As in [34], we state some important results of Poisson stochastic calculus which we use (a detailed treatment of the following results can be found in [14]):

$$\mathbb{E}[dP_i(s)] = \rho_i ds, \tag{33}$$

$$\text{Var}[dP_i(s)] = \rho_i ds, \tag{34}$$

where $\rho_i(s) > 0$ is the i^{th} jump rate or jump density and $\rho_i ds$ is the mean count of the i^{th} Poisson process in the time interval $(t, t + ds]$. We also have the following results:

$$\text{Cov}[dP_i(s_j) dP_i(s_k)] = \text{Var}[dP_i(s_j)] \delta_{k,j} = \rho_i(s_j) ds \delta_{k,j}, \tag{35}$$

where $\delta_{k,j}$ is the kroneck-delta function. Moreover, let r and m be continuous parameters then we have:

$$\text{Cov}[dP_i(r) dP_i(m)] = \rho_i(m_j) dm \delta(m - r) dr. \tag{36}$$

For the Poisson differential vector dP, we have $\text{Var}[dP] = \text{diag}(\rho_1, \ldots, \rho_m)$ and for non-independent Poisson increments we have $\text{Var}[dP] = \sigma_p dt$. As in [34], we note also that since both the processes P_i and dP_i are Poisson distributed, we can clearly write the following expressions:

$$\text{Prob}(P_i(s) = k) = \exp(-\nu_i)(\nu_i^k)/k!, \tag{37}$$

and similarly,

$$\text{Prob}(dP_i(s) = k) = \exp(-\rho_i)(\rho_i ds)^k/k!, \tag{38}$$

By the zero-one-law, for the calculation of the probability in (38), for the jumps in (38) we can write the following:

$$\text{Prob}(dP_i(s) = k) = (1 - \rho_i(s) ds)\delta_{k,0} + \rho_i(s) ds \delta_{k,1} \tag{39}$$

Proof of results

Proof of Lemma 1. Remark 1. Note that Theorem 1 implies that on D:

$$\inf_{u\in\mathcal{U}}\left[\partial_s\phi(s,X_s)+\mathscr{L}^{\hat{u}}\phi(s,X_s)+f(s,X_s,\hat{u})\right]=0,\qquad \forall X\in\mathbb{R}^p \tag{40}$$

Applying Itō's lemma to $\phi(x)$, inserting the expression for the running cost function f c.f. (2) and (6) and, by Theorem 1, we readily compute that:

$$\begin{aligned}0&=\inf_{u\in\mathcal{U}} f(s,X_s,u)+\partial_s\phi(s,X_s)+\langle\nabla_x\phi(s,X_s),\hat{\mu}(s,X_s,u)\rangle\\&\quad+\frac{1}{2}\mathrm{tr}(\nabla^2_{xx}\phi(s,X_s)\sigma(s,X_s)[\sigma(s,X_s)]^T)+D_{\mathrm{jump}}\phi(s,X_s)\\&=f(s,X^{t,x,\hat{u}}_s)+\frac{1}{2}\hat{u}^TR_u\hat{u}+\partial_s\phi(s,X_s)+(\nabla_x\phi(s,X_s))^T(\mu(s,X_s)\\&\quad+G(s,X_s)\hat{u}+\frac{1}{2}\mathrm{tr}(\nabla^2_{xx}\phi(s,X_s)\cdot\sigma(s,X_s)[\sigma(s,X_s)]^T)+D_{\mathrm{jump}}\phi(s,X_s)),\end{aligned} \tag{41}$$

from which we readily compute that the optimal control $\hat{u}\in\mathcal{U}$ for player 1 is given by:

$$\hat{u}(s,x)=-R_u^{-1}[G(s,x)]^T\nabla_x\phi(s,x), \tag{42}$$

Reinserting (42) back into (41) we arrive at the following:

$$-\frac{\partial\phi}{\partial s}(x)-(\mathscr{A}\phi(x)+f(x))=0, \tag{43}$$

where $\mathscr{A}$ is given by:

$$\begin{aligned}&\mathscr{A}\phi(x)\\&:=\mu(s,x)\nabla_x\phi(x)+\sum_{i,j=1}^{d}(\sigma\sigma^T)_{ij}\nabla_{xx}\phi(x)+\nabla_x\phi(x)G(s,x)R^{-1}G(s,x)\nabla_x\phi(x)\\&\qquad\qquad+D_{\mathrm{jump}}\phi(x)\end{aligned} \tag{44}$$

□

Proof of Theorem 2. The proof of the theorem follows immediately from (42) and the definition of the continuation region D. Indeed, whenever the process X exits D it is optimal for player 2 to terminate the game — the remainder of the theorem then follows from applying the definition of D. □

Proof of Lemma 3. Since the stochastic part of the process is Gaussian distributed with variance σ_{s_j} the transition probability is given by the following expression:

$$\mathrm{Prob}(X_{s_{j+1}}|X_{s_j})=\frac{1}{((2\pi)^l\cdot|\sigma_{s_j}|)^{\frac{1}{2}}}\exp\left(-\frac{1}{2}\|X_{s_{j+1}}-X_{s_j}-\mu(s_j,X_{s_j})ds\|^2_{\sigma^{-1}_{s_j}}\right)\quad,$$

where $\|\cdot\|^2$ is the Mahalanobis norm defined by $\|v\|_M^2 := v^T M v$ for some $v \in GL(p, R)$ and $M \in S_p$. After performing some standard steps (see [33, 36]), we can rewrite the PI (31) in our case as:

$$\begin{aligned}\psi(x) = \lim_{ds\to 0} \int \frac{1}{\Pi_{j=1}^{N-1}(2\pi)^l \cdot |\sigma_{s_j}|)^{\frac{1}{2}}} \left(\sum_{j=1}^{N-1} f(s_j, X_{s_j} + K(X_\tau^{t,x_0,\hat{u}})\right) \\ \cdot \exp\left(-\frac{1}{2}\sum_{j=1}^{N-1} \left\|X_{s_{j+1}} - X_{s_j} - \mu(s_j, X_{s_j})dt\right\|^2_{\sigma_{s_j}^{-1}}\right) ds_j \\ = \lim_{ds\to 0} \int \frac{1}{D(\tau)} S(s_j, X_{s_j}) \exp\left(Z(s_j, X_{s_j})\right) ds_i, \end{aligned} \tag{45}$$

where Z, S and D are defined by:

$$\begin{aligned} Z(x) &:= -\frac{1}{2}\sum_{j=1}^{N-1} \left\|X_{s_{j+1}} - X_{s_j} - \mu(s_j, X_{s_j})ds\right\|^2_{\sigma_{s_j}^{-1}}, \\ S(x) &= \sum_{j=1}^{N-1} f(s_j, X_{s_j}) + K(X_\tau^{t,x_0,\hat{u}}) \end{aligned}$$

and

$$D(\tau) := \Pi_{j=1}^{N-1}((2\pi)^l \cdot |\sigma_{s_j}|)^{\frac{1}{2}}. \tag{46}$$

Inserting (45) into (42), the expression for the optimal control $\hat{u}$, we find that:

$$\hat{u}(s_j, x) = -\lim_{ds\to 0} R^{-1} G(s_j, x)^T \nabla_x \left(\int \frac{1}{D(s_i)} S(s_i, x) \exp(Z(s_i, x))\right). \tag{47}$$

If we now suppose that the integrand of (47) is continuously differentiable in X_{s_j}, then we can readily compute that:

$$\begin{aligned}\hat{u}(s_j, x) &= -\lim_{ds\to 0} R^{-1} G(s_j, x)^T \cdot \nabla_x \left(\int \frac{1}{D(\tau)} S(s_i, x) \exp(Z(s_i, x)) ds_i\right) \\ &= -R^{-1} \lim_{ds\to 0} G(s_j, x)^T \\ &\quad \cdot \left(\frac{1}{D(\tau)} \exp(Z(s_i, x))(\nabla_x S(s_i, x) + S(s_i, x)\nabla_x Z(s_i, x)) ds_i\right) \end{aligned}$$

which is the required result. □

Proof of Theorem 3. To prove an equivalence of the two problems, we must prove two facts: i) the objective function of the game of control and stopping corresponds to the objective of the SP where the adversary chooses the measure $\mathbb{Q}$. ii) an equivalence between the (optimal) VF for the game of control and stopping and the (optimal) VF for the SP.

Our first task is to construct an equivalent measure $\mathbb{Q}$ which is selected by the adversary. Using Girsanov's theorem (see [27]), for the passive dynamics in (30), we have the following relation for $\mathbb{Q}$:

$$\frac{d\mathbb{Q}}{d\mathbb{P}} = e^{-\xi(u)} \tag{48}$$

where the quantity in (48) is the Radon-Nikodym derivative and $\xi(u)$ is:

$$\xi(u) = \frac{1}{2}\sigma \int_0^\tau u_s^T u_s ds + \sqrt{\sigma} \int_0^\tau u^T dB_s \tag{49}$$

where $u \in \mathcal{U}$ is as in Sec.3.1 and $\mathbb{P}$ is the reference measure for the passive dynamics in (4). With the transformation (48) the (controlled) dynamics is:

$$dX_s = \mu(s, X_s)ds + \sigma dB_s, \tag{50}$$

We now show that the performance function for the SP reproduces the objective function of the game of control stopping (c.f. (3) and (6)), indeed:

$$\begin{aligned}
&\log\left\{\mathbb{E}_{\mathbb{Q}}\left[\exp\left(Y_\tau\right)\right]\right\} = \log\left\{\mathbb{E}_{\mathbb{P}}\left[\exp\left(Y_\tau\right)\frac{d\mathbb{Q}}{d\mathbb{P}}\right]\right\} \\
&\geq \mathbb{E}_{\mathbb{P}}\left[\log\left(\exp\left(Y_\tau\right)\frac{d\mathbb{Q}}{d\mathbb{P}}\right)\right] \\
&= \mathbb{E}_{\mathbb{P}}\left[\log\left(\exp\left(Y_\tau\right)\right) + \log\left(\frac{d\mathbb{Q}}{d\mathbb{P}}\right)\right] \\
&= \mathbb{E}_{\mathbb{P}}\left[Y_\tau + \xi(u)\right] = \mathbb{E}_{\mathbb{P}}\left[Y_\tau - \frac{1}{2}\sigma\int_0^\tau u_s^T u_s ds\right],
\end{aligned} \tag{51}$$

where we have used that $\mathbb{E}\left[u^T\sigma_i(s,x)dB_s\right] = 0$ by the standard properties of Brownian motion. At optimum, when $u = \hat{u}$ we have equality in (51).

This proves the equivalence between the two objectives (now $R \equiv \sigma 1_d$). It remains to prove that the VF of the game is equivalent for the VF of the SP:

$$\sup_{\tau\in\mathcal{T}}\left(\inf_{\mathbb{Q}\in\mathcal{M}_a}\mathbb{E}_{\mathbb{Q}}\left[Y_\tau\right]\right) = \inf_{u\in\mathcal{U}}\left(\sup_{\tau\in\mathcal{T}}\mathbb{E}\left[Y_\tau^u\right]\right). \tag{52}$$

The result follows directly from the existence of a value of the game of control and stopping and the Girsanov theorem. To see this we note by Theorem 1:

$$\inf_{\tau\in\mathcal{T}}\left(\sup_{u\in\mathcal{U}}\mathbb{E}\left[\bar{Y}_\tau^u\right]\right) = \sup_{u\in\mathcal{U}}\left(\inf_{\tau\in\mathcal{T}}\mathbb{E}\left[\bar{Y}_\tau^u\right]\right), \tag{53}$$

where $\mathbb{E}$ is taken under the controlled diffusion (5). Lastly, by Girsanov's theorem (c.f. (48)), we have that $\sup_{u\in\mathcal{U}}\mathbb{E}\left[\bar{Y}_\tau^u\right] = \sup_{\mathbb{Q}}\mathbb{E}_{\mathbb{Q}}\left[\bar{Y}_\tau\right]$; setting $\bar{Y}_{(\cdot)} \equiv -Y_{(\cdot)}$ then reveals the equivalence, hence

$$\arg\inf_{\tau\in\mathcal{T}}\left(\sup_{u\in\mathcal{U}}\mathbb{E}\left[Y_\tau^u\right]\right) = \arg\inf_{\tau\in\mathcal{T}}\left(\mathbb{E}\left[Y_\tau^{\hat{u}}\right]\right) = \hat{\tau}, \tag{54}$$

which yields the required result. □

Bibliography

[1] Anderson, B.D., Moore, J.B.: Optimal control: linear quadratic methods. Courier Corporation (2007)

[2] Baghery, F., Haadem, S., Øksendal, B., Turpin, I.: Optimal stopping and stochastic control differential games for jump diffusions. Stochastics An International Journal of Probability and Stochastic Processes 85(1), 85–97 (2013)

[3] Bardi, M., Raghavan, T., Parthasarathy, T.: Stochastic and differential games: theory and numerical methods, vol. 4. Springer Science & Business Media (2012)

[4] Bayraktar, E., Hu, X., Young, V.R.: Minimizing the probability of lifetime ruin under stochastic volatility. Insurance: Mathematics and Economics 49(2), 194–206 (2011)

[5] Carr, P., Madan, D.B.: A note on sufficient conditions for no arbitrage. Finance Research Letters 2(3), 125–130 (2005)

[6] Chancelier, J.P., Øksendal, B., Sulem, A.: Combined stochastic control and optimal stopping, and application to numerical approximation of combined stochastic and impulse control. Preprint series: Pure mathematics http://urn. nb. no/URN: NBN: no-8076 (2000)

[7] Dragulescu, A.A., Yakovenko, V.M.: Probability distribution of returns in the heston model with stochastic volatility. Quantitative finance 2(6), 443–453 (2002)

[8] Dynkin, E.: Game variant of a problem on optimal stopping. In: Soviet Math. Dokl. vol. 10, pp. 270–274 (1967)

[9] Efimov, D., Cieslak, J., Zolghadri, A., Henry, D.: Actuator fault detection in aircraft systems: Oscillatory failure case study. Annual Reviews in Control 37(1), 180–190 (2013)

[10] Fleming, W.H., Souganidis, P.E.: On the existence of value functions of two-player, zero-sum stochastic differential games. Indiana University Mathematics Journal 38(2), 293–314 (1989)

[11] Fleming, W.H., et al.: Risk sensitive stochastic control and differential games. Communications in Information & systems 6(3), 161–177 (2006)

[12] Friedman, A.: Differential games. Courier Corporation (2013)

[13] Hagan, P.S., Kumar, D., Lesniewski, A.S., Woodward, D.E.: Managing smile risk. The Best of Wilmott 1, 249–296 (2002)

[14] Hanson, F.B.: Applied stochastic processes and control for Jump-diffusions: modeling, analysis, and computation, vol. 13. Siam (2007)

[15] Hill, T.P.: On the existence of good markov strategies. Transactions of the American Mathematical Society 247, 157–176 (1979)

[16] Ikeda, N., Watanabe, S.: Stochastic differential equations and diffusion processes, vol. 24. Elsevier (2014)

[17] Jegadeeshwaran, R., Sugumaran, V.: Fault diagnosis of automobile hydraulic brake system using statistical features and support vector machines. Mechanical Systems and Signal Processing 52, 436–446 (2015)
[18] Jiang, C., Sheng, Z.: Case-based reinforcement learning for dynamic inventory control in a multi-agent supply-chain system. Expert Systems with Applications 36(3), 6520–6526 (2009)
[19] Kappen, H.J.: Linear theory for control of nonlinear stochastic systems. Physical review letters 95(20), 200201 (2005)
[20] Karatzas, I., Sudderth, W.: Stochastic games of control and stopping for a linear diffusion. In: Random Walk, Sequential Analysis And Related Topics: A Festschrift in Honor of Yuan-Shih Chow, pp. 100–117. World Scientific (2006)
[21] Koppejan, R., Whiteson, S.: Neuroevolutionary reinforcement learning for generalized helicopter control. In: Proceedings of the 11th Annual conference on Genetic and evolutionary computation. pp. 145–152. ACM (2009)
[22] Mahadevan, S., Marchalleck, N., Das, T.K., Gosavi, A.: Self-improving factory simulation using continuous-time average-reward reinforcement learning. In: Machine Learning-International Workshop then Conference-. pp. 202–210. Morgan Kaufmann Publishers, inc. (1997)
[23] Mguni, D.: Optimal capital injections with the risk of ruin: A stochastic differential game of impulse control and stopping approach. arXiv preprint arXiv:1805.01578 (2018)
[24] Mguni, D.: A viscosity approach to stochastic differential games of control and stopping involving impulsive control. arXiv preprint arXiv:1803.11432 (2018)
[25] Mguni, D.: Cutting your losses: Learning fault-tolerant control and optimal stopping under adverse risk. arXiv preprint arXiv:1902.05045 (2019)
[26] Munos, R., Bourgine, P.: Reinforcement learning for continuous stochastic control problems. In: Advances in neural information processing systems. pp. 1029–1035 (1998)
[27] Øksendal, B.: Stochastic differential equations. In: Stochastic differential equations, pp. 65–84. Springer (2003)
[28] Øksendal, B.K., Sulem, A.: Applied stochastic control of jump diffusions, vol. 498. Springer (2005)
[29] Peng, S., Wang, F.: Bsde, path-dependent pde and nonlinear feynman-kac formula. Science China Mathematics 59(1), 19–36 (2016)
[30] Peskir, G., Shiryaev, A.: Optimal stopping and free-boundary problems. Springer (2006)
[31] Pham, H.: Feynman-kac representation of fully nonlinear pdes and applications. Acta Mathematica Vietnamica 40(2), 255–269 (2015)
[32] Thapa, D., Jung, I.S., Wang, G.N.: Agent based decision support system using reinforcement learning under emergency circumstances. In: International Conference on Natural Computation. pp. 888–892. Springer (2005)
[33] Theodorou, E., Buchli, J., Schaal, S.: A generalized path integral control approach to reinforcement learning. Journal of Machine Learning Research 11(Nov), 3137–3181 (2010)

[34] Theodorou, E.A., Todorov, E.: Stochastic optimal control for nonlinear markov jump diffusion processes. In: American Control Conference (ACC), 2012. pp. 1633–1639. IEEE (2012)
[35] Tsitsiklis, J.N., Van Roy, B.: Optimal stopping of markov processes: Hilbert space theory, approximation algorithms, and an application to pricing high-dimensional financial derivatives. IEEE Transactions on Automatic Control 44(10), 1840–1851 (1999)
[36] Wio, H.: Application of path integration to stochastic processes: an introduction. Fundamentals and Applications of Complex Systems, Nueva Ed. Univ., UN San Luis 253 (1999)