Cutting Your Losses: Learning Fault-Tolerant Control and Optimal Stopping under Adverse Risk

David Mguni[1]

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

Abstract. Recently, there has been a surge in interest in safe and robust techniques within reinforcement learning (RL). Current notions of risk in RL fail to capture the potential for systemic failures such as abrupt stoppages from system failures or surpassing of safety thresholds and the appropriate responsive controls in such instances. We propose a novel approach to risk minimisation within RL in which, in addition to taking actions that maximise its expected return, the controller learns a policy that is robust against stoppages due to an adverse event such as an abrupt failure. The results of the paper cover fault-tolerant control in worst-case scenarios under random stopping and optimal stopping, all in unknown environments. By demonstrating that the class of problems is represented by a variant of stochastic games, we prove the existence of a solution which is a unique fixed point equilibrium of the game and characterise the optimal controller behaviour. We then introduce a value function approximation algorithm that converges to the solution through simulation in unknown environments.

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

1 Introduction

A significant amount of focus within reinforcement learning (RL) is now being placed on safe execution, robust control, fault-tolerance and risk-minimisation (Garcıa and Fernández, 2015). Driving this interest is an increase in application of RL in real-world environments and industrial applications such as traffic light control (Arel et al., 2010), robotics (Deisenroth et al., 2013), autonomous vehicles (Shalev-Shwartz et al., 2016) and healthcare (Gottesman et al., 2019). Applying RL in various environments requires safe operation of autonomous agents is ensured. At present however, such frameworks within RL are restricted to models in which the agent modifies the state process using an expectation measure which is altered to accommodate a predefined notion of risk e.g. H_∞ control (Morimoto and Doya, 2001). Other notions of risk include coherent risk, conditional value at risk (CVar) (Tamar et al., 2015).

There are numerous instances in which controllers are required to act in systems which suffer the potential for random stoppages or failures that produce catastrophic outcomes (Garcıa and Fernández, 2015). Examples include, in finance, optimal trading under random counterparty risk (Jiao and Pham, 2011), and in control systems, optimal robotic control with random sensor failure and helicopter control under engine failure (Abbeel et al., 2010). Consequently, when using RL in environments that present potential issues of safety, the important question of how to control the system in a way that is robust against faults that lead to catastrophic events arises. An additional issue in matters of safety is when to optimally stop the system with concern for risk of adverse events (e.g. when to sell all asset holdings with concern for financial ruin). Despite its importance however, current notions of risk do not offer a method of mitigating risk by selectively stopping the system.

To this end, we for the first time construct a method that enables an RL controller to determine an optimal sequence of actions that is robust against failures that lead to adverse events. In order to find the optimal control policy, it is necessary to determine a stopping criterion that stops the system which produces a worst-case scenario. Secondly, we construct a method that enables an RL agent to determine when to stop the system in order to maximise its expected payoff in the presence of adverse risk. As we show, each problem admits a two-player stochastic game (SG) representation in which one of the players is delegated the task of modifying the system dynamics through its actions and the other player has the task of stopping the system under an adversarial criterion.

We perform a formal analysis of an SG between a 'controller' and a 'stopper'. Under this interpretation, the outcome is determined by a controller that affects the state process through its actions whilst playing against an adversary that has the right to choose when to stop the game. This produces a framework that finds an optimal sequence of actions that is robust against stoppages at times that pose adverse risk. The notion of risk is defined in the worst-case scenario sense — given the complete set of probability distributions, the agent considers the worst-case in assessing the expected payoff.

These results tackle optimal stopping problems (OSPs) under worst-case scenarios. OSPs are a subclass 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 (Peskir and Shiryaev, 2006). Despite the fundamental relevance of risk in RL, current iterative methods in OSPs in unknown environments are restricted to risk-neutral settings (Tsitsiklis and Van Roy, 1999) and do not permit the inclusion of a controller that modifies the dynamics. Introducing a notion of risk (generated adversarially) adds considerable difficulty as the solution concept is now an SG saddle point equilibrium, the existence of which must be established.

As we show, our framework provides an iterative method of solving worst-case scenario OSPs in unknown environments. The framework is developed through a series of theoretical results: first, we establish the existence of a value of the game which characterises the payoff for the (saddle point) equilibrium. Second, we prove a contraction mapping property of a Bellman operator of the game and that the value is a unique fixed point of the operator. Third, we prove the existence and characterise the optimal stopping time. We then prove an equivalence between the game of control and stopping and worst-case OSPs and show that the fixed point solution of the game solves the OSP.

Finally, using an approximate dynamic programming method, we develop a simulation-based iterative scheme that computes the optimal controls. The method applies in settings in which neither the system dynamics nor the reward function are known. Hence, the agent need only observe its realised rewards by interacting with the environment.

1.1 Related Work

The coverage of FT within RL is extremely limited. In (Zhang and Gao, 2018) RL is applied to tackle systems in which faults might occur with the occurrence of a fault incurring a large cost. Similarly, RL is applied to a problem in (Yasuda et al., 2006) in which an RL method for Bayesian discrimination which is used to segment the state and action spaces. Unlike these methods in which infrequent faults from the environment generate negative feedback, our method uses a game-theoretic framework to simulate faults leading to an FT trained policy.

Our main results are centered around a minimax proof that establishes the existence of a value of the game. This is necessary for simulating the stopping action to induce fault-tolerance. Although minimax proofs are well-known in game theory (Shapley, 1953; Maitra and Parthasarathy, 1970; Filar et al., 1991), replacing a player's action set with stopping times necessitates a minimax proof (which now relies on a construction of open sets) which markedly differs to the standard methods within game theory. Additionally, crucial to our analysis is the characterisation of the adversary optimal stopping time (Theorem 3).

A relevant framework is a two-player optimal stopping game (Dynkin game) in which each player chooses one of two actions; to stop the game or continue (Dynkin, 1967). Dynkin games have generated a vast literature since the setting requires a markedly different analysis from standard stochastic game theory. In the case with one stopper and one controller such as we are concerned with, the minimax proof requires a novel construction using open sets to cope with the stopping problem for the minimax result.

Presently, the study of optimal control that combines control and stopping is limited to a few studies e.g. (Chancelier et al., 2002). Similarly, games of control and stopping have been analysed in continuous-time in specific contexts e.g. linear diffusions (Karatzas and Sudderth, 2006), geometric Brownian motion (Bayraktar et al., 2011) and jump-diffusions (Baghery et al., 2013; Mguni, 2018). In these analyses, all aspects of the environment are known and the controller affects the dynamics of a continuous diffusion process. In general, under these methods, solving these problems requires computing analytic solutions to non-linear partial differential equations which are typically insoluble.

There is a plethora of work on OSPs in continuous and discrete-time (Peskir and Shiryaev, 2006). Tsitsiklis and Van Roy (1999) use approximate dynamic programming methods to construct an iterative scheme to compute the solution of an OSP. Our results generalise existing analyses to strategic settings with both a controller and an adversarial stopper which tackles risk within OSPs.

1.2 Organisation

The paper is organised as follows: in Sec. 2, we introduce some relevant mathematical preliminaries and give a canonical description of both the fault-tolerant RL problem and the OSP under worst-case scenarios. In Sec. 3, we provide illustrative examples for each problem within the context of finance and RL. In Sec. 4, we introduce the underlying SG framework which we use within the main theoretical analysis which we perform in Sec. 5. Lastly in Sec. 6, we develop an approximate dynamic programming approach that enables the optimal controls to be computed through simulation, followed by some concluding remarks.

2 Canonical Description

In this setting, the state of the system is determined by a stochastic process $\{s_t|t = 0, 1, 2, \ldots\}$ whose values are drawn from a state space $\mathcal{S} \subseteq \mathbb{R}^p$ for some $p \in \mathbb{N}$. The state space is defined on a probability space $(\Omega, \mathcal{B}, P)$, where Ω is the sample space, $\mathcal{B}$ is the set of events and P is a map from events to probabilities. We denote by $\mathcal{F} = (\mathcal{F}_n)_{n\geq 0}$ the filtration over $(\Omega, \mathcal{B}, P)$ which is an increasing family of $\sigma-$algebras generated by the random variables $s_1, s_2, \ldots$.

We operate in a Hilbert space $\mathcal{V}$ of real-valued functions on $\mathbb{L}_2$, i.e. a complete[1] vector space which we equip with a norm $\|\cdot\| : \mathcal{V} \to \mathbb{R}_{>0} \times \{0\}$ given by $\|f\|_\mu := \sqrt{\mathbb{E}_\mu[f^2(s)]}$ and its inner product $\langle f, f^T\rangle_\mu := \mathbb{E}_\mu\left[f(s)f^T(s)\right]$ where $\mu : \mathcal{B}(\mathbb{R}^n) \to [0, 1]$ is a probability measure. The problem occurs over a time interval $[0, K]$ where $K \in \mathbb{N} \times \{\infty\}$ is the time horizon. A stopping time is defined as a random variable $\tau \in \{0, 1, 2, \ldots\}$ for which $\{\omega \in \Omega|\tau(\omega) \leq t\} \in \mathcal{F}_t$ for any $t \in [0, K]$ — this says that given the information generated by the state process, we can determine if the stopping criterion has occurred.

We now describe the two problems with which we are concerned that is, FT RL and OSPs under worst-case scenarios. We later prove an equivalence between the two problems and characterise the solution of each problem.

2.1 Fault-Tolerant Reinforcement Learning

We concern ourselves with finding control policy that copes with abrupt system stoppages and failures at the worst times in problems. In the current setting, the reward and transition functions are assumed a priori unknown. Unlike standard methods in RL and game theory that have fixed time horizons (or purely random exit times) in the following, the process is stopped by a fictitious adversary that uses a stopping strategy to decide when to stop given its observations of the state. In order to generate an FT control, we simulate the adversary's action whilst the controller seeks to determine its optimal policy. This as we show, induces a form of control that is an FT best-response control.

A formal description is as follows: an agent exercises actions that influence the sequence of states visited by the system. At each state, the agent receives a

[1] A vector space is complete if it contains the limit points of all its Cauchy sequences.

reward which is dependent on the state. The agent's actions are selected by a policy $\pi : \mathcal{S} \times \mathcal{A} \to [0,1]$ — a map from the set of states $\mathcal{S}$ and the set of actions $\mathcal{A}$ to a probability. We assume that the action set is a discrete compact set and that the agent's policy π is drawn from a compact policy set Π. The horizon of the problem is at most T but the process may be terminated earlier at some ($\mathcal{F}-$measurable) stopping time at which point the agent receives a terminal reward.

The agent's performance function is given by:

$$J^{k,\pi}[s] = \mathbb{E}\left[\sum_{t=0}^{k \wedge T} \gamma^t R(s_t, a_t) + \gamma^{k \wedge T} G(s_{k \wedge T}) \middle| s_0 = s\right], \tag{1}$$

where $a \wedge b := \min\{a, b\}$, $\mathbb{E}$ is taken w.r.t. the transition function P and the controller's policy $\pi \in \Pi$. The performance function (1) consists of a reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ which quantifies the agent's immediate reward when the system transitions from one state to the next, a bequest function $G : \mathcal{S} \to \mathbb{R}$ which quantifies the reward received by the agent when the system is stopped and $\gamma \in [0, 1[$, a discount factor. We assume R and G are bounded and measurable.

The problem we consider is the following:

Find $(\hat{k}, \hat{\pi}) \in \mathcal{T} \times \Pi$ and $J^{\hat{k},\hat{\pi}}[s]$ s.th.

$$\max_{\pi \in \Pi}\left(\min_{k \in \mathcal{T}} J^{k,\pi}[s]\right) = J^{\hat{k},\hat{\pi}}[s], \tag{2}$$

where the minimisation is taken pointwise and $\mathcal{T} \subseteq \{0, 1, 2 \ldots\}$ is a set of ($\mathcal{F}-$measurable) stopping times.

Hereon, we employ the following shorthand $R(s, a) \equiv R_s^a$ for any $s \in \mathcal{S}, a \in \mathcal{A}$.

The dual objective problem in (2) consists of finding both a stopping time that minimises J and an optimal policy that maximises J. By considering the tasks as being delegated to two individual players, the problem becomes an SG. The SG occurs between a 'controller' that seeks to maximise J by manipulating state visitations through its actions and an adversary or 'stopper' that chooses a time to stop the process to minimise J (i.e. at the worst possible time). The structure of the game combines an OSP and a Markov decision process (MDP). We consider a setting in which neither player has up-front knowledge of the transition model or objective function but each only observes their realised rewards.

2.2 Robust Optimal Stopping

The second problem we consider is robust optimal stopping. In OSPs, the goal is to determine a criterion for stopping the system at a time that maximises some state-dependent payoff. OSPs are ubiquitous in finance e.g. for options pricing (Pham, 1997) and in economics for characterising optimal market entry/exit strategies (Kruse and Strack, 2015). OSPs are closely related to multi-armed bandits and clinical trials (Jennison and Turnbull, 2013).

OSPs in worst-case scenarios regularly arise in economic decision-making when an agent seeks to determine an optimal time to exit the financial market

(Young, 2004) or terminate some costly industrial process (Zhao and Chen, 2009) under worst-case scenarios. Examples of worst-case OSPs are agents that seek to determine when to arrest a costly industrial process or experiment (e.g. clinical trials) and, within finance, investors that seek to determine market entry/exit decisions; each under worst-case scenarios.

We later prove an equivalence of SGs of control and stopping and robust OSPs, the latter of which we now introduce:

The problem involves an agent that seeks to find an optimal stopping time $\hat{\tau}$ under the adverse non-linear expectation $\mathscr{E}_P := \min_{\pi \in \Pi} \mathbb{E}_{P,\pi}$ such that:

$$\hat{\tau} \in \arg\max_{\tau \in \mathcal{T}} \mathscr{E}_P \left[Y_\tau \right] = \arg\max_{\tau \in \mathcal{T}} \left(\min_{\pi \in \Pi} \mathbb{E}_{P,\pi} \left[Y_\tau \right] \right), \tag{3}$$

where $Y_k := \sum_{t=0}^{k \wedge T} \gamma^t R(s_t, a_t) + \gamma^{k \wedge T} G(s_{k \wedge T})$.

The problem describes an agent that seeks to find an optimal stopping time under a worst-case scenario.

3 Examples

To elucidate the ideas, we now provide applications of the problems.

As the following example illustrates, the framework applies to actuator failure within RL applications.

3.1 Example: Control with random actuator failure

Consider an adaptive learner, for example a robot that uses a set of actutors to perform actions. Given full operability of its set of actuators, the agent's actions are determined by a policy $\pi : S \times A \to [0,1]$ which maps from the state space S and the set of actions A to a probability. In many systems, there exists some risk of actuator failure at which point the agent thereafter can affect the state transitions by operating only a subset of its actuators. In this instance, the agent's policy determines actions using only a subset of its action space $\hat{A} \subset A$. In this scenario, the agent is now restricted to policies $\pi_{\text{partial}} : S \times \hat{A} \to [0,1]$ which map from from a subset of operative actuators — thereafter its expected return is given by the value function $V^{\pi_{\text{partial}}}$. In order to perform robustly against actuator failure, it is therefore necessary to consider a set of stopping times $\mathcal{T} \subseteq \{0,1,2,\ldots\}$ after which, the robot can no longer select actions that require functionality of the full set of actuators. In particular, in order to construct a robust policy against catastrophic outcomes, it is useful to consider actuator failure in worst-case scenarios.

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

$$\min_{k' \in \mathcal{T}} \left(\max_{\pi' \in \Pi} \mathbb{E} \left[H^{\pi',k'}(s) \right] \right) = \mathbb{E} \left[H^{\hat{\pi},\hat{\tau}}(s) \right],$$

where $a_t \sim \pi'$ and $H^{\pi,k}(s) := \sum_{t=0}^{k\wedge\infty} \gamma^t R(s_t, a_t) + \gamma^{k\wedge\infty} V^{\pi_{\text{partial}}}(s_{k\wedge\infty})$. The resulting policy $\hat{\pi}$ is robust against actuator failure in worst-case scenarios.

3.2 Example: Optimal selling in an adversarial market

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.[2] 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:

$$\max_{k\in\mathcal{T}} \mathbb{E}_P\left[\gamma^k X_k\right]. \tag{4}$$

In (4), the expectation is taken with respect to a risk-neutral measure P. However, the above formulation does not include the adversarial effect of the market. To accommodate this, we modify the objective to the following:

$$\max_{k\in\mathcal{T}}\left(\min_{\pi\in\Pi} \mathbb{E}_{P,\pi}\left[\gamma^k X_k\right]\right). \tag{5}$$

In (5), the worst-case dynamics are induced by choice of adversarial probability measure π that alters the neutral measure P over which the objective expectation is defined. This captures the observed effect that financial markets adversarially eliminate investment opportunities. Now the goal of the agent is to find an optimal time to exit a financial market under an adversarial market scenario.

4 Stochastic games

Embedded within problem (2) is an interdependence between the actions of the players — that is, the solution to the problem is jointly determined by the actions of both players and their responses to each other. The appropriate framework to tackle this problem is therefore an SG (Shapley, 1953). An SG is an augmented MDP which proceeds by two players tacking actions that jointly manipulate the transitions of a system over K rounds which may be infinite. At each round, the players receive some immediate reward or cost which is a function of the players' joint actions. The framework is zero-sum so that a reward for player 1 simultaneously represents a cost for player 2.

Formally, a two-player zero-sum stochastic game is a 6−tuple $\langle \mathcal{S}, \mathcal{A}_{i\in\{1,2\}}, P, R, \gamma\rangle$ where $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$ is a set of states, $\mathcal{A}_i$ is an action set for each player $i \in \{1,2\}$. The map $P : \mathcal{S}\times\mathcal{A}_1\times\mathcal{A}_2\times\mathcal{S} \to [0,1]$ is a Markov transition probability matrix i.e. $P(s'; s, a_1, a_2)$ is the probability of the state s' being the next state given the system is in state s and actions $a_1 \in \mathcal{A}_1$ and $a_2 \in \mathcal{A}_2$ are applied by player 1 and player 2 (resp.). The function $R : \mathcal{S}\times\mathcal{A}_1\times\mathcal{A}_2$

[2] This is the no arbitrage principle (Carr and Madan, 2005).

is the one-step reward for player 1 and represents one-step cost for player 2 when player 1 takes action $a_1 \in \mathscr{A}_1$ and player 2 takes action $a_2 \in \mathscr{A}_2$ and $\gamma \in [0,1[$ is a discount factor. The goal of each player is to maximise its expected cumulative return — since the game is antagonistic, the total expected reward received by player 1 which we denote by J, represents a total expected cost for player 2.

Denote by $H_{t \leq K}$ the set of all finite histories and by $\mathscr{H} \equiv \cup_{t \leq K} H_t$ so that for each $h_{j \leq K} = ((s_0,(a_1,a_2)),(s_1,(a_1,a_2)),\ldots,((s_j,(a_1,a_2)) \in H_j$ which is a sequence of state and joint action pairs. For each player $i \in \{1,2\}$, a pure strategy is a map $\pi_i : (\mathscr{H}) \times \mathscr{A}_i \to [0,1]$ that assigns to every finite history $h_{j \leq K} \in \mathscr{H}$ an action $\pi(h)$ in $\mathscr{A}_i$. Similarly, for each player $i \in \{1,2\}$, a behavioural strategy is a map $\pi_i : \mathscr{H} \times \mathscr{A}_i \to [0,1]$ that assigns to every finite history $h \in \mathscr{H}$ a probability distribution $\pi(h)$ in $\mathscr{A}_i$. We denote the space of strategies for each player $i \in \{1,2\}$ by Π_i.

For SGs with Markovian transition dynamics, we can safely dispense with path dependencies in the space of strategies.[3] Consequently, w.log. we restrict ourselves to the class of behavioural strategies that depend only on the current state and round, namely Markov strategies, hence for each player i, the strategy space Π_i consists of strategies of the form $\pi_i : \mathcal{S} \times \mathscr{A}_i \to [0,1]$. It is well-known that for SGs, an equilibrium exists in Markov strategies even when the opponent can draw from non-Markovian strategies (Hill, 1979).

In SGs, it is usual to consider the case $\mathscr{A}_1 = \mathscr{A}_2$ so that the players' actions are drawn from the same set. We depart from this model and consider a game in which player 2 can choose a time to stop the process so that the action set for player 2 is the set $\mathscr{T} \subseteq \{0,1,2,\ldots\}$ which consists of ($\mathscr{F}-$measurable) stopping times. In this setting, player 1 can manipulate the system dynamics by taking actions drawn from $\mathscr{A}_1$ (we hereon use $\mathscr{A}$) and at each point, player 2 can decide to intervene to stop the game.

Let us define by $\text{val}^+[J] := \min\limits_{k \in \mathscr{T}} \max\limits_{\pi \in \Pi} J^{k,\pi}$ the upper value function and by $\text{val}^-[J] := \max\limits_{\pi \in \Pi} \min\limits_{k \in \mathscr{T}} J^{k,\pi}$, the lower value function. The upper (lower) value function represents the minimum payoff that player I (player II) can guarantee itself irrespective of the actions of the opponent.

The value of the game exists if we can commute the max and min operators:

$$\text{val}^-[J] = \max_{\pi \in \Pi} \min_{k \in \mathscr{T}} J^{k,\pi}[\cdot] = \min_{k \in \mathscr{T}} \max_{\pi \in \Pi} J^{k,\pi}[\cdot] = \text{val}^+[J]. \qquad (6)$$

We denote the value by $J^\star := \text{val}^+[J] = \text{val}^-[J]$ and denote by $(\hat{k},\hat{\pi}) \in \mathscr{T} \times \Pi$ the pair that satisfies $J^{\hat{k},\hat{\pi}} \equiv J^\star$. The value, should it exist, is the minimum payoff each player can guarantee itself under the equilibrium strategy. In general, the functions $\text{val}^+[J]$ and $\text{val}^-[J]$ may not coincide.

Should the value $J^\star$ 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

[3] There are some exceptions for games with payoff structures not considered here for example, limiting average (Ergodic) payoffs (Blackwell and Ferguson, 1968).

zero-sum games. Thus the central task to establish an equilibrium involves unambiguously assigning a value to the game, that is proving the existence of $J^\star$.

5 Main Analysis

In this section, we present the key results and perform the main analysis of the paper. In the main analysis, our first task is to establish the existence of a value of the game then secondly, we perform analyses that enables us to construct an approximate dynamic programming method. In particular, we construct a Bellman operator for the game and show that the operator is a contraction mapping. We show that the value is unique and that the value coincides with a fixed point of the Bellman operator. Using these results, we construct an equivalence between robust OSPs and games of control and stopping.

Our results develop the theory of risk within RL to cover instances in which the agent has concern for stopping the process at an optimal time. We develop the theory of SGs to cover games of control and stopping when neither player has up-front environment knowledge. In particular, we establish the existence of a value of the game in a discrete-time setting and show that the value can be obtained using a value-iterative method. This, as we show in Sec. 9, underpins a simulation-based scheme that is suitable for settings in which the transition model and reward function is a priori unknown.

5.1 Theoretical Analysis

The purpose of this section is to twofold: our first task is to establish the existence of a value of the game. Secondly, we perform analyses that enables us to construct an approximate dynamic programming method. In particular, we construct a Bellman operator for the game and show that the operator is a contraction mapping. We show that the value is unique and that the value coincides with a fixed point of the Bellman operator. Using these results, we construct an equivalence between robust OSPs and games of control and stopping. We defer some of the proofs to the appendix.

We now introduce concepts that relate to estimates on the operators of the game. These concepts will be useful for proving the existence of a fixed point.

Definition 1. An operator $T : \mathscr{V} \to \mathscr{V}$ is said to be a contraction w.r.t a norm $\|\cdot\|$ if there exists a constant $c \in [0,1[$ s.th for any $V_1, V_2 \in \mathscr{V}$ we have that:

$$\|TV_1 - TV_2\| \leq c\|V_1 - V_2\|. \tag{7}$$

A central task is to prove that the Bellman operator for the game is a contraction mapping. Thereafter, we prove convergence to the unique value. Consider a Borel measurable function which is absolutely integrable w.r.t. the transition kernel $P^\cdot$ then $\mathbb{E}\left[J[s'] | \mathscr{F}_t\right] = \int_{\mathcal{S}} J[s'] P^a_{ss'}$, where $P^a_{ss'} \equiv P(s'; s, a)$ is the probability of the state s' being the next state given the action $a \in \mathscr{A}$ and the current state is s . In this paper, we denote by $(PJ)(s) := \int_{\mathcal{S}} J[s'] P^a_{sds'}$.

We now introduce the operator of the game which is of central importance:

$$TJ := \min_{\tau\in\mathcal{T}}\left\{\max_{a\in A} R^a_s + \gamma \sum_{s'\in\mathcal{S}} P^a_{ss'} J^{\tau,\pi}[s'], G(S_\tau)\right\}. \tag{8}$$

The operator T enables the game to be broken down into a sequence of sub minimax problems. It will later play a crucial role in establishing a value iterative method for computing the value of the game.

Before performing the analysis on T, we prove the existence of the value:

Theorem 1.

$$\text{val}^+[J] = \text{val}^-[J] \equiv J^\star. \tag{9}$$

Theorem 1 establishes the existence of the game which permits commuting the max and min operators of the objective (2). Crucially, the theorem secures the existence of an equilibrium pair $(\hat{\tau}, \hat{\pi}) \in \mathcal{T} \times \mathit{\Pi}$, the value of which is given by $J^\star$, the computation of which, is the subject of the next section.

In turn, we can now establish the optimal strategies for each player. To this end, we now define best-response strategies which shall be useful for further characterising the equilibrium:

Definition 2. The set of best-response (BR) strategies for player 1 against the stopping time $\tau \in \mathcal{T}$ (BR strategies for player II against the control policy $\pi \in \mathit{\Pi}$) is defined by $\hat{\pi} \in \underset{\pi'\in\mathit{\Pi}}{\arg\max}\, \mathbb{E}[J^{\tau,\pi'}[\cdot]]$ (resp., $\hat{\tau} \in \underset{\tau'\in\mathcal{T}}{\arg\min}\, \mathbb{E}[J^{\tau',\pi}[\cdot]]$).

The question of computing the value of the game remains. To this end, we now prove the contraction mapping property of the operator T. We then show that repeatedly applying T produces a sequence that converges to the value.

Proposition 1. The operator T in (8) is a contraction.

Prop. 1 underscores a fixed point property which is stated in the following:

Theorem 2. 1. The sequence $(T^n J)^\infty_{n=0}$ converges (in $\mathbb{L}_2$).
2. There exists a unique function $J^\star \in \mathbb{L}_2$ s.th.:

$$J^\star = TJ^\star \text{ and } \lim_{n\to\infty} T^n J = J^\star. \tag{10}$$

Theorem 2 establishes the existence of a fixed point of T. Crucially, it underpins a value iterative method which we formally develop in Sec. 6.

Definition 3. The pair $(\hat{\tau}, \hat{\pi}) \in \mathcal{T} \times \mathit{\Pi}$ is a saddle point equilibrium iff $\forall s \in \mathcal{S}$:

$$J^{\hat{\tau},\hat{\pi}}[s] = \max_{\pi\in\mathit{\Pi}} J^{\hat{\tau},\pi}[s] = \min_{\tau\in\mathcal{T}} J^{\tau,\hat{\pi}}[s]. \tag{11}$$

A saddle point equilibrium therefore defines a strategic configuration in which both players play their BR strategies.

Proposition 2. The pair $(\hat{\tau}, \hat{\pi}) \in \mathcal{T} \times \mathit{\Pi}$ consists of BR strategies and constitutes a saddle point equilibrium.

By Prop. 2, when the pair $(\hat{\tau}, \hat{\pi})$ is played, each player executes its BR strategy in response to their opponent. In the context of the problem in Sec. 2, the strategic response induces risk minimising behaviour by the controller. We now turn to the existence and characterising the optimal stopping time for player 2. The following result establishes its existence.

Theorem 3. There exists an $\mathcal{F}$-measurable stopping time:
$\hat{\tau} = \min \left\{ k \in \mathcal{T} \middle| G(s_k) \leq \min\limits_{k \in \mathcal{T}} \max\limits_{\pi \in \mathit{\Pi}} J^{k,\pi}[s_k] \right\}, a.s.$

Theorem 3 establishes the existence and characterises the player 2 optimal stopping time. The stopping time $\hat{\tau}$ is a best-response for player 2 against the equilibrium policy played by player 1. The theorem plays a vital role in the robust optimal stopping problem of Sec. 4.

Having shown the existence of the optimal stopping time $\tau^\star$, by Theorem 3 and Theorem 1, we find the following:

Theorem 4. Let $\hat{\tau}$ be the player 2 optimal stopping time defined in (3) and let $\tau^\star$ be the optimal stopping time for the robust OSP (c.f. (3)) then $\tau^\star = \hat{\tau}$.

Theorem 4 establishes an equivalence between the robust OSP and the SG of control and stopping. In particular, any method that computes the optimal stopping time with the SG provides a solution to the robust OSP.

6 Simulation-Based Value Iteration

We now develop a simulation-based value-iterative scheme. We show that the method produces an iterative sequence that converges to the value of the game from which the optimal controls can be extracted. The method is suitable for environments in which the transition model and reward functions are not known to either player. Our approach is related to approximated dynamic programming methods e.g. (Bertsekas, 2008). However, our problem requires generalisation to an SG involving a controller and stopper which alters the proofs throughout.

The fixed point property of the game established in Theorem 2 immediately suggests a solution method for finding the value. In particular, we may seek to solve the fixed point equation (FPE) $J^\star = TJ^\star$. Direct approaches at solving the FPE are not generally fruitful as closed solutions are typically unavailable.

To compute the value function, we develop an iterative method that tunes weights of a set of basis functions $\{\phi_k : \mathbb{R}^p \to \mathbb{R} | k \in 1, 2, \ldots D\}$ to approximate $J^\star$ through simulated system trajectories and associated costs. Algorithms of this type were first introduced by Watkins (Watkins and Dayan, 1992) as an approximate dynamic programming method and have since been augmented to cover various settings. Therefore the following can be considered as a generalised Q-learning algorithm for zero-sum controller stopper games.

Let us denote by $\varPhi r := \sum_{j=1}^{D} r(j)\phi_j$ an operator representation of the basis expansion. The algorithm is initialised with weight vector $r_0 = (r_0(1), \ldots, r_0(P))' \in \mathbb{R}^d$. Then as the trajectory $\{s_t | t = 0, 1, 2, \ldots\}$ is simulated, the algorithm produces an updated series of vectors $\{r_t | t = 0, 1, 2, \ldots\}$ by the update:

$$r_{t+1} = r_t + \gamma\phi(s_t)\Big(\max_{a\in\mathcal{A}} R^a_{s_t} + \gamma \min\left\{(\phi r_t)(s_{t+1}), G(s_{t+1})\right\} - (\phi r_t)(s_t)\Big).$$

Theorem 5 demonstrates that the method converges to an approximation of $J^\star$. We provide a bound for the approximation error in terms of the basis choice.

We define the function $Q^\star$ which the algorithm approximates by:

$$Q^\star(s) = \max_{a\in\mathcal{A}} R^a_s + \gamma P J^\star, \qquad \forall s \in \mathcal{S} \tag{12}$$

We later show that $Q^\star$ serves to approximate the value $J^\star$. In particular, we show that the algorithm generates a sequence of weights r_n that converge to a vector $r^\star$ and that $\varPhi r^\star$, in turn approximates $Q^\star$. To complete the connection, we then provide a bound between the outcome of the game when the players use controls generated by the algorithm.

First, we introduce our player 2 stopping criterion which now takes the form: $\hat{\tau} = \min\{t | G(s_t) \leq Q^\star(s_t)\}$. We define a orthogonal projection $\varPi$ and the function F by the following:$\varPi Q := \underset{\bar{Q}\in\{\varPhi r | r\in\mathbb{R}^p\}}{\arg\min} \|\bar{Q} - Q\|, FQ := \max_{a\in\mathcal{A}} R^a_s + \gamma P \min\{G, Q\}$. We now state the main results of the section:

Theorem 5. Under (12), r_n converges to $r^\star$ where $r^\star$ is the unique solution: $\varPi F(\varPhi r^\star) = \varPhi r^\star, \qquad a.e.$

Theorem 6. Let $\hat{\tau} = \min\Big\{k \in \mathcal{T} \Big| G(s_k) \leq (\varPhi r^\star)(s_k)\Big\}$, then the following hold:

1. $\|\varPhi r^\star - Q^\star\| \leq \left(\sqrt{1-\gamma^2}\right)^{-1} \|\varPi Q^\star - Q^\star\|$,
2. $\mathbb{E}\left[J^\star[s] - J^{\tilde{\tau},\tilde{\pi}}[s]\right] \leq \frac{2}{\left[(1-\gamma)\sqrt{1-\gamma^2})\right]} \|\varPi Q^\star - Q^\star\|$.

Theorem 6 says the error bound in algorithm approximation of the value is determined by the goodness of the projection.

Conclusion

In this paper, we tackled the problem of risk within an RL setting in which the controller seeks to obtain a fault-tolerant control that is robust to catastrophic failures. To formally analyse the problem and characterise the optimal behaviour, we performed an in-depth analysis of a stochastic game (SG) of control and stopping. We established the existence of an equilibrium value then, using a contraction mapping argument, showed that the game can be solved by iterative application of a Bellman operator. We proved that the method leads to an approximate dynamic programming algorithm so that the game can be solved by simulation. By proving an equivalence between the SG and robust optimal stopping problems, we showed that the method developed in the paper serves to compute solutions to optimal stopping problems in worst-case scenarios.

Bibliography

Christopher JCH Watkins and Peter Dayan. 1992. Q-learning. Machine learning 8, 3-4 (1992), 279–292.

Pieter Abbeel, Adam Coates, and Andrew Y Ng. 2010. Autonomous helicopter aerobatics through apprenticeship learning. The International Journal of Robotics Research 29, 13 (2010), 1608–1639.

Toshiyuki Yasuda, Kazuhiro Ohkura, and Kanji Ueda. 2006. A homogeneous mobile robot team that is fault-tolerant. Advanced Engineering Informatics 20, 3 (2006), 301–311.

Dapeng Zhang and Zhiwei Gao. 2018. Reinforcement learning–based fault-tolerant control with application to flux cored wire system. Measurement and Control 51, 7-8 (2018), 349–359.

Enhancing R&D in science-based industry: An optimal stopping model for drug discovery. International Journal of Project Management 27, 8 (2009), 754–764.

Jerzy A Filar, Todd A Schultz, Frank Thuijsman, and OJ Vrieze. 1991. Nonlinear programming and stationary equilibria in stochastic games. Mathematical Programming 50, 1-3 (1991), 227–237.

A Maitra and T Parthasarathy. 1970. On stochastic games. Journal of Optimization Theory and Applications 5, 4 (1970), 289–300.

Itamar Arel, Cong Liu, T Urbanik, and AG Kohls. 2010. Reinforcement learning-based multi-agent system for network traffic signal control. IET Intelligent Transport Systems 4, 2 (2010), 128–135.

Fouzia Baghery, Sven Haadem, Bernt Øksendal, and Isabelle Turpin. 2013. Optimal stopping and stochastic control differential games for jump diffusions. Stochastics An International Journal of Probability and Stochastic Processes 85, 1 (2013), 85–97.

Erhan Bayraktar, Xueying Hu, and Virginia R Young. 2011. Minimizing the probability of lifetime ruin under stochastic volatility. Insurance: Mathematics and Economics 49, 2 (2011), 194–206.

Dimitri P Bertsekas. 2008. Approximate dynamic programming. (2008).

David Blackwell and Tom S Ferguson. 1968. The big match. The Annals of Mathematical Statistics 39, 1 (1968), 159–163.

Peter Carr and Dilip B Madan. 2005. A note on sufficient conditions for no arbitrage. Finance Research Letters 2, 3 (2005), 125–130.

Jean-Philippe Chancelier, Bernt Øksendal, and Agnès Sulem. 2002. Combined stochastic control and optimal stopping, and application to numerical approximation of combined stochastic and impulse control. 237, 0 (2002), 149–172.

Marc Peter Deisenroth, Gerhard Neumann, Jan Peters, et al. 2013. A survey on policy search for robotics. Foundations and Trends® in Robotics 2, 1–2 (2013), 1–142.

EB Dynkin. 1967. Game variant of a problem on optimal stopping. In Soviet Math. Dokl., Vol. 10. 270–274.

Javier Garcia and Fernando Fernández. 2012. Safe exploration of state and action spaces in reinforcement learning. Journal of Artificial Intelligence Research 45 (2012), 515–564.

Javier Garcıa and Fernando Fernández. 2015. A comprehensive survey on safe reinforcement learning. Journal of Machine Learning Research 16, 1 (2015), 1437–1480.

Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. 2019. Guidelines for reinforcement learning in healthcare. Nature medicine 25, 1 (2019), 16–18.

Feng Guo, Carl R Chen, and Ying Sophie Huang. 2011. Markets contagion during financial crisis: A regime-switching approach. International Review of Economics & Finance 20, 1 (2011), 95–109.

Theodore Preston Hill. 1979. On the existence of good Markov strategies. Trans. Amer. Math. Soc. 247 (1979), 157–176.

Christopher Jennison and Bruce W Turnbull. 2013. Interim monitoring of clinical trials: Decision theory, dynamic programming and optimal stopping. Kuwait Journal of Science 40, 2 (2013).

Ying Jiao and Huyên Pham. 2011. Optimal investment with counterparty risk: a default-density model approach. Finance and Stochastics 15, 4 (2011), 725–753.

Ioannis Karatzas and William Sudderth. 2006. 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. World Scientific, 100–117.

Thomas Kruse and Philipp Strack. 2015. Optimal stopping with private information. Journal of Economic Theory 159 (2015), 702–727.

David Mguni. 2018. A Viscosity Approach to Stochastic Differential Games of Control and Stopping Involving Impulsive Control. arXiv preprint arXiv:1803.11432 (2018).

Jun Morimoto and Kenji Doya. 2001. Robust reinforcement learning. In Advances in Neural Information Processing Systems. 1061–1067.

Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay V Vazirani. 2007. Algorithmic game theory. Cambridge University Press.

Goran Peskir and Albert Shiryaev. 2006. Optimal stopping and free-boundary problems. Springer.

Huyên Pham. 1997. Optimal stopping, free boundary, and American option in a jump-diffusion model. Applied Mathematics and Optimization 35, 2 (1997), 145–164.

Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. 2016. Safe, multi-agent, reinforcement learning for autonomous driving. arXiv preprint arXiv:1610.03295 (2016).

Lloyd S Shapley. 1953. Stochastic games. Proceedings of the national academy of sciences 39, 10 (1953), 1095–1100.

Aviv Tamar, Yinlam Chow, Mohammad Ghavamzadeh, and Shie Mannor. 2015. Policy gradient for coherent risk measures. In Advances in Neural Information Processing Systems. 1468–1476.

John N Tsitsiklis and Benjamin Van Roy. 1999. Optimal stopping of Markov processes: Hilbert space theory, approximation algorithms, and an application to pricing high-dimensional financial derivatives. IEEE Trans. Automat. Control 44, 10 (1999), 1840–1851.

Ka-Fai Cedric Yiu. 2004. Optimal portfolios under a value-at-risk constraint. Journal of Economic Dynamics and Control 28, 7 (2004), 1317–1334.

Virginia R Young. 2004. Optimal investment strategy to minimize the probability of lifetime ruin. North American Actuarial Journal 8, 4 (2004), 106–126.

Guozhen Zhao and Wen Chen. 2009. Enhancing R&D in science-based industry: An optimal stopping model for drug discovery. International Journal of Project Management 27, 8 (2009), 754–764.

Supplementary Material (Appendix)

Assumptions

Our results are built under the following assumptions:

Assumption A.1. Stationarity: the expectations $\mathbb{E}$ are taken w.r.t. a stationary distribution so that for any measurable function f we have $\mathbb{E}\left[f(s)\right] = \mathbb{E}\left[f(s_k)\right]$ for any $k \geq 0$.

Assumption A.2. Ergodicity: i) Any invariant random variable of the state process is $P-$almost surely ($P-$a.s.) a constant.

Assumption A.3. Markovian transition dynamics: the transition probability function P satisfies the following equality: $P(s_{k+1} \in A|\mathcal{F}_k) = P(s_{k+1}, A)$ for any $A \in \mathcal{B}(\mathbb{R}^p)$.

Assumption A.4. The constituent functions $\{R, G\}$ in (1) are square integrable: that is, $R, G \in \mathbb{L}_2(\mu)$.

Additional Lemmata

We begin the analysis with some preliminary lemmeta which are useful for proving the main results.

Definition D.1. An operator $T : \mathcal{V} \to \mathcal{V}$ is non-expansive if $\forall V_1, V_2 \in \mathcal{V}$ we have:

$$\|TV_1 - TV_2\| \leq \|V_1 - V_2\|. \tag{13}$$

Definition D.2. The residual of a vector $V \in \mathcal{V}$ w.r.t the operator $T : \mathcal{V} \to \mathcal{V}$ is:

$$\epsilon_T(V) := \|TV - V\|. \tag{14}$$

Lemma B.1. Define $\text{val}^+[f] := \min_{b\in\mathbb{B}} \max_{a\in\mathbb{A}} f(a,b)$ and define $\text{val}^-[f] := \max_{a\in\mathbb{A}} \min_{b\in\mathbb{B}} f(a,b)$, then for any $b \in \mathbb{B}$ we have that for any $f, g \in \mathbb{L}$ and for any $c \in \mathbb{R}_{>0}$:

$$\left|\max_{a\in\mathbb{A}} f(a,b) - \max_{a\in\mathbb{A}} g(a,b)\right| \leq c \implies \left|\text{val}^-[f] - \text{val}^-[g]\right| \leq c.$$

Lemma B.2. For any $f, g, h \in \mathbb{L}$ and for any $c \in \mathbb{R}_{>0}$ we have that:

$$\|f - g\| \leq c \implies \|\min\{f,h\} - \min\{g,h\}\| \leq c.$$

Lemma B.3. Let the functions $f, g, h \in \mathbb{L}$ then

$$\|\max\{f,h\} - \max\{g,h\}\| \leq \|f - g\|. \tag{15}$$

The following estimates provide bounds on the value $J^\star$ which we use later in the development of the iterative algorithm. We defer the proof of the results to the appendix.

Lemma B.4. Let $T : \mathcal{V} \to \mathcal{V}$ be a contraction mapping in $\|\cdot\|$ and let $V^\star$ be a fixed point so that $TJ^\star = J^\star$ then there exists a constant $c \in [0, 1[$ s.th:

$$\|J^\star - J\| \leq (1 - c)^{-1}\epsilon_T(J). \tag{16}$$

Lemma B.5. Let $T_1 : \mathcal{V} \to \mathcal{V}, T_2 : \mathcal{V} \to \mathcal{V}$ be contraction mappings and suppose there exists vectors $J_1^\star, J_2^\star$ s.th $T_1 J_1^\star = J_1^\star$ and $T_2 J_2^\star = J_2^\star$ (i.e. $J_1^\star, J_2^\star$ are fixed points w.r.t T_1 and T_2 respectively) then $\exists c_1, c_2 \in [0, 1[$ s.th:

$$\|J_1^\star - J_2^\star\| \leq (1 - \{c_1 \wedge c_2\})^{-1} (\epsilon_{T_1}(J) - \epsilon_{T_2}(J)) .$$

Lemma B.6. The operator T satisfies the following:

1. (Monotonicity) For any $J_1, J_2 \in \mathbb{L}_2$ s.th. $J_1(s) \leq J_2(s)$ then $TJ_1 \leq TJ_2$.
2. (Constant shift) Let $I(s) \equiv 1$ be the unit function, then for any $J \in \mathbb{L}_2$ and for any scalar $\alpha \in \mathbb{R}$, T satisfies $T(J + \alpha I)(s) = TJ(s) + \alpha I(s)$.

The proof uses an application of Lemma B.2.

Proof of Results

Proof of Lemma B.1.

Proof. We begin by noting the following inequality for any $f : \mathcal{V} \times \mathcal{V} \to \mathbb{R}, g : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$ s.th. $f, g \in \mathbb{L}$ we have that for all $b \in \mathcal{V}$:

$$\left|\max_{a\in\mathcal{V}} f(a, b) - \max_{a\in\mathcal{V}} g(a, b)\right| \leq \max_{a\in\mathcal{V}} |f(a, b) - g(a, b)| . \tag{17}$$

From (17) we can straightforwardly derive the fact that for any $b \in \mathcal{V}$:

$$\left|\min_{a\in\mathcal{V}} f(a, b) - \min_{a\in\mathcal{V}} g(a, b)\right| \leq \max_{a\in\mathcal{V}} |f(a, b) - g(a, b)| , \tag{18}$$

(this can be seen by negating each of the functions in (17) and using the properties of the max operator).

Assume that for any $b \in \mathcal{V}$ the following inequality holds:

$$\max_{a\in\mathcal{V}} |f(a, b) - g(a, b)| \leq c \tag{19}$$

Since (18) holds for any $b \in \mathcal{V}$ and, by (17), we have in particular that

$$\begin{aligned} &\left|\max_{b\in\mathcal{V}} \min_{a\in\mathcal{V}} f(a, b) - \max_{b\in\mathcal{V}} \min_{a\in\mathcal{V}} g(a, b)\right| \\ &\leq \max_{b\in\mathcal{V}} \left|\min_{a\in\mathcal{V}} f(a, b) - \min_{a\in\mathcal{V}} g(a, b)\right| \\ &\leq \max_{b\in\mathcal{V}} \max_{a\in\mathcal{V}} |f(a, b) - g(a, b)| \leq c, \end{aligned} \tag{20}$$

whenever (19) holds which gives the required result. □

Proof of Theorem 1

Proof. We begin by noting the following inequality holds:

$$\mathrm{val}^+[J] = \min_{\tau\in\mathcal{T}} \max_{\pi\in\mathit{\Pi}} \mathbb{E}[J^{\tau,\pi}[s]] \geq \max_{\pi\in\mathit{\Pi}} \min_{\tau\in\mathcal{T}} \mathbb{E}[J^{\tau,\pi}[s]] = \mathrm{val}^-[J]. \tag{21}$$

The inequality follows by noticing $J^{k,\pi} \leq \max\limits_{\pi\in\mathit{\Pi}} J^{k,\pi}$ and thereafter applying the $\min_{k\in\mathcal{T}}$ and $\max_{\pi\in\mathit{\Pi}}$ operators.

The proof can now be settled by reversing the inequality in (21). To begin, choose a sequence of open intervals $\{D_m\}_{m=1}^{\infty}$ s.th. for each $m = 1, 2, \ldots$ $\bar{D}_m$ is compact and $\bar{D}_m \supset \bar{D}_{m+1}$ and $[0, T] = \cap_{m=1}^{\infty} \bar{D}_m$ and define $\tau_D(m) := \inf_{k\in D_m} \mathbb{E}[J^{k,\pi}[s_0]]$.

We now observe that:

$$\begin{aligned}
\mathbb{E}[J^{\tau,\hat{\pi}}[s]] &= \max_{\pi\in\mathit{\Pi}} \mathbb{E}\left[\sum_{t=0}^{\tau_D(m)} \gamma^t(R(s_t,a_t) + G(s_{\tau_D(m)}))\right] - \mathbb{E}\left[\sum_{t=\tau}^{\tau_D(m)} \gamma^t(R(s_t,a_t) + G(s_{\tau_D(m)}))\right] \\
&\geq \mathbb{E}\left[J^{\tau_D(m),\pi}[s]\right] - \left|\mathbb{E}\left[\sum_{t=\tau}^{\tau_D(m)} \gamma^t(R(s_t,a_t) + G(s_{\tau_D(m)}))\right]\right| \\
&\geq \mathbb{E}\left[J^{\tau_D(m),\pi}[s]\right] - \sum_{t=\tau}^{\tau_D(m)} \gamma^t \left|\mathbb{E}[R(s_t,a_t)] + \mathbb{E}\left[G(s_{\tau_D(m)})\right]\right| \\
&\geq \mathbb{E}\left[J^{\tau_D(m),\pi}[s]\right] - \sum_{t=\tau}^{\tau_D(m)} \gamma^t \left(\mathbb{E}\left[|R(s_0,\cdot)|\right] + \mathbb{E}\left[|G(s_0)|\right]\right) \\
&= \mathbb{E}\left[J^{\tau_D(m),\pi}[s]\right] + \gamma^{\tau_D(m)+1}\frac{1-\gamma^{\tau-\tau_D(m)}}{1-\gamma}c \\
&= \lim_{m\to\infty} \inf \mathbb{E}[J^{\tau_D(m),\pi}[s]] + \lim_{m\to\infty}\left[\gamma^{\tau_D(m)+1}\frac{1-\gamma^{\tau-\tau_D(m)}}{1-\gamma}\right]c \geq \mathbb{E}[J^{\tau,\pi}[s]],
\end{aligned}$$

where we have used the stationarity property and, in the limit $m \to \infty$ and, in the last line we used the Fatou lemma. The constant c is given by $c := (\mathbb{E}[R(s_0,\cdot)] + \mathbb{E}[G(s_0)]) \in \mathbb{L}$.

Hence, we now find that

$$\mathbb{E}[J^{\tau,\hat{\pi}}[s]] \geq \mathbb{E}[J^{\tau,\pi}[s]]. \tag{22}$$

Now since (22) holds $\forall \pi \in \mathit{\Pi}$ we find that:

$$\mathbb{E}[J^{\tau,\hat{\pi}}[s]] \geq \max_{\pi\in\mathit{\Pi}} \mathbb{E}[J^{\tau,\pi}[s]]. \tag{23}$$

Lastly, applying min operator we observe that:

$$\mathbb{E}[J^{\hat{\tau},\hat{\pi}}[s]] \geq \min_{\tau\in\mathcal{T}} \max_{\pi\in\mathit{\Pi}} \mathbb{E}[J^{\tau,\pi}[s]] = \mathrm{val}^+[J]. \tag{24}$$

It now remains to show the reverse inequality holds:

$$\mathbb{E}[J^{\hat{\tau},\hat{\pi}}[s]] \leq \max_{\pi\in\Pi}\min_{\tau\in\mathcal{T}} \mathbb{E}[J^{\tau,\pi}[s]] = \text{val}^{-}[J]. \tag{25}$$

Indeed, we observe that

$$\mathbb{E}\left[J^{\hat{\tau},\hat{\pi}}[s]\right] \leq \min_{\tau\in\mathcal{T}} \mathbb{E}\left[J^{\tau\wedge m,\hat{\pi}}[s]\right] + \mathbb{E}\left[\sum_{t=m}^{\infty}\gamma^{t}\left(|R(s_t,a_t)|+|G(s_t)|\right)\right] \tag{26}$$

$$\leq \lim_{m\to\infty}\left[\min_{\tau\in\mathcal{T}} \mathbb{E}\left[J^{\tau\wedge m,\hat{\pi}}[s]\right] + c(m)\right] \tag{27}$$

$$= \min_{\tau\in\mathcal{T}} \mathbb{E}\left[J^{\tau,\hat{\pi}}[s]\right] \leq \max_{\pi\in\Pi}\min_{\tau\in\mathcal{T}} \mathbb{E}\left[J^{\tau,\pi}[s]\right], \tag{28}$$

since $\gamma \in [0,1[$,where $c(m) := \frac{\gamma^m}{1-\gamma}(\mathbb{E}[|R(s_0,\cdot)|] + \mathbb{E}[|G(s_0)|])$ (using the stationarity of the state process) and where we have used Lebesgue's Dominated Convergence Theorem in the penultimate step.

Hence, by (28) we have that:

$$\mathbb{E}\left[J^{\hat{\tau},\hat{\pi}}[s]\right] \leq \max_{\pi\in\Pi}\min_{\tau\in\mathcal{T}} \mathbb{E}\left[J^{\tau,\pi}[s]\right] = \text{val}^{-}[J]. \tag{29}$$

Hence putting (24) and (29) together gives:

$$\begin{aligned}\text{val}^{-}[J] &= \max_{\pi\in\Pi}\min_{\tau\in\mathcal{T}} \mathbb{E}\left[J^{\tau,\pi}[s]\right] \\ &\geq \mathbb{E}[J^{\hat{\tau},\hat{\pi}}[s]] \geq \min_{\tau\in\mathcal{T}}\max_{\pi\in\Pi} \mathbb{E}[J^{\tau,\pi}[s]] = \text{val}^{+}[J].\end{aligned} \tag{30}$$

After combining (30) with (21) we deduce the thesis. □

Proof of Theorem 3.

Proof. For any $m \in \mathbb{N}$ we have that

$$\max_{\pi\in\Pi} J^{\tau,\pi}[s] \geq \max_{\pi\in\Pi} J^{\tau\wedge m,\pi}[s] - \sum_{t=m}^{\infty}\gamma^{t}\max_{\pi\in\Pi}\left(|R(s_t,a_t)|+|G(s_t)|\right).$$

We now apply the min operator to both sides of (31) which gives:

$$\min_{\tau\in\mathcal{T}}\max_{\pi\in\Pi} J^{\tau,\pi}[s] \geq \min_{\tau\in\mathcal{T}}\max_{\pi\in\Pi} J^{\tau\wedge m,\pi}[s] - \sum_{t=m}^{\infty}\gamma^{t}\max_{\pi\in\Pi}\left(|R(s_t,a_t)|+|G(s_t)|\right). \tag{31}$$

After taking expectations, we find that:

$$\mathbb{E}\left[\min_{\tau\in\mathcal{T}}\max_{\pi\in\Pi} J^{\tau,\pi}[s]\right] \tag{32}$$

$$\geq \mathbb{E}\left[\min_{\tau\in\mathcal{T}}\max_{\pi\in\Pi} J^{\tau\wedge m,\pi}[s]\right] - \sum_{t=m}^{\infty}\gamma^{t}\mathbb{E}\left[\max_{\pi\in\Pi}\left(|R(s_t,a_t)|+|G(s_t)|\right)\right]. \tag{33}$$

Now by Jensen's inequality and, using the stationarity of the state process (recall the expectation is taken under π) we have that:

$$\begin{aligned}&\mathbb{E}\left[\max_{\pi\in\Pi}\ (|R(s_t,a_t)|+|G(s_t)|)\right]\\&\geq\max_{\pi\in\Pi}\mathbb{E}\left[(|R(s_t,a_t)|+|G(s_t)|)\right]=\mathbb{E}[|R(s_0,\cdot)|]+\mathbb{E}[|G(s_0)|].\end{aligned}\tag{34}$$

By standard arguments of dynamic programming, the value of the game with horizon n can be obtained from n iterations of the dynamic recursion; in particular, we have that:

$$\min_{\tau\in\mathcal{T}}\max_{\pi\in\Pi}J^{\tau\wedge m,\pi}[s]=T^mG(s),\tag{35}$$

Inserting (34) and (35) into (33) gives:

$$\begin{aligned}&\mathbb{E}\left[\min_{\tau\in\mathcal{T}}\max_{\pi\in\Pi}J^{\tau,\pi}[s]\right]\qquad(36)\\&\geq\mathbb{E}\left[T^mG(s)\right]-c(m)=\lim_{m\to\infty}\left[\mathbb{E}\left[T^mG(s)\right]-c(m)\right]=\mathbb{E}\left[J^{\hat{\tau},\hat{\pi}}[s]\right]\end{aligned}$$

where $c(m):=\frac{\gamma^m}{1-\gamma}(\mathbb{E}[|R(s_0,\cdot)|]+\mathbb{E}[|G(s_0)|])$ so that $\lim_{m\to\infty}c(m)=0$. Hence, we find that:

$$\mathbb{E}\left[J^{\hat{\tau},\hat{\pi}}[s]\right]\leq\mathbb{E}\left[\min_{\tau\in\mathcal{T}}\max_{\pi\in\Pi}J^{\tau,\pi}[s]\right].\tag{37}$$

We deduce the result after noting that by definition of G we have that $G(s_\tau)=J^{\tau,\cdot}[s_\tau]$. □

The following lemma is a required result for proving the contraction mapping property of the operator T.

Lemma 1. The probability transition kernel P is non-expansive, that is:

$$\|PV_1-PV_2\|\leq\|V_1-V_2\|.\tag{38}$$

Proof of Lemma 1.

Proof. The proof is standard, we give the details for the sake of completion. Indeed, using the Tonelli-Fubini theorem and the iterated law of expectations, we have that:

$$\begin{aligned}&\|PJ\|^2=\mathbb{E}\left[(PJ)^2[s_0]\right]\\&=\mathbb{E}\left(\left[\mathbb{E}\left[J[s_1]|s_0\right]\right)^2\right]\leq\mathbb{E}\left[\mathbb{E}\left[J^2[s_1]|s_0\right]\right]=\mathbb{E}\left[J^2[s_1]\right]=\|J\|^2,\end{aligned}$$

where we have used Jensen's inequality to generate the inequality. This completes the proof. □

Proof of Lemma B.6.

Proof. Part 2 immediately follows from the properties of the max and min operators. It remains only to prove part 1.
We seek to prove that for any $s \in \mathcal{S}$, if $J \leq \bar{J}$ then

$$\min_{\tau \in \mathcal{T}} \left\{ \max_{a \in A} R_s^a + \gamma \sum_{s' \in \mathcal{S}} P^a_{ss'} J^{\tau,\pi}[s'], G(S_\tau) \right\} \\ - \min_{\tau \in \mathcal{T}} \left\{ \max_{a \in A} R_s^a + \gamma \sum_{s' \in \mathcal{S}} P^a_{ss'} \bar{J}^{\pi}[s'], G(S_\tau) \right\} \leq 0 \tag{39}$$

We begin by firstly making the following observations:
1. For any $x, y, h \in \mathcal{V}$

$$x \leq y \implies \min\{x, h\} \leq \min\{y, h\}. \tag{40}$$

2. For any $f, g, h \in \mathbb{L}_2$

$$\left| \max_{x \in \mathcal{V}} f(x) - \max_{x \in \mathcal{V}} g(x) \right| \leq \max_{x \in \mathcal{V}} \left| f(x) - g(x) \right|. \tag{41}$$

Assume that $J \leq \bar{J}$, then we observe that:

$$\max_{a \in \mathcal{A}} \left\{ R_s^a + \gamma \sum_{s' \in \mathcal{S}} P^a_{ss'} J^{\tau,\pi}[s'] \right\} - \max_{a \in \mathcal{A}} \left\{ R_s^a + \gamma \sum_{s' \in \mathcal{S}} P^a_{ss'} \bar{J}^{\pi}[s'] \right\} \tag{42}$$

$$\leq \gamma \max_{a \in A} \left\{ \sum_{s' \in \mathcal{S}} P^a_{ss'} \left(J^{\tau,\pi}[s'] - \bar{J}^{\pi}[s'] \right) \right\}$$

$$= \gamma \left((PJ) - \left(P\bar{J} \right) \right) \leq J - \bar{J} \leq 0,$$

where we have used (41) in the penultimate line. The result immediately follows after applying (40). □

Proof of Proposition 1 .

Proof. We wish to prove that:

$$\| TJ - T\bar{J} \|_\pi \leq \gamma \| J - \bar{J} \|. \tag{43}$$

Firstly, we observe that:

$$\left\| \max_{a \in A} \left\{ R_s^a + \gamma \sum_{s' \in \mathcal{S}} P^a_{ss'} J^{\tau,\pi}[s'], G(s_k) \right\} - \left(\max_{a \in A} \left\{ R_s^a + \gamma \sum_{s' \in \mathcal{S}} P^a_{ss'} \bar{J}^{\pi}[s'], \bar{G}(s_k) \right\} \right) \right\|$$

$$\leq \gamma \max_{a \in A} \left\| \sum_{s' \in \mathcal{S}} P^a_{ss'} \left(J^{\tau,\pi}_{s-1}[s'] - \bar{J}^{\pi}_{s-1}[s'] \right) \right\| \leq \gamma \left\| J^{\tau,\pi}_{s-1} - \bar{J}^{\pi}_{s-1} \right\|,$$

using Cauchy-Schwartz (and that $\gamma \in [0, 1[$) and (41). The result follows after applying Lemma B.2 and Lemma B.3. □

Proof of Theorem 2

Proof. Part 1: We note that the contraction property of T (c.f. Prop. 1) allows us to demonstrate that the game has a unique fixed point to which a sequence $(T^nJ)_{n=0}^{\infty}$ converges (in $\mathbb{L}_2$). In particular, by Prop. 1 we have that $\|T^2J - TJ\| \leq \gamma\|TJ - J\|$ which proves that the sequence $(T^nJ)_{n=0}^{\infty}$ converges to a fixed point.

Part 2: We observe that the fixed point is unique since if $\exists J, M \in \mathbb{L}_2$ s.th. $TJ = J$ and $TM = M$ we find that $\|M - J\| = \|TM - TJ\| = \gamma\|M - J\|$, so that $M = J$ (since $\gamma \in [0, 1[$) which gives the desired result.

Adopting notions in dynamic programming, denote by:

$$T^nJ[s] = \min_{\tau\in\mathcal{T}} \max_{\pi_0,\pi_1,\ldots,\pi_{n-1}} \mathbb{E}\left[\sum_{t=0}^{\{n-1\wedge\tau\}} \gamma^t R(s_t, a_t) + \gamma^n J(s_{n\wedge\tau})\right].$$

We begin the proof by invoking similar reasoning as (26) - (27) to deduce that:

$$\mathbb{E}\left[J^{\hat{\tau},\hat{\pi}}[s]\right] \leq \min_{\tau\in\mathcal{T}} \mathbb{E}\left[J^{\tau\wedge n,\hat{\pi}}[s]\right] + \frac{\gamma^n}{1-\gamma}c,$$

where $c := (\mathbb{E}[|R(s_0, \cdot)|] + \mathbb{E}[|G(s_0)|])$. Hence,

$$T^nJ[s] \leq \max_{\pi\in\Pi} \min_{\tau\in\mathcal{T}} \mathbb{E}\left[J^{\tau,\pi}[s]\right] + \frac{\gamma^n}{1-\gamma}c = J^{\star}[s] + \frac{\gamma^n}{1-\gamma}c. \tag{44}$$

By analogous reasoning we can deduce that:

$$T^nJ[s] \geq \min_{\tau\in\mathcal{T}} \max_{\pi\in\Pi} \mathbb{E}\left[J^{\tau,\pi}[s]\right] - \frac{\gamma^n}{1-\gamma}c = J^{\star}[s] - \frac{\gamma^n}{1-\gamma}c. \tag{45}$$

Putting (44) and (45) together implies:

$$J^{\star}[s] - \frac{\gamma^n}{1-\gamma}c \leq T^nJ[s] \leq J^{\star}[s] + \frac{\gamma^n}{1-\gamma}c. \tag{46}$$

By Lemma B.6, i.e. invoking the monotonicity and constant shift properties of T, we can apply T to (46) and preserve the inequalities to give:

$$TJ^{\star}[s] - \frac{\gamma^n}{1-\gamma}c \leq T^{n+1}J[s] \leq TJ^{\star}[s] + \frac{\gamma^n}{1-\gamma}c. \tag{47}$$

After taking the limit in (47) and, using the sandwich theorem of calculus, we deduce the result. □

Proof of Theorem 3

Proof. For any $m \in \mathbb{N}$ we have that:

$$\max_{\pi\in\Pi} J^{\tau,\pi}[s] \geq \max_{\pi\in\Pi} J^{\tau\wedge m,\pi}[s] - \sum_{t=m}^{\infty} \gamma^t \max_{\pi\in\Pi} \left(|R(s_t, a_t)| + |G(s_t)|\right). \tag{48}$$

We now apply the min operator to both sides of (48) which gives:

$$\min_{\tau\in\mathcal{T}}\max_{\pi\in\Pi} J^{\tau,\pi}[s] \geq \min_{\tau\in\mathcal{T}}\max_{\pi\in\Pi} J^{\tau\wedge m,\pi}[s] - \sum_{t=m}^{\infty}\gamma^t \max_{\pi\in\Pi}\left(|R(s_t,a_t)|+|G(s_t)|\right).$$

After taking expectations, we find that:

$$\mathbb{E}\left[\min_{\tau\in\mathcal{T}}\max_{\pi\in\Pi} J^{\tau,\pi}[s]\right] \tag{49}$$

$$\geq \mathbb{E}\left[\min_{\tau\in\mathcal{T}}\max_{\pi\in\Pi} J^{\tau\wedge m,\pi}[s]\right] - \sum_{t=m}^{\infty}\gamma^t\mathbb{E}\left[\max_{\pi\in\Pi}\left(|R(s_t,a_t)|+|G(s_t)|\right)\right]. \tag{50}$$

Now by Jensen's inequality and, using the stationarity of the state process (recall the expectation is taken under π) we have that:

$$\mathbb{E}\left[\max_{\pi\in\Pi}\left(|R(s_t,a_t)|+|G(s_t)|\right)\right]$$

$$\geq \max_{\pi\in\Pi}\mathbb{E}\left[\left(|R(s_t,a_t)|+|G(s_t)|\right)\right] = \mathbb{E}[|R(s_0,\cdot)|]+\mathbb{E}[|G(s_0)|]. \tag{51}$$

By standard arguments of dynamic programming, the value of the game with horizon n can be obtained from n iterations of the dynamic recursion; in particular, we have that:

$$\min_{\tau\in\mathcal{T}}\max_{\pi\in\Pi} J^{\tau\wedge m,\pi}[s] = T^m G(s). \tag{52}$$

Inserting (51) and (52) into (50) gives:

$$\mathbb{E}\left[\min_{\tau\in\mathcal{T}}\max_{\pi\in\Pi} J^{\tau,\pi}[s]\right] \geq \mathbb{E}\left[T^m G(s)\right] - c(m)$$

$$= \lim_{m\to\infty}\left[\mathbb{E}\left[T^m G(s)\right] - c(m)\right] = \mathbb{E}\left[J^{\hat{\tau},\hat{\pi}}[s]\right], \tag{53}$$

where $c(m) := \frac{\gamma^m}{1-\gamma}(\mathbb{E}[|R(s_0,\cdot)|]+\mathbb{E}[|G(s_0)|])$ so that $\lim_{m\to\infty} c(m) = 0$. Hence, we find that:

$$\mathbb{E}\left[J^{\hat{\tau},\hat{\pi}}[s]\right] \leq \mathbb{E}\left[\min_{\tau\in\mathcal{T}}\max_{\pi\in\Pi} J^{\tau,\pi}[s]\right], \tag{54}$$

we deduce the result after noting that $G(s_\tau) = J^{\tau,\cdot}[s_\tau]$ by definition of G. □

Proof of Lemma B.4.

Proof. The proof follows almost immediately from the triangle inequality, indeed for any $J \in \mathbb{L}_2$:

$$\|J^\star - J\| = \|TJ^\star - J\| \leq \gamma\|J^\star - J\| + \|TJ - J\|, \tag{55}$$

where we have added and subtracted TJ to produce the inequality. The result then follows after inserting the definition of $\epsilon_T(J)$. □

Proof of Lemma B.5.

Proof. The proof follows directly from Lemma B.4. Indeed, we observe that for any $J \in \mathbb{L}_2$ we have

$$\|J_1^\star - J_2^\star\| \leq \|J_1^\star - J\| + \|J_2^\star - J\|, \tag{56}$$

where we have added and subtracted J to produce the inequality. The result then follows from Lemma B.4. □

Proof of Proposition 2.

Proof. The proposition follows from the fact that if either player plays a Markov strategy then their opponent's best-response is a Markov strategy. Moreover, by Theorem 2, $\hat{\tau}$ is a BR strategy for player 2 (recall Definition 3). Moreover, by Theorem 1 (commuting the max and min operators) we observe that $\hat{\pi}$ is a BR strategy for player 1. □

The proofs of the results in Sec. 9 are constructed in a similar fashion that in (Bertsekas, 2008) (approximate dynamic programming). However, the analysis incorporates some important departures due to the need to accommodate the actions of two players that operate antagonistically.

We now prove the first of the two results of Sec. 9.
Proof of Theorem 5.

Proof. We firstly notice the construction of $\hat{\tau}$ given by

$$\hat{\tau} = \min\{t|G(s_t) \leq Q^\star\}, \tag{57}$$

is sensible since we observe that

$$\begin{aligned} &\min\{t|G(s_t) \leq J^\star\} \\ =&\min\{t|G(s_t) \leq \min\{G(s_t), Q^\star(s_t)\}\} \\ =&\min\{t|G(s_t) \leq Q^\star\}. \end{aligned}$$

<u>Result 1</u>
<u>Step 1</u> Our first step is to prove the following bound:

$$\left\|FQ - F\bar{Q}\right\| \leq \gamma \left\|Q - \bar{Q}\right\|. \tag{58}$$

Proof.

$$\begin{aligned} &\left\|\max_{a\in\mathcal{A}} R_s^a + \gamma P \min\{G, Q\} - \left(\max_{a\in\mathcal{A}} R_s^a + \gamma P \min\{G, \bar{Q}\}\right)\right\| \\ &= \gamma \left\|P \min\{G, Q\} - P \min\{G, \bar{Q}\}\right\| \\ &\leq \gamma \left\|\min\{G, Q\} - \min\left\{G, \bar{Q}\right\}\right\| \\ &\leq \gamma \left\|Q - \bar{Q}\right\|. \end{aligned}$$

which is the required result. □

Step 2

Our next task is to prove that the quantity $Q^\star$ is a fixed point of F and hence we can apply the operator F to achieve the approximation of the value.

Proof. Using the definition of T (c.f. (13) we find that:

$$\begin{aligned}
&J^\star = TJ^\star \iff \max_{a\in\mathcal{A}} R_s^a + \gamma P J^\star \\
&= \max_{a\in\mathcal{A}} R_s^a + \gamma P \min\left\{\max_{a\in\mathcal{A}} R_s^a + \gamma P J, G\right\} \\
&\iff \\
&Q^\star = \max_{a\in\mathcal{A}} R_s^a + \gamma P \min\left\{Q^\star, G\right\} \\
&\iff \\
&\qquad\qquad Q^\star = FQ^\star.
\end{aligned}$$

□

Step 3

We now prove that the operator ΠF is a contraction on Q, that is the following inequality holds:

$$\left\|\Pi FQ - \Pi F\bar{Q}\right\| \leq \gamma \left\|Q - \bar{Q}\right\|.$$

Proof. The proof follows straightforwardly by the properties of a projection mapping:

$$\left\|\Pi FQ - \Pi F\bar{Q}\right\| \leq \left\|FQ - F\bar{Q}\right\| \leq \gamma \left\|Q - \bar{Q}\right\|.$$

□

Step 4

$$\left\|\Phi r^\star - Q^\star\right\| \leq \frac{1}{\sqrt{1-\gamma^2}} \left\|\Pi Q^\star - Q^\star\right\|. \tag{59}$$

The result is proven using the orthogonality of the (orthogonal) projection and by the Pythagorean theorem. Indeed, we have that:

Proof.

$$\begin{aligned}
&\left\|\Phi r^\star - Q^\star\right\|^2 = \left\|\Phi r^\star - \Pi Q^\star\right\|^2 + \left\|\Pi Q^\star - Q^\star\right\|^2 \\
&= \left\|\Pi F \Phi r^\star - \Pi Q^\star\right\|^2 + \left\|\Pi Q^\star - Q^\star\right\|^2 \\
&= \left\|\Pi F \Phi r^\star - \Pi Q^\star\right\|^2 + \left\|\Pi Q^\star - Q^\star\right\|^2 \\
&\leq \gamma^2 \left\|\Phi r^\star - Q^\star\right\|^2 + \left\|\Pi Q^\star - Q^\star\right\|^2.
\end{aligned}$$

Hence, we find that

$$\|\Phi r^\star - Q^\star\| \leq \frac{1}{\sqrt{1-\gamma^2}} \left\|\Pi Q^\star - Q^\star\right\|,$$

which is the required result. □

<u>Result 2</u>

$$\mathbb{E}\left[J^\star[s]\right] - \mathbb{E}\left[J^{\tilde{\tau},\tilde{\pi}}[s]\right] \leq \frac{2}{[(1-\gamma)\sqrt{1-\gamma^2}]} \|\Pi Q^\star - Q^\star\|. \tag{60}$$

Proof. The proof by Jensen's inequality, stationarity and the non-expansive property of P. In particular, we have

$$\begin{aligned}
&\mathbb{E}\left[J^\star[s]\right] - \mathbb{E}\left[J^{\tilde{\tau},\tilde{\pi}}[s]\right] \\
&= \mathbb{E}\left[PJ^\star[s]\right] - \mathbb{E}\left[PJ^{\tilde{\tau},\tilde{\pi}}[s]\right] \\
&\leq \left|\mathbb{E}\left[PJ^\star[s]\right] - \mathbb{E}\left[PJ^{\tilde{\tau},\tilde{\pi}}[s]\right]\right| \\
&\leq \|PJ - PJ^{\tilde{\tau},\tilde{\pi}}\|.
\end{aligned} \tag{61}$$

Inserting the definitions of $Q^\star$ and $\tilde{Q}$ into (61) then gives:

$$\mathbb{E}\left[J^\star[s]\right] - \mathbb{E}\left[J^{\tilde{\tau},\tilde{\pi}}[s]\right] \leq \frac{1}{\gamma}\|Q^\star - \tilde{Q}\|. \tag{62}$$

It remains therefore to place a bound on the term $\|Q^\star - \tilde{Q}$. We observe that by the triangle inequality and the fixed point properties of F on Q and $\tilde{F}$ on $\tilde{Q}$ we have

$$\begin{aligned}
\|Q^\star - \tilde{Q}\| &\leq \|Q^\star - F(\Phi r^\star)\| + \|\tilde{Q} - F(\Phi r^\star)\| && (63)\\
&\leq \gamma\left\{\|Q^\star - \Phi r^\star\| + \|\tilde{Q} - \Phi r^\star\|\right\} && (64)\\
&\leq \gamma\left\{2\|Q^\star - \Phi r^\star\| + \|Q^\star - \tilde{Q}\|\right\}. && (65)
\end{aligned}$$

So that

$$\|Q^\star - \tilde{Q}\| \leq \frac{2\gamma}{1-\gamma}\|Q^\star - \Phi r^\star\|. \tag{66}$$

The result then follows after substituting the result of step 4 (59). □

Let us now define the following quantity:

$$HQ(s) := \begin{cases} G(s) & \text{if } G(s) \leq (\Phi r^\star)(s) \\ Q(s) & \text{otherwise,} \end{cases} \tag{67}$$

and

$$\tilde{F}Q := \max_{a\in\mathcal{A}} R_s^a + \gamma PHQ. \tag{68}$$

<u>Step 5</u>

$$\left\|\tilde{F}Q - \tilde{F}\bar{Q}\right\| \leq \gamma \left\|Q - \bar{Q}\right\| \tag{69}$$

Proof.

$$\begin{aligned}
&\left\|\tilde{F}Q - \tilde{F}\bar{Q}\right\| = \left\|\max_{a\in\mathcal{A}} R_s^a + \gamma PHQ - \left(\max_{a\in\mathcal{A}} R_s^a + \gamma PH\bar{Q}\right)\right\| \\
&= \gamma \left\|PHQ - PH\bar{Q}\right\| \\
&\leq \gamma \left\|HQ - H\bar{Q}\right\| \\
&= \gamma \left\|\min\{G, Q\} - \min\{G, \bar{Q}\}\right\| \\
&\leq \gamma \left\|Q - \bar{Q}\right\|.
\end{aligned}$$

We now prove that $\tilde{Q} = \max\limits_{a\in\mathcal{A}} R_s^a + \gamma PJ^{\pi,\tilde{\tau}}$ is a fixed point.

$$\begin{aligned}
H\tilde{Q} &= H\left(\max_{a\in\mathcal{A}} R_s^a + \gamma PJ^{\pi,\tilde{\tau}}\right) \\
&= \begin{cases} G(s) & \text{if } G(s) \leq (\varPhi r^\star)(s) \\ \max\limits_{a\in\mathcal{A}} R_s^a + \gamma PJ^{\pi,\tilde{\tau}} & \text{otherwise} \end{cases} \\
&= J^{\pi,\tilde{\tau}}
\end{aligned}$$

□

Let us now define the following quantity:

$$s(z,r) := \phi(s)\left(\max_{a\in\mathcal{A}} R_s^a + \gamma \min\left\{(\varPhi r)(y), G(y)\right\} - (\varPhi r)(s)\right).$$

Additionally, we define $\bar{s}$ by the following:

$$\bar{s}(z,r) := \mathbb{E}\left[s(z_0, r)\right].$$

The components of $s(z,r)$ are then given by:

$$s_k \equiv \mathbb{E}\left[\phi_k(s_0)\left(\max_{a\in\mathcal{A}} R_s^a + \gamma \min\left\{(\phi r)(s_0), G(s_0)\right\} - (\phi r)(s_0)\right)\right].$$

We now observe that s_k can be described in terms of an inner product. Indeed, using the iterated law of expectations we have that

$$\begin{aligned}
s_k &\equiv \mathbb{E}\left[\Phi_k(s_0)\left(\max_{a\in\mathcal{A}} R^a_s + \gamma \min\left\{(\Phi r)(s_0), G(s_0)\right\} - (\Phi r)(s_0)\right)\right] \\
&= \mathbb{E}\left[\Phi_k(s_0)\left(\max_{a\in\mathcal{A}} R^a_s + \gamma \mathbb{E}\left[\min\left\{(\Phi r)(s_0), G(s_0)\right\}|s_0\right] - (\Phi r)(s_0)\right)\right] \\
&= \mathbb{E}\left[\Phi_k(s_0)\left(\max_{a\in\mathcal{A}} R^a_s + \gamma P \min\left\{(\Phi r)(s_0), G(s_0)\right\} - (\Phi r)(s_0)\right)\right] \\
&= \left\langle \Phi_k, F(\Phi r) - F(\Phi r)\right\rangle .
\end{aligned}$$

□

Proof of Theorem 6

Step 5 enables us to use classic arguments for approximate dynamic programming. In particular, following step 5, Theorem 6 follows directly from Theorem 2 in (Tsitsiklis & Van Roy, 1999) with only a minor adjustment in substituting the max operator with min.