FAULT TOLERANT REINFORCEMENT LEARNING VIA A MARKOV GAME OF CONTROL AND STOPPING

Anonymous authors
Paper under double-blind review

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 fault-tolerance within RL in which the controller learns a policy can cope with adversarial attacks and random stoppages that lead to failures of the system subcomponents. The results of the paper also cover fault-tolerant (FT) control so that the controller learns to avoid states that carry risk of system failures. By demonstrating that the class of problems is represented by a variant of SGs, 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.

1 INTRODUCTION

Reinforcement learning (RL) provides the promise of adaptive agents being able to discover solutions merely through repeated interaction with their environment. RL has been deployed in a number of real-world settings in which, using RL, an adaptive agent learns to perform complex tasks, often in environments shared by human beings. Large scale factory industrial applications, traffic light control (Arel et al., 2010), robotics (Deisenroth et al., 2013) and autonomous vehicles (Shalev-Shwartz et al., 2016) are notable examples of settings to which RL methods have been applied.

Numerous automated systems are however, susceptible to failures and unanticipated outcomes. Moreover, many real-world systems amenable to RL suffer the potential for random stoppages and abrupt failures; actuator faults, failing mechanical system components, sensor failures are few such examples. In these settings, executing preprogrammed behaviours or policies that have been trained in idealised simulated environments can prove vastly inadequate for the task of ensuring the safe execution of tasks. Consequently, in the presence of such occurrences, the deployment of RL agents introduces a risk of catastrophic outcomes whenever the agent is required to act so as to avoid adverse outcomes in unseen conditions. The important question of how to control the system in a way that is both robust against systemic faults and, minimises the risk of faults or damage therefore arises.

In response to the need to produce RL algorithms that execute tasks with safety guarantees, a significant amount of focus has recently been placed on safe execution, robust control and risk-minimisation (Garcıa and Fernández, 2015). Examples include H_∞ control (Morimoto and Doya, 2001), coherent risk, conditional value at risk (Tamar et al., 2015). In general, these methods introduce an objective[1] defined with an expectation measure that either penalises actions that lead to greater uncertainty or embeds a more pessimistic view of the world (for example, by biasing the transition predictions towards less desirable states). In both cases, the resulting policies act more cautiously over the horizon of the problem as compared to policies trained with a standard objective function.

Despite the recent focus on safe methods within RL, the question of how to train an RL agent that can cope with random failures remains unaddressed. In particular, at present the question of how to produce an RL policy that can cope with an abrupt failure of some system subcomponent has received

[1] With a Lagrangian approach, constraints are captured in the construction of the Lagrangian.

no systematic treatment. Similarly, the task of addressing how to produce RL policies that account for the risk of states in which such failures occur has not been addressed.

In this paper, we for the first time produce a method that learns optimal policies in response to random and adversarial systems attacks that lead to stoppages of system (sub)components that may produce adverse events. Our method works by introducing an adversary that seeks to determine a stopping criterion to stop the system at states that lead to the worst possible (overall) outcomes for the controller. Using a game-theoretic construction, we then show how a policy that is robust against adversarial attacks that lead to abrupt failure can be learned by an adaptive agent using an RL updating method. In particular, the introduction of an adversary that performs attacks at states that lead to worst outcomes generates experiences for the adaptive RL agent to learn a *best-response policy* against such scenarios.

To tackle this problem, we construct a novel two-player stochastic game (SG) in which one of the players, the controller, is delegated the task of learning to modify the system dynamics through its actions that maximise its payoff and an adversary or 'stopper' that enacts a strategy that stops the system in such a way that maximises the controller's costs. This produces a framework that finds optimal policies that are robust against stoppages at times that pose the greatest risk of catastrophe.

The main contribution of the paper is to perform the first systematic treatment of the problem of robust control under worst-case failures. In particular, we perform a formal analysis of the game between the controller and the stopper. 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 rules 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 rule (Theorem 3).

Our results tackle optimal stopping problems (OSPs) under worst-case transitions. OSPs are a subclass of optimal stochastic control (OSC) problems in which the goal is to determine a criterion for stopping at a time that maximises some state-dependent payoff (Peskir and Shiryaev, 2006).

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 (SPE). 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

At present, the coverage of FT within RL is limited. In (Zhang and Gao, 2018) RL is *applied* to tackle systems in which faults might occur and subsequently incur 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 introduces an adversary that performs the task of simulating high-cost stoppages (hence, modelling faults) that induce an FT trained policy.

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 SG 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 (Bayraktar et al., 2011; Baghery et al., 2013; Mguni, 2018). In these analyses, all

aspects of the environment are known and in general, solving these problems requires computing analytic solutions to non-linear partial differential equations which are often analytically insoluble and whose solutions can only be approximated numerically at very low dimensions.

Current iterative methods in OSPs (and approximated dynamic programming methods e.g. (Bertsekas, 2008)) in unknown environments are restricted to risk-neutral settings (Tsitsiklis and Van Roy, 1999) — introducing a notion of risk (generated adversarially) adds considerable difficulty as it requires generalisation to an SG involving a controller and stopper which alters the proofs throughout. In particular, the solution concept is now an SG SPE, the existence of which must be established. As we show, our framework provides an iterative method of solving OSPs with worst-case transitions in unknown environments and hence, generalises existing OSP analyses to incorporate a notion of risk.

Organisation

The paper is organised as follows: we firstly give a formal description of the FT RL problem we tackle and the OSP with worst-case transitions and give a concrete example to illustrate an application of the problem. In Sec. 2, we introduce the underlying SG framework which we use within the main theoretical analysis which we perform in Sec. 3. Lastly, in Sec. 4, we develop an approximate dynamic programming approach that enables the optimal controls to be computed through simulation, followed by some concluding remarks.

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

1.2 Fault-Tolerant Reinforcement Learning

We concern ourselves with finding a control policy that copes with abrupt system stoppages and failures at the worst possible states. 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 or *rule* to decide when to stop given its state observations. In order to generate an FT control, we simulate the adversary's action whilst the controller determines 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 reward which is dependent on the state and the chosen action. 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 $T \in \mathbb{N} \times \{\infty\}$. However, at any given point $\tau_S \leq T$ the system may stop (randomly) and the problem terminates where $\tau_S \sim f(\{0, \dots, T\})$ is a measurable, random exit time and f is some distribution on $\{0, \dots, T\}$. If after $k \leq T$ time steps the system stops, the agent incurs a cost of $G(S_k)$ and the process terminates.

For any $s \in \mathcal{S}$ and for any $\pi \in \Pi$, the agent's **performance function** is given by:

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

where $a \wedge b := \min\{a, b\}$, $\mathbb{E}$ is taken w.r.t. the transition function P. 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 penalty incurred 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 FT control problem which we tackle is one in which the controller acts both with concern for abrupt system failures and stoppages. In particular, the analysis is performed in sympathy with addressing the problem of how the controller should act in two scenarios — the first involves acting in environments that are susceptible to adversarial attacks or random stoppages in high costs states. Such situations are often produced in various real-world scenarios such as engine failures in autonomous vehicles, network power failures and digital (communication) networks attacks. The second scenario involves a controller that seeks to avoid system states that yield a high likelihood

of systemic (subcomponent) failure. Examples of this case include an agent that seeks to avoid performing tasks that increase the risk of some system failure, for example increasing stress that results in component failure or breakages within robotics.

To produce a control that is robust in these scenarios, it is firstly necessary to determine a stopping rule that stops the system at states that incur the highest overall costs. Applying this stopping rule to the system subsequently induces a response by the controller that is robust against systemic faults at states in which stopping inflicts the greatest overall costs. This necessitates a formalism that combines an OSP to determine an optimal (adversarial) stopping rule and secondly, a RL problem. Hence, problem we consider is the following:

Find $(\hat{k}, \hat{\pi}) \in \mathscr{V} \times \Pi$ and $J^{\hat{k},\hat{\pi}}$ s.th.

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

where the minimisation is taken pointwise and $\mathscr{V}$ is a set of stochastic processes of the form $v : \Omega \to \mathscr{T}$ where $\mathscr{T} \subseteq \{0, 1, 2 \ldots\}$ is a set of stopping times.

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

The dual objective (2) consists of finding both a stopping rule 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 between a controller that seeks to maximise J by manipulating state visitations through its actions and an adversarial stopper that chooses a stopping rule to stop the process in order to minimise J. We later 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.

The results of this paper also tackle OSPs under a worst-case transitions — problems in which the goal is to find a stopping rule $\hat{\tau}$ under the adverse *non-linear expectation* $\mathscr{C}_P := \min_{\pi \in \Pi} \mathbb{E}_{P,\pi}$ s.th.

$$\hat{\tau} \in \arg\max_{k \in \mathscr{V}} \mathscr{C}_P \left[\sum_{t=0}^{k \wedge T} \gamma^t R(s_t, a_t) + \gamma^{k \wedge T} G(s_{k \wedge T}) \right]. \tag{3}$$

Here, the agent seeks to find an optimal stopping time in a problem in which the system transitions according to an adversarial (worst-case) probability measure.

1.3 Example: Control with random actuator failure

To elucidate the ideas, we now provide a concrete practical example namely that of actuator failure within RL applications.

Consider an adaptive learner, for example a robot that uses a set of actuators 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 can only execute actions drawn from a subset of its action space $\hat{A} \subset A$ and hence, the agent is now restricted to policies of the form $\pi_{\text{partial}} : S \times \hat{A} \to [0, 1]$ — thereafter its expected return is given by the value function $V^{\pi_{\text{partial}}}$ (this plays the role of the bequest function G in (1)). In order to perform robustly against actuator failure, it is therefore necessary to consider a set of stopping times $\mathscr{T} \subseteq \{0, 1, 2, \ldots\}$ and a stopping criterion $\hat{\tau} : \Omega \to \mathscr{T}$ which determines the worst states for the agent's functionality to be impaired so that it can only use some subset of its set of actuators.

The problem involves finding a pair $(\hat{\tau}, \hat{\pi}) \in \mathscr{V} \times \Pi$ — a stopping time and control policy s.th.

$$\min_{k' \in \mathscr{V}} \left(\max_{\pi' \in \Pi} \mathbb{E}\left[H^{\pi', k'}(s) \right] \right) = \mathbb{E}\left[H^{\hat{\pi}, \hat{\tau}}(s) \right]; \qquad \forall s \in \mathcal{S},$$

where $s := s_0, a_t \sim \pi'$ and $H^{\pi,k}(s) := \sum_{t=0}^{k \wedge \infty} \gamma^t R(s_t, a_t) + \gamma^{k \wedge \infty} V^{\pi_{\text{partial}}}(s_{k \wedge \infty})$. Hence the role of the adversary is to determine and execute the stopping action $\hat{\tau}$ that leads to the greatest reduction in the controller's overall payoff. The controller in turn learns to execute the policy $\hat{\pi}$ which involves

playing a policy $\hat{\pi}_{\text{partial}} \in \arg\max V^{\pi_{\text{partial}}}$ after the adversary has executed its stopping action. The resulting policy $\hat{\pi}$ is hence robust against actuator failure at the worst possible states.

Embedded within problem (4) 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).

2 DISCRETE-TIME STOCHASTIC GAMES OF CONTROL AND STOPPING

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[2] 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 $\{0, \ldots K\}$ where $K \in \mathbb{N} \times \{\infty\}$ is the time horizon. A **stopping time** is defined as a random variable $\tau : \Omega \to \{0, \ldots, K\}$ for which $\{\omega \in \Omega | \tau(\omega) \leq t\} \in \mathcal{F}_t$ for any $t \in \{0, \ldots, K\}$ — this says that given the information generated by the state process, we can determine if the stopping criterion has occurred.

An SG is an augmented Markov decision process 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 I simultaneously represents a cost for player II.

Formally, a two-player zero-sum SG 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 $n \in \mathbb{N}$ 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 I and player II (resp.). The function $R : \mathcal{S} \times \mathcal{A}_1 \times \mathcal{A}_2$ is the one-step reward for player I and represents one-step cost for player II when player I takes action $a_1 \in \mathcal{A}_1$ and player II takes action $a_2 \in \mathcal{A}_2$ 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 I which we denote by J, represents a total expected cost for player II.

Denote by Π_i, the space of strategies for each player $i \in \{1, 2\}$. 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 $\pi_i : \mathcal{S} \times \mathcal{A}_i \to [0, 1]$. 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 $\mathcal{A}_1 = \mathcal{A}_2$ so that the players' actions are drawn from the same set. We depart from this model and consider a game in which player II can choose a strategy which determines a time to stop the process contained within the set $\mathcal{T} \subseteq \{0, 1, 2, \ldots\}$ which consists of $\mathcal{F}-$ measurable stopping times. In this setting, player I can manipulate the system dynamics by taking actions drawn from $\mathcal{A}_1$ (we hereon use $\mathcal{A}$) and at each point, player II can decide to intervene to stop the game.

Let us define by $\text{val}^+[J] := \min_{k\in\mathcal{V}} \max_{\pi\in\Pi} J^{k,\pi}$ the *upper value function* and by $\text{val}^-[J] := \max_{\pi\in\Pi} \min_{k\in\mathcal{V}} J^{k,\pi}$, 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.

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

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

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

$$\mathrm{val}^{-}[J] = \max_{\pi\in\Pi}\min_{k\in\mathscr{V}} J^{k,\pi} = \min_{k\in\mathscr{V}}\max_{\pi\in\Pi} J^{k,\pi} = \mathrm{val}^{+}[J]. \tag{4}$$

We denote the value by $J^\star := \mathrm{val}^{+}[J] = \mathrm{val}^{-}[J]$ and denote by $(\hat{k},\hat{\pi}) \in \mathscr{V}\times\Pi$ 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 $\mathrm{val}^{+}[J]$ and $\mathrm{val}^{-}[J]$ may not coincide. Should $J^\star$ exist, it constitutes an SPE 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. Thus the central task to establish an equilibrium involves unambiguously assigning a value to the game, that is proving the existence of $J^\star$.

3 Main Analysis

In this section, we present the key results and perform the main analysis of the paper. Our first task is to prove the existence of a value of the game. This establishes a fixed or stable point which describes the equilibrium policies enacted by each player. Crucially, the equilibrium describes the maximum payoff that the controller can expect in an environment that is subject to adversarial attacks that stop the system or some subcomponent. Unlike standard SGs with two controllers, introducing a stopping criterion requires an alternative analysis in which i) an equilibrium with Markov strategies in which one of the players uses a stopping criterion is determined and ii) the stopping criterion is characterised. It is well-known that introducing a stopping action to one of the players alters the analysis of SGs the standard methods of which cannot be directly applied (c.f. Dynkin games (Dynkin, 1967)).

Our second task is to perform an analysis that enables us to construct an approximate dynamic programming method. This enables the value function to be computed through simulation. This, as we show in Sec. 4, underpins a simulation-based scheme that is suitable for settings in which the transition model and reward function is a priori unknown. Lastly, we construct an equivalence between robust OSPs and games of control and stopping. We defer some of the proofs to the appendix.

Our results develop the theory of risk within RL to cover instances in which the agent has concern the process at a catastrophic system state. Consequently, we develop the theory of SGs to cover games of control and stopping when neither player has up-front environment knowledge. We prove an equivalence between robust OSPs and games of control and stopping and demonstrate how each problem can be solved in unknown environments.

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:

$$TJ[s] := \min\left\{\max_{a\in A} R^a_s + \gamma\sum_{s'\in\mathcal{S}} P^a_{ss'}J^{\tau,\pi}[s'], G(s)\right\}, \qquad \forall s\in\mathcal{S} \tag{5}$$

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.

We now briefly discuss strategies. A player strategy is a map from the opponent's policy set to the player's own policy set. In general, in two player games the player who performs an action first employs the use of a strategy. Typically, this allows the player to increase its rewards since their action is now a function of the other player's later decisions. Markov controls use only information about the current state and duration of the game rather than using information about the opponent's decisions or the game history. Seemingly, limiting the analysis to Markov controls in the current game may restrict the abilities of the players to perform optimally.

Our first result however proves the existence of the value in Markov controls:

Theorem 1.

$$\mathrm{val}^{+}[J] = \mathrm{val}^{-}[J] \equiv J^\star. \tag{6}$$

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 $(\hat{\tau}, \hat{\pi}) \in \mathscr{V} \times \Pi$, where $\hat{\pi} \in \Pi$ is the controller's optimal **Markov** policy when it faces adversarial attacks that stop the system. Additionally, Theorem 1 establishes the existence of a given by $J^\star$, the computation of which, is the subject of the next section.

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 1. *The set of **best-response (BR) strategies** for player I against the stopping time $\tau \in \mathscr{V}$ (BR strategies for player II against the control policy $\pi \in \Pi$) is defined by:*

$$\hat{\pi} \in \underset{\pi' \in \Pi}{\arg\max}\, \mathbb{E}[J^{\tau,\pi'}[s]] \quad (\textit{resp.,} \hat{\tau} \in \underset{\tau' \in \mathscr{V}}{\arg\min}\, \mathbb{E}[J^{\tau',\pi}[s]]), \qquad \forall s \in \mathcal{S}. \tag{7}$$

The question of computing the value of the game remains. To this end, we now prove that repeatedly applying T produces a sequence that converges to the value. In particular, the game has a *fixed point property* which is stated in the following:

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

$$J^\star = TJ^\star \;\; \textit{and} \;\; \lim_{n\to\infty} T^n J = J^\star. \tag{8}$$

Theorem 2 establishes the existence of a fixed point of T and that the fixed point coincides with the value of the game. Crucially, it suggests that $J^\star$ can be computed by an iterative application of the Bellman operator which underpins a value iterative method. We study this aspect in Sec. 4 where we develop an iterative scheme for computing $J^\star$.

Definition 2. *The pair $(\hat{\tau}, \hat{\pi}) \in \mathscr{V} \times \Pi$ is an SPE iff:*

$$J^{\hat{\tau},\hat{\pi}}[s] = \max_{\pi \in \Pi} J^{\hat{\tau},\pi}[s] = \min_{\tau \in \mathscr{V}} J^{\tau,\hat{\pi}}[s], \qquad \forall s \in \mathcal{S}. \tag{9}$$

An SPE therefore defines a strategic configuration in which both players play their BR strategies. With reference to the FT RL problem, an SPE describes a scenario in which the controller optimally responds against stoppages at the set of states that inflict the greatest costs to the controller. In particular, we will demonstrate that $\hat{\pi} \in \Pi$ is a BR to a system that undergoes adversarial attacks.

Proposition 1. *The pair $(\hat{\tau}, \hat{\pi}) \in \mathscr{V} \times \Pi$ consists of BR strategies and constitutes an SPE.*

By Prop. 1, when the pair $(\hat{\tau}, \hat{\pi})$ is played, each player executes its BR strategy. The strategic response then induces FT behaviour by the controller. We now turn to the existence and characterising the optimal stopping time for player II. The following result establishes its existence.

Theorem 3. *There exists an $\mathscr{F}$-measurable stopping time:*

$$\hat{\tau} = \min\left\{ k \in \mathcal{T} \,\middle|\, G(s_k) \leq \min_{v \in \mathscr{V}} \max_{\pi \in \Pi} J^{v,\pi}[s_k] \right\}, \;\; a.s.$$

The theorem characterises and establishes the existence of the player II optimal stopping time which, when executed by the adversary, induces an FT control by the controller.

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

Theorem 4. *Let $\hat{\tau}$ be the player II 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 hence, any method that computes $\hat{\tau}$ for the SG yields a solution to the robust OSP.

4 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.

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

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

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:

$$Q^\star(s) = \max_{a\in\mathcal{A}} R^a_s + \gamma P J^\star[s], \qquad \forall s \in \mathcal{S} \tag{10}$$

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 $\Phi r^\star$, in turn approximates $Q^\star$. To complete the connection, we provide a bound between the outcome of the game when the players use controls generated by the algorithm.

We introduce our player II stopping criterion which now takes the form:

$$\hat{\tau} = \min\{t | G(s_t) \le Q^\star(s_t)\}. \tag{11}$$

Let us define a orthogonal projection Π and the function F by the following:

$$\Pi Q := \underset{\bar{Q}\in\{\Phi r | r\in\mathbb{R}^p\}}{\arg\min} \|\bar{Q} - Q\|, FQ := \max_{a\in\mathcal{A}} R^a_s + \gamma P \min\{G, Q\}. \tag{12}$$

We now state the main results of the section:

Theorem 5. *r_n converges to $r^\star$ where $r^\star$ is the unique solution: $\Pi F(\Phi r^\star) = \Phi r^\star$.*

The following results provide approximation bounds when employing the projection Π:

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

$$\|\Phi r^\star - Q^\star\| \le \left(\sqrt{1-\gamma^2}\right)^{-1} \|\Pi Q^\star - Q^\star\|, \tag{13}$$

$$\mathbb{E}\left[J^\star - J^{\tilde{\tau},\tilde{\pi}}\right] \le 2\Big[(1-\gamma)\sqrt{1-\gamma^2})\Big]^{-1} \|\Pi Q^\star - Q^\star\|. \tag{14}$$

Hence the error bound in approximation of $J^\star$ is determined by the goodness of the projection.

Theorem 5 and Theorem 6 thus enable the FT RL problem to be solved by way of simulating the behaviour of the environment and using the update rule (10) to approximate the value function. Applying the stopping rule in (11), by Theorem 6 and Theorem 2, means the pair $(\tilde{\tau}, \tilde{\pi})$ is generated where the policy $\tilde{\pi}$ approximates the policy $\hat{\pi}$ which is FT against adversarial stoppages and faults.

CONCLUSION

In this paper, we tackled the problem of fault-tolerance within RL in which the controller seeks to obtain a control that is robust against catastrophic failures. To formally characterise the optimal behaviour, we constructed a new discrete-time 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 and constructed an approximate dynamic programming algorithm so that the game can be solved by simulation.

REFERENCES

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. RL–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 SGs. *Mathematical Programming* 50, 1-3 (1991), 227–237.

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

Itamar Arel, Cong Liu, T Urbanik, and AG Kohls. 2010. RL-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 RL. *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 RL 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. SGs 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 RL. 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, RL for autonomous driving. *arXiv preprint arXiv:1610.03295* (2016).

Lloyd S Shapley. 1953. SGs. *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.

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 $k \geq 0$ where $s := s_0$.

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 J are square integrable: that is, $R, G \in \mathbb{L}_2(\mu)$.

ADDITIONAL LEMMATA

We begin the analysis with some preliminary lemmata and definitions which are useful for proving the main results.

Definition A.1. *An operator $T : \mathcal{V} \to \mathcal{V}$ 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 $V_1, V_2 \in \mathcal{V}$ we have that:*

$$\|TV_1 - TV_2\| \leq c\|V_1 - V_2\|. \tag{15}$$

Definition A.2. *An operator $T : \mathcal{V} \to \mathcal{V}$ is **non-expansive** if $\forall V_1, V_2 \in \mathcal{V}$ we have:*

$$\|TV_1 - TV_2\| \leq \|V_1 - V_2\|. \tag{16}$$

Definition A.3. *The **residual** of a vector $V \in \mathcal{V}$ w.r.t the operator $T : \mathcal{V} \to \mathcal{V}$ is:*

$$\epsilon_T(V) := \|TV - V\|. \tag{17}$$

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

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

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

$$\|\max\{f, h\} - \max\{g, h\}\| \leq \|f - g\|. \tag{18}$$

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

Lemma A.4. *The probability transition kernel P is non-expansive, that is:*

$$\|PV_1 - PV_2\| \leq \|V_1 - V_2\|. \tag{19}$$

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.

Proposition A.1. *The operator T in (5) is a contraction.*

Lemma A.5. *Let $T : \mathcal{V} \to \mathcal{V}$ be a contraction mapping in $\|\cdot\|$ and let $J^\star$ be a fixed point so that $TJ^\star = J^\star$ then there exists a constant $c \in [0, 1[$ s.th:*

$$\|J^\star - J\| \leq (1-c)^{-1}\epsilon_T(J). \tag{20}$$

Lemma A.6. *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:*

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

Lemma A.7. *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 $I(s) \equiv \mathbf{1}$ be the unit function, then for any $J \in \mathbb{L}_2$ and for any scalar $\alpha \in \mathbb{R}$, T satisfies $T(J + \alpha I)(s) = TJ(s) + \alpha I(s)$.*

PROOF OF RESULTS

Proof of Lemma A.1. 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 $b \in \mathcal{V}$:

$$\left|\max_{a\in\mathcal{V}} f(a,b) - \max_{a\in\mathcal{V}} g(a,b)\right| \leq \max_{a\in\mathcal{V}} \left|f(a,b) - g(a,b)\right|. \tag{21}$$

From (21) we can straightforwardly derive the fact that for any $b \in \mathcal{V}$:

$$\left|\min_{a\in\mathcal{V}} f(a,b) - \min_{a\in\mathcal{V}} g(a,b)\right| \leq \max_{a\in\mathcal{V}} \left|f(a,b) - g(a,b)\right|, \tag{22}$$

(this can be seen by negating each of the functions in (21) and using the properties of the $\max$ operator).

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

$$\max_{a\in\mathcal{V}} \left|f(a,b) - g(a,b)\right| \leq c \tag{23}$$

Since (22) holds for any $b \in \mathcal{V}$ and, by (21), we have in particular that

$$\begin{aligned}
&\left|\max_{b\in\mathcal{V}} \min_{a\in\mathcal{V}} f(a,b) - \max_{b\in\mathcal{V}} \min_{a\in\mathcal{V}} g(a,b)\right| \\
&\leq \max_{b\in\mathcal{V}} \left|\min_{a\in\mathcal{V}} f(a,b) - \min_{a\in\mathcal{V}} g(a,b)\right| \\
&\leq \max_{b\in\mathcal{V}} \max_{a\in\mathcal{V}} \left|f(a,b) - g(a,b)\right| \leq c,
\end{aligned} \tag{24}$$

whenever (23) holds which gives the required result. □

Lemma A.2 and Lemma A.3 are given without proof but can be straightforwardly checked.

Proof of Lemma A.4. 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 Proposition A.1. We wish to prove that:

$$\|TJ - T\bar{J}\|_\pi \leq \gamma \|J - \bar{J}\|. \tag{25}$$

Firstly, we observe that:

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

using Cauchy-Schwartz (and that $\gamma \in [0,1[$) and (30). The result follows after applying Lemma A.2 and Lemma A.3. □

Proof of Lemma A.5. The proof follows almost immediately from the triangle inequality, indeed for any $J \in \mathbb{L}_2$:

$$\|J^\star - J\| = \|TJ^\star - J\| \leq \gamma \|J^\star - J\| + \|TJ - J\|, \tag{26}$$

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 A.6. The proof follows directly from Lemma A.5. Indeed, we observe that for any $J \in \mathbb{L}_2$ we have

$$\|J_1^\star - J_2^\star\| \leq \|J_1^\star - J\| + \|J_2^\star - J\|, \tag{27}$$

where we have added and subtracted J to produce the inequality. The result then follows from Lemma A.5. □

Proof of Lemma A.7. 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

$$\begin{aligned}\min_{\tau\in\mathcal{T}}\left\{ \max_{a\in A} R_s^a + \gamma \sum_{s'\in\mathcal{S}} P_{ss'}^a J^{\tau,\pi}[s'], G(S_\tau)\right\} \\ - \min_{\tau\in\mathcal{T}}\left\{ \max_{a\in A} R_s^a + \gamma \sum_{s'\in\mathcal{S}} P_{ss'}^a \bar{J}^{\pi}[s'], G(S_\tau)\right\} \leq 0\end{aligned} \tag{28}$$

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

$$x \leq y \implies \min\{x,h\} \leq \min\{y,h\}. \tag{29}$$

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

$$\left|\max_{x\in\mathcal{V}} f(x) - \max_{x\in\mathcal{V}} g(x)\right| \leq \max_{x\in\mathcal{V}} |f(x) - g(x)|. \tag{30}$$

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

$$\begin{aligned}&\max_{a\in\mathcal{A}}\left\{ R_s^a + \gamma \sum_{s'\in\mathcal{S}} P_{ss'}^a J^{\tau,\pi}[s']\right\} - \max_{a\in\mathcal{A}}\left\{ R_s^a + \gamma \sum_{s'\in\mathcal{S}} P_{ss'}^a \bar{J}^{\pi}[s']\right\} \\ &\leq \gamma \max_{a\in A}\left\{ \sum_{s'\in\mathcal{S}} P_{ss'}^a \left(J^{\tau,\pi}[s'] - \bar{J}^{\pi}[s']\right)\right\} \\ &= \gamma\left((PJ) - \left(P\bar{J}\right)\right) \leq J - \bar{J} \leq 0,\end{aligned} \tag{31}$$

where we have used (30) in the penultimate line. The result immediately follows after applying (29). □

Proof of Theorem 1. We begin by noting the following inequality holds:

$$\mathrm{val}^{+}[J] = \min_{\tau\in\mathcal{T}} \max_{\pi\in\Pi} \mathbb{E}[J^{\tau,\pi}[s]] \geq \max_{\pi\in\Pi} \min_{\tau\in\mathcal{T}} \mathbb{E}[J^{\tau,\pi}[s]] = \mathrm{val}^{-}[J]. \tag{32}$$

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

The proof can now be settled by reversing the inequality in (32). 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:

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

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

$$\mathbb{E}[J^{\tau,\hat{\pi}}[s]] \geq \mathbb{E}[J^{\tau,\pi}[s]]. \tag{33}$$

Now since (33) holds $\forall \pi \in \Pi$ we find that:

$$\mathbb{E}[J^{\tau,\hat{\pi}}[s]] \geq \max_{\pi\in\Pi} \mathbb{E}[J^{\tau,\pi}[s]]. \tag{34}$$

Lastly, applying $\min$ operator we observe that:

$$\mathbb{E}[J^{\hat{\tau},\hat{\pi}}[s]] \geq \min_{\tau\in\mathcal{T}} \max_{\pi\in\Pi} \mathbb{E}[J^{\tau,\pi}[s]] = \mathrm{val}^{+}[J]. \tag{35}$$

It now remains to show the reverse inequality holds:

$$\mathbb{E}[J^{\hat{\tau},\hat{\pi}}[s]] \leq \max_{\pi\in\Pi} \min_{\tau\in\mathcal{T}} \mathbb{E}[J^{\tau,\pi}[s]] = \mathrm{val}^{-}[J]. \tag{36}$$

Indeed, we observe that

$$\mathbb{E}\left[J^{\hat{\tau},\hat{\pi}}[s]\right] \leq \min_{\tau\in\mathcal{T}} \mathbb{E}\left[J^{\tau\wedge m,\hat{\pi}}[s]\right] + \mathbb{E}\left[\sum_{t=m}^{\infty} \gamma^t \left(|R(s_t,a_t)| + |G(s_t)|\right)\right] \tag{37}$$

$$\leq \lim_{m\to\infty} \left[\min_{\tau\in\mathcal{T}} \mathbb{E}\left[J^{\tau\wedge m,\hat{\pi}}[s]\right] + c(m)\right] \tag{38}$$

$$= \min_{\tau\in\mathcal{T}} \mathbb{E}\left[J^{\tau,\hat{\pi}}[s]\right] \leq \max_{\pi\in\Pi} \min_{\tau\in\mathcal{T}} \mathbb{E}\left[J^{\tau,\pi}[s]\right], \tag{39}$$

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 (39) we have that:

$$\mathbb{E}\left[J^{\hat{\tau},\hat{\pi}}[s]\right] \leq \max_{\pi\in\Pi}\min_{\tau\in\mathcal{T}} \mathbb{E}\left[J^{\tau,\pi}[s]\right] = \mathrm{val}^{-}[J]. \tag{40}$$

Hence putting (35) and (40) together gives:

$$\begin{aligned}\mathrm{val}^{-}[J] &= \max_{\pi\in\Pi}\min_{\tau\in\mathcal{T}} \mathbb{E}\left[J^{\tau,\pi}[s]\right] \\ &\geq \mathbb{E}[J^{\hat{\tau},\hat{\pi}}[s]] \geq \min_{\tau\in\mathcal{T}}\max_{\pi\in\Pi} \mathbb{E}[J^{\tau,\pi}[s]] = \mathrm{val}^{+}[J].\end{aligned} \tag{41}$$

After combining (41) with (32) we deduce the thesis. □

Proof of Prop. 1. 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, $\hat{\tau}$ 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. □

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

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

We begin the proof by invoking similar reasoning as (37) - (38) 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,

$$T^n J[s] \leq \max_{\pi\in\Pi}\min_{\tau\in\mathcal{T}} \mathbb{E}\left[J^{\tau,\pi}[s]\right] + \frac{\gamma^n}{1-\gamma}c = J^{\star}[s] + \frac{\gamma^n}{1-\gamma}c. \tag{42}$$

By analogous reasoning we can deduce that:

$$T^n J[s] \geq \min_{\tau\in\mathcal{T}}\max_{\pi\in\Pi} \mathbb{E}\left[J^{\tau,\pi}[s]\right] - \frac{\gamma^n}{1-\gamma}c = J^{\star}[s] - \frac{\gamma^n}{1-\gamma}c. \tag{43}$$

Putting (42) and (43) together implies:

$$J^{\star}[s] - \frac{\gamma^n}{1-\gamma}c \leq T^n J[s] \leq J^{\star}[s] + \frac{\gamma^n}{1-\gamma}c. \tag{44}$$

By Lemma A.7, i.e. invoking the monotonicity and constant shift properties of T, we can apply T to (44) and preserve the inequalities to give:

$$TJ^{\star}[s] - \frac{\gamma^n}{1-\gamma}c \leq T^{n+1} J[s] \leq TJ^{\star}[s] + \frac{\gamma^n}{1-\gamma}c. \tag{45}$$

After taking the limit in (45) and, using the sandwich theorem of calculus, we deduce the result. □

Proof of Theorem 3. For any $m \in \mathbb{N}$ we have that:

$$\max_{\pi\in\Pi} J^{\tau,\pi}[s] \geq \max_{\pi\in\Pi} J^{\tau\wedge m,\pi}[s] - \sum_{t=m}^{\infty} \gamma^t \max_{\pi\in\Pi} \left(|R(s_t, a_t)| + |G(s_t)|\right). \tag{46}$$

We now apply the $\min$ operator to both sides of (46) 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:

$$\mathbb{E}\left[\min_{\tau\in\mathcal{T}}\max_{\pi\in\Pi} J^{\tau,\pi}[s]\right] \tag{47}$$

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

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

$$\begin{aligned}&\mathbb{E}\left[\max_{\pi\in\Pi}\left(|R(s_t,a_t)| + |G(s_t)|\right)\right]\\ &\geq \max_{\pi\in\Pi}\mathbb{E}\left[\left(|R(s_t,a_t)| + |G(s_t)|\right)\right] = \mathbb{E}[|R(s_0,\cdot)|] + \mathbb{E}[|G(s_0)|].\end{aligned} \tag{49}$$

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:

$$\min_{\tau\in\mathcal{T}}\max_{\pi\in\Pi} J^{\tau\wedge m,\pi}[s] = T^m G(s). \tag{50}$$

Inserting (49) and (50) into (48) gives:

$$\begin{aligned}\mathbb{E}\left[\min_{\tau\in\mathcal{T}}\max_{\pi\in\Pi} J^{\tau,\pi}[s]\right] &\geq \mathbb{E}\left[T^m G(s)\right] - c(m)\\ &= \lim_{m\to\infty}\left[\mathbb{E}\left[T^m G(s)\right] - c(m)\right] = \mathbb{E}\left[J^{\hat{\tau},\hat{\pi}}[s]\right],\end{aligned} \tag{51}$$

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:

$$\mathbb{E}\left[J^{\hat{\tau},\hat{\pi}}[s]\right] \leq \mathbb{E}\left[\min_{\tau\in\mathcal{T}}\max_{\pi\in\Pi} J^{\tau,\pi}[s]\right], \tag{52}$$

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

The proofs of the results in Sec. 4 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. 4.

Proof of Theorem 5. We firstly notice the construction of $\hat{\tau}$ given by

$$\hat{\tau} = \min\{t|G(s_t) \leq Q^\star\}, \tag{53}$$

is sensible since we observe that

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

Result 1
Step 1 Our first step is to prove the following bound:

$$\left\|FQ - F\bar{Q}\right\| \leq \gamma\left\|Q - \bar{Q}\right\|. \tag{54}$$

Proof.

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

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:

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

□

Step 3

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

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

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

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

□

Step 4

$$\left\|\Phi r^\star - Q^\star\right\| \leq \frac{1}{\sqrt{1-\gamma^2}} \left\|\Pi Q^\star - Q^\star\right\|. \tag{55}$$

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

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

which is the required result. □

Result 2

$$\mathbb{E}\left[J^{\star}[s]\right] - \mathbb{E}\left[J^{\tilde{\tau},\tilde{\pi}}[s]\right] \leq \frac{2}{[(1-\gamma)\sqrt{1-\gamma^2}]}\|\Pi Q^{\star} - Q^{\star}\|. \tag{56}$$

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

$$\begin{aligned}
&\mathbb{E}\left[J^{\star}[s]\right] - \mathbb{E}\left[J^{\tilde{\tau},\tilde{\pi}}[s]\right] \\
&= \mathbb{E}\left[PJ^{\star}[s]\right] - \mathbb{E}\left[PJ^{\tilde{\tau},\tilde{\pi}}[s]\right] \\
&\leq \left|\mathbb{E}\left[PJ^{\star}[s]\right] - \mathbb{E}\left[PJ^{\tilde{\tau},\tilde{\pi}}[s]\right]\right| \\
&\leq \|PJ - PJ^{\tilde{\tau},\tilde{\pi}}\|.
\end{aligned} \tag{57}$$

Inserting the definitions of $Q^{\star}$ and $\tilde{Q}$ into (57) then gives:

$$\mathbb{E}\left[J^{\star}[s]\right] - \mathbb{E}\left[J^{\tilde{\tau},\tilde{\pi}}[s]\right] \leq \frac{1}{\gamma}\|Q^{\star} - \tilde{Q}\|. \tag{58}$$

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

$$\|Q^{\star} - \tilde{Q}\| \leq \|Q^{\star} - F(\Phi r^{\star})\| + \|\tilde{Q} - F(\Phi r^{\star})\| \tag{59}$$

$$\leq \gamma\left\{\|Q^{\star} - \Phi r^{\star}\| + \|\tilde{Q} - \Phi r^{\star}\|\right\} \tag{60}$$

$$\leq \gamma\left\{2\|Q^{\star} - \Phi r^{\star}\| + \|Q^{\star} - \tilde{Q}\|\right\}. \tag{61}$$

So that

$$\|Q^{\star} - \tilde{Q}\| \leq \frac{2\gamma}{1-\gamma}\|Q^{\star} - \Phi r^{\star}\|. \tag{62}$$

The result then follows after substituting the result of step 4 (55). □

Let us now define the following quantity:

$$HQ(s) := \begin{cases} G(s) & \text{if } G(s) \leq (\Phi r^{\star})(s) \\ Q(s) & \text{otherwise,} \end{cases} \tag{63}$$

and

$$\tilde{F}Q := \max_{a \in \mathcal{A}} R_s^a + \gamma PHQ. \tag{64}$$

Step 5

$$\left\|\tilde{F}Q - \tilde{F}\bar{Q}\right\| \leq \gamma\left\|Q - \bar{Q}\right\| \tag{65}$$

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 $\tilde{Q} = \max_{a \in \mathcal{A}} R_s^a + \gamma PJ^{\pi,\tilde{\tau}}$ is a fixed point.

$$H\tilde{Q} = H\left(\max_{a \in \mathcal{A}} R_s^a + \gamma PJ^{\pi,\tilde{\tau}}\right)$$

$$= \begin{cases} G(s) & \text{if } G(s) \leq (\Phi r^\star)(s) \\ \max\limits_{a \in \mathcal{A}} R_s^a + \gamma P J^{\pi,\tilde{\tau}} & \text{otherwise} \end{cases}$$
$$= J^{\pi,\tilde{\tau}}$$

□

Let us now define the following quantity:

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

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:

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

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

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

□

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 $\min$. □