

Part I

Appendix

Table of Contents

9	Full Algorithms	13
10	Fuel Budget	15
11	Ablation studies	15
11.1	Ablation Study 2: Testing Different functional forms of the intervention costs . . .	15
11.2	Ablation Study 3: Benefits of LICRA in Smaller Intervention Region	18
12	Analysis of LICRA Q-learning Variant	19
13	Hyperparameter Settings	20
14	Notation & Assumptions	21
15	Proof of Technical Results	21

9 FULL ALGORITHMS

In this supplementary section, we explicitly define two versions of the LICRA algorithm. Algorithm 2 describes the version where both agents use PPO. PPO_update() subroutine is a standard PPO gradient update done as in Algorithm 1 of (Schulman et al., 2017) with clipping surrogate objective with parameter ϵ . The gradient update utilises batch size B , stepsize α and performs T update steps per episode. Algorithm 3 defines the LICRA_SAC version, utilising SAC for the switching agent. Here SAC_update() is analogously a standard soft actor-critic update done as in Algorithm 1 of (Haarnoja et al., 2018), where B , α and T play the same role as in the PPO update.

Algorithm 2: LICRA with PPO

```

1: Input: Stepsize  $\alpha$ , batch size  $B$ , episodes  $K$ , steps per episode  $T$ , mini-epochs  $e$ ,
   clipping-parameter  $\epsilon$ 
2: Initialise: Policy network (acting)  $\pi$ , Policy network (switching)  $\mathfrak{g}$ ,
   Critic network (acting)  $V_\pi$ , Critic network (switching)  $V_g$ 
3: Given reward objective function,  $R$ , initialise Rollout Buffers  $\mathcal{B}_\pi, \mathcal{B}_g$ 
4: for  $N_{episodes}$  do
5:   Reset state  $s_0$ , Reset Rollout Buffers  $\mathcal{B}_\pi, \mathcal{B}_g$ 
6:   for  $t = 0, 1, \dots$  do
7:     Sample  $a_t \sim \pi(\cdot|s_t)$ 
8:     Sample  $g_t \sim \mathfrak{g}(\cdot|s_t)$ 
9:     if  $g_t = 1$  then
10:      Apply  $a_t$  so  $s_{t+1} \sim P(\cdot|a_t, s_t)$ ,
11:      Receive rewards  $r_t = \mathcal{R}(s_t, a_t)$ 
12:      Store  $(s_t, a_t, s_{t+1}, r_t)$  in  $\mathcal{B}_\pi$ 
13:     else
14:      Apply the null action so  $s_{t+1} \sim P(\cdot|0, s_t)$ ,
15:      Receive rewards  $r_t = \mathcal{R}(s_t, 0)$ .
16:     end if
17:     Store  $(s_t, g_t, s_{t+1}, r_t)$  in  $\mathcal{B}_g$ 
18:   end for
19:   // Learn the individual policies
20:   PPO_update( $\pi, V_\pi, \mathfrak{B}_\pi, B, \alpha, T$ )
21:   PPO_update( $\mathfrak{g}, V_g, \mathfrak{B}_g, B, \alpha, T$ )
22: end for

```

Algorithm 3: LICRA with SAC

```

1: Input: Step size  $\alpha$ , batch size  $B$ , episodes  $K$ , steps per episode  $T$ , mini-epochs  $e$ 
2: Initialise: Policy network (acting)  $\pi$ , Policy network (switching)  $\mathfrak{g}$ ,
   Critic network (acting)  $V_\pi$ , Q-Critic network (switching)  $Q_\mathfrak{g}$ , V-Critic network (switching)  $V_\mathfrak{g}$ 
3: Given reward objective function,  $R$ , initialise Rollout Buffer  $\mathcal{B}_\pi$  Replay Buffer  $\mathcal{B}_\mathfrak{g}$ 
4: for  $N_{episodes}$  do
5:   Reset state  $s_0$ , Reset Rollout Buffer  $\mathcal{B}_\pi$ 
6:   for  $t = 0, 1, \dots$  do
7:     Sample  $a_t \sim \pi(\cdot|s_t)$ 
8:     Sample  $g_t \sim \mathfrak{g}(\cdot|s_t)$ 
9:     if  $g_t = 1$  then
10:      Apply  $a_t$  so  $s_{t+1} \sim P(\cdot|a_t, s_t)$ ,
11:      Receive rewards  $r_t = \mathcal{R}(s_t, a_t)$ 
12:      Store  $(s_t, a_t, s_{t+1}, r_t)$  in  $\mathcal{B}_\pi$ 
13:     else
14:      Apply the null action so  $s_{t+1} \sim P(\cdot|0, s_t)$ ,
15:      Receive rewards  $r_t = \mathcal{R}(s_t, 0)$ .
16:     end if
17:     Store  $(s_t, g_t, s_{t+1}, r_t)$  in  $\mathcal{B}_\mathfrak{g}$ 
18:   end for
19:   // Learn the individual policies
20:   PPO_update( $\pi, V_\pi, \mathfrak{B}_\pi, B, \alpha, T$ )
21:   Sample a batch of  $|\mathcal{B}_\mathfrak{g}|$  transitions  $B_\mathfrak{g}$  from  $\mathcal{B}_\mathfrak{g}$ 
22:   SAC_update( $\mathfrak{g}, V_\mathfrak{g}, Q_\mathfrak{g}, B_\mathfrak{g}, B, \alpha, T$ )
23: end for

```

10 FUEL BUDGET

We test the ability of LICRA (LICRA_SAC) in a budgeted Drive environment. We modified the Drive environment in Sec. 7 so that there is an additional scarce fuel level for the controlled vehicle. This fuel level decreases in proportion to the magnitude of the action taken by the agent. If the controlled vehicle runs out of fuel then it receives a large negative reward, and therefore it is important that the agent learns to use minimal acceleration/braking to reach the final destination.

Fig. 5 shows the performance of LICRA and corresponding baselines. The majority of the baselines fail to learn anything in the environment, which we expect is due to the difficult of exploration due to low fuel levels (LICRA can counter this by exploring using the null action) and how easy it is to run out of fuel. In some seeds CPO is able to solve the environment to the same level as LICRA, but is not as stable over seeds and it is not as fast as LICRA in improving performance.

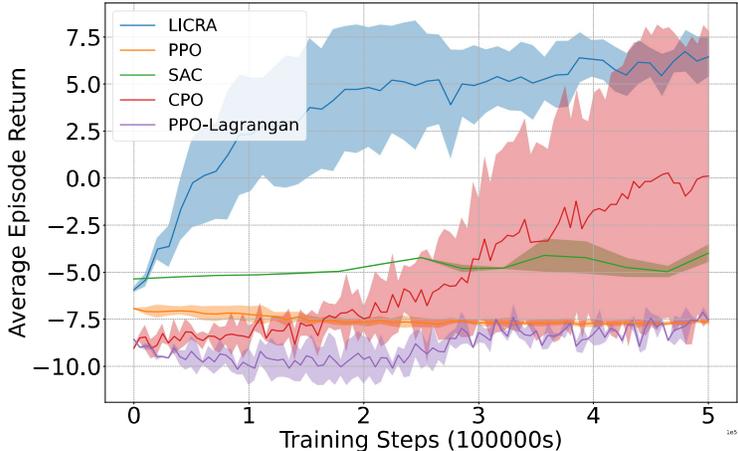


Figure 5

11 ABLATION STUDIES

Fig. 3 shows the performance of LICRA variants and baselines when $c(s, a) \equiv 5$. In this section, we analyse LICRA’s ability to handle various functional forms of the intervention cost.

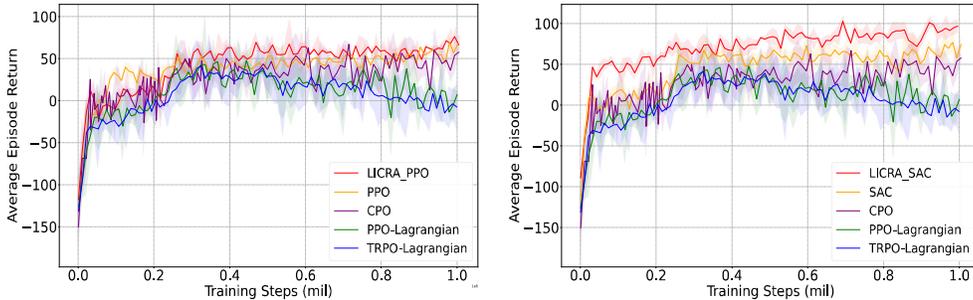
11.1 ABLATION STUDY 2: TESTING DIFFERENT FUNCTIONAL FORMS OF THE INTERVENTION COSTS

Intervention costs of the form $c(s, a) \equiv 0$.

In Sec. 4, we claimed that LICRA’s impulse control mechanism which first determines the set of states to perform actions then only learns the optimal actions at these states can induce a much quicker learning process. In particular, we claimed that LICRA enables the RL agent to rapidly learn which states to focus on to learn optimal actions.

In our last ablation analysis, we test LICRA’s ability to handle the case in which the agent faces no fixed costs so $c(s, a) \equiv 0$, therefore deviating from the form of the impulse control objective (1). In doing so, we test the ability of LICRA to prioritise the most important states for learning the correct actions in a general RL setting. As shown in Fig. 6 we present the average returns when $c(s, a) \equiv 0$. For this case, LICRA_SAC both learns the fastest indicating the benefits of the impulse control framework even in the setting in which the agent does not face fixed minimal costs for each action. Strikingly, LICRA also achieves the highest performance which demonstrates that LICRA also improves overall performance in complex tasks.

Intervention costs of the form $c(s, a) \equiv k > 0$.

Figure 6: Ablation Study when $c(s, a) \equiv 0$ in Lunar Lander.

We first analyse the behaviour of LICRA and leading baselines when the environment has intervention costs of the form $c(s, a) = k$ where the fixed part $k > 0$ is a strictly positive constant. Intervention costs of this kind are frequently found within economic settings where they are characterised as ‘menu costs’. This name derives from the fixed costs associated to a vendor adjusting their prices and serves as an explanation for price rigidities within the macroeconomy (Caplin & Spulber, 1987; Mguni, 2018c).

We next test the average returns when $c(s, a) \equiv 10$, $c(s, a) \equiv 20$, as shown in Fig. 7. Note that the action space is discrete and the intervention costs only occur when $a \neq 0$. In this case, LICRA_SAC both learns the fastest and achieves the highest performance. Moreover, unlike PPO which produces declining performance and does not converge, LICRA_PPO converges to a high reward stable point.

Intervention costs of the form $c(s, a) = k + \lambda a$.

We next analyse the behaviour of LICRA and leading baselines when the environment has intervention costs of the form $c(s, a) = k + \lambda a$ where the fixed part and proportional part $\lambda, k > 0$ are strictly positive constants. Intervention costs of this kind are frequently found within financial settings in which an investor incurs a fixed cost for investment decisions e.g. broker costs (Davis & Norman, 1990; Mguni, 2018b). Note that the action space is discrete and the intervention costs only occur when $a \neq 0$.

We present the average returns when $c(s, a) = 5 + |a|$, $c(s, a) = 5 + 5 \cdot |a|$ in Fig. 7. As with previous case, LICRA_SAC both learns the fastest and achieves the highest performance. Moreover, unlike PPO which produces declining performance and does not converge, LICRA_PPO converges to a high reward stable point.

Intervention costs of the form $c(s, a) = k + f(s, a)$.

In general, the intervention costs incurred by an agent for each intervention can be allowed to be a function of the state. For example, activating an actuator under adverse environment conditions may incur greater wear to machinery than in other conditions. The functional form of the cost function is the most general and produces the most complex decision problem out of the aforementioned cases. To capture this general case, we lastly analysed the behaviour of LICRA and leading baselines when the environment has intervention costs of the form $c(s, a) = k + f(s, a)$ where the fixed part $k > 0$ and $f : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}_{>0}$ is a positive function.

As shown in Fig. 7, for this case we present the average returns when $c(s, a) = 5 + |d_s - d_{\text{target}}| \cdot |a|$, where $|d_s - d_{\text{target}}|$ represents the distance between the current position to the destination (determined by the state s). As before, the action space is discrete and the intervention costs only occur when $a \neq 0$. As with previous cases, LICRA_SAC both learns the fastest and achieves the highest performance, demonstrating LICRA’s ability to solve the more complex task. Moreover, as in the previous cases, unlike PPO which produces declining performance and does not converge, LICRA_PPO converges to a high reward stable point.

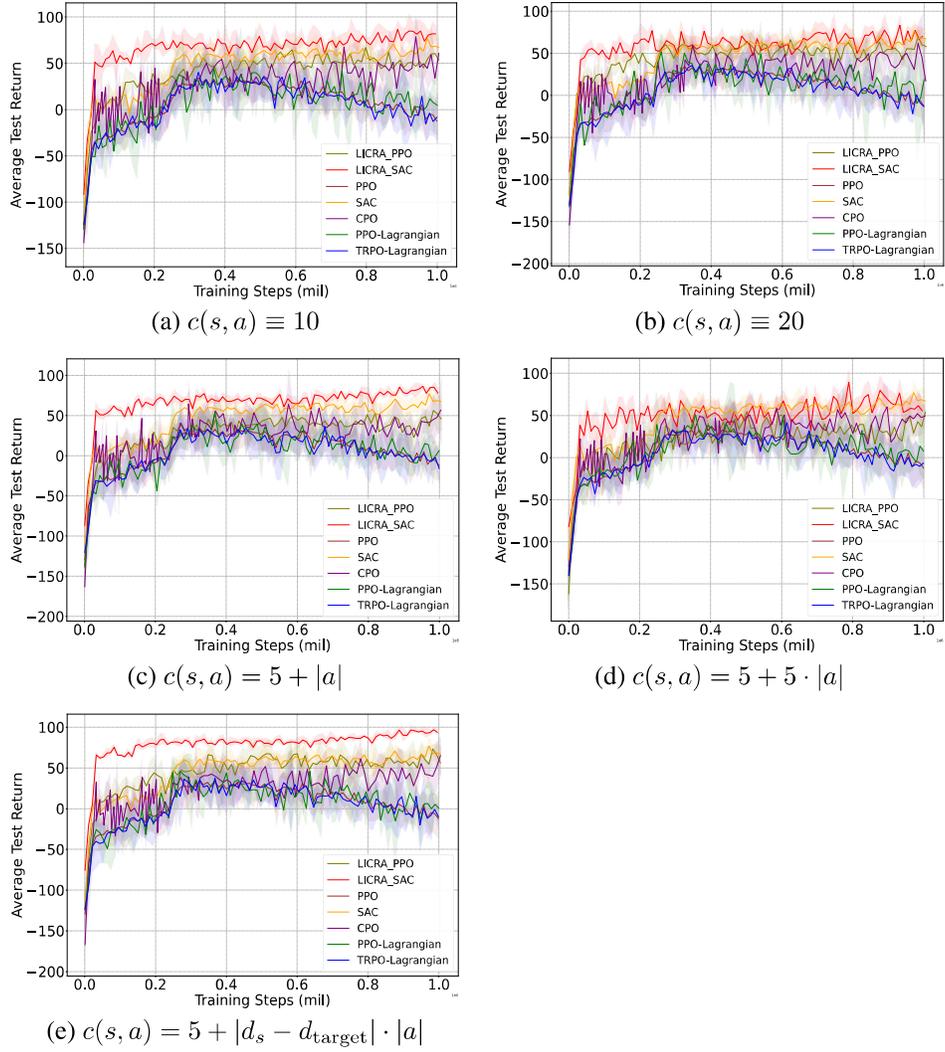


Figure 7: Ablation Study when (a) $c(s, a) \equiv 10$, (b) $c(s, a) \equiv 20$, (c) $c(s, a) = 5 + |a|$, (d) $c(s, a) = 5 + 5 \cdot |a|$ and (e) $c(s, a) = 5 + |d_s - d_{\text{target}}| \cdot |a|$ in Lunar Lander.

11.2 ABLATION STUDY 3: BENEFITS OF LICRA IN SMALLER INTERVENTION REGION

In Sec. 4, we claimed that LICRA enables efficient learning in setting in which the set of states in which the agent should act, which we call the *intervention region* is a subset of the state space. Moreover, we claimed that this advantage over existing methods is increased as the intervention region becomes small in relation to the entire state space.

To test these claims, we modified the Drive environment in Sec. 7 to a problem setting in which the agent is required to bring a moving vehicle to rest in a particular subregion of the lane or *stop gap* which we denote by $\mathcal{S}_I \subset \mathcal{S}$. If the agent brings the vehicle to rest within \mathcal{S}_I the agent receives a reward $R > 0$ and the episode terminates. If however the agent brings the vehicle to rest outside of \mathcal{S}_I the agent receives a lower reward $r < R$ the episode terminates. Lastly, if the agent fails to bring the vehicle to rest before the end of the lane the episode terminates and the agent receives 0 reward. The length of the entire lane \mathcal{S} is 500 units and we ablate the size of the region \mathcal{S}_I in which the agent is required to stop to receive the maximum reward. Now the agent gets to decide a magnitude which decelerates the vehicle i.e. how heavily to brake (at any given point, the agent can also choose not to brake at all). Each deceleration $a \in [0, 1]$ incurs a fixed minimal cost of $c(s, a) = \kappa + \lambda a$ where $\kappa, \lambda > 0$.

Fig. 8 shows the results of the ablation on the stop gap \mathcal{S}_I when \mathcal{S}_I is a length of 50 units or 10% of the entire state space \mathcal{S} through to $\mathcal{S}_I \equiv \mathcal{S}$ i.e. when the intervention region is the entire state space.

As can be observed in Fig. 8, when the intervention region is comparatively small, LICRA_PPO produces a significant performance advantage over the base learner PPO. This performance gap is gradually decreased as the size of the stop gap increases and eventually becomes the entire state space (which is the case when the stop gap is 500). Interestingly, LICRA_PPO still maintains a performance advantage over PPO even when $\mathcal{S}_I \equiv \mathcal{S}$.

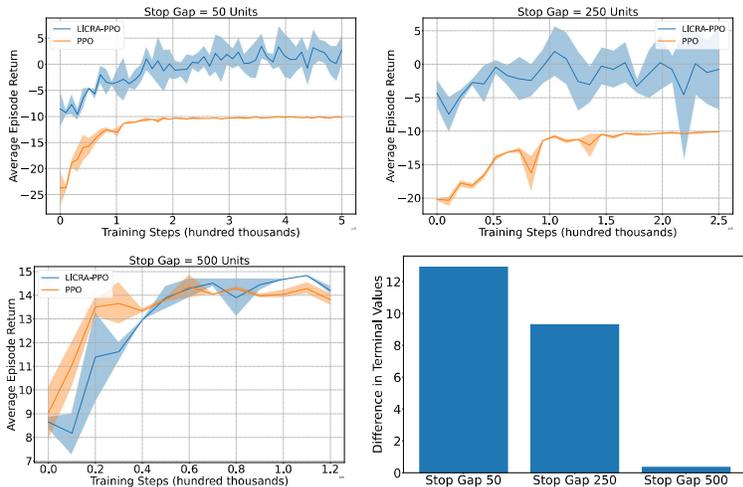


Figure 8: Ablation on the intervention region. The 'Stop Gap' represents the intervention region in our modified Drive environment. When the Stop Gap is 50 units, the intervention region is 10% of the entire state space. When the Stop Gap is 509 units, the intervention region is the entire state space.

12 ANALYSIS OF LICRA Q-LEARNING VARIANT

In order to validate the convergence of the Q-learning variant of LICRA (c.f. (4)), we ran an experiment where the LICRA Q-learning variant given in (4) can choose whether to act or not to act in a given discrete state (and continuous action space). In keeping with the problem setting we consider, there is a cost associated with any non-zero action. We consider a 6-by-6 grid world problem, where the start state is $(0, 0)$ and the algorithm receives a positive reward only if it reaches the point $(5, 5)$. The action space is two-dimensional and continuous and the agent is moved in x and y dimensions by the number of grid tiles corresponding to the value of action projected to the nearest integer. Every non-zero action is associated with a cost of 1. Additionally, the region described by $0 \leq x \leq 2$ and $0 \leq y \leq 2$ is "windy" and the agent is pushed by one grid tile at each step in both dimensions. Hence the optimal policy is to wait for the first two time steps so that the agent is moved to $(2, 2)$ without incurring any cost and then perform one final jump to $(5, 5)$. LICRA decides whether to act or not to act based on a tabular Q-learning rule given in (4) (using a normalised advantage function (NAF) to handle the continuous action space), where we store expected value for non acting in each cell. Fig. 9 shows that the TD-error of this tabular Q-learning setting converges to zero which validates the results of Theorem 2 and Theorem 3.

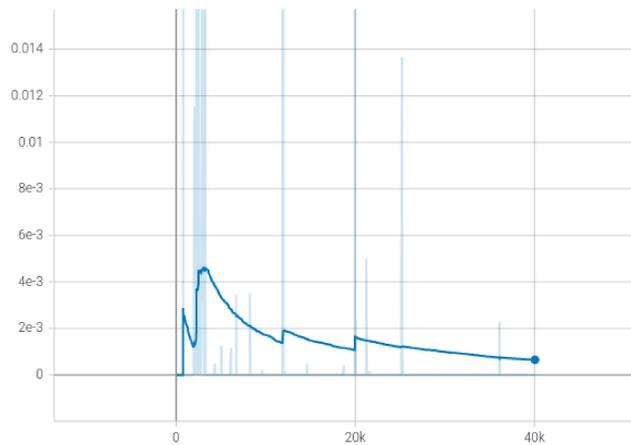


Figure 9: TD error of tabular version of LICRA (smoothed).

13 HYPERPARAMETER SETTINGS

In the table below we report all hyperparameters used in Merton Investment problem experiments.

Hyperparameter	Value (PPO methods)	Value (SAC)
Clip Gradient Norm	0.5	None
Discount Factor	0.99	0.99
Learning rate	1×10^{-3}	1×10^{-4}
Batch size	32	1024
Steps per epoch	2000	2000
Optimisation algorithm	ADAM	ADAM

In the next table, we report hyperparameters used in remaining experiments.

Hyperparameter	Value
Clip Gradient Norm	1
γ_E	0.99
λ	0.95
Learning rate	1×10^{-4}
Number of minibatches	4
Number of optimisation epochs	4
Number of parallel actors	16
Optimisation algorithm	ADAM
Rollout length	128
Sticky action probability	0.25
Use Generalized Advantage Estimation	True

14 NOTATION & ASSUMPTIONS

We assume that \mathcal{S} is defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and any $s \in \mathcal{S}$ is measurable with respect to the Borel σ -algebra associated with \mathbb{R}^p . We denote the σ -algebra of events generated by $\{s_t\}_{t \geq 0}$ by $\mathcal{F}_t \subset \mathcal{F}$. In what follows, we denote by $(\mathcal{V}, \|\cdot\|)$ any finite normed vector space and by \mathcal{H} the set of all measurable functions.

The results of the paper are built under the following assumptions which are standard within RL and stochastic approximation methods:

Assumption 1. The stochastic process governing the system dynamics is ergodic, that is the process is stationary and every invariant random variable of $\{s_t\}_{t \geq 0}$ is equal to a constant with probability 1.

Assumption 2. The function R is in L_2 .

Assumption 3. For any positive scalar c , there exists a scalar μ_c such that for all $s \in \mathcal{S}$ and for any $t \in \mathbb{N}$ we have: $\mathbb{E}[1 + \|s_t\|^c | s_0 = s] \leq \mu_c(1 + \|s\|^c)$.

Assumption 4. There exists scalars C_1 and c_1 such that for any function J satisfying $|v(s)| \leq C_2(1 + \|s\|^{c_2})$ for some scalars c_2 and C_2 we have that: $\sum_{t=0}^{\infty} |\mathbb{E}[v(s_t) | s_0 = s] - \mathbb{E}[v(s_0)]| \leq C_1 C_2 (1 + \|s\|^{c_1 c_2})$.

Assumption 5. There exists scalars c and C such that for any $s \in \mathcal{S}$ we have that: $|\mathcal{R}(s, \cdot)| \leq C(1 + \|s\|^c)$.

15 PROOF OF TECHNICAL RESULTS

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

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

$$\|TV_1 - TV_2\| \leq c\|V_1 - V_2\|. \quad (8)$$

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

$$\|TV_1 - TV_2\| \leq \|V_1 - V_2\|. \quad (9)$$

Lemma 2 For any $f : \mathcal{V} \rightarrow \mathbb{R}, g : \mathcal{V} \rightarrow \mathbb{R}$, we have that:

$$\left| \max_{a \in \mathcal{V}} f(a) - \max_{a \in \mathcal{V}} g(a) \right| \leq \max_{a \in \mathcal{V}} \|f(a) - g(a)\|. \quad (10)$$

Proof 1 We restate the proof given in (Mguni, 2019):

$$f(a) \leq \|f(a) - g(a)\| + g(a) \quad (11)$$

$$\implies \max_{a \in \mathcal{V}} f(a) \leq \max_{a \in \mathcal{V}} \{\|f(a) - g(a)\| + g(a)\} \leq \max_{a \in \mathcal{V}} \|f(a) - g(a)\| + \max_{a \in \mathcal{V}} g(a). \quad (12)$$

Deducting $\max_{a \in \mathcal{V}} g(a)$ from both sides of (12) yields:

$$\max_{a \in \mathcal{V}} f(a) - \max_{a \in \mathcal{V}} g(a) \leq \max_{a \in \mathcal{V}} \|f(a) - g(a)\|. \quad (13)$$

After reversing the roles of f and g and redoing steps (11) - (12), we deduce the desired result since the RHS of (13) is unchanged.

Lemma 3 A.4 The probability transition kernel P is non-expansive, that is:

$$\|PV_1 - PV_2\| \leq \|V_1 - V_2\|. \quad (14)$$

Proof 2 The result is well-known e.g. (Tsitsiklis & Van Roy, 1999). We give a proof using the Tonelli-Fubini theorem and the iterated law of expectations, we have that:

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

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

PROOF OF THEOREM 1

Proof 3 (Proof of Lemma 1) *In what follows, we employ the following shorthands:*

$$\mathcal{P}_{ss'}^a =: \sum_{s' \in \mathcal{S}} P(s'; a, s), \quad \mathcal{P}_{ss'}^\pi =: \sum_{a \in \mathcal{A}} \pi(a|s) \mathcal{P}_{ss'}^a, \quad \mathcal{R}^\pi(s_t) := \sum_{a_t \in \mathcal{A}} \pi(a_t|s) R(s_t, a_t)$$

For notational simplicity, where it will not cause confusion we also drop the dependence of the functions $v^{\pi, \mathfrak{g}}, Q^{\pi, \mathfrak{g}}$ on the policy pair (π, \mathfrak{g}) . With a slight abuse of notation we will write $\mathcal{M}v^{\pi, \mathfrak{g}}(s_{\tau_k}) := \sup_{a \in \mathcal{A}} \{ \mathcal{R}(s_{\tau_k}, a) - c(s_{\tau_k}, a) + \gamma \sum_{s' \in \mathcal{S}} P(s'; a, s) v^{\pi, \mathfrak{g}}(s') \}$ in which case the Bellman operator T acting on a function $v^{\pi, \mathfrak{g}} : \mathcal{S} \rightarrow \mathbb{R}$ by

$$Tv^{\pi, \mathfrak{g}}(s) := \max \left\{ \mathcal{M}[v^{\pi, \mathfrak{g}}](s), \mathcal{R}(s, 0) + \gamma \sum_{s' \in \mathcal{S}} P(s'; 0, s) v^{\pi, \mathfrak{g}}(s') \right\}, \quad \forall s \in \mathcal{S}. \quad (15)$$

To prove that T is a contraction, we consider the three cases produced by (15), that is to say we prove the following statements:

- i) $\left| \mathcal{R}(s_t, a_t) + \gamma \max_{a \in \mathcal{A}} \mathcal{P}_{s't}^a v(s') - \left(\mathcal{R}(s_t, a_t) + \gamma \max_{a \in \mathcal{A}} \mathcal{P}_{s't}^a v'(s') \right) \right| \leq \gamma \|v - v'\|$
- ii) $\|\mathcal{M}v - \mathcal{M}v'\| \leq \gamma \|v - v'\|$, (and hence \mathcal{M} is a contraction).
- iii) $\left\| \mathcal{M}v - \left[\mathcal{R}(\cdot, a) + \gamma \max_{a \in \mathcal{A}} \mathcal{P}^a v' \right] \right\| \leq \gamma \|v - v'\|$.

We begin by proving i).

Indeed, for any $a \in \mathcal{A}$ and $\forall s_t \in \mathcal{S}, \forall s' \in \mathcal{S}$ we have that

$$\begin{aligned} & \left| \mathcal{R}(s_t, a_t) + \gamma \max_{a \in \mathcal{A}} \mathcal{P}_{s't}^a v(s') - \left[\mathcal{R}(s_t, a_t) + \gamma \max_{a \in \mathcal{A}} \mathcal{P}_{s't}^a v'(s') \right] \right| \\ & \leq \max_{a \in \mathcal{A}} \left| \gamma \mathcal{P}_{s't}^a v(s') - \gamma \mathcal{P}_{s't}^a v'(s') \right| \\ & \leq \gamma \|Pv - Pv'\| \\ & \leq \gamma \|v - v'\|, \end{aligned}$$

again using the fact that P is non-expansive and Lemma 2.

We now prove ii).

For any $\tau \in \mathcal{F}$, define by $\tau' = \inf\{t > \tau | s_t \in \mathcal{S}_I, \tau \in \mathcal{F}_t\}$. Now using the definition of \mathcal{M} we have that for any $s_\tau \in \mathcal{S}$

$$\begin{aligned} & |(\mathcal{M}v - \mathcal{M}v')(s_\tau)| \\ & \leq \max_{a_\tau \in \mathcal{A}} \left| \mathcal{R}(s_\tau, a_\tau) + c(s_\tau, a_\tau) + \gamma \mathcal{P}_{s's_\tau}^\pi \mathcal{P}^a v(s_\tau) - (\mathcal{R}(s_\tau, a_\tau) + c(s_\tau, a_\tau) + \gamma \mathcal{P}_{s's_\tau}^\pi \mathcal{P}^a v'(s_\tau)) \right| \\ & = \gamma \left| \mathcal{P}_{s's_\tau}^\pi \mathcal{P}^a v(s_\tau) - \mathcal{P}_{s's_\tau}^\pi \mathcal{P}^a v'(s_\tau) \right| \\ & \leq \gamma \|Pv - Pv'\| \\ & \leq \gamma \|v - v'\|, \end{aligned}$$

using the fact that P is non-expansive. The result can then be deduced easily by applying max on both sides.

We now prove iii). We split the proof of the statement into two cases:

Case I:

$$\mathcal{M}v(s_\tau) - \left(\mathcal{R}(s_\tau, a_\tau) + \gamma \max_{a \in \mathcal{A}} \mathcal{P}_{s's_\tau}^a v'(s') \right) < 0. \quad (16)$$

We now observe the following:

$$\begin{aligned}
& \mathcal{M}v(s_\tau) - \mathcal{R}(s_\tau, a_\tau) + \gamma \max_{a \in \mathcal{A}} \mathcal{P}_{s'_t s_\tau}^a v'(s') \\
& \leq \max \left\{ \mathcal{R}(s_\tau, a_\tau) + \gamma \mathcal{P}_{s'_t s_\tau}^\pi \mathcal{P}^a v(s'), \mathcal{M}v(s_\tau) \right\} - \mathcal{R}(s_\tau, a_\tau) + \gamma \max_{a \in \mathcal{A}} \mathcal{P}_{s'_t s_\tau}^a v'(s') \\
& \leq \left| \max \left\{ \mathcal{R}(s_\tau, a_\tau) + \gamma \mathcal{P}_{s'_t s_\tau}^\pi \mathcal{P}^a v(s'), \mathcal{M}v(s_\tau) \right\} - \max \left\{ \mathcal{R}(s_\tau, a_\tau) + \gamma \max_{a \in \mathcal{A}} \mathcal{P}_{s'_t s_\tau}^a v'(s'), \mathcal{M}v(s_\tau) \right\} \right| \\
& \quad + \max \left\{ \mathcal{R}(s_\tau, a_\tau) + \gamma \max_{a \in \mathcal{A}} \mathcal{P}_{s'_t s_\tau}^a v'(s'), \mathcal{M}v(s_\tau) \right\} - \mathcal{R}(s_\tau, a_\tau) + \gamma \max_{a \in \mathcal{A}} \mathcal{P}_{s'_t s_\tau}^a v'(s') \\
& \leq \left| \max \left\{ \mathcal{R}(s_\tau, a_\tau) + \gamma \max_{a \in \mathcal{A}} \mathcal{P}_{s'_t s_\tau}^a v(s'), \mathcal{M}v(s_\tau) \right\} - \max \left\{ \mathcal{R}(s_\tau, a_\tau) + \gamma \max_{a \in \mathcal{A}} \mathcal{P}_{s'_t s_\tau}^a v'(s'), \mathcal{M}v(s_\tau) \right\} \right| \\
& \quad + \left| \max \left\{ \mathcal{R}(s_\tau, a_\tau) + \gamma \max_{a \in \mathcal{A}} \mathcal{P}_{s'_t s_\tau}^a v'(s'), \mathcal{M}v(s_\tau) \right\} - \mathcal{R}(s_\tau, a_\tau) + \gamma \max_{a \in \mathcal{A}} \mathcal{P}_{s'_t s_\tau}^a v'(s') \right| \\
& \leq \gamma \sup_{a \in \mathcal{A}} \left| \mathcal{P}_{s'_t s_\tau}^\pi \mathcal{P}^a v(s') - \mathcal{P}_{s'_t s_\tau}^\pi \mathcal{P}^a v'(s') \right| \\
& \quad + \left| \max \left\{ 0, \mathcal{M}v(s_\tau) - \left(\mathcal{R}(s_\tau, a_\tau) + \gamma \max_{a \in \mathcal{A}} \mathcal{P}_{s'_t s_\tau}^a v'(s') \right) \right\} \right| \\
& \leq \gamma \|Pv - Pv'\| \\
& \leq \gamma \|v - v'\|,
\end{aligned}$$

where we have used the fact that for any scalars a, b, c we have that $|\max\{a, b\} - \max\{b, c\}| \leq |a - c|$ and the non-expansiveness of P .

Case 2:

$$\mathcal{M}v(s_\tau) - \left(\mathcal{R}(s_\tau, a_\tau) + \gamma \max_{a \in \mathcal{A}} \mathcal{P}_{s'_t s_\tau}^a v'(s') \right) \geq 0.$$

For this case, first recall that for any $\tau \in \mathcal{F}$, $-c(s_\tau, a_\tau) > \lambda$ for some $\lambda > 0$.

$$\begin{aligned}
& \mathcal{M}v(s_\tau) - \left(\mathcal{R}(s_\tau, a_\tau) + \gamma \max_{a \in \mathcal{A}} \mathcal{P}_{s'_t s_\tau}^a v'(s') \right) \\
& \leq \mathcal{M}v(s_\tau) - \left(\mathcal{R}(s_\tau, a_\tau) + \gamma \max_{a \in \mathcal{A}} \mathcal{P}_{s'_t s_\tau}^a v(s') \right) - c(s_\tau, a_\tau) \\
& \leq \mathcal{R}(s_\tau, a_\tau) + c(s_\tau, a_\tau) + \gamma \mathcal{P}_{s'_t s_\tau}^\pi \mathcal{P}^a v(s') \\
& \quad - \left(\mathcal{R}(s_\tau, a_\tau) + c(s_\tau, a_\tau) + \gamma \max_{a \in \mathcal{A}} \mathcal{P}_{s'_t s_\tau}^a v'(s') \right) \\
& \leq \gamma \max_{a \in \mathcal{A}} \left| \mathcal{P}_{s'_t s_\tau}^\pi \mathcal{P}^a (v(s') - v'(s')) \right| \\
& \leq \gamma |v(s') - v'(s')| \\
& \leq \gamma \|v - v'\|,
\end{aligned}$$

again using the fact that P is non-expansive. Hence we have succeeded in showing that for any $v \in L_2$ we have that

$$\left\| \mathcal{M}v - \max_{a \in \mathcal{A}} [v(\cdot, a) + \gamma \mathcal{P}^a v'] \right\| \leq \gamma \|v - v'\|. \quad (17)$$

Gathering the results of the three cases gives the desired result.

To prove part ii), we make use of the following result:

Theorem 5 (Theorem 1, pg 4 in (Jaakkola et al., 1994)) Let $\Xi_t(s)$ be a random process that takes values in \mathbb{R}^n and given by the following:

$$\Xi_{t+1}(s) = (1 - \alpha_t(s)) \Xi_t(s) + \alpha_t(s) L_t(s), \quad (18)$$

then $\Xi_t(s)$ converges to 0 with probability 1 under the following conditions:

- i) $0 \leq \alpha_t \leq 1, \sum_t \alpha_t = \infty$ and $\sum_t \alpha_t < \infty$
- ii) $\|\mathbb{E}[L_t|\mathcal{F}_t]\| \leq \gamma\|\Xi_t\|$, with $\gamma < 1$;
- iii) $\text{Var}[L_t|\mathcal{F}_t] \leq c(1 + \|\Xi_t\|^2)$ for some $c > 0$.

To prove the result, we show (i) - (iii) hold. Condition (i) holds by choice of learning rate. It therefore remains to prove (ii) - (iii). We first prove (ii). For this, we consider our variant of the Q-learning update rule:

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha_t(s_t, a_t) \left[\max \left\{ \mathcal{M}Q_t(s_t, a_t), \mathcal{R}(s_t, a_t) + \gamma \sup_{a' \in \mathcal{A}} Q_t(s_{t+1}, a') \right\} - Q_t(s_t, a_t) \right].$$

After subtracting $Q^*(s_t, a_t)$ from both sides and some manipulation we obtain that:

$$\begin{aligned} \Xi_{t+1}(s_t, a_t) &= (1 - \alpha_t(s_t, a_t))\Xi_t(s_t, a_t) \\ &\quad + \alpha_t(s_t, a_t) \left[\max \left\{ \mathcal{M}Q_t(s_t, a_t), \mathcal{R}(s_t, a_t) + \gamma \sup_{a' \in \mathcal{A}} Q_t(s_{t+1}, a') \right\} - Q^*(s_t, a_t) \right], \end{aligned}$$

where $\Xi_t(s_t, a_t) := Q_t(s_t, a_t) - Q^*(s_t, a_t)$.

Let us now define by

$$L_t(s_{\tau_k}, a) := \max \left\{ \mathcal{M}Q(s_{\tau_k}, a), \mathcal{R}(s_{\tau_k}, a) + \gamma \sup_{a' \in \mathcal{A}} Q(s', a') \right\} - Q^*(s_t, a).$$

Then

$$\Xi_{t+1}(s_t, a_t) = (1 - \alpha_t(s_t, a_t))\Xi_t(s_t, a_t) + \alpha_t(s_t, a_t) [L_t(s_{\tau_k}, a)]. \quad (19)$$

We now observe that

$$\begin{aligned} \mathbb{E}[L_t(s_{\tau_k}, a)|\mathcal{F}_t] &= \sum_{s' \in \mathcal{S}} P(s'; a, s_{\tau_k}) \max \left\{ \mathcal{M}Q(s_{\tau_k}, a), \mathcal{R}(s_{\tau_k}, a) + \gamma \sup_{a' \in \mathcal{A}} Q(s', a') \right\} - Q^*(s_{\tau_k}, a) \\ &= T_\phi Q_t(s, a) - Q^*(s, a). \end{aligned} \quad (20)$$

Now, using the fixed point property that implies $Q^* = T_\phi Q^*$, we find that

$$\begin{aligned} \mathbb{E}[L_t(s_{\tau_k}, a)|\mathcal{F}_t] &= T_\phi Q_t(s, a) - T_\phi Q^*(s, a) \\ &\leq \|T_\phi Q_t - T_\phi Q^*\| \\ &\leq \gamma \|Q_t - Q^*\|_\infty = \gamma \|\Xi_t\|_\infty. \end{aligned} \quad (21)$$

using the contraction property of T established in Lemma 1. This proves (ii).

We now prove (iii), that is

$$\text{Var}[L_t|\mathcal{F}_t] \leq c(1 + \|\Xi_t\|^2). \quad (22)$$

Now by (20) we have that

$$\begin{aligned} \text{Var}[L_t|\mathcal{F}_t] &= \text{Var} \left[\max \left\{ \mathcal{M}Q_t(s_t, a_t), \mathcal{R}(s_t, a_t) + \gamma \sup_{a' \in \mathcal{A}} Q_t(s_{t+1}, a') \right\} - Q^*(s_t, a) \right] \\ &= \mathbb{E} \left[\left(\max \left\{ \mathcal{M}Q(s_{\tau_k}, a), \mathcal{R}(s_{\tau_k}, a) + \gamma \sup_{a' \in \mathcal{A}} Q(s', a') \right\} \right. \right. \\ &\quad \left. \left. - Q^*(s_t, a) - (TQ_t(s, a) - Q^*(s, a)) \right)^2 \right] \\ &= \mathbb{E} \left[\left(\max \left\{ \mathcal{M}Q(s_{\tau_k}, a), \mathcal{R}(s_{\tau_k}, a) + \gamma \sup_{a' \in \mathcal{A}} Q(s', a') \right\} - TQ_t(s, a) \right)^2 \right] \\ &= \text{Var} \left[\max \left\{ \mathcal{M}Q_t(s_t, a_t), \mathcal{R}(s_t, a_t) + \gamma \sup_{a' \in \mathcal{A}} Q_t(s_{t+1}, a') \right\} - TQ_t(s, a) \right]^2 \\ &\leq c(1 + \|\Xi_t\|^2), \end{aligned}$$

for some $c > 0$ where the last line follows due to the boundedness of Q (which follows from Assumptions 2 and 4). This concludes the proof of the Theorem.

PROOF OF CONVERGENCE WITH LINEAR FUNCTION APPROXIMATION

First let us recall the statement of the theorem:

Theorem 3 *LICRA converges to a limit point r^* which is the unique solution to the equation:*

$$\Pi \mathfrak{F}(\Phi r^*) = \Phi r^*, \quad a.e. \quad (23)$$

where we recall that for any test function $\psi \in \mathcal{V}$, the operator \mathfrak{F} is defined by $\mathfrak{F}\psi := \mathcal{R} + \gamma P \max\{\mathcal{M}\psi, \psi\}$.

Moreover, r^* satisfies the following:

$$\|\Phi r^* - Q^*\| \leq c \|\Pi Q^* - Q^*\|. \quad (24)$$

The theorem is proven using a set of results that we now establish. To this end, we first wish to prove the following bound:

Lemma 4 *For any $Q \in \mathcal{V}$ we have that*

$$\|\mathfrak{F}Q - Q'\| \leq \gamma \|Q - Q'\|, \quad (25)$$

so that the operator \mathfrak{F} is a contraction.

Proof 4 *Recall, for any test function ψ , a projection operator Π acting ψ is defined by the following*

$$\Pi\psi := \arg \min_{\bar{\psi} \in \{\Phi r \mid r \in \mathbb{R}^p\}} \|\bar{\psi} - \psi\|.$$

Now, we first note that in the proof of Lemma 1, we deduced that for any $\psi \in L_2$ we have that

$$\left\| \mathcal{M}\psi - \left[\mathcal{R}(\cdot, a) + \gamma \sup_{a \in \mathcal{A}} P^a \psi' \right] \right\| \leq \gamma \|\psi - \psi'\|,$$

(c.f. Lemma 1).

Setting $\psi = Q$ and $\psi' = Q'$ it can be straightforwardly deduced that for any $Q, \hat{Q} \in L_2$: $\|\mathcal{M}Q - \hat{Q}\| \leq \gamma \|Q - \hat{Q}\|$. Hence, using the contraction property of \mathcal{M} , we readily deduce the following bound:

$$\max \left\{ \|\mathcal{M}Q - \hat{Q}\|, \|\mathcal{M}Q - \mathcal{M}\hat{Q}\| \right\} \leq \gamma \|Q - \hat{Q}\|, \quad (26)$$

We now observe that \mathfrak{F} is a contraction. Indeed, since for any $Q, Q' \in L_2$ we have that:

$$\begin{aligned} \|\mathfrak{F}Q - \mathfrak{F}Q'\| &= \|\mathcal{R} + \gamma P \max\{\mathcal{M}Q, Q\} - (\mathcal{R} + \gamma P \max\{\mathcal{M}Q', Q'\})\| \\ &= \gamma \|P \max\{\mathcal{M}Q, Q\} - P \max\{\mathcal{M}Q', Q'\}\| \\ &\leq \gamma \|\max\{\mathcal{M}Q, Q\} - \max\{\mathcal{M}Q', Q'\}\| \\ &\leq \gamma \|\max\{\mathcal{M}Q - \mathcal{M}Q', Q - \mathcal{M}Q', \mathcal{M}Q - Q', Q - Q'\}\| \\ &\leq \gamma \max\{\|\mathcal{M}Q - \mathcal{M}Q'\|, \|Q - \mathcal{M}Q'\|, \|\mathcal{M}Q - Q'\|, \|Q - Q'\|\} \\ &= \gamma \|Q - Q'\|, \end{aligned}$$

using (26) and again using the non-expansiveness of P .

We next show that the following two bounds hold:

Lemma 5 *For any $Q \in \mathcal{V}$ we have that*

$$i) \quad \|\Pi \mathfrak{F}Q - \Pi \mathfrak{F}\bar{Q}\| \leq \gamma \|Q - \bar{Q}\|,$$

$$ii) \quad \|\Phi_{r^*} - Q^*\| \leq \frac{1}{\sqrt{1-\gamma^2}} \|\Pi Q^* - Q^*\|.$$

Proof 5 The first result is straightforward since as Π is a projection it is non-expansive and hence:

$$\|\Pi \mathfrak{F}Q - \Pi \mathfrak{F}\bar{Q}\| \leq \|\mathfrak{F}Q - \mathfrak{F}\bar{Q}\| \leq \gamma \|Q - \bar{Q}\|,$$

using the contraction property of \mathfrak{F} . This proves i). For ii), we note that by the orthogonality property of projections we have that $\langle \Phi_{r^*} - \Pi Q^*, \Phi_{r^*} - \Pi Q^* \rangle$, hence we observe that:

$$\begin{aligned} \|\Phi_{r^*} - Q^*\|^2 &= \|\Phi_{r^*} - \Pi Q^*\|^2 + \|\Phi_{r^*} - \Pi Q^*\|^2 \\ &= \|\Pi \mathfrak{F}\Phi_{r^*} - \Pi Q^*\|^2 + \|\Phi_{r^*} - \Pi Q^*\|^2 \\ &\leq \|\mathfrak{F}\Phi_{r^*} - Q^*\|^2 + \|\Phi_{r^*} - \Pi Q^*\|^2 \\ &= \|\mathfrak{F}\Phi_{r^*} - \mathfrak{F}Q^*\|^2 + \|\Phi_{r^*} - \Pi Q^*\|^2 \\ &\leq \gamma^2 \|\Phi_{r^*} - Q^*\|^2 + \|\Phi_{r^*} - \Pi Q^*\|^2, \end{aligned}$$

after which we readily deduce the desired result.

Lemma 6 Define the operator H by the following: $HQ(s, a) = \begin{cases} \mathcal{M}Q(s, a), & \text{if } \mathcal{M}Q(s, a) > \Phi_{r^*}, \\ Q(s, a), & \text{otherwise,} \end{cases}$

where we define $\tilde{\mathfrak{F}}$ by: $\tilde{\mathfrak{F}}Q := \mathcal{R} + \gamma PHQ$.

For any $Q, \bar{Q} \in L_2$ we have that

$$\|\tilde{\mathfrak{F}}Q - \tilde{\mathfrak{F}}\bar{Q}\| \leq \gamma \|Q - \bar{Q}\| \quad (27)$$

and hence $\tilde{\mathfrak{F}}$ is a contraction mapping.

Proof 6 Using (26), we now observe that

$$\begin{aligned} \|\tilde{\mathfrak{F}}Q - \tilde{\mathfrak{F}}\bar{Q}\| &= \|\mathcal{R} + \gamma PHQ - (\mathcal{R} + \gamma PH\bar{Q})\| \\ &\leq \gamma \|HQ - H\bar{Q}\| \\ &\leq \gamma \|\max\{\mathcal{M}Q - \mathcal{M}\bar{Q}, Q - \bar{Q}, \mathcal{M}Q - \bar{Q}, \mathcal{M}\bar{Q} - Q\}\| \\ &\leq \gamma \max\{\|\mathcal{M}Q - \mathcal{M}\bar{Q}\|, \|Q - \bar{Q}\|, \|\mathcal{M}Q - \bar{Q}\|, \|\mathcal{M}\bar{Q} - Q\|\} \\ &\leq \gamma \max\{\gamma \|Q - \bar{Q}\|, \|Q - \bar{Q}\|, \|\mathcal{M}Q - \bar{Q}\|, \|\mathcal{M}\bar{Q} - Q\|\} \\ &= \gamma \|Q - \bar{Q}\|, \end{aligned}$$

again using the non-expansive property of P .

Lemma 7 Define by $\tilde{Q} := \mathcal{R} + \gamma Pv^{\tilde{\pi}}$ where

$$v^{\tilde{\pi}}(s) := \mathcal{R}(s_{\tau_k}, a) + \gamma \sup_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} P(s'; a, s_{\tau_k}) \Phi_{r^*}(s'), \quad (28)$$

then \tilde{Q} is a fixed point of $\tilde{\mathfrak{F}}\tilde{Q}$, that is $\tilde{\mathfrak{F}}\tilde{Q} = \tilde{Q}$.

Proof 7 We begin by observing that

$$\begin{aligned} H\tilde{Q}(s, a) &= H(\mathcal{R}(s, \cdot) + \gamma Pv^{\tilde{\pi}}) \\ &= \begin{cases} \mathcal{M}Q(s, a), & \text{if } \mathcal{M}Q(s, a) > \Phi_{r^*}, \\ Q(s, a), & \text{otherwise,} \end{cases} \\ &= \begin{cases} \mathcal{M}Q(s, a), & \text{if } \mathcal{M}Q(s, a) > \Phi_{r^*}, \\ \mathcal{R}(s, \cdot) + \gamma Pv^{\tilde{\pi}}, & \text{otherwise,} \end{cases} \\ &= v^{\tilde{\pi}}(s). \end{aligned}$$

Hence,

$$\tilde{\mathfrak{F}}\tilde{Q} = \mathcal{R} + \gamma PH\tilde{Q} = \mathcal{R} + \gamma Pv^{\tilde{\pi}} = \tilde{Q}. \quad (29)$$

which proves the result.

Lemma 8 *The following bound holds:*

$$\mathbb{E} [v^{\hat{\pi}}(s_0)] - \mathbb{E} [v^{\tilde{\pi}}(s_0)] \leq 2 \left[(1 - \gamma) \sqrt{(1 - \gamma^2)} \right]^{-1} \|\Pi Q^* - Q^*\|. \quad (30)$$

Proof 8 *By definitions of $v^{\hat{\pi}}$ and $v^{\tilde{\pi}}$ (c.f. (28)) and using Jensen's inequality and the stationarity property we have that,*

$$\begin{aligned} \mathbb{E} [v^{\hat{\pi}}(s_0)] - \mathbb{E} [v^{\tilde{\pi}}(s_0)] &= \mathbb{E} [Pv^{\hat{\pi}}(s_0)] - \mathbb{E} [Pv^{\tilde{\pi}}(s_0)] \\ &\leq |\mathbb{E} [Pv^{\hat{\pi}}(s_0)] - \mathbb{E} [Pv^{\tilde{\pi}}(s_0)]| \\ &\leq \|Pv^{\hat{\pi}} - Pv^{\tilde{\pi}}\|. \end{aligned} \quad (31)$$

Now recall that $\tilde{Q} := \mathcal{R} + \gamma Pv^{\hat{\pi}}$ and $Q^ := \mathcal{R} + \gamma Pv^{\pi^*}$, using these expressions in (31) we find that*

$$\mathbb{E} [v^{\hat{\pi}}(s_0)] - \mathbb{E} [v^{\tilde{\pi}}(s_0)] \leq \frac{1}{\gamma} \|\tilde{Q} - Q^*\|.$$

Moreover, by the triangle inequality and using the fact that $\mathfrak{F}(\Phi r^) = \tilde{\mathfrak{F}}(\Phi r^*)$ and that $\mathfrak{F}Q^* = Q^*$ and $\tilde{\mathfrak{F}}\tilde{Q} = \tilde{Q}$ (c.f. (30)) we have that*

$$\begin{aligned} \|\tilde{Q} - Q^*\| &\leq \|\tilde{Q} - \mathfrak{F}(\Phi r^*)\| + \|\mathfrak{F}(\Phi r^*) - \tilde{\mathfrak{F}}(\Phi r^*)\| \\ &\leq \gamma \|\tilde{Q} - \Phi r^*\| + \gamma \|Q^* - \Phi r^*\| \\ &\leq 2\gamma \|\tilde{Q} - \Phi r^*\| + \gamma \|Q^* - \tilde{Q}\|, \end{aligned}$$

which gives the following bound:

$$\|\tilde{Q} - Q^*\| \leq 2(1 - \gamma)^{-1} \|\tilde{Q} - \Phi r^*\|,$$

from which, using Lemma 5, we deduce that $\|\tilde{Q} - Q^\| \leq 2 \left[(1 - \gamma) \sqrt{(1 - \gamma^2)} \right]^{-1} \|\tilde{Q} - \Phi r^*\|$, after which by (32), we finally obtain*

$$\mathbb{E} [v^{\hat{\pi}}(s_0)] - \mathbb{E} [v^{\tilde{\pi}}(s_0)] \leq 2 \left[(1 - \gamma) \sqrt{(1 - \gamma^2)} \right]^{-1} \|\tilde{Q} - \Phi r^*\|,$$

as required.

Let us rewrite the update in the following way:

$$r_{t+1} = r_t + \gamma_t \Xi(w_t, r_t),$$

where the function $\Xi : \mathbb{R}^{2d} \times \mathbb{R}^p \rightarrow \mathbb{R}^p$ is given by:

$$\Xi(w, r) := \phi(s) (\mathcal{R}(s, \cdot) + \gamma \max \{(\Phi r)(s'), \mathcal{M}(\Phi r)(s')\} - (\Phi r)(s)),$$

for any $w \equiv (s, s') \in \mathcal{S}^2$ and for any $r \in \mathbb{R}^p$. Let us also define the function $\Xi : \mathbb{R}^p \rightarrow \mathbb{R}^p$ by the following:

$$\Xi(r) := \mathbb{E}_{w_0 \sim (\mathbb{P}, \mathbb{P})} [\Xi(w_0, r)]; w_0 := (s_0, z_1).$$

Lemma 9 *The following statements hold for all $z \in \{0, 1\} \times \mathcal{S}$:*

- i) $(r - r^*) \Xi_k(r) < 0, \quad \forall r \neq r^*,$
- ii) $\Xi_k(r^*) = 0.$

Proof 9 To prove the statement, we first note that each component of $\Xi_k(r)$ admits a representation as an inner product, indeed:

$$\begin{aligned}\Xi_k(r) &= \mathbb{E}[\phi_k(s_0)(\mathcal{R}(s_0, a_0) + \gamma \max\{\Phi r(s_1), \mathcal{M}\Phi(s_1)\}) - (\Phi r)(s_0)] \\ &= \mathbb{E}[\phi_k(s_0)(\mathcal{R}(s_0, a_0) + \gamma \mathbb{E}[\max\{\Phi r(s_1), \mathcal{M}\Phi(s_1)\} | z_0]) - (\Phi r)(s_0)] \\ &= \mathbb{E}[\phi_k(s_0)(\mathcal{R}(s_0, a_0) + \gamma P \max\{(\Phi r, \mathcal{M}\Phi)\}(s_0) - (\Phi r)(s_0))] \\ &= \langle \phi_k, \mathfrak{F}\Phi r - \Phi r \rangle,\end{aligned}$$

using the iterated law of expectations and the definitions of P and \mathfrak{F} .

We now are in position to prove i). Indeed, we now observe the following:

$$\begin{aligned}(r - r^*) \Xi_k(r) &= \sum_{l=1} (r(l) - r^*(l)) \langle \phi_l, \mathfrak{F}\Phi r - \Phi r \rangle \\ &= \langle \Phi r - \Phi r^*, \mathfrak{F}\Phi r - \Phi r \rangle \\ &= \langle \Phi r - \Phi r^*, (\mathbf{1} - \Pi)\mathfrak{F}\Phi r + \Pi\mathfrak{F}\Phi r - \Phi r \rangle \\ &= \langle \Phi r - \Phi r^*, \Pi\mathfrak{F}\Phi r - \Phi r \rangle,\end{aligned}$$

where in the last step we used the orthogonality of $(\mathbf{1} - \Pi)$. We now recall that $\Pi\mathfrak{F}\Phi r^* = \Phi r^*$ since Φr^* is a fixed point of $\Pi\mathfrak{F}$. Additionally, using Lemma 5 we observe that $\|\Pi\mathfrak{F}\Phi r - \Phi r^*\| \leq \gamma\|\Phi r - \Phi r^*\|$. With this we now find that

$$\begin{aligned}\langle \Phi r - \Phi r^*, \Pi\mathfrak{F}\Phi r - \Phi r \rangle &= \langle \Phi r - \Phi r^*, (\Pi\mathfrak{F}\Phi r - \Phi r^*) + \Phi r^* - \Phi r \rangle \\ &\leq \|\Phi r - \Phi r^*\| \|\Pi\mathfrak{F}\Phi r - \Phi r^*\| - \|\Phi r^* - \Phi r\|^2 \\ &\leq (\gamma - 1) \|\Phi r^* - \Phi r\|^2,\end{aligned}$$

which is negative since $\gamma < 1$ which completes the proof of part i).

The proof of part ii) is straightforward since we readily observe that

$$\Xi_k(r^*) = \langle \phi_l, \mathfrak{F}\Phi r^* - \Phi r \rangle = \langle \phi_l, \Pi\mathfrak{F}\Phi r^* - \Phi r \rangle = 0,$$

as required and from which we deduce the result.

To prove the theorem, we make use of a special case of the following result:

Theorem 6 (Th. 17, p. 239 in (Benveniste et al., 2012)) Consider a stochastic process $r_t : \mathbb{R} \times \{\infty\} \times \Omega \rightarrow \mathbb{R}^k$ which takes an initial value r_0 and evolves according to the following:

$$r_{t+1} = r_t + \alpha \Xi(s_t, r_t), \quad (32)$$

for some function $s : \mathbb{R}^{2d} \times \mathbb{R}^k \rightarrow \mathbb{R}^k$ and where the following statements hold:

1. $\{s_t | t = 0, 1, \dots\}$ is a stationary, ergodic Markov process taking values in \mathbb{R}^{2d}
2. For any positive scalar q , there exists a scalar μ_q such that $\mathbb{E}[1 + \|s_t\|^q | s \equiv s_0] \leq \mu_q(1 + \|s\|^q)$
3. The step size sequence satisfies the Robbins-Monro conditions, that is $\sum_{t=0}^{\infty} \alpha_t = \infty$ and $\sum_{t=0}^{\infty} \alpha_t^2 < \infty$
4. There exists scalars c and q such that $\|\Xi(w, r)\| \leq c(1 + \|w\|^q)(1 + \|r\|)$
5. There exists scalars c and q such that $\sum_{t=0}^{\infty} \|\mathbb{E}[\Xi(w_t, r) | z_0 \equiv z] - \mathbb{E}[\Xi(w_0, r)]\| \leq c(1 + \|w\|^q)(1 + \|r\|)$
6. There exists a scalar $c > 0$ such that $\|\mathbb{E}[\Xi(w_0, r)] - \mathbb{E}[\Xi(w_0, \bar{r})]\| \leq c\|r - \bar{r}\|$
7. There exists scalars $c > 0$ and $q > 0$ such that $\sum_{t=0}^{\infty} \|\mathbb{E}[\Xi(w_t, r) | w_0 \equiv w] - \mathbb{E}[\Xi(w_0, \bar{r})]\| \leq c\|r - \bar{r}\|(1 + \|w\|^q)$
8. There exists some $r^* \in \mathbb{R}^k$ such that $\Xi(r)(r - r^*) < 0$ for all $r \neq r^*$ and $\bar{s}(r^*) = 0$.

Then r_t converges to r^* almost surely.

In order to apply the Theorem 6, we show that conditions 1 - 7 are satisfied.

Proof 10 Conditions 1-2 are true by assumption while condition 3 can be made true by choice of the learning rates. Therefore it remains to verify conditions 4-7 are met.

To prove 4, we observe that

$$\begin{aligned} \|\Xi(w, r)\| &= \|\phi(s) (\mathcal{R}(s, \cdot) + \gamma \max \{(\Phi r)(s'), \mathcal{M}\Phi(s')\} - (\Phi r)(s))\| \\ &\leq \|\phi(s)\| \|\mathcal{R}(s, \cdot) + \gamma (\|\phi(s')\| \|r\| + \mathcal{M}\Phi(s'))\| + \|\phi(s)\| \|r\| \\ &\leq \|\phi(s)\| (\|\mathcal{R}(s, \cdot)\| + \gamma \|\mathcal{M}\Phi(s')\|) + \|\phi(s)\| (\gamma \|\phi(s')\| + \|\phi(s)\|) \|r\|. \end{aligned}$$

Now using the definition of \mathcal{M} , we readily observe that $\|\mathcal{M}\Phi(s')\| \leq \|\mathcal{R}\| + \gamma \|\mathcal{P}_{s't}^\pi \Phi\| \leq \|\mathcal{R}\| + \gamma \|\Phi\|$ using the non-expansiveness of P .

Hence, we lastly deduce that

$$\begin{aligned} \|\Xi(w, r)\| &\leq \|\phi(s)\| (\|\mathcal{R}(s, \cdot)\| + \gamma \|\mathcal{M}\Phi(s')\|) + \|\phi(s)\| (\gamma \|\phi(s')\| + \|\phi(s)\|) \|r\| \\ &\leq \|\phi(s)\| (\|\mathcal{R}(s, \cdot)\| + \gamma \|\mathcal{R}\| + \gamma \|\phi\|) + \|\phi(s)\| (\gamma \|\phi(s')\| + \|\phi(s)\|) \|r\|, \end{aligned}$$

we then easily deduce the result using the boundedness of ϕ and \mathcal{R} .

Now we observe the following Lipschitz condition on Ξ :

$$\begin{aligned} &\|\Xi(w, r) - \Xi(w, \bar{r})\| \\ &= \|\phi(s) (\gamma \max \{(\Phi r)(s'), \mathcal{M}\Phi(s')\} - \gamma \max \{(\Phi \bar{r})(s'), \mathcal{M}\Phi(s')\}) - ((\Phi r)(s) - \Phi \bar{r}(s))\| \\ &\leq \gamma \|\phi(s)\| \|\max \{\phi'(s')r, \mathcal{M}\Phi'(s')\} - \max \{\phi'(s')\bar{r}, \mathcal{M}\Phi'(s')\}\| + \|\phi(s)\| \|\phi'(s)r - \phi(s)\bar{r}\| \\ &\leq \gamma \|\phi(s)\| \|\phi'(s')r - \phi'(s')\bar{r}\| + \|\phi(s)\| \|\phi'(s)r - \phi'(s)\bar{r}\| \\ &\leq \|\phi(s)\| (\|\phi(s)\| + \gamma \|\phi(s)\| \|\phi'(s') - \phi'(s')\|) \|r - \bar{r}\| \\ &\leq c \|r - \bar{r}\|, \end{aligned}$$

using Cauchy-Schwarz inequality and that for any scalars a, b, c we have that $|\max\{a, b\} - \max\{b, c\}| \leq |a - c|$.

Using Assumptions 3 and 4, we therefore deduce that

$$\sum_{t=0}^{\infty} \|\mathbb{E} [\Xi(w, r) - \Xi(w, \bar{r}) | w_0 = w] - \mathbb{E} [\Xi(w_0, r) - \Xi(w_0, \bar{r})]\| \leq c \|r - \bar{r}\| (1 + \|w\|^l). \quad (33)$$

Part 2 is assured by Lemma 5 while Part 4 is assured by Lemma 8 and lastly Part 8 is assured by Lemma 9. This result completes the proof of Theorem 1.

PROOF OF PROPOSITION 1

Proof 11 First let us recall that the intervention time τ_k is defined recursively $\tau_k = \inf\{t > \tau_{k-1} | s_t \in A, \tau_k \in \mathcal{F}_t\}$ where $A = \{s \in \mathcal{S}, g(s_t) = 1\}$. The proof is given by establishing a contradiction. Therefore suppose that $\mathcal{M}\psi(s_{\tau_k}) \leq \psi(s_{\tau_k})$ and suppose that the intervention time $\tau'_1 > \tau_1$ is an optimal intervention time. Construct the $\pi^l \in \Pi$ and $\tilde{\pi} \in \Pi$ policy switching times by $(\tau'_0, \tau'_1, \dots)$ and (τ_0, τ_1, \dots) respectively. Define by $l = \inf\{t > 0; \mathcal{M}\psi(s_t) = \psi(s_t)\}$ and $m = \sup\{t; t < \tau'_1\}$. By construction we have that

$$\begin{aligned} &v^{\pi^l}(s) \\ &= \mathbb{E} \left[\mathcal{R}(s_0, a_0) + \mathbb{E} \left[\dots + \gamma^{l-1} \mathbb{E} \left[\mathcal{R}(s_{\tau_1-1}, a_{\tau_1-1}) + \dots + \gamma^{m-l-1} \mathbb{E} \left[\mathcal{R}(s_{\tau'_1-1}, a_{\tau'_1-1}) + \gamma \mathcal{M}^{\pi^l, \pi^l} v^{\pi^l}(s', I(\tau'_1)) \right] \right] \right] \right] \\ &< \mathbb{E} \left[\mathcal{R}(s_0, a_0) + \mathbb{E} \left[\dots + \gamma^{l-1} \mathbb{E} \left[\mathcal{R}(s_{\tau_1-1}, a_{\tau_1-1}) + \gamma \mathcal{M}^{\tilde{\pi}} v^{\pi^l}(s_{\tau_1}) \right] \right] \right] \end{aligned}$$

We now use the following observation

$$\mathbb{E} \left[\mathcal{R}(s_{\tau_1-1}, a_{\tau_1-1}) + \gamma \mathcal{M}^{\tilde{\pi}} v^{\pi'}(s_{\tau_1}) \right] \quad (34)$$

$$\leq \max \left\{ \mathcal{M}^{\tilde{\pi}} v^{\pi'}(s_{\tau_1}), \max_{a_{\tau_1} \in \mathcal{A}} \left[\mathcal{R}(s_{\tau_k}, a_{\tau_k}) + \gamma \sum_{s' \in \mathcal{S}} P(s'; a_{\tau_1}, s_{\tau_1}) v^{\pi}(s') \right] \right\}. \quad (35)$$

Using this we deduce that

$$\begin{aligned} v^{\pi'}(s) &\leq \mathbb{E} \left[\mathcal{R}(s_0, a_0) + \mathbb{E} \left[\dots \right. \right. \\ &\quad \left. \left. + \gamma^{l-1} \mathbb{E} \left[\mathcal{R}(s_{\tau_1-1}, a_{\tau_1-1}) + \gamma \max \left\{ \mathcal{M}^{\tilde{\pi}} v^{\pi'}(s_{\tau_1}), \max_{a_{\tau_1} \in \mathcal{A}} \left[\mathcal{R}(s_{\tau_k}, a_{\tau_k}) + \gamma \sum_{s' \in \mathcal{S}} P(s'; a_{\tau_1}, s_{\tau_1}) v^{\pi}(s') \right] \right\} \right] \right] \right] \\ &= \mathbb{E} \left[\mathcal{R}(s_0, a_0) + \mathbb{E} \left[\dots + \gamma^{l-1} \mathbb{E} \left[\mathcal{R}(s_{\tau_1-1}, a_{\tau_1-1}) + \gamma [T v^{\tilde{\pi}}](s_{\tau_1}) \right] \right] \right] = v^{\tilde{\pi}}(s), \end{aligned}$$

where the first inequality is true by assumption on \mathcal{M} . This is a contradiction since π' is an optimal policy for Player 2. Using analogous reasoning, we deduce the same result for $\tau'_k < \tau_k$ after which deduce the result. Moreover, by invoking the same reasoning, we can conclude that it must be the case that $(\tau_0, \tau_1, \dots, \tau_{k-1}, \tau_k, \tau_{k+1}, \dots)$ are the optimal switching times.