Part I

Appendix

Table of Contents

9 ALGORITHM

Algorithm 1: **L**earnable **I**ntrinsic-Reward **G**eneration **S**election algorithm (LIGS)

Input: Environment E

Initial agent policies $\boldsymbol{\pi}_0 = (\pi_0^1, \dots \pi_0^N)$ with parameters $\theta_{\pi_0^1}, \dots \theta_{\pi_N^1}$, Initial Generator switch policy $\mathfrak{g}_{c_0}$ with parameters $\theta_{\mathfrak{g}_{c_0}}$, Initial Generator action policy g_0 with parameters θ_{g_0}, Randomly initialised fixed neural network $\phi(\cdot, \cdot)$, Neural networks h (fixed) and $\hat{h}$ for Augmented RND with parameter $\theta_{\hat{h}}$, Buffer B, Number of rollouts N_r, rollout length T, Number of mini-batch updates N_u, Switch cost c, discount factor γ, learning rate α.

Output: Optimised agent policies $\boldsymbol{\pi}^\star = (\pi^{\star,1}, \dots, \pi^{\star,N})$

$\boldsymbol{\pi} = (\pi^1, \dots, \pi^N), g, \mathfrak{g}_c \leftarrow \boldsymbol{\pi}_0, g_0, \mathfrak{g}_{c_0}$

for $n = 1, N_r$ **do**

  // Collect rollouts

  for $t = 1, T$ **do**

    Get environment states s_t from E

    Sample $\boldsymbol{a}_t = (a_t^1, \dots, a_t^N)$ from $(\pi^1(s_t), \dots, \pi^N(s_t))$

    Apply action $\boldsymbol{a}_t$ to environment E, get rewards $\boldsymbol{r}_t = (r_t^1, \dots, r_t^N)$ and next state s_{t+1}

    Sample q_t from $\mathfrak{g}_c(s_t)$ **// Switching control**

    if $q_t = 1$ **then**

      Sample θ_t^c from $g(s_t)$

      Sample θ_{t+1}^c from $g(s_{t+1})$

      $f_t^i = \gamma\theta_{t+1}^c - \theta_t^c$ **// Calculate** $F(\theta_t^c, \theta_{t+1}^c)$

    else

      $\theta_t^c, f_t^i = 0, 0$ **// Dummy values**

    Append $(s_t, \boldsymbol{a}_t, g_t, \theta_t^c, \boldsymbol{r}_t, f_t^i, s_{t+1})$ to B

  for $u = 1, N_u$ **do**

    Sample data $(s_t, \boldsymbol{a}_t, g_t, \theta_t^c, \boldsymbol{r}_t, f_t^i, s_{t+1})$ from B

    if $g_t = 1$ **then**

      Set reward to $\boldsymbol{r}_t^s = \boldsymbol{r}_t + f_t^i$

    else

      Set reward to $\boldsymbol{r}_t^s = \boldsymbol{r}_t$

    // Update Augmented RND

    $\text{Loss}_{\text{RND}} = ||h(s_t, \boldsymbol{a}_t) - \hat{h}(s_t, \boldsymbol{a}_t)||^2$

    $\theta_{\hat{h}} \leftarrow \theta_{\hat{h}} - \alpha\nabla\text{Loss}_{\text{RND}}$

    // Update Generator

    $l_t = ||h(s_t, \boldsymbol{a}_t) - \hat{h}(s_t)||^2$ **// Compute** $L(s_t, \boldsymbol{a}_t)$

    $c_t = cg_t$

    Compute Loss_g using $(s_t, a_t, g_t, c_t, \boldsymbol{r}_t, f_t^i, l_t, s_{t+1})$ using PPO loss **// Section 4.1**

    Compute $\text{Loss}_{\mathfrak{g}_c}$ using $(s_t, a_t, g_t, c_t, \boldsymbol{r}_t, f_t^i, l_t, s_{t+1})$ using PPO loss **// Section 4.1**

    $\theta_g \leftarrow \theta_g - \alpha\nabla\text{Loss}_g$

    $\theta_{\mathfrak{g}_c} \leftarrow \theta_{\mathfrak{g}_c} - \alpha\nabla\text{Loss}_{\mathfrak{g}_c}$

    // Update agent j**, for each** $j \in 1, \dots, N$

    Compute Loss_{π^j} using $(s_t, \boldsymbol{a}_t, r_t^{j,s} := r_t^j + f_t^i, s_{t+1})$ using PPO loss **// Section 4.1**

    $\theta_{\pi^j} \leftarrow \theta_{\pi^j} - \alpha\nabla\text{Loss}_{\pi^j}$

10 Ablation study: Plug & Play

In order to validate our claim that LIGS freely adopts RL learners, we tested the ability of LIGS to boost performance in a complex coordination task using independent Proximal policy optimization algorithm (IPPO) (Schulman et al., 2017) as the base learner. In this experiment, two agents are spawned at opposite sides of the grid. The red agent is spawned in the left hand side and the blue agent is spawned in the right hand side of the grid in Fig. 5 (right). The goal of the agents is to arrive at their corresponding goal states (indicated by the coloured square, where the colour corresponds to the agent whose goal state it is) at the other side of the grid. Upon arriving at their goal state the agents receive their reward. However, the task is made difficult by the fact that only one agent can pass through the corridor at a time. Therefore, in this setup, the only way for the agents to complete the task is for the agents to successfully coordinate, i.e. one agent is required to allow the other agent to pass through before attempting to traverse the corridor.

It is known that independent learners in general, struggle to solve such tasks since their ability to coordinate systems of RL learners is lacking (Yang et al., 2020). This is demonstrated in Fig. 5 (left) which displays the performance curve of for IPPO which fails to score above 0. As claimed, when incorporated into the LIGS framework, the agents succeed in coordinating to solve the task. This is indicated by the performance of IPPO + LIGS (blue).

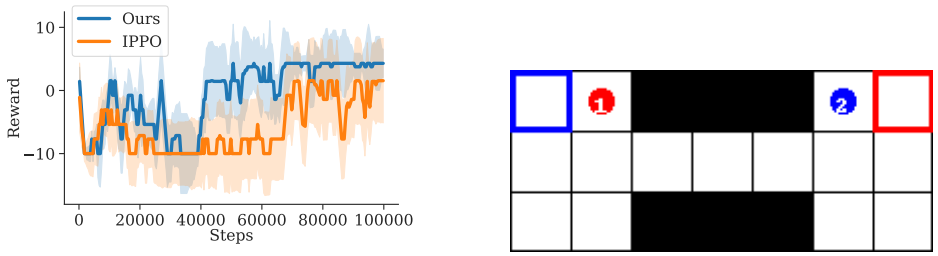

Figure 5: *Left.* Performance curves for IPPO and IPPO with LIGS. *Right.* Coordination environment.

11 Ablation Study: The Utility of Switching Controls

A core component of LIGS is the switching control mechanism. This component enables the Generator to selectively add intrinsic rewards only at the set of states most relevant for improving learning outcomes while avoiding adding intrinsic rewards where they are not necessary. To evaluate the impact of this component of LIGS, we compared the performance of LIGS with a version in which the switching control was replaced with an equal-chances Bernoulli Random Variable (i.e., at any given state, the Generator adds or does not add intrinsic rewards with equal probability), and, a version where it always adds intrinsic rewards. Figure 6 shows the performance of these three versions of LIGS. We added vanilla MAPPO as a baseline reference. We examined the performance of the variants of LIGS on the coordination task described in Section 10. As can be seen in the plot, incorporating learned switching controls in LIGS (labelled "LIGS") leads to superior performance compared to simply adding intrinsic rewards at random (line labelled "LIGS with Random Switching") and adding intrinsic rewards everywhere (labelled "LIGS with Always Adding intrinsic Rewards"). In fact, adding intrinsic rewards at random is detrimental to performance as demonstrated by the fact that the performance of LIGS with Random Switching is worse than that of vanilla MAPPO.

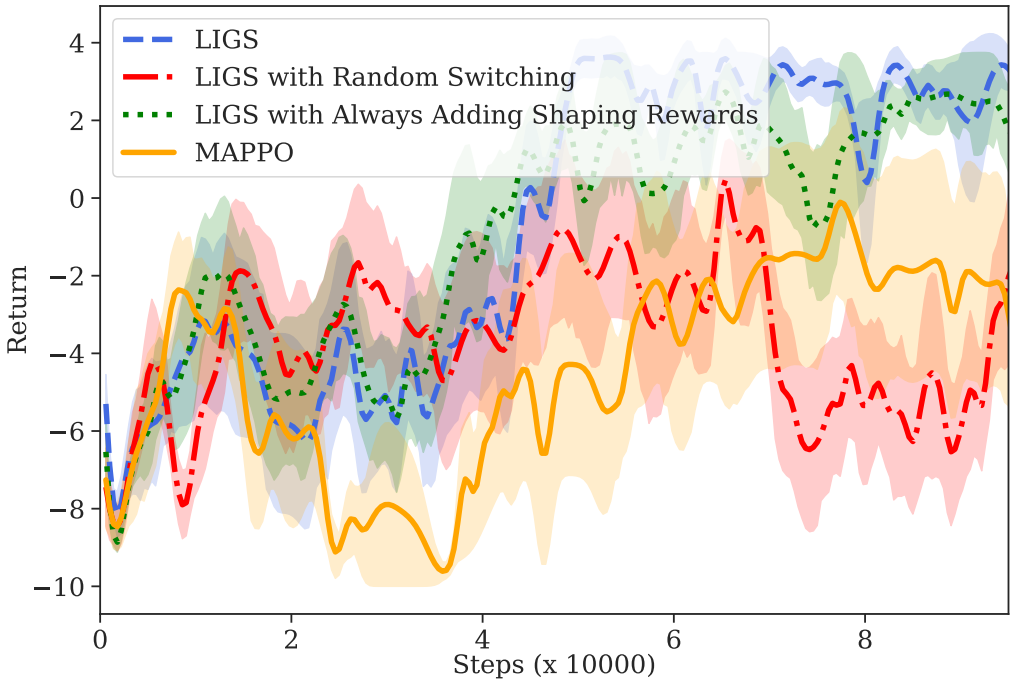

Figure 6: Ablation of the switching control mechanism. Learned switching controls ("LIGS") outperform versions where intrinsic rewards are added at random ("LIGS with Random Switching") and where intrinsic rewards are always added ("LIGS with Always Added intrinsic Rewards").

12 FLEXIBILITY OF LIGS TO ACCOMMODATE DIFFERENT EXPLORATION BONUS TERMS L

To demonstrate the robustness of our method to different choices of exploration bonus terms in Generator's objective, we conducted an Ablation study on the L-term (c.f. Equation 3) where we replaced the RND L term with a basic count-based exploration bonus. To exemplify the high degree of flexibility, we replaced the RND with a simple exploration bonus term $L(s) = \frac{1}{\text{Count(s)}+1}$ for any given state $s \in \mathcal{S}$ where Count(s) refers to a simple count of the number of times the state s has been visited. We conducted the Ablation study on all three Foraging environments presented in Sec. 6.1. We note that despite the simplicity of the count-based measure, generally the performance of both versions of LIGS is comparable and in fact the count-based variant is superior to the RND version for the joint exploration environment.

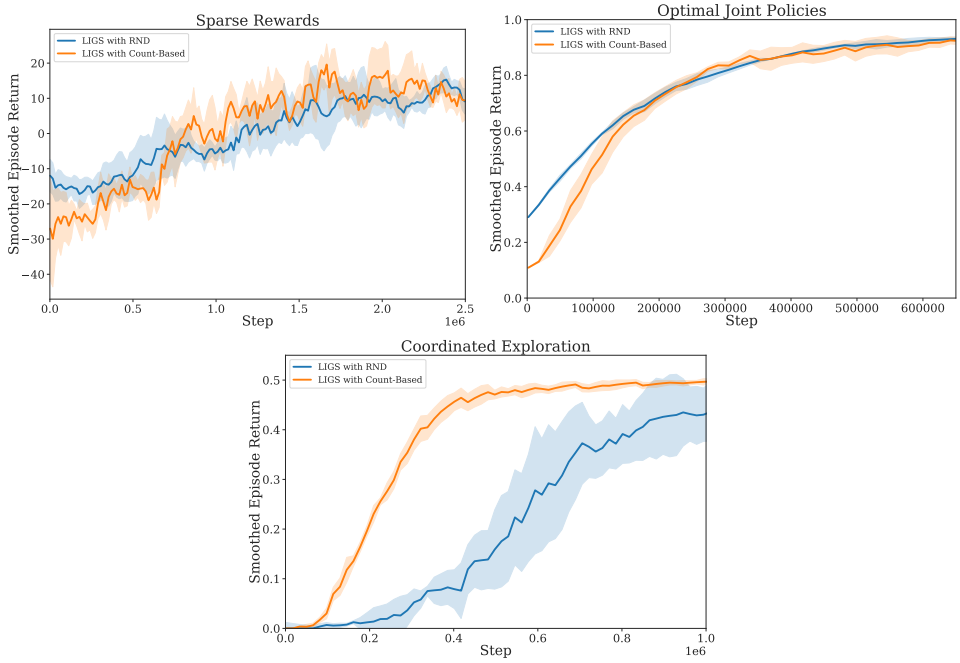

Figure 7: Performance of LIGS compared with the exploration bonus replaced by count-based method on the three tasks in the Foraging environment.

13 Further Experiment Demonstrating LIGS improved use of Exploration Bonuses.

As we have shown above, LIGS can accommodate a variety of exploration bonuses and perform well. Here, we did a experiment to further justify using LIGS against simpler exploration bonus methods. We compared LIGS against and MAPPO with an RND intrinsic reward in the agents' objectives (MAPPO+RND) and vanilla MAPPO. Fig. 8 shows performance of these two methods on coordination environment shown in Fig. 5. We note that LIGS markedly outperforms both MAPPO+RND and vanilla MAPPO. Due to the added benefit of switching controls and intrinsic reward selection performed by the Generator, we observe that LIGS is able to significantly augment the benefits of applying RND directly to the agents' objectives.

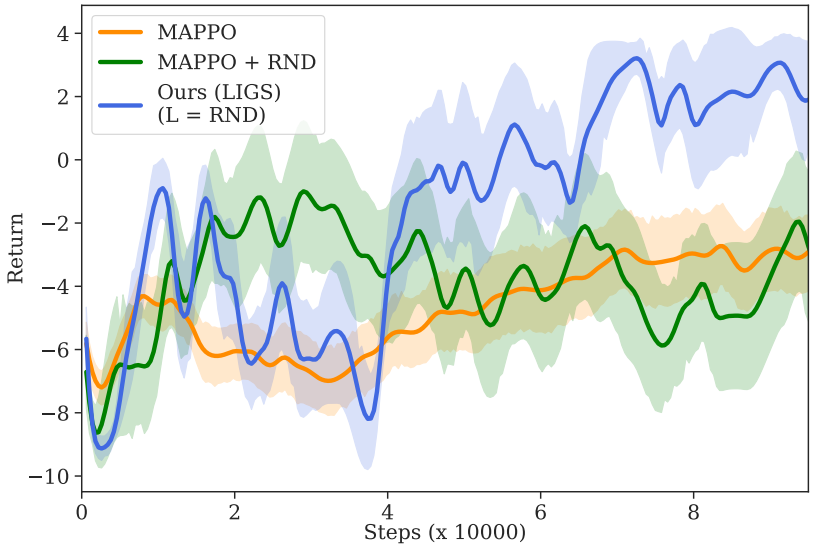

Figure 8: Performance curves for LIGS, MAPPO with RND intrinsic rewards and vanilla MAPPO. The additional machinery of switching-controls and intrinsic reward selection allows LIGS to make better use of exploration bonuses. In this case, LIGS demonstrates significant improvement over MAPPO with RND intrinsic rewards.

14 FURTHER IMPLEMENTATION DETAILS

Details of the Generator and F (intrinsic-reward)

Object	Description
Θ	Discrete action set which is size of output of f, i.e., Θ is set of integers $\{1, ..., m\}$
g	Fixed feed forward NN that maps $\mathbb{R}^d \mapsto \mathbb{R}^m$ [512, `ReLU`, 512, `ReLU`, 512, m]
F	$\gamma\theta^c_{t+1}$ - θ^c_t, $\gamma = 0.95$

d=Dimensionality of states; $m \in \mathbb{N}$ - tunable free parameter.

In all experiments we used the above form of F as follows: a state s_t is input to the g network and the network outputs logits p_t. we softmax and sample from p_t to obtain the action θ^c_t. This action is one-hot encoded. In this way the policy of the Generator chooses the intrinsic-reward.

14.1 HYPERPARAMETER SETTINGS

In the table below we report all hyperparameters used in our experiments. Hyperparameter values in square brackets indicate ranges of values that were used for performance tuning.

Clip Gradient Norm	1
γ_E	0.99
λ	0.95
Learning rate	1x10^{-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
Coefficient of extrinsic reward	[1, 5]
Coefficient of intrinsic reward	[1, 2, 5, 10, 20, 50]
Generator discount factor	0.99
Probability of terminating option	[0.5, 0.75, 0.8, 0.9, 0.95]
L function output size	[2, 4, 8, 16, 32, 64, 128, 256]

15 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 $(\mathcal{V}, |||)$ any finite normed vector space and by $\mathcal{H}$ the set of all measurable functions. Where it will not cause confusion (and with a minor abuse of notation) for a given function h we use the shorthand $h^{(\pi^i,\pi^{-i})}(s) = h(s, \pi^i, \pi^{-i}) \equiv \mathbb{E}_{\pi^i,\pi^{-i}}[h(s, a^i, a^{-i})]$.

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 constituent functions of the agents' objectives R, F and L are 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: $\mathbb{E}\left[1 + \|s_t\|^c | s_0 = s\right] \leq \mu_c(1 + \|s\|^c)$.

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

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

We also make the following finiteness assumption on set of switching control policies for the Generator:

Assumption 6 For any policy $\mathfrak{g}_c$, the total number of interventions is $K < \infty$.

We lastly make the following assumption on L which can be made true by construction:

Assumption 7 Let $n(s)$ be the state visitation count for a given state $s \in \mathcal{S}$. For any $\boldsymbol{a} \in \boldsymbol{\mathcal{A}}$, the function $L(s, \boldsymbol{a}) = 0$ for any $n(s) \geq M$ where $0 < M \leq \infty$.

16 PROOF OF TECHNICAL RESULTS

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

Given a $V^{\boldsymbol{\pi},g} : \mathcal{S} \times \mathbb{N} \to \mathbb{R},\ \forall \boldsymbol{\pi} \in \mathbf{\Pi}$ and g, $\forall s_{\tau_k} \in \mathcal{S}$, we define the Generator intervention operator $\mathcal{M}^{\boldsymbol{\pi},g}V^{\boldsymbol{\pi},g}$ by

$$\mathcal{M}^{\boldsymbol{\pi},g}V^{\boldsymbol{\pi},g}(s_{\tau_k}, I_{\tau_k}) := R(s_{\tau_k}, \boldsymbol{a}_{\tau_k}) + F^{(\theta_{\tau_k}, \theta_{\tau_{k-1}})} - \delta^{\tau_k}_{\tau_k} + \gamma \sum_{s' \in \mathcal{S}} P(s'; \boldsymbol{a}_{\tau_k}, s) V^{\boldsymbol{\pi},g}(s', I(\tau_{k+1})), \tag{4}$$

where $\boldsymbol{a}_{\tau_k} \sim \boldsymbol{\pi}(\cdot|s_{\tau_k})$, $\theta_{\tau_k} \sim g(\cdot|s_{\tau_k})$ and τ_k is a Generator switching time. We define the Bellman operator T of $\mathcal{G}$ by

$$TV^{\boldsymbol{\pi},g}(s_t, I_t) := \max\left\{\mathcal{M}^{\boldsymbol{\pi},g}V^{\boldsymbol{\pi},g}(s_t, I_t), R(s_t, \boldsymbol{a}_t) + \gamma \max_{\boldsymbol{a} \in \boldsymbol{\mathcal{A}}} \sum_{s' \in \mathcal{S}} P(s'; \boldsymbol{a}, s_t) V^{\boldsymbol{\pi},g}(s', I_t)\right\}. \tag{5}$$

Definition 1 *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 $c \in [0, 1[$ such that for any $V_1, V_2 \in \mathcal{V}$ we have that:*

$$\|TV_1 - TV_2\| \leq c\|V_1 - V_2\|. \tag{6}$$

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

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

Lemma 1 *For any $f : \mathcal{V} \to \mathbb{R}, g : \mathcal{V} \to \mathbb{R}$, we have that:*

$$\left\|\max_{a \in \mathcal{V}} f(a) - \max_{a \in \mathcal{V}} g(a)\right\| \leq \max_{a \in \mathcal{V}} \|f(a) - g(a)\|. \tag{8}$$

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

$$f(a) \leq \|f(a) - g(a)\| + g(a) \tag{9}$$

$$\implies \max_{a \in \mathcal{V}} f(a) \leq \max_{a \in \mathcal{V}} \{\|f(a) - g(a)\| + g(a)\} \leq \max_{a \in \mathcal{V}} \|f(a) - g(a)\| + \max_{a \in \mathcal{V}} g(a). \tag{10}$$

Deducting $\max_{a \in \mathcal{V}} g(a)$ from both sides of (10) yields:

$$\max_{a \in \mathcal{V}} f(a) - \max_{a \in \mathcal{V}} g(a) \leq \max_{a \in \mathcal{V}} \|f(a) - g(a)\|. \tag{11}$$

After reversing the roles of f and g and redoing steps (9) - (10), we deduce the desired result since the RHS of (11) is unchanged. □

Lemma 2 *A.4 The probability transition kernel P is non-expansive, that is:*

$$\|PV_1 - PV_2\| \leq \|V_1 - V_2\|. \tag{12}$$

Proof: 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:

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

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

PROOF OF PROP. 1

Proof: To prove (i) of the proposition it suffices to prove that the term $\sum_{t=0}^{T} \gamma^t F(\theta^c_t, \theta^c_{t-1}) I(t)$ converges to 0 in the limit as $T \to \infty$. As in classic potential-based reward shaping (Ng et al., 1999), central to this observation is the telescoping sum that emerges by construction of F.

First recall $\tilde{v}^{\boldsymbol{\pi},\mathfrak{g}}(s, I_0)$, for any $(s, I_0) \in \mathcal{S} \times \{0, 1\}$ is given by:

$$\tilde{v}^{\boldsymbol{\pi},\mathfrak{g}}(s, I_0) = \mathbb{E}_{\boldsymbol{\pi},g}\left[\sum_{t=0}^{\infty} \gamma^t \left\{R(s_t, \boldsymbol{a}_t) + F(\theta^c_t, \theta^c_{t-1})I_t\right\}\right] \tag{13}$$

$$= \mathbb{E}_{\boldsymbol{\pi},g}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, \boldsymbol{a}_t) + \sum_{t=0}^{\infty} \gamma^t F(\theta^c_t, \theta^c_{t-1})I_t\right] \tag{14}$$

$$= \mathbb{E}_{\boldsymbol{\pi},g}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, \boldsymbol{a}_t)\right] + \mathbb{E}_{\boldsymbol{\pi},g}\left[\sum_{t=0}^{\infty} \gamma^t F(\theta^c_t, \theta^c_{t-1}))I_t\right]. \tag{15}$$

Hence it suffices to prove that $\mathbb{E}_{\boldsymbol{\pi},g}\left[\sum_{t=0}^{\infty} \gamma^t F(\theta^c_t, \theta^c_{t-1}))I_t\right] = 0$.

Recall there a number of time steps that elapse between τ_k and τ_{k+1}, now

$$\begin{aligned}
&\sum_{t=0}^{\infty} \gamma^t F(\theta^c_t, \theta^c_{t-1}))I(t) \\
&= \sum_{t=\tau_1+1}^{\tau_2} \gamma^t \theta^c_t - \gamma^{t-1}\theta^c_{t-1} + \gamma^{\tau_1}\theta^c_{\tau_1} + \sum_{t=\tau_3+1}^{\tau_4} \gamma^t \theta^c_t - \gamma^{t-1}\theta^c_{t-1} + \gamma^{\tau_3}\theta^c_{\tau_3} \\
&\quad + \ldots + \sum_{t=\tau_{(2k-1)}+1}^{\tau_{2k}} \gamma^t \theta^c_t - \gamma^{t-1}\theta^c_{t-1} + \gamma^{\tau_1}\theta^c_{\tau_{2k+1}} + \ldots + \\
&= \sum_{t=\tau_1}^{\tau_2-1} \gamma^{t+1}\theta^c_{t+1} - \gamma^t\theta^c_t + \gamma^{\tau_1}\theta^c_{\tau_1} + \sum_{t=\tau_3}^{\tau_4-1} \gamma^{t+1}\theta^c_{t+1} - \gamma^t\theta^c_t + \gamma^{\tau_3}\theta^c_{\tau_3} \\
&\quad + \ldots + \sum_{t=\tau_{(2k-1)}}^{\tau_{2K-1}} \gamma^t \theta^c_t - \gamma^{t-1}\theta^c_{t-1} + \gamma^{\tau_{2k-1}}\theta^c_{\tau_{2k-1}} + \ldots + \\
&= \sum_{k=1}^{\infty} \sum_{t=\tau_{2k-1}}^{\tau_{2K-1}} \gamma^{t+1}\theta^c_{t+1} - \gamma^t\theta^c_t - \sum_{k=1}^{\infty} \gamma^{\tau_{2k-1}}\theta^c_{\tau_{2k-1}} \\
&= \sum_{k=1}^{\infty} \gamma^{\tau_{2k}}\theta^c_{\tau_{2k}} - \sum_{k=1}^{\infty} \gamma^{\tau_{2k-1}}\theta^c_{\tau_{2k-1}} \\
&= \sum_{k=1}^{\infty} \gamma^{\tau_{2k}} 0 - \sum_{k=1}^{\infty} \gamma^{\tau_{2k-1}} 0 = 0,
\end{aligned}$$

where we have used the fact that by construction $\theta^c_t \equiv 0$ whenever $t = \tau_1, \tau_2, \ldots$.

We now note that it is easy to see that $\hat{v}^{\boldsymbol{\pi},\mathfrak{g}}_c(s_0, I_0)$ is bounded above, indeed using the above we have that

$$\hat{v}_c^{\boldsymbol{\pi},\mathfrak{g}}(s_0, I_0) = \mathbb{E}_{\boldsymbol{\pi},g}\left[\sum_{t=0}^{\infty}\gamma^t\left(\hat{R} - \sum_{k\geq 1}\delta^t_{\tau_{2k-1}} + L_n(s_t)\right)\right] \tag{16}$$

$$= \mathbb{E}_{\boldsymbol{\pi},g}\left[\sum_{t=0}^{\infty}\gamma^t\left(R - \sum_{k\geq 1}\delta^t_{\tau_{2k-1}} + L_n(s_t)\right) + \sum_{t=0}^{\infty}\gamma^t F I_t\right] \tag{17}$$

$$\leq \mathbb{E}_{\boldsymbol{\pi},g}\left[\sum_{t=0}^{\infty}\gamma^t\left(R + L_n(s_t)\right)\right] \tag{18}$$

$$\leq \left|\mathbb{E}_{\boldsymbol{\pi},g}\left[\sum_{t=0}^{\infty}\gamma^t\left(R + L_n(s_t)\right)\right]\right| \tag{19}$$

$$\leq \mathbb{E}_{\boldsymbol{\pi},g}\left[\sum_{t=0}^{\infty}\gamma^t\left\|R + L_n\right\|\right] \tag{20}$$

$$\leq \sum_{t=0}^{\infty}\gamma^t\left(\|R\| + \|L_n\|\right) \tag{21}$$

$$= \frac{1}{1-\gamma}\left(\|R\| + \|L\|\right), \tag{22}$$

using the triangle inequality, the definition of $\hat{R}$ and the (upper-)boundedness of L and R (Assumption 5). We now note that by the dominated convergence theorem we have that $\forall(s_0, I_0) \in \mathcal{S}\times\{0,1\}$

$$\lim_{n\to\infty}\hat{v}_c^{\boldsymbol{\pi},\mathfrak{g}}(s_0, I_0) = \lim_{n\to\infty}\mathbb{E}_{\boldsymbol{\pi},g}\left[\sum_{t=0}^{\infty}\gamma^t\left(\hat{R} - \sum_{k\geq 1}\delta^t_{\tau_{2k-1}} + L_n(s_t)\right)\right] \tag{23}$$

$$= \mathbb{E}_{\boldsymbol{\pi},g}\lim_{n\to\infty}\left[\sum_{t=0}^{\infty}\gamma^t\left(\hat{R} - \sum_{k\geq 1}\delta^t_{\tau_{2k-1}} + L_n(s_t)\right)\right] \tag{24}$$

$$= \mathbb{E}_{\boldsymbol{\pi},g}\left[\sum_{t=0}^{\infty}\gamma^t\left(\hat{R} - \sum_{k\geq 1}\delta^t_{\tau_{2k-1}}\right)\right] \tag{25}$$

$$= \mathbb{E}_{\boldsymbol{\pi},g}\left[\sum_{t=0}^{\infty}\gamma^t\left(R - \sum_{k\geq 1}\delta^t_{\tau_{2k-1}}\right)\right] = -\frac{K}{1-\gamma} + v^{\pi}(s_0), \tag{26}$$

using Assumption 6 in the last step, after which we deduce (i).

To deduce (ii) we simply note that $\hat{v}_c^{\boldsymbol{\pi},\mathfrak{g}}(s_0, I_0)$ and $v^{\pi}(s_0)$ differ by only a constant and hence share the same optimisation.

□

PROOF OF THEOREM 1

Proof: Theorem 1 is proved by firstly showing that when the players jointly maximise the same objective there exists a fixed point equilibrium of the game when all players use Markov policies and Generator uses switching control. The proof then proceeds by showing that the MG $\mathcal{G}$ admits a dual representation as an MG in which jointly maximise the same objective which has a stable point that can be computed by solving an MDP. Thereafter, we use both results to prove the existence of a fixed point for the game as a limit point of a sequence generated by successively applying the Bellman operator to a test function.

Therefore, the scheme of the proof is summarised with the following steps:

I) Prove that the solution to Markov Team games (that is games in which both players maximise *identical objectives*) in which one of the players uses switching control is the limit point of a sequence of Bellman operators (acting on some test function).

II) Prove that for the MG $\mathcal{G}$ that is there exists a function $B^{\boldsymbol{\pi},\mathfrak{g}} : \mathcal{S} \times \{0,1\} \to \mathbb{R}$ such that[5] $v^{\boldsymbol{\pi},\mathfrak{g}}(z) - v^{\boldsymbol{\pi}',\mathfrak{g}}(z) = B^{\boldsymbol{\pi},\mathfrak{g}}(z) - B^{\boldsymbol{\pi}',\mathfrak{g}}(z), \quad \forall z \equiv (s, I_0) \in \mathcal{S} \times \{0,1\}, \ \forall \mathfrak{g}$, and $\hat{v}_c^{\boldsymbol{\pi},\mathfrak{g}}(z) - \hat{v}_c^{\boldsymbol{\pi},\mathfrak{g}'}(z) = B^{\boldsymbol{\pi},\mathfrak{g}}(z) - B^{\boldsymbol{\pi}',\mathfrak{g}}(z), \quad \forall z \equiv (s, I_0) \in \mathcal{S} \times \{0,1\}, \forall \boldsymbol{\pi} \in \boldsymbol{\Pi}$,

III) Prove that the MG $\mathcal{G}$ has a dual representation as a *Markov Team Game* which admits a representation as an MDP.

PROOF OF PART **I**

Our first result proves that the operator T is a contraction operator. First let us recall that the *switching time* τ_k is defined recursively $\tau_k = \inf\{t > \tau_{k-1} | s_t \in A, \tau_k \in \mathcal{F}_t\}$ where $A = \{s \in \mathcal{S}, m \in M | \mathfrak{g}_c(m|s_t) > 0\}$. To this end, we show that the following bounds holds:

Lemma 3 *The Bellman operator T is a contraction, that is the following bound holds:*

$$\|T\psi - T\psi'\| \leq \gamma \|\psi - \psi'\|.$$

Proof: Recall we define the Bellman operator T_ψ of $\mathcal{G}$ acting on a function $\Lambda : \mathcal{S} \times \mathbb{N} \to \mathbb{R}$ by

$$T_\psi \Lambda(s_{\tau_k}, I(\tau_k)) := \max \left\{ \mathcal{M}^{\boldsymbol{\pi},g} \Lambda(s_{\tau_k}, I(\tau_k)), \left[\psi(s_{\tau_k}, \boldsymbol{a}) + \gamma \max_{\boldsymbol{a} \in \boldsymbol{\mathcal{A}}} \sum_{s' \in \mathcal{S}} P(s'; \boldsymbol{a}, s_{\tau_k}) \Lambda(s', I(\tau_k)) \right] \right\} \tag{27}$$

In what follows and for the remainder of the script, we employ the following shorthands:

$$\mathcal{P}^{\boldsymbol{a}}_{ss'} =: \sum_{s' \in \mathcal{S}} P(s'; \boldsymbol{a}, s), \quad \mathcal{P}^{\boldsymbol{\pi}}_{ss'} =: \sum_{\boldsymbol{a} \in \boldsymbol{\mathcal{A}}} \boldsymbol{\pi}(\boldsymbol{a}|s) \mathcal{P}^{\boldsymbol{a}}_{ss'}, \quad \mathcal{R}^{\boldsymbol{\pi}}(z_t) := \sum_{\boldsymbol{a}_t \in \boldsymbol{\mathcal{A}}} \boldsymbol{\pi}(\boldsymbol{a}_t|s) \hat{R}(z_t, \boldsymbol{a}_t, \theta_t, \theta_{t-1})$$

To prove that T is a contraction, we consider the three cases produced by (27), that is to say we prove the following statements:

i) $\left| \Theta(z_t, \boldsymbol{a}, \theta^c_t, \theta^c_{t-1}) + \gamma \max_{\boldsymbol{a} \in \boldsymbol{\mathcal{A}}} \mathcal{P}^{\boldsymbol{a}}_{s's_t} \psi(s', \cdot) - \left(\Theta(z_t, \boldsymbol{a}, \theta^c_t, \theta^c_{t-1}) + \gamma \max_{\boldsymbol{a} \in \boldsymbol{\mathcal{A}}} \mathcal{P}^{\boldsymbol{a}}_{s's_t} \psi'(s', \cdot) \right) \right| \leq \gamma \|\psi - \psi'\|$

ii) $\|\mathcal{M}^{\boldsymbol{\pi},g} \psi - \mathcal{M}^{\boldsymbol{\pi},g} \psi'\| \leq \gamma \|\psi - \psi'\|$, (and hence $\mathcal{M}$ is a contraction).

iii) $\left\| \mathcal{M}^{\boldsymbol{\pi},g} \psi - \left[\Theta(\cdot, \boldsymbol{a}) + \gamma \max_{\boldsymbol{a} \in \boldsymbol{\mathcal{A}}} \mathcal{P}^{\boldsymbol{a}} \psi' \right] \right\| \leq \gamma \|\psi - \psi'\|$. where $z_t \equiv (s_t, I_t) \in \mathcal{S} \times \{0,1\}$.

We begin by proving i).

Indeed, for any $\boldsymbol{a} \in \boldsymbol{\mathcal{A}}$ and $\forall z_t \in \mathcal{S} \times \{0,1\}, \forall \theta_t, \theta_{t-1} \in \Theta, \forall s' \in \mathcal{S}$ we have that

$$\begin{aligned}
&\left| \Theta(z_t, \boldsymbol{a}, \theta^c_t, \theta^c_{t-1}) + \gamma \mathcal{P}^{\pi}_{s's_t} \psi(s', \cdot) - \left[\Theta(z_t, \boldsymbol{a}, \theta^c_t, \theta^c_{t-1}) + \gamma \max_{\boldsymbol{a} \in \boldsymbol{\mathcal{A}}} \mathcal{P}^{\boldsymbol{a}}_{s's_t} \psi'(s', \cdot) \right] \right| \\
&\leq \max_{\boldsymbol{a} \in \boldsymbol{\mathcal{A}}} \left| \gamma \mathcal{P}^{\boldsymbol{a}}_{s's_t} \psi(s', \cdot) - \gamma \mathcal{P}^{\boldsymbol{a}}_{s's_t} \psi'(s', \cdot) \right| \\
&\leq \gamma \|P\psi - P\psi'\| \\
&\leq \gamma \|\psi - \psi'\|,
\end{aligned}$$

again using the fact that P is non-expansive and Lemma 1.

We now prove ii).

[5] This property is analogous to the condition in Markov potential games (Macua et al., 2018; Mguni et al., 2021)

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

$$
\begin{aligned}
&|(\mathcal{M}^{\boldsymbol{\pi},g}\psi - \mathcal{M}^{\boldsymbol{\pi},g}\psi')(s_\tau, I(\tau))| \\
&\leq \max_{\boldsymbol{a}_\tau, \theta^c_\tau, \theta^c_{\tau-1} \in \boldsymbol{\mathcal{A}} \times \Theta^2} \Big| \Theta(z_\tau, \boldsymbol{a}_\tau, \theta^c_\tau, \theta^c_{\tau-1}) - \delta^\tau_t + \gamma \mathcal{P}^{\boldsymbol{\pi}}_{s's_\tau} \mathcal{P}^{\boldsymbol{a}} \psi(s_\tau, I(\tau')) \\
&\qquad\qquad - \left(\Theta(z_\tau, \boldsymbol{a}_\tau, \theta^c_\tau, \theta^c_{\tau-1}) - \delta^\tau_t + \gamma \mathcal{P}^{\boldsymbol{\pi}}_{s's_\tau} \mathcal{P}^{\boldsymbol{a}} \psi'(s_\tau, I(\tau'))\right) \Big| \\
&= \gamma \left| \mathcal{P}^{\boldsymbol{\pi}}_{s's_\tau} \mathcal{P}^{\boldsymbol{a}} \psi(s_\tau, I(\tau')) - \mathcal{P}^{\boldsymbol{\pi}}_{s's_\tau} \mathcal{P}^{\boldsymbol{a}} \psi'(s_\tau, I(\tau')) \right| \\
&\leq \gamma \left\| P\psi - P\psi' \right\| \\
&\leq \gamma \left\| \psi - \psi' \right\|,
\end{aligned}
$$

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 1:

$$
\mathcal{M}^{\boldsymbol{\pi},g}\psi(s_\tau, I(\tau)) - \left(\Theta(z_\tau, \boldsymbol{a}_\tau, \theta^c_\tau, \theta^c_{\tau-1}) + \gamma \max_{\boldsymbol{a} \in \boldsymbol{\mathcal{A}}} \mathcal{P}^{\boldsymbol{a}}_{s's_\tau} \psi'(s', I(\tau))\right) < 0. \tag{28}
$$

We now observe the following:

$$
\begin{aligned}
&\mathcal{M}^{\boldsymbol{\pi},g}\psi(s_\tau, I(\tau)) - \Theta(z_\tau, \boldsymbol{a}_\tau, \theta^c_\tau, \theta^c_{\tau-1}) + \gamma \max_{\boldsymbol{a} \in \boldsymbol{\mathcal{A}}} \mathcal{P}^{\boldsymbol{a}}_{s's_\tau} \psi'(s', I(\tau)) \\
&\leq \max\left\{\Theta(z_\tau, \boldsymbol{a}_\tau, \theta^c_\tau, \theta^c_{\tau-1}) + \gamma \mathcal{P}^{\boldsymbol{\pi}}_{s's_\tau} \mathcal{P}^{\boldsymbol{a}} \psi(s', I(\tau)), \mathcal{M}^{\boldsymbol{\pi},g}\psi(s_\tau, I(\tau))\right\} \\
&\qquad - \Theta(z_\tau, \boldsymbol{a}_\tau, \theta^c_\tau, \theta^c_{\tau-1}) + \gamma \max_{\boldsymbol{a} \in \boldsymbol{\mathcal{A}}} \mathcal{P}^{\boldsymbol{a}}_{s's_\tau} \psi'(s', I(\tau)) \\
&\leq \Bigg| \max\left\{\Theta(z_\tau, \boldsymbol{a}_\tau, \theta^c_\tau, \theta^c_{\tau-1}) + \gamma \mathcal{P}^{\boldsymbol{\pi}}_{s's_\tau} \mathcal{P}^{\boldsymbol{a}} \psi(s', I(\tau)), \mathcal{M}^{\boldsymbol{\pi},g}\psi(s_\tau, I(\tau))\right\} \\
&\qquad\qquad - \max\left\{\Theta(z_\tau, \boldsymbol{a}_\tau, \theta^c_\tau, \theta^c_{\tau-1}) + \gamma \max_{\boldsymbol{a} \in \boldsymbol{\mathcal{A}}} \mathcal{P}^{\boldsymbol{a}}_{s's_\tau} \psi'(s', I(\tau)), \mathcal{M}^{\boldsymbol{\pi},g}\psi(s_\tau, I(\tau))\right\} \\
&\qquad + \max\left\{\Theta(z_\tau, \boldsymbol{a}_\tau, \theta^c_\tau, \theta^c_{\tau-1}) + \gamma \max_{\boldsymbol{a} \in \boldsymbol{\mathcal{A}}} \mathcal{P}^{\boldsymbol{a}}_{s's_\tau} \psi'(s', I(\tau)), \mathcal{M}^{\boldsymbol{\pi},g}\psi(s_\tau, I(\tau))\right\} \\
&\qquad\qquad - \Theta(z_\tau, \boldsymbol{a}_\tau, \theta^c_\tau, \theta^c_{\tau-1}) + \gamma \max_{\boldsymbol{a} \in \boldsymbol{\mathcal{A}}} \mathcal{P}^{\boldsymbol{a}}_{s's_\tau} \psi'(s', I(\tau)) \Bigg| \\
&\leq \Bigg| \max\left\{\Theta(z_\tau, \boldsymbol{a}_\tau, \theta^c_\tau, \theta^c_{\tau-1}) + \gamma \max_{\boldsymbol{a} \in \boldsymbol{\mathcal{A}}} \mathcal{P}^{\boldsymbol{a}}_{s's_\tau} \psi(s', I(\tau)), \mathcal{M}^{\boldsymbol{\pi},g}\psi(s_\tau, I(\tau))\right\} \\
&\qquad\qquad - \max\left\{\Theta(z_\tau, \boldsymbol{a}_\tau, \theta^c_\tau, \theta^c_{\tau-1}) + \gamma \max_{\boldsymbol{a} \in \boldsymbol{\mathcal{A}}} \mathcal{P}^{\boldsymbol{a}}_{s's_\tau} \psi'(s', I(\tau)), \mathcal{M}^{\boldsymbol{\pi},g}\psi(s_\tau, I(\tau))\right\} \Bigg| \\
&\qquad + \Bigg| \max\left\{\Theta(z_\tau, \boldsymbol{a}_\tau, \theta^c_\tau, \theta^c_{\tau-1}) + \gamma \max_{\boldsymbol{a} \in \boldsymbol{\mathcal{A}}} \mathcal{P}^{\boldsymbol{a}}_{s's_\tau} \psi'(s', I(\tau)), \mathcal{M}^{\boldsymbol{\pi},g}\psi(s_\tau, I(\tau))\right\} \\
&\qquad\qquad - \Theta(z_\tau, \boldsymbol{a}_\tau, \theta^c_\tau, \theta^c_{\tau-1}) + \gamma \max_{\boldsymbol{a} \in \boldsymbol{\mathcal{A}}} \mathcal{P}^{\boldsymbol{a}}_{s's_\tau} \psi'(s', I(\tau)) \Bigg| \\
&\leq \gamma \max_{a \in \mathcal{A}} \left| \mathcal{P}^{\boldsymbol{\pi}}_{s's_\tau} \mathcal{P}^{\boldsymbol{a}} \psi(s', I(\tau)) - \mathcal{P}^{\boldsymbol{\pi}}_{s's_\tau} \mathcal{P}^{\boldsymbol{a}} \psi'(s', I(\tau)) \right| \\
&\qquad + \left| \max\left\{0, \mathcal{M}^{\boldsymbol{\pi},g}\psi(s_\tau, I(\tau)) - \left(\Theta(z_\tau, \boldsymbol{a}_\tau, \theta^c_\tau, \theta^c_{\tau-1}) + \gamma \max_{\boldsymbol{a} \in \boldsymbol{\mathcal{A}}} \mathcal{P}^{\boldsymbol{a}}_{s's_\tau} \psi'(s', I(\tau))\right)\right\}\right| \\
&\leq \gamma \left\| P\psi - P\psi' \right\| \\
&\leq \gamma \|\psi - \psi'\|,
\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:

$$\mathcal{M}^{\boldsymbol{\pi},g}\psi(s_\tau, I(\tau)) - \left(\Theta(z_\tau, \boldsymbol{a}_\tau, \theta^c_\tau, \theta^c_{\tau-1}) + \gamma \max_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}} \mathcal{P}^{\boldsymbol{a}}_{s's_\tau}\psi'(s', I(\tau))\right) \geq 0.$$

$$\begin{aligned}
&\mathcal{M}^{\boldsymbol{\pi},g}\psi(s_\tau, I(\tau)) - \left(\Theta(z_\tau, \boldsymbol{a}_\tau, \theta^c_\tau, \theta^c_{\tau-1}) + \gamma \max_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}} \mathcal{P}^{\boldsymbol{a}}_{s's_\tau}\psi'(s', I(\tau))\right) \\
&\leq \mathcal{M}^{\boldsymbol{\pi},g}\psi(s_\tau, I(\tau)) - \left(\Theta(z_\tau, \boldsymbol{a}_\tau, \theta^c_\tau, \theta^c_{\tau-1}) + \gamma \max_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}} \mathcal{P}^{\boldsymbol{a}}_{s's_\tau}\psi'(s', I(\tau))\right) + \delta^\tau_t \\
&\leq \Theta(z_\tau, \boldsymbol{a}_\tau, \theta^c_\tau, \theta^c_{\tau-1}) - \delta^\tau_t + \gamma\mathcal{P}^{\boldsymbol{\pi}}_{s's_\tau}\mathcal{P}^{\boldsymbol{a}}\psi(s', I(\tau')) \\
&\qquad\qquad - \left(\Theta(z_\tau, \boldsymbol{a}_\tau, \theta^c_\tau, \theta^c_{\tau-1}) - \delta^\tau_t + \gamma \max_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}} \mathcal{P}^{\boldsymbol{a}}_{s's_\tau}\psi'(s', I(\tau))\right) \\
&\leq \gamma \max_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}} \left|\mathcal{P}^{\boldsymbol{\pi}}_{s's_\tau}\mathcal{P}^{\boldsymbol{a}}\left(\psi(s', I(\tau')) - \psi'(s', I(\tau))\right)\right| \\
&\leq \gamma \left|\psi(s', I(\tau')) - \psi'(s', I(\tau))\right| \\
&\leq \gamma \left\|\psi - \psi'\right\|,
\end{aligned}$$

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

$$\left\|\mathcal{M}^{\boldsymbol{\pi},g}\Lambda - \max_{\boldsymbol{a}\in\boldsymbol{\mathcal{A}}} \left[\psi(\cdot, a) + \gamma\mathcal{P}^{\boldsymbol{a}}\Lambda'\right]\right\| \leq \gamma \left\|\Lambda - \Lambda'\right\|. \tag{29}$$

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

PROOF OF PART **II**

To prove Part **II**, we prove the following result:

Proposition 3 *For any $\pi \in \Pi$ and for any Generator policy $\mathfrak{g}$, there exists a function $B^{\boldsymbol{\pi},\mathfrak{g}} : \mathcal{S} \times \{0,1\} \to \mathbb{R}$ such that*

$$v_i^{\boldsymbol{\pi},\mathfrak{g}} - v_i^{\boldsymbol{\pi}',\mathfrak{g}} = B^{\boldsymbol{\pi},\mathfrak{g}}(z) - B^{\boldsymbol{\pi}',\mathfrak{g}}(z), \quad \forall z \equiv (s, I_0) \in \mathcal{S} \times \{0,1\} \tag{30}$$

where in particular the function B is given by:

$$B^{\boldsymbol{\pi},\mathfrak{g}}(s_0, I_0) = \mathbb{E}_{\boldsymbol{\pi},\mathfrak{g}}\left[\sum_{t=0}^{\infty}\gamma^t R\right], \tag{31}$$

for any $(s_0, I_0) \in \mathcal{S} \times \{0,1\}$.

Proof: Note that by the deduction of (ii) in Prop 1, we may consider the following quantity for the Generator expected return:

$$\hat{v}_c^{\boldsymbol{\pi},\mathfrak{g}}(s_0, I_0) = \mathbb{E}_{\boldsymbol{\pi},\mathfrak{g}}\left[\sum_{t=0}^{\infty}\gamma^t\left(R - \sum_{k\geq 1}\delta^t_{\tau_{2k-1}}\right)\right]. \tag{32}$$

Therefore, we immediately observe that

$$\hat{v}_c^{\boldsymbol{\pi},\mathfrak{g}}(s_0, I_0) = B^{\boldsymbol{\pi},\mathfrak{g}}(s_0, I_0) - K, \quad \forall (s_0, I_0) \in \mathcal{S} \times \{0,1\}. \tag{33}$$

We therefore immediately deduce that for any two Generator policies $\mathfrak{g}$ and $\mathfrak{g}'$ the following expression holds $\forall (s_0, I_0) \in \mathcal{S} \times \{0,1\}$:

$$\hat{v}_c^{\boldsymbol{\pi},\mathfrak{g}}(s_0, I_0) - \hat{v}_c^{\boldsymbol{\pi},\mathfrak{g}'}(s_0, I_0) = B^{\boldsymbol{\pi},\mathfrak{g}}(s_0, I_0) - B^{\pi,\mathfrak{g}'}(s_0, I_0). \tag{34}$$

Our aim now is to show that the following expression holds $\forall (s_0, I_0) \in \mathcal{S} \times \{0,1\}$:

$$\hat{v}_c^{\boldsymbol{\pi},\mathfrak{g}}(I_0, s_0) - \hat{v}_c^{\boldsymbol{\pi}',\mathfrak{g}}(I_0, s_0) = B^{\boldsymbol{\pi},\mathfrak{g}}(I_0, s_0) - B^{\boldsymbol{\pi}',\mathfrak{g}}(I_0, s_0),$$

This is manifest from the construction of B. □

PROOF OF PART III

To prove Part **III**, we firstly define precisely the notion of a stable point of the MG, $\mathcal{G}$:

Definition 3 *A policy profile $\boldsymbol{\sigma}^{\star} = (g^{\star}, \pi_i^{\star}, \pi_{-i}^{\star}) \in \boldsymbol{\Pi}$ is a Markov perfect equilibrium (MPE) in Markov strategies if the following condition holds for any $i \in \mathcal{N} \times \{0\}$:*

$$v_i^{(g^{\star},\pi_i^{\star},\pi_{-i}^{\star})}(z) \geq v_i^{g^{\star},(\pi_i',\pi_{-i}^{\star})}(z),\ \forall z \equiv (s_0, I_0) \in \mathcal{S} \times \{0,1\},\ \forall \pi_i' \in \Pi_i. \tag{35}$$

$$v_c^{(g^{\star},\pi_i^{\star},\pi_{-i}^{\star})}(z) \geq v_c^{g',(\pi_i,\pi_{-i}^{\star})}(z),\ \forall z \equiv (s_0, I_0) \in \mathcal{S} \times \{0,1\},\ \forall g'. \tag{36}$$

The condition characterises strategic configurations which are stable points of the MG, $\mathcal{G}$. In particular, an MPE is achieved when at any state no agent can improve their expected cumulative rewards by unilaterally deviating from their current policy. We denote by $NE\{\mathcal{G}\}$ the set of MPE strategies for the MG, $\mathcal{G}$.

Next we prove that the set of maxima of the function B are the MPE of the MG $\mathcal{G}$:

Proposition 4 *The following implication holds:*

$$\boldsymbol{\sigma} \in \underset{g',\boldsymbol{\pi}' \in \boldsymbol{\Pi}}{\arg\sup}\, B^{g',\boldsymbol{\pi}'}(s) \implies \boldsymbol{\sigma} \in NE\{\mathcal{G}\}. \tag{37}$$

where B is the function in Prop. 3.

Prop. 4 indicates that the game has an equivalent representation in which all agents maximise the same function and thus play a *team game*.

Proof: We do the proof by contradiction. Let $\boldsymbol{\sigma} = (\pi_1, \ldots, \pi_N, g) \in \underset{\boldsymbol{\pi}' \in \boldsymbol{\Pi}, g'}{\arg\sup}\, B^{\boldsymbol{\pi}', g'}(s)$ for any $s \in \mathcal{S}$. Let us now therefore assume that $\boldsymbol{\sigma} \notin NE\{\mathcal{G}\}$, hence there exists some other policy profile $\tilde{\boldsymbol{\sigma}} = (\pi_1, \ldots, \tilde{\pi}_i, \ldots, \pi_N, g)$ which contains at least one profitable deviation by one of the agents $i \in \mathcal{N} \times \{0, \}$. For now let us consider the case in which the profitable deviation is for a agent $i \in \mathcal{N}$ so that $\pi_i' \neq \pi_i$ for $i \in \mathcal{N}$ i.e. $v_i^{(\pi_i',\pi_{-i}),g}(s) > v_i^{(\pi_i,\pi_{-i}),g}(s)$ (using the preservation of signs of integration). Prop. 3 however implies that $B^{(\pi_i',\pi_{-i}),g}(s) - B^{(\pi_i,\pi_{-i}),g}(s) > 0$ which is a contradiction since $\boldsymbol{\sigma} = (\pi_i, \pi_{-i}, g)$ is a maximum of B. The proof can be straightforwardly adapted to cover the case in which the deviating agent is the Generator after which we deduce the desired result. □ The last result completes the proof of Theorem 1. □

PROOF OF PROPOSITION 2

Proof: The proof is given by establishing a contradiction. Therefore suppose that $\mathcal{M}^{\boldsymbol{\pi},g}\psi(s_{\tau_k}, I(\tau_k)) \leq \psi(s_{\tau_k}, I(\tau_k))$ and suppose that the switching time $\tau_1' > \tau_1$ is an optimal switching time. Construct the Generator g' and $\tilde{g}$ policy switching times by $(\tau_0', \tau_1', \ldots,)$ and g'^2 policy by $(\tau_0', \tau_1, \ldots)$ respectively. Define by $l = \inf\{t > 0; \mathcal{M}^{\boldsymbol{\pi},g}\psi(s_t, I_0) = \psi(s_t, I_0)\}$ and $m = \sup\{t; t < \tau_1'\}$. By construction we have that

$$\begin{aligned}
&v_c^{\boldsymbol{\pi},g'}(s, I_0) \\
&= \mathbb{E}\left[R(s_0, \boldsymbol{a}_0) + \mathbb{E}\left[\ldots + \gamma^{l-1}\mathbb{E}\left[R(s_{\tau_1-1}, \boldsymbol{a}_{\tau_1-1}) + \ldots + \gamma^{m-l-1}\mathbb{E}\left[R(s_{\tau_1'-1}, \boldsymbol{a}_{\tau_1'-1}) + \gamma\mathcal{M}^{\boldsymbol{\pi},\mathfrak{g}}v_c^{\boldsymbol{\pi},g'}(s', I(\tau_1'))\right]\right]\right]\right] \\
&< \mathbb{E}\left[R(s_0, \boldsymbol{a}_0) + \mathbb{E}\left[\ldots + \gamma^{l-1}\mathbb{E}\left[R(s_{\tau_1-1}, \boldsymbol{a}_{\tau_1-1}) + \gamma\mathcal{M}^{\boldsymbol{\pi},\tilde{g}}v_c^{\boldsymbol{\pi},g'}(s_{\tau_1}, I(\tau_1))\right]\right]\right]
\end{aligned}$$

We now use the following observation $\mathbb{E}\left[R(s_{\tau_1-1}, \boldsymbol{a}_{\tau_1-1}) + \gamma\mathcal{M}^{\boldsymbol{\pi},\tilde{g}}v_c^{\boldsymbol{\pi},g'}(s_{\tau_1}, I(\tau_1))\right]$

$$\leq \max\left\{\mathcal{M}^{\boldsymbol{\pi},\tilde{g}}v_c^{\boldsymbol{\pi},g'}(s_{\tau_1}, I(\tau_1)), \max_{a_{\tau_1} \in \mathcal{A}}\left[R(s_{\tau_k}, \boldsymbol{a}_{\tau_k}) + \gamma \textstyle\sum_{s' \in \mathcal{S}} P(s'; \boldsymbol{a}_{\tau_1}, s_{\tau_1}) v_c^{\boldsymbol{\pi},g}(s', I(\tau_1))\right]\right\}.$$

Using this we deduce that

$$v_2^{\boldsymbol{\pi},g'}(s,I_0) \leq \mathbb{E}\Bigg[R(s_0,\boldsymbol{a}_0) + \mathbb{E}\Bigg[\dots$$
$$+\gamma^{l-1}\mathbb{E}\left[R(s_{\tau_1-1},\boldsymbol{a}_{\tau_1-1}) + \gamma\max\left\{\mathcal{M}^{\boldsymbol{\pi},\tilde{g}}v_c^{\boldsymbol{\pi},g'}(s_{\tau_1},I(\tau_1)),\ \max_{a_{\tau_1}\in\mathcal{A}}\left[R(s_{\tau_k},\boldsymbol{a}_{\tau_k}) + \gamma\sum_{s'\in\mathcal{S}}P(s';\boldsymbol{a}_{\tau_1},s_{\tau_1})v_c^{\boldsymbol{\pi},g}(s',I(\tau_1))\right]\right.\right.$$
$$= \mathbb{E}\left[R(s_0,\boldsymbol{a}_0) + \mathbb{E}\left[\dots + \gamma^{l-1}\mathbb{E}\left[R(s_{\tau_1-1},\boldsymbol{a}_{\tau_1-1}) + \gamma\left[Tv_c^{\boldsymbol{\pi},\tilde{g}}\right](s_{\tau_1},I(\tau_1))\right]\right]\right] = v_c^{\boldsymbol{\pi},\tilde{g}}(s,I_0))$$

where the first inequality is true by assumption on $\mathcal{M}$. This is a contradiction since g' is an optimal policy for the Generator. 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.

□

PROOF OF THEOREM 2

Proof: The proof which is done by contradiction follows from the definition of v_c. Denote by $v_i^{\boldsymbol{\pi},g\equiv\mathbf{0}}$ value function an agent $i \in \mathcal{N}$ *excluding the Generator* and its intrinsic-reward function. Indeed, let $(\hat{\boldsymbol{\pi}}, \hat{g})$ be the policy profile induced by the Nash equilibrium policy profile and assume that the intrinsic-reward F leads to a decrease in payoff for agent i. Then by construction $v^{\boldsymbol{\pi},g}(s) < v^{\boldsymbol{\pi},g\equiv\mathbf{0}}(s)$ which is a contradiction since $(\hat{\boldsymbol{\pi}}, \hat{g})$ is an MPE profile. □

PROOF OF THEOREM 3

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

Theorem 4 (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:*

$$\Xi_{t+1}(s) = (1-\alpha_t(s))\,\Xi_t(s)\alpha_t(s)L_t(s), \tag{38}$$

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) $\mathrm{Var}\,[L_t|\mathcal{F}_t] \leq c(1+\|\Xi_t\|^2)$ *for some* $c > 0$.

Proof: 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:

$$Q_{t+1}(s_t,I_t,\boldsymbol{a}_t) = Q_t(s_t,I_t,\boldsymbol{a}_t)$$
$$+\alpha_t(s_t,I_t,\boldsymbol{a}_t)\left[\max\left\{\mathcal{M}^{\boldsymbol{\pi},g}Q(s_{\tau_k},I_{\tau_k},\boldsymbol{a}),\phi(s_{\tau_k},\boldsymbol{a}) + \gamma\max_{a'\in\mathcal{A}}Q(s',I_{\tau_k},\boldsymbol{a'})\right\} - Q_t(s_t,I_t,\boldsymbol{a}_t)\right].$$

After subtracting $Q^\star(s_t,I_t,\boldsymbol{a}_t)$ from both sides and some manipulation we obtain that:

$$\Xi_{t+1}(s_t,I_t,\boldsymbol{a}_t)$$
$$= (1-\alpha_t(s_t,I_t,\boldsymbol{a}_t))\Xi_t(s_t,I_t,\boldsymbol{a}_t)$$
$$+\alpha_t(s_t,I_t,\boldsymbol{a}_t))\left[\max\left\{\mathcal{M}^{\boldsymbol{\pi},g}Q(s_{\tau_k},I_{\tau_k},\boldsymbol{a}),\phi(s_{\tau_k},\boldsymbol{a}) + \gamma\max_{a'\in\mathcal{A}}Q(s',I_{\tau_k},\boldsymbol{a'})\right\} - Q^\star(s_t,I_t,\boldsymbol{a}_t)\right],$$

where $\Xi_t(s_t,I_t,\boldsymbol{a}_t) := Q_t(s_t,I_t,\boldsymbol{a}_t) - Q^\star(s_t,I_t,\boldsymbol{a}_t)$.

Let us now define by

$$L_t(s_{\tau_k}, I_{\tau_k}, \boldsymbol{a}) := \max\left\{\mathcal{M}^{\boldsymbol{\pi},g}Q(s_{\tau_k}, I_{\tau_k}, \boldsymbol{a}), \phi(s_{\tau_k}, \boldsymbol{a}) + \gamma \max_{a' \in \mathcal{A}} Q(s', I_{\tau_k}, \boldsymbol{a'})\right\} - Q^\star(s_t, I_t, a).$$

Then

$$\Xi_{t+1}(s_t, I_t, \boldsymbol{a}_t) = (1 - \alpha_t(s_t, I_t, \boldsymbol{a}_t))\Xi_t(s_t, I_t, \boldsymbol{a}_t) + \alpha_t(s_t, I_t, \boldsymbol{a}_t))\left[L_t(s_{\tau_k}, a)\right]. \tag{39}$$

We now observe that

$$\begin{aligned}\mathbb{E}\left[L_t(s_{\tau_k}, I_{\tau_k}, \boldsymbol{a})|\mathcal{F}_t\right] &= \sum_{s' \in \mathcal{S}} P(s'; a, s_{\tau_k}) \max\left\{\mathcal{M}^{\boldsymbol{\pi},g}Q(s_{\tau_k}, I_{\tau_k}, \boldsymbol{a}), \phi(s_{\tau_k}, \boldsymbol{a}) + \gamma \max_{a' \in \mathcal{A}} Q(s', I_{\tau_k}, \boldsymbol{a'})\right\} - Q^\star(s_{\tau_k}, a) \\ &= T_\phi Q_t(s, I_{\tau_k}, \boldsymbol{a}) - Q^\star(s, I_{\tau_k}, \boldsymbol{a}). \end{aligned} \tag{40}$$

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

$$\begin{aligned}\mathbb{E}\left[L_t(s_{\tau_k}, I_{\tau_k}, \boldsymbol{a})|\mathcal{F}_t\right] &= T_\phi Q_t(s, I_{\tau_k}, \boldsymbol{a}) - T_\phi Q^\star(s, I_{\tau_k}, \boldsymbol{a}) \\ &\leq \|T_\phi Q_t - T_\phi Q^\star\| \\ &\leq \gamma \left\|Q_t - Q^\star\right\|_\infty = \gamma \left\|\Xi_t\right\|_\infty. \end{aligned} \tag{41}$$

using the contraction property of T established in Lemma 3. This proves (ii).

We now prove iii), that is

$$\mathrm{Var}\left[L_t|\mathcal{F}_t\right] \leq c(1 + \|\Xi_t\|^2). \tag{42}$$

Now by (40) we have that

$$\begin{aligned}\mathrm{Var}\left[L_t|\mathcal{F}_t\right] &= \mathrm{Var}\left[\max\left\{\mathcal{M}^{\boldsymbol{\pi},g}Q(s_{\tau_k}, I_{\tau_k}, \boldsymbol{a}), \phi(s_{\tau_k}, \boldsymbol{a}) + \gamma \max_{a' \in \mathcal{A}} Q(s', I_{\tau_k}, \boldsymbol{a'})\right\} - Q^\star(s_t, I_t, a)\right] \\ &= \mathbb{E}\Bigg[\Bigg(\max\left\{\mathcal{M}^{\boldsymbol{\pi},g}Q(s_{\tau_k}, I_{\tau_k}, \boldsymbol{a}), \phi(s_{\tau_k}, \boldsymbol{a}) + \gamma \max_{a' \in \mathcal{A}} Q(s', I_{\tau_k}, \boldsymbol{a'})\right\} \\ &\qquad - Q^\star(s_t, I_t, a) - (T_\Phi Q_t(s, I_{\tau_k}, \boldsymbol{a}) - Q^\star(s, I_{\tau_k}, \boldsymbol{a}))\Bigg)^2\Bigg] \\ &= \mathbb{E}\left[\left(\max\left\{\mathcal{M}^{\boldsymbol{\pi},g}Q(s_{\tau_k}, I_{\tau_k}, \boldsymbol{a}), \phi(s_{\tau_k}, \boldsymbol{a}) + \gamma \max_{a' \in \mathcal{A}} Q(s', I_{\tau_k}, \boldsymbol{a'})\right\} - T_\Phi Q_t(s, I_{\tau_k}, \boldsymbol{a})\right)^2\right] \\ &= \mathrm{Var}\left[\max\left\{\mathcal{M}^{\boldsymbol{\pi},g}Q(s_{\tau_k}, I_{\tau_k}, \boldsymbol{a}), \phi(s_{\tau_k}, \boldsymbol{a}) + \gamma \max_{a' \in \mathcal{A}} Q(s', I_{\tau_k}, \boldsymbol{a'})\right\} - T_\Phi Q_t(s, I_{\tau_k}, \boldsymbol{a}))^2\right] \\ &\leq c(1 + \|\Xi_t\|^2), \end{aligned}$$

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 Function Approximation

First let us recall the statement of the theorem:

Theorem 3 *LIGS converges to a limit point $r^\star$ which is the unique solution to the equation:*

$$\Pi\mathfrak{F}(\Phi r^\star) = \Phi r^\star, \qquad a.e. \tag{43}$$

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

Moreover, $r^\star$ satisfies the following:

$$\|\Phi r^\star - Q^\star\| \leq c\,\|\Pi Q^\star - Q^\star\|. \tag{44}$$

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*

$$\|\mathfrak{F}Q - Q'\| \leq \gamma\,\|Q - Q'\|, \tag{45}$$

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

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

$$\Pi\Lambda := \underset{\bar{\Lambda}\in\{\Phi r|r\in\mathbb{R}^p\}}{\arg\min} \left\|\bar{\Lambda} - \Lambda\right\|.$$

Now, we first note that in the proof of Lemma 3, we deduced that for any $\Lambda \in L_2$ we have that

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

(c.f. Lemma 3).

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

$$\max\left\{\left\|\mathcal{M}Q - \hat{Q}\right\|, \left\|\mathcal{M}Q - \mathcal{M}\hat{Q}\right\|\right\} \leq \gamma \left\|Q - \hat{Q}\right\|, \tag{46}$$

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

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

using (46) 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) $\left\|\Pi\mathfrak{F}Q - \Pi\mathfrak{F}\bar{Q}\right\| \leq \gamma \left\|Q - \bar{Q}\right\|,$

ii) $\|\Phi r^\star - Q^\star\| \leq \frac{1}{\sqrt{1-\gamma^2}} \|\Pi Q^\star - Q^\star\|.$

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

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

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

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

after which we readily deduce the desired result. □

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

and $\tilde{\mathfrak{F}}$ by: $\tilde{\mathfrak{F}}Q := \Theta + \gamma PHQ$.

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

$$\left\|\tilde{\mathfrak{F}}Q - \tilde{\mathfrak{F}}\bar{Q}\right\| \leq \gamma \left\|Q - \bar{Q}\right\| \tag{47}$$

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

Proof: Using (46), we now observe that

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

again using the non-expansive property of P. □

Lemma 7 *Define by* $\tilde{Q} := \Theta + \gamma P v^{\tilde{\boldsymbol{\pi}}}$ *where*

$$v^{\tilde{\boldsymbol{\pi}}}(z) := \Theta(s_{\tau_k}, a) + \gamma \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} P(s'; a, s_{\tau_k}) \Phi r^\star(s', I(\tau_k)), \tag{48}$$

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

Proof: We begin by observing that

$$\begin{aligned}H\tilde{Q}(z) &= H\left(\Theta(z) + \gamma P v^{\tilde{\boldsymbol{\pi}}}\right) \\ &= \begin{cases}\mathcal{M}Q(z), & \text{if } \mathcal{M}Q(z) > \Phi r^\star, \\ Q(z), & \text{otherwise,}\end{cases} \\ &= \begin{cases}\mathcal{M}Q(z), & \text{if } \mathcal{M}Q(z) > \Phi r^\star, \\ \Theta(z) + \gamma P v^{\tilde{\boldsymbol{\pi}}}, & \text{otherwise,}\end{cases} \\ &= v^{\tilde{\boldsymbol{\pi}}}(z).\end{aligned}$$

Hence,

$$\tilde{\mathfrak{F}}\tilde{Q} = \Theta + \gamma PH\tilde{Q} = \Theta + \gamma P v^{\tilde{\boldsymbol{\pi}}} = \tilde{Q}. \tag{49}$$

which proves the result. □

Lemma 8 *The following bound holds:*

$$\mathbb{E}\left[v^{\hat{\boldsymbol{\pi}}}(z_0)\right] - \mathbb{E}\left[v^{\tilde{\boldsymbol{\pi}}}(z_0)\right] \leq 2\left[(1-\gamma)\sqrt{(1-\gamma^2)}\right]^{-1} \left\|\Pi Q^\star - Q^\star\right\|. \tag{50}$$

Proof: By definitions of $v^{\hat{\boldsymbol{\pi}}}$ and $v^{\tilde{\boldsymbol{\pi}}}$ (c.f (48)) and using Jensen's inequality and the stationarity property we have that,

$$\begin{aligned}\mathbb{E}\left[v^{\hat{\boldsymbol{\pi}}}(z_0)\right] - \mathbb{E}\left[v^{\tilde{\boldsymbol{\pi}}}(z_0)\right] &= \mathbb{E}\left[P v^{\hat{\boldsymbol{\pi}}}(z_0)\right] - \mathbb{E}\left[P v^{\tilde{\boldsymbol{\pi}}}(z_0)\right] \\ &\leq \left|\mathbb{E}\left[P v^{\hat{\boldsymbol{\pi}}}(z_0)\right] - \mathbb{E}\left[P v^{\tilde{\boldsymbol{\pi}}}(z_0)\right]\right| \\ &\leq \left\|P v^{\hat{\boldsymbol{\pi}}} - P v^{\tilde{\boldsymbol{\pi}}}\right\|. \end{aligned} \tag{51}$$

Now recall that $\tilde{Q} := \Theta + \gamma P v^{\tilde{\boldsymbol{\pi}}}$ and $Q^\star := \Theta + \gamma P v^{\boldsymbol{\pi}^\star}$, using these expressions in (51) we find that

$$\mathbb{E}\left[v^{\hat{\boldsymbol{\pi}}}(z_0)\right] - \mathbb{E}\left[v^{\tilde{\boldsymbol{\pi}}}(z_0)\right] \leq \frac{1}{\gamma}\left\|\tilde{Q} - Q^\star\right\|.$$

Moreover, by the triangle inequality and using the fact that $\mathfrak{F}(\Phi r^\star) = \tilde{\mathfrak{F}}(\Phi r^\star)$ and that $\mathfrak{F}Q^\star = Q^\star$ and $\mathfrak{F}\tilde{Q} = \tilde{Q}$ (c.f. (50)) we have that

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

which gives the following bound:

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

from which, using Lemma 5, we deduce that $\left\|\tilde{Q} - Q^\star\right\| \leq 2\left[(1-\gamma)\sqrt{(1-\gamma^2)}\right]^{-1}\left\|\tilde{Q} - \Phi r^\star\right\|$, after which by (52), we finally obtain

$$\mathbb{E}\left[v^{\hat{\boldsymbol{\pi}}}(z_0)\right] - \mathbb{E}\left[v^{\tilde{\boldsymbol{\pi}}}(z_0)\right] \leq 2\left[(1-\gamma)\sqrt{(1-\gamma^2)}\right]^{-1}\left\|\tilde{Q} - \Phi r^\star\right\|,$$

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 $\Xi : \mathbb{R}^{2d} \times \mathbb{R}^p \to \mathbb{R}^p$ is given by:

$$\Xi(w, r) := \phi(z)\left(\Theta(z) + \gamma \max\left\{(\Phi r)(z'), \mathcal{M}(\Phi r)(z')\right\} - (\Phi r)(z)\right),$$

for any $w \equiv (z, z') \in (\mathbb{N} \times \mathcal{S})^2$ where $z = (t, s) \in \mathbb{N} \times \mathcal{S}$ and $z' = (t, s') \in \mathbb{N} \times \mathcal{S}$ and for any $r \in \mathbb{R}^p$. Let us also define the function $\boldsymbol{\Xi} : \mathbb{R}^p \to \mathbb{R}^p$ by the following:

$$\boldsymbol{\Xi}(r) := \mathbb{E}_{w_0 \sim (\mathbb{P}, \mathbb{P})}\left[\Xi(w_0, r)\right]; w_0 := (z_0, z_1).$$

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

i) $(r - r^\star)\boldsymbol{\Xi}_k(r) < 0, \qquad \forall r \neq r^\star,$

ii) $\boldsymbol{\Xi}_k(r^\star) = 0.$

Proof: To prove the statement, we first note that each component of $\boldsymbol{\Xi}_k(r)$ admits a representation as an inner product, indeed:

$$\begin{aligned}
\boldsymbol{\Xi}_k(r) &= \mathbb{E}\left[\phi_k(z_0)(\Theta(z_0) + \gamma \max\left\{\Phi r(z_1), \mathcal{M}\Phi(z_1)\right\} - (\Phi r)(z_0)\right] \\
&= \mathbb{E}\left[\phi_k(z_0)(\Theta(z_0) + \gamma \mathbb{E}\left[\max\left\{\Phi r(z_1), \mathcal{M}\Phi(z_1)\right\}|z_0\right] - (\Phi r)(z_0)\right] \\
&= \mathbb{E}\left[\phi_k(z_0)(\Theta(z_0) + \gamma P \max\left\{(\Phi r, \mathcal{M}\Phi)\right\}(z_0) - (\Phi r)(z_0)\right] \\
&= \left\langle \phi_k, \mathfrak{F}\Phi r - \Phi r\right\rangle,
\end{aligned}$$

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:

$$\begin{aligned}
\left(r - r^\star\right)\boldsymbol{\Xi}_k(r) &= \sum_{l=1}\left(r(l) - r^\star(l)\right)\left\langle \phi_l, \mathfrak{F}\Phi r - \Phi r\right\rangle \\
&= \left\langle \Phi r - \Phi r^\star, \mathfrak{F}\Phi r - \Phi r\right\rangle \\
&= \left\langle \Phi r - \Phi r^\star, (\mathbf{1} - \Pi)\mathfrak{F}\Phi r + \Pi\mathfrak{F}\Phi r - \Phi r\right\rangle \\
&= \left\langle \Phi r - \Phi r^\star, \Pi\mathfrak{F}\Phi r - \Phi r\right\rangle,
\end{aligned}$$

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

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

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

$$\boldsymbol{\Xi}_k(r^\star) = \left\langle \phi_l, \mathfrak{F}\Phi r^\star - \Phi r\right\rangle = \left\langle \phi_l, \Pi\mathfrak{F}\Phi r^\star - \Phi r\right\rangle = 0,$$

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

$$r_{t+1} = r_t + \alpha \Xi(s_t, r_t), \tag{52}$$

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

1. $\{s_t | t = 0, 1, \ldots\}$ *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* $\mathbb{E}\left[1 + \|s_t\|^q | s \equiv s_0\right] \leq \mu_q \left(1 + \|s\|^q\right)$
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* $\|\Xi(w, r)\| \leq c\left(1 + \|w\|^q\right)\left(1 + \|r\|\right)$
5. *There exists scalars* c *and* q *such that* $\sum_{t=0}^{\infty} \left\|\mathbb{E}\left[\Xi(w_t, r) | z_0 \equiv z\right] - \mathbb{E}\left[\Xi(w_0, r)\right]\right\| \leq c\left(1 + \|w\|^q\right)\left(1 + \|r\|\right)$
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* $\sum_{t=0}^{\infty} \left\|\mathbb{E}\left[\Xi(w_t, r) | w_0 \equiv w\right] - \mathbb{E}\left[\Xi(w_0, \bar{r})\right]\right\| \leq c\|r - \bar{r}\|\left(1 + \|w\|^q\right)$
8. *There exists some* $r^\star \in \mathbb{R}^k$ *such that* $\boldsymbol{\Xi}(r)(r - r^\star) < 0$ *for all* $r \neq r^\star$ *and* $\bar{s}(r^\star) = 0$.

Then r_t *converges to* $r^\star$ *almost surely.*

In order to apply the Theorem 5, we show that conditions 1 - 7 are satisfied.

Proof: 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

$$\begin{aligned}
\|\Xi(w, r)\| &= \|\phi(z)\left(\Theta(z) + \gamma \max\left\{(\Phi r)(z'), \mathcal{M}\Phi(z')\right\} - (\Phi r)(z)\right)\| \\
&\leq \|\phi(z)\| \|\Theta(z) + \gamma\left(\|\phi(z')\| \|r\| + \mathcal{M}\Phi(z')\right)\| + \|\phi(z)\| \|r\| \\
&\leq \|\phi(z)\|\left(\|\Theta(z)\| + \gamma\|\mathcal{M}\Phi(z')\|\right) + \|\phi(z)\|\left(\gamma\|\phi(z')\| + \|\phi(z)\|\right)\|r\|.
\end{aligned}$$

Now using the definition of $\mathcal{M}$, we readily observe that $\|\mathcal{M}\Phi(z')\| \leq \|\Theta\| + \gamma\|\mathcal{P}^{\pi}_{s's_t}\Phi\| \leq \|\Theta\| + \gamma\|\Phi\|$ using the non-expansiveness of P.

Hence, we lastly deduce that

$$\begin{aligned}
\|\Xi(w, r)\| &\leq \|\phi(z)\|\left(\|\Theta(z)\| + \gamma\|\mathcal{M}\Phi(z')\|\right) + \|\phi(z)\|\left(\gamma\|\phi(z')\| + \|\phi(z)\|\right)\|r\| \\
&\leq \|\phi(z)\|\left(\|\Theta(z)\| + \gamma\|\Theta\| + \gamma\|\psi\|\right) + \|\phi(z)\|\left(\gamma\|\phi(z')\| + \|\phi(z)\|\right)\|r\|,
\end{aligned}$$

we then easily deduce the result using the boundedness of ϕ, Θ and ψ.

Now we observe the following Lipschitz condition on Ξ:

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

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

$$\sum_{t=0}^{\infty} \|\mathbb{E}\left[\Xi(w,r) - \Xi(w,\bar{r})|w_0 = w\right] - \mathbb{E}\left[\Xi(w_0,r) - \Xi(w_0,\bar{r})\|\right] \leq c\,\|r-\bar{r}\|\,(1+\|w\|^l). \quad (53)$$

Part 2 is assured by Lemma 5 while Part 4 is assured by Lemma 8 and lastly Part 8 is assured by Lemma 9. □