A Missing statements and proofs

A.1 Statements for Section 3.1

Claim A.1. Let a two-player Markov game where both players affect the transition. Further, consider a correlated policy $\boldsymbol{\sigma}$ and its corresponding marginalized product policy $\boldsymbol{\pi}^{\boldsymbol{\sigma}} = \boldsymbol{\pi}_1^{\boldsymbol{\sigma}} \times \boldsymbol{\pi}_2^{\boldsymbol{\sigma}}$. Then, for any $\boldsymbol{\pi}'_1, \boldsymbol{\pi}'_2$,

$$V_{k,1}^{\boldsymbol{\pi}'_1, \boldsymbol{\sigma}_{-1}}(s_1) = V_{k,1}^{\boldsymbol{\pi}'_1, \boldsymbol{\pi}_2^{\boldsymbol{\sigma}}}(s_1),$$
$$V_{k,2}^{\boldsymbol{\sigma}_{-2}, \boldsymbol{\pi}'_2}(s_1) = V_{k,2}^{\boldsymbol{\pi}_1^{\boldsymbol{\sigma}}, \boldsymbol{\pi}'_2}(s_1).$$

Proof. We will effectively show that the problem of best-responding to a correlated policy $\boldsymbol{\sigma}$ is equivalent to best-responding to the marginal policy of $\boldsymbol{\sigma}$ for the opponent. The proof follows from the equivalence of the two MDPs.

As a reminder,

$$\pi_{1,h}(a|s) = \sum_{b \in \mathcal{A}_2} \boldsymbol{\sigma}_h(a, b|s)$$
$$\pi_{2,h}(b|s) = \sum_{a \in \mathcal{A}_1} \boldsymbol{\sigma}_h(a, b|s)$$

As we have seen in Section 2.1, in the case of unilateral deviation from joint policy $\boldsymbol{\sigma}$, an agent faces a single agent MDP. More specifically, agent 2, best-responds by optimizing a reward function $\bar{r}_{2,h}(s, b)$ under a transition kernel $\bar{\mathbb{P}}_2$ for which,

$$\bar{r}_{2,h}(s, b) = \mathbb{E}_{b \sim \boldsymbol{\sigma}}\left[r_{2,h}(s, a, b)\right] = \mathbb{E}_{b \sim \boldsymbol{\pi}_1^{\boldsymbol{\sigma}}}\left[r_{2,h}(s, a, b)\right] = r_{2,h}(s, \boldsymbol{\pi}_1^{\boldsymbol{\sigma}}, b).$$

Similarly,

$$\bar{r}_{1,h}(s, b) = r_{1,h}(s, a, \boldsymbol{\pi}_2^{\boldsymbol{\sigma}}).$$

Analogously, for each of the transition kernels,

$$\bar{\mathbb{P}}_{2,h}(s'|s, b) = \mathbb{E}_{a \sim \boldsymbol{\sigma}}\left[\mathbb{P}_{2,h}(s'|s, a, b)\right] = \mathbb{E}_{a \sim \boldsymbol{\pi}_2^{\boldsymbol{\sigma}}}\left[\mathbb{P}_{2,h}(s'|s, a, b)\right] = \mathbb{P}_{2,h}(s'|s, \boldsymbol{\pi}_1^{\boldsymbol{\sigma}}, b),$$

as for agent 1,

$$\bar{\mathbb{P}}_{1,h}(s'|s, a) = \mathbb{P}_{1,h}(s'|s, a, \boldsymbol{\pi}_2^{\boldsymbol{\sigma}}).$$

Hence, it follows that, $V_{2,1}^{\boldsymbol{\sigma}_{-2} \times \boldsymbol{\pi}'_2}(s_1) = V_{2,1}^{\boldsymbol{\pi}_1^{\boldsymbol{\sigma}} \times \boldsymbol{\pi}'_2}(s_1), \ \forall \boldsymbol{\pi}'_2$ and $V_{1,1}^{\boldsymbol{\pi}'_1 \times \boldsymbol{\sigma}_{-1}}(s_1) = V_{1,1}^{\boldsymbol{\pi}'_1 \times \boldsymbol{\pi}_2^{\boldsymbol{\sigma}}}(s_1), \ \forall \boldsymbol{\pi}'_2$.

□

Before that, given a (possibly correlated) joint policy $\boldsymbol{\sigma}$ we define a nonlinear program, (P_{BR}), whose optimal solutions are best-response policies of each agent k to $\boldsymbol{\sigma}_{-k}$ and the values for each state s and timestep h:

A.2 Proof of Theorem 3.2

The best-response program. First, we state the following lemma that will prove useful for several of our arguments,

Lemma A.1 (Best-response LP)**.** Let a (possibly correlated) joint policy $\hat{\boldsymbol{\sigma}}$. Consider the following linear program with variables $\boldsymbol{w} \in \mathbb{R}^{n \times H \times S}$,

$$(\mathrm{P}_{\mathrm{BR}}) \qquad \begin{aligned} \min \quad & \sum_{k\in[n]} w_{k,s}(s_1) - \boldsymbol{e}_{s_1}^\top \sum_{h=1}^{H} \left(\prod_{\tau=1}^{h} \mathbb{P}_\tau(\hat{\boldsymbol{\sigma}}_\tau) \right) \boldsymbol{r}_{k,h}(\hat{\boldsymbol{\sigma}}_h) \\ \text{s.t.} \quad & w_{k,h}(s) \geq r_{k,h}(s,a,\hat{\boldsymbol{\sigma}}_{-k,h}) + \mathbb{P}_h(s,a,\hat{\boldsymbol{\sigma}}_{-k,h})\boldsymbol{w}_{k,h+1}, \\ & \qquad \forall s\in\mathcal{S}, \forall h\in[H], \forall k\in[n], \forall a \in \mathcal{A}_k; \\ & w_{k,H}(s) = 0, \; \forall k\in[n], \forall s\in\mathcal{S}. \end{aligned}$$

The optimal solution $\boldsymbol{w}^\dagger$ of the program is unique and corresponds to the value function of each player $k\in[n]$ when player k best-responds to $\hat{\boldsymbol{\sigma}}$.

Proof. We observe that the program is separable to n independent linear programs, each with variables $\boldsymbol{w}_k \in \mathbb{R}^{n\times H}$,

$$\begin{aligned} & \min \; w_{k,1}(s_1) \\ & \text{s.t. } w_{k,h}(s) \geq r_{k,h}(s,a,\hat{\boldsymbol{\sigma}}_{-k,h}) + \mathbb{P}_h(s,a,\hat{\boldsymbol{\sigma}}_{-k,h})\boldsymbol{w}_{k,h+1}, \\ & \qquad \forall s\in\mathcal{S}, \forall h\in[H], \forall a\in\mathcal{A}_k; \\ & \quad w_{k,H}(s) = 0, \; \forall k\in[n], \forall s\in\mathcal{S}. \end{aligned}$$

Each of these linear programs describes the problem of a single agent MDP (Neu and Pike-Burke, 2020, Section 2) —that agent being k— which, as we have seen in Best-response policies, is equivalent to the problem of finding a best-response to $\hat{\boldsymbol{\sigma}}_{-k}$. It follows that the optimal $\boldsymbol{w}_k^\dagger$ for every program is unique (each program corresponds to a set of Bellman optimality equations). □

Properties of the NE program. Second, we need to prove that the minimum value of the objective function of the program is nonnegative.

Lemma A.2 (Feasibility of ($\mathrm{P}'_{\mathrm{NE}}$) and global optimum)**.** The nonlinear program ($\mathrm{P}'_{\mathrm{NE}}$) is feasible, has a nonnegative objective value, and its global minimum is equal to 0.

Proof. Analogously to the finite-horizon case, for the feasibility of the nonlinear program, we invoke the theorem of the existence of a Nash equilibrium. We let a NE product policy, $\boldsymbol{\pi}^\star$, and a vector $\boldsymbol{w}^\star \in \mathbb{R}^{n\times S}$ such that $w_k^\star(s) = V_k^{\dagger,\boldsymbol{\pi}^\star_{-k}}(s), \; \forall k\in[n]\times\mathcal{S}$.

By Lemma A.1, we know that $(\boldsymbol{\pi}^\star, \boldsymbol{w}^\star)$ satisfies all the constraints of (P_{NE}). Additionaly, because $\boldsymbol{\pi}^\star$ is a NE, $V_{k,h}^{\boldsymbol{\pi}^\star}(s_1) = V_{k,h}^{\dagger,\boldsymbol{\pi}^\star_{-k}}(s_1)$ for all $k\in[n]$. Observing that,

$$w_{k,1}^\star(s_1) - \boldsymbol{e}_{s_1}^\top \sum_{h=1}^{H} \left(\prod_{\tau=1}^{h} \mathbb{P}_\tau(\boldsymbol{\pi}_\tau^\star) \right) \boldsymbol{r}_{k,h}(\boldsymbol{\pi}_h^\star) = V_{k,h}^{\dagger,\boldsymbol{\pi}^\star_{-k}}(s_1) - V_{k,h}^{\boldsymbol{\pi}^\star}(s_1) = 0,$$

concludes the argument that a NE attains an objective value equal to 0.

Continuing, we observe that due to (1) the objective function can be equivalently rewritten as,

$$\begin{aligned} & \sum_{k\in[n]} \left(w_{k,1}(s_1) - \boldsymbol{e}_{s_1}^\top \sum_{h=1}^{H} \left(\prod_{\tau=1}^{h} \mathbb{P}_\tau(\boldsymbol{\pi}_\tau) \right) \boldsymbol{r}_{k,h}(\boldsymbol{\pi}_h) \right) \\ & \quad = \sum_{k\in[n]} w_{k,1}(s_1) - \boldsymbol{e}_{s_1}^\top \sum_{h=1}^{H} \left(\prod_{\tau=1}^{h} \mathbb{P}_\tau(\boldsymbol{\pi}_\tau) \right) \sum_{k\in[n]} \boldsymbol{r}_{k,h}(\boldsymbol{\pi}_h) \\ & \quad = \sum_{k\in[n]} w_{k,1}(s_1). \end{aligned}$$

Next, we focus on the inequality constraint

$$w_{k,h}(s) \geq r_{k,h}(s,a,\boldsymbol{\pi}_{-k,h}) + \mathbb{P}_h(s,a,\boldsymbol{\pi}_{-k,h})\boldsymbol{w}_{k,h+1}$$

which holds for all $s\in\mathcal{S}$, all players $k\in[n]$, all $a\in\mathcal{A}_k$, and all timesteps $h\in[H-1]$.

By summing over $a \in \mathcal{A}_k$ while multiplying each term with a corresponding coefficient $\pi_{k,h}(a|s)$, the display written in an equivalent element-wise vector inequality reads:

$$\boldsymbol{w}_{k,h} \geq \boldsymbol{r}_{k,h}(\boldsymbol{\pi}_h) + \mathbb{P}_h(\boldsymbol{\pi}_h)\boldsymbol{w}_{k,h+1}.$$

Finally, after consecutively substituting $\boldsymbol{w}_{k,h+1}$ with the element-wise lesser term $\boldsymbol{r}_{k,h+1}(\boldsymbol{\pi}_{h+1}) + \mathbb{P}_{h+1}(\boldsymbol{\pi}_{h+1})\boldsymbol{w}_{k,h+2}$, we end up with the inequality:

$$\boldsymbol{w}_{k,1} \geq \sum_{h=1}^{H} \left(\prod_{\tau=1}^{h} \mathbb{P}_\tau(\boldsymbol{\pi}_\tau) \right) \boldsymbol{r}_{k,h}(\boldsymbol{\pi}_h). \tag{5}$$

Summing over k, it holds for the s_1-th entry of the inequality,

$$\sum_{k\in[n]} w_{k,1} \geq \sum_{k\in[n]} \sum_{h=1}^{H} \left(\prod_{\tau=1}^{h} \mathbb{P}_\tau(\boldsymbol{\pi}_\tau) \right) \boldsymbol{r}_{k,h}(\boldsymbol{\pi}_h) = 0.$$

Where the equality holds due to the zero-sum property, (1). $\square$

An approximate NE is an approximate global minimum. We show that an ϵ-approximate NE, $\boldsymbol{\pi}^\star$, achieves an $n\epsilon$-approximate global minimum of the program. Utilizing Lemma A.1, setting $w_k^\star(s_1) = V_{k,1}^{\dagger,\boldsymbol{\pi}^\star_{-k}}(s_1)$, and the definition of an ϵ-approximate NE we see that,

$$\begin{aligned}
\sum_{k\in[n]} \left(w_{k,1}^\star(s_1) - \boldsymbol{e}_{s_1}^\top \sum_{h=1}^{H} \left(\prod_{\tau=1}^{h} \mathbb{P}_\tau(\boldsymbol{\pi}_\tau^\star) \right) \boldsymbol{r}_{k,h}(\boldsymbol{\pi}_h^\star) \right) &= \sum_{k\in[n]} \left(w_{k,1}^\star(s_1) - V_{k,1}^{\boldsymbol{\pi}^\star}(s_1) \right) \\
&\leq \sum_{k\in[n]} \epsilon = n\epsilon.
\end{aligned}$$

Indeed, this means that $\boldsymbol{\pi}^\star, \boldsymbol{w}^\star$ is an $n\epsilon$-approximate global minimizer of (P_{NE}).

An approximate global minimum is an approximate NE. For the opposite direction, we let a feasible ϵ-approximate global minimizer of the program (P_{NE}), $(\boldsymbol{\pi}^\star, \boldsymbol{w}^\star)$. Because a global minimum of the program is equal to 0, an ϵ-approximate global optimum must be at most $\epsilon > 0$. We observe that for every $k \in [n]$,

$$w_{k,1}^\star(s_1) \geq \boldsymbol{e}_{s_1}^\top \sum_{h=1}^{H} \left(\prod_{\tau=1}^{h} \mathbb{P}_\tau(\boldsymbol{\pi}_\tau^\star) \right) \boldsymbol{r}_{k,h}(\boldsymbol{\pi}_h^\star), \tag{6}$$

which follows from induction on the inequality constraint over all h similar to (5).

Consequently, the assumption that

$$\epsilon \geq \sum_{k\in[n]} \left(w_{k,1}^\star(s_1) - \boldsymbol{e}_{s_1}^\top \sum_{h=1}^{H} \left(\prod_{\tau=1}^{h} \mathbb{P}_\tau(\boldsymbol{\pi}_\tau^\star) \right) \boldsymbol{r}_{k,h}(\boldsymbol{\pi}_h^\star) \right),$$

and Equation (6), yields the fact that

$$\begin{aligned}
\epsilon &\geq w_{k,1}^\star(s_1) - \boldsymbol{e}_{s_1}^\top \sum_{h=1}^{H} \left(\prod_{\tau=1}^{h} \mathbb{P}_\tau(\boldsymbol{\pi}_\tau^\star) \right) \boldsymbol{r}_{k,h}(\boldsymbol{\pi}_h^\star) \\
&\geq V_{k,1}^{\dagger,\boldsymbol{\pi}^\star_{-k}}(s_1) - V_{k,1}^{\boldsymbol{\pi}^\star}(s_1),
\end{aligned}$$

where the second inequality holds from the fact that $\boldsymbol{w}^\star$ is feasible for (P_{BR}). The latter concludes the proof, as the display coincides with the definition of an ϵ-approximate NE.

A.3 Proof of Claim 3.1

Proof. The value function of s_1 for $h = 1$ of players 1 and 2 read:

$$\begin{aligned}V_{1,1}^{\boldsymbol{\sigma}}(s_1) &= \boldsymbol{e}_{s_1}^\top \left(\boldsymbol{r}_1(\boldsymbol{\sigma}) + \mathbb{P}(\boldsymbol{\sigma})\boldsymbol{r}_1(\boldsymbol{\sigma})\right) \\ &= -\frac{9\sigma(a_1,b_1|s_1)}{20} + \frac{\sigma(a_1,b_2|s_1)}{20} + \frac{\left(1-\sigma(a_1,b_1|s_1)\right)\left(\sigma(a_1,b_1|s_2)+\sigma(a_1,b_2|s_2)\right)}{20},\end{aligned}$$

and,

$$\begin{aligned}V_{2,1}^{\boldsymbol{\sigma}}(s_1) &= \boldsymbol{e}_{s_1}^\top \left(\boldsymbol{r}_2(\boldsymbol{\sigma}) + \mathbb{P}(\boldsymbol{\sigma})\boldsymbol{r}_2(\boldsymbol{\sigma})\right) \\ &= -\frac{9\sigma(a_1,b_1|s_1)}{20} + \frac{\sigma(a_2,b_2|s_1)}{20} + \frac{\left(1-\sigma(a_1,b_1|s_1)\right)\left(\sigma(a_1,b_1|s_2)+\sigma(a_2,b_1|s_2)\right)}{20}.\end{aligned}$$

We are indifferent to the corresponding value function of player 3 as they only have one available action per state and hence, cannot affect their rewards. For the joint policy $\boldsymbol{\sigma}$, the corresponding value functions of both players 1 and 2 are $V_{1,1}^{\boldsymbol{\sigma}}(s_1) = V_{2,1}^{\boldsymbol{\sigma}}(s_1) = \frac{1}{20}$.

Deviations. We will now prove that no deviation of player 1 manages to accumulate a reward greater than $\frac{1}{20}$. The same follows for player 2 due to symmetry.

When a player deviates unilaterally from a joint policy, they experience a single agent Markov decision process (MDP). It is well-known that MDPs always have a deterministic optimal policy. As such, it suffices to check whether $V_{1,1}^{\boldsymbol{\pi}_1,\boldsymbol{\sigma}_{-1}}(s_1)$ is greater than $\frac{1}{20}$ for any of the four possible deterministic policies:

- $\boldsymbol{\pi}_1(s_1) = \boldsymbol{\pi}_1(s_2) = (1 \quad 0)$,
- $\boldsymbol{\pi}_1(s_1) = \boldsymbol{\pi}_1(s_2) = (0 \quad 1)$,
- $\boldsymbol{\pi}_1(s_1) = (1 \quad 0)\,,\ \boldsymbol{\pi}_1(s_2) = (0 \quad 1)$,
- $\boldsymbol{\pi}_1(s_1) = (0 \quad 1)\,,\ \boldsymbol{\pi}_1(s_2) = (1 \quad 0)$.

Finally, the value function of any deviation $\boldsymbol{\pi}_1'$ writes,

$$V_{1,1}^{\boldsymbol{\pi}_1'\times\boldsymbol{\sigma}_{-1}}(s_1) = -\frac{\pi_1'(a_1|s_1)}{5} - \frac{\pi_1'(a_1|s_2)\left(\pi_1'(a_1|s_1)-2\right)}{40}.$$

We can now check that for all deterministic policies $V_{1,1}^{\boldsymbol{\pi}_1'\times\boldsymbol{\sigma}_{-1}}(s_1) \le \frac{1}{20}$. By symmetry, it follows that $V_{2,1}^{\boldsymbol{\pi}_2'\times\boldsymbol{\sigma}_{-2}}(s_1) \le \frac{1}{20}$ and as such $\boldsymbol{\sigma}$ is indeed a CCE. $\square$

A.4 Proof of Claim 3.2

Proof. In general, the value functions of each player 1 and 2 are:

$$V_{1,1}^{\boldsymbol{\pi}_1\times\boldsymbol{\pi}_2}(s_1) = -\frac{\pi_1(a_1|s_1)\pi_2(b_1|s_1)}{2} + \frac{\pi_1(a_1|s_1)}{20} - \frac{\pi_1(a_1|s_2)\left(\pi_1(a_1|s_1)\pi_2(b_1|s_1)-1\right)}{20},$$

and

$$V_{2,1}^{\boldsymbol{\pi}_1\times\boldsymbol{\pi}_2}(s_1) = -\frac{\pi_1(a_1|s_1)\pi_2(b_1|s_1)}{2} + \frac{\pi_1(b_1|s_1)}{20} - \frac{\pi_1(b_1|s_2)\left(\pi_1(a_1|s_1)\pi_2(b_1|s_1)-1\right)}{20}.$$

Plugging in $\boldsymbol{\pi}_1^\sigma, \boldsymbol{\pi}_2^\sigma$ yields $V_{1,1}^{\boldsymbol{\pi}_1^\sigma\times\boldsymbol{\pi}_2^\sigma}(s_1) = V_{2,1}^{\boldsymbol{\pi}_1^\sigma\times\boldsymbol{\pi}_2^\sigma}(s_1) = -\frac{13}{160}$. But, if player 1 deviates to say $\pi_1'(s_1) = \pi_1'(s_2) = (0 \quad 1)$, they get a value equal to 0 which is clearly greater than $-\frac{13}{160}$. Hence, $\boldsymbol{\pi}_1^\sigma \times \boldsymbol{\pi}_2^\sigma$ is not a NE. $\square$

A.5 Proof of Theorem 3.4

Proof. The proof follows from the game of Example 1, and Claims 3.1 and 3.2. $\square$

B Proofs for infinite-horizon Zero-Sum Polymatrix Markov Games

In this section we will explicitly state definitions, theorems and proofs relating to the infinite-horizon discounted zero-sum polymatrix Markov games.

B.1 Definitions of equilibria for the infinite-horizon

Let us restate the definition specifically for infinite-horizon Markov games. They are defined as a tuple $\Gamma(H, \mathcal{S}, \{\mathcal{A}_k\}_{k\in[n]}, \mathbb{P}, \{r_k\}_{k\in[n]}, \gamma, \boldsymbol{\rho})$.

- $H = \infty$ denotes the *time horizon*
- $\mathcal{S}$, with cardinality $S := |\mathcal{S}|$, stands for the state space,
- $\{\mathcal{A}_k\}_{k\in[n]}$ is the collection of every player's action space, while $\mathcal{A} := \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$ denotes the *joint action space*; further, an element of that set —a joint action— is generally noted as $\boldsymbol{a} = (a_1, \ldots, a_n) \in \mathcal{A}$,
- $\mathbb{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition probability function,
- $r_k : \mathcal{S}, \mathcal{A} \to [-1, 1]$ yields the reward of player k at a given state and joint action,
- a discount factor $0 < \gamma < 1$,
- an initial state distribution $\boldsymbol{\rho} \in \Delta(\mathcal{S})$.

Policies and value functions. In infinite-horizon Markov games policies can still be distinguished in two main ways, *Markovian/non-Markovian* and *stationary/nonstationary*. Moreover, a joint policy can be a *correlated* policy or a *product* policy.

Markovian policies attribute a probability over the simplex of actions solely depending on the running state s of the game. On the other hand, *non-Markovian* policies attribute a probability over the simplex of actions that depends on any subset of the history of the game. *I.e.*, they can depend on any sub-sequence of actions and states up until the running timestep of the horizon.

Stationary policies are those that will attribute the same probability distribution over the simplex of actions for every timestep of the horizon. *Nonstationary* policies, on the contrary can change depending on the timestep of the horizon.

A joint Markovian stationary policy $\boldsymbol{\sigma}$ is said to be *correlated* when for every state $s \in \mathcal{S}$, attributes a probability distribution over the simplex of joint actions $\mathcal{A}$ for all players, *i.e.*, $\boldsymbol{\sigma}(s) \in \Delta(\mathcal{A})$. A Markovian stationary policy $\boldsymbol{\pi}$ is said to be a *product* policy when for every $s \in \mathcal{S}$, $\boldsymbol{\pi}(s) \in \prod_{k=1}^{n} \Delta(\mathcal{A}_k)$. It is rather easy to define *correlated/product* policies for the case of non-Markovian and nonstationary policies.

Given a Markovian stationary policy $\boldsymbol{\pi}$, the value function for an infinite-horizon discounted game is defined as,

$$V_k^{\boldsymbol{\pi}}(s_1) = \mathbb{E}_{\boldsymbol{\pi}}\left[\sum_{h=1}^{H} \gamma^{h-1} r_{k,h}(s_h, \boldsymbol{a}_h) \big| s_1\right] = \boldsymbol{e}_{s_1}^\top \sum_{h=1}^{H} \left(\gamma^{h-1} \prod_{\tau=1}^{h} \mathbb{P}_\tau(\boldsymbol{\pi}_\tau)\right) \boldsymbol{r}_{k,h}(\boldsymbol{\pi}_h).$$

It is possible to express the value function of each player k in the following way,

$$V_k^{\boldsymbol{\pi}}(s_1) = \boldsymbol{e}_{s_1}^\top \left(\mathbf{I} - \gamma\, \mathbb{P}(\boldsymbol{\pi})\right)^{-1} \boldsymbol{r}(\boldsymbol{\pi}).$$

Where $\mathbf{I}$ is the identity matrix of appropriate dimensions. Also, when the initial state is drawn from the initial state distribution, we denote, the value function reads $V_k^{\boldsymbol{\pi}}(\boldsymbol{\rho}) = \boldsymbol{\rho}^\top \left(\mathbf{I} - \gamma\, \mathbb{P}(\boldsymbol{\pi})\right)^{-1} \boldsymbol{r}(\boldsymbol{\pi})$.

Best-response policies. Given an arbitrary joint policy $\boldsymbol{\sigma}$ (which can be either a correlated or product policy), a best-response policy of a player k is defined to be $\boldsymbol{\pi}_k^\dagger \in \Delta(\mathcal{A}_k)^S$ such that $\boldsymbol{\pi}_k^\dagger \in \arg\max_{\boldsymbol{\pi}'_k} V_k^{\boldsymbol{\pi}'_k \times \boldsymbol{\sigma}_{-k}}(s)$. Also, we will denote $V_k^{\dagger, \boldsymbol{\sigma}_{-k}}(s) = \max_{\boldsymbol{\pi}'_k} V_k^{\boldsymbol{\pi}'_k, \boldsymbol{\sigma}_{-k}}(s)$. It is rather straightforward to see that the problem of computing a best-response to a given policy is equivalent to solving a single-agent MDP problem.

Notions of equilibria. Now that best-response policies have been defined, it is straightforward to define the different notions of equilibria. First, we define the notion of a coarse-correlated equilibrium.

Definition B.1 (CCE—infinite-horizon)**.** *A joint (potentially correlated) policy $\boldsymbol{\sigma} \in \Delta(\mathcal{A})^S$ is an ϵ-approximate coarse-correlated equilibrium if it holds that for an ϵ,*

$$V_k^{\dagger,\boldsymbol{\sigma}_{-k}}(\boldsymbol{\rho}) - V_k^{\boldsymbol{\sigma}}(\boldsymbol{\rho}) \leq \epsilon, \ \forall k \in [n].$$

Second, we define the notion of a Nash equilibrium. The main difference of the definition of the coarse-correlated equilibrium, is the fact that a NE Markovian stationary policy is a *product policy*.

Definition B.2 (NE—infinite-horizon)**.** *A joint (potentially correlated) policy $\boldsymbol{\pi} \in \prod_{k\in[n]} \Delta(\mathcal{A}_k)^S$ is an ϵ-approximate coarse-correlated equilibrium if it holds that for an ϵ,*

$$V_k^{\dagger,\boldsymbol{\pi}-k}(\boldsymbol{\rho}) - V_k^{\boldsymbol{\pi}}(\boldsymbol{\rho}) \leq \epsilon, \ \forall k \in [n].$$

As it is folklore by now, infinite-horizon discounted Markov games have a stationary Markovian Nash equilibrium.

C Main results for infinite-horizon games

The workhorse of our arguments in the following results is still the following nonlinear program with variables $\boldsymbol{\pi}, \boldsymbol{w}$,

$$(\mathrm{P}'_{\mathrm{NE}}) \qquad \begin{aligned} \min \ & \sum_{k\in[n]} \boldsymbol{\rho}^\top \left(\boldsymbol{w}_k - (\mathbf{I} - \gamma\,\mathbb{P}(\boldsymbol{\pi}))^{-1}\boldsymbol{r}_k(\boldsymbol{\pi})\right) \\ \text{s.t. } & w_k(s) \geq r_k(s,a,\boldsymbol{\pi}_{-k}) + \gamma\,\mathbb{P}(s,a,\boldsymbol{\pi}_{-k})\boldsymbol{w}_k, \\ & \qquad \forall s \in \mathcal{S}, \forall k \in [n], \forall a \in \mathcal{A}_k; \\ & \boldsymbol{\pi}_k(s) \in \Delta(\mathcal{A}_k), \\ & \qquad \forall s \in \mathcal{S}, \forall k \in [n], \forall a \in \mathcal{A}_k. \end{aligned}$$

As we will prove, approximate NE's correspond to approximate global minima of ($\mathrm{P}'_{\mathrm{NE}}$) and vice-versa. Before that, we need some intermediate lemmas. The first lemma we prove is about the best-response program.

The best-response program. Even for the infinite-horizon, we can define a linear program for the best-responses of all players. That program is the following, with variables $\boldsymbol{w}$,

$$(\mathrm{P}'_{\mathrm{BR}}) \qquad \begin{aligned} \min \ & \sum_{k\in[n]} \boldsymbol{\rho}^\top \left(\boldsymbol{w}_k - (\mathbf{I} - \gamma\,\mathbb{P}(\hat{\boldsymbol{\sigma}}))^{-1}\boldsymbol{r}_k(\hat{\boldsymbol{\sigma}})\right) \\ \text{s.t. } & w_k(s) \geq r_k(s,a,\hat{\boldsymbol{\sigma}}_{-k}) + \mathbb{P}(s,a,\hat{\boldsymbol{\sigma}}_{-k})\boldsymbol{w}_k, \\ & \qquad \forall s \in \mathcal{S}, \forall k \in [n], \forall a \in \mathcal{A}_k. \end{aligned}$$

Lemma C.1 (Best-response LP—infinite-horizon)**.** Let a (possibly correlated) joint policy $\hat{\boldsymbol{\sigma}}$. Consider the linear program ($\mathrm{P}'_{\mathrm{BR}}$). The optimal solution $\boldsymbol{w}^\dagger$ of the program is unique and corresponds to the value function of each player $k \in [n]$ when player k best-responds to $\hat{\boldsymbol{\sigma}}$.

Proof. We observe that the program is separable to n independent linear programs, each with variables $\boldsymbol{w}_k \in \mathbb{R}^n$,

$$\begin{aligned} \min \ & \boldsymbol{\rho}^\top \boldsymbol{w}_k \\ \text{s.t. } & w_k(s) \geq r_k(s,a,\hat{\boldsymbol{\sigma}}_{-k}) + \gamma\,\mathbb{P}(s,a,\hat{\boldsymbol{\sigma}}_{-k})\boldsymbol{w}_k, \\ & \qquad \forall s \in \mathcal{S}, \forall a \in \mathcal{A}_k. \end{aligned}$$

Each of these linear programs describes the problem of a single agent MDP —that agent being k. It follows that the optimal $\boldsymbol{w}_k^\dagger$ for every program is unique (each program corresponds to a set of Bellman optimality equations). □

Properties of the NE program. Second, we need to prove that the minimum value of the objective function of the program is nonnegative.

Lemma C.2 (Feasibility of ($\mathrm{P}'_{\mathrm{NE}}$) and global optimum)**.** The nonlinear program ($\mathrm{P}'_{\mathrm{NE}}$) is feasible, has a nonnegative objective value, and its global minimum is equal to 0.

Proof. For the feasibility of the nonlinear program, we invoke the theorem of the existence of a Nash equilibrium. *i.e.*, let a NE product policy, $\boldsymbol{\pi}^\star$, and a vector $\boldsymbol{w}^\star \in \mathbb{R}^{n\times H\times S}$ such that $w^\star_{k,s}(s) = V_k^{\dagger,\boldsymbol{\pi}^\star_{-k}}(s),\ \forall k \in [n]\times\mathcal{S}$.

By Lemma C.1, we know that $(\boldsymbol{\pi}^\star, \boldsymbol{w}^\star)$ satisfies all the constraints of ($\mathrm{P}'_{\mathrm{NE}}$). Additionally, because $\boldsymbol{\pi}^\star$ is a NE, $V_k^{\boldsymbol{\pi}^\star}(\boldsymbol{\rho}) = V_k^{\dagger,\boldsymbol{\pi}^\star_{-k}}(\boldsymbol{\rho})$ for all $k \in [n]$. Observing that,

$$\boldsymbol{\rho}^\top\left(\boldsymbol{w}^\star_k - (\mathbf{I} - \gamma\,\mathbb{P}(\boldsymbol{\pi}^\star))^{-1}\boldsymbol{r}_k(\boldsymbol{\pi}^\star)\right) = V_k^{\dagger,\boldsymbol{\pi}^\star_{-k}}(\boldsymbol{\rho}) - V_k^{\boldsymbol{\pi}^\star}(\boldsymbol{\rho}) = 0,$$

concludes the argument that a NE attains an objective value equal to 0.

Continuing, we observe that due to (1) the objective function can be equivalently rewritten as,

$$\begin{aligned}
&\sum_{k\in[n]}\left(\boldsymbol{\rho}^\top\boldsymbol{w}_k - \boldsymbol{\rho}^\top(\mathbf{I} - \gamma\,\mathbb{P}(\boldsymbol{\pi}))^{-1}\boldsymbol{r}_k(\boldsymbol{\pi})\right)\\
&\qquad = \sum_{k\in[n]}\boldsymbol{\rho}^\top\boldsymbol{w}_k - \boldsymbol{\rho}^\top(\mathbf{I} - \gamma\,\mathbb{P}(\boldsymbol{\pi}))^{-1}\sum_{k\in[n]}\boldsymbol{r}_k(\boldsymbol{\pi}_h)\\
&\qquad = \sum_{k\in[n]}\boldsymbol{\rho}^\top\boldsymbol{w}_k.
\end{aligned}$$

Next, we focus on the inequality constraint

$$w_k(s) \geq r_k(s, a, \boldsymbol{\pi}_{-k}) + \gamma\,\mathbb{P}(s, a, \boldsymbol{\pi}_{-k})\boldsymbol{w}_k$$

which holds for all $s \in \mathcal{S}$, all players $k \in [n]$, and all $a \in \mathcal{A}_k$.

By summing over $a \in \mathcal{A}_k$ while multiplying each term with a corresponding coefficient $\pi_k(a|s)$, the display written in an equivalent element-wise vector inequality reads:

$$\boldsymbol{w}_k \geq \boldsymbol{r}_{k,h}(\boldsymbol{\pi}) + \gamma\,\mathbb{P}(\boldsymbol{\pi})\boldsymbol{w}_k.$$

Finally, after consecutively substituting $\boldsymbol{w}_k$ with the element-wise lesser term $\boldsymbol{r}_k(\boldsymbol{\pi}) + \gamma\,\mathbb{P}_(\boldsymbol{\pi})\boldsymbol{w}_k$, we end up with the inequality:

$$\boldsymbol{w}_k \geq (\mathbf{I} - \gamma\,\mathbb{P}(\boldsymbol{\pi}))^{-1}\,\boldsymbol{r}_k(\boldsymbol{\pi}). \tag{9}$$

We note that $\mathbf{I} + \gamma\,\mathbb{P}(\boldsymbol{\pi}) + \gamma^2\,\mathbb{P}^2(\boldsymbol{\pi}) + \cdots = (\mathbf{I} - \gamma\,\mathbb{P}(\boldsymbol{\pi}))^{-1}$.

Summing over k, it holds for the s_1-th entry of the inequality,

$$\sum_{k\in[n]}\boldsymbol{w}_k \geq \sum_{k\in[n]}(\mathbf{I} - \gamma\,\mathbb{P}(\boldsymbol{\pi}))^{-1}\,\boldsymbol{r}_k(\boldsymbol{\pi}) = (\mathbf{I} - \gamma\,\mathbb{P}(\boldsymbol{\pi}))^{-1}\sum_{k\in[n]}\boldsymbol{r}_k(\boldsymbol{\pi}) = 0.$$

Where the equality holds due to the zero-sum property, (1). □

Theorem C.1 (NE and global optima of ($\mathrm{P}'_{\mathrm{NE}}$)—infinite-horizon)**.** *If $(\boldsymbol{\pi}^\star, \boldsymbol{w}^\star)$ yields an ϵ-approximate global minimum of ($\mathrm{P}'_{\mathrm{NE}}$), then $\boldsymbol{\pi}^\star$ is an $n\epsilon$-approximate NE of the infinite-horizon zero-sum polymatrix switching controller MG, Γ. Conversely, if $\boldsymbol{\pi}^\star$ is an ϵ-approximate NE of the MG Γ with corresponding value function vector $\boldsymbol{w}^\star$ such that $w^\star_k(s) = V_k^{\boldsymbol{\pi}^\star}(s)\forall(k, s) \in [n]\times\mathcal{S}$, then $(\boldsymbol{\pi}^\star, \boldsymbol{w}^\star)$ attains an ϵ-approximate global minimum of ($\mathrm{P}'_{\mathrm{NE}}$).*

Proof.

An approximate NE is an approximate global minimum. We show that an ϵ-approximate NE, $\boldsymbol{\pi}^\star$, achieves an $n\epsilon$-approximate global minimum of the program. Utilizing Lemma C.1 by setting $\boldsymbol{w}^\star_k = \mathbf{V}^{\dagger,\boldsymbol{\pi}^\star_{-k}}(\boldsymbol{\rho})$, feasibility , and the definition of an ϵ-approximate NE we see that,

$$\sum_{k\in[n]}\left(\boldsymbol{\rho}^\top \boldsymbol{w}_k^\star - \boldsymbol{\rho}^\top\left(\mathbf{I}-\gamma\,\mathbb{P}(\boldsymbol{\pi}^\star)\right)^{-1}\boldsymbol{r}_k(\boldsymbol{\pi}^\star)\right) = \sum_{k\in[n]}\left(\boldsymbol{\rho}^\top \boldsymbol{w}_k^\star - V_k^{\boldsymbol{\pi}^\star}(\boldsymbol{\rho})\right)$$
$$\leq \sum_{k\in[n]} \epsilon = n\epsilon.$$

Indeed, this means that $\boldsymbol{\pi}^\star, \boldsymbol{w}^\star$ is an $n\epsilon$-approximate global minimizer of ($\mathrm{P}'_{\mathrm{NE}}$).

An approximate global minimum is an approximate NE. For this direction, we let a feasible ϵ-approximate global minimizer of the program ($\mathrm{P}'_{\mathrm{NE}}$), $(\boldsymbol{\pi}^\star, \boldsymbol{w}^\star)$. Because a global minimum of the program is equal to 0, an ϵ-approximate global optimum must be at most $\epsilon > 0$. We observe that for every $k \in [n]$,

$$\boldsymbol{\rho}^\top \boldsymbol{w}_k^\star \geq \boldsymbol{\rho}^\top\left(\mathbf{I}-\gamma\,\mathbb{P}(\boldsymbol{\pi}^\star)\right)^{-1}\boldsymbol{r}_k(\boldsymbol{\pi}^\star), \tag{10}$$

which follows from induction on the inequality constraint (9).

Consequently, the assumption that

$$\epsilon \geq \boldsymbol{\rho}^\top \boldsymbol{w}_k^\star - \boldsymbol{\rho}^\top\left(\mathbf{I}-\gamma\,\mathbb{P}(\boldsymbol{\pi}^\star)\right)^{-1}\boldsymbol{r}_k(\boldsymbol{\pi}^\star)$$

and Equation (10), yields the fact that

$$\begin{aligned}\epsilon &\geq \boldsymbol{\rho}^\top \boldsymbol{w}_k^\star - \boldsymbol{\rho}^\top\left(\mathbf{I}-\gamma\,\mathbb{P}(\boldsymbol{\pi}^\star)\right)^{-1}\boldsymbol{r}_k(\boldsymbol{\pi}^\star)\\ &\geq V_k^{\dagger,\boldsymbol{\pi}^\star_{-k}}(\boldsymbol{\rho}) - V_k^{\boldsymbol{\pi}^\star}(\boldsymbol{\rho}),\end{aligned}$$

where the second inequality holds from the fact that $\boldsymbol{w}^\star$ is also feasible for ($\mathrm{P}'_{\mathrm{BR}}$). The latter concludes the proof, as the display coincides with the definition of an ϵ-approximate NE. □

Theorem C.2 (CCE collapse to NE in polymatrix MG—infinite-horizon)**.** *Let a zero-sum polymatrix switching-control Markov game, i.e., a Markov game for which Assumptions 1 and 2 hold. Further, let an ϵ-approximate CCE of that game $\boldsymbol{\sigma}$. Then, the marginal product policy $\boldsymbol{\pi}^\sigma$, with $\boldsymbol{\pi}_k^{\boldsymbol{\sigma}}(a|s) = \sum_{\boldsymbol{a}_{-k}\in\mathcal{A}_{-k}} \boldsymbol{\sigma}(a, \boldsymbol{a}_{-k}),\ \forall k \in [n]$ is an $n\epsilon$-approximate NE.*

Proof. Let an ϵ-approximate CCE policy, $\boldsymbol{\sigma}$, of game Γ. Moreover, let the best-response value-vectors of each agent k to joint policy $\boldsymbol{\sigma}_{-k}$, $\boldsymbol{w}_k^\dagger$.

Now, we observe that due to Assumption 1,

$$\begin{aligned}w_k^\dagger(s) &\geq r_k(s, a, \boldsymbol{\sigma}_{-k}) + \mathbb{P}_h(s, a, \boldsymbol{\sigma}_{-k})\boldsymbol{w}_k^\dagger\\ &= \sum_{j\in\mathrm{adj}(k)} r_{(k,j),h}(s, a, \boldsymbol{\pi}_j^{\boldsymbol{\sigma}}) + \mathbb{P}(s, a, \boldsymbol{\sigma}_{-k})\boldsymbol{w}_k^\dagger.\end{aligned}$$

Further, due to Assumption 2,

$$\mathbb{P}(s, a, \boldsymbol{\sigma}_{-k})\boldsymbol{w}_k^\dagger = \mathbb{P}(s, a, \boldsymbol{\pi}^{\boldsymbol{\sigma}}_{\mathrm{argctrl}(s)})\boldsymbol{w}_k^\dagger,$$

or,

$$\mathbb{P}(s, a, \boldsymbol{\sigma}_{-k})\boldsymbol{w}_k^\dagger = \mathbb{P}(s, a, \boldsymbol{\pi}^{\boldsymbol{\sigma}})\boldsymbol{w}_k^\dagger.$$

Putting these pieces together, we reach the conclusion that $(\boldsymbol{\pi}^\sigma, \boldsymbol{w}^\dagger)$ is feasible for the nonlinear program ($\mathrm{P}'_{\mathrm{NE}}$).

What is left is to prove that it is also an ϵ-approximate global minimum. Indeed, if $\sum_k \boldsymbol{\rho}^\top \boldsymbol{w}_k^\dagger \leq \epsilon$ (by assumption of an ϵ-approximate CCE), then the objective function of ($\mathrm{P}'_{\mathrm{NE}}$) will attain an ϵ-approximate global minimum. In turn, due to Theorem C.1 the latter implies that $\boldsymbol{\pi}^{\boldsymbol{\sigma}}$ is an $n\epsilon$-approximate NE. □

C.1 No equilibrium collapse with more than one controllers per-state

Example 2. *We consider the following 3-player Markov game that takes place for a time horizon $H = 3$. There exist three states, s_1, s_2, and s_3 and the game starts at state s_1. Player 3 has a single action in every state, while players 1 and 2 have two available actions $\{a_1, a_2\}$ and $\{b_1, b_2\}$ respectively in every state. The initial state distribution $\boldsymbol{\rho}$ is the uniform probability distribution over $\mathcal{S}$.*

Reward functions. *If player* 1 *(respectively, player* 2*) takes action* a_1 *(resp.,* b_1*), in either of the states* s_1 *or* s_2*, they get a reward equal to* $\frac{1}{20}$*. In state* s_3*, both players get a reward equal to* $-\frac{1}{2}$ *regardless of the action they select. Player* 3 *always gets a reward that is equal to the negative sum of the reward of the other two players. This way, the* zero-sum polymatrix property *of the game is ensured (Assumption 1).*

Transition probabilities. *If players* 1 *and* 2 *select the joint action* (a_1, b_1) *in state* s_1*, the game will transition to state* s_2*. In any other case, it will transition to state* s_3*. The converse happens if in state* s_2 *they take joint action* (a_1, b_1)*; the game will transition to state* s_3*. For any other joint action, it will transition to state* s_1*. From state* s_3*, the game transition to state* s_1 *or* s_2 *uniformally at random.*

At this point, it is important to notice that two players control the transition probability from one state to another. In other words, Assumption 2 does not hold.

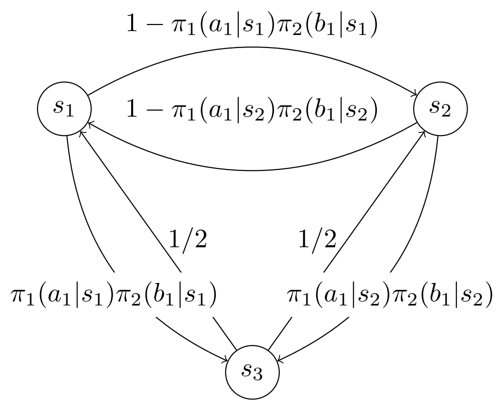

Figure 2: A graph of the state space with transition probabilities parametrized with respect to the policy of each player.

Next, we consider the joint policy $\boldsymbol{\sigma}$*,*

$$\boldsymbol{\sigma}(s_1) = \boldsymbol{\sigma}(s_2) = \begin{array}{c c} & \begin{array}{cc} b_1 & \;\; b_2 \end{array} \\ \begin{array}{c} a_1 \\ a_2 \end{array} & \begin{pmatrix} 0 & 1/2 \\ 1/2 & 0 \end{pmatrix} \end{array}.$$

Claim C.1. The joint policy $\boldsymbol{\sigma}$ that assigns probability $\frac{1}{2}$ to the joint actions (a_1, b_2) and (a_2, b_1) in both states s_1, s_2 is a CCE and $V_1^{\boldsymbol{\sigma}}(\boldsymbol{\rho}) = V_2^{\boldsymbol{\sigma}}(\boldsymbol{\rho}) = -\frac{1}{10}$.

Proof.

$$\begin{aligned} V_1^{\boldsymbol{\sigma}}(\boldsymbol{\rho}) &= \boldsymbol{\rho}^\top \left(\mathbf{I} - \gamma\, \mathbb{P}(\boldsymbol{\sigma})\right)^{-1} \boldsymbol{r}_1(\boldsymbol{\sigma}) \\ &= \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix} \begin{pmatrix} \frac{9}{5} & \frac{6}{5} & 0 \\ \frac{6}{5} & \frac{9}{5} & 0 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{40} \\ \frac{1}{40} \\ -\frac{1}{2} \end{pmatrix} \\ &= -\frac{1}{10}. \end{aligned}$$

We check every deviation,

- $\boldsymbol{\pi}_1(s_1) = \boldsymbol{\pi}_1(s_2) = \begin{pmatrix} 1 & 0 \end{pmatrix}, V^{\boldsymbol{\pi}_1 \times \boldsymbol{\sigma}_{-1}}(\boldsymbol{\rho}) = -\frac{2}{5}$,
- $\boldsymbol{\pi}_1(s_1) = \boldsymbol{\pi}_1(s_2) = \begin{pmatrix} 0 & 1 \end{pmatrix}, V^{\boldsymbol{\pi}_1 \times \boldsymbol{\sigma}_{-1}}(\boldsymbol{\rho}) = -\frac{1}{6}$,
- $\boldsymbol{\pi}_1(s_1) = \begin{pmatrix} 1 & 0 \end{pmatrix},\ \boldsymbol{\pi}_1(s_2) = \begin{pmatrix} 0 & 1 \end{pmatrix}, V^{\boldsymbol{\pi}_1 \times \boldsymbol{\sigma}_{-1}}(\boldsymbol{\rho}) = -\frac{5}{16}$,
- $\boldsymbol{\pi}_1(s_1) = \begin{pmatrix} 0 & 1 \end{pmatrix},\ \boldsymbol{\pi}_1(s_2) = \begin{pmatrix} 1 & 0 \end{pmatrix}, V^{\boldsymbol{\pi}_1 \times \boldsymbol{\sigma}_{-1}}(\boldsymbol{\rho}) = -\frac{5}{16}$.

For every such deviation the value of player 1 is smaller than $-\frac{1}{10}$. For player 2, the same follows by symmetry. Hence, $\boldsymbol{\sigma}$ is indeed a CCE.

$\square$

Yet, the marginalized product policy of $\boldsymbol{\sigma}$ which we note as $\boldsymbol{\pi}_1^{\boldsymbol{\sigma}} \times \boldsymbol{\pi}_2^{\boldsymbol{\sigma}}$ does not constitute a NE. The components of this policy are,

$$\begin{cases} \boldsymbol{\pi}_1^{\boldsymbol{\sigma}}(s_1) = \boldsymbol{\pi}_1^{\boldsymbol{\sigma}}(s_2) = \begin{array}{cc} a_1 & a_2 \\ \multicolumn{2}{c}{\begin{pmatrix} 1/2 & 1/2 \end{pmatrix}} \end{array}, \\ \\ \boldsymbol{\pi}_2^{\boldsymbol{\sigma}}(s_1) = \boldsymbol{\pi}_2^{\boldsymbol{\sigma}}(s_2) = \begin{array}{cc} b_1 & b_2 \\ \multicolumn{2}{c}{\begin{pmatrix} 1/2 & 1/2 \end{pmatrix}} \end{array}. \end{cases}$$

I.e., the product policy $\boldsymbol{\pi}_1^{\boldsymbol{\sigma}} \times \boldsymbol{\pi}_2^{\boldsymbol{\sigma}}$ selects any of the two actions of each player in states s_1, s_2 independently and uniformally at random. With the following claim, it can be concluded that in general when more than one player control the transition the set of equilibria do not collapse.

Claim C.2. The product policy $\boldsymbol{\pi}_1^{\boldsymbol{\sigma}} \times \boldsymbol{\pi}_2^{\boldsymbol{\sigma}}$ is not a NE.

Proof. For $\boldsymbol{\pi}^{\sigma} = \boldsymbol{\pi}_1^{\boldsymbol{\sigma}} \times \boldsymbol{\pi}_2^{\boldsymbol{\sigma}}$ we get,

$$\begin{aligned} V_1^{\boldsymbol{\pi}^{\sigma}} &= \boldsymbol{\rho}^\top \left(\mathbf{I} - \gamma \, \mathbb{P}(\boldsymbol{\pi}^{\boldsymbol{\sigma}})\right)^{-1} \boldsymbol{r}_1(\boldsymbol{\pi}^{\boldsymbol{\sigma}}) \\ &= \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix} \begin{pmatrix} \frac{34}{21} & \frac{20}{21} & \frac{3}{7} \\ \frac{20}{21} & \frac{34}{21} & \frac{3}{7} \\ \frac{6}{7} & \frac{6}{7} & \frac{9}{7} \end{pmatrix} \begin{pmatrix} \frac{1}{40} \\ \frac{1}{40} \\ -\frac{1}{2} \end{pmatrix} \\ &= -\frac{3}{10}. \end{aligned}$$

But, for the deviation $\boldsymbol{\pi}_1(a_1|s_1) = \boldsymbol{\pi}_1(a_1|s_2) = 0$, the value funciton of player 1, is equal to $-\frac{1}{6}$. Hence, $\boldsymbol{\pi}^{\sigma}$ is not a NE. $\square$

In conclusion, Assumption 1 does not suffice to ensure equilibrium collapse.

Theorem C.3 (No collapse—infinite-horizon)**.** *There exists a zero-sum polymatrix Markov game (Assumption 2 is not satisfied) that has a CCE which does not collapse to a NE.*

Proof. The proof follows from the game of Example 2, and Claims C.1 and C.2. $\square$