Stochastic Potential Games

David Mguni†

Abstract

Computing the Nash equilibrium (NE) for $N$-player non-zero-sum stochastic games is a formidable challenge. Currently, algorithmic methods in stochastic game theory are unable to compute NE for stochastic games (SGs) for settings in all but extreme cases in which the players either play as a team or have diametrically opposed objectives in a two-player setting. This greatly impedes the application of the SG framework to numerous problems within economics and practical systems of interest. In this paper, we provide a method of computing Nash equilibria in nonzero-sum settings and for populations of players more than two. In particular, we identify a subset of SGs known as stochastic potential games (SPGs) for which the (Markov perfect) Nash equilibrium can be computed tractably and in polynomial time. Unlike SGs for which, in general, computing the NE is $\text{PSPACE}$-hard, we show that SGs with the potential property are $\text{P}$-Complete. We further demonstrate that for each SPG there is a dual Markov decision process whose solution coincides with the MP-NE of the SPG. We lastly provide algorithms that tractably compute the MP-NE for SGs with more than two players.

1 Introduction

Game theory (GT) is a mathematical framework that is used for making predictions about outcomes in systems with self-interested and interacting players. In these settings, each player reasons about the environment and the actions of other players to decide on actions that maximise their own rewards. In the simplest setting, the situation is a one-off interaction in which the players’ rewards for each joint action are described by a matrix or a normal form game [OR94]. Stochastic games (SGs) [Sha53] generalise normal form games to a dynamic setting in which actions determine both an immediate reward and influence probabilistic transitions of a system across a sequence of states. The system state, in turn, determines the reward structure of the current interaction.

SGs have found widespread applications, economics problems [Mgu18; CG10], robotics [MG07], evolutionary biology [TH09] and computer networks [NAB09], smart grids [PA17] are a just few such examples. However, despite the success of GT in providing a formal tool for modelling various practical settings, a key challenge is computing the solution concept to games, i.e., the Nash equilibrium (NE) and its variants in non-zero-sum games [GS09]. Indeed, computing the Nash equilibrium for SGs has remained an open challenge. The absence of tractable techniques has prevented application of stochastic game theory to many real-world scenarios.

In stochastic settings with Markovian transitions, the appropriate solution concept is a Markov perfect Nash equilibrium (MP-NE) whose computation remains a formidable challenge in all but a handful of simple classes of games. Existing solutions of such MP-NE refer to extreme circumstances
namely two-player zero-sum games and so-called team games in which all players share the same reward function. In particular, tractable algorithms have been advanced for zero-sum SGs and team SGs [WS03]. In practice these settings are usually far away from reality of systems of interest. Indeed apart from a few restrictive cases in which neither the joint action (of all players) nor the current state influences the state transitions (so-called separable reward state independent games), to date there exists no reliable algorithm for computing NE for non-zero-sum SGs [RF91]. As such, the challenge of tractably computing MP-NE for N-player non-zero-sum SGs remains an unaddressed yet crucial problem. The obstacle in computing the NE in SGs is partly as a consequence of the fact that without a known structure to represent the problem tractably, the problem lies in a complexity class known as PPAD (Polynomial Parity Arguments on Directed graphs) which prohibits brute force and exhaustive search solution methods.

In this paper, we tackle the problem of computing the MP-NE of nonzero-sum SGs in a tractable fashion. For the systems we solve, the players’ interaction need not be perfectly adversarial nor are they required to act as a team (however there are cases of games in these classes also covered by our setting). In this sense, the problem solved by our invention covers solving MP-NE in for systems that better fit with most physical systems. Therefore to our knowledge, this paper provides the first method that enables the MP-NE (best-response) strategies in systems with populations of self-interested players to be computed tractably.

To perform this task, we study a subclass of nonzero-sum games that admit a particular structure on each stage game, namely games that satisfy a potential game property. Potential games (PGs) are a class of games that have been widely studied within GT in large part due to the fact they model various real-world scenarios [LCS16]. PGs are endowed with special properties that provide convergence guarantees to (pure strategy) NE. In particular, a number of algorithms such as best-response, fictitious play and best reply are all known to converge in PGs [You04].

Generalising PGs to SGs offers the prospect of conferring to SGs the convergence properties and the numerous computational benefits observed in static PGs. However, the PG framework has thus far resisted attempts to generalise to stochastic settings. Attempts to extend the potentiality property to SGs have thus far required a number of strong restrictive assumptions that impair the generality of the game.

We propose a dynamic formulation of PGs which generalises PGs to a stochastic setting with Markovian transitions. By introducing a novel construction of the stochastic generalisation, we show that exact solutions of the MP-NE of non-zero-sum SGs in this class can be computed tractably. To achieve this, we show that for this class of games, there exists a dual Markov decision process (MDP) the solutions of which coincide with the MP-NE of the SPG. As we demonstrate, this property induces a vast reduction in computational complexity for the task of computing the MP-NE.

Our analysis also reveals a continuity property of the MP-NE in SPGs for changes in the reward function. This result is particularly beneficial when large numbers of games must be analysed, a need which commonly arises in in mechanism design and principal-agent problems [Cai+18].

The contributions of the paper are as follows: i) We extend the notion of PGs to a stochastic setting from an extrapolation of the PG property at the stage game to cover the multi-stage SG. ii) We show that this allows for the problem of solving the game to be transformed into a team game which has an MDP representation. We then give some results on the resulting reduction in complexity for the problem of computing MP-NE, therefore demonstrating that solving MP-NE in SPGs is of a lower complexity than for general SGs. iii) We then give two algorithms which provably converge to MP-NE for SPGs. The first algorithm is a centralised algorithm which computes the MP-NE using a centralised dynamic programming method. The second algorithm is a distributed method which enables tractable computation of the MP-NE even when the player population grows to be large. iv) Lastly, we show that MP-NE of SPGs exhibit a continuity property under changes to the game reward.
functions. This property ensures that games that have similar properties all have MP-NE that all lie within some neighbourhood of each other.

2 Related work

PGs have been widely studied within GT. The first systematic treatment of PGs appeared in [MS96] in which it was shown that PGs are guaranteed to have a pure NE — a property which is not guaranteed in general by the Nash theorem. Since then a large body of literature has been dedicated to analysing the computational properties of various algorithms which seek to find the NE [LCS16]. Additionally, the study of PGs has been extended to settings with unknown rewards and in which the players’ observe noisy feedback [LM11; HCM17], infinite populations [San01; CH17] and multi-agent (mean field) reinforcement learning [MJC18].

However, the analysis of dynamic PGs is at present extremely sparse. Among the few, a dynamic PG with a deterministic transition function is studied in [Zaz+15]. It considers open-loop policies in which the players’ decisions take no account of the state but is a function of time. Generalising to stochastic settings is complicated by the fact that players must now execute policies that depend on the state and may need to directly take into account the actions of other players.

In this direction, [GH13] consider a SG with a potentiality property in which they derive sufficiency conditions for an NE in a Markov game. However the treatment requires two limiting assumptions: firstly each player’s reward function must be a concave function of the state. Secondly, the transition function is required to be invertible in order to express the policy in reduced form. Additionally, the solution method relies on verifying that a parameterised policy satisfies a set of sufficiency conditions which imposes further difficulties given the size of the space of functions. Macua et al. 2018 initiated the study of potential games within a stochastic setting thus paving the way for exporting the benefits for potential games to Markov games. In [MZZ18] however the potentiality condition is imposed directly on the value functions. This formulation requires that the players’ policies depend only on disjoint state-components meaning that players can only strategically respond to the actions of a local subset of players. This restriction prohibits non-local interactions between players.

In our analysis, we show by introducing the potentiality condition at the stage game, we can extrapolate the potentiality property to cover SGs. We then subsequently derive the generalised versions of the PG properties for the SG which reveals that the equilibrium of the game in pure strategies can be found by merely solving an MDP. However naïve attempts to solve such an MDP would lead to a combinatorial explosion in the number of players since the computation of optimal strategy in the MDP setting appears in a centralised form in which optimisations are performed over the joint action space. After demonstrating a method by which the potential can be computed, we provide an algorithm based on dynamic programming method in which the computation is distributed among the players which in turn avoids any combinatorial explosion.

This paper generalises the results of the static state-based PG in [Mar12] and the results in the deterministic dynamic PG in [Zaz+15] (with open loop controls) to now cover stochastic settings with state dependent (closed loop) controls. Additionally, unlike [MZZ18], our framework avoids the need to impose the disjointness assumption, in which the players’ strategies must depend only on disjoint subsets of components of the state.

Notation

We denote the set of players by \(\mathcal{N} := \{1, \ldots, N\}\) where \(N \in \mathbb{N}\). Given a metric space \(X\), we denote by \(d_X : X \times X \rightarrow \mathbb{R}_{\geq 0}\) the metric associated to \(X\) and by \(B_\alpha(x) \triangleq \{y \in X : \|x - y\| < \alpha\}\)
the open ball with radius \( \alpha > 0 \) around \( \mathbf{x} \in \mathbf{X} \). We denote by \((x_i)_{i \in \mathcal{N}} \equiv (x_1, \ldots, x_N)\) and by \(f_{-i} = (f_1, \ldots, f_{i-1}, f_{i+1}, \ldots, f_N)\) for a given set of functions \(\{f_i\}_{1 \leq i \leq N}\).  

## 3 Potential games

PGs are a class of games that have been widely studied within GT in large part due to the fact they model various real-world scenarios [LCS16]. Examples include cooperative control for consensus based problems e.g. dynamic sensor coverage, distributed control, smart grid [Bau16; Saa+12; MAS09], traffic network problems and spectrum sharing [Zaz+15] (more generally, congestion games), labelling within computer vision [YB95], cooperative control (team games) [FFC18], network resource allocation problems [SBP06].

A game is defined by a tuple \(\mathcal{M} = (\mathcal{N}, (\mathcal{A}_i)_{i \in \mathcal{N}}, (R_i)_{i \in \mathcal{N}})\) where \(\mathcal{N} := \{1, \ldots, N\}\) is the set of players, \(\mathcal{A}_i\) is an action set for each player \(i \in \mathcal{N}\) and \(R_i : x_{i \in \mathcal{N}} \mathcal{A}_i \rightarrow \mathbb{R}\) is the reward function for player \(i\). Each player \(i\) employs a strategy \(\pi_i \in \Pi_i\) which it uses to decide its action \(a^i \in \mathcal{A}_i\) where \(\Pi_i\) is the player \(i\) strategy space. We denote by \(\mathcal{A} = \times_{i \in \mathcal{N}} \mathcal{A}_i\) and \(\Pi := \times_{i \in \mathcal{N}} \Pi_i\) (the joint strategy set).

**Definition 1** A game \(\mathcal{M}\) is an (exact) PG if there exists a measurable function \(\phi : S \times \mathcal{A} \rightarrow \mathbb{R}\) such that the following holds for any \((a_i, a_{-i}), (a_i', a_{-i}) \in \mathcal{A}, \forall i \in \mathcal{N}, \forall s_t \in S:\)

\[
R_i(s_t, (a_i', a_i^{-i})) - R_i(s_t, (a_i, a_i^{-i})) = \phi(s_t, (a_i, a_i^{-i})) - \phi(s_t, (a_i', a_i^{-i})), \tag{1}
\]

where \(a_i^{-i} := (a_1^i, \ldots, a_{i-1}^i, a_{i+1}, \ldots, a_N^i)\).

Condition (1) says that the difference in payoff induced by a single deviation by one of the players is exactly quantified by a difference in \(\phi\), a function over state and joint actions.

The condition is satisfied in a vast range of games. Common examples include congestion games, network games, team games, identical interest games (in which players have identical reward functions) and classic games such as the prisoner’s dilemma [LCS16].

Since the functions \((R_i)_{i \in \mathcal{N}}\) are differentiable in the action inputs, we first we note that the following \(\forall a^i, a'^i \in \mathcal{A}_i, a^{-i} \in \mathcal{A}_{-i}, \forall s \in S:\)

\[
R_i(s, a^i, a^{-i}) - R_i(s, a'^i, a^{-i}) = \int_{a^i}^{a'^i} \frac{\partial R_i(s, a, a^{-i})}{\partial a} da \tag{2}
\]

and

\[
\phi(s, a^i, a^{-i}) - \phi(s, a'^i, a^{-i}) = \int_{a^i}^{a'^i} \frac{\partial \phi(s, a, a^{-i})}{\partial a} da \tag{3}
\]

Hence we quickly deduce that \(\forall a^i, a'^i \in \mathcal{A}_i, a^{-i} \in \mathcal{A}_{-i}, \forall s \in S:\)

\[
\frac{\partial R_i(s, a, a^{-i})}{\partial a} = \frac{\partial \phi(s, a, a^{-i})}{\partial a} \tag{4}
\]

It has long been understood that normal form PGs admit a team game representation [LCS16]. Consequently, the problem of finding a NE is reduced to finding some (global) maximum of a function, namely the potential of the game.\(^1\) In the case of Fig. 2, the NE for the prisoner’s dilemma \(((D, D))\) and coordination game \(((H, H))\) can be inferred immediately from the PG. This benefit is even more

\(^1\)For games with nonconvex potentials, other NE exist and correspond to local maxima of the potential.
apparent in games with larger actions sets and more players, to exemplify, an $N$–player game with binary action sets has $2^N$ possible outcomes.

In general, computing the NE in normal form games requires methods such as vertex enumeration, Lemke-Howson algorithm which is $\mathcal{PSPACE}$-complete [GPS13; Von02; M+96]. This is computational issue becomes further exacerbated in games beyond two players. In games for which a potential exists, reducing the problem to finding some maximum entry of a potential (matrix) leads to a vast reduction in complexity as illustrated in Fig. 1 and Fig. 2.

![Figure 1: Classic potential games](image1)

|   | $H$ | $P$ |
|---|-----|-----|
| $H$ | 10, 10 | 0, 0 |
| $P$ | 0, 0 | 5, 5 |

(a) Coordination Game

![Figure 2: Potential game representations](image2)

|   | $C$ | $D$ |
|---|-----|-----|
| $C$ | -1, -1 | -6, 0 |
| $D$ | 0, -6 | -4, -4 |

(b) Prisoner's Dilemma

The main contribution of this paper is to extend this computational advantage to a class of SGs with a potential property.

4 Stochastic Potential Games

We now discuss SPGs which is the main subject of the paper. First we give a description of SGs.

![Figure 3: A simple network (left) can be naturally described in terms of normal form games, but may be inadequate for modelling real world scenarios which involve changes of state over time. For larger and more complex networks (right), in which players traverse nodes over time, a more appropriate description is an SG. Although SGs are reducible to normal form games, the normal form representation yields an exponential growth in complexity in the number of states and players, rendering this approach intractable for games with even modestly sized player populations and state spaces. For example, in the case of the congestion game described by a network graph, reducing the game to a normal form description leads to exponential scaling in the size of the network.](image3)
An SG is an augmented MDP which proceeds by two or more players taking actions that \textit{jointly} manipulate the transitions of a system over $T \in \mathbb{N}$ rounds which may be infinite. At each round, the players receive some immediate reward or cost which is a function of the players’ current joint actions and the current state. In an SG, at a given time, the players simultaneously play one of many possible \textit{stage games} $\mathcal{M}$ which are indexed by states that lie within some state space $\mathcal{S}$. The outcome of each stage game $\mathcal{M}(s)$ depends on the joint action executed by the players where $a^i_s \in \mathcal{A}_i$ is the action taken by player $i \in \mathcal{N}$ and $s \in \mathcal{S}$ is the state of the world and $\mathcal{A}_i$ is the action set for player $i$. Given some stage game $\mathcal{M}(s)$ for $s \in \mathcal{S}$, the players simultaneously execute a joint action $a_s = (a^1_s, a^2_s, \ldots, a^N_s) \in \mathcal{A}$ and immediately thereafter, each player $i \in \mathcal{N}$ receives a payoff $R(i, a_s)$, the state then transitions to $s' \in \mathcal{S}$ with probability $P(s'|s, a_s)$ where the game $\mathcal{M}(s')$ is played.

Formally, we consider an SG defined by a tuple $\mathcal{G} = \langle \mathcal{N}, \mathcal{S}, (\mathcal{A}_i)_{i \in \mathcal{N}}, P, (R_i)_{i \in \mathcal{N}}, \gamma \rangle$ where $\mathcal{N} := \{1, \ldots, N\}$ is the set of players for some $N \in \mathbb{N}$, $\mathcal{S}$ is a finite set of states, $\mathcal{A}_i$ is an action set for each player $i \in \mathcal{N}$ and the function $R_i : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is the one-step reward for player $i$. The map $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ is a Markov transition probability matrix i.e. $P(s'|s, a_s)$ is the probability of the state $s'$ being the next state given the system is in state $s$ and the joint action $a_s \in \mathcal{A}$ is played. Therefore the SG proceeds as follows: given some stage game $\mathcal{M}(s) = \langle (\mathcal{A}_i)_{i \in \mathcal{N}}, (R_i)_{i \in \mathcal{N}}, \mathcal{N} \rangle$, the players simultaneously execute a joint action $a_s := (a^1_s, \ldots, a^N_s) \in \mathcal{A}$ and immediately thereafter, each player $i \in \mathcal{N}$ receives a payoff $R_i(s, a_s)$, the state then transitions to $s' \in \mathcal{S}$ with probability $P(s'|s, a_s)$ where the game $\mathcal{M}(s')$ is played in which the players receive a reward discounted by $\gamma$.

Now, each player employs a strategy, $\pi_i \in \Pi_i$ to decide its action at $s \in \mathcal{S}$. For an SG, $\mathcal{G}$, the goal of each player $i \in \mathcal{N}$ is to determine a policy $\pi_i \in \Pi_i$ that maximises the following quantity:

$$v^*_{i, \pi, \gamma}(s) = \mathbb{E}_{\pi_i, \pi_{-i}, s_t \sim P} \left[ \sum_{t \geq 0} \gamma^t R_i(s_t, a_t) | s_0 = s \right].$$

For each player $i \in \mathcal{N}$, a \textit{pure strategy} is a map $\pi_i : \mathcal{S} \rightarrow \mathcal{A}_i$ that assigns to every state an action in $\mathcal{A}_i$. Similarly, for each player $i \in \mathcal{N}$, a \textit{behavioural strategy} is a map $\pi_i : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$ that assigns to every state $s \in \mathcal{S}$ a probability distribution in $\mathcal{A}_i$.\footnote{By Kuhn’s theorem [Rit+02], if the player retains a history of previous actions and states, each behavioural strategy has an equivalent mixed strategy.} We denote the space of behavioural strategies (pure strategies) for each player $i \in \mathcal{N}$ by $\Pi_i (\Pi_i^\rho)$. Note that pure strategies are a degenerate class of behavioural strategies which assign to any state $s \in \mathcal{S}$ the Dirac measure with its probability mass concentrated at a single point.

\textbf{Remark 1} \textit{Note that a team game settings (e.g. [AY16]) correspond to the degenerate case of our setting when $R_1 = R_2 = \ldots = R_N$. Similarly, it can be shown PGs also encompass some zero-sum games (e.g. Ex. 4 in [Bal+18]).}

In order to describe the stable outcomes in an SG, we adopt the following variant of the NE which is appropriate for SGs with Markov transitions [FT91]:

\textbf{Definition 2} \textit{A strategy profile $\pi^* = (\pi^*_i, \pi^*_{-i}) \in \Pi$ is a Markov perfect Nash equilibrium (MP-NE) in Markov strategies if the following condition holds for any $i \in \mathcal{N}$:}

$$v^{(\pi^*_i, \pi^*_{-i})}_i (s) \geq v^{(\pi^*_i, \pi^*_{-i})}_i (s), \forall s \in \mathcal{S}, \forall \pi'_i \in \Pi_i.$$  

The condition characterises strategic configurations in which at any state no player can improve their expected cumulative rewards by unilaterally deviating from their current strategy.
For infinite horizon SGs with Markov transitions we can safely dispense with path dependencies in the strategies. Moreover, for MP-NE attention can be restricted to stationary stochastic policies (behavioural strategies) that depend only on the current state, namely stationary Markov strategies (Ergodic) payoffs in dynamic setting with Markovian transitions.

A pure strategy MP-NE (PS-MP-NE) is when condition (5) holds in pure strategies. Unlike mixed strategies, PS-MP-NE give predictions of the game outcomes which do not involves randomness introduced by the players. This endows the PS-MP-NE concept with strong predictive properties [OR94]. However, the existence of an NE is in general only guaranteed when the players execute mixed strategies. We denote by $NE\{G\}$ the set of MP-NE strategies for the game $G$.

Determining whether a (pure-strategy) NE exists in an SG is $\mathcal{PSPACE}$-hard even in the case in which the time horizon $T$ is finite [CS08]. In this paper, we show that for a class of SGs with a PG property, computing the MP-NE in pure strategies is reducible to solving an MDP and inherits the P-complete complexity class.

We now adapt the potentiality condition to accommodate a state-based setting with the following definition:

**Definition 3** A stage game $M(s)$ is an (exact) state-based potential game (PG) if there exists a function $\phi : S \times A \rightarrow \mathbb{R}$ ($\phi \in \mathcal{H}$) such that the following holds for any $(a_i, a_{-i}), (a'_i, a_{-i}) \in A$ where $a_t^{-i} := (a_1^{-i}, \ldots, a_{i-1}^{-i}, a_{i+1}^{-i}, \ldots, a_N^{-i})$ $\forall i \in \mathcal{N}, \forall s_t \in S$:

$$R_i(s_t, (a'_i, a_t^{-i})) - R_i(s_t, (a_i, a_t^{-i})) = \phi(s_t, (a'_i, a_t^{-i})) - \phi(s_t, (a_i, a_t^{-i})).$$ (6)

Condition (6) says that the difference in payoff induced by a single deviation by one of the players is exactly quantified by a difference in a function $\phi$, a function over state and joint actions.

Where convenient, we will use the shorthand $\phi_{\pi^i, \pi_{-i}}(s) \equiv \mathbb{E}_{\pi_i, \pi_{-i}}[\phi(s, a_i^t, a_{-i}^{-t})]$ and $R_i^{\pi^i, \pi_{-i}}(s) \equiv \mathbb{E}_{\pi_i, \pi_{-i}}[R_i(s, a_i^t, a_{-i}^{-t})]$ for any $s \in S$ and for any $\pi_i \in \Pi_i, \pi_{-i} \in \Pi_{-i}$ and for each $i \in \mathcal{N}$.

**Definition 4** An SPG, $G$ is an SG which each stage game $M$ is a PG. SPGs generalise PGs to the dynamic setting with Markovian transitions.

**Definition 5** A stage game $M(s)$ is state transitive if there exists a function $\phi : S \times A \rightarrow \mathbb{R}$ ($\phi \in \mathcal{H}$) such that the following holds for any $(a_i, a_{-i}) \in A$ where $\forall i \in \mathcal{N}, \forall s_t, s'_t \in S$:

$$R_i(s_t, (a'_i, a_t^{-i})) - R_i(s'_t, (a'_i, a_t^{-i})) = \phi(s_t, (a'_i, a_t^{-i})) - \phi(s'_t, (a'_i, a_t^{-i})).$$ (7)

The intuition is the following: consider a metric space $(S, d)$, then the reduction in reward faced by agent $i \in \mathcal{N}$ that seeks a goal state $s_i^* \in S$ but arrives at state $s_i' \in S$ is equal to the reduction in reward that agent $j \in \mathcal{N}$ faces when it seeks to arrive at goal state $s_j^* \in S$ but arrives at state $s_j' \in S$ whenever $d(s_i^*, s_i') = d(s_j^*, s_j')$ for a given metric $d$.

Having introduced an SPG, we now proceed to analysing the construction and demonstrating its properties.

In the next section, we perform the main analysis. There we prove that solving SPGs reduces to solving an MDP and inherits the P-complete complexity class.

---

3There are some exceptions for games with payoff structures not considered here for example, limiting average (Ergodic) payoffs [BF68].

4This condition is a milder condition than the construction of the PG in [Mar12] for which the PF is assumed to be non-decreasing along any action invariant state trajectory.
5 Main Analysis

We now turn to proving the main results. The first result reduces the task of finding the MP-NE of the game (which generally involves finding a fixed point) to solving an MDP. This vastly reduces the complexity of the problem and unlocks dynamic programming methods for computing the solution. Following this, we study the associated MDP and show that the MP-NE exhibits a continuity property w.r.t changes in the reward functions of the game.

Assumptions

The results contained within the paper are built under the following assumptions:

Assumption A.1. The functions \( \{R_i\}_{i \in \mathcal{N}} \) are bounded, measurable, differentiable functions in the action and state inputs.

Assumption A.2. The strategy spaces \( (\Pi_i)_{i \in \mathcal{N}} \) are compact.

Assumption A.3. The state transitivity assumption holds.

The results of Sec. 7 are built under this assumption:

Assumption A.4. The functions \( (R_i, \theta)_{i \in \mathcal{N}} \) satisfy assumption A.1 and are Lipschitz in the parameter \( \theta \), that is there exists constants \( (c_{R_i})_{i \in \mathcal{N}} > 0 \) s.th. for any \( \theta, \theta' \in \Theta \) and for any \( i \in \mathcal{N} \), we have that \( |R_{i, \theta} - R_{i, \theta'}| \leq c_{R_i} d(\theta, \theta') \).

We now state the main results of the paper:

**Theorem 1** There exists a PS-MP-NE whenever \( \forall s \in \mathcal{S} \)

\[
\sup_{\pi' \in \Pi} B^\pi'(s) \neq \emptyset
\]  

We prove the theorem through a set of results that first show the game has an equivalent representation in which all players maximise the same function and thus, play a team game. In particular, the PS-MP-NE of \( \mathcal{G} \) correspond to the extreme points of some functional.

**Theorem 2** SPGs are reducible to single player MDPs. In particular, for any SPG \( \mathcal{G} \), there exists a measurable function (which depends on \( \mathcal{G} \)) \( B : \Pi \times \mathcal{S} \to \mathbb{R} \) whose maxima are attained when a policy \( \pi^* \in \text{NE}(\mathcal{G}) \) is executed.

Theorem 2 reduces the problem of finding a fixed point in the space of strategies to finding optimal joint policy of an MDP. The theorem therefore reveals that unlike general SGs, games of this type lie in a lower complexity class.

Since solving the game is reducible to solving an MDP, we deduce the following which is a consequence of Theorem 1 in [PT87]:

**Corollary 1** Computing the PS-MP-NE for SPGs is P-Complete.

In Lemma 3, we give an exact characterisation of the function \( B \) by way of computing a line integral. Moreover, Algorithm 3 provides a means to compute \( B \) without needing to compute a line integral.

**Proposition 1** For any PG, there exists a function \( B : \Pi \times \mathcal{S} \to \mathbb{R} (B \in \mathcal{H}) \) such that the following holds for any \( i \in \mathcal{N} \):

\[
R_i(s, a_i^t, a_{-i}^t) = \phi(s, a_i^t, a_{-i}^t) + F_{-i}(s, a_{-i}^t),
\]

where \( F_i \) satisfies the following condition

\[
F_{-i}(s, a_{-i}^t) = F_{-i}(s', a_{-i}^t).
\]
The result generalises dummy-coordination separability known in PGs to a state-based setting [Sla94; 

**Proof 1** To establish the forward implication, we make the following observation which is straightforward:

\begin{align}
R_i(s, a^i_t, a^{-i}_t) - R_i(s, a^i_t, a^{-i}_t) \\
= \phi(s, a^i_t, a^{-i}_t) + F_{-i}(s, a^{-i}_t) - \left(\phi(s, a^i_t, a^{-i}_t) + F_{-i}(s, a^{-i}_t)\right)
\end{align}

(12)

To prove the reverse, and assume that the game is state-based potential. Let us now define the function

\begin{align}
T_i(s, a^i_t, a^{-i}_t) := R_i(s, a^i_t, a^{-i}_t) - \phi(s, a^i_t, a^{-i}_t),
\end{align}

then we observe that:

\begin{align}
R_i(s, a^i_t, a^{-i}_t) - R_i(s, a^i_t, a^{-i}_t) = \phi(s, a^i_t, a^{-i}_t) - \phi(s, a^i_t, a^{-i}_t)
\end{align}

(14)

\[\iff\]

\begin{align}
R_i(s, a^i_t, a^{-i}_t) - \phi(s, a^i_t, a^{-i}_t) = R_i(s, a^i_t, a^{-i}_t) - \phi(s, a^i_t, a^{-i}_t),
\end{align}

(16)

and hence

\begin{align}
T_i(s, a^i_t, a^{-i}_t) = T_i(s, a^i_t, a^{-i}_t),
\end{align}

(17)

which implies that \(T_i(s, a^i_t, a^{-i}_t) \equiv K_i(s, \pi^{-i}).\) In a similar way, writing \(T_i(s, a^i_t, a^{-i}_t) := R_i(s, a^i_t, a^{-i}_t) - \phi(s, a^i_t, a^{-i}_t)\) and using the state transitive property, we deduce that \(T_i(s', a^i_t, a^{-i}_t) = T_i(s, a^i_t, a^{-i}_t)\) which settles the proof.

**Proposition 2** For any joint strategy \((\pi_i, \pi_{-i}) \in \Pi,\) define by \(v_{i,k}\) the value function for the finite horizon game of length \(k \in \mathbb{N}\) (i.e. \(v_{i,k}(\pi_i, \pi_{-i})(s) := \mathbb{E}_{s_t \sim P, \pi_i, \pi_{-i}} \left[ \sum_{t=0}^{k} \gamma^t R_i(s_t, a_t) | s = s_0 \right] \) for any \(i \in \mathcal{N}\) and \(k < \infty\). Then there exists a measurable function \(B_k : \Pi \times S \rightarrow \mathbb{R}\) such that the following holds for any \(i \in \mathcal{N}\) and \(\forall \pi_i, \pi'_i \in \Pi_i, \forall \pi_{-i} \in \Pi_{-i}\) and \(\forall s \in S:\)

\begin{align}
\mathbb{E}_{s \sim P(\cdot)} \left[ v_{i,k}(\pi_i, \pi_{-i})(s) - v_{i,k}(\pi'_i, \pi_{-i})(s) \right] = \mathbb{E}_{s \sim P(\cdot)} \left[ B_k(\pi_i, \pi_{-i})(s) - B_k(\pi'_i, \pi_{-i})(s) \right].
\end{align}

(18)

Prop. 2 extends the potentiality property to finite horizon SGs. In doing so it shows that the useful properties of PGs are preserved within an MG setting. Unlike [GH13; MZZ18], Prop. 2 requires no further assumptions beyond the potentiality of each stage game.

The proof of the proposition is quite lengthy and is therefore deferred to the appendix.

Thus far we have established the relation (18) holds only for the finite horizon case. We now extend the coverage to the infinite horizon case in which we can recover the use of stationary strategies. In order to perform the extension, we first require some preliminary results (Lemma 1 and Lemma 2) which study the limiting behaviour of terms involving \(B_i\).

**Lemma 1** For any \(t' < \infty\), define by

\begin{align}
B^{(\pi_i, \pi_{-i})}(s) := \mathbb{E}_{s_t \sim P, \pi_i, \pi_{-i}} \left[ \sum_{t=0}^{t'} \gamma^t \phi(s_t, a_t) | s = s_0 \right]
\end{align}

then \(\exists B^{(\pi_i, \pi_{-i})} \) such that \(\forall s \in S\) and for any \((\pi_i, \pi_{-i}) \in \Pi,\)

\begin{align}
\lim_{t \to \infty} B^{(\pi_i, \pi_{-i})}(s) = B^{(\pi_i, \pi_{-i})}(s), \quad s = s_0,
\end{align}

(19)

where for any \(\pi_i \in \Pi_i, \pi_{-i} \in \Pi_{-i},\) the function \(B^{(\pi_i, \pi_{-i})}\) is given by:

\begin{align}
B^{(\pi_i, \pi_{-i})}(s) := \mathbb{E}_{P, \pi} \left[ \sum_{t=0}^{\infty} \gamma^t \phi(s_t, a_t) | s = s_0 \right].
\end{align}

(20)
The result is proven by showing that the sequence $B_{i_{n+1}}^{(\pi_i, \pi_{-i})}$, $B_{i_{n+1}}^{(\pi_i, \pi_{-i})}$, ... converges uniformly, that is the sequence is a Cauchy sequence.

**Proof 2** We prove the result by showing that the sequence $B_{i_{n+1}}^{(\pi_i, \pi_{-i})}$, $B_{i_{n+1}}^{(\pi_i, \pi_{-i})}$, ... converges uniformly, that is the sequence is a Cauchy sequence. In particular, we show that $\forall \epsilon > 0$, $\exists T' > 0$ s.t. $\forall t', t'' > T'$ and for any $\pi_i \in \Pi_i$, $\pi_{-i} \in \Pi_{-i}$

$$
\left\| B_{i_{t'}}^{(\pi_i, \pi_{-i})} - B_{i_{t''}}^{(\pi_i, \pi_{-i})} \right\| < \epsilon. \tag{21}
$$

Firstly, we deduce that the function $\phi$ is bounded since each $R_i$ is bounded also (see (45) in the appendix). Now w.l.o.g., consider the case when $t' \geq t''$. We begin by observing the fact that

$$
B_{i_{t'}}^{(\pi_i, \pi_{-i})}(s) - B_{i_{t''}}^{(\pi_i, \pi_{-i})}(s)
= \mathbb{E}_{s_t \sim \mathcal{P}(i_{t-1}, a_{t-1}), \pi_i, \pi_{-i}} \left[ \sum_{t=0}^{t'} \gamma^t \phi_t(s_t, a_t) - \sum_{t=0}^{t''} \gamma^t \phi_t(s_t, a_t) \right]
= \mathbb{E}_{s_t \sim \mathcal{P}(i_{t-1}, a_{t-1}), \pi_i, \pi_{-i}} \left[ \sum_{t=t'}^{t''} \gamma^t \phi_t(s_t, a_t) \right]. \tag{22}
$$

Hence, we find that

$$
\left| B_{i_{t'}}^{(\pi_i, \pi_{-i})}(s) - B_{i_{t''}}^{(\pi_i, \pi_{-i})}(s) \right|
= \left| \mathbb{E}_{s_t \sim \mathcal{P}(i_{t-1}, a_{t-1}), \pi_i, \pi_{-i}} \left[ \sum_{t=t'}^{t''} \gamma^t \phi_t(s_t, a_t) \right] \right|
\leq \sum_{t=t'}^{t''} \gamma^t \| \phi \|_\infty
\leq |\gamma| \left| \frac{\gamma^{t''} - \gamma^{t'}}{1 - \gamma} \right| \| \phi \|_\infty \tag{26}
\leq \left| \frac{\gamma^{t''}}{1 - \gamma} \right| \| \phi \|_\infty = e^{t''} \ln |\phi| \frac{\| \phi \|_\infty}{1 - \gamma} \tag{27}
= e^{-t'|\ln |\phi|} \left( \frac{\| \phi \|_\infty}{1 - \gamma} \right) \leq e^{-T'|\ln |\phi|} \left( \frac{\| \phi \|_\infty}{1 - \gamma} \right), \tag{28}
$$

using Cauchy-Schwarz and since $t' \geq t'' > T'$ and $\gamma \in [0, 1]$. The inequality of the proposition is true whenever $T'$ is chosen to satisfy

$$
T' \geq \left| \ln (\epsilon)(\ln (\gamma)) \left( \frac{\| \phi \|_\infty}{1 - \gamma} \right) \right|^{-1}, \tag{31}
$$

hence the result is proven.

**Lemma 2**

$$
\lim_{t \to \infty} \left| \mathbb{E}_{s_t \sim \mathcal{P}(i)} \left[ B_{i,t}^{(\pi_i, \pi_{-i})}(s) - B_{i,t}^{(\pi_i, \pi_{-i})}(s) \right] \right| < \infty.
$$
Proof 3 Our first task is to establish that the quantity
\[ \lim_{t \to \infty} \mathbb{E}_{s \sim P(i)} \left[ B_{i,t}^{(\pi_i, \pi_{-i})}(s) - B_{i,t}^{(\pi'_i, \pi_{-i})}(s) \right] \]

is in fact, well-defined for any \( s \in S \) and \( \forall i \in \mathcal{N} \).

This is true since by (\ref{eq:bound}) for any \( t > 0 \) we have that
\[ \left| \mathbb{E}_{s \sim P(i)} \left[ B_{i,t}^{(\pi_i, \pi_{-i})}(s) - B_{i,t}^{(\pi'_i, \pi_{-i})}(s) \right] \right| < \infty. \]

and hence we have that
\[ \left| \mathbb{E}_{s \sim P(i)} \left[ B^{(\pi_i, \pi_{-i})}(s) - B^{(\pi'_i, \pi_{-i})}(s) \right] \right| < \infty. \]

To see this, we firstly observe that by the boundedness of \( R_i \), \( \exists c > 0 \) s.th. \( \forall t \in \mathbb{N}, \forall i \in \mathcal{N} \) and for any \( \pi_i \in \Pi_i, \pi_{-i} \in \Pi_{-i} \)
\[ \left| v_{i,t}^{(\pi_i, \pi_{-i})}(s) - v_{i,t}^{(\pi'_i, \pi_{-i})}(s) \right| < c. \]

This is true since for any \( k < \infty \) we have
\[ v_{i,k}^{(\pi_i, \pi_{-i})}(s) - v_{i,k}^{(\pi'_i, \pi_{-i})}(s) \]
\[ \begin{align*}
= & \mathbb{E}_{s_t \sim P, \pi_i, \pi_{-i}} \left[ \sum_{t=0}^{k} \gamma^t R_i(s_t, \alpha_t) \right] - \mathbb{E}_{s_t \sim P, \pi'_i, \pi_{-i}} \left[ \sum_{t=0}^{k} \gamma^t R_i(s_t, \alpha_t) \right] \\
\leq & \left| \mathbb{E}_{s_t \sim P, \pi_i, \pi_{-i}} \left[ \sum_{t=0}^{k} \gamma^t R_i(s_t, \alpha_t) \right] - \mathbb{E}_{s_t \sim P, \pi'_i, \pi_{-i}} \left[ \sum_{t=0}^{k} \gamma^t R_i(s_t, \alpha_t) \right] \right| \\
\leq & \sum_{t=0}^{k} \gamma^t \| R_i \|_\infty = 2 \frac{1 - \gamma^k}{1 - \gamma} \| R_i \|_\infty.
\end{align*} \]

Therefore, by the bounded convergence theorem we have that
\[ \lim_{t \to \infty} \left| \mathbb{E}_{s \sim P(i)} \left[ v_{i,t}^{(\pi_i, \pi_{-i})}(s) - v_{i,t}^{(\pi'_i, \pi_{-i})}(s) \right] \right| < \infty. \]

Now, using (\ref{eq:bound}), we deduce that for any \( \epsilon > 0 \), the following statement holds:
\[ \left| \mathbb{E}_{s \sim P(i)} \left[ B_{i,t}^{(\pi_i, \pi_{-i})}(s) - B_{i,t}^{(\pi'_i, \pi_{-i})}(s) \right] \right| < \left| \mathbb{E}_{s \sim P(i)} \left[ v_{i,t}^{(\pi_i, \pi_{-i})}(s) - v_{i,t}^{(\pi'_i, \pi_{-i})}(s) \right] \right| + \epsilon, \]

which after taking the limit as \( t \to \infty \) and using (\ref{eq:bound}), Lemma 1 and the dominated convergence theorem, we find that
\[ \lim_{t \to \infty} \left| \mathbb{E}_{s \sim P(i)} \left[ B_{i,t}^{(\pi_i, \pi_{-i})}(s) - B_{i,t}^{(\pi'_i, \pi_{-i})}(s) \right] \right| < \infty. \]

We now extend the finite horizon dynamic potential property (\ref{eq:finite_horizon}) to the infinite horizon case:

Proposition 3 There exists a measurable function \( B : \Pi \times S \to \mathbb{R} \) such that the following holds for any \( i \in \mathcal{N} \) and \( \forall \pi_i, \pi'_i \in \Pi_i, \forall \pi_{-i} \in \Pi_{-i} \) and \( \forall s \in S \):
\[ \mathbb{E}_{s \sim P(i)} \left[ v_{i}^{(\pi_i, \pi_{-i})}(s) - v_{i}^{(\pi'_i, \pi_{-i})}(s) \right] = \mathbb{E}_{s \sim P(i)} \left[ B^{(\pi_i, \pi_{-i})}(s) - B^{(\pi'_i, \pi_{-i})}(s) \right]. \]
Proof 4 The result is proven by contradiction. To this end, let us firstly assume there exists $c \neq 0$ such that

$$
\mathbb{E}_{s \sim P(\cdot)} \left[ v_i^{(\pi_i, \pi_{-i})}(s) - v_i^{(\pi'_i, \pi_{-i})}(s) \right] - \mathbb{E}_{s \sim P(\cdot)} \left[ B_i^{(\pi_i, \pi_{-i})}(s) - B_i^{(\pi'_i, \pi_{-i})}(s) \right] = c.
$$

Let us now define the following quantities for any $s \in S$ and for each $\pi_i \in \Pi_i$ and $\pi_{-i} \in \Pi_{-i}$ and $\forall i \in N$:

$$
v_{i,T'}^{(\pi_i, \pi_{-i})}(s) := \sum_{t=0}^{T'} \mu(s_0)\pi_i(a_0^i, s_0)\pi_{-i}(a_0^{-i}, s_0) \prod_{j=0}^{t-1} \gamma^j P(s_{j+1}|s_j, a_j^i, a_j^{-i}) \cdot \pi_i(a_j^i|s_j)\pi_{-i}(a_j^{-i}|s_j) R_i(s_t, a_t^i, a_t^{-i}),
$$

and

$$
B_{T'}^{(\pi_i, \pi_{-i})}(s) := \sum_{t=0}^{T'} \mu(s_0)\pi_i(a_0^i, s_0)\pi_{-i}(a_0^{-i}, s_0) \prod_{j=0}^{t-1} \gamma^j P(s_{j+1}|s_j, a_j^i, a_j^{-i}) \cdot \pi_i(a_j^i|s_j)\pi_{-i}(a_j^{-i}|s_j) \phi(s_t, a_t^i, a_t^{-i}),
$$

so that the quantity $v_{i,T'}^{(\pi_i, \pi_{-i})}(s)$ measures the expected cumulative return until the point $T' < \infty$.

Hence, we straightforwardly deduce that

$$
v_i^{(\pi_i, \pi_{-i})}(s) \equiv v_{i,\infty}^{(\pi_i, \pi_{-i})}(s) = v_{i,T'}^{(\pi_i, \pi_{-i})}(s) + \gamma^T \mu(s_0)\pi_i(a_0^i, s_0)\pi_{-i}(a_0^{-i}, s_0) \prod_{j=0}^{T-1} \gamma^j P(s_{j+1}|s_j, a_j^i, a_j^{-i}) \cdot \pi_i(a_j^i|s_j)\pi_{-i}(a_j^{-i}|s_j) v_{i,T'}^{(\pi_i, \pi_{-i})}(s_{T'}).$$

Next we observe that:

$$
c = \mathbb{E}_{s \sim P(\cdot)} \left[ \left( v_i^{(\pi_i, \pi_{-i})} - v_i^{(\pi'_i, \pi_{-i})} \right)(s) \right] - \mathbb{E}_{s \sim P(\cdot)} \left[ \left( B_i^{(\pi_i, \pi_{-i})} - B_i^{(\pi'_i, \pi_{-i})} \right)(s) \right]
$$

$$
= \mathbb{E}_{s \sim P(\cdot)} \left[ \left( v_i^{(\pi_i, \pi_{-i})} - v_i^{(\pi'_i, \pi_{-i})} \right)(s) \right] - \mathbb{E}_{s \sim P(\cdot)} \left[ \left( B_i^{(\pi_i, \pi_{-i})} - B_i^{(\pi'_i, \pi_{-i})} \right)(s) \right]
$$

$$
+ \gamma^T \mathbb{E}_{s_{T'} \sim P(\cdot)} \left[ \mu(s_0)\pi_i(a_0^i, s_0)\pi_{-i}(a_0^{-i}, s_0) \prod_{j=0}^{T-1} P(s_{j+1}|s_j, a_j^i, a_j^{-i}) \cdot \pi_i(a_j^i|s_j)\pi_{-i}(a_j^{-i}|s_j) \left( v_i^{(\pi_i, \pi_{-i})}(s_{T'}) - B_i^{(\pi_i, \pi_{-i})}(s_{T'}) \right) \right]
$$

$$
- \mu(s_0)\pi'_i(a_0^i, s_0)\pi_{-i}(a_0^{-i}, s_0) \prod_{j=0}^{T'-1} P(s_{j+1}|s_j, a_j^i, a_j^{-i}) \cdot \pi'_i(a_j^i|s_j)\pi_{-i}(a_j^{-i}|s_j) \left( v_i^{(\pi'_i, \pi_{-i})}(s_{T'}) - B_i^{(\pi'_i, \pi_{-i})}(s_{T'}) \right) \right].
$$

Considering the last expectation and its coefficient and denoting the product by $\kappa$, using the fact that by the Cauchy-Schwarz inequality we have $\|AX - BY\| \leq \|A\|\|X\| + \|B\|\|Y\|$, moreover whenever
\[ A, B \text{ are non-expansive}^5 \text{ we have that } \|AX - BY\| \leq \|X\| + \|Y\|, \text{ hence we observe the following bound:} \]
\[
\kappa \leq \|\kappa\| \leq 2\gamma^{T'} (\|v_i\| + \|B\|). \tag{41}
\]

Since we can choose \(T'\) freely and \(\gamma \in ]0, 1[\), we can choose \(T'\) to be sufficiently large so that
\[
\gamma^{T'} (\|v_i\| + \|B\|) < \frac{1}{4} |c|. \tag{42}
\]

This then implies that
\[
\left| \mathbb{E}_{s \sim P(\cdot)} \left[ \left( v_i^{(\pi_i, \pi_{-i})} - v_i^{(\pi_i', \pi_{-i})} \right)(s) - \left( B^{(\pi_i, \pi_{-i})} - B^{(\pi_i', \pi_{-i})} \right)(s) \right] \right| > \frac{1}{2} c,
\]

which is a contradiction since thanks to Prop. 2, we have proven that for any finite \(T'\) it is the case that
\[
\mathbb{E}_{s \sim P(\cdot)} \left[ \left( v_i^{(\pi_i, \pi_{-i})} - v_i^{(\pi_i', \pi_{-i})} \right)(s) - \left( B^{(\pi_i, \pi_{-i})} - B^{(\pi_i', \pi_{-i})} \right)(s) \right] = 0,
\]

and hence we deduce the thesis.

Prop. 3 indicates that the potentiality property is generalisable to SGs. In particular, there is a functional which mimics the behaviour of the potential function in SGs.

In a similar way to (4), we can obtain the following differential equation for \(B\) for any \(\pi \in \Pi, \forall i \in \mathcal{N}, \forall s \in S:\n\]
\[
\frac{\partial v_i^{\pi}}{\partial \pi_i}(s) = \frac{\partial B^{\pi}}{\partial \pi_i}(s). \tag{43}
\]

Our next result shows that the set of joint policies that maximise \(B\) are in fact NE policies of the game.\(^6\) We will use this result to show that computing the NE for SPGs is a has a significantly reduced computational complexity in comparison to SGs without this property.

**Proposition 4** There exists a measurable function \(B : S \times \Pi \to \mathbb{R}\) such that for any \(s \in S\) we have that
\[
\pi \in \arg \sup_{\pi' \in \Pi} B^{\pi'}(s) \implies \pi \in NE\{\mathcal{G}\}. \tag{44}
\]

**Proof 5** We do the proof by contradiction. Let \(\pi = (\pi_1, \ldots, \pi_N) \in \arg \sup_{\pi' \in \Pi} B^{\pi'}(s)\). Let us now therefore assume that \(\pi \notin NE\{\mathcal{G}\}, \text{ hence there exists some other strategy profile } \pi' = (\pi'_1, \ldots, \pi'_i, \ldots, \pi_N) \text{ which contains at least one profitable deviation by one of the players so that } \pi'_i \neq \pi_i \text{ for } i \in \mathcal{N} \text{ i.e.} \]
\[
v_i^{(\pi_i', \pi_{-i})}(s) > v_i^{(\pi_i, \pi_{-i})}(s) \text{ (using the preservation of signs of integration). Prop. 3 however implies that} \]
\[
B^{(\pi'_i, \pi_{-i})}(s) - B^{(\pi_i, \pi_{-i})}(s) > 0 \text{ which is a contradiction since } \pi = (\pi_i, \pi_{-i}) \text{ is a maximum of } B.
\]

Having reduced the problem of finding a fixed point MP-NE to maximising \(B\), we now characterise \(B\). We later show that \(B\) can be obtained using an iterative algorithm.

\(^5\)An operator \(T : \mathcal{V} \to \mathcal{V}\) is non-expansive if \(\forall V_1, V_2 \in \mathcal{V}\) we have: \(\|TV_1 - TV_2\| \leq \|V_1 - V_2\|\).

\(^6\)More accurately, the MP-NE of the game correspond to local maxima of \(B\).
Lemma 3 The function $B$ is given by the following expression for $s \in S, \forall \pi \in \Pi$:

$$B^\pi(s) - B'^\pi(s) = \mathbb{E}_{s_t \sim \pi} \left[ \sum_{t=0}^{\infty} \sum_{i \in \mathcal{N}} \gamma^t \int_0^1 \gamma'(z) \frac{\partial R_i}{\partial \pi_i}(s_t, \gamma(z)) \bigg| s = s_0 \right],$$

where $\gamma(z)$ is a continuous differentiable path in $\Pi$ connecting two strategy profiles $\pi \in \Pi$ and $\pi' \in \Pi$.

Proof 6 To ascertain the result, we note that from equation (2), using the gradient theorem of vector calculus, it is straightforward to deduce that the potential function $\phi$ can be computed from the reward functions $(R_i)_{i \in \mathcal{N}}$ via the following expression [MS96]:

$$\phi^\pi(s) = \phi'^\pi(s) + \sum_{i \in \mathcal{N}} \int_0^1 \gamma'(z) \frac{\partial R_i}{\partial \pi_i}(s_t, \gamma(z)), \quad (45)$$

where $\gamma(z)$ is a continuous differentiable path in $\Pi$ connecting two strategy profiles $\pi \in \Pi$ and $\pi' \in \Pi$.

We then deduce the result for the finite case after inserting (45) into (20).

For the team Markov game case, the function $B$ is straightforward to construct. In particular, let $J$ be the global performance function for the team game $G$, then we have that $B^\pi(s) := \mathbb{E}_{s_t \sim \pi, \pi'} \left[ \sum_{t=0}^{\infty} \gamma^t J(s_t, a_t) \bigg| s \equiv s_0 \right]$.

Proof 7 (Proof of Theorem 2) Combining Prop. 4 with Prop. 3 proves Theorem 2.

The characterisation of $B$ provided by Lemma 3 yields a function $B$ which is not unique. What remains is to investigate any relation between such functions. The following result indicates that the space of functions $B$ contains functions that differ only by a constant:

Lemma 4 Let $B_1$ and $B_2$ be dynamic potentials that satisfy relation (3). Then the following relation is satisfied:

$$B_1^{\pi_i, \pi_{-i}}(s) - B_2^{\pi_i, \pi_{-i}}(s) = c(\lambda), \quad \forall (\pi_i, \pi_{-i}) \in \Pi, \forall s \in S,$$

where $\lambda$ is the discount factor of $G$.

The result is proven after a straightforward extension of the static case (Lemma 2.7. in [MS96]).

6 Computing the equilibria

We now propose a set of algorithms to compute the MP-NE for the game iteratively. The first algorithm is a centralised method which performs joint updates of the players’ actions with a known potential function. The second algorithm computes the MP-NE in a decentralised fashion so that each player performs independent updates. Unlike the first algorithm, this algorithm does not require the potential function $\phi$ to be known in advance.

Having reduced the game to a standard MDP, we now introduce the following Bellman operator from which, by the above analysis, it follows can be used to compute the solution to the game.

We now introduce the following recursive dynamic programming relationship:

$$B_k(s) = \max_{a \in \mathcal{A}} \left[ \phi(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s, a) B_{k-1}(s') \right] \quad (46)$$

The following result is a direct consequence of Prop. 4 and the fact that thanks to Lemma 1 for any $s \in \mathcal{S}$, $B(s)$ may be expressed as $\lim_{t \to \infty} B_t^{(\pi_i, \pi_{-i})}(s) = B^{(\pi_i, \pi_{-i})}(s)$. 

14
Algorithm 1
1: For a given SG $G$, require $B$, $\phi$, initial strategy profile $\pi$, constant $\epsilon$
2: Initialise $B_0(s)$ arbitrarily $\forall s \in S$.
3: repeat
4:   $k \leftarrow 0$
5:   for $s \in S$, $i = 0, 1, \ldots, N$ do
6:     Build auxiliary game $\phi(s, a) + \gamma \sum_{s' \in S} P(s'|s, a) B_0(s')$
7:     return $B_1(s) = \max_{a \in A} [\phi(s, a) + \gamma \sum_{s' \in S} P(s'|s, a) B_0(s')]$
8:     $B_k(s) \leftarrow \max_{a \in A} [\phi(s, a) + \gamma \sum_{s' \in S} P(s'|s, a) B_{k-1}(s')]$
9:   $k \leftarrow k + 1$
10: end for
11: until $|B_k(s) - B_{k-1}(s)| < \epsilon, \forall s \in S$.

Lemma 5 Consider the sequence $T^k B^{\pi'}(\cdot), T^{k+1} B^{\pi'}(\cdot), \ldots$ and let $\pi^\infty \equiv (\pi_1, \ldots, \pi_N) \in \Pi$ be the joint policy extracted from the limit of the sequence then $\pi^* \in NE\{G\}$.

Algorithm 1 follows as a direct consequence of Lemma 5 which indicates that finding the maximum of $B$ can be obtained by dynamic programming.

Algorithm 1 is a dynamic programming method in which the game is treated as an MDP with value function $B$. However, such a method computes the PS-MP-NE in a way that involves centralised calculations within the joint action space. It additionally requires that the potential function $\phi$ be known in advance or that $B$ be computed using a line integral which may not be feasible in all cases.

We now provide a method of decentralising the computation in order to calculate the solution to each stage game. Additionally, the following algorithm avoids the need to compute a line integral. The following method relies on computing the NE for each stage game then computing the function $B$.

Integral to the algorithm is a generalised weakened fictitious play (GWFP) step. GWFP updates the weights each player places on an action in the direction of the best response to their GWFP has strong convergence guarantees for PGs [LC06].

Definition 6 A path is a sequence of elements $\bar{\pi} = (\pi(k))^{\infty}_1$ of elements in $\Pi$.

Note the the function obeys the following Bellman equation:

$$B_k(s) = \max_{a \in A} \left[ \phi(s, a) + \gamma \sum_{s' \in S} P(s'|s, a) B_{k-1}(s') \right], \quad k = 1, 2, \ldots, (47)$$

Algorithm 2 can be broken up into three components. The first computes the potential function $\phi$ (for each $s \in S$) given the set of player rewards $(R_i)_{i \in N}$. Second, the algorithm computes the MP-NE for each auxiliary game defined by $B_k^{(i)}$ using GWFP for $0 < k < \infty$. Note that each auxiliary game is constructed using step 8 (so that the entries of this game contain the future stream of payoffs under the current strategy). Lastly, the algorithm computes $B^{\pi^*}$ using a value iterative approach. The first component generalises current methods (e.g. the traverse procedure in [LCS16]) to extract the potential to the case in which the reward functions are state dependent.

Steps 9 - 10 involve a generalised weakened fictitious play (GWFP) step. GWFP updates the weights each player places on an action in the direction of the best response to their opponents’ actions. GWFP has strong convergence guarantees for PGs [LC06].

Theorem 3 Algorithm 2 converges to $\pi^*$. 

15
Algorithm 2
1: For a given SG $G$, require $B$, initial strategy profile $\pi$, constant $\epsilon$
2: compute potential function $\phi$
3: For each $s \in S$ find $\phi$ using Algorithm 3.
4: Initialise $B_0(s)$ arbitrarily $\forall s \in S$.
5: repeat
6: $k \leftarrow 0$
7: for $s \in S$, $i = 0, 1, \ldots, N$ do
8: Build auxiliary game $\phi(s, a) + \gamma \sum_{s' \in S} P(s'|s, a) B_0(s')$
9: Compute optimal joint action using actor-critic generalised weakened fictitious play
10: return $B_1(s) = \max_{a \in A} [\phi(s, a) + \gamma \sum_{s' \in S} P(s'|s, a) B_0(s')]$
11: $B_k(s) \leftarrow \max_{a \in A} [\phi(s, a) + \gamma \sum_{s' \in S} P(s'|s, a) B_{k-1}(s')]$
12: $k \leftarrow k + 1$
13: end for
14: until $|B_k(s) - B_{k-1}(s)| < \epsilon, \forall s \in S$

Proof 8 The proof of the algorithm follows from the following observations: i) by Prop. 2, for each $k$th iteration $B_k$ has a finite improvement path for which GWFP converges to $\max_{\pi} [\phi(s, a) + \gamma \sum_{s' \in S} P(s'|s, a) B_k(s')]$ ii) by Lemma 3, $\forall s \in S$, the sequence $T^k B^\pi(s), T^{k+1} B^\pi(s), \ldots$ converges to the $B^\pi(s)$ where $\pi^* \in NE\{G\}$.

This completes the proof.

7 Regularity of Stochastic Potential Games

In general, MP-NE are highly sensitive to even small changes in the game [FT91]. In general, SGs have the property that small changes in the structure of the game can incur large changes to the equilibrium of the game. Consequently, the equilibria of a SG are not informative about the outcomes of other close SGs. The implications are that even intentional modification to multiplayer systems can lead to highly undesirable outcomes [Tir+16].

The following sets of results establish that unlike SGs in general, SPGs enjoy smoothness properties with respect to changes in its components.

First, we formalise the notion of continuity of the MP-NE with the following definition:

Definition 7 (Essentiality) Given a metric space $X$, let $B_\alpha(x) \triangleq \{ y \in X : \|x - y\| < \alpha \}$ denote the open ball with radius $\alpha > 0$ and denote by $G(\theta)$ the game $G(\theta) = (N, S, (A_i)_{i \in N}, P, (R_i)_{i \in N}, \gamma)$ where $\{R_i, \theta\}_{i \in N}$ are parameterised reward functions and $\theta \in \Theta$. Then $x \in NE\{G(\theta)\}$ is essential in $\theta \in \Theta$ if for any $\epsilon > 0$, $\exists \delta > 0 : \theta' \in B_\epsilon(\theta) \implies x' \in B_\delta(x)$, for any $x' \in NE\{G(\theta')\}$.

Definition 7 says that for any $\theta \in \Theta$ a game $G(\theta)$ is essential if small changes in $\theta$ produce only a small change in the NE of the game.

Theorem 4 Let $\{G_\theta\}_{\theta \in \Theta} := \{R, S, A, P, R_\theta, \gamma\}_{\theta \in \Theta}$ be a family of SPGs parameterised by $\theta \in \Theta \subset \mathbb{R}^w$. Then the MP-PS-NE are essential in $\theta \in \Theta$.

The theorem is proven in two steps. Firstly, it is shown that the value functions of MDP exhibit a Lipschitz property in parameterisations in their reward functions (see Prop. 5 in the appendix).
Proposition 5 Let \( \{M_\theta\}_{\theta \in \Theta} := \langle R, S, A, P, R_\theta, \gamma \rangle_{\theta \in \Theta} \) be a family of MDPs and denote by \( v^{\pi^*(\theta)} \) the optimal value function for the MDP \( M_\theta \), where \( R_\theta \) is a reward function which is Lipschitz in the parameter \( \theta \), then for any \( \theta, \theta' \in \Theta \):

\[
d_1 \left( v^{\pi^*(\theta)}, v^{\pi^*(\theta')}, \theta, \theta' \right) \leq \frac{2\|R\|_\infty}{1 - \gamma} L_{R \theta} (\theta, \theta'),
\]

Proof 9 Consider two MDPs, \( M_1 \) and \( M_2 \) and let us suppose \( v^{\pi^*, \theta} := \sup_{\pi \in \Pi} v^{\pi, \theta} \) and \( v^{\pi^*, \theta'} := \sup_{\pi \in \Pi} v^{\pi, \theta'} \) are the optimal value functions associated to \( M_1 \) and \( M_2 \) respectively. Our task is to show that given two reward functions \( R_\theta \) and \( R_{\theta'} \), the value functions satisfy the following estimate:

\[
d \left( v^{\pi^*, \theta}, v^{\pi^*, \theta'} \right) \leq c d (\theta, \theta'),
\]

for some \( c > 0 \) which shows that the MDPs are Lipschitz in \( \theta \in \Theta \).

Consider the MDPs \( M_1 \) and \( M_2 \) which differ only in their reward functions, i.e. \( M_1 := \langle R_\theta, S, A, P, \gamma \rangle \) and \( M_2 := \langle R_{\theta'}, S, A, P, \gamma \rangle \).

To prove the result, using Lemma 6, we have that

\[
\left\| \sup_{\pi \in \Pi} v^{\pi^*, \theta} - \sup_{\pi \in \Pi} v^{\pi^*, \theta'} \right\| \leq \sup_{\pi \in \Pi} \left\| v^{\pi^*, \theta} - v^{\pi^*, \theta'} \right\|
\]

\[
= \sup_{\pi \in \Pi} \left\| \sum_{t=0}^{T} \gamma^t R_\theta(s_t, a_t) - \sum_{t=0}^{T} \gamma^t R_{\theta'}(s_t, a_t) \right\|
\]

\[
= \sup_{\pi \in \Pi} \sum_{t=0}^{T} \gamma^t \left\| R_\theta(s_t, a_t) - R_{\theta'}(s_t, a_t) \right\|
\]

\[
\leq \tilde{c}_i \| \theta - \theta' \|
\]

where \( \tilde{c}_i := \frac{L_{R_\theta}}{1 - \gamma} \| R\|_\infty \) using the boundedness and the Lipschitz condition on \( R_\theta \) and \( \gamma > 0 \) which concludes the proof of the proposition.

Proposition 6 Let \( \pi^* \in NE\{G(\theta)\} \) be a PS-MP-NE joint policy for the players playing the game \( G(\theta) \) for some \( \theta \in \Theta \), then \( B \) is Lipschitz continuous in \( \theta \), in particular

\[
d_1 \left( B^{\pi^*(\theta)}, B^{\pi^*(\theta')}, \theta, \theta' \right) \leq 2 \frac{N}{1 - \gamma} L_{B \infty} d_2 (\theta, \theta').
\]

where \( L_{B \theta} \) are Lipschitz and L-Lipschitz constants.

Proof 10 For any \( \theta \in \Theta \) let the joint policy \( \pi^*(\theta) \in \Pi \) be any PS-MP-NE joint policy for the game \( G(\theta) \) i.e. \( \pi^*(\theta) \in NE\{G(\theta)\} \) then using Prop. 4 it suffices to deduce the existence of a constant \( c \) s.th.

\[
\left\| \sup_{\pi \in \Pi} B^{\pi(\theta)},, B^{\pi(\theta')}, \theta, \theta' \right\| \leq c \| \theta - \theta' \|.
\]

where the constant is later specified.

Indeed, by Lemma 6 and Lemma 3, we have that

\[
\left\| \sup_{\pi \in \Pi} B^{\pi(\theta)}, B^{\pi(\theta')}, \theta, \theta' \right\|
\]
\[ \leq \sup_{\pi \in \Pi} \left\| B^{\pi(\theta)}, \theta - B^{\pi(\theta')}, \theta' \right\| \]

\[ \begin{aligned}
&= \sup_{\pi \in \Pi} \left\| E_{s_t \sim P, \pi} \left[ \sum_{t=0}^{\infty} \sum_{i \in N} \gamma^t \int_0^1 \gamma'(z) \left( \frac{\partial R_i, \theta(s_t, a)}{\partial \pi_i}(\gamma(z)) - \frac{\partial R_i, \theta'(s_t, a)}{\partial \pi_i}(\gamma(z)) \right) \right] \right\| \\
&= \sup_{\pi \in \Pi} \left\| \sum_{t=0}^{\infty} \sum_{i \in N} \gamma^t E_{s_t \sim P, \pi} \left[ \int_0^1 \gamma'(z) \left( \frac{\partial R_i, \theta(s_t, a)}{\partial \pi_i}(\gamma(z)) - \frac{\partial R_i, \theta'(s_t, a)}{\partial \pi_i}(\gamma(z)) \right) \right] \right\| \\
&\leq \sup_{\pi \in \Pi} \left\| \sum_{t=0}^{\infty} \sum_{i \in N} \gamma^t E_{s_t \sim P, \pi} \left[ L \frac{\partial R_i, \theta}{\partial \pi_i} \int_0^1 \gamma'(z) \| \theta - \theta' \| (\gamma(z)) \right] \right\| \\
&\leq \frac{NL \frac{\partial R_i, \theta}{\partial \pi_i}}{1 - \gamma} \| \theta - \theta' \| ,
\end{aligned} \]

and hence we deduce the statement.

8 Conclusion

In this paper, we provided a first method of solving MP-NE for SGs with nonzero-sum payoffs. In particular, we identified a subset of SGs, namely stochastic potential games that can be solved tractably. The construction of this subset involved a new notion of potential games in the stochastic setting. We showed that computing the MP-NE in pure strategies is reducible to solving a cooperative game with an MDP representation. Although computing the NE for a non-zero-sum SG is in general \(PSPACE\)-hard, we have shown that in the special case of a SPG the solution is \(P\)-Complete. To find the corresponding MP-NE for the SPGs we provided two algorithms which provably converge. Finally, we further investigated the properties of the MP-NE, showing that it satisfies a continuity property under changes of the game reward functions.
References

[Sha53] Lloyd S Shapley. “Stochastic games”. In: Proceedings of the national academy of sciences 39.10 (1953), pp. 1095–1100.

[BF68] David Blackwell and Tom S Ferguson. “The big match”. In: The Annals of Mathematical Statistics 39.1 (1968), pp. 159–163.

[Hil79] Theodore Preston Hill. “On the existence of good Markov strategies”. In: Transactions of the American Mathematical Society 247 (1979), pp. 157–176.

[PT87] Christos H Papadimitriou and John N Tsitsiklis. “The complexity of Markov decision processes”. In: Mathematics of operations research 12.3 (1987), pp. 441–450.

[FT91] Drew Fudenberg and Jean Tirole. “Tirole: Game Theory”. In: MIT Press 726 (1991), p. 764.

[RF91] TES Raghavan and Jerzy A Filar. “Algorithms for stochastic games—a survey”. In: Zeitschrift für Operations Research 35.6 (1991), pp. 437–472.

[OR94] Martin J Osborne and Ariel Rubinstein. A course in game theory. MIT press, 1994.

[Sla94] Margaret E Slade. “What does an oligopoly maximize?” In: The Journal of Industrial Economics (1994), pp. 45–61.

[YB95] S. Yu and M. Berthod. “A Game Strategy Approach for Image Labeling”. In: Computer Vision and Image Understanding 61.1 (1995), pp. 32–37. ISSN: 1077-3142. DOI: https://doi.org/10.1006/cviu.1995.1003. URL: http://www.sciencedirect.com/science/article/pii/S107731428571003X.

[MS96] Dov Monderer and Lloyd S Shapley. “Potential games”. In: Games and economic behavior 14.1 (1996), pp. 87–142.

[MS96] Dov Monderer and Lloyd S Shapley. “Potential games”. In: Games and economic behavior 14.1 (1996), pp. 87–142.

[Ui00] Takashi Ui. “A Shapley value representation of potential games”. In: Games and Economic Behavior 31.1 (2000), pp. 121–135.

[San01] William H Sandholm. “Potential games with continuous player sets”. In: Journal of Economic theory 97.1 (2001), pp. 81–108.

[Rit+02] Klaus Ritzberger et al. “Foundations of non-cooperative game theory”. In: OUP Catalogue (2002).

[Von02] Bernhard Von Stengel. “Computing equilibria for two-person games”. In: Handbook of game theory with economic applications 3 (2002), pp. 1723–1759.

[WS03] Xiaofeng Wang and Tuomas Sandholm. “Reinforcement learning to play an optimal Nash equilibrium in team Markov games”. In: Advances in neural information processing systems. 2003, pp. 1603–1610.

[You04] H Peyton Young. Strategic learning and its limits. OUP Oxford, 2004.

[LC06] David S Leslie and Edmund J Collins. “Generalised weakened fictitious play”. In: Games and Economic Behavior 56.2 (2006), pp. 285–298.

[SBP06] Gesualdo Scutari, Sergio Barbarossa, and Daniel P Palomar. “Potential games: A framework for vector power control problems with coupled constraints”. In: 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings. Vol. 4. IEEE. 2006, pp. IV–IV.
[MG07] Chris Murray and Geoffrey J Gordon. “Multi-robot negotiation: approximating the set of subgame perfect equilibria in general-sum stochastic games”. In: Advances in neural information processing systems. 2007, pp. 1001–1008.

[CS08] Vincent Conitzer and Tuomas Sandholm. “New complexity results about Nash equilibria”. In: Games and Economic Behavior 63.2 (2008), pp. 621–641.

[GS09] Sam Ganzfried and Tuomas Sandholm. “Computing equilibria in multiplayer stochastic games of imperfect information”. In: Twenty-First International Joint Conference on Artificial Intelligence. 2009.

[MAS09] Jason R Marden, Gürdal Arslan, and Jeff S Shamma. “Cooperative control and potential games”. In: IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 39.6 (2009), pp. 1393–1407.

[NAB09] Kien C Nguyen, Tansu Alpcan, and Tamer Basar. “Stochastic games for security in networks with interdependent nodes”. In: 2009 International Conference on Game Theory for Networks. IEEE. 2009, pp. 697–703.

[TH09] Arne Traulsen and Christoph Hauert. “Stochastic evolutionary game dynamics”. In: Reviews of nonlinear dynamics and complexity 2 (2009), pp. 25–61.

[CG10] Sofia Ceppi and Nicola Gatti. “An algorithmic game theory framework for bilateral bargaining with uncertainty”. In: Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1. 2010, pp. 1489–1490.

[LM11] David S Leslie and Jason R Marden. “Equilibrium selection in potential games with noisy rewards”. In: International Conference on NETwork Games, Control and Optimization (NetGCooP 2011). IEEE. 2011, pp. 1–4.

[Mar12] Jason R Marden. “State based potential games”. In: Automatica 48.12 (2012), pp. 3075–3088.

[Saa+12] Walid Saad et al. “Game-theoretic methods for the smart grid: An overview of microgrid systems, demand-side management, and smart grid communications”. In: IEEE Signal Processing Magazine 29.5 (2012), pp. 86–105.

[GPS13] Paul W Goldberg, Christos H Papadimitriou, and Rahul Savani. “The complexity of the homotopy method, equilibrium selection, and Lemke-Howson solutions”. In: ACM Transactions on Economics and Computation (TEAC) 1.2 (2013), pp. 1–25.

[GH13] David González-Sánchez and Onésimo Hernández-Lerma. Discrete-time stochastic control and dynamic potential games: the Euler–Equation approach. Springer Science & Business Media, 2013.

[Zaz+15] Santiago Zazo et al. “Dynamic potential games in communications: Fundamentals and applications”. In: arXiv preprint arXiv:1509.01313 (2015).

[AY16] Gürdal Arslan and Serdar Yüksel. “Decentralized Q-learning for stochastic teams and games”. In: IEEE Transactions on Automatic Control 62.4 (2016), pp. 1545–1558.

[Bau16] Dario Bauso. Game theory with engineering applications. SIAM, 2016.

[LCS16] Quang Duy Lạ, Yong Huat Chew, and Boon-Hee Soong. Potential Game Theory. Springer, 2016.

[Tir+16] Tawin Tiratanapakhom et al. “Braess’ paradox in the uncertain demand and congestion assumed stochastic transportation network design problem”. In: KSCE Journal of Civil Engineering 20.7 (2016), pp. 2928–2937.
Appendix

Algorithm 3

1: For a given game $\mathcal{M}(s), s \in S$. Require $\phi(s, \cdot)$, initial strategy profile $\pi$, constant $\kappa$
2: $\phi(s, \cdot) \leftarrow \kappa$
3: while training do
4: Visiting all profiles in $\mathcal{M}$ starting from $\pi$
5: for $i = 0, \ldots, N$ do
6: $i \leftarrow$ deviating player
7: $\pi_a \leftarrow$ new joint strategy profile
8: $\pi_b \leftarrow$ new joint strategy profile
9: $\phi^{\pi_b}(s) \leftarrow \phi^{\pi_a}(s) + R^{\pi_b}_{i}(s) - R^{\pi_a}_{i}(s)$
10: end for
11: end while
Proofs

Lemma 6 For any \( f : X \times X \to \mathbb{R}, g : X \times X \to \mathbb{R} \), we have that:

\[
\| \max_{a \in X} f(a) - \max_{a \in X} g(a) \| \leq \max_{a \in X} \| f(a) - g(a) \|. \tag{50}
\]

Proof

\[
f(a) \leq \| f(a) - g(a) \| + g(a) \tag{51}
\]

\[
\implies \max_{a \in X} f(a) \leq \max_{a \in X} \{ \| f(a) - g(a) \| + g(a) \} \tag{52}
\]

\[
\leq \max_{a \in X} \| f(a) - g(a) \| + \max_{a \in X} \| g(a) \|. \tag{53}
\]

Deducting \( \max_{a \in X} g(a) \) from both sides of (53) yields:

\[
\max_{a \in X} f(a) - \max_{a \in X} g(a) \leq \max_{a \in X} \| f(a) - g(a) \|. \tag{54}
\]

After reversing the roles of \( f \) and \( g \) and redoing steps (51) - (53), we deduce the desired result since the RHS of (54) is unchanged.

Proof 8.1 (Proof of Prop. 3) Recall the proposition asserts the existence of a measurable function \( B : \Pi \times S \to \mathbb{R} \) s.t. the following holds for any \( i \in \mathcal{N} \)

\[
\mathbb{E}_{s \sim P(\cdot)} \left[ v_i^{(\pi_i, \pi_{-i})}(s) - v_i^{(\pi_i', \pi_{-i})}(s) \right] = \mathbb{E}_{s \sim P(\cdot)} \left[ B(\pi_i, \pi_{-i})(s) - B(\pi_i', \pi_{-i})(s) \right]. \tag{55}
\]

For the finite horizon case, the result is proven by induction on the number of time steps until the end of the game. Unlike the infinite horizon case, for the finite horizon case the value function and policy have an explicit time dependence.

We consider the case of the proposition at time \( T - 1 \) that is we evaluate the value functions at the penultimate time step. In this case, we have that:

\[
\mathbb{E}_{s_{T-1} \sim P(\cdot)} \left[ v_i^{(\pi_i, \pi_{-i})}(s_{T-1}) - v_i^{(\pi_i', \pi_{-i})}(s_{T-1}) \right]
\]

\[
= \mathbb{E}_{s_{T-1} \sim P(\cdot)} \left[ \sum_{a_{T-1}^{i} \in A_i} \sum_{a_{T-1}^{-i} \in A_{-i}} \pi_i(a_{T-1}^{i}, s_{T-1}^{-i}) \pi_{-i}(a_{T-1}^{-i}, s_{T-1}) R_i(s_{T-1}, a_{T-1}^{i}, a_{T-1}^{-i}) \right.
\]

\[
+ \gamma \sum_{s_T \in S} \sum_{a_{T-1}^{i} \in A_i} \sum_{a_{T-1}^{-i} \in A_{-i}} \pi_i(a_{T-1}^{i}, s_{T-1}^{-i}) \pi_{-i}(a_{T-1}^{-i}, s_{T-1}) P(s_T; a_{T-1}^{i}, a_{T-1}^{-i}) v_i^{(\pi_i', \pi_{-i})}(s_T) \right]
\]

\[
- \left( \sum_{a_{T-1}^{i} \in A_i} \sum_{a_{T-1}^{-i} \in A_{-i}} \pi_i(a_{T-1}^{i}, s_{T-1}^{-i}) \pi_{-i}(a_{T-1}^{-i}, s_{T-1}) R_i(s_{T-1}, a_{T-1}^{i}, a_{T-1}^{-i}) \right.
\]

\[
+ \gamma \sum_{s_T \in S} \sum_{a_{T-1}^{i} \in A_i} \sum_{a_{T-1}^{-i} \in A_{-i}} \pi_i(a_{T-1}^{i}, s_{T-1}^{-i}) \pi_{-i}(a_{T-1}^{-i}, s_{T-1}) P(s_T; a_{T-1}^{i}, a_{T-1}^{-i}) v_i^{(\pi_i', \pi_{-i})}(s_T) \right]\]
\[
\mathbb{E}_{s_{T-1} \sim P(i)} \left[ \sum_{a_{T-1}^i \in A_i, a_{T-1}^{-i} \in A_{-i}} \pi_i(a_{T-1}^i; s_{T-1}) \pi_{-i}(a_{T-1}^{-i}, s_{T-1}) \phi(s_{T-1}, a_{T-1}^i, a_{T-1}^{-i}) \right. \\
- \sum_{a_{T-1}^i \in A_i, a_{T-1}^{-i} \in A_{-i}} \sum_{a_{T-1}^{'i} \in A_i, a_{T-1}^{''-i} \in A_{-i}} \pi_i(a_{T-1}^{i'}; s_{T-1}) \pi_{-i}(a_{T-1}^{-i'}, s_{T-1}) \phi(s_{T-1}, a_{T-1}^{i'}, a_{T-1}^{-i'}) \\
+ \sum_{s_T \in S} \sum_{a_{T-1}^i \in A_i, a_{T-1}^{-i} \in A_{-i}} \pi_i(a_{T-1}^i, s_{T-1}) \pi_{-i}(a_{T-1}^{-i}, s_{T-1}) P(s_T; a_{T-1}^i, a_{T-1}^{-i}) v_i^{(\pi_i, \pi_{-i})}(s_T) \\
- \gamma \sum_{s_T \in S} \sum_{a_{T-1}^i \in A_i, a_{T-1}^{-i} \in A_{-i}} \pi_i(a_{T-1}^{i'}, s_{T-1}) \pi_{-i}(a_{T-1}^{-i'}, s_{T-1}) P(s_T; a_{T-1}^{i'}, a_{T-1}^{-i'}) v_i^{(\pi_i, \pi_{-i})}(s_T) \right] \\
= \mathbb{E}_{s_{T-1} \sim P(i)} \left[ \sum_{a_{T-1}^i \in A_i, a_{T-1}^{-i} \in A_{-i}} \pi_i(a_{T-1}^i; s_{T-1}) \pi_{-i}(a_{T-1}^{-i}, s_{T-1}) \phi(s_{T-1}, a_{T-1}^i, a_{T-1}^{-i}) \right. \\
- \sum_{a_{T-1}^i \in A_i, a_{T-1}^{-i} \in A_{-i}} \sum_{a_{T-1}^{i'} \in A_i, a_{T-1}^{''-i} \in A_{-i}} \pi_i(a_{T-1}^{i'}; s_{T-1}) \pi_{-i}(a_{T-1}^{''-i}, s_{T-1}) \phi(s_{T-1}, a_{T-1}^{i'}, a_{T-1}^{''-i}) \\
+ \sum_{s_T \in S} \sum_{a_{T-1}^i \in A_i, a_{T-1}^{-i} \in A_{-i}} \pi_i(a_{T-1}^i, s_{T-1}) \pi_{-i}(a_{T-1}^{-i}, s_{T-1}) P(s_T; a_{T-1}^i, a_{T-1}^{-i}) v_i^{(\pi_i, \pi_{-i})}(s_T) \\
- \gamma \sum_{s_T \in S} \sum_{a_{T-1}^i \in A_i, a_{T-1}^{-i} \in A_{-i}} \pi_i(a_{T-1}^{i'}, s_{T-1}) \pi_{-i}(a_{T-1}^{''-i}, s_{T-1}) P(s_T; a_{T-1}^{i'}, a_{T-1}^{''-i}) v_i^{(\pi_i, \pi_{-i})}(s_T) \right] \\
\]

We now observe that for any \( \pi_i \in \Pi_i \) and for any \( \pi_{-i} \in \Pi_{-i} \) we have that \( \forall i \in \mathcal{N}, v_i^{\pi_i, \pi_{-i}}(s_T) = \mathbb{E}_{s_{T-1} \sim P(i)} \left[ \sum_{a_i \in A_i} \sum_{a_{-i} \in A_{-i}} \pi_i(a_{s_T}^i, s_T) \pi_{-i}(a_{s_T}^{-i}, s_T) R_i(a_{s_T}^i, a_{s_T}^{-i}, s_T) \right]. \)

By Prop. 1 we have that

\[
\mathbb{E}_{s_{T-1} \sim P(i)} \left[ \sum_{a_{T-1}^i \in A_i, a_{T-1}^{-i} \in A_{-i}} \pi_i(a_{T-1}^i, s_{T-1}) \pi_{-i}(a_{T-1}^{-i}, s_{T-1}) R_i(s_T, a^i, a^{-i}) \right] \\
= \mathbb{E}_{s_{T-1} \sim P(i)} \left[ \sum_{a_{T-1}^i \in A_i, a_{T-1}^{-i} \in A_{-i}} \pi_i(a_{T-1}^i, s_{T-1}) \pi_{-i}(a_{T-1}^{-i}, s_{T-1}) R_i(s_T, a^i, a^{-i}) \right. \\
- \sum_{a_{T-1}^i \in A_i, a_{T-1}^{-i} \in A_{-i}} \sum_{a_{T-1}^{i'} \in A_i, a_{T-1}^{''-i} \in A_{-i}} \pi_i(a_{T-1}^{i'}, s_{T-1}) \pi_{-i}(a_{T-1}^{''-i}, s_{T-1}) R_i(s_T, a^{i'}, a^{''-i}) \\
= \sum_{s_T \in S} \sum_{a_{T-1}^i \in A_i, a_{T-1}^{-i} \in A_{-i}} \pi_i(a_{T-1}^i, s_{T-1}) \pi_{-i}(a_{T-1}^{-i}, s_{T-1}) P(s_T; a^i, a^{-i}) \left[ \phi(s_T, a^i, a^{-i}) + F_i(a^{-i}) \right] \\
- \gamma \sum_{s_T \in S} \sum_{a_{T-1}^i \in A_i, a_{T-1}^{-i} \in A_{-i}} \pi_i(a_{T-1}^i, s_{T-1}) \pi_{-i}(a_{T-1}^{-i}, s_{T-1}) P(s_T; a^i, a^{-i}) \left[ \phi(s_T, a^i, a^{-i}) + F_i(s_T, a^{-i}) \right] \\
= \sum_{s_T \in S} \sum_{a_{T-1}^i \in A_i, a_{T-1}^{-i} \in A_{-i}} \pi_i(a_{T-1}^i, s_{T-1}) \pi_{-i}(a_{T-1}^{-i}, s_{T-1}) P(s_T; a^i, a^{-i}) \phi(s_T, a^i, a^{-i}) \\
\]
We now show the last two summations add to 0. Indeed, we have that

\[
\sum_{s_T \in S} \sum_{a_{T-1} \in \mathbb{A}} \sum_{a_{T-1} \in \mathbb{A}} P(s_T; a_{T-1}^i, a_{T-1}^-) \cdot \left[ \pi_i(a_{T-1}^i, s_{T-1}) \pi_{-i}(a_{T-1}^-, s_{T-1}) F_i(a^{-i}) - \pi_i(a_{T-1}^i, s_{T-1}) \pi_{-i}(a_{T-1}^-, s_{T-1}) F_i(a^{-i}) \right]
\]

= \sum_{s_T \in S} \sum_{a_{T-1} \in \mathbb{A}} \sum_{a_{T-1} \in \mathbb{A}} \sum_{a_{T-1} \in \mathbb{A}} \left[ \pi_i(a_{T-1}^i, s_{T-1}) - \pi_i(a_{T-1}^i, s_{T-1}) \right] \cdot P(s_T; a_{T-1}^i, a_{T-1}^-) \pi_{-i}(a_{T-1}^-, s_{T-1}) F_i(a^{-i})

= \sum_{s_T \in S} \sum_{a_{T-1} \in \mathbb{A}} \sum_{a_{T-1} \in \mathbb{A}} \pi_{-i}(a_{T-1}^-, s_{T-1}) \left( P(s_T; a_{T-1}^i, a_{T-1}^-) - P(s_T; a_{T-1}^i, a_{T-1}^-(k+1)) \right) F_i(a^{-i})

= 0

We therefore find that

\[
\sum_{s_T \in S} \sum_{a_{T-1} \in \mathbb{A}} \sum_{a_{T-1} \in \mathbb{A}} \pi_i(a_{T-1}^i, s_{T-1}) \pi_{-i}(a_{T-1}^-, s_{T-1}) P(s_T; a^i, a^{-i}) \phi(s_T, a^i, a^{-i})
\]

\[
- \sum_{s_T \in S} \sum_{a_{T-1} \in \mathbb{A}} \sum_{a_{T-1} \in \mathbb{A}} \pi_i(a_{T-1}^i, s_{T-1}) \pi_{-i}(a_{T-1}^-, s_{T-1}) P(s_T; a^i, a^{-i}) \phi(s_T, a^i, a^{-i})
\]

\[
+ \sum_{s_T \in S} \sum_{a_{T-1} \in \mathbb{A}} \sum_{a_{T-1} \in \mathbb{A}} \pi_i(a_{T-1}^i, s_{T-1}) \pi_{-i}(a_{T-1}^-, s_{T-1}) P(s_T; a^i, a^{-i}) F_i(a^{-i})
\]

\[
- \sum_{s_T \in S} \sum_{a_{T-1} \in \mathbb{A}} \sum_{a_{T-1} \in \mathbb{A}} \pi_i(a_{T-1}^i, s_{T-1}) \pi_{-i}(a_{T-1}^-, s_{T-1}) P(s_T; a^i, a^{-i}) F_i(a^{-i})
\]

\[
= \mathbb{E}_{s_T \sim P(\cdot)} \left[ \phi^{(\pi, \pi_{-1})}(s_T) - \phi^{(\pi', \pi_{-1})}(s_T) \right]
\]

Hence substituting (79) into (66), we find that

\[
\mathbb{E}_{s_{T-1} \sim P(\cdot)} \left[ v_{i, T-1}^{(\pi_i, \pi_{-1})}(s_{T-1}) - v_{i, T-1}^{(\pi_i', \pi_{-1})}(s_{T-1}) \right]
\]

24
Now we consider the case when we evaluate the expression (83). Our inductive hypothesis is the expression holds for some \(0 < k \leq T\), that is for any \(0 < k \leq T\) we have that:

\[
\mathbb{E}_{s_{T-k} \sim P(\cdot)} \left[ v_{i,k}^{(\pi,\pi^{-i})}(s_{T-k}) - v_{i,k}^{(\pi',\pi^{-i})}(s_{T-k}) \right] = \mathbb{E}_{s_{T-k} \sim P(\cdot)} \left[ B_{T_{i,k}}^{(\pi,\pi^{-i})}(s_{T-k}) - B_{T_{i,k}}^{(\pi',\pi^{-i})}(s_{T-k}) \right].
\]  

(83)

It remains to show that the expression holds for \(k + 1\) time steps prior to the end of the horizon. The result can be obtained using the dynamic programming principle and the base case \((k = 1)\) result.

First, we note that it is easy to see that given (83) and by Prop. 1, it must be the case that:

\[
\mathbb{E}_{s_{T-k} \sim P(\cdot)} \left[ v_{i,k}^{(\pi,\pi^{-i})}(s_{T-k}) \right] = \mathbb{E}_{s_{T-k} \sim P(\cdot)} \left[ B_{T_{i,k}}^{(\pi,\pi^{-i})}(s_{T-k}) + G_{i,k}^{(\pi,\pi^{-i})}(s_{T-k}) \right].
\]  

(84)

where $G_{i,k}^{\pi-i}(s) := \mathbb{E}_{P,\pi-i} \left[ \sum_{t=0}^{k} \gamma^{t} F_{-i}(s, a_{t}^{-i}) \right]$. 

Moreover, we recall that $F_{-i}$ satisfies the condition $F_{-i}(s, a_{t}^{-i}) = F_{-i}(s', a_{t}^{-i})$, hence $G_{i,k}^{\pi-i}(s) = G_{i,k}^{\pi-i}(s')$ so from now on we drop the dependence on $s$ and write $G_{i,k}^{\pi-i}$.

We now observe that

$$
\mathcal{E}_{s_{T-k} \sim P(\cdot)} \left[ v_{i,k+1}^{(\pi_{i},\pi_{-i})} (s_{T-(k+1)}) - v_{i,k+1}^{(\pi_{i},\pi_{-i})} (s_{T-(k+1)}) \right] 
$$

$$
= \mathcal{E}_{s_{T-k} \sim P(\cdot)} \left[ \sum_{a_{T-(k+1)} \in A_{i}} \sum_{a_{T-(k+1)} \in A_{-i}} \pi_{i}(a_{T-(k+1)}, s_{T-(k+1)}) \pi_{-i}(a_{T-(k+1)}, s_{T-(k+1)}) 
\cdot R_{i}(s_{T-(k+1)}, a_{T-(k+1)}, a_{T-(k+1)}) 
\right] 
$$

$$
+ \gamma \sum_{s_{T-k} \in S} \sum_{a_{T-(k+1)} \in A_{i}} \sum_{a_{T-(k+1)} \in A_{-i}} \pi_{i}(a_{T-(k+1)}, s_{T-(k+1)}) \pi_{-i}(a_{T-(k+1)}, s_{T-(k+1)}) 
\cdot P(s_{T-k}; a_{T-(k+1)}, a_{T-(k+1)}) \nu_{i,k}^{(\pi_{i},\pi_{-i})}(s_{T-k}) 
$$

$$
- \sum_{a_{T-(k+1)} \in A_{i}} \sum_{a_{T-(k+1)} \in A_{-i}} \pi_{i}(a_{T-(k+1)}, s_{T-(k+1)}) \pi_{-i}(a_{T-(k+1)}, s_{T-(k+1)}) 
\cdot R_{i}(s_{T-(k+1)}, a_{T-(k+1)}, a_{T-(k+1)}) 
$$

$$
+ \gamma \sum_{s_{T-k} \in S} \sum_{a_{T-(k+1)} \in A_{i}} \sum_{a_{T-(k+1)} \in A_{-i}} \pi_{i}(a_{T-(k+1)}, s_{T-(k+1)}) \pi_{-i}(a_{T-(k+1)}, s_{T-(k+1)}) 
\cdot P(s_{T-k}; a_{T-(k+1)}, a_{T-(k+1)}) \nu_{i,k}^{(\pi_{i},\pi_{-i})}(s_{T-k}) 
$$

$$
= \mathcal{E}_{s_{T-k} \sim P(\cdot)} \left[ \sum_{a_{T-(k+1)} \in A_{i}} \sum_{a_{T-(k+1)} \in A_{-i}} \pi_{i}(a_{T-(k+1)}, s_{T-(k+1)}) \pi_{-i}(a_{T-(k+1)}, s_{T-(k+1)}) 
\cdot R_{i}(s_{T-(k+1)}, a_{T-(k+1)}, a_{T-(k+1)}) 
\right] 
$$

$$
+ \gamma \sum_{s_{T-k} \in S} \sum_{a_{T-(k+1)} \in A_{i}} \sum_{a_{T-(k+1)} \in A_{-i}} \pi_{i}(a_{T-(k+1)}, s_{T-(k+1)}) \pi_{-i}(a_{T-(k+1)}, s_{T-(k+1)}) 
\cdot P(s_{T-k}; a_{T-(k+1)}, a_{T-(k+1)}) \nu_{i,k}^{(\pi_{i},\pi_{-i})}(s_{T-k}) 
$$

$$
- \sum_{a_{T-(k+1)} \in A_{i}} \sum_{a_{T-(k+1)} \in A_{-i}} \pi_{i}(a_{T-(k+1)}, s_{T-(k+1)}) \pi_{-i}(a_{T-(k+1)}, s_{T-(k+1)}) 
\cdot R_{i}(s_{T-(k+1)}, a_{T-(k+1)}, a_{T-(k+1)}) 
$$

$$
\cdot \sum_{s_{T-k} \in S} P(s_{T-k}; a_{T-(k+1)}, a_{T-(k+1)}) \nu_{i,k}^{(\pi_{i},\pi_{-i})}(s_{T-k}) 
$$

26
Studying the terms under the first expression, we observe that by construction, we have that:

\[ \mathbb{E}_{s_{T-k} \sim P(-)} \left[ \sum_{a_{i,T-(k+1)} \in A_i} \sum_{a_{(-i),T-(k+1)} \in A_{-i}} \pi_i(a_{i,T-(k+1)}, s_{T-(k+1)}) \pi_{-i}(a_{(-i),T-(k+1)}, s_{T-(k+1)}) \cdot R_i(s_{T-(k+1)}, a_{i,T-(k+1)}, a_{(-i),T-(k+1)}) \right] \]

(90)

\[ = \mathbb{E}_{s_{T-k} \sim P(-)} \left[ \sum_{a_{i,T-(k+1)} \in A_i} \sum_{a_{(-i),T-(k+1)} \in A_{-i}} \pi_i(a_{i,T-(k+1)}, s_{T-(k+1)}) \pi_{-i}(a_{(-i),T-(k+1)}, s_{T-(k+1)}) \cdot \phi(s_{T-(k+1)}, a_{i,T-(k+1)}, a_{(-i),T-(k+1)}) \right] \]

\[ \cdot \phi(s_{T-(k+1)}, a_{i,T-(k+1)}, a_{(-i),T-(k+1)}) \]

(91)

\[ \cdot \phi(s_{T-(k+1)}, a_{i,T-(k+1)}, a_{(-i),T-(k+1)}) \]

(92)

We now study the terms within the second expectation.

Using (83) (i.e. the inductive hypothesis), we find that:

\[ \sum_{s_{T-k} \in S} \sum_{a_{i,T-(k+1)} \in A_i} \sum_{a_{(-i),T-(k+1)} \in A_{-i}} \sum_{\pi_i(a_{i,T-(k+1)}, s_{T-(k+1)}) \pi_{-i}(a_{(-i),T-(k+1)}, s_{T-(k+1)}) \cdot v_{i,k}^{(\pi_i, \pi_{-i})}(s_{T-k})} \]

\[ = \sum_{s_{T-k} \in S} \sum_{a_{i,T-(k+1)} \in A_i} \sum_{a_{(-i),T-(k+1)} \in A_{-i}} \sum_{\pi_i(a_{i,T-(k+1)}, s_{T-(k+1)}) \pi_{-i}(a_{(-i),T-(k+1)}, s_{T-(k+1)})} \cdot v_{i,k}^{(\pi_i, \pi_{-i})}(s_{T-k}) \]

\[ = \sum_{s_{T-k} \in S} \sum_{a_{i,T-(k+1)} \in A_i} \sum_{a_{(-i),T-(k+1)} \in A_{-i}} \sum_{\pi_i(a_{i,T-(k+1)}, s_{T-(k+1)}) \pi_{-i}(a_{(-i),T-(k+1)}, s_{T-(k+1)})} \cdot v_{i,k}^{(\pi_i, \pi_{-i})}(s_{T-k}) \]

\[ = \sum_{s_{T-k} \in S} \sum_{a_{i,T-(k+1)} \in A_i} \sum_{a_{(-i),T-(k+1)} \in A_{-i}} \sum_{\pi_i(a_{i,T-(k+1)}, s_{T-(k+1)}) \pi_{-i}(a_{(-i),T-(k+1)}, s_{T-(k+1)})} \cdot v_{i,k}^{(\pi_i, \pi_{-i})}(s_{T-k}) \]

\[ = \sum_{s_{T-k} \in S} \sum_{a_{i,T-(k+1)} \in A_i} \sum_{a_{(-i),T-(k+1)} \in A_{-i}} \sum_{\pi_i(a_{i,T-(k+1)}, s_{T-(k+1)}) \pi_{-i}(a_{(-i),T-(k+1)}, s_{T-(k+1)})} \cdot v_{i,k}^{(\pi_i, \pi_{-i})}(s_{T-k}) \]

\[ = \sum_{s_{T-k} \in S} \sum_{a_{i,T-(k+1)} \in A_i} \sum_{a_{(-i),T-(k+1)} \in A_{-i}} \sum_{\pi_i(a_{i,T-(k+1)}, s_{T-(k+1)}) \pi_{-i}(a_{(-i),T-(k+1)}, s_{T-(k+1)})} \cdot v_{i,k}^{(\pi_i, \pi_{-i})}(s_{T-k}) \]

\[ = \sum_{s_{T-k} \in S} \sum_{a_{i,T-(k+1)} \in A_i} \sum_{a_{(-i),T-(k+1)} \in A_{-i}} \sum_{\pi_i(a_{i,T-(k+1)}, s_{T-(k+1)}) \pi_{-i}(a_{(-i),T-(k+1)}, s_{T-(k+1)})} \cdot v_{i,k}^{(\pi_i, \pi_{-i})}(s_{T-k}) \]

\[ = \sum_{s_{T-k} \in S} \sum_{a_{i,T-(k+1)} \in A_i} \sum_{a_{(-i),T-(k+1)} \in A_{-i}} \sum_{\pi_i(a_{i,T-(k+1)}, s_{T-(k+1)}) \pi_{-i}(a_{(-i),T-(k+1)}, s_{T-(k+1)})} \cdot v_{i,k}^{(\pi_i, \pi_{-i})}(s_{T-k}) \]
\[ + \sum_{s_{T-k} \in S} \sum_{a_{T-(k+1)} \in A} \sum_{a_{T-(k+1)} \in A^{-i}} P(S_{T-k}; a_{T-(k+1)}^i, a_{T-(k+1)}^{-i}) \]
\[ \cdot \left[ \pi_i^i(a_{T-(k+1)}^i, s_{T-(k+1)}) \pi_{-i}(a_{T-(k+1)}^{-i}, s_{T-(k+1)}) G_{i,k}^{\pi_{-i}} \right. \]
\[ \left. - \pi_i^i(a_{T-(k+1)}^i, s_{T-(k+1)}) \pi_{-i}(a_{T-(k+1)}^{-i}, s_{T-(k+1)}) G_{i,k}^{\pi_{-i}} \right] \]
\[
Now
\[
\sum_{s_{T-k} \in S} \sum_{a_{T-(k+1)} \in A} \sum_{a_{T-(k+1)} \in A^{-i}} P(S_{T-k}; a_{T-(k+1)}^i, a_{T-(k+1)}^{-i}) \]
\[ \cdot \left[ \pi_i^i(a_{T-(k+1)}^i, s_{T-(k+1)}) \pi_{-i}(a_{T-(k+1)}^{-i}, s_{T-(k+1)}) G_{i,k}^{\pi_{-i}} \right. \]
\[ \left. - \pi_i^i(a_{T-(k+1)}^i, s_{T-(k+1)}) \pi_{-i}(a_{T-(k+1)}^{-i}, s_{T-(k+1)}) G_{i,k}^{\pi_{-i}} \right] \]
\[ = \sum_{s_{T-k} \in S} \sum_{a_{T-(k+1)} \in A} \sum_{a_{T-(k+1)} \in A^{-i}} \left[ \pi_i^i(a_{T-(k+1)}^i, s_{T-(k+1)}) - \pi_i^i(a_{T-(k+1)}^i, s_{T-(k+1)}) \right] \]
\[ \cdot \left( P(S_{T-k}; \pi_i, a_{T-(k+1)}^{-i}) - P(S_{T-k}; \pi_i^i, a_{T-(k+1)}^{-i}) \right) G_{i,k}^{\pi_{-i}} \]
\[ = 0 \]
\[ \]
We therefore find that:
\[
\sum_{s_{T-k} \in S} \sum_{a_{T-(k+1)} \in A} \sum_{a_{T-(k+1)} \in A^{-i}} P(s; a_{T-(k+1)}^i, a_{T-(k+1)}^{-i}) \]
\[ \cdot \left[ \pi_i^i(a_{T-(k+1)}^i, s_{T-(k+1)}) \pi_{-i}(a_{T-(k+1)}^{-i}, s_{T-(k+1)}) v_{i,k}^{(\pi_{-i})} (s_{T-k}) \right. \]
\[ \left. - \pi_i^i(a_{T-(k+1)}^i, s_{T-(k+1)}) \pi_{-i}(a_{T-(k+1)}^{-i}, s_{T-(k+1)}) v_{i,k}^{(\pi_{-i})} (s_{T-k}) \right] \]
\[ = \sum_{s_{T-k} \in S} \sum_{a_{T-(k+1)} \in A} \sum_{a_{T-(k+1)} \in A^{-i}} P(s; a_{T-(k+1)}^i, a_{T-(k+1)}^{-i}) \]
\[ \cdot \left[ \pi_i^i(a_{T-(k+1)}^i, s_{T-(k+1)}) \pi_{-i}(a_{T-(k+1)}^{-i}, s_{T-(k+1)}) B_{k}^{\pi_{-i}} (s_{T-k}) \right. \]
\[ \left. - \pi_i^i(a_{T-(k+1)}^i, s_{T-(k+1)}) \pi_{-i}(a_{T-(k+1)}^{-i}, s_{T-(k+1)}) B_{k}^{\pi_{-i}} (s_{T-k}) \right) \]
\[ \text{using (83) (i.e. the inductive hypothesis).} \]
Now combining (92) and (99) leads to the fact that:

$$
\mathbb{E}_{s_{T-k} \sim P(\cdot)} \left[ v_{i,k+1}^{(\pi_i, \pi_i^{-1})}(s_{T-(k+1)}) - v_{i,k+1}^{(\pi_i', \pi_i^{-1})}(s_{T-(k+1)}) \right]
= \sum_{s_{T-k} \in S} \sum_{a_{T-(k+1)}' \in A_i} \sum_{a_{T-(k+1)}^{-i} \in A_{-i}} P(s_{T-k}; a_{T-(k+1)}', a_{T-(k+1)}^{-i})
\cdot \pi_i(a_{T-(k+1)}', s_{T-(k+1)}) \pi_i^{-1}(a_{T-(k+1)}^{-i}, s_{T-(k+1)}) B_k^{\pi_i, \pi_i^{-1}}(s_{T-k})
- \pi_i'(a_{T-(k+1)}', s_{T-(k+1)}) \pi_i^{-1}(a_{T-(k+1)}^{-i}, s_{T-(k+1)}) B_k^{\pi_i', \pi_i^{-1}}(s_{T-k})
$$

(100)

$$
+ \mathbb{E}_{s_{T-(k+1)} \sim P(\cdot)} \left[ \sum_{a_{T-(k+1)}' \in A_i} \pi_i'(a_{T-(k+1)}', s_{T-(k+1)}) \pi_i^{-1}(a_{T-(k+1)}^{-i}, s_{T-(k+1)}) \cdot \phi(s_{T-(k+1)}; a_{T-(k+1)}', a_{T-(k+1)}^{-i})
- \sum_{a_{T-(k+1)}' \in A_i} \sum_{a_{T-(k+1)}^{-i} \in A_{-i}} \pi_i'(a_{T-(k+1)}', s_{T-(k+1)}) \pi_i^{-1}(a_{T-(k+1)}^{-i}, s_{T-(k+1)}) \cdot \phi(s_{T-(k+1)}; a_{T-(k+1)}', a_{T-(k+1)}^{-i}) \right],
$$

which immediately suggests that

$$
\mathbb{E}_{s_{T-(k+1)} \sim P(\cdot)} \left[ v_{i,k+1}^{(\pi_i, \pi_i^{-1})}(s_{T-(k+1)}) - v_{i,k+1}^{(\pi_i', \pi_i^{-1})}(s_{T-(k+1)}) \right]
= \mathbb{E}_{s_{T-(k+1)} \sim P(\cdot)} \left[ B_k^{(\pi_i, \pi_i^{-1})}(s_{T-(k+1)}) - B_{k+1}^{(\pi_i', \pi_i^{-1})}(s_{T-(k+1)}) \right],
$$

(101)

where

$$
B_k^{(\pi_i, \pi_i^{-1})}(s) = \mathbb{E}_{\pi_i, \pi_i^{-1}} \left[ \phi(s_{k}, a_{k}', a_{k}^{-i}) + \gamma \sum_{s' \in S} P(s' ; s, a_{k}', a_{k}^{-i}) B_{k-1}^{\pi_i, \pi_i^{-1}}(s') \right],
$$

(102)

from which we deduce the result.
This figure "commodity_routing.png" is available in "png" format from:

http://arxiv.org/ps/2005.13527v1
This figure "commodity_routin2.png" is available in "png" format from:

http://arxiv.org/ps/2005.13527v1