Algorithms for Nash Equilibria in General-Sum Stochastic Games

Prasad H.L. ∗1, Prashanth L A ‡2 and Shalabh Bhatnagar ‡1

1Department of Computer Science and Automation, Indian Institute of Science, Bangalore, INDIA
2INRIA Lille - Nord Europe, Team SequeL, FRANCE.

Abstract

Over the past few decades the quest for algorithms to compute Nash equilibria in general-sum stochastic games has intensified and several important algorithms (cf. [9], [12], [16], [7]) have been proposed. However, they suffer from either lack of generality or are intractable for even medium sized problems or both. In this paper, we first formulate a non-linear optimization problem for stochastic games and then break it down into simpler sub-problems that ensure there is no Bellman error for a given state and agent. Next, we derive a set of novel necessary and sufficient conditions for solution points of these sub-problems to be Nash equilibria of the underlying game. Using these conditions, we develop two novel algorithms - OFF-SGSP and ON-SGSP, respectively. OFF-SGSP is an off-line centralized algorithm which assumes complete information of the game. On the other hand, ON-SGSP is an online decentralized algorithm that works with simulated transitions of the stochastic game. Both algorithms are guaranteed to converge to Nash equilibrium strategies for general-sum (discounted) stochastic games.

Keywords: General-Sum Discounted Stochastic Games, Nash Equilibrium, Stochastic Algorithms, On-line learning, Decentralized learning

1 Introduction

Game theoretic approaches have been useful for analyzing multi-agent scenarios. Since the seminal work of Shapley [22], stochastic games have been an important class of models for multi-agent systems, with important applications in oligopolistic economics. In the past, zero-sum stochastic games have been modelled and solved for Nash equilibria using the standard techniques of Markov decision processes. General-sum stochastic games, on the contrary, have posed a difficulty and the quest for algorithms to compute Nash equilibria in this setting has intensified over the past few decades. Some important algorithms in this direction are stochastic tracing procedure [9], NashQ [12], FFQ [16], and their generalised representations such as the optimization problem formulations for various reward structures [7]. However, these suffer either from lack of generality or are intractable for even medium sized problems or both.

In our quest for deriving algorithms for stochastic games, we start with a non-linear optimization problem and then provide a characterization of those solution points which correspond to Nash equilibria of the underlying game. As a result, we also derive a set of necessary and sufficient conditions, henceforth referred to as SG-SP (Stochastic Game - Sub-Problem) conditions. These are obtained by breaking down the main optimization problem into several sub-problems. Each sub-problem can be seen as ensuring no Bellman error, for a particular choice of the state \( x \in S \) and agent \( i \in \{1, \ldots, N\} \), where \( S \) is the state space and \( N \) is the number of agents.

---

∗hlprasu@csa.iisc.ernet.in
†prashanth.la@inria.fr
‡shalabh@csa.iisc.ernet.in
of the stochastic game considered. We also establish the kinship of these conditions to KKT conditions for the sub-problems.

Using SG-SP conditions, we propose two algorithms, OFF-SGSP and ON-SGSP, respectively. Here OFF-SGSP is an off-line iterative algorithm that works in a centralized setting and assumes complete information of the game in order to compute a Nash equilibrium strategy. On the other hand, ON-SGSP is an on-line decentralized algorithm which does not require any knowledge of the structure of the state transition or the reward schemes. Here the system model can be described as in Figure 1. The learning is localized to each agent with one instance of ON-SGSP running in each agent. Each agent presents its action to the environment and gets its reward and the next state. We provide a formal proof of convergence that both our algorithms converge to Nash equilibrium strategies, in the supplementary material.

Related Work The field of stochastic games has been actively pursued over the last seven decades with several important applications in economics. Many problems like Fishery games, Advertisement games and several oligopolistic situations can be modelled as stochastic games [7, 18, 19, 20, 2]. A comprehensive treatment of stochastic games under various pay-off criteria is given by Filar and Vrieze [7]. Singh et al. [23] observed that in a two-agent iterated general-sum game, Nash convergence is assured either in strategies or in the very least in average payoffs. Later by Hu and Wellman [11], stochastic game framework was seen as an extension of the well studied Markov decision theory [3] to multi-agent scenarios. An interesting Q-learning algorithm based on reinforcement learning (RL) [4, 24], has been proposed by Hu and Wellman [11] for general-sum games. An extension, NashQ, to the above algorithm was proposed by Hu and Wellman [12] which showed improvement in performance. However, the algorithm of Hu and Wellman [11] as well as NashQ assure convergence only if the game has exactly one Nash equilibrium. Litman [16] proposed a friend-or-foe Q-learning (FFQ) algorithm as an improvement over the NashQ algorithm with assured convergence, though not necessarily to a Nash equilibrium. Moreover, the FFQ algorithm is applicable to a restricted class of games where either full co-operation between agents is ensured or the game is zero-sum. Zinkevich et al. [26] show that the traditional Q-learning approach need not be sufficient to compute a Nash equilibrium for a typical general-sum stochastic game. Mac Dermed and Isbell [17] formulate intermediate optimization problems, called Multi-Objective Linear Programs (MOLPs), to compute Nash equilibria as well as Pareto optimal solutions. However, their algorithm is tractable only for small sized problems. Herings and Peeters [9] propose an algorithm, where a homotopic path between equilibrium points of \( N \) independent MDPs and the \( N \) player stochastic game in question, is traced numerically. This, in turn, gives a Nash equilibrium point of the stochastic game of interest. However, this approach is off-line with a computational complexity that is exponential in \( N \).

A popular algorithm with guaranteed convergence to Nash equilibrium in general-sum stochastic games is rational learning, proposed by [13]. In their algorithm, each agent maintains a prior on what he believes to be other agents’ strategy and updates it in a Bayesian manner. Combining this with certain assumptions of absolute continuity and grain of truth, the algorithm there is shown to converge to Nash equilibrium. Our decentralized algorithm, ON-SGSP, operates in a similar setting as that considered by [13], with the exception that the knowledge of reward functions is not assumed. The strategy of any agent in ON-SGSP does not depend upon Bayesian estimates of other agents’ strategies and hence, the absolute continuity/grain of truth assumptions from [13] do not apply. Instead, adopting the ordinary differential equations (ODEs) approach, we provide a novel proof of (asymptotic) convergence of our algorithms to Nash equilibrium.

A more recent approach for the computation of Nash equilibria is given by Akchurina [1]. The approach adopted there is to solve a system of ODEs using numerical methods. This is unlike some of the multi-agent RL type algorithms mentioned above, which use dynamic programming principles in arriving at a Nash strategy. Further, a system of ODEs given by Akchurina [1], also found in [25, pp. 189], turns out to be similar to a portion of the system of ODEs that are tracked by our algorithms. Though their experiments do show convergence in a large group of randomly generated games, unlike us, a formal proof of convergence is not provided.

2 Formal Definitions

A stochastic game can be seen to be an extension of the single-agent Markov decision process. A discounted reward stochastic game is described by a tuple \( < N, S, A, p, r, \beta > \), where \( N \) represents the number of agents, \( S \) denotes
the state space, $A^i(x)$ denotes the action space for the $i^{th}$ agent, $i = 1, 2, \ldots, N$ and $A = \bigcup_{x \in S} A(x)$ is the aggregate action space, where $A(x) = \times_{i=1}^{N} A^i(x)$ is the Cartesian product of action spaces of individual agents when the state of the game is $x \in S$. Further, $p(y|x, a)$ denotes the probability of going from the current state $x \in S$ to $y \in S$ when the vector of actions $a \in A(x)$ (of the $N$ players) is chosen and $r(x, a) = \langle r^i(x, a) : i = 1, 2, \ldots, N \rangle$ denotes the vector of reward functions of all agents when the state is $x \in S$ and the vector of actions $a \in A(x)$ is chosen. Also, $0 < \beta < 1$ denotes the discount factor.

We assume that the game is in normal form, i.e., in any given state $x \in S$, all agents act simultaneously with an action vector $a \in A(x)$ resulting in transition to the next state $y \in S$ according to the transition probability $p(y|x, a)$. In other words, no agent gets to know what the other agents’ actions are before selecting its own action.

A strategy $\pi^i_n$ of the $i^{th}$ player prescribes the action to be performed in each state at the $n^{th}$ time instant by that player. We denote by $\pi^i_n(\cdot)$ the action prescribed for the $i^{th}$ agent by the strategy $\pi^i_n$ at the $n^{th}$ time instant. The quantity `~' in $\pi^i_n(\cdot)$, in general, corresponds to the entire history of states and actions of all agents up to the $(n - 1)^{th}$ instant and the current system state at the $n^{th}$ instant. Let the set of all possible strategies for the $i^{th}$ player be denoted by $\mathcal{F}^i$. A strategy $\pi^i_n$ is said to be a Markov strategy if $\pi^i_n$ depends only on the current state $x_n \in S$ at time $n$, and may in general vary with $n$ (i.e., could prescribe a different action, for the same state, at different times). Thus, for a Markov strategy, $\pi^i_n(x) \in A^i(x)$, $\forall n \geq 0, x \in S, i = 1, 2, \ldots, N$. If the action chosen in any state for a Markov strategy $\pi^i_n$ is independent of the time instant $n$, viz., $\pi^i_n \equiv \pi^i$, $\forall n \geq 0, i = 1, 2, \ldots, N$, then the strategy is said be stationary.

Let $\mathcal{P}(A(x))$ (resp. $\mathcal{P}(A^i(x))$) denote the set of all probability measures on $A(x)$ (resp. $A^i(x)$). A randomized Markov strategy is specified via the sequence of maps $\phi^i_n : S \rightarrow \mathcal{P}(A^i(x))$, $x \in S$, $n \geq 0$, $i = 1, 2, \ldots, N$, $n$. Thus, $\phi^i_n(x)$ is a distribution on the set of actions $A^i(x)$ and in general depends on $n$. We say that $\phi^i_n$ is a stationary randomized strategy or simply a randomized strategy for player $i$ if $\phi^i_n \equiv \phi^i$. By an abuse of notation, we denote by $\pi \equiv (\pi^1, \pi^2, \ldots, \pi^N)$, a stationary randomized strategy that we also (many times) call a strategy, since from now on, we shall only work with randomized strategies.

**Notation.** $\langle \cdots \rangle$ represents a column vector and $\mathbb{1}_m$ is a vector of ones with $m$ elements. The various constituents of the stochastic game considered are denoted as follows: 1. **Action:** $a = \langle a^1, a^2, \ldots, a^N \rangle \in A(x)$ is the aggregate action, $a^{-i}$ is the action vector of all agents except agent $i$ and $A^{-i}(x) = \prod_{j \neq i} A^j(x)$ is the action space in state $x \in S$ of all agents except agent $i$. 2. **Policy:** $\pi^i(x, a^i)$ is the probability of picking action $a^i \in A^i(x)$ by agent $i$ in state $x \in S$, $\pi^i(x) = \langle \pi^i(x, a^i) : a^i \in A^i(x) \rangle$ is the randomized policy vector in state $x \in S$ for the agent $i$, $\pi^i = \langle \pi^i(x) : x \in S \rangle$, $\pi = \langle \pi^i : i = 1, 2, \ldots, N \rangle$ is the strategy-tuple and $\pi^{-i} = \langle \pi^j : j = 1, 2, \ldots, N, j \neq i \rangle$ is the strategy-tuple of all agents except agent $i$. 3. **Transition Probability:** $p(y|x, \pi) = \sum_{a \in A(x)} \left( p(y|x, a) \prod_{i=1}^{N} \pi^i(x, a^i) \right)$ represents the Markov transition probability from state $x \in S$ to state

![Figure 1: ON-SGSP's on-line learning model with $N$ agents](image-url)
3 Problem Formulation

Here we generalize the non-linear optimization problem formulated in Filar and Vrieze [7] for two agents, whose solution gives Nash equilibria of the underlying game. Next, we form sub-problems from the main optimization problem and derive a set of necessary and sufficient conditions that corresponding to Nash equilibria of the underlying stochastic game.

An Optimization Problem: The objective is to obtain the value function for all agents such that in any state \( x \), the value of each agent is equal to the maximum of the sum of single stage reward and the discounted expected value of the next state, averaged over the other agents’ strategies. The latter operation, i.e., of averaging over the other agents’ strategies, is done naturally because each agent \( i \) selects actions there using its randomized strategy \( \pi^i \), \( i = 1, \ldots, N \). Thus, given the knowledge of the other agents’ randomized strategies, agent \( i \) would select a strategy to maximize its expected pay-off. Let \( f(\mathbf{v},\pi) := \sum_{i=1}^{N} \sum_{x \in \mathcal{S}} [v^i(x) - r^i(x,\pi) - \beta \sum_{y \in \mathcal{U}(x)} p(y|x,\pi) v^i(y)] \).

Then, the optimization problem is given by

\[
\begin{align*}
\min_{\mathbf{v},\pi} f(\mathbf{v},\pi) \\
\text{s.t. } & \forall x \in \mathcal{S}, i = 1, 2, \ldots, N, \\
& \bar{p}^i(x,\pi^{-i},a^i) + \beta \sum_{y \in \mathcal{U}(x)} \bar{p}(y|x,\pi^{-i},a^i)v^i(y) \leq v^i(x), \forall a^i \in \mathcal{A}^i(x), \\
& \sum_{a^i \in \mathcal{A}^i(x)} \pi^i(x,a^i) = 1, \\
& \pi^i(x,a^i) \geq 0, \forall a^i \in \mathcal{A}^i(x).
\end{align*}
\]
In the above, $U(x) \subseteq S$ denotes the set of all possible next states when the current state is $x \in S$, $v^i(x)$ denotes the value of state $x \in S$ for agent $i$, $v^i = \{v^i(y) : y \in S\}$ denotes the value vector for agent $i$ (over all states) and $v = \{v^i : i = 1, 2, \ldots, N\}$ the vector of value vectors of all agents. Prasad and Bhatnagar [21] consider a similar optimization problem as (1) and derive a set of verifiable necessary and sufficient conditions that they call KKT-SP conditions (see Appendix A for a description). We however derive an alternative easier set of conditions that we call SG-SP (Stochastic Game - Sub-Problem) conditions. SG-SP conditions will also be shown to be necessary and sufficient for Nash equilibrium of the underlying stochastic game.

**Sub-problems of (1):** We form sub-problems from the main optimization problem (1) along the lines of [21], for each state $x \in S$ and each agent $i \in \{1, 2, \ldots, N\}$. The sub-problems are formed with the objective of ensuring that there is no Bellman error (see $g^i_{x,z}(\theta)$ below). Let $m = |A^i(x)|$. Then, the sub-problems are formulated as follows: For all $x \in S$ and $i \in \{1, 2, \ldots, N\}$

\[
\begin{align*}
\min_{\theta, p} h_z(\theta, p) &= \sum_{z=1}^m p_z \left[-g^i_{x,z}(\theta)\right] \\
\text{s.t.} & \quad g^i_{x,z}(\theta) \leq 0, \quad z = 1, 2, \ldots, m, \\
& \quad -p_z \leq 0, \quad z = 1, 2, \ldots, m,
\end{align*}
\]

(2)

where,

\[
\begin{align*}
\theta &= \{v^i, \pi^{-i}(x)\}, z \in \{1, 2, \ldots, m\}, \\
\theta^i_{x,z}(\theta) &= \pi^i(x, \pi^{-i}, a^i_z) + \beta \sum_{y \in U(x)} \pi^i(y|x, \pi^{-i}, a^i_z)v^i(y) - v^i(x), \\
p_z &= \pi^i(x, a^i_z), z = 1, 2, \ldots, m, \\
p &= \{p_z : z = 1, 2, \ldots, m\}.
\end{align*}
\]

The Lagrangian corresponding to (2) can be written as

\[
k(\theta, p, \lambda, \delta, s, t) = h_z(\theta, p) + \sum_{z=1}^m \lambda_z \left( g^i_{x,z}(\theta) + s^2_z \right) + \sum_{z=1}^m \delta_z \left( -p_z + t^2_z \right),
\]

(3)

where $\lambda_z$ and $\delta_z$ are the Lagrange multipliers and $s_z$ and $t_z, z = 1, 2, \ldots, m$ are the slack variables, corresponding to the first and second constraint of the sub-problem (2), respectively. While Prasad and Bhatnagar [21] proposed certain KKT-SP conditions for $N = 2$, we generalize them next and also derive a simpler set of SG-SP conditions in our setting.

**SG-SP conditions:**

**Definition 1 (SG-SP Point).** A point $(v^*, \pi^*)$ of the optimization problem (1) is said to be an SG-SP point, if it is a feasible point of (1) and for every sub-problem, i.e., for all $x \in S$ and $i \in \{1, 2, \ldots, N\}$

\[
p_z^i g^i_{x,z}(\theta^*) = 0, \quad \forall z = 1, 2, \ldots, m.
\]

(4)

The above conditions, which define a point to be an SG-SP point, are called SG-SP conditions. The connection between SG-SP points and Nash equilibria of the underlying game can be seen intuitively as follows: Recall that the objective of the sub-problem is to ensure that there is no Bellman error, which in turn implies that the value estimates $(v)$ are correct with respect to the policy $(\pi)$ of all agents. The feasibility requirement above then ensures that players rationality constraints are met (no player wishes to change policies). In this light, the main result below states that the pair of value-function and the policy tuple $(v^*, \pi^*)$ is Nash if and only if the value-function is correct and the joint-policy is an equilibrium with respect to the value function.
Proposition 3. Then for each descent direction in tuple. Combining the above with extension of value iteration \([3, 24]\) type schemes for on-policy computation of game. Further, it operates in a centralized setting.

As a consequence of the above, a simpler set of necessary and sufficient conditions for Nash equilibrium can be used as compared to KKT-SP conditions proposed by \([21]\) (for \(N = 2\)). Before proceeding onto the algorithms, we would like to remark here that, unlike KKT-SP points, SG-SP points do not impose a linear independence requirement to ensure they are Nash points (see Appendix \([A]\) for a detailed discussion).

4 An offline algorithm (OFF-SGSP) for Nash equilibria

The analysis of the previous section suggests that it is sufficient to find an SG-SP point in order to compute Nash strategies of the underlying stochastic game. We now derive a descent direction for both our algorithms using this fact about SG-SP points.

For every sub-problem, substitute and eliminate \(x_z = p_z\) and \(\delta_z = g_{x,z}^i(\theta), z = 1, 2, \ldots, m\). Then, complementary slackness conditions for optimality in the sub-problems reduce to SG-SP conditions. Further, differentiating the Lagrangian \([3]\) of the sub-problem w.r.t. the slack variable \(t_z\) and equating it to zero, we obtain

\[
\frac{\partial k}{\partial t_z} = 2t_z\delta_z = 2t_z[-g_{x,z}^i(\theta)] = 0, \quad z = 1, 2, \ldots, m.
\]

Now, \(t_z^2\) being a slack variable to \(-p_z \leq 0\), we have \(-p_z + t_z^2 = 0\) or \(t_z = \sqrt{-p_z}\). Note that \(\frac{\partial k}{\partial t_z}\) is the steepest ascent direction in the value of \(k\) along \(t_z\). Since \(t_z = \sqrt{-p_z}\), an ascent direction in \(t_z\) is also an ascent direction in \(p_z\). Thus, \(\frac{\partial k}{\partial t_z} = 2t_z[-g_{x,z}^i(\theta)]\), or simply \(t_z[-g_{x,z}^i(\theta)]\), is an ascent direction in \(p_z\) along the Lagrangian \(k\). Since our aim is to minimize the objective, we need to descend in \(p_z\). Thus, \(\sqrt{-p_z}g_{x,z}^i(\theta)\) is a possible candidate for descent direction in \(p_z\) along the Lagrangian \(k\). The following proposition makes this claim precise.

Proposition 3. Suppose the value \(\pi^i(x)\) is maintained as a probability over \(A^i(x)\) for all \(x \in S, i = 1, 2, \ldots, N\). Then for each \(i = 1, 2, \ldots, N, x \in S, a^i \in A^i(x)\), we have that \(\sqrt{\pi^i(x, a^i)}g_{x,a,i}^i(\pi^i, \pi^{-i})\) is a non-ascent, and in particular a descent direction if \(\sqrt{\pi^i(x, a^i)}g_{x,a,i}^i(\pi^i, \pi^{-i}) \neq 0\), in the objective of the main optimization problem \([\text{I}]\). Here \(g_{x,a,i}^i(\pi^i, \pi^{-i}) \leq 0\) represents the first constraint in \([\text{I}]\).

Later it will be shown that the direction suggested in Proposition 3 is the key to convergence to a Nash strategy-tuple. Combining the above with extension of value iteration \([3, 24]\) type schemes for on-policy computation of \(v\), we propose first an off-line algorithm, that we call OFF-SGSP, for the computation of an SG-SP point. This algorithm assumes the knowledge of the transition dynamics \(p\) and reward structure \(r\) of the underlying stochastic game. Further, it operates in a centralized setting.

Algorithm 1 OFF-SGSP

| Input: | The underlying general-sum stochastic game \((N, S, A, p, r, \beta)\), step-sizes \(\{b(n), c(n)\}_{n \geq 1}\), initial point \(\theta_0 = (v_0, \pi_0)\), \(\epsilon > 0\) and \(\delta > 0\) |
|--------|---------------------------------------------------------------------------------------------------|
| Output: | \((v^\ast, \pi^\ast)\): An \(\epsilon\)-Nash equilibrium with \(\epsilon \leq f(v^\ast, \pi^\ast)\) |
| Loop:  | \(n \leftarrow 1\), the iteration index |
| Loop:  | for \(x \in S\) do |
| Loop:  | for \(i = 1\) to \(N\) do |
| Loop:  | if \(\text{mod}(n, Q) = 0\) then |
| Loop:  | \(\hat{\pi}_n^i(x) \leftarrow \text{perturb}(\pi_n^i(x), \delta)\) |
| Loop:  | else |
| Loop:  | \(\hat{\pi}_n^i(x) \leftarrow \pi_n^i(x)\) |
| Loop:  | end if |
| Loop:  | end for |
| Loop:  | end for |
end for
for \( i = 1 \) to \( N \) do
  for \( a^i \in \mathcal{A}^i(x) \) do
    \[
    \pi^i_{n+1}(x, a^i) = \Gamma(\pi^i_n(x, a^i) + b(n)\sqrt{\pi^i_n(x, a^i)}g^i_{x,a^i}(v^i_n, \hat{\pi}^{-1}_n))
    \]
  end for
  \[
  v^i_{n+1}(x) = v^i_n(x) + c(n) \sum_{a^i \in \mathcal{A}^i(x)} \hat{\pi}^i_n(x, a^i)g^i_{x,a^i}(v^i_n, \hat{\pi}^{-1}_n)
  \]
end for
if \( f(v_{n+1}, \pi_{n+1}) \leq \epsilon (1 - \beta) \) then
  \((\mathbf{v}^*, \pi^*) \leftarrow (v_{n+1}, \pi_{n+1})\)
  Terminate algorithm.
end if
\( n \leftarrow n + 1 \)
end loop

For every \( Q > 0 \) iterations, \( \text{perturb}(\cdot, \delta) \) is used to derive a \( \delta \)-offset policy for picking actions, i.e., \( \hat{\pi}^i(x) \) is used instead of \( \pi^i(x) \), where

\[
\hat{\pi}^i(x, a^i) = \frac{\pi(x, a^i) + \delta}{\sum_{a^i \in \mathcal{A}^i(x)} (\pi(x, a^i) + \delta)}, a^i \in \mathcal{A}^i(x).
\]

The perturbation of the policy is performed every \( Q \) iterations in order to push the policy \( \pi \) out of the domain of attraction of any local equilibrium.

The step-sizes \( \{b(n)\}, \{c(n)\} \) satisfy

\[
\sum_{n=1}^{\infty} b(n) = c(n) = \infty, \sum_{n=1}^{\infty} (b^2(n) + c^2(n)) < \infty.
\]

Further, \( \frac{b(n)}{c(n)} \to 0 \) as \( n \to \infty \). These requirements on the step-sizes are standard stochastic approximation conditions and in particular, ensure that the \( \pi \)-recursion proceeds on a slower timescale in comparison to the \( v \)-recursion.

5 An online (decentralized) algorithm (ON-SGSP) for Nash equilibria

Though OFF-SGSP is suitable for only off-line learning of Nash strategies, it is amenable for extension to the general on-line setting where neither the transition probability \( p \) nor the reward function \( r \) are explicitly known. ON-SGSP uses the stochastic game as a generative model and is a stochastic approximation algorithm that is in the spirit of learning approaches for MDPs. The value \( v \) update can be seen to be similar to an on-policy value iteration used in MDPs where the reward and transition probability structures are not known. However, the novelty lies in its extension to the multi-agent setting with policy \( \pi \) update scheme shown to converge to Nash strategies for the underlying general-sum stochastic game.

Algorithm 2 ON-SGSP

Input: the starting state \( x_0 \), initial point \( \theta^i_0 = (v^i_0, \pi^i_0) \), step-sizes \( \{b(n), c(n)\}_{n \geq 1} \), number of iterations to run \( M \gg 0 \) and \( \text{pickAtRandom}(\cdot) \), for picking an action according to randomized policy

Output: \((v^*, \pi^*)\): Terminal \((v^i, \pi^i)\) obtained after \( M \) iterations of the algorithm.

\( n \leftarrow 1 \), the iteration index
\( \theta^i \leftarrow \theta^i_0 \), current estimate of an optimal point in \([1]\)
\( x \leftarrow x_0 \), current state

loop
if \( \text{mod}(n, Q) = 0 \) then 
\[ \hat{\pi}^i(x) \leftarrow \text{perturb}(\pi^i(x), \delta) \]
else 
\[ \hat{\pi}^i(x) \leftarrow \pi^i(x) \]
end if 
\[ a^i \leftarrow \text{pickAtRandom}(\hat{\pi}^i(x), \mathcal{A}_t^i(x)) \]

Play action \( a^i \) along with other agents in current state \( x \in \mathcal{S} \) to get environment response as next state \( y \in \mathcal{S} \) and reward \( r^i \in \mathbb{R} \).
\[ g^i_{x,a^i} \leftarrow r^i + \beta v^i(y) - v^i(x) \]
\[ \pi^i(x,a^i) \leftarrow \Gamma(\pi^i(x,a^i) + b(n)\sqrt{\pi^i(x,a^i)g^i_{x,a^i}}) \]  \hspace{1cm} (8)
\[ v^i(x) \leftarrow v^i(x) + c(n)g^i_{x,a^i} \]  \hspace{1cm} (9)
if \( n = M \) then 
\[ (v^i, \pi^i) \leftarrow (v^i, \pi^i) \]
Terminate algorithm.
end if 
\[ x \leftarrow y \]
\[ n \leftarrow n + 1 \]
end loop

Note that, unlike OFF-SGSP, the above algorithm is a decentralized reinforcement learning scheme with learning localized to each agent \( i \in \{1, 2, \ldots, N\} \). Every iteration in algorithm 2 represents a discrete-time instant of transaction with the environment. We assume that the environment synchronizes all agents’ iterations. Each agent learns his policy \( \pi^i \) according to algorithm 2 and the theoretical convergence guarantees given subsequently establish that a Nash equilibrium for the underlying general-sum discounted stochastic game will be attained.

**Remark 2.** It was shown in Proposition 3 that \( \sqrt{\pi^i(x,a^i)g^i_{x,a^i}(v^i, \pi^{-i})} \) is a descent direction in \( \pi^i(x,a^i) \) for every \( i = 1, 2, \ldots, N, x \in \mathcal{S}, a^i \in \mathcal{A}_t^i(x) \). Since \( \{\pi^i(x,a^i)\}^\alpha \geq 0 \) for any \( \alpha \geq 0 \),
\[ \{\pi^i(x,a^i)\}^\alpha \sqrt{\pi^i(x,a^i)g^i_{x,a^i}(v^i, \pi^{-i})} \]
can also be seen to be a descent direction in \( \pi^i(x,a^i) \) for every \( i = 1, 2, \ldots, N, x \in \mathcal{S}, a^i \in \mathcal{A}_t^i(x) \). So, the policy updates in algorithms 1 and 2 can be generalized to
\[ \pi^i(x,a^i) \leftarrow \Gamma(\pi^i(x,a^i) + b(n)(\pi^i(x,a^i))^\beta g^i_{x,a^i}(v^i, \pi^{-i})) \]
\[ \pi^i(x,a^i) \leftarrow \Gamma(\pi^i(x,a^i) + b(n)(\pi^i(x,a^i))^\beta g^i_{x,a^i}) \]
respectively, where \( \beta \geq \frac{1}{2} \).

**Remark 3. (Complexity Analysis)** Let \( A \) be the typical number of actions available to any agent in any given state and let \( U \) be the typical number of next states for any state \( s \in \mathcal{S} \). Then, the typical number of multiplications in OFF-SGSP per iteration is \( N \times ((U + 1) \times A^N + 4A) \times |\mathcal{S}| \). Thus, the computational complexity grows exponentially in terms of the number of agents while being linear in the size of the state space. Note that the exponential behaviour in \( N \) appears because of the computation of expectation over possible next states and strategies of agents. This computation is avoided in our on-line ON-SGSP algorithm. In a typical iteration of ON-SGSP, \( (2A + 1) \) is the number of multiplications considering the computation of the projection operator \( \Gamma \). Thus, per-iteration complexity of OFF-SGSP is \( \Theta(2^N) \) while that of ON-SGSP is \( \Theta(1) \). In comparison, the stochastic tracing procedure of Herings and Peeters [9] has a complexity of \( O(|\mathcal{S}| \times A^N) \) per iteration which is similar to that of OFF-SGSP.

However, per-iteration complexity alone is not sufficient and an analysis of the number of iterations required is necessary to complete the picture. However, convergence rate results for general multi-timescale stochastic approximation schemes are not available, see however, [14] for rate results of two timescale schemes with linear recursions.
Remark 4. We would also like to point out that our algorithms are not proven to find Nash equilibrium of a stochastic game in polynomial time and hence, do not contradict existing game-theoretic complexity results. A well-known result in this context is that finding the Nash equilibrium of a two player (or even $N$ player) game is PPAD-complete.

6 Outline of Convergence Proof

We provide here a sketch of the convergence analysis of the two proposed algorithms. For both the algorithms, we assume that the underlying Markov chain with transition probabilities $p(y|x, \pi)$, $x, y \in S$, corresponding to the general-sum discounted stochastic game, is irreducible and positive recurrent for all possible strategies $\pi$.

Both our algorithms employ two time-scale stochastic approximation [5, Chapter 6]. That is, they comprise of iteration sequences that are updated using two different time-scales or step-size schedules defined via $\{b(n)\}$ and $\{c(n)\}$, respectively. Here, $\{c(n)\}$ is the faster of the two schedules that is used for update of the value $v$ while the slower time-scale $\{b(n)\}$ is used to update $\pi$. Hence, $\pi$ appears quasi-static for updates of $v$. On the other hand, for updates of $\pi$, $v$ seems to have converged to $v_{\pi}$ which is the value $v$ given a particular $\pi$. Convergence analyses of the three algorithms involve the following steps:

1. We first show that the updates of $v$, that are on the faster time-scale, converge to a limit point of the following system of ODEs: $\forall x \in S, i = 1, 2, \ldots, N$,

$$\frac{dv^i(x)}{dt} = r^i(x, \pi) + \beta \sum_{y \in U(x)} p(y|x, \pi)v^j(y) - v^i(x),$$

where $\pi$ is assumed to be time-invariant. We will also see that the system of ODEs above has a unique limit point, henceforth referred to as $v_{\pi}$, which is stable.

2. Next, using the converged values of $v$ corresponding to strategy update $\pi_n$, i.e., $v_{\pi_n}$ on the slower time-scale, we show that updates of $\pi$ converge to a limit point of the following system of ODEs: $\forall a^i \in A^i(x), x \in S, i = 1, 2, \ldots, N$,

$$\frac{d\pi^i(x, a^i)}{dt} = \bar{\Gamma} \left( \sqrt{\pi^i(x, a^i)}g_{x,a^i}(v^i_{\pi}, \pi^{-i}) \right),$$

where for any bounded and continuous function $\epsilon(\cdot)$, the operator $\bar{\Gamma}$ is defined by

$$\bar{\Gamma}(\epsilon(\pi)) = \lim_{\alpha \downarrow 0} \frac{\Gamma(\pi + \alpha \epsilon(\pi)) - \pi}{\alpha}.$$

3. Finally, we will show that each stable limit point of the system of ODEs in Step 2 above is an SG-SP point. Thus, the strategy $\pi$ corresponding to each stable limit of the above system of ODEs gives a Nash equilibrium of the underlying general-sum discounted stochastic game.

7 Simulation Results

We define a simple general-sum discounted stochastic game, named “Stick Together Game” or in short STG. Informally, STG is a game where participating agents located on a rectangular terrain are supposed to come together and stay close to each other. For the purpose of illustration we fix $N = 2$ and define the tuple $(N, S, A, p, r, \beta)$ as follows.

1. **State Space, $S$.** Let $M > 0$ and $O = \{(x, y)|x, y \in \mathbb{Z}\}$. Define $W = \{s = (x, y) \in O|0 \leq x, y < M\}$ as the possible positions of an agent. Then the state space is given by the Cartesian product $S = W \times W$.

2. **Action Space, $A$.** For $s \in W$, let $||s||_1 = ||x|| + ||y||$ be its $L_1$ norm. Then, $A(s) = \{a \in O||s + a|| \leq 1\}$ represents the actions available for an agent to move to one-step neighbouring positions of $s \in W$. The action space is then defined by $A = \bigcup_{s^1, s^2 \in W} A(s^1) \times A(s^2)$. Let $U(s) = \{s' \in W||s' - s||_1 \leq 1\}$ represent the set of
all next states for an agent in state \( s \in W \).

3. **Transition probability**, \( p \). We assume that state transitions of individual agents are independent. Let \( q(s'|s,a) \) represent, for agent \( i \), the probability of transition from state \( s \in W \) to \( s' \in W \) upon taking action \( a^i \in A^i(s) \). We define

\[
q(s'|s,a) = \frac{2^{-\|s'-a\|_1}}{\sum_{s'' \in U(s)} 2^{-\|s''-a\|_1}}.
\]

Then, the transition probability is given by \( p((s'^1, s'^2)|(s^1, s^2), (a^1, a^2)) = q(s'^1|s^1, a^1)q(s'^2|s^2, a^2) \). This transition probability function has the highest value towards that next state to which the action points to.

4. **Reward**, \( r \). The reward for the two agents is defined as

\[
r^i(s^i, a^i) = 1 - e^{\|s^1-s^2\|_1},
\]

for state \( (s^1, s^2) \in S \) and action \( (a^1, a^2) \in A(s) \times A(s) \). Thus, the reward is zero if the distance between the two agents is zero. Otherwise, it is a negative and monotonically decreasing function with respect to the distance between the two agents.

For verification of the solution of on-line algorithms in this game, we define a simple average distance tracking iteration as follows: \( d_{n+1} = d_n + a \times (d_n - d_n) \), where \( d_n \) is the distance between the agents at simulation instant \( n \), \( a > 0 \) is a small constant representing iterate step-size and \( d_n \) is the average distance at simulation instant \( n \).

**Results:** In the case of OFF-SGSP, we provide the evolution of the objective function \( f \) as a function of the number of iterations in Figure 2(a). Note that \( f \) should go to zero for a Nash equilibrium point. Figure 2(b) shows the evolution of \( d_n \) with the number of iterations in the case of ON-SGSP applied to the STG. One would expect the average distance \( d_n \) would reduce to zero. However, considering that the transition probability is non-zero in various directions for each agent, the average is non-zero but small in value. All our simulation results are run on a 1.83GHz Intel Core2Duo, 2GB DDR2 667MHz machine. It took \( \approx 50 \) minutes to run 10,000 iterations of OFF-SGSP while it took \( \approx 2.5 \) seconds for 100,000 iterations of ON-SGSP.

Now, we show simulation results for a more realistic version of the STG game where \( M = 30 \). The number of states here is \( |S| = 810,000 \). We provide in this case the results only with on-line algorithms because OFF-SGSP’s computational complexity grows exponentially with \( M \). The evolution of the average distance \( d_n \) between the two agents as a function of the number of iterations of ON-SGSP is shown in Figure 2(c). To complete \( 6 \times 10^7 \) iterations, ON-SGSP took \( \approx 50 \) minutes. It is evident that ON-SGSP iterates converged in \( 2 \times 10^7 \) iterations, implying an average \( \frac{2 \times 10^7}{|S|} \approx 21 \) iterations per state.

**Simple function approximation for STG:** OFF-SGSP and ON-SGSP represent two extreme cases of learning where OFF-SGSP assumes full information of the game while ON-SGSP assumes that neither rewards nor state transition probabilities are known. Here, we explore an intermediate information case albeit restricted to STG where a partial structure of rewards is made known. In particular, let it be known that the reward depends on the difference in positions between the two agents. Then, we can do an approximation about the value function \( v \) and strategy \( \pi \) as follows. Let \( t = ([x_1-x_2], |y_1-y_2|) \in S \) where \( s_1 = (x_1, y_1), s_2 = (x_2, y_2) \in S \) are positions of the two agents. With the partial information about the reward being function of \( t \), we could approximate \( v^i(t) \approx v^i(t) \) and possibly \( \pi^i(s) = \pi^i(t) \), \forall s \in S \). Thus, the algorithms need to compute \( \hat{v} \) and \( \hat{\pi} \) which is on the subspace \( W \subset S \). However, the result would apply to the entire state space \( S \). We present the simulation result of ON-SGSP in Figure 2(d) under this partial information setting for the case where \( M = 30 \). Note that for \( M = 30 \), \( |W| = 900 \). ON-SGSP took \( \approx 15 \) seconds to complete 600,000 iterations. The solution has converged by 200,000 iterations which suggests that it took on an average \( \frac{200,000}{|W|} \approx 22 \) iterations per \( t \in W \) to converge.

### 8 Conclusions

In this paper, we presented two novel algorithms - OFF-SGSP and ON-SGSP, for the computation of Nash equilibrium strategies of any given discounted general-sum stochastic game. OFF-SGSP is an offline centralized algorithm that uses full information about stochastic game structure. On the other hand, ON-SGSP is an online
Figure 2: Performance of our algorithms for STG
decentralized scheme that only uses the stochastic game as a generative model and follows a learning approach. Both the algorithms are shown to converge to a Nash equilibrium in stationary strategies. Further, experimental evaluation suggests that convergence is relatively quick.

There are several future directions to be explored such as handling (i) equilibrium selection based on some specified criteria, (ii) very large state spaces, and (iii) incomplete state information.

A Appendix

A KKT-SP conditions

Recall that the Lagrangian corresponding to (2), can be written as

\[ k(\theta, p, \lambda, s, t) = h_x(\theta, p) + \sum_{z=1}^{m} \lambda_z \left( g_{x,z}^i(\theta) + s_z^2 \right) + \sum_{z=1}^{m} \delta_z \left( -p_z + t_z^2 \right), \]

where \( s_z \) and \( t_z, z = 1, 2, \ldots, m \) are the slack variables.

Using the Lagrangian, the associated KKT conditions for the sub-problem (2) corresponding to a state \( x \in S \) and agent \( i \in \{1, \ldots, N\} \) at a point \((\theta^*, p^*)\) are the following:

\[
\begin{align*}
(a) \; \nabla_{\theta} h_x(\theta^*, p^*) + \sum_{z=1}^{m} \lambda_z \nabla_{\theta} g_{x,z}^i(\theta^*) &= 0, \\
(b) \; \frac{\partial h_x(\theta^*, p^*)}{\partial p_z} - \delta_z + \delta_m &= 0, \quad z = 1, 2, \ldots, m, \\
(c) \; \delta_z p_z^* &= 0, \quad z = 1, 2, \ldots, m, \\
(d) \; \lambda_z g_{x,z}^i(\theta^*) &= 0, \quad z = 1, 2, \ldots, m, \\
(e) \; \lambda_z &\geq 0, \quad z = 1, 2, \ldots, m, \\
(f) \; \delta_z &\geq 0. \quad z = 1, 2, \ldots, m.
\end{align*}
\]

KKT-SP conditions were shown by Prasad and Bhatnagar [21] to be necessary and sufficient for \((v^*, \pi^*)\) to represent a Nash equilibrium point of the underlying stochastic game and \( \pi^* \) to be a Nash strategy-pair in the case of \( N = 2 \). However, this was shown under an assumption that for each sub-problem, \( \{ \nabla_{\theta} g_{x,z}^i(\theta^*) : z = 1, 2, \ldots, m \} \) is a set of linearly independent vectors. On the other hand, we formulated an alternative easier set of conditions that we called SG-SP conditions (see Definition 1). An added advantage with these conditions is that they do not impose any linear independence requirement to ensure that the solution points of the sub-problems correspond to Nash equilibria, unlike the KKT-SP conditions.

We now establish the kinship between SG-SP and KKT-SP conditions.

Lemma 4 (KKT-SP \(\Rightarrow\) SG-SP). A KKT-SP point is also an SG-SP point.

Proof. A KKT-SP point \((v^*, \pi^*)\) is a feasible point of (1). For every sub-problem, substitute and eliminate \( \lambda_z^* = p_z^* \) and \( \delta_z^* = -g_{x,z}^i(\theta^*), z = 1, 2, \ldots, m. \) Then

1. Conditions (11(a)) and (11(b)) are satisfied;
2. Conditions (11(c)) and (11(d)) reduce to (4), and
3. Conditions (11(e)) and (11(f)) are satisfied as the point \((v^*, \pi^*)\) is assumed to be feasible.
B Proofs of Section

Theorem 5. A feasible point \((v^*, \pi^*)\) of the optimization problem (1) provides a Nash equilibrium in stationary strategies to the corresponding general-sum discounted stochastic game if and only if \(f(v^*, \pi^*) = 0\).

Proof. See [7, Theorem 3.8.2] for a proof in the case of \(N = 2\). The proof works in a similar manner for general \(N\).

Lemma 6 (SG-SP \(\Rightarrow\) Nash). An SG-SP point \((v^*, \pi^*)\) gives Nash strategy-tuple for the underlying stochastic game.

Proof. The objective function value \(f(v^*, \pi^*)\) of the optimization problem (1) can be expressed as a summation of terms of the form \(p^*_z[-g^*_z, z(\theta^*)]\) over \(z = 1, 2, \ldots, m\) and over all sub-problems. Condition (4) suggests that each of these terms is zero which implies \(f(v^*, \pi^*) = 0\). From Filar and Vrieze [7, Theorem 3.8.2, page 132], since \((v^*, \pi^*)\) is a feasible point of (1) and \(f(v^*, \pi^*) = 0\), \((v^*, \pi^*)\) corresponds to Nash equilibrium of the underlying stochastic game.

Lemma 7 (Nash \(\Rightarrow\) SG-SP). A strategy \(\pi^*\) is Nash if \((v^*, \pi^*)\) for the corresponding optimization problem (1) is an SG-SP point.

Proof. From Filar and Vrieze [7, Theorem 3.8.2, page 132], if a strategy \(\pi^*\) is Nash, then a feasible point \((v^*, \pi^*)\) exists for the corresponding optimization problem (1), where \(f(v^*, \pi^*) = 0\). From the constraints of (1), it is clear that for a feasible point, \(p^*_z[-g^*_z, z(\theta^*)] \geq 0\), for \(z = 1, 2, \ldots, m\), for every sub-problem. Since the sum of all these terms, i.e., \(f(v^*, \pi^*)\), is zero, each of these terms is zero, i.e., \((v^*, \pi^*)\) satisfies (4). Thus, \((v^*, \pi^*)\) is an SG-SP point.

Theorem 8 (Nash \(\Leftrightarrow\) SG-SP). A strategy \(\pi^*\) is Nash if and only if \((v^*, \pi^*)\) for the corresponding optimization problem (1) is an SG-SP point.

Proof. The proof follows from a combination of Lemmas 6 and 7.

C Assumptions

We make the following assumptions for the analysis of our algorithms.

Assumption 2. A Nash equilibrium in stationary randomized strategies exists for general-sum discounted stochastic games.

Assumption 3. The step-sizes \(\{b(n)\}, \{c(n)\}\) are such that \(b(n), c(n) > 0 \forall n\),

\[
\sum_{n=1}^{\infty} b(n) = \infty, \quad \sum_{n=1}^{\infty} c(n) = \infty,
\]

\[
\sum_{n=1}^{\infty} (b^2(n) + c^2(n)) < \infty, \quad \frac{b(n)}{c(n)} \to 0.
\]

Assumption 4. The underlying Markov chain with transition probabilities \(p(y|x, \pi)\), \(x, y \in S\), corresponding to the general-sum discounted stochastic game, is irreducible and positive recurrent for all possible strategies \(\pi\).

D Proofs of Section

Proposition 9. Suppose the value \(\pi^i(x)\) is maintained as a probability over \(A^i(x)\) for all \(x \in S, i = 1, 2, \ldots, N\). Then for each \(i = 1, 2, \ldots, N\), \(x \in S, a^i \in A^i(x)\), we have that \(\sqrt{\pi^i(x, a^i)g^i_{x, a^i}(\psi^i, \pi^{-i})}\) is a non-ascent, and in particular a descent direction if \(\sqrt{\pi^i(x, a^i)g^i_{x, a^i}(\psi^i, \pi^{-i})} \neq 0\), in the objective of the main optimization problem (7). Here \(g^i_{x, a^i}(\psi^i, \pi^{-i}) \leq 0\) represents the first constraint (1).
Proof. The objective function of the optimization problem \((1)\) can be rewritten in terms of \(g^i_{x,a^i}(v^i, \pi^{-i})\) as

\[
 f(v, \pi) = \sum_{i=1}^{N} \sum_{x \in S} \sum_{a^i \in A^i(x)} \left\{ \pi^i(x, a^i) \left[ -g^i_{x,a^i}(v^i, \pi^{-i}) \right] \right\}.
\]

Consider an \(a^i \in A^i(x)\) for some \(x \in S\) and \(i \in \{1, 2, \ldots, N\}\). We show that \(\sqrt{\pi^i(x, a^i)}g^i_{x,a^i}(v^i, \pi^{-i})\) is a descent direction by the following Taylor series argument. Let

\[
 \hat{\pi}^i(x, a^i) = \pi^i(x, a^i) + \delta \sqrt{\pi^i(x, a^i)}g^i_{x,a^i}(v^i, \pi^{-i}),
\]

for a small \(\delta > 0\). Define \(\hat{\pi}\) to be the same as \(\pi\) except with the probability of picking action \(a^i\) in state \(x\) by agent \(i\) being changed to \(\hat{\pi}^i(x, a^i)\) (and the rest staying the same). Then by Taylor series expansion of \(f(v, \hat{\pi})\) till the first order term for small \(\delta\),

\[
 f(v, \hat{\pi}) = f(v, \pi) + \delta \sqrt{\pi^i(x, a^i)}g^i_{x,a^i}(v^i, \pi^{-i}) \frac{\partial f(v, \pi)}{\partial \pi^i(x, a^i)} + o(\delta).
\]

It is easy to see that

\[
 \frac{\partial f(v, \pi)}{\partial \pi^i(x, a^i)} = -g^i_{x,a^i}(v^i, \pi^{-i}).
\]

So,

\[
 f(v, \hat{\pi}) = f(v, \pi) - \delta \sqrt{\pi^i(x, a^i)} \left( g^i_{x,a^i}(v^i, \pi^{-i}) \right)^2 + o(\delta).
\]

Thus, for \(a' \in A^i(x), x \in S\) and \(i \in \{1, 2, \ldots, N\}\) where \(\pi^i(x, a^i) > 0\) and \(g^i_{x,a^i}(v^i, \pi^{-i}) \neq 0\), \(f(v, \hat{\pi}) < f(v, \pi)\) for small enough \(\delta\), while for the remaining cases, \(f(v, \hat{\pi}) \approx f(v, \pi)\). Note that if \(f(v, \pi) > 0\) which implies that solution is not yet achieved (see Theorem\(\text{5}\)), there is at least one former case.

\[\square\]

E Convergence analysis

We provide a proof of convergence of the two proposed algorithms - OFF-SGSP and ON-SGSP, respectively. For both the algorithms, we assume that the underlying Markov chain with transition probabilities \(p(y|x, \pi), x, y \in S,\) corresponding to the general-sum discounted stochastic game, is irreducible and positive recurrent for all possible strategies \(\pi\).

Both our algorithms employ two time-scale stochastic approximation\[1,\) Chapter 6\]. That is, they comprise of iteration sequences that are updated using two different time-scales or step-size schedules defined via \{\(b(n)\)\} and \{\(c(n)\)\}, respectively. Here, \{\(c(n)\)\} is the faster of the two schedules that is used for update of the value \(v\) while the slower time-scale \{\(b(n)\)\} is used to update \(\pi\). Hence, \(\pi\) appears quasi-static for updates of \(v\). On the other hand, for updates of \(\pi, v\) seems to have converged to \(v_\pi\) which is the value \(v\) given a particular \(\pi\). Convergence analyses of the three algorithms involve the following steps:

1. We first show that the updates of \(v\), that are on the faster time-scale, converge to a limit point of the following system of ODEs:

\[
 \frac{dv^i(x)}{dt} = r^i(x, \pi) + \beta \sum_{y \in U^i(x)} p(y|x, \pi)v^i(y) - v^i(x), \forall x \in S, i = 1, 2, \ldots, N, \tag{12}
\]

where \(\pi\) is assumed to be time-invariant. We will also see that the system of ODEs \((12)\) has a unique limit point which is stable. We will refer to the value \(v\) at that limit point as \(v_\pi\).

2. Next, using the converged values of \(v\) corresponding to strategy update \(\pi_n\), i.e., \(v_{\pi_n}\) on the slower time-scale, we show that updates of \(\pi\) converge to a limit point of the following system of ODEs:

\[
 \frac{d\pi^i(x, a^i)}{dt} = \Gamma \left( \sqrt{\pi^i(x, a^i)}g^i_{x,a^i}(v^i_{\pi_n} - \pi^{-i}) \right), \forall a^i \in A^i(x), x \in S, i = 1, 2, \ldots, N, \tag{13}
\]

14
Lemma 11. For a given \( \pi \) of ODEs (12). We now show that (6) is trivially satisfied. Thus by Borkar and Meyn [6, Theorem 2.2], asymptotically stable equilibrium for the ODE (15). Assumption (A1) of Borkar and Meyn [6] is thus satisfied. Since the updates of \( v^i = [I - \beta P_{\pi}]^{-1} R^i_\pi, i = 1, 2, \ldots, N. \) (14)

Proof. The system of ODEs (12) can be re-written in vector form as given below.

\[
\frac{dv^i}{dt} = R^i_\pi + \beta P_{\pi} v^i - v^i.
\] (15)

Rearranging terms, we get

\[
\frac{dv^i}{dt} = R^i_\pi + (\beta P_{\pi} - I)v^i,
\]

where \( I \) is the identity matrix of suitable dimension. Note that for a fixed \( \pi \), this ODE is linear in \( v^i \) with state transition matrix as \((\beta P_{\pi} - I)\). Since \( P_{\pi} \) is a stochastic matrix, the magnitude of all its eigen-values is upper bounded by 1. Hence all the eigen-values of the state transition matrix \((\beta P_{\pi} - I)\) have negative real parts and the matrix \((\beta P_{\pi} - I)\) is in particular non-singular. Thus by standard linear systems theory, the above ODE has a unique globally asymptotically stable limit point which can be computed by setting \( \frac{dv^i}{dt} = 0, i = 1, 2, \ldots, N, \) i.e.,

\[
R^i_\pi + (\beta P_{\pi} - I)v^i = 0, i = 1, 2, \ldots, N.
\]

The trajectories of the ODE (15) converge to the above point starting from any initial condition in lieu of the above.

3. Finally, we will show that each stable limit point of the system of ODEs (13) is an SG-SP point. Thus, the strategy \( \pi \) corresponding to each stable limit of the system of ODEs (13) gives a Nash equilibrium of the underlying general-sum discounted stochastic game.

E.1 OFF-SGSP

Let \( R^i_\pi = \langle r^i(x, \pi), x \in S \rangle \) be a column vector of rewards to agent \( i \) and \( P_{\pi} = [p(y|x, \pi), x \in S, y \in S] \) be the transition probability matrix, both for a given \( \pi \).

Lemma 10. The system of ODEs (12) has a unique globally asymptotically stable limit point given by

\[ v^i_\pi = [I - \beta P_{\pi}]^{-1} R^i_\pi, i = 1, 2, \ldots, N. \]

Proof. The system of ODEs (12) can be re-written in vector form as given below.

\[ \frac{dv^i}{dt} = R^i_\pi + \beta P_{\pi} v^i - v^i. \] (15)

Rearranging terms, we get

\[ \frac{dv^i}{dt} = R^i_\pi + (\beta P_{\pi} - I)v^i, \]

where \( I \) is the identity matrix of suitable dimension. Note that for a fixed \( \pi \), this ODE is linear in \( v^i \) with state transition matrix as \((\beta P_{\pi} - I)\). Since \( P_{\pi} \) is a stochastic matrix, the magnitude of all its eigen-values is upper bounded by 1. Hence all the eigen-values of the state transition matrix \((\beta P_{\pi} - I)\) have negative real parts and the matrix \((\beta P_{\pi} - I)\) is in particular non-singular. Thus by standard linear systems theory, the above ODE has a unique globally asymptotically stable limit point which can be computed by setting \( \frac{dv^i}{dt} = 0, i = 1, 2, \ldots, N, \) i.e.,

\[ R^i_\pi + (\beta P_{\pi} - I)v^i = 0, i = 1, 2, \ldots, N. \]

The trajectories of the ODE (15) converge to the above point starting from any initial condition in lieu of the above.

For a given \( \pi \), the updates of \( v \) in equation (6) (OFF-SGSP) can be seen as Euler discretization of the system of ODEs (12). We now show that \( v_n \) in equation (6) of OFF-SGSP converges to \( v_\pi \) as given in equation (14).

Lemma 11. For a given \( \pi \), i.e., with \( \pi^1_n \equiv \pi^i_\pi \), updates of \( v \) governed by (6) (OFF-SGSP) satisfy \( v_n \rightarrow v_\pi \), as \( n \rightarrow \infty \), where \( v_\pi \) is the globally asymptotically stable equilibrium point of the system of ODEs (12).

Proof. We verify here assumptions (A1) and (A2) of Borkar and Meyn [6] in order to use their result [6, Theorem 2.2]. Let \( h(v^i) = R^i_\pi + (\beta P_{\pi} - I)v^i \). Since \( h(v^i) \) is linear in \( v^i \), it is Lipschitz continuous. Let \( h_r(v^i) = \frac{h(rv^i)}{r} \) for a scalar real number, \( r > 0 \). It is easy to see that \( h_r(v^i) = \frac{R^i_\pi + (\beta P_{\pi} - I)v^i}{r} \). Now, \( h_\infty(v^i) = \lim_{r \rightarrow \infty} h_r(v^i) = (\beta P_{\pi} - I)v^i \). Now since \((\beta P_{\pi} - I)\) has negative real parts, the ODE \( \frac{dv^i}{dt} = h_\infty(v^i) \) has the origin as its unique globally asymptotically stable equilibrium. Further, as shown in Lemma 10, \( v^i_\pi \) is the unique globally asymptotically stable equilibrium for the ODE (15). Assumption (A1) of Borkar and Meyn [6] is thus satisfied. Since the updates of \( v \) in equation (6) do not have any noise term in them, assumption (A2) of Borkar and Meyn [6] is trivially satisfied. Thus by Borkar and Meyn [6, Theorem 2.2], \( v_n \) in equation (6) converges to the globally asymptotically stable limit point \( v_\pi \) given in equation (14).
Thus on the faster time-scale \( \{ c(n) \} \), the updates of \( \nu \) obtained from (16) converge to \( \nu_x \), as given by (14). Let

\[
K = \left\{ \pi \left| \Gamma \left( \sqrt{\pi^i(x,a^i)g_{x,a^i}^i(v_{\pi}^i,\pi^{-i})} \right) = 0, \forall a^i \in \mathcal{A}^i(x), x \in S, i = 1, 2, \ldots, N \right\},
\]

be the set of all limit points of the ODEs (13) and let

\[
L = \{ \pi | \pi(x) \text{ is a probability vector over } \mathcal{A}^i(x), \forall x \in S \}.
\]

**Lemma 12.** \( \pi \in K \) if and only if \( \pi \in L \) and

\[
\sqrt{\pi^i(x,a^i)g_{x,a^i}^i(v_{\pi}^i,\pi^{-i})} = 0, \forall a^i \in \mathcal{A}^i(x), x \in S, i = 1, 2, \ldots, N.
\]  \( \text{(16)} \)

**Proof.** If part:

If \( \pi \in L \) and equation (16) is true, then by definition of operators \( \Gamma \) and \( \bar{\Gamma} \), the result follows.

Only if part:

The operator \( \bar{\Gamma} \), by definition, ensures that \( \pi \in L \). Suppose for some \( a^i \in \mathcal{A}^i(x), x \in S \) and \( i \in \{1, 2, \ldots, N\} \), we have \( \bar{\Gamma}(\sqrt{\pi^i(x,a^i)g_{x,a^i}^i(v_{\pi}^i,\pi^{-i})}) = 0 \) but \( \sqrt{\pi^i(x,a^i)g_{x,a^i}^i(v_{\pi}^i,\pi^{-i})} \neq 0 \). Then, \( g_{x,a^i}^i(v_{\pi}^i,\pi^{-i}) \neq 0 \) and since \( \pi \in L, 1 \geq \pi^i(x,a^i) > 0 \). We analyze this condition by considering the following three cases.

1. Case when \( 1 > \pi^i(x,a^i) > 0 \) and \( g_{x,a^i}^i(v_{\pi}^i,\pi^{-i}) \neq 0 \). In this case, it is possible to find a \( \Delta > 0 \) such that for all \( \delta \leq \Delta \),

\[
1 > \pi^i(x,a^i) + \delta \sqrt{\pi^i(x,a^i)g_{x,a^i}^i(v_{\pi}^i,\pi^{-i})} > 0.
\]

This implies that

\[
\bar{\Gamma}(\sqrt{\pi^i(x,a^i)g_{x,a^i}^i(v_{\pi}^i,\pi^{-i})}) = \sqrt{\pi^i(x,a^i)g_{x,a^i}^i(v_{\pi}^i,\pi^{-i})} \neq 0,
\]

which contradicts the initial supposition.

2. Case when \( \pi^i(x,a^i) = 1 \) and \( g_{x,a^i}^i(v_{\pi}^i,\pi^{-i}) \neq 0 \). Since \( v_{\pi}^i \) is solution of the system of ODEs (12), the following should hold:

\[
\sum_{\hat{a}^i \in \mathcal{A}^i(x)} \pi^i(x,\hat{a}^i)g_{x,\hat{a}^i}^i(v_{\pi}^i,\pi^{-i}) = \pi^i(x,a^i)g_{x,a^i}^i(v_{\pi}^i,\pi^{-i}) = 0.
\]

This again leads to a contradiction.

The result follows. \( \square \)

From Lemma 12 the set \( K \) can be redefined as follows:

\[
K = \left\{ \pi \in L \left| \sqrt{\pi^i(x,a^i)g_{x,a^i}^i(v_{\pi}^i,\pi^{-i})} = 0, \forall a^i \in \mathcal{A}^i(x), x \in S, i = 1, 2, \ldots, N \right. \right\}.
\]

Let

\[
G = \left\{ \pi \in L \left| g_{x,a^i}^i(v_{\pi}^i,\pi^{-i}) \leq 0, \forall a^i \in \mathcal{A}^i(x), x \in S, i = 1, 2, \ldots, N \right. \right\}
\]

represent the set of all feasible points of the optimization problem (1). The set \( K \) can be partitioned using the feasible set \( G \) as \( K = K_1 \cup K_2 \) where \( K_1 = K \cap G \) and \( K_2 = K \setminus K_1 \). We now show that the set \( K_1 \) is the set of all locally asymptotically stable equilibrium points and that \( K_2 \) is the set of locally unstable equilibrium points of the system of ODEs (13).

**Lemma 13.** All \( \pi^* \in K_1 \) are locally asymptotically stable equilibrium points of the system of ODEs (13).
Proof. Let $V(\pi) = f(v_\pi, \pi)$. We show that $V(\pi)$ is a Lyapunov function for every equilibrium point $\pi^* \in K_1$. Consider a particular equilibrium point $\pi^* \in K_1$. With a suitable norm $\| \cdot \|$ defined over the set $L$, let $B_\delta(\pi^*) = \{ \pi \in L | \| \pi - \pi^* \| < \delta \}$ be a small neighbourhood of $\pi^*$ in set $L$ such that $B_\delta(\pi^*) \setminus K$ is non-empty. Then, there exist $\Delta > 0$ such that for all $\delta < \Delta$, the following properties hold.

1. $V(\pi) > 0$ for $\pi \in B_\delta(\pi^*) \setminus K$ and $V(\pi^*) = 0$. This follows from the fact that $f(v_\pi, \pi) \geq 0$ for $\pi \in G$ and that for every $\pi \in K, f(v_\pi, \pi) = 0$.

2. $\frac{dV(\pi)}{dt} < 0$ for $\pi \in B_\delta(\pi^*) \setminus K$. Consider

\[
\frac{dV(\pi)}{dt} = \nabla_\pi V(\pi)^T \frac{d\pi}{dt} = \sum_{i=1}^{N} \sum_{x \in S} \sum_{a_i \in A_i(x)} \frac{\partial V(\pi)}{\partial \pi^i(x, a^i)} \frac{d\pi^i(x, a^i)}{dt}.
\]

By Proposition 5, the direction chosen for update of $\pi^i(x, a^i)$, i.e., $\sqrt{\pi^i(x, a^i)} g^i_{x,a^i}(v^i_\pi, \pi^{-i})$ is a descent direction in $\pi^i(x, a^i)$ along the objective $f(v, \pi)$. Since, by definition, $V(\pi) = f(v_\pi, \pi)$, $\frac{\partial V(\pi)}{\partial \pi^i(x, a^i)}$ is a steepest ascent direction in $\pi^i(x, a^i)$ along the objective $f(v, \pi)$. Now since $\pi^i(x, a^i)$ is a scalar variable, it follows that

\[
\frac{\partial V(\pi)}{\partial \pi^i(x, a^i)} \frac{d\pi^i(x, a^i)}{dt} \leq 0, \forall a^i \in A^i(x), x \in S, i = 1, 2, \ldots, N.
\]

Thus, $\frac{dV(\pi)}{dt} \leq 0$. Since $\pi \notin K$, there exists some $a^i \in A^i(x), x \in S, i \in \{1, 2, \ldots, N\}$ such that $\pi^i(x, a^i) > 0$ and $g^i_{x,a^i}(v^i_\pi, \pi^{-i}) \neq 0$. Also, $\pi^i(x, a^i) < \delta$ by definition of $B_\delta(\pi^*)$. Thus, this corresponds to a non-boundary point which implies that $\Gamma(\sqrt{\pi^i(x, a^i)} g^i_{x,a^i}(v^i_\pi, \pi^{-i})) \neq 0$ with same sign as that of $g^i_{x,a^i}(v^i_\pi, \pi^{-i})$. Also, $\Delta > 0$ can be chosen such that $\frac{\partial V(\pi)}{\partial \pi^i(x, a^i)} 
eq 0$ for all $\pi \in B_\delta(\pi^*) \setminus K$. Hence,

\[
\frac{dV(\pi)}{dt} < 0.
\]

\[\square\]

Lemma 14. All $\pi^* \in K_2$ are unstable equilibrium points of the system of ODEs (13).

Proof. For any $\pi^* \in K_2$, there exists some $a^i \in A^i(x), x \in S, i \in \{1, 2, \ldots, N\}$, such that $g^i_{x,a^i}(v^i_\pi, \pi^{-i}) > 0$ and $\pi^i(x, a^i) = 0$ because $K_2$ is not in the feasible set $G$. Let $B_\delta(\pi^*) = \{ \pi \in L | \| \pi - \pi^* \| < \delta \}$. Choose $\delta > 0$ such that $g^i_{x,a^i}(v^i_\pi, \pi^{-i}) > 0$ for all $\pi \in B_\delta(\pi^*) \setminus K$. So, $\Gamma(\sqrt{\pi^i(x, a^i)} g^i_{x,a^i}(v^i_\pi, \pi^{-i})) > 0$ for any $\pi \in B_\delta(\pi^*) \setminus K$ which suggests that $\pi^i(x, a^i)$ will increase when moving away from $\pi^*$. Thus, $\pi^*$ is an unstable equilibrium point of the system of ODEs (13).

\[\square\]

Lemma 15. The update of $\pi$, in OFF-SGSP, given by (5), converges to $\pi^* \in K$, almost surely.

Proof. The updates of $\pi$ given by (5) on the slower time-scale $\{b(n)\}$ can be rewritten as

\[
\pi^i_{n+1}(x, a^i) = \Gamma(\pi^i_n(x, a^i) + b(n) \sqrt{\pi^i_n(x, a^i)} g^i_{x,a^i}(v^i_\pi, \pi^{-i}) + b(n) \chi_n),
\]

\[
\forall a^i \in A^i(x), x \in S, i = 1, 2, \ldots, N, (17)
\]

where $\chi_n$ is an $o(1)$ error term that goes to zero as $n \to \infty$. The set of iterations in (17) can be viewed as an Euler discretization of the system of ODEs (13) with an extra error term $\chi_n$ which however (as mentioned before) is $o(1)$. Using a standard stochastic approximation argument, see Kushner and Clark [15, pp. 191-196], and the Hirsch’s Lemma [10, Theorem 1, pp. 339], it is easy to see that updates of $\pi$ in (17) converge to an equilibrium point for the system of ODEs (13).

\[\square\]
Note that from the foregoing, the set $K$ comprises of both stable and unstable attractors and in principle from Lemma 13 the iterates $\pi_n^i$ governed by (13) can converge to an unstable equilibrium. In most practical scenarios, however, a gradient descent scheme is observed to converge to a stable equilibrium. In fact, the $\delta$-offset policy computed using perturb($\cdot, \delta$) for every $Q > 0$ iterations in OFF-SGSP ensures numerically that as $n \to \infty$, $\pi_n \to \pi^* \in K_2$. In other words, convergence of the strategy sequence $\pi_n$ obtained from algorithm 1 is to the stable set $K_1$. Let

$$K' = \{ \pi \in L | \pi \text{ is an SG-SP point of the underlying stochastic game} \}.$$  

We now show that $K_1 = K'$.

**Theorem 16.** All stable limit points of the system of ODEs (13) correspond to SG-SP points of the underlying discounted stochastic game and the policy in every SG-SP point is a stable limit point of the system of ODEs (13).

**Proof.** The proof follows simply from the definition of SG-SP point and the set $K_1$.

To summarize, OFF-SGSP converges to a stable limit point of the system of ODEs (12,13) and every stable limit point corresponds to a Nash-equilibrium of the underlying general-sum discounted stochastic game.

### E.2 ON-SGSP

**Theorem 17.** The updates of $v$ in ON-SGSP converge to the globally asymptotically stable limit point of the system of ODEs (12).

**Proof.** Fix a state $x \in S$. Let $\{\tilde{n}\}$ represent a sub-sequence of iterations in ON-SGSP when the state is $x \in S$. Also, let $Q_n = \{ \tilde{n} : \tilde{n} < n \}$. For a given $\pi$, the updates of $v$ on the slower time-scale $\{c(n)\}$ given in equation (9) can be re-written as

$$v_{\tilde{n}+1}^i(x) = v_{\tilde{n}}^i(x) + c(n) \left[ \sum_{a^i \in A^i(x)} \pi_{\tilde{n}}^i(x, a^i) g_{x,a}^i(v_{\tilde{n}}^i, \pi_{\tilde{n}}^{-i}) + \tilde{\chi}_{\tilde{n}} \right],$$

where $\tilde{\chi}_{\tilde{n}} = \tilde{g}_{x,a}^i - \sum_{a^i \in A^i(x)} \pi_{\tilde{n}}^i(x, a^i) g_{x,a}^i(v_{\tilde{n}}^i, \pi_{\tilde{n}}^{-i})$, is the noise term. Let $\tilde{M}_n = \sum_{m \in Q_n} c(m) \tilde{\chi}_n$. Note that by assumption 4 every state $x \in S$ is visited infinitely often. Then, $\tilde{M}_n, n \geq 0$ is a convergent martingale sequence by the martingale convergence theorem (since $\sum_n c^2(\tilde{n}) < \infty$ and $\|g_i\| = \|g_{x,a}^i(\cdot, \cdot)\| < \infty$). The above equation can now be seen as an Euler discretization of the system of ODEs (12) using Hirsch’s Lemma [10, Theorem 1, pp. 339]. Thus, $v$ converges to the globally asymptotically stable limit point $v_\pi$ (see equation (14)) of the system of ODEs (12).

**Theorem 18.** The updates of $\pi$ in ON-SGSP converge to a limit point of the system of ODEs (13), almost surely.

**Proof.** The update of $\pi$ on the slower time-scale can be re-written as

$$\pi_{\tilde{n}+1}^i(x, a^i) = \Gamma(\pi_{\tilde{n}}^i(x, a^i) + b(\tilde{n}) \sqrt{\pi_{\tilde{n}}^i(x, a^i)} g_{x,a}^i(\pi_{\tilde{n}}^i, \pi_{\tilde{n}}^{-i}) + b(\tilde{n}) \tilde{\chi}_{\tilde{n}} + b(\tilde{n}) \tilde{\chi}_{\tilde{n}}), \forall i = 1, 2, \ldots, N,$$

where $\tilde{\chi}_{\tilde{n}} = o(1)$ is an error term as in (17) and $\tilde{\chi}_{\tilde{n}} = \sqrt{\pi_{\tilde{n}}^i(x, a^i)} \left[ \tilde{g}_{x,a}^i - g_{x,a}^i(v_{\tilde{n}}^i, \pi_{\tilde{n}}^{-i}) \right]$. Let $\tilde{M}_n = \sum_{m \in Q_n} b(m) \tilde{\chi}_n$. Then, following similar argument as previously, $\tilde{M}_n$ is an almost surely convergent martingale sequence by the martingale convergence theorem. Now again using Hirsch’s Lemma, the updates of $\pi$ can be viewed as noisy Euler discretization of the system of ODEs (13). Thus, updates of $\pi$ in ON-SGSP converge to a stable limit point of the system of ODEs (13).
References

[1] N. Akchurina. Multi-agent Reinforcement learning: Algorithm converging to Nash equilibrium in general-sum stochastic games. In 8th International Conference on Autonomous Agents and Multi-agent Systems, 2009.

[2] D. Bergemann and J. Välimäki. Learning and strategic pricing. *Econometrica: Journal of the Econometric Society*, pages 1125–1149, 1996.

[3] D. P. Bertsekas. *Dynamic Programming and Optimal Control*, volume II. Athena Scientific, Belmont, Massachusetts, 3rd edition, 2007. ISBN 1-886529-30-2.

[4] D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA, 1996. ISBN 1-886529-10-8.

[5] V.S. Borkar. *Stochastic approximation: a dynamical systems viewpoint*. Cambridge Univ Press, 2008.

[6] V.S. Borkar and S.P. Meyn. The ode method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2):447–469, 2000.

[7] J. Filar and K. Vrieze. *Competitive Markov Decision Processes*. Springer-Verlag, New York, Inc., 1st edition, November 2004. ISBN 0-387-94805-8.

[8] J. Flesch, F. Thuijsman, and K. Vrieze. Cyclic markov equilibria in stochastic games. *International Journal of Game Theory*, 26(3):303–314, 1997.

[9] P. Herings and R.J.A.P. Peeters. Stationary equilibria in stochastic games: Structure, selection, and computation. *Journal of Economic Theory*, 118(1):32–60, 2004.

[10] M.W. Hirsch. Convergent activation dynamics in continuous time networks. *Neural Networks*, 2(5):331–349, 1989.

[11] J. Hu and M. P. Wellman. Multiagent reinforcement learning: Theoretical framework and an algorithm. In *Proceedings of 15th International Conference on Machine Learning*, pages 242–250, 1999.

[12] J. Hu and M. P. Wellman. Nash q-learning for general-sum stochastic games. In *Journal of Machine Learning Research*, volume 4, pages 1039–1069, 2003.

[13] E. Kalai and E. Lehrer. Rational learning leads to nash equilibrium. *Econometrica: Journal of the Econometric Society*, pages 1019–1045, 1993.

[14] V. R. Konda and J. N. Tsitsiklis. Convergence rate of linear two-time-scale stochastic approximation. *Annals of Applied Probability*, pages 796–819, 2004.

[15] H.J. Kushner and D.S. Clark. *Stochastic approximation methods for constrained and unconstrained systems*. Springer-Verlag, 1978. ISBN 0-387-90341-0.

[16] M. L. Littman. Friend-or-Foe Q-Learning in General Sum Games. In *Proceedings of the 18th International Conference on Machine Learning*, pages 322–328. Morgan Kaufmann, 2001.

[17] L. Mac Dermed and C. L. Isbell. Solving stochastic games. *Advances in Neural Information Processing Systems*, 22:1186–1194, 2009.

[18] G.S. Olley and A. Pakes. The dynamics of productivity in the telecommunications equipment industry. Technical report, National Bureau of Economic Research, 1992.

[19] A. Pakes and R. Ericson. Empirical implications of alternative models of firm dynamics. *Journal of Economic Theory*, 79(1):1–45, March 1998.
[20] A. Pakes and P. McGuire. Stochastic algorithms, symmetric Markov perfect equilibrium, and the curse of dimensionality. *Econometrica*, 69(5):1261–1281, 2001.

[21] H. L. Prasad and S. Bhatnagar. General-sum stochastic games: Verifiability conditions for nash equilibria. *Automatica*, 48(11):2923–2930, nov 2012.

[22] L. S. Shapley. Stochastic games. In *Proceedings of the National Academy of Sciences*, volume 39, pages 1095–1100, 1953.

[23] S. Singh, M. Kearns, and Y. Mansour. Nash convergence of gradient dynamics in general-sum games. In *Proc. of the 16th Conference on Uncertainty in Artificial Intelligence*, pages 541–548, 2000.

[24] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998. ISBN 0-262-19398-1.

[25] J. W. Weibull. *Evolutionary Game Theory*. MIT Press, 1996.

[26] M. Zinkevich, A. Greenwald, and M. Littman. Cyclic equilibria in markov games. *Advances in Neural Information Processing Systems*, 18:1641–1648, 2006.