Journal of Dynamics and Games
©American Institute of Mathematical Sciences
Volume **1**, Number **3**, July **2014**
doi:10.3934/jdg.2014.1.347
pp. **347–361**

POLICY IMPROVEMENT FOR PERFECT INFORMATION ADDITIVE REWARD AND ADDITIVE TRANSITION STOCHASTIC GAMES WITH DISCOUNTED AND AVERAGE PAYOFFS

Matthew Bourque

Department of Mathematics and Statistics
Loyola University Chicago
1032 W. Sheridan Road
Chicago, IL 60660, USA

T. E. S. Raghavan

Department of Mathematics, Statistics, and Computer Science
University of Illinois at Chicago
322 Science and Engineering Offices (M/C 249)
851 S. Morgan Street
Chicago, IL 60607-7045, USA

Abstract. We give a policy improvement algorithm for additive reward, additive transition (ARAT) zero-sum two-player stochastic games for both discounted and average payoffs. The class of ARAT games includes perfect information games.

1. Preliminaries.

1.1. Introduction.

A stochastic game, introduced in a seminal paper of Shapley, [18], is a system which evolves over discrete time steps, representing an ongoing interaction between two players in which each of the players' choices at each time step not only determine an immediate payoff for each player, but also influence the actions and payoffs which will be available at the next time step.

The notion of value in a stochastic game depends on how the players evaluate their infinite stream of payoffs. Shapley proved that stochastic games possess a discounted value (when players evaluate their payoff streams by taking a discounted sum). Mertens and Neyman [14] proved that these games also have an undiscounted or limiting average value (when players evaluate their payoff streams by taking a limiting average).

Stochastic games with rational rewards and transitions need not have rational values [7, example 3.2.1]. It is clear that for such a game, no finite-step algorithm using only arithmetic operations can give an exact answer, and we can only hope for such an algorithm when the game has the order-field property defined in [15]. Fortunately, many natural classes of games have been shown to possess this property [7]. However, even among zero-sum games with the order field property, the general existence of finite-step algorithms for computing the value and optimal strategies

2010 *Mathematics Subject Classification.* Primary: 91A15, 91A05; Secondary: 90C40.
Key words and phrases. Stochastic games, additive reward additive transition, perfect information, policy iteration, Markov decision process.

with respect to discounted or undiscounted criteria is an open question. The finite-step algorithms which have been found for some categories of stochastic games are generally of two types: via a linear program, or policy improvement.

In this paper we give a policy improvement algorithm for solving two player zero-sum stochastic games with additive rewards and additive transitions (ARAT games) with respect to limiting average payoffs. The family of policy improvement algorithms begins with Howard's policy improvement algorithm for discounted Markov decision processes (MDPs) [10]. Policy improvement algorithms are widely recognized as fast in practice for solving MDPs and other related problems [3, 20, 5, 17]; indeed, this speed is suggested by the fact that under certain regularity conditions, policy improvement for MDPs is equivalent to Newton's method [6].

A policy improvement algorithm for zero sum stochastic games of perfect information with respect to the discounted criterion was given by Raghavan and Syed [17]. Syed [19] extended the algorithm to ARAT games with discounted payoff. Their proof techniques were based on an induction on the total number of actions available in all states for the two players. They also observed empirically that the discounted values of MDPs along the algorithmic path were monotonic. A formal proof of this empirical observation, called the Patience Theorem (Theorem 2.3), is key to our proof technique. It provides an alternative proof to the algorithm for discounted games. It also allows us to provide a proof for a new algorithm for average payoff ARAT games.

In contrast to our goal of solving for optimal strategies for games with average payoffs, a more ambitious approach is to search for uniform optimal strategies for stochastic games. This would also yield the optimal strategies for average payoffs at one blow.

Finding uniform optimal strategies is a more difficult problem even for Markov decision processes. Inspired by Jeroslow, [11], Hordijk et al. [9] proposed an algorithm to find uniform optimal strategies for a Markov decision process. This is a simplex-type algorithm on the non-Archimedian ordered field of rational functions.

Using this parametric linear programming method, Avrachenkov et al. [2] propose two algorithms for solving for uniform optimal strategies for two player zero sum stochastic games of perfect information. The first is more involved, but the authors provide a proof of termination along the lines of the proof of the discounted-case algorithm by Raghavan and Syed [17]. The second, unlike Raghavan and Syed's or the one presented in this paper, is a best-response algorithm for both players, and appears to run faster in numerical simulations. However, the termination of this second algorithm in general is still an open question.

To the best of our knowledge, the first policy improvement algorithm for solving stochastic games of perfect information with average payoffs was given by Cochet-Terrasson and Gaubert [4], with a more thorough treatment in [1]. The main proof technique in these papers hinges on Kohlberg's theorem on invariant half-lines of nonexpansive piecewise linear transformations [13]. Though producing a related algorithm, our work is independent of these results, and the techniques are quite different, following a more purely game-theoretic approach beginning with Shapley [18], through Blackwell [3], Veinott [20], and Raghavan and Syed [17]. Another distinguishing feature of this paper is that we have focused on ARAT games, introduced in [16]. This class of games includes perfect information games as a special case.

1.2. **Definition of a stochastic game.** A **stochastic game**

$$\Gamma = \langle S, \mathbf{A}^1, \mathbf{A}^2, \mathbf{R}^1, \mathbf{R}^2, \mathbf{P} \rangle$$

comprises

- a finite set of states $S = \{1, 2, \ldots, N\}$;
- a set of finite nonempty action sets $\mathbf{A}^i = \{A^i(s)\}_{s \in S}$ for player $i \in \{1, 2\}$;
- a set of rewards

$$\mathbf{R}_i = \left\{ r_i(s, a^1, a^2) \in \mathbf{R} \middle| a^1 \in A^1(s), a^2 \in A^2(s), s \in S \right\}$$

  for player i, $i \in \{1, 2\}$;
- and a set of Markovian transition probabilities

$$\mathbf{P} = \left\{ p(s' \mid s, a^1, a^2) \in [0, 1] \middle| a^1 \in A^1(s), a^2 \in A^2(s), s, s' \in S \right\},$$

  where $\sum_{s'=1}^{N} p(s' \mid s, a^1, a^2) = 1$ for all $a^1 \in A^1(s), a^2 \in A^2(s)$ for all states s.

We will refer to the players in the game as player 1 and player 2. (When distinct pronouns enhance readability, we will use the convention that player 1 is male and player 2 is female.) We interpret the reward $r_i(s, a^1, a^2)$ as the immediate payoff to player i when player 1 chooses action a^1 and player 2 chooses action a^2 in state s, and interpret $p(s' \mid s, a^1, a^2)$ as the Markovian probability of transition to state s' in the next time step when the state at the current time step is s, player 1 chooses action a^1 and player 2 chooses action a^2.

We will restrict our attention to zero-sum stochastic games with additive rewards and additive transitions, defined below.

Definition 1.1. A stochastic game Γ is a **zero-sum** game if

$$r_1(s, a^1, a^2) = -r_2(s, a^1, a^2)$$

for all actions $a^1 \in A^1(s), a^2 \in A^2(s)$ available in state s.

For simplicity, since this paper deals exclusively with zero-sum games, we will use a single payoff function $r(s, a^1, s^2) \equiv r_1(s, a^1, a^2)$ with the understanding that, with respect to the discounted or average payoffs corresponding to r, player 1 is a maximizer and player 2 is a minimizer.

Definition 1.2. A stochastic game has the additive reward, additive transition (ARAT) property if for every pair of actions $a^1 \in A^1(s)$ and $a^2 \in A^2(s)$ in every state s,

$$r(s, a^1, a^2) = r^1(s, a^1) + r^2(s, a^2)$$

and

$$p(s' \mid s, a^1, a^2) = p^1(s' \mid s, a^1) + p^2(s' \mid s, a^2)$$

for every state s', with $p^i(s' \mid s, a^i) \geq 0$ for all actions $a^i \in A^i(s)$, for all states $s, s' \in S$, and for both players $i = 1, 2$.

Definition 1.3. A stochastic game Γ has **perfect information** if at most one of $A^1(s)$ and $A^2(s)$ has more than one element for all states s.

For simplicity, when we speak of a "game," we will mean a zero-sum stochastic game with the ARAT property, unless otherwise specified. Note that the class of ARAT stochastic games includes the perfect information stochastic games.

1.3. **Strategies.** A "strategy" for a player is a (possibly randomized) rule for selecting an action at each time step, possibly depending on the current state and the entire history of the game. Strategies which choose an action deterministically in each state, independent of the history of the game, form an important subclass, called pure stationary strategies.

Definition 1.4. A **pure stationary strategy** f for player i in a game Γ is an element of $A^i(1) \times A^i(2) \times \cdots \times A^i(N)$, where $f(s)$ represents the action which player i will choose in state s.

We will denote by F (respectively G) the set of pure stationary strategies for player 1 (respectively player 2), with these strategy sets' dependence on the game understood.

Given a pair of strategies $(f, g) \in F \times G$ for the players in a game, Let $P(f, g)$ be the stationary transition matrix whose (s, s') entry is $p\big(s' \mid s, f(s), g(s)\big)$. We also define a column N-vector $\boldsymbol{r}(f, g)$ whose s-th entry is $r\big(s, f(s), g(s)\big)$.

Note that when one player in game fixes his or her strategy, the induced game for the opponent is a Markov decision process (MDP). We will denote by $\Gamma|_f$ the MDP resulting from the game Γ when one player's strategy is fixed at f. If f is a strategy for player 2, then $\Gamma|_f$ is an MDP for player 1 where payoffs are viewed as rewards and player 1 is a maximizer. If f is a strategy for player 1, then $\Gamma|_f$ is an MDP for player 2 where payoffs are viewed as costs and player 2 is a minimizer.

2. Games With discounted payoffs.

2.1. **Discounted payoffs and discounted value.** The discounted payoff with a discount factor $\beta \in [0, 1)$ for a stochastic game models a situation in which players are more concerned with the relatively short term, with β inversely proportional to myopia.

Definition 2.1. For a given discount factor $\beta \in [0, 1)$, and a pure stationary strategy pair (f, g), the ***β*-discounted payoff** is

$$\begin{aligned} \boldsymbol{v}_\beta(f, g) &= \sum_{t=0}^{\infty} \beta^t P^t(f, g) \boldsymbol{r}(f, g) \\ &= \big[I - \beta P(f, g)\big]^{-1} \boldsymbol{r}(f, g). \end{aligned}$$

Definition 2.2. A pure stationary strategy pair (f^*, g^*) for a game Γ is ***β*-optimal** (over pure stationary strategies) for a discount factor $\beta \in [0, 1)$ if there exists a ***β*-discounted value** $\boldsymbol{v}_\beta(\Gamma)$ such that for all states s,

$$\boldsymbol{v}_\beta(\Gamma) = \max_{f \in F} \boldsymbol{v}_\beta(f, g^*)$$

and

$$\boldsymbol{v}_\beta(\Gamma) = \min_{g \in G} \boldsymbol{v}_\beta(f^*, g),$$

with the max and min taken componentwise.

Raghavan et al. [16] show that ARAT games may be played optimally over pure stationary strategies , and so there is no loss in limiting ourselves to such strategies for these games. We will use the word "strategy" to refer to a pure stationary strategy unless otherwise specified.

2.2. **Policy improvement for discounted games.** In this section, we present a proof of a result empirically observed by Raghavan and Syed [17]. One can view their policy improvement algorithm as solving a sequence of discounted MDPs for player 1. They observed that, in computational examples, the discounted value of these MDPs decreases monotonically when the payoff vectors are compared coordinatewise. Here we prove that this property holds in general for the sequence of strategy pairs produced by their algorithm. In fact we will see later that a similar monotonicity property will hold for the strategy pair sequence produced by the policy improvement algorithm in the average-payoff case. We call this result the Patience Theorem, because it provides an intuitive justification of the "patient" approach the algorithm takes to improving strategies for player 2. We begin with some notation.

For a strategy pair (f, g) and a fixed discount factor β, let $G_\beta(s, g \mid f)$ be the (possibly empty) set of actions $a \in A^2(s)$ (actions for player 2) which satisfy

$$r(s, f(s), a) + \beta \sum_{s'=1}^{N} p\big(s' \mid s, f(s), a\big) v_\beta(s', f, g) < v_\beta(s, f, g), \tag{1}$$

and let $G_\beta(g \mid f)$ be the (possibly empty) set of all pure strategies $g' \neq g$ for player 2 such that, for each state s, either $g'(s) \in G_\beta(s, g \mid f)$ or $g'(s) = g(s)$. The sets $G_\beta(s, g \mid f)$ and $G_\beta(g \mid f)$ are subsets of, respectively, actions and strategies for player 2; we define the sets $G_\beta(s, f \mid g)$ and $G_\beta(f \mid g)$ for player 1 in an analogous fashion, with the inequality in (1) reversed.

Theorem 2.3 (Patience Theorem)**.** *Suppose (f^0, g^0) is a strategy pair and g^1 a strategy for player 2 for a game Γ with a fixed discount factor β such that $G_\beta(f^0 \mid g^0)$ is empty and $g^1 \in G_\beta(g^0 \mid f^0)$. Then $\boldsymbol{v}_\beta(\Gamma|_{g^0}) > \boldsymbol{v}_\beta(\Gamma|_{g^1})$.*

In our proof of this theorem, and in the rest of the paper, we will write $\boldsymbol{x} > \boldsymbol{y}$ for two comparable N-vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ when $x(s) \geq y(s)$ for all $s \in S$, and the inequality is strict for some s.

Proof. Consider the MDP $\Gamma|_g$ for some $g \in G$. The solution of such an MDP by a linear program is well known [7]. It is the solution to the LP

$$\text{minimize} \sum_{s=1}^{N} \gamma(s) v(s)$$

subject to

$$v(s) \geq r\big(s, a, g(s)\big) + \beta \sum_{s'=1}^{N} p\big(s' \mid s, a, g(s)\big) v(s') \text{ for all } a \in A^1(s), s \in S$$

where $\boldsymbol{\gamma}$ is any positive N-vector with $\sum_s \gamma(s) = 1$.

Since $G_\beta(f^0 \mid g^0)$ is empty, f^0 is β-optimal for the MDP $\Gamma|_{g^0}$, and so $\boldsymbol{v}_\beta(f^0, g^0)$ is the optimal solution to the LP corresponding to $\Gamma|_{g^0}$. In particular, it is feasible for this LP. We will show that $\boldsymbol{v}_\beta(f^0, g^0) - \boldsymbol{\epsilon}_X$ is also feasible for the LP corresponding to $\Gamma|_{g^1}$, where $\boldsymbol{\epsilon}_X$ is an N-vector whose coordinates are strictly positive for states contained in some nonempty set X (to be defined later), and zero otherwise. This will show that, for any appropriate $\boldsymbol{\gamma}$,

$$\sum_{s=1}^{N} \gamma(s)\big(v_\beta(s, f^0, g^0) - \boldsymbol{\epsilon}_X(s)\big) \geq \sum_{s=1}^{N} \gamma(s) v_\beta(s, \Gamma|_{g^1}). \tag{2}$$

From here, the theorem is proved as follows: for any state s, we choose a sequence of positive vectors $\{\boldsymbol{\gamma}_n\}$ with $\sum_t \gamma_n(t) = 1$ for all n, and such that $\gamma_n(s) \to 1$ as $n \to \infty$. Replacing $\boldsymbol{\gamma}$ with $\boldsymbol{\gamma}_n$ in (2) and taking a limit as $n \to \infty$ on both sides, yields for all states s, $v_\beta(s, f^0, g^0) - \boldsymbol{\epsilon}_X(s) \geq v_\beta(s, \Gamma|_{g^1})$. Since $\boldsymbol{\epsilon}_X(s)$ is positive for some s and $\boldsymbol{v}_\beta(\Gamma|_{g^0}) = \boldsymbol{v}_\beta(f^0, g^0)$, this gives the theorem. We now proceed to prove that $\boldsymbol{v}_\beta(f^0, g^0) - \boldsymbol{\epsilon}_X$ is feasible for the LP corresponding to $\Gamma|_{g^1}$. Let $v^0(s) = v_\beta(s, f^0, g^0)$. We will define $\boldsymbol{\epsilon}_X$ appropriately along the way.

Let X be the set of states s for which $g^1(s) \in G_\beta(s, g^0 \mid f^0)$.

When $g^1(s) = g^0(s)$ (that is, for states s in which g^0 is not changed),

$$v^0(s) \geq r\big(s, a, g^1(s)\big) + \beta \sum_{s'=1}^{N} p\big(s' \mid s, a, g^1(s)\big) v^0(s') \tag{3}$$

for $a \in A^1(s)$.

For any $s \in X$,

$$v^0(s) > r\big(s, f^0(s), g^1(s)\big) + \beta \sum_{s'=1}^{N} p\big(s' \mid s, f^0(s), g^1(s)\big) v^0(s') \tag{4a}$$

or, by the ARAT property,

$$v^0(s) > \big[r^1\big(s, f^0(s)\big) + r^2\big(s, g^1(s)\big)\big] + \beta \sum_{s'=1}^{N} \big[p^1\big(s' \mid s, f^0(s)\big) + p^2\big(s' \mid s, g^1(s)\big)\big] v^0(s') \tag{4b}$$

We want to show that this inequality holds when we replace $f^0(s)$ by any action $a \in A^1(s)$ in (4a) or, equivalently, in (4b).

Toward this end, for any action $a \in A^1(s)$, since v^0 is optimal (and so feasible) for the LP corresponding to Γ_{g^0} we have

$$v^0(s) \geq \big[r^1\big(s, a\big) + r^2\big(s, g^0(s)\big)\big] + \beta \sum_{s'=1}^{N} \big[p^1(s' \mid s, a) + p^2\big(s' \mid s, g^0(s)\big)\big] v^0(s') \tag{5a}$$

Now the above is an equality when $a = f^0(s)$. Replacing $v^0(s)$ with this equivalent expression, making use of the ARAT property, and taking a difference yields

$$0 \geq \big[r^1\big(s, a\big) - r^1\big(s, f^0(s)\big)\big] + \beta \sum_{s'=1}^{N} \big[p^1(s' \mid s, a) - p^1\big(s' \mid s, f^0(s)\big)\big] v^0(s') \tag{5b}$$

Now summing (4b) and (5b) we have

$$v^0(s) > \big[r^1\big(s, a\big) + r^2\big(s, g^1(s)\big)\big] + \beta \sum_{s'=1}^{N} \big[p^1(s' \mid s, a) + p^2\big(s' \mid s, g^1(s)\big)\big] v^0(s') \tag{6}$$

for $s \in X$, as desired.

Now by the strict inequality in (6), for any $s \in X$ we can choose $\epsilon(s)$ small enough that, for any action a for player 1 in state s, the strict inequality is preserved when we replace $v_\beta(s, f^0, g^0)$ by $v_\beta(s, f^0, g^0) - \epsilon(s)$. We now let $\epsilon_X(s) = \epsilon(s)$ for $s \in X$ and $\epsilon_X(s) = 0$ otherwise. As before, this argument along with (3) and (6) yield

the feasibility of $v_\beta(f^0, g^0) - \epsilon_X$ for the LP corresponding to $\Gamma|_{g^1}$, and the proof is complete. □

3. Games with average payoffs.

3.1. Average payoffs and average value.

The limiting average payoff for a strategy pair models a situation in which players are concerned only with long term payoffs.

Definition 3.1. For a given pure stationary strategy pair (f, g), the limiting average or Cesaro average payoff (also called simply the average payoff or undiscounted payoff) is

$$\boldsymbol{v}(f,g) = \lim_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} P^t(f,g)\boldsymbol{r}(f,g)$$

Definition 3.2. A strategy pair (f^*, g^*) for a game Γ is **average optimal** for the game if there exists an **average value** $\boldsymbol{v}(\Gamma)$ such that for all states s,

$$\boldsymbol{v}(s,\Gamma) = \max_{f\in F} \boldsymbol{v}(f, g^*)$$

and

$$\boldsymbol{v}(\Gamma) = \min_{g\in G} \boldsymbol{v}(f^*, g)$$

Another kind of optimality which will interest us is uniform optimality, which models situations in which players do not necessarily agree on a discount factor for the game, but want strategies which will do well for all discount factors sufficiently close to 1.

Definition 3.3. A pure stationary strategy pair $(\hat{f}, \hat{g})$ is **uniform optimal** for a given game Γ if it is β-optimal for all β sufficiently close to 1.

A uniform optimal strategy for an MDP is average optimal, but as [3, example 1] shows, an average optimal strategy may not be β-optimal for any value of β.

Raghavan et al. [16] proved the existence of β-optimal pure stationary strategies for all discount factors β for ARAT games. Further, in the case of MDPs, Blackwell [3] showed that the β-discounted payoff for any fixed pure stationary strategy is a rational function of β, and proved that a pure stationary uniform optimal strategy exists for any MDP. A similar argument applies for any stochastic game with pure stationary β-optimal strategy pairs, and so we may conclude that ARAT games admit a pure stationary uniform optimal strategy pair. As Blackwell showed for MDPs, such a strategy pair also serves as an average optimal strategy pair. This justifies limiting ourselves to pure stationary strategy pairs when defining average and uniform optimal strategy pairs.

3.2. Properties of average payoffs.

In this subsection, we collect a few results which will be useful later. We will make extensive use of the following theorem relating β-discounted and average payoffs, which is a restatement of [3, Theorem 4, part (a)].

Theorem 3.4. *For a game Γ and strategy pair $(f, g) \in F \times G$*

$$\boldsymbol{v}_\beta(f,g) = \frac{\boldsymbol{v}(f,g)}{1-\beta} + \boldsymbol{y}(f,g) + \boldsymbol{\epsilon}(f,g,\beta)$$

where $\boldsymbol{y}(f,g)$ *is the unique solution to*

$$\big(I - P(f,g)\big)y = \boldsymbol{r}(f,g) - \boldsymbol{v}(f,g) \text{ and } P^*(f,g)y = 0$$

and $\boldsymbol{\epsilon}(f,g,\beta)$ *is a* $\mathbf{R}^N$*-valued function which approaches* 0 *in every coordinate as* $\beta \to 1$.

We will follow the notation used in this theorem consistently, using $\boldsymbol{\epsilon}$ to denote a function which depends on β and which goes to zero as β increases to 1. Also, we will refer to the vector $\boldsymbol{y}(f,g)$ as the **deviation** of the strategy pair (f,g).

Some results from Markov chain theory will be helpful in proofs or for computation in implementing algorithms. We collect them here; details may be found in [3] and [12].

Lemma 3.5. *In the case of a pure stationary strategy pair* (f,g),

1. *the average payoff* $\boldsymbol{v}(f,g)$ *may be computed as*

$$\boldsymbol{v}(f,g) = P^*(f,g)\boldsymbol{r}(f,g),$$

and is the unique solution to

$$\big(I - P(f,g)\big)\boldsymbol{v} = 0 \text{ and } P^*(f,g)\boldsymbol{v} = \boldsymbol{v};$$

where $P^*(f,g)$ *is the Cesaro limit of the matrix* $P(f,g)$.

2. *The vector* $\boldsymbol{y}(f,g)$ *may be computed as*

$$\boldsymbol{y}(f,g) = D(f,g)\boldsymbol{r}(f,g),$$

where $D(f,g) = \big(I - P(f,g) + P^*(f,g)\big)^{-1} - P^*(f,g)$.

3.3. **Policy improvement for games with average payoffs.** In this section, we present a policy improvement algorithm and prove its correctness for solving zero sum ARAT stochastic games. We begin by developing some notation.

First we define sets which are the average-payoff analogues of $G_\beta(s, g \mid f)$ and $G_\beta(g \mid f)$ for a given strategy pair (f,g) for a game Γ. Let $G(s, g \mid f)$ be the (possibly empty) set of actions $a \in A^2(s)$ which satisfy

$$\sum_{s'=1}^{N} p\big(s' \mid s, f(s), a\big)v(s', f, g) \le v(s, f, g) \tag{7}$$

and in case (7) holds with equality

$$r\big(s, f(s), a\big) + \sum_{s'=1}^{N} p\big(s' \mid s, f(s), a\big)y(s', f, g) < v(s, f, g) + y(s, f, g). \tag{8}$$

Let $G(g \mid f)$ be the (possibly empty) set of all pure strategies $g' \neq g$ for player 2 such that for each state s, $g'(s) \in G(s, g \mid f)$ or $g'(s) = g(s)$. We define the sets $G(s, f \mid g)$ and $G(f \mid g)$ (for player 1) analogously, with the inequalities in (7) and (8) reversed. By a theorem in [3], if $G(f \mid g)$ is empty, then f is average optimal for the MDP $\Gamma|_g$.

It will be noted that the sets $G(s, f \mid g)$ and $G(f \mid g)$ are straightforward extensions to the competitive case of the set $G(s, f)$ corresponding to a state s and a strategy f for a single player in an MDP introduced in [3] (our notation differs from Blackwell's in making the sets' dependence on β explicit in the discounted case). We further introduce a competitive version of the set $H(f)$ defined in [20]. The theorem afterward will make clear its significance in terms of 1-optimal strategies, and its utility in our proof will be seen further along.

Given a strategy pair (f, g) for a game Γ, let $H(s, g \mid f)$ be the (possibly empty) set of actions $a \in A^2(s)$ for which (7) and (8) are both equalities and furthermore

$$\sum_{s'=1}^{N} p\big(s' \mid s, f(s), a\big) z(s', f, g) < y(s, f, g) + z(s, f, g) \tag{9}$$

where $\boldsymbol{z}(f, g)$ is the unique solution to

$$\big(I - P(f, g)\big)\boldsymbol{z} = -\boldsymbol{y}(f, g) \text{ and } P^*(f, g)\boldsymbol{z} = 0.$$

Let $H(g \mid f)$ be the set of pure strategies $g' \neq g$ for player 2 with $g'(s) \in H(s, g \mid f)$ or $g'(s) = g(s)$ for all s. As before, we define the sets $H(s, f \mid g)$ and $H(f \mid g)$ for player 1 analogously, with the inequality in (9) reversed.

The following is a restatement of a theorem of Veinott [20, Theorem 6] to the competitive setting.

Theorem 3.6. *For a strategy pair (f, g)*

1. *if $f' \in G(f \mid g)$ then $\boldsymbol{v}(f', g) \geq \boldsymbol{v}(f, g)$ and if $\boldsymbol{v}(f', g) = \boldsymbol{v}(f, g)$, then $\boldsymbol{y}(f', g) > \boldsymbol{y}(f, g)$;*
2. *if $f' \in H(f \mid g)$ then $\boldsymbol{v}(f', g) = \boldsymbol{v}(f, g)$, $\boldsymbol{y}(f', g) \geq \boldsymbol{y}(f, g)$, and if $\boldsymbol{y}(f', g) = \boldsymbol{y}(f, g)$, then $\boldsymbol{z}(f', g) > \boldsymbol{z}(f, g)$;*
3. *if $G(f \mid g)$ is empty, then f is average optimal for the MDP $\Gamma|_g$;*
4. *if $G(f \mid g) \cup H(f \mid g)$ is empty, then $\boldsymbol{y}(f, g) \geq \boldsymbol{y}(f', g)$ for all $f' \in F$ which are average optimal for $\Gamma|_g$. We say that f is* **1-optimal** *for $\Gamma|_g$.*

This theorem holds with the inequalities reversed when g is a strategy for player 1 and f is a strategy for player 2. In particular, the "reversed" version of Theorem 3.6, part 3 states that if $G(g \mid f)$ is empty, then g is average optimal for $\Gamma|_f$. This yields an immediate corollary characterizing average optimal strategy pairs for a game.

Corollary 1. *If $G(f \mid g)$ and $G(g \mid f)$ are both empty, then (f, g) is an average optimal strategy pair for the game.*

We are now ready to present our main result.

Algorithm 1 Policy improvement for ARAT Games With Average Payoffs

1: Choose an arbitrary initial strategy pair (f^0, g^0), and let $k = 0$.
2: **while** $G(f^k \mid g^k) \cup H(f^k \mid g^k)$ or $G(g^k \mid f^k)$ is nonempty **do**
3: Choose f for player 1 (the maximizer) so that $G(f \mid g^k) \cup H(f \mid g^k)$ is empty. Let $f^{k+1} = f$.
4: **if** $G(g^k \mid f^{k+1})$ is nonempty **then**
5: Update the strategy for player 2: choose $g^{k+1} \in G(g^k \mid f^{k+1})$.
6: **else**
7: Let $g^{k+1} = g^k$
8: **end if**
9: Increment k.
10: **end while**
11: When $G(f^k \mid g^k) \cup H(f^k \mid g^k)$ and $G(g^k \mid f^k)$ are both empty, return the strategy pair $(f^*, g^*) = (f^k, g^k)$.

Before we discuss the termination and correctness of this algorithm, a note on the implementation of line 3 is in order. In this step, for a fixed g^k for player 2,

we must find an f for player 1 such that $G(f \mid g^k) \cup H(f \mid g^k)$ is empty. This is achieved with the policy improvement algorithm for a 1-optimal strategy in MDPs in [20]. For details on policy iteration in MDPs, see [3, 20].

For any game Γ and starting strategy pair (f^0, g^0) for which the algorithm terminates, the resulting pair (f^*, g^*) is such that the sets $G(f^* \mid g^*)$ and $G(g^* \mid f^*)$ are both empty. Hence, by Corollary 1, the pair is average optimal. Since there are only finitely many pure stationary strategy pairs, our goal will be to prove that the algorithm cannot cycle. In the discounted case this can be done via the Patience Theorem, showing that each improvement by player 2 presents player 1 with a new MDP whose optimal value is strictly smaller than the last. Here, instead of looking directly at the average value of the MDPs for player 1 corresponding to the strategies chosen by player 2 in line 5 of Algorithm 1, we will rely on the same discounted result; showing that for β sufficiently close to 1, the β-discounted value of these MDPs is strictly decreasing. Our method will be to show that the sequence of strategies $(f^0, g^0), (f^1, g^1), \ldots$ can be "shadowed" by a sequence for the β-discounted algorithm, in which the strategies f^k chosen by the algorithm may be taken to be a uniform optimal $\hat{f}^k$. Then the Patience Theorem may be applied to show that for all discount factors $\beta < 1$ sufficiently close to 1, $\boldsymbol{v}_\beta(\Gamma|_{g^k}) < \boldsymbol{v}_\beta(\Gamma|_{g^{k-1}})$. The next three lemmas will help us make this precise.

First, we will show that although the strategy that player 1 finds in line 3 may not itself be uniform optimal for $\Gamma|_{g^k}$, it must have the same value and deviation as the uniform optimal strategy for this MDP.

Lemma 3.7. *Given a strategy pair (f, g) for a game Γ, if $G(f \mid g) \cup H(f \mid g)$ is empty then for any f^* uniform optimal for $\Gamma|_g$, $\boldsymbol{v}(f^*, g) = \boldsymbol{v}(f, g)$ and $\boldsymbol{y}(f^*, g) = \boldsymbol{y}(f, g)$.*

Proof. If $G(f \mid g) \cup H(f \mid g)$ is empty then by Theorem 3.6, $\boldsymbol{v}(f, g) \geq \boldsymbol{v}(f^*, g)$ and if $\boldsymbol{v}(f^*, g) = \boldsymbol{v}(f, g)$ then $\boldsymbol{y}(f, g) \geq \boldsymbol{y}(f^*, g)$. Now for all β sufficiently close to 1, f^* is β-optimal for $\Gamma|_g$, so $\boldsymbol{v}_\beta(f^*, g) - \boldsymbol{v}_\beta(f, g) \geq 0$ or, rewriting using Theorem 3.4:

$$\frac{1}{1-\beta}\big[\boldsymbol{v}(f^*, g) - \boldsymbol{v}(f, g)\big] + \big[\boldsymbol{y}(f^*, g) - \boldsymbol{y}(f, g)\big] + \boldsymbol{\epsilon}(f', f, g, \beta) \geq 0.$$

Since the above inequality holds for all β sufficiently close to 1, $\boldsymbol{v}(f^*, g) \geq \boldsymbol{v}(f, g)$, so $\boldsymbol{v}(f^*, g) = \boldsymbol{v}(f, g)$. But then, since $\boldsymbol{\epsilon}(f', f, g, \beta) \to 0$ as $\beta \nearrow 1$, $\boldsymbol{y}(f^*, g) \geq \boldsymbol{y}(f, g)$, and we conclude that $\boldsymbol{y}(f^*, g) = \boldsymbol{y}(f, g)$. □

Next we show that any improvement for player 2 in the average sense is also a "uniform improvement," that is, it is an improvement in the discounted sense for all discount factors sufficiently close to 1.

Lemma 3.8. *Given a strategy pair (f, g), for $\beta < 1$ sufficiently close to 1, $G(g \mid f) \subset G_\beta(g \mid f)$.*

Proof. Take any $g' \in G(g \mid f)$, and fix any state s with $g'(s) \in G(s, g \mid f)$. For such a state, the inequalities (7) and (8) must hold (by definition of the set $G(s, g \mid f)$). These inequalities may be rewritten as follows:

$$\big[P(f, g')\boldsymbol{v}(f, g) - \boldsymbol{v}(f, g)\big]_s \leq 0 \tag{10}$$

and if this holds with equality, then

$$\big[\boldsymbol{r}(f, g') + P(f, g')\boldsymbol{y}(f, g) - \boldsymbol{v}(f, g) - \boldsymbol{y}(f, g)\big]_s < 0. \tag{11}$$

We want to show that for $\beta < 1$ sufficiently close to 1, $g'(s)$ is also an element of $G_\beta(s, g \mid f)$. Recall that this is so precisely when (1) holds, or, equivalently, when

$$\big[\boldsymbol{r}(f,g') + \beta P(f,g')\boldsymbol{v}_\beta(f,g) - \boldsymbol{v}_\beta(f,g)\big]_s < 0. \tag{12}$$

But by Theorem 3.4 we have

$$\boldsymbol{v}_\beta(f,g) = \frac{\boldsymbol{v}(f,g)}{1-\beta} + \boldsymbol{y}(f,g) + \boldsymbol{\epsilon}(\beta,f,g),$$

where $\boldsymbol{\epsilon}(\beta,f,g)$ goes to zero in all coordinates at β increases to 1. Using this, we may rewrite the condition (12) as

$$\begin{aligned}\frac{1}{1-\beta}\big[P(f,g')\boldsymbol{v}(f,g) - \boldsymbol{v}(f,g)\big]_s \\ + \big[\boldsymbol{r}(f,g') + P(f,g')\boldsymbol{y}(f,g) - P(f,g')\boldsymbol{v}(f,g) - \boldsymbol{y}(f,g')\big]_s \\ + \big[(\beta-1)P(f,g')\boldsymbol{y}(f,g) + \beta P(f,g')\boldsymbol{\epsilon}(\beta,f,g,g') - \boldsymbol{\epsilon}(\beta,f,g')\big]_s < 0,\end{aligned} \tag{13}$$

Now if (10) holds strictly, then we may choose β near enough to 1 that (13) holds. If (10) holds with equality, then $\big[P(f,g')\boldsymbol{v}(f,g)\big]_s = v(s,f,g)$ and (11) is strict, and so we again choose β near enough to 1 that the negative value of the second bracketed expression in (13) guarantees the negative value of the entire expression as the third bracketed expression approaches zero.

This argument holds for any state s with $g'(s) \neq g(s)$, so for all $\beta < 1$ near enough to 1, we have $g' \in G_\beta(g \mid f)$. □

We need one more lemma crucial to our main result. Here we show that the set of improvements in the average sense for a strategy g for player 2 is unchanged by changing the fixed strategy for player 1 to another which has the same average payoff and deviation against g.

Lemma 3.9. *Let f^1 and f^2 be pure stationary strategies for player 1 and g a pure stationary strategy for player 2 for a game Γ with $\boldsymbol{v}(f^1,g) = \boldsymbol{v}(f^2,g)$ and $\boldsymbol{y}(f^1,g) = \boldsymbol{y}(f^2,g)$. Then $G(g \mid f^1) = G(g \mid f^2)$.*

Proof. Since the roles of f^1 and f^2 are completely interchangeable, it suffices to show that $G(g \mid f^1) \subseteq G(g \mid f^2)$. Write $\boldsymbol{v}$ for the common value of $\boldsymbol{v}(f^i,g)$, $i = 1,2$ and $\boldsymbol{y}$ for the common value of $\boldsymbol{y}(f^i,g)$, $i = 1,2$.

By the definition of the set $G(g \mid f^1)$, for a strategy g' contained in this set and any state s with $g'(s) \neq g(s)$, we must have

$$\sum_{t=1}^{N} p\big(t \mid s, f^1(s), g'(s)\big)v(t) \leq v(s) \tag{14}$$

and, if this holds with equality,

$$r\big(s, f^1(s), a\big) + \sum_{t=1}^{N} p\big(t \mid s, f(s), g'(s)\big)y(t) < v(s) + y(s) \tag{15}$$

Now observe that by Lemma 3.5, we have

$$P(f^i,g)\boldsymbol{v} = \boldsymbol{v} \text{ and } \boldsymbol{r}(f^i,g) + P(f^i,g)\boldsymbol{y} = \boldsymbol{v} + \boldsymbol{y}$$

for $i = 1,2$, and so

$$P(f^1,g)\boldsymbol{v} = P(f^2,g)\boldsymbol{v} \tag{16}$$

and

$$\boldsymbol{r}(f^1,g) + P(f^1,g)\boldsymbol{y} = \boldsymbol{r}(f^2,g) + P(f^2,g)\boldsymbol{y}. \tag{17}$$

Now fix an s with $g'(s) \neq g(s)$. The sth row of (16) is

$$\sum_{s'=1}^{N} p\big(s' \mid s, f^1(s), g(s)\big)v(s') = \sum_{s'=1}^{N} p\big(s' \mid s, f^2(s), g(s)\big)v(s')$$

Using the ARAT property of the game,

$$\sum_{s'=1}^{N} p^1\big(s' \mid s, f^1(s)\big)v(s') + \sum_{s'=1}^{N} p^2\big(s' \mid s, g(s)\big)v(s') \\ = \sum_{s'=1}^{N} p^1\big(s' \mid s, f^2(s)\big)v(s') + \sum_{s'=1}^{N} p^2\big(s' \mid s, g(s)\big)v(s')$$

and so

$$\sum_{s'=1}^{N} p^1\big(s' \mid s, f^1(s)\big)v(s') = \sum_{s'=1}^{N} p^1\big(s' \mid s, f^2(s)\big)v(s')$$

and we can add $\sum_{s'=1}^{N} p^2\big(s' \mid s, g'(s)\big)v(s')$ to both sides of this equation, yielding

$$\sum_{s'=1}^{N} p\big(s' \mid s, f^1(s), g'(s)\big)v(s') = \sum_{s'=1}^{N} p\big(s' \mid s, f^2(s), g'(s)\big)v(s') \tag{18}$$

Similarly, from the sth row of (17) and the ARAT property, we obtain

$$r(s, f^1(s), a) + \sum_{s'=1}^{n} p\big(s' \mid s, f^1(s), a\big)y(s') \\ = r(s, f^2(s), a) + \sum_{s'=1}^{n} p\big(s' \mid s, f^2(s), a\big)y(s') \tag{19}$$

By virtue of (18) and (19), we can replace all superscripts of 1 with 2 in (14) and (15) for any s with $g'(s) \neq g(s)$. This is precisely the requirement for g' to be an element of $G(g \mid f^2)$. This shows $G(g \mid f^1) \subseteq G(g \mid f^2)$, and so the lemma. □

We now come to our main result. The bulk of the work for this result has already been done in the preceding three lemmas.

Theorem 3.10. *Algorithm 1 will terminate, and the returned strategy pair (f^*, g^*) is average optimal for the game Γ.*

Proof. Suppose that the algorithm does not stop after the kth time through the while loop. Consider the strategy pairs (f^{k+1}, g^k) after the $(k+1)$-st execution of line 3, so that $G(f^{k+1} \mid g^k) \cup H(f^{k+1} \mid g^k)$ is empty. Next the algorithm will choose $g^{k+1} \in G(g^k \mid f^{k+1})$ for player 2. We will show the following monotonicity claim: for any discount factor β sufficiently close to 1, $\boldsymbol{v}_\beta(\Gamma|_{g^k}) > \boldsymbol{v}_\beta(\Gamma|_{g^{k+1}})$.

Let $\hat{f}^{k+1}$ be a uniform optimal strategy for player 1 in the Markov decision process $\Gamma|_{g^k}$. This will not, in general, be the strategy actually chosen by the algorithm. However, by Lemma 3.7, the algorithm's choice of a 1-optimal f^{k+1} has the same average value and deviation as $\hat{f}^{k+1}$ against g^k. Therefore, by Lemma 3.9, $G(g^k \mid f^{k+1})$ and $G(g^k \mid \hat{f}^{k+1})$ are in fact the *same set*, so g^{k+1} is an improvement for g^k against any uniform optimal choice for player 1. Finally, by Lemma 3.8,

$G(g^k \mid \hat{f}^{k+1}) \subseteq G_\beta(g^k \mid \hat{f}^{k+1})$ for all $\beta < 1$ sufficiently close to 1. Putting this all together, we have that for all $\beta < 1$ sufficiently close to 1

$$g^{k+1} \in G(g^k \mid f^{k+1}) = G(g^k \mid \hat{f}^{k+1}) \subseteq G_\beta\big(g^k \mid \hat{f}^{k+1}\big).$$

Therefore, since $g^{k+1} \in G_\beta(g^k \mid f^{k+1})$ for all $\beta < 1$ sufficiently close to 1, the Patience Theorem (Theorem 2.3) proves our monotonicity claim. Since there are only a finite number of pure stationary strategies available to player 2, this monotonicity demands that the algorithm must terminate, returning (f^*, g^*). When the algorithm terminates, $G(f^* \mid g^*)$ and $G(g^* \mid f^*)$ are both empty. The average optimality of the pair then follows from Corollary 1. □

4. **Implementation of the algorithms and numerical results.** A prototype implementation of Algorithm 1, along with its discounted-case analogue, written in Python, can be found in the repository at https://github.com/mattjbourque/stochgame. The Python module includes various functions supporting these algorithms, including a class for representing stochastic matrices with a method for computing the Cesaro limit of a stochastic matrix, as well as functions for generating random stochastic games and for reading and writing stochastic game data to text files.

When choosing an element of $G(f \mid g)$ or $G(g \mid f)$, one has to decide whether to choose only adjacent strategies (that is, those for which only one action is changed) or to allow choices which make changes in more than one state. Experiments on randomly generated games suggest that changing strategies in several states at once is usually faster, but it seems likely that there are examples for which restricting to adjacent improvements would be faster. Our current implementation is not restricted to adjacent improvements.

Each time a strategy is updated to a pair (f, g), we must compute the Cesaro limit matrix in order to find the average payoff $\boldsymbol{v}(f,g) = P^*(f,g)\boldsymbol{r}(f,g)$ as well as the deviation vector $\boldsymbol{y}(f,g) = D(f,g)\boldsymbol{r}(f,g)$ and the vector $\boldsymbol{z}(f,g) = -D^2(f,g)\boldsymbol{r}(f,g)$. In order to compute $P^*(f,g)$ we use use the algorithm in [8] for computing the ergodic classes of $P(f,g)$.

Furthermore, the Python implementation follows the suggestion in Veinott [20] in implementing the policy improvement algorithm for one-optimal policies for MDPs at line 3: we do not check condition (9) until we arrive at a strategy for player 1 for which neither (7) nor (8) holds strictly for any action a for player 1 in any state. In other words, the algorithm is limited at first to finding a strategy for player 1 which is average optimal, and only when this is achieved does it attempt to progress to a 1-optimal strategy. Empirical results suggest that this is a more efficient way to proceed, and that for randomly generated games, it is rarely necessary to use condition (9).

4.1. **Numerical results.** Figure 4.1 displays the results of a applying Algorithm 1 to randomly generated ARAT games. Each game in the simulation has five actions available for each player. The number of states varies from 5 to 50. Ten games of each size were solved. Figure 4.1 shows the average number of iterations required, counting all iterations required by Veinott's policy iteration algorithm for player 1 in executing line 3 and for player 2 in line 5, with bars displaying the maximum and minimum number of iterations required at each games size.

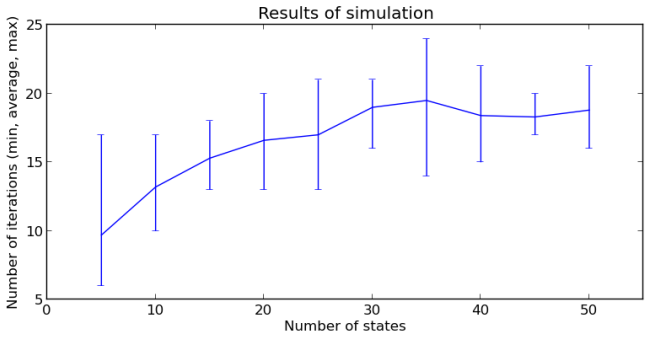

FIGURE 1. Maximum, minimum, and average number of iterations required for solving ARAT games with 5 actions for each player and from 5 - 50 states.

Acknowledgments. We are indebted to Stéphane Gaubert for pointing out an error in an earlier version of the paper, and to Lorenzo Maggi for bringing to our attention his related work with Konstantin Avrachenkov and Laura Cottatellucci on computing uniform optimal strategies. Our thanks also go to the anonymous reviewers for their help in clarifying some of the proofs.

REFERENCES

[1] M. Akian, J. Cochet-Terrasson, S. Detournay and S. Gaubert, Policy iteration algorithm for zero-sum multichain stochastic games with mean payoff and perfect information, preprint arXiv:1208.0446

[2] K. Avrachenkov, L. Cottatellucci and L. Maggi, Algorithms for uniform optimal strategies in two-player zero-sum stochastic games with perfect information, *Operations Research Letters*, **40** (2012), 56–60.

[3] D. Blackwell, Discrete dynamic programming, *The Annals of Mathematical Statistics*, **33** (1962), 719–726.

[4] J. Cochet-Terrasson and S. Gaubert, A policy iteration algorithm for zero-sum stochastic games with mean payoff, *Comptes Rendus Mathematique*, **343** (2006), 377–382.

[5] J. Cochet-Terrasson, G. Cohen, S. Gaubert, M. Mc Gettrick and J.-p. Quadrat, Numerical computation of spectral elements in max-plus algebra, in *Proceedings SSC'98 (IFAC Conference on System Structure and Control)*, Pergamon, Nantes, France, (1998), 667–674.

[6] J. A. Filar and B. Tolwinski, On the algorithm of Pollatschek and Avi-ltzhak, in *Stochastic Games And Related Topics* (eds. T. E. S. Raghavan, T. S. Ferguson, T. Parthasarathy, O. J. Vrieze, W. Leinfellner, G. Eberlein and S. H. Tijs), Theory and Decision Library, 7, Springer, Netherlands, 1991, 59–70.

[7] J, A. Filar and K, Vrieze, *Competitive Markov Decision Processes*, Springer, 1997.

[8] B. L. Fox and D. M. Landi, An algorithm for identifying the ergodic subchains and transient states of a stochastic matrix, *Commun. ACM*, **11** (1968), 619–621.

[9] A. Hordijk, R. Dekker and L. C. M. Kallenberg, Sensitivity-analysis in discounted Markovian decision problems, *OR Spektrum*, **7** (1985), 143–151.

[10] R. A. Howard, *Dynamic Programming and Markov Process*, The Technology Press of M.I.T., Cambridge, Mass.; John Wiley & Sons, Inc., New York-London, 1960.

[11] R. G. Jeroslow, Asymptotic linear programming, *Operations Research*, **21** (1973), 1128–1141.

[12] J. G. Kemeny and J. L. Snell, *Finite Markov Chains*, The University Series in Undergraduate Mathematics, D. Van Nostrand Co., Inc., Princeton, N.J.-Toronto-London-New York, 1960.

[13] E. Kohlberg, Invariant half-lines of nonexpansive piecewise-linear transformations, *Mathematics of Operations Research*, **5** (1980), 366–372.

[14] J.-F. Mertens and A. Neyman, Stochastic games have a value, *Proceedings of the National Academy of Sciences of the United States of America*, **79** (1982), 2145–2146.

[15] T. Parthasarathy and T. E. S. Raghavan, An orderfield property for stochastic games when one player controls transition probabilities, *Journal of Optimization Theory and Applications*, **33** (1981), 375–392.
[16] T. E. S. Raghavan, S. H. Tijs and O. J. Vrieze, On stochastic games with additive reward and transition structure, *Journal of Optimization Theory and Applications*, **47** (1985), 451–464.
[17] T. E. S. Raghavan and Z. Syed, A policy-improvement type algorithm for solving zero-sum two-person stochastic games of perfect information, *Mathematical Programming*, **95** (2003), 513–532.
[18] L. S. Shapley, Stochastic games, *Proceedings of the National Academy of Sciences of the United States of America*, **39** (1953), 1095–1100.
[19] Z. U. Syed, *Algorithms for Stochastic Games and Related Topics*, Ph.D thesis, University of Illinois at Chicago, Chicago, IL, USA, 1999.
[20] A. F. Veinott, On finding optimal policies in discrete dynamic programming with no discounting, *The Annals of Mathematical Statistics*, **37** (1966), 1284–1294.

Received December 2012; revised December 2013.

E-mail address: mbourque@luc.edu
E-mail address: terctu@gmail.com