ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ
ΤΟΜΕΑΣ ΣΤΑΤΙΣΤΙΚΗΣ ΚΑΙ ΕΠΙΧΕΙΡΗΣΙΑΚΗΣ
ΕΡΕΥΝΑΣ

Μεταπτυχιακό Δίπλωμα Ειδίκευσης
στη Στατιστική και Επιχειρησιακή Έρευνα

Διπλωματική Εργασία

Αδιαχώριστα Στοχαστικά Παιχνίδια

Παΐζης Γεράσιμος

Επιβλέπων Καθηγητής
Μηλολιδάκης Κωνσταντίνος

Αθήνα 2013

Θα ήθελα να ευχαριστήσω θερμά τον καθηγητή μου, Κωνσταντίνο Μηλολιδάκη για τη βοήθεια του και το χρόνο που αφιέρωσε σε μένα. Ευχαριστώ επίσης τους καθηγητές του μεταπτυχιακού προγράμματος, οι οποίοι ήταν πάντα πρόθυμοι να λύσουν κάθε απορία μου και να μεταδώσουν τις γνώσεις τους. Τέλος θα ήθελα να ευχαριστήσω όλους τους συμφοιτητές μου που βοήθησαν να περατωθεί η εργασία μου.

Γεράσιμος Παΐζης

Στην οικογένειά μου

Περιεχόμενα

Κεφάλαιο 1

Εισαγωγή

1.1 Εισαγωγή

Τα στοχαστικά παιχνίδια εισήγαγε το 1953 ο Shapley. Η περιοχή αυτή έκτοτε έχει επεκταθεί δραστικά, έχει αναπτυχθεί σημαντική θεωρία και έχουν πρόσφατα εκδοθεί αρκετά βιβλία που τη συνοψίζουν (π.χ. J. Filar και K. Vrieze (1997), A. Maitra και W. Sudderth (1996), A. Neyman και S. Sorin (2003)), συνδέοντάς την με τη Θεωρία Πιθανοτήτων, τις Ελεγχόμενες Διαδικασίες Markov κ.λ.π. Σε αυτή την εργασία θα ασχοληθούμε με ορισμένες περιοχές από τα στοχαστικά παιχνίδια 2-παιχτών μηδενικού αθροίσματος. Η εργασία αποτελείται από τέσσερα κεφάλαια.

Στο πρώτο κεφάλαιο περιγράφουμε το μοντέλο ενός στοχαστικού παιχνιδιού 2-παιχτών μηδενικού αθροίσματος και τους διάφορους τύπους στρατηγικών των παιχτών. Στη συνέχεια παρουσιάζουμε το μοντέλο μιας Ελεγχόμενης Διαδικασίας Markov. Έπειτα, περιγράφουμε δύο διαφορετικούς τρόπους υπολογισμού της ολικής πληρωμής ενός στοχαστικού παιχνιδιού. Συγκεκριμένα θα αναφερθούμε στην "αποπληθωρισμένη" ολική πληρωμή και στη "μέση οριακή" πληρωμή των παιχτών ενός στοχαστικού παιχνιδιού. Τέλος, θα δούμε πως μπορεί ένας παίχτης να εξασφαλίσει το μέγιστο δυνατό κέρδος για τον εαυτό του όταν ο άλλος παίχτης σταθεροποιήσει μία στρατηγική του, δηλαδή με άλλα λόγια όταν ο δεύτερος παίχτης έχει δηλώσει από την αρχή του παιχνιδιού τη στρατηγική που θα ακολουθήσει σε όλη τη διάρκεια του παιχνιδιού.

Στο δεύτερο κεφάλαιο αναφέρουμε εν περιλήψει χρήσιμα συμπεράσματα από τη Μαρκοβιανή θεωρία και πιο συγκεκριμένα κάποια αποτελέσματα για τις Μαρκοβιανές αλυσίδες, για τις ιδιότητες των καταστάσεων του στοχαστι-

κού παιχνιδιού, για τον πίνακα πιθανοτήτων μεταβάσεων των καταστάσεων ενός συστήματος (στοχαστικό παιχνίδι), για το Σέζαρο-όριο του πίνακα πιθανοτήτων μεταβάσεων των καταστάσεων και τέλος για τη στάσιμη κατανομή.

Στο τρίτο κεφάλαιο μελετούμε τα αποπληθωρισμένα στοχαστικά παιχνίδια. Η θεωρία των αποπληθωρισμένων στοχαστικών παιχνιδιών εισήχθηκε το 1953 από τον Shapley ο οποίος μελέτησε στοχαστικά παιχνίδια με θετική πιθανότητα σταματήματος. Τα αποπληθωρισμένα στοχαστικά παιχνίδια μπορούν να θεωρηθούν ως ειδική περίπτωση των στοχαστικών παιχνιδιών με θετική πιθανότητα σταματήματος. Ο Shapley απέδειξε ότι ένα τέτοιο στοχαστικό παιχνίδι έχει τιμή και ότι και οι δύο παίχτες έχουν βέλτιστες (στάσιμες) στρατηγικές. Εμείς, θα δείξουμε την υπάρξη τιμής και βέλτιστων (στάσιμων) στρατηγικών για τα αποπληθωρισμένα στοχαστικά παιχνίδια. Στη συνέχεια θα δώσουμε παραδείγματα αποπληθωρισμένων στοχαστικών παιχνιδιών και τέλος θα αναφερθούμε εν συντομία στην οριακή αποπληθωρισμένη εξίσωση, η οποία αποτελεί χρήσιμο εργαλείο για τη μελέτη της ασυμπτωτικής συμπεριφοράς των λύσεων των αποπληθωρισμένων στοχαστικών παιχνιδιών όταν ο "συντελεστής αποπληθωρισμού" είναι πολύ κοντά στο 1.

Στο τέταρτο κεφάλαιο της εργασίας θα μελετήσουμε στοχαστικά παιχνίδια 2-παιχτών μηδενικού αθροίσματος με κριτήριο πληρωμής την οριακή μέση πληρωμή. Τη θεωρία αυτών των παιχνιδιών εισήγαγε το 1957 ο Gillette. Για ένα μεγάλο χρονικό διάστημα ήταν ανοιχτό ερώτημα εάν αυτά τα παιχνίδια διαθέτουν τιμή. Το ερώτημα απαντήθηκε ανεξάρτητα από τους Mertens και Neyman (1980) και Monash (1979). Στη συγκεκριμένη εργασία θα ασχοληθούμε με τα αδιαχώριστα στοχαστικά παιχνίδια 2-παιχτών μηδενικού αθροίσματος, δηλαδή στοχαστικά παιχνίδια που μπορούν από οποιαδήποτε κατάσταση να βρεθούν σε οποιαδήποτε άλλη κατάσταση με θετική πιθανότητα οπότε οι παίχτες περιορίζονται σε στάσιμες στρατηγικές. Θα αποδείξουμε την ύπαρξη τιμής και βέλτιστων (στάσιμων) στρατηγικών αυτών των παιχνιδιών, θα κάνουμε μια σύνδεση με την οριακή αποπληθωρισμένη εξίσωση και τέλος θα δώσουμε έναν αλγόριθμο για τον υπολογισμό της τιμής και των βέλτιστων (στάσιμων) στρατηγικών των δύο παιχτών.

1.2 Βασικές Έννοιες

Ένα στοχαστικό παιχνίδι δύο παιχτών μηδενικού-αθροίσματος είναι μια καθορισμένη πεντάδα $\langle S, \{A^1(s), s \in S\}, \{A^2(s), s \in S\}, r, p\rangle$, όπου $S, A^1(s)$ και $A^2(s)$ είναι πεπερασμένα μη κενά σύνολα, r είναι μία συνάρτηση πάνω στο σύνολο $H := \{(s, a^1, a^2), s \in S, a^1 \in A^1(s), a^2 \in A^2(s)\}$ και p είναι μία απεικόνιση $p : H \to \mathcal{P}(S)$ όπου $\mathcal{P}(S)$ είναι η οικογένεια των κατανομών πιθανότητας πάνω στο χώρο S.

Τα παραπάνω σύνολα έχουν τις εξής ονομασίες (που επίσης ερμηνεύουν τη λειτουργία τους).

- Το $S = \{1, \ldots, N\}$ ονομάζεται χώρος καταστάσεων.
- Το $A^1(s) = \{1, 2, \ldots, m^1(s)\}$ ονομάζεται χώρος αποφάσεων του παίχτη I όταν το παιχνίδι βρίσκεται στην κατάσταση s.
- Το $A^2(s) = \{1, 2, \ldots, m^2(s)\}$ ονομάζεται χώρος αποφάσεων του παίχτη II όταν το παιχνίδι βρίσκεται στην κατάσταση s.
- Η συνάρτηση $r : H \to \mathbb{R}$ ονομάζεται συνάρτηση πληρωμής. Αν στην κατάσταση s ο παίχτης I αποφασίσει $a^1 \in A^1(s)$ και ο παίχτης II αποφασίσει $a^2 \in A^2(s)$, τότε ο παίχτης II πληρώνει στον I το ποσό $r(s, a^1, a^2)$. Αν $r(s, a^1, a^2) < 0$ τότε ο παίχτης II λαμβάνει $-r(s, a^1, a^2)$ από τον παίχτη I.
- Η συνάρτηση $p : H \to \mathcal{P}(S)$ ονομάζεται απεικόνιση μεταβάσεων ή νόμος κίνησης του συστήματος. Το $\mathcal{P}(S)$ μπορεί να ταυτιστεί με το σύνολο

$$\left\{ x \mid x \in \mathbb{R}^N, x_s \geq 0 \, \gamma\iota\alpha \, \kappa\alpha\theta\epsilon \, s \in S \, \kappa\alpha\iota \sum_{s=1}^{N} x_s = 1 \right\}$$

  Επιπλέον, για κάθε $(s, a^1, a^2) \in H$, προσδιορίζουμε το $\mathbf{p}(s, a^1, a^2)$ με το διάνυσμα $(p(1 \mid s, a^1, a^2), p(2 \mid s, a^1, a^2), \ldots, p(N \mid s, a^1, a^2))$. Η ποσότητα $p(s' \mid s, a^1, a^2)$ αντιπροσωπεύει την πιθανότητα το σύστημα να μεταπηδά στην κατάσταση s' αν στην κατάσταση s ο παίχτης I αποφασίσει $a^1 \in A^1(s)$ και ο παίχτης II αποφασίσει $a^2 \in A^2(s)$. Ισχύει ότι $p(s' \mid s, a^1, a^2) \geq 0$ για κάθε $s' \in S$ και $\sum_{s'=1}^{N} p(s' \mid s, a^1, a^2) = 1$. Στην παρούσα εργασία την πιθανότητα $p(s' \mid s, a^1, a^2)$ θα τη συμβολίζουμε $p_{ss'}(a^1, a^2)$.

Ένα στοχαστικό παιχνίδι εξετάζεται σε διακριτές χρονικές στιγμές. Σε αυτές τις χρονικές στιγμές και οι δύο παίχτες παίρνουν αποφάσεις και έτσι επηρεάζουν την εξέλιξη του παιχνιδιού. Οι χρονικές στιγμές καλούνται στιγμές αποφάσεων ή στάδια. Εμείς θα θεωρήσουμε ότι το παιχνίδι παίζεται σε άπειρο ορίζοντα και έτσι το σύνολο των στιγμών αποφάσεων ταυτίζεται με το σύνολο $\mathbb{N} = \{0, 1, 2, \dots\}$.

Το παιχνίδι παίζεται ως εξής. Υποθέτουμε ότι η αρχική κατάσταση s_0 του παιχνιδιού τη στιγμή απόφασης 0 είναι γνωστή και στους δύο παίχτες. Οι παίχτες παίρνουν ταυτόχρονα και ανεξάρτητα ο ένας από τον άλλον μία απόφαση $a_0^1 \in A^1(s_0)$ και $a_0^2 \in A^2(s_0)$ αντίστοιχα. Τώρα δύο πράγματα συμβαίνουν και τα δύο εξαρτώνται από την παρούσα κατάσταση s_0 και τις ταυτόχρονα επιλεγμένες αποφάσεις a_0^1 και a_0^2.

(i) Καταγράφεται το ποσό $r(s_0, a_0^1, a_0^2)$ λογιστικά ως "πληρωμή" του I από τον II, το οποίο καλείται *τρέχουσα πληρωμή*.

(ii) Το παιχνίδι μεταπηδά στην επόμενη κατάσταση s_1 σύμφωνα με το αποτέλεσμα ενός τυχαίου πειράματος. Η πιθανότητα η επόμενη κατάσταση να ειναι η s' είναι $p_{s_0 s'}(a_0^1, a_0^2)$.

Ακολούθως, πριν την στιγμή απόφασης 1 και οι δύο παίχτες πληροφορούνται τις προηγούμενες αποφάσεις του άλλου παίχτη καθώς και την κατάσταση s_1. Στην στιγμή απόφασης 1, η διαδικασία επαναλαμβάνεται παρόμοια.

Υποθέτουμε ότι το παιχνίδι είναι τέλειας ανάμνησης, δηλαδή σε κάθε στιγμή απόφασης κάθε παίχτης θυμάται όλες τις προεπιλεγείσες αποφάσεις και των δύο παιχτών και όλες τις προηγούμενες καταστάσεις που έχουν συμβεί (με δύο λόγια την *ιστορία* του παιχνιδιού).

Σημειώνουμε ότι για ένα μηδενικού-αθροίσματος στοχαστικό παιχνίδι δύο παιχτών κάθε κατάσταση ταυτίζεται με ένα πινακοπαιχνίδι, με την έννοια ότι η ποσότητα $r(s, a^1, a^2)$ (πιθανώς αρνητική) συμβολίζει το ποσό που πρέπει να πληρώσει ο παίχτης II στον παίχτη I αν το παιχνίδι βρίσκεται στην κατάσταση s και οι παίχτες επιλέξουν a^1, a^2 αντίστοιχα. Βέβαια, στο μοντέλο έχουμε μεταπηδήσεις από πινακοπαιχνίδι σε πινακοπαιχνίδι σύμφωνα με το μέτρο πιθανότητας $p_{s\cdot}(a^1, a^2)$. Έτσι, προκειμένου να πάρει μία απόφαση σε μια συγκεκριμένη κατάσταση ένας παίχτης δεν λαμβάνει υπ' όψην του μόνο

την τρέχουσα πληρωμή αλλά και τις δυνατότητες που έχει από τις πληρωμές μελλοντικών καταστάσεων.

Ένα στοχαστικό παιχνίδι δύο παιχτών (όχι μηδενικού-αθροίσματος) ορίζεται με τον ίδιο τρόπο μου ορίστηκε προηγουμένως το μηδενικού-αθροίσματος στοχαστικό παιχνίδι με τη διαφορά ότι οι πληρωμές των δύο παιχτών είναι διαφορετικές συναρτήσεις και δεν αθροίζουν στο μηδέν. Ο κάθε παίχτης επιθυμεί να μεγιστοποιήσει τη δικία του συνάρτηση πληρωμής $r^k : H \to \mathbb{R}$ με $k = 1, 2$, χωρίς να ενδιαφέρεται για την πληρωμή που θα πάρει ο άλλος παίχτης. Συνεπώς, αν στην κατάσταση s ο παίχτης I αποφασίσει $a^1 \in A^1(s)$ και ο παίχτης II αποφασίσει $a^2 \in A^2(s)$, τότε ο παίχτης I πληρώνεται το ποσό $r^1(s, a^1, a^2)$ και ο παίχτης II πληρώνεται το ποσό $r^2(s, a^1, a^2)$.

Στρατηγικές για τους παίχτες είναι κανόνες (πλήρη σχέδια δράσης) που ορίζουν τον τρόπο λήψης απόφασης σε κάθε δυνατή περίπτωση. Η επιλογή σε μια συγκεκριμένη στιγμή απόφασης μπορεί να εξαρτάται από την ιστορία του παιχνιδιού μέχρι εκείνη τη στιγμή. Επιπλέον, η επιλογή μιας απόφασης μπορεί να συμβαίνει με τυχαιοποιημένο τρόπο, δηλαδή να ορίζεται από ένα διάνυσμα πιθανότητας πάνω στο χώρο αποφάσεων των παιχτών και έτσι η επόμενη απόφαση να είναι το αποτέλεσμα ενός τυχαίου πειράματος σύμφωνα με το συγκεκριμένο διάνυσμα πιθανότητας.

Ορισμός 1.2.1. *Το σύνολο όλων των δυνατών ιστοριών μέχρι τη στιγμή απόφασης t αποτελείται από όλες τις ακολουθίες*

$$h_t = (s_0, a_0^1, a_0^2, s_1, a_1^1, a_1^2, \ldots, s_{t-1}, a_{t-1}^1, a_{t-1}^2)$$

που μπορούν να συμβούν μέχρι τη στιγμή $t, t \geq 1$. Εδώ, s_k συμβολίζει την κατάσταση και a_k^1, a_k^2 τις αποφάσεις των δύο παιχτών αντίστοιχα, τη στιγμή $k, k = 0, 1, \ldots, t-1$.

Επιπροσθέτως, θα συμβολίζουμε με H^t το σύνολο όλων των δυνατών ιστοριών μέχρι τη χρονική στιγμή απόφασης t.

Στη συνέχεια περιγράφουμε διαφορετικούς τύπους στρατηγικών που μπορεί να χρησιμοποιήσει κάποιος παίχτης και μετά θα δώσουμε τον ακριβή ορισμό τους.

Μία *καθαρή στρατηγική* σ^1 του παίχτη I ορίζει για κάθε στιγμή απόφασης t, για κάθε κατάσταση s_t και κάθε ιστορία h_t μία απόφαση πάνω στο χώρο αποφάσεων $A^1(s_t)$ του παίχτη I στην κατάσταση s_t. Ο χώρος των

καθαρών στρατηγικών μπορεί να εφοδιαστεί με κατάλληλη τοπολογία και μπορούν να οριστούν μέτρα πιθανότητας πάνω στο χώρο αυτό. Τα μέτρα αυτά ονομάζονται *μεικτές στρατηγικές*.

Μία *συμπεριφορική στρατηγική* π^1 του παίχτη I ορίζει για κάθε στιγμή απόφασης t, για κάθε κατάσταση s_t και για κάθε ιστορία h_t μία κατανομή πιθανότητας $\pi_t^1(h_t, s_t)$ πάνω στο χώρο αποφάσεων $A^1(s_t)$ του παίχτη I στην κατάσταση s_t. Τότε $\pi_t^1(a^1 \mid h_t, s_t)$ είναι η πιθανότητα με την οποία ο παίχτης I διαλέγει την απόφαση $a^1 \in A^1(s_t)$ στη στιγμή t δεδομένου της κατάστασης s_t και της ιστορίας h_t.

Ειδική, απλούστερη μορφή συμπεριφορικής στρατηγικής για τον παίχτη I αποτελεί η *ημι-μαρκοβιανή στρατηγική*, η οποία εξαρτάται από την ιστορία μόνο μέσω της αρχικής κατάστασης s_0 και της χρονικής στιγμής t και έτσι θα είναι της μορφής $\pi_t^1(s_0, s_t)$.

Μία *μαρκοβιανή στρατηγική* για τον παίχτη I είναι μία ημι-μαρκοβιανή στρατηγική που δεν εξαρτάται από την αρχική κατάσταση s_0 και έτσι θα είναι της μορφής $\pi_t^1(s_t)$.

Μία *στάσιμη στρατηγική* για τον παίχτη I είναι μία μαρκοβιανή στρατηγική που δεν εξαρτάται από τη στιγμή απόφασης t, δηλαδή είναι της μορφής $\pi^1(s_t)$. Τη στάσιμη στρατηγική για τον παίχτη I θα τη συμβολίζουμε με $\mathbf{f}$. Τότε, $\mathbf{f} = (f(1), f(2), \ldots, f(N))$, όπου $f(s)$ είναι το μέτρο πιθανότητας πάνω στο χώρο αποφάσεων $A^1(s)$ για κάθε $s \in S$. Έτσι, $f(s) \in \mathcal{P}(A^1(s))$. Αν ο παίχτης I αποφασίσει να παίξει τη στάσιμη στρατηγική $\mathbf{f}$, τότε κάθε στιγμή που το παιχνίδι βρίσκεται στην κατάσταση s, ο παίχτης I θα παίρνει την καθαρή απόφαση σύμφωνα με την $f(s)$.

Μία στάσιμη στρατηγική θα λέγεται *ντετερμινιστική* αν $f(s)$ είναι καθαρή απόφαση για κάθε $s \in S$, δηλαδή δίνει βάρος πιθανότητας 1 σε μια συγκεκριμένη απόφαση $a_s^1 \in A^1(s)$ και 0 στις υπόλοιπες.

Οι στρατηγικές για τον παίχτη II ορίζονται ανάλογα. Για τον παίχτη II, η συμπεριφορική στρατηγική θα συμβολίζεται με π^2 και η στάσιμη στρατηγική με $\mathbf{g}$.

Παρακάτω δίνουμε αυστηρά τους ορισμούς των προαναφερθέντων στρατηγικών.

Ορισμός 1.2.2. *Μία συμπεριφορική στρατηγική* π^1 *για τον παίχτη* I *είναι μία ακολουθία* $\pi_0^1, \pi_1^1, \pi_2^1, \ldots$ *όπου* $\pi_0^1 \in \bigtimes_{s=1}^{N} \mathcal{P}(A^1(s))$ *και* $\pi_t^1 : H_t \to \bigtimes_{s=1}^{N} \mathcal{P}(A^1(s))$ *για* $t \geq 1$.
Μία ημι-μαρκοβιανή στρατηγική μ *για τον παιχτη* I *είναι μία ακολουθία* $\mu_0, \mu_1, \mu_2, \ldots$ *όπου* $\mu_0 \in \bigtimes_{s=1}^{N} \mathcal{P}(A^1(s))$ *και* $\mu_t : S \to \bigtimes_{s=1}^{N} \mathcal{P}(A^1(s))$ *για* $t \geq 1$.
Μία μαρκοβιανή στρατηγική μ *για τον παιχτη* I *είναι μία ακολουθία* $\mu_0, \mu_1, \mu_2, \ldots$ *όπου* $\mu_t \in \bigtimes_{s=1}^{N} \mathcal{P}(A^1(s))$ *για* $t \geq 0$.
Μία στάσιμη στρατηγική **f** *για τον παιχτη* I *είναι ένα στοιχείο του* $\bigtimes_{s=1}^{N} \mathcal{P}(A^1(s))$.
Μία ντετερμινιστική στάσιμη στρατηγική **f** *για τον παιχτη* I *είναι ένα στοιχείο του* $\bigtimes_{s=1}^{N} A^1(s)$.
Οι στρατηγικές για τον παίχτη II *ορίζονται ανάλογα*.

Πρέπει να σημειωθεί ότι το σύνολο των συμπεριφορικών στρατηγικών δεν είναι το πιο γενικό σύνολο στρατηγικών που υπάρχει. Εάν αναπαραστήσουμε ένα στοχαστικό παιχνίδι σε εκτεταμένη μορφή[1], τότε αυτό μπορεί να παράγει ένα δέντρο άπειρου μήκους όταν το παιχνίδι παίζεται σε άπειρο ορίζοντα. Σε αυτό το δέντρο μπορούν να οριστούν καθαρές και μεικτές στρατηγικές[2]. Αυτή η διαδικασία θα οδηγούσε σε μία κλάση στρατηγικών στην οποία το σύνολο των συμπεριφορικών στρατηγικών θα ήταν κατάλληλο υποσύνολο. Παρόλ΄ αυτά, από το θεώρημα του R. Aumann (1964) προκύπτει ότι κάτω από κάποιες κανονιστικές συνθήκες σε παιχνίδια τέλειας πληροφόρησης άπειρου μήκους μπορούμε να περιοριστούμε σε συμπεριφορικές στρατηγικές. Στα στοχαστικά παιχνίδια που εξετάζουμε, οι υποθέσεις του θεωρήματος Aumann ισχύουν και επομένως στο εξής θα περιοριζόμαστε σε συμπεριφορικές στρατηγικές.

Δύο στρατηγικές ενός παίχτη καλούνται *ισοδύναμες*, αν για όλες τις αποφάσεις του άλλου παίχτη και για κάθε αρχική κατάσταση, και οι δύο στρατηγικές αποδίδουν, σε κάθε στιγμή απόφασης, την ίδια αναμενόμενη πληρωμή στον παίχτη.

[1] ο αναγνώστης μπορεί να ανατρέξει στους Von Neuman και Morgenstern (1944)
[2] ο αναγνώστης μπορεί να ανατρέξει στους Kuhn (1953) και Aumann (1964)

1.3 Διαδικασίες Αποφάσεων Markov (MDP)

Στην εξέταση των στοχαστικών παιχνιδιών, που είναι το αντικείμενο της παρούσας εργασίας, θα χρειαστούμε ορισμένα στοιχεία από τις διαδικασίες αποφάσεων Markov, τις οποίες εισάγουμε σε αυτή την παράγραφο.

Ορισμός 1.3.1. *Μία πεπερασμένη Μαρκοβιανή διαδικασία λήψης αποφάσεων είναι μία καθορισμένη τετράδα* $\langle S, \{A(s), s \in S\}, r, p\rangle$*, όπου το σύνολο* S *είναι πεπερασμένο και ονομάζεται χώρος καταστάσεων, το σύνολο* $A(s)$ *είναι επίσης πεπερασμένο και καλείται χώρος αποφάσεων όταν το σύστημα βρίσκεται στην κατάσταση* s*,* r *είναι μια συνάρτηση τρέχουσας πληρωμής και* p *η απεικόνιση μεταβάσεων. Μία Μαρκοβιανή Διαδικασία Αποφάσεων (Markov Decision Process) θα καλείται για συντομία,* MDP*.*

Η ερμηνεία των παραπάνω παραμέτρων είναι ίδια με αυτή που δόθηκε για το στοχαστικό παιχνίδι δύο παιχτών μηδενικού-αθροίσματος με μόνη διαφορά ότι τώρα έχουμε έναν ελεγκτή της διαδικασίας (αντί δύο παίχτες). Σε διακριτές χρονικές στιγμές (στιγμές απόφασης ή στάδια) η εξέλιξη της διαδικασίας επηρεάζεται από μία επιλεχθείσα απόφαση, η οποία ανήκει στο σύνολο αποφάσεων που εξαρτάται από την παρούσα κατάσταση. Αυτή η απόφαση έχει ως αποτέλεσμα μία άμεση πληρωμή και καθορίζει την επόμενη κατάσταση σύμφωνα με ένα τυχαίο πείραμα. Αυτό το τυχαίο πείραμα εξαρτάται μόνο από την παρούσα κατάσταση και από την επακόλουθη απόφαση που πήρε ο ελεγκτής της διαδικασίας. Όπως και στα στοχαστικά παιχνίδια, το σύνολο των στιγμών αποφάσεων ταυτίζεται με το σύνολο $\mathbb{N} = \{0, 1, 2, \dots\}$.

Επομένως, μία Μαρκοβιανή διαδικασία αποφάσεων (MDP) μπορεί να θεωρηθεί ως στοχαστικό παιχνίδι με ένα μόνο παίχτη.

Οι στρατηγικές για τη MDP ορίζονται με ανάλογο τρόπο όπως στα στοχαστικά παιχνίδια αλλά θα αναφέρονται ως πολιτικές, όπως έχει επικρατήσει στη βιβλιογραφία. Όμως, θα κρατήσουμε τον όρο "στρατηγικές" όταν αναφερόμαστε στα στοχαστικά παιχνίδια. Οι διαφορετικοί τύποι πολιτικών που θα μας χρειαστούν εδώ είναι συμπεριφορικές (π), ημιμαρκοβιανές ή μαρκοβιανές (μ), στάσιμες (f) και ντετερμινιστικές στάσιμες (σ).

Όπως θα δούμε και για τα στοχαστικά παιχνίδια στο επόμενο εδάφιο, η ολική πληρωμή είναι συνάρτηση της ροής των τρεχουσών πληρωμών. Έτσι, οι MDP μπορούν να εξεταστούν υπό το πρίσμα διαφόρων κριτηρίων βελτιστότητας ανάλογα με τον τρόπο υπολογισμού της ροής των άμεσων αναμενόμενων πληρωμών. Σε μία MDP, το πρόβλημα που αντιμετωπίζει ο ελεγκτής της

διαδικασίας είναι η μεγιστοποίηση της συνάρτησης ολικής πληρωμής πάνω στο σύνολο των στρατηγικών. Ας δούμε λοιπόν δύο βασικές συναρτήσεις ολικής πληρωμής.

Πρώτα σημειώνουμε ότι στον ορισμό που ακολουθεί, η ποσότητα $\upsilon^t(s_0, \pi)$ ισούται με την αναμενόμενη τρέχουσα πληρωμή κατά τη στιγμή απόφασης t όταν η αρχική κατάσταση του συστήματος είναι η s_0 και ο ελεγκτής ακολουθεί την πολιτική π.

Ορισμός 1.3.2. *Μία αποπληθωρισμένη MDP με συντελεστή αποπληθωρισμού β, όπου $\beta \in (0,1)$, είναι μία μαρκοβιανή διαδικασία λήψης αποφάσεων για την οποία η ολική πληρωμή για κάθε αρχική κατάσταση s_0 υπολογίζεται από τον τύπο*

$$\upsilon_\beta(s_0, \pi) := \sum_{t=0}^{\infty} \beta^t \upsilon^t(s_0, \pi).$$

Ορισμός 1.3.3. *Μία MDP με πληρωμή αποτιμημένη μέσω του οριακού μέσου όρου (για συντομία μέσης πληρωμής) είναι μία μαρκοβιανή διαδικασία αποφάσεων για την οποία η συνολική πληρωμή για κάθε αρχική κατάσταση s_0 υπολογίζεται από τον τύπο:*

$$\upsilon_\alpha(s_0, \pi) := \liminf_{T \to \infty} \frac{1}{T+1} \sum_{t=0}^{T} \upsilon^t(s_0, \pi).$$

Ορισμός 1.3.4. *Έστω $G(s, \pi)$ η συνάρτηση υπολογισμού της πληρωμής για τη MDP για ένα από τα δύο παραπάνω κριτήρια υπολογισμού της πληρωμής. Για $\varepsilon \geq 0$, μία πολιτική π_ε θα λέγεται ε-βέλτιστη, εάν για κάθε $s \in S$:*

$$G(s, \pi_\varepsilon) \geq \sup_\pi G(s, \pi) - \varepsilon$$

Για $\varepsilon = 0$, η πολιτική καλείται βέλτιστη.

Αναφέρουμε τώρα κάποια αποτελέσματα που προκύπτουν από τις MDP. Στα ακόλουθα θεωρήματα τα $r(s, f(s))$ και $p_{ss'}(f(s))$ για μία στάσιμη πολιτική f, ορίζονται ως εξής

$$r(s, f(s)) := \sum_{a \in A(s)} r(s, a) f(s, a)$$

$$p_{ss'}(f(s)) := \sum_{a \in A(s)} p_{ss'}(a) f(s, a)$$

όπου $f(s, a)$ είναι η πιθανότητα ο παίχτης να πάρει την απόφαση $a \in A(s)$ όταν το σύστημα βρίσκεται στην κατάσταση s.

Θεώρημα 1.3.1. *Για μία αποπληθωρισμένη MDP άπειρου ορίζοντα, το διάνυσμα $\mathbf{v}_\beta \in \mathbb{R}^N$, ορισμένο ως $v_\beta(s) := \sup_\pi v_\beta(s,\pi)$ για κάθε κατάσταση $s \in S$, είναι η μοναδική λύση των ακόλουθων εξισώσεων*

$$x(s) = \max_{a \in A(s)} \left\{ r(s,a) + \beta \sum_{s'=1}^{N} p_{ss'}(a) x(s') \right\} \quad , \quad s \in S.$$

Μία στάσιμη πολιτική $\mathbf{f}^* = (f^*(1), f^*(2), \ldots, f^*(N))$ είναι βέλτιστη αν και μόνο αν για κάθε $s \in S$:

$$v_\beta(s) = r(s, f^*(s)) + \beta \sum_{s'=1}^{N} p_{ss'}(f^*(s)) v_\beta(s').$$

Επίσης υπάρχει μία βέλτιστη ντετερμινιστική στάσιμη πολιτική.

Μία απόδειξη αυτού του θεωρήματος μπορεί να βρεθεί στον Blackwell (1965) και στον Derman (1970).

Θεώρημα 1.3.2. *Για μία MDP μέσης πληρωμής άπειρου ορίζοντα, αν υπάρχει συνάρτηση $w(s)$, $s \in S$ και μία σταθερά v τέτοια ώστε*

$$v + w(s) = \max_a \left[r(s,a) + \sum_{s'=1}^{N} p_{ss'}(a) w(s') \right] \qquad (1)$$

τότε υπάρχει μία στάσιμη πολιτική f^* τέτοια ώστε

$$v = v_\alpha(s, f^*) = \max_f v_\alpha(s, f)$$

για κάθε $s \in S$ και η f^* είναι μία οποιαδήποτε πολιτική που για κάθε $s \in S$ επιλέγει μία απόφαση $a \in A(s)$ η οποία μεγιστοποιεί τη δεξιά ποσότητα της (1).

Μία απόδειξη αυτού του θεωρήματος μπορεί να βρεθεί στον Ross (1982), κεφάλαιο 5.

Θεώρημα 1.3.3. *Αν υπάρχει $k < \infty$ και $s_0 \in S$ έτσι ώστε*

$$| v_\beta(s) - v_\beta(s_0) | < k$$

για κάθε $\beta < 1$ και για κάθε κατάσταση $s \in S$, τότε

(i) υπάρχει μία συνάρτηση $w(s)$ και μία σταθερά v που να ικανοποιούν τη σχέση (1) του θεωρήματος 1.3.2

(ii) για μια ακολουθία $\beta_n \to 1$, έχουμε ότι $w(s) = \lim_{n\to\infty} [v_{\beta_n}(s) - v_{\beta_n}(s_0)]$

(iii) $\lim_{\beta\to 1}(1-\beta) v_\beta(s_0) = v$

Μία απόδειξη αυτού του θεωρήματος μπορεί να βρεθεί στον Ross (1982), κεφάλαιο 5.

Επειδή στην παρούσα εργασία θα ασχοληθούμε με αδιαχώριστα στοχαστικά παιχνίδια με πεπερασμένο χώρο καταστάσεων, το επόμενο πόρισμα μας είναι χρήσιμο.

Πόρισμα 1.3.1. *Εάν ο χώρος καταστάσεων του συστήματος είναι πεπερασμένος και κάθε στάσιμη πολιτική επάγει μία αδιαχώριστη Μαρκοβιανή αλυσίδα, τότε η ποσότητα* $v_\beta(s) - v_\beta(s_0)$ *είναι ομοιόμορφα φραγμένη για μια σταθεροποιημένη κατάσταση* $s_0 \in S$ *και για κάθε κατάσταση* $s \in S$ *και έτσι οι συνθήκες του θεωρήματος 1.3.3 ικανοποιούνται.*

Μία απόδειξη αυτού του πορίσματος μπορεί να βρεθεί στον Ross (1982), κεφάλαιο 5.

Πόρισμα 1.3.2. *Έστω* MDP *με πεπερασμένο χώρο καταστάσεων όπου κάθε στάσιμη πολιτική επάγει μία αδιαχώριστη Μαρκοβιανή αλυσίδα. Εάν υπάρχει σταθερά* c *και συνάρτηση* $w(s)$ *έτσι ώστε*

$$\min_a \left[r(s,a) + \sum_{s'=1}^{N} p_{ss'}(a) w(s') \right] \geq c + w(s)$$

για κάθε $s \in S$*, τότε η τιμή* $v := \min_f v_\alpha(s,f)$ *ικανοποιεί την* $v \geq c$*.*

Απόδειξη

Κάτω από τις υποθέσεις (βλ. Derman (1970)) η τιμή v είναι λύση του προβλήματος γραμμικού προγραμματισμού

$$\max u$$

$$\kappa.\,\alpha. \qquad r(s,a) + \sum_{s'=1}^{N} p_{ss'}(a) h(s') \geq u + h(s) \quad \forall s \in S,\ \forall a \in A(s)$$

ως προς τις μεταβλητές $(u, h(1), h(2), \ldots, h(N))$.

$\boxtimes$

1.4 Κριτήρια υπολογισμού της πληρωμής ενός στοχαστικού παιχνιδιού

Για ένα στοχαστικό παιχνίδι 2-παιχτών, ένα ζευγάρι στρατηγικών (π^1, π^2) για μία σταθεροποιημένη αρχική κατάσταση s_0 και κάθε στιγμή απόφασης t επάγει ένα μέτρο πιθανότητας $\mathbb{P}_{s_0\pi^1\pi^2}(t)$ στον πεπερασμένο καρτεσιανό χώρο H^t. Από το θεώρημα επέκτασης του Kolmogorov (1933) η ακολουθία $\mathbb{P}_{s_0\pi^1\pi^2}(0), \mathbb{P}_{s_0\pi^1\pi^2}(1), \ldots$ μπορεί να επεκταθεί σε ένα μοναδικό μέτρο πιθανότητας $\mathbb{P}_{s_0\pi^1\pi^2}$ πάνω στο μη-πεπερασμένο χώρο των άπειρων ιστοριών H^∞.

Δεδομένου ότι οι παίχτες I και II παίζουν τις στρατηγικές π^1 και π^2 αντίστοιχα, ορίζουμε τις ακόλουθες τυχαίες μεταβλητές

$S_{t,\pi^1\pi^2} \equiv S_t$ αντιπροσωπεύει την κατάσταση τη στιγμή απόφασης t.
$A^1_{t,\pi^1\pi^2} \equiv A^1_t$ αντιπροσωπεύει την απόφαση του παίχτη I τη στιγμή απόφασης t.
$A^2_{t,\pi^1\pi^2} \equiv A^2_t$ αντιπροσωπεύει την απόφαση του παίχτη II τη στιγμή απόφασης t.

Ολοφάνερα, οι περιθώριες κατανομές των S_t, A^1_t, A^2_t για κάθε $t \in \mathbb{N}$, καθορίζονται από το $\mathbb{P}_{s_0\pi^1\pi^2}$. Για μία αρχική κατάσταση s_0 η αναμενόμενη πληρωμή στη στιγμή απόφασης t δίνεται από τον τύπο

$$\mathbf{v}^t(s_0, \pi^1, \pi^2) := \mathbb{E}_{s_0}\left\{r(S_t, A^1_t, A^2_t)\right\} =$$

$$\sum_{(s,a^1,a^2)\in H} r(s, a^1, a^2)\mathbb{P}_{s_0\pi^1\pi^2}(S_t = s, A^1_t = a^1, A^2_t = a^2).$$

Ο τρόπος με τον οποίο υπολογίζεται η ροή των πληρωμών μέσω της ολικής πληρωμής προσδιορίζει ένα συγκεκριμένο παιχνίδι. Οι δύο γνωστότεροι τρόποι υπολογισμού της ολικής πληρωμής παρουσιάζονται στη συνέχεια.

Ορισμός 1.4.1. *Ένα αποπληθωρισμένο στοχαστικό παιχνίδι μηδενικού αθροίσματος 2-παιχτών με συντελεστή αποπληθωρισμού $\beta \in (0, 1)$ είναι ένα παιχνίδι στο οποίο η ροή των πληρωμών υπολογίζεται ως εξής*

$$\mathbf{v}_\beta(s_0, \pi^1, \pi^2) := \sum_{t=0}^{\infty} \beta^t \mathbf{v}^t(s_0, \pi^1, \pi^2).$$

Δηλαδή, σύμφωνα με τον παραπάνω ορισμό, το $\mathbf{v}_\beta(s_0, \pi^1, \pi^2)$ ισούται με τη συνολική αποπληθωρισμένη αναμενόμενη πληρωμή του παιχνιδιού όταν ο συντελεστής αποπληθωρισμού είναι $\beta, \beta \in (0,1)$, η αρχική κατάσταση είναι η $s_0 \in S$ και οι παίχτες παίζουν τις στρατηγικές π^1 και π^2 αντίστοιχα. Εφόσον έχουμε υποθέσει ότι ο χώρος καταστάσεων και ο χώρος αποφάσεων είναι πεπερασμένα σύνολα, η τιμή $\mathbf{v}_\beta(s_0, \pi^1, \pi^2)$ υπάρχει για όλα τα ζευγάρια στρατηγικών (π^1, π^2), αφού η σειρά φράσσεται απολύτως.

Ορισμός 1.4.2. *Ένα μέσης πληρωμής στοχαστικό παιχνίδι μηδενικού αθροίσματος 2-παιχτών είναι ένα παιχνίδι στο οποίο η ροή των πληρωμών υπολογίζεται ως εξής*

$$\mathbf{v}_\alpha(s_0, \pi^1, \pi^2) := \liminf_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} \mathbf{v}^t(s_0, \pi^1, \pi^2).$$

Δηλαδή, σύμφωνα με τον παραπάνω ορισμό, το $\mathbf{v}_\alpha(s_0, \pi^1, \pi^2)$ ισούται με το μικρότερο οριακό σημείο της ακολουθίας των μέσων αναμενόμενων πληρωμών όταν η αρχική κατάσταση είναι η $s_0 \in S$ και οι παίχτες παίζουν τις στρατηγικές π^1 και π^2 αντίστοιχα. Προφανώς, η τιμή $\mathbf{v}_\alpha(s_0, \pi^1, \pi^2)$ υπάρχει για όλα τα ζευγάρια στρατηγικών (π^1, π^2). Η τιμή αυτή συχνά καλείται στη βιβλιογραφία *οριακή μέση πληρωμή*.

Οποιαδήποτε αν είναι η συνάρτηση υπολογισμού των πληρωμών, ο παίχτης I θέλει να μεγιστοποιήσει αυτή τη συνάρτηση ενώ ο παίχτης II θέλει να την ελαχιστοποιήσει.

Ορισμός 1.4.3. *Έστω G μία συνάρτηση υπολογισμού των πληρωμών για ένα στοχαστικό παιχνίδι 2-παιχτών μηδενικού-αθροίσματος. Το παιχνίδι θα λέμε ότι έχει τιμή κάτω από αυτήν τη συνάρτηση πληρωμής αν για κάθε αρχική κατάσταση $s_0 \in S$ ισχύει:*

$$\sup_{\pi^1} \inf_{\pi^2} G(s_0, \pi^1, \pi^2) = \inf_{\pi^2} \sup_{\pi^1} G(s_0, \pi^1, \pi^2).$$

Έστω $G^* \in \mathbb{R}^N$ η τιμή ενός παιχνιδιού (που διαθέτει τιμή). Τότε, μία στρατηγική π^1_ε του παίχτη I καλείται ε-βέλτιστη, $\varepsilon \geq 0$, αν για κάθε αρχική κατάσταση $s_0 \in S$:

$$\inf_{\pi^2} G(s_0, \pi^1_\varepsilon, \pi^2) \geq G^*(s_0) - \varepsilon.$$

Για $\varepsilon = 0$ η στρατηγική καλείται βέλτιστη.
Ανάλογα ορίζονται και οι ε-βέλτιστες στρατηγικές του παίχτη II.

Θεώρημα 1.4.1. *Έστω G η συνάρτηση υπολογισμού των πληρωμών για ένα στοχαστικό παιχνίδι 2-παιχτών μηδενικού-αθροίσματος. Εάν υπάρχει διάνυσμα $\mathbf{v} \in \mathbb{R}^N$ και στρατηγικές $\tilde{\pi}^1$ και $\tilde{\pi}^2$ τέτοιες ώστε $G(s_0, \pi^1, \tilde{\pi}^2) \leq \mathbf{v}(s_0) \leq G(s_0, \tilde{\pi}^1, \pi^2)$ για κάθε στρατηγικές π^1 και π^2 των δύο παιχτών και για κάθε αρχική κατάσταση $s_0 \in S$, τότε το $\mathbf{v}$ ισούται με την τιμή του παιχνιδιού και οι στρατηγικές $\tilde{\pi}^1$ και $\tilde{\pi}^2$ είναι βέλτιστες για τους παίχτες I και II αντίστοιχα.*

Θεώρημα 1.4.2. *Έστω G η συνάρτηση υπολογισμού των πληρωμών για ένα στοχαστικό παιχνίδι 2-παιχτών μηδενικού-αθροίσματος. Εάν, για κάθε $\varepsilon > 0$, υπάρχουν στρατηγικές π^1_ε και π^2_ε τέτοιες ώστε για κάθε στρατηγικές π^1 και π^2 των δύο παιχτών και για κάθε αρχική κατάσταση $s_0 \in S$ να ισχύει:*

$$G(s_0, \pi^1, \pi^2_\varepsilon) - \varepsilon \leq G(s_0, \pi^1_\varepsilon, \pi^2_\varepsilon) \leq G(s_0, \pi^1_\varepsilon, \pi^2) + \varepsilon,$$

τότε η τιμή του παιχνιδιού υπάρχει και δεδομένου ότι η αρχική κατάσταση του παιχνιδιού είναι η $s_0 \in S$ η τιμή ισούται με $\lim_{\varepsilon \to 0} G(s_0, \pi^1_\varepsilon, \pi^2_\varepsilon)$.

Οι αποδείξεις αυτών των θεωρημάτων είναι άμεσες (Μηλολιδάκης (2009), κεφάλαιο 4).

1.5 Παίζοντας εναντίον μιας σταθεροποιημένης στρατηγικής

Σε αυτή την παράγραφο εξετάζουμε πως μπορεί ο παίχτης I να επωφεληθεί για να μεγιστοποιήσει την πληρωμή του όταν γνωρίζει τη στρατηγική που σκοπεύει να παίξει ο παίχτης II.

Όταν σε ένα πεπερασμένο μηδενικού-αθροίσματος 2-παιχτών στοχαστικό παιχνίδι ο παίχτης II παίζει μία συμπεριφορική στρατηγική, τότε ο Monash (1979, θεώρημα 1, σελίδα 6) απέδειξε ότι ο παίχτης I μπορεί να περιοριστεί σε μη-τυχαιοποιημένες στρατηγικές. Παρακάτω εξετάζουμε τι συμβαίνει όταν ο παίχτης II σταθεροποιεί μία ημι-μαρκοβιανή ή μία στάσιμη στρατηγική.

Θεώρημα 1.5.1. *Για ένα 2 παιχτών μηδενικού-αθροίσματος στοχαστικό παιχνίδι, έστω ν μία ημι-μαρκοβιανή στρατηγική για τον παίχτη II. Τότε για κάθε συμπεριφορική στρατηγική π^1 του παίχτη I, υπάρχει μία ημι-μαρκοβιανή στρατηγική μ για τον παίχτη I τέτοια ώστε:*

$$\mathbf{v}^t(s_0, \pi^1, \nu) = \mathbf{v}^t(s_0, \mu, \nu)$$

για κάθε αρχική κατάσταση $s_0 \in S$ και στιγμή σπόφασης $t = 0, 1, \ldots$.

Απόδειξη

Έστω ότι η αρχική κατάσταση του παιχνιδιού είναι $s_0 \in S$ και έστω μ μία συμπεριφορική στρατηγική του παίχτη I. Χάρην συντομογραφίας αντί για $\mathbb{P}_{s_0,\pi^1,\nu}, A^1_{t,\pi^1\nu}, A^2_{t,\pi^1\nu}, S_{t,\pi^1\nu}$ θα γράφουμε $\mathbb{P}_{s_0}, A^1_t, A^2_t, S_t$ αντίστοιχα. Για κάθε στιγμή απόφασης $t = 0, 1, \ldots$ και για κάθε $(s, a^1, a^2) \in H$ έχουμε:

$$\mathbb{P}_{s_0}(S_t = s, A^1_t = a^1, A^2_t = a^2) =$$

$$= \mathbb{P}_{s_0}(A^1_t = a^1 \mid S_t = s, A^2_t = a^2)\mathbb{P}_{s_0}(S_t = s, A^2_t = a^2) \qquad (2)$$

Αφού η ν είναι ημι-μαρκοβιανή στρατηγική του παίχτη II, οι τυχαιες μεταβλητές A^1_t, A^2_t δεδομένου των s_0 και s είναι ανεξάρτητες. Τότε έχουμε

$$\mathbb{P}_{s_0}(A^1_t = a^1 \mid S_t = s, A^2_t = a^2) = \mathbb{P}_{s_0}(A^1_t = a^1 \mid S_t = s)$$

Επομένως η σχέση (1) γίνεται:

$$\mathbb{P}_{s_0}(S_t = s, A^1_t = a^1, A^2_t = a^2) =$$

$$= \mathbb{P}_{s_0}(A^1_t = a^1 \mid S_t = s)\mathbb{P}_{s_0}(S_t = s, A^2_t = a^2) \qquad (3)$$

Τώρα, ορίζουμε μία ημι-μαρκοβιανή στρατηγική για τον παίχτη I ως εξής.Αν η αρχική κατάσταση του παιχνιδιού είναι η s_0 και η κατάσταση τη στιγμή απόφασης t είναι η s, τότε ο παίχτης I παίρνει την αποφαση a^1 με πιθανότητα $\mathbb{P}_{s_0}(A_t^1 = a^1 \mid S_t = s)$, δηλαδή $\mu = (\mu_0, \mu_1, \dots)$ όπου $\mu_t = \mathbb{P}_{s_0}(A_t^1 = a^1 \mid S_t = s)$. Για συντομία θα γράφουμε $\mathbb{P}^*_{s_0}$ αντί για $\mathbb{P}^*_{s_0\mu\nu}$. Με επαγωγή θα δείξουμε ότι:

$$\mathbb{P}_{s_0}(S_t = s, A_t^1 = a^1, A_t^2 = a^2) = \mathbb{P}^*_{s_0}(S_t = s, A_t^1 = a^1, A_t^2 = a^2) \qquad (4)$$

Πράγματι, η (3) ισχύει για $t = 0$. Έστω ότι ισχύει για κάποιο t, τότε:

$$\mathbb{P}_{s_0}(S_{t+1} = s') =$$

$$= \sum_{s,a^1,a^2} \mathbb{P}_{s_0}(S_t = s, A_t^1 = a^1, A_t^2 = a^2) p_{ss'}(a^1, a^2) =$$

$$= \sum_{s,a^1,a^2} \mathbb{P}^*_{s_0}(S_t = s, A_t^1 = a^1, A_t^2 = a^2) p_{ss'}(a^1, a^2) =$$

$$= \mathbb{P}^*_{s_0}(S_{t+1} = s') \qquad (5)$$

Επειδή η στρατηγική ν του παίχτη II είναι ημι-μαρκοβιανή από τη σχέση (5) προκύπτει ότι:

$$\mathbb{P}_{s_0}(S_{t+1} = s', A_{t+1}^2 = a^2) = \mathbb{P}^*_{s_0}(S_{t+1} = s', A_{t+1}^2 = a^2) \qquad (6)$$

Από τον ορισμό της ημι-μαρκοβιανής στρατηγικής μ του παίχτη I, τη σχέση (3) για $t+1$ και τη σχέση (6) έχουμε:

$$\mathbb{P}_{s_0}(S_{t+1} = s, A_{t+1}^1 = a^1, A_{t+1}^2 = a^2) =$$

$$\mathbb{P}^*_{s_0}(A_{t+1}^1 = a^1 \mid S_{t+1} = s) \mathbb{P}^*_{s_0}(S_{t+1} = s, A_{t+1}^2 = a^2) =$$
$$= \mathbb{P}^*_{s_0}(S_{t+1} = s, A_{t+1}^1 = a^1, A_{t+1}^2 = a^2)$$

Συνεπώς αποδείξαμε την (4) και άρα έχουμε:

$$\mathbf{v}^t(s_0, \pi^1, \nu) =$$

$$= \sum_{s,a^1,a^2} r(s, a^1, a^2) \mathbb{P}_{s_0\pi^1\nu}(S_t = s, A_t^1 = a^1, A_t^2 = a^2) =$$

$$= \sum_{s,a^1,a^2} r(s, a^1, a^2) \mathbb{P}^*_{s_0\mu\nu}(S_t = s, A_t^1 = a^1, A_t^2 = a^2)$$

$$= \mathbf{v}^t(s_0, \mu, \nu)$$

$\boxtimes$

Θεώρημα 1.5.2. *Θεωρούμε ένα στοχαστικό παιχνίδι 2-παιχτών μηδενικού-αθροίσματος με κριτήριο πληρωμής είτε της αποπληθωρισμένης είτε της μέσης πληρωμής. Υποθέτουμε ότι στο παιχνίδι που οι παίχτες περιορίζονται στο να παίζουν ημι-μαρκοβιανές στρατηγικές, η τιμή υπάρχει. Τότε και για το μη περιορισμένο παιχνίδι η τιμή υπάρχει και ισούται με την τιμή στο περιορισμένο παιχνίδι. Επιπροσθέτως, μία ε-βέλτιστη στρατηγική,$\varepsilon \geq 0$ για ένα παίχτη στο περιορισμένο παιχνίδι είναι και ε-βέλτιστη στρατηγική στο αρχικό παιχνίδι.*

Απόδειξη

Έστω G η συνάρτηση πληρωμής με κριτήριο υπολογισμού της πληρωμής είτε της αποπληθωρισμένης είτε της μέσης πληρωμής για το αρχικό παιχνίδι. Έστω G^* η τιμή του περιορισμένου παιχνιδιού. Έστω ν_ε μία ε-βέλτιστη ημι-μαρκοβιανή στρατηγική του παίχτη II στο περιορισμένο παιχνίδι. Μία τέτοια στρατηγική υπάρχει από τη στιγμή που υπάρχει η τιμή του παιχνιδιού. Από το θεώρημα 1.5.1 για κάθε συμπεριφορική στρατηγική π^1 του παίχτη I, υπάρχει μία ημι-μαρκοβιανή στρατηγική μ τέτοια ώστε για κάθε $s \in S$

$$G(s, \pi^1, \nu_\varepsilon) = G(s, \mu, \nu_\varepsilon).$$

Επίσης, για κάθε $s \in S$ έχουμε

$$G(s, \mu, \nu_\varepsilon) \leq G^*(s) + \varepsilon$$

οπότε

$$G(s, \pi^1, \nu_\varepsilon) \leq G^*(s) + \varepsilon \qquad (7)$$

για κάθε π^1 και $s \in S$. Παρόμοια έχουμε

$$G^*(s) - \varepsilon \leq G(s, \mu_\varepsilon, \pi^2) \qquad (8)$$

για κάθε π^2 και $s \in S$, όπου μ_ε είναι μία ε-βέλτιστη ημι-μαρκοβιανή στρατηγική του παίχτη I στο περιορισμένο παιχνίδι. Αφού $\varepsilon > 0$ είναι αυθαίρετο, εφαρμόζοντας το θεώρημα 1.4.2 σε συνδυασμό με τις σχέσεις (7) και (8) προκύπτει ότι η τιμή του αρχικού παιχνιδιού υπάρχει και ισούται με $\lim_{\varepsilon \to 0} G(s, \mu_\varepsilon, \nu_\varepsilon) = G^*(s), s \in S$.

Επίσης, από τις σχέσεις (7) και (8) προκύπτει ότι οι στρατηγικές μ_ε και ν_ε είναι ε-βέλτιστες στο αρχικό παιχνίδι.

$\boxtimes$

Τώρα ερχόμαστε να αναλύσουμε τι συμβαίνει όταν ένας παίχτης σταθεροποιεί μία στάσιμη στρατηγική του. Έστω ότι ο παίχτης II σταθεροποιεί τη στάσιμη στρατηγική του $\mathbf{g}$. Τότε ορίζουμε το ακόλουθο μαρκοβιανό πρόβλημα αποφάσεων $MDP(\mathbf{g})$ ως εξής:

$$MDP(\mathbf{g}) = \langle \tilde{S}, \{\tilde{A}^1(s), s \in S\}, \tilde{r}, \tilde{p} \rangle$$

όπου

$\tilde{S} := S$
$\tilde{A}^1(s) := A^1(s),\ s \in S$
$\tilde{r}(s, a^1) := \sum_{a^2 \in A^2(s)} r(s, a^1, a^2) g(s, a^2),\ a^1 \in \tilde{A}^1(s),\ s \in \tilde{S}$
$\tilde{p}_{ss'}(a^1) := \sum_{a^2 \in A^2(s)} p_{ss'}(a^1, a^2) g(s, a^2),\ a^1 \in \tilde{A}^1(s),\ s \in \tilde{S}$

Θεώρημα 1.5.3. *Έστω ότι σε ένα στοχαστικό παιχνίδι 2-παιχτών μηδενικού-αθροίσματος ο παίχτης II σταθεροποιεί μία στάσιμη στρατηγική του $\mathbf{g}$. Τότε είτε για το κριτήριο της αποπληθωρισμένης πληρωμής είτε για αυτό της μέσης πληρωμής ο παίχτης I μπορεί να απαντήσει βέλτιστα επιλέγοντας μία βέλτιστη στάσιμη στρατηγική στο $MDP(\mathbf{g})$.*

Απόδειξη

Γνωρίζουμε ότι είτε για το κριτήριο αποπληθωρισμένης πληρωμής (Ross, θεώρημα 2.2, σελίδα 32) είτε της μέσης πληρωμής (Ross, θεώρημα 2.1, σελίδα 93) ότι για ένα MDP μία βέλτιστη στρατηγική μπορει να βρεθεί στη κλάση των στάσιμων στρατηγικών.

Παρατηρούμε ότι υπάρχει ένα προς ένα αντιστοιχία για τον παίχτη I μεταξύ του συνόλου των στάσιμων στρατηγικών στο $MDP(\mathbf{g})$ και του συνόλου των στάσιμων στρατηγικών στο αρχικό παιχνίδι. Έστω $\tilde{G}(\mathbf{f})$ η πληρωμή του I όταν αυτός ακολουθεί την πολιτική $\mathbf{f}$ στο $MDP(\mathbf{g})$ για οποιοδήποτε από τα παραπάνω δύο κριτήρια πληρωμής. Διευκρινίζουμε ότι το $\tilde{G}(\mathbf{f})$ είναι ένα $N-$διάστατο διάνυσμα. Αν μία πολιτική $\mathbf{f}^*$ είναι βέλτιστη στο $MDP(\mathbf{g})$ τότε θα ισχύει:

$$\sup_{\mathbf{f}}(\tilde{G}(\mathbf{f})) = \tilde{G}(\mathbf{f}^*) \qquad (9)$$

Θα αποδείξουμε ότι $G(\mathbf{f}, \mathbf{g}) = \tilde{G}(\mathbf{f})$ για κάθε στάσιμη πολιτική $\mathbf{f}$. Δηλαδή θα αποδείξουμε ότι οποιαδήποτε στάσιμη πολιτική $\mathbf{f}$ επιλέξει ο παίχτης I στο

$MDP(\mathbf{g})$ θα πάρει την ίδια πληρωμή με αυτή που θα έπαιρνε αν έπαιζε την **f** στο αρχικό παιχνίδι δεδομένου ότι ο παίχτης II παίζει στο αρχικό παιχνίδι τη στάσιμη στρατηγική **g**.

Αυτό προκύπτει από το γεγονός ότι η στάσιμη στρατηγική **g** δεν εξαρτάται από την ιστορία του παιχνιδιού εκτός από την παρούσα κατάσταση. Με επαγωγή θα δείξουμε ότι:

$$\mathbb{P}_{s_0\mathbf{fg}}(S_t = s) = \tilde{\mathbb{P}}_{s_0\mathbf{f}}(\tilde{S}_t = s)$$

Πράγματι, ισχύει για $t = 0$ και έστω ότι ισχύει για κάποιο t. Τότε:

$$\mathbb{P}_{s_0\mathbf{fg}}(S_{t+1} = s') =$$

$$\sum_{s\in S} \mathbb{P}_{s_0\mathbf{fg}}(S_t = s)\mathbb{P}_{s_0\mathbf{fg}}(S_{t+1} = s' \mid S_t = s) =$$

$$\sum_{s\in S} \tilde{\mathbb{P}}_{s_0\mathbf{f}}(\tilde{S}_t = s)\mathbb{P}_{s_0\mathbf{fg}}(S_{t+1} = s' \mid S_t = s) =$$

$$\sum_{s\in S} \tilde{\mathbb{P}}_{s_0\mathbf{f}}(\tilde{S}_t = s) \sum_{a^1,a^2} p_{ss'}(a^1, a^2)\mathbb{P}_{s_0\mathbf{fg}}(A_t^1 = a^1 \mid S_t = s)g(s, a^2) =$$

$$\sum_{s\in S} \tilde{\mathbb{P}}_{s_0\mathbf{f}}(\tilde{S}_t = s) \sum_{a^1} \left(\sum_{a^2} p_{ss'}(a^1, a^2)g(s, a^2) \right) \tilde{\mathbb{P}}_{s_0\mathbf{f}}(\tilde{A}_t^1 = a^1 \mid \tilde{S}_t = s) =$$

$$\sum_{s\in S} \tilde{\mathbb{P}}_{s_0\mathbf{f}}(\tilde{S}_t = s) \sum_{a^1} \tilde{p}_{ss'}(a^1)\tilde{\mathbb{P}}_{s_0\mathbf{f}}(\tilde{A}_t^1 = a^1 \mid \tilde{S}_t = s) =$$

$$\tilde{\mathbb{P}}_{s_0\mathbf{f}}(\tilde{S}_{t+1} = s')$$

Συνεπώς έχουμε:

$$\sum_{s,a^1,a^2} r(s, a^1, a^2)\mathbb{P}_{s_0\mathbf{fg}}(S_t = s, A_t^1 = a^1, A_t^2 = a^2) =$$

$$\sum_{s,a^1} \sum_{a^2} r(s, a^1, a^2)g(s, a^2)\mathbb{P}_{s_0\mathbf{fg}}(S_t = s)\mathbb{P}_{s_0\mathbf{fg}}(A_t^1 = a^1 \mid S_t = s) =$$

$$\sum_{s,a^1} \tilde{r}(s, a^1)\tilde{\mathbb{P}}_{s_0\mathbf{f}}(\tilde{S}_t = s)\tilde{\mathbb{P}}_{s_0\mathbf{f}}(\tilde{A}_t^1 = a^1 \mid \tilde{S}_t = s)$$

Η παραπάνω σχέση δείχνει ότι όταν ο παίχτης II χρησιμοποιεί τη στάσιμη στρατηγική **g**, τότε η αναμενόμενη τρέχουσα πληρωμή του παίχτη I είναι ίδια στο στοχαστικό παιχνίδι και στο $MDP(\mathbf{g})$. Επομένως, λόγω του τρόπου

ορισμού τους, είτε χρησιμοποιούμε τον αποπληθωρισμένο τρόπο υπολογισμού της ολικής αμοιβής είτε χρησιμοποιούμε την οριακή μέση πληρωμή, οι ολικές πληρωμές στο στοχαστικό παιχνίδι και στο $MDP(\mathbf{g})$ θα συμπίπτουν, δηλαδή

$$G(\mathbf{f}, \mathbf{g}) = \tilde{G}(\mathbf{f})$$

Επομένως καταλήγουμε έτσι στο

$$\sup_{\mathbf{f}} G(\mathbf{f}, \mathbf{g}) = \sup_{\mathbf{f}} \tilde{G}(\mathbf{f})$$

και από τη σχέση (9) προκύπτει ότι

$$G(\mathbf{f}^*, \mathbf{g}) = \tilde{G}(\mathbf{f}^*)$$

$\boxtimes$

Πόρισμα 1.5.1. *Εάν για ένα στοχαστικό παιχνίδι 2-παιχτών μηδενικού-αθροίσματος το κριτήριο υπολογισμού των πληρωμών είναι είτε της αποπληθωρισμένης είτε της μέσης πληρωμής, τότε η βέλτιστη απάντηση ενός παίχτη απέναντι σε μια σταθεροποιημένη στάσιμη στρατηγική του άλλου παίχτη είναι ντετερμινιστική στάσιμη στρατηγική.*

Η απόδειξη προκύπτει από το προηγούμενο θεώρημα σε συνδυασμό με τα θεωρήματα 1.3.1 και 1.3.2.

Με το προηγούμενο θεώρημα δείξαμε ότι όταν ένας παίχτης περιορίζεται σε στάσιμες στρατηγικές, τότε και ο άλλος παίχτης μπορει να περιοριστεί σε στάσιμες στρατηγικές απαντώντας βέλτιστα στη στάσιμη στρατηγική του πρώτου παίχτη. Επιπλέον έχουμε

Θεώρημα 1.5.4. *Αν ένα στοχαστικό παιχνίδι 2-παιχτών μηδενικού αθροίσματος με συνάρτηση πληρωμής G, είτε της αποπληθωρισμένης είτε της μέσης πληρωμής, έχει ΣΣΙ όταν οι παίχτες περιορίζονται σε στάσιμες στρατηγικές, τότε αυτό είναι ΣΣΙ και στο αρχικό παιχνίδι και οι βέλτιστες στάσιμες στρατηγικές στο περιορισμένο παιχνίδι είναι και βέλτιστες στρατηγικές στο αρχικό παιχνίδι.*

Απόδειξη

Έστω $(\mathbf{f}^*, \mathbf{g}^*)$ ένα ΣΣΙ στο περιορισμένο παιχνίδι. Τότε

$$G(\mathbf{f}, \mathbf{g}^*) \leq G(\mathbf{f}^*, \mathbf{g}^*) \leq G(\mathbf{f}^*, \mathbf{g}) \qquad (10)$$

για κάθε στάσιμες στρατηγικές $\mathbf{f}$ και $\mathbf{g}$ των δύο παιχτών.

Όταν ο παίχτης II περιορίζεται σε στάσιμη στρατηγική, τότε από το θεώρημα 1.5.3 ο παίχτης I απαντάει βέλτιστα με στάσιμη στρατηγική. Θέτοντας $\mathbf{g} = \mathbf{g}^*$ έχουμε

$$G(\mathbf{f}, \mathbf{g}^*) = \max_{\pi^1} G(\pi^1, \mathbf{g}^*) \geq G(\pi^1, \mathbf{g}^*)$$

για κάθε συμπεριφορική στρατηγική π^1. Επομένως, σε συνδυασμό με τη σχέση (10) προκύπτει

$$G(\pi^1, \mathbf{g}^*) \leq G(\mathbf{f}^*, \mathbf{g}^*)$$

για κάθε συμπεριφορική στρατηγική π^1. Αντίστοιχα προκύπτει

$$G(\mathbf{f}^*, \mathbf{g}^*) \leq G(\mathbf{f}^*, \pi^2)$$

για κάθε συμπεριφορική στρατηγική π^2. Άρα το $(\mathbf{f}^*, \mathbf{g}^*)$ είναι ΣΣΙ στο αρχικό παιχνίδι και επιπλέον οι στρατηγικές $\mathbf{f}^*$ και $\mathbf{g}^*$ είναι βέλτιστες στο αρχικό παιχνίδι.

$\boxtimes$

Κεφάλαιο 2

Μαρκοβιανή Θεωρία

2.1 Μαρκοβιανές αλυσίδες

Η ιδέα των μαρκοβιανών αλυσίδων βασίζεται στο ότι αν η περιγραφή της κατάστασης ενός συστήματος σε μια συγκεκριμένη χρονική στιγμή είναι αρκετά πλούσια, τότε η κατάσταση του συστήματος στην επόμενη στιγμή παρατήρησης είναι (κατά μια στοχαστική έννοια) καθορισμένη από αυτή την περιγραφή της κατάστασης. Πιο μαθηματικά, έστω $\mathbb{N}$ το σύνολο των φυσικών αριθμών, αρχίζοντας από το 0, το οποίο θα αναπαριστά τις στιγμές παρατήρησης και έστω $S = \{1, 2, \ldots, N\}$ ο χώρος καταστάσεων του συστήματος. Θεωρούμε ότι για οποιαδήποτε $t \in \mathbb{N}$, το σύστημα βρίσκεται σε κάποια κατάσταση $s \in S$. Θεωρούμε ότι οι μεταπηδήσεις από κατάσταση σε κατάσταση γίνονται τυχαία και έστω S_t, $t = 0, 1, \ldots$ η μεταβλητή που αναπαριστά την κατάσταση του συστήματος τη στιγμή παρατήρησης t.

Ορισμός 2.1.1. *Λέμε ότι η $\{S_t, t = 0, 1, 2, \ldots\}$ είναι μια μαρκοβιανή αλυσίδα αν*

$$\mathbb{P}\{S_{t+1} = s_{t+1} | S_t = s_t, S_{t-1} = s_{t-1}, \ldots, S_0 = s_0\} = \mathbb{P}\{S_t = s_t | S_{t-1} = s_{t-1}\}$$

για οποιαδήποτε $t \geq 1$ και οποιαδήποτε $s_{t+1}, s_t, s_{t-1}, \ldots, s_0 \in S$. Εάν επιπλέον, $\mathbb{P}\{S_{t+1} = s' | S_t = s\}$ δεν εξαρτάται από το t , τότε η μαρκοβιανή αλυσίδα ονομάζεται ομογενής.

Για συντομία θα συμβολίζουμε $\mathbb{P}\{S_{t+1} = s' | S_t = s\} = p_{ss'}$. Αυτές οι πιθανότητες ονομάζονται πιθανότητες μεταβάσεων.

Σε αυτήν την εργασία θα θεωρούμε μόνο συστήματα με πεπερασμένο χώρο καταστάσεων S. Ολοφάνερα, μία μαρκοβιανή αλυσίδα σε ένα τέτοιο

σύστημα καθορίζεται από τους αριθμούς $p_{ss'}, (s'|s) \in S \times S$, και αυτοί οι αριθμοί μπορούν να τοποθετηθούν σε ένα $N \times N$ πίνακα, στον οποίο η γραμμή s θα αποτελείται από τους αριθμούς $p_{s1}, p_{s2}, \ldots, p_{sN}$. Έτσι, στη γραμμή s βρίσκουμε τις πιθανότητες με τις οποίες το σύστημα μεταπηδά σε κάθε κατάσταση όταν η τωρινή κατάσταση είναι η s. Έστω P ο συγκεκριμένος πίνακας πιθανοτήτων μεταβάσεων. Τότε ο P έχει όλα τα στοιχεία του μη αρνητικά και κάθε γραμμή αθροίζει στη μονάδα, δηλαδή είναι ένας στοχαστικός πίνακας. Η ιδιότητα κάθε γραμμή να αθροίζει στη μονάδα είναι συνέπεια του γεγονότος ότι σε κάθε στιγμή παρατήρησης το σύστημα είναι σε μία από τις καταστάσεις του S.

Πρόταση 2.1.1. *Για κάθε $(s, s') \in S \times S$, το στοιχείο (s, s') του πίνακα P^t περιέχει την t-βήματος πιθανότητα $p^{(t)}_{ss'}$, δηλαδή την πιθανότητα το σύστημα να βρεθεί στην κατάσταση s' μετά από t βήματα ξεκινώντας από την κατάσταση s.*

Απόδειξη

Επαγωγή στο t. Για $t = 1$ ισχύει. Έστω ότι ισχύει για $t = n - 1$. Τότε, για $t = n$ έχουμε

$$p^{(n)}_{ss'} = \sum_{k \in S} p_{sk} p^{(n-1)}_{ks'}$$

από όπου συμπαιρένουμε ότι $P^{(n)} = P^n$.

$\boxtimes$

Ορισμός 2.1.2. *Μία κατάσταση s' είναι προσπελάσιμη (προσιτή) από μία κατάσταση s όταν για κάποιο $t \geq 1, p^t_{ss'} > 0$. Επίσης, λέμε ότι οι καταστάσεις s και s' επικοινωνούν όταν η s είναι προσπελάσιμη από την s' και η s' είναι προσπελάσιμη από την s.*

Ορισμός 2.1.3. *Ένα σύνολο καταστάσεων $K \subset S$ ονομάζεται κλειστό, αν οποιαδήποτε κατάσταση $s' \in S \setminus K$ είναι απροσπέλαστη από οποιαδήποτε κατάσταση $s \in K$.*

Ορισμός 2.1.4. *Ένα κλειστό σύνολο ονομάζεται αδιαχώριστο αν όλες οι καταστάσεις του επικοινωνούν.*

Ορισμός 2.1.5. *Αν ένα κλειστό σύνολο αποτελείται μόνο από μία κατάσταση τότε αυτή ονομάζεται κατάσταση απορρόφησης.*

Ορισμός 2.1.6. *Μία κατάσταση ονομάζεται έμμονη αν η πιθανότητα να επιστρέψουμε σε κάποιο βήμα σε αυτήν είναι 1. Διαφορετικά αν η πιθανότητα είναι αυστηρά μικρότερη του 1 ονομάζεται μεταβατική.*

Ορισμός 2.1.7. *Για μία έμμονη κατάσταση αν ο μέσος χρόνος επανόδου στην κατάσταση είναι πεπερασμένος τότε ονομάζεται έμμονη θετική. Διαφορετικά ονομάζεται έμμονη μηδενική.*

Ορισμός 2.1.8. *Περίοδος μιας κατάστασης s είναι ο μέγιστος κοινός διαιρέτης των βημάτων t τέτοια ώστε $p_{ss}^t > 0, t = 1, 2, \dots$. Αν η περίοδος είναι 1 τότε η κατάσταση ονομάζεται απεριοδική.*

Ορισμός 2.1.9. *Μία κατάσταση που είναι έμμονη, θετική και απεριοδική καλείται εργοδική.*

Αφού ο χώρος καταστάσεων S είναι κλειστό σύνολο τότε εξ' ορισμού η μαρκοβιανή αλυσίδα έχει τουλάχιστον ένα αδιαχώριστο σύνολο καταστάσεων. Έστω $S_1, S_2, \dots, S_L$ όλα τα αδιαχώριστα σύνολα καταστάσεων μιας μαρκοβιανής αλυσίδας. Τότε ο πίνακας P μπορεί να γραφτεί στη μορφή

$$\begin{pmatrix} P_1 & 0 & \dots & 0 & 0 \\ 0 & P_2 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & 0 \\ 0 & 0 & \dots & P_L & 0 \\ P_{L+11} & P_{L+12} & \dots & P_{L+1L} & P_{L+1} \end{pmatrix}$$

Οι $P_1, P_2, \dots, P_L$ είναι τετραγωνικοί πίνακες που αντιστοιχούν στις καταστάσεις των συνόλων $S_1, S_2, \dots, S_L$. Επίσης, P_{L+1} είναι ένας τετραγωνικός πίνακας που αντιστοιχεί στις υπόλοιπες καταστάσεις του S_{L+1}. Τα στοιχεία των πινάκων P_{L+1l} είναι πιθανότητες μεταπηδήσεων ενός βήματος στα αδιαχώριστα σύνολα $S_l, l = 1, 2, \dots, L$ αντίστοιχα. Οι καταστάσεις που ανήκουν στα αδιαχώριστα σύνολα είναι έμμονες ενω οι υπόλοιπες καταστάσεις είναι μεταβατικές. Επιπλέον,

$$P^t = \begin{pmatrix} P_1^t & 0 & \dots & 0 & 0 \\ 0 & P_2^t & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & 0 \\ 0 & 0 & \dots & P_L^t & 0 \\ P_{L+11}^{(t)} & P_{L+12}^{(t)} & \dots & P_{L+1L}^{(t)} & P_{L+1}^t \end{pmatrix}$$

όπου $P_{L+1l}^{(t)} = \sum_{k=1}^{t} P_{L+1}^{t-k} P_{L+1l} P_l^{k-1}, l = 1, 2, \dots, L$.

Συμβατικά, ορίζουμε P_l^0 ως το μοναδιαίο πίνακα. Σημειώνουμε ότι για οποιαδήποτε κατάσταση $s \in S_{L+1}$ τουλάχιστον μία από τις καταστάσεις $s' \in S_l$ για κάποιο $l \in \{1, 2, \dots, L\}$ μπορεί να γίνει προσιτή. Σε αντίθετη περίπτωση, το S_{L+1} θα αποτελούσε κλειστό σύνολο καταστάσεων.

2.2 Στάσιμη Κατανομή

Όταν μελετούμε μαρκοβιανές αλυσίδες, ενδιαφερόμαστε κυρίως για το μέσο μακροπρόθεσμο αριθμό επισκέψεων μιας κατάστασης. Ο αναμενόμενος αριθμός επισκέψεων της κατάστασης s' κατά τη διάρκεια T βημάτων όταν το σύστημα ξεκινά από την κατάσταση s ισούται με $\sum_{t=0}^{T} p_{ss'}^t$ όπου $p_{ss'}^0 = 1$ αν $s' = s$ και 0 διαφορετικά και τότε το αναμενόμενο ποσοστό του χρόνου που το σύστημα απασχολείται στην κατάσταση s' ισούται με $\frac{1}{T+1}\sum_{t=0}^{T} p_{ss'}^t$. Έστω $q_{ss'}$ σημείο συσσώρευσης της ακολουθίας $\frac{1}{T+1}\sum_{t=0}^{T} p_{ss'}^t$, που υπάρχει λόγω ακολουθιακής συμπάγιας (τα σημεία της ακολουθίας ανήκουν στο $[0,1]$) και $\mathbf{q}_s = (q_{s1}, \ldots, q_{sN})$. Παρατηρούμε ότι

$$\mathbf{q}_s^T P = \mathbf{q}_s^T$$

διότι

$$\lim_{T\to\infty} \frac{1}{T+1}\sum_{t=0}^{T} P^t = \lim_{T\to\infty} \frac{1}{T+1}\sum_{t=1}^{T} P^t = \left(\lim_{T\to\infty} \frac{1}{T+1}\sum_{t=0}^{T-1} P^t\right) P$$

όποτε υπάρχει το όριο.

Παρατηρούμε επίσης ότι το $\mathbf{q}_s$ είναι ένα διάνυσμα πιθανότητας καθώς

$$\sum_{s'\in S}\left(\frac{1}{T+1}\sum_{t=0}^{T} p_{ss'}^t\right) = 1$$

για όλα τα $T = 0, 1, 2, \ldots$ και όλα τα $s \in S$.

Όταν $s \in S_l$, όπου S_l είναι ένα αδιαχώριστο σύνολο, τότε η εξίσωση $\mathbf{q}_s^T P = \mathbf{q}_s^T$ μπορεί να γραφτεί και ως

$$\mathbf{q}_s^T(l) P_l = \mathbf{q}_s^T(l)$$

όπου $\mathbf{q}_s^T(l)$ είναι ο περιορισμός του $\mathbf{q}_s^T$ στις συνιστώσες που αντιστοιχούν στο σύνολο S_l (ολοφάνερα, $q_{ss'} = 0$ για $s' \notin S_l$.)

Πρόταση 2.2.1. *Η εξίσωση $\mathbf{q}_s^T(l) P_l = \mathbf{q}_s^T(l)$ έχει μοναδική λύση για ένα αδιαχώριστο σύνολο S_l για όλα τα $s \in S_l$, κάτω από τον περιορισμό ότι $\mathbf{q}_s^T(l)$ είναι διάνυσμα πιθανότητας.*

Η απόδειξη της πρότασης δίνεται παρακάτω στο λήμμα 2.2.1 .

Μια σημαντική συνέπεια είναι ότι το $\mathbf{q}_s^T(l)$ είναι το ίδιο για κάθε $s \in S_l$. Έτσι, θα γράφουμε $\mathbf{q}(l)$ για το μοναδικό διάνυσμα πιθανότητας που επιλύει την εξίσωση $\mathbf{q}P_l = \mathbf{q}$. Το διάνυσμα $\mathbf{q}(l)$ ονομάζεται στάσιμη κατανομή του P_l.

Πρόταση 2.2.2.

1. $\lim_{T\to\infty} P_{L+1}^T = 0$
2. $\sum_{t=0}^{\infty} P_{L+1}^t = (I - P_{L+1})^{-1}$

Απόδειξη

1. Επειδή ο πίνακας P_{L+1} περιέχει μεταβατικές καταστάσεις, έχουμε ότι για κάθε γραμμή s του P_{L+1} υπάρχει ακέραιος $T(s)$ τέτοιος ώστε το άθροισμα της s γραμμής του πίνακα $P_{L+1}^{T(s)}$ να είναι αυστηρά μικρότερο του 1. Επομένως, υπάρχει κάποιος αριθμός $\alpha \in (0,1)$ και ένας ακέραιος T τέτοιοι ώστε ο πίνακας P_{L+1}^T να έχει άθροισμα γραμμών το πολύ α για όλες τις γραμμές. Τότε ο πίνακας $\left(P_{L+1}^T\right)^2$ έχει άθροισμα γραμμών το πολύ α^2 κ.ο.κ. οπότε ο πίνακας $\left(P_{L+1}^T\right)^n$ έχει άθροισμα γραμμών το πολύ α^n και καθώς $n \to \infty, \alpha^n \to 0$ οπότε $\left(P_{L+1}^T\right)^n = 0$ και άρα $\lim_{T\to\infty} P_{L+1}^T = 0$.

2. Προκύπτει από τις ισότητες

$$(I - P_{L+1}) \left(\sum_{t=0}^{T} P_{L+1}^t \right) = \left(\sum_{t=0}^{T} P_{L+1}^t \right) (I - P_{L+1}) = I - P_{L+1}^{T+1}$$

που ισχύουν για οποιοδήποτε T. Παίρνοντας όρια για $T \to \infty$ προκύπτει το αποτέλεσμα.

⊠

Πρόταση 2.2.3. *Το Σέζαρο-όριο,* $\lim_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} P^t$ *υπάρχει και ορίζει ένα πίνακα* Q *της μορφής:*

$$Q = \begin{pmatrix} Q_1 & 0 & \dots & 0 & 0 \\ 0 & Q_2 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & 0 \\ 0 & 0 & \dots & Q_L & 0 \\ Q_{L+11} & Q_{L+12} & \dots & Q_{L+1L} & 0 \end{pmatrix}$$

όπου Q_l, $l = 1, 2, \ldots, L$ *έχει ίδιες γραμμές, που η κάθε μία ισούται με* $\mathbf{q}(l)$. *Η* $\mathbf{q}(l)$ *είναι φυσικά η στάσιμη κατανομή που αντιστοιχεί στον* Q_l. *Επίσης ισχύει*

$$Q_{L+1l} = (I - P_{L+1})^{-1} P_{L+1l} Q_l,\ l = 1, 2, \ldots, L.$$

Ο πίνακας Q *ονομάζεται Σέζαρο-όριο του πίνακα* P.

Αυτό το αποτέλεσμα (ειδικότερα το δεύτερο μέρος της πρότασης) δεν θα μας απασχολείσει σε αυτή την εργασία, καθώς ασχολούμαστε με αδιαχώριστα σύνολα καταστάσεων δηλαδή σύνολα που είναι στοχαστικά κλειστά και όλες οι καταστάσεις επικοινωνούν. Για την απόδειξη της πρότασης ο αναγνώστης ας ανατρέξει στους Filar και Vrieze (1997), Παράρτημα M.

Πρόταση 2.2.4. *Ισχύει η ακόλουθη ταυτότητα:*

$$QP = PQ = QQ = Q$$

Απόδειξη

Πράγματι,

$$Q = \lim_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} P^t = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} P^t =$$

$$= P \left(\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} P^t \right) = \left(\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} P^t \right) P$$

και

$$Q \lim_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} P^t = \lim_{T\to\infty} \frac{1}{T+1} Q \sum_{t=0}^{T} P^t = \lim_{T\to\infty} \frac{(T+1)\,Q}{T+1} = Q$$

⊠

Έστω ένα στοχαστικό παιχνίδι με χώρο καταστάσεων $S = \{1, 2, \ldots, N\}$. Υπενθυμίζουμε ότι ένα σύνολο καταστάσεων μιας αλυσίδας Markov λέγεται αδιαχώριστο αν είναι στοχαστικά κλειστό και όλες οι καταστάσεις του επικοινωνούν. Έστω $\mathbf{f}$ και $\mathbf{g}$, στάσιμες στρατηγικές των δύο παιχτών αντίστοιχα, και έστω ότι ο πίνακας πιθανοτήτων μεταβάσεων $P(\mathbf{f}, \mathbf{g})$ (δηλαδή ο $N \times N$ πίνακας πιθανοτήτων μεταβάσεων του παιχνιδιού από οποιαδήποτε κατάσταση s σε οποιαδήποτε κατάσταση s' όπου $s, s' \in S$ όταν οι δύο παίχτες

χρησιμοποιούν τις στάσιμες στρατηγικές τους **f** και **g**) επάγει μια αδιαχώριστη μαρκοβιανή αλυσίδα (δηλαδή ο χώρος καταστάσεων είναι αδιαχώριστο σύνολο). Καθώς ο χώρος καταστάσεων είναι πεπερασμένο σύνολο έπεται ότι όλες οι καταστάσεις είναι έμμονες θετικές.

Λήμμα 2.2.1. *Αν οι στάσιμες στρατηγικές* **f** *και* **g** *είναι τέτοιες ώστε ο πίνακας πιθανοτήτων μεταβάσεων* $P(\mathbf{f},\mathbf{g})$ *να επάγει μια αδιαχώριστη μαρκοβιανή αλυσίδα, τότε η εξίσωση*

$$\mathbf{q}^T = \mathbf{q}^T P(\mathbf{f},\mathbf{g})$$

διαθέτει λύση **q**. *Επιπλέον, εάν απαιτήσουμε το* **q** *να είναι διάνυσμα πιθανότητας, τότε η λύση είναι μοναδική.*

Απόδειξη

Η απόδειξη αποτελείται από δύο μέρη. Στο πρώτο μέρος θα δείξουμε την ύπαρξη μιας λύσης και στο δεύτερο θα δείξουμε τη μοναδικότητα της λύσης.

Μέρος Πρώτο

Έστω P_{NN-1} και P_{N-1} οι περικομμένοι πίνακες από τον $P(\mathbf{f},\mathbf{g})$ όπου έχουμε διαγράψει την τελευταία στήλη στον πρώτο και την τελευταία στήλη και τελευταία γραμμή στον δεύτερο. Έστω $\mathbf{p}_{N-1}^T$ η τελευταία γραμμή του πίνακα $P(\mathbf{f},\mathbf{g})$ χωρίς το τελευταίο στοιχείο. Έστω $\mathbf{q}_{N-1}$ το διάνυσμα **q** χωρίς το τελευταίο στοιχείο του, $q(N)$, δηλαδή $\mathbf{q}^T = (\mathbf{q}_{N-1}^T, q(N))$. Τότε, μπορούμε να γράψουμε τις πρώτες $N-1$ εξισώσεις της σχέσης $\mathbf{q}^T = \mathbf{q}^T P(\mathbf{f},\mathbf{g})$ ως εξής

$$\mathbf{q}_{N-1}^T = \left(\mathbf{q}_{N-1}^T, q(N)\right) P_{NN-1} = \mathbf{q}_{N-1}^T P_{N-1} + q(N)\mathbf{p}_{N-1}^T$$

Η σχέση αυτή φυσικά είναι ισοδύναμη με την

$$\mathbf{q}_{N-1}^T (I_{N-1} - P_{N-1}) = q(N)\mathbf{p}_{N-1}^T \qquad (1)$$

Ισχυριζόμαστε ότι ο πίνακας $I_{N-1} - P_{N-1}$ είναι αντιστρέψιμος και ότι

$$(I_{N-1} - P_{N-1})^{-1} = \sum_{t=0}^{\infty} P_{N-1}^t \qquad (2)$$

Πράγματι, όταν θεωρούμε το σύνολο των καταστάσεων $\{1, 2, \ldots, N-1\}$ σαν ένα ξεχωριστό υποσύστημα ελεγχόμενο από τον υποστοχαστικό πίνακα (δηλαδή πίνακα με γραμμές που αθροίζουν το πολύ στο 1 και μη αρνητικά

στοιχεία) P_{N-1}, τότε το υποσύστημα είναι μεταβατικό. Αυτή η παρατήρηση είναι άμεση συνέπεια του γεγονότος ότι, άσχετα από την αρχική κατάσταση του υποσυστήματος $\{1, 2, \ldots, N-1\}$, με πιθανότητα 1 αυτό το σύστημα θα εγκαταληφθεί κάτω από τον πίνακα μεταβάσεων P_{N-1} αφού στο αρχικό σύστημα η κατάσταση N είναι προσιτή με πιθανότητα 1. Επομένως, υπάρχει $K > 0$ με $P_{N-1}^K \mathbf{1}_{N-1} \leq \delta \mathbf{1}_{N-1}$ για κάποιο $\delta \in (0,1)$. Αλλά τότε $\lim_{K\to\infty} P_{N-1}^K = 0$. Η σχέση (2) τότε προκύπτει ανάλογα με το (2) της πρότασης 2.2.2 .

Στη συνέχεια, θέτουμε $q(N) = \tilde{q} \in (0,1)$ αυθαίρετο. Με τη βοήθεια της (2) μπορούμε να λύσουμε την (1):

$$\mathbf{q}_{N-1}^T(\tilde{q}) = \tilde{q}\mathbf{p}_{N-1}^T \left(I_{N-1} - P_{N-1}\right)^{-1}. \qquad (3)$$

Παρατηρούμε ότι $\mathbf{q}_{N-1}(\tilde{q}) \geq \mathbf{0}_{N-1}$, αφού όλες οι ποσότητες στο δεξί μέρος της τελευταίας σχέσης είναι μη-αρνητικές. Παρόλο που το διάνυσμα $\left(\mathbf{q}_{N-1}(\tilde{q}), \tilde{q}\right)$ ικανοποιεί την (1), γενικά δεν είναι διάνυσμα πιθανότητας. Παρόλαυτα, εύκολα αποδεικνύεται ότι το κανονικοποιημένο διάνυσμα

$$\mathbf{q}^* := \frac{1}{L}\left(\mathbf{q}_{N-1}(\tilde{q}), \tilde{q}\right)$$

όπου

$$0 < L := \tilde{q} + \sum_{s=1}^{N-1} \left(\mathbf{q}_{N-1}(\tilde{q})\right)_s$$

ικανοποιεί κι αυτό την (1). Απομένει να δείξουμε ότι το $\mathbf{q}^*$ ικανοποιεί και την τελευταία εξίσωση της σχέσης $\mathbf{q}^T = \mathbf{q}^T P(\mathbf{f}, \mathbf{g})$. Αφού ο πίνακας $P(\mathbf{f}, \mathbf{g})$ είναι στοχαστικός, η τελευταία του γραμμή μπορεί να γραφεί ως εξής

$$\left(I_N - (P_{NN-1}, \mathbf{0}_N)\right)\mathbf{1}_N,$$

όπου $(P_{NN-1}, \mathbf{0}_N)$ ισούται με τον P εκτός από την τελευταία στήλη που έχει αντικατασταθεί από ένα μηδενικό διάνυσμα. Τότε

$$\begin{aligned}
{\mathbf{q}^*}^T \left(I_N - (P_{NN-1}, \mathbf{0}_N)\right)\mathbf{1}_N &= \left[\left(\mathbf{q}_{N-1}^{*T}, q^*(N)\right) - \left({\mathbf{q}^*}^T P_{NN-1}, 0\right)\right]\mathbf{1}_N \\
&= \left(\mathbf{0}^T, q^*(N)\right)\mathbf{1}_N \\
&= q^*(N)
\end{aligned}$$

Έτσι, ολοκληρώθηκε η απόδειξη του πρώτου μέρους.

Μέρος Δεύτερο

Πρώτα, παρατηρούμε ότι για οποιαδήποτε λύση της $\mathbf{q}^T = \mathbf{q}^T P(\mathbf{f}, \mathbf{g})$, όπου $\mathbf{q}$ είναι ένα διάνυσμα πιθανότητας, ισχύει ότι $q(s) > 0$ για κάθε $s \in S$. Πράγματι, έστω $\bar{S} := \{s \in S : q(s) = 0\}$ και υποθέτουμε ότι $\bar{S} \neq \emptyset$. Επειδή το S είναι αδιαχώριστο, θα υπάρχει $\bar{s} \in \bar{S}$ και $s \in S \backslash \bar{S}$ τέτοια ώστε $p_{s\bar{s}}(f, g) > 0$. Αλλά τότε

$$\sum_{s=1}^{N} q(s) p_{s\bar{s}}(f, g) > 0 = q(\bar{s})$$

το οποίο δείχνει ότι το $\mathbf{q}$ δεν μπορεί να είναι λύση της εξίσωσης $\mathbf{q}^T = \mathbf{q}^T P(\mathbf{f}, \mathbf{g})$. Άτοπο, άρα $q(s) > 0$ για κάθε $s \in S$.

Τώρα, υποθέτουμε ότι δύο διανύσματα $\bar{\mathbf{q}}$ και $\tilde{\mathbf{q}}$ διαφορετικά μεταξύ τους, ικανοποιούν την σχέση $\mathbf{q}^T = \mathbf{q}^T P(\mathbf{f}, \mathbf{g})$. Από την ανάλυση του πρώτου μέρους, προκύπτει ότι μόλις η τελευταία συνιστώσα μιας λύσης $\mathbf{q}$ σταθεροποιηθεί, τότε οι υπόλοιπες $N - 1$ συνιστώσες καθορίζονται από την (3). Έτσι, έχουμε:

$$\bar{\mathbf{q}}_{N-1}^T = \bar{q}\,(N)\, \mathbf{p}_{N-1}^T \left(I - P_{N-1}\right)^{-1}$$

και

$$\tilde{\mathbf{q}}_{N-1}^T = \tilde{q}\,(N)\, \mathbf{p}_{N-1}^T \left(I - P_{N-1}\right)^{-1}$$

Από τη μη αρνητικότητα του $\mathbf{p}_{N-1}^T \left(I - P_{N-1}\right)^{-1}$ έπεται ότι αν $\bar{q}(N) \geq \tilde{q}(N)$ τότε $\bar{\mathbf{q}}_{N-1}^T \geq \tilde{\mathbf{q}}_{N-1}^T$ και αντίστροφα. Έτσι είτε $\bar{\mathbf{q}} \geq \tilde{\mathbf{q}}$ είτε $\bar{\mathbf{q}} \leq \tilde{\mathbf{q}}$ και αφού και για τα δύο το άθροισμα των συνιστωσών ισούται με 1, συμπεραίνουμε ότι $\bar{\mathbf{q}} = \tilde{\mathbf{q}}$.

$\boxtimes$

Η μοναδική λύση $\mathbf{q}$ της εξίσωσης $\mathbf{q}^T = \mathbf{q}^T P(\mathbf{f}, \mathbf{g})$ καλείται στάσιμη κατανομή του πίνακα $P(\mathbf{f}, \mathbf{g})$. Μπορεί να ερμηνευτεί ως εξής: Υποθέτουμε ότι η αρχική κατάσταση του συστήματος δεν είναι γνωστή με πιθανότητα 1, αλλά ότι $q(s), s \in S$ είναι η πιθανότητα η κατάσταση s να είναι η αρχική κατάσταση. Τότε, για κάθε $s' \in S$, η πιθανότητα ότι στην επόμενη στιγμή απόφασης, η κατάσταση θα είναι s' ισούται με:

$$\sum_{s=1}^{N} p_{ss'}(\mathbf{f}, \mathbf{g}) q(s) = q(s')$$

και έτσι με επαγωγή προκύπτει ότι για κάθε στιγμή απόφασης οι πιθανότητες εμφάνισης της κάθε κατάστασης δίνονται από το διάνυσμα $\mathbf{q}$. Επιπλέον,

παρατηρούμε ότι από τον ορισμό του, το $q(s)$ μπορεί να ερμηνευτεί ως η αναμενόμενη συχνότητα επισκέψεων στην κατάσταση $s \in S$ κατά τη διάρκεια ενός παιχνιδιού.

Κεφάλαιο 3

Αποπληθωρισμένα στοχαστικά παιχνίδια ($\beta-$**Discounted stochastic games)**

3.1 Θεωρία

Έστω Γ_β αποπληθωρισμένο στοχαστικό παιχνίδι 2-παιχτών. Έστω $S = \{1, 2, \ldots, N\}$ ο χώρος καταστάσεων, $A^1(s), A^2(s)$ τα σύνολα αποφάσεων των δύο παιχτών αντίστοιχα και $r^1(s, a^1, a^2)$, $r^2(s, a^1, a^2)$ οι αντίστοιχες τρέχουσες πληρωμές τους σε κάποια στιγμή απόφασης που το παιχνίδι βρίσκεται στην κατάσταση $s \in S$ και οι παίχτες παίρνουν τις αποφάσεις $a^1 \in A^1(s), a^2 \in A^2(s)$. Έστω $p_{ss'}(a^1, a^2) = \mathbb{P}(S_{t+1} = s' \mid S_t = s, A_t^1 = a^1, A_t^2 = a^2)$ οι πιθανότητες μεταβάσεων από την κατάσταση s στην κατάσταση s' σε μια χρονική στιγμή απόφασης $t, t = 0, 1, 2, \ldots$ δεδομένου ότι ο παίχτης I επέλεξε την απόφαση $a^1 \in A^1(s)$ και ο παίχτης II επέλεξε $a^2 \in A^2(s)$, όπου $s, s' \in S$. Βεβαίως, S_t είναι η κατάσταση του παιχνιδιού στη στιγμή απόφασης t, A_t^1, A_t^2 είναι οι αποφάσεις που παίρνουν οι παίχτες I και II αντίστοιχα στη στιγμή t.

Έστω το σύνολο όλων των στάσιμων στρατηγικών του παίχτη I

$$\mathbf{F} := \{\mathbf{f} = (\mathbf{f}(1), \mathbf{f}(2), \ldots, \mathbf{f}(N)) \mid f(s, a^1) \geq 0$$

$$\kappa\alpha\iota \sum_{a^1=1}^{m^1(s)} f(s, a^1) = 1, a^1 \in A^1(s), s \in S\}$$

και αντίστοιχα του παίχτη II

$$\mathbf{G} := \{\mathbf{g} = (\mathbf{g}(1), \mathbf{g}(2), \ldots, \mathbf{g}(N)) \mid g(s, a^2) \geq 0$$

$\kappa\alpha\iota \sum_{a^2=1}^{m^2(s)} g(s,a^2) = 1, a^2 \in A^2(s), s \in S\}$

Έστω $\mathbf{f} = (\mathbf{f}(1), \mathbf{f}(2), \ldots, \mathbf{f}(N))$ και $\mathbf{g} = (\mathbf{g}(1), \mathbf{g}(2), \ldots, \mathbf{g}(N))$ ένα ζευγάρι στάσιμων στρατηγικών για τους παίχτες I και II αντίστοιχα. Στην παρούσα εργασία κάνουμε την σύμβαση ότι $\mathbf{f}$ είναι ένα διάνυσμα-γραμμή ενώ $\mathbf{g}$ ένα διάνυσμα-στήλη. Έτσι, αν ορίσουμε $m^1 := \sum_{s=1}^{N} m^1(s)$ και $m^2 := \sum_{s=1}^{N} m^2(s)$ τότε $\mathbf{f}$ είναι ένα m^1-διάστατο διάνυσμα-γραμμή και $\mathbf{g}$ είναι ένα m^2-διάστατο διάνυσμα-στήλη, όπου $\mathbf{f}(s)$ είναι $m^1(s)$-διάστατο διάνυσμα πιθανότητας και $\mathbf{g}(s)$ $m^2(s)$-διάστατο αντίστοιχα, όπου $m^1(s) = |A^1(s)|$ και $m^2(s) = |A^2(s)|$ οι αντίστοιχοι πληθάριθμοι. Επίσης υπενθυμίζουμε ότι

- $f(s, a^1)$ είναι η πιθανότητα ο παίχτης I να πάρει την απόφαση $a^1 \in A^1(s)$ όταν το σύστημα βρίσκεται στην κατάσταση $s \in S$.
- $g(s, a^2)$ είναι η πιθανότητα ο παίχτης II να πάρει την απόφαση $a^2 \in A^2(s)$ όταν το σύστημα βρίσκεται στην κατάσταση $s \in S$.

Επιπλέον, αναφέρουμε ότι αφού η συνιστώσα $\mathbf{f}(s)$ είναι $m^1(s)$-διάστατο διάνυσμα πιθανότητας θα ισχύει ότι $\mathbf{f}(s) \in \mathbb{P}^{m^1(s)}$ δηλαδή θα ανήκει στο Simplex πιθανοτήτων που είναι συμπαγές και κυρτό σύνολο. Επομένως, το σύνολο $\mathbf{F}$, που είναι το καρτεσιανό γίνομενο των Simplex πιθανοτήτων $\bigtimes_{s=1}^{N} \mathbb{P}^{m^1(s)}$, είναι συμπαγές και κυρτό σύνολο. Δηλαδή, το σύνολο $\mathbf{F}$ των στάσιμων στρατηγικών του παίχτη I είναι συμπαγές και κυρτό. Ομοίως το σύνολο $\mathbf{G}$ των στάσιμων στρατηγικών του παίχτη II είναι συμπαγές και κυρτό.

Επιπροσθέτως, για οποιοδήποτε $s \in S$ το Simplex πιθανοτήτων $\mathbb{P}^{m^1(s)}$ είναι φραγμένο αφού περιέχεται στη μοναδιαία σφαίρα του $\mathbb{R}^{m^1(s)}$. Συνεπώς και το σύνολο $\mathbf{F}$ θα είναι φραγμένο. Ομοίως το σύνολο $\mathbf{G}$ θα είναι φραγμένο.

Οι επόμενες ποσότητες που θα ορίσουμε θα είναι χρήσιμες για την εργασία. Για $s, s' \in S, a^1 \in A^1(s), a^2 \in A^2(s)$ έχουμε

(i) $p_{ss'}(\mathbf{f}, a^2) := \sum_{a^1=1}^{m^1(s)} p_{ss'}(a^1, a^2) f(s, a^1)$

(ii) $p_{ss'}(a^1, \mathbf{g}) := \sum_{a^2=1}^{m^2(s)} p_{ss'}(a^1, a^2) g(s, a^2)$

(iii) $p_{ss'}(\mathbf{f}, \mathbf{g}) := \sum_{a^1=1}^{m^1(s)} \sum_{a^2=1}^{m^2(s)} p_{ss'}(a^1, a^2) f(s, a^1) g(s, a^2)$

(iv) $r(s, \mathbf{f}, a^2) := \sum_{a^1=1}^{m^1(s)} r(s, a^1, a^2) f(s, a^1)$

(v) $r(s, a^2, \mathbf{g}) := \sum_{a^2=1}^{m^2(s)} r(s, a^1, a^2) g(s, a^2)$

(vi) $r(s, \mathbf{f}, \mathbf{g}) := \sum_{a^1=1}^{m^1(s)} \sum_{a^2=1}^{m^2(s)} r(s, a^1, a^2) f(s, a^1) g(s, a^2)$

(vii) το $N-$διάστατο διάνυσμα-στήλη των τρέχουσων πληρωμών θα είναι το $\mathbf{r}(\mathbf{f}, \mathbf{g}) := (r(1, \mathbf{f}, \mathbf{g}), r(2, \mathbf{f}, \mathbf{g}), \ldots, r(N, \mathbf{f}, \mathbf{g}))^T$

Επιπλέον, η αναμενόμενη πληρωμή του παίχτη k, $k = 1, 2$ τη στιγμή απόφασης t όταν η αρχική κατάσταση του παιχνιδιού είναι s_0 και οι παίχτες παίζουν τις στάσιμες στρατηγικές $\mathbf{f}$ και $\mathbf{g}$, είναι $\mathbb{E}_{s_0 \mathbf{fg}}(R_t^k)$ όπου R_t^k είναι η πληρωμή του παίχτη k τη στιγμή t. Συνεπώς, η συνολική αποπληθωρισμένη αναμενόμενη πληρωμή για τον παίχτη k για ένα ζευγάρι στάσιμων στρατηγικών $(\mathbf{f}, \mathbf{g}) \in \mathbf{F} \times \mathbf{G}$ δίνεται από τη σχέση:

$$v_\beta^k(s_0, \mathbf{f}, \mathbf{g}) = \sum_{t=0}^{\infty} \beta^t \mathbb{E}_{s_0 \mathbf{fg}}(R_t^k)$$

όπου $\beta \in (0, 1)$ ο συντελεστής αποπληθωρισμού.

Προς το παρόν θα περιορίσουμε τους παίχτες να επιλέγουν αποκλειστικά στάσιμες στρατηγικές, δηλαδή περιορίζουμε το στοχαστικό παιχνίδι πάνω στο σύνολο $\mathbf{F} \times \mathbf{G}$. Έχουμε δείξει (θεώρημα 1.5.4) ότι για τους δύο συγκεκριμένους τρόπους υπολογισμού της ολικής πληρωμής που εξετάζουμε (αποπληθωρισμένη και οριακού μέσου όρου), εάν το περιορισμένο παιχνίδι έχει τιμή, τότε και το αρχικό παιχνίδι (αυτό όπου οι παίχτες δεν περιορίζονται αποκλειστικά σε στάσιμες στρατηγικές) έχει τιμή και τιμές και βέλτιστες στρατηγικές συμπίπτουν στα δύο παιχνίδια.

Όμως, ενώ το αποπληθωρισμένο περιορισμένο παιχνίδι, όπως θα δούμε, έχει πάντα τιμή, το περιορισμένο παιχνίδι οριακής αναμενόμενης πληρωμής δεν έχει πάντα τιμή.

Υποθέτουμε ότι οι παίχτες αποφασίζουν τις στρατηγικές τους εντελώς ανεξάρτητα και μυστικά και ότι το ενδιαφέρον τους έγκειται στο να μεγιστοποιήσει ο καθένας τη δική του συνολική πληρωμή. Αν επιπλέον, οι παίχτες διαθέτουν ακριβή γνώση για την παρουσία του άλλου στο παιχνίδι και για τις συναρτήσεις πληρωμών τους, τότε η λύση του παιχνιδιού που έχει επικρατήσει στη βιβλιογραφία είναι γνωστή ως *Σημείο Στρατηγικής Ισορροπίας* ή *Σημείο Nash*. Για συντομία όταν αναφερόμαστε σε αυτό, θα το γράφουμε ως ΣΣΙ.

Ορισμός 3.1.1. *Λέμε ότι ένα ζευγάρι στρατηγικών* $\left(\mathbf{f}^0, \mathbf{g}^0\right) \in \mathbf{F} \times \mathbf{G}$ *είναι ΣΣΙ του περιορισμένου παιχνιδιού* Γ_β *όταν*

$$\mathbf{v}_\beta^1(\mathbf{f}, \mathbf{g}^0) \leq \mathbf{v}_\beta^1(\mathbf{f}^0, \mathbf{g}^0) \; \gamma\iota\alpha \; \kappa\acute{\alpha}\theta\varepsilon \; \mathbf{f} \in \mathbf{F}$$

και

$$\mathbf{v}_\beta^2(\mathbf{f}^0, \mathbf{g}^{)} \leq \mathbf{v}_\beta^2(\mathbf{f}^0, \mathbf{g}^0) \; \gamma\iota\alpha \; \kappa\acute{\alpha}\theta\varepsilon \; \mathbf{g} \in \mathbf{G}$$

Το σημαντικότερο χαρακτηριστικό του παραπάνω ορισμού είναι ότι μονομερείς αποκλίσεις από το ΣΣΙ $(\mathbf{f}^0, \mathbf{g}^0)$, είτε του παίχτη I είτε του II, δεν έχουν κίνητρο να εμφανιστούν. Από τη στιγμή που δεν επιτρέπεται η συννενόηση μεταξύ των παιχτών, δεν υπάρχει νόημα για κανέναν από τους δύο παίχτες να αποκλίνει από το $(\mathbf{f}^0, \mathbf{g}^0)$. Το μειονέκτημα του παραπάνου (απλοϊκού) επιχειρήματος πηγάζει από το γεγονός ότι, γενικά, μπορούν να υπάρχουν πολλά ΣΣΙ με διαφορετικές πληρωμές για τους παίχτες. Το πρόβλημα παύει να υφίσταται στα μηδενικού-αθροίσματος στοχαστικά παιχνίδια.

Ένα αποπληθωρισμένο στοχαστικό παιχνίδι θα καλείται μηδενικού αθροίσματος αν

$$r^1(s, a^1, a^2) + r^2(s, a^1, a^2) = 0$$

για κάθε $s \in S, a^1 \in A^1(s), a^2 \in A^2(s)$. Έτσι προκύπτει ότι

$$r(s, a^1, a^2) := r^1(s, a^1, a^2) = -r^2(s, a^1, a^2)$$

για κάθε $s \in S, a^1 \in A^1(s), a^2 \in A^2(s)$. Από τον παραπάνω ορισμό συμπεραίνουμε ότι:

$$\mathbf{v}_\beta(\mathbf{f}, \mathbf{g}) := \mathbf{v}_\beta^1(\mathbf{f}, \mathbf{g}) = -\mathbf{v}_\beta^2(\mathbf{f}, \mathbf{g})$$

για κάθε $(\mathbf{f}, \mathbf{g}) \in \mathbf{F} \times \mathbf{G}$.

Επιπλέον, αν $\left(\mathbf{f}^0, \mathbf{g}^0\right) \in \mathbf{F} \times \mathbf{G}$ είναι ΣΣΙ, τότε τα δύο σύνολα ανισοτήτων στον ορισμό του ΣΣΙ παίρνουν τη μορφή

$$\mathbf{v}_\beta(\mathbf{f}, \mathbf{g}^0) \leq \mathbf{v}_\beta(\mathbf{f}^0, \mathbf{g}^0) \leq \mathbf{v}_\beta(\mathbf{f}^0, \mathbf{g}) \qquad (1)$$

για κάθε $\mathbf{f} \in \mathbf{F}$ και $\mathbf{g} \in \mathbf{G}$. Σε αυτήν την περίπτωση, οι στρατηγικές $\mathbf{f}^0, \mathbf{g}^0$ καλούνται βέλτιστες στάσιμες στρατηγικές για τους παίχτες I και II αντίστοιχα.

Από την (1) προκύπτει άμεσα ότι αν $(\mathbf{f}^*, \mathbf{g}^*) \in \mathbf{F} \times \mathbf{G}$ είναι ένα άλλο ζευγάρι βέλτιστων στρατηγικών, τότε:

$$\mathbf{v}_\beta(\mathbf{f}^*, \mathbf{g}^*) = \mathbf{v}_\beta(\mathbf{f}^0, \mathbf{g}^0).$$

Συνεπώς, τα διανύσματα των αποπληθωρισμένων τιμών του παιχνιδιού για όλα τα ζευγάρια βέλτιστων στρατηγικών συμπίπτουν. Ένα τέτοιο διάνυσμα εφεξής θα καλείται διάνυσμα τιμή του (μηδενικού-αθροίσματος) παιχνιδιού Γ_β και θα συμβολίζεται ως

$$\mathbf{v}_\beta = (\upsilon_\beta(1), \upsilon_\beta(2), \ldots, \upsilon_\beta(N))^T.$$

Από τον ορισμό της συνολικής αποπληθωρισμένης αναμενόμενης πληρωμής, εύκολα προκύπτει ότι για στάσιμες στρατηγικές $\mathbf{f}$ και $\mathbf{g}$ το διάνυσμα $\mathbf{v}_\beta(\mathbf{f}, \mathbf{g})$ των πληρωμών ικανοποιεί την αναδρομική σχέση

$$\mathbf{v}_\beta(\mathbf{f}, \mathbf{g}) = \mathbf{r}(\mathbf{f}, \mathbf{g}) + \beta P(\mathbf{f}, \mathbf{g}) \mathbf{v}_\beta(\mathbf{f}, \mathbf{g})$$

από την οποία προκύπτει ότι

$$[I - \beta P(\mathbf{f}, \mathbf{g})]\, \mathbf{v}_\beta(\mathbf{f}, \mathbf{g}) = \mathbf{r}(\mathbf{f}, \mathbf{g})$$

Ο πίνακας $[I - \beta P(\mathbf{f}, \mathbf{g})]$ είναι αντιστρέψιμος. Πράγματι, έχουμε

$$(I - \beta P)\left(\sum_{t=0}^{T} \beta^t P^t\right) = (I - \beta P)(I + \beta P + \beta^2 P^2 + \cdots + \beta^T P^T) =$$

$$I + \beta P + \beta^2 P^2 + \cdots + \beta^T P^T - \beta P - \beta^2 P^2 - \cdots - \beta^{T+1} P^{T+1} =$$

$$I - \beta^{T+1} P^{T+1}$$

Επιπλέον

$$\left(\sum_{t=0}^{T} \beta^t P^t\right)(I - \beta P) = (I + \beta P + \beta^2 P^2 + \cdots + \beta^T P^T)(I - \beta P) =$$

$$I + \beta P + \beta^2 P^2 + \cdots + \beta^T P^T - \beta P - \beta^2 P^2 - \cdots - \beta^{T+1} P^{T+1} =$$

$$I - \beta^{T+1} P^{T+1}$$

Συνεπώς

$$(I - \beta P)\left(\sum_{t=0}^{T} \beta^t P^t\right) = \left(\sum_{t=0}^{T} \beta^t P^t\right)(I - \beta P) = I - \beta^{T+1} P^{T+1}$$

και παίρνοντας όρια για $T \to \infty$ προκύπτει

$$(I - \beta P)\left(\sum_{t=0}^{\infty} \beta^t P^t\right) = \left(\sum_{t=0}^{\infty} \beta^t P^t\right)(I - \beta P) = I - \lim_{T\to\infty} \beta^{T+1} P^{T+1} = I$$

διότι ο πίνακας P^{T+1} για $T \to \infty$ έχει πεπερασμένα στοιχεία (αφού είναι στοχαστικός πίνακας) και $\beta \in (0,1)$. Άρα ο πίνακας $(I - \beta P)$ είναι αντιστρέψιμος και ισχύει

$$(I - \beta P)^{-1} = \sum_{t=0}^{\infty} \beta^t P^t$$

Επομένως, θα ισχύει ότι

$$\mathbf{v}_\beta(\mathbf{f}, \mathbf{g}) = [I - \beta P(\mathbf{f}, \mathbf{g})]^{-1} \mathbf{r}(\mathbf{f}, \mathbf{g}) \qquad (2)$$

Τέλος, στα αποπληθωρισμένα παιχνίδια για τον υπολογισμό της τιμής για μία αρχική κατάσταση s_0 δεδομένου ότι οι παίχτες παίζουν στάσιμες στρατηγικές $\mathbf{f}$ και $\mathbf{g}$, αντίστοιχα, εισάγουμε ένα παράγοτα κανονικοποίησης $(1-\beta)$ ως εξής:

$$\upsilon_\beta(s_0, \mathbf{f}, \mathbf{g}) = (1-\beta) \sum_{t=0}^{\infty} \beta^t \mathbb{E}_{s_0 \mathbf{fg}} \left[r(S_t, A_t^1, A_t^2)\right] \qquad (3)$$

Εισάγοντας τον παραπάνω παράγοντα κανονικοποίησης, η αποπληθωρισμένη συνολική πληρωμή του στοχαστικού παιχνιδιού είναι κυρτός συνδυασμός των ποσοτήτων $\mathbb{E}_{s_0 \mathbf{fg}} \left[r(S_t, A_t^1, A_t^2)\right], t = 0, 1, \ldots$ και έτσι έχουμε:

$$\mid \upsilon_\beta(s_0, \mathbf{f}, \mathbf{g}) \mid \leq \max_{s, a^1, a^2} \mid r(s, a^1, a^2) \mid .$$

Στην ουσία ο λόγος που πολλαπλασιάζουμε με την ποσότητα $1-\beta$ αφορά την αναγωγή της ροής των πληρωμών σε παρούσες τιμές, ώστε να έχει νόημα η σύγκριση διαφορετικών αναμενόμενων ροών που θα προκύπτουν από διαφορετικές στρατηγικές. Για παράδειγμα, αν πάρουμε την ακολουθία πληρωμών $1, 1, 1, \ldots$ στο παιχνίδι Γ_β, τότε η παρούσα αξία θα είναι $1+\beta+\beta^2+\cdots = \frac{1}{1-\beta}$. Πολλαπλασιάζοντας με την ποσότητα $1-\beta$ προκύπτει ότι η τιμή του παιχνιδιού είναι 1 που είναι ακριβώς η πληρωμή σε κάθε στιγμή απόφασης.

Επίσης, συμπληρώνουμε ότι βάσει της σχέσης (3), η σχέση (2) γίνεται

$$\mathbf{v}_\beta(\mathbf{f}, \mathbf{g}) = (1-\beta)\,[I - \beta P(\mathbf{f}, \mathbf{g})]^{-1} r(\mathbf{f}, \mathbf{g}) \qquad (4)$$

Λήμμα 3.1.1. *Αν το αποπληθωρισμένο στοχαστικό παιχνίδι διαθέτει τιμή* $\mathbf{v}_\beta = (\upsilon_\beta(1), \upsilon_\beta(2), \ldots, \upsilon_\beta(N))^T$ *και αν και οι δύο παίχτες διαθέτουν βέλτιστες στάσιμες στρατηγικές, τότε το διάνυσμα* $\mathbf{v}_\beta$ *ικανοποιεί το σύνολο των εξισώσεων:*

$$\upsilon_\beta(s) = val\left[(1-\beta)r(s,a^1,a^2) + \beta\sum_{s'\in S} p_{ss'}(a^1,a^2)\upsilon_\beta(s')\right]$$

όπου η ποσότητα μέσα στην αγκύλη είναι ένας $m^1(s)\times m^2(s)$ *πίνακας (με γραμμές* $a^1 = 1,2,\ldots,m^1(s)$ *και στήλες* $a^2 = 1,2,\ldots,m^2(s)$*).*

Απόδειξη

Έστω ότι $\mathbf{v}_\beta = (\upsilon_\beta(1), \upsilon_\beta(2), \ldots, \upsilon_\beta(N))^T$ είναι το διάνυσμα τιμή του παιχνιδιού Γ_β και έστω ότι και οι δύο παίχτες διαθέτουν βέλτιστες στάσιμες στρατηγικές. Έστω $\mathbf{f}^*$ μία βέλτιστη στάσιμη στρατηγική του παίχτη I. Υποθέτουμε ότι ο παίχτης I παίζει $\mathbf{f}^*$, ο παίχτης II παίζει αυθαίρετα μία στρατηγική $\mathbf{g}$ στη χρονική στιγμή απόφασης 0 και ότι μετά την αρχική κατάσταση $s\in S$ το σύστημα μεταπηδά σε μια κατάσταση s' στη στιγμή απόφασης $t=1$. Τότε, ο παίχτης II, ο οποίος από την υπόθεση διαθέτει βέλτιστη στάσιμη στρατηγική, μπορεί να εμποδίσει την αποπληθωρισμένη πληρωμή στην κατάσταση s' στη στιγμή απόφασης 1 να ξεπεράσει την $\upsilon_\beta(s')$. Αφού η πιθανότητα να φτάσουμε στην κατάσταση s' ισούται με $p_{ss'}(\mathbf{f}^*,\mathbf{g})$, έχουμε ότι η συνολική αναμενόμενη πληρωμή μπορει να είναι το πολύ:

$$(1-\beta)r(s,\mathbf{f}^*,\mathbf{g}) + \beta\sum_{s'\in S} p_{ss'}(\mathbf{f}^*,\mathbf{g})\upsilon_\beta(s')$$

το οποίο μπορει να είναι τουλάχιστον τόσο μεγάλο όσο το $\upsilon_\beta(s)$ διότι η στρατηγική $\mathbf{f}^*$ είναι βέλτιστη. Έτσι έχουμε

$$\min_{\mathbf{g}}\left\{(1-\beta)r(s,\mathbf{f}^*,\mathbf{g}) + \beta\sum_{s'\in S} p_{ss'}(\mathbf{f}^*,\mathbf{g})\upsilon_\beta(s')\right\} \geq \upsilon_\beta(s)$$

και ανάλογα μπορούμε να συμπεράνουμε ότι

$$\max_{\mathbf{f}}\left\{(1-\beta)r(s,\mathbf{f},\mathbf{g}^*) + \beta\sum_{s'\in S} p_{ss'}(\mathbf{f},\mathbf{g}^*)\upsilon_\beta(s')\right\} \leq \upsilon_\beta(s)$$

για μια βέλτιστη στάσιμη στρατηγική $\mathbf{g}^*$ του παίχτη II. Συνδυάζοντας τις δύο παραπάνω ανισότητες προκύπτει για κάθε $s\in S$:

$$\upsilon_\beta(s) = val\left[(1-\beta)r(s,a^1,a^2) + \beta\sum_{s'\in S} p_{ss'}(a^1,a^2)\upsilon_\beta(s')\right]$$

όπου η ποσότητα μέσα στην αγκύλη είναι ο πίνακας με γραμμές $a^1 = 1, 2, \ldots, m^1(s)$ και στήλες $a^2 = 1, 2, \ldots, m^2(s)$.

$\boxtimes$

Πρωτού αναφέρουμε το θεώρημα για την ύπαρξη τιμής και βέλτιστων στρατηγικών για τα αποπληθωρισμένα παιχνίδια, θα δώσουμε δύο χρήσιμα για την απόδειξη του θεωρήματος λήμματα.

Λήμμα 3.1.2. *Αν* $\mathbf{v} \in \mathbb{R}^N$ *και* $\mathbf{f}$ *και* $\mathbf{g}$ *είναι στάσιμες στρατηγικές τέτοιες ώστε*

$$\mathbf{v} \leq (1-\beta)\mathbf{r}(\mathbf{f},\mathbf{g}) + \beta P(\mathbf{f},\mathbf{g})\mathbf{v},$$

τότε $\mathbf{v} \leq \mathbf{v}_\beta(\mathbf{f},\mathbf{g})$. *Αν στην πρώτη σχέση η ανισότητα ισχύει αντίστροφα, τότε η ανισότητα ισχύει αντίστροφα και στη δεύτερη σχέση.*

Απόδειξη

Επαναλαμβάνοντας την παραπάνω ανισότητα k φορές έχουμε:

$$\begin{aligned}\mathbf{v} \leq (1-\beta)r(\mathbf{f},\mathbf{g}) + \beta(1-\beta)P(\mathbf{f},\mathbf{g})r(\mathbf{f},\mathbf{g}) + \beta^2(1-\beta)P^2(\mathbf{f},\mathbf{g})r(\mathbf{f},\mathbf{g}) + \\ + \beta^3(1-\beta)P^3(\mathbf{f},\mathbf{g})r(\mathbf{f},\mathbf{g}) + \cdots + \beta^{k-1}(1-\beta)P^{k-1}(\mathbf{f},\mathbf{g})r(\mathbf{f},\mathbf{g}) + \\ + \beta^k P^k(\mathbf{f},\mathbf{g})\mathbf{v} = \\ = (1-\beta)\left[I + \beta P + \cdots + \beta^{k-1}P^{k-1}(\mathbf{f},\mathbf{g})\right] r(\mathbf{f},\mathbf{g}) + \beta^k P^k(\mathbf{f},\mathbf{g})\mathbf{v}\end{aligned}$$

και παίρνοντας όριο για $k \to \infty$ έπεται

$$\mathbf{v} \leq (1-\beta)\left(I - \beta P(\mathbf{f},\mathbf{g})\right)^{-1} r(\mathbf{f},\mathbf{g}).$$

Όμως από τη σχέση (4) γνωρίζουμε ότι $(1-\beta)\left(I - \beta P(\mathbf{f},\mathbf{g})\right)^{-1} r(\mathbf{f},\mathbf{g}) = \mathbf{v}_\beta(\mathbf{f},\mathbf{g})$ και έτσι έπεται το ζητούμενο

$$\mathbf{v} \leq \mathbf{v}_\beta(\mathbf{f},\mathbf{g})$$

$\boxtimes$

Λήμμα 3.1.3. *Αν* A *και* B *είναι πίνακες ίδιας διάστασης, τότε:*

$$\mid val[A] - val[B] \mid \leq max_{i,j} \mid a_{ij} - b_{ij} \mid$$

Απόδειξη

Για κάθε i, j ισχύει:

$$a_{ij} - b_{ij} \leq \max_{i,j} |a_{ij} - b_{ij}|$$

$$\Rightarrow a_{ij} \leq b_{ij} + \max_{i,j} |a_{ij} - b_{ij}|$$

οπότε προκύπτει ότι:

$$val[A] \leq val[B] + \max_{i,j} |a_{ij} - b_{ij}|$$

δηλαδή

$$val[A] - val[B] \leq \max_{i,j} |a_{ij} - b_{ij}|$$

Αντιμεταθέτοντας τώρα τους πίνακες A και B στην παραπάνω σχέση συμπεραίνουμε ότι

$$| \, val[A] - val[B] \, | \leq max_{i,j} \mid a_{ij} - b_{ij} \mid .$$

$\boxtimes$

Θεώρημα 3.1.1.

1. *Τα αποπληθωρισμένα στοχαστικά παιχνίδια μηδενικού-αθροίσματος 2-παιχτών διαθέτουν (διάνυσμα) τιμή* $\mathbf{v}_\beta$.

2. *Το διάνυσμα τιμή είναι η μοναδική λύση στο παρακάτω σύνολο από N εξισώσεις (στην κάθε κατάσταση* s *αντιστοιχεί μία εξίσωση):*

$$x(s) = val \left[(1-\beta) r(s, a^1, a^2) + \beta \sum_{s' \in S} p_{ss'}(a^1, a^2) x(s') \right] \qquad (5)$$

   όπου η ποσότητα μέσα στην αγκύλη είναι ο $m^1(s) \times m^2(s)$ *πίνακας με γραμμές* $a^1 = 1, 2, \ldots, m^1(s)$ *και στήλες* $a^2 = 1, 2, \ldots, m^2(s)$.

3. *Βέλτιστες στάσιμες στρατηγικές μπορούν να κατασκευαστούν από βέλτιστες αποφάσεις στα πινακοπαιχνίδια* (5). *Δηλαδή, οι* $\mathbf{f}$ *και* $\mathbf{g}$ *είναι βέλτιστες στάσιμες στρατηγικές στο στοχαστικό παιχνίδι όταν για κάθε* $s \in S$ *οι συνιστώσες τους* $\mathbf{f}(s)$ *και* $\mathbf{g}(s)$ *είναι βέλτιστες στρατηγικές στο πινακοπαιχνίδι*

$$\left[r(s, a^1, a^2) + \beta \sum_{s' \in S} p_{ss'}(a^1, a^2) \upsilon_\beta(s') \right].$$

Απόδειξη

Η απόδειξη θα δοθεί σε τρία μέρη:

Μέρος I

Πρώτα θα αποδείξουμε ότι η εξίσωση (5) έχει μοναδική λύση στο $\mathbb{R}^N$. Θεωρούμε την απεικόνιση $U : \mathbb{R}^N \to \mathbb{R}^N$ ορισμένη ως εξής:

$$(U\mathbf{v})(s) := val\left[(1-\beta)r(s,a^1,a^2) + \beta\sum_{s'\in S} p_{ss'}(a^1,a^2)\upsilon(s')\right]$$

για κάθε $s \in S$ και για κάθε $\mathbf{v} \in \mathbb{R}^N$. Παρατηρούμε ότι σε συμπαγή μορφή οι εξισώσεις (5) γράφονται $U\mathbf{x} = \mathbf{x}$ και επομένως για να δείξουμε ότι οι (5) έχουν λύση αρκεί να δείξουμε ότι ο τελεστής U έχει σταθερό σημείο. Θα δείξουμε ότι η απεικόνιση U είναι τελεστής συστολής ως προς τη νόρμα $||\mathbf{v}|| = max_s|\upsilon(s)|$. Πράγματι,

$$||U\mathbf{v} - U\mathbf{w}|| \leq$$

$$\max_s\{\max_{a^1,a^2} \mid (1-\beta)r(s,a^1,a^2) + \beta\textstyle\sum_{s'\in S} p_{ss'}(a^1,a^2)\upsilon(s') -$$

$$(1-\beta)r(s,a^1,a^2) - \beta\textstyle\sum_{s'\in S} p_{ss'}(a^1,a^2)w(s') \mid\} =$$

$$\max_{s,a^1,a^2} \mid \beta\sum_{s'\in S} p_{ss'}(a^1,a^2)\left(\upsilon(s') - w(s')\right) \mid \leq$$

$$\max_{s,a^1,a^2} \beta\sum_{s'\in S} p_{ss'}(a^1,a^2) \mid \upsilon(s') - w(s') \mid \leq$$

$$\max_{s,a^1,a^2} \beta\sum_{s'\in S} p_{ss'}(a^1,a^2) \mid\mid \mathbf{v} - \mathbf{w} \mid\mid =$$

$$\beta \mid\mid \mathbf{v} - \mathbf{w} \mid\mid .$$

Επομένως, η απεικόνιση U είναι τελεστής συστολής και από το θεώρημα συστολής Banach προκύπτει ότι η U έχει μοναδικό σταθερό σημείο.

Μέρος II

Έστω $\mathbf{v}^*$ η μοναδική λύση των εξισώσεων (5). Θα δείξουμε ότι υπάρχουν στάσιμες στρατηγικές $\mathbf{f}^*$ και $\mathbf{g}^*$ τέτοιες ώστε για κάθε $\mathbf{f}$ και $\mathbf{g}$

$$\mathbf{v}_\beta(\mathbf{f},\mathbf{g}^*) \leq \mathbf{v}^* \leq \mathbf{v}_\beta(\mathbf{f}^*,\mathbf{g})$$

Έστω $\mathbf{f}^*$ τέτοια ώστε $\mathbf{f}^*(s)$ να είναι βέλτιστη στρατηγική του παίχτη I στο πινακοπαιχνίδι

$$\left[(1-\beta)r(s,a^1,a^2)+\beta\sum_{s'\in S}p_{ss'}(a^1,a^2)v^*(s')\right]$$

το οποίο έχει τιμή $v^*(s)$. Έτσι, για όλα τα $a^2=1,2,\ldots,m^2(s)$ έχουμε

$$(1-\beta)r(s,\mathbf{f}^*,a^2)+\beta\sum_{s'\in S}p_{ss'}(\mathbf{f}^*,a^2)v^*(s')\geq v^*(s).$$

Αλλά τότε, για κάθε $\mathbf{g}$

$$(1-\beta)r(s,\mathbf{f}^*,\mathbf{g})+\beta\sum_{s'\in S}p_{ss'}(\mathbf{f}^*,\mathbf{g})v^*(s')\geq v^*(s),$$

ή σε μορφή πινάκων

$$(1-\beta)r(\mathbf{f}^*,\mathbf{g})+\beta P(\mathbf{f}^*,\mathbf{g})\mathbf{v}^*\geq\mathbf{v}^*,$$

το οποίο από το λήμμα 3.1.2 δίνει ότι:

$$\mathbf{v}_\beta(\mathbf{f}^*,\mathbf{g})\geq\mathbf{v}^*.$$

Αντίστοιχα, προκύπτει ότι $\mathbf{v}_\beta(\mathbf{f},\mathbf{g}^*)\leq\mathbf{v}^*$ όταν η $\mathbf{g}^*$ ορίζεται ανάλογα με την $\mathbf{f}^*$. Άρα το παιχνίδι, περιορισμένο σε στάσιμες στρατηγικές, έχει τιμή $\mathbf{v}^*$.

Μέρος III

Αφού το παιχνίδι περιορισμένο σε στάσιμες στρατηγικές έχει τιμή $\mathbf{v}^*$, τότε από το θεώρημα 1.5.4 και το αρχικό παιχνίδι θα έχει τιμή $\mathbf{v}^*$ και οι βέλτιστες στρατηγικές $\mathbf{f}^*$ και $\mathbf{g}^*$ θα είναι και βέλτιστες στρατηγικές στο αρχικό παιχνίδι.

$\boxtimes$

Η απεικόνιση U χρησιμοποιείται στον αλγόριθμο διαδοχικών προσεγγίσεων (Successive Approximation Algorithm) όπως φαίνεται παρακάτω.

Ορίζουμε επαγωγικά, $U^n=U(U^{n-1}\mathbf{v})$ όπου $U^1\equiv U$. Αφού $\mathbf{v}_\beta=U\mathbf{v}_\beta$ συμπεραίνουμε ότι $\mathbf{v}_\beta=U^n\mathbf{v}_\beta$. Άρα για κάθε $\mathbf{v}\in\mathbb{R}^N$ έχουμε

$$||U^n\mathbf{v}-\mathbf{v}_\beta||=||U^n\mathbf{v}-U^n\mathbf{v}_\beta||\leq\beta||U^{n-1}\mathbf{v}-U^{n-1}\mathbf{v}_\beta||$$

Επομένως,

$$||U^n\mathbf{v}-\mathbf{v}_\beta||\leq\beta^n||\mathbf{v}-\mathbf{v}_\beta||$$

οπότε το $U^n\mathbf{v}$ συγκλίνει γεωμετρικά στο $\mathbf{v}_\beta$.

Επιπλέον, μπορούμε να εφαρμόσουμε τον εξής αλγόριθμο αναδρομής πάνω στις πολιτικές (Policy Iteration) στα αποπληθωρισμένα στοχαστικά παιχνίδια.

Έστω $\mathbf{v}_n$ η παρούσα εκτίμηση της τιμής του παιχνιδιού. Κατασκευάζουμε στρατηγικές $\mathbf{f}_n$ και $\mathbf{g}_n$ τέτοιες ώστε για κάθε $s \in S$ οι αποφάσεις $\mathbf{f}_n(s)$ και $\mathbf{g}_n(s)$ να είναι βέλτιστες στο πινακοπαιχνίδι

$$\left[(1-\beta)r(s,a^1,a^2) + \beta \sum_{s' \in S} p_{ss'}(a^1,a^2)v_n(s')\right]$$

και παίρνουμε $\mathbf{v}_{n+1} = \mathbf{v}_\beta(\mathbf{f}_n, \mathbf{g}_n)$.

Δυστυχώς ο συγκεκριμένος αλγόριθμος δεν συγλίνει πάντα (Filar και Vrieze (1996), Παράδειγμα 4.3.3). Παρολ΄ αυτά σε πολλές περιπτώσεις ο αλγόριθμος μπορεί να εφαρμοστεί.

Επιγραμματικά αναφέρουμε ότι ο γραμμικός προγραμματισμός γενικά δεν μπορεί να λύσει τα αποπληθωρισμένα στοχαστικά παιχνίδια 2-παιχτών μηδενικού αθροίσματος. Αυτό μπορούμε να το δούμε κατασκευάζοντας ένα αποπληθωρισμένο παιχνίδι 2-παιχτών μηδενικού-αθροίσματος με ρητούς αριθμούς ως συντεταγμένες των πινάκων-καταστάσεων, του οποίου η τιμή να μπορεί να υπολογιστεί με κλειστό τρόπο και να είναι άρρητος αριθμός (Παράδειγμα 3.3 παρακάτω). Δηλαδή, τα αποπληθωρισμένα στοχαστικά παιχνίδια δεν έχουν την *Ordered Field Property* (OFP). Αν όμως επιδέχονταν επίλυση μέσω γραμμικού προγραμματισμού θα είχαν την OFP, όπως συμβαίνει π.χ. με την τιμή ενός πινακοπαιχνιδιού.

Πάντως σε τρεις τουλάχιστον ειδικές περιπτώσεις (Filar και Vrieze (1997), Κεφάλαιο 3.2) η λύση (τιμή) μπορεί να βρεθεί μέσω γραμμικού προγραμματισμού. Αυτές είναι οι περιπτώσεις

1. που οι πιθανότητες μεταβάσεων εξαρτώνται μόνο από τις αποφάσεις του ενός μόνο παίχτη (Single-Controller Discounted Stochastic Games),

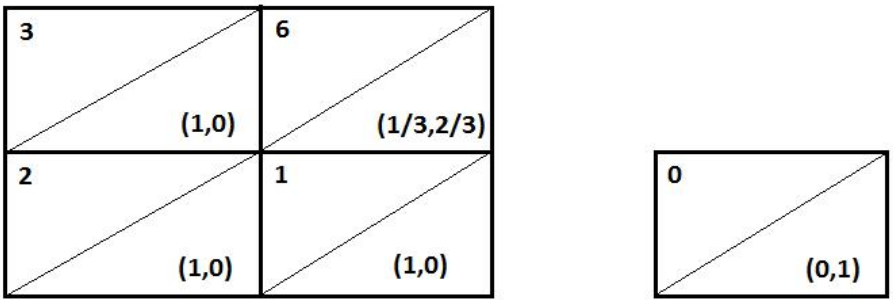
2. που (a) η τρέχουσα πληρωμή $r(s,a^1,a^2)$ χωρίζεται στο άθροισμα δύο συναρτήσεων όπου η μία εξαρτάται αποκλειστικά από την κατάσταση s και η άλλη εξαρτάται αποκλειστικά από το ζεύγος των αποφάσεων (a^1,a^2) και (b) οι μεταπηδήσεις είναι της μορφής $p_{\cdot s'}(a^1,a^2)$, δηλαδή είναι ανεξάρτητες από την εκάστοτε κατάσταση s (Separable Reward State Independent Trasition Discounted Stochastic Games, SER-SIT),

3. το σύνολο των καταστάσεων S διαμερίζεται στα σύνολα S^1 και S^2 τα οποία έχουν την ιδιότητα για $s \in S^1$ να ισχύει $p_{ss'}(a^1, a^2) = p_{ss'}(a^1)$ και για $s \in S^2$ να ισχύει $p_{ss'}(a^1, a^2) = p_{ss'}(a^2)$ (Switching Controller Discounted Stochastic Games).

3.2 Παραδείγματα αποπληθωρισμένων στοχαστικών παιχνιδιών

Παράδειγμα 3.1.

Έστω ένα στοχαστικό παιχνίδι 2-παιχτών μηδενικού-αθροίσματος με χώρο καταστάσεων $S = \{1, 2\}$, χώρο αποφάσεων $A^1(1) = A^2(1) = \{1, 2\}$, $A^1(2) = A^2(2) = \{1\}$, συντελεστή αποπληθωρισμού $\beta = \frac{3}{4}$ και πίνακα πληρωμών και πιθανοτήτων μετάβασης



όπου οι πιθανότητες μετάβασης $(\frac{1}{3}, \frac{2}{3})$ στην κατάσταση 1 δηλώνουν ότι με πιθανότητα $\frac{1}{3}$ θα παραμείνει στην κατάσταση 1 και με πιθανότητα $\frac{2}{3}$ θα μεταβεί στην κατάσταση 2 όταν ο παίχτης I διαλέγει την απόφαση $1 \in A^1(1)$ και ο παίχτης II διαλέγει την απόφαση $2 \in A^2(1)$. Η τρέχουσα πληρωμή (από τον παίχτη II στον παίχτη I) είναι $r(1, 1, 2) = 6$. Επιπλέον, η κατάσταση 2 είναι απορροφητική με πληρωμή 0, δηλαδή όταν το παιχνίδι βρεθεί στην κατάσταση 2 θα παραμείνει για πάντα εκεί και οι δύο παίχτες θα έχουν πληρωμή 0.

Λύνουμε τις εξισώσεις

$$\upsilon_\beta(1) = val \begin{bmatrix} (1-\beta)3 + \beta\upsilon_\beta(1) & (1-\beta)6 + \beta\frac{1}{3}\upsilon_\beta(1) + \beta\frac{2}{3}\upsilon_\beta(2) \\ (1-\beta)2 + \beta\upsilon_\beta(1) & (1-\beta)1 + \beta\upsilon_\beta(1) \end{bmatrix}$$

και

$$\upsilon_\beta(2) = (1-\beta)0 + \beta\upsilon_\beta(2).$$

Από τη δεύτερη προκύπτει άμεσα ότι

$$\upsilon_\beta(2) = 0$$

οπότε η πρώτη εξίσωση για $\beta = \frac{3}{4}$ γίνεται

$$v_{\frac{3}{4}}(1) = val \begin{bmatrix} \frac{3}{4} + \frac{3}{4}v_{\frac{3}{4}}(1) & \frac{6}{4} + \frac{1}{4}v_{\frac{3}{4}}(1) \\ \frac{2}{4} + \frac{3}{4}v_{\frac{3}{4}}(1) & \frac{1}{4} + \frac{3}{4}v_{\frac{3}{4}}(1) \end{bmatrix}$$

Παρατηρούμε ότι προφανώς $v_{\frac{3}{4}}(1) > 0$. Διακρίνουμε τρεις περιπτώσεις
1. Αν $\frac{6}{4} + \frac{1}{4}v_{\frac{3}{4}}(1) \geq \frac{3}{4} + \frac{3}{4}v_{\frac{3}{4}}(1) \Leftrightarrow v_{\frac{3}{4}}(1) \leq \frac{3}{2}$ τότε το πάνω αριστερά στοιχείο είναι σαγματικό σημείο και επομένως $v_{\frac{3}{4}}(1) = \frac{3}{4} + \frac{3}{4}v_{\frac{3}{4}}(1) \Rightarrow v_{\frac{3}{4}}(1) = 3$. Άτοπο, διότι αντιφάσκει με την $v_{\frac{3}{4}}(1) \leq \frac{3}{2}$.
2. Αν $\frac{1}{4} + \frac{3}{4}v_{\frac{3}{4}}(1) \geq \frac{6}{4} + \frac{1}{4}v_{\frac{3}{4}}(1) \Leftrightarrow v_{\frac{3}{4}}(1) \geq \frac{5}{2}$ τότε το κάτω δεξιά στοιχείο είναι σαγματικό σημείο και επομένως $v_{\frac{3}{4}}(1) = \frac{1}{4} + \frac{3}{4}v_{\frac{3}{4}}(1) \Rightarrow v_{\frac{3}{4}}(1) = 1$.
Αλλά $1 < \frac{5}{2}$ και επομένως ούτε αυτή η περίπτωση μπορεί να ισχύσει.
3. Αν $\frac{3}{4} + \frac{3}{4}v_{\frac{3}{4}}(1) > \frac{6}{4} + \frac{1}{4}v_{\frac{3}{4}}(1) > \frac{1}{4} + \frac{3}{4}v_{\frac{3}{4}}(1) \Leftrightarrow \frac{3}{2} < v_{\frac{3}{4}}(1) < \frac{5}{2}$ τότε το πάνω δεξιά στοιχείο είναι σαγματικό σημείο και επομένως $v_{\frac{3}{4}}(1) = \frac{6}{4} + \frac{1}{4}v_{\frac{3}{4}}(1) \Rightarrow v_{\frac{3}{4}}(1) = 2$. Η λύση γίνεται δεκτή αφού ανήκει στο $(\frac{3}{2}, \frac{5}{2})$.

Οι βέλτιστες στάσιμες στρατηγικές για τους δύο παίχτες θα είναι

$$\mathbf{f}^*_{\frac{3}{4}} = ((1), (1))$$

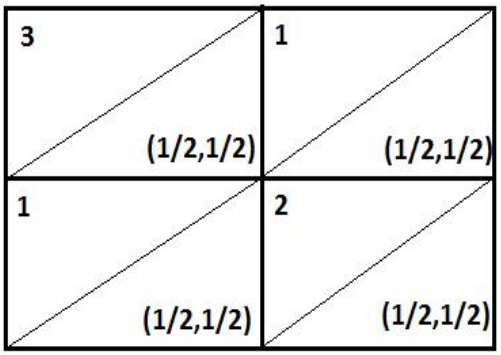
και

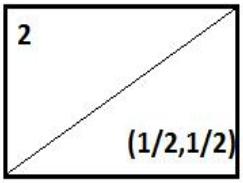
$$\mathbf{g}^*_{\frac{3}{4}} = ((2), (1))$$

⊠

Παράδειγμα 3.2.

Έστω ένα στοχαστικό παιχνίδι 2-παιχτών μηδενικού αθροίσματος με χώρο καταστάσεων $S = \{1, 2\}$, χώρο αποφάσεων $A^1(1) = A^2(1) = \{1, 2\}, A^1(2) = A^2(2) = \{1\}$ και πίνακα πληρωμών και πιθανοτήτων μετάβασης





Λύνουμε τις εξισώσεις

$$\upsilon_\beta(1) = val\begin{bmatrix} 3(1-\beta)+\beta\frac{1}{2}\upsilon_\beta(1)+\beta\frac{1}{2}\upsilon_\beta(2) & 1-\beta+\beta\frac{1}{2}\upsilon_\beta(1)+\beta\frac{1}{2}\upsilon_\beta(2) \\ 1-\beta+\beta\frac{1}{2}\upsilon_\beta(1)+\beta\frac{1}{2}\upsilon_\beta(2) & 2(1-\beta)+\beta\frac{1}{2}\upsilon_\beta(1)+\beta\frac{1}{2}\upsilon_\beta(2) \end{bmatrix}$$

και

$$\upsilon_\beta(2) = 2(1-\beta)+\beta\frac{1}{2}\upsilon_\beta(1)+\beta\frac{1}{2}\upsilon_\beta(2)$$

Η πρώτη γίνεται

$$\upsilon_\beta(1) = \beta\frac{1}{2}\upsilon_\beta(1)+\beta\frac{1}{2}\upsilon_\beta(2)+(1-\beta)val\begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix}$$

Η τιμή του πίνακα στο τέλος υπολογίζεται με εξισωτικές στρατηγικές (προκύπτουν οι $(\frac{1}{3},\frac{2}{3})$ και για τους δύο παίχτες) και προκύπτει ίση με $\frac{5}{3}$ οπότε τελικά προκύπτουν οι εξισώσεις

$$\upsilon_\beta(1) = \beta\frac{1}{2}\upsilon_\beta(1)+\beta\frac{1}{2}\upsilon_\beta(2)+(1-\beta)\frac{5}{3}$$

$$\upsilon_\beta(2) = 2(1-\beta)+\beta\frac{1}{2}\upsilon_\beta(1)+\beta\frac{1}{2}\upsilon_\beta(2)$$

και τελικά προκύπτει ότι

$$\upsilon_\beta(1) = \frac{10+\beta}{6}$$

$$\upsilon_\beta(2) = \frac{12-\beta}{6}$$

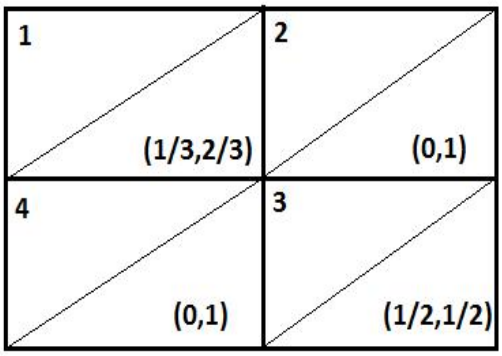
Οι βέλτιστες στάσιμες στρατηγικές για τους δύο παίχτες θα είναι

$$\mathbf{f}^*_\beta = \left((\frac{1}{3},\frac{2}{3}),(1)\right) = \mathbf{g}^*_\beta$$

$\boxtimes$

Παράδειγμα 3.3.

Έστω ένα στοχαστικό παιχνίδι μηδενικού-αθροίσματος με $S=\{1,2\}$, $A^1(1)=A^2(1)=\{1,2\}$, $A^1(2)=A^2(2)=\{1\}$ και πίνακα πληρωμών και πιθανοτήτων μετάβασης



Λύνουμε τις εξισώσεις

$$\upsilon_\beta(1) = val\begin{bmatrix}(1-\beta)1+\beta\frac{1}{3}\upsilon_\beta(1)+\beta\frac{2}{3}\upsilon_\beta(2) & (1-\beta)2+\beta 1\upsilon_\beta(2)\\ (1-\beta)4+\beta 1\upsilon_\beta(2) & (1-\beta)3+\beta\frac{1}{2}\upsilon_\beta(1)+\beta\frac{1}{2}\upsilon_\beta(2)\end{bmatrix}$$

και

$$\upsilon_\beta(2) = (1-\beta)2+\beta 1\upsilon_\beta(2).$$

Από τη δεύτερη προκύπτει άμεσα ότι

$$\upsilon_\beta(2) = 2$$

οπότε η πρώτη εξίσωση γίνεται

$$\upsilon_\beta(1) = val\begin{bmatrix}1+\frac{1}{3}\beta+\frac{\beta}{3}\upsilon_\beta(1) & 2\\ 4-2\beta & 3-2\beta+\frac{\beta}{2}\upsilon_\beta(1)\end{bmatrix}$$

Είναι προφανές ότι $\upsilon_\beta(1) > \upsilon_\beta(2)$ αφού ο παίχτης I έχει μια στρατηγική που του εξασφαλίζει πάνω από 2 (και $2 = \upsilon_\beta(2)$). Η στρατηγική αυτή είναι να παίζει πάντα τη δεύτερη γραμμή όσο είναι στον πίνακα της καταστασης 1. Τότε η πληρωμή του, ό,τι και να γίνει, θα είναι μεγαλύτερη του 2. Άρα, $3-2\beta+\frac{\beta}{2}\upsilon_\beta(1) \geq 2$. Αν $4-2\beta \geq 3-2\beta+\frac{\beta}{2}\upsilon_\beta(1) \Leftrightarrow \beta\upsilon_\beta(1) \leq 2$, τότε το κάτω δεξιά στοιχείο είναι σαγματικό σημείο και επομένως $\upsilon_\beta(1) = 3-2\beta+\frac{\beta}{2}\upsilon_\beta(1) \Rightarrow \upsilon_\beta(1) = \frac{3-2\beta}{1-\frac{\beta}{2}}$. Ελέγχουμε αν η λύση που βρήκαμε ικανοποιεί την υπόθεση που κάναμε

$$\beta\upsilon_\beta(1) \leq 2 \Leftrightarrow \beta\frac{3-2\beta}{1-\frac{\beta}{2}} \leq 2 \Leftrightarrow (\beta-1)^2 \geq 0$$

που είναι αληθές, ενώ επίσης αληθές είναι ότι $3-2\beta+\frac{\beta}{2}\frac{3-2\beta}{1-\frac{\beta}{2}} \geq 2$. Η μοναδικότητα της λύσης εξασφαλίζει ότι δεν είναι δυνατό να πάρουμε άλλη

ανισότητα με συμβατή (άλλη) λύση στην πρώτη εξίσωση. Άρα τελικά έχουμε ότι

$$\upsilon_\beta(1) = \frac{6-4\beta}{2-\beta}$$

και

$$\upsilon_\beta(2) = 2$$

Οι βέλτιστες στάσιμες στρατηγικές για τους δύο παίχτες θα είναι

$$\mathbf{f}^*_\beta = ((2),(1))$$

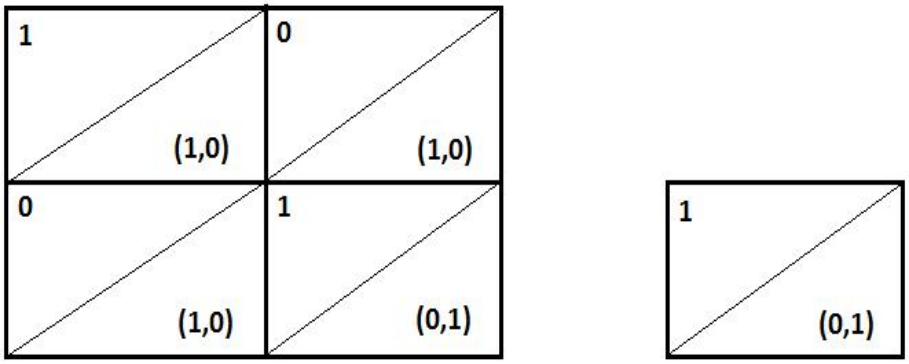
και

$$\mathbf{g}^*_\beta = ((2),(1))$$

⊠

Παράδειγμα 3.4.

Έστω ένα στοχαστικό παιχνίδι μηδενικού-αθροίσματος με $S = \{1,2\}, A^1(1) = A^2(1) = \{1,2\}, A^1(2) = A^2(2) = \{1\}$ και πίνακα πληρωμών και πιθανοτήτων μετάβασης



Γράφουμε τις εξισώσεις

$$\upsilon_\beta(1) = val\begin{bmatrix} 1-\beta+\beta\upsilon_\beta(1) & \beta\upsilon_\beta(1) \\ \beta\upsilon_\beta(1) & 1-\beta+\beta\upsilon_\beta(2) \end{bmatrix}$$

και

$$\upsilon_\beta(2) = 1-\beta+\beta\upsilon_\beta(2).$$

Από τη δεύτερη προκύπτει άμεσα ότι

$$\upsilon_\beta(2) = 1$$

οπότε η πρώτη εξίσωση γίνεται

$$\upsilon_\beta(1) = \begin{bmatrix} 1-\beta+\beta\upsilon_\beta(1) & \beta\upsilon_\beta(1) \\ \beta\upsilon_\beta(1) & 1 \end{bmatrix}$$

Αν $\beta\upsilon_\beta(1) \geq 1$, τότε το πάνω δεξιά στοιχείο είναι σαγματικό σημείο και επομένως $\upsilon_\beta(1) = \beta\upsilon_\beta(1) \Rightarrow \upsilon_\beta(1) = 0$ πράγμα άτοπο. Άρα $\beta\upsilon_\beta(1) < 1$ και επομένως δεν υπάρχει σαγματικό σημείο. Λύνοντας με εξισωτικές στρατηγικές, δηλαδή βάζοντας βάρος πιθανότητας x στην πρώτη στρατηγική του παίχτη I και $1-x$ στη δεύτερη και εξισώνοντας τις πληρωμές του παίχτη I για τις δύο στήλες του παίχτη II, παίρνουμε

$$\upsilon_\beta(1) = \frac{1-x}{1-x\beta}$$

και

$$\upsilon_\beta(1) = x$$

οπότε προκύπτει:

$$x = \upsilon_\beta(1) = \frac{1 \pm \sqrt{1-\beta}}{\beta}$$

Ελέγχουμε αν ικανοποιούν τον περιορισμό $\beta\upsilon_\beta(1) < 1$
$\beta\frac{1+\sqrt{1-\beta}}{\beta} < 1 \Rightarrow \sqrt{1-\beta} < 0$ που δεν είναι αληθές και
$\beta\frac{1-\sqrt{1-\beta}}{\beta} < 1 \Rightarrow \sqrt{1-\beta} > 0$ που είναι αληθές. Άρα τελικά έχουμε ότι

$$\upsilon_\beta(1) = \frac{1-\sqrt{1-\beta}}{\beta}$$

και

$$\upsilon_\beta(2) = 1$$

Οι βέλτιστες στάσιμες στρατηγικές για τους δύο παίχτες θα είναι

$$\mathbf{f}^*_\beta = \left(\left(\frac{1-\sqrt{1-\beta}}{\beta}, 1 - \frac{1-\sqrt{1-\beta}}{\beta} \right), (1) \right)$$

και

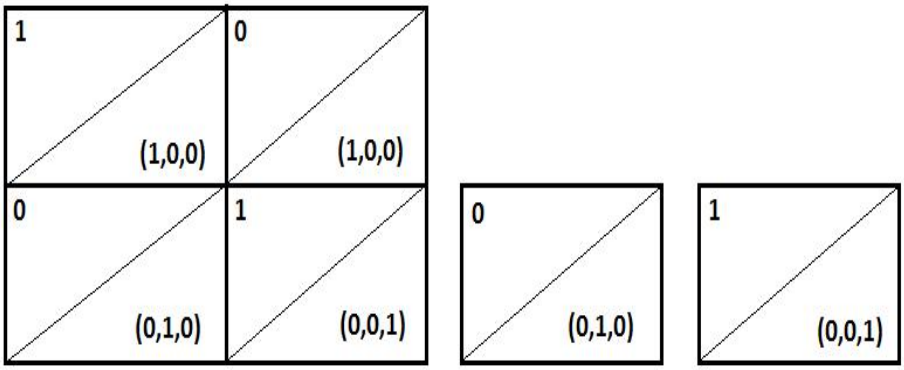
$$\mathbf{g}^*_\beta = \left(\left(\frac{1-\sqrt{1-\beta}}{\beta}, 1 - \frac{1-\sqrt{1-\beta}}{\beta} \right), (1) \right)$$

⊠

Το επόμενο παράδειγμα είναι διάσημο και γνωστό στη βιβλιογραφία ως the Big Match.

Παράδειγμα 3.5.

Έστω ένα στοχαστικό παιχνίδι μηδενικού-αθροίσματος 2-παιχτών με $S = \{1,2,3\}, A^1(1) = A^2(1) = \{1,2\}, A^1(2) = A^2(2) = \{1\}, A^1(3) = A^2(3) = \{1\}$ και πίνακα πληρωμών και πιθανοτήτων μετάβασης



Γράφουμε τις εξισώσεις

$$v_\beta(1) = \begin{bmatrix} 1-\beta+\beta v_\beta(1) & \beta v_\beta(1) \\ \beta v_\beta(2) & 1-\beta+\beta v_\beta(3) \end{bmatrix}$$

$$v_\beta(2) = \beta v_\beta(2)$$

$$v_\beta(3) = 1-\beta+\beta v_\beta(3).$$

Άμεσα προκύπτει ότι

$$v_\beta(2) = 0$$

$$v_\beta(3) = 1$$

οπότε η πρώτη εξίσωση γίνεται

$$v_\beta(1) = \begin{bmatrix} 1-\beta+\beta v_\beta(1) & \beta v_\beta(1) \\ 0 & 1 \end{bmatrix}$$

Αν $\beta v_\beta(1) \geq 1$, τότε το πάνω δεξιά στοιχείο είναι σαγματικό σημείο και επομένως $v_\beta(1) = \beta v_\beta(1) \Rightarrow v_\beta(1) = 0$. Άτοπο. Άρα $\beta v_\beta(1) < 1$ και επομένως δεν υπάρχει σαγματικό σημείο. Λύνοντας με εξισωτικές στρατηγικές καταλήγουμε στις σχέσεις

$$v_\beta(1) = \frac{1-x}{1-x\beta}$$

και

$$\upsilon_\beta(1) = \frac{x - x\beta}{1 - x\beta}$$

οπότε προκύπτει

$$\upsilon_\beta(1) = \frac{1}{2}$$

και

$$x = \frac{1}{2 - \beta}$$

Επομένως η τιμή του παιχνιδιού είναι:

$$\mathbf{v}_\beta = (\frac{1}{2}, 0, 1)$$

με βέλτιστες στάσιμες στρατηγικές

$$\mathbf{f}^*_\beta = \left((\frac{1}{2 - \beta}, \frac{1 - \beta}{2 - \beta}), (1), (1) \right).$$

και

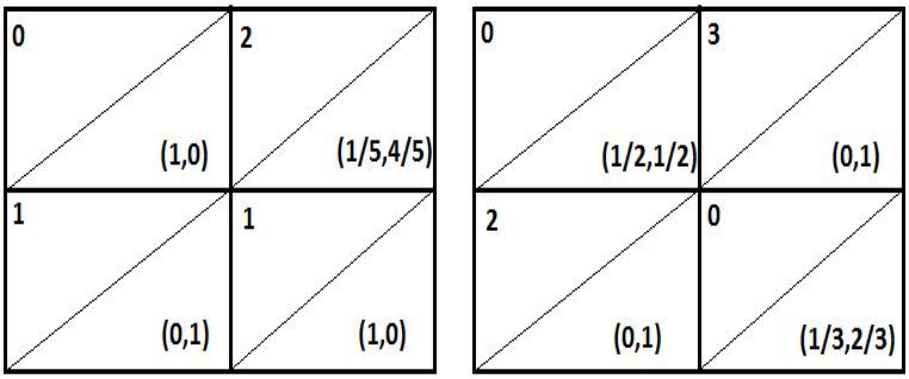
$$\mathbf{g}^*_\beta = \left((\frac{1}{2}, \frac{1}{2}), (1), (1) \right).$$

⊠

Βεβαίως η επίλυση των αποπληθωρισμένων στοχαστικών παιχνιδιών δεν είναι πάντα εύκολη όπως διαπιστώνεται στο ακόλουθο παράδειγμα.

Παράδειγμα 3.6.

Έστω ένα στοχαστικό παιχνίδι μηδενικού-αθροίσματος 2-παιχτών με $S = \{1, 2\}, A^1(1) = A^2(1) = \{1, 2\}, A^1(2) = A^2(2) = \{1, 2\}$ και πίνακα πληρωμών και πιθανοτήτων μετάβασης



Γράφουμε τις εξισώσεις

$$\upsilon_\beta(1) = val\begin{bmatrix} \beta\upsilon_\beta(1) & (1-\beta)2 + \beta\frac{1}{5}\upsilon_\beta(1) + \beta 1\frac{4}{5}\upsilon_\beta(2) \\ 1-\beta+\beta\upsilon_\beta(2) & 1-\beta+\beta\upsilon_\beta(1) \end{bmatrix}$$

και

$$\upsilon_\beta(2) = val\begin{bmatrix} \beta\frac{1}{2}\upsilon_\beta(1) + \beta\frac{1}{2}\upsilon_\beta(2) & (1-\beta)3 + \beta\upsilon_\beta(2) \\ (1-\beta)2+\beta\upsilon_\beta(2) & \beta\frac{1}{3}\upsilon_\beta(1) + \beta\frac{2}{3}\upsilon_\beta(2) \end{bmatrix}$$

Παρατηρούμε ότι και οι δύο πίνακες είναι εντελώς μπλεγμένοι και επομένως είναι πιο δύσκολη η λύση των εξισώσεων. Επίσης, είναι πολύ πιθανό να προκύψουν παραπάνω από μία λύσεις οπότε θα πρέπει να ελέγξουμε ποια από αυτές ικανοποιεί τη μοναδική λύση των αρχικών εξισώσεων.

⊠

3.3 Οριακή αποπληθωρισμένη εξίσωση (Limit Discount Equation)

Σε αυτή την παράγραφο θα αναφερθούμε πολύ επιγραμματικά στην οριακή αποπληθωρισμένη εξίσωση. Για το σκοπό αυτό εισάγουμε το σώμα των σειρών Puiseux που αποδεικνύεται πολύ χρήσιμο στη μελέτη της ασυμπτωτικής συμπεριφοράς των λύσεων των β-αποπληθωρισμένων στοχαστικών παιχνιδιών όταν το β τείνει στο 1.

Για M θετικό ακέραιο, ορίζουμε F_M το σύνολο όλων των $\sum_{k=K}^{\infty} c_k x^{\frac{k}{M}}$ όπου K είναι ακέραιος, $c_k \in \mathbb{R}$ είναι τέτοια ώστε η σειρά $\sum_{k=K}^{\infty} c_k x^{\frac{k}{M}}$ να συγκλίνει για κάθε αρκετά μικρό θετικό πραγματικό αριθμό x.

Έτσι, τα στοιχεία που ανήκουν στο F_M είναι δυναμοσειρές ως προς $x^{\frac{1}{M}}$. Η πρόσθεση και ο πολλαπλασιασμός στο σύνολο F_M ορίζονται με το συνήθη τρόπο για δυναμοσειρές, δηλαδή

$$\sum_{k=K_1}^{\infty} c_k x^{\frac{k}{M}} + \sum_{k=K_2}^{\infty} d_k x^{\frac{k}{M}} := \sum_{k=\min(K_1,K_2)}^{\infty} (c_k + d_k) x^{\frac{k}{M}}$$

και

$$\left(\sum_{k=K_1}^{\infty} c_k x^{\frac{k}{M}}\right)\left(\sum_{k=K_2}^{\infty} d_k x^{\frac{k}{M}}\right) := \sum_{k=K_1+K_2}^{\infty} \left(\sum_{i+j=k} c_i d_j\right) x^{\frac{k}{M}}$$

(Παρατήρηση: στην πρόσθεση, αν $K_1 > K_2$, ορίζουμε $c_k = 0$ για $k = K_2, \ldots, K_1 - 1$).

Επιπλέον, ορίζουμε $\sum_{k=K}^{\infty} c_k x^{\frac{k}{M}} > 0$ αν και μόνο αν $c_{k^*} > 0$, όπου k^* είναι ο μικρότερος ακέραιος k τέτοιος ώστε $c_k \neq 0$. Με στοιχειώδη ανάλυση προκύπτει ότι το F_M είναι ένα διατεταγμένο σώμα. Έστω $F := \bigcup_{M=1}^{\infty} F_M$. Τότε το F είναι και αυτό ένα διατεταγμένο σώμα και καλείται σώμα των πραγματικών σειρών Puiseux.

Ορισμός 3.3.1. *Το σύνολο των N εξισώσεων (μία για κάθε κατάσταση $s \in S$) με άγνωστο τη μεταβλητή $\mathbf{v} = (v(1), v(2), \ldots, v(N)) \in F^N$*

$$v(s) = val\left[xr(s, a^1, a^2) + (1-x)\sum_{s'=1}^{N} p_{ss'}(a^1, a^2) v(s')\right]$$

όπου η αγκύλη παριστάνει τον $m^1(s) \times m^2(s)$ πίνακα με γραμμές $a^1 = 1, \ldots, m^1(s)$ και στήλες $a^2 = 1, \ldots, m^2(s)$, και όπου $x \in (0,1)$, καλείται οριακή αποπληθωρισμένη εξίσωση (limit discount equation).

Θεώρημα 3.3.1. *Η οριακή αποπληθωρισμένη εξίσωση έχει μοναδική λύση πάνω στο καρτεσιανό γινόμενο F^N της μορφής*

$$\mathbf{v}^* = \sum_{k=0}^{\infty} \mathbf{c}_k x^{\frac{k}{M}} \in F^N \quad \mu\epsilon \quad \mathbf{c}_k \in \mathbb{R}^N.$$

Η απόδειξη του θεωρήματος είναι αρκετά πολύπλοκη και δεν θα παρατεθεί σε αυτήν την εργασία. Η πρώτη απόδειξη δόθηκε από τους T. Bewley και E. Kohlberg (1976a, 1976b).

Θεώρημα 3.3.2. *Έστω ένα στοχαστικό παιχνίδι στο οποίο και οι δύο παίχτες διαθέτουν βέλτιστες στάσιμες στρατηγικές σύμφωνα με το κριτήριο οριακής μέσης πληρωμής. Τότε η λύση της οριακής αποπληθωρισμένης εξίσωσης έχει την ιδιότητα:*

$$\mathbf{c}_1 = \mathbf{c}_2 = \cdots = \mathbf{c}_{M-1} = \mathbf{0}.$$

Η απόδειξη του συγκεκριμένου θεωρήματος επίσης παραλείπεται. Ο αναγνώστης μπορεί να ανατρέξει στους Filar και Vrieze (1997), Κεφάλαιο 4.3.4.

Πόρισμα 3.3.1. *Έστω $\mathbf{c}_0 = \upsilon \mathbf{1}_N$, όπου $\upsilon \in \mathbb{R}$, δηλαδή ο πρώτος όρος της λύσης της οριακής αποπληθωρισμένης εξίσωσης είναι ο ίδιος, υ, για κάθε κατάσταση και έστω ότι και οι δύο παίχτες διαθέτουν βέλτιστες στάσιμες στρατηγικές σύμφωνα με το κριτήριο της οριακής μέσης πληρωμής. Τότε ισχύει η ακόλουθη σχέση που συνδέει τα $\mathbf{c}_0 = \upsilon \mathbf{1}_N$ και $\mathbf{c}_M$*

$$\upsilon + c_M(s) = val\left[r(s, a^1, a^2) + \sum_{s'=1}^{N} p_{ss'}(a^1, a^2) c_M(s')\right]$$

για κάθε $s \in S$.

Το πόρισμα προκύπτει από το προηγούμενο θεώρημα, από την υπόθεση ότι $\mathbf{c}_0$ είναι σταθερά και από τον πολλαπλασιασμό της οριακής αποπληθωρισμένης εξίσωσης με το x^{-1}.

Προφανώς, από το παραπάνω πόρισμα προκύπτει ότι η οριακή μέση τιμή ισούται με $\upsilon \mathbf{1}_N$.

Κεφάλαιο 4

Αδιαχώριστα Στοχαστικά Παιχνίδια Αναμενόμενης Μέσης Πληρωμής (**Average Reward Irreducible Stochastic Games)**

4.1 Θεωρία

Υπενθυμίζουμε η οριακή μέση συνολική πληρωμή για ένα ζευγάρι (συμπεριφορικών) στρατηγικών (π^1, π^2) των δύο παιχτών και μία αρχική κατάσταση $s \in S$ για ένα στοχαστικό παιχνίδι 2-παιχτών μηδενικού-αθροίσματος ορίζεται ως

$$v_\alpha(s, \pi^1, \pi^2) = \liminf_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{\infty} \mathbb{E}_{s\pi^1\pi^2}[r(S_t, A_t^1, A_t^2)]$$

Κατά τα γνωστά, το εξεταζόμενο παιχνίδι θα έχει τιμή αν

$$\sup_{\pi^1} \inf_{\pi^2} v_\alpha(s, \pi^1, \pi^2) = \inf_{\pi^2} \sup_{\pi^1} v_\alpha(s, \pi^1, \pi^2).$$

Αν υπάρχει η τιμή, έστω $\mathbf{v}_\alpha = (v_\alpha(1), \ldots, v_\alpha(N))$, το επόμενο ερώτημα είναι αν υπάρχουν βέλτιστες στρατηγικές για τους δύο παίχτες ή έστω ε-βέλτιστες στρατηγικές.

Οι στρατηγικές $\tilde{\pi}^1, \tilde{\pi}^2$ είναι βέλτιστες όταν για κάθε στρατηγική π^1 του παίχτη I και για κάθε στρατηγική π^2 του παίχτη II ισχύει

$$v_\alpha(s, \pi^1, \tilde{\pi}^2) \leq v_\alpha(s) \leq v_\alpha(s, \tilde{\pi}^1, \pi^2).$$

Αντίστοιχα, οι στρατηγικές $\pi^1_\varepsilon, \pi^2_\varepsilon$ είναι ε-βέλτιστες όταν

$$\upsilon_\alpha(s, \pi^1, \pi^2_\varepsilon) - \varepsilon \leq \upsilon_\alpha(s) \leq \upsilon_\alpha(s, \pi^1_\varepsilon, \pi^2) + \varepsilon$$

για οποιεσδήποτε στρατηγικές π^1, π^2.

Στη γενική περίπτωση, οι παίχτες δεν διαθέτουν βέλτιστες στρατηγικές για το οριακό μέσο κριτήριο πληρωμής. Ένας καλός τρόπος προσέγγισης, είναι να εξετάσουμε μία ακολουθία $\beta-$αποπληθωρισμένων παιχνιδιών με $\beta \to 1$ και να μελετήσουμε τις οριακές ιδιότητες αυτής της ακολουθίας.

Στην παρούσα εργασία εμείς θα ασχοληθούμε μόνο με τα αδιαχώριστα στοχαστικά παιχνίδια. Επιπροσθέτως, τα θεωρήματα 1.5.3 και 1.5.4 μας δίνουν τη δυνατότητα να εξετάσουμε τα συγκεκριμένα παιχνίδια όταν οι παίχτες περιορίζονται σε στάσιμες στρατηγικές.

Υπενθυμίζουμε ότι ένα στοχαστικό παιχνίδι είναι αδιαχώριστο όταν για κάθε ζευγάρι στάσιμων στρατηγικών $(\mathbf{f}, \mathbf{g})$ των δύο παιχτών, ο αντίστοιχος πίνακας πιθανοτήτων μετάβασης $P(\mathbf{f}, \mathbf{g})$ σχηματίζει μια μαρκοβιανή αλυσίδα η οποία έχει ένα μόνο αδιαχώριστο σύνολο καταστάσεων που περιέχει όλο το χώρο καταστάσεων.

Θεώρημα 4.1.1. *Αν ένα ζευγάρι στάσιμων στρατηγικών $(\mathbf{f}, \mathbf{g})$ των δύο παιχτών είναι τέτοιο ώστε ο αντίστοιχος πίνακας πιθανοτήτων μετάβασης $P(\mathbf{f}, \mathbf{g})$ να επάγει μια αδιαχώριστη μαρκοβιανή αλυσίδα, τότε*

1. *Η $s-$οστη συνιστώσα της στάσιμης κατανομής του πίνακα $P(\mathbf{f}, \mathbf{g})$, $q(s, \mathbf{f}, \mathbf{g}) = \lim_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} \mathbb{P}_{s_0 \mathbf{fg}}[S_t = s]$ υπάρχει για οποιαδήποτε κατάσταση $s \in S$ και είναι ανεξάρτητη από την αρχική κατάσταση s_0.*
2. *$q(s, \mathbf{f}, \mathbf{g}) > 0$, για κάθε $s \in S$.*
3. *Αν $Q(\mathbf{f}, \mathbf{g}) := \lim_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} P^t(\mathbf{f}, \mathbf{g})$ το Σέζαρο όριο του $P(\mathbf{f}, \mathbf{g})$, τότε κάθε γραμμή του πίνακα $Q(\mathbf{f}, \mathbf{g})$ ισούται με $\mathbf{q}^T$ και επιπλέον*

$$Q(\mathbf{f}, \mathbf{g}) = Q(\mathbf{f}, \mathbf{g}) P(\mathbf{f}, \mathbf{g})$$

4. *$\upsilon_\alpha(s_0, \mathbf{f}, \mathbf{g}) = \sum_{s=1}^{N} q(s, \mathbf{f}, \mathbf{g}) r(s, \mathbf{f}, \mathbf{g})$ για κάθε $s_0 \in S$.*
5. *Για κάθε ζευγάρι $(\upsilon, \mathbf{w})$ με $\upsilon \in \mathbb{R}$ και $\mathbf{w} \in \mathbb{R}^N$ που ικανοποιεί την εξίσωση*

$$\mathbf{w} + \upsilon \mathbf{1}_N = r(\mathbf{f}, \mathbf{g}) + P(\mathbf{f}, \mathbf{g}) \mathbf{w}, \qquad (1)$$

ισχύει ότι $\upsilon_\alpha(s_0, \mathbf{f}, \mathbf{g}) = \upsilon$ για κάθε $s \in S$.

6. Ο πίνακας $(I - P + Q)$ *είναι αντιστρέψιμος και το ζευγάρι* $(\upsilon, \mathbf{w})$ *με*

$$\upsilon \mathbf{1}_N = \mathbf{v}_\alpha(\mathbf{f}, \mathbf{g})$$

και

$$\mathbf{w} = (I - P(\mathbf{f}, \mathbf{g}) + Q(\mathbf{f}, \mathbf{g}))^{-1} (r(\mathbf{f}, \mathbf{g}) - \upsilon \mathbf{1}_N)$$

ικανοποιεί τη σχέση (1).

Απόδειξη

1. Η ύπαρξη του ορίου έπεται από το λήμμα 2.2.1. Για οποιαδήποτε αρχική κατάσταση η στάσιμη κατανομή ικανοποιεί τη σχέση $\mathbf{q}^T = \mathbf{q}^T P(f, g)$, η οποία έχει μοναδική λύση. Έτσι η στάσιμη κατανομή είναι ανεξάρτητη από την αρχική κατάσταση.

2. Έχει αποδειχτεί ήδη στο λήμμα 2.2.1.

3. Ένα γνωστό αποτέλεσμα στη θεωρία μαρκοβιανών αλυσίδων είναι ότι

$$\mathbb{P}_{s_0 \mathbf{fg}}[S_t = s] = (P^t(\mathbf{f}, \mathbf{g}))_{s_0 s}$$

όπου $(P^t(\mathbf{f}, \mathbf{g}))_{s_0 s}$ είναι το $(s_0, s)-$στοιχείο του πίνακα μεταβάσεων σε t βήματα και ο ισχυρισμός έπεται από το *1* του θεωρήματος. Η σχέση $Q(\mathbf{f}, \mathbf{g}) = Q(\mathbf{f}, \mathbf{g}) P(\mathbf{f}, \mathbf{g})$ έχει αποδειχτεί στην πρόταση 2.2.4.

4. Δεδομένου ότι $q(s, \mathbf{f}, \mathbf{g}) = \lim_{T \to \infty} \frac{1}{T+1} \sum_{t=0}^{T} \mathbb{P}_{s_0 \mathbf{fg}}[S_t = s]$ για κάθε $s \in S$ και αφού το S είναι πεπερασμένο, θα ισχύει

$$\lim_{T \to \infty} \left[\frac{1}{T+1} \sum_{t=0}^{T} \mathbb{P}_{s_0 \mathbf{fg}}[S_t = 1] r(1, \mathbf{f}, \mathbf{g}) + \cdots + \sum_{t=0}^{T} \mathbb{P}_{s_0 \mathbf{fg}}[S_t = N] r(N, \mathbf{f}, \mathbf{g}) \right] =$$

$$q(1, \mathbf{f}, \mathbf{g}) r(1, \mathbf{f}, \mathbf{g}) + \cdots + q(N, \mathbf{f}, \mathbf{g}) r(N, \mathbf{f}, \mathbf{g}) =$$

$$\sum_{s \in S} q(s, \mathbf{f}, \mathbf{g}) r(s, \mathbf{f}, \mathbf{g})$$

Αλλά

$$v_\alpha(s_0, \mathbf{f}, \mathbf{g}) := \liminf_{T \to \infty} \frac{1}{T+1} \sum_{t=0}^{T} \sum_{s \in S} \mathbb{P}_{s_0 \mathbf{fg}}[S_t = s] r(s, \mathbf{f}, \mathbf{g})$$

Άρα

$$v_\alpha(s_0, \mathbf{f}, \mathbf{g}) = \sum_{s \in S} q(s, \mathbf{f}, \mathbf{g}) r(s, \mathbf{f}, \mathbf{g}).$$

5. Πολλαπλασιάζοντας τη σχέση (1) με τον πίνακα $Q(\mathbf{f},\mathbf{g})$ από τα αριστερά και χρησιμοποιώντας τα *3* και *4* του θεωρήματος, έχουμε

$$Q(\mathbf{f},\mathbf{g})\mathbf{w} + \upsilon Q(\mathbf{f},\mathbf{g})\mathbf{1}_N = Q(\mathbf{f},\mathbf{g})r(\mathbf{f},\mathbf{g}) + Q(\mathbf{f},\mathbf{g})P(\mathbf{f},\mathbf{g})\mathbf{w}$$

$$\Rightarrow Q(\mathbf{f},\mathbf{g})\mathbf{w} + \upsilon Q(\mathbf{f},\mathbf{g})\mathbf{1}_N = \mathbf{v}_\alpha(\mathbf{f},\mathbf{g}) + Q(\mathbf{f},\mathbf{g})\mathbf{w}$$

$$\Rightarrow \upsilon Q(\mathbf{f},\mathbf{g})\mathbf{1}_N = \mathbf{v}_\alpha(\mathbf{f},\mathbf{g})$$

$$\Rightarrow \upsilon \mathbf{1}_N = \mathbf{v}_\alpha(\mathbf{f},\mathbf{g})$$

και συνεπώς $\upsilon_\alpha(s,\mathbf{f},\mathbf{g}) = \upsilon$ για κάθε $s \in S$.

6. Σε αυτό το κομμάτι της απόδειξης, για απλούστευση δεν σημειώνουμε την εξάρτηση των πινάκων $P(\mathbf{f},\mathbf{g})$ και $Q(\mathbf{f},\mathbf{g})$ από τις στρατηγικές $\mathbf{f}$ και $\mathbf{g}$ των δύο παιχτών. Πρώτα θα δείξουμε ότι ο πίνακας $(I - P + Q)$ είναι αντιστρέψιμος. Πράγματι, έστω ότι δεν είναι αντιστρέψιμος. Τότε υπάρχει $\boldsymbol{\lambda} \in \mathbb{R}^N,\ \boldsymbol{\lambda} \neq \mathbf{0}$ τέτοιο ώστε

$$(I - P + Q)\ \boldsymbol{\lambda} = \mathbf{0}$$

$$\Rightarrow Q(I - P + Q)\ \boldsymbol{\lambda} = Q\ \boldsymbol{\lambda} = \mathbf{0}.$$

Αλλά τότε

$$(I - P)\ \boldsymbol{\lambda} = \mathbf{0} \quad \eta \quad \boldsymbol{\lambda} = P\ \boldsymbol{\lambda}$$

Έστω $\tilde{S}$ το σύνολο των καταστάσεων στις οποίες παίρνεται η μέγιστη συντεταγμένη του διανύσματος $\boldsymbol{\lambda}$, δηλαδή $\tilde{S} := \{s \in S : \lambda(s) = \max_{s'} \lambda(s')\}$. Θα δείξουμε ότι $\tilde{S} = S$. Έστω ότι $\tilde{S} \neq S$. Επειδή το S είναι αδιαχώριστο, θα υπάρχουν $s \in \tilde{S}$ και $s' \in S \setminus \tilde{S}$ τέτοια ώστε $p_{ss'}(\mathbf{f},\mathbf{g}) > 0$. Αλλά τότε, αφού $\boldsymbol{\lambda} = P\boldsymbol{\lambda}$

$$\lambda(s) = \sum_{s'=1}^{N} p_{ss'}(\mathbf{f},\mathbf{g})\lambda(s') < \max_{s'} \lambda(s') = \lambda(s)$$

που είναι αντίφαση. Έτσι $\tilde{S} = S$, δηλαδή το $\boldsymbol{\lambda}$ έχει όλες τις συντεταγμένες ίσες και επομένως $\boldsymbol{\lambda} = \tilde{\lambda}\mathbf{1}_N$ για κάποιο $\tilde{\lambda} \in \mathbb{R}$. Χρησιμοποιώντας ότι $Q\ \boldsymbol{\lambda} = \mathbf{0}$ παίρνουμε ότι

$$Q\tilde{\lambda}\mathbf{1}_N = \mathbf{0} \Rightarrow \tilde{\lambda}\sum_{s\in S} q(s) = \mathbf{0} \Rightarrow \tilde{\lambda}\mathbf{1}_N = \mathbf{0} \Rightarrow \tilde{\lambda} = 0$$

και άρα $\boldsymbol{\lambda} = \mathbf{0}$ το οποίο οδηγεί σε άτοπο. Άρα, ο πίνακας $(I - P + Q)$ είναι αντιστρέψιμος.

Απομένει να δείξουμε ότι το $(v, \mathbf{w})$ ικανοποιεί τη σχέση (1). Έστω $v\mathbf{1}_N = \mathbf{v}_\alpha$ (το διάνυσμα $\mathbf{v}_\alpha$ έχει όλες τις συντεταγμένες ίσες αφού $\mathbf{v}_\alpha(s_0) = Q\mathbf{r}$ ανεξάρτητο του s_0) και $\mathbf{w} = (I - P + Q)^{-1}(\mathbf{r} - v\mathbf{1}_N)$. Διερωτόμαστε αν

$$(I - P + Q)^{-1}(\mathbf{r} - v\mathbf{1}_N) + v\mathbf{1}_N = \mathbf{r} + P(I - P + Q)^{-1}(\mathbf{r} - v\mathbf{1}_N) \Leftrightarrow$$

$$\mathbf{r} - v\mathbf{1}_N + (I - P + Q)v\mathbf{1}_N = (I - P + Q)\mathbf{r} + (I - P + Q)P(I - P + Q)^{-1}(\mathbf{r} - v\mathbf{1}_N)$$

Αλλά $Pv\mathbf{1}_N = v\mathbf{1}_N$ (ο P είναι στοχαστικός πίνακας και άρα $P\mathbf{1}_N = \mathbf{1}_N$) και $Q\mathbf{1}_N = \mathbf{1}_N$ (ο Q είναι στοχαστικός πίνακας και άρα $Q\mathbf{1}_N = \mathbf{1}_N$). Άρα η σχέση που διερωτόμαστε αν ισχύει γίνεται

$$\mathbf{r} = \mathbf{r} - P\mathbf{r} + Q\mathbf{r} + (I - P + Q)P(I - P + Q)^{-1}(\mathbf{r} - v\mathbf{1}_N)$$

Τώρα παρατηρούμε ότι $(I - P + Q)P = P(I - P + Q)$. Πράγματι

$$(I - P + Q)P = P(I - P + Q) \Leftrightarrow P - P^2 + QP = P - P^2 + PQ \Leftrightarrow QP = PQ$$

που είναι αληθές από την πρόταση 2.2.4. Άρα, η σχέση που διερωτόμαστε αν ισχύει γίνεται

$$\mathbf{0} = -P\mathbf{r} + Q\mathbf{r} + P\mathbf{r} - Pv\mathbf{1}_N$$

Αλλά $Pv\mathbf{1}_N = v\mathbf{1}_N$ και άρα η σχέση είναι ισοδύναμη με $\mathbf{v}_\alpha = Q\mathbf{r}$ που είναι αληθές.

$\boxtimes$

Τα αδιαχώριστα στοχαστικά παιχνίδια είναι εύκολα επιλύσιμα επειδή η οριακή μέση πληρωμή είναι συνεχής ως προς τις στρατηγικές $\mathbf{f}$ και $\mathbf{g}$ των δύο παιχτών.

Λήμμα 4.1.1. *Η οριακή μέση πληρωμή $\mathbf{v}_\alpha(\mathbf{f}, \mathbf{g})$ είναι συνεχής συνάρτηση πάνω στο χώρο των στάσιμων στρατηγικών $\mathbf{f}$ και $\mathbf{g}$ των δύο παιχτών.*

Απόδειξη

Έστω $(\mathbf{f}_n, \mathbf{g}_n) \to (\mathbf{f}, \mathbf{g})$. Τότε, $P(\mathbf{f}_n, \mathbf{g}_n) \to P(\mathbf{f}, \mathbf{g})$. Έστω $\mathbf{q}$ σημείο συσσώρευσης της ακολουθίας $\mathbf{q}_n, n = 1, 2, \ldots$. Αφού $\mathbf{q}_n P(\mathbf{f}_n, \mathbf{g}_n) = \mathbf{q}_n$ έχουμε ότι $\mathbf{q}P(\mathbf{f}, \mathbf{g}) = \mathbf{q}$. Επομένως, $\mathbf{q}$ ισούται με τη μοναδική στάσιμη κατανομή του πίνακα $P(\mathbf{f}, \mathbf{g})$, αποδεικνύοντας ότι η ακολουθία $\mathbf{q}_n$ συγκλίνει στο $\mathbf{q}$. Αν υπήρχε και δεύτερο σημείο συσσώρευσης $\mathbf{q}'$ της $\mathbf{q}_n$, τότε με τον ίδιο τρόπο $\mathbf{q}'P = \mathbf{q}'$. Αλλά υπάρχει μόνο ένα διάνυσμα πιθανότητας, έστω $\mathbf{u}$, που να ικανοποιεί την $\mathbf{u}P = \mathbf{u}$. Άρα $\mathbf{u} = \mathbf{q} = \mathbf{q}'$, δηλαδή η ακολουθία έχει μοναδικό σημείο συσσώρευσης. Το συμπέρασμα έπεται από το *4* του θεωρήματος 4.1.1.

$\boxtimes$

Στη συνέχεια ακολουθεί το βασικό θεώρημα για τα αδιαχώριστα στοχαστικά παιχνίδια.

Θεώρημα 4.1.2.

1. *Το διάνυσμα της τιμής υπάρχει για το οριακό μέσο κριτήριο πληρωμής και η τιμή είναι η ίδια για κάθε αρχική κατάσταση.*
2. *Και οι δύο παίχτες διαθέτουν βέλτιστες στάσιμες στρατηγικές.*
3. *Η λύση της οριακής αποπληθωρισμένης εξίσωσης μπορεί να γραφτεί ως εξής:*

$$\mathbf{v}^*(x) = \upsilon_\alpha \mathbf{1}_N + \sum_{k=M}^{\infty} \mathbf{c}_k x^{\frac{k}{M}},$$

όπου $\upsilon_\alpha \mathbf{1}_N$ ισούται με την τιμή του οριακού μέσου παιχνιδιού.

Απόδειξη

1. Θεωρούμε τις ανταποκρίσεις βέλτιστης απάντησης των δύο παιχτών δηλαδή, έστω $MDP(\mathbf{f})$ το μαρκοβιανό πρόβλημα αποφάσεων που αντιμετωπίζει ο παίχτης II, όταν ο παίχτης I χρησιμοποιεί τη στάσιμη στρατηγική $\mathbf{f}$ και έστω αντίστοιχα το $MDP(\mathbf{g})$ για τον παίχτη I. Η ανταπόκριση βέλτιστης απάντησης του παίχτη II, όταν το παιχνίδι περιοριστεί σε στάσιμες στρατηγικές, σύμφωνα με το μέσο κριτήριο πληρωμής, αποτελείται από όλες τις βέλτιστες στάσιμες στρατηγικές του παίχτη II στο $MDP(\mathbf{f})$ (θεώρημα 1.5.3) και το συμβολίζουμε με $\mathbf{G}(\mathbf{f})$. Τα ακόλουθα είναι βασικά αποτελέσματα στις μαρκοβιανές διαδικασίες αποφάσεων:

(i) $\mathbf{G}(\mathbf{f})$ είναι συμπαγές και κυρτό σύνολο.

(ii) $\mathbf{G}(\mathbf{f})$ είναι το ίδιο για κάθε αρχική κατάσταση.

Όσο αφορά το (i) πρώτα θα δείξουμε ότι το σύνολο $\mathbf{G}(\mathbf{f})$ είναι κλειστό. Πράγματι, έστω μια ακολουθία $\mathbf{g}_n \in \mathbf{G}(\mathbf{f})$ και έστω ότι $\mathbf{g}_n \to \mathbf{g}_0$. Θα δείξουμε ότι $\mathbf{g}_0 \in \mathbf{G}(\mathbf{f})$. Αφού $\mathbf{g}_n \in \mathbf{G}(\mathbf{f})$ έχουμε

$$\mathbf{v}_\alpha(\mathbf{f}, \mathbf{g}_n) = \max_{\mathbf{u} \in \mathbf{G}} \mathbf{v}_\alpha(\mathbf{f}, \mathbf{u}) \leq \mathbf{v}_\alpha(\mathbf{f}, \mathbf{u})$$

για οποιαδήποτε στάσιμη στρατηγική $\mathbf{u} \in \mathbf{G}$ του παίχτη II. Γνωρίζουμε από το λήμμα 4.1.1 ότι η συνάρτηση $\mathbf{v}_\alpha : \mathbf{F} \times \mathbf{G} \to \mathbb{R}$ είναι συνεχής. Επομένως, παίρνοντας όρια για $n \to \infty$ έχουμε

$$\mathbf{v}_\alpha(\mathbf{f}, \mathbf{g}_0) = \lim_{n \to \infty} \mathbf{v}_\alpha(\mathbf{f}, \mathbf{g}_n) = \lim_{n \to \infty} \max_{\mathbf{u} \in \mathbf{G}} \mathbf{v}_\alpha(\mathbf{f}, \mathbf{u}) = \max_{\mathbf{u} \in \mathbf{G}} \mathbf{v}_\alpha(\mathbf{f}, \mathbf{u})$$

Άρα $\mathbf{g}_0 \in \mathbf{G}(\mathbf{f})$ και άρα $\mathbf{G}(\mathbf{f})$ κλειστό σύνολο.

Επίσης το σύνολο $\mathbf{G}(\mathbf{f})$ είναι φραγμένο ως υποσύνολο του φραγμένου συνόλου $\mathbf{G}$. Άρα το σύνολο $\mathbf{G}(\mathbf{f})$ είναι συμπαγές ως κλειστό και φραγμένο.

Στη συνέχεια θα δείξουμε ότι το σύνολο $\mathbf{G}(\mathbf{f})$ είναι κυρτό. Έστω $\mathbf{g}_1 \in \mathbf{G}(\mathbf{f})$ δηλαδή $\mathbf{v}_\alpha(\mathbf{f}, \mathbf{g}_1) = \max_{\mathbf{u}\in\mathbf{G}} \mathbf{v}_\alpha(\mathbf{f}, \mathbf{u})$ και έστω $\mathbf{g}_2 \in \mathbf{G}(\mathbf{f})$ οπότε $\mathbf{v}_\alpha(\mathbf{f}, \mathbf{g}_2) = \max_{\mathbf{u}\in\mathbf{G}} \mathbf{v}_\alpha(\mathbf{f}, \mathbf{u})$. Θα δείξουμε ότι για κάθε $\lambda \in [0, 1]$ ισχύει ότι $\lambda\mathbf{g}_1 + (1 - \lambda)\mathbf{g}_2 \in \mathbf{G}(\mathbf{f})$ δηλαδή ότι

$$\mathbf{v}_\alpha(\mathbf{f}, \lambda\mathbf{g}_1 + (1 - \lambda)\mathbf{g}_2) = \max_{\mathbf{u}\in\mathbf{G}} \mathbf{v}_\alpha(\mathbf{f}, \mathbf{u})$$

Έστω $\lambda \in [0, 1]$. Ισχύει ότι

$$\mathbf{v}_\alpha(\mathbf{f}, \mathbf{g}) = \liminf_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} \mathbb{E}_{s_0\mathbf{fg}} R_t(s, a^1, a^2)$$

Επειδή το στοχαστικό παιχνίδι είναι αδιαχώριστο θα ισχύει

$$\mathbf{v}_\alpha(\mathbf{f}, \mathbf{g}) = \lim_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} \mathbb{E}_{s_0\mathbf{fg}} R_t(s, a^1, a^2)$$

Από τη γραμμικότητα του τελεστή τη μέσης τιμής έχουμε

$$\mathbb{E}_{s_0\mathbf{f}(\lambda\mathbf{g}_1+(1-\lambda)\mathbf{g}_2)} R_t(s, a^1, a^2) = \lambda\mathbb{E}_{s_0\mathbf{fg}_1} R_t(s, a^1, a^2) + (1 - \lambda)\mathbb{E}_{s_0\mathbf{fg}_2} R_t(s, a^1, a^2)$$

και από τη γραμμικότητα του τελεστή του ορίου προκύπτει

$$\lim_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} \mathbb{E}_{s_0\mathbf{f}(\lambda\mathbf{g}_1+(1-\lambda)\mathbf{g}_2)} R_t(s, a^1, a^2) =$$

$$= \lambda \lim_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} \mathbb{E}_{s_0\mathbf{fg}_1} R_t(s, a^1, a^2) + (1-\lambda) \lim_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} \mathbb{E}_{s_0\mathbf{fg}_2} R_t(s, a^1, a^2)$$

δηλαδή

$$\mathbf{v}_\alpha(\mathbf{f}, \lambda\mathbf{g}_1 + (1 - \lambda)\mathbf{g}_2) = \lambda\mathbf{v}_\alpha(\mathbf{f}, \mathbf{g}_1) + (1 - \lambda)\mathbf{v}_\alpha(\mathbf{f}, \mathbf{g}_2) =$$

$$\lambda \max_{\mathbf{u}\in\mathbf{G}} \mathbf{v}_\alpha(\mathbf{f}, \mathbf{u}) + (1 - \lambda) \max_{\mathbf{u}\in\mathbf{G}} \mathbf{v}_\alpha(\mathbf{f}, \mathbf{u}) =$$

$$= \max_{\mathbf{u}\in\mathbf{G}} \mathbf{v}_\alpha(\mathbf{f}, \mathbf{u})$$

Άρα το σύνολο $\mathbf{G}(\mathbf{f})$ είναι κυρτό.

Το (ii) είναι άμεσο αφού το στοχαστικό παιχνίδι είναι αδιαχώριστο και οι παίχτες χρησιμοποιούν στάσιμες στρατηγικές.

Τα ίδια αποτελέσματα ισχύουν αντίστοιχα και για το $\mathbf{F}(\mathbf{g})$ το οποίο ορίζεται ως η ανταπόκριση βέλτιστης απάντησης του παίχτη I, όταν το παιχνίδι περιοριστεί σε στάσιμες στρατηγικές, σύμφωνα με το μέσο κριτήριο πληρωμής, δηλαδή όλες οι βέλτιστες στάσιμες στρατηγικές του παίχτη I στο $MDP(\mathbf{g})$ (θεώρημα 1.5.3).

Στη συνέχεια θα δείξουμε ότι η απεικόνιση $U(\mathbf{f},\mathbf{g}) := (\mathbf{F}(\mathbf{g}),\mathbf{G}(\mathbf{f}))$ είναι άνω ημισυνεχής στο χώρο των στάσιμων στρατηγικών. Πρώτα θα δείξουμε ότι η ανταπόκριση $\mathbf{F}(\mathbf{g})$ από το $\mathbf{G}$ στο $\mathbf{F}$ είναι άνω ημισυνεχής.

Καταρχήν παρατηρούμε ότι για κάθε στάσιμη στρατηγική $\mathbf{g} \in \mathbf{G}$ του παίχτη II η συνέχεια της $\mathbf{v}_\alpha(\cdot,\mathbf{g})$ πάνω στο συμπαγές $\mathbf{F}$ εξασφαλίζει ότι το $\mathbf{F}(\mathbf{g})$ είναι μη κενό σύνολο. Έστω λοιπόν ακολουθία $(\mathbf{f}_n,\mathbf{g}_n) \to (\mathbf{f}_0,\mathbf{g}_0)$ με $\mathbf{f}_n \in \mathbf{F}(\mathbf{g}_n)$, $n \in \mathbb{N}$. Αρκεί να δείξουμε ότι $\mathbf{f}_0 \in \mathbf{F}(\mathbf{g}_0)$.

Έστω μια στάσιμη στρατηγική $\mathbf{f}^* \in \mathbf{F}$ του παίχτη I τέτοια ώστε

$$\mathbf{v}_\alpha(\mathbf{f}^*,\mathbf{g}_0) = \max_{\mathbf{u}\in\mathbf{F}} \mathbf{v}_\alpha(\mathbf{u},\mathbf{g}_0)$$

Από τον ορισμό των $(\mathbf{f}_n,\mathbf{g}_n)$ προκύπτει

$$\mathbf{v}_\alpha(\mathbf{f}^*,\mathbf{g}_n) \le \mathbf{v}_\alpha(\mathbf{f}_n,\mathbf{g}_n)$$

και από τη συνέχεια της $\mathbf{v}_\alpha$ πάνω στο $\mathbf{F}\times\mathbf{G}$ έχουμε

$$\max_{\mathbf{u}\in\mathbf{F}} \mathbf{v}_\alpha(\mathbf{u},\mathbf{g}_0) \le \mathbf{v}_\alpha(\mathbf{f}_0,\mathbf{g}_0)$$

άρα

$$\mathbf{v}_\alpha(\mathbf{f}_0,\mathbf{g}_0) = \max_{\mathbf{u}\in\mathbf{F}} \mathbf{v}_\alpha(\mathbf{u},\mathbf{g}_0)$$

δηλαδή $\mathbf{f}_0 \in \mathbf{F}(\mathbf{g}_0)$. Συνεπώς η ανταπόκριση $\mathbf{F}(\mathbf{g})$ είναι άνω ημισυνεχής.

Παρόμοια αποδεικνύεται ότι αν μια ακολουθία $(\mathbf{f}_n,\mathbf{g}_n) \to (\mathbf{f}_0,\mathbf{g}_0)$ με $\mathbf{g}_n \in \mathbf{G}(\mathbf{f}_n)$ τότε $\mathbf{g}_0 \in \mathbf{G}(\mathbf{f}_0)$ άρα η $\mathbf{G}(\mathbf{f})$ άνω ημισυνεχής.

Έστω $(\mathbf{f}_n,\mathbf{g}_n) \to (\mathbf{f}_0,\mathbf{g}_0)$ με $\mathbf{f}_n \in \mathbf{F}(\mathbf{g}_n)$ και $\mathbf{g}_n \in \mathbf{G}(\mathbf{f}_n)$. Από τα προηγούμενα $\mathbf{f}_0 \in \mathbf{F}(\mathbf{g}_0)$ και $\mathbf{g}_0 \in \mathbf{G}(\mathbf{f}_0)$. Άρα το γράφημα της ανταπόκρισης $U(\mathbf{f},\mathbf{g})$ είναι κλειστό και επομένως αυτή είναι άνω ημισυνεχής.

Αφού οι χώροι των στάσιμων στρατηγικών είναι συμπαγή και κυρτά σύνολα, όπως έχουμε δείξει νωρίτερα, εφαρμόζουμε το θεώρημα σταθερού σημείου του Kakutani για να συμπεράνουμε την ύπαρξη ενός ζευγαριού στρατηγικών $(\mathbf{f}^*, \mathbf{g}^*)$ με $\mathbf{f}^* \in \mathbf{F}(\mathbf{g}^*)$ και $\mathbf{g}^* \in \mathbf{G}(\mathbf{f}^*)$. Συνεπώς, έχουμε

$$\mathbf{v}_\alpha(\mathbf{f}, \mathbf{g}^*) \leq \mathbf{v}_\alpha(\mathbf{f}^*, \mathbf{g}^*) \leq \mathbf{v}_\alpha(\mathbf{f}^*, \mathbf{g}). \qquad (4)$$

για οποιεσδήποτε στάσιμες στρατηγικές $\mathbf{f}$ και $\mathbf{g}$ των δύο παιχτών αντίστοιχα. Έτσι αποδεικνύεται η ύπαρξη του διανύσματος τιμής του παιχνιδιού όταν οι παίχτες περιορίζονται σε στάσιμες στρατηγικές και η τιμή ισούται με $\mathbf{v}_\alpha(\mathbf{f}^*, \mathbf{g}^*)$. Επικαλούμενοι το θεώρημα 1.5.4 συμπεραίνουμε ότι η παραπάνω τιμή είναι και η τιμή του αρχικού παιχνιδιού (δηλαδή όταν οι παίχτες δεν περιορίζονται σε στάσιμες στρατηγικές).

Αφού $\mathbf{v}_\alpha(\mathbf{f}^*, \mathbf{g}^*) = \upsilon_\alpha \mathbf{1}_N$ για κάποια $\upsilon_\alpha \in \mathbb{R}$, από τα *1* και *4* του θεωρήματος 4.1.1 προκύπτει ότι η τιμή είναι ανεξάρτητη από την αρχική κατάσταση.

2. Προκύπτει άμεσα από την ισότητα (4) ότι οι στάσιμες στρατηγικές $\mathbf{f}^*$ και $\mathbf{g}^*$ είναι βέλτιστες και από το θεώρημα 1.5.4 προκύπτει ότι είναι και βέλτιστες στο αρχικό παιχνίδι.

3. Αυτή η ιδιότητα προκύπτει από το θεώρημα 3.3.2.

$\boxtimes$

Θεώρημα 4.1.3. *Έστω $\mathbf{f}_\beta$ και $\mathbf{g}_\beta$ $\beta-$αποπληθωρισμένες βέλτιστες στάσιμες στρατηγικές για ένα αδιαχώριστο στοχαστικό παιχνίδι. Τότε, για $\varepsilon > 0$, οι στρατηγικές $\mathbf{f}_\beta$ και $\mathbf{g}_\beta$ είναι $\varepsilon-$βέλτιστες στην εκδοχή του παιχνιδιού όπου η ολική πληρωμή αποτιμάται σύμφωνα με τον οριακό μέσο όρο για όλα τα β αρκετά κοντά στο 1. Επίσης, εάν $\mathbf{f}_1$ και $\mathbf{g}_1$ είναι τέτοιες ώστε $\mathbf{f}_1 = \lim_{\beta\to 1} \mathbf{f}_\beta$ και $\mathbf{g}_1 = \lim_{\beta\to 1} \mathbf{g}_\beta$ για μια ακολουθία από β που τείνει στο 1, τότε οι στρατηγικές $\mathbf{f}_1$ και $\mathbf{g}_1$ είναι βέλτιστες ως προς το κριτήριο της οριακής μέσης πληρωμής.*

Απόδειξη

Από το θεώρημα 4.1.2, η οριακή αποπληθωρισμένη εξίσωση γίνεται

$$\upsilon_\alpha + (1-\beta)c_M(s) + \varepsilon(\beta, s) =$$
$$val\left[(1-\beta)r(s, a^1, a^2) + \beta \sum_{s'=1}^{N} p_{ss'}(a^1, a^2)(\upsilon_\alpha + (1-\beta)c_M(s') + \varepsilon(\beta, s'))\right]$$

όπου $\lim_{\beta\to 1} \frac{\varepsilon(\beta,s)}{1-\beta} = 0$.

Παρατηρούμε ότι οι παραπάνω εξισώσεις είναι βεβαίως οι εξισώσεις βελτιστότητας του Shapley. Επομένως, αν $\mathbf{f}_\beta$ είναι μια στάσιμη βέλτιστη στρατηγική του παίχτη I για το $\beta-$αποπληθωρισμένο παιχνίδι, θα ισχύει

$$\upsilon_\alpha \mathbf{1}_N + (1-\beta)\mathbf{c}_M + \boldsymbol{\varepsilon}(\beta) \leq (1-\beta)\mathbf{r}(\mathbf{f}_\beta, \mathbf{g}) + \beta P(\mathbf{f}_\beta, \mathbf{g})(\upsilon_\alpha \mathbf{1}_N + (1-\beta)\mathbf{c}_M + \boldsymbol{\varepsilon}(\beta))$$

για κάθε στάσιμη στρατηγική $\mathbf{g}$ του παίχτη II, και όπου $\lim_{\beta\to 1} \frac{1}{1-\beta}\boldsymbol{\varepsilon}(\beta) = 0$. Από εδώ προκύπτει ότι

$$(1-\beta)\upsilon_\alpha \mathbf{1}_N + (1-\beta)\mathbf{c}_M + \boldsymbol{\varepsilon}(\beta) \leq$$

$$(1-\beta)\mathbf{r}(\mathbf{f}_\beta, \mathbf{g}) + \beta(1-\beta)P(\mathbf{f}_\beta, \mathbf{g})\mathbf{c}_M + \beta P(\mathbf{f}_\beta, \mathbf{g})\boldsymbol{\varepsilon}(\beta)$$

και επομένως

$$\upsilon_\alpha \mathbf{1}_N + \mathbf{c}_M \leq \mathbf{r}(\mathbf{f}_\beta, \mathbf{g}) + \beta P(\mathbf{f}_\beta, \mathbf{g})\mathbf{c}_M + \tilde{\boldsymbol{\varepsilon}}(\beta)$$

όπου $\tilde{\boldsymbol{\varepsilon}}(\beta) = (\beta P(\mathbf{f}_\beta, \mathbf{g}) - I)\frac{\boldsymbol{\varepsilon}(\beta)}{1-\beta}$ και προφανώς $\lim_{\beta\to 1} \tilde{\boldsymbol{\varepsilon}}(\beta) = 0$. Πολλαπλασιάζοντας από αριστερά με τον πίνακα $Q(\mathbf{f}_\beta, \mathbf{g})$ παρατηρούμε τα εξής

(a) $Q\upsilon_\alpha \mathbf{1}_N = \upsilon_\alpha \mathbf{1}_N$,

(b) $QP = Q$,

(c) $Q\tilde{\boldsymbol{\varepsilon}}(\beta) = (\mathbf{q}\tilde{\boldsymbol{\varepsilon}}(\beta), \ldots, \mathbf{q}\tilde{\boldsymbol{\varepsilon}}(\beta))^T$ και $|\mathbf{q}\tilde{\boldsymbol{\varepsilon}}(\beta)| \leq ||\mathbf{q}|| \cdot ||\tilde{\boldsymbol{\varepsilon}}(\beta)|| \leq ||\tilde{\boldsymbol{\varepsilon}}(\beta)||$.

Επομένως καταλήγουμε ότι

$$\upsilon_\alpha \mathbf{1}_N + Q(\mathbf{f}_\beta, \mathbf{g}))\mathbf{c}_M \leq Q(\mathbf{f}_\beta, \mathbf{g})\mathbf{r}(\mathbf{f}_\beta, \mathbf{g}) + \beta Q(\mathbf{f}_\beta, \mathbf{g}))\mathbf{c}_M + ||\tilde{\boldsymbol{\varepsilon}}(\beta)||\mathbf{1}_N \qquad (7)$$

το οποίο αποδεικνύει ότι η στρατηγική $\mathbf{f}_\beta$ είναι $\varepsilon-$βέλτιστη για β αρκετά κοντά στο 1 ως προς το μέσο κριτήριο πληρωμής.

Ανάλογα, αποδεικνύεται το συμμετρικό αποτέλεσμα για την $\varepsilon-$βελτιστότητα της $\mathbf{g}_\beta$ στο παιχνίδι με οριακή μέση πληρωμή.

Ο δεύτερος ισχυρισμός του θεωρήματος προκύπτει λόγω συνέχειας (λήμμα 4.1.1) με χρήση της ανισότητας (7)

$$\mathbf{v}_\alpha(\mathbf{f}_1, \mathbf{g}) = \lim_{\beta\to 1} \mathbf{v}_\alpha(\mathbf{f}_\beta, \mathbf{g}) \geq \lim_{\beta\to 1} \left(\upsilon_\alpha - ||\tilde{\boldsymbol{\varepsilon}}(\beta)||\right)\mathbf{1}_N = \upsilon_\alpha \mathbf{1}_N.$$

$\boxtimes$

Στη συνέχεια δίνουμε έναν αλγόριθμο για τον υπολογισμό της τιμής και των βέλτιστων στάσιμων στρατηγικών των δύο παιχτών ενός αδιαχώριστου στοχατικού παιχνιδιού σύμφωνα με το κριτήριο της οριακής μέσης πληρωμής.

Αλγόριθμος

Βήμα 1
Έστω $\mathbf{f}_0$ αυθαίρετη στάσιμη στρατηγική του παίχτη I.

Βήμα 2
Έστω t φυσικός αριθμός,
έστω $\mathbf{f}_t$ ορισμένο επαγωγικά όπως στο βήμα 3,
έστω $\upsilon_t \mathbf{1}_N$ η τιμή μέσης οριακής πληρωμής της $MDP(\mathbf{f}_t)$,
έστω $\mathbf{w}_t \in \mathbb{R}^N$ τέτοιο ώστε για κάθε $s \in S$ να ισχύει

$$\upsilon_t + w_t(s) = \min_{a^2} \left[r(s, \mathbf{f}_t, a^2) + \sum_{s'=1}^{N} p_{ss'}(\mathbf{f}_t, a^2) w_t(s') \right] \qquad (8)$$

Βήμα 3
Για κάθε $s \in S$, έστω $f_{t+1}(s)$ η βέλτιστη απόφαση του παίχτη I στο πινακοπαιχνίδι

$$\left[r(s, a^1, a^2) + \sum_{s'=1}^{N} p_{ss'}(a^1, a^2) w_t(s') \right] \qquad (9)$$

Θέτουμε $\mathbf{f}_{t+1} = (f_{t+1}(1), f_{t+1}(2), \ldots, f_{t+1}(N))$.

Βήμα 4
Αν

$$val \left[r(s, a^1, a^2) + \sum_{s'=1}^{N} p_{ss'}(a^1, a^2) w_t(s') \right] = \upsilon_t + w_t(s)$$

για κάθε $s \in S$ τότε σταματάμε, αλλιώς θέτουμε $t = t + 1$ και επιστρέφουμε στο Βήμα 2.

Θεώρημα 4.1.4. *Στον παραπάνω αλγόριθμο για ένα αδιαχώριστο στοχαστικό παιχνίδι.*

1. $\upsilon_{t+1} \geq \upsilon_t$ *για κάθε* $t \in \mathbb{N}$ *και η στρατηγική* $\mathbf{f}_t$ *εγγυάται μέση πληρωμή* $\upsilon_t \mathbf{1}_N$ *στον παίχτη* I.

2. $\lim_{t \to \infty} \upsilon_t = \upsilon_\alpha$.

3. *Αν $v_{t+1} = v_t$, το οποίο συμβαίνει όταν ο αλγόριθμος σταματά, τότε $v_t = v_\alpha$ και η $\mathbf{f}_t$, όπως και η $\mathbf{f}_{t+1}$ είναι μία βέλτιστη στάσιμη στρατηγική μέσης πληρωμής για τον παίχτη I. Μία βέλτιστη στάσιμη στρατηγική για τον παίχτη II μπορεί να προκύψει από τις βέλτιστες στρατηγικές του παίχτη II στα πινακοπαιχνίδια:*

$$\left[r(s, a^1, a^2) + \sum_{s'=1}^{N} p_{ss'}(a^1, a^2) w_t(s') \right]$$

για $s = 1, 2, \ldots, N$.

Απόδειξη

1. Η v_t είναι η ελάχιστη πληρωμή που θα δώσει ο παίχτης II στον παίχτη I, αν ο II απαντήσει βέλτιστα στην $\mathbf{f}_t$. Άρα ο I χρησιμοποιώντας τη στάσιμη στρατηγική του $\mathbf{f}_t$ σίγουρα θα εξασφαλίσει πληρωμή v_t (ανεξάρτητα της αρχικής κατάστασης s_0).

Έστω $\tilde{v}_t(s)$ η τιμή του πινακοπαιχνιδιού (9). Λόγω της (8) η στρατηγική $\mathbf{f}_t$ εξασφαλίζει στον I $v_t + w_t(s)$ στο πινακοπαιχνίδι (9). Άρα για κάθε $s \in S$

$$\tilde{v}_t(s) \geq v_t + w_t(s) \qquad (10)$$

Αν ισχύει $\tilde{v}_t(s) = v_t + w_t(s)$ για κάθε $s \in S$, αφού η στρατηγική $\mathbf{f}_{t+1}$ είναι βέλτιστη στο πινακοπαιχνίδι (9), θα ισχύει

$$\min_{a^2} \left[r(s, \mathbf{f}_{t+1}, a^2) + \sum_{s'=1}^{N} p_{ss'}(\mathbf{f}_{t+1}, a^2) w_t(s') \right] = v_t + w_t(s) \qquad (11)$$

Από το θεώρημα 1.3.2, η ισχύς της (11) συνεπάγεται ότι η τιμή της $MDP(\mathbf{f}_{t+1})$ είναι η v_t, δηλαδή $v_{t+1} = v_t$, οπότε κατόπιν και $v_{t+2} = v_t$ κ.ο.κ.

Αν υπάρχει $s^* \in S$ τέτοιο ώστε $\tilde{v}_t(s^*) > v_t + w_t(s^*)$, δηλαδή αν στην ανισότητα (10) δεν ισχύει πάντα η ισότητα, τότε η (11) ισχύει με ανισότητα και επομένως $v_{t+1} \geq v_t$ (λόγω του πορίσματος 1.3.2, όπου θεωρούμε τη $MDP(\mathbf{f}_{t+1})$ και θέτουμε $v = v_{t+1}$ και $c = v_t$).

2. Η αύξουσα ακολουθία v_t οριακά φράσσεται άνω και κάτω από την v_α. Για την απόδειξη ο αναγνώστης παραπέμπεται στους Hoffman και Karp (1966).

3. Από το *1* έχουμε δείξει ότι ο αλγόριθμος σταματά όταν $v_{t+1} = v_t$ και

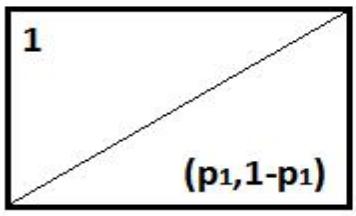
από το 2 έπεται άμεσα ότι $v_t = v_\alpha$. Επιπλέον, οι $\mathbf{f}_t$ και $\mathbf{f}_{t+1}$ είναι βέλτιστες στάσιμες στρατηγικές του I στα πινακοπαιχνίδια (9) και αφού ο αλγόριθμος σταματά θα είναι βέλτιστες στάσιμες στρατηγικές του I στο στοχαστικό παιχνίδι σύμφωνα με το κριτήριο της μέσης πληρωμής. Τέλος, μία βέλτιστη στάσιμη στρατηγική του II προκύπτει από τα πινακοπαιχνίδια (9) στην τελευταία επανάληψη του αλγορίθμου.

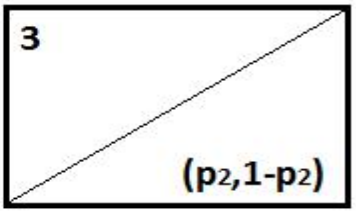
$\boxtimes$

4.2 Παραδείγματα στοχαστικού παιχνιδιού μέσης οριακής πληρωμής

Παράδειγμα 4.1.

Έστω ένα στοχαστικό παιχνίδι 2-παιχτών μηδενικού-αθροίσματος με δύο καταστάσεις, $S = 1, 2$, στο οποίο και οι δύο παίχτες έχουν μόνο μία διαθέσιμη απόφαση και στις δύο καταστάσεις.





Ο πίνακας πιθανοτήτων μετάβασης είναι

$$P = \begin{pmatrix} p_1 & 1-p_1 \\ p_2 & 1-p_2 \end{pmatrix}$$

Όταν $p_1 \in [0, 1)$ και $p_2 \in (0, 1]$ τότε το στοχαστικό παιχνίδι είναι αδιαχώριστο και επομένως γράφουμε τις εξισώσεις:

$$\mathbf{q}^T P = \mathbf{q}^T$$

και

$$\mathbf{q}^T \mathbf{1} = 1$$

δηλαδή έχουμε:

$$\begin{pmatrix} q(1) & , & q(2) \end{pmatrix} \begin{pmatrix} p_1 - 1 & 1-p_1 & 1 \\ p_2 & -p_2 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix}$$

$$\Rightarrow \begin{pmatrix} q(1) & , & q(2) \end{pmatrix} \begin{pmatrix} p_1 - 1 & 1 \\ p_2 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 1 \end{pmatrix}$$

Ο πίνακας

$$\begin{pmatrix} p_1 - 1 & 1 \\ p_2 & 1 \end{pmatrix}$$

είναι αντιστρέψιμος. Πράγματι, η ορίζουσά του είναι $p_1 - 1 - p_2$ που είναι 0 μόνο όταν $p_1 - 1 = p_2$ (αδύνατο).

Ως γνωστόν,

$$\begin{pmatrix} p_1-1 & 1 \\ p_2 & 1 \end{pmatrix}^{-1} = \frac{1}{p_1-p_2-1}\begin{pmatrix} 1 & -1 \\ -p_2 & p_1-1 \end{pmatrix}$$

οπότε

$$\begin{pmatrix} q(1) & , & q(2) \end{pmatrix} = \begin{pmatrix} 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & -1 \\ -p_2 & p_1-1 \end{pmatrix}\frac{1}{p_1-p_2-1}$$

$$\Rightarrow \begin{pmatrix} q(1) & , & q(2) \end{pmatrix} = \begin{pmatrix} \frac{-p_2}{p_1-p_2-1} & \frac{p_1-1}{p_1-p_2-1} \end{pmatrix}$$

Επιπλέον, όταν ο παίχτης I παίζει τη στάσιμη στρατηγική $\mathbf{f} = ((1),(1))$ και ο παίχτης II τη στάσιμη στρατηγική $\mathbf{g} = ((1),(1))$, τότε χρησιμοποιώντας το *4* του θεωρήματος 4.1.1 υπολογίζουμε την πληρωμή του παίχτη I για οποιαδήποτε αρχική κατάσταση $s_0 = 1, 2$ ως εξής:

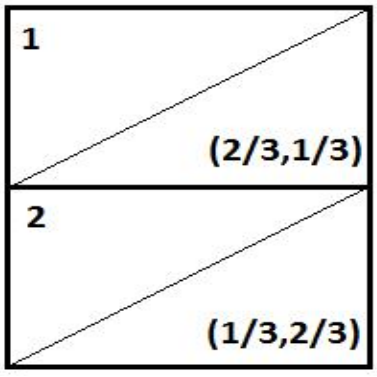
$$v_\alpha(1,\mathbf{f},\mathbf{g}) = v_\alpha(2,\mathbf{f},\mathbf{g}) = q(1)r(1,\mathbf{f},\mathbf{g}) + q(2)r(2,\mathbf{f},\mathbf{g})$$

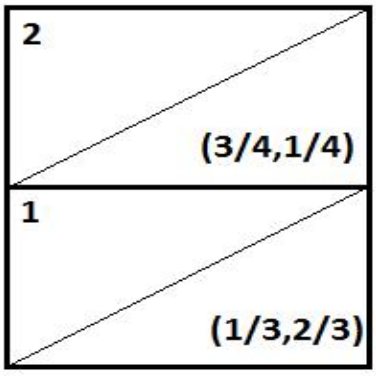
$$\Rightarrow v_\alpha(1,\mathbf{f},\mathbf{g}) = v_\alpha(2,\mathbf{f},\mathbf{g}) = \frac{-p_2}{p_1-p_2-1}1 + \frac{p_1-1}{p_1-p_2-1}3$$

⊠

Παράδειγμα 4.2.

Έστω ένα αδιαχώριστο στοχαστικό παιχνίδι 2-παιχτών μηδενικού αθροίσματος με χώρο καταστάσεων $S = \{1,2\}$, χώρο αποφάσεων $A^1(1) = A^1(2) = \{1,2\}$, $A^2(1) = A^2(2) = \{1\}$ και πίνακα πληρωμών και πιθανοτήτων μετάβασης





Εφαρμόζουμε τον αλγόριθμο που περιγράψαμε στο τέλος της προηγούμενης παραγράφου για την εύρεση της τιμής του παιχνιδιού και των βέλτιστων στάσιμων στρατηγικών των δύο παιχτών.

Βήμα 1
Αυθαίρετα παίρνουμε $\mathbf{f}_0 = ((1), (1))$.

Βήμα 2
Λύνουμε τη $MDP(\mathbf{f}_0)$ κάνοντας χρήση του θεωρήματος 1.3.2, δηλαδή βρίσκουμε ένα διάνυσμα $\mathbf{w}_0 \in \mathbb{R}^2$ τέτοιο ώστε για κάθε $s \in S$ να ισχύει

$$\upsilon_0 + w_0(s) = \min_{a^2} \left[r(s, \mathbf{f}_0(s), a^2) + \sum_{s'=1,2} p_{ss'}(\mathbf{f}_0(s), a^2) w_0(s') \right]$$

δηλαδή

$$\upsilon_0 + w_0(1) = 1 + \frac{2}{3} w_0(1) + \frac{1}{3} w_0(2)$$
$$\upsilon_0 + w_0(2) = 2 + \frac{3}{4} w_0(1) + \frac{1}{4} w_0(2)$$

Θέτοντας $w_0(1) = 0$ παίρνουμε $w_0(2) = \frac{12}{13}$ και $\upsilon_0 = \frac{17}{13}$

Βήμα 3
Αναζητούμε τη βέλτιστη στρατηγική $\mathbf{f}_1$ του παίχτη I στα πινακοπαιχνίδια

$$\left[r(s, a^1, a^2) + \sum_{s'=1,2} p_{ss'}(a^1, a^2) w_0(s') \right]$$

όπου η ποσότητες μέσα στις αγκύλες είναι οι $m^1(s) \times m^2(s)$ πίνακες με γραμμές $a^1 = 1, 2, \ldots, m^1(s)$ και στήλες $a^2 = 1, 2, \ldots, m^2(s)$ για $s = 1, 2$. Αντικαθιστώντας προκύπτει

$$\begin{bmatrix} \frac{17}{13} \\ \frac{34}{13} \end{bmatrix}$$

και

$$\begin{bmatrix} \frac{29}{13} \\ \frac{21}{13} \end{bmatrix}$$

για $s = 1, 2$ αντίστοιχα. Άρα η ζητούμενη στρατηγική του I είναι η $\mathbf{f}_1 = ((2), (1))$.

Βήμα 4
Για $s = 1$

$$val \begin{bmatrix} \frac{17}{13} \\ \frac{34}{13} \end{bmatrix} = \frac{34}{13} > \upsilon_0 + w_0(1) = \frac{17}{13}$$

Για $s = 2$

$$val \begin{bmatrix} \frac{29}{13} \\ \frac{21}{13} \end{bmatrix} = \frac{29}{13} = \upsilon_0 + w_0(2) = \frac{29}{13}$$

Άρα ξαναπηγαίνουμε στο βήμα 2.

Βήμα 2
Λύνουμε τη $MDP(\mathbf{f}_1)$ χρησιμοποιώντας το θεώρημα 1.3.2 οπότε παρόμοια με πριν (βήμα 2) προκύπτουν

$$\upsilon_1 + w_1(1) = 2 + \frac{1}{3}w_1(1) + \frac{2}{3}w_1(2)$$

$$\upsilon_1 + w_1(2) = 2 + \frac{3}{4}w_1(1) + \frac{1}{4}w_1(2)$$

Θέτοντας $w_1(1) = 0$ παίρνουμε $w_1(2) = 0$ και $\upsilon_1 = 2$.

Βήμα 3
Παρόμοια με πριν (βήμα 3) έχουμε

$$\begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

και

$$\begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

για $s = 1, 2$ αντίστοιχα. Άρα έχουμε $\mathbf{f}_3 = ((2), (1)) = \mathbf{f}_2$.

Βήμα 4
Για $s = 1$

$$val \begin{bmatrix} 1 \\ 2 \end{bmatrix} = 2 = \upsilon_1 + w_1(1) = 2$$

Για $s = 2$

$$val \begin{bmatrix} 2 \\ 1 \end{bmatrix} = 2 = \upsilon_1 + w_1(2) = 2$$

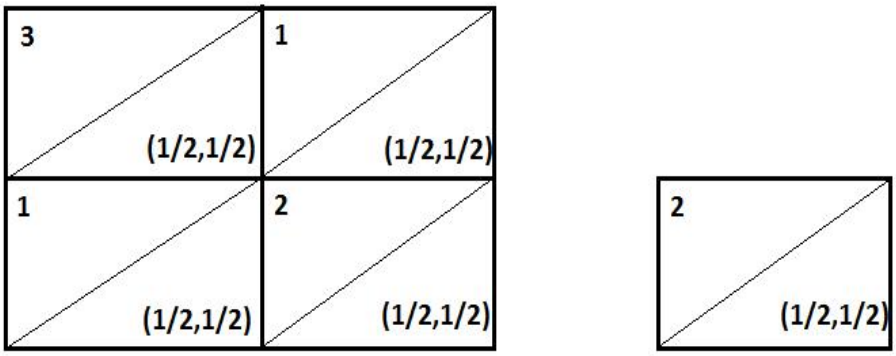
οπότε σταματάμε. Επομένως $\upsilon_\alpha = \upsilon_1 = 2$ για οποιαδήποτε αρχική κατάσταση και οι βέλτιστες στάσιμες στρατηγικές των δύο παιχτών είναι $\mathbf{f}^* = ((2), (1))$ και $\mathbf{g}^* = ((1), (1))$.

$\boxtimes$

Τέλος, το παράδειγμα που ακολουθεί είναι το παράδειγμα 3.2 που έχουμε ήδη λύσει βρίσκοντας την αποπληθωρισμένη τιμή του. Εδώ θα υπολογίσουμε την τιμή του για τη μέση οριακή πληρωμή με δύο τρόπους.

Παράδειγμα 4.3.

Έστω το στοχαστικό παιχνίδι 2-παιχτών μηδενικού αθροίσματος που ορίσαμε στο παράδειγμα 3.2 δηλαδή το παιχνίδι με χώρο καταστάσεων $S = \{1, 2\}$, χώρο αποφάσεων $A^1(1) = A^2(1) = \{1, 2\}, A^1(2) = A^2(2) = \{1\}$ και πίνακα πληρωμών και πιθανοτήτων μετάβασης



Το συγκεκριμένο στοχαστικό παιχνίδι είναι αδιαχώριστο αφού και οι δύο καταστάσεις του παιχνιδιού επικοινωνούν για οποιεσδήποτε στάσιμες στρατηγικές των δύο παιχτών.

Ένας τρόπος υπολογισμού της τιμής του παιχνιδιού χρησιμοποιώντας το κριτήριο της μέσης ολικής πληρωμής είναι

$$\upsilon_\alpha(1) = \lim_{\beta \to 1} \upsilon_\beta(1) = \lim_{\beta \to 1} \frac{12 - \beta}{6} = \frac{11}{6}$$

$$\upsilon_\alpha(2) = \lim_{\beta \to 1} \upsilon_\beta(2) = \lim_{\beta \to 1} \frac{10 + \beta}{6} = \frac{11}{6}$$

Παρατηρούμε ότι $\upsilon_\alpha(1) = \upsilon_\alpha(2)$ (όπως άλλωστε υπογραμμίζει το *4* του θεωρήματος 4.1.1).

Ένας δεύτερος τρόπος για να υπολογίσουμε την τιμή του παιχνιδιού χρησιμοποιώντας το κριτήριο της μέσης ολικής πληρωμής είναι να εφαρμόσουμε τον αλγόριθμο που περιγράψαμε στο τέλος της προηγούμενης παραγράφου.

Βήμα 1
Αυθαίρετα παίρνουμε $\mathbf{f}_0 = ((1),(1))$.

Βήμα 2
Λύνουμε τη $MDP(\mathbf{f}_0)$ χρησιμοποιώντας το θεώρημα 1.3.2, δηλαδή θα βρούμε ένα διάνυσμα $\mathbf{w}_0 \in \mathbb{R}^2$ τέτοιο ώστε για κάθε $s \in S$ να ισχύει

$$\upsilon_0 + w_0(1) = \min\left\{3 + \frac{1}{2}w_0(1) + \frac{1}{2}w_0(2), 1 + \frac{1}{2}w_0(1) + \frac{1}{2}w_0(2)\right\}$$

$$\upsilon_0 + w_0(2) = 2 + \frac{1}{2}w_0(1) + \frac{1}{2}w_0(2)$$

Θέτοντας $w_0(1) = 0$ παίρνουμε $w_0(2) = 1$ και $\upsilon_0 = \frac{3}{2}$

Βήμα 3
Αναζητούμε τη βέλτιστη στρατηγική $\mathbf{f}_1$ του παίχτη I στα πινακοπαιχνίδια

$$val\begin{bmatrix} 3+\frac{1}{2} & 1+\frac{1}{2} \\ 1+\frac{1}{2} & 2+\frac{1}{2} \end{bmatrix}$$

και

$$val\left[2+\tfrac{1}{2}\right]$$

Στο πρώτο λύνοντας με εξισωτικές στρατηγικές βρίσκουμε τιμή $\frac{13}{6}$ και στο δεύτερο βρίσκουμε άμεσα τιμή $\frac{5}{2}$. Άρα η ζητούμενη στρατηγική του I είναι η $\mathbf{f}_1 = ((\frac{1}{3},\frac{2}{3}),(1))$.

Βήμα 4
Για $s = 1$

$$\frac{13}{6} > \upsilon_0 + w_0(1) = \frac{3}{2}$$

Για $s = 2$

$$\frac{5}{2} = \upsilon_0 + w_0(2) = \frac{5}{2}$$

Άρα ξαναπηγαίνουμε στο βήμα 2.

Βήμα 2
Λύνουμε τη $MDP(\mathbf{f}_1)$ χρησιμοποιώντας το θεώρημα 1.3.2, δηλαδή θα βρούμε ένα διάνυσμα $\mathbf{w}_1 \in \mathbb{R}^2$ τέτοιο ώστε για κάθε $s \in S$ να ισχύει

$$\upsilon_1 + w_1(1) = \min\{\ \frac{1}{3}(3 + \frac{1}{2}w_1(1) + \frac{1}{2}w_1(2)) + \frac{2}{3}(1 + \frac{1}{2}w_1(1) + \frac{1}{2}w_1(2)),$$

$$\tfrac{1}{3}\left(1+\tfrac{1}{2}w_1(1)+\tfrac{1}{2}w_1(2)\right)+\tfrac{2}{3}\left(2+\tfrac{1}{2}w_1(1)+\tfrac{1}{2}w_1(2)\right)\}$$

$$v_1+w_1(2)=2+\frac{1}{2}w_1(1)+\frac{1}{2}w_1(2)$$

Θέτοντας $w_1(1)=0$ παίρνουμε $w_1(2)=\frac{1}{3}$ και $v_1=\frac{11}{6}$.

Βήμα 3
Αναζητούμε τη βέλτιστη στρατηγική $\mathbf{f}_1$ του παίχτη I στα πινακοπαιχνίδια

$$val\begin{bmatrix}3+\frac{1}{2}\frac{1}{3} & 1+\frac{1}{2}\frac{1}{3}\\ 1+\frac{1}{2}\frac{1}{3} & 2+\frac{1}{2}\frac{1}{3}\end{bmatrix}$$

και

$$val\left[2+\tfrac{1}{2}\tfrac{1}{3}\right]$$

Στο πρώτο λύνοντας με εξισωτικές στρατηγικές βρίσκουμε τιμή $\frac{11}{6}$ και στο δεύτερο βρίσκουμε άμεσα τιμή $\frac{13}{6}$. Άρα η ζητούμενη στρατηγική του I είναι η $\mathbf{f}_2=((\frac{1}{3},\frac{2}{3}),(1))=\mathbf{f}_1$.

Βήμα 4
Για $s=1$

$$\frac{11}{6}=v_1+w_1(1)=\frac{11}{6}$$

Για $s=2$

$$\frac{13}{6}=v_1+w_1(2)=\frac{13}{6}$$

οπότε σταματάμε. Επομένως $v_\alpha=v_1=\frac{11}{6}$ για οποιαδήποτε αρχική κατάσταση και οι βέλτιστες στάσιμες στρατηγικές των δύο παιχτών είναι $\mathbf{f}^*=((\frac{1}{3},\frac{2}{3}),(1))$ και $\mathbf{g}^*=((\frac{1}{3},\frac{2}{3}),(1))$. Η βέλτιστη στάσιμη στρατηγική $\mathbf{g}^*$ προκύπτει με χρήση εξισωτικών στρατηγικών στα πινακοπαιχνίδια της τελευταίας επανάληψης του βήματος 3 του αλγορίθμου.

$\boxtimes$

Βιβλιογραφία

[1] Aumann, R.J. (1964), Mixed and Behavior Strategies in Infinite Extensive Games, Advances in Game Theory, Annals of Mathematics Studies 52, 627-650, Princenton University Press.

[2] Bewley, T. and Kohlberg, E. (1976), The Asymptotic Solution of a Recursive Equation Arising in Stochastic Games, Mathematics of Operations Research, 3:104-125.

[3] Filar, J. and Vrieze, K. (1997), Competitive Markov Decision Processes, Springer.

[4] Kolmogorv, A. (1933), Grundbegriffe der Wahrscheinlichkeitsrechnung, Ergebnisse der Mathematik 2, no. 3, Springer Verlag, Berlin.

[5] Blackwell, D. (1965), Discounted Dynamic Programming, Ann. of Math. Stat. 36, 226-235.

[6] Derman, C. (1970), Finite State Markovian Decision Processes, Academic Press, New York.

[7] Hoffman, A. J. and Karp, R. M. (1966), On Nonterminating Stochastic Games, Management Science, Vol. 12, No. 5, Series A, Sciences(Jan), pp. 359-370, Informs.

[8] Kuhn, H. W. (1953), Extensive Games and the Problem of Information, Contribution to the Theory of Games II, Annals of Mathematics Studies 28, 193-216, Princeton University Press (Επανέκδοση: Kuhn, H. W. (ed.) (1997)).

[9] Maitra, A. and Sudderth, W. (1996), Discrete Gambling and Stochastic Games, Applications of Mathematics, Stochastic Modelling and Applied Probability 32, Springer.

[10] Monash, C. A. (1979), Stochastic Games: The Minimax Theorem, Thesis submitted to the dep. of Math. of the Harvard University, Cambridge, Massachusetts.

[11] Neyman, A. and Sorin, S. (eds.) (2003), Stochastic Games and Applications, Series C: Mathematical and Physical Sciences, Vol. 570, Kluwer Academic Publishers.

[12] von Neumann, J. and Morgestern, O. (1944), Theory of Games and Economic Behavior, Wiley.

[13] Vrieze, O.J. (1983), Stochastic Games with Finite State and Action Spaces. CWI Tracts 33, Amsterdam.

[14] Ross, S. (1983), Introduction to Stochastic Dynamic Programming, Academic Press, University of California, Berkeley.

[15] Μηλολιδάκης, Κ. (2009), Θεωρία Παιγνίων, Μαθηματικά Μοντέλα Σύγκρουσης και Συνεργασίας, Εκδόσεις Σοφία.