ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΘΕΣΣΑΛΟΝΙΚΗΣ

ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

«ΣΤΑΤΙΣΤΙΚΗ ΚΑΙ ΜΟΝΤΕΛΟΠΟΙΗΣΗ»

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Επίλυση στοχαστικών παιγνίων με μεθόδους γραμμικού προγραμματισμού

Κωνσταντίνος Ζυγογιάννης

ΕΠΙΒΛΕΠΟΥΣΑ: Αλεξάνδρα Παπαδοπούλου

Αναπ. Καθηγ. Α.Π.Θ.

Θεσσαλονίκη, Δεκέμβριος 2022

ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΘΕΣΣΑΛΟΝΙΚΗΣ

ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

«ΣΤΑΤΙΣΤΙΚΗ ΚΑΙ ΜΟΝΤΕΛΟΠΟΙΗΣΗ»

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Επίλυση στοχαστικών παιγνίων με μεθόδους γραμμικού προγραμματισμού

Κωνσταντίνος Ζυγογιάννης

ΕΠΙΒΛΕΠΟΥΣΑ: Αλεξάνδρα Παπαδοπούλου

Αναπ. Καθηγ. Α.Π.Θ.

Εγκρίθηκε από την Τριμελή Εξεταστική Επιτροπή:

……	……	…….
Γ.Κωνσταντινίδης	Α.Παπαδοπούλου	Χ. Πελέκης
Διδάκτορας Α.Π.Θ.	Αναπ. Καθηγήτρια Α.Π.Θ.	Επίκ. Καθηγητής ΑΠΘ

Θεσσαλονίκη, Δεκέμβριος 2022

…….

Κωνσταντίνος Ζυγογιάννης

Πτυχιούχος Μαθηματικός Α.Π.Θ.

ΠΡΟΛΟΓΟΣ

Η παρούσα διπλωματική εργασία εκπονήθηκε στο πλαίσιο του Μεταπτυχιακού Προγράμματος Σπουδών «Στατιστική και Μοντελοποίηση» του τμήματος Μαθηματικών του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης. Θα ήθελα να ευχαριστήσω θερμά την κυρία Αλεξάνδρα Παπαδοπούλου για τις πολύ χρήσιμες συμβουλές, την άψογη συνεργασία και καθοδήγησή της καθ' όλη τη διάρκεια προετοιμασίας και υλοποίησης της εργασίας αυτής. Ακόμη, θα ήθελα να ευχαριστήσω την οικογένεια και τους φίλους μου, για την αδιάκοπη και ανιδιοτελή στήριξή τους όλα αυτά τα χρόνια.

ΠΕΡΙΛΗΨΗ

Η παρούσα εργασία περιέχει κατ' αρχήν μια ανασκόπηση με κεντρικό αντικείμενο τα Στοχαστικά Παίγνια. Το πρώτο κεφάλαιο αποτελεί μια εισαγωγή στις Μαρκοβιανές διαδικασίες απόφασης , ένα πολύ ισχυρό εργαλείο για την μελέτη των πιο «πολύπλοκων» στοχαστικών παιγνίων, τα οποία αποτελούν αντικείμενο μελέτης του δεύτερου κεφαλαίου. Σημαντικό μέρος της θεωρίας που αναπτύσσεται στα δύο πρώτα κεφάλαια αντλήθηκε από το βιβλίο των Jerzy Filar και Koos Vrieze με τίτλο Competitive Markov Decision Processes (1997). Το τρίτο κεφάλαιο βασίστηκε στην εργασία των S. Sinha και K. G. Bakshi με τίτλο Zero-Sum Two Person Perfect Information Semi-Markov Games: A Reduction (2022). Στην εργασία αναδεικνύεται ο ρόλος του Γραμμικού Προγραμματισμού, στην ανάλυση και επίλυση προβλημάτων που μοντελοποιούνται είτε ως διαδικασίες απόφασης είτε ως παίγνια. Όπως θα γίνει αντιληπτό, μελετώντας συγκεκριμένες κατηγορίες παιγνίων ως προς τη δομή τους, εν τέλει, ο υπολογισμός μιας βέλτιστης πολιτικής η οποία θα αποφέρει το μέγιστο αναμενόμενο κέρδος (ή την ελάχιστη ζημία) αποτελεί πρόβλημα που ισοδυναμεί με την επίλυση κατάλληλα κατασκευασμένων γραμμικών προγραμμάτων. Στο τρίτο κεφάλαιο γενικεύονται τα παίγνια που προαναφέρθηκαν, τα οποία πλέον θα καλούνται Ημιμαρκοβιανά Παίγνια. Πραγματοποιείται ανάλυση μιας συγκεκριμένης κλάσης αυτής της ιδιαίτερα ευρείας κατηγορίας παιγνίων, τα Ημιμαρκοβιανά Παίγνια Τέλειας Πληροφόρησης. Τέλος, γίνεται υλοποίηση του σχετικού αλγορίθμου επίλυσης τέτοιων παιγνίων με το λογισμικό MATLAB και παρατίθενται τα σχετικά αποτελέσματα.

ABSTRACT

The present paper initially includes a review with Stochastic Games as the main subject. The first chapter is an introduction to Markov Decision Processes, a significant tool in order to study the more complicated case of Stochastic Games, which are going to be the subject of the next chapter. A big part of the theory developed in the first two chapters is due to the book Competitive Markov Decision Processes (1997), written by Jerzy Filar και Koos Vrieze. The third chapter is based on a research paper by S. Sinha και K. G. Bakshi, titled as Zero-Sum Two Person Perfect Information Semi-Markov Games: A Reduction (2022). Through this paper, the role of Linear Programming stands out in analyzing and solving problems which can be modelled as either decision processes or stochastic games. When analyzing specific subclasses of games regarding their structure, it is concluded that finding the policy that maximizes the expected profit (or minimizes the expected damage) is a problem equivalent to solving suitably constructed linear programs. In the third chapter, semi-Markov games, a generalization of the games mentioned so far, are being introduced. The main subject of this chapter is a specific category of semi-Markov games, called Perfect Information semi-Markov games. Finally, a specific algorithm regarding the solution of such games is implemented through the MATLAB software and its results are briefly discussed.

Περιεχόμενα

Κεφάλαιο 1

Μαρκοβιανές Διαδικασίες Απόφασης

1.0 Εισαγωγή

Το 1ο κεφάλαιο της εργασίας αφορά στις Μαρκοβιανές Διαδικασίες Απόφασης και, συγκεκριμένα, πώς οι διαδικασίες αυτές μπορούν να μελετηθούν ως Στοχαστικά παίγνια, τα οποία αποτελούν αντικείμενο μελέτης του επόμενου κεφαλαίου, όπου υφίσταται μόνον ένας παίκτης-ελεγκτής. Τόσο στο παρόν κεφάλαιο όσο και γενικότερα στα πλαίσια της εργασίας αυτής, θα μας απασχολήσουν διαδικασίες απόφασης όπου ο χώρος καταστάσεων και ενεργειών του/των παίκτη/παικτών είναι πεπερασμένος. Γίνεται αναφορά σε συγκεκριμένες κατηγορίες διαδικασιών απόφασης, οι οποίες διαφέρουν ως προς περιορισμούς που πιθανόν να συνδέονται με τη δομή αυτών ή/και το κριτήριο πληρωμής του ελεγκτή στην εκάστοτε διαδικασία. Παράλληλα, θα γίνει αντιληπτός ο τρόπος που μπορούν να επιλυθούν τέτοιου τύπου μοντέλα απόφασης με μεθόδους που προκύπτουν από το αντικείμενο του Γραμμικού Προγραμματισμού.

1.1 Η διαδικασία με συντελεστή προεξόφλησης και η τερματιζόμενη διαδικασία

Θα συμβολίζουμε με S_t την τυχαία μεταβλητή που παίρνει τιμές από το σύνολο $S = \{1,2, \dots, N\}$, το οποίο καλούμε χώρο καταστάσεων. Όταν λέμε ότι η διαδικασία βρίσκεται στην κατάσταση $s \in S$ στον χρόνο t, θα εννοούμε ότι πραγματοποιείται το γεγονός $\{S_t = s\}$.

Η διαδικασία απόφασης ελέγχεται από έναν «decision-maker» (d-m), ο οποίος, δεδομένου ότι το σύστημα βρίσκεται στην κατάσταση $s \in S$, επιλέγει μια ενέργεια $a \in A(s) = \{1,2, \dots, m(s)\}$. Η επιλογή της $a \in A(s)$ είναι συνδεδεμένη με μία άμεσα παραγόμενη αμοιβή $r(s,a)$, που θα καλούμε αμοιβή ενός βήματος, την οποία λαμβάνει ο decision-maker και έπειτα η διαδικασία μεταπηδά σε μία νέα κατάσταση $s' \in S$ και εξελίσσεται με τον ίδιο τρόπο από την νέα κατάσταση s'.

Ορίζουμε:

$$p(s'|s,a) = \mathbb{P}\{S_{t+1} = s'|S_t = s, A_t = a\}$$

να είναι η πιθανότητα μετάβασης στην κατάσταση s', δεδομένου ότι βρισκόμαστε στην κατάσταση s και επιλέγεται η ενέργεια $a \in A(s)$. Όπως είναι φανερό, έχουμε ομοιογένεια ως προς το χρόνο, δηλαδή οι πιθανότητες μετάβασης είναι ανεξάρτητες του t.

Μία *στρατηγική* $\boldsymbol{f} = [\boldsymbol{f}(1), \boldsymbol{f}(2), \dots, \boldsymbol{f}(N)]$ είναι ένα διάνυσμα-γραμμή, όπου το s-οστό στοιχείο του είναι το μη αρνητικό διάνυσμα γραμμής:

$\boldsymbol{f}(s) = \big[f(s,1), f(s,2), \dots, f\big(s, m(s)\big)\big]$ για το οποίο ισχύει ότι :

$$\sum_{a=1}^{m(s)} f(s,a) = 1$$

$f(s,a) \doteqdot$ Η πιθανότητα να επιλεγεί η ενέργεια $a \in A(s)$, δεδομένου ότι η διαδικασία βρίσκεται στην κατάσταση s.

Μία στρατηγική $\boldsymbol{f}$ θα καλείται *καθαρή (ή ντετερμινιστική)* αν

$f(s,a) \in \{0,1\}$, για όλα τα $a \in A(s), \mathrm{s} \in \mathrm{S}$

Οπότε μία στρατηγική $\boldsymbol{f}$ ορίζει έναν $N \times N$ πίνακα πιθανοτήτων μετάβασης

$$\boldsymbol{P}(\boldsymbol{f}) = \big(p(s'|s,\boldsymbol{f})\big)_{s,s' \in S}$$

τα στοιχεία του οποίου ορίζονται ως :

$$p(s'|s,\boldsymbol{f}) = \sum_{a=1}^{m(s)} p(s'|s,a) f(s,a)$$

Φυσικά, για συγκεκριμένο $s \in S$, για όλα τα $a \in A(s)$, θα έχουμε ότι :

$$\sum_{s'=1}^{N} p(s'|s,a) = 1$$

Δηλαδή, ο πίνακας $P(\boldsymbol{f})$ είναι στοχαστικός, που σημαίνει ότι όλες του οι γραμμές αθροίζουν στη μονάδα, και συνεπώς, ορίζει μια μοναδική Μαρκοβιανή Αλυσίδα. Φυσικά, μια διαφορετική στρατηγική $\boldsymbol{f}'$ ορίζει διαφορετική Μαρκοβιανή Αλυσίδα μέσω του $P(\boldsymbol{f}')$.

1.1.1 Μαρκοβιανό μοντέλο απόφασης Γ_β με συντελεστή προεξόφλησης β

Έστω $\{R_t\}_{t=0}^{\infty}$ η ακολουθία των αμοιβών, όπου R_t θα είναι η αμοιβή για τον d-m που αντιστοιχεί στην χρονική περίοδο $[t, t+1)$. Εφόσον οριστεί η αρχική κατάσταση s και μία στρατηγική $\boldsymbol{f}$, τότε γνωρίζουμε την κατανομή πιθανότητας της R_t και η μέση τιμή αυτής είναι καλά ορισμένη. Θα συμβολίζουμε τη μέση τιμή με :

$$E_{s\boldsymbol{f}}[R_t] \doteqdot E_{\boldsymbol{f}}[R_t | S_0 = s]$$

Η συνολική αναμενόμενη τιμή με προεξόφληση για την $S_0 = s$ μέσω της $\boldsymbol{f}$ θα είναι :

$$v_\beta(s, \boldsymbol{f}) = \sum_{t=0}^{\infty} \beta^t E_{s\boldsymbol{f}}[R_t], \qquad \text{όπου } \beta \in [0,1) \text{ είναι ο συντελεστής προεξόφλησης} \quad (1.1.1)$$

Η φυσική ερμηνεία αυτού είναι πως: Η απόδοση μιας μονάδας στο χρόνο $t+1$ 'αξίζει' μόνο β φορές απ'ότι άξιζε στο χρόνο t ($\beta < 1$).

Στη συνέχεια, ορίζουμε το διάνυσμα αναμενόμενης αμοιβής ενός βήματος ως:

$$\boldsymbol{r}(\boldsymbol{f}) = [r(1, \boldsymbol{f}), r(2, \boldsymbol{f}), \dots, r(N, \boldsymbol{f})]^{\boldsymbol{T}}$$

όπου, για κάθε $s \in S$

$$r(s, \boldsymbol{f}) \doteqdot \sum_{a \in A(s)} r(s, a) f(s, a)$$

Για αυθαίρετο $s \in S$, υπολογίζουμε:

$$E_{s\boldsymbol{f}}[R_0] = r(s, \boldsymbol{f}) = [\boldsymbol{r}(\boldsymbol{f})]_s$$

$$E_{s\boldsymbol{f}}[R_1] = \sum_{s'=1}^{N} p(s'|s, \boldsymbol{f}) r(s', \boldsymbol{f}) = [P(\boldsymbol{f})\boldsymbol{r}(\boldsymbol{f})]_s$$

...

$$E_{s\boldsymbol{f}}[R_t] = \sum_{s'=1}^{N} p_t(s'|s, \boldsymbol{f}) r(s', \boldsymbol{f}) = [P^t(\boldsymbol{f})\boldsymbol{r}(\boldsymbol{f})]_s$$

όπου με $[\boldsymbol{a}]_s$ συμβολίζουμε το s-οστό στοιχείο του διανύσματος $\boldsymbol{a}$ και με $p_t(s'|s, \boldsymbol{f})$ την πιθανότητα μετάβασης t-βήματος από την κατάσταση s στην s' για την αλυσίδα που ορίζει η $\boldsymbol{f}$.

Από τη θεωρία γνωρίζουμε ότι η t-δύναμη του πίνακα $P(\boldsymbol{f})$ περιγράφει ακριβώς την πιθανότητα μετάβασης t-βήματος. Δηλαδή :

$$\boldsymbol{P}^t(\boldsymbol{f}) = \left(p_t(s'|s, \boldsymbol{f})\right)_{s,s'=1}^{N}$$

Αντικαθιστώντας στην (1.1.1) , για $\boldsymbol{v}_\beta(\boldsymbol{f}) = \left(v(1, \boldsymbol{f}), v(2, \boldsymbol{f}), \dots, v(N, \boldsymbol{f})\right)^T$ παίρνουμε ότι:

$$\boldsymbol{v}_\beta(\boldsymbol{f}) = \sum_{t=0}^{\infty} \beta^t P^t(\boldsymbol{f})\, \boldsymbol{r}(\boldsymbol{f}) \qquad (1.1.2)$$

όπου ορίζουμε $\boldsymbol{P}^0(\boldsymbol{f}) \doteqdot \boldsymbol{I}_N$ (με $\boldsymbol{I}_N$ συμβολίζουμε τον μοναδιαίο πίνακα διάστασης $N \times N$)

Ο πίνακας $[\boldsymbol{I} - \beta \boldsymbol{P}(\boldsymbol{f})]$ είναι αντιστρέψιμος και

$$[\boldsymbol{I}-\beta\boldsymbol{P}(\boldsymbol{f})]^{-1}=\boldsymbol{I}+\beta\boldsymbol{P}(\boldsymbol{f})+\beta^{\boldsymbol{2}}\boldsymbol{P}^{\boldsymbol{2}}(\boldsymbol{f})+\cdots$$

Απόδειξη

Αρχικά παρατηρούμε ότι ο πίνακας $\beta\boldsymbol{P}(\boldsymbol{f})$ είναι υποστοχαστικός, εφόσον ο $\boldsymbol{P}(\boldsymbol{f})$ είναι στοχαστικός και $0<\beta<1$. Οπότε θα ισχύει ότι :

$$\lim_{n\to\infty}(\beta\boldsymbol{P}(\boldsymbol{f}))^n=0$$

Ισχύει ότι: $\boldsymbol{I}^n-\big(\beta\boldsymbol{P}(\boldsymbol{f})\big)^n=\big(\boldsymbol{I}-\beta\boldsymbol{P}(\boldsymbol{f})\big)\Big(\boldsymbol{I}+\beta\boldsymbol{P}(\boldsymbol{f})+\big(\beta\boldsymbol{P}(\boldsymbol{f})\big)^2+\cdots+\big(\beta\boldsymbol{P}(\boldsymbol{f})\big)^{n-1}\Big)$

Παίρνοντας όρια στην παραπάνω σχέση, λαμβάνουμε:

$$\lim_{n\to\infty}\Big(\boldsymbol{I}^n-\big(\beta\boldsymbol{P}(\boldsymbol{f})\big)^n\Big)=\lim_{n\to\infty}\big(\boldsymbol{I}-\beta\boldsymbol{P}(\boldsymbol{f})\big)\Big(\boldsymbol{I}+\beta\boldsymbol{P}(\boldsymbol{f})+\big(\beta\boldsymbol{P}(\boldsymbol{f})\big)^2+\cdots+\big(\beta\boldsymbol{P}(\boldsymbol{f})\big)^{n-1}\Big)$$

$$\Rightarrow\boldsymbol{I}=\big(\boldsymbol{I}-\beta\boldsymbol{P}(\boldsymbol{f})\big)\sum_{n=0}^{\infty}\big(\beta\boldsymbol{P}(\boldsymbol{f})\big)^n \qquad (1.1.3)$$

Παίρνοντας ορίζουσες στην (1.1.3) , έχουμε ότι :

$$1=\det\big(\boldsymbol{I}-\beta\boldsymbol{P}(\boldsymbol{f})\big)\det\left(\sum_{n=0}^{\infty}\big(\beta\boldsymbol{P}(\boldsymbol{f})\big)^n\right)$$

Από την τελευταία σχέση συμπεραίνουμε πως $\det\big(\boldsymbol{I}-\beta\boldsymbol{P}(\boldsymbol{f})\big)\neq 0$, δηλαδή ορίζεται ο πίνακας $\big(\boldsymbol{I}-\beta\boldsymbol{P}(\boldsymbol{f})\big)^{-1}$.

Οπότε, από την (1.1.3), πολλαπλασιάζοντας επί $\big(\boldsymbol{I}-\beta\boldsymbol{P}(\boldsymbol{f})\big)^{-1}$:

$$\big(\boldsymbol{I}-\beta\boldsymbol{P}(\boldsymbol{f})\big)^{-1}=\sum_{n=0}^{\infty}\big(\beta\boldsymbol{P}(\boldsymbol{f})\big)^n$$

δηλαδή, αποδείχθηκε το ζητούμενο.

■

Τέλος, αντικαθιστώντας στην (1.1.2) ,λαμβάνουμε ότι :

$$\boldsymbol{v}_\beta(\boldsymbol{f})=[\boldsymbol{I}-\beta\boldsymbol{P}(\boldsymbol{f})]^{-1}\boldsymbol{r}(\boldsymbol{f})$$

ή, ισοδύναμα, ότι:

$$\boldsymbol{v}_\beta(\boldsymbol{f})=\boldsymbol{r}(\boldsymbol{f})+\beta\boldsymbol{P}(\boldsymbol{f})\boldsymbol{v}_\beta(\boldsymbol{f}) \qquad (1.1.4)$$

1.1.2 Τερματιζόμενο Μαρκοβιανό μοντέλο Απόφασης Γ_T

Σε αυτή τη μοντελοποίηση της διαδικασίας απόφασης, κάνουμε την υπόθεση πως :

$$\sum_{s'=1}^{N} p(s'|s,a) < 1\,, \qquad \text{για όλα τα } a \in A(s), s \in S$$

Σύμφωνα με την υπόθεση αυτή, για κάθε ενέργεια $a \in A(s)$ και για κάθε κατάσταση s, υπάρχει μια θετική πιθανότητα απορρόφησης :

$$p(0|s,a) \doteqdot 1 - \sum_{s'=1}^{N} p(s'|s,a) > 0$$

η οποία σηματοδοτεί τον τερματισμό της διαδικασίας (ή την απορρόφησή της από μία 'τεχνητή' κατάσταση $s_T = 0$)

Ανάλογα με την περίπτωση της διαδικασίας με προεξόφληση, για το διάνυσμα αξίας έχουμε ότι:

$$\boldsymbol{v}_T(\boldsymbol{f}) = \sum_{t=0}^{\infty} \boldsymbol{P^t}(\boldsymbol{f})\, \boldsymbol{r}(\boldsymbol{f}) = [\boldsymbol{I} - \boldsymbol{P}(\boldsymbol{f})]^{-1}\boldsymbol{r}(\boldsymbol{f})$$

1.2 Η Μαρκοβιανή Διαδικασία Απόφασης Πεπερασμένου Ορίζοντα

Στην προηγούμενη παράγραφο θεωρήσαμε πως ο χρονικός ορίζοντας των μοντέλων ήταν ,εν γένει, άπειρος. Στο σημείο αυτό θα θεωρήσουμε πεπερασμένο χρονικό ορίζοντα, έχοντας ότι

$$t \in \{0,1,2,\dots,T\}$$

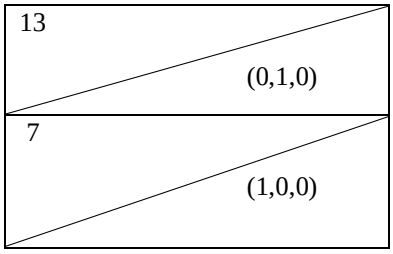
Οι στρατηγικές που έχουμε ορίσει μέχρι στιγμής χαρακτηρίζονται ως *στάσιμες*, με την έννοια ότι δεν μας ενδιαφέρει η χρονική στιγμή t κατά την οποία επισκεπτόμαστε την κατάσταση $s \in S$. Θα θέλαμε να επεκταθούμε σε μια γενικότερη κλάση στρατηγικών. Ο λόγος για αυτή την γενίκευση είναι πως η επιλογή μιας ενέργειας ενδέχεται να είναι μη ευνοϊκή στα πρώτα στάδια της διαδικασίας, επειδή θα μπορούσε να οδηγήσει σε μία 'κακή' κατάσταση, ενώ θα μπορούσε να είναι ευνοϊκή σε μια μεταγενέστερη χρονικά στιγμή, εφόσον δεν θα υπάρχει αρκετός χρόνος προτού τερματίσει η διαδικασία ώστε να επισκεφτούμε αυτή την 'κακή' κατάσταση. Αυτό σημαίνει πως η αξία μιας ενέργειας είναι πλέον συνάρτηση του χρόνου που απομένει ως τον τερματισμό, και έτσι, εικάζουμε πως και μια βέλτιστη στρατηγική οφείλει και αυτή να εξαρτάται από τον υπολειπόμενο χρόνο έως τον τερματισμό της διαδικασίας.

Παράδειγμα 1.2.1

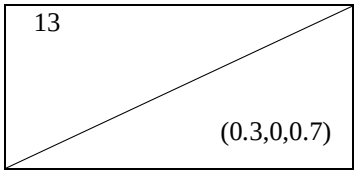
Έστω $S = \{1,2,3\},\ A(1) = \{1,2\}, A(2) = \{1\} = A(3), T = 2$ (δηλαδή $t \in \{0,1,2\}$)

Ενώ οι αμοιβές και οι πιθανότητες μετάβασης δίνονται παρακάτω:

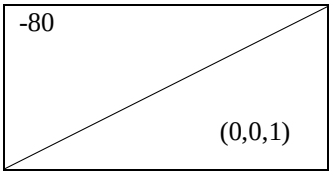
(Σε κάθε ορθογώνιο, πάνω αριστερά αναγράφονται οι αμοιβές ενός βήματος και μέσα στην παρένθεση αναγράφονται οι πιθανότητες μετάβασης σε όλες τις καταστάσεις)



Κατάσταση 1



Κατάσταση 2



Κατάσταση 3

Υποθέτουμε ότι ο ελεγκτής προσπαθεί να μεγιστοποιήσει το άθροισμα των αναμενόμενων αμοιβών στους χρόνους $t = 0,1,2$. Εφόσον ο ελεγκτής έχει περισσότερες από μία ενέργειες μόνο στην κατάσταση 1, κάθε στάσιμη στρατηγική θα είναι της μορφής :

$$\boldsymbol{f}_p = \big((p, 1-p), (1), (1)\big), \quad \text{με } p \in (0,1)$$

Στο συγκεκριμένο παράδειγμα, το σύνολο των καθαρών στάσιμων στρατηγικών θα είναι το $F_D = \{\boldsymbol{f}_0, \boldsymbol{f}_1\}$. Θέλουμε τώρα να υπολογίσουμε το άθροισμα των αμοιβών για τα στάδια 0,1,2, ξεκινώντας από την κατάσταση 1 και ακολουθώντας την στρατηγική $\boldsymbol{f}_p$.

$$\begin{aligned}\boldsymbol{v}_2\big(1, \boldsymbol{f}_p\big) &= [13p + 7(1-p)] + \{(1-p)[13p + 7(1-p)] + 13p\} \\ &\quad + \big\{((1-p)^2 + 0.3p)\big(13p + 7(1-p)\big) + 13p(1-p) - 56p\big\} \\ &= 21 - 30.9p - 22.2p^2 + 6p^3\end{aligned}$$

Η παραπάνω συνάρτηση του p λαμβάνει ακρότατο (μέγιστο) στο $p = 0$, με $\boldsymbol{v}_2(1, \boldsymbol{f}_0) = 21$. Συνεπώς, το καλύτερο που μπορεί να πετύχει ο ελεγκτής ,στο σύνολο των στάσιμών στρατηγικών, είναι να επιλέξει την στρατηγική $\boldsymbol{f}_0$.

Εάν, ωστόσο, θεωρήσουμε τη νέα στρατηγική $\boldsymbol{\pi} = (\boldsymbol{f}_0, \boldsymbol{f}_1, \boldsymbol{f}_1)$, που σημαίνει ότι θα χρησιμοποιήσουμε την $\boldsymbol{f}_0$ στο $t = 0$ και την $\boldsymbol{f}_1$ για $t = 1,2$, τότε προκύπτει πως :

$$\boldsymbol{v}_2(1, \boldsymbol{\pi}) = 7 + 13 + 13 = 33$$

Μπορούμε να ερμηνεύσουμε το παραπάνω ως εξής:

Ο ελεγκτής μπορεί να έχει μεγαλύτερο κέρδος εάν επιλέξει την ενέργεια 1 στην κατάσταση 1 στον χρόνο $t = 1$, παρά στον χρόνο $t = 0$, εφόσον δεν θα χρειάζεται να ανησυχεί για το κόστος 80 μονάδων στην κατάσταση 3, αφού δεν θα υπάρχει αρκετός χρόνος ώστε η διαδικασία να φτάσει στην κατάσταση 3.

Το παραπάνω παράδειγμα μας οδηγεί να ορίσουμε μία στρατηγική ως μία πεπερασμένη ακολουθία

$$\boldsymbol{\pi} = (\boldsymbol{f}_0, \boldsymbol{f}_1, \boldsymbol{f}_2, \dots, \boldsymbol{f}_T)$$

όπου όλα τα $\boldsymbol{f}_t$ είναι στάσιμες στρατηγικές ($\boldsymbol{f}_t \in F_s$). Θα συμβολίζουμε το σύνολο αυτών των στρατηγικών με F_M^T και θα τις καλούμε *Μαρκοβιανές στρατηγικές* της διαδικασίας απόφασης πεπερασμένου ορίζοντα.

Για κάθε $\boldsymbol{\pi} \in F_M^T$, ορίζουμε: $E_{s\boldsymbol{\pi}}[R_t] \doteqdot E_{\boldsymbol{\pi}}[R_t | S_0 = s]$

Η μέση τιμή είναι καλά ορισμένη για κάθε $t = 0,1, \dots, T$ και η αξία Τ-σταδίου για την $\boldsymbol{\pi}$ θα είναι

$$v_T(s, \boldsymbol{\pi}) = \sum_{t=0}^{T} E_{s\boldsymbol{\pi}}[R_t], \qquad \text{για κάθε } s \in S$$

και το αντίστοιχο διάνυσμα για την $\boldsymbol{\pi}$, για όλες τις καταστάσεις, θα είναι :

$$\boldsymbol{v}_T(\boldsymbol{\pi}) = \big(v_T(1, \boldsymbol{\pi}), v_T(2, \boldsymbol{\pi}), \dots, v_T(N, \boldsymbol{\pi})\big)$$

Ένας από τους πιο διαδεδομένους τρόπους προκειμένου να υπολογίσουμε την αξία αλλά και την βέλτιστη στρατηγική της διαδικασίας, συνοψίζεται στον αλγόριθμο που θα ακολουθήσει, γνωστό και ως «*Προς τα πίσω αναδρομή του Δυναμικού Προγραμματισμού».* Η βασική ιδέα στην οποία στηρίζεται ο αλγόριθμος είναι *η αρχή της βελτιστοποίησης* , σύμφωνα με την οποία:

'Εάν υποθέσουμε πως η μέγιστη αμοιβή είναι γνωστή όταν απομένουν $(n-1)$ στάδια για τον τερματισμό της διαδικασίας, τότε για να βρούμε τη μέγιστη αμοιβή ενώ απομένουν n στάδια για τον τερματισμό, αρκεί να μεγιστοποιήσουμε το άθροισμα της αμοιβής του τωρινού σταδίου με την μέγιστη αναμενόμενη πληρωμή για τα υπόλοιπα $(n-1)$ στάδια που απομένουν.'

Θα συμβολίζουμε με $\underset{z \in Z}{arg\max}\ h(z)$ την τιμή του z για την οποία η πραγματική συνάρτηση $h(z)$ λαμβάνει το μέγιστο στο σύνολο Z.

Αλγόριθμος 1.2.1 – Προς τα πίσω αναδρομή του Δυναμικού Προγραμματισμού

Βήμα 1 (Αρχικοποίηση)

Θέτουμε $V_{-1}(s) = 0$ για όλα τα $s \in S$ και ορίζουμε:

$$f_T^*(s) \doteqdot a_s^T = \underset{a \in A(s)}{arg\max} \left\{ r(s,a) + \sum_{s'=1}^{N} p(s'|s,a)\, V_{-1}(s') \right\}$$

και

$$V_0(s) \doteqdot r(s, a_s^T) = \max_{a \in A(s)} \{r(s,a) + 0\}$$

Βήμα 2 (Αναδρομή)

Για κάθε $n = 1,2, \dots, T$ υπολογίζουμε, για κάθε $s \in S$:

$$f_{T-n}^* = a_s^{T-n} = arg\max_{a \in A(s)} \left\{ r(s,a) + \sum_{s'=1}^{N} p(s'|s,a)\, V_{n-1}(s') \right\}$$

και

$$V_n(s) \doteqdot r(s, a_s^{T-n}) + \sum_{s'=1}^{N} p(s'|s, a_s^{T-n})\, V_{n-1}(s')$$

Βήμα 3

Κατασκευάζουμε την στρατηγική $\boldsymbol{\pi}^* = (\boldsymbol{f}_0^*, \boldsymbol{f}_1^*, \dots, \boldsymbol{f}_T^*) \in F_M^T$

Σχόλια

- Η φράση 'προς τα πίσω αναδρομή' οφείλεται στα επαναληπτικά βήματα 2 και 3, όπου η $\boldsymbol{\pi}^*$ κατασκευάζεται αναδρομικά από τις $\boldsymbol{f}_T^*, \boldsymbol{f}_{T-1}^*, \dots, \boldsymbol{f}_0^*$.
- Για κάθε $s \in S$, για κάθε $\boldsymbol{\pi} \in F_M^T$ και $n = 0,1, \dots, T$, ορίζουμε:

  $$V_n(s, \boldsymbol{\pi}) = \sum_{t=T-n}^{\boldsymbol{T}} E_{\boldsymbol{\pi}}[R_{\boldsymbol{t}} | S_{\boldsymbol{T-n}} = s]$$

  οπότε το $V_n(s, \boldsymbol{\pi})$ αναπαριστά την αναμενόμενη αμοιβή για τα τελευταία n-στάδια, δεδομένου ότι στο χρόνο $T - n$ βρισκόμαστε στην s. Για $n = T$, λαμβάνουμε τη συνολική αναμενόμενη αμοιβή, δηλαδή :

  $$V_T(s, \boldsymbol{\pi}) = v_T(s, \boldsymbol{\pi}), \qquad \text{για όλα τα } s \in S, \boldsymbol{\pi} \in F_M^T$$

Θεώρημα 1.2.1

Θεωρούμε την Μαρκοβιανή διαδικασία απόφασης Τ-ορίζοντα Γ_T και έστω $\boldsymbol{\pi}^* \in F_M^T$ μια στρατηγική κατασκευασμένη όπως υποδεικνύεται από τον Αλγόριθμο 2.2.1. Τότε, η $\boldsymbol{\pi}^*$ είναι βέλτιστη στρατηγική και, για όλα τα $n = 0,1, \dots, T, s \in S$:

$$V_n(s) = \max_{a \in A(s)} \left\{ r(s,a) + \sum_{s'=1}^{N} p(s'|s,a)\, V_{n-1}(s') \right\} \qquad (1.2.1)$$

Απόδειξη

Η εξίσωση (1.2.1) είναι συνέπεια του ορισού του $V_n(s)$ από το Βήμα 2 του παραπάνω Αλγορίθμου. Μένει να δείξουμε ότι η $\boldsymbol{\pi}^*$ είναι βέλτιστη.

Για $n = 0$ και μία αυθαίρετη στρατηγική $\boldsymbol{\pi} = (f_0, f_1, \dots, f_T) \in F_M^T$, από το Βήμα 1 του αλγορίθμου, για κάθε $s' \in S$, λαμβάνουμε ότι:

$$V_0(s', \boldsymbol{\pi}) = E_{\boldsymbol{\pi}}[R_{\boldsymbol{T}}|S_{\boldsymbol{T}} = s'] \leq \max_{a \in A(s')} \{r(s', a\} = E_{\boldsymbol{\pi}^*}\,[R_{\boldsymbol{T}}|S_{\boldsymbol{T}} = s'] = V_0(s')$$

Επαγωγικά, υποθέτουμε ότι το ζητούμενο ισχύει για $n-1$, δηλαδή ότι:

$$V_{n-1}(s', \boldsymbol{\pi}) = \sum_{t=T-n+1}^{T} E_{\boldsymbol{\pi}}[R_t|S_{T-n+1} = s'] \leq \sum_{t=T-n+1}^{T} E_{\boldsymbol{\pi}^*}[R_t|S_{T-n+1} = s'] = V_{n-1}(s', \boldsymbol{\pi}^*)$$

$$= V_{n-1}(s') \qquad (1.2.2)$$

για όλα τα $s' \in S$. Ακόμη, έχουμε ότι:

$$V_n(s'', \boldsymbol{\pi}) = E_{\boldsymbol{\pi}}\left[\sum_{t=T-n}^{T} (R_{\boldsymbol{t}}|S_{T-n} = s'')\right]$$

$$= \sum_{a \in A(s'')} E_{\boldsymbol{\pi}}\left[\sum_{t=T-n}^{T} (R_{\boldsymbol{t}}|S_{T-n} = s'', A_{T-n} = a)\right] f_{T-n}(s'', a) \qquad (1.2.3)$$

Τώρα, θα γράψουμε τον όρο με τη μέση τιμή της (1.2.3) ως:

$$E_{\boldsymbol{\pi}}\left[\sum_{t=T-n}^{T} (R_{\boldsymbol{t}}|S_{T-n} = s'', A_{T-n} = a)\right] =$$

$$= r(s'', a) + \sum_{s'=1}^{N} p(s'|s'', a)\, E_{\boldsymbol{\pi}}\left[\sum_{t=T-n+1}^{T} (R_t|S_{T-n} = s'', A_{T-n} = a, S_{T-n+1} = s')\right]$$

και επειδή η κατανομή του $\sum_{t=T-n+1}^{T} R_t$, δεδομένου του γεγονότος $\{S_{T-n+1} = s'\}$ και της $\boldsymbol{\pi}$, είναι ανεξάρτητη από όσα έχουν προηγηθεί στη διαδικασία πριν από το $(T-n+1)$ - οστό στάδιο,

$$= r(s'', a) + \sum_{s'=1}^{N} p(s'|s'', a) E_{\boldsymbol{\pi}}\left[\sum_{t=T-n+1}^{T} (R_{\boldsymbol{t}}|S_{T-n+1} = s')\right]$$

$$= r(s'', a) + \sum_{s'=1}^{N} p(s'|s'', a)\, V_{n-1}(s', \boldsymbol{\pi})$$

Αντικαθιστώντας την παραπάνω στην (1.2.3) , παίρνουμε:

$$V_n(s'', \boldsymbol{\pi}) = \sum_{a \in A(s'')} f_{T-n}(s'', a) \left(r(s'', a) + \sum_{s'=1}^{N} p(s'|s'', a)\, V_{n-1}(s', \boldsymbol{\pi}) \right)$$

Τώρα θα χρησιμοποιήσουμε την επαγωγική υπόθεση, οπότε:

$$V_n(s'', \boldsymbol{\pi}) \leq \sum_{a \in A(s'')} f_{T-n}(s'', a) \left(r(s'', a) + \sum_{s'=1}^{N} p(s'|s'', a)\, V_{n-1}(s', \boldsymbol{\pi}) \right)$$

$$\leq \sum_{a \in A(s'')} f_{T-n}(s'', a) \left\{ r(s'', a_{s''}^{T-n}) + \sum_{s'=1}^{N} p(s'|s'', a_{s''}^{T-n})\, V_{n-1}(s') \right\}$$

(λόγω του βήματος 2 του αλγορίθμου)

$$= \sum_{a \in A(s'')} f_{T-n}(s'', a)\, V_{\boldsymbol{n}}(s'') = V_{\boldsymbol{n}}(s'')$$

για όλα τα $s'' \in S$ και $\boldsymbol{\pi} \in F_M^T$, το οποίο ολοκληρώνει την απόδειξη.

■

Παράδειγμα 1.2.2

Σε αυτό το παράδειγμα θα υπολογίσουμε μια βέλτιστη στρατηγική για το Παράδειγμα 1.2.1. Θα δούμε ότι, χρησιμοποιώντας τον Αλγόριθμο 1.2.1, θα είμαστε σε θέση να επιβεβαιώσουμε ότι η στρατηγική $\boldsymbol{\pi} = (\boldsymbol{f_0}, \boldsymbol{f_1}, \boldsymbol{f_2})$ που χρησιμοποιήσαμε στο παράδειγμα, είναι πράγματι μια 'καλή' στρατηγική.

Αρχικοποιούμε τον αλγόριθμο με τις τιμές:

$$\boldsymbol{V}_0 = \big(V_0(1), V_0(2), V_0(3)\big)^T = (13{,}13, -80)^T, \qquad f_2^*(1) = f_2^*(2) = f_2^*(3) = 1$$

Για $n = 1$,

$$f_1^*(1) = a_1^{2-1} = argmax \left(r(1{,}1) + \sum_{s'} p(s'|1{,}1) V_0(s'), r(1{,}2) + \sum_{s'} p(s'|1{,}2) V_0(s') \right)$$

$$= argmax\ (13 + 13{,}7 + 13) = argmax\ (26) = 1$$

$$V_1(1) = r(1,1) + \sum_{s'} p(s'|1,1)V_0(s') = 13 + 13 = 26$$

$$f_1^*(2) = a_2^{2-1} = argmax\left(r(2,1) + \sum_{s'} p(s'|2,1)V_0(s') \right)$$
$$= argmax\ (13 + 13(0.3) - 80(0.7)) = argmax\ (-39.1) = 1$$
$$V_1(2) = r(2,1) + \sum_{s'} p(s'|2,1)V_0(s') = 13 + 3.9 - 56 = -39.1$$

$$f_1^*(3) = a_3^{2-1} = argmax\left(r(3,1) + \sum_{s'} p(s'|3,1)V_0(s') \right) = argmax\ (-80 + (-80))$$
$$= argmax\ (-160) = 1$$
$$V_1(3) = r(3,1) + \sum_{s'} p(s'|3,1)V_0(s') = -160$$

Για $n = 2$,

$$f_0^*(1) = a_1^{2-2} = argmax\left(r(1,1) + \sum_{s'} p(s'|1,1)V_1(s'), r(1,2) + \sum_{s'} p(s'|1,2)V_1(s') \right)$$
$$= argmax\ (13 + (-39.1), 7 + 26) = argmax\ (33) = 2$$
$$V_2(1) = r(1,2) + \sum_{s'} p(s'|1,2)V_1(s') = 7 + 26 = 33$$

$$f_0^*(2) = a_2^{2-2} = argmax\left(r(2,1) + \sum_{s'} p(s'|2,1)V_1(s') \right)$$
$$= argmax\ (13 + 26(0.3) - 160(0.7)) = argmax\ (-91.2) = 1$$
$$V_2(2) = r(2,1) + \sum_{s'} p(s'|2,1)V_1(s') = -91.2$$

$$f_0^*(3) = a_3^{2-2} = argmax\left(r(3,1) + \sum_{s'} p(s'|3,1)V_1(s') \right) = argmax\ (-80 + (-160))$$
$$= argmax\ (-240) = 1$$

$$V_2(3) = r(3,1) + \sum_{s'} p(s'|3,1)V_1(s') = -80 - 160 = -240$$

Έχουμε πλέον δείξει ότι $f_0^*(1) = 2$, που σημαίνει ότι $f_0 = \boldsymbol{f}_0$ και

$f_1^*(1) = f_2^*(1) = 1$, που σημαίνει ότι $f_1 = f_2 = \boldsymbol{f}_1$

Δηλαδή έχουμε επιβεβαιώσει ότι η $\boldsymbol{\pi} = (\boldsymbol{f_0}, \boldsymbol{f_1}, \boldsymbol{f_1})$ του παραδείγματος 1.2.1 είναι πράγματι μια βέλτιστη στρατηγική.

Σχόλια

- Παρά την απλότητα στην έκφραση του Αλγορίθμου 1.2.1 και του αντίστοιχου θεωρήματος, ίσως μπορεί να διαισθανεί κανείς πως η χρήση τους ενδέχεται να επιφέρει υπολογιστικές δυσκολίες. Το αναδρομικό βήμα του αλγορίθμου εμπεριέχει $T \times N$ μεγιστοποιήσεις, αριθμός που προφανώς είναι ανάλογος τόσο του χρονικού ορίζοντα T, όσο και του πλήθους των καταστάσεων N. Ακόμη, εάν έχουμε έναν 'μεγάλο' χώρο ενεργειών $A(s)$, τότε καθε μία από αυτές τις μεγιστοποιήσεις ενδέχεται να είναι αρκετά χρονοβόρα. Λόγω αυτών, ο παραπάνω αλγόριθμος και το σχετικό θεώρημα, φαίνεται να μην είναι ο καταλληλότερος τρόπος αντιμετώπισης προβλημάτων με πολλές καταστάσεις, πολλές ενέργειες ή/και μεγάλο χρονικό ορίζοντα.
- Ένα πρακτικό πρόβλημα που συναντάται στη μοντελοποίηση με πεπερασμένο ορίζοντα, απορρέει από το γεγονός πως, εν γένει, οι βέλτιστες στρατηγικές είναι μη-στάσιμες. Το βήμα 3 του αλγορίθμου υποδεικνύει πως μία βέλτιστη Μαρκοβιανή στρατηγική προϋποθέτει από τον decision-maker να θυμάται και την τρέχουσα κατάσταση της διαδικασίας , αλλά και το τρέχον στάδιο. Παρόλο που αυτό φαίνεται απολύτως λογικό από μαθηματικής σκοπιάς, σε πολλές πρακτικές εφαρμογές, αυτό το επίπεδο πολυπλοκότητας της βέλτιστης στρατηγικής ενδέχεται να αποθαρρύνει τον d-m από τη χρήση της, ο οποίος θα προτιμούσε μία αρκετά 'απλούστερη' βέλτιστη πολιτική.

1.3 Γραμμικός Προγραμματισμός και Αθροίσιμες Μαρκοβιανές Διαδικασίες Απόφασης

Στην παράγραφο 1.1 θεωρήσαμε τη διαδικασία Γ_β με συντελεστή προεξόφλησης β. Το πρόβλημα βελτιστοποίησης που είναι συνδεδεμένο με την διαδικασία αυτή, αφορά στη μεγιστοποίηση της τιμής, ενώ βρισκόμαστε στο σύνολο των στάσιμων στρατηγικών, δηλαδή:

$$\max \boldsymbol{v}_\beta(\boldsymbol{f})$$

με περιορισμούς (I)

$$\boldsymbol{f} \in F_s$$

Μία στρατηγική $\boldsymbol{f}_0 \in F_s$ για την οποία επιτυγχάνεται το παραπάνω μέγιστο, θα καλείται βέλτιστη και ,τότε, το διάνυσμα αξίας για την $\boldsymbol{f}_0$ θα είναι:

$$\boldsymbol{v}_\beta \div \boldsymbol{v}_\beta(\boldsymbol{f}_0) = \max_{\boldsymbol{f}} \boldsymbol{v}_\beta(\boldsymbol{f})$$

Φυσικά ακόμη δεν γνωρίζουμε καν εάν υπάρχει τέτοια βέλτιστη $\boldsymbol{f}_0$ και το αντίστοιχο διάνυσμα τιμής. Θα δούμε στη συνέχεια αυτής της παραγράφου όχι μόνο ότι υπάρχουν τέτοιες βέλτιστες στρατηγικές, αλλά και ότι αντιστοιχούν σε λύσεις κατάλληλα κατασκευασμένων γραμμικών προγραμμάτων.

Έχουμε δείξει τη σχέση

$$\boldsymbol{v}_\beta(\boldsymbol{f}) = [\boldsymbol{I} - \beta \boldsymbol{P}(\boldsymbol{f})]^{-1} \boldsymbol{r}(\boldsymbol{f}) \qquad (1.3.1)$$

Λόγω της μη γραμμικότητας της παραπάνω σχέσης ως προς $\boldsymbol{f}$, δεν μπορούμε να γνωρίζουμε με βεβαιότητα την ύπαρξη του *max* στο (I) , πόσο μάλλον την υπολογιστική δυσκολία αυτού του εγχειρήματος.

Πολλαπλασιάζοντας την (1.3.1) επί $\boldsymbol{I} - \beta \boldsymbol{P}(\boldsymbol{f})$, παίρνουμε:

$$[\boldsymbol{I} - \beta \boldsymbol{P}(\boldsymbol{f})]\boldsymbol{v}_\beta(\boldsymbol{f}) = \boldsymbol{r}(\boldsymbol{f}) \Rightarrow \boldsymbol{v}_\beta(\boldsymbol{f}) = \boldsymbol{r}(\boldsymbol{f}) + \beta \boldsymbol{P}(\boldsymbol{f})\boldsymbol{v}_\beta(\boldsymbol{f}) \qquad (1.3.2)$$

Συνεπώς, εάν υπάρχουν τα $\boldsymbol{f}_0 \in F_s$ και $\boldsymbol{v}_\beta(\boldsymbol{f}_0)$, θα πρέπει να ικανοποιούν την (1.3.2).
Ακόμη, από την αρχή της βελτιστοποίσης, θα πρέπει το $\boldsymbol{v}_\beta$ να ικανοποιεί και την:

$$v_\beta(s) = \max_{a \in A(s)} \left\{ r(s,a) + \beta \sum_{s'=1}^{N} p(s'|s,a)\, v_\beta(s') \right\} \qquad (1.3.3)$$

όπου με $v_\beta(s)$ συμβολίζουμε το s-οστό στοιχείο του $\boldsymbol{v}_\beta$.

Η τελευταία ισότητα γεννά ένα σύνολο γραμμικών ανισοτήτων. Έστω $\boldsymbol{v}$ ένα αυθαίρετο διάνυσμα που τις ικανοποιεί, δηλαδή έστω:
$\boldsymbol{v} = \big(v(1), v(2), \dots, v(N)\big)^T$, τέτοιο ώστε

$$v(s) \geq r(s, \boldsymbol{f}) + \beta \sum_{s'=1}^{N} p(s'|s, \boldsymbol{f})\, v(s'), \qquad \text{για κάθε } s \in S$$

ή, σε μορφή πινάκων:

$$\boldsymbol{v} \geq \boldsymbol{r}(\boldsymbol{f}) + \beta \boldsymbol{P}(\boldsymbol{f})\boldsymbol{v}$$

Αντικαθιστώντας την παραπάνω ανισότητα στον εαυτό της k-φορές , παίρνουμε ότι:

$$\boldsymbol{v} \geq \boldsymbol{r}(\boldsymbol{f}) + \beta \boldsymbol{P}(\boldsymbol{f})\boldsymbol{r}(\boldsymbol{f}) + \cdots + \beta^{k-1}\boldsymbol{P}^{k-1}(\boldsymbol{f})\boldsymbol{r}(\boldsymbol{f}) + \beta^{k}\boldsymbol{P}^{k}(\boldsymbol{f})\boldsymbol{v}$$

Αφήνουμε τώρα $k \to \infty$, και προκύπτει:

$$\boldsymbol{v} \geq [\boldsymbol{I} - \beta \boldsymbol{P}(\boldsymbol{f})]^{-1}\boldsymbol{r}(\boldsymbol{f}) = \boldsymbol{v}_{\beta}(\boldsymbol{f})$$

Δηλαδή, το αυθαίρετο $\boldsymbol{v}$ που επιλέξαμε να ικανοποιεί τις παραπάνω ανισότητες, αποτελεί ένα άνω φράγμα για την τιμή με προεξόφληση , στο σύνολο των στάσιμων στρατηγικών F_s.

Ακριβώς αυτή η τελευταία παρατήρηση μας επιτρέπει να αρχίσουμε να υποψιαζόμαστε πως το διάνυσμα τιμής $\boldsymbol{v}_{\beta}$ της διαδικασίας Γ_{β} θα μπορούσε να αποτελεί την βέλτιστη λύση για το γραμμικό πρόβλημα:

$$\min \sum_{s=1}^{N} \frac{1}{N} v(s)$$

με περιορισμούς (P_{β})

$$v(s) \geq r(s,a) + \beta \sum_{s'=1}^{N} p(s'|s,\alpha)\, v(s')\,, \quad a \in A(s), s \in S$$

Σχόλιο

Οι συντελεστές $\frac{1}{N}$ στην αντικειμενική συνάρτηση του (P_{β}) ερμηνεύονται από το γεγονός πως θεωρούμε ισοπίθανο η αρχική κατάσταση να είναι οποιαδήποτε από τις $\{1,2,\dots,N\}$.

Πριν προχωρήσουμε στη διατύπωση του δυϊκού προβλήματος που αντιστοιχεί στο (P_{β}), ας δούμε πρώτα τον τρόπο με τον οποίο μπορούμε να κατασκευάσουμε ένα δυϊκό πρόβλημα γραμμικού προγραμματισμού, στηριζόμενοι στο αντίστοιχο πρωτεύον αυτού.

Αρχικά, θέλουμε το πρωτεύον πρόγραμμα να είναι γραμμένο σε *κανονική μορφή*. Δηλαδή , να είναι της μορφής:

$$\begin{aligned} &\max \boldsymbol{c}^{T}\boldsymbol{x} \\ &\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b} \qquad (P) \\ &\boldsymbol{x} \geq \boldsymbol{0} \end{aligned}$$

όπου $\boldsymbol{c} \in \mathbb{R}_{n\times 1}, \boldsymbol{x} \in \mathbb{R}_{n\times 1}, \boldsymbol{b} \in \mathbb{R}_{m\times 1}, \boldsymbol{A} \in \mathbb{R}_{m\times n}$

Τότε, το *δυϊκό* πρόβλημα (D) που αντιστοιχεί στο (P) θα έχει τη μορφή:

$$
\begin{aligned}
&min\ \boldsymbol{b}^T\boldsymbol{w} \\
&\boldsymbol{A}^T\boldsymbol{w} \geq \boldsymbol{c} \qquad\qquad (D) \\
&\boldsymbol{w} \geq \boldsymbol{0}
\end{aligned}
$$

όπου $\boldsymbol{w} \in \mathbb{R}_{m\times 1}$.

Η διαδικασία την οποία ακολουθούμε για να κατασκευάσουμε τους περιορισμούς και τις μεταβλητές του (D) είναι η εξής:

(i) Ορίζουμε μια δυϊκή μεταβλητή w_i για καθέναν από τους m περιορισμούς του πρωτεύοντος.
(ii) Για κάθε μία από τις n μεταβλητές του πρωτεύοντος, ορίζεται ένας δυϊκός περιορισμός.
(iii) Οι συντελεστές των μεταβλητών στους περιορισμούς του δυϊκού είναι ίσοι με τους συντελεστές της συνδεόμενης μεταβλητής του πρωτεύοντος (ανεστραμμένοι). Το δεξιό μέλος των περιορισμών του δυϊκού θα είναι ίσο με τους συντελεστές της αντικειμενικής συνάρτησης του πρωτεύοντος.
(iv) Οι συντελεστές της αντικειμενικής συνάρτησης του δυϊκού είναι ίσοι με τα δεξιά μέλη των περιορισμών του πρωτεύοντος.

Παράδειγμα 1.3.1

Έστω το παρακάτω πρόγραμμα γραμμικού προγραμματισμού:

$$\max z = 5x_1 + 6x_2 + 2x_3$$

με περιορισμούς:

$$4x_1 + 3x_2 + 2x_3 \leq 20$$

$$2x_1 + 5x_2 + x_3 \geq 8$$

$$0.5x_1 + 3.5x_2 + 2x_3 \leq 12$$

Για να υπολογίσουμε το δυϊκό του παραπάνω προβλήματος, πρέπει πρώτα να μετατρέψουμε τον 2ο περιορισμό , έτσι ώστε να συμφωνεί με την κανονική μορφή. Οπότε ο 2ος περιορισμός θα γίνει:

$$-2x_1 - 5x_2 - x_3 \leq -8$$

Οπότε το δυϊκό πρόβλημα θα έχει τη μορφή:

$$\min u = 20w_1 - 8w_2 + 12w_3$$

με περιορισμούς:

$$4w_1 - 2w_2 + 0.5w_3 \geq 5$$

$$3w_1 - 5w_2 + 3.5w_3 \geq 6$$

$$2w_1 - w_2 + 2w_3 \geq 2$$

Ωστόσο, πολλές φορές η διαδικασία υπολογισμού της κανονικής μορφής ενός γραμμικού προβλήματος είναι ιδιαίτερα χρονοβόρα. Για αυτές τις περιπτώσεις, κατασκευάζουμε το δυϊκό πρόγραμμα σύμφωνα με τον παρακάτω πίνακα:

Πρωτεύον		Δυϊκό
maximize z		minimize u
i-περιορισμός μορφής $=$		i-μεταβλητή $w_i \geq 0$
i-περιορισμός μορφής $\leq$		i-μεταβλητή $w_i \in \mathbb{R}$
i-περιορισμός μορφής $\geq$	$\Leftrightarrow$	i-μεταβλητή $w_i \leq 0$
i-μεταβλητή $x_i \geq 0$		i-περιορισμός μορφής $\geq$
i-μεταβλητή $x_i \in \mathbb{R}$		i-περιορισμός μορφής $=$
i-μεταβλητή $x_i \leq 0$		i-περιορισμός μορφής $\leq$

Δηλαδή, ο παραπάνω πίνακας ερμηνεύεται ως εξής:

- Αν μια μεταβλητή του ενός δεν έχει περιορισμό στο πρόσημο, ο αντίστοιχος περιορισμός του άλλου είναι εξίσωση και αντίστροφα.
- Αν μια μεταβλητή του ενός είναι μη θετική, τότε ο αντίστοιχος περιορισμός του άλλου είναι ανίσωση με φορά αντίθετη της αναμενόμενης και αντίστροφα.

Τα σύνολα $\{\boldsymbol{x} \in \mathbb{R}^n | \boldsymbol{Ax} \leq \boldsymbol{b}, \boldsymbol{x} \geq \boldsymbol{0}\}$ και $\{\boldsymbol{w} \in \mathbb{R}^m | \boldsymbol{w}^T\boldsymbol{A} \geq \boldsymbol{c}^T, \boldsymbol{w} \geq \boldsymbol{0}\}$ ονομάζονται *εφικτά σύνολα* και, όταν αυτά είναι μη-κενά, το αντίστοιχο πρόβλημα θα καλείται *εφικτό*.

Ορισμός 1.3.1.

Ένα στοιχείο $\boldsymbol{x}$ του εφικτού συνόλου $\{\boldsymbol{x} \in \mathbb{R}^n | \boldsymbol{Ax} \leq \boldsymbol{b}, \boldsymbol{x} \geq \boldsymbol{0}\}$ θα καλείται *εφικτή λύση*. Μία εφικτή λύση $\boldsymbol{x}$ θα καλείται *βασική εφικτή λύση* οποτεδήποτε είναι δυνατό να επιλέξουμε έναν $r \times r$ υποπίνακα A_r του A τέτοιο ώστε:

(i) Ο $\boldsymbol{A}_r$ είναι μη-ιδιάζων
(ii) $\boldsymbol{A}_r\boldsymbol{x}_r = \boldsymbol{b}_r$
(iii)$x_j = 0$ για όλα τα j πέρα των r επιλεγμένων στηλών

Παρατηρούμε ότι , όποτε υπάρχει μία βασική εφικτή λύση, τότε αυτή καθορίζεται πλήρως από τα $\boldsymbol{A}_r$ και $\boldsymbol{b}_r$. Η επόμενη πρόταση φανερώνει την σημασία της ύπαρξης τέτοιων βασικών εφικτών λύσεων.

Πρόταση 1.3.1.

(i) Εάν το εφικτό σύνολο είναι μη-κενό, τότε υπάρχει βασική εφικτή λύση
(ii) Εάν υπάρχει βέλτιστη εφικτή λύση, τότε υπάρχει βέλτιστη βασική εφικτή λύση

Η σημασία του (ii) της παραπάνω πρότασης έγκειται στο γεγονός πως αναζητώντας μια βέλτιστη λύση, θα μπορούσαμε να περιοριστούμε στο πεπερασμένο σύνολο των βασικών εφικτών λύσεων και έπειτα, μέσω της γνωστής μεθόδου *Simplex* να φτάσουμε στο ζητούμενο αποτέλεσμα.

Η γεωμετρική ερμηνεία του εφικτού συνόλου είναι η ακόλουθη. Κάθε γραμμικός περιορισμός της μορφής:

$$\sum_{j=1}^{n} \alpha_{ij} x_j \leq b_i \quad , \quad i = 1,2, \dots, m$$

ορίζει ένα κλειστό 'ημι-επίπεδο' στον $\mathbb{R}^n$. Τότε, μπορούμε να δούμε το εφικτό σύνολο ως την τομή m στο πλήθος τέτοιων 'ημι-επιπέδων', το οποίο εάν είναι μη-κενό, θα καλείται *πολύεδρο.* Τα ακραία σημεία του πολυέδρου αυτού βρίσκονται σε 1-1 αντιστοιχία με τις βασικές εφικτές λύσεις.

Πρόταση 1.3.2. (Δυϊκό Θεώρημα)

(i) Εάν είτε το Πρωτεύον είτε το Δυϊκό πρόβλημα έχει πεπερασμένη βέλτιστη λύση, τότε το ίδιο ισχύει και για το άλλο πρόβλημα, και οι αντίστοιχες τιμές της αντικειμενικής συνάρτησης είναι ίσες.
(ii) Εάν είτε το Πρωτεύον είτε το Δυϊκό πρόβλημα είναι μη φραγμένο, τότε το άλλο πρόβλημα δεν έχει εφικτές λύσεις.

Πρόταση 1.3.3. (Συμπληρωματική Χαλαρότητα)

Μια αναγκαία και ικανή συνθήκη ώστε ένα ζεύγος εφικτών λύσεων $(\boldsymbol{w}, \boldsymbol{x})$ να είναι βέλτιστο είναι, για όλα τα i, j :

$$(a) \quad w_i > 0 \Longrightarrow \sum_{i=1}^{m} a_{ij} w_i = c_j$$

$$(b) \quad \sum_{j=1}^{n} a_{ij} x_j < b_i \Longrightarrow w_i = 0$$

$$(c) \quad x_j > 0 \Longrightarrow \sum_{i=1}^{m} a_{ij} w_i = c_j$$

$$(d) \quad \sum_{i=1}^{m} a_{ij} w_i > c_j \Longrightarrow x_j = 0$$

Είμαστε πλέον σε θέση να συνεχίσουμε με την ανάλυση του (P_β). Εάν θεωρήσουμε το (P_β) ως ένα πρωτεύον (Primal) πρόβλημα γραμμικού προγραμματισμού, και συσχετίσουμε με κάθε περιορισμό τη δυϊκή μεταβλητή x_{sa}, τότε το δυϊκό (Dual) πρόγραμμα (D_β) του (P_β) θα έχει την ακόλουθη μορφή :

$$\max \sum_{s=1}^{N} \sum_{a=1}^{m(s)} r(s,a) x_{sa}$$

με περιορισμούς (D_β)

$$\sum_{s=1}^{N} \sum_{a=1}^{m(s)} [\delta_{ss'} - \beta p(s'|s,a)]\, x_{sa} = \frac{1}{N} \ , s' \in S$$

$$x_{sa} \geq 0 \ , a \in A(s), s \in S$$

Θεώρημα 1.3.1

(i) Τα γραμμικά προγράμματα (P_β) και (D_β) έχουν πεπερασμένες βέλτιστες λύσεις

(ii) Έστω $\boldsymbol{v}^0 = \left(v^0(1), v^0(2), \dots, v^0(N)\right)^T$ μία βέλτιστη λύση του (P_β). Τότε $\boldsymbol{v}^0 = \boldsymbol{v}_\beta$, όπου $\boldsymbol{v}_\beta$ είναι το διάνυσμα-τιμής της διαδικασίας Γ_β.

(iii) Έστω $x^0 = \{x_{sa}^0 | a \in A(s), s \in S\}$ μία βέλτιστη λύση του (D_β) και έστω

$$x_0^s \doteqdot \sum_{a=1}^{m(s)} x_{sa}^0 \ \ \text{, για κάθε } s \in S$$

Τότε $x_0^s > 0$ και η $\boldsymbol{f}^0 \in F_s$ που ορίζεται ως:

$$\boldsymbol{f}^0(s,a) \doteqdot \frac{x_{sa}^0}{x_s^0} \ \ \text{, για όλα τα } a \in A(s), s \in S$$

είναι μία βέλτιστη στάσιμη στρατηγική για την διαδικασία Γ_β.

Πόρισμα 1.3.1

(i) Το διάνυσμα τιμής $\boldsymbol{v}_\beta$ είναι η μοναδική λύση της εξίσωσης βελτιστοποίησης (1.3.3).

(ii) Για κάθε $s \in S$, έστω μια ενέργεια $a_s \in A(s)$ για την οποία επιτυγχάνεται το *max* στην (1.3.3), δηλαδή:

$$v(s) = r(s,a_s) + \beta \sum_{s'=1}^{N} p(s'|s,a_s)v(s') = \max_{a \in A(s)} \left\{ r(s,a) + \beta \sum_{s'=1}^{N} p(s'|s,a)\, v(s') \right\}$$

όπου $\boldsymbol{v} = \big(v(1), v(2), \dots, v(N)\big)$ είναι η λύση της (1.3.3). Ορίζουμε την $\boldsymbol{f}^* \in F_s$ ως εξής:

$$\boldsymbol{f}^*(s,a) = \begin{cases} 1, & \alpha\nu\ a = a_s \\ 0, & \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases} \ \ \text{, για όλα τα } s \in S$$

Τότε η $\boldsymbol{f}^*$είναι μία βέλτιστη στάσιμη καθαρή (ντετερμινιστική) λύση.

Παρατηρήσεις

- Η εγκυρότητα της εξίσωσης βελτιστοποίησης (1.3.3) υποδεικνύει πως το πρόβλημα εύρεσης βέλτιστης στρατηγικής είναι πλήρως αντιμετωπίσιμο από τη στιγμή που γίνει γνωστό το διάνυσμα $\boldsymbol{v}_\beta$, εφόσον το μόνο που απαιτείται είναι ο υπολογισμός N στο πλήθος *maxima* , όπως φαίνεται από το (ii) του Πορίσματος
- Η ανάλυση που προηγήθηκε για το μοντέλο της διαδικασίας Γ_β, ισχύει και στην περίπωση μοντελοποίησης της τερματιζόμενης διαδικασίας Γ_T. Η ειδοποιός διαφορά είναι πως στην ανάλυση της Γ_T θα παραλείπεται ο παράγοντας προεξόφλησης β από τις σχετικές εξισώσεις.

Για παράδειγμα, το αντίστοιχο πρωτεύον γραμμικό πρόγραμμα του (P_β) θα είναι το:

$$\min \sum_{i=1}^{N} \frac{1}{N} v(s)$$

με περιορισμούς (P_T)

$$v(s) = r(s,a) + \sum_{s'=1}^{N} p(s'|s,a)v(s'), \quad a \in A(s), s \in S$$

1.4 Η Αδιαχώριστη διαδικασία Οριακού Μέσου

Στην παράγραφο αυτή θα χρησιμοποιήσουμε ένα νέο κριτήριο απόδοσης το οποίο είναι διαφορετικό τόσο από το κριτήριο με προεξόφληση όσο και από εκείνο της τερματιζόμενης διαδικασίας. Θα θέλαμε να εστιάσουμε στη μέση απόδοση μιας στρατηγικής μακροπρόθεσμα (in the long run).

Συγκεκριμένα, ορίζουμε την *τιμή οριακού μέσου* της στάσιμης στρατηγικής $\boldsymbol{f}$, ξεκινώντας από μια αρχική κατάσταση s, ως εξής:

$$v_a(s, \boldsymbol{f}) = \lim_{T\to\infty} \left(\left(\frac{1}{T+1} \right) \sum_{t=0}^{T} E_{s\boldsymbol{f}}[R_t] \right) \qquad (1.3.4)$$

Το μοντέλο που χρησιμοποιεί το παραπάνω ως κριτήριο απόδοσης θα το ονομάσουμε *Μαρκοβιανή Διαδικασία Απόφασης Οριακού Μέσου* και θα το συμβολίζουμε με Γ_a.

Κατ' αναλογία με τις προηγούμενες παραγράφους, ορίζουμε το διάνυσμα τιμής: $\boldsymbol{v}_a(\boldsymbol{f}) \div \left((v_a(1, \boldsymbol{f}), v_a(2, \boldsymbol{f}), \dots, v_a(N, \boldsymbol{f}) \right)^T$

Μπορούμε τώρα να συνδέσουμε το Γ_a με το πρόβλημα βελτιστοποίησης:

$$\max \boldsymbol{v}_a(\boldsymbol{f})$$

με περιορισμούς:

$$\boldsymbol{f} \in F_s$$

Η σχέση (1.3.4) μπορεί να γραφεί σε διανυσματική μορφή:

$$\boldsymbol{v}_a(\boldsymbol{f}) = \lim_{T\to\infty} \left(\frac{1}{T+1} \sum_{t=0}^{T} \boldsymbol{P^t}(\boldsymbol{f}) \boldsymbol{r}(\boldsymbol{f}) \right) = \left(\lim_{T\to\infty} \frac{1}{T+1} \sum_{t=0}^{T} P^t(\boldsymbol{f}) \right) \boldsymbol{r}(\boldsymbol{f}) = \boldsymbol{Q}(\boldsymbol{f}) \boldsymbol{r}(\boldsymbol{f})$$

Η τελευταία ισότητα προκύπτει από γνωστή ιδιότητα των Μαρκοβιανών Αλυσίδων, η οποία εξασφαλίζει την ύπαρξη ενός πίνακα $\boldsymbol{Q}(\boldsymbol{f})$ τέτοιου ώστε:

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

Ο πίνακας $\boldsymbol{Q}(\boldsymbol{f})$ ονομάζεται *όριο κατά Cesaro* πίνακας του $\boldsymbol{P}(\boldsymbol{f})$. Αυτός ο πίνακας θα μας απασχολήσει ιδιαίτερα στο 3ο Κεφάλαιο.

Οπότε για $\boldsymbol{f} \in F_s$ έχουμε ότι:

$$\boldsymbol{v}_a(\boldsymbol{f}) = \boldsymbol{Q}(\boldsymbol{f})\boldsymbol{r}(\boldsymbol{f})$$

Ωστόσο, ένα πρόβλημα που θα μπορούσε να δημιουργηθεί σε αυτή την διαδικασία, αποτελεί η ύπαρξη απορροφητικών καταστάσεων. Είναι πιθανόν, ανάλογα με την επιλογή της στρατηγικής, η διαδικασία να εγκλωβιστεί σε μια τέτοια κατάσταση. Προς αποφυγή αυτού, αρχικά θα περιοριστούμε σε *αδιαχώριστες* Μαρκοβιανές Αλυσίδες, υποθέτοντας ότι για κάθε $\boldsymbol{f} \in F_s$**,** ο πίνακας πιθανοτήτων μετάβασης $\boldsymbol{P}(\boldsymbol{f})$ ορίζει μια *αδιαχώριστη* Μαρκοβιανή Αλυσίδα.

Από μαθηματική σκοπιά, αυτό σημαίνει ότι για οποιεσδήποτε δύο καταστάσεις (s, s') , υπάρχει θετικός ακέραιος t τέτοιος ώστε $[\boldsymbol{P}^t(\boldsymbol{f})]_{s,s'} > 0$. Προφανώς το t εξαρτάται από την επιλογή των s, s'. Ουσιαστικά, μια αδιαχώριστη αλυσίδα συνεπάγεται για μία στρατηγική $\boldsymbol{f}$,ότι η διαδικασία θα επισκεφθεί κάθε κατάσταση του συστήματος άπειρες το πλήθος φορές. Στη συνέχεια παραθέτουμε ένα βασικό λήμμα που αφορά σε αυτή την κατηγορία μαρκοβιανών αλυσίδων.

Λήμμα 1.4.1.

Έστω *P* ο πίνακας πιθανοτήτων μετάβασης μιας αδιαχώριστης ΜΑ και $\boldsymbol{Q}$ το αντίστοιχο *όριο κατά Cesaro* του $\boldsymbol{P}$. Τότε

(i) Οι γραμμές του $\boldsymbol{Q}$ ταυτίζονται

(ii) Έστω $\boldsymbol{q} = (q_1, q_2, \dots, q_N)$ μία γραμμή του $\boldsymbol{Q}$. Τότε κάθε στοιχείο του $\boldsymbol{q}$ θα είναι αυστηρά θετικό και $\boldsymbol{q}$ είναι η μοναδική λύση του συστήματος γραμμικών εξισώσεων:

$$\boldsymbol{x}\, P = \boldsymbol{x}$$

$$\boldsymbol{x}\, \mathbf{1} = 1$$

όπου με $\mathbf{1}$ συμβολίζουμε το N- διάστατο διάνυσμα στήλης που έχει μονάδα σε κάθε στοιχείο του. Το διάνυσμα $\boldsymbol{q}$ καλείται *στάσιμη κατανομή* της αδιαχώριστης ΜΑ.

Ας θεωρήσουμε μια αδιαχώριστη ΜΑ που προσδιορίζεται από κάποια $\boldsymbol{f} \in F_s$ μέσω του πίνακα μεταβάσεων $\boldsymbol{P}(\boldsymbol{f})$ και έστω $\boldsymbol{q}(\boldsymbol{f})$ η στάσιμη κατανομή αυτής, όπως στο παραπάνω Λήμμα. Για κάθε $a \in A(s), s \in S$ ορίζουμε:

$$x_{sa}(\boldsymbol{f}) \doteqdot q_s(f) f(s,a) \qquad (1.4.1)$$

$$x_s(\boldsymbol{f}) \doteqdot \sum_{a \in A(s)} x_{sa}(\boldsymbol{f}) = q_s(f) \quad \left(\text{αφού} \sum_{a \in A(s)} f(s,a) = 1 \right)$$

Ερμηνεύουμε το $x_s(\boldsymbol{f})$ ως τη συχνότητα επίσκεψης της κατάστασης s μακροπρόθεσμα και το $x_{sa}(\boldsymbol{f})$ ως τη συχνότητα εμφάνισης του ζεύγους (s,a) μακροπρόθεσμα. Ακόμη, θα συμβολίζουμε με $\boldsymbol{x}(\boldsymbol{f})$ το διάνυσμα συχνότητας εμφάνισης του ζεύγους (κατάσταση-ενέργεια) μακροπρόθεσμα, του οποίου το s-οστό μπλοκ θα είναι το διάνυσμα στήλη:

$$x_s(\boldsymbol{f}) = \left(x_{s1}(\boldsymbol{f}), x_{s2}(\boldsymbol{f}), \dots, x_{sm(s)}(\boldsymbol{f}) \right)^{\boldsymbol{T}}$$

Τέλος, ορίζουμε το διάνυσμα-γραμμή συχνότητας εμφάνισης κάθε κατάστασης μακροπρόθεσμα:

$$\overline{\boldsymbol{x}}(\boldsymbol{f}) = \left(x_1(\boldsymbol{f}), x_2(\boldsymbol{f}), \dots, x_N(\boldsymbol{f}) \right)$$

Από τη σχέση (1.4.1) μπορούμε να ορίσουμε μια απεικόνιση

$$\begin{cases} M: F_s \to \mathbb{R}^m \\ M(f) \to \boldsymbol{x}(f) \end{cases}, \text{όπου } m \doteqdot \sum_{s=1}^{N} m(s)$$

Από το Λήμμα 1.4.1 έχουμε ότι:

$$\boldsymbol{q}(\boldsymbol{f})\boldsymbol{P}(\boldsymbol{f}) = \boldsymbol{q}(\boldsymbol{f}) \Rightarrow \boldsymbol{q}(\boldsymbol{f})\left(\boldsymbol{I} - \boldsymbol{P}(\boldsymbol{f})\right) = \boldsymbol{0}$$

Η τελευταία σχέση μπορεί να γραφεί σε μη-διανυσματική μορφή ως

$$\sum_{s=1}^{N} \left(\delta_{ss'} - p(s'|s, \boldsymbol{f}) \right) q_s(\boldsymbol{f}) = 0, \quad s, s' \in S$$

Αναλύοντας το $p(s'|s, \boldsymbol{f})$ στις επιμέρους ενέργειες της $\boldsymbol{f}$, παίρνουμε:

$$\sum_{s=1}^{N} \left(\delta_{ss'} - p(s'|s,a) \right) q_s(\boldsymbol{f}) f(s,a) = 0 \Rightarrow \sum_{s=1}^{N} \sum_{a \in A(s)} (\delta_{ss'} - p(s'|s,a)) x_{sa}(\boldsymbol{f}) = 0$$

Επιπλέον, έχουμε ότι

$$\sum_{s=1}^{N} \sum_{a \in A(s)} x_{sa}(\boldsymbol{f}) = \sum_{s=1}^{N} \sum_{a \in A(s)} q_s(\boldsymbol{f}) f(s,a) = \sum_{s=1}^{N} q_s(\boldsymbol{f}) = 1$$

Οπότε τελικά καταλήγουμε στο πολύεδρο σύνολο X που ορίζεται από τους παρακάτω περιορισμούς:

$$\sum_{s=1}^{N} \sum_{a \in A(s)} (\delta_{ss'} - p(s'|s,a)) x_{sa} = 0, \;\; s' \in S$$

$$\sum_{s=1}^{N} \sum_{a \in A(s)} x_{sa} = 1$$

$$x_{sa} \geq 0, \qquad a \in A(s), \qquad s \in S$$

Εναλλακτικά, σε μορφή πινάκων:

$$X = \{\boldsymbol{x} | \boldsymbol{W}\boldsymbol{x} = \boldsymbol{0}, \boldsymbol{1}^T \boldsymbol{x} = \boldsymbol{1}, \boldsymbol{x} \geq \boldsymbol{0}\}$$

Όπου τα $\boldsymbol{x}, \boldsymbol{1}$, είναι m-διάστατα διανύσματα, και $\boldsymbol{W}$ είναι ένας $N \times m$ πίνακας του οποίου το$(s', (s,a))$-οστό στοιχείο είναι:

$$w_{s'}(s,a) \doteqdot \delta_{ss'} - p(s'|s,a)$$

Στο σημείο αυτό θα θέλαμε να δείξουμε ότι το σύνολο X αντιστοιχεί στο 'χώρο συχνοτήτων' του Γ_α , και στη συνέχεια ότι ο χώρος αυτός βρίσκεται σε 1-1 αντιστοιχία με το χώρο των στάσιμων στρατηγικών F_s. Παραθέτουμε, χωρίς απόδειξη, ορισμένα 'τεχνικά' αποτελέσματα.

Λήμμα 1.4.2

Έστω Γ_α μια αδιαχώριστη διαδικασία οριακού μέσου και X το αντίστοιχο πολύεδρο όπως ορίζεται από τα τρία παραπάνω σύνολα περιορισμών. Έστω $\boldsymbol{x} \in X$ και ορίζουμε το $\overline{\boldsymbol{x}} = (x_1, x_2, \dots, x_N)$ ως εξής:

$$x_s = \sum_{a \in A(s)} x_{sa}\,, \quad s \in S$$

Τότε θα ισχύει ότι $\overline{\boldsymbol{x}} > \boldsymbol{0}$ (δηλαδή $x_s > 0$ για κάθε $s \in S$)

Θεώρημα 1.4.1

Έστω Γ_α αδιαχώριστο μοντέλο, το σύνολο X και η συνάρτηση $M: F_s \to \mathbb{R}^m$, όπως ακριβώς έχουν οριστεί. Τότε η M είναι αντιστρέψιμη απεικόνιση από το F_s στο X και η αντίστροφη της είναι η:

$M^{-1}(x) = f_x$, όπου

$$f_x(s,a) \doteqdot \frac{x_{sa}}{x_s}, \qquad a \in A(s), s \in S$$

Το τελευταίο θεώρημα υποδεικνύει πως μπορούμε να μετασχηματίσουμε το πρόβλημα

(I) $\begin{cases} \max v_a(\boldsymbol{f}) \\ \boldsymbol{f} \in F_s \end{cases}$, που ορίστηκε στην αρχή της παραγράφου, στο πρόβλημα

$\begin{cases} \max v_a\left(M^{-1}(\boldsymbol{x})\right) \\ \boldsymbol{x} \in X \end{cases}$ πάνω στο χώρο συχνοτήτων X.

Το επόμενο αποτέλεσμα υποδεικνύει τον τρόπο εύρεσης βέλτιστης λύσης για το (I) από μία βέλτιστη λύση ενός γραμμικού προγράμματος.

Θεώρημα 1.4.2

Έστω Γ_α το αδιαχώριστο μοντέλο, το σύνολο X και η συνάρτηση $M: F_s \to \mathbb{R}^m$ όπως στο Θεώρημα 1.4.1. Έστω $\boldsymbol{x}_0$ μια βέλτιστη λύση στο γραμμικό πρόγραμμα

$$\max \sum_{s=1}^{N} \sum_{a \in A(s)} r(s,a) x_{sa}$$

με περιορισμούς

$$\boldsymbol{W}\boldsymbol{x} = \boldsymbol{0}$$

$$\boldsymbol{1}^T\boldsymbol{x} = \boldsymbol{0}$$

$$\boldsymbol{x} \geq \boldsymbol{0}$$

Τότε, η $f^0 \doteqdot f_{x_0} = M^{-1}(\boldsymbol{x}_0)$ αποτελεί μια βέλτιστη στρατηγική για το αρχικό πρόβλημα βελτιστοποίησης (I).

1.5 Συμπεριφορικές και Μαρκοβιανές στρατηγικές

Όπως μπορεί να αντιληφθεί κανείς, στα προβλήματα που έχουμε έως τώρα δει , δεν είναι αρκετό να περιοριστούμε μόνο στη χρήση καθαρών στρατηγικών , καθ' ότι κρίνεται αναγκαία η ύπαρξη στοχαστικότητας στην πολιτική που ακολουθείται, προκειμένου να φτάσουμε στη βέλτιστη λύση. Σε ακόμη γενικότερο πλαίσιο, θα μπορούσε κανείς να αναρωτηθεί εάν και το σύνολο F_s των στάσιμων στρατηγικών είναι επαρκές ή αρκετά μεγάλο, όσο θα θέλαμε.

Αντικείμενο αυτής της παραγράφου αποτελεί η εισαγωγή δύο νέων συνόλων από στρατηγικές, ευρύτερων από το F_s, καθώς και η σύγκριση της απόδοσης του μοντέλου για στρατηγικές που ανήκουν στα δύο αυτά νέα σύνολα.

Θα συμβολίζουμε με $h_t = (s_0, a_0, s_1, a_1, \dots, a_{t-1}, s_t)$ την *ιστορία* της διαδικασίας μέχρι το χρόνο t, και με H_t *το σύνολο όλων των πιθανών ιστοριών* μέχρι το χρόνο t. Έστω $A = \bigcup_{s \in S} A(s)$ ο χώρος όλων των ενεργειών της διαδικασίας και $P(A)$ το σύνολο όλων των κατανομών πιθανότητας πάνω στο (πεπερασμένο) σύνολο Α.

Ορίζουμε ως *κανόνα απόφασης* στο χρόνο t, τη συνάρτηση:

$f_t: H_t \longrightarrow P(A)$ τέτοια ώστε

$$f_t(h_t, a) = \begin{cases} \mathbb{P}_{f_t}(A_t = a | h_t), \alpha\nu\ a \in A(s_t) \\ 0, \quad \alpha\nu\ a \notin A(s_t) \end{cases}$$

Τότε, ορίζουμε ως *συμπεριφορική στρατηγική* $\boldsymbol{\pi}$ να είναι η ακολουθία από κανόνες απόφασης:

$$\boldsymbol{\pi} \div (f_0, f_1, f_2, \dots)$$

Θα συμβολίζουμε με F_B την κλάση όλων των συμπεριφορικών στρατηγικών. Μια συμπεριφορική στρατηγική $\boldsymbol{\pi}$ θα καλείται *Μαρκοβιανή* εάν κάθε κανόνας απόφασης f_t εξαρτάται μόνο από την παρούσα κατάσταση, για κάθε $t = 0,1,2, \dots$, δηλαδή:

$$f_t(s_t, a) = \mathbb{P}_{f_t}(A_t = a | S_t = s_t) = \mathbb{P}_{f_t}(A_t = a | S_0 = s_0, A_o = a_o, \dots, A_{t-1} = a_{t-1}, S_t = s_t) \\ = f_t(h_t, a)$$

για όλες τις ιστορίες $h_t = (s_0, a_0, s_1, a_1, \dots, a_{t-1}, s_t) \in H_t$

Η κλάση των *Μαρκοβιανών* στρατηγικών συμβολίζεται με F_M.

Η κλάση F_M είναι μεγαλύτερη από την κλάση F_S των στάσιμων στρατηγικών. Πράγματι, μια στάσιμη στρατηγική είναι μια Μαρκοβιανή στρατηγική για την οποία όλοι οι κανόνες απόφασης είναι ανεξάρτητοι από το χρόνο t, δηλαδή $f_t = f$ για κάθε t, όπου:

$$f(s, a) = \mathbb{P}_f((A_t = a | S_t = s)\ , \text{για κάθε } t = 0,1,2, \dots$$

Υπενθυμίζουμε ότι μία καθαρή στάσιμη στρατηγική $\boldsymbol{f}$ είναι μία στάσιμη στρατηγική για την οποία, για κάθε $s \in S$ υπάρχει $a_s \in A(s)$ τέτοιο ώστε $f(s, a) = 0$ για κάθε $a \neq a_s$. Το σύνολο των καθαρών στάσιμων στρατηγικών θα συμβολίζεται με F_D. Φυσικά αυτό το σύνολο είναι ένα πεπερασμένο σύνολο, με πληθικό αριθμό:

$$|F_D| = \prod_{s=1}^{N} m(s)$$

Από τον τρόπο που κατασκευάστηκαν όλες οι παραπάνω κλάσεις, γίνεται φανερή η σχέση εγκλεισμού που τις συνδέει. Ισχύει δηλαδή ότι:

$$F_B \supseteq F_M \supseteq F_S \supseteq F_D$$

Σχόλιο

Παρατηρούμε ότι, ανεξάρτητα από την κλάση στρατηγικών, το γεγονός ότι ο χώρος καταστάσεων και ενεργειών είναι πεπερασμένα σύνολα, μας εξασφαλίζει ότι οι αναμενόμενες αμοιβές για κάθε t είναι καλά ορισμένες και ικανοποιούν την :

$$E_{s_0\boldsymbol{\pi}}[R_{\boldsymbol{t}}] = \sum_{s=1}^{\boldsymbol{N}} \sum_{a\in A(s)} \mathbb{P}_{\boldsymbol{\pi}}(S_t = s, A_t = a|S_0 = s_0) r(s,a) \quad , \boldsymbol{\pi} \in F_B, s_0 \in S$$

Ο ορισμός για την τιμή μιας στρατηγικής με συντελεστή προεξόφλησης, καθώς και για την τιμή του τερματιζόμενου μοντέλου, επεκτείνονται με φυσικό τρόπο στο σύνολο στρατηγικών F_B.

Για την τιμή στο μοντέλο οριακού μέσου, ωστόσο, θα χρειαστεί να τροποποιήσουμε ελαφρώς τον ορισμό, κατά τον εξής τρόπο:

Για κάθε $\boldsymbol{\pi} \in F_B$ και αρχική κατάσταση s, ορίζουμε:

$$v_a(s, \boldsymbol{\pi}) \doteqdot \liminf_{T\to\infty} \left(\frac{1}{T+1} \sum_{t=0}^{T} E_{s\boldsymbol{\pi}}[R_{\boldsymbol{t}}] \right)$$

Το επόμενο Θεώρημα μας λέει ότι μπορούμε να 'προσαρμόσουμε' σε κάθε συμπεριφορική στρατηγική μία Μαρκοβιανή στρατηγική

Θεώρημα 1.5.1

Έστω $\boldsymbol{\pi} \in F_B$ μια αυθαίρετη συμπεριφορική στρατηγική. Τότε, για κάθε αρχική κατάσταση $s_0 \in S$, υπάρχει Μαρκοβιανή στρατηγική $\bar{\boldsymbol{\pi}} \in F_M$ τέτοια ώστε, για $a \in A, s \in S, t = 0,1,2, \ldots,$

$$\mathbb{P}_{\boldsymbol{\pi}}(S_t = s, A_t = a|S_0 = s_0) = \mathbb{P}_{\bar{\boldsymbol{\pi}}}(S_t = s, A_t = a|S_0 = s_0) \qquad (1.5.1)$$

(Εν γένει, η $\bar{\boldsymbol{\pi}}$ εξαρτάται από την αρχική κατάσταση)

Απόδειξη

Παρατηρούμε ότι η ιστορία μέχρι τον χρόνο $t = 0$ είναι απλά $h_0 = (s_0)$.

Εάν $\boldsymbol{\pi} = (f_0, f_1, \ldots, f_t, \ldots)$ και ορίσουμε την $\bar{\boldsymbol{\pi}} \doteqdot (\bar{f}_0, \bar{f}_1, \ldots, \bar{f}_t, \ldots)$, όπου $\bar{f}_0 \doteqdot f_0$, τότε προφανώς η σχέση $(1.5.1)$ ισχύει για $t = 0$.

Θα δείξουμε ότι αν ορίσουμε την $\bar{f}_t$ ως:

$$\bar{f}_t(s_t, a) = \mathbb{P}_{\boldsymbol{\pi}}[A_t = a|\, S_0 = s_0, S_t = s_t] \qquad (1.5.2)$$

για όλα τα $a \in A(s_t),\ \ s_t, s_0 \in S, t = 1,2,3, \ldots$

τότε η σχέση (1.5.1) ισχύει για όλους τους μη-αρνητικούς ακεραίους t, με επαγωγή. Πράγματι, είδαμε ότι ισχύει για $t = 0$ και υποθέτουμε ότι ισχύει για όλα τα $t = 0, 1, 2, \dots, T-1$.

Το αριστερό μέλος της (1.5.1) για $t = T$ γράφεται ως εξής:

$$\mathbb{P}_{\boldsymbol{\pi}}(S_T = s_T, A_T = a | S_0 = s_0) = \sum_{s=1}^{N} \sum_{a \in A(s)} \mathbb{P}_{\boldsymbol{\pi}} \left[(S_{T-1} = s, A_{T-1} = a | S_0 = s_0)\right] p(s_T | s, a)$$

$$= \sum_{s=1}^{N} \sum_{a \in A(s)} \mathbb{P}_{\bar{\boldsymbol{\pi}}} \left[(S_{T-1} = s, A_{T-1} = a | S_0 = s_0)\right] p(s_T | s, a)$$

$$= \mathbb{P}_{\bar{\boldsymbol{\pi}}}[S_T = s_T | S_0 = s_0]$$

για όλα τα s_T, s_0. Συνεπώς, η (1.5.1) ισχύει και για την περίπτωση $t = T$, το οποίο ολοκληρώνει την απόδειξη.

■

Πόρισμα 1.5.1.

Έστω $s \in S$ και $\boldsymbol{\pi} \in F_B$ μια αυθαίρετη συμπεριφορική στρατηγική. Θεωρούμε την $\bar{\boldsymbol{\pi}}$ η οποία κατασκευάζεται από την $\boldsymbol{\pi}$ όπως στη σχέση (1.5.2). Έστω $\Gamma_B, \Gamma_T, \Gamma_a$ το μοντέλο με προεξόφληση, το τερματιζόμενο και το μοντέλο οριακού μέσου, αντίστοιχα, όπως αυτά έχουν οριστεί.

Τότε, χωρίς περιορισμό της γενικότητας, μπορούμε να περιοριστούμε στο σύνολο στρατηγικών F_M, γιατί:

(i) $v_\beta(s, \boldsymbol{\pi}) = v_\beta(s, \bar{\boldsymbol{\pi}}),\ v_T(s, \boldsymbol{\pi}) = v_T(s, \bar{\boldsymbol{\pi}}),\ v_a(s, \boldsymbol{\pi}) = v_a(s, \bar{\boldsymbol{\pi}})$

(ii) $\sup_{F_B} v_\beta(s, \boldsymbol{\pi}) = \sup_{F_M} v_\beta(s, \boldsymbol{\pi}), \sup_{F_B} v_T(s, \boldsymbol{\pi}) = \sup_{F_M} v_T(s, \boldsymbol{\pi}), \sup_{F_B} v_a(s, \boldsymbol{\pi}) = \sup_{F_M} v_a(s, \boldsymbol{\pi})$

Απόδειξη

Αρχικά παρατηρούμε ότι οποιοδήποτε από τα τρία κριτήρια απόδοσης χρησιμοποιεί την ακολουθία αναμενόμενων αμοιβών $\left\{E_{s,\boldsymbol{\pi}}[R_t]\right\}_{t=0}^{\infty},\ \boldsymbol{\pi} \in F_B, s \in S$

Από το προηγούμενο θεώρημα έχουμε:

$$E_{s_0,\boldsymbol{\pi}}[R_t] = \sum_{s=1}^{N} \sum_{a \in A(s)} \mathbb{P}_{\boldsymbol{\pi}} \left[(S_t = s, A_t = a | S_0 = s_0)\right] r(s, a)$$

$$= \sum_{s=1}^{N} \sum_{a \in A(s)} \mathbb{P}_{\bar{\boldsymbol{\pi}}} \left[(S_t = s, A_t = a | S_0 = s_0)\right] r(s, a)$$

$$= E_{s_0,\bar{\boldsymbol{\pi}}}[R_t]$$

για όλα τα $s_0 \in S$. Οπότε τα (i),(ii) έπονται άμεσα από τους ορισμούς των αντίστοιχων κριτηρίων απόδοσης και την τελευταία ισότητα.

■

Σχόλιο

Συνοψίζοντας, το Θεώρημα 1.5.1 και το Πόρισμα 1.5.1 μας επιτρέπουν να περιοριστούμε στην κλάση των Μαρκοβιανών στρατηγικών, χωρίς αυτό να περιορίζει την γενικότητα.

Κεφάλαιο 2

Στοχαστικά παίγνια και γραμμικός προγραμματισμός

2.0 Εισαγωγή

Τα στοχαστικά παιχνίδια αποτελούν γενίκευση των Μαρκοβιανών διαδικασιών απόφασης. Πλέον, η εξέλιξη της διαδικασίας ελέγχεται από δύο ή περισσότερους ελεγκτές, οι οποίοι θα καλούνται παίκτες. Στα πλαίσια της εργασίας αυτής, θα μας απασχολήσουν μόνον παίγνια δύο παικτών και μηδενικού αθροίσματος, το οποίο πολύ απλά ερμηνεύεται ως "το κέρδος του ενός εκ των δύο παικτών θα ισούται με τη ζημία του δεύτερου παίκτη". Τα παιχνίδια που θα μελετήσουμε είναι μη-συνεργατικά (ανταγωνιστικά), δηλαδή ο κάθε παίκτης επιδιώκει να παίξει με τέτοιο τρόπο ώστε να πετύχει την μεγιστοποίηση του κέρδους του ή την ελαχιστοποίηση της ζημίας του. Φυσικά, όπως και στο προηγούμενο κεφάλαιο, θα μας απασχολήσουν περιπτώσεις όπου τα σύνολα καταστάσεων και ενεργειών είναι πεπερασμένα. Και πάλι το ενδιαφέρον μας θα στραφεί σε συγκεκριμένες κατηγορίες παιγνίων, τα οποία διαφέρουν ως προς τη δομή ή/και το κριτήριο πληρωμής των παικτών. Ο περιορισμός αυτός κρίνεται απαραίτητος , εφόσον δεν είναι ακόμα γνωστό εάν υπάρχει αναλυτική λύση για οποιοδήποτε γενικό Στοχαστικό παίγνιο, πόσο μάλλον όταν περιοριζόμαστε στα παίγνια εκείνα που επιδέχονται λύση με μεθόδους Γραμμικού Προγραμματισμού.

Αξιοσημείωτο είναι το γεγονός πως η θεωρία περί Στοχαστικών Παιγνίων αναπτύχθηκε ανεξάρτητα και, μάλιστα, προγενέστερα της θεωρίας των Μαρκοβιανών Διαδικασιών απόφασης ως ένα εντελώς ξεχωριστό αντικείμενο, προτού βρεθεί η σύνδεση μεταξύ των δύο κλάδων.

2.1 Στοχαστικά παίγνια με συντελεστή προεξόφλησης

Στην παράγραφο αυτή θα γενικεύσουμε το Μαρκοβιανό μοντέλο απόφασης Γ_β με συντελεστή προεξόφλησης β, όπως αυτό ορίστηκε στην παράγραφο 1.1, για την περίπτωση που έχουμε 2 παίκτες-ελεγκτές.

Θεωρούμε πως, αν η διαδικασία βρίσκεται στην κατάσταση $s \in S = \{1,2,\dots,N\}$ στον χρόνο t, οι παίκτες επιλέγουν *ανεξάρτητα* ο ένας από τον άλλον ενέργειες $a^1 \in A^1(s)$ και $a^2 \in A^2(s)$ και λαμβάνουν αμοιβές $r^1(s,a^1,a^2)$ και $r^2(s,a^1,a^2)$, αντίστοιχα.

Επιπλέον, οι πιθανότητες μετάβασης γενικεύονται έτσι, ώστε:

$$p(s'|s,a^1,a^2) \doteqdot \mathbb{P}\{S_{t+1}=s'|S_t=s,A_t^1=a_1,A_t^2=a_2\} \quad ,\text{για κάθε } t=0,1,2,\dots$$

Σε αναλογία με τον συμβολισμό του προηγούμενου κεφαλαίου, θα συμβολίζουμε με F_s το σύνολο των στάσιμων στρατηγικών του παίκτη 1 και με G_s το αντίστοιχο σύνολο για τον παίκτη

2. Δηλαδή, αν $\boldsymbol{g} = (g(1), g(2), \dots, g(N)) \in G_s$, τότε κάθε $g(s)$ είναι ένα $m^2(s)$-διάστατο διάνυσμα πιθανοτήτων, όπου $m^2(s) = |A^2(s)|$.

Ακόμη, θα συμβολίζουμε με R_t^1 (R_t^2) τη συνολική αμοιβή του παίκτη 1 (παίκτη 2) στο χρόνο t, ενώ με $r^1(\boldsymbol{f}, \boldsymbol{g})$ ($r^{\boldsymbol{2}}(\boldsymbol{f}, \boldsymbol{g})$) την αναμενόμενη αμοιβή ενός βήματος του παίκτη 1 (παίκτη 2) που αντιστοιχεί στο ζεύγος στρατηγικών $(\boldsymbol{f}, \boldsymbol{g}) \in F_{\boldsymbol{s}} \times G_{\boldsymbol{s}}$. Η αναμενόμενη αμοιβή του παίκτη k στο στάδιο t για τις στρατηγικές $(\boldsymbol{f}, \boldsymbol{g})$ και αρχική κατάσταση s, θα συμβολίζεται με

$E_{s\,\boldsymbol{fg}}[R_t^k]$. Συνεπώς, η συνολική τιμή με προεξόφληση για ένα ζεύγος στρατηγικών $(\boldsymbol{f}, \boldsymbol{g}) \in F_{\boldsymbol{s}} \times G_{\boldsymbol{s}}$ για τον παίκτη k ($k = 1$ ή $k = 2$) θα δίνεται από την:

$$v_\beta^k(s, \boldsymbol{f}, \boldsymbol{g}) \doteqdot \sum_{t=0}^{\infty} \beta^t \, E_{s\,\boldsymbol{fg}}[R_t^k], \quad \text{όπου } b \in [0,1) \text{ και } k \in \{1,2\}$$

Θα θέλαμε στο σημείο αυτό να βρούμε, με κάποιο τρόπο, ένα ζευγάρι στάσιμων στρατηγικών $(\boldsymbol{f}, \boldsymbol{g})$, το οποίο να αποτελεί 'λύση' του παιχνιδιού. Στο αντίστοιχο σημείο του προηγούμενου κεφαλαίου είχαμε ορίσει ως λύση για τον παίκτη 1 την βελτιστοποίηση του προβλήματος:

$$\begin{cases} max\ v_\beta^1(\boldsymbol{f}, \boldsymbol{g}) \\ \quad \boldsymbol{f} \in F_s \end{cases}$$

Προφανώς μια τέτοια αντιμετώπιση θα κρίνονταν ανεπαρκής στην περίπτωση των δύο παικτών, εφόσον η λύση του παιχνιδιού ,εν γένει, εξαρτάται και από την στρατηγική $\boldsymbol{g} \in G_s$ του 2ου παίκτη.

Η κλασική υπόθεση που κάνουμε στα μη-συνεργατικά παίγνια τα οποία θα μας απασχολήσουν, έγκειται στο ότι οι παίκτες επιλέγουν ανεξάρτητα (και κρυφά) τις ενέργειές τους, καθώς και ότι μοναδικός τους στόχος είναι η μεγιστοποίηση της συνολικής προσωπικής τους αμοιβής. Εάν ακόμη υποθέσουμε ότι οι παίκτες γνωρίζουν επακριβώς το σύνολο στρατηγικών και τη συνάρτηση πληρωμής των υπόλοιπων παικτών, τότε η πλέον διαδεδομένη λύση του παιχνιδιού που ψάχνουμε είναι γνωστή στη βιβλιογραφία ως *σημείο Nash* ή *σημείο Ισορροπίας κατά Nash (Nash Equilibrium).*

Ορισμός 2.1.1.

Θα λέμε ότι το ζεύγος στρατηγικών $(\boldsymbol{f}^0, \boldsymbol{g}^0) \in F_{\boldsymbol{s}} \times G_{\boldsymbol{s}}$ αποτελεί ένα *σημείο Ισορροπίας κατά Nash* για το στοχαστικό παίγνιο Γ_β με συντελεστή προεξόφλησης, εάν:

$$\boldsymbol{v}_\beta^1(\boldsymbol{f}, \boldsymbol{g}^0) \leq \boldsymbol{v}_\beta^1(\boldsymbol{f}^0, \boldsymbol{g}^0) \quad \text{, για όλα τα } \boldsymbol{f} \in F_{\boldsymbol{s}}$$

και

$$\boldsymbol{v}_\beta^2(\boldsymbol{f}^0, \boldsymbol{g}) \leq \boldsymbol{v}_\beta^2((\boldsymbol{f}^0, \boldsymbol{g}^0) \quad \text{, για όλα τα } \boldsymbol{g} \in G_{\boldsymbol{s}}$$

Από τον παραπάνω ορισμό μπορούμε να δούμε πως η καταλληλότητα του σημείου Nash ως λύση του παιχνιδιού, έγκειται στο γεγονός πως κανένας παίκτης δεν ωφελείται εάν επιχειρήσει να παρεκκλίνει από την στρατηγική του. Με άλλα λόγια, εάν $(\boldsymbol{f}^0, \boldsymbol{g}^0)$ είναι ένα σημείο Nash και ο παίκτης 1 γνωρίζει πως ο παίκτης 2 θα παίξει την στρατηγική $\boldsymbol{g}^0$, τότε το καλύτερο που μπορεί να κάνει ο παίκτης 1 για να μεγιστοποιήσει το κέρδος του, είναι να επιλέξει την στρατηγική $\boldsymbol{f}^0$.

Ωστόσο, ένα μεγάλο πρόβλημα που ενδέχεται να συναντήσουμε σε πολλά παίγνια, είναι η ύπαρξη πολλαπλών σημείων Nash με διαφορετικές πληρωμές για τους παίκτες.

Μία ευρεία κατηγορία παιχνιδιών στην οποία θα εστιάσουμε για το υπόλοιπο της εργασίας, αποτελούν τα παιχνίδια *μηδενικού αθροίσματος*. Θα δούμε, μάλιστα, πως για αυτή την κλάση παιγνίων, το πρόβλημα της ύπαρξης πολλαπλών σημείων Nash δεν υφίσταται.

Ένα στοχαστικό παίγνιο θα καλείται *μηδενικού-αθροίσματος* εάν:

$r^1(s, a^1, a^2) + r^2(s, a^1, a^2) = 0$, για κάθε $s \in S, a^1 \in A^1(s), a^2 \in A^2(s)$

Οπότε μπορούμε να απλοποιήσουμε ελαφρώς τον συμβολισμό και να έχουμε:

$$\begin{cases} r(s, a^1, a^2) \doteqdot r^1(s, a^1, a^2) = -r^2(s, a^1, a^2)\,, & s \in S, a^1 \in A^1(s), a^2 \in A^2(s) \\ \boldsymbol{v}_\beta(\boldsymbol{f}, \boldsymbol{g}) \doteqdot \boldsymbol{v}_\beta^1(\boldsymbol{f}, \boldsymbol{g}) = -\boldsymbol{v}_\beta^2(\boldsymbol{f}, \boldsymbol{g})\,, & (\boldsymbol{f}, \boldsymbol{g}) \in F_s \times G_s \end{cases}$$

Έστω ότι $(\boldsymbol{f}^0, \boldsymbol{g}^0)$ είναι ένα σημείο Nash. Τότε, σύμφωνα με τα παραπάνω, οι δύο ανισότητες του ορισμού του σημείου Nash ανάγονται στην :

$$\boldsymbol{v}_\beta(\boldsymbol{f}, \boldsymbol{g}^0) \leq \boldsymbol{v}_\beta(\boldsymbol{f}^0, \boldsymbol{g}^0) \leq \boldsymbol{v}_\beta(\boldsymbol{f^0}, \boldsymbol{g}), \quad \boldsymbol{f} \in F_s, \boldsymbol{g} \in G_s \qquad (2.1.1)$$

Θα καλούμε την στρατηγική $\boldsymbol{f^0}$ βέλτιστη στάσιμη στρατηγική για τον παίκτη 1 (Αντίστοιχα για τον παίκτη 2).

Στη συνέχεια παραθέτουμε ένα ιδιαίτερα χρήσιμο θεώρημα το οποίο θα απαντήσει στο πρόβλημα περί ύπαρξης πολλαπλών σημείων Nash σε ένα παίγνιο με 2 παίκτες.

Θεώρημα 2.1.1.

Έστω ότι η ανισότητα (2.1.1) ικανοποιείται για τα ζεύγη στρατηγικών $(\boldsymbol{f}^0, \boldsymbol{g}^0)$ και $(\boldsymbol{f}', \boldsymbol{g}')$ που ανήκουν στο $F_s \times G_s$. Τότε:

(i) $\boldsymbol{v}_\beta = \boldsymbol{v}_\beta(\boldsymbol{f}^0, \boldsymbol{g}^0) = \boldsymbol{v}_\beta(\boldsymbol{f}', \boldsymbol{g}') = \boldsymbol{v}_\beta(\boldsymbol{f}^0, \boldsymbol{g}') = \boldsymbol{v}_\beta(\boldsymbol{f}', \boldsymbol{g}^0)$

(ii) $\boldsymbol{v}_\beta(s, \boldsymbol{f}^0, \boldsymbol{g}^0) = \max\limits_{F_s} \min\limits_{G_s} \boldsymbol{v}_\beta(s, \boldsymbol{f}, \boldsymbol{g}) = \min\limits_{G_s} \max\limits_{F_s} \boldsymbol{v}_\beta(s, \boldsymbol{f}, \boldsymbol{g})\,,$ για κάθε $s \in S$

(iii) Αντίστροφα, εάν υπάρχουν τα $\max_{F_s}\min_{G_s}\boldsymbol{v}_\beta(s,\boldsymbol{f},\boldsymbol{g})$, $\min_{G_s}\max_{F_s}\boldsymbol{v}_\beta(s,\boldsymbol{f},\boldsymbol{g})$ και είναι ίσα για κάποιο $s_0 \in S$, τότε υπάρχουν στάσιμες στρατηγικές $\boldsymbol{f}^0, \boldsymbol{g}^0$ οι οποίες ικανοποιούν την (2.1), για την κατάσταση s_0.

Απόδειξη

(i) Υποθέτουμε ότι τα ζεύγη $(\boldsymbol{f}^0, \boldsymbol{g}^0)$ και $(\boldsymbol{f}', \boldsymbol{g}')$ ικανοποιούν την (2.1.1). Αυτό σημαίνει πως:

$$\boldsymbol{v}_\beta(\boldsymbol{f}, \boldsymbol{g}^0) \le \boldsymbol{v}_\beta(\boldsymbol{f}^0, \boldsymbol{g}^0) \le \boldsymbol{v}_\beta(\boldsymbol{f^0}, \boldsymbol{g}), \qquad \text{για όλα τα } \boldsymbol{f} \in F_s, \boldsymbol{g} \in G_s$$

Οπότε θα πρέπει να ισχύει και ότι:

$$\boldsymbol{v}_\beta(\boldsymbol{f}', \boldsymbol{g}^0) \le \boldsymbol{v}_\beta(\boldsymbol{f}^0, \boldsymbol{g}^0) \le \boldsymbol{v}_\beta(\boldsymbol{f^0}, \boldsymbol{g}') \qquad (2.1.2)$$

Από την άλλη μεριά, για το ζεύγος $(\boldsymbol{f}', \boldsymbol{g}')$ παίρνουμε ότι:

$$\boldsymbol{v}_\beta(\boldsymbol{f}, \boldsymbol{g}') \le \boldsymbol{v}_\beta(\boldsymbol{f}', \boldsymbol{g}') \le \boldsymbol{v}_\beta(\boldsymbol{f}', \boldsymbol{g}), \qquad \text{για όλα τα } \boldsymbol{f} \in F_s, \boldsymbol{g} \in G_s$$

Οπότε θα πρέπει να ισχύει και ότι:

$$\boldsymbol{v}_\beta(\boldsymbol{f}^0, \boldsymbol{g}') \le \boldsymbol{v}_\beta(\boldsymbol{f}', \boldsymbol{g}') \le \boldsymbol{v}_\beta(\boldsymbol{f}', \boldsymbol{g}^0) \qquad (2.1.3)$$

Συνδυάζοντας τις (2.1.2) και (2.1.3) λαμβάνουμε το ζητούμενο αποτέλεσμα.

(ii) Από την (2.1.1) γνωρίζουμε ότι :

$$\max_{\boldsymbol{f}\in F_s}\boldsymbol{v}_\beta(s,\boldsymbol{f},\boldsymbol{g}^0) = \boldsymbol{v}_\beta(s,\boldsymbol{f}^0,\boldsymbol{g}^0) = \min_{\boldsymbol{g}\in G_s}(s,\boldsymbol{f}^0,\boldsymbol{g})$$

Επιπλέον:

$$\max_{\boldsymbol{f}\in F_s}\min_{\boldsymbol{g}\in G_s}\boldsymbol{v}_\beta(s,\boldsymbol{f},\boldsymbol{g}) \ge \min_{\boldsymbol{g}\in G_s}\boldsymbol{v}_\beta(s,\boldsymbol{f}^*,\boldsymbol{g})$$

και

$$\min_{\boldsymbol{g}\in G_s}\max_{\boldsymbol{f}\in F_s}\boldsymbol{v}_\beta(s,\boldsymbol{f},\boldsymbol{g}) \le \max_{\boldsymbol{f}\in F_s}\boldsymbol{v}_\beta(s,\boldsymbol{f},\boldsymbol{g}^*)$$

για όλα τα $\boldsymbol{f}^* \in F_s$ και $\boldsymbol{g}^* \in G_s$.

Οπότε παίρνουμε ότι:

$$\max_{\boldsymbol{f}\in F_s}\min_{\boldsymbol{g}\in G_s}\boldsymbol{v}_\beta(s,\boldsymbol{f},\boldsymbol{g}) \ge \min_{\boldsymbol{g}\in G_s}\boldsymbol{v}_\beta(s,\boldsymbol{f}^0,\boldsymbol{g}) = \max_{\boldsymbol{f}\in F_s}\boldsymbol{v}_\beta(s,\boldsymbol{f},\boldsymbol{g}^0) \ge \min_{\boldsymbol{g}\in G_s}\max_{\boldsymbol{f}\in F_s}\boldsymbol{v}_\beta(s,\boldsymbol{f},\boldsymbol{g})$$

και:

$$\min_{\boldsymbol{g}\in G_s}\max_{\boldsymbol{f}\in F_s}\boldsymbol{v}_\beta(s,\boldsymbol{f},\boldsymbol{g}) \ge \min_{\boldsymbol{g}\in G_s}\boldsymbol{v}_\beta(s,\boldsymbol{f}^0,\boldsymbol{g}) = \max_{\boldsymbol{f}\in F_s}\boldsymbol{v}_\beta(s,\boldsymbol{f},\boldsymbol{g}^0) \ge \max_{\boldsymbol{f}\in F_s}\min_{\boldsymbol{g}\in G_s}\boldsymbol{v}_\beta(s,\boldsymbol{f},\boldsymbol{g})$$

Οι δύο παραπάνω ανισοτικές σχέσεις συνεπάγουν την ζητούμενη ισότητα.

(iii) Ορίζουμε τις ακόλουθες ποσότητες:

$$\boldsymbol{v}_\beta \doteqdot \boldsymbol{v}_\beta(s^0, \boldsymbol{f}^0, \boldsymbol{g}^0), \qquad F_\beta(\boldsymbol{f}) \doteqdot \min_{\boldsymbol{g}\in G_s} \boldsymbol{v}_\beta(s^0, \boldsymbol{f}, \boldsymbol{g}), \qquad G_\beta(\boldsymbol{g}) \doteqdot \max_{\boldsymbol{f}\in F_s} \boldsymbol{v}_\beta(s^0, \boldsymbol{f}, \boldsymbol{g})$$

όπου $\boldsymbol{f} \in F_s, \boldsymbol{g} \in G_s$

Από την υπόθεσή μας γνωρίζουμε ότι θα υπάρχει ένα $\boldsymbol{f}^0 \in F_s$, τέτοιο ώστε $F_\beta(\boldsymbol{f}^0) = \boldsymbol{v}_\beta$

Αυτό σημαίνει ότι:

$$\boldsymbol{v}_\beta = \min_{\boldsymbol{g}\in G_s} \boldsymbol{v}_\beta(s^0, \boldsymbol{f}^0, \boldsymbol{g}) \le \boldsymbol{v}_\beta(s^0, \boldsymbol{f}^0, \boldsymbol{g}), \qquad \text{για όλα τα } \boldsymbol{g} \in G_s$$

Με ανάλογο επιχείρημα, γνωρίζουμε ότι θα υπάρχει $\boldsymbol{g}^0 \in G_s$ ώστε:

$$\boldsymbol{v}_\beta = G_\beta(\boldsymbol{g}^0) = \max_{\boldsymbol{f}\in F_s} \boldsymbol{v}_\beta(s^0, \boldsymbol{f}, \boldsymbol{g}^0) \ge \boldsymbol{v}_\beta(s^0, \boldsymbol{f}, \boldsymbol{g}^0), \qquad \text{για όλα τα } \boldsymbol{f} \in F_s$$

Οπότε το ζεύγος στρατηγικών $(\boldsymbol{f}^0, \boldsymbol{g}^0)$ αποτελεί ένα ζεύγος στάσιμων βέλτιστων στρατηγικών.

■

Σχόλιο

Από το προηγούμενο θεώρημα συμπεραίνουμε πως τα διανύσματα προεξοφλητικής τιμής ταυτίζονται για όλα τα ζεύγη βέλτιστων στρατηγικών. Αυτό το διάνυσμα θα καλείται *διάνυσμα τιμής του παιγνίου μηδενικού αθροίσματος* Γ_β, και θα συμβολίζεται με :

$$\boldsymbol{v}_\beta = (v_\beta(1), v_\beta(2), \dots, v_\beta(N))^T$$

Έως τώρα έχουμε δει τα στοχαστικά παίγνια με προεξόφληση από την σκοπιά της θεωρίας αποφάσεων. Ορίστηκε, δηλαδή, το Γ_β ως μια γενίκευση της Μαρκοβιανής διαδικασίας απόφασης για περισσότερους από έναν ελεγκτές-παίκτες, και συγκεκριμένα για την περίπτωση των δύο παικτών. Μια εναλλακτική σκοπιά, από την οποία θα μπορούσαμε να κοιτάξουμε τα εν λόγω παίγνια, αποτελεί εκείνη των *στατικών πινακοπαιχνιδιών* (matrix games). Μπορούμε, με αυτό το σκεπτικό, να ορίσουμε το παίγνιο Γ_β ως γενίκευση αυτών των πινακοπαιχνιδιών σε πολλαπλά στάδια.

Έστω $m^1(s)$ η πληθικότητα του $A^1(s)$ (αντίστοιχα $m^2(s)$ η πληθικότητα του $A^2(s)$) , με $s \in S$. Τότε, ορίζουμε N στο πλήθος πινακοπαιχνίδια ως εξής:

$$R(s) = [r(s, a^1, a^2)]_{a^1=1, a^2=1}^{m^1(s), m^2(s)}$$

τα οποία βρίσκονται σε 1-1 αντιστοιχία με τις καταστάσεις του Γ_β.

Ένα πινακοπαιχνίδι $R(s)$ προσδιορίζεται πλήρως από έναν πίνακα $\boldsymbol{M}$ διάστασης $m^1(s) \times m^2(s)$, όπου το m_{ij} στοιχείο του πίνακα είναι η αμοιβή που θα λάβει ο παίκτης 1 εάν αυτός επιλέξει την ενέργεια i και ο παίκτης 2 επιλέξει την ενέργεια j. Φυσικά, εφόσον αναφερόμαστε σε παίγνια δύο παικτών και μηδενικού αθροίσματος, από τον πίνακα $\boldsymbol{M}$ είναι γνωστές και οι αμοιβές του δεύτερου παίκτη. Η i-γραμμή του πίνακα $\boldsymbol{M}$ αντιπροσωπεύει την i-ενέργεια του παίκτη 1, ενώ η j-στήλη αντιπροσωπεύει την j-ενέργεια του παίκτη 2. Τα πινακοπαιχνίδια αποτελούν την πιο απλή κατηγορία παιχνιδιών και ιστορικά αποτέλεσαν τα πρώτα παιχνίδια που μελετήθηκαν. Για παράδειγμα, το «Πέτρα-Ψαλίδι-Χαρτί» αποτελεί ένα πινακοπαιχνίδι, όπου ο πίνακας $\boldsymbol{M}$ είναι διάστασης 3×3.

Θα λέμε ότι ένα πινακοπαιχνίδι δύο παικτών και μηδενικού αθροίσματος έχει τιμή $\boldsymbol{v}$, αν και μόνο αν

$$\sup_{\boldsymbol{f}} \inf_{\boldsymbol{g}} \boldsymbol{r}(\boldsymbol{f}, \boldsymbol{g}) = \boldsymbol{v} = \inf_{\boldsymbol{g}} \sup_{\boldsymbol{f}} \boldsymbol{r}(\boldsymbol{f}, \boldsymbol{g})$$

Για ένα τέτοιο παιχνίδι με τιμή $\boldsymbol{v}$ οι στρατηγικές $\boldsymbol{f}_\varepsilon, \boldsymbol{g}_\varepsilon$ θα καλούνται ε-βέλτιστες , $\varepsilon \geq 0$, αν $\inf_{\boldsymbol{g}} \boldsymbol{r}(\boldsymbol{f}_\varepsilon, \boldsymbol{g}) \geq \boldsymbol{v} - \varepsilon$ και $\sup_{\boldsymbol{f}} \boldsymbol{r}(\boldsymbol{f}, \boldsymbol{g}_\varepsilon) \leq \boldsymbol{v} - \varepsilon$. Οι $0-$ βέλτιστες στρατηγικές θα καλούνται απλά βέλτιστες.

Η ακόλουθη πρόταση που αφορά στα πινακοπαιχνίδια και της βέλτιστες στρατηγικές οφείλεται στον J. von Neumann (1928).

Πρόταση 2.1.1.

Για όλους τους πίνακες $\boldsymbol{A}$, το πινακοπαιχνίδι $[\boldsymbol{A}]$ έχει τιμή και οι παίκτες διαθέτουν βέλτιστες στρατηγικές.

Φυσικά, οι στρατηγικές δεν θα είναι υποχρεωτικά ντετερμινιστικές, αλλά ,εν γένει, μεικτές.

Μπορούμε να ερμηνεύσουμε το ‘παίξιμο’ ενός μόνο σταδίου, στην κατάσταση s, σαν οι παίκτες να παίζουν το πινακοπαιχνίδι $R(s)$ μία φορά, όπου οι επιλογές των ενεργειών $a^1 \in A^1(s)$ και $a^2 \in A^2(s)$ προσδιορίζουν τόσο την αμοιβή ενός βήματος $r(s, a^1, a^2)$ όσο και την πιθανότητα μετάβασης $p(s'|s, a^1, a^2)$ στο επόμενο πινακοπαιχνίδι $R(s')$ που θα παιχτεί στο επόμενο στάδιο.

Ακολουθώντας όμοια συλλογιστική πορεία με τη θεωρία των διαδικασιών απόφασης, τότε εάν υποθέσουμε πως η τιμή του παιχνιδιού, $\boldsymbol{v}_\beta$, υπάρχει και ότι γνωρίζουμε ποιο είναι το βέλτιστο παίξιμο από το επόμενο στάδιο και έπειτα, τότε στο παρόν στάδιο έχουμε να αντιμετωπίσουμε το ακόλουθο βοηθητικό πινακοπαιχνίδι:

$$R(s, \boldsymbol{v}_\beta) = \left[r(s, a^1, a^2) + \beta \sum_{s' \in S} p(s'|s, a^1, a^2)\, v_\beta(s') \right]_{a^1=1, a^2=1}^{m^1(s), m^2(s)} \qquad (2.1.4)$$

Μπορούμε τώρα να υποθέσουμε ότι η σχέση (1.2.1) του 1ου Κεφαλαίου γενικεύεται και στην περίπτωση των στοχαστικών παιχνιδιών.

Εάν, λοιπόν, υπάρχει η τιμή $\boldsymbol{v}_\beta$ του παιχνιδιού, τότε για κάθε $s \in S$ θα πρέπει να ικανοποιείται ότι:

$$v_\beta(s) = val[R(s, \boldsymbol{v}_\beta)] \qquad (2.1.5)$$

όπου με $val[A]$ θα συμβολίζουμε την τιμή του πινακοπαιχνιδιού A, ενώ $v_\beta(s)$ είναι το s-οστό στοιχείο του διανύσματος $\boldsymbol{v}_\beta$.

Όλα τα παραπάνω συνοψίζονται στο επόμενο βασικό θεώρημα, το οποίο παραθέτουμε χωρίς απόδειξη.

Θεώρημα 2.1.2. (Shapley)

Το στοχαστικό παίγνιο μηδενικού αθροίσματος Γ_β με συντελεστή προεξόφλησης β, έχει ως διάνυσμα τιμής, $\boldsymbol{v}_\beta$, την μοναδική λύση των εξισώσεων:

$$v(s) = val[R(s, \boldsymbol{v})] \text{ , για όλα τα } s \in S, \text{ όπου } \boldsymbol{v}^T = \big(v(1), v(2), \dots, v(N)\big)^T$$

Ακόμη, εάν $\big(f^0(s), g^0(s)\big)$ είναι ένα ζευγάρι βέλτιστων στρατηγικών (εν γένει, όχι καθαρών) στο πινακοπαιχνίδι $R(s, \boldsymbol{v}_\beta)$ για κάθε $s \in S$, τότε :

Η $\boldsymbol{f}^0 = \big(\boldsymbol{f}^0(1), \boldsymbol{f}^0(2), \dots, \boldsymbol{f}^0(N)\big)$ είναι μια βέλτιστη στάσιμη στρατηγική για τον παίκτη 1 στο Γ_β και η $\boldsymbol{g}^0 = \big(\boldsymbol{g}^0(1), \boldsymbol{g}^0(2), \dots, \boldsymbol{g}^0(N)\big)$ είναι μια βέλτιστη στάσιμη στρατηγική για τον παίκτη 2 στο Γ_β.

Κλείνοντας αυτή την παράγραφο, επεκτείνουμε τον συμβολισμό του προηγούμενου κεφαλαίου και παραθέτουμε τον ορισμό κάποιων ποσοτήτων που θα χρησιμοποιήσουμε εν συνεχεία.

Έστω ένα σταθεροποιημένο ζεύγος στάσιμων στρατηγικών $\boldsymbol{f} = \big(\boldsymbol{f}(1), \boldsymbol{f}(2), \dots, \boldsymbol{f}(N)\big)$ και $\boldsymbol{g} = \big(\boldsymbol{g}(1), \boldsymbol{g}(2), \dots, \boldsymbol{g}(N)\big)$ των παικτών 1 και 2 αντίστοιχα. Θα κάνουμε τη σύμβαση πως η $\boldsymbol{f}$ είναι ένα μπλοκ διάνυσμα-γραμμή, ενώ η $\boldsymbol{g}$ θα είναι ένα μπλοκ διάνυσμα-στήλη.

Συνεπώς , αν

$$m^2 \doteqdot \sum_{s=1}^{N} m^2(s)$$

τότε η $\boldsymbol{g}$ θα είναι ένα m^2-διάσταστο διάνυσμα-στήλη όπου το s-οστό μπλοκ $g(s)$ θα είναι $m^2(s)$-διάστατο.

Στη συνέχεια ορίζουμε :

(i) Τις πιθανότητες μετάβασης στην επόμενη κατάσταση:

$$p(s'|s, \boldsymbol{f}, a^2) \doteqdot \sum_{a^1=1}^{m^1(s)} p(s'|s, a^1, a^2) f(s, a^1)\,, \quad s, s' \in S, a^2 \in A^2(s)$$

$$p(s'|s, a^1, \boldsymbol{g}) \doteqdot \sum_{a^2=1}^{m^2(s)} p(s'|s, a^1, a^2) g(s, a^2)\,, \quad s, s' \in S, a^1 \in A^1(s)$$

$$p(s'|s, \boldsymbol{f}, \boldsymbol{g}) \doteqdot \sum_{a^1=1}^{m^1(s)} \sum_{a^2=1}^{m^2(s)} p(s'|s, a^1, a^2) f(s, a^1) g(s, a^2)\,, \quad s, s' \in S$$

(ii) Τον πίνακα πιθανοτήτων μετάβασης που επάγεται από τις $(\boldsymbol{f}, \boldsymbol{g})$:

$$\boldsymbol{P}(\boldsymbol{f}, \boldsymbol{g}) \doteqdot [p(s'|s, \boldsymbol{f}, \boldsymbol{g})]_{s,s' \in S}$$

(iii) Τις αναμενόμενες αμοιβές ενός βήματος:

$$r(s, \boldsymbol{f}, a^2) \doteqdot \sum_{a^1=1}^{m^1(s)} r(s, a^1, a^2) f(s, a^1) = [\boldsymbol{f}(s) R(s)]_{a^2}\,, \quad s \in S, a^2 \in A^2(s)$$

$$r(s, a^1, \boldsymbol{g}) \doteqdot \sum_{a^2=1}^{m^2(s)} r(s, a^1, a^2) g(s, a^2) = [R(s) \boldsymbol{g}(s)]_{a^1}\,, \quad s \in S, a^1 \in A^1(s)$$

$$r(s, \boldsymbol{f}, \boldsymbol{g}) \doteqdot \sum_{a^1=1}^{m^1(s)} \sum_{a^2=1}^{m^2(s)} r(s, a^1, a^2) f(s, a^1) g(s, a^2) = \boldsymbol{f}(s) R(s) \boldsymbol{g}(s)\,, \quad s \in S$$

(iv)Το N-διάστατο διάνυσμα-στήλη αμοιβών ενός βήματος:

$$\boldsymbol{r}(\boldsymbol{f},\boldsymbol{g}) \doteqdot \big(r(1,\boldsymbol{f},\boldsymbol{g}), r(2,\boldsymbol{f},\boldsymbol{g}), \dots, r(N,\boldsymbol{f},\boldsymbol{g})\big)^T$$

(v) Το N-διάστατο διάνυσμα τιμής με προεξόφληση για το ζεύγος $(\boldsymbol{f},\boldsymbol{g})$:

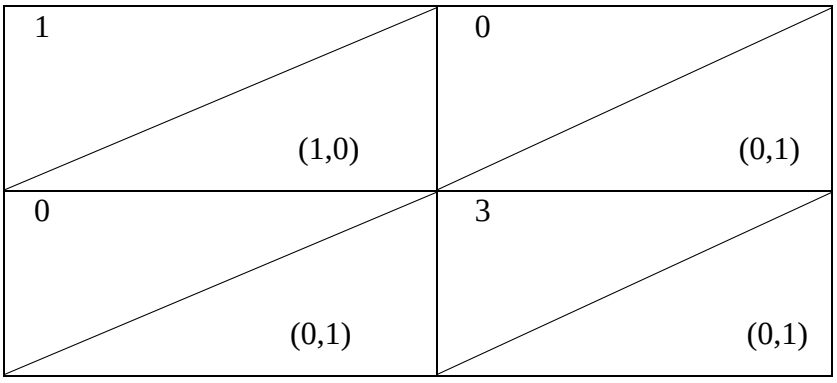
$$\boldsymbol{v}_\beta(\boldsymbol{f},\boldsymbol{g}) \doteqdot \big(\boldsymbol{I} - \beta \boldsymbol{P}(\boldsymbol{f},\boldsymbol{g})\big)^{-1} \boldsymbol{r}(\boldsymbol{f},\boldsymbol{g})$$

2.2 Γραμμικός Προγραμματισμός και Στοχαστικά Παίγνια με συντελεστή προεξόφλησης

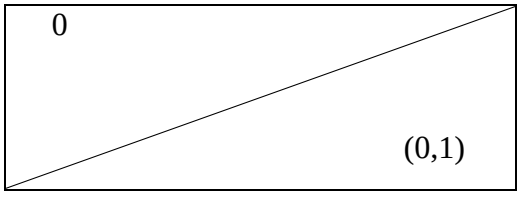
Όπως έχουμε ήδη διαπιστώσει, ο γραμμικός προγραμματισμός αποτελεί ένα ιδιαιτέρως χρήσιμο εργαλείο για την επίλυση Μαρκοβιανών διαδικασιών Απόφασης. Όπως θα γίνει φανερό μέσω του επόμενου παραδείγματος, δεν θα πρέπει να αναμένουμε να φτάσουμε στη λύση οποιουδήποτε στοχαστικού παιγνίου με μεθόδους γραμμικού προγραμματισμού.

Παράδειγμα 2.2.1

Έστω $S = \{1,2\}$, $A^1(1) = A^2(1) = \{1,2\}$, $A^1(2) = A^2(2) = \{1\}$, $\beta = \frac{1}{2}$, ενώ οι αμοιβές και οι πιθανότητες μετάβασης φαίνονται στο επόμενο γράφημα:



Κατάσταση 1



Κατάσταση 2

Σκοπός του παραπάνω παραδείγματος είναι να δείξουμε πως, πράγματι, οι μέθοδοι γραμμικού προγραμματισμού κρίνονται ανεπαρκείς για την επίλυση ενός οποιουδήποτε στοχαστικού παιγνίου. Για το λόγο αυτό, χωρίς να δούμε καθόλου το θεωρητικό υπόβαθρο που απαιτείται για

τον υπολογισμό της τιμής του παιγνίου αυτού, παραθέτουμε απλώς το αποτέλεσμα. Αναφέρουμε, μόνο, πως το αποτέλεσμα αυτό είναι συνέπεια ενός θεωρήματος γνωστού στη βιβλιογραφία ως Shapley-Snow Theorem. Από το θεώρημα προκύπτει ότι τελικά το ζητούμενο v θα είναι λύση της εξίσωσης:

$$v = \frac{(1+\frac{1}{2}v)(3+\frac{1}{2}v)}{4+v} \Rightarrow 3v^2 + 8v - 12 = 0 \Longrightarrow v = \frac{1}{3}\left(-4+2\sqrt{13}\right)$$

Οπότε το διάνυσμα τιμής θα είναι το :

$$\boldsymbol{v} = \left(\frac{1}{3}\left(-4+2\sqrt{13}\right), 0\right)^T$$

Σχόλιο

Αξίζει να παρατηρήσουμε πως, εάν το παραπάνω παράδειγμα επιδέχονταν λύση με μεθόδους γραμμικού προγραμματισμού, με συντελεστές που είναι ρητές συναρτήσεις των (ρητών) δεδομένων του προβλήματος, τότε η βέλτιστη λύση θα έπρεπε να βρίσκεται σε κάποιο οριακό σημείο της εφικτής περιοχής και θα είχε υποχρεωτικά ρητή τιμή. Συνεπώς, το γεγονός ότι στο συγκεκριμένο παράδειγμα καταλήξαμε σε διάνυσμα τιμής με άρρητους αριθμούς , υποδηλώνει πως δεν είναι εφικτή η επίλυση μέσω γραμμικού προγραμματισμού. Το ερώτημα που θα μας απασχολήσει στη συνέχεια είναι το εξής: « Μήπως υπάρχουν ορισμένες ειδικές κλάσεις στοχαστικών παιγνίων, τα οποία μπορούν να αντιμετωπιστούν με μεθόδους γραμμικού προγραμματισμού;»

2.2.1. Παίγνια με προεξόφληση με 1 ελεγκτή

Για αυτή την κατηγορία παιγνίων, η πιθανότητα μετάβασης μεταξύ των καταστάσεων εξαρτάται από τις ενέργειες που επιλέγει μόνον ο ένας παίκτης

Θα συμβολίζουμε με $\Gamma_\beta(1)$ το παιχνίδι που 'ελέγχεται' από τον παίκτη 1 και που ορίζεται από την ιδιότητα:

$$p(s'|s,a^1,a^2) \equiv p(s'|s,a^1) \text{ ,όπου } s,s' \in S,\ a^1 \in A^1(s),\ a^2 \in A^2(s)$$

Με εντελώς όμοιο τρόπο ορίζεται και το παιχνίδι $\Gamma_\beta(2)$.

Δοθέντος του παιχνιδιού $\Gamma_\beta(1)$, μπορούμε να 'ελαφρύνουμε' τον συμβολισμό που έχουμε δει ως εξής:

$$\boldsymbol{P}(\boldsymbol{f},\boldsymbol{g}) = \boldsymbol{P}(\boldsymbol{f}) \text{ και}$$

$$\boldsymbol{v}_\beta(\boldsymbol{f},\boldsymbol{g}) = \left(\boldsymbol{I} - \beta\boldsymbol{P}(\boldsymbol{f})\right)^{-1}\boldsymbol{r}(\boldsymbol{f},\boldsymbol{g}) \text{ ,για όλα τα } (\boldsymbol{f},\boldsymbol{g}) \in F_s \times G_s$$

Λόγω της κατασκευής του $\Gamma_\beta(1)$, είναι λογικό να αναμένουμε αυτό να συμπεριφέρεται περισσότερο σαν Μαρκοβιανή διαδικασία Απόφασης του παίκτη 1, παρά σαν γενικό στοχαστικό παίγνιο με προεξόφληση.

Υποθέτουμε πως ο παίκτης 2 ακολουθεί μία στάσιμη στρατηγική $\boldsymbol{g} \in G_S$. Τότε, κατ'αναλογία με τα γραμμικά προγράμματα (P_β) και (D_β) του προηγούμενου κεφαλαίου , θεωρούμε το ακόλουθο ζεύγος (Πρωτεύοντος – Δυϊκού) προβλημάτων :

$$\min \sum_{s'=1}^{N} \frac{1}{N} v(s')$$

με περιορισμούς: $\left(P_\beta(1)\right)$

$$v(s) \geq [R(s)\boldsymbol{g}(\boldsymbol{s})]_{a^1} + \beta \sum_{s'=1}^{N} p(s'|s,a^1)v(s'), \qquad s \in S, a^1 \in A^1(s)$$

$$\sum_{a^2 \in A^2(s)} g(s,a^2) = 1\,, \qquad s \in S$$

$$g(s,a^2) \geq 0, \;\; s \in S, a^2 \in A^2(s)$$

$$\max \sum_{s'=1}^{N} z(s)$$

με περιορισμούς: $\left(D_\beta(1)\right)$

$$\sum_{s=1}^{N} \sum_{a^1 \in A^1(s)} \left(\delta_{ss'} - \beta p(s'|s,a^1)\right) x_{sa^1} = \frac{1}{N}\;, \qquad s' \in S$$

$$z(s) \leq [\boldsymbol{x}(s)R(s)]_{a^2}, \;\; s \in S, a^1 \in A^1(s)$$

$$x(s,a^1) \geq 0, \qquad s \in S, a^1 \in A^1(s)$$

όπου $\boldsymbol{x}(\boldsymbol{s}) = \left(x(s,1), x(s,2), \dots, x\left(s, m^1(s)\right)\right)$ για κάθε $s \in S$.

Σχόλιο

Όπως και στην παράγραφο 1.3, τα παραπάνω επιχειρήματα θα ήταν έγκυρα , ακόμη και αν αντικαθιστούσαμε τους συντελεστές $\frac{1}{N}$ με κάποιες θετικές πιθανότητες εκκίνησης $\gamma(s')$, οι οποίες αθροίζουν στη μονάδα.

Το προηγούμενο ζεύγος γραμμικών προγραμμάτων μπορεί να χρησιμοποιηθεί για την επίλυση του στοχαστικού παιγνίου με προεξόφληση με έναν ελεγκτή.

Η επόμενη παρατήρηση θα φανεί ιδιαίτερα χρήσιμη στην συνέχεια του κεφαλαίου.

Παρατήρηση

Παρατηρούμε πως ορισμένοι από τους περιορισμούς στο $\big(P_\beta(1)\big)$ και $\big(D_\beta(1)\big)$ είναι σε ομάδες που αντιστοιχούν ακριβώς στη μπλοκ δομή των στάσιμων στρατηγικών $\boldsymbol{f}$ ή $\boldsymbol{g}$ του παίκτη 1 ή 2. Επειδή αυτή ακριβώς η δομή πρόκειται να εμφανιστεί επανειλημμένα στη συνέχεια, είναι βολικό να χρησιμοποιούμε τη φράση «αναμιγνύοντας τους περιορισμούς ως προς την $\boldsymbol{f}$ ή την $\boldsymbol{g}$». Αυτή η φράση θα σημαίνει ότι κάθε περιορισμός μιας τέτοιας ομάδας που αντιστοιχεί στην (s, a^1) (αντίστοιχα (s, a^2)) πολλαπλασιάζεται επί $f(s, a^1)$ (αντίστοιχα επί $g(s, a^2)$) και ότι όλοι οι περιορισμοί αυτής της ομάδας θα αθροίζονται για όλα τα $a^1 \in A^1(s)$ (αντίστοιχα $a^2 \in A^2(s)$).

Για παράδειγμα, έστω μια αυθαίρετη $\boldsymbol{f} \in F_s$ και αναμιγνύουμε την πρώτη ομάδα περιορισμών στο $\big(P_\beta(1)\big)$ ως προς την $\boldsymbol{f}$. Τότε, για $s \in S$, θα πάρουμε:

$$\sum_{a^1 \in A^1(s)} v(s) f(s, a^1) \geq \boldsymbol{f}(s) R(s) \boldsymbol{g}(s) + \beta \sum_{s'=1}^{N} p(s'|s, \boldsymbol{f}) v(s')$$

ή, ισοδύναμα, αφού $\sum_{a^1 \in A^1(s)} f(s, a^1) = 1$

$$v(s) \geq r(s, \boldsymbol{f}, \boldsymbol{g}) + \beta [\boldsymbol{P}(\boldsymbol{f}) \boldsymbol{v}]_s$$

(όπου $\boldsymbol{v} = \big(v(1), v(2), \dots, v(N)\big)^T$)

Θεώρημα 2.2.1

Θεωρούμε το στοχαστικό παίγνιο με προεξόφληση με 1-ελεγκτή $\Gamma_\beta(1)$ και το ζεύγος $\big(P_\beta(1)\big)$ και $\big(D_\beta(1)\big)$ γραμμικών προγραμμάτων. Ακόμη, έστω $(\boldsymbol{v}^0, \boldsymbol{g}^0)$ μια βέλτιστη λύση του $\big(P_\beta(1)\big)$ και $(\boldsymbol{z}^0, \boldsymbol{x}^0)$ μια βέλτιστη λύση του $\big(D_\beta(1)\big)$. Τότε:

(i) Το διάνυσμα αξίας του $\Gamma_\beta(1)$ είναι $\boldsymbol{v}^0$ και η $\boldsymbol{g}^0$ είναι μία βέλτιστη στρατηγική για τον παίκτη 2

(ii) Εάν θέσουμε:

$$x_s^o \doteqdot \sum_{a^1 \in A^1(s)} x_{sa^1}^o$$

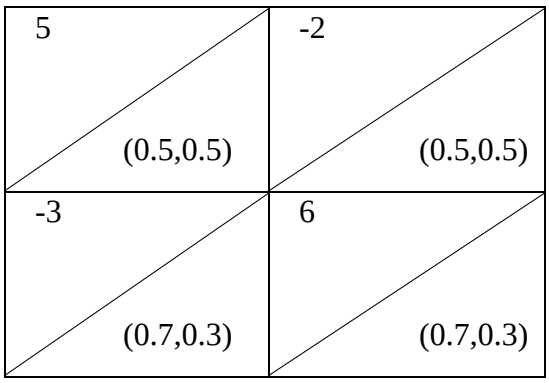
και μία στάσιμη στρατηγική $\boldsymbol{f}^0$ για τον παίκτη 1 ορίζεται ως:

$$\boldsymbol{f}^0(s,a^1)=\frac{x^o_{sa^1}}{x^o_s}\quad ,\qquad s\in S, a^1\in A^1(s)$$

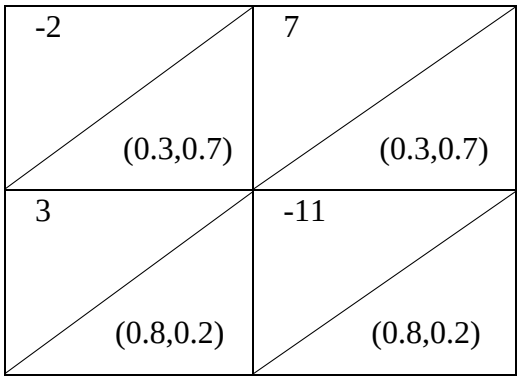
τότε η $\boldsymbol{f}^0$ είναι μια βέλτιστη στρατηγική του παίκτη 1

Παράδειγμα 2.2.2

Έστω $S=\{1,2\}, A^1(s)=A^2(s)=\{1,2\}$ για $s=1,2$, $\beta=0.6$ και οι αμοιβές και οι πιθανότητες μετάβασης είναι:



Κατάσταση 1



Κατάσταση 2

Από τη δομή των πιθανοτήτων μετάβασης μεταξύ των καταστάσεων, γίνεται φανερό πως το παραπάνω παίγνιο ελέγχεται από τον παίκτη 1. Το πρωτεύον γραμμικό πρόγραμμα $P_{0.6}(1)$ θα έχει την ακόλουθη μορφή:

$$\min\left[\frac{1}{2}v(1)+\frac{1}{2}v(2)\right]$$

με περιορισμούς:

$v(1)\geq 5g(1,1)-2g(1,2)+0.3v(1)+0.3v(2)$

$v(1)\geq -3g(1,1)+6g(1,2)+0.42v(1)+0.18v(2)$

$v(2)\geq -2g(2,1)+7g(2,2)+0.18v(1)+0.42v(2)$

$v(2)\geq 3g(2,1)-11g(2,2)+0.48v(1)+0.12v(2)$

$g(1,1)+g(1,2)=1$

$g(2,1)+g(2,2)=1$

$g(1,1), g(1,2), g(2,1), g(2,2)\ \geq 0$

Χρησιμοποιώντας το λογισμικό MATLAB, παίρνουμε τα ακόλουθα αποτελέσματα:

Αντικειμενική συνάρτηση	1.9225
$\mathrm{v}(1)$	2.7393
$v(2)$	1.1058
$g(1,1)$	0.5123
$g(1,2)$	0.4877
$g(2,1)$	0.7613
$g(2,2)$	0.2387

Οπότε συμπεραίνουμε ότι :

$$\boldsymbol{v}_{0.6}^{0} = (2.7393, 1.1058) \text{ και } \boldsymbol{g}^{0} = \big((0.5123, 0.4877), (0.7613\ , 0.2387)\big)$$

Το δυϊκό πρόβλημα θα έχει τη μορφή:

$$\max[z(1) + z(2)]$$

με περιορισμούς:

$$0.7x_{11} + 0.58x_{12} - 0.18x_{21} - 0.48x_{22} = \frac{1}{2}$$

$$-0.3x_{11} - 0.18x_{12} + 0.58x_{21} + 0.88x_{22} = \frac{1}{2}$$

$$z(1) \leq 5x_{11} - 3x_{12}$$

$$z(1) \leq -2x_{11} + 6x_{12}$$

$$z(2) \leq -2x_{21} + 3x_{22}$$

$$z(2) \leq 7x_{21} - 11x_{22}$$

$$x_{11}, x_{12}, x_{21}, x_{22} \geq 0$$

Χρησιμοποιώντας και πάλι το MATLAB, λαμβάνουμε τα αποτελέσματα:

Αντικειμενική συνάρτηση	1.9225
$z(1)$	1.9740
$z(2)$	-0.0515
x_{11}	0.7403
x_{12}	0.5758
x_{21}	0.7207
x_{22}	0.4633

Σύμφωνα με το (ii) του Θεωρήματος 2.2.1, η βέλτιστη στρατηγική $\boldsymbol{f}^0$ θα είναι:

$$\boldsymbol{f}^0 = \left(\left(\frac{x_{11}^0}{x_{11}^0 + x_{11}^0}, \frac{x_{12}^0}{x_{11}^0 + x_{12}^0}\right), \left(\frac{x_{21}^0}{x_{21}^0 + x_{22}^0}, \frac{x_{22}^0}{x_{21}^0 + x_{22}^0}\right)\right)$$

$$= \big((0.5625, 0.4375), (0.6087, 0.3913)\big)$$

2.2.2. Παίγνια με προεξόφληση διαχωρίσιμης αμοιβής (SEparable Reward) και ανεξάρτητα της κατάστασης μετάβασης (State Independent Transition) (SER-SIT)

Σ' αυτή την κλάση παιχνιδιών, θα κάνουμε ορισμένες υποθέσεις που αφορούν στη δομή του παιχνιδιού, ούτως ώστε να μπορέσουμε να αξιοποιήσουμε αποτελέσματα από τη θεωρία του γραμμικού προγραμματισμού και να φτάσουμε στον υπολογισμό του διανύσματος-τιμής καθώς και βέλτιστων στρατηγικών των παικτών.

Συγκεκριμένα, η δομή του παιχνιδιού θέλουμε να έχει την ακόλουθη μορφή:

$$r(s, a^1, a^2) = c(s) + \rho(a^1, a^2)\,, \;\; s \in S, a^1 \in A^1(s), a^2 \in A^2(s) \qquad (SER)$$

$$p(s'|s, a^1, a^2) = p(s'|a^1, a^2)\,, \;\; s \in S, a^1 \in A^1(s), a^2 \in A^2(s) \qquad (SIT)$$

Προφανώς, για να έχει νόημα η 2η υπόθεση (SIT) , θα πρέπει υποχρεωτικά να ισχύει ότι $m^1(s) = \mu$ και $m^2(s) = \nu$ για όλα τα $s \in S$. Έτσι, ένα ζεύγος ενέργειων a^1, a^2 θα καθορίζει την ίδια μετάβαση, ανεξαρτήτως από την παρούσα κατάσταση.

Η 1η υπόθεση (SER) καθιστά όλες τις αμοιβές των παικτών να προκύπτουν από το άθροισμα μιας ποσότητας που εξαρτάται μόνο από την παρούσα κατάσταση ($c(s)$) και μιας άλλης που εξαρτάται μόνο από το ζεύγος των ενεργειών που επιλέχθηκαν ($\rho(a^1, a^2)$).

Θα συμβολίζουμε : $\boldsymbol{c}^T = \big(c(1), c(2), \dots, c(N)\big)$

Ο συλλογισμός που θα ακολουθήσουμε για την επίλυση παιγνίων τύπου SER-SIT έχει ως εξής: Αντιστοιχίζουμε στο διάνυσμα $\boldsymbol{c}$ το βοηθητικό πινακοπαιχνίδι:

$$R(\boldsymbol{c}) \doteqdot \left[\rho(a^1, a^2) + \beta \sum_{s' \in S} p(s' \mid a^1, a^2) c(s')\right]_{\alpha^1=1, \alpha^2=1}^{\mu, \nu}$$

Έστω $\rho \doteqdot val[R(\boldsymbol{c})]$ και $\boldsymbol{x}^0 = \big(x_1^0, x_2^0, \dots, x_\mu^0\big), \boldsymbol{y}^0 = (y_1^0, y_2^0, \dots, y_\nu^0)$ ένα ζεύγος από βέλτιστες στρατηγικές για το παιχνίδι $R(\boldsymbol{c})$.

Θεωρούμε τώρα τη στρατηγική $\boldsymbol{g}^0 \in G_s$ (αντίστοιχα $\boldsymbol{f}^0 \in F_s$) η οποία ορίζεται ως: $\boldsymbol{g}^0(s) = \boldsymbol{y}^0$ (αντίστοιχα $\boldsymbol{f}^0(s) = \boldsymbol{x}^0$) για $s \in S$.

Οι στρατηγικές αυτές είναι βέλτιστες για το SER-SIT παίγνιο και μάλιστα θα δείξουμε ότι ισχύει:

$$\boldsymbol{v}_\beta = \boldsymbol{c} + \left(\frac{\rho}{1-\beta}\right)\boldsymbol{1} \qquad (2.2.1)$$

Αρχικά, θα αναδιατυπώσουμε το παιχνίδι χρησιμοποιώντας τη σχέση (1.1.4). Θα έχουμε, δηλαδή, ότι :

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

Λόγω των περιορισμών SER και SIT , για κάθε $s \in S$ θα έχουμε:

$$v_\beta(s, \boldsymbol{f}, \boldsymbol{g}) = r(s, \boldsymbol{f}, \boldsymbol{g}) + \beta \sum_{s' \in S} p(s', \boldsymbol{f}, \boldsymbol{g})\, v_\beta(s, \boldsymbol{f}, \boldsymbol{g})$$

$$= \sum_{a^1} \sum_{a^2} r(s, a^1, a^2) f(s, a^1) g(s, a^2) + \beta \sum_{s' \in S} p(s', \boldsymbol{f}, \boldsymbol{g})\, v_\beta(s, \boldsymbol{f}, \boldsymbol{g})$$

$$= \sum_{a^1} \sum_{a^2} \big(c(s) + \rho(a^1, a^2)\big) f(s, a^1) g(s, a^2) + \beta \sum_{s' \in S} p(s', \boldsymbol{f}, \boldsymbol{g})\, v_\beta(s, \boldsymbol{f}, \boldsymbol{g})$$

$$= c(s) \sum_{a^1} \sum_{a^2} f(s, a^1) g(s, a^2) + \sum_{a^1} \sum_{a^2} \rho(a^1, a^2)\, f(s, a^1) g(s, a^2)$$
$$+ \beta \sum_{s' \in S} p(s', \boldsymbol{f}, \boldsymbol{g})\, v_\beta(s, \boldsymbol{f}, \boldsymbol{g})$$

$$= c(s) + \sum_{a^1}\sum_{a^2}\left([R(c)]_{a^1,a^2} - \beta \sum_{s'\in S} p(s'|\, a^1, a^2)c(s') \right) f(s,a^1)g(s,a^2)$$
$$+ \beta \sum_{s'\in S} p(s', \boldsymbol{f}, \boldsymbol{g})\, v_\beta(s, \boldsymbol{f}, \boldsymbol{g})$$

$$= c(s) + \boldsymbol{f}(s)R(c)\boldsymbol{g}(s) - \beta \sum_{a^1} f(s,a^1) \sum_{a^2} g(s,a^2) \sum_{s\prime\in S} p(s'|a^1,a^2)c(s')$$
$$+ \beta \sum_{s'\in S} p(s', \boldsymbol{f}, \boldsymbol{g})\, v_\beta(s, \boldsymbol{f}, \boldsymbol{g})$$

$$= c(s) + \boldsymbol{f}(s)R(c)\boldsymbol{g}(s) + \beta \sum_{s\prime\in S} p(s'|\boldsymbol{f}, \boldsymbol{g}) \left(\boldsymbol{v}_\beta(s', \boldsymbol{f}, \boldsymbol{g}) - c(s') \right)$$

Εάν θέσουμε $r(\boldsymbol{c}, \boldsymbol{f}, \boldsymbol{g}) \doteqdot \boldsymbol{f}(s)R(\boldsymbol{c})\boldsymbol{g}(s)$, για $s \in S$, τότε η παραπάνω σχέση σε διανυσματική μορφή γίνεται:

$$\boldsymbol{v}_\beta(\boldsymbol{f}, \boldsymbol{g}) = \boldsymbol{c} + r(\boldsymbol{c}, \boldsymbol{f}, \boldsymbol{g})\mathbf{1} + \beta \boldsymbol{P}(\boldsymbol{f}, \boldsymbol{g})\big(\boldsymbol{v}_\beta(\boldsymbol{f}, \boldsymbol{g}) - \boldsymbol{c}\big)$$

και λύνοντας ως προς $\boldsymbol{v}_\beta(\boldsymbol{f}, \boldsymbol{g})$, λαμβάνουμε:

$$\big(\boldsymbol{I} - \beta \boldsymbol{P}(\boldsymbol{f}, \boldsymbol{g})\big)\boldsymbol{v}_\beta(\boldsymbol{f}, \boldsymbol{g}) = \big(\boldsymbol{I} - \beta \boldsymbol{P}(\boldsymbol{f}, \boldsymbol{g})\big)\boldsymbol{c} + \boldsymbol{r}(\boldsymbol{c}, \boldsymbol{f}, \boldsymbol{g})\mathbf{1}$$
$$\Rightarrow \boldsymbol{v}_\beta(\boldsymbol{f}, \boldsymbol{g}) = \boldsymbol{c} + [\boldsymbol{I} - \beta \boldsymbol{P}(\boldsymbol{f}, \boldsymbol{g})]^{-1}[r(\boldsymbol{c}, \boldsymbol{f}, \boldsymbol{g})\mathbf{1}] \qquad (2.2.2)$$

Όπως είδαμε και στο 1ο κεφάλαιο, ισχύει ότι :

$$[\boldsymbol{I} - \beta \boldsymbol{P}(\boldsymbol{f}, \boldsymbol{g})]^{-1} = \sum_{t=0}^{\infty} \beta^t\, \boldsymbol{P}^t(\boldsymbol{f}, \boldsymbol{g})$$

Όμως, λόγω της δομής SER-SIT, θα έχουμε ότι για κάθε t, $\boldsymbol{P}^t(\boldsymbol{f}, \boldsymbol{g}) = \boldsymbol{P}(\boldsymbol{f}, \boldsymbol{g})$, οπότε:

$$[\boldsymbol{I} - \beta \boldsymbol{P}(\boldsymbol{f}, \boldsymbol{g})]^{-1} = \boldsymbol{P}(\boldsymbol{f}, \boldsymbol{g}) \sum_{t=0}^{\infty} \beta^t = \frac{1}{1-\beta} \boldsymbol{P}(\boldsymbol{f}, \boldsymbol{g})$$

Συνεπώς, έχουμε καταλήξει στη σχέση:

$$\boldsymbol{v}_\beta(\boldsymbol{f}, \boldsymbol{g}) = \boldsymbol{c} + \frac{1}{1-\beta} \boldsymbol{P}(\boldsymbol{f}, \boldsymbol{g})[r(\boldsymbol{c}, \boldsymbol{f}, \boldsymbol{g})\mathbf{1}]$$

Οπότε αν θεωρήσουμε το ζεύγος βέλτιστων στρατηγικών $(\boldsymbol{f}, \boldsymbol{g})$ τέτοιες ώστε $\boldsymbol{f}(s) = \boldsymbol{x}^T$ και $\boldsymbol{g}(s) = \boldsymbol{y}$ για όλα τα $s \in S$, και επειδή ο πίνακας $\boldsymbol{P}$ είναι στοχαστικός, παίρνουμε :

$$\boldsymbol{v}_{\beta}(\boldsymbol{f},\boldsymbol{g}) = \boldsymbol{c} + \frac{r(\boldsymbol{c},\boldsymbol{f},\boldsymbol{g})}{1-\beta}\mathbf{1} = \boldsymbol{c} + \frac{\boldsymbol{x}^T R(c)\boldsymbol{y}}{1-\beta}\mathbf{1} = \boldsymbol{c} + \left(\frac{\rho}{1-\beta}\right)\mathbf{1}$$

και έτσι, έχουμε καταλήξει στο ζητούμενο.

2.2.3 Παίγνια με προεξόφληση εναλλασσόμενου ελεγκτή

Η κατηγορία αυτή αποτελεί την 3η και τελευταία κλάση παιγνίων με συντελεστή προεξόφλησης που θα μας απασχολήσει σε αυτή την εργασία. Πρόκειται για γενίκευση των παιχνιδιών με έναν ελεγκτή που αναλύθηκε προηγουμένως. Για αυτά τα παίγνια, οι πιθανότητες μετάβασης μεταξύ των καταστάσεων εξαρτώνται από τις ενέργειες του ενός παίκτη για ένα σύνολο καταστάσεων, και από τις ενέργειες του άλλου παίκτη, για τις υπόλοιπες καταστάσεις.

Για το λόγο αυτό, καλούμαστε να διαμερίσουμε το σύνολο καταστάσεων S. Θα έχουμε δηλαδή:

$S = S^1 \cup S^2$, όπου τα S^1, S^2 είναι μη-κενά και ξένα μεταξύ τους σύνολα, έτσι ώστε για κάθε $a^1 \in A^1(s), a^2 \in A^2(s)$, να ισχύει ότι :

$$p(s'|s,a^1,a^2) = \begin{cases} p(s'|s,a^1)\,, s \in S^1 \\ p(s'|s,a^2)\,, s \in S^2 \end{cases} \qquad (SW)$$

Προκειμένου να αντιμετωπιστεί ένα τέτοιο παίγνιο εναλλασσόμενου ελεγκτή μέσω γραμμικού προγραμματισμού, θα χρειαστεί πρώτα να λυθεί ένα (πεπερασμένο) σύνολο γραμμικών προγραμμάτων, σε αντίθεση με την κατηγορία του ενός ελεγκτή όπου η λύση ενός γραμμικού προγράμματος ήταν αρκετή.

Προτού δοθεί ο σχετικός αλγόριθμος, αξίζει να παρατηρήσουμε μια ιδιότητα των παιγνίων τύπου *SW*. Έστω ότι ο παίκτης 1 σταθεροποιεί την στρατηγική $\hat{\boldsymbol{f}}(s)$ για κάθε $s \in S^1$. Τότε, μπορούμε να ορίσουμε ένα παίγνιο $\Gamma_{\beta}(2, \hat{\boldsymbol{f}})$ με ελεγκτή τον παίκτη 2, χώρο καταστάσεων και ενεργειών ίδιο με του Γ_{β} , αμοιβές και πιθανότητες μετάβασης ίδιες για τα $s \in S^2$, αλλά τροποποιημένες για τα $s \in S^1$, ως εξής :

$$\hat{r}(s,a^1,a^2) \doteq \sum_{a^1 \in A^1(s)} r(s,a^1,a^2)\hat{f}(s,a^1)$$

$$\hat{p}(s'|s,a^1,a^2) \doteq \sum_{a^1 \in A^1(s)} p(s'|s,a^1,a^2)\,\hat{f}(s,a^1)$$

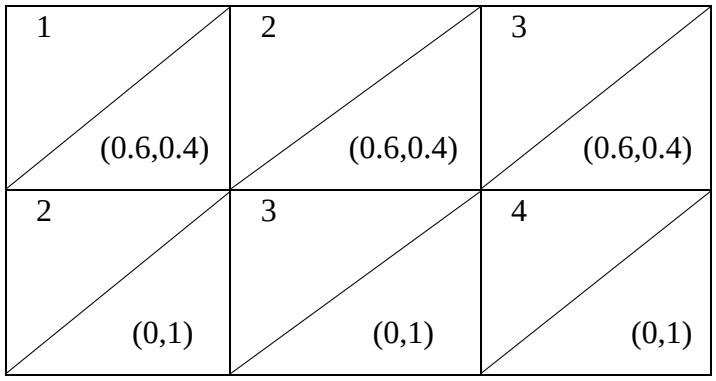
για όλα τα $a^1 \in A^1(s), a^2 \in A^2(s), s \in S^1, s' \in S^2$

Φυσικά, για να ισχύουν τα παραπάνω, θα πρέπει όλες οι ενέργειες του παίκτη 1 στο $\Gamma_\beta(2,\hat{\boldsymbol{f}})$, για $s \in S^1$, να ταυτίζονται · όσον αφορά τόσο τις μεταβάσεις όσο και τις αμοιβές, ακριβώς όπως φαίνεται στο επόμενο παράδειγμα.

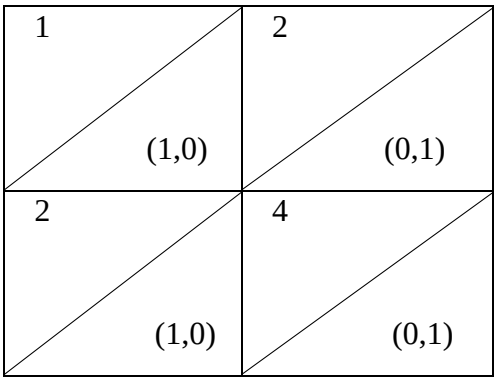
Παράδειγμα 2.2.3.

Έστω $S = \{1,2\}, A^1(1) = A^1(2) = A^2(2) = \{1,2\}, A^2(1) = \{1,2,3\}$

ενώ οι αμοιβές ενός βήματος και οι μεταβάσεις φαίνονται στο παρακάτω διάγραμμα.

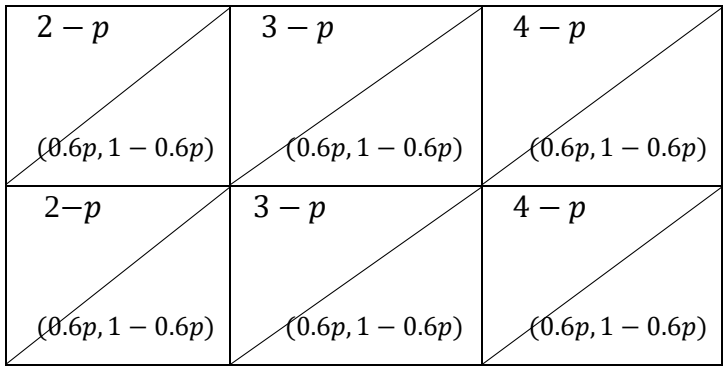


Κατάσταση 1

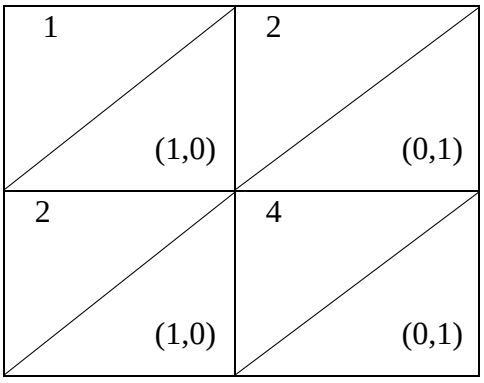


Κατάσταση 2

Βλέπουμε πως ο παίκτης 1 ελέγχει τις μεταβάσεις στο σύνολο $S^1 = \{1\}$, ενώ ο πάικτης 2 στο σύνολο $S^2 = \{2\}$. Ακόμη, εάν ο παίκτης 1 σταθεροποιήσει μία στρατηγική για την κατάσταση 1, έστω την $\hat{\boldsymbol{f}}(1) = (p, 1-p)$, τότε σύμφωνα με τα προηγούμενα, μπορούμε να θεωρήσουμε το ακόλουθο παιχνίδι $\Gamma_\beta(2,\hat{\boldsymbol{f}}(1))$, με ελεγκτή τον παίκτη 2 :



Κατάσταση 1



Κατάσταση 2

Ο αλγόριθμος που θα χρησιμοποιήσουμε προκειμένου να φτάσουμε στη λύση ενός παιγνίου τύπου SW, όπως κατασκευάστηκε παραπάνω, είναι ο ακόλουθος:

Αλγόριθμος 2.2.1.

Βήμα 1: Θέσε $k = 0$, επίλεξε ένα αυθαίρετο $\boldsymbol{v}^0 = (v^0(1), \dots, v^0(N))^T$ και βρες μια βέλτιστη στρατηγική $\boldsymbol{f}^0(s)$ για τον παίκτη 1 στο πινακοπαιχνίδι $R(s, \boldsymbol{v}^0)$, για κάθε $s \in S^1$.

Βήμα 2: Θέσε $k = k + 1$. Λύσε το παίγνιο $\Gamma_\beta(2, \hat{\boldsymbol{f}}^{k-1})$ με ελεγκτή τον παίκτη 2. Η τιμή του θα συμβολίζεται με $\boldsymbol{v}_\beta$.
Θέσε $\boldsymbol{v}^k = \boldsymbol{v}_\beta$.

Βήμα 3: Αν $v^k(s) = val[R(s, v^k)]$, για κάθε $s \in S$, τότε σταματάμε.
Αλλιώς, βρες μια βέλτιστη στρατηγική $\boldsymbol{f}^k(s)$ για τον παίκτη 1 στο πινακοπαιχνίδι $R(s, \boldsymbol{v}^k)$, για κάθε $s \in S^1$ και επίστρεψε στο Βήμα 2.

2.3 Στοχαστικά παίγνια Οριακού Μέσου

Στην ανάλυση που έγινε στο προηγούμενο κεφάλαιο, πάνω στις Μαρκοβιανές διαδικασίες απόφασης, είδαμε πως το κριτήριο πληρωμής οριακού μέσου θεωρείται 'πιο δύσκολο' στην αντιμετώπιση, συγκριτικά με το κριτήριο προεξόφλησης. Παρ' όλα αυτά, ξεπερνώντας ορισμένες τεχνικές δυσκολίες, καταλήξαμε στη διατύπωση σημαντικών αποτελεσμάτων και για τα δύο κριτήρια απόδοσης, όπως είναι η ύπαρξη βέλτιστων στάσιμων στρατηγικών, η εύρεση των οποίων προκύπτει να είναι η λύση σε κατάλληλα κατασκευασμένα γραμμικά προγράμματα.

Με αυτό το σκεπτικό, θα μπορούσε κανείς να εικάσει πως, με ίσως ορισμένες τεχνικές δυσκολίες, θα μπορούσαμε να γενικεύσουμε τα αποτελέσματα των Μαρκοβιανών διαδικασιών στα στοχαστικά παίγνια.

Δυστυχώς όμως, μια τέτοια γενίκευση δεν είναι εφικτή και αυτό θα γίνει φανερό μέσα από ένα πολύ γνωστό αντιπαράδειγμα, το οποίο απαντάται στη βιβλιογραφία ως 'The Big Match'. Προτού περάσουμε στην ανάλυση του Big Match , ας ορίσουμε το στοχαστικό παίγνιο με κριτήριο πληρωμής τον οριακό μέσο, Γ_α, περιορισμένοι στο σύνολο των στάσιμων στρατηγικών, με τρόπο ανάλογο της πληρωμής με προεξόφληση.

Η διαφοροποίηση των δύο παιγνίων έγκειται ουσιαστικά στον ορισμό της συνάρτησης απόδοσης των παικτών. Ορίζουμε, λοιπόν, την πληρωμή που θα λάβει ο παίκτης 1 από τον παίκτη 2 , που αντιστοιχεί σε ζεύγος στρατηγικών $(\boldsymbol{f}, \boldsymbol{g}) \in F^s \times G^s$ ως εξής:

$$v_a(s, \boldsymbol{f}, \boldsymbol{g}) \doteqdot \lim_{T\to\infty} \left[\left(\frac{1}{T+1} \right) \sum_{t=0}^{T} E_{s\,\boldsymbol{fg}}[R_t] \right] = [\boldsymbol{Q}(\boldsymbol{f}, \boldsymbol{g}) \boldsymbol{r}(\boldsymbol{f}, \boldsymbol{g})]_s$$

για κάθε $s \in S$, όπου $\boldsymbol{Q}(\boldsymbol{f}, \boldsymbol{g})$ είναι το κατά Cesaro όριο του πίνακα $\boldsymbol{P}(\boldsymbol{f}, \boldsymbol{g})$.

Φυσικά, θα λέμε ότι $\boldsymbol{f}_0 \in F_s$ και $\boldsymbol{g}_0 \in G_s$ είναι βέλτιστες, στάσιμες στρατηγικές εάν , για κάθε $s \in S, \boldsymbol{f} \in F_s, \boldsymbol{g} \in G_s$, ισχύει ότι :

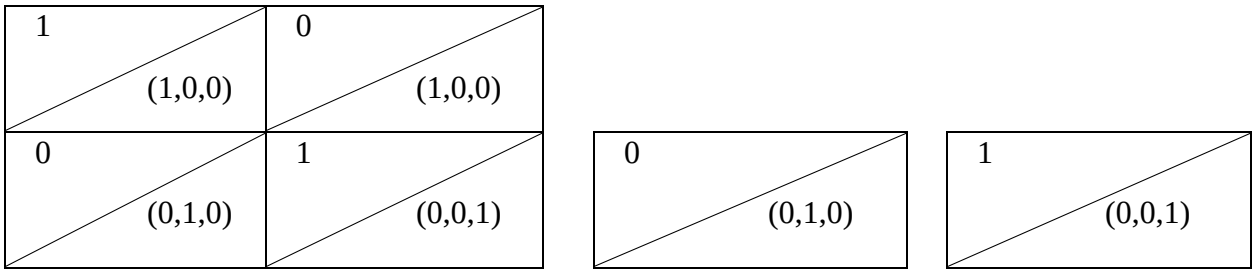
$$v_a(s, \boldsymbol{f}, \boldsymbol{g}_0) \leq v_a(s, \boldsymbol{f}_0, \boldsymbol{g}_0) \leq v_a(s, \boldsymbol{f}_0, \boldsymbol{g})$$

Εάν υπάρχει ένα τέτοιο ζεύγος στρατηγικών $(\boldsymbol{f}_0, \boldsymbol{g}_0)$, τότε υπάρχει και η μη-προεξοφλητική τιμή του παιχνιδιού Γ_α, η οποία ορίζεται να είναι:

$$\boldsymbol{v}_a \doteqdot \boldsymbol{v}_a(\boldsymbol{f}_0, \boldsymbol{g}_0)$$

Παράδειγμα 2.3.1. (The Big Match)

Έστω $S = \{1,2,3\},\ A^1(1) = A^2(1) = \{1,2\},\ A^1(s) = A^2(s) = \{1\},\ \text{για } s = 2,3$
και οι αμοιβές και πιθανότητες μετάβασης δίνονται στο ακόλουθο σχήμα:



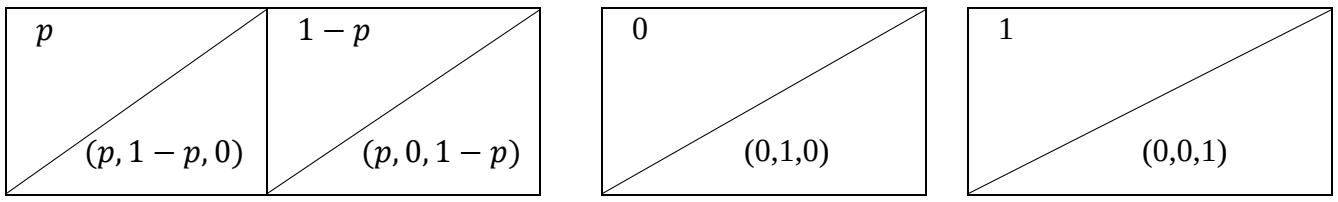
Κατάσταση 1 Κατάσταση 2 Κατάσταση 3

Βλέπουμε πως οι καταστάσεις 2 και 3 είναι απορροφητικές, συνεπώς $v_a(2) = 0$ και $v_a(3) = 1$

Έστω ότι ο παίκτης 1 διαθέτει μία βέλτιστη στάσιμη στρατηγική:

$$\boldsymbol{f}_p = \big((p, 1-p), (1), (1)\big)$$

για κάποιο σταθερό $p \in [0,1]$. Έχοντας να αντιμετωπίσει αυτή τη στρατηγική, ο παίκτης 2 στην ουσία προσπαθεί να ελαχιστοποιήσει το κόστος για την Μαρκοβιανή διαδικασία απόφασης:



Κατάσταση 1 Κατάσταση 2 Κατάσταση 3

Διαχωρίζουμε 2 περιπτώσεις:

A. $p = 1$, δηλαδή ο παίκτης 1 δεν ρισκάρει και επιλέγει να παραμείνει στην κατάσταση 1. Ωστόσο, αυτό σημαίνει ότι εναντίον της στρατηγικής $\boldsymbol{g}_0 = \big((0,1),(1),(1)\big)^T$, ο παίκτης 1 θα λαμβάνει σχεδόν πάντα 0, και έτσι, $v_a(1, \boldsymbol{f}_1, \boldsymbol{g}_0) = 0$.

B. $0 \leq p < 1$, δηλαδή ο παίκτης 1 παίρνει το ρίσκο να επιλέξει την 2η ενέργεια στην κατάσταση 1 με πιθανότητα $1 - p > 0$ κάθε φορά που η διαδικασία επισκέπτεται την 1η κατάσταση. Ωστόσο, σε αυτή την περίπτωση ο παίκτης 2 μπορεί να χρησιμοποιήσει την στρατηγική $\boldsymbol{g}_1 = \big((1,0),(1),(1)\big)^T$ και έτσι η διαδικασία θα εγκλωβιστεί στην κατάσταση 2 με πιθανότητα 1. Οπότε και εδώ πάλι θα έχουμε ότι $v_a\big(1, \boldsymbol{f}_p, \boldsymbol{g}_1\big) = 0$.

Δηλαδή, για όλα τα $p \in [0,1]$ μπορούμε να συμπεράνουμε πως

$$\min_{\boldsymbol{g} \in G_s} v_a\big(1, \boldsymbol{f}_p, \boldsymbol{g}\big) = 0$$

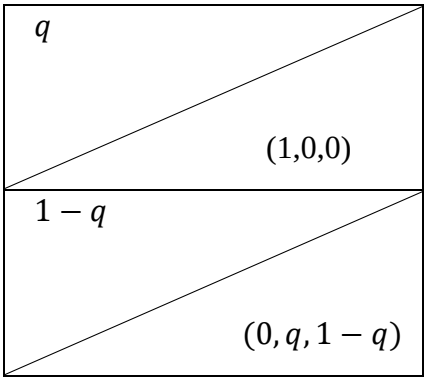
Απ' την άλλη μεριά, εάν ο παίκτης 2 χρησιμοποιήσει την στρατηγική $\boldsymbol{g}^* = \left(\left(\frac{1}{2}, \frac{1}{2}\right), (1), (1)\right)^T$, βλέπουμε πως, ανεξάρτητα από την επιλογή του παίκτη 1 στην κατάσταση 1 , θα έχουμε ότι:

$$v_a(1, \boldsymbol{f}, \boldsymbol{g}^*) = \frac{1}{2}$$

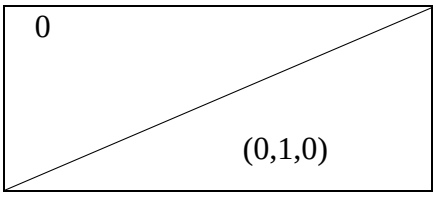
Ακόμη, παρατηρούμε ότι κάθε στάσιμη στρατηγική του παίκτη 2 μπορεί να γραφεί στη μορφή:

$$\boldsymbol{g}_q = \big((q, 1-q),(1),(1)\big)^T, \qquad q \in [0,1]$$

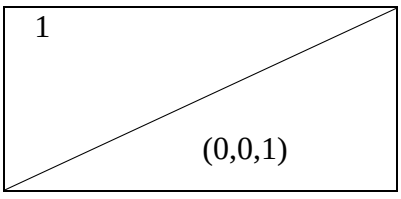
Φυσικά, όπως και πριν, για μια σταθεροποιημένη στρατηγική $\boldsymbol{g}_q$, ο παίκτης 1 καλείται να αντιμετωπίσει την διαδικασία απόφασης:



Κατάσταση 1



Κατάσταση 2

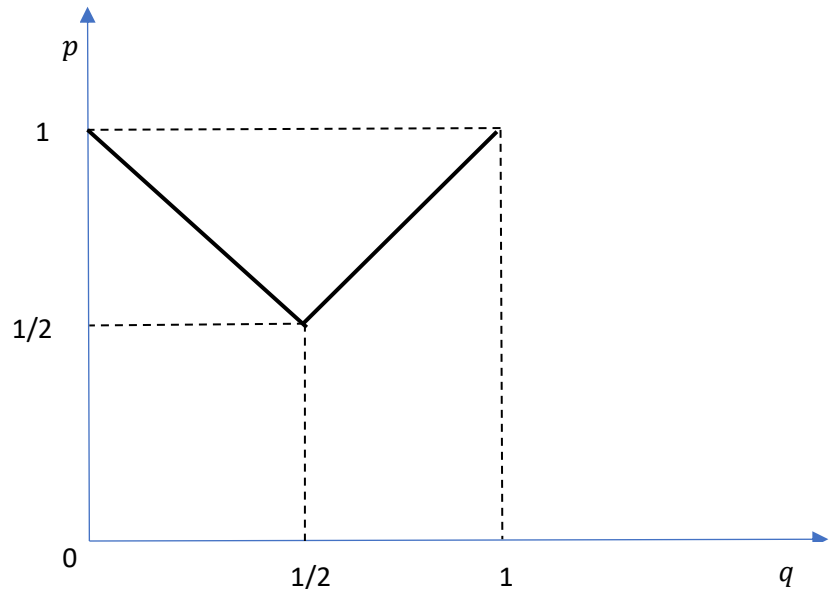


Κατάσταση 3

Παρατηρούμε πως, εάν ο παίκτης 1 χρησιμοποιεί την στρατηγική $\boldsymbol{f}_p$ με $p < 1$, τότε η απορρόφηση στις καταστάσεις 2 και 3 θα συμβεί με πιθανότητα q και $1-q$, αντίστοιχα. Εάν $p = 1$, τότε η κατάσταση 1 θα επαναλαμβάνεται συνεχώς. Συνεπώς, λαμβάνουμε ότι:

$$v_a(1, \boldsymbol{f}_p, \boldsymbol{g}_q) = \begin{cases} q, & \text{εάν } p = 1 \\ 1-q, & \text{εάν } p < 1 \end{cases}$$

Οπότε, το γράφημα της $\max_{\boldsymbol{f} \in F_s} v_a(1, \boldsymbol{f}, \boldsymbol{g}_q)$ φαίνεται παρακάτω:



Το ελάχιστο της παραπάνω συνάρτησης έχει τιμή $\frac{1}{2}$. Από την έως τώρα ανάλυση έχουμε στην ουσία δείξει ότι :

$$0 = \max_{\boldsymbol{f} \in F_s} \min_{\boldsymbol{g} \in G_s} v_a(1, \boldsymbol{f}, \boldsymbol{g}) < \frac{1}{2} = \min_{\boldsymbol{g} \in G_s} \max_{\boldsymbol{f} \in F_s} v_a(1, \boldsymbol{f}, \boldsymbol{g})$$

Ακριβώς λόγω αυτής της αυστηρής ανισοτικής σχέσης, μπορούμε να συμπεράνουμε πως δεν υπάρχουν βέλτιστες στάσιμες στρατηγικές για το Big Match.

Ακόμη και αν θέλαμε να αναζητήσουμε ένα ζεύγος βέλτιστων στρατηγικών στο, εν γένει, ευρύτερο σύνολο των Μαρκοβιανών στρατηγικών, θα βλέπαμε ότι ούτε και αυτό είναι εφικτό.

Πράγματι, έστω $\pi^1 = (\boldsymbol{f}_0, \boldsymbol{f}_1, \dots, \boldsymbol{f}_t, \dots) \in F_M$ και $\pi^2 = (\boldsymbol{g}_0, \boldsymbol{g}_1, \dots, \boldsymbol{g}_t, \dots) \in G_M$ ένα ζεύγος Μαρκοβιανών στρατηγικών των παικτών, αντίστοιχα.

Εάν η πιθανότητα για τον παίκτη 1 να επιλέξει την 2η ενέργεια είναι ίση με 1 , τότε παίρνοντας $\boldsymbol{g}_t = (1,0)$ για όλα τα t, θα έχουμε ότι $v_a(1, \pi^1, \pi^2) = 0$

Εάν όχι, τότε έστω $m \geq 0$ ο μικρότερος αριθμός από τον οποίο και έπειτα ο παίκτης 1 επιλέγει την 2η ενέργεια με θετική πιθανότητα, έστω $\varepsilon > 0$. Τώρα, ορίζουμε $\boldsymbol{\pi}_{\varepsilon}^{2}$ να είναι η στρατηγική κατά την οποία επιλέγεται η $\boldsymbol{g} = (1,0)$ για τα πρώτα m στάδια, έπειτα $\boldsymbol{g}_{m+1} = (0,1)$ και $\boldsymbol{g} = (\frac{1}{2}, \frac{1}{2})$ από εκεί και πέρα. Θα έχουμε, τότε, ότι $v_a(1, \pi^1, \pi_{\varepsilon}^{2}) = \frac{(1-\varepsilon)}{2}$.

Από την άλλη, παίρνοντας $\boldsymbol{g}_t = \left(\frac{1}{2}, \frac{1}{2}\right)$ για όλα τα t, μας δίνει $v_a(1, \pi^1, \pi^2) = \frac{1}{2}$ για όλα τα $\pi^1 \in F_M$. Εάν θεωρήσουμε την $\boldsymbol{g}_t = (q, 1-q)$ όπου $q \leq \frac{1}{2}$ για όλα τα t, τότε η π^1 για την οποία $\boldsymbol{f}_t = (1,0)$ για όλα τα t, μας δίνει ότι $v_a(1, \pi^1, \pi^2) \geq \frac{1}{2}$. Εάν τώρα σε κάποιο στάδιο n έχουμε ότι $q > \frac{1}{2}$ για πρώτη φορά, επιλέγοντας την στρατηγική $\boldsymbol{\pi}^1$ όπου $\boldsymbol{f}_t = (1,0)$ για $t < n$ και $\boldsymbol{f}_n = (0,1)$, δίνει ότι $v_a(1, \pi^1, \pi^2) > \frac{1}{2}$.

Η παραπάνω κατασκευαστική απόδειξη δείχνει πως ακόμα και στο σύνολο των Μαρκοβιανών στρατηγικών, δεν μπορούμε να βρούμε ένα ζεύγος βέλτιστων στρατηγικών για τους 2 παίκτες.

Σχόλιο

Στην πραγματικότητα, από την παραπάνω ανάλυση συμπεραίνουμε πως η τιμή του Big Match είναι ίση με $\frac{1}{2}$ και ότι ο 2ος παίκτης διαθέτει βέλτιστη στρατηγική, την $\boldsymbol{g}_t = \left(\frac{1}{2}, \frac{1}{2}\right)$. Ωστόσο, όπως δείξαμε, δεν είναι εφικτό να βρεθεί στρατηγική του παίκτη 1 που να του εξασφαλίζει $\frac{1}{2}$, αλλά μόνο ότι μπορεί να φτάσει «ε-κοντά» στην τιμή αυτή.

Στη συνέχεια, θα μας απασχολήσει να δούμε πώς μπορούν να βρεθούν βέλτιστες (στάσιμες) στρατηγικές , για εκείνα τα παιχνίδια με πληρωμή οριακού μέσου, για τα οποία υφίστανται τέτοιες στρατηγικές.

2.4 Στοχαστικά Παίγνια οριακού μέσου με έναν ελεγκτή

Το ερώτημα στο οποίο καλούμαστε να απαντήσουμε στην τελευταία παράγραφο αυτού του κεφαλαίου είναι το εξής: «Υπάρχει κάποια υποκατηγορία παιγνίων οριακού μέσου τα οποία επιδέχονται λύση με μεθόδους γραμμικού προγραμματισμού;». Εάν η απάντηση στο παραπάνω είναι θετική, τότε θα είμαστε σε θέση να αξιοποιήσουμε ήδη υπάρχοντες αλγόριθμους προκειμένου να φτάσουμε στη λύση των παιγνίων αυτών.

Η αναμενόμενη ιδέα τώρα είναι να κάνουμε την υπόθεση πως μόνον ο ένας εκ των δύο παικτών ‘ελέγχει’ τις πιθανότητες μετάβασης, ακριβώς όπως πράξαμε και για την περίπτωση με προεξόφληση.

Σκοπός μας είναι να διατυπώσουμε έναν αλγόριθμο επίλυσης στοχαστικών παιγνίων οριακού μέσου με ελεγκτή τον παίκτη 1, καθώς θα γίνει και μια προσπάθεια επιβεβαίωσης της εγκυρότητάς του.

Αρχικά, υπενθυμίζουμε πως για ένα στοχαστικό παίγνιο οριακού μέσου Γ_α και ένα ζεύγος στάσιμων στρατηγικών $(\boldsymbol{f}, \boldsymbol{g})$, η τιμή του παιχνιδιού, για αυτό το ζεύγος, είναι:

$$\boldsymbol{v}_a(\boldsymbol{f}, \boldsymbol{g}) = \boldsymbol{Q}(\boldsymbol{f}, \boldsymbol{g})\boldsymbol{r}(\boldsymbol{f}, \boldsymbol{g})$$

όπου με $\boldsymbol{Q}(\boldsymbol{f}, \boldsymbol{g})$ συμβολίζουμε το *Cesaro-όριο* του $\boldsymbol{P}(\boldsymbol{f}, \boldsymbol{g})$.

Το παίγνιο $\Gamma_\alpha(1)$ προκύπτει με φυσιολογικό τρόπο από το Γ_α, θέτοντας τον επιπλέον περιορισμό:

$$p(s'|s, a^1, a^2) \equiv p(s'|s, a^1), \text{ όπου } a^1 \in A^1(s),\ a^2 \in A^2(s),\ s, s' \in S$$

Προφανώς, ανάλογα μπορεί να οριστεί και το παίγνιο $\Gamma_\alpha(2)$.

Προτού αναπτύξουμε την θεωρία για τον σχετικό αλγόριθμο, θα χρειαστεί να παραθέσουμε τον εξής συμβολισμό:

Συμβολισμός

Για κάθε κατάσταση $s \in S$, ορίζουμε έναν $N \times m(s)$ μπλοκ πίνακα $\boldsymbol{W}_s$, του οποίου το $(s', (s, a))$ -οστό στοιχείο θα είναι:

$$w_{s'(s,a)} \doteqdot \delta_{ss'} - p(s'|s, a), \qquad s' \in S, \qquad a = 1,2, \dots, m(s) \in A(s)$$

Για αυτό το s-οστό μπλοκ, ορίζουμε τα ακόλουθα διανύσματα στήλης, διάστασης $m(s) \times 1$:

$$\boldsymbol{x}_s = (x_{s1}, x_{s2}, \dots, x_{sm(s)})^T$$

$$\boldsymbol{y}_s = (y_{s1}, y_{s2}, \dots, y_{sm(s)})^T$$

$$\boldsymbol{r}_s = (r(s,1), r(s,2), \dots, r\big(s, m(s)\big)^T$$

$$\mathbf{1}_s = (1,1, \dots,1), \qquad \mathbf{0}_s = (0,0, \dots,0)$$

και συνθέτοντας αυτά τα μπλοκ, παίρνουμε:

$$\boldsymbol{x}^T = (\boldsymbol{x}_1^T, \boldsymbol{x}_2^T, \dots, \boldsymbol{x}_N^T)$$

$$\boldsymbol{y}^T = (\boldsymbol{y}_1^T, \boldsymbol{y}_2^T, \dots, \boldsymbol{y}_N^T)$$

$$\boldsymbol{r}^T = (\boldsymbol{r}_1^T, \boldsymbol{r}_2^T, \dots, \boldsymbol{r}_N^T)$$

Ορίζουμε τους μπλοκ-πίνακες:

$$\boldsymbol{J}_1^T = (\mathbf{1}_1^T, \mathbf{0}_2^T, \dots, \mathbf{0}_N^T)$$

$$\boldsymbol{J}_2^T = (\mathbf{0}_1^T, \mathbf{1}_2^T, \dots, \mathbf{0}_N^T)$$

$$\dots$$

$$\boldsymbol{J}_N^T = (\mathbf{0}_1^T, \mathbf{0}_2^T, \dots, \mathbf{1}_N^T)$$

Επιπλέον, θα χρησιμοποιήσουμε τους $N \times m$ -διάστατους πίνακες:

$$\boldsymbol{W} \doteqdot (\boldsymbol{W}_1 \vdots \boldsymbol{W}_2 \vdots \cdots \vdots \boldsymbol{W}_N)$$

$$\boldsymbol{J} \doteqdot (\boldsymbol{J}_1 \vdots \boldsymbol{J}_2 \vdots \cdots \vdots \boldsymbol{J}_N)^T$$

καθώς και τα διανύσματα-γραμμή διάστασης $1 \times N$:

$$\boldsymbol{v}^T = \big(v(1), \dots, v(N)\big)$$

$$\boldsymbol{u}^T = \big(u(1), \dots, u(N)\big)$$

$$\boldsymbol{\gamma}^T = \big(\gamma(1), \dots, \gamma(N)\big)$$

όπου $\gamma(s) > 0$ και $\sum_{s=1}^{N} \gamma(s) = 1$

Εφόσον οι αμοιβές εξαρτώνται από τις ενέργειες και των δύο παικτών, θα ορίσουμε έναν μπλοκ-διαγώνιο πίνακα

$$\boldsymbol{R} = diag[R(1), R(2), \dots, R(N)]$$

όπου $R(s) = [r(s, a^1, a^2)]_{a^1=1, a^2=1}^{m^1(s), m^2(s)}$

Ο πίνακας R έχει διάσταση $m^1 \times m^2$, όπου με m^k συμβολίζουμε το άθροισμα του πλήθους των ενεργειών του παίκτη k ($k = 1$ ή $k = 2$) για όλες τις καταστάσεις.

Για μια στάσιμη στρατηγική του παίκτη 2 θα έχουμε:

$$\boldsymbol{Rg} = \Big[\big(R(1)\boldsymbol{g}(1)\big)^T, \big(R(2)\boldsymbol{g}(2)\big)^T, \dots, \big(R(N)\boldsymbol{g}(N)\big)^T\Big]^T$$

το οποίο είναι ένα $m^1 \times 1$ μπλοκ διάνυσμα στήλης.

Όμοια, για μια στάσιμη στρατηγική $\boldsymbol{f}$ του παίκτη 1 θα έχουμε:

$$\boldsymbol{fR} = [\boldsymbol{f}(1)R(1), \boldsymbol{f}(2)R(2), \dots, \boldsymbol{f}(N)R(N)]$$

το οποίο είναι ένα $1 \times m^2$ μπλοκ διάνυσμα γραμμής.

Είμαστε πλέον σε θέση να ορίσουμε ένα ζεύγος γραμμικών προγραμμάτων (Primal-Dual), $P_a(1)$ και $D_a(1)$, το οποίο αντιστοιχεί στο στοχαστικό παίγνιο οριακού μέσου με έναν ελεγκτή, $\Gamma_a(1)$.

$$\min [\gamma^T \boldsymbol{v}]$$

με περιορισμούς:

$P_a(1)$

a) $(\boldsymbol{u}^T, \boldsymbol{v}^T, \boldsymbol{g}^T)\begin{pmatrix} \boldsymbol{W} & \vdots & \boldsymbol{0} \\ \dots & \vdots & \dots \\ \boldsymbol{J} & \vdots & \boldsymbol{W} \\ \dots & \vdots & \dots \\ -\boldsymbol{R^T} & \vdots & \boldsymbol{0} \end{pmatrix} \geq (\boldsymbol{0}^T, \boldsymbol{0}^T)$

b) $\boldsymbol{1}^T \boldsymbol{g}(s) = 1,\ s \in S$

c) $\boldsymbol{g}(s) \geq \boldsymbol{0},\ s \in S$

Για το δυϊκό πρόγραμμα, τα διανύσματα δυϊκών μεταβλητών $\boldsymbol{x}, \boldsymbol{y}$ αντιστοιχούν στα δύο μπλοκ περιορισμών στο *α)* και η δυϊκή μεταβλητή $\boldsymbol{z}$ αντιστοιχεί στους περιορισμούς του *β)*.

$$\max[\boldsymbol{1}^T \boldsymbol{z}]$$

με περιορισμούς: $D_a(1)$

d) $\begin{pmatrix} \boldsymbol{W} & \vdots & \boldsymbol{0} \\ \dots & \vdots & \dots \\ \boldsymbol{J} & \vdots & \boldsymbol{W} \end{pmatrix}\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix} = \begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{\gamma} \end{pmatrix}$

e) $[-R^T \boldsymbol{x}]_s + z_s \boldsymbol{1}_{m^2(s)} \leq \boldsymbol{0}_{m^2(s)},\ s \in S$

f) $\boldsymbol{x} \geq \boldsymbol{0}\,, \boldsymbol{y} \geq \boldsymbol{0}$

Η ποσότητα $[-\boldsymbol{R}^T\boldsymbol{x}]_s = -R(s)^T\boldsymbol{x}(s)$ είναι ένα $(m^2(s) \times 1)$ – διάσταστο διάνυσμα για κάθε $s \in S$.

Τέλος, παρουσιάζουμε τον αλγόριθμο επίλυσης του $\Gamma_a(1)$ μέσω αυτού του ζεύγους γραμμικών προγραμμάτων.

Αλγόριθμος 2.4.1.

Βήμα 1: Βρες μια βέλτιστη λύση $(\hat{\boldsymbol{u}}^T, \hat{\boldsymbol{v}}^T, \hat{\boldsymbol{g}}^T)$ του $P_a(1)$ και μια βέλτιστη λύση $(\hat{\boldsymbol{x}}^T, \hat{\boldsymbol{y}}^T, \hat{\boldsymbol{z}}^T)$ του $D_a(1)$.

Βήμα 2: Όρισε το σύνολο καταστάσεων:

$$S^* \doteqdot \{s \in S | \hat{x}_s \doteqdot \sum_{a^1 \in A^1(s)} \hat{x}_{sa^1} > 0\}$$

Βήμα 3: Κατασκεύασε μια στάσιμη στρατηγική:

$$\hat{\boldsymbol{f}} = \left(\hat{\boldsymbol{f}}(1), \hat{\boldsymbol{f}}(2), \dots, \hat{\boldsymbol{f}}(N)\right)$$

σύμφωνα με:

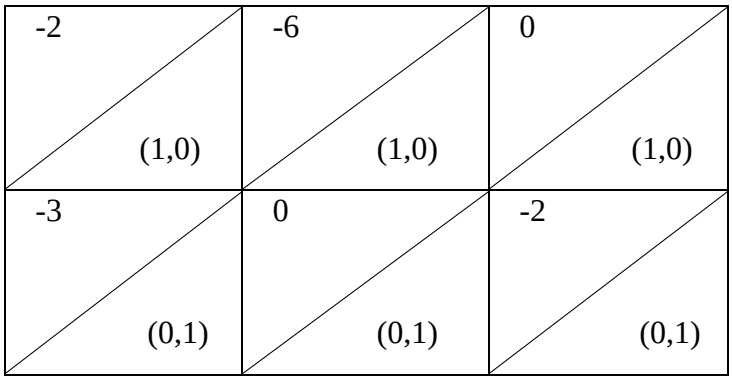
$$\hat{f}(s, a^1) = \begin{cases} \dfrac{\hat{x}_{sa^1}}{\hat{x}_s}, & s \in S^*, a^1 \in A^1(s) \\ \dfrac{\hat{y}_{sa^1}}{\hat{y}_s}, & s \in S \backslash S^*, a^1 \in A^1(s) \end{cases}$$

όπου:

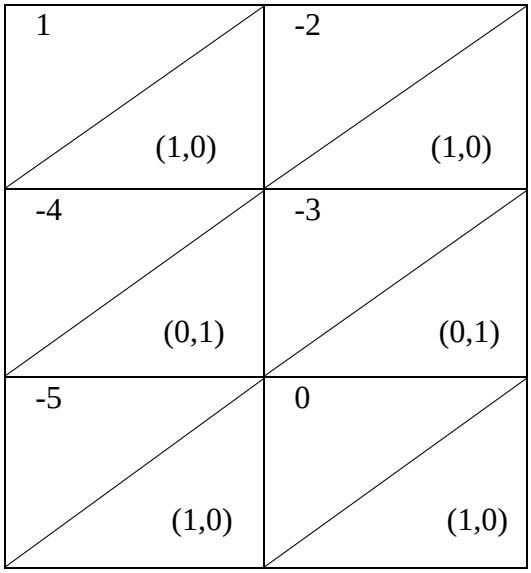
$$\hat{y}_s \doteqdot \sum_{a^1 \in A^1(s)} \hat{y}_{sa^1}$$

Παράδειγμα 2.4.1.

Έστω $S = \{1,2\}, A^1(1) = A^2(2) = \{1,2\}, A^1(2) = A^2(1) = \{1,2,3\}$, ενώ οι αμοιβές και οι πιθανότητες μετάβασης φαίνονται στο ακόλουθο σχήμα:



Κατάσταση 1



Κατάσταση 2

Παίρνουμε το διάνυσμα $\gamma^T = (\frac{1}{2}, \frac{1}{2})$

Τότε το πρωτεύον πρόβλημα ανάγεται στο :

$$\min\left[\frac{1}{2}v(1) + \frac{1}{2}v(2)\right]$$

με περιορισμούς: $P_a(1)$

$$a) \begin{pmatrix} 0 & 1 & -1 & 0 & -1 & \vdots & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 1 & \vdots & 0 & 0 & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & 1 & 0 & 0 & 0 & \vdots & 0 & 1 & -1 & 0 & -1 \\ 0 & 0 & 1 & 1 & 1 & \vdots & 0 & -1 & 1 & 0 & 1 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 2 & 3 & 0 & 0 & 0 & \vdots & 0 & 0 & 0 & 0 & 0 \\ 6 & 0 & 0 & 0 & 0 & \vdots & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & \vdots & 0 & 0 & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \vdots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & -1 & 4 & 5 & \vdots & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 3 & 0 & \vdots & 0 & 0 & 0 & 0 & 0 \end{pmatrix}^T \begin{pmatrix} u(1) \\ u(2) \\ v(1) \\ v(2) \\ g(1,1) \\ g(1,2) \\ g(1,3) \\ g(2,1) \\ g(2,2) \end{pmatrix} \geq 0$$

b) $g(1,1) + g(1,2) + g(1,3) = 1$
$g(2,1) + g(2,2) = 1$

c) $g(s, a^2) \geq 0$, $s \in \{1,2\}$, $a^2 \in A^2(s)$

Χρησιμοποιώντας το λογισμικό MATLAB, παίρνουμε τα ακόλουθα αποτελέσματα:

Αντικειμενική συνάρτηση	-2.0909
$u(1)$	0
$u(2)$	0.8409
$v(1)$	-2.0909
$v(2)$	-2.0909
g_{11}	0.9773
g_{12}	0.0227
g_{13}	0
g_{21}	0.25
g_{22}	0.75

Το αντίστοιχο δυϊκό πρόγραμμα θα είναι ως εξής:

$$\max[z(1) + z(2)]$$

με περιορισμούς: $D_a(1)$

d)
$$\begin{pmatrix} 0 & 1 & -1 & 0 & -1 & \vdots & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 1 & \vdots & 0 & 0 & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & 1 & 0 & 0 & 0 & \vdots & 0 & 1 & -1 & 0 & -1 \\ 0 & 0 & 1 & 1 & 1 & \vdots & 0 & -1 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_{11} \\ x_{12} \\ x_{21} \\ x_{22} \\ x_{23} \\ y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{23} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1/2 \\ 1/2 \end{pmatrix}$$

e) $2x_{11} + 3x_{12} + z(1) \leq 0$

$6x_{11} + z(1) \leq 0$

$2x_{12} + z(1) \leq 0$

$-x_{21} + 4x_{22} + 5x_{23} + z(2) \leq 0$

$$2x_{21} + 3x_{22} + z(2) \leq 0$$

f) $x_{ij} \geq 0, y_{ij} \geq 0$ για $i = 1,2$, $j = 1,2,3$

Χρησιμοποιώντας το λογισμικό MATLAB, παίρνουμε τα ακόλουθα αποτελέσματα:

Αντικειμενική συνάρτηση	-2.0909
$z(1)$	-1.6364
$z(2)$	-0.4545
x_{11}	0.2727
x_{12}	0.3636
x_{21}	0.2273
x_{22}	0
x_{23}	0.1364
y_{11}	0
y_{12}	0
y_{21}	0.1364
y_{22}	0
y_{23}	0

Από το Βήμα 2 του Αλγορίθμου:

$$S^* = \{1,2\} = S$$

Χρησιμοποιώντας τα παραπάνω αποτελέσματα, μπορούμε τώρα να υπολογίσουμε την στρατηγική $\hat{\boldsymbol{f}}$:

$$\hat{f}(1,1) = \frac{x_{11}}{x_{11} + x_{12}} = 0.4286$$

$$\hat{f}(1,2) = \frac{x_{12}}{x_{11} + x_{12}} = 0.5714$$

$$\hat{f}(2,1) = \frac{x_{21}}{x_{21} + x_{22} + x_{23}} = 0.625$$

$$\hat{f}(2,2) = \frac{x_{22}}{x_{21} + x_{22} + x_{23}} = 0$$

$$\hat{f}(2,3) = \frac{x_{23}}{x_{21} + x_{22} + x_{23}} = 0.375$$

Φυσικά, η βέλτιστη στρατηγική $\widehat{\boldsymbol{g}}$, για τον παίκτη 2 , δίνεται στα αποτελέσματα της λύσης του $P_a(1)$.

Συνεπώς , έχουμε καταλήξει σε ένα ζεύγος βέλτιστων στρατηγικών για το αρχικό παίγνιο

$$(\widehat{\boldsymbol{f}}, \widehat{\boldsymbol{g}}) = \Big(\big((0.4286, 0.5714), (0.625, 0, 0.375)\big), \big((0.9773, 0.0227, 0), (0.25, 0.75)\big)\Big)$$

Για το υπόλοιπο αυτής την ενότητας, θα ασχοληθούμε με τον σχολιασμό του Αλγορίθμου 2.4.1.. Αρχικά, θα δούμε γιατί ο αλγόριθμος αυτός είναι καλά ορισμένος.

Πράγματι, ας πάρουμε ένα αυθαίρετο $\boldsymbol{g} \in G_s, \boldsymbol{u} = \boldsymbol{0}$ και $\boldsymbol{v} = M\boldsymbol{1}_N$, όπου:

$$M \doteqdot max_{s,a^1,a^2}\{|r(s, a^1, a^2)|\}$$

Ισχυριζόμαστε πως θα καταλήξουμε σε μια εφικτή λύση του $P_a(1)$.

Βλέπουμε πως οι περιορισμοί *b)* και *c)* ικανοποιούνται τετριμμένα

Το δεύτερο μπλοκ περιορισμών του *α)* ανάγεται σε:

$$\boldsymbol{v}^T\boldsymbol{W} = M\boldsymbol{1}^T\boldsymbol{W} \geq \boldsymbol{0}^T$$

(λόγω της κατασκευής του $\boldsymbol{W}$, πολλαπλασιάζονται μόνο τα διαγώνια στοιχεία , τα οποία έχουν μη αρνητικό πρόσημο)

ενώ το πρώτο μπλοκ γίνεται:

$$\boldsymbol{u}^T\boldsymbol{W} + \boldsymbol{v}^T\boldsymbol{J} - \boldsymbol{g}^T\boldsymbol{R}^T = \boldsymbol{0}^T + M[\boldsymbol{1}_N^T\boldsymbol{J}] - [\boldsymbol{R}\boldsymbol{g}]^T$$

η τελευταία έκφραση θα είναι πάντα μεγαλύτερη ή ίση με $\boldsymbol{0}^T$, αφού:

$$[M\boldsymbol{1}_{m^1(s)}]^T - [R(s)\boldsymbol{g}(s)]^T \geq \boldsymbol{0}_{m^1(s)}^T \ , \quad \text{για κάθε } s \in S$$

Προκειμένου να εξασφαλιστεί η ύπαρξη πεπερασμένης βέλτιστης λύσης για το $P_a(1)$, και άρα λόγω δυϊκού θεωρήματος και για το $D_a(1)$, μένει να δείξουμε πως το $\boldsymbol{v}^T$-μπλοκ κάθε εφικτής λύσης του $P_a(1)$ είναι κάτω φραγμένο. Το αποτέλεσμα αυτό θα προκύψει ως πόρισμα της ακόλουθης πρότασης.

Πρόταση 2.4.1.

Έστω $(\overline{\boldsymbol{u}}^T, \overline{\boldsymbol{v}}^T, \overline{\boldsymbol{g}}^T)$ μια αυθαίρετη εφικτή λύση του $P_a(1)$ και $\boldsymbol{f}$ οποιαδήποτε στάσιμη στρατηγική του παίκτη 1. Τότε:

$$\overline{\boldsymbol{v}} \geq \boldsymbol{v}_a(\boldsymbol{f}, \overline{\boldsymbol{g}})$$

Απόδειξη

Αρχικά θεωρούμε το μπλοκ περιορισμών από το *α) :*

$$\overline{\boldsymbol{v}}^T \boldsymbol{W} \geq \boldsymbol{0}_{m^1}^T$$

Από τη στιγμή που το $\boldsymbol{W}$ έχει την ίδια μπλοκ δομή με την $\boldsymbol{f}$, αναμιγνύοντας το s- οστό μπλοκ του $\boldsymbol{W}$ ως προς την $\boldsymbol{f}$, παίρνουμε ότι για όλα τα $s \in S$:

$$v(s) \geq [\boldsymbol{P}(\boldsymbol{f})\overline{\boldsymbol{v}}]_s$$

ή, ισοδύναμα, σε διανυσματική μορφή:

$$\overline{\boldsymbol{v}} \geq \boldsymbol{P}(\boldsymbol{f})\overline{\boldsymbol{v}}, \quad \text{όπου } \boldsymbol{f} \in F_s$$

Ισχυριζόμαστε ότι:

$$\overline{\boldsymbol{v}} \geq \boldsymbol{P}(\boldsymbol{f})\overline{\boldsymbol{v}} \Rightarrow \overline{\boldsymbol{v}} \geq \boldsymbol{Q}(\boldsymbol{f})\overline{\boldsymbol{v}}$$

Πράγματι, έχουμε ότι:

$$\overline{\boldsymbol{v}} \geq \boldsymbol{P}(\boldsymbol{f})\overline{\boldsymbol{v}} \geq \boldsymbol{P}^2(\boldsymbol{f})\overline{\boldsymbol{v}} \geq \cdots \geq \boldsymbol{P}^k(\boldsymbol{f})\overline{\boldsymbol{v}}$$

Ακόμη, για κάθε $T \in \mathbb{N}$, παίρνουμε ότι:

$$T\overline{\boldsymbol{v}} \geq \sum_{t=1}^{T} \boldsymbol{P}^t(\boldsymbol{f})\overline{\boldsymbol{v}} \Rightarrow T\overline{\boldsymbol{v}} + \overline{\boldsymbol{v}} \geq \sum_{t=1}^{T} \boldsymbol{P}^t(\boldsymbol{f})\overline{\boldsymbol{v}} + \overline{\boldsymbol{v}} \Rightarrow \overline{\boldsymbol{v}} \geq \frac{1}{T+1}\sum_{t=0}^{T} \boldsymbol{P}^t(\boldsymbol{f})\overline{\boldsymbol{v}}$$

Παίρνοντας $T \to \infty$, προκύπτει το ζητούμενο αποτέλεσμα:

$$\overline{\boldsymbol{v}} \geq \lim_{T\to\infty}\left(\frac{1}{T+1}\sum_{t=0}^{T} \boldsymbol{P}^t(\boldsymbol{f})\overline{\boldsymbol{v}}\right) = \boldsymbol{Q}(\boldsymbol{f})\overline{\boldsymbol{v}} \qquad (2.4.1)$$

Όμοια, αναμιγνύοντας ως προς την $\boldsymbol{f}$ το s-οστό μπλοκ περιορισμών της:

$$\overline{\boldsymbol{u}}^T \boldsymbol{W} + \overline{\boldsymbol{v}}^T \boldsymbol{J} - \overline{\boldsymbol{g}}^T \boldsymbol{R}^T \geq \boldsymbol{0}_{m^1}^T$$

λαμβάνουμε ότι:

$$\bar{v}(s) + \bar{u}(s) \geq \boldsymbol{f}(s)R(s)\overline{\boldsymbol{g}}(s) + [P(\boldsymbol{f})\overline{\boldsymbol{u}}]_s$$

για κάθε $s \in S$. Ισοδύναμα, σε διανυσματική γραφή:

$$\overline{\boldsymbol{v}} + \overline{\boldsymbol{u}} \geq \boldsymbol{r}(\boldsymbol{f}, \overline{\boldsymbol{g}}) + \boldsymbol{P}(\boldsymbol{f})\overline{\boldsymbol{u}}$$

Πολλαπλασιάζοντας και τα 2 μέλη επί $\boldsymbol{Q}(\boldsymbol{f})$, η τελευταία σχέση δίνει:

$$\boldsymbol{Q}(\boldsymbol{f})\overline{\boldsymbol{v}} + \boldsymbol{Q}(\boldsymbol{f})\overline{\boldsymbol{u}} \geq \boldsymbol{Q}(\boldsymbol{f})\boldsymbol{r}(\boldsymbol{f}, \overline{\boldsymbol{g}}) + \boldsymbol{Q}(\boldsymbol{f})\boldsymbol{P}(\boldsymbol{f})\overline{\boldsymbol{u}}$$

Όμως γνωρίζουμε ότι $\boldsymbol{Q}P = \boldsymbol{Q}$, οπότε:

$$\boldsymbol{Q}(\boldsymbol{f})\overline{\boldsymbol{v}} \geq \boldsymbol{Q}(\boldsymbol{f})\boldsymbol{r}(\boldsymbol{f}, \overline{\boldsymbol{g}})$$

και χρησιμοποιώντας τη σχέση (2.4.1) παίρνουμε το ζητούμενο, δηλαδή ότι:

$$\overline{\boldsymbol{v}} \geq \boldsymbol{v}_a(\boldsymbol{f}, \overline{\boldsymbol{g}}) \qquad (2.4.2)$$

■

Πόρισμα 2.4.1

(i) Η αντικειμενική συνάρτηση $[\gamma^T \boldsymbol{v}]$ στο $P_a(1)$ είναι κάτω φραγμένη.
(ii) Τα προβλήματα $P_a(1)$ και $D_a(1)$ έχουν πεπερασμένες βέλτιστες λύσεις.

Απόδειξη

(i) Έστω $M_L \doteqdot \min_{s,a^1,a^2}\{r(s, a^1, a^2)\}$. Από τον ορισμό του κριτηρίου οριακού μέσου, μπορούμε να δούμε ότι για όλα τα $(\boldsymbol{f}, \boldsymbol{g}) \in F_{\boldsymbol{s}} \times G_s$ ισχύει:

$$\boldsymbol{v}_a(\boldsymbol{f}, \boldsymbol{g}) \geq M_L \mathbf{1}_N$$

Τότε, από την (2.5.2) , για οποιαδήποτε εφικτά $(\overline{\boldsymbol{u}}^T, \overline{\boldsymbol{v}}^T, \overline{\boldsymbol{g}}^T)$ για το $P_a(1)$, έπεται:

$$[\gamma^T \overline{\boldsymbol{v}}^T] = \sum_{s \in S} \gamma(s)\bar{v}(s) \geq M_L \left(\sum_{s \in S} \gamma(s) \right) = M_L$$

(ii) Εφόσον το $P_a(1)$ είναι ένα εφικτό και φραγμένο γραμμικό πρόγραμμα , θα επιδέχεται πεπερασμένη βέλτιστη λύση. Από το Δυϊκό θεώρημα του γραμμικού προγραμματισμού, συμπεραίνουμε πως το ίδιο ισχύει και για το $D_a(1)$.

■

Έχοντας πλέον δείξει το ζητούμενο αποτέλεσμα, συνεχίζοντας την ανάλυση του Αλγορίθμου 2.5.1., παρατηρούμε ότι το δεύτερο μπλοκ περιορισμών του *d)* , δηλαδή $J\boldsymbol{x} + W\boldsymbol{y} = \gamma$, έχει ως s-οστή είσοδο (μετά από αναδιάταξη όρων) :

$$x_{s\prime} + y_{s\prime} = \left[\sum_{s\in S}\sum_{a^1\in A^1(s)} p(s'|s,a^1)y_{sa^1}\right] + \gamma(s') \geq \gamma(s') > 0\,, \qquad s' \in S$$

Οπότε, θα έχουμε ότι $y_{s\prime} > 0$ κάθε φορά που $x_{s\prime} = 0$, και έτσι, η $\hat{f}$ του Βήματος 3 του Αλγορίθμου είναι καλώς ορισμένη.

Πρόταση 2.4.2.

Έστω $(\widehat{\boldsymbol{u}}^T, \widehat{\boldsymbol{v}}^T, \widehat{\boldsymbol{g}}^T)$ και $(\widehat{\boldsymbol{x}}^T, \widehat{\boldsymbol{y}}^T, \widehat{\boldsymbol{z}}^T)$ να είναι ένα ζεύγος δυϊκών βέλτιστων λύσεων στα $P_a(1)$ και $D_a(1)$, αντίστοιχα. Έστω, ακόμα $\hat{\boldsymbol{f}} \in F_s$ μία στάσιμη στρατηγική του παίκτη 1 και $S^* \subset S$ κατασκευασμένα όπως στο Βήμα 3 και Βήμα 2 του Αλγορίθμου, αντίστοιχα. Τότε:

(i) Το S^* είναι το σύνολο των επαναλαμβανόμενων καταστάσεων της Μαρκοβιανής αλυσίδας που επάγεται από τον $P(\hat{\boldsymbol{f}})$.

(ii) $\widehat{\boldsymbol{v}} = \boldsymbol{P}(\hat{\boldsymbol{f}})\widehat{\boldsymbol{v}} = \boldsymbol{Q}(\hat{\boldsymbol{f}})\widehat{\boldsymbol{v}}$

(iii) $\hat{v}(s) + \left[\left(I - \boldsymbol{P}(\hat{\boldsymbol{f}})\right)\widehat{\boldsymbol{u}}\right]_s = \hat{\boldsymbol{f}}(s)R(s)\widehat{\boldsymbol{g}}(s),\ s \in S^*$

Έστω τώρα $\boldsymbol{Q}(\hat{\boldsymbol{f}}) = \left[q(s'|s,\hat{\boldsymbol{f}})\right]_{s,s\prime\in S}$. Από το (i) της προηγούμενης πρότασης παίρνουμε ότι :

$$q(s'|s,\hat{\boldsymbol{f}}) = 0, \ \text{ αν } s' \notin S^*$$

Ακόμη, για όλα τα $s \in S$ έχουμε ότι:

$$v_a(s,\hat{\boldsymbol{f}},\widehat{\boldsymbol{g}}) = \left[\boldsymbol{Q}(\hat{\boldsymbol{f}})\boldsymbol{r}(\hat{\boldsymbol{f}},\widehat{\boldsymbol{g}})\right]_s = \sum_{s\prime\in S^*} q(s'|s,\hat{\boldsymbol{f}})r(s,\hat{\boldsymbol{f}},\widehat{\boldsymbol{g}})$$

Από το (iii) της πρότασης, η παραπάνω σχέση γίνεται:

$$v_a(s,\hat{\boldsymbol{f}},\widehat{\boldsymbol{g}}) = \sum_{s\prime\in S^*} q(s'|s,\hat{\boldsymbol{f}})\hat{v}(s) + \sum_{s\prime\in S^*} q(s'|s,\hat{\boldsymbol{f}})\left[\left(I - P(\hat{\boldsymbol{f}})\right)\widehat{\boldsymbol{u}}\right]_s$$

$$= [\boldsymbol{Q}(\hat{\boldsymbol{f}})\widehat{\boldsymbol{v}}]_s + \left[\boldsymbol{Q}(\hat{\boldsymbol{f}})\left(\boldsymbol{I} - \boldsymbol{P}(\hat{\boldsymbol{f}})\right)\widehat{\boldsymbol{u}}\right]_s = \hat{v}(s)$$

όπου η τελευταία ισότητα προκύπτει από το (ii) και το γεγονός ότι $\boldsymbol{Q}(\hat{\boldsymbol{f}}) = \boldsymbol{Q}(\hat{\boldsymbol{f}})\boldsymbol{P}(\hat{\boldsymbol{f}})$

Έτσι, έχουμε δείξει πως μια βέλτιστη λύση $(\widehat{\boldsymbol{u}}^T, \widehat{\boldsymbol{v}}^T, \widehat{\boldsymbol{g}}^T)$ του $P_a(1)$ ικανοποιεί την:

$$\widehat{\boldsymbol{v}} = \boldsymbol{v}_a(\hat{\boldsymbol{f}}, \widehat{\boldsymbol{g}})$$

Σε συνδυασμό με την Πρόταση 2.4.1. , έχουμε μέχρι τώρα δείξει το ένα σκέλος της συνθήκης βελτιστοποίησης, δηλαδή ότι:

$$\boldsymbol{v}_a(\boldsymbol{f}, \widehat{\boldsymbol{g}}) \leq \boldsymbol{v}_a(\hat{\boldsymbol{f}}, \widehat{\boldsymbol{g}}), \quad \text{με } \boldsymbol{f} \in F_s \qquad (2.4.3)$$

Για το δεύτερο σκέλος της συνθήκης βελτιστοποίησης θα χρειαστούμε μία τελευταία πρόταση, την οποία παραθέτουμε χωρίς απόδειξη.

Πρόταση 2.4.3.

Έστω $(\widehat{\boldsymbol{u}}^T, \widehat{\boldsymbol{v}}^T, \widehat{\boldsymbol{g}}^T)$, $(\hat{\boldsymbol{x}}^T, \hat{\boldsymbol{y}}^T, \hat{\boldsymbol{z}}^T)$ και $\hat{\boldsymbol{f}} \in F_s$ όπως είναι στη διατύπωση της Πρότασης 2.4.2. Τότε:

(i) Για όλα τα $\boldsymbol{g} \in G_s$:

$$\sum_{s \in S} \hat{z}_s = \gamma^T \boldsymbol{v}_a(\hat{\boldsymbol{f}}, \widehat{\boldsymbol{g}}) \leq \gamma^T \boldsymbol{v}_a(\hat{\boldsymbol{f}}, \boldsymbol{g})$$

(ii) Για όλα τα $\boldsymbol{g} \in G_s$:

$$\boldsymbol{v}_a(\hat{\boldsymbol{f}}, \widehat{\boldsymbol{g}}) \leq \boldsymbol{v}_a(\hat{\boldsymbol{f}}, \boldsymbol{g})$$

Φυσικά, το (ii) της Πρότασης μαζί με την (2.4.3) συμπληρώνουν τα δύο μέλη της ζητούμενης ανισότητας. Έχουμε καταλήξει δηλαδή στην :

$$\boldsymbol{v}_a(\boldsymbol{f}, \widehat{\boldsymbol{g}}) \leq \boldsymbol{v}_a(\hat{\boldsymbol{f}}, \widehat{\boldsymbol{g}}) \leq \boldsymbol{v}_a(\hat{\boldsymbol{f}}, \boldsymbol{g})$$

για όλα τα $\boldsymbol{f} \in F_s$ και $\boldsymbol{g} \in G_s$.

Κεφάλαιο 3

Ημιμαρκοβιανά παίγνια τέλειας πληροφόρησης

3.0 Εισαγωγή

Το τελευταίο κεφάλαιο της εργασίας αυτής πραγματεύεται ημιμαρκοβιανά παίγνια τέλειας πληροφόρησης. Πρόκειται για μια γενίκευση των "κλασικών" στοχαστικών παιγνίων, όπου πλέον θεωρούμε ότι η επιλογή ενεργειών από τους παίκτες συνεπάγεται και έναν χρόνο παραμονής στην παρούσα κατάσταση, προτού πραγματοποιηθεί μετάβαση στην επόμενη κατάσταση και ούτω καθεξής. Ο χρόνος παραμονής θα προέρχεται από γνωστή κατανομή, χωρίς να επεκταθούμε παραπάνω στο συγκεκριμένο ζήτημα, εφόσον το ενδιαφέρον μας στρέφεται γύρω από τη μέση τιμή της παραμονής αυτής. Πιο συγκεκριμένα, θα γίνει ανάλυση των παιγνίων που διαθέτουν το χαρακτηριστικό της "τέλειας πληροφόρησης". Αφού δοθεί το κατάλληλο θεωρητικό υπόβαθρο σχετικά με αυτήν την κατηγορία παιγνίων, παρατίθεται ο αλγόριθμος επίλυσης ενός τέτοιου παιγνίου μέσω γραμμικού προγραμματισμού, το οποίο πρώτα έχει υποβιβαστεί σε ημιμαρκοβιανή διαδικασία απόφασης. Τέλος, γίνεται εφαρμογή του συγκεκριμένου αλγορίθμου σε αριθμητικό παράδειγμα, το οποίο επιλύουμε με δύο διαφορετικούς τρόπους.

3.1 Πεπερασμένα ημιμαρκοβιανά παίγνια δύο παικτών και μηδενικού αθροίσματος

Ας ξεκινήσουμε από τον ορισμό ενός ημιμαρκοβιανού παιγνίου 2 παικτών και μηδενικού αθροίσματος, όπου τόσο ο χώρος καταστάσεων όσο και ο χώρος των διαθέσιμων ενεργειών για κάθε παίκτη, είναι πεπερασμένα σύνολα. Ορίζουμε ,λοιπόν, ένα τέτοιο παίγνιο Γ , ως μία συλλογή από 7 αντικείμενα :

$\Gamma = < S, \{A^1(s) \mid s \in S\}, \{A^2(s) \mid s \in S\}, p, Q, r >$, όπου:

$S = \{1,2, \dots, N\}$ είναι το μη-κενό πεπερασμένο σύνολο καταστάσεων

$A^1(s) = \{1,2, \dots, m^1(s)\}$ είναι το μη-κενό πεπερασμένο σύνολο των επιτρεπτών ενεργειών για τον παίκτη 1 στην κατάσταση s

$A^2(s) = \{1,2, \dots, m^2(s)\}$ είναι το μη-κενό πεπερασμένο σύνολο των επιτρεπτών ενεργειών για τον παίκτη 2 στην κατάσταση s

$K = \{(s, a^1, a^2) \mid s \in S, a^1 \in A^1(s), a^2 \in A^2(s)\}$ είναι το σύνολο που περιλαμβάνει όλες τις επιτρεπτές 3-αδες καταστάσεων-ενεργειών των 2 παικτών

$p(. \mid s, a^1, a^2)$ είναι η πιθανότητα μετάβασης στην επόμενη κατάσταση (νόμος της κίνησης)

$Q_{ss'}(. \mid a^1, a^2)$ είναι μια κατανομή πιθανότητας στο $[0, \infty)$, δεδομένου του $K \times S$, την οποία θα καλούμε κατανομή των ενδιάμεσων χρόνων παραμονής

r είναι η πραγματική συνάρτηση πληρωμής ενός βήματος για τον πάικτη 1, που ορίζεται πάνω στο K (η συνάρτηση πληρωμής ενός βήματος για τον παίκτη 2 θα είναι η $r_2(s, a^1, a^2) = -r(s, a^1, a^2)$)

Το ημιμαρκοβιανό παίγνιο εξελίσσεται επ' άπειρον ως εξής:
Στην μηδενική εποχή απόφασης, το παιχνίδι ξεκινάει από την αρχική κατάσταση $s_0 \in S$ και οι παίκτες 1 και 2 ταυτόχρονα και ανεξάρτητα ο ένας απ' τον άλλο, επιλέγουν ενέργειες.

$a_0^1 \in A^1(s_0)$ και $a_0^2 \in A^2(s_0)$. Συνεπώς, οι παίκτες λαμβάνουν αμοιβές ενός βήματος $r(s_0, a_0^1, a_0^2)$ και $r_2(s_0, a_0^1, a_0^2)$, αντίστοιχα και το παίγνιο μεταβαίνει στην επόμενη κατάσταση s_1 με πιθανότητα μετάβασης $p(s_1 \mid s_0, a_0^1, a_0^2)$. Ο χρόνος παραμονής στην κατάσταση s_0 έως ότου γίνει η μετάβαση στην κατάσταση s_1 ορίζεται από τη συνάρτηση κατανομής $Q_{s_0 s_1}(. \mid a_0^1, a_0^2)$. Εφόσον φτάσουμε στην κατάσταση s_1 στην επόμενη εποχή απόφασης, επαναλαμβάνεται η ίδια διαδικασία με την κατάσταση s_0 να έχει πλέον αντικατασταθεί από την s_1.

Στρατηγικές

Στο 1ο κεφάλαιο ορίσαμε μία *συμπεριφορική* στρατηγική για τον παίκτη 1 να είναι μια ακολουθία

$$\boldsymbol{f} \div (f_0, f_1, f_2, \dots, f_t, \dots)$$

όπου τα f_t είναι κανόνες απόφασης από το σύνολο των ιστοριών της διαδικασίας, H_t, στο $P(A^1)$, το οποίο είναι το σύνολο των κατανομών πιθανότητας πάνω στο πεπερασμένο σύνολο ενεργειών A^1.

Με όμοιο τρόπο μπορεί να οριστεί η συμπεριφορική στρατηγική $\boldsymbol{g}$ για τον παίκτη 2.

Οι χώροι όλων των συμπεριφορικών στρατηγικών των παικτών θα συμβολίζονται με F_B^1 και F_B^2, αντίστοιχα. Οι συμπεριφορικές στρατηγικές αποτελούν την πιο ευρεία κλάση στρατηγικών που θα μας απασχολήσουν.

Εν συνεχεία, θα καλούμε μία συμπεριφορική στρατηγική $\boldsymbol{f} = \{f_t\}_{t=0}^{\infty}$ του παίκτη 1 *ημιμαρκοβιανή στρατηγική* εάν για κάθε t, το f_t εξαρτάται από τα s_0, s_t και την εποχή απόφασης t. Ανάλογα ορίζεται και η *ημιμαρκοβιανή στρατηγική* $\boldsymbol{g} = \{g_t\}_{t=0}^{\infty}$ για τον παίκτη 2.

Όπως έχουμε ήδη δει, μία *στάσιμη στρατηγική* θα εξαρτάται μόνο από την παρούσα κατάσταση. Δηλαδή, μία στάσιμη στρατηγική του παίκτη 1 (ανάλογα και για τον παίκτη 2) ορίζεται ως μία N-άδα,

$\boldsymbol{f} = (\boldsymbol{f}(1), \boldsymbol{f}(2), \dots, \boldsymbol{f}(N))$, όπου $\boldsymbol{f}(s) = \left(f(s,1), f(s,2), \dots, f\left(s, m^1(s)\right)\right)$ και το $f(s,i)$ δηλώνει την πιθανότητα να επιλεγεί η ενέργεια i στην κατάσταση s. Θα συμβολίζουμε με F_s^1και F_s^2 τα σύνολα των στάσιμων στρατηγικών των παικτών, αντίστοιχα.

Υπενθυμίζουμε πως μία στάσιμη στρατηγική θα καλείται *καθαρή* εάν σε κάθε κατάσταση επιλέγεται μία ενέργεια με πιθανότητα 1 (και όλες οι υπόλοιπες με πιθανότητα 0). Θα συμβολίζουμε με F_{sp}^1 και F_{sp}^2 τα σύνολα των στάσιμων καθαρών στρατηγικών των παικτών , αντίστοιχα.

Μία *ημιστάσιμη στρατηγική* είναι μία ημιμαρκοβιανή στρατηγική, η οποία είναι ανεξάρτητη της εποχής απόφασης t. Θα συμβολίζουμε με ξ_1 και ξ_2τα σύνολα των ημιστάσιμων στρατηγικών των παικτών.

Ορισμός 3.1.1.

Ένα ημιμαρκοβιανό παίγνιο μηδενικού αθροίσματος και δύο παικτών $\Gamma = < S, \{A^1(s) \mid s \in S\}, \{A^2(s) \mid s \in S\}, p, Q, r >$ θα καλείται παίγνιο *τέλειας πληροφόρησης*, εάν ισχύουν τα ακόλουθα :

(i) $S = S_1 \cup S_2,\ S_1 \cap S_2 = \emptyset$
(ii) $|A^2(s)| = 1$, για όλα τα $s \in S_1$, δηλαδή στο S_1 ο παίκτης 2 έχει μόνο μία διαθέσιμη ενέργεια
(iii) $|A^1(s)| = 1$, για όλα τα $s \in S_2$, δηλαδή στο S_2 ο παίκτης 1 έχει μόνο μία διαθέσιμη ενέργεια

3.2. Ημιμαρκοβιανά παίγνια με κριτήριο πληρωμής τον οριακό λόγο των μέσων

Έστω $(X_1, A_1^1, A_1^2, X_2, A_2^1, A_2^2, \dots)$ μία διατεταγμένη ακολουθία στο $S \times (A^1 \times A^2 \times S)^\infty$. Δοθέντος ζεύγους συμπεριφορικών στρατηγικών $(\boldsymbol{f}, \boldsymbol{g}) \in F_B^1 \times F_B^2$ και αρχικής κατάστασης s, ορίζουμε ως αναμενόμενη πληρωμή τον οριακό λόγο των μέσων ,για τον παίκτη 1, να είναι :

$$v_a(s, \boldsymbol{f}, \boldsymbol{g}) = \lim_{t\to\infty} inf \frac{E_{\boldsymbol{f},\boldsymbol{g}}[\sum_{m=1}^{t}(r(X_m, A_m^1, A_m^2) \mid X_1 = s]}{E_{\boldsymbol{f},\boldsymbol{g}}[\sum_{m=1}^{t}(\bar{\tau}(X_m, A_m^1, A_m^2) \mid X_1 = s]} \qquad (3.2.1)$$

Όπου το $\bar{\tau}(s, a^1, a^2) = \sum_{s' \in S} \left(p(s' \mid a^1, a^2) \int_0^\infty t\, dQ_{ss'}(t \mid a^1, a^2) \right)$ δηλώνει τον αναμενόμενο χρόνο παραμονής στην κατάσταση s για το ζευγάρι ενεργειών $(a^1, a^2) \in A^1(s) \times A^2(s)$.

Ορισμός 3.2.1.

Για ένα ζεύγος στάσιμων στρατηγικών $(\boldsymbol{f},\boldsymbol{g}) \in F_s^1 \times F_s^2$ ορίζουμε τον πίνακα πιθανοτήτων μετάβασης ως:

$$\boldsymbol{P}(\boldsymbol{f},\boldsymbol{g}) = [p(s'|\, s,\boldsymbol{f},\boldsymbol{g})]_{N\times N}$$

όπου

$$p(s'|\, s,\boldsymbol{f},\boldsymbol{g}) = \sum_{a^1\in A^1(s)} \sum_{a^2\in A^2(s)} p(s'|\, s,a^1,a^2) \text{ δηλώνει την πιθανότητα}$$

ότι, ξεκινώντας από την κατάσταση s, η επόμενη κατάσταση θα είναι η s' για το ζεύγος στρατηγικών $(\boldsymbol{f},\boldsymbol{g})$ των παικτών, αντίστοιχα.

Για τις στάσιμες στρατηγικές $(\boldsymbol{f},\boldsymbol{g})$, θα γράφουμε την συνολική πληρωμή των παικτών ως :

$$v_a(s,\boldsymbol{f},\boldsymbol{g}) = \liminf_{t\to\infty} \frac{\sum_{m=1}^{t} r^m(\, s,\boldsymbol{f},\boldsymbol{g})}{\sum_{m=1}^{t} \bar{\tau}^m(\, s,\boldsymbol{f},\boldsymbol{g})} \quad \text{, για κάθε } s \in S.$$

όπου τα $r^m(s,\boldsymbol{f},\boldsymbol{g})$ και $\bar{\tau}^m(s,\boldsymbol{f},\boldsymbol{g})$ είναι αντίστοιχα η αναμενόμενη αμοιβή για τον παίκτη 1 και ο αναμενόμενος χρόνος παραμονής στην m-οστή εποχή απόφασης, όταν βρισκόμαστε στην κατάσταση s και έχουν επιλεγεί οι στρατηγικές $\boldsymbol{f},\boldsymbol{g}$ από τους παίκτες. Ορίζουμε :

$$r(\boldsymbol{f},\boldsymbol{g}) = [r(s,\boldsymbol{f},\boldsymbol{g})]_{N\times 1}\,, \qquad \bar{\tau}(\boldsymbol{f},\boldsymbol{g}) = [\bar{\tau}(s,\boldsymbol{f},\boldsymbol{g})]_{N\times 1}, \qquad v_a(\boldsymbol{f},\boldsymbol{g}) = [v_a(s,\boldsymbol{f},\boldsymbol{g})]_{N\times 1}$$

ως τα διανύσματα-στήλες της αναμενόμενης αμοιβής ενός βήματος, του αναμενόμενου χρόνου παραμονής και της αναμενόμενης πληρωμής του οριακού λόγου των μέσων αντίστοιχα, για το ζεύγος στρατηγικών $(\boldsymbol{f},\boldsymbol{g}) \in F_s^1 \times F_s^2$. Έχουμε, ακόμη, ότι:

$$r^m(s,\boldsymbol{f},\boldsymbol{g}) = \sum_{s'\in S} \mathbb{P}_{\boldsymbol{f},\boldsymbol{g}}(X_m = s'|X_1 = s)\, r(s',\boldsymbol{f},\boldsymbol{g}) = \sum_{s'\in S} r(s',\boldsymbol{f},\boldsymbol{g})\, p^{m-1}(s'|s,\boldsymbol{f},\boldsymbol{g})$$

$$= [\boldsymbol{P}^{m-1}(\boldsymbol{f},\boldsymbol{g}) r(\boldsymbol{f},\boldsymbol{g})](s)$$

$$\bar{\tau}^m(s,\boldsymbol{f},\boldsymbol{g}) = \sum_{s'\in S} \mathbb{P}_{\boldsymbol{f},\boldsymbol{g}}\left((X_m = s'|X_1 = s)\bar{\tau}(s',\boldsymbol{f},\boldsymbol{g}\right) = \sum_{s'\in S} \bar{\tau}(s',\boldsymbol{f},\boldsymbol{g})\, p^{\boldsymbol{m-1}}(s'|s,\boldsymbol{f},\boldsymbol{g})$$

$$= [\boldsymbol{P}^{m-1}(\boldsymbol{f},\boldsymbol{g})\bar{\tau}(\boldsymbol{f},\boldsymbol{g})](s)$$

Στο σημείο αυτό παραθέτουμε το ακόλουθο Λήμμα, όπως αυτό είναι διατυπωμένο στο [7] (σελ.175)

Λήμμα 3.2.1.

Έστω $\boldsymbol{P} = [p(s'|s)]_{n\times n}$ να είναι ένας πίνακας πιθανοτήτων μετάβασης . Τότε υπάρχει στοχαστικός πίνακας $\boldsymbol{P}^* = [p^*(s'|s)]_{n\times n}$, τον οποίο θα καλούμε *όριο κατά Cesaro* του πίνακα $\boldsymbol{P}$ και ο οποίος ορίζεται να είναι :

$$p^*(s'|s) = \lim_{m\to\infty} \frac{1}{k} \sum_{k=1}^{m} p^k(s'|s) \quad , \qquad s, s' = 1,2, \dots, n$$

Εφόσον ο πίνακας $\boldsymbol{P}(\boldsymbol{f}, \boldsymbol{g})$ είναι ένας τέτοιος πίνακας πιθανοτήτων μετάβασης, έχουμε ότι το όριο

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

υπάρχει και είναι ίσο με $\boldsymbol{P}^*(\boldsymbol{f}, \boldsymbol{g})$, όπου με $\boldsymbol{P}^*(\boldsymbol{f}, \boldsymbol{g})$ συμβολίζουμε τον πίνακα που προκύπτει ως όριο κατά Cesaro από τον $\boldsymbol{P}(\boldsymbol{f}, \boldsymbol{g})$

Επιπλέον έχουμε ότι:

$$\lim_{t\to\infty} \frac{1}{t} \sum_{m=1}^{t} r^m(\boldsymbol{f}, \boldsymbol{g}) = [\boldsymbol{P}^*(\boldsymbol{f}, \boldsymbol{g})\boldsymbol{r}(\boldsymbol{f}, \boldsymbol{g})](s)$$

και

$$\lim_{t\to\infty} \frac{1}{t} \sum_{m=1}^{t} \bar{\tau}^m(\boldsymbol{f}, \boldsymbol{g}) = [\boldsymbol{P}^*(\boldsymbol{f}, \boldsymbol{g})\bar{\tau}(\boldsymbol{f}, \boldsymbol{g})](s)$$

Οπότε, τελικά, για ένα ζεύγος στάσιμων στρατηγικών $(\boldsymbol{f}, \boldsymbol{g}) \in F_s^1 \times F_s^2$, θα έχουμε ότι :

$$v_a(s, \boldsymbol{f}, \boldsymbol{g}) = \frac{[\boldsymbol{P}^*(\boldsymbol{f}, \boldsymbol{g})\boldsymbol{r}(\boldsymbol{f}, \boldsymbol{g})](s)}{[\boldsymbol{P}^*(\boldsymbol{f}, \boldsymbol{g})\bar{\boldsymbol{\tau}}(\boldsymbol{f}, \boldsymbol{g})](s)} \quad \text{, για όλα τα } s \in S \qquad (3.2.2)$$

Ορισμός 3.2.2.

Θα λέμε ότι το ημιμαρκοβιανό παίγνιο με πληρωμές τον οριακό λόγο των μέσων , δύο-παικτών και μηδενικού αθροίσματος, θα έχει διάνυσμα *τιμής* $\boldsymbol{v}_a = [v_a(s)]_{N\times 1}$ αν

$$sup_{\boldsymbol{f}\in \boldsymbol{F}_b^1}\, inf_{\boldsymbol{g}\in \boldsymbol{F}_b^2}\; v_a(s, \boldsymbol{f}, \boldsymbol{g}) = v_a(s) = inf_{\boldsymbol{g}\in \boldsymbol{F}_b^2}\, sup_{\boldsymbol{f}\in \boldsymbol{F}_b^1} v_a(s, \boldsymbol{f}, \boldsymbol{g}) \quad \text{, για όλα τα } s \in S$$

Ένα ζεύγος στρατηγικών $(\boldsymbol{f}^*, \boldsymbol{g}^*) \in \boldsymbol{F_b^1} \times \boldsymbol{F_b^2}$ θα λέμε ότι είναι *βέλτιστο* για τους παίκτες, εάν

$$v_a(s, \boldsymbol{f}^*, \boldsymbol{g}) \geq v_a(s) \geq v_a(s, \boldsymbol{f}, \boldsymbol{g}^*) \text{ , για όλα τα } s \in S \text{ και όλα τα } (\boldsymbol{f}, \boldsymbol{g}) \in \boldsymbol{F_b^1} \times \boldsymbol{F_b^2}$$

3.3 Πεπερασμένες ημιμαρκοβιανές διαδικασίες απόφασης

Κατ' αναλογία με την προηγούμενη παράγραφο καθώς και με την σχετική παράγραφο του κεφαλαίου 1 για μαρκοβιανές διαδικασίες, θα περιγράψουμε εν συντομία την διαδικασία εξέλιξης μιας τέτοιας ημιμαρκοβιανής διαδικασίας.

Ορίζουμε λοιπόν μια τέτοια πεπερασμένη διαδικασία (πεπερασμένος χώρος καταστάσεων και ενεργειών) ως μία συλλογή 5 αντικειμένων ως εξής :

$$\hat{\Gamma} = < S, \hat{A} = \{A(s) | \ s \in S\}, \hat{p}, \hat{Q}, \hat{r} >$$

Ο συμβολισμός που θα χρησιμοποιήσουμε ταυτίζεται με τον αντίστοιχο για ημιμαρκοβιανά παίγνια. Θα χρησιμοποιούμε τον συμβολισμό $(\hat{\cdot})$ για να δηλώσουμε την διαφοροποίηση από τον αντίστοιχο συμβολισμό των παιγνίων.

Η διαδικασία και πάλι εξελίσσεται στο διηνεκές, ακριβώς όπως και στο ανάλογο παίγνιο , με τη μόνη διαφορά να είναι πως τώρα οι δύο παίκτες αντικαθίστανται από έναν που λαμβάνει αποφάσεις.

Φυσικά οι ορισμοί που αφορούν τους χώρους στρατηγικών είναι ακριβώς οι ίδιοι με εκείνους των ημιμαρκοβιανών παιχνιδιών. Θα συμβολίζουμε με F_B, F_S, F_{SP}, ξ τα σύνολα των συμπεριφορικών, στάσιμων, καθαρών στάσιμων και ημιστάσιμων στρατηγικών, αντίστοιχα, του decision maker.

Για μια συμπεριφορική στρατηγική $\boldsymbol{f} \in F_B$, η αναμενόμενη πληρωμή του οριακού λόγου των μέσων ορίζεται από την:

$$\widehat{v_a}(s, \boldsymbol{f}) = \liminf_{t \to \infty} \frac{E_{\boldsymbol{f}}[\sum_{m=1}^{t} (\hat{r}(X_m, A_m) | \ X_1 = s]}{E_{\boldsymbol{f}}[\sum_{m=1}^{t} (\bar{\tau}(X_m, A_m) | \ X_1 = s]} \text{ , για όλα τα } s \in S \qquad (3.3.1)$$

όπου

$$\bar{\tau}(s, a) = \sum_{s' \in S} \left(\hat{p}(s' | a) \int_0^{\infty} t \, d\hat{Q}_{ss'}(t \mid a) \right)$$

είναι ο αναμενόμενος χρόνος παραμονής στην κατάσταση s όταν ο decision-maker έχει επιλέξει την ενέργεια $a \in \hat{A}(s)$.

Παρατήρηση

Ο τρόπος που ορίσαμε την αναμενόμενη πληρωμή των παικτών, μέσω της (3.2.1) για τα παίγνια και μέσω της (3.3.1) για διαδικασίες απόφασης, προφανώς δεν είναι μοναδικός. Ωστόσο, είναι αυτός που απαντάται συχνότερα στη βιβλιογραφία και θα παραθέσουμε ορισμένους λόγους για τους οποίους αυτό συμβαίνει. Ως εναλλακτική συνάρτηση πληρωμής των παικτών θα μπορούσαμε να θεωρήσουμε την:

$$\varphi(s, \boldsymbol{f}) = E_{\boldsymbol{f}}\left[\liminf_{\boldsymbol{n}\to\infty} \frac{\sum_{m=0}^{n} r(X_m, A_m)}{\sum_{m=0}^{n} \tau(X_m, A_m)} | X_0 = s\right]$$

ή και την :

$$\psi(s, \boldsymbol{f}) = \liminf_{\boldsymbol{n}\to\infty} E_{\boldsymbol{f}}\left[\frac{\sum_{m=0}^{n} r(X_m, A_m)}{\sum_{m=0}^{n} \tau(X_m, A_m)} | X_0 = s\right]$$

Ο κυριότερος λόγος που επιλέγουμε η πληρωμή να ορίζεται από την σχέση (3.3.1), έναντι των φ και ψ, είναι πως γνωρίζουμε την ύπαρξη βέλτιστης ημι-στάσιμης στρατηγικής για αυτό το κριτήριο, από την εργασία [14] , ενώ κάτι τέτοιο δεν είναι ακόμα γνωστό για τα δύο εναλλακτικά κριτήρια πληρωμής. Επιπλέον, η σχέση (3.3.1) φαίνεται να είναι πιο 'εύκολη' στον χειρισμό , εφόσον τόσο ο αριθμητής όσο και ο παρονομαστής μπορούν να υπολογιστούν ακολουθώντας τη λογική μιας Μαρκοβιανής διαδικασίας πεπερασμένου ορίζοντα, όπως αυτή ορίστηκε στο 1ο Κεφάλαιο, και έπειτα η συνολική αναμενόμενη πληρωμή θα προκύπτει ως το $liminf$ του πηλίκου που υπολογίστηκε. Απ'την άλλη, κάτι τέτοιο δεν είναι εφικτό για την φ, αφού θα πρέπει να υπολογιστεί το $liminf$ για κάθε άπειρου μήκους ιστορία της διαδικασίας και έπειτα να πάρουμε την μέση τιμής αυτής, ενώ για το κριτήριο ψ θα πρέπει να υπολογιστεί η μέση τιμή για μια άπειρη ακολουθία τυχαίων μεταβλητών, ώστε να πάρουμε το $liminf$ αυτής.

Τέλος, ένας εναλλακτικός ορισμός θα ήταν να αντικαταστήσουμε στη σχέση (3.3.1) (και στην (3.2.1)) το $liminf$ με $limsup$. Αυτό δεν θα είχε καμία διαφορά στην κλάση των στάσιμων στρατηγικών (επειδή το συγκεκριμένο όριο υπάρχει) , αλλά αποτελεί, εν γένει, ένα διαφορετικό κριτήριο πληρωμής για την ευρύτερη κλάση των συμπεριφορικών στρατηγικών.

Ορισμός 3.3.1

Μία στρατηγική $\boldsymbol{f}^*$θα καλείται *βέλτιστη*, με κριτήριο πληρωμής τον οριακό λόγο των μέσων, εάν

$$\widehat{v_a}(s, \boldsymbol{f}^*) \geq \widehat{v_a}(s, \boldsymbol{f})$$

για όλα τα $\boldsymbol{f} \in F_B$ και όλα τα $s \in S$. Έστω

$$\widehat{v_a}(s) = sup_{\boldsymbol{f}\in F_B}\{\widehat{v_a}(s, \boldsymbol{f})\}$$

Τότε, η τιμή $\widehat{v_a}(s)$ θα καλείται η *τιμή* της ημιμαρκοβιανής διαδικασίας απόφασης για την αρχική κατάσταση s, ενώ το διάνυσμα

$$\widehat{\boldsymbol{v}_a} = [\widehat{v_a}(s)]_{s\in S}$$

θα είναι το *διάνυσμα τιμής* για την διαδικασία.

Θεώρημα 3.3.1

Σε μια πεπερασμένη ημιμαρκοβιανή διαδικασία απόφασης με κριτήριο πληρωμής τον οριακό λόγο των μέσων, υπάρχει καθαρή ημιστάσιμη στρατηγική $\boldsymbol{f}_P^* \in \xi$, η οποία είναι βέλτιστη για τον decision-maker.

Το κυριότερο αποτέλεσμα αυτού του κεφαλαίου συνοψίζεται στο επόμενο σημαντικό θεώρημα. Αρχικά διατυπώνουμε το θεώρημα και έπειτα παραθέτουμε την απόδειξή του. Η συλλογιστική πορεία της απόδειξης θα φανεί χρήσιμη στη συνέχεια, όταν θα είμαστε σε θέση να εφαρμόσουμε τον αλγόριθμο επίλυσης τέτοιου τύπου παιγνίων.

Θεώρημα 3.3.2

Οποιοδήποτε ημιμαρκοβιανό παίγνιο δύο παικτών μηδενικού αθροίσματος και τέλειας πληροφόρησης, με κριτήριο πληρωμής τον οριακό λόγο των μέσων, έχει λύση στο σύνολο των καθαρών ημιστάσιμων στρατηγικών.

Απόδειξη

Θα δείξουμε το παραπάνω θεώρημα δείχνοντας την στρατηγική ισοδυναμία μεταξύ του ημιμαρκοβιανού παιγνίου και της αντίστοιχης ημιμαρκοβιανής διαδικασίας απόφασης. Έστω λοιπόν το ημιμαρκοβιανό παίγνιο

$$\Gamma =< S = S_1 \cup\ S_2, A^1 = \{A^1(s) \mid s \in S_1\}, A^2 = \{A^2(s) \mid s \in S_2\}, p, Q, r >$$

δύο παικτών, μηδενικού αθροίσματος και τέλειας πληροφόρησης με πληρωμή που προκύπτει από τον οριακό λόγο των μέσων, όπου σε $|S_1|$ στο πλήθος καταστάσεις ο παίκτης 2 διαθέτει μόνο μία ενέργεια για να επιλέξει, ενώ το ίδιο ισχύει για τον παίκτη 1 για τις καταστάσεις $\{|S_1| + 1, |S_1| + 2, \dots, |S_1| + |S_2|\}$. Από αυτό το ημιμαρκοβιανό παίγνιο θα κατασκευάσουμε μία διαδικασία απόφασης $\hat{\Gamma}$ με τον εξής τρόπο. Έστω :

$$\hat{\Gamma} =< S = S_1 \cup\ S_2, \hat{A} = \left\{\hat{A}(s) = A^1(s) \,\middle|\, s \in S_1\right\} \cup \left\{\hat{A}(s) = A^2(s) \,\middle|\, s \in S_2\right\}, \hat{p} = p, \hat{Q} = Q, \hat{r} >$$

όπου το $\hat{r}(s, a)$ ορίζεται ως:

$$\hat{r}(s,a) = \begin{cases} r(s,a)\text{ , εάν } s \in S_1\text{ , } a \in \hat{A}(s) = A^1(s) \\ -r(s,a)\text{ , εάν } s \in S_2\text{ , } a \in \hat{A}(s) = A^2(s) \end{cases}$$

Από το Θεώρημα 3.3.1 γνωρίζουμε πως υπάρχει βέλτιστη καθαρή ημιστάσιμη στρατηγική στο $\hat{\Gamma}$.Έστω $\widehat{\boldsymbol{f}^*}$ μια καθαρή ημιστάσιμη στρατηγική , η οποία είναι βέλτιστη για τη διαδικασία $\hat{\Gamma}$. Έστω s_0 μία αρχική κατάσταση, οπότε :

$$\widehat{\boldsymbol{f}^*} = \{\widehat{\boldsymbol{f}^*}(s_0,s) \mid \widehat{\boldsymbol{f}^*}(s_0,s) \in P\left(\hat{A}(s)\right) \text{ για κάθε } s_0, s \in S\}$$

όπου με $P\left(\hat{A}(s)\right)$ συμβολίζουμε το σύνολο όλων των κατανομών πιθανότητας πάνω στο χώρο $\hat{A}(s)$. Στη συνέχεια, κατασκευάζουμε ένα ζεύγος καθαρών ημιστάσιμων στρατηγικών $(\boldsymbol{f}^*, \boldsymbol{g}^*)$ στο παίγνιο Γ, που προκύπτουν από την $\widehat{\boldsymbol{f}^*}$, ως εξής:

$$f^*(s_0,s) = \begin{cases} \widehat{f^*}(s_0,s) & s \in S_1 \\ 1 & s \in S_2 \end{cases}$$

$$g^*(s_0,s) = \begin{cases} 1 & s \in S_1 \\ \widehat{f^*}(s_0,s) & s \in S_2 \end{cases}$$

Συμβολίζουμε με v_a και $\widehat{v_a}$ τις συναρτήσεις πληρωμών για το παίγνιο Γ (για τον παίκτη 1) και για την διαδικασία $\hat{\Gamma}$, αντίστοιχα.

Από το θεώρημα 9 του [11] γνωρίζουμε ότι

Εάν υπάρχει ζεύγος καθαρών ημιστάσιμων στρατηγικών $(\boldsymbol{f}^*, \boldsymbol{g}^*)$ τέτοιο ώστε για όλα τα $s \in S$ να ισχύει:

$$v_a(s, \boldsymbol{f}, \boldsymbol{g}^*) \le v_a(s, \boldsymbol{f}^*, \boldsymbol{g}^*) \le v_a(s, \boldsymbol{f}^*, \boldsymbol{g}) \text{ για όλα τα } (\boldsymbol{f}, \boldsymbol{g}) \in \xi_1 \times \xi_2$$

τότε το $(\boldsymbol{f}^*, \boldsymbol{g}^*)$ είναι ένα ζεύγος βέλτιστων στρατηγικών του παιχνιδιού. Μπορούμε, λοιπόν, να εστιάσουμε στο σύνολο των ημιστάσιμων στρατηγικών αντί του μεγαλύτερου συνόλου των συμπεριφορικών στρατηγικών.

Έστω τώρα ότι σταθεροποιούμε την στρατηγική $\boldsymbol{g}^*$ για τις καταστάσεις $\{|S_1|+1, |S_1|+2, \dots, |S_1|+|S_2|\}$ στην διαδικασία απόφασης $\hat{\Gamma}$. Τότε αναγόμαστε σε ένα νέο μοντέλο ημιμαρκοβιανής διαδικασίας απόφασης $\hat{\Gamma}_1$, όπου για τις καταστάσεις $\{|S_1|+1, |S_1|+2, \dots, |S_1|+|S_2|\}$, ο decision-maker ακολουθεί την στρατηγική $\boldsymbol{g}^*(s_0, |S_1|+j)$,όπου $j \in \{1,2,\dots,|S_2|\}$. Ομοίως, μπορούμε να θεωρήσουμε την διαδικασία $\hat{\Gamma}_2$, σταθεροποιώντας την στρατηγική $\boldsymbol{f}^*$ για τις καταστάσεις $\{1,2,\dots,|S_1|\}$ της διαδικασίας $\hat{\Gamma}$. Εάν μπορέσουμε να αποδείξουμε ότι η $\boldsymbol{f}^*$ είναι μία βέλτιστη καθαρή ημιστάσιμη στρατηγική για την $\hat{\Gamma}_1$, με σταθεροποιημένη την $\boldsymbol{g}^*$, και ότι η $\boldsymbol{g}^*$ είναι μία βέλτιστη καθαρή ημιστάσιμη στρατηγική για την $\hat{\Gamma}_2$, με σταθεροποιημένη την $\boldsymbol{f}^*$, τότε θα έχουμε καταλήξει στο γεγονός ότι το $(\boldsymbol{f}^*, \boldsymbol{g}^*)$ αποτελεί ένα ζεύγος βέλτιστων καθαρών ημιστάσιμων στρατηγικών για το παίγνιο Γ.

Λήμμα 3.3.1

Η $\boldsymbol{f}^*$ είναι βέλτιστη καθαρή ημιστάσιμη στρατηγική για τον παίκτη 1 στην διαδικασία $\widehat{\Gamma_1}$.

Απόδειξη (Λήμμα 3.3.1)

Με απαγωγή σε άτοπο. Ας υποθέσουμε ότι η $\boldsymbol{f}^*$ δεν είναι βέλτιστη καθαρή ημιστάσιμη στρατηγική στην $\widehat{\Gamma_1}$. Έστω $\widehat{v_1}$ η συνάρτηση συνολικής πληρωμής για την διαδικασία $\widehat{\Gamma_1}$. Έστω, ακόμη, ότι η $\boldsymbol{f}_1^*$ είναι μία βέλτιστη καθαρή ημιστάσιμη στρατηγική στην $\widehat{\Gamma_1}$.Τότε, θα έχουμε ότι :

$$\widehat{v_1}(s, \boldsymbol{f}_1^*) \geq \widehat{v_1}(s, \boldsymbol{f}^*) \quad \forall\, s \in (S_1 \cup S_2)$$

και θα πρέπει να υπάρχει τουλάχιστον μία κατάσταση για την οποία η ανισότητα είναι αυστηρή. Ας υποθέσουμε ότι για την $s_1 \in S$ ισχύει ότι $\widehat{v_1}(s_1, \boldsymbol{f}_1^*) > \widehat{v_1}(s_1, \boldsymbol{f}^*)$

Αυτό όμως σημαίνει ότι $v_a(s_1, \boldsymbol{f}_1^*, \boldsymbol{g}^*) > v_a(s_1, \boldsymbol{f}^*, \boldsymbol{g}^*)$

Μπορούμε τώρα να κατασκευάσουμε την καθαρή ημιστάσιμη στρατηγική $\widehat{\boldsymbol{f}_1}^*$, η οποία θα ταυτίζεται με την $\boldsymbol{f}_1^*$ στο S_1 και με την $\boldsymbol{g}^*$στο S_2, δηλαδή :

$$\hat{f}_1{}^*(s_0, s) = \begin{cases} f_1^*(s_0, s) & , \ s \in S_1 \\ g^*(s_0, s) & , \ s \in S_2 \end{cases}$$

Όμως, τότε έχουμε ότι $\widehat{v_a}\left(s_1, \widehat{\boldsymbol{f}_1}^*\right) > \widehat{v_a}\left(s_1, \widehat{\boldsymbol{f}^*}\right)$, το οποίο είναι αντίφαση, εφόσον γνωρίζουμε ότι η $\widehat{\boldsymbol{f}^*}$ είναι βέλτιστη για την $\hat{\Gamma}$.

Οπότε συμπεραίνουμε πως, πράγματι, η $\boldsymbol{f}^*$είναι βέλτιστη καθαρή ημιστάσιμη στρατηγική στην $\widehat{\Gamma_1}$.■

Σχόλιο

Με όμοια επιχειρήματα μπορεί να δείξει κανείς ότι η στρατηγική $\boldsymbol{g}^*$ είναι μια βέλτιστη καθαρή ημιστάσιμη στρατηγική για την ημιμαρκοβιανή διαδικασία $\widehat{\Gamma_2}$.

Τέλος, χρησιμοποιώντας το Λήμμα (3.3.1) και το παραπάνω σχόλιο, μπορούμε πλέον να συμπεράνουμε πως το $(\boldsymbol{f}^*, \boldsymbol{g}^*)$ αποτελεί ζεύγος βέλτιστων καθαρών ημιστάσιμων στρατηγικών και για τους δύο παίκτες στο ημιμαρκοβιανό παίγνιο τέλειας πληροφόρησης $\hat{\Gamma}$.■

3.4 Αλγόριθμος επίλυσης ημιμαρκοβιανών παιγνίων τέλειας πληροφόρησης

Στη βιβλιογραφία έχει αναπτυχθεί ένας αλγόριθμος επίλυσης ημιμαρκοβιανών διαδικασιών απόφασης με κριτήριο πληρωμής τον οριακό λόγο των μέσων. Για την επίλυση ενός παιγνίου με τα χαρακτηριστικά που έχουμε μελετήσει, αρχικά θα το υποβιβάσουμε σε ημιμαρκοβιανή διαδικασία και έπειτα θα χρησιμοποιήσουμε τον αλγόριθμο από το [12]. Προκειμένου να υπολογιστεί η τιμή του παιχνιδιού καθώς και βέλτιστες στρατηγικές των παικτών, θα ακολουθήσουμε τη λογική της απόδειξης του προηγούμενου Θεωρήματος.

Αλγόριθμος 3.4.1.

Έστω s_0 μια αυθαίρετη αρχική κατάσταση. Θεωρούμε το ακόλουθο πρόβλημα γραμμικού προγραμματισμού για τις μεταβλητές $v(s_0),\ \ g = (g_s \mid s \in S),\ \ h = (h_s \mid s \in S)$ ως εξής :

$$LP : \min v(s_0)$$

υπό τους περιορισμούς

$$g_s \geq \sum_{s' \in S} \hat{p}(s'|s,a) g_{s'} \quad \forall s \in S, \forall\, a \in \hat{A}(s)$$

$$g_s + h_s \geq r(s,a) - v(s_0)\bar{\tau}(s,a) + \sum_{s' \in S} \hat{p}(s'|s,a) h_{s'} \quad \forall s \in S, \forall a \in \hat{A}(s)$$

$$g_{s_0} \leq 0$$

Οι μεταβλητές $v(s_0),\ \ (g_s \mid s \in S \backslash \{s_0\}),\ \ (h_s \mid s \in S)$ δεν έχουν περιορισμό στο πρόσημο.

Το αντίστοιχο δυϊκό γραμμικό πρόγραμμα για τις μεταβλητές

$$x = \left(x_{sa} \mid s \in S, a \in \hat{A}(s)\right), y = \left(y_{sa} \mid s \in S, a \in \hat{A}(s)\right) \quad \text{και } t \text{ , θα είναι :}$$

$$DLP : \max R_s$$

$$\text{όπου } R_s = \sum_{s \in S} \sum_{a \in \hat{A}(s)} \hat{r}(s,a)\, x_{sa} \qquad (3a)$$

υπό τους περιορισμούς :

$$\sum_{s\in S}\sum_{a\in\hat{A}(s)}\{\delta_{ss'}-\hat{p}(s'|s,a)\}x_{sa}=0,\qquad \forall s'\in S \tag{3b}$$

$$\sum_{a\in\hat{A}(s)}x_{s'a}+\sum_{s\in S}\sum_{a\in\hat{A}(s)}\{\delta_{ss'}-\hat{p}(s'|s,a)\}y_{sa}=0\,,\qquad \forall s'\in S\backslash\{s_0\} \tag{3c}$$

$$\sum_{a\in\hat{A}(s_0)}x_{s_0a}+\sum_{s\in S}\sum_{a\in\hat{A}(s)}\{\delta_{ss_0}-\hat{p}(s_0|s,a)\}y_{sa}-t\ =0 \tag{3d}$$

$$\sum_{s\in S}\sum_{a\in\hat{A}(s)}\bar{\tau}(s,a)\,x_{sa}=1 \tag{3e}$$

$x_{sa}\geq 0, y_{sa}\geq 0\ \forall s\in S$

$a\in\hat{A}(s)$

$t\geq 0$

(όπου το $\delta_{ss'}$ είναι το Δέλτα του Kronecker)

Για την εύρεση μιας εφικτής λύσης (x,y,t) του *DLP* ,ορίζουμε τα ακόλουθα σύνολα :

$$S_x=\left\{s\in S\mid \sum_{a\in\hat{A}(s)}x_{sa}>0\right\}$$

$$S_y=\left\{s\in S\mid \sum_{a\in\hat{A}(s)}x_{sa}=0\ \text{και}\ \sum_{a\in\hat{A}(s)}y_{sa}>0\right\}$$

$$S_{xy}=\left\{s\in S\mid \sum_{a\in\hat{A}(s)}x_{sa}=0\ \text{και}\ \sum_{a\in\hat{A}(s)}y_{sa}=0\right\}$$

Οπότε θα ισχύει ότι $S=S_x\cup S_y\cup S_{xy}$, και τα σύνολα S_x,S_y και S_{xy} είναι ανά δύο ξένα μεταξύ τους. Μία καθαρή στάσιμη στρατηγική που αντιστοιχεί στην εφικτή λύση (x,y,t) του *DLP* ορίζεται ως $\boldsymbol{f}_{xyt}^{ps_0}$, όπου s_0 είναι η αυθαίρετη αρχική κατάσταση που σταθεροποιήσαμε εξ αρχής.

$\boldsymbol{f}_{xyt}^{ps_0}(s)=a_s,\ s\in S$, τέτοια ώστε:

$$a_s=\begin{cases} a, & \text{εάν } s\in S_x \text{ και } x_{sa}>0\\ a', & \text{εάν } s\in S_y \text{ και } y_{sa'}>0\\ \alpha\upsilon\theta\alpha\acute{\iota}\rho\varepsilon\tau o, & \text{εάν } s\in S_{xy}\end{cases}$$

Στο σημείο αυτό, θα χρειαστούμε το παρακάτω σημαντικό αποτέλεσμα, το οποίο παρατίθεται στο [12].

Θεώρημα 3.4.1

Έστω (x^*, y^*, t^*) μία βέλτιστη λύση του *DLP*. Τότε η $\boldsymbol{f}^{ps_0}_{x^*y^*t^*}$ είναι μία βέλτιστη καθαρή στάσιμη στρατηγική της ημιμαρκοβιανής διαδικασίας , για την αρχική κατάσταση s_0.

Από το παραπάνω θεώρημα εξάγουμε το συμπέρασμα πως μία βέλτιστη καθαρή ημιστάσιμη στρατηγική μιας ημιμαρκοβιανής διαδικασίας απόφασης μπορεί να βρεθεί μέσω της βέλτιστης λύσης του δυϊκού γραμμικού προγράμματος. Για το λόγο αυτό, ο συγκεκριμένος αλγόριθμος είναι ιδιαίτερα χρήσιμος, καθώς θα μας επιτρέψει να εξάγουμε βέλτιστες στρατηγικές και για τους δύο παίκτες στο ημιμαρκοβιανό παίγνιο τέλειας πληροφόρησης, ακολουθώντας την συλλογιστική πορεία της απόδειξης του Θεωρήματος 3.3.2.

Για μία σταθεροποιημένη αρχική κατάσταση s_0, έχουμε μία βέλτιστη καθαρή στάσιμη στρατηγική, την $\boldsymbol{f}^{ps_0}_{x^*y^*t^*}$, όπου $\boldsymbol{f}^{ps_0}_{x^*y^*t^*}(s) = a^*_s$ και το a^*_s είναι όπως ορίστηκε προηγουμένως. Έτσι, για διαφορετικές αρχικές καταστάσεις, λαμβάνουμε μία βέλτιστη στρατηγική για τον decision-maker, στην διαδικασία απόφασης. Τέλος, θα χρειαστούμε ακόμη ένα πόρισμα , το Πόρισμα 1 από το [12].

Πόρισμα 3.4.1

Έστω $\boldsymbol{f}^{p^*}_{s_0}$να είναι μία καθαρή στάσιμη στρατηγική που εξήχθη από μία βέλτιστη λύση του *DLP* για αρχική κατάσταση $s_0 \in S$. Τότε η καθαρή ημιστάσιμη στρατηγική

$$\xi^{p^*} = (f_1^{p^*}, f_2^{p^*}, \dots, f_N^{p^*})$$

θα είναι βέλτιστη για την ημιμαρκοβιανή διαδικασία με κριτήριο τον οριακό λόγο των μέσων.

Οπότε από το τελευταίο πόρισμα , λαμβάνουμε μία βέλτιστη στρατηγική $\hat{\boldsymbol{f}}^*$ του decision-maker στην διαδικασία $\hat{\Gamma}$, όπου :

$$\hat{\boldsymbol{f}}^* = (\boldsymbol{f}_1^*, \boldsymbol{f}_2^*, \dots, \boldsymbol{f}_s^*, \dots, \boldsymbol{f}_N^*)$$

και $\boldsymbol{f}_s^* = \boldsymbol{f}^{ps}_{x^*y^*t^*}$ αποτελεί μία βέλτιστη καθαρή στάσιμη στρατηγική της ημιμαρκοβιανής διαδικασίας με αρχική κατάσταση s.

Το ζευγάρι $(\boldsymbol{f}^*, \boldsymbol{g}^*)$ βέλτιστων καθαρών στρατηγικών για τους δύο παίκτες στο παίγνιο Γ, δίνεται όπως είδαμε και παραπάνω από :

$$\boldsymbol{f}^*(s_0,s)=\begin{cases}\widehat{\boldsymbol{f}^*}(s_0,s) & s\in S_1\\ 1 & s\in S_2\end{cases}$$

$$\boldsymbol{g}^*(s_0,s)=\begin{cases}1 & s\in S_1\\ \widehat{\boldsymbol{f}^*}(s_0,s) & s\in S_2\end{cases}$$

Εάν R_s είναι η αντικειμενική συνάρτηση του *DLP* για αρχική κατάσταση s , τότε το

$$\hat{\boldsymbol{v}}=(\hat{v}(1),\hat{v}(2),\dots,\hat{v}(s),\dots,\hat{v}(N))$$

θα είναι το διάνυσμα τιμής της διαδικασίας $\hat{\Gamma}$, όπου $\hat{v}(s)=\max R_s$.

Το διάνυσμα τιμής για το παίγνιο Γ θα δίνεται από :

$$\boldsymbol{v}=\left(\hat{v}(1),\hat{v}(2),\dots,\hat{v}(|S_1|),-\hat{v}(|S_1|+1),\dots,-\hat{v}(|S_1|+|S_2|)\right)$$

Το αρνητικό πρόσημο, που αφορά στις καταστάσεις $\{|S_1|+1,\dots,|S_1|+|S_2|\}$, δικαιολογείται από το γεγονός ότι ο παίκτης 2 θέλει να ελαχιστοποιήσει το κόστος για αυτές τις καταστάσεις.

Υπολογισμός βέλτιστης στρατηγικής μέσω πλήρους απαρίθμησης

Σύμφωνα με την μέθοδο της πλήρους απαρίθμησης, θα υπολογίσουμε την τιμή της ημιμαρκοβιανής διαδικασίας για μία καθαρή στάσιμη στρατηγική $\boldsymbol{f}$, κατ' αναλογία με την σχέση (3.2.1) που αφορά τα παίγνια, ως εξής:

$$\hat{v}(s,\boldsymbol{f})=\frac{[\boldsymbol{P}^*(\boldsymbol{f})\hat{\boldsymbol{r}}(\boldsymbol{f})](s)}{[\boldsymbol{P}^*(\boldsymbol{f})\bar{\boldsymbol{\tau}}(\boldsymbol{f})](s)},\qquad \text{για όλα τα } s\in S$$

όπου $\hat{\boldsymbol{r}}(\boldsymbol{f})$ είναι το διάνυσμα αμοιβής και $\bar{\boldsymbol{\tau}}(\boldsymbol{f})$ είναι το διάνυσμα του αναμενόμενου χρόνου παραμονής.

Εν συνεχεία, παραθέτουμε σε βήματα τον αλγόριθμο υπολογισμού του πίνακα ορίου κατά Cesaro για μια ημιμαρκοβιανή διαδικασία με *N* καταστάσεις,

Είσοδος: Ο πίνακας πιθανοτήτων μετάβασης $\boldsymbol{P}\in M_N(\mathbb{R})$

Έξοδος: Ο πίνακας όριο κατά Cesaro $\boldsymbol{P}^*\in M_N(\mathbb{R})$ του $\boldsymbol{P}$

Βήμα 1: Ορίζουμε το χαρακτηριστικό πολυώνυμο $Ch_P(z) = |P - zI_n|$

Βήμα 2: Διαιρούμε το $Ch_P(z)$ με το πολυώνυμο $(z-1)^{m(1)}$, όπου $m(1)$ είναι η αλγεβρική πολλαπλότητα της ιδιοτιμής $z_0 = 1$. Θα συμβολίζουμε το πηλίκο της διαίρεσης με $D(z)$.

Βήμα 3: Υπολογίζουμε τον πίνακα $\boldsymbol{R} = D(\boldsymbol{Q})$

Βήμα 4: Ορίζουμε τον πίνακα όριο $\boldsymbol{P}^*$ διαιρώντας τον πίνακα $\boldsymbol{R}$ με το άθροισμα των στοιχείων μιας αυθαίρετης γραμμής του.

Κλείνοντας αυτό το κεφάλαιο, παραθέτουμε ένα παράδειγμα, αναλυτικά λυμένο στο προγραμματιστικό περιβάλλον MATLAB. Αρχικά, θα γίνει επίλυση μέσω γραμμικού προγραμματισμού και έπειτα θα συγκρίνουμε τη λύση με εκείνη που θα λάβουμε από τη μέθοδο πλήρους απαρίθμησης, Φυσικά, αναμένεται τα αποτελέσματα όσον αφορά τόσο την τιμή του παιγνίου όσο και τη βέλτιστη στρατηγική, να ταυτίζονται.

3.5 Επίλυση μέσω γραμμικού προγραμματισμού και μέσω πλήρους απαρίθμησης

Έστω το ακόλουθο παίγνιο Γ, όπου οι παίκτες 1 και 2 θα παίζουν σύμφωνα με τους παρακάτω πίνακες:

1.5 $(\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4})$ 1
1 $(\frac{1}{3},0,\frac{2}{3},0)$ 0.8

Κατάσταση 1

3 $(\frac{1}{2},\frac{1}{2},0,0)$ 1.5
2 $(\frac{1}{3},\frac{1}{3},\frac{1}{3},0)$ 1

Κατάσταση 2

3 $(\frac{1}{2},0,\frac{1}{2},0)$ 1	7 $(\frac{1}{2},0,0,\frac{1}{2})$ 2

Κατάσταση 3

2 $(\frac{1}{4},\frac{1}{2},\frac{1}{4},0)$ 2	1 $(\frac{1}{2},\frac{1}{4},\frac{1}{4},0)$ 1.2

Κατάσταση 4

Η δομή του παραπάνω παιγνίου ερμηνεύεται ως εξής:

r $(p_1,p_2,p_3,1-p_1-p_2-p_3)$ $\bar{\tau}$

όπου με r συμβολίζεται η αμοιβή που θα λάβει ο παίκτης 1 (και ο παίκτης 2 θα λαμβάνει αμοιβή $-r$), με p_i συμβολίζεται η πιθανότητα η επόμενη μετάβαση να γίνει στην κατάσταση i (υποχρεωτικά, για κάθε κατάσταση, τα p_i θα έχουν άθροισμα ίσο με μονάδα) και με $\bar{\tau}$ συμβολίζεται η αναμενόμενη τιμή του χρόνου παραμονής στην κατάσταση που βρίσκεται η διαδικασία.

Αρχικά, προκειμένου να γίνει η ανάλυση του παραπάνω παιγνίου, θα το μετατρέψουμε πρώτα σε διαδικασία απόφασης για τον παίκτη 1. Με ανάλογο τρόπο μπορεί να γίνει η μετατροπή σε διαδικασία απόφασης για τον δεύτερο παίκτη. Ορίζεται, λοιπόν η διαδικασία $\widehat{\Gamma_1}$, με τον τρόπο που περιγράφεται στην απόδειξη του Θεωρήματος 3.3.2., ως εξής:

1.5 $(\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4})$ 1	3 $(\frac{1}{2},\frac{1}{2},0,0)$ 1.5	-3 $(\frac{1}{2},0,\frac{1}{2},0)$ 1	-2 $(\frac{1}{4},\frac{1}{2},\frac{1}{4},0)$ 2
1 $(\frac{1}{3},0,\frac{2}{3},0)$ 0.8	2 $(\frac{1}{3},\frac{1}{3},\frac{1}{3},0)$ 1	-7 $(\frac{1}{2},0,0,\frac{1}{2})$ 2	-1 $(\frac{1}{2},\frac{1}{4},\frac{1}{4},0)$ 1.2
Κατάσταση 1	Κατάσταση 2	Κατάσταση 3	Κατάσταση 4

α) <u>Επίλυση μέσω γραμμικού προγραμματισμού</u>

Ακολουθώντας τα βήματα του αλγορίθμου που περιγράφηκαν παραπάνω, θα αναζητήσουμε λύση της διαδικασίας $\widehat{\Gamma_1}$ επιλύοντας το αντίστοιχο δυϊκό πρόβλημα γραμμικού προγραμματισμού.

Το ζητούμενο είναι η μεγιστοποίηση της R_s, η οποία από την σχέση (3α) προκύπτει να είναι:

$$R_s = 1.5x_{11} + x_{12} + 3x_{21} + 2x_{22} - 3x_{31} - 7x_{32} - 2x_{41} - x_{42}$$

Το σύνολο περιορισμών που προκύπτουν από την $(3b)$ είναι:

$9x_{11} + 8x_{12} - 6x_{12} - 4x_{22} - 6x_{31} - 6x_{32} - 3x_{41} - 6x_{42} = 0$

$-3x_{11} + 6x_{21} + 8x_{22} - 6x_{41} - 3x_{42} = 0$

$-3x_{11} - 8x_{12} - 4x_{22} + 6x_{31} + 12x_{32} - 3x_{41} - 3x_{42} = 0$

$-x_{11} - 2x_{32} + 4x_{41} + 4x_{42} = 0$

Το σύνολο περιορισμών που προκύπτει από τις $(3c)$ και $(3d)$ είναι:

$$12x_{11} + 12x_{22} + 9y_{11} + 8y_{12} - 6y_{21} - 4y_{22} - 6y_{31} - 6y_{32} - 3y_{41} - 6y_{42} - 12\delta_{s_0 1}t = 0$$

$$12x_{21} + 12x_{22} - 3y_{11} + 6y_{21} + 8y_{22} - 6y_{41} - 3y_{42} - 12\delta_{s_0 2}t = 0$$

$$12x_{31} + 12x_{32} - 3y_{11} - 8y_{12} - 4y_{22} + 6y_{31} + 12y_{32} - 3y_{41} - 3y_{42} - 12\delta_{s_0 3}t = 0$$

$$4x_{41} + 4x_{42} - y_{11} - 2y_{32} + 4y_{41} + 4y_{42} - 4$$

Τέλος, από την $(3e)$ λαμβάνουμε τον περιορισμό:

$$x_{11} + 0.8x_{12} + 1.5x_{21} + x_{22} + x_{31} + 2x_{32} + 2x_{41} + 1.2x_{42} = 1$$

και

$t \geq 0\,, x_{sa} \geq 0, y_{sa} \geq 0$ για κάθε $s \in S, a \in \hat{A}(s)$

Χρησιμοποιώντας την συνάρτηση linprog, λαμβάνουμε από το MATLAB , για κάθε πιθανή αρχική κατάσταση, τα ακόλουθα αποτελέσματα:

(i) $s_0 = 1$

x_{11}	0	y_{11}	0	t	0.7724
x_{12}	0.3239	y_{12}	1.0590		
x_{21}	0	y_{21}	0		
x_{22}	0.0498	y_{22}	0		
x_{31}	0	y_{31}	0		
x_{32}	0.2658	y_{32}	0.4651		
x_{41}	0	y_{41}	0.0997		
x_{42}	0.1329	y_{42}	0		

$\max R_1 = 1.5698$

(ii) $s_0 = 2$

x_{11}	0	y_{11}	0	t	0.7724
x_{12}	0.3239	y_{12}	0.2554		
x_{21}	0	y_{21}	0		
x_{22}	0.0498	y_{22}	1.0839		
x_{31}	0	y_{31}	0		
x_{32}	0.2658	y_{32}	0.2658		
x_{41}	0	y_{41}	0		
x_{42}	0.1329	y_{42}	0		

$\max R_2 = 1.5698$

(iii) $s_0 = 3$

x_{11}	0	y_{11}	0	t	0.7724
x_{12}	0.3239	y_{12}	0		
x_{21}	0	y_{21}	0		
x_{22}	0.0498	y_{22}	0		
x_{31}	0	y_{31}	0.1329		
x_{32}	0.2658	y_{32}	0.4651		
x_{41}	0	y_{41}	0.0997		
x_{42}	0.1329	y_{42}	0		

$\max R_3 = 1.5698$

(iv) $s_0 = 4$

x_{11}	0	y_{11}	0	t	0.7724
x_{12}	0.3239	y_{12}	0		
x_{21}	0	y_{21}	0		
x_{22}	0.0498	y_{22}	0.4049		
x_{31}	0	y_{31}	0.0581		
x_{32}	0.2658	y_{32}	0		
x_{41}	0	y_{41}	0.6395		
x_{42}	0.1329	y_{42}	0		

$\max R_4 = 1.5698$

Από τα παραπάνω αποτελέσματα, γίνεται φανερό ότι το διάνυσμα τιμής της διαδικασίας $\widehat{\Gamma_1}$ θα είναι ίσο με :

$$\widehat{\boldsymbol{\varphi}} = (1.5698, 1.5698, 1.5698, 1.5698)$$

Έστω τώρα $\widehat{\boldsymbol{f_1^*}} = (\boldsymbol{f}_1^*, \boldsymbol{f}_2^*, \boldsymbol{f}_3^*, \boldsymbol{f}_4^*)$ μια βέλτιστη καθαρή ημι-στάσιμη στρατηγική της διαδικασίας $\widehat{\Gamma_1}$, , όπου με $\boldsymbol{f}_i^*, i = 1,2,3,4$, θα συμβολίζουμε μια βέλτιστη στρατηγική της διαδικασίας, δεδομένου ότι η αρχική κατάσταση είναι η i.

Τότε, σύμφωνα με την σχέση $f_{x^*y^*t^*}^{ps_0}(s) = a_s^*$, υπολογίζεται ότι :

$$\boldsymbol{f}_1^* = f_{x^*y^*t^*}^{p1}(s) = a_{s1}^* = (2,2,2,2)$$

$$\boldsymbol{f}_2^* = f_{x^*y^*t^*}^{p2}(s) = a_{s2}^* = (2,2,2,2)$$

$$\boldsymbol{f}_3^* = f_{x^*y^*t^*}^{p3}(s) = a_{s3}^* = (2,2,2,2)$$

$$\boldsymbol{f}_4^* = f_{x^*y^*t^*}^{p4}(s) = a_{s4}^* = (2,2,2,2)$$

Υπολογίστηκε, λοιπόν, η βέλτιστη στρατηγική για την διαδικασία να είναι η

$$\widehat{\boldsymbol{f_1^*}} = (\boldsymbol{f}_1^*, \boldsymbol{f}_2^*, \boldsymbol{f}_3^*, \boldsymbol{f}_4^*) = (\boldsymbol{f}_1^*, \boldsymbol{f}_1^*, \boldsymbol{f}_1^*, \boldsymbol{f}_1^*)$$

και η $\boldsymbol{f}_1^*$ είναι η καθαρή στρατηγική που υποδεικνύει την επιλογή της $2^{ης}$ διαθέσιμης ενέργειας , για κάθε κατάσταση $s \in \{1,2,3,4\}$, για κάθε αρχική κατάσταση s_0.

β) <u>Επίλυση μέσω της μεθόδου πλήρους απαρίθμησης</u>

Σε αυτή την μέθοδο και πάλι ζητούμενο είναι ο υπολογισμός της αξίας του παιχνιδιού και η εύρεση μιας βέλτιστης στρατηγικής για κάθε πιθανής αρχική κατάσταση s_0, για την διαδικασία $\widehat{\Gamma_1}$.

Αρχικά, παρατηρούμε ότι υπάρχουν συνολικά 2^4 διαθέσιμες καθαρές στρατηγικές για την διαδικασία $\widehat{\Gamma_1}$, ακριβώς όσοι είναι οι τρόποι με τους οποίους μπορούμε να κατανείμουμε τις 2 διαθέσιμες ενέργειες κάθε κατάστασης σε ένα 4-διάστατο διάνυσμα, κάθε συντεταγμένη του οποίου αντιστοιχεί σε μία εκ των 4 καταστάσεων της διαδικασίας.

Υλοποιώντας τον αλγόριθμο που αφορά στον υπολογισμό του ορίου κατά Cesaro ενός πίνακα, και χρησιμοποιώντας την σχέση (3.2) για τον υπολογισμό της αναμενόμενης τιμής, λαμβάνονται τα ακόλουθα αποτελέσματα:

- Στρατηγική $\boldsymbol{f}_1 = (1,1,1,1)$

$$\boldsymbol{P}^*(\boldsymbol{f}_1) = \begin{pmatrix} 0.3810 & 0.2857 & 0.2381 & 0.0952 \\ 0.3810 & 0.2857 & 0.2381 & 0.0952 \\ 0.3810 & 0.2857 & 0.2381 & 0.0952 \\ 0.3810 & 0.2857 & 0.2381 & 0.0952 \end{pmatrix}$$

Τότε, για την $\boldsymbol{f}_1$, παίρνουμε:

$$\widehat{v_a}(\boldsymbol{f}_1) = (0.4231,0.4231,0.4231,0.4231)$$

- Στρατηγική $\boldsymbol{f}_2 = (2,1,1,1)$

$$\boldsymbol{P}^*(\boldsymbol{f}_2) = \begin{pmatrix} 0.4286 & 0 & 0.5714 & 0 \\ 0.4286 & 0 & 0.5714 & 0 \\ 0.4286 & 0 & 0.5714 & 0 \\ 0.4286 & 0 & 0.5714 & 0 \end{pmatrix}$$

Τότε, για την $\boldsymbol{f}_2$, παίρνουμε:

$$\widehat{v_a}(\boldsymbol{f}_2) = (1.4062,1.4063,1.4063,1.4062)$$

- Στρατηγική $\boldsymbol{f}_3 = (1,2,1,1)$

$$\boldsymbol{P}^*(\boldsymbol{f}_3) = \begin{pmatrix} 0.3556 & 0.2 & 0.3556 & 0.0889 \\ 0.3556 & 0.2 & 0.3556 & 0.0889 \\ 0.3556 & 0.2 & 0.3556 & 0.0889 \\ 0.3556 & 0.2 & 0.3556 & 0.0889 \end{pmatrix}$$

Τότε, για την $\boldsymbol{f}_3$, παίρνουμε:

$$\widehat{v_a}(\boldsymbol{f}_3) = (0.2857,0.2857,0.2857,0.2857)$$

- Στρατηγική $\boldsymbol{f}_4 = (2,2,1,1)$

$$\boldsymbol{P}^*(\boldsymbol{f_4}) = \begin{pmatrix} 0.4286 & 0 & 0.5714 & 0 \\ 0.4286 & 0 & 0.5714 & 0 \\ 0.4286 & 0 & 0.5714 & 0 \\ 0.4286 & 0 & 0.5714 & 0 \end{pmatrix}$$

Τότε, για την $\boldsymbol{f}_4$, παίρνουμε:

$$\widehat{v_a}(\boldsymbol{f}_4) = (1.4062,1.4062,1.4063,1.4062)$$

- Στρατηγική $\boldsymbol{f}_5 = (1,1,2,1)$

$$\boldsymbol{P}^*(\boldsymbol{f}_5) = \begin{pmatrix} 0.3684 & 0.3421 & 0.1316 & 0.1579 \\ 0.3684 & 0.3421 & 0.1316 & 0.1579 \\ 0.3684 & 0.3421 & 0.1316 & 0.1579 \\ 0.3684 & 0.3421 & 0.1316 & 0.1579 \end{pmatrix}$$

Τότε, για την $\boldsymbol{f}_5$, παίρνουμε:

$$\widehat{v_a}(\boldsymbol{f}_5) = (0.2342,0.2342,0.2342,0.2342)$$

- Στρατηγική $\boldsymbol{f}_6 = (2,1,2,1)$

$$\boldsymbol{P}^*(\boldsymbol{f}_6) = \begin{pmatrix} 0.3962 & 0.1509 & 0.3019 & 0.1509 \\ 0.3962 & 0.1509 & 0.3019 & 0.1509 \\ 0.3962 & 0.1509 & 0.3019 & 0.1509 \\ 0.3962 & 0.1509 & 0.3019 & 0.1509 \end{pmatrix}$$

Τότε, για την $\boldsymbol{f}_6$, παίρνουμε:

$$\widehat{v_a}(\boldsymbol{f}_6) = (1.0807,1.0807,1.0807,1.0807)$$

- Στρατηγική $\boldsymbol{f}_7 = (1,2,2,1)$

$$\boldsymbol{P}^*(\boldsymbol{f}_7) = \begin{pmatrix} 0.3265 & 0.2653 & 0.2177 & 0.1905 \\ 0.3265 & 0.2653 & 0.2177 & 0.1905 \\ 0.3265 & 0.2653 & 0.2177 & 0.1905 \\ 0.3265 & 0.2653 & 0.2177 & 0.1905 \end{pmatrix}$$

Τότε, για την $\boldsymbol{f}_7$, παίρνουμε:

$$\widehat{v_a}(\boldsymbol{f}_7) = (0.6280,0.6280,0.6280,0.6280)$$

- Στρατηγική $\boldsymbol{f}_8 = (2,2,2,1)$

$$\boldsymbol{P}^*(\boldsymbol{f}_8) = \begin{pmatrix} 0.3750 & 0.1250 & 0.3333 & 0.1667 \\ 0.3750 & 0.1250 & 0.3333 & 0.1667 \\ 0.3750 & 0.1250 & 0.3333 & 0.1667 \\ 0.3750 & 0.1250 & 0.3333 & 0.1667 \end{pmatrix}$$

Τότε, για την $\boldsymbol{f}_8$, παίρνουμε:

$$\widehat{v_a}(\boldsymbol{f}_8) = (1.4327,1.4327,1.4327,1.4327)$$

- Στρατηγική $\boldsymbol{f}_9 = (1,1,1,2)$

$$\boldsymbol{P}^*(\boldsymbol{f}_9) = \begin{pmatrix} 0.4000 & 0.2500 & 0.2500 & 0.1000 \\ 0.4000 & 0.2500 & 0.2500 & 0.1000 \\ 0.4000 & 0.2500 & 0.2500 & 0.1000 \\ 0.4000 & 0.2500 & 0.2500 & 0.1000 \end{pmatrix}$$

Τότε, για την $\boldsymbol{f}_9$, παίρνουμε:

$$\widehat{v_a}(\boldsymbol{f}_9) = (0.4367,0.4367,0.4367,0.4367)$$

- Στρατηγική $\boldsymbol{f}_{10} = (2,1,1,2)$

$$\boldsymbol{P}^*(\boldsymbol{f}_{10}) = \begin{pmatrix} 0.4286 & 0 & 0.5714 & 0 \\ 0.4286 & 0 & 0.5714 & 0 \\ 0.4286 & 0 & 0.5714 & 0 \\ 0.4286 & 0 & 0.5714 & 0 \end{pmatrix}$$

Τότε, για την $\boldsymbol{f}_{10}$, παίρνουμε:

$$\widehat{\boldsymbol{v}_a}(\boldsymbol{f}_{10}) = (1.4062,1.4063,1.4063,1.4063)$$

- Στρατηγική $\boldsymbol{f}_{11} = (1,2,1,2)$

$$\boldsymbol{P}^*(\boldsymbol{f}_{11}) = \begin{pmatrix} 0.3765 & 0.1765 & 0.3529 & 0.0941 \\ 0.3765 & 0.1765 & 0.3529 & 0.0941 \\ 0.3765 & 0.1765 & 0.3529 & 0.0941 \\ 0.3765 & 0.1765 & 0.3529 & 0.0941 \end{pmatrix}$$

Τότε, για την $\boldsymbol{f}_{11}$, παίρνουμε:

$$\widehat{\boldsymbol{v}_a}(\boldsymbol{f}_{11}) = (0.2309,0.2309,0.2309,0.2309)$$

- Στρατηγική $\boldsymbol{f}_{12} = (2,2,1,2)$

$$\boldsymbol{P}^*(f_{12}) = \begin{pmatrix} 0.4286 & 0 & 0.5714 & 0 \\ 0.4286 & 0 & 0.5714 & 0 \\ 0.4286 & 0 & 0.5714 & 0 \\ 0.4286 & 0 & 0.5714 & 0 \end{pmatrix}$$

Τότε, για την $\boldsymbol{f}_{12}$, παίρνουμε:

$$\widehat{\boldsymbol{v}_a}(\boldsymbol{f}_{12}) = (1.4062,1.4062,1.4063,1.4062)$$

- Στρατηγική $\boldsymbol{f}_{13} = (1,1,2,2)$

$$\boldsymbol{P}^*(f_{13}) = \begin{pmatrix} 0.4000 & 0.2857 & 0.1429 & 0.1714 \\ 0.4000 & 0.2857 & 0.1429 & 0.1714 \\ 0.4000 & 0.2857 & 0.1429 & 0.1714 \\ 0.4000 & 0.2857 & 0.1429 & 0.1714 \end{pmatrix}$$

Τότε, για την $\boldsymbol{f}_{13}$, παίρνουμε:

$$\widehat{\boldsymbol{v}_a}(\boldsymbol{f}_{13}) = (0.2165,0.2165,0.2165,0.2165)$$

- Στρατηγική $\boldsymbol{f}_{14} = (2,1,2,2)$

$$\boldsymbol{P}^*(\boldsymbol{f}_{14}) = \begin{pmatrix} 0.4286 & 0.0816 & 0.3265 & 0.1633 \\ 0.4286 & 0.0816 & 0.3265 & 0.1633 \\ 0.4286 & 0.0816 & 0.3265 & 0.1633 \\ 0.4286 & 0.0816 & 0.3265 & 0.1633 \end{pmatrix}$$

Τότε, για την $\boldsymbol{f}_{14}$, παίρνουμε:

$$\widehat{\boldsymbol{v}_a}(\boldsymbol{f}_{14}) = (1.3509,1.3509,1.3509,1.3509)$$

- Στρατηγική $\boldsymbol{f}_{15} = (1,2,2,2)$

$$\boldsymbol{P}^*(\boldsymbol{f}_{15}) = \begin{pmatrix} 0.3714 & 0.2143 & 0.2143 & 0.2000 \\ 0.3714 & 0.2143 & 0.2143 & 0.2000 \\ 0.3714 & 0.2143 & 0.2143 & 0.2000 \\ 0.3714 & 0.2143 & 0.2143 & 0.2000 \end{pmatrix}$$

Τότε, για την $\boldsymbol{f}_{15}$, παίρνουμε:

$$\widehat{\boldsymbol{v}_a}(\boldsymbol{f}_{15}) = (0.5695,0.5695,0.5695,0.5695)$$

- Στρατηγική $\boldsymbol{f}_{16} = (2,2,2,2)$

$$\boldsymbol{P}^*(\boldsymbol{f}_{16}) = \begin{pmatrix} 0.4194 & 0.0645 & 0.3441 & 0.1720 \\ 0.4194 & 0.0645 & 0.3441 & 0.1720 \\ 0.4194 & 0.0645 & 0.3441 & 0.1720 \\ 0.4194 & 0.0645 & 0.3441 & 0.1720 \end{pmatrix}$$

Τότε, για την $\boldsymbol{f}_{16}$, παίρνουμε:

$$\widehat{\boldsymbol{v}_a}(\boldsymbol{f}_{16}) = (1.5698,1.5698,1.5698,1.5698)$$

Όπως ήταν αναμενόμενο, η βέλτιστη στρατηγική για τη διαδικασία $\widehat{\Gamma_1}$ είναι η $\boldsymbol{f}_{16} = (2,2,2,2)$, δηλαδή το βέλτιστο 'παίξιμο' είναι να επιλέγεται η 2η ενέργεια για κάθε κατάσταση $s \in \{1,2,3,4\}$, και το διάνυσμα τιμής της διαδικασίας, ταυτίζεται με εκείνο που λάβαμε από τον αλγόριθμο γραμμικού προγραμματισμού και είναι ίσο με:

$$\widehat{\boldsymbol{v}_a}(\boldsymbol{f}_{16}) = (1.5698,1.5698,1.5698,1.5698)$$

Σχόλια-Παρατηρήσεις

- Οπότε έχουμε καταλήξει και με τις 2 μεθόδους στο ίδιο διάνυσμα τιμής για την διαδικασία $\widehat{\Gamma_1}$ και επιστρέφοντας στο αρχικό παίγνιο Γ, εάν $(\boldsymbol{f}^*, \boldsymbol{g}^*)$ είναι ένα ζεύγος βέλτιστων στρατηγικών, τότε η $\boldsymbol{f}^*$ υποδεικνύει στον $1^{ο}$ παίκτη να επιλέγει την $2^{η}$ ενέργεια για κάθε αρχική κατάσταση για τις δύο πρώτες καταστάσεις του παιγνίου και η $\boldsymbol{g}^*$ υποδεικνύει στον $2^{ο}$ παίκτη να επιλέγει την $2^{η}$ ενέργεια για κάθε αρχική κατάσταση για τις δύο τελευταίες καταστάσεις του παιγνίου.

- Κατά την υλοποίηση των παραπάνω μέσω του λογισμικού MATLAB , έγινε μια προσπάθεια εκτίμησης του υπολογιστικού χρόνου που απαιτείται για την επίλυση του συγκεκριμένου παραδείγματος με τις δύο διαφορετικές μεθόδους, κάνοντας χρήση της εντολής "tic-tac". Οι μετρήσεις έδειξαν πως για τη μέθοδο πλήρους απαρίθμησης, οι υπολογισμοί ολοκληρώνονταν κατά μέσο όρο περίπου 10 φορές ταχύτερα από εκείνους της μεθόδου γραμμικού προγραμματισμού. Ενώ αυτό φαίνεται ως ένα μη αναμενόμενο αποτέλεσμα, μπορεί να αποδοθεί στις "μικρές" διαστάσεις του προβλήματος. Σε ανάλογες μετρήσεις που έγιναν σε αριθμητικό παράδειγμα με 8 καταστάσεις, παρατηρήσαμε πως ο χρόνος εκτέλεσης της μεθόδου πλήρους απαρίθμησης, αυτή τη φορά ήταν κατά μέσο όρο περίπου 3 φορές πιο γρήγορος από εκείνον της μεθόδου γραμμικού προγραμματισμού. Για το λόγο αυτό και γνωρίζοντας πως η μέθοδος Simplex ενδείκνυται υπολογιστικά για προβλήματα μεγάλης διάστασης, εικάζουμε, πώς όσο μεγαλώνει η διάσταση του προβλήματος, τόσο περισσότερο υπολογιστικά συμφέρουσα θα είναι η μέθοδος αυτή. Έτσι, θεωρούμε πως θα υπάρχει μία διάσταση προβλήματος για την οποία οι δύο τρόποι υπολογισμού που συγκρίνονται παραπάνω θα έχουν το ίδιο υπολογιστικό κόστος και πως από εκείνο το σημείο και έπειτα, η μέθοδος πλήρους απαρίθμησης θα αρχίσει να είναι ιδιαίτερα χρονοβόρα, ώστε να αποθαρρύνει κάποιον από τη χρήση της. Το γεγονός αυτό αποτελεί αντικείμενο μελλοντικής έρευνας.

Παράρτημα

Για την υλοποίηση της εφαρμογής της παραγράφου 3.5, έγινε χρήση των παρακάτω, μέσω του λογισμικού MATLAB:

i) Συνάρτηση linprog για επίλυση μέσω γραμμικού προγραμματισμού

```
S = 4; % Number of States
d = zeros(S,1);
e = eye(S);
A = [];
b = [];
Aeq = @(d)[9 8 -6 -4 -6 -6 -3 -6 0 0 0 0 0 0 0 0 0
    -3 0 6 8 0 0 -6 -3 0 0 0 0 0 0 0 0 0
    -3 -8 0 -4 6 12 -3 -3 0 0 0 0 0 0 0 0 0
    -1 0 0 0 0 -2 4 4 0 0 0 0 0 0 0 0 0
    12 12 0 0 0 0 0 0 9 8 -6 -4 -6 -6 -3 -6 -12*d(1)
    0 0 12 12 0 0 0 0 -3 0 6 8 0 0 -6 -3 -12*d(2)
    0 0 0 0 12 12 0 0 -3 -8 0 -4 6 12 -3 -3 -12*d(3)
    0 0 0 0 0 0 4 4 -1 0 0 0 0 -2 4 4 -4*d(4)
    1 0.8 1.5 1 1 2 2 1.2 0 0 0 0 0 0 0 0 0];

beq = [0 0 0 0 0 0 0 0 1];
f = [1.5 1 3 2 -3 -7 -2 -1 0 0 0 0 0 0 0 0 0];

s = size(f,2); %number of variables
lb = zeros([1,s]); % lower bound of variables
ub = inf*ones([1,s]); % upper bound of variables
```

```
for i = 1:S
d = e(i,:);
[x,fval] = linprog(f,A,b,Aeq(d),beq,lb,ub);
disp(x)
disp(fval)
end
```

ii) Μέθοδος πλήρους απαρίθμησης

```
state_actions = [2,2,2,2];

rewards = { [1.5,1]
            [3,2]
            [-3,-7]
            [-2,-1]
            };

sojourn  = {[1,0.8]
            [1.5,1]
            [1,2]
            [2,1.2]
            };

probs = {[1/4,1/4,1/4,1/4; 1/3, 0, 2/3,0]
        [1/2, 1/2, 0, 0 ; 1/3,1/3,1/3,0]
        [1/2,0,1/2,0; 1/2,0,0,1/2]
        [1/4,1/2,1/4,0 ; 1/2,1/4,1/4,0]
```

```
        };

states_count = size(state_actions,2);
policies = [ones(1,states_count)];
new_p = [];
for i = 1:states_count
    temp =[];
    for j = 2:state_actions(i)
        new_p = policies;
        new_p(:,i) = new_p(:,i) + j-1;
        temp = cat(1,temp,new_p);
     end
    policies = cat(1,policies,temp);
end
policies = int32(policies);
%%

phi = zeros(size(policies));

for i = 1:size(policies,1)
    pol = policies(i,:);
    P = zeros(states_count);
    re = zeros(states_count,1);
    soj = zeros(states_count,1);

    for j = 1:states_count
        P(j,:) = probs{j}(pol(j),:);
        re(j) = rewards{j}(pol(j));
```

```
    soj(j) = sojourn{j}(pol(j));
  end

  Pstar = cesaro(P);
  disp(Pstar);
  phi(i,:) = ((Pstar*re)./(Pstar*soj))';


end
[val, argmax] = max(phi,[],1);
disp('------phi------')
disp(phi);
```

iii) Συνάρτηση eigval

```
function [evalues, repeats] = eigval(A)

% eigval  Eigenvalues and their algebraic multiplicity.
%
% evalues = eigval(A) returns the distinct eigenvalues of A,
% with duplicates removed and sorted in decreasing order.
%
% [evalues, repeats] = eigval(A) also returns the row vector
% repeats that gives the multiplicity of each eigenvalue.
% The sum of the multiplicities is n.
%
% Examples: Let A = eye(n) and B = diag([3 4]).
% For A, evalues is 1 and repeats is n.
```

```
% For B, evalues is [4; 3]  and repeats is [1 1].


tol = sqrt(eps);
lambda = sort(eig(A));
lambda = round(lambda/tol) * tol;
%
% lambda gives all n eigenvalues (repetitions included).
%
evalues = unique(lambda);
evalues = flipud(evalues);
n = length(lambda);
d = length(evalues);
A = ones(n, 1) * evalues';
B = lambda * ones(1, d);
MATCH = abs(A-B) <= tol;
%
% MATCH is an n by d zero matrix except
% MATCH(i,j) = 1 when lambda(i) = evalues(j).
% Summing the columns gives the row vector repeats.
%
repeats = sum(MATCH);
```

iv) Συνάρτηση Cesaro

```
function ces = cesaro(t)
   % get unique eigenvalues (ev) and their multiplicity (mu)
   [ev,mu] = eigval(t);
   eig1 = find(ev==1);
```

```
    %remove eigenvalue equal to 1 and its multiplicity from ev and rep
    ev(eig1) = [];
    mu(eig1) = [];

    I = eye(size(t));
    R = 1;
    for i = 1:size(ev,1)
        R = R*(t-ev(i)*I)^mu(i);
    end
    % divide R by the sum of its first row
    ces = R/sum(R(1,:));
end
```

Βιβλιογραφία

[1] Alexandru Lazari, Dmitrii Lozovanu. (2020). New algorithms for finding the limiting. *Buletinul Academiei de Stiinte a Moldevei*, pp. 92(1):75–88.

[2] D. Blackwell, T. Ferguson. (1968). The big match. *The Annals of Mathematical Statistics*, pp. 39: 159-163.

[3] *Function "eig"*. (Introduced before R2006a). Retrieved from MathWorks: https://www.mathworks.com/help/matlab/ref/eig.html

[4] *Function "linprog"*. (Introduced before R2006a). Retrieved from MathWorks: https://www.mathworks.com/help/optim/ug/linprog.html

[5] Howard, Ronald A. (1971). Dynamic Probabilistic Systems Volume II: *Semi-Markov and decision Processes.* John Wiley & Sons.

[6] Inc., T. M. (2022). MATLAB. version 9.13.0.2114483 (R2022b). Natick, Massachusetts.

[7] J.L.Doob. (1953). *Stochastic Processes.* John Willey & Sons.

[8] Jerzy Filar, Koos Vrieze (1997). *Competitive Markov decision processes.* New York: Springer.

[9] *MATLAB Teaching Codes*. (n.d.). Retrieved from web.mit.edu: http://web.mit.edu/18.06/www/Course-Info/Tcodes.html?fbclid=IwAR2C8oO2pmQeTm6edQwNc-yFpCkYSAeiabfwNrkKzwrZVTPUAcmQMKylAfo

[10] Mondal, P. (2015). Linear Programming and Zero-Sum Two-Person Undiscounted Semi-Markov Games. *Asia-Pacific Journal of Operational Research*, p. 150043(20).

[11] Mondal, P. (2017). On zero-sum two-person undiscounted semi-markov games. *Advances in Applied Probability*, pp. 49: 826-849.

[12] Mondal, P. (2017). Computing semi-stationary optimal policies for multichain semi-Markov decision processes. *Annals of Operations Research*, pp. 843-865.

[13] S. Sinha, K. G.Bakshi (2022). *Zero-Sum Two Person Perfect Information Semi-Markov Games: A Reduction*. Retrieved from arXiv:2201.00179

[14] Sagnik Sinha, P. Mondal (2017). Semi-Markov decision processes with limiting ratio average. *Journal of Mathematical Analysis and Applications*, pp. 864-871.

[15] Strang, G. (2006). *Linear Algebra and Its Applications, 4th Edition.* Cengage Learning.

[16] Κ.Α.Μηλολιδάκης (2009). *Θεωρία Παιγνίων.* Αθήνα: Εκδόσεις Σοφία.

[17] Ν.Δ. Τσάντας, Π. -Χ.Γ.Βασιλείου (2000). *Εισαγωγή στην επιχειρησιακή έρευνα.* Θεσσαλονίκη: Εκδόσεις Ζήτη.