ΠΡΟΛΟΓΟΣ

Οι μερικά παρατηρήσιμες Μαρκοβιανές διαδικασίες αποφάσεων (POMDP) αποτελούν γενίκευση των Μαρκοβιανών διαδικασιών αποφάσεων (MDP),στις οποίες οι καταστάσεις του συστήματος δεν είναι παρατηρήσιμες. Ο decision maker λαμβάνει κάποιο μήνυμα από ένα σύνολο μηνυμάτων στην αρχή κάθε χρονικής περιόδου και ακολούθως παίρνει μια απόφαση από ένα σύνολο εναλλακτικών αποφάσεων.

Εκκινώντας από ένα διάνυσμα πληροφορίας (μία κατανομή πιθανότητας για τις καταστάσεις του συστήματος), αυτό τροποποιείται στην αρχή κάθε χρονικής περιόδου με την έλευση ενός μηνύματος μέσω του τύπου του Bayes,με βάση τον πίνακα μετάβασης καταστάσεων και τον πίνακα μηνυμάτων που αντιστοιχούν στην απόφαση που είχε ληφθεί την προηγούμενη χρονική περίοδο. Το διάνυσμα πληροφορίας ενσωματώνει όλη την πληροφορία της ιστορίας του συστήματος που είναι αναγκαία για την επιλογή μιας απόφασης στην αντίστοιχη χρονική περίοδο. Για προβλήματα κόστους (εσόδων) τα άμεσα κόστη (κέρδη) εξαρτώνται από την κατάσταση του συστήματος και από την απόφαση που επιλέγεται σε μία χρονική περίοδο. Σκοπός είναι ο υπολογισμός του ελάχιστου (μέγιστου) αναμενόμενου ολικού εκπίπτοντος κόστους (κέρδους) για πεπερασμένο ή άπειρο χρονικό ορίζοντα και ο προσδιορισμός της άριστης πολιτικής.

Παρόλο που οι POMDP αποτελούν κατάλληλα υποδείγματα για πολλούς τομείς της ανθρώπινης δραστηριότητας, οι υπολογιστικές δυσκολίες καθιστούν την χρήση τους οριακή. Σε αυτό το πλαίσιο οι κύριοι στόχοι της διατριβής αυτής είναι οι ακόλουθοι:

Πρώτον, η ανάπτυξη ευέλικτων αλγόριθμων για την εύρεση άριστων ή σχεδόν άριστων λύσεων τόσο για πεπερασμένο όσο και για άπειρο χρονικό ορίζοντα.Δεύτερον, γενίκευση της συνθήκης Sondik, που εξασφαλίζει ότι μια στάσιμη πολιτική επάγει Μαρκοβιανή διαμέριση στον χώρο των διανυσμάτων πληροφορίας. Έτσι αν μία πολιτική ικανοποιεί αυτή τη συνθήκη, τότε η συνάρτηση του αναμενόμενου ολικού εκπίπτοντος κόστους για άπειρο χρονικό ορίζοντα είναι κατά τμήματα γραμμική και ο

υπολογισμός της ανάγεται στην επίλυση ενός γραμμικού συστήματος εξισώσεων. Τρίτον, εφαρμογή της POMDP σε προβλήματα συντήρησης/αντικατάστασης συστήματος όπου η κατάσταση (επίπεδο χειροτέρευσης) δεν είναι παρατηρήσιμη, αλλά λαμβάνονται μηνύματα που εξαρτώνται από την κατάσταση μέσω ενός μηχανισμού ελέγχου.

Τέταρτον, εφαρμογή της POMDP σε προβλήματα επιλογής διδακτικών μεθόδων,όπου η μαθησιακή κατάσταση της τάξης (βαθμός αφομοίωσης της διδασκόμενης ύλης) δεν είναι παρατηρήσιμη, αλλά λαμβάνονται μηνύματα τύπου επιτυχία/αποτυχία σε test.

Abstract

A Partially Observable Markov Decision Process (POMDP) is a natural extension of the Markov Decision processs (MDP).In POMDPs the state of the system is not observable and therefore unknown. Instead, the decision maker receives a random signal that depends on the state of the system at the beginning of each epoch and then he chooses an action from a finite set of actions.

Starting with an initial prior information vector, belief state,(i.e a probability distribution on the state space), it is updated at beginning of each time epoch just after the arrival of a signal.The new information vector (or belief state) is the posterior distribution on the state space using Bayes' rule that involves the transition and observation matrices assigned to the action selected at the previous time epoch.

It is well known that information vector incoporates the information of the history of the system when choosing an action at a time epoch. The immediate costs (rewards) depend on the current state and action.

The objective is the calculation of the optimal expected total discounted cost (reward) with respect to finite or infinite horizon and the determination of the optimal policy. Although POMDP may provide as suitable model for many applications they may be severely limited due to the computational complexity. Within this context the main goals of this thesis are as follows:

Firstly,the development of flexible algorithms for the determination on optimal or near optimal policies, as well as approximations of the optimal reward or (cost) functions for finite or infinite horizon.

Secondly, to find alternative conditions or generalize known conditions that ensure that a given stationary policy induces a Markovian partition of the belief state space. In this case the reward (or cost) function for infinite horizon is piecewise linear function and its evaluation is significantly simplified.

Thirdly, application of the POMDP model in problems of repair /replacement of the system. It is assumed that the system is monitored incompletely by a certain mechanism which gives the decision maker some information about the exact state of the system.

Fourthly, modeling a teaching methods selection problems as POMDP, where the state of the class (the degree of comprehension teaching material) is unknown to the teacher, and instead signals of success/failure type in tests are received.

ΠΕΡΙΛΗΨΗ

Η διατριβή οργανώνεται ως εξής**:**

Στο κεφάλαιο 1 περιγράφουμε τις Μαρκοβιανές διαδικασίες αποφάσεων (MDP), τις μερικά παρατηρήσιμες Μαρκοβιανές διαδικασίες αποφάσεων (POMDP) και συνοψίζουμε βασικά αποτελέσματα αναφορικά με τα διάφορα κριτήρια βελτιστοποίησης.Επίσης δίνουμε μία σύντομη ανασκόπηση της σχετικής βιβλιογραφίας.

Στο κεφάλαιο 2 περιγράφουμε την μέθοδο Smallwood-Sondik-Lovejoy για τον υπολογισμό της άριστης συνάρτησης κόστους (κέρδους) σε πεπερασμένο χρονικό ορίζοντα και επισημαίνουμε τις υπολογιστικές δυσκολίες αυτής της προσέγγισης.

Στο κεφάλαιο 3 παρουσιάζουμε μία νέα μέθοδο υπολογισμού της βέλτιστης συνάρτησης του αναμενόμενου ολικού εκπίπτοντος κόστους (ή κέρδους) για πεπερασμένο χρονικό ορίζοντα (αλγόριθμος των ακροτάτων σημείων). Σε γενικές γραμμές η βέλτιστη συνάρτηση κόστους (ή κέρδους) αντιπροσωπεύεται από ένα πεπερασμένο σύνολο διανυσμάτων *(«gradient vectors»), το* οποίο δεν είναι γνωστό από την αρχή, αλλά «χτίζεται» σε διαδοχικά βήματα. Σε κάθε βήμα το σύνολο αυτό

εμπλουτίζεται με νέα *«gradient vectors»* και υπολογίζεται το μέγιστο σφάλμα προσέγγισης, που είναι φθίνουσα συνάρτηση του αριθμού των βημάτων. Η διαδικασία συνεχίζεται μέχρις ότου το μέγιστο σφάλμα προσέγγισης μηδενισθεί ή γίνει αρκούντως μικρό (δηλαδή μικρότερο ή ίσο από ένα *προκαθορισμένο σφάλμα*).Ο αλγόριθμος των ακροτάτων σημείων αναφέρεται στο πέρασμα από έναν χρονικό ορίζοντα στον επόμενο. *Το συσσωρευμένο σφάλμα προσέγγισης* για κάθε χρονικό ορίζοντα υπολογίζεται μέσω απλής αναγωγικής σχέσης. Το προκαθορισμένο σφάλμα προσέγγισης στον αλγόριθμο των ακροτάτων σημείων επιλέγεται έτσι ώστε το συσσωρευμένο σφάλμα προσέγγισης για οποιοδήποτε χρονικό ορίζοντα να μην υπερβαίνει ένα επιθυμητό φράγμα.

Στο κεφάλαιο 4 εξετάζονται προσεγγίσεις της βέλτιστης συνάρτησης του αναμενόμενου ολικού εκπίπτοντος κόστους (ή κέρδους) σε άπειρο χρονικό ορίζοντα και προσδιορίζονται σχεδόν άριστες πολιτικές εφαρμόζοντας επαναληπτικά τον αλγόριθμο των ακροτάτων σημείων. Μια παρεμφερής μέθοδος που μελετάμε, είναι η επιλογή άνω και κάτω φραγμάτων ως αρχικών προσεγγίσεων της άριστης συνάρτησης κόστους (ή κέρδους), η οποία συνοδεύεται από την επαναληπτική εφαρμογή του αλγορίθμου των ακροτάτων σημείων με σκοπό τη δημιουργία νέων προσεγγίσεων. Από τις προβαλλόμενες συναρτήσεις ελέγχου των νέων προσεγγίσεων, κατασκευάζονται σχεδόν άριστες πολιτικές. Επιτάχυνση της διαδικασίας είναι δυνατή αν η απόσταση των αρχικών φραγμάτων είναι μικρή. Σε κάθε περίπτωση υπολογίζεται ο απαιτούμενος αριθμός επαναλήψεων καθώς και το προκαθορισμένο σφάλμα του αλγορίθμου των ακροτάτων σημείων, έτσι ώστε να επιτυγχάνεται προσέγγιση με οποιαδήποτε επιθυμητή ακρίβεια.

Στο κεφάλαιο 5 παρουσιάζεται η επαναληπτική μέθοδος πολιτικής (policy-iteration) για τις POMDP αναφορικά με το κριτήριο βελτιστοποίησης του αναμενόμενου ολικού εκπίπτοντος κόστους (ή κέρδους) για άπειρο χρονικό ορίζοντα

Στο πλαίσιο αυτό, ο υπολογισμός της συνάρτησης κόστους (ή κέρδους) για άπειρο χρονικό ορίζοντα που αντιστοιχεί σε μία πολιτική απλουστεύεται σημαντικά αν η πολιτική αυτή επάγει Μαρκοβιανή διαμέριση στον χώρο των διανυσμάτων πληροφορίας. Στην περίπτωση αυτή η παραπάνω συνάρτηση είναι κατά τμήματα γραμμική και ο υπολογισμός της ανάγεται στην επίλυση ενός γραμμικού συστήματος

εξισώσεων. Μία γνωστή ικανή συνθήκη ώστε μία πολιτική να επάγει Μαρκοβιανή διαμέριση είναι:η πολιτική να είναι *πεπερασμένα μεταβατική* (συνθήκη του Sondik). Παρουσιάζουμε μια νέα ικανή συνθήκη: η πολιτική να είναι *περιοδική* . Επιπλέον διατυπώνουμε μία γενικότερη συνθήκη από την πεπερασμένη μεταβατικότητα και περιοδικότητα προκειμένου μια πολιτική να επάγει Μαρκοβιανή διαμέριση στον χώρο των διανυσμάτων πληροφορίας.

Στο κεφάλαιο 6 δίνεται γεωμετρική ερμηνεία της δομής της άριστης πολιτικής για άπειρο χρονικό ορίζοντα σε ένα πρόβλημα POMDP συντήρησης/αντικατάστασης συστήματος με πεπερασμένο πλήθος καταστάσεων (επιπέδων χειροτέρευσης), το οποίο παρατηρείται ατελώς μέσω μηνυμάτων ενός μηχανισμού ελέγχου και το οποίο μελετήθηκε από τους Ohnishi-Ibaraki. Η γεωμετρική ερμηνεία αναφέρεται σε τρία επίπεδα χειροτέρευσης στο πλαίσιο της μερικής διάταξης του λόγου πιθανοφανειών.

Στο κεφάλαιο 7 μελετάται μία ειδική περίπτωση του προβλήματος συντήρησης/ αντικατάστασης συστήματος των Ohnishi-Ibaraki που περιγράφεται στο κεφάλαιο 6, με δύο καταστάσεις και δύο μηνύματα. Ειδικότερα εξετάζεται η κλάση των control-limit πολιτικών στην οποία ανήκει και η άριστη πολιτική αναφορικά με το κριτήριο βελτιστοποίησης για άπειρο χρονικό ορίζοντα. Διερευνώνται συνθήκες κάτω από τις οποίες μία control-limit πολιτική είναι πεπερασμένα μεταβατική ή περιοδική. Δίνουμε αριθμητικά παραδείγματα περιοδικών πολιτικών..

Στο κεφάλαιο 8, εξετάζεται το πρόβλημα συντήρησης/αντικατάστασης συστήματος των Ohnishi-Ibaraki ως προς το κριτήριο του μακροπρόθεσμου μέσου κόστους αν μονάδα χρόνου. Μελετάται η δομή της άριστης πολιτικής με το παραπάνω κριτήριο στο ειδικότερο πρόβλημα με δύο καταστάσεις και δύο μηνύματα.

Στο κεφάλαιο 9 μελετάται ένα πρόβλημα επιλογής ανάμεσα σε δύο διδακτικές μεθόδους,μια συμβατική και φθηνή και μια εξειδικευμένη (π.χ.ενισχυτική, υποστηριζόμενη από υπολογιστές κ.λ.π.) και δαπανηρή. Το πρόβλημα τίθεται στην μορφή POMDP με δύο δυνατές μαθησιακές καταστάσεις αναφορικά με το βαθμό αφομοίωσης της διδασκόμενης ύλης από την τάξη και δύο μηνύματα (π.χ. επιτυχία/αποτυχία σε test).Υπολογίζεται αναλυτικά η συνάρτηση του ελάχιστου αναμενόμενου ολικού εκπίπτοντος κόστους για άπειρο χρονικό ορίζοτα και

προσδιορίζεται η άριστη πολιτική επιλογής διδακτικών μεθόδων σε δύο περιπτώσεις: α)περίπτωση πλήρους αβεβαιότητας, όπου το μήνυμα (π.χ. το αποτέλεσμα ενός test) είναι ανεξάρτητο από την μαθησιακή κατάσταση της τάξης είτε επιλέγεται η συμβατική είτε η εξειδικευμένη μέθοδος διδασκαλίας και β)περίπτωση πλήρους αβεβαιότητας όταν επιλέγεται η συμβατική μέθοδος και μερικής πληροφόρησης όταν επιλέγεται η εξειδικευμένη μέθοδος.

ΚΕΦΑΛΑΙΟ 1

Εισαγωγή

Περίληψη

Σε ένα σύστημα, όπου καλούμεθα να πάρουμε αποφάσεις, οι αποφάσεις αυτές μπορεί να βασίζονται πάνω σε πολλούς παράγοντες: Γνώση των άμεσων συνθηκών, για κάποιο εξειδικευμένο πρόγραμμα, προφανή εμπειρία περί των συνεπειών των ποικίλων αποφάσεων,εμπειρικοί-κανόνες, εγκατεστημένα πρωτόκολλα κ.λ.π. Για πολλές περιπτώσεις, αυτή η προσέγγιση εργάζεται καλά, ή τουλάχιστον αρκετά καλά, ώστε να μην υπάρχει λόγος για αλλαγή στον τρόπο και την συλλογιστική που λαμβάνονται οι αποφάσεις. Ωστόσο, όσο τα συστήματα γίνονται περισσότερο πολύπλοκα, η λήψη αποφάσεων δεν είναι απλή υπόθεση, διότι υπάρχει μεγάλη αλληλεπίδραση ανάμεσα σε πολλές παραμέτρους.Αυτή η δυσκολία αυξάνεται δραματικά σε συστήματα, όταν υπάρχει υψηλός βαθμός αβεβαιότητας. Για να αντιμετωπισθεί λοιπόν η παραπάνω δυσκολία, προβλήθηκε η ανάγκη αναζήτησης νέων θεωρητικών μοντέλων, που βασίζονται στη θεωρία πιθανοτήτων, ώστε να καλυφθεί η παραπάνω αβεβαιότητα.

Το μοντέλο των μερικά παρατηρήσιμων Μαρκοβιανών διαδικασιών απόφασης σύντομα POMDP αποτελεί ένα από τα κύρια θεωρητικά μοντέλα που έχουν σαν αντικείμενο τη λήψη αποφάσεων σε συνθήκες αβεβαιότητας. Στις POMDP οι καταστάσεις του συστήματος δεν είναι παρατηρήσιμες. Στη θέση τους ο decision maker λαμβάνει μηνύματα που συνδέονται πιθανοθεωρητικά με τις καταστάσεις του συστήματος. Επομένως με αυτή την έννοια υπάρχει μερική πληροφόρηση. Τα μοντέλα POMDP έχουν πολλές εφαρμογές στον σχεδιασμό και την αντιμετώπιση πολύπλοκων συστημάτων, με ατελή πληροφόρηση.

Στην ενότητα 1.1 περιγράφουμε το μοντέλο της Μαρκοβιανής διαδικασίας αποφάσεων σύντομα M D P.

Στις ενότητες 1.2 και 1.3 κάνουμε μια ιστορική αναδρομή και περιγράφουμε το μοντέλο των POMDPs.

Στην ενότητα 1.4 ορίζονται τα δ.π τα οποία ενσωματώνουν την ιστορία του συστήματος που είναι αναγκαία για τη λήψη αποφάσεων και παρέχεται η επικαιροποίησή τους με τον κανόνα του Bayes. Επίσης ορίζονται οι τελεστές που συνδέονται με συναρτήσεις ελέγχου,ο τελεστής ελαχιστοποίησης (για προβλήματα κόστους),ο τελεστής μεγιστοποίησης (για προβλήματα εσόδων)και δίνονται οι ιδιότητές τους. Στην ενότητα 1.5 παρουσιάζουμε διάφορους τύπους πολιτικών (κανόνων αποφάσεων) καθώς και τα βασικά κριτήρια βελτιστοποίησης για πεπερασμένο και άπειρο χρονικό ορίζοντα.

1.1.Μαρκοβιανές διαδικασίες αποφάσεων με πεπερασμένο πλήθος καταστάσεων και αποφάσεων.

Μία Μαρκοβιανή διαδικασία αποφάσεων (Markov Decision Process),σύντομα MDP, είναι ένα απλό στοχαστικό υπόδειγμα , στο οποίο σε κάθε χρονική περίοδο έχουμε γνώση της κατάστασης του συστήματος πριν από τη λήψη αποφάσεων. Οι MDPs έχουν μελετηθεί στα πλαίσια του άριστου στοχαστικού ελέγχου (optimal stochastic control) και του στοχαστικού δυναμικού προγραμματισμού από τους Howard [51], Derman [24], Ross [104], Bellman [9], Puterman [98], Bertseka [11] κ.α.

Θα περιγράψουμε τώρα το παραπάνω μοντέλο των MDPs. Θεωρούμε προς τούτο ένα δυναμικό σύστημα ,του οποίου η κατάσταση επιθεωρείται στις χρονικές περιόδους (time epochs) t=0,1,2,… .

Το σύνολο S των δυνατών καταστάσεων θεωρείται πεπερασμένο,

$$S=\{1,2,\ldots,N\}.$$

Σε κάθε χρονική περίοδο, αφού παρατηρηθεί η κατάσταση του συστήματος,ο decision maker επιλέγει μία απόφαση, από ένα πεπερασμένο σύνολο εναλλακτικών αποφάσεων το οποίο συμβολίζουμε με A.

Έστω X_t η κατάσταση του συστήματος στον χρόνο t και Y_t η απόφαση που επιλέγεται στον χρόνο t. Η στοχαστική διαδικασία $\{X_t, t \in \mathbb{N}_0\}$ περιγράφεται από $N \times N$ πίνακα μετάβασης $P^{\alpha}=(p_{ij}^{a})$, $\alpha \in A$,σύμφωνα με την ακόλουθη σχέση:

Για $i,j \in S, \alpha \in A$,

$$P[X_{t+1}=j \ / \ X_t=i , X_{t-1},\ldots, X_0; Y_t=\alpha, Y_{t-1},\ldots, Y_0]$$

$$= P[X_{t+1}=j \ / \ X_t=i , Y_t=\alpha] \equiv p_{ij}^{a} , \ t \in \mathbb{N}_0.$$

Με άλλα λόγια η πιθανότητα μετάβασης του συστήματος σε μια κατάσταση την επόμενη χρονική περίοδο, εξαρτάται αποκλειστικά από την κατάσταση και την απόφαση που επιλέχθηκε στην τρέχουσα χρονική περίοδο (Μαρκοβιανή ιδιότητα). Επίσης εισάγεται μια δομή κέρδους (εσόδων) ή κόστους ανάλογα με το πρόβλημα. Για τα προβλήματα κέρδους θεωρούμε ότι $q(i,a)$ είναι το άμεσο κέρδος (immediate reward) στον χρόνο t ,όταν η κατάσταση του συστήματος είναι i και επιλέγεται η απόφαση α.

Για τα προβλήματα κόστους αντίστοιχα θεωρούμε ότι $c(i,a)$ είναι το άμεσο κόστος (immediate cost) στον χρόνο t ,όταν η κατάσταση είναι i *και επιλέγεται η απόφαση* a. Συνοψίζοντας μια MDP αναφορικά με πρόβλημα εσόδων, περιγράφεται από την τετράδα $(S,A,(P^a)_{a\in A}, q(.,.))$.Ανάλογα περιγράφεται μια MDP,που αναφέρεται σε πρόβλημα κόστους.

Η ιστορία του συστήματος στον χρόνο t συμβολίζεται με h_t και περιλαμβάνει τις καταστάσεις του συστήματος, καθώς και τις αποφάσεις που επιλέχθηκαν στους χρόνους 0,1,…,t,

δηλαδή

$$h_t = (X_0, Y_0, X_1, Y_1, \ldots, X_t, Y_t), \quad t = 1,2,\ldots$$

$$h_0 = (X_0, Y_0).$$

Το πεδίο τιμών της ιστορίας h_t είναι το σύνολο

$$H_t = (S \times A)^{t+1}.$$

Μία πολιτική ή στρατηγική (policy-strategy) ορίζεται ένας μηχανισμός λήψης αποφάσεων στις χρονικές περιόδους $t = 0,1,2,\ldots$. Σε πλήρη γενικότητα η επιλογή της

απόφασης στον χρόνο t μέσω της πολιτικής δ γίνεται σύμφωνα με μια κατανομή πιθανότητας , που εξαρτάται από το ζεύγος (h_{t-1}, X_t):

$$\{\delta_t(a/h_{t-1}, X_t) : \alpha \in \mathrm{A}\}, \qquad \textbf{1.1.1}$$

όπου $\delta_t(a/h_{t-1}, X_t) \geq 0 \quad \forall a \in A$ και $\sum_{a \in A} \delta_t(a/h_{t-1}, X_t) = 1.$

Με D συμβολίζουμε την κλάση όλων των πολιτικών.

Ορισμός 1.1.1:Μια πολιτική δ καλείται αμνήμων (memoryless policy),αν σε κάθε χρονική περίοδο t η κατανομή πιθανότητας (1.1.1) που επάγεται από τη δ εξαρτάται μόνο από την κατάσταση X_t, δηλαδή για κάθε $a \in A$, $h_{t-1} \in H_{t-1}$,

$$\delta_t(a/h_{t-1}, X_t) = \delta_t(a/X_t).$$

Με D_A συμβολίζουμε το σύνολο των αμνημόνων πολιτικών.

Ορισμός 1.1.2:Μια αμνήμων πολιτική δ καλείται γνήσια ή μη τυχαιοποιημένη (nonrandomized policy) αν σε κάθε χρονική περίοδο t η κατανομή πιθανότητας $\{\delta_t(a/X_t) : a \in A\}$ είναι εκφυλισμένη, δηλαδή $\delta_t(a/X_t)$=0 ή 1, $a \in A$.

Σημειώνουμε ότι η παραπάνω εκφυλισμένη κατανομή πιθανότητας μπορεί να εκφρασθεί ως συνάρτηση ελέγχου (control function) $\delta_t : S \to A$

με $\delta_t(i) = a^* \Leftrightarrow \delta_t(a^*/X_t = i) = 1.$

Συμπεραίνουμε ότι μία γνήσια πολιτική δ μπορεί να θεωρηθεί ως χρονική ακολουθία συναρτήσεων ελέγχου $\{\delta_t : t \in \square_0\}$ και παριστάνεται ως

$$\delta = (\delta_0, \delta_1, ...)$$

Με D_Γ συμβολίζουμε το σύνολο των γνήσιων πολιτικών.

Ορισμός 1.1.3:Μια γνήσια πολιτική δ καλείται στάσιμη (stationary policy), αν οι συναρτήσεις ελέγχου στις χρονικές περιόδους t=0,1,2,… ταυτίζονται:

$$\delta_t = \delta_0 \quad \forall t = 1, 2, ..., \text{ δηλαδή } \delta = (\delta_0, \delta_0, ...).$$

Συνήθως μία γνήσια στάσιμη πολιτική συμβολίζεται

$$\delta^{\infty} = (\delta, \delta, ...),$$

όπου δ είναι συνάρτηση ελέγχου (control function) $\delta : S \to A$.

Με D_Σ συμβολίζουμε το σύνολο των γνήσιων στάσιμων πολιτικών.

Προφανώς

$$D_\Sigma \subset D_\Gamma \subset D_A \subset D.$$

Θα περιγράψουμε εν συντομία δύο κριτήρια βελτιστοποίησης. Για περισσότερη ανάλυση βλέπε Bertseka [11] και Derman [24].

Περιοριζόμαστε σε MDP για προβλήματα εσόδων. Για προβλήματα κόστους έχουμε ανάλογη αντιμετώπιση.

1) Κριτήριο βελτιστοποίησης για πεπερασμένο χρονικό ορίζοντα

Θεωρούμε τον χρονικό ορίζοντα $T \geq 1$. Το αναμενόμενο ολικό εκπίπτον κέρδος για τον χρονικό ορίζοντα T, όταν η αρχική κατάσταση του συστήματος είναι $X_0 = i$ και εφαρμόζουμε την πολιτική δ, γράφεται:

$$E_\delta \left[\sum_{t=0}^{T-1} \beta^t .q(X_t, Y_t) + \beta^T .q(X_T) / X_0 = i \right], i \in S, \qquad \textbf{1.1.2}$$

όπου $\beta > 0$ είναι ο συντελεστής έκπτωσης (discount factor) και q(j) είναι το (άμεσο) κέρδος τερματισμού (terminal reward), όταν η κατάσταση του συστήματος στον χρόνο περάτωσης T είναι j.

Επιθυμούμε να μεγιστοποιήσουμε την (1.1.2) πάνω στην κλάση όλων των πολιτικών D και να καθορίσουμε την άριστη πολιτική για την οποία επιτυγχάνεται το μέγιστο.(Για την ύπαρξη άριστης πολιτικής βλέπε Derman [24] και Denardo [23]).

Έστω $V_n(i)$ το βέλτιστο (μέγιστο) αναμενόμενο ολικό εκπίπτον κέρδος,όταν απομένουν $n \leq T$ χρονικές περίοδοι μέχρι το πέρας του χρονικού ορίζοντα T και η κατάσταση του συστήματος στον χρόνο T-n είναι i $(X_{T-n} = i)$. Η συνάρτηση $V_n(i), i \in S$ καλείται βέλτιστη συνάρτηση τιμών για χρονικό ορίζοντα n και υπολογίζεται από την ακόλουθη αναγωγική σχέση του δυναμικού προγραμματισμού.

$$V_n(i) = \max_a \{q(i,\alpha) + \beta . \sum_{j=1}^{N} p_{ij}^a .V_{n-1}(j)\}, i \in S. \qquad \textbf{1.1.3}$$

$$V_0(i) = q(i), i \in S$$

Η παράσταση εντός της αγκύλης στην (1.1.3)

$$q(i,\alpha)+\beta.\sum_{j=0}^{N} p_{ij}^{a}.V_{n-1}(j)$$

εκφράζει το αναμενόμενο ολικό εκπίπτον κέρδος όταν απομένουν n χρονικές περίοδοι μέχρι το πέρας του χρονικού ορίζοντα T, στη χρονική περίοδο T-n η κατάσταση του συστήματος είναι i $(X_{T-n}=i)$, επιλέγεται η απόφαση α $(Y_{T-n}=a)$ και ακολουθούμε βέλτιστη πορεία για τις εναπομένουσες n-1 χρονικές περιόδους. Είναι φανερό ότι η άριστη πολιτική για τον πεπερασμένο χρονικό ορίζοντα T είναι η γνήσια μη στάσιμη πολιτική (non stationary policy) $\delta^{*}=(\delta_{T}^{*}, \delta_{T-1}^{*},..., \delta_{1}^{*})$, όπου η συνάρτηση ελέγχου δ_{n}^{*} υπολογίζεται από τη σχέση:

$$\delta_{n}^{*}(i)=\arg\max_{a}\{q(i,\alpha)+\beta.\sum_{j=1}^{N} p_{ij}^{a}.V_{n-1}(j)\}, i\in S, \qquad \underline{\mathbf{1.1.4}}$$

$n=1,2,\ldots,T.$

2) Κριτήριο βελτιστοποίησης για άπειρο χρονικό ορίζοντα

Το αναμενόμενο ολικό εκπίπτον κέρδος για άπειρο χρονικό ορίζοντα, όταν η αρχική κατάσταση του συστήματος $X_0 = i$ και εφαρμόζουμε την πολιτική δ, γράφεται:

$$V_{\delta}(i)=E_{\delta}\left[\ \sum_{t=0}^{\infty}\beta^{t}.q(X_t,Y_t)/X_0=i\ \right], i\in S, \qquad \underline{\mathbf{1.1.5}}$$

όπου για τον συντελεστή εκπτώσεως υποθέτουμε ότι $\beta\in[0,1)$.Αποδεικνύεται εύκολα ότι για κάθε $\delta\in D$,

$$\left|\ V_{\delta}(i)\ \right|\le\frac{\Lambda}{1-\beta}\quad\forall\ i\in S,$$

όπου $$\Lambda\equiv\max_{i,a}\left|q(i,a)\ \right|.$$

Η συνάρτηση $V_{\delta}(i)$, $i\in S$ αναφέρεται ως συνάρτηση τιμών για την πολιτική δ^{∞}. Επιθυμούμε να μεγιστοποιήσουμε την (1.1.5) πάνω στην κλάση όλων των πολιτικών D και να καθορίσουμε την άριστη πολιτική για την οποία επιτυγχάνεται το μέγιστο. Αποδεικνύεται ότι υπάρχει γνήσια στάσιμη πολιτική που είναι άριστη (δηλαδή μεγιστοποιεί την (1.1.5)).(βλέπε και Maitra (1968) [79]).

Επομένως μπορούμε να περιορισθούμε στην κλάση των γνήσιων στάσιμων πολιτικών D_Σ.

Θεωρούμε την γνήσια στάσιμη πολιτική $\delta^\infty = (\delta, \delta, ..., \delta)$. Η συνάρτηση τιμών V_δ για την πολιτική δ^∞ είναι η μοναδική λύση της εξίσωσης βελτιστοποίησης

$$V_\delta(i) = q(i,\delta(i)) + \beta . \sum_{j=1}^{N} p_{ij}^{\delta(i)} . V_\delta(j), \quad i=1,2,\ldots,N . \qquad \textbf{1.1.6}$$

Η βέλτιστη συνάρτηση τιμών

$$V^*(i) = \sup_{\delta \in D} V_\delta(i) = \sup_{\delta^\infty \in D_\Sigma} V_\delta(i) \ , i \in S .$$

Είναι η μοναδική λύση της εξίσωσης βελτιστοποίησης

$$V^*(i) = \max_a \{ q(i,\alpha) + \beta . \sum_{j=1}^{N} p_{ij}^a . V^*(j) \} \ , i \in S . \qquad \textbf{1.1.7}$$

Η παράσταση εντός της αγκύλης στην (1.1.7)

$$q(i,\alpha) + \beta . \sum_{j=1}^{N} p_{ij}^a . V^*(j)$$

εκφράζει το αναμενόμενο ολικό εκπίπτον κέρδος, όταν στον χρόνο t=0 η κατάσταση είναι $X_0 = i$,επιλέγεται η απόφαση $Y_0=\alpha$ και κατόπιν ακολουθείται βέλτιστη πορεία. Η συνάρτηση ελέγχου της άριστης πολιτικής $(\delta^*)^\infty = (\delta^*, \delta^*, ...)$ προσδιορίζεται από τη σχέση:

$$\delta^*(i) = \arg \max_a \ [q(i,\alpha) + \beta . \sum_{j=1}^{N} p_{ij}^a . V^*(j)], i \in S . \qquad \textbf{1.1.8}$$

Επομένως για τον προσδιορισμό της άριστης πολιτικής $(\delta^*)^\infty$ είναι αναγκαίος ο υπολογισμός της άριστης συνάρτησης τιμών V^*.Παραθέτουμε δύο μεθόδους υπολογισμού της V^*.

Α) Μέθοδος των διαδοχικών προσεγγίσεων ή επαναληπτική μέθοδος τιμών (method of successive approximations,value-iteration method).

Με τη μέθοδο αυτή επιλέγουμε αυθαίρετα μία συνάρτηση u (i), $i \in S$ και ακολούθως υπολογίζουμε την ακολουθία των συναρτήσεων $\{ V_n, n \in \square \}$ μέσω της αναγωγικής σχέσης

$$V_n(i) = \max_a \{q(i,\alpha) + \beta . \sum_{j=1}^{N} p_{ij}^a . V_{n-1}(j)\}, i \in S \qquad \textbf{1.1.9}$$

$$V_0(i) = u(i), i \in S$$

Η συνάρτηση V_n είναι η βέλτιστη συνάρτηση τιμών για χρονικό ορίζοντα n με συνάρτηση κέρδους τερματισμού u.

Αποδεικνύεται ότι:

$$V_n \underset{n\to\infty}{\to} V^*$$

(βλέπε και Ross [107]).

Ορισμός 1.1.4: Μια πολιτική δ^∞ καλείται $\varepsilon - \acute{\alpha}\rho\iota\sigma\tau\eta$, όπου $\varepsilon > 0$, αν η συνάρτηση τιμών για την δ^∞, V_δ, αποτελεί προσέγγιση της άριστης συνάρτησης τιμών V^* με μέγιστο σφάλμα προσέγγισης μικρότερο ή ίσο του ε, δηλ.

$$\max_{i \in S} |V_\delta(i) - V^*(i)| \le \varepsilon$$

Πρόταση 1.1.1: Θεωρούμε τις συναρτήσεις V_n ,n=0,1,2,… οι οποίες υπολογίζονται μέσω της αναγωγικής σχέσης (1.1.9).

Αν για κάποιο $n \ge 1$,

$$\max_{i \in S} |V_n(i) - V_{n-1}(i)| \le \varepsilon \quad ,\text{όπου} \quad \varepsilon > 0,$$

τότε η πολιτική δ^∞ με συνάρτηση ελέγχου που υπολογίζεται από τη σχέση

$$\delta(i) = \arg\max_a [q(i,\alpha) + \beta . \sum_{j=1}^{N} p_{ij}^a . V_{n-1}(j)], i \in S$$

είναι $\frac{2\beta\varepsilon}{1-\beta} - \acute{\alpha}\rho\iota\sigma\tau\eta.$ (βλέπε και Bellman [9]).

Αλγόριθμος A_1 (Value-iteration) **(**Howard [51],Bellman [9]**)**

1. Input: Μιά αρχική αυθαίρετη συνάρτηση τιμών V_n με n=0,και μια παράμετρος $\varepsilon > 0$,για την επίτευξη ε-βέλτιστης πολιτικής.

2. Improve value function: Αυξάνουμε το n και για κάθε $i \in S$,

$$V_n(i) = \max_a q(i,\alpha) + \beta . \sum_j p_{ij}^a . V_{n-1}(j),$$

3.Convergence -test: Αν

$$\max_{i \in S} |V_n(i) - V_{n-1}(i)| \le \frac{\varepsilon.(1-\beta)}{2\beta}$$

πήγαινε στο βήμα 4. Άλλως πήγαινε στο βήμα 2.

4.Output:Μια ε-βέλτιστη πολιτική δ^∞ με συνάρτηση ελέγχου

$$\delta(i) := \arg \max_a [q(i,\alpha) + \beta . \sum_j p_{ij}^a . V_{n-1}(j)],\ i \in S .$$

Β) Επαναληπτική μέθοδος πολιτικής (policy iteration method).

Η μέθοδος αυτή αναπτύχθηκε από τους Howard [51] και Blackwell [15], [16] και περιλαμβάνει δύο στάδια. Κατά το πρώτο στάδιο (policy evaluation) υπολογίζεται η συνάρτηση τιμών V_δ μιας γνήσιας στάσιμης πολιτικής δ^∞,μέσω του συστήματος των εξισώσεων (1.1.6).Κατά το δεύτερο στάδιο (policy improvement) εντοπίζεται μία βελτιωμένη πολιτική $(\delta')^\infty$.Κατόπιν εφαρμόζεται το πρώτο στάδιο στην πολιτική $(\delta')^\infty$,και η διαδικασία συνεχίζεται μέχρις ότου καταλήξουμε σε άριστη πολιτική. Η πρόταση που ακολουθεί διευκρινίζει το δεύτερο στάδιο της βελτιωμένης πολιτικής.

Πρόταση 1.1.2: Εστω δ^{∞} γνήσια στάσιμη πολιτική και V_δ η αντίστοιχη συνάρτηση τιμών .Θεωρούμε την ακόλουθη συνάρτηση ελέγχου δ΄ :

$$\delta'(i) := \arg\max_a \{ q(i,\alpha) + \beta . \sum_{j=1}^{N} p_{ij}^{a} . V_\delta(j) , \quad i \in S .$$

Τότε για την συνάρτηση τιμών $V_{\delta'}$ της πολιτικής $(\delta')^{\infty}$ ισχύει:

$$V_{\delta'}(i) = q(i,\delta(i)) + \beta . \sum_{j=1}^{N} p_{ij}^{\delta(i)} . V_\delta(j) , \quad i=1,2,\ldots,N .$$

$$V_{\delta'}(i) \geq V_\delta(i) , \forall\ i \in S ,$$

.

και εάν $V_{\delta'}(i) = V_\delta(i) \ \forall\ i \in S$ τότε:

$V_{\delta'} = V_\delta = V^*$ (άριστη συνάρτηση τιμών).

Βλέπε και Howard [51].

Αλγόριθμος A_2 (policy-Iteration)

1.Input: Μια αρχική γνήσια στάσιμη πολιτική $\delta^{\infty} = (\delta, \delta, ..., \delta)$.

2.Policy evaluation: Υπολογίζουμε την συνάρτηση τιμών V_δ, για την πολιτική δ^{∞}, λύνοντας το σύνολο των εξισώσεων (1.1.6).

3.Policy improvement: Για κάθε κατάσταση $i \in S$, αν υπάρχει κάποια απόφαση $\alpha \in A$ ώστε: $q(i,\alpha) + \beta . \sum_j p_{ij}^{a} . V_\delta(j) > V_\delta(i)$,

τότε $\delta'(i)=\alpha$, αλλιώς , $\delta'(i)=\delta(i)$.

4.Convergence -test: Αν δ' είναι η ίδια με την δ, τότε πάμε στο βήμα 5. Αλλιώς θέτουμε δ=δ' και πάμε στο βήμα 2

5. Output: Μία άριστη πολιτική δ^{∞} και άριστη συνάρτηση τιμών $V^* = V_\delta$.

Επειδή τα σύνολα S,A υποτέθηκαν πεπερασμένα ,το πλήθος των δυνατών συναρτήσεων ελέγχου είναι πεπερασμένο ($|A|^N$).Επομένως και το σύνολο των γνήσιων στάσιμων

πολιτικών D_Σ είναι πεπερασμένο με πληθάριθμο $|D_\Sigma| = |A|^N$.Επειδή σε κάθε επανάληψη βελτιώνεται η πολιτική στο βήμα 3,και το πλήθος των δυνατών γνήσιων στάσιμων πολιτικών είναι πεπερασμένο, ο αλγόριθμος A_2 τερματίζεται σε πεπερασμένο πλήθος επαναλήψεων.

1.2. Μερικά παρατηρήσιμες Μαρκοβιανές διαδικασίες αποφάσεων, (ιστορική αναδρομή).

Μια μερικά παρατηρήσιμη Μαρκοβιανή διαδικασία,σύντομα POMDP,είναι μια γενικευμένη (Markov decision process), που επιτρέπει ατελή πληροφόρηση του συστήματος των καταστάσεων. Η γενίκευση αυτή, είναι σημαντική σε προβλήματα όπου η αβεβαιότητα ως προς την κατάσταση είναι το κεντρικό και ουσιώδες. Στην πραγματικότητα έχουμε ένα ευρύ πεδίο εφαρμογών του μοντέλου POMDP, Goulionis [36,37,38,44,45] και το κοινό σημείο όλων αυτών των εφαρμογών, είναι η αβεβαιότητα ως προς την κατάσταση στην οποία βρίσκεται το σύστημα, και η επίδραση αυτής της αβεβαιότητας στην επιλογή μιας βέλτιστης πολιτικής. Το μοντέλο POMDP επίσης εξαναγκάζει τον ερευνητή να κάνει μια ξεκάθαρη διάκριση μεταξύ πραγματικών καταστάσεων και μηνυμάτων.

Παράλληλα ανάλογα με την «κατάσταση» που φαίνεται να βρίσκεται το σύστημα (belief-state) λαμβάνεται μια απόφαση. Όταν κάποιος καλείται να πάρει αποφάσεις βασίζεται σε ολόκληρη την ιστορία του συστήματος, ένα σύνολο δηλαδή από αποφάσεις και μηνύματα,που έχουν ήδη ληφθεί. Η POMDP συνήθως μετατρέπεται σε μια ισοδύναμη MDP,όπου ο χώρος καταστάσεων είναι η δεσμευμένη πιθανότητα κατανομής της κατάστασης του συστήματος, δοσμένης της ιστορίας του (Astrom 1965) [5],[6].

Έρευνα για τις POMDPs άρχισε την δεκαετία του 1960 από τους Howard [51] και Drake [26],που ανέπτυξαν το πιο απλό μοντέλο. Το 1965 ο Astrom [5] διατύπωσε το μοντέλο για τις μερικά παρατηρήσιμες MDPs σε πεπερασμένο χρονικό ορίζοντα.

Οι θεμελιωτές όμως της θεωρίας POMDP είναι οι Smallwood και Sondik [117,118,119,120],και κυρίαρχα ο δεύτερος.Αυτοί έδωσαν το έναυσμα για ικανοποιητικούς αλγόριθμους. Τα κύρια σημεία του μοντέλου είναι ο ορισμός μίας στοχαστικής διαδικασίας καταστάσεων (core-process)και μιας στοχαστικής διαδικασίας μηνυμάτων. Η στοχαστική διαδικασία καταστάσεων σχηματίζει μια Μαρκοβιανή διαδικασία, και δεν μπορεί να παρατηρηθεί απευθείας. Η στοχαστική διαδικασία μηνυμάτων είναι μια ακολουθία από καταστάσεις, που πραγματικά παρατηρούνται, αποφασίζεται μέσω της (core-process) και δεν είναι απαραίτητα Μαρκοβιανή.

Ο Sondik το 1971 [117] διατύπωσε τον αλγόριθμο ενός βήματος (one –pass-algorithm).Απέδειξε δύο ουσιαστικά πράγματα, που έκαναν την μέχρι τότε σχεδόν άβολη υπολογιστική διαδικασία αρκετά εφικτή, και την θεωρία γόνιμη και ρεαλιστική στην αντιμετώπιση προβλημάτων. Πρώτα απέδειξε ότι η βέλτιστη συνάρτηση τιμών σε πεπερασμένο χρονικό ορίζοντα, έχει δύο σημαντικές ιδιότητες, δηλαδή είναι κατά τμήματα γραμμική και κυρτή (piecewise –linear and convex) p.w.l.c. Κατόπιν ήλθε σαν άμεσο αποτέλεσμα της *κατά τμήματα γραμμικότητας και κυρτότητας*, ότι η παραπάνω συνάρτηση για κάθε χρονικό ορίζοντα Τ, μπορεί να αντιπροσωπευθεί χρησιμοποιώντας τα λεγόμενα «gradients vectors».

Επεκτείνοντας τις σκέψεις του στο πρόβλημα του άπειρου χρονικού ορίζοντα, εισήγαγε την κλάση των πεπερασμένα μεταβατικών πολιτικών, και ανέπτυξε προσεγγίσεις για κάθε στάσιμη πολιτική, που βασίζονται ακριβώς στις πεπερασμένα μεταβατικές πολιτικές, δείχνοντας παράλληλα ότι οι τομές των συναρτήσεων οφέλους, που βασίζονται σε μια τέτοια προσέγγιση, μπορούν να συμπεριληφθούν στον αλγόριθμο του Howard [51] (policy-improvement) με επακόλουθη σύγκλιση. Δηλαδή οι πεπερασμένα μεταβατικές πολιτικές παίζουν έναν ρόλο κλειδί, διότι γενικεύουν δυναμικές ισοδύναμες με εκείνες των MDPs. Ο Denardo [23] έδωσε μια πιο βολική μορφή στα «gradients-vectors» και εισήγαγε τους τελεστές H_{δ},Η στην αντιπροσώπευση της βέλτιστης συνάρτησης τιμών σε κάθε χρονικό ορίζοντα.

Πολλοί ερευνητές ασχολήθηκαν με την επίτευξη ενός πιο λειτουργικού από την άποψη μοντέλου, διότι όπως απέδειξε ο Mukherjee [87] ο αριθμός των « gradient-

vectors» αυξάνει εκθετικά καθώς αυξάνεται ο χρονικός ορίζοντας, με αποτέλεσμα η όλη διαδικασία του δυναμικού προγραμματισμού να είναι υπολογιστικά ανέφικτη και το μοντέλο μη γόνιμο στην αντιμετώπιση ρεαλιστικών προβλημάτων.

Ο Eagle (1984) [29] χρησιμοποίησε την POMDP προκειμένου να μελετήσει κινούμενο στόχο.

Ο Albright (1979) [1] έδωσε συνθήκες, ώστε ή βέλτιστη πολιτική για ένα σύστημα δύο καταστάσεων να είναι μονότονη ως προς την κατανομή πιθανότητας των καταστάσεων του συστήματος.

Ο Platzman (1980) [96] ανέπτυξε τις συνθήκες για να είναι καλά ορισμένο το πρόβλημα σε άπειρο χρονικό ορίζοντα (undiscounded -infinite - horizon POMDP).

Ο Lovejoy (1987)[76] παρείχε ικανοποιητικές συνθήκες, που αποφέρουν μονότονες βέλτιστες πολιτικές για το πρόβλημα POMDPs σε πεπερασμένο χρονικό ορίζοντα.

Ο Littman(1995) [69],[70],[71],[72] έδωσε μια πιο τυποποιημένη μορφή στο πρόβλημα του άπειρου χρονικού ορίζοντα, και απέδειξε πολύ απλά την ιδιότητα της κατά τμήματα γραμμικής συνάρτησης, ενώ σύνδεσε το εργαλείο που λέγεται δυναμικός προγραμματισμός, με τα (policy - trees).

Η απλούστερη έκδοση του μοντέλου που θα αναπτυχθεί είναι το μοντέλο Μαρκοβιανής εξέλιξης. Η πρώτη διατύπωσή του έγινε με τις εργασίες του Bellman το 1957 [9], μολονότι προηγήθηκαν οι εργασίες του Pollock [97] που αφορούσαν τα στοχαστικά παίγνια.

Βέβαια οι εργασίες των Howard [51] εφαρμόζοντας δυναμικό προγραμματισμό, Manne[80] εφαρμόζοντας γραμμικό προγραμματισμό(linear-programming formulation), Blackwell [14] που επέκτεινε το πρόβλημα σε αυθαίρετους χώρους καταστάσεων και Ross [104] ήταν οι θεμέλιοι λίθοι. Το μοντέλο των POMDPs είναι ένα τμήμα του δυναμικού προγραμματισμού και δίνει χρήσιμα και επιτυχημένα εργαλεία σε επιχειρήσεις, που ασχολούνται με ένα τέτοιο είδος πολύπλοκων αποφάσεων. Πέρα από αυτό,βρίσκεται στις παρυφές της (artificial-intelligence-community),με την συλλογιστική, ότι αρκετές φορές απαιτείται, ή είναι επιθυμητό, να πάρουμε μια ακολουθία αποφάσεων χωρίς την συμμετοχή ανθρώπινου παράγοντα, βλέπε Madani [78].Υπάρχει στενή σύνδεση ανάμεσα σε επιχειρησιακή έρευνα (δυναμικό

προγραμματισμό) και τεχνητή νοημοσύνη (artificial-intelligence) βλέπε και Zhang [142].

1.3.Μερικά παρατηρήσιμη Μαρκοβιανή διαδικασία αποφάσεων πεπερασμένου πλήθους καταστάσεων, περιγραφή.

Στην ενότητα αυτή θα περιγράψουμε την μερικά παρατηρήσιμη Μαρκοβιανή διαδικασία αποφάσεων πεπερασμένου πλήθους καταστάσεων (finite state partially observable Markov decision process) ή σύντομα POMDP.

Το σύνολο A των εναλλακτικών αποφάσεων, που έχει στη διάθεσή του ο decision maker (action space) θεωρείται πεπερασμένο.Οι αποφάσεις επιλέγονται σε διακριτούς χρόνους t=0,1,2,3…

Η στοχαστική διαδικασία των αποφάσεων συμβολίζεται με $\{Y_t, t\in \square_0\}$.

Θεωρούμε ότι το σύνολο των καταστάσεων του συστήματος είναι S={1,2,3…,N}.Η στοχαστική διαδικασία $\{X_t, t\in \square_0\}$ των καταστάσεων καλείται διαδικασία πυρήνα (core-process) και υποτίθεται είναι μια (πεπερασμένη) Μαρκοβιανή διαδικασία που περιγράφεται από έναν $N\times N$ πίνακα μετάβασης $P^{\alpha} = (p_{ij}^{a})$ $\alpha\in A$, σύμφωνα με την ακόλουθη σχέση:

Για $i, j \in S$,$\alpha\in A$,

$$p[X_{t+1}=j / X_t=i, X_{t-1},\ldots,X_0; Y_t=\alpha, Y_{t-1},\ldots,Y_0] = p[X_{t+1}=j / X_t=i, Y_t=\alpha] \equiv p_{ij}^{a},\ t\in \square_0.$$

Με άλλα λόγια η πιθανότητα μετάβασης του συστήματος σε μία κατάσταση κάποια χρονική περίοδο (time epoch),εξαρτάται αποκλειστικά από την κατάσταση του συστήματος καθώς και από την απόφαση που επιλέχθηκε την προηγούμενη περίοδο. Θεωρούμε ότι η διαδικασία πυρήνα δεν είναι άμεσα παρατηρήσιμη,δηλαδή ο decision maker δεν λαμβάνει γνώση της κατάστασης του συστήματος στον χρόνο t=0,1,2,…

Ο decision maker λαμβάνει ωστόσο στον χρόνο t ένα μήνυμα από ένα σύνολο μηνυμάτων Θ={1,2,3,…M}.Η στοχαστική διαδικασία μηνυμάτων $\{Z_t, t\in \square_0\}$συνδέεται με τη διαδικασία πυρήνα $\{X_t, t\in \square_0\}$ μέσω της ακόλουθης σχέσης:

Για $i, j \in S$, $\theta \in \Theta$, $\alpha \in A$,

$p[Z_{t+1}=\theta/Z_t,\ldots, Z_1;X_{t+1}=\iota,X_t,\ldots,X_0;Y_t=\alpha,Y_{t-1},\ldots,Y_0]=p[Z_{t+1}=\theta/X_{t+1}=\iota ,Y_t=\alpha] \equiv r_{i\theta}^{\alpha}$,t$\in$ $_0$.

Με άλλα λόγια η πιθανότητα με την οποία λαμβάνεται ένα μήνυμα κάποια χρονική περίοδο (time epoch),εξαρτάται αποκλειστικά από την κατάσταση του συστήματος την ίδια χρονική περίοδο και την απόφαση που επιλέχθηκε την προηγούμενη περίοδο. Οι στοχαστικοί $N \times M$ πίνακες $R^a = (r_{i\theta}^{\alpha})$,καλούνται πίνακες μηνυμάτων. Σημειώνουμε ότι σε κάθε χρονική περίοδο, η απόφαση λαμβάνεται μετά τη λήψη του μηνύματος. Αναλυτικότερα η σειρά με την οποία συμβαίνουν τα γεγονότα θεωρείται η ακόλουθη:Αρχικά για (t=0) το σύστημα βρίσκεται στην κατάσταση X_0,επιλέγεται μια απόφαση Y_0 και στην αρχή της χρονικής περιόδου t=1 το σύστημα μεταβαίνει στην κατάσταση X_1, λαμβάνεται ένα μήνυμα Z_1 και ακολούθως επιλέγεται η απόφαση Y_1. Γενικά στην περίοδο t το σύστημα βρίσκεται στην κατάσταση X_t,λαμβάνεται ένα μήνυμα Z_t και κατόπιν επιλέγεται μία απόφαση Y_t .

Στην αρχή της περιόδου t+1 το σύστημα μεταβαίνει στην κατάσταση X_{t+1},λαμβάνεται μήνυμα Z_{t+1},κατόπιν επιλέγεται η απόφαση Y_{t+1} κ.ο.κ.

Για να γίνουν κατανοητά τα παραπάνω, δίνουμε ένα τυπικό παράδειγμα POMDP. Όταν εξετάζουμε την κατάσταση ενός ασθενούς που πάσχει από στεφανιαία νόσο, το αποτέλεσμα του λεγόμενου τέστ κόπωσης, το επίπεδο ισχαιμίας, καθώς και ο πόνος στο στήθος είναι μηνύματα, που συνυφαίνονται με την ζωτική κατάσταση του ασθενούς. Ωστόσο, δεν γνωρίζουμε την κατάσταση στην οποία βρίσκεται ο εν λόγω ασθενής.

Τέλος εισάγεται μία δομή κέρδους (εσόδων) ή δομή κόστους ,ανάλογα με το πρόβλημα. Για τα προβλήματα κέρδους, θεωρούμε ότι q(i,α) είναι το άμεσο κέρδος (immediate reward) στον χρόνο t, όταν η κατάσταση του συστήματος είναι *i* και λαμβάνεται η απόφαση α.Το διάνυσμα άμεσου κέρδους που αντιστοιχεί στην απόφαση α συμβολίζεται με q^{α} και θεωρείται διάνυσμα στήλη.

$$q^{\alpha}=(q(1,\alpha),q(2,\alpha),\ldots,q(N,\alpha))^T.$$

Με ανάλογο τρόπο σε προβλήματα κόστους εισάγεται το άμεσο κόστος (immediate cost) στον χρόνο t, όταν η κατάσταση του συστήματος είναι *i* και λαμβάνεται η απόφαση α. Το διάνυσμα άμεσου κόστους που αντιστοιχεί στην απόφαση α συμβολίζεται

$$c^{\alpha} = (c(1,\alpha),c(2,\alpha),\ldots,c(N,\alpha))^{T}.$$

Συνοψίζοντας, μία **POMDP**, αναφορικά με πρόβλημα εσόδων περιγράφεται από την εξάδα $(S,A,\Theta,(P^{\alpha})_{\alpha\in A}, (R^{\alpha})_{\alpha\in A}, (q^{\alpha})_{\alpha\in A})$. Για προβλήματα κόστους έχουμε ανάλογη αντιμετώπιση.

1.4. Μετατροπή μίας POMDP σε πλήρως παρατηρήσιμη MDP

Θεωρούμε μία **POMDP** όπως περιγράφθηκε στην ενότητα 1.3. Η κατανομή πιθανότητας της αρχικής κατάστασης του συστήματος θεωρείται γνωστή στον decision maker και συμβολίζεται με

$$\pi(0)= (\pi_1(0),\ldots, \pi_N(0))$$

όπου $\pi_i(0) \equiv p[X_0 = i], i = 1,2,...,N$.

Η ιστορία του συστήματος στον χρόνο t, συμβολίζεται με h_t και περιλαμβάνει όλη την πληροφορία (δεδομένα) που είναι διαθέσιμη πριν από τη λήψη απόφασης στον χρόνο t. Συγκεκριμένα η ιστορία h_t περιλαμβάνει την κατανομή πιθανότητας $\pi(0)$ της αρχικής κατάστασης, τα μηνύματα που πήραμε στους χρόνους 1,2,…,t καθώς και τις αποφάσεις που επιλέχθηκαν στους χρόνους 0,1,…,t-1 ,δηλαδή

$$h_t = (\pi(0), Y_0, Z_1, ..., Y_{t-1}, Z_t), t = 1,2,...$$

$$h_0 = \pi(0).$$

Προφανώς $h_t = (h_{t-1}, Y_{t-1}, Z_t), t = 1,2,...$

Το πεδίο τιμών της ιστορίας h_t είναι το σύνολο

$$H_t = \Pi \times (A \times \Theta)^t$$

όπου $\Pi = \{x \in \mathbb{R}^N : \sum_{i=1}^{N} x_i = 1, x_i \geq 0, i = 1,2,...,N\}$

(το σύνολο των κατανομών πιθανότητας στον χώρο S).

As θεωρήσουμε ένα πρόβλημα εσόδων για πεπερασμένο χρονικό ορίζοντα $T \geq 1$. Αν η συνάρτηση $\upsilon_t(h_t), h_t \in H_t$ δηλώνει το βέλτιστο (μέγιστο) αναμενόμενο ολικό εκπίπτον όφελος από τον χρόνο t μέχρι το πέρας του χρονικού ορίζοντα T ($t \leq T$) ,τότε:

$$\upsilon_t(h_t) = \max_{Y_t \in A}\{E[q(X_t, Y_t) + \beta.\upsilon_{t+1}(h_{t+1}) / h_t, Y_t]\}$$

$$= \max_{Y_t \in A} \{E\,[q(X_t, Y_t) / h_t, Y_t] + \beta.E\,[\upsilon_{t+1}(h_{t+1}) / h_t, Y_t]$$

$$= \max_{Y_t \in A}\left\{\sum_{i=1}^{N} p\left(X_t = i \middle| h_t\right).q(i, Y_t) + \beta \cdot \sum_{\theta \in \Theta} p\left(Z_{t+1}=\theta \middle| h_t, Y_t\right) \cdot \upsilon_{t+1}\left(h_t, Y_t, Z_{t+1} = \theta\right)\right\},$$

$$h_t \in H_t\ ,\ t=0,\ldots,T\text{-}1 \qquad \textbf{1.4.1}$$

όπου β είναι ο παράγοντας έκπτωσης (discount factor).Υποθέτουμε ότι $\beta > 0$.Για τον χρόνο περατώσεως t=T παίρνουμε:

$$\upsilon_T(h_T) = \sum_{i=1}^{N} p[\,X_T = i / h_T\,].\,q(i)\,, \quad h_T \in H_T\,, \qquad \textbf{1.4.2}$$

όπου q(i) είναι το άμεσο κέρδος τερματισμού (terminal -reward),όταν η κατάσταση του συστήματος είναι η ι .

Οι πιθανότητες που υπεισέρχονται στις (1.4.1),(1.4.2) υπολογίζονται στη συνέχεια αυτής της ενότητας.

Αν αντιμετωπίζουμε πρόβλημα κόστους και $\upsilon_t\left(h_t\right), h_t \in H_t$ είναι το βέλτιστο (ελάχιστο) αναμενόμενο ολικό εκπίπτον κόστος από τον χρόνο t έως τον χρονικό ορίζοντα T ($t \le T$),τότε παίρνουμε ανάλογες εκφράσεις με την (1.4.1) με τις προφανείς αλλαγές $c(X_t,Y_t)$ αντί $q(X_t,Y_t)$ και $\min_{Y_T \in A}$ αντί $\max_{Y_T \in A}$. Τέλος αν c(i) είναι το άμεσο κόστος περατώσεως (terminal –cost ή salvage- cost) όταν η κατάσταση του συστήματος είναι η ι παίρνουμε αντίστοιχη προς την (1.4.2) σχέση για τον χρόνο t=T.

Ο υπολογισμός των συναρτήσεων υ_t γίνεται κατά την ανάδρομη χρονική φορά. Πρώτα υπολογίζεται η υ_T μέσω της (1.4.2) και κατόπιν υπολογίζονται αναγωγικά οι συναρτήσεις $\upsilon_{T-1},\upsilon_{T-2},...,\upsilon_0$ μέσω της (1.4.1).Σημειώνουμε ότι η συνάρτηση υ_t στον χρόνο t πρέπει να υπολογισθεί για κάθε δυνατή ιστορία $h_t \in H_t$. Για δοσμένη κατανομή πιθανότητας π(0), το πλήθος των δυνατών ιστοριών στον χρόνο t είναι $\left|H_t\right| = (\left|A\right|.\left|\Theta\right|)^t$.

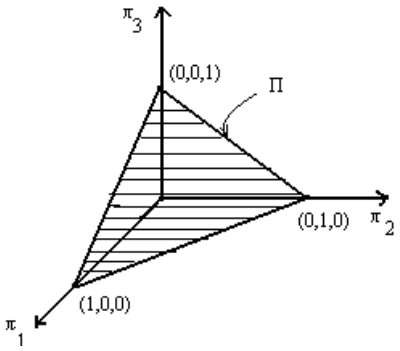
Οι υπολογιστικές απαιτήσεις ενός τέτοιου DP(dynamic-programming) αλγόριθμου ενδέχεται να είναι τεράστιες.(Οι λεπτομέρειες αυτού του προβλήματος σχολιάζονται στον Bertseka [11] και Bertseka-Shreve [115]).

Έστω $\pi_i(t) \equiv p[X_t = i / h_t]$,$i = 1, 2, ...,N.$

και $\pi(t) = (\pi_1(t), \pi_2(t), ..., \pi_N(t))$,

όπου $\sum_{i=1}^{N} \pi_i(t) = 1$, $0 \le \pi_i(t) \le 1$.

Η κατανομή πιθανότητας π(t) της κατάστασης του συστήματος στον χρόνο t καλείται διάνυσμα πληροφορίας «διάνυσμα πληροφορίας» information vector ή (belief-state) σύντομα δ.π .



Σχήμα 1.1:*Το σύνολο Π για N=3.*

Η ακολουθία $\{\pi(t), t \in \mathbb{N}_0\}$ αποτελεί στοχαστική διαδικασία επειδή εξαρτάται από τη στοχαστική διαδικασία της ιστορίας $\{h_t : t \in \mathbb{N}_0\}$.Αποδεικνύεται (Dynkin [28]) ότι:

$$\pi_i(t+1) = p\left(X_{t+1} = i \middle| h_t, Y_t, Z_{t+1}\right) = p\left(X_{t+1} = i \middle| \pi(t), Y_t, Z_{t+1}\right), i \in S$$

και

$$\Pr\left(Z_{t+1} = \theta \middle| h_t, Y_t\right) = \Pr\left(Z_{t+1} = \theta \middle| \pi(t), Y_t\right), \theta \in \Theta.$$

Αν είναι γνωστό το δ.π και η απόφαση στο χρόνο t ,καθώς και το μήνυμα στον χρόνο t+1,μπορούμε να υπολογίσουμε το δ.π στον χρόνο t+1.Πιο συγκεκριμένα, αν $\pi(t) = \pi, Y_t = \alpha, Z_{t+1} = \theta,$ εφαρμόζοντας τον κανόνα Bayes' παίρνουμε:

$$T_j(\pi,\theta,\alpha) \equiv \pi_j(t+1) = \Pr(X_{t+1} = j / \pi(t) = \pi, Y_t = a, Z_{t+1} = \theta)$$

$$= \frac{p(Z_{t+1} = \theta / X_{t+1} = j, Y_t = a).p(X_{t+1} = j / \pi(t) = \pi, Y_t = a)}{p(Z_{t+1} = \theta / \pi(t), Y_t = a)}$$

$$= \frac{r_{j\theta}^{\alpha} \sum_{i=1}^{N} \cdot p_{ij}^{\alpha} .\pi_i}{\sum_{K=1}^{N} r_{k\theta}^{\alpha} \sum_{i=1}^{N} \cdot p_{iк}^{\alpha} .\pi_i}, \quad j=1,2,\ldots,N. \qquad \textbf{1.4.3}$$

Το διάνυσμα πληροφορίας στον χρόνο t+1

$$T(\pi, \theta, \alpha) = (T_1(\pi(t),\theta,\alpha),\ldots,T_N(\pi(t),\theta,\alpha),$$

γράφεται σε μορφή πινάκων

$$T(\pi,\theta,\alpha) = \frac{\pi \cdot P^{\alpha} \cdot R_{\theta}^{\alpha}}{\pi \cdot P^{\alpha} \cdot R_{\theta}^{\alpha} \cdot 1} \qquad \textbf{1.4.4}$$

όπου R_{θ}^{α} είναι ο $N \times N$ διαγώνιος πίνακας με τα διαγώνια στοιχεία (j,j) ίσα με $r_{j\theta}^{\alpha}$, δηλαδή:

$$R_{\theta}^{\alpha} = diag(r_{1\theta}^{\alpha}, r_{2\theta}^{\alpha}, \ldots r_{N\theta}^{\alpha}) = \begin{pmatrix} r_{1\theta}^{\alpha} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & r_{N\theta}^{\alpha} \end{pmatrix}$$

και **1** είναι το $N \times 1$ διάνυσμα στήλη με όλα τα στοιχεία του 1.

Ο παρανομαστής των 1.4.3) και (1.4.4) δηλώνει την πιθανότητα το επόμενο μήνυμα να είναι θ, δεδομένου ότι το τρέχον δ.π είναι π, και η τρέχουσα απόφαση είναι η α. και συμβολίζεται με {θ/π,α},δηλ.

$$\{\theta/\pi,\alpha\} \equiv p(Z_{t+1} = \theta / \pi(t) = \pi, Y_t = a)$$

$$= \sum_{k=1}^{N} p(Z_{t+1} = \theta / X_{t+1} = k, Y_t = a).\, p(X_{t+1} = k / \pi(t) = \pi, Y_t = a)$$

$$= \sum_{k=1}^{N} r_{\kappa\theta}^{\alpha} \cdot \sum_{i=1}^{N} p_{ik}^{a} .\pi_i$$

$$= \pi.P^{\alpha}.R^{\alpha}{}_{\theta}.\mathbf{1} \qquad \textbf{1.4.5}$$

Ισοδύναμα η ποσότητα {θ/π,α} δηλώνει την πιθανότητα το επόμενο δ.π να είναι T(π,θ,α) δοσμένου ότι το τρέχον δ.π είναι , π, και η τρέχουσα απόφαση είναι α, δηλαδή:

$$\{\theta/\pi,\alpha\} = p[\,\pi(t+1) = T(\pi,\theta,\alpha)\,/\,\pi(t) = \pi, Y_t = a\,]\,.$$

Το δ.π, π(t), είναι μια επαρκής στατιστική για την ιστορία h_t ,δηλαδή ενσωματώνει όλη την αναγκαία πληροφορία, προκειμένου να επιλεγεί μια απόφαση στον χρόνο t, βλέπε και (*Bertsekas* [11], *Monahan* [85],*Sondik* [117], *Striebel* [121]).

Επιπλέον ισχύει το ακόλουθο θεώρημα:

Θεώρημα 1.4.1:Για κάθε σταθερή ακολουθία αποφάσεων $Y_0, Y_1, \ldots$, η στοχαστική διαδικασία $\{\pi(t), t \in \square_0\}$είναι Μαρκοβιανή, δηλαδή αν $\Gamma \subset \Pi$,τότε:

$$p\left(\pi(t+1) \in \Gamma \,|\, \pi(0), \pi(1), \ldots \pi(t), Y_t\right) = Pr\left(\pi(t+1) \in \Gamma \,|\, \pi(t), Y_t\right).$$

Βλέπε και Aoki [3], Astrom [6]. □

Με βάση τα παραπάνω αποτελέσματα, ένα πρόβλημα POMDP μετατρέπεται σε ένα ισοδύναμο (πλήρως παρατηρήσιμο) πρόβλημα MDP με χώρο καταστάσεων το σύνολο Π των κατανομών πιθανότητας στον χώρο S, το οποίο είναι το (N-1)-simplex του χώρου $\square^N$.

Μια σημαντική ιδιότητα της συνάρτησης μεταφοράς $T(\pi,\theta,\alpha), \pi \in \Pi$ είναι ότι μετασχηματίζει ευθύγραμμα τμήματα σε ευθύγραμμα τμήματα. Συγκεκριμένα ισχύει η ακόλουθη πρόταση.

Πρόταση 1.4.1: Έστω $\theta \in \Theta$,$\alpha \in A$,$\pi^1, \pi^2 \in \Pi$, $0 \le \lambda \le 1$.

Τότε

$$T(\lambda.\pi^1 + (1-\lambda).\pi^2), \theta, \alpha) = \nu.\, T(\pi^1,\theta,\alpha) + (1-\nu).T(\pi^2,\theta,\alpha)\ ,$$

όπου $\nu = \lambda.\{\theta/\pi^1,\alpha\} / \lambda.\{\theta/\pi^1,\alpha\} + (1-\lambda).\{\theta/\pi^2,\alpha).$ □

Με βάση την παραπάνω πρόταση, το ευθύγραμμο τμήμα $\lambda.\pi^1 + (1-\lambda).\pi^2$, $0 \le \lambda \le 1$ του χώρου Π μετασχηματίζεται μέσω της $T(.,\theta,\alpha)$ *στο ευθύγραμμο τμήμα*

$$\nu.T(\pi^1,\theta,\alpha) + (1-\nu).T(\pi^2,\theta,\alpha),\ \ 0 \le \nu \le 1.$$

Στη συνέχεια της ενότητας αυτής θα ορίσουμε τελεστές, που αποτελούν πολύ χρήσιμα εργαλεία στη μελέτη της POMDP.

Με $F(\Pi)$ συμβολίζουμε το σύνολο των πραγματικών συναρτήσεων με πεδίο ορισμού το σύνολο Π,

Με $B(\Pi)$ συμβολίζουμε το σύνολο των φραγμένων πραγματικών συναρτήσεων με πεδίο ορισμού το σύνολο Π,

$$B(\Pi) \subset F(\Pi).$$

Με $\|.\|$ συμβολίζουμε τη νόρμα supremum :

Για $u \in F(\Pi)$, $\|u\| := \sup_{\pi \in \Pi} |u(\pi)|$.

Θεωρούμε τη συνάρτηση $h : \Pi \times A \times B(\Pi) \to \square$,η οποία για προβλήματα εσόδων ορίζεται ως

$$h(\pi,\alpha,u) := \pi.q^{\alpha} + \beta. \sum_{\theta} \{\theta/\pi,\alpha\}.u(T(\pi,\theta,\alpha)),$$

$$(\pi, \alpha,u) \in \Pi \times A \times B(\Pi)$$

ενώ για προβλήματα κόστους ορίζεται ως

$$h(\pi,\alpha,u) := \pi.c^{\alpha} + \beta. \sum_{\theta} \{\theta/\pi,\alpha\}.u(T(\pi,\theta,\alpha)),$$

$$(\pi, \alpha,u) \in \Pi \times A \times B(\Pi),$$

όπου $\beta > 0$ είναι ο συντελεστής έκπτωσης .

Εισάγουμε τώρα τους ακόλουθους τελεστές:

1) *Θεωρούμε τη συνάρτηση ελέγχου* $\delta: \Pi \to A$.

Ο τελεστής $H_{\delta} : B(\Pi) \longrightarrow F(\Pi)$

ορίζεται ως εξής: Για $u \in B(\pi)$,

$$H_{\delta}u(\pi) := h(\pi, \delta(\pi), u), \pi \in \Pi.$$

Στην ειδική περίπτωση όπου η συνάρτηση ελέγχου δ είναι σταθερή, $\delta(\pi)=\alpha$ $\forall \pi \in \Pi (\alpha \in A)$, ο τελεστής συμβολίζεται με H_{α} και έχουμε

$$H_{\alpha}u(\pi) := h(\pi, \alpha, u), \pi \in \Pi.$$

2) Ο τελεστής $H : B(\Pi) \longrightarrow F(\Pi)$

για προβλήματα εσόδων ορίζεται ως εξής: Για $u \in B(\pi)$,

$$Hu(\pi) := \max_{\alpha \in A} H_{\alpha}u(\pi)$$

$$= \max_{a \in A} \{\pi.q^{\alpha} + \beta \sum_{\theta} \{\theta / \pi, a\} .u\ (T(\pi/\theta,\alpha))\},\ \pi \in \Pi.$$

(τελεστής μεγιστοποίησης).

Για προβλήματα κόστους ορίζεται ως

$$Hu(\pi) := \min_{a \in A} H_{\alpha} u(\pi)$$

$$= \min_{a \in A} \{\pi.c^{\alpha} + \beta. \sum_{\theta} \{\theta / \pi, a\} .u\ (T(\pi/\theta,\alpha))\},\ \pi \in \Pi.$$

(τελεστής ελαχιστοποίησης).

Αποδεικνύεται εύκολα ότι οι τελεστές H_{δ},H έχουν τις ακόλουθες ενδιαφέρουσες και χρήσιμες ιδιότητες:είναι φραγμένοι, ισότονοι και συστολές modulus β.(βλέπε Bertsekas [11]).Πιο συγκεκριμένα,

Αν $L= H_{\delta}$,H τότε:

i) Φραγμένο: $\| \ Lu \ \| \leq \Lambda + \beta . \|u\| < \infty \quad \forall u \in B(\Pi)$,

όπου $\Lambda := \max_{i \in S, a \in A} |q(i,a)|$ για προβλήματα εσόδων,

και $\Lambda := \max_{i \in S, a \in A} |c(i,a)|$ για προβλήματα κόστους.

Επομένως $Lu \in B(\Pi)$, $\forall\ u \in B(\Pi)$.

ii) Ισοτονία: Αν $\upsilon, u \in B(\Pi)$ *με* $u \geq \upsilon$ *, τότε* $Lu \geq L\upsilon$.

iii) Συστολή: $\|Lu - L\upsilon \| \leq \beta . \|u - \upsilon\| \ \forall\ \upsilon, u \in B(\Pi)$.

Αν για τον συντελεστή έκπτωσης ισχύει $0 < \beta < 1$, συμπεραίνουμε ότι οι τελεστές H_{δ} , H έχουν μοναδικά σταθερά σημεία (fixed points).

Έστω $w \in B(\Pi)$ το σταθερό σημείο του τελεστή L

(όπου $L = H_{\delta}$,H), δηλαδή $w = Lw$.

Θεωρώντας την επαναληπτική σχέση

$$w_n = Lw_{n-1},\ n = 1, 2, ...$$

η ακολουθία $\{w_n\}$ συγκλίνει ομαλά όταν $n \to \infty$ στο σταθερό σημείο w ,ανεξάρτητα από την επιλογή της αρχικής συνάρτησης $w_0 \in B(\Pi)$.

1.5. Πολιτικές και κριτήρια βελτιστοποίησης για προβλήματα POMDP.

Μία πολιτική ή στρατηγική (policy,strategy) σε μία POMDP *ορίζεται ως ένας μηχανισμός λήψης αποφάσεων στις χρονικές περιόδους t=0,*1,….Σε πλήρη γενικότητα η επιλογή της απόφασης στον χρόνο t μέσω της πολιτικής δ γίνεται με μία κατανομή πιθανότητας η οποία εξαρτάται από την ιστορία του συστήματος $h_t \in H_t$

$$\delta_t(a/h_t), a \in A,$$

όπου $\delta_t(a/h_t) \geq 0, a \in A,$ και $\sum_{a \in A} \delta_t(a/h_t) = 1.$

Μπορούμε να θεωρήσουμε ισοδύναμα , ότι η κατανομή πιθανότητας εξαρτάται από το δ.π $\pi(t) \in \Pi$.

$$\delta_t(a/\pi(t)), a \in A.$$

επειδή το π(t) ενσωματώνει όλη την πληροφορία σχετικά με την ιστορία του συστήματος στον χρόνο t.

Με D συμβολίζουμε την κλάση όλων των πολιτικών.

Ορισμός 1.5.1:Μία πολιτική λέγεται γνήσια ή μη τυχαιοποιημένη(nonrandomized) αν για κάθε χρονική περίοδο t η κατανομή πιθανότητας $\delta_t(a/\pi(t)), a \in A.$ είναι εκφυλισμένη, δηλαδή

$$\delta_t(a/\pi(t)) = 0 \ \acute{\eta} \ 1, \ a \in A.$$

Σημειώνουμε ότι η παραπάνω εκφυλισμένη κατανομή πιθανότητας μπορεί να εκφρασθεί ως συνάρτηση ελέγχου (control function) $\delta_t : \Pi \to \mathrm{A}$

με $\delta_t(\pi) = \alpha^* \Leftrightarrow \delta_t(a^*/\pi(t) = \pi) = 1.$

Συμπεραίνουμε ότι μια γνήσια πολιτική δ μπορεί να θεωρηθεί ως χρονική ακολουθία συναρτήσεων ελέγχου $\{\delta_t : t \in \square_0\}$ και παριστάνεται ως

$$\delta = (\delta_0, \delta_1, ...).$$

Με D_Γ συμβολίζουμε το σύνολο των γνήσιων πολιτικών.

Ορισμός 1.5.2:Μία γνήσια πολιτική δ καλείται στάσιμη (stationary) αν οι συναρτήσεις ελέγχου στις χρονικές περιόδους t=0,1,… ταυτίζονται: $\delta_t = \delta_0 \ \forall t = 1, 2, ...,$δηλαδή

$$\delta = (\delta_0, \delta_0, ...).$$

Συνήθως μία γνήσια στάσιμη πολιτική συμβολίζεται $\delta^{\infty} = (\delta, \delta, ...)$,όπου δ είναι συνάρτηση ελέγχου $\delta : \Pi \to A$.

Με D_Σ συμβολίζουμε το σύνολο των γνήσιων στάσιμων πολιτικών. Προφανώς

$$D_\Sigma \subset D_\Gamma \subset D.$$

Θα περιγράψουμε εν συντομία δύο κριτήρια βελτιστοποίησης. Περιοριζόμαστε σε POMDP για προβλήματα εσόδων. Η περιγραφή των κριτηρίων αυτών για προβλήματα κόστους είναι ανάλογη.

1) <u>Κριτήριο βελτιστοποίησης για πεπερασμένο χρονικό ορίζοντα.</u>

Θεωρούμε τον χρονικό ορίζοντα $T \geq 1$.Το αναμενόμενο ολικό εκπίπτον κέρδος για τον χρονικό ορίζοντα T , όταν το αρχικό δ.π είναι π(ο)=π και εφαρμόζουμε την πολιτική δ γράφεται:

$$V_T(\pi / \delta) \equiv \mathrm{E}_\delta \left[\sum_{t=0}^{T-1} \beta^t .q(X_t, Y_t) + \beta^{\mathrm{T}} .q(X_T) / \pi(o) = \pi \right], \pi \in \Pi. \qquad \underline{\mathbf{1.5.1}}$$

όπου $\beta > 0$ είναι ο συντελεστής έκπτωσης και q(j) είναι το (άμεσο) κέρδος τερματισμού (terminal reward),όταν η κατάσταση του συστήματος στον χρόνο περάτωσης T είναι j. Επιθυμούμε να μεγιστοποιήσουμε την (1.5.1), πάνω στην κλάση όλων των πολιτικών **D** και να καθορίσουμε την άριστη πολιτική για την οποία επιτυγχάνεται το παραπάνω μέγιστο.

Έστω $V_n(\pi)$ το βέλτιστο (μέγιστο) αναμενόμενο ολικό εκπίπτον κέρδος, όταν απομένουν $n \leq T$ χρονικές περίοδοι μέχρι το πέρας του χρονικού ορίζοντα T και το δ.π. στον χρόνο T-n είναι π (π(T-n)=π).

Η συνάρτηση $V_n(\pi)$, $\pi \in \Pi$ καλείται βέλτιστη συνάρτηση τιμών για χρονικό ορίζοντα n και υπολογίζεται από την ακόλουθη σχέση του δυναμικού προγραμματισμού. Για n=1,2,…,T

$$V_n(\pi) = \mathrm{H}V_{n-1}(\pi)$$

$$= \max_a \{\pi.q^a + \beta.\sum_\theta \{\theta / \pi, \alpha\}.V_{n-1}(T(\pi, \theta, \alpha))\}, \pi \in \Pi \ . \underline{\mathbf{1.5.2}}$$

$$V_0(\pi) = \pi.q$$

όπου $q=(q(1),\ldots,q(N))^T$ είναι το διάνυσμα των άμεσων κερδών τερματισμού.

Η παράσταση εντός της αγκύλης στην (1.5.2)

$$\pi.q^{a} + \beta.\sum_{\theta}\{\theta / \pi, \alpha\}.V_{n-1}(T(\pi, \theta, \alpha))$$

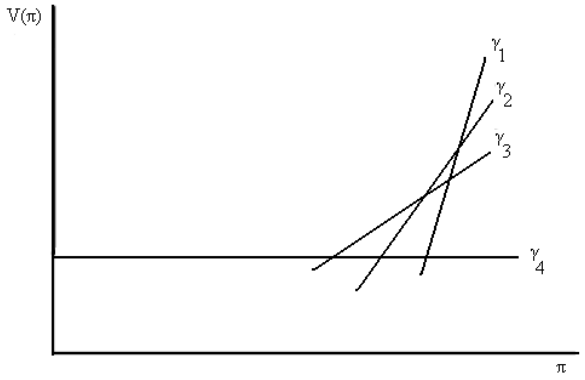
εκφράζει το αναμενόμενο ολικό εκπίπτον κέρδος, όταν απομένουν n χρονικές περίοδοι μέχρι το πέρας του χρονικού ορίζοντα Τ, στην χρονική περίοδο Τ-n ,το δ.π. είναι το π (π(Τ-n)=π),επιλέγεται η απόφαση α ,($Y_{t-n} = a$) και ακολουθείται βέλτιστη πορεία για τις εναπομένουσες n-1 χρονικές περιόδους. Είναι φανερό ότι η άριστη πολιτική για τον πεπερασμένο χρονικό ορίζοντα Τ είναι η γνήσια μη στάσιμη πολιτική $\delta^{*} = (\delta_{\mathrm{T}}^{*}, \delta_{\mathrm{T}-1}^{*}, ..., \delta_{1}^{*})$, όπου η συνάρτηση ελέγχου δ_{n}^{*} υπολογίζεται από τη σχέση:

$$\delta_{n}^{*}(\pi) = \arg\max_{a}\{\pi.q^{\alpha} + \beta.\sum_{\theta}\{\theta / \pi, \alpha\}V_{n-1}(T(\pi, \theta, \alpha))\}, \pi \in \Pi, n = 1, 2, ..., T,$$

Οι Smallwood and Sondik [119] έδειξαν ότι για $n = 1, 2, ..., T$,η βέλτιστη συνάρτηση τιμών $V_n(\pi)$,$\pi \in \Pi$ είναι συνεχής, κατά τμήματα γραμμική και κυρτή (piecewise linear and convex)(p.w.l.c),δηλαδή:

$$V_{n}(\pi) = \max\{\pi.\gamma : \gamma \in \Gamma_{n}\}, \pi \in \Pi \qquad \textbf{1.5.3}$$

όπου Γ_n είναι πεπερασμένο σύνολο διανυσμάτων του χώρου $\mathbb{R}^{N}$.Τα διανύσματα $\gamma \in \Gamma_n$ θεωρούνται διανύσματα στήλες. Ο Lovejoy [74] εισήγαγε τον όρο «gradients-vectors» για τα διανύσματα γ. Με βάση τη σχέση (1.5.3) η συνάρτηση $V_n(\pi)$, $\pi \in \Pi$ είναι η άνω επιφάνεια (επιγραφή) των υπερεπιπέδων $\pi.\gamma$, $\gamma \in \Gamma_n$.



Σχήμα1.2:Κατά τμήματα γραμμική και κυρτή συνάρτηση με δύο καταστάσεις.

Στα προβλήματα κόστους η συνάρτηση $V_n(\pi)$, $\pi \in \Pi$ του ελάχιστου αναμενόμενου ολικού εκπίπτοντος κόστους *(βέλτιστη συνάρτηση τιμών)* για χρονικό ορίζοντα $n \geq 1$ υπολογίζεται με ανάλογο τρόπο:

$$V_n(\pi) = HV_{n-1}(\pi)$$

$$= \min_a \{\pi.c^a + \beta.\sum_\theta \{\theta/\pi,\alpha\}.V_{n-1}(T(\pi,\theta,\alpha))\}, \pi \in \Pi \qquad \textbf{1.5.4}$$

$$V_0(\pi) = \pi.c\,,$$

όπου c = (c(1),c(2),…,c(N))^T είναι το διάνυσμα του άμεσου κόστους τερματισμού. Η άριστη πολιτική για χρονικό ορίζοντα T είναι η γνήσια μη στάσιμη πολιτική δ*=($\delta_T^*, \delta_{T-1}^*, \ldots, \delta_1^*$), όπου η συνάρτηση ελέγχου δ_n^* υπολογίζεται από τη σχέση:

$$\delta_n^*(\pi) = \arg\min_a \{\pi.c^\alpha + \sum_\theta \{\theta/\pi,\alpha\} V_{n-1}(T(\pi,\theta,\alpha))\}, \pi \in \Pi, n = 1,2,...,T\,.$$

Επίσης η συνάρτηση $V_n(\pi)$ είναι συνεχής, κατά τμήματα γραμμική και κοίλη (piecewise linear and concave).

Δηλαδή

$$V_n(\pi) = \min\{\pi.\gamma : \gamma \in \Gamma_n\}, \pi \in \Pi\,. \qquad \textbf{1.5.5}$$

Θα ασχοληθούμε με το κριτήριο αυτό και ειδικότερα με αλγόριθμους υπολογισμού της $V_n(\pi)$,$\pi \in \Pi$ στα κεφάλαια 2 και 3.

2) <u>Κριτήριο βελτιστοποίησης για άπειρο χρονικό ορίζοντα.</u>

Το αναμενόμενο ολικό εκπίπτον κέρδος για άπειρο χρονικό ορίζοντα, όταν το αρχικό δ.π είναι π(ο)=π και εφαρμόζουμε την πολιτική δ γράφεται:

$$V(\pi/\delta) := \mathrm{E}_\delta\left[\sum_{t=0}^{\infty} \beta^t.q(X_t,Y_t)/\pi(o) = \pi\right], \pi \in \Pi. \qquad \textbf{1.5.6}$$

όπου για τον συντελεστή εκπτώσεως υποθέτουμε ότι $\beta \in (0,1)$.Αποδεικνύεται εύκολα ότι για κάθε $\delta \in D$

$$\left| V(\pi/\delta) \right| \le \frac{\Lambda}{1-\beta}, \quad \forall \pi \in \Pi\,.$$

όπου $\Lambda \equiv \max_{i,a} |q(i,a)|$.

Η συνάρτηση $V(\pi/\delta), \pi \in \Pi$ *αναφέρεται ως συνάρτηση τιμών για την πολιτική δ.* Αποδεικνύεται (Blackwell [15]), ότι υπάρχει γνήσια στάσιμη πολιτική ,η οποία είναι άριστη, δηλαδή μεγιστοποιεί την (1.5.4). Επιπλέον η βέλτιστη συνάρτηση τιμών

$$V^*(\pi) := \sup_{\delta \in D} V(\pi/\delta) = \sup_{\delta^\infty \in D_\Sigma} V(\pi/\delta), \pi \in \Pi,$$

είναι η μοναδική λύση της εξίσωσης βελτιστοποίησης

$$V^*(\pi) = \max_a \{\pi.q^a + \beta.\sum_\theta \{\theta/\pi,\alpha\}.V^*\ (T(\pi,\theta,\alpha))\}, \pi \in \Pi. \qquad \textbf{1.5.7}$$

Η παράσταση εντός της αγκύλης στην (1.5.7)

$$\pi.q^a + \beta.\sum_\theta \{\theta/\pi,\alpha\}.V^*\ (T(\pi,\theta,\alpha))$$

εκφράζει το αναμενόμενο ολικό εκπίπτον κέρδος, όταν στον χρόνο t=0 , το δ.π. είναι το π(0)=π, επιλέγεται η απόφαση α,($Y_0 = a$) και κατόπιν ακολουθείται άριστη πορεία. Η συνάρτηση ελέγχου στην άριστη πολιτική $(\delta^*)^\infty = (\delta^*, \delta^*, ...,)$, προσδιορίζεται από τη σχέση:

$$\delta^*(\pi) = \arg \max_a \{\pi.q^\alpha + \sum_\theta \{\theta/\pi,\alpha\} V^*(T(\pi,\theta,\alpha))\}, \pi \in \Pi.$$

Χρησιμοποιώντας τον τελεστή μεγιστοποίησης Η (βλέπε ενότητα 1.4), η σχέση (1.5.7) γράφεται

$$V^* = HV^* \qquad \textbf{1.5.8}$$

Επομένως η βέλτιστη συνάρτηση τιμών V^*για άπειρο χρονικό ορίζοντα είναι το σταθερό σημείο του τελεστή Η. Λαμβάνοντας υπόψη την ιδιότητα της συστολής modulus β του τελεστή Η και την επαναληπτική σχέση (1.5.2), η ακολουθία βέλτιστων συναρτήσεων τιμών για πεπερασμένους χρονικούς ορίζοντες $\{V_n\}$συγκλίνει ομαλά όταν $n \to \infty$στην συνάρτηση V^*(βλέπε ενότητα 1.4). Επιπλέον οι ιδιότητες της κυρτότητας και της συνέχειας μεταφέρονται στο όριο. Με άλλα λόγια η συνάρτηση V^*είναι κυρτή και συνεχής.

Με ανάλογο τρόπο στα προβλήματα κόστους η βέλτιστη συνάρτηση τιμών V^* για άπειρο χρονικό ορίζοντα ικανοποιεί την εξίσωση αριστοποίησης:

$$V^*(\pi) = \min_a \{\pi.c^a + \beta.\sum_\theta \{\theta/\pi,\alpha\}.V^*(T(\pi,\theta,\alpha))\}, \pi \in \Pi. \qquad \textbf{1.5.9}$$

Η (1.5.9) γράφεται:

$$V^* = HV^* \qquad \textbf{1.5.10}$$

όπου Η είναι τελεστής ελαχιστοποίησης.

Επομένως η συνάρτηση V^* είναι το σταθερό σημείο του τελεστή Η. Επίσης η V^* είναι συνεχής και κοίλη συνάρτηση. Σημειώνουμε ακόμη ότι η γνήσια στάσιμη πολιτική

$$(\delta^*)^\infty = (\delta^*, \delta^*, ...)$$

με συνάρτηση ελέγχου: $\delta^*(\pi) = \arg \min_a \{\pi.c^\alpha + \sum_\theta \{\theta / \pi, \alpha\} V^*(T(\pi, \theta, \alpha))\}, \pi \in \Pi$,

είναι άριστη πολιτική. Με μεθόδους προσέγγισης της συνάρτησης V^* και της άριστης πολιτικής $(\delta^*)^\infty$ θα ασχοληθούμε στο κεφάλαιο 4. Καλές αρχικές αναφορές για την δομή των MDPs και επέκταση αυτών στις POMDPs υπάρχουν στις εργασίες των Puterman [98],[99],Bertseka [12],Monahan [85], Lovejoy [74] και κύρια του White [129,130, 131,132,133].

ΣΥΜΠΕΡΑΣΜΑΤΑ

Στο κεφάλαιο αυτό δώσαμε μία σύντομη ανασκόπηση της βιβλιογραφίας των μοντέλων MDP και POMDP.Επίσης περιγράψαμε τα παραπάνω μοντέλα, παρουσιάσαμε τα κύρια κριτήρια βελτιστοποίησης, και συνοψίσαμε βασικά αποτελέσματα που θα χρησιμοποιηθούν στην συνέχεια