«ΣΠΟΥΔΑΙ», Τόμος 55, Τεύχος 3ο, (2005) / «SPOUDAI», Vol. 55, No 3, (2005), University of Piraeus, pp. 55-75

ΜΕΡΙΚΑ ΠΑΡΑΤΗΡΗΣΙΜΕΣ ΑΛΥΣΙΔΕΣ ΜΑΡΚΟΒΙΑΝΗΣ ΑΠΟΦΑΣΗΣ ΜΕ ΟΜΟΙΟΜΟΡΦΗ ΚΑΤΑΝΟΜΗ ΜΗΝΥΜΑΤΩΝ

Υπό

Γκουλιώνη Ιωάννη
Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιώς

Abstract

PARTIALLY OBSERVABLE MARKOV DECISION PROCESSES WITH UNIFORMLY DISTRIBUTED SIGNAL PROCESSES

A Partially observed Markov decision process (P.O.M.D.P.) is a sequential decision problem where information concerning parameters of interest is incomplete, and possible actions include sampling, surveying, or otherwise collecting additional information. Such problems can theoretically be solved as dynamics programs, but the relevant state space is infinite which inhibits algorithmic solution. We formulate a (P.O.M.D.P.) with a continous signal space and a method to convert a problem with uniformly distributed signal processes. We discussed how to solve (P.O.M.D.P.) problems with continous signal processes. However, in order to obtain a value function which is close to the optimal value function, we might need to construct a step function with large number of signals.

Keywords: Maintenance, dynamic-programming, P.O.M.D.P.

1. Εισαγωγή

Μια P.O.M.D.P. (Μερικά παρατηρήσιμη Μαρκοβιανή αλυσίδα) είναι μια γενικευμένη αλυσίδα Μαρκοβιανής απόφασης, που επιτρέπει μιαν 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.

Επειδή λοιπόν σε ένα πρόβλημα P.O.M.D.P. δεν ξέρουμε ακριβώς την κατάσταση του συστήματος σε κάθε χρόνο, μέσω ενός μηχανισμού ελέγχου λαμβάνουμε ένα μήνυμα που συνδέεται στοχαστικά με την πραγματική

κατάσταση. Για παράδειγμα, όταν εξετάζουμε την κατάσταση ενός ασθενούς που πάσχει από στεφανιαία νόσο, το αποτέλεσμα του λεγόμενου τεστ κόπωσης, το επίπεδο ισχαιμίας, καθώς και ο πόνος στο στήθος είναι μηνύματα που συνυφαίνονται με την ζωτική κατάσταση του ασθενούς. Αρκετοί ερευνητές ασχολήθηκαν με το πρόβλημα εύρεσης βέλτιστων στρατηγικών σε πεπερασμένο και σε άπειρο χρονικόν ορίζοντα με διακριτό χώρο μηνυμάτων όπως οι Sondik [6,7,8], Denardo [2], Littman [5], Lovejoy [4], Howard [3], Bertsekas [1].

Στην εργασία αυτή επεκτείνουμε τα αποτελέσματα, στην περίπτωση όπου ο χώρος μηνυμάτων είναι συνεχής, με ομοιόμορφη κατανομή μηνυμάτων.

Η εργασία οργανώνεται ως εξής: Στην παράγραφο 2 περιγράφουμε το μοντέλο P.O.M.D.P. με διακριτό χώρο μηνυμάτων. Στην παράγραφο 3 δίνουμε τα ήδη γνωστά αποτελέσματα, που αφορούν την ύπαρξη βέλτιστης στάσιμης στρατηγικής για το πρόβλημα του άπειρου χρονικού ορίζοντα με διακριτό αριθμό μηνυμάτων. Οι ιδιότητες της συνάρτησης τιμών σε πεπερασμένο χρονικόν ορίζοντα είναι ουσιαστικές, διότι επιτρέπουν την επανάληψη στον συνεχή χώρο των καταστάσεων μιας P.O.M.D.P.

Στην παράγραφο 4 επεκτείνουμε τις τεχνικές που ισχύουν για το πρόβλημα P.O.M.D.P. με διακριτό χώρο μηνυμάτων. Κλειδί για την παραπάνω επέκταση είναι η διαμέριση του χώρου μηνυμάτων, με βάση ορισμένες υποθέσεις, σε τρόπον ώστε κάθε κελλί της διαμέρισης να μπορεί να θεωρηθεί σαν διακριτό μήνυμα.

Τέλος δίνουμε μια μέθοδο που μετατρέπει το πρόβλημα με συνεχή χώρο μηνυμάτων σε ένα πρόβλημα με διακριτό χώρο μηνυμάτων, καθώς και μια εφαρμογή για να γίνουν κατανοητές οι παρακάτω σκέψεις.

2. Περιγραφή του μοντέλου των μερικά παρατηρήσιμων Μαρκοβιανών αλυσίδων

Μια P.O.M.D.P. περιγράφει μια στοχαστική διαδικασία που τυπικά αντιστοιχεί στην εξάδα (S, A, Θ, P, R,q).

• Το $S = \{1,2,3...N\}$ είναι το σύνολο των δυνατών καταστάσεων του συστήματος.

• Το A είναι το σύνολο των αποφάσεων,που υποτίθεται ορισμένο. Η απόφαση που εκλέγεται στον χρόνο t ορίζεται σαν Y_t.

Στον χρόνο t, αν το σύστημα βρίσκεται στην κατάσταση i, και επιλέξουμε την απόφαση α, τότε είναι γνωστό ένα σύνολο από πιθανότητες $\{p^{\alpha}_{ij}(t)$ για j=1,2,...,N $\}$,όπου $(\sum_j P^a_{ij}(t)=1$ για κάθε κατάσταση του συστήματος i).

$$P^a_{ij}(t) = \Pr\left[x_{t+1} = j \middle| x_t = i, Y_t = a\right] \quad (1)$$

Υποθέτουμε ότι οι πιθανότητες μεταφοράς είναι ανεξάρτητες του χρόνου. Υποτίθεται ότι γνωρίζουμε τον πίνακα μεταφοράς P^a

• Ένα ορισμένο όφελος $q^a(i)$ προκύπτει, όταν το σύστημα βρίσκεται στην κατάσταση i και εκλέγεται η απόφαση α. Το q^a είναι ένα N-διάστατο διάνυσμα στήλη της μορφής $q^a=(q^a(1),q^a(2),\ldots, q^a(N))$.

Αν γνωρίζουμε την τρέχουσα κατάσταση του συστήματος, πριν εκλέξουμε μιαν απόφαση τότε το πρόβλημα είναι ένα πρόβλημα πλήρως παρατηρήσιμης M.D.P. Μπορούμε τότε να εκλέξουμε μιαν απόφαση, που βασίζεται στην τρέχουσα κατάσταση του συστήματος, μεγιστοποιώντας το ολικό προσδοκώμενο αποπληθωρισμένο όφελος.

Οι POMDPS έχουν το χαρακτηριστικό ότι η κατάσταση του συστήματος δεν μπορεί να παρατηρηθεί απευθείας. Αντί για την κατάσταση, μέσω ενός μηχανισμού ελέγχου αυτός/ή που αποφασίζει λαμβάνει ένα τυχαίο μήνυμα θ.

$$r^a_{i\theta} = \Pr[z_t = \theta \,|\, x_t = \iota,\ Y_{t-1} = a] \forall t \quad (2)$$

Μολονότι το μήνυμα είναι τυχαίο, γνωρίζουμε την παραπάνω υπό συνθήκη πιθανότητα να λάβουμε το μήνυμα θ, όταν η κατάσταση του συστήματος ,καθώς και η απόφαση είναι γνωστές. Υποθέτοντας ότι οι υπό συνθήκη πιθανότητες δεν μεταβάλλονται με τον χρόνο μπορούμε να θεωρήσουμε τον πίνακα $R^a = [r^a_{i\theta}]$

Υποθέτοντας ότι το αντικείμενο του μοντέλου απόφασης είναι να μεγιστοποιήσει το ολικό προσδοκώμενο εκπίπτον όφελος, αυτός/ ή που αποφασίζει, αφού δεν γνωρίζει την πραγματική κατάσταση του συστήματος παίρνει μια απόφαση βασιζόμενος σε όλη την ιστορία της εξέλιξης. Κριτήρια για την επιλογή βέλτιστων στρατηγικών σε πεπερασμένο και άπειρο χρονικόν ορίζοντα, είναι η μεγιστοποίηση του ολικού προσδοκώμενου εκπίπτοντος

οφέλους, που υλοποιείται μέοω των παρακάτω σχέσεων (3), (4) για πεπερασμένο και άπειρο χρονικόν ορίζοντα αντίστοιχα:

$$E_{\pi(0)}\{\sum_{t=0}^{t=T-1}\beta^t.q[s(t),\alpha(t)]+\beta^t q[s(T)]\} \quad (3)$$

$$E_{\pi(0)}\{\sum_{t=0}^{t=\infty}\beta^t.q[s(t),\alpha(t)]\} \quad (4)$$

3. Η βέλτιστη συνάρτηση τιμών

Η ιστορία της εξέλιξης στον χρόνο t, δηλώνεται σαν I_t, όπου

$I_0 = [\pi(0), Z_0]$

$I_t = \pi(0), Z_0, Y_0, Z_1, Y_1,...,Z_{t-1}, Y_{z-1}, Z_t]$ και

$I_{t+1} = [I_t, Y_t, Z_{t+1}]$

Σημειώνουμε ότι {It, t=0,1,2,...} είναι μια αλυσίδα Μαρκοβιανής απόφασης σύντομα, (M.D.P).

Αν τώρα, $\upsilon_t(.)$, δηλώνει το βέλτιστο ολικό προσδοκώμενο όφελος από τον χρόνο t μέχρι το τέλος του σχεδιασθέντος ορίζοντα αποφάσεων.

τότε το δυναμικό πρόγραμμα (επαναληπτική εξίσωση D.P.) μπορεί να γράφει σαν:

$$\upsilon_t(I_t) = \max_{Y_t \in A} E\left\{q^{Yt}(x_t)+\beta.\upsilon_{t+1}(I_{t+1})\middle|I_t, Y_t\right\}$$

$$= \max_{Y_t \in A} E\left\{E\left(q^{Yt}(x_t)\middle|I_t, Y_t\right)+\beta.E\left(\upsilon_{t+1}(I_{t+1})\middle|I_t, Y_t\right)\right\}$$

$$= \max_{Y_t \in A} E\left\{\sum_{i=1}^{N}\Pr(x_t = i|I_t).q^{Yt}(i)+\beta.\sum_{\theta\in\theta}\Pr\left(z_{t+1}=\theta|I_t, Y_t\right).\upsilon_{t+1}\left(\left[I_t, Y_t, Z_{t+1}=\theta\right]\right)\right\} \quad (5)$$

όπου β είναι ένας παράγοντας έκπτωσης (discount factor).

Υποθέτουμε ότι $0.<\beta<1$ για τα προβλήματα άπειρου χρονικού ορίζοντα, ενώ $0<\beta$ για τα προβλήματα του πεπερασμένου χρονικού ορίζοντα.

Για να λύσουμε την επαναληπτική εξίσωση (5), οι συναρτήσεις τιμών στον χρόνο t πρέπει να υπολογισθούν για κάθε πιθανή ιστορία I_t. Όταν υπάρχει ένας μεγάλος αριθμός από πιθανές αποφάσεις, ένας μεγάλος αριθμός μηνυμάτων, και ο χρονικός ορίζοντας είναι μεγάλος, τότε ο αριθμός των πιθανών ιστοριών I_t είναι μεγάλος (ή άπειρος) και αυξάνει γραμμικά με το t. Οι υπολογιστικές απαιτήσεις ενός τέτοιου DP (dynamic - programming) αλγόριθμου είναι αληθινά τεράστιες (Οι λεπτομέρειες αυτού του προβλήματος σχολιάζονται στον Bertseka (1976).

Επομένως, μπορούμε να θεωρήσουμε όλες τις πιθανές μεταβλητές του μοντέλου. Ας είναι $\pi_i(t)=\Pr(x_t=i|I_t)$ και $\pi(t) = [\pi_i(t), \pi_2(t),...,\pi_N(t),$

όπου $\sum_{i=1}^{N}\pi_i(t) = 1 \qquad 0 \le \pi i\ (t) \le 1$ για $i = 1, 2, ...,N.$

Είναι εύκολο να δείξουμε $\Pr\left(x_{t+1}=i\middle|I_t, Y_t, Z_{t+1}\right)=$
$\Pr\left(x_{t+1}=i\middle|\pi(t), Y_t, Z_{t+1}\right)$

και $\Pr\left(z_{t+1}=\theta\middle|I_t, Y_t\right)=\Pr\left(z_{t+1}=\theta\middle|\pi_{(t)}, Y_t\right)$

Επομένως, αν δοθούν $I_t, Y_t=\alpha, z_{t+1}=\theta$ από τον κανόνα Bayes'

$$\pi_j(t+1)=\Pr(x_{t+1}=j|I_t, Y_t=\alpha, z_{t+1}=\theta)=\Pr(x_{t+1}=j|\pi(t), Y_t=\alpha, z_{t+1}=\theta)$$

$$=\frac{\Pr(x_{t+1}=j, z_{t+1}=\theta|\pi(t), Y_t=\alpha)}{\Pr(z_{t+1}=\theta|\pi(t), Y_t=\alpha)}=\frac{\sum_{i=1}^{N}\pi_i(t).P^a_{i,j}.r^a_{j,\theta}}{\sum_{k=1}^{N}\sum_{i=1}^{N}\pi_i(t).P^a_{i,k}.r^a_{x,\theta}} \qquad (6)$$

ή σε μορφή πινάκων

$$T(\pi(t),\theta,\alpha\equiv\Pr(x_{t+1}|I_t, Y_t=a, z_{t+1}=\theta)=\Pr(x_{t+1}|\pi(t), Y_t=a, z_{t+1}=\theta)=$$
$$=\frac{\pi(t).P^a.R^a_\theta}{\pi(t).P^a.R^a_\theta.1}$$

όπου R^{α}_{θ} είναι ένας διαγώνιος πίνακας που έχει $r^{a}_{j\theta}$ σαν διαγώνια στοιχεία, και 1 είναι ένας πίνακας στήλη με όλα τα στοιχεία του 1.

Είναι καλά γνωστό, ότι π(t) είναι μια επαρκής στατιστική για την ιστορία της εξέλιξης I_t, όταν εκλέγουμε μια απόφαση σε χρόνο t. Συγκεκριμένα π(t) περιέχει όλη την αναγκαία πληροφορία, όσον αφορά την ιστορία της εξέλιξης, προκειμένου να εκλεγεί μια απόφαση στον χρόνο t *(Bertsekas 1976, Monahan 1982, Sondik 1971, Striebel 1965)*.

Ας είναι $\Pi \equiv \{\pi \in R^N : \sum_{i=1}^{N} \pi_i = 1$ και $\pi_i \geq 0$ για i=1, 2,..., N$\}$. Το ακόλουθο αποτέλεσμα οφείλεται στον (Sondik 1971): Για κάθε ακολουθία αποφάσεων $Y_1, Y_2, ..., Y_t \in A$, η ακολουθία των πιθανοτήτων $\{\pi(\kappa)\}$ όπου $\kappa = 0, 1, ..., t$ είναι μια Μαρκοβιανή αλυσίδα, που σημαίνει ότι, αν $\Gamma \subset \Pi$, τότε:

$$\Pr(\pi(t+1) \in \Gamma \mid \pi(0),\ \pi(1),\ \ldots\ \pi(t) Y_t) = \Pr(\pi(t+1) \in \Gamma \mid \pi(t),\ Y_t)$$

Επομένως, $\forall \pi \in \Pi$,

$$\upsilon_t(\pi) = \max_{a \in A} E\{q^a(x_t) + \beta . \upsilon_{t+1}(T(\pi, \alpha, \theta)) | \pi, \alpha\}$$

$$= \max_{\alpha \in A} \left\{ \sum_{i=1}^{N} \pi_i . q^a(i) + \beta . \sum_{\theta \in \Theta} \Pr(z_{t+1} = \theta | \pi, \alpha) . \upsilon_{t+1}(T(\pi, \alpha, \theta)) \right\} \qquad (7)$$

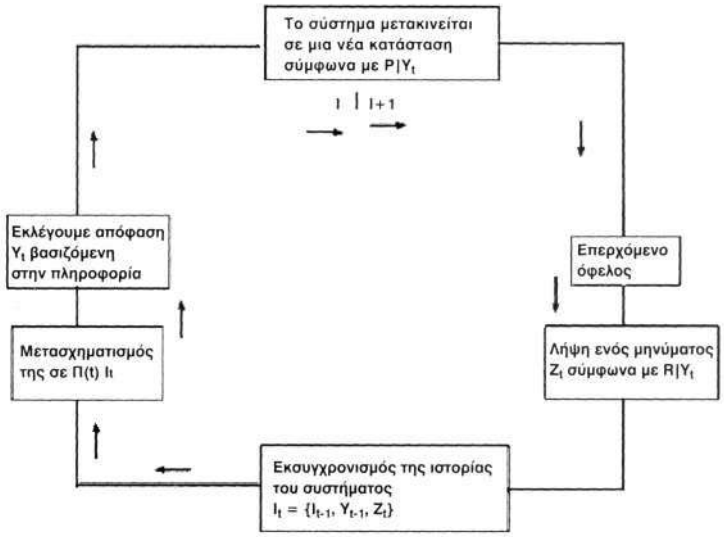
Σημειώνουμε ότι η εξίσωση (7) είναι μια επαναληπτική εξίσωση, που θεωρητικά τουλάχιστον μπορεί να βρει μια βέλτιστη στρατηγική, με βάση το κριτήριο του μέσου αποπληθωρισμένου οφέλους. Ωστόσο επειδή ο χώρος Π είναι συνεχής, δεν επιτρέπονται οι επαναλήψεις, οπότε το να βρούμε μια βέλτιστη στρατηγική δεν είναι εύκολο έργο. Ευτυχώς όμως οι POMDPS έχουν κάποιες θαυμάσιες ιδιότητες, που μας βοηθούν να αναπτύξουμε υπολογιστικούς αλγόριθμους για την εύρεση μιας βέλτιστης στρατηγικής. Αυτές τις ιδιότητες παραθέτουμε στην επόμενη παράγραφο.

4. Συμβολισμοί και τελεστές

Στην παράγραφο αυτή, κάποιοι συμβολισμοί και τελεστές εισάγονται. Εισάγουμε, την νόρμα $\|.\|$ του Supremum. Έστω B(π) το σύνολο των φραγμένων συναρτήσεων (πραγματικών τιμών) στο Π.

ΣΧΗΜΑ 1

Διάγραμμα απόφασης για τις μερικά παρατηρήσιμες Μαρκοβιανές αλυσίδες



Τότε $\|\upsilon\| = \sup\{\upsilon(\pi) | : \pi \in \Pi\}$

Έστω τώρα συνάρτηση πραγματικών τιμών h: Π Χ Α Χ Β (Π) $\rightarrow R$

$$h(\pi, \alpha, \upsilon) = \pi.q^{\alpha} + \beta \cdot \sum_{\theta \in \Theta} \Pr(\theta | \pi, \alpha) . \upsilon(T(\pi, \alpha, \theta)) \qquad (8)$$

Τότε μέσω της h η εξίσωση (7) μπορεί να ξαναγραφεί σαν:

$$\upsilon_t(\pi) = \max_{\alpha \in A} h(\pi, \alpha, \upsilon_{t+1}).$$

Αν, δ, είναι τώρα μια στρατηγική, δ:Π→Α, και Δ το σύνολο όλων των στάσιμων στρατηγικών, μπορούμε να γενικεύσουμε μέσω ενός τελεστού H_{δ} στο Β(Π), $\forall \delta \in \Delta$.

$$[H_{\delta}(\upsilon)](\pi) = h(\pi, \delta(\pi), \upsilon) .$$

Αν $\delta(\pi) = \alpha$ για $\forall \pi \in \Pi$, τότε χρησιμοποιούμε την έκφραση H_{α} αντί για την H_{δ}.

Τελικά ορίζεται ένας τελεστής βελτιστοποίησης σαν: $Hυ = \max_{δ∈Δ}[H_δ υ]$

Ιδιότητες τελεστών Hδ, H

α) Είναι φραγμένοι *(Whitt 1978)*

β) Έχουν την ιδιότητα της μονοτονίας

γ) Έχουν την ιδιότητα της συστολής για το πρόβλημα του άπειρου χρονικού ορίζοντα.

Στις ιδιότητες αυτές των παραπάνω τελεστών θα πρέπει να προσθέσουμε, ότι η βέλτιστη συνάρτηση τιμών, σε πεπερασμένο χρονικόν ορίζοντα Τ, είναι κατά τμήματα γραμμική και μάλιστα κυρτή *(Sondik 1971)*.

Μια συνάρτηση υ καλείται κατά τμήματα γραμμική και κυρτή αν υπάρχει ένα ορισμένο σύνολο από Ν-διάστατα "gradients", Γ, ώστε:

$$υ(π) = \max_{i=1,2,3...x} \{π.γ^i : \text{όπου } γ^i ∈ Γ\}$$

Για $π∈Π$, $α∈A$ και $θ∈Θ$, ορίζουμε τό σύνολο:

$$Γ_{π,α,θ} = \{\bar{γ} ∈ Γ : T(π(α,θ)\cdot\bar{γ} ≥ T(π,α,θ)\cdot γ \ ∀γ ∈ Γ\}$$

Τότε $υ(T(π,α,θ)) = T(π,α,θ)\cdot\bar{γ} \quad ∀γ ∈ Γ_{π,α,θ}$.

Αν $γ_{π,α,θ}$ είναι ένα "gradient" στο σύνολο $Γ_{π,α,θ}$, τότε, Hυ(π) μπορεί να γραφεί σαν:

$$H_α υ(π) = π\cdot q^α + β\sum_{θ∈Θ} Pr(θ|π,α)\cdot υ(T(π,α,θ))$$

$$= π\, q^α + β\sum_{θ∈Θ}\left(π\cdot P^α\cdot R_θ^α\cdot 1\right)\frac{π\cdot P^α R_θ^α}{π\cdot P^α\cdot R_θ^α\cdot 1}\,γ_{π,α,θ} = π\cdot\left[q^α + β\cdot\sum_{θ∈Θ} P^α\cdot R_θ^α\cdot γ_{π,α,θ}\right] \quad (9)$$

Επειδή $q^α + β\sum_{θ∈Θ} P^α\cdot R_θ^α\cdot γ_{π,α,θ}$ είναι ένα Ν-διάστασης "gradient" η (9) μπορεί να απλοποιηθεί σαν: $H_α υ(π) = π.γ_{π,α}$

όπου $γ_{π,α} = q^α + β\cdot P^α\sum_{θ∈Θ} R_θ^α\cdot γ_{π,α,θ}$

Επιπλέον, $H\upsilon(\pi) = \max_{\alpha \in A} H_\alpha \upsilon(\pi)$, τότε:

$$H\upsilon(\pi) = \max_{\alpha \in A}\left[H_\alpha \upsilon(\pi)\right] = \max_{\alpha \in A}\left\{\pi \cdot \gamma_{\pi,\alpha}\right\} = \pi \cdot \gamma_\pi$$

όπου $\gamma_\pi = q^\alpha + \beta \sum_{\theta \in \Theta} P^{\alpha^*} R_\theta^{\alpha^*} \cdot \gamma_{\pi,\alpha^*,\theta}$ (10)

Όπου, α*, είναι μια βέλτιστη απόφαση για το π.

Επειδή μάλιστα οι τελεστές H_δ, H είναι συστολές, υπάρχουν μοναδικά "fixed-point" στο B(π), ας τα πούμε υ_δ, και υ* αντίστοιχα ώστε:

$H_\delta \ \upsilon_\delta = \upsilon_\delta$

και

$H \ \upsilon^* = \upsilon^*$ (*Elgolic* 1964)

Μια στρατηγική $\delta \in \Delta$ για την οποία ισχύει $H_\delta \upsilon^* = H\upsilon^*$ είναι μια βέλτιστη στρατηγική. Οι Sawaragi και Yoshikawa (1970) έδειξαν, ότι αν ο χώρος A των αποφάσεων είναι ορισμένος, τότε υπάρχει μια βέλτιστη στάσιμη στρατηγική. Αυτό εκφράζει ότι, αν έχουμε ένα αρχικό (i.v) (information-vector), π, και η στρατηγική $\delta \in \Delta$ (στάσιμη) χρησιμοποιείται για το πρόβλημα του άπειρου χρονικού ορίζοντα, τότε το ολικό προσδοκώμενο εκπίπτον όφελος θα είναι $\upsilon_\delta(\pi)$.

Από το ορισμό τώρα του τελεστή H,

$\upsilon^*(\pi) = \max_{\delta \in \Delta}\{\upsilon_\delta(\pi)\}$ για όλα τα $\pi\Pi$.

Άρα, υ* είναι η βέλτιστη συνάρτηση τιμών.

Επειδή μια P.O.M.D.P. έχει έναν συνεχή χώρο καταστάσεων, είναι πολύ δύσκολο να βρούμε την υ_δ για $\delta \varepsilon \Delta$ ή το υ*. Επομένως, φράγματα για τις τιμές των παραπάνω συναρτήσεων αναζητούνται σε αρκετούς ερευνητές.

Συνοπτικά μπορούμε να πούμε, ότι οι αλγόριθμοι που αφορούν το πρόβλημα του πεπερασμένου χρονικού ορίζοντα, και βασίζονται σε μια επαναληπτική διαδικασία "value-iteration"χρησιμοποιούνται επίσης στο βήμα "policy-improvement" για το πρόβλημα του άπειρου χρονικού ορίζοντα.

Επομένως ουσιαστικό βήμα στην κατασκευή αλγορίθμων για το πρόβλημα του άπειρου χρονικού ορίζοντα είναι ο υπολογισμός Hu για μια δεδομένη κατά τμήματα γραμμική και κυρτή συνάρτηση υ.

5. Μερικά παρατηρήσιμες Μαρκοβιανές αλυσίδες με συνεχή χώρο μηνυμάτων

Το σχέδιο αυτής της παραγράφου έχει ως ακολούθως: υποθέσεις, συμβολισμοί και διατύπωση μιας P.O.M.D.P. με συνεχή χώρο μηνυμάτων, μια μέθοδο που να μετατρέπει ένα πρόβλημα P.O.M.D.P. με συνεχή χώρο μηνυμάτων σε μια P.O.M.D.P. με ορισμένο αριθμό μηνυμάτων.

Οι βασικές υποθέσεις είναι ίδιες με αυτές που αναπτύχθηκαν στην παράγραφο (3). Η μόνη διαφορά είναι στην φύση των μηνυμάτων. Στην παράγραφο αυτή υποθέτουμε, ότι οι κατανομές μηνυμάτων έχουν συναρτήσεις πυκνότητας, ενώ στην παράγραφο (3) υποθέσαμε ότι είναι διακριτές. Οι παράμετροι αυτών των συναρτήσεων πυκνότητας εξαρτώνται από την κατάσταση του συστήματος, καθώς επίσης και από την απόφαση που λαμβάνεται στον συγκεκριμένο χρόνο. Για να γίνουμε πιο κατανοητοί, για κάθε κατάσταση του συστήματος $i \varepsilon S$, κάθε απόφαση $\alpha \varepsilon A$, και κάθε χρόνο $t=0$, 1, 2,... υπάρχει μια συνάρτηση πυκνότητας πιθανότητας $ff^{a}_{i,t}(.)$ στο σύνολο των μηνυμάτων.

Για το πρόβλημα του άπειρου χρονικού ορίζοντα, υποθέτουμε ότι η συνάρτηση πυκνότητας πιθανότητας είναι ανεξάρτητη από τον χρόνο και επομένως η εξάρτηση από τον χρόνο παραλείπεται στον συμβολισμό.

Ας είναι $\pi \varepsilon \Pi$, όπως ακριβώς ορίσθηκε στην παράγραφο (3). Δεδομένου ότι το τρέχον (i.v) είναι π και η απόφαση α χρησιμοποιείται, η υπό συνθήκη πυκνότητα πιθανότητας είναι $\overline{f}(\theta|\pi,\alpha)=\sum_{K=1}^{N}\sum_{j=1}^{N}\pi_j \cdot P^{\alpha}_{j,K} \cdot f^{\alpha}_{K}(\theta)$ ή σε μορφή πινάκων

$$\overline{f}(\theta|\pi,\alpha)=\pi.P^{\alpha}\,\overline{R}^{\alpha}_{\theta}\cdot 1 \qquad (11)$$

όπου $\overline{R}^{a}_{\theta}$ είναι ένας διαγώνιος πίνακας που έχει διαγώνια στοιχεία τα $f^{a}_{i}(\theta)$ και **1** είναι ένα N-διάστασης διάνυσμα στήλη με όλα τα στοιχεία αποτελούμενα από 1.

Ανάλογα με τον ορισμό του $T(\pi, \alpha, \theta)$, ορίζουμε $\overline{T}(\pi, \alpha, \theta)$ την κατανομή πιθανότητας, που αφορά την κατάσταση του συστήματος, στην επόμενη χρονική περίοδο, δεδομένου ότι η κατανομή πιθανότητας για την τρέχουσα κατάσταση του συστήματος είναι π, λαμβάνεται η απόφαση α,και παρατηρείται το μήνυμα θ στην αρχή της επόμενης περιόδου.

$$\overline{T}_i(\pi, \alpha, \theta) = \Pr\left(x_{t+1} = i \middle| \pi, Y_t = \alpha, Z_{t+1} = \theta\right),$$

και

$$\overline{T}_i(\pi, \alpha, \theta) = \Pr\left(x_{t+1} = i \middle| \pi, Y_{t+1} = \alpha, Z_{t+1} = \theta\right) = \frac{\sum_{j=1}^{N} \pi_j \cdot P_{j,i}^{\alpha} \cdot f_i^{\alpha}(\theta)}{\sum_{\kappa=1}^{N} \sum_{j=1}^{N} \pi_j \cdot P_{j,\kappa}^{\alpha} \cdot f_{\kappa}^{\alpha}(\theta)} \qquad (12)$$

ή, in matrix form, $$\overline{T}(\pi, \alpha, \theta) = \frac{\pi \cdot P^{\alpha} \cdot \overline{R}_{\theta}^{\alpha}}{\bar{f}(\theta | \pi, \alpha)} = \frac{\pi \cdot P^{\alpha} \cdot \overline{R}_{\theta}^{\alpha}}{\pi \cdot P^{\alpha} \cdot \overline{R}_{\theta}^{\alpha} \cdot 1} \qquad (13)$$

Αναλογικά με την παράγραφο (3), ας είναι υ_t και υ_{t+1} κυρτές, φραγμένες και συνεχείς συναρτήσεις τιμών στις χρονικές περιόδους t και t+1, αντίστοιχα. Τότε, για $\pi \in \Pi$,

$$\upsilon_t(\pi) = \max_{a \in A} \; E\{q^a(x_t) + \beta . \upsilon_{t+1}(\overline{T}(\pi, \alpha, \theta)) | \pi, \alpha\}.$$

$$= \max_{a \in A} \left\{ \sum_{i=1}^{N} \pi_i . q^a(i) + \beta . \int_{\theta \in \Theta} \bar{f}(\theta | \pi, \alpha) \, . \, \upsilon_{t+1}(\overline{T}(\pi, \alpha, \theta) d\theta \right\} \qquad (14)$$

Αν υ_{t+1} είναι κατά τμήματα γραμμική συνάρτηση, με πεδίο ορισμού το Π, τότε ο τύπος (14) μπορεί να ξαναγραφεί σαν:

$$\upsilon_t(\pi) = \max_{\alpha \in A} \left\{ \sum_{i=1}^{N} \pi_i \cdot q^{\alpha}(i) + \beta \cdot \int_{\theta \in \Theta} \pi \cdot P^{\alpha} \cdot \overline{R}_{\theta}^{\alpha} \cdot \gamma^{l(\pi, \alpha, \theta)} \cdot d\theta \right\} \qquad (15)$$

όπου $\gamma^{l(\pi,\alpha,\theta)}$ είναι το "gradient" του υ_{t+1}, ώστε:

$$T(\pi, \alpha, \theta) . \gamma^{l(\pi,\alpha,\theta)} \geq \overline{T}(\pi, \alpha, \theta) . \gamma^{x}$$

για όλα τα γ^{x} που υποστηρίζουν την υ_{t+1}.

Σημειώνουμε ότι ο τύπος (15) είναι παρόμοιος με τον αντίστοιχο τύπο που αφορά την περίπτωση που έχουμε διακριτό αριθμό μηνυμάτων. Η μόνη διαφορά είναι ότι το άθροισμα αντικαθίσταται με ένα ολοκλήρωμα.

Οι White and Harrington (1980) έδειξαν, ότι αν υ_t+1 είναι μια κυρτή συνάρτηση, το ίδιο ισχύει και για την υ_t. Εν τούτοις, σε αντιδιαστολή με την περίπτωση που έχουμε ορισμένο αριθμό μηνυμάτων, υ_t δεν είναι κατά τμήματα γραμμική ακόμη και αν υ_t+1 είναι κατά τμήματα γραμμική. Αυτή η ιδιότητα καθιστά δυσκολότερο το πρόβλημα όπου έχουμε συνεχή κατανομή μηνυμάτων, σε σύγκριση με εκείνο, όπου έχουμε διακριτή κατανομή μηνυμάτων.

Μια ομοιόμορφη κατανομή χρησιμοποιείται σε μοντέλα όπου δεν έχουμε αρκετή πληροφόρηση. Σχολιάζουμε αυτή την κατανομή, διότι μας δίνει την δυνατότητα να αντιμετωπίσουμε το πρόβλημα όπου έχουμε συνεχή χώρο μηνυμάτων, εφαρμόζοντας τις μεθόδους που αφορούν τον διακριτό χώρο μηνυμάτων.

Ας είναι Θ ο χώρος μηνυμάτων. Επίσης στο χρόνο t,ας είναι Θ(t,i,α) ο χώρος μηνυμάτων για την εξέλιξη, δεδομένου ότι η κατάσταση του συστήματος είναι η i και η απόφαση που λαμβάνεται στο χρόνο t είναι η α. Η συνάρτηση πυκνότητας πιθανότητας για το μήνυμα στον χώρο μηνυμάτων υποθέτουμε ότι έχει την ομοιόμορφη κατανομή.

Μια τετριμμένη περίπτωση έχουμε, όταν $\Theta(t, i, \alpha)\ \ \Theta(t,\ \ i, \alpha) = 0$ για όλα τα ζεύγη των καταστάσεων i και j. Τότε μπορούμε ξεκάθαρα να διατυπώσουμε το πρόβλημα, στα πλαίσια μιας πλήρως παρατηρήσιμης M D P. Για παράδειγμα, υποθέτουμε ότι έχουμε ένα πρόβλημα με δύο καταστάσεις, σε τρόπο ώστε μια μηχανή να βρίσκεται στην καλή κατάσταση όταν έχει λειτουργικό κόστος από 100-250 ευρώ την ημέρα, ενώ βρίσκεται στην κακή κατάσταση, όταν το λειτουργικό κόστος είναι από 300-450 ευρώ την ημέρα. Έτσι αν το τρέχον λειτουργικό κόστος είναι 350 ευρώ την ημέρα, τότε είναι προφανές ότι το σύστημα βρίσκεται στην κακή κατάσταση.

Βέβαια, η παραπάνω τεχνική αποτυγχάνει, όταν έχουμε αλληλοεπικάλυψη στις κατανομές μηνυμάτων. Εν τούτοις, αν οι κατανομές μηνυμάτων έχουν την ομοιόμορφη κατανομή, τότε το πρόβλημα μπορεί να αναδιατυπωθεί στα πλαίσια μιας POMDP με ορισμένο αριθμό μηνυμάτων.

Ας είναι $\Theta'(t,\mathrm{i},\alpha)=\Theta-\Theta(\mathrm{t},\mathrm{i},\alpha)$ και $\overline{\Theta}(\mathrm{t},\mathrm{i},\alpha)=\{\Theta(\mathrm{t},\mathrm{i},\alpha),\Theta'(t,\mathrm{i},\alpha)\}$

Τότε $\overline{\Theta}(t, \mathrm{i}, \alpha)$ είναι μια διαμέριση του χώρου των μηνυμάτων Θ. Ας είναι τώρα

$$\Theta_t = \left\{ \hat{\Theta}_{(t,1)}, \hat{\Theta}_{(t,2)} \ldots \hat{\Theta}_{(t,k)} \right\}$$

το γινόμενο των διαμερίσεων των $\overline{\Theta}(t, \mathrm{i}, \alpha)$ για όλες τις καταστάσεις του συστήματος i και όλες τις αποφάσεις α.

Με την παραπάνω μέθοδο χωρίζουμε τον παραμετρικό χώρο Θ που είναι συνεχής σε κ το πλήθος κελιά. Κάθε κελί μπορεί να θεωρηθεί σαν ένα διακριτό μήνυμα. Τα μηνύματα είναι όμως ομοιόμορφα κατανεμημένα στον παραμετρικό χώρο Θ, οπότε οι πιθανότητες για καθένα από τα μηνύματα που ορίσθηκαν με τον νέο αυτό τρόπο, μπορούν να υπολογισθούν σαν το ολοκλήρωμα της συνάρτησης πυκνότητας σε κάθε περιοχή του Θt.

Επειδή υπάρχει μόνον ορισμένος αριθμός N, από καταστάσεις, το σύνολο Θt έχει ορισμένο αριθμό από στοιχεία. Το κλειδί για την μετατροπή ενός προβλήματος P.O.M.D.P. με συνεχή χώρο μηνυμάτων, που όμως οι κατανομές μηνυμάτων (Signal distributions) είναι ομοιόμορφα καταναμεμημένες ,σε ένα πρόβλημα P.O.M.D.P. με διακριτό αριθμό μηνυμάτων, είναι το γεγονός, ότι η μόνη πληροφορία που παρέχεται από το μήνυμα είναι το κελί της διαμέρισης στην οποίαν αυτό ανήκει.

Κάθε στοιχείο του Θ_t, μπορεί να θεωρηθεί σαν ένα μήνυμα, οπότε πλέον έχουμε ένα πρόβλημα P.O.M.D.P. με διακριτό αριθμό μηνυμάτων. Το κλειδί για την μετατροπή μιας P.O.M.D.P. με ομοιόμορφη κατανομή μηνυμάτων, σε ένα πρόβλημα P.O.M.D.P. με ορισμένο διακριτό αριθμό μηνυμάτων είναι ότι το T (π, α, .) είναι σταθερό πάνω σε κάθε στοιχείο του Θ_t. Ας δείξουμε αυτό το αποτέλεσμα στα λήμματα που ακολουθούν.

Για να γίνουν κατανοητά τα παραπάνω, θεωρούμε μια μηχανή με λειτουργικό κόστος που εξαρτάται από την κατάστασή της. Αν η μηχανή είναι σε άριστη κατάσταση, το λειτουργικό κόστος είναι ομοιόμορφα κατανε-μημένο από 100-250 ευρώ ανά ημέρα. Αν τώρα η μηχανή είναι σε κατάσταση επισκευής, το λειτουργικό κόστος είναι ομοιόμορφα κατανεμημένο από 200-350 ευρώ ανά ημέρα. Τέλος αν η μηχανή είναι σε άθλια κατάσταση, το λειτουργικό κόστος είναι ομοιόμορφα κατανεμημένο από 300-450 ευρώ ανά ημέρα. Άρα το κόστος μπορεί να διαιρεθεί σε πέντε περιοχές.

100-200
200-250

250-300
300-350
350-450

Κάθε περιοχή μπορούμε να την δοῦμε σαν διακριτό (ξεχωριστό μήνυμα). Από το γεγονός ότι τα μηνύματα είναι ομοιόμορφα κατανεμημένα στον χώρο των μηνυμάτων, οι πιθανότητες του καθενός από τα νέα μηνύματα όπως τα ορίσαμε, μπορούν να ορισθούν με ολοκλήρωμα της συνάρτησης πυκνότητας πιθανότητας σε κάθε μία από τις περιοχές του Θ_t. Για παράδειγμα, υπάρχει 66,67% πιθανότητα το λειτουργικό κόστος να είναι μεταξύ 100-200 ευρώ και 33,33% το λειτουργικό κόστος να είναι μεταξύ 200-250 ευρώ, αν η μηχανή βρίσκεται σε εξαιρετική κατάσταση. Όμοια η πιθανότητα για κάθε διάστημα κόστους μπορεί να υπολογισθεί για την κατάσταση επισκευής, καθώς και για την χειρότερη κατάσταση.

Το κλειδί όμως για την μετατροπή μιας P.O.M.D.P. με ομοιόμορφη κατανομή μηνυμάτων, σε μία P.O.M.D.P. με ορισμένο διακριτό αριθμό μηνυμάτων βασίζεται στο αποτέλεσμα ότι η $\bar{T}$ (π, α, .) είναι σταθερή πάνω σε κάθε στοιχείο του Θ_t. Προτείνουμε λοιπόν το ακόλουθο λήμμα και θεώρημα που αποδεικνύουν το αποτέλεσμα.

Λήμμα 3.1: Για κάθε κατάσταση του συστήματος $i \in S, f_i^a$ είναι σταθερή για κάθε στοιχείο του Θ_t.

Απόδειξη: Ας είναι θ_1 και θ_2 δύο αυθαίρετα μηνύματα σε ένα οποιοδήποτε $\hat{\Theta}(t, j) \in \Theta_t$. Σύμφωνα με την μέθοδο που αφορά την γενίκευση των στοιχείων του συνόλου Θ_t θα ισχύουν: $\hat{\Theta}(t, j) \cap \Theta(t, i, \alpha) = \varnothing$ είτε $\hat{\Theta}(t, j) \subseteq \Theta(t, i, \alpha)$. Αυτό μάλιστα θα ισχύει για όλες τις καταστάσεις i.

Αν $\hat{\Theta}(t, j) \cap \Theta(t, i, \alpha) = \varnothing$ τότε $f_i^{\alpha}(\theta_1) = f_i^{\alpha}(\theta_2) = 0$. Αν $\hat{\Theta}(t, j) \subseteq \hat{\Theta}(t, i, \alpha)$ τότε, από την υπόθεση της ομοιόμορφης κατανομής μηνυμάτων $f_i^{\alpha}(\theta_1) = f_i^{\alpha}(\theta_2)$.

Το λήμμα είναι ουσιαστικό, διότι κάνει εμφανή την μέθοδο με την οποία μία P.O.M.D.P. με ομοιόμορφη κατανομή μηνυμάτων, αναδιαμορφώνεται σε

μία P.O.M.D.P. με ορισμένο διακριτό αριθμό μηνυμάτων. Επιπλέον, στην απόδειξη του λήμματος απαιτούμε ότι $\hat{\Theta}(t,j) \subseteq \Theta(t,i,\alpha)$ ή $\hat{\Theta}(t,j) \cap \Theta(t,i,\alpha) = \varnothing$ για όλες τις καταστάσεις του συστήματος i, και επίσης ότι οι στοχαστικές διαδικασίες μηνυμάτων είναι ομοιόμορφα κατανεμημένες στο $\hat{\Theta}(t,j) \in \Theta_t$. Στο σημείο αυτό επισημαίνουμε, ότι μπορούμε να επεκτείνουμε αυτό το αποτέλεσμα σε περισσότερο γενικές περιπτώσεις, όπου οι στοχαστικές διαδικασίες (signal-processes) είναι step-functions.

Από το δεδομένο ότι $f_i^a(.)$ είναι σταθερή για κάθε στοιχείο του Θ_t, $\overline{R}_\theta^a$ είναι το ίδιο για όλα τα θ που ανήκουν στο ίδιο στοιχείο (κελί) της διαμέρισης του Θ_t. Παράλληλα, $\overline{f}(\theta|\pi,\alpha) = \pi \cdot P^\alpha \cdot R_\theta^\alpha \cdot 1$, επομένως, $\overline{f}(\cdot|\pi,\alpha)$ είναι σταθερό σε κάθε στοιχείο του Θ_t. Τώρα θα αποδείξουμε ότι είναι σταθερό σε κάθε στοιχείο του Θ_t.

Θεώρημα 3.2: $\overline{T}(\pi,\alpha,.)$ είναι σταθερό σε κάθε στοιχείο του Θ_t.

Απόδειξη: Ας είναι $\hat{\Theta}(t,j) \in \Theta_t$, ένα αυθαίρετο στοιχείο του Θ_t και θ_1, θ_2 δύο αυθαίρετα μηνύματα στο $\hat{\Theta}(t,j)$. Αφού $\overline{R}_{\theta_1}^\alpha = \overline{R}_{\theta_2}^\alpha$ και $\overline{f}(\theta_1|\pi,\alpha) = \overline{f}(\theta_2,\pi,\alpha)$ από την (11). Οπότε πλέον θά έχουμε ότι $\overline{T}(\pi,\alpha,\theta_1) = \overline{T}(\pi,\alpha,\theta_2)$ λόγω της (13). Το αποτέλεσμα λοιπόν αποδείχθηκε.

Τώρα ας δείξουμε, ότι αν οι στοχαστικές διαδικασίες μηνυμάτων είναι ομοιόμορφα κατανεμημένες, τότε η εξίσωσή (14) μπορεί να ξαναγραφεί μορφολογικά εντελώς όμοια με την (7).

Ας θεωρήσουμε από τη σχέση (15) πρώτα το τμήμα του ολοκληρώματος. Ας είναι θ_i ένα αυθαίρετο μήνυμα $(\theta_i \in \hat{\Theta}(t,i))$

Από το θεώρημα $\int_{\theta\in\hat{\Theta}(t,i)} \bar{f}(\theta|\pi,\alpha).\upsilon_{t+1}(\bar{T}(\pi,\alpha,\theta)\,d\theta$

$$= \upsilon_{t+1}\left(\bar{T}(\pi,\alpha,\theta)\right).\int_{\theta\in\hat{\Theta}(t,i)} \bar{f}(\theta|\pi,\alpha).\,d\theta = \upsilon_{t+1}\left(\bar{T}(\pi,\alpha,\theta_i)\right).\Pr\left(\theta\in\hat{\Theta}(t,i)|\pi,\alpha\right)$$

$$= \Pr\left(\theta\in\hat{\Theta}(t,i)|\pi,\alpha\right).\upsilon_{t+1}\left(\bar{T}(\pi,\alpha,\theta_i)\right)$$

Από το θεώρημα (3.2), κάθε στοιχείο του Θ_t μπορούμε να το δούμε σαν ένα ξεχωριστό μήνυμα. Ορίζοντας τώρα $T\left(\pi,\alpha\hat{\Theta}(t,i)\right)=\bar{T}(\pi,\alpha,\theta_i)$ για $\left(\theta_i\in\hat{\Theta}(t,i)\right)$ και λαμβάνοντας υπόψη ότι έχουμε μονάχα ορισμένον αριθμό από στοιχεία του Θt, η εξίσωση (15) μπορεί να ξαναγραφεί σαν:

$$\upsilon_t(\pi)=\max_{\alpha\in A}\left\{\sum_{i=1}^{N}\pi_i\cdot q^{\alpha}(i)+\beta\cdot\int_{\theta\in\Theta_t}\bar{f}\left(\theta\,|\,\pi,\alpha\right)\upsilon_{t+1}\left(\bar{T}(\pi,\alpha,\theta)\right)\,d\theta\right\}$$

$$=\max_{\alpha\in A}\left\{\sum_{i=1}^{N}\pi_i.\,q^{\alpha}(i)+\beta.\sum_{\Theta(t,i)\in\Theta_t}\int_{\theta\in\Theta_{(t,i)}}\bar{f}(\theta|\pi,\alpha).\upsilon_{t+1}(\bar{T}(\pi,\alpha,\theta)d\theta\right\}$$

$$=\max_{\alpha\in A}\left\{\sum_{i=1}^{N}\pi_i.\,q^{\alpha}(i)+\beta.\sum_{\hat{\Theta}(t,i)\in\Theta_t}\Pr\left(\hat{\Theta}(t,1)|\pi,\alpha\right).\upsilon_{t+1}\left(T(\pi,\alpha,\Theta(t,i))\right)\right\}$$

όπου

$$\Pr\left(\hat{\Theta}(t,1)|\pi,\alpha\right)=\int_{\theta\in\hat{\Theta}(t,1)}\bar{f}(\theta|\pi,\alpha)d\theta$$

Η τελευταία εξίσωση είναι ίδια μορφολογικά με την εξίσωση (7). Επομένως η (15) μπορεί να ξαναγραφεί όπως ακριβώς η (7), **που πρακτικά σημαίνει ότι μία P.O.M.D.P. με ομοιόμορφα καταναμεμημένες τις στοχαστικές διαδικασίες μηνυμάτων, μπορεί να αναδιατυπωθεί σαν μία P.O.M.D.P. με ορισμένο διακριτό αριθμό μηνυμάτων.**

6. Μέθοδοι για την λύση των P.O.M.D.P.S. με συνεχή χώρο μηνυμάτων

Στην παράγραφο αυτή θα υπολογίσουμε την τιμή και τα gradients του Hυ(π) για αυθαίρετο π∈Π, αν υ είναι κατά τμήματα γραμμική συνάρτηση (υπολογισμός μέσω της (15)). Ας είναι Γ ένα σύνολο από "gradients" του υ. Ας είναι τώρα $\bar{\gamma}$ ένα gradient, ($\bar{\gamma} \in \Gamma$), και α μία απόφαση ($\alpha \in A$), τότε ορίζουμε:

Ορισμός $\Theta_{\pi,\alpha,\bar{\gamma}} = \{\theta \in \Theta : \pi \cdot \mathrm{P}^{\alpha} \cdot \overline{R}_{\theta}^{\alpha} \cdot \bar{\gamma} \geq \pi \cdot \mathrm{P}^{\alpha} \cdot \overline{R}_{\theta}^{\alpha} \cdot \gamma, \forall \gamma \in \Gamma\}$ (16)

Τότε η (15) λόγω της (16) μπορεί να ξαναγραφεί σαν:

$$Hυ(\pi) = \max_{\alpha \in A} \left\{ \sum_{i=1}^{N} \pi_i \cdot q^{\alpha}(i) + \beta \int_{\theta \in \Theta} \pi \cdot \mathrm{P}^{\alpha} \cdot \overline{R}_{\theta}^{\alpha} \cdot \gamma^{\mathrm{I}(\pi(\alpha,\theta)} \cdot d\theta \right\} =$$

$$\max_{\alpha \in A} \{ \sum_{i=1}^{N} \pi_i q^{\alpha}(i) + \beta . \sum_{\bar{\gamma} \in \Gamma} \int_{\theta \in \Theta_{\pi,\alpha,\bar{\gamma}}} \pi \cdot \mathrm{P}^{\alpha} \cdot \overline{R_{\theta}^{\alpha}} \cdot \bar{\gamma} \cdot d\theta \quad =$$

$$= \max_{\alpha \in A} \left\{ \sum_{i=1}^{N} \pi_i \cdot q^{\alpha}(i) + \beta \cdot \sum_{\gamma \in \Gamma} \pi \cdot \mathrm{P}^{\alpha} \int_{\theta \in \Theta_{\pi,\alpha,\bar{\gamma}}} \overline{R_{\theta}^{\alpha}} \cdot \bar{\gamma} \cdot d\theta \right\} \quad (17)$$

Προκειμένου να απλοποιήσουμε τον υπολογισμό του $\int_{\theta \in \Theta_{\pi,\alpha,\bar{\gamma}}} \pi \cdot \mathrm{P}^{\alpha} \cdot \overline{R_{\theta}^{\alpha}} \cdot \bar{\gamma} \cdot d\theta$ ας υποθέσουμε ότι για ένα δεδομένο π∈Π, α∈Α, και $\bar{\gamma} \in \Gamma$,υπάρχει μονάχα ένας ορισμένος αριθμός από επικοινωνούντα σύνολα στο $\Theta_{\pi,\alpha,\bar{\gamma}}$

Επιπλέον, μπορούμε νά υποθέσουμε ότι τα σύνορα του $\Theta_{\pi,\alpha,\bar{\gamma}}$ έχουν μέτρο 0.

Αφού $\overline{R_{\theta}^{a}}$ είναι ένας διαγώνιος πίνακας που έχει σαν ι-διαγώνιο στοιχείο το $f_i^{a}(\theta)$ και $\bar{\gamma}$ είναι ένα Ν-διάστατο διάνυσμα, που έχει $\bar{\gamma}_i$ το i-στοιχείο του. Παραστατικά $\bar{\gamma} = (\bar{\gamma}_1, \bar{\gamma}_2, \ldots, \bar{\gamma}_N)$.

Τότε με σκοπό την απλοποίηση των παραστάσεων, και για να φανεί η μορφολογική ομοιότητα θεωρούμε

$$\xi^{\alpha}(\pi,\bar{\gamma})=\begin{pmatrix}\int\limits_{\theta\in\Theta_{\pi,\alpha,\bar{\gamma}}} f_1^{\alpha}(\theta)\cdot\bar{\gamma}_1\cdot d\theta\\ \int\limits_{\theta\in\Theta_{\pi,\alpha,\bar{\gamma}}} f_2^{\alpha}(\theta)\cdot\bar{\gamma}_2\cdot d\theta\\ \vdots\\ \int\limits_{\theta\in\Theta_{\pi,\alpha,\bar{\gamma}}} f_N^{\alpha}(\theta)\cdot\bar{\gamma}_N\cdot d\theta\end{pmatrix} \tag{18}$$

Επομένως, η (17) μπορεί να ξαναγραφεί σαν

$$\begin{aligned}Hu(\pi)&=\max_{\alpha\in A}\left\{\sum_{i=1}^{N}\pi_i\cdot q^{\alpha}(i)+\beta\sum_{\Theta_{\pi,\alpha,\bar{\gamma}}}\pi P^{\alpha}\cdot\int_{\theta\in\Theta_{\pi,\alpha,\bar{\gamma}}}\overline{R_{\theta}^{\alpha}}\cdot\bar{\gamma}\cdot d\theta\right\}\\ &=\max_{\alpha\in A}\left\{\sum_{i=1}^{N}\pi_i\cdot q^{\alpha}(i)+\beta\cdot\sum_{\Theta_{\pi,\alpha,\bar{\gamma}}}\pi\cdot P^{\alpha}\cdot\xi^{\alpha}\left(\pi,\bar{\gamma}\right)\right\}\\ &=\max_{\alpha\in A}\left\{\pi\cdot\left[q^{\alpha}+\beta\cdot\sum_{\Theta_{\pi,\alpha,\bar{\gamma}}}P^{\alpha}\cdot\xi^{\alpha}\left(\pi,\bar{\gamma}\right)\right]\right\}\end{aligned} \tag{19}$$

Παρατηρούμε ότι η παράσταση στις αγκύλες $q^{\alpha}+\beta\sum_{\Theta_{\pi,\alpha,\bar{\gamma}}}P^{\alpha}\cdot\xi^{\alpha}\left(\pi,\bar{\gamma}\right)$ είναι ένα N-διάστατο διάνυσμα (gradient) και ένα "gradient" επομένως του τελεστού Hu στο π.

Μολονότι οι υποθέσεις σχετικά με το $\Theta_{\pi,\alpha,\bar{\gamma}}$ είναι ισχυρές, αυτές οι υποθέσεις είναι στην πραγματικότητα αληθείς για πολλές κατανομές που χρησιμοποιούνται συνήθως. Πριν ωστόσο αρχίσουμε τους υπολογισμούς, θεωρείται απαραίτητο να επιβεβαιώσουμε αν οι δεδομένες κατανομές μηνυμάτων

έχουν έναν ορισμένο αριθμό από επικοινωνούντα σύνολα στο $\Theta_{\pi,\alpha,\bar{\gamma}}$.

Για παράδειγμα, αν οι στοχαστικές διαδικασίες μηνυμάτων είναι εκθετικές και υπάρχουν μόνον δυο καταστάσεις στο σύστημά μας, μπορούμε να δείξουμε ότι οι υποθέσεις ισχύουν. Τώρα ας χρησιμοποιήσουμε εκθετική κατανομή στο παράδειγμα που ακολουθεί.

Παράδειγμα

Για να διευκρινίσουμε την μέθοδο παραθέτουμε το παράδειγμα

$$P^1=\begin{bmatrix}0.8 & 0.2\\ 0.5 & 0.5\end{bmatrix} \qquad q^1=\begin{bmatrix}-4\\ 4\end{bmatrix}$$

$$P^2=\begin{bmatrix}0.5 & 0.5\\ 0.4 & 0.6\end{bmatrix} \qquad q^2=\begin{bmatrix}0\\ 3\end{bmatrix}$$

Οι στοχαστικές διαδικασίες μηνυμάτων έχουν τις ακόλουθες συναρτήσεις πυκνότητας πιθανότητας: Για $\theta>0$

$$f_1^1(\theta)=e^{-\theta}, \qquad f_2^1(\theta)=10\cdot e^{-10\theta}$$

$$f_1^2(\theta)=3\cdot e^{-3\theta} \qquad f_2^2(\theta)=2\cdot e^{-2\theta}$$

και $f_1^1(\theta)=f_2^1(\theta)=f_1^2(\theta)=f_2^2(\theta)=0 \quad \text{για}\ \theta\le 0$

Τώρα ας υποθέσουμε ότι $\beta=0.9$ $\pi=(0,1)$ και $\Gamma=\{\gamma^1,\gamma^2\}$ όπου $\gamma^1=[-4,4]^T$ $\gamma^2=[0,3]^T$. Υπολογίζουμε πρώτα το $\Theta_{\pi,1,\gamma 1}$, έχουμε

$$[0,1]\cdot\begin{bmatrix}0.8 & 0.2\\ 0.5 & 0.5\end{bmatrix}\cdot\begin{bmatrix}e^{-\theta} & 0\\ 0 & 10\cdot e^{-10\theta}\end{bmatrix}\cdot\begin{bmatrix}-4\\ 4\end{bmatrix}\ge[0,1]\cdot\begin{bmatrix}0.8 & 0.2\\ 0.5 & 0.5\end{bmatrix}\cdot\begin{bmatrix}e^{-\theta} & 0\\ 0 & 10\cdot e^{-10\theta}\end{bmatrix}\cdot\begin{bmatrix}0\\ 3\end{bmatrix}$$

Λύνοντας την ανισότητα έχουμε: $\Theta_{\pi,1,\gamma 1}=\{0<\theta\le 0.1\}$. Άρα $\Theta_{\pi,1,\gamma 2}=\{0<\theta<0.2\}$. Όμοια βρίσκουμε $\Theta_{\pi,2,\gamma 1}=\{\Theta<-177\}$, που είναι έξω από το πεδίον ορισμού των συναρτήσεων πυκνότητας, οπότε $\Theta_{\pi,2,\gamma 1}=\varnothing$ και $\Theta_{\pi,2,\gamma 2}=\{\theta>0\}$. Κατόπιν χρησιμοποιώντας τα $\Theta_{\pi,1,\gamma 1}$, $\Theta_{\pi,1,\gamma 2}$, $\Theta_{\pi,2,\gamma 1}$, $\Theta_{\pi,2,\gamma 2}$ στην (19), πετυχαίνουμε:

$$Hv([0,1]) = \max[0,1]\cdot\left\{\begin{bmatrix}-4\\4\end{bmatrix}+0.9\cdot\begin{bmatrix}0.8 & 0.2\\0.5 & 0.5\end{bmatrix}\cdot\left(\begin{bmatrix}\int_0^{0.1} -4\cdot e^{-\theta}\cdot d\theta\\ \int_0^{0.1} 4\cdot 10\cdot e^{-10\theta}\cdot d\theta\end{bmatrix}+\begin{bmatrix}0\\ \int_{0.1}^{\infty} 3\cdot 10\cdot e^{-10\theta}\cdot d\theta\end{bmatrix}\right)\right\};$$

$$[0,1]\left\{\begin{bmatrix}0\\3\end{bmatrix}+0.9\cdot\begin{bmatrix}0.5 & 0.5\\0.4 & 0.6\end{bmatrix}\cdot\begin{bmatrix}0\\ \int_0^{\infty} 3\cdot 2\cdot e^{-2\theta}\cdot d\theta\end{bmatrix}\right\}$$

$$=\max\left\{[0,1]\cdot\begin{bmatrix}-3.62\\5.46\end{bmatrix};\ [0,1]\cdot\begin{bmatrix}1.35\\4.62\end{bmatrix}\right\}=5.46$$

Επομένως Hu στο [0, 1] είναι 5.46 και το "gradient" που αντιστοιχεί στην κατάσταση αυτή είναι $[-3.62,\ 5.46]^T$.

Συμπεράσματα

Αναπτύξαμε μια μέθοδο που μπορεί να λύσει ένα πρόβλημα μερικά παρατηρήσιμων Μαρκοβιανών αλυσίδων με συνεχή χώρο μηνυμάτων. Ακολουθήσαμε την ίδια πορεία, με εκείνη που ακολούθησαν αρκετοί ερευνητές και αφορούσε το πρόβλημα P.O.M.D.P. με διακριτό αριθμό μηνυμάτων. Επεκτείναμε τα αποτελέσματα για την περίπτωση που ο χώρος μηνυμάτων είναι συνεχής. Κλειδί για την επέκταση αυτή είναι το λήμμα (3.1) και το θεώρημα (3.2) που ουσιαστικά διαμερίζουν τον χώρο μηνυμάτων, Θ, σε τρόπον ώστε κάθε κελλί της διαμέρισης που προτείνεται, να μπορεί να θεωρηθεί σαν ένα ξεχωριστό διακριτό μήνυμα.

Επομένως μπορούμε να αναδιατυπώσουμε το πρόβλημα P.O.M.D.P. με συνεχή χώρο μηνυμάτων, ώστε να αντιμετωπισθεί με μια μέθοδο επίλυσης ανάλογη με εκείνη που λύνει το πρόβλημα με διακριτό αριθμό μηνυμάτων.

Η παραπάνω επέκταση κρίθηκε αναγκαία, διότι σε αρκετά προβλήματα της καθημερινότητας ο χώρος μηνυμάτων είναι συνεχής. Για παράδειγμα η θερμοκρασία μιας μηχανής είναι συνεχής μεταβλητή.

References

1. Bertsekas, D.P. (1975) Dynamic programming and stochastic control. Academic Press New York.

2. Denardo, E.V. (1967) Contraction mapping in theory underlying dynamic programming. SIAM Reviews 9, (165-177).

3. Howard, R.A. (1960) Dynamic programming and Markov processes. MIT.

4. Lovejoy, W.S. (1991)Computationally feasible bounds for P.O.M.D.P. Operation research, vol 39, No 1, 1991 (162-176).

5. Littman, M.L. Algorithms for sequential decision making PhD thesis, Department of computer science, Brown University (1996).

6. Sondik E.J. (1978) Optimal control of P.O.M.D.P. over the infinite horizon: discount costs, Operation research, vol 2, March 1978 (282-303).

7. Sondik E.J. (1973) Optimal control of P.O.M.D.P. over the infinite horizon: discount costs, Operation research, 21 (1071-1088).

8. Sondik E.J. (1971) Optimal control of P.O.M.D.P.. PhD thesis. Department of electrical engineering, Stanford University 1971.

9. White C.C. and Harrington D.P. (1980) Appication of Jensen's Inequality to adaptive Suboptimal Design, Journal of optimization theory and applications 32, (89-99).

10. Whitt,W. (1978). Approximations of dynamics programs, I Mathematics of operation research 3, (231-243).

11. Abraham Grosfeld-Nir (1996) A Two-state partially observable Markov decision process with uniformly distributed observations. Operation research vol 44, no 3, (458-462).