Κεφάλαιο 1

Μαρκοβιανές διαδικασίες αποφάσεων

1.1 Διαδικασίες Markov με αμοιβές

Έστω μια αδιαχώριστη διαδικασία Markov (X_n) με χώρο καταστάσεων $\mathbb{S}$ και πίνακα πιθανοτήτων μετάβασης P_{ij}. Θα υποθέσουμε ότι κάθε φορά που η διαδικασία βρίσκεται στην κατάσταση j παίρνουμε μια αμοιβή $g(j)$ όπου $g : \mathbb{S} \mapsto \mathbb{R}$ μια δεδομένη συνάρτηση. Έστω επίσης ότι η παρούσα αξία την χρονική στιγμή 0 μιας μοναδιαίας αμοιβής που θα ληφθεί την χρονική στιγμή 1 είναι ίση με $\alpha \in (0, 1]$, όπου α ο συντελεστής απόσβεσης. Αν υποθέσουμε ότι την χρονική στιγμή 0 η αλυσίδα βρίσκεται στην κατάσταση i τότε η παρούσα αξία της συνολικής αμοιβής την χρονική στιγμή 0 των αμοιβών που θα πάρουμε τις πρώτες $n-1$ χρονικές στιγμές είναι $\sum_{k=0}^{n-1} \alpha^k g(X_k)$. Αν η συνάρτηση g είναι φραγμένη (μια συνθήκη που ικανοποιείται αυτόματα αν ο χώρος καταστάσεων είναι πεπερασμένος) τότε έχει σίγουρα νόημα να εξετάσουμε την παρούσα αξία των συνολικών αμοιβών όταν ο χρονικός ορίζοντας είναι άπειρος, δηλαδή η διαδικασία Μάρκωφ δεν σταματά ποτέ. Σ′ αυτή την περίπτωση η παρούσα αξία της συνολικής αμοιβής είναι $\sum_{k=0}^{\infty} \alpha^k g(X_k)$ και είναι ασφαλώς μια τυχαία μεταβλητή.

Ας συμβολίσουμε τώρα με $V(i)$ την μέση τιμή αυτής της τυχαίας μεταβλητής. Συμβολίζοντας με $E_i Y$ την δεσμευμένη μέση τιμή $E[Y|X_0 = i]$ όπου Y είναι μια οποιαδήποτε τυχαία μεταβλητή έχουμε ότι

$$\begin{aligned} V(i) &= E_i \sum_{k=0}^{\infty} \alpha^k g(X_k) = \sum_{k=0}^{\infty} \alpha^k E[g(X_k)|X_0 = i] \\ &= \sum_{k=0}^{\infty} \alpha^k \sum_{j \in \mathbb{S}} P_{ij}^k g(j) = \sum_{j \in \mathbb{S}} g(j) \sum_{k=0}^{\infty} \alpha^k P_{ij}^k \end{aligned}$$

Παρατηρείστε ότι η συνάρτηση $V : \mathbb{S} \mapsto \mathbb{R}$ μπορεί να εκφραστεί βάσει της g καθώς και του πίνακα

$$R_{ij}^{\alpha} := \sum_{k=0}^{\infty} \alpha^k P_{ij}^k. \tag{1.1}$$

Συγκεκριμένα έχουμε ότι

$$V(i) = \sum_{j \in \mathbb{S}} R_{ij}^{\alpha} g(j). \tag{1.2}$$

Η σχέση (1.1 μπορεί να εκφραστεί και ως εξής

$$R^{\alpha} = I + \alpha P + \alpha^2 P^2 + \cdots + \alpha^n P^n + \cdots .$$

Από την παραπάνω σχέση έχουμε, πολλαπλασιάζοντας και τα δύο μέλη από τα δεξιά με αP,

$$\alpha P R^{\alpha} = \alpha P + \alpha^2 P^2 + \cdots + \alpha^n P^n + \cdots .$$

Αφαιρώντας κατά μέλη την δεύτερη από την πρώτη σχέση έχουμε

$$(I - \alpha P) R^{\alpha} = I. \tag{1.3}$$

Κάτω από επιπλέον συνθήκες (για παράδειγμα αν ο χώρος καταστάσεων $\mathbb{S}$ είναι πεπερασμένος και συνεπώς οι πίνακες P και R^{α} είναι πεπερασμένοι τετραγωνικοί πίνακες) από την (1.3) προκύπτει ότι

$$R^{\alpha} = (I - \alpha P)^{-1}. \tag{1.4}$$

Η εξίσωση (1.2) τότε γράφεται στη μορφή

$$V = R^{\alpha} g = (I - \alpha P)^{-1} g. \tag{1.5}$$

Μια ισοδύναμη μέθοδος για τον προσδιορισμό της V βασίζεται στην ανάλυση του πρώτου βήματος και στην μαρκοβιανή ιδιότητα που μας επιτρέπει να γράψουμε την εξής σχέση

$$V(i) = g(i) + \alpha \sum_{j \in \mathbb{S}} P_{ij} V(j). \tag{1.6}$$

Δεν είναι δύσκολο να διαπιστώσουμε την ισοδυναμία των (1.4) και (1.6). Η δικαιολόγηση της (1.6) βασιζεται στο εξής επιχείρημα. Την χρονική στιγμή μηδέν εισπράτουμε $g(i)$ και στη συνέχεια μεταβαίνουμε στην κατάσταση j με πιθανότητα P_{ij}. Η παρούσα αξία των συνολικών απολαβών ξεκινώντας από την κατάσταση j είναι εξ' ορισμού $V(j)$, αλλά θα πρέπει να πολλαπλασιαστεί με τον παράγοντα απόσβεσης α.

Παράδειγμα 1. Εστω μια διαδικασία Markov με χώρο καταστάσεων $\mathbb{S} = \{0, 1, 2, 3\}$ και πίνακα πιθανοτήτων μετάβασης

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

Αν $g(i)$ για $i = 0, 1, 2, 3$ είναι δεδομένοι αριθμοί, και ο συντελεστής απόσβεσης είναι $\alpha < 1$, να ευρεθεί η παρούσα αξία $V(i)$ των συνολικών αμοιβών για κάθε αρχικό σημείο εκκίνησης i. Σ' αυτή την περίπτωση το σύστημα (1.6) γίνεται

$$\begin{aligned} V(0) &= g(0) + \alpha V(1) \\ V(1) &= g(1) + \frac{\alpha}{2}(V(0) + V(2)) \\ V(2) &= g(2) + \frac{\alpha}{2}(V(1) + V(3)) \\ V(3) &= g(3) + \alpha V(2). \end{aligned}$$

1.2 Μέσος ρυθμός αμοιβής

Σε αντίθεση με την προηγούμενη παράγραφο όπου θεωρήσαμε την παρούσα αξία όλων των μελλοντικών αμοιβών, σε πολλά προβλήματα έχουμε διαδικασίες Markov για τις οποίες έχουμε κάποιο κόστος ή κέρδος $g(i)$ όταν βρισκόμαστε στην κατάσταση i και μας ενδιαφέρει το μέσο κόστος (ή κέρδος) ανά μονάδα χρόνου για κάποιο μεγάλο χρονικό ορίζοντα. Θα θεωρήσουμε ότι η αλυσίδα Markov (X_n) είναι αδιαχώριστη, θετικά επαναληπτική. Το μέσο κόστος αυτό εκφράζεται ως

$$\lim_{n\to\infty} \frac{1}{n} E_i \sum_{k=0}^{n-1} g(X_k). \tag{1.7}$$

Η παραπάνω σχέση, εναλλάσοντας το άθροισμα με την δεσμευμένη μέση τιμή γράφεται και ως

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} E[g(X_k)|X_0 = i] &= \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \sum_{j\in\mathbb{S}} P_{ij}^k g(j) \\ &= \sum_{j\in\mathbb{S}} g(j) \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} P_{ij}^k \\ &= \sum_{j\in\mathbb{S}} g(j)\pi_j. \end{aligned} \tag{1.8}$$

Η εναλλαγή του ορίου και του αθροίσματος στον χώρο καταστάσεων $\mathbb{S}$ στην δεύτερη εξίσωση πιο πάνω χρειάζεται μια πιο προσεκτική δικαιολόγηση, επιτρέπεται πάντα όμως όταν ο $\mathbb{S}$ είναι πεπερασμένος. Στην τρίτη εξίσωση χρησιμοποιήσαμε το βασικό οριακό θεώρημα για αδιαχώριστες θετικά επαλαλληπτικές αλυσίδες Markov που μας εξασφαλίζει ότι

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} P_{ij}^k = \pi_j \tag{1.9}$$

για κάθε $i \in \mathbb{S}$, όπου π είναι η στάσιμη κατανομή που προκύπτει ως η (μοναδική μετά την κανονικοποίηση) λύση του συστήματος

$$\pi_j = \sum_{i \in \mathbb{S}} \pi_i P_{ij}. \tag{1.10}$$

1.3 Αμοιβές που εξαρτώνται από μεταβάσεις

Εδώ θα εξετάσουμε την περίπτωση που οι αμοιβές εξαρτώνται από τις μεταβάσεις και όχι μονο από την εκάστοτε κατάσταση στην οποία βρίσκεται η αλυσίδα Markov. Θα θεωρήσουμε δηλαδή ότι υπάρχει μια συνάρτηση $\gamma : \mathbb{S} \times \mathbb{S} \mapsto \mathbb{R}$ που προσδιορίζει την αμοιβή $\gamma(i, j)$ για κάθε μετάβαση από την κατάσταση i στην κατάσταση j. Στην περίπτωση αυτή η παρούσα αξία της συνολικής μέσης αμοιβής ξεκινώντας από την κατάσταση i και με συντελεστή απόσβεσης α είναι

$$V(i) = \sum_{j \in \mathbb{S}} P_{ij}\gamma(i, j) + \alpha \sum_{j \in \mathbb{S}} P_{ij} V(j). \tag{1.11}$$

Παρατηρείστε ότι, αν θέσουμε

$$g(i) := \sum_{j \in \mathbb{S}} P_{ij}\gamma(i, j), \tag{1.12}$$

η (1.11) είναι ίδια με την (1.2).

Παρομοίως, το μέσο κόστος μακροπρόθεσμα δίδεται από την αντίστοιχη έκφραση της (1.7) δηλαδή

$$\lim_{n \to \infty} \frac{1}{n} E_i \sum_{k=0}^{n-1} \gamma(X_k, X_{k+1}). \tag{1.13}$$

Λαμβάνοντας υπ' όψιν ότι

$$\begin{aligned} E_i[\gamma(X_k, X_{k+1})] &= \sum_{(j,l) \in \mathbb{S} \times \mathbb{S}} P(X_{k+1} = l, X_k = j | X_0 = i)\gamma(j, l) \\ &= \sum_{(j,l) \in \mathbb{S} \times \mathbb{S}} P_{ij}^k P_{jl}\gamma(j, l) \\ &= \sum_{j \in \mathbb{S}} P_{ij}^k \sum_{l \in \mathbb{S}} P_{jl}\gamma(j, l) \\ &= \sum_{j \in \mathbb{S}} P_{ij}^k g(j) \end{aligned}$$

όπου στην τελευταία σχέση χρησιμοποιήσαμε τον ορισμό (1.12). Συνεπώς η (1.13) γράφεται ως

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1}\sum_{j\in\mathbb{S}}P_{ij}^k g(j) &= \sum_{j\in\mathbb{S}} g(j)\lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1}P_{ij}^k = \sum_{j\in\mathbb{S}} g(j)\pi_j \\ &= \sum_{(j,l)\in\mathbb{S}\times\mathbb{S}} \pi_j P_{jl}\gamma(j,l).\end{aligned}$$

1.4 Μαρκοβιανές Διαδικασίες Αποφάσεων

Θεωρούμε δύο σύνολα, το σύνολο καταστάσεων, $\mathbb{S}$ που είναι πεπερασμένο ή αριθμήσιμο και το οποίο συνήθως θα ταυτίζουμε με κάποιο υποσύνολο των φυσικών $\{0,1,2,\ldots,\}$ και το σύνολο αποφάσεων $A=\{a,b,c,\ldots\}$ το οποίο θεωρούμε πεπερασμένο. Θεωρούμε επίσης μια *οικογένεια* από πίνακες πιθανοτήτων μετάβασης, $\{P_{ij}(a), a\in A\}$ καθώς και μια συνάρτηση κόστους $C:\mathbb{S}\times A\mapsto\mathbb{R}$. Συγκεκριμένα, $C(i,a)$ είναι το κόστος που πληρώνουμε όταν βρισκόμαστε στην κατάσταση i και επιλέξουμε την απόφαση a. Τέλος έστω α ο συντελεστής απόσβεσης βάσει του οποίου υπολογίζουμε την παρούσα αξία κάποιου μελλοντικού κόστους. Αν $X_k=i$ είναι η κατάσταση την χρονική στιγμή k και $Y_n=a$ η απόφαση που παίρνουμε την ίδια χρονική στιγμή, τότε η επόμενη κατάσταση είναι $X_{k+1}=j$ με πιθανότητα $P(X_{k+1}=j|X_k=i,Y_k=a)=P_{ij}(a)$. Η παρούσα αξία του συνολικού κόστους αυτής της διαδικασίας θα είναι τότε

$$\sum_{k=0}^{\infty}\alpha^k C(X_k,Y_k).$$

Σκοπός μας στο προκείμενο πρόβλημα είναι να βρούμε μια ακολουθία αποφάσεων Y_k, $k=0,1,2,\ldots$ τέτοια ώστε να ελαχιστοποιεί το μέσο συνολικό κόστος (ή να μεγιστοποιεί την μέση συνολική αμοιβή, στην περίπτωση που τα $C(i,a)$ αντιπροσωπεύουν αμοιβές).

Θα ονομάζουμε *πολιτική* κάθε συνάρτηση $f:\mathbb{S}\mapsto A$ από τον χώρο καταστάσεων στον χώρο αποφάσεων. Μια πολιτική με άλλα λόγια είναι ένας κανόνας που υπαγορεύει μια απόφαση, έστω a, κάθε φορά που η αλυσίδα Markov βρίσκεται στην κατάσταση i. Έτσι έχουμε $f(i)=a$. Η υιοθέτηση μιας πολιτικής περιορίζει την ελευθερία που έχουμε να διαλέγουμε τις αποφάσεις μας. Μας αναγκάζει κάθε φορά που βρισκόμαστε σε μια συγκεκριμένη κατάσταση να διαλέγουμε την ίδια πάντα απόφαση. Αποδεικνύεται όμως ότι, λόγω της μαρκοβιανής φύσης της διαδικασίας, υπάρχει *βελτιστη πολιτική* η οποία εξασφαλίζει συνολικό μέσο κόστος εξίσου χαμηλό ή χαμηλότερο από οποιαδήποτε άλλη ακολουθία αποφάσεων που δεν προκύπτει από κάποια πολιτική. Συνεπώς είναι αρκετό

[0]Παραπέμπουμε τους ενδιαφερόμενους για αυτήν, και για όλες τις προτάσεις χωρίς απόδειξη που θα ακολουθήσουν, στον Sheldon Ross, (1970). *Applied Probability Models with Optimization Applications,* Εκδόσεις Dover.

να αναζητήσουμε την βέλτιστη πολιτική χωρίς να χρειάζεται να ασχοληθούμε όλες τις δυνατές ακολουθίες αποφάσεων. Παρατηρούμε πρώτα απ' όλα ότι, αν χρησιμοποιήσουμε μια οποιαδήποτε πολιτική, f, η διαδικασία που προκύπτει ειναι μαρκοβιανή με πίνακα πιθανοτήτων μετάβασης $P_{ij}(f(i))$.

Έστω $V_f(i)$ η παρούσα αξία του συνολικού κόστους όταν ξεκινάμε από την κατάσταση i και εφαρμόζουμε την πολιτική f. Τότε, από την (1.6) έχουμε ότι

$$V_f(i) = C(i, f(i)) + \alpha \sum_{j \in \mathbb{S}} P_{ij}(f(i)) V_f(j). \tag{1.14}$$

Η συνάρτηση V_f ονομάζεται και συνάρτηση αξίας (value function).

1.4.1 Παράδειγμα: Ένα απλό πρόβλημα συντήρησης εξοπλισμού

Πριν προχωρήσουμε στην ανάλυσή μας ας δούμε ένα απλό παράδειγμα. Θεωρούμε μία μηχανή το οποίο μπορεί να βρίσκεται σε δύο καταστάσεις, την κατάσταση ομαλής λειτουργίας, έστω 1 και την κατάσταση κακής λειτουργίας, έστω 0. Εδώ λοιπόν ο χώρος καταστάσεων είναι $\mathbb{S} = \{0, 1\}$. Έστω ότι ο χώρος των αποφάσεων έχει επίσης δύο σημεία, επισκευή r, ή μή επισκευή, n. Συνεπώς $A = \{n, r\}$. Ας υποθέσουμε ακόμη ότι

$$P(n) = \begin{bmatrix} 1 & 0 \\ 1/3 & 2/3 \end{bmatrix}, \qquad P(r) = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}. \tag{1.15}$$

Η έννοια των παραπάνω πινάκων πιθανοτήτων μετάβασης είναι ότι, χωρίς επισκευή μια μηχανή, αν μεν βρίσκεται σε κακή κατάσταση, τότε παραμένει σε κακή κατάσταση, αν δε βρίσκεται σε καλή κατάσταση, τότε με πιθανότητα $1/3$ την επόμενη χρονική στιγμή θα βρεθεί σε κακή κατάσταση. Αντίθετα με επισκευή, η μηχανή την επόμενη χρονική στιγμή βρίσκεται πάντα σε καλή κατάσταση. Τέλος θεωρούμε δεδομένα και τα κόστη $C(1, n) = 0$, $C(1, r) = c$, $C(0, n) = d$, και $C(0, r) = c + d$. Τα κόστη αυτά υπολογίζονται με την παραδοχή ότι κάθε επισκευή κοστίζει c ανεξαρτήτως της κατάστασης της μηχανής και κάθε χρονική περίοδος που η μηχανή βρίσκεται σε κακή κατάσταση κοστίζει d. Θα υποθέσουμε επιπλέον ότι $c < d$ άλλως το πρόβλημα είναι τετριμμένο: αν η επισκευή κοστίζει περισότερο από την κακή λειτουργία τότε δεν επισκευάζουμε ποτέ.

Έστω f_1 η πολιτική σύμφωνα με την οποία $f_1(0) = r$, $f_1(1) = n$, δηλαδή επισκευάζουμε πάντα όταν η μηχανή είναι χαλασμένη και ποτέ όταν είναι σε καλή λειτουργία. Με αυτή την πολιτική η μαρκοβιανή διαδικασία αποφάσεων γίνεται μια απλή αλυσίδα Markov με πίνακα πιθανοτήτων μετάβασης

$$P = \begin{bmatrix} 0 & 1 \\ 1/3 & 2/3 \end{bmatrix}.$$

Η παρούσα αξία του συνολικού μέσου κόστους στην περίπτωση αυτή δίνεται από το σύστημα

$$\begin{aligned} V_{f_1}(0) &= c + d + \alpha V_{f_1}(1) \\ V_{f_1}(1) &= \alpha \frac{V_{f_1}(0) + 2V_{f_1}(1)}{3} \end{aligned}$$

απ' όπου προκύπτει ότι

$$\begin{aligned} V_{f_1}(0) &= \frac{(3-2\alpha)}{(1-\alpha)(3+\alpha)}(c+d) \\ V_{f_1}(1) &= \frac{\alpha}{(1-\alpha)(3+\alpha)}(c+d). \end{aligned} \tag{1.16}$$

Έστω τώρα f_2, f_3 οι πολικτικές $f_2(0) = f_2(1) = n$ και $f_3(0) = f_3(1) = r$ που αντιστοιχούν, η μεν πρώτη στην περίπτωση που δεν επισκευάζουμε ποτέ, η δε δεύτερη στην περίπτωση που επισκευάζουμε πάντα. Οι αντίστοιχοι πίνακες πιθανοτήτων μετάβασης είναι οι $P(n)$ και $P(r)$ της εξίσωσης (1.15). Στην πρώτη περίπτωση έχουμε το σύστημα

$$\begin{aligned} V_{f_2}(0) &= d + \alpha V_{f_2}(0) \\ V_{f_2}(1) &= \alpha \frac{V_{f_2}(0) + 2V_{f_2}(1)}{3} \end{aligned}$$

ενω στην δεύτερη το σύστημα

$$\begin{aligned} V_{f_3}(0) &= c + d + \alpha V_{f_3}(1) \\ V_{f_3}(1) &= c + \alpha V_{f_3}(1). \end{aligned}$$

Λύνοντας αυτά τα συστήματα έχουμε

$$\begin{aligned} V_{f_2}(0) &= \frac{1}{1-\alpha} d \\ V_{f_2}(1) &= \frac{\alpha}{(1-\alpha)(3-2\alpha)} d. \end{aligned} \tag{1.17}$$

και

$$\begin{aligned} V_{f_3}(0) &= d + c\frac{1}{1-\alpha} \\ V_{f_3}(1) &= \frac{1}{1-\alpha} c. \end{aligned} \tag{1.18}$$

Συγκρίνοντας ανάμεσα στις λύσεις αυτές μπορούμε στην προκείμενη περίπτωση να βρούμε την βέλτιστη πολιτική για δεδομενες τιμές των α, c, d. Στην γενική περίπτωση όμως, όταν ο αριθμός των διαθέσιμων αποφάσεων $|A|$, και ο αριθμός των καταστάσεων του συστήματος, $|\mathbb{S}|$ είναι πιό μεγάλος, ο αριθμός όλων των πολιτικών είναι $|\mathbb{S}|^{|A|}$ και συνεπώς η απαρίθμησή τους και ο υπολογισμός όλων των τιμών για κάθε πολιτική δέν είναι πρακτικά εφικτός. (Παρατηρείστε ότι στο παράδειγμα συντήρησης εξοπλισμού που είδαμε, εξετάσαμε μόνο τρείς πολιτικές. Η τέταρτη που είναι 'επισκευή όταν η μηχανή είναι σε καλή κατάσταση και μη επισκευή όταν είναι σε κακή' δεν εξετάσθηκε διότι θεωρήθηκε εκ προϊμίου μη βέλτιστη.)

1.4.2 Η εξίσωση της βέλτιστης πολιτικής

Έστω ότι η f είναι μια βέλτιστη πολιτική για την μαρκοβιανή διαδικασία αποφάσεων που περιγράψαμε και V_f η αντίστοιχη συνάρτηση αξίας όπως προκύπτει από την εξίσωση (1.14). Η V_f θα πρέπει τότε να ικανοποιεί την ακόλουθη εξίσωση

$$V_f(i) = \min_{b \in A} \left(C(i,b) + \alpha \sum_{j \in \mathbb{S}} P_{ij}(b) V_f(j) \right). \tag{1.19}$$

Η αιτιολόγηση για την παραπάνω εξίσωση είναι αρκετά απλή: Αν διαλέξουμε το b ως την πρώτη απόφαση και στη συνέχεια ακολουθήσουμε την βέλτιστη πολιτική η παρούσα αξία του συνολικού κόστους θα είναι $C(i,b) + \alpha \sum_{j \in \mathbb{S}} P_{ij}(b) V_f(j)$. Η καλύτερη δυνατή επιλογή για το b είναι η επιλογή του έτσι ώστε να ελαχιστοποιεί αυτή την ποσότητα και αυτή η ιδιότητα χαρακτηρίζει την βέλτιστη πολιτική.

1.5 Αλγόριθμος Βελτίωσης Πολιτικής (Policy Improvement)

Ο αλγόριθμος αυτός βασίζεται στην παρατήρηση ότι, αν με κάποιο τρόπο γνωρίζαμε την συνάρτηση αξίας για την βέλτιστη πολιτική, V_f, θα μπορούσαμε να προσδιορίσουμε και την βέλτιστη πολιτική.

Ορίζουμε μια αρχική πολιτική f_1 ως εξής

$$f_1(i) = \arg\min_{b \in A} \left(C(i,b) \right). \tag{1.20}$$

Το νόημα αυτής της σχέσης είναι ότι η $f_1(i)$ είναι ίση με την τιμή εκείνη του b που ελαχιστοποιεί την $C(i,b)$. (Σε περίπτωση που περισότερες από μια αποφάσεις ελαχιστοποιούν την $C(i,\cdot)$ θεωρούμε ότι το σύνολο των αποφάσεων A είναι διατεταγμένο και διαλέγουμε την μικρότερη από αυτές.) Στην συνέχεια υπολογίζουμε την συνάρτηση αξίας V_1 που προκύπτει όταν χρησιμοποιούμε την πολιτική f_1 και που υπολογίζεται από το σύστημα

$$V_1(i) = C(f_1(i), i) + \alpha \sum_{j \in \mathbb{S}} P_{ij}(f_1(i)) V_1(j). \tag{1.21}$$

Στην συνέχεια υπολογίζουμε μια καινούργια πολιτική f_2 από την σχέση

$$f_2(i) = \arg\min_{b \in A} \left(C(i,b) + \alpha \sum_{j \in \mathbb{S}} P_{ij}(f_1(i)) V_1(j) \right) \tag{1.22}$$

και την καινούργια συνάρτηση αξίας με βάση την πολιτική f_2 που συμβολίζουμε με V_2 και που υπολογίζουμε από το σύστημα

$$V_2(i) = C(f_2(i), i) + \alpha \sum_{j \in \mathbb{S}} P_{ij}(f_2(i)) V_2(j). \tag{1.23}$$

Δεν είναι δύσκολο να δούμε ότι $V_2(i) \leq V_1(i)$ για κάθε $i \in \mathbb{S}$ και συνεπώς η καινούργια πολιτική είναι καλύτερη από την προηγούμενη.

Φανταζόμαστε ότι έχουμε επαναλάβει αυτή την διαδικασία n φορές και μετά από αυτές τις n επαναλήψεις έχουμε καταλήξει σε μια πολιτική f_n και μια αντίστοιχη συνάρτηση αξίας V_n. Η πολιτική f_{n+1} ορίζεται τότε από την σχέση

$$f_{n+1}(i) = \arg\min_{b \in A} \left(C(i, b) + \alpha \sum_{j \in \mathbb{S}} P_{ij}(f_n(i)) V_n(j) \right) \tag{1.24}$$

και η αντίστοιχη συνάρτηση αξίας V_{n+1} από το σύστημα

$$V_{n+1}(i) = C(f_{n+1}(i), i) + \alpha \sum_{j \in \mathbb{S}} P_{ij}(f_{n+1}(i)) V_{n+1}(j). \tag{1.25}$$

Η συνάρτηση αξίας V_{n+1} είναι με τη σειρά της μικρότερη από την V_n και επομένως η πολική f_{n+1} καλύτερη από την f_n. Δεν είναι δύσκολο να διαπιστώσει κανείς ότι, μετά από πεπερασμένο αριθμό βημάτων ο αλγόριθμος συγκλίνει, δηλαδή ότι υπάρχει N τέτοιο ώστε $f_{N+1} = f_N$ και συνεπώς $V_{N+1} = V_N$. Αυτό είναι φυσική συνέπεια του γεγονότος ότι ο χώρος των πολιτικών είναι ένα πεπερασμένο σύνολο και συνεπώς δεν είναι δυνατόν οι πολιτικές να βελτιώνονται επ΄ άπειρον. Ας συμβολίσουμε λοιπόν με f^* την πολιτική f_N στην οποία ο αλγόριθμος σταματά και με V^* την αντίστοιχη συνάρτηση αξίας. Δεδομένου ότι $f_N = f_{N+1} = f^*$, η σχέση (1.24) με $n = N$ δίνει

$$f^*(i) = \arg\min_{b \in A} \left(C(i, b) + \alpha \sum_{j \in \mathbb{S}} P_{ij}(f^*(i)) V^*(j) \right). \tag{1.26}$$

Παρατηρείστε ότι η f^* ικανοποιεί την εξίσωση βέλτιστης πολιτικής (1.19).

1.6 Παράδειγμα: Πώληση ακινήτου

Υποθέτουμε ότι έχουμε ένα ακίνητο προς πώληση. Κάθε μέρα υποθέτουμε ότι έχουμε μια προσφορά ύψους $i = 0, 1, 2, \ldots, N$ με πιθανότητα p_i για το ακίνητο αυτό. Οι προσφορές θεωρούνται ανεξάρτητες, ισόνομες τυχαίες μεταβλητές. (Η υπόθεση αυτή είναι πιο ρεαλιστική απ΄ ότι φαίνεται αρχικά: Μέρες χωρίς καθόλου προσφορές αντιστοιχούν σε προσφορές ύψους 0 ενώ για τις μέρες που έχουμε περισότερες από μία προσφορές

παίρνουμε την μέγιστη.) Υποθέτουμε επίσης ότι κάθε μέρα που δεν πουλάμε το ακίνητο υποχρεούμεθα να πληρώσουμε κόστος συντήρησης c καθώς και ότι ο συντελεστής απόσβεσης είναι α.

Ξεκινάμε θέτωντας το πρόβλημα αυτό στο γενικό πλαίσιο των μαρκοβιανών διαδικασιών απόφασης. Ο χώρος αποφάσεων έχει δύο στοιχεία, $A = \{s, r\}$, όπου το s είναι πώληση και το r είναι απόρριψη της προσφοράς. Ο χώρος καταστάσεων είναι το μέγεθος της εκάστοτε προσφοράς μαζί με την κατάσταση -1 που είναι η κατάσταση στην οποία βρισκόμαστε μετά την πώληση του ακινήτου. Έτσι έχουμε $\mathbb{S} = \{-1, 0, 1, 2, \ldots, N\}$. Τέλος η συνάρτηση κόστους είναι $C(i, r) = c$, $C(i, s) = -i$, για $i = 0, 1, 2, \ldots, N$ και $C(-1, r) = C(-1, s) = 0$. Οι πίνακες πιθανοτήτων μετάβασης δίδονται από τους

$$P(r) = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & p_0 & p_1 & \cdots & p_{N-1} & p_N \\ 0 & p_0 & p_1 & \cdots & p_{N-1} & p_N \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & p_0 & p_1 & \cdots & p_{N-1} & p_N \\ 0 & p_0 & p_1 & \cdots & p_{N-1} & p_N \end{bmatrix}, \qquad P(s) = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}.$$

(Παρατηρείστε ότι $p_0 + p_1 + \cdots + p_N = 1$.)

Θα συμβολίσουμε την βέλτιστη πολιτική με f^* και την αντίστοιχη συνάρτηση αξίας με V^*. Η συνάρτηση αξίας όταν χρησιμοποιούμε την βέλτιστη πολιτική πρέπει να ικανοποιεί την σχέση

$$V^*(i) = \min\{-i, c + \alpha \sum_{j=0}^{N} p_j V^*(j)\} \tag{1.27}$$

και η βέλτιστη πολιτική την

$$f^*(i) = \arg\min\{-i, c + \alpha \sum_{j=0}^{N} p_j V^*(j)\}. \tag{1.28}$$

(Η παραπάνω εξίσωση πρεπει να ερμηνευθεί ως εξής: αν το πρωτο από τα δύο ορισματα είναι το μικρότερο τότε $f^*(i) = s$, αν το δεύτερο είναι το μικρότερο, τότε $f^*(i) = r$.) Αν θέσουμε

$$i^* = \min\{i : -i < c + \alpha \sum_{j=0}^{N} p_j V^*(j)\} \tag{1.29}$$

βλέπουμε οτι η βέλτιστη πολιτική συνίσταται στην αποδοχή κάθε προσφοράς μεγαλύτερης ή ίσης με i^* και στην απόρριψη κάθε προσφοράς μικρότερης από i^*. Πράγματι, από την (1.27), αν $i < i^*$ τότε $V^*(i) = c + \alpha \sum_{j=0}^{N} p_j V^*(j) < -i$. Συνεπώς, από την (1.28) έχουμε για αυτό το i οτι $\min\{-i, V^*(i)\} = V^*(i)$ και $f^*(i) = r$. Με ανάλογα επιχειρήματα βλέπουμε ότι αν $i \geq i^*$ τότε $f^*(i) = s$.

Μέχρι στιγμής έχουμε καταφέρει να βρούμε όχι την ίδια την βέλτιστη πολιτική αλλά την δομή της: υπάρχει ένας ακέραιος i^* τέτοιος ώστε αν μια προσφορά είναι μεγαλύτερη ή ίση από i^* να πρέπει να σταματήσουμε ενώ διαφορετικά να συνεχίσουμε. Από την στιγμή όμως που η δομή της βέλτιστης πολιτικής είναι γνωστή δεν είναι δύσκολο να την προσδιορίσουμε πλήρως. Πράγματι, έστω f_i η πολιτική που αποδέχεται κάθε προσφορά μεγαλύτερη ή ίση με i. Αν συμβολίσουμε σ΄ αυτή την περίπτωση με T τον αριθμό των προσφορών που απορρίπουμε μέχρι να αποδεχτούμε μια προσφορά είναι εύκολο να δουμε ότι

$$P(T = k) = P(X < i)^{k-1} P(X \geq i), \quad k = 0, 1, 2, \ldots,$$

όπου X μια τυχαία μεταβλητή με κατανομή $P(X = i) = p_i$. Η παρούσα αξία του μέσου κόστους στην συγκεκριμένη περίπτωση δίδεται από την σχέση

$$E\left[c + \alpha c + \cdots + \alpha^T c - \alpha^{T+1} E[X|X \geq i]\right] = c\frac{1 - E[\alpha^{T+1}]}{1 - \alpha} - E[\alpha^{T+1}]E[X|X \geq i]. \tag{1.30}$$

Δεδομένου ότι

$$E\alpha^T = \frac{P(X \geq i)}{1 - \alpha P(X < i)}$$

(πιθανογεννήτρια της γεωμετρικής κατανομής) το δεξί μέλος της (1.30) γίνεται

$$\begin{aligned} G(i) &= \frac{c}{1 - \alpha P(X < i)} - \alpha \frac{P(X \geq i)}{1 - \alpha P(X < i)} E[X|X \geq i] \\ &= \frac{c - \alpha E[X; X \geq i]}{1 - \alpha P(X < i)} \\ &= \frac{c - \alpha \sum_{j=i}^{N} j p_j}{1 - \alpha \sum_{j=0}^{i-1} p_j}. \end{aligned}$$

Το πρόβλημα πλέον ανάγεται στην ελαχιστοποίηση της συνάρτησης $G(i)$.