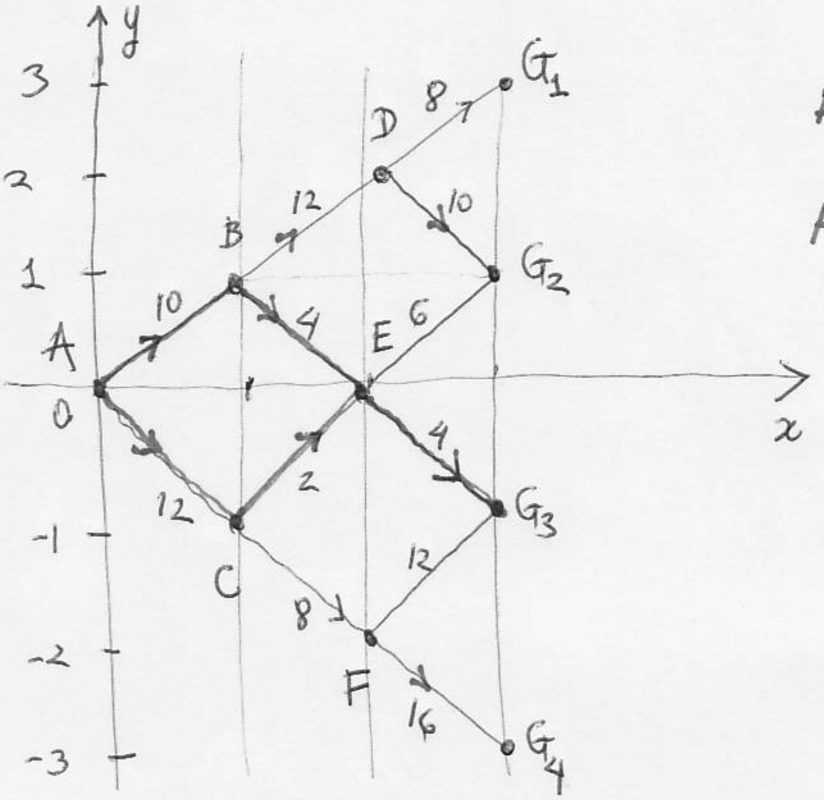
ΕΝΑ ΣΤΟΧΑΣΤΙΚΟ ΠΡΟΒΛΗΜΑ ΔΙΑΔΡΟΜΗΣ ΕΛΑΧΙΣΤΟΥ ΚΟΣΤΟΥΣ.



$A \xrightarrow{u} B \xrightarrow{d} E \xrightarrow{d} G_3$

$A \xrightarrow{d} C \xrightarrow{u} E \xrightarrow{d} G_3$

οι μη στοχαστικές άριστες πολιτικές

Θέλουμε να μεταβούμε από την κορυφή A στην "ευθεία" $\{G_i\}$ με ελάχιστο κόστος. Υπάρχει όμως αβεβαιότητα στο γεγονός εάν ακολουθείται η βέλτιστη διαδρομή. Δηλαδή ενώ σε κάθε κορυφή $(x,y) = S_t$ εκδίδουμε την οδηγία $\beta_t^* \in \{u,d\}$ η οδηγία ακολουθείται με πιθανότητα:

$$p_t(x,y) = P\{\alpha_t = u \mid \beta_t^* = u\} =$$
$$= P\{\alpha_t = d \mid \beta_t^* = d\}$$

και δεν ακολουθείται με πιθανότητα:

$$1 - p_t(x,y) = q_t(x,y) = P\{\alpha_t = d \mid \beta_t^* = u\}$$
$$= P\{\alpha_t = u \mid \beta_t^* = d\}$$

$$V(x,y) = V_t(s_t) = \min_{\beta_t \in \{u,d\}} \left\{ \mathbb{E}\left[c(\alpha_t, s_t) + V_{t+1}(s_{t+1}(\alpha_t, s_t)) \mid \beta_t \right] \right\}$$

$$= \min \Big\{ \mathbb{E}\left[c(\alpha_t, s_t) + V_{t+1}(s_{t+1}(\alpha_t, s_t)) \mid \beta_t = u \right],$$

$$\mathbb{E}\left[c(\alpha_t, s_t) + V_{t+1}(s_{t+1}(\alpha_t, s_t)) \mid \beta_t = d \right] \Big\}$$

$$= \min \Big\{ P\{\alpha_t = u \mid \beta_t = u\} \cdot \big(c(u, s_t) + V_{t+1}(s_{t+1}(u, s_t)) \big) +$$

$$+ P\{\alpha_t = d \mid \beta_t = u\} \cdot \big(c(d, s_t) + V_{t+1}(s_{t+1}(d, s_t)) \big),$$

$$P\{\alpha_t = u \mid \beta_t = d\} \cdot \big(c(u, s_t) + V_{t+1}(s_{t+1}(u, s_t)) \big) +$$

$$+ P\{\alpha_t = d \mid \beta_t = d\} \cdot \big(c(d, s_t) + V_{t+1}(s_{t+1}(d, s_t)) \big) \Big\}$$

$$= \min \Big\{ p_t(x,y) \cdot \big(u(x,y) + V(x+1, y+1) \big) +$$

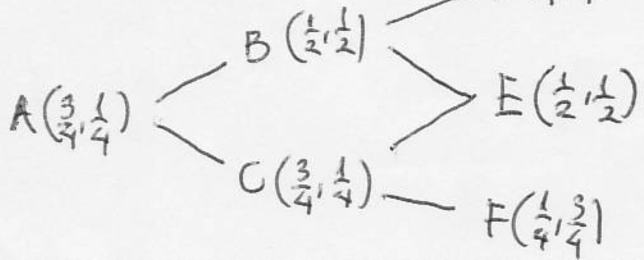
$$+ q_t(x,y) \cdot \big(d(x,y) + V(x+1, y-1) \big),$$

$$q_t(x,y) \cdot \big(u(x,y) + V(x+1, y+1) \big) +$$

$$+ p_t(x,y) \cdot \big(d(x,y) + V(x+1, y-1) \big) \Big\}$$

με συνοριακές συνθήκες: $V(3, 3-2i) = V(G_{i+1}) = 0$, $i = 0, \dots, 3$

και πιθανότητες:



$\underline{\underline{x=2}}$:

$$V(D) = \min\{ p_2(D)(u(D)+\overset{\circ}{V}(G_1)) + q_2(D)(d(D)+\overset{\circ}{V}(G_2)),$$
$$p_2(D)(d(D)+\overset{\circ}{V}(G_2)) + q_2(D)(u(D)+\overset{\circ}{V}(G_1))\}$$
$$= \min\left\{ \frac{17}{2}, \frac{19}{2} \right\} = \frac{17}{2}, \quad \boxed{\beta_2^* = u}$$

$$V(E) = \min\{ p_1(E)(u(E)+V(G_2)) + q_1(E)(d(E)+V(G_3)),$$
$$p_1(E)(d(E)+V(G_3)) + q_1(E)(u(E)+V(G_2))\}$$
$$= \min\{5, 5\} = 5, \quad \boxed{\beta_2^* \in \{u, d\}}$$

$$V(F) = \min\{ p_1(F)(u(F)+V(G_3)) + q_1(F)(d(F)+V(G_4)),$$
$$p(F)(d(F)+V(G_4) + q_1(F)(u(F)+V(G_3))\}$$
$$= \min\{15, 13\} = 13, \quad \boxed{\beta_2^* = d.}$$

$\underline{\underline{x=1}}$:

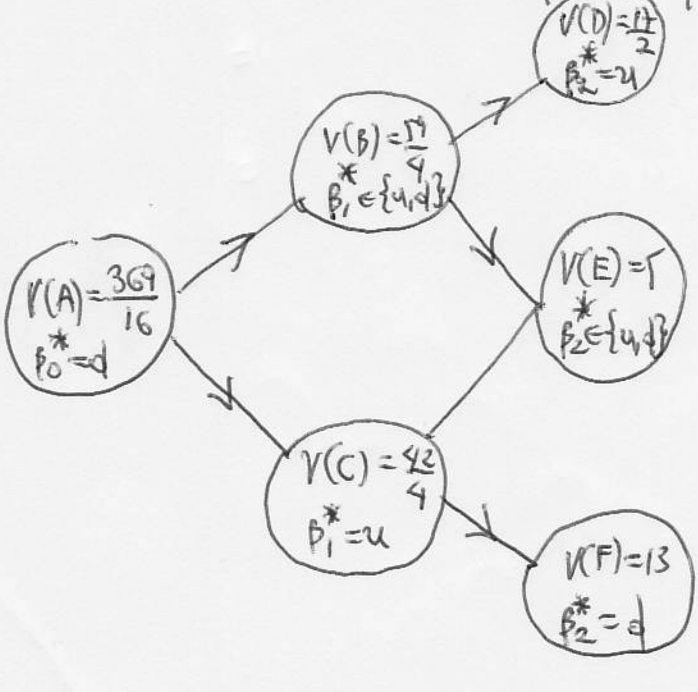
$$V(B) = \min\left\{ \frac{59}{4}, \frac{59}{4} \right\} = \frac{59}{4}, \quad \boxed{\beta_1^* \in \{u, d\}}.$$

$$V(C) = \min\left\{ \frac{42}{4}, \frac{70}{4} \right\} = \frac{42}{4}, \quad \boxed{\beta_1^* = u}$$

$\underline{\underline{x=0}}$:

$$V(A) = \min\left\{ \frac{387}{16}, \frac{369}{16} \right\} = \frac{369}{16}, \quad \boxed{\beta_0^* = d}$$

Έτσι η στοχαστικά άριστη τροχιά είναι:



ΕΝΑ ΑΙΤΙΟΚΡΑΤΙΚΟ ΠΡΟΒΛΗΜΑ ΦΟΡΤΩΣΗΣ

(Η' ΠΩΣ ΝΑ ΛΥΝΟΥΜΕ ΠΡΟΒΛΗΜΑΤΑ ΑΚΕΡΑΙΟΥ ΠΡΟΓΡΑΜΜΑΤΙΣΜΟΥ ΧΡΗΣΙΜΟΠΟΙΩΝΤΑΣ ΔΥΝΑΜΙΚΟ ΠΡΟΓΡΑΜΜ.)

Έστω W = σταθ το μέγιστο δυνατό φορτίο που μπορεί να μεταφερθεί (από ένα ας πούμε αεροσκάφος), και N είδη αντικειμένων τα οποία θέλουμε να μεταφέρουμε. Θέτουμε

u_t = αξία του t-αντικειμένου $1 \leq t \leq N$.

w_t = το βάρος του t-αντικειμένου

Θέλουμε η αξία των αντικειμένων που τελικά θα μεταφερθεί να είναι μέγιστη. Δηλαδή θέλουμε να λύσουμε το εξής πρόβλημα ακεραίου προγρ.

$$\max_{(\alpha_1,\dots,\alpha_N)} \sum_{t=1}^{N} u_t \alpha_t = \sum_{t=1}^{N} u_t \alpha_t^* = V(\alpha_1^*,\dots,\alpha_N^*)$$

<u>έτσι ώστε</u>: $\sum_{t=1}^{N} w_t \alpha_t \leq W$ <u>και</u> $\alpha_t \in \mathbb{Z}^{\geq 0}$, $1 \leq t \leq N$

υπό τον περιορισμό.

Για την εξίσωση Bellman έχουμε:

$$V_t(s_t) = \max_{\alpha_t \in A_t(s_t)} \left\{ b_t(\alpha_t, s_t) + V_{t+1}\left(s_{t+1}(\alpha_t, s_t)\right) \right\}$$

s_t = κατάστασης $\in \{0, 1, 2, \dots, W\}$

$A_t(s_t) = \{0, 1, \dots, \lfloor s_t / w_t \rfloor\} \ni \alpha_t$ = # κομματιών από το t-είδος

$b_t(\alpha_t, s_t) = \text{benefit} = \alpha_t \cdot u_t$

$s_{t+1}(\alpha_t, s_t) = s_t - \alpha_t \cdot w_t \quad , \quad 1 \le t < N$

$$t = N \Rightarrow \begin{cases} V_N(s_N) = \alpha_N \cdot u_N \quad , \quad \alpha_N \in A_N(s_N) = \{ \lfloor s_N / w_N \rfloor \} \\ s_N \in \{0, 1, \dots, W\} \qquad \text{(η συνοριακή συνθήκη)} \end{cases}$$

Αριθμητικό παράδειγμα: Για $W = 5$,

Δίνεται:

t	w_t	u_t
3	1	30
2	3	80
1	2	65

$$\Longleftrightarrow \begin{cases} \max\limits_{(\alpha_1, \alpha_2, \alpha_3)} \overbrace{(65\alpha_1 + 80\alpha_2 + 30\alpha_3)}^{v(\alpha_1, \alpha_2, \alpha_3)} \\ \text{υτπ: } 2\alpha_1 + 3\alpha_2 + \alpha_3 \le \underset{\uparrow W}{5} \end{cases}$$

$t=3$:

$V_3(s_3) = u_3 \cdot \lfloor s_3 / w_3 \rfloor$

$s_3 = \{0, \dots, 5\}$

s_3	0	1	2	3	4	5
$V_3(s_3)$	0	30	60	90	120	150
α_3^*	0	1	2	3	4	5

$t=2$: $V_2(s_2) = \max\limits_{\alpha_2 \in A_2(s_2)} \{ \alpha_2 u_2 + V_3(s_2 - w_2 \alpha_2) \} = \max\limits_{\alpha_2 \in A_2(s_2)} \{ 80\alpha_2 + V_3(s_2 - 3\alpha_2) \}$

$s_2 \in \{0, \dots, 5\} \Rightarrow s_2 = 0 \Rightarrow A_2(0) = \{0\} \Rightarrow V_2(0) = 80 \cdot 0 + V_3(0) = 0 \quad \alpha_2^* = 0$

$s_2 = 1 \Rightarrow A_2(1) = \{0\} \Rightarrow V_2(1) = 80 \cdot 0 + V_3(1) = 30 \quad \alpha_2^* = 0$

$s_2 = 2 \Rightarrow A_2(2) = \{0\} \Rightarrow V_2(2) = 80 \cdot 0 + V_3(2) = 60 \quad \alpha_2^* = 0$

$s_2 = 3 \Rightarrow A_2(3) = \{0, 1\} \Rightarrow V_2(3) = \max\{90, 80\} = 90 \quad \alpha_2^* = 0$

$s_2 = 4 \Rightarrow A_2(4) = \{0, 1\} \Rightarrow V_2(4) = \max\{120, 110\} = 120 \quad \alpha_2^* = 0$

$s_2 = 5 \Rightarrow A_2(5) = \{0, 1\} \Rightarrow V_2(5) = \max\{150, 140\} = 150 \quad \alpha_2^* = 0$

s_2	0	1	2	3	4	5
$V_2(s_2)$	0	30	60	90	120	150
α_2^*	0	0	0	0	0	0

$\underline{t=1}$ Επειδη πάντα έχουμε ολόκληρο το βάρος

$$S_1 \in \{W=5\} \Rightarrow V_1(S_1) = \max_{\alpha_1 \in A_1(S_1)} \{u_1\alpha_1 + V_2(S_1 - w_1\alpha_1)\}$$

$$= \max_{\alpha_1 \in A_1(S_1)} \{65\alpha_1 + V_2(5 - 2\alpha_1)\}$$

$$A_1(S_1) = A_1(5) = \{0, \dots, \lfloor 5/\underbrace{w_1}_{2} \rfloor\} = \{0,1,2\}$$

$$\Downarrow$$

$$V_1(S_1) = \max\{0 + V_2(5),\ 65 + V_2(3),\ 130 + V_2(1)\} = 160$$

κι $\alpha_1^* = 2$

<u>Βέλτιστη τροχιά</u>

$V_1(5) = 160$ | $V_2(1) = 30$ | $V_3(1) = 30$

$\alpha_1^* = 2 \Rightarrow S_2^* = \underbrace{S_1^*}_{5} - w_1\alpha_1^* = 1$ | $\alpha_2^* = 0 \Rightarrow S_3^* = \underbrace{S_2^*}_{1} - w_2\alpha_2^* = 1$ | $\alpha_3^* = 1$

$$\Updownarrow$$

$$\begin{cases} \max(65\alpha_1 + 80\alpha_2 + 30\alpha_3) = V(2,0,1) = 160 \\ (\alpha_1,\alpha_2,\alpha_3) \in \text{Ακέραια} \\ 2\alpha_1 + 3\alpha_2 + \alpha_3 \le 5 \end{cases}$$

MARKOV CHAIN DECISION PROCESSES

Έστω η Markovian διαδικασία $\{\xi_t\}$ διακριτού χρόνου κ' διακριτού χώρου καταστάσεων $\xi_t(\Omega) = S = \{1, 2, \dots, r\}$, $r < \infty$

Ένα σύστημα το οποίο βασίζεται πάνω σε μια Μ.Ρ. χαρακτηρίζεται πλήρως από το παρόν δηλαδή

$$P\{\xi_{t+1} = y \mid \xi_t = x\} = P\{\xi_{t+1} = y \mid \xi_t = x, \xi_{t-1} = x_{t-1}, \dots, \xi_0 = x_0\}$$

για $\forall\ x, y, x_0, \dots, x_{t-1} \in S$.

<u>Θέτουμε:</u> $P_{xy}(\alpha) = P\{\xi_{t+1} = y \mid \xi_t = \alpha(x)\}$, $x, \alpha(x), y \in S$

όπου $\alpha : S \to S$ μια πολιτική. Επιπρόσθετα ζητάμε η συν/ση $P_{xy}(\alpha)$ να μην εξαρτάται από το χρόνο (δηλ. ζητάμε η $\{\xi_t\}$ να είναι χρονικά ομογενής - stationary) ή ότι $P\{\xi_{t+1} = y \mid \xi_t = \alpha(x)\} = P\{\xi_1 = y \mid \xi_0 = \alpha(x)\}$.

Από τα προηγούμενα η (στοχαστική) εξίσωση Bellman θα έχει την εξής γενική μορφή

$$V_t(s_t) = \sup_{\alpha_t \in A_t(s_t)} \Big\{ \underbrace{\mathbb{E}_{\xi}\big(b_t(\alpha_t, s_t)\big)}_{\substack{\text{αναμενόμενο κέρδος} \\ \text{(expected payoff)}}} + \beta \underbrace{\mathbb{E}_{\xi}\Big(V_{t+1}\big(s_{t+1}(\alpha_t, s_t)\big)\Big)}_{\substack{\text{αναμενόμενη συνέχιση} \\ \text{(expected continuation payoff)}}} \Big\}$$

Απλοποιούμε την προηγούμενη εξίσωση θεωρώντας οι συν/σεις $b_t(\alpha_t, S_t)$ να είναι μη στοχαστικές κ' επειδή $\alpha_t \in S = \xi_t(\Omega)$, $A_t(S_t) = A(S_t)$.

$$\mathbb{E}_\xi \left(V_{t+1}(S_{t+1}(\alpha_t, x_t)) \right) = \mathbb{E}\left[V_{t+1}(\xi_{t+1}) \mid \xi_t = \alpha(x) \right] =$$

$$= \sum_{y \in S'} P\{ \xi_{t+1} = y \mid \xi_t = \alpha(x) \} V_{t+1}(y) = \sum_{y \in S'} P_{xy}(\alpha) V_{t+1}(y) : (2.1)$$

Έτσι η εξίσωση Bellman γίνεται:

$$V_t(x) = \sup_{\alpha \in A(x)} \left\{ b_t(\alpha(x), x) + \beta \sum_{y \in S'} P_{xy}(\alpha) V_{t+1}(y) \right\} : (2.2)$$

Σε τέτοιου είδους προβλήματα μας ενδιαφέρει η συμπεριφορά της $V_t(x)$ όταν $t \to \infty$ (infinite horizon problems). Γνωρίζουμε άλλωστε ότι μια εργοδική (aperiodic κ' positive recurrent) κ' ανάγωγη αλυσίδα Markov, συγκλίνει όταν το $t \to \infty$, σε μια αναλλοίωτη κατανομή ανεξάρτητη της αρχικής συνθήκης του συστήματος. Αυτή η ιδιότητα είναι που μας επιτρέπει να πούμε με πια πιθανότητα το σύστημα θα βρίσκεται σε μια συγκεκριμένη κατάσταση εαν ακολουθήσουμε μια συγκεκριμένη πολιτική.

Δεχόμαστε την ύπαρξη του ορίου $V_t(x) \xrightarrow[t\to\infty]{} V(x)$, $\forall x \in S$

κ' η σχέση (2.2) μας δίνει:

$$V(x) = \sup_{\alpha \in A} \left\{ b(\alpha(x), x) + \beta \sum_{y \in S} P_{xy}(\alpha) V(y) \right\}, \quad \forall x \in S$$

που είναι μια <u>συναρτησιακή εξίσωση</u> ως προς V.

ΛΥΝΟΝΤΑΣ ΩΣ ΠΡΟΣ V

<u>Α' τρόπος</u> (Policy iteration).

<u>Βήμα 1</u>: Διαλέγουμε αυθαίρετα ένα κανόνα που αντιστοιχεί μια ενέργεια $\alpha(x)$ σε κάθε κατάσταση $x \in S$ (the policy rule) δηλ. $\alpha(x) = x'$, $x, x' \in S = \{1, 2, \dots, r\}$

<u>Βήμα 2</u>: Χρησιμοποιούμε τις r εξισώσεις

$$V(x) = b(\alpha(x), x) + \beta \sum_{y \in S} P_{xy}(\alpha) V(y)$$

για να λύσουμε ως προς τους αγνώστους $V(x)$, $x \in S = \{1, \dots, r\}$

<u>Βήμα 3</u>: Για κάθε μια από τις r καταστάσεις ελέγχουμε εάν υπάρχει άλλο policy rule $\tilde{\alpha}(x) = x''$ που να μας οδηγεί σε μεγαλύτερο $V(x)$ δηλ. για την κατάσταση x ελέγχουμε:

$$b(\tilde{\alpha}(x), x) + \beta \sum_{y \in S} P_{xy}(\tilde{\alpha}) V(y) > b(\alpha(x), x) + \beta \sum_{y \in S} P_{xy}(\alpha) V(y)$$

εάν υπάρχει τέτοιο $\tilde{\alpha}$ θέτουμε $\alpha(x) \equiv \tilde{\alpha}(x)$ κ' ξαναγυρνά-

-με στο βήμα 2. Εάν δεν υπάρχει τότε έχουμε βρει την V πάνω στο S' κ' την βέλτιστη πολιτική α πάνω στο S'. (optimal policy rule)

<u>Παράδειγμα</u> Αντικατάσταση εργαλείων (Machine replacement)

Στην αρχή κάθε εβδομάδας μια μηχανή βρίσκεται σε μια απο 4 καταστάσεις $S' = \{E = 1, G = 2, A = 3, B = 4\}$.

excellent good average bad

Το εβδομαδιαίο κέρδος απο μία μηχανή σε κατάσταση $x \in S'$ είναι

x'	$b(x', x)$
E	100
G	80
A	50
B	10

Ελέγχοντας την μηχανή στην αρχή κάθε εβδομάδας είμαστε σε θέση να πούμε σε τί κατάσταση βρίσκεται. Έχουμε την επιλογή να την αλλάξουμε με μια νέα μηχανή (που θα είναι σε κατάσταση E) με κόστος 200.

Η ποιότητα της μηχανής από εβδομάδα φθίνει σύμφωνα με το πίνακα μετάβασης πιθανοτήτων της Markovian chain $\{\xi_t\}$ με $\xi_t(\Omega) = \{E, G, A, B\} = \{x_1, x_2, x_3, x_4\}$ $(\tau = 4)$

$$P = \left(P(x_i, x_j)\right) = \left(P\{\xi_{t+1} = x_j \mid \xi_t = x_i\}\right) =$$

present

$$= \begin{array}{c|cccc} E & 0.7 & 0.3 & 0 & 0 \\ G & 0 & 0.7 & 0.3 & 0 \\ A & 0 & 0 & 0.6 & 0.4 \\ B & 0 & 0 & 0 & 1 \\ \hline x_i / x_j & E & G & A & B \end{array} \quad \leftarrow \text{future}$$

$$P\{\xi_{t+1} = G \mid \xi_t = G\} = 0.7$$

$$P\{\xi_{t+1} = B \mid \xi_t = A\} = 0.4$$

$$P\{\xi_{t+1} = B \mid \xi_t = B\} = 1.0 ,$$

Ενώ δίνεται ο παράγοντας προεξόφλησης $\beta = 0.9$

<u>Αυθαίρετα λοιπόν</u> διαλέγουμε το policy rule:

$$\alpha = \begin{pmatrix} E & G & A & B \\ E & G & E & E \end{pmatrix} \iff$$

Αλλάζουμε με νέα μηχανή την παλιά μηχανή, μόνο εάν βρίσκεται σε κατάσταση A ή B.

Το σύστημα των γραμμικών εξισώσεων στο βήμα 2 γίνεται:

$$V(E) = b(\overset{\alpha(E)}{E}, E) + \beta\left\{P\{\xi_{t+1} = E \mid \xi_t = \overset{\alpha(E)}{E}\}V(E) + P\{\xi_{t+1} = G \mid \xi_t = \overset{\alpha(E)}{E}\}V(G)\right\}$$

$$V(G) = b(G, G) + \beta\left\{P\{\xi_{t+1} = G \mid \xi_t = \overset{\alpha(G)}{G}\}V(G) + P\{\xi_{t+1} = A \mid \xi_t = \overset{\alpha(G)}{G}\}V(A)\right\}$$

$$V(A) = b(\overset{\alpha(A)}{E}, E) + \beta\left\{P\{\xi_{t+1} = E \mid \xi_t = \overset{\alpha(A)}{E}\}V(E) + P\{\xi_{t+1} = G \mid \xi_t = E\}V(G)\right\}$$

$$V(B) = b(\overset{\alpha(B)}{E}, E) + \beta\left\{P\{\xi_{t+1} = E \mid \xi_t = \overset{\alpha(B)}{E}\}V(E) + P\{\xi_{t+1} = G \mid \xi_t = E\}V(G)\right\}$$

$$\iff \begin{bmatrix} V(E) \\ V(G) \\ V(A) \\ V(B) \end{bmatrix} = \begin{bmatrix} 100 \\ 80 \\ -100 \\ -100 \end{bmatrix} + (0.9)\begin{bmatrix} 0.7 & 0.3 & 0 & 0 \\ 0 & 0.7 & 0.3 & 0 \\ 0.7 & 0.3 & 0 & 0 \\ 0.7 & 0.3 & 0 & 0 \end{bmatrix}\begin{bmatrix} V(E) \\ V(G) \\ V(A) \\ V(B) \end{bmatrix} \Rightarrow$$

$$\Rightarrow \begin{bmatrix} V(E) \\ V(G) \\ V(A) \\ V(B) \end{bmatrix} = \begin{bmatrix} 687.81 \\ 572.19 \\ 487.81 \\ 487.81 \end{bmatrix} : (6.1)$$

Στην γενική περίπτωση το γραμμικό σύστημα για τους αγνώστους $\underline{V} = [V(1), \ldots, V(n)]^T$ θα είναι

$$V(x) = b(\alpha(x), x) + \beta \sum_{y=1}^{n} P_{xy}(\alpha) V(y) \quad ; \quad x = 1, 2, \ldots, n$$

<u>με λύση</u>: $\underline{V} = (\delta_{ij} - \beta P_{ij}(\alpha))^{-1} \cdot \underline{b}$

<u>οπου</u> $\underline{b} = (b(\alpha(1), 1), \ldots, b(\alpha(n), n))^T$

<u>Ελέγχουμε τώρα εαν η προηγουμένως αυθαίρετα επιλεγμένη policy rule είναι optimal:</u>

(i) Εαν η μηχανή βρίσκεται σε κατάσταση E, δεν χρειάζεται να την αντικαταστήσουμε άρα <u>$\alpha(E) = E$ είναι optimal</u>

(ii) Εαν η μηχ. βρίσκεται σε G κ' επίσης την αλλάξουμε σε E (δηλ δοκιμάζουμε το εναλλακτικό policy rule $\tilde{\alpha}(G) = E$) θα έχουμε:

$$\tilde{V}(G) = -100 + (0.9)((0.7)\cdot(687.81) + (0.3)\cdot(572.19))$$

$$= 487.81 < 572.19 = V(G) \Rightarrow$$

$\Rightarrow$ <u>Μένουμε στο παλιό policy rule.</u> $\Leftrightarrow$ <u>$\alpha(G) = G$ optimal</u>

(iii) Εαν η μηχ. βρίσκεται σε A κατάσταση κ' εμείς εναλλακτικά δεν την αλλάξουμε ($\tilde{\alpha}(A) = A$) τότε

$$\tilde{V}(A) = 50 + (0.9) \cdot \left[(0.6) \cdot (487.81) + (0.4) \cdot (487.81) \right]$$

$$= 489.03 > 487.81 = V(A) \Rightarrow$$

$\Rightarrow$ Αλλάζουμε σε νέο policy rule $\Leftrightarrow$ $\alpha(A) = A$ optimal.

(iv) Εαν η μηχ. βρίσκεται σε B κατάσταση κ' εμείς εναλλακτικά δεν την αλλάξουμε ($\tilde{\alpha}(B) = B$) τότε

$$\tilde{V}(B) = 10 + (0.9) \cdot \left[(487.81)(1) \right] = 449.03 < 487.81 = V(B)$$

$\Rightarrow$ Μένουμε στο παλιό policy rule $\Leftrightarrow$ $\alpha(B) = E$ optimal

Έτσι η νέα policy θα είναι: $\alpha = \begin{pmatrix} E & G & A & B \\ E & G & A & E \end{pmatrix} \Rightarrow$

$$\Rightarrow \underline{V} = \begin{bmatrix} 100 \\ 80 \\ 50 \\ -100 \end{bmatrix} + (0.9) \begin{bmatrix} 0.7 & 0.3 & 0 & 0 \\ 0 & 0.7 & 0.3 & 0 \\ 0 & 0 & 0.6 & 0.4 \\ 0.7 & 0.3 & 0 & 0 \end{bmatrix} \underline{V} \Rightarrow$$

$$\Rightarrow \underline{V} = \begin{bmatrix} 690.23 \\ 575.50 \\ 492.36 \\ 490.23 \end{bmatrix}$$

Θα πρέπει να κάνουμε έλεγχο για optimality στις νέες τιμές:

(i) Σαφώς $\alpha(E) = E$.

(ii) Θέτοντας $\tilde{\alpha}(G) = E \Rightarrow \tilde{V}(G) = -100 + (0.9)[(0.7)(690.23) +$
$+ (0.3)(575.50)] = 490.23 < V(G) = 575.50 \Rightarrow$
$\Rightarrow$ Εμμένουμε σε $\alpha(G) = G$.

(iii) Κατά την ίδια έννοια εμμένουμε σε $\alpha(A) = A$.

(iv) Κ' εμμένουμε σε $\alpha(B) = E$.

<u>Τώρα μπορούμε να πούμε ότι</u>:

βέλτιστη πολιτική : $\alpha = \begin{pmatrix} E & G & A & B \\ E & G & A & E \end{pmatrix}$

και $\underline{V}^T = (690.23,\ 575.50,\ 492.36,\ 490.23)$.

<u>β' Τρόπος</u> Προς τα εμπρός επαναλήψεις της $\underline{V}$
(Value function iterations).

<u>Βήμα 1</u>: Επιλέγουμε αυθαίρετα μια $\underline{V}^{(0)}$ συν/ση
$\underline{V}^{(0)} = (V^{(0)}(1), \dots, V^{(0)}(n))^T$ (συνήθως την μηδενική)

<u>Βήμα 2</u>: $\begin{cases} V^{(n)}(x) = \max\limits_{\alpha \in A(x)} \left\{ b(\alpha(x), x) + \beta \sum\limits_{y=1}^{r} P_{xy}(\alpha) V^{(n-1)}(y) \right\} \\ \forall x \in \{1, \dots, r\} = S \end{cases}$

Βήμα 3 Εάν $\sup_{x \in S} \left| V^{(n)}(x) - V^{(n-1)}(x) \right| \le \varepsilon$

όπου ε προκαθορισμένο σφάλμα, σταματούμε κ' $\underline{V^{(n)}}$ είναι η άγνωστη συν/ση αλλιώς αυξάνουμε το n σε $n+1$ κ' πηγαίνουμε στο Βήμα 2.

Παράδειγμα (Το προηγούμενο παράδειγμα με Value-function iteration).

αρχικά θέτω $V^{(0)}(E) = V^{(0)}(G) = V^{(0)}(A) = V^{(0)}(B) = 0$

State \ Decision	r (replace)	r' (do not replace)	$V^{(1)}$(State)	$\alpha^{(1)*}$
E	-100	100	100	r'
G	-100	80	80	r'
A	-100	50	50	r'
B	-100	10	10	r'

$$V^{(1)}(E) = \max \Big\{ -100 + \beta \cdot \sum_{y \in S} P(\overbrace{r(E)}^{E}, y) \overbrace{V^{(0)}(y)}^{0},$$

$$100 + \beta \sum_{y \in S} P(\underbrace{r'(E)}_{E}, y) \underbrace{V^{(0)}(y)}_{0} \Big\} = 100$$

$$V^{(1)}(G) = \max\Big\{ -100 + \beta \sum_{y \in S} P(\overbrace{r(G)}^{E}, y)\, V^{(0)}(y),$$

$$80 + \beta \sum_{y \in S} P(\underbrace{r'(G)}_{G}, y)\, V^{(0)}(y) \Big\} = 80$$

<u>όμοια</u>: $\begin{cases} V^{(1)}(A) = 50 \\ \alpha^{(1)*}(A) = A \end{cases}$ κ' $\begin{cases} V^{(1)}(B) = 10 \\ \alpha^{(1)*}(B) = B \end{cases}$

State \ Decision	r	r'	$V^{(2)}(\text{State})$	$\alpha^{(2)*}$
E	-15.4	184.6	184.6	r'
G	-15.4	143.9	143.9	r'
A	-15.4	80.6	80.6	r'
B	-15.4	19.0	19.0	r'

$$V^{(2)}(E) = \Big\{ -100 + \beta \sum_{y \in S} P(\overbrace{r(E)}^{E}, y)\, V^{(1)}(y),\ 100 + \beta \sum_{y \in S} P(\overbrace{r'(E)}^{E}, y)\, V^{(1)}(y) \Big\}$$

$$= \Big\{ -100 + \beta \big[\underbrace{P(E,E)}_{0.7} \underbrace{V^{(1)}(E)}_{100} + \underbrace{P(E,G)}_{0.3} \underbrace{V^{(1)}(G)}_{80} + \underbrace{P(E,A)}_{0} V^{(1)}(A) + \underbrace{P(E,B)}_{0} V^{(1)}(B)$$

$$100 + \beta \big[P(E,E) V^{(1)}(E) + P(E,G) V^{(1)}(G) + P(E,A) V^{(1)}(A) + P(E,B) V^{(1)}(B)$$

$$= \{ -15.4,\ 184.6 \} = 184.6, \quad \alpha^{(2)*} = r'$$

State \ Decision	r	r'	$V^{(3)}(\text{State})$	$\alpha^{(3)*}$
E	55.15	255.15	255.15	r'
G	55.15	192.42	192.42	r'
A	55.15	100.36	100.36	r'
B	55.15	27.10	55.15	r

⋮

State \ Decision	r	r'	$V^{(100)}(\text{State})$	$\alpha^{(100)*}$
E	490.22	690.22	690.22	r'
G	490.22	575.49	575.49	r'
A	490.22	492.34	492.34	r'
B	490.22	451.19	490.22	r

<u>Παρατήρηση</u>: Εαν ο αριθμός των states κ' των δυνατών actions είναι μεγάλος οι προς τα εμπρός επαναλήψεις της $\underline{V}$ θα είναι η καλύτερη μέθοδος