ε.δε.μ²
ΕΠΙΣΤΗΜΗ ΔΕΔΟΜΕΝΩΝ & ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΕΡΓΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ

Αίθουσα 005 - Νέα Κτίρια ΣΗΜΜΥ Ε.Μ.Π.

Ενισχυτική Μάθηση - Δυναμικός Προγραμματισμός:

1. Markov Decision Processes
2. Bellman's Optimality Criterion
3. Αλγόριθμος Policy Iteration
4. Αλγόριθμος Value Iteration

καθ. Βασίλης Μάγκλαρης
maglaris@netmode.ntua.gr
www.netmode.ntua.gr

Πέμπτη 9/5/2019

ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ

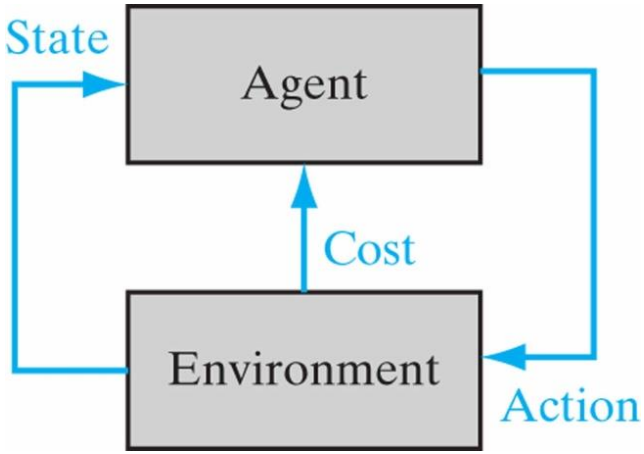
Reinforcement Learning - Markov Decision Processes

Reinforcement Learning

- Αποφάσεις (***actions***) από εξωτερικό ***agent*** σε ορίζοντα N βημάτων που επηρεάζουν την εξέλιξη καταστάσεων (***states***) στοχαστικού περιβάλλοντος και το συνεπαγόμενο κόστος

- Έμφαση σε σχεδιασμό πολιτικής (***policy***) σαν ζεύγη καταστάσεων – αποφάσεων (***states*** – ***actions***) από τον ***agent*** για μέσο - μακροπρόθεσμο στόχο κόστους/οφέλους

- Κύρια εργαλεία βελτιστοποίησης: Δυναμικός προγραμματισμός (***Dynamic Programming***) και στοχαστικές διαδικασίες αποφάσεων ***Markov Decision Processes***



ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ

Reinforcement Learning - Markov Decision Processes (1/2)

Ορισμοί Markov Decision Processes

- Πεπερασμένος Δειγματικός Χώρος $\mathcal{X}$ διακριτών καταστάσεων (***states***) περιβάλλοντος σε διακριτές χρονικές στιγμές (βήματα) $n = 0,1 \dots, N$:
  Τυχαία μεταβλητή $X_n \in \mathcal{X}$ που λαμβάνει τιμές $X_n = i$
- Πεπερασμένος Δειγματικός Χώρος $\mathcal{A}_i$ διακριτών αποφάσεων (***actions***) που ορίζει ο ***agent*** όταν το περιβάλλον βρίσκεται στη κατάσταση $X_n = i$:
  Τυχαία μεταβλητή $A_n \in \mathcal{A}_i$ απόφασης στο n, με τιμές a_{ik} όταν $X_n = i$
- Μεταβάσεις ***Markov*** $p_{ij}(a)$ από κατάσταση περιβάλλοντος i σε κατάσταση j υπό την επήρεια της απόφασης του ***agent*** a στα διακριτά βήματα $n = 0,1 \dots, N$
  $p_{ij}(a) = P(X_{n+1} = j | X_n = i, A_n = a),\ p_{ij}(a) \geq 0,\ \sum_\iota p_{ij}(a) = 1$
- Άμεσο κόστος (***observed cost***) του ***agent*** στο βήμα n όταν παίρνει απόφαση a_{ik} που οδηγεί σε μετάβαση $(X_n = i) \rightarrow (X_{n+1} = j)$:
  $g(i, a_{ik}, j)$ και με ***απόσβεση*** $\gamma^n g(i, a_{ik}, j)$ με συντελεστή $0 \leq \gamma < 1$ (***discount factor***)
  - ✓ Αν $\gamma = 0$ ο ***agent*** δεν ενδιαφέρεται για μελλοντικές επιπτώσεις αποφάσεών του (***myopic***)
  - ✓ Όσο $\gamma \rightarrow 1$ οι αποφάσεις του ***agent*** καθορίζονται σημαντικά από μελλοντικές επιπτώσεις
- Πολιτική (***policy***): $\pi = \{\mu_0, \mu_1, \dots, \mu_n, \dots\}$ όπου μ_n συνάρτηση που στο βήμα n απεικονίζει την κατάσταση του περιβάλλοντος $X_n = i$ στις αποφάσεις $A_n = a$ του ***agent***
  $\mu_n(i) \in \mathcal{A}_i$ για όλες τις καταστάσεις $i \in \mathcal{X}$ (π ***admissible policies***)

Αν $\mu_n(i) = \mu(i) = a$ ανεξάρτητα από το βήμα n η πολιτική π είναι χρονοσταθερή (***stationary***) και οι μεταβάσεις $p_{ij}(a)$ ορίζουν ***αλυσίδα Markov*** $(X_n = i) \rightarrow (X_{n+1} = j)$

ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ

Reinforcement Learning - Markov Decision Processes (2/2)

Ορισμοί Βελτιστοποίησης Δυναμικού Προγραμματισμού

Το συνολικό κόστος αθροίζεται σε πεπερασμένα βήματα (***Finite-Horizon***) ή απεριόριστα (***Infinite-Horizon***) από τα ***άμεσα κόστη των μεταβάσεων Markov*** $X_n \to X_{n+1}$ λόγω $\mu_n(X_n)$:

$$g(X_n, \mu_n(X_n), X_{n+1})$$

Το συνολικό αναμενόμενο κόστος σε απεριόριστο ορίζοντα λόγω πολιτικής $\pi = \{\mu_0, \mu_1, \dots, \mu_n, \dots\}$ με αρχική κατάσταση $X_0 = i$ και απόσβεση γ (***Total Discounted Expected Cost-to-Go over Infinite Horizon***) είναι:

$$J^{\pi}(i) = \mathrm{E}\left[\sum_{n=0}^{\infty} \gamma^n g(X_n, \mu_n(X_n), X_{n+1}) | X_0 = i\right]$$

Ζητείται πολιτική π ελαχιστοποίησης του $J^{\pi}(i)$: $J^*(i) = \min_{\pi} J^{\pi}(i)$

Η ανωτέρω πολιτική π είναι άπληστη (***greedy***) με την έννοια του ότι ο ***agent*** επιλέγει αποφάσεις που ελαχιστοποιούν το ***Expected Cost-to-Go*** $J^{\pi}(i)$ από την αρχική κατάσταση $X_0 = i$ αδιαφορώντας για πιθανές αρνητικές συνέπειες που μπορεί να έχουν μελλοντικά

Αν η πολιτική περιορίζεται σε χρονοσταθερές αποφάσεις $\pi = \{\mu, \mu, \dots\}$ τότε $J^{\pi}(i) \triangleq J^{\mu}(i)$ και το τελικό ζητούμενο είναι η βέλτιστη συνάρτηση $\mu(X_n)$ που ελαχιστοποιεί τα $J^{\mu}(i) = J^*(i)$ για όλες τις αρχικές καταστάσεις $X_0 = i$

Σημείωση: Εναλλακτικά με το κριτήριο ***Total Discounted Expected Cost-to-Go*** μπορεί να οριστεί κριτήριο χωρίς απόσβεση π.χ. ***Expected Average Cost*** ανά βήμα σε ***Infinite Horizon*** (Sheldon Ross, "**Applied Probability Models with Optimization**", Dover, 1992)

ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ

Principle of Optimality (*Bellman 1957*) – Finite Horizon Problem

Έστω διαδικασία αποφάσεων Markov σε ορίζοντα πεπερασμένων βημάτων $n \leq K$ με κόστη $g_n(X_n, \mu_n(X_n), X_{n+1}) \triangleq \gamma^n g(X_n, \mu_n(X_n), X_{n+1}), n < K$ και κόστος τερματικής κατάστασης $g_K(X_K) \triangleq \gamma^K g(X_K)$. Μια βέλτιστη πολιτική $\pi^* = \{\mu_0^*, \mu_1^*, \mu_2^*, \dots, \mu_{K-1}^*\}$ οδηγεί το περιβάλλον στη κατάσταση X_n μετά από n βήματα. Τότε η περικομμένη (***truncated***) πολιτική $\{\mu_n^*, \mu_{n+1}^*, \dots, \mu_{K-1}^*\}$ είναι βέλτιστη για την υπολειπόμενη διαδικασία $\{X_{n+1}, X_{n+2}, \dots, X_K\}$ με κατάσταση εκκίνησης X_n. Το υπολειπόμενο αναμενόμενο κόστος (***Expected Cost-to-Go***) είναι:

$$J_n(X_n) = \mathrm{E}\left[\left\{g_K(X_K) + \sum_{k=n}^{K-1} g_k(X_k, \mu_k(X_k), X_{k+1})\right\} \mid X_n\right]$$

Προσδιορισμός Βέλτιστης Πολιτικής $\pi^* = \{\mu_0^*, \mu_1^*, \mu_2^*, \dots, \mu_{K-1}^*\}$

1. Εύρεση βέλτιστης πολιτικής μ_{K-1}^* για το τελικό βήμα $X_{K-1} \to X_K$
2. Για τα δύο τελικά βήματα $X_{K-2} \to X_{K-1} \to X_K$ εύρεση της μ_{K-2}^* με αναλλοίωτη την μ_{K-1}^*
3. Επανάληψη μέχρι το βήμα $n = 0$ και προσδιορισμός της μ_0^* που συμπληρώνει την π^*

Αλγόριθμος Δυναμικού Προγραμματισμού

1. Εκκίνηση με $J_K(X_K) = g_K(X_K)$ για όλες τις τελικές καταστάσεις X_K
2. Υπολογισμός των $J_n(X_n)$ για όλες τις καταστάσεις X_n με τον ***Αναδρομικό Τύπο*** άπληστων (***greedy***) αποφάσεων:

$$J_n(X_n) = \min_{\mu_n} \mathrm{E}[g_n(X_n, \mu_n(X_n), X_{n+1}) | X_n] \quad \text{για } n = (K-1), (K-2), \dots, 1, 0$$

3. Τελικός προσδιορισμός των $J_0(X_0)$ για όλες τις αρχικές καταστάσεις X_0 και της βέλτιστης $\pi^* = \{\mu_0^*, \mu_1^*, \dots, \mu_{K-1}^*\}$ των αποφάσεων μ_n^* που ελαχιστοποιούν τον αναδρομικό τύπο

ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ

Optimality Equation – Infinite Horizon Problem

Έστω διαδικασία αποφάσεων ***Markov*** με άπειρο ορίζοντα εξέλιξης, πεπερασμένες καταστάσεις $X_n \in \{1,2,\dots,N\}$, κόστη $g_n(X_n, \mu_n(X_n), X_{n+1}) \triangleq \gamma^n g(X_n, \mu_n(X_n), X_{n+1})$ με απόσβεση $0 < \gamma < 1$ και αρχική κατάσταση X_0. Ζητείται η βέλτιστη πολιτική ανάμεσα σε χρονοσταθερές πολιτικές $\pi = \{\mu, \mu, \dots\}$ ελάχιστο ***Expected Cost over Infinite Horizon***.

Με επαναδιατύπωση του ***Αναδρομικού Τύπου Δυναμικού Προγραμματισμού*** και ***αναστροφή της χρονικής εξέλιξης*** σε $n = 0,1,2,\dots$ έχουμε για πεπερασμένο ορίζοντα K:

$$J_{n+1}(X_0) = \min_{\mu} \mathrm{E}[(g(X_0, \mu(X_0), X_1) + \gamma J_n(X_1))|X_0] \text{ και αρχική συνθήκη } J_0(X), \forall X$$

Για άπειρο ορίζοντας η βέλτιστη πολιτική δίνει κόστη $J^*(i) = \lim_{K\to\infty} J_K(i)\,, \forall i = X_0 \Rightarrow$

$$J^*(i) = \min_{\mu} \mathrm{E}[(g(i, \mu(i), X_1) + \gamma J^*(X_1))|X_0 = i]$$

Ορίζουμε το ***άμεσο αναμενόμενο κόστος*** κατάστασης $X_0 = i$ με πολιτική $\mu(X_0) \to X_1 = j$:

$$c\big(i, \mu(i)\big) \triangleq \mathrm{E}[g(i, \mu(i), X_1 = j)|X_0 = i] = \sum_{j=1}^{N} p_{ij} g(i, \mu(i), j)$$

Η βέλτιστη πολιτική δίνει αναμενόμενο κόστος στο 1ο βήμα $\mathrm{E}[J^*(X_1)|X_0 = i] = \sum_{j=1}^{N} p_{ij} J^*(j)$

Τελικά προκύπτουν N εξισώσεις βελτιστοποίησης (***Bellman's Optimality Equations***):

$$J^*(i) = \min_{\mu} \left(c\big(i, \mu(i)\big) + \gamma \sum_{j=1}^{N} p_{ij} J^*(j) \right), i = 1,2,\dots,N$$

Από την επίλυση των N εξισώσεων προκύπτουν τα βέλτιστα αναμενόμενα κόστη από την $X_0 = i$ με απόσβεση σε άπειρο ορίζοντα και η βέλτιστη πολιτική $j = \mu(i)$. Αλγοριθμικά, η βέλτιστη πολιτική ανιχνεύεται με τους βασικούς αλγορίθμους ***Policy*** & ***Value Iteration***

ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ

Αλγόριθμος Policy Iteration (1/2)

Ορισμός Q-factor

Έστω χρονοσταθερή πολιτική $\pi = \{\mu, \mu, \dots\}$ που οδηγεί σε γνωστά ***costs-to-go*** $J^{\mu}(i), \forall i \in \mathcal{X}$ (καταστάσεις του ***περιβάλλοντος***) με αποφάσεις του ***agent*** $a = \mu(i) \in \mathcal{A}_i$

Για κάθε ζεύγος (i, a) στο υπό εξέταση βήμα και πολιτική για τα υπολειπόμενα βήματα $\pi = \{\mu, \mu, \dots\}$ ορίζω τους ***Q-factors*** σαν μέτρο κατάταξης εναλλακτικών άμεσων αποφάσεων $a \in \mathcal{A}_i$ του ***agent***

$$Q^{\mu}(i, a) \triangleq c(i, a) + \gamma \sum_{j=1}^{N} p_{ij}(a)\, J^{\mu}(j)$$

Μια πολιτική $\pi = \{\mu, \mu, \dots\}$ ικανοποιεί τις συνθήκες απληστίας (***greedy conditions***) σε σχέση με τα ***costs-to-go*** $J^{\mu}(i)$ όταν

$$Q^{\mu}(i, \mu(i)) = \min_{a \in \mathcal{A}_i} Q^{\mu}(i, a)$$

Μια πολιτική $\pi^* = \{\mu^*, \mu^*, \dots\}$ είναι βέλτιστη αν ικανοποιεί τις συνθήκες απληστίας (***greedy conditions***) του δυναμικού προγραμματισμού:

$$Q^{\mu^*}(i, \mu^*(i)) = \min_{a \in \mathcal{A}_i} Q^{\mu^*}(i, a)$$

Σημείωση: Όταν τα άμεσα αναμενόμενα κόστη $c(i, a)$ αντικαθίστανται από ***rewards*** $r(i, a)$, τα ***costs-to-go*** $J^{\mu}(i)$ αποκαλούνται ***Value Functions*** $V^{\mu}(i)$ και έχουμε κατ' αντιστοιχία:

$$Q^{\mu}(i, a) \triangleq r(i, a) + \gamma \sum_{j=1}^{N} p_{ij}(a)\, V^{\mu}(j) \text{ και } Q^{\mu^*}(i, \mu^*(i)) = \max_{a \in \mathcal{A}_i} Q^{\mu^*}(i, a)$$

ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ

Αλγόριθμος Policy Iteration (2/2)

Αλγόριθμος Reinforcement Learning

(Αρχιτεκτονική ***Actor – Critic***)

Επαναλήψεις $n = 1,2, \dots$ από δύο βήματα μέχρι σύγκλισης πολιτικής $\pi_n = \pi_{n+1}$

Βήμα 1. **Policy Evaluation** (ο ***critic*** αναλύει τις αποφάσεις του ***agent***):

Με βάση την παρούσα πολιτική $\pi_n = \{\mu_n, \mu_n, \dots\}$ υπολογίζονται τα ***costs-to-go***

$J^{\mu_n}(i) = c(i,a) + \gamma \sum_{j=1}^{N} p_{ij}(a)\, J^{\mu_n}(j)$ για $i = 1,2, \dots, N$

και οι ***Q-factors*** $Q^{\mu_n}(i,a) = c(i,a) + \gamma \sum_{j=1}^{N} p_{ij}(a)\, J^{\mu_n}(j)$ για $i = 1,2, \dots, N$ και $a \in \mathcal{A}_i$

Βήμα 2. **Policy Improvement** (ο ***actor*** καθοδηγεί τις αποφάσεις του ***agent***):

Η πολιτική π_n βελτιώνεται σε π_{n+1} μέσω της $\mu_{n+1}(i) = \arg\min_{a \in \mathcal{A}_i} Q^{\mu_n}(i,a)$ για $i = 1,2, \dots, N$

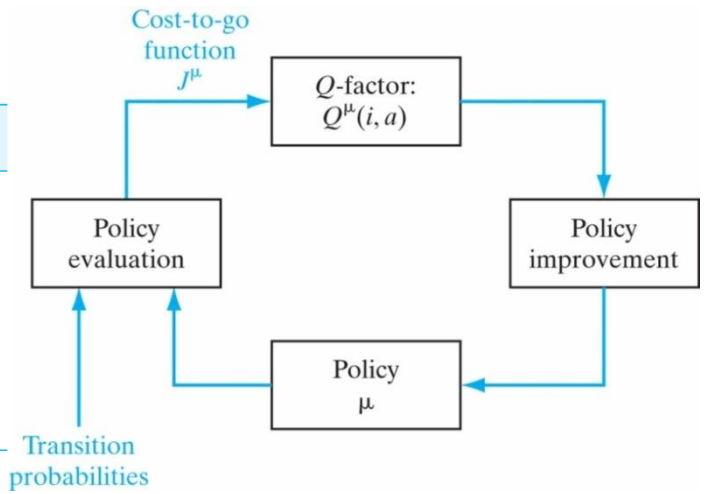
$\arg\min_x f(x)$: Η τιμή της x που οδηγεί την $f(x)$ σε ελάχιστο

TABLE 12.1 Summary of the Policy Iteration Algorithm

1. Start with an arbitrary initial policy μ_0.
2. For $n = 0, 1, 2, \dots$, compute $J^{\mu_n}(i)$ and $Q^{\mu_n}(i,a)$ for all states $i \in \mathcal{X}$ and actions $a \in \mathcal{A}_i$.
3. For each state i, compute

$$\mu_{n+1}(i) = \arg\min_{a \in \mathcal{A}_i} Q^{\mu_n}(i,a)$$

4. Repeat steps 2 and 3 until μ_{n+1} is not an improvement on μ_n, at which point the algorithm terminates with μ_n as the desired policy.



Ο αλγόριθμος συγκλίνει σε βέλτιστη πολιτική σε πεπερασμένα βήματα n λόγω πεπερασμένου πλήθους καταστάσεων N και επιλογών αποφάσεων

ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ

Value Iteration Algorithm

Εκτίμηση των Συναρτήσεων Cost-to-Go μέσω Διαδοχικών Προσεγγίσεων $J_n(i) \to J_{n+1}(i)$

- Εκκίνηση με αυθαίρετες τιμές $J_0(i) \;\; \forall i$
- Επαναλήψεις $n \to n+1$ μέχρι ***ανεκτή σύγκλιση*** (θεωρητικά $n \to \infty$) μέσω σχέσεων ***backup***:

  $J_{n+1}(i) = \min_{a \in \mathcal{A}_i} \{ c(i,a) + \gamma \sum_{j=1}^{N} p_{ij}(a) J_n(j) \}$ για $i = 1,2,\dots,N$ (από εξισώσεις ***Bellman***)
- Τελικός υπολογισμός των βέλτιστων ***Costs-to-Go***

  $$J^*(i) = \lim_{n \to \infty} J_n(i), \; Q^*(i,a) = c(i,a) + \gamma \sum_{j=1}^{N} p_{ij}(a) J^*(j)$$

  και προσδιορισμός της ***βέλτιστης πολιτικής*** $\mu^*(i) = \arg\min_{a \in \mathcal{A}_i} Q^*(i,a)$ για $i = 1,2,\dots,N$

TABLE 12.2 Summary of the Value Iteration Algorithm

1. Start with arbitrary initial value $J_0(i)$ for state $i = 1, 2, \dots, N$.
2. For $n = 0, 1, 2, \dots$, compute

$$J_{n+1}(i) = \min_{a \in \mathcal{A}_i} \left\{ c(i,a) + \gamma \sum_{j=1}^{N} p_{ij}(a) J_n(j), \right\}, \qquad \begin{matrix} a \in \mathcal{A}_i \\ i = 1, 2, \dots, N \end{matrix}$$

Continue this computation until

$$|J_{n+1}(i) - J_n(i)| < \epsilon \qquad \text{for each state } i$$

where ϵ is a prescribed tolerance parameter. It is presumed that ϵ is sufficiently small for $J_n(i)$ to be close enough to the optimal cost-to-go function $J^*(i)$. We may then set

$$J_n(i) = J^*(i) \qquad \text{for all states } i$$

3. Compute the Q-factor

$$Q^*(i,a) = c(i,a) + \gamma \sum_{j=1}^{N} p_{ij}(a) J^*(j) \qquad \begin{matrix} \text{for } a \in \mathcal{A}_i \text{ and} \\ i = 1, 2, \dots, N \end{matrix}$$

Hence, determine the optimal policy as a greedy policy for $J^*(i)$:

$$\mu^*(i) = \arg\min_{a \in \mathcal{A}_i} Q^*(i,a)$$

Ο αλγόριθμος ***Value Iteration*** συνήθως συγκλίνει ικανοποιητικά και θεωρείται αποτελεσματικότερος του ***Policy Iteration*** καθώς αποφεύγει υπολογισμούς όλων των ***Costs-to-Go*** $J^{\mu_n}(i)$ σε κάθε βήμα

ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ & ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ

Παράδειγμα Δυναμικού Προγραμματισμού: Βελτιστοποίηση Δρομολόγησης

Εύρεση Δρόμων Ελάχιστου Κόστους από Κόμβο A *σε* Κόμβο J μέσω του μονο-κατευθυντικού γράφου όπως στο σχήμα με κατεύθυνση γραμμών ***Δ → Α***

Ενδεικτικό κόστος γραμμών: $A \to B: 2, B \to A: \infty$

$B \to F: 4, F \to B: \infty$

Ενδεικτικό κόστος δρόμου: Δρόμος $\{A, B, F, I, J, Q\}$: $2 + 4 + 3 + 4 = 13$

Κατάσταση Περιβάλλοντος: Κόμβος σε παρούσα διερεύνηση $\{A, B, \dots, J\}$

Αποφάσεις Agent: Επόμενος κόμβος για διερεύνηση $\{up, down, staight\}$

Αναδρομικός Υπολογισμός *Q-Factors*:

$Q(H, down) = 3 \quad Q(I, up) = 4$

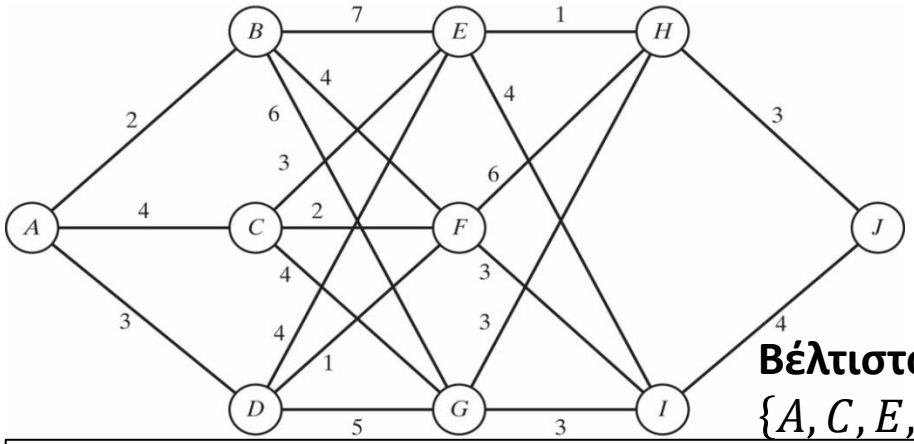
$Q(E, staight) = 1 + 3 = 4 \quad Q(E, down) = 4 + 4 = 8$

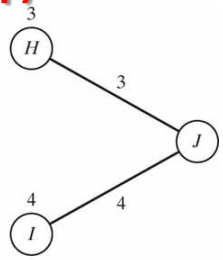
$Q(F, up) = 6 + 3 = 9 \quad Q(F, down) = 3 + 4 = 7$

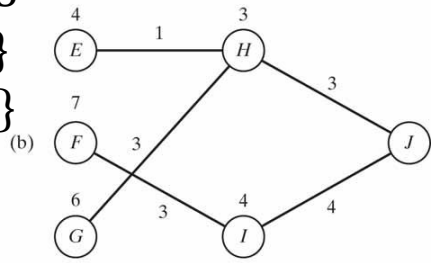
.....

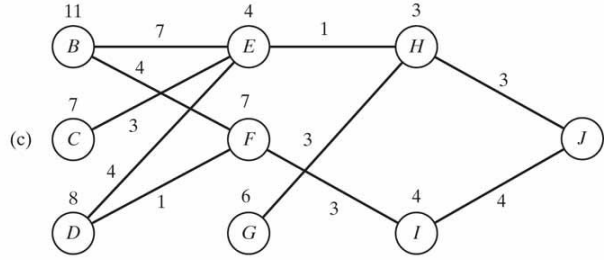
Κατεύθυνση Γραμμών

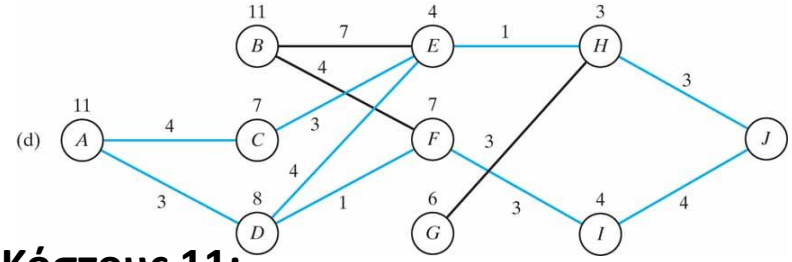
Δ ⟶ ***Α***











Βέλτιστοι Δρόμοι Κόστους 11:

$\{A, C, E, H, J\}, \{A, D, E, H, J\}, \{A, D, F, I, J\}$

Αλγόριθμοι Δυναμικού Προγραμματισμού ***Bellman-Ford*** στηρίζουν την δρομολόγηση ***Border Gateway Protocols*** (***BGP***) ανάμεσα στα ~62,000 Αυτόνομα Συστήματα (***Autonomous Systems, AS***) στο ***Internet*** (~750,000 γνωστά δίκτυα)