Άσκηση 1.1

1. Έχουμε ότι $v[n] = x[n] * h_1[n]$ και $y[n] = v[n] * h_2[n]$. Από τον ορισμό της συνέλιξης, $v[n] = \sum_{k=-\infty}^{\infty} x[k]\, h_1[n-k]$. Επομένως, $y[n] = \left(\sum_{k=-\infty} x[k]\, h_1[n-k] \right) *$

$$* \, h_2[n] = \sum_{\ell=-\infty}^{\infty} \left(\sum_{k=-\infty}^{\infty} x[k]\, h_1[\ell-k] \right) h_2[n-\ell] = \sum_{k=-\infty}^{\infty} x[k] \sum_{\ell=-\infty}^{\infty} h_1[\ell-k]\, h_2[n-\ell]$$

Θέτουμε $m = \ell - k$ και λαμβάνουμε ότι $y[n] = \sum_{k=-\infty}^{\infty} x[k] \sum_{m=-\infty}^{\infty} h_1[m]\, h_2[n-(k+m)]$

$$\Rightarrow y[n] = \sum_{k=-\infty}^{\infty} x[k] \sum_{m=-\infty}^{\infty} h_1[m]\, h_2[(n-k)-m] = \sum_{k=-\infty}^{\infty} x[k] \left(h_1[n-k] * h_2[n-k] \right) =$$

$$= \sum_{k=-\infty}^{\infty} x[k]\, (h_1 * h_2)[n-k] = x[n] * (h_1 * h_2)[n] = x[n] * h[n] \Rightarrow$$

$\Rightarrow y[n] = x[n] * h[n]$ όπου $h[n] = h_1[n] * h_2[n]$

2. Ισχύει ότι $h_1[n] * h_2[n] = \sum_{k=-\infty}^{\infty} h_1[k]\, h_2[n-k]$. Θέτουμε $\ell = n-k$ και έχουμε $h_1[n] * h_2[n] = \sum_{\ell=\infty}^{-\infty} h_1[n-\ell]\, h_2[\ell] = \sum_{\ell=-\infty}^{\infty} h_2[\ell]\, h_1[n-\ell] = h_2[n] * h_1[n]$

$\Rightarrow h_1[n] * h_2[n] = h_2[n] * h_1[n]$

3. Έχουμε ότι $v[n] = x[n] * h_1[n] \overset{Z}{\Rightarrow} V(z) = X(z)\, H_1(z)$ όπου $H_1(z) = \sum_{r=0}^{M} b_r z^{-r}$

Αντικαθιστώντας, $V(z) = X(z) \sum_{r=0}^{M} b_r z^{-r} = \sum_{r=0}^{M} b_r z^{-r} X(z) \Rightarrow V(z) = \sum_{r=0}^{M} b_r z^{-r} X(z) \overset{Z^{-1}}{\Rightarrow}$

$v[n] = \sum_{r=0}^{M} b_r\, x[n-r]$ από την ιδιότητα $x[n-m] \overset{Z}{\rightleftarrows} z^{-m} X(z)$. Ομοίως

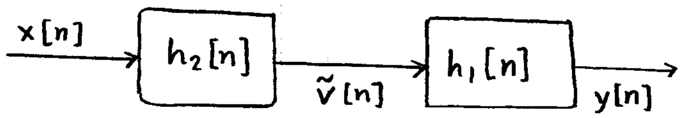
έχουμε ότι $y[n] = v[n] * h_2[n] \overset{Z}{\Rightarrow} Y(z) = V(z)\, H_2(z)$ όπου $H_2(z) = \dfrac{1}{1 - \sum_{k=1}^{N} a_k z^{-k}}$

$$Y(z) = V(z) \frac{1}{1 - \sum_{k=1}^{N} a_k z^{-k}} \Rightarrow Y(z)\left[1 - \sum_{k=1}^{N} a_k z^{-k}\right] = V(z) \Rightarrow Y(z) - Y(z) \sum_{k=1}^{N} a_k z^{-k} =$$

$$= Y(z) - \sum_{k=1}^{N} a_k z^{-k} Y(z) = V(z) \overset{Z^{-1}}{\Rightarrow} y[n] - \sum_{k=1}^{N} a_k\, y[n-k] = v[n] \Rightarrow$$

$$y[n] - \sum_{k=1}^{N} a_k\, y[n-k] = \sum_{r=0}^{M} b_r\, x[n-r]$$

4. Σε αυτή την περίπτωση, έχουμε ότι $\tilde{v}[n] = x[n] * h_2[n]$



Επομένως $\tilde{V}(z) = X(z)\, H_2(z) =$

$$= X(z) \frac{1}{1 - \sum_{k=1}^{N} a_k z^{-k}} \Rightarrow \tilde{V}(z)\left[1 - \sum_{k=1}^{N} a_k z^{-k}\right] = X(z) \Rightarrow \tilde{V}(z) - \sum_{k=1}^{N} a_k z^{-k} \tilde{V}(z) =$$

$$= X(z) \overset{Z^{-1}}{\Rightarrow} \tilde{v}[n] - \sum_{k=1}^{N} a_k\, \tilde{v}[n-k] = x[n]$$

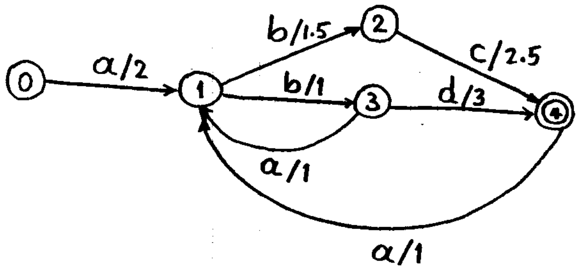
Επίσης, $y[n] = \tilde{v}[n] * h_1[n] \overset{Z}{\Rightarrow} Y(z) = \tilde{V}(z)\, H_1(z) = \tilde{V}(z) \sum_{r=0}^{M} b_r z^{-r}$

Επομένως, $Y(z) = \sum_{r=0}^{M} b_r z^{-r} \tilde{V}(z) \xrightarrow{Z^{-1}} y[n] = \sum_{r=0}^{M} b_r \tilde{v}[n-r]$. Τελικά,

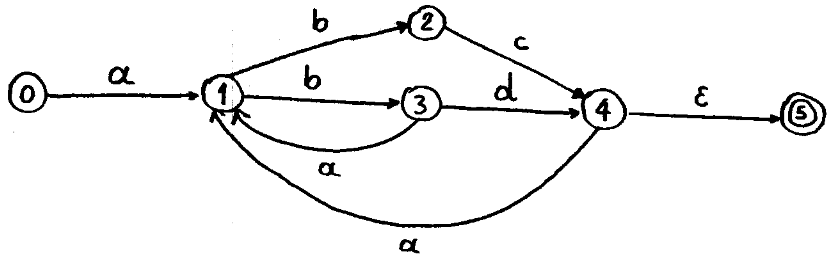
$$\tilde{v}[n] - \sum_{k=1}^{N} a_k \tilde{v}[n-k] = x[n] \quad \text{και} \quad y[n] = \sum_{r=0}^{M} b_r \tilde{v}[n-r]$$

Οι δύο μορφές των εξισώσεων διαφορών είναι ισοδύναμες.

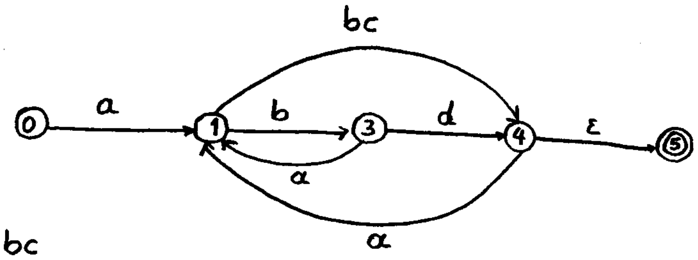
Άσκηση 1.3



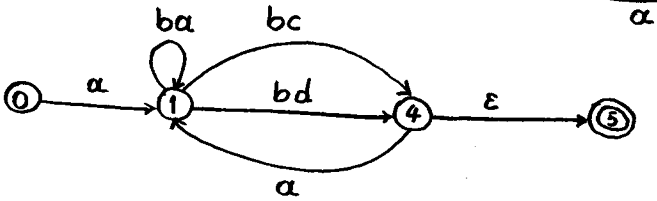
1. Για να βρούμε την κανονική έκφραση που αντιστοιχεί στην μηχανή, θα εφαρμόσουμε τον αλγόριθμο απαλοιφής καταστάσεων. Η αρχική κατάσταση δεν έχει εισερχόμενες ακμές, επομένως δεν χρειάζεται να κάνουμε κάτι. Η τελική κατάσταση έχει εξερχόμενες ακμές, επομένως θεωρούμε νέα τελική κατάσταση ως εξής:



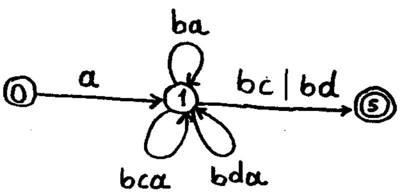
Αρχικά, απαλοίφουμε την κατάσταση 2:



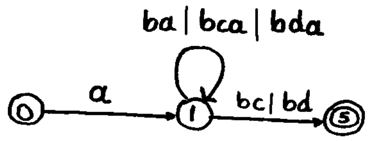
Έπειτα, την κατάσταση 3:



Μετά, την κατάσταση 4:



ή



Η ζητούμενη κανονική έκφραση είναι η $a(ba|bca|bda)*(bc|bd)$

2. Αναζητούμε την γραμματοσειρά ελαχίστου συνολικού κόστους εφόσον χρησιμοποιούμε τροπικό ημιδακτύλιο. Εύκολα βλέπουμε ότι αρκεί να εξετάσουμε τις γραμματοσειρές abc, abd καθώς οποιαδήποτε άλλη θα είναι της μορφής $a(ba \mid bca \mid bda)+(bc \mid bd)$ και θα έχει υποχρεωτικά μεγαλύτερο κόστος. Το κόστος της abc είναι $2+1.5+2.5=6$ και το κόστος της abd είναι $2+1+5+3=11$ καθώς η κατάσταση 3 έχει κόστος 5. Συνεπώς η πιο πιθανή γραμματοσειρά είναι η abc.

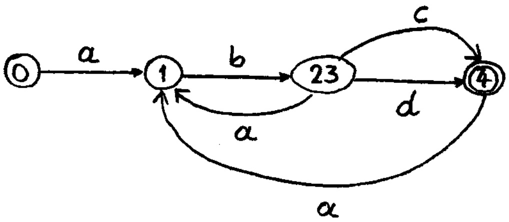
3. Η μοναδική ακολουθία καταστάσεων που αντιστοιχεί στην γραμματοσειρά abcababd είναι η $0 \xrightarrow[2]{a} 1 \xrightarrow[1.5]{b} 2 \xrightarrow[2.5]{c} 4 \xrightarrow[1]{a} 1 \xrightarrow[1]{b} 3 \xrightarrow[1]{a} 1 \xrightarrow[1]{b} 3 \xrightarrow[3]{d} 4$

με κόστος $2+1.5+2.5+1+1+1+1+3=13$

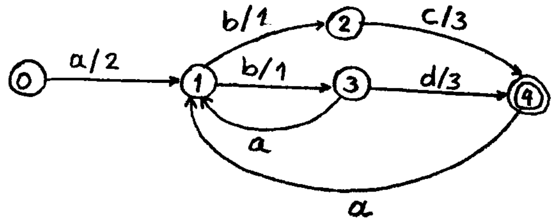
4. Η μηχανή που δίνεται στην εκφώνηση είναι μη ντετερμινιστική, καθώς από την κατάσταση 1 με είσοδο b μεταβαίνουμε είτε στην κατάσταση 2 είτε στην κατάσταση 3. Θα χρησιμοποιήσουμε τον αλγόριθμο μετατροπής από μη ντετερμινιστική σε ντετερμινιστική μηχανή.

ΚΑΤΑΣΤΑΣΗ \ ΕΙΣΟΔΟΣ	a	b	c	d
{0}	{1}	–	–	–
{1}	–	{2,3}	–	–
{2,3}	{1}	–	{4}	{4}
{4}	{1}	–	–	–

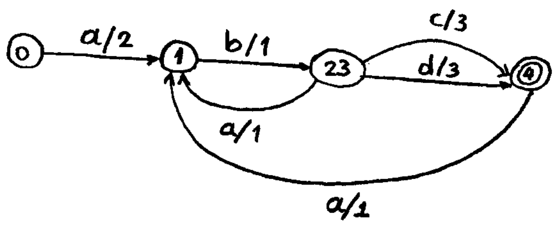
Με βάση τον παραπάνω πίνακα, η ντετερμινιστική μηχανή που προκύπτει είναι η εξής:



5. Παρατηρούμε ότι μπορούμε να μεταφέρουμε 0.5 μονάδες κόστους από την ακμή 1→2 στην 2→4 χωρίς να αλλάξει το κόστος οποιασδήποτε γραμματοσειράς που γίνεται αποδεκτή, λαμβάνοντας

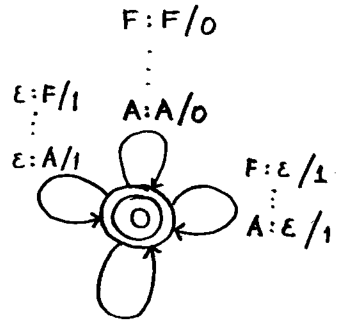


Συγχωνεύοντας τις καταστάσεις 2,3 όπως στο προηγούμενο ερώτημα λαμβάνουμε την ισοδύναμη ντετερμινιστική μηχανή με κόστος:



Άσκηση 1.4

α. Ο ζητούμενος μετατροπέας είναι ο εξής:



$X:Y/1$ όπου $(X,Y) \in \Sigma \times \Sigma$ και $X \neq Y$

β. Α΄ ΤΡΟΠΟΣ με χρήση δυναμικού προγραμματισμού. Εκφράζουμε το κόστος αντιστοίχισης από την γραμματοσειρά S="EDBAEDC" στην T="CDFABEA" μέσω της εξής αναδρομικής σχέσης:

$$D(1..i,\ 1..j) = \min \begin{cases} D(1..i,\ 1..j-1)+1 & \text{(εισαγωγή)} \\ D(1..i-1,\ 1..j)+1 & \text{(διαγραφή)} \\ D(1..i-1,\ 1..j-1) + d(S(i), T(j)) & \text{(αντικατάσταση)} \end{cases}$$

όπου $D(1..i,\ 1..j)$ δηλώνει το ελάχιστο κόστος αντιστοίχισης από την $S(1..i)$ στην $T(1..j)$, και προκύπτει είτε αντιστοιχίζοντας την $S(1..i)$ στην $T(1..j-1)$ και εισάγοντας τον χαρακτήρα $T(j)$, είτε αντιστοιχίζοντας την $S(1..i-1)$ στην $T(1..j)$ και διαγράφοντας τον χαρακτήρα $S(i)$, είτε αντιστοιχίζοντας την $S(1..i-1)$ στην $T(1..j-1)$ και αντικαθιστώντας τον $S(i)$ με $T(j)$ αν $S(i) \neq T(j)$.

SOURCE

C	7	6	6	6	5	5	5	5
D	6	6	5	5	4	4	4	4
E	5	5	4	4	3	3	3	4
A	4	4	3	3	2	3	4	4
B	3	3	2	2	3	3	4	5
D	2	2	1	2	3	4	5	6
E	1	1	2	3	4	5	5	6
#	0	1	2	3	4	5	6	7
	#	C	D	F	A	B	E	A

TARGET

Η βέλτιστη αντιστοίχιση έχει κόστος $D(1..7, 1..7) = 5$ και προκύπτει μέσω 5 αντικαταστάσεων και 2 no edits (7 βήματα):

E D B A E D C
↓ ↓ ↓ ↓ ↓
C D F A B E A

3. Η δεύτερη καλύτερη αντιστοίχιση κόστους 5 προκύπτει από διαφορετικό μονοπάτι (διακεκομμένα βέλη) και χρειάζεται 8 βήματα: 3 αντικαταστάσεις, 2 no edits, 1 εισαγωγή και 1 διαγραφή:

E D B A ε E D C
↓ | | ↓ ↓
C D F A B E A ε

Β΄ ΤΡΟΠΟΣ Με χρήση του Levenshtein transducer και του λογισμικού Openfst, βρίσκουμε τα δυο συντομότερα μονοπάτια στην μηχανή T o L o S όπου S ο αποδοχέας της EDBAEDC και T ο αποδοχέας της CDFABEA.

Άσκηση 2.1

1. Έχουμε ότι $E_x = \sum_{n=0}^{N-1+p} (x[n] - \hat{x}[n])^2$ όπου $\hat{x}[n] = \sum_{k=1}^{p} a_{xk}\, x[n-k]$

Για $a_{x0} = -1$, $\hat{x}[n] = \sum_{k=0}^{p} a_{xk}\, x[n-k] + x[n]$. Επομένως,

$$E_x = \sum_{n=0}^{N-1+p} \left(\cancel{x[n]} - \sum_{k=0}^{p} a_{xk}\, x[n-k] - \cancel{x[n]}\right)^2 = \sum_{n=0}^{N-1+p} \left(\sum_{k=0}^{p} a_{xk}\, x[n-k]\right)\left(\sum_{k'=0}^{p} a_{xk'}\, x[n-k']\right)$$

$$= \sum_{k=0}^{p} a_{xk} \sum_{k'=0}^{p} a_{xk'} \sum_{n=0}^{N-1+p} x[n-k]\, x[n-k'] \qquad (1)$$

Επίσης, $\sum_{n=0}^{N-1+p} x[n-k]\, x[n-k'] \overset{n'=n-k}{=} \sum_{n'=-k}^{N-1-k+p} x[n']\, x[n'+k-k']$

Όμως, $x[n'] = 0$ για $n' < 0$ και $x[n'+k-k'] = 0$ για $n'+k-k' > N$

Επομένως $n'_{min} = 0$ και $n'_{max} + k - k' = N-1 \Rightarrow n'_{max} = N-1-k+k' \leq N-1-k+p$

καθώς $k' \leq p$. Συνεπώς, $\sum_{n'=-k}^{N-1-k+p} x[n']\, x[n'+k-k'] = \sum_{n'=0}^{N-1-k+k'} x[n']\, x[n'+(k-k')] =$

$= \sum_{n'=0}^{N-1-(k-k')} x[n']\, x[n'+(k-k')] = R_x(k-k') = R_x(k'-k)$ λόγω συμμετρίας, καθώς μπορούμε

να εναλλάξουμε τα k και k'. Επομένως, $E_x = \sum_{k=0}^{p} a_{xk} \sum_{k'=0}^{p} a_{xk'}\, R_x(|k-k'|) =$

$$= \sum_{k=0}^{p} a_{xk}\, \vec{a}_x \begin{bmatrix} R_x(|k|) \\ R_x(|k-1|) \\ \vdots \\ R_x(|k-p|) \end{bmatrix} = \vec{a}_x \sum_{k=0}^{p} a_{xk} \begin{bmatrix} R_x(|k|) \\ R_x(|k-1|) \\ \vdots \\ R_x(|k-p|) \end{bmatrix} = \vec{a}_x \begin{bmatrix} R_x(0) & R_x(1) & & R_x(p) \\ R_x(1) & R_x(0) & \cdots & R_x(p-1) \\ \vdots & \vdots & & \vdots \\ R_x(p) & R_x(p-1) & & R_x(0) \end{bmatrix} \begin{bmatrix} a_{x0} \\ a_{x1} \\ \vdots \\ a_{xp} \end{bmatrix} =$$

$= \vec{a}_x R_x \vec{a}_x^T \Rightarrow E_x = \vec{a}_x R_x \vec{a}_x^T$ όπως θέλαμε να δείξουμε.

2. Εργαζόμαστε με τον ίδιο τρόπο, χρησιμοποιώντας τους συντελεστές $\vec{a}_y$ και προκύπτει

ότι $E_{xy} = \vec{a}_y R_x \vec{a}_y^{\,T}$

3. Εξ ορισμού η ενέργεια του σφάλματος είναι ελάχιστη για τους συντελεστές

γραμμικής πρόβλεψης $\vec{a}_x$ και οποιοιδήποτε άλλοι συντελεστές θα αντιστοιχούν σε ίση

ή μεγαλύτερη ενέργεια σφάλματος, δηλαδή $E_{xy} \geq E_x \Rightarrow \frac{E_{xy}}{E_x} \geq 1$

Άσκηση 2.2

Για $t=1$, έχουμε ότι $\delta_1(i) = P(q_1 = i, O_1 | \lambda) = P(q_1 = i | \lambda)\, P(O_1 | q_1 = i, \lambda)$

$$\Rightarrow \delta_1(i) = \pi_i\, P(O_1 | q_1 = i, \lambda)$$

Για $t > 1$, έχουμε ότι $\delta_t(i) = \max\limits_{q_1, \dots q_{t-1}} P(q_1 q_2, \dots q_t = i, O_1 O_2 \dots O_t | \lambda) =$

$$= \max_{q_1, \dots q_{t-2}, j} P(q_1 q_2 \dots q_{t-1} = j, O_1 O_2 \dots O_{t-1} | \lambda)\, P(q_t = i, O_t | \lambda, q_1 q_2 \dots q_{t-1} = j, O_1 \dots O_{t-1})$$

$$= \max_{q_1, \dots q_{t-2}, j} P(q_1 q_2 \dots q_{t-1} = j, O_1 O_2 \dots O_{t-1} | \lambda)\, P(q_t = i, O_t | q_{t-1} = j, \lambda)$$

$$= \max_{q_1, \dots q_{t-2}, j} P(q_1 q_2 \dots q_{t-1} = j, O_1 O_2 \dots O_{t-1} | \lambda)\, P(q_t = i | q_{t-1} = j, \lambda)\, P(O_t | q_t = i, q_{t-1} = j, \lambda)$$

$$= \max_{1 \le j \le 4} \delta_{t-1}(j)\, a_{ji}\, P(O_t | q_t = i, \lambda) \Rightarrow \delta_t(i) = P(O_t | q_t = i, \lambda) \max_{1 \le j \le 4} \delta_{t-1}(j)\, a_{ji}$$

Έχουμε λοιπόν ότι $\delta_1(1) = 0.25\, P(O_1 = U | q_1 = 1) = 0.25 \cdot 0.4 = 0.1$

$\delta_1(2) = 0.25\, P(O_1 = U | q_1 = 2) = 0.25 \cdot 0.3 = 0.075$

$\delta_1(3) = 0.25\, P(O_1 = U | q_1 = 3) = 0.25 \cdot 0.8 = 0.2$

$\delta_1(4) = 0.25\, P(O_1 = U | q_1 = 4) = 0.25 \cdot 0.75 = 0.1875$

$$\delta_2(1) = P(O_2 = V | q_2 = 1) \max\{\delta_1(1) a_{11}, \delta_1(2) a_{21}, \delta_1(3) a_{31}, \delta_1(4) a_{41}\} =$$

$$= 0.6 \max\{0.1 \cdot 0.15, 0.075 \cdot 0.3, 0.2 \cdot 0.3, 0.1875 \cdot 0.25\} =$$

$$= 0.6 \max\{0.015, 0.0225, 0.06, 0.046875\} = 0.6 \cdot 0.06 = 0.036$$

Θεωρούμε επίσης την συνάρτηση $\psi_t(i) = \underset{1\le j\le 4}{\operatorname{argmax}} \; \delta_{t-1}(j)\, a_{ji}$, $t>1$ ώστε να κρατάμε την βέλτιστη κατάσταση j που μας οδήγησε στην κατάσταση i. Έχουμε $\psi_2(1)=3$.

$$\delta_2(2) = P(O_2 = V \mid q_2 = 2)\, \max\{\delta_1(1)a_{12},\, \delta_1(2)a_{22},\, \delta_1(3)a_{32},\, \delta_1(4)a_{42}\} =$$

$$= 0.7 \max\{0.1\cdot 0.3,\ 0.075\cdot 0.15,\ 0.2\cdot 0.25,\ 0.1875\cdot 0.3\} =$$

$$= 0.7 \max\{0.03,\ 0.01125,\ 0.05,\ 0.05625\} = 0.7\cdot 0.05625 = 0.039375,\quad \psi_2(2)=4$$

Εργαζόμαστε κατ' αυτό τον τρόπο και προκύπτει τελικά ο ακόλουθος πίνακας:

$\delta_t(i)$ — ΧΡΟΝΟΣ t

ΚΑΤΑΣΤΑΣΗ	1	2	3	4	5	6	7	8	9	10
1	0.1	0.036 (3)	0.0047 (2)	0.0017 (3)	0.00029 (2)	4.28e-5 (2)	1.54e-5 (3)	1.8e-6 (2)	6.6e-7 (3)	7.77e-8 (2)
2	0.075	0.039 (4)	0.00324 (1)	0.0016 (3)	0.00035 (1)	2.67e-5 (4)	1.50e-5 (3)	1.38e-6 (1)	6.48e-7 (3)	5.99e-8 (1)
3	0.2	0.009 (4)	0.00945 (2)	0.00036 (4)	0.00010 (1)	8.57e-5 (2)	3.34e-6 (4)	3.7e-6 (1)	1.44e-7 (4)	1.599e-7 (1)
4	0.1875	0.015 (3)	0.00738 (2)	0.0007 (3)	0.00010 (1)	6.69e-5 (2)	6.42e-6 (3)	2.89e-6 (1)	2.77e-7 (3)	1.249e-7 (1)

όπου στις παρενθέσεις αναγράφεται η αντίστοιχη τιμή της συνάρτησης ψ.

2. Έχουμε $q_{10}^* = \underset{1\le j\le 4}{\operatorname{argmax}}\, \delta_{10}(j) = 3$, $q_9^* = \psi_{10}(q_{10}^*) = 1$, $q_8^* = \psi_9(q_9^*) = 3$, $q_7^* = \psi_8(q_8^*) = 1$

$q_6^* = \psi_7(q_7^*) = 3$, $q_5^* = \psi_6(q_6^*) = 2$, $q_4^* = \psi_5(q_5^*) = 1$, $q_3^* = \psi_4(q_4^*) = 3$, $q_2^* = \psi_3(q_3^*) = 2$

$q_1^* = \psi_2(q_2^*) = 4$. Επομένως, $Q^* = (4,2,3,1,2,3,1,3,1,3)$

3. $P^*(Q^*, O \mid \lambda) = \max\limits_{1\le j\le 4} \delta_{10}(j) \approx 1.599\cdot 10^{-7} \Rightarrow P^*(Q^*, O\mid\lambda) \approx 1.599\cdot 10^{-7}$

Άσκηση 2.6

α. Έχουμε ότι $\dfrac{\partial y}{\partial x} = \dfrac{\partial \sigma(Wx)}{\partial (Wx)} \dfrac{\partial (Wx)}{\partial x} = \dfrac{\partial \sigma(u)}{\partial u} W$ όπου $u = Wx \in \mathbb{R}^n$

Επίσης $\dfrac{\partial \sigma(u)}{\partial u} = \begin{bmatrix} \frac{\partial \sigma(u_1)}{\partial u_1} & \frac{\partial \sigma(u_1)}{\partial u_2} & \cdots & \frac{\partial \sigma(u_1)}{\partial u_n} \\ \vdots & & \ddots & \\ \frac{\partial \sigma(u_n)}{\partial u_1} & \frac{\partial \sigma(u_n)}{\partial u_2} & & \frac{\partial \sigma(u_n)}{\partial u_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial \sigma(u_1)}{\partial u_1} & 0 & \cdots & 0 \\ & & & \\ 0 & 0 & \cdots & \frac{\partial \sigma(u_n)}{\partial u_n} \end{bmatrix} = \begin{bmatrix} \sigma'(u_1) & 0 & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & \sigma'(u_n) \end{bmatrix}$

Επομένως, $\dfrac{\partial \sigma(u)}{\partial u} = diag(\sigma'(u))$ και $\dfrac{\partial y}{\partial x} = diag(\sigma'(Wx))\, W$

β. Δίνεται ότι $L = \sum_{t=0}^{T} L_t$. Κάθε επιμέρους χρονικό σφάλμα L_t εξαρτάται από την κατάσταση h_t, επομένως $\dfrac{\partial L_t}{\partial W} = \dfrac{\partial L_t}{\partial h_t} \dfrac{\partial h_t}{\partial W}$. Κάθε κατάσταση h_t εξαρτάται από τις προηγούμενες, και έτσι $\dfrac{\partial h_t}{\partial W} = \sum_{k=0}^{t} \dfrac{\partial h_t}{\partial h_k} \dfrac{\partial h_k}{\partial W} = \dfrac{\partial h_t}{\partial h_0} \cancel{\dfrac{\partial h_0}{\partial W}}^{0} + \sum_{k=1}^{t} \dfrac{\partial h_t}{\partial h_k} \dfrac{\partial h_k}{\partial W}$ Έχουμε λοιπόν ότι $\dfrac{\partial L_t}{\partial W} = \dfrac{\partial L_t}{\partial h_t} \sum_{k=1}^{t} \dfrac{\partial h_t}{\partial h_k} \dfrac{\partial h_k}{\partial W} = \sum_{k=1}^{t} \dfrac{\partial L_t}{\partial h_t} \dfrac{\partial h_t}{\partial h_k} \dfrac{\partial h_k}{\partial W}$. Επομένως, προκύπτει ότι

$$\frac{\partial L}{\partial W} = \sum_{t=0}^{T} \frac{\partial L_t}{\partial W} \Rightarrow \frac{\partial L}{\partial W} = \sum_{t=0}^{T} \sum_{k=1}^{t} \frac{\partial L_t}{\partial h_t} \frac{\partial h_t}{\partial h_k} \frac{\partial h_k}{\partial W}$$

γ. Για $T=3$, έχουμε ότι $\dfrac{\partial L}{\partial W} = \sum_{t=0}^{T} \dfrac{\partial L_t}{\partial h_t} \sum_{k=1}^{t} \dfrac{\partial h_t}{\partial h_k} \dfrac{\partial h_k}{\partial W} = \dfrac{\partial L_1}{\partial h_1} \dfrac{\partial h_1}{\partial W} + \dfrac{\partial L_2}{\partial h_2} \left(\dfrac{\partial h_2}{\partial h_1} \dfrac{\partial h_1}{\partial W} + \dfrac{\partial h_2}{\partial W} \right) +$

$$+ \frac{\partial L_3}{\partial h_3} \left(\frac{\partial h_3}{\partial h_1} \frac{\partial h_1}{\partial W} + \frac{\partial h_3}{\partial h_2} \frac{\partial h_2}{\partial W} + \frac{\partial h_3}{\partial W} \right) = \frac{\partial L_1}{\partial h_1} \frac{\partial h_1}{\partial W} + \frac{\partial L_2}{\partial h_2} \left(\frac{\partial h_2}{\partial h_1} \frac{\partial h_1}{\partial W} + \frac{\partial h_2}{\partial h_1} \frac{\partial h_1}{\partial W} \right) +$$

$$+ \frac{\partial L_3}{\partial h_3} \left(\frac{\partial h_3}{\partial h_2} \frac{\partial h_2}{\partial h_1} \frac{\partial h_1}{\partial W} + \frac{\partial h_3}{\partial h_2} \frac{\partial h_2}{\partial h_1} \frac{\partial h_1}{\partial W} + \frac{\partial h_3}{\partial h_2} \frac{\partial h_2}{\partial h_1} \frac{\partial h_1}{\partial W} \right) = \frac{\partial L_1}{\partial h_1} \frac{\partial h_1}{\partial W} + 2 \frac{\partial L_2}{\partial h_2} \frac{\partial h_2}{\partial h_1} \frac{\partial h_1}{\partial W} + 3 \frac{\partial L_3}{\partial h_3} \frac{\partial h_3}{\partial h_2} \frac{\partial h_2}{\partial h_1} \frac{\partial h_1}{\partial W}$$

Κατά την εκτέλεση backpropagation σε n+1 το πλήθος χρονικές στιγμές, χρειάζεται ο υπολογισμός του γινομένου $\frac{\partial h_{n+1}}{\partial h_n} \cdot \frac{\partial h_{n-1}}{\partial h_{n-2}} \cdots \frac{\partial h_2}{\partial h_1}$ προκειμένου να βρεθεί το $\frac{\partial h_n}{\partial W}$

Έχουμε ότι $\frac{\partial h_i}{\partial h_{i-1}} = \frac{\partial \sigma(W h_{i-1} + U x_t)}{\partial h_{i-1}} = \text{diag}\, \sigma'(W h_{i-1} + U x_t)\, W$ Επομένως χρειάζεται να πολ/σιάσουμε τον $\text{diag}\, \sigma'(W h_{i-1} + U x_t) W$ με τον εαυτό του n φορές.

δ. Έχουμε ότι $M = Q \Lambda Q^{-1}$. Τότε $\prod_{i=1}^{n} M = \underbrace{(Q\Lambda Q^{-1})(Q\Lambda Q^{-1}) \ldots (Q\Lambda Q^{-1})}_{n \text{ φορές}} =$

$= Q \Lambda^n Q^{-1} \Rightarrow \quad M^n = Q \Lambda^n Q^{-1}$.

ε. Είναι $W^{30} = Q \Lambda^{30} Q^{-1} = \begin{pmatrix} -0.8 & 0.6 \\ 0.6 & 0.8 \end{pmatrix} \begin{pmatrix} 0.3^{30} & 0 \\ 0 & 0.55^{30} \end{pmatrix} \begin{pmatrix} -0.8 & 0.6 \\ 0.6 & 0.8 \end{pmatrix}$

Τελικά, $W^{30} \simeq \begin{pmatrix} 5.85 \cdot 10^{-9} & 7.8 \cdot 10^{-9} \\ 7.8 \cdot 10^{-9} & 1.04 \cdot 10^{-8} \end{pmatrix}$. Βλέπουμε ότι οι τιμές του W είναι πολύ κοντά στο μηδέν όταν οι ιδιοτιμές έχουν μέτρο μικρότερο της μονάδας.

Αυτό μπορεί να δημιουργήσει vanishing gradients, καθώς τα gradients που θα πολ/στουν με W^{30} θα είναι πολύ κοντά στο μηδέν. Στην γενική περίπτωση έχουμε τις εξής περιπτώσεις:

i. Εάν $|\lambda_i| < 1$ τότε $\lim_{n\to\infty} |\lambda_i|^n = 0$. Τα στοιχεία του W^n τείνουν στο μηδέν και υπάρχει το πρόβλημα των vanishing gradients.

ii. Εάν $|\lambda_i| > 1$ τότε $\lim_{n\to\infty} |\lambda_i|^n = \infty$. Τα στοιχεία του W^n τείνουν στο άπειρο και υπάρχει το πρόβλημα των exploding gradients.

iii. Εάν $|\lambda_i| = 1$ τότε $\lim_{n\to\infty} |\lambda_i|^n = 1$. Τα στοιχεία του W^n δεν παρουσιάζουν μεταβολή στο μέγεθος τους.

στ. Η πύλη f_t (forget gate) αφορά στο κατά πόσο θα "λησμονηθεί" η προηγούμενη κατάσταση του cell state. Αν έχει τιμές κοντά στο μηδέν, δεν διατηρείται η προηγούμενη κατάσταση του cell state. Η πύλη i_t (input gate) καθορίζει πόσο θα ληφθεί υπόψη το υποψήφιο cell state $\tilde{C}_t$ κατά την ανανέωση του cell state. Η πύλη o_t (output gate) καθορίζει την επιρροή του cell state στην κρυφή κατάσταση.

ζ. Οι πύλες f_t, i_t, o_t έχουν μη αρνητικές τιμές στο διάστημα $[0, 1]$ λόγω του ορισμού τους με σιγμοειδή συνάρτηση.

η. Ναι, καθώς το C_t εξαρτάται από τα $f_t, i_t, \tilde{C}_t$ που με την σειρά τους εξαρτώνται από το προηγούμενο cell state.

θ. Έχουμε ότι $C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t$. Για $f_t = 1$ και $i_t = 0$, προκύπτει ότι $C_t = 1 \odot C_{t-1} + 0 \odot \tilde{C}_t \Rightarrow C_t = C_{t-1}$. Επομένως, $\frac{\partial C_t}{\partial C_k} = \prod_{i=k+1}^{t} \frac{\partial C_i}{\partial C_{i-1}} = \prod_{i=k+1}^{t} 1 = 1 \Rightarrow$

$\frac{\partial C_t}{\partial C_k} = 1$.

ι. Έχουμε ότι

$$\frac{\partial C_t}{\partial C_{t-1}} = \frac{\partial\left(f_t \odot C_{t-1} + i_t \odot \tilde{C}_t\right)}{\partial C_{t-1}} = \frac{\partial(f_t \odot C_{t-1})}{\partial C_{t-1}} + \frac{\partial(i_t \odot \tilde{C}_t)}{\partial C_{t-1}} =$$

$$= \frac{\partial f_t}{\partial C_{t-1}} \odot C_{t-1} + f_t \odot \frac{\partial C_{t-1}}{\partial C_{t-1}} + \frac{\partial i_t}{\partial C_{t-1}} \odot \tilde{C}_t + i_t \odot \frac{\partial \tilde{C}_t}{\partial C_{t-1}} =$$

$$= \frac{\partial f_t}{\partial C_{t-1}} \odot C_{t-1} + f_t + \frac{\partial i_t}{\partial C_{t-1}} \odot \tilde{C}_t + i_t \odot \frac{\partial \tilde{C}_t}{\partial C_{t-1}}$$

Υπολογίζουμε ξεχωριστά τις επιμέρους παραγώγους.

$$\frac{\partial f_t}{\partial C_{t-1}} = \frac{\partial \sigma(W_f h_{t-1} + U_f x_t)}{\partial C_{t-1}} = \frac{\partial \sigma(W_f h_{t-1} + U_f x_t)}{\partial h_{t-1}} \frac{\partial h_{t-1}}{\partial C_{t-1}} = \sigma'(W_f h_{t-1} + U_f x_t) \cdot W_f \frac{\partial h_{t-1}}{\partial C_{t-1}}.$$

Επίσης,

$$\frac{\partial h_{t-1}}{\partial C_{t-1}} = \frac{\partial(o_{t-1} \odot \tanh(C_{t-1}))}{\partial C_{t-1}} = \cancelto{0}{\frac{\partial o_{t-1}}{\partial C_{t-1}}} \odot \tanh(C_{t-1}) +$$

$$+ \ o_{t-1} \odot \frac{\partial \tanh(C_{t-1})}{\partial C_{t-1}} = o_{t-1} \odot \tanh'(C_{t-1}) \Rightarrow \frac{\partial h_{t-1}}{\partial C_{t-1}} = o_{t-1} \odot \tanh'(C_{t-1}).$$

Συνεπώς, $\frac{\partial f_t}{\partial C_{t-1}} = \sigma'(W_f h_{t-1} + U_f x_t) W_f \underbrace{\left(o_{t-1} \odot \tanh'(C_{t-1}) \right)}_{\delta} = \sigma'(W_f h_{t-1} + U_f x_t) W_f \delta$

Ακόμα, $\frac{\partial i_t}{\partial C_{t-1}} = \frac{\partial \sigma(W_i h_{t-1} + U_i x_t)}{\partial C_{t-1}} = \frac{\partial \sigma(W_i h_{t-1} + U_i x_t)}{\partial h_{t-1}} \frac{\partial h_{t-1}}{\partial C_{t-1}} = \sigma'(W_i h_{t-1} + U_i x_t) W_i \delta$

$\Rightarrow \frac{\partial i_t}{\partial C_{t-1}} = \sigma'(W_i h_{t-1} + U_i x_t) W_i \delta$. Επίσης, $\frac{\partial \tilde{C}_t}{\partial C_{t-1}} = \frac{\partial \tanh(W_c h_{t-1} + U_c x_t)}{\partial C_{t-1}} =$

$= \frac{\partial \tanh(W_c h_{t-1} + U_c x_t)}{\partial h_{t-1}} \frac{\partial h_{t-1}}{\partial C_{t-1}} = \tanh'(W_c h_{t-1} + U_c x_t) W_c \delta \Rightarrow$

$\frac{\partial i_t}{\partial C_{t-1}} = \tanh'(W_c h_{t-1} + U_c x_t) W_c \delta$. Συνολικά, έχουμε ότι $\frac{\partial C_t}{\partial C_{t-1}} = \frac{\partial f_t}{\partial C_{t-1}} \odot C_{t-1} +$

$+ f_t + \frac{\partial i_t}{\partial C_{t-1}} \odot \tilde{C}_t + i_t \odot \frac{\partial \tilde{C}_t}{\partial C_{t-1}} = \sigma'(W_f h_{t-1} + U_f x_t) W_f (\delta \odot C_{t-1}) + f_t +$

$+ \ \sigma'(W_i h_{t-1} + U_i x_t) W_i (\delta \odot \tilde{C}_t) + i_t \odot \tanh'(W_c h_{t-1} + U_c x_t) W_c \delta \Rightarrow$

$$\frac{\partial C_t}{\partial C_{t-1}} = \sigma'(W_f h_{t-1} + U_f x_t) W_f \delta \odot C_{t-1} + f_t + \sigma'(W_i h_{t-1} + U_i x_t) W_i \delta \odot \tilde{C}_t + i_t \odot \tanh'(W_c h_{t-1} + U_c x_t) W_c \delta$$

Είναι καλύτερο να χρησιμοποιείται το cell state λόγω της μορφής της $\frac{\partial C_t}{\partial C_{t-1}}$, που δεν ευνοεί την εμφάνιση vanishing gradients σε αντίθεση με το hidden state των RNNs.