Q-Learning

Νευρο-Ασαφής Έλεγχος και Εφαρμογές, ΣΗΜΜΥ, ΕΜΠ

Νικόλαος Τσιλιβής, ΑΜ: 03114078

Φεβρουάριος 2019

1 Σκοπός άσκησης

Σκοπός της άσκησης είναι η επίλυση ενός LQR προβλήματος βέλτιστου ελέγχου για δοσμένο σύστημα του οποίου η δυναμική θεωρείται, για τις προσομειώσεις, άγνωστη. Πιο συγκεκριμένα, θα βρεθεί ο βέλτιστος νόμος ελέγχου με χρήση της τεχνικής που είναι γνωστή με το όνομα Q learning.

2 Προδιαγραφές συστήματος

Το σύστημα διακριτού χρόνου περιγράφεται απ' τις εξισώσεις:

$$x_{k+1} = Ax_k + Bu_k,$$

όπου $A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$ και $B = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$, ενώ το προς ελαχιστοποίηση κριτήριο κόστους είναι:

$$J = \sum_{k=0}^{\infty} (x_k^T x_k + \rho u_k^2)$$

3 Αναλυτική εύρεση βέλτιστου κέρδους

Αρχικά, θα βρούμε τη τιμή του βέλτιστου κέρδους k για ελεγκτή της μορφής $u_k = kx_k$, επιλύοντας την εξίσωση Riccati του συστήματος:

$$P = I_{3\times 3} + A^T PA - A^T PB(\rho + B^T PB)^{-1} B^T PA \Leftrightarrow$$

$$\Leftrightarrow \begin{bmatrix} p_1 & p_{12} & p_{13} \\ p_{12} & p_2 & p_{23} \\ p_{13} & p_{23} & p_{33} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 + p_1 - \frac{p_{13}^2}{\rho + p_3} & p_{12} - \frac{p_{13}p_{23}}{\rho + p_3} \\ 0 & p_{12} - \frac{p_{13}p_{23}}{\rho + p_3} & 1 + p_2 - \frac{p_{23}^2}{\rho + p_3} \end{bmatrix} \Leftrightarrow$$

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

Επομένως, ο βέλτιστος νόμος ελέγχου είναι:

$$\begin{aligned} k &= -(I + B^T PB)^{-1} B^T PA = \\ &= -\frac{3}{\rho + 3} \begin{bmatrix} 0 & 0 & 0 \end{bmatrix} = \\ &= \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}. \end{aligned}$$

4 Κατάστρωση Q learning

Γνωρίζουμε ότι η τεχνική Q learning αποτελεί μια επαναληπτική μέθοδο υπολογισμού του πίνακα

$$H = \begin{bmatrix} M + A^T PA & A^T PB \\ B^T PA & R + B^T PB \end{bmatrix},$$

μέσω της σχέσης:

$$Q_{i+1}(x_k, u_k) = J^*(x_k, u_k) - Q_i(x_k, u_k),$$

με

$$Q_i(x_k, u_k) = \begin{bmatrix} x_k^T & u_k^T \end{bmatrix} H_i \begin{bmatrix} x_k \\ u_k \end{bmatrix}$$

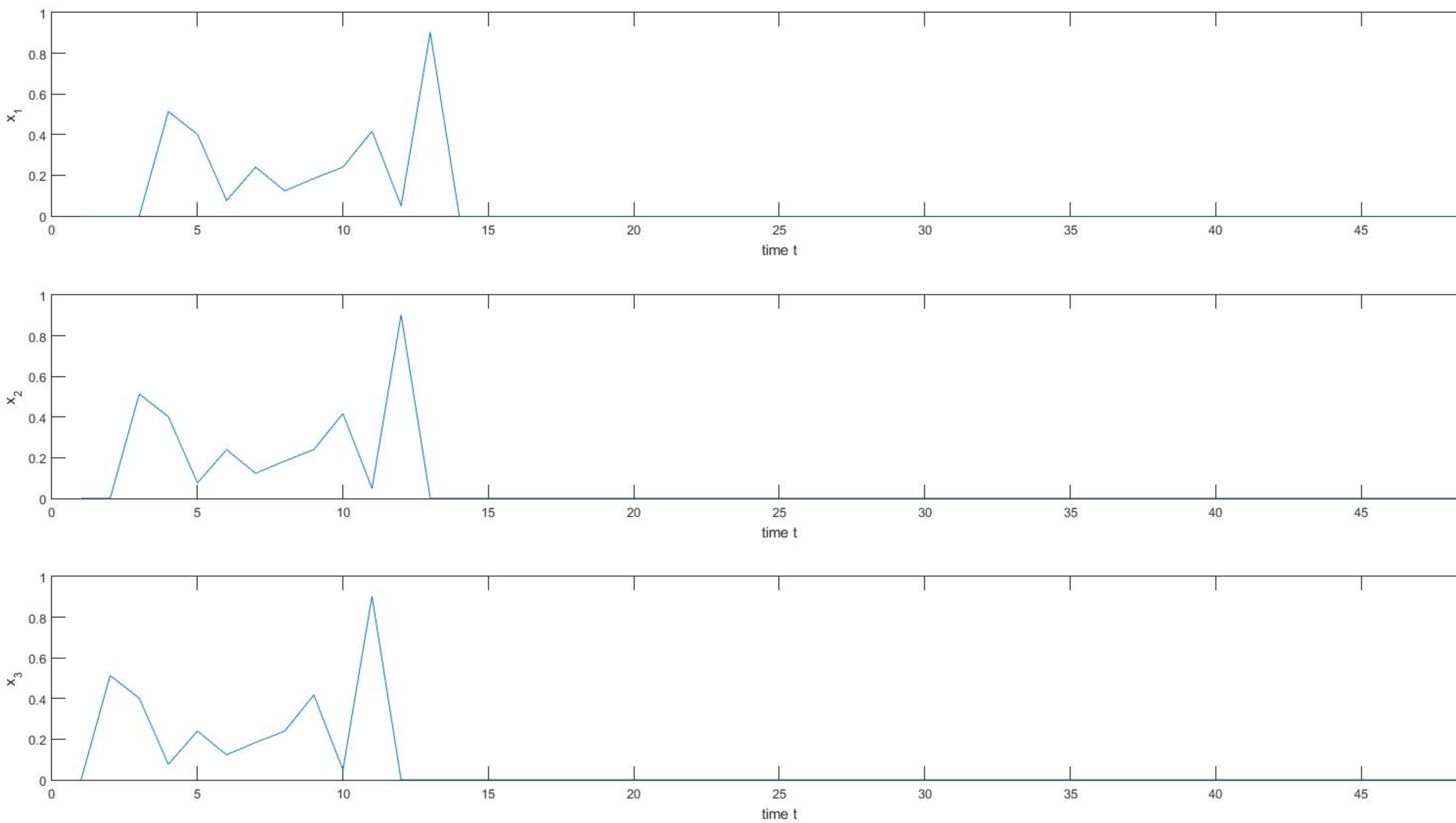
όπου $J^*(x_k, u_k) = x_k^T x_k + \rho u_k^2$ και το k ανήκει σε ένα αρκούντως μεγάλο διάστημα, έτσι ώστε το προκύπτον γραμμικό σύστημα να είναι ακριβώς επιλύσιμο. Με άλλα λόγια, παίρνουμε αρκετές μετρήσεις απ' το σύστημα μας, έτσι ώστε να αντιστρέφεται ο πίνακας, που η κάθε γραμμή του μοιάζει ως εξής:

$$Z[i,:] = \begin{bmatrix} x_1^2(i) & x_2^2(i) & x_3^2(i) & x_1x_2(i) & x_2x_3(i) & x_1x_3(i) & u^2(i) & 2ux_1(i) & 2ux_2(i) & 2ux_3(i) \end{bmatrix}$$

Στον κώδικα μας στο Matlab παίρνουμε 10 αυθαίρετες τιμές στο διάστημα $[0, 1]$ για την είσοδο, ελέγχοντας πως αυτές δίνουν full rank στον, πλέον, 10×10 πίνακα Z. Σημειώνουμε, ότι στο πρόβλημα που μελετάμε ο πίνακας H έχει διάσταση 4×4, ενώ αναγκαία για την επίλυση του συστήματος είναι η διαπίστωση πως ο $Z[1:3, 1:3]$ είναι συμμετρικός, μιας και ο πίνακας κόστους για το x_k (ο $I_{3\times3}$) είναι προφανώς συμμετρικός.

5 Κώδικας και αποτελέσματα

Τρέχουμε τον κώδικα του Matlab που βρίσκεται στο συνημμένο αρχείο και βλέπουμε πως υπολογίζει, ταχύτατα, την βέλτιστη τιμή του κέρδους (το 0 που βρίσκει ο αλγόριθμος, κάποιες φορές απεικονίζεται ως $e - 015$ κατά την εκτέλεση του κώδικα). Φυσικά, όπως διαπιστώσαμε και απ' την θεωρητική ανάλυση, η τιμή του ρ δεν επηρεάζει την τελική τιμή του νόμο ελέγχου. Στη συνέχεια, δείχνουμε και τις ζητούμενες γραφικές παραστάσεις του state (η είσοδος γυσικά πάει στο 0):



Σχήμα 1: State συστήματος