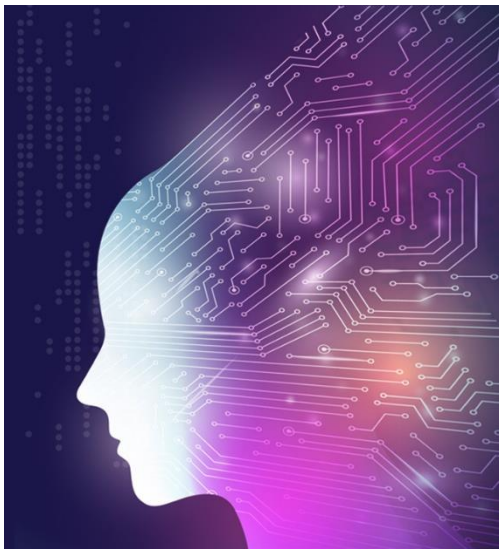
ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

Σχολή Χημικών Μηχανικών

Τομέας Ανάλυσης, Σχεδιασμού και Ανάπτυξης Διεργασιών και Συστημάτων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Κολυνδρίνη Βασιλική Μαρία



Ανάπτυξη συστημάτων αυτόματης ρύθμισης με τεχνολογίες βαθιάς ενισχυτικής μάθησης

Επιβλέπων καθηγητής: Χ. Σαρίμβεης

Αθήνα, Ιούνιος 2023

Ευχαριστίες

Η παρούσα διπλωματική εργασία εκπονήθηκε στη Μονάδα Αυτόματης Ρύθμισης και Πληροφορικής της Σχολής Χημικών Μηχανικών του Εθνικού Μετσόβιου Πολυτεχνείου (Ε.Μ.Π.) κατά την περίοδο Φεβρουαρίου 2023 έως τον Ιούνιο 2023.

Αρχικά θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα της διπλωματικής εργασίας κ. Χαράλαμπο Σαρίμβεη, Καθηγητή Ε.Μ.Π., για την εμπιστοσύνη που μου έδειξε αναθέτοντάς μου το συγκεκριμένο θέμα καθώς και για την υπομονή, την υποστήριξη και την καθοδήγησή του κατά την διάρκεια εκπόνησης της διπλωματικής αυτής εργασίας. Η παρουσία του και η επαγγελματική του εμπειρία με ενέπνευσαν και με βοήθησαν να ξεπεράσω δυσκολίες και προκλήσεις.

Επίσης, είμαι πολύ ευγνώμων για την ευκαιρία που είχα να συνεργαστώ με την υποψήφια διδάκτορα Αργυρή Καρδαμάκη που με βοήθησε να αναπτύξω σφαιρική κατανόηση του θέματος και να προχωρήσω στην επίλυση των προβλημάτων που αντιμετώπισα. Η συνεργασία μας ήταν πολύτιμη και ουσιαστική, και θα ήθελα να εκφράσω την βαθιά ευγνωμοσύνη για την αφιέρωση και την εμπειρογνωμοσύνη που μου παρείχε.

Ευχαριστώ, εν συνεχεία, τον κ. Φίλιππο Δογάνη για την ενασχόλησή του με την διπλωματική αυτή εργασία και την συνεισφορά του στην διαλεύκανση αποριών.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένεια μου αλλά και τους φίλους μου και συμφοιτητές, που με βοήθησαν σε όλο το ταξίδι μου κατά την διάρκεια της φοίτησης μου. Ανάμεσα σε αυτούς ξεχώρισε ο Ηλιάκης Ευστάθιος, η Βασιλική Αγγάνη και ο Αλέξανδρος Γιαννακόπουλος των οποίων η παρουσία ήταν ζωτικής σημασίας για όλη την πορεία μου ως αυτό το στάδιο.

Περιεχόμενα

Περίληψη

Η παρούσα διπλωματική εργασία επικεντρώνεται στην μελέτη και τον σχεδιασμό ενός συστήματος αυτόματου ελέγχου για έναν αντιδραστήρα με τη χρήση παραδοσιακών μεθόδων ρύθμισης και τεχνολογιών Ενισχυτικής Μάθησης (Reinforcement Learning, RL). Το σύστημα προς μελέτη αποτελείται από έναν αντιδραστήρα συνεχούς λειτουργίας και πλήρους ανάμιξης (Continuous Stirred Tank Reactor, CSTR) με ρυθμιζόμενες μεταβλητές τη θερμοκρασία και τη συγκέντρωση του ρεύματος του προϊόντος και μεταβλητές εκ χειρισμού την παροχή της τροφοδοσίας και την θερμοκρασία του ψυκτικού. Στόχος είναι η ανάπτυξη μιας στρατηγικής ελέγχου που να βελτιστοποιεί την απόδοση του αντιδραστήρα καθοδηγώντας τις δύο μεταβλητές ελέγχου στις επιθυμητές τους τιμές, ελαχιστοποιώντας παράλληλα τις αποκλίσεις και τις διαταραχές. Για να επιτευχθεί αυτό, χρησιμοποιούνται δύο παραδοσιακοί ελεγκτές τύπου PI με αναλογικό και ολοκληρωτικό μέρος, ένας για τον έλεγχο της θερμοκρασίας και ένας για τον έλεγχο της συγκέντρωσης. Στο σύστημα ελέγχου ενσωματώνεται ένας RL πράκτορας που εκπαιδεύεται με χρήση κατάλληλου αλγορίθμου ενισχυτικής μάθησης για να προσδιορίσει την βέλτιστη στρατηγική ελέγχου. Ο RL πράκτορας αλληλεπιδρά συνεχώς με το περιβάλλον του αντιδραστήρα και αποκτάει εμπειρία μέσα από τις ενέργειες που κάνει και την επίδραση που έχουν στην κατάσταση του συστήματος. Η αξιολόγηση των αποφάσεων του πράκτορα γίνεται μέσω της συνάρτησης ανταμοιβής η οποία διαμορφώνεται κατάλληλα έτσι ώστε να αποτυπώνει τους στόχους της ρύθμισης. Μέσω ενός συστήματος επιβράβευσης και τιμωρίας, ο πράκτορας αναπροσαρμόζει την πολιτική του με στόχο να βελτιστοποίηση την απόδοσή του και την μεγιστοποίηση της μακροπρόθεσμης ανταμοιβής του. Στην παρούσα εργασία, μελετήθηκαν δύο ξεχωριστές περιπτώσεις αλγορίθμων, μια που να βελτιώνει την απόδοση όταν οι ρυθμιστές PI πετυχαίνουν τον στόχο τους και μια που να σταθεροποιεί το σύστημα στις επιθυμητές τιμές όταν οι PI οδηγούν σε ταλαντωτική συμπεριφορά. Τα αποτελέσματα δείχνουν πως η χρήση αλγορίθμου μηχανικής μάθησης με κατάλληλη αρχιτεκτονική και ορθά σχεδιασμένη συνάρτηση ανταμοιβής μπορεί να βελτιώσει σημαντικά την ρύθμιση του συστήματος σε σύγκριση με παραδοσιακές μεθόδους ελέγχου που αγνοούν τις αλληλεπιδράσεις

ανάμεσα στις μεταβλητές του συστήματος, μειώνοντας τον χρόνο αποκατάστασης και την υπέρβασης και εξαλείφοντας φαινόμενα ταλάντωσης.

Abstract

This diploma thesis focuses on the implementation and design of an automated control system for a reactor by combining traditional control methods with Reinforcement Learning (RL) approaches. The system consists of a Continuously Stirred Tank Reactor (CSTR) where the temperature and concentration of product stream are the controlled variables, and the feed flow rate and cooling temperature are the manipulated variables. The objective is to develop a control strategy that optimizes the CSTR's performance by driving the controlled variables to their desired setpoints while minimizing deviations and rejecting disturbances. To achieve this, two PI controllers are employed, one for temperature control and another for concentration control. An RL agent is integrated into the control system and trained to "learn" from the environment and find the optimal control policy. The RL agent interacts with the environment of the reactor and gains experience through its actions and their effect on the system. By continuously exploring and exploiting the environment of the reactor, it adapts its policy and learns to take control actions that maximize the predefined reward function, which reflects the desired behavior and objectives of the control system. In this study, two reward functions were designed in order to train two different agents, one that seeks to improve the control performance of the PIs and another one that aims to stabilize the system in steady states that appear to be unstable. Results show that a well-designed reward function can improve the performance of the control system, reduce deviations from setpoints and enhance overall system efficiency. The proposed control system aims to showcase the benefits of RL-based control strategies in complex and dynamic processes like CSTRs. The diploma project involves designing the RL algorithm, implementing the control system, training the RL agent, and evaluating its performance in comparison to traditional control approaches.

Ευρετήριο εικόνων

Ευρετήριο Διαγραμμάτων

Ευρετήριο Πινάκων

Κεφάλαιο 1. Εισαγωγή

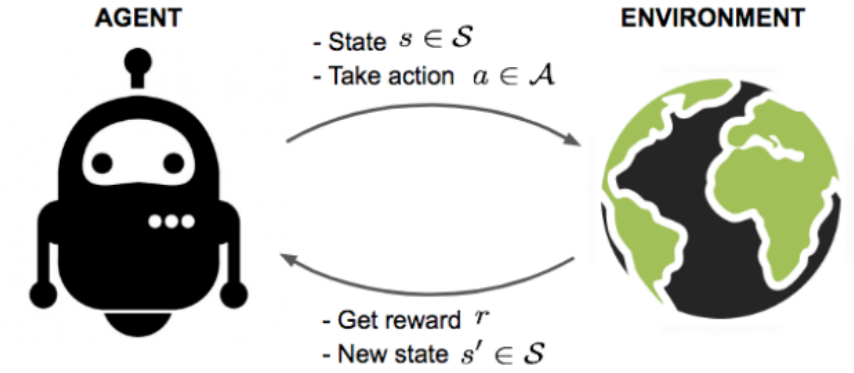
1.1. Εισαγωγή στην Ενισχυτική Μάθηση

Η μηχανική μάθηση αποτελεί ένα υποσύνολο της τεχνητής νοημοσύνης (AI) που ασχολείται με την ανάπτυξη αλγορίθμων με στόχο την εύρεση μοτίβων και συσχετίσεων σε σύνολα δεδομένων που μπορούν να υποβοηθήσουν την λήψη αποφάσεων. Στόχος της μηχανικής μάθησης είναι οι αλγόριθμοι να μπορούν να «εκπαιδεύσουν» μοντέλα από τα διαθέσιμα δεδομένα και να κάνουν προβλέψεις ή να λάβουν αποφάσεις ακόμα και αν δεν έχουν προγραμματιστεί ρητά για αυτό τον σκοπό. Ο τομέας της μηχανικής μάθησης γνωρίζει ιδιαίτερη άνθηση τα τελευταία χρόνια και διακρίνεται σε τρεις επιμέρους κλάδους:

1. **Επιβλεπόμενη μάθηση (supervised learning):** το σύστημα εκπαιδεύεται να προβλέπει εξόδους ή να συνδέει κομμάτια σε «παζλ» με βάση δεδομένα που του παρέχονται κατά τη διαδικασία και συνοδεύονται από αναλυτική περιγραφή (ετικέτες)
2. **Μη επιβλεπόμενη μάθηση (unsupervised learning):** ο αλγόριθμος μαθαίνει να ανακαλύπτει κρυπτογραφημένες δομές, πρότυπα ή μοτίβα σε δεδομένα χωρίς ετικέτες και προσπαθεί να τα κατηγοριοποιήσει, χωρίσει, αναλύσει, χωρίς να γνωρίζει περαιτέρω πληροφορίες όπως τις ετικέτες.
3. **Ενισχυτική μάθηση (reinforcement learning):** ο αλγόριθμος βασίζεται στην αλληλεπίδραση του με το περιβάλλον και μαθαίνει ποιες κινήσεις πρέπει να κάνει μέσα από ένα σύστημα επιβράβευσης και τιμωρίας. Ο κλάδος αυτός αποτελεί το επίκεντρο της παρούσας εργασίας και θα αναλυθεί εκτενώς παρακάτω.

Η Ενισχυτική Μάθηση είναι κλάδος της μηχανικής μάθησης ανάμεσα στην επιβλεπόμενη και την μη επιβλεπόμενη μηχανική μάθηση [1] και συχνά αναφέρεται και ως ημι-επιβλεπόμενη μηχανική μάθηση [2]. Στο διάγραμμα 1 αποτυπώνεται

μέσω μιας απλοποιημένης σχηματικής αναπαράστασης ο τρόπος λειτουργίας ενός συστήματος ενισχυτικής μάθησης.



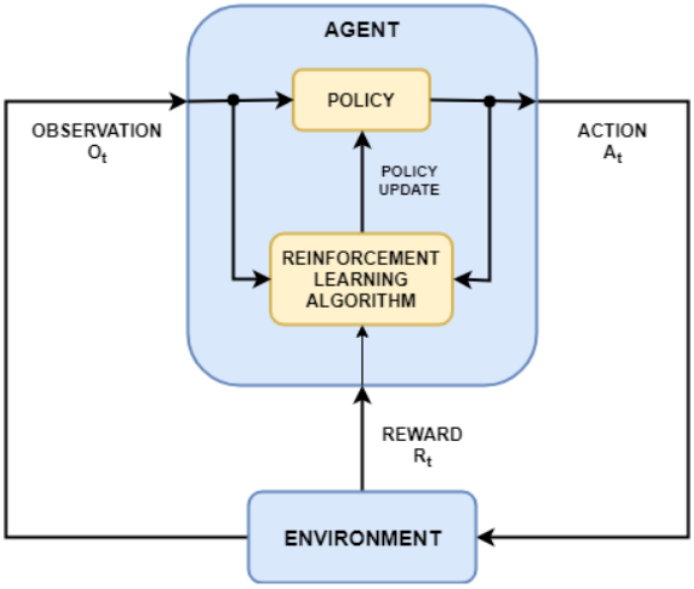
Διάγραμμα 1. Απλοποιημένη αναπαράσταση του συστήματος agent-environment.

Το περιβάλλον του συστήματος μοντελοποιείται από μια Μαρκοβιανή αλυσίδα. Αυτό σημαίνει ότι χαρακτηρίζεται από ένα πλήθος μεταβλητών, των οποίων η τρέχουσες τιμές αποτελούν την κατάσταση (state). Ο **πράκτορας** (agent) ενσωματώνει το σύνολο των αλγορίθμων που επεξεργάζονται τα σήματα από το περιβάλλον (observations και reward) και λαμβάνουν τις αποφάσεις (actions). Οι αποφάσεις που θα λάβει ο πράκτορας οδηγούν το σύστημα σε μια καινούρια κατάσταση. Είναι σημαντικό να τονιστεί μια σημαντική ιδιότητα των Μαρκοβιανών αλυσίδων, σύμφωνα με την οποία η καινούρια κατάσταση εξαρτάται μόνο από την προηγούμενη κατάσταση και την ενέργεια που θα αποφασίσει ο πράκτορας.

Στο διάγραμμα 2 [3] φαίνεται πιο ξεκάθαρα πως ο agent, ως πυρήνας λειτουργικότητας του συστήματος, αποτελείται στην ουσία από τον μηχανισμό της πολιτικής (policy) που αντιστοιχεί τις παρατηρήσεις O_t (observations) με τις ενέργειες A_t (actions) μέσω σύνθετων δομών όπως τα νευρωνικά δίκτυα και τον αλγόριθμο μάθησης που ανανεώνει τις παραμέτρους της πολιτικής με βάση τις ενέργειες, τις παρατηρήσεις και την συνάρτηση ανταμοιβής, R_t (reward).

Η επιβράβευση R_t είναι ο μοχλός κίνησης του συστήματος καθώς αποτελεί το μέτρο επιτυχίας της ενέργειας. Μικρή επιβράβευση σημαίνει ότι το σύστημα απέτυχε να προσεγγίσει τις απαιτούμενες προδιαγραφές και συνεπώς χρειάζεται αναπροσαρμογή της πολιτικής, ενώ μεγάλη επιβράβευση δηλώνει πως το σύστημα

έχει σχεδιαστεί σωστά. Η δουλειά του agent, τελικά, είναι να παραλαμβάνει τις παρατηρήσεις και την επιβράβευση και να στέλνει στο περιβάλλον (environment) τις ενέργειες με μία διαδικασία επιβράβευσης-τιμωρίας στο ρόλο της ανταμοιβής, η οποία οδηγεί στην επίτευξη της βέλτιστης δυνατής επιθυμητής συμπεριφοράς [3].



Διάγραμμα 2. Αναπαράσταση ενός γενικού σεναρίου Ενισχυτικής μάθησης.

Στο παρακάτω σχεδιάγραμμα [4] (διάγραμμα 3) αποτυπώνεται με την μορφή διαδοχικών βημάτων η συλλογιστική πορεία επίλυσης που ακολουθείται στην ενισχυτική μάθηση. Σε πρώτη φάση, είναι απαραίτητη η κατανόηση και η σαφής διατύπωση του προβλήματος που καλείται να αντιμετωπίσει ο agent. Αφού προσδιοριστεί ο σκοπός της εκπαίδευσης, διαμορφώνεται το περιβάλλον εργασίας (εν προκειμένω το σύστημα με τον CSTR) και καθορίζεται η συνάρτηση ανταμοιβής του πράκτορα. Η κατάστρωση μιας αποδοτικής συνάρτησης ανταμοιβής αποτελεί ένα από τα βασικότερα γρανάζια κατά την εκπαίδευση και είναι το σημείο στο οποίο θα δοθεί περισσότερη έμφαση στην παρούσα εργασία. Στο επόμενο στάδιο καθορίζεται η μορφή και το είδος του πράκτορα και ακολουθεί η εκπαίδευση και

επικύρωση του με βάση τα κριτήρια τερματισμού που κρίνουν και την τελική του απόδοση.

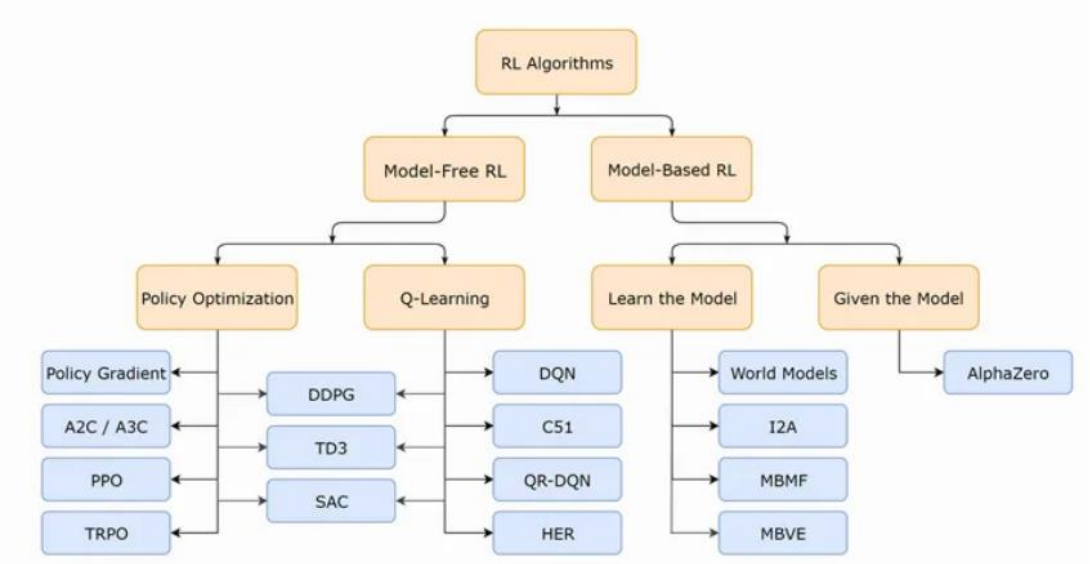
Διάγραμμα 3. Διάγραμμα ροής Ενισχυτικής μάθησης.

1.1.1. Κατηγορίες αλγορίθμων ενισχυτικής μάθησης

Όπως φαίνεται και στο διάγραμμα 4, οι αλγόριθμοι ενισχυτικής μάθησης χωρίζονται σε επιμέρους κατηγορίες ανάλογα με το είδος της προσέγγισης που χρησιμοποιούν. Η προσέγγιση μπορεί είτε να βασίζεται σε κάποιο μοντέλο (model-based) είτε όχι (model-free), δημιουργώντας έτσι τις δύο μεγάλες υποκατηγορίες αλγορίθμων του σχεδιαγράμματος.

Στην ενισχυτική μάθηση με μοντέλο ο πράκτορας προσπαθεί να κατανοήσει το περιβάλλον του και να διαμορφώσει ένα μοντέλο για αυτό με βάση τις αλληλεπιδράσεις μεταξύ του. Το μοντέλο αυτό του δίνει τη δυνατότητα μελλοντικά να μπορεί να προβλέψει την ανταμοιβή που αντιστοιχεί σε μια πράξη χωρίς να χρειαστεί να την εκτελέσει εκ των προτέρων. Από την άλλη πλευρά, οι αλγόριθμοι χωρίς μοντέλο επικεντρώνονται στις επιπτώσεις των πράξεων τους όπως αυτές αποτυπώνονται μέσω της εμπειρίας. Στις περιπτώσεις αυτές, οι αλγόριθμοι θα πρέπει να εκτελέσουν μια ενέργεια και να «μάθουν» από το περιβάλλον αν αυτή η ενέργεια είχε θετικές ή αρνητικές συνέπειες όπως αυτές ποσοτικοποιούνται μέσω της ανταμοιβής. Στην παρούσα διπλωματική εργασία, η ανάλυση θα επικεντρωθεί

γύρω από τους αλγορίθμους model-free και συγκεκριμένα εκείνους που εμπίπτουν στην κατηγορία Q-learning.



Διάγραμμα 4. Κατηγοριοποίηση RL αλγορίθμων.

Σύμφωνα με το παραπάνω διάγραμμα, οι model-free αλγόριθμοι ενισχυτικής μάθησης μπορεί να ακολουθούν δυο διαφορετικές προσεγγίσεις ή και συνδυασμό των δύο. Στην περίπτωση των **policy-based** αλγορίθμων η βελτιστοποίηση γίνεται μέσω άμεσης παρέμβασης στην πολιτική του αλγορίθμου που μεγιστοποιεί την ανταμοιβή. Αντίθετα, η **value-based** προσέγγιση δίνει έμφαση στην ποσοτικοποίηση της τιμής (value) που αντιστοιχεί σε κάθε ζεύγος κατάστασης και πράξης (state-action) την εκτίμηση της συνολικής ανταμοιβής για την οποία χρησιμοποιείται ο όρος «επιστροφή». Παρακάτω, αναλύονται πιο λεπτομερώς οι δυο διαφορετικές προσεγγίσεις.

1.1.1.1. Value-based προσέγγιση της ενισχυτικής μάθησης

Η value-based προσέγγιση είναι μέθοδος που βοηθά τον agent να βρει τη βέλτιστη πολιτική που πρέπει να ακολουθήσει μέσω του υπολογισμού της επίδρασης που έχει κάθε πράξη με βάση την τρέχουσα κατάσταση του συστήματος. Η ουσιαστική ιδέα

της μεθόδου βασίζεται στην εκτίμηση της αξίας διαφορετικών σταδίων ή σετ σταδίου-ενέργειας με σκοπό την λήψη απόφασης που μεγιστοποιεί την αθροιστική επιβράβευση. Για την περιγραφή της, χρησιμοποιείται μία «συνάρτηση αξίας», η (value-function), η οποία εκτιμά την αθροιστική επιβράβευση που θα μπορούσε να λάβει ο agent αν επιλέξει μία ενέργεια σε μία κατάσταση.

Υπάρχουν δύο είδη χρησιμοποιούμενων συναρτήσεων αξίας:

- **State value function, V(s) (V-function):** αφορά την κατάσταση s και υπολογίζει την αθροιστική επιβράβευση από το στάδιο s και μετά, ενώ αποτελεί ένδειξη της μακροπρόθεσμης αξίας σε ένα στάδιο.
- **Action value function, Q(s, a) (Q-function):** αφορά την αθροιστική επιβράβευση σε μία κατάσταση s αν συμβεί μία ενέργεια a, και αντίστοιχα αντιπροσωπεύει την μακροπρόθεσμη αξία σε μία κατάσταση s αν ληφθεί μια ενέργεια a.

Ο υπολογισμός τους μπορεί να γίνει αναδρομικά ως προς την αξία των επόμενων καταστάσεων μέσω της προσθήκης στην τωρινή τιμή, την τιμή των μελλοντικών καταστάσεων ή των σετ κατάστασης-ενέργειας με έναν παράγοντα μείωσης της σημασίας των επόμενων σταδίων για να υπάρχει ισορροπία ανάμεσα στα άμεσα και τα μελλοντικά στάδια. Για να γίνει αυτός ο υπολογισμός, χρησιμοποιείται η συνάρτηση **Bellman**.

1. State value function (V-function):

$$V(s) = R(s) + \gamma * \Sigma [P(s'|s,a) * V(s')] \text{ (1.1)}$$

2. Action value function (Q-function):

$$Q(s,a) = R(s,a) + \gamma * \Sigma [P(s'|s,a) * \Sigma [\pi(a'|s') * Q(s',a')]] \text{ (1.2)}$$

<u>Επεξήγηση συμβόλων:</u>

- R(s)=επιβράβευση στην κατάσταση s.
- R (s, a) =επιβράβευση στην κατάσταση s αν ληφθεί η ενέργεια a.
- P (s'|s, a) = η πιθανότητα το σύστημα να οδηγηθεί στην κατάσταση s’ αν ληφθεί η ενέργεια a στο στάδιο s.

- π(a'|s')= πολιτική υπεύθυνη για την λήψη της ενέργειας a’ στην κατάσταση s’.
- γ= ο συντελεστής που αναφέρθηκε προηγουμένως.

Τελικώς, η ενημέρωση των συναρτήσεων αυτών οδηγεί στην εύρεση της βέλτιστης πολιτικής που μεγιστοποιεί τη συσσωρευμένη ανταμοιβή.

1.1.1.2. Policy-based προσέγγιση της ενισχυτικής μάθησης

Οι μέθοδοι που είναι βασισμένες στην πολιτική χρησιμοποιούνται για να καθορίσουν της συμπεριφορά ενός πράκτορα και αντί να εκτιμούν την αξία κάθε κατάστασης ή σετ κατάστασης και ληφθείσας ενέργειας βελτιστοποιούν απευθείας την πολιτική τους. Στόχο αυτής της προσέγγισης αποτελεί η εύρεση της πολιτικής που μεγιστοποιεί κάποια ένδειξη της πορείας όπως η επιστροφή (η αθροιστική επιβράβευση σε βάθος χρόνου). Αυτές οι μέθοδοι έχουν αρκετά πλεονεκτήματα καθώς μπορούν να διαχειριστούν χώρους κατάστασης που χαρακτηρίζονται από συνέχεια και φυσικά να επιδοθούν στην εξερεύνηση του χώρου κατάστασης χρησιμοποιώντας στοχαστικές πολιτικές.

Η κατηγοριοποίηση των μεθόδων βασισμένων στην πολιτική μπορεί να γίνει στις εξής δύο κατηγορίες, στοχαστικές και ντετερμινιστικές.

- Οι **στοχαστικές** μέθοδοι λαμβάνουν υπόψη την αβεβαιότητα του περιβάλλοντος και εξάγουν μία πιθανοτική κατανομή για τις πιθανές ενέργειες σε κάθε στάδιο. Στη συνέχεια ο πράκτορας επιλέγει ενέργειες από την κατανομή για να εκμεταλλευτεί τη γνώση που έχει συλλέξει και να εξερευνήσει το περιβάλλον του.
- Οι **ντετερμινιστικές** μέθοδοι που αντιστοιχούν καταστάσεις σε ενέργειες χωρίς αβεβαιότητα.

Οι εξισώσεις που χαρακτηρίζουν αυτή την στοχαστική προσέγγιση είναι οι εξής:

Αν αναλογιστεί κανείς μία στοχαστική πολιτική π(a|s), ο απώτερος σκοπός είναι η μεγιστοποίηση της επιστροφής, δηλαδή η εκτιμώμενη αξία της ανταμοιβής, J(θ) όπου θ είναι οι παράμετροι της πολιτικής.

Η επιστροφή J(θ) δίνεται από την εξίσωση:

$$J(\theta) \ = \ E[R(t) \mid \theta] \ = \ \sum \pi(a(t)|s(t), \theta) \ R(t) \text{ (1.3)}$$

Η πτώση κλίσης παρουσιάζεται έτσι:

$$\nabla J(\theta) \ \approx \ E[R(t) \ \nabla log \ \pi(a(t)|s(t), \theta)] \text{ (1.4)}$$

Επεξήγηση συμβόλων:

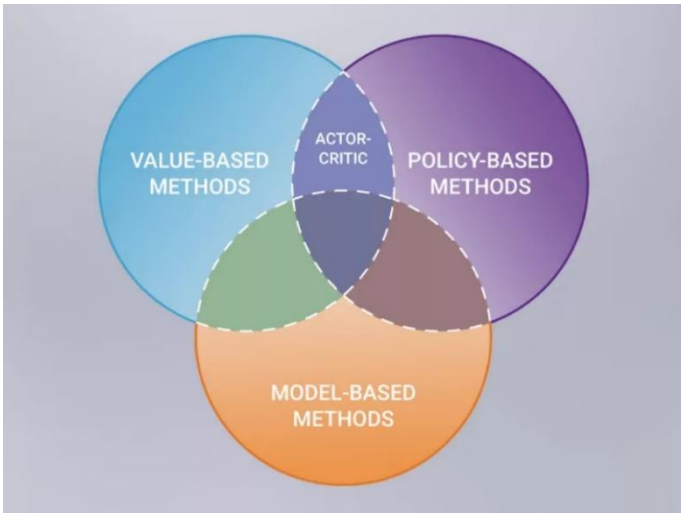
- R(t)= η αθροιστική ελαττωμένη επιβράβευση από χρόνο t και έπειτα.
- S(t)= η κατάσταση s την χρονική στιγμή t.
- A(t)= ενέργεια που λαμβάνεται την χρονική στιγμή t.

Η ανανέωση των παραμέτρων γίνεται με την μέθοδο Monte Carlo που περιλαμβάνει την κλίση

$$\nabla J(\theta): \theta \ \leftarrow \ \theta \ + \ \alpha \ \nabla J(\theta) \text{ (1.5)}$$

1.1.1.3. Συνδυασμός των προσεγγίσεων: Actor-Critic αλγόριθμοι

Στο παρακάτω σχεδιάγραμμα [5] αποδίδεται παραστατικά η διάκριση των αλγορίθμων RL στις επιμέρους κατηγορίες με βάση τα εσωτερικά μοντέλα πολιτικής, αξίας ή περιβάλλοντος όπως αναλύθηκε σε προηγούμενες παραγράφους. Στην τομή των δύο μεθόδων εντοπίζεται ο αλγόριθμος Actor-Critic ο οποίος αξιοποιεί και τις δύο προσεγγίσεις για την επίλυση του προβλήματος.



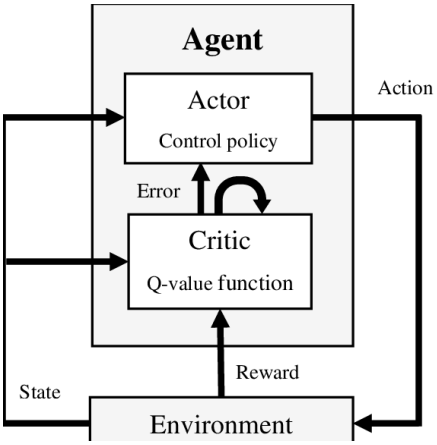
Εικόνα 1. Οι διαφορετικές προσεγγίσεις των αλγορίθμων RL

Η Actor-Critic προσέγγιση της ενισχυτικής μάθησης αποτελείται από δύο βασικούς πυλώνες, τον Actor και τον Critic. Ο Actor είναι policy-based αλγόριθμος και στοχεύει στον μηχανισμό λήψης αποφάσεων ενώ ο Critic, ως αλγόριθμος value-based, αποτελεί έναν μηχανισμό αξιολόγησης των αποφάσεων [1].

Για την λήψη αποφάσεων, ο Actor χρησιμοποιεί ως πολιτική ρύθμισης (control policy/ policy-based method) ένα νευρωνικό δίκτυο που αποτελείται από τα σώματα εισόδου και εξόδου (input, output layers) και τυχόν κρυμμένα στρώματα (hidden layers). Ως είσοδο λαμβάνει την παρούσα κατάσταση του συστήματος και σαν έξοδο έχει μία κατανομή πιθανών ενεργειών. Αυτό του δίνει τη δυνατότητα να διαλέγει την καλύτερη ενέργεια βασισμένο στην διδαχθείσα πολιτική και με τέλεια ισορροπία πάνω στην εκμετάλλευση-εξερεύνηση (exploration-exploitation). Η δομή του νευρωνικού δικτύου εξαρτάται από τις ανάγκες του προβλήματος σε κάθε περίπτωση. Ορισμένες δομές νευρωνικών δικτύων που ξεχωρίζουν είναι το Feedforward NN, Recurrent NN, Convolutional NN.

Ο Critic, που είναι το σύστημα αξιολόγησης των ενεργειών, παρέχει σε μορφή ανατροφοδότησης το σήμα ανταμοιβής. Για την αξιολόγηση χρησιμοποιεί μία συνάρτηση αξίας (value-function/ value-based method) η οποία επίσης είναι ένα νευρωνικό δίκτυο συχνά αποκαλούμενο ως value-function approximator, και υπολογίζει την αθροιστική επιβράβευση την οποία επιστρέφει στο σύστημα [6]. Το νευρωνικό δίκτυο λαμβάνει ως είσοδο την παρούσα κατάσταση του συστήματος και μέσω επιπλέον στρωμάτων κατανοεί τις αλληλεπιδράσεις και τις σχέσεις που

επικρατούν στο σύστημα. Σαν έξοδο έχει μία μεταβλητή που παρουσιάζει την εκτιμώμενη τιμή. Μία απλοποιημένη αναπαράσταση του αλγορίθμου Actor-Critic παρουσιάζεται στην εικόνα 2.



Εικόνα 2. Σχηματική αναπαράσταση του Actor-Critic αλγορίθμου

1.1.2. DQN και DDPG αλγόριθμοι

Στο σημείο αυτό, γίνεται πιο λεπτομερής αναφορά σε δυο χαρακτηριστικές περιπτώσεις αλγορίθμων μηχανικής μάθησης, του Deep Q-Network (DQN) και του Deep Deterministic Policy Gradient (DDPG). Ο DDPG θα χρησιμοποιεί ως αλγόριθμος επίλυσης στο πρόβλημα ρύθμισης που πραγματεύεται η παρούσα εργασία.

Ο αλγόριθμος **DQN** ανήκει στην κατηγορία των αλγορίθμων Q-learning, είναι value-based. Οι αρχικές προσπάθειες συνδυασμού της ενισχυτικής μάθησης με τα τεχνητά νευρωνικά δίκτυα (artificial neural networks) αποτύγχαναν λόγω διαφόρων αστοχιών όπως πολύ εύκολες αποκλίσεις, υπερεκτιμήσεις και μη επαρκή δεδομένα. Ο πρώτος επιτυχημένος συνδυασμός ήρθε με την εμφάνιση του Deep-Q Network που χρησιμοποιεί ένα συνελικτικό νευρωνικό δίκτυο, συνδυάζοντας την λογική των αλγορίθμων Q-learning (αλγόριθμοι που μαθαίνουν συναρτήσεις ενέργειας-τιμής) με βαθιά νευρωνικά δίκτυα. Είναι ιδανικό για διακριτούς χώρους δράσης.

Τα κύρια χαρακτηριστικά του DQN περιλαμβάνουν:

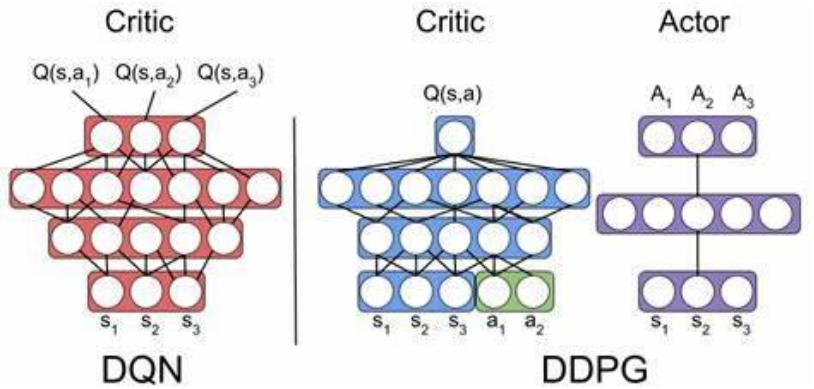
1. **Αναπαραγωγή εμπειριών**: χρησιμοποιεί μία αποθήκη δεδομένων (εμπειριών) που συλλέγονται κατά την εκτέλεση του αλγόριθμου. Έτσι διευκολύνεται η αποτελεσματικότερη χρήση των δεδομένων, καθώς διακόπτεται η ακολουθιακή συσχέτισή τους, με αποτέλεσμα να μειώνεται το σφάλμα που οφείλεται στην εκπαίδευση σε διαδοχικά δείγματα. Ταυτόχρονα επιταχύνεται η διαδικασία εκπαίδευσης καθώς δεν είναι απαραίτητη η δημιουργία νέων δεδομένων σε κάθε επανάληψη.
2. **Epsilon-greedy:** ο πράκτορας επιλέγει ενέργειες είτε ανάλογα με την τιμή Q είτε τυχαία με μία πιθανότητα epsilon ώστε να δίνεται η δυνατότητα για εξερεύνηση. Η τιμή epsilon (δηλαδή ουσιαστικά η πιθανότητα να επιλέγεται μια τυχαία ενέργεια αντί για αυτή που αντιστοιχεί στη μέγιστη τιμή Q), μπορεί να ελαττώνεται όσο προχωρούν οι επαναλήψεις του αλγόριθμου.

Ο αλγόριθμος **DDPG** ανήκει στην κατηγορία των Actor-Critic αλγορίθμων καθώς συνδυάζει Q-leaning με policy gradients. Αποτελεί ουσιαστικά επέκταση του DQN σε συνεχείς χώρους δράσης και είναι κατάλληλο για σχεδιασμό συστημάτων ελέγχου.

Τα κύρια χαρακτηριστικά του DDPG περιλαμβάνουν:

1. **Αρχιτεκτονική Actor-Critic**: διατηρεί δύο νευρωνικά δίκτυα, έναν actor και έναν critic που αναλύονται σε επόμενο κεφάλαιο.
2. **Ντετερμινιστική πολιτική κλίσης**: σε αντίθεση με παραδοσιακές μεθόδους που χρησιμοποιούν στοχαστική πολιτική, το DDPG μαθαίνει ντετερμινιστική πολιτική.
3. **Αποθήκευση εμπειριών**: παρόμοια με τον DQN αποθηκεύει εμπειρίες για την σταθεροποίηση της εκπαίδευσης.
4. **Δίκτυα στόχου**: χρησιμοποιεί δίκτυα στόχου τόσο για τον actor όσο και για τον critic. Τα δίκτυα αυτά ενημερώνονται σε πραγματικό χρόνο παράλληλα με την εκπαίδευση παρέχοντας πιο αξιόπιστες τιμές.

Στην παρακάτω εικόνα παρουσιάζεται μία σχηματική σύγκριση των δύο αλγορίθμων.

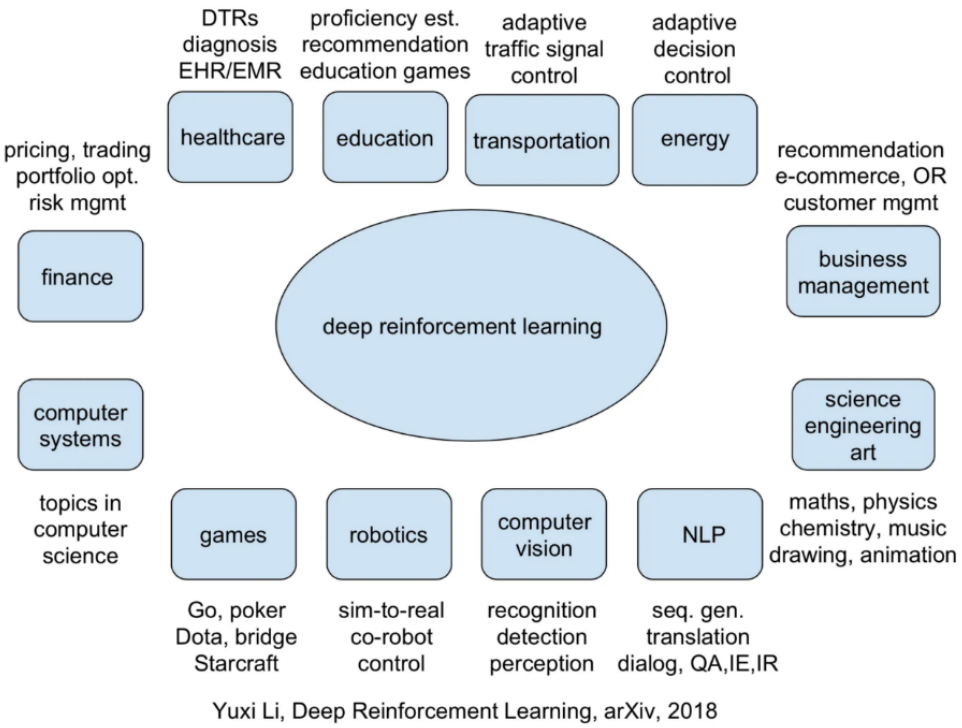


Εικόνα 3. Σχηματική σύγκριση των αλγορίθμων DQN και DDPG.

1.2. Εφαρμογές στη βιομηχανία

Η ενισχυτική μάθηση έχει εφαρμοστεί επιτυχία σε διάφορους κλάδους της βιομηχανίας και ξεχωρίζει για τις εφαρμογές της στον τομέα της παραγωγής, την ρομποτική, τον υγειονομικό κλάδο, τον οικονομικό τομέα [7], τα μέσα ενημέρωσης και τη διαφήμιση , την ενέργεια, την μεταφορά, την αυτόματη οδήγηση, την βιομηχανία παιχνιδιών αλλά και στον κατασκευαστικό και οικοδομικό τομέα με απώτερο σκοπό την βελτιστοποίηση αρχικά, τη ρύθμιση αργότερα και τέλος τον έλεγχο και τη διατήρηση [7].

Οι εφαρμογές παρουσιάζονται αναλυτικά στην εικόνα 4.



Εικόνα 4. Εφαρμογές της Ενισχυτικής Μάθησης

Αναλυτικότερα:

- **Παιχνιδοβιομηχανία (game playing):** η ενισχυτική μάθηση έχει καταφέρει πολύ σημαντικά πράγματα στον κλάδο της παιχνιδοβιομηχανίας. Σημαντικά παραδείγματα είναι το AlphaGo και το AlphaZero στα οποία χρησιμοποιήθηκαν RL τεχνικές για να εκπαιδευτεί ένας AI αλγόριθμος καταφέρνοντας να υπερνικηθούν οι ανθρώπινες δυνατότητες στα συγκεκριμένα προβλήματα.
- **Ρομποτική (robotics):** η ενισχυτική μάθηση έχει το ρόλο του εκπαιδευτή του ρομπότ, με απώτερο σκοπό να επιτύχει τους στόχους του και να μάθει πολιτικές ελέγχου μέσα από αλληλεπίδραση με το περιβάλλον του.
- **Αυτόματα οχήματα (autonomous vehicles):** η ενισχυτική μάθηση μπορεί να ενσωματωθεί στον εγκέφαλο αυτοκινούμενων οχημάτων για να εκπαιδευτούν σε δύσκολες καταστάσεις κίνησης στους δρόμους, στην λήψη αποφάσεων και στην βελτιστοποίηση των πολιτικών τους με δεδομένα πραγματικού χρόνου.

- **Συμβουλευτικά συστήματα (Recommendation systems):** χρησιμοποιείται η ενισχυτική μάθηση για την κατανόηση των προτιμήσεων διαφόρων χρηστών έτσι ώστε να παρέχονται προσωποποιημένες προτάσεις για διάφορα θέματα όπως είναι τα Cookies.
- **Natural language processing (NLP):** ουσιαστικά χρησιμοποιείται σε συστήματα που απαντά μηχανή σε συζητήσεις με χρήστες, σε συστήματα παραγωγής κειμένων και σε συστήματα μεταφράσεων
- **Οικονομικός τομέας (finance):** αξιοποιείται για τον έλεγχο του διεθνούς εμπορίου, τη βελτιστοποίηση επενδυτικών δαπανών και την διαχείριση ρίσκων μέσω της λήψης αποφάσεων εμπορίου με βάση ιστορικά στοιχεία marketing.
- **Τομέας υγείας (Healthcare):** η ενισχυτική μάθηση μπορεί να βοηθήσει στο κομμάτι της θεραπείας ασθενών, της δοσολογίας φαρμάκων αλλά λειτουργεί επικουρικά και στην κατανόηση κλινικών εικόνων ασθενών, την εύρεση νέων φαρμάκων και στην βελτιστοποίηση των κλινικών δοκιμών.
- **Διαχείριση ενέργειας (energy management):** επιστρατεύεται για την βέλτιστη διαχείριση της ενέργειας σε δυναμικά και ενεργειακά συστήματα, δηλαδή την ισορροπία ανάμεσα στην παραγωγή και την κατανάλωση ενέργειας ελαχιστοποιώντας το περιβαλλοντικό και οικονομικό κόστος.
- **Εκπαίδευση (education):** με το χαρακτηριστικό της να μπορεί να προσαρμόζεται στις εκάστοτε απαιτήσεις της εποχής, η ενισχυτική μάθηση χτίζει εξειδικευμένο περιβάλλον εκπαίδευσης στοχευμένο στα προβλήματα του εκάστοτε μαθητή, ενώ πλέον χρησιμοποιούνται για την κατασκευή ιδιαιτέρως έξυπνων συστημάτων εκπαίδευσης.
- **Συστήματα ρύθμισης (control systems):** χρησιμοποιείται για την κατασκευή συστημάτων ρύθμισης μέσα από την διδαχή βέλτιστων πολιτικών ρύθμισης ενώ πλέον είναι δυνατή η προσαρμοστική ρύθμιση που μπορεί να διαχειριστεί πολύπλοκα και δυναμικά συστήματα.

1.2.1. Συστήματα ρύθμισης

Η ενισχυτική μάθηση έχει διεισδύσει αποτελεσματικά και στον τομέα της αυτόματης ρύθμισης καθώς διέπεται από αρχές που προσεγγίζουν την θεωρία βέλτιστου ελέγχου. Στον τομέα της ρύθμισης στόχος είναι η κατάλληλη διαχείριση των μεταβλητών εκ χειρισμού με στόχο την επίτευξη ενός συγκεκριμένου στόχου ενώ στην ενισχυτική μάθηση το ζητούμενο είναι η εύρεση μιας βέλτιστης πολιτικής που να μεγιστοποιεί την ανταμοιβή.

Οι στόχοι που τίθενται στα συστήματα βέλτιστου ελέγχου όπως η ελαχιστοποίηση της χρησιμοποιούμενης ενέργειας, του υπολογιστικού χρόνου και κόστους και η μεγιστοποίηση του ρυθμού παραγωγής και του κέρδους είναι αντικείμενα που μπορεί να διαπραγματευτεί και ένας αλγόριθμος ενισχυτικής μάθησης ύστερα από κατάλληλη διαμόρφωση του περιβάλλοντος εργασίας και της συνάρτησης ανταμοιβής.

Ένα παράδειγμα εφαρμογής της Ενισχυτικής μάθησης στην ρύθμιση διεργασιών θα μπορούσε να είναι η δράση του RL σε ένα σύστημα ρύθμισης με χημικό αντιδραστήρα. Συγκεκριμένα, για αντιδραστήρες συνεχούς έργου πλήρους ανάμιξης (CSTR) που χρησιμοποιούνται ευρέως στους τομείς φαρμάκων, πολυμερών, τροφίμων και διαφόρων χημικών, η εφαρμογή μεθόδων ενισχυτικής μάθησης είναι ιδιαίτερα αποτελεσματική. Η παρουσία ενός RL agent σε ένα τέτοιο σύστημα αποσκοπεί στον βέλτιστο έλεγχο μέσω της αλληλεπίδρασης του agent με το περιβάλλον του και τις εμπειρίες που ποσοτικοποιούνται μέσα από την συνάρτηση ανταμοιβής. Μέσω της επιβράβευσης και της τιμωρίας, ο πράκτορας «μαθαίνει» την πολιτική ενέργειας και κατευθύνει το σύστημα στην επιθυμητή απόκριση που είναι συνήθως η παραγωγή προϊόντος συγκεκριμένων προδιαγραφών.

Μία γενική περιγραφή ενός τέτοιου συστήματος είναι η εξής:

Ο RL agent λαμβάνει σαν πληροφορίες- παρατηρήσεις κάποια στοιχεία του συστήματος όπως η θερμοκρασία των αντιδραστηρίων, η παροχή της τροφοδοσίας ή η θερμοκρασία του ψυκτικού, και επιλέγει μια ενέργεια. Αυτή η αντίδραση «βαθμολογείται» από το σύστημα επιβράβευσης-ποινής και καθώς ο agent έχει σαν

σκοπό τη μεγιστοποίηση της επιβράβευσης, τροποποιεί αναλόγως την επόμενη ενέργεια του [8].

1.2.2. Σύγκριση τεχνικών RL με την κλασική ρύθμιση μέσω PID

Τόσο οι αλγόριθμοι ενισχυτικής μάθησης όσο και οι κλασικοί, ευρέως καθιερωμένοι PID (Proportional-Integral-Derivative) ρυθμιστές μπορούν να χρησιμοποιούν στον τομέα της αυτόματης ρύθμισης και των συστημάτων ελέγχου.

Οι PID είναι ευρέως διαδεδομένοι στον χώρο της βιομηχανίας και στα μηχανικά συστήματα. Λειτουργούν με βάση την ανάδραση, λαμβάνουν δηλαδή πληροφορίες για την τρέχουσα κατάσταση του συστήματος και προσαρμόζουν την ενέργεια που θα δώσουν στο σύστημα με σκοπό να το σταθεροποιήσουν στις επιθυμητές συνθήκες. Αυτή η ενέργεια ελέγχου προέρχεται από τη σύγκριση της τρέχουσας κατάστασης με την επιθυμητή κατάσταση ή το σημείο αναφοράς.

Χρησιμοποιούν ένα προκαθορισμένο μαθηματικό μοντέλο του συστήματος που απαιτεί τη γνώση της δυναμικής με τη μορφή συναρτήσεων μεταφοράς ή μοντέλων μεταβλητών κατάστασης. Για τη βαθμονόμηση των ρυθμιστών PID, υπάρχουν ποικίλλες μέθοδοι, όπως η Ziegler-Nickols.

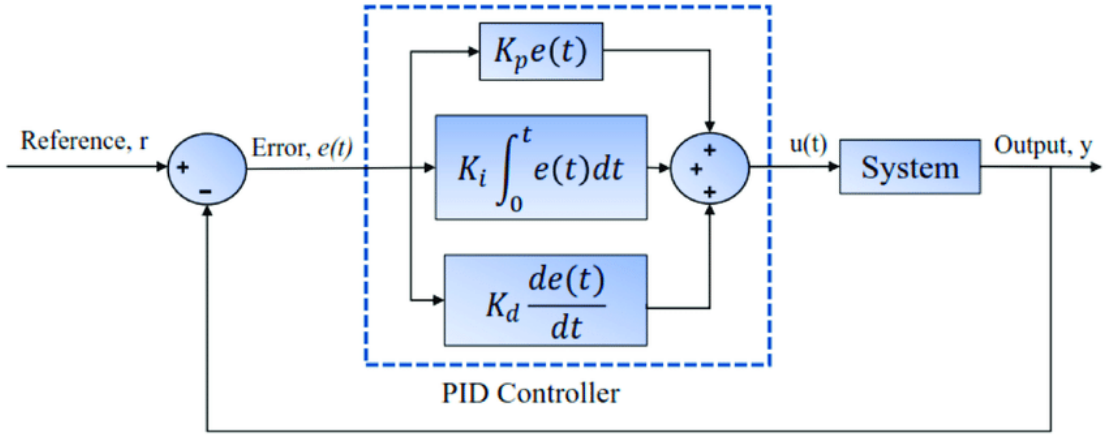
Η δομή των PID διακρίνεται από τρία βασικά δομικά χαρακτηριστικά, τα στοιχεία P, I και D και τα διαφορετικά είδη ρυθμιστών προκύπτουν από συνδυασμούς των χαρακτηριστικών αυτών.

Ο **αναλογικός όρος P** (Proportional) συνεισφέρει στη ενέργεια ελέγχου ανάλογα με τη διαφορά μεταξύ της τρέχουσας κατάστασης και της επιθυμητής κατάστασης.

Ο **ολοκληρωτικός όρος I** (Integral) λαμβάνει υπόψιν το αθροιστικό σφάλμα σε βάθος χρόνου ολοκληρώνοντας το σήμα του σφάλματος. Αυτός ο όρος εξασφαλίζει ότι το σύστημα θα σταθεροποιηθεί σε μια νέα ισορροπία με μηδενικό σφάλμα.

Ο **διαφορικός όρος D** (Derivative) προσφέρει αντισταθμιστικό έλεγχο που εξαρτάται από την ρυθμό αλλαγής της κατάστασης.

Η δομή του PID μπορεί να δοθεί σχηματικά στην εικόνα 5.



Εικόνα 5. Σχηματική αναπαράσταση ενός PID σε ένα σύστημα.

Οι παρακάτω εξισώσεις που εμφανίζονται και στην εικόνα 4, αναφέρονται στην κλασική δομή των PID:΄

- Για το αναλογικό τμήμα:

$$P = K_p * e(t) \quad (1.6)$$

Όπου:

- ✓ K_p το αναλογικό βάρος (proportional gain)
- ✓ $e(t)$ το σφάλμα τη χρονική στιγμή t ανάμεσα στην επιθυμητή τιμή και την μετρούμενη τιμή της διεργασίας

- Για το ολοκληρωτικό τμήμα:

$$I = K_i * \int e(t) \quad (1.7)$$

Όπου:

- ✓ K_i το ολοκληρωτικό βάρος (integral gain)
- ✓ $\int e(t)$ το ολοκλήρωμα του σφάλματος ως προς τον χρόνο

- Για το διαφορικό τμήμα:

$$D = K_d * \frac{de(t)}{dt} \quad (1.8)$$

Όπου:

- ✓ K_d το διαφορικό βάρος (derivative gain)
- ✓ $\frac{de(t)}{dt}$ το διαφορικό του σφάλματος ως προς τον χρόνο

Σύμφωνα με τα παραπάνω, η συνάρτηση μεταφοράς ενός PID ελεγκτή μπορεί να αποδοθεί ως εξής:

$$u(t) = K_p * e(t) + K_i * \int e(t) + K_d * \frac{de(t)}{dt} \quad (1.9)$$

Παρόλα αυτά, προκύπτουν ορισμένοι προβληματισμοί από τη χρήση ρυθμιστών PID οι οποίοι συνοψίζονται στους εξής:

Υπόθεση Γραμμικού Μοντέλου: Οι ελεγκτές PID υποθέτουν μια γραμμική σχέση μεταξύ της ελεγχόμενης μεταβλητής και της εξόδου του συστήματος. Ωστόσο, πολλές πραγματικές διεργασίες, συμπεριλαμβανομένων των συστημάτων CSTR, εμφανίζουν μη γραμμική συμπεριφορά, καθιστώντας δύσκολο για τους ελεγκτές PID να αποτυπώσουν και να ελέγξουν ακριβώς την δυναμική του συστήματος.

Πολυπλοκότητα Ρύθμισης: Η επίτευξη βέλτιστης απόδοσης των ελεγκτών PID απαιτεί προσεκτική επιλογή των παραμέτρων του (αναλογικός, ολοκληρωτικός, και διαφορικός). Αυτή η διαδικασία βαθμονόμησης μπορεί να απαιτεί πολύ χρόνο και είναι ιδιαίτερα απαιτητική, ειδικά για συστήματα με πολύπλοκη δυναμική όπως οι CSTR.

Αλληλεπίδραση και Συσχέτιση: Τα συστήματα CSTR συχνά εμφανίζουν ισχυρή αλληλεπίδραση και συσχέτιση μεταξύ των μεταβλητών εκ χειρισμού και των μεταβλητών που ελέγχονται. Οι ελεγκτές PID συνήθως αντιμετωπίζουν κάθε σχέση εισόδου-εξόδου ανεξάρτητα, παραβλέποντας τις διασυνδέσεις σε επίπεδο συστήματος. Αυτό μπορεί να οδηγήσει σε μη βέλτιστη απόδοση ελέγχου, δυσκολία στην βέλτιστη επίτευξη της τιμής στόχου και την απόρριψη των διαταραχών.

Περιορισμένη Προσαρμοστικότητα σε Μεταβαλλόμενα Συστήματα: Οι ελεγκτές PID δεν προσαρμόζονται εύκολα σε συστήματα με μεταβαλλόμενη δυναμική ή αλλαγές στις συνθήκες λειτουργίας. Αν το σύστημα CSTR υποστεί σημαντικές αλλαγές στις

παραμέτρους ή τις συνθήκες λειτουργίας, η απόδοση του ελεγκτή PID μπορεί να επιδεινωθεί.

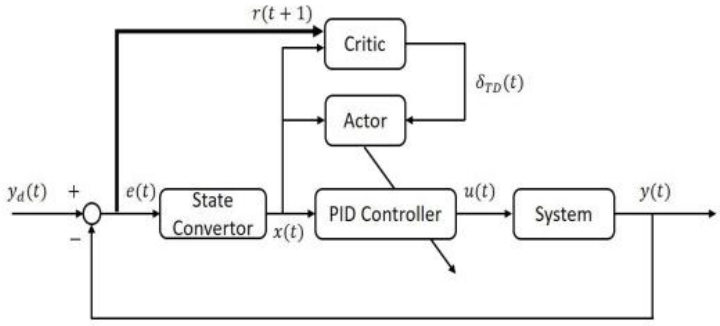
Μη Βέλτιστη Απόδοση Ελέγχου: Οι ελεγκτές PID σχεδιάζονται για τη βέλτιστη απόδοση σε ένα συγκεκριμένο στόχο ελέγχου, συνήθως σταθερότητα, ακολουθία τιμής αναφοράς ή απόρριψη διαταραχών. Ωστόσο, η επίτευξη βέλτιστης απόδοσης σε πολλούς στόχους ταυτόχρονα, όπως διατήρηση της θερμοκρασίας και της συγκέντρωσης εντός των επιθυμητών ευρών ενώ ταυτόχρονα ελαχιστοποιείται η κατανάλωση ενέργειας, μπορεί να αποτελέσει πρόκληση με τους ελεγκτές PID.

Για να αντιμετωπιστούν αυτοί οι περιορισμοί μπορούν να χρησιμοποιηθούν προηγμένες τεχνικές ελέγχου όπως ο ρυθμιστής προβλεπτικού ελέγχου (Model Predictive Control - MPC) ή αλγόριθμοι ενισχυτικής μάθησης (Reinforcement Learning - RL). Αυτές οι προσεγγίσεις προσφέρουν προσαρμοστικότητα, μη γραμμική μοντελοποίηση και τη δυνατότητα αντιμετώπισης πολύπλοκης δυναμικής συστήματος, οδηγώντας σε βελτιωμένη απόδοση ελέγχου στα συστήματα CSTR [9, 10, 11, 12, 13].

Ένα σύστημα ελέγχου με αλγόριθμο RL χρησιμοποίει για τους στόχους της ρύθμισης την αλληλεπίδραση με το περιβάλλον του με σκοπό να βελτιστοποιήσει τις ενέργειες του. Το προτέρημά του είναι ότι δεν απαιτεί μαθηματικό μοντέλο και γνώσεις της δυναμικής του συστήματος αλλά μέσω δοκιμής και σφάλματος μπορεί να αντιμετώπισε αποτελεσματικά ακόμα και μη γραμμικά, πολύπλοκα συστήματα.

Συνολικά, οι ρυθμιστές PID είναι η ιδανική επιλογή για τη ρύθμιση συστημάτων για τα οποία είναι γνωστή η δυναμική και είναι διαθέσιμο ένα μαθηματικό μοντέλο. Σε συστήματα τα οποία είναι περίπλοκα, μεταβατικά ή μη γραμμικά ένα σύστημα RL μπορεί να επιφέρει καλύτερα αποτελέσματα, εφόσον βέβαια διατίθενται και οι κατάλληλοι υπολογιστικοί πόροι [14].

Η σύζευξη των δύο μεθόδων με στόχο την αξιοποίηση των εκατέρωθεν προτερημάτων μπορεί να αποδοθεί γραφικά στο διάγραμμα 5. Στην περίπτωση αυτή, η μεθοδολογία ελέγχου περιλαμβάνει έναν ρυθμιστή PID και έναν αλγόριθμο RL Actor-Critic ο οποίος συνεργάζεται με τον κλασικό ρυθμιστή. Στόχος αυτής της διασύνδεσης είναι ο αρχικός έλεγχος των βασικών παραμέτρων σε ένα πρώτο βαθμό από τον PID και έπειτα η βελτιστοποίηση του συστήματος με την παρουσία του RL διορθώνοντας τυχόν αστοχίες του PID [14].



Διάγραμμα 5. Απλοποιημένη απεικόνιση συστήματος με συνδυασμό ρύθμισης RL-PI

Κεφάλαιο 2. Παρουσίαση του συστήματος ελέγχου

Αντικείμενο της παρούσης διπλωματικής εργασίας είναι η αξιοποίηση της τεχνολογίας ενισχυτικής μάθησης σε συστήματα ελέγχου. Το σύστημα που θα μελετηθεί παρακάτω αποτελείται από έναν αντιδραστήρα CSTR. Το προτεινόμενο σχήμα ελέγχου καταρτίζεται από συνδυασμό των μεθόδων κλασικής ρύθμισης με την παρουσία PID ρυθμιστών και της ενισχυτικής μάθησης με την παρουσία ενός RL πράκτορα. Ακολουθεί μια σύντομη ανάλυση σχετικά με τις δυναμικές που διέπουν την λειτουργία του αντιδραστήρα.

2.1. Εισαγωγή στο πρόβλημα του CSTR

Ο αντιδραστήρας συνεχούς έργου πλήρους ανάμιξης, ή εν συντομία CSTR, είναι από τους πιο ευρέως χρησιμοποιούμενους χημικούς αντιδραστήρες στη βιομηχανία. Αποτελείται από μία καλά αναμεμιγμένη δεξαμενή με αναδευτήρα που εξασφαλίζει ομοιόμορφες συνθήκες σε όλα τα σημεία του αντιδραστήρα. Σε ένα τέτοιο σύστημα, η τροφοδοσία των προϊόντων είναι συνεχής ενώ τα προϊόντα απομακρύνονται καθ' όλη τη διάρκεια της διαδικασίας. Η δυναμική του συστήματος διαμορφώνεται από την συμπεριφορά μεταβλητών όπως η θερμοκρασία και η συγκέντρωση των αντιδρώντων και των προϊόντων μες στον αντιδραστήρα. Οι απαιτήσεις σε ρύθμιση συνήθως συνοψίζονται στον έλεγχο των δύο παραπάνω μεταβλητών στο ρεύμα εξόδου με σκοπό την επίτευξη των επιθυμητών τιμών-στόχων (setpoints) ανάλογα με τις προδιαγραφές της παραγωγής, εξασφαλίζοντας παράλληλα αποδοτικές κινητικές για τον αντιδραστήρα και με βέλτιστους ρυθμούς σύγκλισης.

Η δυναμική συμπεριφορά της θερμοκρασίας σε έναν τέτοιο περιβάλλον καθορίζεται από διαφόρους παράγοντες όπως η μεταφορά θερμότητας, είτε μέσω παραγωγής ή κατανάλωσης θερμότητας κατά την αντίδραση (εξώθερμες ή ενδόθερμες αντιδράσεις) είτε μέσω ανταλλαγής θερμότητας με το περιβάλλον του αντιδραστήρα. Παρόλο που η θερμοκρασία δεν τίθεται ως πρωταρχικός στόχος της

ρύθμισης, ο έλεγχός της κρίνεται αναγκαίος καθώς, τυχόν διακυμάνσεις έξω από τα επιτρεπτά όρια μπορεί να έχουν αρνητικές συνέπειες στην ποιότητα και την ασφάλεια των προϊόντων και τις συνθήκες της διεργασίας.

Από την άλλη πλευρά, στην παρακολούθηση της συγκέντρωσης δίνεται συνήθως περισσότερο βάρος καθώς σχετίζεται άμεσα με την εξέλιξη της αντίδρασης και τις προδιαγραφές των προϊόντων. Έτσι, επιδιώκεται η ρύθμιση και η σταθεροποίηση της συγκέντρωσης σε συγκεκριμένα επιθυμητά επίπεδα με μεγιστοποίηση της απόδοσης και βελτιστοποίηση της ποιότητας των προϊόντων.

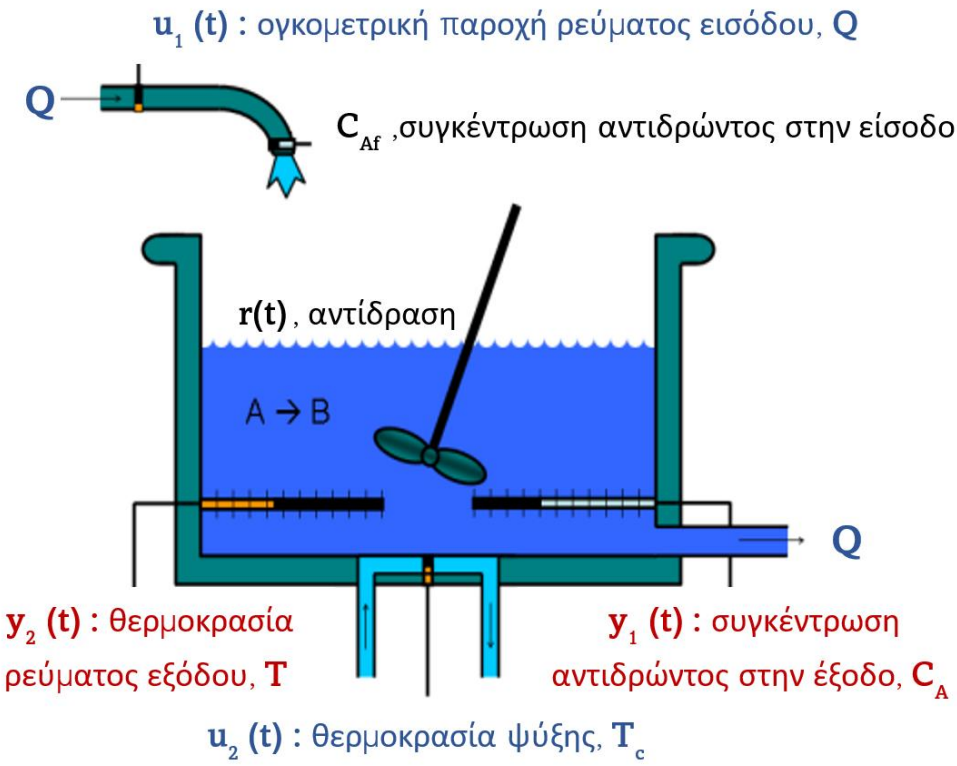
Στις παραδοσιακές μεθόδους ρύθμισης, όπως οι ευρέως καθιερωμένοι ρυθμιστές PID, ανάλογα με την πολυπλοκότητα των συστημάτων που μελετώνται προκύπτουν διάφοροι περιορισμοί και αστοχίες που επιδέχονται βελτίωση. Τα συχνότερα προβλήματα που συναντώνται αφορούν τους παρακάτω τομείς:

1. Μη γραμμικές δυναμικές: λόγω πολύπλοκων κινητικών, φαινομένων μεταφοράς θερμότητας και διαφοροποιήσεων στις συγκεντρώσεις των αντιδρώντων.
2. Χρονικές καθυστερήσεις: τα συστήματα με CSTR αντιμετωπίζουν πρόβλημα με τις χρονικές καθυστερήσεις, λόγω χρόνο-χώρου αντιδραστήρα, τα οποία δεν μπορεί να επιλύσει ένας απλός PID με αποτέλεσμα να παρουσιάζονται ταλαντώσεις και αποκλίσεις στην τελική τιμή.
3. Αβεβαιότητες και διαταραχές: υπάρχουν αβεβαιότητες, συχνά, στην κινητική, στις σταθερές της μεταφοράς θερμότητας της αντίδρασης και στις συγκεντρώσεις της τροφοδοσίας.
4. Επιδράσεις σύζευξης: οι δυναμικές της θερμοκρασίας και της συγκέντρωσης, σε ένα τέτοιο σύστημα, είναι αλληλένδετες. Αυτό κυρίως το πρόβλημα εντοπίζεται σε μεγάλο βαθμό στη συγκεκριμένη εργασία και γίνεται προσπάθεια επίλυσης καθώς οποιαδήποτε αλλαγή στη θερμοκρασία έχει αντίκτυπο στην συγκέντρωση του τελικού προϊόντος, φαινόμενο που αγνοείται από τις κλασσικές μεθόδους ρύθμισης.

Για την αντιμετώπιση των ανωτέρω προβλημάτων προτείνονται προηγμένες μέθοδοι ρύθμισης όπως ο προγνωστικός έλεγχος (MPC) που χρησιμοποιεί προβλεπτικό

μοντέλο για να διαχειριστεί συγχρόνως πολλαπλές μεταβλητές και σύγχρονες προσεγγίσεις μηχανικής μάθησης, όπως η ενισχυτική μάθηση (Reinforcement learning – RL) την οποία πραγματεύεται η παρούσα εργασία.

Ένα τυπικό σύστημα CSTR παρουσιάζεται παρακάτω στην εικόνα 6 [4].



Εικόνα 6. Σύστημα CSTR.

Όπως φαίνεται και στην σχηματική αναπαράσταση, ο αντιδραστήρας τροφοδοτείται με ρεύμα εισόδου ογκομετρικής παροχής Q και συγκέντρωσης C_{Af} ως προς το αντιδρών. Μες στον αντιδραστήρα επικρατούν συνθήκες πλήρους ανάδευσης και θεωρούμε ότι λαμβάνει χώρα η μονόδρομη αντίδραση 1-1 που οδηγεί σε προϊόν B.

$$A \rightarrow B$$

Η αντίδραση είναι εξώθερμη και η ρύθμιση της θερμοκρασίας γίνεται μέσω συστήματος ψύξης με θερμοκρασία ψυκτικού ίση με T_c.Στην έξοδο λαμβάνεται ρεύμα με ογκομετρική παροχή ίση με εκείνη της εισόδου, θερμοκρασία T που αντιστοιχεί και στην θερμοκρασία του αντιδραστήρα λόγω ομοιόμορφης κατανομής στο εσωτερικό και συγκεντρώσεις C_A και C_B για το αντιδρών και το προϊόν

αντίστοιχα. Στην συγκεκριμένη περίπτωση, τα ισοζύγια μάζας και ενέργειας θα διαμορφωθούν με αναφορά στη συγκέντρωση του αντιδρώντος.

ΙΣΟΖΥΓΙΟ ΜΑΖΑΣ

Η γενική μορφή του ισοζυγίου μάζας [15] σε ένα σύστημα δίνεται ως εξής:

$$\text{Είσοδος} - \text{Έξοδος} + \text{Παραγωγή} - \text{Κατανάλωση} = \text{Συσσώρευση}$$

Όταν το ισοζύγιο αναφέρεται σε αντιδρώντα μηδενίζεται ο ρυθμός παραγωγής με αποτέλεσμα το ισοζύγιο να αναδιατυπώνεται ως εξής:

$$\text{Συσσώρευση}^1 = \text{Είσοδος} - \text{Έξοδος} - \text{Κατανάλωση} \quad (2.1)$$

Θεωρώντας σταθερό τον όγκο του υγρού μες στον αντιδραστήρα το ισοζύγιο μάζας για το συστατικό Α σε συνθήκες τέλειας ανάμειξης δίνεται με την παρακάτω εξίσωση (2.2) η οποία ενσωματώνει τους ρυθμούς συσσώρευσης, εισροής, εκροής, κατανάλωσης:

$$V\frac{dC_A}{dt} = Q * C_{Af}(t) - Q * C_A(t) - V * r(t) \Rightarrow \frac{dC_A}{dt} = \frac{Q}{V}\left(C_{Af}(t) - C_A(t)\right) - r(t) \quad (2.2)$$

Επεξήγηση συμβόλων:

- C_{Af}, C_A η συγκέντρωση του Α στην τροφοδοσία και τον αντιδραστήρα αντίστοιχα
- V ο όγκος του αντιδραστήρα
- Q η ογκομετρική παροχή των ρευμάτων εισόδου και εξόδου
- $r(t)$ ο ρυθμός της αντίδρασης

Ο ρυθμός της αντίδρασης $r(t)$ περιγράφεται μέσω της εξίσωσης Arrhenius [16]:

$$r(t) = k_0\, e^{\frac{-E_a}{RT(t)}} C_A(t) \quad (2.3)$$

Επεξήγηση συμβόλων:

- E_A η ενέργεια ενεργοποίησης

[1] Στην περίπτωση μόνιμης κατάστασης, η συσσώρευση θεωρείται μηδενική.

- R η σταθερά Boltzmann
- k_0 ο προεκθετικός παράγοντας
- T η θερμοκρασία του αντιδραστήρα και του ρεύματος εξόδου

Συνδυάζοντας τις εξισώσεις (2.2) και (2.3) διαμορφώνεται η τελική μορφή του ισοζυγίου μάζας:

$$\frac{dC_A}{dt} = \frac{Q}{V}\left(C_{Af}(t) - C_A(t)\right) - k_0\, e^{\frac{-E_a}{RT(t)}} C_A(t) \qquad (2.4)$$

<u>ΙΣΟΖΥΓΙΟ ΕΝΕΡΓΕΙΑΣ</u>

Το ισοζύγιο ενέργειας εκφράζει μαθηματικά την μεταφορά ενέργειας ανάμεσα στο σύστημα και το περιβάλλον του. Όπως και στην περίπτωση του ισοζυγίου μάζας, η γενική μορφή για το ισοζύγιο ενέργειας δίνεται από την εξίσωση (2.5). Επειδή η $\mathrm{A} \rightarrow \mathrm{B}$ είναι εξώθερμη λόγω της αντίδρασης απελευθερώνεται ενέργεια και έτσι μηδενίζεται ο όρος της παραγωγής [15].

$$\text{Συσσώρευση} = \text{Είσοδος} - \text{Έξοδος} + \text{Παραγωγή} \qquad (2.5)$$

Σε έναν αντιδραστήρα CSTR με συνθήκες πλήρους ανάδευσης και ομοιόμορφη κατανομή της θερμοκρασίας το ισοζύγιο ενέργειας διατυπώνεται ως εξής:

$$V\rho C_p \frac{dT}{dt} = Q\rho C_p\left(T_f(t) - T(t)\right) + UA\left(T_c(t) - T(t)\right) + V\Delta H\, r(t) \qquad (2.6)$$

<u>Επεξήγηση συμβόλων:</u>

- ΔH η ενθαλπία της αντίδρασης
- C_p η ειδική θερμοχωρητικότητα
- ρ η πυκνότητα
- U ο συντελεστής μεταφοράς θερμότητας
- A η επιφάνεια μεταφοράς θερμότητας

Επιλύοντας στην (2.6) ως προς την παράγωγο της θερμοκρασίας και αντικαθιστώντας τον ρυθμό της αντίδρασης από τη σχέση (2.3) προκύπτει η τελική μορφή του ισοζυγίου ενέργειας που χρησιμοποιείται στους υπολογισμούς:

$$\frac{dT}{dt} = \frac{Q}{V}\left(T_f(t) - T(t)\right) + \frac{UA}{V\rho C_p}\left(T_c(t) - T(t)\right) + \frac{\Delta H}{\rho C_p}\ k_0\ e^{\frac{-E_a}{RT(t)}} C_A(t) \quad (2.7)$$

Στον πίνακα που ακολουθεί παρατίθενται οι επιλογές για τις σχεδιαστικές παραμέτρους του συστήματος. Τα μεγέθη του πίνακα 1 διατηρούνται σταθερά κατά τους υπολογισμούς.

Πίνακας 1. Δεδομένα για την επίλυση της εξίσωσης 2.7

V	100	m^3
C_p	0.239	J/kg-K
P	1000	Kg/m^3
k_0	7.2e10	s^{-1}
E_α/R	8750	K
ΔH_{rxn}	5e4	J/mol
UA	5e4	W/K

Τελικά, το μη γραμμικό σύστημα εξισώσεων που διαμορφώνεται είναι το ακόλουθο:

$$\frac{dC_A}{dt} = \frac{Q}{V}\left(C_{Af}(t) - C_A(t)\right) - k_0\ e^{\frac{-E_a}{RT(t)}} C_A(t) \quad (2.8)$$

$$\frac{dT}{dt} = \frac{Q}{V}\left(T_f(t) - T(t)\right) + \frac{UA}{V\rho C_p}\left(T_c(t) - T(t)\right) + \frac{\Delta H}{\rho C_p}\ k_0\ e^{\frac{-E_a}{RT(t)}} C_A(t) \quad (2.9)$$

Η μοντελοποίηση του συστήματος και η επίλυση των εξισώσεων αυτών γίνεται υπολογιστικά μέσω του κώδικα του παραρτήματος που αναφέρεται στον αντιδραστήρα (Reactor Set-up) [17].

2.1.1. Ανάλυση μεταβλητών συστήματος

Στο περιβάλλον ρύθμισης, οι μεταβλητές διακρίνονται στις μεταβλητές εκ χειρισμού που συντονίζει ο ρυθμιστής και τις ρυθμιζόμενες μεταβλητές οι οποίες παρακολουθούνται και καθοδηγούνται μέσω των σημάτων ελέγχου.

Για το συγκεκριμένο σύστημα ο επιμερισμός των μεταβλητών γίνεται ως εξής:

- ✓ **<u>Ρυθμιζόμενες μεταβλητές:</u>** $y_1(t) = C_A, y_2(t) = T$

Ως ρυθμιζόμενες μεταβλητές επιλέγονται η συγκέντρωση του αντιδρώντος Α στο ρεύμα εξόδου, C_A και η θερμοκρασία του ρεύματος αυτού, T.

Η συγκέντρωση σχετίζεται άμεσα με τις προδιαγραφές του προϊόντος και αποτελεί έμμεσα ένα δείκτη της απόδοσης του συστήματος. Σε περιπτώσεις μεγιστοποίησης της παραγωγικότητας στόχος είναι η ελαχιστοποίηση της συγκέντρωση του αντιδρώντος, κάτι που μεταφράζεται σε μεγιστοποίηση του παραγόμενου προϊόντος. Ωστόσο, υπάρχουν και περιπτώσεις που οι προδιαγραφές για την συγκέντρωση δεν εμπίπτουν στην ελάχιστη ή μέγιστη τιμή και αποτελούν μια συγκεκριμένη τιμή-στόχο με βάση τις ανάγκες εκμετάλλευσης του προϊόντος.

Η παρακολούθηση της θερμοκρασίας κατά την ρύθμιση του συστήματος είναι ιδιαίτερα σημαντική όχι μόνο επειδή σχετίζεται άμεσα με την συγκέντρωση λόγω της εξίσωσης (2.3) αλλά και επειδή μια υψηλή θερμοκρασία μπορεί να προκαλέσει προβλήματα ασφάλειας στον αντιδραστήρα. Η θερμοκρασία στο εσωτερικό του αντιδραστήρα αποτελεί μοχλό ελέγχου της ταχύτητας των χημικών αντιδράσεων που λαμβάνουν χώρα στον αντιδραστήρα. Απότομες διακυμάνσεις στο προφίλ της θερμοκρασίας μπορεί να οδηγήσουν σε απώλεια προϊόντων, παραμορφώσεις στον έλεγχο της αντίδρασης και ανεπιθύμητες παρενέργειες, όπως η αύξηση των παράγωγων προϊόντων ή η δημιουργία παρεμπιπτόντων προϊόντων ακόμα και η καταστροφή των αντιδραστηρίων.

- **<u>Μεταβλητές εκ χειρισμού:</u>** $u_1(t) = Q, u_2(t) = T_c$

Ως μεταβλητές εκ χειρισμού ορίζονται η παροχή του ρεύματος εισόδου, Q και η θερμοκρασία του ψυκτικού, T_c.

<u>Θερμοκρασία ψυκτικού:</u>

Η θερμοκρασία του ψυκτικού επιλέγεται ως μεταβλητή εκ χειρισμού καθώς επηρεάζει άμεσα – σε συνδυασμό με την συγκέντρωση, C_A - την θερμοκρασία του αντιδραστήρα και του προϊόντος και έτσι διαμορφώνει τις συνθήκες στις οποίες διενεργείται η αντίδραση, όπως φαίνεται και στην εξίσωση (2.7).

Ακολουθεί μια εξήγηση για το πώς λειτουργεί το σύστημα ψύξης:

Κατανόηση του αντιδραστήρα: Ο αντιδραστήρας είναι ένας χώρος όπου λαμβάνουν χώρα χημικές αντιδράσεις. Αν οι αντιδράσεις είναι εξώθερμες, , η θερμοκρασία του αντιδραστήρα αυξάνεται.

Αιτίες υπερθέρμανσης: Υπάρχουν πολλοί παράγοντες που μπορούν να προκαλέσουν υπερθέρμανση στον αντιδραστήρα, όπως υψηλή εξωτερική θερμοκρασία, αντιδράσεις με υψηλή ενεργειακή ελευθέρωση και ανεπαρκής ροή ψύξης.

Σύστημα ψύξης: Το σύστημα ψύξης περιλαμβάνει έναν ψυκτήρα ή έναν εναλλάκτη θερμότητας. Ο ψυκτήρας λειτουργεί με τέτοιο τρόπο ώστε να απορροφά τη θερμότητα που παράγεται στον αντιδραστήρα και να την απομακρύνει από το σύστημα.

Ροή Τροφοδοσίας:

Αντίστοιχα, η παροχή του ρεύματος εισόδου έχει καθοριστικό ρόλο και στα δύο ισοζύγια και σε συνδυασμό με τον ρυθμό της αντίδρασης διαμορφώνει το τελικό προφίλ της συγκέντρωσης.

Η ροή τροφοδοσίας αποτελεί σημαντικό παράγοντα για τον έλεγχο της συγκέντρωσης των επιθυμητών ειδών στον αντιδραστήρα. Ο έλεγχος της ροής τροφοδοσίας μπορεί να επηρεάσει την ταχύτητα των χημικών αντιδράσεων, τον χρόνο παραμονής των ειδών στον αντιδραστήρα και την ποσότητα των παραγόμενων προϊόντων. Όταν ο ρυθμός τροφοδοσίας αυξάνεται, αυξάνεται και η ποσότητα των αρχικών αντιδρώντων που εισάγονται στον αντιδραστήρα, επιτρέποντας έτσι μεγαλύτερη ποσότητα του επιθυμητού προϊόντος να παραχθεί αν οι συνθήκες είναι κατάλληλες σύμφωνα με σχετικό άρθρο [18]. Αυτή η επίδραση του ρυθμού τροφοδοσίας στην συγκέντρωση του επιθυμητού είδους είναι κρίσιμη για τον έλεγχο της απόδοσης του αντιδραστήρα CSTR και για την επίτευξη των επιθυμητών αποτελεσμάτων.

Συνολικά, η ψύξη και η ροή τροφοδοσίας παίζουν κρίσιμο ρόλο στην επίτευξη επιθυμητού ελέγχου θερμοκρασίας και συγκέντρωσης στον CSTR. Η κατάλληλη ρύθμιση και λειτουργία αυτών των παραμέτρων εξασφαλίζει την αποτελεσματική λειτουργία του αντιδραστήρα και την παραγωγή επιθυμητών προϊόντων.

2.1.2. Συνθήκες μόνιμης κατάστασης

Για τον προσδιορισμό συνθηκών μόνιμης κατάστασης μηδενίζονται οι παράγωγοι στις σχέσεις (2.4) και (2.7), δίνονται τιμές στις μεταβλητές εκ χειρισμού και λύνονται οι αλγεβρικές εξισώσεις που προκύπτουν ως προς τις ρυθμιζόμενες μεταβλητές.

Σύμφωνα με την αρχική διατύπωση του ισοζυγίου ενέργειας (2.7) οι επιμέρους όροι της εξίσωσης μπορούν να αποδοθούν ως εξής:

- Συσσώρευση: $V\rho C_p \frac{dT}{dt}$
- Εισροή/Εκροή ενέργειας από το περιβάλλον: $Q\rho C_p \left(T_f(t) - T(t)\right)$
- Εισροή/Εκροή ενέργειας από το σύστημα ψύξης: $UA \left(T_c(t) - T(t)\right)$
- Παραγωγή ενέργειας λόγω της αντίδρασης: $V\Delta H\, r(t)$

Η ενέργεια που μεταφέρεται λόγω Εισροής-Εκροής στο σύστημα μπορεί να ενσωματωθεί σε έναν ενιαίο όρο με συμβολισμό Q_T.

$$Q_T = Q\rho C_p \left(T_f(t) - T(t)\right) + UA \left(T_c(t) - T(t)\right) \quad (2.10)$$

Αντίστοιχα, η ενέργεια που απελευθερώνεται λόγω της αντίδρασης είναι ίση με:

$$Q_R = V\Delta H\, r(t) = V\Delta H\, k_0\, e^{\frac{-E_a}{RT(t)}} C_A(t) \quad (2.11)$$

Σε συνθήκες μόνιμης κατάστασης η συσσώρευση είναι μηδενική και έτσι ο πρώτος όρος μηδενίζεται. Το ισοζύγιο ενέργειας τότε λαμβάνει την μορφή:

$$0 = Q_T + Q_R \Rightarrow Q_T = -Q_R \quad (2.12)$$

Για να υπολογίσει κανείς την μόνιμη κατάσταση του συστήματος για προκαθορισμένες τιμές των μεταβλητών εκ χειρισμού αρκεί να σχεδιάσει την γραφική παράσταση των Q_T και $-Q_R$ για διάφορες θερμοκρασίες μες στον αντιδραστήρα και να βρει το σημείο τομής των δύο καμπυλών. Το σημείο τομής δίνει την θερμοκρασία T_{ss} που επικρατεί στην μόνιμη κατάσταση και μέσω της εξίσωσης Arrhenius οδηγεί και στον υπολογισμό της συγκέντρωσης $C_{A,ss}$.

Σε γραμμικά συστήματα η λύση που προκύπτει είναι μοναδική και οι συνθήκες λειτουργίας οδηγούν σε μια συγκεκριμένη μόνιμη κατάσταση. Ωστόσο, τα περισσότερα φυσικά συστήματα, όπως αυτό του CSTR, είναι ιδιαίτερα πολύπλοκα και μπορεί να παρουσιάσουν περισσότερες από μια μόνιμες κατάστασης για ένα ζεύγος τιμών $\mathrm{Q} - \mathrm{T_c}$.

Γενικά, ένας CSTR είναι σχεδιασμένος για να λειτουργεί σε μέγιστη απόδοσης σε συνθήκες μόνιμης κατάστασης όπου η είσοδος του αντιδρώντος είναι σε ισορροπία με την έξοδο του προϊόντος. Ωστόσο, πολλές φορές είναι πιθανόν να υπάρχουν περισσότερες από μια μόνιμες καταστάσεις που δεν είναι δυνατόν να παρατηρηθούν χωρίς την χρήση μαθηματικών μοντέλων. Η παρουσία πολλών μόνιμων καταστάσεων αποτελεί μία πρόκληση για την ρύθμιση του συστήματος και τη βελτιστοποίηση του καθώς στις περιοχές αυτές οσοδήποτε μικρή μεταβολή στις λειτουργικές συνθήκες μπορεί να αποσταθεροποιήσει πλήρως το σύστημα. Οι δυναμικές του συστήματος θα πρέπει να λαμβάνονται υπόψη κατά τον σχεδιασμό του συστήματος ελέγχου με στόχο την αντιμετώπιση τέτοιων φαινομένων [19].

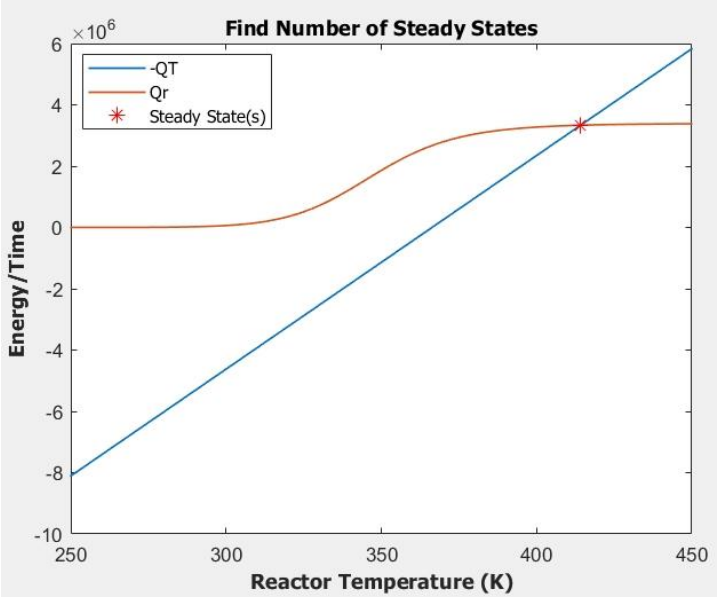
Στην πολλαπλή λύση του συστήματος – όσο και σε μερικές μοναδικές λύσεις - είναι δυνατόν μία ή περισσότερες από τις λύσεις να οδηγούν σε ασταθείς μόνιμες καταστάσεις. Η συμπεριφορά αυτή λαμβάνεται υπόψη κατά τον σχεδιασμό του προβλήματος και την εκπαίδευση του πράκτορα RL. Συγκεκριμένα, οι πολλαπλές λύσεις θεωρείται πως στο μεγαλύτερο μέρος τους οδηγούν σε αστάθεια και απομονώνονται από τις μοναδικές κατά την εκπαίδευση[2]. Παρακάτω, δίνονται τα διαγράμματα με τις καμπύλες Q_T και $-Q_R$ και επισημαίνεται σε αυτά η λύση ή οι λύσεις που δίνουν τα σημεία τομής.

Επιλέγοντας ως τιμές στις μεταβλητές εκ χειρισμού τις ακόλουθες προκύπτει μοναδική λύση για το σύστημα:

- $T_c = 372.7631\ K$

[2] Ασταθείς μόνιμες κατάστασης προκύπτουν και από τις μοναδικές λύσεις όχι όμως σε τόσο μεγάλο βαθμό που να κρίνεται απαραίτητος ο διαχωρισμός τους.

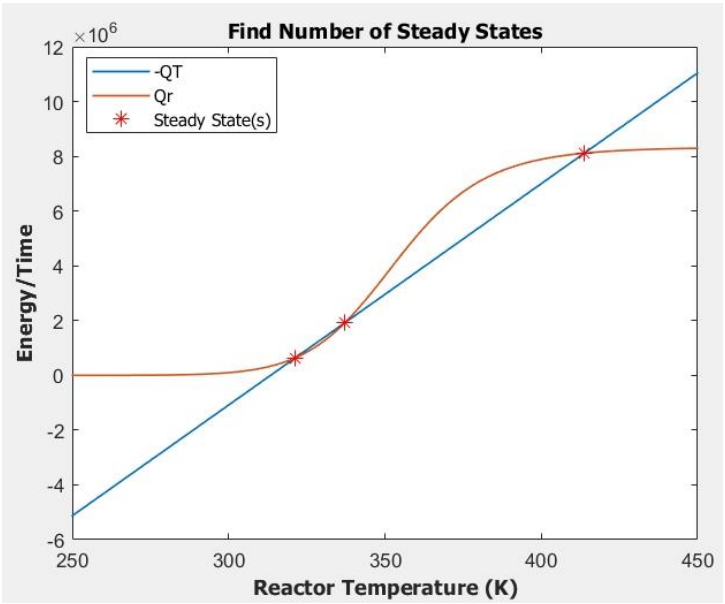
- $Q_f = 82.3738\, m^3/s$



Διάγραμμα 6. Λύση που οδηγεί σε μια μόνιμη κατάσταση

Αντίστοιχα, για το παρακάτω ζεύγος τιμών η λύση που προκύπτει είναι τριπλή αφού οι καμπύλες τέμνονται σε τρεις διαφορετικές θερμοκρασίες.

- $T_c = 290.8301\, K$
- $Q_f = 129.1190\, m3/s$



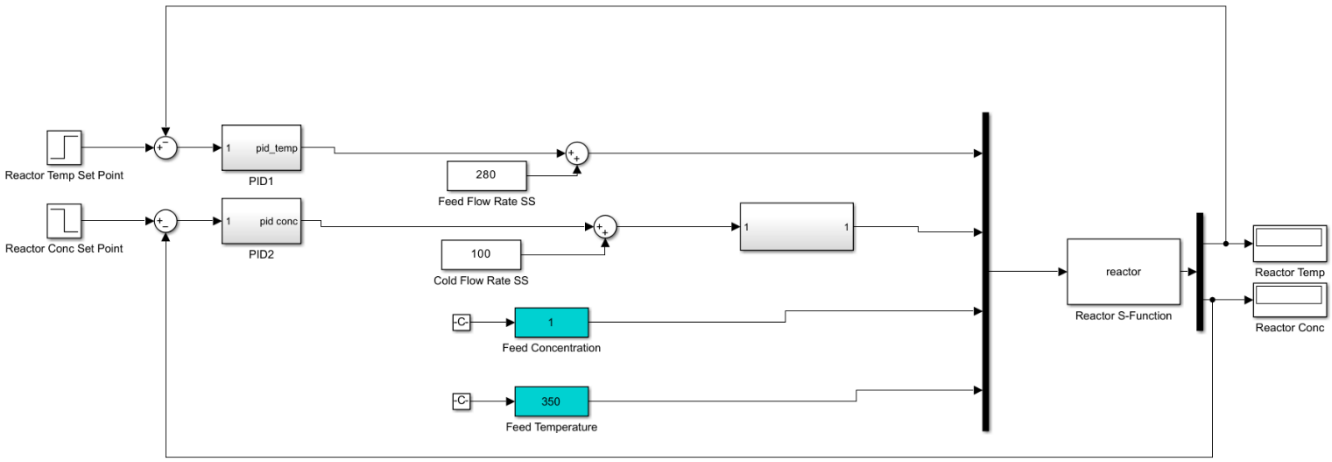
Διάγραμμα 7. Τριπλή λύση που οδηγεί σε πολλαπλές μόνιμες καταστάσεις

Θα πρέπει να αναφερθεί ότι αν επιλεγεί τυχαία ένα ζεύγος επιθυμητών τιμών για τις δύο ρυθμιζόμενες μεταβλητές, δεν είναι πάντα εφικτός ο προσδιορισμός μεταβλητών εκ χειρισμού που μπορούν να οδηγήσουν τις ρυθμιζόμενες μεταβλητές στις επιθυμητές τιμές.

2.2. Ρύθμιση του συστήματος αποκλειστικά με PI

Προτού παρουσιαστεί η μορφή του τελικού σχήματος ελέγχου, παρατίθεται μια σύντομη ανάλυση της ρύθμισης του συστήματος του CSTR αποκλειστικά με την χρήση PI ρυθμιστών. Λόγω της ιδιότητας των PI να διαχειρίζονται αποκλειστικά μια ρυθμιζόμενη μεταβλητή, επιστρατεύονται δύο ρυθμιστές, ένας που μεταβάλλει την παροχή του ρεύματος τροφοδοσίας με στόχο την ρύθμιση της συγκέντρωσης στο εσωτερικό του αντιδραστήρα και ένας δεύτερος ρυθμιστής που καθορίζει την θερμοκρασία του ψυκτικού με στόχο τον έλεγχο της θερμοκρασίας.

Το σύστημα ελέγχου με τον CSTR και τους PI διαμορφώνεται στο περιβάλλον προσομοίωσης του MATLAB-Simulink και έχει την ακόλουθη μορφή.



Εικόνα 7. Σύστημα ελέγχου με τον αντιδραστήρα CSTR και τους ρυθμιστές PI.

Ο ρυθμιστής του ζεύγους Q – C_A (PI1) λαμβάνει ως είσοδο την διαφορά της τωρινής τιμής της συγκέντρωσης από το επιθυμητό setpoint και έχει ως έξοδο την παροχή του

ρεύματος τροφοδοσίας ως μεταβλητή απόκλισης από την αρχική μόνιμη κατάσταση. Αντίστοιχα, ο ρυθμιστής του ζεύγους T_c – T (PI2) έχει ως είσοδο τη διαφορά της θερμοκρασίας από το setpoint της και ως έξοδο την θερμοκρασία του ψυκτικού ως μεταβλητή απόκλισης από την αρχική μόνιμη κατάσταση.

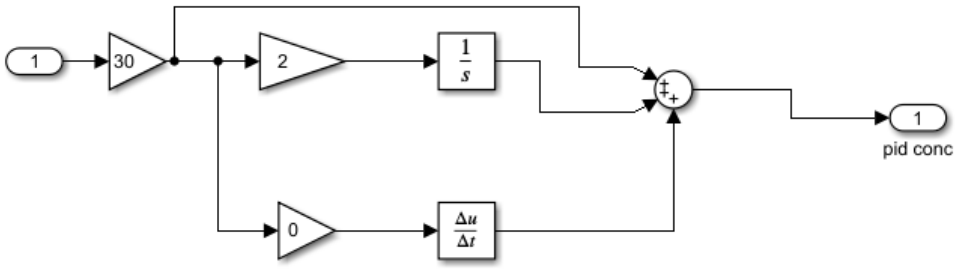
Όπως γίνεται κατανοητό, οι μεταβλητές του προβλήματος δεν είναι ανεξάρτητες μεταξύ τους και η λειτουργία του ενός PI επηρεάζει σημαντικά την λειτουργία του άλλου. Καθένας όμως θεωρείται υπεύθυνος αποκλειστικά για την ρύθμιση μίας μεταβλητής και δεν λαμβάνει υπόψη τις αλληλεπιδράσεις ανάμεσα στις μεταβλητές του συστήματος με αποτέλεσμα σε μερικές περιπτώσεις να δρα ανταγωνιστικά σε σχέση με τον άλλο.

Επιχειρείται μια αρχική βαθμονόμηση των ρυθμιστών με κατάλληλη επιλογή των στοιχείων K_p, K_i, K_d. Συγκεκριμένα, για κάθε ρυθμιστή γίνονται οι εξής επιλογές:

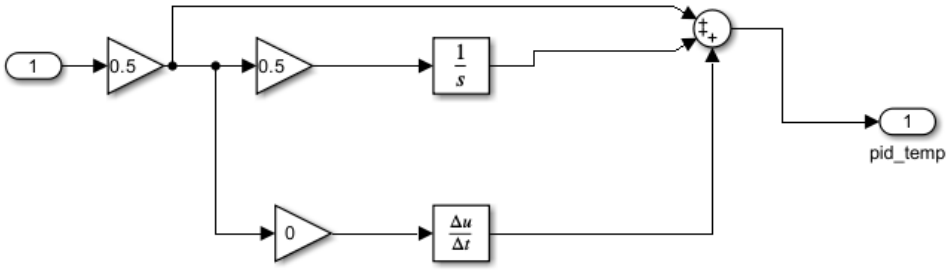
Πίνακας 2. Βαθμονόμηση PI ρυθμιστών

PI θερμοκρασίας		**PI συγκέντρωσης**	
P στοιχείο	0.5	P στοιχείο	30
I στοιχείο	0.25	I στοιχείο	60
D στοιχείο	0	D στοιχείο	0

Στις παρακάτω εικόνες φαίνεται διαγραμματικά μες στο περιβάλλον του MATLAB-Simulink η εσωτερική δομή κάθε ρυθμιστή με βάση τις επιλογές του πίνακα 2.



Εικόνα 8. Δομή PI που είναι υπεύθυνος για τη ρύθμιση της συγκέντρωσης



Εικόνα 9. Δομή PI που είναι υπεύθυνος για τη ρύθμιση της θερμοκρασίας

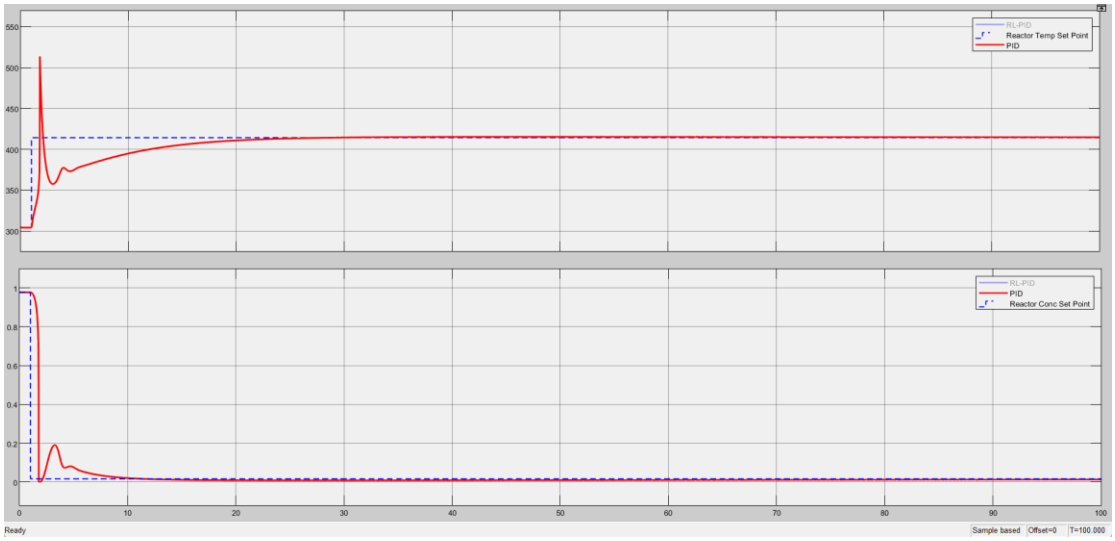
Όπως αποτυπώνεται και στις εξισώσεις λειτουργίας των PI η επιλογή των σταθερών σχετίζεται άμεσα με την στρατηγική ελέγχου και τους στόχους της ρύθμισης σε κάθε πρόβλημα.

Μεγάλη τιμή στο P στοιχείο όπως αυτό της συγκέντρωσης υποδηλώνει ότι ενισχύεται η απόκριση του ελέγχου σε σφάλματα της συγκέντρωσης με αποτέλεσμα μία πιο επιθετική διόρθωση. Παρόμοια επιθετική διόρθωση στο σφάλμα σταθερής κατάστασης στη συγκέντρωση προκύπτει από το μεγάλο βάρος στο ολοκληρωτικό στοιχείο του PID της συγκέντρωσης ενώ η απουσία D στοιχείου υποδηλώνει την απουσία άμεσης αντίδρασης σε γρήγορες μεταβολές της συγκέντρωσης και έτσι οι ρυθμιστές εδώ μπορούν να ονομάζονται PI. Στον PI της θερμοκρασίας έχουν χρησιμοποιηθεί μικρότερες τιμές γεγονός που υποδηλώνει μία πιο προσεκτική απόκριση σε σφάλματα και μεταβολές. Η απουσία διαφορικού στοιχείου αντίστοιχα υποδηλώνει αδράνεια σε γρήγορες μεταβολές της θερμοκρασίας.

Συνολικά, για τον έλεγχο της συγκέντρωσης, επιδιώκεται μια πιο επιθετική απόκριση με υψηλότερα βάρη στα στοιχεία P και I, ενώ δεν χρησιμοποιείται διαφορική δράση. Αντίθετα, ο έλεγχος θερμοκρασίας χρησιμοποιεί χαμηλότερα βάρη P και I, αντανακλώντας μια πιο προσεκτική απόκριση και επίσης δεν χρησιμοποιεί διαφορική δράση. Οι επιλογές αυτές σχετίζονται άμεσα με το γεγονός πως η συγκέντρωση αποτελεί τον βασικότερο παράγοντα απόδοσης του σύστηματος ενώ η θερμοκρασία συμπεριλαμβάνεται στο σύστημα ελέγχου για λόγους παρακολούθησης και ελέγχου.

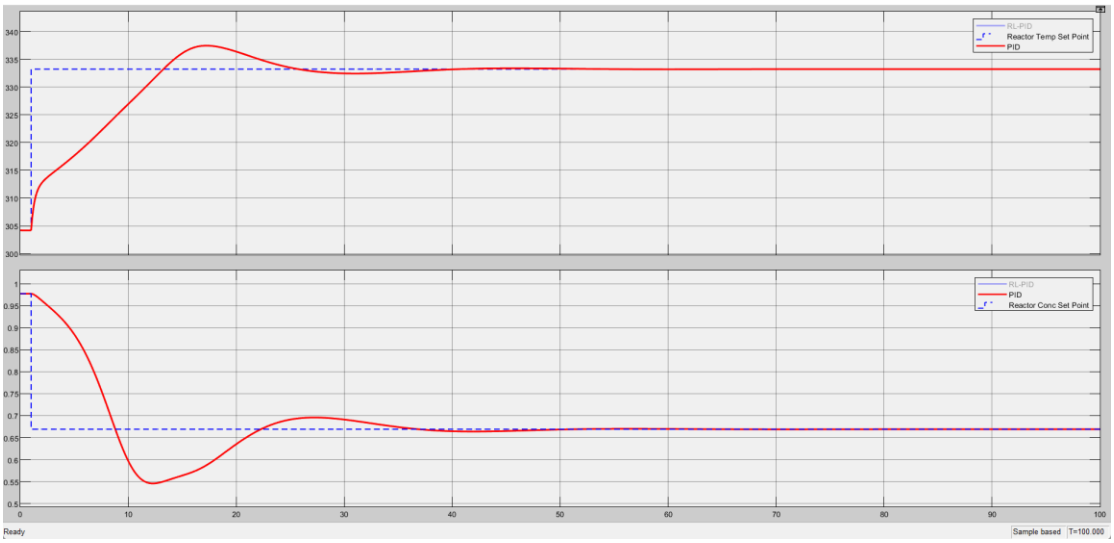
Στο σημείο αυτό, θα πρέπει να σημειωθεί πως η βαθμονόμηση των PI είναι μια χρονοβόρα, επαναληπτική διαδικασία που απαιτεί προσεκτική εξέταση της δυναμικής του συστήματος, των χαρακτηριστικών απόκρισης και των στόχων ελέγχου. Ο στόχος της παρούσης εργασίας δεν είναι η ανάλωση στην εύρεση των καταλληλότερων τιμών για τα στοιχεία ρύθμισης αλλά η διερεύνηση της συνεισφοράς του RL στο αρχικό σύστημα ελέγχου.

Το υπάρχον σύστημα ελέγχου δουλεύει αποδοτικά σε αρκετές περιπτώσεις. Δύο ενδεικτικά παραδείγματα από την καλή εφαρμογή των PI στο παρόν σύστημα με τον CSTR φαίνονται στα διαγράμματα 8 και 9. Στο διάγραμμα 8 ζητείται επίτευξη των τιμών $T_{sp} = 414.2\ K$ και $C_{A,sp} = 0.016812\ mol/m^3$.



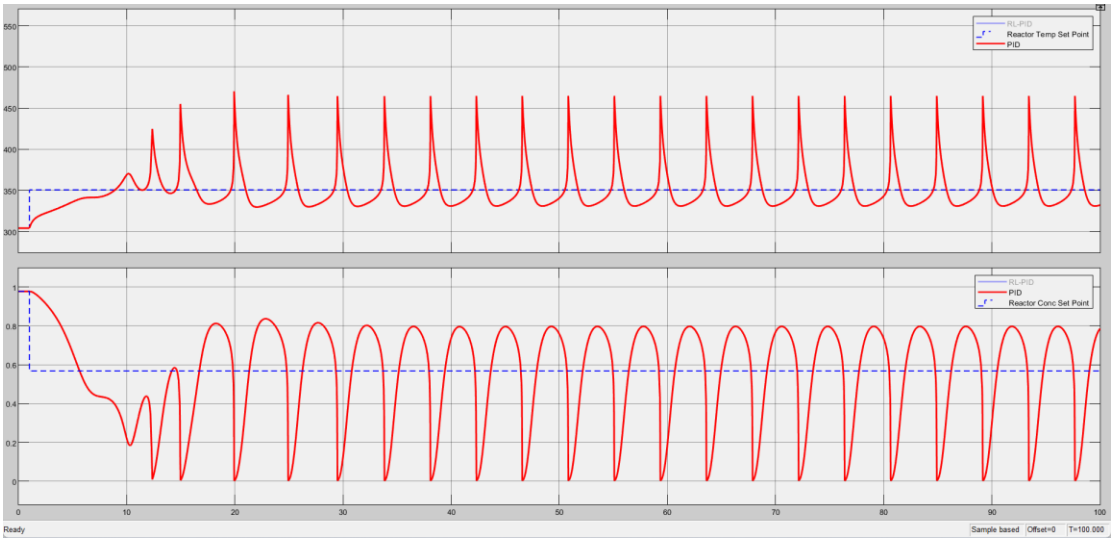
Διάγραμμα 8. Απόδοση συστήματος PI σε χαμηλές συγκεντρώσεις ευσταθών Μ.Κ.

Αντίστοιχα, στο διάγραμμα 9 τα ζητούμενα setpoints είναι $T_{sp} = 333.23\ K$ και $C_{A,sp} = 0.66902\ mol/m^3$.

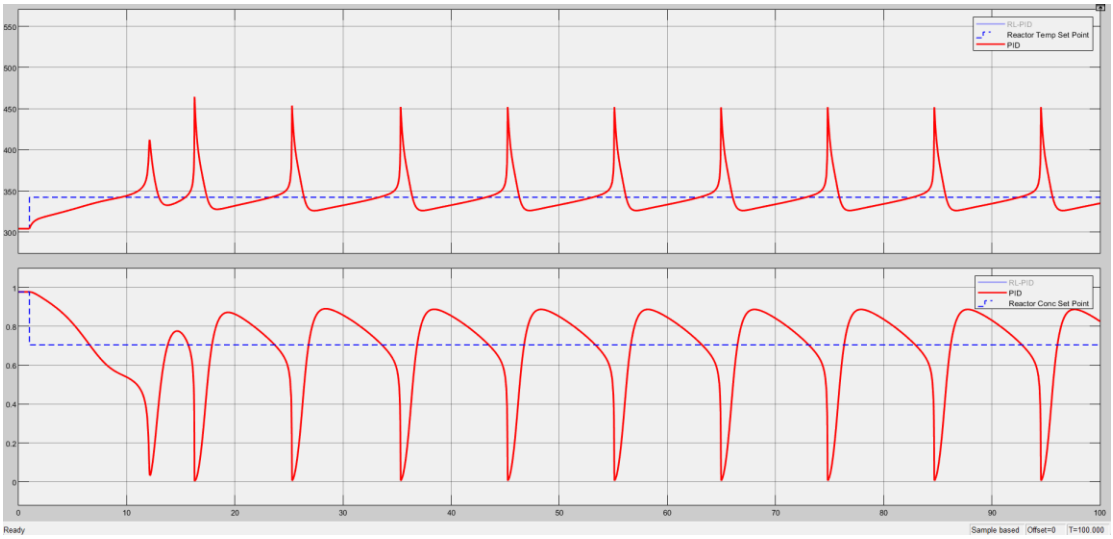


Διάγραμμα 9.Απόδοση συστήματος PI σε μέσες συγκεντρώσεις ευσταθών Μ.Κ.

Ωστόσο, παρουσιάζονται και κάποιες περιπτώσεις στις οποίες ο συνδυασμός των δύο ρυθμιστών PI αποτυγχάνει να προσεγγίσει σωστά τις επιθυμητές τιμές και οι αποτυχημένες του απόπειρες οδηγούν το σύστημα σε ταλαντώσεις σταθερού πλάτους, όπως φαίνεται στα διαγράμματα 10 και 11. Στο διάγραμμα 10 επιχειρείται προσέγγιση των τιμών $T_{sp} = 350.71\ K$ και $C_{A,sp} = 0.56815\ mol/m^3$ ενώ στο διάγραμμα 11 οι τιμές-στόχοι είναι $T_{sp} = 342.33\ K$ και $C_{A,sp} = 0.70428\ mol/m^3$



Διάγραμμα 10. Απόδοση συστήματος PI σε μέσες συγκεντρώσεις ασταθή Μ.Κ.



Διάγραμμα 11. Απόδοση συστήματος PI σε υψηλές συγκεντρώσεις ασταθή Μ.Κ.

2.3. Ανάπτυξη συνεργατικού συστήματος ελέγχου PI-RL

Η προηγούμενη παράγραφος κατέδειξε πως οι PI από μόνοι τους δεν μπορούν να κατευθύνουν αποτελεσματικά το σύστημα σε όλα τα ζητούμενα setpoints και δυσκολεύονται ιδιαίτερα σε εκείνα που οδηγούν σε ασταθείς μόνιμες καταστάσεις και προκύπτουν κατά κύριο λόγο από τριπλές ρίζες του συστήματος.

Για την βελτίωση του παρόντος συστήματος ρύθμισης στο σχήμα ελέγχου προστίθεται ένας RL agent ο οποίος αναλαμβάνει να διορθώσει το νόμο ελέγχου που προκύπτει από τους PI ρυθμιστές. Σε αντίθεση με τους PI, ο RL πράκτορας έχει τη δυνατότητα να μεταβάλλει συγχρόνως τις δύο μεταβλητές, ή και περισσότερες, και να παρατηρεί επιπλέον μεταβλητές πέρα από τα σφάλματα των ρυθμιζόμενων μεταβλητών.

Από την αλληλεπίδραση με το περιβάλλον λαμβάνει σαν είσοδο την κατάσταση του συστήματος – όπως αυτή καθορίζεται από πλήθος μεταβλητών που επιλέγει ο χρήστης - και παρεμβαίνει για να καθορίσει την επόμενη κατάσταση με τις ενέργειες του ανάλογα με το feedback που παρέχει η συνάρτηση ανταμοιβής.

Σκοπός του RL πράκτορα είναι η μεγιστοποίηση της συνάρτησης ανταμοιβής που θα σημάνει την ορθή λειτουργία του συστήματος σε περίπτωση που έχουν ληφθεί σωστές επιλογές από τον χρήστη για τον σχεδιασμό της.

Με βάση, λοιπόν, τα παραπάνω, οι PI και ο RL στο σύστημα λειτουργούν παράλληλα και ο ένας επηρεάζει την λειτουργία του άλλου. Ο PI παρέχει τα αρχικά σήματα ελέγχου με βάση τις παραμέτρους βαθμονόμησης και καθορίζει σε μεγάλο βαθμό την κατεύθυνση στην οποία οδηγείται το σύστημα. Ο RL προσπαθεί να τροποποιήσει αυτή την κατεύθυνση όταν η λειτουργία των PI δεν είναι ικανοποιητική, προσαρμόζοντας συνεχώς και τελειοποιώντας την πολιτική του μέσα από τις εμπειρίες που αποκτά από την αλληλεπίδραση με το σύστημα.

Ο PI δεν έχει άμεση πρόσβαση στις ενέργειες του RL όμως μπορεί να αποκτήσει μια εικόνα της συνεισφοράς του μέσα από τις αλλαγές στα σήματα εισόδου. Ο RL, αντιθέτως, μπορεί να παρακολουθεί τις επιλογές που κάνουν οι PI, όπως αυτές αποτυπώνονται στα σήματα ελέγχου, καθώς και την πορεία των ρυθμιζόμενων μεταβλητών, αποκτώντας έτσι μια συνολική εικόνα για το σύστημα. Ορισμένες ή και όλες αυτές οι πληροφορίες μπορούν να χρησιμοποιηθούν από τον πράκτορα RL σαν παρατηρήσεις του περιβάλλοντος για να αξιολογήσει την αποτελεσματικότητα των ενεργειών του και να λάβει τεκμηριωμένες αποφάσεις.

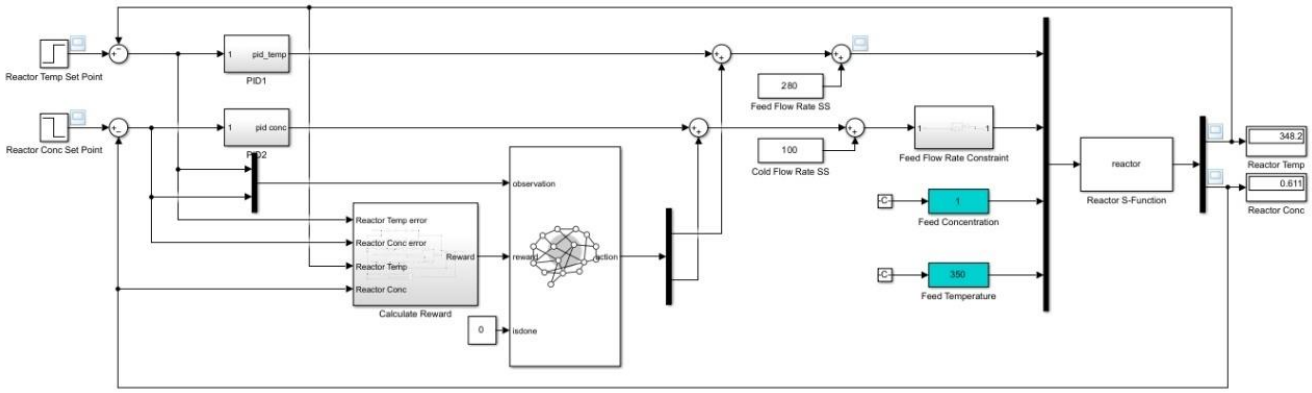
Η συνάρτηση ανταμοιβής καθορίζει τον τρόπο με τον οποίο αξιολογούνται οι ενέργειες του RL. Η ανταμοιβή μπορεί να βασίζεται σε προκαθορισμένους στόχους όπως η επίτευξη επιθυμητών τιμών θερμοκρασίας και συγκέντρωσης, καθώς και στη μείωση των αποκλίσεων από αυτές τις τιμές αλλά και τον χρόνο που αυτές προσεγγίζονται [20].

Συνοψίζοντας, στο σύστημα CSTR, οι PI ελεγκτές, ο RL πράκτορας, αλληλεπιδρούν για τον έλεγχο της θερμοκρασίας και της συγκέντρωσης του επιθυμητού είδους.

Οι PI ελεγκτές χρησιμοποιούνται για να ρυθμίσουν το σύστημα και να διατηρήσουν τη θερμοκρασία και τη συγκέντρωση σε επιθυμητά επίπεδα. Οι ελεγκτές αυτοί λαμβάνουν μετρήσεις από το σύστημα και προσαρμόζουν τις μεταβλητές εκ χειρισμού για να διορθώσουν τυχόν αποκλίσεις από τις επιθυμητές τιμές [10].

Ο RL πράκτορας χρησιμοποιείται για να βελτιώσει τη στρατηγική ελέγχου από την αλληλεπίδρασή του με το περιβάλλον. Ο πράκτορας αξιολογεί τις καταστάσεις και λαμβάνει αποφάσεις σχετικά με τις ενέργειες που πρέπει να προκαλέσει για να ρυθμίσει το σύστημα [1].

Στην εικόνα 10 παρουσιάζεται η συνδεσμολογία του τελικού συστήματος ελέγχου. Οι PI ρυθμιστές δέχονται τις τιμές των σφαλμάτων από τις τιμές-στόχους και παράγουν τα σήματα ελέγχου. Η ενέργεια του RL προστίθεται στα υπάρχοντα σήματα ελέγχου διαμορφώνοντας το τελικό προφίλ των μεταβλητών εκ χειρισμού. Η δομή του πράκτορα διαμορφώνεται από την εκπαίδευση δύο νευρωνικών δικτύων (actor-critic) όπως αναφέρθηκε προηγούμενα. Η διαδικασία εκπαίδευσης παρουσιάζεται αναλυτικά στη συνέχεια της εργασίας.



Εικόνα 10. Εκδοχή του συνολικού τελικού συστήματος για μία επιλεχθείσα συνάρτηση ανταμοιβής.

Το περιβάλλον του πράκτορα αναπτύσσεται μέσω του κώδικα (environment definition) που επισυνάπτεται στο παράρτημα και σε αυτό καθορίζονται τα σήματα των παρατηρήσεων (observations) και τα σήματα των ενεργειών (actions). Ως ενέργειες ορίζονται οι μεταβλητές απόκλισης για την παροχή και την θερμοκρασία του ψυκτικού και ως παρατηρήσεις τα σήματα των σφαλμάτων. Στο σημείο αυτό, ενσωματώνονται τυχόν περιορισμοί για τις μεταβλητές του συστήματος. Τέλος, στο αντίστοιχο function του παραρτήματος (Local Reset Function/ LocalResetFcn) περιλαμβάνεται ο κώδικας που ανανεώνει το περιβάλλον μετά από κάθε επεισόδιο, καθορίζοντας νέο ζεύγος setpoint από εκείνα του πίνακα υπολογισμών.

2.3.1. Δομή πράκτορα

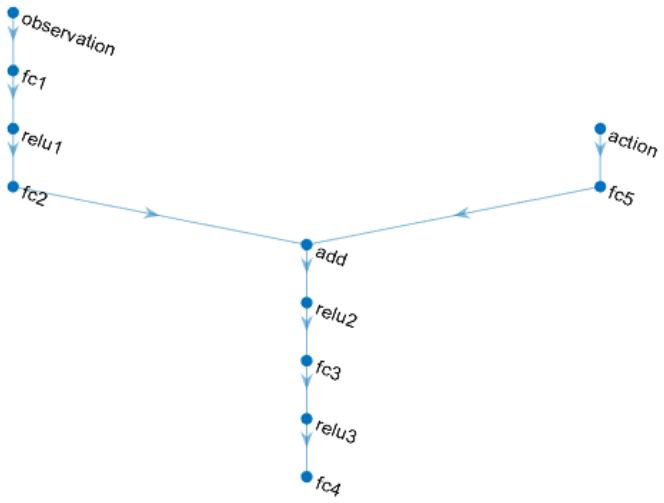
Ως αλγόριθμος για την αρχιτεκτονική του πράκτορα επιλέγεται ο αλγόριθμος DDPG που αναλύθηκε εκτενώς στην παράγραφο στην παράγραφο 1.1.2 και ανήκει στην κατηγορία των Actor-Critic αλγορίθμων. Σύμφωνα με την ανάλυση αυτή, ο Actor-Critic δομείται από δύο νευρωνικά δίκτυα, τον Actor και τον Critic.

Ο Critic που είναι υπεύθυνος για την εκτίμηση της αξίας των ενεργειών αποτελείται από δύο πλήρως συνδεδεμένα επίπεδα με 20 νευρώνες το καθένα και ένα άλλο πλήρως συνδεδεμένο επίπεδο με έναν μόνο νευρώνα.

Ο Actor που αναλαμβάνει την βελτιστοποίησης της πολιτικής σχεδιάζεται με τρία πλήρως συνδεδεμένα επίπεδα 20 νευρώνων και ένα επίπεδο εξόδου που δίνει τελικά την προτεινόμενη ενέργεια του πράκτορα.

Ως συνάρτηση ενεργοποίησης χρησιμοποιείται η ReLU, εκτός από το προτελευταίο επίπεδο του actor όπου χρησιμοποιείται υπερβολική εφαπτομένη.

Παρουσιάζονται σχηματικά στα διαγράμματα 12 και 13.



Διάγραμμα 12. Νευρωνικό δίκτυο για τον Critic του πράκτορα

Διάγραμμα 13. Νευρωνικό δίκτυο για τον Actor του πράκτορα

Όπως είναι φανερό, ο Critic λαμβάνει τόσο τις παρατηρήσεις όσο και τις ενέργειες που διαμορφώνουν την μελλοντική κατάσταση του συστήματος και επιστρέφει μία εκτίμηση της τιμής της ανταμοιβής ενώ ο Actor δέχεται τις παρατηρήσεις και επιστρέφει την ενέργεια με βάση την τρέχουσα πολιτική του.

Ο πράκτορας DDPG ενσωματώνει τόσο την εξερεύνηση του χώρου κατάστασης όσο και εκμετάλλευση των ευρημάτων του. Η εξερεύνηση διευκολύνεται μέσω της προσθήκης θορύβου στις ενέργειες του actor. Στον κώδικα που σχεδιάζει τον πράκτορα (agent definition), η παράμετρος agentOptions.NoiseOptions.Variance ορίζεται ίση με 0.1, υποδηλώνοντας την παρουσία θορύβου που προστίθεται στην επιλογή της ενέργειας. Αυτός ο θόρυβος βοηθά στην εξερεύνηση διάφορων ενεργειών και αποτρέπει τον πράκτορα από το να μένει ακινητοποιημένος σε μια υποβέλτιστη πολιτική.

Η εκμετάλλευση επιτυγχάνεται μέσω του Critic δικτύου, το οποίο εκτιμά τις τιμές Q που συνδέονται με τα τρέχοντα ζεύγη κατάστασης-ενέργειας. Το δίκτυο του Actor χρησιμοποιεί στη συνέχεια αυτές τις εκτιμώμενες τιμές Q για να επιλέξει ενέργειες που αναμένεται να παράγουν υψηλότερες ανταμοιβές. Μέσω της ταυτόχρονης

βελτιστοποίησης των δικτύων του ηθοποιού και του κριτικού, ο πράκτορας ισορροπεί την εξερεύνηση και την εκμετάλλευση.

Επιπλέον, η παράμετρος agentOptions.TargetSmoothFactor (ορισμένη σε 1e-3) ελέγχει τον ρυθμό με τον οποίο ενημερώνονται τα δίκτυα στόχων (target networks). Με τον αργό ρυθμό ενημέρωσης των δικτύων στόχων προς τα τρέχοντα δίκτυα του actor και του critic, ο πράκτορας μπορεί να σταθεροποιηθεί και να βελτιώσει την απόδοσή του με τον χρόνο, επιτυγχάνοντας μια ισορροπία μεταξύ εξερεύνησης και εκμετάλλευσης.

Συνολικά, ο παρεχόμενος κώδικας ενσωματώνει την εξερεύνηση και την εκμετάλλευση έμμεσα μέσω του αλγορίθμου DDPG, της προσθήκης θορύβου και της βελτιστοποίησης των δικτύων του actor και του critic. Αυτό επιτρέπει στον πράκτορα να εξερευνήσει διάφορες ενέργειες και να εκμεταλλευτεί τις εμπειρίες του για να βελτιώσει την απόδοσή του στο προκείμενο περιβάλλον.

Κεφάλαιο 3. Αποτελέσματα και συζήτηση

3.1. Σχεδιασμός της συνάρτησης ανταμοιβής

Ύστερα από την κατάλληλη διαμόρφωση του περιβάλλοντος εργασίας και της αρχιτεκτονικής του πράκτορα ακολουθεί ένα από τα κρισιμότερα αντικείμενα της μελέτης, ο σχεδιασμός της συνάρτησης ανταμοιβής.

Ο στόχος της συνάρτησης ανταμοιβής, όπως αναφέρθηκε και προηγουμένως, είναι να καθοδηγήσει τον πράκτορα στην εύρεση της βέλτιστης πολιτικής που θα ικανοποιεί τις απαιτήσεις ρύθμισης του συστήματος. Σε πρώτη φάση, σκοπός είναι να επιτευχθούν οι τιμές των setpoints με το μικρότερο δυνατό κόστος ελέγχου και σε όσο το δυνατόν λιγότερο χρόνο. Συγχρόνως, προκύπτουν και άλλοι παράπλευροι στόχοι όπως η αποφυγή των ταλαντώσεων, η μείωση της υπέρβασης και η βελτίωση της συμπεριφοράς σε σχέση με το αρχικό σύστημα ελέγχου των PI.

Όπως αναλύθηκε εκτενώς στην παράγραφο 2.1.2, για να βελτιωθούν οι συνθήκες εκπαίδευσης κατασκευάστηκαν δύο ξεχωριστά πακέτα τιμών-στόχων με βάση την προέλευση τους είτε από την μοναδική είτε από την τριπλή λύση της εξίσωσης (2.12). Οι συναρτήσεις ανταμοιβής που σχεδιάστηκαν προσαρμόστηκαν πάνω στις απαιτήσεις και τα προβλήματα των δύο διαφορετικών ειδών setpoints.

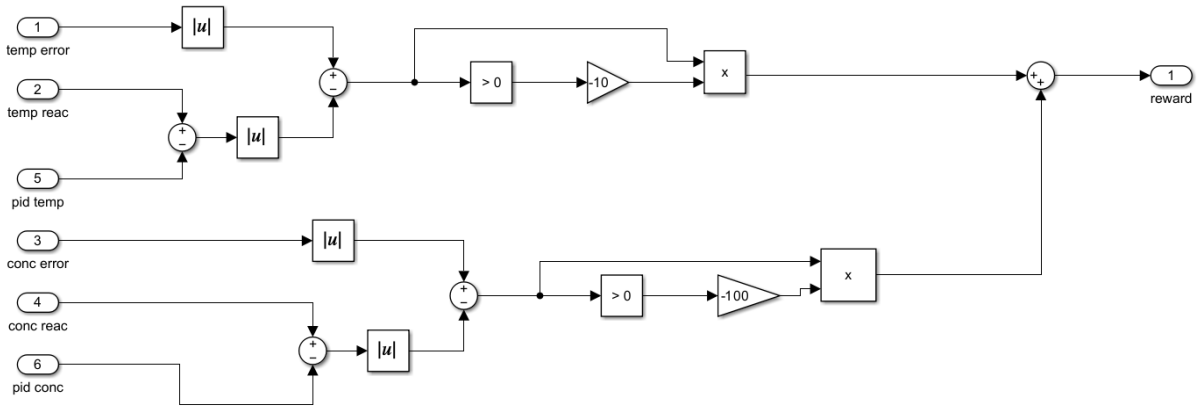
Τα setpoints που προήλθαν από την τριπλή λύση του συστήματος κατά κύριο λόγο οδηγούσαν τους PI σε έντονα ταλαντωτικές συμπεριφορές γύρω από τις τιμές-στόχους όπως φάνηκε καθαρά και στα διαγράμματα 10 και 11. Στις περιπτώσεις αυτές, η ρύθμιση του συστήματος είναι πιο απαιτητική καθώς ο RL καλείται να παρέμβει και να διορθώσει την ασταθή συμπεριφορά των κλασικών ρυθμιστών.

Αντιθέτως, στα setpoints που αποτελούσαν μοναδική λύση του συστήματος η σταθεροποίηση του συστήματος από τους PI ήταν πιο εύκολη. Εδώ ο RL καλείται να βελτιώσει ακόμα περισσότερο την συμπεριφορά των ρυθμιστών είτε μειώνοντας τον χρόνο αποκατάστασης είτε ελαχιστοποιώντας την υπέρβαση.

Κατά την αξιολόγηση του τελικού συστήματος ελέγχου εξετάζονται κυρίως οι ακόλουθοι παράγοντες:

- **<u>Χρόνος αποκατάστασης</u>**: το χρονικό διάστημα που χρειάζεται το σύστημα ρύθμισης για να φτάσει την απόλυτη απόκλιση στο 2% της τιμής στόχου
- **<u>Υπέρβαση</u>**: η μέγιστη απόσταση της μεταβλητής εξόδου από την τελική της κατάσταση
- **<u>Φαινόμενα ταλάντωσης</u>**: η εμφάνιση ταλαντωτικής συμπεριφοράς είτε στις ρυθμιζόμενες μεταβλητές είτε στις μεταβλητές εκ χειρισμού

Γενικά, ο σχεδιασμός του συνδυαστικού συστήματος ελέγχου κρίνεται επιτυχημένος όταν ο RL οδηγεί σε μικρότερους χρόνους αποκατάστασης και πιο ομαλές μεταβάσεις σε σχέση με το σύστημα των PI. Παράλληλα, στην αξιολόγηση μελετάται και η συμπεριφορά τόσο της παροχής της τροφοδοσίας όσο και της θερμοκρασίας του ψυκτικού καθώς δεν είναι επιθυμητές τιμές εκτός ορίων και φαινόμενα ταλάντωσης. Οι συναρτήσεις ανταμοιβής που σχεδιάστηκαν παρουσιάζονται στις εικόνες 11 και 12.



Εικόνα 11. Συνάρτηση ανταμοιβής προσαρμοσμένη στις μοναδικές λύσεις του συστήματος

Η συνάρτηση ανταμοιβής της εικόνας 11 σχεδιάστηκε με αποκλειστικό σκοπό την επίδειξη συμπεριφοράς καλύτερης από εκείνη του PI, όταν αυτή είναι ήδη αρκετά ικανοποιητική. Η λογική πίσω από τον σχεδιασμό της έγκειται στο γεγονός ότι ο RL πρέπει να πλησιάσει κοντά στην τιμή στόχο πιο γρήγορα από τους PI και να απομακρυνθεί από αυτή όσο λιγότερο είναι εφικτό.

Αναλυτικότερα στον πίνακα 3 παρέχεται επεξήγηση των σημάτων:

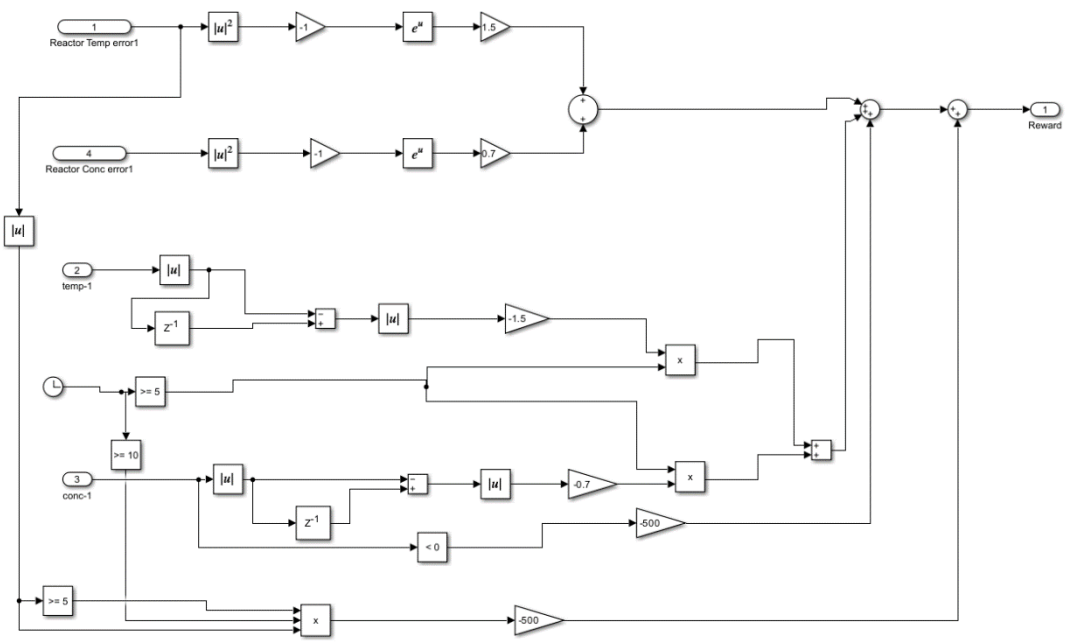
Πίνακας 3. Συνάρτηση ανταμοιβής για Ευσταθής τιμές.

α/α Σημάτων	Επεξήγηση
1. Temp error	Το σφάλμα της θερμοκρασίας, όπως προκύπτει από τις ενέργειες RL-PI.
2. Temp reac	Η θερμοκρασία του αντιδραστήρα, όπως προκύπτει από τις ενέργειες RL-PI.
3. Conc error	Το σφάλμα της συγκέντρωσης, όπως προκύπτει από τις ενέργειες RL-PI.
4. Conc reac	Η συγκέντρωση του αντιδραστήρα, όπως προκύπτει από τις ενέργειες RL-PI.
5. PID Temp (PI Temp)	Η θερμοκρασία του αντιδραστήρα, όπως προκύπτει από τις ενέργειες μόνο του PI.
6. PID Conc (PI Conc)	Η συγκέντρωση του αντιδραστήρα, όπως προκύπτει από τις ενέργειες μόνο του PI.

Η συγκεκριμένη συνάρτηση ανταμοιβής σχεδιάστηκε πάνω σε δύο βασικούς κλάδους λειτουργίας, βασισμένους στην ίδια λογική, ένα για τη θερμοκρασία και ένα για τη συγκέντρωση.

- ✓ Πραγματοποιούνται δύο συγκρίσεις ανάμεσα στις απόλυτες τιμές των σφαλμάτων που προκύπτουν από τις δύο μεθόδους ρύθμισης. Η σύγκριση γίνεται μέσω αφαίρεσης του απόλυτου σφάλματος της δράσης του PI από το απόλυτο σφάλμα της συνεργασίας RL-PI. Αν το αποτέλεσμα της σύγκρισης είναι θετικό, που σημαίνει ότι η απόκλιση της συνεργασίας είναι μεγαλύτερη από εκείνο του PI, η συνεισφορά του RL κρίνεται αποτυχημένη και επιβαρύνεται η διαφορά με αρνητικό πολλαπλασιαστή (-10 για τη θερμοκρασία και -100 για τη συγκέντρωση).

Με την παραπάνω διαδικασία, ενθαρρύνονται ενέργειες από τον RL που οδηγούν σε πιο γρήγορη σύγκλιση των μεταβλητών εξόδου καθώς σκοπό αποτελεί μία συμπεριφορά καλύτερη από τον PI στο θέμα αποκλίσεων.



Εικόνα 12. Συνάρτηση ανταμοιβής προσαρμοσμένη στις πολλαπλές λύσεις του συστήματος.

Η συνάρτηση ανταμοιβής της εικόνας 12 αποτέλεσε προϊόν εκτενούς μελέτης της συμπεριφοράς του συστήματος και χτίστηκε μέσω προσαρμογών ύστερα από προβλήματα που ανέδειξε η εκπαίδευση.

Η βασική ιδέα σχεδιασμού είναι η γρήγορη επίτευξη της τιμής στόχου χωρίς ταλαντωτικές συμπεριφορές γεγονός που επιτυγχάνεται με την προσθήκη των εκθετικών όρων.

Αναλυτικότερα στον πίνακα 4 παρουσιάζεται επεξήγηση των σημάτων:

Πίνακας 4. Συνάρτηση ανταμοιβής για Ασταθείς τιμές

α/α Σημάτων	Επεξήγηση
Reactor Temp error1	Το σφάλμα της θερμοκρασίας, όπως προκύπτει από τις ενέργειες RL-PI.
Reactor Conc error1	Το σφάλμα της συγκέντρωσης, όπως προκύπτει από τις ενέργειες RL-PI.
Temp-1	Η θερμοκρασία, όπως προκύπτει από τις ενέργειες RL-PI για την προηγούμενη χρονική στιγμή.

Conc-1	Η συγκέντρωση, όπως προκύπτει από τις ενέργειες RL-PI για την προηγούμενη χρονική στιγμή.
Clock	Ο χρόνος της προσομοίωσης.

Εντοπίζονται τέσσερις βασικοί όροι στη δομή αυτής της συνάρτησης ανταμοιβής.

- ✓ Το τετράγωνο των σφαλμάτων της συγκέντρωσης και της θερμοκρασίας πολλαπλασιασμένο με μείον 1 και υψωμένο σε εκθετική δύναμη έτσι ώστε να επιτευχθούν ομαλές αλλαγές χωρίς ταλαντώσεις και αποκλίσεις και να κατευθυνθεί το σφάλμα πιο γρήγορα σε χαμηλές τιμές λόγω της ιδιότητας της συνάρτησης $\exp(-x^2)$ να δίνει τις μέγιστες τιμές καθώς οι τιμές του x πλησιάζουν στο 0 ενώ οι τιμές μειώνονται συμμετρικά καθώς οι τιμές του x απομακρύνονται από το 0. Έπειτα πολλαπλασιάζονται με θετικά βάρη (1.5 για τη θερμοκρασία και 0.7 για τη συγκέντρωση) και δίνεται ελαφρώς μεγαλύτερο βάρος στη θερμοκρασία καθώς εκεί παρατηρούνται πιο ακραίες μεταβολές.
- ✓ Λαμβάνονται οι τιμές των δύο προηγούμενων χρονικών στιγμών για τη θερμοκρασία και τη συγκέντρωση και υπολογίζεται η διαφορά τους. Η απόλυτη τιμή της τελευταίας επιβαρύνεται με αρνητικό πολλαπλασιαστή (-1.5 για τη θερμοκρασία και -0.7 για τη συγκέντρωση). Οι όροι αυτοί προσμετρώνται για χρόνους μεγαλύτερους των 5 μονάδων χρόνου και σκοπό έχουν να αποφεύγονται μεγάλες αποκλίσεις ανάμεσα σε επόμενη και προηγούμενη χρονική στιγμή ενώ όσο περισσότερο προσεγγίζεται το τέλος του επεισοδίου χρονικά, τόσο επιβαρύνεται αρνητικά η συνάρτηση ανταμοιβής για να αποφευχθούν αποκλίσεις στο τέλος της προσομοίωσης.
- ✓ Για να αποφευχθεί μία αρνητική τιμή στη συγκέντρωση, προστίθεται ένας όρος με μεγάλη αρνητική επιβάρυνση (-500) στη συνάρτηση, αν η συγκέντρωση πέσει κάτω από το 0.
- ✓ Ο τελευταίος όρος επικεντρώνεται στο σφάλμα της θερμοκρασίας όπου αποθαρρύνεται, με μεγάλο αρνητικό βάρος (-500), μία υπέρβαση της θερμοκρασίας για χρόνο μεγαλύτερο των 10 μονάδων χρόνου.

3.2. Εκπαίδευση αλγορίθμου

Ο αλγόριθμος εκπαιδεύτηκε με στόχο την βελτιστοποίηση του συστήματος ελέγχου αξιοποιώντας τις συναρτήσεις ανταμοιβής που σχεδιάστηκαν. Οι παράμετροι της εκπαίδευσης και τα κριτήρια τερματισμού αποδόθηκαν από τον αντίστοιχο κώδικα που επισυνάπτεται στο παράρτημα (Train Agent). Σε κάθε επεισόδιο εκπαίδευσης χρησιμοποιήθηκε διαφορετικό ζεύγος setpoint και το περιβάλλον ανανεώνεται με την ολοκλήρωση ή τον τερματισμό του επεισοδίου. Η επιλογή των setpoints γίνεται από την συνάρτηση Reset Function που αναφέρθηκε παραπάνω και επιλέγει τις τιμές στόχους από δύο ξεχωριστούς πίνακες, έναν που περιλαμβάνει όλες τις μοναδικές λύσεις και έναν που εμπεριέχει τις πολλαπλές λύσεις για τυχαία ζεύγη μεταβλητών εκ χειρισμού που επιλέχθηκαν.

Κριτήρια τερματισμού:

Ένας κρίσιμος παράγοντας κατά την εκπαίδευση του αλγορίθμου και της αξιολόγησης της συνάρτησης ανταμοιβής είναι η κατάλληλη επιλογή των συνθηκών τερματισμού, όπως ο μέγιστος αριθμός επεισοδίων ή ο στόχος για την τιμή ανταμοιβής. Με βάση τον σχεδιασμό, ο χρήστης καλείται να κάνει μια εκτίμηση για την μέση τιμή ανταμοιβής (average reward) στην οποία ο πράκτορας θα έχει πλήρως διαμορφώσει την πολιτική του, εκμεταλλεύοντας με κατάλληλο τρόπο το περιβάλλον του. Μέσα στο παράθυρο αυτό, ο αλγόριθμος θα πρέπει να έχει προλάβει να δει αρκετά διαφορετικά επεισόδια ώστε να είναι μπορεί να θεωρηθεί ικανοποιητικά ενημερωμένος για το περιβάλλον του. Με άλλα λόγια, θα πρέπει να μην είναι biased και να έχει μην εκπαιδευτεί μόνο πάνω σε εύκολα επεισόδια που δεν του δημιουργούν προκλήσεις.

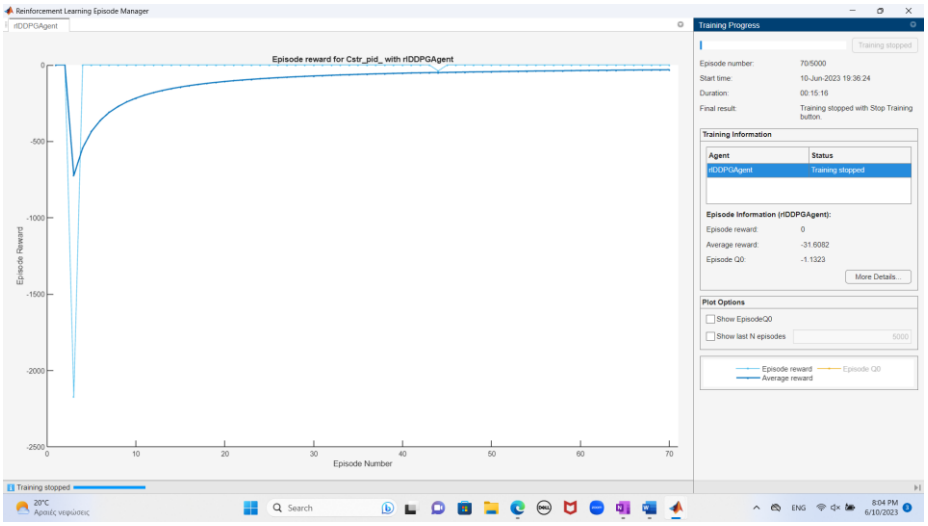
Αν η τιμή αυτή – Stop Training Value – είναι τόσο μεγάλη που στην πράξη δεν μπορεί να επιτευχθεί ο αλγόριθμος θα εξερευνά διαρκώς σενάρια χωρίς να έχει κάτι παραπάνω να αποκομίσει από αυτά και ενδέχεται μάλιστα να απορρίψει επιτυχημένες πολιτικές επειδή δεν ανταποκρίνονται στις ακραίες απαιτήσεις της εκπαίδευσης. Αντίστοιχα, πολύ μικρή τιμή θα τερματίσει σύντομα την εκπαίδευση με

αποτέλεσμα ο πράκτορας να έχει πολύ μικρή εικόνα για το περιβάλλον και τις απαιτήσεις του.

Συνεπώς, σημαντικό κομμάτι της εκπαίδευσης, ίσως το ίδιο κρίσιμο με τον σχεδιασμό της συνάρτησης ανταμοιβής, αποτελεί η εύρεση της χρυσής τομής ανάμεσα στο σωστό αριθμό επεισοδίων και την συνθήκη τερματισμού της εκπαίδευσης με βάση την συνολική ανταμοιβή.

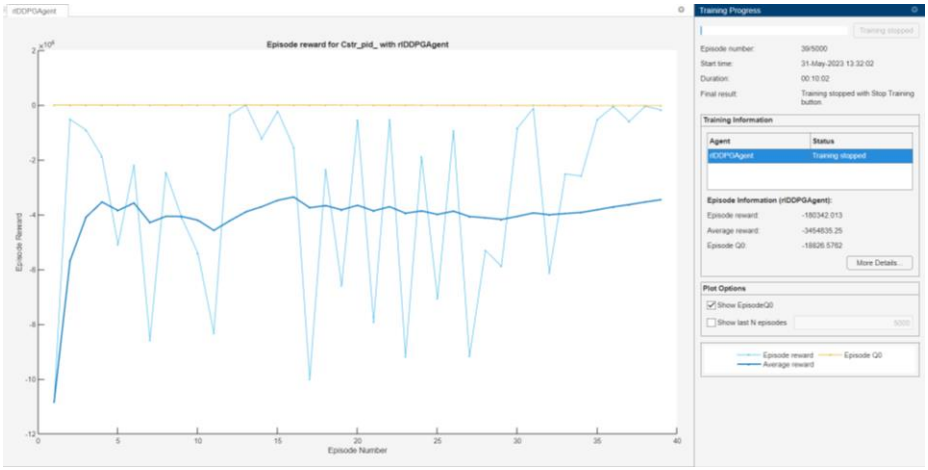
Επιπλέον, είναι στην ευχέρεια του χρήστη να επέμβει και να τερματίσει χειροκίνητα την εκπαίδευση αν βάσει της εμπειρίας του και της παρακολούθησης της πορείας των επεισοδίων αντιλαμβάνεται πως έχει ήδη βρεθεί μια καλή πολιτική. Αυτό συνέβη και στις δύο περιπτώσεις που εξετάστηκαν καθώς η συμπεριφορά του συστήματος ήταν ήδη αρκετά ικανοποιητική [3].

Η εκπαίδευση που πραγματοποιήθηκε για την επίτευξη των μονών λύσεων φαίνεται στο διάγραμμα 14 και για τις τριπλές λύσεις στο διάγραμμα 15.



Διάγραμμα 14. Η επιβράβευση για τα επεισόδια που πραγματοποιήθηκαν για την επίτευξη των μονών λύσεων.

[3] Ο αριθμός των επεισοδίων μπορεί ορισμένες φορές να είναι αρκετά μεγάλος και παρόλα αυτά η εκπαίδευση να μην έχει θετική έκβαση. Μπορεί ωστόσο το σύστημα να μην έχει εκπαιδευτεί για μεγάλο αριθμό επεισοδίων και να έχει ανακαλύψει τη σωστή πολιτική.



Διάγραμμα 15. Η επιβράβευση για τα επεισόδια που πραγματοποιήθηκαν για την επίτευξη των τριπλών λύσεων.

Όπως είναι φανερό από τα διαγράμματα 15 και 16, με μία καλή συνάρτηση ανταμοιβής δεν είναι απαραίτητος ο μεγάλος υπολογιστικός χρόνος καθώς το σύστημα, ήδη από τα πρώτα επεισόδια, αντιλαμβάνεται ποια είναι η σωστή πολιτική που πρέπει να ακολουθήσει έτσι ώστε να πετύχει αποδοτικά τις τιμές – στόχους.

Στους πίνακες 5 και 6 παρουσιάζονται τα βήματα και τα επεισόδια που εκτελέστηκαν αναλυτικά για κάθε πακέτο τιμών, η μέση επιβράβευση και η τιμή της Q-value.

Πίνακας 5. Πληροφορίες εκπαίδευσης επεισοδίων μονών λύσεων.

	rlDDPGAgent
Status	Training stopped
Episode number	70
Episode reward	0
Episode steps	1000
Total agent steps	70000
Average reward	-31.6082
Average steps	1000
Episode Q0	-1.1323
Averaging window length	100
Training stopped by	Stop Training button
Training stopped at	Episode 70

Πίνακας 6. Πληροφορίες εκπαίδευσης επεισοδίων τριπλών λύσεων.

	rlDDPGAgent
Status	Training stopped
Episode number	39
Episode reward	-180342.013
Episode steps	1000
Total agent steps	39000
Average reward	-3454835.25
Average steps	1000
Episode Q0	-18826.5762
Averaging window length	100
Training stopped by	Stop Training button
Training stopped at	Episode 39

3.3. Ανάλυση αποτελεσμάτων

Οι δύο βασικότεροι στόχοι ενός συστήματος ελέγχου είναι:

- **Παρακολούθηση επιθυμητών τιμών (Set-point tracking):** η προσέγγιση των τιμών στόχων που καθορίζονται από τον χρήστη με μείωση του υπολογιστικού κόστους και του χρόνου σύγκλισης
- **Απόρριψη διαταραχών (Disturbance rejection):** η απόρριψη της επίδρασης τυχόν διαταραχών στο σύστημα που μπορεί να αφορούν σχεδιαστικές παραμέτρους του συστήματος που πρέπει να παραμένουν σταθερές όπως η θερμοκρασία και η συγκέντρωση του ρεύματος τροφοδοσίας.

Για την αξιολόγηση των πρακτόρων εξετάζεται η απόδοση τους στους δυο αυτούς στόχους [21].

3.3.1. Αποτελέσματα για μοναδικές μόνιμες καταστάσεις

Οι μοναδικές λύσεις του συστήματος οδηγούν κατά κύριο λόγο σε ευσταθείς μόνιμες καταστάσεις. Για τις περιπτώσεις αυτές ύστερα από πολλές δοκιμές διαμορφώθηκε τελικά η συνάρτηση ανταμοιβής της εικόνας 11. Για την αξιολόγηση της γίνεται έλεγχος σχετικά με:

1) Την βελτίωση της συμπεριφοράς του συνεργατικού συστήματος σε σχέση με το αρχικό σύστημα των PI μέσω καλύτερων χρόνων και πιο ομαλών μεταβάσεων.
2) Την ανταπόκριση του συστήματος σε παρουσία διαταραχών τύπου ±5.7% για τη θερμοκρασία και ±5% για την συγκέντρωση.

Τα αποτελέσματα συγκεντρώνονται σε διαγράμματα όπου αποτυπώνουν την συμπεριφορά του συνεργατικού συστήματος RL-PI καθώς και εκείνη του αρχικού συστήματος με τους PI για λόγος σύγκρισης. Παρουσιάζονται τρεις διαφορετικές εκδοχές για την περίπτωση 1 για τις μοναδικές λύσεις της εξίσωσης 2.10 και τρεις για τις πολλαπλές λύσεις της εξίσωσης έτσι ώστε να καλύπτεται αρκετά το φάσμα των επιθυμητών συγκεντρώσεων εξόδου ως προς το αντιδρών ως προς το οποίο επιλέγεται η ανάλυση.

Ακόμη, περιλαμβάνονται ορισμένα αποτελέσματα από τις μεταβλητές εκ χειρισμού έτσι ώστε να γίνει αντιληπτή η συνεισφορά του RL και στην ενεργειακή κατανάλωση του συστήματος συγκριτικά πάντα με τον παραδοσιακό σύστημα ρύθμισης.[4]

[4] Ο RL εκπαιδεύτηκε για χρόνο ίσο με 100 μονάδες χρόνου, παρόλα αυτά σε κάποιες περιπτώσεις, τα αποτελέσματα είναι εμφανή στο μισό χρόνο και έτσι ο χρόνος προσομοίωσης ελαττώνεται στο μισό.

3.3.1.1. Αποτελέσματα για set-point tracking

Τα επεισόδια που παρουσιάζονται καλύπτουν ένα μεγάλο εύρος των συγκεντρώσεων και θερμοκρασιών και για την περίπτωση του set-point tracking στις μοναδικές λύσεις συνοψίζονται οι περιγραφές τους στον ακόλουθο πίνακα 7.

Πίνακας 7. Μοναδικές Μόνιμες Καταστάσεις.

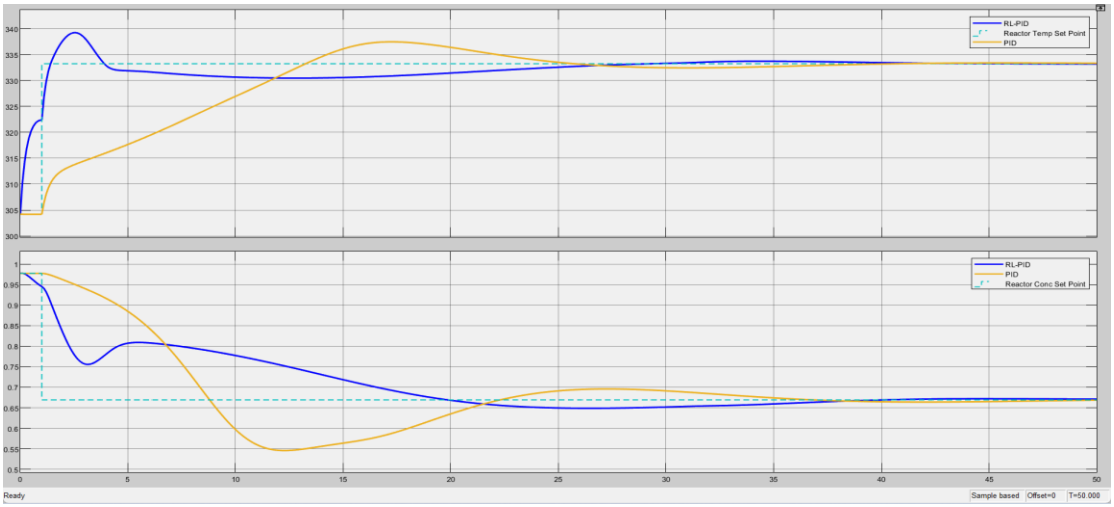
α/α Διαγράμματος	α/α Επεισοδίου	Θερμοκρασία Sp (K)	Συγκέντρωση Sp (mol/m^3)
16	62	333.23	0.66902
17	1	414.20	0.016812
18	1500	488.92	0.0010121
19	62	333.23	0.66902
20	1	414.20	0.016812
21	62	333.23	0.66902
22	1500	488.92	0.0010121

Στο διάγραμμα 16, η συμπεριφορά του RL είναι σημαντικά καλύτερη από εκείνη των PI. Ειδικά για τη ρύθμιση της συγκέντρωσης, ο RL αποφεύγει η υπέρβαση των PIs και συγκλίνει πιο γρήγορα. Αντίστοιχα, θεωρείται επιτυχής και ο έλεγχος της θερμοκρασίας που είναι αρκετά πιο γρήγορος αλλά περισσότερο επιθετικός.

Για το συγκεκριμένο επεισόδιο, υπολογίζονται οι χρόνοι που απαιτούνται έτσι ώστε το σύστημα να φτάσει στο 2% της επιθυμητής τιμής καθώς και οι μέγιστες αποκλίσεις που παρατηρούνται. Τα αποτελέσματα παρουσιάζονται στον πίνακα 8 μόνο για τη θερμοκρασία καθώς η συγκέντρωση έχει παρόμοια συμπεριφορά.

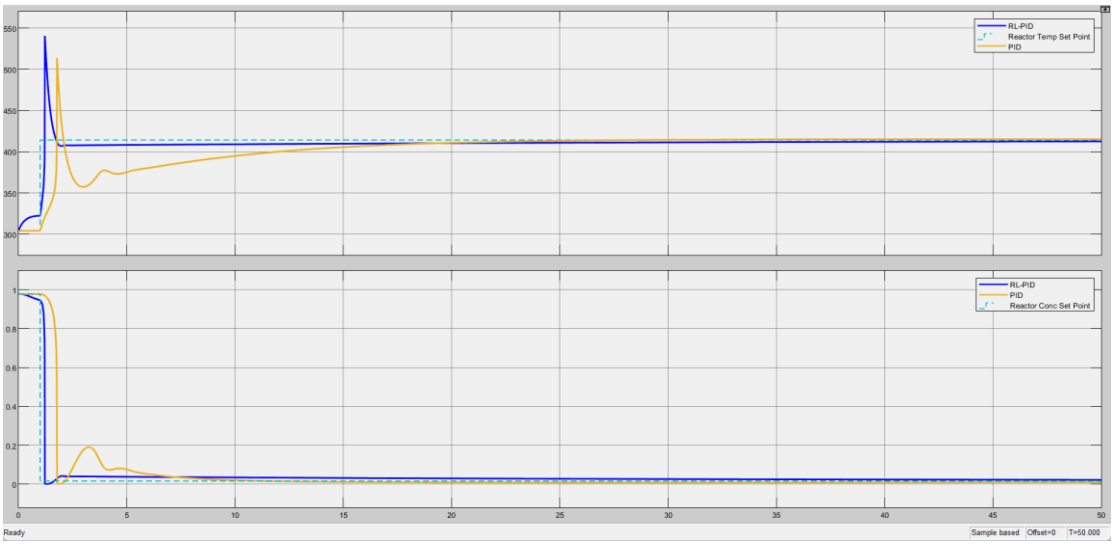
Πίνακας 8. Αποτελέσματα υπέρβασης και χρόνος για επίτευξη 2% της επιθυμητής τιμής.

α/α	Χρόνος υπέρβασης (RL)	Υπέρβαση (RL)	Χρόνος υπέρβασης (PI)	Υπέρβαση (PI)	Χρόνος για 2% (RL)-MX	Χρόνος για 2% (PI)-MX
Διάγραμμα 18	2.5000	339.23	17.200	337.46	1.0940	9.7987



Διάγραμμα 16. Σενάριο σύγκρισης RL με PI (ευσταθής τιμή)

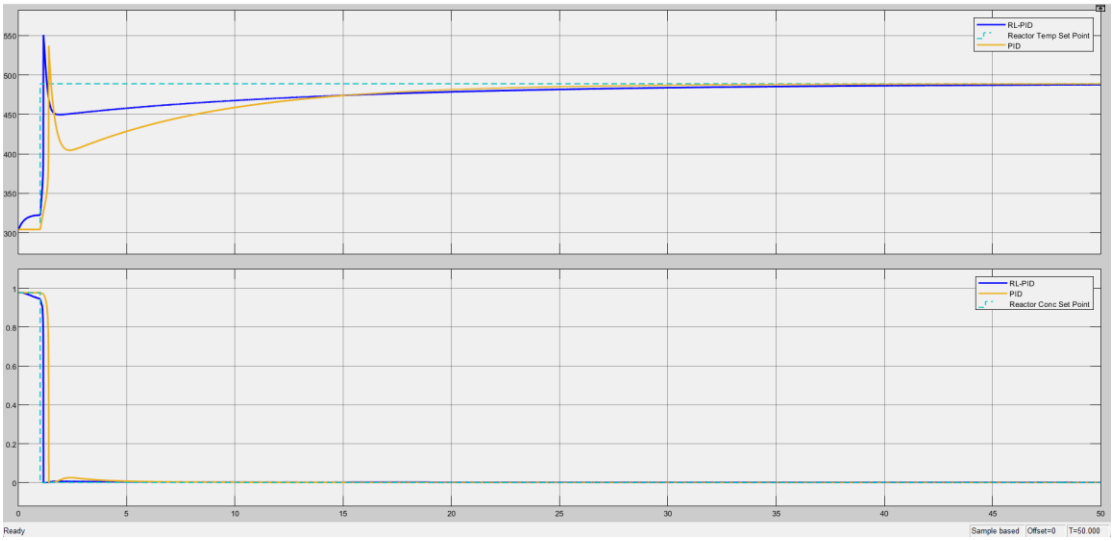
Για τα υπόλοιπα διαγράμματα της κατηγορίας, ισχύουν παρόμοια αποτελέσματα όσον αφορά τους χρόνους που απαιτούνται για την επίτευξη του 2% και έτσι δεν παρουσιάζονται τα αποτελέσματα.



Διάγραμμα 17. Σενάριο σύγκρισης RL με PI (ευσταθής τιμή)

Στο διάγραμμα 17, φαίνεται πως ο RL καταφεύγει σε μια πιο επιθετική προσέγγιση για να επιταχύνει την σύγκλιση και τελικά το καταφέρνει σε λιγότερο από πέντε μονάδες χρόνου.

Όσον αφορά τη συγκέντρωση, ο έλεγχος της είναι σημαντικά καλύτερος καθώς όχι μόνο προσεγγίζει την τιμή-στόχο σχεδόν αμέσως αλλά αποφεύγει κιόλας την υπέρβαση που κάνει το σύστημα των PI προτού τελικά συγκλίνει.



Διάγραμμα 18. Σενάριο σύγκρισης του RL με τον PI (ευσταθής τιμή)

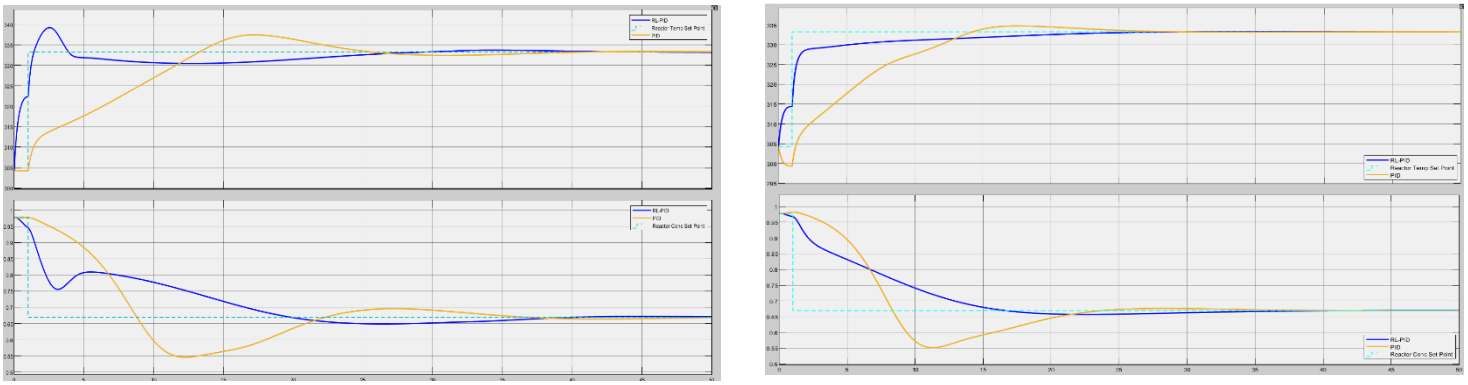
Παρόμοια αποτελέσματα προκύπτουν και στο διάγραμμα 18 όπου ο RL είναι πιο γρήγορος με κόστος μια πιο επιθετική πολιτική ρύθμισης.

Παρατηρείται ότι για μεγάλες τιμές της θερμοκρασίας που συνοδεύονται από πολύ μικρές συγκεντρώσεις καθώς όλο το αντιδρών έχει σχεδόν μετατραπεί σε προϊόν η σύγκλιση επιτυγχάνεται σχετικά εύκολα και γρήγορα και για τα δυο συστήματα ρύθμισης. Αυτές οι καταστάσεις ευνοούνται και θερμοδυναμικά και έτσι το σύστημα τείνει να σταθεροποιηθεί από μόνο του σε αυτές τις περιοχές χαμηλής συγκέντρωσης. Κατά ανάλογο τρόπο, θερμοδυναμικά ευνοούνται και οι συγκεντρώσεις που βρίσκονται πολύ κοντά στο 1 και συνδυάζονται με μικρές τιμές θερμοκρασιών. Αυτό συμβαίνει διότι οι καταστάσεις αυτές βρίσκονται πολύ κοντά στην αρχική κατάσταση του συστήματος και δεν απαιτούν ιδιαίτερη προσπάθεια από το σύνολο των ρυθμιστών για να μετακινήσουν το σύστημα. Στις περιπτώσεις που αναφέρθηκαν παραπάνω, είναι δύσκολο να διευκρινιστεί ποιο σύστημα ελέγχου είναι περισσότερο αποδοτικό αφού η σύγκλιση είναι ήδη πολύ γρήγορη. Ο RL στην

προσπάθεια του να βελτιώσει ακόμα περισσότερο την ταχύτητα ελέγχου αναγκάζεται να καταφύγει σε πιο επιθετικά σήματα ρύθμισης.

3.3.1.2. Αποτελέσματα για disturbance rejection στη θερμοκρασία

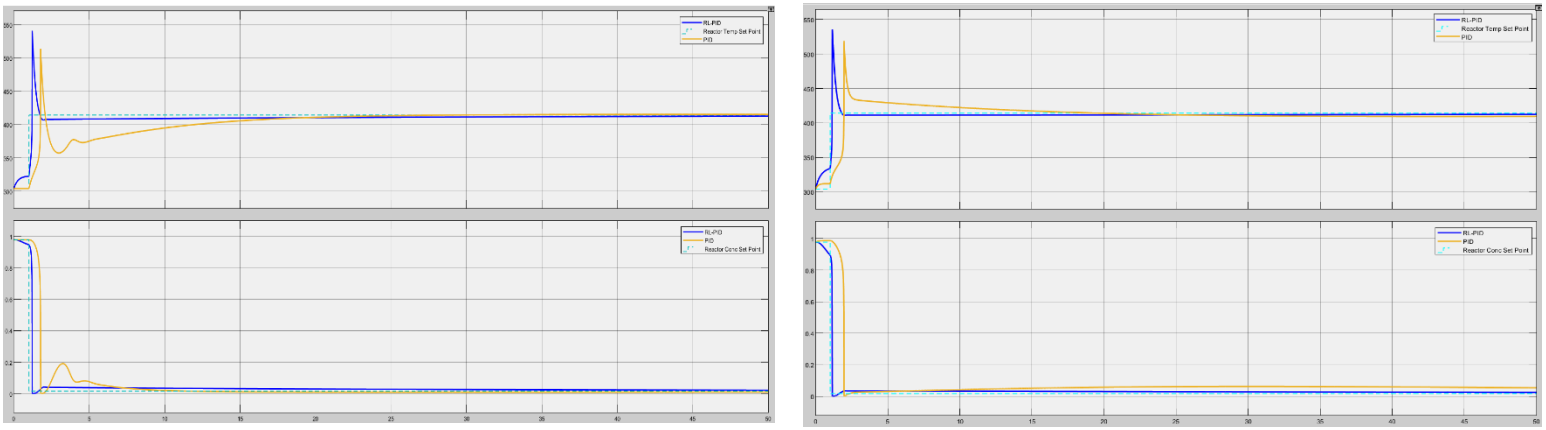
Στο σημείο αυτό, εξετάζεται η συμπεριφορά του συστήματος ρύθμισης όταν για αδιευκρίνιστες συνθήκες μεταβληθεί κάποια από τις σχεδιαστικές παραμέτρους. Για παράδειγμα, θεωρείται διαταραχή της θερμοκρασίας τροφοδοσίας με μείωση κατά 20 βαθμούς K. Τα αποτελέσματα παρουσιάζονται στο διάγραμμα 19 (αριστερά πριν τη διαταραχή και δεξιά μετά) σύμφωνα με τον πίνακα 7.



Διάγραμμα 19. Διαταραχή μείον 20 (-20) βαθμούς K στη θερμοκρασία παροχής (ευσταθής τιμή).

Σημαντικά καλύτερη παρατηρείται να είναι η συμπεριφορά του RL συγκριτικά με εκείνη του PI ακόμα και υπό την επίδραση διαταραχών, παρόλο που η μελέτη αυτών δεν είχε συμπεριληφθεί στην εκπαίδευση του. Προσεγγίζεται σαφώς καλύτερα η επιθυμητή τιμή στη θερμοκρασία ενώ παρατηρείται και ομαλότερη προσέγγιση της συγκέντρωσης.

Με αντίστοιχο τρόπο, δίνεται διαταραχή συν 20/ +20 βαθμοί K στην θερμοκρασία τροφοδοσίας και τα αποτελέσματα φαίνονται στο διάγραμμα 20 (αριστερά πριν τη διαταραχή και δεξιά μετά) (πίνακας 7).



Διάγραμμα 20. Διαταραχή συν 20 (+20) βαθμούς K στη θερμοκρασία παροχής (ευσταθής τιμή)

Όπως και προηγουμένως, έτσι και εδώ ο RL αποκρίνεται πολύ αποδοτικά ενώ ο PI παρουσιάζει ελαφρώς χειρότερη συμπεριφορά με την παρουσία των διαταραχών καθώς, όσον αφορά την συγκέντρωση, αποκλίνει.

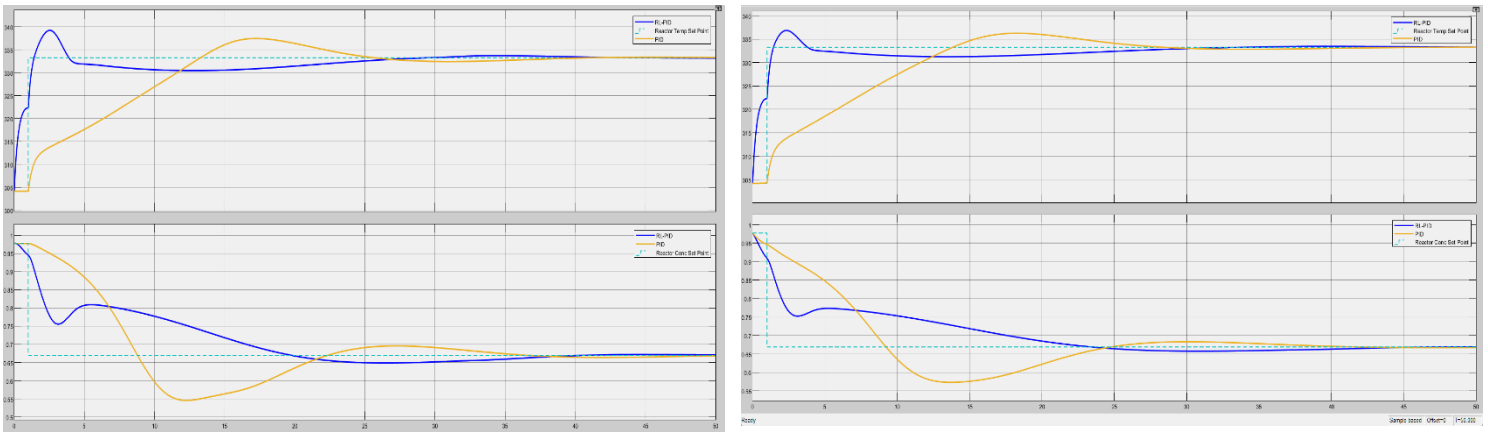
Αυτό συμβαίνει γιατί ο ελεγκτής PI λειτουργεί εντός περιορισμένου εύρους ελέγχου που ορίζεται από τις δυνατότητες εξόδου του. Εάν η θερμοκρασία διαταραχής υπερβαίνει αυτό το εύρος ελέγχου, ο ελεγκτής PI ενδέχεται να μην είναι σε θέση να παράγει σήμα ελέγχου αρκετά ισχυρό ώστε να εξουδετερώνει αποτελεσματικά τη διαταραχή. Στις περιπτώσεις αυτές, ο ελεγκτής μπορεί να φθάσει τα μέγιστα ή ελάχιστα όρια εξόδου του, περιορίζοντας την ικανότητά του να ελέγχει το σύστημα.

Αντιθέτως, παρόλο που ήταν πολύ πιθανό και ο RL πράκτορας να αποτύγχανε σε αντίστοιχη περίπτωση, τα καταφέρνει χάρη στη γενίκευση των εμπειριών που απέκτησε κατά τη διάρκεια της εκπαίδευσης και την συνάρτηση ανταμοιβής που διαμόρφωσε κατάλληλα στην στρατηγική ελέγχου. Τέλος, η εξερεύνηση κατά την διάρκεια των επεισοδίων προσφέρει μελέτη της δυναμικής του συστήματος και

αυξάνει την προσαρμοστικότητα του σε καταστάσεις εκτός των αρχικών ορίων εκπαίδευσης.

3.3.1.3. Αποτελέσματα εκπαίδευσης για disturbance rejection στη συγκέντρωση

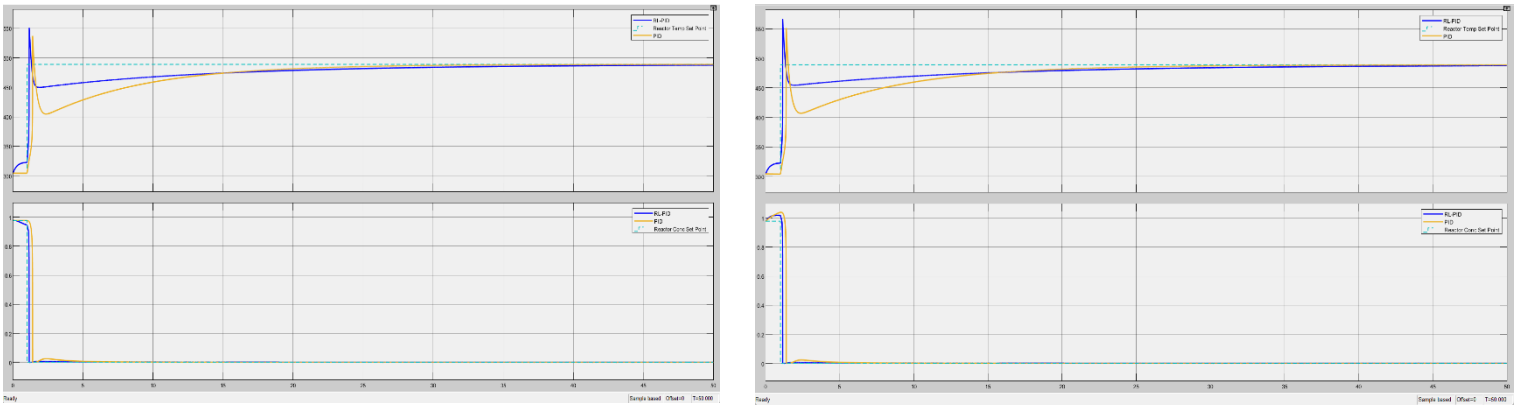
Στο σημείο αυτό, εξετάζεται η επίδραση των διαταραχών στην συγκέντρωση της τροφοδοσίας. Στην πρώτη περίπτωση, δίνεται διαταραχή στη συγκέντρωση εισόδου ίση με μείον 0.05/ -0.05 mol/m^3 και τα αποτελέσματα παρουσιάζονται στο διάγραμμα 21 (αριστερά πριν τη διαταραχή και δεξιά μετά) (πίνακας 7).



Διάγραμμα 21. Διαταραχή μείον 0.05 (-0.05) mol/m^3 στη συγκέντρωση εισόδου (ευσταθής τιμή)

Φαίνεται πως η διαταραχή στη συγκέντρωση όταν αυτή είναι αρνητική, έχει επίδραση στην καλύτερη συμπεριφορά του PI ενώ ο RL δεν επηρεάζεται.

Με ανάλογα τρόπο, για το επεισόδιο επιβάλλεται θετική διαταραχή ίσου μεγέθους (+0.05 mol/m^3) στη συγκέντρωση και τα αποτελέσματα φαίνονται στο διάγραμμα 22 του (Πίνακας 7).



Διάγραμμα 22. Σενάριο διαταραχής συν 0.05 (+0.05) mol/m³ στη συγκέντρωση εισόδου για την σύγκριση RL με PI (ευσταθής τιμή)

Παρόμοια με τις υπόλοιπες περιπτώσεις, η συμπεριφορά του RL δεν μεταβάλλεται σημαντικά, ενώ φαίνεται να ακολουθεί την συμπεριφορά της επιθυμητής τιμής χωρίς σημαντική απομάκρυνση από εκείνη στις πρώτες μονάδες χρόνου.

3.3.2. Αποτελέσματα για πολλαπλές μόνιμες καταστάσεις

Το συγκεκριμένο κομμάτι του σχεδιασμού αποτέλεσε την μεγαλύτερη πρόκληση στα πλαίσια της διπλωματικής εργασίας. Το σύστημα αδυνατούσε να σταθεροποιηθεί σε αυτές τις καταστάσεις και οι αποκρίσεις εμφάνιζαν έντονη ταλαντωτική συμπεριφορά και υψηλές αποκλίσεις, ενώ οι μεταβλητές εκ χειρισμού καθιστούσαν αδύνατο αυτό το σύστημα να λειτουργήσει εκτός των ορίων προσομοίωσης εξαιτίας των έντονων ταλαντωτικών συμπεριφορών που είναι απαγορευτικές για κλίμακα βιομηχανίας.

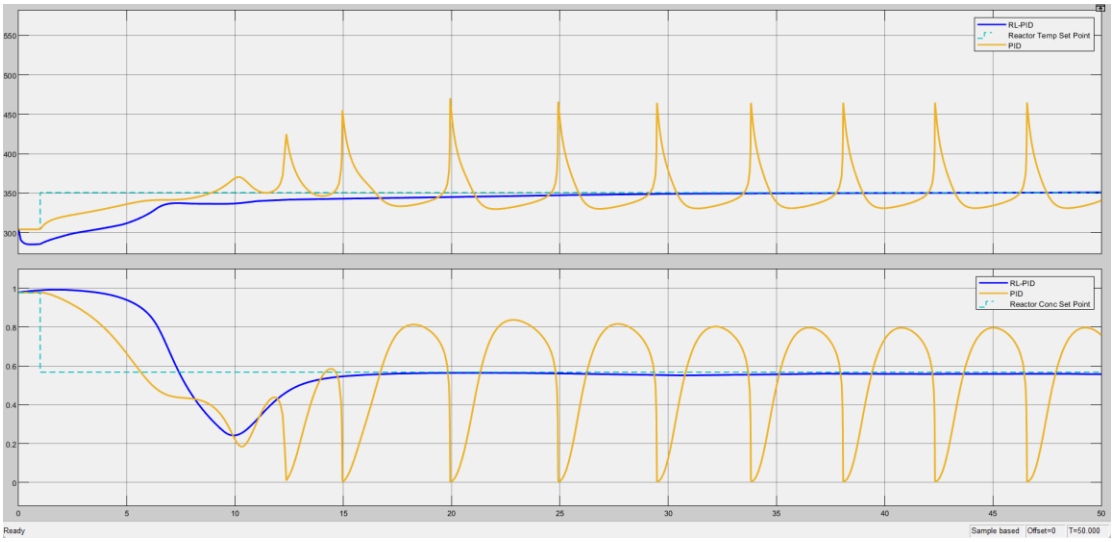
Καθώς ο RL λειτουργεί επικουρικά και διορθώνει την συμπεριφορά του PI, επηρεάζεται πολύ άμεσα από τα σφάλματα εκείνου και οδηγείται σε περιοχές με υψηλή αστάθεια. Όσο χειρότερη είναι η συμπεριφορά των PI, τόσο μεγαλύτερη προσπάθεια χρειάζεται να καταβάλλει ο RL για να αντισταθμίσει αυτή την απόδοση και να σταθεροποιήσει το σύστημα. Η διαμόρφωση της συνάρτησης ανταμοιβής

μέσω της προσθήκης κατάλληλων όρων και σωστά ζυγισμένων βαρών είναι ύψιστης σημασίας ενώ θα πρέπει να δοθεί και ιδιαίτερη προσοχή στον αριθμό επεισοδίων και στις συνθήκες τερματισμού.

3.3.2.1. Αποτελέσματα της εκπαίδευσης για set-point tracking

Πίνακας 9. Πολλαπλές Μόνιμες Καταστάσεις.

α/α Διαγράμματος	α/α Επεισοδίου	Θερμοκρασία Sp (K)	Συγκέντρωση Sp (mol/m^3)
23	11	350.71	0.56815
24	1623	353.91	0.43001
25	1705	348.15	0.61074
26	1705	348.15	0.61074
27	1623	353.91	0.43001
28	1623	353.91	0.43001
29	1623	353.91	0.43001



Διάγραμμα 23. Σενάριο σύγκρισης RL με PI (ασταθής τιμή)

Η σύγκριση της απόδοσης των δυο συστημάτων είναι σαφώς πιο εύκολη σε αυτές τις περιπτώσεις. Για παράδειγμα στο διάγραμμα 23 είναι έκδηλη η απαγορευτική συμπεριφορά του PI που δεν παρουσιάζει καμία τάση σύγκλισης και παγιδεύεται σε ταλάντωση σταθερού πλάτους.

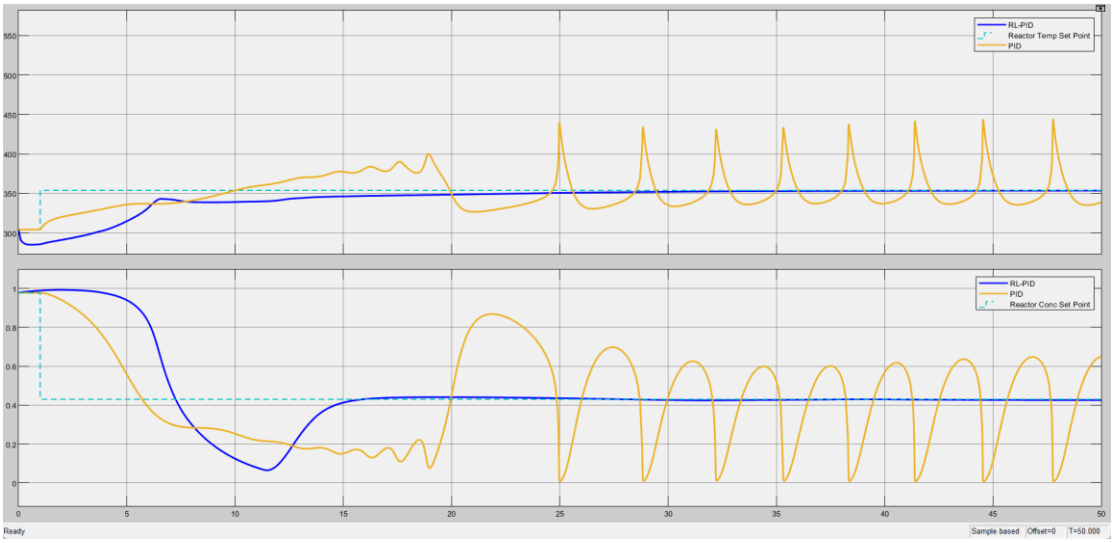
Παρόλα αυτά ο RL αντισταθμίζει την άσχημη αυτή συμπεριφορά και προσεγγίζει την επιθυμητή τιμή αποφεύγοντας με επιτυχία τις περιοχές αστάθειας. Αυτό καταδεικνύει τον ορθό σχεδιασμό της συνάρτησης ανταμοιβής και την προσαρμοστικότητα του αλγορίθμου RL ο οποίος μέσω αλληλεπίδρασης με το περιβάλλον «μαθαίνει» ποιες περιοχές πρέπει να αποφύγει αφού εκεί γνωρίζει θα τιμωρηθεί μέσω των penalty που είναι ενσωματωμένα στη συνάρτηση ανταμοιβής.

Με ανάλογο τρόπο με προηγουμένως, έτσι και για το διάγραμμα 23 υπολογίζονται οι χρόνοι που απαιτούνται έτσι ώστε το σύστημα υπό την επιρροή του ελέγχου να φτάσει στο 2% της επιθυμητής τιμής. Ο υπολογισμός της υπέρβασης κρίνεται άσκοπος καθώς ο RL προσεγγίζει την επιθυμητή τιμή δίχως να την ξεπεράσει και η υπέρβαση είναι μηδενική, ενώ για τον PI το overshoot είναι το πλάτος της ταλάντωσης. Τα αποτελέσματα παρουσιάζονται στον πίνακα 10 μόνο για τη θερμοκρασία καθώς η συγκέντρωση έχει παρόμοια συμπεριφορά.

Πίνακας 10. Αποτελέσματα χρόνου για επίτευξη 2% της επιθυμητής τιμής.

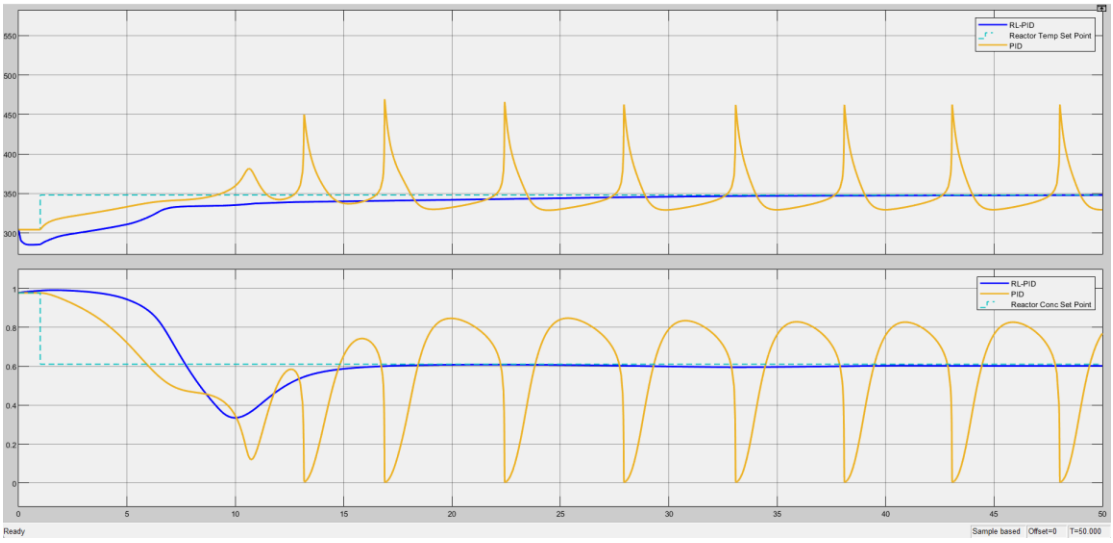
α/α	Χρόνος για 2% (RL)-MX	Χρόνος για 2% (PI)-MX
Διάγραμμα 26	16.636	Inf

Είναι ξεκάθαρη η υπεροχή του RL σε σχέση με τους PI αφού σε διάστημα λιγότερο από 17 μονάδες χρόνου έχει σταθεροποιήσει το σύστημα, κάτι που οι PI δεν καταφέρνουν ποτέ.



Διάγραμμα 24. Σενάριο σύγκρισης RL με PI (ασταθής τιμή)

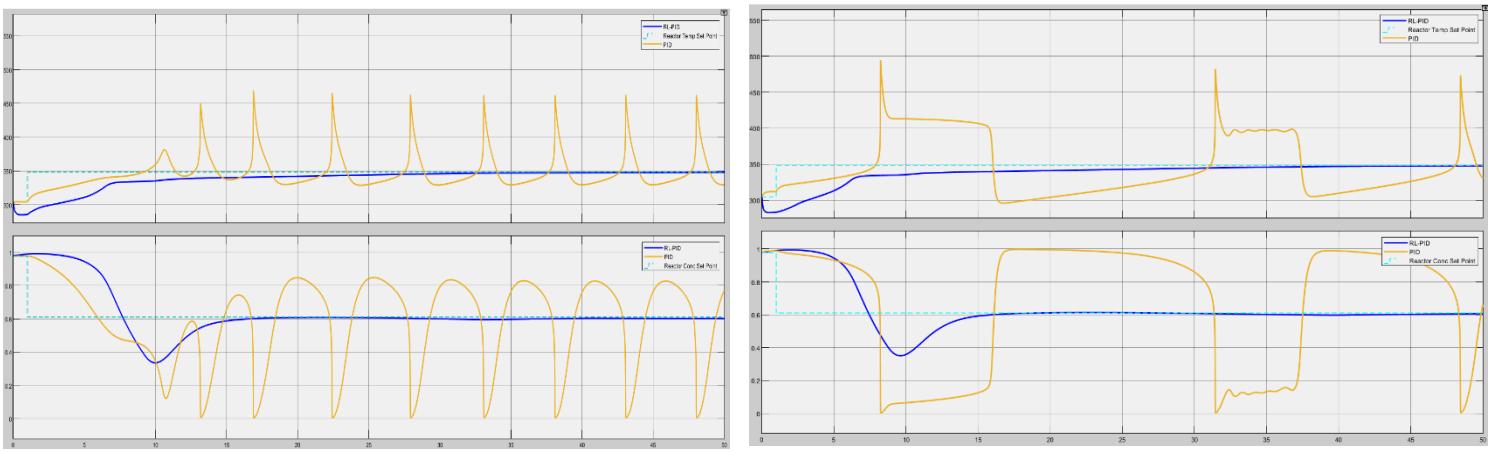
Παρόμοια με το διάγραμμα 23, στο διάγραμμα 24, ο RL συγκλίνει πολύ αποδοτικά ενώ η συμπεριφορά του PI είναι έντονα ταλαντωτική με μεγάλο εύρος ταλάντωσης όπως ακριβώς και στο διάγραμμα 25. Έτσι φαίνεται πως για τα ενδιάμεσα εύρη συγκεντρώσεων και στα οποία είναι πιο πιθανό να παρουσιαστούν οι αστάθειες ο συγκεκριμένος RL λειτουργεί καλύτερα από ότι ο PI και καταφέρνει να συγκλίνει το σύστημα και να γίνει εφικτό και πρακτικά.



Διάγραμμα 25. Σενάριο σύγκρισης RL με PI (ασταθής τιμή)

3.3.2.2. Αποτελέσματα της εκπαίδευσης για disturbance rejection στη θερμοκρασία

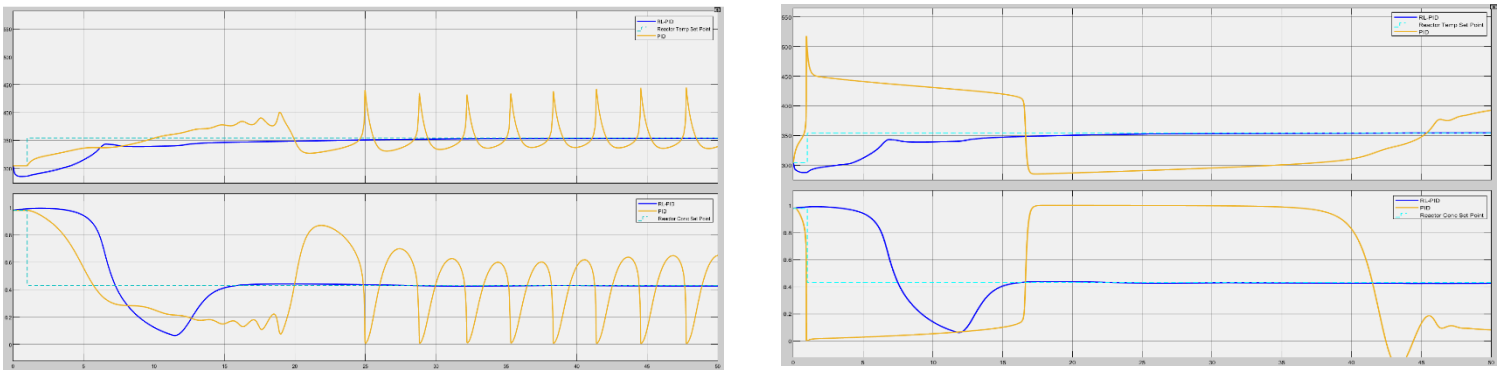
Επιβάλλεται διαταραχή μείον 20/-20 βαθμών K στη θερμοκρασία εισόδου και τα αποτελέσματα φαίνονται στο διάγραμμα 26 (αριστέρα πριν τη διαταραχή, δεξία μετά) (πίνακας 9).



Διάγραμμα 26. Επίδραση διαταραχής μείον 20 (-20) βαθμών K στη θερμοκρασία (ασταθής τιμή)

Ακόμα και υπό την επίδραση διαταραχής που ο RL δεν έχει εκπαιδευτεί να διαχειρίζεται, παρουσιάζει συμπεριφορά ιδανική και πρακτικά λειτουργική.

Με διαταραχή συν 20 (+20) βαθμούς K στη θερμοκρασία τα αποτελέσματα φαίνονται στο διάγραμμα 27 (αριστερά πριν τη διαταραχή και δεξιά μετά) (πίνακας 9).

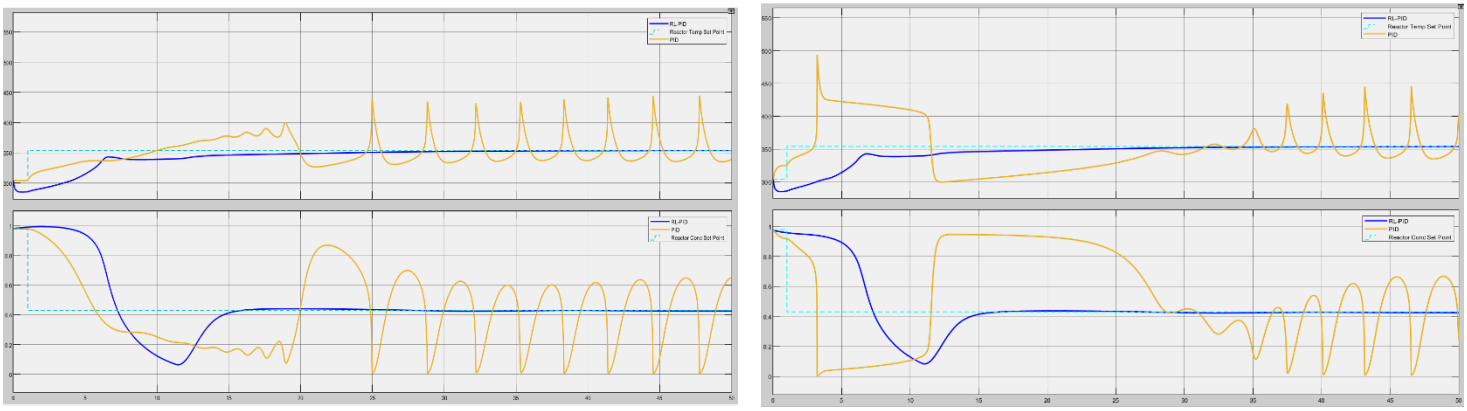


Διάγραμμα 27. Διαταραχή συν 20 (+20) βαθμών Κ στην θερμοκρασία (ασταθής τιμή)

3.3.2.3. Αποτελέσματα εκπαίδευσης για disturbance rejection στη συγκέντρωση

Τέλος, εξετάζεται και η επίδραση των διαταραχών στην συγκέντρωση.

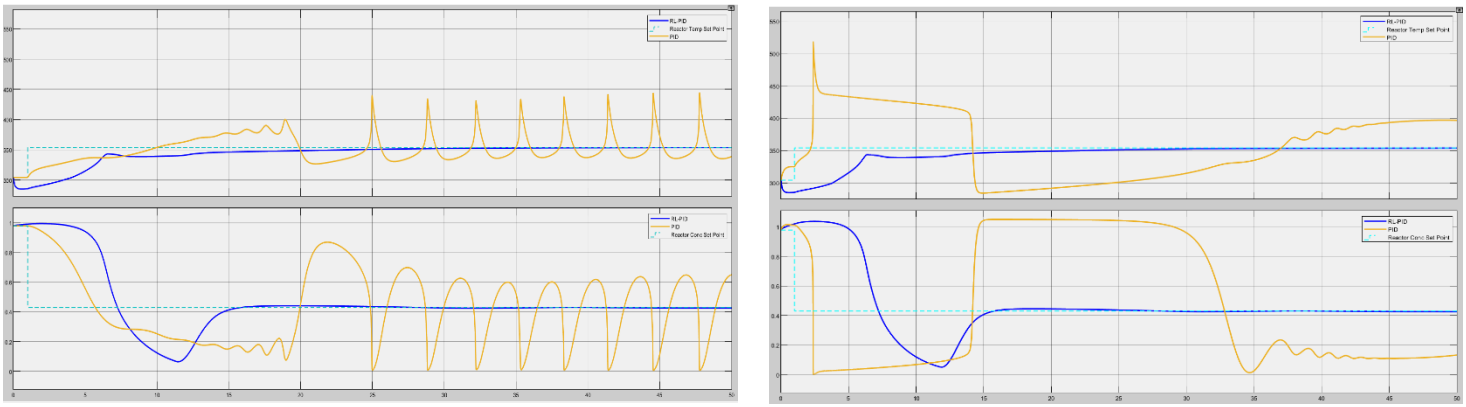
Αρχικά, δίνεται μεταβολή μείον 0.05/ -0.05 mol/m^3 στη συγκέντρωση εισόδου και τα αποτελέσματα δίνονται στο διάγραμμα 28 (αριστερά πριν τη διαταραχή και δεξιά μετά) (πίνακας 9).



Διάγραμμα 28. Διαταραχή μείον 0.05 (-0.05) mol/m^3 στη συγκέντρωση εισόδου (ασταθής τιμή)

Η υπεροχή του συστήματος με τον RL είναι ξεκάθαρη και σε αυτό το διάγραμμα αφού με παρόμοια συμπεριφορά και κόστος μόνο ένα undershoot στη συγκέντρωση επιφέρει τελικά τη σύγκλιση και των δύο μεταβλητών.

Αντίστοιχα πραγματοποιείται θετική διαταραχή πάλι στη συγκέντρωση εισόδου ίση με συν 0.05/ +0.05 mol/m^3 και τα αποτελέσματα παρουσιάζονται στο διάγραμμα 29 (αριστερά πριν τη διαταραχή και δεξιά μετά) (πίνακας 9). Τα αποτελέσματα επιβεβαιώνουν την υπεροχή του RL έναντι του κλασικού συστήματος ρύθμισης όπως είναι αυτό βαθμονομημένο.



Διάγραμμα 29. Διαταραχή συν 0.05 (+0.05) mol/m^3 στη συγκέντρωση εισόδου (ασταθής τιμή)

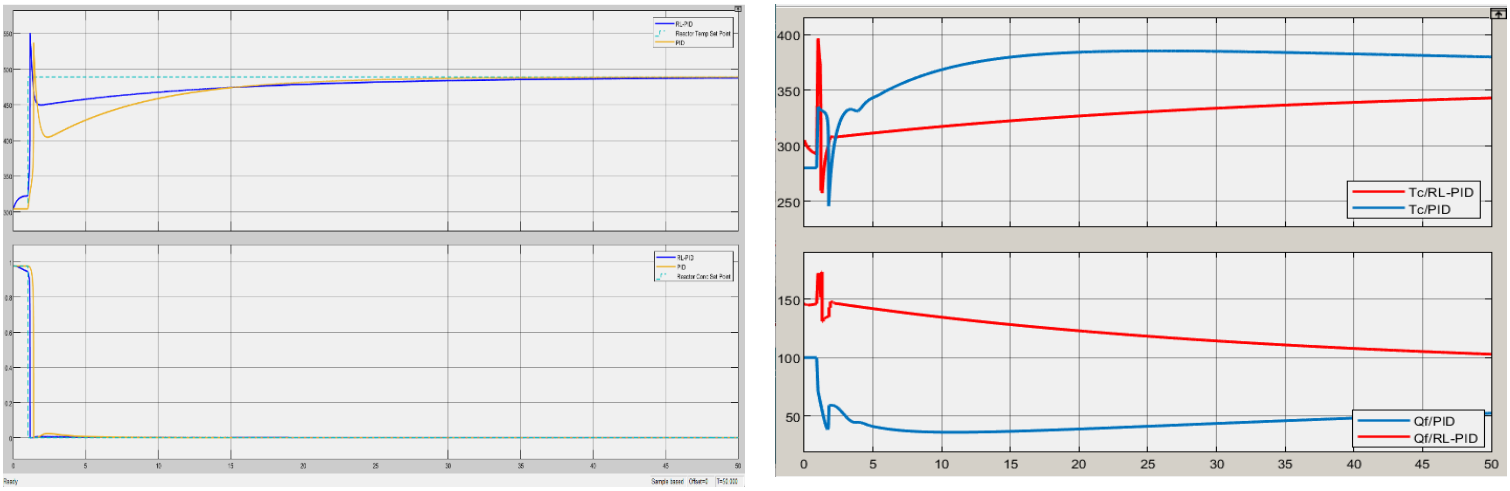
Το εγχείρημα της ρύθμισης του συστήματος στις περιοχές των ασταθών μόνιμων καταστάσεων κρίνεται επιτυχές αφού μέσω του RL δίνεται λύση στο πρόβλημα που δεν μπορούσαν να αντιμετωπίσουν οι PI. Ο πράκτορας καταφέρνει τελικά να βρει τις κατάλληλες ακολουθίες σημάτων ελέγχου που δύνανται να σταθεροποιήσουν το σύστημα στις επιθυμητές τιμές-στόχους.

3.3.3. Παρουσίαση συμπεριφοράς μεταβλητών εκ χειρισμού

Όπως προαναφέρθηκε, για κάποιες περιπτώσεις ο RL εμφανίζει κατά βάση καλύτερη συμπεριφορά από τον PI. Αυτό απεικονίζεται και στην συμπεριφορά των μεταβλητών

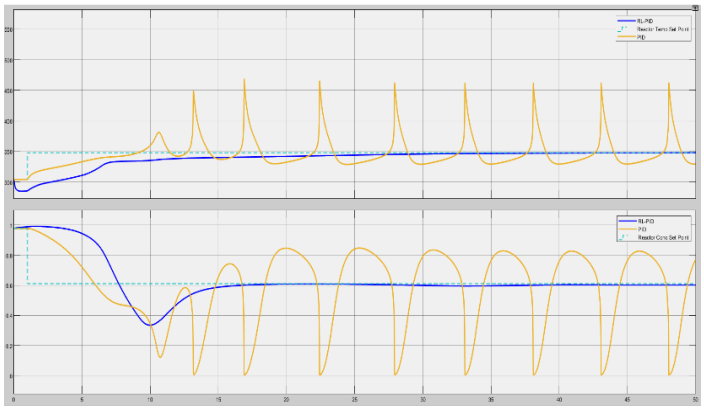
εκ χειρισμού οι οποίες πλέον δεν χρειάζεται να παγιδεύονται σε ταλαντωτικές συμπεριφορές και αποφεύγουν ακραίες ή αρνητικές τιμές υπό περιπτώσεις.

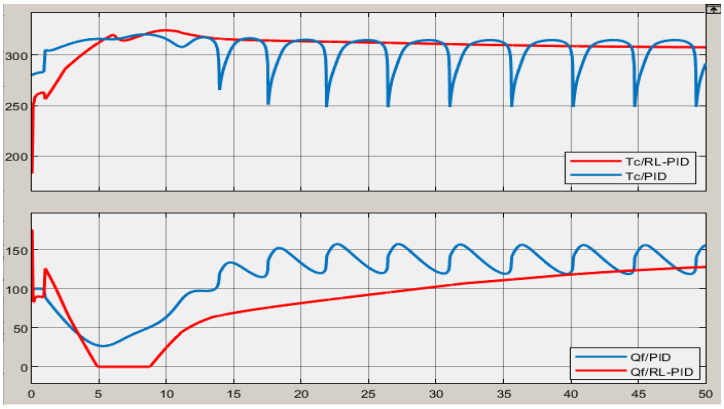
Μάλιστα, στα περισσότερα setpoint που αφορούν ασταθείς μόνιμες καταστάσεις, οι PI αποτυγχάνουν πλήρως και η συμπεριφορά των μεταβλητών εκ χειρισμού με την προσθήκη του πράκτορα παρουσιάζεται δραστικά βελτιωμένη. Τα αποτελέσματα συνοψίζονται στα διαγράμματα 30 (Πίνακας 7 – επεισόδιο 1500) και 31 (Πίνακας 9 – επεισόδιο 1705)[5] όπου στο πρώτο γίνεται αναφορά σε ευσταθή μόνιμη κατάσταση και στο δεύτερο σε ασταθή (αριστερά φαίνεται το επεισόδιο για το οποίο πραγματοποιείται σχολιασμός και δεξιά οι αντίστοιχες μεταβλητές εκ χειρισμού).



Διάγραμμα 30. Μεταβλητές εκ χειρισμού για ευσταθές σημείο ισορροπίας.

[5] Στο πάνω διάγραμμα παρουσιάζεται η παροχή τροφοδοσίας και στο κάτω παρουσιάζεται η παροχή του ψυκτικού.





Διάγραμμα 31. Μεταβλητές εκ χειρισμού για ασταθές σημείο ισορροπίας

Συμπεράσματα

Σύμφωνα με τα ευρήματα της παρούσας διπλωματικής εργασίας, η προσθήκη του RL agent στο σύστημα του CSTR για τη βελτίωση της λειτουργίας των ρυθμιστών PI κρίνεται επιτυχής. Αποδεικνύεται ότι ο ρυθμιστής RL μπορεί όχι μόνο να οδηγήσει σύστημα ελέγχου σε μία καλύτερη συμπεριφορά αλλά ακόμα και να διορθώσει τα λάθη που προκαλούν οι PI δρώντας ανεξάρτητα και αντισταθμίζοντας όλες τις λανθασμένες ενέργειές τους. Έτσι, η επιτυχής εφαρμογή του RL για τον έλεγχο του συστήματος CSTR με δύο PI ελεγκτές αποδεικνύει τη δυνατότητα αξιοποίησης των μηχανισμών ενισχυτικής μάθησης σε περιπτώσεις όπου οι παραδοσιακές μέθοδοι ελέγχου φαίνονται ανεπαρκείς ή και δεν μπορούν να ανταπεξέλθουν στις απαιτήσεις ρύθμισης πολύπλοκων συστημάτων με ισχυρές αλληλεπιδράσεις.

Κατά την εκπαίδευση του πράκτορα καίριο ρόλο διαδραματίζει ο σχεδιασμός της συνάρτησης ανταμοιβής μέσω της οποίας ο αλγόριθμος επιχειρεί να βελτιστοποιήσει την πολιτική του. Με επιλογή κατάλληλων όρων και των αντίστοιχων βαρών τους, δίνεται προσοχή στα κρίσιμα σημεία και τις αστοχίες των παραδοσιακών ρυθμιστών PI με αποτέλεσμα την αύξηση της απόδοσης του συστήματος ελέγχου. Μέσω προσεκτικής μελέτης και αναπροσαρμογής, η συνάρτηση ανταμοιβής μπορεί να τροποποιηθεί κατάλληλα ώστε να δίνει προτεραιότητα σε σημαντικούς στόχους, όπως η διατήρηση των επιθυμητών επιπέδων θερμοκρασίας και συγκέντρωσης, με ελαχιστοποίηση της κατανάλωσης ενέργειας ή μεγιστοποίηση της απόδοσης παραγωγής. Παράλληλα, γίνεται κατανοητό πως μία ορθά δομημένη συνάρτηση ανταμοιβής μπορεί να αποβεί πολύ χρήσιμη και για συστήματα υπό την επίδραση διαταραχών που δεν αποτέλεσαν παράγοντες της εκπαίδευσης του αλγορίθμου. Μάλιστα, η παρουσία των διαταραχών είναι ένα αντικείμενο που μπορεί να ενσωματωθεί στην εκπαίδευση για καλύτερα αποτελέσματα μέσα από την ανανέωση του περιβάλλοντος.

Συνοψίζοντας, η ενσωμάτωση του αλγορίθμου RL στο σύστημα CSTR με χρήση Matlab και περιβάλλον προσομοίωσης Simulink οδηγεί σε ένα αποτελεσματικό συνολικό σύστημα ελέγχου της διεργασίας και αποτελεί μία πρακτική και άμεσα εφαρμόσιμη λύση για συστήματα με αντίστοιχη δυναμική, σε βιομηχανική κλίμακα.

Τα αποτελέσματα που προέκυψαν ενθαρρύνουν περαιτέρω έρευνα και πειραματισμό πάνω στην εξερεύνηση πιο απλών ή σύνθετων αλγορίθμων RL, εναλλακτική διαμόρφωση του περιβάλλοντος προσομοίωσης, επανασχεδιασμό της συνάρτησης ανταμοιβής και προσθήκη μεταβλητών εκ χειρισμού για την ενίσχυση του ελέγχου των συστημάτων CSTR. Αυτό θα συμβάλει στην επέκταση της γνώσης και της κατανόησης των προσεγγίσεων ελέγχου που πραγματοποιούνται από την μηχανική μάθηση σε εφαρμογές χημικής μηχανικής.

Παράρτημα

Reactor set-up

```
function [sys,x0,str,ts,simStateCompliance] = reactor(t,x,u,flag)

switch flag,

    %%%%%%%%%%%%%%%%%%
    % Initialization %
    %%%%%%%%%%%%%%%%%%
    case 0,
        [sys,x0,str,ts,simStateCompliance]=mdlInitializeSizes();

        %%%%%%%%%%%%%%%
        % Derivatives %
        %%%%%%%%%%%%%%%
    case 1,
        sys=mdlDerivatives(t,x,u);

        %%%%%%%%%%
        % Update %
        %%%%%%%%%%
    case 2,
        sys=mdlUpdate(t,x,u);

        %%%%%%%%%%
        % Outputs %
        %%%%%%%%%%
    case 3,
        sys=mdlOutputs(t,x,u);

        %%%%%%%%%%%%%%%%%%%%%%
        % GetTimeOfNextVarHit %
        %%%%%%%%%%%%%%%%%%%%%%
    case 4,
        sys=mdlGetTimeOfNextVarHit(t,x,u);

        %%%%%%%%%%%%
        % Terminate %
        %%%%%%%%%%%%
    case 9,
        sys=mdlTerminate(t,x,u);

        %%%%%%%%%%%%%%%%%%%%
        % Unexpected flags %
        %%%%%%%%%%%%%%%%%%%%
    otherwise
        DAStudio.error('Simulink:blocks:unhandledFlag', num2str(flag));

end

% end sfuntmpl

%
%================================================================================
==
% mdlInitializeSizes
% Return the sizes, initial conditions, and sample times for the S-function.
%================================================================================
==
%
```

```
function [sys,x0,str,ts,simStateCompliance]=mdlInitializeSizes()

%
% call simsizes for a sizes structure, fill it in and convert it to a
% sizes array.
%
% Note that in this example, the values are hard coded.  This is not a
% recommended practice as the characteristics of the block are typically
% defined by the S-function parameters.
%
sizes = simsizes;

sizes.NumContStates  = 2;
sizes.NumDiscStates  = 0;
sizes.NumOutputs     = 2;
sizes.NumInputs      = 4;
sizes.DirFeedthrough = 0;
sizes.NumSampleTimes = 1;   % at least one sample time is needed

sys = simsizes(sizes);
%
% initialize the initial conditions
x0 = [304.2; 0.9774]; % for Tc = 280
%x0  = [324.475443431599; 0.87725294608097]; % for Tc = 300

%
% str is always an empty matrix
%
str = [];

%
% initialize the array of sample times
%
ts  = [0 0];

% Specify the block simStateCompliance. The allowed values are:
%    'UnknownSimState', < The default setting; warn and assume
DefaultSimState
%    'DefaultSimState', < Same sim state as a built-in block
%    'HasNoSimState',   < No sim state
%    'DisallowSimState' < Error out when saving or restoring the model sim
state
simStateCompliance = 'UnknownSimState';

% end mdlInitializeSizes

%=============================================================
% mdlDerivatives
% Return the derivatives for the continuous states.
%=============================================================================
==
%
function sys=mdlDerivatives(t,x,u)
%
% CSTR model from
%
% Michael A. Henson and Dale E. Seborg.  Nonlinear Process Control.
%      Prentice Hall PTR, Upper Saddle River, New Jersey, 1997.

% Description:
% Continuously Stirred Tank Reactor with energy balance and reaction A->B.
%   The temperature of the cooling jacket is the manipulated variable.

% Inputs (4):
% Temperature of cooling jacket (K)
Tc = u(1);  % nominal = 300
% Volumetric Flowrate (m^3/sec)
q = u(2);   % nominal = 100
```

```
% Feed Concentration (mol/m^3)
Caf = u(3); % nominal = 1
% Feed Temperature (K)
Tf = u(4);  % nominal = 350

% States (2):
% Temperature in CSTR (K)
T = x(1);
% Concentration of A in CSTR (mol/m^3)
Ca = x(2);

%Tc = 300;
%Ca_ss = 0.87725294608097;
%T_ss = 324.475443431599;

% Parameters:
% Volume of CSTR (m^3)
V = 100;
% Density of A-B Mixture (kg/m^3)
rho = 1000;
% Heat capacity of A-B Mixture (J/kg-K)
Cp = .239;
% Heat of reaction for A->B (J/mol)
mdelH = 5e4;
% E - Activation energy in the Arrhenius Equation (J/mol)
% R - Universal Gas Constant = 8.31451 J/mol-K
EoverR = 8750;
% Pre-exponential factor (1/sec)
k0 = 7.2e10;
% U - Overall Heat Transfer Coefficient (W/m^2-K)
% A - Area - this value is specific for the U calculation (m^2)
UA = 5e4;

% Compute xdot:
sys(1,1) = q/V*(Tf - T) ...
    + mdelH/(rho*Cp)*k0*exp(-EoverR/T)*Ca ...
    + UA/V/rho/Cp*(Tc-T);
sys(2,1) = q/V*(Caf - Ca) ...
    - k0*exp(-EoverR/T)*Ca;

% end mdlDerivatives

%
%=============================================================================
==
% mdlUpdate
% Handle discrete state updates, sample time hits, and major time step
% requirements.
%=============================================================================
==
%
function sys=mdlUpdate(t,x,u)

sys = [];

% end mdlUpdate

%
%=============================================================================
==
% mdlOutputs
% Return the block outputs.
%=============================================================================
==
%
function sys=mdlOutputs(t,x,u)
x1        = x(1);
x2        = x(2);
```

```
sys = [x1;x2];

% end mdlOutputs

%
%=============================================================================
==
% mdlGetTimeOfNextVarHit
% Return the time of the next hit for this block.  Note that the result is
% absolute time.  Note that this function is only used when you specify a
% variable discrete-time sample time [-2 0] in the sample time array in
% mdlInitializeSizes.
%=============================================================================
==
%
function sys=mdlGetTimeOfNextVarHit(t,x,u)

sampleTime = 1;    %  Example, set the next hit to be one second later.
sys = t + sampleTime;

% end mdlGetTimeOfNextVarHit

%
%=============================================================================
==
% mdlTerminate
% Perform any end of simulation tasks.
%=============================================================================
==
%
function sys=mdlTerminate(t,x,u)

sys = [];

% end mdlTerminate
```

Create Environment

```
% Open the model
load('Setpoints_cut_v2.mat','setpoints_mss_cut')
setpoints=setpoints_mss_cut;
ep=(the desired one to see the simulation);
Tsp = setpoints(ep,1)
Csp = setpoints(ep,2)
mdl = 'Cstr_pid_';
open_system(mdl);
agentblk = [mdl '/RL Agent'];

% Create Environment Interface

% ObservtionInfo
observationInfo = rlNumericSpec([2 1],'LowerLimit',-
inf*ones(2,1),'UpperLimit',inf*ones(2,1));
observationInfo.Name = 'observations';
observationInfo.Description = 'information deviations from steady state';
%ActionInfo

actionInfo=rlNumericSpec([2 1],'LowerLimit',[-500; -500],'UpperLimit',[100;
100]);
actionInfo.Name = 'action';
actionInfo.Description = 'feed Flow Rate Correction Cold Flow Temperature';
```

```
% Create the environment interface
env = rlSimulinkEnv(mdl,agentblk,observationInfo,actionInfo);

env.ResetFcn = @(in)localResetFcn(in,setpoints_mss_cut);


rng(0)
```

Reset Function

```
function in = localResetFcn(in,setpoints_mss_cut)
% Select random episode
%ep = randi([1,size(setpoints_cut,1)]);
%Run through all 5000 episodes
persistent episode
if isempty(episode)
episode = 0;
end
episode = episode + 1;

T = setpoints_mss_cut(episode,1);
C = setpoints_mss_cut(episode,2);


blk1= 'Cstr_pid_/Reactor Temp Set Point';
in=setBlockParameter(in,blk1,'After',num2str(T));

blk2= 'Cstr_pid_/Reactor Conc Set Point';
in=setBlockParameter(in,blk2,'After', num2str(C));

end
```

Create DDPG agent

```
numberofInputs = observationInfo.Dimension(1); % number of inputs (2)
neuronsNumber = 20; % number of neurons
numberofOutputs = actionInfo.Dimension(1); % number of outputs(2)

% Critic


statePath = [
featureInputLayer(numberofInputs,'Normalization','none','Name','observation'
)
fullyConnectedLayer(neuronsNumber,'Name','fc1')
reluLayer('Name','relu1')
fullyConnectedLayer(neuronsNumber,'Name','fc2')
additionLayer(2,'Name','add')
reluLayer('Name','relu2')
fullyConnectedLayer(numberofOutputs,'Name','fc3');
reluLayer('Name','relu3')
fullyConnectedLayer(1,'Name','fc4')];

actionPath= [
```

```
featureInputLayer(numberofOutputs,'Normalization','none','Name','action')
fullyConnectedLayer(neuronsNumber,'Name','fc5')];

criticNetwork = layerGraph(statePath);
criticNetwork = addLayers(criticNetwork,actionPath);

criticNetwork = connectLayers(criticNetwork,'fc5','add/in2');

figure
plot(criticNetwork)

criticOptions = rlRepresentationOptions('LearnRate',1e-
3,'GradientThreshold',1,'L2RegularizationFactor',1e-4);

critic =
rlQValueRepresentation(criticNetwork,observationInfo,actionInfo,'Observation
',{'observation'},'Action',{'action'},criticOptions);

% Actor


actorNetwork = [

featureInputLayer(numberofInputs,'Normalization','none','Name','observation'
)
    fullyConnectedLayer(neuronsNumber,'Name','fc1')
    reluLayer('Name','relu1')
    fullyConnectedLayer(neuronsNumber,'Name','fc2')
    reluLayer('Name','relu2')
    fullyConnectedLayer(neuronsNumber,'Name','fc3')
    reluLayer('Name','relu3')
    fullyConnectedLayer(numberofOutputs,'Name','fc4')
    tanhLayer('Name','tanh1')

scalingLayer('Name','ActorScaling1','Scale',[actionInfo.UpperLimit(1,1)])];

actorOptions = rlRepresentationOptions('LearnRate',1e-
4,'GradientThreshold',1,'L2RegularizationFactor',1e-2);

actor =
rlDeterministicActorRepresentation(actorNetwork,observationInfo,actionInfo,'
Observation',{'observation'},'Action',{'ActorScaling1'},actorOptions);

Ts=0.1;

% Agent Options
agentOptions = rlDDPGAgentOptions(...
    'SampleTime',Ts,...
    'TargetSmoothFactor',1e-3,...
    'ExperienceBufferLength',1e6,...
    'DiscountFactor',0.99,...
    'MiniBatchSize',128);
agentOptions.NoiseOptions.Variance =0.1;
% [0.1;0.6;0.6];
agentOptions.NoiseOptions.VarianceDecayRate = 1e-5;
unstable3= rlDDPGAgent(actor,critic,agentOptions);
```

Train Agent

```
Ts=.1;


Tf=100;
```

```
maxepisodes = (by user);

maxsteps = ceil(Tf/Ts);

trainingOpts = rlTrainingOptions(...

'MaxEpisodes',maxepisodes, ...
'MaxStepsPerEpisode',maxsteps, ...
'ScoreAveragingWindowLength',(by user), ...
'Verbose',false, ...
'Plots','training-progress',...
'StopTrainingCriteria','AverageReward',...
'StopTrainingValue',(by user) ...
);

%trainingOpts = rlTrainingOptions(...
%    'MaxEpisodes',maxepisodes, ...
%    'MaxStepsPerEpisode',maxsteps, ...
%    'Verbose',false, ...
%    'StopOnError', ...
%    'Plots','training-progress', ...
%    'StopTrainingCriteria','AverageReward', ...
%    'StopTrainingValue',-5)
%    'SaveAgentDirectory',pwd + "\MyAgents");

% doTraining = false;
doTraining = true;

if doTraining
% Train the agent.
trainingStats = train(unstable3,env,trainingOpts);
% save(opt.SaveAgentDirectory + "/TotalDDPGAgent.mat",'agent')
%save TotalOuterDDPGAgent1.mat TotalOuterAgent
else
% Load the pretrained agent for the example.
% load('Agent_Definition.mat','agent')
end
% save TotalDDPGAgent.mat agent
%save(trainingOpts.SaveAgentDirectory +
"/TotalOuterDDPGAgent1.mat",'TotalOuterAgent')
```

Βιβλιογραφία

[1] R.S. Sutton, A.G. Barto, Reinforcement learning: An introduction (Second edition), Cambridge, Massachusetts: The MIT Press, 2018.

[2] C. Gary, «Medium,» 05 08 2018. [Ηλεκτρονικό]. Available: https://towardsdatascience.com/applications-of-reinforcement-learning-in-real-world-1a94955bcd12. [Πρόσβαση 14 06 2023].

[3] F. Ruan De Rezende, C. Bruno Didier Olivier, S. Argimiro Resende και D. S. Maurício B., «Where Reinforcement Learning Meets Process Control,» *Processes,* τόμ. 10, αρ. 11, p. 2311, 2022.

[4] «Reinforcement Learning Toolbox Documentation,» [Ηλεκτρονικό]. Available: https://www.mathworks.com/help/reinforcement-learning/index.html. [Πρόσβαση 14 06 2023].

[5] «What is Reinforcement Learning? – Overview of How it Works | Synopsys,» 14 6 2023. [Ηλεκτρονικό]. Available: https://www.synopsys.com/ai/what-is-reinforcement-learning.html.

[6] V. R. Konda και J. N. Tsitsiklis, «On actor-critic algorithms,» *SIAM Journal on Control and Optimization,* τόμ. 42, αρ. 4, pp. 1143-1166, 2003.

[7] P. Abbeel και A. Y. Ng, «Apprenticeship Learning via Inverse Reinforcement Learning,» *Proceedings of the Machine Learning (ICML),* 2004.

[8] N. Krakauer και E. D. Sontag, «Reinforcement Learning for Optimal Control of Chemical Processes,» *Journal ofm Process Control,* τόμ. 11, αρ. 5, 2001.

[9] M. Thomas E., Process control: designing processes and control systems for dynamic perfomance, Boston: McGraw-Hill, 2000.

[10] A. Karl J. και H. Tore, PID controllers, Research Triangle Park, N.C: International Society for Measurement and Control, 1995.

[11] F. Gene F., P. J. David και E.-N. Abbas, Feedback control of dynamic systems, Boston: Pearson, 2015.

[12] D. Richard C. και B. Robert H., Modern control systems, Boston: Pearson, 2011.

[13] O. Katsuhiko, Modern control engineering, Boston: Prentice-Hall, 2010.

[14] Z. Guan και T. Yamamoto, «Design of a Reinforcement Learning PID Controller,» σε *International Joint Conference on Neural Networks (IJCNN)*, Glasgow, 2020.

[15] H. S. Fogler, Μηχανική Χημικών Αντιδράσεων και Σχεδιασμός Αντιδραστήρων, Αθήνα: Τζιόλα, 2018.

[16] J. J. A. A. B. Smith, «Kinetic Study of the Decomposition of Hydrogen Peroxide Using Iron(III) Nitrate as Catalyst,» *Journal of Chemical Kinetics,* τόμ. 42, αρ. 3, pp. 235-242, 2015.

[17] G. Miroslav, «Thermodynamics and its applications,» *The Canadian Journal of Chemical Engineering,* τόμ. 75, αρ. 6, pp. 1165-1166, 1997.

[18] T. Hoben, «A binomial mixture model for classification performance,» *Journal of Experimental Child Psychology,* τόμ. 48, αρ. 3, pp. 423-430, 1989.

[19] E. Patrick M. και M. Bruce, «LEARNING ABOUT EDUCATION,» *Economic Inquiry,* τόμ. 56, αρ. 1, pp. 263-277, 2018.

[20] Π. Παπαδημητρίου και Α. Παπαδημητρίου, Μοντελοποίηση και Ελέγχος Βιομηχανικών Συστημάτων με Εφαρμογές στη Μηχανική Τροφίμων, Αθήνα, 2018.

[21] G. F. Franklin, J. D. Powell και A. Emami-Naeini, Feedback control of dynamic systems, Boston: Pearson, 2015.