Πολυτεχνείο Κρήτης
Σχολή Ηλεκτρονικών Μηχανικών και Μηχανικών Υπολογιστών

JazzMate

Ανάπτυξη Πράκτορα Ενισχυτικής Μάθησης για Διαδραστικό Μουσικό Αυτοσχεδιασμό

Αυτόνομοι Πράκτορες

Ιωάννης Μπουρίτης
2021030173

Χειμερινό Εξάμηνο
February 6, 2026

Contents

1 Εισαγωγή και Στόχος

Η παρούσα εργασία πραγματεύεται την ανάπτυξη του **JazzMate**, ενός αυτόνομου πράκτορα (autonomous agent) ικανού να παράγει Jazz αυτοσχεδιασμούς (solo) σε πραγματικό χρόνο. Στόχος ήταν η δημιουργία ενός συστήματος που δεν βασίζεται σε άκαμπτους κανόνες (hard-coded rules), αλλά μαθαίνει να "αισθάνεται" τη μουσική μέσω **Ενισχυτικής Μάθησης (Reinforcement Learning - RL)**.

Ο πράκτορας καλείται να αντιμετωπίσει ένα δυναμικό περιβάλλον όπου η αρμονία (συγχορδίες) αλλάζει συνεχώς και πρέπει να επιλέξει τις κατάλληλες νότες, παύσεις και διάρκειες ώστε το αποτέλεσμα να είναι μουσικά ευχάριστο και δημιουργικό.

2 Αρχιτεκτονική Συστήματος

2.1 Αλγόριθμος και Εργαλεία

Για την εκπαίδευση του πράκτορα επιλέχθηκε ο αλγόριθμος **Deep Q-Network (DQN)**, με χρήση της πολιτικής `MultiInputPolicy` από τη βιβλιοθήκη *Stable-Baselines3*. Ο αλγόριθμος εκπαιδεύτηκε για 200,000 βήματα (timesteps) ώστε να επιτύχει σταθερή σύγκλιση.

- **Γλώσσα:** Python
- **Βιβλιοθήκες:** Gymnasium (Custom Environment), Stable-Baselines3 (DQN), Mido (MIDI I/O), FluidSynth (Synthesis).

2.2 Μοντελοποίηση Προβλήματος (MDP)

Το πρόβλημα ορίστηκε ως μια Διαδικασία Λήψης Αποφάσεων Markov:

- **State (Observation Space):** Ένας συνδυασμένος χώρος (Dict) που περιέχει:
  - **Chord Tones:** Διάνυσμα 12 θέσεων (Multi-hot encoding) που υποδεικνύει ποιες νότες ανήκουν στην τρέχουσα συγχορδία.
  - **Step Progress:** Η σχετική θέση μέσα στο κομμάτι ($0 \to 1$).
  - **Last Action & Held Duration:** Η προηγούμενη επιλογή και η διάρκεια διατήρησής της.
  - **Style Seed:** Έναν τυχαίο παράγοντας για την ενθάρρυνση διαφορετικών προσεγγίσεων.
- **Action Space:** 38 διακριτές επιλογές:
  - 0-35: Νότες σε εύρος 3 οκτάβων (C3 έως B5).
  - 36: Παύση (Rest).
  - 37: Κράτημα Νότας (Hold).

3 Τεχνικές Λεπτομέρειες Εκπαίδευσης

3.1 Υπερπαράμετροι Αλγορίθμου

Η εκπαίδευση του DQN πράκτορα πραγματοποιήθηκε με τις ακόλουθες υπερπαραμέτρους:

Παράμετρος	**Τιμή**
Συνολικά Timesteps	200,000
Learning Rate (α)	1×10^{-4}
Replay Buffer Size	50,000 transitions
Batch Size	32 (default)
Target Network Update (τ)	1.0 (hard update)
Target Network Update Frequency	10,000 steps
Gamma (γ) - Discount Factor	0.99 (default)
Exploration Fraction	0.4 (40% των timesteps)
Initial Exploration (ϵ_{start})	1.0
Final Exploration (ϵ_{end})	0.05

Table 1: Υπερπαράμετροι του DQN αλγορίθμου

3.2 Στρατηγική Εξερεύνησης

Χρησιμοποιήθηκε η κλασική ϵ**-greedy** στρατηγική με γραμμική φθίνουσα εξερεύνηση:

$$\epsilon(t) = \max\left(\epsilon_{end}, \epsilon_{start} - \frac{t}{T_{explore}} \cdot (\epsilon_{start} - \epsilon_{end})\right) \tag{1}$$

όπου $T_{explore} = 0.4 \times 200000 = 80000$ timesteps.

Αυτό σημαίνει ότι:

- **Timesteps 0–80,000:** Το ϵ μειώνεται γραμμικά από 1.0 σε 0.05, επιτρέποντας στον πράκτορα να εξερευνά ενεργά νέες στρατηγικές.
- **Timesteps 80,000–200,000:** Το ϵ παραμένει σταθερό στο 0.05, δίνοντας έμφαση στην εκμετάλλευση (exploitation) της μαθημένης πολιτικής.

3.3 Δομή Επεισοδίων

Κάθε επεισόδιο (episode) αποτελείται από μία ολοκληρωμένη μουσική εκτέλεση:

- **Μήκος Progression:** Κάθε φορά που καλείται το `reset()`, δημιουργείται μια τυχαία ακολουθία 8 συγχορδιών, με κάθε συγχορδία να διαρκεί 16 steps.
- **Συνολικά Steps ανά Episode:** $8 \times 16 = 128$ steps/episode.
- **Αριθμός Επεισοδίων:** Για 200,000 συνολικά timesteps:

$$N_{episodes} = \frac{200000}{128} \approx 1562 \text{ επεισόδια} \tag{2}$$

Κατά την εκπαίδευση, ο πράκτορας έπαιξε συνολικά **περίπου 1562 διαφορετικά μουσικά κομμάτια**, εξερευνώντας χιλιάδες διαφορετικούς συνδυασμούς συγχορδιών.

3.4 Experience Replay και Neural Network

Το DQN χρησιμοποιεί:

- **Replay Buffer:** Αποθηκεύονται οι τελευταίες 50,000 μεταβάσεις $(s, a, r, s', done)$.

- **Mini-batch Updates:** Σε κάθε βήμα εκπαίδευσης, δειγματοποιούνται τυχαία 32 transitions για gradient descent.
- **Target Network:** Ανανεώνεται πλήρως (hard update) κάθε 10,000 steps για σταθεροποίηση της εκπαίδευσης.

Η νευρωνική αρχιτεκτονική (`MultiInputPolicy`) χειρίζεται αυτόματα το Dict observation space, εξάγοντας χαρακτηριστικά από κάθε στοιχείο ξεχωριστά (chord tones, progress, action history) και στη συνέχεια τα συνδυάζει σε ένα κοινό latent representation.

3.5 Μετρικές Παρακολούθησης

Η εκπαίδευση παρακολουθήθηκε μέσω του **Monitor wrapper** του Gymnasium, το οποίο καταγράφει:

- **Episode Reward (R_{total}):** Το άθροισμα των rewards σε κάθε επεισόδιο.
- **Episode Length:** Πάντα 128 steps (σταθερό μήκος επεισοδίου).
- **Timestep:** Ο συνολικός αριθμός βημάτων μέχρι στιγμής.

Τα δεδομένα αποθηκεύονται στο αρχείο `training_logs/monitor.csv` και απεικονίζονται στη γραφική παράσταση της Ενότητας 4.4.

4 Η Διαδικασία Εκπαίδευσης: Reward Shaping

Η συνάρτηση ανταμοιβής (R) σχεδιάστηκε με στόχο την ισορροπία μεταξύ αρμονικής ορθότητας και μελωδικής ελευθερίας.

4.1 Αρμονία και Μελωδική Ροή

Ο πράκτορας λαμβάνει $+1.0$ για νότες εντός συγχορδίας και -0.6 για "λάθος" νότες. Ταυτόχρονα, πριμοδοτούνται τα μικρά μελωδικά διαστήματα (διαστήματα 1-2 ημιτονίων λαμβάνουν $+0.8$), ενώ τιμωρούνται αυστηρά τα μεγάλα πηδήματα (διάστημα > 9 λαμβάνει -1.5), προωθώντας τη δημιουργία ομαλών μελωδικών γραμμών.

4.2 Καταπολέμηση της Μονοτονίας (Anti-Pattern Detection)

Ένα από τα σημαντικότερα κομμάτια του κώδικα είναι η αποτροπή των "λούπων".

- **Exact Note Repeats:** Η επανάληψη της ίδιας νότας τιμωρείται άμεσα (-2.0 για την πρώτη επανάληψη, -10.0 για κάθε επόμενη).
- **Sequence Loops:** Εισήχθη μηχανισμός ελέγχου ιστορικού (History). Αν ο πράκτορας επαναλάβει μοτίβα 2, 3 ή 4 νοτών (π.χ. A-B-A-B), δέχεται κλιμακωτές ποινές από -5.0 έως -15.0.

4.3 Ενθάρρυνση "Riffs" και Φρασεολογίας

Για να μην γίνει ο πράκτορας υπερβολικά παθητικός λόγω των ποινών:

- **Riff Bonus:** Μια αλληλουχία 4-5 διαφορετικών νοτών ανταμείβεται με $+2.5$ έως $+4.0$.

- **Musical Phrasing:** Δίνεται η μέγιστη ανταμοιβή ($+3.0$) όταν ο πράκτορας επιλέξει παύση (**Rest**) αμέσως μετά από ένα πετυχημένο riff, διδάσκοντάς του να ”αναπνέει” μουσικά.

4.4 Ανάλυση Καμπύλης Εκμάθησης

Παρακάτω παρουσιάζεται η πορεία εκπαίδευσης του πράκτορα για 200.000 βήματα.

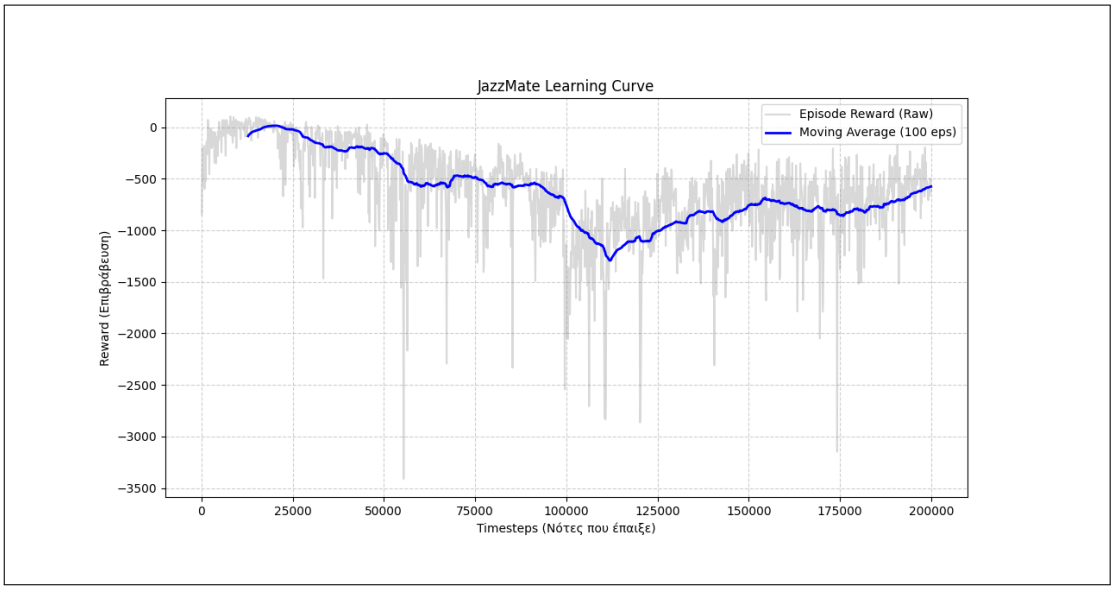

Figure 1: **Καμπύλη Εκμάθησης (Learning Curve):** Η μπλε γραμμή αναπαριστά τον κυλιόμενο μέσο όρο (moving average) της επιβράβευσης. Παρατηρούμε μια αρχική πτώση (dip) γύρω στα 100k βήματα, η οποία οφείλεται στην εισαγωγή αυστηρών ποινών και στην εξερεύνηση. Στη συνέχεια, ο πράκτορας προσαρμόζεται, μαθαίνει να αποφεύγει τις επαναλήψεις και η απόδοση ανακάμπτει σταθερά.

Ερμηνεία Φάσεων Εκπαίδευσης:

1. **Αρχική Εξερεύνηση (0–50k):** Υψηλή διακύμανση λόγω τυχαίας εξερεύνησης ($\epsilon \approx 1.0 \rightarrow 0.4$). Μέσος όρος reward περίπου -200 έως -400.
2. **Φάση Προσαρμογής (50k–100k):** Παρατηρείται πτώση στην απόδοση καθώς εφαρμόζονται οι ποινές anti-spam. Ο πράκτορας δοκιμάζει νέες στρατηγικές.
3. **Ανάκαμψη (100k–150k):** Εμφανής βελτίωση καθώς ο πράκτορας μαθαίνει να συνδυάζει αρμονία, ποικιλία και phrasing. Το ϵ έχει πέσει στο 0.05.
4. **Σταθεροποίηση (15a0k–200k):** Σύγκλιση σε σταθερή απόδοση με μέσο reward γύρω στο -400 έως -600, που αντιστοιχεί σε μουσικά αποδεκτά solos.

5 Διαδραστικότητα και Υλοποίηση Player

Το σύστημα αναπαραγωγής (`play_jazz.py`) μετατρέπει τις αποφάσεις του AI σε ζωντανή μουσική.

5.1 Humanization και Στυλ

Εφαρμόζεται αλγόριθμος **Swing** με αυξομείωση του χρόνου των steps (1.3x / 0.7x) και δυναμικές εντάσεις (velocity) ανάλογα με το τονικό ύψος. Ο χρήστης μπορεί να επιλέξει μεταξύ **Simple Block Chords** και **Arpeggio** για τη συνοδεία του αριστερού χεριού.

5.2 Jam Mode

Το σύστημα υποστηρίζει πλήρως εξωτερικούς MIDI Controllers. Μέσω ενός callback μηχανισμού, το πρόγραμμα ενημερώνεται ακαριαία για την αλλαγή συγχορδίας που προκαλεί ο χρήστης πατώντας νότες-βάσεις στο κλαβιέ, επιτρέποντας έναν διαδραστικό αυτοσχεδιασμό ”ανθρώπου-μηχανής”.

6 Συμπέρασμα και Μελλοντικές Επεκτάσεις

Το JazzMate αποδεικνύει ότι η Ενισχυτική Μάθηση μπορεί να κωδικοποιήσει σύνθετες καλλιτεχνικές έννοιες μέσω προσεκτικού Reward Shaping. Ο πράκτορας παρουσιάζει συνοχή, αποφεύγει τη στατικότητα και ανταποκρίνεται αρμονικά στις αλλαγές.

1. **Ολοκληρωμένη ”Ορχήστρα”:** Επέκταση του συστήματος για την ταυτόχρονη παραγωγή μουσικής από πολλαπλά όργανα (π.χ. πιάνο, μπάσο και τύμπανα), δημιουργώντας ένα πλήρες αυτοσχεδιαστικό σύνολο.
2. **Δυναμική Αναγνώριση Συγχορδιών:** Αντικατάσταση της στατικής αντιστοίχισης πλήκτρου-συγχορδίας με έναν αλγόριθμο αναγνώρισης σε πραγματικό χρόνο, επιτρέποντας στον χρήστη να παίζει ελεύθερα σύνθετες συγχορδίες τις οποίες το μοντέλο θα αναγνωρίζει αυτόματα.
3. **Πολυπρακτορική Υλοποίηση (Multi-Agent):** Εκπαίδευση διαφορετικών πρακτόρων με διαφορετικές ”μουσικές προσωπικότητες” και κανόνες. Ένα ανώτερο σύστημα αποφάσεων θα μπορεί να εναλλάσσει τους πράκτορες, αποφεύγοντας τη μονοτονία και εξασφαλίζοντας ότι το αποτέλεσμα θα ακούγεται πάντα φρέσκο.
4. **RL από Ανθρώπινη Ανατροφοδότηση (RLHF):** Ενσωμάτωση ενός μηχανισμού όπου ο μουσικός-συνεργάτης θα βαθμολογεί την απόδοση του πράκτορα σε πραγματικό χρόνο. Αυτό θα επιτρέψει στο μοντέλο να προσαρμόζεται δυναμικά στο υποκειμενικό γούστο και το στυλ του εκάστοτε συνεργάτη.