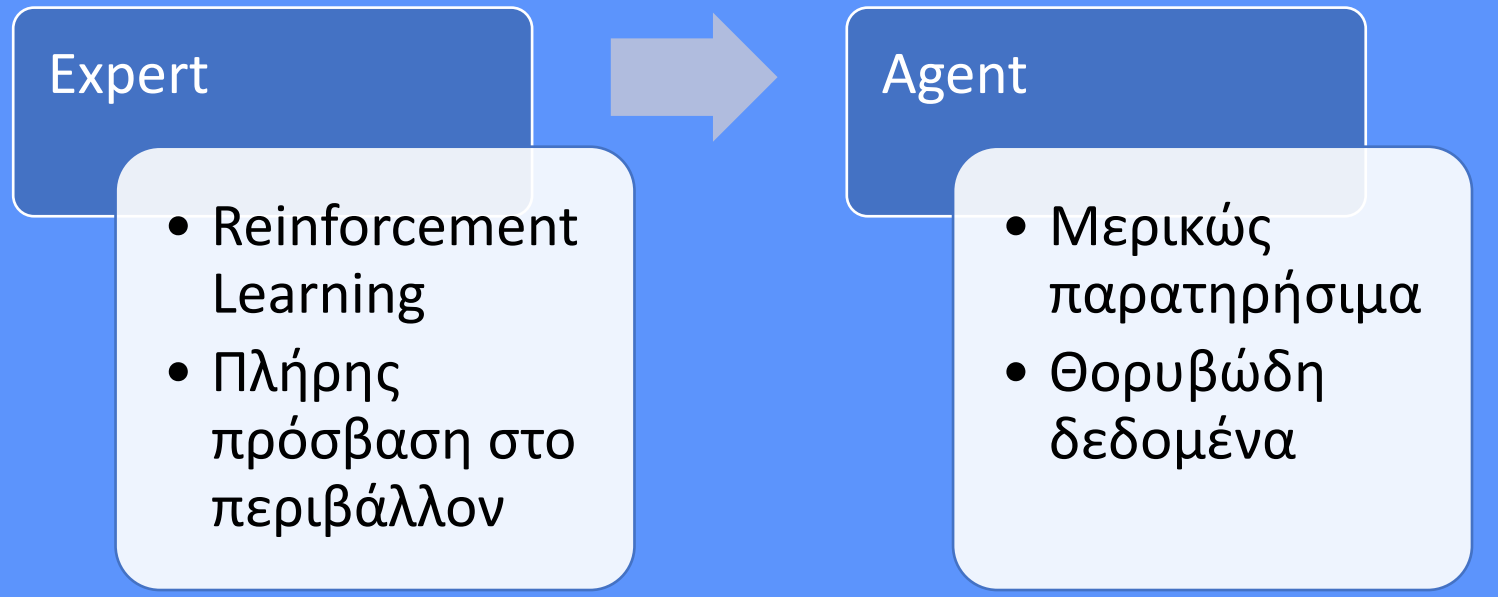
Imitation Learning in Super Mario Bros: Behavior Cloning and DAgger Using Privileged Information

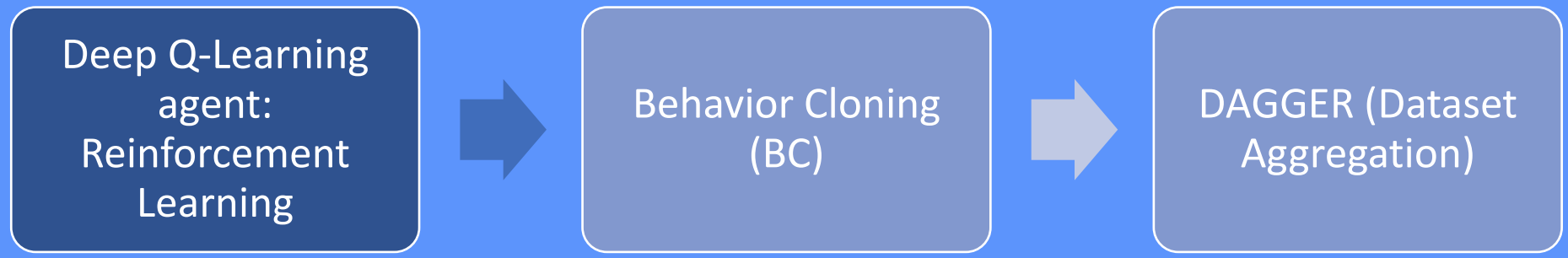
Μαρία - Νίκη Ζωγράφου up1096060@ac.upatras.gr

Νικόλαος Γέροντας up1092813@ac.upatras.gr

Στόχος Εργασίας



Εκπαίδευση των Agents

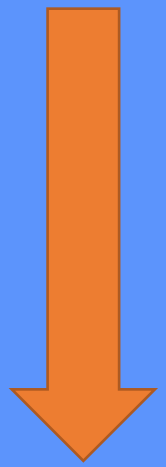


Expert: Deep Q-Learning

Q-learning:

$$Q(s,a) = Q(s,a) + a * (r + \gamma \, max_z Q(s',a') - Q(s,a))$$

Q: Q-table
s: state
a: action
s': next state
a': action with max future reward

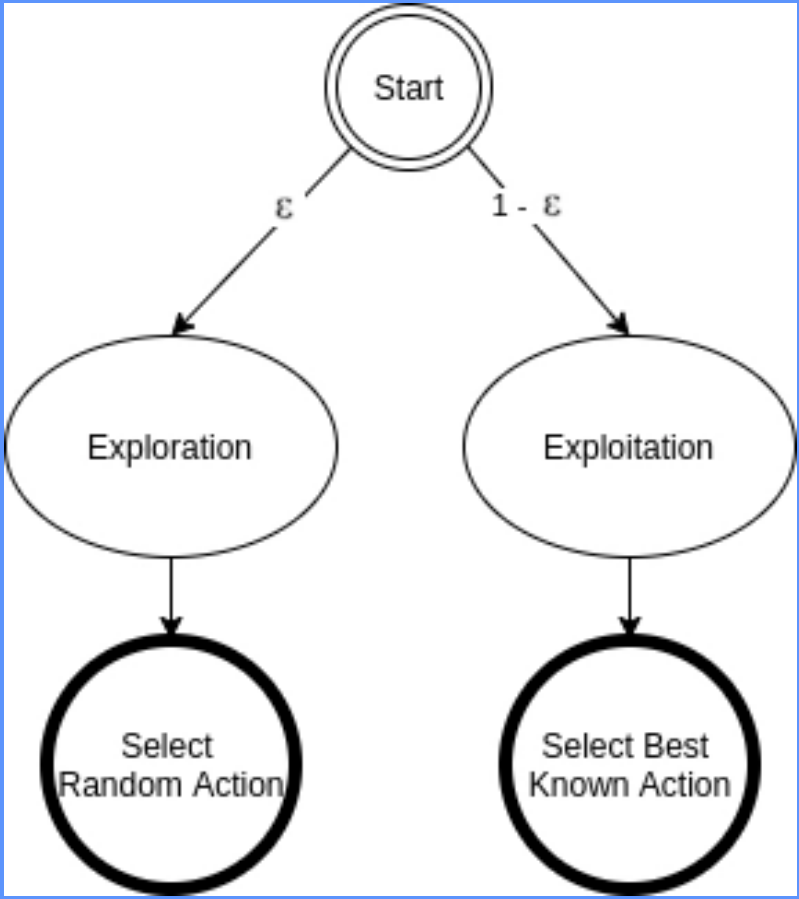


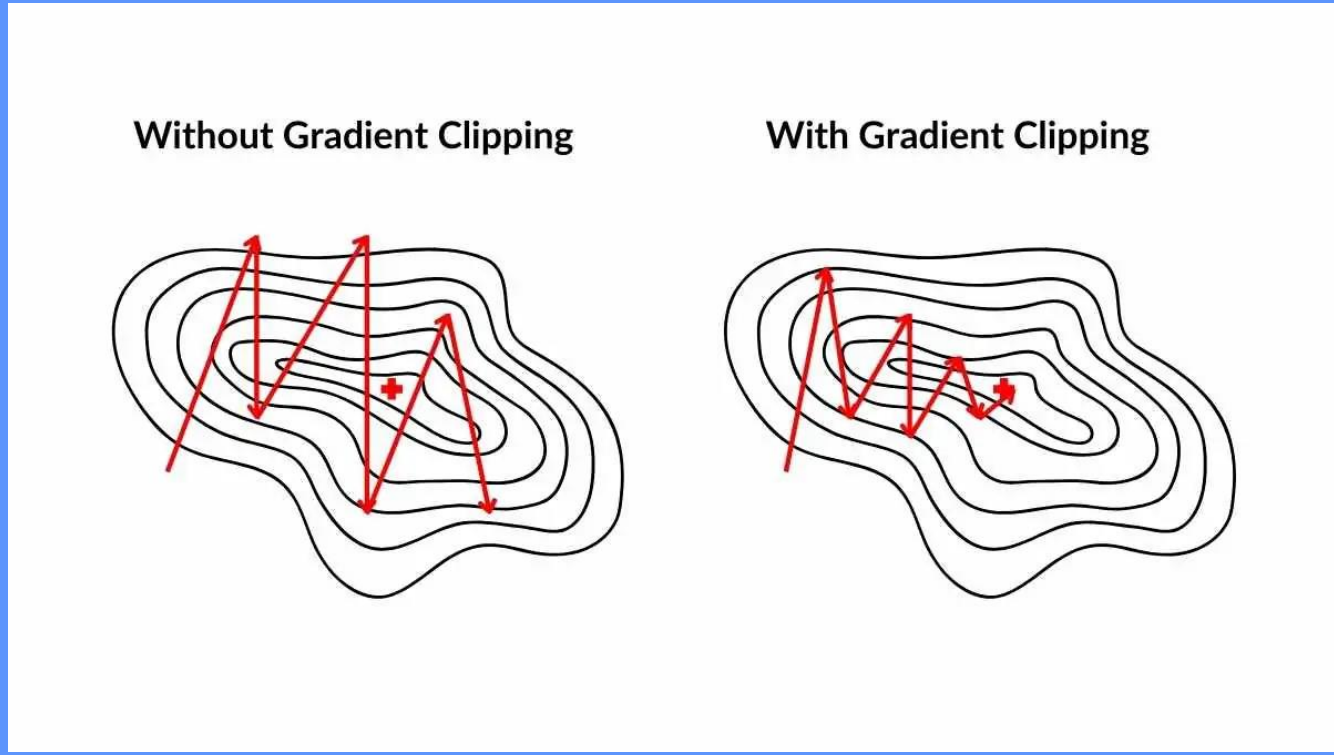
Αντικατάσταση Q-table
με νευρωνικό δίκτυο

Deep Q-learning:

Q – value function: $Q(s,a;\theta)$, όπου θ οι παράμετροι του νευρωνικού

ε - greedy policy / Gradient clipping



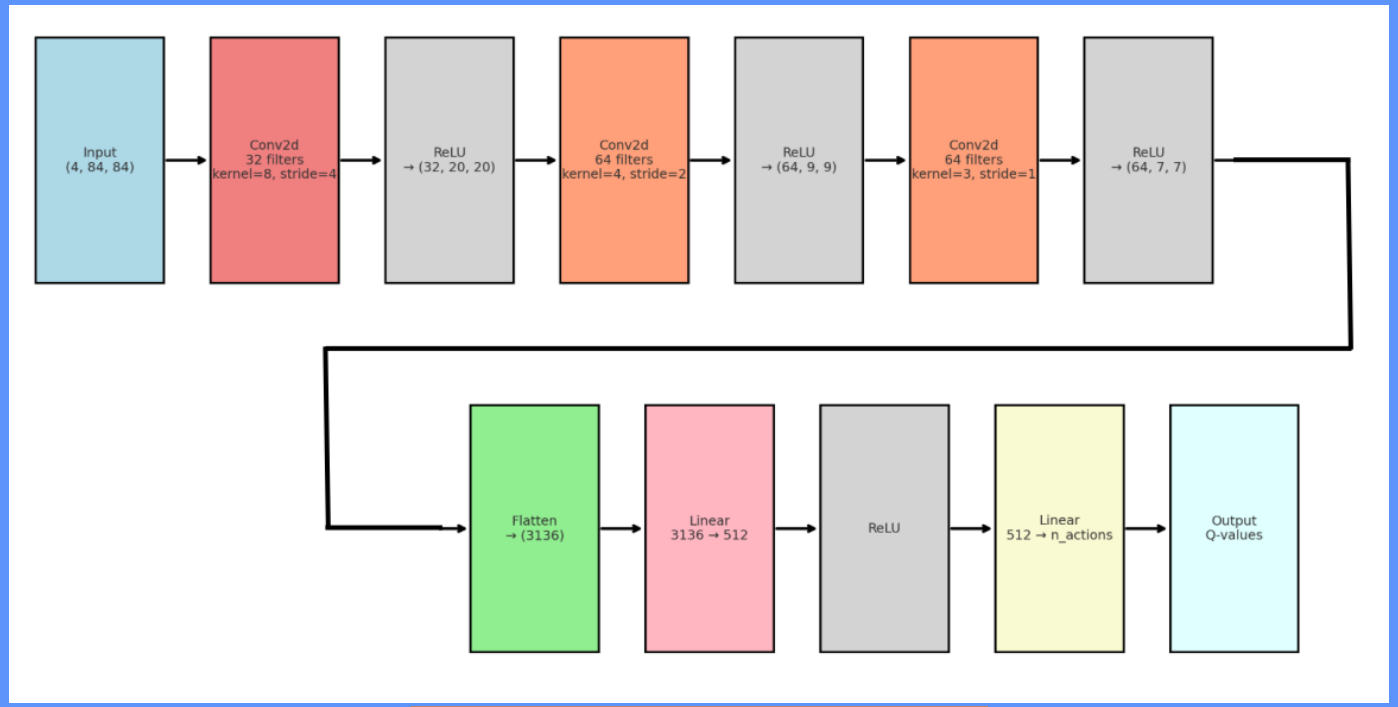


Expert: Deep Q-Learning

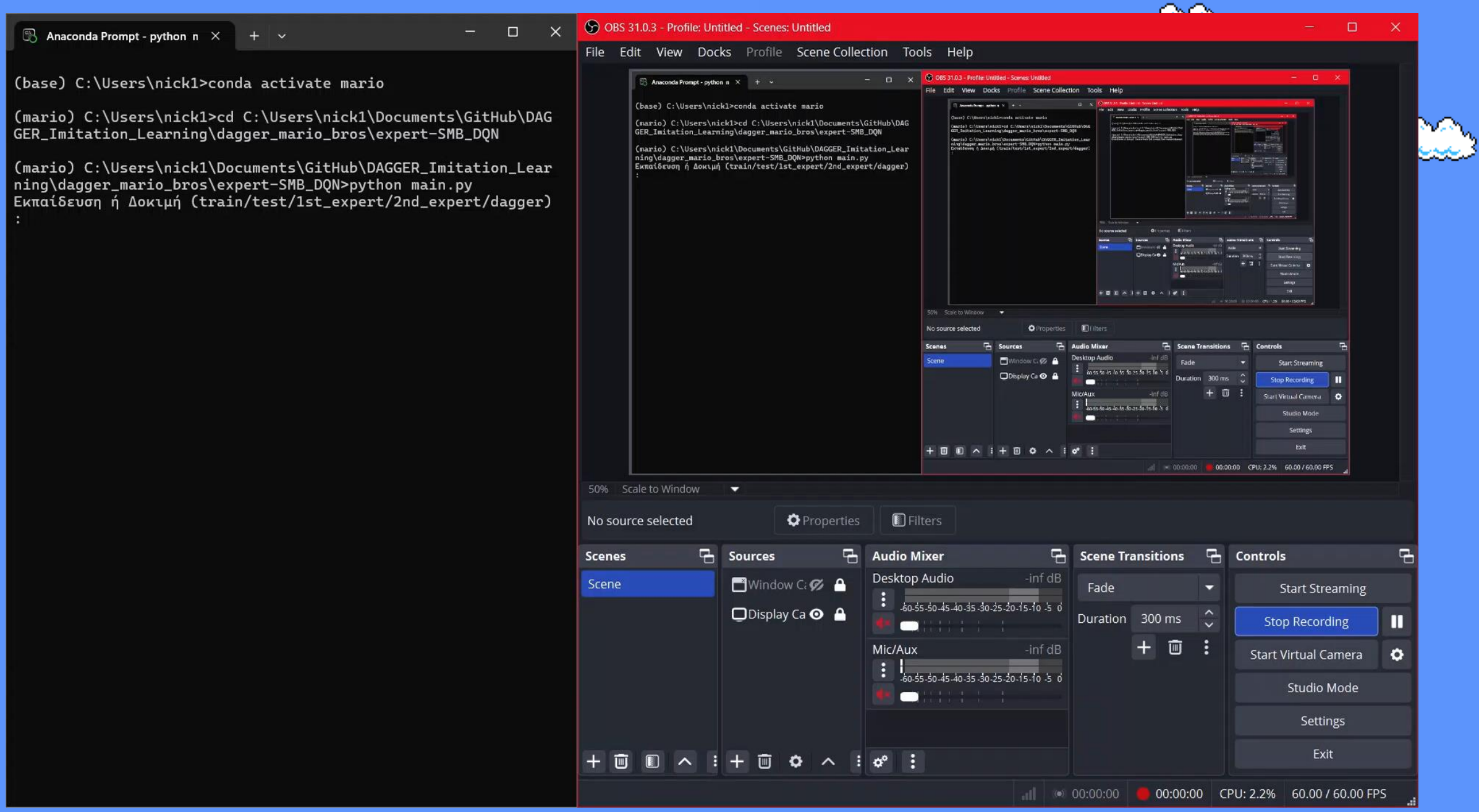
Input
- 4 τελευταία frames
- μετατροπή σε grayscale
- 84x84 pixels
- frame skipping

Reward Shaping
- Πρόοδος προς τα δεξιά +0.1
- Χρονικό penalty -0.1
- Ποινή για θάνατο -10
- Τερματισμός +100

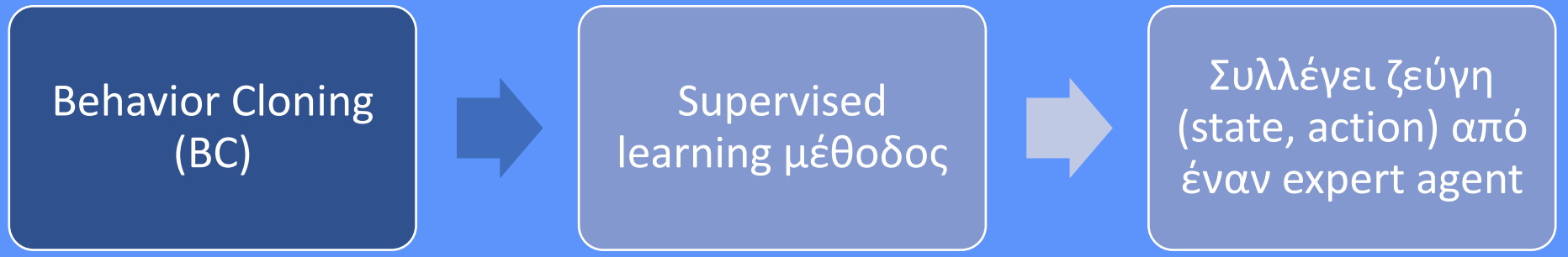
Loss function: **MSE** Loss



Αρχιτεκτονική νευρωνικού


Anaconda Prompt - python n
(base) C:\Users\nick1>conda activate mario
(mario) C:\Users\nick1>cd C:\Users\nick1\Documents\GitHub\DAGGER_Imitation_Learning\dagger_mario_bros\expert-SMB_DQN
(mario) C:\Users\nick1\Documents\GitHub\DAGGER_Imitation_Learning\dagger_mario_bros\expert-SMB_DQN>python main.py
Εκπαίδευση ή Δοκιμή (train/test/1st_expert/2nd_expert/dagger)
:
OBS 31.0.3 - Profile: Untitled - Scenes: Untitled
File Edit View Docks Profile Scene Collection Tools Help
50% Scale to Window
No source selected
Properties
Filters
Scenes
Scene
Sources
Window C
Display Ca
Audio Mixer
Desktop Audio -inf dB
Mic/Aux -inf dB
Scene Transitions
Fade
Duration 300 ms
Controls
Start Streaming
Stop Recording
Start Virtual Camera
Studio Mode
Settings
Exit
00:00:00
00:00:00
CPU: 2.2%
60.00 / 60.00 FPS

Imitation Learning



Σύνηθες πρόβλημα:
Φτάνει σε καταστάσεις που δεν έχει δει ο expert

Anaconda Prompt

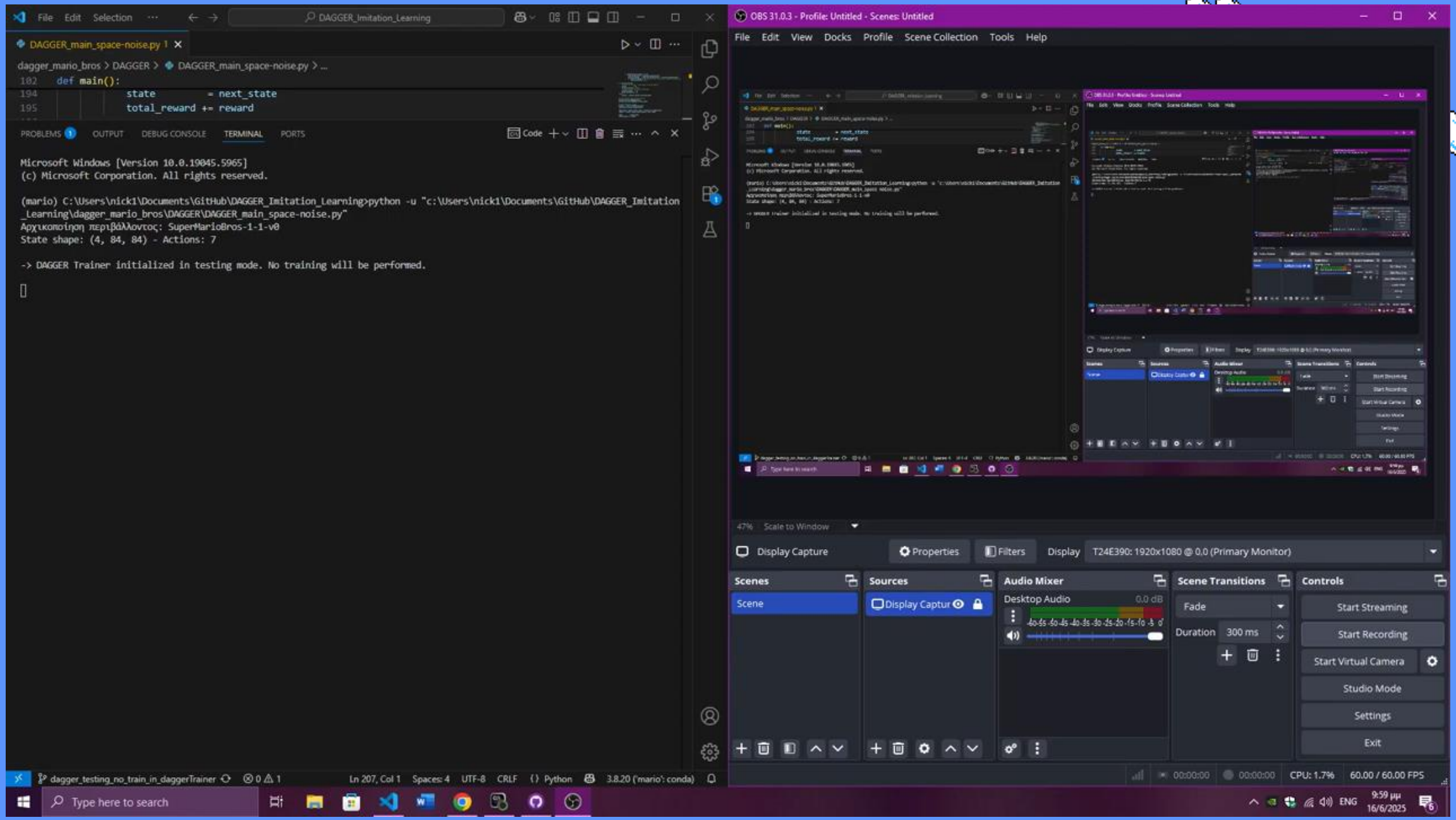
```
BehaviorCloningNetwork(
  (conv): Sequential(
    (0): Conv2d(4, 32, kernel_size=(8, 8), stride=(4, 4))
    (1): ReLU()
    (2): Conv2d(32, 64, kernel_size=(4, 4), stride=(2, 2))
    (3): ReLU()
    (4): Conv2d(64, 64, kernel_size=(3, 3), stride=(1, 1))
    (5): ReLU()
  )
  (classifier): Sequential(
    (0): Linear(in_features=3136, out_features=512, bias=True)
    (1): ReLU()
    (2): Dropout(p=0.5, inplace=False)
    (3): Linear(in_features=512, out_features=256, bias=True)
    (4): ReLU()
    (5): Dropout(p=0.3, inplace=False)
    (6): Linear(in_features=256, out_features=7, bias=True)
  )
)
[DEBUG] BEHAVIOR_CLONING.PY 325 calling load_model
C:\Users\Mania\Documents\GitHub\DAGGER_Imitation_Learning\dagger_mario_bros\CLONING\behavior_cloning.py:331: FutureWarning: You are using `torch.load` with
`weights_only=False` (the current default value), which uses the default pickle module implicitly. It is possible to construct malicious pickle data which w
ill execute arbitrary code during unpickling (See https://github.com/pytorch/pytorch/blob/main/SECURITY.md#untrusted-models for more details). In a future r
elease, the default value for `weights_only` will be flipped to `True`. This limits the functions that could be executed during unpickling. Arbitrary object
s will no longer be allowed to be loaded via this mode unless they are explicitly allowlisted by the user via `torch.serialization.add_safe_globals`. We rec
ommend you start setting `weights_only=True` for any use case where you don't have full control of the loaded file. Please open an issue on GitHub for any i
ssues related to this experimental feature.
  checkpoint = torch.load(filepath, map_location=self.device)
⚠️  Loading weights only (raw state_dict).
tester.py:55: FutureWarning: You are using `torch.load` with `weights_only=False` (the current default value), which uses the default pickle module implicit
ly. It is possible to construct malicious pickle data which will execute arbitrary code during unpickling (See https://github.com/pytorch/pytorch/blob/main/
SECURITY.md#untrusted-models for more details). In a future release, the default value for `weights_only` will be flipped to `True`. This limits the functio
ns that could be executed during unpickling. Arbitrary objects will no longer be allowed to be loaded via this mode unless they are explicitly allowlisted b
y the user via `torch.serialization.add_safe_globals`. We recommend you start setting `weights_only=True` for any use case where you don't have full control
 of the loaded file. Please open an issue on GitHub for any issues related to this experimental feature.
  state_dict = torch.load(filename, map_location=agent.device)
🎬 Παρακολούθηση του agent...

✅ Ολοκληρώθηκε! Score: 1886.0, X: 1977, Flag: False

(mario) C:\Users\Mania\Documents\GitHub\DAGGER_Imitation_Learning\dagger_mario_bros\CLONING>python tester.py
```

DAgger (Dataset Aggregation)

Ο DAGGER υπερβαίνει τους περιορισμούς του Behavior Cloning επειδή διορθώνει το distributional shift!


File Edit Selection
DAGGER_Imitation_Learning
DAGGER_main_space-noise.py 1
dagger_mario_bros > DAGGER > DAGGER_main_space-noise.py > ...
102 def main():
194 state = next_state
195 total_reward += reward
PROBLEMS 1 OUTPUT DEBUG CONSOLE TERMINAL PORTS
Code
Microsoft Windows [Version 10.0.19045.5965]
(c) Microsoft Corporation. All rights reserved.
(mario) C:\Users\nick1\Documents\GitHub\DAGGER_Imitation_Learning>python -u "c:\Users\nick1\Documents\GitHub\DAGGER_Imitation_Learning\dagger_mario_bros\DAGGER\DAGGER_main_space-noise.py"
Αρχικοποίηση περιβάλλοντος: SuperMarioBros-1-1-v0
State shape: (4, 84, 84) - Actions: 7
-> DAGGER Trainer initialized in testing mode. No training will be performed.
dagger_testing_no_train_in_daggerTrainer
Ln 207, Col 1 Spaces: 4 UTF-8 CRLF Python 3.8.20 ('mario': conda)
OBS 31.0.3 - Profile: Untitled - Scenes: Untitled
File Edit View Docks Profile Scene Collection Tools Help
47% Scale to Window
Display Capture
Properties
Filters
Display
T24E390: 1920x1080 @ 0,0 (Primary Monitor)
Scenes
Scene
Sources
Display Captur
Audio Mixer
Desktop Audio
0.0 dB
Scene Transitions
Fade
Duration
300 ms
Controls
Start Streaming
Start Recording
Start Virtual Camera
Studio Mode
Settings
Exit
00:00:00
00:00:00
CPU: 1.7%
60.00 / 60.00 FPS
Type here to search
ENG
9:59 μμ
16/6/2025

DAgger με Behaviour Cloning Warm up

Dagger + Σύντομο στάδιο behavior cloning → Expert agreement → Επιτάχυνση σύγκλισης του Dagger!

Σταθεροποίηση εκμάθησης της πολιτικής με ~60 επαναλήψεις, σε αντίθεση με τις ~1400 του «απλού» Dagger.

Στατιστική Αξιολόγηση (final loss)

Plain DAGGER

- Median: 0.41
- 25%-75%: 0.34 - 0.42
- 10%-90%: 0.24 - 0.45
- Min-Max: 0.04 - 0.48

DAGGER + BC Warmup

Improvement: -55.5%

- Median: 0.18
- 25%-75%: 0.07 - 0.33
- 10%-90%: 0.05 - 0.38
- Min-Max: 0.04 - 0.42

STAGE COMPLETION:

Plain DAGGER: 70.0% (20 runs)
BC Warmup DAGGER: 100.0% (20 runs)
Improvement: +30.0%

Σας ευχαριστούμε για τον χρόνο σας!