



SAPIENZA
UNIVERSITÀ DI ROMA

DEPARTMENT OF COMPUTER, CONTROL AND
MANAGEMENT ENGINEERING

Atari Breakout with LTL_f/LDL_f Goals

ELECTIVE IN ARTIFICIAL INTELLIGENCE:
REASONING ROBOTS

Professor:
Giuseppe De Giacomo

Students:
Ivan Bergonzani
Michele Cipriano
Armando Nania

Contents

1	Introduction	2
2	Reinforcement Learning	3
2.1	Q-Learning	3
2.2	SARSA	4
3	LTL_f/LDL_f Non-Markovian Rewards	6
3.1	Theoretical Background	6
3.2	Examples	6
4	OpenAI Gym	8
4.1	Framework	8
4.2	Examples	8
5	Atari Breakout	10
5.1	PyGame Breakout	10
5.2	Arcade Learning Environment	11
5.3	Implementation	11
5.3.1	Robot Features Extractor	11
5.3.2	Goal Features Extractor	13
5.3.3	Temporal Goals	14
5.4	Experiments	15
6	Conclusion	16

1 Introduction

Introduction to the whole project, structure of the report and summary of the work.

2 Reinforcement Learning

Reinforcement learning [1] is an area of machine learning which aims at studying how to develop agents that can interact with their environment maximizing a cumulative reward. The environment can be formally defined as a Markov Decision Process (MDP), which is a tuple $\langle S, A, \delta, R \rangle$ where S is a finite set of states that can represent the environment, A is a finite set of actions that can be performed by an agent in the environment, δ is a probability function modeling the transition from a state to another when performing a certain action and R is a reward function which models the reward received by the environment when performing a certain action which makes the agent move from a state to another.

An interesting property of the MDP is that it satisfies the Markov property, hence, the future states that will be reached by the agent do not depend on the past interaction of the environment, but just on the current state. This makes it possible to define the transition and the reward function depending only on the current state (and of course the action and the future state of interest).

This section considers two common reinforcement learning algorithms, namely Q-Learning and SARSA, which have been used in our experiments in order to train an agent interacting with an Atari Breakout environment (section 5.4).

2.1 Q-Learning

Q-Learning is a temporal difference (TD) algorithm that directly approximates the optimal action-value function. This method guarantees to find an optimal behaviour under the assumption that all state-action pairs are updated infinitely many times. It is defined [1] by the following equation:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t) \right] \quad (1)$$

Let's briefly discuss the implementation used in our project by studying the Python implementation (Algorithm 1). The algorithm is defined by the class `QLearning` that extends the abstract class `TDBrain`. The constructor of the class (lines 2-4) simply calls its parent constructor that will initialize the parameters of the object, hence, the observation space and the action space (`gym` objects), the strategy used by the policy function, which is ε -greedy by default and the hyperparameters γ , α and λ of the upper class. The abstract method inherited from `TDBrain` is `update_Q`, which should be implemented in order to define how to update the action-state table. The method (lines 6-21) simply follows Eq. 1.

TODO: eligibility.

Algorithm 1: Q-Learning algorithm Python implementation.

```
1 class QLearning(TDBrain):
2     def __init__(self, observation_space:Discrete,
3         ↪ action_space, policy:Policy=EGreedy(),
4         ↪ gamma=0.99, alpha=None, lambda_=0):
5         super().__init__(observation_space, action_space,
6         ↪ policy, gamma, alpha, lambda_)
7
8     def update_Q(self, obs:AgentObservation):
9         state, action, reward, state2 = obs.unpack()
10
11         action2 = self.choose_action(state2)
12         Qa = np.max(self.Q[state2])
13         actions_star = np.argwhere(self.Q[state2] == Qa).
14         ↪ flatten().tolist()
15
16         delta = reward + self.gamma * Qa - self.Q[state][
17         ↪ action]
18         for (s, a) in set(self.eligibility.traces.keys()):
19             self.Q[s][a] += self.alpha.get(s,a) * delta *
20             ↪ self.eligibility.get(s, a)
21             if action2 in actions_star:
22                 self.eligibility.update(s, a)
23             else:
24                 self.eligibility.to_zero(s, a)
25
26         return action2
```

2.2 SARSA

A similar TD algorithm is the SARSA algorithm, which name comes from the fact that at each timestep a quintuple $\langle S_t, A_t, R_{t+1}, S_{t+1}, A_{t+1} \rangle$ is considered. As before, SARSA converges to an optimal action-value function under the assumption that all state-action pairs are updated infinitely many times. It is defined [1] by the following equation:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha [R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)] \quad (2)$$

Let's briefly discuss the implementation used in our project by studying the implementation (Algorithm 2), as done before with the Q-Learning algorithm. Again, the algorithm is defined by the class **Sarsa** that extends the abstract class **TDBrain**. The constructor of the class (lines 2-4) calls its parent constructor initializing the parameters of the upper class exactly in the same way as the class **QLearning**. The class implements the inherited method **update_Q** by following Eq. 2 (lines 6-17). TODO: eligibility.

Algorithm 2: SARSA algorithm Python implementation.

```
1 class Sarsa(TDBrain):
2     def __init__(self, observation_space:Discrete,
3         ↪ action_space, policy:Policy=EGreedy(),
4         ↪ gamma=0.99, alpha=None, lambda_=0.0):
5         super().__init__(observation_space, action_space,
6         ↪ policy, gamma, alpha, lambda_)
7
8     def update_Q(self, obs:AgentObservation):
9         state, action, reward, state2 = obs.unpack()
10
11         action2 = self.choose_action(state2)
12         Qa = self.Q[state2][action2]
13
14         delta = reward + self.gamma * Qa - self.Q[state][
15         ↪ action]
16
17         for (s, a) in set(self.eligibility.traces.keys()):
18             self.Q[s][a] += self.alpha.get(s,a) * delta *
19             ↪ self.eligibility.get(s, a)
20             self.eligibility.update(s, a)
21
22         return action2
```

3 LTL_f/LDL_f Non-Markovian Rewards

Recently, non-Markovian reward decision processes (NMRDPs) has attracted interest in the scientific community because of the possibility of specifying them as MDPs with LTL_f/LDL_f non-Markovian rewards [4]. In particular, it is possible to model the problem with two separate representations of the world, one for the agent (low-level) and one for the goal (expressed in terms of high-level fluents).

This section presents the approach used in [4], where an efficient method has been developed in order to work with NMRDPs. The theory behind the main idea is quickly described and an example on a theoretical Breakout environment is discussed in order to be used in the following sections easily.

3.1 Theoretical Background

Before describing the problem, let's give a formal definition of NMRDP. A non-Markovian reward decision process is a tuple $M = \langle S, A, \delta, \bar{R} \rangle$, with S finite set of states that can represent the environment, A is a finite set of actions that can be performed by an agent in the environment, δ is a probability function modeling the transition from a state to another when performing a certain action and $\bar{R} : (S \times A)^* \rightarrow \mathbb{R}$ is a function from finite state-action sequences (traces) to real-values that represents the reward given by the environment when performing a certain state-action sequence. Specifying a non-Markovian reward function explicitly is difficult even when considering a finite number of traces. Luckily, the LTL_f/LDL_f formalism allows to specify \bar{R} implicitly using a set of pairs $\{(\phi, r)\}$ with ϕ boolean proposition over the components of the state vector and r such that, given a current trace $\pi = \langle s, a_1, \dots, s_{n-1}, a_n \rangle$, the agent receives at s_n a reward r if ϕ_i is satisfied by π , hence:

$$\bar{R}(\pi) = \begin{cases} r & \text{if } \pi \models \phi \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Introduction to the research paper.

3.2 Examples

How it can be used to train a RL model.

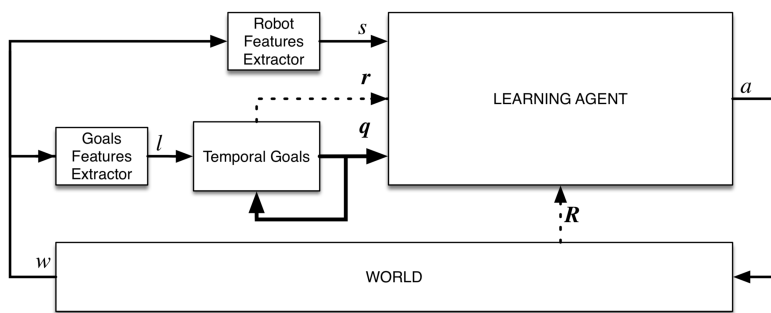


Figure 1: TODO: description.

4 OpenAI Gym

OpenAI gym [2] is a toolkit for developing and comparing reinforcement learning algorithms, without making assumptions about the structure of the agent interacting with the environment, in order to keep development flexible to updates on both sides.

4.1 Framework

The framework of gym allows to interact easily with an environment, giving the developers to tools they need to perform actions and to observe the state of the environment itself. In this way it is possible to focus more on the development of the agent without spending time on the structure of the world.

gym makes it possible to interact with multiple kinds of environments. Among these, the authors of the framework developed the support for Arcade Learning Environment [3], which includes all the classing Atari games, including Breakout, which has been used in this project.

4.2 Examples

Let's consider a simple example to understand how gym works and how the framework can be used to interact with an environment. The description will follow Algorithm 3.

Algorithm 3: Example of a random interaction with the gym environment BreakoutNoFrameskip-v4, used also in our experiments of subsection 5.4.

```
1 | import gym
2 |
3 | env = gym.make("BreakoutNoFrameskip-v4")
4 | env.reset()
5 |
6 | for _ in range(1000):
7 |     env.render()
8 |     action = env.action_space.sample() # takes random actions
9 |     observation, reward, done, info = env.step(action)
10 |    if done == True:
11 |        env.reset()
12 |
13 | env.close()
```

Initially (line 1) the framework is imported. Then (line 3-4) an environment is created specifying its name and initializing it. The program makes a random agent interact randomly with the environment for 1000 episodes (lines 6-11) before closing the environment. Line 7 renders the current observation of the

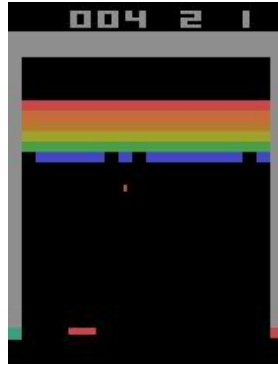


Figure 2: Observation of a frame of the environment `BreakoutNoFrameskip-v4`.

environment on screen, line 8-9 performs a random action between those available in this Brekout version, note that the method `step` return an `observation` (shown in Fig. 2), which is an array of pixels that represent the current state of the environment, a `reward`, which is a value return by the game after performing the specified action `action`, a boolean value `done`, which is `True` is the game is over, `False` otherwise, and `info` which contains extra information about the game. Lines 10-11 handles the case when the game is over, resetting the environment.

5 Atari Breakout

This section contains the main part of the project, describing in detail our starting point, how the program has been developed, the results achieved and the comparison with the original implementation, which uses a non-Atari version of the game Breakout [4] with the same number of bricks used in the Atari implementation of `gym`, already introduced in section 5.2.

Initially, the non-Atari version of Breakout (built on PyGame) is introduced, its implementation is discussed and the results on the training with a brick matrix of dimension 6×18 are presented. Then the `gym` environment `BreakoutNoFrameskip-v4` is presented and compared with the PyGame breakout. A detailed description of the implementation of the project is described, discussing in detail how the features have been extracted from the environment (both the robot and the goal features), how temporal goals have been used to evaluate the state of the bricks and how everything is connected together in order to make it work with `gym`. In the end, the main experiments performed on `BreakoutNoFrameskip-v4` are presented.

5.1 PyGame Breakout

As introduced above, in [4] a PyGame version of the Breakout game has been used in order to test the algorithms introduced in the paper. The implementation of this Breakout easily allows to determine the state of the environment, saving the status of the bricks (either in the scene or broken), the position of the ball, the direction of the ball and the position of the paddle. This makes it possible to reduce a lot the computational time of the implementation of the agent since the data it receives are already preprocessed in order to have a complete overview of the environment, allowing to focus on higher-level reasoning tasks.

Originally, the paper focused on a Breakout environment with a brick matrix of dimension 4×5 . Since the Atari version of Breakout is dealing with a brick matrix of dimension 6×18 , a new test has been performed to make the two environments comparable, in order to better understand the potentiality of LTL_f/LDL_f , the use of non-Markovian rewards and how to approach complex `gym` environments in the future.

The method managed to correctly break all the bricks in around one hour of training. TODO: add three figures with initial state, middle, final. TODO: add plot of rewards (shall this go to Experiments subsec?).

5.2 Arcade Learning Environment

This work aims at comparing the `gym` environment `BreakoutNoFrameskip-v4`, already introduced in section with the non-Atari Breakout used in [4]. In last years, the reinforcement learning community has grown a lot thanks to the introduction of `gym` and deep reinforcement learning algorithms [5] that managed to easily solve complex games that are considered difficult also for humans, often achieving better results than expert human gamers. More and more algorithms are introduced every year, exploiting GPU resources and managing to solve harder games like Montezuma Revenge [6]. The popularity of deep reinforcement learning begun with the introduction of Arcade Learning Environment (ALE) that includes most famous arcade Atari games [3].

The main characteristics of `gym` (or ALE) environments is that the world can be observed only from the pixels of the screen, putting the algorithms at the same level of the human, that can only observe the display while playing. This makes the game a lot more complex since a more abstract reasoning strategy is needed in order to solve the game. This hypothesis should make `gym` games a lot harder than the non-Atari version of Breakout that has been used to test algorithms that work with non-Markovian rewards.

5.3 Implementation

Intro to implementation.

5.3.1 Robot Features Extractor

`RobotFeatureExtractor` (OpenCV). Extracts features of the robot (robot and ball positions).

Algorithm 4: Robot feature extractor Python implementation.

```
1 class BreakoutNRobotFeatureExtractor(  
2     ↳ BreakoutRobotFeatureExtractor):  
3     def __init__(self, obs_space):  
4         robot_feature_space = Tuple((  
5             Discrete(287),  
6             Discrete(157),  
7         ))  
8  
9         self.prev_ballX = 0  
10        self.prev_ballY = 0  
11        self.prev_paddleX = 0  
12        self.still_image = True  
13  
14        super().__init__(obs_space, robot_feature_space)
```

```

15
16     def _extract(self, input, **kwargs):
17         self.still_image = not self.still_image
18         if self.still_image:
19             return (self.prev_ballX-self.prev_paddleX+143,
20                     ↪ self.prev_ballY)
21         # Extract position of the paddle:
22         paddle_img = input[189:193,8:152,:]
23         gray = cv2.cvtColor(paddle_img, cv2.COLOR_RGB2GRAY)
24         thresh = cv2.threshold(gray, 60, 255, cv2.
25             ↪ THRESH_BINARY)[1]
26         cnts = cv2.findContours(thresh.copy(), cv2.
27             ↪ RETR_EXTERNAL, cv2.CHAIN_APPROX_SIMPLE)
28         cnts = cnts[0] if imutils.is_cv2() else cnts[1]
29         min_distance = np.inf
30         paddleX = self.prev_paddleX
31         for c in cnts:
32             M = cv2.moments(c)
33             if M["m00"] == 0:
34                 continue
35             pX = int(M["m10"] / M["m00"])
36             if abs(self.prev_paddleX - pX) < min_distance:
37                 min_distance = abs(self.prev_paddleX - pX)
38                 paddleX = pX
39
40         # Extract position of the ball:
41         ballX = self.prev_ballX
42         ballY = self.prev_ballY
43         ballspace_img = input[32:189,8:152,:]
44         lower = np.array([200, 72, 72], dtype=np.uint8)
45         upper = np.array([200, 72, 72], dtype=np.uint8)
46         mask = cv2.inRange(ballspace_img, lower, upper)
47         cnts = cv2.findContours(mask.copy(), cv2.
48             ↪ RETR_EXTERNAL, cv2.CHAIN_APPROX_SIMPLE)
49         cnts = cnts[0] if imutils.is_cv2() else cnts[1]
50         for c in cnts:
51             M = cv2.moments(c)
52             # Avoid to compute position of the ball if M["
53             ↪ m00"] is zero:
54             if M["m00"] == 0:
55                 continue
56             # Calculate the centroid
57             cX = int(M["m10"] / M["m00"])
58             cY = int(M["m01"] / M["m00"])
59             # Check that the centroid is actually part of
60             ↪ the ball:
61             left_black = False
62             right_black = False
63             if cX > 3:
64                 if ballspace_img[cY][cX-3][0] != 200 or \
65                     ballspace_img[cY][cX-3][1] != 72 or \
66                     ballspace_img[cY][cX-3][2] != 72:
67                     left_black = True
68             else:

```

```

63         if ballspace_img[cY][cX+3][0] != 200 or \
64            ballspace_img[cY][cX+3][1] != 72 or \
65            ballspace_img[cY][cX+3][2] != 72:
66             right_black = True
67         if left_black or right_black:
68             ballX = cX
69             ballY = cY
70
71         self.prev_ballX = ballX
72         self.prev_ballY = ballY
73         self.prev_paddleX = paddleX
74
75         return (self.prev_ballX - self.prev_paddleX + 143,
76                 ↪ self.prev_ballY)

```

5.3.2 Goal Features Extractor

GoalFeatureExtractor (OpenCV). Extracts 6x18 table representation of the bricks in order to evaluate a formula.

Algorithm 5: Goal feature extractor Python implementation.

```

1 class BreakoutGoalFeatureExtractor(FeatureExtractor):
2     def __init__(self, obs_space, bricks_rows=6,
3         ↪ bricks_cols=18):
4         self.bricks_rows = bricks_rows
5         self.bricks_cols = bricks_cols
6         output_space = Box(low=0, high=1, shape=(
7             ↪ bricks_cols, bricks_rows), dtype=np.uint8)
8         super().__init__(obs_space, output_space)
9
10    def _extract(self, input, **kwargs):
11        bricks_features = np.ones((self.bricks_cols, self.
12            ↪ bricks_rows))
13        for row, col in itertools.product(range(self.
14            ↪ bricks_rows), range(self.bricks_cols)):
15            # Pixel of the observation to check:
16            px_upper_left = int( 8 + 8 * col)
17            py_upper_left = int(57 + 6 * row)
18            px_upper_right = int(15 + 8 * col)
19            py_upper_right = int(57 + 6 * row)
20
21            # Checking max because the input has 3 channels
22            ↪ :
23            if max(input[py_upper_left][px_upper_left]) ==
24                ↪ 0 or \
25                max(input[py_upper_right][px_upper_right])
26                ↪ == 0:
27                bricks_features[col][row] = 0
28
29        return bricks_features

```

*Ext used to improve implementation.

5.3.3 Temporal Goals

LTL_f/LDL_f implementation (with Marco Favorito libraries).

Algorithm 6: LTL_f/LDL_f formulas Python implementation.

```

1 def get_breakout_lines_formula(lines_symbols):
2     # Generate the formula string
3     # E.g. for 3 line symbols:
4     # "<(!10 & !11 & !12)*;(10 & !11 & !12);(10 & !11 & !12
        ↳ )*(10 & 11 & !12); (10 & 11 & !12)*; 10 & 11 &
        ↳ 12>tt"
5     pos = list(map(str, lines_symbols))
6     neg = list(map(lambda x: "!" + str(x), lines_symbols))
7
8     s = "(%s)*" % " & ".join(neg)
9     for idx in range(len(lines_symbols)-1):
10        step = " & ".join(pos[:idx + 1]) + " & " + " & ".
            ↳ join(neg[idx + 1:])
11        s += ";({0});({0})*".format(step)
12    s += ";(%s)" % " & ".join(pos)
13    s = "<%s>tt" % s
14
15    return s
16
17 class BreakoutCompleteLinesTemporalEvaluator(
18     ↳ TemporalEvaluator):
19     """Breakout temporal evaluator for delete columns from
20     ↳ left to right"""
21
22     def __init__(self, input_space, bricks_cols=3,
23     ↳ bricks_rows=3, lines_num=3, gamma=0.99,
24     ↳ on_the_fly=False):
25         assert lines_num == bricks_cols or lines_num ==
26         ↳ bricks_rows
27         self.line_symbols = [Symbol("l%s" % i) for i in
28         ↳ range(lines_num)]
29         lines = self.line_symbols
30
31         parser = LDLfParser()
32
33         string_formula = get_breakout_lines_formula(lines)
34         print(string_formula)
35         f = parser(string_formula)
36         reward = 10000
37
38         super().__init__(BreakoutGoalFeatureExtractor(
39         ↳ input_space, bricks_cols=bricks_cols,
40         ↳ bricks_rows=bricks_rows),
41         ↳ set(lines),
42         ↳ f,
43         ↳ reward,
44         ↳ gamma=gamma,
45         ↳ on_the_fly=on_the_fly)

```

```

39
40     @abstractmethod
41     def fromFeaturesToPropositional(self, features, action,
42         ↪ *args, **kwargs):
43         ↪ """map the matrix bricks status to a propositional
44         ↪ formula
45         ↪ first dimension: columns
46         ↪ second dimension: row
47         ↪ """
48         matrix = features
49         lines_status = np.all(matrix == 0.0, axis=kwargs["
50         ↪ axis"])
51         result = set()
52         sorted_symbols = reversed(self.line_symbols) if
53         ↪ kwargs["is_reversed"] else self.line_symbols
54         for rs, sym in zip(lines_status, sorted_symbols):
55             if rs:
56                 result.add(sym)
57
58         return frozenset(result)
59
60 class BreakoutCompleteRowsTemporalEvaluator(
61     ↪ BreakoutCompleteLinesTemporalEvaluator):
62     ↪ """Temporal evaluator for complete rows in order"""
63
64     def __init__(self, input_space, bricks_cols=3,
65         ↪ bricks_rows=3, bottom_up=True, gamma=0.99,
66         ↪ on_the_fly=False):
67         ↪ super().__init__(input_space, bricks_cols=
68         ↪ bricks_cols, bricks_rows=bricks_rows,
69         ↪ lines_num=bricks_rows, gamma=gamma,
70         ↪ on_the_fly=on_the_fly)
71         ↪ self.bottom_up = bottom_up
72
73     def fromFeaturesToPropositional(self, features, action,
74         ↪ *args, **kwargs):
75         ↪ """complete rows from bottom-to-up or top-to-down,
76         ↪ depending on self.bottom_up"""
77         ↪ return super().fromFeaturesToPropositional(features
78         ↪ , action, axis=0, is_reversed=self.bottom_up
79         ↪ )

```

Atari wrappers (OpenAI).

5.4 Experiments

Results with 6x18 non-ATARI Breakout (+CODE).

Results with our experiments (+CODE).

6 Conclusion

Why it does not work.

Summary + differences between the two environments.

Future works (neural networks and parallel computation).

References

- [1] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. The MIT Press, second ed., 2018.
- [2] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, “OpenAI Gym,” 2016.
- [3] M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling, “The Arcade Learning Environment: An Evaluation Platform for General Agents,” *Journal of Artificial Intelligence Research*, vol. 47, pp. 253–279, jun 2013.
- [4] G. De Giacomo, L. Iocchi, M. Favorito, and F. Patrizi, “Reinforcement Learning for LTLf/LDLf Goals,” *CoRR*, vol. abs/1807.06333, 2018.
- [5] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, “Human-level control through deep reinforcement learning,” *Nature*, vol. 518, pp. 529–533, Feb. 2015.
- [6] “Montezuma’s Revenge Solved by Go-Explore, a New Algorithm for Hard-Exploration Problems (Sets Records on Pitfall, Too).” <https://eng.uber.com/go-explore/>.