

ÉCOLE POLYTECHNIQUE FEDERALE DE LAUSANNE
MT SECTION
FAURY Louis

SEMESTER PROJECT REPORT
LASA



ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Learning from noisy demonstrations: the
role of compliance in exploration-exploitation
trade-off

Laboratory Learning Algorithms and Systems Laboratory
Supervisors Mahdi Khoramshahi, Sutcliffe Andrew
Semester Spring

Contents

1	Introduction	2
1.1	Motivations	2
1.2	Background	2
1.2.1	Learning from demonstration	3
1.2.2	Transfer learning	3
1.3	Proposed approach	4
2	Reinforcement Learning	5
2.1	Formulation	5
2.1.1	Notations and first definitions	5
2.1.2	Markov decision process	5
2.1.3	State and action value function	6
2.1.4	Optimal policies	6
2.2	Dynamic Programming	7
2.2.1	Generalized policy iteration	7
2.2.2	The value iteration algorithm	7
2.3	Temporal differences methods	8
2.3.1	On-policy method : SARSA	8
2.3.2	Off-policy method : Q-learning	8
2.3.3	Eligibility traces	9
2.4	Grid-world examples	10
2.4.1	Dynamic Programming solving	10
2.4.2	SARSA solving	11
2.4.3	Q-learning solving	11
3	Compliant-based reinforcement learning	13
3.1	Problem selection	13
3.2	First naive approaches	14
3.2.1	Constant compliance	14
3.2.2	Constantly decreasing compliance	15
3.3	Adaptive compliance approach	18
3.3.1	Actor-critic approach	18
3.3.2	Action value approach	18
4	Results	20
4.1	Actor-critic approach	20
4.1.1	Practical considerations	20
4.1.2	Results	20
4.1.3	Discussion	21
4.2	Action value approach	22
4.2.1	Practical considerations	22
4.2.2	Results	22
4.2.3	Discussion	23
4.3	Method comparison	23
Conclusion		25

Chapter 1

Introduction

1.1 Motivations

For long and complex tasks, usual machine learning algorithms are often unable to come up with feasible solutions, or at least fail to do so in acceptable computational times. A natural way to accelerate the learning process is to provide a learning algorithm with prior belief on its environment, as well as recommandations with respect to the task it is trying to learn.

Hence, we provide the learning algorithm with demonstrations, much like we would do with a child in the process of learning a skill (dancing, kicking a ball,..). We will use such an example in the following paragraph in order to highlight the different sources of difficulties such an approach might arise. Because the teacher doesn't have the same physical abilities than the child, or is not an expert in the skill he is trying to teach, he might give the child a suboptimal way of performing the task. In a ideal learning process, the child would use the prior information given by the teacher, and use it to practice its skill. He will soon be aware of its own abilities, and can then start learning by itself, exploring around the way he has learnt so far. He might find that, by slightly changing how he performs the tasks, he is able to perform better than the teacher. Therefore, the learning child will first be *compliant* with the teacher, before trying things out by itself once it has become pretty good at performing the learned skill.

This semester project aims at introducing a theoretical formulation of such a compliance-based behavior, and experimentally test its performance on simple problems and compare it to other usual algorithm. We will define our problem in a reinforcement learning framework, which allows a fairly good definition of our initial problem.

The underlying goal behind this objective is to get better intuition about how interactive learning between human and robots can be achieved. Indeed, we would like to be able to teach a robot from demonstration not by only providing it a large number of trajectories that he will then use as a motion generator, but by directly interacting with him. By showing a robotic arm, at different times of its learning procedure, concrete examples on how to avoid its closest obstacle and grab a given object, seems like a more natural way of teaching the robot a task, while still enabling him to correct some small sub-optimality we could have introduced to it.

Our approach also allows use to tackle a somehow different problem, related to transfer learning. Indeed, the teacher might not be a human but another learner, only better trained than the current learner. In such a case, we would like our learner to quickly find if it can trust its mentor, and if not where it should focus its computations to overcome its mentor sub-optimality. Such questions will therefore be tackled in this report.

1.2 Background

We hereinafter introduce some notions from the Learning from Demonstration (LfD) framework, as well as some transfer learning metrics that will show useful later on.

1.2.1 Learning from demonstration

As discussed before, Learning from demonstration (LfD) is a framework where a robot can learn from interacting with a human. It particularly focuses on the case where a mentor (human) provides demonstration on how to perform a task. Such an approach is also known as Programming from demonstration (PbD) or imitation learning.

LfD mostly lies on the principle that the learning robot can be taught new tasks by end-users, without having to be programmed again. We therefore are dealing with robots that are able to generalize from demonstration, be it only to understand what is the task that they must learn. There are many underlying problematics and opened questions on the subject, detailed in [BCD16].

One approach of LfD combines with reinforcement learning techniques. A could argument for this is that LfD could easily speed-up the learning process by providing a finite or infinite set of good solutions. However, this assumes some kind of reproducibility by the robot, as well as a similar enough context between demonstration and reproduction.

Under such assumptions, RL techniques can use LfD to speed up their learning process and discover new control policies through a directed exploration of its action-state space. Indeed, demonstrations are used to guide the RL algorithm exploration, reducing the time and iterations needed for an agent to discover a new control policy, and eventually overcoming the teacher's apparent policy sub-optimalities.

There are many ways of performing such a demonstration-based learning. We remind here a few of them, taken from [ACVB09], all related to Imitation Reinforcement Learning (IRL) :

- policy derivations techniques : directly approximate the robot's mapping from state to action by reproducing and generalizing the teacher's one.
- use demonstration data to learn a model from the world's dynamic (or transition probability model) to compute the optimal policy (see [PB03]).
- use demonstration data and additional user intention to learn the rule that associates a set of pre and post-conditions which each action as well as a sparse transition model to plan for actions.

The framework we described hereinbefore can easily be extended to more than one teacher. Moreover, such teachers can actually be other learning algorithms with different levels of training. Quickly, the problem drifts toward the underlying issue of performing transfer skill between agents. As this will also be tackled during this project, the next section reminds a few useful facts about the theory of transfer learning.

1.2.2 Transfer learning

Transfer learning deals with speeding a learning process thanks to another learning experience - for instance training an agent based on the previous training on another agent, performing a similar task. The literature of transfer learning in reinforcement learning is pretty large, and for a full description of the different problematics (homogeneous / inhomogeneous sets of actions, multi task learning, ..) we'll refer to [TS09].

As we said before, one of our goal is to assess our method as a transfer learning one, and up to a point, consider our teacher to be a trained (fully or not) agent, before considering many different agents as potential mentors. We would consider a really simple case of transfer learning, where all agents display the same state-action state, with the same abilities and constraints.

Hence, we remind here some metrics to evaluate transfer learning methods that we would have to use to compare our different approaches :

- *Jump start* : the initial performance of an agent can be improved by transfer.
- *Asymptotic performance* : how close is the asymptotic learned policy from the optimal policy ?
- *Total or averaged reward* : averaged reward may be improved by transfer

- *Transfer ratio* : Ratio of total reward accumulated by the transfer learner and the total of reward accumulated by a non-transfer learner.
- *Time to threshold* : the time required by a transfer learning to reach a pre-specified level of skills in achieving its task.

All those metrics actually compared different aspects of how well a RL algorithm learns. Comparing two RL methods is actually still an open question, and hence we'll need to address several of those metrics to decide on the wellness of our derived TL methods.

1.3 Proposed approach

A reinforcement learning approach makes a lot of sense in our situation. The next chapter of this report is dedicated to derive some background on this framework, therefore we will hereinafter only describe our proposed approach of imitation learning with a reinforcement learning formulation.

In the most general case, we will consider the teacher's demonstrations to be a bijection between the state space \mathcal{S} of the robot and its action set $\mathcal{A}(s)$, $\forall s \in \mathcal{S}$, denoted π_m :

$$\begin{aligned} \pi_m : \mathcal{S} &\rightarrow \mathcal{A} \\ s &\rightarrow a_m(s) \end{aligned} \tag{1.1}$$

where $a_m(s)$ is the *recommended action* of the mentor at state s . Some other specifications could be brought later on - like considering that the mentor's recommendations don't cover the whole state space, or introducing *inhomogeneous settings*. Given those informations, we want to come up with a learning algorithm that comes up with its own mapping π , by observing π_m and eventually becoming better (we will see later how we define that notion) than its mentor at the demonstrated task.

Practically speaking, we are going to start with a fairly simple environment and explicit task in order to address the problem of compliance in learning from a teacher. We hope that a simple enough state space will allow us to better understand how compliance in learning by demonstration should be tackled with a reinforcement learning formulation. We also hope we would be able to generalize to more complex situations once the understanding on a simple but generic model is mastered.

Hence, after defining a simple two dimensional grid-world state space with a simple action set, we will quickly study how well the different classical reinforcement learning policy search algorithms performs on such a space. Once this done, we will introduce a new exploration policy, based on the learned confidence the learner has in its teacher and compare its efficiency with the latter algorithms. We will also focus on the relation between the learner and the prior information on its tasks it gains through its mentors recommendation. Especially, we will study how well a learner can overcome its mentor own sub-optimality.

Chapter 2

Reinforcement Learning

2.1 Formulation

2.1.1 Notations and first definitions

Reinforcement learning is a framework in which an *agent* (or a *learner*) learns its actions from interaction with its environment. The environment generates scalar values called *rewards*, that the agent is seeking to maximize over time.

Let \mathcal{S} denote the state space in which our agent evolves (the localization of a robot on a grid for instance), and $\forall s \in \mathcal{S}$ we will define the action state $\mathcal{A}(s)$, describing all possible action that can be taken by the agent at state s . When taking an action from a state s_t , the agent finds itself in a new state s_{t+1} where it receives a reward $r_{t+1} \in \mathbb{R}$. The action taken is sampled over a probability distribution from the joint space of state and action :

$$\begin{aligned}\pi : \mathcal{S} \times \mathcal{A}(s) &\rightarrow [0, 1] \\ (s, a) &\rightarrow \pi(s, a)\end{aligned}\tag{2.1}$$

where $\pi(s, a)$ is the probability of picking action a in state s . Such a distribution is called the agent's *policy*. The key goal of reinforcement learning is teaching an agent on how to change its policy to maximize its reward on the long run.

The agent indeed seeks to maximize the *expected return* R_t mapping the reward sequence into \mathbb{R} . A commonly used expression for this value employs a *discount factor* $\gamma \in [0, 1]$, allowing to make the agent's more sensible to rewards it will get in a close future :

$$R_t = \sum_{i=0}^T \gamma^i r_{t+1+i}\tag{2.2}$$

This also allows to adapt this formulation to continuous tasks, where there are no terminal states and the task goes on indefinitely (there are no *episodes* in the learning).

2.1.2 Markov decision process

To make the problem tractable, we ask for the state signal to comply with Markov's property, hence to be *memory-less*. For instance, we want to be able to write that, in a stochastic environment, $\forall s' \in \mathcal{S}$:

$$\mathbb{P}(s_{t+1} = s' | a_t, s_t, \dots, a_1, s_1) = \mathbb{P}(s_{t+1} = s' | s_t, a_t)\tag{2.3}$$

Hence, every reinforcement learning problem can be represented by a *Markov Decision Process*, that consists in a 5-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}(\cdot, \cdot), \mathcal{R}(\cdot), \gamma)$ where :

- ▷ \mathcal{S} is the agent's state space
- ▷ \mathcal{A} is the agent's action space
- ▷ $\forall s, s' \in \mathcal{S}, \forall a \in \mathcal{A}(s), \mathcal{P}_a(s, s') = \mathbb{P}(s_{t+1} = s' | s_t = s, a_t = a)$ is the probability that action a in state s will lead the agent to transitioning to state s' .

- ▷ $\forall s, s' \in \mathcal{S}, \forall a \in \mathcal{A}(s), \mathcal{R}_a(s, s')$ is the immediate reward perceived by the agent when transitioning from state s to s' when taking action a .
- ▷ γ is the discount factor.

A *finite Markov decision process* designates a MDP for which both the action and state space are finite.

2.1.3 State and action value function

Most of the reinforcement learning algorithms are based on evaluation value function. A value function is a function mapping the state space in \mathbb{R} , estimating how good (in terms of expected future reward) it is for the agent to be in a given space. More precisely, a value function $V^\pi(\cdot)$ evaluates the expected return of a state when following the policy π . $V^\pi(\cdot)$ is called the **state-value function**.

$$\forall s \in \mathcal{S}, \quad V^\pi(s) = \mathbb{E}_\pi [R_t | s_t = s] \quad (2.4)$$

The **action-value function** evaluates the value of taking a given action, and then following the policy π :

$$\forall s, a \in \mathcal{S} \times \mathcal{A}(s), \quad Q^\pi(s, a) = \mathbb{E}_\pi [R_t | s_t = s, a_t = a] \quad (2.5)$$

Both those functions satisfy particular recursive relationships known as the *Bellman equations*. It is shown that (see TODO quote Sutton) we have the following results :

Bellman equations for Markov Decision Process

- ▷ Bellman equation for the state-value function : $\forall s \in \mathcal{S}$

$$V^\pi(s) = \sum_{a \in \mathcal{A}} \pi(s, a) \sum_{s'} \mathcal{P}_a(s, s') [\mathcal{R}_a(s, s') + \gamma V^\pi(s')] \quad (2.6)$$

- ▷ Bellman equation for the action value function : $\forall s, a \in \mathcal{S} \times \mathcal{A}(s)$:

$$Q^\pi(s, a) = \sum_{s'} \mathcal{P}_a(s, s') [\mathcal{R}_a(s, s') + \gamma V^\pi(s')] \quad (2.7)$$

2.1.4 Optimal policies

The value functions define a partial ordering in the policy space. A policy π is therefore said to be better than π' (or $\pi \geq \pi'$) if $\forall s \in \mathcal{S}, V^\pi(s) \geq V^{\pi'}(s)$. We are looking for π^* so that :

$$\forall \pi, \quad \pi^* \geq \pi \quad (2.8)$$

TODO quote showed that for finite MDPs, there is always at least one policy that is better or equal to all others, and therefore is called the *optimal policy* π^* . As shown in TODO quote, the state-value and action-value function verify the *Bellman optimality equations*.

Bellman optimality equations

- ▷ Bellman optimality equation for the state-value function : $\forall s \in \mathcal{S}$

$$V^\pi(s) = \max_{a \in \mathcal{A}(s)} \pi(s, a) \sum_{s'} \mathcal{P}_a(s, s') [\mathcal{R}_a(s, s') + \gamma V^\pi(s')] \quad (2.9)$$

- ▷ Bellman optimality equation for the action value function : $\forall s, a \in \mathcal{S} \times \mathcal{A}(s)$:

$$Q^\pi(s, a) = \sum_{s'} \mathcal{P}_a(s, s') \left[\mathcal{R}_a(s, s') + \max_{a' \in \mathcal{A}(s')} Q(s', a') \right] \quad (2.10)$$

Those relations are essential in understanding the solving algorithms that will be presented later.

There exists several ways of solving (i.e computing the optimal policy) of a Markov Decision Process, that can generically be separated in two categories : *model-based* and *model-free* methods.

2.2 Dynamic Programming

Dynamic programming is a mathematically well-developed theory. It requires the complete and accurate model of the environment, making it a model-based method.

Dynamic programming methods aims at computing the optimal value function at every state of state space. This could, of course, be done by solving the $|\mathcal{S}|$ equations of $|\mathcal{S}|$ unknowns that are the Bellman equations for a given policy, and then evolve that policy toward a better one, based on the current value function. Of course, this approach is computationally intractable for large state space and therefore needs to be adapted, but gives a first approach of the idea behind dynamic programming.

2.2.1 Generalized policy iteration

The generalized policy iteration methods rely on alternating two processes, known as **policy evaluation** and **policy improvement**.

- ▷ Policy evaluation deals with estimating the value function of a given policy π , without directly solving the full system given by Bellman equations. The idea is actually pretty simple : use Bellman's equation as an update rule, using that the value function is a fixed point for this update rule. The algorithm, after setting the tabled value function to an initial value, iterate by performing what is called *full Bellman backups* :

$$\forall s \in \mathcal{S}, \quad V_{k+1}^{\pi}(s) = \sum_{a \in \mathcal{A}} \pi(s, a) \sum_{s'} \mathcal{P}_a(s, s') [\mathcal{R}_a(s, s') + V_k^{\pi}(s')] \quad (2.11)$$

This algorithm converges under the same assumptions that guarantee the existence of the value function, and has the generic name of *iterative policy evaluation*. They are many refining for speeding up that process (reduced backups, prioritized sweeping) that we won't address here.

- ▷ Policy evaluation is a process that from a given policy value function, returns a better or equal policy compared to the latter. The simplest way to do that is for every state $s \in \mathcal{S}$, consider every action-value functions :

$$Q(s, a) = \sum_{s'} \mathcal{P}_a(s, s') [\mathcal{R}_a(s, s') + \gamma V^{\pi}(s')], \quad a \in \mathcal{A}(s) \quad (2.12)$$

and to build π' to be *greedy* with respect to those actions-values :

$$\pi'(s) = \operatorname{argmax}_{a \in \mathcal{A}(s)} \{Q(s, a)\} \quad (2.13)$$

The policy improvement theorems then ensures that $\pi' \geq \pi$.

Hence, generalized policy improvement are a set of methods that iteratively combine those two sub-methods to compute the optimal policy for a given MDP. Of course, one does not have to perform all sweeps of value evaluation before improving the policy to converge toward an optima (indeed, many times our sweeps won't have any affect on the greedy policy). They are many ways to combine the two (prioritized sweeping, asynchronous dynamic programming), but the most used and one of the most quickest way to converge is to use the value iteration algorithm.

2.2.2 The value iteration algorithm

The value iteration algorithm takes the limit of the behavior we just described, and stops the value evaluation procedure after only *one state space sweep*. It therefore performs a simple backup procedure :

$$\forall s \in \mathcal{S}, \quad V_{k+1} = \max_{a \in \mathcal{A}} \sum_{s'} \mathcal{P}_a(s, s') [\mathcal{R}_a(s, s') + \gamma V_k^{\pi}(s')] \quad (2.14)$$

For any arbitrary V_0 , it is shown that $V_k \rightarrow V^*$ as $k \rightarrow \infty$, under the same hypothesis that ensure the existence of the optimal value function V^* . As one can notice, it actually implements the *Bellman optimality conditions* as an update rule !

2.3 Temporal differences methods

Temporal difference methods can be seen as a combination of dynamic programming and another kind of learning called Monte Carlo methods, where the expected return are approximated via sampling. Like dynamic programming, TD methods are said to bootstrap (meaning that they build their estimators through already estimated values), but are *model-free* methods and learn from experience.

The justification, proof of convergences and literature and those models is pretty wide, hence we will not cover them in this report. However, full description of those methods can be found in TODO quote.

2.3.1 On-policy method : SARSA

The SARSA algorithm is an *on-policy* control method, meaning that the algorithm updates the value function and improves the current policy it is following. At state s_t , it chooses an action a_t from its policy and follows it. After observing the reward r_{t+1} and the next state s_{t+1} , it again chooses an action a_{t+1} using a soft policy and performs a one-step backup :

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha [r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)] \quad (2.15)$$

It therefore relies on a 5-tuple $(s_t, a_t, r_{t+1}, s_{t+1}, a_{t+1})$ to perform the update, giving it the State Action Reward State Action (SARSA) name.

The General Sarsa Algorithm

1. Initialize $Q(s, a)$ arbitrarily $\forall (s, a) \in \mathcal{S} \times \mathcal{A}(s)$
2. Repeat for each episode :
 - Initialize s
 - Choose $a \in \mathcal{A}(s)$ using a soft policy derived from Q (typically ε -greedy)
 - Repeat for each step of the current episode :
 - Take a , observe r, s'
 - Choose a' from s' using policy derived from Q
 - $Q(s, a) \leftarrow Q(s, a) + \alpha[r + \gamma Q(s', a') - Q(s, a)]$
 - $a \leftarrow a', s \leftarrow s'$
 - until $s \in \mathcal{S}^+$.

The convergence properties of SARSA depend on the nature of the policy's dependency on Q . Indeed, SARSA converges with probability 1 to the optimal policy as long as all the state and actions pairs are visited an infinite number of time, and the policy converges in the limit to the greedy policy. This is done, for instance, by turning the temperate of a softmax based policy to 0, or by having $\varepsilon \rightarrow 0$ for a ε -greedy policy. For SARSA to converge, we also as the learning rate to comply with the stochastic approximation conditions :

$$\sum_k \alpha_k(a) = +\infty \quad \text{and} \quad \sum_k \alpha_k(a)^2 < +\infty \quad (2.16)$$

where $\alpha_k(a)$ is the learning rate for the k^{th} visit of the pair (s, a) .

2.3.2 Off-policy method : Q-learning

The Q-learning algorithm is an off-policy method who learns to directly approximate Q^* , independently of the policy being followed. Its update rule is given by :

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_{t+1} + \gamma \max_{a \in \mathcal{A}(s_{t+1})} Q(s_{t+1}, a) - Q(s_t, a_t) \right] \quad (2.17)$$

The actual policy being followed still has an effect in that it determines which state-actions pairs are visited and updated. However all that is required for convergence it that all pairs continue to be updated.

Q-Learning Algorithm

1. Initialize $Q(s, a)$ arbitrarily $\forall (s, a) \in \mathcal{S} \times \mathcal{A}(s)$
2. Repeat for each episode :
 - Initialize s
 - Repeat for each step of the current episode :
 - Choose $a \in \mathcal{A}(s)$ using arbitrary policy
 - Take a , observe r, s'
 - $$Q(s, a) \leftarrow Q(s, a) + \alpha [r + \gamma \max_{a' \in \mathcal{A}(s')} Q(s', a') - Q(s, a)]$$
 - $$s \leftarrow s'$$
 - until $s \in \mathcal{S}^+$.

Along with this hypothesis and a slight variation in the usual stochastic approximation conditions, the learned action value function by Q-learning has been shown to converge to Q^* with probability 1.

2.3.3 Eligibility traces

In TD(0) approach (described in the latest section), we update the value function in the direction of the *one-step return* :

$$\Delta V_t(s_t)^{(1)} = \alpha [r_t + \gamma V_t(s_{t+1}) - V_t(s_t)] \quad (2.18)$$

The idea behind eligibility traces is to expand that update rule in order to steer the value function towards the *n-step return* (or at least until a terminal state is reached) :

$$\begin{aligned} \Delta V_t(s_t)^{(n)} &= \alpha [r_t + \gamma r_{t+1} + \dots + \gamma^n V_t(s_{t+n}) - V_t(s_t)] \\ &= \alpha [R_t^{(n)} - V_t(s_t)] \end{aligned} \quad (2.19)$$

The backups can not only be done toward any n-step return, but toward any average of such returns, as long as the corresponding weights sum-up to one. In this way, the TD(λ) algorithm can be understood as a particular way of averaging n-steps returns. With $\lambda < 1$, the resulting backup is known as the *λ -return* :

$$R_t^\lambda = (1 - \lambda) \sum_{n=1}^{\infty} \lambda^{n-1} R_t^{(n)} \quad (2.20)$$

where the weights are fading with n . When the runs are episodic, we can write this return as :

$$R_t^\lambda = (1 - \lambda) \sum_{n=1}^{T-t-1} \lambda^{n-1} R_t(n) + \lambda^{T-t-1} R_t^{(T)} \quad (2.21)$$

Sarsa(λ) algorithm

1. Initialize $Q(s, a)$ arbitrarily $\forall (s, a) \in \mathcal{S} \times \mathcal{A}(s)$
2. Repeat (for each episode) :
 - Initialize s, a
 - Repeat for each step of the current episode :
 - Take action a , observe r, s' .
 - Choose $a' \in \mathcal{A}(s')$ using soft policy derived from Q
 - $\delta \leftarrow r + \gamma Q(s', a') - Q(s, a)$
 - $e(s, a) \leftarrow e(s, a) + 1$
 - For all s, a :
 - $Q(s, a) \leftarrow Q(s, a) + \alpha \delta e(s, a)$
 - $e(s, a) \leftarrow \gamma \lambda e(s, a)$
 - $s \leftarrow s', a \leftarrow a'$
 - until $s \in \mathcal{S}^+$.

Such a formulation of eligibility traces is known as the *forward view* of $\text{TD}(\lambda)$, and shows how eligibility traces build the bridge between $\text{TD}(0)$ methods and Monte-Carlo one. It is not implementable as is since it is non-causal. There exist a more mechanistic view, equivalent to the forward view (see [SB98]), known as the *backward-view*. It gives birth to causal version of the $\text{TD}(\lambda)$ method. We give as an example the pseudo-code for the $\text{SARSA}(\lambda)$ algorithm above. Its Q-learning equivalent can be found in [SB98].

2.4 Grid-world examples

We hereinafter describe two grid-world space, on which we will apply the learning rule derived in the latest section. Such example are trivial and are displayed here just to show convergence and behavior of the different algorithms.

We'll consider the two following state space :

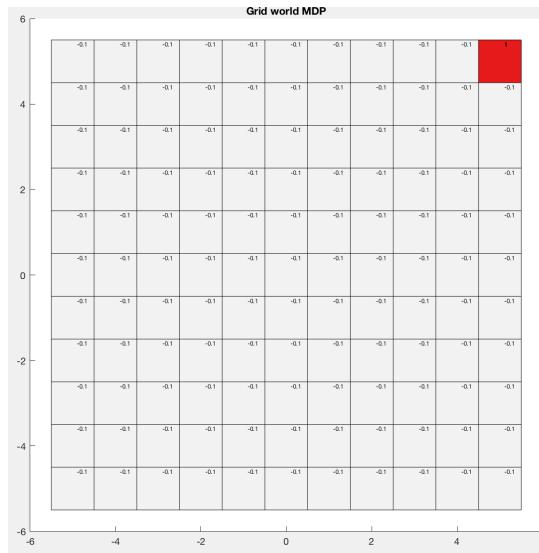


Figure 2.1: The *free_grid* state space

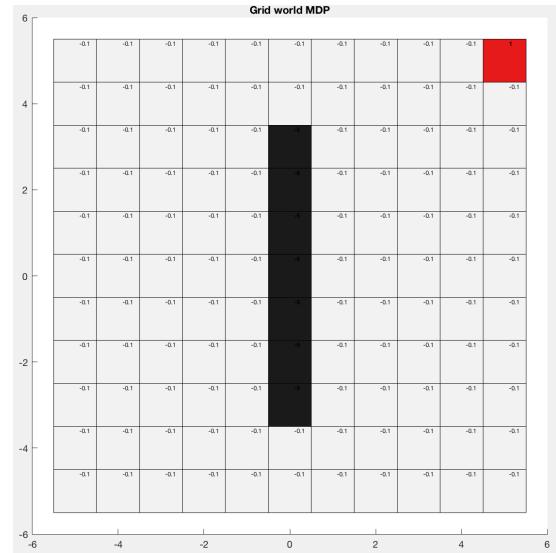


Figure 2.2: The *bar_grid* state space

2.4.1 Dynamic Programming solving

Let us run the DP algorithm on such grid worlds. We'll consider a stochastic environment, with the transition probability :

$$\mathcal{P}_{s,s'}^a = \begin{cases} 0.9 & \text{if } s' = s(a) \\ 0.1 & \text{otherwise} \end{cases} \quad (2.22)$$

Running the value iteration algorithm (assuming we now the environment model), we obtain the following policies and learning curves. The stopping criterion adresses the maximum absolute change brought to the value function as the sweeping goes through the state space :

$$\text{If } \max_{s \in \mathcal{S}} |V_{k+1}(s) - V_k| < \delta \text{ then stop} \quad (2.23)$$

One can notice that for the *bar_grid* environment, the agent undergoes a longer trajectory than necessary, at the left of the obstacle. This is because of the stochastic nature of the environment, causing the agent to learn to takes its distance with the obstacle in order not to accidentally hit it (and then receive a very negative reward).

The learned policy are indeed optimal. The next algorithms (SARSA and Q-learning) will try to reproduce them without a model for the environment.

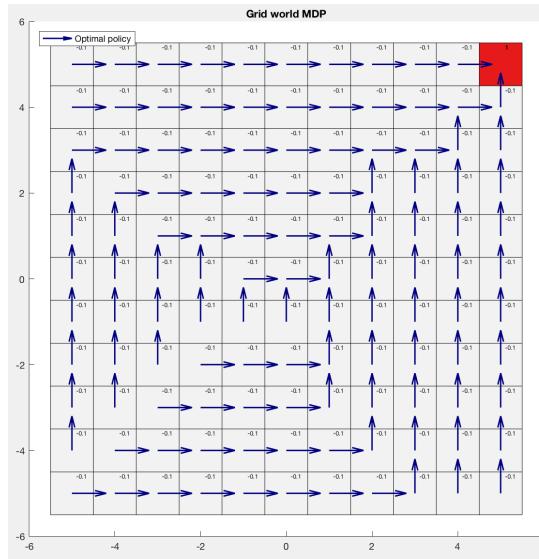


Figure 2.3: The *free_grid* learned optimal policy

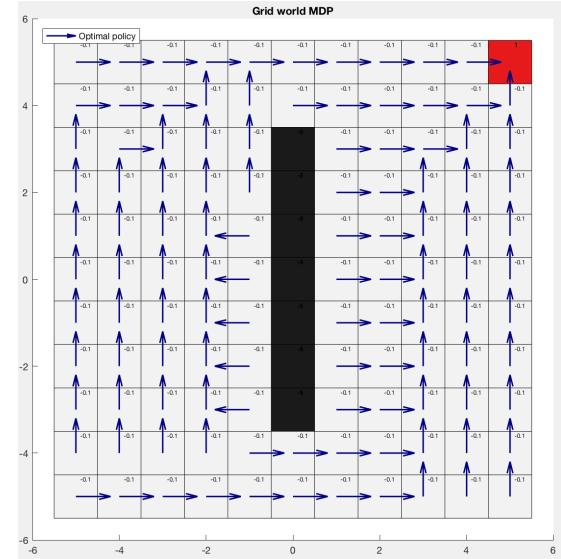


Figure 2.4: The *bar_grid* learned optimal policy

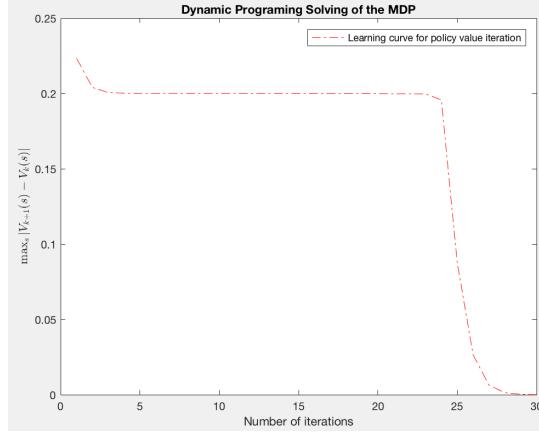


Figure 2.5: The *free_grid* value iteration learning curve

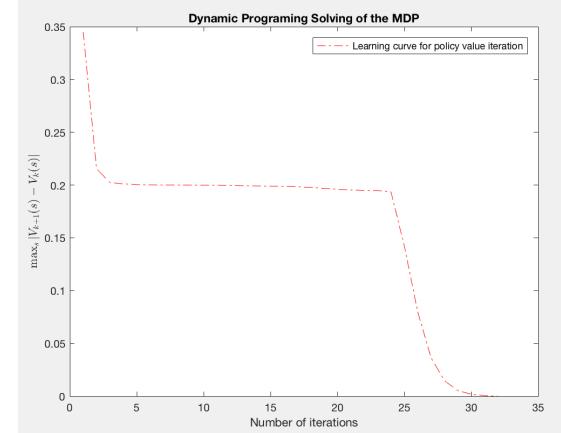


Figure 2.6: The *bar_grid* value iteration learning curve

2.4.2 SARSA solving

We display here the learning curves for the *free_grid* state space using SARSA. The algorithm manages to learn the optimal policy and the right action-value functions. We use *optimistic initialization* to encourage exploration, and Gibbs sampling for following a soft-policy : $\forall(s, a) \in \mathcal{S} \times \mathcal{A}(s)$

$$\pi(s, a) = \frac{e^{Q(s, a)/\tau}}{\sum_{a' \in \mathcal{A}(s)} e^{Q(s, a')/\tau}} \quad (2.24)$$

We'll tune the distribution's temperature τ to zero, in order to converge toward the greedy policy w.r.t the learned action-value function.

Following this strategy and tuning our learning rate to comply with the stochastic approximation conditions, we obtain the following learning curves. Again, our stopping criterion addresses the *maximum change in the action-value function over all the trajectories of a mini-batch*.

2.4.3 Q-learning solving

We display here the learning curves for the *bar_grid* state space using Q-learning. Again, the algorithm manages to learn the optimal policy and the right action-value functions. We use Gibbs sampling for the

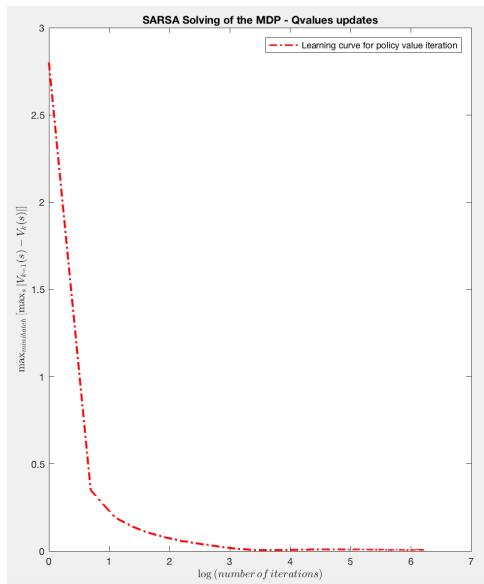


Figure 2.7: Learning curve for SARSA on *free_grid*

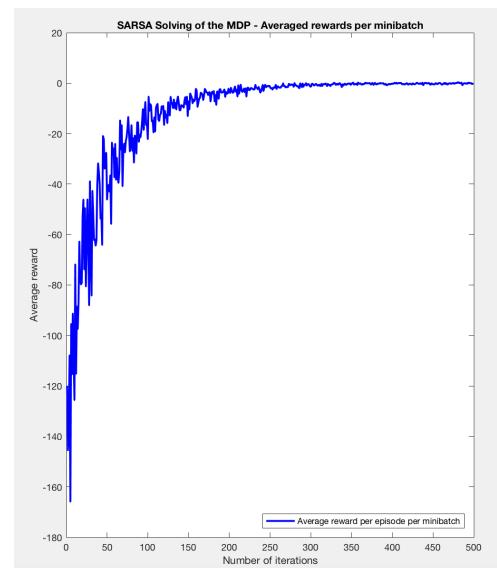


Figure 2.8: Averaged rewards over minibatch for SARSA on *free_grid*

behavior policy, without any tuning for the temperature (the behavior policy doesn't need to be greedy in limit). The learning curves obtained are given hereinafter.

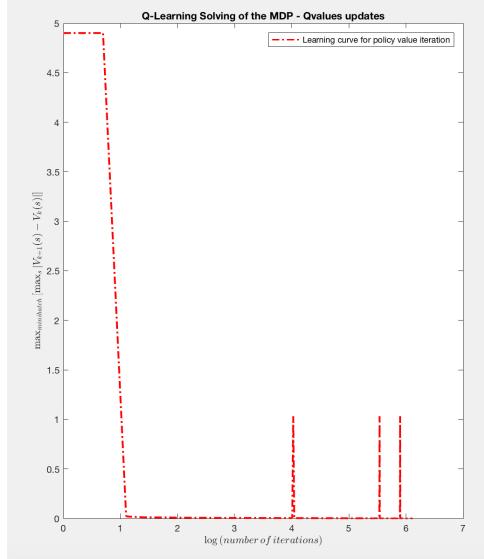


Figure 2.9: Learning curve for Q-learning on *bar_grid*

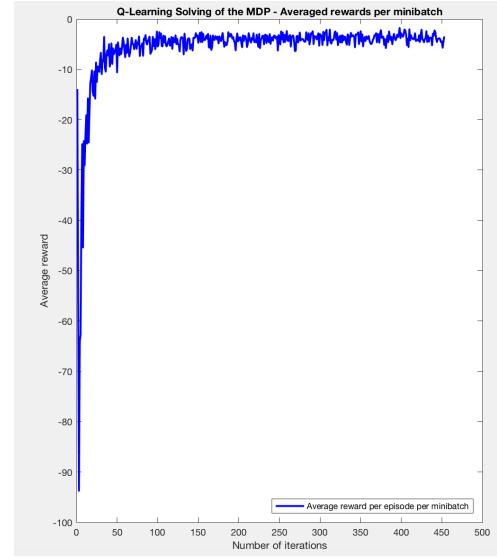


Figure 2.10: Averaged rewards over minibatch for Q-learning on *bar_grid*

Chapter 3

Compliant-based reinforcement learning

3.1 Problem selection

With our approach, we wish to tackle two topics : on one hand, we wish to display a imitation based learning framework that *accelerate* the learning process. Also, we wish to have a shifting compliance based behavior so that an agent can *overtake a suboptimal teacher*. To test the effectiveness of the solutions we would propose, we decided to provide ourself with a model that would stay fixed all along the experiments, in order to compare the different algorithms we could come up with.

We therefore needed to provide an environment that, with the algorithms described before, cannot solve its MDP quickly, and can sometimes get stuck in local minima. We chose the state space displayed in figure (3.1). In this environment, all black cells are obstacles. They give out highly negative rewards ($r = -10$). Whenever an agent take the action to enter such a cell, it immediately perceives the negative reward but stays in its current cell. The only positive reward in this environment is at the middle of the grid ($r = 10$), the only terminal state. Any trajectory of the learner starts at one of the corner of the grid (green cells), and every step spent on a non-terminal cell gives out a small negative reward ($r = -0.1$). The environment is stochastic, with the transition probability model :

$$s' = \begin{cases} a(s) \text{ with proba 0.95} \\ s'' \neq a(s) \text{ otherwise, uniformly sampled} \end{cases} \quad (3.1)$$

It is big enough for the usual algorithms to learn rather slowly, even if they are greatly enhanced by the use of eligibility traces. Also, all tested algorithms (SARSA, Q-learning, SARSA(λ) and Watkins Q(λ)), because they do not perform infinite exploitation / exploration moves, renders slightly suboptimal policies.

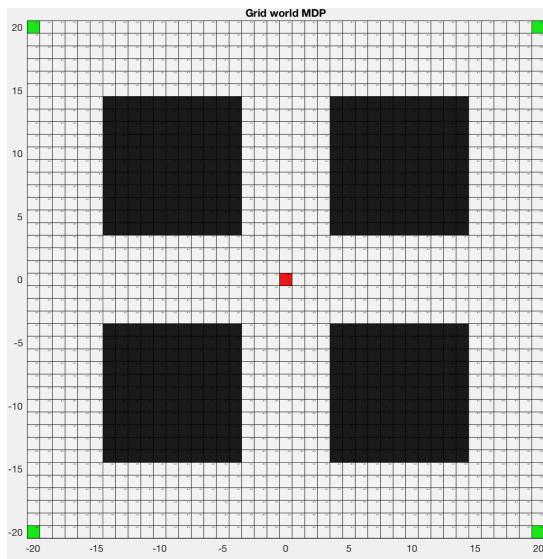


Figure 3.1: The *maze_grid* state space

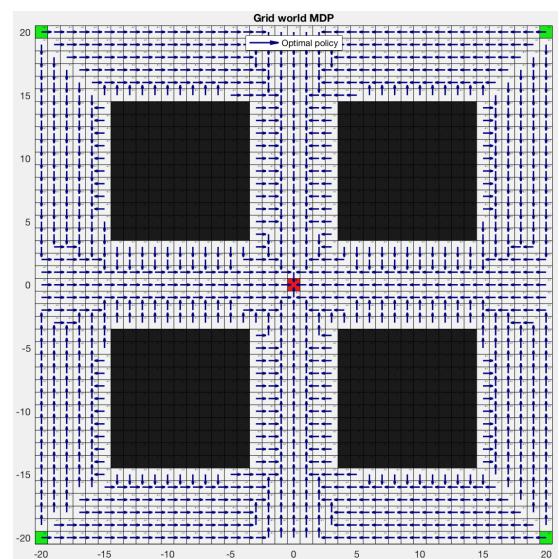


Figure 3.2: Optimal policy, computed with DP

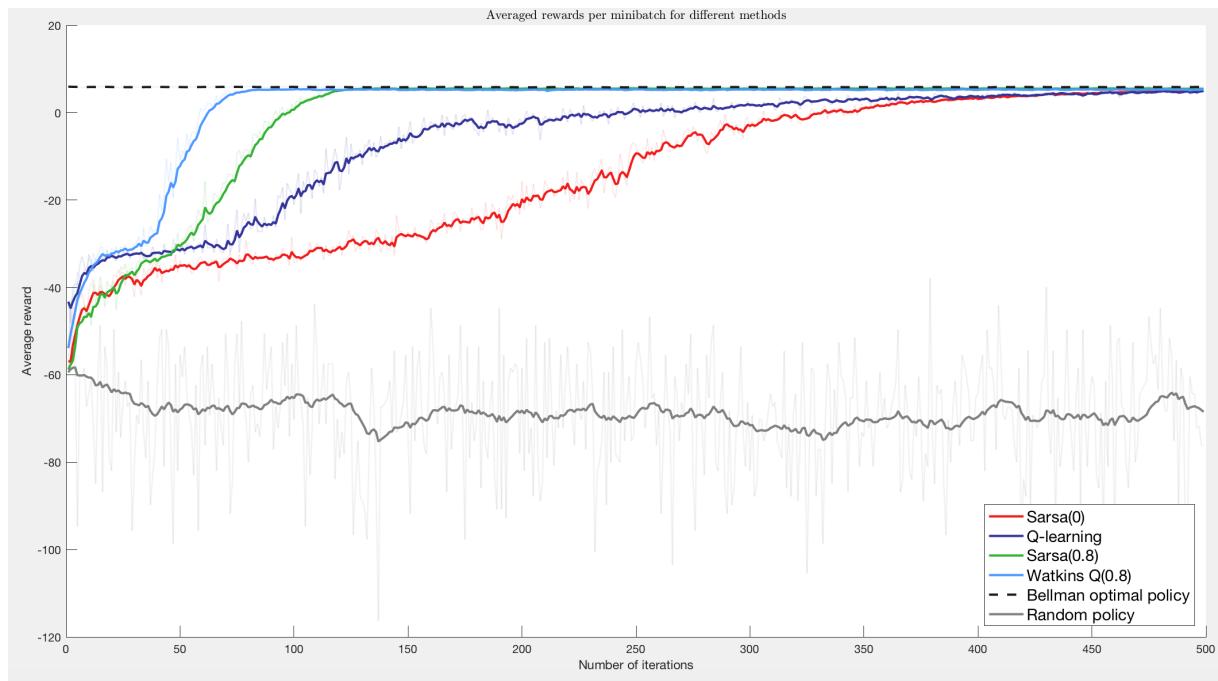


Figure 3.3: Average rewards on minibatch for learned policies, optimal policy and random policy.

Figure (3.3) displays the convergence (expressed as average reward on minibatch) for Qlearning, Sarsa(λ), Sarsa(0) and Watkins Q(λ). By average reward on minibatch, we mean that at every iteration, the reward is average around a given number of trajectories, following the same *exploratory* policy. This allows to reduce the stochasticity of trajectories while learning and give a smoother estimation of how well the algorithm is learning. As a way of comparing them to the optimal and random policies, we also plot the average rewards perceived by the latest along many trajectories. For all the learning algorithms tested, we used optimistic action-value initialisation to promote exploration. This explains why many negative rewards are perceived in the beginning.

As explained before, this environment is considered as a base line in all that follows, and will serve as a sandbox model where we will test our different methods for an imitation learning based approach.

3.2 First naive approaches

3.2.1 Constant compliance

In order to grasp as well as possible some intuition behind a compliance-based learning, we decided to first implement a really simple method, which consist in following the teacher's recommandation with a constant probability p (close from 1), defined globally (the same for all the states). More formally, the *action selection* process is defined as follows. For all $s \in \mathcal{S}$

$$a = \begin{cases} a_m \text{ with probability } p \text{ independent of } s \\ a \in \mathcal{A}(s) \text{ (softmax action picking based on current Qvalues)} \end{cases} \quad (3.2)$$

The learning is done with a succession of action selection based on a five tuple s, a, r, s', a' (the SARSA framework) and the usual TD(0) updates : $\forall s \in \mathcal{S}, a \in \mathcal{A}(s)$

$$Q(s, a) \leftarrow Q(s, a) + \alpha [r + Q(s', a') - Q(s, a)] \quad (3.3)$$

Therefore, the learner *complies* with the teacher with probability p , and decide to choose its own action (that **could include a_m**) with probability $1 - p$. Only the action selection (hence the *exploratory policy*) is impacted by the presence of the mentor.

The intuition behind this approach is rather simple : we direct the exploration of the learner with the recommandations from the teacher, while still enabling it to search around them, based on its current Q-values. The softmax based exploration is regulated by a temperature factor, making it greedy in limit.

The figures (3.4) and (3.5) display the learning obtained (cumulated rewards averaged over mini-batches) when fed with, respectively, the optimal policy for the MDP and a slightly suboptimal one.

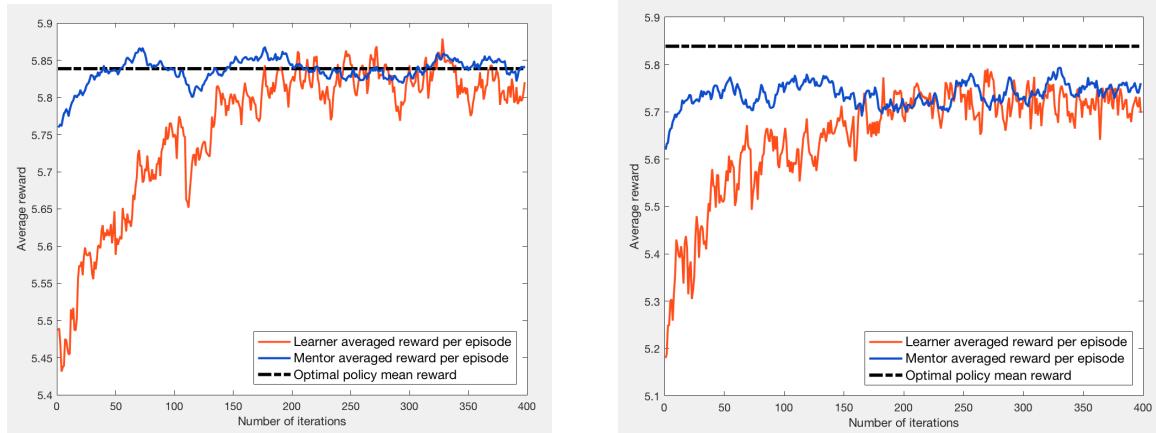


Figure 3.4: Constant compliance learning, $p = 0.9$, with the optimal mentor

Figure 3.5: Constant compliance learning, $p = 0.9$, with a slightly suboptimal mentor

Figure (3.4) shows that from its exploration, the learner is able to quickly learn his way thanks to the optimal mentor. As expected, it eventually follows the mentor's action, wether it complies or not, since the recommended action holds the best action-values. On the other hand, figure (3.5) : because of the high confidence the learner initially have in its mentor, its is never able to reach optimal performance, and the policy its renders actually mimic its mentor suboptimal one. Still, one could expect the learned policy to be slightly better than its suboptimal teacher's one, even if not optimal. However, achieving this comes with a lot of effort in the tuning of p and of the softmax distribution temperature decrease coefficient.

This remark actually pinpoints a major downside of this method, that is the need of fine tuning of the parameter p . But obviously, there is even a bigger downside, that would after become a major game killer and is related to how we generate suboptimal policies from learner. Indeed, mentor policies are taken by generating greedy policies (argmax over the Q-values) of a learner that hasn't completed its learning. Therefore, some mentors can be suboptimal enough to only show a good direction of exploitation, but not be able to reach the target. On a learning from demonstration point of view, such a phenomenon could happen if the human teacher only show the robot part of the task it has to do. Figure (3.6) displays such a policy, where the mentor's policy creates loops and does not always leads to the positive reward. Applying the latest method with such a teacher is fatal for the learning, as the learner discovers that following the mentor's action yields largely negative rewards, but is not able to bypass them as it keeps a constant confidence in its teacher. This leads the learner to build up low Q-values in the directions of the teacher, and trying to follow an opposite path from the mentor. This is highly counter-productive since the mentor still gives the right exploration direction.

3.2.2 Constantly decreasing compliance

One of the downside of the constant compliance approach is that the exploratory behavior is always biased by the mentor's recommandation. This breaks the need of this policy to be *greedy in limit* (which is a specification of the SARSA algorithm). In the case of a sub-oboptimal mentor, this means that the optimal behavior could never be reached.

We now decide to comply with SARSA's exigence to be *greedy in limit*. Therefore, p is *constantly decreased along the learning procedure* :

$$p_{t+1} = \beta p_t \quad (3.4)$$

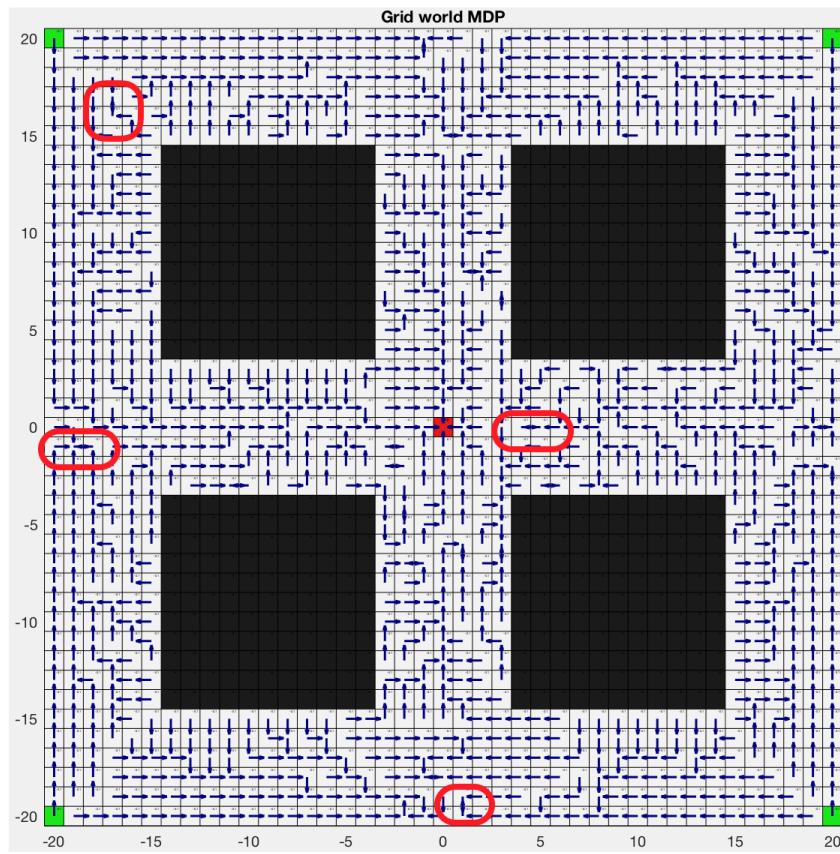


Figure 3.6: An exemple of suboptimal mentor policy that doesn't always lead to the positive reward

with $\beta < 1$. As before, the temperature of the Gibb's softmax sampling also goes to 0 as the learning goes on.

The action-selection is therefore biased by the mentor's recommandation in the beginning of the learning only, and slowly decides to take its own choices, based on its current Q-values estimates. This approach sounds more promising as the learner is more likely to quickly discover the location of high rewards (following the teacher policy with p close to 1), and will then makes it own exploration along those trajectory, to end up in a setting where the teacher's actions are now longer considered.

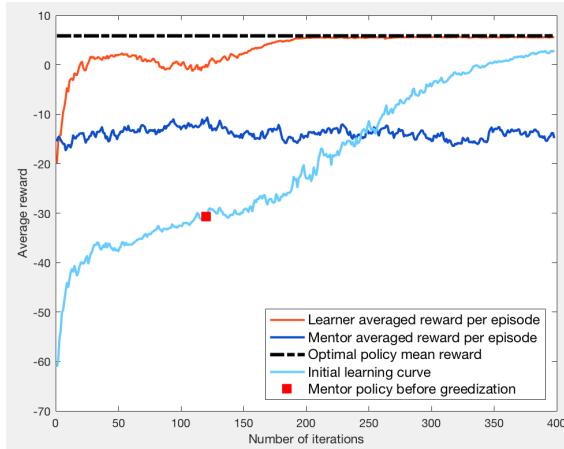


Figure 3.7: Constantly decreasing compliance learning, $\beta = 0.99$

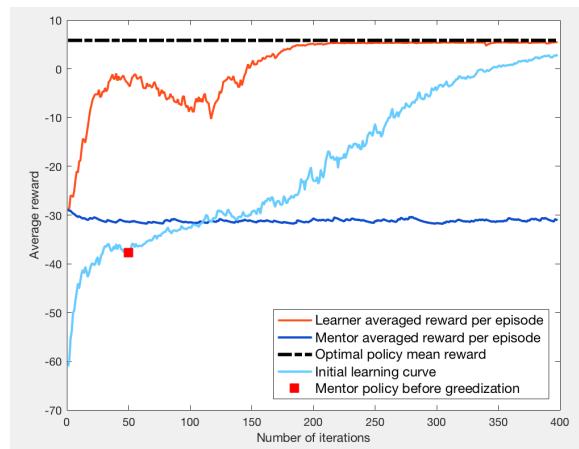


Figure 3.8: Constantly decreasing compliance learning, $\beta = 0.97$

Figures (3.7) and (3.8) display the result on two sub-optimal policies, that were derived from the Q-values learned at a given moment (denoted through a red square) in a learning process. They show that with this approach, while the learner is compliant with the teacher, it steadily increases its accumulated reward thanks to its exploratory actions, where he discard the teacher's recommendation. However, it then comes to a plateau (or even an undershoot) when the guidance by the teacher becomes too weak as p reaches a critical level. The learner then mostly relies on its exploratory actions (which volatility are guided by its learning processus' temperature) to explore the space where the teacher's action never took it before. Once it has somehow learnt the action value function related to this part of its state space, and as its temperature goes down, the learner's accumulated reward goes up until it reaches the level of the optimal policy.

With this technique, the learner is therefore able to learn the optimal policy even if following its teacher recommendation never takes it to the target. It is interesting to see, in such case, what policy the learner actually learns. In figure (3.9), we display the policy learnt from the mentor's policy showed in figure (3.6). One can notice that the learner actually learns to avoid place where the mentor was wrong or created loops, and forks to part of the state space where the teacher was more accurate.

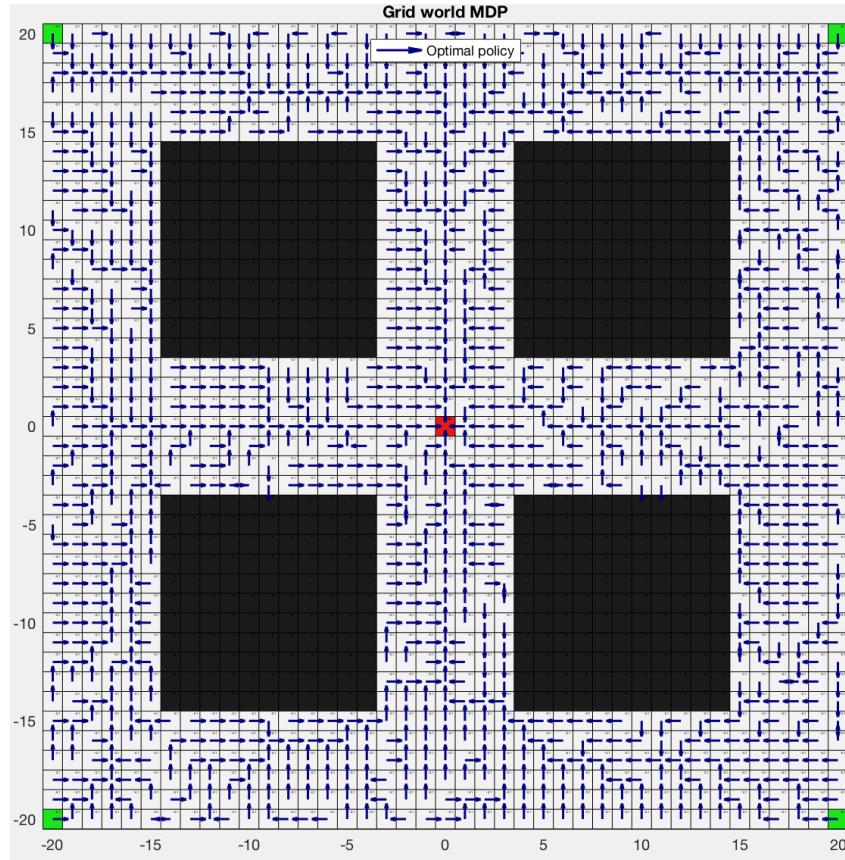


Figure 3.9: Policy learnt from a suboptimal mentor

If this approach seems to work relatively well, there are still some critical downsides to it. First of all, a lot of tuning is required in order to find the hyper-parameters (p_0 , β , temperature evolution, ..) that lead to a fast learning. Also, there is an undershoot in the learning that seems to slow it down, due to the exploring phase that happens when p becomes too low. Such exploration could be avoided or reduced if the learner is able to figure out quickly that indeed, the mentor was right in its recommendation.

3.3 Adaptive compliance approach

Up to this point, it makes sense to somehow *learn* the optimal value of the confidence parameter p , so that we don't have to manually tune its evolution, and so that it can store how right the teacher recommendations are. We hereinafter describe two method which attempt to do so. Their respective experimental results as well as further discussion will be found in the chapter (4).

3.3.1 Actor-critic approach

Let us define the *compliance* term p locally at every state of the state space. For each $s \in \mathcal{S}$, $p(s)$ is given a Beta prior : $p(s) \sim \beta(\alpha(s), \beta(s))$ and represents the trust we have in a mentor's recommandation at state s . The initial values α_0 and β_0 define the initial prior belief we have over p (for instance, $\alpha = 0.5$ and $\beta = 0.055$ define a prior belief that the teacher is most probably right).

As in before, we perform a *p-greedy policy* with respect to the teacher recommandation :

$$\forall s \in \mathcal{S}, \quad \pi_p(s) = \begin{cases} a_m \text{ with probability } p(s) \\ a \in \mathcal{A}(s) \text{ with probability } (1 - p(s)) \end{cases} \quad (3.5)$$

The following describe how we wish to learn the value of $p(s)$, $\forall s \in \mathcal{S}$. Based on the 5-tuple (s, a, r, s', a') obtained thanks to the action selection process we just described, we compute at every step the following temporal difference value :

$$\delta_t = r + \gamma Q(s', a') - Q(s, a_m) \quad (3.6)$$

which compares, *in average*, the advantage (our drawback) of following the state-action pair (s, a) rather than the one indicated by the teacher, according to the current Q-values estimates.

We then apply the following update rule to $p(s)$:

$$\begin{aligned} \alpha_t(s) &\leftarrow \alpha_t(s) + \mathbb{1}_{a=a_m} \delta_t \varepsilon_t \\ \beta_t(s) &\leftarrow \beta_t(s) + \mathbb{1}_{a \neq a_m} \delta_t \varepsilon_t \end{aligned} \quad (3.7)$$

The intuition behind this update rule is simple : if we see that the expected return for the mentor action increases (resp. decreases), then we increase (resp. decrease) α which results in a shift of p toward a larger (resp. smaller) confidence. A similar reasoning holds for the β term. ε_t is the update rule's learning rate, which value and dynamic will be discussed later.

The Q-values for the initial MDP are updated with SARSA algorithm, as it is usually done. Only the action selection is different (for instance, it will be in the usual case derived directly from the learned action function).

3.3.2 Action value approach

We consider here a somewhat similar approach, where our listen vs discard exploration policy is computed according to the current estimated values of the actions '*listen*' and '*discard*'. Let us introduce the action spaces :

$$\forall s \in \mathcal{S}, \mathcal{A}_c(s) = \{'listen', 'discard'\} \quad (3.8)$$

to which we assign the action values $Q_c(\cdot, l)$ and $Q_c(\cdot, d)$ (where l denotes the action of listening and d the action of discarding the teacher recommendation).

The exploration is done by computing a soft policy derived from $\{Q_c(s, l), Q_c(s, d)\}$ for all $s \in \mathcal{S}$. We do this using a Gibbs softmax distribution, which yields :

$$\forall s \in \mathcal{S}, \quad \pi_c(s) = \begin{cases} 'l' \text{ with probability } p = \sigma\left(\frac{Q_c(s, l) - Q_c(s, d)}{\tau}\right) \\ 'd' \text{ with probability } 1 - p \end{cases} \quad (3.9)$$

where $\sigma(\cdot)$ is the logistic sigmoid and τ is a temperate decreasing to 1 (greedy policy in limit).

The update rule for the initial MDP is performed as usual, using TD updates : $\forall(s, a) \in \mathcal{S} \times \mathcal{A}(s)$

$$Q(s, a) \leftarrow (1 - \alpha)Q(s, a) + \alpha [r + \gamma Q(s', a')] \quad (3.10)$$

After each of those update, we also make the following update :

$$\begin{cases} Q_c(s, l) \leftarrow \beta Q_c(s, l) + (1 - \beta)Q(s, a_m) \\ Q_c(s, d) = \beta Q_c(s, d) + (1 - \beta) \max_{a \neq a_m} Q(s, a) \end{cases} \quad (3.11)$$

with $\beta > 0$ the update rule's learning rate.

We can introduce some prior confidence in the teacher by setting, $\forall s \in \mathcal{S}$:

$$Q_c^0(s, l) - Q_c^0(s, d) = \log\left\{\frac{p}{1-p}\right\} \quad (3.12)$$

where p is the retained probability of initially choosing to listen to the teacher.

Chapter 4

Results

4.1 Actor-critic approach

We hereinafter give some practical considerations for the actor-critic approach of compliance-based reinforcement learning, and discuss the experimental results we obtained.

4.1.1 Practical considerations

One of the main specification of the SARSA learning paradigm is that the exploration policy must be *greedy in limit* so that a fixed point can emerge (hopefully the set of Q-values related to the optimal policy). In an actor-critic approach, the usual way to bring the *critic* term in a stationary regime is to modify its learning rate to take the probability of taking an action into account (see [SB98]). This is the approach we chose in order to pick the learning rate ε_t of the update rule (3.7), which now becomes :

$$\begin{aligned}\alpha_t(s) &\leftarrow \alpha_t(s) + \mathbb{1}_{a=a_m} \gamma \delta_t p(s) \\ \beta_t(s) &\leftarrow \beta_t(s) + \mathbb{1}_{a \neq a_m} \gamma \delta_t (1 - p(s))\end{aligned}\tag{4.1}$$

with $\gamma > 0$. This update rule can still be simplified, by approximating p by its mean value, which gives the update rule that we applied in practice :

$$\begin{aligned}\alpha_t(s) &\leftarrow \alpha_t(s) + \mathbb{1}_{a=a_m} \frac{\gamma \alpha_t(s)}{\alpha_t(s) + \beta_t(s)} \delta_t \\ \beta_t(s) &\leftarrow \beta_t(s) + \mathbb{1}_{a \neq a_m} \frac{\gamma \beta_t(s)}{\alpha_t(s) + \beta_t(s)} \delta_t\end{aligned}\tag{4.2}$$

This ensures us that we will direct the update toward a fixed point, and end up with a greedy policy (in limit).

In this method, p is given a Beta prior distribution. To perform the action selection, we would therefore need to sample $p(s)$ from its current distribution, and then sample a Bernoulli random variable of parameter $p(s)$. Assuming that the Beta distribution is sharply peaked around its mean (which we will guarantee by choosing an appropriate prior and learning rate γ). Then $p(s)$ can be approximated by its mean value $\mathbb{E}[p(s)] = \frac{\alpha(s)}{\alpha(s) + \beta(s)}$. We therefore only need to sample a Bernoulli random variable of parameter $\mathbb{E}[p(s)]$ at every state to complete the action selection process.

The tuning that needs to be done is therefore left to γ , the prior and the temperature. In this framework, tuning is made much simpler and one only has to check that the updates are of the same order of magnitude with the prior (in order for the posterior distribution to adapt to observations, but also to retain the memory of the prior).

4.1.2 Results

Figures (4.1) and (4.2) show the learning curves for an agent using the actor-critic compliant-based method, with two different suboptimal teachers. As before, both of them were generated by observing the Q-values of a "normal" learner (without mentor) and by creating the greedy policy with respect to them.

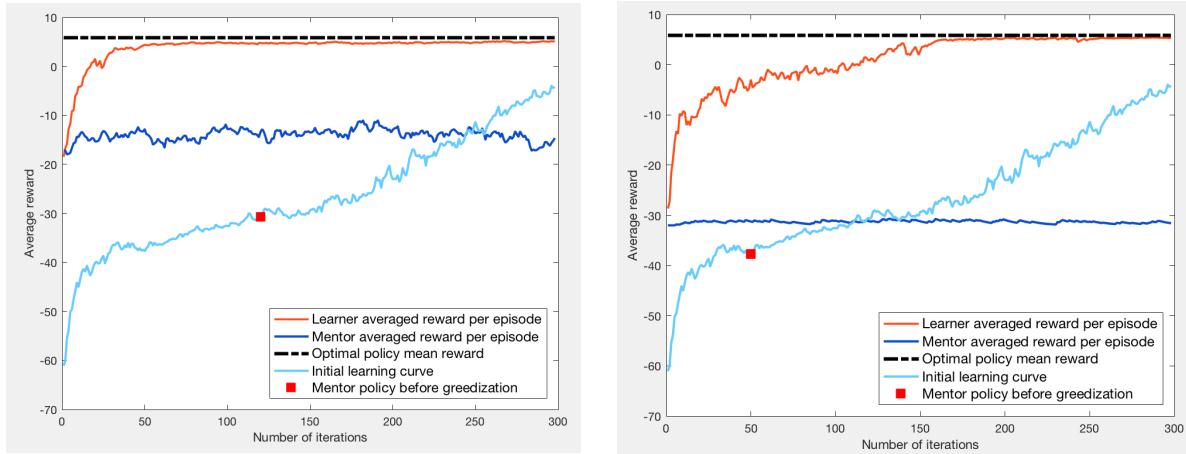


Figure 4.1: Actor-critic learnt compliance learning curve for
 Figure 4.2: Actor-critic learnt compliance learning curve

The first remark we can make is that by learning the confidence, we were able to avoid the undershoot that we observed for the constantly decreasing compliance (see figures (3.7) and (3.8)). This results in a faster learning and more natural learning curve. The different methods will be further compared in (4.3).

4.1.3 Discussion

It would be interesting to see what is the *posterior distribution* of the $p(s)$, $s \in \mathcal{S}$. We expect our final policy to be greedy, and hence the posterior distribution of the compliance term to be sharply peaked around 0 or 1, with proportion given by how good the mentor is.. Figures (4.3) and (4.4) display the histogram of the repartition of the mean of $p(s)$, $s \in \mathcal{S}$ (again, we approximate the Beta distribution by its mean to have an understandable visualization), for two different mentors (figure (4.3) uses the mentor of (4.1) and figure (4.4) the mentor of (4.2)) . As expected, most of the means are either close to 0 and 1, and they are more means close to 0 (poor confidence) when using the second mentor that is far worse than the first.

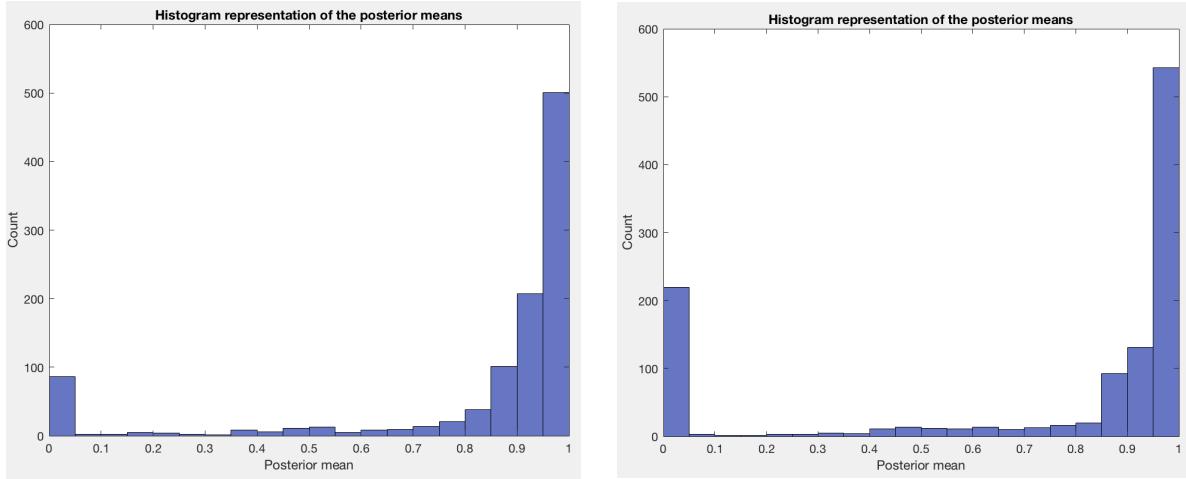


Figure 4.3: Histogram representation of the posterior means for a suboptimal teacher
 Figure 4.4: Histogram representation of the posterior means for a even more suboptimal teacher

Also, to better understand how this algorithm work, being able to visualize the areas where the learner rejects or listen to its mentor. We expect the learner to discard the mentor's recommandation where those are wrong, and to trust its mentor where it is right. Figures (4.5) and (4.6) display, for two different

mentors (the same two ones as we just used), the mentor's policy and the heat-map of the confidence the learner as acquired with respect to its teacher's recommandation at the end of the learning.

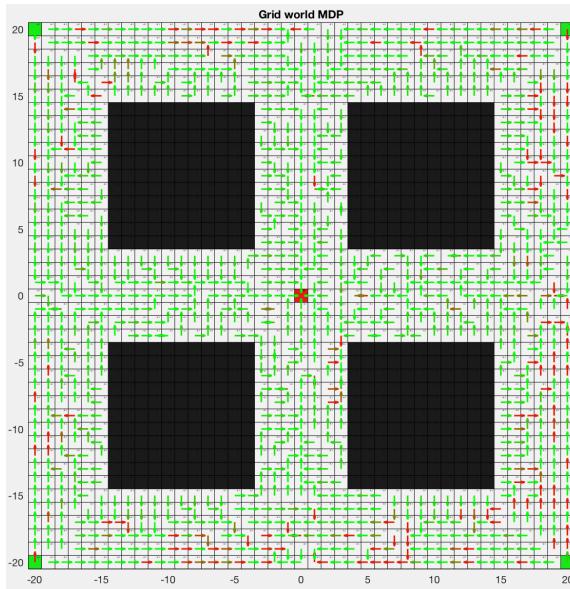


Figure 4.5: Learnt confidence : green arrows show near 1 posterior mean, red arrows near 0

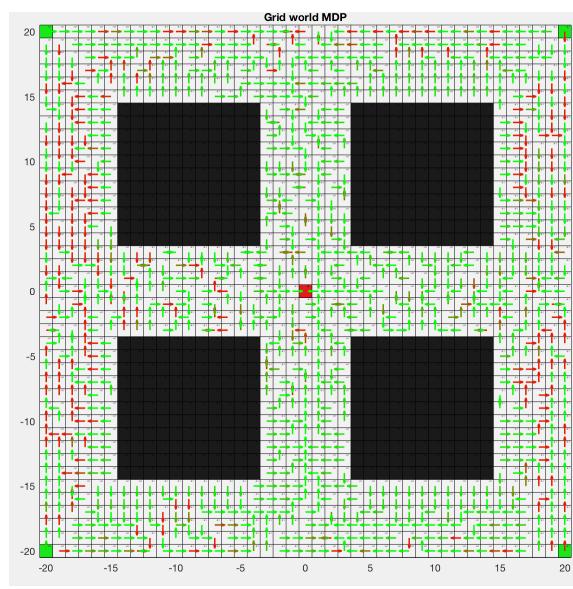


Figure 4.6: Learnt confidence : green arrows show near 1 posterior mean, red arrows near 0

The first observation one can make is that most of the suboptimal mentor actions are indeed well classified by the learner (red arrows). Also, most of the actions leading to such actions are also classified as poor recommandation, because they have a tendency of leading to deadlocks, or suboptimal actions. On the other hand, some optimal actions are also classified as bad recommandations, as due to the greedisation, the learner will somehow prefer another equivalently optimal action.

4.2 Action value approach

4.2.1 Practical considerations

The implementation of this method is rather straight forward and contain no major difficulties. However, some hyper-parameters have to be tuned, like the two temperatures (one for the initial MDP, the other for the above action selection) as well as their dynamics, and the update rule of (3.11).

In practice, this tuning is rather easy to perform, as long as one make sure that the initial value for $Q_c(s, l)$ and $Q_c(s, d)$ ($s \in \mathcal{S}$) don't absorb the observations but also retain some prior knowledge along the learning (the learning rate β has to be related to those initial values in some way).

4.2.2 Results

Figures (4.7) and (4.8) show the learning curves for an agent using the action-value compliant-based method, with two different suboptimal teachers. As before, both of them were generated by observing the Q-values of a "normal" learner (without mentor) and by creating the greedy policy with respect to them.

As for the actor-critic method, we are able to reduce or even suppress the undershoot the constantly decreasing compliant approach displayed, and to obtain fast convergence.

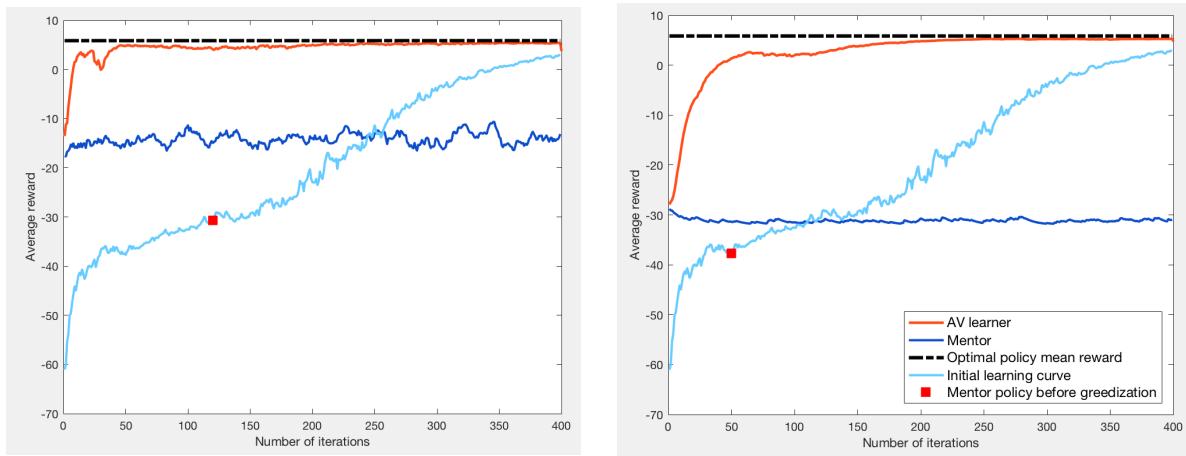


Figure 4.7: Actor-critic learnt compliance learning curve for
Figure 4.8: Actor-critic learnt compliance learning curve

4.2.3 Discussion

Similar plots of posterior result as for the actor-critic can now be drawn. Figures (4.9) and (4.10) display the histograms distributions for the action 'listen' and 'discard' (computed by the sigmoid value at end temperature). Figures (4.11) and (4.12) show the learnt decision over the action-state space.

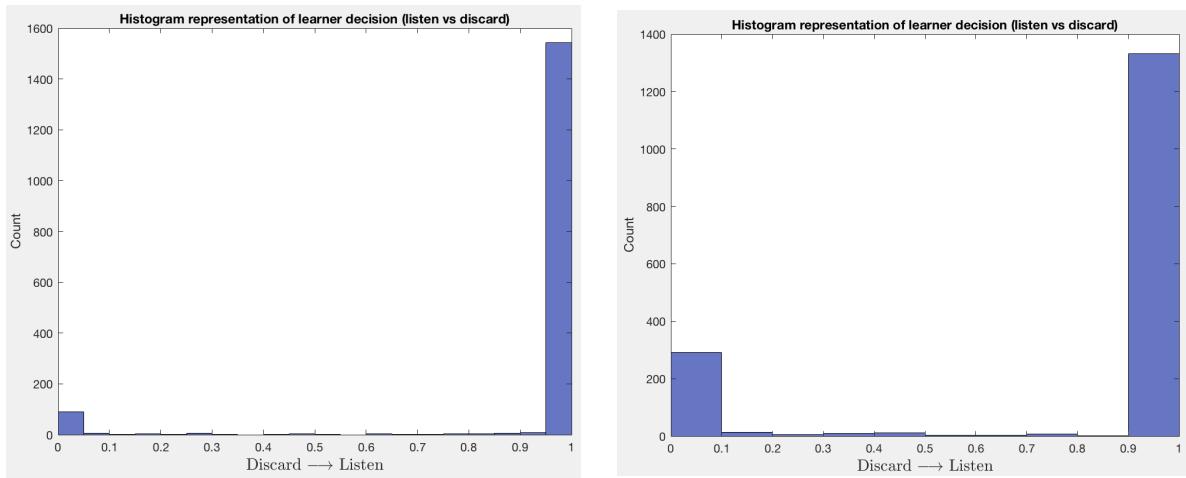


Figure 4.9: Histogram representation of the posterior decisions for a suboptimal teacher
Figure 4.10: Histogram representation of the posterior decisions for a even more suboptimal teacher

Again, one can see that most wrong recommendations are well-classified. However, this method as a tendency to consider as wrong action all trajectories in the mentor's policy leading to a sub-optimality. This as for consequence that the learner will try to always to go round the mentor's suboptimal recommendation, and sometime by making some detours.

4.3 Method comparaison

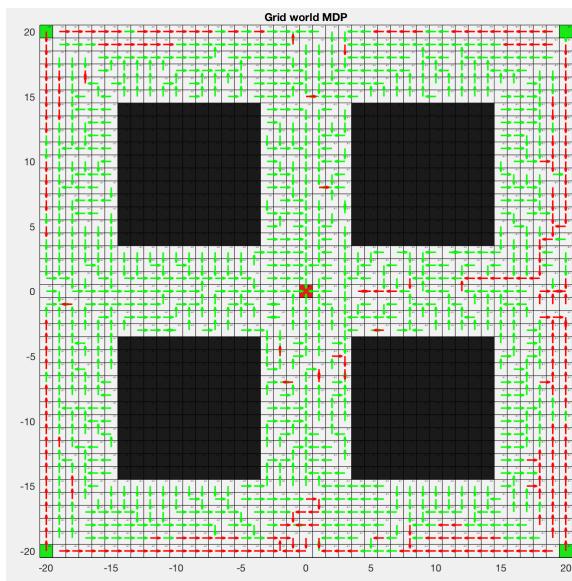


Figure 4.11: Learnt decisions : green arrows show listening, red arrows discarding

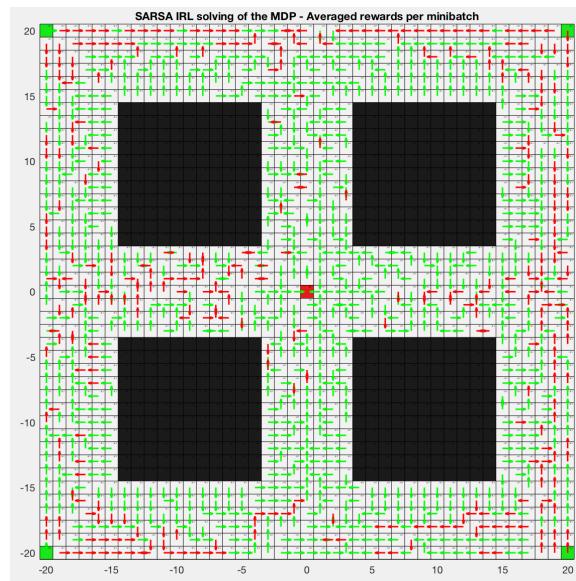


Figure 4.12: Learnt decisions : green arrows show listening, red arrows discarding

Conclusion

Bibliography

- [ACVB09] Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483, 2009.
- [BCD16] Aude G. Billard, Sylvain Calinon, and Rüdiger Dillmann. *Learning from Humans*, pages 1995–2014. Springer International Publishing, Cham, 2016.
- [PB03] Bob Price and Craig Boutilier. Accelerating reinforcement learning through implicit imitation. *Journal of Artificial Intelligence Research*, 19:569–629, 2003.
- [SB98] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning : An Introduction*. MIT Press, 1998.
- [TS09] Matthew E Taylor and Peter Stone. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10(Jul):1633–1685, 2009.