Reinforcement Learning, Master MVA, 2017/2018

Curriculum Learning vs Intrinsically-Motivated Reinforcement Learning

Karl Hajjar & Léonard Hussenot-Desenonges

firstname.name@polytechnique.edu

1 Introduction

This review is the result of our work for the MVA class of Reinforcement Learning. Our main objective is to present and connect different ideas in Reinforcement Learning (RL) around **Curriculum Learning**, **Intrinsic Motivation** and **Surprise-Based** exploration strategies. Our work is based on our reading of articles [1], [2], [3], [4], [5] and our main focus is to discover the **potential links** between curriculum learning strategies and intrinsic motivation. What we will see is that both are a form of enhanced exploration and that an intrinsically motivated RL strategy can automatically induce a form of curriculum in the learning process of the agent.

We will first present the different notions involved via a brief review of each article, and will then try to explore possible correlations between Curriculum Learning and Intrinsic Motivation.

2 Review of the different approaches

2.1 Curriculum Learning

This part is based on our reading of article [1] on Curriculum Learning. The idea of behind Curriculum Learning comes from the fact that the way we, humans, learn along our lives is very systematic as we understood that we could learn much faster by "starting small". By beginning with simple concepts, simple examples before moving to more complex ones, we accelerate the process of learning. That was the idea used in **shaping**, when researchers trained animals to achieve increasingly harder tasks.

This is the concept that motivated **curriculum learning**, a hierarchical organization of learning tasks in machine learning.

There are several ways to interpret the good effect of curriculum learning:

- When trying to learn a deep neural network for a complex task, we can see unsupervised pre-training [Erhan et al.] as a form of curriculum learning. We first train our network to solve a simpler problem than for example, our final classification tasks. First, we teach our network, using it as an autoencoder, to retrieve the original input. This simpler learning task will hopefully lead our parameters to be in a region that will be a much better initialization for our classification task. Indeed, such an initialization will place parameters in a region that will result in a faster optimization and in a convergence to a local minimum that will generally have better generalization properties. Training error will not generally be much better than random initialization, but test error will very probably be. In that respect, we can say that the curriculum strategy acts as a **regularizer**.

- Curriculum Learning can also be seen as a continuation method. Continuation method were introduced to solve non-convex optimization problems. Instead of directly optimizing our non-convex criterion C, we introduce a family of criteria $C_\lambda(\theta)$ such that C_0 is easily solved (because it is convex, for example) and $C_1 = C$. We then gradually increase λ from 0 to 1, keeping θ at the minimum of C_λ. Curriculum learning can be seen as a **continuation method** as we first solve a simpler task, i.e., we solve a

simpler optimization problem (the same loss-function but on simpler data) in the hope that it will be a good initialization for our more complex task (the same loss-function but on more complex data).

Curriculum learning succeeded in simple examples to outperform the classical strategy but we have not find yet a way to unify what we conceive as a "simple example" and more complex ones. It is, for the moment, up to everyone to derive this strategy for each problem. Although a general approach can be given in the supervised setting : learn to do a task on different samples and remove of the dataset the examples were you went wrong, then retrain from scratch on those examples only, use what has been learned to then learn to do the task on the more complex examples removed at first. However, in the unsupervised setting, this kind of procedure cannot be applied as you cannot directly measure the performance of your learned algorithm as you do not have a baseline (labels for examples) to compare the results.

2.2 Intrinsically Motivated Reinforcement Learning

While there are evident common goals, this new strategy , Intrinsically Motivated Reinforcement Learning, introduced in [2], is not a priori directly linked to curriculum learning as it is a reinforcement learning framework and not a supervised learning one but we will explore the link between those two in Section 3.

The objective of this framework is to be able to learn progressively higher level skills, in order to combine those high-level skills already learned when facing a new problem instead of learning to solve everything from scratch.

While the common reinforcement learning framework divides into 2 parts : the environment, distributing rewards, and the agent, receiving rewards and taking actions, we are now going to separate the agent itself into two entities.

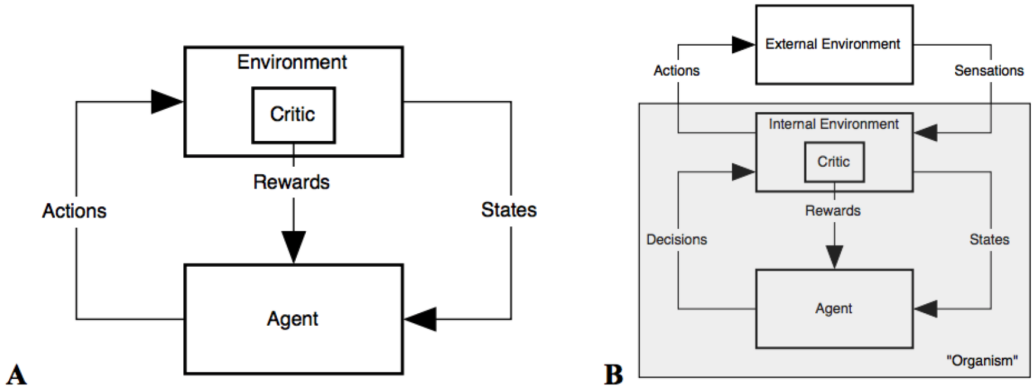

Figure 1: A: Common Environment vs Agent framework. B: Splitting the agent in two

We now consider an organism, that contains the former agent ('the body') and the internal

environment ('the mind'). This factorization came after Sutton and Barto [11] pointed out that one's reward was determined not only by the external state but also by an internal one. One very same external state may lead one to have different rewards as rewards are processed by its brain and not by the environment. The point of departure is thus to note that the internal environment contains, among other things, the organism's motivational system, which needs to be a sophisticated system that should not have to be redesigned for different problems.

The goal will eventually be to develop **skills**. Skills can be seen as options : that is a set of different actions, which when concatenated, lead to a specific result called an option. Options are composed of an option policy that directs the agent's behavior for a subset of the environment states, initiation set, that determines the states where the option might begin and a termination set that determines the states where it may end. Thanks to *option models* and *intra-option Learning Methods*, one can learn those subroutines.

In order to learn these options, Intrinsically motivated RL will give the agent intrinsic rewards. Those intrinsic rewards will model the surprise he gets when encountering a new event.

In the intrinsically motivated RL, the agent is encouraged to reproduce the steps leading to a "salient" event, i.e. events that have granted him "joy" or that he has found to be interesting (surprise related) in themselves and not for their extrinsic reward. In this context, since simpler events tend to happen before more complicated ones which need a chain of several actions to be accomplished (as shown in Section 4 4, playing with the toy monkey happens only if light is off, music on, etc.), the agent first learns to attain these simpler states (light on, music on, ...) before the intrinsic reward associated to these events gets lower as the agent gets "bored" because these events no longer present any novelty (which was the reason why they appeared interesting in the first place). In a way, when a salient event is encountered for the first time, the action value function is updated giving high intrinsic reward for this state-action, thereby encouraging the agent to return to this state to gain reward, but as this state is reached increasingly over time, its intrinsic reward decreases, thus encouraging the agent to explore new states, and possibly discover new, more difficult to attain salient events. If those new events need to be reached passing through salient events already met, the agent will already know how to get to these first salient events, before learning to go from there to obtaining repeatedly a new salient event. The agent will in turn be able to reach easily this harder event and thus to keep reaching this state as long as it receives intrinsic reward for it.

Thus, a complex task that needs many steps to be achieved is learned a lot quicker when the tasks leading to each of those steps are already learned by the agent.

2.3 Surprise Based Reinforcement Learning

This part here is very closely related to the previous part on Intrinsically Motivated Reinforcement learning and is mainly based on the article [3]. Indeed, we have already seen that one way to derive an intrinsic reward is to base it on surprise : high when seeing an event for the first time (regardless of the fact that it is good or bad), and decreasing each time we see this same event again. We will see that this surprised-based rewards leads to great

results as it pushes the agent the learn tasks in a hierarchical order, from to simplest to the hardest.

This form of intrinsic motivation based on surprised has been introduced as a way of enhancing exploration in problems of continuous control with sparse rewards where the classical ϵ-greedy approach fails to learn any reward-granting behavior for reward signals are encountered too infrequently.

Adding intrinsic reward is thus a way to encourage exploration. There are 3 different types of intrinsic motivations :

(i) *empowerment* : level of control of the future (knowing what will happen)
(ii) *surprise* : outcome of an action giving unexpected results (contrary to what has been previously observed)
(iii) *novelty* : discovering new, unexplored states

In the article, the kind of approach which is proposed is to learn the transition probabilities concurrently with the policy, using surprise as a way to encourage further exploration. In this kind of model, the intrinsic motivation is represented by the level of surprise of the agent which is defined as the KL-divergence between the true transition probability distribution and a current learned transition distribution. Unfortunately, the true expected Kl-divergence cannot be computed explicitly since the true transition probability distribution is unknown. Therefore, this Kl-divergence has to be approximated, and it can be done in two ways using a concept of surprise. The learning is done in 2 steps : i) learning the transition probabilities - using Deep-NN to solve regularized Maximum Likelihood on a dataset generated with the current learned policy, ii) maximizing a trade-off finite-horizon cumulative reward between true reward and intrinsic reward.

3 Discovering Correlations

3.1 Enhancing Exploration

The process of enhancing exploration through surprised-based intrinsic reward is quite clear but is not completely clear how we can link this to a form of curriculum, i.e. step based learning process where each step is more difficult to take than the precedent. For instance, in [2], it is shown that intrinsic motivation encourages to explore more in the sense that the novelty of a state or an action is directly linked to the reward. Indeed, an event that has occurred only a very small number of times will be known by the agent to bring a good reward because of the intrinsic reward it produces, and so the agent will be encouraged to re-visit states or actions that have only been seen a little.

This result is summarized in [4] as "explore what surprise you". In fact, the latter article first makes the following statement : "in spite of their pleasant theoretical guarantees, count-based methods have not played a role in the contemporary successes of reinforcement learning (e.g. Mnih et al., 2015). Instead, most practical methods still rely on simple rules such as ϵ-greedy. The issue is that visit counts are not directly useful in large domains, where states are rarely visited more than once." What the paper shows is that "that intrinsic motivation and count-based exploration are but two sides of the same coin", which is to say

that the classical approach of exploration via the counts of action-states already visited can be seen equivalently as a form of intrinsic motivation incentive.

In a similar way, even though Curriculum Learning is not *per se* a Reinforcement Learning framework, the fact that the examples in the dataset, or equivalently the task presented to the learning algorithm, are presented in a gradually more difficult order can be seen as a way to explore the "environment" of the learning algorithm, starting by exploring the easier states, and then going on gradually to the more difficult ones, re-using what he has learned on the first examples.

Thus, both Curriculum Learning and intrinsic motivation can be seen as a form of enhanced exploration.

3.2 Building Curricula through Intrinsic Motivation and Surprised-Based Strategies

Let us dive now in what is really the main point we highlighted concerning the link between Curriculum Learning and Intrinsic Motivation. What we aim to show here is that, in an RL setting, intrinsic motivation strategies (surprised based or not) can induce a form of curriculum on the tasks that are presented to the agent. If we look back at the playroom problem in [2], we see that motivating learning via an intrinsic reward leads to a set of task that appear to the agent, and thus that are learned by him, in a form of curriculum. Indeed, at the beginning when new events come up, they are associated with simpler tasks to solve since they require facing only one event, and are thus the first thing the agent will learn. Building on these small learned bricks, the agent will progressively face events involving more difficult problems to solve from scratch as they themselves appear as a result of a chain of events. Thus, naturally, the intrinsic motivation incentive has built a structured curriculum for the agent, which therefore learns better and faster, as assessed in [2].

We thus believe the crucial element linking curriculum strategies to intrinsic motivation strategies is the fact that an agent which is motivated by intrinsic reward will automatically, and quite naturally learn to achieve tasks that are gradually more difficult. Whereas [4] presents quantitative and formal results to show that intrinsic motivation is strongly linked with enhanced exploration, there has not been (at least not that we know of) until this day a paper building some theory or some quantitative results showing that an intrinsic motivation strategy in RL indeed induces automatically a curriculum of tasks of increasing difficulty or organized in some form of hierarchy. Yet, in that respect, [5] is able to produce a more quantitative approach to show that, in fact, by asymmetric self-play, the intrinsic reward can be precisely built to enforce a curriculum of tasks. This quantifies the fact that intrinsic reward can be carefully chosen to be formulated as an incentive for the agent to take only a few steps beyond its current capabilities regarding his its environment. More detailed about this asymmetric self-play are given in part 4.2.

We therefore see that curriculum learning and intrinsic motivation are in fact closely related and that there is an increasing amount of research being done to keep exploring the relationship between the two and get a better grasp of how they are linked.

4 Examples

4.1 The playground

In his paper, *Intrinsically Motivated Reinforcement Learning* [2], Andrew Bartow brings an interesting example that may help us enlighten the link it has with Curriculum example. Andrew Bartow and his team invented a toy example, a 5x5 gridworld, and objects in it : a hand and a eye, that the agent may move, a ball, some buttons, a bell and a monkey that may interact with the agent. Those interactions are the "salient events" we introduced in first part. Some of them are simple : a button turns on/off the light and another turns on/off the music. On another hand, some of them are more complex : The toy monkey makes frightened sounds if simultaneously the room is dark and the music is on and the bell is rung. It thus requires a long sequence of action to get discovered or reproduced. Here are, in Figure 2, the results of the experience.

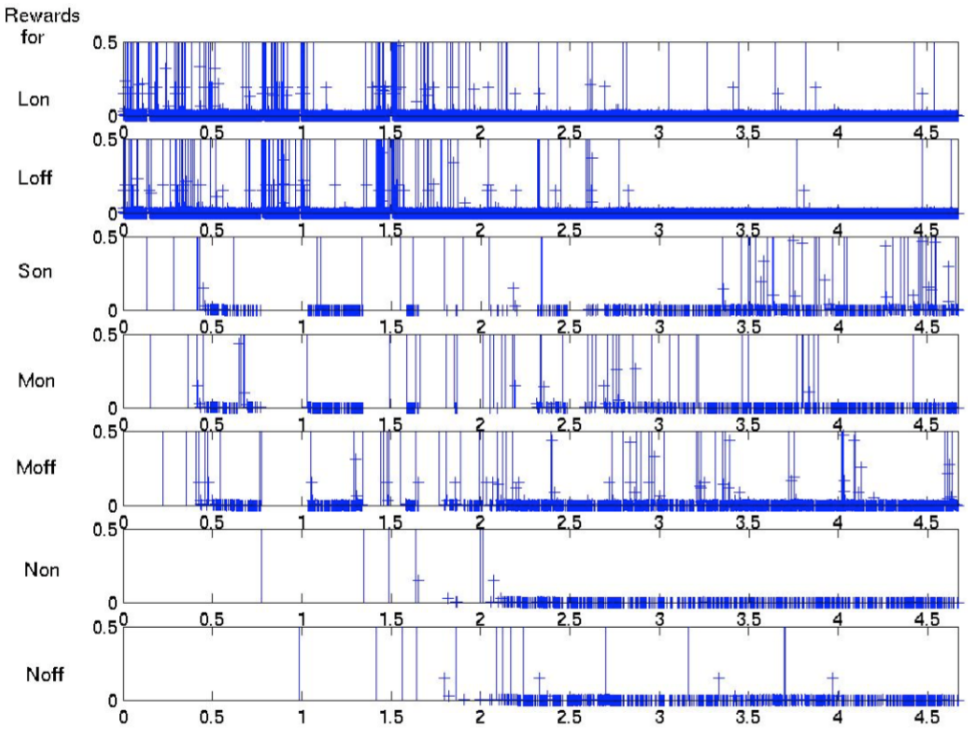

Figure 2: Occurrence of each action ($\times 10^5$); L: light, : bell sound, M: music, N : monkey noise

Let us first note that the salient events that are simpler to achieve occur earlier in time. For example, Lon (light turning on) and Loff (light turning off) are the simplest salient events, and the agent makes these happen quite early. The agent is going to repeat these actions as long as it gets intrinsic reward for it, before getting bored and trying other states. Naturally, these skills of achieving a simple task are learned first. Of course, the events keep

happening despite their diminished capacity to reward because they are needed to achieve the more complex events. Consequently, the agent continues to turn the light on and off even after it has learned this skill because this is a step along the way toward turning on the music, as well as along the way toward turning on the monkey noise. Finally note that the more complex skills are learned relatively quickly once the required sub-skills are in place, as one can see by the few rewards the agent receives for them.

This behavior is definitely equivalent to having a curriculum learning framework where we would first put the agent in a environment where it can just turn on and off the music, then the music and the light, then the music and the light and the bell etc...

The simple difference is that the "training examples" would be artificially/manually ordered in a hierarchical way in the curriculum learning setting, while, when doing Intrinsically Motivated RL, we directly put the agent in an environment where all actions are available and use the intrinsic rewards to force him to show interest to actions in a hierarchical order.

4.2 Alice & Bob

This paradigm was introduced by Facebook researchers in [5]. It is also a gridworld, where the agent needs to learn to do a sequence of action depending on the location of some objects. They introduced a framework that works for any restartable or reversible environment. The goal is to do an **unsupervised pre-training** of the agent, to familiarize it with the environment before asking him to do precise tasks in it.

They introduce Alice and Bob, two versions (or two "minds") of the agent that will play against each other in the gridworld (thus the term "self-play" since the agent is playing against itself). This pre-training is called *unsupervised* as the agent (both Alice and Bob) will not receive any external reward signal from the environment. They will just be self-rewarding themselves, thanks to intrinsic motivation. During this self-play unsupervised pre-training, these two versions will play a game where Alice sets tasks by altering the state via interaction with its objects (key, door, light) and then hands control over to Bob. Bob must return the environment to its original state (in the case of a reversible environment) or get in the same state starting from Alice's initial state (in the case of a restartable one) to receive an internal reward. After this pre-training, Bob's policy will be used to control the agent, with him receiving an external reward if he visits the flag.

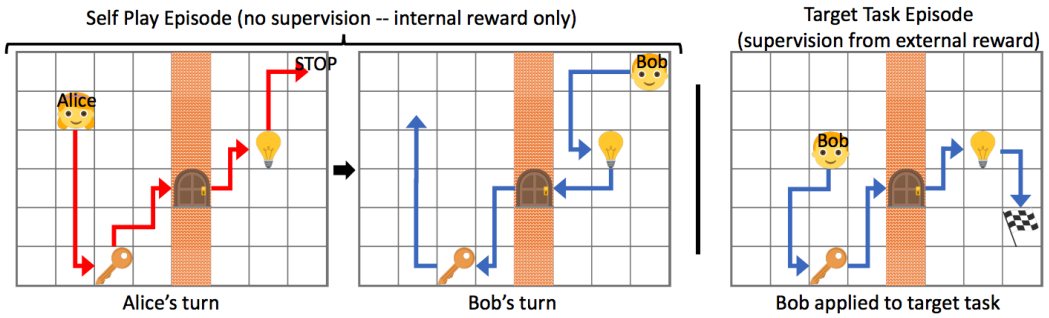

Figure 3: Illustration of the framework in the context of reversible environment

During the pre-training, denoting R_A (resp. R_B) Alice's (resp. Bob's) total reward over one turn of self-play, we have :

$$R_B = -\gamma t_B$$

and

$$R_A = \gamma \max(0, t_B - t_A)$$

with γ a normalization constant to balance intrinsic and external rewards, t_A the time needed by Alice to execute her actions, and t_B the time taken by Bob, trying to reproduce Alice's actions. The total length of an episode is limited to t_{max}, so if Bob fails to complete the task in time we set $R_A = t_{max} - t_A$.

We hereby see that Alice is rewarded if Bob fails to reproduce her actions but that the negative term $-t_A$ will push her to take as few step as possible (and thus a simpler sequence of action) to make Bob fail. Alice will indeed be pushed to **find the simplest task Bob cannot realize** and will thus, thanks to the introduced intrinsic reward, automatically build a curriculum of increasingly challenging tasks for Bob. This will facilitate Bob's learning as he will always have to realize a task slightly above its current abilities. *The self-regulating feedback between Alice and Bob allows them to automatically construct a curriculum for exploration.*

5 Limits and differences

As we showed how correlated are Surprise based Reinforcement Learning and Curriculum Learning, we can now wonder : is there any limit to their resemblance? Let consider the situation, that may appear often in a close future, where you have a personal assistant as an artificial intelligence, that keeps learning when helping you. In such a context, curriculum learning would provide a safe framework for your assistant to learn along its experience interacting with you, learning simple tasks at the beginning and then using this knowledge to help you achieving harder tasks. While surprise-based RL may lead to a similar result (in term of acquired knowledge), the learning process is, in this framework, much more based on a "trial and error" learning. Would you want your assistant to keep doing something you consider as an error as long as this error "surprises" it? Or would you prefer it to progress gradually, tackling problems from the simplest to the hardest?

Were hereby see that, even if Intrinsic Reward Reinforcement Learning provides is a powerful tool to learn complex tasks, Curriculum Learning may, in certain cases, appears to be a much safer framework to learn these tasks.

References

[1] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. *Curriculum Learning*. ICML 2009.

[2] Satinder Singh, Andrew G. Barto, and Nuttapong Chentanez. *Intrinsically Motivated Reinforcement Learning*. NIPS 2004.

[3] Joshua Achiam and Shankar Sastry. *Surprised-Based Intrinsic Motivation For Deep Reinforcement Learning*.

[4] Marc G. Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. *Unifying Count-Based Exploration and Intrinsic Motivation*.

[5] Sainbayar Sukhbaatar, Zeming Lin, Ilya Kostrikov, Gabriel Synnaeve, Arthur Szlam and Rob Fergus. *Intrinsic Motivation and Automatic Curricula via Asymmetric Self-Play*. ICLR 2018.