Relaxed Exploration Constrained Reinforcement Learning

Extended Abstract

Shahaf S. Shperberg
Ben-Gurion University
Be'er Sheva, Israel
shperbsh@bgu.ac.il

Bo Liu
The University of Texas at Austin
Austin, TX, USA
bliu@cs.utexas.edu

Peter Stone
The University of Texas at Austin
Sony AI
Austin, TX, USA
pstone@cs.utexas.edu

ABSTRACT

This extended abstract introduces a novel setting of reinforcement learning with constraints, called Relaxed Exploration Constrained Reinforcement Learning (RECRL). As in standard constrained reinforcement learning (CRL), the aim is to find a policy that maximizes environmental return subject to a set of constraints. However, in RECRL there is an initial training phase in which the constraints are relaxed, thus the agent can explore the environment more freely. When training is done, the agent is deployed in the environment and is required to fully satisfy all constraints. As an initial approach to RECRL problems, we introduce a curriculum-based approach, named CLiC, that can be applied to existing CRL algorithms to improve their exploration during the training phase while allowing them to gradually converge to a policy that satisfies the full set of constraints. Empirical evaluation shows that CLiC produces policies with a higher return during deployment than policies learned when training is done using only the strict set of constraints.

KEYWORDS

Constrained Reinforcement Learning; Curriculum Learning

ACM Reference Format:
Shahaf S. Shperberg, Bo Liu, and Peter Stone. 2023. Relaxed Exploration Constrained Reinforcement Learning: Extended Abstract. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

1 INTRODUCTION

In reinforcement learning, the main objective is to optimize the return received from the environment. However, in many cases, agents are subject to various constraints (e.g., safety, fairness, smoothness of the policy, etc.). To this end, the constrained reinforcement learning (CRL) problem setting was introduced, in which the objective is to optimize return subject to a given set of constraints.

Most work on CRL aims at quickly identifying and focusing exploration on policies that adhere to the constraints, while continually improving the environment return. Some lines of work even further restrict the exploration and aim to completely avoid constraint violations while learning a policy (this is often referred to as "safe exploration", [3]). Adhering to the constraints while learning a policy limits the exploration, which often results in suboptimal policies. This strongly-restricted exploration is important in some settings. Nonetheless, given the rapid development of fast and accurate simulators, it is reasonable to only require these constraints during deployment, while letting the agent explore more freely during training. Alternatively, agents can be initially trained in lab conditions (e.g., with additional supervision), which allow for more lenient constraints than during deployment. To model such situations, we introduce the problem of relaxed exploration constrained reinforcement learning (RECRL), in which constraints can be relaxed during an initial training phase, while the aim is to find a policy that would comply with the constraints while maximizing environmental return during deployment (i.e., after the training phase is over). In the following sections, we briefly describe the RECRL framework, present a curricilum-based approach for solving RECRL problems, and show experimental study on the safe-RL benchmark [4], in which agents operate under safety constraints.

2 RELAXED EXPLORATION CRL

The CRL problem is concerned with finding a policy that maximizes the environment return subject to the given set of constraints. As a result, algorithms designed for solving such problems focus on searching in the space of feasible policies which obey the constraints. To account for scenarios in which constraints can be alleviated when the policy is being trained, we introduce the *Relaxed Exploration Constrained Reinforcement Learning* problem, or RECRL. In this setting, agents face two different phases, a *training phase* and a *deployment phase*. The deployment phase is consistent with standard CRL, i.e agents are allowed to consider only feasible policies that do not validate the given constraints. By contrast, in the training phase, the constraints (cost limits) are relaxed. Formally, a RECRL problem consists of a training budget B and two constrained Markov decision process [2] (CMDPs), $M_t = (\mathcal{S}, \mathcal{A}, T, \gamma, R, C, d_t)$ and $M_d = (\mathcal{S}, \mathcal{A}, T, \gamma, R, C, d_d)$, corresponding to the training and deployment phases, respectively. Both CMDPs are identical with the exception of their cost limits. To capture the desire for relaxed constraints during training, we require that $d_{t_i} \geq d_{d_i}$ for all $1 \leq i \leq k$. Note that in the case where the agent is trained using a simulator, we can set $d_{t_i} = \infty$ for all $1 \leq i \leq k$, effectively reducing M_1 to an MDP. The objective in RECRL is to find a policy for M_d that optimize the return subject to the constraint, $\pi^* = \text{argmax}_{\pi \in \Pi(C, d_d)} J(\pi)$. However, in contrast to CRL, for the first B episodes, the agent operates on M_t and is therefore allowed to explore policies in $\Pi(C, d_t)$.

3 SOLVING RECRL PROBLEMS

While agents that solve RECRL problems can benefit from more lenient constraints during training, they still need to converge to a policy that satisfies the strict constraints during deployment. Since

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

an optimal policy for M_t is not a valid solution for M_d, CRL algorithms cannot be simply executed on M_t to learn a policy for M_d in the RECRL setting. While CRL algorithms can learn a valid policy when directly being applied on M_d, they will fail to utilize the advantages of training with relaxed constraints. In this section, we introduce a method based on a Cost-Limit Curriculum (CLiC) that aims at adapting any CRL algorithm to benefit from the additional exploration allowed for by RECRL problems. Instead of training the agent on M_t throughout the entire training phase, the agent is presented with a curriculum, i.e., a sequence of models, $\mathcal{M} = M_{t_1} \ldots M_{t_n}$, that differ in their cost limits. We restrict the space of curricula to sequences with non-ascending cost limits, in which the final task is the the deployment model M_d. This formulation enables agents to learn a policy using constraints that gradually become tighter, ensuring that the final policy fits the deployment constraints, while supporting better exploration opportunities during the training process. The restriction on the cost limits solves the sequencing problem of determining the order in which the different source tasks are presented to the agent. However, to generate a curriculum the teacher needs to determine: 1) what are the source tasks, i.e. which cost limit to choose for each model presented to the agent, and 2) for how many time steps to train the agent on each task. To this end, we consider two types of curricula.

Static Curricula. *static curricula* (static CLiC). Such curricula are predetermined and do not require any runtime information. Three types of static curricula are studied, based on the most common scheduling strategies: linear decay, cosine decay, and exponential decay. These curricula change the cost limit at every episode based on the training progress with respect to the training budget B, the training costs d_t, and the deployment cost d_d.

By using a static curriculum, students benefit from improved exploration while bounding both the maximum and the cumulative worst-case cost violations experienced during training. Yet, static curricula are not without flaws. First, since the cost limits at each iteration are predetermined, the curriculum cannot reflect the actual costs observed by the student. For example, models in the curriculum can have a cost limit that is much higher than the costs experienced by the student, which can result in "wasted" iterations, in which no progress is made towards converging to a policy that satisfies d_d. Moreover, the return of the student is also not taken into consideration, thus the student can be constantly introduced to new cost limits without ever learning a stable policy. Finally, the student is at risk of converging to a policy that does not satisfy d_d at the end of the training phase.

Dynamic Curriculum. To mitigate the above weaknesses of the static CLiC, we introduce a new teacher, capable of generating a curriculum *dynamically* (Dynamic CLiC) based on the recent history of a student's experience. In Dynamic CLiC , the cost limit in the first model is initialized to be d_t. The teacher provides the student with the same model, while observing the performance of the agent in a moving window of W episodes. Once the agent has experienced at least W episodes on the current cost limit and has converged to a policy both with respect to the environment return and the costs (as determined by two thresholds, ϵ_r and ϵ_c), the cost limits are reduced linearly in each dimension, with respect to the difference between the current limits and d_d, and the number remaining episodes; this part of the dynamic curriculum addresses the first two issues of the static curricula. In addition, if the costs observed by the agent are higher than the cost limit, noise is added to the policy of the student in order to encourage exploration, in an attempt to escape possible local optima.

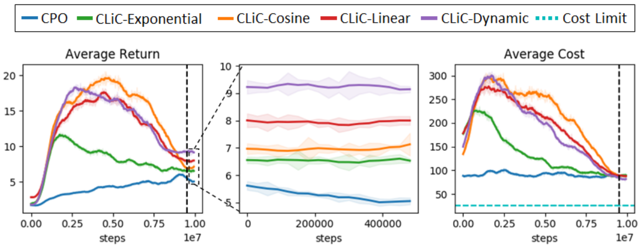

Figure 1: Static and Dynamic CLiC applied to CPO

4 EMPIRICAL EVALUATION

To create RECRL instances and to evaluate the effect of the curriculum-based approaches, we utilize the Safety Gym benchmark [4]. In this benchmark, a robot operates in an environment in the presence of unsafe elements. In the reported experiment, we considered the *car* robot, a wheeled robot with differential drive control, that aims to press a highlighted button. The outcomes considered as unsafe are entering dangerous areas, touching dangerous objects (either movable or immovable, and either stationary or moving), and pressing the wrong button. The cost at every step is a binary function which indicates whether at least one unsafe outcome has occurred. The agent interacts with the environment for $1e7$ steps. The first 95% of the steps are the training phase, in which the agent can be trained using relaxed constraints, while the last 5% of the steps (500k) are the deployment phase, where the agent must adhere to the full set of constraints. We applied the CLiC methods on the Constrained policy optimization (CPO, [1]) algorithm, which enforces constraints throughout training by solving trust region optimization problems at each policy update. The results that shown in Figure 1, in which the solid lines are the mean values (returns or costs), the dashed horizontal line (cyan) indicates the deployment cost-limit (d_d), and the dashed vertical line (black) indicates the transition between the training phase and the deployment phase. In addition, the shaded areas in the plots depict the standard error.

The results show that all CLiC methods improved over the base algorithms, obtaining policies with better returns that induce similar costs, and that the Dynamic CLiC outperformed all static curricula.

5 CONCLUSION

We introduced a setting for constrained reinforcement learning (CRL) that enables agents to train with constraints that are more lenient than during their deployment. To solve such problems, we introduced a curriculum-based (CLiC) approach that can be applied to existing CRL algorithms, with two types of curricula, static and dynamic. The different CLiC methods were shown to significantly boost performance when applied to CPO in an empirical study on the Safe-RL benchmark.

ACKNOWLEDGMENTS

This work has taken place in the Learning Agents Research Group (LARG) at the Artificial Intelligence Laboratory, The University of Texas at Austin. LARG research is supported in part by the National Science Foundation (CPS-1739964, IIS-1724157, FAIN-2019844), the Office of Naval Research (N00014-18-2243), Army Research Office (W911NF-19-2-0333), DARPA, General Motors, Bosch, and Good Systems, a research grand challenge at the University of Texas at Austin. The views and conclusions contained in this document are those of the authors alone. Peter Stone serves as the Executive Director of Sony AI America and receives financial compensation for this work. The terms of this arrangement have been reviewed and approved by the University of Texas at Austin in accordance with its policy on objectivity in research.

REFERENCES

[1] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. 2017. Constrained Policy Optimization. In *ICML (Proceedings of Machine Learning Research, Vol. 70)*. PMLR, 22–31.

[2] Eitan Altman. 1999. *Constrained Markov decision processes*. Vol. 7. CRC Press.

[3] Javier García and Fernando Fernández. 2012. Safe Exploration of State and Action Spaces in Reinforcement Learning. *J. Artif. Intell. Res.* 45 (2012), 515–564.

[4] Alex Ray, Joshua Achiam, and Dario Amodei. 2019. Benchmarking safe exploration in deep reinforcement learning. *arXiv preprint arXiv:1910.01708* 7 (2019), 1.