Stable Reinforcement Learning for Robotic Interaction Tasks*

Shahbaz A. Khader[1,2], Hang Yin[1], Pietro Falco[3], and Danica Kragic[1]

[1] RPL, KTH Royal Institute of Technology, Stockholm, Sweden
{shahak, hyin, dani}@kth.se
[2] ABB Future Labs, Baden, Switzerland (hosted at ABB CRC, Vasteras, Sweden)
[3] ABB CRC, Vasteras, Sweden
pietro.falco@se.abb.com

Abstract. Reinforcement learning (RL) has shown impressive results in robotic manipulation, but in most cases the important property of control stability is neglected. In the context of interaction tasks, or tasks that involve contact with the environment, stability is also related to certain structures of the control policies, for example variable impedance control (VIC). In this work, we aim for RL with VIC policies and stability guarantees. Our contribution is a novel model-free RL method that conforms to the paradigm of Evolution Strategy. Experimental studies on a 7-DOF robotic manipulation task (peg-in-hole) show that not only do we achieve stable exploration but also remarkable sample efficiency.

Keywords: Reinforcement Learning · Evolution Strategy · Control stability · Interaction control

1 Introduction

Interaction tasks, or manipulation tasks that involve contacts with the environment, benefit immensely from RL. Unlike the case of contact-free manipulation, interaction tasks involve highly complex contact dynamics that is often impossible to obtain a priori by analytic means. Instead, RL provides a data-driven solution to learn control policies. Recently, it has been shown that a VIC structure in the policy is advantageous [3]. An additional advancement could be a guarantee of stability, which we address in our work [1].

First, we introduce the term *all-the-time-stability* to mean an absolute guarantee of stability throughout the RL process. Second, we leverage a previously proposed modeling framework for discrete motions [2] as the policy parameterization. Finally, we develop a novel model-free RL method based on the paradigm of Evolution Strategy that can achieve *all-the-time-stability.* Experimental studies on the classical benchmark problem of peg-in-hole demonstrate that not only *all-the-time-stability* is achieved but with very low sample complexity.

* This work was supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

2 i-MOGIC: A VIC Policy with Stability Guarantee

Khansari et al. proposed i-MOGIC as a modeling framework for discrete motions, in the form of a mixture of a base spring-damper element (superscript 0) and several other spring-damper elements (superscript k) [2]. It can be interpreted as a parameterized policy (with VIC structure) [2]:

$$\mathbf{u} = \pi_\theta(s, \dot{s}) \qquad \theta = \{\boldsymbol{S}^0, \boldsymbol{D}^0, \boldsymbol{S}^k, \boldsymbol{D}^k, s^k, l^k\} \text{ for } k = 1, ..., K, \tag{1}$$

where $\mathbf{u}$ is the action, s and $\dot{s}$ are position and velocity of the robot, $\boldsymbol{S}$ and $\boldsymbol{D}$ are stiffness and damping matrices, s^k is attractor point for element k, and l^k is a scalar quantity. The system ensures a unique equilibrium point at the final goal (set to the origin). Using a uniquely constructed Lyapunov function, $V(s, \dot{s})$, the policy was shown to be globally asymptotically stable (GAS) at the origin if:

$$\boldsymbol{S}^0 = (\boldsymbol{S}^0)^T \succ 0,\ \boldsymbol{D}^0 \succ 0,\ \boldsymbol{S}^k = (\boldsymbol{S}^k)^T \succeq 0,\ \boldsymbol{D}^k \succeq 0,\ l^k > 0 \quad \forall k = 1, ..., K \tag{2}$$

If used in RL, rollouts will be bounded in state space and will tend to the goal.

3 Stability-Guaranteed RL

It follows from the previous section that our goal is to maximize the RL objective by optimizing θ in (1) subjected to the constraints in (2). A suitable approach is Evolution Strategy (ES) in which the solution to the problem is represented by a sampling distribution of θ and is iteratively refined using performance weighted samples that was generated (rollouts) from it. A special case of ES is Cross-Entropy Method (CEM) but its easy application is limited to a Gaussian sampling distribution of $\theta \in \mathbb{R}^N$. Since in our case $\theta \notin \mathbb{R}^N$ and has additional structure, we take a novel approach and construct a unique sampling distribution that directly conforms to the structure in (1) and inherently satisfies (2). This leads to two subproblems: how to model the sampling distribution and how to iteratively update it until convergence?

3.1 Stability-Aware Sampling Distribution

We define a sampling distribution of the form:

$$q(\theta|\boldsymbol{\Phi}) = q(\mathbf{s}|\boldsymbol{\Phi}_\mathbf{s})q(\boldsymbol{S}^0|\boldsymbol{\Phi}_{\boldsymbol{S}^0})q(\boldsymbol{D}^0|\boldsymbol{\Phi}_{\boldsymbol{D}^0}) \prod_{k=1}^{K} q(\boldsymbol{S}^k|\boldsymbol{\Phi}_{\boldsymbol{S}^k})q(\boldsymbol{D}^k|\boldsymbol{\Phi}_{\boldsymbol{D}^k})q(l^k|\boldsymbol{\Phi}_{l^k}), \tag{3}$$

where $q(\theta|\boldsymbol{\Phi})$ is the joint probability distribution of all the elements in (1). With $\mathbf{s} = [(s^1)^T, ..., (s^K)^T]^T$, $q(\mathbf{s}|\boldsymbol{\Phi}_\mathbf{s})$ is a multivariate Gaussian distribution with parameters $\boldsymbol{\Phi}_\mathbf{s} = \{\mu_\mathbf{s}, \Sigma_\mathbf{s}\}$; and $q(\boldsymbol{M}|\boldsymbol{\Phi}_{\boldsymbol{M}})$ is a Wishart distributions with parameters $\boldsymbol{\Phi}_{\boldsymbol{M}} = \{W_M, \nu_M\}$ where $\boldsymbol{M} \in \{\boldsymbol{S}^0, \boldsymbol{D}^0, \boldsymbol{S}^k, \boldsymbol{D}^k, l^k\}$ for $k = 1, ..., K$. Wishart distribution is a distribution of symmetric positive definite (SPD) matrices and is fully defined by two parameters ν and W. Note that modeling all elements in $\boldsymbol{M}$ as SPD is more conservative than (2).

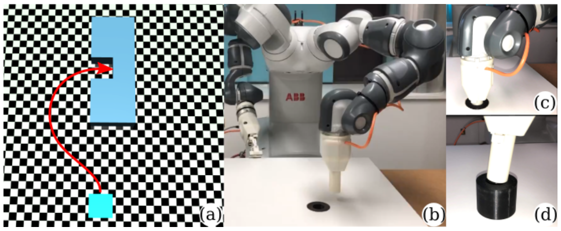

Fig. 1: **Experimental setup (a)** 2D block-insertion environment (MuJoCo simulator): 2 mm insertion clearance, 2 seconds execution time **(b-d)** The real-world manipulator environment: 7-DOF YuMi robot, cylindrical peg and hole, 0.5 mm insertion clearance and and 5 seconds of execution time. **(c)** Successful insertion position. **(d)** 3D printed cylindrical peg and hole.

3.2 Updating the Stability-Aware Sampling Distribution

Since the individual elements in θ are modeled as independent random variables, each of them can be updated independently. The update of the Gaussian $q(\mathbf{s}|\boldsymbol{\Phi}_\mathbf{s})$ is performed by maximum likelihood estimation (MLE) just as in the case of CEM. In contrast, due to the unavailability of analytic MLE for the Wishart distribution, we propose the novel update rule (see [1] for derivation):

$$W^{i+1} = \frac{1}{N_e}\sum_{m=1}^{N_e} \bar{S}_m^i; \qquad \nu^{i+1} = \nu^i \exp\left(\gamma\beta\left(\frac{R_e^i - R_b^i}{R_b^i}\right)\right) \tag{4}$$

where N_e is the number of high performing samples (*elites*); i is the iteration variable; $\bar{S}_m^i$ is the scaled version of SPD samples; γ is a fixed constant; β is the learning rate of the method; and R_e^i and R_b^i are rewards associated to samples. The significance of our contribution is that it is now possible to replace a potentially complex numerical MLE for Wishart distribution with a simple analytic rule. It can also be shown that convergence is guaranteed [1].

4 Experiments

A simulated block of mass is controlled in 2D (no rotation) by applying forces on it with the goal of inserting it into a slot (Fig. 1a). The challenge for the RL algorithm is not only to discover a path but also the right interaction control during contact. In Fig. 2a-b, we see that the RL algorithm converges in under 50 iterations (750 trials). It is guaranteed that $V(s, \dot{s})$ has a unique minimum at the goal and $\dot{V}(s, \dot{s}) < 0$ for $t > 0$ for all iterations. However, RL is allowed to reshape it to a richer function that is optimal for the task (Fig. 2c-d).

A real-world peg-in-hole task is set up for the next experiment (Fig. 1b-d). The peg-in-hole task is a classical benchmark for interaction tasks; the goal is to insert a rigidly held peg into a low clearance hole. The main challenge is control during unexpected contacts. Our method is able to learn the task successfully (Fig. 2e-f) in about 20 iterations (300 trials). Unlike the previous experiment, here we introduced a principled method for initializing the sampling distribution

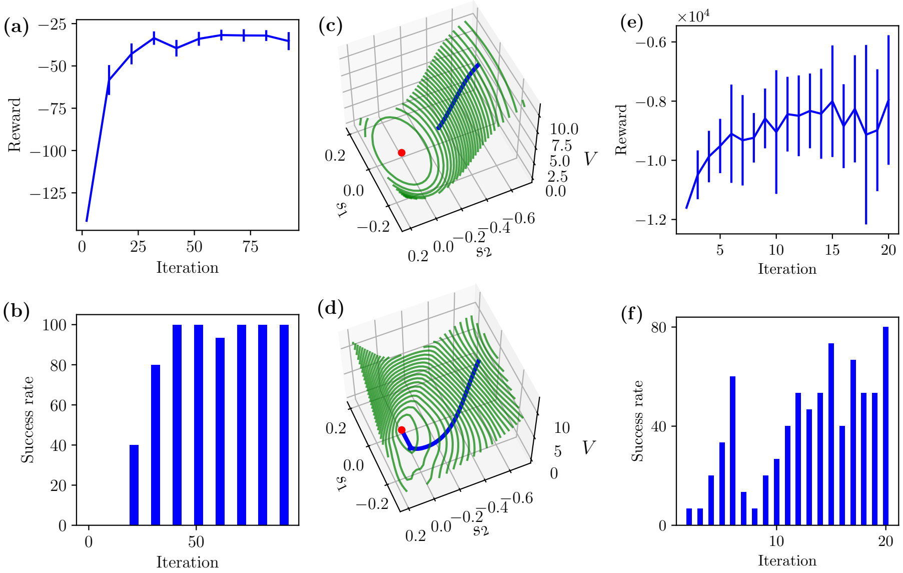

Fig. 2: **Experimental results. (a-b)** shows the RL progress for the 2D block-insertion task. The success rate is calculated as the % success within an iteration. **(c-d)** depicts the transformation of $V(s, \dot{s})_{|\dot{s}=0}$ before (top) and after (bottom) RL. Trajectories are overlaid on the contour plots and goal position is indicated by a red dot. **(e-f)** shows the RL progress for the peg-in-hole task.

[1]. To the best of our knowledge, the RL of peg-in-hole with *all-the-time-stability* is unprecedented.

5 Conclusion

We have proposed a novel Evolution Strategy based RL method (inspired by CEM) and together with a previously proposed specialized policy (i-MOGIC) with stability properties, we succeeded in achieving *all-the-time-stability*. The results are significant because of the unprecedented demonstration of a stable RL of the peg-in-hole task and also the very low sample complexity of the method.

References

1. Khader, S.A., Yin, H., Falco, P., Kragic, D.: Stability-guaranteed reinforcement learning for contact-rich manipulation. arXiv preprint arXiv:2004.10886 (2020)
2. Khansari-Zadeh, S.M., Kronander, K., Billard, A.: Modeling robot discrete movements with state-varying stiffness and damping: A framework for integrated motion generation and impedance control. Proceedings of Robotics: Science and Systems X (RSS 2014) **10**, 2014 (2014)
3. Martín-Martín, R., Lee, M., Gardner, R., Savarese, S., Bohg, J., Garg, A.: Variable impedance control in end-effector space. an action space for reinforcement learning in contact rich tasks. In: Proceedings of the International Conference of Intelligent Robots and Systems (IROS) (2019)