Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

Model-based contextual policy search for data-efficient generalization of robot skills

Andras Kupcsik[a,b,*], Marc Peter Deisenroth[e], Jan Peters[c,d], Loh Ai Poh[a], Prahlad Vadakkepat[a], Gerhard Neumann[c]

[a] National University of Singapore, Department of Electrical and Computer Engineering, 4 Engineering Drive 3, Singapore 118571, Singapore
[b] National University of Singapore, School of Computing, 13 Computing Drive, Singapore 117417, Singapore
[c] Technische Universität Darmstadt, Fachbereich Informatik, Fachgebiet Intelligente Autonome Systeme, Hochschulstr. 10, D-64289 Darmstadt, Germany
[d] Max-Planck Institute for Intelligent Systems, Spemannstrasse 38, 72076 Tübingen, Germany
[e] Imperial College London, Department of Computing, 180 Queen's Gate, London SW7 2AZ, United Kingdom

ARTICLE INFO

Article history:
Received in revised form 29 September 2014
Accepted 24 November 2014
Available online xxxx

Keywords:
Robotics
Reinforcement learning
Contextual policy search
Model-based policy search
Robot skill generalization
Gaussian processes
Movement primitives
Robot table tennis
Robot hockey

ABSTRACT

In robotics, lower-level controllers are typically used to make the robot solve a specific task in a fixed context. For example, the lower-level controller can encode a hitting movement while the context defines the target coordinates to hit. However, in many learning problems the context may change between task executions. To adapt the policy to a new context, we utilize a hierarchical approach by learning an upper-level policy that generalizes the lower-level controllers to new contexts. A common approach to learn such upper-level policies is to use policy search. However, the majority of current contextual policy search approaches are model-free and require a high number of interactions with the robot and its environment. Model-based approaches are known to significantly reduce the amount of robot experiments, however, current model-based techniques cannot be applied straightforwardly to the problem of learning contextual upper-level policies. They rely on specific parametrizations of the policy and the reward function, which are often unrealistic in the contextual policy search formulation. In this paper, we propose a novel model-based contextual policy search algorithm that is able to generalize lower-level controllers, and is data-efficient. Our approach is based on learned probabilistic forward models and information theoretic policy search. Unlike current algorithms, our method does not require any assumption on the parametrization of the policy or the reward function. We show on complex simulated robotic tasks and in a real robot experiment that the proposed learning framework speeds up the learning process by up to two orders of magnitude in comparison to existing methods, while learning high quality policies.

1. Introduction

Learning is a successful alternative to hand-designing robot controllers to solve complex tasks in robotics. Algorithms that learn such controllers need to take several important challenges into consideration. First, robots typically operate in

[*] Corresponding author.
E-mail addresses: kupcsik@comp.nus.edu.sg (A. Kupcsik), m.deisenroth@imperial.ac.uk (M.P. Deisenroth), peters@ias.tu-darmstadt.de (J. Peters), elelohap@nus.edu.sg (A.P. Loh), prahlad@nus.edu.sg (P. Vadakkepat), neumann@ias.tu-darmstadt.de (G. Neumann).

http://dx.doi.org/10.1016/j.artint.2014.11.005

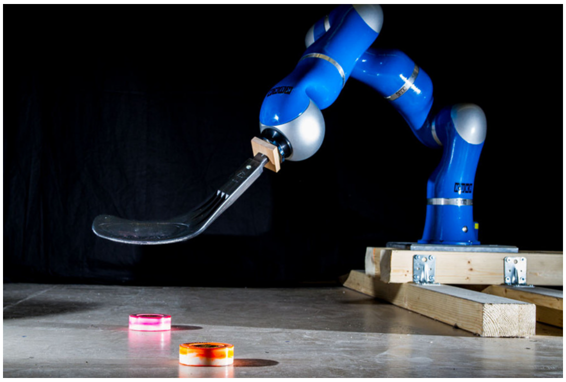

Fig. 1. KUKA lightweight arm shooting hockey pucks.

high-dimensional continuous state-action spaces. Thus, the learning algorithm has to scale well to higher dimensional robot tasks. Second, running experiments with real robots typically has a high cost. An experiment rollout is time consuming, usually requires expert supervision and it might lead to robot damage. Thus, the learning algorithm is required to operate with a limited number of evaluations. Furthermore, when learning with real robots, safety becomes an important factor. To avoid robot and environmental damage, the learning algorithm has to provide robot controllers that generate robot trajectories close to the already explored and, therefore, safe trajectory space. Lastly, robot skills have to be able to adapt to changing environmental conditions. For example, if the task is defined as throwing a ball at varying target positions, the controller has to be adapted to the current target position. In the following, we will refer to such task variables as context $\boldsymbol{s}$. In the throwing example, the context is represented as the target position to throw to. In this paper, we introduce a new model-based policy search method to generalize a learned skill to a new context. For example, if we have learned to throw a ball to a specific location, we want to generalize this skill such that we can throw the ball to multiple locations.

Policy Search (PS) methods are one of the most successful Reinforcement Learning (RL) algorithms for learning complex movement tasks in robotics [31,38,22,23,21,8,30,20,27,13]. PS algorithms typically optimize the parameters $\boldsymbol{\omega}$ of a parametrized control policy, which generates the control commands for the robot, such that the policy obtains maximum reward. A common approach to parametrize the policy is to use a compact representation of a movement with a moderate amount of parameters, such as movement primitives [17,21]. In many approaches to movement primitives, the parameters $\boldsymbol{\omega}$ specify the shape of a desired trajectory. The policy is then defined as trajectory tracking controller that follows this desired trajectory. Such a desired trajectory, represented by a single parameter vector $\boldsymbol{\omega}$, can be used to solve one specific task, characterized by the context vector $\boldsymbol{s}$. The goal in contextual policy search is to learn how to choose the parameter vector $\boldsymbol{\omega}$ of the control policy as a function of the context $\boldsymbol{s}$. To do so, it is convenient to define two different levels of policies that are used in policy search. At the lower level, the control policy specifies the controls of the robot as a function of its state. The lower level policy is parametrized by the parameter vector $\boldsymbol{\omega}$. The lower-level policy can, for example, be implemented as movement primitive [17]. On the upper-level, a policy that chooses the parameters $\boldsymbol{\omega}$ of the lower-level policy is used. We will denote this policy as upper-level policy. Given the current task description $\boldsymbol{s}$, the upper-level policy chooses the parameters $\boldsymbol{\omega}$ of the lower-level policy. The lower level policy is subsequently executed with the given parameters $\boldsymbol{\omega}$ for the whole episode. Although PS algorithms can be applied to learn a large variety of robot skills, in this paper we focus on learning stroke-based movements, such as throwing, hitting, etc.

Most of the existing contextual policy search methods are model-free [20,27], i.e., they try to optimize the policy without estimating a model of the robot and the environment. Model-free PS algorithms execute rollouts on the real robot to evaluate parameter vectors $\boldsymbol{\omega}$. These evaluations are finally used to improve the policy. Most model-free PS algorithms require hundreds if not thousands of real robot interactions until converging to a high quality policy. For many robot learning problems, such data inefficiency is impractical, as executing real robot experiments is time consuming, requires expert supervision and it might lead to robot wear, or even robot damage. It has been shown that the data-efficiency of policy search methods can be considerably improved by learning forward models of the robot and its environment. These models are used to predict the experiment outcome, which allows more robust and efficient policy updates. We refer to such algorithms as model-based policy search algorithms [10,1,4,34,2,18,29]. However, current model-based policy search methods such as PILCO [9,12] suffer from severe limitations that make it hard to apply these methods for learning generalized robot skills. PILCO uses computationally demanding deterministic approximate inference techniques that assume a specific structure of the reward function as well as of the used lower-level policy. These assumptions do not hold for many applications that occur in contextual policy search and have hindered the use of model-based policy search for learning contextual upper level policies. Moreover, the deterministic approximate inference method adds a bias in the prediction of the experiment outcome. Recently, the PILCO algorithm has been extended for learning generalized lower-level controllers [11]. Promising results have been demonstrated for learning robot controllers for hitting and box stacking tasks. Still, PILCO suffers from the restrictions on the structure of the used reward function and lower-level controllers.