Neural Networks 57 (2014) 128–140

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Model-based policy gradients with parameter-based exploration by least-squares conditional density estimation

Voot Tangkaratt [a,*], Syogo Mori [a], Tingting Zhao [a], Jun Morimoto [b], Masashi Sugiyama [a]

[a] Tokyo Institute of Technology, Japan

[b] ATR Computational Neuroscience Labs, Japan

ARTICLE INFO

Article history:
Received 11 July 2013
Received in revised form 17 April 2014
Accepted 11 June 2014
Available online 21 June 2014

Keywords:
Reinforcement learning
Transition model estimation
Conditional density estimation

ABSTRACT

The goal of reinforcement learning (RL) is to let an agent learn an optimal control policy in an unknown environment so that future expected rewards are maximized. The model-free RL approach directly learns the policy based on data samples. Although using many samples tends to improve the accuracy of policy learning, collecting a large number of samples is often expensive in practice. On the other hand, the model-based RL approach first estimates the transition model of the environment and then learns the policy based on the estimated transition model. Thus, if the transition model is accurately learned from a small amount of data, the model-based approach is a promising alternative to the model-free approach. In this paper, we propose a novel model-based RL method by combining a recently proposed model-free policy search method called *policy gradients with parameter-based exploration* and the state-of-the-art transition model estimator called *least-squares conditional density estimation*. Through experiments, we demonstrate the practical usefulness of the proposed method.

1. Introduction

Reinforcement learning (RL) is a framework to let an agent learn an optimal control policy in an unknown environment so that expected future rewards are maximized (Kaelbling, Littman, & Moore, 1996). The RL methods developed so far can be categorized into two types: *policy iteration* where policies are learned based on value function approximation (Lagoudakis & Parr, 2003; Sutton & Barto, 1998) and *policy search* where policies are learned directly to maximize expected future rewards (Dayan & Hinton, 1997; Kakade, 2002; Sehnke et al., 2010; Sutton, McAllester, Singh, & Mansour, 2000; Williams, 1992; Zhao, Hachiya, Tangkaratt, Morimoto, & Sugiyama, 2013).

1.1. Policy iteration vs. policy search

A value function represents expected future rewards as a function of a state or a state–action pair. In the policy iteration framework, approximation of the value function for the current policy and improvement of the policy based on the learned value function are iteratively performed until an optimal policy is found. Thus, accurately approximating the value function is a challenge in the value function based approach. So far, various machine learning techniques have been employed for better value function approximation, such as least-squares approximation (Lagoudakis & Parr, 2003), manifold learning (Sugiyama, Hachiya, Towell, & Vijayakumar, 2008), efficient sample reuse (Hachiya, Akiyama, Sugiyama, & Peters, 2009), active learning (Akiyama, Hachiya, & Sugiyama, 2010), and robust learning (Sugiyama, Hachiya, Kashima, & Morimura, 2010).

However, because policy functions are learned indirectly via value functions in policy iteration, improving the quality of value function approximation does not necessarily yield a better policy function. Furthermore, because a small change in value functions can cause a big change in policy functions, it is not safe to use the value function based approach for controlling expensive dynamic systems such as a humanoid robot. Another weakness of the value function approach is that it is difficult to handle continuous actions because a maximizer of the value function with respect to an action needs to be found for policy improvement.

On the other hand, in the policy search approach, policy functions are determined so that expected future rewards are directly maximized. A popular policy search method is to update policy functions via gradient ascent. However, a classic policy gradient

* Corresponding author. Tel.: +81 8044090125.
E-mail addresses: voot@sg.cs.titech.ac.jp (V. Tangkaratt), mori@sg.cs.titech.ac.jp (S. Mori), tingting@sg.cs.titech.ac.jp (T. Zhao), xmorimo@atr.jp (J. Morimoto), sugi@cs.titech.ac.jp (M. Sugiyama).

http://dx.doi.org/10.1016/j.neunet.2014.06.006
0893-6080/

method called REINFORCE (Williams, 1992) tends to produce gradient estimates with large variance, which results in unreliable policy improvement (Peters & Schaal, 2006). More theoretically, it was shown that the variance of policy gradients can be proportional to the length of an agent's trajectory, due to the stochasticity of policies (Zhao, Hachiya, Niu, & Sugiyama, 2012). This can be a critical limitation in RL problems with long trajectories.

To cope with this problem, a novel policy gradient method called *policy gradients with parameter-based exploration* (PGPE) was proposed (Sehnke et al., 2010). In PGPE, deterministic policies are used to suppress irrelevant randomness and useful stochasticity is introduced by drawing policy parameters from a prior distribution. Then, instead of policy parameters, hyper-parameters included in the prior distribution are learned from data. Thanks to this prior-based formulation, the variance of gradient estimates in PGPE is independent of the length of an agent's trajectory (Zhao et al., 2012). However, PGPE still suffers from an instability problem in small sample cases. To further improve the practical performance of PGPE, an efficient sample reuse method called *importance-weighted PGPE* (IW-PGPE) was proposed recently and demonstrated to achieve the state-of-the-art performance (Zhao et al., 2013).

1.2. Model-based vs. model-free

The RL methods reviewed above are categorized into the *model-free* approach, where policies are learned without explicitly modeling the unknown environment (i.e., the transition probability of the agent in the environment). On the other hand, an alternative approach called the *model-based* approach explicitly models the environment in advance and uses the learned environment model for policy learning (Deisenroth & Rasmussen, 2011; Wang & Dietterich, 2003). In the model-based approach, no additional sampling cost is necessary to generate artificial samples from the learned environment model.

Model-based methods are the predominant approach for fast and data-efficient learning. For example, given a fixed budget for data collection, IW-PGPE requires us to determine the *sampling schedule* in advance. More specifically, we need to decide, e.g., whether many samples are gathered in the beginning or only a small batch of samples are collected for a longer period. However, optimizing the sampling schedule in advance is not possible without strong prior knowledge. Thus, we need to just blindly design the sampling schedule in practice, which can cause significant performance degradation. On the other hand, the model-based approach does not suffer from this problem because we can draw as many trajectory samples as we want from the learned transition model without additional sampling costs.

Another advantage of the model-based approach lies in *baseline subtraction*. In the gradient-based policy search methods such as REINFORCE and PGPE, subtraction of a baseline from a gradient estimate is a vital technique to reduce the estimation variance of policy gradients (Peters & Schaal, 2006; Zhao et al., 2013). If the baseline is estimated from samples that are statistically independent of samples used for the estimation of policy gradients, variance reduction can be carried out without increasing the estimation bias. However, such independent samples are not available in practice (if available, they should be used for policy gradient estimation), and thus variance reduction by baseline subtraction is practically performed at the expense of bias increase. On the other hand, in the model-based scenario, we can draw as many trajectory samples as we want from the learned transition model without additional sampling costs. Therefore, two statistically independent sets of samples can be generated and they can be separately used for policy gradient estimation and baseline estimation.

1.3. Transition model learning by least-squares conditional density estimation

If the unknown environment is accurately approximated, the model-based approach can fully enjoy all the above advantages. However, accurately estimating the transition model from a limited amount of trajectory data in multi-dimensional continuous state and action spaces is highly challenging. Although the model-based method that does not require an accurate transition model was developed (Abbeel, Quigley, & Ng, 2006), it is only applicable to deterministic environments, which significantly limits its range of applications in practice. On the other hand, a recently proposed model-based policy search method called PILCO (Deisenroth & Rasmussen, 2011) learns a probabilistic transition model by the Gaussian process (GP) (Rasmussen & Williams, 2006), and explicitly incorporates long-term model uncertainty. However, PILCO requires states and actions to follow Gaussian distributions and the reward function to be a particular exponential form to ensure that the policy evaluation is performed in a closed form and policy gradients are computed analytically for policy improvement. These strong requirements make PILCO practically restrictive.

To overcome such limitations of existing approaches, we propose a highly practical policy-search algorithm by extending the model-free PGPE method to the model-based scenario. In the proposed model-based PGPE (M-PGPE) method, the transition model is learned by the state-of-the-art non-parametric conditional density estimator called *least-squares conditional density estimation* (LSCDE) (Sugiyama, Takeuchi et al., 2010), which can handle multi-modal distributions directly. LSCDE has various superior properties:

- It can directly handle multi-dimensional inputs and outputs.
- It achieves the optimal convergence rate (Kanamori, Suzuki, & Sugiyama, 2012).
- It has high numerical stability (Kanamori, Suzuki, & Sugiyama, 2013).
- It is robust against outliers (Sugiyama, Suzuki, & Kanamori, 2012).
- Its solution can be analytically and efficiently computed just by solving a system of linear equations (Kanamori, Hido, & Sugiyama, 2009).
- Generating samples from the learned conditional density is straightforward.

Through experiments, we demonstrate that the proposed M-PGPE method is a promising approach.

The rest of this paper is structured as follows. In Section 2, we formulate the RL problem and review model-free RL methods including PGPE. We then propose the model-based PGPE method in Section 3, and experimentally demonstrate its usefulness in Section 4. Finally, we conclude in Section 5.

2. Problem formulation and model-free policy search

In this section, we first formulate our RL problem and review existing model-free policy search methods.

2.1. Formulation

Let us consider a Markov decision problem consisting of the following elements:

- $\mathcal{S}$: A set of continuous states.
- $\mathcal{A}$: A set of continuous actions.
- $p(\boldsymbol{s})$: The (unknown) probability density of initial states.
- $p(\boldsymbol{s}'|\boldsymbol{s}, a)$: The (unknown) conditional probability density of visiting state $\boldsymbol{s}'$ from state $\boldsymbol{s}$ by action a.
- $R(\boldsymbol{s}, a, \boldsymbol{s}')$: The immediate reward function for the transition from $\boldsymbol{s}$ to $\boldsymbol{s}'$ by a.