Robotics and Autonomous Systems 106 (2018) 107–116

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

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

Mirror descent search and its acceleration☆

Megumi Miyashita [a], Shiro Yano [b,*], Toshiyuki Kondo [b]

[a] Department of Computer and Information Sciences, Graduate School of Engineering, Tokyo University of Agriculture and Technology, Tokyo, Japan
[b] Division of Advanced Information Technology and Computer Science, Institute of Engineering, Tokyo University of Agriculture and Technology, Tokyo, Japan

HIGHLIGHTS

- We proposed algorithms both for black-box-opt and reinforcement learning problems
- We showed advanced methods in optimization theories can be applied to RL algorithms
- We revealed the relation between existing reinforcement learning methods

ARTICLE INFO

Article history:
Received 22 October 2017
Received in revised form 18 April 2018
Accepted 28 April 2018
Available online 5 May 2018

Keywords:
Reinforcement learning
Mirror descent
Bregman divergence
Accelerated mirror descent
Policy improvement with path integrals

ABSTRACT

In recent years, attention has been focused on the relationship between black-box optimization problem and reinforcement learning problem. In this research, we propose the Mirror Descent Search (MDS) algorithm which is applicable both for black box optimization problems and reinforcement learning problems. Our method is based on the mirror descent method, which is a general optimization algorithm. The contribution of this research is roughly twofold. We propose two essential algorithms, called MDS and Accelerated Mirror Descent Search (AMDS), and two more approximate algorithms: Gaussian Mirror Descent Search (G-MDS) and Gaussian Accelerated Mirror Descent Search (G-AMDS). This research shows that the advanced methods developed in the context of the mirror descent research can be applied to reinforcement learning problem. We also clarify the relationship between an existing reinforcement learning algorithm and our method. With two evaluation experiments, we show our proposed algorithms converge faster than some state-of-the-art methods.

1. Introduction

Similarity between black-box optimization problem and reinforcement learning (RL) problem inspires recent researchers to develop novel RL algorithms [1–3]. The objective of a black box optimization problem is to find the optimal input $x^* \in \mathcal{X}$ of an unknown function $f : \mathcal{X} \to \mathbb{R}$. Because the objective function f is unknown, we usually solve the black box optimization problem without gradient information $\nabla_x f$. Such is the case with RL problem. The objective of an RL problem is to find the optimal policy that maximizes the expected cumulative reward [4]. As is the case in a black-box optimization problem, the agent does not know the problem formulation initially, so he is required to tackle the lack of information. In this research, we propose RL algorithms from a standpoint of a black-box optimization problem.

RL algorithm has been categorized into a value-based method and a policy-based method, roughly. In the value-based method, the agent learns the value function of some action in some state. On the other hand, in the policy-based method, the agent learns policy from the observation directly. Moreover, RL algorithm has been divided into a model-free approach and a model-based approach. In the model-based approach, first, the agent gains the model of a system from the sample. Then, it learns policy or the value using the model. In contrast, in the model-free approach, the agent learns the policy or value without the model. RL algorithms usually employ the assumption that the behavior of environment is well approximated by Markov Decision Process (MDP).

Recently, KL divergence regularization plays a key role in policy search algorithms. KL divergence is one of the essential metrics between two distributions. Past methods [5–10] employ KL divergence regularization to find a suitable distance between a new distribution and a referential distribution. It is important to note that there exists two types of KL divergence: KL and reverse-KL (RKL) divergence [11,12]. The past researches mentioned above are clearly divided into the algorithms with KL divergence [5,7] and

☆ The research was partially supported by JSPS KAKENHI (Grant numbers JP26120005, JP16H03219, and JP17K12737).

* Corresponding author.
E-mail address: syano@cc.tuat.ac.jp (S. Yano).

https://doi.org/10.1016/j.robot.2018.04.009

RKL divergence [6,8–10]. We review details of these algorithms afterward.

Bregman divergence is the general metric which includes both of KL and RKL divergence [13] (see Appendix A). Moreover, it includes Euclidean distance, Mahalanobis distance, Hellinger distance and so on. Mirror Descent (MD) algorithm employs the Bregman divergence to regularize the learning steps of decision variables; it includes a variety of gradient methods [14]. Accelerated mirror descent [15] is one of the recent advance applicable for the MD algorithms universally.

In this study, we propose four reinforcement learning algorithms on the basis of MD method. Proposed algorithms can be applied in the non-MDP setting. We propose two essential algorithms and two approximate algorithms of them. We propose mirror descent search (MDS) and accelerated mirror descent search (AMDS) as the essential algorithms, and Gaussian mirror descent search (G-MDS) and Gaussian accelerated mirror descent search (G-AMDS) as the approximate algorithms. G-AMDS showed significant improvement in convergence speed and optimality in two benchmark problems. If other existing reinforcement learning algorithms can be reformulated as the MDS form, they would also get the benefit from the acceleration. We also clarify the relationship between existing reinforcement learning algorithms and our method. As an example, we show the relationship between MDS and Policy Improvement with Path Integrals (PI^2) [16,17] in Section 5.

2. Related works

This section will proceed in the order described below. First of all, we introduce the concept of KL and RKL divergences. Then we refer the two types of RL algorithms: RL with KL divergence [5,7] and RL with RKL divergence [6,8–10]. We also refer the RL algorithm PI^2; we show the relation between PI^2 and our method afterward. We conclude this section with a comment on other MD-based RL algorithms.

The KL divergence between $\boldsymbol{x}$ and $\boldsymbol{x}'$ is represented as follows.

$$\mathrm{KL}\left(\boldsymbol{x}, \boldsymbol{x}'\right) = \sum_{j=1}^{m} x_j \log \frac{x_j}{x'_j} \left(\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^m, x_j, x'_j > 0\right). \tag{1}$$

We call $\mathrm{KL}\left(\boldsymbol{x}', \boldsymbol{x}\right)$ Kullback Leibler divergence under the condition that we determine $\boldsymbol{x}$ by reference to the fixed $\boldsymbol{x}'$; we call $\mathrm{KL}\left(\boldsymbol{x}, \boldsymbol{x}'\right)$ reverse-KL divergence [12]. Bregman divergence includes both of KL and RKL divergence [13], so we expect it provides an unified formulation of above-mentioned algorithms.

Let us introduce the RKL-based RL algorithms. Relative Entropy Policy Search (REPS) [6] is one of the pioneering algorithms focusing on the information loss during the policy search process. The information loss is defined as the relative entropy, also known as the RKL divergence, between the old policy and the new policy. The new policy is determined under the upper bound constraints of the RKL divergence. Episode-based REPS also considers information loss bound with regard to the upper-level policy [10]. The method is proposed as an extension of REPS to be an episode-based algorithm. The paper [9] discussed the similarity between Episode-based REPS and the proximal point algorithm; they proposed the Online-REPS algorithm as an theoretically guaranteed one. MOdel-based Relative Entropy stochastic search (MORE) also employed RKL divergence [8], which extends the episode-based REPS to be a model-based RL algorithm. These algorithms employ RKL divergence in their formulation.

There are some methods employing KL divergence. Trust Region Policy Optimization (TRPO) [5], which is one of the suitable algorithms to solve deep reinforcement learning problem, updates the policy parameters under the KL divergence bound. The research [7] showed that KL divergence between policies plays a key role to derive the well-known heuristic algorithm: Co-variance Matrix Adaptation Evolutionary Strategy (CMA-ES) [18]. Authors named the method Trust-Region Co-variance Matrix Adaptation Evolution Strategy (TR-CMA-ES). TR-CMA-ES is similar to episode-based REPS but uses the KL divergence. Proximal Policy optimization (PPO) algorithm also introduces KL divergence in their penalized objective [19].

PI^2 [17,20] would be one of the worth mentioning RL algorithm. PI^2 encouraged researchers [21,2] to focus on the relationship between RL algorithms and black box optimization. For example, [21] proposes a reinforcement learning algorithm PI^{BB} on the basis of black box optimization algorithm: CMA-ES. The authors [22,20] discussed the connection between PI^2 and KL control. We further discuss PI^2 from a viewpoint of our proposed methods at Section 5.

Previous studies also proposed reinforcement learning algorithms on the basis of MD method [23,24]. Mirror Descent TD(λ) (MDTD) [23] is a value based RL algorithm. The paper [23] employs Minkowski distance with Euclidean space rather than KL divergence. By contrast, we basically employ the Bregman divergences on the simplex space, i.e. non-Euclidean space. Mirror Descent Guided Policy Search (MDGPS) [24] is also associated with our proposed method. They showed mirror descent formulation improved the Guided Policy Search (GPS) [25]. MDGPS has a distinctive feature that it depends both on KL divergence and RKL divergence. However, as is shown in [26], there are the variety of Bregman divergences on simplex space other than KL divergence and RKL divergence. Moreover, it plays an important role in accelerating the mirror descent [15]. So we explicitly use Bregman divergence in this research.

3. Mirror descent search and its variants

3.1. Problem statement

In this section, we mainly explain our algorithm as a method for the black box optimization problem. Consider the problem of minimizing the original objective function J defined on subspace $\Omega \subseteq \mathbb{R}^l$, i.e. $J : \Omega \to \mathbb{R}$. We represent the decision variable by $\boldsymbol{\omega} \in \Omega$. Rather than dealing with decision variable $\boldsymbol{\omega} \in \Omega$ directly, we consider the continuous probability density function of $\boldsymbol{\omega}$. Let us introduce the probability space. The probability space is defined as $(\Omega, \mathcal{F}, P)$, where $\mathcal{F}$ is the σ-field of Ω and P is a probability measure over $\mathcal{F}$.

In this paper, we introduce the continuous probability density function $p(\boldsymbol{\omega})$ as the alternative decision variable defined on the probability space. We also define the alternative objective function by the expectation of the original objective function $J(\boldsymbol{\omega})$:

$$\mathcal{J} = \int_{\Omega} J(\boldsymbol{\omega})\, p(\boldsymbol{\omega})\, d\boldsymbol{\omega} \tag{2}$$

Therefore, we search the following domain:

$$p(\boldsymbol{\omega}) \geq 0 \tag{3}$$

$$\int_{\Omega} p(\boldsymbol{\omega})\, d\boldsymbol{\omega} = 1 \tag{4}$$

Let us introduce the set $\mathcal{P}_{\mathrm{all}}$ consists of all probability density functions defined on the probability space. The optimal generative probability is

$$p^*(\boldsymbol{\omega}) = \underset{p(\boldsymbol{\omega}) \in \mathcal{P}_{\mathrm{all}}}{\arg\min} \left\{ \int_{\Omega} J(\boldsymbol{\omega})\, p(\boldsymbol{\omega})\, d\boldsymbol{\omega} \right\} = \underset{p(\boldsymbol{\omega}) \in \mathcal{P}_{\mathrm{all}}}{\arg\min}\, \mathcal{J}. \tag{5}$$

From the viewpoint of the black box optimization problems, the algorithm aims at obtaining the optimal decision variable $p^*(\boldsymbol{\omega})$ to optimize the alternative objective function $\mathcal{J}$. From the viewpoint