Neural Networks 52 (2014) 43–61

Contents lists available at ScienceDirect

Neural Networks

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

Policy oscillation is overshooting☆

Paul Wagner

Department of Information and Computer Science, Aalto University, FI-00076 Aalto, Finland

ARTICLE INFO

Article history:
Received 4 January 2013
Received in revised form 16 October 2013
Accepted 3 January 2014

Keywords:
Reinforcement learning
Approximate dynamic programming
Policy gradient
Natural gradient
Policy oscillation
Policy chattering

ABSTRACT

A majority of approximate dynamic programming approaches to the reinforcement learning problem can be categorized into greedy value function methods and value-based policy gradient methods. The former approach, although fast, is well known to be susceptible to the policy oscillation phenomenon. We take a fresh view to this phenomenon by casting, within the context of non-optimistic policy iteration, a considerable subset of the former approach as a limiting special case of the latter. We explain the phenomenon in terms of this view and illustrate the underlying mechanism with artificial examples. We also use it to derive the constrained natural actor-critic algorithm that can interpolate between the aforementioned approaches. In addition, it has been suggested in the literature that the oscillation phenomenon might be subtly connected to the grossly suboptimal performance in the Tetris benchmark problem of all attempted approximate dynamic programming methods. Based on empirical findings, we offer a hypothesis that might explain the inferior performance levels and the associated policy degradation phenomenon, and which would partially support the suggested connection. Finally, we report scores in the Tetris problem that improve on existing dynamic programming based results by an order of magnitude.

1. Introduction

We consider the reinforcement learning problem in which one attempts to find a good policy for controlling a stochastic non-linear dynamical system. Many approaches to the problem are value-based and build on the methodology of simulation-based approximate dynamic programming (Bertsekas, 2005; Bhatnagar, Sutton, Ghavamzadeh, & Lee, 2009; Buşoniu, Babuška, De Schutter, & Ernst, 2010; Peters & Schaal, 2008; Sutton & Barto, 1998; Szepesvári, 2010). In this setting, there is no fixed set of data to learn from, but instead the target system, or typically a simulation of it, is actively sampled during the learning process so as to obtain the information needed for policy improvement. The sampling policy is often chosen to be the current policy itself or some slightly perturbed variation of it. This learning setting is often described as interactive learning (e.g., Szepesvári, 2010, Section 3).

The majority of these methods can be categorized into greedy value function methods (critic-only) and value-based policy gradient methods (actor-critic) (e.g., Konda & Tsitsiklis, 2004; Szepesvári, 2010). The former approach, although fast, is susceptible to potentially severe policy oscillations in the presence of approximations. This phenomenon is known as the policy oscillation (or policy chattering) phenomenon (Bertsekas, 2011; Bertsekas & Tsitsiklis, 1996). The latter approach has better convergence guarantees, with the strongest case being for Monte Carlo evaluation with 'compatible' value function approximation. In this case, convergence w.p.1 to a local optimum can be established under mild assumptions (Konda & Tsitsiklis, 2004; Peters & Schaal, 2008; Sutton, Mcallester, Singh, & Mansour, 2000).

Bertsekas has recently called attention to the currently not well understood policy oscillation phenomenon (Bertsekas, 2011). He suggests that a better understanding of it is needed and that such understanding "has the potential to alter in fundamental ways our thinking about approximate DP". He also notes that little progress has been made on this topic in the past decade (see also Sutton, 1999). In this paper, we will try to shed more light on this topic. The motivation is twofold. First, the policy oscillation phenomenon is intimately connected to some aspects of the learning dynamics at the very heart of approximate dynamic programming; the lack of understanding in the former implies a lack of understanding in the latter. In the long run, this state might well be holding back important theoretical developments in the field. Second, methods not susceptible to oscillations have a much better suboptimality bound (Bertsekas, 2011), which gives also immediate value to a better understanding of oscillation-predisposing conditions.

The involved problematic aspects of the learning dynamics arise from the interactive nature of the setting, in which the available set of information about the target system changes during learning in a way that depends on the learning process itself. The sampling distribution is conditional on the current intermediate solution (the current policy), which in turn is based

☆ A shorter version of the paper appeared in NIPS 2011 (Wagner, 2011).
E-mail address: paul.wagner@aalto.fi.

0893-6080/$ – see front matter
http://dx.doi.org/10.1016/j.neunet.2014.01.002

on the currently available set of information about the target system. This set of information, again, depends on the sampling distribution. The resulting learning timescale feedback loop is a source of considerable complications for establishing convergence in the presence of approximations (e.g., in the presence of value function approximation or imperfect state estimation). Solving a control problem under such approximations using sampled data will give different solutions for different sampling distributions, just as fitting a low-capacity function approximator to high information content data will give different results for different subsets of the data. The target system will appear as changing during the learning process, suggesting different solutions on different moments due to the current policy being changed. In a sense, the question changes whenever the answer is updated. This can become a problem with respect to stability in the case of incautious exploitation (instantaneous over-use) of momentarily perceived opportunities. It is possible that there are such pairs (or groups) of policies that lead to such samples and (an implicit) model that always suggest the other policy as the optimal one.

The policy oscillation phenomenon is strongly associated in the literature with the popular Tetris benchmark problem. This problem has been used in numerous studies to evaluate different learning algorithms (see Szita & Lörincz, 2006; Thiery & Scherrer, 2009a). Several studies, including those by Bertsekas and Ioffe (1996), Desai, Farias, and Moallemi (2009), Farias and Van Roy (2006), Kakade (2002), Petrik and Scherrer (2008), Szita and Lörincz (2006), Thiery and Scherrer (2009b), have been conducted using a standard set of features that were originally proposed by Bertsekas and Ioffe (1996). This setting has posed considerable difficulties to some approximate dynamic programming methods. Impressively fast initial improvement followed by severe degradation was reported by Bertsekas and Ioffe (1996) using a greedy approximate policy iteration method. This degradation has been taken in the literature as a manifestation of the policy oscillation phenomenon (Bertsekas & Ioffe, 1996; Bertsekas & Tsitsiklis, 1996).

Policy gradient and greedy approximate value iteration methods have shown much more stable behavior in the Tetris problem (Kakade, 2002; Petrik & Scherrer, 2008), although it has seemed that this stability tends to come at the price of speed (see esp. Kakade, 2002). Still, the performance levels reached by even these methods fall way short of what is known to be possible. The typical performance levels obtained with approximate dynamic programming methods have been around 5000 points (Bertsekas & Ioffe, 1996; Bertsekas & Tsitsiklis, 1996; Farias & Van Roy, 2006; Kakade, 2002), while an improvement to around 20,000 points has been obtained by Petrik and Scherrer (2008) by considerably *lowering* the discount factor. On the other hand, performance levels between 300,000 and 900,000 points were obtained recently with the very same features using the cross-entropy method (Szita & Lörincz, 2006; Thiery & Scherrer, 2009b). It has been hypothesized by Bertsekas (2011) that this grossly suboptimal performance of even the best-performing approximate dynamic programming methods might also have some subtle connection to the oscillation phenomenon. In this paper, we investigate also these potential connections.

The structure of the paper is as follows. After providing a background in Section 2, we discuss the policy oscillation phenomenon in Section 3 along with three examples, one of which is novel and generalizes the others. We develop a novel view to the policy oscillation phenomenon in Sections 4 and 5. We validate the view also empirically in Section 6, after which we proceed to look for the suggested connection between the oscillation phenomenon and the convergence issues in the Tetris problem. In Section 6.2, we report empirical evidence that indeed suggests a shared explanation to the policy degradation observed by Bertsekas and Ioffe (1996), Bertsekas and Tsitsiklis (1996) and the early stagnation of all the rest of the attempted approximate dynamic programming methods.

2. Background

A Markov decision process (MDP) is defined by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r)$, where $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces. $S_t \in \mathcal{S}$ and $A_t \in \mathcal{A}$ denote random variables on time t, and $s, s' \in \mathcal{S}$ and $a, b \in \mathcal{A}$ denote state and action instances. $\mathcal{P}(s, a, s') = \mathbb{P}(S_{t+1} = s'|S_t = s, A_t = a)$ defines the transition dynamics and $r(s, a) \in \mathbb{R}$ defines the expected immediate reward function. A (soft-)greedy policy $\pi^*(a|s, Q)$ is a (stochastic) mapping from states to actions and is based on the value function Q. A parameterized policy $\pi(a|s, \theta)$ is a stochastic mapping from states to actions and is based on the parameter vector θ. Note that we use π^* to denote a (soft-)greedy policy, not an optimal policy. The action value functions $Q(s, a)$ and $A(s, a)$ are estimators of the γ-discounted cumulative reward $\sum_t \gamma^t \mathbb{E}[r(S_t, A_t)|S_0 = s, A_0 = a, \pi]$ that follows some (s, a) under some π. The state value function $V(s)$ is an estimator of such cumulative reward that follows some s.

In policy iteration, the current policy is fully evaluated, after which a policy improvement step is taken based on this evaluation. In optimistic policy iteration, policy improvement is based on an incomplete evaluation. In value iteration, just a one-step lookahead improvement is made at a time.

In greedy value function reinforcement learning (e.g., Bertsekas, 2005; Buşoniu et al., 2010), the current policy on iteration k is usually implicit and is greedy (and thus deterministic) with respect to the value function Q_{k-1} of the previous policy:

$$\pi^*(a|s, Q_{k-1}) = \begin{cases} 1 & \text{if } a = \arg\max_b Q_{k-1}(s, b) \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Improvement is obtained by estimating a new value function Q_k for this policy, after which the process repeats. Soft-greedy iteration is obtained by slightly softening π^* in some way so that $\pi^*(a|s, Q_{k-1}) > 0, \ \forall a, s$, the Gibbs soft-greedy policy class with a temperature τ (Boltzmann exploration) being a common choice:

$$\pi^*(a|s, Q_{k-1}) \propto e^{Q_{k-1}(s,a)/\tau}. \tag{2}$$

We note that (1) becomes approximated by (2) arbitrarily closely as $\tau \to 0$ and that this corresponds to scaling the action values toward infinity.

A common choice for approximating Q is to obtain a least-squares fit using a linear-in-parameters approximator $\tilde{Q}$ with the feature basis ϕ^*:

$$\tilde{Q}(s, a, w_k) = w_k^\top \phi^*(s, a) \approx Q_k(s, a). \tag{3}$$

For the soft-greedy case, one option is to use an approximator that will obtain an approximation of an advantage function. The use of an advantage value function, in which the action values are centered around some per-state reference value, was analyzed in-depth first in Baird (1993). For a related analysis of optimal baselines, see Peters (2007, Section 4.3.2) and references therein. We use the following definition from Sutton et al. (2000):

$$\tilde{A}(s, a, w_k) = w_k^\top \left(\phi^*(s, a) - \sum_b \pi^* \left(b|s, \tilde{A}(w_{k-1}) \right) \phi^*(s, b) \right) \approx A_k(s, a). \tag{4}$$

Convergence properties depend on how the estimation is performed and on the function approximator class with which Q is being approximated. For greedy approximate policy iteration in the general case, policy convergence is guaranteed only up to bounded sustained oscillation (Bertsekas, 2005). Optimistic variants can permit asymptotic convergence in parameters, although the corresponding policy can manifest sustained oscillation even then (Bertsekas, 2005, 2011; Bertsekas & Tsitsiklis, 1996).