ADAPTIVE DYNAMIC PROGRAMMING

Gerhard Hübner
University of Hamburg, Germany.

Keywords: adaptive, dynamic programming, decision process, average reward, discounted, estimation and control, nonstationary value iteration, policy iteration, applications

Contents

Summary

Adaptive dynamic programming is the problem of finding an optimal (or nearly optimal) control policy for a (discrete time) valuated stochastic process whose local rewards and transitions depend on unknown parameters. This problem is formalized by a family of Markov decision processes using the valuations of long run average and of asymptotic discounted total value. (For simplicity we restrict to finite or countable state spaces.) To get information on the unknown parameters, these are estimated with increasing precision, while the process evolves in time. At each stage, the presently best estimates are used to improve controlling actions.

Several solution methods are presented, ranging from completely exploiting the recent estimates and needing a great many calculations , e.g. the "principle of estimation and control", to those with a slow adaptation to changing estimates and using few calculations, e.g. "non-stationary policy/value iterations".

Finally, some frequently used estimation procedures are presented. In addition, we

discuss some typical applications and related concepts.

1. Introduction

The essential features of adaptive dynamic programming may be demonstrated by the following example from inventory management:

The owner of a little shop selling running shoes states, at the end of each week, the numbers of sold shoes and remaining shoes. Then he orders new ones according to his knowledge and feeling about future demands. The size of his order may be calculated by using models and methods of *dynamic programming*.

During the following weeks his experience and information about future demands and prices may change. Consequently, his decisions regarding the amounts to be ordered may change too. This type of behavior, i.e. including latest information into decisions, is called *adaptive*.

The situation of the shop owner described above - without "adaptive" aspects - is an example from the class of *Markov Decision Processes* which are also called *Stochastic Dynamic Programs* when stressing algorithmic aspects.

The additional feature above, called *adaptive*, is the fact that (some of) the parameters influencing the present reward and/or the future development of the process are not (exactly) known to the decision maker. Thus, the situation changes according to the knowledge of these parameters.

It is the aim of the decision maker to make improved estimates of the unknown parameters and to find a policy (prescribing his actions in all possible situations) such that his long run expected reward is maximized.

The article is structured as follows:

In section 2, the basic components and notions are presented. At first, we introduce a family of stationary Markov decision models, depending on an unknown parameter, which represents the missing knowledge. Then we define the notion of (deterministic non-Markov) policies and the pertinent stochastic processes. Finally, we discuss two standard global value functions to measure the quality of different policies. These are: the long run average of the expected rewards per stage, and the sum of discounted expected rewards (discounted to the starting time).

In sections 3 and 4, the commonly used algorithms for solving these problems and the pertinent estimation methods are presented. Section 5 contains some typical applications. Some related problems are discussed in section 6.

2. Basic Models and Valuations

2.5 Stationary Adaptive Markov Decision Models

The situation described in the beginning of section 1 will now be formalized as a "Macros Decision Model". This model contains all structural and numerical information needed to describe the system. It consists of the following components:

1. The evolution of the system will be observed and decided about at a set T of discrete time points which are numbered $n = 0, 1, 2, ...$ and may be thought to be equidistant.
2. Let S be the set of possible states of the system, the so-called *state space*. For simplicity we assume that S will be countable (finite or countably infinite). In our example, $i \in S$ may be the number of shoes in stock.
3. Let A be the (nonempty) set of all possible actions (or decisions) called *action space*. Here $a \in A$ may be the number of shoes to be ordered.
4. Not all actions may be allowed in every state (e.g. if there are restricted resources). Therefore we introduce the nonempty subsets $A(i)$ of actions feasible in state $i \in S$.
5. If $i \in S$ is the state of the system (at some point of time) and action $a \in A(i)$ is chosen, then $p(i, a, j)$ is the probability that the next state will be $j \in S$. The mapping p is called the *transition law*. In case of inventory management p may be derived from the random demand.
6. Assigned to each feasible state-action pair (i, a) there is a reward $r(i, a)$. These rewards are later summarized to an over-all value function. The mapping r is called the (one-step) *reward function*, which is assumed to be bounded.
7. To compare rewards obtained at different time points a *discount factor* $ß$ with $0 < ß \leq 1$ is introduced. The most important cases are $ß < 1$ where $ß = (1 + \rho)^{-1}$ reflects the interest rate ρ, and $ß = 1$ when long run averages are considered.

So far we have defined a Markov Decision Model which is stationary, i.e. all components do not depend on the time points $n \in T$. This model may be summarized as $(S, A, p, r, ß)$.

However, this model does not take into account the "adaptive" aspect, i.e. the varying knowledge about the probability law p and the reward function r.

For this reason, we introduce a family of stationary Markov decision models depending on a parameter θ. The set Θ of all parameters θ reflects the different states of information.

Definition

A *Stationary Adaptive Markov Decision Model* is a family

$$\left((S, A, p^{\theta}, r^{\theta}, \beta), \theta \in \Theta\right) \tag{1}$$

of stationary Markov decision models as described above where Θ is assumed to be a compact subset of a metric space (allowing e.g. Θ to be finite or countably infinite).

(For more details see *Markov Decision Processes*.)

2.6 Policies and Value Functions

The above concept of a Markov Decision *Model* describes only the local behavior of a system. There is no connection between the available information and the actions to be chosen. This connection will be established by the concept of a policy, i.e. a prescription of actions in advance for all stages and all possible states of information. This prescription (e.g. by a decision maker) determines totally the probabilistic behavior of the system and is formalized as follows.

Definition

a. Denote by $h_n := (i_0, i_1, i_2, ..., i_n)$ $(n \geq 0)$ any *history* of the system up to time n.
b. The functions f_n assigning to a history h_n an action $a \in A(i_n)$ are called *decision rules* (at time n). These rules are deterministic and non-Markov.
c. Any sequence $\pi := (f_0, f_1, f_2, ...)$ of decision rules is called a *policy*, more precisely, a deterministic non-Markov policy with infinite horizon. Let Π be the set of all such policies.
d. Using a policy π the transition probabilities will be non-stationary and non-Markov, since action a at time n is chosen according to the decision rule f_n which depends on the history h_n. Similarly, the one stage rewards depend on h_n. Thus we write for $h_n = (i_0, i_1, ..., i_n)$, $i_{n+1} \in S$

$$p^{\theta}_{n\pi}(h_n, i_{n+1}) := p^{\theta}(i_n, f_n(h_n), i_{n+1}), \tag{2}$$

$$r^{\theta}_{n\pi}(h_n) := r^{\theta}(i_n, f_n(h_n)). \tag{3}$$

e. If we fix a parameter θ, a state i, and a policy $\pi \in \Pi$, then we obtain a stochastic process $X_0, X_1, X_2, ...$, describing the development of the states X_n of the system over time. The pertinent probability measure $\mathbb{P}^{\theta}_{\pi i}$ on the set of infinite histories $h = (i_0, i_1, ...)$ is defined (according to the theorem of IONESCU-TULCEA) by

$$\begin{aligned} &\mathbb{P}^{\theta}_{\pi i}(X_0 = i_0, X_1 = i_1, ..., X_n = i_n) \\ &= \delta_{ii_0}\, p^{\theta}_{0\pi}(i_0, i_1)\, p^{\theta}_{1\pi}(h_1, i_2) ... p^{\theta}_{n-1,\pi}(h_{n-1}, i_n),\ n = 1, 2, ..., \\ &\qquad \text{where } \delta_{ij} = 1, \text{ if } i = j, \text{ else } \delta_{ij} = 0. \end{aligned} \tag{4}$$

We will consider (for fixed θ, some starting state i and some policy π) two different valuations of the behavior of the system over time:

In case of $ß = 1$ we use the long-run average expected reward $G^{\theta}_{\pi i}$ (see (5) below), whereas in case of $ß < 1$ we consider the expected discounted total reward $V^{\theta}_{\pi}(i)$ (see (16) below), both depending on θ, i and π.

Our problem is to maximize these functions over all policies, if possible. A "solution" of this problem consists in either case of two components: the maximal value *and* a policy by which this value is obtained.

In the following two sections we show how to find solutions if the parameter θ is *known*. However, we are interested in policies, which are optimal for the *true, but unknown* parameter.

Therefore, while the process is running, the true parameter θ will be sequentially estimated based on improving information. These estimates are used to derive policies which are optimal (or nearly optimal) for the true θ. This concept will be implemented in Section 3.

(For more details on policies and value functions see *Markov Decision Processes*.)

-
-
-

Bibliography

Hernández-Lerma O. (1989). *Adaptive Markov Control Processes*, 148 pp. New York: Springer-Verlag. [Graduate textbook on Markov decision processes under uncertainty.]

Kurano M. (1987). Learning Algorithms for Markov Decision Processes. J*ournal of Applied Probability*, **24**, 270-276. Applied Probability Trust, Sheffield. [Interesting paper on adaptive control using relative frequencies.]

Mandl P. (1974). Estimation and Control in Markov Chains. *Advances of Applied Probability*, **6**, 40-60. Applied Probability Trust, Sheffield. [Interesting paper, one of the first on adaptive dynamic programming, containing also detailed proofs.]

Puterman M.L. (1994). *Markov Decision Processes*, 649 pp. New York: Wiley. [Textbook on Markov decision processes including many details on algorithms, contraction properties and error bounds.]

Rieder U. (1975) Bayesian Dynamic Programming. *Advances of Applied Probability*, **7**, 330-348. Applied Probability Trust, Sheffield. [Basic paper on Bayesian methods in stochastic dynamic programming.]

Biographical Sketch

Gerhard Hűbner is Professor of Mathematics at the University of Hamburg, Germany, in the Center of Mathematical Statistics and Stochastic Processes. He received his doctoral degree and his research degree (Habilitation) from the University of Hamburg. His research area includes stochastic dynamic programming with emphasis on algorithms, approximations and adaptive methods. He has been co-editor of the journal *Operations Research Spektrum*, and is author of an application-oriented textbook on stochastics (in German).