Physica A xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Physica A

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

On the Bellman's principle of optimality

Eitan Gross *

Department of Physics, University of Arkansas, Fayetteville, AR 72701, USA

HIGHLIGHTS

- A new proof for Bellman's equation of optimality is presented.
- Our proof rests its case on the availability of an explicit model of the environment that embodies transition probabilities and associated costs.
- Contrary to previous proofs, our proof does not rely on $\mathcal{L}$-estimates of the distribution of stochastic integrals.

ARTICLE INFO

Article history:
Received 14 April 2016
Available online xxxx

Keywords:
Dynamic programming
Markov decision processes
Principle of optimality

ABSTRACT

Bellman's equation is widely used in solving stochastic optimal control problems in a variety of applications including investment planning, scheduling problems and routing problems. Building on Markov decision processes for stationary policies, we present a new proof for Bellman's equation of optimality. Our proof rests its case on the availability of an explicit model of the environment that embodies transition probabilities and associated costs.

1. Introduction

Pontryagin's maximum principle describes necessary conditions to find a strong maximum in a non-classical variational problem in the mathematical theory of optimal control. Hence, it is commonly used to find the best possible control for taking a dynamical system from one state to another, especially in the presence of constraints for the state or input controls. Specifically, it accounts to the fact that a dynamic system may evolve stochastically and that key variables may be unknown or imperfectly observed. The theory underlies fundamental concepts in physics including in quantum mechanics [1–3], inference and statistical physics [1–3] and large deviations [4].

Often, however, the real problem does not have a single criterion by which a solution can be judged. A key idea that the current paper deals with is that optimization over time can often be regarded as 'optimization in stages'. One trades off the desire to obtain the lowest possible cost at the present stage against the implication this would have for costs at future stages. The best action minimizes the sum of the cost incurred at the current stage and the least total cost that can be incurred from all subsequent stages, consequent on this decision. This course of action is known as the Principle of Optimality, which for the case of discrete time variables can be mathematically formalized by the dynamic programming equation, or as it commonly known, the Bellman's equation [5]. Dynamic programming has been used in a variety of applications including investment planning [6–8], scheduling problems [9–14] and routing problems [15–18]. The relationship between Pontryagin's maximum principle and Bellman's dynamic programming has been discussed by several researchers [19–23]. Together, the two methods are the most important tools in solving stochastic optimal control problems.

* Correspondence to: Statistical Analyses and Bioinformatics, 211 Cambridge Place Drive, Little Rock, AR, 72227, USA.
E-mail address: dreitangross@gmail.com.

http://dx.doi.org/10.1016/j.physa.2016.06.083
0378-4371/

More specifically, dynamic programming is a mathematical technique that deals with decisions that are made in stages, with the outcome of each decision being predictable to some extent before the next decision is made. A key aspect of such situations is that decisions cannot be made in isolation. Rather, the desire for a low cost at the present must be balanced against the undesirability of high cost in the future. This is a credit assignment problem, because credit or blame must be assigned to each one of a set of interacting decisions. For optimal planning, it is necessary to have an efficient tradeoff between immediate and future costs. Such a tradeoff is indeed captured by the formalism of dynamic programming. In particular, dynamic programming addresses the question of how can an agent or a decision maker improve its long-term performance in a stochastic environment when the attainment of this improvement may require having to sacrifice short-term performance? To address this issue, in Section 2, we build a model around Markov decision processes. Given the initial state of a dynamic system, a Markov decision process will provide the mathematical basis for choosing a sequence of decisions that will maximize the returns from an N-stage decision-making process. In Section 3, we proceed to prove Bellman's dynamic programming equation. Previous proofs of Bellman's equation [24] have used Λ-estimates of the distribution of stochastic integrals and theorems on passage to a limit under the action of a non-linear differential operator.

2. Derivation of Bellman's equation

Let $g(X_n,\ \mu_n(X_n),\ X_{n+1})$ represent the observed cost sustained as the result of the transition from state X_n to state X_{n+1} due to policy $\mu_n(X_n)$. The total expected cost in an infinite-horizon problem, starting from an initial state $X_0 = i$ and due to policy $\pi = \{\mu_n\}$, is defined by

$$J^{\pi}(i) = E\left[\sum_{n=0}^{\infty} \gamma^n g\left(X_n,\ \mu_n(X_n),\ X_{n+1} | X_o = i\right)\right], \tag{1}$$

where the expected value is calculated with respect to the Markov chain $\{X_1,\ X_2, \ldots\}$ and γ is the discount factor. The function $J^{\pi}(i)$ is the cost-to-go function under policy π starting from state i. Its optimal value, denoted by $J^*(i)$, is defined by

$$J^*(i) = \min_{\pi} J^{\pi}(i), \tag{2}$$

where policy π is optimal if, and only if, it is greedy with respect to $J^*(i)$. We use the term "greedy" here to describe the case when the agent or algorithm seeks to minimize the immediate next cost irrespective of the possibility that such an action may not provide access to better alternatives in the future. When the policy π is stationary, i.e. $\pi = \{\mu,\ \mu,\ \mu, \ldots\}$, we will use the notation $J^{\pi}(i)$ instead of $J^*(i)$ and say that π is optimal if

$$J^{\pi}(i) = J^*(i) \quad \text{for all initial states } i. \tag{3}$$

The dynamic-programming technique rests on Bellman's principle of optimality which states that an optimal policy possesses the property that whatever the initial state and initial decision are, the decisions that will follow must create an optimal policy starting from the state resulting from the first decision. Here, we use the term "decision" to indicate a choice of control at a particular time, and the term "policy" to indicate the entire control sequence or control function. To formulate the principle of optimality in mathematical terms, we consider a finite horizon problem for which the cost-to-go function is defined by

$$J_o(X_o) = E\left[g_k(X_k) + \sum_{n=0}^{K-1} g_n(X_n,\ \mu_n(X_n),\ X_{n+1})\right] \tag{4}$$

where K is the planning horizon (i.e., the number of stages) and $g_k(X_k)$ is the terminal cost. Given X_0, the expectation in Eq. (4) is with respect to the remaining states $X_1, \ldots, X_{K-1}$. We can thus state the principle of optimality as follows: Let $\pi^* = \left\{\mu_0^*,\ \mu_1^*, \ldots, \mu_{k-1}^*\right\}$ be an optimal policy for the basic finite-horizon problem and assume that when using the optimal policy π^*, a given state X_n occurs with positive probability. Consider the sub-problem where the environment is in state X_n at time n, and suppose we wish to minimize the corresponding cost-to-go function:

$$J_n(X_n) = E\left[g_k(X_k) + \sum_{k=n}^{K-1} g_k(X_k,\ \mu_k(X_k),\ X_{k+1})\right], \tag{5}$$

for $n = 0, 1, \ldots, K-1$. Then, the truncated policy $\left\{\mu_n^*,\ \mu_{n+1}^*,\ \ldots,\ \mu_{k-1}^*\right\}$ is optimal for the sub-problem. One may justify the principle of optimality by saying that if the truncated policy $\left\{\mu_n^*,\ \mu_{n+1}^*, \ldots, \mu_{k-1}^*\right\}$ was not optimal, then once the state X_n is reached at time n, we could reduce the cost-to-go function $J_n(X_n)$ simply by switching to a policy that is optimal for the sub-problem. To summarize, the principle of optimality builds on the adage of "divide and conquer" in which an optimal policy for a complex multistage control problem is constructed by the following procedure: (i) construct an optimal policy for the "tail sub-problem" involving only the last stage of the system; (ii) extend the optimal policy to the "tail sub-problem" involving the last two stages of the system; (iii) continue the procedure in this manner until the entire problem has been solved.

Please cite this article in press as: E. Gross, On the Bellman's principle of optimality, Physica A (2016), http://dx.doi.org/10.1016/j.physa.2016.06.083