On the convexity of the value function in Bayesian optimal control problems*

Yaw Nyarko
Department of Economics, New York University, New York, NY 10003, USA

Received: July 16, 1991; revised version January 18, 1993

Summary. I study the question on the convexity of the value function and Blackwell (1951)'s Theorem and relate this to the uniqueness of optimal policies. The main results will conclude that strict convexity and a strict inequality in Blackwell's Theorem will hold if and only if from different priors different optimal actions may be chosen.

I. Introduction

I study the question of the convexity of the value function and Blackwell's Theorem (1951) and relate this to the uniqueness of optimal policies. The main results will conclude that strict convexity and a strict inequality in Blackwell's Theorem will hold if and only if from different priors different optimal actions may be chosen. The principal purpose of this paper is to provide simple and accessible proofs to economists of the above results. Most of the proofs of these results in the literature either use special assumptions (e.g., finiteness of the set of signals and/or states); or, because of a greater interest in maximal generality, they rely on arguments which make them inaccessible to many economists. For further work on the convexity of the value function and Blackwell's Theorem see Kihlstrom (1984) or LeCam (1964). For applications in economics see Rothschild (1974), Grossman et al. (1977), Kiefer and Nyarko (1989), and Nyarko (1991).

II. The decision problem

An agent does not know the true value of a parameter \( \theta \), in a parameter space \( \Theta \). The agent's (prior) beliefs about \( \theta \) in the first period, date 1, are represented by the prior probability \( \mu_0 \) over \( \Theta \). The agent must choose at each date \( t \) an action \( a_t \) in an action space \( A \). The action results in an observation (or signal) \( y_t \) in the set \( Y \),

* Financial support from the C. V. Starr Center and the Research Challenge Fund at New York University is gratefully acknowledged. I thank Professor Tara Vishwanathan whose questioning resulted in this paper.
with probability distribution $P(dy_i|a_i, \theta)$ which depends upon both the action $a_i$ and the parameter $\theta$. The action and observation result in a date $t$ utility of $u(a_i, y_t)$. We suppose that $\Theta, A$ and $Y$ are complete and separable metric spaces. Given any metric space $S$ we let $\mathcal{P}(S)$ denote the set of all (Borel) probability measures on $S$. $\mathcal{P}(S)$ will, unless otherwise stated, be endowed with the topology of weak convergence (see Billingsley (1968) for more on this). Suppose that at some date $t$ the agent begins the period with beliefs about the parameter $\theta$ represented by the posterior probability $\mu_{t-1}$. The agent uses the observation $y_t$ resulting from the action $a_t$ to update the posterior via Bayes’ rule (i.e., the laws of probability). We let $B: A \times Y \times \mathcal{P}(\Theta) \to \mathcal{P}(\Theta)$ be the Bayes’ rule operator, so that for all $t \geq 1$

$$\mu_t = B(a_t, y_t, \mu_{t-1})$$

(2.1)

II.a. Histories and policies

A date $n$ partial history is any sequence of observations, actions and prior probabilities at all dates preceding $n$; i.e., $h_n \equiv (\mu_0, \{(a_t, y_t, \mu_t)\}_{t=1}^{n-1}) \in \mathcal{P}(\Theta) \times \Pi_{t=1}^{n-1} [A \times Y \times \mathcal{P}(\Theta)] = H_n$. A policy is a sequence $\pi = \{\pi_t\}_{t=1}^{\infty}$, where for each $t \geq 1$, $\pi_t: H_t \to A$ specifies the date $t$ action $a_t = \pi_t(h_t)$, as a (measurable) function of the partial history. We let $D$ denote the set of all policies. Any policy, $\pi$, given initial beliefs $\mu_0$, then generates a sequence of actions, observations and posterior beliefs, $\{(a_t, y_t, \mu_{t-1})\}_{t=1}^{\infty}$, via the conditional probability $P(dy\|a, \theta)$ and the Bayesian updating rule (2.1). Any policy $\pi$ results in a sum of expected discounted rewards with discount factor $\delta \in (0, 1)$, $V_\pi(\mu_0) \equiv E[\sum_{t=1}^{\infty} \delta^{t-1} u(a_t, y_t)\|\mu_0, \pi]$. We define the value function by $V(\mu) \equiv \sup_{\pi \in D} V_\pi(\mu)$. A policy is optimal if it attains this supremum. We assume the existence of an optimal policy and we assume the utility function is uniformly bounded so our value function is well defined. (See Kiefer and Nyarko (1989) for conditions which ensure this and for other details.)

II.b. *-Policies

We define a date $n$ *-partial history, $h^*_n$, to be any sequence of past actions and observations (i.e., the same as date $n$ partial histories without a specification of the history of posterior beliefs); i.e., any $h^*_n \equiv \{(a_t, y_t)\}_{t=1}^{n-1} \in \prod_{t=1}^{n-1} A \times Y = H^*_n$. A date $n$ *-policy is any function $\pi_n^*: H^*_n \to A$ and a *-policy is any collection of date $n$ policies, $\pi^* = \{\pi_n^*\}_{n=1}^{\infty}$. Let $D^*$ denote the set of all *-policies. At date one there is no history so $h^*_1$ is a “null” history and $\pi^*_1$ is identified with an action (i.e., $\pi^*_1 \in A$). Any parameter value $\theta$ and *-policy $\pi^*$ generates the sequence of actions and observations, $\{(a_n, y_n)\}_{n=1}^{\infty}$, defined inductively via the following relations:

$$a_1 = \pi^*_1, \quad y_1 \sim P(dy\|\theta, a_1) \quad \text{and} \quad y_n \sim P(dy\|\theta, a_n)$$

where by the notation $y_n \sim P$ we mean that $y_n$ is a random variable with distribution given by the probability $P$. $\pi^*$ and $\theta$ also generate a conditional expected discounted sum of returns

$$L(\pi^*, \theta) = E[\sum_{n=1}^{\infty} \delta^{n-1} u(a_n, y_n)\|\theta, \pi^*]$$

(2.2)

Recall that a date $n$ policy as defined in section II.a. is allowed to be a function also of the history of posterior distributions. Hence every *-policy may also be