On the Convexity of the Value Function in Bayesian Optimal Control Problems
Author(s): Yaw Nyarko
Source: *Economic Theory*, Vol. 4, No. 2 (Mar., 1994), pp. 303-309
Published by: Springer
Stable URL: http://www.jstor.org/stable/25054763
Accessed: 09/05/2010 18:41

Econ. Theory 4, 303–309 (1994)

Economic Theory

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 θ, in a parameter space Θ. The agent's (prior) beliefs about θ in the first period, date 1, are represented by the prior probability μ_0 over Θ. 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 $P(dy_t|a_t, \theta)$ which depends upon both the action a_t and the parameter θ. The action and observation result in a date t utility of $u(a_t, y_t)$. We suppose that Θ, A and Y are complete and separable metric spaces. Given any metric space S we let $\mathscr{P}(S)$ denote the set of all (Borel) probability measures on S. $\mathscr{P}(S)$ 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 θ represented by the posterior probability μ_{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 $B: A \times Y \times \mathscr{P}(\Theta) \to \mathscr{P}(\Theta)$ be the Bayes' rule operator, so that for all $t \geq 1$

$$\mu_t = B(a_t, y_t, \mu_{t-1}) \tag{2.1}$$

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., $h_n \equiv (\mu_0, \{(a_t, y_t, \mu_t)\}_{t=1}^{n-1}) \in \mathscr{P}(\Theta) \times \Pi_{t=1}^{n-1} [A \times Y \times \mathscr{P}(\Theta)] \equiv H_n$. 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 (measureable) function of the partial history. We let D denote the set of all policies. Any policy, π, given initial beliefs μ_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 $P(dy|\theta, a)$ and the Bayesian updating rule (2.1). Any policy π 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 $V(\mu) \equiv \mathrm{Sup}_{\pi \in D} V_\pi(\mu)$. 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 π_1^* is identified with an action (i.e., $\pi_1^* \in A$). Any parameter value θ and *-policy π^* generates the sequence of actions and observations, $\{a_n, y_n\}_{n=1}^{\infty}$, defined inductively via the following relations:

$a_1 = \pi_1^*$, $y_1 \sim P(dy|\theta, a_1)$ and given the *-partial history h_n^*, $a_n = \pi_n^*(h_n^*)$ and $y_n \sim P(dy|\theta, a_n)$, where by the notation $y_n \sim P$ we mean that y_n is a random variable with distribution given by the probability P. π^* and θ also generate a conditional expected discounted sum of returns

$$L(\pi^*, \theta) = E[\textstyle\sum_{n=1}^{\infty} \delta^{n-1} u(a_n, y_n)|\theta, \pi^*] \tag{2.2}$$

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 considered a policy, but not vice versa. However, a *fixed* prior μ and policy π "induce" a *-policy as follows: Define $\pi_1^* = \pi_1(\mu)$; for any $n \geq 2$, any *-partial history h_n^* will generate via Bayes' rule (2.1) a unique sequence of posteriors $\{\mu_t\}_{t=1}^{\infty}$ and hence a unique partial history h_n; so define $\pi_n^*(h_n^*) = \pi_n(h_n)$. The constructed *-policy, π^*, will generate the same sequence of actions and observations as the policy π from fixed initial prior μ. (Of course, a *-history, h_n^*, is any member of H_n^*, and in particular may not be consistent with Bayes rule (2.1) from initial prior μ. Such *-partial histories have a zero probability of being observed under the fixed prior μ and policy π. At such *-partial histories, h_n^*, we are free to arbitrarily define the corresponding date n *-policy, π_n^*.)

It should be fairly obvious from the above constructions that for an agent with *fixed* prior μ, choosing a policy is equivalent to choosing a *-policy. The argument is as follows: Every *-policy may be considered a policy – one which is independent of the posteriors. Hence choosing a policy (as opposed to a *-policy) can not result in lower sum of discounted payoffs. Further, for *fixed* prior, we argued above that any policy induces a *-policy which results in the same distribution of actions and observations, and hence the same payoffs. So, for *fixed* prior, for each policy there exists a *-policy which attains the same sum of discounted payoffs. This shows the required equivalence. Now fix any policy π and initial prior μ and let π^{**} be the *-policy induced by π and μ. We may therefore conclude that

$$V_\pi(\mu) = \int L(\pi^{**}, \theta)\mu(d\theta) \tag{2.3}$$

and

$$\mathrm{Sup}_{\pi \in D} V_\pi(\mu) = V(\mu) = \mathrm{Sup}_{\pi^* \in D^*} \int L(\pi^*, \theta)\mu(d\theta) \tag{2.4}$$

Any *-policy that attains the supremum on the right-hand-side of (2.4) shall be called an optimal *-policy for the given initial prior, μ.

III. The (weak) convexity of the value function

Let M be the vector space of all finite signed measures on the parameter space Θ endowed with the total variation norm defined by $\|\mu\| = \mathrm{Sup}_A|\mu(A)|$, where the supremum is over measureable subsets A of Θ. For any *-policy, π^*, and any $\mu \in M$, we may define (with a slight abuse of notation – see (2.2)):

$$L(\pi^*, \mu) = \int L(\pi^*, \theta)\mu(d\theta). \tag{3.1}$$

Note that the function $L(\pi^*, \mu)$ is a function of μ only directly through the second argument and not indirectly through π^*, (which is indeed why we work with *-policies as opposed to policies)! The lemma below is easy to check and implies Proposition 3.2, the convexity to the value function.

Lemma 3.1. *(a) $L(\pi^*, \mu)$ is linear in μ for fixed *-policy π^*. (b) $L(\pi^*, \mu)$ is continuous on M under the total variation norm, uniformly in π^*; i.e., $\exists K > 0$ such that for all μ' and μ'' in M, $\mathrm{Sup}_{\pi^* \in D^*}|L(\pi^*, \mu') - L(\pi^*, \mu'')| \leq K\|\mu' - \mu''\|$.*

Proposition 3.2. *The value function V is (a) continuous on M when M is endowed with the total variation norm and (b) convex on M.*

Proof. The value function, V, can of course be defined on the set of finite signed measures M (as opposed being defined on the set of probability measures) e.g., by (2.3) and (2.4). From (2.4) and Lemma 3.1(b) it is easy to check that V is continuous. Further, it is easy to check that the supremum of an arbitrary collection of linear real-valued functions is convex. Hence from (2.4) and the linearity of $L(\pi^*, \mu)$ in μ we conclude that the value function is convex.

IV. Strict convexity of the value function and uniqueness of optimal actions

We now show that the value function is *NOT* strictly convex between two priors if and only if all the priors in the convex hull of those two priors share a common optimal *-policy function. In particular, if one can check that all the priors in the convex hull of two priors do *NOT* share a common *date one* optimal action then the value function must be strictly convex between those two priors.

Proposition 4.1. *Fix any $\mu', \mu'' \in M$ and $\lambda \in (0,1)$. Denote their convex hull by $C(\mu', \mu'') \equiv \{\mu \in M \mid \exists \phi \in [0,1]$ s.t. $\mu = \phi\mu' + (1-\phi)\mu''\}$. Then $V(\lambda\mu' + (1-\lambda)\mu'') = \lambda V(\mu') + (1-\lambda)V(\mu'')$ if and only if $\exists$ a *-policy π^*, such that from all initial priors $\mu \in C(\mu', \mu'')$, π^* is an optimal *-policy from that prior μ.*

V. The comparison of experiments

Any action and observation combination $(a, y) \in A \times Y$ provides information on the unknown parameter θ and hence may be referred to as an **experiment** on θ. We shall say that Experiment A is **sufficient** for Experiment B if the observation in Experiment B is the observation of Experiment A perturbed by some noise. Formally we have:

Definition 5.1. Fix any actions a and a' in A and let y and y' denote the observations resulting from those actions. The experiment $P(dy \mid a, \theta)$ is sufficient for $P(dy' \mid a', \theta)$ if there exists a conditional probability $M(dy' \mid y)$ (of y' conditional on y) such that $P(dy' \mid a', \theta) = M(dy' \mid y) \cdot P(dy \mid a, \theta)$ for all θ in Θ; i.e., for any bounded (measurable) function $g: Y \to R$ and any $\theta \in \Theta$, $\int_Y g(y')P(dy' \mid a', \theta) = \int_Y \int_Y g(y')M(dy' \mid y) \cdot P(dy \mid a, \theta)$. In particular, if $y^{*\prime}$ is the random variable generated via $M(dy^{*\prime} \mid y) \cdot P(dy \mid a, \theta)$ then conditional on θ, y' and $y^{*\prime}$ have the same distribution.

Example 5.2. Suppose A and Θ are both subsets of the real line and $y = \theta a + \varepsilon$ with ε normally distributed with zero mean and unit variance. Fix any a and $a' \in A$ with $|a| > |a'| > 0$. Then one may show that $P(dy \mid a, \theta)$ is sufficient for $P(dy' \mid a', \theta)$, with $y^{*\prime} = (ya'/a) + \varepsilon^*$ where ε^* is independent of ε and normally distributed with mean zero and variance equal to $[1 - (a'/a)^2]$.

Proposition 5.3. *(Blackwell's Theorem). If the experiment $P(dy \mid a, \theta)$ is sufficient for $P(dy' \mid a', \theta)$ then $\forall \mu \in \mathscr{P}(\Theta)$, $\int_{\Theta \times Y} V(B(a', y', \mu))P(dy' \mid a', \theta)\mu(d\theta) \leq \int_{\Theta \times Y} V(B(a, y, \mu)) \cdot P(dy \mid a, \theta)\mu(d\theta)$.*

VI. The 'strict comparison' of experiments

We seek to determine when the inequality in Proposition 5.3 can be made strict. As one may guess from Proposition 4.1, the conclusion will be of the form that

Proposition 5.3 holds with equality if and only if the actions from the two experiments are somehow the "same." Fix the initial prior μ and let $P(dy|a,\theta)$ be sufficient for $P(dy'|a',\theta)$. Let $y^{*\prime}$ be as in the definition (5.1) of sufficient experiments. Let Q denote the joint probability distribution of $y^{*\prime}$, y and θ generated by letting θ have distribution μ, letting y have distribution $P(dy|a,\theta)$, and letting $y^{*\prime}$ have distribution $M(dy^{*\prime}|y)$; (i.e., given any real-valued function f on $Y \times Y \times \Theta$, $\int f dQ = \int_{Y\times Y\times\Theta} f(y^{*\prime}, y, \theta) M(dy^{*\prime}|y)P(dy|a,\theta)\mu(d\theta)$). Suppose that at date one, two initially identical agents with the same prior μ choose the actions a and a' respectively. Upon making their observations, y and y' respectively, they will choose date 2 actions given by $a_2(y)$ and $a_2'(y')$ respectively. Let $a_2^{*\prime}(y^{*\prime})$ be the optimal date 2 action of the agent who observes only $y^{*\prime}$. Given any probability P on a metric space let Supp P denote the support of P (i.e., the smallest closed set with P-probability one). Then we have,

Proposition 6.1. *Suppose Proposition 5.3 holds with equality. Then Q-a.e., (a) $Q(\{a_2 = a_2^{*\prime}\}|y^{*\prime}) = 1$ for all $y^{*\prime}$; (i.e., outside of a set of $(y, y^{*\prime})$ values with Q-probability zero, for all $y^{*\prime}$, $a_2(y) = a_2'(y^{*\prime})$ for all $y \in \mathrm{Supp}\, Q(\cdot|y^{*\prime})$), and (b) a_2 and a_2' have the same distribution conditional on θ; i.e., if C is any (measurable) subset of the action space A then $Q(\{a_2 \in C\}|\theta) = Q(\{a_2' \in C\}|\theta) \forall \theta$.*

Remark. Suppose that the observations take values on the real line and that the conditional probability $M(dy^{*\prime}|y)$ and the marginal distribution on Y of Q, both admit strictly positive density functions, $q(y, y^{*\prime})$ and $q(y)$ respectively (as in Ex. 5.2). Then using Bayes' rule it is easy to see that the marginal distribution on Y of $Q(\cdot|y^{*\prime})$ has a strictly positive density function given by $q(y, y^{*\prime})/q(y)$. Then $\{y \in \mathrm{Supp}\, Q(dy|y^{*\prime})\} = Y$. Proposition 6.1(a) concludes that there is a common action which is optimal from all observations y. Hence, if the utility function of the agent is such that different observations of y lead to different optimal actions, Proposition 5.3 must hold with strict inequality.

Appendix: The Proofs

Proof of Lemma 3.1. (a) Obvious. (b) Given any $\mu \in M$ define μ^+ and μ^- to be, resp., the positive and negative parts of μ; then both μ^+ and μ^- are non-negative measures and $\mu = \mu^+ - \mu^-$ (see Dunford and Schwartz (1957)). By assumption the utility function is uniformly bounded so we may suppose without loss of generality that $L(\pi^*, \theta)$ is non-negative and bounded above uniformly in (π^*, θ) by some $K > 0$. Then for any *-policy π^*, and any $\mu', \mu'' \in M$,

$$\begin{aligned}|L(\pi^*, \mu') - L(\pi^*, \mu'')| &= |L(\pi^*, \mu' - \mu'')| = |\textstyle\int L(\pi^*, \theta) d(\mu' - \mu'')| \\ &= |\textstyle\int L(\pi^*, \theta) d(\mu' - \mu'')^+ - \int L(\pi^*, \theta) d(\mu' - \mu'')^-| \le \int L(\pi^*, \theta) d(\mu' - \mu'')^+ \\ &\quad + L(\pi^*, \theta) d(\mu' - \mu'')^- \le K[(\mu' - \mu'')^+(\Theta) + (\mu' - \mu'')^-(\Theta)] \le 2K \|\mu' - \mu''\|\end{aligned}$$

which implies part (b) of lemma.

Proof of Proposition 4.1: Suppose there exists $\mu', \mu'' \in \mathscr{P}(\Theta)$ and a $\lambda \in (0, 1)$, such that $V(\lambda\mu' + (1-\lambda)\mu'') = \lambda V(\mu') + (1-\lambda)V(\mu'')$. Let π^* be any *-policy optimal from initial prior $\lambda\mu' + (1-\lambda)\mu''$. Then $V(\mu') \ge L(\pi^*, \mu')$ and $V(\mu'') \ge L(\pi^*, \mu'')$. Suppose, per absurdum, that from one of the priors μ' or μ'', the *-policy π^* is not optimal;

then one of these inequalities hold strictly. Then,

$$V(\lambda\mu' + (1-\lambda)\mu'') = \lambda V(\mu') + (1-\lambda)V(\mu'') > \lambda L(\pi^*, \mu') + (1-\lambda)L(\pi^*, \mu'')$$
$$= L(\pi^*, \lambda\mu' + (1-\lambda)\mu'') = V(\lambda\mu' + (1-\lambda)\mu''),$$

a contradiction. Hence the *-policy π^* is optimal from initial priors μ' and μ''. Now fix any $\phi \in [0, 1]$ and suppose that from initial prior $\phi\mu' + (1-\phi)\mu''$, π^* is not optimal. Then $V(\phi\mu' + (1-\phi)\mu'') > L(\pi^*, \phi\mu' + (1-\phi)\mu'') = \phi L(\pi^*, \mu') + (1-\phi)L(\pi^*, \mu'') = \phi V(\mu') + (1-\phi)V(\mu'')$ which is a contradiction to the convexity of V. This proves the "only if" part of the proposition.

Next, for fixed probability measures μ' and μ'' on Θ suppose that there exists a *-policy π^*, such that from all initial priors μ in $C(\mu', \mu'')$, π^* is an optimal *-policy from that prior. Then for all such μ, $V(\mu) = L(\pi^*, \mu)$. The "if" part of the proposition then follows from linearity of L in μ.

Lemma 5.3.1. (Jensen's Inequality): Let S be a complete and separable metric space and let q be any joint (Borel) probability on $\Omega = \Theta \times S$. Let μ and $\mu(\cdot|s)$ denote respectively the marginal and conditional distribution of q on Θ. If W is any continuous and convex function on M (the set of finite signed measures on Θ endowed with the total variation norm) then $W(\mu) \leq \int_s W(\mu(\cdot|s))dq$.

Proof of lemma 5.3.1.: This is a straightforward application of the version of Jensen's inequality typically found in mathematical analysis textbooks. (See Woodroofe (1982)).

Proof of Proposition 5.3. Let μ, a, a', y, and y' be as in Proposition 5.3. Let $y^{*\prime}$ be as in the definition 5.1 of sufficient experiments. Let Q denote the joint probability distribution of $(y^{*\prime}, y, \theta)$ as defined in section VI; and let $\mu(\cdot|y)$, $\mu(\cdot|y^{*\prime})$ and $\mu(\cdot|y^{*\prime}, y)$, denote the marginals on Θ of the probability Q conditional on y, $y^{*\prime}$ and $\{y^{*\prime}, y\}$ respectively. Fix any $y^{*\prime} \in Y$, and consider an agent with initial prior $\mu' = \mu(\cdot|y^{*\prime})$, obtained from the observation of $y^{*\prime}$ only. Suppose such an agent then observes y. The posterior distribution over Θ will be $\mu(\cdot|y^{*\prime}, y)$. Since the value function V is continuous and convex (Proposition 3.2), an application of Jensen's inequality (Lemma 5.3.1) implies that

$$V(\mu(\cdot|y^{*\prime})) \leq \int_Y V(\mu(\cdot|y^{*\prime}, y))dQ(\cdot|y^{*\prime}). \tag{5.3.2}$$

Let LHS, RHS denote respectively the left and right hand side of (5.3.2). Since by construction the distribution of y' conditional on θ is that same as that of $y^{*\prime}$ conditional on θ, $\int(\text{LHS})dQ \equiv \int V(\mu(\cdot|y^{*\prime}))dQ = E[V(B(a', y', \mu))]$. Since conditional on y, $y^{*\prime}$ is independent of θ, $\mu(\cdot|y^{*\prime}, y) = B(a, y, \mu)$, so $\int(\text{RHS})dQ = E[V(B(a, y, \mu))]$. Hence (5.3.2) implies Proposition 5.3.

Proof of Proposition 6.1. If Proposition 5.3 holds with equality then (5.3.2) in the proof of Proposition 5.3 must hold with equality (for Q-a.e. $y^{*\prime}$). For any such $y^{*\prime}$, using arguments similar to those used in the proof of Proposition 4.1, it is easy to show that if the *-policy $\pi^*(y^{*\prime})$ is optimal for the initial prior $\mu(\cdot|y^{*\prime})$ then that *-policy is also optimal for the initial prior $\mu(\cdot|y^{*\prime}, y)$ for $Q(\cdot|y^{*\prime})$ almost every y. Since $\mu(\cdot|y^{*\prime}, y) = \mu(\cdot|y)$, this means that conditional on $y^{*\prime}$, the optimal action from

$\mu(\cdot|y^{*\prime})$ is equal to the optimal action from $\mu(\cdot|y)$, for $Q(\cdot|y^*)$ a.e. y, from which part (a) follows. Next, let C be any (measureable) subset of the action space A, and let E_Q denote expectations with respect to Q. Then $Q(\{a_2 \in C\}|\theta) = E_Q[Q(\{a_2 \in C\}|y^{*\prime}, \theta)|\theta]$ (by iterated conditioning) $= E_Q[Q(\{a_2^{*\prime} \in C\}|y^{*\prime}, \theta)|\theta]$ (from part (a)) $= Q(\{a_2^{*\prime} \in C\}|\theta)$ (upon integration) $= Q(\{a'_2 \in C\}|\theta)$ (since $y^{*\prime}$ has same distribution as y').

References

Billingsley, P.: Convergence of probability measures. New York: Wiley 1968

Blackwell, D.: The comparison of experiments. In: Proceedings of the Second Berkeley Symposium on Statistics and Probability, University of California Press, Berkeley (1951)

Dunford, N., Schwartz, J.: Linear operators. New York: Wiley 1957

Grossman, S., Kihlstrom, R., Mirman, L.: A Bayesian approach to the production of information and learning by doing. Rev. Econ. Stud. **44**, 533–547 (1977)

Kiefer, N., Nyarko, Y.: Optimal control of an unknown linear process with learning. Int. Econ. Rev. **30**, 571–586 (1989)

Kihlstrom, R.: A Bayesian exposition of Blackwell's theorem on the comparison of experiments. In: M. Boyer and R. Kihlstrom (eds). Bayesian models of economic theory. Amsterdam: Elsevier Science Publishers B.V. 1984

LeCam, L.: Sufficiency and approximate sufficiency. Ann. Math. Stat. **35**, 1419–1455 (1964)

Nyarko, Y.: On the convergence of Bayesian posterior processes in linear economic models. J. Econ. Dynam. Control **15**, 687–713 (1991)

Rothschild, M.: A two-armed bandit theory of market pricing. J. Econ. Theory **9**, 185–202 (1974)

Woodroofe, M.: Sequential allocation with covariates. Sankhya Ind. J. Stat. [Ser. A] **44**, pp. 403–414 (1982)