THE VALUE OF A STOCHASTIC INFORMATION STRUCTURE

YARON AZRIELI AND EHUD LEHRER[†][*]

November 16, 2004

Abstract. Upon observing a signal, a Bayesian decision maker updates her probability distribution over the state space, chooses an action, and receives a payoff that depends on the state and the action taken. An information structure determines the set of possible signals and the probability of each signal given a state. For a fixed decision problem (consisting of a state space, action set and utility function) the value of an information structure is the maximal expected utility that the decision maker can get when the observed signals are governed by this structure.

This note studies the functions defined over information structures that measure their value. It turns out that two conditions play a major role in the characterization of these functions: additive separability and convexity.

[†] School of Mathematical Sciences, Tel Aviv University, Tel Aviv 69978, Israel. e-mail: azrieliy@post.tau.ac.il, lehrer@post.tau.ac.il.

[*] Dinah Rosenberg is acknowledged for helpful conversations.

1. INTRODUCTION

Blackwell (1951, 1953) compared different information structures. He defined two orders over the set of information structures: one in terms of decision problems and one in purely probabilistic terms. He showed the equivalence of these two orders. In this note we discuss the same model of stochastic information structures but deal with another issue: the value of information.

A decision maker (DM) has a prior distribution over the true state of nature. Before taking a decision she receives a stochastic signal that depends on the state realized. The set of possible signals and the probability of each signal given the state are determined by the particular information structure of the case. Upon receiving a signal the DM updates her belief and takes an action that maximizes her expected utility.

An outside observer collects data about the DM. He cannot observe the DM's prior distribution nor her actions, while he knows through which information structure the DM receives her signals and the utility associated with each structure. The question arises as to what kind of observations can be rationalized within the Bayesian paradigm. In other words, what conditions the data should satisfy in order to be consistent with a behavior of a utility maximizer in a Bayesian decision problem?

More specifically, a decision problem is defined by a state space, a prior distribution, an action set and a utility function. For a given decision problem, different information structures determine potentially different maximal achievable expected utilities. Thus, a decision problem implicitly induces a function that attaches to any information structure its corresponding value. We deal here with these functions (defined over information structures) and provide conditions that characterize them. These conditions can be used by an outside observer to tell whether the observations are consistent with the behavior of a rational agent in a Bayesian model.

The problem of weather the data collected by an outside observer is consistent with the Bayesian model is similar in spirit to a question answered by Afriat (1967). Afriat (1967) dealt with a database that contains different prices and their corresponding consumption bundles. He found the conditions that such a database should satisfy in order to be consistent with a behavior of a utility maximizing consumer in a competitive market.

One property of such information functions is already well-known. Blackwell (1951, 1953) defined one information structure, say I, as *better than* another (I') if, whatever the decision problem is, the expected utility of the DM is higher when the information structure is I. He showed that I is better than I' if and only if the signal produced by I can be used to simulate the signal produced by I' (I is *more informative* than I'). Thus, every information function is monotonic with respect to the *more informative* order.

Gilboa and Lehrer (1991) investigated properties of information functions whose domain is restricted to the set of deterministic information structures. In such information structures the signal observed by the DM is uniquely determined by the state of nature. Therefore, every deterministic information structure generates a partition of the state space, where an atom of the partition corresponds to a signal. This allows one to translate the model into terms of cooperative games: states of nature take the role of players, and atoms of the partition take the role of coalitions. Furthermore, the worth of a coalition is the maximal utility achievable on the corresponding atom.

When the information structure is stochastic, the translation to cooperative games is not possible anymore. Instead, an information function is expressed by means of another function defined over posteriors. It turns out that the key characteristic of an information function is that this function (defined over the set of posteriors) is *convex*.

The note is organized as follows. In Section 2 the model is presented and the notion of information function is defined. Section 3 discusses an important property of information functions: additive separability. In Section 4 the main result of this note is proved. Section 5 is dedicated to decision problems having finite action set. We conclude with Section 6 which addresses the relation to the entropy and other functions that measure the contents of information.

2. THE MODEL

Let $\Omega = \{\omega_1, \ldots, \omega_n\}$ be a finite set of states of nature, and let μ be the prior probability over Ω. It is assumed that μ assigns a positive probability to any state. That is, $\mu(\omega) > 0$ for every $\omega \in \Omega$. The set of actions available to the DM is denoted by A. The utility of the DM when she takes the action $a \in A$ and when the state of nature is $\omega \in \Omega$ is denoted by $u(a, \omega)$.

An information structure is a pair $I = (S, M)$, where S is the set of signals and M is a collection of distributions on S, one for each state. M can be thought of as a stochastic matrix with n rows; the i-th row of M (for $1 \leq i \leq n$) is the distribution over signals given the state ω_i. Stated differently, the cell M_{is} of the matrix M is the probability of receiving the signal $s \in S$ given that the state of nature is ω_i. Note that the number of columns in M coincides with the number of signals in S. For the sake of simplicity, we always write m instead of $|S|$, where no confusion can arise.

Denote by $\mathcal{I}$ the set of all information structures.

For a given information structure $I = (S, M)$, denote by $\pi_I = (\pi_I^s)_{s \in S}$ the distribution on S induced by M. That is, $\pi_I^s = \sum_{i=1}^n \mu(\omega_i) M_{is}$ is the probability of observing s. Also, for $s \in S$, let $q_{I,s} = (q_{I,s}^1, \ldots, q_{I,s}^n)$ be the distribution on Ω given that the observed signal is s. Formally, $q_{I,s}^i = \mathbb{P}_I(\omega_i | s) = \frac{\mu(\omega_i) M_{is}}{\pi_I^s}$.

A (pure) strategy of the DM is a function $\sigma : S \to A$ which dictates the action to be chosen after observing each of the signals. If the information structure is $I = (S, M)$ and the DM follows a strategy σ, then her expected utility is

$$E_{I,\sigma} = \sum_{i=1}^{n} \mu(\omega_i) \sum_{s\in S} M_{is} u(\sigma(s), \omega_i) = \sum_{s \in S} \pi_I^s \sum_{i=1}^{n} q_{I,s}^i u(\sigma(s), \omega_i).$$

$\hat{\sigma}_I$ is an *optimal strategy*, subject to the information structure I, if it maximizes $E_{I,\sigma}$. In other words, $\hat{\sigma}_I$ is optimal if for every $s \in S$, $\hat{\sigma}_I(s) = a$ implies that $\sum_{i=1}^{n} q_{I,s}^i u(a, \omega_i) \geq \sum_{i=1}^{n} q_{I,s}^i u(b, \omega_i)$ for every $b \in A$. We assume that an optimal strategy exists.

For a certain information structure I and an optimal strategy $\hat{\sigma}_I$, let $v_I(s) = \sum_{i=1}^{n} q_{I,s}^i u(\hat{\sigma}_I(s), \omega_i)$. $v_I(s)$ is the maximal expected utility that the DM can obtain upon observing s. Notice that $v_I(s)$ depends only on the posterior distribution $q_{I,s}$: if $q_{I,s_1} = q_{I,s_2}$, then $v_I(s_1) = v_I(s_2)$.

2.1. **Information functions.**

Consider a decision problem characterized by Ω, μ, A and u. The DM may obtain information about the realized state through various information structures. Each information structure entails a different maximal achievable expected payoff. The main issue of this note is measuring the value of information structures.

The value of an information structure is defined as the maximal expected utility achievable when signals are received according to it. The information function attaches to each information structure its value. Formally,

Definition 1. *A function* $V : \mathcal{I} \to \mathbb{R}$ *is an* information function *if there exist a set of actions* A *and a utility function* $u : A \times \Omega \to \mathbb{R}$ *such that* $V(I) = \sum_{s\in S} \pi_I^s v_I(s)$ *for every* $I = (S, M) \in \mathcal{I}$.

The goal is to study information functions. More specifically, we are interested in characterizing those functions defined over information structures that are information functions of some decision problem.

3. ADDITIVE SEPARABILITY

Definition 2. *$V : \mathcal{I} \to \mathbb{R}$ is* additively separable *if there exist a function $v : \Delta(\Omega) \to \mathbb{R}$ such that $V(I) = \sum_{s\in S} \pi_I^s v(q_{I,s})$ for every $I \in \mathcal{I}$. If v is such a function we will say that v* corresponds *to V.*

It is obvious that any information function is additively separable. We are about to describe natural properties of the function V that are equivalent to additive separability. For that purpose we need the following notation.

Notation 1. *For any information structure $I = (S, M)$ and a subset of signals $T \subseteq S$,*
(a) $\underline{I}(T) = (\underline{S}(T), \underline{M}(T))$ is the information structure defined as follows: $\underline{S}(T) = (S \setminus T) \cup \{t\}$. If $s \in S \setminus T$ then $\underline{M}(T)_{is} = M_{is}$, and if $s = t$ then $\underline{M}(T)_{is} = \sum_{s'\in T} M_{is'}$, for every $1 \leq i \leq n$.
(b) $\overline{I}(T) = (\overline{S}(T), \overline{M}(T))$ is the information structure defined as follows: Every signal $t \in T$ is replaced by a set of n signals $S_t = \{t_1, \ldots, t_n\}$, so the new set of signals is $\overline{S}(T) = (S \setminus T) \bigcup (\cup_{t\in T} S_t)$. For $s \in S \setminus T$, $\overline{M}(T)_{is} = M_{is}$ $(1 \leq i \leq n)$. For $t_k \in S_t$ (for some $t \in T$), if $i = k$ then $\overline{M}(T)_{it_k} = M_{i,t}$ and if $i \neq k$ then $\overline{M}(T)_{it_k} = 0$.
(c) For two disjoint sets of signals $T_1, T_2 \subseteq S$, $\underline{I}(T_1, T_2) = \underline{\underline{I}(T_1)}(T_2)$.

In words, $\underline{I}(T)$ is the information structure which differs from I only in that the columns corresponding to the signals in T are summed up to form one column. Instead of being informed separately of the signals in T, the DM is informed that one of the signals in T occurred. That is, the signals in T are lumped together. On the other hand, $\overline{I}(T)$ is the information structure which is identical to I on $S \setminus T$, and any column corresponding to some $s \in T$ is replaced by a diagonal $n \times n$ matrix.

Definition 3. *A function $V : \mathcal{I} \to \mathbb{R}$ is* Independent of Irrelevant Signals *(IIS) if*

$$V(\underline{I}(T_1)) + V(\underline{I}(T_2)) = V(\underline{I}(T_1, T_2)) + V(I) \tag{1}$$

for every information structure $I = (S, M)$ *and for every two disjoint subsets of signals* $T_1, T_2 \subseteq S$*.*

To justify the term *IIS*, notice that equation (1) can be rewritten as $V(I) - V(\underline{I}(T_1)) = V(\underline{I}(T_2)) - V(\underline{I}(T_1, T_2))$. The left-hand side of this equation is equal to the loss incurred to the DM due to coarsening the information structure: instead of being informed of each signal in T_1 separately, the signals of T_1 are lumped together. This is the value of the information embedded in the set T_1 when the information structure is I. If equation (1) holds for every information structure I and for every subset of signals T_2, it means that this value is independent of I. This is so, because when $T_2 = S \setminus T_1$, the right-hand side of equation (1) depends only on T_1. Therefore, the left-hand side is constant across all information structures that contain T_1. This implies that the contribution of a set of signals (columns in the stochastic matrix) to the value of information is independent of the informational structure out of this set.

Definition 4. *A function* $V : \mathcal{I} \to \mathbb{R}$ *is* reducible *if for every information structure* $I = (S, M)$ *and for any pair of signals* $s_1, s_2 \in S$ *such that* $q_{I,s_1} = q_{I,s_2}$*,* $V(\underline{I}(\{s_1, s_2\})) = V(I)$*.*

Proposition 1. $V : \mathcal{I} \to \mathbb{R}$ *is additively separable iff it is IIS and reducible.*

Proof. Assume first that V is additively separable. Then there exists $v : \Delta(\Omega) \to \mathbb{R}$ such that $V(I) = \sum_{s \in S} \pi_I^s v(q_{I,s})$ for every $I \in \mathcal{I}$. It is straightforward to see that V is reducible. To check that V is IIS, fix some $I \in \mathcal{I}$ and let T_1, T_2 be two disjoint subsets of signals. We have,

$$
\begin{aligned}
& V(\underline{I}(T_1)) + V(\underline{I}(T_2)) = \sum_{s \in \underline{I}(T_1)} \pi^s_{\underline{I}(T_1)} v\left(q_{\underline{I}(T_1),s}\right) + \sum_{s \in \underline{I}(T_2)} \pi^s_{\underline{I}(T_2)} v\left(q_{\underline{I}(T_2),s}\right) = \\
& = \sum_{s \notin T_1} \pi^s_I v\left(q_{I,s}\right) + \pi^{t_1}_{\underline{I}(T_1)} v\left(q_{\underline{I}(T_1),t_1}\right) + \sum_{s \notin T_2} \pi^j_M v\left(q_{M,j}\right) + \pi^{t_2}_{\underline{I}(T_2)} v\left(q_{\underline{I}(T_2),t_2}\right) = \\
& = V(I) + \sum_{s \notin T_1 \cup T_2} \pi^s_I v\left(q_{I,s}\right) + \pi^{t_1}_{\underline{I}(T_1)} v\left(q_{\underline{I}(T_1),t_1}\right) + \pi^{t_2}_{\underline{I}(T_2)} v\left(q_{\underline{I}(T_2),t_2}\right) = \\
& = V(I) + V(\underline{I}(T_1, T_2)).
\end{aligned}
$$

So V is IIS.

Conversely, assume that V is IIS and reducible. We need to prove the existence of a function $v : \Delta(\Omega) \to \mathbb{R}$ such that $V(I) = \sum_{s \in S} \pi^s_I v(q_{I,s})$ for every $I \in \mathcal{I}$.

In order to find an appropriate function v, we first need to define two auxiliary information structures for every vector $x = (x_1, \ldots, x_n)$ with $0 \le x_i \le 1,\ i = 1, 2, \ldots, n$. The first one is denoted $B^{x,1}$ and has $n+1$ signals. The stochastic matrix is

$$
\begin{pmatrix}
x_1 & 0 & 0 & \ldots & 1 - x_1 \\
0 & x_2 & 0 & \ldots & 1 - x_2 \\
\vdots & \vdots & \ddots & & \vdots \\
0 & 0 & 0 & x_n & 1 - x_n
\end{pmatrix}
$$

The second information structure is denoted $B^{x,2}$ and has only 2 signals. The distribution over signals is defined by the matrix

$$
\begin{pmatrix}
x_1 & 1 - x_1 \\
x_2 & 1 - x_2 \\
\vdots & \vdots \\
x_n & 1 - x_n
\end{pmatrix}
$$

Finally, let I_d denote the deterministic information structure with n signals, under which the DM is fully informed about the true state of nature.

Notice that, since the prior distribution on Ω is μ, $q_{I,s}$ is always of the form[1] $q_{I,s} = \frac{\mu \circ x}{\|\mu \circ x\|_1}$, where x is a vector as above. We define $v\left(\frac{\mu \circ x}{\|\mu \circ x\|_1}\right) =$

---

[1] For any two vectors $x = (x_1, ..., x_n)$ and $y = (y_1, ..., y_n)$, $x \circ y$ denotes the vector of the same length whose i-th coordinate is equal to $x_i y_i$.

$V(I_d) - \frac{V(B^{x,1})-V(B^{x,2})}{\|\mu \circ x\|_1}$. To finish the proof, it only remains to check that if v is defined as above then $V(I) = \sum_{s\in S} \pi_I^s v(q_{I,s})$ for every $I \in \mathcal{I}$. Indeed, using IIS we have for any $I = (S, M)$ (recall Notation 1(b)),

$$\begin{aligned}
\sum_{s\in S} \pi_I^s v(q_{I,s}) &= \sum_{s\in S} \pi_I^s \left(V(I_d) - \frac{V(B^{M_{\cdot s},1}) - V(B^{M_{\cdot s},2})}{\|\mu \circ M_{\cdot s}\|_1} \right) = \\
&= V(I_d) - \sum_{s\in S} \left(V(B^{M_{\cdot s},1}) - V(B^{M_{\cdot s},2}) \right) = \\
&= V(I_d) - \sum_{s\in S} \left(V(\overline{I}(\{s\})) - V(I) \right).
\end{aligned}$$

Repeated use of IIS gives $\sum_{s\in S} \left(V(\overline{I}(\{s\})) - V(I) \right) = V(\overline{I}(S)) - V(I)$. Therefore, $\sum_{s\in S} \pi_I^s v(q_{I,s}) = V(I_d) - V(\overline{I}(S)) + V(I)$. It only remains to check that $V(I_d) = V(\overline{I}(S))$. However, since V is reducible we are done. ■

We conclude this section with a short discussion on the uniqueness of the function v. A function $v : \Delta(\Omega) \to \mathbb{R}$ uniquely determines a function $V : \mathcal{I} \to \mathbb{R}$ via the equation $V(I) = \sum_{s\in S} \pi_I^s v(q_{I,s})$. However, given some additively separable function V, the corresponding v is not unique. The following proposition states that v_1 and v_2 both correspond to V iff $v_1 - v_2$ is a linear function which vanish at the prior distribution μ.

Proposition 2. *Assume that $V : \mathcal{I} \to \mathbb{R}$ is additively separable with v_1 corresponding to it. Then, v_2 also corresponds to V, if and only if there exists a vector $x \in \mathbb{R}^n$ such that*[2] *$x\mu = 0$ and $v_1(q) - v_2(q) = xq$ for every $q \in \Delta(\Omega)$.*

Proof. Assume that for a certain $x \in \mathbb{R}^n$ and for every $q \in \Delta(\Omega)$, $v_2(q) = v_1(q) - xq$. Moreover, assume that $x\mu = 0$. We show that v_2 corresponds to V.

---

[2] For any two vectors $x = (x_1, ..., x_n)$ and $y = (y_1, ..., y_n)$, xy denotes the inner product $\sum x_i y_i$.

For every $I \in \mathcal{I}$,

$$\begin{aligned}\sum_{s\in S}\pi_I^s v_2(q_{I,s}) &= \sum_{s\in S}\pi_I^s(v_1(q_{I,s})-xq_{I,s}) = V(I)-\sum_{s\in S}\pi_I^s x q_{I,s} = \\ &= V(I) - \sum_{s\in S}\pi_I^s\sum_{i=1}^n x^i \frac{\mu(\omega_i)M_{is}}{\pi_I^s} = \\ &= V(I)-\sum_{i=1}^n x^i\mu(\omega_i)\sum_{s\in S}M_{is} = V(I)-x\mu = V(I).\end{aligned}$$

Therefore, v_2 also corresponds to V.

In the other direction, assume that v_2 corresponds to V and define $v = v_1 - v_2$. Let $q_j = (q_j^1, ..., q_j^n)$, $j = 1, 2$, be two distributions over Ω, and let $\alpha \in [0, 1]$. For $i = 1, ..., n$, define $r_1^i = c\alpha\frac{q_1^i}{\mu(\omega_i)}$ and $r_2^i = c(1-\alpha)\frac{q_2^i}{\mu(\omega_i)}$, where c is a positive constant that satisfies $r_1^i + r_2^i \leq 1$ for every $i = 1, ..., n$. Finally, define $r_3 = (1, ..., 1) - r_1 - r_2$. Consider the information structure $I = (S, M)$, where $S = \{s_1, s_2, s_3\}$, and M is the $n \times 3$ matrix whose j'th column is r_j, $j = 1, 2, 3$. Since both, v_1 and v_2, correspond to V, we have for $j = 1, 2$, $V(I) = \sum_{s\in S}\pi_I^s v_j(q_{I,s}) = c\alpha v_j(q_1) + c(1-\alpha)v_j(q_2) + (1-c)v_j(q_3)$, where $q_3 = \frac{r_3 \circ \mu}{\|r_3 \circ \mu\|_1}$. Reorganizing the terms yields,

$$c\alpha v(q_1) + c(1-\alpha)v(q_2) = (c-1)v(q_3). \tag{2}$$

Set $T = \{s_1, s_2\}$. For $j = 1, 2$ we obtain, $V(\underline{I}(T)) = cv_j(\alpha q_1 + (1-\alpha)q_2) + (1-c)v_j(q_3)$, which is equivalent to

$$cv(\alpha q_1 + (1-\alpha)q_2) = (c-1)v(q_3). \tag{3}$$

¿From (2) and (3) it follows that $\alpha v(q_1) + (1-\alpha)v(q_2) = v(\alpha q_1 + (1-\alpha)q_2)$ for every two distributions q_1, q_2 and for every $\alpha \in [0, 1]$. In other words, $v = v_1 - v_2$ is linear on $\Delta(\Omega)$. Thus, there is $x \in \mathbb{R}^n$ such that $v_1(q) - v_2(q) = xq$. Finally, since v_1 and v_2 agree on the information structure with only one signal, $x\mu = v_1(\mu) - v_2(\mu) = 0$. ■

4. Characterization of information functions

Definition 5. *A function $V : \mathcal{I} \to \mathbb{R}$ is* convex *if $V(\underline{I}(T)) \leq V(I)$ for every information structure $I = (S, M)$ and for every subset of signals $T \subseteq S$.*

Remark 1. Blackwell (1951, 1953) defined the *more informative* partial order over information structures. Let $I = (S, M)$ and $I' = (S', M')$ be two information structures. I is *more informative* than I', if there is a stochastic matrix,[3] say C, such that $M' = MC$. That is, M' can be obtained by multiplying M with a stochastic matrix. We say that a function V defined over $\mathcal{I}$ has the *Blackwell property*, if $V(I) \geq V(I')$, whenever I is more informative than I'. The information structure I is more informative than $\underline{I}(T)$, and therefore, if V has the Blackwell property, then $V(I) \geq V(\underline{I}(T))$. That is, if V has the Blackwell property, then it is convex. Blackwell (1951, 1953) showed that any information function has the Blackwell property and is, therefore, convex.■

We conclude that any information function is additively separable and convex. In order to show that these conditions are also sufficient, we first need to prove the following lemma which relates the convexity of V with the convexity of v that corresponds to it.

Lemma 1. *Let $V : \mathcal{I} \to \mathbb{R}$ be additively separable function and v corresponds to V. Then, V is convex if and only if v is convex on $\Delta(\Omega)$.*

Proof. Let $q_j = (q_j^1, ..., q_j^n),\ j = 1, 2$ be two distributions over Ω, and let $\alpha \in [0, 1]$. We start by showing that if V is convex then $\alpha v(q_1) + (1 - \alpha)v(q_2) \geq v(\alpha q_1 + (1 - \alpha)q_2)$.

For $i = 1, ..., n$, define $r_1^i = c\alpha \frac{q_1^i}{\mu(\omega_i)}$ and $r_2^i = c(1 - \alpha)\frac{q_2^i}{\mu(\omega_i)}$, where c is a positive constant that satisfies $r_1^i + r_2^i \leq 1$ for every $i = 1, ..., n$. Let $r_3 = (1, ..., 1) - r_1 - r_2$ and consider the information structure

---

[3] A stochastic matrix is a matrix whose entries are all non-negative and the sum of each row is 1.

$I = (S, M)$, where $S = \{s_1, s_2, s_3\}$, and M is an $n \times 3$ matrix whose j-th column is r_j, $j = 1, 2, 3$.

Note that $V(I) = \sum_{s \in S} \pi_I^s v(q_{I,s}) = c\alpha v(q_1) + c(1-\alpha)v(q_2) + (1-c)v(q_3)$, where $q_3 = \frac{r_3 \circ \mu}{\|r_3 \circ \mu\|_1}$. If $T = \{s_1, s_2\}$ then $V(\underline{I}(T)) = cv(\alpha q_1 + (1-\alpha)q_2) + (1-c)v(q_3)$. Since V is convex, $V(I) \geq V(\underline{I}(T))$. Thus, $c\alpha v(q_1) + c(1-\alpha)v(q_2) + (1-c)v(q_3) \geq cv(\alpha q_1 + (1-\alpha)q_2) + (1-c)v(q_3)$, which implies that $\alpha v(q_1) + (1-\alpha)v(q_2) \geq v(\alpha q_1 + (1-\alpha)q_2)$.

In the other direction, assume that v is convex and let $T \subseteq S$ be a subset of signals of some information structure $I = (S, M)$. Recall that the set of signals in the information structure $\underline{I}(T)$ is $(S \setminus T) \cup \{t\}$ and the column corresponding to the signal t is the sum of columns of the signals in T. By convexity of v we obtain

$$\begin{aligned} V(\underline{I}(T)) &= \sum_{s \in S \setminus T} \pi_I^s v(q_{I,s}) + \pi^t v(q_t) \\ &\leq \sum_{s \in S \setminus T} \pi_I^s v(q_{I,s}) + \pi^t \sum_{s \in T} \frac{\pi_I^s}{\pi^t} v(q_{I,s}) = V(I). \end{aligned}$$

Therefore, V is convex. ■

Proposition 3. *If $V : \mathcal{I} \to \mathbb{R}$ is additively separable and convex, then it is an information function.*

Proof. Let V be an additively separable and convex function, and let v correspond to V. By Lemma 1, v is convex on $\Delta(\Omega)$. Therefore, at every point r in the simplex, there is a vector $x_r = (x_r^1, ..., x_r^n)$ (that defines the tangent to the graph of v at the point r) such that $v(q) \geq qx_r$ for every q with equality when $q = r$. In particular $v(q) = qx_q = \max_r qx_r$.

Define the set of actions, A, to be identical to the simplex. When the state realized is ω_i and the action taken is r, the utility, $u(r, \omega_i)$, is defined to be x_r^i. Thus, when the distribution over states is $q = (q^1, ..., q^n)$ and the action taken is r, the expected utility is $\sum_i q^i u(r, \omega_i) = qx_r$. Hence, when the posterior distribution over states is q, the optimal action is q and the expected utility is $qx_q = v(q)$. ■

Remark 2. The action set A defined in the previous proof is the simplex of distributions over Ω, which is a compact set. Since v is convex, it is almost surely differentiable. It implies that qx_r, as a function of r, is almost surely continuous. (Recall that x_r defines a tangent to the graph of v at r. Thus, as a function of r, it may have discontinuity at the kink points of v.) Therefore, the utility function u defined on $A \times \Omega$ is almost surely continuous. $\blacksquare$

We therefore proved the following theorem:

Theorem 1. *$V : \mathcal{I} \to \mathbb{R}$ is an information function if and only if it is additively separable and convex.*

Remark 3. (i) By Lemma 1, Theorem 1 can be rephrased as follows: *$V : \mathcal{I} \to \mathbb{R}$ is an information function if and only if it is additively separable and every function v that corresponds to it is convex.*
(ii) As indicated in Remark 1, if a function V has the Blackwell property, then it is convex. Therefore, Theorem 1 can be rephrased as follows: *V is an information function if and only if it is additively separable and has the Blackwell property.*$\blacksquare$

Remark 4. Theorem 1 refers to the case where the state space, Ω, and the prior distribution μ are known to the outside observer. It characterizes those observations that are consistent with a rational behavior of a decision maker in a Bayesian setting, given that the prior is μ. One may ask a similar question for an unknown prior. That is, when are the observations consistent with a rational behavior of a decision maker with *some* prior μ?

Regarding Ω, if we assume that the outside observer can see the information structure, then we implicitly assume that he can also see the state space (or at least its cardinality). However, the observer need not know the prior beliefs of the DM.

The properties *IIS* and *convexity* of an information function V do not depend on the prior distribution μ. The *reducibility* condition, although phrased in terms of the posteriors $q_{I,s}$, could be rephrased without resorting to any particular prior: for any pair of signals $s_1, s_2 \in S$ whose corresponding columns are proportional, $V(\underline{I}(\{s_1, s_2\})) = V(I)$. This implies that if V has these three properties, then for any distribution μ, as long as it has a full support (all the states are assigned positive probability), V is an information function of a decision problem with μ being its prior. In other words, V is an information function with *a certain* prior having full support if and only if it is an information function with *any* prior having full support. It should be noted, however, that the corresponding v might change with the prior.■

5. FINITE ACTION SET

Theorem 1 ensures the existence of an action set, typically an infinite one. The question arises as to when a function is an information function of a decision problem having a finite set of actions.

Definition 6. *A real function v defined on the simplex is* piecewise-linear *if there are finitely many disjoint sets $W_1, ..., W_k$ in the simplex, such that*[4] $\cup_{i=1}^{k} \mathrm{cl} W_i$ *covers the entire simplex and v is linear over W_i, $i = 1, ..., k$.*

Theorem 2. *A function $V : \mathcal{I} \to \mathbb{R}$ is an information function of a decision problem with finitely many actions if and only if it is additively separable and any v corresponding to it is a piecewise-linear and convex function.*

Proof. Let A be a finite action set. For a fixed $a \in A$, the function $u(a, q) = \sum_i q^i u(a, \omega_i)$ is a linear function of q. Furthermore, $v(q) = \max_{a \in A} u(a, q)$. Thus, v is a maximum of finitely many linear functions, and is, therefore, piecewise-linear and convex. It follows, that if v corresponds to V, where the decision problem has a finite actions set, then v must be piecewise-linear and convex.

---

[4] $\mathrm{cl} W$ is the closure of W.

Conversely, if v is a piecewise-linear and convex, then v is the maximum of finitely many linear functions. As in the proof of Theorem 1, by identifying the set of actions with this finite set of linear functions, one can show that V is an information function. ■

6. Measuring information by entropy

The entropy function that measures the contents of information plays a major role in various fields such as information theory, Ergodic theory and probability. The entropy of the distribution q, denoted $e(q)$, is defined as $\sum_i -q_i \log(q_i)$. This function is concave and attains its maximum where the uncertainty is maximal: at the uniform distribution. Using this function one can define the entropy of an information structure $I = (S, M)$ as,

$$e(I) = \sum_{s \in S} \pi_I^s e(q_{I,s}). \tag{4}$$

$e(I)$ is the weighted average of the entropy of all the posteriors. The greater the entropy the greater the uncertainty about the state. Like the entropy of distributions, the entropy over information structures attains its maximum when the structure is perfectly non-informative; that is, when the DM obtains no information about the state realized.

It turns out that the entropy measurement does not entirely capture the essence of being more informative. As shown in the following example, the entropy of the information structure I may be smaller than that of I', and yet, I is *not* more informative than I'. It should be emphasized though that due to the concavity of the entropy, the converse is true: If I is more informative than I', then its entropy is smaller or equal to that of I'.

Example 1: Let Ω be consisting of two equally likely states, $S = \{a, b, c\}$ and $S' = \{a, b\}$. Set

$$M = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{3} & \frac{2}{3} & 0 \end{pmatrix} \text{ and } M' = \begin{pmatrix} \frac{1}{4} & \frac{3}{4} \\ \frac{3}{4} & \frac{1}{4} \end{pmatrix}.$$

Define $I = (S, M)$ and $I' = (S', M')$. $e(I) = -\frac{1}{4}\log(\frac{3}{7}) - \frac{1}{3}\log(\frac{4}{7})$ and $e(I') = -\frac{1}{4}\log(\frac{1}{4}) - \frac{3}{4}\log(\frac{3}{4})$. Thus, $e(I) < e(I')$ and therefore I contains more information than I' with respect to the entropy measurement. However, there is no stochastic matrix C that satisfies $M' = MC$. We conclude that I is not more informative than I' (recall Remark 1), although the entropy of I is smaller than the entropy of I'.

■

Theorem 1 refers to convexity and it is therefore more convenient in this context to refer to the convex function $-e$ (minus entropy) rather than to the entropy itself. The convexity of $-e$ (which implies that if I is more informative than I', then $-e(I) \geq -e(I')$) is its only important property[5]. Thus, in order to measure the informational contents of a distribution one could replace $-e$ by any other convex function defined over distributions, say w. The domain of w can be extended to information structures in a similar way it was done in (4). This extension, like $-e$, has the property that if I is more informative than I', then $w(I)$ is greater than or equal to $w(I')$.

A convex function w induces a complete order over information structures in the following manner:

Definition 7. *Let w be a convex function defined over $\Delta(\Omega)$ and let I and I' be two information structures. I* contains at least as much information *as I'* with respect to w*, if $w(I) \geq w(I')$.*

Theorem 1 implies that the partial order *'more informative'* as defined by Blackwell (1951, 1953) is the intersection of the complete orders *'contains at least as much information with respect to w*, when the intersection is taken over all convex functions w. Formally,

---

[5] The entropy plays a significant role in other fields due to its concavity and to other properties which are irrelevant to the subject of information structures.

Corollary 1. *I is more informative than I' if and only if for any convex function w, I contains at least as much information as I' with respect to w.*

7. References

Afriat, S. (1967) "The Construction of a Utility Function from Expenditure Data," *International Economic Review,* **8**, 67-77.

Blackwell, D. (1951) "Comparison of Experiments," *Proc. Second Berkeley Symposium on Math. Stat and Prob.*, University of California Press, 93-102.

Blackwell, D. (1953) "Equivalent Comparison of Experiments," *Ann. Math. Stat.*, **24**, 265-272.

Gilboa, I. and E. Lehrer (1991) "The value of information – An axiomatic approach," *J. Math. Economics*, **20**, 443-459.