ENTROPY AS RELEVANCE:

MAXIMIZING ENTROPY AND MINIMIZING LOSS

Sophia Alexandra Malamud

A THESIS

in

Mathematics

Presented to the Faculties of the University of Pennsylvania

in Partial Fulfillment of the Requirements for the Degree of Master of Arts

2006

Thesis Committee: Murray Gerstenhaber (Supervisor), Department of Mathematics

Dean P. Foster, Department of Statistics

Acknowledgements

I want to dedicate this thesis to my mother, Alexandra, my father, Arkadiy, and my grandmother, Zoya – the people who taught me, by example, that math is fun, and that I can learn to have fun with it. Thank you for that, and for all your support!

I owe a debt of gratitude to Irina Markovna Zonn, who in 1991-1993 showed me how (enormously) much math can be taught to very young and eager students by a very talented teacher.

I'm grateful to my teachers at the UPenn Mathematics Department, who taught me much math during my undergraduate and graduate education. I'm also grateful to Harry Tamvakis, and to Scott Weinstein, who at different times taught me more math than I thought I could possibly learn, and made me work harder than ever before.

I also want to thank Tony Kroch, Robin Clark, and Maribel Romero at the UPenn Linguistics Department, who were so enthusiastic about the idea that I should get an M.A. in mathematics (and about this Decision Theory project in particular), that they twisted my arm when I wavered.

Teddy Seidenfeld of Carnegie Mellon University provided loads of information and inspiration during my visit there.

Last, but very far from least, I am very grateful to my committee members – Murray Gerstenhaber, for constant encouragement and lots of patience, and Dean Foster, for explaining those bits and pieces that I just couldn't get on my own.

Table of Contents

1 Shannon entropy and information theory

1.1 Background

The concept of entropy and the definition of the information content of a message was developed independently by Shannon (Shannon 1948) and Alan Turing (during his work at Bletchley Park in WWII), but came to be associated with Shannon's name.

If the probability of a certain message m_i is $p(m_i)$, then the information content of a message is defined as follows (cf. e.g. Cover and Thomas 1991):

Definition 1.1 ***Information content*** of message m_i is $I(m_i) = - \log p(m_i)$ (1)

The negative log function makes $I(m_i)$ a positive function, giving greater values to less probable (less predictable, more surprising) messages.

Shannon entropy (which Turing called weight of evidence) is then defined as the mean information content per message (cf. e.g. Cover and Thomas 1991):

Definition 1.2 ***Shannon entropy*** of a set of messages $\{m_i\}$ is

$$H = \sum_i p(m_i) I(m_i) = - \sum_i p(m_i) \log p(m_i) \quad (2)$$

If we replace the message m with a value x for a random variable X, then the entropy of X is the measure of uncertainty inherent in X. With P the distribution of X, and $p(x)$ the corresponding mass density function, we can write (cf. e.g. Cover and Thomas 1991)

$$H(X) = H(P) = \sum_{x \in X} p(x) \log p(x) \qquad (3)$$

The entropy of X is thus an expected value for $\log (1/p(X))$, with values of X are drawn according to the probability mass function $p(x)$. So, another notation for the value of entropy of X is $E_P\{-\log p(X)\}$. It describes how much information, on average, is necessary to describe the distribution of the random variable (cf. e.g. Cover and Thomas 1991). The entropy is usually measured in bits, and the *log* is then taken to base *2*.

Example 1.1 The entropy of a fair coin toss is *1* bit.

Definition 1.3 ***Relative entropy*** (or ***Kullback Leibler distance*** or ***I-divergence***) between two probability mass functions $p(x)$ and $q(x)$ is defined as

$$D(p \| q) = \sum_{x \in \chi} p(x) \log \left(p(x)/q(x) \right) \qquad (4)$$

(cf. e.g. Topsøe 1979, Cover and Thomas 1991, Berger 1985)

Relative entropy is not really a metric, but it shares some nice properties of metrics. We list some of them now:

Theorem 1.1 Properties of I-divergence: (cf. e.g. Khudanpur 1999, references therein)

Non-negativity. $D(p \| q)$ is always non-negative, and is zero if and only if $p = q$.

Lower semicontinuity. For a sequence of probability mass functions (p_n, q_n), $n = 1,2,\ldots$, which converges to (p,q),

$$\lim_{n\to\infty} \inf D(p_n \| q_n) \geq D(p \| q) \tag{5}$$

If $Q(x)>0$ for each x, then $D(p \| q)$ is continuous in the pair (p,q)

Convexity. For any $a \in [0,1]$, and distributions p_1, q_1, p_2, q_2,

$$aD(p_1 \| q_1)+(1-a)\, D(p_2 \| q_2) \geq D(ap_1+(1-a)p_2 \| aq_1+(1-a)q_2) \tag{6}$$

Partition inequality. Suppose $\boldsymbol{A} = \{A_1, \ldots, A_K\}$ is a partition of the space χ, that is $\chi = \cup_{i=1}^{K} A_i$, and for all $i \neq j$ $A_i \cap A_j = \emptyset$.

Suppose also that

$$p_{\mathbf{A}}(i) = \sum_{x\in A_i} p(x), \qquad i = 1, \ldots, K,$$

$$q_{\mathbf{A}}(i) = \sum_{x\in A_i} q(x), \qquad i = 1, \ldots, K.$$

Then $D(p \| q) \geq D(p_{\mathbf{A}} \| q_{\mathbf{A}})$, with equality iff $P(x \mid x\in A_i) = Q(x \mid x \in A_i)$, for each i. (7)

Pinsker's inequality. The variational distance between distributions,

$d(p,q) = \sum_{x\in\chi} |p(x)-q(x)|$, is bounded above by the I-divergence:

$$D(p \| q) \geq \tfrac{1}{2}\, d^2(p,q) \tag{8}$$

Information inequality. Let $p(x)$, $q(x)$, $x \in X$, be two probability mass functions. Then

$D(p \| q) \geq 0$ with equality if and only if $p(x) = q(x)$ for all x. (9)

Before we can prove these properties, we need a lemma: (cf. e.g. Khudanpur 1999)

Lemma 1.1 ***Log-Sum Inequality.*** Let $\{a_i\}^n_{i=1}$ and $\{b_i\}^n_{i=1}$ be sequences of nonnegative numbers. Let $a=\sum^n_{i=1} a_i$ and $b=\sum^n_{i=1} b_i$. Then

$$\sum^n_{i=1} a_i \log a_i/b_i \geq a \log a/b \qquad (10)$$

With equality iff $a_i/b_i = c$ for every i, where c is some constant

Proof. It suffices to prove this for $a_i > 0$, since when we drop from (10) the a_i and b_i for which $a_i=0$, this doesn't change the left side of the inequality and can only increase the right side by (maybe) reducing b. Also, it suffices to prove (10) for $b_i>0$; otherwise, the left side is infinity and there is nothing to prove. Finally, it suffices to prove this for $a=b$. The inequality is invariant to a scaling of the b_i's because

$\sum^n_{i=1} a_i \log a_i/kb_i = \sum^n_{i=1} a_i \log a_i/b_i + a_i \log 1/k,$ and

$a \log a/\sum kb_i = a \log a/b + a \log 1/k.$

Recall that $\log t \leq t-1$ for positive t, with equality iff $t=1$. Thus, setting $t_i = b_i/a_i$, we get

$$\sum^n_{i=1} a_i \log a_i/b_i = -\sum^n_{i=1} a_i \log t_i \geq -\sum^n_{i=1} a_i (t_i - 1) = -\sum^n_{i=1} (b_i - a_i)$$

This proves the desired inequality, together with condition for equality.

Proof of Theorem 1.1[1]

Non-negativity: Nonnegativity of $D(p\|q)$ follows trivially from the log-sum inequality, as does the condition for equality.

Lower semicontinuity: Consider the following three cases.

Case 1. For each x, $q(x) > 0$.

[1] Except where indicated, the proof of this theorem follows the one given in Khudanpur (1999).

In this case, since $q_n(x) \rightarrow q(x)$, $q_n(x) > 0$ for all $n > n_0$ for some $n_0(x)$. Then, $p_n(x) \log p_n(x)/q_n(x) \rightarrow p(x) \log p(x)/q(x) < +\infty$,

And since $D(p_n \| q_n)$ is a sum of a finite number of these terms, each of which converges to a finite limit, then $D(p_n \| q_n) \rightarrow D(p \| q)$.

Case 2. There is an x' such that $q(x') = 0$ and $p(x') > 0$.

In this case, $D(p \| q) = +\infty$, and

$D(p_n \| q_n) =$

$$= p_n(x') \log p_n(x')/q_n(x') + \sum_{x \neq x'} p_n(x) \log p_n(x)/q_n(x) =$$

$$= \begin{cases} p_n(x') \log p_n(x')/q_n(x') + 0 & \text{if } p_n(x')=1 \\ p_n(x') \log p_n(x')/q_n(x') + \infty & \text{if } p_n(x')<1 \text{ and } q_n(x')=1 \\ p_n(x') \log p_n(x')/q_n(x') + & \text{otherwise} \\ \quad + \sum_{x \neq x'} p_n(x) \left[\log \frac{p_n(x)}{1- p_n(x)} \Big/ \frac{q_n(x)}{1- q_n(x)} + \log \frac{1- p_n(x)}{1- q_n(x)}\right] & \end{cases}$$

Since $q_n(x') \rightarrow 0$, we can verify that the first term above goes to infinity, and in all three cases, the second term does not go to negative infinity.

So, $D(p_n \| q_n) \rightarrow +\infty$

Case 3. Every time when $q(x') = 0$, $p(x') = 0$ as well.

In this case $D(p_n \| q_n) =$

$$= \sum_{x,\ s.t.\ q(x)>0} p_n(x) \log p_n(x)/q_n(x) + \sum_{x',\ s.t.\ q(x')=0} p_n(x') \log p_n(x')/q_n(x')$$

$$= \sum_{x,\ s.t.\ q(x)>0} p_n(x) \log p_n(x)/q_n(x) + \sum_{x',\ s.t.\ q(x')=0} p_n(x') \log p_n(x') - p_n(x') \log q_n(x')$$

$\geq \sum_{x,\ s.t.\ q(x)>0} p_n(x) \log p_n(x)/q_n(x) + \sum_{x',\ s.t.\ q(x')=0} p_n(x') \log p_n(x')$

Again, we can check that the first term goes to $D(p_n \| q_n)$, and the second term goes to zero. So, $lim_{n \to \infty} inf\ D(p_n \| q_n) \geq D(p \| q)$

Convexity: Since $ap_1(x) \log p_1(x)/q_1(x) + (1-a)p_2(x) \log p_2(x)/q_2(x) =$

$= ap_1(x) \log ap_1(x)/aq_1(x) + (1-a)p_2(x) \log (1-a)p_2(x)/(1-a)q_2(x) \geq$

$\geq [ap_1(x) + (1-a)p_2(x)] \log [ap_1(x) + (1-a)p_2(x)]/[aq_1(x) + (1-a)q_2(x)]$

Then from log-sum inequality we get that

$aD(p_1 \| q_1)+(1-a)\ D(p_2 \| q_2) \geq D(ap_1 + (1-a)p_2 \| aq_1 + (1-a)q_2)$.

Partition inequality: For a partition $\boldsymbol{A} = \{A_1, \ldots, A_K\}$,

$D(p \| q) = \sum_{x \in \chi} p(x) \log (p(x)/q(x)) = \sum_{i=1}^{K} \sum_{x \in A_i} p(x) \log (p(x)/q(x)) \geq$

$\geq \sum_{i=1}^{K} [\sum_{x \in A_i} p(x)] \log (\sum_{x \in A_i} p(x) / \sum_{x \in A_i} q(x)) = \sum_{i=1}^{K} p_{\mathbf{A}}(i) \log p_{\mathbf{A}}(i)/q_{\mathbf{A}}(i) = D(p_{\mathbf{A}} \| q_{\mathbf{A}})$

Since the log-sum inequality was used for every i, then for this to be an equality, we must have, for every i, that $p(x)/q(x)=c_i$, so that $\sum_{x \in A_i} p(x) = c_i \sum_{x \in A_i} q(x)$, equivalently

$p(x \in A_i) / q(x \in A_i) = c_i$

This gives us the conditions for equality.

Pinsker's inequality: Let $\boldsymbol{A}=\{A_1, A_2\}$, where $A_1 = \{x \mid p(x) \geq q(x)\}$ and $A_2 = \{x \mid p(x) < q(x)\}$

Then, $d(p,q) = \sum_{x \in \chi} |p(x)-q(x)| = \sum_{x \in A_1} (p(x)-q(x)) - \sum_{x \in A_2} (p(x)-q(x)) =$

$= (p_{\mathbf{A}}(1)-q_{\mathbf{A}}(1)) + (q_{\mathbf{A}}(2)-p_{\mathbf{A}}(2)) = d(p_{\mathbf{A}}, q_{\mathbf{A}})$

Then, from the partition inequality it follows that $D(p\|q) \geq D(p_A\|q_A)$, so we only have to prove Pinsker's inequality for $|\chi| = 2$.

For $\chi = \{0,1\}$, let $p = (r,1-r)$ and $q=(s,1-s)$, and

let $f(s) = r \log r/s + (1-r)\log [1-r]/[1-s] - 4c(r-s)^2$, treating r and c as constants.

Then, $f(r) = 0$, and for $s \neq 0$ and $s \neq 1$,

the partial derivative $f_c'(s) = -r/s + [1-r]/[1-s] + 8c(r-s) = (s-r)\,[\,1/(1-s)s - 8c]$

Since r and s are probabilities, $s(1-s) \leq \frac{1}{4}$ and $f(s)$ achieves a minimum at $s = r$ when the constant $c \leq \frac{1}{2}$. Thus, for $c \leq \frac{1}{2}$ we have

$$f_c(s) = D(p\|q) - c(|r-s| + |(1-r) - (1-s)|)^2 = D(p\|q) - cd^2(p,q) \geq 0.$$

When $c = \frac{1}{2}$, we get Pinsker's inequality.

Information inequality: (proof as in Cover and Thomas 1991) We can rewrite the information inequality as follows:

$\sum_{x \in X} p(x) \log (p(x)/q(x)) \geq 0$	Starting with (9), we use
$\sum_{x \in X} p(x) [\log p(x) - \log q(x)] \geq 0$	$\log(a/b)=\log a - \log b$
$\sum_{x \in X} [p(x) \log p(x) - p(x) \log q(x)] \geq 0$	
$\sum_{x \in X} [p(x) \log p(x)] - \sum_{x \in X} [p(x) \log q(x)] \geq 0$	
$-\sum_{x \in X} [p\,(x) \log q(x)] \geq -\sum_{x \in X} [p(x) \log p(x)]$	Note that the right-hand side
$E_P \{-\log q(x)\} \geq E_P \{-\log p(x)\}$	is just the entropy $H(P)$.

Thus, another way to write the information inequality is (11)

$$E_P\{-\log q(x)\} \geq E_P\{-\log p(x)\} = H(P) \quad (11)$$

Shannon proceeded to develop his theory of information with entropy as the natural measure of efficient minimal description length, formulating the coding theorems that help calculate the capacity of a message channel, and in general provide theoretical foundations for designing telecommunication systems. We will return to code length in section 4.2 below.

1.2 The problem of maximizing entropy

A particular, and important, problem in Information Theory is that of maximizing the informational content per message, that is, maximizing entropy.

If P assigns probability 1 to a particular value x_0 in χ, and 0 to all other values, then there is no uncertainty at all – the entropy of such P is null ($0\log 0$ is taken to be 0 by convention here). On the other hand, if all values of χ are equally probable, the entropy (and the uncertainty) is highest possible (cf. e.g. Berger 1985).

$$H(P) = -\sum_n p(x_i) \log p(x_i) = -(1/n)(\log 1/n) = (\log n)/n \quad (12)$$

A detour: why maximize entropy?

To provide a motivation for maximizing entropy of probability distributions, consider a simple (and simplistic) version of a general problem.

Example 1.2 (Blum and Kalai 1999, pointed out by Dean Foster, personal communication). The decision maker Dina is a student of linguistics, (13)

who has two friends who are investment specialists.

One of them, Wendy, is from Wharton Business School;

the other, Harry, is from Harvard Business School.

Since Dina knows nothing about finance, she is unable to judge which of her two friends will do better in terms of long-term accumulation of wealth. However, she wants her own means to grow at about the same rate as the wealth of the better of the two. What should she do?

Some non-solutions. Some strategies that won't achieve Dina's goal of imitating the rate of growth of the better of her two friends:

1. Dina should do whatever Wendy does initially, then, depending on who does better after some time, switch to the better person.

   This approach is just about the worst way to do it, since Dina will be missing the good bets each friend makes, and switching to him or her after a period of success, and not during that period. This way, Dina will do worse than either of them.

   The flaw of this approach is its excessive adaptability to change. So, maybe adapting should be abandoned altogether:

2. Dina should take a wild guess and stick with either Wendy or Harry.

Of course, Dina is not guaranteed from any disasters, so if she picks Wendy and Wendy has a negative growth rate (i.e., she keeps losing money), while Harry grows 10% a year, Dina is very far from her objective.

So, maybe she should guarantee herself some balance between the two friends

3. With half her money Dina should do whatever Wendy does, and follow Harry with the other half.

This approach is better than the previous two: so, if Wendy grows 10% a year, and Harry loses in value 10% a year, Dina will be safely at 0%. This, however, doesn't do much to approximate Wendy's 10%.

Solution. The best way to approximate the growth rate of the more successful friend is to take all the wealth Dina wants to invest, and give half of it to Wendy, and half to Harry, for them to include in their own portfolios.

This is distinct from the non-solution #3, since the worst Dina can do is lose what she has invested – Dina can never lose more than she had. So, suppose after the term of the investment, Harry does badly and loses half Dina's wealth, while Wendy does better. Her average growth rate is $\log w/t$. Then, Dina's average growth rate is $\log(w/2)/t = \log w/t - \log 2/t$. So, Dina's growth rate is the same as Wendy's, minus a very small correction factor $\log 2/t$.

Generalization. Generalizing this problem to any number of friends, we can interpret the fraction of Dina's wealth that she gives to each friend as a probability – the subjective probability she assigns to this friend's chances of success. The best way to distribute this wealth, in the case when Dina is completely uncertain about her friends' chances, is to distribute it as widely as possible, maximally avoiding concentration of wealth in any one

friend's hands. If we interpret the fraction of Dina's wealth give to each friend x from the set of friends χ as a probability $p(x)$, then, to achieve a close approximation of the most successful friend, Dina has to maximize the entropy over χ.

This method of limiting one's losses and maximizing growth is used, e.g., in actual financial management by funds that invest in other managed funds.

In this particular case, when the distribution is not constrained in any further way, the maximal-entropy distribution is the uniform one: give each friend an equal proportion of money.

More generally, when decisions have to be made based on a probability distribution that is not completely known, it is desirable to select a distribution which is as unbiased (and thus as uncertain about outcomes) as possible, subject to the constraints imposed by what is known (Berger 1985).

Thus, maximizing the entropy of a distribution, subject to some constraints, is an important problem of statistical decision theory.

Now that we have some reasons to maximize entropy, let us flesh out the problem of maximization.

Suppose a class Γ of probability distributions P over the finite sample space χ is given.

We wish to select a particular distribution P^* that will maximize the entropy over Γ.

That is, the distribution P^* in the class Γ results in the maximum possible value of

$H(P) = -\sum_{x \in \chi} p(x) \log p(x) = E_P\{-\log p(X)\}$ where p is the probability mass function of P.

If A is the set of all probability mass functions defined over χ, by information inequality (11) (cf. e.g. Cover and Thomas 1991), it follows that, for any distribution P, $\inf_{q \in A} E_P\{-\log q(X)\}$ is achieved uniquely at $q=p$, where it takes the value $H(P)$.

So,

$$H(P) = \inf_{q \in A} E_P\{-\log q(X)\}, \tag{14}$$

So the maximum entropy can be written as

$$\sup_{P \in \Gamma} H(P) = \sup_{P \in \Gamma} \inf_{q \in A} E_P\{-\log q(X)\} \tag{15}$$

2 Game theory and minimax strategies

Every game consists of decision problems. A decision problem is the problem of choosing among a set of alternative actions, in the cases when certain choices of actions and their consequences can be unambiguously defined, the consequences of joint choices can be precisely specified, and the choosers have distinct preferences among the outcomes (cf. e.g. Rapoport 1966). The simple concepts used here merit precise definition.

Definition 2.1 (cf. e.g. Rapoport 1966) An ***action*** or ***move*** by player i, denoted a_i is a choice that player can make. The entire set of actions available to player i is called an ***action set*** $A_i = \{a_i\}$.

An ordered set $<a_i>, (i = 1, \ldots , n)$ of one action for each of the n players in the game is called an ***action combination***. (16)

"A particular game is defined when the choices open to the players in each situation, the situations defining the end of a play, and the payoffs associated with each play-terminating situation have been specified" (Rapoport 1966).

Definition 2.2 The payoffs, when they are specified on an interval scale (i.e., a scale impervious to order-preserving linear transformations), are called ***utilities***. (17)

That is, given any three outcomes of a game, A, B, and C, a player can specify the rank order among them (e.g., $A>B>C$), as well as the ratio of the differences among the preferences (e.g., $(B-C)/(A-B) = 2$). If the payoffs are interpreted as money, this means that a player's choices are unchanged if the stakes are changed from pennies to thousands of dollars, or if regardless of the outcome, the player has to pay a certain fee.

The players (decision-makers) strive to maximize their expected gains in utility in making their choices. A negative of utility is a loss. Thus, loss is that quantity whose expected gain is attempted to be minimized by a decision-maker. Knowledge of possible consequence of decisions can thus be incorporated in the form of a loss function into statistical analysis (Wald 1950).

A two-player zero-sum game can now be defined.

Definition 2.3 (cf. e.g. Rasmusen 1989, Berger 1985) A game in which the payoffs for all players always sum up to zero is called a ***zero-sum game***. (18)

Definition 2.4 (cf. e.g. Rasmusen 1989) A (pure) ***strategy*** s_i is a particular way for a player i to play the game, uniquely specifying a particular choice for i in each move. (19)

More formally, a player's strategy is a function from a sequence of (previous) events/actions (called an information set) into this player's action set.

The distinction between pure and mixed strategies will be drawn shortly.

For example, consider a zero-sum two-player game of 'rock, paper, scissors' where the players display their choice of rock, paper, or scissors simultaneously three times in a row. The person who wins two or three of the three rounds wins. Then, 'always choose rock' describes one possible strategy; 'choose rock first, then choose whatever the other player chose on previous move' describes another.

Definition 2.5 (cf. e.g. Rasmusen 1989) The set of strategies available to each player i is that player's ***strategy set*** or ***strategy space*** $S_i = \{s_i\}$. A ***strategy combination*** $s = <s_1, ...,s_n>$ is an ordered set consisting of one strategy for each of the n players in the game. (20)

Zero-sum two-player games can be represented in ***normal form***.

Definition 2.6 (cf. e.g. Rasmusen 1989) *Normal form* is the form which consists of all possible strategy combinations $s^1, s^2, ..., s^p$; and payoff functions mapping s^i onto the payoff n-vector $\pi^i (i = 1,2,...,p)$. (21)

The normal form for a two-person game can be drawn up as a table, in which one player's strategy set is listed in the first column of a table, all strategies available to the other player are listed in the first row, and each cell of the table contains the payoffs resulting from the particular row-column strategy combination (the payoffs for column player are always listed before the payoffs for the row player). A simple example will illustrate:

Example 2.1 One-round game of "rock, paper, scissors" where the two players move simultaneously can be represented in normal form as in Table 1 below. The highlighted cell reads that in case Player 1 plays "paper" while Player 2 plays "scissors," Player 1 will receive *-1* units of utility while Player 2 will receive *1*. (22)

Table 1

Player 1 ↓ Player 2 →	*Rock*	*Paper*	*Scissors*
Rock	*0,0*	*-1,1*	*1,-1*
Paper	*1,-1*	*0,0*	***-1,1***
Scissors	*-1,1*	*1,-1*	*0,0*

Note that in a game where it is impossible to base actions on prior information, like in the one-round "rock, paper, scissors", there is no distinction between actions and strategies.

It is clear that payoffs for each player are functions of strategy combinations, so we can denote the contents of highlighted cell above as

π_1 (<paper,scissors>) = -1 , π_2 (<paper,scissors>) = 1

A solution concept for a game is called an ***equilibrium***.

Definition 2.7 (cf. e.g. Rasmusen 1989) An ***equilibrium*** $s^* = \langle s_1^*, \ldots, s_n^* \rangle$ is a strategy combination consisting of a best strategy for each of the n players of the game. (23)

Definition 2.8 (cf. e.g. Rasmusen 1989, Berger 1985) ***Mixed strategy*** for (24)

a player *i* is a function that maps a sequence of actions/events (a possible information set w_i) to a probability distribution over actions: $s_i : w_i \rightarrow m(a_i)$, where $m \geq 0$, $\int_A m(a_i)da_i = 1$.

Compare this with a pure strategy, which maps information sets to actions:

$s_i : w_i \rightarrow a_i$

Example 2.2 (Rasmusen 1989). ***The Welfare Game*** (25)

This game models a pauper who starts looking for work only if he knows he cannot get government aid, and a government that wished to aid a pauper if he searches for work but not otherwise. The payoffs are shown in a table below.

Table 2

Government↓ Pauper→	*Try to work*	*Be idle*
Aid	*3,2*	*-1,3*
No aid	*-1,1*	*0,0*

If players are allowed to mix-and-match their strategies, how often should they play each one in order to maximize their payoffs?

If the government plays *Aid* with probability P_a, and the pauper plays *Try to Work* with probability P_w, their expected payoffs are

$$E\pi_{goverment} = P_a[3P_w + (-1)(1-P_w)] + [1-P_a][(-1)P_w + 0(1-P_w)] = P_a[5P_w - 1] - P_w$$

To maximize the government's payoffs as a function of its actions, we take the derivative of $E\pi_{goverment}$ with respect to the probability of aid P_a and set it equal to zero:

$0 = dE\pi_{goverment} / dP_a = 5P_w - 1$.

Solving for P_w, we find $P_w = 0.2$

Similarly, for pauper's expected payoffs,

$E\pi_{pauper} = P_a[2P_w + 3(1-P_w)] + [1-P_a][1P_w + 0(1-P_w)] = -P_w[2P_a - 1] + 3P_a$

Maximizing the payoffs, we get $0 = dE\pi_{pauper} / dP_w = 2P_a - 1$.

Solving for P_a, we find $P_a = 0.5$

So, both players' payoffs are maximized when pauper plays *Try to work* with probability *0.2* and the government plays *Aid* with probability *0.5*.

In fact, these two mixed strategies form a ***Nash equilibrium***, the most frequently used equilibrium concept (Nash 1951): (Nash equilibria are sets of strategies for players in a noncooperative game such that no single one of them would be better off switching strategies unless others did.)

Another equilibrium concept is based on a notion of maximin strategies.

Definition 2.7 (cf. e.g. Rasmusen 1989) A ***maximin strategy*** for player *i* gives *i* the highest possible payoff in the case when all other players pick strategies to make *i*'s payoff as low as possible. (26)

For a two-player game (to simplify notation),

a maximin strategy for player 1, denoted $s_1^{ma\text{-}mi}$, is the one that solves

$max_{S1}\, min_{S2}\, \pi_1(s_1, s_2)$ or $sup_{S1}\, inf_{S2}\, \pi_1(s_1, s_2)$ (27)

The maximin strategy doesn't have to be unique, and it could be a pure or a mixed strategy.

Definition 2.8 (cf. e.g. Rasmusen 1989) A ***minimax strategy*** in a two-player game is one chosen by one player so as to keep the other player's payoff as low as possible. A minimax strategy for player 1, denoted $s_1^{mi\text{-}ma}$, is the one that solves (28)

$min_{S2}\, max_{S1}\, \pi_1(s_1, s_2)$ or $inf_{S2}\, sup_{S1}\, \pi_1(s_1, s_2)$

As Rasmusen puts it, "In non-zero-sum games, minimax is for sadists and maximin for paranoids. In zero-sum games, the players are merely neurotic: minimax is for optimists and maximin for pessimists" (Rasmusen 1989).

In the Welfare Game above, if we restrict the players to pure strategies, pauper's maximin strategy is *Try to work* (pauper's minimum payoff is *1*), and his minimax strategy is *Be idle* (government's maximum payoff is *0*).

A ***minimax equilibrium*** is a strategy combination consisting of minimax strategies for each player; similarly, a ***maximin equilibrium*** is a strategy combination consisting of maximin strategies for each player.

Theorem 2.1 (Von Neumann 1928)**.** In every two-person zero-sum game, there is a minimax equilibrium in pure or mixed strategies, and it is identical to the maximin equilibrium. (29)

In our notation, allowing for mixed strategies $m(s_1)$ and $m(s_2)$, this means that in a two-person zero-sum game

$$\inf_{m(S2)} \sup_{m(S1)} \pi_1(s_1, s_2) = \sup_{m(S1)} \inf_{m(S2)} \pi_1(s_1, s_2) \tag{29'}$$

And since the limit is actually reached by a pure or mixed strategy,

$$\min_{m(S2)} \max_{m(S1)} \pi_1(s_1, s_2) = \max_{m(S1)} \min_{m(S2)} \pi_1(s_1, s_2) \tag{29"}$$

Coming back to our problem of maximizing entropy, the minimax result from game theory (29') suggests that the following equality might be true (Grünwald and Dawid 2003):

$$\sup_{P \in \Gamma} \inf_{q \in A} E_P\{-\log q(X)\} = \inf_{q \in A} \sup_{P \in \Gamma} E_P\{-\log q(X)\} \tag{30}$$

While the supremum on the left-hand side is achieved when the entropy over Γ is maximized, the meaning of the infimum on the right-hand side will become clear from the following section.

3 Decision-making as games with Nature

Decision theory presents a detailed analysis of a subset of cases studied by Game Theory. In particular, consider a Decision Maker who has to take some action a selected from the given *action space* A, after which Nature will reveal the value $x \in \chi$ of a quantity X, and Decision Maker will then suffer a loss $L(x,a)$ in $(-\infty, \infty]$ (Grünwald and Dawid 2003, Berger 1985).

Since the value x is independent of any action chosen by Decision Maker, this can be considered as a zero-sum game between Nature and Decision Maker. The two players move simultaneously, after which the moves are revealed and Decision Maker pays Nature $L(x,a)$.

If Decision Maker knows that X comes from a probability distribution P (objective probability), or if Decision Maker is using P to represent uncertainty about X (subjective probability), then the undesirability to Decision Maker of any act $a \in A$ will be assessed by means of its *Bayesian expected loss*,

$$L(P,a) = \sum_x P(x)\, L(x,a) = E_P\{L(X,a)\} \quad [1] \qquad (31)$$

The *expected value* or expected loss method in decision-making was known from the 17th century. In 1738 Daniel Bernoulli published an influential paper entitled *Exposition of a New Theory on the Measurement of Risk* in which he defines a utility

[1] I am not considering non-finite cases, where loss would be defined as the corresponding integral.

function and computes expected utility in presenting a solution to a decision problem. In the last century, Frank Ramsey, Bruno de Finetti, Leonard Savage and others developed and applied the concept of expected utility/loss in situations where all probabilities are subjective (Wikipedia: Decision Theory).

In non-Bayesian or Frequentist schools of statistical decision theory, expected loss function is different from *L(P,a)* above. First, we need to distinguish states of nature $x \in \chi$ from the random variable *y* (dependent on *x*) representing sample information, or observation/measurement. Then, we can define the notion of *decision rule* (*procedure*): (cf. e.g. Berger 1985)[2]

Definition 3.1 A (nonrandomized) *decision rule* $\delta(y)$ is a (measurable) function (32) from *Y* into *A*. If *Y=y* is the observed value of the sample information, then $\delta(y)$ is the action that will be taken. If $P_x(\delta_1(Y)=\delta_2(Y))=1$ for all states of nature *x*, the two decision rules δ_1 and δ_2 are equivalent.
In the absence of data, the decision rule is 'no action.'

The *risk function* of a decision rule is the expected loss for repeated use of $\delta(Y)$ with varying *Y*:

$$R(x,\delta)= \sum_y P(y/x)\, L(x,\delta(y)) = Ey\{L(x,\delta(Y))\} \quad (33)$$

For a no-data problem, $R(x,\delta) \equiv L(P,a)$

[2] The entire section following this definition is drawn from Berger's classic textbook on Statistical Decision Theory (Berger 1985), so that definitions, theorems, and proofs till the end of the section are cited from there, unless otherwise indicated.

The following partial ordering serves as the first step to determining a good decision rule:

Definition 3.2 A decision rule δ_1 is ***R-better*** than a decision rule δ_2 if (34)

$R(x,\delta_1) \leq R(x,\delta_2)$ for all x (with strict < for some x). The two rules are

R-equivalent if $R(x,\delta_1) = R(x,\delta_2)$ for all x.

On the basis of this partial ordering we can define the notion of rule admissibility:

Definition 3.3 A decision rule δ is *admissible* if there is no *R*-better decision rule. (35)

A decision rule is *inadmissible* otherwise.

Finally, the notion of ***Bayes risk*** is defined by averaging both over x and over y.

Definition 3.4 Given a prior distribution P on χ, *Bayes risk of decision rule* δ is (36)

defined as $r(P,\delta) = E_P\{R(x,\delta)\}$

Definition 3.5 A ***randomized decision rule*** $\delta^*(y,.)$ for each y is a probability (37)

distribution on A, stating that if y is observed, $\delta^*(y,B)$ is the probability that an

action in $B \subseteq A$ will be chosen.

Definition 3.6 A ***randomized act/move*** for a player is a choice represented as a

(38)

probability distribution on this player's action space, and is the same as a randomized decision rule for a no-data problem.

A non-randomized rule is then a special case of randomized ones, choosing a specific action a for each *x* with probability *1*.

The loss function and risk function of a randomized rule are defined in terms of expected loss.

$$L(x,\delta^*(y,.))=E_{a\sim\delta^*(y,.)}\{L(x,a)\}, \quad R(x,\delta^*)=E_Y\{L(x,\delta^*(Y,.))\} \tag{39}$$

For randomized rules, as for non-randomized rules, in a no-data problem, the risk is just the loss.

Thus, for Nature, a non-randomized move involves picking value *x* from the sample space χ, while a randomized move involves picking a distribution *P* over χ. For Decision Maker, a non-randomized move involves picking an action *a* from *A*, while a randomized move involves picking a distribution *z* over *A*.

Applying the definition in (21) to this case, $L(x,z) = \sum_{a\in A} z(a)\, L(x,a) = E_{A\sim z}\{L(X,a)\}$

and thus $L(P,z) = \sum_a \sum_x z(a)P(x)\, L(x,a) = E_{(X,A)\sim P\times z}\{L(X,a)\}$

For obvious reasons, the attention of scholars here is restricted to cases when the loss function is defined: to the set W of probability functions *P* such that *L(P,a)* is defined

for all $a \in A$, and to the set $\boldsymbol{Z}$ of probability functions z such that $L(P,z)$ is defined for all $P \in \boldsymbol{W}$. Then the loss function is well-defined:

Lemma 3.1 (Grünwald and Dawid 2003) For all $P \in \boldsymbol{W}$ and $z \in \boldsymbol{Z}$, (40)

$L(P,z) = E_{X \sim P}\{L(X,z)\} = E_{A \sim z}\{L(P,A)\}$

Proof. Suppose L is finite. Then, by Fubini's theorem, orders of summation or integration can be interchanged, $E_{X \sim P}\{L(X,z)\} = E_{A \sim z}\{L(P,A)\}$

For the case $L(P,z)=\infty$, consider the first equality $L(P,z) = E_{X \sim P}\{L(X,z)\}$.

Suppose $L(P,z) \geq 0$ everywhere.

If $L(x,z) = \infty$ for x in a subset of χ that has a positive P-measure, then both sides are infinite and the first equality in (40) holds.

If $L(x,z)$ is infinite on a subset of χ that has a non-positive (that is, zero) P-measure, it is finite almost surely (that is, the loss associated with events whose probability is zero can be considered to be zero as well).

If $E_{X \sim P}\{L(X,z)\}$ were finite, then by Fubini it would be the same as $L(P,z)$. So it is not finite, and again both sides are infinite and the first equality in (40) holds.

If L can be negative at some points, note that its negative part must be finite; and the reasoning for $L(P,z) = -\infty$ is parallel to that for the positive ∞ case.

The whole argument can be repeated changing x into a and P into z to prove the second equality $L(P,z) = E_{A \sim z}\{L(P,A)\}$ q.e.d.

Corollary 3.2 (Grünwald and Dawid 2003) For all $P \in \mathcal{W}$, (41)

$\inf_{z \in \mathbf{Z}} L(P,z) = \inf_{a \in A} L(P,a)$

Proof. First, note that $\inf_{z \in \mathbf{Z}} L(P,z) \leq \inf_{a \in A} L(P,a)$

If $\inf_{a \in A} L(P,a) = -\infty$, then $\inf_{z \in \mathbf{Z}} L(P,z) = \inf_{a \in A} L(P,a) = -\infty$;

otherwise for any $z \in \mathbf{Z}$ $L(P,z) = E_{A \sim z}\{L(P,A)\}$, but

$E_{A \sim z}\{L(P,A)\} \geq \inf_{a \in A} L(P,a)$, so $\inf_{z \in \mathbf{Z}} L(P,z) = \inf_{a \in A} L(P,a)$.

This result, together with Theorem 3.2 below will allow us to concentrate on non-randomized acts for Decision Maker in our attempts to minimize loss.

Definition 3.7 (cf. e.g. Berger 1985, see footnote 3) A set $S \subseteq R^m$ is convex if for (42)

any two points $\boldsymbol{x}$ and $\boldsymbol{y}$ in S, the segment joining them ($ax+(1-a)y$ for $0 \leq a \leq 1$) is in S.

Definition 3.8 A real-valued function $g(\boldsymbol{x})$ defined on a convex set S is ***convex*** if (43)

for all $\boldsymbol{x}, \boldsymbol{y} \in S$, $g(a\boldsymbol{x}+(1-a)\boldsymbol{y}) \leq ag(\boldsymbol{x})+(1-a)g(\boldsymbol{y})$, strictly so if the inequality is strict.

$g(\boldsymbol{x})$ is ***concave*** if for all $\boldsymbol{x}, \boldsymbol{y} \in S$, $ag(\boldsymbol{x})+(1-a)g(\boldsymbol{y}) \leq g(a\boldsymbol{x}+(1-a)\boldsymbol{y})$, strictly so if the inequality is strict.

Definition 3.9 A ***hyperplane*** in R^m is a set of the form

(44)

$H(d,k) = \{ \boldsymbol{z} \in R^m: d^t\boldsymbol{z}=\sum^m_{i=1} d_i z_i=k\}$, where k is some real number, $d \in R^m$ and $d \neq 0$.

A hyperplane has the following property: if l is a line in R^m and z^0 a point in R^m, and y^1 and y^2 any distinct point of l, the hyperplane passing through z^0 and perpendicular to l is $H(d,k)$ where $d=y^1-y^1$ and $k=(z^0)^t(y^1-y^2)$. (45)

Definition 3.10 A ***supporting hyperplane*** to a set S in R^m at a boundary (46)
point $\boldsymbol{s}$ of S is a hyperplane $H(d,k)$ ($d \neq 0$) which contains $\boldsymbol{s}$ (that is, $d^t\boldsymbol{s}=k$) and for which $d^t\boldsymbol{s} \geq k$ when $\boldsymbol{s} \in S$. A ***separating hyperplane*** for sets S_1 and S_2 in R^m is a hyperplane, $H(d,k)$ $(d \neq 0)$ such that $d^t s^1 \geq k$ for $s^1 \in S_1$ and $d^t s^2 \leq k$ for $s^2 \in S_2$.

A supporting hyperplane to a set S, according the definition in (46) is tangent to the S, and S lies completely to one side of the hyperplane. A separating hyperplane is then one for which one set lies entirely in the half-space $\{z \in R^m \mid d^t z \geq k\}$ on one side of the hyperplane, and the other set lies entirely in the half-space $\{z \in R^m \mid d^t z \leq k\}$ on the other side of the hyperplane.

Before we can prove a few important theorems in the minimax analysis, we need the following lemma:

Lemma 3.2 If S is a closed convex set in R^m and $x^0 \notin S$, then there is a hyperplane (47) separating S and x^0.

Proof. The proof will be geometric.

First, we show that there is a unique $s^0 \in S$ nearest to x^0, such that $|s^0 - x^0| = inf_{s \in S}|s - x^0|$.

Existence of s^0:

Choose a sequence $s^n \in S$ so that $|s^n - x^0| \rightarrow inf_{s \in S}|s - x^0|$. It is easy to show that $\{s^n\}$ can be chosen to be a bounded sequence, so that, by the Bolzano-Weierstrass theorem, $\{s^n\}$ has a convergent subsequence $\{s^{n(i)}\}$ with a limit point s^0. Since S is closed, $s^0 \in S$.

Then, $|s^0 - x^0| = lim_{n \rightarrow \infty}|s^{n(i)} - x^0| = inf_{s \in S}|s - x^0|$.

Uniqueness of s^0:

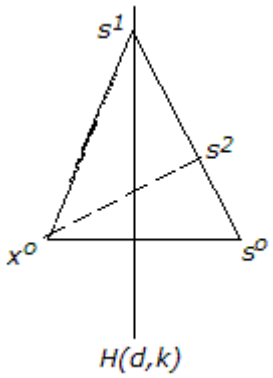
Assume there is an $s' \in S$ such that $s' \neq s^0$ and $|s' - x^0| = |s^0 - x^0|$. Then the points x, s^0, and s' form an isosceles triangle with the base $s's^0$. Now, the midpoint s'' of this base is closer to x^0 than either s' or s^0, and since S is convex, $s'' \in S$. This contradicts the assumption that s^0 is closest to x^0. So s^0 must be unique.

Second, we define $d = (s^0 - x^0)$ and $k = (|s^0|^2 - |s^0|^2)/2$. Because of the property (45) of hyperplanes, we know that $H(d,k)$ is the perpendicular to the line segment x^0s^0 and passing through the midpoint $(x^0 + s^0)/2$ of that line segment. Also, since $0 < \frac{1}{2}|s^0 - x^0| = \frac{1}{2}(d^t s^0 - d^t x^0)$, then $d^t x0 < d^t s^0$.

Consequently, $d^t x^0 < \frac{1}{2}(d^t x^0 + d^t s^0) = k < d^t s^0$.

So, $H(d,k)$ separates x^0 and s^0.

Third, we show that $H(d,k) \cap S$ is empty. Assume it is not, that is, suppose $s^1 \in S$ and $d^t s^1 = k$. Because $H(d,k)$ is the perpendicular bisector of $x^0 s^0$, the triangle joining the points x^0, s^0, and s^1 is isosceles, with $x^0 s^0$ as the base. Let s^2 be the point on $s^0 s^1$ for which $x^0 s^2$ is perpendicular to $s^0 s^1$. Then $|x^0 - s^2| < |x^0 - s^0|$, and since S is convex, $s^2 \in S$, again contradicting the assumption that s^0 is closest to x^0. So $H(d,k) \cap S$ is empty.



Finally, we show that $d^t s > k$ for $s \in S$, so that $H(d,k)$ separates x^0 and S.

Assume that $s^1 \in S$ and $d^t s^1 = a_1 \leq k$. From the second step, we know that $d^t s^0 = a_0 > k$.

Now, let $\lambda = (k - a_1)/(a_0 - a_1)$, and notice that $0 \leq \lambda < 1$. By convexity, $s^2 = (\lambda s^0 + (1-\lambda) s^1) \in S$, but then $d^t s^2 = \lambda a_0 + (1-\lambda) a_1 = a_1 + \lambda(a_0 - a_1) = k$.

This contradicts the previous (third) step, and so we are done. q.e.d.

The next two steps, building up to some of the central theorems of minimax analysis are the following two lemmas, called Supporting Hyperplanes Theorem and Separating Hyperplanes Theorem.

Lemma 3.3 (Supporting Hyperplanes Theorem). If s^0 is a boundary point of a convex set S, then there is a supporting hyperplane to S at s^0. (48)

Proof. S and *closure(S)* have the same boundary points, so that there are points $x^m \notin closure(S)$ such that $x^m \to s^0$. By Lemma 3.2 in (47), there are hyperplanes $H(d^m, k_m)$, such that

$$(d^m)^t s \geq k_m \text{ for } s \in closure(S), \text{ and } (d^m)^t x^m \leq k_m \tag{49}$$

These hyperplanes (and the inequalities in (49)) are not affected if d^m and k_m are replaced by $t^m = d^m/|d^m|$ *and* $k'_m = k_m/|d^m|$.

So, for $H(t^m, k'_m)$, we also have

$$(t^m)^t s \geq k'_m \text{ for } s \in closure(S), \text{ and } (t^m)^t x^m \leq k'_m \tag{50}$$

Now, since $|t^m|=1$ for all m, the Bolzano-Weierstrass theorem implies that the sequence $\{t^m\}$ has a convergent subsequence $\{t^{m(i)}\}$. Let $t^0 = lim_{t \to \infty} t^{m(i)}$.

Now we want to show that the sequence $\{k'_{m(i)}\}$ has a convergent subsequence. Note that for any fixed $s \in S$,

$$k'_m \leq (t^m)^t s \leq |t^m||s| = |s|. \tag{50'}$$

$$\text{Similarly, } k'_m \geq (t^m)^t x^m \geq -|t^m||x^m| = -|x^m|. \tag{50''}$$

Since $x^m \to s^0$, (50') and (50'') imply that $\{k'_{m(i)}\}$ is a bounded sequence. So, by Bolzano-Weierstrass, it has a convergent subsequence $\{k'_{m(i(j))}\}$, whose limit we shall call k_0.

Since by (50) $k'_{m(i(j))} \leq (t^{m(i(j))})^t s$ for all $s \in closure(S)$, it follows that

$k_0 \leq (t^0)^t s$ for all $s \in closure(S)$ (51)

Similarly, since $k'_{m(i(j))} \geq (t^{m(i(j))})^t x^{m(i(j))}$, it follows that $k_0 \geq (t^0)^t s^0$.

Together with (51), this yields $(t^0)^t s^0 = k_0$. (51')

From (51) and (51') we have that $H(t^0, k_0)$ is a supporting hyperplane to S at s^0. q.e.d.

Lemma 3.4 (Separating Hyperplanes Theorem). If S_1 and S_2 are disjoint convex (52)

subsets of R^m, then

there is a vector $d \in R^m$ $(d \neq 0)$ such that $d^t s^1 \geq d^t s^2$ for all $s^1 \in S_1$ and $s^2 \in S_2$. (52')

Indeed defining $k = sup_{s^2 \in S_2} d^t s^2$, the hyperplane $H(d,k)$ separates S_1 and S_2. (52")

Proof. Let $A = \{x = s^1 - s^2 : s^1 \in S_1 \text{ and } s^2 \in S_2\}$

A is convex; also $\mathbf{0}$ is not in A, since S_1 and S_2 are disjoint. There are two possibilities.

First, assume $\mathbf{0} \in closure(A)$, that is, $\mathbf{0}$ is a boundary point of A. Then, by Lemma 3.3 in (48) there is a supporting hyperplane $H(d,k)$ to A at $\mathbf{0}$. Therefore, $k = d^t \mathbf{0} = 0$, and

$d^t x \geq k = \mathbf{0}$ for $x \in A$.

Then, for $s^1 \in S_1$ and $s^2 \in S_2$, it follows that $d^t(s^1 - s^2) \geq 0$. This proves (52').

Observe now that $d^t s^1 \geq sup_{s^2 \in S_2} d^t s^2 \geq d^t s^2$, which proves (52").

Second, assume $\mathbf{0} \notin closure(A)$. Then, by Lemma 3.2 in (47) there is a hyperplane separating $\mathbf{0}$ and $closure(A)$. The rest of the argument is as above. q.e.d.

The final lemma focusing on a property of convex sets is given below:

Lemma 3.5 Let $\boldsymbol{X}$ be an m-variate random vector such that (53)

$E[|\boldsymbol{X}|]<\infty$ and $P(\boldsymbol{X}\in S)=1$, where S is a convex subset of R^m. Then $E[\boldsymbol{X}]\in S$.

Proof. Define $\boldsymbol{Y}=\boldsymbol{X}-E[\boldsymbol{X}]$, and let $S'=S-E[\boldsymbol{X}] = \{y \mid y = x-E[\boldsymbol{X}] \text{ for some } x\in S\}$.

Note that S' is convex, that $P(\boldsymbol{Y}\in S')=1$, and that $E[\boldsymbol{Y}]=0$. Showing that $E[\boldsymbol{X}]\in S$ is equivalent to showing that $\boldsymbol{0}\in S'$. This will be proved by induction on m.

When $m=0$, $\boldsymbol{Y}$ is degenerate to a point, so that $E[\boldsymbol{Y}]=0$.

Now suppose the result holds for all dimensions up to and including $m-1$.

For dimension m, assume, by way of contradiction, that $\boldsymbol{0}\notin S'$.

Then, by Lemma 3.4 in (52) there is a vector $d\neq 0$ in R^m such that $d^t y \geq d^t\boldsymbol{0} = 0$ for all $y\in S'$. Let $\boldsymbol{Z}=d^t\boldsymbol{Y}$, then $P(\boldsymbol{Z}\geq 0)=1$. However, $E[\boldsymbol{Z}] = d^t E[\boldsymbol{Y}] = 0$, so that $P(\boldsymbol{Z}=0)=1$. So, with certainty, $\boldsymbol{Y}$ lies in the hyperplane defined by $d^t y=0$.

Now let $S''=S'\cap \{y \mid d^t y=0\}$, and observe that S'' is a convex subset of an $(m-1)$-dimensional Euclidean space, and that $P(\boldsymbol{Y}\in S'')=1$ and $E[\boldsymbol{Y}]=0$.

By induction hypothesis, $\boldsymbol{0}\in S''$. Since S'' is a subset of S', this contradicts the supposition that $0\notin S'$, completing the proof. q.e.d.

Now, we are ready to prove the first important theorem, called Jensen's inequality.

Theorem 3.1 (Jensen's inequality). Let $g(\boldsymbol{x})$ be a convex real-valued function

(54)

defined on a convex subset S of R^m, and X an m-variate random vector for which $E[|X|]< \infty$, and $P(X \in S) = 1$. Then, $g(E\{X\}) \leq E\{g(X)\}$, strictly so if g is strictly convex and X is not concentrated at a point.

$g(E\{X\})$ is well-defined, since when $E[|X|]< \infty$, and $P(X \in S)=1$, $E[X] \in S$.

Proof. By induction on m: For $m=0$, theorem holds trivially since S is a point. Assume theorem holds for dimensions up to $m-1$.

Define $\boldsymbol{B}=\{(\boldsymbol{x}^t,y)^t \subseteq R^{m+1}: \boldsymbol{x} \in S, y \in R^1,$ and $y \geq g(\boldsymbol{x})\}$.

First, $\boldsymbol{B}$ is convex in R^{m+1} : If $(\boldsymbol{x}^t, y_1)^t$ and $(\boldsymbol{z}^t,y_2)$ are two points in $\boldsymbol{B}$, then

$k(\boldsymbol{x}^t, y_1)^t +(1-k)(\boldsymbol{z}^t,y_2)=([k\boldsymbol{x}+(1-k)\boldsymbol{z}]^t, ky_1+(1-k)y_2)$

But since S is convex, $[k\boldsymbol{x}+(1-k)\boldsymbol{z}] \in S$ for $0 \leq k \leq 1$,

and since $y_1 \geq g(\boldsymbol{x})$ and $y_2 \geq g(\boldsymbol{z})$ and g is convex,

$ky_1+(1-k)y_2 \geq k\, g(\boldsymbol{x})+(1-k)\, g(\boldsymbol{z}) \geq g(k\boldsymbol{x}+(1-k)\boldsymbol{z})$

Thus, $k(\boldsymbol{x}^t, y_1)^t +(1-k)(\boldsymbol{z}^t,y_2) \in \boldsymbol{B}$ for $0 \leq k \leq 1$.

Second, from Lemma 3.5 in (53) above, $E[\boldsymbol{X}] \in S$, so it follows that $\boldsymbol{b} = (E[\boldsymbol{X}]^t, g(E[\boldsymbol{X}]))^t$ is a boundary point of $\boldsymbol{B}$.

Third, let $H(d,k)$ be the supporting hyperplane to $\boldsymbol{B}$ at $\boldsymbol{b}$. Writing $d=(v^t,r)^t,$ where $v \in R^m$ and $r \in R^1$, it follows that $d^t b = v^t E[\boldsymbol{X}]+rg(E[\boldsymbol{X}])=k$

and $d^t(\boldsymbol{x}^t,y)^t = v^t\boldsymbol{x} + ry \geq k$ for $\boldsymbol{x} \in S$ and $y \geq g(\boldsymbol{x})$

Now, if we set $\boldsymbol{x} = \boldsymbol{X}$ and $y = g(\boldsymbol{X})$, then with probability one,

$v^t\boldsymbol{X} + rg(\boldsymbol{X}) \geq v^tE[\boldsymbol{X}]+rg(E[\boldsymbol{X}])$

Fourth, note that $r \geq 0$ (otherwise $d^t(\boldsymbol{x}^t,y)^t = v^t\boldsymbol{x} + ry \geq k$ for $\boldsymbol{x} \in S$ and $y \geq g(\boldsymbol{x})$ will be wrong because y will be infinite). If $r > 0$, $g(E\{\boldsymbol{X}\}) \leq E\{g(\boldsymbol{X})\}$ follows by taking the expectations.

If $r = 0$, from $v^t\boldsymbol{X} + rg(\boldsymbol{X}) \geq v^tE[\boldsymbol{X}]+rg(E[\boldsymbol{X}])$ we have $h(\boldsymbol{X}) = v^t(\boldsymbol{X}- E[\boldsymbol{X}]) \geq 0$.

Also, $E[h(\boldsymbol{X})] = 0$, so $h(\boldsymbol{X}) = 0$ with probability one, which we can rewrite as

$P(v^t\boldsymbol{X}=k) = 1$.

Thus, $\boldsymbol{X}$ is concentrated on the hyperplane $H(v,k)$ with probability one.

Then let $Sh = S \cap H(v,k)$. Since intersection of convex sets is convex, Sh is convex.

$P(\boldsymbol{X} \in Sh) = 1$. But Sh is an $(m-1)$-dimensional set, so the induction hypothesis applies to give the desired conclusion.

Next, we prove a theorem that lets us eliminate randomized rules from consideration in certain cases.

Theorem 3.2 Assume that A is a convex subset of R^m, and that for each $x \in \chi$ (55)

the loss function $L(x,\boldsymbol{a})$ is a convex function of $\boldsymbol{a}$. Let δ^* be a randomized decision rule in D^* for which $E^{\delta^*(y,.)}[|\boldsymbol{a}|] < \infty$ for all $y \in Y$. Then (subject to measurability conditions) the nonrandomized rule $\delta(y) = E^{\delta^*(y,.)}[\boldsymbol{a}]$

has $L(x,\delta(y)) \leq L(x,\delta^*(y,.))$ for all y and x.

Proof. From Lemma 3.5 in (53) above, $\boldsymbol{\delta(y) \in A}$.

From Jensen's inequality we have that

$L(x,\delta(y))=L(x,\ E^{\delta^*(y,.)}[\boldsymbol{a}]) \ \leq E^{\delta^*(y,.)}[L(x,\boldsymbol{a})]= L(x,\delta^*(y,.))$ q.e.d.

So, when the loss function is convex, only nonrandomized decision rules need be considered.

For a randomized rule $\delta^* \in D^*$, the quantity $sup_{x\in\chi}\ R(x,\delta^*)$ is the worst possible outcome if δ^* is used. A desire to protect oneself from the worst possible scenario is reflected in the ***Minimax principle***:

A decision rule δ_1^* is preferred to δ_2^* if $sup_{x\in\chi}\ R(x,\delta_1^*) \leq sup_{x\in\chi}\ R(x,\delta_2^*)$. (56)

Definition 3.11 A decision rule δ^{*M} is a **minimax decision rule** if it minimizes (57)

$\sup_{x\in\chi} R(x,\delta^*)$ among all randomized rules in D^*, that is, if

$sup_{x\in\chi}\ R(x,\delta^{*M})=inf_{\delta^*\in D^*}\ sup_{x\in\chi}\ R(x,\delta^*)$.

This is equivalent to a minimax mixed strategy (28) if decision-making is interpreted as a zero-sum game with Nature.

Note that minimaxity, according to Definition 3.11, is decided regardless of the distribution of X or any prior knowledge about this distribution. For minimaxity specific to a particular class of distributions, decision rules (procedures), and the notion of Bayes risk $r(P,\delta)$ will be used (Definition 3.4).

Definition 3.12 If D is a probability distribution, then a ***Bayes act against D*** or (58)

D-minimax act is the action a that minimizes the average expected loss $L(D,a)$

over all $a \in A$.

A Bayes act need not exist for every distribution/loss function. A more universally applicable notion is that of a Bayes loss:

Definition 3.13 The ***Bayes loss*** of a distribution $P \in \mathrm{P}$, denoted $H(P) \in [-\infty, \infty]$, (59)

defined by $H(P) = \inf_{a \in A} L(P,a)$

From (41), (55) we know that we will get the same result whether using (58) or extending the definition to randomized acts.

In a frequentist Bayes approach, when a class of distributions Γ is given, a distribution is often chosen based on minimization of Γ-Bayes risk (Definition 3.14).

Definition 3.14 The Γ-Bayes risk of a procedure δ is $r_\Gamma(\delta) = \sup_{P \in \Gamma} r(P,\delta)$ (60)

Correspondingly, the Γ-minimax principle is:

A decision rule δ_1^* is preferred over a rule δ_2^* if $r_\Gamma(\delta_1^*) < r_\Gamma(\delta_2^*)$ (61)

Definition 3.15 A rule δ^* is said to be Γ-minimax if (62)

$r_\Gamma(\delta^*) = \inf_{\delta^* \in D^*} r_\Gamma(\delta^*) = \inf_{\delta^* \in D^*} \sup_{P \in \Gamma} r(P,\delta^*)$

Now consider the 'log loss game' (Good 1952), in which Decision Maker has to specify some $q \in A$, and her ensuing loss, if Nature then reveals $X=x$, is measured by $-\log q(x)$. In this game, $H(P)$ is just the Shannon entropy of P. A reinterpretation of this game will be given below in section 4.2.

In addition, the infimum on the right-hand side of (30), repeated below in the next section, is then a *Γ-minimax* act (Berger 1985), also termed a *robust Bayes* act against Γ. Such an act describes a maximin strategy (28) for Decision Maker for the log loss decision problem given Γ.

4 Maximal entropy as minimal loss

4.1 Proving the result for the special case

The following theorem is key to showing the connection between maximal entropy and the acts that minimize loss:

Theorem 4.1 (Grünwald and Dawid 2004) Let $P \in$ P, and suppose $H(P)$ is finite. (63)

Then the following three conditions hold:

1. $z_P \in \mathbf{Z}$ is Bayes against P iff $E_P\{L(X,a)-L(X,z_P)\} \in [0, \infty]$ for all $a \in A$.
2. z_P is Bayes against P iff $L(P,z_P) = H(P)$.
3. If P admits some randomized Bayes act, then P also admits some non-randomized Bayes act.

The first two conditions of this theorem claim that given a distribution P, a Bayes act against P can be found by comparing its loss with those of the available non-randomized acts (condition 1), or by finding an act whose loss is the entropy of P (condition 2).
The search for a Bayes act may be restricted to non-randomized acts, if desired (condition 3).

Proof. From corollary in (41) and the fact that $H(P)$ is finite, we have 1. and 2.

Let $f(P,a)$ be defined as $L(P,a)-H(P)$.

Then $f(P,a) \geq 0$ for all a, while $E_{A \sim z_P} f(P,A) = L(P,z_P) - H(P) = 0$. It follows that the set

$\{a \in A: f(P,a)=0\}$ has probability 1 under z_P and so must be non-empty. q.e.d

And so, we hope to successfully follow (30), repeated below, in our search for a Bayes act against Γ in the log-loss game.

$$\sup_{P\in\Gamma} \inf_{q\in A} E_P\{-\log q(X)\} = \inf_{q\in A} \sup_{P\in\Gamma} E_P\{-\log q(X)\} \qquad (30)$$

When Γ is closed and convex, (30) does indeed hold under very general conditions. Moreover the infimum on the right-hand side is achieved, uniquely, for $q=p^*$, the probability mass function of the maximum entropy distribution P^*.

Thus, in this game between Decision Maker and Nature, the maximum entropy distribution P^* may be viewed simultaneously as defining both Nature's maximin and Decision Maker's minimax strategy. In other words, maximum entropy is robust Bayes against Γ.

Note that the acts q available to Decision Maker are not restricted to those corresponding to a distribution in the class Γ: that the optimal act p^* does indeed turn out to have this property follows from the analysis, and is not a case of circular reasoning.

Suppose Γ is described by mean-value constraints (Jaynes 1989, Cziszár 1991):

$\Gamma=\{P : E_P(T) = \tau\}$, where $T=t(X)\in R^k$ is a real-valued statistic. (64)

Grünwald (1998) shows that for such constraints the property (30) is particularly easy to show. That is, while the Decision Maker cannot be sure which distribution Nature will choose, a particular exponential distribution will be a possible (and an optimal) choice for Nature.

Namely, in presenting a general theory of exponential families, Barndorff-Nielsen (1978) shows that under some mild consistency conditions on τ (such as requiring that the probability function integrate to *1*), there will exist a distribution P^* which satisfies the constraint $E_P(T) = \tau$, and has probability mass function of the form

$$p^*(x)=exp\{\alpha_0 +\alpha^T t(x)\}= exp\{\alpha_0\}\ exp\{\alpha^T t(x)\} \text{ for some } \alpha \in R^k,\ \alpha_0 \in R. \quad (65)$$

Then for any $P \in \Gamma$,

$$E_P\{ - log\ p^*(X)\} = -\alpha_0 - \alpha^T \tau = H(P^*) \quad (66)$$

Thus p^* has the same expected loss under any $P \in \Gamma$.

P^* maximizes entropy, since for any $P \in \Gamma$,

$$H(P) = inf_{q \in A}\ E_P\{-log\ q(X)\} \leq E_P\{-log\ p^*(X)\} = H(P^*) \quad (67)$$

by the previous rule.

To demonstrate that p^* is a Robust Bayes act against Γ and that (30) holds, first observe that for any $q \in A$,

$$sup_{P \in \Gamma} E_P\{-log\ q(X)\} \geq E_{P^*}\{-log\ q(X)\} \geq E_{P^*}\{-log\ p^*(X)\} = H(P^*),$$

where the second inequality is the information inequality (11) (Cover and Thomas 1991). So

$$H(P^*) \leq inf_{q \in A}\ sup_{P \in \Gamma} E_P\{-log\ q(X).\} \quad (68)$$

From the definition of p^* in (65), we have, trivially, that

$$sup_{P \in \Gamma} E_P\{-log\ p^*(X)\} = H(P^*) \quad (69)$$

A related theorem concerning the existence of minimax strategies is stated below as Theorem 4.2.

Definition 4.1 (cf. e.g. Berger 1985). For each randomized strategy δ^* (70)

let $R_i(\delta^*) = L(x_i,\delta^*)$, and $\mathbf{R}(\delta^*)=(R_1(\delta^*), R_2(\delta^*), ..., R_m(\delta^*))$.

The set of all ***risk points*** $\mathbf{R}(\delta^*)$, $S = \{ \mathbf{R}(\delta^*)\}$ is called ***the risk set*** of the game. (71)

This set is convex.

Theorem 4.2 (Minimax Theorem). (cf. e.g. Berger 1985) Consider a two-person (72)

zero-sum game in which $\chi = \{x_1, x_2, ..., x_m\}$.

Assume that the risk set S is bounded from below. Then the game has value

$$V=\sup_{P\in\Gamma}\inf_{\delta^*\in D^*} L(P,\delta^*)= \inf_{\delta^*\in D^*}\sup_{P\in\Gamma} L(P,\delta^*) \tag{73}$$

and a maximin strategy P^M exists (for Nature).

Moreover, if S is closed from below, then a minimax strategy δ^{*M} exists,

and $L(P^M, \delta^{*M}) = V$.

In general, when the essential assumptions that the risk set is closed from below and bounded from below are true, the game will usually have a value, and a minimax strategy will exist.

4.2 Games, entropy, and codeword length

Here, we will consider a reinterpretation of the zero-sum log-loss game between Decision Maker and Nature described in the previous section. This reinterpretation was proposed by Topsøe (Topsøe 1979, Topsøe 1993). Probability distributions are defined on the finite space χ (for Topsøe, discrete countable χ). Γ is the set of *consistent* probability distributions on χ. The Decision Maker is now the Observer, whose goal is, as before, to make inferences about an unknown probability distribution based on the incomplete information that this distribution belongs to Γ. While the Nature chooses the true distribution from Γ, the observer chooses a way of observing χ, formalized by a *code* (defined below in 4.2).

Definition 4.2 A code is a map $k: \chi \rightarrow [0,\infty]$ such that $\Sigma_{x\in\chi} e^{-k_x} = 1$ (74)

This definition can be interpreted as existence of *codebook* corresponding to k, with all the codewords corresponding to the possible states x. The length of the codeword corresponding to $x \in \chi$ is k_x, which may be thought of as the length of time needed by the Observer to discover the true state of the system when this state is x, or as the time needed to communicate x to someone else.

The goal of the Observer is to minimize this observation time, which is equal to the *expected code length*, whereas Nature is trying to maximize this quantity.

From (74) it is clear that there is a natural bijection between the set M of all probability distributions over χ and the set K of all codes. This bijection is given by

$$k \leftrightarrow P: \quad k_x = -\log p(x) \text{ and } p(x) = e^{-k_x} \tag{75}$$

When (75) holds, we say that k is adapted to P, or that P is associated with k. The cost function for the game is defined on $K \times \Gamma$ with values in $[0, \infty]$ given by

$$(k,P) \mapsto \langle k,P \rangle = E_P\{k\} \qquad \text{where } (k,P) \in K \times \Gamma$$

The cost function thus gives the expected code length (or observation time).

From the above definitions it is clear that

$$\sup_{P \in \Gamma} \inf_{k \in K} \langle k,P \rangle \leq \inf_{k \in K} \sup_{P \in \Gamma} \langle k,P \rangle \tag{76}$$

When the equality obtains, $\alpha = \sup_{P \in \Gamma} \inf_{k \in K} \langle k,P \rangle = \inf_{k \in K} \sup_{P \in \Gamma} \langle k,P \rangle$ is the value of the game.

For a particular code k, the risk is defined as follows:

Definition 4.3 Risk of code k, $R(k) = R(k \mid \Gamma) = \sup_{P \in \Gamma} \langle k,P \rangle$ (77)

Thus, the minimum risk is

Definition 4.4 $R_{min}(\Gamma) = \inf_{k \in K} R(k) = \inf_{k \in K} \sup_{P \in \Gamma} \langle k,P \rangle$ (78)

The distribution P^* is *optimal strategy* for Nature when $P^* \in \Gamma$ and

$inf_{k\in K} <k,P^*> = sup_{P\in\Gamma} inf_{k\in K} <k,P>$ (79)

The quantity $inf_{k\in K} <k,P^*>$ is just the entropy of P^*, $H(P^*) = -\sum_{x\in\chi} p(x) \log p(x)$.

Thus, the optimal strategy for nature is just the maximum entropy distribution $P^*\in\Gamma$ with

$H(P^*) = sup_{P\in\Gamma} H(P)$ (79')

At the same time, the code k^* is *optimal strategy* for Observer when $k^*\in K$ and

$sup_{P\in\Gamma} <k^*,P> = inf_{k\in K} sup_{P\in\Gamma} <k,P>$ (80)

Thus, the value of the game, when it exists, can be expressed as

$\alpha = sup_{P\in\Gamma} H(P) = inf_{k\in K} R(k)$, or $R_{min}(\Gamma) = H_{max}(\Gamma)$, (81)

or minimum risk equals maximum entropy, as in the equation (30) above.

Definition 4.5 A code k is cost-stable iff $<k,P>$ is finite and independent (82)

of P for $P\in\Gamma$.

We will also recall the notion of relative entropy or Kullback-Leibler distance between two probability mass functions $p(x)$ and $q(x)$, defined in (4), and repeated below

$D(p\|q) = \sum_{x\in\chi} p(x) \log (p(x)/q(x))$ (4)

Considered from the point of view of the Observer in our game, this quantity denotes the average improvement resulting from the information that a distribution that Observer believed to be Q has changed to P:

$$D(P\|Q) = \sum_{x\in\chi} p(x) \log (p(x)/q(x)) = \sum_{x\in\chi} p(x) (\log p(x) - \log q(x)) = \quad (83)$$

$$= \sum_{x\in\chi} p(x) (l_x - k_x) ,$$

where k_x and l_x are codes adapted to P and Q, respectively. Rewriting the above formula in a way that connects code length, entropy, and relative entropy, we get

$$\langle l,P\rangle = D(P\|Q) + H(P) \quad (84)$$

Lemma 4.1 Assume that Γ is convex and that $H_{max}(\Gamma) < \infty$. (85)

Then there exists a unique distribution P^* such that $P_n \to P^*$ for every sequence (P_n) such that $H(P_n) \to H_{max}(\Gamma)$.

Proof. First, for any $P_1, \ldots, P_n$ and a probability vector $Q = (Q_1, \ldots, Q_n)$, we have

$$H(\Sigma Q_i P_i) = \Sigma Q_i H(P_i) + \Sigma Q_i D(P_i\|\Sigma Q_j P_j). \quad (86)$$

This is because, taking k to be the code adapted to $v=\Sigma Q_i P_i$, from (84) we have

$$H(v)=\langle k,v\rangle=\langle k,\Sigma Q_i P_i\rangle=\Sigma Q_i\langle k,P_i\rangle=\Sigma Q_i[D(P_i\|v) + H(P_i)]= \Sigma Q_i D(P_i\|v)+ \Sigma Q_i H(P_i).$$

Then, from Pinsker's inequality (8) for every n and m we get

$$H_{max}(\Gamma)\geq H(\tfrac{1}{2}P_n+\tfrac{1}{2}P_m) = \tfrac{1}{2}H(P_n)+\tfrac{1}{2}H(P_m)+ \tfrac{1}{2}D(P_n\|\tfrac{1}{2}P_n+\tfrac{1}{2}P_m)+\tfrac{1}{2}D(P_m\|\tfrac{1}{2}P_n+\tfrac{1}{2}P_m) \geq$$

$$\geq \tfrac{1}{2}H(P_n)+\tfrac{1}{2}H(P_m)+ \tfrac{1}{4} d^2(P_n,\tfrac{1}{2}P_n+\tfrac{1}{2}P_m)+ \tfrac{1}{4} d^2(P_m,\tfrac{1}{2}P_n+\tfrac{1}{2}P_m)=$$

$$= \tfrac{1}{2}H(P_n)+\tfrac{1}{2}H(P_m)+ \tfrac{1}{2}||P_n-P_m||^2$$

Thus, when $n, m \to \infty$, we have $\| P_n - P_m \| \to 0$. So, the sequence converges to some $P^* \in \Gamma$, independent of the choice of a particular sequence (P_n). q.e.d.

Lemma 4.2 Assume that Γ is convex and that $H_{max}(\Gamma) < \infty$. (87)

Then, for every $P \in \Gamma$, the following inequality holds:

$H(P) + D(P \| P^*) \leq H_{max}(\Gamma)$

Proof. Choose a sequence $(P_n) \subseteq \Gamma$ such that $n[Hmax(\Gamma) - H(Pn)] \to 0$

For each $P \in \Gamma$, assemble a sequence of distributions $P_n^* = (1 - 1/n)P_n + 1/n\, P,$ where $n \geq 1$.

Then, since each $P_n^* \in \Gamma$, so $H(P_n^*) \leq H_{max}(\Gamma)$. Then, from (86) we have

$H(P_n^*) = (1-1/n)H(P_n) + 1/n\, H(P) + (1-1/n)\, D(P_n \| P_n^*) + 1/n\, D(P_n \| P_n^*) \geq$

$\geq (1 - 1/n)\, H(P_n) + 1/n\, H(P) + 1/n\, D(P_n \| P_n^*)$

So, we have $H(P) + D(P_n \| P_n^*) \leq n[Hmax(\Gamma) - H(P_n)] + H(P_n)$. (88)

Now, from the lower semicontinuity property of divergence (5) it follows that

$\lim_{n \to \infty} \inf D(P_n \| P_n^*) \geq D(P \| P^*)$ (89)

Combining (88) and (89) we get $H(P) + D(P \| P^*) \leq H_{max}(\Gamma)$.

In the log-loss game recast in terms of codes, the theorem connecting minimum risk and maximum entropy can be states as follows:

Theorem 4.3 Let Γ be convex with $H_{max}(\Gamma) < \infty$. Then, (90)

1. The value of the log-loss game on Γ exists and is $H_{max}(\Gamma)$,
2. The Observer has an optimal strategy code k^* adapted to the "center of attraction" distribution P^* as in the preceding lemmas.
3. This optimal strategy is unique, and
4. for any $k \in K$, $P \in \Gamma$ adapted to each other, $\sup_{P\in\Gamma} \langle k,P\rangle \geq H_{max}(\Gamma) + D(P^* \| P)$
5. Nature has an optimal strategy iff $P^* \in \Gamma$ and $H(P^*) = H_{max}(\Gamma)$

Proof. Combining (84) and Lemma 4.2 (87), we get

$$\inf_{k\in K} \sup_{P\in\Gamma} \langle k,P\rangle \leq \sup_{P\in\Gamma} \langle k^*,P\rangle = \sup_{P\in\Gamma} [H(P) + D(P \| P^*)] \leq H_{max}(\Gamma) \quad (91)$$

At the same time, as (86) states, $H_{max}(\Gamma) \leq \inf_{k\in K} \sup_{P\in\Gamma} \langle k,P\rangle$ (92)

Combining (91) and (92) we get equality throughout, showing that the value of the game exists and is $H_{max}(\Gamma)$ (condition 1) and that k^* is the optimal strategy for the observer (condition 2).

To prove condition 3, let k∈K, let v be the distribution associated with k, and let (Pn) be a sequence in Γ for which $H(P_n) \rightarrow H_{max}(\Gamma)$.

From (84), Lemma 4.1 (85), and lower semicontinuity of divergence (5), we get

$$\sup_{P\in\Gamma} \langle k,P\rangle \geq \lim_{n\rightarrow\infty} \inf [H(P_n) + D(P_n \| v)] \geq H_{max}(\Gamma) + D(P^* \| v)$$

(proving condition 4).

Now, since $k=k^*$ iff $v = P^*$ iff $D(P^* \| v)=0$, so from condition 4, it follows that

$sup_{P \in \Gamma} <k,P> > H_{max}(\Gamma)$ unless $k=k^*$. So, k^* is the unique optimal strategy.

Condition 5 follows from $<l,P> = D(P \| Q) + H(P)$ (84), which implies the following observation:

for any P (in or outside of Γ), $min_{k \in K} <k,P> = H(P)$ and the minimum is achieved for the code adapted to P. When the entropy of P is not infinite, the minimum is not achieved for any other code. But then, for $H_{max}(\Gamma) = sup_{P \in \Gamma} H(P)$, it follows that an optimal strategy for Nature in the game is the distribution $P \in \Gamma$ for which $H(P) = H_{max}(\Gamma)$ (condition 5).

REFERENCES

Barndorff-Nielsen, Ole. 1978. *Information and Exponential Families in Statistical Theory*. New York: John Wiley.

Berger, James O. 1985. *Statistical decision theory and Bayesian analysis* (2nd ed.) New York: Springer-Verlag.

Blum, Avrim and Adam Kalai. 1999. 'Universal Portfolios With and Without Transaction Costs'. *Machine Learning* 35:3, pp 193-205.

Cover, Thomas M. and Thomas, Joy A. 1991. *Elements of information theory.* New York: Wiley Interscience.

Csiszár, Imre. 1991. 'Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems'. *Annals of Statistics* 19: 2032-2066.

Good, Irving J. 1952. 'Rational decisions' *Journal of the Royal Statistical Society B*14:107-114.

Grünwald, Peter D. 1998. *The Minimum Description Length Principle and Reasoning under Uncertainty*. Ph.D. thesis, University of Amsterdam.

Grünwald, Peter D. and A. Philip Dawid. 2004. 'Game Theory, Maximum Entropy, Minimum Discrepancy, and Robust Bayesian Decision Theory'. *Annals of Statistics* 32, 1367–1433.

Jaynes, Edwin T. 1989. *Clearing up Mysteries in Maximum Entropy and Bayesian Methods*. J. Skilling (ed.). Kluwer.

Khudanpur, Sanjeev. 1999.'Properties of I-divergence'. Lecture at *Information Theoretic Methods in Statistics.* Center for Language&Speech Processing, Johns Hopkins University

http://www.clsp.jhu.edu/~sanjeev/520.447/Spring00/I-divergence-properties.pdf

Myerson, Roger B. 1991. *Game Theory: Analysis of Conflict.* Cambridge, MA: Harvard University Press

Nash, John. 1951. 'Non-cooperative games'. *The Annals of Mathematics*, 2nd Ser., Vol. 54, No. 2. (Sep., 1951): 286-295.

von Neumann, John. 1928. Zur Theorie der Gesellschaftspiele. *Mathematische Annalen* 100: 295-320.

Rapoport, Anatol. 1966. *Two-person game theory.* Ann Arbor: University of Michigan Press.

Raiffa, Howard. 1968. *Decision Analysis: Introductory Lectures on Choices under Uncertainty.* Reading, MA: Addison-Wesley.

Rasmusen, Eric. 1989. *Games and information: an introduction to game theory*. Cambridge: Cambridge University Press.

Shannon, Claude E. 1948. 'A mathematical theory of communication'. *Bell Systems Technical Journal 27*.

Topsøe, Fleming. 1979. 'Information theoretical optimization techniques'. *Kybernetika* 15, 8-27.

Topsøe, Fleming. 1993. 'Game theoretical equilibrium, maximum entropy and minimum information discrimination'. In A. Mohammad-Djafari and G. Demoments (eds.) *Maximum Entropy and Bayesian Methods*. Kluwer, 15-23.

Wald, Abraham. 1950. *Statistical Decision Functions*. Wiley, New York.