A General Minimax Result for Relative Entropy

*David Haussler**
UC Santa Cruz

December 29, 1996

University of California Technical Report UCSC-CRL-96-26
Baskin Center for Computer Science and Computer Engineering
UC Santa Cruz, CA 96064

Abstract: Suppose Nature picks a probability measure P_θ on a complete separable metric space X at random from a measurable set $\mathcal{P}_\Theta = \{P_\theta : \theta \in \Theta\}$. Then, without knowing θ, a statistician picks a measure Q on X. Finally, the statistician suffers a loss $D(P_\theta||Q)$, the relative entropy between P_θ and Q. We show that the minimax and maximin values of this game are always equal, and there is always a minimax strategy in the closure of the set of all Bayes strategies. This generalizes previous results of Gallager, and Davisson and Leon-Garcia.

Index terms: minimax theorem, minimax redundancy, minimax risk, Bayes risk, relative entropy, Kullback-Leibler divergence, density estimation, source coding, channel capacity, computational learning theory

1 Introduction

Consider a sequential estimation game in which a statistician is given n independent observations $Y_1, \ldots, Y_n$ distributed according to an unknown distribution $\tilde{P}_\theta$ chosen at random by Nature from the set $\{\tilde{P}_\theta : \theta \in \Theta\}$ according to a known prior distribution μ on Θ. For each time t between 1 and n, the statistician must produce an estimate $\hat{P}_t$ for the unknown distribution $\tilde{P}_\theta$, based on the previous $t-1$ observations $Y_1, \ldots, Y_{t-1}$. At the end of this time period, the statistician suffers a *cumulative relative entropy loss* $\sum_{t=1}^{n} D(\tilde{P}_\theta||\hat{P}_t)$, which measures the quality of the sequential estimates made. Variants of this game have been studied by several authors (see e.g. [15, 9, 3, 4, 14]). If the observations are restricted to a finite alphabet, then the actions of the statistician can be interpreted as adaptive source coding for an unknown source. For fixed θ, the average loss suffered is the redundancy [6, 3]. When we also average over the random choice of θ according to the prior μ, we get the *risk* for this game, which is then average redundancy. This risk is called *cumulative relative entropy risk* in statistics.

The product of the sequence of conditional distributions $\{\hat{P}_t(Y_t|Y_1, \ldots, Y_{t-1}) : 1 \leq t \leq n\}$ forms a joint distribution Q on the n-fold product of the space of observations. Let us call this n-fold product space X. In this way, the actions of the statistician can be interpreted as choosing a joint distribution Q on the space X. Let $P_\theta = \tilde{P}_\theta^n$ be the "true" joint distribution for the observations. Using the chain rule for relative entropy, it is seen that the risk for this game reduces to $\int D(P_\theta||Q)d\mu(\theta)$ [3]. This leads us to a simpler, more general game: Nature picks a prior μ on Θ and then picks a probability measure P_θ on a space X at random (according to μ) from a set $\{P_\theta : \theta \in \Theta\}$. Then, knowing μ but not knowing θ, a statistician picks a measure Q on X. Finally, the statistician suffers a loss $D(P_\theta||Q)$.

We show that the minimax and maximin values of this game are always equal, and there is always a minimax strategy in the closure of the set of all Bayes strategies. This generalizes results of Gallager [11], and Davisson and Leon-Garcia [7], which were restricted to the case when the observations are chosen from

*Supported by NSF grant IRI-9123692. Computer and Information Sciences, UC Santa Cruz, Santa Cruz, CA 95064. Email addresses: haussler@cse.ucsc.edu

a finite set of symbols. The proof of the general result closely follows that of Theorem 2, page 85 in [10], which is based on earlier results of Le Cam [2], with one fairly simple extension to handle the case when $\{P_\theta : \theta \in \Theta\}$ is not uniformly tight (Lemma 4 below).

In the source coding interpretation of this game, the minimax value is the capacity of the channel from Θ to X [7, 4]. A similar interpretation applies in computational learning theory, where the cumulative relative entropy risk is interpreted as the average additional loss suffered by an adaptive algorithm that predicts each observation before it arrives, based on the previous observations, as compared to an algorithm that makes predictions knowing the true distribution [12, 13]. Here, to get this interpretation, we assume that the observation at time t is predicted by the "predictive" probability distribution $\hat{P}_t$, formed by the adaptive algorithm using the previous $t-1$ observations, and that when this tth observation arrives, the loss is the negative logarithm of its probability under $\hat{P}_t$. The game has interpretations in other fields as well. For example, in mathematical finance and gambling theory, the cumulative relative entropy risk measures the expected reduction in the logarithm of compounded wealth due to lack of knowledge of the true distribution [1, 3].

2 Preliminary Definitions

We first briefly review some basic facts about probability measures on complete separable metric spaces; proofs of these can be found in e.g. [8]. Let (X, ρ) be a complete separable metric space and let $A(X)$ be the set of all probability measures defined on the σ-field generated by the open sets of X (i.e. the Borel subsets of X). For any real-valued function f on X, $||f||_\infty = \sup_x |f(x)|$, $||f||_L = \sup_{x \neq y} |f(x) - f(y)|/\rho(x,y)$, and $||f||_{BL} = ||f||_\infty + ||f||_L$. For any two probability measures $P, Q \in A(X)$, let

$$\beta(P,Q) = \sup\{|\int f dP - \int f dQ| : ||f||_{BL} \leq 1\}.$$

It is known that $\beta(P,Q)$ is a metric and $(A(X), \beta)$ is a complete separable metric space. For measures P and P_n, $n \geq 1$, we say P_n converges weakly to P, denoted $P_n \to P$, provided that

$$\int f(x) dP_n(x) \to \int f(x) dP(x)$$

for every bounded continuous real-valued function f on X. Then $P_n \to P$ iff $\beta(P_n, P) \to 0$. Finally, a set of measures $\mathcal{P} \subseteq A(X)$ is *uniformly tight* iff for every $\epsilon > 0$ there is a compact set $K \subseteq X$ such that $P(K) > 1 - \epsilon$ for all $P \in \mathcal{P}$. It is known that $\mathcal{P}$ is uniformly tight iff $\mathcal{P}$ is totally bounded with respect to the metric β.

Next, we look at how relative entropy and mutual information can be defined in the above setting. For any measures $P, Q \in A(X)$, the *relative entropy* or *Kullback-Leibler (KL) divergence* between P and Q can be defined by

$$D(P||Q) = \sup_{\{E_i\} \in \Pi(X)} \sum_{i=1}^{k} P(E_i) \log \frac{P(E_i)}{Q(E_i)},$$

where $\Pi(X)$ is the set of all finite partitions of X into Borel sets. (Throughout the paper we define $0 \log 0 = 0$.) An equivalent definition (see [17]) is

$$D(P||Q) = \int \left(\log \frac{dP(x)}{dQ(x)} \right) dP(x),$$

where dP and dQ are Radon-Nikodym derivatives with respect to a suitable dominating measure for P and Q.

Let Θ be a set and P_θ be a measure in $A(X)$ for every $\theta \in \Theta$. Each θ is called a possible "state of Nature." We assume that $\mathcal{P}_\Theta = \{P_\theta : \theta \in \Theta\}$ is a Borel subset of $A(X)$. The set Θ inherits the metric β on $A(X)$ in the obvious way: $\beta(\theta_1, \theta_2) = \beta(P_{\theta_1}, P_{\theta_2})$ for $\theta_1, \theta_2 \in \Theta$. Let $\mathcal{A}(\Theta)$ denote the set of all measures

μ defined on the Borel subsets of Θ. We will refer to these as *priors*. For any prior $\mu \in \mathcal{A}(\Theta)$, there is a corresponding measure $P_\mu \in A(X)$ defined by letting

$$P_\mu(S) = \int P_\theta(S) d\mu(\theta)$$

for all Borel sets $S \subseteq X$.

For any prior $\mu \in \mathcal{A}(\Theta)$ and any measure $Q \in A(X)$, the *cross information* between Θ under prior μ and X under measure Q is denoted by

$$I(\mu, Q) = \int D(P_\theta || Q) d\mu(\theta),$$

and the *mutual information* between Θ and X under the prior μ is defined by

$$I(\mu) = I(\mu, P_\mu).$$

3 Result

Consider the game in which first Nature chooses a state $\theta \in \Theta$ at random according to a prior $\mu \in \mathcal{A}(\Theta)$, then without knowing θ, a statistician chooses a measure $Q \in A(X)$, and finally the statistician suffers a loss $D(P_\theta || Q)$. The average loss, or *risk*, of this game for the statistician is $I(\mu, Q)$. Using the convexity of $-log(x)$, it follows easily from Jensen's inequality that

$$I(\mu) \leq I(\mu, Q)$$

for all $Q \in A(X)$. Hence for a given μ, choosing $Q = P_\mu$ is the best strategy for the statistician (the *Bayes optimal strategy*), and $I(\mu)$ is the minimal (or *Bayes*) risk.

Let $\mathcal{M}_\Theta = \{P_\mu : \mu \in \mathcal{A}(\Theta)\}$, i.e. all measures on X that can be obtained as mixtures of the P_θ measures. This is the set of all Bayes strategies for the statistician. Let $\overline{\mathcal{M}_\Theta} \subseteq A(X)$ be the closure of $\mathcal{M}_\Theta$ in the topology of weak convergence, i.e. the topology of the metric β.

Lemma 1 *If Θ is totally bounded then $\overline{\mathcal{M}_\Theta}$ is compact.*

Proof: We will use two properties of the metric β: For any $P_1, P_2, Q_1, Q_2 \in A(X)$ and $0 \leq \lambda \leq 1$,

$$\beta(\lambda P_1 + (1-\lambda)P_2, \lambda Q_1 + (1-\lambda)Q_2) \leq \lambda \beta(P_1, Q_1) + (1-\lambda)\beta(P_2, Q_2), \tag{1}$$

i.e. β is convex in both its arguments, and for any $0 \leq \lambda' \leq 1$

$$\beta(\lambda P_1 + (1-\lambda)P_2, \lambda' P_1 + (1-\lambda')P_2) \leq 2|\lambda - \lambda'|. \tag{2}$$

Each of these are easily verified.

In a complete metric space, a set is compact iff it is closed and totally bounded. Hence it suffices to show that $\overline{\mathcal{M}_\Theta}$ is totally bounded. Since Θ is totally bounded, for any $\epsilon > 0$ we can find $\Theta^\epsilon = \{\theta_1, \ldots, \theta_k\} \subseteq \Theta$ such that for all $\theta \in \Theta$ there exists $g(\theta) \in \Theta^\epsilon$ with $\beta(P_{g(\theta)}, P_\theta) \leq \epsilon$, and such that g is measurable.

Let $\mathcal{A}(\Theta^\epsilon)$ be the set of all probability mass functions on Θ^ϵ and $\mathcal{M}^\epsilon_\Theta = \{P_\mu : \mu \in \mathcal{A}(\Theta^\epsilon)\}$. Suppose $P_\mu = \int P_\theta d\mu(\theta) \in \mathcal{M}_\Theta$. Let $P = \int P_{g(\theta)} d\mu(\theta) \in \mathcal{M}^\epsilon_\Theta$. Then by the convexity of β,

$$\beta(P_\mu, P) \leq \int \beta(P_\theta, P_{g(\theta)}) d\mu(\theta) \leq \epsilon.$$

It follows that for any $P \in \mathcal{M}_\Theta$ these is a $Q \in \mathcal{M}^\epsilon_\Theta$ with $\beta(P, Q) \leq \epsilon$, and thus the same is true for $\overline{\mathcal{M}_\Theta}$. Hence, it suffices to show that $\mathcal{M}^\epsilon_\Theta$ is totally bounded. However, using (2), if $\{\lambda_i\}$ and $\{\lambda'_i\}$ are probability mass functions on Θ^ϵ with $|\lambda_i - \lambda'_i| \leq \epsilon/k$ then

$$\beta(\sum_i \lambda_i P_{\theta_i}, \sum_i \lambda'_i P_{\theta_i}) \leq \epsilon.$$

Thus since the k-dimensional simplex is totally bounded, so is $\mathcal{M}_\Theta^\epsilon$ $\Box$

We define the following values associated with the above game: The *minimax value* is defined by

$$\overline{V} = \inf_{Q \in A(X)} \sup_{\mu \in \mathcal{A}(\Theta)} I(\mu, Q).$$

The *maximin value* is defined by

$$\underline{V} = \sup_{\mu \in \mathcal{A}(\Theta)} \inf_{Q \in A(X)} I(\mu, Q).$$

The *minimax-Bayes value* is defined by

$$V^* = \inf_{Q \in \overline{\mathcal{M}_\Theta}} \sup_{\mu \in \mathcal{A}(\Theta)} I(\mu, Q).$$

This last value represents the smallest possible worst-case loss the statistician can guarantee if she restricts herself to strategies that are in the closure of the set of Bayes strategies.

Some obvious equalities are

$$\overline{V} = \inf_{Q \in A(X)} \sup_{\theta \in \Theta} D(P_\theta || Q)$$

and

$$V^* = \inf_{Q \in \overline{\mathcal{M}_\Theta}} \sup_{\theta \in \Theta} D(P_\theta || Q),$$

since for any choice of Q by the statistician, Nature maximizes the risk of the statistician by putting all the probability in the prior μ on the worst θ, and

$$\underline{V} = \sup_{\mu \in \mathcal{A}(\Theta)} I(\mu),$$

since $I(\mu)$ is the Bayes risk for prior μ, as discussed above.

The following lemma is a minor variant of the standard minimax theorem for finite Θ from [10] (see Theorem 1, page 82). We include the proof for completeness.

Lemma 2 *If Θ is finite then $V^* \leq \underline{V}$.*

Proof: Suppose $\Theta = \{\theta_1, \ldots, \theta_k\}$. Let

$$S = \{(D(P_{\theta_1} || Q), \ldots, D(P_{\theta_k} || Q)) : Q \in \overline{\mathcal{M}_\Theta}\},$$

and let $co(S)$ be the convex hull of S. By Helley's theorem (see e.g. [10], Lemma 1, page 65), for every $\mathbf{z} \in co(S)$ there exist $\mathbf{s}_1, \ldots, \mathbf{s}_{k+1} \in S$ and $\lambda_1, \ldots, \lambda_{k+1}$ with $\lambda_j \geq 0$ and $\sum_{j=1}^{k+1} \lambda_j = 1$ such that $\sum_{j=1}^{k+1} \lambda_j \mathbf{s}_j = \mathbf{z}$. Let $\mathbf{s}_j = (D(P_{\theta_1} || Q_j), \ldots, D(P_{\theta_k} || Q_j))$ for $Q_j \in \overline{\mathcal{M}_\Theta}$ and $Q = \sum_{j=1}^{k+1} \lambda_j Q_j \in \overline{\mathcal{M}_\Theta}$. By Jensen's inequality ([5]), for all θ,

$$\sum_{j=1}^{k+1} \lambda_j D(P_\theta || Q_j) \geq D(P_\theta || Q).$$

Thus for all $\mathbf{z} \in co(S)$ there exists $\mathbf{s} \in S$ with $s_i \leq z_i$ for all i.

For each real a let $L_a = \{(z_1, \ldots, \mathbf{z}_k) : \mathbf{z}_i \leq a, 1 \leq i \leq k\}$. Let $V = lub\{a : L_a \cap co(S) = \emptyset\}$. From the observations above, it follows that for every $n \geq 1$ there exists $Q_n \in \overline{\mathcal{M}_\Theta}$ such that

$$D(P_{\theta_i} || Q_n) \leq V + \frac{1}{n} \text{ for all } 1 \leq i \leq k.$$

Thus $V^* \leq V$. It suffices to show $V \leq \underline{V}$.

Let L_V^0 denote the interior of L_V. Since L_V^0 and $co(S)$ are disjoint convex sets there exists a hyperplane that separates them, i.e. there exist real p_i, $1 \leq i \leq k$, and c, such that

$$\sum_{i=1}^{k} p_i z_i \leq c \text{ for all } \mathbf{z} \in L_V^0 \tag{3}$$

and

$$\sum_{i=1}^{k} p_i z_i \geq c \text{ for all } \mathbf{z} \in co(S). \tag{4}$$

Each p_i must be nonnegative, because if $p_i < 0$ for some i then keeping $\mathbf{z} \in L_V^0$ we can let $z_i \to -\infty$ holding the other coordinates fixed, which contradicts (3). In addition, since we must have $\sum p_i > 0$, we can assume wlog that $\sum p_i = 1$ (dividing p_i and c by $\sum p_i$ if necessary).

From (3) it follows that $V = \sum_{i=1}^{k} p_i V \leq c$. From (4) it follows that for all $Q \in \overline{\mathcal{M}_\Theta}$,

$$\sum_{i=1}^{k} p_i D(P_{\theta_i}||Q) \geq c \geq V.$$

Hence

$$\underline{V} = \sup_{\mu \in \mathcal{A}(\Theta)} \inf_{Q \in \overline{\mathcal{M}_\Theta}} \sum_{i=1}^{k} \mu(\theta_i) D(P_{\theta_i}||Q) \geq \inf_{Q \in \overline{\mathcal{M}_\Theta}} \sum_{i=1}^{k} p_i D(P_{\theta_i}||Q) \geq V.$$

□

The lemma below extends the above result to the case of infinite Θ. This generalizes similar results of Gallager [11], and Davisson and Leon-Garcia [7]. As in the latter result, the proof closely follows that of Theorem 2, page 85 in [10]. Again, we include it only for completeness[1].

Before we can prove the lemma, we need a few more preliminary definitions. A real-valued function f on a topological space X is *lower semicontinuous* if for all real r, $\{x : f(x) > r\}$ is open. Any lower semicontinuous function defined on a compact set achieves its infimum on that set, and if $\mathcal{F}$ is any set of lower semicontinuous functions, then $g(x) = \sup\{f(x) : f \in \mathcal{F}\}$ is lower semicontinuous (see [10]). Finally, Posner has shown that the function $D(P||Q)$ is lower semicontinuous in both its arguments with respect to the topology of weak convergence (or equivalently, w.r.t. the β metric) [18].

Lemma 3 *If $\mathcal{P}_\Theta$ is uniformly tight then $V^* = \overline{V} = \underline{V}$, and moreover there exists a minimax strategy in $\overline{\mathcal{M}_\Theta}$, i.e. there exists $Q_0 \in \overline{\mathcal{M}_\Theta}$ such that $\overline{V} = \sup_{\mu \in \mathcal{A}(\Theta)} I(\mu, Q_0)$.*

Proof: It is obvious that $V^* \geq \overline{V}$, and it is easily verified that $\overline{V} \geq \underline{V}$. These inequalities hold for any game. Thus it suffices to show that $V^* \leq \underline{V}$ and that there is a measure Q_0 in $\overline{\mathcal{M}_\Theta}$ such that $V^* = \sup_{\mu \in \mathcal{A}(\Theta)} I(\mu, Q_0)$. The latter is equivalent to showing there is a $Q_0 \in \overline{\mathcal{M}_\Theta}$ such that $\inf_{Q \in \overline{\mathcal{M}_\Theta}} \sup_{\theta \in \Theta} D(P_\theta||Q) = \sup_{\theta \in \Theta} D(P_\theta||Q_0)$. To verify this claim, first note that if $V^* = \infty$ then any $Q \in \overline{\mathcal{M}_\Theta}$ will do for Q_0. So wlog, assume $V^* < \infty$. Since $\mathcal{P}_\Theta$ is uniformly tight, it is totally bounded, and hence $\overline{\mathcal{M}_\Theta}$ is compact by Lemma 1. Thus since $\sup_{\theta \in \Theta} D(P_\theta||Q)$ is lower semicontinuous, it achieves its minimum over $\overline{\mathcal{M}_\Theta}$ at some $Q_0 \in \overline{\mathcal{M}_\Theta}$. The claim follows.

We now show $V^* \leq \underline{V}$. Suppose $V < V^*$. For each $\theta \in \Theta$, let $S_\theta(V) = \{Q \in \overline{\mathcal{M}_\Theta} : D(P_\theta||Q) > V\}$. Since $D(P_\theta||Q)$ is lower semicontinuous in Q, $S_\theta(V)$ is an open subset of $\overline{\mathcal{M}_\Theta}$ for every θ. In addition, it is easily verified that for every $Q \in \overline{\mathcal{M}_\Theta}$ there exists a $\theta \in \Theta$ such that $D(P_\theta||Q) > V$: just find θ such that

$$D(P_\theta||Q) > \sup_\theta D(P_\theta||Q) - (V^* - V) \geq \inf_{Q \in \overline{\mathcal{M}_\Theta}} \sup_\theta D(P_\theta||Q) - (V^* - V) = V.$$

Hence $\{S_\theta(V) : \theta \in \Theta\}$ is an open cover of $\overline{\mathcal{M}_\Theta}$. Since $\overline{\mathcal{M}_\Theta}$ is compact, there exists a finite set

$$\Theta_V = \{\theta_1, \ldots, \theta_k\} \subseteq \Theta$$

such that $\{S_{\theta_1}(V), \ldots, S_{\theta_k}(V)\}$ covers $\overline{\mathcal{M}_\Theta}$. Hence for every $Q \in \overline{\mathcal{M}_\Theta}$ there exists i, $1 \leq i \leq k$, such that $D(P_{\theta_i}||Q) > V$, and hence $\max_{1 \leq i \leq k} D(P_{\theta_i}||Q) > V$. It follows that

$$\inf_{Q \in \overline{\mathcal{M}_\Theta}} \max_{1 \leq i \leq k} D(P_{\theta_i}||Q) \geq V. \tag{5}$$

[1] The key here is that the convex closure of the set of measures $\mathcal{P}_\Theta$ forms an "essentially complete class", as defined in Ferguson's text, and compactness of this set follows from Lemma 1. This, and the comments below, give us the set-up needed to apply Ferguson's theorem, which is based on [2].

Let $\mathcal{A}(\Theta_V) \subseteq \mathcal{A}(\Theta)$ be the set of all priors (probability mass functions) over Θ_V and let $\mathcal{M}_{\Theta_V} = \{P_\mu : \mu \in \mathcal{A}(\Theta_V)\}$. It follows from (5) that

$$\inf_{Q \in \mathcal{M}_{\Theta_V}} \max_{1 \le i \le k} D(P_{\theta_i}||Q) \ge V. \tag{6}$$

Since $\mathcal{M}_{\Theta_V} = \overline{\mathcal{M}_{\Theta_V}}$, by Lemma 2

$$\inf_{Q \in \mathcal{M}_{\Theta_V}} \max_{1 \le i \le k} D(P_{\theta_i}||Q) = \sup_{\mu \in \mathcal{A}(\Theta_V)} \inf_{Q \in \mathcal{M}_{\Theta_V}} I(\mu, Q) \tag{7}$$

Since the Bayes optimal strategy for the statistician is mixture derived from a prior in $\mathcal{A}(\Theta_V)$, we also have

$$\sup_{\mu \in \mathcal{A}(\Theta_V)} \inf_{Q \in \mathcal{M}_{\Theta_V}} I(\mu, Q) = \sup_{\mu \in \mathcal{A}(\Theta_V)} \inf_{Q \in A(X)} I(\mu, Q) \le \sup_{\mu \in \mathcal{A}(\Theta)} \inf_{Q \in A(X)} I(\mu, Q) = \underline{V}. \tag{8}$$

By (6), (7), and (8), it follows that for all $V < V^*$, $V \le \underline{V}$. Hence $V^* \le \underline{V}$. □

Our final lemma examines the case when $\mathcal{P}_\Theta$ is not uniformly tight.

Lemma 4 *If $\mathcal{P}_\Theta$ is not uniformly tight, then $\underline{V} = \infty$.*

Proof: Since $\mathcal{P}_\Theta$ is not uniformly tight, there exists $\epsilon > 0$ such that for every compact $K \subseteq X$, there is a measure $P \in \mathcal{P}_\Theta$ such that $P(K) \le 1 - \epsilon$. Let $\delta = \epsilon/2$. Then we can construct an infinite sequence $\{X_i\}$ of disjoint Borel subsets of X and a corresponding sequence $\{P_i\}$ of measures in $\mathcal{P}_\Theta$ such that $P_i(X_i) \ge \delta$ and $S_i = \bigcup_{j=1}^{i} X_i$ is compact for all i. The construction proceeds as follows: First let P_1 be any measure in $\mathcal{P}_\Theta$ and X_1 be any compact set such that $P_1(X_1) \ge \delta$. As mentioned above, a set of probability measures on X is uniformly tight iff it is totally bounded, so in particular, any single probability measure P in $\mathcal{P}_\Theta$ is tight in the sense that for any $\epsilon' > 0$ there exists a compact $K \subseteq X$ with $P(K) > 1 - \epsilon'$. So this first part of the construction is possible. Now, assume we have completed the ith step of the construction. Since $\mathcal{P}_\Theta$ is not uniformly tight and S_i is compact, we can find $P_{i+1} \in \mathcal{P}_\Theta$ such that $P_{i+1}(S_i) \le 1 - \epsilon$. But since P_{i+1} (by itself) is tight, we can find a compact set K with $P_{i+1}(K) > 1 - \delta$. Let $S_{i+1} = K \cup S_i$, and hence $X_{i+1} = K - S_i$. Clearly S_{i+1} is compact and $P_{i+1}(X_{i+1}) \ge \delta$. Thus by induction, the construction is possible.

Now we make a few simple claims about the relative entropy of finite distributions that are easily verified. First, assuming that log denotes the natural logarithm, then for any finite probability mass functions $\{p_i\}$ and $\{q_i\}$,

$$\sum_i p_i \log \frac{p_i}{q_i} \ge \left(\sum_{i : \log(p_i/q_i) > 0} p_i \log \frac{p_i}{q_i} \right) - \frac{1}{e}. \tag{9}$$

Second, if $\{q_i\}$ is a probability mass distribution and $\{p_i\}$ is any set of nonnegative numbers then

$$\sum_i p_i \log \frac{p_i}{q_i} \ge \left(\sum_i p_i \right) \log \sum_i p_i. \tag{10}$$

The second claim follows directly from Jensen's inequality, and is a special case of the *log sum inequality* given in [5]. The first claim follows from the fact that $x \log x \ge -1/e$ for all nonnegative x, and the observation that

$$\sum_{i : \log(p_i/q_i) < 0} p_i \log \frac{p_i}{q_i} = \sum_{i : \log(p_i/q_i) < 0} q_i \left(\frac{p_i}{q_i} \log \frac{p_i}{q_i} \right).$$

Returning to our construction, for each $n \ge 1$ let $\{E_j\}$ be the finite partition of X obtained by $E_j = X_j$, $1 \le j \le n$, and $E_{n+1} = X - S_n$, and let $Q = \frac{1}{n} \sum_{i=1}^{n} P_i$. Then using the definitions and the above claims and construction,

$$\begin{aligned} \underline{V} &= \sup_{\mu \in \mathcal{A}(\Theta)} I(\mu, P_\mu) \\ &\ge \sup_n \frac{1}{n} \sum_{i=1}^{n} D(P_i||Q) \end{aligned}$$

$$
\begin{aligned}
&\geq \sup_n \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n+1} P_i(E_j) \log \frac{P_i(E_j)}{Q(E_j)} \\
&\geq -\frac{1}{e} + \sup_n \frac{1}{n} \sum_{i=1}^{n} P_i(E_i) \log \frac{P_i(E_i)}{Q(E_i)} \\
&\geq -\frac{1}{e} + \sup_n \frac{1}{n} \left(\sum_{i=1}^{n} P_i(E_i) \right) \log \sum_{i=1}^{n} P_i(E_i) \\
&\geq -\frac{1}{e} + \sup_n \delta \log(\delta n) \\
&= \infty.
\end{aligned}
$$

□

Putting these lemmas together, we have

Theorem 1 *$V^* = \overline{V} = \underline{V}$, and moreover there exists a minimax strategy in $\overline{\mathcal{M}_\Theta}$, i.e. there exists $Q_0 \in \overline{\mathcal{M}_\Theta}$ such that $\overline{V} = \sup_{\mu \in \mathcal{A}(\Theta)} I(\mu, Q_0)$.*

Proof: If $\mathcal{P}_\Theta$ is uniformly tight the result follows from Lemma 3. Otherwise, $\underline{V} = \infty$ by Lemma 4. However, since we always have $V^* \geq \overline{V} \geq \underline{V}$, in this case these values are trivially all equal, and since the minimax value is infinite, any strategy is minimax. □

4 Discussion

Here we show that the minimax and maximin values for this game are equal, but we do not give general bounds on this value. For the general source coding/cumulative relative entropy risk case in which $\mathcal{P}_\Theta$ is a family of smooth parametric n-fold product distributions, bounds on the Bayes Risk and the minimax value of the game that hold asymptotically for large n are given in [4]. These have a long history, also described there. Bounds on these quantities in a more abstract n-fold product setting are obtained in [14], using the results of this paper. We are unaware of any general bounds for the case when the distributions in Θ are not product distributions.

In [16] it is shown that the minimax value of this game is nearly a lower bound on the loss that must be suffered by the statistician for "most" states of Nature, where "most" is defined with respect to a limit of priors that achieve the maximin bound. Related results are given in [14]. Both [4] and [16] also investigate the limiting value of this game in the above case as $n \to \infty$. It is shown in [4] that in this limit Jeffreys' prior achieves the maximin = minimax value asymptotically for smooth, parametric distributions. It would be interesting to know the structure of the corresponding "asymptotically least favorable" prior in more abstract settings.

Acknowledgements

We thank Andrew Barron, Robert Gallager and Neri Merhav for pointing out related work, and Manfred Opper and Nick Littlestone for valuable discussions.

References

[1] A. Barron and T. Cover. A bound on the financial value of information. *IEEE Trans. on Information Theory*, 34:1097–1100, 1988.

[2] L. L. Cam. An extension of Wald's theory of statistical decision functions. *Annals of Mathematical Statistics*, 26:69–81, 1955.

[3] B. Clarke and A. Barron. Information-theoretic asymptotics of Bayes methods. *IEEE Transactions on Information Theory*, 36(3):453–471, 1990.

[4] B. Clarke and A. Barron. Jefferys' prior is asymptotically least favorable under entropy risk. *J. Statistical Planning and Inference*, 41:37–60, 1994.

[5] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley, 1991.

[6] L. Davisson. Universal noisless coding. *IEEE transactions on information theory*, IT-19:783–795, 1973.

[7] L. Davisson and A. Leon-Garcia. A source matching approach to finding minimax codes. *IEEE transactions on information theory*, IT-26:166–174, 1980.

[8] R. M. Dudley. *Real Analysis and Probability*. Wadsworth, 1989.

[9] S. Y. Efroimovich. Information contained in a sequence of observations. *Problems in Information Transmission*, 15:178–189, 1980.

[10] T. Ferguson. *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, 1967.

[11] R. Gallager. Source coding with side information and universal coding. Technical Report LIDS-P-937, MIT Laboratory for Information and Decision Systems, 1979.

[12] D. Haussler and A. Barron. How well do Bayes methods work for on-line prediction of $\{+1, -1\}$ values? In *Proceedings of the Third NEC Symposium on Computation and Cognition*. SIAM, 1992.

[13] D. Haussler, M. Kearns, and R. E. Schapire. Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension. *Machine Learning*, 14(1):83–113, 1994.

[14] D. Haussler and M. Opper. Mutual information, metric entropy, and cumulative relative entropy risk. *Annals of Statistics*, 1997. to appear.

[15] I. Ibragimov and R. Hasminskii. On the information in a sample about a parameter. In *Second Int. Symp. on Information Theory*, pages 295–309, 1972.

[16] N. Merhav and M. Feder. A strong version of the redundancy-capacity theorem of universal coding. *IEEE Trans. Info Th.*, 41(3):714–, 1995.

[17] M. S. Pinsker. *Information and Information Stability of Random Variables and Processes (Transl.)*. Holden Day, 1964.

[18] E. Posner. Random coding strategies for minimum entropy. *IEEE Trans. Info. Th.*, IT-21:388–391, 1975.