Unified Approach to Optimization Techniques in Shannon Theory

Peter Harremoës and Flemming Topsøe

October 12, 2001

1 Theory

1.1 The model

To focus on the main ideas and avoid technical problems we will assume that we work with alphabets $A, B$ and $E$ of finite size. Consider an information system
\[(A, B, E, M, \Gamma, c).\]
Let $M : A \rightarrow M_1^+(B \times E)$ be a Markov kernel. Let $\Gamma$ be a subset of $M_1^+(A)$. Let $c : \Gamma \rightarrow \mathbb{R}$ be a convex function. A coding strategy $\kappa$ is a map from $B \times E$ to $[0, \infty]$ which satisfy a variant of Krafft’s inequality
\[
\sum_{b \in B} \exp \left( -\kappa (b \mid e) \right) \leq 1 \text{ for all } e \in E.
\]

The interpretation is as follows. For each $e \in E$ we code the elements in $B$ in such way that the code length of letter $b \in B$ given $e \in E$ is $\kappa (b \mid e)$. Let a 2-persons 0-sum game between Alice and Eve be defined by that Alice plays a distribution $P$ from $\Gamma$ and Eve plays a coding strategy $\kappa$. The object function should be
\[
\langle \kappa, P \rangle_e = \sum_{e \in E} M (P) (e) \langle \kappa (\cdot \mid e), M (P) (\cdot \mid e) \rangle - c (P).
\]

The cost function $c$ is introduced such that the model will cover as many examples as possible. At this state we may think of $c (P)$ as a cost for Alice of sending $P$ through the channel given by $M$. Alice wants the object function to be big and Eve wants it to be small. One interpretation of the model is that Alice sends a secret message $P$ and an eavesdropper Eve wants to extract as much information as possible. Alice wants to prevent Eve from extract information.

*Peter Harremoës is supported by a post-doctoral stipend from the Villum Kann-Rasmussen Foundation and both authors are supported by the Danish Natural Science Research Council and by INTAS (project 00-738).
Alice is only allowed to send $P \in \Gamma$, because otherwise Bob will not be able to recover the encrypted information. Define

$$H_c(P) = \inf_{\kappa} \langle \kappa, P \rangle_c$$

$$R(\kappa) = \sup_{P \in \Gamma} \langle \kappa, P \rangle_c$$

and

$$H_c(\Gamma) = \sup_{P \in \Gamma} H_c(P)$$

$$R_{\min} = \inf_{\delta} R(\kappa)$$

With these definitions one automatically has

$$H_c(\Gamma) \leq R_{\min}.$$  

The information divergence $D$ shall always mean the mean of the conditional divergence on the output side given the letter in $E$. The divergence from a distribution to a code/coding strategy shall denote the divergence from the the distribution to the distribution/kernel corresponding to the code/coding strategy.

1.2 Results

**Proposition 1** $\langle \kappa, P \rangle_c = \langle \kappa_P, P \rangle_c + D(P \parallel \kappa)$ and therefore $H_c(P) = \langle \kappa_P, P \rangle_c$.

**Proposition 2** The function $P \rightarrow H(P)$ is continuous.

**Proof.** First remark that

$$H(P) = \langle \kappa_P, P \rangle_c = \sum_{e \in E} M(P)(e) \cdot H((MP)(\cdot | e))$$

Assume $P_\lambda \rightarrow P$. Then $M(P_\lambda)(e) \rightarrow M(P)(e)$. If $M(P)(e) \neq 0$ then $$(MP_\lambda)(\cdot | e) \rightarrow (MP)(\cdot | e)$$

and therefore

$$H((MP_\lambda)(\cdot | e)) \rightarrow H((MP)(\cdot | e)).$$

If $M(P_\lambda)(e) \rightarrow M(P)(e)$ and $M(P)(e) = 0$ then

$$|M(P_\lambda)(e) \cdot H((MP_\lambda)(\cdot | e))| \leq M(P_\lambda)(e) \cdot \log |B|$$

$\rightarrow 0$

**Lemma 3** Assume that $c$ is continuous, and that $\Gamma$ is convex and compact. Then there exists an input distribution $P_{opt}$ such that

$$H_c(P) + D(P \parallel P_{opt}) \leq H_c(\Gamma).$$
Proof. The map $P \to H_c(P)$ is continuous and by compactness of $\Gamma$ there exists an input distribution $P_{\text{opt}}$ such that

$$H_c(P_{\text{opt}}) = H_c(\Gamma)$$

Let $P \in \Gamma$ be an arbitrary input distribution. By convexity we have

$$H_c(\Gamma)$$
\begin{align*}
&\geq H_c((1 - \alpha) P_{\text{opt}} + \alpha P) \\
&\geq (1 - \alpha) H_c(P_{\text{opt}}) + \alpha H_c(P) \\
&+ (1 - \alpha) D(P_{\text{opt}} \| (1 - \alpha) P_{\text{opt}} + \alpha P) + \alpha D(P \| (1 - \alpha) P_{\text{opt}} + \alpha P) \\
&\geq (1 - \alpha) H_c(\Gamma) + \alpha H_c(P) + \alpha D(P \| (1 - \alpha) P_{\text{opt}} + \alpha P)
\end{align*}

and therefore

$$0 \geq (-\alpha) H_c(\Gamma) + \alpha H_c(P) + \alpha D(P_{\text{opt}} \| (1 - \alpha) P_{\text{opt}} + \alpha P)$$

$$H_c(\Gamma) \geq H_c(P) + D(P \| (1 - \alpha) P_{\text{opt}} + \alpha P)$$

Let $\alpha \to 0$ and use lower semi continuity to obtain

$$H_c(P) + D(P \| P_{\text{opt}}) \leq H_c(\Gamma)$$

$$\langle \kappa_{P_{\text{opt}}}, P \rangle_c \leq H_c(\Gamma)$$

Theorem 4 Assume that $c$ is continuous and that $\Gamma$ is convex and compact. Then $H_c(\Gamma) < \infty$, and the value of the game exists and equals $H_c(\Gamma)$. Further Alice has an optimal input distribution $P_{\text{opt}}$ and Eve has an optimal coding strategy $\kappa_{\text{opt}}$. The trivial inequality

$$H_c(P) \leq H_c(\Gamma) \leq R(\kappa)$$

can be improved to

$$H_c(P) + D(P \| \kappa_{\text{opt}}) \leq H_c(\Gamma) \leq R(\kappa) - D(P_{\text{opt}} \| \kappa)$$

Proof. The equation $H_c(P) + D(P \| P_{\text{opt}}) = \langle \kappa_{P_{\text{opt}}}, P \rangle_c$ shows that

$$\langle \kappa_{P_{\text{opt}}}, P \rangle_c \leq H_c(\Gamma)$$

This shows that $\kappa_{P_{\text{opt}}}$ is an optimal strategy for Eve.

Further we have

$$\langle \kappa, P_{\text{opt}} \rangle_c = D(P_{\text{opt}} \| \kappa) + H_c(P_{\text{opt}})$$

$$\geq H_c(\Gamma)$$

To demonstrate the last inequality we write

$$R(\kappa) \geq \langle \kappa, P_{\text{opt}} \rangle_c$$

We remark that the inequality (1) is the condition for $\kappa_{\text{opt}}$ to be a Nash equilibrium coding strategy.

3
**Condition 5 (Kuhn-Tucker conditions)** If $\Gamma$ is compact, then a necessary and sufficient condition for $\kappa$ to be optimal is that there exists input distributions $P_1, P_2, \cdots, P_n \in \Gamma$ such that $\kappa$ is induced by a convex combination of $P_1, P_2, \cdots, P_n$ and such that $R(\kappa) \geq \langle \kappa, P_i \rangle_c$ for all $i$.

For the value of the game to exist neither continuity of $c$ nor compactness of $\Gamma$ is necessary, and one can get

**Theorem 6** Assume that $\Gamma$ is convex and that $H_c(\Gamma) < \infty$. Then the value of the game exists and equals $H_c(\Gamma)$. Further Eve has an optimal coding strategy $\kappa_{opt}$.

2 Examples

2.1 Maximum Entropy Principle

First we consider the code length game where the cost function is zero and there are no side conditions. In this game $\mathcal{A} = \mathcal{B}$, and $\mathcal{M}, \mathcal{E}$ and $c$ are trivial. We arrive at a game where $H_c(\Gamma)$ is the maximum entropy (if it exists). This game was considered in great detail in [3].

2.2 Minimal free energy

In thermodynamics one defines Helmholtz’ free energy $A = U - TS$ where $U$ is the inner energy, $T$ is the absolute temperature and $S$ is the thermodynamic entropy. In a physical or chemical system where the volume and temperature are kept fixed Helmholtz’ free energy will tend to a minimum. The thermodynamic entropy $S$ is related to the information theoretic entropy $H$ by the relation $S = nkH$ where $n$ is the number of molecules and $k$ is Boltzmann’s constant. Now, consider a system where the molecules can be in states with energies $E_1$, $E_2$, $\ldots$, $E_n$, and let $p_i$ denote the probability that a molecule is in state $E_i$. Then we have to minimize

$$n \sum p_i E_i - nkTH(p_1, p_2, \ldots, p_n)$$

or, equivalently maximize

$$H(p_1, p_2, \ldots, p_n) - \sum p_i \cdot \frac{E_i}{kT}.$$ 

This corresponds to a game with $\mathcal{A} = \mathcal{B}$, and $\mathcal{M}, \mathcal{E}$ are trivial, and the cost function $c$ is given by $c(p) = \sum p_i \cdot \frac{E_i}{kT}$. Thus the cost for the system to use state $i$ is proportional to the energy $E_i$ of the state. The Kuhn-Tucker conditions show that an optimal code is obtained if $\kappa(i) = \frac{E_i}{kT} + \gamma$ where $\gamma$ is a suitable chosen constant. This is in agreement with the solution found in textbooks [5], where the distribution with minimal free energy is shown to have point probabilities proportional to $\exp \left( -\frac{E_i}{kT} \right)$. 

4
2.3 Minimum Information Principle

Again we consider a game with no side conditions. In this game $A = B$, and $M$, $E$ are trivial. A reference distribution $Q$ in $M^1_+$ ($A$) is given and the cost function is given by $c(P) = \left< \log \frac{1}{Q}, P \right>$. The the object function is $\left< \kappa (i) - \log \frac{1}{Q(i)}, P \right>$ and we get the relative game described in [6]. If $\Gamma$ is convex and compact the optimal strategy for Alice is the information projection of $Q$ on $\Gamma$, and the optimal strategy of Eve is the corresponding code. This amounts to the minimum information principle. The Kuhn-Tucker conditions leads naturally to the exponential families well-known from statistics.

2.4 Discrete Memoryless Channel

To model a discrete memoryless channel without side information, put $\Gamma = M^1_+ (A)$ and $E$ trivial. $M$ is the Markov kernel defining the channel as usual. The cost function is given by $c(P) = H (M (P))$. Then the equilibrium of the game is the Gallager-Ryabko Theorem which states that maximal transmission rate equals minimal redundancy. The conditions 5 are the well-known Kuhn-Tucker conditions for DMC’s [2, p. 191].

If $E$ is non trivial this corresponds to having a channel with side conditions. The model actually also covers rate distortion theory as described in [1].

2.5 Exact Prediction

Some problems in prediction theory can be solved exactly. Here we shall only consider the simplest prediction problem where we have to guess the next letter in a sequence of letters from an alphabet $B = \{a, b\}$. We consider a sequence of length 2, and assume that first and second letter are independent and identically distributed. We have to specify conditional probabilities $P (a | a), P (b | a), P (a | b)$ and $P (b | b)$, and we will measure the performance of the predictor $P$ by the supremum of

$$D (Q \parallel P) = \sum Q (i, j) \log \frac{Q(i | j)}{P(i | J)}$$

where the supremum is taken over all i.i.d. over $B^2$. To use our general model put $E = B$, $A = [0, 1]$ and $\Gamma = M^1_+ (A)$. For $q \in [0, 1]$ $M$ is given by

$$M (a, a) = q^2, M (a, b) = q (1 - q), M (b, a) = (1 - q) q, M (a, a) = (1 - q)^2$$

5
If $Q$ is given by a number $q = Q(a)$ then

$$D(Q \parallel P) = q^2 \log \frac{q}{P(a | a)} + q (1 - q) \log \frac{q}{P(a | b)} + (1 - q) \log \frac{1 - q}{P(b | a)} + (1 - q)^2 \log \frac{1 - q}{P(b | b)}$$

$$= \left\langle \log \frac{1}{P(i | j)} \right\rangle Q - H(q, 1 - q) = \left\langle \log \frac{1}{P(i | j)} \right\rangle + c(q)$$

where $c(q) = H(q, 1 - q)$. We state that the predictor given by

$$P(a | a) = \frac{4}{5}, P(b | a) = \frac{1}{5}, P(a | b) = \frac{1}{5}, P(b | b) = \frac{4}{5}$$

is optimal. For this predictor

$$D(Q \parallel P) = \left( q^2 + (1 - q)^2 \right) \log \frac{1}{4} + \log 5 - H(q, 1 - q) \leq \log \frac{5}{4}$$

with equality for $q \in \{0, \frac{1}{4}, 1\}$. For the inequality see [4]. Using the Kuhn-Tucker conditions we just have to show that $P$ is induced by a mixture of probability distributions $Q$ with $q = 0, q = \frac{1}{4}$ and $q = 1$.

References


