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 $\mathbb{A}, \mathbb{B}$ and $\mathbb{E}$ of finite size. Consider an information system $(\mathbb{A}, \mathbb{B}, \mathbb{E}, \mathcal{M}, \Gamma, c)$. Let $\mathcal{M} : \mathbb{A} \to \mathbb{M}_1^+ (\mathbb{B} \times \mathbb{E})$ be a Markov kernel. Let Γ be a subset of $M_1^+ (\mathbb{A})$. Let $c : \Gamma \to \mathbb{R}$ be a convex function. A coding strategy κ is a map from $\mathbb{B} \times \mathbb{E}$ to $[0; \infty]$ which satisfy a variant of Krafft's inequality

$$\sum_{b \in \mathbb{B}} \exp \left(-\kappa \left(b \mid e\right)\right) \leq 1 \text{ for all } e \in \mathbb{E} .$$

The interpretation is as follows. For each $e \in \mathbb{E}$ we code the elements in $\mathbb{B}$ in such way that the code length of letter $b \in \mathbb{B}$ given $e \in \mathbb{E}$ is $\kappa \left(b \mid e\right)$. Let a 2-persons 0-sum game between Alice and Eve be defined by that Alice plays a distribution P from Γ and Eve plays a coding strategy κ. The object function should be

$$\left\langle \kappa, P \right\rangle_c = \sum_{e \in \mathbb{E}} \mathcal{M} \left(P\right) \left(e\right) \left\langle \kappa \left(\cdot \mid e\right), \mathcal{M} \left(P\right) \left(\cdot \mid e\right) \right\rangle - c \left(P\right) \ .$$

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 $\mathcal{M}$. 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-dos stipend from the Villum Kann-Rasmussen Foundation and both authors are supported by the Danish Natural Science Research Counsil 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

$$H_c(\Gamma) = \sup_{P \in \Gamma} H_c(P)$$
$$R_{\min} = \inf_{\kappa} R(\kappa)$$

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 $\mathbb{E}$. The divergence from a distribution to a code/ciding strategy shall denote the divergence from the the distribution to the distribution/kernel corresponind 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* $P \to H(P)$ *is continuous.*

Proof. First remark that

$$H(P) = \langle \kappa_P, P \rangle$$
$$= \sum_{e \in \mathbb{E}} \mathcal{M}(P)(e) \cdot H((\mathcal{M}P)(\cdot \mid e))$$

Assume $P_\lambda \to P$. Then $\mathcal{M}(P_\lambda)(e) \to \mathcal{M}(P)(e)$. If $\mathcal{M}(P)(e) \neq 0$ then $(\mathcal{M}P_\lambda)(\cdot \mid e) \to (\mathcal{M}P_\lambda)(\cdot \mid e)$ and therefore

$$H((\mathcal{M}P_\lambda)(\cdot \mid e)) \to H((\mathcal{M}P)(\cdot \mid e)) .$$

If $\mathcal{M}(P_\lambda)(e) \to \mathcal{M}(P)(e)$ and $\mathcal{M}(P)(e) = 0$ then

$$|\mathcal{M}(P_\lambda)(e) \cdot H((\mathcal{M}P_\lambda)(\cdot \mid e))| \leq \mathcal{M}(P_\lambda)(e) \cdot \log |B|$$
$$\to 0$$

■

Lemma 3 *Assume that* c *is continuous, and that* Γ *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 Γ there exists an input distribution P_{opt} such that

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

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

$$\begin{aligned}
&H_c(\Gamma) \\
&\geq H_c((1-\alpha) P_{opt} + \alpha P) \\
&\geq (1-\alpha) H_c(P_{opt}) + \alpha H_c(P) \\
&+ (1-\alpha) D(P_{opt} \parallel (1-\alpha) P_{opt} + \alpha P) + \alpha D(P \parallel (1-\alpha) P_{opt} + \alpha P) \\
&\geq (1-\alpha) H_c(\Gamma) + \alpha H_c(P) + \alpha D(P \parallel (1-\alpha) P_{opt} + \alpha P)
\end{aligned}$$

and therefore

$$\begin{aligned}
0 &\geq (-\alpha) H_c(\Gamma) + \alpha H_c(P) + \alpha D(P_{opt} \parallel (1-\alpha) P_{opt} + \alpha P) \\
H_c(\Gamma) &\geq H_c(P) + D(P \parallel (1-\alpha) P_{opt} + \alpha P)
\end{aligned}$$

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

$$\begin{aligned}
H_c(P) + D(P \parallel P_{opt}) &\leq H_c(\Gamma) \\
\langle \kappa_{P_{opt}}, P \rangle_c &\leq H_c(\Gamma)
\end{aligned}$$

■

Theorem 4 *Assume that c is continuous and that Γ 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_{opt} and Eve has an optimal coding strategy κ_{opt}. The trivial inequality*

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

can be improved to

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

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

$$\langle \kappa_{P_{opt}}, P \rangle_c \leq H_c(\Gamma) \tag{1}$$

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

Further we have

$$\begin{aligned}
\langle \kappa, P_{opt} \rangle_c &= D(P_{opt} \parallel \kappa) + H_c(P_{opt}) \\
&\geq H_c(\Gamma)
\end{aligned}$$

To demonstrate the last inequality we write

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

■

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

Condition 5 (Kuhn-Tucker conditions) *If Γ is compact, then a necessary and sufficient condition for κ to be optimal is that there exists input distributions $P_1, P_2, \cdots, P_n \in \Gamma$ such that κ 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 Γ is necessary, and one can get

Theorem 6 *Assume that Γ 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 κ_{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 $\mathbb{A} = \mathbb{B}$, and $\mathcal{M}$,$\mathbb{E}$ 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 Helmholz' 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 Helmholz' 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 Bolzmann's constant. Now, consider a system where the molecules can be in states with energies E_1, E_2,...,E_n , and let p_i denote the probability that a molecule is in state E_i . Then we have to minimize

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

or, equivalently maximize

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

This corresponds to a game with $\mathbb{A} = \mathbb{B}$, and $\mathcal{M}$,$\mathbb{E}$ 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 shows that an optimal code is obtained if $\kappa(i) = \frac{E_i}{kT} + \gamma$ where γ 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 $\exp\left(-\frac{E_i}{kT}\right)$.

2.3 Minimum Information Principle

Again we consider a game with no side conditions. In this game $\mathbb{A} = \mathbb{B}$, and $\mathcal{M}$, $\mathbb{E}$ are trivial. A reference distribution Q in $M_+^1(\mathbb{A})$ is given and the cost function is given by $c(P) = \left\langle \log \frac{1}{Q(i)}, P \right\rangle$. The the object function is $\left\langle \kappa(i) - \log \frac{1}{Q(i)}, P \right\rangle$ and we get *the relative game* described in [6]. If Γ is convex and compact the optimal strategy for Alice is the information projection of Q on Γ, 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 $\Gamma = M_1^+(\mathbb{A})$ and $\mathbb{E}$ trivial. M is the Markov kernel defining the channel as usual. The cost function is given by $c(P) = H(\mathcal{M}(P))$. 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 $\mathbb{E}$ 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 $\mathbb{B} = \{a, b\}$. 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 $P(a \mid a)$, $P(b \mid a)$, $P(a \mid b)$ and $P(b \mid b)$, and we will measure the performance of the predictor P by the supremum of

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

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

$$\mathcal{M}(a,a) = q^2, \mathcal{M}(a,b) = q(1-q), \mathcal{M}(b,a) = (1-q)q, \mathcal{M}(a,a) = (1-q)^2$$

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

$$\begin{aligned}
&D(Q \parallel P) \\
&= q^2 \log \frac{q}{P(a \mid a)} + q(1-q) \log \frac{q}{P(a \mid b)} \\
&+ (1-q) q \log \frac{1-q}{P(b \mid a)} + (1-q)^2 \log \frac{1-q}{P(b \mid b)} \\
&= \left\langle \log \frac{1}{P(i \mid j)}, Q \right\rangle - H(q, 1-q) = \left\langle \log \frac{1}{P(i \mid j)}, Q \right\rangle + c(q)
\end{aligned}$$

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

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

is optimal. For this predictor

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

with equality for $q \in \left\{0, \frac{1}{2}, 1\right\}$. 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 $q = 0$, $q = \frac{1}{2}$ and $q = 1$.

References

[1] M. Chiang and T. Cover. Unified duality between channel capacity and rate distortion with state information. In *Proceedings 2001 IEEE Int. Symp. on Inform. Theory*, page 301, 2001.

[2] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, New York, 1991.

[3] P. Harremoës and F. Topsøe. Maximum entropy fundamentals. *Entropy*, (3):191 – 226, Sept. 2001. ONLINE http://www.mdi.org/entropy/.

[4] Peter Harremoës and Flemming Topsøe. Information diagrams: Entropy, index of coincidence and probability of error. In *Proceedings IEEE International Symposium on Information Theory*, page 43. IEEE, June 2001.

[5] P. T. Landsberg. *Thermodynamics and Statistical Mechanics*. Dover, New York, 1990.

[6] Flemming Topsøe. Information theoretic optimization technics. *Kybernetika*, 15(1), 1979.