Most Informative Quantization Functions

V. Chandar (MIT Lincoln) and A. Tchamkerten (Telecom ParisTech)

***Abstract*—This note provides some insights to an intriguing quantization problem recently posed by Kumar and Courtade. Any feedback is welcome.**

I. The problem

Throughout this note X^n denotes a randomly and uniformly chosen vector in $\{0,1\}^n$ and Y^n denotes a random observation of X^n through a binary symmetric channel with crossover probability $p \in [0, 1/2)$. Given an integer $k \geq 1$ we want to find the k-bit quantization of X^n which induces the largest mutual information with Y^n, that is we are interested in

$$I(k,n) \overset{\text{def}}{=} \max_{f \in \mathcal{S}_{n,k}} I(f(X^n); Y^n)$$

where

$$\mathcal{S}_{n,k} \overset{\text{def}}{=} \{f : \{0,1\}^n \to \{0,1\}^k\}.$$

This problem was posed by Kumar and Courtade in [1] for $k = 1$, where it is conjectured that

$$I(1,n) = 1 - h(p)$$

for any $n \geq 1$, and $h(p) \overset{\text{def}}{=} -p \log p - (1-p) \log(1-p)$ where the logarithm is to the base 2. If the conjecture is true, then any binary function of the form $f(X^n) = X_i$, $1 \leq i \leq n$, achieves the maximum.

More generally we could ask whether $I(k,n)/k$ equals to $1 - h(p)$ for any integers $k \geq 1$ and $n \geq 1$. Surprisingly perhaps, the answer turns out to be negative.

II. Properties of $I(k,n)$ and positive rates limits

We make some preliminary observations regarding the function $I(k,n)$.

Since $I(n,k) \leq k$ and since, for fixed k, $I(k,n)$ is a non-decreasing function of n (we can always extend a function defined for a given n_1 to $n_2 > n_1$ by ignoring the extra symbols $X_{n_1+1}^{n_2}$ and achieve the same mutual information) the function

$$I(k) \overset{\text{def}}{=} \lim_{n \to \infty} I(k,n)$$

is well defined for any $k \geq 1$ and also satisfies

$$I(k) = \sup_{n \geq 1} I(k,n).$$

This function is also super-additive, *i.e.*, for any $k \geq 1$ and $l \geq 1$

$$I(k+l) \geq I(k) + I(l).$$

This property reflects the fact that, given some input X_1^{2n}, we can always quantize the first n bits to k bits and the remaining n bits to l bits. Fekete's lemma then guarantees that

$$\lim_{k \to \infty} I(k)/k$$

is well defined and is equal to

$$\sup_{k \geq 1} \frac{I(k)}{k}.$$

It follows that

$$\lim_{k \to \infty} \lim_{n \to \infty} \frac{I(k,n)}{k} = \sup_{k,n} \frac{I(k,n)}{k},$$

which is lower bounded by the value of $\frac{I(k,n)}{k}$ for any k and n.

The following theorem can be attributed to Erkip and Cover [2, Section IV] and to earlier work from Witsenhausen and Wyner [3]. In the former work the problem is studied in the context of horse race betting with Y^n interpreted as a sequence of outcomes of a two-horse race and $f(X^n)$ interpreted as side information provided to a bettor. An alternate proof that appeals to rate distortion theory is provided thereafter.

Theorem 1: For any integers $k \geq 1$ and $n \geq 1$ we have[1]

$$\frac{I(k,n)}{k} \leq \frac{1 - h(p \star h^{-1}(1-R))}{R}$$

where $R \overset{\text{def}}{=} k/n$. Moreover, for any fixed $0 < R \leq 1$

$$\lim_{n \to \infty} \frac{I(k = Rn, n)}{n} = 1 - h(p \star h^{-1}(1-R)).$$

It can be checked that the function

$$\frac{1 - h(p \star h^{-1}(1-R))}{R}$$

is a monotonically decreasing function of R, which is equal to $1 - h(p)$ for $R = 1$ and satisfies

$$\lim_{R \downarrow 0} \frac{1 - h(p \star h^{-1}(1-R))}{R} = (1-2p)^2.$$

Hence

$$\frac{I(k,n)}{k} \leq (1-2p)^2 + \varepsilon(R)$$

with $\varepsilon(R) \downarrow 0$ as $R \to 0$. Since $I(k,n)$ is non-decreasing in n we deduce the following corollary:

Corollary 1: For any integers $k \geq 1$ and $n \geq 1$ we have

$$\frac{I(k,n)}{k} \leq (1-2p)^2.$$

Moreover,

$$\lim_{k \to \infty} \lim_{n \to \infty} \frac{I(k,n)}{k} = (1-2p)^2.$$

The previous results imply that for k and n large enough $I(k,n)/n$ surpasses $1 - h(p)$, but how large should these be? A partial answer to this question is provided by the

[1] $p \star q \overset{\text{def}}{=} p(1-q) + (1-p)q$ and $h^{-1}(q)$ is defined as the value in $[0, 1/2]$ whose binary entropy is equal to q.

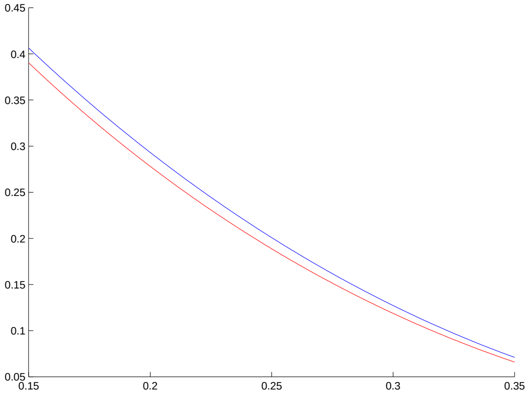

Fig. 1. The lower curve is $1 - h(p)$ and the upper curve represents the (normalized) mutual information attained from the quantization map derived from the $[n = 23, k = 12, d = 7]$ Golay code.

$[n = 15, k = 11, d = 3]$ Hamming code and the $[n = 23, k = 12, d = 7]$ Golay code which surpass the $1 - h(p)$ bound for a large range of p.[2] For the Hamming code, the range is approximately $.05 \leq p \leq .5$, and for the Golay code, the range is approximately $.04 \leq p \leq .5$ (see Fig. 1)). Both the Hamming and the Golay code are perfect codes, however not all perfect codes exceed the bound; for instance, the $[7, 4, 3]$ Hamming code does not. It would be interesting to exhibit other quantization functions, possibly derived from linear codes, which surpass the bound with values of $k < 11$.[3]

As we shall see, Theorem 1 can be extended to some other channels since the basic ingredient of its proof is a convexity argument which holds for some other channels as well.

Proof of Theorem 1: We first establish the upper bound on $I(k, n)$. Since Y^n is uniformly distributed,

$$I(f(X^n); Y^n) = n - H(Y^n|f(X^n)) \,.$$

Consider a particular k-bit string z, and assume that the preimage of z under f

$$f^{-1}(z) \overset{\text{def}}{=} \{x^n : f(x^n) = z\}$$

has size $2^{ns(z)}$. Then, because X^n is uniform, Mrs. Gerber's lemma implies that

$$H(Y^n|f(X^n) = z) \geq nh(p \star h^{-1}(s(z))) \,.$$

Therefore,

$$\begin{aligned} H(Y^n|f(X^n)) &= \sum_{z \in \{0,1\}^k} H(Y^n|f(X^n) = z) 2^{n(s(z)-1)} \\ &\geq \sum_{z \in \{0,1\}^k} nh(p \star h^{-1}(s(z)) 2^{n(s(z)-1)} \qquad (1) \end{aligned}$$

where

$$p \star q \overset{\text{def}}{=} p(1-q) + (1-p)q.$$

We now show that the above final expression is minimized when $s(z) = 1 - k/n$ for all $z \in \{0,1\}^k$, *i.e.*, when f is uniformly distributed over $\{0,1\}^k$. Specifically, the function

$$ah\left(p \star h^{-1}\left(\frac{\log a}{n}\right)\right)$$

is convex over the range $1 \leq a \leq 2^n$. This can be seen as follows. The function

$$F(x) \overset{\text{def}}{=} h\left(p \star h^{-1}(x)\right)$$

is monotonic increasing and convex over the range $0 \leq x \leq 1$ [4]. Now,

$$\begin{aligned} &\frac{d^2}{da^2} ah\left(p \star h^{-1}\left(\frac{\log a}{n}\right)\right) \\ &\qquad = \left(2\frac{\log e}{an} - \frac{\log e}{an}\right)\frac{dF}{dx} + \frac{(\log e)^2}{an^2}\frac{d^2F}{dx^2}, \end{aligned}$$

is positive since $a \geq 1$ and since both $\frac{dF}{dx}$ and $\frac{d^2F}{dx^2}$ are positive. Finally, since

$$ah\left(p \star h^{-1}\left(\frac{\log a}{n}\right)\right)$$

is convex in a, it follows that

$$\sum_{z \in \{0,1\}^k} nh(p \star h^{-1}(s(z)) 2^{n(s(z)-1)}$$

is minimized when $s(z)$ is the same for all values of z.

Hence

$$\sum_{z \in \{0,1\}^k} nh(p \star h^{-1}(s(z)) 2^{n(s(z)-1)} \geq nh(p \star h^{-1}(1 - k/n))$$

and we deduce that

$$\frac{I(k,n)}{k} \leq \frac{1 - h(p \star h^{-1}(1-R))}{R}$$

where $R \overset{\text{def}}{=} k/n$.

To prove the second part of the theorem it suffices to prove that

$$\liminf_{n \to \infty} \frac{I(k = Rn, n)}{n} \geq 1 - h(p \star h^{-1}(1-R)). \qquad (2)$$

We do this through rate-distortion theory. Specifically, we choose the function to be the encoder of a random code for the rate-distortion problem of quantizing a binary symmetric source under Hamming distortion d_H. From standard results on the rate-distortion problem [5] it is known that, for any $0 < R \leq 1$,

$$\Pr\left\{d_H(X^n, f(X^n)) > nh^{-1}(1-R) + \sqrt{3n\log(n)}\right\} = 0$$

[2] Recall that a quantization map can be derived from an (n, k) code by mapping X^n to the message sequence whose codeword is the closest to X^n.

[3] We consider perfect codes because the symmetry of the quantizer regions makes an exact computation of $I(f(X^n); Y^n)$ simpler. Specifically, $H(Y^n|f(X^n) = c)$ is independent of the value c for a perfect code, and simply corresponds to the output entropy when the input to the channel is uniform over a Hamming ball with radius equal to the error-correction radius of the code, *i.e.*, radius 1 for Hamming codes and radius 3 for the Golay code. Because of permutation symmetry, the output distribution assigns the same probability to all strings of a given Hamming weight. The output distribution can thus be characterized by $n + 1$ numbers instead of the 2^n required for an arbitrary probability distribution which allows to efficiently compute the mutual information. The expressions for the mutual information are unwieldy but straightforward to obtain and are thus omitted in this note.

where the randomness comes from both X^n and f. For such a random function f we can bound $H(Y^n|f(X^n)=c)$ for any value c as follows. Given that $f(X^n)=c$, the conditional distribution of X^n is uniform over a subset of a Hamming ball with radius $nh^{-1}(1-R)+\sqrt{3n\log(n)}$. Therefore, Y^n is concentrated on a Hamming ball with radius approximately $n(p\star h^{-1}(1-R))$. Proceeding formally, let E be the indicator random variable for the event

$$\{d_H(Y^n, f(X^n)=c) > (p\star h^{-1}(1-R)) + \sqrt{10n\log(n)}\}.$$

Then,

$$\begin{aligned} H(Y^n|f(X^n)=c) &\le H(Y^n,E|f(X^n)=c)\\ &= H(E|f(X^n)=c)\\ &\quad + \Pr\{E=1\}H(Y^n|E=1,f(X^n)=c)\\ &\quad + \Pr\{E=0\}H(Y^n|E=0,f(X^n)=c)\\ &\le 1+n\Pr\{E=1\}+H(Y^n|E=0,f(X^n)=c). \end{aligned} \tag{3}$$

From Azuma's inequality

$$\Pr\{E=1\}<\frac{1}{n}. \tag{4}$$

The term $H(Y^n|E=0,f(X^n)=c)$ is at most the logarithm of the number of strings in a Hamming ball of radius

$$n(p\star h^{-1}(1-R))+\sqrt{10n\log(n)}$$

which can be approximated using standard bounds on binomial coefficients to obtain

$$\begin{aligned} &H(Y^n|E=0,f(X^n)=c)\\ &\qquad \le nh(p\star h^{-1}(1-R))+O(\sqrt{n\log(n)}). \end{aligned} \tag{5}$$

From (3), (4), and (5) we get

$$H(Y^n|f(X^n))\le nh(p\star h^{-1}(1-R))+O(\sqrt{n\log(n)}).$$

Hence

$$\begin{aligned} I(Rn,n) &\ge I(f(X^n);Y^n)\\ &\ge n(1-h(p\star h^{-1}(1-R)))-O(\sqrt{n\log(n)}), \end{aligned} \tag{6}$$

which implies the desired inequality (2). ■

Remark 1: Observe that Theorem 1 can be extended to other channels as well. In particular, the key ingredient in the proof of the upper bound is the convexity property of the function $F(x)$. The corresponding function, for instance, for binary input symmetric channel with inputs $\{0,1\}$, is the function [6]

$$F_{P_{Y|X}}(x)\overset{\text{def}}{=}H(P_{Y|X}\star h^{-1}(x))$$

where $H(\cdot)$ denotes the usual entropy and where

$$P_{Y|X}\star h^{-1}(x)=P_{Y|x=0}(y)h^{-1}(u)+P_{Y|x=1}(y)h^{-1}(1-h^{-1}(u))$$

denotes the output distribution of the channel when

$$\mathbb{P}(X=0)=h^{-1}(u)=1-\mathbb{P}(X=1).$$

III. Suboptimality of lex functions

Conjecture 2 in [1] states that subject to a given cardinality constraint $|f^{-1}(0)|$, or equivalently, a given bias $\Pr(f(X^n)=0)$, the lexicographic function maximizes the mutual information. We show that this claim is incorrect:

Theorem 2: For fixed $0<p<1/2$ and any n large enough there exists cardinalities $|f^{-1}(0)|$ for which lexicographic functions achieve a strictly lower mutual information than suitable functions (with the same cardinality constraint).

Proof: Fix $0<\alpha<1/2$ and let

$$\ell\overset{\text{def}}{=}\lfloor nh(\alpha)\rfloor.$$

Consider the lex function $f_{lex}(X^n)$ corresponding to the set of all strings where the first ℓ bits are equal to 0, hence $|f_{lex}^{-1}(0)|=2^{n-\ell}$. Let w be the smallest value such that

$$\binom{n}{w}\ge 2^{n-\ell}.$$

Let $\mathcal{S}$ be an arbitrary subset of size $2^{n-\ell}$ of the set of strings with Hamming weight w and define the pseudo-"threshold" function f_t so that $f_t(X^n)=0$ if and only if X^n belongs to this subset. Hence we have

$$\mathbb{P}(f_{lex}(X^n)=0)=\mathbb{P}(f_t(X^n)=0)=2^{-\ell}.$$

We show that

$$H(Y^n|f_{lex}(X^n))>H(Y^n|f_t(X^n))$$

for sufficiently large n, which yields the desired result since $H(Y^n)=n$.

First we expand the entropies as

$$\begin{aligned} H(Y^n|f_{lex}(X^n)) =&2^{-\ell}H(Y^n|f_{lex}(X^n)=0)\\ &+(1-2^{-\ell})H(Y^n|f_{lex}(X^n)=1) \end{aligned} \tag{7}$$

$$\begin{aligned} H(Y^n|f_t(X^n)) =&2^{-\ell}H(Y^n|f_t(X^n)=0)\\ &+(1-2^{-\ell})H(Y^n|f_t(X^n)=1) \end{aligned} \tag{8}$$

We now lower bound $H(Y^n|f_{lex}(X^n))$. It is easy to check that

$$H(Y^n|f_{lex}(X^n)=0)=n-\ell+\ell h(p).$$

Now since

$$H(X^n|f_{lex}(X^n)=1)=n+\log(1-2^{-\ell})>n-2^{-\ell}2\log(e)$$

for sufficiently large ℓ (or large enough n), Mrs. Gerber's lemma trivially implies that

$$H(Y^n|f_{lex}(X^n)=1)\ge n-2^{-\ell}2\log(e).$$

Therefore

$$\begin{aligned} H(Y^n|f_{lex}(X^n)) \ge&2^{-\ell}\left(n-\ell+\ell h(p)\right)\\ &+\left(1-2^{-\ell}\right)\left(n-2^{-\ell}2\log(e)\right)\\ &=n-n2^{-\ell}h(\alpha)\left(1-h(p)\right)\\ &-2^{-\ell}2\log(e)(1-2^{-\ell}) \end{aligned} \tag{9}$$

for n large enough.

We now upper bound $H(Y^n|f_t(X^n))$. We upper bound $H(Y^n|f_t(X^n)=1)$ as

$$H(Y^n|f_t(X^n)=1) \leq n.$$

To upper bound $H(Y^n|f_t(X^n)=0)$, let $wt(Y^n)$ denote the weight of Y^n and let E denote the indicator function of the event

$$\{wt(Y^n) \geq \hat{w}\}$$

where

$$\hat{w} \overset{\text{def}}{=} n\left(\frac{w}{n} \star p\right) + \sqrt{2n\log_e(n)}.$$

We have

$$\begin{aligned}
&H(Y^n|f_t(X^n)=0) \\
&\quad \leq H(Y^n, E|f_t(X^n)=0) \\
&\quad = H(E|f_t(X^n)=0) + H(Y^n|E, f_t(X^n)=0) \\
&\quad = H(E|f_t(X^n)=0) \\
&\qquad\quad + \Pr(E=0)H(Y^n|E=0, f_t(X^n)=0) \\
&\qquad\quad + \Pr(E=1)H(Y^n|E=1, f_t(X^n)=0).
\end{aligned}$$

Note that when the input X^n has Hamming weight w, the expected Hamming weight of the output Y^n obtained by passing X^n through the BSC(p) is exactly $n\left(\frac{w}{n} \star p\right)$. Therefore, Azuma's inequality implies that

$$\Pr\{E=1\} \leq e^{-\log_e(n)} = \frac{1}{n}.$$

Thus, we obtain the upper bound

$$\begin{aligned}
&H(Y^n|f_t(X^n)=0) \\
&\qquad \leq 1 + H(Y^n|E=0, f_t(X^n)=0) + \frac{1}{n}n \\
&\qquad \leq 2 + \log\left(\mathrm{Vol}(n, \hat{w})\right)
\end{aligned}$$

where $\mathrm{Vol}(n, \hat{w})$ denotes the number of strings with Hamming weight at most $\hat{w}$ in the n-dimensional Hamming space. To bound $\log\left(\mathrm{Vol}(n, \hat{w})\right)$, first observe that Stirling's approximation

$$n^n e^{-n}\sqrt{2\pi n} \leq n! \leq n^n e^{-n} e\sqrt{n}$$

implies that w satisfies

$$w = nh^{-1}(1-h(\alpha)) + O(\log(n)),$$

where the constant hidden in the big-O only depends on α. Therefore, standard bounds on binomial coefficients imply that

$$\log\left(\mathrm{Vol}(n, \hat{w})\right) \leq nh\left(h^{-1}(1-h(\alpha)) \star p\right) + O\left(\sqrt{n\log(n)}\right),$$

where the constant hidden in the big-O depends only on α and p. We conclude that

$$\begin{aligned}
&H(Y^n|f_t(X^n)) \\
&\quad \leq 2^{-\ell}\left(nh\left(h^{-1}(1-h(\alpha)) \star p\right) + O\left(\sqrt{n\log(n)}\right)\right) \\
&\qquad\qquad + n(1-2^{-\ell}) \\
&\quad = n - 2^{-\ell}n\left(1 - h\left(h^{-1}(1-h(\alpha)) \star p\right)\right) \\
&\qquad\qquad\qquad + O\left(2^{-\ell}\sqrt{n\log(n)}\right).
\end{aligned}$$

Comparing the above upper bound with (9) we see that

$$H(Y^n|f_{lex}(X^n)) > H(Y^n|f_t(X^n))$$

for sufficiently large n provided that α satisfies

$$h\left(h^{-1}(1-h(\alpha)) \star p\right) < 1 - h(\alpha) + h(\alpha)h(p). \tag{10}$$

Now because of the strict convexity of

$$h(h^{-1}(q) \star p)$$

as a function of q (see, *e.g.*, Lemma 2 of [4]), it follows that the above inequality is in fact always satisfied for any $0 < \alpha < 1$ and $0 \leq p < 1/2$. Indeed, inequality (10) has the following interpretation. Consider an arbitrary distribution over X^n with entropy at least $n(1-h(\alpha))$. Then, the left-hand side of (10) is the minimum entropy of Y^n given by Mrs. Gerber's lemma, and this minimum is achieved by a distribution over X^n consisting of i.i.d. bits with bias $h^{-1}(1-h(\alpha))$. The right hand side, however, is the entropy of Y^n when X^n has a distribution such that the first $n(1-h(\alpha))$ bits are uniform, and the remaining bits are identically 0.

To summarize, we have shown that for any $0 < p < 1/2$, if n is sufficiently large (as a function of p and α), then

$$I(f_{lex}(X^n); Y^n) < I(f_t(X^n); Y^n).$$

The analysis can be refined to show that lex functions are suboptimal not just for the special case where $|f_{lex}^{-1}(0)|$ is precisely a power of two sufficiently small compared to n, but in fact are generally suboptimal for n large enough for a broad range, *e.g.*, $2^{.00001n} \leq |f_{lex}^{-1}(0)| \leq q2^n$ for some constant q independent of n. This is left as an exercise for the reader. ■

Remark 2: The above proof shows that lex functions are suboptimal by considering small enough cardinalities $|f^{-1}(0)|$. However a simple calculation using Mrs. Gerber's lemma to lower bound conditional entropy terms shows that if a function f is such that

$$|f^{-1}(0)| = b2^n,$$

then

$$I(f(X^n); Y^n) \leq H(b)(1-2p)^2.$$

This in turn places bounds on the range of the bias b where any counterexamples to the main conjecture can be found. It is easily verified that

$$\inf_p \frac{1-h(p)}{(1-2p)^2} = \frac{\log(e)}{2},$$

hence any

$$\alpha < h^{-1}\left(\frac{\log(e)}{2}\right)$$

cannot violate the conjecture.

Our previous analysis therefore only proves the suboptimality of lex functions in a regime where no function can violate the conjecture anyway. However our analysis suggests an interesting phase transition between two types of behavior; for biased functions, pseudo-threshold functions are asymptotically optimal, while for nearly balanced functions, dictator functions appear to be optimal.

IV. An Entropy Power inequality for Renyi Entropies?

We observe that the problem of maximizing mutual information is closely related to the following problem.

Problem 1: Find

$$\min H(Y_1, \ldots, Y_n)$$

subject to

$$H_\infty(X_1, \ldots, X_n) > h.$$

In Problem 1, H_∞ denotes the min-entropy,[4] and as usual Y_i is obtained by passing X_i through a BSC with crossover probability p. The relationship of Problem 1 to the original problem of maximizing the mutual information is the following. We have

$$\begin{aligned} I(f(X^n); Y^n) = H(Y^n) - \mathbb{P}(f(X^n) = 0) H(Y^n | f(X^n) = 0) \\ - \mathbb{P}(f(X^n) = 1) H(Y^n | f(X^n) = 1), \end{aligned} \tag{11}$$

and each of the conditional entropy terms can be lower bounded in terms of the solution to Problem 1 where h is set to

$$n + \log(\mathbb{P}(f(X^n) = 0))$$

and

$$n + \log(\mathbb{P}(f(X^n) = 1)),$$

respectively.

The entropy minimization in Problem 1 is reminiscent of Mrs. Gerber's lemma; if we replace the min-entropy constraint on the input distribution with a Shannon entropy constraint, then Mrs. Gerber's lemma provides the exact solution. More generally, we can consider

Problem 2:

$$\min H_\alpha(Y_1, \ldots, Y_n)$$

subject to

$$H_\beta(X_1, \ldots, X_n) > h.$$

In the statement of Problem 2, H_α and H_β denote the Renyi entropies of order α and β, respectively. Problem 1 corresponds to the case $\alpha = 1$ and $\beta = \infty$ in Problem 2. Since Mrs. Gerber's lemma, which corresponds to the $\alpha = \beta = 1$ case of Problem 2, can be viewed as the binary version of the entropy power inequality, we can interpret the entropy minimization Problem 2 as an entropy power inequality for Renyi entropies. Recall that the Renyi entropy is monotonically non-increasing with respect to the order of the entropy. Hence we can derive lower bounds on the solution to Problem 2 with $\alpha = 1$ and $\beta = \infty$ by relaxing the problem, *e.g.*, by solving Problem 2 for a larger value of α and/or a smaller value of β. A detailed study of Problem 2 for general α, β, and h is beyond the scope of this note, but we make a few preliminary observations below.

$\alpha = 1$, relax β: Mrs. Gerber's gives the solution to Problem 2 for $\alpha = \beta = 1$. Interestingly, when $\alpha = 1$, the relevant regime for the parameter h is $h > n - C$, for some constant C independent of n. The reason is that if h is small enough, then there is almost no difference between the Shannon entropy ($\beta = 1$) and the entropy of any other order β; the solution given by Mrs. Gerber's lemma is an essentially tight lower bound to the solution for general $\beta > 1$ if h is small enough. The proof is straightforward. It suffices to exhibit a distribution on $(X_1, X_2, \ldots, X_n)$ such that

$$H_\infty(X_1, \ldots, X_n) \geq h$$

and

$$H_1(Y_1, \ldots, Y_n) \approx nh(h^{-1}(\frac{h}{n}) \star p).$$

One such distribution is the uniform distribution over the strings with Hamming weight $nh^{-1}(\frac{h}{n})$. This can be shown through a similar calculation to the one used to prove that threshold functions have higher mutual information than lex functions.

Hence, if $h \leq n - C$, then the optimal input distribution is approximately given by a uniform distribution over a Hamming ball. Assuming the validity of the original Boolean conjecture that the maximum achievable mutual information is $1 - h(p)$, this suggests the same conclusion as the one at the end of Section III; for sufficiently biased functions, the most informative function is a threshold function based on the Hamming weight, while for nearly balanced functions, the most informative function is simply the first bit.

Analogously to the proof of Mrs. Gerber's lemma, another direction is to consider the function[5]

$$F_{\alpha,\beta}(x) \triangleq h_\alpha(p \star h_\beta^{-1}(x)).$$

The function $F_{1,1}$ corresponds to the function F in Mrs. Gerber's lemma. A crucial step in the proof of this lemma is to show that the function g is a convex function. Interestingly, $F_{\alpha=1,\beta}(x)$ quickly becomes concave when β increases. For instance, for $\beta > 1.6$, we have observed numerically that $F_{\alpha=1,\beta}(x)$ is concave for all $p < .49$. In particular, the case $\beta = 2$ would be of interest since Fourier techniques can potentially be brought to bear in this case.

Note that when $F_{\alpha=1,\beta}(x)$ is concave, if we add the extra constraint that the input variables X_i should be independent, then it can be shown that the optimal solution is the following: the bias of the Bernoulli distribution of each X_i is either $1/2$ (fair coin flip) or 0 (deterministic) and the number of random variables X_i with bias $1/2$ is h so as to satisfy the input entropy constraint. In particular, if $h = n - 1$ or, equivalently, $\mathbb{P}(f(X^n) = 0) = \mathbb{P}(f(X^n) = 1) = 1/2$, then there is only one deterministic X_i, equivalently, f is a dictator function.

Relax α, $\beta = 1$: A natural relaxation for α is to set $\alpha = 2$, since Fourier techniques can potentially be brought to bear in this case. Unfortunately, the best bound for Problem 2 when $\alpha = 2$ and $\beta = \infty$ is worse than the $\alpha = 1$ and $\beta = 1$ case.

[4] Given a discrete random variable X the min-entropy

$$H_\infty \stackrel{\text{def}}{=} \inf_i \log \frac{1}{p_i}$$

where $p_i \stackrel{\text{def}}{=} \mathbb{P}(X = i)$.

[5] h_β denotes the binary Renyi entropy of order β.

Specifically, consider $h = n - 1$. Then Mrs. Gerber's Lemma gives the lower bound

$$H_1(Y_1, \dots, Y_n) \geq n - (1-2p)^2.$$

By considering the uniform distribution over all strings whose first coordinate is 0, however, we find that

$$H_2(Y_1, \dots, Y_n) \leq n - 1 - \log(p^2 + (1-p)^2).$$

It can be shown that

$$n - 1 - \log(p^2 + (1-p)^2) \leq n - (1-2p)^2,$$

hence, in order to improve on the mutual information bound given by Mrs. Gerber's Lemma, we cannot relax the output entropy constraint significantly but must instead consider α close to 1.

Minimizing error probability instead of maximizing mutual information

Given the channel output Y^n suppose we wish to guess the value of a binary function $f(X^n)$. What function minimizes the error probability? In this section we show that dictator functions, *i.e.*, functions of the form $f(X^n) = X_i$, are optimal among balanced function. The arguments here closely follow the arguments that show that dictator functions maximize stability (see, *e.g.*, [7, Proposition 49, Section 2.4]).

Theorem 3: Dictator functions minimize the error probability among balanced function.

Proof: Define the Fourier transform of a function

$$f : \mathcal{Z}_2^n \to \mathcal{C}$$

as

$$\mathcal{F}(f)(x) \stackrel{\text{def}}{=} 2^{-n} \sum_{y \in \mathcal{Z}_2^n} f(y)(-1)^{x \cdot y},$$

where $x \cdot y$ denotes the inner (dot) product of x and y, viewed as $\{0, 1\}$ vectors of length n.

Now, for a binary function f taking values 0 and 1 equiprobably, it can be checked that the minimum probability of error for guessing $f(X)$ given $Y \in \mathcal{Z}_2^n$ is

$$\frac{1}{2} - \frac{||p_0 - p_1||_1}{4} \tag{12}$$

where

$$p_0(Y) \stackrel{\text{def}}{=} P(Y|f(X) = 0)$$

and

$$p_1(Y) \stackrel{\text{def}}{=} P(Y|f(X) = 1)$$

and where $|| \cdot ||_1$ denotes the ℓ_1 norm. Since the Fourier transform is unitary, the Cauchy-Schwartz inequality gives

$$\frac{||p_0 - p_1||_1}{4} \leq 2^n \frac{||\hat{p_0} - \hat{p_1}||_2}{4}. \tag{13}$$

Denote by S the subset of $\mathcal{Z}_2^n$ such that $f(x) = 1$ for any $x \in \mathcal{Z}_2^n$, and define the uniform probability distribution

$$u_f(x) = \begin{cases} \frac{1}{|S|} & x \in S \\ 0 & \text{else.} \end{cases}$$

Observe that $p_1(y)$ can be written as the convolution[6]

$$p_1(y^n) = (u_f \star v)(y)$$

where v denotes the noise distribution of the channel, *i.e.*

$$v(y) = p^{wt(y)}(1-p)^{n - wt(y)}$$

where $wt(y)$ denotes the Hamming weight of y. Hence, since

$$\hat{v}(y) = (1-2p)^{wt(y)}$$

we have

$$\hat{p_1}(y) = \hat{u_f}(y)(1-2p)^{wt(y)}.$$

Using that f is balanced, we have

$$\frac{1}{2}(p_0(y) + p_1(y)) = u(y)$$

where $u(y)$ denotes the uniform distribution over $\mathcal{Z}_2^n$. Hence,

$$\hat{p_0}(y) = 2 \cdot \hat{u}(y) - \hat{u_f}(y)(1-2p)^{wt(y)}.$$

Therefore

$$\begin{aligned} & ||\hat{p_0} - \hat{p_1}||_2^2 \\ &= \sum_{y \in \mathcal{Z}_2^n} (2 \cdot \hat{u}(y) - 2\hat{u_f}(y)(1-2p)^{wt(y)})^2 \\ &= \sum_{y \neq 0^n} 4|\hat{u_f}(y)|^2 (1-2p)^{2wt(y)} \end{aligned} \tag{14}$$

since $\hat{u_f}(0^n) = \hat{u}(0^n)$. Because the Fourier transform is unitary,

$$\sum_y |\hat{u_f}(y)|^2 = \frac{2^{-n}}{|S|} = 2^{1-2n}$$

is fixed regardless of the choice of the balanced set S. Therefore, to maximize the right-hand side of (14), f should concentrate its Fourier mass on coefficients with low Hamming weight. Because $\hat{u_f}(0^n) = 2^{-n}$ is fixed regardless of S, the best scenario is to concentrate the remaining Fourier mass

$$\sum_{y \neq 0^n} |\hat{u_f}(y)|^2 = 2^{1-2n} - 2^{-2n} = 2^{-2n}$$

on the Fourier coefficients with Hamming weight 1. Hence we deduce that

$$\sum_{y \neq 0^n} 4|\hat{u_f}(y)|^2(1-2p)^{2wt(y)} \leq 2^{-2n}(4(1-2p)^2). \tag{15}$$

From (12), (13), (14), and (15) the probability of error for guessing a balanced function f from Y^n is at least

$$\frac{1}{2} - \frac{2(1-2p)}{4} = p,$$

with equality if and only if the set S is such that the function f concentrates all of its mass of Fourier coefficients with Hamming weight 0 or 1. It is a simple exercise to see that the function

$$f(X^n) = X_i$$

has the property that the Fourier coefficients are non-zero only for the strings 0^n and e_i, the all-zero vector with a single 1 at the ith coordinate. Thus, the bound is achievable. ■

[6] $f \star g(y) \stackrel{\text{def}}{=} \sum_{x \in \mathcal{Z}_2^n} f(x)g(y \oplus x)$.

References

[1] G. R. Kumar and T. A. Courtade, “Which boolean functions are most informative?” *CoRR*, vol. abs/1302.2512, 2013.

[2] E. Erkip and T. M. Cover, “The efficiency of investment information,” *Information Theory, IEEE Transactions on*, vol. 44, no. 3, pp. 1026–1040, 1998.

[3] H. S. Witsenhausen and A. D. Wyner, “A conditional entropy bound for a pair of discrete random variables,” *Information Theory, IEEE Transactions on*, vol. 21, no. 5, pp. 493–501, 1975.

[4] A. Wyner and J. Ziv, “A theorem on the entropy of certain binary sequences and applications–i,” *Information Theory, IEEE Transactions on*, vol. 19, no. 6, pp. 769–772, Nov 1973.

[5] G. Cohen, “A nonconstructive upper bound on covering radius,” *Information Theory, IEEE Transactions on*, vol. 29, no. 3, pp. 352–353, 1983.

[6] N. Chayat and S. Shamai, “Extension of an entropy property for binary input memoryless symmetric channels,” *Information Theory, IEEE Transactions on*, vol. 35, no. 5, pp. 1077–1079, 1989.

[7] R. ODonnell, “Analysis of boolean functions,” *Lecture Notes. Available online at http://www. cs. cmu. edu/odonnell/boolean-analysis*, vol. 9, pp. 13–49, 2007.