Information Theory and Signal Processing
Gastpar, Telatar, Urbanke

Ecole Polytechnique Fédérale, Lausanne: Fall 2018
November 2, 2018

Problem Set 4 — *Due Friday, November 16, before class starts*

For the Exercise Sessions on Nov 2 and 9

Last name	First name	SCIPER Nr	Points

Problem 1: Elias coding

Let 0^n denote a sequence of n zeros. Consider the code (the subscript U a mnemonic for 'Unary'), $\mathcal{C}_U : \{1, 2, \dots\} \to \{0, 1\}^*$ for the positive integers defined as $\mathcal{C}_U(n) = 0^{n-1}$.

(a) Is $\mathcal{C}_U$ injective? Is it prefix-free?

Consider the code (the subscript B a mnenonic for 'Binary'), $\mathcal{C}_B : \{1, 2, \dots\} \to \{0, 1\}^*$ where $\mathcal{C}_B(n)$ is the binary expansion of n. I.e., $\mathcal{C}_B(1) = 1$, $\mathcal{C}_B(2) = 10$, $\mathcal{C}_B(3) = 11$, $\mathcal{C}_B(4) = 100$, ... Note that

$$\operatorname{length} \mathcal{C}_B(n) = \lceil \log_2(n + 1) \rceil = 1 + \lfloor \log_2 n \rfloor.$$

(b) Is $\mathcal{C}_B$ injective? Is it prefix-free?

With $k(n) = \operatorname{length} \mathcal{C}_B(n)$, define $\mathcal{C}_0(n) = \mathcal{C}_U(k(n))\mathcal{C}_B(n)$.

(c) Show that $\mathcal{C}_0$ is a prefix-free code for the positive integers. To do so, you may find it easier to describe how you would recover $n_1, n_2, \dots$ from the concatenation of their codewords $\mathcal{C}_0(n_1)\mathcal{C}_0(n_2)\dots$.

(d) What is $\operatorname{length}(\mathcal{C}_0(n))$?

Now consider $\mathcal{C}_1(n) = \mathcal{C}_0(k(n))\mathcal{C}_B(n)$.

(e) Show that $\mathcal{C}_1$ is a prefix-free code for the positive integers, and show that $\operatorname{length}(\mathcal{C}_1(n)) = 2 + 2\lfloor \log(1 + \lfloor \log n \rfloor) \rfloor + \lfloor \log n \rfloor \le 2 + 2 \log(1 + \log n) + \log n$.

Suppose U is a random variable taking values in the positive integers with $\Pr(U = 1) \ge \Pr(U = 2) \ge \dots$.

(f) Show that $E[\log U] \le H(U)$, [Hint: first show $i \Pr(U = i) \le 1$], and conclude that

$$E[\operatorname{length} \mathcal{C}_1(U)] \le H(U) + 2 \log(1 + H(U)) + 2.$$

Problem 2: Universal codes

Suppose we have an alphabet $\mathcal{U}$, and let Π denote the set of distributions on $\mathcal{U}$. Suppose we are given a family of S of distributions on $\mathcal{U}$, i.e., $S \subset \Pi$. For now, assume that S is finite.

Define the distribution $Q_S \in \Pi$

$$Q_S(u) = Z^{-1} \max_{P \in S} P(u)$$

where the normalizing constant $Z = Z(S) = \sum_u \max_{P \in S} P(u)$ ensures that Q_S is a distribution.

(a) Show that $D(P\|Q) \leq \log Z \leq \log |S|$ for every $P \in S$.

(b) For any S, show that there is a prefix-free code $\mathcal{C} : \mathcal{U} \to \{0,1\}^*$ such that for any random variable U with distribution $P \in S$,

$$E[\text{length}\, \mathcal{C}(U)] \leq H(U) + \log Z + 1.$$

(Note that $\mathcal{C}$ is designed on the knowledge of S alone, it cannot change on the basis of the choice of P.) [Hint: consider $L(u) = -\log_2 Q_S(u)$ as an 'almost' length function.]

(c) Now suppose that S is not necessarily finite, but there is a finite $S_0 \subset \Pi$ such that for each $u \in \mathcal{U}$, $\sup_{P \in S} P(u) \leq \max_{P \in S_0} P(u)$. Show that $Z(S) \leq |S_0|$.

Now suppose $\mathcal{U} = \{0,1\}^m$. For $\theta \in [0,1]$ and $(x_1, \ldots, x_m) \in \mathcal{U}$, let

$$P_\theta(x_1, \ldots, x_n) = \prod_i \theta^{x_i} (1-\theta)^{1-x_i}.$$

(This is a fancy way to say that the random variable $U = (X_1, \ldots, X_n)$ has i.i.d. Bernoulli θ components). Let $S = \{P_\theta : \theta \in [0,1]\}$.

(d) Show that for $u = (x_1, \ldots, x_m) \in \{0,1\}^m$

$$\max_\theta P_\theta(x_1, \ldots, x_m) = P_{k/m}(x_1, \ldots, x_m)$$

where $k = \sum_i x_i$.

(e) Show that there is a prefix-free code $\mathcal{C} : \{0,1\}^m \to \{0,1\}^*$ such that whenever $X_1, \ldots, X_n$ are i.i.d. Bernoulli,

$$\frac{1}{m} E[\text{length}\, \mathcal{C}(X_1, \ldots, X_m)] \leq H(X_1) + \frac{1 + \log_2(1+m)}{m}.$$

Problem 3: Prediction and coding

After observing a binary sequence $u_1, \ldots, u_i$, that contains $n_0(u^i)$ zeros and $n_1(u^i)$ ones, we are asked to estimate the probability that the next observation, u_{i+1} will be 0. One class of estimators are of the form

$$\hat{P}_{U_{i+1}|U^i}(0|u^i) = \frac{n_0(u^i) + \alpha}{n_0(u^i) + n_1(u^i) + 2\alpha} \quad \hat{P}_{U_{i+1}|U^i}(1|u^i) = \frac{n_1(u^i) + \alpha}{n_0(u^i) + n_1(u^i) + 2\alpha}.$$

We will consider the case $\alpha = 1/2$, this is known as the Krichevsky–Trofimov estimator. Note that for $i = 0$ we get $\hat{P}_{U_1}(0) = \hat{P}_{U_1}(1) = 1/2$.

Consider now the joint distribution $\hat{P}(u^n)$ on $\{0,1\}^n$ induced by this estimator,

$$\hat{P}(u^n) = \prod_{i=1}^{n} \hat{P}_{U_i|U^{i-1}}(u_i|u^{i-1}).$$

(a) Show, by induction on n that, for any n and any $u^n \in \{0,1\}^n$,

$$\hat{P}(u_1, \ldots, u_n) \geq \frac{1}{2\sqrt{n}} \Big(\frac{n_0}{n}\Big)^{n_0} \Big(\frac{n_1}{n}\Big)^{n_1},$$

where $n_0 = n_0(u^n)$ and $n_1 = n_1(u^n)$.

(b) Conclude that there is a prefix-free code $\mathcal{C} : \mathcal{U} \to \{0,1\}^*$ such that

$$\operatorname{length} \mathcal{C}(u_1, \ldots, u_n) \leq n h_2\left(\frac{n_0(u^n)}{n}\right) + \frac{1}{2}\log n + 2,$$

with $h_2(x) = -x \log x - (1-x)\log(1-x)$.

(c) Show that if $U_1, \ldots, U_n$ are i.i.d. Bernoulli, then

$$\frac{1}{n} E[\operatorname{length} \mathcal{C}(U_1, \ldots, U_n)] \leq H(U_1) + \frac{1}{2n}\log n + \frac{2}{n}$$

Problem 4: Lempel Ziv 78

Suppose $\ldots, U_{-1}, U_0, U_1, \ldots$ is a stationary process, i.e., for any $k = 1, 2, \ldots$, any $u_0, \ldots, u_{k-1}$, and any $n = \ldots, -1, 0, 1, \ldots$

$$\Pr(U_n \ldots U_{n+k-1} = u_0 \ldots u_{k-1}) = \Pr(U_0 \ldots U_{k-1} = u_0 \ldots u_{k-1}).$$

Suppose also that U is a recurrent process, i.e., any letter u_0 with $\Pr(U_0 = u_0) > 0$, the event $A = \{\text{there exists } i \geq 0 \text{ and } j > 0 \text{ such that } U_i = U_{-j} = u_0\}$ has $\Pr(A) = 1$. (That is, a positive probability letter u_0 will occur infinitely often.)

Fix u_0 with $\Pr(U_0 = u_0) > 0$. For $i \geq 0$ and $j < 0$, let

$$A_{ij} = \{U_i = u_0\} \cap \{U_{-j} = u_0\} \cap \bigcap_{k=-j+1}^{i-1} \{U_k \neq u_0\}$$

denote the event that j is the last time before time 0 that u_0 was seen and i was the first time after time 0 that u_0 is seen.

(a) Show that $\sum_{i \geq 0, j > 0} \Pr(A_{ij}) = \Pr(A) = 1$.

(b) Show that $\Pr(A_{ij}) = f(i + j)$, where

$$f(k) = \Pr(U_{-k} = u_0,\, U_{-l} \neq u_0 \text{ for } l = 1, \ldots, k-1, U_0 = u_0).$$

(c) Using (a) and (b), show that

$$1 = \sum_{k \geq 1} k f(k) = 1.$$

(d) Let $K = \inf\{k > 0 : U_{-k} = u_0\}$ (i.e., the negative index of the most recent time before time 0 u_0 was seen). Observe that the event $\{K = k, U_0 = u_0\}$ is the event whose probability is $f(k)$. Using (c) show that

$$E[K | U_0 = u_0] = 1/\Pr(U_0 = u_0)$$

and that $E[\log K] \leq H(U_0)$.

Suppose we have a stationary and ergodic source $\ldots, X_{-1}, X_0, X_1, \ldots$. This means, in particular, that for any $n > 0$, the process $\{U_i\}$ defined by $U_i = (X_i, X_{i+1}, X_{i+n-1})$ is stationary and recurrent.

Fix a sequence $x_0, \ldots, x_{n-1}$ with $\Pr\big((X_0 \ldots X_{n-1}) = (x_0 \ldots x_{n-1})\big) > 0$. Let

$$K = \inf\{k > 0 : (X_{-k} \ldots X_{-k+n-1}) = (x_0 \ldots x_{n-1})\}.$$

(e) Show that $E[\log K] \leq H(X_0 \ldots X_{n-1})$.

(f) Consider the following data compression method. Assuming that the encoder has already described the infinite past $\ldots, X_{-2}, X_{-1}$ to the decoder, he describes $X_0, \ldots, X_{n-1}$ by (i) finding the most recent occurrence $X_0 \ldots X_{n-1}$ in the past, (ii) describing the index K of this occurrence by the method of problem 1(f). Now that the decoder knows $\ldots, X_{n-1}$, the encoder describes $X_n \ldots X_{2n-1}$ is the same way, etc. Show that this method uses fewer than

$$\frac{1}{n} H(X_0 \ldots X_{n-1}) + \frac{2}{n} \log(1 + H(X_0 \ldots X_{n-1})) + \frac{2}{n}$$

bits per letter on the average.

Problem 5: Quantization with two criteria

Suppose U^n has i.i.d. components with distribution P. We want to describe U^n at rate R, i.e., we want to design a function $f : \mathcal{U}^n \to \{1, \dots, 2^{nR}\}$.

We are given two distortion measures $d_1 : \mathcal{U} \times \mathcal{V}_1 \to \mathbb{R}$ and $d_2 : \mathcal{U} \times \mathcal{V}_2 \to \mathbb{R}$, and we wish to ensure that from $i = f(U^n)$ we can reconstruct $V_1^n = g_1(i) \in \mathcal{V}_1^n$ and $V_2^n = g_2(i) \in \mathcal{V}_2^n$ so that

$$E[d_1(U^n, V_1^n)] \le D_1 \quad \text{and} \quad E[d_2(U^n, V_2^n)] \le D_2$$

with given distortion criteria D_1 and D_2. (As in class $d(U^n, V^n) = \frac{1}{n} \sum_{i=1}^{n} d(U_i, V_i)$.)

(a) What is the rate distortion function $R(D_1, D_2)$?

(b) Suppose $R_1(D_1)$ is the rate distortion function with the first distortion criterion alone, and $R_2(D_2)$ is the rate distortion function with the second criterion alone. What relationship exists between $R(D_1, D_2)$ and $R_1(D_1) + R_2(D_2)$?