Spring 2016

6.441 - Information Theory

Homework 4

Due: Thur, Mar 3, 2016 (in class)

`Prof. Y. Polyanskiy`

1 Reading (optional)

1. Read [1, Chapters 5,6,13]

2 Exercises

NOTE: Each exercise is 10 points. Only 3 exercises per assignment will be graded. If you submit more than 3 solved exercises please indicate which ones you want to be graded.

1 Consider the conditional distribution $P_{Y^m|X} : [0,1] \mapsto \{0,1\}^m$, where given $x \in [0,1]$, Y^m is i.i.d. Bern(x). Define the capacity:

$$C(m) \stackrel{\triangle}{=} \max_{P_X} I(X; Y^m).$$

Our goal is to show that $C(m) = \frac{1}{2}\log m + O(1)$.

1. Let $S = \sum_{i=1}^m Y_i$. Prove that $C(m) = \max_{P_X} I(X;S)$.
2. Show that
$$C(m) = \min_{Q_S} \sup_{0 \le x \le 1} D(\text{Binom}(m,x)||Q_S).$$
where Q_S is a distribution on $\{0,\dots,m\}$. *Hint:* Capacity saddle point, be sure to check conditions!
3. Choosing uniform Q_S show $C(m) \le \frac{1}{2}\log m + O(1)$ as $m \to \infty$.
4. Choosing uniform P_X show $C(m) \ge \frac{1}{2}\log m + O(1)$ as $m \to \infty$. *Hint:* Show $H(\text{Binom(n, p)}) = \frac{1}{2}\log(np(1-p)) + O(1)$ with $O(1)$ *uniform* in $p \in [0,1]$.

2 Consider a ternary fixed length (almost lossless) compression $\mathcal{X} \to \{0,1,2\}^k$ with an additional requirement that the string in $w^k \in \{0,1,2\}^k$ should satisfy

$$\sum_{j=1}^k w_j \le \frac{k}{2} \tag{1}$$

For example, $(0,0,0,0)$, $(0,0,0,2)$ and $(1,1,0,0)$ satisfy the constraint but $(0,0,1,2)$ does not.

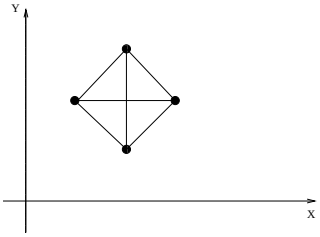
Let $\epsilon^*(S^n, k)$ denote the minimum probability of error among all possible compressors of $S^n = \{S_j, j = 1,\dots,n\}$ with i.i.d. entries of finite entropy $H(S) < \infty$. Compute

$$\lim_{n\to\infty} \epsilon^*(S^n, nR)$$

as a function of $R \ge 0$.

Hint: Relate to $\mathbb{P}[\ell(f^*(S^n)) \ge \gamma n]$ and use Stirling's formula (or Theorems 11.1.1, 11.1.3 in [1]) to find γ.

3 Consider a particle walking randomly on the graph with 4 nodes as shown below:



(each edge is taken with equal probability; particle does not stay in the same node). Alice observes the X coordinate and Bob observes the Y coordinate. How many bits per step (in the long run) does Bob need to send to Alice so that Alice will be able to reconstruct the particle's trajectory with vanishing probability of error?

4 *Mismatched compression.* Let P, Q be distributions on some discrete alphabet $\mathcal{A}$.

1. Let $f_P^* : \mathcal{A} \mapsto \{0,1\}$ denote the optimal variable-length lossless compressor for $X \sim P$. Show that under Q,
$$\mathbb{E}_Q[l(f_P^*(X))] \leq H(Q) + D(Q||P).$$
Hint: For any positive random variable U, $\mathbb{E}[U] = \int_0^\infty \mathbb{P}[U \geq u]\,\mathrm{d}u$.
2. The Shannon code for $X \sim P$ is a prefix code f_P with the code length $l(f_P(a)) = \lceil \log_2 \frac{1}{P(a)} \rceil, a \in \mathcal{A}$. Show that if X is distributed according to Q instead, then
$$H(Q) + D(Q||P) \leq \mathbb{E}_Q[l(f_P(X))] \leq H(Q) + D(Q||P) + 1\,\texttt{bit}.$$

5 *Krichevsky-Trofimov codes.* From Kraft's inequality we know that any probability distribution Q_{X^n} on $\{0,1\}^n$ gives rise to a prefix code f such that $l(f(x^n)) = \left\lceil \log_2 \frac{1}{Q_{X^n}(x^n)} \right\rceil$ for all x^n. Consider the following Q_{X^n} defined by the factorization $Q_{X^n} = Q_{X_1} Q_{X_2|X_1} \cdots Q_{X_n|X^{n-1}}$,

$$Q_{X_1}(1) = \frac{1}{2}, \quad Q_{X_{t+1}|X^t}(1|x^t) = \frac{n_1(x^t) + \frac{1}{2}}{t+1}, \tag{2}$$

where $n_1(x^t)$ denotes the number of ones in x^t. Denote the prefix code corresponding to this Q_{X^n} by $f_{\mathrm{KT}} : \{0,1\}^n \to \{0,1\}^*$.

1. Prove that for any n and any $x^n \in \{0,1\}^n$,
$$Q_{X^n}(x^n) \geq \frac{1}{2} \frac{1}{\sqrt{n_0 + n_1}} \left(\frac{n_0}{n_0 + n_1} \right)^{n_0} \left(\frac{n_1}{n_0 + n_1} \right)^{n_1}.$$
where $n_0 = n_0(x^n)$ and $n_1 = n_1(x^n)$ denote the number of zeros and ones in x^n.
Hint: Use induction on n.
2. Conclude that the K-T code length satisfies:
$$l(f_{\mathrm{KT}}(x^n)) \leq nh\left(\frac{n_1}{n}\right) + \frac{1}{2}\log n + 1, \quad \forall x^n \in \{0,1\}^n.$$
3. Conclude that for K-T codes :
$$\sup_{0 \leq \theta \leq 1} \{\mathbb{E}[l(f_{\mathrm{KT}}(S_\theta^n))] - nh(\theta)\} \leq \frac{1}{2}\log n + O(1).$$
This value is known as the *redundancy* of a universal code. It turns out that $\frac{1}{2}\log n + O(1)$ is optimal for the class of all Bernoulli sources (see lectures).

Comments:

1. The probability assignment (2) is known as the "add-$\frac{1}{2}$" estimator: Upon observing x^t which contains n_1 number of ones, a natural probability assignment to $x_{t+1} = 1$ is the empirical average $\frac{n_1}{t}$. Instead, K-T codes assign probability $\frac{n_1+\frac{1}{2}}{t+1}$, or equivalently, adding $\frac{1}{2}$ to both n_0 and n_1. This is a crucial modification to Laplace's "add-one estimator".[1]
2. By construction, the probability assignment Q_{X^n} can be sequentially computed, which allows us implement sequential encoding and encode a stream of bits on the fly. This is a highly desirable feature of the K-T codes. Of course, we need to resort to construction other than the one in Kraft's inequality construction, e.g., arithmetic coding.

6 (Combinatorial meaning of conditional entropy)

1 Fix $n \geq 1$, a sequence $x^n \in \mathcal{X}^n$ and define

$$N_{x^n}(a,b) = |\{(x_i, x_{i+1}) : x_i = a, x_{i+1} = b, i = 1, \ldots, n\}|\,,$$

where we define $x_{n+1} = x_1$ (cyclic continuation). Show that $\frac{1}{n}N_{x^n}(\cdot,\cdot)$ defines a probability distribution $P_{A,B}$ on $\mathcal{X} \times \mathcal{X}$ with equal marginals $P_A = P_B$. Conclude that $H(A|B) = H(B|A)$.

2 Let $T^{(2)}_{x^n}$ (Markov type-class of x^n) be defined as

$$T^{(2)}_{x^n} = \{\tilde{x}^n \in \mathcal{X}^n : N_{\tilde{x}^n} = N_{x^n}\}\,.$$

Show that elements of $T^{(2)}_{x^n}$ can be identified with cycles in the complete directed graph G on $\mathcal{X}$, such that for each $(a,b) \in \mathcal{X} \times \mathcal{X}$ the cycle passes $N_{x^n}(a,b)$ times through edge (a,b).

3 Show that each such cycle can be uniquely specified by indentifying the first node and by choosing at each vertex of the graph the order in which the outgoing edges are taken. From this and Stirling's approximation conclude that

$$\log |T^{(2)}_{x^n}| = nH(x_{T+1}|x_T) + O(\log n)\,, \qquad T \sim \mathrm{Unif}[n]\,.$$

Check that $H(x_{T+1}|x_T) = H(A|B) = H(B|A)$.

4 Show that for any stationary Markov chain X^n we have

$$\log P_{X^n}(X^n \in T^{(2)}_{x^n}) = -nD(P_{B|A}\|P_{X_2|X_1}|P_A) + O(\log n)\,.$$

References

[1] T. Cover and J. Thomas, *Elements of Information Theory,* Second Edition, Wiley, 2006

[1] For interesting readers, see *Laplace's rule of succession* and the sunrise problem https://en.wikipedia.org/wiki/Rule_of_succession.

6.441 Information Theory

Spring 2016