Spring 2017
STAT 364 - Information Theory
Homework 4
Due: Mar 9, 2017
`Prof. Yihong Wu`

1 Exercises

In the following all logarithms are with respect to base 2.

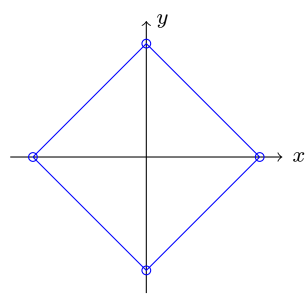
1. *Mismatched compression.* Let P, Q be distributions on some discrete alphabet $\mathcal{A}$.

   (a) Let $f_P^*: \mathcal{A} \mapsto \{0,1\}$ denote the optimal variable-length lossless compressor for $X \sim P$. Show that under Q,
   $$\mathbb{E}_Q[l(f_P^*(X))] \le H(Q) + D(Q\|P).$$

   (b) 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
   $$H(Q) + D(Q\|P) \le \mathbb{E}_Q[l(f_P(X))] \le H(Q) + D(Q\|P) + 1\,\texttt{bit}.$$

   Comments: This shows that when a compressor designed for P is applied to a source whose distribution is in fact Q, the suboptimality incurred by this mismatch can be related to divergence $D(Q\|P)$ and hence stable.

2. Draw n random points independently and uniformly from the vertices of the following square.



   Denote the coordinates by $(X_1, Y_1), \ldots, (X_n, Y_n)$. Suppose Alice only observes X^n and Bob only observes Y^n. They want to encode their observation using R_X and R_Y bits per symbol respectively and send the codewords to Charlie who will be ablve to reconstruct the sequence of pairs.

   (a) Find the optimal rate region for (R_X, R_Y).

   (b) What if the square is rotated by 45°?

3. *Arithmetic Coding.* Suppose we want to compress a sequence of English letters $X = (X_1, \ldots, X_n)$ drawn iid from some distribution P on $\{\mathsf{a}, \ldots, \mathsf{z}\}$. Let us agree upon some ordering on the alphabet (e.g. $\mathsf{a} < \mathsf{b} < \cdots < \mathsf{z}$). Then we can order the sequences lexigraphically as follows: For $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$, we say $x < y$ if $x_i < y_i$ for the first i such that $x_i \neq y_i$ (e.g., baba < babb). Define $p(x) = \prod_{i=1}^n P(x_i)$ and $q(x) = \sum_{y<x} p(y)$.

(a) Show that if $x' = (x_1, \ldots, x_{n-1})$, then

$$q(x) = q(x') + p(x') \sum_{\alpha < x_n} P(\alpha).$$

Conclude that $q(x)$ can be computed in $O(n)$ steps sequentially.

(b) Associate to each x an interval $I_x = [q(x), q(x) + p(x))$. Show that these intervals are disjoint subintervals of $[0, 1)$. Therefore each x can be represented uniquely by any point in the interval I_x.

(c) *Encoding*. Show that one can choose a rational number $c_x \in I(x)$, whose binary expansion consists of $\ell(x) = \left\lceil \log_2 \frac{1}{p(x)} \right\rceil + 1$ bits, such that the binary expansion of c_x gives us a prefix code. (*Warning*: Verify the prefix condition explicitly. This is not about checking Kraft's inequality.)

(d) *Decoding*. Upon receipt of the codeword, compute the number c_x. Divide the unit interval according to the distribution P, i.e., partition $[0, 1)$ into disjoint subintervals $I_{\mathsf{a}}, \ldots, I_{\mathsf{z}}$. Output the index that contains c_x. Show that this gives the first symbol x_1. Continue in this fashion by dividing I_{x_1} into $I_{x_1,\mathsf{a}}, \ldots, I_{x_1,\mathsf{z}}$ and etc. Argue that x can be decoded losslessly. How many steps are needed?

(e) Suppose $P_X(\mathsf{e}) = 0.5, P_X(\mathsf{o}) = 0.3, P_X(\mathsf{t}) = 0.2$. Encode etoo (write the binary codewords) and describe how to decode.

(f) Show that the average length of this code satisfies

$$nH(P) \le \mathbb{E}[l(c_X)] \le nH(P) + 2 \quad \texttt{bits}.$$

(g) Discuss the pros and cons of arithmetic coding in comparison with Huffman coding.

(h) Assume that $X = (X_1, \ldots, X_n)$ is not iid but $P_{X_1}, P_{X_2|X_1}, \ldots, P_{X_n|X^{n-1}}$ are known. How would you modify the scheme so that we have

$$H(X) \le \mathbb{E}[l(c_X)] \le H(X) + 2 \quad \texttt{bits}.$$

4. Recall the total variation distance

$$\mathrm{TV}(P, Q) \triangleq \sup_E \left(P[E] - Q[E]\right) .$$

(a) Prove that

$$\mathrm{TV}(P, Q) = \sup_{0 \le \alpha \le 1} \{\alpha - \beta_\alpha(P, Q)\}.$$

Explain how to read the value $\mathrm{TV}(P, Q)$ from the region $\mathcal{R}(P, Q)$. Does it equal half the maximal vertical segment in $\mathcal{R}(P, Q)$?

(b) (Bayesian criteria) Fix a prior $\pi = (\pi_0, \pi_1)$ such that $\pi_0 + \pi_1 = 1$ and $0 < \pi_0 < 1$. Denote the optimal average error probability by

$$P_e \triangleq \inf_{P_{Z|X^n}} \pi_0 \pi_{1|0} + \pi_1 \pi_{0|1}.$$

Prove that if $\pi = (\frac{1}{2}, \frac{1}{2})$, then

$$P_e = \frac{1}{2}(1 - \mathrm{TV}(P, Q)).$$

Find the optimal test.

(c) Find the optimal test for general prior π (not necessarily equiprobable).

(d) Why is it always sufficient to focus on deteministic test in order to minimize the Bayesian error probability?

5. (a) Consider the binary hypothesis test:

$$H_0 : X \sim \mathcal{N}(0,1) \text{ v.s. } H_1 : X \sim \mathcal{N}(\mu,1).$$

Compute the region $\mathcal{R}(\mathcal{N}(0,1), \mathcal{N}(\mu,1))$.

(b) Now suppose we have n samples and we want to test

$$H_0 : X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1) \text{ v.s. } H_1 : X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu,1).$$

Compute the region $\mathcal{R}(\mathcal{N}(0,1)^n, \mathcal{N}(\mu,1)^n)$. As the sample size increases, describe how the region evolves and provides an interpretation. *Hint: Consider sufficient statistics.*

2 Optional reading

1. Read [1, Chapters 5]

References

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