EE 376A/Stat 376A
Information Theory
Prof. T. Cover

Handout #21
Saturday, February 14, 2009

EE376A WINTER 08-09 MIDTERM EXAM REVIEW SESSION NOTES

BY WILLIAM WU

Contents

Date: Opened on February 12, 2009. Last updated on 2009-03-15 18:55.
Author's contact: `willywu AT stanford DOT edu` .

1. Outline

(1) Algebra of Information Theory (Chapter 2)

(2) AEP (Chapter 3)

(3) Entropy Rates and 2nd Law (Chapter 4)

(4) Gambling (Chapter 6)

(5) Huffman Coding (Chapter 5)

2. General Studying Tips

- Do the sample exams online.
- Do problems in the book (many problems there are from old exams).
- Go over the homework (most problems you have done already are from old exams).
- Rederive results from scratch.
- Mark up the book.
- Study with a friend. If you want help finding someone in the course to study with, you can send me your e-mail and I can help hook you up.

3. Algebra of Information Theory

Three major quantities so far: H, I, and D.

3.1. **Entropy.**

3.1.1. *Properties.*

(1) $H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x)$.

(2) Measure of uncertainty in a random variable.

(3) Only a function of the probability vector. Compare with variance.

(4) $0 \leq H(X) \leq \log m$. Maximized by uniform distribution.

(5) $H(p)$ notation. Graph of $H(p)$ vs. p. Concave in p.

(6) Conditional entropy: $H(X|Y) = \sum_{y \in \mathcal{Y}} p(y) H(X|Y = y)$.

(7) Conditioning reduces entropy: $H(X|Y) \leq H(X)$, with equality iff $X \perp\!\!\!\perp Y$.

(8) Joint entropy: just a vector version of original definition.

(9) Chain rule:

$$H(X^n) = \sum_{i=1}^{n} H(X_i|X_1^{i-1})$$

(10) $H(X) = H(f(X))$ if and only if f is a one-to-one function.

3.1.2. *Questions that $H(X)$ is The Answer To.*

(1) (20 Questions) Average number of yes-no questions to determine X.

(2) (Compression) Lower bound on the average description length of X; achievable within one bit by Huffman coding:

$$H(X) \leq \mathbf{E}[L] \leq H(X) + 1.$$

(3) (AEP) If we generate n i.i.d. random variables, the number of typical sequences is on the order of $2^{nH(X)}$. (Compare this to $|\mathcal{X}|^n = 2^{n \log |\mathcal{X}|}$.)

(4) (Kolmogorov) Apparently the entropy answers questions similar to those that Kolmogorov asked. For i.i.d. integer-valued random variables, $H \leq \frac{1}{n}\mathbf{E}[K(X^n|n)] \leq H + |\mathcal{X}|\frac{\log n}{n} + \frac{c}{n}$.

3.2. **Mutual Information.**

(1) $I(X,Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$.

(2) $I(X,Y) = D(p(x,y)||p(x)p(y))$.

(3) Venn Diagram (tells you everything!)

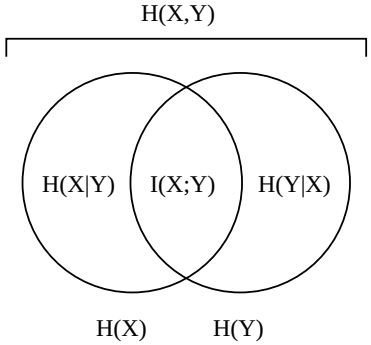


FIGURE 1. Incredibly Important Venn Diagram.

(4) $I(X;Y)$ is the reduction in the uncertainty of X, given Y.

(5) It is a measure of the dependence between X and Y.

(6) $I(X;Y) = H(X) - H(X|Y)$

(7) $I(X;Y) = H(Y) - H(Y|X)$.

(8) $I(X;Y) = H(X) + H(Y) - H(X,Y)$.

(9) $H(X) = I(X;Y) + H(X|Y)$.

(10) $I(X;Y) = I(Y;X)$.

(11) $I(X;X) = H(X)$.

(12) $I(X;Y) \geq 0$ iff $X \perp\!\!\!\perp Y$.

(13) Chain Rule: $I(X_1^n;Y) = \sum_{i=1}^n I(X_i;Y|X_1,X_2,\ldots,X_{i-1})$

(14) $I(X;Y|Z) = H(X|Z) - H(X|Y,Z)$.

3.2.1. *Questions that $I(X;Y)$ is The Answer To.*

(1) (Gambling) Increase in the doubling rate due to side information Y for horse race X.

(2) (Channel Capacity) $C = \max_{p(x)} I(X;Y)$.

3.3. **Frequently Used Techniques for Entropy and Mutual Information.** What to do when you need to prove an algebraic statement involving entropies and mutual information?

"Show that LHS = RHS".

(1) Apply chain rules directly, and work with it.

(2) Order in the arguments to $H(\cdot)$ doesn't matter.

(3) You can break up $H(X_1^n)$ in any which way you want so as to exploit constraints. Examples:

$$I(X;Q_1,A_1,Q_2,A_2) = I(X;Q_1,A_1) + I(X;Q_2,A_2|Q_1,A_1).$$

$$H(X_1,X_2,X_3,X_4) = H(X_1,X_3) + H(X_4|X_1,X_3) + H(X_2|X_1,X_3,X_4).$$

(4) Expand same expression in two different ways and set equal to each other.

(5) Starting from left-hand side (LHS), derive equivalence to the right-hand side (RHS).

(6) Starting from RHS, get to LHS.

(7) Or start somewhere in the middle, with something that is neither LHS nor RHS, and break it in different ways till you can make it to both ends.

3.4. **Inequalities.**

(1) Data-Processing: If $X \to Y \to Z$ forms a Markov Chain, then $I(X;Y) \geq I(X;Z)$. Can also write it in this useful form: $H(X|Z) \geq H(X|Y)$.

(2) Independence Bound:

$$H(X^n) \leq \sum_{i=1}^n H(X_i).$$

(3) Fano: $H(P_e) + P_e \log |\mathcal{X}| \geq H(X|Y)$

(4) Jensen: $\mathbf{E}[f(X)] \geq f(\mathbf{E}[X])$.

3.5. **Relative Entropy.**

3.5.1. *Properties.*

(1) $D(\underline{p}||\underline{q}) = \sum_x p(x) \log \frac{p(x)}{q(x)}$.

(2) Technically not a distance, but we like to think of it as one.

(3) Chain Rule: $D(p(x,y)||q(x,y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x))$.

(4) $D(\underline{p}||\underline{q}) \geq 0$ with equality iff $p(x) = q(x)$ for all $x \in \mathcal{X}$.

3.5.2. *Interpretations of $D(\underline{p}||\underline{q})$.*

(1) Has to do with the consequences of having a mismatch between your perception of the probabilities and the true probabilities.

(2) (HW5:) Suppose random variable has average description length $H(\underline{p})$. If we instead used a code corresponding to distribution $\underline{q}$, then we need $H(\underline{p})+D(\underline{p}||\underline{q})$ bits on average.

3.5.3. *"The Relative Entropy Optimization Technique".*

(1) Say you want to optimize a sum of logarithms with respect to a vector $\underline{\mathrm{b}}$.

(2) One strategy is Lagrange Multipliers – you should review and know how to do this.

(3) An alternative, more information-theoretic strategy is to play with the objective function until you have isolated the optimization variable in a term of the form $D(\underline{y}||\underline{\mathrm{b}})$, where y is *something*. Then the optimum is hopefully achieved by setting $\underline{\mathrm{b}} = \underline{y}$ so as to make $D(\underline{y}||\underline{\mathrm{b}}) = 0$.

(4) "Playing with the expression" usually includes the following tricks:

   (a) Normalizing by a constant to supply valid probability distributions to the relative entropy operator.

   (b) Multiplying and dividing by the same thing.

   (c) Using a negative sign to reciprocate the argument to a logarithm.

   (d) Using $\log(uv) = \log u + \log v$ to separate out constant terms that have nothing to do with the optimization variable of concern.

(5) Example: HW5, Problem 7. Minimize $C = \sum_{i=1}^{m} p_i c_i l_i$ over all $l_1, \ldots, l_m$ such that $\sum 2^{-l_i} = 1$.

$$\begin{aligned} C &= \sum p_i c_i l_i \\ &\overset{(c)}{=} \sum p_i c_i \log \frac{1}{r_i} \\ &\overset{(d)}{=} (\sum_j p_j c_j) \sum \frac{p_i c_i}{\sum_j p_j c_j} \log \frac{1}{r_i} \\ &= Q \sum q_i \log \frac{q_i}{r_i} \frac{1}{q_i} \\ &= Q \left(\sum q_i \log \frac{q_i}{r_i} \right) + Q \sum q_i \log \frac{1}{q_i} \\ &= Q \left(D(\underline{q} || \underline{r}) + H(\underline{q}) \right). \end{aligned}$$

Reasoning:

(a) The optimization variables are the l_i's. The goal is to get these variables into a relative entropy term.

(b) Since relative entropy requires probabilities as input, we must transform the l_i's into some variables which sum to 1.

(c) The Kraft constraint tells us that $\sum 2^{-l_i} = 1$, so that's a natural choice. Set $r_i = 2^{-l_i}$. Then $l_i = \log \frac{1}{r_i}$.

(d) The $p_i c_i$ terms must be made into probabilities. So use the normalization trick, dividing by $Q := \sum_j p_j c_j$.

Thus, to minimize, choose $\underline{r} = \underline{q}$; that is,

$$\begin{aligned} 2^{-l_i^*} = q_i &= \frac{p_i c_i}{\sum_j p_j c_j} \\ l_i^* &= -\log q_i = -\log \frac{p_i c_i}{\sum p_j c_j}. \end{aligned}$$

Then $C^* = QH(\underline{q})$.

Lastly, to show the relationship $C^* \leq C_{Huffman} \leq C^* + Q$, note that since $l_i = \lceil l_i \rceil$, we have

$$-\log q_i \leq l_i < -\log q_i + 1$$

Multiplying by $p_i c_i$ and summing over i, we get the relationship

$$C^* \leq C_{Huffman} < C^* + Q.$$

(6) Other examples that use this strategy: proving that the expected description length is lower bounded by the entropy, proving that Kelly proportional gambling is log-optimal, proving that the uniform distribution maximizes the entropy, etc.

3.6. **Summary of Nonnegativity Arguments.** A surprising number of results in this course are derived from three seemingly small and innocent facts:

(1) $D(p||q) \geq 0$.

Proof. $D(p||q) = \mathbf{E}_p\left[\log \frac{p(X)}{q(X)}\right]$. Apply Jensen's Inequality. □

(2) $I(X;Y) \geq 0$

Proof. $I(X;Y) = D(p(x,y)||p(x)p(y)) \geq 0$ by (1). □

(3) $H(X|Y) \leq H(X)$

Proof. $H(X) - H(X|Y) = I(X;Y) \geq 0$ by (2). □

So apparently everything follows from convexity.

The two pillars of Information theory are convexity (Jensen's Inequality) and the Law of Large Numbers; we will address the latter in Section 4.

3.7. **Concept Map of Ideas.** First, we established definitions for $H(X), H(X|Y), I(X;Y), I(X;Y|Z)$, and $D(p||q)$, along with chain rules for H, I, D. Then we showed the relations in Figure 2.

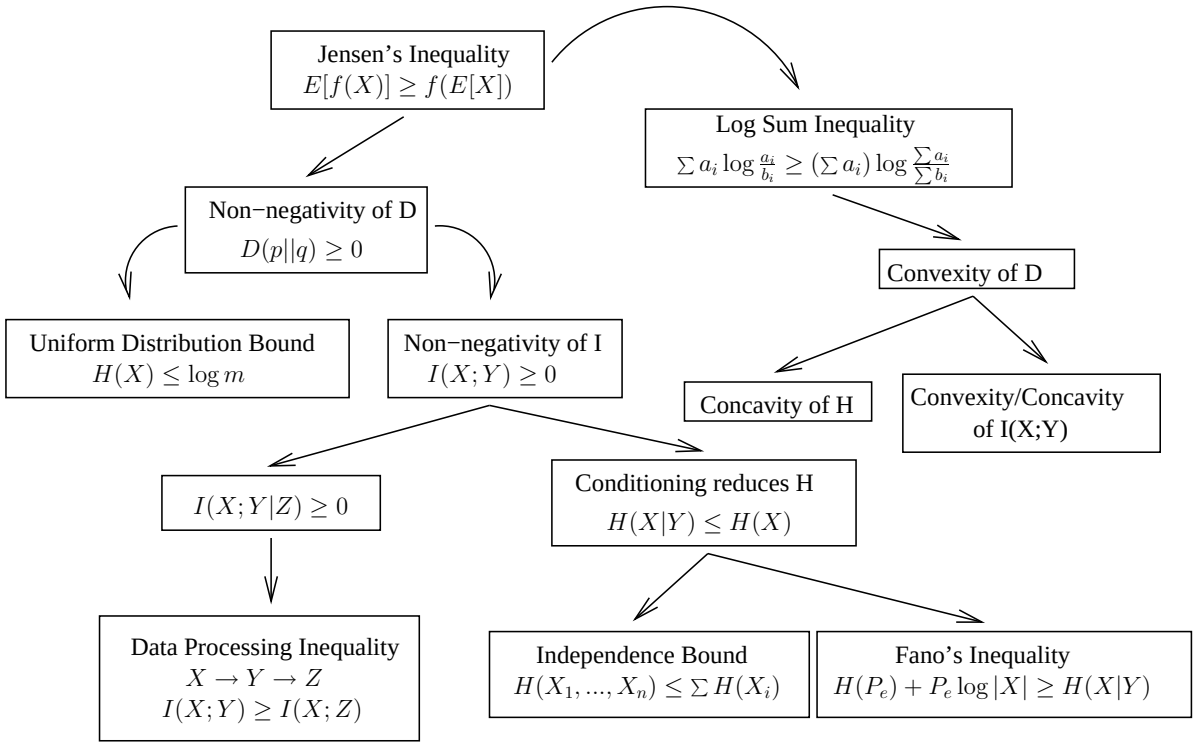


FIGURE 2. Chain of ideas in Chapter 2.

4. The AEP (Asymptotic Equipartition Property)

- Sequence of i.i.d. random variables $X_1, \ldots, X_n$.
- There are $|\mathcal{X}|^n$ possible sequences we could observe.
- Main Point #1: We can split the set of all observed sequences into two sets:

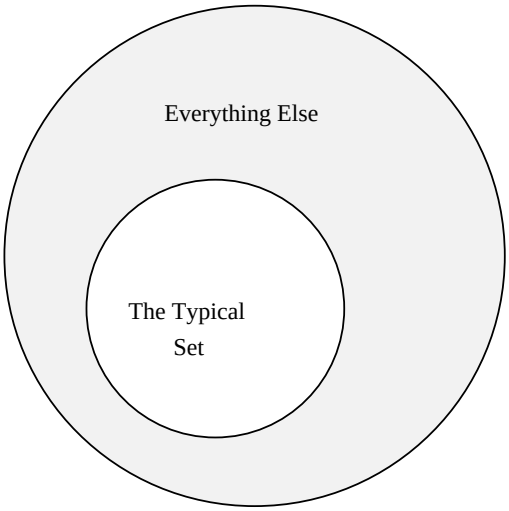


FIGURE 3. Purely Suggestive Illustration of the AEP

(1) "The Typical Set" $A_\epsilon^{(n)}$.
   - It's no larger than $2^{n(H+\epsilon)}$.
   - The probability that we see a sequence from this set approaches 1
   - Every sequence in the typical set is almost equally likely, occuring with probability about 2^{-nH}.

(2) "The Rest."
   - The probability we see something from this set approaches zero.

- Main Point #2: We can give a brute-force description of all the sequences in the typical set. Since $|A_\epsilon^{(n)}| \doteq 2^{nH}$, and all the atypical sequences practically never show up. So the ratio of bits that we use to describe the sequence, compared to the number of bits the other guy uses, is
$$\frac{nH}{n \log |\mathcal{X}|} = \frac{H}{\log |\mathcal{X}|}.$$
- The relative size ratio of the typical set to all the sequences is
$$\frac{2^{nH}}{2^{n \log |\mathcal{X}|}} = \frac{1}{2^{n(\log |\mathcal{X}| - H)}}$$
which is approaching zero.
- One should keep these points in mind when looking at suggestive pictures such as those in Figure 3. If we compare the cardinality of the typical set with that of the whole set, then we have an exponentially increasing difference in size. So if we were to compare the typical set against the whole set using only the number of sequences in these sets as a metric, then the typical set should look like a very tiny sub-blob of the whole set.

  However, if we compare the probability mass of the typical set with that of the original entire set, they would both be about 1 (the latter would, of course, be exactly 1).

- Intuition: Keep flipping coin with bias 1/4. The sequence reveals a structure: about a fourth of the coins are heads, and the rest are tails. Any sequence that deviates from this heads/tails ratio becomes increasingly unlikely as the number of flips goes up.
- Definition:

$$A_\epsilon^{(n)} := \{(x_1, \ldots, x_n) \in \mathcal{X}^n : 2^{-n(H(X)+\epsilon)} \leq p(x_1, \ldots, x_n) \leq 2^{-n(H(X)-\epsilon)}\}.$$

Theorem 4.1. The AEP.

(1) If $(x_1, \ldots, x_n) \in A_\epsilon^{(n)}$*, then* $H(X) - \epsilon \leq -\frac{1}{n} \log p(x_1, x_2, \ldots, x_n) \leq H(X) + \epsilon$.

(2) $\mathbf{P}\left[A_\epsilon^{(n)}\right] > 1 - \epsilon$ *for* n *sufficiently large.*

(3) $|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$

(4) $|A_\epsilon^{(n)}| \geq (1-\epsilon)2^{n(H(X)-\epsilon)}$.

Theorem 4.2. Typical Set Coding. *There exists a code that maps sequences* x^n *of length* n *into binary strings such that the mapping is one-to-one and*

$$\mathbf{E}\left[l(X^n)\right] \leq n(H(X) + \epsilon).$$

5. Entropy Rate

5.1. Motivations.

(1) Want to know how the entropy of a sequence of random variables grows with n.

(2) Want to know if we can compress more general stochastic processes (not just i.i.d. sequences) using an AEP-style argument. Turns out you can do so for stationary ergodic processes.

5.2. Preliminary Definitions.

- Stationary: Joint distribution of any subset of the sequence is invariant w.r.t shifts in time index.
  - Consequence: $H(X_1^n) = H(X_2^{n+1})$
- Markov: $\mathbf{P}\left[X_{n+1}|x_{n+1}|X_n = x_n\right]$
  - Consequence: $H(X_n|X_1^{n-1}) = H(X_n|X_{n-1})$.
  - Strategy: Draw a simple line-graph depicting the Markov relationships.
  - $X \to Y \to Z$ implies $X \leftarrow Y \leftarrow Z$. So all the arrows are really double arrows. It could perhaps more descriptively be written as $X \leftrightarrow Y \leftrightarrow Z$.
  - Recall that if g is a deterministic function, we can write $Y \to g(Y)$.
- Stationary Distribution of a Markov Chain: μ such that $\mu_j = \sum_i \mu_i P_{ij}$ for all j. Eigenvector of P with eigenvalue 1; if MC is aperiodic and irreducible, it is unique and is also the limiting distribution.

5.3. **Entropy Rate Calculations.**

- Entropy rate is defined as $H(\mathcal{X}) := \lim_{n\to\infty} \frac{H(X_1^n)}{n}$
- For a stationary stochastic process, $H(X_n|X_{n-1},\ldots,X_1)$ is nonincreasing in n and has a limit.
- For a stationary stochastic process, $H(\mathcal{X})$ exists, and is also equal to $\lim_{n\to\infty} H(X_n|X_1^n)$. (Cesaro convergence technique)
- For a stationary Markov Chain,
$$H(\mathcal{X}) = H(X_2|X_1).$$
- By generalized AEP in Chapter 16.8, if process is stationary and ergodic, there are $2^{nH(\mathcal{X})}$ typical sequences. So we can represent sequences of length n using $nH(\mathcal{X})$ bits.
- Entropy rate for stationary Markov Chain $\{X_i\}$ where $X_1 \sim \mu$, the stationary distribution:
$$H(\mathcal{X}) = -\sum_{ij} \mu_i P_{ij} \log P_{ij}.$$
- Nice formula for entropy rate for random walk on a graph:
$$H(\mathcal{X}) = \log 2E - H\left(\frac{E_1}{2E}, \frac{E_2}{2E}, \ldots, \frac{E_m}{2E}\right).$$

5.4. **Second Law of Thermodynamics.** Does entropy always increase? No. Start an MC in a uniform distribution, and watch it converge to a nonuniform stationary distribution. So it depends.

(1) $D(\mu_n||\mu_n')$ decreases.

$$D(p(x_n,x_{n+1})||q(x_n,x_{n+1})) = D(p(x_n)||q(x_n)) + \cancel{D(p(x_{n+1}|x_n)||q(x_{n+1}|x_n))}$$
$$D(p(x_n,x_{n+1})||q(x_n,x_{n+1})) = D(p(x_{n+1})||q(x_{n+1})) + D(p(x_n|x_{n+1})||q(x_n|x_{n+1}))$$

(2) $D(\mu_n||\mu)$ decreases. (Use (1).)

(3) $H(X_n)$ increases if the stationary distribution is uniform. (Use (2).)

(4) $H(X_n|X_1)$ increases for a stationary Markov chain. (easy)

(5) $H(X_0|X_n)$ increases for any Markov chain. (easy)

6. Gambling

(1) Goal: Choose asset allocation vector $\underline{b}$ to maximize doubling rate of wealth; that is, maximize your wealth to first order in the exponent.

(2) Doubling Rate:

- $W(\underline{b},\underline{p}) = \mathbf{E}\left[\log S(X)\right] = \sum_{k=1}^m p_k \log b_k o_k$
- $W(\underline{b},\underline{p}) = D(\underline{p}||\underline{r}) - D(\underline{p}||\underline{b})$. Difference between distance of the bookie's estimate from the true distribution and the distance of the gambler's estimate from the true distribution.

(3) Proportional Gambling is log-optimal: $W^*(\underline{p}) = \max_b W(\underline{b}, \underline{p}) = \sum_i p_i \log o_i - H(\underline{p})$, and is achieved by $\underline{b}^* = \underline{p}$.

(4) Growth rate: Wealth grows as $S_n \doteq 2^{nW^*(\underline{p})}$.

(5) Conservation Law: For uniform fair odds,

$$H(\underline{p}) + W^*(\underline{p}) = \log m.$$

(6) Side information: For horse race X, the increase in doubling rate due to side information Y is $\Delta W = I(X;Y)$.

7. Huffman Coding

7.1. Terms To Know.

- Nonsingular: $C(x_i) \neq C(x_j)$ if $x_i \neq x_j$. $C(\cdot)$ is 1-1.
- Uniquely Decodable: A sequence of concatenated codewords $C(x_1)C(x_2)\dots C(x_n)$ is decodable in only one unambiguous way. In other words, only one way to parse it.
- Instantaneous / Prefix Code / Self-Punctuating: You can parse the concatenated codewords unambiguously on the fly.

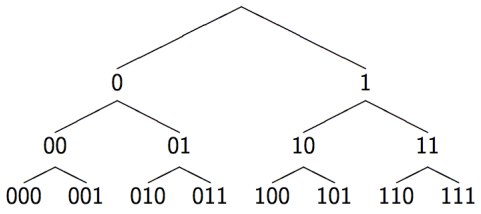
7.2. Running the Huffman Algorithm.

- Combine two smallest probabilities. Draw a tree branch joint that connects them.
- Rinse wash and repeat until you have the whole tree.
- To get the codewords, label the edges of the tree 0 and 1, and read off the numbers you see when traversing the tree from the root down to the leaf.

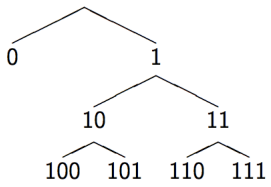
7.3. Main Theorems.

- Kraft Inequality: $D^{-l_i} \leq 1 \iff$ Instantaneous codes.
  - Any prefix-free code must satisfy the inequality.
  - But given only some integers that satisfy the inequality, you can also create a prefix-free code!

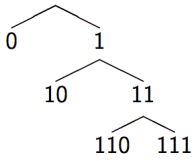
    Suppose we are given $(1, 2, 3, 3)$, where $D = 2$. Start with



After taking the first available spot on the first level...



After taking the second available spot on the second level ...



Done.

- By using the $\Longrightarrow$ direction, you can also take any Huffman code and turn it into a slice code. Take the codeword lengths and do the procedure directly above.
- Alternative interpretation of Kraft in terms of partitioning the unit interval $[0, 1]$: Let codeword $y_1 y_2 \cdots y_{l_i}$ correspond to the subinterval $[0.y_1 y_2 \cdots y_{l_i}, [0.y_1 y_2 \cdots y_{l_i} + \frac{1}{D^{l_i}})$. Prefix-free condition means that these subintervals must be disjoint. Thus, sum of the lengths of the subintervals cannot be larger than 1. Each interval has length D^{l_i}.

- McMillan Inequality: $D^{-l_i} \leq 1 \Longleftrightarrow$ Uniquely Decodable Codes
- Entropy Bound on Data Compression:

  Theorem 7.1. *The expected length L of any instantaneous D-ary code for a random variable X is greater than or equal to the entropy $H_D(X)$; that is,*

  $$L := \sum p_i l_i \geq H_D(X)$$

  with equality if and only if $D^{-l_i} = p_i$.

  Proof. Using the relative entropy optimization technique, that $L - H_D(X) = D(\underline{p}||\underline{r}) + \log_D \frac{1}{\sum D^{-l_i}}$, where $r_i := \frac{D^{-l_i}}{\sum_j D^{-l_j}}$. Verify to yourself that this is zero if and only if $p_i = D^{-l_i}$. □

  Interpretation: A good code is finding the D-adic distribution $\underline{r}$ that is closest (in relative entropy sense) to the true distribution $\underline{p}$

- Shannon Code:

  $$l_i = \lceil \log_D \frac{1}{p_i} \rceil$$
  $$H_D(X) \leq L \leq H_D(X) + 1$$

- Huffman Code:

  $$L^* = \min_{\sum D^{-l_i} \leq 1} \sum p_i l_i$$
  $$H_D(X) \leq L^* < H_D(X) + 1$$

- Wrong code:
$$H(p) + D(p||q) \leq L < H(p) + D(p||q) + 1$$
- Stochastic processes
$$\frac{H(X_1, \ldots, X_n)}{n} \leq L_n < \frac{H(X_1, \ldots, X_n)}{n} + \frac{1}{n}$$
- Stationary Processes:
$$L_n \to H(\mathcal{X})$$
- Competitive optimality: Shannon code $l(x) = \lceil \log \frac{1}{p(x)} \rceil$ versus any other code:
$$\mathbf{P}\left[l(X) \geq l'(X) + c\right] \leq \frac{1}{2^{c-1}}$$

8. Legal Disclaimer

These review session notes were mostly typed over the course of five days, with little sleep. Therefore, it is entirely possible – and in fact highly likely – that they contain a few errors. I believe the likelihood of conceptual errors is very low, but the likelihood of stupid typos is probably high. It's just not reasonable for me to promise that there are absolutely none.

These notes have been provided as a complimentary service for the students, and are not part of my required duties as a teaching assistant. They should not be put up to the same standard as the required textbook, which was written by top authorities in the field, has been read by many people, and has withstood the tests of time.

In short, I am humbly requesting to not bear any culpability. The reader relinquishes all rights to litigate the author should an error in these notes somehow adversely affect him or her on the final examination or, God forbid, even later in life.