Logistics

Review

Class Road Map - IT-I

- L1 (9/26): Overview, Communications, Information, Entropy
- L2 (10/1): Props. Entropy, Mutual Information
- L5 (10/10): AEP, Compression
- L6 (10/15): Compression, Method of Types
- L7 (10/17): Types, U. Coding., Stoc Processes, Entropy rates
- L8 (10/22): Entropy rates, HMMs, Coding, Kraft
- L9 (10/24): Kraft, Shannon Codes, Huffman, Shannon/Fano/Elias
- L10 (10/28): Huffman, Shannon/Fano/Elias
- L11 (10/29): Shannon Games
- LXX (10/31): Midterm, in class.
- L12 (11/7): Arith. Coding, Channel Capacity
- L13 (11/12): Channel Capacity
- L14 (11/14): Channel Capacity, Shannon’s 2nd thm
- L15 (11/19): Shannon’s 2nd thm, zero error codes, feedback
- L16 (11/21): Joint thm, coding, hamming, diff. entropy
- L17
- L18
- L19

Finals Week: December 12th–16th.
Cumulative Outstanding Reading

- Read chapters 1 and 2 in our book (Cover & Thomas, “Information Theory”) (including Fano’s inequality).
- Chapter 3 in our book (Cover & Thomas, “Information Theory”).
- Section 11.1 (method of types).
- Chapter 4 and 5 in our book (Cover & Thomas, “Information Theory”)
- Read stream code chapter 6 in “Information Theory, Inference, and Learning Algorithms” by David J.C. MacKay (available online http://www.inference.phy.cam.ac.uk/mackay/itila/)
- Read Chapter 7 in our book (Cover & Thomas, “Information Theory”).
- Read Chapter 8 in our book (Cover & Thomas, “Information Theory”).

Homework

- Homework 6 on our web page (http://j.ee.washington.edu/~bilmes/classes/ee514a_fall_2013/), due tonight, at 11:45pm.
Logistics

Review

Announcements

- Office hours, every week, now Thursdays 4:30-5:30pm. Can also reach me at that time via a canvas conference.
- Final assignment, need to upload pdf scan we send you.

Channel Coding Theorem (Shannon 1948): more formally

Theorem 16.2.2

All rates below \( C \triangleq \max_p(x) I(X;Y) \) are achievable. Specifically, \( \forall R < C, \) there exists a sequence of \((2^{nR}, n)\) codes with maximum probability of error \( \lambda^{(n)} \to 0 \) as \( n \to \infty \). Conversely, any \((2^{nR}, n)\) sequence of codes with \( \lambda^{(n)} \to 0 \) as \( n \to \infty \) must have that \( R < C \).

- Implications: as long as we do not code above capacity we can, for all intents and purposes, code with zero error.
- This is true for all noisy channels representable under this model.
- We’re talking about discrete channels now, but we generalize this to continuous channels in the coming weeks.
Zero Error Codes

- If \( P_e^{(n)} = 0 \), then \( H(W|Y^n) = 0 \) (no uncertainty)
- For simplicity, assume \( H(W) = nR = \log M \) (i.e., uniform dist. over \( \{1, 2, \ldots, M\} \). Sufficient since this is max rate under \( M \) messages.
- First lets consider the case if \( P_e^{(n)} = 0 \), in such case it is easy to show that \( R \leq C \).

\[
nR = H(W) = H(W|Y^n) + I(W; Y^n) = I(W; Y^n)
\]

(16.21)

\[
\leq I(X^n; Y^n) \quad \text{//Since } W \rightarrow X^n \rightarrow Y^n \text{ and data proc. ineq.} 
\]

(16.22)

\[
= H(Y^n) - H(Y^n|X^n) = H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_{1:i-1}, X^n)
\]

(16.23)

But \( Y_i \perp \perp \{Y_{1:i-1}, X_{1:i-1}, X_{i+1:n}\}|X_i \), so we can continue as

\[
= H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) \leq \sum_{i} [H(Y_i) - H(Y_i|X_i)]
\]

(16.24)

\[
= \sum_{i} I(Y_i; X_i) \leq nC
\]

(16.25)

Sequence of codes w. vanishing error must have \( R < C \).

Proof that \( \lambda^{(n)} \rightarrow 0 \) as \( n \rightarrow \infty \Rightarrow R < C \).

\[
nR = H(W) = H(W|Y^n) + I(W; Y^n)
\]

(16.30)

\[
\leq H(W|Y^n) + I(X^n(W); Y^n) \quad \text{//Since } W \rightarrow X^n \rightarrow Y^n
\]

(16.31)

\[
\leq 1 + P_e^{(n)} nR + I(X^n(W); Y^n) \quad \text{//by Fano}
\]

(16.32)

\[
\leq 1 + P_e^{(n)} nR + nC \quad \text{//by lemma ??}
\]

(16.33)

\[
\Rightarrow R \leq P_e^{(n)} R + 1/n + C
\]

(16.34)

Now as \( n \rightarrow \infty \), \( P_e^{(n)} \rightarrow 0 \), and \( 1/n \rightarrow 0 \) as well. Thus

\[
\Rightarrow R < C
\]

(16.35)
Sequence of codes w. vanishing error must have $R < C$.

Lower bound on error:

$$P_e^{(n)} \geq 1 - \frac{C}{R} \quad (16.31)$$

generates this plot (lower bound on error):

[Diagram: Log $P_e$ vs $R$, with $0$ at lower bound and increasing curve as $R$ increases from $C$.]

Zero-error capacity

- What if we insist on $R = C$ and $P_e = 0$. In such case, what are the requirements of any such code.

\[
\begin{align*}
nR &= H(W) = H(X^n|W) \quad \text{// if codewords distinct} \\
&= H(X^n|Y^n) + I(X^n; Y^n) = I(X^n; Y^n) \quad (16.33) \\
&= 0 \quad \text{since } P_e = 0 \quad (16.34) \\
&= H(Y^n) - H(Y^n|X^n) \quad (16.35) \\
&= H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) \quad (16.36) \\
&= \sum_{i} H(Y_i) - \sum_{i} H(Y_i|X_i) \quad \text{// if all } Y_i \text{'s are indep} \quad (16.37) \\
&= \sum_{i} I(X_i; Y_i) \quad (16.38) \\
&= nC \quad \text{// if we choose } p^*(x) \in \arg\max_{p(x)} I(X; Y) \quad (16.39)
\end{align*}
\]
Feedback for DMC

**Definition 16.2.3 ((2nR, n) feedback code)**

Such a code is the encoder \( X_i(W, Y_{1:i-1}) \), a decoder \( g : Y^n \to \{1, 2, \ldots, 2^{nR}\} \), and \( P_e^{(n)} = \Pr(g(Y^n) \neq W) \) for \( H(W) = nR \) (uniform).

**Definition 16.2.4 (Capacity)**

The capacity with feedback \( C_{FB} \) of a DMC is the max of all rates achievable by feedback codes.

**Theorem 16.2.5**

\[
C_{FB} = C = \max_{p(x)} I(X;Y) \quad \text{for a DMC} \quad (16.33)
\]

---

Joint Source/Channel Theorem

- Data compression: We now know that it is possible to achieve error free compression if our average rate of compression, \( R \), measured in units of bits per source symbol, is such that \( R > H \) where \( H \) is the entropy of the generating source distribution.
- Data Transmission: We now know that it is possible to achieve error free communication and transmission of information if \( R < C \), where \( R \) is the average rate of information sent (units of bits per channel use), and \( C \) is the capacity of the channel.
- Q: Does this mean that if \( H < C \), we can reliably send a source of entropy \( H \) over a channel of capacity \( C \)?
- This seems intuitively reasonable.
Joint Source/Channel Theorem: process

The process would go something as follows:
1. Compress a source down to its entropy, using Huffman, LZ, arithmetic coding, etc.
2. Transmit it over a channel.
3. If all sources could share the same channel, would be very useful.
4. I.e., perhaps the same channel coding scheme could be used regardless of the source, if the source is first compressed down to the entropy. The channel encoder/decoder need not know anything about the original source (or how to encode it).
5. Joint source/channel decoding as in the following figure:

```
source encoder channel encoder
noise
channel source
encoder
channel decoder
receiver
```

6. Maybe obvious now, but at the time (1940s) it was a revolutionary idea!

Joint Source/Channel Theorem

- Source: $V \in \mathcal{V}$ that satisfies AEP (e.g., stationary ergodic).
- Send $V_{1:n} = V_1, V_2, \ldots, V_n$ over channel, entropy rate $H(\mathcal{V})$ of stochastic process (if i.i.d., $H(\mathcal{V}) = H(V_i), \forall i$).
- $V_{1:n} \rightarrow \text{Encoder} \rightarrow X^n \rightarrow \text{Channel} \rightarrow Y^n \rightarrow \text{Decoder} \rightarrow \hat{V}_{1:n}$
- Error probability and setup:

$$P_e^{(n)} = P(V_{1:n} \neq \hat{V}_{1:n})$$
$$= \sum_{y_{1:n}, v_{1:n}} \Pr(v_{1:n}) \Pr(y_{1:n} | X^n(v_{1:n})) 1\{g(y_{1:n}) \neq v_{1:n}\} \quad \text{(16.1)}$$
Theorem 16.3.1 (Source/Channel Coding Theorem)

If \( V_{1:n} \) satisfies AEP, then \( \exists \) a sequence of \( (2^{nR}, n) \) codes with \( P_e^{(n)} \to 0 \) if \( H(V) < C \). Conversely, if \( H(V) > C \), then \( P_e^{(n)} > 0 \) for all \( n \) and cannot send with arbitrarily low probability of error.

Proof.

- If \( V \) satisfies AEP, then \( \exists \) a set \( A_e^{(n)} \) with \( |A_e^{(n)}| \leq 2^{n(H(V)+\epsilon)} \) (\( A_e^{(n)} \) has all the probability).
- We only encode the typical set, and signal an error otherwise. This contributes \( \epsilon \) to \( P_e \).
- We index elements of \( A_e^{(n)} \) as \( \{1, 2, \ldots, 2^n(H+\epsilon)\} \), so need \( n(H + \epsilon) \) bits.
- This gives a rate of \( R = H(V) + \epsilon \). If \( R < C \) then the error \( < \epsilon \) which we can make as small as we wish.

... proof continued.

- Then
  
  \[
  P_e^{(n)} = \Pr(V_{1:n} \neq \hat{V}_{1:n}) 
  \leq \Pr(V_{1:n} \notin A_e^{(n)}) + \Pr(g(Y^n) \neq V^n | V^n \in A_e^{(n)}) < \epsilon \text{ since } R < C 
  \]

  \[
  \leq \epsilon + \epsilon = 2\epsilon 
  \]

  And the first part of the theorem is proved.

- To show the converse, show that \( P_e^{(n)} \to 0 \Rightarrow H(V) \leq C \) for source channel codes.
Joint Source/Channel Theorem

... proof continued.

- Define:
  \[ X^n(V^n) : V^n \to X^n \]  
  //encoder  
  (16.6)  
  \[ g^n(Y^n) : Y^n \to V^n \]  
  //decoder  
  (16.7)  

- Now recall, original Fano says \( H(X|Y) \leq 1 + P_e \log |X| \).
- Here we have
  \[ H(V^n|\hat{V}^n) \leq 1 + P_e^{(n)} \log |V^n| = 1 + nP_e \log |X| \]  
  (16.8)

We get the following derivation

\[
H(\mathcal{V}) \leq \frac{H(V_1, V_2, \ldots, V_n)}{n} = \frac{H(V_{1:n})}{n} \\
= \frac{1}{n} H(V_{1:n}|\hat{V}_{1:n}) + \frac{1}{n} I(V_{1:n}; \hat{V}_{1:n}) \\
\leq \frac{1}{n} \left( 1 + P_e^{(n)} \log |\mathcal{V}| \right) + \frac{1}{n} I(V_{1:n}; \hat{V}_{1:n}) \quad \text{by Fano} \\
\leq \frac{1}{n} \left( 1 + P_e^{(n)} \log |\mathcal{V}| \right) + \frac{1}{n} I(X_{1:n}; Y_{1:n}) \quad \text{by } V \to X \to Y \to \hat{V} \text{ and DPP} \\
\leq \frac{1}{n} + P_e^{(n)} \log |\mathcal{V}| + C \quad \text{memoryless} \\
\leq \frac{1}{n} + P_e^{(n)} \log |\mathcal{V}| + C \quad \text{memoryless}  \\
\]

- Letting \( n \to \infty, \frac{1}{n} \text{ and } P_e \to 0 \) which leaves us with \( H(\mathcal{V}) \leq C \).
Coding and Codes

- Shannon’s theorem says that there exists a sequence of codes such that if $R < C$ the error goes to zero.
- It doesn’t provide such a code, nor does it offer much insight on how to find one.
- In all cases, we add enough redundancy to a message so that the original message can be decoded unambiguously.

Physical Solution to Improve Coding

- It is possible to communicate more reliably by changing physical properties to decrease the noise (e.g., decrease $p$ in a BSC).
- Use more reliable and expensive circuitry
- Improve environment (e.g., control thermal conditions, remove dust particles or even air molecules)
- In compression, use more physical area/volume for each bit.
- In communication, use higher power transmitter, use more energy thereby making noise less of a problem.
- These are not IT solutions which is what we want.
Rather than send message \( x_1 x_2 \ldots x_k \) we repeat each symbol \( K \) times redundantly.

Recall our example of repeating each word in a noisy analog radio connection.

Message becomes \( x_1 x_1 \ldots x_1 x_2 x_2 \ldots x_2 \ldots \)

For many channels (e.g., BSC \((p < 1/2)\)), error goes to zero as \( K \to \infty \).

Easy decoding: when \( k \) is odd, take a majority vote (which is optimal for a BSC)

On the other hand, \( R \propto 1/k \to 0 \) as \( k \to \infty \)

This is really a pre-1948 way of thinking code.

Thus, this is not a good code.

---

(From D. Mackay) Consider sending message \( s = 0 \, 0 \, 1 \, 0 \, 1 \, 1 \, 0 \)

One scenario

\[
\begin{array}{cccccccc}
  s & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\
  t & 000 & 000 & 111 & 000 & 111 & 111 & 000 \\
  n & 000 & 001 & 000 & 000 & 101 & 000 & 000 \\
  r & 000 & 001 & 111 & 000 & 010 & 111 & 000 \\
\end{array}
\]

Another scenario

\[
\begin{array}{cccccccc}
  s & 0 & 0 & 1 & 0 & 1 & 1 & 0 \\
  t & 000 & 000 & 111 & 000 & 111 & 111 & 000 \\
  n & 000 & 001 & 000 & 000 & 101 & 000 & 000 \\
  r & 000 & 001 & 111 & 000 & 010 & 111 & 000 \\
  \hat{s} & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\
\end{array}
\]

corrected errors \( \rightarrow * \)
detected but uncorrected errors \( \rightarrow * \)

Thus, can only correct one bit error not two.
**Simple Parity Check Code**

- Binary input/output alphabets $\mathcal{X} = \mathcal{Y} = \{0, 1\}$.
- Block sizes of $n-1$ bits: $x_{1:n-1}$.
- $n^{th}$ bit is an indicator of an odd number of 1 bits in $x_{1:n-1}$.
- I.e., $x_n \leftarrow \text{mod} \left( \sum_{i=1}^{n-1} x_i, 2 \right)$.
- Thus a necessary condition for valid code word is: $\text{mod} \left( \sum_{i=1}^{n} x_i, 2 \right) = 0$.
- Any instance of an odd number of errors (bit swaps) won’t pass this condition, and such an error is hence detected.
- although an even number of errors will pass the condition (error goes undetected).
- can not correct all errors, and moreover only detects some of the kinds of errors (odd number of swaps).
- On the other hand, parity checks form the basis for many sophisticated coding schemes (e.g., low-density parity check (LDPC) codes, Hamming codes etc.).
- We study Hamming codes next.

**Hamming Codes**

- Best illustrated by an example.
- Let $\mathcal{X} = \mathcal{Y} = \{0, 1\}$.
- Fix the desired rate at $R = 4/7$ bit per channel use.
- Thus, in order to send 4 data bits, we need to use the channel 7 times.
- Let the four data bits be denoted $x_0, x_1, x_2, x_3 \in \{0, 1\}$.
- When we send these 4 bits, we are also going to send 3 additional parity or redundancy bits, named $x_4, x_5, x_6$.
- Note: all arithmetic in the following will be mod 2. I.e. $1 + 1 = 0, 1 + 0 = 1, 1 = 0 - 1 = -1$, etc.
(7, 4, 3) Hamming Codes

- Parity bits determined by the following equations:
  \begin{align*}
  x_4 &\equiv x_1 + x_2 + x_3 \pmod 2 \quad (16.14) \\
  x_5 &\equiv x_0 + x_2 + x_3 \pmod 2 \quad (16.15) \\
  x_6 &\equiv x_0 + x_1 + x_3 \pmod 2 \quad (16.16)
  \end{align*}

- I.e., if \((x_0, x_1, x_2, x_3) = (0110)\) then \((x_4, x_5, x_6) = (011)\) and complete 7-bit codeword sent over channel would be \((0110011)\).

- We can also describe this using linear equalities as follows (all \(\pmod 2\)).
  \begin{align*}
  x_1 + x_2 + x_3 + x_4 &= 0 \\
  x_0 + x_2 + x_3 + x_5 &= 0 \\
  x_0 + x_1 + x_3 + x_6 &= 0 \quad (16.17)
  \end{align*}

- Or alternatively, as \(Hx = 0\) where \(x^\top = (x_1, x_2, \ldots, x_7)\) and
  \[
  H = \begin{pmatrix}
  0 & 1 & 1 & 1 & 1 & 0 & 0 \\
  1 & 0 & 1 & 1 & 0 & 1 & 0 \\
  1 & 1 & 0 & 1 & 0 & 0 & 1
  \end{pmatrix} \quad (16.18)
  \]

- Codewords lie in null-space of \(H\)

- Notice that \(H\) is a column permutation of all seven non-zero length-3 column vectors.

- Thus the code words are defined by the null-space of \(H\). I.e., \(\{x : Hx = 0\}\).

- Since the rank of \(H\) is 3, the null-space is 4, and we expect there to be \(16 = 2^4\) binary vectors in this null space.
Hamming Codes

- The 16 vectors in the nullspace (i.e., \( \{ x : Hx = 0 \} \)) are as follows:

  \[
  \begin{align*}
  0000000 & \ 0100101 & \ 1000011 & \ 1100011 \\
  0001111 & \ 0101010 & \ 1001100 & \ 1101100 \\
  0010110 & \ 0110011 & \ 1010101 & \ 1110101 \\
  0011001 & \ 0111100 & \ 1011010 & \ 1111010 \\
  \end{align*}
  \] (16.19) (16.20) (16.21) (16.22)

- Note the first (highest order) four bits of these vectors range from 0 to 15 in binary (i.e., all bit strings of length 4).
- These first four bits are the data bits, and the 2nd three bits are the redundancy bits.
- The vectors constitute the codewords, any codeword must be one of the above.

Thus, any valid codeword is in \( C = \{ x : Hx = 0 \} \).

Thus, if \( v_1, v_2 \in C \) then \( H(v_1 + v_2) = Hv_1 + Hv_2 = 0 \) and thus \( v_1 + v_2 \in C \).

Also, \( v_1 - v_2 \in C \) due to linearity (codewords closed under addition and subtraction).

Minimum number of ones in any non-zero codeword is 3. This is called the weight of a code.

Why weight 3? Suppose there was a weight-two code word with non-zeros at position \( i \) and \( j \). Thus, sum of columns \( i \) and \( j \) would be zero. But since columns of \( H \) are all different, sum of any two columns is non-zero. Hence, can’t have any weight-2 codeword.

Q: why can’t we have a weight 1 code word?

Can have weight 3 codeword, since sum of two columns will equal another column, and sum of two equal binary vectors is zero (mod 2).
Hamming Codes: Distance

- Thus, any codeword is in \( C = \{ x : Hx = 0 \} \).
- minimum distance of a code is also 3, which is minimum number of differences between any two codewords.
- Another way of saying this: if \( v_1, v_2 \in C \) then \( d_H(v_1, v_2) \geq 3 \) where \( d_H(x, y) = \sum_i 1 \{ x(i) \neq y(i) \} \) is the Hamming distance.
- Why? Suppose \( v_1, v_2 \in C \) differ in only two places. Then \( H(v_1 - v_2) \) will be a difference or sum of two columns of \( H \) (mod 2). But given \( v_1, v_2 \in C \Rightarrow (v_1 - v_2) \in C \). Can’t have difference or sum, (1+1 = 1-1 mod 2) of any two columns equaling zero, so \( v_1 - v_2 \) can’t differ in only two places.
- In general, codes with large minimum distance is good because then it is possible to correct errors. I.e., if \( \hat{v} \) is received codeword, then we can find \( i \in \text{argmin}_i d_H(\hat{v}, v_i) \) as the decoding procedure.

Hamming Codes: BSC

- Now a BSC(\( p \)) (crossover probability \( p \)) will chance some of the bits (noise), meaning a 0 might change to a 1 and vice versa.
- So if \( x = (x_0, x_2, \ldots, x_6) \) is transmitted, what is received is

\[
y = x + z = (x_0 + z_0, x_1 + z_1, \ldots, x_6 + z_6)
\]

where \( z = (z_0, z_2, \ldots, z_6) \) is the noise vector.
- Receiver knows \( y \) but wants to know \( x \). We then compute

\[
s = Hy = H(x + z) = \underbrace{Hx}_{=0} + Hz = Hz
\]

- \( s \) is called the syndrome of \( y \). If \( s = 0 \), then all parity checks are satisfied by \( y \) and is a necessary condition for a correct codeword.
Hamming Codes : BSC

- Moreover, we see that $s$ is a linear combination of columns of $H$

$$s = z_0 \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} + z_1 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + z_2 \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} + \cdots + z_6 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \quad (16.25)$$

- Since $y = x + z$, we know $y$, so if we know $z$ we know $x$.
- We only need to solve for $z$ in $s = H z$, 16 possible solutions.
- Ex: Suppose that $y^T = 0111001$ is received (which is not a codeword), then $s = H y = (101)^T$ and the 16 solutions are:

<table>
<thead>
<tr>
<th>0100000</th>
<th>0010011</th>
<th>0101111</th>
<th>1001001</th>
</tr>
</thead>
<tbody>
<tr>
<td>1100011</td>
<td>0001010</td>
<td>1000110</td>
<td>1111010</td>
</tr>
<tr>
<td>0000101</td>
<td>0111001</td>
<td>1110101</td>
<td>0011100</td>
</tr>
<tr>
<td>0110110</td>
<td>1010000</td>
<td>1101100</td>
<td>1011111</td>
</tr>
</tbody>
</table>

- 16 is better than 128 (possible $z$ vectors) but still many.
- What is the probability of each solution? Since we are assuming a BSC($p$) with $p < 1/2$, the most probable solution for $z$ has the least weight. Any solution with weight $k$ has probability $p^k$.
- Notice that there is only one possible solution with weight 1, and this is most probable solution.
- In previous example, most probable solution is $z^T = (01000000)$ and in $y = x + z$ with $y^T = 0111001$ this leads to codeword $x = 0011001$ and information bits 0011.
- In fact, for any $s$, there is a unique minimum weight solution for $z$ in $s = H z$ (in fact, this weight is no more than 1)!
- If $s = (000)$ then the unique solution is $z = (0000000)$.
- For any other $s$, then $s$ must be equal to one of the columns of $H$, so we can generate $s$ by flipping the corresponding bit of $z$ on (giving weight 1 solution).
Hamming Decoding Procedure

Here is the final decoding procedure on receiving $y$:

1. Compute the syndrome $s = Hy$.
2. If $s = (000)$ set $z \leftarrow (0000000)$ and goto 4.
3. Otherwise, locate unique column of $H$ equal to $s$ form $z$ all zeros but with a 1 in that position.
4. Set $x \leftarrow y + z$.
5. Output $(x_0, x_1, x_2, x_3)$ as the decoded string.

This procedure can correct any single bit error, but fails when there is more than one error.

Hamming Decoding: Venn Diagrams

- We can visualize the decoding procedure using Venn Diagrams

- Here, first four bits to be sent $(x_0, x_1, x_2, x_3)$ are set as desired and parity bits $(x_4, x_5, x_6)$ are also set. Figure shows $(x_0, x_2, \ldots, x_6) = (1, 0, 0, 0, 1, 0, 1)$ with parity check bits:

$$x_4 \equiv x_0 + x_1 + x_2 \mod 2 \quad (16.26)$$
$$x_5 \equiv x_1 + x_2 + x_3 \mod 2 \quad (16.27)$$
$$x_6 \equiv x_0 + x_2 + x_3 \mod 2 \quad (16.28)$$
Hamming Decoding: Venn Diagrams

- The syndrome can be seen as a condition where the parity conditions are not satisfied.
- Above we argued that for $s \neq (0,0,0)$ there is always a one bit flip that will satisfy all parity conditions.

Example: Here, $z_1$ can be flipped to achieve parity.
Example: Here, $z_4$ can be flipped to achieve parity.

![Venn Diagram](c)

Example: And here, $z_2$ can be flipped to achieve parity.

![Venn Diagram](d)
Example: And here, there are two errors, \( y_6 \) and \( y_2 \) (each of which are marked with a *).

Flipping \( y_1 \) will achieve parity, but this will lead to three errors (i.e., we will switch to a wrong codeword, and since codewords have minimum Hamming distance of 3, we’ll get 3 bit errors).

Many other coding algorithms.
- Reed Solomon Codes (used by CD players).
- Bose, Ray-Chaudhuri, Hocquenghem (BCH) codes.
- Convolutional codes
- Turbo codes (two convolutional codes with permutation network)
- Low Density Parity Check (LDPC) codes.
- All developed on our journey to find good codes with low rate that achieve Shannon’s promise.
- We may discuss LDPC and Turbo codes a bit more next quarter (but there are a few things we need to do first, such as . . .)
Entropy

- $H(X) = - \sum_x p(x) \log p(x)$
- All entropic quantities we’ve encountered in IT-I have been discrete.
- The world is continuous, channels are continuous, noise is continuous,
- We need a theory of compression, entropy, and channel capacity that applies to such continuous domains.
- We explore this next.

Continuous/Differential Entropy

- Let $X$ now be a continuous r.v. with cumulative distribution
  
  $F(x) = \Pr(X \leq x)$

  and $f(x) = \frac{d}{dx} F(x)$ is the density function.
- Let $S = \{x : f(x) > 0\}$ be the support set. Then

Definition 16.6.1 (differential entropy $h(X)$)

\[
  h(X) = - \int_S f(x) \log f(x) dx
\]

- Since we integrate over only the support set, no worries about $\log 0$.
- Perhaps it is best to do some examples.
Continuous Entropy Of Uniform Distribution

- Here, $X \sim U[0, a]$ with $a \in \mathbb{R}^+$. 
- Then
  \[
  h(X) = - \int_0^a \frac{1}{a} \log \frac{1}{a} \, dx = - \log \frac{1}{a} \tag{16.31}
  \]

- Note: continuous entropy can be both positive or negative.
- How can entropy (which we know to mean “uncertainty”, or “information”) be negative?
- In fact, entropy (as we’ve seen perhaps once or twice) can be interpreted as the exponent of the “volume” of a typical set.
- Example: $2^H(X)$ is the number of things that happen, on average, and can have $2^H(X) \ll |X|$. 
- Consider a uniform r.v. $Y$ such that $2^H(X) = |Y|$. 
- Thus having a negative exponent just means the volume is small.

Continuous Entropy Of Normal Distribution

- Normal (Gaussian) distributions are very important. 
- We have:
  \[
  X \sim N(0, \sigma^2) \iff f(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{x^2}{2\sigma^2}} \tag{16.32}
  \]
- Lets compute the entropy of $f$ in nats.
  \[
  h(X) = - \int f \ln f = - \int f(x) \left[ -\frac{x^2}{2\sigma^2} - \ln \sqrt{2\pi\sigma^2} \right] \, dx \tag{16.33}
  \]
  \[
  = \frac{1}{2} \ln e + \frac{1}{2} \ln(2\pi\sigma^2) = \frac{1}{2} \ln(2\pi e\sigma^2) \text{nats} \times \left( \frac{1}{\ln 2} \text{bits/nats} \right) 
  \]
  \[
  = \frac{1}{2} \ln(2\pi e\sigma^2) \text{bits} \tag{16.35}
  \]
- Note: only a function of the variance $\sigma^2$, not the mean. Why?
- So entropy of a Gaussian is monotonically related to the variance.
AEP lives

- We even have our own AEP in the continuous case, but before that a bit more intuition.
- In the discrete case, we have \( \Pr(x_1, x_2, \ldots, x_n) \approx 2^{-nH(X)} \) for big \( n \) and \( |A(\epsilon)^{(n)}| = 2^{nH} = (2^H)^n. \)
- Thus, \( 2^H \) can be seen like a “side length” of an \( n \)-dimensional hypercube, and \( 2^{nH} \) is like the volume of this hypercube (or volume of the typical set).
- So \( H \) being negative could mean small side length (small \( 2^H \) but still positive).

---

AEP

- Things are similar for the continuous case. Indeed

**Theorem 16.6.2**

Let \( X_1, X_2, \ldots, X_n \) be a sequence of r.v.’s, i.i.d. \( \sim f(x) \). Then

\[
\frac{1}{n} \log f(X_1, X_2, \ldots, X_n) \rightarrow E[-\log f(X)] = h(X)
\] (16.36)

- this follows via the weak law of large numbers (WLLN) just like in the discrete case.

**Definition 16.6.3**

\[ A(\epsilon)^{(n)} = \{x_{1:n} \in S^n : | -\frac{1}{n} \log f(x_1, \ldots, x_n) - h(X)| \leq \epsilon \} \]

- Note: \( f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i) \).
- Thus, we have upper/lower bounds on the probability

\[
2^{-n(h+\epsilon)} \leq f(x_{1:n}) \leq 2^{-n(h-\epsilon)}
\] (16.37)
**AEP**

- The volume of $A \subseteq \mathbb{R}^n$ is well defined as:
  \[ \text{Vol}(A) = \int_A dx_1 dx_2 \ldots dx_n \]  
  \[ (16.38) \]

- then we have

**Theorem 16.6.4**

1. $Pr(A_{\epsilon}^{(n)}) > 1 - \epsilon$
2. $(1 - \epsilon)2^{n(h(X) - \epsilon)} \leq \text{Vol}(A_{\epsilon}^{(n)}) \leq 2^{n(h(X) + \epsilon)}$

Note this is a bound on volume $\text{Vol}(A_{\epsilon}^{(n)})$ of typical set.

In discrete AEP, we bound cardinality of typical set $|A_{\epsilon}^{(n)}|$, and never need entropy to be negative — $H(X) \geq 0$ suffices to limit $|A_{\epsilon}^{(n)}|$ down to its lowest sensible value, namely 1.

**proof of theorem 16.6.4.**

1: First,

\[ p(A_{\epsilon}^{(n)}) = \int_{x_1:n \in A_{\epsilon}^{(n)}} f(x_1, \ldots, x_n) dx_1 \ldots dx_n \]  
\[ = Pr \left( \left| -\frac{1}{n} f(x_1, x_2, \ldots, x_n) - h(X) \right| \leq \epsilon \right) \geq 1 - \epsilon \]  
\[ (16.40) \]

for big enough $n$ which follows from the WLLN.

2: Next, we have

\[ 1 = \int_{S^n} f(x_1, \ldots, x_n) dx_1 \ldots dx_n \geq \int_{A_{\epsilon}^{(n)}} f(x_1, \ldots, x_n) dx_1 \ldots dx_n \]  
\[ \geq \int_{A_{\epsilon}^{(n)}} 2^{-n(h(X) + \epsilon)} dx_1 : n \]  
\[ = 2^{-n(h(X) + \epsilon)} \text{Vol}(A_{\epsilon}^{(n)}) \]  
\[ (16.42) \]

\[ \Rightarrow \text{Vol}(A_{\epsilon}^{(n)}) \leq 2^{n(h(X) + \epsilon)}. \]
proof of theorem 16.6.4 cont.

Similarly,

\[ 1 - \epsilon \leq \Pr(A_{\epsilon}^{(n)}) = \int_{A_{\epsilon}^{(n)}} f(x_{1:n})dx_{1:n} \tag{16.43} \]

\[ \leq \int_{A_{\epsilon}^{(n)}} 2^{-n(h(X) - \epsilon)}dx_{1:n} = 2^{-n(h(X) - \epsilon)} \text{Vol}(A_{\epsilon}^{(n)}) \tag{16.44} \]

- Like in the discrete case, \( A_{\epsilon}^{(n)} \) is the smallest volume that contains, essentially, all of the probability, and that volume is \( \approx 2^{nh} \).
- If we look at \( (2^{nh})^{1/n} \), we get a “side length” of \( 2^h \).
- So, \( -\infty < h < \infty \) is a meaningful range for entropy since it is the exponent of the equivalent side length of the \( n \)-D volume.
- Large negative entropy just means small volume.

Differential vs. Discrete Entropy

- Let \( X \sim f(x) \), and divide the range of \( X \) up into bins of length \( \Delta \).
- E.g., quantize the range of \( X \) using \( n \) bits, so that \( \Delta = 2^{-n} \).
- We can then view this as follows:

\[ f(x) \]

\[ \Delta = 2^{-n} \]

Mean value theorem, i.e., that if continuous within bin \( \exists x_i \) such that

\[ f(x_i) = \frac{1}{\Delta} \int_{i\Delta}^{(i+1)\Delta} f(x)dx \tag{16.45} \]
Joint Thm  Coding  Hamming Codes  Differential Entropy

Differential vs. Discrete Entropy

- Create a quantized random variable $X^\Delta$ having those values so that $X^\Delta = x_i$ if $i\Delta \leq X < (i + 1)\Delta$ (16.46)
- This gives a discrete distribution
  \[
  \Pr(X^\Delta = x_i) = p_i = \int_{i\Delta}^{(i+1)\Delta} f(x) dx = \Delta f(x_i)
  \] (16.47)
  and we can calculate the entropy
  \[
  H(X^\Delta) = -\sum_{i=-\infty}^{\infty} p_i \log p_i = -\sum_i f(x_i) \Delta \log(f(x_i)\Delta) = -\sum_i \Delta f(x_i) \log f(x_i) - \sum_i f(x_i) \Delta \log \Delta
  \] (16.48)
  \[
  = -\sum_i \Delta f(x_i) \log f(x_i) - \log \Delta
  \] (16.49)
  \[
  = -\sum_i \Delta f(x_i) \log f(x_i) - \log \Delta
  \] (16.50)

This follows since (as expected)

\[
\sum_i \Delta f(x_i) = \Delta \sum_i \frac{1}{\Delta} \int_{i\Delta}^{(i+1)\Delta} f(x) dx = \Delta \frac{1}{\Delta} \int f(x) dx = 1
\] (16.51)

- Also, as $\Delta \to 0$, we have $-\log \Delta \to \infty$ and (assuming all is integrable in a Riemannian sense)

\[
-\sum_i \Delta f(x_i) \log f(x_i) \to -\int f(x) \log f(x) dx
\] (16.52)

- So, $H(X^\Delta) + \log \Delta \to h(f)$ as $\Delta \to 0$.
- Loosely, $h(f) \approx H(X^\Delta) + \log \Delta$ and for an $n$-bit quantization with $\Delta = 2^{-n}$, we have

\[
H(X^\Delta) \approx h(f) - \log \Delta = h(f) + n
\] (16.53)

- This means that as $n \to \infty$, $H(X^\Delta)$ gets larger. Why?

Prof. Jeff Bilmes  EE514a/Fall 2013/Information Theory I – Lecture 16 - Nov 21st, 2013  L16 F51/58 (pg.51/58)
Differential vs. Discrete Entropy

- This makes sense. We start with a continuous random variable $X$ and quantize it at an $n$-bit accuracy.
- For a discrete representation to represent $2^n$ values, we expect the entropy to go up with $n$, and as $n$ gets large so would the entropy, but then adjusted by $h(X)$.
- $H(X^\Delta)$ is the number of bits to describe this $n$-bit equally spaced quantization of the continuous random variable $X$.
- $H(X^\Delta) \approx h(f) + n$ says that it might take either more than $n$ bits to describe $X$ at $n$-bit accuracy, or less than $n$ bits to describe $X$ at $n$-bit accuracy, depending on the concentration of $X$.
- If $X$ is very concentrated $h(f) < 0$ then fewer bits. If $X$ is very spread out, then more than $n$ bits.

Joint Differential Entropy

- Like discrete case, we have entropy for vectors of r.v.s
- The joint differential entropy is defined as:

  $$h(X_1, X_2, \ldots, X_n) = - \int f(x_1:n) \log f(x_1:n) dx_1:n$$  \hspace{1cm} (16.54)

- Conditional differential entropy

  $$h(X|Y) = - \int f(x, y) \log f(x|y) dx dy = h(X, Y) - h(Y)$$  \hspace{1cm} (16.55)
Entropy of a Multivariate Gaussian

- When $X$ is distributed according to a multivariate Gaussian distribution, i.e.,

$$X \sim \mathcal{N}(\mu, \Sigma) = \frac{1}{|2\pi\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

then the entropy of $X$ has a nice form, in particular

$$h(X) = \frac{1}{2} \log \left[(2\pi e)^n |\Sigma|\right] \text{ bits} \quad (16.57)$$

- Notice that the entropy is monotonically related to the determinant of the covariance matrix $\Sigma$ and is not at all dependent on the mean $\mu$.
- The determinant is a form of spread, or dispersion of the distribution.

**Entropy of a Multivariate Gaussian: Derivation**

$$h(X) = - \int f(x) \left[ -\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu) - \ln \left((2\pi)^{n/2} |\Sigma|^{1/2}\right) \right]$$

$$= \frac{1}{2} E_f \left[ \text{tr} \left((x - \mu)^T \Sigma^{-1}(x - \mu)\right) \right] + \frac{1}{2} \ln \left((2\pi)^n |\Sigma|\right) \quad (16.58)$$

$$= \frac{1}{2} E_f \left[ \text{tr}((x - \mu)(x - \mu)^T \Sigma^{-1}) \right] + \frac{1}{2} \ln \left((2\pi)^n |\Sigma|\right) \quad (16.59)$$

$$= \frac{1}{2} \text{tr} E_f \left[(x - \mu)(x - \mu)^T\right] \Sigma^{-1} + \frac{1}{2} \ln \left((2\pi)^n |\Sigma|\right) \quad (16.60)$$

$$= \frac{1}{2} \text{tr} \Sigma \Sigma^{-1} + \frac{1}{2} \ln \left((2\pi)^n |\Sigma|\right) \quad (16.61)$$

$$= \frac{1}{2} \text{tr} I + \frac{1}{2} \ln \left((2\pi)^n |\Sigma|\right) \quad (16.62)$$

$$= \frac{n}{2} + \frac{1}{2} \ln \left((2\pi)^n |\Sigma|\right) = \frac{1}{2} \ln \left((2\pi e)^n |\Sigma|\right) \quad (16.63)$$

This uses the “trace trick”, that $\text{tr}(ABC) = \text{tr}(CAB)$. 


The relative entropy (or Kullback-Leibler divergence) for continuous distributions also has a familiar form

\[ D(f||g) = \int f(x) \log \frac{f(x)}{g(x)} \, dx \geq 0 \] (16.65)

We can, like in the discrete case, use Jensen’s inequality to prove the non-negativity of \( D(f||g) \).

Mutual Information:

\[ D(f(X,Y)||f(X)f(Y)) = I(X;Y) = h(X) - h(X|Y) \] (16.66)

\[ = h(Y) - h(Y|X) \geq 0 \] (16.67)

Thus, since \( I(X;Y) \geq 0 \) we have again that conditioning reduces entropy, i.e., \( h(Y) \geq h(Y|X) \).

We still have chain rules

\[ h(X_1, X_2, \ldots, X_n) = \sum_i h(X_i|X_{1:i-1}) \] (16.68)

And bounds of the form

\[ \sum_i h(X_i|X_{1:n}\{i\}) \leq h(X_1, X_2, \ldots, X_n) \leq \sum_i h(X_i) \] (16.69)

For discrete entropy, we have monotonicity. I.e.,

\[ H(X_1, X_2, \ldots, X_k) \leq H(X_1, X_2, \ldots, X_k, X_{k+1}) \]. More generally

\[ f(A) = H(X_A) \] (16.70)

is monotonic non-decreasing in set \( A \) (i.e., \( f(A) \leq f(B), \forall A \subseteq B \)).

Is \( f(A) = h(X_A) \) monotonic? No, consider Gaussian entropy with diagonal \( \Sigma \) with small diagonal values. So \( h(X) = \frac{1}{2} \log \left( (2\pi e)^n |\Sigma| \right) \) can get smaller with more random variables.

Similarly, when some variables independent, adding independent variables with negative entropy can decrease overall entropy.