EE 231A Handout #18, Solution Set 6
Information Theory Wednesday, May 13, 2020
Instructor: Rick Wesel

90 pts
Reading: Section 2.10 and Chapters 9 and 10

1. (12 pts) *A truncated Gaussian.*

   (a) Prove that the Normal density $\phi(x)$ maximizes differential entropy for a fixed second moment σ^2.
   $$\phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/2\sigma^2} \tag{1}$$
   See the proof of Theorem 8.6.5 on pages 254-255 of Cover and Thomas.

   (b) Use this fact to show that a Gaussian input maximizes the mutual information $I(X;Y)$ on the memoryless Gaussian channel $Y = X + Z$ where Z is a Gaussian independent of X and Y.
   $$\begin{aligned} I(X;Y) &= h(Y) - h(Y|X) &(2)\\ &= h(Y) - h(Y - X|X) &(3)\\ &= h(Y) - h(Z) &(4)\\ &\leq \frac{1}{2}\log(2\pi eP) - h(Z) &(5)\end{aligned}$$
   with equality when X is a Gaussian inducing a Gaussian distribution on Y.

   (c) Prove that a truncated Normal density $\tau(x)$ maximizes differential entropy for a fixed second moment under a peak limitation constraint. For example, suppose that the peak limitation constraint is that $x \in (-1, 1)$. The truncated Normal density $\tau(x)$ that maximizes the entropy for a fixed second moment σ^2 is described as follows:
   $$\tau(x) = \begin{cases} \frac{\frac{1}{\sqrt{2\pi\gamma^2}} e^{-x^2/2\gamma^2}}{K} & \text{if } x \in (-1, 1) \\ 0 & \text{otherwise} \end{cases}, \tag{6}$$
   where
   $$K = \int_{-1}^{1} \frac{1}{\sqrt{2\pi\gamma^2}} e^{-x^2/2\gamma^2} dx\,, \tag{7}$$
   and γ^2 is chosen so that
   $$\frac{1}{K}\int_{-1}^{1} x^2 \frac{1}{\sqrt{2\pi\gamma^2}} e^{-x^2/2\gamma^2} dx = \sigma^2\,. \tag{8}$$
   Add limits to all the integrals in the proof of Theorem 8.6.5 on pages 254-255 of Cover and Thomas and note that it all still works.

(d) Now consider a channel that is characterized by the peak limitation $x \in (-1, 1)$ on the transmit signal and additive noise that is a truncated Normal with distribution $\tau(x)$ as described above. Is the mutual information of this channel obviously maximized by a truncated Gaussian input or not? Please explain the issues.

It is not obvious that the mutual information of this channel obviously maximized by a truncated Gaussian input. The key issue is that the sum of two truncated Gaussian's is not a truncated Gaussian so it is not clear that $h(Y)$ is maximized by this choice of input.

2. (10 pts) *Shaping Gain.*

Consider the additive white Gaussian noise channel with power constrained input X and noise $Z \sim \mathcal{N}(0, N)$. For this channel the optimal input distribution is a Gaussian $X_g \sim \mathcal{N}(0, P)$. For practical reasons, a suboptimal X_s that is not a Gaussian but does have $EX_s^2 = P$ is often used.

The maximum shaping gain is that performance improvement that could theoretically be obtained by using X_g rather than X_s. i.e.

$$\text{maximum shaping gain} = I(X_g; Y_g) - I(X_s; Y_s)\,, \tag{9}$$

where $Y_g = X_g + Z$ and $Y_s = X_s + Z$.

Show that the maximum shaping gain may be expressed succinctly as a relative entropy involving f_s (the p.d.f. for Y_s) and f_g (the p.d.f. for Y_g).

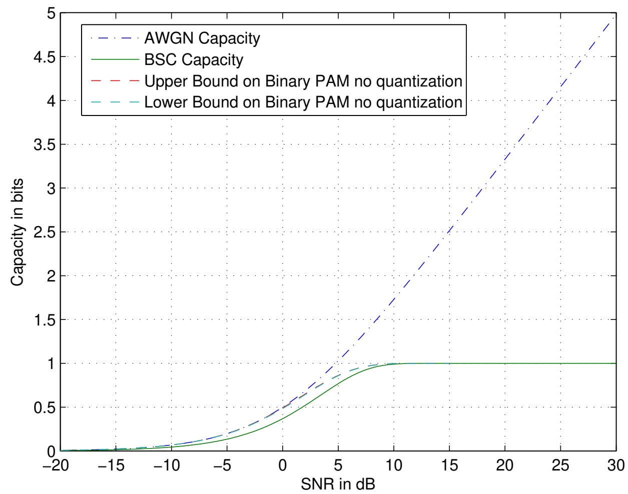
For simplicity we use the natural logarithm below.

$$\begin{aligned}
I(X-g; Y_g) - I(X_s; Y_s) &= h(Y_g) - h(Y_g|X_g) - h(Y_s) + h(Y_s|X_s) && (10)\\
&= h(Y_g) - h(Z) - h(Y_s) + h(Z) && (11)\\
&= h(Y_g) - h(Y_s)) && (12)\\
&= -E_{f_g} \ln f_g(Y) + E_{f_s} \ln f_s(Y) && (13)\\
&= -E_{f_g}\left[\frac{1}{2}\ln\frac{1}{2\pi(P+N)} + \frac{Y^2}{2(P+N)}\right] + E_{f_s} \ln f_s(Y) && (14)\\
&= -E_{f_s}\left[\frac{1}{2}\ln\frac{1}{2\pi(P+N)} + \frac{Y^2}{2(P+N)}\right] + E_{f_s} \ln f_s(Y) && (15)\\
&= -E_{f_s} \ln f_g(Y) + E_{f_s} \ln f_s(Y) && (16)\\
&= D(f_s||f_g) && (17)\\
& && (18)
\end{aligned}$$

The key step is from (14) to (15), which utilizes the fact that Y_s and Y_g have the same second moment.

3. (10 pts) *BSC vs. Gaussian Channel*

   (a) Here is the single plot for (a), (b), and (c) with the Matlab code that generated it.




```
SNR = -20:0.1:30;
snr = 10.^(SNR ./ 10);
C_awgn = 0.5 * log2(1 + snr);
C_bsc = 1 - Hp(Q(sqrt(snr)));
SNR2 = -20:0.1:15;
upper = BINCAPUP(SNR2);
lower = BINCAPLO(SNR2);
clf
h= plot (SNR, C_awgn, '-.', SNR, C_bsc, '-', SNR2, upper, '--', SNR2, lower, '--')
axis ([-20 30 0 5]);
grid on
xlabel ('SNR in dB');
ylabel ('Capacity in bits');
hh=legend(h, 'AWGN Capacity', 'BSC Capacity', 'Upper Bound on Binary PAM no quantization', 'L
axes(hh); refresh;

function [H] = Hp(p)
% function [H] = Hp(p)
% computes the entropy of the binary pmf (p, 1-p).
H = -p .* log2(p) - (1-p) .* log2(1-p);
```

   (b) The quantized output is a function of the unquantized output so the unquantized maximum mutual information curve lies above the quantized curve by the data processing inequality.

Even for the channel with unquantized output , we have

$$
\begin{aligned}
I(X;Y) &= H(X) - H(X|Y) && (19)\\
&\leq H(X) && (20)\\
&\leq \log|X| && (21)\\
&= 1 && (22)
\end{aligned}
$$

Thus the maximum mutual information is always upper bounded by 1. However, its maximum mutual information is lower bounded by the quantized (BSC) channel capacity which converges to 1. Thus, this maximum mutual information curve must also converge to 1. See the plot on the previous page for an exact picture of what is going on.

4. (12 pts) *Shannon, Sensors, A/D convertors*

(a) Write down the capacity of the discrete-time additive white Gaussian noise channel where $Z \sim \mathcal{N}(0, N)$ and transmit energy is constrained to P joules per symbol. Noise energy N is also in units of joules. X, Y, and Z are real-valued. What are the units of capacity?
$C = \frac{1}{2}\log\left(1 + \frac{P}{N}\right)$ bits/symbol.

(b) Write down the capacity of the continuous-time bandlimited AWGN channel with bandwidth W Hz, i.e. signal constrained to $(-W, W)$. The additive noise Z is a white Gaussian process with power spectral density $N_0/2$ watts/Hz. Transmit power is constrained to P watts. Again, what are the units of capacity?

$C = W\log\left(1 + \frac{P}{WN_0}\right)$ bits/second.

(c) Find the capacity of the infinite bandwidth AWGN channel with noise psd and transmitter power constraint as in (b).

$\lim_{W\to\infty} W\log\left(1 + \frac{P}{WN_0}\right) = \log_2 e \frac{P}{N_0}$ as in lecture.

(d) The channel between a remote sensor and a base station is modeled by a discrete time AWGN channel with $N = 1$ joule. Give a lower bound on the amount of transmitter energy required per transmitted bit.

Because capacity increases logarithmically with expended energy, the largest ratio of capacity/energy occurs in the limit of zero energy and zero rate. With $N = 1$ we have
$\lim_{P\to 0} \frac{P}{1/2\log(1+P)} = \frac{2}{\log_2 e}$ (using L'Hospital). The answer for general N is $\frac{2N}{\log_2 e}$.

(e) Actually, the remote sensor has a continuous time AWGN channel to the base station, and must select a bandwidth. What is the correct choice of bandwidth in order to minimize energy per bit? How does the efficiency compare with the discrete time AWGN channel? What are the practical limitations to achieving this ultimate efficiency?

The situation is actually very similar to the prevcious case. To make the situation completely analogous, let $N_0 = 1$ Now the ratio can now be written as

$$\frac{P/W}{\log(1 + P/W)} . \tag{23}$$

We see that P/W plays the role P played in the previous part. Thus the maximum energy efficiency occurs when P/W approaches zero, i.e. when we transmit with infinite bandwidth. The limit for general N_0 is $\frac{N_0}{\log_2 e}$. This is identical in form to the discrete time limit with $N_0/2$ playing the role of N. One practical limitation to achieving infinite (or very large) bandwidth is the implementation of a very fast A/D converter. (see next question)

(f) An A/D converter is designed to have 1 bit of resolution at the highest possible sampling rate on a continuous AWGN channel where the noise psd is $N_0/2$ with $N_0 = 10^{-9}$ watts/Hz and the received signal power is 3 watts. This A/D converter may employ extensive processing and require extensive latency. Furthermore, the transmitted waveform is carefully designed to allow the A/D converter the best possible performance. What is the highest sampling rate that is theoretically possible while still truly achieving one bit of resolution? You may assume that the A/D is preceded by an ideal lowpass filter with cutoff frequency W at half the sampling rate.

This question is really asking you to find the largest symbol rate $2W$ for which the band-limited continuous-time capacity is one bit per *symbol*, with $P = 3$ and $N_0 = 10^{-9}$.

$$1/2 \log\left(1 + \frac{P}{N_0 W}\right) \geq 1 \tag{24}$$

$$\frac{3}{10^{-9} W} \geq 3 \tag{25}$$

$$10^9 \geq W \tag{26}$$

The highest possible sampling frequency is $2W$ or $2 \times 10^9 Hz$.

5. (8 pts) *Parallel channels and waterfilling.* By the result of Section 10.4, it follows that we will put all the signal power into the channel with less noise until the total power of noise + signal in that channel equals the noise power in the other channel. After that, we will split any additional power evenly between the two channels.

Thus the combined channel begins to behave like a pair of parallel channels when the signal power is equal to the difference of the two noise powers, i.e., when $2P = \sigma_1^2 - \sigma_2^2$.

6. (8 pts) *Split Ends.*

(a) Using the sufficient statistic Z, recognize that the channel between X and Z is an erasure channel with erasure probability of 0.5 Thus $C = 1 - 0.5 = 0.5$.

(b) We have $X \longrightarrow Y \longrightarrow Z$ and we want to show $X \longrightarrow Z \longrightarrow Y$.

There is no need to worry about characterizing $P(Y|X,Z)$ for incompatible values of X and Z such as $X = 0$ and $Z = c$. For all compatible cases we have the following:

$$\begin{aligned} P(Y|X,Z) &= \begin{cases} \frac{1}{2} & \text{for } y \in \{0,1\} \text{ when } z = a \\ \frac{1}{2} & \text{for } y \in \{2,3\} \text{ when } z = b \\ \frac{1}{2} & \text{for } y \in \{4,5\} \text{ when } z = c \\ 0 & \text{otherwise} \end{cases} && (27) \\ &= P(Y|Z) && (28) \end{aligned}$$

Thus Z is a sufficient statistic of Y for X.

7. (8 pts) *A channel with two independent looks at Y.*

(a)

$$\begin{aligned} I(X;Y_1,Y_2) &= H(Y_1,Y_2) - H(Y_1,Y_2|X) && (29) \\ &= H(Y_1) + H(Y_2) - I(Y_1;Y_2) - H(Y_1|X) - H(Y_2|X) && (30) \\ &\quad \text{(since } Y_1 \text{ and } Y_2 \text{ are conditionally independent given } X) && (31) \\ &= I(X;Y_1) + I(X;Y_2) - I(Y_1;Y_2) && (32) \\ &= 2I(X;Y_1) - I(Y_1;Y_2) \quad \text{(since } Y_1 \text{ and } Y_2 \text{ are conditionally identically distributed)} && (33) \end{aligned}$$

(b) The capacity of the single look channel $X \to Y_1$ is

$$C_1 = \max_{p(x)} I(X;Y_1).$$

The capacity of the channel $X \to (Y_1,Y_2)$ is

$$\begin{aligned} C_2 &= \max_{p(x)} I(X;Y_1,Y_2) && (34) \\ &= \max_{p(x)} 2I(X;Y_1) - I(Y_1;Y_2) && (35) \\ &\leq \max_{p(x)} 2I(X;Y_1) && (36) \\ &= 2C_1. && (37) \end{aligned}$$

Hence, two independent looks cannot be more than twice as good as one look.

8. (9 pts) *Mutual information for correlated normals.*

$$\left[\begin{array}{c} X \\ Y \end{array}\right] \sim \mathcal{N}_2\left(\mathbf{0}, \left[\begin{array}{cc} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{array}\right]\right) \tag{38}$$

Using the expression for the entropy of a multivariate normal derived in class

$$h(X,Y) = \frac{1}{2}\log(2\pi e)^2|K| = \frac{1}{2}\log(2\pi e)^2\sigma^4(1-\rho^2). \tag{39}$$

Since X and Y are individually normal with variance σ^2,

$$h(X) = h(Y) = \frac{1}{2}\log 2\pi e\sigma^2. \tag{40}$$

Hence

$$I(X;Y) = h(X) + h(Y) - h(X,Y) = -\frac{1}{2}\log(1-\rho^2). \tag{41}$$

(a) $\rho = 1$. In this case, $X = Y$, and knowing X implies perfect knowledge about Y. Hence the mutual information is infinite, which agrees with the formula.

(b) $\rho = 0$. In this case, X and Y are independent, and hence $I(X;Y) = 0$, which agrees with the formula.

(c) $\rho = -1$. In this case, $X = -Y$, and again the mutual information is infinite as in the case when $\rho = 1$.

9. (9 pts) *The two look Gaussian channel.*

A quick way to solve this problem is to realize the special cases implied by the ρ values of interest. When $|\rho| = 0$, Z_1 and Z_1 are independent. When $|\rho| = 1$ there is a deterministic linear relationship between Z_1 and Z_2 which simplifies to $Z_1 = \pm Z_2$ when Z_1 and Z_1 and Z_2 have the same mean and variance. See, for example, section 4.8 of [**?**] if you don't know how to deduce these special cases from the covariance matrix K.

(a) When $\rho = 1$, $Z_1 = Z_2$ so $Y_1 = Y_2$. This is equivalent to the single-look Gaussian channel. Let's do the math.

$$\begin{aligned} I(X;Y_1,Y_2) &= h(Y_1,Y_2) - h(Y_1,Y_2|X) && (42)\\ &= h(Y_1) + h(Y_2|Y_1) - h(Y_1|X) - h(Y_2|Y_1,X) && (43)\\ &= h(Y_1) + h(Y_2|Y_1) - h(Y_1|X) - h(Y_2|Y_1) && (44)\\ &= H(Y_1) - h(Z_1) && (45)\end{aligned}$$

So the capacity is simply the 1-D Gaussian channel capacity.

$$C = \frac{1}{2}\log\left(1 + \frac{P}{N}\right)$$

(b) When $\rho = 0$ Z_1 and Z_2 are independent. We showed in class that $Y_1 + Y_2$ is a sufficient statistic in this case. The channel with input X and output $Y_1 + Y_2 = 2X + Z_1 + Z_2$ is a 1-D Gaussian channel with power constraint 4P (since X is scaled by 2) and noise variance 2N (since independent Gaussian variances add). Again the capacity is a 1-D Gaussian channel capacity with the signal and noise powers adjusted appropriately.

$$C = \frac{1}{2}\log\left(1 + \frac{4P}{2N}\right) \tag{46}$$

$$= \frac{1}{2}\log\left(1 + \frac{2P}{N}\right) \tag{47}$$

(c) When $\rho = -1$, $Z_1 = -Z_2$. Consider

$$g(Y_1, Y_2) = \frac{Y_1 + Y_2}{2} = X$$

The noiseless channel with input X and output $g(Y_1, Y2)$ has infinite capacity since X is a continuous random variable which can represent an infinite number of bits with one value.

From the data processing inequality, $I(X; Y_1, Y_2) \geq I(X, g(Y_1, Y_2))$. Thus for this two-look channel, the second look is extremely useful and

$$C = \infty\,.$$

If you want to solve the problem for general K, here is how to do it: The input distribution that maximizes the capacity is always $X \sim \mathcal{N}(0, P)$. You can show this in the usual way. Evaluating the mutual information for this distribution,

$$\begin{aligned} C_2 &= \max I(X; Y_1, Y_2) && (48)\\ &= h(Y_1, Y_2) - h(Y_1, Y_2|X) && (49)\\ &= h(Y_1, Y_2) - h(Z_1, Z_2|X) && (50)\\ &= h(Y_1, Y_2) - h(Z_1, Z_2) && (51) \end{aligned}$$

Now since

$$(Z_1, Z_2) \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} N & N\rho \\ N\rho & N \end{bmatrix}\right),$$

we have

$$h(Z_1, Z_2) = \frac{1}{2}\log(2\pi e)^2 |K_Z| = \frac{1}{2}\log(2\pi e)^2 N^2(1-\rho^2).$$

Since $Y_1 = X + Z_1$, and $Y_2 = X + Z_2$, we have

$$(Y_1, Y_2) \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} P+N & P+\rho N \\ P+\rho N & P+N \end{bmatrix}\right),$$

and

$$h(Y_1, Y_2) = \frac{1}{2}\log(2\pi e)^2 |K_Y| = \frac{1}{2}\log(2\pi e)^2 (N^2(1-\rho^2) + 2PN(1-\rho)).$$

Hence the capacity is

$$C_2 = h(Y_1, Y_2) - h(Z_1, Z_2) \tag{52}$$

$$= \frac{1}{2} \log \left(1 + \frac{2P}{N(1+\rho)} \right). \tag{53}$$