CS3236: Solutions to Tutorial 5
(Continuous Channels)

1. **[Differential Entropies]**

   (a) Assume that a continuous-valued random variable Z has a probability density that is 0 except in the interval $[-a, a]$. Show that the differential entropy $h(Z)$ is upper bounded by $1 + \log_2 a$, with equality if and only if Z is uniformly distributed between $[-a, a]$.

   (Hint: Use a similar trick based on KL divergence to the proof of $H(X) \le \log_2 |\mathcal{X}|$ in the discrete setting.)

   (b) Suppose that X is a Laplace random variable, i.e., $f_X(x) = \frac{1}{2b}\exp(-|x|/b)$ for $x \in \mathbb{R}$. Find the differential entropy $h(X)$.

   (Hint: This distribution satisfies $\mathbb{E}[|X|] = b$.)

   Solution.

   (a) Let f be the density of X, and let g be the uniform density on $[-a, a]$, so that $g(x) = \frac{1}{2a}$ for all $x \in [-a, a]$. Then note that

$$\begin{aligned}\int_{-a}^{a} f(x)\log_2 \frac{f(x)}{g(x)}dx &= \int_{-a}^{a} f(x)\log_2\big(2a \cdot f(x)\big)\\ &= \log_2(2a) + \int_{-a}^{a} f(x)\log_2 f(x)dx\\ &= 1 + \log_2 a - h(X).\end{aligned}$$

   Since the left-hand side is a KL divergence, it is non-negative with equality if and only if $f = g$. Specialized to the above choices of f and g, we get $1 + \log_2 a - h(X) \ge 0$ with equality if and only if f is uniform, as desired.

   (b) We have

$$\begin{aligned}h(X) &= \mathbb{E}\bigg[\log_2 \frac{1}{f_X(X)}\bigg]\\ &= \log_2(2b) + \mathbb{E}[|x|/b]\log_2 e\\ &= \log_2(2be),\end{aligned}$$

   where the last step uses the hint.

2. **[Typical Set in the Continuous Setting]**

   Let $X_1, \dots, X_n$ be i.i.d. continuous random variables with density function f_X. Similarly to the discrete setting, we can define the *typical set*:

$$\mathcal{T}_n(\epsilon) = \Big\{\mathbf{x} \in \mathbb{R}^n \,:\, 2^{-n(h(X)+\epsilon)} \le f_{\mathbf{X}}(\mathbf{x}) \le 2^{-n(h(X)-\epsilon)}\Big\},$$

   where $f_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} f_X(x_i)$. Prove the following properties:

(a) (Equivalent definition) We have $\mathbf{x} \in \mathcal{T}_n(\epsilon)$ if and only if

$$h(X) - \epsilon \le \frac{1}{n}\sum_{i=1}^{n} \log_2 \frac{1}{f_X(x_i)} \le h(X) + \epsilon.$$

(b) (High probability) $\mathbb{P}[\mathbf{X} \in \mathcal{T}_n(\epsilon)] \to 1$ as $n \to \infty$.

(c) (Volume upper bound) $\mathrm{Vol}(\mathcal{T}_n(\epsilon)) \le 2^{n(h(X)+\epsilon)}$.

(d) (Volume lower bound) $\mathrm{Vol}(\mathcal{T}_n(\epsilon)) \ge (1 - o(1))2^{n(h(X)-\epsilon)}$, where $o(1)$ represents a term that vanishes as $n \to \infty$.

Here the volume of a set $\mathcal{A}$ is defined as $\mathrm{Vol}(\mathcal{A}) = \int_{\mathbf{x}\in\mathcal{A}} d\mathbf{x}$.

Solution. *The first part follows by simple re-arranging of the equation in the definition, and the second part follows from the law of large numbers.*

Third property: By the definition of the typical set, if $\mathbf{x} \in \mathcal{T}_n(\epsilon)$ then $f_{\mathbf{X}}(\mathbf{x}) \ge 2^{-n(h(X)+\epsilon)}$. Since any probability is at most one, we have

$$\begin{aligned}
1 &\ge \mathbb{P}[\mathbf{X} \in \mathcal{T}_n(\epsilon)] \\
&= \int_{\mathbf{x}\in\mathcal{T}_n(\epsilon)} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \\
&\ge \int_{\mathbf{x}\in\mathcal{T}_n(\epsilon)} 2^{-n(h(X)+\epsilon)} d\mathbf{x} \\
&= \mathrm{Vol}(\mathcal{T}_n(\epsilon)) \cdot 2^{-n(h(X)+\epsilon)}.
\end{aligned}$$

Re-arranging gives the third property.

Fourth property: By the definition of the typical set, if $\mathbf{x} \in \mathcal{T}_n(\epsilon)$ then $f_{\mathbf{X}}(\mathbf{x}) \le 2^{-n(h(X)-\epsilon)}$. Writing property 2 as $\mathbb{P}[\mathbf{X} \in \mathcal{T}_n(\epsilon)] = 1 - o(1)$, we obtain

$$\begin{aligned}
1 - o(1) &= \mathbb{P}[\mathbf{X} \in \mathcal{T}_n(\epsilon)] \\
&= \int_{\mathbf{x}\in\mathcal{T}_n(\epsilon)} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \\
&\le \int_{\mathbf{x}\in\mathcal{T}_n(\epsilon)} 2^{-n(h(X)-\epsilon)} d\mathbf{x} \\
&= \mathrm{Vol}(\mathcal{T}_n(\epsilon)) \cdot 2^{-n(h(X)-\epsilon)}.
\end{aligned}$$

Re-arranging gives the forth property.

3. **[Fading Channel]**

Consider a model where we have additive *and* multiplicative noise:

$$Y = XV + Z,$$

where Z and V are both "noise" random variables. The presence of V is often referred to as *fading*, with applications in wireless communication.

Suppose that X and V are independent. Argue that knowledge of V increases the channel capacity by showing that $I(X; Y|V) \ge I(X; Y)$.

Solution. *We have*

$$\begin{aligned}
I(X; Y|V) &= h(X|V) - h(X|V, Y) \\
&= h(X) - h(X|V, Y) \\
&\ge h(X) - h(X|Y) \\
&= I(X; Y)
\end{aligned}$$

where the second equality is because X is independent of V and the inequality is due to $h(X|V,Y) \leq h(X|Y)$ (conditioning reduces entropy). Hence, the capacity of the fading channel with knowledge of the fading factor V generally improves the capacity.

4. **[Discrete-Input Continuous-Output Channel]**

   Consider the additive channel $Y = X + Z$ with $Z \sim \text{Uniform}[0,a]$ for some $a > 1$, and where X can only take values in $\{0,1\}$ (discrete inputs). The channel capacity formula $C = \max_{P_X} I(X;Y)$ still holds in this case, where $I(X;Y) = H(X) - H(X|Y)$ (in terms of regular entropy) and also $I(X;Y) = h(Y) - h(Y|X)$ (in terms of differential entropy).

   (Hint: The assumption $a > 1$ is very important. When can we say that $H(X|Y=y) = 0$?)

   (a) Let $X = 1$ with probability p, and therefore $X = 0$ with probability $1-p$. Calculate $H(X)$ and $H(X|Y)$, and deduce an expression for $I(X;Y)$ *(Note: As an optional extra, you could try doing this also via $h(Y)$ and $h(Y|X)$)*

   Solution.

   We clearly have $H(X) = H_2(p)$, the binary entropy function. For $H(X|Y)$, first note that if $Y < 1$ we are certain that $X = 0$ and similarly if $Y > a$ we are certain that $X = 1$. So $H(X|Y=y)$ is non-zero only for $y \in [1,a]$ (recall that $a > 1$ by assumption). For any such y, by Bayes' rule, we have

$$\begin{aligned}
\mathbb{P}[X=1|Y=y] &= \frac{\mathbb{P}[X=1]f_{Y|X}(y|1)}{\mathbb{P}[X=0]f_{Y|X}(y|0) + \mathbb{P}[X=1]f_{Y|X}(y|1)} \\
&= \frac{pf_{Y|X}(y|1)}{pf_{Y|X}(y|0) + (1-p)f_{Y|X}(y|1)} \\
&= p,
\end{aligned}$$

   where the last line uses the fact that (by the uniform noise model) $f_{Y|X}(y|x)$ equals the same value wherever it is non-zero, and it is non-zero for both $x = 0$ and $x = 1$ when $y \in [1,a]$.

   Combining the above, we deduce that

$$H(X|Y) = \mathbb{P}[Y \in [1,a]]H(X|Y=y \text{ where } y \in [1,a]) = \frac{a-1}{a}H_2(p),$$

   where we used $\mathbb{P}[Y \in [1,a]] = \frac{a-1}{a}$ (by adding the probability $(1-p)\cdot\frac{a-1}{a}$ event for $x = 0$, and the probability $p \cdot \frac{a-1}{a}$ event for $x = 1$). Finally,

$$I(X;Y) = H(X) - H(X|Y) = \frac{1}{a}H_2(p),$$

   since $\frac{a-1}{a} = 1 - \frac{1}{a}$.

   (b) Maximize your answer from part (a) over p to deduce the channel capacity.

   Solution. *This just amounts to maximizing $H_2(p)$. So the maximizing value is $p = \frac{1}{2}$, and the capacity is $\frac{1}{a}$.*

5. **[Parallel Gaussian Channel]**

   Consider a channel with two inputs (X_1, X_2) and two outputs (Y_1, Y_2), where:

   - $Y_1 = X_1 + Z_1$ with $Z_1 \sim N(0,1)$;
   - $Y_2 = X_2 + Z_2$ with $Z_2 \sim N(0,10)$, and Z_2 is independent of Z_1.

Since the noises are independent, we can think of this as transmitting information separately over two "sub-channels": One from X_1 to Z_1, and the other from $X_1 \to Z_2$. However, we only have an *overall* power constraint of P (i.e., we have some freedom in how to allocate powers $P_1 \geq 0$ and $P_2 \geq 0$ to the two sub-channels, but these must satisfy $P_1 + P_2 \leq P$).

Prof. Smith states the following: "The second channel is a lot noisier, so there is no point in using it. Setting $P_1 = P$ and $P_2 = 0$ is the best we can ever hope to do, regardless of the value of P.". Is he correct?

Solution. *With powers P_1 and P_2, using the Gaussian channel capacity result from the lecture, we can achieve an overall rate of*

$$R = \frac{1}{2}\log_2\left(1 + \frac{P_1}{1}\right) + \frac{1}{2}\log_2\left(1 + \frac{P_1}{10}\right).$$

For small enough P, it can actually be shown that Prof. Smith has the right idea. However, the overall claim is incorrect, because it fails for large P.

For instance, suppose that $P = 100$. If we let $P_1 = 100$ and $P_2 = 0$, we get $R \approx 3.329$ bits/use. But letting $P_1 = 90$ and $P_2 = 10$ gives $R \approx 3.75$ bits/use.

For a complete optimal power allocation solution (to a more general version of this problem), see Section 9.4 of Cover and Thomas.

6. **(Advanced) [Two-Look Gaussian Channel]**

Consider the Gaussian channel with two correlated looks at X; specifically, we have $Y = (Y_1, Y_2)$, where

$$Y_1 = X + Z_1$$
$$Y_2 = X + Z_2,$$

where (Z_1, Z_2) are jointly Gaussian with mean zero, equal variance $\mathbb{E}[Z_1^2] = \mathbb{E}[Z_2^2] = N$, and correlation $\mathbb{E}[Z_1 Z_2] = \sigma^2\rho$ for some correlation coefficient $\rho \in [-1, 1]$.

(a) Show that the channel capacity is $\frac{1}{2}\log_2\left(1 + \frac{2P}{\sigma^2(1+\rho)}\right)$.

(Hint: You may use that the capacity-achieving input distribution is $X \sim N(0, P)$, even if you don't prove it. Also note that we gave a formula for the differential entropy of a multivariate Gaussian in the lecture – you may use the fact that in the bivariate case with covariance matrix $\begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix}$, it simplifies to

$$h(Z_1, Z_2) = \frac{1}{2}\log_2((2\pi e)^2\sigma^4(1 - \rho^2))$$

without having to prove this. The identity $1 - \rho^2 = (1 - \rho)(1 + \rho)$ is also useful.)

(b) Specialize the answer in (a) to $\rho = -1$, $\rho = 0$, and $\rho = 1$, and try to interpret the capacities obtained.

Solution. *(a) As stated in the hint, the input distribution that maximizes the capacity is $X \sim N(0, P)$. Evaluating the mutual information for this distribution, we get*

$$C = I(X; Y_1, Y_2) = h(Y_1, Y_2) - h(Y_1, Y_2|X) = h(Y_1, Y_2) - h(Z_1, Z_2),$$

where the last step follows because the noise variables are independent of the signal (this step is similar to the lecture with only one noise variable). Now since

$$(Z_1, Z_2) \sim N\left(0, \begin{bmatrix} \sigma^2 & \sigma^2\rho \\ \sigma^2\rho & \sigma^2 \end{bmatrix}\right),$$

we have from the hint that

$$h(Z_1, Z_2) = \frac{1}{2}\log_2((2\pi e)^2\sigma^4(1-\rho^2))$$

Since $Y_j = X + Z_j$ *for* $j = 1, 2$*, we also have*

$$(Y_1, Y_2) \sim N\left(0, \begin{bmatrix} P+\sigma^2 & P+\sigma^2\rho \\ P+\sigma^2\rho & P+\sigma^2 \end{bmatrix}\right).$$

We now apply the hint with σ^2 *replaced by* $P+\sigma^2$ *and* ρ *replaced by* $\frac{P+\sigma^2\rho}{P+\sigma^2}$ *to get*

$$\begin{aligned} h(Y_1, Y_2) &= \frac{1}{2}\log_2\left((2\pi e)^2(P+\sigma^2)^2\left(1-\frac{(P+\sigma^2\rho)^2}{(P+\sigma^2)^2}\right)\right) \\ &= \frac{1}{2}\log_2\left((2\pi e)^2\left((P+\sigma^2)^2-(P+\sigma^2\rho)^2\right)\right) \\ &= \frac{1}{2}\log_2\left((2\pi e)^2\left[\sigma^4(1-\rho^2)+2P\sigma^2(1-\rho)\right]\right) \end{aligned}$$

Combining the entropy calculations above with the capacity equation, we get:

$$\begin{aligned} C &= \frac{1}{2}\log_2\left(1+\frac{2P\sigma^2(1-\rho)}{\sigma^4(1-\rho^2)}\right) \\ &= \frac{1}{2}\log_2\left(1+\frac{2P}{\sigma^2(1+\rho)}\right), \end{aligned}$$

where we have used the fact that $1-\rho^2 = (1-\rho)(1+\rho)$.

(b) When $\rho = 1$,

$$C_{\rho=1} = \frac{1}{2}\log_2\left(1+\frac{P}{\sigma^2}\right)$$

which makes sense $C_{\rho=1}$ *is the capacity of the single look channel. This is not surprising, since in this case* $Z_1 = Z_2$ *and hence* $Y_1 = Y_2$*, so once we have one output, the other one tells us nothing extra.*

When $\rho = 0$,

$$C_{\rho=0} = \frac{1}{2}\log_2\left(1+\frac{2P}{\sigma^2}\right)$$

which corresponds to using twice the power in a single look. The capacity is the same as the capacity of the channel $X \to (Y_1+Y_2)$*; this is because* $Y_1+Y_2 = 2X+Z'$ *with* $Z' \sim N(0, 2\sigma^2)$*, so the noise power is* $2\sigma^2$ *and the signal power (including the multiplication by two) is* $4P$.

When $\rho = -1$,

$$C_{\rho=-1} = \infty$$

which is not surprising since in this case $Z_1 = -Z_2$*, so if we add* Y_1 *and* Y_2*, we can recover* X *exactly. Recovering a continuous quantity exactly amounts to recovering infinitely many bits (e.g., consider the binary expansion, which runs on forever).*

7. **[Infinite Capacity]**

   Consider the additive channel $Y = X + Z$ with the usual power constraint $\mathbb{E}[X^2] \le P$, but with a not-so-usual noise distribution:

$$Z = \begin{cases} Z_{\text{Gaussian}} & \text{with probability } 0.9 \\ 0 & \text{with probability } 0.1, \end{cases}$$

where $Z_{\text{Gaussian}} \sim N(0, \sigma^2)$. By studying $\max_{P_X : \mathbb{E}[X^2] \le P} I(X;Y)$, it can be shown that the channel capacity is infinite. Give an alternative proof of this fact by describing a simple coding scheme to transmit infinitely many bits.

(Hint: (i) A single real number can carry infinitely many bits, since its binary expansion goes on indefinitely such as $0.011010101110101010\ldots$*; (ii) If* U *is a Gaussian random variable with non-zero variance, then* $\mathbb{P}[U = u] = 0$ *for any* u *(whereas* $\mathbb{P}[u_1 \le U \le u_2] > 0$ *provided that* $u_2 > u_1$*))*

Solution. *Encode infinitely many bits into a real number* $X = 0.000 \star\star\star\ldots$ *where* $\star\star\star\ldots$ *are the bits, and the zeros are included at the start to ensure satisfying the power constraint (e.g., if* $P = 1$ *then any number starting with* $0.\star$ *is feasible).*

Use the channel n *times, with the same input* X *being used on every use. To decode, look for two symbols* $i \ne j$ *such that* $Y_i = Y_j$*, and declare this value to be the input* X*.*

The probability of getting $Y_i = Y_j$ *by coincidence under* Z_{Gaussian} *is zero, because the Gaussian distribution is continuous. Moreover, for large enough* n*, the probability of there existing* (i,j) *with* $Y_i = Y_j$ *due to the 0-noise event can be made arbitrarily close to one. In this case, the decoder clearly correctly recovers* X*. Therefore, we can recover infinitely many bits in finitely many channel uses, so the capacity is infinite.*