Lecture 15 — March 2

Lecturer: David Tse *Scribe: Vihan L, Burak B, Aykan O, Tung-Yu W, David Z*

15.1 Outline

- Differential Entropy, Divergence, and Mutual Information
- Entropy Maximization
- Capacity of Gaussian Channels

15.2 Recap - Differential Entropy

Last lecture, we introduced a notion of entropy in the setting of continuous random variables known as *differential entropy*. For a continuous random variable Y, the differential entropy of Y is defined as

$$h(Y) \triangleq E\left[\log \frac{1}{f(Y)}\right]$$

and, for another continuous random variable X, we defined the conditional differential entropy as

$$h(Y|X) \triangleq E\left[\log \frac{1}{f(Y|X)}\right].$$

Recall that $h(Y), h(Y|X)$ have units unlike their discrete counterpart. Let's now check other interesting properties.

15.2.1 Mutual Information: Label Invariance

Theorem 1. For a constant a, $I(aX;Y) = I(X;Y)$.

Proof. Recall from last time,

$$I(aX;Y) = h(aX) - h(aX|Y).$$

Furthermore, we showed last time time that $h(aX) = h(X) + \log|a|$. Therefore,

$$I(aX;Y) = h(X) + \log|a| - h(aX|Y).$$

To proceed from here, we need to get a handle on the term $h(aX|Y)$. It will help to reason by analogy and remember that in the discrete case we have

$$H(X|Y) = \sum_y H(X|Y=y)p(y).$$

For continuous random variables, recall that sums are replaced by integrals and probabilities by density functions. Making these modifications, we arrive at the following definition for conditional differential entropy:

$$h(X|Y) = \int_y h(X|Y=y)f(y)dy.$$

By the same reasoning through which we showed $h(aX) = h(X) + \log|a|$ (see the previous lecture) we can conclude that $h(aX|Y=y) = h(X|Y=y) + \log|a|$. Therefore,

$$\int_y h(aX|Y=y)f(y)dy = \int_y \left(h(X|Y=y) + \log|a|\right) f(y)dy = h(X|Y) + \log|a|.$$

Plugging these quantities into the definition of mutual information we have

$$\begin{aligned} I(aX;Y) &= h(aX) - h(aX|Y) \\ &= h(X) + \log|a| - h(X|Y) - \log|a| \\ &= h(X) - h(X|Y) = I(X;Y). \end{aligned}$$

□

As an exercise, the reader is encouraged to consider whether the above result holds for *any* one-to-one function on X or Y. Why would such a result be important? Remember that the capacity of a channel can be described in terms of mutual information. Thus, the aforementioned result would imply that the capacity remains the same if we perform a lossless pre or post-processing to the data we communicate.

15.3 Properties

15.3.1 Chain Rule for Differential Entropy

Let's continue our exploration of differential entropy and ask another very natural question: Does the chain rule hold? That is, does $h(X_1, X_2) = h(X_1) + h(X_2|X_1)$? As it turns out, this result is true and the proof is nearly identical to our derivation of the original chain rule for discrete random variables:

$$\begin{aligned} h(X_1, X_2) &= E\left[\log \frac{1}{f(X_1, X_2)}\right] \\ &= E\left[\log \frac{1}{f(X_1)f(X_2|X_1)}\right] \\ &= E\left[\log \frac{1}{f(X_1)}\right] + E\left[\frac{1}{f(X_2|X_1)}\right] \\ &= h(X_1) + h(X_2|X_1). \end{aligned}$$

15.3.2 Differential Entropy and Divergence

What about divergence? As a refresher, we defined the divergence between two discrete distributions p and q as

$$D(p||q) \triangleq E\left[\log \frac{p(X)}{q(X)}\right] \quad \text{for } X \sim p.$$

A very natural idea is to define divergence the same way in the continuous case, substituting density functions into the appropriate places:

$$D(f||g) \triangleq E\left[\log \frac{f(X)}{g(X)}\right] \quad \text{for } \ X \sim f.$$

One of the most important properties of relative entropy in the discrete case is non-negativity, which we leveraged a number of times in proving other bounds. Thus, we would like to know if the same result holds for relative differential entropy, and the answer turns out to be an affirmative.

Theorem 2 (Cover & Thomas, page 252)**.**

$$D(f||g) \geq 0.$$

Proof. The proof will be very similar to the discrete version we saw in an earlier lecture. Let f and g be two probability density functions and let S denote the support of f. Then,

$$\begin{aligned}
-D(f||g) &= E\left[-\log \frac{g(X)}{f(X)}\right] \\
&\geq -\log E\left[\frac{g(X)}{f(X)}\right] \qquad \text{Jensen's Inequality} \\
&= -\log \int f(x)\frac{g(x)}{f(x)}dx \\
&\geq -\log 1 = 0.
\end{aligned}$$

□

Corollary 1. For continuous random variables X and Y, $I(X;Y) \geq 0$.

Proof. Recall that we can write mutual information as a divergence between a joint density function and the product of the marginal densities. Therefore,

$$I(X;Y) = D(f_{X,Y}(x,y)||f_X(x) \cdot f_Y(y)) \geq 0$$

by Theorem 2 □

Corollary 2. For continuous random variables X and Y, $h(X|Y) \leq h(X)$.

Note that $h(X|Y) = h(X)$ if and only if X and Y are independent.

15.3.3 Concavity of Differential Entropy

Theorem 3. $h(X)$ is concave in the distribution of X.

Proof. Let X be a continuous random variable. Now, let's define an auxiliary random variable $Y \sim Bern(\lambda)$ and let

$$X = \begin{cases} X_1 & \text{if } Y = 0 \\ X_2 & \text{if } Y = 1 \end{cases}.$$

Thus, we have

$$\begin{aligned} h(X) &\geq h(X|Y) \\ &= h(X|Y=1)p(Y=1) + h(X|Y=0)p(Y=0) \\ &= \lambda h(X_1) + (1-\lambda)h(X_2) \end{aligned}$$

□

15.4 Entropy Maximization: Discrete Case

We now transition to entropy maximization. In this section, we review a result that we have seen before regarding the discrete distribution that maximizes entropy. Let X be a discrete random variable on the support set $\{1, 2, \ldots, k\}$. Amongst all discrete distributions p of X on $\{1, 2, \ldots, k\}$, which one has maximum entropy?

As we've seen, the answer is the discrete uniform distribution $U(1, 2, \ldots, k)$. In a previous homework, we showed this result using concavity and label invariance. In this section, we will present another argument using the definition of entropy directly – a line of reasoning we will use again later in this lecture when finding the entropy maximizing distribution in the continuous case.

Theorem 4. For a given alphabet, uniform distribution achieves the maximum entropy.

Proof. Let U denote the discrete uniform distribution on $\{1, 2, \ldots, k\}$ as defined above and let p denote an arbitrary discrete distribution on X. By the non-negativity of relative entropy, we note that

$$D(p||U) \geq 0.$$

Thus,

$$D(p||U) = \sum_x p(x) \log \frac{p(x)}{U(x)} = -H(X) - \sum_x p(x) \log U(x).$$

Now, note that $U(x) = \frac{1}{k}$ for all values of $X \in \{1, 2, \ldots k\}$. Thus, $U(x)$ is in fact a constant so we can pull it out of the above summation and write

$$D(p||U) = -H(X) - \log U(x) \sum_x p(x).$$

Since p is a probability distribution, we observe that $\sum_x p(x) = 1$. Therefore, can substitute another expression equal to 1 in place of $\sum_x p(x)$. The key insight is to substitute in $\sum_x U(x)$ which is also equal to 1 since it is also a probability distribution. Applying this substitution gives us

$$D(p||U) = -H(X) - \log U(x) \sum_x U(x) = -H(X) + H(U)$$

where we define $H(U)$ to be the entropy of a uniformly distributed random variable. Up to this point, we have shown that

$$D(p||U) = -H(X) + H(U)$$

and using the fact that $D(p||U) \geq 0$ we can conclude that

$$H(U) \geq H(X).$$

Since our choice of a probability distribution on X was arbitrary, this completes the proof. □

15.5 Entropy Maximization: Continuous Case

Let's now ask the same question for the continuous setting. That is, we would like to solve the optimization problem

$$\max_f \quad h(X)$$

for a continuous random variable X. After a bit of thought, one realizes that this problem is not very well formulated since the differential entropy can become arbitrary large. In particular, if we take the density function f to be uniform on longer and longer intervals, the differential entropy will approach ∞. To remedy this issue, we introduce an additional constraint. Intuitively, the problem we ran into with our first attempt at formulating the optimization problem is that we could keep "spreading out" a uniform distribution over longer and longer intervals, which thereby increases the differential entropy. We observe that as we stretch a uniform distribution over longer intervals, the variance of the distribution increases as well. Thus, if we introduce a constraint that could control the variance of the density function, then we might be able to circumvent the problem. To that end, we will introduce a *second moment constraint*. Since the second moment $E[X^2]$ is intimately tied to variance, this will help us rectify our earlier issue. Our new optimization problem is now

$$\begin{aligned} \max_f \quad & h(X) \\ \text{subject to} \quad & E[X^2] = \alpha. \end{aligned}$$

for some constant α. Now, we are ready to state and prove the main result of this section.

Theorem 5. The Gaussian distribution achieves maximum differential entropy subject to the second moment constraint.

Proof. We'll follow a similar outline to our prove that the uniform distribution achieves maximum entropy in the discrete case. As we did previously, let's start with divergence. Let $\phi(x)$ denote the Gaussian density function with 0 mean and unit variance. Thus, $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. Let f denote an arbitrary density function on X. The divergence is then

$$D(f||\phi) = E\left[\log \frac{f(X)}{\phi(X)}\right], \quad X \sim f.$$

Using the definition of differential entropy this becomes

$$\begin{aligned} D(f||\phi) &= -h(X) + E\left[\log \frac{1}{\phi(X)}\right] \\ &= -h(X) + E\left[\frac{1}{\sqrt{2\pi}} e^{-X^2/2}.\right] \end{aligned}$$

Simplifying the second term of the above sum gives us

$$D(f||\phi) = -h(X) + \log\sqrt{2\pi} + \frac{\log e}{2} E\left[X^2\right].$$

Observe that the term inside the expectation in the above expression is a constant by our second moment constraint. Using the same trick we employed in the discrete case proof, we can thus change the distribution as long as we preserve the second moment. Thus, we will replace f by a Gaussian density function. To make this substitution explicit, let X_G denote the random variable X now under the Gaussian distribution with mean zero and variance α. With this substitution, observe that

$$E\left[\log \frac{1}{\phi(X)}\right] = E\left[\log \frac{1}{\phi(X_G)}\right] = h(X_G).$$

Therefore, we see that

$$D(f||\phi) = -h(X) + h(X_G) \geq 0$$

by the non-negativity of divergence. Therefore,

$$h(X_G) \geq h(X).$$

□

Remark: Also note that the Gaussian distribution is essentially the unique differential entropy optimizer. Slightly more formally, if f maximizes the differential entropy subject to the second moment constraint, then f is equal to the Gaussian distribution almost everywhere.

15.6 The Gaussian Channel Revisited

With our understanding of maximum entropy distributions, let's turn our attention back to the communication problem and conclude this lecture by proving the result stated during the previous class that the capacity of a Gaussian channel is $C = \frac{1}{2}\log(1 + \frac{P}{\sigma^2})$.

Recall the setup of the Gaussian channel. For each input, X_i, the channel outputs Y_i where $Y_i = X_i + Z_i$ where $Z_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$, also $Z_i \perp X$. Furthermore, we impose a power constraint for a block length n that $\frac{1}{n} E\left[\sum_{i=1}^n X_i^2\right] \leq P$. The key insight we now make is that this power constraint is nothing more than a second moment constraint that we worked with in the previous section. If we formulate the channel capacity as an optimization problem as we have done before, we have

$$\begin{aligned} \max_{f_X} \quad & I(X;Y) \\ \text{subject to} \quad & E[X^2] = P. \end{aligned}$$

Expanding the definition of mutual information, we have

$$I(X;Y) = h(Y) - h(Y|X) = h(Y) - h(Z)$$

where the last equality follows from the fact that Z does not depend on X. Thus, we can rewrite our optimization problem as

$$\begin{aligned} \max_{f} \quad & h(Y) \\ \text{subject to} \quad & E[X^2] = P. \end{aligned}$$

which we can rewrite as

$$\begin{aligned} \max_{f} \quad & h(X+Z) \\ \text{subject to} \quad & E[X^2 + Z^2] = P + \sigma^2. \end{aligned}$$

From the result of the previous section, we know that maximum entropy will be achieved if we set $X + Z$ be normally distributed. Since the sum of two Gaussians is a Gaussian, we can achieve this by setting $X \sim \mathcal{N}(0, P)$ (recall that Z is Gaussian by assumption). As before, let X_G denote the random variable X under the Gaussian distribution. Applying this result to the mutual information between X and Y we have

$$I(X;Y) = h(X_G + Z) - h(Z) = \frac{1}{2}\log 2\pi e(P + \sigma^2) - \frac{1}{2}\log 2\pi e\sigma^2$$

which simplifies to

$$\frac{1}{2}\log(1 + \frac{P}{\sigma^2}).$$

Therefore, the capacity of the Gaussian channel is

$$C = \max_{E[X^2] \leq P} I(X;Y) = \frac{1}{2}\log(1 + \frac{P}{\sigma^2})$$

as we stated last lecture.