ECE 6960: Adv. Random Processes & Applications
Lecture Notes, Fall 2010

Lecture 4

Today: (1) Transforms, (2) Entropy, (3) Differential Entropy

- HW 1 due today. HW Set 2 due Sept. 9.

1 Transforms

It is often useful to be able to generate the pdf of a sum of random variables, and this section presents several transforms which have the property that the transform of the pdf, $\mathcal{F}\{f_{X_1}(x) \star f_{X_2}(x)\} = \mathcal{F}\{f_{X_1}(x)\}\mathcal{F}\{f_{X_2}(x)\}$. These transforms also give us an alternate method to calculate the moments of a r.v.

1.1 Characteristic Function

The characteristic function (CF) of a random variable X is defined as $\Phi_x(\omega) = E_X\left[e^{j\omega X}\right]$, equivalently for continuous r.v.s,

$$\begin{aligned}\Phi_x(\omega) &= \int_{-\infty}^{\infty} f_X(x)e^{j\omega x}dx \\ f_X(x) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \Phi_x(\omega)e^{-j\omega x}d\omega\end{aligned}$$

Typically in ECE we use the $e^{-j\omega x}$ in the forward Fourier transform, and $e^{j\omega x}$ in the inverse. Also, in math, they would put $1/\sqrt{2\pi}$ in both. Honestly, it doesn't matter, as long as you are consistent. However, most sources follow this definition of the CF. Table 4.1 has CFs for several r.v.s.

For a discrete r.v., we can also define the CF, as done on page 185 of Leon-Garcia.

1.2 Moment Generating Function

We mentioned in the previous lecture that we can also define the moment generating function (MGF) as $MGF_X(s) = E_X\left[e^{sX}\right] = \Phi_X(-js)$.

Moments can be calculated by taking the derivative of the MGF w.r.t. s, and when done, plugging in $s = 0$. We write this as:

$$E_X\left[X^n\right] = \frac{\partial^n}{\partial s^n} MGF_X(s)\Big|_{s=0}$$

We can also do this directly from the CF, but it is more complex (pardon the pun):

$$E_X\left[X^n\right] = \frac{1}{j^n}\frac{\partial^n}{\partial\omega^n}\Phi_X(\omega)\Big|_{\omega=0}$$

Example: What is the MGF of a Laplace r.v.? Recall the Laplace pdf is $f_X(x) = \frac{\lambda}{2}e^{-\lambda|x|}$.

You know the solution from Table 4.1: $\Phi_X(\omega) = \frac{\lambda^2}{\omega^2+\lambda^2}$, so $MGF_X(s) = \frac{\lambda^2}{-s^2+\lambda^2}$. You can also use the definition.

Solution: Using the definition of the MGF, and assuming that $s < \lambda$,

$$\begin{aligned} MGF_X(s) &= E_X\left[e^{sX}\right] = \int_{-\infty}^{\infty} e^{sx}\frac{\lambda}{2}e^{-\lambda|x|}dx \\ &= \frac{\lambda}{2}\left[\int_0^{\infty} e^{(s-\lambda)x}dx + \int_{-\infty}^{0} e^{(s+\lambda)x}dx\right] \\ &= \frac{\lambda}{2}\left[\frac{1}{(s-\lambda)}\left(e^{(s-\lambda)x}\Big|_0^{\infty} + \frac{1}{(s+\lambda)}\left(e^{(s+\lambda)x}\Big|_{-\infty}^{0}\right]\right. \\ &= \frac{\lambda}{2}\left[\frac{1}{(s-\lambda)}(-1) + \frac{1}{(s+\lambda)}(1)\right] = \frac{\lambda^2}{\lambda^2 - s^2} \end{aligned}$$

Example: Use the MGF to find the first four moments of the standard normal.

Solution: We know that for X unit-variance zero-mean Gaussian,

$$MGF_X(s) = e^{s^2/2}$$

Taking derivatives and then evaluating at $s = 0$,

$$E_X\left[X\right] = \frac{\partial}{\partial s}MGF_X(s)\big|_{s=0} = se^{s^2/2}\Big|_{s=0} = 0.$$

Next,

$$E_X\left[X^2\right] = \frac{\partial}{\partial s}se^{s^2/2}\Big|_{s=0} = s^2e^{s^2/2} + e^{s^2/2}\Big|_{s=0} = 1.$$

Next,

$$\begin{aligned} E_X\left[X^3\right] &= \frac{\partial}{\partial s}(s^2+1)e^{s^2/2}\Big|_{s=0} \\ &= (s^2+1)se^{s^2/2} + (2s)e^{s^2/2}\Big|_{s=0} = 0. \end{aligned}$$

Next,

$$\begin{aligned} E_X\left[X^4\right] &= \frac{\partial}{\partial s}(s^3+3s)e^{s^2/2}\Big|_{s=0} \\ &= (s^3+3s)se^{s^2/2} + (3s^2+3)e^{s^2/2}\Big|_{s=0} = 3. \end{aligned}$$

1.3 Probability Generating Function

For non-negative, integer-valued discrete random variables, we typically use the "probability generating function", that is, the z-transform of the pmf (again, with a sign change in the exponent. Specifically, for r.v. N we define $G_N(z) = E_N\left[z^N\right]$,

$$G_N(z) = \sum_{k=0}^{\infty} p_N(k) z^k$$

The pmf can be generated from the PGF as follows:

$$p_N(k) = \frac{\partial^n}{\partial z^k} G_N(z)\Big|_{z=0}$$

The book notes that the mean and variance can be calculated from the derivatives of the PGF as well.

Example: Find the PGF for N, a uniform r.v. with $S_N = \{1, 2, ..., L\}$.

Solution:

$$G_N(z) = \sum_{k=1}^{L} \frac{1}{L} z^k = \frac{1}{L} \sum_{k=1}^{L} z^k = \frac{1}{L} \frac{z - z^{L+1}}{1 - z} = \frac{z}{L} \frac{1 - z^L}{1 - z}$$

Example: Find the PGF for N, a geometric r.v., that is, the number of trials up to and including the first success, when trials are independent with success probability p. Then derive the PGF for r.v. M, the number of trials up to and including the rth success (called a negative binomial r.v.).

Solution: Let $q = 1 - p$. From the definition of the PGF,

$$G_N(z) = \sum_{k=1}^{\infty} p(1-p)^{k-1} z^k = pz \sum_{k=1}^{\infty} (qz)^{k-1} = pz \sum_{k=0}^{\infty} (qz)^k = \frac{pz}{1 - qz}$$

Since $M = N_1 + N_2 + \cdots + N_r$ for r independent geometric r.v.s, we should have in the transform domain the product of all the PGFs:

$$G_M(z) = [G_N(z)]^r = \left[\frac{pz}{1 - qz}\right]^r.$$

2 Entropy

Information entropy is the same physical quantity as thermodynamic entropy in physics [1]. Entropy measures disorder or uncertainty in a random variable. Leon-Garcia Section 4.10 does a good job of showing that the entropy, in bits, is the minimum average number of binary questions (bits) required to be answered to establish the outcome of X. Please read this section. Answers to binary questions is essentially what a digital communication system is designed to convey.

Def'n: *Entropy*
Let X be a discrete random variable with pmf $p_X(x_i) = P[X = x]$. Here, there is a finite or countably infinite set S_X, and $x \in S_X$. We will shorten the notation by using p_i as follows:

$$p_i = p_X(x_i) = P[X = x_i]$$

where $\{x_1, x_2, \ldots\}$ is an ordering of the possible values in S_X. Then the entropy of X, in units of bits, is defined as,

$$H[X] = -\sum_i p_i \log_2 p_i \tag{1}$$

Notes:

- $H[X]$ is an operator on a random variable, not a function of a random variable. It returns a (deterministic) number, not another random variable. This it is like $E[X]$, another operator on a random variable.
- Entropy of a discrete random variable X is calculated using the probability values of the pmf of X, p_i. Nothing else is needed.
- Use that $0 \log 0 = 0$. This is true in the limit of $x \log x$ as $x \to 0^+$.
- All "log" functions are log-base-2 in information theory unless otherwise noted. Keep this in mind when reading a book on information theory. The "reason" the units are bits is because of the base-2 of the log. Actually, when theorists use $\log_e$ or the natural log, they express information in "nats", short for "natural" digits.
- I recommend reading Shannon's original discussion in [3].

Example: Binary r.v.

A binary (Bernoulli) r.v. has pmf,

$$p_X(x) = \begin{cases} s, & x = 1 \\ 1-s, & x = 0 \\ 0, & o.w. \end{cases}$$

What is the entropy $H[X]$ as a function of s?
Solution: Entropy is given by (1) and is:

$$H[X] = -s \log_2 s - (1-s) \log_2 (1-s)$$

The solution is plotted in Figure 1.

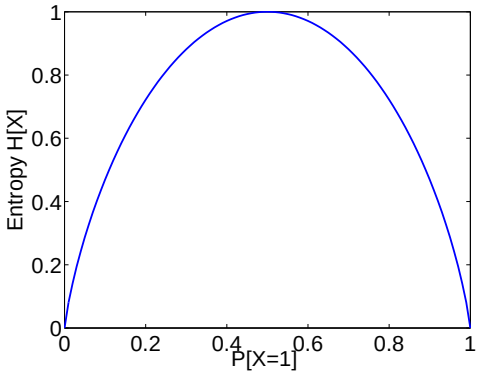

Figure 1: Entropy of a binary r.v.

Example: Non-uniform source with five messages
Some signals are more often close to zero (*e.g.*, audio). Model the r.v. X to have pmf

$$p_X(x) = \begin{cases} 1/16, & x = 2 \\ 1/4, & x = 1 \\ 1/2, & x = 0 \\ 1/8, & x = -1 \\ 1/16, & x = -2 \\ 0, & o.w. \end{cases}$$

What is its entropy $H[X]$? Note it is not $\log_2 5 \approx 2.3$.
Solution:

$$\begin{aligned} H[X] &= \frac{1}{2}\log_2 2 + \frac{1}{4}\log_2 4 + \frac{1}{8}\log_2 8 + 2\frac{1}{16}\log_2 16 \\ &= \frac{15}{8} = 1.875 \text{ bits.} \end{aligned} \tag{2}$$

Other questions:

1. Do you need to know what the symbol set S_X is?

2. Would multiplying X by 2 change its entropy?
3. Would an arbitrary one-to-one function change the entropy of X?

Answers: no!

2.1 Joint Entropy

Def'n: *Joint Entropy*
The joint entropy of two random variables X_1, X_2 with event sets S_{X_1} and S_{X_2} is defined as

$$H[X_1, X_2] = -\sum_{x_1 \in S_{X_1}} \sum_{x_2 \in S_{X_2}} p_{X_1,X_2}(x_1, x_2) \log_2 p_{X_1,X_2}(x_1, x_2) \quad (3)$$

For N joint random variables, $X_1, \ldots, X_N$, entropy is

$$H[X_1, \ldots, X_N] = -\sum_{x_1 \in S_{X_1}} \cdots \sum_{x_N \in S_{X_N}} p_{X_1,\ldots,X_N}(x_1, \ldots, x_N) \log_2 p_{X_1,\ldots,X_N}(x_1, \ldots, x_N)$$

What is the entropy for N i.i.d. random variables? You can show that

$$H[X_1, \ldots, X_N] = -N \sum_{x_1 \in S_{X_1}} p_{X_1}(x_1) \log_2 p_{X_1}(x_1) = NH(X_1)$$

The entropy of N i.i.d. random variables has N times the entropy of any one of them. In addition, the entropy of any N independent (but possibly with different distributions) r.v.s is just the sum of the entropy of each individual r.v.

When r.v.s are not independent, the joint entropy of N r.v.s is less than N times the entropy of one of them. The difference between the entropy assuming the N are independent and the actual joint entropy is called the "mutual information".

Intuitively, if you know some of the N, because of the dependence or correlation, the rest that you don't know become less random. For example, in an image, since pixels are correlated in space, the joint r.v. of several neighboring pixels will have less entropy than the sum of the individual pixel entropies.

2.2 Conditional Entropy

How much additional entropy is in the joint random variables X_1, X_2 compared just to one of them? This is often an important question because it answers the question, "How much additional information do I get from both, compared to just one of them?". We call this difference the conditional entropy, $H[X_2|X_1]$:

$$H[X_2|X_1] = H[X_2, X_1] - H[X_1]. \quad (4)$$

In general, the multi-variate conditional entropy is:

$$H[X_N|X_{N-1},\ldots,X_1] = -\sum_{x_{N-1}\in S_{X_{N-1}}} \cdots \sum_{x_1\in S_{X_1}} p_{X_1,\ldots,X_N}(x_1,x_N)\log_2 p_{X_N|X_{N-1},\ldots,X_1}(x_N|x_{N-1},\ldots,x_1)$$

which is the additional entropy (or information) contained in the Nth random variable, given the values of the $N-1$ previous random variables.

2.3 Entropy Rate

Typically, we're interested in discrete-time random processes, in which we have random variables $X_1, X_2, \ldots$. Since there are infinitely many of them, the joint entropy of all of them may go to infinity as $N \to \infty$. For this case, we are more interested in the rate. How many additional bits, in the limit, are needed for the average r.v. as $N \to \infty$?

Def'n: *Entropy Rate*
The entropy rate of a stationary discrete-time random process, in units of bits per random variable (a.k.a. source output), is defined as

$$H = \lim_{N\to\infty} H[X_N|X_{N-1},\ldots,X_1].$$

It can be shown that entropy rate can equivalently be written as

$$H = \lim_{N\to\infty} \frac{1}{N} H[X_1, X_2, \ldots, X_N].$$

Example: Entropy of English text
Let X_i be the ith letter or space in a common English sentence. What is the sample space S_{X_i}? Is X_i uniform on that space?

What is $H[X_i]$? Solution: I had Matlab read in the text of Shakespeare's *Romeo and Juliet*. See Figure 2(a). For this pmf, I calculated an entropy of $H = 4.1199$. The Proakis & Salehi book [2] says this value for general English text is about 4.3.

What is $H[X_i, X_{i+1}]$? Solution: Again, using Matlab on Shakespeare's *Romeo and Juliet*, I calculated the entropy of the joint pmf of each two-letter combination. This gives me the two-dimensional pmf shown in Figure 2(b). I calculate an entropy of 7.46, which is $2 \cdot 3.73$. For the three-letter combinations, the joint entropy was $10.04 = 3 \cdot 3.35$. For four-letter combinations, the joint entropy was $11.98 = 4 \cdot 2.99$. You can see that the average entropy in bits per letter is decreasing quickly. What is the entropy rate, H? Solution: For $N = 10$, we have $H = 1.3$ bits/letter [2, Section 6.2].

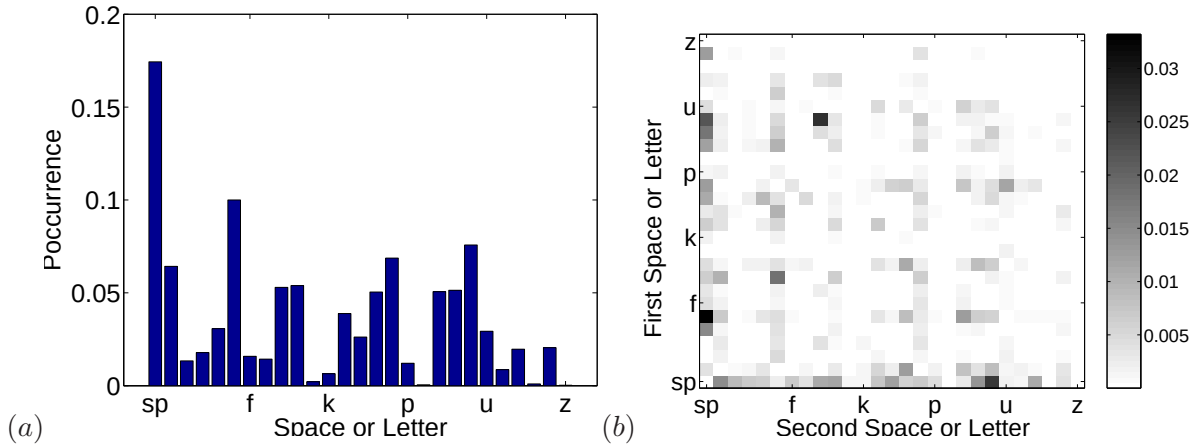

Figure 2: PMF of (a) single letters and (b) two-letter combinations (including spaces) in Shakespeare's *Romeo and Juliet*.

3 Differential Entropy

For continuous random variables, we can't use the definition of entropy in (1). Why not? Well, the result would be an infinite entropy for any cts. r.v., and thus just wouldn't be useful. But we still can compare the entropy of two different random variables using "differential entropy", which is defined as,

$$h(X) = -\int_{-\infty}^{\infty} f_X(x) \ln f_X(x) dx = -E_X\left[\ln f_X(x)\right]. \qquad (5)$$

Note that I use lowercase $h(X)$ instead of $H(X)$, even though the book uses capital H for differential entropy as well. This is more typical of the literature – people don't want to confuse the two concepts. One should not say that $h(X)$ is entropy. Why not? Differential entropy:

- *does not* have an interpretation as the number of binary questions one might answer to convey X.
- *is not* non-negative.
- *is changed* by a one-to-one function $Y = g(X)$.

Example: What is the differential entropy of X, which is uniform between 0 and 1/2?
Solution:

$$h(X) = -\int_0^{0.5} 2 \ln 2 dx = -\ln 2 \approx -0.69$$

However, differential entropy can tell us which of two random variables has more entropy. It can also give us information about mutual information between two cts r.v.s, which is a difference between two differential entropies.

References

[1] E. T. Jaynes. Information theory and statistical mechanics. *Physical Review Series II*, 106(4):620–630, 1957.

[2] J. G. Proakis and M. Salehi. *Communication System Engineering*. Prentice Hall, 2nd edition, 2002.

[3] C. Shannon. A mathematical theory of communications. *Bell System Technical Journal*, 27:379–423 and 623–656, 1948. Available at `http://www.cs.bell-labs.com/cm/ms/what/shannonday/paper.html`.