ECE 7680
Lecture 3 – Some more bounds

Objective: We finish Chapter 2 of C&T with more inequalities that will be useful to us in our later work. Hang in there!

The log sum inequality

In this section we introduce an inequality which will allow us to deduce the concavity (or convexity) of some many useful functions.

Theorem 1 *(Log-sum inequality) For non-negative numbers* $a_1, a_2, \ldots, a_n$ *and* $b_1, b_2, \ldots, b_n$,

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i \right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$$

with equality iff $a_i/b_i = constant$.

Note that the conditions on this theorem are much weaker than for Jensen's inequality, since it is not necessary to have the sets of numbers add up to 1.

Proof The function $f(t) = t \log t$ is strictly convex. (Why?). By Jensen's inequality,

$$\sum_i \alpha_i f(t_i) \geq f(\sum_i \alpha_i t_i)$$

for $\alpha_i/geq0$, $\sum_i \alpha = 1$. Let $\alpha_i = b_i / \sum_{j=1}^{n} b_j$ and $t_i = a_i/b_i$.

$$\sum_i \frac{b_i}{\sum_j b_j} \frac{a_i}{b_i} \log \frac{a_i}{b_i} = \frac{a_i}{\sum_j b_j} \log \frac{a_i}{b_i} \geq \sum_i \frac{a_i}{\sum_j b_j} \log \sum_j \frac{a_i}{\sum_j b_j}$$

□

Using this inequality, we can prove a convexity statement about the relative entropy function.

Theorem 2 *If* (p_1, q_1) *and* (p_2, q_2) *are pairs of probability mass functions then*

$$D(\lambda p_1 + (1-\lambda)p_2 \| \lambda q_1 + (1-\lambda)q_2) \leq \lambda D(p_1 \| q_1) + (1-\lambda) D(p_2 \| q_2)$$

for all $0 \leq \lambda \leq 1$. *That is,* $D(p\|q)$ *is convex in the pair* (p, q).

Proof Recall that

$$D(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)}.$$

Writing out the LHS of the theorem and applying the log sum inequality,

$$\begin{aligned} \sum_x \lambda p_1(x) + (1-\lambda)p_2(x) \log \frac{\lambda p_1(x) + (1-\lambda)p_2(x)}{\lambda q_1(x) + (1-\lambda)q_2(x)} \\ \leq \sum_x \lambda p_1(x) \log \frac{\lambda p_1(x)}{\lambda q_1(x)} + (1-\lambda)p_2(x) \log \frac{(1-\lambda)p_2(x)}{(1-\lambda)q_2(x)} \\ = \lambda D(p_1\|q_2) + (1-\lambda) D(p_2\|q_2). \quad (1) \end{aligned}$$

□

Theorem 3 *$H(p)$ is a concave function of p.*

Proof

$$H(p) = \log |\mathcal{X}| - D(p\|u)$$

where u is the uniform distribution. Since D is convex, H must be concave. □

Proof Here is another more direct proof. Let $X_1 \sim p_1$ and $X_2 \sim p_2$ be two random variables taking values on the same set A. Let θ be a random variable defined as

$$\theta = \begin{cases} 1 & \text{with probability } \lambda \\ 2 & \text{with probability } 1-\lambda. \end{cases}$$

Let $Z = X_\theta$. That is, Z is one or the other of X_1 or X_2 selected at random. Then the distribution of Z is $\lambda p_1 + (1-\lambda)p_2$. (This is the key observation.) Now

$$H(Z) \geq H(Z|\theta),$$

since conditioning reduces entropy. Recalling the definition of conditional entropy,

$$\begin{aligned} H(Z|\theta) &= \sum_\theta p(\theta) H(Z|\theta) = \lambda H(Z|\theta=1) + (1-\lambda) H(Z|\theta=2) \\ &= \lambda H(p_1) + (1-\lambda) H(p_2). \end{aligned}$$

That is,

$$H(\lambda p_1 + (1-\lambda)p_2) \geq \lambda H(p_1) + (1-\lambda) H(p_2).$$

□

The following theorem is important and will be used several times throughout the quarter.

Theorem 4 *Let $(X,Y) \sim p(x,y) = p(x)p(y|x)$. The mutual information $I(X;Y)$ is a concave function of $p(x)$ for fixed $p(y|x)$ and a convex function of $p(y|x)$ for fixed $p(x)$.*

Proof Recall that

$$p(y) = \sum_x p(x,y) = \sum_x p(x)p(y|x),$$

which is a linear function of $p(x)$ for fixed $p(y|x)$. The mutual information can be expanded as

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - \sum_x p(x) H(Y|X=x)$$

By the observation above, if $p(y|x)$ is fixed, then $p(y)$ is a linear function of $p(x)$ and hence $H(Y)$, which is a concave function of $p(y)$ is a concave function of $p(x)$. The second term in the summation above is a linear function of $p(x)$, which does not alter the concavity, hence the difference is a concave function of $p(x)$.

Proof of the second part is accomplished by explicitly forming convex combinations. Fix $p(x)$. Consider $p_1(y|x)$ and $p_2(y|x)$ with corresponding joint distributions $p_1(x,y) = p_1(y|x)p(x)$ and $p_2(x,y) = p_2(y|x)p(x)$ with their respective marginals $p(x), p_1(y)$ and $p(x), p_2(y)$. Let

$$p_\lambda(y|x) = \lambda\lambda p_1(y|x) + (1-\lambda) p_2(y|x)$$

and note that the corresponding joint distribution is also a mixture (since $p(x)$ is fixed)

$$p_\lambda(x,y) = \lambda\lambda p_1(x,y) + (1-\lambda)p_2(x,y).$$

The distribution of Y is a mixture:

$$p_\lambda(y) = \lambda p_1(y) + (1-\lambda)p_2(y).$$

Recall that the mutual information is the relative entropy between the joint distribution and the product distribution. Therefore consider the product distribution,

$$\begin{aligned} q_\lambda(x) &= p(x)p_\lambda(y) = p(x)(\lambda p_1(y) + (1-\lambda)p_2(y)) \\ &\stackrel{\triangle}{=} \lambda q_1(x,y) + (1-\lambda)q_2(x,y). \end{aligned}$$

Then

$$I(X;Y) = D(p_\lambda \| q_\lambda).$$

Since D is a convex function of (p,q), the mutual information must be a convex function of the conditional distribution. □

The data processing inequality

The data processing inequality is a simple but interesting theorem that states (in essence) the following: no matter what processing you do on some data, you cannot get more information out of a set of data than was there to begin with. In a sense, it provide a bound on how much can be accomplished with signal processing.

Definition 1 Random variable X, Y, and Z are said to form a **Markov chain** in that order, denoted by $X \to Y \to Z$ if the conditional distribution of Z depends only on Y and is independent of X. (That is, if we know Y, then knowing X also does not tell us any more than if we only know Y.) If X, Y and Z form a Markov chain then the joint distribution can be written

$$p(x,y,z) = p(x)p(y|x)p(z|y).$$

□

A Markov chain is at the heart of the "state" idea in differential equations and is used commonly in controls. The concept of a state is that *knowing the present state, the future of the system of independent of the past.* In other words, the state provides all the information necessary to move into the future: the necessary initial conditions of the differential equations.

The "conditional independence" idea means

$$p(x,z|y) = \frac{p(x,y,z)}{p(y)} = \frac{p(x,y)p(z|y)}{p(y)} = p(x|y)p(z|y).$$

Note that if $Z = f(Y)$ then $X \to Y \to Z$.

Theorem 5 *(Data processing inequality) If $X \to Y \to Z$ then*

$$I(X;Y) \geq I(X;Z)$$

Interpretation: If we think of Z as being the result of some processing that is done on the data Y, that is, $Z = f(Y)$ for some function, deterministic or random, then there is no function that can increase the amount of information that Y tells about X.

Proof By the chain rule for mutual information we can write

$$\begin{aligned} I(X;Y,Z) &= I(X;Z) + I(X;Y|Z) \\ &= I(X;Y) + I(X;Z|Y) \end{aligned}$$

By the Markov property, since X and Z and independent given Y,

$$I(X;Z|Y) = 0.$$

Since $I(X;Y|Z) \geq 0$ we have

$$I(X;Y) \geq I(X;Z).$$

□

Fano's Inequality

A fundamental operation in communications is estimating a value based on some measurement. That is, suppose a value X is sent through a channel (where it is corrupted) and a value Y is received. Based on that received value we want to determine an estimate of X by performing some function on the observed value Y. Denote the estimate of X by $\hat{X}$:

$$\hat{X} = g(Y).$$

A question of performance now arises naturally: what is the probability that we have estimated the correct value of X. This can be explored in a variety of ways. One of the ways that will be fruitful to us in this class is by Fano's inequality, which relates the probability of error to the conditional entropy $H(X|Y)$. Intuitively, if there is little uncertainty about X when we know Y, then the probability of error should be small. In fact, when $H(X|Y) = 0$, then the probability of error should be zero: there is no uncertainty left over after we observe Y. Fano's inequality makes a quantitative statement to this effect.

Let

$$P_e = \text{probability of error} = \Pr\{\hat{X} \neq X\}.$$

Theorem 6 *(Fano's inequality)*

$$\boxed{H(P_e) + P_e \log(|\mathcal{X}| - 1) \geq H(X|Y).}$$

The inequality can be weakened to

$$1 + P_e \log(|\mathcal{X}|) \geq H(X|Y).$$

Note that if $P_e = 0$ then $H(X|Y) = 0$.

Proof Define the random variable E by

$$E = \begin{cases} 1 & \text{if } \hat{X} \neq X \\ 0 & \text{if } \hat{X} = X. \end{cases}$$

Now expand $H(E, X|Y)$ two different ways using the chain rule:

$$\begin{aligned} H(E, X|Y) &= H(X|Y) + H(E|X, Y) \\ &= H(E|Y) + H(X|E, Y) \end{aligned}$$

Note the following:

$$\begin{aligned} H(E|X, Y) &= 0 \qquad \text{(if we know both } X \text{ and } Y \text{ then we don't make errors)} \\ H(E|Y) &\leq H(E) = H(P_e) \qquad \text{(conditioning reduces entropy)} \\ H(X|E, Y) &= \Pr(E = 0)H(X|Y, E = 0) + \Pr(E = 1)H(X|Y, E = 1) \\ &\leq (1 - P_e)0 + P_e \log(|\mathcal{X}| - 1) \qquad \text{(how many ways to make an error?)} \end{aligned}$$

□