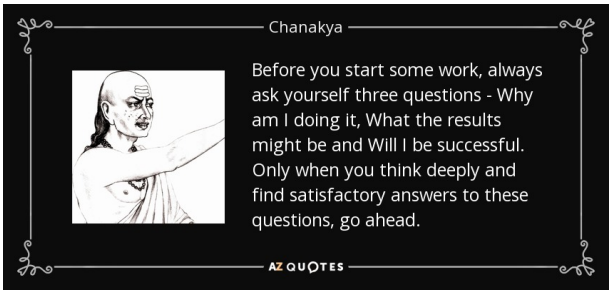

CMI/BVR Probability notes 14

Weierstrass theorem:

Given any real valued continuous function f on $[0,1]$, there is a sequence of polynomials $\{P_n\}$ such that

$$\sup\{|f(x) - P_n(x)| : \ 0 \leq x \leq 1\} \to 0$$

That is, the sequence (P_n) converges to f uniformly on $[0,1]$.

equivalently, given a real valued continuous function f on $[0,1]$ and given an $\epsilon > 0$; we can get a polynomial P such that

$$|f(x) - P(x)| < \epsilon; \quad \text{for all} \quad 0 \leq x \leq 1$$

Shall follow Bernstein's proof. Consider the polynomials

$$P_n(x) = \sum_0^n f(k/n) \binom{n}{k} x^k (1-x)^{n-k}$$

Since $f(k/n)$ is a number the above is a polynomial in x, for each fixed n. These are called Bernstein polynomials associated with f.

Fix $\epsilon > 0$.

Since f is uniformly continuous on $[0,1]$ fix $\delta > 0$ such that; $|f(u) - f(v)| < \epsilon/4$ whenever $|u - v| < \delta$.

Since f is bounded, fix a number $M > 0$ such that $|f(u)| < M$ for all $u \in [0,1]$.

Choose large N such that

$$\frac{1}{N} \frac{M}{\delta^2} < \frac{\epsilon}{4}$$

We now show that $|f(x) - P_n(x)| < \epsilon$ for any $n > N$ and for any $x \in [0,1]$. This will complete the proof.

Accordingly fix any x and any $n > N$. Since binomial probabilities add to one, we see

$$f(x) - P_n(x) = \sum_k \big[f(x) - f(k/n)\big] \quad \binom{n}{k} x^k (1-x)^{n-k}$$

Hence

$$|f(x) - P_n(x)| \le \sum_k \big|f(x) - f(k/n)\big| \quad \binom{n}{k} x^k (1-x)^{n-k}$$

$$= \sum_{k:|x-(k/n)|\le\delta} \quad + \sum_{k:|x-(k/n)|>\delta}$$

$$\le \frac{\epsilon}{4} \sum_{k:|x-(k/n)|\le\delta} \binom{n}{k} x^k (1-x)^{n-k} \quad + 2M \sum_{k:|x-(k/n)|>\delta} \binom{n}{k} x^k (1-x)^{n-k}$$

$$\le \frac{\epsilon}{4} \times 1 \quad + \quad 2M \frac{nx(1-x)}{n^2\delta^2}$$

where we used: choice of δ in the first sum and Chebyshev inequality in the second sum; just note that the second sum is binomial $B(n,x)$ probabilities added over k with $|k - nx| > n\delta$.

$$\le \frac{\epsilon}{4} \quad + \quad \frac{1}{n} \frac{M}{2\delta^2}$$

where we used: $x(1-x) \le 1/4$.

$$\le \frac{\epsilon}{4} \quad + \quad \frac{1}{N} \frac{M}{2\delta^2}$$

where we used: $n > N$.

$$\le \frac{\epsilon}{4} \quad + \quad \frac{\epsilon}{4}$$

where we used: choice of N.

$$< \epsilon$$

Completes proof.

Moments determine f:

Suppose that f is a continuous function on $[0,1]$. Do all the moments, that is, all the numbers $\{\int_0^1 x^n f(x)dx : n \geq 0\}$ determine f?

In other words, if f and g are continuous functions on $[0,1]$ then is the following true?

$$\int_0^1 x^n f(x)dx = \int_0^1 x^n g(x)dx \quad \forall n \geq 0 \quad \longrightarrow \quad f = g$$

Equivalently (considering $f - g$) if h is a continuous function on $[0,1]$

$$\int_0^1 x^n h(x)dx = 0 \quad \forall n \geq 0 \quad \longrightarrow \quad h = 0 \quad ?$$

Yes, it is true. Indeed the hypothesis, by linearity of integral, implies

$$\int_0^1 P(x)h(x)dx = 0$$

for all polynomials P. In particular if you take a sequence (P_n) of polynomials uniformly converging to h on $[0,1]$ then $P_n h$ converges to h^2 uniformly on $[0,1]$ and so by property of integrals,

$$\int_0^1 h^2 = \lim_n \int_0^1 P_n h = 0$$

But since h^2 is a continuous function which is always non-negative, the above equation implies

$$h \equiv 0.$$

Moment generating function:

For a random variable X we define its moment generating function (abbreviated as mgf) by

$$M_X(t) = E(e^{tX})$$

In other words if X is discrete; takes values $(x_i : i \geq 0)$ with probabilities $(p_i : i \geq 0)$ respectively then

$$M_X(t) = \sum_i e^{tx_i}\, p_i$$

If X has density f then

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f(x)dx$$

The above expectation is finite for $t = 0$ and $M_X(0) = 1$.

There are several possibilities.

For every $t \neq 0$ this may be infinite.

This may be finite for every $t \in R$.

Or this may be finite for t in a bounded interval only.

Or this may be finite for only $t \leq 0$ etc etc.

Let us assume the following from now on. There is an $a > 0$ such that $M_X(t) < \infty$ for at least $-a < t < a$.

In such a case it is possible to justify 'pushing differentiation under expectation sign' and conclude the following. For $k \geq 1$ the k-th moment of X exists and is indeed given by

$$E(X^k) = M_X^{(k)}(t)\Big|_{t=0}$$

where the superscript denotes the k-th derivative of M_X. Leaving aside the justification, note that

$$M_X^{(k)}(t) = \left(\frac{d}{dt}\right)^k E(e^{tX}) = E\left[\left(\frac{d}{dt}\right)^k e^{tX}\right] = E(X^k e^{tX})$$

so that

$$M_X^{(k)}(t)\Big|_{t=0} \quad = \quad E(X^k)$$

Because you can generate all moments by taking derivatives at zero, this is called *moment generating* function.

This is similar to probability generating function; which of course, is only for non-negative integer valued random variables; it generates the probabilities.

Markov Chains:

Hope you appreciated the importance of simulation, what is it; why is it useful and in fact why in some cases this is the only way to find answers to questions. You have seen in Professor Karandikar's lectures a nice Markov chain, where the number of points of the state space is not completely understood.

Consider a $K \times K$ chess board. A configuration is a way of placing balls on the squares so that if one box is occupied then all the neighbouring eight

squares should be unoccupied. The collection of such configurations is the state space. This is a finite set but we do not know number of elements in the set and we can not list all states.

However, we have been able to define a Markov chain on this state space. The markov chain has simple transitions: goes from one configuration to nearby (close) configurations. Yet, moves all over. Such dynamics is called Glauber Dynamics. We have been able to simulate the chain and find expected number of occupied sites in a randomly selected configuration.

Back to entropy:

Recall that for a probability $P = (p_1, p_2, \cdots, p_k)$, we define entropy $H(P) = -\sum p_i \log p_i$. Let X be a random variable taking k values with probabilities $\{p_1, \cdots, p_k\}$. We define $H(X) = -\sum p_i \log p_i$. In other words entropy of a random variable is entropy of its distribution (we are considering random variables taking only finitely many values).

Similarly, if we have two random variables (X, Y) then we define $H(X, Y)$ as the entropy of their joint distribution. Thus if the bivariate table consists of the pairs

$$\{(x_i, y_j) : 1 \leq i \leq m; 1 \leq j \leq n\} \quad \text{probabilities} \quad \{p_{ij}\}$$

then

$$H(X, Y) = -\sum p_{ij} \log p_{ij}$$

similarly we can define $H(X_1, X_2, \cdots, X_n)$ for n random variables; simply take the entropy of their joint distribution. This is called joint entropy of $(X_1, X_2, \cdots, X_n)$.

It may bother you that the values seem to not matter for entropy. It depends only on the distribution vector of probabilities. For example consider a cancer drug which may cure or the patient may die on administration of the drug. Suppose for one drug the possibilities are: Cure with probability 0.9 and Die with probability 0.1. Consider another drug for which the possibilities are: Cure with probability 0.1 and Die with probability 0.9. As you see the entropy for both is same.

But intuitively you feel that first drug should have less entropy because it cures with very high probability and should have less uncertainty. The second drug should have high uncertainty or entropy because the chance that it

cures with only very small probability. But such feelings are just consequence of our emotions attached to the possibility of cure. As far as uncertainty is concerned, emotions or meanings of outcomes have no role to play. Only the distribution matters.

If you are unhappy with state of affairs, let us remember that this notion was invented in the context of communication and transmission of messages. Telegraph lines do not know meanings of words; they only recognize letters (through the codes) and some times add noise and send the letters! We have already seen how it makes its appearance in the context of coding.

This is how it appears in transmission. You send one-by-one your letters following a probabilistic rule. There is a noise added and at the other end, the receiver may receive the letter sent with high probability but there is a small chance that noise of the channel may convert it to another letter. To avoid this possibility, the common sense idea is to send each letter three or four times. the receiver has higher chance of correctly decoding the signal. But to achieve higher and higher accuracy you need to repeat each letter a large number of times.

In the early days it was felt that if the transmission rate increases then errors increase. Shannon introduced the concept of Channel capacity and showed, surprisingly, that as long as you are below the limits of channel capacity, you can communicate as error-free as you want. It is precisely this channel capacity that depends on entropy.

Interestingly enough, this concept turned out to be fundamental and is useful in several areas. Returning to entropy, we shall discuss some of its properties and close our discussion.

Fact 1: If P is the K-length probability vector with each entry $1/K$ then $H(P) = \log K$. That is

$$H(\frac{1}{K}, \frac{1}{K}, \cdots, \frac{1}{K}) = \log K$$

Since we are talking about probability vectors, I need not tell you how many times $1/K$ appears in the above display.

Fact 2: For any probability vector of length K, entropy is at most $\log K$; with equality when and only when the vector is the above one. That is

$$H(p_1, p_2, \cdots, p_K) \leq \log K$$

with equality iff each $p_i = 1/K$.

to see this note for $a > 0$; we have $\log a \leq a - 1$ with equality iff $a = 1$. As noted earlier already, for two probability vectors P and Q,

$$\sum p_i \log q_i \leq \sum p_i \log p_i$$

with equality iff the vectors P and Q are same.

This is intuitively obvious. An experiment with 100 outcomes where all outcomes are equally likely must give us 'more uncertain' feeling as compared to experiment with 100 outcomes with some having higher probabilities. You have a suspicion that those with higher probabilities are more likely to occur.

Fact 3: For coin tossing; X takes two values H and T with probabilities p and $(1-p)$, then

$$H(X) = -[p \log p + (1-p) \log(1-p)]$$

Thus if we have fair coin toss, then its entropy is $\log 2$.

$\log 2$ is called a bit, short for binary digit. Thus an experiment with two equally likely outcomes yes/no or 1/0 or H/T has one bit information. We shall return to meaning of this sentence later.

Fact 4: $H(X,Y) \leq H(X) + H(Y)$ with equality iff X, Y are independent.

Decipher the notation, here is the argument. Consider the two probabilities $\{p_{ij}\}$ and $\{p_{i\bullet}p_{\bullet j}\}$

$$\sum_{i,j} p_{ij} \log p_{i\bullet}p_{\bullet j} \leq \sum_{i,j} p_{ij} \log p_{ij}$$

with equality iff $p_{ij} = p_{i\bullet}p_{\bullet j}$ for all (i,j). Thus

$$H(X,Y) = -\sum_{i,j} p_{ij} \log p_{ij}$$

$$\leq -\sum_i (\sum_j p_{ij}) \log p_{i\bullet} - \sum_j (\sum_i p_{ij}) \log p_{\bullet j}$$

$$= -\sum p_{i\bullet} \log p_{i\bullet} - \sum_j p_{\bullet j} \log p_{\bullet j} = H(X) + H(Y)$$

the same argument also shows the following.

Fact 5: If $X_1, X_2, \cdots, X_n$ are random variables on a probability space then

$$H(X_1, X_2, \cdots, X_n) \leq H(X_1) + H(X_2) + \cdots + H(X_n)$$

with equality iff $X_1, X_2, \cdots, X_n$ are independent.

We can define conditional entropy.

Suppose that X, Y are random variables. Given $X = i$ the conditional distribution of Y is

value j with probability $p_{ij}/p_{i\bullet}$.

(Instead of thinking of values of X and Y as x_i, y_j etc we are just thinking of i and j. If you wish instead of i, j you can put x_i, y_j.)

Thus Given $X = i$ the conditional entropy of Y is

$$H(Y|X = i) = -\sum_j \frac{p_{ij}}{p_{i\bullet}} \log \frac{p_{ij}}{p_{i\bullet}}$$

The conditional entropy of Y given X is defined as their average,

$$H(Y|X) = \sum_i p_{i\bullet} \;\; H(Y|X = i)$$

This is same as

$$= -\sum_i \sum_j \;\; p_{i\bullet} \frac{p_{ij}}{p_{i\bullet}} \log \frac{p_{ij}}{p_{i\bullet}}$$

$$= -\sum_{ij} p_{ij} \log \frac{p_{ij}}{p_{i\bullet}}$$

Note that conditional entropy of Y given X is not a function, it is a number. This is in contrast with conditional expectation of Y given X. This last one is a function, which for an outcome ω is calculated as follows. First calculate $X(\omega)$ If it si i then calculate conditional distribution of Y given $X = i$ and calculate expectation of this conditional distribution.

Thus conditional entropy is not a function, it is a number.

Fact 6:

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

This is simple

$$H(X,Y) = -\sum_{ij} p_{ij} \log p_{ij} = -\sum_{ij} p_{ij} \log \left(\frac{p_{ij}}{p_{i\bullet}} \, p_{i\bullet} \right)$$

simplify.

puzzles:

A M Yaglom and I M Yaglom have a book on information theory. I shall discuss three examples from that. They attribute these to Kordemskii. I found these amusing and just puzzle solving arguments. But recently I learnt, from Uma, that this philosophy is more serious than what I thought. it can be used successfully to get bounds for some complicated counting arguments.

Example 1: There are two towns A and B; Every one in A always tells the truth; Every one in B always tells lie. You enter one of these towns and you do not know which town it is. You meet a person. There is a serious problem: you do not know to which town he belongs, there is movement between the towns.

How do you find out where you are. You are allowed to ask questions with only yes/no answers.

Of course one way is to ask two questions:
is $2 \times 2 = 4$
is this town A.

But let us think again. Is it necessary to ask two questions? You want to know answer to: is this A? yes/no. Without any prior information we feel that the chances of the town being $(A)/(\text{not}A)$ are 1/2 and 1/2. Thus entropy is one bit. if you ask one question, you should be able to get one bit of information. So it should be possible to find out with one question.

Yes, ask: Do you live here. Everyone will answer Yes, if this is town A; everyone will answer No if this is B. This one question does.

Of course if you want to know 'which town this is and which town does the person belong' then one question does not suffice. After all there are

four possibilities; two for the town and two for the person. Thus without any prior information, each has chance 1/4 and entropy is two bits. single question gets you only one bit of information.

Example 2: Think of a number from one to ten. How many questions should I ask to know the number? Only yes/no answers will be given.

Without any prior info we believe each of the ten numbers is equally likely to be the number thought; thus ten outcomes each with chance 1/10. Entropy is $\log 10$ (remember base 2); between 3 to 4 bits ($2^3 = 8; 2^4 = 16$). Thus you need to ask four questions. You can consider binary expansion and keep asking for the digits; at most four digits. What is the thought process?

Since the questions have only yes/no answers. The first question should be: is the number in A where $A \subset \{1, 2, \cdots, 10\}$. What A would you select.

Suppose A has k elements. Since each number is equally likely to be selected your question would have Yes answer with probability $k/10 = p_1$ and No answer with probability $(10 - k)/10 = p_2$. if you want to get maximum information you should select k in such a way that p_1, p_2 are nearly equal. Thus $k = 5$ is best. Thus for example Q1 is: is it among $\{1, 3, 5, 7, 9\}$, of course, you can take any subset of five elements.

You can continue this way.

Example 3: suppose there are 25 coins of which 24 are of equal weight and one lighter. How many weighings are necessary to locate the counterfeit coin. You have only a beam balance and no weights available. you can only weigh by putting coins in two scales.

To start with, since any one of the 25 coins could be the lighter one, entropy or uncertainty in the experiment is $\log 25$.

Each experiment that we can do consists of taking m and n coins in the two scales and weighing. The outcome has three possibilities, not two:

left scale is heavier/lighter/same as the other.

Thus each experiment gives $\log 3$ bits of information. Thus each weighing

can clear at most $\log 3$ bits of uncertainty. So you need at least

$$\frac{\log 25}{\log 3} = \log_3(25) \sim 3$$

weighings.

If we want to get maximum possible information with a weighing the outcomes should have nearly equal probability. Firstly, it makes no sense to put different number of coins in the scales, you get no information. Think about it.

Thus if you put n coins in each of the scales the outcomes are: (i) left scale (i) lighter (ii) heavier (iii) same as other. This last possibility means lighter coin is in the remaining ones. The corresponding probabilities are respectively $n/25$; $n/25$ and $(25-2n)/25$. To make these as equal as possible; you should put 8 coins in each scale. Thus first weigh with 8 coins in each scale. Understand the consequence. Take the correct group of 8/8/9 coins and repeat. Two more weighings will reveal the coin.