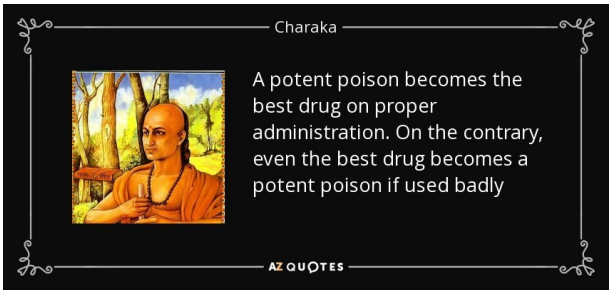

CMI/BVR Probability notes 13

Renewal equation:

Let us now prove the renewal equation. As stated earlier, there is no new idea, we only need to recall what we did in case of random walk and realize that exactly the same works here too. Wish to show the following.

$$\blacksquare \quad p_{ij}^{(n)} = \sum_{m=0}^{n} f_{ij}^{(m)}\, p_{jj}^{(n-m)} \qquad\qquad n \geq 1 \quad \blacksquare$$

Let us define events

$$A = \{X_n = j\}$$
$$A_1 = \{X_1 = j;\ \ X_n = j\}$$
$$A_2 = \{X_1 \neq j, X_2 = j;\ \ X_n = j\}$$

In general

$$A_m = \{X_1 \neq j, \cdots, X_{m-1} \neq j, X_m = j;\ \ X_n = j\}$$

Finally

$$A_n = \{X_1 \neq j, \cdots, X_{n-1} \neq j, X_n = j\}$$

Not difficult to see $(A_m : 1 \leq m \leq n)$ are disjoint events and their union is A. Also

$$P_i(A) = p_{ij}^{(n)}$$
$$P_i(A_m) = P_i\{X_1 \neq j, \cdots, X_{m-1} \neq j, X_m = j\}$$
$$\times P_i\{X_n = j \| X_1 \neq j, \cdots, X_{m-1} \neq j, X_m = j\}$$
$$= f_{ij}^{(m)} p_{jj}^{(n-m)}.$$

Completes proof. Note that $m = 0$ term in the sum is zero and you need not worry.

■ Theorem:

Recurrence is a class property, that is, if a state i is recurrent and j is in the same class as i then j is also recurrent.

Obviously, the above implies that transience is also class property. ■

Fix two integers l and m such that

$$p_{ji}^{(l)} > 0 \qquad p_{ij}^{(m)} > 0 \qquad (\bullet)$$

this is possible because j is in the same class as i. By Chapman-Kolmogorov

$$p_{jj}^{(l+n+m)} \geq p_{ji}^{(l)} p_{ii}^{(n)} p_{ij}^{(m)}.$$

Thus

$$\sum_n p_{jj}^{(l+n+m)} \geq p_{ji}^{(l)} \left\{ \sum_n p_{ii}^{(n)} \right\} p_{ij}^{(m)}.$$

Right side is infinite because i is recurrent and $(\bullet)$. Thus we have

$$\sum_n p_{jj}^{(n)} = \infty$$

completes proof that j is recurrent.

[To start with you can assume that $j \neq i$ in above proof, though that proof works even when $j = i$. Also, since you do not know a priori finiteness of the sums involved (in fact they are not) you can take partial sums, get inequality and take limits.]

Shall now state a result and justify that recurrent state is indeed visited infinitely many times. But before that let us digress and understand the meaning of the quantity

$$\sum p_{ii}^{(n)}.$$

Let us define random variables Z_n as follows: $Z_n = 1$ if $X_n = i$; and $Z_n = 0$ if $X_n \neq i$. In other words Z_n indicates whether you are at i or not on day n. Clearly

$$p_{ii}^{(n)} = P_i(X_n = i) = E_i(Z_n)$$

Thus

$$\sum_n p_{ii}^{(n)} = \sum_n E_i(Z_n) = E_i(\sum_n Z_n)$$

We have interchanged infinite sum and expectation. This is valid for non-negative random variables. What is $\sum Z_n$. It is simply the number of visits to the state i. Thus $\sum_n p_{ii}^{(n)}$ is just the expected number of visits to the state i, starting at i.

[I was a little loose in not specifying whether sum over n extends from $n = 0$ onwards or $n = 1$ onwards. In the first case, yes, it is number of visits including the initial visit. In the second case it is number of returns to i; or number of visits, not counting initial visit. But since our later statements are about this being finite/infinite, this makes no difference. Nonetheless, you should know what you are talking about.]

Thus upshot of the above discussion is the following.

■ Theorem:
A state is recurrent if Expected number of visits to that state — starting in that state — is infinity (same as expected number of returns to that state is infinity).

A state is transient if Expected number of visits to that state — starting from that state — is finite (equivalently, expected number of returns to that state is finite). ■

Define

$$f_{ij}^* = \sum f_{ij}^{(n)}$$

This is the probability of ever returning to j starting from i. Of course, in case $j \neq i$ this is same as probability of ever visiting j starting from i. The quantity f_{ii}^* is the chance of ever returning to i.

I shall define just one more quantity, this will be the last one in Markov chains.

We define for $n \geq 1$,

$$g_{ij}^{(n)} = P_i\{X_m = j \quad \text{for at least } n \text{ values of } m \geq 1\}$$

This is the probability of at least n returns to j starting at i. It makes sense and is useful to make the convention $g_{ij}^{(0)} = 1$

We now have the following.

■ Theorem:

$$(\spadesuit) \quad g_{ij}^{(n)} = f_{ij}^* g_{jj}^{(n-1)}. \qquad\qquad n \geq 1 \quad ■$$

The meaning is that to return to j at least n times, you must return first time and from then on you should return at least $(n-1)$ times. Let us first see the implication of the theorem. Using induction on n, we can easily show the following. Remember $g_{ii}^{(1)} = f_{ii}^*$.

■ Theorem:

$$g_{ii}^{(n)} = (f_{ii}^*)^n \qquad\qquad n \geq 1 \quad ■$$

In particular we have the following.

■ Theorem:
$\lim_n g_{ii}^{(n)}$ is one or zero according as $f_{ii}^* = 1$ or $f_{ii}^* < 1$; in other words, according as i is recurrent or transient. ■

Note that the events

$$\{X_m = j \quad \text{for at least } n \text{ values of } m \geq 1\}$$

are decreasing as n increases. Their intersection is the event

$$\{X_m = j \quad \text{for infinitely many values of } m\}$$

But $A_n \downarrow A$ implies $P_i(A_n) \downarrow P_i(A)$. As a consequence, $\lim g_{ii}^{(n)}$ is nothing but the chances of infinite number of returns to i starting at i.

Thus the previous theorem leads us to the following.

■ Theorem:
Let i be a recurrent state. Then starting from i, the chances of returning to i infinitely many times equals one; that is, we surely return to i infinitely often.

Let i be a transient state. Then starting from i, the chances of returning to i infinitely many times equals zero; that is, we return to i only finitely often. ■

Returning to (♠), we shall not completely prove, here is the idea of proof.

Observe that for $n = 1$ it is simply definition of f_{ij}^*.

Proof for $n = 1$:

$g_{ij}^{(2)}$ is the probability of the union of the following disjoint events (union over $m \geq 1$):

there is a first return to j on day m and after day m there is a visit to j.

Probability of this event equals

P_i(first return to j on day m) $\times$
P_i(at least one visit to j after day $m \| $ you are at j on day m).

$$= f_{ij}^{(m)} \times g_{jj}^{(1)}$$

Now summing over m you get the result.

Proof for general n is simple induction. If you can not get, read and understand the above first.

Though we shall not prove the convergence of P^n in the aperiodic irreducible case, we shall prove a weaker result which is equally important.

■ Theorem:
Let P be any stochastic matrix of finite order.
(i) Then there is a stochastic matrix Q such that

$$\frac{I + P + P^2 + \cdots + P^{n-1}}{n} \longrightarrow Q$$

Convergence is entrywise.
(ii) $PQ = QP = QQ = Q$
(iii) Each row of Q is an invariant distribution for the markov chain.
(iv) If P is irreducible then all rows of Q are same. ■.

Existence of limiting matrix Q is also true in case of countably infinite state space. The limiting matrix may turn out to be zero matrix. The rows may thus be not probability vectors.

Proof of (i):
Let us put

$$Q_n = \frac{I + P + P^2 + \cdots + P^{n-1}}{n}$$

Since P is stochastic, so are its powers and so are their averages. Thus each Q_n is a stochastic matrix. In particular, all entries of Q_n are between zero and one. Thus we can get

$$n_1 < n_2 < n_3 < \cdots\cdots$$

such that the sequence

$$Q_{n_1}(i,j), \quad Q_{n_2}(i,j), \quad Q_{n_3}(i,j), \cdots\cdots$$

converges for each (i,j). In other words

$$Q_{n_1}, \quad Q_{n_2}, \quad Q_{n_3}, \cdots\cdots \to Q$$

for a matrix Q.

Note

$$Q_n P = \frac{P + P^2 + \cdots + P^n}{n} = \frac{P^n - I}{n} + Q_n$$

Thus

$$Q_{n_i} P = \frac{P^{n_i} - I}{n_i} + Q_{n_i}$$

Taking limits we get (note our matrices are of fixed finite order)

$$QP = Q$$

Similarly

$$PQ_n = \frac{P + P^2 + \cdots + P^n}{n} = \frac{P^n - I}{n} + Q_n$$

Thus

$$PQ_{n_i} = \frac{P^{n_i} - I}{n_i} + Q_{n_i}$$

showing, after taking limits

$$PQ = Q$$

Once more the same idea

$$Q_n Q = \frac{Q + PQ + P^2 Q + \cdots + P^{n-1} Q}{n} = Q$$

Last equality is from what was proved above. Thus

$$Q_n Q = Q$$

Taking limits we get

$$QQ = Q$$

Finally we intend to show that the full sequence, not just a subsequence, converges. Suppose not so. So there is a pair (i,j) such that the sequence $\{Q_n(i,j) : n \geq 1\}$ does not converge to $Q(i,j)$. Since any way we have

bounded entries, you can get a subsequence of the above which converges to a number different from $Q(i,j)$. Now take a further subsequence of this in which all matrix entries converge. In other words get

$$m_1 < m_2 < m_3 < \cdots\cdots$$

such that

$$Q_{m_2}, \quad Q_{m_2}, \quad Q_{m_3}, \cdots\cdots \to R$$

Of course what ever we proved about Q remains true for R with same proof ands hence we have

$$PR = RP = RR = R$$

We now use the same idea that appeared above, twice more.

$P^iR = R$ so that for each n, $Q_nR = R$. In particular $Q_{n_i}R = R$ and taking limits we get

$$QR = R$$

Again $QP^i = Q$ so that for each m, $QQ_m = Q$. In particular $QQ_{m_i} = Q$ and taking limits we get

$$QR = Q$$

the two displays above show $Q = R$, but remember the (i,j)-th entries (where i, j are fixed earlier) of Q and R differ.

This contradiction completes proof that the full sequence $\{Q_n\}$ converges to Q.

[instead of going via contradiction, you can use a positive argument. Show every subsequence converges and hence the full sequence converges. But then how do you prove this last statement?]

(ii) is already part of the proof above.

(iii) restates $QP = Q$.

(iv) is clear because in the irreducible case there is only one invariant probability. ∎

Entropy:

We shall now discuss a different topic, *Entropy* which is important. This was introduced by Shannon in his fundamental paper; a work that lead to

three theories — information theory, coding and communication. This concept is also related to the concept of Entropy in physics (introduced by Classius?)

For a probability vector (of finite length)

$$P = (p_1, \cdots, p_k)$$

we define its entropy

$$H(P) = -\sum p_i \log p_i$$

Meaningful convention: $0 \log 0 = 0$.

There are several reasons for this definition. This is supposed to measure thee information in an experiment or uncertainty in an experiment. Roughly the idea is that uncertainty in an experiment is just the uncertainty removed by knowing the outcome of the experiment. But how do you make mathematics out of this idea?

Suppose that there are two outcomes of an experiment A, B. The chance of A is 0.99 and chance of B is 0.01. If somebody told you that A happened you are likely to say, I expected it. On the other hand if some one told you that B happened, you will feel surprised and seem to get information from this announcement. Think about it.

For example if some one told you that BVRao is teaching a course in calculus, it would not surprise you and also you do not consider it as much of an information. Imagine some one telling you that BVRao is teaching a course in biology. It should be a surprise to you and you may consider it as an information.

In other words, occurring of an outcome with small probability causes more surprise. It is common practice to assume that outcome with high probability would occur (after all, this is at the basis of the concept of probability); but there is always the possibility that an outcome with a small probability may actually occur (after all every outcome with positive probability does indeed occur if you keep on doing the experiment).

Thus, let us assume that an outcome with probability p causes surprise $\log(1/p)$. Then if our experiment has outcomes $\{1, 2, \cdots, n\}$ with probabilities $\{p_1, \cdots, p_k\}$ respectively then the average surprise in the experiment

is

$$\sum p_i \log(1/p_i) = -\sum p_i \log p_i$$

Why do you measure surprise as $\log(1/p)$? The appearance of $1/p$ is justified from what was explained above. But why did we select logarithmic scale? This is called addative model for surprise, you assume that surprise for independent outcomes adds up and arrive at this. Or you go back to Gibbs factors in physics which use exponentials and identification of entropy as 'logarithm of the number of micro states'. We shall not dwell on these matters. But just for fun, think of using $1/p_i$ itself and see what happens to average surprize.

Here is a another reason for the choice of definition. Instead of outcomes, let us think of an alphabet consisting of m letters. We want to send words through a communication channel to a receiver. Word means a finite sequences of letters. We want to code the letters by strings of zeros and ones. Obviously, different letters must have distinct strings.

Suppose you have three letters a, b, c and you code them by $0, 1, 00$ respectively. If you send 00 then receiver does not know if it is the word aa or the letter c. So it is not enough to have distinct codes for different letters.

The code strings should be such that any code word is uniquely decipherable. This means the following. Given a code for letters, extend it to finite words by obvious concatenation. Then this map from words to strings should be one-to-one. Such codes for letters are called uniquely decipherable codes.

Let us assume that there are m letters. You can choose integer k such that $2^{k-1} < m \leq 2^k$. Consider m distinct strings each of length k and use them to code the m letters. This is uniquely decipherable. If you receive a code word, cut it into blocks of length k and read the letters.

You can, in practice, do better. In Morse code, the frequently occuring letter 'e' is coded by a dot '$\cdot$' and the infrequently occuring 'q' is coded by dash dash dot dash '$- - \cdot -$' . Choose smaller strings for frequently occuring letters. This is how probability enters the discussion.

Let the m letters be $a_1, \cdots, a_m$. Let p_i be the probability of occurrence of the letter a_i. Suppose you code the letter a_i by a string of length n_i. Then your average code length for letters is $\widetilde{n} = \sum n_i p_i$.

Question: how small can this average length $\widetilde{n}$ be?

Answer: not less than $H = -\sum p_i \log p_i$ where log is to the base 2.

We shall now see why. Let us note, in passing, if we are coding by strings from $\{0, 1, 2 \cdots, d-1\}$ it is beneficial to use d as base of the log.

We start with an observation.
If $a > 1$, then by mean value theorem,

$$(\log a - \log 1)/(a-1) < 1$$

so that

$$\log a < a - 1.$$

If $0 < a < 1$,

$$(\log 1 - \log a)/(1-a) > 1$$

so that

$$\log a < a - 1$$

again. Thus

$$\log a \leq a - 1; \qquad \text{for all} \quad a > 0$$

with equality occuring only when $a = 1$. Thus for two positive numbers x and y;

$$\log(x/y) \leq (x/y) - 1$$

with equality iff $x = y$. If we have two probability vectors $\underline{p}$ and $\underline{q}$ of the same length, then

$$\sum p_i \log(q_i/p_i) \leq \sum (q_i - p_i) = 0$$

so that

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

with equality iff $q_i = p_i$ for all i. Note that this holds even if we interpret log as taken to base 2, instead of e. This is because both differ by a multiplicative constant.