Appendix G (g) second attempt PhL 7.29.12

In this second attempt (the sequence section got moved from (f) to (g) and a new section (f) was added about how to compute summed-out p functions) I first write down the useful ensemble approach. A third attempt makes it better, see separate doc.

(g) Experiments with Sequences of Pulse Train Amplitudes

Our pulse trains have a set of amplitudes y_n which form a sequence. If the sequence has N elements, then a sequence can be written $[y_1, y_2.....y_N]$. When many pulse trains are generated in some line code, one finds that the values of y_n in position n of the sequence can vary, and that there is some probability distribution associated with the parameter y_n. Thus, the position n in the pulse train or sequence is associated with a random variable we must call Y_n which takes values y_n.

Here is our first experiment. We have an Apparatus which generates sequences of length N according to some set of rules implemented within the Apparatus (perhaps these are the rules for generating AMI linecode sequences, and the AMI encoder inputs random input streams). Every sequence it generates is "legal" according to its rules. Earlier we discussed an ensemble experiment in which M dice were rolled one at a time and were left on the craps table for study and from what we saw on the table, we were able to construct a distribution function. The ensemble was M dice. In our current **ensemble experiment**, we let the Apparatus crank away and generate an ensemble of M sequences each of length N, and these M sequences are left sitting on the same craps table for us to inspect. This is our "preliminary" ensemble of sequences.

The notation Y_n leads to some confusion. When we had a random variable H which took values h, we might have enumerated the set of h values of the sample space as $\{h_i\}$. To maintain clarity, if we have random variable Y_n which takes values y_n, we should enumerate those values as $(y_n)_i$. Here, y_n is the name of a parameter, just as h was the name of a parameter. Thus, to write down a specific sequence "i" we really should say

$$[(y_1)_i, (y_2)_i.....(y_N)_i] \quad = \text{sequence "i"}$$
$$\text{or} \qquad (G.39)$$
$$[y_1{}^{(i)}, y_2{}^{(i)}..... y_2{}^{(i)}] \quad = \text{sequence "i"}$$

where $(y_2)_i = y_2{}^{(i)}$ is some specific symbol value like "5". We used the latter notation when talking about an ensemble of pulse trains in Section 35.

The Useful Ensemble. To construct a "useful" ensemble of pulse trains, we first imagine generating some number M of pulse trains of length N from our Apparatus as described above. We then take this ensemble of M pulse trains and considerably increase the size of the ensemble by including in it all cyclic permutations (rotations) of the M pulse trains, so now the enlarged ensemble contains I = M*N pulse trains.

Depending on the rules of the Apparatus generating the pulse trains, these rotated pulse trains might be "illegal" at the boundary where the pulse train wraps around on itself. For example with N = 5 we obtain this sequence plus its rotations,

A B C D E

B C D E|**A**
C D E|**A** B
D E|**A** B C
E|**A** B C D . (G.39)

Here the seam is marked by a bar | . If the seam were "illegal", we could probably repair the seam by replacing A with some other symbol Q to get a legal sequence of symbols,

A B C D E
B C D E Q
C D E Q B
D E Q B C
E Q B C D (G.40)

Perhaps several symbols near the seam would have to be "repaired" in this manner, but we shall assume just one as shown above. Notice that each column has at most one bad symbol Q, and that is 1/N of the symbols in the column. If N = 100, then at most 1% of the symbols in a column are wrong after adjustment. If a 2-symbol repair were needed, then at most 2% in a column would be wrong.

We imagine doing this for all M of our initial pulse trains to get our final set of I = M*N pulse trains. In this "useful ensemble", in each vertical column, 1/N of the symbols might be wrong. We of course have in mind that N is very large, but our examples will always have N very small, such as N = 5.

Now we associate each column of our ensemble of sequences with a random variable. For (G.40),

$Y_1\ Y_2\ Y_3\ Y_4\ Y_5$ (G.41)

Fact: The expected value $\langle y_n \rangle \equiv E(Y_n)$ is independent of n. (G.42)

Proof: Each column has at most 1 bad symbol like Q for each initial ensemble sequence. So if there are now I = M*N pulse trains in the ensemble, then in each column there will be at most N bad symbols. When the expected value is computed by summing the elements of the column n and dividing by I,

$$E(Y_n) = (1/I) \sum_{i=1}^{I} y_n^{(i)} = (1/I) \sum_{i=1}^{I} (y_n)_i \qquad \text{// two notations for the same thing}$$

one finds that the sums for different n are almost identical with an error on the order of 1/N. In our example the first column sums to A+B+C+D+E whereas the other columns are Q+B+C+D+E . Of course the exact error depends on what symbol Q was used to do the repair and what the palette of symbol values is, and so on, but the general order is 1/N. As N is made large, this error approaches 0 and then all the columns have the same expected value. **QED**

Now recall the probability distribution for our N=5 ensemble, written in two ways as in (G.5),

$p(a,b,c,d,e) \equiv p(Y_1=a, Y_2= b, Y_3= c, Y_4= d, Y_5= e)$. (G.43)

This is the height of a specific "bar" in the 6D bar chart that is $p(y_1,y_2,y_3,y_4,y_5)$, see Fig G.12

Fact: For our useful ensemble, the probability distribution p(a,b,c ...) has cyclic symmetry. (G.44)

Example: p(a,b,c,d,e) = p(e,a,b,c,d) = p(d,e,a,b,c) = p(e,a,b,c,d)

Proof: This might seem obvious, but a proof is warranted. We show this for the case N = 5; the general case should then be obvious. Using the notation introduced in (G.37) we can imagine evaluating p(a,b,c,d,e) this way, where [....] indicates an ordered sequence of numbers,

$$p(a,b,c,d,e) = (1/I)\,\Sigma_{i=1}^{I}\,(\,[\,a,b,c,d,e] = [\,(y_1)_i, (y_2)_i, (y_3)_i, (y_4)_i, (y_5)_i\,]\,)\,.$$

Since our ensemble contains all cyclic rotations of $[\,(y_1)_i, (y_2)_i, (y_3)_i, (y_4)_i, (y_5)_i\,]$, we don't change the above sum by replacing this [...] with $[\,(y_2)_i, (y_3)_i, (y_4)_i, (y_5)_i, (y_1)_i\,]$, we merely re-order the sum. In our example above, we can rewrite the 5 sequences on the left in this fancier notation,

		$(y_1)_i$	$(y_2)_i$	$(y_3)_i$	$(y_4)_i$	$(y_5)_i$
	$Y_1Y_2Y_3Y_4Y_5$					
α	A B C D E	$(y_1)_1$	$(y_2)_1$	$(y_3)_1$	$(y_4)_1$	$(y_5)_1$
β	B C D E A	$(y_1)_2$	$(y_2)_2$	$(y_3)_2$	$(y_4)_2$	$(y_5)_2$
δ	C D E A B	$(y_1)_3$	$(y_2)_3$	$(y_3)_3$	$(y_4)_3$	$(y_5)_3$
γ	D E A B C	$(y_1)_4$	$(y_2)_4$	$(y_3)_4$	$(y_4)_4$	$(y_5)_4$
ε	E A B C D	$(y_1)_5$	$(y_2)_5$	$(y_3)_5$	$(y_4)_5$	$(y_5)_5$

We now rotate the above columns one column to the left to get

		$(y_2)_i$	$(y_3)_i$	$(y_4)_i$	$(y_5)_i$	$(y_1)_i$
	$Y_1Y_2Y_3Y_4Y_5$					
β	B C D E A	$(y_2)_1$	$(y_3)_1$	$(y_4)_1$	$(y_5)_1$	$(y_1)_1$
δ	C D E A B	$(y_2)_2$	$(y_3)_2$	$(y_4)_2$	$(y_5)_2$	$(y_1)_2$
γ	D E A B C	$(y_2)_3$	$(y_3)_3$	$(y_4)_3$	$(y_5)_3$	$(y_1)_3$
ε	E A B C D	$(y_2)_4$	$(y_3)_4$	$(y_4)_4$	$(y_5)_4$	$(y_1)_4$
α	A B C D E	$(y_2)_5$	$(y_3)_5$	$(y_4)_5$	$(y_5)_5$	$(y_1)_5$

As it most easily seen on the left, this rotation by one column has just reordered the ensemble sequences from the original order α,β,γ,δ,ε to the new order β,γ,δ,ε,α. The same set of samples is included in the sum above regardless of how they are ordered.

Consider then our starting expression,

$$p(a,b,c,d,e) = (1/I)\,\Sigma_{i=1}^{I}\,(\,[\,a,b,c,d,e] = [\,(y_1)_i, (y_2)_i, (y_3)_i, (y_4)_i, (y_5)_i]\,)\,.$$

Note that c is a specific number, as is $(y_5)_3$. Neither of these is a variable. Now suppose the ensemble has the property just noted above, that we can reorder the sum this way,

$$p(a,b,c,d,e) = (1/I)\,\Sigma_{i=1}^{I}\,(\,[\,a,b,c,d,e] \;=\; [(y_2)_i, (y_3)_i, (y_4)_i, (y_5)_i, (y_1)_i]\,)\,.$$

The equality of the two sequences of numbers shown in the truth test [...] = [...] does not change if we alter each sequence in the exact same way, so rewrite again as

$$p(a,b,c,d,e) = (1/I)\,\Sigma_{i=1}^{I}\,(\,[\,e,a,b,c,d] \;=\; [(y_1)_i, (y_2)_i, (y_3)_i, (y_4)_i, (y_5)_i,]\,)\,.$$

But the expression on the right is the definition of p(e,a,b,c,d). Thus we have shown that

$$p(a,b,c,d,e) = p(e,a,b,c,d)\,.$$

We can then show in the same way that p(e,a,b,c,d) = p(d,e,a,b,c) and so on, so in the end all cyclic permutations (rotations) of p(a,b,c,d,e) are equal for our useful ensemble. **QED**

Comment: In the special case that all the sequence random variables Y_n are statistically independent, we can write

$$p(y_1,y_2,y_3,y_4,y_5) = p(y_1)\,p(y_2)\,p(y_3)\,p(y_4)\,p(y_5)\,.$$

In this special case it is obvious that $p(y_1,y_2,y_3,y_4,y_5)$ is symmetric under cyclic permutations, as it is under *any* permutation. A point to note is that we are *not* assuming statistical independence of the Y_n and that therefore there can exist "correlation" between pairs of random variables in our sequences (pulse train amplitudes). Nevertheless, in this general case $p(y_1,y_2,y_3,...)$ for our useful ensemble has cyclic symmetry, and this ensemble is a realizable ensemble from our experiments for large N with only some very small error. As $N\to\infty$, that error goes to 0.

Fact: The probability distribution $p(Y_n = x)$ is independent of n. (G.45)

Proof by Example: When we write $p(Y_n = x) = p(x)$, then p(x) is deceptively independent of n, but we must show this in the true full notation where n dependence is not concealed. So consider using our N= 5 example, where we use (G.7) to obtain $p(Y_3 = x)$ by summing over all the "other variables" of the full 5th order joint probability distribution $p(y_1,y_2,y_3,y_4,y_5)$,

$p(Y_3 = x)$

$= \Sigma_{y1,y2,y4,y5}\; p(Y_1=y_1,Y_2= y_2,Y_3 = x,Y_4= y_4,Y_5= y_5)$ // definition of $p(Y_3 = x)$

$= \Sigma_{y2,y3,y5,y1}\; p(Y_1=y_2,Y_2= y_3,Y_3= x,Y_4= y_5,Y_5= y_1)$ // rename dummy summation variables

$= \Sigma_{y1,y2,y3,y5}\; p(Y_1=y_2,Y_2= y_3,Y_3= x,Y_4= y_5,Y_5= y_1)$ // trivially reorder sums on Σ

$= \Sigma_{y1,y2,y3,y5}\; p(Y_1=y_1,Y_2= y_2,Y_3= y_3,Y_4= x,Y_5= y_5)$ // use cyclic property of p function

$= p(Y_4 = x)$ // definition of $p(Y_4 = x)$

Thus we have shown that $p(Y_3 = x) = p(Y_4 = x)$ and either by iterating or using other cyclic permutations in the middle step above we conclude that all the $p(Y_n = x)$ are equal and thus $p(Y_n = x)$ is independent of n. As usual, this is for our "useful distribution".

Corollary 1: $\langle y_n \rangle \equiv E(Y_n)$ is independent of n, since $E(Y_n) = (1/I)\Sigma_x \; x \; p(Y_n = x)$ and $p(Y_n = x)$ is independent of n. Thus we have an alternate proof of (G.42) above. (G.46)

Corollary 2: $\langle y_n^2 \rangle \equiv E(Y_n^2)$ is independent of n, since $E(Y_n^2) = (1/I)\Sigma_x \; x^2 \; p(Y_n = x)$ and $p(Y_n = x)$ is independent of n. (G.47)

Corollary 3: $\langle f(y_n) \rangle \equiv E(f(Y_n))$ is independent of n, since $E(f(Y_n)) = (1/I)\Sigma_x \; f(x) \; p(Y_n = x)$, etc. See (G.22). (G.48)

Fact: The probability distribution $p(Y_n = x, Y_m = z)$ depends only on m-n mod N. (G.49)

Proof by Example: Using the same example as above, and again applying (G.7), we find that

$p(Y_2 = x, Y_4 = z)$ // m-n = 4-2 = 2

$= \Sigma_{y1,y3,y5} \; p(Y_1=y_1, Y_2= x, Y_3 = y_3, Y_4= z, Y_5= y_5)$ // definition of $p(Y_3 = x)$

$= \Sigma_{y3,y5,y1} \; p(Y_1=y_3, Y_2= x, Y_3 = y_5, Y_4= z, Y_5= y_1)$ // rename dummy sum indices

$= \Sigma_{y1,y3,y5} \; p(Y_1=y_3, Y_2= x, Y_3 = y_5, Y_4= z, Y_5= y_1)$ // trivially reorder sums on Σ

$\qquad\qquad\qquad\qquad\quad \backslash \quad \backslash \quad \backslash \quad \backslash$

$= \Sigma_{y1,y3,y5} \; p(Y_1=y_1, Y_2= y_3, Y_3 = x, Y_4= y_5, Y_5= z)$ // use cyclic property of p function

$= p(Y_3 = x, Y_5 = z)$ // m - n = 5-3 = 2

In general, one has

$$p(Y_n = x, Y_m = z) = p(Y_{n+k} = x, Y_{m+k} = z)$$

again for our useful ensemble.

Corollary: $\langle y_n y_m \rangle \equiv E(Y_n Y_m)$ for $m \neq n$ depends only on m-n mod N. (G.50)

$$E(Y_n Y_m) = (1/I) \; \Sigma_{x,z} \; (xz) \; p(Y_n = x, Y_m = z) = (1/I) \; \Sigma_{x,z} \; (xz) \; p(Y_{n+k} = x, Y_{m+k} = z)$$

$$= E(Y_{n+k} Y_{m+k})$$