Computational Neuroscience
Dr. Sharba Bandyopadhyay
Department of Electronics and Electrical Communication Engineering
Indian Institute of Technology Kharagpur
Week – 06
Lecture – 28

Lecture 28: Basics of Information Theory - I

Welcome. So as introduced in the last lecture, we spoke about taking help of information theory in order to understand the stimulus and response relationship or the transformation from the stimulus to the response. Because as we said that we can only go so far in terms of systems level model or even network models to understand the entire transformation totally biologically and because the higher order models that we will that may be used are require huge amounts of data and it grows actually exponentially as the complexity of the model increases. So we take help of information theory not that it requires less data but it goes and takes it to the limit in the sense that with information theory we can actually provide or quantify complete dependence between two random variables. So remember our stimulus we are treating it as a random variable, the response is also a random variable and in this transformation we have two random variables that are dependent on each other or could be independent of each other if a particular pathway is not encoding for any information about the stimulus in those responses. So this overall dependence, the complete dependence between the two can be quantified using techniques in information theory and that is what we will introduce today.

So if we go back to our block box model we have stimulus as one side input to the black box and with the number of processing elements we are getting to the response. So our random variables are S which is the stimulus and R the response. So in order to find the relationship or the mapping from the stimulus to the response we may need to understand first that how much is it that the response really cares about the stimulus. That is if let us say if this is the entire sort of information so to speak available about the stimulus how much of it is represented in R or in the response.

So this how much or this quantification is what we will talk about. So first we start with the idea of how much randomness or uncertainty there is in a particular random variable. That is if we have let us say stimulus then before the stimulus arrives and we get a response to the neuron there is a huge amount of uncertainty present or if it is only a few set of stimuli that the neuron often sees then it is the uncertainty associated with those stimuli that are present because we do not know

what the stimulus is going to be beforehand a priori unless there is some historical dependence based on previous stimuli and so on. So in any case we that is also part of the process that if there is dependence then we will have the stimulus space in a different way. So let us now just think of this fact that the stimulus is coming once at a time one of the stimuli and there is associated uncertainty about it and we are now observing the response.

Now on observing the response the uncertainty about the stimulus may be reduced. We will get into this in a minute and that is essentially what can be quantified in information theory. So in order to look at how much uncertainty is reduced we must have to define what uncertainty about a random variable in this case stimulus is. So let us say the stimulus is being a random variable it takes on values S1, S2, S3 and so on up to Sn let us say. Now each of these have a probability associated with it P1, P2, P3 and Pn.

This is assuming that the stimulus is a set of discrete elements that is the stimulus can be one of these discrete stimuli S1 to Sn and each of these are have a probability of occurrence P1 to Pn being at random variable. So this can either be defined by the experimenter in the sense that we are interested in understanding only a set of stimulus let us say in the auditory system a particular set of sounds whether a neuron is encoding the presence of a set of sounds only and whether the responses encode in different manner each of the sounds. So in that case it is those S1 to Sn that we will be using in the experiment and recording responses for each of those. And now the probability of S1 and probability of S2 and all the stimuli are in the hands of the experimenter. And in that case usually the idea is to give a uniform probability to each of them S1 to Sn that is 1 over n to each of them.

However we can also I mean depending on the question we can also approach it in this way that in reality in the natural world these stimuli S1 to Sn may be occurring with certain probabilities only based on how the stimulus statistics are over in general in the real world. And then we can associate the probabilities accordingly to each of those stimuli S1 to Sn. So that is one side of it and so depending on how we choose the P1, P2 to Pn that depends on the question and the uncertainty of the stimulus is dependent on that. And so in earlier work I mean in the middle of the last century around 1950s, 1952 Shannon introduced this idea of entropy and further on the whole field of information theory developed from those. And with a priori I mean with some axiomatic properties of uncertainty or measures of uncertainty it could be shown that the measure of an uncertainty which we will call measure of a uncertainty of a random variable which we will say entropy, entropy of the random variable S is how we will write it can be

defined to be the negative sum of P_i log(P_i) where i is varying from 1 to n which are the n stimuli that we have and P_i being the probabilities associated with each of the stimuli.

So this H_S entropy is essentially provides the amount of uncertainty present in a random variable. Now why do we say that I mean to get an idea of why it is so think about it in this way for a moment. Let us say so another additional thing is that if our P_i any of the P_i's is 0 or P_j is 0 then we define that P_i log(P_i), P_j log(P_j) is 0. So when we have the random variable S taking on the value S1 with probability P_1 and all the other stimuli have a probability of 0 that is P_2 to P_n are 0 then given that the sum of P_1 to P_n must be 1 given that it is a distribution our P_1 has to be 1 and this uncertainty based on this formula turns out to be 0 log of 1 log of 1 which is 0. So now you can imagine that if the stimulus can take on only one value which is S1 then there is no uncertainty about the stimulus.

So indeed the entropy is 0. Now in case if you think of it in this way that what is the maximum possible uncertainty that this particular stimulus can have. So if you think of it in this way then we must say that there should not be any of the stimuli occurring with chance less than the any other of them because if something is occurring with less chance than the other or higher chance than the other than any other then there is a difference in uncertainty that is the higher or lower uncertainty about one of the stimuli over the other. So there is some predictability associated with one of the stimuli at least if it has a higher probability of occurrence than the rest of the stimuli. So if so that means the uncertainty is maximum when all the stimuli have equal probability of occurrence that is P_1 equals P_2 equals up to P_n and indeed you can actually show in here that that is the case when that entropy of H_S is maximum with when it has a uniform distribution when S has a uniform distribution and as you can imagine when all of them are equally likely then in the beginning before the experiment is done or before the stimulus takes on a particular value all of them have an equal probability of occurrence and so that gives us the maximum possible uncertainty because if one of them has a higher probability of occurrence that means there is a reduced uncertainty.

So that is a sort of how we can physically think of how this measure entropy depicts or quantifies the amount of uncertainty associated with the stimulus S. So similarly if we now have an additional random variable let us say R then what we can come up with is what is the uncertainty of the stimulus S given the response R that is conditional entropy that is we know that a certain R has occurred that is R is equal to a particular R_j. So for that matter first let us define the random variable R to take on values from R_1, R_2 up to R_j so sorry R_m let us say R_m.

So when the when the response let us say we are doing the experiment and the stimuli is in our hand and as we have said the stimuli can take on values S1 up to Sn with the probabilities P_1 up to P_n and here the response has a probability of some occurrence of let us say Q_1, Q_2 up to Q_m based on an overall observation. So if we are given that the response is a particular R_j then the associated stimuli have a set of the associated stimuli that can give rise to the response R_j need not be the entire set of stimuli only a few of the stimuli may be giving rise to the response R_j or a different probability is associated with each of the each of the stimuli being possible to produce the R_j.

So here in order to go to the conditional case it is defined by essentially the average this whole thing is the average of the uncertainty of S given R equals one particular response R_j and then we multiply or do the weighted average probability of R equals R_j and sum this over the different j's j equal 1 to n. So that is the average uncertainty of the stimulus given the response. So this is the uncertainty in S given a particular response R_j and the average of this whole thing is this conditional entropy H of S given R. So in other words here we are looking at the entropy of S given that we know what response has occurred that means it is essentially showing us the uncertainty remaining in S given that on average the response has occurred. So we started that with the idea that the stimulus has an uncertainty S and once we observe the response given R the uncertainty in the stimulus H of S given R this is the uncertainty remaining in S once R is observed and the difference between the two is the uncertainty of S that is reduced by knowing the response R and that is what we call the mutual information between S and R.

And you can also show in your reading material you will see that this is symmetric in the sense that this is H of R minus H of R given S. So in this case what we are essentially having is okay so we have two random variables we may be observing one of them and by observing one of them we are trying to guess the first one. So we are observing the response and based on the response we are trying to find out which stimulus has occurred with what probability each of the stimuli may have occurred and that is the that is basically this dependence that is there between S and R and the quantification of this relationship that is the entire dependence between the S and R is done by this reduction in uncertainty or mutual information $I_{S,R}$. So in order to understand this a little more we have to think of basically the joint distribution of the two random variables in this case that is S and R. So let us say a toy example there are possible stimuli this side is S and this side is R let us say R can be 0 or 1 that is in our case let us say that the neuron either produces a spike or does not produce a spike the simplest sort of response

space that we can think of in response to stimuli and let us say that there are only three possible stimuli S1, S2 and S3.

So if we design the experiment in such a way and that let us say our S1, S2 and S3 are equally probable then the marginal probabilities of S that is P_S is one third one third and one third. Now if our probability of response being 0 and 1 are equal if stimulus 1 occurs then we have the conditional probability of R equals 0 given S equals S1 that is equal to half and probability of R equals 1 given S equals S1 is also half. So that means the joint probabilities you simply multiply it with the probability of S equals S1 that is $P_{R,S}(R = r, S = s) = P(R = r|S = s) * P(S = s)$ this is from the definition of joint and conditional probability so we have one sixth and one sixth here. Let us say when stimulus S2 occurs mostly it produces a spike that is let us say for probability R equals 1 given S equals S2 is equal to 5 by or let us say 3 by 4 75 percent of the time it produces a spike that is 3 by 4 and let us say obviously then $P(R = 0|S = S_2)$ is one fourth. So again using by multiplying by the probability of S equals S2 we have one fourth here and one twelfth here and let us say now that S3 is such a stimulus that probability of response equal to 1 given S equal to S3 is very small or let us say 3 by 8 not very small or let us make it even smaller 3 by 16 and similarly the probability of R equals 0 given S equals S3 is basically 13 by 16.

So now this here is then multiplying by one third is 1 by 16 and here we have 3 13 by 48. So now if I ask you the question that you do not know what the stimulus is but you know that one of these stimulus S1 S2 and S3 has been played to the neuron or the system and you see that there is a spike. What would be your guess as to which stimulus has occurred? Obviously you will say that the one with the highest probability associated with the spike is the most likely one to have occurred which is S2 that is with a 25 percent chance there is a probability of a spike if S2 occurs but for the other ones there is a lower probability. So the second most probable is S1 and the third most probable is S3. So by choosing S2 you will be most often right compared to by choosing S1 all the time or S3 all the time.

Similarly the opposite way if now if you observe no spike on presentation of stimulus S1 or S2 or S3 and you are asked that which stimulus might have occurred obviously your answer is going to be S3 because that is the most likely one that produces no spike as compared to S2 and S1 the next one being S1 and the least possible one is S2. So in that sense initially before you observe the response 0 or 1 if you were asked what stimulus is going to be played or presented you would have said that each of them has equal likelihood of occurrence because you know a priori that each of the stimuli have a probability one third of occurrence.

So the entropy associated with the stimulus is the entropy associated with this distribution one third one third and one third and when we know the response that is so this is our H_S is the entropy associated with this distribution and once we know the response S equal to either S1 or S2 or S3 then we have $H(S|R = 0) + H(S|R = 1)$ and the average of this that is multiplying by $P(R = 0)$ and this multiplied by $P(R = 1)$ that is our $H(S|R)$. So what you will find is that since now given the stimulus the given the response we know that certain stimuli have higher probability of having been presented than the other ones there is a reduction in uncertainty about the stimulus and that is quantified by this $H(S|R)$ and in order to compute this $H(S|R)$ we need to compute $H(S|R = 0)$ and $H(S|R = 1)$ and we need to have the probability of R equals 0 and R equals 1 which are the marginals by adding the probabilities in the column probability of R equals 0 comes here probability of R equals 1 comes here and with these we can compute this $H(S|R)$ and so with both of these together we can also compute $I(S, R)$ which will be $H(S) - H(S|R)$ which is the mutual information between the stimulus and response.

So here we have talked about primarily random variable stimuli and responses that are discrete in nature we can also have stimulus that is in the continuous domain or a continuous random variable and R is also can be discrete or continuous or even let us say both are continuous because we can say that rate over possible time windows is a continuous random variable that is number of spikes per second. So in that case we have a similar definition that is $I(S, R)$ is given by $H(S) - H(S|R)$. So what we mean by little H is for continuous random variables differential entropy and it is defined similar to the discrete entropy case in this case we have let us say S has the distribution $P(S)$ or density $P(S)$. So in that case $h(S)$ or the differential entropy of stimulus is given by the integral of $P(S)$ negative integral of $P(S)$ log($P(S)$) dS and so this is this differential entropy is not physically the same as what entropy is that is the amount of uncertainty associated with a random variable we cannot say that but because the entropy of a continuous random variable $H(S)$ actually goes off to is undefined goes off to infinity because of we can show that by discretizing the stimulus space with very small bin size and taking in the limit that bin size goes to 0 we can show that the entropy or capital $H(S)$ is becomes infinite. Similarly for $P(R)$ also we have the same definition $h(R)$ and that also is not exactly the uncertainty associated with R there is a difference between them between differential entropy and entropy.

However when we take a look at the mutual information or $I_{S,R}$ in this case the meaning is intact because in this case the reduction in uncertainty remains meaningful because in the limit when we the discrete bin size of the two cases

cancel each other out and we still remain with the meaning of mutual information which is the reduction in the uncertainty by knowing one random variable of the other random variable. And so based on this ideas of mutual information and entropy we will go forward with few other ideas from information theory to go into applications of information theory in our concept of neuronal encoding and decoding. Thank you.