Computational Neuroscience
Dr. Sharba Bandyopadhyay
Department of Electronics and Electrical Communication Engineering
Indian Institute of Technology Kharagpur
Week – 06
Lecture – 30

Lecture 30: Maximally Informative Dimensions

Welcome. So we have covered the basic concepts of information theory, the definitions of mutual information, entropy, KL distance or relative entropy. We have conditional entropy and shown how mutual information is a way to quantify dependence between two random variables that are not necessarily linearly dependent but can be dependent in some other way. And in fact could be even for cases where the random variables are not really continuous or parameterizable and can have discrete kind of elements as their values. So in going forward using to use this in terms of understanding the stimulus response relationship of neurons, we would like to talk a little bit about what we exactly mean by this dependence. So we have looked at our stimulus response correlations.

So stimulus and we have a response. These based on this we have used the CSR, the stimulus response correlation, cross correlation. And this has been the basis for estimating the linear kernel or the spike triggered average. All of these are based on this cross correlation between stimulus and response.

So that means a linear dependent that is a linear correlation between stimulus and response. So in order to understand a little more basically if we think of one stimulus element, forget about a vector of stimulus as we have done in the past. Let us say there is only one stimulus element that is a scalar and we have a rate that is on this axis. So S value on this axis and R value or some version of R value on the other axis. So let us say we know that the stimulus and response are related in a linear manner such that when the stimulus is varying from minus 1 to 1, the response varies from let us say on a in a linear manner with a particular slope M and then intercept C.

So the rate increases in this linear manner. Then we have then this way of determining the model or a linear kernel is useful and we will be able to use that to predict responses to other stimuli. So an extension of this to higher dimensional stimulus space is what we were doing. So here the CSR, the correlation is simply the correlation between the response and the stimulus with this kind of a relationship and it is dependent on how much noise there is in our rate responses because remember the responses rate responses are stochastic and so if we use a number of

stimulus values on this axis, we may have multiple responses at the same stimulus values and so essentially we may get a spread of rates, a cloud of rate responses. If we use each stimulus and each response pair and do a scatter plot, it may be forming this kind of a cloud around that straight line.

What CSR is doing is providing the correlation between R and S. Now so if we have the response to the stimulus between minus 1 and 1 to be of this particular form, the underlying form is let us say quadratic kind of relationship. So that is our response R is basically S^2. So we can remember we are saying that the stimulus has only one element, it is not a vector just a scalar and the response is some factor also must be there. So this rate is equal to AS^2 and so with some additive white Gaussian noise or what have you, I mean the way you want to model, let us say for simplicity it is n that is n that is distributed, 0 mean and variance σ^2 Gaussian normal distribution.

In that case, if we look at the correlation between CSR which is the expectation of R times S minus our expectation of R and expectation of S. In this particular case, if we say that the S is 0 mean, let us say uniformly distributed between minus 1 to 1 and the A is going to be immaterial. So if I instead of R we plug in AS^2 plus noise times S and let us say because expectation of S is 0, this term goes to 0. What we have is expectation of AS^3 plus expectation of n times S and this is expectation of A times expectation of S^3 plus expectation of n times expectation of S. This goes to 0 because our noise is 0 mean Gaussian and so this turns out to be the expectation of S^3 which we will see is very much close to 0 under if we assume certain, if we assume that it is S is 0 mean uniform over 0 to 1, like minus 1 to 1.

So obviously here if now we had plotted our stimulus many we did the experiment many times for different stimulus values. For each of them we obtained a multiple number of rates and if we now plotted a scatter that would appear a cloud like this in this experiment with points scattered around this parabola. Then we will be able to see that there is a clear relationship between them. However, our CSR is providing a 0 value. So there is a clear dependence but no correlation.

No correlation in the sense that it is linear dependence is absent. So that is why for the general case here we have taken a simple example of a quadratic function. There can be arbitrarily many complicated functions and so that is why we need to know if there is any dependence between R and S beyond over and above the linear dependence and that is where the mutual information based quantification of dependence comes in. So in this particular example if we use mutual information we will find that it is indeed non-zero and so that will show us the power of information theory capturing dependence. So the other aspect of it is that here we

have simplified the problem easily in the quadratic case with simply one factor AS^2.

We have removed any linear term there and any constant term there and we have S as a single element vector or rather a scalar. So we have only one parameter A that would be required to be estimated in order to know the dependence between R and S if had we tried to model R with a quadratic function. But as you can imagine as soon as you make S two element vector there are going to be interactions between the different elements of S that has to be included and there will be four factors four parameters that will require to be estimated. And if we go on further with S increasing in dimensions to about a hundred let us say for an auditory neuron generally people or you know scientists look into hundred time bins of one millisecond each preceding particular time point. So if we have hundred time point values of the stimulus then we have ten thousand interaction factors that is quadratic terms and then of course the hundred linear terms and one single scalar constant term.

So you can imagine how the number of parameters will blow up with increase in the stimulus space. So for this reason what people have done is or rather this is based on work by Sharpey et al. in 2004 and neural computation. So here the method that was developed is the assumption is okay so let us say if this is our stimulus space S where we have multiple dimensions S_1, S_2 and so on. Obviously we cannot show all the dimensions in this let us imagine that it is an n dimensional stimulus space and actually only a small subspace of it is what matters to the neuron that is what they call as the relevant subspace.

So what we mean by that is let us say if we are considering an auditory neuron and looking at its encoding of different frequencies. So if we have this as the frequency axis then let us say that this is 1 kilohertz and this is 10 kilohertz or let us say 2, 4, 8 kilohertz. So 1 kilohertz and 8 kilohertz up to here 2, 4 in our log scale. This is the range of frequencies to which a neuron response in the auditory system or if we have a visual neuron or if it is a somatosensory neuron where this represents the entire skin or this rectangle represents the entire visual field in front. Then let us say we have this region which has an about a hundred elements.

We are not considering the time factor in here we are just thinking of static stimuli and that is the number of pixels in this area that can be independently stimulated in order to create responses of a neuron that we are recording from and we want to find out its receptive field model. So that is the overall space. So the number of frequency bins that we will have here if it is let us say 8 bins per octave then we have essentially 32 so this is 3 octaves so we have 24 frequency bins. So basically our stimulus is 24 dimensional in this simple case of frequency

based receptive fields of an auditory neuron. And if we have hundred pixels in that circular region here then we have hundred dimensions of the stimulus.

So the values along each of these dimensions there are the stimuli can have any value along those dimensions or some range of values along each of these dimensions independently or in the auditory case for the 24 here in the hundred. And so each stimulus is a single point in the hundred dimensional or 24 dimensional stimulus space and that is what we are representing by S_1, S_2, S_2 up to S_n. And now let us say within this there is only a few dimensions or a few sub sub region in this over space in this overall space that is in a subspace the responses are really determined based on the stimulus in that particular subspace is a relevant subspace. So we need to be able to find that subspace that is basically providing the responses. So we need to actually determine what kind of stimuli on which the responses are dependent and how can we capture the total dependence between the stimulus and response.

Now when as soon as we are saying dependence we are saying that we will be quantifying the dependence based on mutual information using information theory. So the idea is fairly simple in the sense that we start off with a single vector V in this multi dimensional stimulus space. Let us say this vector maybe I will use a different color to this is the starting vector V. So this is the responses let us say I mean we arbitrarily choose one vector. We want to now vary this V in a manner such that when we look at the relation of the stimulus projected onto this dimension V and the response then the mutual information between them is maximized.

So going over it once more so let us say we have different stimuli sorry let us say we have different stimuli S multiple number of them that are played and there are corresponding probabilities associated with spiking. Let us say the response measure is spike I mean this can be extended to any other thing. So let us say it is a 0 1 yes or no 0 or 1 probability of spike. So that is we have a probability of spike associated with a stimulus. So let us say this is S_1 S_2 and so on.

So this is each of those has a certain probability of spike associated with it and the probability of spike overall probability of spike is the average rate response of the neuron over the entire period of all the stimuli. So this S_1 S_2 can be like the sliding parts of a stimulus throughout a continuous stimulus or they may be distinct stimuli depending on how we are setting up the problem or whatever system we are trying to analyze. So now given this pairs of stimulus and spike what we can we do have is that the probability of spike given a stimulus and we have an overall probability of spike and also a probability of stimulus which is in our design. So this probability of spike given a stimulus is simply a proxy for remem-

ber for mutual information we need only the joint distribution that is probability of spike and each stimulus pair the combination this joint distribution is what is required and the marginals are can be obtained here. So now in this case we can also look at the probability of a particular stimulus given a spike and we can in either case we can come up with the information associated or the mutual information gained by observing a spike and that is simply given by our P of stimulus given spike and $log_2 P$ of stimulus given spike divided by the P of stimulus and this has to be integrated over all possible stimuli and ds or summed if the stimulus is discrete then summed over all of this.

So note that I have changed it a little bit from the usual definition and this can be replaced by the joint and then divided by P of spike that is what we have written here and this integration of S is taking care of the divided by probability of spike that will be needed to get the joint distribution outside in the mutual information formula. So this is defined as our I_{spike} that is information associated with a single spike when there is a set of stimuli. Now what is now the relationship how does the V come in? So we want to change this V in a manner such that if we convert this I_{spike} as a function of V we can keep on increasing the I_{spike} as a function of V. So remember that if we from the data processing inequality if we have let us say stimulus to function of the stimulus and then a response here then our $I(R, S)$ is going to be less than equal to $I(R, S)$. So here we can actually have the arrows in the other way and then get the same result from the data processing.

So here what we are saying that if we are transforming the stimulus into some other space which is we will be what we will be doing is projecting each of the stimuli on to the vector V which is simply the dot product of S and V and we will call this X. So this is our $F(S)$ that we have defined here this X is our $F(S)$. So now using this X instead of S we have the same thing that we have I_{spike} as a function of V the vector V and after having projected the stimulus S on to that particular vector V we have over the entire X now we have a single dimension because X is a scalar ah dX similarly from the S we can get obtain the probability of we can obtain the probability of X given spike just as we obtain the probability of S given spike and log_2 probability of X given spike divided by probability of X. So using this idea what if we take if we now keep on changing the V in a manner such that this $I_{spike}(V)$ keeps on increasing if the $I_{spike}(V)$ keeps on increasing and reaches a maximum then we have found the V or the relevant subspace if it is one dimensional we would have found the relevant subspace that is important to the neuron that will capture the overall dependence in this at least in one particular dimension. So if we use any kind of optimization method like a gradient descent by obtaining the gradient of $I_{spike}(V)$ we can basically now update the V in every

step so we start with an arbitrary V_0 we obtain the $I_{spike}(V)$ at that particular V_0.

So $I_{spike}(V_0)$ at this point we have a gradient of the mutual information over the direction of V. So we over the some direction in the stimulus space we update the V to V_1 such that our we move up the gradient we get $I_{spike}(V_1)$ now and so we keep on moving and finally till we reach a maximum and that final V is what is going to be the relevant subspace had we had only one dimension. Now the transformation from the stimulus to the response is basically that we stimulus dot V then there is a function along over it that is providing the spike or probability of spike and so on or rate in this case. So this function is not available all we are getting is if the model is based on only one dimension along the entire stimulus space then that dimension provides that V particular V provides the maximum mutual information between the stimulus and response had provided we are going through this one dot vector one dot product step. So now this can be extended to simultaneously have multiple V's to start off with and each of those can be optimized to reach a higher mutual information.

Again we are only obtaining the subspace V one particular V or multiple V's depending on how we set up the problem. However that function also is important and that function will have to be determined only empirically from the data that is after having projected the S onto V we get those X's and from the data estimate what rates we were getting for each of the X's and get a numerical value of the function over the different X values or essentially different S values. So that function cannot be obtained through the mutual information idea analytically. So it has to be in empirical form. So and now if we extend that to multiple vectors it gradually again runs into the same problem of estimating many parameters and so on and the optimization becomes more tricky with multiple V's.

However there are examples in the literature where two or even three dimensions have been optimized directly. So I do understand that for the purposes of this course this is a little advanced so this would not be for the entire group of you. Some of you who are interested more in this direction can look into this as a way to model systems. For others you should understand the idea behind it as to how mutual information or information theory is being used in this particular case to model the stimulus response relationship. In later lectures we will be still using information theory and looking at coding and decoding but in a different way not from the perspective of modeling more from the perspective of discrimination by responses of two different stimuli of multiple different stimuli.

So we will take up those studies in our next lectures where we extend the ideas of information theory in understanding neural coding and decoding. Thank you.