

Is Statistical Independence a Proper Goal for Neural Network Preprocessors?

William B Levy and Dawn M. Adelsberger-Mangan
 Department of Neurological Surgery Box 420
 University of Virginia Health Sciences Center
 Charlottesville, Virginia 22908

1 Abstract

It is generally believed that maintaining information in a neural network is important and that minimizing statistical dependency is also beneficial. Here we show some examples in which predictions generated at individual neurons benefit from incomplete subsets of the possible inputs (which results in lost information) and an example showing the benefits of statistical dependency in the input space. Finally, we introduce the idea that a locally controlled synaptogenesis might select subsets from an input space, thereby allowing a neuron to conditionally preprocess its inputs in a local and biologically plausible manner.

2 Introduction

Quite some time ago, Barlow (59) suggested that the purpose of processing in the brain is to produce representations of lower redundancy or, to put it another way, of lower statistical dependency (59). This theme has been continued in Barlow's laboratory (Földiák 90) and also seized upon by many other researchers (Levy 85, Linsker 90, Adelsberger-Mangan and Levy 93). Included among the motivations for minimal statistical dependency is superior performance of a preprocessed input set that will be used for prediction in a supervised system downstream from the preprocessor (Becker 91). The purpose of this report is to point out that removing statistical dependence is not always a good thing.

In place of a general statistical independence, we wish to emphasize the importance of conditional statistical independence. As a result, and with the constraint that a prediction generating neuron will only be using linear (or log linear) computations to produce its predictions, we show that allowing statistical dependence to remain in the initial preprocessed input environment can be advantageous. Specifically, it can be advantageous when input subset selection is allowed at the prediction generating neuron. Furthermore, we propose - but do not prove - that there is a form of adaptive synaptogenesis which might perform appropriate subset selection.

3 The neural computations

We have previously proposed (Levy 89, Levy and Deliç 93,94) several maximum entropy formulations that are neurally plausible in the sense that they only require a linear computation and locally implementable, associative synaptic modification. Here we work with the simplest version of the maximum entropy generating neuron. We use binary inputs designated X_i for an individual input or X for the entire set of inputs or X_S for the subset of inputs that are used by a neuron to create a prediction of its binary state, $Z_j = 0$ or 1 . Using the locally available pairwise expectations and the locally available average activity of the postsynaptic cell, the maximum entropy form (P^*) of the probability that Z_j is active is given as (Levy 89, Levy, Colbert and Desmond 90).

$$P^*(Z_j = 1|X_s) = \frac{\prod_i P(X_i = x|Z_j = 1) * P(Z_j = 1)}{\prod_i P(X_i = x|Z_j = 1) * P(Z_j = 1) + \prod_i P(X_i = x|Z_j = 0) * P(Z_j = 0)} \quad (1)$$

This calculation is derived using Bayes theorem and the conditional independence forced by maximum entropy on the subset of input neurons (X_s) innervating output neuron j . That is, the output neuron

approximates $P(X_s|Z_j = 1)$ and $P(X_s|Z_j = 0)$ with $\prod_i (P(X_i = x|Z_j = 1))$ and $\prod_i (P(X_i = x|Z_j = 0))$, respectively.

We demonstrate in the following examples that, when output neuron j selects a subset of X that is not conditionally independent with respect to the firing of the output neuron, there is significant error in the neuron's approximation of $P^*(Z_j = 1|X)$. Indeed, we demonstrate that the error introduced by this violation of the independence assumption can be of such magnitude that the neuron generates more accurate predictions by using fewer, less statistically dependent input neurons.

4 Examples

For each input pattern X , there are three probabilities generated: 1) the true conditional probability $P^T(Z_j = 1|X)$, 2) the maximum entropy (ME) (inferred probability from the subset of inputs $P^*(Z_j = 1|X_s)$) and 3) the ME inferred probability based on the complete set of inputs X , $P^*(Z_j|X)$. The probability designated by P^T is the true probability and the probabilities designated by P^* are the ME inferred probabilities. For each input pattern the absolute value of the difference between the true probability and the ME inferred probabilities is determined (that is, $|P^T(Z_j = 1|X) - P^*(Z_j|X_s)|$ and $|P^T(Z_j = 1|X) - P^*(Z_j|X)|$). These differences are then averaged over the entire input environment. In the simulation with 50 output neurons, the reported performance measures are the mean probability differences averaged over the 50 output neurons.

4.1 A small, illustrative example

In this example there are two input neurons and 10 input patterns. There is one output neuron whose state will be predicted using the inputs. The connectivity between the input layer and output layer (in all the examples presented here) is feed-forward. The network is trained with the following input/output vectors (the format is $x_1, x_2 : z_1$): 1) 1,1:1, 2) 1,1:1, 3) 0,0:1, 4) 1,1:0, 5) 1,1:0, 6) 0,0:0, 7) 0,0:0, 8) 0,0:0, 9) 0,0:0, and 10) 0,0:0. The input vectors average 0.97 bits of representational information and 0.97 bits of statistical dependence.

In this simple example we compare the results obtained by the output neuron when it generates probabilities based on input x_1 ($P^*(Z_j = 1|X_1)$) versus the results obtained when the neuron samples both neurons of the input layer ($P^*(Z_j|X)$). The following table illustrates the differences between the true conditional probability and the two ME inferred probabilities for each of the 2 unique input patterns.

Input Pattern	$P^T(Z_j = 1 X)$	$P^*(Z_j = 1 X_s)$	$P^*(Z_j = 1 X)$
11	0.50	0.50	0.70
00	0.167	0.167	0.085

As can be seen in the above table, the calculation based on the single input ($P^*(Z_j = 1|X_1)$) is closer to the true probability ($P^T(Z_j = 1|X)$) than the calculation that is based on input from both input neurons ($P^*(Z_j = 1|X_1, X_2)$). Over the 10 input patterns, the mean absolute difference between the true probability and the ME inferred probability using input 1 is zero. However, the average absolute difference between the true probability and the ME inferred probability using inputs 1 and 2 is 0.129. The ME prediction is exact only when the conditioning occurs on the single input; that is, when the conditional statistical dependence is zero. With the incorporation of the second input the conditional statistical dependence increases to 1.0 bit and the ME neuron fails to generate accurate predictions. Clearly, subsets can be better than the full input. In the next example the superior subsets have less unconditional information.

4.2 A large, complex input

Now we present the results obtained from a more complex input. In this example there are 4 input neurons, 50 output neurons, and 100 input patterns. The input pattern space (X) contains 2.31 bits of representational information and 1.354 bits of statistical dependence. Because there are too many input/output mappings to list here, we present the summary statistics: $P(X_1 = 1|Z_j = 1)$ equals 0.400,

$P(X_1 = 1|Z_j = 0)$ equals 0.286, $P(X_2 = 1|Z_j = 1)$ equals 0.567, $P(X_2 = 1|Z_j = 0)$ equals 0.243, $P(X_3 = 1|Z_j = 1)$ equals 0.633, $P(X_3 = 1|Z_j = 0)$ equals 0.243, $P(X_4 = 1|Z_j = 1)$ equals 0.533, $P(X_4 = 1|Z_j = 0)$ equals 0.214, and $P(Z_j = 1)$ is 0.30.

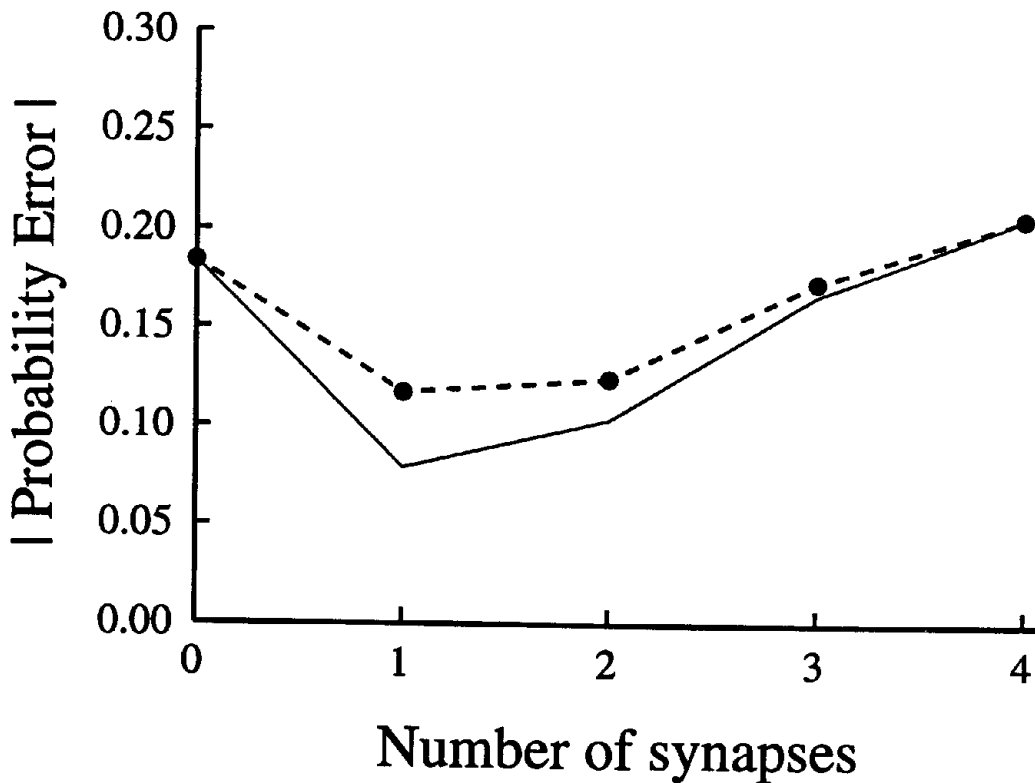


Figure 1: Prediction accuracy as a function of increased synaptic connections. There are four inputs neurons and 100 input patterns. Average error equals the $E[|P^T(Z_j = 1|X) - P^*(Z_j = 1|X_S)|]$ where averaging occurs over entire input and all appropriate subsets. Each of 50 output neurons makes the designated number of synapses with the input layer (there can be at most one synapse between an input and output neuron). The input connections to the output neurons are randomly determined. The filled circles represent the mean absolute value of the probability differences averaged over the 50 output neurons. The solid line represents the smallest probability difference obtained over the output neurons, i.e., the score of the input subset producing the best performance.

Figure 1 illustrates the average performance of the output neurons as a function of the number of connections with the input layer. Average error is smallest when each output layer neuron is innervated by one input neuron. The largest probability error is generated when the output neurons are innervated by the entire input layer. Indeed, there is *less error* when the output neuron receives *no innervation* from the input layer than when the output neuron is innervated by the entire input layer. Thus, inherent in these results are many examples where input subsets with less unconditional information outperform input subsets with more information.

As can be also be seen in Figure 1, there are output neurons that perform significantly better than average. These output neurons are innervated by subsets that contain a small fraction of the information of the input space, but which carry a large fraction of the conditional information needed for the ME prediction. It is this conditional information that confers upon an output neuron the ability to generate accurate predictions. The role of conditional dependence versus unconditional dependence is illustrated in the next example.

4.3 When and where statistical dependence is good

In this final example we will consider an input world of three dimensions that is being used to predict about two output neurons. The accompanying table lists the probabilities of all the input events and the probability of the conditional predictions. Note that the input space has a statistical dependence of nearly one bit. That is, based on the first two dimensions, the third dimension can be perfectly predicted. We can easily preprocess this space to lower statistical dependence (by discarding the third dimension), but this would be a mistake.

Input Vector (X_1, X_2, X_3)	Output Vector (Z_1, Z_2)	Probability
000	10	0.12
001	00	0.00
010	00	0.00
011	11	0.18
100	00	0.00
101	11	0.28
110	10	0.42
111	00	0.00

Based on this probability distribution, $E[X_1|Z_1 = 1] = 0.70$, $E[X_2|Z_1 = 1] = 0.60$ and $E[X_3|Z_2 = 1] = 1$. Inputs X_1 and X_2 are conditionally independent based on Z_1 and conditionally dependent based on Z_2 .

With regard to the first output neuron, the first two dimensions create perfect prediction and the third dimension just adds to the statistical dependency of this conditioned prediction and therefore degrades performance. However, for the second output neuron, perfect prediction is obtained by using the third dimension and addition of any of the other two dimensions degrades performance. Moreover, using either of the other dimensions alone results in poorer performance. Thus, the best performance is obtained by *not removing statistical dependence* from the input space. Instead, we allow some subset selection procedure at each prediction generating neuron to conditionally preprocess the inputs.

5 Concluding remarks

Perhaps the most basic theorem in information theory concerning a conditional probability considers the effect of increasing the number of conditioning variables. The uncertainty about the outcome in one particular variable can only decrease when conditioning on an additional variable. That is, $H(Z|X, Y) \leq H(Z|X)$. However, this theorem implicitly assumes that all the information in the conditioning variables is used. In neural networks, and more to the point here, in single neurons, there are computational constraints on how information can be used. Therefore, this theorem need not hold.

Removing all statistical dependence from an input space can be a mistake. If information processing in a succeeding layer of the network depends on a linear process as computed on a neuron by neuron basis, it is easily conceivable that different neurons will require different configurations of the input environment to optimally apply their restricted computation. That is, the input coding that might be optimal for one neuron may not be optimal for another neuron.

The process of subset selection is not commonly considered in the neural network literature. However, we have considered subset selection for unsupervised neural networks under the terminology of synaptogenesis and synaptic shedding (e.g., Adelsberger-Mangan and Levy 94, Levy and Desmond 86). Previously, using biologically inspired local mechanisms to control a random synaptogenesis, we have shown that synaptogenesis can produce networks that are tuned for the traditional preprocessing that lowers statistical dependence while maintaining as much information as possible (Adelsberger-Mangan and Levy 93,94). In the present context we consider the problem of conditional statistical dependence and conditional entropies. The adaptive synaptogenesis rules that we proposed previously will have to be modified. In particular, adaptive synaptogenesis must now be made conditional on each postsynaptic neuron. The condition which we propose is inactivity or low activity in the prediction compartment that is followed by high activity in the compartment being predicted on each neuron. This mismatch condition within each neuron would result in an increased receptivity to new innervation as appropriate

to performance. Sensibly, a recently active presynaptic input should have a higher probability of forming a new synapse with a postsynaptic structure than should a less recently active input.

We hypothesize that such conditional adaptive synaptogenesis will lead to improved performance in prediction networks that benefit from subset selection of their inputs.

6 Acknowledgements

This work supported by NIH MH48161, MH00622, RR07864 and EPRI RP8030-08 to WBL, and by the Department of Neurosurgery, Dr. John A. Jane, Chairman.

References

1. Adelsberger-Mangan, D.M. and Levy, W.B (1993) "Adaptive synaptogenesis constructs networks that maintain information and reduce statistical dependence", *Biol. Cybern.*, 70:81-87.
2. Adelsberger-Mangan, D.M. and Levy, W.B (1994) "The influence of limited presynaptic growth and synapse removal on adaptive synaptogenesis", *Biol. Cybern.*, 71:461-468.
3. Barlow, H. (1959) "Sensory mechanisms, the reduction of redundancy, and intelligence". In: *National Physical Laboratory Symposium Number 10, The Mechanization of Thought Processes*, London: Her Majesty's Stationery Office, 537-559.
4. Becker, S. (1991) "Unsupervised learning procedures for neural networks", *Int. J. Neural Systems*, 2:17-33.
5. Földiák, P. (1990) "Forming sparse representations by local anti-Hebbian learning", *Biol. Cybern.*, 64:165-170.
6. Levy, W.B (1985) "An information/computation theory of hippocampal function", *Society of Neuroscience Abstracts*, 11:493.
7. Levy, W.B (1989) "A computational approach to hippocampal function", In: *Computational Models of Learning in Simple Neural Systems*, (R.D. Hawkins and G.H Bower, Eds.), New York: Academic Press, 243-305.
8. Levy, W.B, Colbert, C. and Desmond, N.L (1990) "Elemental adaptive processes of neurons and synapses: A statistical/computational perspective", In: *Neuroscience and Connectionists Models*, (M.A Gluck and D.E. Rumelhart, Eds.), Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., 187-235.
9. Levy, W.B and Delic, H. (1993) "A generalized theory of maximum entropy prediction by neurons", *INNS World Congress on Neural Networks*, II:131-134.
10. Levy, W.B and Delic, H. (1994) "Maximum entropy aggregation of individual opinions", *IEEE Trans. Sys. Man and Cybern.*, 24:606-613.
11. Levy, W.B and Desmond, N.L (1986) "The rules of elemental synaptic plasticity", In: *Synaptic Modification, Neuron Selectivity and Nervous system Organization*, (W.B Levy, J. Anderson and S. Lehmkuhle, Eds.), Hillsdale NJ: Lawrence Erlbaum Associates Inc., 105-121.
12. Linsker, R. (1990) "Perceptual neural organization: Some approaches based on network models and information theory", *Ann. Rev. Neurosci.*, 13:257-281.

WCNN'95 – WASHINGTON, D.C.

WORLD CONGRESS ON NEURAL NETWORKS

**1995 International
Neural Network Society
Annual Meeting**

**Renaissance Hotel
Washington, D.C., USA
July 17 - 21, 1995**

Volume I

Sponsored by the International Neural Network Society

Copyright © 1995 by Lawrence Erlbaum Associates, Inc., and INNS Press, held jointly. All rights reserved. No part of this book may be reproduced in any form, by photostat, microform, retrieval system, or by any other means, without the prior written permission of the publisher.

Lawrence Erlbaum Associates, Inc., Publishers
10 Industrial Avenue
Mahwah, New Jersey 07430

ISBN 0-8058-2125-2

Books published by Lawrence Erlbaum Associates are printed on acid-free paper, and their bindings are chosen for strength and durability.

Printed in the United States of America