Commentary on: Andy Clark & Chris Thornton (1997). Trading spaces: Computation, representation, and the limits of uninformed learning. Behavioral and Brain Sciences, 20, 57-66.

Constraining Solution Space to Improve Generalization

John A. Bullinaria

Centre for Speech and Language, Department of Psychology,
Birkbeck College, Malet Street, London WC1E 7HX, UK

johnbull@ed.ac.uk

Abstract: I suggest that the difficulties inherent in discovering the hidden regularities in realistic (type-2) problems can often be solved by learning algorithms employing simple constraints (such as symmetry and the importance of local information) that are natural from an evolutionary point of view. Neither "heavy-duty nativism" nor "representational recoding" appear to offer totally appropriate descriptions of such natural learning processes.

I agree with the main conclusion drawn by Clark & Thornton that successful generalization in the case of realistic mappings often requires something more than the simplest statistical analysis. However, I would like to suggest that the case for representational recoding may be somewhat overstated, and that simple constraints on the solution space are often sufficient on their own to lead to good generalization.

Let us consider again the four bit parity problem cited by Clark & Thornton. One can explore the solution space in this case without making unnecessary assumptions concerning the properties of particular learning algorithms by performing a Monte Carlo search for solutions in weight space. The minimal network to solve this problem requires four hidden units (and hence 25 degrees of freedom) so we use that architecture. We choose sets of network weights at random (in the range -16 to +16) and check to see if they solve the four bit parity problem for 15 of the 16 training patterns in the sense that each output unit activation is to the correct side of 0.5. To find 20 solutions took 11.8 billion attempts. Each solution generalized incorrectly to the missing training pattern, which is what we would expect given that random hyper-planes in input space are likely to cut off the missing pattern with its closest neighbours which all produce the opposite output.

We have to ask why we consider one particular generalization to be better than the others. In the sense of Occam's razor, such as embodied in Bayesian model comparison (e.g. MacKay

1992), the best (or "most correct") generalization is the one provided by the simplest model (e.g. the one with the least "free" parameters). In fact, smaller networks are well known to provide superior generalization (e.g. Baum & Haussler 1989). In this respect the arguments of Clark & Thornton would have been more convincing if six or more bit parity were used, so that the mapping could be carried out with fewer free parameters (i.e. weights) than training patterns. Since avoiding local minima in minimal six (or more) bit parity networks is extremely difficult and since it is unlikely that real brains employ minimal networks we shall pass over this point.

One natural way to achieve model simplification is by constraining the search space, and one natural constraint might be the imposition of symmetry, i.e. start learning assuming maximal symmetry and only relax that assumption as each level of symmetry is found not to exist. This will automatically reduce the effective number of free parameters. For example, imposing a symmetry on the weights is sufficient to give good generalization for the four bit parity problem. Here we constrain the weight solutions to lie on the hyper-planes in weight space corresponding to weights that are symmetric with respect to the input units. This might be implemented in a learning network by constraining the weight changes to be the same for each input unit. This reduced the problem to 13 degrees of freedom and required only 16.3 million random attempts to find 20 solutions. The symmetry guarantees that all these solutions will generalize correctly. Such "weight sharing" is known to improve generalization more generally (e.g. Nowlan & Hinton 1992).

Another natural constraint we may impose is to assume that local information is more important than distant information until such an assumption is proven incorrect. We may view this to be at work in Elman's grammar acquisition network discussed by Clark & Thornton. Elman (1993) implemented these constraints by incremental learning schemes. In fact this is another poor example, since the network not only fails to generalize but also has insufficient processing power to even learn the raw training data (Elman 1993, p. 76). A more powerful recurrent network, or a network with appropriate input buffers or time delay lines, should not have this problem, but there is no reason to suppose that this would improve generalization as well. In time buffered networks we can constrain solutions to make maximal use of local information by having a smaller learning rates for weights corresponding to longer range dependencies. This approach has also, for examples, been shown to improve generalization in past tense acquisition models for which the inflection is usually, but not always, determined by the final phoneme of the stem and in models of reading aloud for which long range dependencies are relatively rare (Bullinaria 1994). Similar constraints may be implemented by weight decay and are also known to improve

generalization (e.g. Krogh & Hertz 1992).

Simple constraints on the weight space may not be sufficient to improve generalization for all type-2 problems, but the examples given above indicate that it does have a range of applicability. One might argue that such constraints are just a convenient way to implement the representational recodings of Clark & Thornton, but if that is the case we would seem to have a continuous spectrum of constraints and their type-1/type-2 distinction becomes rather fuzzy.

References

Baum, E.B. & Haussler, D. (1989) What size net gives valid generalization? *Neural Computation* 1: 151-160.

Bullinaria, J.A. (1994) Modelling reading, spelling and past tense learning with artificial neural networks. *Brain and Language*, in press.

Elman, J.L. (1993) Learning and development in neural networks: The importance of starting small. *Cognition* 48: 71-99.

Krogh, A. & Hertz, J.A. (1992) A simple weight decay can improve generalization. In J.E. Moody, S.J. Hanson & R.P. Lippman (Eds.) *Advances in Neural Information Processing Systems 4*, 950-957. Morgan Kauffmann.

MacKay, D.J.C. (1992) Bayesian interpolation. *Neural Computation* 4: 415-447.

Nowlan, S.N. & Hinton, G.E. (1992) Simplifying neural networks by soft weight sharing. *Neural Computation* 4: 473-493.