Reading the Generalizer's Mind

Chris Thornton and Andy Clark

May 21, 2003

Abstract In his new commentary, Damper re-emphasises his claim that parity is not a generalisation problem. But when proper account is taken of the arguments he puts forward, we find that the proposed conclusion is not the only one that can be drawn.

In adding the word 'still' to the title of his ongoing commentary, Damper re-emphasises his claim that 'parity is not a generalisation problem.' His view is that in our response we failed to accept or even properly address his argument. However, as we hinted in the first paragraph of R6, the interpretation of his claim is not a straightforward matter. The name 'parity' refers to the truth function whose rule is that the output is true if an odd number of the inputs are true. Parity functions may be of any arity but we are particularly familiar with the 2-place variant, also known as 'exclusive-or' (XOR).

Now, of course, a function is a function is a function. It is not, in itself, a 'problem'. And incontrovertibly, therefore, it cannot be a 'generalisation problem'. However the parity function (like any other function) is easily used as the *basis* for a generalisation problem. The procedure is straightforward: we take the complete mapping for a given parity function (e.g., all 16 input/output associations for the 4-bit parity function) and we present a *subset* of these cases to the generalisation mechanism, e.g., supervised learner. We then test the generalisation performance of the mechanism by examining its response on the unseen cases, or the 'validation' set as Damper calls it. This is the standard method for presenting a generalisation problem to a supervised learning mechanism. And we note that in both his commentaries, Damper describes the way he used the method to test the generalisation abilities of a backpropagation network. His account in his first commentary is particularly clear: he describes how he provided his backpropagation network with just 'the first three lines of the [xor] truth table' so as to see whether it could 'generalise on the 2-variable parity (XOR) problem.' Clearly, then, Damper is familiar with the procedure via which a parity mapping is used as the basis for the presentation of a generalisation problem. His intention thus cannot be to argue that parity cannot be used as the basis of a generalisation problem. What, then, are we to make of his declaration that 'parity is not a generalisation problem.'

A clue to his intentions comes in his second commentary. Here he suggests that generalisation 'refers to the fitting of a smooth function to an input-output mapping'. For Damper, this implies that 'parity is not a generalisation problem

because binary-to-binary mappings are inherently discontinuous.' The argument here is that since parity involves a particular operation — namely the 'fitting of a smooth function to an input-output mapping' — the setting up of a generalisation problem in which this operation cannot be applied, inevitably results in failure. In experimental terms, the procedure should be deemed a pointless and vacuous exercise and 'parity generalisation' classified as an oxymoron.

Several aspects of Damper's commentaries suggest that this is, indeed, his intended argument and that we should therefore treat 'parity [still] isn't a generalisation problem' as a claim about the impossibility of *solving* parity-generalisation problems. And yet, nagging doubts remain. Can Damper's view that generalisation involves the fitting of a 'smooth function' to an input-output mapping be taken seriously? A significant proportion of contemporary generalisation models are symbolic in nature and do not trade in *any* sort of numerical representation. Is it Damper's intention that these should be ruled out of court? His attitude to the performance of backpropagation on parity-generalisation also presents problems. We expect Damper to take the failure of *any* generalisation method on parity generalisation to corroborate his view that parity-generalisation cannot be regarded as a genuine problem. And yet in his conclusion, Damper suggests that 'C&T are wrong to read too much into [backpropagation's] failures' on the parity-generalisation task.

And what are we to make of his treatment of the generalisation method described by Berkeley? Damper notes that Berkeley's method provides a 'solution' to a parity generalisation task and we naturally expect Damper's position to be an emphatic rejection of Berkeley's claim that the method performs anything approximating 'genuine generalisation.' And yet, paradoxically, Damper's view is that Berkeley's method *satisfactorily* accomplishes the task it is set even though — as he puts it — the method may appear to 'mind-read by discovering just that solution which happens to be in the experimenter's mind.' This sounds suspiciously like a muted round of applause. And in fact it turns out that Damper's view is that 'the lesson of all this is that learning is not homogenous — different algorithms learn different things.'

Reading this we struggle with the contrast between the Damper who thinks parity generalisation is not a genuine learning problem and the Damper who considers that certain methods satisfactorily solve parity generalisation problems. But to get hung up on these apparent contradictions may be a mistake. Rather we should try to determine Damper's intended meaning by carefully reading between the lines.

In appearing to present inconsistent views with respect to generalisation's technicalities, Damper may be cunningly shaking out the thorny problem which lies at the subject's core, namely Hume's problem, or the 'problem of induction'. This is the observation that since inductive generalisations do not have (by definition) a logical derivation, they can never be regarded as entirely certain. Any inductively-acquired knowledge (e.g., scientific knowledge) is thus necessarily uncertain.

An interesting corollary is that, since generalisation *products* are always uncertain, we should arguably treat all generalisation *methods* as being of equal

status. And, indeed, this thesis has recently been given a mathematical foundation in the form of the No-Free-Lunch theorem of Wolpert [1; 2] and the **Conservation Law** of Schaffer [3].[1]

Damper's assertion that 'backpropagation has no special status' seems to confirm our suspicion that his underlying aim is not so much to demonstrate that parity isn't a generalisation problem but rather to demonstrate that the performance of particular learning methods on particular problems does not tell us very much. But if this *is* his intention then all parties can breathe a sigh of relief. There is nothing about our position which would cause us to do anything but wholeheartedly concur.

Recall that our paper used a probability argument to show that inductive generalisations may be justified either though type-1 (statistical) effects or through type-2 (relational) effects. We demonstrated that learning which depends exclusively on the exploitation of type-1 effects cannot deal with relational problems because such problems do not present exploitable, type-1 effects. (In both his commentaries, Damper takes time to illustrate what this means in the context of the parity mapping.) Following the presentation of the type-1/type-2 distinction we then introduced a case study involving the backpropagation method. This was intended merely to provide an illustrative example. Our suggestion was that the in-principle intractability that relational problems present to methods relying on type-1 effects 'may help to explain why backpropagation ... often fails to solve low-order parity problems when presented as generalisation problems.' In other words, we were speculating that backpropagation may be bad at parity generalisation as a result of depending too heavily on the exploitation of type-1 effects. We could easily have made the same remark about any other method adopting the same strategy.

We believe that wires may have become crossed over this reference to backpropagation partly because of the unconventional way in which the type-1/type-2 distinction was formulated. But we hope that the ensuing commentary has clarified the fact that the distinction introduced is uncontentious and that it is in fact one which has been expressed in a wide variety of ways over a large number of years. As it turns out, it can even be formulated in terms of Damper's own 'smooth function' concept.

To formulate the distinction in these terms we first need to visualise the generalisation process and associated input/output mapping in pictorial terms. We view the input/output mapping in terms of an input space whose datapoints correspond to individual inputs. The label attached to each datapoint is then the output associated with the input; a method solves a generalisation problem by successfully using a sample of labelled datapoints to predict the labels of inputs not included in the sample.

In a 'smooth' input/output mapping — the type that Damper believes presents a genuine generalisation problem — the labeling of datapoints varies

[1] These make use of the fact that when we average the performance of a generalisation method over all possible scenarios, we inevitably find that each particular generalisation is correct just as often as it is incorrect. The counter-intuitive effect is that all generalisation methods have an average performance which is identical to that achieved by random guessing.

smoothly across the space. Datapoints with the same label cluster together and there is a gradual transition between different labels as we move across the space. Generalisation methods must associate groups of inputs with specific outputs and in the 'smooth mapping' context, this is easily accomplished. Because of the way inputs with the same output tend to cluster together, the process of separating them can be straightforwardly accomplished by introducing simple bounding constructs (planes, spheres etc.) into the space. If the input/output mapping is not smooth and datapoints with the same label do not cluster together then separation of groups of inputs requires the introduction of more complex bounding constructs.

These observations might lead us to introduce a 'new' fundamental distinction between smooth and non-smooth input/output mappings and to point out that only smooth input/output mappings allow for learning/generalisation processes based on the introduction of simple bounding constructs. Particular learning methods could then be divided up according to whether they utilise simple or complex bounding constructs. Key members of the 'simple' camp would turn out to be the Perceptron method [4] which introduces a single, planar boundary, ID3 [5] which adds an arbitrary number of axis-aligned, extreme boundaries, Backpropagation [6] which manipulates a fixed number of linear boundaries, LVQ [7] which manipulates a fixed number of spherical boundaries and the k-nearest-neighbours method [8] which utilises the implicit planar boundaries between datapoints. Key members of the 'complex' camp would turn out to be methods such as AQ15 [9], CIGOL [10], and FOIL [11] which utilise background knowledge of one form or another for the purposes of forming complex separations among classes of inputs.

But in working through this argument we would, of course, simply be rehashing the type-1/type-2 distinction introduced in our paper. Input/output mappings are 'smooth' just in case datapoints with the same label cluster together. This occurs if absolute input values (i.e., datapoint coordinates) are significant for the prediction of output. If input values are not significant, then there is no reason to expect datapoints with the same label to occupy the same part of the input space; there is no clustering and no smoothness. Damper suggests that expecting a method to generalise in this context — when absolute values are insignificant for the prediction of output (as they are in the parity mapping) — amounts to 'expecting the [method] to be a mind-reader.' But although absolute values may not be significant, the relationship(s) among them may be. Generalisation then does not require mind-reading but merely an accurate identification of the relationship underlying the mapping. This brings us more or less back to the original point around which our paper was based. Mappings in which the underlying input/output rule is **relational** or **type-2** cannot be generalised by methods which utilise simpler bounding constructs and thus implicitly assume a 'smooth' (type-1) mapping.

As we suggested in our initial response, Damper is quite right to observe that absolute input values (datapoint coordinates) cannot be used as a basis for predicting outputs in parity mappings. But rather than demonstrating that parity cannot be treated as a generalisation problem, it actually demonstrates that

parity forms the basis for a particular type of generalisation problem, namely a **relational** problem in which the successful prediction of outputs involves the discovery of the relational rule underlying the mapping. Damper's correction of his title should thus not involve the insertion of the word 'still' but rather the insertion of 'type-1', thus producing the correct conclusion 'Parity is not a *type-1* generalisation problem.'

References

[1] Wolpert, D. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation*, *8*, No. 7.

[2] Wolpert, D. (1996). The existence of a priori distinctions between learning algorithms. *Neural Computation*, *8*, No. 7.

[3] Ross, T., Noviskey, M., Axtell, M., Gadd, D. and Goldman, J. (1994). Pattern theoretic feature extraction and constructive induction. *Proceedings of ML-COLT'94.*

[4] Minsky, M. and Papert, S. (1988). *Perceptrons: An Introduction to Computational Geometry* (expanded edn). Cambridge, Mass.: MIT Press.

[5] Quinlan, J. (1983). Learning efficient classification procedures and their application to chess end games. In R. Michalski, J. Carbonell and T. Mitchell (Eds.), *Machine Learning: An Artificial Intelligence Approach.* Palo Alto: Tioga.

[6] Rumelhart, D., Hinton, G. and Williams, R. (1986). Learning representations by back-propagating errors. *Nature*, *323* (pp. 533-6).

[7] Kohonen, T., Barna, G. and Chrisley, R. (1990). Statistical pattern recognition with neural networks: benchmarking studies. In J.A. Anderson, A. Pellionisz and E. Rosenfeld (Eds.), *Neuocomputing 2* (pp. 516-523). The MIT Press.

[8] Duda, R. and Hart, P. (1973). *Pattern Classification and Scene Analysis.* New York: Wiley.

[9] Michalski, R., Mozetic, I., Hong, J. and Lavrac, N. (1986). The multipurpose incremental learning system AQ15 and its testing application to three medical domains. *Proceedings of the Fifth National Conference on Artificial Intelligence* (pp. 1041-1045). Philadelphia, PA:: Morgan Kaufmann.

[10] Muggleton, S. and Buntine, W. (1988). Machine invention of first order predicates by inverting resolution. In J. Laird (Ed.), *Proceedings of the Fifth International Conference on Machine Learning* (pp. 339-352). Morgan Kaufmann.

[11] Quinlan, J. (1990). Learning logical definitions from relations. *Machine Learning*, *5* (pp. 239-266).