Back-propagation under non-optimal supervision: Developmental dyslexia in a computational model of reading.

John E. McEneaney
Indiana University South Bend

Back-propagation has been successfully applied in a variety of engineering and computer science settings and in an equally wide array of cognitive modeling applications. The purpose of this paper is to explore the consequences of non-optimal supervision for learning in back-propagation networks where non-optimal supervision refers to noise and/or error introduced as an output-level error term that is back-propagated through the network in adjusting connection weights. It will be argued that, from a cognitive modeling perspective, non-optimal supervision provides a more plausible model of the fallible supervision generally available to human learners (especially in circumstances of self-supervision) and that this concept provides a more adequate framework for modeling aspects of developmental reading disabilities, a topic of growing interest among researchers applying back-propagation in cognitive theory.

1. Introduction

Additive noise, usually attenuated across learning cycles, is commonly employed in training neural networks in an effort to enhance learning. Among other benefits, the use of attenuated noise (often referred to as simulated annealing) is a process that tends to result in networks that are more likely to stabilize in global energy minima, have an enhanced capacity to generalize, and are quicker to recover from damage in retraining.

One common strategy for introducing noise to a system is by adding it to training stimuli prior to presentation (Holmström & Koistinen, 1992). Another approach is to introduce a noise factor directly into the equations used in adjusting synaptic weights (Burton & Mpitsos, 1992). From a cognitive modeling perspective, noise introduced at the input level can be interpreted either as external noise inherent in the input patterns to be learned or as internal noise resulting from lower-level coding systems (i.e. sensory systems). Noise introduced in the equations responsible for synaptic adjustments, on the other hand, is more plausibly considered internal to the learning system, perhaps reflecting computational limits of implementation-level entities.

The research reported in this paper will explore a third way noise can be introduced to a learning system that may help in conceptualizing developmental dyslexias - as noise added to training patterns that are employed in determining error terms at the output level. As in the case of input level noise, this approach can be interpreted in at least two ways. One way to view noise introduced at the output level, like its analog on the input level, is as error resulting from lower-level systems (i.e. garbled feedback). A second way of viewing this kind of noise is that the feedback or supervision provided to the learning system is simply in error, that the "teacher" responsible for providing feedback is wrong. This way of viewing output level noise, as non-optimal supervision, will be the focus of simulation trials.

The mapping of this interpretation of non-optimal supervision onto classroom learning is founded on the observation that beginning readers require many more exposures to new words than can normally be provided in direct instruction by teachers. A consequence of this is that beginning readers are frequently obliged to rely on self-generated feedback (Jorm & Share, 1983), particularly in the earliest stages of reading acquisition. The purpose of the simulation trials in this study is
to explore the consequences of this kind of error-prone feedback on long-term learning.

Although a number of recent investigators (e.g. Hinton & Shallice, 1991; Patterson, Seidenberg, & McClelland, 1989; Seidenberg, 1992) have modeled reading disabilities in neural network systems, this prior work has focused on acquired disabilities simulated through training and subsequent destruction of hidden nodes or synaptic weights. While this kind of modeling may help explain disabilities which can be traced to brain injury or dysfunction, this approach does little or nothing to help us understand the origins of developmental dyslexias which cannot be localized yet which account for a much larger percentage of children with severe learning difficulties in school settings. The purpose of this paper is to explore the explanatory power of a network model in accounting for specifically developmental dyslexias.

The work described in this paper is framed within a broader research program that integrates traditional rule-based cognition with connectionist processing through a loosely coupled system of modules like that pictured in Figure 1. According to this model, the interaction of rule-based and connectionist modules can occur either through a common workspace like that postulated in blackboard models or through self-supervised internal feedback.

A central premise of the model is that both symbolic and connectionist processes are available at all stages of word reading and that word reading ultimately involves a balance between (at least) two different processing systems that employ radically different forms of orthographic knowledge. Although strategic decoding of words through grapheme-phoneme correspondence (GPC) rules is slow, in the earliest stages of reading acquisition GPC rules provide a more reliable approach than connectionist processing under conditions of limited experience. As the experience of a reader grows, however, the reliability of connectionist processing increases and the disadvantages of strategic rule-based processing become more important. These assumptions suggest that observed developmental trends in reading acquisition may be accounted for by a shift in dominance from primarily strategic rule-based cognition in early word reading, through a phase of mixed rule-based/connectionist processing, and finally into a primarily connectionist (automatized) phase of mature word reading. These assumptions also suggest a possible model for developmental dyslexia that may arise as a consequence of persistent non-optimal supervision within a learning system.

3. Simulation Trials
Three different learning tasks were developed for the purpose of this study. One task involved auto-associative learning of 3-bit input-output pairs (i.e. 010-010, 111-111, etc.). Experimental trials consisted of training groups of networks across four levels of supervision error (0.0, 0.10, 0.30, 0.50) which specified the probability that any given digit in the training pattern was in error. One-way ANOVAs with follow-up Bonferroni tests were carried out with average error/pattern as the dependent variable at the conclusion of 100 cycles of training in order to determine whether levels differed significantly from the learning that occurs under conditions of...
optimal supervision (no error).

A second more complex learning task involved training groups of networks across 4 levels of persistent pattern error on a 512-item data set that simulated learning a reading vocabulary. Binary input values represented 14 visual feature elements (Rumelhart & Siple, 1974) across three letter positions for a total of 42 input units. Binary output nodes represented 12 phonemic feature elements across three letter positions for a total of 36 output units. A total of 256 CVC "words" were included in the training data set. Of those 256 words, 128 were high frequency (X3) types for a total of 512 training exposures/cycle. Hidden layers consisted of 30 units resulting in a 42:30:36 fully-connected architecture.

The third learning task was identical to the second but input/output training pairs were shuffled so that formerly regular sound/symbol correspondences of the original training data set were lost. The resulting training data set was an essentially arbitrary association of letter and phoneme triples. Network architectures in this set of trials were exactly as in the second task.

All networks were based on C++ network and layer classes described by Rao and Rao (1993) that were adapted to incorporate a pattern error variable. In addition, command-line parameters were added so that multiple trials across variables of interest could be carried out as batch routines. Learning rates were adjusted between task 1 (0.9) and tasks 2 and 3 (0.001) so that development in tasks 2 and 3 could be followed since the size of the data set in tasks 2 and 3 tended to obscure early stages of development within the first cycle of training. Variable bias terms, momentum and other numerical techniques designed to enhance learning were not employed in training.

3. Results

Task 1 learning was carried out under two pattern error conditions. One condition employed attenuated pattern error, corresponding to simulated annealing where noise is gradually reduced across training cycles. A second condition employed persistent pattern error which was not reduced across cycles. Four levels of pattern error were simulated. Data depicted in Figures 2 and 3, report the mean average error/pattern for each cycle for all networks within pattern error levels.

As indicated in Figure 2, long-term learning under attenuated pattern error is indistinguishable from learning under no error. No significant (p<.05) differences between levels of pattern error were revealed in one-way ANOVAs at cycle 100. Although pattern error tends to induce occasional error peaks proportional to the current level of error, these peaks appear to be a result of the attenuation process since peaks tend to correspond with the error attenuation steps that occurred at cycles 10, 30, 50, and 70.

![Figure 2](image)

**Figure 2.** Task 1 auto-associative learning under 4 levels of attenuated pattern error.

When pattern error was persistent, however, long-term effects on learning were evident (See Figure 3). An ANOVA employing mean error at cycle 100 and level of persistent error as dependent and
independent variables respectively revealed statistically significant group differences, $F(3,76) = 4.4$, $p < .01$) with follow-up Bonferroni tests indicating significant pair-wise differences between groups with error levels of 0.0 and 0.5.

Figure 3. Task 1 auto-associative learning under 4 levels of persistent pattern error.

Figure 4 depicts results of task 2 word learning under 4 levels of persistent pattern error. A one-way ANOVA with error at cycle 20 as the dependent variable revealed significant group differences, $F(3,60) = 389.9$, $p < .0001$. Follow-up Bonferroni analyses indicated pair-wise differences between all pairs of groups except those with error levels of .3 and .5.

Figure 4. Task 2 word learning under 4 levels of persistent pattern error.

Results of task 3 word learning did not differ significantly from the outcomes observed in Task 2 suggesting that the increased demand of learning the randomized training data set was still well within the capacity of the network design.

4. Conclusions

Non-optimal supervision appears to influence learning outcomes in a manner similar to input-level noise although its relatively coarse "bit-wise" grain results in less smooth learning curves than comparable levels of noise. Although attenuated pattern error does not appear distinguishable from optimal supervision (i.e. no error), persistent pattern error has measurable consequences for long-term learning which gradually become apparent over extended periods of training and appear to be related to the complexity of the learning task. Measures of learning outcomes on the simpler auto-associative learning task (Task 1) were less sensitive to influence by error, presumably as a consequence of the greater within-group variation reflected in the spiky learning curves. The more complex learning tasks (Tasks 2 and 3) were more likely to evince measurably different learning outcomes.

While the present investigation does not answer any specific empirical questions concerning the etiology of developmental dyslexias it does appear to offer a plausible, readily interpretable model that is consistent with recent empirical findings (Nicholson & Fawcett, 1989) suggesting that the concept of automaticity may play a central role in the learning difficulties characteristic of dyslexics. The more general framework upon which the present investigation is based also finds support in models of skill acquisition which are increasingly making use of encapsulated modules (Stanovich, 1990) and have frequently distinguished between strategic and automatic processing (Schneider & Fisk, 1983; Schiffren & Dumais, 1981; Sincoff & Sternberg, 1989).
Finally, it may be that the concept of non-optimal supervision has utility beyond simulating aspects of human cognition. Researchers with more traditional AI orientations have been interested in the capacity of hybrid systems for some time and in such systems the interaction of potentially fallible modules is an issue of central importance. It is also interesting to note that in one recent model (Shavlik & Towell, 1990) purely functional considerations led investigators to a developmental framework within which strategic rule-based processing precedes and lays the ground work for subsequent connectionist processing. It also seems likely that genuinely autonomous learning will require some capacity for self-supervision which, in an imperfect world, will probably be subject to error.

References


Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 111-140). Hillsdale, NJ: Lawrence Erlbaum.
