Chapter 1

Two-layer networks, non-linear separation, and human learning

R. Harald Baayen and Peter Hendrix
University of Tübingen

Ever since the criticism of the perceptron by Minsky & Papert (1969), two-layer networks have been regarded as far too restricted for classification tasks requiring more than the simplest linear separation. We discuss an example of a classification task that in $\mathbb{R} \times \mathbb{R}$ is not only not linearly separable, but also not non-linearly separable. Yet, this classification task can be carried out with error-free performance. To do so, it is mandatory to step outside the box of $\mathbb{R} \times \mathbb{R}$, and we discuss how several state-of-the-art methods from machine learning achieve this. We also show that a two-layer network that makes use of the learning rule of Rescorla and Wagner (1972) can solve this classification task, with different degrees of success (up to 100% accuracy) depending on the representations chosen for the input units. The excellent classificatory performance of our two-layer network helps explain why wide learning with two-layer networks with thousand and even tens of thousands of input and output units is so succesful in predicting aspects of human implicit learning, including the consequences of trial-by-trial learning for response latencies in the visual lexical decision task.

1 Introduction

Several computational modeling studies suggest that two-layer networks with connection weights estimated with the learning rule of Rescorla & Wagner (1972) capture non-trivial aspects of lexical processing. In what follows, these networks will be referred to as 'wide learning' networks, as they typically comprise just two layers with, however, many tens of thousands of units.

Baayen et al. (2011) observed that the activations of lexical output units (conceptualized as pointers to semantic vectors in Milin et al. (2016)) in wide learning networks with sublexical input units (e.g., letter bigrams) closely mirrored reaction times in the visual lexical decision task. Regression models, one fitted to the

R. Harald Baayen & Peter Hendrix. 2016. Two-layer networks, non-linear separation, and human learning. In Martijn Wieling, Gosse Bouma & Gertjan van Noord (eds.), *From semantics to dialectometry*, 1–11. Groningen: University of Groningen.

reaction times, and one fitted to the reciprocally transformed activations, produced very similar results. The same predictors reached significance, with very similar relative effect sizes. Even though the network did not have any form units for morphemes or words, it correctly predicted the facilitatory effects of constituent frequency and word frequency typically observed for English. When the same kind of network was trained on Vietnamese, it correctly predicted the inhibitory effect of constituent frequency that surprisingly characterizes compound reading in this language (Pham & Baayen 2015). These results show that wide networks with sublexical input units capture frequency effects that, depending on low-level orthographic distributional structure, can work out in very different ways, facilitatory in English but inhibitory in Vietnamese. Wide learning networks have also been found to provide superior prediction for the brain's electrophysiological response to linguistic stimuli (Hendrix, Bolger & Baayen 2016) and the details of eye-movements during reading (Hendrix 2015).

In classical accounts of lexical processing, the presence of a frequency effect is treated as a litmus test for the existence of form representations. For instance, a frequency effect for complex words is taken as proof of the existence in the mind of form representations for complex words. Typically, such representations are associated with resting activation levels that are assumed to depend on frequency of use, and that are assumed to underlie the frequency effects observed in tasks tapping into lexical processing. However, theories that posit such form units have to explain how such units are accessed. This question seldom is reflected on, probably because we are so familiar with being able to look up words in a dictionary, or to search for patterns in files, that we take for granted that accessing units is trivial. However, models such as the interactive activation model (McClelland & Rumelhart 1981) were developed precisely because human look-up has all kinds of properties that are foreign to look-up with the algorithms implemented on our computers. Wide learning networks offer an alternative algorithm that, like the interactive activation model, targets an algorithmic approximation of human look-up. Importantly, frequency effects (and also similarity effects) come for free with wide learning, and arise as a consequence of continuous error-driven optimization of lexical discrimination. There no longer is a need for positing counters in the head such as resting activation levels.

Current research is revealing that a range of quantitative measures derived from the connection weights of wide learning networks also generate precise predictions about trial-to-trial learning in the visual lexical decision task. Figure 1 presents three measures of goodness of fit, the AIC, the ML score, and the (adjusted) R-squared, for generalized additive models fitted to the data of a partici-

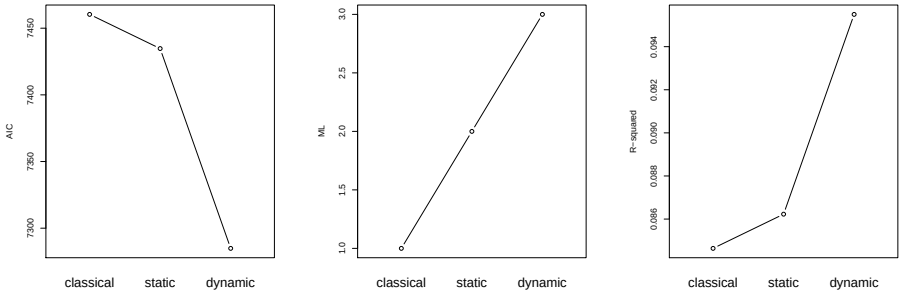

Figure 1: Three measures of goodness for the GAMs fitted to lexical decision reaction times.

pant contributing responses to the British Lexicon Project (Keuleers et al. 2012). Goodness of fit is worst for a classical model with as predictors the conventional measures of frequency of occurrence and neighborhood density. Model predictions improve when these classical measures are replaced by measures derived from the 15,106 $\times$ 30,117 weight matrix of a network trained on the British National Corpus. Performance is best when this network is allowed to continue learning as it is presented with words and nonwords, made available in exactly the same order as in the lexical decision experiment.

These results indicate that wide networks with the Rescorla-Wagner learning rule provide a useful computational window on human lexical processing, complementing the strong support for this learning rule in the literature on animal learning (Siegel & Allan 1996) and more recently also in computational evolutionary biology (Trimmer et al. 2012).

However, in the light of the criticism by Minsky & Papert (1969) of simple two-layer perceptrons as being incapable of approximating a wide range of useful functions, the excellent performance of wide learning networks is surprising. Would performance have been better with deep learning, or with Bayesian updating? Is a two-layer network, however wide, actually too simple to be taken seriously for the computational modeling of lexical processing?

In what follows, this issue is addressed by investigating a simple but non-trivial classification problem and comparing the performance of wide learning with three state-of-the-art classifiers: support vector machines, deep learning, and gradient boosting machines.

2 A non-linearly separable classification problem

The left panel of Figure 2 presents a classification problem that is not only not linearly separable, but also not non-linearly separable. In a grid of 50×50 pixels, 260 (highlighted in black) belong to class *A* and the remaining 2240 (in gray) belong to class *B*. In the $\mathbb{R} \times \mathbb{R}$ input space, this classification problem cannot be solved by means of linear or non-linear boundary functions. There is no straight line that separates the grey dots from the black dots, nor is there a sensible curve that would achieve separation.

Three machine learning techniques were applied to this classification task. Deep learning, using the `h2o` package (Fu et al. 2015), which provides an adaptive learning rate per neuron as well as regularization through shrinkage and dropout, with 1 layer of 100 hidden units, reached an accuracy of 99.4%. A gradient boosting machine (fit with 20 trees with a maximum tree depth of 20, using `xgboost` package (Chen, He & Benesty 2015)) provided perfect classification with only minor deterioration under 10-fold cross-validation. A support vector machine (using `svm` in the `e1071` package (Meyer et al. 2015), with a second-order polynomial as a kernel) performed quite well on the full data, but performance dropped below that of the other two methods under 10-fold cross-validation (cf. Table 1). The success of the support vector machine indicates that there exists a transformation of the $\mathbb{R} \times \mathbb{R}$ input space in which the two classes of data points are to a considerable extent linearly separable.

A very different transformation of the input space is achieved by moving from coordinates in $\mathbb{R} \times \mathbb{R}$ to one-hot encoding for rows and columns, resulting in two sets of 50 units representing row and column identifiers. Moving to this binary 100-dimensional representation (henceforth $\mathbb{B}^{100}$) allows all three abovementioned machine learning techniques to achieve perfect classification on the full data set, and to retain a high accuracy under 10-fold cross-validation (cf. Table 1).

A wide learning network with as cues the row and column identifiers performs, with a single pass through the data, with 96.0% accuracy (F-score 0.81, precision and recall both 0.81), with the expected decrease in performance under 10-fold cross-validation (accuracy: 95.0%; F-score 0.71, precision 0.88, recall 0.61).[1] Although clearly lagging behind the gradient boosting machine and the deep

[1] As wide learning does not make use of nonlinear activation functions at the output layer to obtain a firing versus not firing response, evaluation of model performance proceeded by collecting the activations for all 2500 pixels, and setting a threshold such that the k pixels with the highest support for class *A*, where k is the cardinality of *A*, are assigned to class *A*. The same threshold was used under cross-validation.

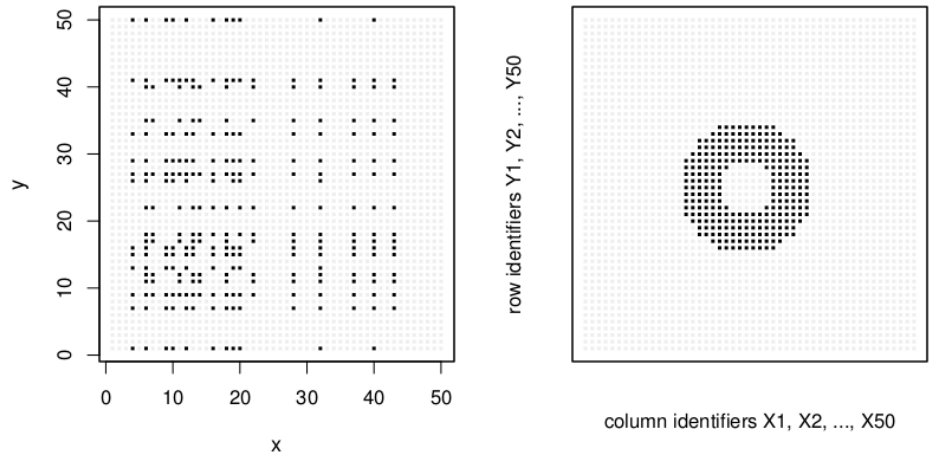

Figure 2: A non-linear classification problem. Left panel: points in a Cartesian grid ($x = 1, 2, \ldots, 50; y = 1, 2, \ldots, 50$). Right panel: the same points in a 100-dimensional space using one-hot encoding, with re-arranged rows and columns.

learning network, it is not the case that wide learning is a total failure — to the contrary, it gets quite far, performing better under 10-fold cross-validation than the support vector machine.

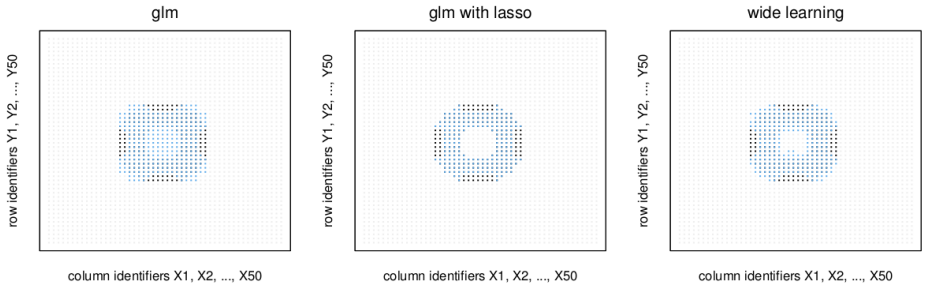

Figure 3: Classification performance for the generalized linear model (left), the generalized linear model with lasso (center), and wide learning (right) predicting class from row and column identifiers.

To appreciate better what wide learning achieves, first note that the move from $\mathbb{R} \times \mathbb{R}$ to $\mathbb{B}^{100}$ renders the classification problem invariant to exchanges of pairs

Table 1: Accuracy for four algorithms, applied to the complete data set and with 10-fold cross-validation, with and without re-ordering of rows and columns. deep: deep learning; gbm: gradient boosting machine; svm: support vector machine; wide: 2-layer wide learning network.

Accuracy		Method	Input Space
complete	cv-10		
0.994	0.986	deep	$\mathbb{R}^2$
1.000	0.994	gbm	$\mathbb{R}^2$
0.982	0.896	svm	$\mathbb{R}^2$
1.000	0.994	deep	$\mathbb{B}^{100}$
1.000	0.994	gbm	$\mathbb{B}^{100}$
1.000	0.949	svm	$\mathbb{B}^{100}$
0.960	0.950	wide	$\mathbb{B}^{100}$
1.000	0.989	wide	hub features

of rows, and to exchanges of pairs of columns. One re-arrangement of rows and columns results in a configuration with all points of class *A* arranged in a circular band, as shown in the right panel of Figure 2. This rearrangement, possible thanks to 'domain knowledge', shows that there is considerably more structure in the data than is apparent to the eye in the scatter in the left panel of Figure 2.

Now consider Figure 3. The left panel presents the predictions (in blue) of a generalized linear model (left), a generalized linear model with lasso correction (center), and a wide learning network (right). Each model was asked to predict the class of a data point from its row and column identifier. A logistic generalized *linear* model correctly detected that the points belonging to class *A* are located within a *circle*, but failed to exclude the points in its center, and lacked precision at the four outer edges of the circular band. Importantly, this linear model achieves considerable separation of the two classes in $\mathbb{B}^{100}$ that, if transformed back into $\mathbb{R} \times \mathbb{R}$, would classify as non-linear.

Improved classification accuracy can be obtained by shrinking the β coefficients of the GLM through lasso (ℓ1-norm) regularization (using `glmnet`, Friedman, Hastie & Tibshirani (2010), run with *maxit* = 300). The center panel shows that all points in the inner disk are now correctly assigned to class *B*. Yet, the model remains

somewhat imprecise at the edges.
The third panel of Figure 3 illustrates the performance of wide learning, which succeeds in correctly assigning most points in the inner circle to class *B*, while reducing slightly the imprecision at the edges that characterizes the `glm` with lasso regularization. Again, we see that a technique that is known to be restricted to linear separation in $\mathbb{R} \times \mathbb{R}$ achieves good separation in $\mathbb{B}^{100}$.
Of special interest is that this classification performance is achieved without knowledge of the topology of the circular band. All that is available to the models is, for each data point, the identifiers for its row and column. In other words, we can re-arrange the rows and colums back into the scatter of the left panel of Figure 2, and a majority of data points would still be correctly classified. This shows that the representation of the problem in $\mathbb{R} \times \mathbb{R}$ is a highly specific one that is restricted to a unique configuration of data points, whereas the representation in $\mathbb{B}^{100}$ covers the full set of $50! \times 50!$ permutations of rows and columns. The classification accuracy of the `glm` and of the wide learning network is exactly the same for all these alternative configurations.
However, the accuracy of wide learning can be improved considerably by making use of the fact that the re-arrangements share underlyingly the topology of the circular band. This topology makes it possible to do error-free classification with just four features. Let a data point be a hub if all of its eight surrounding data points belong to class *A*, and let a data point be a hub neighbor if at least one of the eight surrounding data points is a hub. We now define four features, IS A HUB, IS NOT A HUB, IS A HUB NEIGHBOR, and IS NOT A HUB NEIGHBOR. When each data point is characterized by the values of these four features, a wide learning network yields error-free classification performance with a single pass through the data. Under leave-one-out cross-validation performance remains error-free. Ten-fold cross-validation with hub features requires special care, as missing data make it impossible to maintain the criterion that a hub should have exactly 8 neighbors. When the neighbor count for a hub is relaxed to 7 during training and to 4 during testing, accuracy remains at 99% (F-score 0.95, precision 0.94, recall 0.96).
Figure 4 presents two further classification tasks for which wide learning with hub features performs with a very high accuracy. The pattern in the left panel was explicitly characterized by Minsky & Papert (1969) as impossible for perceptrons to classify, which is correct when the problem is formulated in $\mathbb{R} \times \mathbb{R}$, but not necessarily true when the problem is reformulated in other spaces. A wide learning network with hub features solves this classification problem in its stride, with error free performance also under leave-one-out cross-validation, but just

the representation in $\mathbb{B}^{100}$ already allows for a classification accuracy no less than 96.9% (F-score 0.88, precision 0.89, recall 0.87).
Interesting is also the 'open cross' task in the right panel of Figure 4. We failed to obtain sensible classification performance for $\mathbb{R} \times \mathbb{R}$ and for $\mathbb{B}^{100}$ under cross-validation with gradient boosting machines and support vector machines (all points assigned to the 'baseline' class *B*). Deep learning on $\mathbb{R} \times \mathbb{R}$ with a two layers of hidden layer, the first with 100 units and the second with four units, performed much better (accuracy 94.2%, 93.4% under 10-fold cross-validation), but upon inspection systematically assigned all class *B* data points within the open squares that build the cross to class *A* under cross-validation. Deep learning on $\mathbb{B}^{100}$ was a total failure (F-score = 0.28 under 10-fold cross-validation). Wide learning in $\mathbb{B}^{100}$ failed miserably as well (F-score 0.19), but wide learning with hub features was highly effective, with accuracy above 99% both for the full data set, as well as under leave-one-out cross-validation.

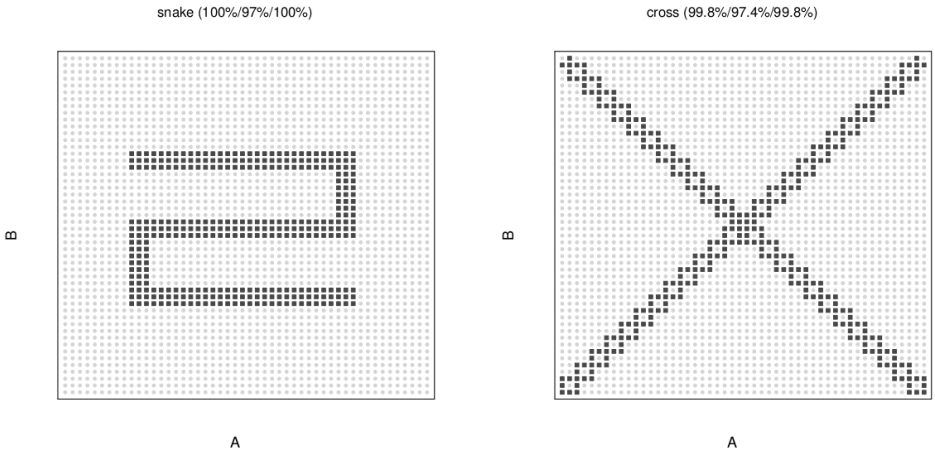

Figure 4: Two further non-linear classification set-ups. Accuracies in parentheses for wide learning with hub features for the full data set, for 10-fold cross-validation, and for leave-one-out cross-validation.

3 Discussion

To solve classification problems in $\mathbb{R} \times \mathbb{R}$ that are not linearly separable and also not non-linearly separable, it is crucial to step outside the box. Gradient

boosting machines achieve this by sidestepping the problem of finding a boundary function, and let classification trees do the best they can with local splits, while letting later trees deal with the classification errors of earlier trees. Support vector machines step outside the box by projecting the data points into a higher-dimensional space in which the classes are linearly separable. With a hundred hidden units, deep learning also finds a solution, thanks to regularization through shrinkage and dropout. (A three-layer network with 100 hidden units trained simply with backpropagation fails to assign any datapoint to the *A* class.) These 100 hidden units constitute a new space that re-represents the original $\mathbb{R} \times \mathbb{R}$ space in such a way that the last two layers of the three-layer network can achieve excellent (linear) separation of the two classes.

Moving from a representation in $\mathbb{R} \times \mathbb{R}$ to a representation with row and column identifiers is yet another way of stepping outside the box. For wide learning with the Rescorla-Wagner rule, this re-representation is a necessary step because input units are restricted to discrete feature detectors that are either on or off. Given this re-representation, wide learning can achieve considerable separation of data points that are not even non-linearly separable, and in this mirrors the performance of a logistic generalized linear model. But deep learning and the support vector machine also also thrive with this re-representation, reaching 100% accuracy on the full data and improved cross-validation scores. Because this re-representation is invariant to order, re-arranging row and column identifiers allows the underlying topology to emerge, which in turn makes an even simpler re-representation with hub features possible.

The present results clarify that it does not make much sense to suppose that a non-trivial wide learning network (such as the abovementioned network with 15,106 input units and 30,117 output units) is handicapped by being limited to 'linear separation'. This handicap holds for $\mathbb{R}^n$, but by re-representing the classification problem in some higher factorial space $\mathbb{B}^{n+m}$, $m \gg n$, a wide learning network can achieve separation that, albeit linear in $\mathbb{B}^{n+m}$, would count as non-linear when projected back into $\mathbb{R}^n$. It follows that what a two-layer wide learning network can or cannot achieve depends crucially on the input representations. Deep learning networks can discover good input representations at their (final) hidden layer, but hand-crafted representations building on domain knowledge can also be highly effective.

Given the strong support for the Rescorla-Wagner learning rule in the literature on animal learning (Siegel & Allan 1996) and evolutionary biology (Trimmer et al. 2012) and its success in predicting details of human lexical processing with input and output features that have a clear and transparent linguistic interpreta-

tion, we think it makes sense to delve deeper into the benefits of going wide for understanding human error-driven learning.

References

Baayen, R. H., P. Milin, D. Filipović Durdević, P. Hendrix & M. Marelli. 2011. An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review* 118. 438–482.

Chen, T., T. He & M. Benesty. 2015. *xgboost: eXtreme Gradient Boosting*. R package version 0.4-0. https://github.com/dmlc/xgboost.

Friedman, J., T. Hastie & R. Tibshirani. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33(1). 1–22.

Fu, A., S. Aiello, A. Rao, A. Wang, T. Kraljevic & P. Maj. 2015. *h2o: H2O R Interface*. R package version 2.8.4.4. https://CRAN.R-project.org/package=h2o.

Hendrix, P. 2015. *Experimental explorations of a discrimination learning approach to language processing*. University of Tübingen PhD thesis.

Hendrix, P., P. Bolger & R. H. Baayen. 2016. Distinct ERP signatures of word frequency, phrase frequency, and prototypicality in speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Keuleers, Emmanuel, Paula Lacey, Kathleen Rastle & M. Brysbaert. 2012. The british lexicon project: lexical decision data for 28,730 monosyllabic and disyllabic english words. *Behavior Research Methods* 44(1). 287–304.

McClelland, J. L. & D. E. Rumelhart. 1981. An interactive activation model of context effects in letter perception: Part I. An account of the basic findings. *Psychological Review* 88. 375–407.

Meyer, D., E. Dimitriadou, K. Hornik, A. Weingessel & F. Leisch. 2015. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.6-7. https://CRAN.R-project.org/package=e1071.

Milin, P., M. Ramscar, R. H. Baayen & Laurie Beth Feldman. 2016. Discrimination in lexical decision. *Manuscript submitted for publication*.

Minsky, M. & S. Papert. 1969. *Perceptrons: An introduction to computational geometry*. Cambridge, MA.

Pham, H. & R. H. Baayen. 2015. Vietnamese compounds show an anti-frequency effect in visual lexical decision. *Language, Cognition, and Neuroscience* 30(9). 1077–1095.

Rescorla, R. A. & A. R. Wagner. 1972. A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H.

Black & W. F. Prokasy (eds.), *Classical conditioning II: Current research and theory*, 64–99. New York: Appleton Century Crofts.

Siegel, S. & L. G. Allan. 1996. The widespread influence of the rescorla-wagner model. *Psychonomic Bulletin & Review* 3(3). 314–321.

Trimmer, P. C., J. M. McNamara, A. I. Houston & J. A. R. Marshall. 2012. Does natural selection favour the Rescorla-Wagner rule? *Journal of Theoretical Biology* 302. 39–52.