# Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping

Rich Caruana
caruana@cs.cmu.edu
CALD, CMU

Steve Lawrence
http://www.research.microsoft.com/~awf
NEC Research Institute

Lee Giles
giles@ist.psu.edu
Information Sciences Penn State University

The conventional wisdom is that backprop nets with excess hidden units generalize poorly due to overfitting and restricting net capacity prevents overfitting because the net has insufficient capacity to learn models that are too complex. This belief is consistent with a VC-dimension analysis of net capacity vs. generalization: The more free parameters in the net the larger the VC-dimension of the hypothesis space, and the less likely the training sample is large enough to select a (nearly) correct hypothesis [1]. So, in this paper it is shown that nets with excess capacity (more hidden units) also generalize well when trained with backprop and early stopping. After applying large nets on training of real data it is noted that it also appeared to generalize like smaller nets (and sometimes even better). One of the report says "*We find only marginal and inconsistent indications that constraining net capacity improves generalization*" [2].

## 1 Overfitting

Overfitting can vary significantly in different regions of the model. Excess capacity allows better fit to regions of high non-linearity, and backprop often avoids overfitting the regions of low non-linearity. As shown in figure 1, We fit polynomial models with orders 2-20 to the data. Underfitting occurs with order 2. The fit is good with order 10. As the order (and number of parameters) increases, however, significant overfitting (poor generalization) occurs. While in case of MLP the smallest net with one hidden unit (HU) (4 weights weights) underfits the data. The fit is good with two HU (7 weights). Unlike polynomials, however, networks with 10 HU (31 weights) and 50 HU (151 weights) also yield good models.

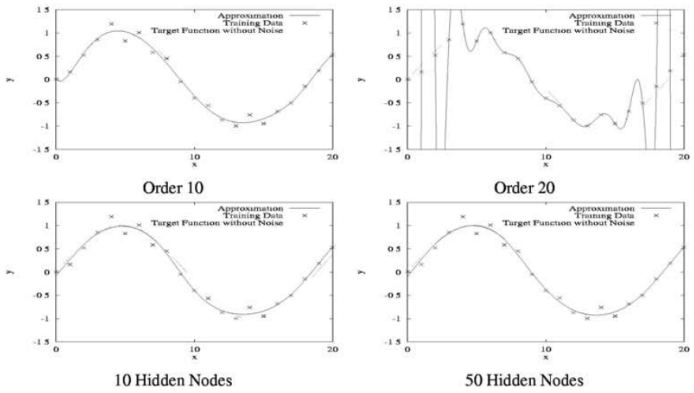


Figure 1: Figure 1: Top: Polynomial fit to data from y = sin( x /3) + v. Order 20 overfits. Bottom: Small and large MLPs fit to same data. The large MLP does not overfit significantly more than the small MLP.

Overfitting can vary significantly in different regions of a model. If we train on small nets it may underfit some regions of graph and perfect on some regions but larger nets, fit the entire function well without significant overfitting in some particular region.

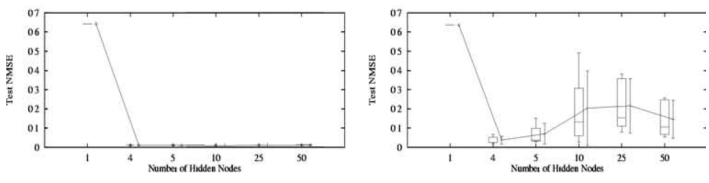


Figure 2: Test Normalized Mean Squared Error for MLPs trained with BP (left) and CG (right). Results are shown with both box-whiskers plots and the mean plus and minus one standard deviation.

Conjugate Gradient (CG) results in lower training error, but overfits significantly. Figure 2 shows results for 10 trials for BP (Back propagation) and CG. Large BP nets generalize better on this problem – even the optimal size CG net is prone to overfitting. The degree of overfitting varies in different regions. When the net is large enough to fit the region of high non-linearity, overfitting is often seen in the region of low non-linearity.

## 2 Generalization, Network Capacity, and Early Stopping

BP nets are less prone to overfitting than expected. But MLPs can and do overfit. We examine overfitting vs net size on seven problem: NETtalk [4], 7 and 12 bit parity, an inverse kinematic model for a robot arm, Base 1 and Base 2: two sonar modeling problems using data collected from a robot wondering hallways at CMU, and vision data used to learn to steer an autonomous car. [3] For each problem we used small training sets (100-1000 points, depending on the problem) so that overfitting was possible. We trained fully connected feedforward MLPs with one hidden layer whose size varied from 2 to 800 HU (about 500-100,000 parameters). All the nets were trained with BP using stochastic updates, learning rate 0.1, and momentum 0.9.

Along with that, we used early stopping for regularization because it doesn't interfere with backprop's ability to control capacity locally. Early stopping combined with backprop is so effective that very large nets can be trained without significant overfitting.

## 3 Results

After examining the results for all seven problems, we observe that on only three (Base 1, Base 2, and ALVINN), do nets that are too large yield worse generalization than smaller networks, but the loss is surprisingly small. Many trials were required before statistical tests confirmed that the differences between the optimal size net and the largest net were significant. Nets trained with conjugate gradient, are more sensitive to net size. Moreover, generalization is hurt more by using a net that is a too small than by using one that is far too large, i.e., it is better to make nets too large than too small.

"*If a BP net with too much capacity would overfit, early stopping could stop training when the model was similar to a model that would have been learned by a smaller net of optimal size.* BP nets appear to be better than CG nets at avoiding overfitting in regions with different degrees of non-linearity (As CG overfits regions of low non-linearity when learning to fit regions of high non-linearity), perhaps because CG is more effective at learning more complex functions that overfit training data, while BP is biased toward learning smoother functions.

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

[2] G.L. Martin and J.A Pittman. Recognizing hand-printed letters and digits using backpropagation learning. In *Neural Computation,*, pages 258–267, 1991.

[3] D.A Pomerleau. An autonomous land vehicle in a neural network. *Advances in Neural Information Processing Systems*, 1:305–313, 1989.

[4] T. Sejnowski and A. W. C. Rosenberg. Parallel networks that learn to pronounce english text. In *Complex Systems*, volume 1, pages 145–168.