Learning curves for stochastic gradient descent in linear feedforward networks

Mohit Kulkarni mohitmk@iitk.ac.in

25th October 2020

1 Learning curves for stochastic gradient descent in linear feedforward networks

Justin Werfel, Xiaohui Xie, H. Sebastian Seung

Abstract: We derive quantitative learning curves for three online training methods used with a linear perceptron: direct gradient descent, node perturbation, and weight perturbation. The maximum learning rate for the stochastic methods scales inversely with the first power of the dimensionality of the noise injected into the system; with sufficiently small learning rate, all three methods give identical learning curves.

1.1 Model

There were N input units and M output units. $y = wx$ was the output to a response x.

$$\begin{aligned} E &= \frac{1}{2}|y-d|^2 = \frac{1}{2}|Wx|^2 \\ E'_{NP} &= \frac{1}{2}|W+\psi x|^2 \\ E'_{WP} &= \frac{1}{2}|(W+\psi)x|^2 \end{aligned} \tag{1}$$

1.1.1 Backpropagation

$$\begin{aligned} \Delta W_{BP} &= -\eta \nabla E = -\eta W x x^T \\ \eta &< \frac{2}{N+2} \quad \text{condition for convergence} \end{aligned} \tag{2}$$

1.1.2 Node Perturbation

$$\begin{aligned} \Delta W_{NP} &= -\frac{\eta}{\sigma^2}(E'_{NP}-E)\xi x^T = -\frac{\eta}{\sigma^2}(\xi^T W x + \frac{1}{2}\xi^T\xi)\xi x^T \\ \eta &< \frac{2}{(N+2)(M+2)} \quad \text{condition for convergence} \end{aligned} \tag{3}$$

1.1.3 Weight Perturbation

$$\begin{aligned} \Delta W_{WP} &= -\frac{\eta}{\sigma^2}(E'_{WP}-E)\psi = -\frac{\eta}{\sigma^2}(x^T\psi^T W x + \frac{1}{2}x^T\psi^T\psi x)\psi \\ \eta &< \frac{2}{(N+2)(MN+2)} \quad \text{condition for convergence} \end{aligned} \tag{4}$$

1. For small η, errors converge at the same rate
2. BP>NP>WP in case of moderate η
3. As long as the output units are small, NP approaches BP in performance

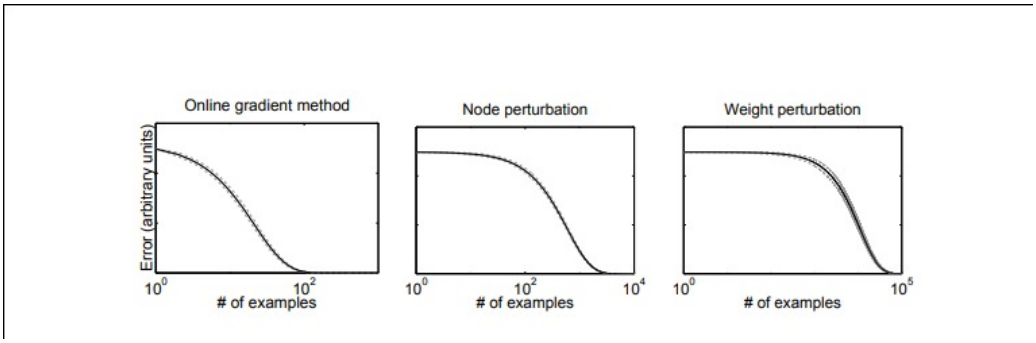

Figure 1: In each case, a network of linear units with N = 20, M = 25, $\sigma = 10^{-}3, and optimal \eta$ was trained on successive input examples for the number of iterations shown. 100 such runs were averaged together in each case; the three gray lines show the mean (solid) and standard deviation (dashed) of squared error among those runs.