CS156b Project1

Letter Recognition — Architecture

Ling Li
Learning Systems Group, Caltech
October 23, 2001

Gradient Descent

- a small learning rate ($\eta = 0.0001$)
- with momentum ($\mu = 0.5$)
- with adaptive learning rate ($\alpha = 1.05$, $\beta = 0.5$)

Line Search

Use method in CS156a (with 5 points). When the 1^{st} step is too large, shrink it first.

Conjugate Gradient

When to reset? After 15 steps, or when $\beta < 0$.

Weighted error function

Two continuous error functions, (W1/W3)

$$e^{(1)} = \frac{1}{2}\sum_i (f_i - y_i)^2, \quad e^{(3)} = e^{(1)} + (f_c - y_c)^2,$$

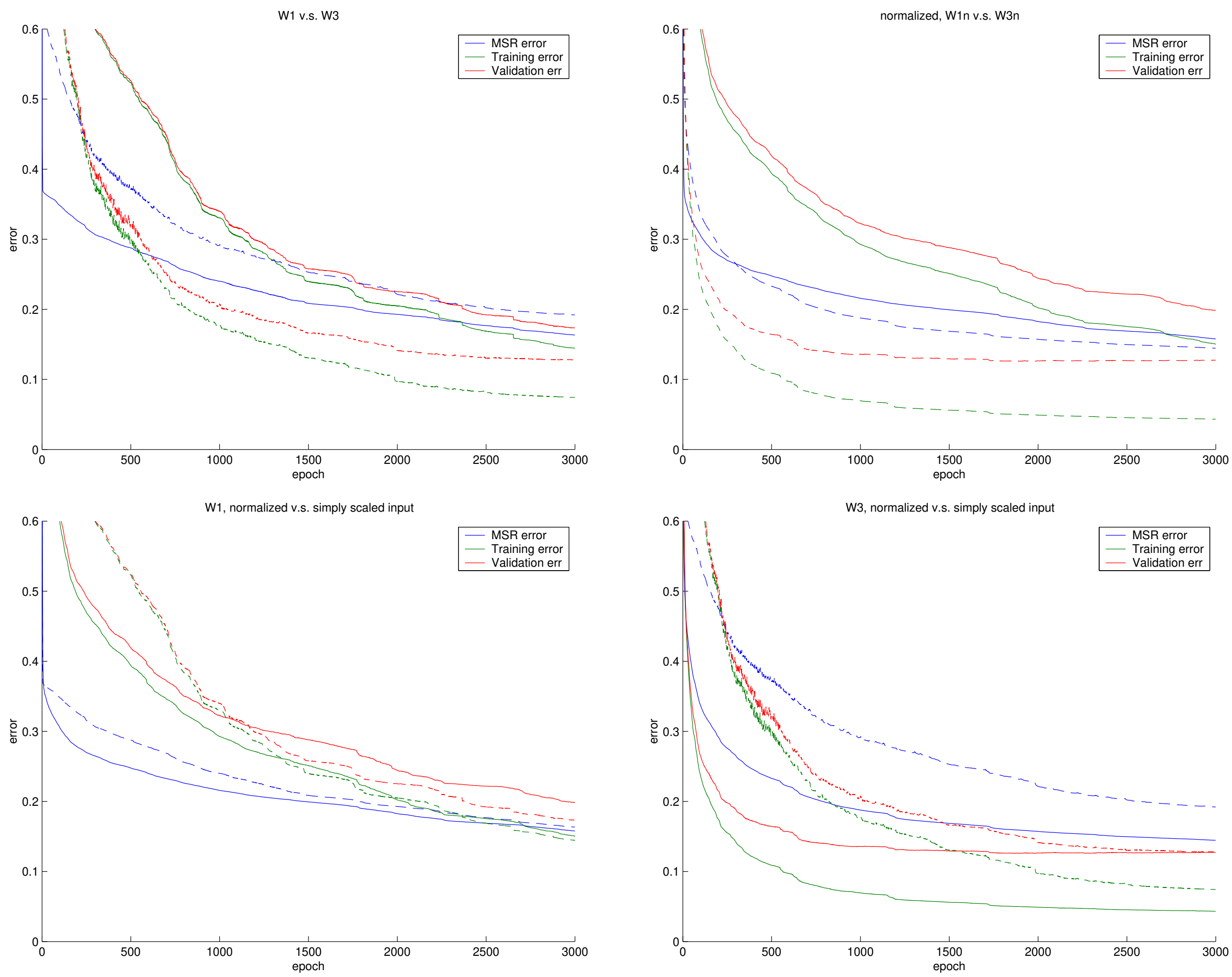
where c is the desired class, are used.

Normalization

1. by mean/variance; 2. by simple scale.

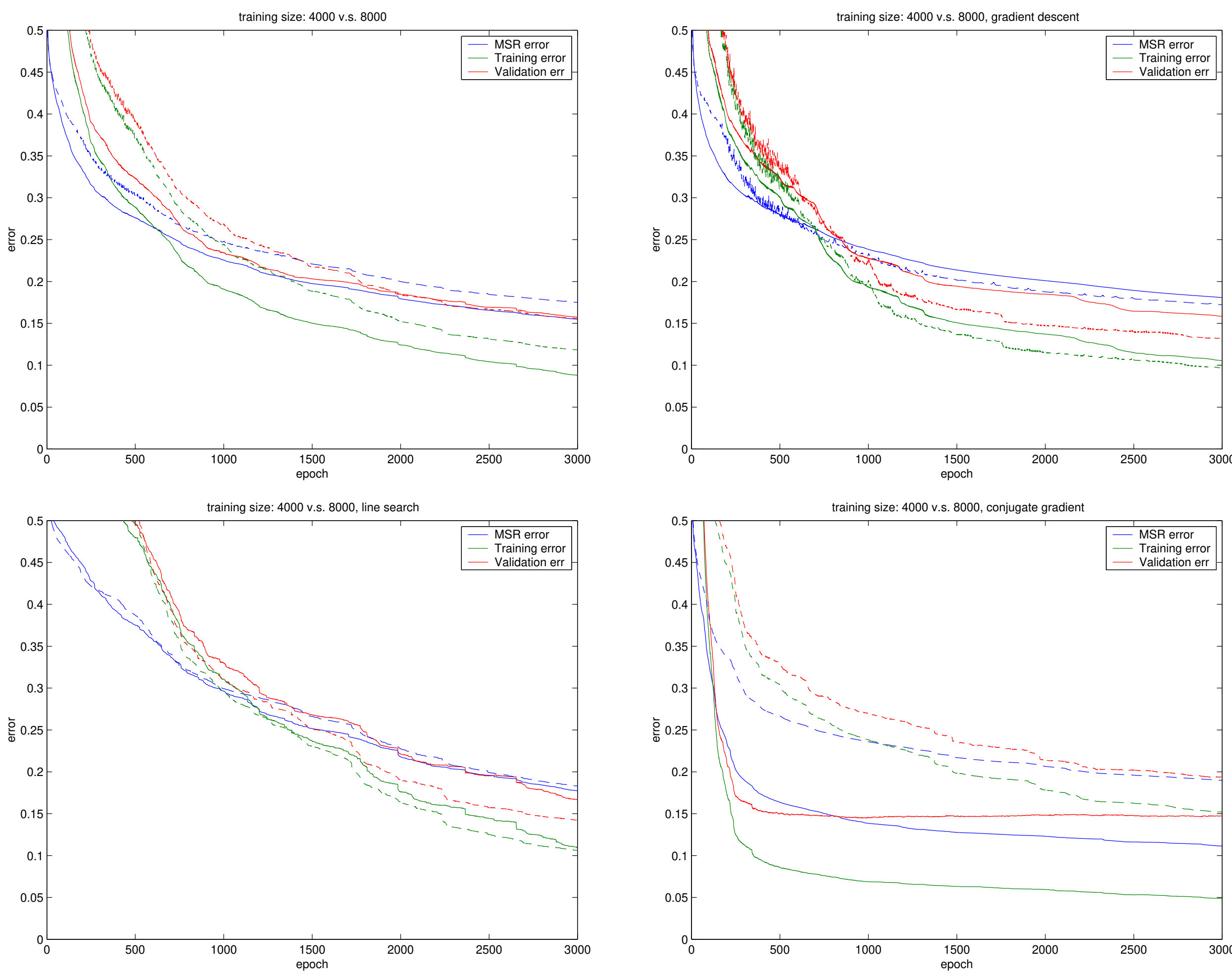
Normalization and Weighted error

Average errors over different training algorithms, different training set size, we compare the effectiveness of normalization and weighted error function.



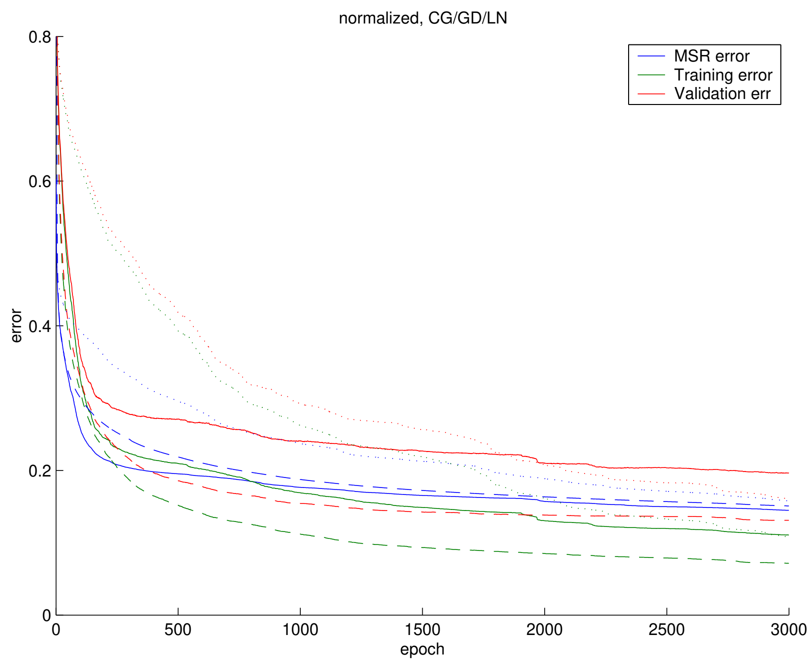
Training data size

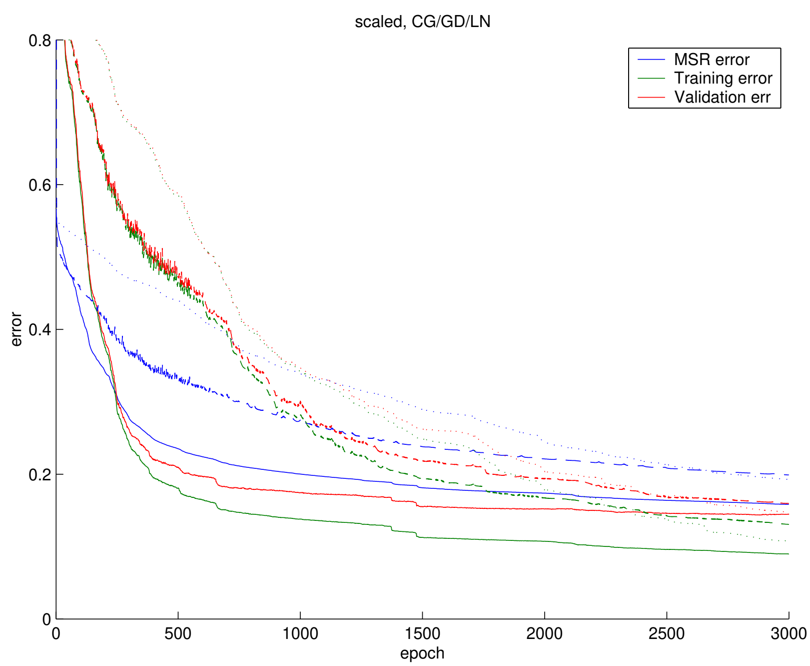
4000 for training, 8000 for validation; and vice versa.



Training algorithms

Note that one run (W1, normalized, 8K, CG) got stuck and the error was really huge (33%).

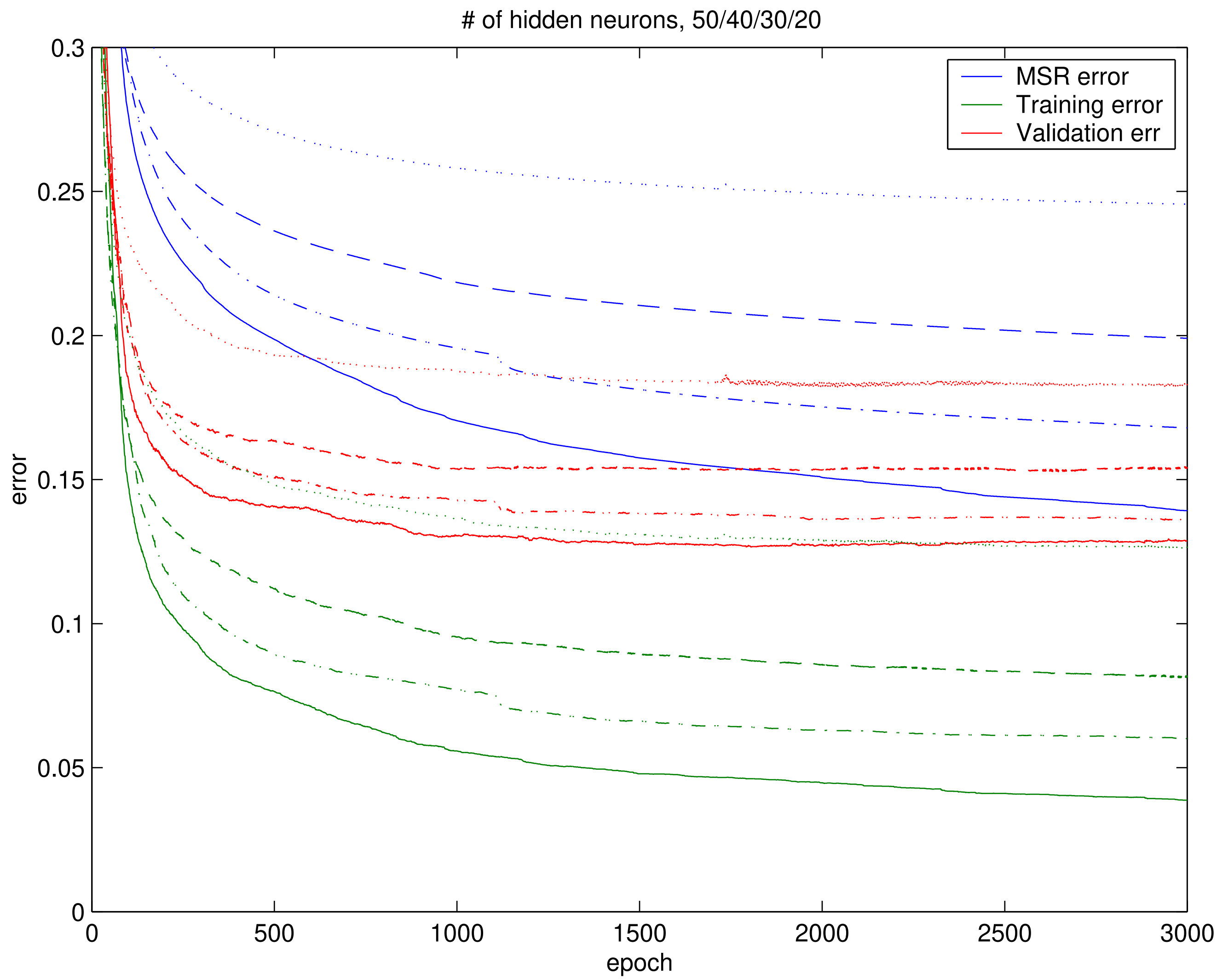




of hidden neurons

Architecture: 16-N_H-26, where $N_H = 20, 30, 40, 50$.

Normalized data and $e^{(3)}$ (W3) error function.



average over CG/GD/LN and 4K/8K