Kronecker factored Newton method

Yaroslav Bulatov

April 10, 2017

Suppose we have a relu network, prediction for a single example, or a group of examples with same ReLU activation pattern, can be written without loss of generality as follows:

$$B'WA$$

The gradient update rule for W with learning rate 1 can be written as

$$W + \hat{B}A' \to W$$

The quantity $\hat{B}$ is known as the matrix of backprop values.
Second derivative of loss with respect to W can be written as:

$$AA' \otimes BB'$$

To compute true Hessian we need to sum up these values over examples i

$$H = \sum_i A_i A_i' \otimes B_i B_i'$$

We obtain Kronecker factorization to obtain the following approximation

$$H \approx (\sum_i A_i A_i') \otimes (\sum_i B_i B_i')$$

Using this as the preconditioner, our gradient descent step becomes as follows:

$$W + G\hat{B}A' \to W$$

Where

$$G = (\sum_i A_i A_i')^{-1} (\sum_i B_i B_i')^{-1}$$