Kfac assumption

Yaroslav Bulatov

April 10, 2017

Suppose we have a left matmul operation, $y = Wx$. The gradient with respect to W evaluated on example i can be written as

$$G_i = b_i a_i'$$

Quantities b and a are column vectors representing backprops and activations. Note that this has the same shape as W.

For purposes of computing covariance, we flatten this gradient

$$g_i = \mathrm{vec}(G_i) = a_i \otimes b_i$$

Covariance can now be expressed as

$$\mathrm{cov} = \frac{1}{n}\sum_i g_i g_i' = \frac{1}{n}\sum_i (a_i a_i') \otimes (b_i b_i')$$

To obtain natural gradient preconditioner we seek a transformation that renders new covariance matrix diagonal.

Let $\hat{a_i} = Ua$ and $\hat{b_i} = Vb$, the corresponding covariance is

$$\bar{\mathrm{cov}} = \frac{1}{n}\sum_i U a_i a_i' U' \otimes V b_i b_i' V'$$

It's easy to find U and V if we assume the quantity above is equalent to

$$\bar{\mathrm{cov}} \approx \frac{1}{n}(\sum_i U a_i a_i' U') \otimes (\sum_i V b_i b_i' V') = \frac{1}{n} UAA'U' \otimes VBB'V$$

From this approximation is follows that we should pick $U = A^{-1}$ and $V = B^{-1}$