CS-GY 9223G / Fall 2020

Homework 2

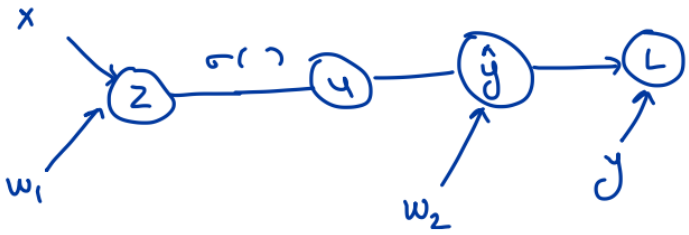
1. **(3 points)** *Exercises in backpropagation*. Consider a one-hidden layer neural network (without biases for simplicity) with sigmoid activations trained with squared-error loss. Draw the computational graph and derive the forward and backward passes used in backpropagation for this network.

$$\hat{y} = W_2\sigma(W_1 x), \qquad \mathcal{L} = \|\hat{y} - y\|_2^2$$

Qualitatively compare the computational complexity of the forward and backward passes. Which pass is more expensive and by roughly how much?

Solution

The computation graph for this network is drawn below.



The forward pass of the network is given by the following set of equations:

$$\begin{aligned} Z &= W_1 x \\ u &= \sigma(Z) \\ \hat{y} &= W_2 u \\ L &= (\hat{y} - y)^2 \end{aligned}$$

The backward pass of the network can be given by the following set of equations (slight abuse of notation, we are representing gradients via the partial derivative symbol ∂):

$$\begin{aligned} \frac{\partial L}{\partial L} &= 1 \\ \frac{\partial L}{\partial \hat{y}} &= 2(\hat{y} - y) \\ \frac{\partial L}{\partial W_2} &= \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial W_2} = 2(\hat{y} - y) \cdot u^T \\ \frac{\partial L}{\partial u} &= \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial u} = 2W_2^T \cdot (\hat{y} - y) \\ \frac{\partial L}{\partial Z} &= \frac{\partial L}{\partial U} \cdot \frac{\partial u}{\partial Z} = 2[W_2^T \cdot (\hat{y} - y)] \odot \sigma'(Z) \\ \frac{\partial L}{\partial W_1} &= \frac{\partial L}{\partial Z} \cdot \frac{\partial Z}{\partial W_1} = \frac{\partial u}{\partial Z} \cdot x^T \end{aligned}$$

Assuming that matrix-multiplies are the dominant cost, we see that the cost of the backward pass is roughly twice as much as the forward pass.

2. **(3 points)** *Slow rate of descent*. Consider a simple function having two weight variables:

$$L(w_1, w_2) = 0.5(aw_1^2 + bw_2^2).$$

   a. Write down the gradient $\nabla L(w)$, and immediately conclude the weights w^* that achieve the minimum value of L.

   b. Instead of simply writing down the optimal weights, let's now try to optimize L using gradient descent. Starting from some arbitrary initialization point $w_1(0), w_2(0)$, write down the gradient descent updates. Show that the updates have the form:

$$w_1(t+1) = \rho_1 w_1(t), \quad w_2(t+1) = \rho_2 w_2(t)$$

   where $w_i(t)$ represent the weights at the t^{th} iteration. Derive the expressions for ρ_1 and ρ_2 in terms of a, b, and the learning rate.

   c. Under what values of the learning rate does gradient descent converge?

   d. Provide a scenario under which the convergence rate of gradient descent is very slow. (*Hint: consider the case where a/b is a very large ratio.*)

Solution

   a. We calculate the gradient as follows:

$$\begin{aligned}\nabla L(w_1, w_2) &= \begin{bmatrix} \frac{\partial L}{\partial w_1} \\ \frac{\partial L}{\partial w_2} \end{bmatrix} \\ &= \begin{bmatrix} aw_1 \\ bw_2 \end{bmatrix}.\end{aligned}$$

   We can see that the gradient becomes zero when $w_1 = 0, w_2 = 0$.

   b. Assuming that the learning rate is η, the descent updates can be written as:

$$\begin{aligned} w_1(t+1) &= w_1(t) - \eta \frac{\partial L}{\partial w_1} = w_1(t) - \eta a w_1(t) = (1 - \eta a) w_1(t) \\ w_2(t+1) &= w_2(t) - \eta \frac{\partial L}{\partial w_2} = w_2(t) - \eta b w_2(t) = (1 - \eta b) w_2(t) \end{aligned}$$

   Therefore,

$$\rho_1 = 1 - \eta a, \qquad \rho_2 = 1 - \eta b$$

   c. For the sequence of updates of w_1 and w_2 to converge, the *magnitudes* of the multiplicative factors ρ_1 and ρ_2 have to be smaller than 1. If either is greater than 1, the sequence diverges. Therefore, $|1 - \eta a| < 1$, which implies that $\eta < 2/a$. Simultaneously, $\eta < 2/b$. Therefore, a sufficient condition for both to hold is given by:

$$\eta < 2/\max(a, b).$$

d. For rapid convergence, we want both ρ_1, ρ_2 to be as close to zero as possible. Assume without loss of generality that $a > b$, and we choose a learning rate $\eta = 1/a$. Then, $\rho_1 = 0$, which means that the iterates for w_1 converge within a single iteration. However, in this case $\rho_2 = 1 - b/a$, which can be very close to 1, if the ratio a/b is very large. Therefore, the iterates for w_2 can take a very long time to converge. (Convince yourself that the above argument holds even for other choices of learning rate.)

3. **(4 points)** Open the (incomplete) Jupyter notebook provided as an attachment to this homework in Google Colab (or other cloud service of your choice) and complete the missing items. Save your finished notebook in PDF format and upload along with your answers to the above theory questions in a single PDF.