Deep Learning (SS2020)
Seminar 4

May 3, 2020

Assignment 1 (Weight initialisation for ReLU networks)
In this assignment we derive a proper weight initialisation for ReLU networks using the same approximation as in Lecture 8, where we derived the initialisation proposed by Glorot et al. (2010) for networks with sigmoid like activations.

Consider a network with randomly initialized weights. Let x^k denote the output vector for the layer k of the ReLU network. Because of the ReLU activations we can not require x^k to have zero mean. We will therefore consider the statistic of the activations $a^k = W^k x^{k-1}$.

a) Prove that variance of the activations $a^k = W x^{k-1}$ in layer k is

$$\mathbb{V}[a^k] = n_{k-1}\mathbb{V}[W^k]\,\mathbb{E}[(x^{k-1})^2]$$

if the weights have zero mean.

b) Prove that the distribution of a^k is symmetric with zero mean, provided the same holds for the distribution of W^k.

c) Conclude that passing the a^k-s through the ReLU-function will lead to $\mathbb{E}[(x^k)^2] = \frac{1}{2}\mathbb{V}[a^k]$.

Collecting the steps, we get

$$\mathbb{V}[a^k] = \frac{1}{2} n_{k-1}\mathbb{V}[W^k]\,\mathbb{V}[a^{k-1}]$$

and obtain the initialisation proposed by He et al. (2015): initialising the weights with zero mean and variance

$$\frac{1}{2} n_{k-1}\mathbb{V}[W^k] = 1.$$

Assignment 2 (Batch Norm)
Batch normalization after a linear layer with a weight matrix W and bias b takes the form:

$$\frac{Wx + b - \mu}{\sigma}\beta + \gamma. \tag{1}$$

Let x_i, $i = 1 \dots m$ inputs to the linear layer in a batch and let $a_i = Wx_i + b$. Then μ is the sample mean of a_i and σ^2 is the sample variance of a_i.

a. Show that the output of batch normalization (1) does not depend on the value of bias b and also does not change when the weight matrix W is scaled by a positive constant.

b. What changes if BN is applied after ReLU?

c. Consider a network without BN. Let μ and σ be statistics of neuron activations a in a particular layer. How to introduce a BN layer at this place so that it does not change the network predictions? I.e. how to initialize β and γ?

Assignment 3 (Prox Problems)
Using the technique of Lagrange multipliers solve the following problems:

a. Optimal step with box trust region: $\min\limits_{\|\Delta x_i\| \le \varepsilon \ \forall i} \langle \nabla f(x_0), \Delta x \rangle$.

Hint: Square the constraints to allow solving via stationary point, e.g. $\Delta x_i^2 \le \varepsilon^2$. Lagrange multipliers to inequality constraints $\Delta x_i^2 - \varepsilon^2 \le 0$ must be non-negative.

b. Optimal step with Mahalanobis trust region: $\min\limits_{\|\Delta x\|_M \le \varepsilon} \langle \nabla f(x_0), \Delta x \rangle$,
where $\|\Delta x\|_M^2 = \langle \Delta x, M \Delta x \rangle$.

Assignment 4 (Mirror Descent)
Solve the MD step proximal problem:

$$\min_x \langle \nabla f(x_0), x - x_0 \rangle + \lambda D(x, x_0),$$

where $x_0 \in (0, 1)$ and $D(x, x_0) = x \log \frac{x}{x_0} + (1 - x) \log \frac{1-x}{1-x_0}$.
Hint: The problem is convex and can be solved by stationary point conditions.