DEEP LEARNING (SS2021)
SEMINAR 4

Assignment 1 (Weight initialisation for ReLU networks)**.** In this assignment we derive a proper weight initialisation for ReLU networks. We will assume that the components of all vectors are statistically independent and identically distributed. Given a vector a, we denote by $\mathbb{E}[a]$ and $\mathbb{V}[a]$ the mean and variance of its components.
Consider a network with randomly initialized weights. Let x^k denote the output vector for the layer k of the ReLU network.

a) Prove that variance of the activations $a^k = Wx^{k-1}$ 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. 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 Normalization)**.** Batch normalization after a linear layer with a weight matrix W and bias b takes the form:

$$\frac{Wx + b - \mu_{\mathcal{B}}}{\sigma_{\mathcal{B}}}\beta + \gamma, \tag{1}$$

where $\mu_{\mathcal{B}}$ and $\sigma_{\mathcal{B}}$ denote the mean and standard deviation of the layer output $a = Wx+b$ taken over a batch.

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

b) Which of the two properties in a) hold if BN is applied after ReLU?

c) Consider a network without BN. Let $\mu_{\mathcal{B}}$ and $\sigma_{\mathcal{B}}$ be the statistics of neuron activations a in a particular layer. We want to introduce a BN layer at this place so that it does not change the network predictions. How shall we initialize β and γ?

Assignment 3 (Dropout, Bernoulli)**.**

a) Dropout noise model can be reformulated for a more convenient implementation. Consider the following Bernoulli noises:

$$Z = \begin{cases} a, & \text{with probability } p \\ 0, & \text{with probability } 1-p \end{cases} \tag{2}$$

What should be the value of a so that $\mathbb{E}[Z] = 1$? This will allow us to avoid rescaling of the weights at the test time and just apply this noise at the training time.

b) Sometimes randomized procedures are used to quantize the gradients for a faster communication in a distributed system. Let the gradient $g \in \mathbb{R}^n$ be computed at the worker. The worker can sends a *quantized* gradient $\tilde{g} \in \{0,1\}^n$ to the server, using only 1 bit per coordinate. The worker can additionally send two real numbers to the server a, b. How to chose the quantization procedure in a randomized way so that $a\mathbb{E}[\tilde{g}]+b = g$ and hence we preserve the guarantee of an unbiased (but more noisy) gradient estimate?

Assignment 4 (Ridge Regression)**.** Consider linear regression model with

$$y_i = w^\mathsf{T} x_i + \varepsilon_i, \tag{3}$$

where i is a data point, $x_i \in \mathbb{R}^n$, $y_i \in \mathbb{R}$, $w \in \mathbb{R}^n$ is the weight vector and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ are independent measurement errors.

a) Formulate the maximum likelihood learning for this problem and express the log likelihood.

Hint: you should get the mean squared error loss function.

Hint: Since $\varepsilon_i = y_i - w^\mathsf{T} x_i$ is normally distributed, we have $p(y_i|x_i) = p_{\mathcal{N}}(y_i - w^\mathsf{T} x_i; 0, \sigma^2)$.

b) Consider now also noises in the input:

$$y_i = w^\mathsf{T}(x_i + \xi_i) + \varepsilon_i, \tag{4}$$

where $\xi_i \sim \mathcal{N}(0, \lambda^2 I_n)$ are independent. Compute the expected value in ξ of the MSE loss function derived above. You should obtain a variant of weight decay regularization term.

c) Formulate the maximum likelihood learning and write the log likelihood for the problem (4). What regularization we get in this case?

Hint: write the likelihood of the observed data point (x_i, y_i) integrating out the unobserved noises ξ_i, ε_i.