DEEP LEARNING (SS2022)
SEMINAR 4

Assignment 1 (Weight initialization for ReLU networks)**.** In this assignment we derive a proper weight initialization for ReLU networks. We will assume that the components of all vectors are statistically independent and identically distributed.

a) Let us consider a single neuron with weight vector w and input vector x. Its pre-activation is $a = w^T x$. Let us denote

$$\mathbb{E}[x_i] = \mu,\ \mathbb{E}[x_i^2] = \chi,\ \mathbb{E}[w_i] = 0, \text{ and } \mathbb{V}[w_i] = v.$$

prove that $\mathbb{E}[a] = 0$ and $\mathbb{V}[a] = nv\chi$, where n is the dimension of the vectors x and w.

b) Show that the distribution of a is symmetric if so is the distribution of w.

c) Consider the neuron output $y = g(a)$, where g denotes the ReLU function. Conclude that $\mathbb{E}[y^2] = \frac{1}{2}\mathbb{V}[a]$.

d) Let us denote $\mathbb{V}[a] = \alpha$ and consider a ReLU network with layers $k = 1, \dots, m$. Collecting the previous steps we get the following recursive relation for the α_k

$$\alpha_k = \frac{1}{2} n_{k-1} v_k \alpha_{k-1}$$

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

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

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) What is the mean and standard deviation of the BN-normalized layer, if we initialize $\beta = 1$, $\gamma = 0$? Assume, we decided to apply BN after each linear layer. Has the weight initialization from Assignment 1 still an effect for the forward pass?

c) Consider a network without BN. Let $\mu_{\mathcal{B}}$ and $\sigma_{\mathcal{B}}$ be the statistics of layer output $a = Wx + b$. 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 (SGD + L2)**.** Consider a regularized loss function $\tilde{L}(\theta) = L(\theta) + \frac{\lambda}{2}\|\theta\|^2$. Let g be a stochastic gradient estimate of L at θ. Notice that the regularization part of the objective, $\frac{\lambda}{2}\|\theta\|^2$, is known in a closed form and so its gradient g_r is non-stochastic.

- Design an SGD algorithm that applies momentum (exponentially weighted averaging) to g only but not to g_r.
- Is it equivalent to an SGD with the momentum applied to both g and g_r but possibly with a different settings of λ, momentum and learning rate?

Assignment 4 (Mixup)**.** The mixup data augmentation draws (x_1, y_1) and (x_2, y_2) at random from data distribution p^*, where y_1 and y_2 are one-hot encoded target labels, and constructs

$$\tilde{x}_\lambda = \lambda x_1 + (1-\lambda)x_2 \tag{2a}$$
$$\tilde{y}_\lambda = \lambda y_1 + (1-\lambda)y_2. \tag{2b}$$

The value of λ is drawn at random from Beta distribution $\mathcal{B}e(\alpha, \alpha)$ with α fixed (*e.g.*, 0.1). The training objective is the expected loss over all such mixup examples:

$$\mathbb{E}_{(x_1,y_1)\sim p^*}\mathbb{E}_{(x_2,y_2)\sim p^*}\mathbb{E}_{\lambda\sim\mathcal{B}e(\alpha,\alpha)} l(\tilde{x}_\lambda, \tilde{y}_\lambda), \tag{3}$$

where $l(x, y)$ is the loss function of neural network predictions with input x with respect to the target y. We will show that in the case of cross-entropy loss l, this it can be reformulated without using y_2, *i.e.*, not mixing labels. Therefore, even unlabeled data may be used for x_2 in the reformulation.

a) Show that the expected mixup loss (3) equals

$$2\mathbb{E}_{(x_1,y_1)\sim p^*}\mathbb{E}_{(x_2)\sim p^*}\mathbb{E}_{\lambda\sim\mathcal{B}e(\alpha,\alpha)} \lambda l(\tilde{x}_\lambda, y_1). \tag{4}$$

Hint: you will need:

- Linearity of the cross-entropy function to show that $l(x, y)$ is linear in y;
- Symmetry of Beta distribution: $\lambda \sim \mathcal{B}e(\alpha, \alpha) \Rightarrow (1-\lambda) \sim \mathcal{B}e(\alpha, \alpha)$;
- Symmetry of the expected loss with respect to swapping (renaming) (x_1, y_1) and (x_2, y_2).

b) Prove that $2\lambda p_{\mathcal{B}e(\alpha,\alpha)}(\lambda) = p_{\mathcal{B}e(\alpha+1,\alpha)}(\lambda)$ and use it to simplify the result. *Hint:* you will need:

- Density of Beta distribution: $p_{\mathcal{B}e(\alpha,\beta)}(\lambda) = \lambda^{\alpha-1}(1-\lambda)^{\beta-1}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$;
- One of the defining properties of Gamma function: $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$.