Deep Learning Exam (OPT8)

Caio Corro, Michèle Sebag

April 22, 2022

When you answer questions, you must give an explanation that shows that you have a good understanding of the answer. Formal notations are mandatory and handwavy explanations should be avoided. You can use examples (and even draw them) if it is beneficial for the explanation. You can answer questions in English or French.

1 Basic questions (10 points)

1. **(2 points)** Describe formally a Multi-Layer Perceptron (MLP) for multi-class classification with two hidden layers.
2. **(2 points)** Show an example of a binary classification dataset that is problematic for a linear model but not for a MLP. What intuition can you give on what a neural network do in this case? (think of the output layer).
3. **(2 points)** What property of activation functions can lead to the vanishing gradient problem? Explain which activation functions are better than which other ones with respect to this property.
4. **(2 points)** Explain formally what is dropout and why it is used.
5. **(2 points)** Minimizing a function $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}$ using standard gradient descent method is simply computing a sequence of arguments as follows:

   $$a^{(t+1)} = a^{(t)} - \epsilon \nabla \mathcal{L}(a^{(t)})$$

   where $\epsilon > 0$ is the stepsize and $a^{(t)} \in \mathbb{R}^d$ is the parameter vector of the model at timestep t. A popular technique to avoid overfitting is weight decay, that is the gradient descent updates are replaced with the following equation:

   $$a^{(t+1)} = (1 - \lambda)a^{(t)} - \epsilon \nabla \mathcal{L}(a^{(t)})$$

   where $\lambda > 0$ is the weight decay parameter. Prove that weight decay is equivalent to adding a L2-regularization term to the objective. In terms of implementation, what is a benefit of the weight decay formulation for deep neural networks?

2 Backpropagation algorithm (20 points)

1. **(4 points)** What is a computational graph? Why is it useful? What information does it contains?
2. **(10 points)** Let $a \in \mathbb{R}$, $b \in \mathbb{R}$ be the parameters of a simple binary linear classifier. Given an input $x \in \mathbb{R}$ and an output label $y \in \{0, 1\}$, the negative log-likelihood for binary classification is defined as follow:

   $$l(x, a, b) = -(ax + \ b) + \log(1 + \exp(ax + b))$$

   Use this simple example to explain the backpropagation algorithm.
3. **(3 points)** Why can in-place operations be problematic? Give an example where in-place operations will make the backpropagation algorithm fail and one example where the use of an in-place operation is not a problem.
4. **(3 points)** In the case of Monte-Carlo approximation of an expectation, explain what is the reparameterization trick and why is it useful. Give an example where it is used.

3 Advanced questions (14 points)

1. **(2 points)** Why would you use a convolutional architecture ? (Several possible answers). Cite an alternative approach.
2. **(4 points)** The image size is $n \times p$ pixels. Your architecture includes K filters. Each filter has $n' \times p'$ pixels as input ($n' << n$; $p' << p$), stride is s (same for both height and width). How many weights do you want to learn ? How many weights should you learn for a same architecture with no weight sharing (not convolutional) ?
3. **(2 points)** Cite several invariance operators (images; sounds).
4. **(2 points)** What is an auto-encoder ? How would you choose the size of the latent space ? (Several possible answers)
5. **(2 points)** How would you use an auto-encoder to characterize an outlier ? (Several possible answers again)
6. **(2 points)** Explain informally the mechanism of a Generative Adversarial Network (add the equation if you want). What can go wrong ? How could you help (modifying the learning setting) ? (Several possible answers).

Cheat sheet

Chain rule

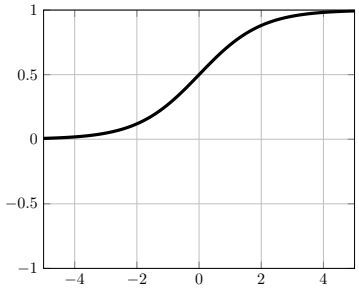
Let $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ be two functions. Assume $a \in \mathbb{R}$, we define $b \in \mathbb{R}$ and $c \in \mathbb{R}$ as follows:

$$b = f(a),$$
$$c = g(b),$$

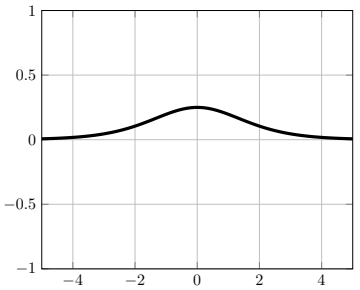
Then:

$$\frac{\partial}{\partial a}c = \left(\frac{\partial}{\partial b}c\right)\left(\frac{\partial}{\partial a}b\right) = g'(b)f'(a)$$

Sigmoid

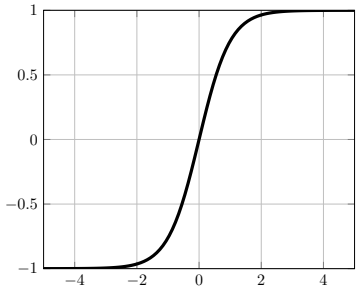


$$\sigma(x) = \frac{1}{1 + \exp(-x)} = \frac{\exp(x)}{1 + \exp(x)}$$

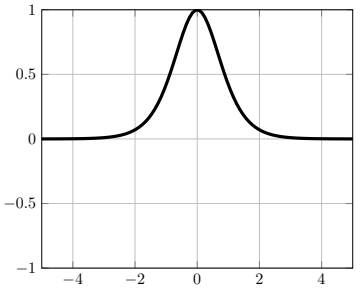


$$\sigma'(x) = \sigma(x)(1 - \sigma(x))$$

Hyperbolic tangent

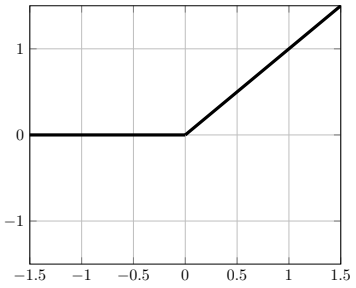


$$\tanh(x) = \frac{1 - \exp(-2x)}{1 + \exp(-2x)}$$

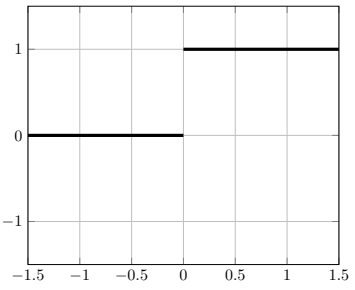


$$\tanh'(x) = 1 - \tanh(x)^2$$

Rectified Linear Unit



$$\text{relu}(x) = \begin{cases} 0 \text{ if } \leq 0 \\ x \text{ otherwise} \end{cases}$$



$$\text{relu}'(x) = \begin{cases} 0 \text{ if } x < 0 \\ 1 \text{ if } x > 0 \\ \text{undefined otherwise} \end{cases}$$