Deep Learning Exam (OPT8)

Caio Corro, Michèle Sebag

November 20, 2019

When you answer questions, you must give an explanation that shows that you have a deep understanding of the answer. You can use examples (and even draw them). You can answer questions in English or French. All three parts are independent of each other.

- Students in the 1st year of master: answer questions in parts 1 to 4 (i.e. you can skip the deep generative models section)
- Students in the 2nd year of master: answer all questions!

Write explicitly on your document if you are in first or second year.

1 (5pts) Neural networks basics

1. **(2pts)** Describe a multilayer perceptron with one hidden layer for classification (e.g. MNIST digit classification) and the associated negative log-likelihood loss used for training. You should introduce the notation you use and describe the different components.
2. **(1pts)** What is the update expression for parameters during training? (we assume we rely on stochastic gradient descent with a single example per batch)
3. **(2pts)** Derive the gradient of the output layer parameters (output projection and output bias) with respect to the negative log-likelihood loss.

2 (10pts) Neural network training

1. **(1pts)** During training, we observe that the training loss is decreasing but the dev loss is increasing. What is happening? ("there is a bug in the code" is not a valid answer!) How can you prevent this issue?
2. **(1pts)** During training, we observe that the training loss is increasing. What is happening? ("there is a bug in the code" is not a valid answer!) How can we prevent this issue?
3. **(1pts)** What is the difference between the perceptron loss and the hinge loss?
4. **(1pts)** What is one problem of the sigmoid activation function for deep neural networks?
5. **(2pts)** Let's consider a (deep) multilayer perceptron using tanh activation functions for each hidden layer. What properties do you want your network to have at initialization? Why?
6. **(1pts)** What are the pros and cons of the relu activation function with respect to its gradient?
7. **(1pts)** What are the benefits of using batches of several instances during training?
8. **(2pts)** A simple way to optimize a neural network is to rely on the stochastic gradient descent algorithm. Let's put aside the fact that this method rely on a Monte-Carlo estimation of the gradient, what are two other limits of this optimization method? Are they easy to circumvent?

3 (6pts) Neural network implementation

1. **(1pts)** What can be problematic in the implementation of the softmax function? How do you fix this problem?
2. **(1pts)** What is a computation graph?
3. **(2pts)** Give a brief description of the back-propagation algorithm.
4. **(2pts)** What kind of information do you need to store to compute the gradient? Why are in-place operations problematic? Give an example.

4 (4pts) Convolutional Neural Networks (CNNs)

1. **(2pts)** What is the main property of CNNs?
2. **(2pts)** Describe the different components of a CNN.

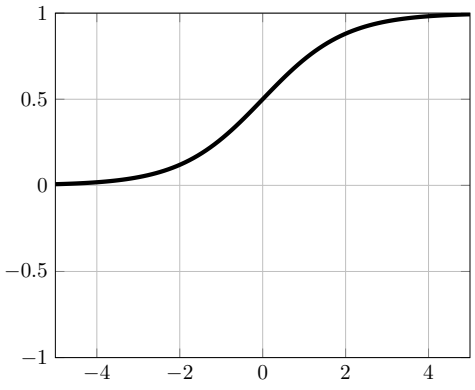
5 (10pts) Deep generative models: Variational Auto-Encoders (VAEs) and Generative Adversarial Networks (GANs)

Student in the 2nd year of master only.

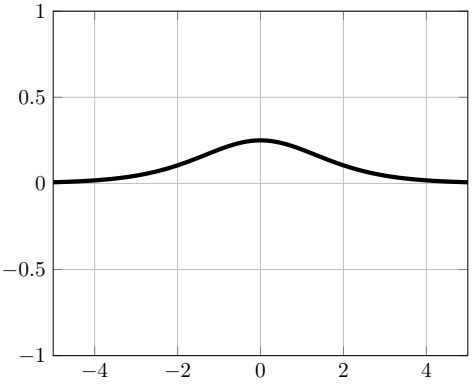
1. **(2pts)** How would you describe the differences between a standard Auto-Encoder and a Variational Auto-Encoder? For which purpose would you use one or the other?
2. **(2pts)** What technical constraints do VAEs impose on the latent variable distribution? Similar question for GANs? That is, is there distribution there cases where you can use one and not the other? (think about the loss functions and how you implement these networks) In which case?
3. **(1pts)** We assume we train a GAN with the binary cross entropy loss for the discriminator. What problem can we encounter when updating the generator parameters? Let $y \in \{0, 1\}$ be the gold label (0 if the discriminator input was generated, 1 if it comes from the dataset) and $w \in [0, 1]$ be the output of the discriminator that was compute with a sigmoid function, the binary cross entropy loss is:
$$-(y \log p + (1 - y) \log(1 - p))$$
4. **(2pts)** VAEs are usually trained by maximizing the Evidence Lower Bound. How can you interpret terms 1 and 2?
$$\max_{\theta,\phi} \quad \underbrace{\mathbb{E}_{q_\phi(z|x)}\left[p_\theta(x|z)\right]}_{1} - \underbrace{\mathrm{KL}\left[q_\phi(z|x), p(z)\right]}_{2}$$
5. **(1pts)** Let $q_\phi(z|x)$ and $p(z)$ be normal distributions. Can you compute these terms?
6. **(1pts)** What is the reparameterization trick and why is it necessary?
7. **(1pts)** Let $q_\phi(z|x)$ and $p(z)$ be categorical distributions (i.e. z is a random variable defined on a finite set). Is the ELBO computation tractable? Explain.

Cheat sheet

Sigmoid

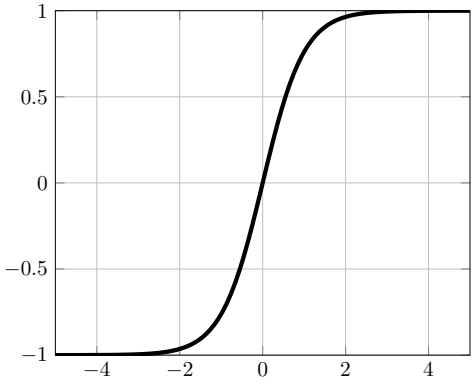


$$\sigma(x) = \frac{1}{1 + \exp(-x)}$$

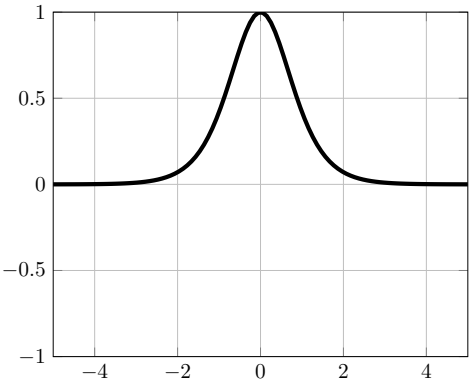


$$\sigma'(x) = \sigma(x)(1 - \sigma(x))$$

Hyperbolic tangent

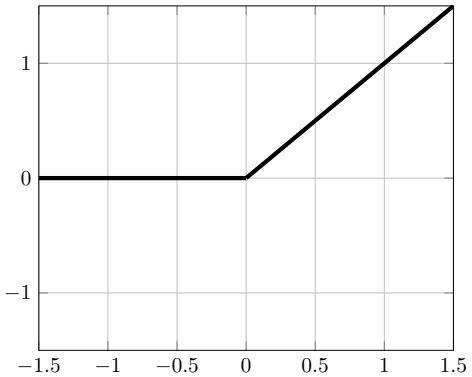


$$\tanh(x) = \frac{1 - \exp(-2x)}{1 + \exp(-2x)}$$

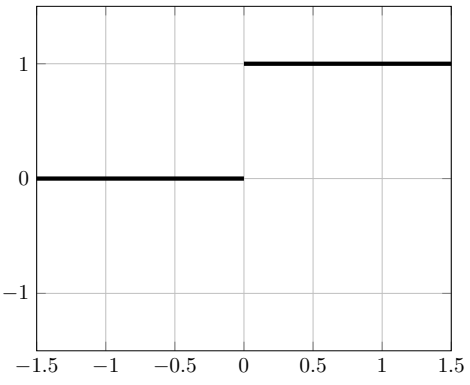


$$\tanh'(x) = 1 - \tanh(x)^2$$

Rectified Linear Unit



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



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