Exercise 1 - Maximum Likelihood Estimation Refresher

Assume you are given datapoints $(\mathbf{x}_i)_{i=1}^N$ with $\mathbf{x}_i \in \mathbb{R}^n$ coming from a Gaussian distribution. Derive the maximum likelihood estimator of its mean.

Exercise 2 - More Gradients

You are an ML Engineer at Googlezon where you are working on an internal ML framework called TorchsorFlow. You are tasked with implementing a new layer known as BatchNormalization. The idea of this layer is as follows:

During training, consider the outputs of the previous layer $\mathbf{a}_i = (a_i^{(1)}, \dots, a_i^{(N)})$ for each element $i \in \{1, \dots, M\}$ of the input batch. Compute the mean μ_j and variance σ_j^2 over each input dimension j. Use the resulting statistics to normalize the output of the previous layer. Finally, rescale the resulting vector with a learned constant γ and shift it by another learned constant β.

(a) Write down the mathematical expression for the BatchNormalization layer. What are its learnable parameters?

(b) Compute the gradient of the loss $\mathcal{L}$ with respect to the input of the BatchNormalization $\mathbf{a}_i$ layer.

(c) At test time, the batch size is usually 1. So, it is not meaningful (or even possible) to compute mean / variance. How would you implement a layer like this?

Exercise 3 - Autodiff Modes

Consider the function $F(x) = f_3(f_2(f_1(x))$ and assume you also know the derivatives f_i' for all f_i.

(a) Apply the chain rule to express $F'(x)$ in terms of f_i's and f_i.

(b) Write down the pseudocode for computing $F'(x)$ using the forward mode and the reverse mode respectively. Assume all functions to be scalar functions of a scalar variable, i.e. $f_i : \mathbb{R} \to \mathbb{R}$.

(c) If you simply ask your interpreter / compiler to evaluate the expression in (a), will the computation be in forward mode, reverse mode, or neither of the modes? Why? You can assume that your interpreter / compiler does not do any caching or optimization and simply evaluates the expression from left to right. Does anything change if you assume that your interpreter caches results that have been computed before?

Exercise 4 - GloVe Embeddings

Open the notebook presented in class and work through it by trying some of the ideas presented therein for different word combinations.