Foundations of Deep Learning **Fall 2024**

Exercise 0: Prerequisites

Lecturer: Aurelien Lucchi

Problem 1 (Operator Norm):

Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ be a rectangular matrix with $m \geq n$.

a) Recall the definition of the operator norm of $\mathbf{A}$:

$$\|\mathbf{A}\|_2 = \sup_{\mathbf{x}\in\mathbb{R}^n\setminus\{0\}} \frac{\|\mathbf{A}\mathbf{x}\|}{\|\mathbf{x}\|}.$$

Write down the definition of a vector norm. Show that the operator norm is indeed a vector norm in the vector space $\mathbb{R}^{m\times n}$ of matrices.

b) By the singular value decomposition (SVD) of $\mathbf{A}$, show that $\|\mathbf{A}\|_2 = s_1(\mathbf{A})$, the largest singular value of $\mathbf{A}$.

c) By SVD of $\mathbf{A}$ again, show that $s_1(\mathbf{A}) = \sqrt{\lambda_1(\mathbf{A}^\top\mathbf{A})}$, where $\lambda_1(\cdot)$ denotes the largest eigenvalue.

Problem 2 (Calculus):

a) Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ be a rectangular matrix, $\mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^m$ be column vectors. Let $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2$. Compute its gradient $\nabla f \in \mathbb{R}^n$.

b) Let $\mathbf{U} \in \mathbb{R}^{n\times p}, \mathbf{V} \in \mathbb{R}^{p\times m}$, and $\mathbf{R} \in \mathbb{R}^{n\times m}$. The matrix factorization problem tries to find an approximation of $\mathbf{R}$ as the product of two matrices with smaller common dimension p, that is

$$\mathbf{R} \approx \mathbf{U}\mathbf{V}.$$

This problem can be solved for instance by minimizing the loss $L(\mathbf{U},\mathbf{V}) := \frac{1}{2}\|\mathbf{U}\mathbf{V} - \mathbf{R}\|_F^2$, where $\|\mathbf{A}\|_F := \sqrt{\sum_i \sum_j |A_{ij}|^2}$ is the Frobenius norm.
Compute the derivative of L w.r.t. $\mathbf{U}$, $\frac{\partial L}{\partial \mathbf{U}}$, and w.r.t. $\mathbf{V}$, $\frac{\partial L}{\partial \mathbf{V}}$, respectively.

c) Consider the problem of non-linear least squares regression with some non-linear function $\ell : \mathbb{R} \to \mathbb{R}$ and n data samples $(\mathbf{x}_i, y_i)$,

$$L(\mathbf{w}) = \frac{1}{2n}\sum_{i=1}^n \left(y_i - \ell(\mathbf{x}_i^\top \mathbf{w})\right)^2.$$

Compute the gradient $\nabla_\mathbf{w} L(\mathbf{w})$ and the Hessian $\nabla^2_\mathbf{w} L(\mathbf{w})$.

Problem 3 (Taylor Expansion):

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice differentiable function with a local minimum $\mathbf{x}^*$.

a) Write down the definition of the Hessian $\mathbf{H}(\mathbf{x}) = \nabla^2 f(\mathbf{x}) \in \mathbb{R}^{n\times n}$ of f and the order-2 Taylor expansion of f at $\mathbf{x}^*$.

b) Using Chain rule or otherwise, prove that for any $\mathbf{x}, \mathbf{v} \in \mathbb{R}^n$, the matrix-vector product $\mathbf{H}(\mathbf{x})\mathbf{v} \in \mathbb{R}^n$ is equal to a limit:

$$\mathbf{H}(\mathbf{x})\mathbf{v} = \lim_{t\to 0} \frac{\nabla f(\mathbf{x} + t\mathbf{v}) - \nabla f(\mathbf{x})}{t}.$$

c) Suppose f is a L-smooth function, that is, there is an $L > 0$ such that

$$\|\nabla f(\mathbf{x}_1) - \nabla f(\mathbf{x}_2)\| \leq L\|\mathbf{x}_1 - \mathbf{x}_2\|, \forall \mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n.$$

Show that each eigenvalue of $\mathbf{H}(\mathbf{x})$ of any $\mathbf{x} \in \mathbb{R}^n$ is upper bounded by L.

d) Recall that $\mathbf{x}^*$ is a local minimum. Using Problem 1c) and 3c) or otherwise, prove that

$$\|\mathbf{H}(\mathbf{x}^*)\|_2 \leq L$$

e) Using Cauchy-Schwartz inequality and Problem 3d) or otherwise, show that

$$f(\mathbf{x}) - f(\mathbf{x}^*) \leq \frac{L}{2} \|\mathbf{x} - \mathbf{x}^*\|^2 + o\Big(\|\mathbf{x} - \mathbf{x}^*\|^3 \Big).$$

Problem 4 (Probability Theory):

a) Use the definition of the expectation

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx$$

and variance $\mathrm{Var}(X) := \mathbb{E}\left[(X - \mathbb{E}[X])^2 \right]$ to verify that the expectation and variance of a normal distributed random variable $X \sim \mathcal{N}(\mu, \sigma)$ with probability density function

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

is indeed $\mathbb{E}[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$.

b) Similarly, using

$$\mathbb{E}[\mathbf{X}] = \int_{-\infty}^{\infty} \mathbf{x} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}$$

and $\mathrm{Cov}(\mathbf{X}) := \mathbb{E}\left[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top \right]$, verify that for a multivariate normal distributed random variable $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\mu} \in \mathbb{R}^k$, $\boldsymbol{\Sigma} \in \mathbb{R}^{k \times k}$ and probability density function

$$f_{\mathbf{X}}(\mathbf{x}) = (2\pi)^{-\frac{k}{2}} \det(\boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right)$$

that $\mathbb{E}[\mathbf{X}] = \mu$ and $\mathrm{Cov}(\mathbf{X}) = \boldsymbol{\Sigma}$.

c) Show the affine transformation rule for Gaussian random variables. That is, let $\mathbf{X}$ be normally distributed with $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} \in \mathbb{R}^k$, $\boldsymbol{\Sigma} \in \mathbb{R}^{k \times k}$ and define an affine transformed random variable $\mathbf{Y} := \mathbf{A}\mathbf{X} + \mathbf{b}$ with $\mathbf{A} \in \mathbb{R}^{p \times k}, \mathbf{b} \in \mathbb{R}^p$. Show that $\mathbf{Y}$ is normally distributed with $\mathbb{E}[\mathbf{Y}] = \mathbf{A}\boldsymbol{\mu} + \mathbf{b}$ and $\mathrm{Cov}(\mathbf{Y}) = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T$.