In class exercises for CW 44

1 Review Calculus

Definition 1 (Hessian)**.** The **Hessian** of a **scalar** function $f\colon \mathbb{R}^n \to \mathbb{R}$ is defined as:

$$\mathrm{H}f = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}\right)_{ij} \tag{1}$$

Fact: If the second partial derivatives $\frac{\partial^2 f}{\partial x_i \partial x_j}$ are all continuous, then $\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$ for all i, j. In this case the Hessian is **symmetric**. We will always assume that this is the cfase if not explicitly stated otherwise.

Remark 2. Note that the Hessian is equal to the gradient of the gradient:

$$\frac{\partial}{\partial x}\left(\frac{\partial}{\partial x} f\right) = \frac{\partial}{\partial x}\begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix} = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdot & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2 \partial x_2} & \cdot & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdot & \frac{\partial^2 f}{\partial x_n \partial x_n} \end{pmatrix} = \mathrm{H}f$$

Theorem 3 (Taylor's theorem)**.** *If $f\colon \mathbb{R}^n \to \mathbb{R}$ is two times continuously differentiable then*

$$\boxed{f(x) \approx f(x^*) + \nabla f[x^*]^\intercal (x - x^*) + \frac{1}{2}(x - x^*)^\intercal \mathrm{H}f[x^*](x - x^*)} \tag{2}$$

*for $x \approx x^*$. (In fact it is an **asymptotic** relationship, i.e. the approximation becomes better the closer x is to x^*)*

Exercise 4. Compute the second order Taylor approx. of $\exp\bigl(-\frac{1}{2}(x_1^2 + x_2^2)\bigr)$ at $x^* = 0$

Exercise 5. Compute the second order Taylor approx. of $\frac{1}{2}x^\intercal A x + b^\intercal x + c$ at $x_0 = 0$.

Definition 6 (symmetric part)**.** Any square matrix $A \in \mathbb{R}^{n \times n}$ can be decomposed **uniquely** into the sum of a symmetric ($A_+^\intercal = A_+$) and an anti-symmetric matrix ($A_-^\intercal = -A_-$).

$$A = A_+ + A_- \qquad A_+ = \frac{1}{2}(A + A^\intercal) \qquad A_- = \frac{1}{2}(A - A^\intercal)$$

A_+ and A_- are called the symmetric and anti-symmetric part of A.

Exercise 7. Given a square matrix A show that for all x holds $x^\intercal A x = x^\intercal A_+ x$. What does this mean for the result of Exercise 5?

Exercise 8. Show that the decomposition is indeed unique, i.e. if $A = B + C$ where B is symmetric and C is anti-symmetric, then $B = A_+$ and $C = A_-$.

Definition 9 (positive/negative definite)**.** For a square matrix $A \in \mathbb{R}^{n\times n}$ define

$$\begin{aligned}
&A \text{ is pos. def.} && (A > 0) \overset{\text{def}}{\iff} x^\intercal A x > 0 \text{ for all } x \iff \text{all EV of } A_+ \text{ are } > 0\\
&A \text{ is neg. def.} && (A < 0) \overset{\text{def}}{\iff} x^\intercal A x < 0 \text{ for all } x \iff \text{all EV of } A_+ \text{ are } < 0\\
&A \text{ is pos. semi-def.} && (A \geq 0) \overset{\text{def}}{\iff} x^\intercal A x \geq 0 \text{ for all } x \iff \text{all EV of } A_+ \text{ are } \geq 0\\
&A \text{ is neg. semi-def.} && (A \leq 0) \overset{\text{def}}{\iff} x^\intercal A x \leq 0 \text{ for all } x \iff \text{all EV of } A_+ \text{ are } \leq 0
\end{aligned}$$

Exercise 10. For which $\alpha \in \mathbb{R}$ is $A = \begin{bmatrix} 1 & \alpha \\ 0 & 1 \end{bmatrix}$ positive definite ?

2 Review Optimization

In machine learning we often want to fit a model to a given dataset. It is therefore important to study non-linear optimization problems. For example for given data x, y we may want to find good parameters θ such a model $\hat{y}(x) = f(x, \theta)$ fits the data well. This leads to an optimization problem.

$$\hat{\theta} = \operatorname*{argmin}_{\theta} \ell(y, \hat{y}) \tag{3}$$

where ℓ, the so called **loss**-function is a measure of how good the model fits the data. Oftentimes a model comes with additional restrictions, for example a parameter might be restricted to certain values or forbidden to become negative. A very general mathematical framework to deal with such problems is called **non-linear programming** which deals with the following optimization problem:

$$\min_x f(x) \quad \text{such that} \quad g(x) = 0 \text{ and } h(x) \geq 0 \tag{4}$$

We will now discuss some fundamental terminology related to this problem.

Theorem 11 (first order necessary condition)**.** *If $f\colon \mathbb{R}^n \to \mathbb{R}$ is cont. diff. then*

$$\boxed{x^* \textit{ is a local extremum} \implies \nabla f(x^*) = 0}$$

Exercise 12.

- Find all local extrema of $x^4 - 2x^2 + 3$
- Show that the converse (" $\Longleftarrow$ ") does not hold in general by giving an example.

Theorem 13 (second order necessary condition)**.** *If $f\colon \mathbb{R}^n \to \mathbb{R}$ is twice cont. diff. then*

$$\begin{aligned}
x \text{ is a local minimum} &\implies \nabla f[x] = 0 \text{ and } \mathrm{H}f[x] \geq 0\\
x \text{ is a local maximum} &\implies \nabla f[x] = 0 \text{ and } \mathrm{H}f[x] \leq 0
\end{aligned}$$

Theorem 14 (second order sufficient condition)**.** *The reverse of Theorem 13 holds if $\mathrm{H}f$ is strictly pos./neg. definite:*

$$\begin{aligned}
\nabla f[x] = 0 \text{ and } \mathrm{H}f[x] > 0 &\implies x \text{ is a local minimum}\\
\nabla f[x] = 0 \text{ and } \mathrm{H}f[x] < 0 &\implies x \text{ is a local maximum}
\end{aligned}$$

Exercise 15.

- Show that $f(x) = e^x - x$ has a local minimum at $x = 0$
- Show that the converse (" $\Longleftarrow$ ") does not hold in general by giving an example.

Definition 16 (convex function)**.** A continuous function $f\colon \mathbb{R}^n \to \mathbb{R}$ is called:

$$\begin{aligned}
&f \text{ convex} &&\overset{\text{def}}{\Longleftrightarrow} f(tx+(1-t)y) \le tf(x)+(1-t)f(y) \text{ for all } t\in(0,1) \text{ and } x,y\\
&f \text{ strictly convex} &&\overset{\text{def}}{\Longleftrightarrow} f(tx+(1-t)y) < tf(x)+(1-t)f(y) \text{ for all } t\in(0,1) \text{ and } x,y\\
&f \text{ concave} &&\overset{\text{def}}{\Longleftrightarrow} f(tx+(1-t)y) \ge tf(x)+(1-t)f(y) \text{ for all } t\in(0,1) \text{ and } x,y\\
&f \text{ strictly concave} &&\overset{\text{def}}{\Longleftrightarrow} f(tx+(1-t)y) > tf(x)+(1-t)f(y) \text{ for all } t\in(0,1) \text{ and } x,y
\end{aligned}$$

Theorem 17 (convexity criterion)**.** *If $f\colon \mathbb{R}^n \to \mathbb{R}$ is one/two times cont. diff. then*

$$\begin{aligned}
&f \text{ convex} &&\Longleftrightarrow f(x) \ge f(y) + \nabla f[y]^\intercal (x-y) \text{ for all } x,y &&\Longleftrightarrow \mathrm{H}f[x] \ge 0 \text{ for all } x\\
&f \text{ strictly convex} &&\Longleftrightarrow f(x) > f(y) + \nabla f[y]^\intercal (x-y) \text{ for all } x,y &&\Longleftrightarrow \mathrm{H}f[x] > 0 \text{ for all } x\\
&f \text{ concave} &&\Longleftrightarrow f(x) \le f(y) + \nabla f[y]^\intercal (x-y) \text{ for all } x,y &&\Longleftrightarrow \mathrm{H}f[x] \le 0 \text{ for all } x\\
&f \text{ strictly concave} &&\Longleftrightarrow f(x) < f(y) + \nabla f[y]^\intercal (x-y) \text{ for all } x,y &&\Longleftrightarrow \mathrm{H}f[x] < 0 \text{ for all } x
\end{aligned}$$

Remark 18. To give some intuition on what the statements in Theorem 17 mean:

1. If one takes two points x, y and draws the straight line connecting $(x, f(x))$ and $(y, f(y))$, then the graph of of f is always below/above that line.
2. The graph of f is always above/below any of its tangent lines (or planes/hyperplanes in the multidimensional setting).
3. The graph of f always 'curves' upwards/downwards in any direction.

Theorem 19. *If f is strictly convex/concave then any local min/max is a global min/max.*

Exercise 20.

- Show that $x^\intercal A x + b^\intercal x + c$ is strictly convex if and only if $A_+ > 0$.
- Show that the sum of two convex functions is convex
- Let $f, g\colon \mathbb{R} \to \mathbb{R}$ be two times cont. diff. Show that if f, g are convex and f is non-decreasing then $f \circ g$ is also convex

Theorem 21 (Lagrange multiplier)**.** *Consider the constrained optimization problem*

$$\max_x f(x) \quad \textit{such that} \quad g(x) = 0 \tag{5}$$

Then if x^* is an optimal value, there exists a λ^* such that (x^*, λ^*) is a stationary point of the Lagrangian

$$\mathcal{L}(x, \lambda) = f(x) - \lambda g(x) \tag{6}$$

(Note: A stationary point is a point at which the gradient is zero.)

Exercise 22. Show that the optimal value of the constrained problem

$$\max_x \|Ax\|_2^2 \quad \text{such that} \quad \|x\|_2^2 = 1 \tag{7}$$

is obtained when x is an eigenvector corresponding to the largest eigenvalue of $A^\intercal A$.