

Sheet 2

For the exercise class on the 15.10.2026 at 12:00.

Exercises from the in-class test

Exercise 1 (Strict & Strong Convexity).

(0 Points)

Prove the following statements

- (i) μ -strong convexity implies strict convexity.
- (ii) For twice differentiable f , the following are equivalent
 - (a) $\nabla^2 f(x) \succeq \mu \mathbb{I}$
 - (b) $z^T \nabla^2 f(x) z \geq \mu \|z\|^2$
 - (c) f is μ -strongly convex

where $\mathbb{I}$ is the identity matrix and

$$A \succeq B : \iff A - B \text{ is (weakly) positive definite.}$$

Exercise 2 (Directional Minima).

(0 Points)

Let f be some differentiable function. For every direction $d \in \mathbb{R}^n$ define

$$g_d(\alpha) := f(x^* + \alpha d).$$

Assume that for every d , g_d is minimized by $\alpha = 0$. Prove that

- (i) We have necessarily $\nabla f(x^*) = 0$.
- (ii) $f(x^*)$ is *not* necessarily a minimum of f .

Hint. Let $0 < p < q$ and define

$$f(y, z) := (z - py^2)(z - qy^2)$$

consider $x^* = (0, 0)$ and prove that $f(y, my^2) < 0$ for $p < m < q$.

Exercise 3 (Convergence to Stationary Point).

(0 Points)

Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be a continuously differentiable function.

- (i) Let $(x_k)_{k \in \mathbb{N}}$ be defined by gradient descent

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k), \quad x_0 \in \mathbb{R}^d$$

with diminishing step size $\alpha_k > 0$ such that $\sum_{k=1}^{\infty} \alpha_k = \infty$. Suppose that $(x_k)_{k \in \mathbb{N}}$ converges to some $x_* \in \mathbb{R}^d$. Prove that x_* is a stationary point of f , i.e. $\nabla f(x_*) = 0$.

Hint. You might want to prove for large enough i, j

$$\langle \nabla f(x_i), \nabla f(x_j) \rangle \geq \|\nabla f(x_*)\|^2 - 2\epsilon \|\nabla f(x_*)\| - \epsilon^2 =: p(\epsilon).$$

- (ii) Assume that f is also L -smooth. Prove for x_n generated by gradient descent with constant step size $\alpha \in (0, \frac{2}{L})$ we have

$$\sum_{k=n}^m \|\nabla f(x_k)\|^2 \leq \frac{f(x_n) - f(x_m)}{\alpha(1 - \frac{L}{2}\alpha)} \leq \frac{f(x_n) - \min_x f(x)}{\alpha(1 - \frac{L}{2}\alpha)}$$

for any $n, m \in \mathbb{N}$. Deduce for the case $\min_x f(x) > -\infty$, that we have

$$\min_{k \leq n} \|\nabla f(x_k)\|^2 \in o(1/n).$$

Exercise 4 (Sub-gradients).

(0 Points)

Let $f, g : \mathbb{R}^d \rightarrow \mathbb{R}$ be convex functions.

- (i) Prove that $\partial f(x)$ is a convex set for any $x \in \mathbb{R}^d$.
(ii) For $h(x) = f(Ax + b)$ prove $\partial h(x) \supseteq A^T \partial f(Ax + b)$. Prove equality for invertible A .

Exercise 5 (PL-Inequality).

(0 Points)

Assume $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is L -smooth and satisfies the Polyak-Łojasiewicz inequality

$$\|\nabla f(x)\|^2 \geq 2c(f(x) - f_*) \quad (\text{PL})$$

for some $c > 0$ and all $x \in \mathbb{R}^d$ with $f_* = \min_x f(x) > -\infty$.

- (i) Prove that gradient descent with fixed step size $\alpha_k = \frac{1}{L}$ converges linearly in the sense

$$f(x_k) - f_* \leq (1 - \frac{c}{L})^k (f(x_0) - f_*).$$

- (ii) Prove that μ -strong-convexity and L -smoothness imply the PL-inequality.

Exercises to be prepared at home for the next exercise class at 05.11.2026

Exercise 6 (Conjugate Gradient Descent).

(0 Points)

Consider a quadratic function

$$f(x) = \frac{1}{2}(x - x_*)^T H(x - x_*)$$

for some symmetric and positive definite H and consider the hilbert space $\mathcal{H} = (\mathbb{R}^d, \langle \cdot, \cdot \rangle_H)$ with

$$\langle x, y \rangle_H = \langle x, Hy \rangle$$

- (i) Prove that $\langle \cdot, \cdot \rangle_H$ is a well-defined scalar product. Show that

$$f(x) = \frac{1}{2} \|x - x_*\|_H^2.$$

- (ii) Determine the derivative $\nabla_H f(x)$ of f in $\mathcal{H}$

Hint. Recall that $\nabla_H f(x)$ is the unique vector satisfying

$$0 = \lim_{v \rightarrow 0} \frac{|f(x+v) - f(x) - \langle \nabla_H f(x), v \rangle_H|}{\|v\|_H}.$$

- (iii) Since gradient descent in the space $\mathcal{H}$ is therefore computationally the Newton method, we want to find a different method of optimization. Consider an arbitrary set of conjugate (H -orthogonal) directions $(v_1, \dots, v_d)$, i.e. $\langle v_i, v_j \rangle_H = \delta_{ij}$, and for some starting point $x_0 \in \mathbb{R}^d$ the following descent algorithm:

$$x_{k+1} = x_k - \alpha_k v_{k+1} \quad \text{with} \quad \alpha_k := \underset{\alpha}{\operatorname{argmin}} f(x_k - \alpha v_{k+1}). \quad (\text{CD})$$

Optimizing over α in this manner is known as “line-search”. Using $y^{(i)} := \langle y, v_i \rangle$ prove that

$$(x_k - x_*) = \sum_{i=k+1}^d (x_0 - x_*)^{(i)} v_i = \underset{x}{\operatorname{argmin}} \{f(x) : x \in x_0 + \operatorname{span}[v_1, \dots, v_k]\} - x_*.$$

Deduce that conjugate descent (CD) converges in d steps.

- (iv) If we had $v_i = \nabla f(x_{i-1})$, then this algorithm would be optimal in the set of algorithms we considered in the previous exercise. Unfortunately the gradients $\nabla f(x_{i-1})$ are generally not conjugate. So while we may select an arbitrary set of conjugate v_i , we cannot select the gradients directly.

Instead we are going to do the next best thing and inductively select v_{k+1} such that

$$\mathcal{K}_k := \operatorname{span}[\nabla f(x_0), \dots, \nabla f(x_k)] = \operatorname{span}[v_1, \dots, v_{k+1}]$$

using the Gram-Schmidt procedure to make v_{k+1} conjugate to $v_1, \dots, v_k$. Since Gram-Schmidt is still computationally too expensive for our tastes, inductively prove

$$\mathcal{K}_k = \operatorname{span}[H^1(x_0 - x_*), \dots, H^{k+1}(x_0 - x_*)]$$

assuming $\mathcal{K}_k$ is $(k+1)$ -dimensional. I.e. $\mathcal{K}_k$ is a “ H -Krylov subspace”.

- (v) Argue that $\nabla f(x_{k+1})$ is orthogonal to every vector in $\mathcal{K}_k$ and inductively deduce either

$$\nabla f(x_{k+1}) = 0$$

which implies $x_{k+1} = x_*$, or $\mathcal{K}_{k+1}$ has full rank. Deduce from the H -Krylov-subspace property, that $\nabla f(x_{k+1})$ is already H -orthogonal to $\mathcal{K}_{k-1}$.

Hint. $x_{k+1} = \underset{x}{\operatorname{argmin}} \{f(x) : x \in \mathcal{K}_k + x_0\}$.

- (vi) Collect the ideas we have gathered to prove the recursively defined

$$v_{k+1} = \nabla f(x_k) - \frac{\langle \nabla f(x_k), v_k \rangle_H}{\|v_k\|_H^2} v_k$$

are H -conjugate and have the same span as the gradients up to $\nabla f(x_k)$.

(vii) To make our procedure truly computable, we want to show

$$\frac{\langle \nabla f(x_k), v_k \rangle_H}{\|v_k\|_H^2} = -\frac{\|\nabla f(x_k)\|^2}{\|\nabla f(x_{k-1})\|^2}.$$

Hint. Proving

$$\nabla f(x_k) = \nabla f(x_{k-1}) - \alpha_{k-1} H v_k$$

should allow you to conclude $\langle \nabla f(x_k), v_k \rangle_h = -\frac{\|\nabla f(x_k)\|^2}{\alpha_{k-1}}$. Then it makes sense to calculate

$$\alpha_{k-1} = -\frac{\langle \nabla f(x_{k-1}), v_k \rangle}{\|v_k\|_H^2}$$

by solving its optimization problem. Finally you may want to consider $v_k = \nabla f(x_{k-1}) - \alpha v_{k-1}$ and $v_{k-1} \in \mathcal{K}_{k-2}$.

(viii) Summarize everything into a pseudo-algorithm for conjugate gradient descent (CGD) and compare it to heavy-ball momentum with

$$\beta_k = \frac{\alpha_k \|\nabla f(x_k)\|^2}{\alpha_{k-1} \|\nabla f(x_{k-1})\|^2}$$

using identical α_k as CGD.