

Solution Sheet 1

For the exercise class on the 01.10.2026 at 12:00.

Exercises to be solved in the exercise class

Exercise 1 (Convex Examples).

(0 Points)

Prove the following functions are convex

- (i) affine linear functions, i.e. $f(x) = a^T x + c$ for $a \in \mathbb{R}^d$, $c \in \mathbb{R}$,

Solution. We have

$$\begin{aligned} f(\lambda x + (1 - \lambda)y) &= a^T(\lambda x + (1 - \lambda)y) + \overbrace{(\lambda + 1 - \lambda)}^1 c \\ &\stackrel{\text{linear}}{=} \lambda \underbrace{(a^T x + c)}_{f(x)} + (1 - \lambda) \underbrace{(a^T y + c)}_{f(y)}. \end{aligned} \quad \square$$

- (ii) norms, i.e. $x \mapsto \|x\|$,

Solution. We have

$$\|\lambda x + (1 - \lambda)y\| \stackrel{\Delta}{\leq} \|\lambda x\| + \|(1 - \lambda)y\| \stackrel{\text{scaling}}{=} \lambda \|x\| + (1 - \lambda) \|y\|. \quad \square$$

- (iii) sums of convex functions f_k , i.e. $f(x) = \sum_{k=1}^n f_k(x)$,

Solution. We have

$$f(\lambda x + (1 - \lambda)y) = \sum_{k=1}^n \underbrace{f_k(\lambda x + (1 - \lambda)y)}_{\leq \lambda f_k(x) + (1 - \lambda)f_k(y)} \leq \lambda \underbrace{\sum_{k=1}^n f_k(x)}_{=f(x)} + (1 - \lambda) \underbrace{\sum_{k=1}^n f_k(y)}_{=f(y)}. \quad \square$$

- (iv) $F(x) := \sup_{f \in \mathcal{F}} f(x)$ for a set of convex functions $\mathcal{F}$.

Solution. We have

$$\begin{aligned} F(\lambda x + (1 - \lambda)y) &\leq \sup_{f \in \mathcal{F}} \lambda f(x) + (1 - \lambda)f(y) \leq \sup_{f \in \mathcal{F}} \lambda f(x) + \sup_{g \in \mathcal{F}} (1 - \lambda)g(y) \\ &= \lambda F(x) + (1 - \lambda)F(y). \end{aligned} \quad \square$$

Exercise 2 (Finite Jensen).**(0 Points)**

Let φ be convex, and $\sum_{i=1}^n \lambda_i = 1$ for $\lambda_i \geq 0$. Prove

$$\varphi\left(\sum_{i=1}^n \lambda_i x_i\right) \leq \sum_{i=1}^n \lambda_i \varphi(x_i)$$

and deduce $(\frac{1}{n} \sum_{i=1}^n x_i)^2 \leq \frac{1}{n} \sum_{i=1}^n x_i^2$.

Solution. For $n = 2$ this is simply the definition of convexity. Assume it holds for n . Then the induction step is given by

$$\begin{aligned} \varphi\left(\sum_{i=1}^{n+1} \lambda_i x_i\right) &= \varphi\left(\lambda_{n+1} x_{n+1} + (1 - \lambda_{n+1}) \sum_{i=1}^n \underbrace{\frac{\lambda_i}{1 - \lambda_{n+1}}}_{=: \tilde{\lambda}_i} x_i\right) \\ &\leq \lambda_{n+1} \varphi(x_{n+1}) + (1 - \lambda_{n+1}) \varphi\left(\sum_{i=1}^n \tilde{\lambda}_i x_i\right) \\ &\leq \sum_{i=1}^{n+1} \lambda_i \varphi(x_i) \quad \underbrace{\qquad}_{\substack{\text{ind.} \\ \leq \sum_{i=1}^n \tilde{\lambda}_i \varphi(x_i)}} \end{aligned}$$

Where it is easy to check that $\tilde{\lambda}_i$ sums to one. For the second part select $\lambda_i = \frac{1}{n}$ and $\varphi(x) = x^2$, which is convex because

$$\begin{aligned} (\lambda x + (1 - \lambda)y)^2 &= \underbrace{\lambda^2}_{=\lambda(1-(1-\lambda))} x^2 + \underbrace{(1 - \lambda)^2}_{=(1-\lambda)-\lambda(1-\lambda)} y^2 + 2\lambda(1 - \lambda)xy \\ &= \lambda x^2 + (1 - \lambda)y^2 - \lambda(1 - \lambda) \underbrace{(x^2 + y^2 - 2xy)}_{=(x-y)^2 \geq 0} \\ &\leq \lambda x^2 + (1 - \lambda)y^2. \end{aligned}$$

□

Exercise 3 (Convexity and Minima).**(0 Points)**

Prove the following statements

- (i) If f is convex, then every local minimum is also a global minimum.

Solution. Let x^* be a local minimum and assume it was not global. Then there exists y such that

$$f(y) < f(x^*).$$

Select λ small enough such that $x^* + \lambda(y - x^*)$ is still in the ϵ neighborhood of x^* where x^* is a local minimum. Then obtain a contradiction

$$f(x^*) \leq f(x^* + \lambda(y - x^*)) \stackrel{\text{conv.}}{\leq} (1 - \lambda)f(x^*) + \underbrace{\lambda}_{>0} \underbrace{f(y)}_{< f(x^*)} < f(x^*). \quad \square$$

- (ii) If f is strictly convex, then there exists at most one minimum.

Solution. Let $x \neq y$ both be minima (which are global by (i)). Then for some $\lambda \in (0, 1)$

$$f(\lambda x + (1 - \lambda)y) \stackrel{\text{strict conv.}}{<} \lambda f(x) + (1 - \lambda)f(y) = \min_z f(z),$$

which is a contradiction. □

(iii) If f convex and differentiable and $\nabla f(x^*) = 0$, then x^* is a minimum.

Solution. By Proposition A.1.8 convexity is equivalent to

$$f(x) \geq f(x^*) + \underbrace{\langle \nabla f(x^*), x - x^* \rangle}_{=0} = f(x^*) \quad \forall x.$$

So x^* is a (global) minimum. □

Exercise 4 (Descent Directions of a Maximum).

(0 Points)

Let $x_* \in \mathbb{R}^d$ be a strict local maximum of $f : \mathbb{R}^d \rightarrow \mathbb{R}$. Prove that every $d \in \mathbb{R}^d \setminus \{0\}$ is a descent direction of f in x_* .

Solution. Let $\epsilon > 0$ be such, that x_* is a strict maximum in $B_\epsilon(x_*) \setminus \{x_*\}$, where existence of such an ϵ is the strict local maximum property. We now have for any direction d that it is a descent direction, because with $\bar{\alpha} = \frac{\epsilon}{\|d\|} > 0$ we have for all $\alpha \in (0, \bar{\alpha}]$

$$f(x_* + \alpha d) < f(x_*),$$

since $x_* + \alpha d \in B_\epsilon(x_*) \setminus \{x_*\}$ as $\|\alpha d\| \leq \bar{\alpha}\|d\| = \epsilon$. □

Exercise 5 (Convergence Speed).

(0 Points)

Proof that

(i) if we have

$$\limsup_{k \rightarrow \infty} \frac{e(x_{k+1})}{e(x_k)} = 0,$$

then $e(x_k)$ converges super-linearly.

Solution. We define $c_n := \sup_{k \geq n} \frac{e(x_{k+1})}{e(x_k)}$. Then

$$\lim_{n \rightarrow \infty} c_n = \limsup_{k \rightarrow \infty} \frac{e(x_{k+1})}{e(x_k)} = 0$$

and by definition

$$e(x_{k+1}) \leq c_k e(x_k).$$

Thus we have super-linear convergence. □

(ii) If for $c \in (0, 1)$ we have

$$\limsup_{k \rightarrow \infty} \frac{e(x_{k+1})}{e(x_k)} < c,$$

then $e(x_k)$ converges linearly with rate c .

Solution. We again define $c_n := \sup_{k \geq n} \frac{e(x_{k+1})}{e(x_k)}$

$$\lim_{n \rightarrow \infty} c_n = \limsup_{k \rightarrow \infty} \frac{e(x_{k+1})}{e(x_k)} < c$$

thus there exists $N \geq 0$ such that for all $n \geq N$ we have $c_n \leq c$ and therefore for all $n \geq N$

$$e(x_{n+1}) \leq c_n e(x_n) \leq c e(x_n). \quad \square$$

(iii) If for $c \in (0, 1)$ we have

$$\limsup_{k \rightarrow \infty} \frac{e(x_{k+1})}{e(x_k)^2} < c,$$

then $e(x_k)$ converges super-linearly with rate c .

Solution. We similarly define $c_n := \sup_{k \geq n} \frac{e(x_{k+1})}{e(x_k)^2}$ and again get $\lim_{n \rightarrow \infty} c_n < c$. Thus there exists $N \geq 0$ such that for all $n \geq K$ we have $c_n \leq c$ and therefore for all $n \geq N$

$$e(x_{n+1}) \leq c_n e(x_n)^2 \leq c e(x_n)^2. \quad \square$$

Exercise 6 (Momentum Matrix).

(0 Points)

let $D = \text{diag}(\lambda_1, \dots, \lambda_d)$, $\alpha, \beta > 0$ and define

$$T = \begin{pmatrix} (1 + \beta)\mathbb{I} - \alpha D & -\beta\mathbb{I} \\ \mathbb{I} & \mathbf{0} \end{pmatrix} \in \mathbb{R}^{2d \times 2d}$$

Prove there exists a regular $S \in \mathbb{R}^{2d \times 2d}$ such that

$$S^{-1}TS = \hat{T} = \begin{pmatrix} T_1 & & \\ & \ddots & \\ & & T_d \end{pmatrix}$$

with

$$T_i = \begin{pmatrix} 1 + \beta - \alpha\lambda_i & -\beta \\ 1 & 0 \end{pmatrix} \in \mathbb{R}^{2 \times 2}.$$

Solution. We simply define for the standard basis $e_i \in \mathbb{R}^d$

$$S = \begin{pmatrix} e_1 & 0 & \dots & e_d & 0 \\ 0 & e_1 & \dots & 0 & e_d \end{pmatrix} \in \mathbb{R}^{2d \times 2d}$$

in particular $S^T = S^{-1}$. $\square$

Exercises to be prepared at home for the next exercise class at 15.10.2026

Exercise 7 (Bregman Divergence).

(0 Points)

The Bregman Divergence $D_f^{(B)}$ of a continuously differentiable function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is defined as the error of the linear approximation and is related to μ -strong convexity and Lipschitz continuous gradients as follows

$$\frac{\mu}{2} \|x - x_0\|^2 \stackrel{\substack{\mu\text{-strongly convex} \\ \text{(definition)}}}{\leq} \underbrace{f(x) - f(x_0) - \langle \nabla f(x_0), x - x_0 \rangle}_{=: D_f^{(B)}(x, x_0)} \stackrel{\substack{\text{linear approximation} \\ L\text{-Lipschitz gradient} \\ \text{(Descent Lemma)}}}{\leq} \frac{L}{2} \|x - x_0\|^2.$$

For $\mu = 0$ this is simply the convexity condition by Prop. A.1.8. So non-negativity of the Bregman divergence implies convexity. The L -Lipschitz gradients provide us with an upper bound on the Bregman divergence on the other hand which immediately results in an upper bound on f

$$f(x) = f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle + \underbrace{D_f^{(B)}(x, x_0)}_{\leq \frac{L}{2} \|x - x_0\|^2}. \quad (1)$$

(i) Prove for functions f with L -Lipschitz gradients, we have for all x_0

$$\min_x f(x) \leq f(x_0) - \frac{1}{2L} \|\nabla f(x_0)\|^2.$$

By minimizing the upper bound (1). What is the minimizer of the upper bound?

Solution. We first solve the directional minimization problem

$$\begin{aligned} \operatorname{argmin}_{x: \|x - x_0\| = r} f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle + \frac{L}{2} \|x - x_0\|^2 &= x_0 + \operatorname{argmin}_{d: \|d\| = r} \langle \nabla f(x_0), d \rangle \\ &= x_0 + \operatorname{argmax}_{d: \|d\| = r} \langle -\nabla f(x_0), d \rangle \\ &= x_0 - \frac{r \nabla f(x_0)}{\|\nabla f(x_0)\|}, \end{aligned} \quad (2)$$

where the last equation is true because by Cauchy-Schwartz

$$\langle -\nabla f(x_0), d \rangle \stackrel{C.S.}{\leq} \|\nabla f(x_0)\| r.$$

So in summary we have

$$\min_x f(x) \leq \min_r \underbrace{\min_{x: \|x - x_0\| = r} f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle + \frac{L}{2} \|x - x_0\|^2}_{\stackrel{(2)}{=} f(x_0) - r \|\nabla f(x_0)\| + \frac{L}{2} r^2}.$$

Minimizing over the length r implies minimizing a convex parabola, so the first order condition is sufficient yielding

$$r^* = \frac{\|\nabla f(x_0)\|}{L}.$$

Reinserting r^* into our upper bound yields the claim and we get the minimizer by inserting r^* into (2) resulting in a gradient descent step

$$x^* = x_0 - \frac{1}{L} \nabla f(x_0). \quad \square$$

(ii) Prove for convex functions f with L -Lipschitz gradients

$$D_f^{(B)}(x, x_0) \geq \frac{1}{2L} \|\nabla f(x) - \nabla f(x_0)\|^2.$$

Solution. To apply (i), we want to prove

$$\phi(x) := D_f^{(B)}(x, x_0) = f(x) - f(x_0) - \langle \nabla f(x_0), x - x_0 \rangle$$

has L -Lipschitz gradient. But the gradient of ϕ is given by

$$\nabla \phi(x) = \nabla f(x) - \nabla f(x_0),$$

which is merely a shift of the gradient of f and therefore still L -Lipschitz. Due to convexity we know the Bregman Divergence is greater than zero. So we have

$$0 = D_f^{(B)}(x_0, x_0) = \min_y \phi(y) \stackrel{(i)}{\leq} \phi(x) - \frac{1}{2L} \|\nabla \phi(x)\|^2.$$

Reordering and inserting the definition we obtain our claim

$$\frac{1}{2L} \|\nabla f(x) - \nabla f(x_0)\|^2 = \frac{1}{2L} \|\nabla \phi(x)\|^2 \leq \phi(x) = D_f^{(B)}(x, x_0). \quad \square$$

(iii) Prove for convex functions f with L -Lipschitz gradients, we have for all x, y

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{1}{L} \|\nabla f(x) - \nabla f(y)\|^2.$$

Solution. We simply apply (ii) twice:

$$\begin{aligned} \frac{1}{L} \|\nabla f(x) - \nabla f(y)\|^2 &\stackrel{2 \times (ii)}{\leq} D_f^{(B)}(y, x) + D_f^{(B)}(x, y) \\ &= \cancel{f(y)} - \cancel{f(x)} - \langle \nabla f(x), y - x \rangle + \cancel{f(x)} - \cancel{f(y)} - \langle \nabla f(y), x - y \rangle \\ &= \langle \nabla f(x) - \nabla f(y), x - y \rangle. \end{aligned} \quad \square$$

(iv) Prove for convex f that the upper bound

$$D_f^{(B)}(x, x_0) \leq \frac{L}{2} \|x - x_0\|^2$$

is sufficient for Lipschitz continuity of the gradient ∇f .

Solution. We deduced this upper bound in (i) and then never used the Lipschitz property again. In (ii) we only needed it to apply (i) and in (iii) we only needed the property to apply (ii). So our premise is sufficient for (iii) without Lipschitz continuity of the gradient. So by (iii) and Cauchy-Schwarz we get

$$\frac{1}{L} \|\nabla f(x) - \nabla f(y)\|^2 \leq \langle \nabla f(x) - \nabla f(y), x - y \rangle \stackrel{\text{C.S.}}{\leq} \|\nabla f(x) - \nabla f(y)\| \|x - y\|.$$

Multiplying this inequality by $L/\|\nabla f(x) - \nabla f(y)\|$ results in the claim. $\square$

(v) In this last part we want to assume f is μ -strongly convex (and its gradient L -Lipschitz). Prove for all x, y

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{\mu L}{L+\mu} \|x - y\|^2 + \frac{1}{L+\mu} \|\nabla f(x) - \nabla f(y)\|^2.$$

Solution. We define

$$g_x(y) := f(y) - \frac{\mu}{2} \|x - y\|^2.$$

for which we have

$$\nabla g_x(y) = \nabla f(y) - \mu(y - x).$$

Therefore we have

$$\begin{aligned} D_{g_x}^{(B)}(y, z) &= g_x(y) - g_x(z) - \langle \nabla g_x(z), y - z \rangle \\ &= f(y) - \frac{\mu}{2} \|x - y\|^2 - f(z) + \frac{\mu}{2} \|x - z\|^2 - \langle \nabla f(z) - \mu(z - x), y - z \rangle \\ &= D_f^{(B)}(y, z) + \frac{\mu}{2} \underbrace{\left[\|x - z\|^2 - \|x - y\|^2 + 2\langle z - x, y - z \rangle \right]}_{= -(\|x - z\|^2 + \|x - z + z - y\|^2 - 2\langle x - z, z - y \rangle)} \\ &= D_f^{(B)}(y, z) - \frac{\mu}{2} \|z - y\|^2. \end{aligned}$$

Convexity is a given because $D_{g_x}^{(B)} \geq 0$ since $D_f^{(B)}(y, z) \geq \frac{\mu}{2} \|y - z\|^2$. Similarly we obtain $L - \mu$ -Lipschitz continuity of $\nabla g_x(y)$ because

$$D_{g_x}^{(B)}(y, z) = D_f^{(B)}(y, z) - \frac{\mu}{2} \|z - y\|^2 \leq \frac{L - \mu}{2} \|z - y\|^2.$$

So we can apply (ii) to $D_{g_x}^{(B)}$ with $z = x$ to obtain

$$\begin{aligned} D_f^{(B)}(y, x) &= D_{g_x}^{(B)}(y, x) + \frac{\mu}{2} \|x - y\|^2 \\ &\stackrel{(ii)}{\geq} \frac{1}{2(L - \mu)} \|\nabla g_x(y) - \nabla g_x(x)\|^2 + \frac{\mu}{2} \|x - y\|^2 \\ &= \frac{1}{2(L - \mu)} \underbrace{\|\nabla f(y) - \mu(y - x) - \nabla f(x)\|^2}_{\|(\nabla f(y) - \mu y) - (\nabla f(x) - \mu x)\|^2} + \frac{\mu}{2} \|x - y\|^2. \end{aligned}$$

Which is symmetric in x and y so we can apply the same trick as in (iii) to get

$$\begin{aligned} \langle \nabla f(y) - \nabla f(x), y - x \rangle &= D_f^{(B)}(y, x) + D_f^{(B)}(y, x) \\ &\geq \frac{1}{(L - \mu)} \underbrace{\|\nabla f(y) - \mu(y - x) - \nabla f(x)\|^2}_{= \|\nabla f(y) - \nabla f(x)\|^2 - 2\mu \langle \nabla f(y) - \nabla f(x), y - x \rangle + \mu^2 \|y - x\|^2} + \mu \|x - y\|^2. \end{aligned}$$

Collecting the scalar product on the left results in

$$\underbrace{\left(1 + \frac{2\mu}{L - \mu}\right)}_{= \frac{L + \mu}{L - \mu}} \langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \frac{1}{(L - \mu)} \|\nabla f(y) - \nabla f(x)\|^2 + \left(\frac{\mu^2}{L - \mu} + \mu\right) \|x - y\|^2.$$

Multiplying both sides by $\frac{L - \mu}{L + \mu}$ we finally get

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \frac{1}{(L + \mu)} \|\nabla f(y) - \nabla f(x)\|^2 + \underbrace{\left(\frac{\mu^2 + \mu(L - \mu)}{L + \mu}\right)}_{= \frac{\mu L}{L + \mu}} \|x - y\|^2. \quad \square$$

Exercise 8 (Optimizing Quadratic Functions).**(0 Points)**

In this exercise we consider functions of type

$$f(x) = x^T A x + b^T x + c,$$

where $x \in \mathbb{R}^d$, $A \in \mathbb{R}^{d \times d}$, $b \in \mathbb{R}^d$, $c \in \mathbb{R}$.

(i) Let $H := A^T + A$ be invertible. Prove that f can be written in the forms

$$f(x) = (x - x_*)^T A (x - x_*) + \tilde{c} \quad (3)$$

$$= \frac{1}{2}(x - x_*)^T \underbrace{(A^T + A)}_{=:H} (x - x_*) + \tilde{c} \quad (4)$$

for some $x_* \in \mathbb{R}^d$ and $\tilde{c} \in \mathbb{R}$. Argue that H is always symmetric. Under which circumstances is x_* a minimum?

Solution. We want for some x_*

$$f(x) \stackrel{!}{=} (x - x_*)^T A (x - x_*) + \tilde{c} = x^T A x \underbrace{- x^T A x_* - x_*^T A x + x_*^T A x_*}_{=-x_*^T (A + A^T) x \stackrel{!}{=} b^T x} + \tilde{c} \stackrel{!}{=} c$$

So we simply select

$$x_* := -(A + A^T)^{-1} b \quad \text{and} \quad \tilde{c} := c - x_*^T A x_*.$$

This proves our first representation (3). For (4) we simply note

$$y^T A y = \langle y, A y \rangle \stackrel{\text{symm.}}{=} \langle A y, y \rangle = y^T A^T y.$$

Applying this to $y = x - x_*$ in (3) we are done.

Symmetry of H follows directly from its definition as $H_{ij} = A_{ji} + A_{ij} = H_{ji}$.

Now x_* is a minimum iff H is positive definite. If it is positive definite, then x_* is a minimum by $\nabla^2 f(x) = H$ and the lecture. If H is not positive definite, we need to show that there exists some x such that $f(x) \leq m$ for all $m \leq \tilde{c}$. Since H is not positive definite, there exists some y such that

$$y^T H y =: -\varepsilon < 0$$

define $x = x_* + y \sqrt{\frac{\tilde{c} - m}{\varepsilon}}$. Then

$$f(x) = \frac{\tilde{c} - m}{\varepsilon} y^T H y + \tilde{c} = m. \quad \square$$

(ii) Argue that the Newton Method (with step size $\alpha_n = 1$) applied to f would jump to x_* in one step and then stop moving.

Solution. Taking the derivative of (4) we get

$$\nabla f(x) = H(x - x_*) \tag{5}$$

So with $\nabla^2 f(x) = H$ and

$$x_* = x - H^{-1}H(x - x_*) = x - [\nabla^2 f(x)]^{-1}\nabla f(x),$$

we have that the Newton Method finds x_* in one step. By (5) we also get $\nabla f(x_*) = 0$ which stops the Newton method afterwards. $\square$

(iii) Let $V = (v_1, \dots, v_d)$ be an orthonormal basis such that

$$H = V \operatorname{diag}[\lambda_1, \dots, \lambda_d] V^T$$

with $0 < \lambda_1 \leq \dots \leq \lambda_d$ and write

$$y^{(i)} := \langle y, v_i \rangle.$$

Express $(x_n - x_*)^{(i)}$ in terms of $(x_0 - x_*)^{(i)}$, where x_n is given by the gradient descent recursion

$$x_{n+1} = x_n - h \nabla f(x_n).$$

For which step size h do all the components $(x_n - x_*)^{(i)}$ converge to zero? Which component has the slowest convergence speed? Find the optimal learning rate h^* and deduce for this learning rate

$$\|x_n - x_*\| \leq (1 - \frac{2}{1+\kappa})^n \|x_0 - x_*\|.$$

with the condition number $\kappa = \frac{\lambda_d}{\lambda_1}$.

Solution. Using the representation (5) of the gradient again and subtracting x_* from our recursion, we get

$$x_{n+1} - x_* = x_n - x_* - hH(x_n - x_*) = [\mathbb{I} - hH](x_n - x_*)$$

Therefore

$$\begin{aligned} (x_{n+1} - x_*)^{(i)} &= \langle [\mathbb{I} - hH](x_n - x_*), v_i \rangle \\ &\stackrel{H \text{ symmetric}}{=} \langle x_n - x_*, [\mathbb{I} - hH]v_i \rangle \\ &\stackrel{\text{eigenvec.}}{=} \langle x_n - x_*, (1 - h\lambda_i)v_i \rangle \\ &= (1 - h\lambda_i)(x_n - x_*)^{(i)} \\ &\stackrel{\text{induction}}{=} (1 - h\lambda_i)^{n+1}(x_0 - x_*)^{(i)}. \end{aligned}$$

For all components to converge we need $|1 - h\lambda_i| < 1$ for all i . Since $1 - h\lambda_i < 1$ is always given, because $h, \lambda_i > 0$, we only need $1 - h\lambda_i > -1$ or $\frac{2}{h} > \lambda_i$ for all i . Since the eigenvalues are sorted, this is equivalent to $\frac{2}{h} > \lambda_d$ or

$$h < \frac{2}{\lambda_d}.$$

Under this condition, all components converge. The component with the slowest convergence is given by

$$\max_i |1 - h\lambda_i| = \max_i \max_{\substack{\leq 1-h\lambda_1 \\ \leq -(1-h\lambda_d)}} \{1 - h\lambda_i, -(1 - h\lambda_i)\} = \max\{1 - h\lambda_1, -(1 - h\lambda_d)\}.$$

To minimize the slowest convergence, we want to take the derivative. The discontinuity is at

$$1 - h\lambda_1 = -(1 - h\lambda_d) \iff 2 = h(\lambda_1 + \lambda_d)$$

so

$$\frac{d}{dh} = \begin{cases} -\lambda_1 & h \leq \frac{2}{\lambda_1 + \lambda_d} \\ \lambda_d & h \geq \frac{2}{\lambda_1 + \lambda_d} \end{cases}$$

So the maximal convergence speed is achieved by $h^* = \frac{2}{\lambda_1 + \lambda_d}$ with

$$r(h^*) := \max_i |1 - h^*\lambda_i| = 1 - \frac{2\lambda_1}{\lambda_1 + \lambda_d} = 1 - \frac{2}{1 + \kappa}$$

where $\kappa = \frac{\lambda_d}{\lambda_1}$ is the condition number. Putting things together, we have

$$\begin{aligned} \|x_n - x_*\|^2 &= \left\| \sum_{i=1}^d (1 - h\lambda_i)^n (x_0 - x_*)^{(i)} v_i \right\|^2 \\ &\stackrel{\text{orthonormal}}{=} \sum_{i=1}^d \underbrace{(1 - h\lambda_i)^{2n}}_{\leq r(h^*)^{2n}} (x_0 - x_*)^{(i)2} = r(h^*)^{2n} \|x_0 - x_*\|^2 \end{aligned}$$

and therefore

$$\|x_n - x_*\| \leq r(h^*)^n \|x_0 - x_*\|. \quad \square$$

Exercise 9 (Lasso).

(0 Points)

Let

$$f(x) = \frac{1}{2} \|x - y\|^2 + \lambda \|x\|_1$$

for $x \in \mathbb{R}^d$ be the Lagrangian form of the least squares LASSO method.

(i) Compute a sub-gradient of f .

Solution. Using $\partial(g + \lambda h)(x) \supseteq \partial g(x) + \lambda \partial h(x)$, we only need to determine the subgradient of $g(x) := \frac{1}{2} \|x - y\|^2$ and

$$h(x) := \|x\|_1 = \sum_{i=1}^d |x_i|.$$

But $\nabla g(x) = x - y$ as g is differentiable. And since it is also convex, we have

$$\partial g(x) = \{\nabla g(x)\}.$$

Now the subgradient of $h_i(x) = |x_i|$ is given by $\text{sgn}(x_i)e_i$, where $\text{sgn}(0) \in [-1, 1]$ can be selected arbitrarily, because

$$\begin{aligned} h_i(x) + \langle \text{sgn}(x_i)e_i, y - x \rangle &= |x_i| + \text{sgn}(x_i)y_i - \underbrace{\text{sgn}(x_i)x_i}_{|x_i|} = \text{sgn}(x_i)y_i \\ &\stackrel{\text{sgn}(x_i) \in [-1, 1]}{\leq} |y_i| = h_i(y). \end{aligned}$$

So again

$$\partial h(x) \supseteq \sum_{i=1}^d \partial h_i(x) \ni (\text{sgn}(x_1), \dots, \text{sgn}(x_n))^T =: s(x).$$

So putting everything together we have

$$\partial f(x) \ni x - y + \lambda s(x). \quad \square$$

(ii) Prove that f is convex.

Solution. As its sets of sub-gradients is nowhere empty, it is convex. $\square$

(iii) Find a global minimum of f .

Solution. By the lecture it is sufficient to find a point x such that $0 \in \partial f(x)$. By the previous exercise we therefore want to solve

$$0 \stackrel{!}{=} x - y + \lambda s(x)$$

entry-wise this implies

$$\begin{aligned} x_i &\stackrel{!}{=} y_i - \lambda \text{sgn}(x_i) = \begin{cases} y_i + \lambda & x_i < 0 \\ y_i - \lambda[-1, 1] & x_i = 0 \\ y_i - \lambda & x_i > 0 \end{cases} \\ &= \begin{cases} y_i + \lambda & y_i + \lambda < 0 \\ 0 & y_i \in [-\lambda, \lambda] \\ y_i - \lambda & y_i - \lambda > 0 \end{cases} \\ &= \begin{cases} y_i + \lambda & y_i < -\lambda \\ 0 & y_i \in [-\lambda, \lambda] \\ y_i - \lambda & y_i > \lambda. \end{cases} \quad \square \end{aligned}$$

Exercise 10 (Weak PL-Inequality).

(0 Points)

Assume $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is L -smooth and satisfies the “weak PL inequality”

$$\|\nabla f(x)\| \geq 2c(f(x) - f_*)$$

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

- (i) Let $a_0 \in [0, \frac{1}{q}]$ for some $q > 0$ and assume for the sequence $(a_n)_{n \in \mathbb{N}}$ that it is positive and satisfies a diminishing contraction

$$0 \leq a_{n+1} \leq (1 - qa_n)a_n \quad \forall n \geq 0.$$

Prove the convergence rate

$$a_n \leq \frac{1}{nq + 1/a_0} \leq \frac{1}{(n+1)q}.$$

Solution. Divide the reordered contraction

$$a_n \geq a_{n+1} + qa_n^2$$

by $a_n a_{n+1}$ to obtain

$$\frac{1}{a_{n+1}} \geq \frac{1}{a_n} + q \underbrace{\frac{a_n}{a_{n+1}}}_{\geq 1} \geq \frac{1}{a_n} + q$$

which leads to

$$\frac{1}{a_n} - \frac{1}{a_0} = \sum_{k=0}^{n-1} \frac{1}{a_{k+1}} - \frac{1}{a_k} \geq nq.$$

Reordering we obtain our claim

$$a_n \leq \frac{1}{nq + \frac{1}{a_0}} \stackrel{a_0 \leq \frac{1}{q}}{\leq} \frac{1}{(n+1)q}.$$

□

- (ii) Prove that f is bounded. More specifically $e(x) := f(x) - f_* \leq \frac{L}{2c^2}$ for all x .

Solution. Using Exercise 1 (i), we get

$$f_* \leq f(x) - \frac{1}{2L} \|\nabla f(x)\|^2$$

and therefore

$$e(x) \geq \frac{1}{2L} \|\nabla f(x)\|^2 \stackrel{\text{weak PL}}{\geq} \frac{4c^2}{2L} e(x)^2.$$

Dividing both sides by $e(x)$ we obtain

$$1 \geq \frac{2c^2}{L} e(x)$$

and thus

$$e(x) \leq \frac{L}{2c^2}.$$

□

- (iii) For gradient descent $x_{n+1} - x_n = -\alpha_n \nabla f(x_n)$ with constant step size $\alpha_k = \frac{1}{L}$ prove the convergence rate

$$f(x_n) - f_* \leq \frac{L}{2c^2(n+1)}.$$

Solution. Using L -smoothness, we have

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2} \|x_{k+1} - x_k\|^2 \\ &\leq f(x_k) - \underbrace{\alpha_k \left(1 - \frac{L}{2} \alpha_k\right)}_{=\frac{1}{2L}} \underbrace{\|\nabla f(x_k)\|^2}_{\geq 4c^2 e(x_k)^2} \end{aligned}$$

If we subtract f_* from both sides and apply our weak PL inequality we get

$$e(x_{k+1}) \leq e(x_k) - \frac{4c^2}{2L} e(x_k)^2 = \left(1 - \frac{2c^2}{L} e(x_k)\right) e(x_k)$$

with $q = \frac{2c^2}{L}$ and $e(x_0) \leq \frac{L}{2c^2} = \frac{1}{q}$ by (ii), we can apply (i) to obtain our claim. $\square$