

Solution Sheet 1

For the exercise class on the 19.09.2024 at 12:00.

Hand in your solutions by 10:15 in the lecture on Tuesday 17.09.2024.

Exercise 1 (Convex Examples).

(2 Points)

Prove the following functions are convex

- (i) affine linear functions, i.e. $f(x) = a^T x + c$ for $a \in \mathbb{R}^d$, $c \in \mathbb{R}$, (0.5 pts)

Solution. We have

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

- (ii) norms, i.e. $x \mapsto \|x\|$, (0.5 pts)

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. $f(x) = \sum_{k=1}^n f_k(x)$, (0.5 pts)

Solution. We have

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

- (iv) $F(x) := \sup_{f \in \mathcal{F}} f(x)$ for a set of convex functions \mathcal{F} . (0.5 pts)

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).**(2 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 $\left(\frac{1}{n} \sum_{i=1}^n x_i\right)^2 \leq \frac{1}{n} \sum_{i=1}^n x_i^2$.

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

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

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 (Strict & Strong Convexity).**(4 Points)**

Prove the following statements

(i) μ -strong convexity implies strict convexity.

(1 pt)

Solution. Assuming $x \neq y$ we have

$$0 < \frac{\mu}{2} \|x - y\|^2 \stackrel{\mu\text{-strong conv.}}{\leq} f(x) - f(y) - \langle \nabla f(y), x - y \rangle$$

But after moving the negative parts to the left we are left with a condition which is equivalent to strict convexity by Proposition A.1.8 in the lecture. □

(ii) For twice differentiable f , the following are equivalent

(3 pts)

- (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

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

Solution. Let us get (a) \iff (b) out of the way first:

$$z^T \nabla^2 f(x) z - \mu \|z\|^2 = \langle z, \nabla^2 f(x) z \rangle - \langle z, \mu \mathbb{I} z \rangle = z^T (\nabla^2 f(x) - \mu \mathbb{I}) z$$

If we have (a), then the rightmost side is positive and thus we also have (b). For the other direction we simply start from the left.

For (c) \implies (a) let

$$g(y) := f(y) - \frac{\mu}{2} \|y\|^2,$$

because f is μ strongly convex

$$\begin{aligned} g(y) &= f(y) - \frac{\mu}{2} \|y\|^2 \\ &\stackrel{f \text{ } \mu\text{-strong conv.}}{\geq} f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2 - \frac{\mu}{2} \|y\|^2 \\ &= \underbrace{f(x) - \frac{\mu}{2} \|x\|^2}_{=g(x)} + \frac{\mu}{2} \|x\|^2 + \underbrace{\langle \nabla f(x) - \mu x, y - x \rangle}_{\nabla g(x)} + \mu \langle x, y - x \rangle - \mu \langle x, y \rangle + \frac{\mu}{2} \|x\|^2 \\ &= g(x) + \langle \nabla g(x), y - x \rangle \end{aligned}$$

g is still convex. So we know by the lecture that

$$\nabla^2 g(x) = \nabla^2 f(x) - \mu \mathbb{I}$$

is positive definite.

On the other hand for (a) \implies (c): if $\nabla^2 f \succeq \mu \mathbb{I}$, we know that $\nabla^2 g$ is positive definite and thus that g is convex by the lecture. Reusing the equations above except for the inequality we obtain

$$\begin{aligned} f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2 - \frac{\mu}{2} \|y\|^2 &= g(x) + \langle \nabla g(x), y - x \rangle \\ &\stackrel{g \text{ conv.}}{\leq} g(y) \\ &= f(y) - \frac{\mu}{2} \|y\|^2 \end{aligned}$$

Adding $\frac{\mu}{2} \|y\|^2$ on both sides results in the definition of μ -strong convexity of f . \square

Exercise 4 (Convexity and Minima).

(3 Points)

Prove the following statements

- (i) If f is convex, then every local minimum is also a global minimum. (1 pt)

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. (1 pt)

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. \square

- (iii) If f convex and differentiable and $\nabla f(x^*) = 0$, then x^* is a minimum. (1 pt)

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. \square

Exercise 5 (Directional Minima).

(3 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 $\nabla f(x^*) = 0$. (1 pt)

Solution. If we select a standard basis vector as the direction, then

$$0 = g'_{e_i}(\alpha) \Big|_{\alpha=0} = \langle \nabla f(x^* + \alpha e_i), e_i \rangle \Big|_{\alpha=0} = \langle \nabla f(x^*), e_i \rangle$$

So the i -th entry of $\nabla f(x^*)$ is zero. As i was arbitrary, we are done. \square

- (ii) $f(x^*)$ is *not* necessarily a minimum of f . (2 pts)

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

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

then for $p < m < q$ we have for $y \neq 0$

$$f(y, my^2) = (my^2 - py^2)(my^2 - qy^2) = \underbrace{y^2}_{>0} \underbrace{(m-p)}_{>0} \underbrace{(m-q)}_{<0} < 0.$$

This means that $f(0, 0) = 0$ is not a minimum since we can get arbitrarily close to $x^* = (0, 0)$ with (y, my^2) with smaller and smaller y .

Let $d = (d_1, d_2)$ be an arbitrary direction. Then

$$\begin{aligned} g_d(\alpha) &= (\alpha d_2 - p(\alpha d_1)^2)(\alpha d_2 - q(\alpha d_1)^2) \\ &= \alpha^2(d_2 - p\alpha d_1^2)(d_2 - q\alpha d_1^2) \\ &= \alpha^2(d_2^2 - \alpha d_2(p + q)d_1^2 + \alpha^2 pq d_1^2) \end{aligned}$$

Thus $g'(0) = 0$ and $g''(0) = 2d_2^2 > 0$ which implies g has a minimum in zero. \square

Exercise 6 (Bregman Divergence).

(10 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? (2 pts)

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

(2 pts)

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

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

(1 pt)

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

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 .

(1 pt)

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 (4 pts)

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

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

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