1. Question: Convex Functions *(elementary)*

1.1. Which of the following functions are convex? (Hint: draw a picture)

(i) $f : \mathbb{R} \longrightarrow \mathbb{R},\ x \mapsto |x|$

(ii) $f : \mathbb{R} \longrightarrow \mathbb{R},\ x \mapsto \cos(x)$

(iii) $f : \mathbb{R} \longrightarrow \mathbb{R},\ x \mapsto x^2$

1.2. Prove that the following functions are convex.

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

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

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

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

Solution:

1.1. $|x|$ and x^2 are both convex. $\cos(x)$ is not convex since we can draw a line at two points (from say $\frac{\pi}{2}$ to $2\pi + \frac{\pi}{2}$) that is not above the function.

Proof that $|x|$ is convex:

$$\begin{aligned} f(\lambda x + (1-\lambda)y) &= |\lambda x + (1-\lambda)y| \\ &\leq \lambda |x| + (1-\lambda)|y| \end{aligned}$$

Proof that x^2 is convex: We begin by examining the inequality

$$\begin{aligned} (\lambda x + (1-\lambda)y)^2 &\leq \lambda x^2 + (1-\lambda)y^2 \\ \lambda^2 x^2 + 2\lambda(1-\lambda)xy + (1-\lambda)^2 y^2 &\leq \lambda x^2 + (1-\lambda)y^2 \\ \lambda^2 x^2 + 2\lambda(1-\lambda)xy + (1-\lambda)^2 y^2 - \lambda x^2 - (1-\lambda)y^2 &\leq 0 \\ \left(\lambda^2 - \lambda\right) x^2 - 2\left(\lambda^2 - \lambda\right) xy + \left(\lambda^2 - \lambda\right) y^2 &\leq 0 \\ \left(\lambda^2 - \lambda\right)\left(x^2 - 2xy + y^2\right) &\leq 0 \\ \left(\lambda^2 - \lambda\right)(x-y)^2 &\leq 0 \end{aligned}$$

Which holds when $\lambda \in [0,1]$, so the inequality is valid and our function is convex.

1.2. (i) We have

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

(ii) We have

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

(iii) We have

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

(iv) We have

$$F(\lambda x+(1-\lambda)y) \leq \sup_{f\in\mathcal{F}} \lambda f(x)+(1-\lambda)f(y) \leq \sup_{f\in\mathcal{F}} \lambda f(x)+\sup_{g\in\mathcal{F}}(1-\lambda)g(y) \\ = \lambda F(x)+(1-\lambda)F(y)$$

2. Question: Lipschitz Continuous Functions *(elementary)*

2.1. Which of the following functions are Lipschitz

(i) $f:[1,2]\to\mathbb{R},\ x\mapsto x^3$

(ii) $f:\mathbb{R}\longrightarrow\mathbb{R},\ x\mapsto x^2$

2.2. Prove the following for Lipschitz functions $f:\mathbb{R}\to\mathbb{R}$ and $g:\mathbb{R}\to\mathbb{R}$.

(i) The composition $f\circ g:\mathbb{R}\to\mathbb{R}$ is Lipschitz.

(ii) The sum $f+g:\mathbb{R}\to\mathbb{R}$ defined by $(f+g)(x)=f(x)+g(x)$ is Lipschitz.

2.3. Show that any Lipschitz function $f:[a,b]\to\mathbb{R}$ defined on an interval of the form $[a,b]$ is a bounded function.

2.4. Show that $h:[0,1]\to\mathbb{R}$ given by $h(x)=\sqrt{x}$, is bounded, but not Lipschitz.

Solution:

2.1. (i) Since $x^3-y^3=(x-y)\left(x^2+xy+y^2\right)$. It follows that

$$|f(x)-f(y)| \leq \left|x^2+xy+y^2\right|\cdot|x-y| \Rightarrow \\ |f(x)-f(y)| \leq \left(x^2+|xy|+y^2\right)\cdot|x-y| \Rightarrow \\ |f(x)-f(y)| \leq (4+4+4)\cdot|x-y|,$$

so f is Lipschitz with constant 12.

(ii) A function $f:\mathbb{R}\longrightarrow\mathbb{R}$ is NOT Lipschitz if for any $L>0$, we can find $x,y\in\mathbb{R}$ such that

$$|f(x)-f(y)|>L|x-y|.$$

In this specific case, take $x=2L$ and $y=L$. Then $|f(x)-f(y)|=3L^2>L^2=L|x-y|$. Therefore, x^2 is not Lipschitz.

2.2. (i) Let L_f and L_g be the Lipschitz constants of f and g respectively. That is $|f(x)-f(y)|\leq L_f|x-y|$ and $|g(x)-g(y)|\leq L_g|x-y|$ for any $x,y\in\mathbb{R}$. Then,

$$\begin{aligned}|(f\circ g)(x)-(f\circ g)(x)| &= |f(g(x))-f(g(y))| \\ &\leq L_f|g(x)-g(y)| \\ &\leq L_f L_g|x-y|,\end{aligned}$$

so $f\circ g$ is Lipschitz with constant $L=L_fL_g$.

(ii) With L_f and L_g as above,

$$\begin{aligned}|(f+g)(x)-(f+g)(y)| &= |f(x)+g(x)-f(y)-g(y)| \\ &= |f(x)-f(y)+g(x)-g(y)| \\ &\leq |f(x)-f(y)|+|g(x)-g(y)| \\ &\leq L_f|x-y|+L_g|x-y| \\ &= (L_f+L_g)\,|x-y|\end{aligned}$$

so $f+g$ is Lipschitz with constant $L=L_f+L_g$.

2.3. For any $x \in [a, b]$,

$$|f(x) - f(a)| \leq L|x - a| \leq L|b - a|,$$

and

$$\begin{aligned} |f(x)| &= |f(x) - f(a) + f(a)| \\ &\leq |f(x) - f(a)| + |f(a)| \\ &\leq L|b - a| + |f(a)|. \end{aligned}$$

so, taking $M = L|b - a| + |f(a)|$, $|f(x)| \leq M$ and, thereby f is bounded.

2.4. If $0 \leq x \leq 1$, then $0 \leq \sqrt{x} \leq 1$ and therefore $|h(x)| \leq 1$. For a constant $0 < L < 1$, take $x = 0$ and $y = 1$ and observe that

$$|h(x) - h(y)| = 1 > L = L|x - y|.$$

For a constant $L > 1$, take $x = 0$ and $y = \frac{1}{4L^2}$ and observe that

$$|h(x) - h(y)| = \frac{1}{2L} > \frac{1}{4L} = L|x - y|.$$

3. Question: Optimization *(elementary)*

Consider an optimization problem

$$\begin{aligned} &\min f(x) \\ &\text{s.t. } x \in \Omega. \end{aligned} \qquad (\star)$$

3.1. Prove that if $\Omega = \mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable, any point $\bar{x}$ that satisfies $\nabla f(\bar{x}) = 0$ is a global minimum.

3.2. Prove that if $f : \mathbb{R}^n \to \mathbb{R}$ is strictly convex on Ω and Ω is a convex set, the optimal solution (assuming it exists) must be unique.

3.3. Consider the optimization problem of $(\star)$ under the additional constraint that $Ax = b$, $A \in \mathbb{R}^{m \times n}$. Prove that if f is a convex function, a point $x \in \mathbb{R}^n$ is optimal to this constrained optimization problem if and only if it is feasible and $\exists \mu \in \mathbb{R}^m$ s.t.

$$\nabla f(x) = A^T \mu.$$

Hint: Start with what the first order condition for convexity tells us about the term $\nabla f^T(x)(y - x)$ for $y :\ Ay = b$ and use the fact that y with $Ay = b$ can be written as $y = x + v$, for $v \in \text{Nul}(A)$.

Solution:

3.1. From the first order characterization of convexity, we have

$$f(y) \geq f(x) + \nabla f^T(x)(y - x), \forall x, y$$

In particular,

$$f(y) \geq f(\bar{x}) + \nabla f^T(\bar{x})(y - x), \forall y$$

Since $\nabla f(\bar{x}) = 0$, we get

$$f(y) \geq f(\bar{x}), \forall y$$

Remarks:

- Recall that $\nabla f(x) = 0$ is always a necessary condition for local optimality in an unconstrained problem. The theorem states that for convex problems, $\nabla f(x) = 0$ is not only necessary, but also sufficient for local and global optimality.

- In absence of convexity, $\nabla f(x) = 0$ is not sufficient even for local optimality (e.g., think of $f(x) = x^3$ and $\bar{x} = 0$).
- Another necessary condition for (unconstrained) local optimality of a point x is that $\boldsymbol{H} f(x)$ is positive semi-definite. Note that a convex function automatically passes this test.

3.2. Suppose there were two optimal solutions $x, y \in \mathbb{R}^n$. This means that $x, y \in \Omega$ and

$$f(x) = f(y) \leq f(z), \forall z \in \Omega. \tag{$*$}$$

But consider $z = \frac{x+y}{2}$. By convexity of Ω, we have $z \in \Omega$. By strict convexity, we have

$$\begin{aligned} f(z) &= f\left(\frac{x+y}{2}\right) \\ &< \frac{1}{2} f(x) + \frac{1}{2} f(y) \\ &= \frac{1}{2} f(x) + \frac{1}{2} f(x) = f(x). \end{aligned}$$

But this contradicts $(*)$.

3.3. Since this is a convex problem, our optimality condition tells us that a feasible x is optimal iff

$$\nabla f^T(x)(y - x) \geq 0, \forall y \text{ s.t. } Ay = b$$

Any y with $Ay = b$ can be written as $y = x + v$, where v is a point in the nullspace of A; i.e., $Av = 0$. Given that $x + v - x = v$, a feasible x is optimal if and only if $\nabla f^T(x) v \geq 0, \forall v$ s.t. $Av = 0$. Since $Av = 0$ implies that $A(-v) = 0$, we also have that $\nabla f^T(x) v \leq 0$. Hence the optimality condition reads

$$\nabla f^T(x) v = 0 \quad \forall v \text{ s.t. } Av = 0.$$

This means that $\nabla f(x)$ is in the orthogonal complement of the nullspace of A which we know from linear algebra equals the row space of A (or equivalently the column space of A^T). Hence $\exists \mu \in \mathbb{R}^m$ s.t. $\nabla f(x) = A^T \mu$.

4. Question: Bregman Divergence
(advanced, to see what Lipschitz continuity can be used for)

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

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

For $\mu = 0$ this is simply the convexity condition. 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

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

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

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

by minimizing the upper bound (UB). What is the minimizer of the upper bound?
Hint: Try minimizing first w.r.t $x : \|x - x_0\| = r$ and then r. Additionally you will need the Cauchy-Schwartz inequality, whereby, for vectors $\mathbf{u}$ and $\mathbf{v}$: $|\langle \mathbf{u}, \mathbf{v}\rangle| \leq \|\mathbf{u}\|\|\mathbf{v}\|$.

Solution:

Solution. We first solve the directional minimization problem

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

where the last equation is true because by Cauchy-Schwartz

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

and

$$\left\langle-\nabla f\left(x_0\right),-\frac{r \nabla f\left(x_0\right)}{\left\|\nabla f\left(x_0\right)\right\|}\right\rangle=\frac{r}{\left\|\nabla f\left(x_0\right)\right\|}\left\langle\nabla f\left(x_0\right), \nabla f\left(x_0\right)\right\rangle=\frac{r}{\left\|\nabla f\left(x_0\right)\right\|}\left\|\nabla f\left(x_0\right)\right\|^2=\left\|\nabla f\left(x_0\right)\right\| r .$$

So, in summary, we have

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

Minimizing over the length r implies minimizing a convex parabola, so the first order condition is sufficient. So setting $f\left(x_0\right) - r\|\nabla f\left(x_0\right)\| + \frac{L}{2}r^2 \overset{!}{=} 0$ yields the minimizer

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

Reinserting r^* into our upper bound yields the claim and we get the minimizer by inserting r^* into (†):

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

If you have any questions or feedback, please feel free to contact me via E-mail at hannah.kuempel@stat.uni-muenchen.de!!

Also, thank you to Felix Benning & Prof. Dr. Simon Weißmann, Andy Hammerlindl, Kevin Jamieson & Anna Karlin, and A.A. Ahmadi whose exercises this sheet was inspired by.