Optimization in Machine Learning **Universität Mannheim**
FSS 2023 Prof. Simon Weißmann, Felix Benning

Solution Sheet 3

For the exercise class on the 30.03.2023.
Hand in your solutions by 12:00 in the exercise on Thursday 30.03.2023.

Exercise 1 (Convergence Speed)**.** **(3 Points)**

Proof that

(i) if we have

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

then $e(x_k)$ converges super-linearly. (1 pt)

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

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

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

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

then $e(x_k)$ converges linearly with rate c. (1 pt)

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

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

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

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

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

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

then $e(x_k)$ converges super-linearly with rate c. (1 pt)

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

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

Exercise 2 (Sub-gradients). **(4 Points)**

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

(i) Prove that $\partial f(x)$ is a convex set for any $x \in \mathbb{R}^d$. (1 pt)

Solution. Let $g_1, g_2 \in \partial f(x)$. Then for any $\lambda \in [0,1]$ and any $y \in \mathbb{R}^d$

$$\begin{aligned} f(y) &= \lambda f(y) + (1-\lambda) f(y) \\ &\overset{g_1,g_2 \in \partial f(x)}{\geq} \lambda\Big(f(x) + \langle g_1, y-x\rangle\Big) + (1-\lambda)\Big(f(x) + \langle g_2, y - x\rangle\Big) \\ &= f(x) + \Big\langle \lambda g_1 + (1-\lambda) g_2, y - x \Big\rangle \end{aligned}$$

thus $\lambda g_1 + (1-\lambda) g_2 \in \partial f(x)$ by definition. $\square$

(ii) Prove for $a > 0$, $\partial(af) = a\partial f$ (1 pt)

Solution. We only need to prove "$\supseteq$". Using $\tilde{f} = af$ with $\tilde{a} = \frac{1}{a}$ the other inclusion immediately follows.

Let $ag_x \in a\partial f(x)$ with $g_x \in \partial f(x)$. We need to show that $ag_x \in \partial(af)(x)$. But this follows immediately

$$\underbrace{a}_{>0} f(y) \overset{g_x \in \partial f(x)}{\geq} a\Big(f(x) + \langle g_x, y - x\rangle\Big) = (af)(x) + \langle a g_x, y - x\rangle. \qquad \square$$

(iii) Prove that $\partial(f_1 + f_2) \supseteq \partial f_1 + \partial f_2$ (1 pt)

Solution. Let $g_i \in \partial f_i(x)$ for $i = 1, 2$. Then we have that $g_1 + g_2 \in \partial(f_1 + f_2)$ because

$$\begin{aligned} (f_1 + f_2)(y) &\geq \Big(f_1(x) + \langle g_1, y - x\rangle\Big) + \Big(f_2(x) + \langle g_2, y - x\rangle\Big) \\ &= (f_1 + f_2)(x) + \langle g_1 + g_2, y - x\rangle. \end{aligned} \qquad \square$$

(iv) For $h(x) = f(Ax + b)$ prove $\partial h(x) \supseteq A^T \partial f(Ax + b)$. Prove equality for invertible A. (1 pt)

Solution. Let $g_x \in \partial f(x)$ i.e. $g_{Ax+b} \in \partial f(Ax + b)$. Then $A^T g_{Ax+b} \in \partial h(x)$ because

$$\begin{aligned} h(y) = f(Ay + b) &\geq f(Ax + b) + \langle g_{Ax+b}, (Ay + b) - (Ax + b)\rangle \\ &= h(x) + \langle A^T g_{Ax+b}, y - x\rangle. \end{aligned}$$

If A is invertible, we have $f(x) = h(A^{-1}x - A^{-1}b)$ so by the previous statement with $\tilde{A} = A^{-1}$ and $\tilde{b} = -A^{-1}b$, we get the other direction. $\square$

Exercise 3 (Lasso). **(6 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. (2 pts)

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 $\operatorname{sgn}(x_i)e_i$, where $\operatorname{sgn}(0) \in [-1,1]$ can be selected arbitrarily, because

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

So again

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

So putting everything together we have

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

(ii) Prove that f is convex. (1 pt)

Solution. As its sets of sub-gradients is nowhere empty, it is convex. □

(iii) Find a global minimum of f. (1 pt)

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

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

entry-wise this implies

$$x_i \overset{!}{=} y_i - \lambda\,\mathrm{sgn}(x_i) = \begin{cases} y_i + \lambda & x_i < 0 \\ y_i - \lambda[-1,1] & x_i = 0 \\ y_i - \lambda & x_i > 0 \end{cases}$$

$$= \begin{cases} y_i + \lambda & y_i + \lambda < 0 \\ 0 & y_i \in [-\lambda, \lambda] \\ y_i - \lambda & y_i - \lambda > 0 \end{cases}$$

$$= \begin{cases} y_i + \lambda & y_i < -\lambda \\ 0 & y_i \in [-\lambda, \lambda] \\ y_i - \lambda & y_i > \lambda. \end{cases}$$ □

(iv) Implement f as a sub-type of "DifferentiableFunction" (even though it is not) by returning a single sub-gradient and apply gradient descent to verify the global minimum `https://classroom.github.com/a/XqNuifmO` (2 pts).

Exercise 4 (Momentum Matrix). **(2 Points)**

let $D = \mathrm{diag}(\lambda_1, \dots, \lambda_d)$, $\alpha, \beta > 0$ 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 $S^T = S^{-1}$. □

Exercise 5 (PL-Inequality). **(5 Points)**

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

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

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

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

Solution. By L-smoothness and the descent lemma, we have

$$f(x_{k+1}) \leq f(x_k) - \tfrac{1}{2L}\|\nabla f(x_k)\|^2 \overset{\text{(PL)}}{\leq} f(x_k) - \tfrac{c}{L}(f(x_k) - f_*).$$

Subtracting f_* from both sides, we get

$$f(x_{k+1}) - f_* \leq (1 - \tfrac{c}{L})(f(x_k) - f_*) \qquad \square$$

(ii) Prove that μ-strong-convexity and L-smoothness imply the PL-inequality. (2 pts)

Solution. Recall by the solution of sheet 1, exercise 6 (iii), and strong convexity we have

$$\begin{aligned}\mu\|x-y\|^2 &\leq D_f^{(B)}(x,y) + D_f^{(B)}(y,x)\\ &= \langle \nabla f(x) - \nabla f(y), x - y\rangle\\ &\overset{\text{C.S.}}{\leq} \|\nabla f(x) - \nabla f(y)\|\|x-y\|\end{aligned}$$

and therefore

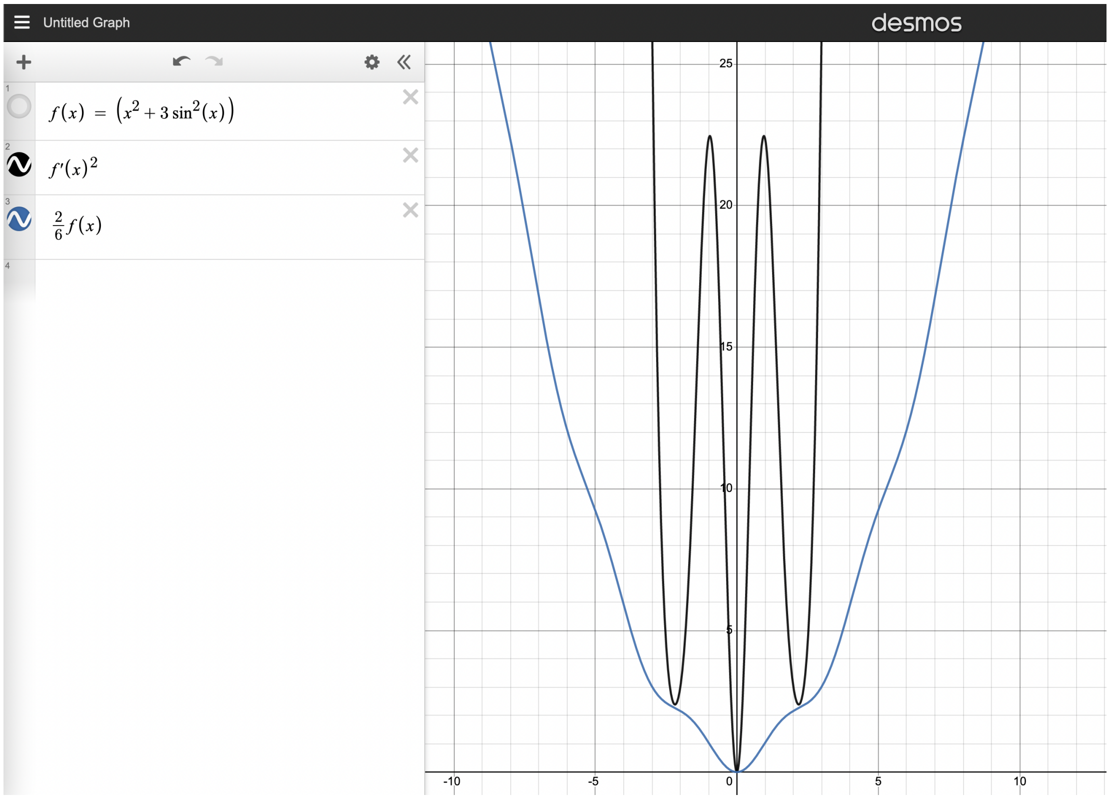
$$\mu\|x-y\| \leq \|\nabla f(x) - \nabla f(y)\|. \tag{1}$$

Finally we know by L-smoothness and $\nabla f(x_*) = 0$ where x_* is the minimum

$$f(x) - f(x_*) \overset{\nabla f(x_*)=0}{=} D_f^{(B)}(x, x_*) \overset{L\text{-smooth}}{\leq} \frac{L}{2}\|x - x_*\|^2 \overset{(1)}{\leq} \frac{L}{2\mu}\|\nabla f(x) - \underbrace{\nabla f(x_*)}_{=0}\|^2. \qquad \square$$

(iii) Use a graphing calculator to find c such that $f(x) = x^2 + 3\sin^2(x)$ satisfies the PL-condition (argue why $x \to \infty$ is not a problem) and prove it is not convex. (2 pts)

Solution. For $c = \frac{1}{6}$ we have the PL-condition



As $f'(x) = 2(x + 3\sin(x)\cos(x))$ and therefore

$$f'(x)^2 = 4(x + 3\underbrace{\sin(x)\cos(x)}_{\in[-1,1]})^2 \overset{|x|\geq 3}{\geq} 4(|x| - 3)^2$$

the x^2 dominates for large x, so if we make c small enough we can ensure the inequality for large x.

f is not convex because

$$f(\tfrac{1}{2}\pi + \tfrac{1}{2}0) = \frac{\pi^2}{4} + 3 > \tfrac{1}{2}\pi^2 = \tfrac{1}{2}f(\pi) + \tfrac{1}{2}f(0).$$

$\square$

Exercise 6 (Weak PL-Inequality). **(4 Points)**

Assume $f : \mathbb{R}^d \to \mathbb{R}$ 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

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

Prove the convergence rate

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

Solution. Divide the reordered contraction

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

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

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

(ii) Prove that f is bounded. More specifically $e(x) := f(x) - f_* \leq \frac{L}{2c^2}$ for all x. (1 pt)

Solution. Using Sheet 1 Exercise 1 (i), we get

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

and therefore

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

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

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

and thus

$$e(x) \leq \tfrac{L}{2c^2}. \qquad \square$$

(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

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

Solution. Using L-smoothness, we have

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

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

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

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. $\square$