Optimization in Machine Learning **Universität Mannheim**
HWS 2024 Prof. Simon Weißmann, Felix Benning

Solution Sheet 2

For the exercise class on the 03.10.2024 at 12:00.
Hand in your solutions by 10:15 in the lecture on Tuesday 01.10.2024.

Exercise 1 (Descent Directions of a Maximum)**.** **(1 Points)**

Let $x_* \in \mathbb{R}^d$ be a strict local maximum of $f : \mathbb{R}^d \to \mathbb{R}$. Prove that every $d \in \mathbb{R}^d$ 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 2 (Convergence to Stationary Point)**.** **(5 Points)**

Let $f : \mathbb{R}^d \to \mathbb{R}$ 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. $\nabla f(x_*) = 0$. (2.5 pts)

Solution. For any $\epsilon > 0$ there exists $n \geq 0$ such that for all $i, j \geq n$ we have by Cauchy-Schwarz and convergence of $\nabla f(x_i)$ to $\nabla f(x_*)$ due to continuity of ∇f

$$\begin{aligned}
&\langle \nabla f(x_i), \nabla f(x_j) \rangle \\
&= \|\nabla f(x_*)\|^2 + \langle \nabla f(x_i) - \nabla f(x_*), \nabla f(x_*) \rangle + \langle \nabla f(x_i), \nabla f(x_j) - \nabla f(x_*) \rangle \\
&\overset{\text{C.S.}}{\geq} \|\nabla f(x_*)\|^2 - \underbrace{\|\nabla f(x_i) - \nabla f(x_*)\|}_{\leq \epsilon} \|\nabla f(x_*)\| - \underbrace{\|\nabla f(x_i)\|}_{\leq \|\nabla f(x_*)\| + \epsilon} \underbrace{\|\nabla f(x_j) - \nabla f(x_*)\|}_{\leq \epsilon} \\
&\geq \|\nabla f(x_*)\|^2 - 2\epsilon \|\nabla f(x_*)\| - \epsilon^2 =: p(\epsilon)
\end{aligned}$$

This results in

$$\begin{aligned}
\|x_n - x_m\|^2 = \Big\| \sum_{k=n}^{m-1} \alpha_k \nabla f(x_k) \Big\|^2 &= \sum_{i,j=n}^{m-1} \alpha_i \alpha_j \langle \nabla f(x_i), \nabla f(x_j) \rangle \\
&\geq \left(\sum_{k=n}^{m-1} \alpha_k \right)^2 p(\epsilon)
\end{aligned}$$

Taking the limit over m results in

$$\infty > \|x_n - x_*\|^2 \geq \underbrace{\left(\sum_{k=n}^{\infty} \alpha_k\right)^2}_{=\infty} p(\epsilon).$$

So we necessarily need $p(\epsilon) \leq 0$. But as ϵ was arbitrary, we have

$$0 \leq \|\nabla f(x_*)\|^2 = \lim_{\epsilon \to 0} p(\epsilon) \leq 0. \qquad \square$$

(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

$$\min_{k \leq n} \|\nabla f(x_k)\|^2 \in o(1/n). \qquad \text{(2.5 pts)}$$

Solution. By L-smoothness of f, we have

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), \overbrace{x_{k+1} - x_k}^{-\alpha \nabla f(x_k)} \rangle + \tfrac{L}{2}\|x_{k+1} - x_k\|^2 \\ &= f(x_k) - (\alpha - \tfrac{L}{2}\alpha^2)\|\nabla f(x_k)\|^2 \end{aligned}$$

and therefore

$$\sum_{k=n}^{m} \|\nabla f(x_k)\|^2 \leq \sum_{k=n}^{m} \frac{f(x_k) - f(x_{k+1})}{\alpha(1 - \frac{L}{2}\alpha)} \overset{\text{telescope}}{=} \frac{f(x_n) - f(x_m)}{\alpha(1 - \frac{L}{2}\alpha)}.$$

Now $a_n := \min_{k \leq n} \|\nabla f(x_k)\|^2$ is non-increasing, therefore

$$n a_{2n} \leq \sum_{k=n}^{2n} a_k \leq \sum_{k=n}^{\infty} a_k \to 0 \quad (n \to \infty).$$

Thus $a_{2n} \in o(1/n)$. And we can simply bound the odd elements of the sequence

$$a_{2n+1} \leq a_{2n} \in o(1/n). \qquad \square$$

Exercise 3 (Optimizing Quadratic Functions)**.** **(9 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

$$f(x) = (x - x_*)^T A(x - x_*) + \tilde{c} \tag{1}$$
$$= \tfrac{1}{2}(x - x_*)^T \underbrace{(A^T + A)}_{=:H}(x - x_*) + \tilde{c} \tag{2}$$

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? (3 pts)

Solution. We want for some x_*

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

So we simply select

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

This proves our first representation (1). For (2) we simply note

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

Applying this to $y = x - x_*$ in (1) 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

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

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

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

(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. (1 pt)

Solution. Taking the derivative of (2) we get

$$\nabla f(x) = H(x - x_*). \tag{3}$$

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 (3) 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 $0 < \lambda_1 \leq \cdots \leq \lambda_d$ 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

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

with the condition number $\kappa = \frac{\lambda_d}{\lambda_1}$. (5 pts)

Solution. Using the representation (3) 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

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

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

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

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

$$\frac{d}{dh} = \begin{cases} -\lambda_1 & h \leq \frac{2}{\lambda_1+\lambda_2} \\ \lambda_d & h \geq \frac{2}{\lambda_1+\lambda_2} \end{cases}$$

So the maximal convergence speed is achieved by $h^* = \frac{2}{\lambda_1+\lambda_2}$ 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

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

and therefore

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

Exercise 4 (Programming exercise)**.** **(9 Points)**

For the Python exercises join the GitHub classroom `https://classroom.github.com/a/8yrTMIm1`. If you are new to git, checkout `https://classroom.github.com/a/dEzm_HGt`