Optimization in Machine Learning **Universität Mannheim**
HWS 2026 Prof. Leif Döring, Benedikt Wille

Solution Sheet 3

For the exercise class on the 05.11.2026 at 12:00.

Exercises to be solved in the exercise class

Exercise 1 (Lower Bounds). **(0 Points)**

In this exercise, we will bound the convergence rates of algorithms which pick their iterates x_{k+1} from

$$\operatorname{span}[\nabla f(x_0), \dots, \nabla f(x_k)] + x_0.$$

We consider the function

$$f_d(x) = \frac{1}{2}(x^{(1)} - 1)^2 + \frac{1}{2}\sum_{i=1}^{d-1}(x^{(i)} - x^{(i+1)})^2$$

(i) To understand our function f_d better, we want to view it as a potential on a graph. For this consider the undirected graph $G = (V, E)$ with vertices

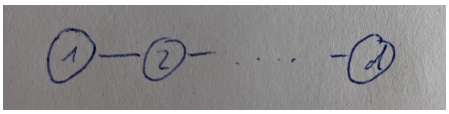
$$V = \{1, \dots, d\}$$

and edges

$$E = \{(i, i+1) : 1 \leq i \leq d-1\}.$$

Draw a picture of this graph.

Solution. The graph is simply a chain



□

(ii) We now interpret $x^{(i)}$ as a quantity (e.g. of heat) at vertex i of our graph G. Our potential f_d decreases, if the quantities at connected vertices i and $i+1$ are of similar size. I.e. if $(x^{(i)} - x^{(i+1)})^2$ is small. Additionally there is a pull for $x^{(1)}$ to be equal to 1. Use this intuition to find the minimizer x_* of f_d.

Solution. The minimizer is $x_* = (1, \dots, 1)^T \in \mathbb{R}^d$ since $f_d(x_*) = 0$ and $f_d(x) \geq 0$. □

(iii) The matrix $A^G \in \mathbb{R}^{d\times d}$ with

$$A^G_{i,j} = \begin{cases} \text{degree of vertex } i & i = j \\ -1 & (i,j) \in E \text{ or } (j,i) \in E \\ 0 & \text{else} \end{cases}$$

is called the "Graph-Laplacian" of G. The degree of vertex i are the number of connecting edges. Calculate A^G for G and prove that

$$\nabla f_d(x) = A^G x + (x^{(1)} - 1)e_1 = (A^G + e_1 e_1^T)x - e_1.$$

Solution. The Graph-Laplacian of G is given by

$$A^G = \begin{pmatrix} 1 & -1 & 0 & \cdots & \cdots & 0 \\ -1 & 2 & -1 & \ddots & & \vdots \\ 0 & -1 & 2 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & -1 & 0 \\ \vdots & & \ddots & -1 & 2 & -1 \\ 0 & \cdots & \cdots & 0 & -1 & 1 \end{pmatrix}$$

Let $i \neq 1, d$ then

$$\frac{\partial f_d}{\partial x_i} = [x^{(i)} - x^{(i+1)}] - [x^{(i-1)} - x^{(i)}] = 2x^{(i)} - x^{(i-1)} - x^{(i+1)} = (A^G x)_i$$

similarly looking at the cases $i = 1, d$ individually immediately reveals

$$\nabla f_d(x) = A^G x + (x^{(1)} - 1)e_1. \qquad \square$$

(iv) Prove that the Hessian $H = \nabla^2 f_d(x)$ is constant and positive definite to show that f_d is convex. Prove that the operator norm of H is smaller than 4. Argue that

$$g_d(x) := \tfrac{L}{4} f_d(x)$$

is therefore L-smooth.

Solution. Taking the derivative of the gradient we calculated previously yields

$$H = \nabla^2 f_d(x) = A^G + e_1 e_1^T.$$

To show positive definiteness, let y be arbitrary

$$y^T H y = (e_1^T y)^2 + y^T A^G y = (y^{(1)})^2 + \sum_{i=1}^{d-1} (y^{(i)} - y^{(i+1)})^2 \geq 0.$$

To find the largest eigenvalue of H we want to calculate the operator norm. For this we use $(a-b)^2 \leq 2(a^2 + b^2)$ to get

$$\langle y, Hy \rangle \leq (y^{(1)})^2 + 2\sum_{i=1}^{d-1} (y^{(i)})^2 + (y^{(i+1)})^2 \leq 4 \sum_{i=1}^{d} (y^{(i)})^2 = 4\|y\|^2.$$

Thus we get

$$\|H\| = \sup_{y : \|y\| = 1} \langle y, Hy \rangle \leq 4.$$

Since the operator norm coincides with the largest absolute eigenvalue for symmetric matrices, this proves our claim. Finally L-smoothness of g_d follows from

$$\|\nabla g_d(x) - \nabla g_d(y)\| = \tfrac{L}{4}\|\nabla f_d(x) - \nabla f_d(y)\| = \tfrac{L}{4}\|H(x-y)\| \leq \underbrace{\tfrac{L}{4}\|H\|}_{\leq L} \|x - y\|. \qquad \square$$

(v) Assume $x_0 = 0$ and and that $(x_n)_{n\in\mathbb{N}}$ is chosen with the restriction

$$x_{n+1} \in \mathcal{K}_n := \operatorname{span}[\nabla g_d(x_0), \dots, \nabla g_d(x_n)].$$

To make notation easier we are going to identify $\mathbb{R}^d$ with an isomorph subset of sequences

$$\mathbb{R}^d := \{x \in \ell^2 : x^{(i)} = 0 \quad \forall i > n\}$$

then $\mathbb{R}^n$ is a subset of $\mathbb{R}^d$ for $n \leq d$. Prove inductively that

$$\mathcal{K}_n \subseteq \mathbb{R}^{n+1} \subseteq \mathbb{R}^d$$

Solution. We have the induction start $n = 0$ by

$$g_d(x_0) = -\tfrac{L}{4}e_1 \in \mathbb{R}^1.$$

Now assume

$$\mathcal{K}_{n-1} \subseteq \mathbb{R}^n,$$

then by our selection process $x_n \in \mathbb{R}^n$. But then

$$\tfrac{4}{L}\nabla g_d(x_n) = \underbrace{A^G x_n}_{\in\mathbb{R}^{n+1}} + \underbrace{(x_n - 1)e_1}_{\in\mathbb{R}^1} \in \mathbb{R}^{n+1}.$$

We therefore have $\mathcal{K}_n = \operatorname{span}[\mathcal{K}_{n-1}, \nabla g_d(x_n)] \subseteq \mathbb{R}^{n+1}$.

Notice how the low connectedness of the graph G limits the spread of our quantity x_n. A higher connectedness would allow for information to travel much quicker. □

(vi) We now want to bound the convergence speed of x_n to x_*. For this we select $d = 2n + 1$.

Note: We may choose a larger dimension d by defining f_{2n+1} on the subset $\mathbb{R}^{2n+1}$ in $\mathbb{R}^d$. The important requirement is therefore $2n + 1 \leq d$. But without loss of generality we assume equality.

Use the knowledge we have collected so far to argue

$$\|x_* - x_n\|^2 \geq d - n \geq \tfrac{1}{2}\|x_* - x_0\|^2.$$

Solution. Since $x_n \in \mathbb{R}^n$ we know that

$$\begin{aligned}
\|x_* - x_n\|^2 &= \sum_{i=1}^{d} (x_*^{(i)} - x_n^{(i)})^2 \\
&\geq \sum_{i=n+1}^{d} (x_*^{(i)})^2 \\
&= d - n \overset{d=2n+1}{=} n + 1 = \tfrac{n+1}{2n+1} d \geq \tfrac{1}{2} d = \frac{1}{2}\sum_{i=1}^{d} 1^2 = \tfrac{1}{2}\|x_* - x_0\|^2.
\end{aligned}$$

□

(vii) To prevent the convergence of the loss $g_d(x_n)$ to $g_d(x_*)$ we need a more sophisticated argument. For this consider

$$\tilde{g}_n(x) := \tfrac{L}{4}[f_n(x) + \tfrac{1}{2}(x^{(n)} - 0)^2].$$

Argue that on $\mathbb{R}^n \subset \mathbb{R}^d$ the functions $\tilde{g}_n$ and g_d are identical. Use this observation to prove

$$g_d(x_n) - \inf_x g_d(x) \geq \inf_x \tilde{g}_n(x).$$

Solution. Let $x \in \mathbb{R}^n$. Then using $x^{(n+1)} = 0$ we have

$$\tilde{g}_n(x) = \frac{L}{8}\Big[(x^{(1)} - 1)^2 + \sum_{i=1}^{n}(x^{(i)} - x^{(i+1)})\Big] = g_d(x)$$

using $x^{(i)} = 0$ for all $i > n$ for the second equality sign. Since $x_n \in \mathbb{R}^n$ we therefore can replace g_d with g_n at will to get

$$g_d(x_n) - \underbrace{\inf_x g_d(x)}_{=0} = \tilde{g}_n(x_n) \geq \inf_x \tilde{g}(x). \qquad \square$$

(viii) Our goal is now to calculate $\inf_x \tilde{g}_n(x)$. Prove convexity of $\tilde{g}_n$ and prove that

$$\hat{x}_n^{(i)} = \begin{cases} 1 - \frac{i}{n+1} & i \leq n+1 \\ 0 & i \geq n+1 \end{cases}$$

is its minimum. Then plug our solution into $\tilde{g}_n$ (or g_d, since $\hat{x}_n$ is in the subset $\mathbb{R}^n$ after all), to obtain the lower bound

$$g_d(x_n) - \inf_x g_d(x) \geq \frac{L\|x_0 - x_*\|^2}{8(n+1)d} \geq \frac{L\|x_0 - x_*\|^2}{16(n+1)^2}.$$

Solution. We have

$$\nabla\tilde{g}_n(x) = \frac{L}{4}[A^{G_n}x + (x^{(1)} - 1)e_1 + (x^{(n)})e_n] = \frac{L}{4}(A^{G_n} + e_1e_1^T + e_ne_n^T)x - e_1$$

where A^{G_n} is the Graph-Laplacian for f_n. Then the Hessian is obviously positive definite

$$\nabla^2\tilde{g}_n(x) = \frac{L}{4}(A^{G_n} + e_1e_1^T + e_ne_n^T)$$

as we could apply the same arguments as for f_n. So $\tilde{g}_n$ is convex. We now plug $\hat{x}_n$ into $\nabla\tilde{g}_n$ to verify the first order condition, proving it is a minimum

$$\begin{aligned}
\frac{4}{L}\frac{\partial\tilde{g}_n}{\partial x_i}(\hat{x}_n) &= (A^{G_n}\hat{x}_n)_i + \underbrace{(\hat{x}_n^{(1)} - 1)}_{=-\frac{1}{n+1}}\delta_{i1} + \underbrace{(\hat{x}_n^{(n)})}_{=\frac{1}{n+1}}\delta_{in} \\
&= \underbrace{[\hat{x}_n^{(i)} - \hat{x}_n^{(i+1)}]}_{-\frac{1}{n+1}}\mathbb{1}_{i\neq n} - \underbrace{[\hat{x}_n^{(i-1)} - \hat{x}_n^{(i)}]}_{-\frac{1}{n+1}}\mathbb{1}_{i\neq 1} - \tfrac{1}{n+1}\delta_{i1} + \tfrac{1}{n+1}\delta_{in} \\
&= 0.
\end{aligned}$$

We now know that

$$\begin{aligned}\inf_x \tilde{g}_n(x) = \tilde{g}_n(\hat{x}_n) &= \frac{L}{8}\Big[\big(-\tfrac{1}{n+1}\big)^2 + \big(1-\tfrac{n}{n+1}\big)^2 + \sum_{i=1}^{n-1}\big(\tfrac{i+1}{n+1}-\tfrac{i}{n+1}\big)^2\Big] \\ &= \frac{L}{8}\sum_{i=0}^{n}\big(\tfrac{1}{n+1}\big)^2 = \frac{L}{8(n+1)} \overset{d=\|x_0-x_*\|^2}{=} \frac{L\|x_0-x_*\|^2}{8(n+1)d} \\ &\geq \frac{L\|x_0-x_*\|^2}{16(n+1)^2}\end{aligned}$$

using $d = 2n+1$ again. □

(ix) Argue that we only needed

$$x_n = x_0 + \sum_{k=0}^{n-1} A_k \nabla f(x_k)$$

with upper triangular matrices A_k to make these bounds work. Since adaptive methods (like Adam) use diagonal matrices A_k, they are therefore covered by these bounds.

Solution. We only needed $\mathcal{K}_n \subseteq \mathbb{R}^{n+1}$ which we proved by induction using only this fact about $\mathcal{K}_{n-1}$. Since upper triangular matrices do not change this fact, we may as well allow them. □

(x) Bask in our glory! For we have proven that ...? Summarize our results into a theorem.

Solution.

Theorem (Nesterov)**.** *Assume there exists upper triangular matrices $A_{k,n}$ such that the sequence $(x_n)_{n\in\mathbb{N}}$ in $\mathbb{R}^d$ is selected by the rule*

$$x_n = x_0 + \sum_{k=0}^{n-1} A_{k,n}\nabla f(x_k)$$

for a convex, L-smooth f to minimize. Then up to $n \leq \frac{d-1}{2}$ there exists a convex, L-smooth function f such that

$$\begin{aligned}\|x_n - x_*\| &\geq \tfrac{1}{\sqrt{2}}\|x_0 - x_*\| \\ f(x_n) - \inf_x f(x) &\geq \frac{L\|x_0-x_*\|^2}{16(n+1)^2}\end{aligned}$$

for $f(x_*) = \inf_x f(x)$.

□