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

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.

(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.

(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

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

(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.

(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

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

(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

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

(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

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

(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

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

(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.

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

Exercises to be prepared at home for the next exercise class at 12.11.2026

Exercise 2 (Swap Integration with Differentiation). **(0 Points)**

(i) What formal requirements on $f : V \times \Omega \to \mathbb{R}$ with $V \subseteq \mathbb{R}$ and measure μ on Ω are needed, for the following argument using the fundamental theorem of calculus (FTC) to work?

$$\begin{aligned}
\frac{\partial}{\partial t}\int_\Omega f(t_0,\omega)d\mu(\omega) &\overset{\text{linear}}{=} \lim_{\epsilon\to 0}\int \frac{f(t_0+\epsilon,\omega)-f(t_0,\omega)}{\epsilon}d\mu(\omega)\\
&\overset{\text{FTC II}}{=} \lim_{\epsilon\to 0}\int \frac{1}{\epsilon}\int_{t_0}^{t_0+\epsilon}\frac{\partial}{\partial t}f(t,\omega)dtd\mu(\omega)\\
&\overset{\text{Fubini}}{=} \lim_{\epsilon\to 0}\frac{1}{\epsilon}\int_{t_0}^{t_0+\epsilon}\int \frac{\partial}{\partial t}f(t,\omega)d\mu(\omega)dt\\
&\overset{\text{def.+lin.}}{=} \frac{d}{dy}\int_{t_0}^{y}\int \frac{\partial}{\partial t}f(t,\omega)d\mu(\omega)dt\Bigg|_{y=t_0}\\
&\overset{\text{FTC I}}{=} \int \frac{\partial}{\partial t}f(t_0,\omega)d\mu(\omega).
\end{aligned}$$

Formulate the corresponding theorem.

(ii) We want to find an example for a function, where you can not swap integration with differentiation. So for a function $f(t,\omega)$ we need some t_0 such that

$$\frac{\partial}{\partial t}\int_\Omega f(t_0,\omega)d\omega \neq \int_\Omega \frac{\partial}{\partial t}f(t_0,\omega)d\omega.$$

For this consider $f(t,\omega) = t^3e^{-t^2\omega}$. Prove the inequality at $t_0 = 0$ and $\Omega = [0,\infty)$. Why is this not a contradiction to (i)?

Hint. *It is helpful to calculate the entire function*

$$t \mapsto \int_0^\infty \frac{\partial}{\partial t}f(t,\omega)d\omega.$$