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

# Solution Sheet 4

For the exercise class on the 27.04.2023.
Hand in your solutions by 12:00 in the exercise on Thursday 27.04.2023.

While there are 38 in total, you may consider all points above the standard 24 to be bonus points.

**Exercise 1** (Lower Bounds)**.** **(13 Points)**

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

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

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

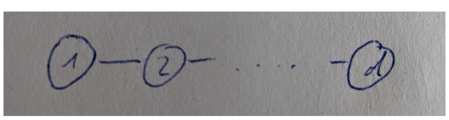
$$V = \{1, \ldots, d\}$$

and edges

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

Draw a picture of this graph. (1 pt)

*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 $f_d$. (1 pt)

*Solution.* The minimizer is $x_* = (1, \ldots, 1)^T \in \mathbb{R}^d$ 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

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

*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. (2 pts)

*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 $n \le d$. Prove inductively that

$$\mathcal{K}_n \subseteq \mathbb{R}^{n+1} \subseteq \mathbb{R}^d \qquad \text{(1 pt)}$$

*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 $2n + 1 \le d$. But without loss of generality we assume equality.

Use the knowledge we have collected so far to argue

$$\|x_* - x_n\|^2 \ge d - n \ge \tfrac{1}{2}\|x_* - x_0\|^2. \qquad \text{(1 pt)}$$

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

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

(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

$$g_d(x_n) - \inf_x g_d(x) \geq \inf_x \tilde{g}_n(x). \qquad \text{(1 pt)}$$

*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

$$g_d(x_n) - \inf_x g_d(x) \geq \frac{L\|x_0 - x_*\|^2}{8(n+1)d} \geq \frac{L\|x_0 - x_*\|^2}{16(n+1)^2}. \qquad \text{(3 pts)}$$

*Solution.* We have

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

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

$$\nabla^2 \tilde{g}_n(x) = \frac{L}{4}(A^{G_n} + e_1 e_1^T + e_n e_n^T)$$

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. (1 pt)

*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. (1 pt)

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

□

(xi) (Bonus) If you wish, you may want to try and repeat those steps for

$$G_d(x) = \frac{L-\mu}{L} g_d(x) + \frac{\mu}{2}\|x\|^2$$

to prove an equivalent result for $\mu$-strongly convex functions. Unfortunately finding $x_*$ is much more difficult in this case. Letting $d \to \infty$ makes this problem tractable again with solution

$$x_*^{(i)} = \Big(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\Big)^i.$$

**Exercise 2** (Conjugate Gradient Descent). **(12 Points)**

Consider a quadratic function

$$f(x) = \tfrac{1}{2}(x - x_*)^T H(x - x_*)$$

for some symmetric and positive definite $H$ and consider the hilbert space $\mathcal{H} = (\mathbb{R}^d, \langle \cdot, \cdot \rangle_H)$ with

$$\langle x, y \rangle_H = \langle x, Hy \rangle$$

(i) Prove that $\langle \cdot, \cdot \rangle_H$ is a well-defined scalar product. Convince yourself that

$$f(x) = \tfrac{1}{2}\|x - x_*\|_H^2. \qquad \text{(1 pt)}$$

*Solution.* Bilinearity is trivial, the positive-definiteness follows from this property of $H$. We have

$$f(x) = \tfrac{1}{2}\langle x - x_*, H(x - x_*)\rangle = \tfrac{1}{2}\langle x - x_*, (x - x_*)\rangle_H = \tfrac{1}{2}\|x - x_*\|_H^2. \qquad \square$$

(ii) Determine the derivative $\nabla_H f(x)$ of $f$ in $\mathcal{H}$ (1 pt)

*Solution.* We need

$$\begin{aligned} 0 &\overset{!}{=} \lim_{v \to 0} \frac{|f(x+v) - f(x) - \langle \nabla_H f(x), v \rangle_H|}{\|v\|_H} \\ &= \lim_{v \to 0} \frac{|f(x+v) - f(x) - \langle H\nabla_H f(x), v \rangle|}{\|v\|} \underbrace{\frac{\|v\|}{\|v\|_H}}_{\geq c}. \end{aligned}$$

We can bound the fraction of norms by a constant $c > 0$ from below due to equivalence of all norms in $\mathbb{R}^d$. This lower bound on the second fraction forces the first fraction to converge to zero. But this implies that

$$\nabla f(x) = H\nabla_H f(x)$$

by the definition (and uniqueness) of $\nabla f(x)$. Thus the gradient we are looking for is

$$\nabla_H f(x) = H^{-1}\nabla f(x). \qquad \square$$

(iii) Since gradient descent in the space $\mathcal{H}$ is therefore computationally the Newton method, we want to find a different method of optimization. Consider an arbitrary set of conjugate ($H$-orthogonal) directions $(v_1, \ldots v_d)$, i.e. $\langle v_i, v_j \rangle_H = \delta_{ij}$, and for some starting point $x_0 \in \mathbb{R}^d$ the following descent algorithm:

$$x_{k+1} = x_k - \alpha_k v_{k+1} \quad \text{with} \quad \alpha_k := \operatorname*{argmin}_{\alpha} f(x_k - \alpha v_{k+1}). \qquad \text{(CD)}$$

Optimizing over $\alpha$ in this manner is known as "line-search". Using $y^{(i)} := \langle y, v_i \rangle$ prove that

$$(x_k - x_*) = \sum_{i=k+1}^{d} (x_0 - x_*)^{(i)} v_i = \operatorname*{argmin}_{x}\{f(x) : x \in x_0 + \operatorname{span}[v_1, \ldots, v_k]\} - x_*.$$

Deduce that conjugate descent (CD) converges in $d$ steps. (2 pts)

*Solution.* We proceed by induction. The induction start with $k = 0$ is obvious. Let us now consider $x_{k+1}$. By its definition we have

$$\begin{aligned}
2f(x_{k+1}) &= \min_\alpha 2f(x_k - \alpha v_{k+1}) \\
&= \min_\alpha \|x_k - \alpha v_{k+1}\|_H \\
&= \min_\alpha \Big\| \sum_{i=1}^d (x_k - x_*)^{(i)} v_i - \alpha v_{k+1} \Big\|_H \\
&= \min_\alpha [(x_k - x_*)^{(k+1)} - \alpha]^2 \|v_{k+1}\|_H^2 + \sum_{i=k+2}^d [(x_k - x_*)^{(i)}]^2 \|v_i\|_H^2 \\
&= \sum_{i=k+2}^d [(x_k - x_*)^{(i)}]^2.
\end{aligned}$$

the minimizer is therefore $\alpha_k = (x_k - x_*)^{(k+1)}$. This removes the $v_{k+1}$ component leaving us with the components $v_{k+2}$ and up. Note that $(x_k - x_*)^{(i)} = (x_0 - x_*)^{(i)}$ for all $i \geq k+1$ by induction. Similarly we can see that this is a minimum in the span of $v_1, \dots, v_{k+1}$, as we have removed those components completely and

$$f(x) = \|x - x_*\|_H^2 = \sum_{i=1}^d [(x - x_*)^{(i)}]^2.$$

Since we can not touch the other components due to $H$-orthogonality, this is the best we can do. □

(iv) If we had $v_i = \nabla f(x_{i-1})$, then this algorithm would be optimal in the set of algorithms we considered in the previous exercise. Unfortunately the gradients $\nabla f(x_{i-1})$ are generally not conjugate. So while we may select an arbitrary set of conjugate $v_i$, we cannot select the gradients directly.

Instead we are going to do the next best thing and inductively select $v_{k+1}$ such that

$$\mathcal{K}_k := \operatorname{span}[\nabla f(x_0), \dots \nabla f(x_k)] = \operatorname{span}[v_1, \dots, v_{k+1}]$$

using the Gram-Schmidt procedure to make $v_{k+1}$ conjugate to $v_1, \dots, v_k$. Since Gram-Schmidt is still computationally too expensive for our tastes, you please inductively prove

$$\mathcal{K}_k = \operatorname{span}[H^1(x_0 - x_*), \dots, H^{k+1}(x_0 - x_*)].$$

assuming $\mathcal{K}_k$ is $(k+1)$-dimensional. I.e. $\mathcal{K}_k$ is a "$H$-Krylov subspace". (2 pts)

*Solution.* The induction start $k = 0$ follows directly from

$$\nabla f(x_0) = H(x_0 - x_*)$$

and the definition of $\mathcal{K}_0$. Assume we have the claim for $k - 1$, then

$$\nabla f(x_k) = H(x_k - x_*) = H(x_{k-1} - \alpha_{k-1} v_k - x_*) = \underbrace{H(x_{k-1} - x_*)}_{= \nabla f(x_{k-1}) \in \mathcal{K}_{k-1}} - \alpha_{k-1} H \underbrace{v_k}_{\in \mathcal{K}_{k-1}}.$$

As $\mathcal{K}_{k-1} = \text{span}[H^1(x_0 - x_*), \dots, H^k(x_0 - x_*)]$ by the induction hypothesis, we therefore have

$$\nabla f(x_k) \in \text{span}[H^1(x_0 - x_*), \dots, H^{k+1}(x_0 - x_*)].$$

Since $\nabla f(x_0), \dots, \nabla f(x_{k-1}) \in \mathcal{K}_{k-1}$ they are by the induction hypothesis also in the span

$$\mathcal{K}_k = \text{span}[\nabla f(x_0), \dots, \nabla f(x_k)] \subseteq \text{span}[H^1(x_0 - x_*), \dots, H^{k+1}(x_0 - x_*)].$$

Since the space on the left is $k+1$ dimensional, we have equality. □

(v) Argue that $\nabla f(x_{k+1})$ is orthogonal to every vector in $\mathcal{K}_k$ and inductively deduce either

$$\nabla f(x_{k+1}) = 0$$

which implies $x_{k+1} = x_*$, or $\mathcal{K}_{k+1}$ has full rank. Deduce from the $H$-Krylov-subspace property, that $\nabla f(x_{k+1})$ is already $H$-orthogonal to $\mathcal{K}_{k-1}$. (2 pts)

*Solution.* By the selection process of $x_{k+1}$, we have

$$x_{k+1} = \underset{x}{\text{argmin}}\{f(x) : x \in \mathcal{K}_k + x_0\}.$$

assume $\nabla f(x_{k+1})$ were not orthogonal to $\mathcal{K}_k$. Then there would exist $v \in \mathcal{K}_k$ such that

$$\langle \nabla f(x_{k+1}), v \rangle > 0$$

By the Taylor approximation we therefore have

$$f(x_{k+1} - \delta v) = f(x_{k+1}) - \delta \underbrace{\langle \nabla f(x_{k+1}), v \rangle}_{>0} + O(\delta^2)$$

so there exists a small $\delta > 0$ such that $f(x_{k+1} - \delta v) < f(x_{k+1})$. But this is a contradiction since $x_{k+1}$ was optimal.

$\nabla f(x_{k+1})$ is therefore orthogonal to $\mathcal{K}_k$. So if it is not zero, $\mathcal{K}_{k+1}$ has (as the span of both) full rank. $\nabla f(x_{k+1})$ being orthogonal to $\mathcal{K}_k$ also implies it is orthogonal to $H\mathcal{K}_{k-1}$, since that is a subspace of $\mathcal{K}_k$ by the Krylov property. But this implies $\nabla f(x_{k+1})$ is $H$-orthogonal to $\mathcal{K}_{k-1}$. □

(vi) Collect the ideas we have gathered to prove the recursively defined

$$v_{k+1} = \nabla f(x_k) - \frac{\langle \nabla f(x_k), v_k \rangle_H}{\|v_k\|_H^2} v_k$$

are $H$-conjugate and have the same span as the gradients up to $\nabla f(x_k)$. (1 pt)

*Solution.* These $v_k$ are the same $v_k$ we would obtain using Gram-Schmidt on the gradients. In fact this is Gram-Schmidt together with the fact that $\nabla f(x_k)$ is already $H$-orthogonal to the $v_1, \dots, v_{k-1} \in \mathcal{K}_{k-2}$. So only the last summand remains. □

(vii) To make our procedure truly computable, we want to show

$$\frac{\langle \nabla f(x_k), v_k\rangle_H}{\|v_k\|_H^2} = -\frac{\|\nabla f(x_k)\|^2}{\|\nabla f(x_{k-1})\|^2}. \qquad \text{(2 pts)}$$

*Solution.* We have

$$\nabla f(x_k) = H(\overbrace{x_{k-1} - \alpha_{k-1} v_k}^{x_k} - x_*) = \nabla f(x_{k-1}) - \alpha_{k-1} H v_k.$$

This implies $v_k = \frac{1}{\alpha_{k-1}} H^{-1}[\nabla f(x_{k-1}) - \nabla f(x_k)]$ and therefore

$$\langle \nabla f(x_k), v_k\rangle_H = \tfrac{1}{\alpha_{k-1}}\langle \nabla f(x_k), [\nabla f(x_{k-1}) - \nabla f(x_k)]\rangle = -\frac{\|\nabla f(x_k)\|^2}{\alpha_{k-1}},$$

where we have used $\langle \nabla f(x_k), \nabla f(x_{k-1})\rangle = 0$, which follows from $\nabla f(x_{k-1}) \in \mathcal{K}_{k-1}$ and $\nabla f(x_k) \perp \mathcal{K}_{k-1}$.

Now we need to find $\alpha_{k-1}$. But the first order condition

$$\begin{aligned} 0 &\overset{!}{=} \frac{d}{d\alpha} f(x_{k-1} - \alpha v_k) \\ &= -\langle \nabla f(x_{k-1} - \alpha v_k), v_k\rangle \\ &= -\langle H(x_{k-1} - x_* - \alpha v_k), v_k\rangle \\ &= -\langle \nabla f(x_{k-1}), v_k\rangle + \alpha \|v_k\|_H^2. \end{aligned}$$

implies

$$\alpha_{k-1} = \frac{\langle \nabla f(x_{k-1}), v_k\rangle}{\|v_k\|_H^2}.$$

Before we put things together, note that by definition of $v_k$

$$\langle \nabla f(x_{k-1}), v_k\rangle = \langle \nabla f(x_{k-1}), \nabla f(x_{k-1}) - c v_{k-1}\rangle = \|\nabla f(x_{k-1})\|^2,$$

since $\nabla f(x_{k-1})$ is orthogonal to $v_{k-1} \in \mathcal{K}_{k-2}$. From this we get

$$\alpha_{k-1} = \frac{\|\nabla f(x_{k-1})\|^2}{\|v_k\|_H^2},$$

So we finally get

$$\frac{\langle \nabla f(x_k), v_k\rangle_H}{\|v_k\|_H^2} = -\frac{\|\nabla f(x_k)\|^2}{\|v_k\|_H^2} \frac{\|v_k\|_H^2}{\|\nabla f(x_{k-1})\|^2} = -\frac{\|\nabla f(x_k)\|^2}{\|\nabla f(x_{k-1})\|^2}. \qquad \square$$

(viii) Summarize everything into a pseudo-algorithm for conjugate gradient descent (CGD) and compare it to heavy-ball momentum with

$$\beta_k = \frac{\alpha_k \|\nabla f(x_k)\|^2}{\alpha_{k-1}\|\nabla f(x_{k-1})\|^2}$$

using identical $\alpha_k$ as CGD. (1 pt)

*Solution.* We set $v_1 = \nabla f(x_0)$ or later

$$v_{k+1} = \nabla f(x_k) + \frac{\|\nabla f(x_k)\|^2}{\|\nabla f(x_{k-1})\|^2} v_k$$

determine the step-size

$$\alpha_k = \operatorname*{argmin}_{\alpha} f(x_k - \alpha v_{k+1})$$

and finally make our step

$$x_{k+1} = x_k - \alpha_k v_{k+1}.$$

Using the fact $v_k = \frac{x_{k-1}-x_k}{\alpha_{k-1}}$ and inserting $v_{k+1}$ into the last equation, we notice

$$\begin{aligned} x_{k+1} &= x_k - \alpha_k \Big[\nabla f(x_k) + \frac{\|\nabla f(x_k)\|^2}{\|\nabla f(x_{k-1})\|^2} \frac{x_{k-1} - x_k}{\alpha_{k-1}}\Big] \\ &= x_k - \alpha_k \nabla f(x_k) + \underbrace{\frac{\alpha_k}{\alpha_{k-1}} \frac{\|\nabla f(x_k)\|^2}{\|\nabla f(x_{k-1})\|^2}}_{=\beta_k} (x_k - x_{k-1}) \end{aligned}$$

that CGD is identical to HBM with certain parameters $\alpha_k$, $\beta_k$. □

**Exercise 3** (Momentum). **(13 Points)**

In this exercise, we take a closer look at heavy-ball momentum

$$x_{k+1} = x_k + \beta_k(x_k - x_{k-1}) + \alpha_k \nabla f(x_k)$$

(i) Find a continuous function $f : \mathbb{R} \to \mathbb{R}$ such that

$$f'(x) = \begin{cases} 25x & x < 1 \\ x + 24 & 1 < x < 2 \\ 25x - 24 & 2 < x. \end{cases}$$

Prove that $f$ is $\mu$-strongly convex with $\mu = 1$, $L$-smooth with $L = 25$ and has a minimum in zero. (2 pts)

*Solution.* We define

$$f(x) = \begin{cases} \frac{25}{2}x^2 & x \le 1 \\ \frac{1}{2}x^2 + 24x - 12 & 1 < x < 2 \\ \frac{25}{2}x^2 - 24x + 36 & 2 \le x, \end{cases}$$

note that it is continuous in 1 and 2 and therefore everywhere, and that it has the correct derivative. Further note that

$$f''(x) = \begin{cases} 1 & 1 < x < 2 \\ 25 & \text{else} \end{cases}$$

is the derivative of $f'(x)$ in the following sense:

$$f'(x) = \int_0^x f''(t)dt,$$

which follows from differentiability of $f'$ on its segments with the fundamental theorem of calculus and continuity between segments. Thus we have

$$\begin{aligned} f(y) &= f(x) + \int_x^y f'(t)dt = f(x) + f'(x)(y-x) + \int_x^y f'(t) - f'(x)dt \\ &= f(x) + f'(x)(y-x) + \int_x^y \int_x^t f''(s)dsdt. \end{aligned}$$

For the Bregman divergence this implies

$$\tfrac{1}{2}\|y-x\|^2 \le D_f^{(B)}(y,x) = \int_x^y \int_x^t f''(s)dsdt \le \tfrac{25}{2}\|y-x\|^2,$$

thus $f$ is $\mu = 1$-strongly convex and $L = 25$-smooth. □

(ii) Recall, we required for convergence of HBM

$$1 > \beta \ge \max\{(1-\sqrt{\alpha\mu})^2, (1-\sqrt{\alpha L})^2\}.$$

Calculate the optimal $\alpha$ and $\beta$ to minimize the rate $\sqrt{\beta}$. (1 pt)

*Solution.* To minimize $\sqrt{\beta}$, we first set

$$\beta = \max\{(1-\sqrt{\alpha\mu})^2, (1-\sqrt{\alpha L})^2\}$$

and then proceed to minimize this over $\alpha$. Which results in

$$\begin{aligned} \alpha^* &= \operatorname*{argmin}_\alpha \max\{(1-\sqrt{\alpha\mu})^2, (1-\sqrt{\alpha L})^2\} \\ &= \operatorname*{argmin}_\alpha \max\{|1-\sqrt{\alpha\mu}|, |1-\sqrt{\alpha L}|\} \\ &= \operatorname*{argmin}_\alpha \max\{(1-\sqrt{\alpha\mu}), -(1-\sqrt{\alpha\mu}), (1-\sqrt{\alpha L}), -(1-\sqrt{\alpha L})\} \\ &= \operatorname*{argmin}_\alpha \max\{(1-\sqrt{\alpha\mu}), -(1-\sqrt{\alpha L})\} \end{aligned}$$

which is monotonously falling for

$$1-\sqrt{\alpha\mu} > \sqrt{\alpha L} - 1$$

and monotonously increasing otherwise. Therefore its minimum is at equality. Thus

$$1-\sqrt{\alpha^*\mu} = \sqrt{\alpha^* L} - 1 \iff 2 = \sqrt{\alpha^*}(\sqrt{\mu}+\sqrt{L}) \iff \alpha^* = \frac{4}{(\sqrt{\mu}+\sqrt{L})^2}.$$

This results in

$$\beta^* = \Big(1 - \frac{2}{1+\sqrt{L/\mu}}\Big)^2.$$ □

(iii) Prove, using heavy ball momentum on $f$ with the optimal parameters results in the recursion (1 pt)

$$x_{k+1} = \tfrac{13}{9}x_k - \tfrac{4}{9}x_{k-1} - \tfrac{1}{9}\nabla f(x_k).$$

*Solution.* Using our previous results about optimal rates we have for $f$

$$\alpha^* = \frac{4}{(1+5)^2} = \frac{1}{9} \qquad \beta^* = (1 - \tfrac{2}{1+5})^2 = \frac{4}{9}.$$

Thus

$$x_{k+1} = \underbrace{x_k + \tfrac{4}{9}(x_k - x_{k-1})}_{=\frac{13}{9}x_k - \frac{4}{9}x_{k-1}} + \tfrac{1}{9}\nabla f(x_k).$$

□

(iv) We want to find a cycle of points $p \to q \to r \to p$, such that for $x_0 = p$ we have

$$x_{3k} = p \quad x_{3k+1} = q \quad x_{3k+2} = r \qquad \forall k \in \mathbb{N}_0.$$

Assume $p < 1$, $q < 1$ and $r > 2$ and use the heavy-ball recursion to create linear equations for $p, q, r$. Solve this linear equation. What does this mean for convergence? (3 pts)

*Solution.* We have

$$\begin{pmatrix} p \\ q \\ r \end{pmatrix} = \begin{pmatrix} 0 & -\frac{4}{9} & \frac{13}{9} \\ \frac{13}{9} & 0 & -\frac{4}{9} \\ -\frac{4}{9} & \frac{13}{9} & 0 \end{pmatrix} \begin{pmatrix} p \\ q \\ r \end{pmatrix} - \frac{1}{9} \begin{pmatrix} \nabla f(r) \\ \nabla f(p) \\ \nabla f(q) \end{pmatrix}$$

Multiplying both sides by $9$, using $\nabla f(r) = 25r - 24$ and $\nabla f(p) = 25p$ and similarly $q$ and reordering, we get

$$\begin{pmatrix} 9 & 4 & 12 \\ 12 & 9 & 4 \\ 4 & 12 & 9 \end{pmatrix} \begin{pmatrix} p \\ q \\ r \end{pmatrix} = \begin{pmatrix} 24 \\ 0 \\ 0 \end{pmatrix}$$

solving this system of equations results in

$$p = \tfrac{792}{1225} \approx 0.65, \quad q = -\tfrac{2208}{1225} \approx -1.80, \quad r = \tfrac{2592}{1225} \approx 2.12.$$

As we have managed to find a cycle of points, HBM does not converge to the minimum at zero in this case. Note: it is also possible to show that this cycle is attractive if you start in an epsilon environment away from these points. □

(v) Implement Heavy-Ball momentum, Nesterov's momentum and CGD `https://classroom.github.com/a/f3PnRxTs`. (6 pts)