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

Solution Sheet 2

For the exercise class on the 15.10.2026 at 12:00.

Exercises from the in-class test

Exercise 1 (Strict & Strong Convexity)**.** **(0 Points)**

Prove the following statements

(i) μ-strong convexity implies strict convexity.

Solution. Assuming $x \neq y$ we have

$$0 < \tfrac{\mu}{2}\|x-y\|^2 \overset{\mu\text{-strong conv.}}{\leq} f(x) - f(y) - \langle \nabla f(y), x-y\rangle$$

But after moving the negative parts to the left we are left with a condition which is equivalent to strict convexity by Proposition A.1.8 in the lecture. □

(ii) For twice differentiable f, the following are equivalent

(a) $\nabla^2 f(x) \succeq \mu \mathbb{I}$
(b) $z^T \nabla^2 f(x) z \geq \mu \|z\|^2$
(c) f is μ-strongly convex

where $\mathbb{I}$ is the identity matrix and

$$A \succeq B :\iff A - B \text{ is (weakly) positive definite.}$$

Solution. Let us get (a)⇔(b) out of the way first:

$$z^T \nabla^2 f(x) z - \mu\|z\|^2 = \langle z, \nabla^2 f(x) z\rangle - \langle z, \mu \mathbb{I} z\rangle = z^T(\nabla^2 f(x) - \mu \mathbb{I}) z$$

If we have (a), then the rightmost side is positive and thus we also have (b). For the other direction we simply start from the left.

For (c)⇒(a) let

$$g(y) := f(y) - \tfrac{\mu}{2}\|y\|^2,$$

because f is μ strongly convex

$$\begin{aligned} g(y) &= f(y) - \tfrac{\mu}{2}\|y\|^2 \\ &\overset{f\ \mu\text{-strong conv.}}{\geq} f(x) + \langle \nabla f(x), y-x\rangle + \tfrac{\mu}{2}\|y-x\|^2 - \tfrac{\mu}{2}\|y\|^2 \\ &= \underbrace{f(x) - \tfrac{\mu}{2}\|x\|^2}_{=g(x)} + \tfrac{\mu}{2}\|x\|^2 + \langle \underbrace{\nabla f(x) - \mu x}_{\nabla g(x)}, y-x\rangle + \mu\langle x, y-x\rangle - \mu\langle x, y\rangle + \tfrac{\mu}{2}\|x\|^2 \\ &= g(x) + \langle \nabla g(x), y-x\rangle \end{aligned}$$

g is still convex. So we know by the lecture that

$$\nabla^2 g(x) = \nabla^2 f(x) - \mu \mathbb{I}$$

is positive definite.

On the other hand for (a)⇒(c): if $\nabla^2 f \succeq \mu \mathbb{I}$, we know that $\nabla^2 g$ is positive definite and thus that g is convex by the lecture. Reusing the equations above except for the inequality we obtain

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

Adding $\frac{\mu}{2}\|y\|^2$ on both sides results in the definition of μ-strong convexity of f. □

Exercise 2 (Directional Minima)**.** **(0 Points)**

Let f be some differentiable function. For every direction $d \in \mathbb{R}^n$ define

$$g_d(\alpha) := f(x^* + \alpha d).$$

Assume that for every d, g_d is minimized by $\alpha = 0$. Prove that

(i) We have necessarily $\nabla f(x^*) = 0$.

Solution. If we select a standard basis vector as the direction, then

$$0 = g'_{e_i}(\alpha)\Big|_{\alpha=0} = \langle \nabla f(x^* + \alpha e_i), e_i \rangle\Big|_{\alpha=0} = \langle \nabla f(x^*), e_i \rangle$$

So the i-th entry of $\nabla f(x^*)$ is zero. As i was arbitrary, we are done. □

(ii) $f(x^*)$ is *not* necessarily a minimum of f.

Solution. Let $0 < p < q$ and define

$$f(y, z) := (z - py^2)(z - qy^2)$$

then for $p < m < q$ we have for $y \neq 0$

$$f(y, my^2) = (my^2 - py^2)(my^2 - qy^2) = \underbrace{y^2}_{>0}\underbrace{(m-p)}_{>0}\underbrace{(m-q)}_{<0} < 0.$$

This means that $f(0, 0) = 0$ is not a minimum since we can get arbitrarily close to $x^* = (0, 0)$ with (y, my^2) with smaller and smaller y.

Let $d = (d_1, d_2)$ be an arbitrary direction. Then

$$\begin{aligned} g_d(\alpha) &= (\alpha d_2 - p(\alpha d_1)^2)(\alpha d_2 - q(\alpha d_1)^2) \\ &= \alpha^2 (d_2 - p\alpha d_1^2)(d_2 - q\alpha d_1^2) \\ &= \alpha^2 (d_2^2 - \alpha d_2 (p+q) d_1^2 + \alpha^2 pq d_1^2) \end{aligned}$$

Thus $g'(0) = 0$ and $g''(0) = 2d_2^2 > 0$ which implies g has a minimum in zero. □

Exercise 3 (Convergence to Stationary Point)**.** **(0 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. $\nabla f(x_*) = 0$.

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

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

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

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

Exercise 4 (Sub-gradients)**.** **(0 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 $x \in \mathbb{R}^d$.

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. □

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

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. □

Exercise 5 (PL-Inequality)**.** **(0 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

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

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.

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

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

Exercises to be prepared at home for the next exercise class at 05.11.2026

Exercise 6 (Conjugate Gradient Descent)**.** **(0 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. Show that

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

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 $\mathcal{H}$

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 $(v_1, \dots v_d)$, 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:

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

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

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

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

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, 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".

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

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

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

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

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 $\mathcal{K}_{k-1}$.

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

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

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 $\nabla f(x_k)$.

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 $v_1, \ldots, v_{k-1} \in \mathcal{K}_{k-2}$. So only the last summand remains. □

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

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

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 α_{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

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

(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 α_k as CGD.

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 α_k, β_k. □