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

Sheet 4

For the exercise class on the 07.11.2024 at 12:00.
Hand in your solutions by 10:15 in the lecture on Tuesday 05.11.2024.

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

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

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

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

Draw a picture of this graph. (1 pt)

(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)

(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

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

(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)

(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

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

(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

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

(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

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

(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

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

(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)

(x) Bask in our glory! For we have proven that ...? Summarize our results into a theorem. (1 pt)

(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 μ-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

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

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

Hint. *Recall that $\nabla_H f(x)$ is the unique vector satisfying*

$$0 = \lim_{v\to 0} \frac{|f(x+v) - f(x) - \langle \nabla_H f(x), v\rangle_H|}{\|v\|_H}.$$

(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:

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

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. (2 pts)

(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)

(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)

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

(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)

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

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

Hint. *Proving*

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

should allow you to conclude $\langle \nabla f(x_k), v_k\rangle_h = -\frac{\|\nabla f(x_k)\|^2}{\alpha_{k-1}}$. *Then it makes sense to calculate*

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

by solving its optimization problem. Finally you may want to consider $v_k = \nabla f(x_{k-1}) - c v_{k-1}$ *and* $v_{k-1} \in \mathcal{K}_{k-2}$.

(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. (1 pt)

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

(ii) Recall, we required for convergence of HBM

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

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

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

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

(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)

(v) Implement Heavy-Ball momentum, Nesterov's momentum and CGD `https://classroom.github.com/a/DXlL27T4`. (6 pts)