Exercise Sheet: Gradient Descent Method

Exercises

Exercise 1 (Implementation for a quadratic function). Write a program (e.g. in MATLAB/Octave/Python) that implements gradient descent for the quadratic function

$$f(x) = \tfrac{1}{2} x^\top A x - b^\top x, \qquad A = \begin{pmatrix} 4 & 1 \\ 1 & 3 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 2 \end{pmatrix},$$

starting from $x^{(0)} = (0, 0)^\top$ with constant step size γ.

1. Derive the explicit formula for $\nabla f(x)$ and the unique minimizer x^*.
2. Run gradient descent for different step sizes $\gamma \in \{0.05, 0.1, 0.3, 0.6\}$ for 100 iterations. Plot $\|x^{(k)} - x^*\|$ versus k in a semilogarithmic plot.
3. Based on the eigenvalues of A, determine the range of constant step sizes for which gradient descent converges and compare with your experiments.

Exercise 2 (Effect of conditioning and step size). Consider the function

$$f(x) = x_1^2 + 10 x_2^2, \qquad x = (x_1, x_2)^\top \in \mathbb{R}^2.$$

1. Implement gradient descent with constant step size γ and starting point $x^{(0)} = (3, 3)^\top$.
2. For $\gamma \in \{0.01, 0.05, 0.1, 0.2\}$, plot the iterates $x^{(k)}$ in the (x_1, x_2)-plane together with level sets of f.
3. Observe how the trajectory changes with γ. For which values of γ does the method converge? For which values does it diverge or oscillate strongly?
4. Comment on the influence of the different curvature in the two directions (the conditioning of the Hessian) on the shape of the gradient descent path.

Exercise 3 (Backtracking line search). Consider the Rosenbrock function

$$f(x_1, x_2) = (1 - x_1)^2 + 100(x_2 - x_1^2)^2.$$

1. Implement gradient descent with *backtracking line search*: starting from a trial step size $\bar{\gamma} > 0$, decrease the step size by a factor $\beta \in (0, 1)$ until the Armijo condition is satisfied. Use for instance $\bar{\gamma} = 1$, $\beta = 0.5$, and $\sigma = 10^{-4}$.
2. Start from $x^{(0)} = (-1.2, 1)^\top$ and run the method until $\|\nabla f(x^{(k)})\| \leq 10^{-4}$ or a maximum number of iterations is reached. Plot $f(x^{(k)})$ versus k and the path of iterates in the plane.
3. Compare your results with gradient descent using a fixed step size $\gamma = 10^{-3}$ and $\gamma = 10^{-4}$. Discuss advantages and disadvantages of backtracking compared to using a fixed step size.

Exercise 4 (Stopping criteria and practical performance). Let $f : \mathbb{R}^n \to \mathbb{R}$ be given by

$$f(x) = \tfrac{1}{2}\|Bx - d\|_2^2,$$

where $B \in \mathbb{R}^{m\times n}$ and $d \in \mathbb{R}^m$ are given (you may generate B and d randomly, e.g. with independent standard normal entries).

1. Implement gradient descent with constant step size $\gamma = \frac{1}{\|B\|_2^2}$, where $\|B\|_2$ denotes the spectral norm (you may approximate it numerically).
2. Implement the following three stopping criteria:
   (a) $\|\nabla f(x^{(k)})\| \le \varepsilon$,
   (b) $\|x^{(k)} - x^{(k-1)}\| \le \varepsilon$,
   (c) $f(x^{(k-1)}) - f(x^{(k)}) \le \varepsilon$,

   with $\varepsilon = 10^{-6}$.
3. For each criterion, record the number of iterations and the final value $f(x^{(k)})$ when starting from the same $x^{(0)}$. Compare the behaviour of the stopping criteria and discuss which one you would prefer in practice and why.

Exercise 5 (Gradient descent with noisy gradients). Consider again the quadratic function from Exercise 17.8. Assume that the gradient is not available exactly, but only via a noisy oracle

$$g(x) = \nabla f(x) + \xi,$$

where ξ is a random vector whose entries are independent and normally distributed with mean 0 and variance σ^2.

1. Implement a *stochastic* gradient descent method

   $$x^{(k+1)} = x^{(k)} - \gamma_k g(x^{(k)}),$$

   with step sizes $\gamma_k = \frac{\gamma_0}{1+k}$ and some $\gamma_0 > 0$.
2. For different noise levels $\sigma \in \{0, 0.1, 0.5\}$ and a fixed γ_0, run the method from the same starting point $x^{(0)}$ and plot $f(x^{(k)}) - f(x^*)$ versus k (averaged over several runs).
3. Compare the convergence behaviour with the noise-free case ($\sigma = 0$) and discuss how noise and the choice of the decreasing step sizes influence convergence to the minimizer.