TUM EI 70360: Machine Learning and Optimization
Fall 2023
Lecturer: Reinhard Heckel
Teaching Assistant: Tobit Klug

Problem Set 4
Issued: Tuesday, Nov. 7, 2023
Due: Thursday, Nov. 16, 2023

Problem 1 (Stepsize of gradient descent for least squares)**.** Suppose we minimize the least squares objective $\hat{R}(\boldsymbol{\theta}) = (1/n)\sum_{i=1}^{n}(\mathbf{x}_i^T\boldsymbol{\theta} - y_i)^2$ with gradient descent. What is a good stepsize to choose and why?

Problem 2 (Gradient descent)**.** As discussed in class, gradient descent is a powerful and standard tool for finding local minima of general functions $f\colon \mathbb{R}^d \to \mathbb{R}$. Gradient descent initialized with $\mathbf{x}^0$ produces the iterates

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha_k \nabla f(\mathbf{x}^k),$$

where α_k is the stepsize. For convex functions, gradient descent provably finds the global minimum, provided the stepsize is chosen sufficiently small. If the stepsize is chosen too small, however, the convergence is slower thus it takes longer to find a minimum. In the absence of prior knowledge on the function class, it is common to choose a decaying stepsize. However, if the stepsize decays too fast, gradient descent might not converge either. In this problem we will examine the choice of stepsize.

1. Let $\mathbf{x}, \mathbf{b} \in \mathbb{R}^d$, and prove that $f(\mathbf{x}) = \|\mathbf{x} - \mathbf{b}\|_2$ is a convex function.

Let $\mathbf{b} = [4.5, 6]$, and consider gradient descent initialized with $\mathbf{x}^0 = \mathbf{0}$ applied to $f(\mathbf{x}) = \|\mathbf{x} - \mathbf{b}\|_2$. For the following choices of stepsizes, does gradient descent converge to the optimal solution? If yes, how many steps are required to take it within 1% of the optimal solution, i.e., how large does k need to be so that

$$\frac{\left\|\mathbf{x}^* - \mathbf{x}^k\right\|_2}{\|\mathbf{x}^*\|_2} \leq 0.01,$$

where $\mathbf{x}^*$ is the minimizer of $f(\mathbf{x})$? You can either prove your claim, or implement gradient descent and answer the question based on simulations. Specifically, you may plot the error $\left\|\mathbf{x}^* - \mathbf{x}^k\right\|_2 / \|\mathbf{x}^*\|_2$ versus the number of steps, and see if and when it is smaller than 0.01.

2. Constant stepsize of $\alpha_k = 1$
3. Decreasing stepsize of $(5/6)^k$
4. Decreasing stepsize of $1/(k+1)$

Next, apply gradient descent to the function $g(\mathbf{x}) = \|\mathbf{x} - \mathbf{b}\|_2^2$. For the following choices of stepsizes, does gradient descent applied to g converge to the optimal solution? If yes, how many steps are required to take it within 1% of the optimal solution:

5. Constant stepsize of $\alpha_k = 0.1$
6. Decreasing stepsize of $(1/6)^k$
7. Decreasing stepsize of $1/(4(k+1))$