ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Prof. D. Kressner
M. Steinlechner

1 ▶ Accelerated Gradient Descent

a) We consider the overdetermined linear system of equations $A\mathbf{x}^* + \mathbf{w} = \mathbf{b}$ with $A \in \mathbb{R}^{n\times p}$ a matrix with full column rank, $\mathbf{x} \in \mathbb{R}^p$ the unknown solution, and $\mathbf{w} \in \mathbb{R}^n$ denotes unknown noise. The least squares estimator for the unknown solution is then given by

$$\mathbf{x}^* = \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^p} f(\mathbf{x}), \quad \text{with } \ f(\mathbf{x}) := \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^p} \|A\mathbf{x} - \mathbf{b}\|_2^2.$$

Show that f is a μ-strongly convex function with μ given by the smallest eigenvalue of $A^\mathsf{T} A$. Furthermore, show that the gradient of f is Lipschitz continuous with the smallest Lipschitz constant L given by the largest eigenvalue of $A^\mathsf{T} A$.

b) Implement Algorithm 4.28, *Accelerated Gradient Descent*. Use the provided template on the lecture homepage, `AGD.m`

c) We will now investigate the convergence rates of standard Gradient Descent (GD) and Accelerated Gradient Descent (AGD) when applied to the cost function f. An implementation of GD is available on the lecture homepage, `GD.m`. Using the provided template `ex9problem1.m`, plot the convergence rates of GD for the optimal choice $\alpha = \frac{2}{L+\mu}$ which exploits the strong convexity of the function. Compare to the convergence rate of AGD using your implementation from **b)**. Try out how the convergence speed changes as you use non-optimal values for L and μ.

Compare the obtained results with the convergence rates given by Theorem 4.23 and 4.29 in the lecture notes.

2 ▶ Binary logistic regression, part 2

In the last exercise sheet, we have seen the *binary logistic regression* approach. Let $\mathbf{a}_i \in \mathbb{R}^n$ be sampling points and b_i be the associated binary class labels. Then, we want to determine the maximum log-likelyhood estimator

$$\operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^p} f(\mathbf{x}), \quad \text{with } f(\mathbf{x}) = -\sum_{i=1}^{n} \log\left(h(b_i \mathbf{a}_i^\mathsf{T} \mathbf{x})\right),$$

where $h(t) = 1/(1+\exp(t))$ is the sigmoid function. In the last exercise sheet, we have shown that the function is convex, but in general not strongly convex, depending on the data matrix $A = \begin{bmatrix}\mathbf{a}_1 & \mathbf{a}_2 & \dots & \mathbf{a}_n\end{bmatrix}^\mathsf{T}$.

Hence, we introduce a regularization term, such that the minimization problem is now given as

$$x^* = \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^p} f_\sigma(\mathbf{x}), \quad \text{with } f_\sigma(\mathbf{x}) = f(\mathbf{x}) + \frac{\sigma}{2}\|x\|_2^2.$$

a) Show that the regularized function f_σ is σ-strongly convex.

b) Show that the gradient of f_σ is Lipschitz continuous with $L = \frac{1}{4}\|A\|_2 + \sigma$.

c) We will now test the performance of GD and AGD applied to binary logistic regression. For simplicity, we choose the regulatization parameter $\sigma = 1$. As a dataset, we will use the `a4a` dataset from `http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html`, a processed version of the "Adult" dataset. This dataset aims to predict from 4781 samples from the1994 US census the probability of a person to earn over 50000 USD per year given 123 attributes such as age, sex, education, etc. Complete the template `ex9problem2.m` using your implementation of AGD from the previous question and the prepared dataset `adult.mat` from the lecture homepage.

3 ► Tangent cones

For the following sets Ω and points $\mathbf{x} \in \Omega$, sketch the tangent cones $T_\Omega(\mathbf{x})$, using Definition 5.4 in the lecture notes.

(a) $\mathbf{x} = (0,1)^T, \quad \Omega = \{\mathbf{x} \in \mathbb{R}^2 \colon (|x_1| + |x_2| - 1)(x_2 - 1) = 0\}$,

(b) $\mathbf{x} = (-1,0,0)^T, \quad \Omega = \{\mathbf{x} \in \mathbb{R}^3 \colon x_1^2 + x_2^2 \leq 1,\ x_3^2 = 0\}$.