Prof. D. Kressner
M. Steinlechner

1 ► Nonlinear CG

Consider a symmetric positive definite matrix A. The smallest eigenvalue is given by the global minimum of the *Rayleigh quotient*:

$$f(\mathbf{x}) = \frac{\mathbf{x}^T A\mathbf{x}}{\mathbf{x}^T\mathbf{x}}.$$

In this exercise we want to solve this important problem using nonlinear CG with Armijo backtracking. The gradient of f at $\mathbf{x}$ is (see also Exercise Sheet 2)

$$\nabla f(\mathbf{x}) = 2\Big(I - \frac{\mathbf{x}\mathbf{x}^T}{\mathbf{x}^T\mathbf{x}}\Big)\frac{A\mathbf{x}}{\mathbf{x}^T\mathbf{x}} = 2(A - f(\mathbf{x})I)\frac{\mathbf{x}}{\mathbf{x}^T\mathbf{x}}.$$

Use the MATLAB template from the homepage and provide the function

$$[\mathtt{X}, \mathtt{fX}, \mathtt{dfX}] = \mathtt{ncg}(\mathtt{f}, \mathtt{df}, \mathtt{x0}, \mathtt{c1}, \mathtt{alpha0}, \mathtt{beta}, \mathtt{tol}, \mathtt{maxiter}, \mathtt{opt}).$$

As usual, `f`, `df` are the function handles for f, ∇f; `x0` is the starting value, and `c1, alpha0, beta` are the parameters of the Armijo backtracking. The iteration should stop when $\|\nabla f(x_n)\| \leq$ `tol` or `maxiter` iterations have been made. The parameter `opt` for the choice of β_k should be a string: `'fr'` for Fletcher-Reeves, `'pr+'` for *modified Polak-Ribière*

$$\beta_k^{PR+} = \max(0, \beta_k^{PR}).$$

The output matrices `X`, `fX`, `dfX` contain all generated iterates, function values, and gradients as columns, respectively.

The template from the homepage compares Fletcher-Reeves CG, Polak-Ribière CG and steepest descent by plotting the error and the norm of the gradients against the number of iterations on a semilogarithmic scale. As A, a $10^2 \times 10^2$ discretization of the 2D Laplace operator is taken, and as starting value the first unit vector. Take instead the starting vector consisting of all ones. What do you observe?

Next, try the more difficult 10×10 prolate matrix from the MATLAB gallery (uncomment the corresponding lines in the template) with both starting values.

2 ► Convex functions

Prove the following simple statements for μ-strongly convex functions.

a) *(Third relation in Lemma 4.22)* Let f be twice differentiable and let $H(\mathbf{x})$ denote the Hessian of f. Then $H(\mathbf{x}) \succcurlyeq \mu I$ if and only if f is μ-strongly convex.

b) Show that for a differentiable μ-strongly convex function, the distance $\|\mathbf{x} - \mathbf{x}^*\|_2$ from the point $\mathbf{x}$ to the minimizer $\mathbf{x}^*$ can be bounded solely by the norm of the gradient, $\|\nabla f(\mathbf{x})\|_2$:

$$\|\mathbf{x} - \mathbf{x}^*\|_2 \leq \frac{2}{\mu}\|\nabla f(\mathbf{x})\|_2.$$

3 ► Binary logistic regression

Logistic regression is an important tool in statistics and has various applications in machine learning and data mining for the classification of data.
The binary logistic model with parameter $\hat{\mathbf{x}} \in \mathbb{R}^p$ yields the probability of the class $b \in \{-1, 1\}$ given a certain sample $\mathbf{a} \in \mathbb{R}^n$:

$$\mathbb{P}(b \,|\, \mathbf{a}) = \frac{1}{1 + \exp(-b\mathbf{a}^T\hat{\mathbf{x}})}$$

Unfortunately, the parameter $\hat{\mathbf{x}}$ is usually unknown and we have to estimate it from data samples. Let $\mathbf{a}_i \in \mathbb{R}^n$ be sampling points and b_i be the associated binary class labels. Then, an approximation of the true parameter $\hat{\mathbf{x}}$ is given by the maximum log-likelyhood estimator

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

where $h(t) = 1/(1+\exp(t))$ is the sigmoid function. Binary classification can hence be cast into an unconstrained optimization problem for the model parameters $\mathbf{x}^*$ *(Note that we have introduced a minus sign to go from a maximization problem to a minimization problem).*

a) Show that for a given data set $\{(\mathbf{a}_1, b_1), (\mathbf{a}_2, b_2), \ldots, (\mathbf{a}_n, b_n)\}$, the objective function f is convex.

b) Is f strongly convex?

c) Show that the Hessian of f is bounded for all $\mathbf{x} \in \mathbb{R}^p$: $\|H(x)\|_2 < C$.

d) What is the smallest Lipschitz constant $L > 0$ you can find such that the gradient ∇f is Lipschitz continuous,

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^p \ ?$$

Hint: *The gradient of f is given by*

$$\nabla f(\mathbf{x}) = -\sum_{i=1}^{n} \left(1 - h(b_i \mathbf{a}_i^T \mathbf{x})\right) b_i \mathbf{a}_i,$$

and the Hessian by

$$H(\mathbf{x}) = A^T D_{\mathbf{x}} A,$$

with the data matrix $A = \begin{bmatrix}\mathbf{a}_1 & \mathbf{a}_2 & \ldots & \mathbf{a}_n\end{bmatrix}^T$ and the diagonal matrix

$$D_{\mathbf{x}} = \operatorname{diag}\left(h(b_1 \mathbf{a}_1^T \mathbf{x})\left(1 - h(b_1 \mathbf{a}_1^T \mathbf{x})\right) \,,\ \ldots\ ,\ h(b_n \mathbf{a}_n^T \mathbf{x})\left(1 - h(b_n \mathbf{a}_n^T \mathbf{x})\right) \right)$$