SOLUTION 8 – Advanced Numerical Analysis

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Prof. D. Kressner
M. Steinlechner

1 ▶ Nonlinear CG

```
function ex9problem2
close all;
clear all

% Poisson matrix
% --------------
n = 10;
I = eye(n^2);
A = gallery('poisson',n);

% Choose a starting value
x0 = [1; zeros(n^2-1,1)];
 %x0 = ones(n^2,1);


% Prolate matrix
% --------------
% n = 10;
% I = eye(n);
% A = gallery('prolate',n);

% Choose a starting value
%x0 = [1; zeros(n-1,1)];
% x0 = ones(n,1);

% Calculate exact solution:
Xexact = min(eig(A));

% Define function handles here
f = @(x) x'*(A*x)/(x'*x);
df = @(x) 2*(I - x*x'/(x'*x))*(A*x)/(x'*x);

% Line search parameters
alpha0 = 1;
beta = 0.5;
c1 = 1e-4;
tol = 1e-8;
maxiter = 10000;

%Run Fletcher-Reeves
[X1,fX1,dfX1] = ncg(f,df,x0,c1,alpha0,beta,tol,maxiter,'fr');

%Run Polak-Ribiere
[X2,fX2,dfX2] = ncg(f,df,x0,c1,alpha0,beta,tol,maxiter,'pr+');

%Run steepest descent
[X3,fX3,dfX3] = steepdesc(f,df,x0,c1,alpha0,beta,tol,maxiter);

%Plot function value and gradient
subplot(1,2,1)
semilogy([0:1:numel(fX1)-1],diag(dfX1'*dfX1).^(1/2),'ro');
hold on
title('Gradient norm')
subplot(1,2,2);
semilogy([0:1:numel(fX1)-1],abs(fX1-Xexact),'ro');
hold on
title('Error on cost function')
subplot(1,2,1)
semilogy([0:1:numel(fX2)-1],diag(dfX2'*dfX2).^(1/2),'ob');
subplot(1,2,2);
semilogy([0:1:numel(fX2)-1],abs(fX2-Xexact),'ob');
```

```
subplot(1,2,1)
semilogy([0:1:numel(fX3)-1],diag(dfX3'*dfX3).^(1/2),'ok');
legend('FR','PR+','SD')
subplot(1,2,2);
semilogy([0:1:numel(fX3)-1],abs(fX3-Xexact),'ok');
legend('FR','PR+','SD')



end

%Required functions

function [X,fX,dfX] = ncg(f,df,x0,c1,alpha0,beta,tol,maxiter,opt)

X(:,1) = x0;
fX(:,1) = f(x0);
dfX(:,1) = df(x0);
k = 1;
xk = x0;
gk = df(xk);
pk = -gk;

while norm(gk) > tol && k<=maxiter

    % start backtracking
    alpha = alpha0;
    while f(xk + alpha*pk) > f(xk) + c1*alpha*gk'*pk
        alpha = alpha*beta;
    end

    %Perform step
    xk = xk + alpha*pk;

    %new gradient
    gknew = df(xk);

    %new search directions
    if strcmp(opt,'fr')
        %Fletcher-Reeves
        betak = norm(gknew)^2 / norm(gk)^2;
    else
        %Polak-Ribiere +
        betak = max(0,gknew'*(gknew - gk) / norm(gk)^2);
    end
    pk = -gknew + betak*pk;

    %continue with
    gk = gknew;
    X(:,k+1) = xk;
    fX(:,k+1) = f(xk);
    dfX(:,k+1) = gk;
    k = k+1;
end
end


function [X,fX,dfX] = steepdesc(f,df,x0,c1,alpha0,beta,tol,maxiter)

X(:,1) = x0;
fX(:,1) = f(x0);
dfX(:,1) = df(x0);
k = 1;
xk = x0;
gk = dfX(:,1);

while norm(gk) > tol && k<=maxiter

    %search directions
```

```
    pk = -gk;

    % start backtracking
    alpha = alpha0;
    while f(xk + alpha*pk) > f(xk) + c1*alpha*gk'*pk
        alpha = alpha*beta;
    end

    %Perform step
    xk = xk + alpha*pk;
    gk = df(xk);
    X(:,k+1) = xk;
    fX(:,k+1) = f(xk);
    dfX(:,k+1) = gk;
    k = k+1;
end

end
```

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.

a) $\Rightarrow$: Let $H(x) \geq \mu I$. Then, f is strongly convex.

Proof: For $x, y \in \mathbb{R}^n$ we have the Taylor-representation

$$f(y) = f(x) + \nabla f(x)^T (y-x) + \frac{1}{2}(y-x)^T H(\xi)(y-x)$$

for some ξ on the line segment between x and y ($\rightsquigarrow$ mean value theorem)

If $H(x) \geq \mu I$, then we directly obtain

$$f(y) \geq f(x) + \nabla f(x)^T (y-x) + \frac{\mu}{2}\|y-x\|_2^2 \qquad \square$$

$\Leftarrow$: If f strongly convex, then $H(x) \geq \mu I$.

Proof: Analogous to Proof of Lemma 4.19, step (4.34) $\Rightarrow$ (4.35), where we replace $x_t = x + ts$

$$\frac{\mu}{2} \leq \frac{1}{t^2}(\nabla f(x_t) - \nabla f(x))^T (x_t - x) \xrightarrow{t \to 0} s^T H(x) s,$$

Hence there can be no eigenvalue of $H(x)$ smaller than μ, see Rayleigh quotient for symmetric matrices. $\square$

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, $\|\nabla f(\mathbf{x})\|_2$:

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

b) Using the strong convexity of f at the point $y = x^*$, we obtain

$$f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x) + \frac{\mu}{2} \|x^* - x\|_2^2$$

$$\geq f(x) - \|\nabla f(x)\|_2 \|x^* - x\|_2 + \frac{\mu}{2} \|x^* - x\|_2^2$$

(Cauchy-Schwarz)

Since x^* is optimal, we have $f(x^*) \leq f(x)$ and therefore

$$0 \geq f(x^*) - f(x) \geq -\|\nabla f(x)\|_2 \|x^* - x\|_2 + \frac{\mu}{2} \|x^* - x\|_2^2$$

$$\Leftrightarrow \quad \|\nabla f(x)\|_2 \|x^* - x\|_2 \geq \frac{\mu}{2} \|x^* - x\|_2^2$$

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

□

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

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

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.

To show that the function

$$f(x) = -\sum_{i=1}^{n} \log\big(h(b_i \mathbf{a}_i^T \mathbf{x})\big)$$

is convex, we will prove that the Hessian of f is positive semidefinite. The Hessian is given on the exercise sheet,

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

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

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

As $h(t) = \frac{1}{1+\exp(-t)}$*, we have that* $g(t) := h(t)(1-h(t)) = \frac{\exp(-t)}{(1+\exp(-t))^2}$*, and it is easy to show that*

$$0 < g(t) \leq \frac{1}{4} < 1, \quad \forall t \in \mathbb{R}.$$

With this definition of g*, we have*

$$D_{\mathbf{x}} = \operatorname{diag}\left(g(b_1\mathbf{a}_1^T\mathbf{x}) \ , \ \dots \ , \ g(b_n\mathbf{a}_n^T\mathbf{x})\right)$$

As g *is strictly positive, we have that* $D_{\mathbf{x}}$ *is a positive definite diagonal matrix. Thus, the square root of it exists, with*

$$D^{\frac{1}{2}} = \operatorname{diag}\left\{\sqrt{g\left(b_i\mathbf{a}_i^T\mathbf{x}\right)}\right\}_{i=1}^{n}.$$

To show that the Hessian $H(x)$ *is positive semidefinite, we need to show that for any vector* $y \in \mathbb{R}^n$ *with* $y \neq 0$ *it holds that*

$$y^T H(x) y \geq 0 \quad \Leftrightarrow \quad y^T A^T D_{\mathbf{x}} A y \geq 0 \quad \Leftrightarrow \quad y^T A^T D^{\frac{1}{2}} D^{\frac{1}{2}} A y \geq 0$$
$$\Leftrightarrow \quad y^T A^T D^{\frac{1}{2}} D^{\frac{1}{2}} A y \geq 0 \quad \Leftrightarrow \quad (D^{\frac{1}{2}} A y)^T (D^{\frac{1}{2}} A y) \geq 0$$
$$\Leftrightarrow \quad \|D^{\frac{1}{2}} A y\|_2 \geq 0.$$

which is clearly true as the norm is always nonnegative.

b) Is f strongly convex?

For f *to be strongly convex, it is necessary (but not sufficient!) that the strict inequality*

$$y^T H(x) y > 0 \quad \Leftrightarrow \quad \|D^{\frac{1}{2}} A y\|_2 > 0.$$

has to hold for all vectors $y \in \mathbb{R}^n$ *with* $y \neq 0$*. In general, the data matrix* A *is not invertible, as its rows (the sample vectors) are not necessarily linearly independent, that is, it can happen that* $\operatorname{rank}(A) < \min\{n, p\}$*. Hence, it may have a non-empty kernel, that is, we can find a* $z \neq 0$*,*

$$z \in \ker(A) \quad \Leftrightarrow \quad Az = 0.$$

and thus also

$$\|D^{\frac{1}{2}} A z\|_2 = 0.$$

Hence, f *is* in general *not strictly convex and thus of course also not strongly convex.*

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

To show that $H(x)$ *is bounded, we use the decomposition of* $H(x)$ *introduced in Exercise 11 above and the fact that all induced matrix norms are submultiplicative, that is,*

$$\|H(x)\| = \|A D_{\mathbf{x}} A\| \leq \|A\| \|D_{\mathbf{x}}\| \|A\| = \|D_{\mathbf{x}}\| \|A\|^2.$$

The diagonal matrix $D_{\mathbf{x}}$ *has entries*

$$g\left(b_i\mathbf{a}_{(i)}^T\mathbf{x}\right)$$

on the diagonal. As $0 < g(t) < \frac{1}{4}$ *for all* $t \in \mathbb{R}$*, we have*

$$\|D_{\mathbf{x}}\| = \lambda_{\max}(D_{\mathbf{x}}) = \max_i g\left(b_i\mathbf{a}_{(i)}^T\mathbf{x}\right) \leq \frac{1}{4}.$$

Hence, the Hessian is bounded by

$$\|H(x)\| \leq \frac{1}{4}\|A\|^2$$

independent of x*, as* A *are the training data samples.*

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

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

A differentiable function is Lipschitz continuous if and only if its derivative is bounded. In this case, the gradient ∇f *is Lipschitz if the second derivative is bounded. This was shown in* ***c)****, with* $L = \frac{1}{4}\|A\|^2$ *a possible Lipschitz constant.*