Conjugate Direction minimization

Lectures for PHD course on
Unconstrained Numerical Optimization

Enrico Bertolazzi

DIMS – Università di Trento

May 2008

Outline

Generic minimization algorithm

In the following we study the convergence rate of the Generic minimization algorithm applied to a quadratic function $\mathsf{q}(\boldsymbol{x})$ with exact line search. The function

$$\mathsf{q}(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{x} + c$$

can be viewed as a n-dimensional generalization of the 1-dimensional parabolic model.

Generic minimization algorithm

Given an initial guess $\boldsymbol{x}_0$, let $k = 0$;
while not converged **do**
  Find a descent direction $\boldsymbol{p}_k$ at $\boldsymbol{x}_k$;
  Compute a step size α_k using a line-search along $\boldsymbol{p}_k$.
  Set $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k\boldsymbol{p}_k$ and increase k by 1.
end while

Assumption (Symmetry)

The matrix $\boldsymbol{A}$ is assumed to be symmetric, in fact,

$$\boldsymbol{A} = \boldsymbol{A}^{Symm} + \boldsymbol{A}^{Skew}$$

where

$$\boldsymbol{A}^{Symm} = \frac{1}{2}[\boldsymbol{A} + \boldsymbol{A}^T], \qquad \boldsymbol{A}^{Symm} = (\boldsymbol{A}^{Symm})^T$$

$$\boldsymbol{A}^{Skew} = \frac{1}{2}[\boldsymbol{A} - \boldsymbol{A}^T], \qquad \boldsymbol{A}^{Skew} = -(\boldsymbol{A}^{Skew})^T$$

moreover

$$\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{A}^{Symm}\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{A}^{Skew}\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{A}^{Symm}\boldsymbol{x}$$

so that only the symmetric part of $\boldsymbol{A}$ contribute to $\mathsf{q}(\boldsymbol{x})$.

Assumption (SPD)

The matrix $\boldsymbol{A}$ is assumed to be symmetric and positive definite, in fact,

$$\nabla \mathsf{q}(\boldsymbol{x})^T = \frac{1}{2}(\boldsymbol{A}+\boldsymbol{A}^T)\boldsymbol{x}-\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}$$

and

$$\nabla^2 \mathsf{q}(\boldsymbol{x}) = \frac{1}{2}(\boldsymbol{A}+\boldsymbol{A}^T) = \boldsymbol{A}$$

From the sufficient condition for a minimum we have that $\nabla \mathsf{q}(\boldsymbol{x}_\star)^T = \boldsymbol{0}$, i.e.

$$\boldsymbol{A}\boldsymbol{x}_\star = \boldsymbol{b}$$

and $\nabla^2 \mathsf{q}(\boldsymbol{x}_\star) = \boldsymbol{A}$ is SPD.

The toy problem (1/3)

- In the following we study the convergence rate of the Steepest Descent and Conjugate Gradient methods applied to

$$\mathsf{q}(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{x} + c$$

  where $\boldsymbol{A}$ is an SPD matrix.
- This assumption simplify the analysis but it is also useful in the non linear case. In fact, by expanding a generic function $\mathsf{f}(\boldsymbol{x})$ near its minimum $\boldsymbol{x}_\star$ we have

$$\begin{aligned}\mathsf{f}(\boldsymbol{x}) = \mathsf{f}(\boldsymbol{x}_\star) + \nabla\mathsf{f}(\boldsymbol{x}_\star)(\boldsymbol{x}-\boldsymbol{x}_\star)\\ +\frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_\star)^T\nabla^2\mathsf{f}(\boldsymbol{x}_\star)(\boldsymbol{x}-\boldsymbol{x}_\star) + \mathcal{O}(\|\boldsymbol{x}-\boldsymbol{x}_\star\|^3)\end{aligned}$$

The toy problem (2/3)

- By setting

$$\begin{aligned}\boldsymbol{A} &= \nabla^2\mathsf{f}(\boldsymbol{x}_\star),\\ \boldsymbol{b} &= \nabla^2\mathsf{f}(\boldsymbol{x}_\star)\boldsymbol{x}_\star - \nabla\mathsf{f}(\boldsymbol{x}_\star)\\ c &= \mathsf{f}(\boldsymbol{x}_\star) - \nabla\mathsf{f}(\boldsymbol{x}_\star)\boldsymbol{x}_\star + \frac{1}{2}\boldsymbol{x}_\star^T\nabla^2\mathsf{f}(\boldsymbol{x}_\star)\boldsymbol{x}_\star\end{aligned}$$

  we have

$$\mathsf{f}(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{x} + c + \mathcal{O}(\|\boldsymbol{x}-\boldsymbol{x}_\star\|^3)$$

- So that we expect that when an iterate $\boldsymbol{x}_k$ is near $\boldsymbol{x}_\star$ then we can neglect $\mathcal{O}(\|\boldsymbol{x}-\boldsymbol{x}_\star\|^3)$ and the asymptotic behavior is the same of the quadratic problem.

The toy problem (3/3)

- we can rewrite the quadratic problem in many different way as follows

$$\begin{aligned}\mathsf{q}(\boldsymbol{x}) &= \frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_\star)^T\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{x}_\star) + c'\\ &= \frac{1}{2}(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b})^T\boldsymbol{A}^{-1}(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}) + c'\end{aligned}$$

  where

$$c' = c + \frac{1}{2}\boldsymbol{x}_\star^T\boldsymbol{A}\boldsymbol{x}_\star$$

- This last forms are useful in the study of the steepest descent method.

Outline

The steepest descent for quadratic functions (1/3)

The steepest descent minimization algorithm

Given an initial guess $\boldsymbol{x}_0$, let $k = 0$;
while not converged **do**
Choose as descent direction $\boldsymbol{p}_k = -\nabla \mathsf{q}(\boldsymbol{x}_k)^T = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k$;
Compute a step size α_k using a line-search along $\boldsymbol{p}_k$.
Set $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k$ and increase k by 1.
end while

Definition (Residual)

The expressions

$$\boldsymbol{r}(\boldsymbol{x}) = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}, \qquad \boldsymbol{r}_k = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k$$

are called the residual. We obviously have $\boldsymbol{r}(\boldsymbol{x}) = -\nabla \mathsf{q}(\boldsymbol{x})^T$ *and* $\boldsymbol{r}(\boldsymbol{x}_\star) = \boldsymbol{0}$.

The steepest descent for quadratic functions (2/3)

Lemma

The solution of the minimization problem:

$$\alpha_k = \arg\min_{\alpha \geq 0} \mathsf{q}(\boldsymbol{x}_k - \alpha \boldsymbol{r}_k) \qquad \textit{is} \qquad \alpha_k = -\frac{\boldsymbol{r}_k^T \boldsymbol{r}_k}{\boldsymbol{r}_k^T \boldsymbol{A} \boldsymbol{r}_k}.$$

Proof.

Because $p(\alpha) = \mathsf{q}(\boldsymbol{x}_k - \alpha \boldsymbol{r}_k)$ the minimum is a stationary point:

$$\begin{aligned}\frac{\mathrm{d}p(\alpha)}{\mathrm{d}\alpha} = \frac{\mathrm{d}\mathsf{q}(\boldsymbol{x}_k - \alpha \boldsymbol{r}_k)}{\mathrm{d}\alpha} &= -\nabla \mathsf{q}(\boldsymbol{x}_k - \alpha \boldsymbol{r}_k)\boldsymbol{r}_k \\ &= \boldsymbol{r}(\boldsymbol{x}_k - \alpha \boldsymbol{r}_k)^T \boldsymbol{r}_k = \big(\boldsymbol{b} - \boldsymbol{A}(\boldsymbol{x}_k - \alpha \boldsymbol{r}_k)\big)^T \boldsymbol{r}_k \\ &= \big(\boldsymbol{r}_k + \alpha \boldsymbol{A} \boldsymbol{r}_k\big)^T \boldsymbol{r}_k = 0\end{aligned}$$

and solving for α the result follows. □

The steepest descent for quadratic functions (3/3)

The steepest descent minimization algorithm

Given an initial guess $\boldsymbol{x}_0$, let $k = 0$;
while not converged **do**
Compute $\boldsymbol{r}_k = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k$;
Compute the step size $\alpha_k = \dfrac{\boldsymbol{r}_k^T \boldsymbol{r}_k}{\boldsymbol{r}_k^T \boldsymbol{A} \boldsymbol{r}_k}$;
Set $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \boldsymbol{r}_k$ and increase k by 1.
end while

Or more compactly

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \frac{\boldsymbol{r}_k^T \boldsymbol{r}_k}{\boldsymbol{r}_k^T \boldsymbol{A} \boldsymbol{r}_k} \boldsymbol{r}_k$$

The steepest descent reduction step (1/4)

The next lemma bound the reduction of $\mathsf{q}(\boldsymbol{x}_{k+1})$ by the value of $\mathsf{q}(\boldsymbol{x}_k)$:

Lemma

Consider the steepest descent for quadratic function, than we have the following estimate

$$\|\boldsymbol{x}_\star - \boldsymbol{x}_{k+1}\|_{\boldsymbol{A}}^2 = \|\boldsymbol{x}_\star - \boldsymbol{x}_k\|_{\boldsymbol{A}}^2 \left(1 - \frac{(\boldsymbol{r}_k^T\boldsymbol{r}_k)^2}{(\boldsymbol{r}_k^T\boldsymbol{A}^{-1}\boldsymbol{r}_k)(\boldsymbol{r}_k^T\boldsymbol{A}\boldsymbol{r}_k)}\right)$$

where

$$\|\boldsymbol{x}\|_{\boldsymbol{A}} = \sqrt{\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}}$$

is the energy norm induced by the SPD matrix $\boldsymbol{A}$.

The steepest descent reduction step (2/4)

Proof. (1/3).

We want bound $\mathsf{q}(\boldsymbol{x}_{k+1})$ by $\mathsf{q}(\boldsymbol{x}_k)$:

$$\begin{aligned}
\mathsf{q}(\boldsymbol{x}_{k+1}) &= \mathsf{q}(\boldsymbol{x}_k + \alpha_k \boldsymbol{r}_k) \\
&= \frac{1}{2}(\boldsymbol{A}\boldsymbol{x}_k + \alpha_k\boldsymbol{A}\boldsymbol{r}_k - \boldsymbol{b})^T \boldsymbol{A}^{-1}(\boldsymbol{A}\boldsymbol{x}_k + \alpha_k\boldsymbol{A}\boldsymbol{r}_k - \boldsymbol{b}) + c' \\
&= \frac{1}{2}(\alpha_k\boldsymbol{A}\boldsymbol{r}_k - \boldsymbol{r}_k)^T \boldsymbol{A}^{-1}(\alpha_k\boldsymbol{A}\boldsymbol{r}_k - \boldsymbol{r}_k) + c' \\
&= \frac{1}{2}\boldsymbol{r}_k^T\boldsymbol{A}^{-1}\boldsymbol{r}_k + \frac{1}{2}\alpha_k^2\boldsymbol{r}_k^T\boldsymbol{A}\boldsymbol{r}_k - \alpha_k\boldsymbol{r}_k^T\boldsymbol{r}_k + c' \\
&= \mathsf{q}(\boldsymbol{x}_k) + \frac{1}{2}\alpha_k\left(\alpha_k\boldsymbol{r}_k^T\boldsymbol{A}\boldsymbol{r}_k - 2\boldsymbol{r}_k^T\boldsymbol{r}_k\right)
\end{aligned}$$

The steepest descent reduction step (3/4)

Proof. (2/3).

Substituting $\alpha_k = \dfrac{\boldsymbol{r}_k^T\boldsymbol{r}_k}{\boldsymbol{r}_k^T\boldsymbol{A}\boldsymbol{r}_k}$ we obtain

$$\mathsf{q}(\boldsymbol{x}_{k+1}) = \mathsf{q}(\boldsymbol{x}_k) - \frac{1}{2}\frac{(\boldsymbol{r}_k^T\boldsymbol{r}_k)^2}{\boldsymbol{r}_k^T\boldsymbol{A}\boldsymbol{r}_k}$$

this shows that the steepest descent method reduce at each step the objective function $\mathsf{q}(\boldsymbol{x})$.

Using the expression $\mathsf{q}(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{r}(\boldsymbol{x})^T\boldsymbol{A}^{-1}\boldsymbol{r}(\boldsymbol{x}) + c'$ we can write:

$$\frac{1}{2}\boldsymbol{r}_{k+1}^T\boldsymbol{A}^{-1}\boldsymbol{r}_{k+1} = \frac{1}{2}\boldsymbol{r}_k^T\boldsymbol{A}^{-1}\boldsymbol{r}_k - \frac{1}{2}\frac{(\boldsymbol{r}_k^T\boldsymbol{r}_k)^2}{\boldsymbol{r}_k^T\boldsymbol{A}\boldsymbol{r}_k}$$

The steepest descent reduction step (4/4)

Proof. (3/3).

or better

$$\boldsymbol{r}_{k+1}^T\boldsymbol{A}^{-1}\boldsymbol{r}_{k+1} = \boldsymbol{r}_k^T\boldsymbol{A}^{-1}\boldsymbol{r}_k\left(1 - \frac{(\boldsymbol{r}_k^T\boldsymbol{r}_k)^2}{(\boldsymbol{r}_k^T\boldsymbol{A}^{-1}\boldsymbol{r}_k)(\boldsymbol{r}_k^T\boldsymbol{A}\boldsymbol{r}_k)}\right)$$

noticing that $\boldsymbol{r}_k = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k = \boldsymbol{A}\boldsymbol{x}_\star - \boldsymbol{A}\boldsymbol{x}_k = \boldsymbol{A}(\boldsymbol{x}_\star - \boldsymbol{x}_k)$ we have

$$\|\boldsymbol{x}_\star - \boldsymbol{x}_{k+1}\|_{\boldsymbol{A}}^2 = \|\boldsymbol{x}_\star - \boldsymbol{x}_k\|_{\boldsymbol{A}}^2 \left(1 - \frac{(\boldsymbol{r}_k^T\boldsymbol{r}_k)^2}{(\boldsymbol{r}_k^T\boldsymbol{A}^{-1}\boldsymbol{r}_k)(\boldsymbol{r}_k^T\boldsymbol{A}\boldsymbol{r}_k)}\right)$$

where

$$\|\boldsymbol{x}\|_{\boldsymbol{A}} = \sqrt{\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}}$$

is the energy norm induced by the SPD matrix $\boldsymbol{A}$.

The estimate of the convergence rate for the steepest descent method is linked to the estimate of the term

$$\frac{(\boldsymbol{r}_k^T\boldsymbol{r}_k)^2}{(\boldsymbol{r}_k^T\boldsymbol{A}^{-1}\boldsymbol{r}_k)(\boldsymbol{r}_k^T\boldsymbol{A}\boldsymbol{r}_k)}$$

in particular we can prove

Lemma (Kantorovic)

Let $\boldsymbol{A}\in\mathbb{R}^{n\times n}$ an SPD matrix then the following inequality is valid

$$1\leq\frac{(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x})(\boldsymbol{x}^T\boldsymbol{A}^{-1}\boldsymbol{x})}{(\boldsymbol{x}^T\boldsymbol{x})^2}\leq\frac{(M+m)^2}{4\,M\,m}$$

for all $\boldsymbol{x}\neq\boldsymbol{0}$. Where $m=\lambda_1$ is the smallest eigenvalue of $\boldsymbol{A}$ and $M=\lambda_n$ is the biggest eigenvalue of $\boldsymbol{A}$.

Proof. (1/5).

STEP 1: problem reformulation. First of all notice that

$$\frac{(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x})(\boldsymbol{x}^T\boldsymbol{A}^{-1}\boldsymbol{x})}{(\boldsymbol{x}^T\boldsymbol{x})^2}=\frac{(\boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y})(\boldsymbol{y}^T\boldsymbol{A}^{-1}\boldsymbol{y})}{(\boldsymbol{y}^T\boldsymbol{y})^2}$$

for all $\boldsymbol{y}=\alpha\boldsymbol{x}$ with $\alpha\neq 0$. Choosing $\alpha=\|\boldsymbol{x}\|^{-1}$ have:

$$\min_{\|\boldsymbol{z}\|=1}(\boldsymbol{z}^T\boldsymbol{A}\boldsymbol{z})(\boldsymbol{z}^T\boldsymbol{A}^{-1}\boldsymbol{z})\leq\frac{(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x})(\boldsymbol{x}^T\boldsymbol{A}^{-1}\boldsymbol{x})}{(\boldsymbol{x}^T\boldsymbol{x})^2}\leq\max_{\|\boldsymbol{z}\|=1}(\boldsymbol{z}^T\boldsymbol{A}\boldsymbol{z})(\boldsymbol{z}^T\boldsymbol{A}^{-1}\boldsymbol{z})$$

Proof. (2/5).

STEP 2: eigenvector expansions. Matrix $\boldsymbol{A}\in\mathbb{R}^{n\times n}$ is an SPD matrix so that there exists $\boldsymbol{u}_1$, $\boldsymbol{u}_2$, …, $\boldsymbol{u}_n$ a complete orthonormal eigenvectors set with $0<\lambda_1\leq\lambda_2\leq\cdots\leq\lambda_n$ corresponding eigenvalues. Let be $\boldsymbol{x}\in\mathbb{R}^n$ then

$$\boldsymbol{x}=\sum_{k=1}^n\alpha_k\boldsymbol{u}_k,\qquad \boldsymbol{x}^T\boldsymbol{x}=\sum_{k=1}^n\alpha_k^2$$

so that $(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x})(\boldsymbol{x}^T\boldsymbol{A}^{-1}\boldsymbol{x})=h(\alpha_1,\ldots,\alpha_n)$ where

$$h(\alpha_1,\ldots,\alpha_n)=\Big(\sum_{k=1}^n\alpha_k^2\lambda_k\Big)\Big(\sum_{k=1}^n\alpha_k^2\lambda_k^{-1}\Big)$$

then the lemma can be reformulated:

- Find maxima and minima of $h(\alpha_1,\ldots,\alpha_n)$
- subject to $\sum_{k=1}^n\alpha_k^2=1$.

Proof. (3/5).

STEP 3: problem reduction. By using Lagrange multiplier maxima and minima are the stationary points of:

$$g(\alpha_1,\ldots,\alpha_n,\mu)=h(\alpha_1,\ldots,\alpha_n)+\mu\Big(\sum_{k=1}^n\alpha_k^2-1\Big)$$

setting $A=\sum_{k=1}^n\alpha_k^2\lambda_k$ and $B=\sum_{k=1}^n\alpha_k^2\lambda_k^{-1}$ we have

$$\frac{\partial g(\alpha_1,\ldots,\alpha_n,\mu)}{\partial\alpha_k}=2\alpha_k\big(\lambda_kB+\lambda_k^{-1}A+\mu\big)=0$$

so that

1. Or $\alpha_k=0$;
2. Or λ_k is a root of the quadratic polynomial $\lambda^2B+\lambda\mu+A$.

in any case there are at most 2 coefficients α's not zero. [a]

[a] the argument should be improved in the case of multiple eigenvalues

Proof. (4/5).

STEP 4: problem reformulation. say α_i and α_j are the only non zero coefficients, then $\alpha_i^2 + \alpha_j^2 = 1$ and we can write

$$
\begin{aligned}
h(\alpha_1,\ldots,\alpha_n) &= (\alpha_i^2\lambda_i + \alpha_j^2\lambda_j)(\alpha_i^2\lambda_i^{-1} + \alpha_j^2\lambda_j^{-1}) \\
&= \alpha_i^4 + \alpha_j^4 + \alpha_i^2\alpha_j^2\left(\frac{\lambda_i}{\lambda_j} + \frac{\lambda_j}{\lambda_i}\right) \\
&= \alpha_i^2(1-\alpha_j^2) + \alpha_j^2(1-\alpha_i^2) + \alpha_i^2\alpha_j^2\left(\frac{\lambda_i}{\lambda_j} + \frac{\lambda_j}{\lambda_i}\right) \\
&= 1 + \alpha_i^2\alpha_j^2\left(\frac{\lambda_i}{\lambda_j} + \frac{\lambda_j}{\lambda_i} - 2\right) \\
&= 1 + \alpha_i^2(1-\alpha_i^2)\frac{(\lambda_i-\lambda_j)^2}{\lambda_i\lambda_j}
\end{aligned}
$$

Proof. (5/5).

STEP 5: bounding maxima and minima. notice that

$$
0 \le \beta(1-\beta) \le \frac{1}{4}, \qquad \forall \beta \in [0,1]
$$

$$
1 \le 1 + \alpha_i^2(1-\alpha_i^2)\frac{(\lambda_i-\lambda_j)^2}{\lambda_i\lambda_j} \le 1 + \frac{(\lambda_i-\lambda_j)^2}{4\lambda_i\lambda_j} = \frac{(\lambda_i+\lambda_j)^2}{4\lambda_i\lambda_j}
$$

to bound $(\lambda_i+\lambda_j)^2/(4\lambda_i\lambda_j)$ consider the function $f(x) = (1+x)^2/x$ which is increasing for $x \ge 1$ so that we have

$$
\frac{(\lambda_i+\lambda_j)^2}{4\lambda_i\lambda_j} \le \frac{(M+m)^2}{4\,M\,m}
$$

and finally

$$
1 \le h(\alpha_1,\ldots,\alpha_n) \le \frac{(M+m)^2}{4\,M\,m}
$$

□

Convergence rate of Steepest Descent

The Kantorovich inequality permits to prove:

Theorem (Convergence rate of Steepest Descent)

Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ an SPD matrix then the steepest descent method:

$$
\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \frac{\boldsymbol{r}_k^T\boldsymbol{r}_k}{\boldsymbol{r}_k^T\boldsymbol{A}\boldsymbol{r}_k}\boldsymbol{r}_k
$$

converge to the solution $\boldsymbol{x}_\star = \boldsymbol{A}^{-1}\boldsymbol{b}$ with at least linear q-rate in the norm $\|\cdot\|_{\boldsymbol{A}}$. Moreover we have the error estimate

$$
\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_\star\|_{\boldsymbol{A}} \le \frac{\kappa-1}{\kappa+1}\|\boldsymbol{x}_k - \boldsymbol{x}_\star\|_{\boldsymbol{A}}
$$

$\kappa = M/m$ is the condition number where $m = \lambda_1$ is the smallest eigenvalue of $\boldsymbol{A}$ and $M = \lambda_n$ is the biggest eigenvalue of $\boldsymbol{A}$.

Proof.

Remember from slide $N^\circ 16$

$$
\|\boldsymbol{x}_\star - \boldsymbol{x}_{k+1}\|_{\boldsymbol{A}}^2 = \|\boldsymbol{x}_\star - \boldsymbol{x}_k\|_{\boldsymbol{A}}^2\left(1 - \frac{(\boldsymbol{r}_k^T\boldsymbol{r}_k)^2}{(\boldsymbol{r}_k^T\boldsymbol{A}^{-1}\boldsymbol{r}_k)(\boldsymbol{r}_k^T\boldsymbol{A}\boldsymbol{r}_k)}\right)
$$

from Kantorovich inequality

$$
1 - \frac{(\boldsymbol{r}_k^T\boldsymbol{r}_k)^2}{(\boldsymbol{r}_k^T\boldsymbol{A}^{-1}\boldsymbol{r}_k)(\boldsymbol{r}_k^T\boldsymbol{A}\boldsymbol{r}_k)} \le 1 - \frac{4\,M\,m}{(M+m)^2} = \frac{(M-m)^2}{(M+m)^2}
$$

so that

$$
\|\boldsymbol{x}_\star - \boldsymbol{x}_{k+1}\|_{\boldsymbol{A}} \le \frac{M-m}{M+m}\|\boldsymbol{x}_\star - \boldsymbol{x}_k\|_{\boldsymbol{A}}
$$

□

Remark (One step convergence)

The steepest descent method can converge in one iteration if $\kappa = 1$ or when $\boldsymbol{r}_0 = \boldsymbol{u}_k$ where $\boldsymbol{u}_k$ is an eigenvector of $\boldsymbol{A}$.

1. *In the first case ($\kappa = 1$) we have $\boldsymbol{A} = \beta \boldsymbol{I}$ for some $\beta > 0$ so it is not interesting.*
2. *In the second case we have*

$$\frac{(\boldsymbol{u}_k^T \boldsymbol{u}_k)^2}{(\boldsymbol{u}_k^T \boldsymbol{A}^{-1} \boldsymbol{u}_k)(\boldsymbol{u}_k^T \boldsymbol{A} \boldsymbol{u}_k)} = \frac{(\boldsymbol{u}_k^T \boldsymbol{u}_k)^2}{\lambda_k^{-1}(\boldsymbol{u}_k^T \boldsymbol{u}_k)\lambda_k(\boldsymbol{u}_k^T \boldsymbol{u}_k)} = 1$$

in both cases we have $\boldsymbol{r}_1 = \boldsymbol{0}$ i.e. we have found the solution.

Outline

Conjugate direction method

Definition (Conjugate vector)

Given two vectors $\boldsymbol{p}$ and $\boldsymbol{q}$ in $\mathbb{R}^n$ are conjugate respect to $\boldsymbol{A}$ if they are orthogonal respect the scalar product induced by $\boldsymbol{A}$; i.e.,

$$\boldsymbol{p}^T \boldsymbol{A} \boldsymbol{q} = \sum_{i,j=1}^{n} A_{ij} p_i q_j = 0.$$

Clearly, n vectors $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots \boldsymbol{p}_n \in \mathbb{R}^n$ that are pair wise conjugated respect to $\boldsymbol{A}$ form a base of $\mathbb{R}^n$.

Problem (Linear system)

Find the minimum of $\mathsf{q}(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} - \boldsymbol{b}^T \boldsymbol{x} + c$ is equivalent to solve the first order necessary condition, i.e.

$$\textit{Find } \boldsymbol{x}_\star \in \mathbb{R}^n \textit{ such that: } \quad \boldsymbol{A}\boldsymbol{x}_\star = \boldsymbol{b}.$$

Observation

Consider $\boldsymbol{x}_0 \in \mathbb{R}^n$ and decompose the error $\boldsymbol{e}_0 = \boldsymbol{x}_\star - \boldsymbol{x}_0$ by the conjugate vectors $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_n \in \mathbb{R}^n$:

$$\boldsymbol{e}_0 = \boldsymbol{x}_\star - \boldsymbol{x}_0 = \sigma_1 \boldsymbol{p}_1 + \sigma_2 \boldsymbol{p}_2 + \cdots + \sigma_n \boldsymbol{p}_n.$$

Evaluating the coefficients $\sigma_1, \sigma_2, \ldots, \sigma_n \in \mathbb{R}$ is equivalent to solve the problem $\boldsymbol{A}\boldsymbol{x}_\star = \boldsymbol{b}$, because knowing $\boldsymbol{e}_0$ we have

$$\boldsymbol{x}_\star = \boldsymbol{x}_0 + \boldsymbol{e}_0.$$

Observation

Using conjugacy the coefficients σ_1, $\sigma_2, \ldots, \sigma_n \in \mathbb{R}$ can be computed as

$$\sigma_i = \frac{\boldsymbol{p}_i^T \boldsymbol{A}\boldsymbol{e}_0}{\boldsymbol{p}_i^T \boldsymbol{A}\boldsymbol{p}_i}, \qquad \text{for } i = 1, 2, \ldots, n.$$

In fact, for all $1 \leq i \leq n$, we have

$$\begin{aligned}\boldsymbol{p}_i^T \boldsymbol{A}\boldsymbol{e}_0 &= \boldsymbol{p}_i^T \boldsymbol{A}\left(\sigma_1\boldsymbol{p}_1 + \sigma_2\boldsymbol{p}_2 + \ldots + \sigma_n\boldsymbol{p}_n\right), \\ &= \sigma_1\boldsymbol{p}_i^T\boldsymbol{A}\boldsymbol{p}_1 + \sigma_2\boldsymbol{p}_i^T\boldsymbol{A}\boldsymbol{p}_2 + \ldots + \sigma_n\boldsymbol{p}_i^T\boldsymbol{A}\boldsymbol{p}_n, \\ &= \sigma_i\boldsymbol{p}_i^T\boldsymbol{A}\boldsymbol{p}_i,\end{aligned}$$

because $\boldsymbol{p}_i^T\boldsymbol{A}\boldsymbol{p}_j = 0$ for $i \neq j$.

The conjugate direction method evaluate the coefficients σ_1, $\sigma_2, \ldots, \sigma_n \in \mathbb{R}$ recursively in n steps, solving for $k \geq 0$ the minimization problem:

Conjugate direction method

Given $\boldsymbol{x}_0$; $k \leftarrow 0$;
repeat
$k \leftarrow k + 1$;
Find $\boldsymbol{x}_k \in \boldsymbol{x}_0 + \mathcal{V}_k$ such that:

$$\boldsymbol{x}_k = \operatorname*{arg\,min}_{\boldsymbol{x} \in \boldsymbol{x}_0 + \mathcal{V}_k} \|\boldsymbol{x}_\star - \boldsymbol{x}\|_{\boldsymbol{A}}$$

until $k = n$

where $\mathcal{V}_k$ is the subspace of $\mathbb{R}^n$ generated by the first k conjugate direction; i.e.,

$$\mathcal{V}_k = \text{SPAN}\{\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_k\}.$$

Step: $\boldsymbol{x}_0 \longrightarrow \boldsymbol{x}_1$

At the first step we consider the subspace $\boldsymbol{x}_0 + \text{SPAN}\{\boldsymbol{p}_1\}$ which consists in vectors of the form

$$\boldsymbol{x}(\alpha) = \boldsymbol{x}_0 + \alpha\boldsymbol{p}_1 \qquad \alpha \in \mathbb{R}$$

The minimization problem becomes:

Minimization step $\boldsymbol{x}_0 \rightarrow \boldsymbol{x}_1$

Find $\boldsymbol{x}_1 = \boldsymbol{x}_0 + \alpha_1\boldsymbol{p}_1$ (i.e., find α_1!) such that:

$$\|\boldsymbol{x}_\star - \boldsymbol{x}_1\|_{\boldsymbol{A}} = \min_{\alpha\in\mathbb{R}} \|\boldsymbol{x}_\star - (\boldsymbol{x}_0 + \alpha\boldsymbol{p}_1)\|_{\boldsymbol{A}},$$

Solving first step method 1

The minimization problem is the minimum respect to α of the quadratic:

$$\begin{aligned}\Phi(\alpha) &= \|\boldsymbol{x}_\star - (\boldsymbol{x}_0 + \alpha\boldsymbol{p}_1)\|_{\boldsymbol{A}}^2, \\ &= \left(\boldsymbol{x}_\star - (\boldsymbol{x}_0 + \alpha\boldsymbol{p}_1)\right)^T \boldsymbol{A}\left(\boldsymbol{x}_\star - (\boldsymbol{x}_0 + \alpha\boldsymbol{p}_1)\right), \\ &= \left(\boldsymbol{e}_0 - \alpha\boldsymbol{p}_1\right)^T \boldsymbol{A}\left(\boldsymbol{e}_0 - \alpha\boldsymbol{p}_1\right), \\ &= \boldsymbol{e}_0^T\boldsymbol{A}\boldsymbol{e}_0 - 2\alpha\boldsymbol{p}_1^T\boldsymbol{A}\boldsymbol{e}_0 + \alpha^2\boldsymbol{p}_1^T\boldsymbol{A}\boldsymbol{p}_1.\end{aligned}$$

minimum is found by imposing:

$$\frac{\mathrm{d}\Phi(\alpha)}{\mathrm{d}\alpha} = -2\boldsymbol{p}_1^T\boldsymbol{A}\boldsymbol{e}_0 + 2\alpha\boldsymbol{p}_1^T\boldsymbol{A}\boldsymbol{p}_1 = 0 \quad \Rightarrow \quad \boxed{\alpha_1 = \frac{\boldsymbol{p}_1^T\boldsymbol{A}\boldsymbol{e}_0}{\boldsymbol{p}_1^T\boldsymbol{A}\boldsymbol{p}_1}}$$

Solving first step method 2 (1/2)

Remember the error expansion:

$$x_\star - x_0 = \sigma_1 p_1 + \sigma_2 p_2 + \cdots + \sigma_n p_n.$$

Let $x(\alpha) = x_0 + \alpha p_1$, the difference $x_\star - x(\alpha)$ becomes:

$$x_\star - x(\alpha) = (\sigma_1 - \alpha) p_1 + \sigma_2 p_2 + \ldots + \sigma_n p_n$$

due to conjugacy the error $\|x_\star - x(\alpha)\|_A$ becomes

$$\begin{aligned}\|x_\star - x(\alpha)\|_A^2 \\ &= \Big((\sigma_1-\alpha)p_1 + \sum_{i=2}^n \sigma_i p_i\Big)^T A \Big((\sigma_1-\alpha)p_1 + \sum_{j=2}^n \sigma_j p_j\Big) \\ &= (\sigma_1-\alpha)^2 p_1^T A p_1 + \sum_{j=2}^n \sigma_j^2 p_j^T A p_j\end{aligned}$$

Solving first step method 2 (2/2)

Because

$$\|x_\star - x(\alpha)\|_A^2 = (\sigma_1 - \alpha)^2 \|p_1\|_A^2 + \sum_{i=2}^n \sigma_2^2 \|p_i\|_A^2,$$

we have that

$$\|x_\star - x(\alpha_1)\|_A^2 = \sum_{i=2}^n \sigma_i^2 \|p_i\|_A^2 \leq \|x_\star - x(\alpha)\|_A^2 \qquad \text{for all } \alpha \neq \sigma_1$$

so that minimum is found by imposing $\alpha_1 = \sigma_1$:

$$\boxed{\alpha_1 = \frac{p_1^T A e_0}{p_1^T A p_1}}$$

This argument can be generalized for all $k > 1$ (see next slides).

Step, $x_{k-1} \to x_k$

For the step from $k-1$ to k we consider the subspace of $\mathbb{R}^n$

$$\mathcal{V}_k = \text{SPAN}\{p_1, p_2, \ldots, p_k\}$$

which contains vectors of the form:

$$x(\alpha^{(1)}, \alpha^{(2)}, \ldots, \alpha^{(k)}) = x_0 + \alpha^{(1)} p_1 + \alpha^{(2)} p_2 + \ldots + \alpha^{(k)} p_k$$

The minimization problem becomes:

Minimization step $x_{k-1} \to x_k$

Find $x_k = x_0 + \alpha_1 p_1 + \alpha_2 p_2 + \ldots + \alpha_k p_k$ (i.e. $\alpha_1, \alpha_2, \ldots, \alpha_k$) such that:

$$\|x_\star - x_k\|_A = \min_{\alpha^{(1)}, \alpha^{(2)}, \ldots, \alpha^{(k)} \in \mathbb{R}} \left\| x_\star - x(\alpha^{(1)}, \alpha^{(2)}, \ldots, \alpha^{(k)}) \right\|_A$$

Solving kth Step: $x_{k-1} \to x_k$ (1/2)

Remember the error expansion:

$$x_\star - x_0 = \sigma_1 p_1 + \sigma_2 p_2 + \cdots + \sigma_n p_n.$$

Consider a vector of the form

$$x(\alpha^{(1)}, \alpha^{(2)}, \ldots, \alpha^{(k)}) = x_0 + \alpha^{(1)} p_1 + \alpha^{(2)} p_2 + \ldots + \alpha^{(k)} p_k$$

the error $x_\star - x(\alpha^{(1)}, \alpha^{(2)}, \ldots, \alpha^{(k)})$ can be written as

$$\begin{aligned} x_\star - x(\alpha^{(1)}, \alpha^{(2)}, \ldots, \alpha^{(k)}) &= x_\star - x_0 - \sum_{i=1}^k \alpha^{(i)} p_i, \\ &= \sum_{i=1}^k (\sigma_i - \alpha^{(i)}) p_i + \sum_{i=k+1}^n \sigma_i p_i. \end{aligned}$$

Solving kth Step: $\boldsymbol{x}_{k-1} \to \boldsymbol{x}_k$ (2/2)

using conjugacy of $\boldsymbol{p}_i$ we obtain the norm of the error:

$$\left\| \boldsymbol{x}_\star - \boldsymbol{x}(\alpha^{(1)}, \alpha^{(2)}, \dots, \alpha^{(k)}) \right\|_{\boldsymbol{A}}^2$$

$$= \sum_{i=1}^{k} \left(\sigma_i - \alpha^{(i)}\right)^2 \|\boldsymbol{p}_i\|_{\boldsymbol{A}}^2 + \sum_{i=k+1}^{n} \sigma_i^2 \|\boldsymbol{p}_i\|_{\boldsymbol{A}}^2 .$$

So that minimum is found by imposing $\alpha_i = \sigma_i$: for $i = 1, 2, \dots, k$.

$$\boxed{\alpha_i = \frac{\boldsymbol{p}_i^T \boldsymbol{A} \boldsymbol{e}_0}{\boldsymbol{p}_i^T \boldsymbol{A} \boldsymbol{p}_i}} \qquad i = 1, 2, \dots, k$$

Successive one dimensional minimization (1/3)

- notice that $\alpha_i = \sigma_i$ and that

$$\boldsymbol{x}_k = \boldsymbol{x}_0 + \alpha_1 \boldsymbol{p}_1 + \cdots + \alpha_k \boldsymbol{p}_k$$
$$= \boldsymbol{x}_{k-1} + \alpha_k \boldsymbol{p}_k$$

- so that $\boldsymbol{x}_{k-1}$ contains $k-1$ coefficients α_i for the minimization.
- if we consider the one dimensional minimization on the subspace $\boldsymbol{x}_{k-1} + \text{SPAN}\{\boldsymbol{p}_k\}$ we find again $\boldsymbol{x}_k$!

Successive one dimensional minimization (2/3)

Consider a vector of the form

$$\boldsymbol{x}(\alpha) = \boldsymbol{x}_{k-1} + \alpha \boldsymbol{p}_k$$

remember that $\boldsymbol{x}_{k-1} = \boldsymbol{x}_0 + \alpha_1 \boldsymbol{p}_1 + \cdots + \alpha_{k-1} \boldsymbol{p}_{k-1}$ so that the error $\boldsymbol{x}_\star - \boldsymbol{x}(\alpha)$ can be written as

$$\boldsymbol{x}_\star - \boldsymbol{x}(\alpha) = \boldsymbol{x}_\star - \boldsymbol{x}_0 - \sum_{i=1}^{k-1} \alpha_i \boldsymbol{p}_i + \alpha \boldsymbol{p}_k$$

$$= \sum_{i=1}^{k-1} (\sigma_i - \alpha_i) \boldsymbol{p}_i + (\sigma_k - \alpha) \boldsymbol{p}_k + \sum_{i=k+1}^{n} \sigma_i \boldsymbol{p}_i .$$

due to the equality $\sigma_i = \alpha_i$ the blue part of the expression is 0.

Successive one dimensional minimization (3/3)

Using conjugacy of $\boldsymbol{p}_i$ we obtain the norm of the error:

$$\|\boldsymbol{x}_\star - \boldsymbol{x}(\alpha)\|_{\boldsymbol{A}}^2 = (\sigma_k - \alpha)^2 \|\boldsymbol{p}_k\|_{\boldsymbol{A}}^2 + \sum_{i=k+1}^{n} \sigma_i^2 \|\boldsymbol{p}_i\|_{\boldsymbol{A}}^2 .$$

So that minimum is found by imposing $\alpha = \sigma_k$:

$$\boxed{\alpha_k = \frac{\boldsymbol{p}_k^T \boldsymbol{A} \boldsymbol{e}_0}{\boldsymbol{p}_k^T \boldsymbol{A} \boldsymbol{p}_k}}$$

Remark

This observation permit to perform the minimization on the k-dimensional space $\boldsymbol{x}_0 + \mathcal{V}_k$ as successive one dimensional minimizations along the conjugate directions $\boldsymbol{p}_k$!.

Problem (one dimensional successive minimization)

Find $\boldsymbol{x}_k = \boldsymbol{x}_{k-1} + \alpha_k \boldsymbol{p}_k$ *such that:*

$$\|\boldsymbol{x}_\star - \boldsymbol{x}_k\|_{\boldsymbol{A}} = \min_{\alpha\in\mathbb{R}} \|\boldsymbol{x}_\star - (\boldsymbol{x}_{k-1} + \alpha \boldsymbol{p}_k)\|_{\boldsymbol{A}},$$

The solution is the minimum respect to α of the quadratic:

$$\begin{aligned}\Phi(\alpha) &= (\boldsymbol{x}_\star - (\boldsymbol{x}_{k-1} + \alpha \boldsymbol{p}_k))^T \boldsymbol{A} (\boldsymbol{x}_\star - (\boldsymbol{x}_{k-1} + \alpha \boldsymbol{p}_k)),\\ &= (\boldsymbol{e}_{k-1} - \alpha \boldsymbol{p}_k)^T \boldsymbol{A} (\boldsymbol{e}_{k-1} - \alpha \boldsymbol{p}_k),\\ &= \boldsymbol{e}_{k-1}^T \boldsymbol{A} \boldsymbol{e}_{k-1} - 2\alpha \boldsymbol{p}_k^T \boldsymbol{A} \boldsymbol{e}_{k-1} + \alpha^2 \boldsymbol{p}_k^T \boldsymbol{A} \boldsymbol{p}_k.\end{aligned}$$

minimum is found by imposing:

$$\frac{\mathrm{d}\Phi(\alpha)}{\mathrm{d}\alpha} = -2\boldsymbol{p}_k^T \boldsymbol{A} \boldsymbol{e}_{k-1} + 2\alpha \boldsymbol{p}_k^T \boldsymbol{A} \boldsymbol{p}_k = 0 \quad \Rightarrow \quad \boxed{\alpha_k = \frac{\boldsymbol{p}_k^T \boldsymbol{A} \boldsymbol{e}_{k-1}}{\boldsymbol{p}_k^T \boldsymbol{A} \boldsymbol{p}_k}}$$

- In the case of minimization on the subspace $\boldsymbol{x}_0 + \mathcal{V}_k$ we have:

$$\alpha_k = \boldsymbol{p}_k^T \boldsymbol{A} \boldsymbol{e}_0 \,/\, \boldsymbol{p}_k^T \boldsymbol{A} \boldsymbol{p}_k$$

- In the case of one dimensional minimization on the subspace $\boldsymbol{x}_{k-1} + \mathrm{SPAN}\{\boldsymbol{p}_k\}$ we have:

$$\alpha_k = \boldsymbol{p}_k^T \boldsymbol{A} \boldsymbol{e}_{k-1} \,/\, \boldsymbol{p}_k^T \boldsymbol{A} \boldsymbol{p}_k$$

- Apparently they are different results, however by using the conjugacy of the vectors $\boldsymbol{p}_i$ we have

$$\begin{aligned}\boldsymbol{p}_k^T \boldsymbol{A} \boldsymbol{e}_{k-1} &= \boldsymbol{p}_k^T \boldsymbol{A} (\boldsymbol{x}_\star - \boldsymbol{x}_{k-1})\\ &= \boldsymbol{p}_k^T \boldsymbol{A} \big(\boldsymbol{x}_\star - (\boldsymbol{x}_0 + \alpha_1 \boldsymbol{p}_1 + \cdots + \alpha_{k-1} \boldsymbol{p}_{k-1})\big)\\ &= \boldsymbol{p}_k^T \boldsymbol{A} \boldsymbol{e}_0 - \alpha_1 \boldsymbol{p}_k^T \boldsymbol{A} \boldsymbol{p}_1 - \cdots - \alpha_{k-1} \boldsymbol{p}_k^T \boldsymbol{A} \boldsymbol{p}_{k-1}\\ &= \boldsymbol{p}_k^T \boldsymbol{A} \boldsymbol{e}_0\end{aligned}$$

- The one step minimization in the space $\boldsymbol{x}_0 + \mathcal{V}_n$ and the successive minimization in the space $\boldsymbol{x}_{k-1} + \mathrm{SPAN}\{\boldsymbol{p}_k\}$, $k = 1, 2, \ldots, n$ are equivalent if $\boldsymbol{p}_i$s are conjugate.
- The successive minimization is useful when $\boldsymbol{p}_i$s are not known in advance but must be computed as the minimization process proceeds.
- The evaluation of α_k is apparently not computable because $\boldsymbol{e}_i$ is not known. However noticing

$$\boldsymbol{A}\boldsymbol{e}_k = \boldsymbol{A}(\boldsymbol{x}_\star - \boldsymbol{x}_k) = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k = \boldsymbol{r}_k$$

  we can write

$$\alpha_k = \boldsymbol{p}_k^T \boldsymbol{A} \boldsymbol{e}_{k-1} \,/\, \boldsymbol{p}_k^T \boldsymbol{A} \boldsymbol{p}_k = \boldsymbol{p}_k^T \boldsymbol{r}_{k-1} \,/\, \boldsymbol{p}_k^T \boldsymbol{A} \boldsymbol{p}_k =$$

- Finally for the residual is valid the recurrence

$$\boldsymbol{r}_k = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k = \boldsymbol{b} - \boldsymbol{A}(\boldsymbol{x}_{k-1} + \alpha_k \boldsymbol{p}_k) = \boldsymbol{r}_{k-1} - \alpha_k \boldsymbol{A}\boldsymbol{p}_k.$$

Conjugate direction minimization

Algorithm (Conjugate direction minimization)

$k \leftarrow 0$; $\boldsymbol{x}_0$ *assigned*;
$\boldsymbol{r}_0 \leftarrow \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_0$;
while *not converged* **do**
  $k \leftarrow k + 1$;
  $\alpha_k \leftarrow \dfrac{\boldsymbol{r}_{k-1}^T \boldsymbol{p}_k^T}{\boldsymbol{p}_k \boldsymbol{A} \boldsymbol{p}_k}$;
  $\boldsymbol{x}_k \leftarrow \boldsymbol{x}_{k-1} + \alpha_k \boldsymbol{p}_k$;
  $\boldsymbol{r}_k \leftarrow \boldsymbol{r}_{k-1} - \alpha_k \boldsymbol{A}\boldsymbol{p}_k$;
end while

Observation (Computazional cost)

The conjugate direction minimization requires at each step one matrix–vector product for the evaluation of α_k *and two update AXPY for* $\boldsymbol{x}_k$ *and* $\boldsymbol{r}_k$.

Monotonic behavior of the error

Remark (Monotonic behavior of the error)

The energy norm of the error $\|e_k\|_A$ is monotonically decreasing in k. In fact:

$$e_k = x_\star - x_k = \alpha_{k+1} p_{k+1} + \ldots + \alpha_n p_n,$$

and by conjugacy

$$\|e_k\|_A^2 = \|x_\star - x_k\|_A^2 = \sigma_{k+1}^2 \|p_{k+1}\|_A^2 + \ldots + \sigma_n^2 \|p_n\|_A^2 .$$

Finally from this relation we have $e_n = 0$.

Outline

Conjugate Gradient method

The Conjugate Gradient method combine the Conjugate Direction method with an orthogonalization process (like Gram-Schmidt) applied to the residual to construct the conjugate directions. In fact, because A define a scalar product in the next slide we prove:

- each residue is orthogonal to the previous conjugate directions, and consequently linearly independent from the previous conjugate directions.
- if the residual is not null is can be used to construct a new conjugate direction.

Orthogonality of the residue r_k respect $\mathcal{V}_k$

- The residue r_k is orthogonal to p_1, p_2, $\ldots, p_k$. In fact, from the error expansion

$$e_k = \alpha_{k+1} p_{k+1} + \alpha_{k+2} p_{k+2} + \cdots + \alpha_n p_n$$

because $r_k = A e_k$, for $i = 1, 2, \ldots, k$ we have

$$\begin{aligned} p_i^T r_k &= p_i^T A e_k \\ &= p_i^T A \sum_{j=k+1}^{n} \alpha_j p_j = \sum_{j=k+1}^{n} \alpha_j p_i^T A p_j \\ &= 0 \end{aligned}$$

Building new conjugate direction (1/2)

- The conjugate direction method build one new direction at each step.
- If $\boldsymbol{r}_k \neq \boldsymbol{0}$ it can be used to build the new direction $\boldsymbol{p}_{k+1}$ by a Gram-Schmidt orthogonalization process

$$\boldsymbol{p}_{k+1} = \boldsymbol{r}_k + \beta_1^{(k+1)}\boldsymbol{p}_1 + \beta_2^{(k+1)}\boldsymbol{p}_2 + \ldots + \beta_k^{(k+1)}\boldsymbol{p}_k,$$

where the k coefficients $\beta_1^{(k+1)}$, $\beta_2^{(k+1)}$, $\ldots, \beta_k^{(k+1)}$ must satisfy:

$$\boldsymbol{p}_i^T\boldsymbol{A}\boldsymbol{p}_{k+1} = 0, \qquad \text{for } i = 1, 2, \ldots, k.$$

Building new conjugate direction (2/2)

(repeating from previous slide)

$$\boldsymbol{p}_{k+1} = \boldsymbol{r}_k + \beta_1^{(k+1)}\boldsymbol{p}_1 + \beta_2^{(k+1)}\boldsymbol{p}_2 + \cdots + \beta_k^{(k+1)}\boldsymbol{p}_k,$$

expanding the expression:

$$\begin{aligned} 0 &= \boldsymbol{p}_i^T\boldsymbol{A}\boldsymbol{p}_{k+1}, \\ &= \boldsymbol{p}_i^T\boldsymbol{A}\big(\boldsymbol{r}_k + \beta_1^{(k+1)}\boldsymbol{p}_1 + \beta_2^{(k+1)}\boldsymbol{p}_2 + \cdots + \beta_k^{(k+1)}\boldsymbol{p}_k\big), \\ &= \boldsymbol{p}_i^T\boldsymbol{A}\boldsymbol{r}_k + \beta_i^{(k+1)}\boldsymbol{p}_i^T\boldsymbol{A}\boldsymbol{p}_i, \end{aligned}$$

$$\Rightarrow \boxed{\beta_i^{(k+1)} = -\frac{\boldsymbol{p}_i^T\boldsymbol{A}\boldsymbol{r}_k}{\boldsymbol{p}_i^T\boldsymbol{A}\boldsymbol{p}_i}} \qquad i = 1, 2, \ldots, k$$

The choice of the residual $\boldsymbol{r}_k \neq \boldsymbol{0}$ for the construction of the new conjugate direction $\boldsymbol{p}_{k+1}$ has three important consequences:

1. simplification of the expression for α_k;
2. Orthogonality of the residual $\boldsymbol{r}_k$ from the previous residue $\boldsymbol{r}_0$, $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_{k-1}$;
3. three point formula and simplification of the coefficients $\beta_i^{(k+1)}$.

this facts will be examined in the next slides.

Simplification of the expression for α_k

Writing the expression for $\boldsymbol{p}_k$ from the orthogonalization process

$$\boldsymbol{p}_k = \boldsymbol{r}_{k-1} + \beta_1^{(k+1)}\boldsymbol{p}_1 + \beta_2^{(k+1)}\boldsymbol{p}_2 + \ldots + \beta_{k-1}^{(k+1)}\boldsymbol{p}_{k-1},$$

using orthogonality of $\boldsymbol{r}_{k-1}$ and the vectors $\boldsymbol{p}_1$, $\boldsymbol{p}_2$, $\ldots, \boldsymbol{p}_{k-1}$, (see slide N.48) we have

$$\begin{aligned} \boldsymbol{r}_{k-1}^T\boldsymbol{p}_k &= \boldsymbol{r}_{k-1}^T\big(\boldsymbol{r}_{k-1} + \beta_1^{(k+1)}\boldsymbol{p}_1 + \beta_3^{(k+1)}\boldsymbol{p}_2 + \ldots + \beta_{k-1}^{(k+1)}\boldsymbol{p}_{k-1}\big), \\ &= \boldsymbol{r}_{k-1}^T\boldsymbol{r}_{k-1}. \end{aligned}$$

recalling the definition of α_k it follows:

$$\alpha_k = \frac{\boldsymbol{e}_{k-1}^T\boldsymbol{A}\boldsymbol{p}_k}{\boldsymbol{p}_k^T\boldsymbol{A}\boldsymbol{p}_k} = \frac{\boldsymbol{r}_{k-1}^T\boldsymbol{p}_k}{\boldsymbol{p}_k^T\boldsymbol{A}\boldsymbol{p}_k} = \boxed{\frac{\boldsymbol{r}_{k-1}^T\boldsymbol{r}_{k-1}}{\boldsymbol{p}_k^T\boldsymbol{A}\boldsymbol{p}_k}}$$

Orthogonally of the residue $\boldsymbol{r}_k$ from $\boldsymbol{r}_0$, $\boldsymbol{r}_1$, $\ldots$, $\boldsymbol{r}_{k-1}$

From the definition of $\boldsymbol{p}_{i+1}$ it follows:

$$\boldsymbol{p}_{i+1} = \boldsymbol{r}_i + \beta_1^{(i+1)}\boldsymbol{p}_1 + \beta_2^{(i+1)}\boldsymbol{p}_2 + \ldots + \beta_i^{(i+1)}\boldsymbol{p}_i,$$

$$\Rightarrow \quad \boldsymbol{r}_i \in \mathrm{SPAN}\{\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_i, \boldsymbol{p}_{i+1}\} = \mathcal{V}_{i+1} \qquad \text{(obvious)}$$

using orthogonality of $\boldsymbol{r}_k$ and the vectors $\boldsymbol{p}_1$, $\boldsymbol{p}_2$, $\ldots$, $\boldsymbol{p}_k$, (see slide N.48) for $i < k$ we have

$$\boldsymbol{r}_k^T\boldsymbol{r}_i = \boldsymbol{r}_k^T\left(\boldsymbol{p}_{i+1} - \sum_{j=1}^{i}\beta_j^{(i+1)}\boldsymbol{p}_j\right),$$

$$= \boldsymbol{r}_k^T\boldsymbol{p}_{i+1} - \sum_{j=1}^{i}\beta_j^{(i+1)}\boldsymbol{r}_k^T\boldsymbol{p}_j = 0.$$

Three point formula and simplification of $\beta_i^{(k+1)}$

From the relation $\quad \boldsymbol{r}_k^T\boldsymbol{r}_i = \boldsymbol{r}_k^T(\boldsymbol{r}_{i-1} - \alpha_i\boldsymbol{A}\boldsymbol{p}_i)$ we deduce

$$\boldsymbol{r}_k^T\boldsymbol{A}\boldsymbol{p}_i = \frac{\boldsymbol{r}_k^T\boldsymbol{r}_{i-1} - \boldsymbol{r}_k^T\boldsymbol{r}_i}{\alpha_i} = \begin{cases} -\boldsymbol{r}_k^T\boldsymbol{r}_k/\alpha_k & \text{if } i = k; \\ 0 & \text{if } i < k; \end{cases}$$

remembering that $\alpha_k = \boldsymbol{r}_{k-1}^T\boldsymbol{r}_{k-1} \;/\; \boldsymbol{p}_k^T\boldsymbol{A}\boldsymbol{p}_k$ we obtain

$$\beta_i^{(k+1)} = -\frac{\boldsymbol{r}_k^T\boldsymbol{A}\boldsymbol{p}_i}{\boldsymbol{p}_i^T\boldsymbol{A}\boldsymbol{p}_i} = \begin{cases} \dfrac{\boldsymbol{r}_k^T\boldsymbol{r}_k}{\boldsymbol{r}_{k-1}^T\boldsymbol{r}_{k-1}} & i = k; \\ 0 & i < k; \end{cases}$$

i.e. there is only one non zero coefficient $\beta_k^{(k+1)}$, so we write $\beta_k = \beta_k^{(k+1)}$ and obtain the three point formula:

$$\boldsymbol{p}_{k+1} = \boldsymbol{r}_k + \beta_k\boldsymbol{p}_k$$

Conjugate gradient algorithm

initial step:
$k \leftarrow 0$; $\boldsymbol{x}_0$ assigned;
$\boldsymbol{r}_0 \leftarrow \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_0$;
$\boldsymbol{p}_1 \leftarrow \boldsymbol{r}_0$;
while $\|\boldsymbol{r}_k\| > \epsilon$ **do**
  $k \leftarrow k + 1$;
  Conjugate direction method
  $\alpha_k \leftarrow \dfrac{\boldsymbol{r}_{k-1}^T\boldsymbol{r}_{k-1}}{\boldsymbol{p}_k^T\boldsymbol{A}\boldsymbol{p}_k}$;
  $\boldsymbol{x}_k \leftarrow \boldsymbol{x}_{k-1} + \alpha_k\boldsymbol{p}_k$;
  $\boldsymbol{r}_k \leftarrow \boldsymbol{r}_{k-1} - \alpha_k\boldsymbol{A}\boldsymbol{p}_k$;
  Residual orthogonalization
  $\beta_k \leftarrow \dfrac{\boldsymbol{r}_k^T\boldsymbol{r}_k}{\boldsymbol{r}_{k-1}^T\boldsymbol{r}_{k-1}}$;
  $\boldsymbol{p}_{k+1} \leftarrow \boldsymbol{r}_k + \beta_k\boldsymbol{p}_k$;
end while

Outline

Polynomial residual expansions (1/6)

Lemma

The residuals and cojugate directions for the Conjugate Gradient iterative scheme of slide 55 can be written as

$$\boldsymbol{r}_k = P_k(\boldsymbol{A})\boldsymbol{r}_0 \qquad k = 0, 1, \ldots, n$$

$$\boldsymbol{p}_k = Q_{k-1}(\boldsymbol{A})\boldsymbol{r}_0 \qquad k = 1, 2, \ldots, n$$

where $P_k(x)$ and $Q_k(x)$ are k-degree polynomial such that $P_k(0) = 1$ for all k.

Proof. (1/2).

The proof is by induction.

Base $k = 0$: $\quad \boldsymbol{p}_1 = \boldsymbol{r}_0$

so that $P_0(x) = 1$ and $Q_0(x) = 1$.

Polynomial residual expansions (2/6)

Proof. (2/2).

Let the expansion valid for $k-1$. Consider the recursion for the residual:

$$\begin{aligned}\boldsymbol{r}_k &= \boldsymbol{r}_{k-1} - \alpha_k \boldsymbol{A}\boldsymbol{p}_k \\ &= P_{k-1}(\boldsymbol{A})\boldsymbol{r}_0 + \alpha_k \boldsymbol{A} Q_{k-1}(\boldsymbol{A})\boldsymbol{r}_0 \\ &= \big(P_{k-1}(\boldsymbol{A}) + \alpha_k \boldsymbol{A} Q_{k-1}(\boldsymbol{A})\big)\boldsymbol{r}_0\end{aligned}$$

then $P_k(x) = P_{k-1}(x) + \alpha_k x Q_{k-1}(x)$ and $P_k(0) = P_{k-1}(0) = 1$. Consider the recursion for the conjugate direction

$$\begin{aligned}\boldsymbol{p}_{k+1} &= P_k(\boldsymbol{A})\boldsymbol{r}_0 + \beta_k Q_{k-1}(\boldsymbol{A})\boldsymbol{r}_0 \\ &= \big(P_k(\boldsymbol{A}) + \beta_k Q_{k-1}(\boldsymbol{A})\big)\boldsymbol{r}_0\end{aligned}$$

then $Q_k(x) = P_k(x) + \beta_k Q_{k-1}(x)$. □

Polynomial residual expansions (3/6)

Corollary

$$\boldsymbol{e}_k = P_k(\boldsymbol{A})\boldsymbol{e}_0.$$

Proof.

$$\begin{aligned}\boldsymbol{e}_k = \boldsymbol{x}_\star - \boldsymbol{x}_k &= \boldsymbol{A}^{-1}\boldsymbol{r}_k \\ &= \boldsymbol{A}^{-1}P_k(\boldsymbol{A})\boldsymbol{r}_0 \\ &= P_k(\boldsymbol{A})\boldsymbol{A}^{-1}\boldsymbol{r}_0 \\ &= P_k(\boldsymbol{A})(\boldsymbol{x}_\star - \boldsymbol{x}_0) \\ &= P_k(\boldsymbol{A})\boldsymbol{e}_0.\end{aligned}$$

□

Polynomial residual expansions (4/6)

Lemma

For the Conjugate Gradient iterative scheme of slide n.55 we have:

$$\mathcal{V}_k = \left\{ p(\boldsymbol{A})\boldsymbol{e}_0 \,|\, p \in \mathbb{P}^k,\ p(0) = 0 \right\}$$

Proof.

Using expansion of slide n.57 and $\boldsymbol{r}_0 = \boldsymbol{A}\boldsymbol{e}_0$ we have:

$$\begin{aligned}\mathcal{V}_k &= \text{SPAN}\{\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots \boldsymbol{p}_k\} \\ &= \left\{ \sum_{i=0}^{k-1} \beta_i Q_i(\boldsymbol{A})\boldsymbol{r}_0 \,\middle|\, (\beta_0, \ldots, \beta_{k-1}) \in \mathbb{R}^{k-1} \right\} \\ &= \left\{ q(\boldsymbol{A})\boldsymbol{A}\boldsymbol{e}_0 \,|\, p \in \mathbb{P}^{k-1} \right\} = \left\{ p(\boldsymbol{A})\boldsymbol{e}_0 \,|\, p \in \mathbb{P}^k,\ p(0) = 0 \right\}\end{aligned}$$

□

Polynomial residual expansions (5/6)

By using the equaility

$$\mathcal{V}_k = \{p(\boldsymbol{A})\boldsymbol{e}_0 \,|\, p \in \mathbb{P}^k,\, p(0) = 0\}$$

The optimality of CG step can be written as

$$\|\boldsymbol{x}_\star - \boldsymbol{x}_k\|_{\boldsymbol{A}} \leq \|\boldsymbol{x}_\star - \boldsymbol{x}\|_{\boldsymbol{A}}, \qquad \forall \boldsymbol{x} \in \boldsymbol{x}_0 + \mathcal{V}_k$$

$$\|\boldsymbol{x}_\star - \boldsymbol{x}_k\|_{\boldsymbol{A}} \leq \|\boldsymbol{x}_\star - (\boldsymbol{x}_0 + p(\boldsymbol{A})\boldsymbol{e}_0)\|_{\boldsymbol{A}}, \qquad \forall p \in \mathbb{P}^k,\, p(0) = 0$$

$$\|\boldsymbol{x}_\star - \boldsymbol{x}_k\|_{\boldsymbol{A}} \leq \|P(\boldsymbol{A})\boldsymbol{e}_0\|_{\boldsymbol{A}}, \qquad \forall P \in \mathbb{P}^k,\, P(0) = 1$$

And using the results of slide 60 and 59 we can write

$$\boldsymbol{e}_k = P_k(\boldsymbol{A})\boldsymbol{e}_0,$$

$$\|\boldsymbol{e}_k\|_{\boldsymbol{A}} = \|P_k(\boldsymbol{A})\boldsymbol{e}_0\|_{\boldsymbol{A}} \leq \|P(\boldsymbol{A})\boldsymbol{e}_0\|_{\boldsymbol{A}} \qquad \forall P \in \mathbb{P}^k,\, P(0) = 1$$

Polynomial residual expansions (6/6)

From previous equations we have the characterization of CG error

$$\|\boldsymbol{e}_k\|_{\boldsymbol{A}} = \inf_{P\in\mathbb{P}^k,\, P(0)=1} \|P(\boldsymbol{A})\boldsymbol{e}_0\|_{\boldsymbol{A}}$$

Thus, an estimate of the form

$$\|\boldsymbol{e}_k\|_{\boldsymbol{A}} \leq C_k \|\boldsymbol{e}_0\|_{\boldsymbol{A}}$$

can be obtained by using estimate on the polynomial of the form

$$\left\{P \in \mathbb{P}^k,\, P(0) = 1\right\}$$

Convergence rate calculation

Lemma

Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ an SPD matrix, and $p \in \mathbb{P}^k$ a polynomial, then

$$\|p(\boldsymbol{A})\boldsymbol{x}\|_{\boldsymbol{A}} \leq \|p(\boldsymbol{A})\|_2 \|\boldsymbol{x}\|_{\boldsymbol{A}}$$

Proof. (1/2).

The matrix $\boldsymbol{A}$ is SPD so that we can write

$$\boldsymbol{A} = \boldsymbol{U}^T\boldsymbol{\Lambda}\boldsymbol{U}, \qquad \boldsymbol{\Lambda} = \mathrm{DIAG}\{\lambda_1, \lambda_2, \dots, \lambda_n\}$$

where $\boldsymbol{U}$ is an orthogonal matrix (i.e. $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$) and $\boldsymbol{\Lambda} \geq \boldsymbol{0}$ is diagonal. We can define the SPD matrix $\boldsymbol{A}^{1/2}$ as follows

$$\boldsymbol{A}^{1/2} = \boldsymbol{U}^T\boldsymbol{\Lambda}^{1/2}\boldsymbol{U}, \qquad \boldsymbol{\Lambda}^{1/2} = \mathrm{DIAG}\{\lambda_1^{1/2}, \lambda_2^{1/2}, \dots, \lambda_n^{1/2}\}$$

and obviously $\boldsymbol{A}^{1/2}\boldsymbol{A}^{1/2} = \boldsymbol{A}$.

Proof. (2/2).

Notice that

$$\|\boldsymbol{x}\|_{\boldsymbol{A}}^2 = \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{A}^{1/2}\boldsymbol{A}^{1/2}\boldsymbol{x} = \left\|\boldsymbol{A}^{1/2}\boldsymbol{x}\right\|_2^2$$

so that

$$\begin{aligned}\|p(\boldsymbol{A})\boldsymbol{x}\|_{\boldsymbol{A}} &= \left\|\boldsymbol{A}^{1/2}p(\boldsymbol{A})\boldsymbol{x}\right\|_2 \\ &= \left\|p(\boldsymbol{A})\boldsymbol{A}^{1/2}\boldsymbol{x}\right\|_2 \\ &\leq \|p(\boldsymbol{A})\|_2 \left\|\boldsymbol{A}^{1/2}\boldsymbol{x}\right\|_2 \\ &= \|p(\boldsymbol{A})\|_2 \|\boldsymbol{x}\|_{\boldsymbol{A}}\end{aligned}$$

Lemma

Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ an SPD matrix, and $p \in \mathbb{P}^k$ a polynomial, then

$$\|p(\boldsymbol{A})\|_2 = \max_{\lambda\in\sigma(\boldsymbol{A})} |p(\lambda)|$$

Proof.

The matrix $p(\boldsymbol{A})$ is symmetric, and for a generic symmetric matrix $\boldsymbol{B}$ we have

$$\|\boldsymbol{B}\|_2 = \max_{\lambda\in\sigma(\boldsymbol{B})} |\lambda|$$

observing that if λ is an eigenvalue of $\boldsymbol{A}$ then $p(\lambda)$ is an eigenvalue of $p(\boldsymbol{A})$ the thesis easily follows. □

- Starting the error estimate

$$\|\boldsymbol{e}_k\|_{\boldsymbol{A}} \leq \inf_{P\in\mathbb{P}^k,\,P(0)=1} \|P(\boldsymbol{A})\boldsymbol{e}_0\|_{\boldsymbol{A}}$$

- Combining the last two lemma we easily obtain the estimate

$$\|\boldsymbol{e}_k\|_{\boldsymbol{A}} \leq \inf_{P\in\mathbb{P}^k,\,P(0)=1} \left[\max_{\lambda\in\sigma(\boldsymbol{A})} |P(\lambda)|\right] \|\boldsymbol{e}_0\|_{\boldsymbol{A}}$$

- The convergence rate is estimated by bounding the constant

$$\inf_{P\in\mathbb{P}^k,\,P(0)=1} \left[\max_{\lambda\in\sigma(\boldsymbol{A})} |P(\lambda)|\right]$$

Finite termination of Conjugate Gradient

Theorem (Finite termination of Conjugate Gradient)

Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ an SPD matrix, the the Conjugate Gradient applied to the linear system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ terminate finding the exact solution in at most n-step.

Proof.

From the estimate

$$\|\boldsymbol{e}_k\|_{\boldsymbol{A}} \leq \inf_{P\in\mathbb{P}^k,\,P(0)=1} \left[\max_{\lambda\in\sigma(\boldsymbol{A})} |P(\lambda)|\right] \|\boldsymbol{e}_0\|_{\boldsymbol{A}}$$

choosing $$P(x) = \prod_{\lambda\in\sigma(\boldsymbol{A})} (x-\lambda) \Big/ \prod_{\lambda\in\sigma(\boldsymbol{A})} (0-\lambda)$$

we have $\max_{\lambda\in\sigma(\boldsymbol{A})} |P(\lambda)| = 0$ and $\|\boldsymbol{e}_n\|_{\boldsymbol{A}} = 0$. □

Convergence rate of Conjugate Gradient

1. The constant

$$\inf_{P\in\mathbb{P}^k,\,P(0)=1} \left[\max_{\lambda\in\sigma(\boldsymbol{A})} |P(\lambda)|\right]$$

is not easy to evaluate,

2. The following bound, is useful

$$\max_{\lambda\in\sigma(\boldsymbol{A})} |P(\lambda)| \leq \max_{\lambda\in[\lambda_1,\lambda_n]} |P(\lambda)|$$

3. in particular the final estimate will be obtained by

$$\inf_{P\in\mathbb{P}^k,\,P(0)=1} \left[\max_{\lambda\in\sigma(\boldsymbol{A})} |P(\lambda)|\right] \leq \max_{\lambda\in[\lambda_1,\lambda_n]} \left|\bar{P}_k(\lambda)\right|$$

where $\bar{P}_k(x)$ is an opportune k-degree polynomial for which $\bar{P}_k(0) = 1$ and it is easy to evaluate $\max_{\lambda\in[\lambda_1,\lambda_n]} \left|\bar{P}_k(\lambda)\right|$.

Chebyshev Polynomials (1/4)

1. The Chebyshev Polynomials of the First Kind are the right polynomial for this estimate. This polynomial have the following definition in the interval $[-1, 1]$:

$$T_k(x) = \cos(k \arccos(x))$$

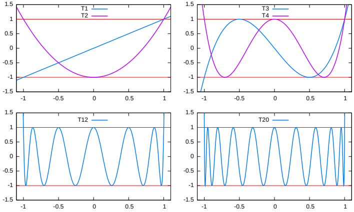
2. Another equivalent definition valid in the interval $(-\infty, \infty)$ is the following

$$T_k(x) = \frac{1}{2}\left[\left(x + \sqrt{x^2-1}\right)^k + \left(x - \sqrt{x^2-1}\right)^k\right]$$

3. In spite of these definition, $T_k(x)$ is effectively a polynomial.

Chebyshev Polynomials (2/4)

Some example of Chebyshev Polynomials.



Chebyshev Polynomials (3/4)

1. It is easy to show that $T_k(x)$ is a polynomial by the use of

$$\cos(\alpha+\beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta$$

$$\cos(\alpha+\beta) + \cos(\alpha-\beta) = 2\cos\alpha\cos\beta$$

let $\theta = \arccos(x)$:
   1. $T_0(x) = \cos(0\,\theta) = 1;$
   2. $T_1(x) = \cos(1\,\theta) = x;$
   3. $T_2(x) = \cos(2\,\theta) = \cos(\theta)^2 - \sin(\theta)^2 = 2\cos(\theta)^2 - 1 = 2x^2 - 1;$
   4. $T_{k+1}(x) + T_{k-1}(x) = \cos((k+1)\theta) + \cos((k-1)\theta)$ $= 2\cos(k\theta)\cos(\theta) = 2\,x\,T_k(x)$
2. In general we have the following recurrence:
   1. $T_0(x) = 1;$
   2. $T_1(x) = x;$
   3. $T_{k+1}(x) = 2\,x\,T_k(x) - T_{k-1}(x).$

Chebyshev Polynomials (4/4)

- Solving the recurrence:
   1. $T_0(x) = 1;$
   2. $T_1(x) = x;$
   3. $T_{k+1}(x) = 2\,x\,T_k(x) - T_{k-1}(x).$
- We obtain the explicit form of the Chebyshev Polynomials

$$T_k(x) = \frac{1}{2}\left[\left(x + \sqrt{x^2-1}\right)^k + \left(x - \sqrt{x^2-1}\right)^k\right]$$

- The translated and scaled polynomial is useful in the study of the conjugate gradient method:

$$T_k(x; a, b) = T_k\left(\frac{a+b-2x}{b-a}\right)$$

where we have $|T_k(x; a, b)| \leq 1$ for all $x \in [a, b]$.

Convergence rate of Conjugate Gradient method

Theorem (Convergence rate of Conjugate Gradient method)

Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ an SPD matrix then the Conjugate Gradient method converge to the solution $\boldsymbol{x}_\star = \boldsymbol{A}^{-1}\boldsymbol{b}$ with at least linear r-rate in the norm $\|\cdot\|_{\boldsymbol{A}}$. Moreover we have the error estimate

$$\|\boldsymbol{e}_k\|_{\boldsymbol{A}} \lesssim 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k \|\boldsymbol{e}_0\|_{\boldsymbol{A}}$$

$\kappa = M/m$ is the condition number where $m = \lambda_1$ is the smallest eigenvalue of $\boldsymbol{A}$ and $M = \lambda_n$ is the biggest eigenvalue of $\boldsymbol{A}$.

The expression $a_k \lesssim b_k$ means that for all $\epsilon > 0$ there exists $k_0 > 0$ such that:

$$a_k \leq (1-\epsilon)b_k, \qquad \forall k > k_0$$

Proof.

From the estimate

$$\|\boldsymbol{e}_k\|_{\boldsymbol{A}} \leq \max_{\lambda\in[m,M]} |P(\lambda)| \, \|\boldsymbol{e}_0\|_{\boldsymbol{A}}, \qquad P \in \mathbb{P}^k,\ P(0)=1$$

choosing $P(x) = T_k(x;m,M)/T_k(0;m,M)$ from the fact that $|T_k(x;m,M)| \leq 1$ for $x \in [m,M]$ we have

$$\|\boldsymbol{e}_k\|_{\boldsymbol{A}} \leq T_k(0;m,M)^{-1}\|\boldsymbol{e}_0\|_{\boldsymbol{A}} = T_k\left(\frac{M+m}{M-m}\right)^{-1}\|\boldsymbol{e}_0\|_{\boldsymbol{A}}$$

observe that $\frac{M+m}{M-m} = \frac{\kappa+1}{\kappa-1}$ and

$$T_k\left(\frac{\kappa+1}{\kappa-1}\right)^{-1} = 2\left[\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^k + \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k\right]^{-1}$$

finally notice that $\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k \to 0$ as $k \to \infty$. □

Outline

Preconditioning

Problem (Preconditioned linear system)

Given $\boldsymbol{A}, \boldsymbol{P} \in \mathbb{R}^{n\times n}$, with $\boldsymbol{A}$ an SPD matrix and $\boldsymbol{P}$ non singular matrix and $\boldsymbol{b} \in \mathbb{R}^n$.

Find $\boldsymbol{x}_\star \in \mathbb{R}^n$ such that: $\quad \boldsymbol{P}^{-T}\boldsymbol{A}\boldsymbol{x}_\star = \boldsymbol{P}^{-T}\boldsymbol{b}.$

A good choice for $\boldsymbol{P}$ should be such that $\boldsymbol{M} = \boldsymbol{P}^T\boldsymbol{P} \approx \boldsymbol{A}$, where $\approx$ denotes that $\boldsymbol{M}$ is an approximation of $\boldsymbol{A}$ in some sense to precise later.

Notice that:

- $\boldsymbol{P}$ non singular imply:

$$\boldsymbol{P}^{-T}(\boldsymbol{b}-\boldsymbol{A}\boldsymbol{x}) = \boldsymbol{0} \quad \Longleftrightarrow \quad \boldsymbol{b}-\boldsymbol{A}\boldsymbol{x} = 0;$$

- $\boldsymbol{A}$ SPD imply $\widetilde{\boldsymbol{A}} = \boldsymbol{P}^{-T}\boldsymbol{A}\boldsymbol{P}^{-1}$ is also SPD (obvious proof).

Now we reformulate the preconditioned system:

Problem (Preconditioned linear system)

Given $\boldsymbol{A},\boldsymbol{P}\in\mathbb{R}^{n\times n}$, with $\boldsymbol{A}$ an SPD matrix and $\boldsymbol{P}$ non singular matrix and $\boldsymbol{b}\in\mathbb{R}^n$ the preconditioned problem is the following:

$$\text{Find } \widetilde{\boldsymbol{x}}_\star\in\mathbb{R}^n \text{ such that:}\qquad \widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{x}}_\star=\widetilde{\boldsymbol{b}}$$

where

$$\widetilde{\boldsymbol{A}}=\boldsymbol{P}^{-T}\boldsymbol{A}\boldsymbol{P}^{-1}\qquad\qquad \widetilde{\boldsymbol{b}}=\boldsymbol{P}^{-T}\boldsymbol{b}$$

notice that if $\boldsymbol{x}_\star$ is the solution of the linear system $\boldsymbol{A}\boldsymbol{x}=\boldsymbol{b}$ then $\widetilde{\boldsymbol{x}}_\star=\boldsymbol{P}\boldsymbol{x}_\star$ is the solution of the linear system $\widetilde{\boldsymbol{A}}\boldsymbol{x}=\widetilde{\boldsymbol{b}}$.

PCG: preliminary version

initial step:
$k\leftarrow 0$; $\boldsymbol{x}_0$ assigned;
$\widetilde{\boldsymbol{x}}_0\leftarrow\boldsymbol{P}\boldsymbol{x}_0$; $\widetilde{\boldsymbol{r}}_0\leftarrow\widetilde{\boldsymbol{b}}-\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{x}}_0$; $\widetilde{\boldsymbol{p}}_1\leftarrow\widetilde{\boldsymbol{r}}_0$;
while $\|\widetilde{\boldsymbol{r}}_k\|>\epsilon$ **do**
  $k\leftarrow k+1$;
  Conjugate direction method
  $\widetilde{\alpha}_k\leftarrow\dfrac{\widetilde{\boldsymbol{r}}_{k-1}^T\widetilde{\boldsymbol{r}}_{k-1}}{\widetilde{\boldsymbol{p}}_k^T\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{p}}_k}$;
  $\widetilde{\boldsymbol{x}}_k\leftarrow\widetilde{\boldsymbol{x}}_{k-1}+\widetilde{\alpha}_k\widetilde{\boldsymbol{p}}_k$;
  $\widetilde{\boldsymbol{r}}_k\leftarrow\widetilde{\boldsymbol{r}}_{k-1}-\widetilde{\alpha}_k\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{p}}_k$;
  Residual orthogonalization
  $\widetilde{\beta}_k\leftarrow\dfrac{\widetilde{\boldsymbol{r}}_k^T\widetilde{\boldsymbol{r}}_k}{\widetilde{\boldsymbol{r}}_{k-1}^T\widetilde{\boldsymbol{r}}_{k-1}}$;
  $\widetilde{\boldsymbol{p}}_{k+1}\leftarrow\widetilde{\boldsymbol{r}}_k+\widetilde{\beta}_k\widetilde{\boldsymbol{p}}_k$;
end while
final step
$\boldsymbol{P}^{-1}\widetilde{\boldsymbol{x}}_k$;

Conjugate gradient algorithm applied to $\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{x}}=\widetilde{\boldsymbol{b}}$ require the evaluation of thing like:

$$\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{p}}_k=\boldsymbol{P}^{-T}\boldsymbol{A}\boldsymbol{P}^{-1}\widetilde{\boldsymbol{p}}_k.$$

this can be done without evaluate directly the matrix $\widetilde{\boldsymbol{A}}$, by the following operations:

1. solve $\boldsymbol{P}\boldsymbol{s}'_k=\widetilde{\boldsymbol{p}}_k$ for $\boldsymbol{s}'_k=\boldsymbol{P}^{-1}\widetilde{\boldsymbol{p}}_k$;
2. evaluate $\boldsymbol{s}''_k=\boldsymbol{A}\boldsymbol{s}'_k$;
3. solve $\boldsymbol{P}^T\boldsymbol{s}'''_k=\boldsymbol{s}''_k$ for $\boldsymbol{s}'''_k=\boldsymbol{P}^{-T}\boldsymbol{s}''$.

Step 1 and 3 require the solution of two auxiliary linear system. This is not a big problem if $\boldsymbol{P}$ and $\boldsymbol{P}^T$ are triangular matrices (see e.g. incomplete Cholesky).

However... we can reformulate the algorithm using only the matrices $\boldsymbol{A}$ and $\boldsymbol{P}$!

Definition

For all $k\geq 1$, we introduce the vector $\boldsymbol{q}_k=\boldsymbol{P}^{-1}\widetilde{\boldsymbol{p}}$.

Observation

If the vectors $\widetilde{\boldsymbol{p}}_1$, $\widetilde{\boldsymbol{p}}_2$, ..., $\widetilde{\boldsymbol{p}}_k$ for all $1\leq k\leq n$ are $\widetilde{\boldsymbol{A}}$-conjugate, then the corresponding vectors $\boldsymbol{q}_1$, $\boldsymbol{q}_2$, ... $\boldsymbol{q}_k$ are $\boldsymbol{A}$-conjugate. In fact:

$$\boldsymbol{q}_j^T\boldsymbol{A}\boldsymbol{q}_i=\underbrace{\widetilde{\boldsymbol{p}}_j^T\boldsymbol{P}^{-T}}_{=\boldsymbol{q}_j^T}\boldsymbol{A}\underbrace{\boldsymbol{P}^{-1}\widetilde{\boldsymbol{p}}_i}_{=\boldsymbol{q}_j^T}=\widetilde{\boldsymbol{p}}_j^T\underbrace{\widetilde{\boldsymbol{A}}}_{=\boldsymbol{P}^{-T}\boldsymbol{A}\boldsymbol{P}^{-1}}\widetilde{\boldsymbol{p}}_i=0,\qquad \text{if } i\neq j,$$

that is a consequence of $\widetilde{\boldsymbol{A}}$-conjugation of vectors $\widetilde{\boldsymbol{p}}_i$.

Definition

For all $k \geq 1$, we introduce the vectors

$$\boldsymbol{x}_k = \boldsymbol{x}_{k-1} + \widetilde{\alpha}_k \boldsymbol{q}_k.$$

Observation

If we assume, by construction, $\widetilde{\boldsymbol{x}}_0 = \boldsymbol{P}\boldsymbol{x}_0$, then we have

$$\widetilde{\boldsymbol{x}}_k = \boldsymbol{P}\boldsymbol{x}_k, \qquad \text{for all } k \text{ with } 1 \leq k \leq n.$$

In fact, if $\widetilde{\boldsymbol{x}}_{k-1} = \boldsymbol{P}\boldsymbol{x}_{k-1}$ (inductive hypothesis), then

$$\begin{aligned}
\widetilde{\boldsymbol{x}}_k &= \widetilde{\boldsymbol{x}}_{k-1} + \widetilde{\alpha}_k \widetilde{\boldsymbol{p}}_k && \text{[preconditioned CG]}\\
&= \boldsymbol{P}\boldsymbol{x}_{k-1} + \widetilde{\alpha}_k \boldsymbol{P}\boldsymbol{q}_k && \text{[inductive Hyp. defs of } \boldsymbol{q}_k\text{]}\\
&= \boldsymbol{P}\left(\boldsymbol{x}_{k-1} + \widetilde{\alpha}_k \boldsymbol{q}_k\right) && \text{[obvious]}\\
&= \boldsymbol{P}\boldsymbol{x}_k && \text{[defs. of } \boldsymbol{x}_k\text{]}
\end{aligned}$$

Observation

Because $\widetilde{\boldsymbol{x}}_k = \boldsymbol{P}\boldsymbol{x}_k$ for all $k \geq 0$, we have the recurrence between the corresponding residue $\widetilde{\boldsymbol{r}}_k = \widetilde{\boldsymbol{b}} - \widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{x}}$ and $\boldsymbol{r}_k = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k$:

$$\widetilde{\boldsymbol{r}}_k = \boldsymbol{P}^{-T}\boldsymbol{r}_k.$$

In fact,

$$\begin{aligned}
\widetilde{\boldsymbol{r}}_k &= \widetilde{\boldsymbol{b}} - \widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{x}}_k, && \text{[defs. of } \widetilde{\boldsymbol{r}}_k\text{]}\\
&= \boldsymbol{P}^{-T}\boldsymbol{b} - \boldsymbol{P}^{-T}\boldsymbol{A}\boldsymbol{P}^{-1}\boldsymbol{P}\boldsymbol{x}_k, && \text{[defs. of } \widetilde{\boldsymbol{b}}, \widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{x}}_k\text{]}\\
&= \boldsymbol{P}^{-T}\left(\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k\right), && \text{[obvious]}\\
&= \boldsymbol{P}^{-T}\boldsymbol{r}_k. && \text{[defs. of } \boldsymbol{r}_k\text{]}
\end{aligned}$$

Definition

For all k, with $1 \leq k \leq n$, the vector $\boldsymbol{z}_k$ is the solution of the linear system

$$\boldsymbol{M}\boldsymbol{z}_k = \boldsymbol{r}_k.$$

where $\boldsymbol{M} = \boldsymbol{P}^T\boldsymbol{P}$. Formally,

$$\boldsymbol{z}_k = \boldsymbol{M}^{-1}\boldsymbol{r}_k = \boldsymbol{P}^{-1}\boldsymbol{P}^{-T}\boldsymbol{r}_k.$$

Using the vectors $\{\boldsymbol{z}_k\}$,

- we can express $\widetilde{\alpha}_k$ and $\widetilde{\beta}_k$ in terms of $\boldsymbol{A}$, the residual $\boldsymbol{r}_k$, and conjugate direction $\boldsymbol{q}_k$;
- we can build a recurrence relation for the $\boldsymbol{A}$-conjugate directions $\boldsymbol{q}_k$.

Observation

$$\widetilde{\alpha}_k = \frac{\widetilde{\boldsymbol{r}}_{k-1}^T\widetilde{\boldsymbol{r}}_{k-1}}{\widetilde{\boldsymbol{p}}_k^T\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{p}}_k} = \frac{\boldsymbol{r}_{k-1}\boldsymbol{P}^{-1}\boldsymbol{P}^{-T}\boldsymbol{r}_{k-1}}{\boldsymbol{q}_k^T\boldsymbol{P}^T\boldsymbol{P}^{-T}\boldsymbol{A}\boldsymbol{P}^{-1}\boldsymbol{P}\boldsymbol{q}_k} = \frac{\boldsymbol{r}_{k-1}\boldsymbol{M}^{-1}\boldsymbol{r}_{k-1}}{\boldsymbol{q}_k\boldsymbol{A}\boldsymbol{q}_k},$$

$$= \boxed{\frac{\boldsymbol{r}_{k-1}\boldsymbol{z}_{k-1}}{\boldsymbol{q}_k\boldsymbol{A}\boldsymbol{q}_k}.}$$

Observation

$$\widetilde{\beta}_k = \frac{\widetilde{\boldsymbol{r}}_k^T\widetilde{\boldsymbol{r}}_k}{\widetilde{\boldsymbol{r}}_{k-1}^T\widetilde{\boldsymbol{r}}_{k-1}} = \frac{\boldsymbol{r}_k^T\boldsymbol{P}^{-1}\boldsymbol{P}^{-T}\boldsymbol{r}_k}{\boldsymbol{r}_{k-1}^T\boldsymbol{P}^{-1}\boldsymbol{P}^{-T}\boldsymbol{r}_{k-1}} = \frac{\boldsymbol{r}_k^T\boldsymbol{M}^{-1}\boldsymbol{r}_k}{\boldsymbol{r}_{k-1}^T\boldsymbol{M}^{-1}\boldsymbol{r}_{k-1}},$$

$$= \boxed{\frac{\boldsymbol{r}_k^T\boldsymbol{z}_k}{\boldsymbol{r}_{k-1}^T\boldsymbol{z}_{k-1}}.}$$

Observation

Using the vector $z_k = M^{-1}r_k$, the following recurrence is true

$$q_{k+1} = z_k + \widetilde{\beta}_k q_k$$

In fact:

$$\widetilde{p}_{k+1} = \widetilde{r}_k + \widetilde{\beta}_k \widetilde{p}_k \quad \text{[preconditioned CG]}$$

$$P^{-1}\widetilde{p}_{k+1} = P^{-1}\widetilde{r}_k + \widetilde{\beta}_k P^{-1}\widetilde{p}_k \quad \text{[left mult } P^{-1}\text{]}$$

$$P^{-1}\widetilde{p}_{k+1} = P^{-1}P^{-T}r_k + \widetilde{\beta}_k P^{-1}\widetilde{p}_k \quad [r_{k+1} = P^{-T}r_{k+1}]$$

$$P^{-1}\widetilde{p}_{k+1} = M^{-1}r_k + \widetilde{\beta}_k P^{-1}\widetilde{p}_k \quad [M^{-1} = P^{-1}P^{-T}]$$

$$q_{k+1} = z_k + \widetilde{\beta}_k q_k \quad [q_k = P^{-1}\widetilde{p}_k]$$

PCG: final version

initial step:
$k \leftarrow 0$; x_0 assigned;
$r_0 \leftarrow b - Ax_0$; $q_1 \leftarrow r_0$;
while $\|z_k\| > \epsilon$ **do**
  $k \leftarrow k + 1$;
  Conjugate direction method
  $\widetilde{\alpha}_k \leftarrow \dfrac{r_{k-1}^T z_{k-1}}{q_k^T A q_k}$;
  $x_k \leftarrow x_{k-1} + \widetilde{\alpha}_k q_k$;
  $r_k \leftarrow r_{k-1} - \widetilde{\alpha}_k A q_k$;
  Preconditioning
  $z_k = M^{-1} r_k$;
  Residual orthogonalization
  $\widetilde{\beta}_k \leftarrow \dfrac{r_k^T z_k}{r_{k-1}^T z_{k-1}}$;
  $q_{k+1} \leftarrow z_k + \widetilde{\beta}_k q_k$;
end while

Outline

Nonlinear Conjugate Gradient extension

1. The conjugate gradient algorithm can be extended for nonlinear minimization.
2. Fletcher and Reeves extend CG for the minimization of a general non linear function $\mathsf{f}(x)$ as follows:
   1. Substitute the evaluation of α_k by an line search
   2. Substitute the residual r_k with the gradient $\nabla \mathsf{f}(x_k)$
3. We also translate the index for the search direction p_k to be more consistent with the gradients. The resulting algorithm is in the next slide

Fletcher and Reeves Nonlinear Conjugate Gradient

initial step:
$k \leftarrow 0$; $\boldsymbol{x}_0$ assigned;
$f_0 \leftarrow \mathsf{f}(\boldsymbol{x}_0)$; $\boldsymbol{g}_0 \leftarrow \nabla \mathsf{f}(\boldsymbol{x}_0)^T$;
$\boldsymbol{p}_0 \leftarrow -\boldsymbol{g}_0$;
while $\|\boldsymbol{g}_k\| > \epsilon$ **do**
  $k \leftarrow k+1$;
  Conjugate direction method
  Compute α_k by line-search;
  $\boldsymbol{x}_k \leftarrow \boldsymbol{x}_{k-1} + \alpha_k \boldsymbol{p}_{k-1}$;
  $\boldsymbol{g}_k \leftarrow \nabla \mathsf{f}(\boldsymbol{x}_k)^T$;
  Residual orthogonalization
  $\beta_k^{FR} \leftarrow \dfrac{\boldsymbol{g}_k^T \boldsymbol{g}_k}{\boldsymbol{g}_{k-1}^T \boldsymbol{g}_{k-1}}$;
  $\boldsymbol{p}_k \leftarrow -\boldsymbol{g}_k + \beta_k^{FR} \boldsymbol{p}_{k-1}$;
end while

1. To ensure convergence and apply Zoutendijk global convergence theorem we need to ensure that $\boldsymbol{p}_k$ is a descent direction.
2. $\boldsymbol{p}_0$ is a descent direction by construction, for $\boldsymbol{p}_k$ we have

$$\boldsymbol{g}_k^T \boldsymbol{p}_k = -\|\boldsymbol{g}_k\|^2 + \beta_k^{FR} \boldsymbol{g}_k^T \boldsymbol{p}_{k-1}$$

   if the line-search is exact than $\boldsymbol{g}_k^T \boldsymbol{p}_{k-1} = 0$ because $\boldsymbol{p}_{k-1}$ is the direction of the line-search. So by induction $\boldsymbol{p}_k$ is a descent direction.
3. Exact line-search is expensive, however if we use inexact line-search with strong Wolfe conditions
   1. sufficient decrease: $\mathsf{f}(\boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k) \leq \mathsf{f}(\boldsymbol{x}_k) + c_1\, \alpha_k \nabla \mathsf{f}(\boldsymbol{x}_k) \boldsymbol{p}_k$;
   2. curvature condition: $|\nabla \mathsf{f}(\boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k) \boldsymbol{p}_k| \leq c_2\, |\nabla \mathsf{f}(\boldsymbol{x}_k) \boldsymbol{p}_k|$.

   with $0 < c_1 < c_2 < 1/2$ then we can prove that $\boldsymbol{p}_k$ is a descent direction.

The previous consideration permits to say that Fletcher and Reeves nonlinear conjugate gradient method with strong Wolfe line-search is globally convergent[1]
To prove globally convergence we need the following lemma:

Lemma (descent direction bound)

Suppose we apply Fletcher and Reeves nonlinear conjugate gradient method to $\mathsf{f}(\boldsymbol{x})$ with strong Wolfe line-search with $0 < c_2 < 1/2$. The the method generates descent direction $\boldsymbol{p}_k$ that satisfy the following inequality

$$-\frac{1}{1-c_2} \leq \frac{\boldsymbol{g}_k^T \boldsymbol{p}_k}{\|\boldsymbol{g}_k\|^2} \leq -\frac{1-2c_2}{1-c_2}, \qquad k = 0, 1, 2, \ldots$$

[1] globally here means that Zoutendijk like theorem apply

Proof. (1/3).

The proof is by induction. First notice that the function

$$t(\xi) = \frac{2\xi - 1}{1 - \xi}$$

is monotonically increasing on the interval $[0, 1/2]$ and that $t(0) = -1$ and $t(1/2) = 0$. Hence, because of $c_2 \in (0, 1/2)$ we have:

$$-1 < \frac{2c_2 - 1}{1 - c_2} < 0. \qquad (\star)$$

base of induction $k = 0$: For $k = 0$ we have $\boldsymbol{p}_0 = -\boldsymbol{g}_0$ so that $\boldsymbol{g}_0^T \boldsymbol{p}_0 / \|\boldsymbol{g}_0\|^2 = -1$. From $(\star)$ the lemma inequality is trivially satisfied.

Proof. (2/3).

Using update direction formula's of the algorithm:

$$\beta_k^{FR} = \frac{\boldsymbol{g}_k^T \boldsymbol{g}_k}{\boldsymbol{g}_{k-1}^T \boldsymbol{g}_{k-1}} \qquad \boldsymbol{p}_k = -\boldsymbol{g}_k + \beta_k^{FR} \boldsymbol{p}_{k-1}$$

we can write

$$\frac{\boldsymbol{g}_k^T \boldsymbol{p}_k}{\|\boldsymbol{g}_k\|^2} = -1 + \beta_k^{FR} \frac{\boldsymbol{g}_k^T \boldsymbol{p}_{k-1}}{\|\boldsymbol{g}_k\|^2} = -1 + \frac{\boldsymbol{g}_k^T \boldsymbol{p}_{k-1}}{\|\boldsymbol{g}_{k-1}\|^2}$$

and by using second strong Wolfe condition:

$$-1 + c_2 \frac{\boldsymbol{g}_{k-1}^T \boldsymbol{p}_{k-1}}{\|\boldsymbol{g}_{k-1}\|^2} \leq \frac{\boldsymbol{g}_k^T \boldsymbol{p}_k}{\|\boldsymbol{g}_k\|^2} \leq -1 - c_2 \frac{\boldsymbol{g}_{k-1}^T \boldsymbol{p}_{k-1}}{\|\boldsymbol{g}_{k-1}\|^2}$$

Proof. (3/3).

by induction we have

$$\frac{1}{1-c_2} \geq -\frac{\boldsymbol{g}_{k-1}^T \boldsymbol{p}_{k-1}}{\|\boldsymbol{g}_{k-1}\|^2} > 0$$

so that

$$\frac{\boldsymbol{g}_k^T \boldsymbol{p}_k}{\|\boldsymbol{g}_k\|^2} \leq -1 - c_2 \frac{\boldsymbol{g}_{k-1}^T \boldsymbol{p}_{k-1}}{\|\boldsymbol{g}_{k-1}\|^2} \leq -1 + c_2 \frac{1}{1-c_2} = \frac{2c_2 - 1}{1-c_2}$$

and

$$\frac{\boldsymbol{g}_k^T \boldsymbol{p}_k}{\|\boldsymbol{g}_k\|^2} \geq -1 + c_2 \frac{\boldsymbol{g}_{k-1}^T \boldsymbol{p}_{k-1}}{\|\boldsymbol{g}_{k-1}\|^2} \geq -1 - c_2 \frac{1}{1-c_2} = -\frac{1}{1-c_2}$$

□

1. The inequality of the the previous lemma can be written as:

$$\frac{1}{1-c_2} \frac{\|\boldsymbol{g}_k\|}{\|\boldsymbol{p}_k\|} \geq -\frac{\boldsymbol{g}_k^T \boldsymbol{p}_k}{\|\boldsymbol{g}_k\| \, \|\boldsymbol{p}_k\|} \geq \frac{1-2c_2}{1-c_2} \frac{\|\boldsymbol{g}_k\|}{\|\boldsymbol{p}_k\|} > 0$$

2. Remembering the Zoutendijk theorem we have

$$\sum_{k=1}^{\infty} (\cos\theta_k)^2 \|\boldsymbol{g}_k\|^2 < \infty, \quad \text{where} \quad \cos\theta_k = -\frac{\boldsymbol{g}_k^T \boldsymbol{p}_k}{\|\boldsymbol{g}_k\| \, \|\boldsymbol{p}_k\|}$$

3. so that if $\|\boldsymbol{g}_k\| / \|\boldsymbol{p}_k\|$ is bounded from below we have that $\cos\theta_k \geq \delta$ for all k and then from Zoutendijk theorem the scheme converge.
4. Unfortunately this bound cant be proved so that Zoutendijk theorem cant be applied directly. However it is possible to prove a weaker results, i.e. that $\liminf_{k\to\infty} \|\boldsymbol{g}_k\| = 0$!

Convergence of Fletcher and Reeves method

Assumption (Regularity assumption)

We assume $\mathsf{f} \in \mathtt{C}^1(\mathbb{R}^n)$ *with Lipschitz continuous gradient, i.e. there exists* $\gamma > 0$ *such that*

$$\left\|\nabla \mathsf{f}(\boldsymbol{x})^T - \nabla \mathsf{f}(\boldsymbol{y})^T\right\| \leq \gamma \|\boldsymbol{x} - \boldsymbol{y}\|, \qquad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$$

Theorem (Convergence of Fletcher and Reeves method)

Suppose the method of Fletcher and Reeves is implemented with strong Wolfe line-search with $0 < c_1 < c_2 < 1/2$. If $f(\boldsymbol{x})$ and $\boldsymbol{x}_0$ satisfy the previous regularity assumptions, then

$$\liminf_{k\to\infty} \|\boldsymbol{g}_k\| = 0$$

Proof. (1/4).

From previous Lemma we have

$$\cos\theta_k \geq \frac{1}{1-c_2}\frac{\|\boldsymbol{g}_k\|}{\|\boldsymbol{p}_k\|} \qquad k = 1, 2, \ldots$$

substituting in Zoutendijk condition we have $\sum_{k=1}^{\infty} \frac{\|\boldsymbol{g}_k\|^4}{\|\boldsymbol{p}_k\|^2} < \infty$.

The proof is by contradiction. in fact if theorem is not true than the series diverge. Next we want to bound $\|\boldsymbol{p}_k\|$.

Proof. (bounding $\|\boldsymbol{p}_k\|$) (2/4).

Using second Wolfe condition and previous Lemma

$$\left|\boldsymbol{g}_k^T \boldsymbol{p}_{k-1}\right| \leq -c_2 \boldsymbol{g}_k^T \boldsymbol{p}_{k-1} \leq \frac{c_2}{1-c_2}\|\boldsymbol{g}_{k-1}\|^2$$

using $\boldsymbol{p}_k = -\boldsymbol{g}_k + \beta_k^{FR}\boldsymbol{p}_{k-1}$ we have

$$\|\boldsymbol{p}_k\|^2 \leq \|\boldsymbol{g}_k\|^2 + 2\beta_k^{FR}\left|\boldsymbol{g}_k^T\boldsymbol{p}_{k-1}\right| + (\beta_k^{FR})^2\|\boldsymbol{p}_{k-1}\|^2$$

$$\leq \|\boldsymbol{g}_k\|^2 + \frac{2c_2}{1-c_2}\beta_k^{FR}\|\boldsymbol{g}_{k-1}\|^2 + (\beta_k^{FR})^2\|\boldsymbol{p}_{k-1}\|^2$$

recall that $\beta_k^{FR} = \|\boldsymbol{g}_k\|^2 / \|\boldsymbol{g}_{k-1}\|^2$ then

$$\|\boldsymbol{p}_k\|^2 \leq \frac{1+c_2}{1-c_2}\|\boldsymbol{g}_k\|^2 + (\beta_k^{FR})^2\|\boldsymbol{p}_{k-1}\|^2$$

Proof. (bounding $\|\boldsymbol{p}_k\|$) (3/4).

setting $c_3 = \frac{1+c_2}{1-c_2}$ and using repeatedly the last inequality we obtain:

$$\|\boldsymbol{p}_k\|^2 \leq c_3\|\boldsymbol{g}_k\|^2 + (\beta_k^{FR})^2\big(c_3\|\boldsymbol{g}_{k-1}\|^2 + (\beta_{k-1}^{FR})^2\|\boldsymbol{p}_{k-2}\|^2\big)$$

$$= c_3\|\boldsymbol{g}_k\|^4\left(\|\boldsymbol{g}_k\|^{-2} + \|\boldsymbol{g}_{k-1}\|^{-2}\right) + \frac{\|\boldsymbol{g}_k\|^4}{\|\boldsymbol{g}_{k-2}\|^4}\|\boldsymbol{p}_{k-2}\|^2$$

$$\leq c_3\|\boldsymbol{g}_k\|^4\left(\|\boldsymbol{g}_k\|^{-2} + \|\boldsymbol{g}_{k-1}\|^{-2} + \|\boldsymbol{g}_{k-2}\|^{-2}\right)$$

$$+ \frac{\|\boldsymbol{g}_k\|^4}{\|\boldsymbol{g}_{k-3}\|^4}\|\boldsymbol{p}_{k-3}\|^2$$

$$\leq c_3\|\boldsymbol{g}_k\|^4\sum_{j=1}^{k}\|\boldsymbol{g}_j\|^{-2}$$

Proof. (4/4).

Suppose now by contradiction there exists $\delta > 0$ such that $\|\boldsymbol{g}_k\| \geq \delta$ [a] by using the regularity assumptions we have

$$\|\boldsymbol{p}_k\|^2 \leq c_3\|\boldsymbol{g}_k\|^4\sum_{j=1}^{k}\|\boldsymbol{g}_j\|^{-2} \leq c_3\|\boldsymbol{g}_k\|^4\delta^{-2}k$$

Substituting in Zoutendijk condition we have

$$\infty > \sum_{k=1}^{\infty}\frac{\|\boldsymbol{g}_k\|^4}{\|\boldsymbol{p}_k\|^2} \geq \frac{\delta^2}{c_4}\sum_{k=1}^{\infty}\frac{1}{k} = \infty$$

this contradict assumption. □

[a] the correct assumption is that there exists k_0 such that $\|\boldsymbol{g}_k\| \geq \delta$ for $k \geq k_0$ but this complicate a little bit the following inequality without introducing new idea.

Weakness of Fletcher and Reeves method

- Suppose that $\boldsymbol{p}_k$ is a bad search direction, i.e. $\cos\theta_k \approx 0$.
- From the descent direction bound Lemma (see slide 91) we have

$$\frac{1}{1-c_2}\frac{\|\boldsymbol{g}_k\|}{\|\boldsymbol{p}_k\|} \geq \cos\theta_k \geq \frac{1-2c_2}{1-c_2}\frac{\|\boldsymbol{g}_k\|}{\|\boldsymbol{p}_k\|} > 0$$

- so that to have $\cos\theta_k \approx 0$ we needs $\|\boldsymbol{p}_k\| \gg \|\boldsymbol{g}_k\|$.
- since $\boldsymbol{p}_k$ is a bad direction near orthogonal to $\boldsymbol{g}_k$ it is likely that the step is small and $\boldsymbol{x}_{k+1} \approx \boldsymbol{x}_k$. If so we have also $\boldsymbol{g}_{k+1} \approx \boldsymbol{g}_k$ and $\beta_{k+1}^{FR} \approx 1$.
- but remember that $\boldsymbol{p}_{k+1} \leftarrow -\boldsymbol{g}_{k+1} + \beta_{k+1}^{FR}\boldsymbol{p}_k$, so that $\boldsymbol{p}_{k+1} \approx \boldsymbol{p}_k$.
- This means that a long sequence of unproductive iterates will follows.

Polack and Ribiére Nonlinear Conjugate Gradient

1. The previous problem can be elided if we restart anew when the iterate stagnate.
2. Restarting is obtained by simply set $\beta_k^{FR} = 0$.
3. A more elegant solution can be obtained with a new definition of β_k due to Polack and Ribiére is the following:

$$\beta_k^{PR} = \frac{\boldsymbol{g}_k^T(\boldsymbol{g}_k - \boldsymbol{g}_{k-1})}{\boldsymbol{g}_{k-1}^T\boldsymbol{g}_{k-1}}$$

4. This definition of β_k^{PR} is identical of β_k^{FR} in the case of quadratic function because $\boldsymbol{g}_k^T\boldsymbol{g}_{k-1} = 0$. The definition differs in non linear case and in particular when there is stagnation i.e. $\boldsymbol{g}_k \approx \boldsymbol{g}_{k-1}$ we have $\beta_k^{PR} \approx 0$, i.e. we have an automatic restart.

Polack and Ribiére Nonlinear Conjugate Gradient

initial step:
$k \leftarrow 0$; $\boldsymbol{x}_0$ assigned;
$f_0 \leftarrow \mathsf{f}(\boldsymbol{x}_0)$; $\boldsymbol{g}_0 \leftarrow \nabla\mathsf{f}(\boldsymbol{x}_0)^T$;
$\boldsymbol{p}_0 \leftarrow -\boldsymbol{g}_0$;
while $\|\boldsymbol{g}_k\| > \epsilon$ **do**
  $k \leftarrow k+1$;
  Conjugate direction method
  Compute α_k by line-search;
  $\boldsymbol{x}_k \leftarrow \boldsymbol{x}_{k-1} + \alpha_k\boldsymbol{p}_{k-1}$;
  $\boldsymbol{g}_k \leftarrow \nabla\mathsf{f}(\boldsymbol{x}_k)^T$;
  Residual orthogonalization
  $\beta_k^{PR} \leftarrow \frac{\boldsymbol{g}_k^T(\boldsymbol{g}_k-\boldsymbol{g}_{k-1})}{\boldsymbol{g}_{k-1}^T\boldsymbol{g}_{k-1}}$;
  $\boldsymbol{p}_k \leftarrow -\boldsymbol{g}_k + \beta_k^{PR}\boldsymbol{p}_{k-1}$;
end while

Weakness of Polack and Ribiére method (1/2)

- Although the modification is minimal, for the Polack and Ribiére method with strong Wolfe line-search it can happen that $\boldsymbol{p}_k$ is not a descent direction.
- If $\boldsymbol{p}_k$ is not a descent direction we can restart i.e. set $\beta_k^{PR} = 0$ or modify β_k^{PR} as follows

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

this new coefficient with a modified Wolfe line-search ensure that $\boldsymbol{p}_k$ is a descent direction.

Weakness of Polack and Ribiére method (2/2)

- Polack and Ribiére choice on the average perform better than Fletcher and Reeves but there is not convergence results!
- Although there is not convergence results there is a negative results due to Powell:

Theorem

Consider the Polack and Ribiére method with exact line-search. There exists a twice continuously differentiable function f : $\mathbb{R}^3 \mapsto \mathbb{R}$ *and a starting point* $\boldsymbol{x}_0$ *such that the sequence of gradients* $\{ \|\boldsymbol{g}_k\| \}$ *is bounded away from zero.*

- However is spite of this results Polack and Ribiére is the first choice among conjugate direction methods.

Other choices

- There are many other modification of the coefficient β_k that collapse to the same coefficient in the case o quadratic function. One important choice is the Hestenes and Stiefel choice

$$\beta_k^{HS} = \frac{\boldsymbol{g}_k^T(\boldsymbol{g}_k - \boldsymbol{g}_{k-1})}{(\boldsymbol{g}_k^T - \boldsymbol{g}_{k-1}^T)\boldsymbol{p}_{k-1}}$$

- For this choice there is similar convergence results of Fletcher and Reeves and similar performance.

References

- J. E. Dennis, Jr. and Robert B. Schnabel
  Numerical Methods for Unconstrained Optimization and Nonlinear Equations
  SIAM, Classics in Applied Mathematics, **16**, 1996.
- J. Nocedal and S. J. Wrigth
  Numerical Optimization
  Springer Series in Operation Research, 1999.
- J. Stoer and R. Bulirsch
  Introduction to numerical analysis
  Springer-Verlag, Texts in Applied Mathematics, **12**, 2002.