Numerical Optimisation: Conjugate gradient methods

f.rullan@cs.uc

Department of Computer Sc
Centre for Medical Image Com
Centre for Inverse Problems
University College London

Lecture 5 & 6

Conjugate gradient: CG

- The linear CG method was proposed by Hestens and Stiefel in 1952 as a direct method for solution of linear systems of equations with positive definite matrix. (It was used to solve 106 difference equations on the Zuse computer at ETH (with a sufficiently accurate answer obtained in 90 iterations each

-

- Renaissance in early 1970 work by John Rei connection to iterative methods. Game ch of CG is determined by the distribution of the e the matrix (preconditioning).

- In top 10 algorithms of 20th century.
- Nonlinear conjugate gradient method proposed by Fletcher and Reeves 1960.

Solution of linear system

$$Ax = b,$$

with A is symmetric positive definite matrix is equivalent to the quad

Both have the same unique solution. In fact

$$\nabla\phi(x) = Ax - b$$

thus the linear system is the 1st order necessary condition (which is also sufficient for strictly convex function ϕ).

A set of non-zero vectors $\{p_1, p_2, \dots, p_\ell\}$ is said to be **conjugate** with respect to the symmetric positive definite matrix A if

/

Conjugate directions are linearly independent

Conjugacy enables us to minimise ϕ in
minimising it along the individual directions in th

Conjugate direction method

Given a starting point x_0 and the set of conjugate directions $\{p_0, p_2, \dots p_{n-1}\}$ let us generate the sequence $\{x_k\}$

$$x_{k+1} = x_k + \alpha p_k,$$

wher

func

$$\alpha_k = -\frac{r_k^{\mathrm{T}} p}{p_k^{\mathrm{T}} A}$$

For any $x_0 \in \mathbb{R}^n$ the sequence converges to the solution $x^\star$ in at most n steps.

Proof: Because $\text{span}\{p_0, p_2, \ldots p_{n-1}\} = \mathbb{R}^n$

$$x^\star - x_0 = \sigma_0 p_0 + \sigma_1 p_1 + \cdots + \sigma_{n-1} p_{n-1}.$$

Multiplying from the left by $p_k^{\mathrm{T}} A$ and using the conjugacy property we obtain σ_k as

$$\sigma_k = \frac{p_k^{\mathrm{T}} A(x^\star - x_0)}{}, \quad k = 0, \ldots, n \quad 1.$$

On th

appr

$$x_k = x_0 + \alpha_0 p_0 + \alpha_1 p_1 +$$

Multiplying from the left by $p_k^{\mathrm{T}} A$ and us

we have $p_k^{\mathrm{T}} A(x_k - x_0) = 0$ and

$$p_k^{\mathrm{T}} A(x^\star - x_0) = p_k^{\mathrm{T}} A(x^\star - x_k) = p_k^{\mathrm{T}}(b - Ax_k) = -p_k^{\mathrm{T}} r_k.$$

Substituting into $\sigma_k = -\frac{p_k^{\mathrm{T}} r_k}{p_k^{\mathrm{T}} A p_k} = \alpha_k$ for $k = 0, \ldots, n-1$.

For any starting point x_0 for the sequence $\{x_k\}$ generated by the conjugate direction method it holds

and $_k$ $\overline{2}$

$\{x : x = x_0 + \mathrm{span}\{p_0,$

Proof: Let's define

$$h(\sigma) = \phi(x_0 + \sigma_0 p_0 + \cdots + \sigma_{k-1} p_{k-1}),$$

where $\sigma = (\sigma_0, \sigma_1, \ldots, \sigma_{k-1})^{\mathrm{T}}$. Since $h(\sigma)$ is a strictly convex quadratic, it has a unique minimiser $\sigma^\star$ that satisfies

$$\frac{\partial h(\sigma^\star)}{}$$

Usin

$$\nabla \phi(x_0 + \sigma^\star p_0 + \cdots + \sigma^\star_{k-1} p_{k-1})^{\mathrm{T}} p_i \qquad 1.$$

Recall that $r(x) = \nabla \phi(x)$, thus for the mini
$\tilde{x} = x_0 + \sigma_0^\star p_0 + \cdots + \sigma^\star_{k-1} p_{k-1}$ on $\{ \quad {}_0 \quad {}_1 \quad\quad {}_{k-1}$
$r(\tilde{x})^{\mathrm{T}} p_i = 0$ as claimed.

By induction:

For $k = 1$, from $x_1 = x_0 + \alpha_0 p_0$ being a minimiser of ϕ along p_0 it follows $r_1^{\mathrm{T}} p_0 = 0$.

Suppose that $r_{k-1}^{\mathrm{T}} p_i = 0$ for $i = 0, 1, \dots, k-2$.

$$r_k = Ax_k - b = A(x_{k-1} + \alpha_{k-1} p_{k-1}) - b = r_{k-1} + \alpha_{k-1} A p_{k-1}.$$

and

by the

For any other $p_i, i = 0, 1, \dots, k-2$ we have

$$p_i^{\mathrm{T}} r_k = p_i^{\mathrm{T}} r_{k-1} + \alpha_{k-1}$$

where the first term disappears because of the indu
and the second because of the conjugacy of p_i. Thus we have shown $r_k^{\mathrm{T}} p_i = 0$ for $i = 0, 1, \dots, k-1$ and the proof is complete.

- So far the discussion was valid for any set of conjugate direction.
- An example are eigenvectors of a symmetric positive definite matrix A which are orthogonal and conjugate w.r.t. A.

it requires to store all the directions to orthogonalise against.

- Conjugate gradient (CG) method has a ver it can compute a new vector p_k u p_{k-1} i.e. it does not need to know the vectors $_{-2}$ while p_k is automatically conjugate to those vectors. This makes CG particularly cheap in terms of computation and memory.

In CG each new direction is chosen as

$$p_k = -r_k + \beta_k p_{k-1},$$

wher

follows from requiring that p_{k-1}, p_k be conjugate
i.e. $p_{k-1}^{\mathrm{T}} A p_k = 0$.

We initialise p_0 with the steepest descent dir

As in the conjugate direction method, we perform successive one dimensional minimisation along each of the search directions.

Given x_0

Set $r_0 = Ax_0 - b$, $p_0 = -r_0$, $k = 0$

while $r_k \neq 0$ **do**

$\beta_{k+1} = \frac{r_{k+1}^{\mathrm{T}} A p_k}{p_k^{\mathrm{T}} A p_k}$

$p_{k+1} = -r_{k+1} + \beta_{k+1} p_k$

$k = k + 1$

end while

Given x_0

Set $r_0 = Ax_0 - b$, $p_0 = -r_0$, $k = 0$

while $r_k \neq 0$ **do**

$$\beta_{k+1} = \frac{r_{k+1}^{\mathrm{T}} r_{k+1}}{r_k^{\mathrm{T}} r_k}$$

$p_{k+1} = -r_{k+1} + \beta_{k+1} p_k$

$k = k + 1$

end while

For the kth iterate of the conjugate gradient method, $x_k \neq x^\star$ the following hold:

$$r_k^T r_i = 0, \quad i = 0, 1, \dots, k-1 \quad (1)$$

$$\text{span}\{r_0, r_1, \dots, r_k\} = \text{span}\{r_0, Ar_0, \dots, A^k r_0\} \quad (2)$$

(3)

(4)

Therefore, the sequence $\{x_k\}$ converge

The proof of this theorem relies on p_0

hold for other choices of p_0.

Note that the gradients r_k are actually orthogonal, while the directions p_k are conjugate, thus the name of conjugate gradients is actually a misnomer.

From the properties of the $k+1$st iterate we have

$$
\begin{aligned}
x_{k+1} &= x_0 + \alpha_0 p_0 + \cdots + \alpha_k p_k && (5)\\
&= x_0 + \gamma_0 r_0 + \gamma_1 A r_0 + \cdots + \gamma_k A^k r_0 && (6)
\end{aligned}
$$

for some γ_i, $i = 0, \ldots, k$.

Let

then

$$x_{k+1} = x_0 + P_k$$

Recall that CG minimises the quadratic function

$x_0 + \mathrm{span}\{p_0, \ldots, p_k\}$ which is the same as

$$
\begin{aligned}
\arg\min \phi(x) &= \arg\min \phi(x) - \phi(x^\star)\\
&= \arg\min \frac{1}{2}(x - x^\star)^{\mathrm{T}} A (x - x^\star) = \arg\min \frac{1}{2}\|x - x^\star\|_A^2
\end{aligned}
$$

CG vs steepest descent with optimal step length

Figure: Wiki: Conjugate gradient method

Thus CG computes the minimising polynomial over all polynomials of degree k

$$\min_{P_k} \|x_0 + P_k(A)r_0 - x^\star\|_A.$$

Observe that similar expressions hold for the error

$$x_k - \ldots \underbrace{(x_0 - x^\star)}_{=r_0} \ldots {}_{k-1} \ldots {}_{0} -$$

and the residual

$$\begin{aligned} r_k &= Ax_k - b = A(x_k - x^\star) = \ldots {}_0 \ldots {}_{k-1} \ldots {}_0 \\ &= \underbrace{A(x_0 - x^\star)}_{=r_0} + AP_{k-1}(A)r_0 = [I + AP_{k-1}(A)]r_0 \end{aligned}$$

Let the eigenvalue decomposition of the symmetric positive definite matrix

$$A = V^{\mathrm{T}}\Lambda V = \sum_{i}^{n} \lambda_i v_i v_i^{\mathrm{T}},$$

with $0 < \lambda_1$ λ_2 λ_n and $v_i, i = 1, \dots, n$ the corre

Sinc

$x_0 -$ $_{i=1}$ $_i$ $_i$

Notice that any eigenvector v_i of A is also a (A) with the corresponding eigenvalue $P_k($

$$P_k(A)v_i = P_k(\lambda_i)v_i, \quad i = 1, \dots, n.$$

Hence

$$x_{k+1} - x^\star = \sum_{i=1}^{n} [1 + \lambda_i P_k(\lambda_i)] \xi_i v_i$$

and

$$\|x_{k+1} - x^\star\|_A^2 = \sum_{i=1}^{n} \lambda_i [1 + \lambda_i P_k(\lambda_i)]^2 \xi_i^2$$

Sinc [illegible]

$$\|x_{k+1} - x^\star\|_A^2 = \min_{P_k} \sum_{i=1} \lambda_i [1 + \ldots]^2 \ldots^2$$

$$\leq \min_{P_k} \max_{1 \leq i \leq n} [1 + \ldots \sum_{i=1} \ldots$$

$$= \min_{P_k} \max_{1 \leq i \leq n} [1 + \lambda_i P_k(\lambda_i)]^2 \|x_0 - x^\star\|_A^2.$$

Theorem If A has only r distinct eigenvalues, then CG will converge to the solution in at most r iterations.

Proof: Suppose the eigenvalues take on distinct r values $\tau_1 < \cdots < \tau_r$ and define a polynomial

$$Q_r(\lambda) = \frac{(-1)^r}{\tau\ \tau\ \ldots \tau}(\lambda - \tau_1)\ldots(\lambda - \tau_r)$$

and r

$r-1$ r

is of degree $r-1$ and we have

$$\begin{aligned} 0 &\leq \min_{P_{r-1}} \max_{1\leq i\leq n} [1 + \lambda_i P_{r-1}(\quad i \\ &\leq \max_{1\leq i\leq n} [1 + \lambda_i \bar{P}_{r-1}(\lambda_i)]^2 = \max_{1\leq i\leq n} Q_r^2(\lambda_i) = 0 \end{aligned}$$

and $\|x_r - x^\star\|_A^2 = 0$ and hence $x_r = x^\star$.

Theorem If A has eigenvalues $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$, we have that

$$\|x_{k+1} - x^\star\|_A^2 \le \left(\frac{\lambda_{n-k} - \lambda_1}{\lambda_{n-k} + \lambda_1}\right)^2 \|x_0 - x^\star\|_A^2.$$

Proof idea: Choose polynomial $\bar{P}_k$ such that

Q_{k+} s

λ_n, λ nd λ_1.

It can b on the

remaining eigenvalues $\lambda_1, \ldots, \lambda_{n-k}$ is $\frac{\lambda_{n-k} - \lambda_1}{}$

Theorem In terms of condition number

$\kappa(A) = \|A\|_2 \|A^{-1}\|_2 = \lambda_n / \lambda_1$, we have tha

$$\|x_k - x^\star\|_A \le 2 \left(\frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1}\right)^k \|x_0 - x^\star\|_A.$$

We can accelerate CG through transformations which cluster eigenvalues. This process is known as **preconditioning**.

We perform a change of variables $\hat{x} = Cx$.

The

Minimising $\hat{\phi}$ is equivalent to solving the syste
equations

$$(C^{-\mathrm{T}} A C^{-1})\hat{x} = (C^{-\mathrm{T}} b)$$

and the convergence rate of CG depends on the eigenvalues of $C^{-\mathrm{T}} A C^{-1}$.

It is not necessary to carry out the transforms explicitly. We can apply CG to $\hat{\phi}$ in terms of $\hat{x}$ and then invert the transformations to reexpress all the equations in terms of the original variable x.

In fact, the **preconditioned CG** algorithm does not use the facto

If we se

The properties of CG generalise, in particular for P

$$r_i^T M^{-1} r_j = 0,$$

Given x_0, preconditioner M

Set $r_0 = Ax_0 - b$,

Solve $My_0 = r_0$

$p_0 = \quad y_0$, $k = 0$

w

$r_{k+1} = r_k + \alpha_k A p_k$

Solve $My_{k+1} = r_{k+1}$

$\beta_{k+1} = \frac{r_{k+1}^T y_{k+1}}{r_k^T y_k}$

$p_{k+1} = -y_{k+1} + \beta_{k+1} p_k$

$k = k + 1$

end while

Nonlinear conjugate gradient

Recall that CG can be interpreted as a minimiser of a quadratic convex function

$$\phi(x) = \frac{1}{2}x^{\mathrm{T}}Ax - x^{\mathrm{T}}b$$

Can t f?

Reca

- step length α_k minimises ϕ along
  For general f: compute α_k using li

$$\alpha_k = \min_{\alpha} f(x_k \quad\quad k$$

- $r = Ax - b = \nabla\phi(x)$.
  For general function f: $r \to \nabla f$

Given x_0
Evaluate $f_0 = f(x_0)$, $\nabla f_0 = \nabla f(x_0)$
Set $p_0 = \quad f_0$, $k = 0$
w /

$$\beta_{k+1} = \frac{\nabla f_{k+1}^{\mathrm{T}} \nabla f_{k+1}}{\nabla f_k^{\mathrm{T}} \nabla f_k}$$

$p_{k+1} = -\nabla f_{k+1} + \beta_{k+1} p_k$
$k = k + 1$

end while

Descent direction

Is p_k a descent direction?

$$\nabla f_k^{\mathrm{T}} p_k = -\nabla f_k^{\mathrm{T}} \nabla f_k + \beta_k \nabla f_k^{\mathrm{T}} p_{k-1} \overset{?}{<} 0$$

If α_{k-1} is a local minimiser along p_{k-1}, $\nabla f_k^{\mathrm{T}} p_{k-1} = 0$ and

thus

If the linear search is not exact, due to the second term $\beta_k \nabla f_k^{\mathrm{T}} p_{k-1}$, p_k may fail to be a descent directi

avoided by requiring that the step length

Wolfe conditions

$$\begin{aligned} f(x_k + \alpha_k p_k) &\leq f(x_k) + c_1 \alpha_k \nabla f_k^{\mathrm{T}} p_k, \\ |\nabla f(x_k + \alpha_k p_k)^{\mathrm{T}} p_k| &\leq -c_2 \nabla f_k^{\mathrm{T}} p_k \end{aligned}$$

with $0 < c_1 < c_2 < \frac{1}{2}$.

Let f be twice continuously differentiable, and the level set $\{x : f(x) \leq f(x_0)\}$ is bounded. If the step length α_k in the FR algorithm satisfies strong Wolfe conditions with $0 < c_2 < \frac{1}{2}$, then the method generates descent directions p_k that satisfy

(7)

Proof: First note that the upper bound (2

monotonically increases for $c_2 \in (0, \frac{1}{2})$ a

$-1 < (2c_2 - 1)/(1 - c_2) < 0$. Thus Lem is a

descent direction $\nabla f_k^{\mathrm{T}} p_k < 0$.

The inequalities can be shown by induction using the form of the update the second strong Wolfe condition.

Induction:

$k = 0: \quad p_0 = -\nabla f_0 \to \frac{\nabla f_0^{\mathrm{T}} p_0}{\|\nabla f_0\|^2} = -1$ and (7) holds.

Assume (7) holds for some $k \geq 1$. From β^{FR}_{k+1} we have

$$\frac{\nabla f_{k+1}^{\mathrm{T}} p_{k+1}}{\|\nabla f_{k+1}\|^2} = -1 + \beta^{FR}_{k+1} \frac{\nabla f_{k+1}^{\mathrm{T}} p_k}{\|\nabla f_{k+1}\|^2} = -1 + \frac{\nabla f_{k+1}^{\mathrm{T}} p_k}{\|\nabla f_k\|^2}$$

Plug $_k$ into

last eq

$$-1 + c_2 \frac{\nabla f_k^{\mathrm{T}} p_k}{\|\nabla f_k\|^2} \leq \frac{\nabla f_{k+1}^{\mathrm{T}} p_{k+1}}{\|\nabla f_{k+1}\|^2}$$

Substituting the lower bound for $\frac{\nabla f_k^{\mathrm{T}} p_k}{\|\nabla f_k\|}$

we obtain (7) for $k+1$

$$-1 - \frac{c_2}{1 - c_2} \leq \frac{\nabla f_{k+1}^{\mathrm{T}} p_{k+1}}{\|\nabla f_{k+1}\|^2} \leq -1 + \frac{c_2}{1 - c_2}.$$

Weakness of FR algorithm

If FR generates a bad direction and a tiny step, then the next direction and the next step are also likely to be poor.

Let $\theta_k = \angle(p_k, -\nabla f_k)$,

$$\cos\theta_k = \frac{-\nabla f_k^T p_k}{\|\nabla f_k\| \, \|p_k\|}.$$

A bad d $_k \approx 0$.

Mult

$$\frac{1-2c_2}{1-c_2}\frac{\|\nabla f_k\|}{\|p_k\|} \leq \cos\theta_k \leq \frac{1}{1-c_2}\frac{\|\nabla f_k\|}{\|p_k\|}$$

Thus $\cos\theta_k \approx 0$ if and only if $\|\nabla f_k\| \ll$

Since p_k is almost orthogonal to $-\nabla f_k$ $_1$ is likely tiny, i.e. $x_{k+1} \approx x_k$. Consequently, $\nabla f_k \approx \nabla f_{k+1}$ then $\beta_{k+1} \approx 1$ and finally given $\|\nabla f_{k+1}\| \approx \|\nabla f_k\| \ll \|p_k\|$, $p_{k+1} \approx p_k$ and the new direction will improve little.

If $\cos\theta_k \approx 0$ holds and the the subsequent step is small, the following updates are unproductive.

Polak-Ribière:

$$\beta_{k+1}^{PR} = \frac{\nabla f_{k+1}^{\mathrm{T}}(\nabla f_{k+1} - \nabla f_k)}{\|\nabla f_k\|^2} \tag{8}$$

If f i
exac

For general nonlinear functions and inexact line s
experience indicates that PR algorithm is more ro
efficient.

As is, the strong Wolfe conditions do not guarantee that p_k is always a descent direction. For $\beta_{k+1} = \max\{\beta_{k+1}^{PR}, 0\}$, simple adaptation of strong Wolfe conditions ensures the descent property.

Hestenes - Stiefel (similar to PR in both theory and practical performance):
Consecutive directions are conjugate wrt *average Hessian* $\bar{G}_k = \int_0^1 \nabla^2 f(x_k + \tau\alpha_k p_k) d\tau$. From Taylor's theorem we have $\nabla f_{k+1} = \nabla f_k + \alpha_k \bar{G}_k p_k$. Solving $p_{k+1}^{\mathrm{T}} \bar{G}_k p_k = 0$ where p_{k+1}

[illegible] (9)

Two competitive with PR choices which guaran descent direction under (standard) Wolfe cond

$$\beta_{k+1} = \frac{\|\nabla f_k}{(\nabla f_{k+1} - \nabla f_k)^{\mathrm{T}} p_k} \quad (10)$$

$$\beta_{k+1} = \left(y_k - 2p_k \frac{\|y_k\|^2}{y_k^{\mathrm{T}} p_k} \right)^{\mathrm{T}} \frac{\nabla f_{k+1}}{y_k^{\mathrm{T}} p_k} \quad \text{with } y_k = \nabla f_{k+1} - \nabla f_k. \quad (11)$$

Restarts

Set $\beta_k = 0$ in every nth step i.e. take steepest descent step. Restarting serves to refresh the algorithm erasing old information that may be not beneficial. Such restarting leads to n step quadratic convergence $\|x_{k+n} - x^\star\| = \mathcal{O}(\|x_k - x^\star\|^2)$.

Consider function which is strongly convex quadratic close to the solut
solut
conj n
steps from the restart (recall that the finite termin
for linear CG only holds if initiated with

In practice, conjugate gradient methods are usu n is large, hence n steps are never taken. Observe that the gradients are mutually orthogonal when f is a quadratic function. Restart when two consecutive gradients are far from orthogonal $\frac{|\nabla f_k^{\mathrm{T}} \nabla f_{k+1}|}{\|\nabla f_k\|^2} \geq \nu$, with ν typically 0.1.

When for some search direction p_k, $\cos\theta_k \approx 0$ and the subsequent step is small, substituting $\nabla f_{k+1} \approx \nabla f_k$ into β^{PR}_{k+1} results in $\beta^{PR}_{k+1} \approx 0$ and the next direction $p_{k+1} \approx -\nabla f_{k+1}$ the steepest descent direction. Therefore the PR algorithm essentially performs a restart after it encounters a bad direction.

The same argument applies to HS, and PR+.

FR alg

Hybrid FR-PR:

Global convergence can be guaranteed if 2.

This suggest following strategy

$$\beta_k = \begin{cases} -\beta_k^{FR}, & \beta_k^{PR} < -\beta_k^{FR} \\ \beta_k^{PR}, & |\beta_k^{PR}| \leq \beta_k^{FR} \\ \beta_k^{FR}, & \beta_k^{PR} > \beta_k^{FR} \end{cases} \tag{12}$$

Assumptions:

i) The level set $= \ x : f(x) \quad f(x_0)$ be bounded.

ii) In s n f

These assumptions imply that there is a constant

$$\|\nabla f(x)\| \leq \gamma, \forall$$

From Zoutenjik's lemma it follows that any line search iteration $x_{k+1} = x_k + \alpha_k p_k$ where p_k is a descent direction and the step length α_k satisfies Wolfe conditions gives the limit

Similarly, to the global convergence for line search, global convergence for **restarted** conjugate gr periodically setting $\beta_k = 0$ (hence cos a subsequence

$$\liminf_{k \to \infty} \|\nabla f_k\| = 0.$$

Theorem: [Al-Baali] Suppose that the assumptions i) and ii) hold and FR algorithm is implemented with line search that satisfies strong Wolfe conditions with $0 < c_1 < c_2 < \frac{1}{2}$. Then

Pro

e into

Zout

k

Lemma [1] recursively to show that the assumed to

sequence in lower bounded by harmonic series wh

hence contradiction.

This global convergence result can be extended to any method satisfying $|\beta_k| \leq \beta_k^{FR}$ for all $k \geq 2$.

If constants $c_4, c_5 > 0$ exist such that

$$\cos\theta_k \geq c_4 \frac{\|\nabla f_k\|}{\|p_k\|}, \quad \frac{\|\nabla f_k\|}{\|p_k\|} \geq c_5 > 0, \quad k = 1, 2, \ldots$$

if follows from Zoutenjik's result that

$$\lim \qquad f_k \quad = 0.$$

This r d

exact line search.

For general nonconvex functions it is not possible e
performs better in practice than FR. PR method ca
infinitely even if ideal line search is used i.e. line search which returns α_k that is the first positive stationary point of $f(x_k + \alpha p_k)$.
Example relies on $\beta_k < 0$ which motivated the modification $\beta_k^+ = \max\{0, \beta_k\}$.