Numerical Optimization

Graduate Course

Unconstrained Smooth Optimization

Part III: Conjugate gradient and inexact Newton methods

Wen Huang

School of Mathematical Sciences
Xiamen University

Compiled on February 14, 2022

Conjugate Gradient Methods

Equivalent to solving a linear system

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} x^T A x - b^T x \iff \text{ find } x \text{ such that } Ax = b$$

where $A \in \mathbb{R}^{n \times n}$ is a symmetric positive definite matrix.

- If A is diagonal, directions?
- If A is not diagonal, directions?

- Conjugate directions $\{p_0, p_1, \dots, p_{n-1}\}$, $p_i^T A p_j = 0$ for all $i \neq j$
- Conjugate direction method:
  1. Given initial x_0; conjugate direction $\{p_i\}_{i=0}^n$; and set $k = 0$
  2. Repeat n steps: $x_{k+1} = x_k + \alpha_k p_k$, where $\alpha_k = -\frac{r_k^T p_k}{p_k^T A p_k}$, $r_k = Ax_k - b$

Theorem 1

For any $x_0 \in \mathbb{R}^n$, the sequence $\{x_k\}$ generated by the conjugate direction algorithm converges to the solution x^* in at most n steps.

Conjugate gradient method is to choose conjugate directions by

- $r_0 = Ax_0 - b$, $p_0 = -r_0$
- $p_k = -r_k + \beta_k p_{k-1}$ such that $p_k^T A p_{k-1} = 0$

Theorem 2

Suppose the k-th iterate generated by the conjugate gradient method is not the solution x^*. Then

$$\begin{aligned} \mathrm{span}(r_0, r_1, \ldots, r_k) &= \mathrm{span}(r_0, Ar_0, \ldots, A^k r_0), \\ \mathrm{span}(p_0, p_1, \ldots, p_k) &= \mathrm{span}(r_0, Ar_0, \ldots, A^k r_0), \\ r_k^T p_i &= 0, \ \textit{for all } i < k, \\ p_k^T A p_i &= 0, \ \textit{for all } i < k, \end{aligned}$$

and x_k is the minimizer of $\frac{1}{2}x^T Ax - b^T x$ over $x_0 + \mathrm{span}(p_0, \ldots, p_{k-1})$.

Therefore, the conjugate gradient method finds x^* in at most n steps.

Linear conjugate gradient method

Input: Initial x_0;
Output: x_k;

1, Set $r_0 \leftarrow Ax_0 - b$, $p_0 \leftarrow -r_0$, $k \leftarrow 0$;
while $r_k \neq 0$ **do**
  2, $\alpha_k \leftarrow \frac{-r_k^T p_k}{p_k^T A p_k}$;
  3, $x_{k+1} \leftarrow x_k + \alpha_k p_k$;
  4, $r_{k+1} \leftarrow Ax_{k+1} - b$;
  5, $\beta_{k+1} \leftarrow \frac{r_{k+1}^T A p_k}{p_k^T A p_k}$;
  6, $p_{k+1} \leftarrow -r_{k+1} + \beta_{k+1} p_k$;
  7, $k \leftarrow k + 1$;
end while

Linear conjugate gradient method (Practical form)

Input: Initial x_0;
Output: x_k;
1, Set $r_0 \leftarrow Ax_0 - b$, $p_0 \leftarrow -r_0$, $k \leftarrow 0$;
while $r_k \neq 0$ **do**
2, $\alpha_k \leftarrow \frac{-r_k^T p_k}{p_k^T A p_k}$; $\Longleftrightarrow \alpha_k \leftarrow \frac{r_k^T r_k}{p_k^T A p_k}$; (by (6))
3, $x_{k+1} \leftarrow x_k + \alpha_k p_k$;
4, $r_{k+1} \leftarrow Ax_{k+1} - b$; $\Longleftrightarrow r_{k+1} \leftarrow r_k + \alpha_k A p_k$;
5, $\beta_{k+1} \leftarrow \frac{r_{k+1}^T A p_k}{p_k^T A p_k}$; $\Longleftrightarrow \beta_{k+1} \leftarrow \frac{r_{k+1}^T r_{k+1}}{r_k^T r_k}$; (by (4) and (6))
6, $p_{k+1} \leftarrow -r_{k+1} + \beta_{k+1} p_k$;
7, $k \leftarrow k + 1$;
end while

Computations of the practical form

- Ap_k, $p_k^T(Ap_k)$, and $r_k^T r_k$
- Main cost on Ap_k

Linear conjugate gradient method (Attempt for nonlinear problems)

Input: Initial x_0;
Output: x_k;

1, Set $r_0 \leftarrow Ax_0 - b$ ($r_0 = \nabla f(x_0)$), $p_0 \leftarrow -r_0$, $k \leftarrow 0$;
while $r_k \neq 0$ **do**
2, $\alpha_k \leftarrow \frac{r_k^T r_k}{p_k^T A p_k}$; $\rightsquigarrow$ exact step size;
3, $x_{k+1} \leftarrow x_k + \alpha_k p_k$;
4, $r_{k+1} = r_k + \alpha_k A p_k$; $\rightsquigarrow$ $r_{k+1} \leftarrow \nabla f(x_{k+1})$;
5, $\beta_{k+1} = \frac{r_{k+1}^T r_{k+1}}{r_k^T r_k}$; $\rightsquigarrow$ Fletcher-Reeves $\beta_{k+1} \leftarrow \frac{\nabla f(x_{k+1})^T \nabla f(x_{k+1})}{\nabla f(x_k)^T \nabla f(x_k)}$;
6, $p_{k+1} \leftarrow -r_{k+1} + \beta_{k+1} p_k$;
7, $k \leftarrow k+1$;
end while

- exact step size is not practical
- inexact step size?

Search direction:

$$p_{k+1} \leftarrow -\nabla f(x_{k+1}) + \beta_{k+1}^{\mathrm{FR}} p_k \text{ with } \beta_{k+1}^{\mathrm{FR}} = \frac{\nabla f(x_{k+1})^T \nabla f(x_{k+1})}{\nabla f(x_k)^T \nabla f(x_k)}$$

- Relax the condition of exact step size
- The strong Wolfe $(0 < c_1 < c_2 < 0.5) \implies p_{k+1}^T \nabla f(x_{k+1}) < 0$

Theorem 3

Let $\{x_k\}$ be the sequence generate by the nonlinear conjugate gradient method with FR scheme and strong Wolfe conditions with $0 < c_1 < c_2 < 0.5$. The the search directions p_k satisfy

$$-\frac{1}{1-c_2} \le \frac{\nabla f(x_k)^T p_k}{\|\nabla f(x_k)\|^2} \le \frac{2c_2 - 1}{1 - c_2}, \forall k \ge 0.$$

Theorem 4

Suppose $\mathcal{N}_{x_0} = \{x : f(x) \leq f(x_0)\}$, $f \in C^1$ *and the gradient* ∇f *is Lipschitz continuous in* $\mathcal{N}_{x_0}$, *i.e.,* $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$ *for all* $x, y \in \mathcal{N}_0$. *Then the FR nonlinear conjugate gradient algorithm either terminates at a stationary point or converges in the sense that*

$$\liminf_{k\to\infty} \|\nabla f(x_k)\| = 0.$$

Search direction:

$$p_{k+1} \leftarrow -\nabla f(x_{k+1}) + \beta_{k+1}^{\mathrm{FR}} p_k \text{ with } \beta_{k+1}^{\mathrm{FR}} = \frac{\nabla f(x_{k+1})^T \nabla f(x_{k+1})}{\nabla f(x_k)^T \nabla f(x_k)}$$

- Global convergence: $\liminf_{k\to\infty} \|\nabla f(x_k)\| = 0$
  - Assumption: $\mathcal{N}_{x_0} = \{x : f(x) \leq f(x_0)\}$ is bounded
  - Assumption: ∇f is Lipschitz in $\mathcal{N}_{x_0}$
- Difficulty: $\cos(\theta_k) \approx 0 \Longrightarrow \cos(\theta_{k+1}) \approx 0$

Search direction in nonlinear conjugate gradient method:

$$p_{k+1} \leftarrow -\nabla f(x_{k+1}) + \beta_{k+1} p_k$$

Remedies for Fletcher-Reeves scheme:

- Polak-Ribiére [PR69]: $\beta_{k+1}^{\mathrm{PR}} = \frac{\nabla f(x_{k+1})^T(\nabla f(x_{k+1}) - \nabla f(x_k))}{\nabla f(x_k)^T \nabla f(x_k)}$
- Hestenes-Stiefel [HS52]: $\beta_{k+1}^{\mathrm{HS}} = \frac{\nabla f(x_{k+1})^T(\nabla f(x_{k+1}) - \nabla f(x_k))}{(\nabla f(x_{k+1}) - \nabla f(x_k))^T p_k}$

Search direction in nonlinear conjugate gradient method:

$$p_{k+1} \leftarrow -\nabla f(x_{k+1}) + \beta_{k+1} p_k$$

Remedies for Fletcher-Reeves scheme:

- Polak-Ribiére [PR69]: $\beta_{k+1}^{\mathrm{PR}} = \frac{\nabla f(x_{k+1})^T(\nabla f(x_{k+1}) - \nabla f(x_k))}{\nabla f(x_k)^T \nabla f(x_k)}$
- Hestenes-Stiefel [HS52]: $\beta_{k+1}^{\mathrm{HS}} = \frac{\nabla f(x_{k+1})^T(\nabla f(x_{k+1}) - \nabla f(x_k))}{(\nabla f(x_{k+1}) - \nabla f(x_k))^T p_k}$

Note that

Exact line search $\Longrightarrow$ (Polak-Ribiére $\Leftrightarrow$ Hestenes-Stiefel)

Search direction in nonlinear conjugate gradient method:

$$p_{k+1} \leftarrow -\nabla f(x_{k+1}) + \beta_{k+1} p_k$$

Remedies for Fletcher-Reeves scheme:

- Polak-Ribiére [PR69]: $\beta_{k+1}^{\mathrm{PR}} = \frac{\nabla f(x_{k+1})^T (\nabla f(x_{k+1}) - \nabla f(x_k))}{\nabla f(x_k)^T \nabla f(x_k)}$
- Hestenes-Stiefel [HS52]: $\beta_{k+1}^{\mathrm{HS}} = \frac{\nabla f(x_{k+1})^T (\nabla f(x_{k+1}) - \nabla f(x_k))}{(\nabla f(x_{k+1}) - \nabla f(x_k))^T p_k}$

Note that

$$\text{Exact line search} \Longrightarrow (\text{Polak-Ribiére} \Leftrightarrow \text{Hestenes-Stiefel})$$

CG with either PR or HS does not even converge globally! [Pow86]

Search direction in nonlinear conjugate gradient method:

$$p_{k+1} \leftarrow -\nabla f(x_{k+1}) + \beta_{k+1} p_k$$

Options for β_{k+1}:

- Many modifications of PR and HS have been proposed
- New schemes with global convergence, see e.g., [HZ06, DLHY15] for a review
- Dai-Yuan [DY99]: $\beta_{k+1}^{\mathrm{DY}} = \frac{\nabla f(x_{k+1})^T \nabla f(x_{k+1})}{(\nabla f(x_{k+1}) - \nabla f(x_k))^T p_k}$

- Assume to use exact step sizes for step size selection
- Even for quadratic convex problem, local convergence can be linear if initial direction is not the negative gradient [Pow76]
- Restarting every n steps $\Longrightarrow$ n-step quadratic convergence in PR and FR nonlinear conjugate gradient methods [Coh72]

$$\hat{x} = Cx: \quad \min_x \ \frac{1}{2}x^T A x - b^T x \Longrightarrow \min_{\hat{x}} \ \frac{1}{2}\hat{x}^T C^{-T} A C^{-1} \hat{x} - (C^{-T} b)^T \hat{x}$$

Linear conjugate gradient method for $\frac{1}{2}\hat{x}^T \hat{A} \hat{x} - \hat{b}^T \hat{x}$

Input: Initial $\hat{x}_0$; $\Longrightarrow x_0 = C^{-1}\hat{x}_0$

Output: $\hat{x}_k$; $\Longrightarrow x_k = C^{-1}\hat{x}_k$

1, Set $\hat{r}_0 \leftarrow \hat{A}\hat{x}_0 - \hat{b}$; $\Longrightarrow r_0 = Ax_0 - b = C^T \hat{r}_0$

2, $\hat{p}_0 \leftarrow -\hat{r}_0$; $\Longrightarrow p_0 = C^{-1}\hat{p}_0 = -C^{-1}\hat{r}_0 = -C^{-1}C^{-T}r_0 = -(C^T C)^{-1} r_0$

3, $k \leftarrow 0$;

while $\hat{r}_k \neq 0$ **do**

4, $\alpha_k \leftarrow \frac{\hat{r}_k^T \hat{r}_k}{\hat{p}_k^T \hat{A} \hat{p}_k}$; $\Longrightarrow \alpha_k = \frac{r_k^T (C^T C)^{-1} r_k}{p_k^T A p_k}$

5, $\hat{x}_{k+1} \leftarrow \hat{x}_k + \alpha_k \hat{p}_k$; $\Longrightarrow x_{k+1} = C^{-1}\hat{x}_{k+1} = x_k + \alpha_k p_k$

6, $\hat{r}_{k+1} \leftarrow \hat{r}_k + \alpha_k \hat{A}\hat{p}_k$; $\Longrightarrow r_{k+1} = C^T \hat{r}_{k+1} = r_k + \alpha_k A p_k$

7, $\beta_{k+1} \leftarrow \frac{\hat{r}_{k+1}^T \hat{r}_{k+1}}{\hat{r}_k^T \hat{r}_k}$; $\Longrightarrow \beta_{k+1} = \frac{r_{k+1}^T (C^T C)^{-1} r_{k+1}}{r_k^T (C^T C)^{-1} r_k}$

8, $\hat{p}_{k+1} \leftarrow -\hat{r}_{k+1} + \beta_{k+1}\hat{p}_k$; $\Longrightarrow p_{k+1} = C^{-1}\hat{p}_{k+1} = -(C^T C)^{-1} r_{k+1} + \beta_{k+1} p_k$

9, $k \leftarrow k + 1$;

end while

- $M = C^T C$
- Linear system $Mu = v$ need be solved inexpensively

Preconditioned linear conjugate gradient method

Input: Initial x_0;
Output: x_k;

1, Set $r_0 \leftarrow Ax_0 - b$;
2, Solve $My_0 = r_0$ for y_0;
3, $p_0 = -y_0$;
4, $k \leftarrow 0$;
while $r_k \neq 0$ **do**
  5, $\alpha_k \leftarrow \frac{r_k^T y_k}{p_k^T A p_k}$;
  6, $x_{k+1} \leftarrow x_k + \alpha_k p_k$;
  7, $r_{k+1} \leftarrow r_k + \alpha_k A p_k$;
  8, Solve $My_{k+1} = r_{k+1}$ for y_{k+1};
  9, $\beta_{k+1} \leftarrow \frac{r_{k+1}^T y_{k+1}}{r_k^T y_k}$;
  10, $p_{k+1} \leftarrow -y_{k+1} + \beta_{k+1} p_k$;
  11, $k \leftarrow k + 1$;
end while

Preconditioning

Linear conjugate gradient to nonlinear conjugate gradient

1, Set $r_0 \leftarrow Ax_0 - b$;
2, Solve $My_0 = r_0$ for y_0;
3, $p_0 = -y_0$, $k \leftarrow 0$;
while $r_k \neq 0$ **do**
    4, $\alpha_k \leftarrow \frac{r_k^T y_k}{p_k^T A p_k}$;
    5, $x_{k+1} \leftarrow x_k + \alpha_k p_k$;
    6, $r_{k+1} \leftarrow r_k + \alpha_k A p_k$;
    7, Solve $My_{k+1} = r_{k+1}$ for y_{k+1};
    8, $\beta_{k+1} \leftarrow \frac{r_{k+1}^T y_{k+1}}{r_k^T y_k}$;
    9, $p_{k+1} \leftarrow -y_{k+1} + \beta_{k+1} p_k$;
    10, $k \leftarrow k + 1$;
end while

- $y_k = M^{-1} r_k \implies y_k = M^{-1} \nabla f(x_k)$
- Nonlinear CG direction: $p_{k+1} = -M(x_{k+1})^{-1} \nabla f(x_{k+1}) + \beta_{k+1} p_k$
- M is an approximation of the Hessian and easy to invert.

One preconditioned CG:
(A preconditioner can be added to other nonlinear CG similarly)

The FR type nonlinear conjugate gradient method

Input: Initial x_0; Parameters $0 < c_1 < c_2 < 1$ for the weak Wolfe condition;
Output: x_k;

1, $y_0 = M(x_0)^{-1}\nabla f(x_0)$, initial search direction $p_0 = -y_0$;
while $r_k \neq 0$ **do**
  2, Find step size α_k satisfying the weak Wolfe conditions;
  3, $x_{k+1} \leftarrow x_k + \alpha_k p_k$;
  4, $y_{k+1} = M(x_{k+1})^{-1}\nabla f(x_{k+1})$;
  4, $\beta_{k+1}^{\mathrm{FR}} \leftarrow \frac{\nabla f(x_{k+1})^T y_{k+1}}{\nabla f(x_k)^T y_k}$;
  5, $p_{k+1} \leftarrow -y_{k+1} + \beta_{k+1}^{\mathrm{FR}} p_k$;
  6, $k \leftarrow k+1$;
end while

Inexact-Newton methods

- Newton's method for root finding $\nabla f(x) = 0$:

$$\nabla f(x+p) \approx \nabla f(x) + \nabla^2 f(x) p.$$

- Find p such that $\nabla f(x+p) \approx 0$:

$$p = -\left(\nabla^2 f(x)\right)^{-1} \nabla f(x)$$

and therefore

$$x_+ = x - \left(\nabla^2 f(x)\right)^{-1} \nabla f(x).$$

Newton's method

Input: Initial iterate x_0;
  Set $k \leftarrow 0$;
  for $k = 0, 1, \ldots$ **do**
    $x_{k+1} \leftarrow x_k - \left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k)$;
    $k \leftarrow k + 1$;
  end for

- Inspiring from solving system $\Longrightarrow$ only converge to a stationary point
- $-p^T \nabla f(x) = \nabla f(x)^T \nabla^2 f(x) \nabla f(x) \overset{?}{\geq} 0$
- Global convergence is not guaranteed

For example:

$$
\begin{aligned}
f(x) =& \frac{1}{4}x^4 - x^2 + 2x \\
f'(x) =& x^3 - 2x + 2 \\
f''(x) =& 3x^2 - 2
\end{aligned}
$$

Choose $x_0 = 0$ or 1.

Newton's Method

Local convergence analysis

- Inspiring from root finding problems $\Longrightarrow$ only converge to a stationary point
- $-p^T \nabla f(x) = \nabla f(x)^T \nabla^2 f(x) \nabla f(x) \overset{?}{\geq} 0$
- Global convergence is not guaranteed
- Fast local convergence

Theorem 5

Let x^* be a minimizer of f. Suppose $f \in C^2$, $\nabla f(x^*) = 0$, $\nabla^2 f(x)$ is positive definite, and the Hessian $\nabla^2 f(x)$ is Lipschitz continuous in a neighborhood Ω_{x^*} of a solution x^*, i.e., $\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L\|x - y\|$ for $x, y \in \Omega_{x^*}$. Then

1. *if x_0 is sufficiently close to x^*, then $\{x_k\}$ by Newton's method converges to x^*; and*
2. *the rate of convergence of $\{x_k\}$ is quadratic.*

Newton direction: $\nabla^2 f(x) p = -\nabla f(x)$

- Such direction p may not exist
- Even it exists, the direction p may not be descent
- Hessian modifications are needed
  - $\nabla^2 f(x)$ positive definite $\Longrightarrow$ accurate enough $p \approx -\nabla^2 f(x)^{-1} \nabla f(x)$
  - $\nabla^2 f(x)$ indefinite $\Longrightarrow$ a descent direction p

Inexact Newton direction: $(\nabla^2 f(x) + E_i)p = -\nabla f(x)$

Inexact Newton's method

Set $k \leftarrow 0$;
for $k = 0, 1, \ldots$ **do**
$p_k \leftarrow -(\nabla^2 f(x_k) + E_k)^{-1} \nabla f(x_k)$, where $E_k = 0$ if $\nabla^2 f(x_k)$ is positive definite; Otherwise, $\nabla^2 f(x_k) + E_k$ is positive definite;
$x_{k+1} \leftarrow x_k + \alpha_k p_k$ with α_k by the Byrd Nocedal condition;
$k \leftarrow k + 1$;
end for

Modifications:

- Eigenvalue modification
- Adding a multiple of the identity
- Modified Cholesky factorization
- Truncated conjugate gradient

$$\nabla^2 f(x) = Q\Lambda Q^T = \sum_{i=1}^{n} \lambda_i q_i q_i^T.$$

Modify the eigenvalues, e.g.,

- $\min_B \|B - \nabla^2 f(x)\|_F$, s.t. B is SPSD $\Rightarrow B = \sum_{i=i}^{n} \max(\lambda_i, 0) q_i q_i^T$
  But $Bp_k = -\nabla f(x_k)$ may not have solution
- $\min_H \|H - \nabla^2 f(x)^{-1}\|_F$, s.t. H is SPSD $\Rightarrow H = \sum_{i=i}^{n} \max(\frac{1}{\lambda_i}, 0) q_i q_i^T$
  and $p_k = -H\nabla f(x_k)$
- Or other norms
- Eigenvalue decomposition is too expensive
- Any computationally efficient modifications

Inexact Newton Methods

Adding a multiple of the identity

$$\nabla^2 f(x) = Q \Lambda Q^T = \sum_{i=1}^{n} \lambda_i q_i q_i^T.$$

- Choose $\tau > 0$ such that $\nabla^2 f(x) + \tau I$ is sufficient SPSD
- τ sufficiently larger than $-\lambda_{\min}$
- $\lambda_{\min}$?

If $\nabla^2 f(x)$ is SPD, then $\nabla^2 f(x) = LDL^T$ unique decomposition

$$\begin{bmatrix} a_{11} & a_{21} & a_{31} \\ a_{21} & a_{22} & a_{32} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix} \begin{bmatrix} d_1 & & \\ & d_2 & \\ & & d_3 \end{bmatrix} \begin{bmatrix} 1 & l_{21} & l_{31} \\ 0 & 1 & l_{32} \\ 0 & 0 & 1 \end{bmatrix}$$

Cholesky Factorization, LDL^T form

for $j = 1, 2, \ldots, n$ **do**
  $c_{jj} \leftarrow a_{jj} - \sum_{s=1}^{j-1} d_s l_{js}^2$;
  $d_j \leftarrow c_{jj}$;
  for $i = j+1, \ldots, n$ **do**
    $c_{ij} \leftarrow a_{ij} - \sum_{s=1}^{j-1} d_s l_{is} l_{js}$;
    $l_{ij} \leftarrow c_{ij}/d_j$;
  end for
end for

- If $\nabla^2 f(x)$ is not SPD, then modify d_i if necessary.

Cholesky Factorization, LDL^T form

for $j = 1, 2, \ldots, n$ **do**
  $c_{jj} \leftarrow a_{jj} - \sum_{s=1}^{j-1} d_s l_{js}^2$;
  $d_j \leftarrow c_{jj}$;
  for $i = j+1, \ldots, n$ **do**
    $c_{ij} \leftarrow a_{ij} - \sum_{s=1}^{j-1} d_s l_{is} l_{js}$;
    $l_{ij} \leftarrow c_{ij}/d_j$;
  end for
end for

Given $\delta > 0$ and $\beta > 0$,

- set $d_j = \max\left(|c_{jj}|, \left(\frac{\max_{j<i\leq n}(|c_{ij}|)}{\beta}\right)^2, \delta\right)$;
- $d_j \geq \delta$ and $|l_{ij}\sqrt{d_j}| \leq \beta$;
- Conditioner number and norm: bounded from above [GMH81]
- Global convergence and local quadratic convergence rate (appropriate δ and β)

- Permutation can be used to $\nabla^2 f$: i.e., $P\nabla^2 f(x)P^T + E$
- Preserve sparsity if $\nabla^2 f$ is sparse
- Computations $O(n^3/3)$
- Storage $O(n^2)$

Newton-CG Method

Truncated conjugate gradient

Use CG to solve the linear system $\nabla^2 f(x)p = -\nabla f(x)$

- $\nabla^2 f$ is SPD $\Leftrightarrow$ CG finds accurate solution
- $\nabla^2 f$ is not SPD $\Leftrightarrow$ CG stops early and guarantee p descent direction
- n-steps CG: computations $2n^3$ on dense $\nabla^2 f \gg$ that of Cholesky decomposition
- Matrix-free method: only need matrix-vector product
- Much smaller than n-steps in early iterations

Truncated conjugate gradient (tCG) for $Bp = -g$
Initializations:
Set $p_0 = 0$, $r_0 = g$, $d_0 = -r_0$,
Then repeat the following loop on j:
Check for negative curvature

if $d_j^T B d_j \leq 0$
  if $j = 0$
    return $p^* \leftarrow -g$;
  else
    return $p^* \leftarrow p_j$;

(Continue on the next page)

Newton-CG Method

Truncated conjugate gradient for the Newton subproblem

Generate next inner iterate

Set $\alpha_j \leftarrow r_j^T r_j / d_j^T B d_j$;

Set $p_{j+1} \leftarrow p_j + \alpha_j d_j$;

Update residual and search direction

Set $r_{j+1} \leftarrow r_j + \alpha_j B d_j$;

Set $\beta_{j+1} \leftarrow r_{j+1}^T r_{j+1} / r_j^T r_j$;

Set $d_{j+1} \leftarrow -r_{j+1} + \beta_{j+1} d_j$;

$j \leftarrow j + 1$;

Check residual

if $\|r_j\| \leq \|r_0\| \min\left(\|r_0\|^\theta, \kappa\right)$ for some prescribed θ and κ

return $p^* \leftarrow p_j$;

A Newotn-CG algorithm

Input: Initial iterate x_0;

Set $k \leftarrow 0$;

while not accurate enough **do**

Compute the search direction p_k by the truncated CG algorithm;

$x_{k+1} \leftarrow x_k + \alpha_k p_k$ with α_k by the Byrd Nocedal condition; Note that 1 is used if it is acceptable;

$k \leftarrow k + 1$;

end while

Newton-CG Method

Local convergence rate

Theorem 6

Let $\{x_k\}$ denote the sequence generated by the Newton-CG method with $\kappa \in (0,1)$ and $\theta > 0$. Suppose that $\nabla^2 f(x)$ exists and is continuous in a neighborhood of a minimizer x^*, with $\nabla^2 f(x^*)$ is positive definite, and that $\{x_k\}$ converges to x^*. Then the convergence rate is superlinear. In addition, if $\nabla^2 f(x)$ is Lipschitz continuous for x near x^*, then the convergence rate is $\min(1+\theta, 2)$.

References I

Arthur I. Cohen.
Rate of convergence of several conjugate gradient algorithms.
SIAM Journal on Numerical Analysis, 9(2):248–259, 1972.

Xiao Liang Dong, Hong Wei Liu, Yu Bo He, and Xi Mei Yang.
A modified Hestenes-Stiefel conjugate gradient method with sufficient descent condition and conjugacy condition.
Journal of Computational and Applied Mathematics, 281:239 – 249, 2015.

Y. H. Dai and Y. Yuan.
A nonlinear conjugate gradient method with a strong global convergence property.
SIAM Journal on Optimization, 10(1):177–182, 1999.

P. E. Gill, W. Murray, and Wright M. H.
Practical Optimization.
Academic Press, 1981.

M. R. Hestenes and E. Stiefel.
Methods of conjugate gradients for solving linear systems.
J. Res. Nat. Bur. Stand., 49:409–436, 1952.

William W. Hager and Hongchao Zhang.
A survey of nonlinear conjugate gradient methods.
Pac. J. Optim., 2:35–58, 2006.

M. J. D. Powell.
Some convergence properties of the conjugate gradient method.
Mathematical Programming, (11):42–49, 1976.

M J. D. Powell.
Convergence properties of algorithms for nonlinear optimization.
SIAM Review, 28(4):487–500, 1986.

References II

E. Polak and G. Ribi'ere.
Note sur la convergence de méthodes de directions conjuguées.
Revue Francaise d'Informatique et de Recherche Operationnelle, 16(16):35–43, 1969.