

## Part 2: Linesearch methods for unconstrained optimization

Nick Gould (RAL)

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad f(x)$$

MSc course on nonlinear optimization

### ITERATIVE METHODS

- in practice very rare to be able to provide explicit minimizer
- iterative method: given starting “guess”  $x_0$ , generate sequence  $\{x_k\}$ ,  $k = 1, 2, \dots$
- **AIM**: ensure that (a subsequence) has some favourable limiting properties:
  - satisfies first-order necessary conditions
  - satisfies second-order necessary conditions

Notation:  $f_k = f(x_k)$ ,  $g_k = g'(x_k)$ ,  $H_k = H(x_k)$ .

### UNCONSTRAINED MINIMIZATION

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad f(x)$$

where the **objective function**  $f : \mathbb{R}^n \rightarrow \mathbb{R}$

- assume that  $f \in C^1$  (sometimes  $C^2$ ) and Lipschitz
- often in practice this assumption violated, but not necessary

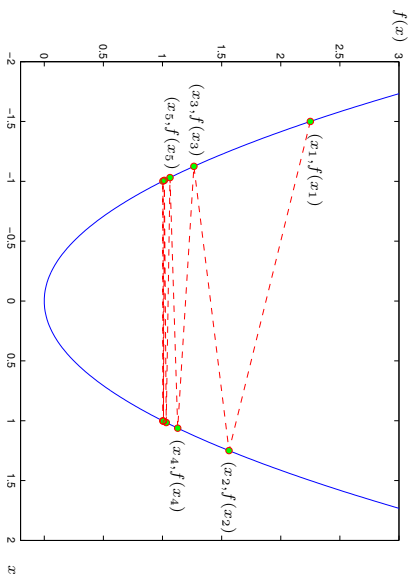
### LINESEARCH METHODS

- calculate a **search direction**  $p_k$  from  $x_k$
- ensure that this direction is a **descent direction**, i.e.,  $g_k^T p_k < 0$  if  $g_k \neq 0$   
so that, for small steps along  $p_k$ , the objective function **will** be reduced
- calculate a suitable **steplength**  $\alpha_k > 0$  so that  $f(x_k + \alpha_k p_k) < f_k$

- computation of  $\alpha_k$  is the **linesearch**—may itself be an iteration
- generic linesearch method:

$$x_{k+1} = x_k + \alpha_k p_k$$

## STEPS MIGHT BE TOO LONG



The objective function  $f(x) = x^2$  and the iterates  $x_{k+1} = x_k + \alpha_k p_k$  generated by the descent directions  $p_k = (-1)^{k+1}$  and steps  $\alpha_k = 2 + 3/2^{k+1}$  from  $x_0 = 2$

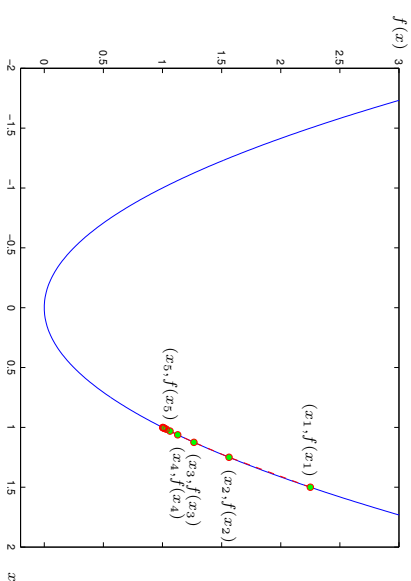
## PRACTICAL LINESEARCH METHODS

- in early days, pick  $\alpha_k$  to minimize

$$f(x_k + \alpha p_k)$$

- rather expensive and certainly not cost effective
- modern methods: **inexact** linesearch
  - ensure steps are neither too long nor too short
  - try to pick “useful” initial stepsize for fast convergence
  - best methods are either
    - “backtracking- Armijo” or
    - “Armijo-Goldstein” based

## STEPS MIGHT BE TOO SHORT



The objective function  $f(x) = x^2$  and the iterates  $x_{k+1} = x_k + \alpha_k p_k$  generated by the descent directions  $p_k = -1$  and steps  $\alpha_k = 1/2^{k+1}$  from  $x_0 = 2$

## BACKTRACKING LINESEARCH

Procedure to find the stepsize  $\alpha_k$ :

Given  $\alpha_{\text{init}} > 0$  (e.g.,  $\alpha_{\text{init}} = 1$ )  
 let  $\alpha^{(0)} = \alpha_{\text{init}}$  and  $l = 0$   
 Until  $f(x_k + \alpha^{(l)} p_k) \leq f_k$   
 set  $\alpha^{(l+1)} = \tau \alpha^{(l)}$ , where  $\tau \in (0, 1)$  (e.g.,  $\tau = \frac{1}{2}$ )  
 and increase  $l$  by 1  
 Set  $\alpha_k = \alpha^{(l)}$

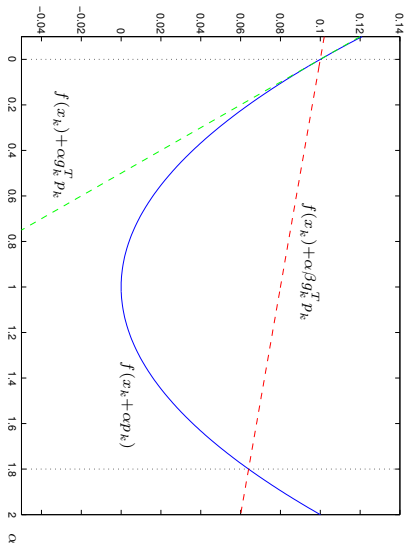
- this prevents the step from getting too small ... but does not prevent too large steps relative to decrease in  $f$
- need to tighten requirement
 
$$f(x_k + \alpha^{(l)} p_k) \leq f_k$$

## ARMJJO CONDITION

In order to prevent large steps relative to decrease in  $f$ , instead require

$$f(x_k + \alpha_k p_k) \leq f(x_k) + \alpha_k \beta g_k^T p_k$$

for some  $\beta \in (0, 1)$  (e.g.,  $\beta = 0.1$  or even  $\beta = 0.0001$ )



## SATISFYING THE ARMJJO CONDITION

**Theorem 2.1.** Suppose that  $f \in C^1$ , that  $g(x)$  is Lipschitz continuous with Lipschitz constant  $\gamma(x)$ , that  $\beta \in (0, 1)$  and that  $p$  is a descent direction at  $x$ . Then the Armijo condition

$$f(x + \alpha p) \leq f(x) + \alpha \beta g(x)^T p$$

is satisfied for all  $\alpha \in [0, \alpha_{\max}(x)]$ , where

$$\alpha_{\max} = \frac{2(\beta - 1)g(x)^T p}{\gamma(x)\|p\|_2^2}$$

## BACKTRACKING-ARMJJO LINESEARCH

Procedure to find the stepsize  $\alpha_k$ :

Given  $\alpha_{\text{init}} > 0$  (e.g.,  $\alpha_{\text{init}} = 1$ )  
 let  $\alpha^{(0)} = \alpha_{\text{init}}$  and  $l = 0$   
 Until  $f(x_k + \alpha^{(l)} p_k) \leq f(x_k) + \alpha^{(l)} \beta g_k^T p_k$   
 set  $\alpha^{(l+1)} = \tau \alpha^{(l)}$ , where  $\tau \in (0, 1)$  (e.g.,  $\tau = \frac{1}{2}$ )  
 and increase  $l$  by 1  
 Set  $\alpha_k = \alpha^{(l)}$

## PROOF OF THEOREM 2.1

Taylor's theorem (Theorem 1.1) +

$$\begin{aligned} \alpha &\leq \frac{2(\beta - 1)g(x)^T p}{\gamma(x)\|p\|_2^2}, \\ \implies f(x + \alpha p) &\leq f(x) + \alpha g(x)^T p + \frac{1}{2} \gamma(x) \alpha^2 \|p\|^2 \\ &\leq f(x) + \alpha g(x)^T p + \alpha(\beta - 1)g(x)^T p \\ &= f(x) + \alpha \beta g(x)^T p \end{aligned}$$

## THE ARMIZO LINESEARCH TERMINATES

**Corollary 2.2.** Suppose that  $f \in C^1$ , that  $g(x)$  is Lipschitz continuous with Lipschitz constant  $\gamma_k$  at  $x_k$ , that  $\beta \in (0, 1)$  and that  $p_k$  is a descent direction at  $x_k$ . Then the stepsize generated by the backtracking-Armijo linesearch terminates with

$$\alpha_k \geq \min \left( \alpha_{\text{init}}, \frac{2\tau(\beta - 1)g_k^T p_k}{\gamma_k \|p_k\|_2^2} \right)$$

## GENERIC LINESEARCH METHOD

Given an initial guess  $x_0$ , let  $k = 0$   
 Until convergence:  
   Find a descent direction  $p_k$  at  $x_k$   
   Compute a stepsize  $\alpha_k$  using a  
   backtracking-Armijo linesearch along  $p_k$   
   Set  $x_{k+1} = x_k + \alpha_k p_k$ , and increase  $k$  by 1

## PROOF OF COROLLARY 2.2

Theorem 2.1  $\implies$  linesearch will terminate as soon as  $\alpha^{(l)} \leq \alpha_{\text{max}}$ .  
 2 cases to consider:

1. May be that  $\alpha_{\text{init}}$  satisfies the Armijo condition  $\implies \alpha_k = \alpha_{\text{init}}$ .
2. Otherwise, must be a last linesearch iteration (the  $l$ -th) for which

$$\alpha^{(l)} > \alpha_{\text{max}} \implies \alpha_k \geq \alpha^{(l+1)} = \tau \alpha^{(l)} > \tau \alpha_{\text{max}}$$

Combining these 2 cases gives required result.

## GLOBAL CONVERGENCE THEOREM

**Theorem 2.3.** Suppose that  $f \in C^1$  and that  $g$  is Lipschitz continuous on  $\mathbb{R}^n$ . Then, for the iterates generated by the Generic Linesearch Method,

either

$$g_l = 0 \text{ for some } l \geq 0$$

or

$$\lim_{k \rightarrow \infty} f_k = -\infty$$

or

$$\lim_{k \rightarrow \infty} \min (|p_k^T g_k|, |p_k^T g_k| / \|p_k\|_2) = 0.$$

### PROOF OF THEOREM 2.3

Suppose that  $g_k \neq 0$  for all  $k$  and that  $\lim_{k \rightarrow \infty} f_k > -\infty$ . Armijo  $\implies$

$$f_{k+1} - f_k \leq \alpha_k \beta p_k^T g_k$$

for all  $k \implies$  summing over first  $j$  iterations

$$f_{j+1} - f_0 \leq \sum_{k=0}^j \alpha_k \beta p_k^T g_k.$$

LHS bounded below by assumption  $\implies$  RHS bounded below. Sum composed of -ve terms  $\implies$

$$\lim_{k \rightarrow \infty} \alpha_k |p_k^T g_k| = 0$$

Let

$$\mathcal{K}_1 \stackrel{\text{def}}{=} \left\{ k \mid \alpha_{\text{init}} > \frac{2\tau(\beta-1)g_k^T p_k}{\gamma \|p_k\|^2} \right\} \& \mathcal{K}_2 \stackrel{\text{def}}{=} \{1, 2, \dots\} \setminus \mathcal{K}_1$$

where  $\gamma$  is the assumed uniform Lipschitz constant.

### METHOD OF STEEPEST DESCENT

The search direction

$$p_k = -g_k$$

gives the so-called **steepest-descent** direction.

○  $p_k$  is a descent direction

○  $p_k$  solves the problem

$$\underset{p \in \mathbb{R}^n}{\text{minimize}} \quad m_k^L(x_k + p) \stackrel{\text{def}}{=} f_k + g_k^T p \quad \text{subject to} \quad \|p\|_2 = \|g_k\|_2$$

Any method that uses the steepest-descent direction is a

**method of steepest descent**.

For  $k \in \mathcal{K}_1$ ,

$$\alpha_k \geq \frac{2\tau(\beta-1)g_k^T p_k}{\gamma \|p_k\|^2}$$

$\implies$

$$\alpha_k p_k^T g_k \leq \frac{2\tau(\beta-1)}{\gamma} \left( \frac{g_k^T p_k}{\|p_k\|} \right)^2 < 0$$

$\implies$

$$\lim_{k \in \mathcal{K}_1 \rightarrow \infty} \frac{|p_k^T g_k|}{\|p_k\|} = 0. \tag{1}$$

For  $k \in \mathcal{K}_2$ ,

$$\alpha_k \geq \alpha_{\text{init}}$$

$\implies$

$$\lim_{k \in \mathcal{K}_2 \rightarrow \infty} |p_k^T g_k| = 0. \tag{2}$$

Combining (1) and (2) gives the required result.

### GLOBAL CONVERGENCE FOR STEEPEST DESCENT

**Theorem 2.4.** Suppose that  $f \in C^1$  and that  $g$  is Lipschitz continuous on  $\mathbb{R}^n$ . Then, for the iterates generated by the Generic Linesearch Method using the steepest-descent direction,

either

$$g_l = 0 \text{ for some } l \geq 0$$

or

$$\lim_{k \rightarrow \infty} f_k = -\infty$$

or

$$\lim_{k \rightarrow \infty} g_k = 0.$$

## PROOF OF THEOREM 2.4

Follows immediately from Theorem 2.3, since

$$\min(|p_k^T g_k|, |p_k^T g_k|/\|p_k\|_2) = \|g_k\|_2 \min(1, \|g_k\|_2)$$

and thus

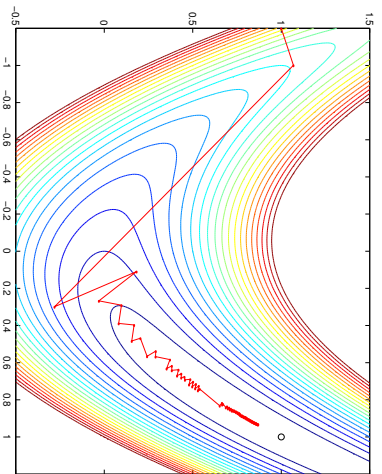
$$\lim_{k \rightarrow \infty} \min(|p_k^T g_k|, |p_k^T g_k|/\|p_k\|_2) = 0$$

implies that  $\lim_{k \rightarrow \infty} g_k = 0$ .

## METHOD OF STEEPEST DESCENT (cont.)

- archetypical globally convergent method
- many other methods resort to steepest descent in bad cases
- not scale invariant
- convergence is usually very (very!) slow (linear)
- numerically often not convergent at all

## STEEPEST DESCENT EXAMPLE



Contours for the objective function  $f(x, y) = 10(y - x^2)^2 + (x - 1)^2$ , and the iterates generated by the Generic Line-search steepest-descent method

## MORE GENERAL DESCENT METHODS

Let  $B_k$  be a symmetric, positive definite matrix, and define the search direction  $p_k$  so that

$$B_k p_k = -g_k$$

Then

- $p_k$  is a descent direction
- $p_k$  solves the problem

$$\underset{p \in \mathbb{R}^n}{\text{minimize}} \quad m_k^Q(x_k + p) \stackrel{\text{def}}{=} f_k + g_k^T p + \frac{1}{2} p^T B_k p$$

- if the Hessian  $H_k$  is positive definite, and  $B_k = H_k$ , this is **Newton's method**

## MORE GENERAL GLOBAL CONVERGENCE

**Theorem 2.5.** Suppose that  $f \in C^1$  and that  $g$  is Lipschitz continuous on  $\mathbb{R}^n$ . Then, for the iterates generated by the Generic Linesearch Method using the more general descent direction, either

$$g_l = 0 \text{ for some } l \geq 0$$

or

$$\lim_{k \rightarrow \infty} f_k = -\infty$$

or

$$\lim_{k \rightarrow \infty} g_k = 0$$

provided that the eigenvalues of  $B_k$  are uniformly bounded and bounded away from zero.

## PROOF OF THEOREM 2.5

Let  $\lambda_{\min}(B_k)$  and  $\lambda_{\max}(B_k)$  be the smallest and largest eigenvalues of  $B_k$ . By assumption, there are bounds  $\lambda_{\min} > 0$  and  $\lambda_{\max}$  such that

$$\lambda_{\min} \leq \lambda_{\min}(B_k) \leq \frac{s^T B_k s}{\|s\|^2} \leq \lambda_{\max}(B_k) \leq \lambda_{\max}$$

and thus that

$$\lambda_{\max}^{-1} \leq \lambda_{\max}^{-1}(B_k) = \lambda_{\min}(B_k^{-1}) \leq \frac{s^T B_k^{-1} s}{\|s\|^2} \leq \lambda_{\max}(B_k^{-1}) = \lambda_{\min}^{-1}(B_k) \leq \lambda_{\min}^{-1}$$

for any nonzero vector  $s$ . Thus

$$|p_k^T g_k| = |g_k^T B_k^{-1} g_k| \geq \lambda_{\min}(B_k^{-1}) \|g_k\|_2^2 \geq \lambda_{\min}^{-1} \|g_k\|_2^2$$

In addition

$$\|p_k\|_2^2 = g_k^T B_k^{-2} g_k \leq \lambda_{\max}(B_k^{-2}) \|g_k\|_2^2 \leq \lambda_{\min}^{-2} \|g_k\|_2^2,$$

$\implies$

$$\|p_k\|_2 \leq \lambda_{\min}^{-1} \|g_k\|_2$$

## MORE GENERAL DESCENT METHODS (cont.)

- may be viewed as “scaled” steepest descent
- convergence is often faster than steepest descent
- can be made scale invariant for suitable  $B_k$

Thus

$$\frac{|p_k^T g_k|}{\|p_k\|_2} \geq \frac{\lambda_{\min}}{\lambda_{\max}} \|g_k\|_2$$

$\implies$

$$\min (|p_k^T g_k|, |p_k^T g_k| / \|p_k\|_2) \geq \frac{\|g_k\|_2}{\lambda_{\max}} \min (\lambda_{\min}, \|g_k\|_2)$$

$\implies$

$$\lim_{k \rightarrow \infty} \min (|p_k^T g_k|, |p_k^T g_k| / \|p_k\|_2) = 0$$

$\implies$

$$\lim_{k \rightarrow \infty} g_k = 0.$$

## CONVERGENCE OF NEWTON'S METHOD

**Theorem 2.6.** Suppose that  $f \in C^2$  and that  $H$  is Lipschitz continuous on  $\mathbb{R}^n$ . Then suppose that the iterates generated by the Generic Linesearch Method with  $\alpha_{\text{mit}} = 1$  and  $\beta < \frac{1}{2}$ , in which the search direction is chosen to be the Newton direction  $p_k = -H_k^{-1}g_k$  whenever possible, has a limit point  $x_*$  for which  $H(x_*)$  is positive definite. Then

- (i)  $\alpha_k = 1$  for all sufficiently large  $k$ ,
- (ii) the entire sequence  $\{x_k\}$  converges to  $x_*$ , and
- (iii) the rate is  $\mathcal{Q}$ -quadratic, i.e, there is a constant  $\kappa \geq 0$ .

$$\lim_{k \rightarrow \infty} \frac{\|x_{k+1} - x_*\|_2}{\|x_k - x_*\|_2^2} \leq \kappa.$$

## PROOF OF THEOREM 2.6

Consider  $\lim_{k \in \mathcal{K}} x_k = x_*$ . Continuity  $\implies H_k$  positive definite for all  $k \in \mathcal{K}$  sufficiently large  $\implies \exists k_0 \geq 0$ :

$$p_k^T H_k p_k \geq \frac{1}{2} \lambda_{\min}(H_*) \|p_k\|_2^2$$

$\forall k_0 \leq k \in \mathcal{K}$ , where  $\lambda_{\min}(H_*)$  = smallest eigenvalue of  $H(x_*) \implies$

$$|p_k^T g_k| = -p_k^T g_k = p_k^T H_k p_k \geq \frac{1}{2} \lambda_{\min}(H_*) \|p_k\|_2^2. \quad (3)$$

$\forall k_0 \leq k \in \mathcal{K}$ , and

$$\lim_{k \in \mathcal{K} \rightarrow \infty} p_k = 0$$

since Theorem 2.5  $\implies$  at least one of the LHS of (3) and

$$\frac{|p_k^T g_k|}{\|p_k\|_2} = -\frac{p_k^T g_k}{\|p_k\|_2} \geq \frac{1}{2} \lambda_{\min}(H_*) \|p_k\|_2$$

converges to zero for such  $k$ .

Taylor's theorem  $\implies \exists z_k$  between  $x_k$  and  $x_k + p_k$  such that

$$f(x_k + p_k) = f_k + p_k^T g_k + \frac{1}{2} p_k^T H(z_k) p_k.$$

Lipschitz continuity of  $H$  &  $H_k p_k + g_k = 0 \implies$

$$\begin{aligned} f(x_k + p_k) - f_k - \frac{1}{2} p_k^T g_k &= \frac{1}{2} (p_k^T g_k + p_k^T H(z_k) p_k) \\ &= \frac{1}{2} (p_k^T g_k + p_k^T H_k p_k) + \frac{1}{2} (p_k^T (H(z_k) - H_k) p_k) \\ &\leq \frac{1}{2} \gamma \|z_k - x_k\|_2 \|p_k\|_2^2 \leq \frac{1}{2} \gamma \|p_k\|_2^3 \end{aligned} \quad (4)$$

Now pick  $k$  sufficiently large so that

$$\gamma \|p_k\|_2 \leq \lambda_{\min}(H_*) (1 - 2\beta).$$

+ (3) + (4)  $\implies$

$$\begin{aligned} f(x_k + p_k) - f_k &\leq \frac{1}{2} p_k^T g_k + \frac{1}{2} \lambda_{\min}(H_*) (1 - 2\beta) \|p_k\|_2^2 \\ &\leq \frac{1}{2} (1 - (1 - 2\beta)) p_k^T g_k = \beta p_k^T g_k \end{aligned}$$

$\implies$  unit stepsize satisfies the Armijo condition for all sufficiently large  $k \in \mathcal{K}$

Now note that  $\|H_k^{-1}\|_2 \leq 2/\lambda_{\min}(H_*)$  for all sufficiently large  $k \in \mathcal{K}$ . The iteration gives

$$\begin{aligned} x_{k+1} - x_* &= x_k - x_* - H_k^{-1} g_k = x_k - x_* - H_k^{-1} (g_k - g(x_*)) \\ &= H_k^{-1} (g(x_*) - g_k - H_k(x_* - x_k)). \end{aligned}$$

But Theorem 1.3  $\implies$

$$\|g(x_*) - g_k - H_k(x_* - x_k)\|_2 \leq \gamma \|x_* - x_k\|_2^2$$

$\implies$

$$\|x_{k+1} - x_*\|_2 \leq \gamma \|H_k^{-1}\|_2 \|x_* - x_k\|_2^2$$

which is (iii) when  $\kappa = 2\gamma/\lambda_{\min}(H_*)$ , for  $k \in \mathcal{K}$ .

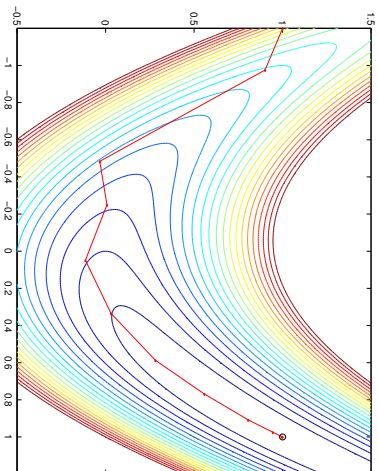
Result (ii) follows since once iterate becomes sufficiently close to  $x_*$ ,

(iii) for  $k \in \mathcal{K}$  sufficiently large implies  $k + 1 \in \mathcal{K} \implies \mathcal{K} = \mathbb{N}$ . Thus

(i) and (iii) are true for all  $k$  sufficiently large.



## NEWTON METHOD EXAMPLE



Contours for the objective function  $f(x, y) = 10(y - x^2)^2 + (x - 1)^2$ , and the iterates generated by the Generic Linesearch Newton method

## QUASI-NEWTON METHODS

Various attempts to approximate  $H_k$ :

- Finite-difference approximations:

$$(H_k)e_i \approx h^{-1}(g(x_k + he_i) - g_k) = (B_k)e_i$$

for some “small” scalar  $h > 0$

- Secant approximations: try to ensure the **secant condition**

$$B_{k+1}s_k = y_k \approx H_{k+1}s_k, \text{ where } s_k = x_{k+1} - x_k \text{ and } y_k = g_{k+1} - g_k$$

- Symmetric Rank-1 method** (but may be indefinite or even fail):

$$B_{k+1} = B_k + \frac{(y_k - B_k s_k)(y_k - B_k s_k)^T}{(y_k - B_k s_k)^T s_k}$$

- BFGS method**: (symmetric and positive definite if  $y_k^T s_k > 0$ ):

$$B_{k+1} = B_k + \frac{y_k y_k^T}{y_k^T s_k} - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k}$$

## MODIFIED NEWTON METHODS

If  $H_k$  is indefinite, it is usual to solve instead

$$(H_k + M_k)p_k \equiv B_k p_k = -g_k$$

where

- $M_k$  chosen so that  $B_k = H_k + M_k$  is “sufficiently” positive definite
- $M_k = 0$  when  $H_k$  is itself “sufficiently” positive definite

Possibilities:

- If  $H_k$  has the spectral decomposition  $H_k = Q_k D_k Q_k^T$  then

$$B_k \equiv H_k + M_k = Q_k \max(\epsilon, |D_k|) Q_k^T$$

- $M_k = \max(0, \epsilon - \lambda_{\min}(H_k))I$

- Modified Cholesky**:  $B_k \equiv H_k + M_k = L_k L_k^T$

## MINIMIZING A CONVEX QUADRATIC MODEL

For convex models ( $B_k$  positive definite)

$$p_k = (\text{approximate}) \arg \min_{p \in \mathbb{R}^n} f_k + p^T g_k + \frac{1}{2} p^T B_k p$$

**Generic convex quadratic problem**: ( $B$  positive definite)

$$(\text{approximately}) \minimize_{p \in \mathbb{R}^n} q(p) = p^T g + \frac{1}{2} p^T B p$$

## MINIMIZATION OVER A SUBSPACE

- $D^i = (d^0 : \dots : d^{i-1})$
- Subspace  $\mathcal{D}^i = \{p \mid p = D^i p_d \text{ for some } p_d \in \mathbb{R}^i\}$
- $p^i = \arg \min_{p \in \mathcal{D}^i} q(p)$
- $\implies D^{i^T} g^i = 0$ , where  $g^i = B p^i + g$
- $p^{i-1} \in \mathcal{D}^i$
- $\implies p^i = p^{i-1} + D^i p_d^i$ , where
- $p_d^i = \arg \min_{p_d \in \mathbb{R}^i} D^{i^T} g^{i-1} + \frac{1}{2} p_d^T D^{i^T} B D^i p_d$
- $= -(D^{i^T} B D^i)^{-1} D^{i^T} g^{i-1} = -d^{i-1^T} g^{i-1} (D^{i^T} B D^i)^{-1} e_i$
- $\implies p^i = p^{i-1} - d^{i-1^T} g^{i-1} D^i (D^{i^T} B D^i)^{-1} e_i$

## CONJUGATE-GRADIENT METHOD

Given  $p^0 = 0$ , set  $g^0 = g$ ,  $d^0 = -g$  and  $i = 0$ .  
 Until  $g^i$  "small" iterate

$$\alpha^i = -g^{i^T} d^i / d^{i^T} B d^i$$

$$p^{i+1} = p^i + \alpha^i d^i$$

$$g^{i+1} = g^i + \alpha^i B d^i$$

$$\beta^i = \|g^{i+1}\|_2^2 / \|g^i\|_2^2$$

$$d^{i+1} = -g^{i+1} + \beta^i d^i$$

and increase  $i$  by 1

Important features

- $d^{i^T} g^{i+1} = 0$  for all  $j = 0, \dots, i \implies \alpha^i = \|g^i\|_2^2 / d^{i^T} B d^i$
- $g^{j^T} g^{i+1} = 0$  for all  $j = 0, \dots, i$
- $g^T p^i < 0$  for  $i = 1, \dots, n \implies$  descent direction for any  $p_k = p^i$

## MINIMIZATION OVER A CONJUGATE SUBSPACE

Minimizer over  $\mathcal{D}^i$ :  $p^i = p^{i-1} - d^{i-1^T} g^{i-1} (D^{i^T} B D^i)^{-1} e_i$

Suppose in addition the members of  $\mathcal{D}^i$  are B-conjugate:

- **B-conjugacy**:  $d_i^T B d_j = 0$  ( $i \neq j$ )

$$\implies p^i = p^{i-1} + \alpha^{i-1} d^{i-1}, \text{ where}$$

$$\alpha^{i-1} = -\frac{d^{i-1^T} g^{i-1}}{d^{i-1^T} B d^{i-1}}$$

### Building a B-conjugate subspace

Since  $g^i$  is independent of  $\mathcal{D}^i$ , let  $d^i = -g^i + \sum_{j=0}^{i-1} \beta^j g^j$

- choose  $\beta^j$  so that  $d^i$  is B-conjugate to  $\mathcal{D}^i$
- $\implies \beta^j = 0$  ( $j < i-1$ ),  $\beta^{i-1} \equiv \beta^i = \frac{\|g^i\|_2^2}{\|g_{i-1}\|_2^2}$

## CONJUGATE GRADIENT METHOD GIVES DESCENT

$$g^{i-1^T} d^{i-1} = d^{i-1^T} (g + B p^{i-1}) = d^{i-1^T} g + \sum_{j=0}^{i-2} \alpha_j d^{i-1^T} B d^j = d^{i-1^T} g$$

$p^i$  minimizes  $q(p)$  in  $\mathcal{D}^i \implies$

$$p^i = p^{i-1} - \frac{g^{i-1^T} d^{i-1}}{d^{i-1^T} B d^{i-1}} d^{i-1} = p^{i-1} - \frac{g^T d^{i-1}}{d^{i-1^T} B d^{i-1}} d^{i-1}.$$

$\implies$

$$g^T p^i = g^T p^{i-1} - \frac{(g^T d^{i-1})^2}{d^{i-1^T} B d^{i-1}},$$

$\implies g^T p^i < g^T p^{i-1} \implies$  (induction)

$$g^T p^i < 0$$

since

$$g^T p^1 = -\frac{\|g\|_2^4}{g^T B g} < 0.$$

$\implies p_k = p^i$  is a descent direction