CS5321 Numerical Optimization

03 Line Search Methods

Line search method

1. Given a point x_k, find a descent direction p_k.
2. Find the step length α to minimize $f(x_k+\alpha p_k)$

- A descent direction p_k means the directional derivative $\nabla f(x_k)^T p_k < 0$
- In Newton's method, $p_k = -H_k g_k$
  - Matrix $H_k=\nabla^2 f(x_k)$ is Hessian; $g_k= \nabla f(x_k)$ is gradient
  - If H_k is positive definite, $\nabla f(x_k)^T p_k= -g_k^T H_k g_k < 0$

Modified Newton's methods

- If the Hessian H is indefinite, ill-conditioned, or singular, the modified Newton's methods compute the inverse of $H+E$, such that $-(H+E)^{-1}\nabla f(x_k)$ gives a descent direction.
- Matrix E can be calculated via
  1. Eigenvalue modification
  2. Diagonal shift
  3. Modified Cholesky factorization
  4. Modified symmetric indefinite factorization

1. Eigenvalue modification

- Modifying the eigenvalues of H such that H becomes positive definite
- Let $H=Q\Lambda Q^{\mathrm{T}}$ be the spectral decomposition of H.
  - $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2,\ldots\lambda_n)$ where $\lambda_1 \geq \ldots \geq \lambda_i > 0 \geq \lambda_{i+1} \ldots \geq \lambda_n$
  - Define $\Delta\Lambda = \mathrm{diag}(0,\ldots,0,\Delta\lambda_{i+1},\ldots,\Delta\lambda_n)$ s.t. $\Lambda + \Delta\Lambda > 0$
  - Matrix $E=Q\Delta\Lambda Q^{\mathrm{T}}$
- Problem: the eigenvalue decomposition is too expensive.

2. Diagonal shift

- If $\Delta\Lambda = \mathrm{diag}(\tau, \ldots, \tau) = \tau I$, $E = Q\Delta\Lambda Q^T = \tau QQ^T = \tau I$
  - To make $H+E$ positive definite, only need to choose $\tau > |\lambda_i|$, where λ_i is minimum negative eigenvalues of H.
- How to know τ without explicitly performing the eigenvalue decomposition?
  - If H is positive definite, H has Cholesky decomposition.
  - Guess a shift τ and try Cholesky decomposition on $H+\tau I$. If fails, increase τ and try again, until succeed
  - The choice of increment is heuristic.

3. Modified Cholesky factorization

- A SPD matrix H can be decomposed as $H=LDL^T$.
  - L is unit triangular matrix
  - D is diagonal with positive elements
  - Relations to Cholesky decomp $H=MM^T$ is $M=LD^{1/2}$
- If A is not SPD, modify the elements of L and D during the decomposition such that
  - $D(i,i) \geq \delta > 0$ and $L(i,j)D(i,i)^{1/2} \leq \beta$.
- The decomposition can be used in solving linear systems.

4. Modified symmetric indefinite factorization

- Symmetric indefinite factorization $PHP^{\mathrm{T}}=LBL^{\mathrm{T}}$
  - better numerical stability than Cholesky decomposition
  - Matrix B has the same *inertia* as H.
    - The inertia of a matrix is the number of positive, zero, and negative eigenvalues of the matrix.
  - Use eigenvalue modification to B st. B+F is positive definite
    - The eigen-decomp of B is cheap since it is block diagonal
  - Thus, $P(H+E)P^{\mathrm{T}}=L(B+F)L^{\mathrm{T}}$ and $E=P^{T}LFL^{T}P$

Step length α

- Assume p_k is a descent direction. Find an optimal step length α.

$$\min_{\alpha>0} \phi(\alpha) = f(x_k + \alpha p_k)$$

- The minimization problem may be difficult to solve. (nonlinear)
- Alternative method is to find an α that satisfies some conditions
  - Wolfe conditions
  - Goldstein conditions

The Goldstein conditions

- With 0< c <1/2,

$$\begin{aligned} f(x_k) + (1-c)\alpha \nabla f_k^T p_k &\leq f(x_k + \alpha p_k) \\ &\leq f(x_k) + c\alpha \nabla f_k^T p_k \end{aligned}$$

- Suitable for Newton-typed methods, but not well suited for quasi-Newton methods

The Wolfe conditions

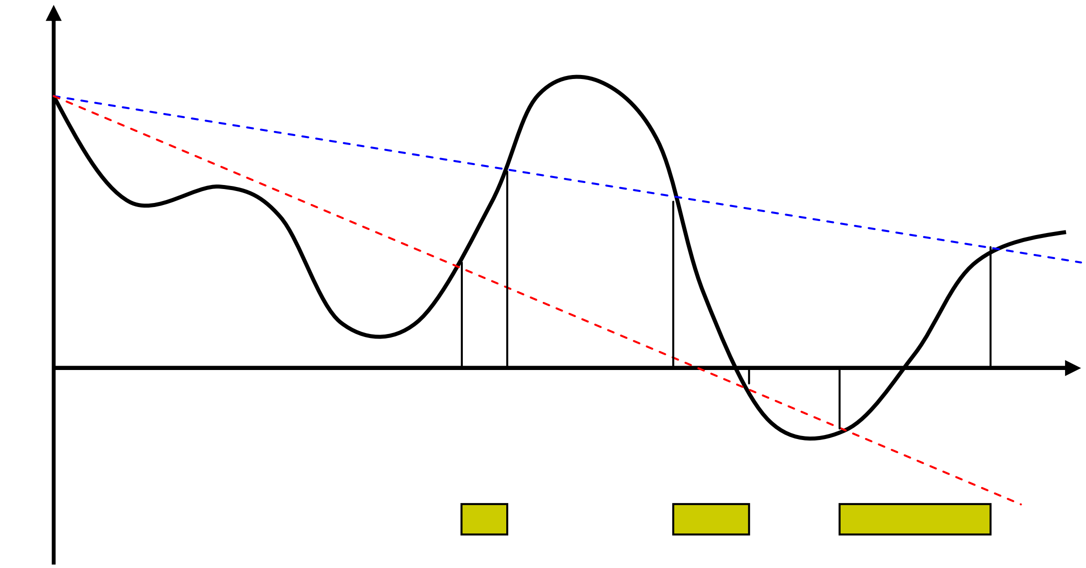
- Sufficient decrease condition: for $c_1 \in (0,1)$

$$f(x_k + \alpha p_k) \leq f(x_k) + c_1 \alpha \nabla f(x_k)^T p_k$$

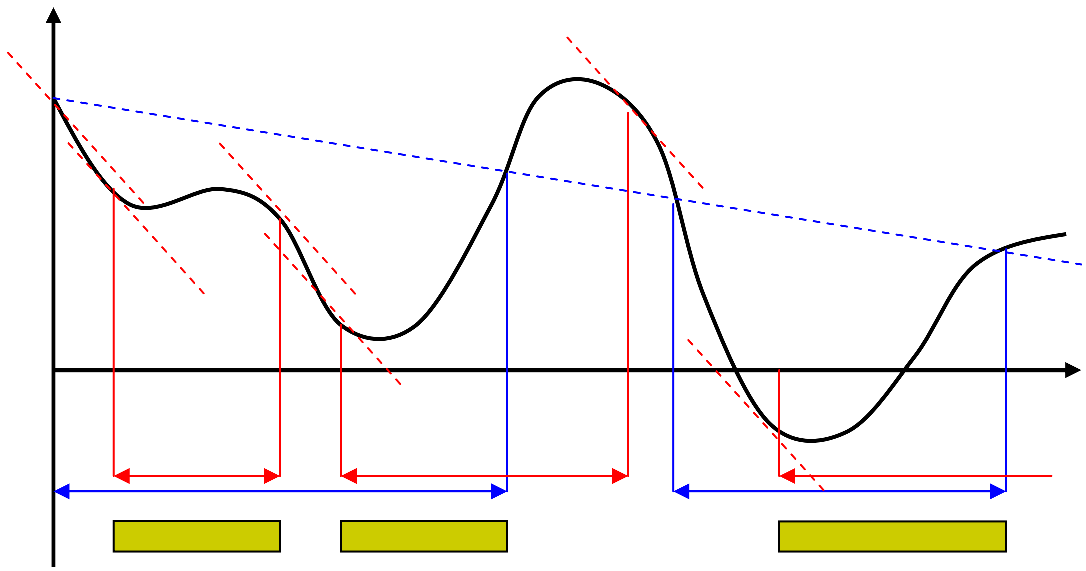
- Curvature condition: for $c_2 \in (0, c_1)$

$$\nabla f(x_k + \alpha p_k)^T p_k \geq c_2 \nabla f(x_k)^T p_k$$

  - Usually choose c_2=0.9 for Newton's method and c_2=0.1 for conjugate gradient method



The Coldstein conditions



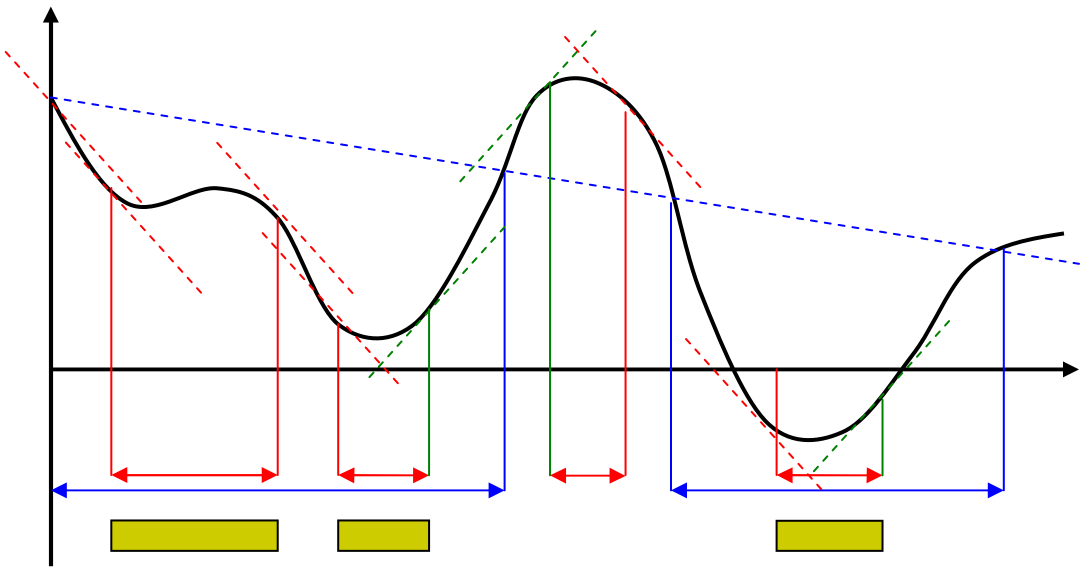
The Wolfe conditions

The strong Wolfe conditions

- Limit the range of $\phi'(\alpha_k)$ from both sides

$$f(x_k + \alpha p_k) \le f(x_k) + c_1 \alpha \nabla f(x_k)^T p_k$$

$$|\nabla f(x_k + \alpha p_k)^T p_k| \le c_2 |\nabla f(x_k)^T p_k|$$



The strong
Wolfe conditions

Convergence of line search

- For a descent direction p_k and a step length α that satisfies the Wolfe condition, if f is bounded below and continuously differentiable in an open neighborhood $\mathcal{N}$ and ∇f is Lipschitz continuous in $\mathcal{N}$, then

$$\cos^2 \theta_k \|\nabla f(x_k)\|^2 \to 0$$

  - θ_k is the angle between p_k and $\nabla f(x_k)$
  - Note that can be either $\cos\theta_k \to 0$ or $\|\nabla f(x_k)\| \to 0$
  - Other two methods have similar results