CS5321
Numerical Optimization

06 Quasi-Newton Methods

Hessian matrix

- The Hessian matrix B is needed in computing Newton's direction $p_k = -B_k^{-1}\nabla f_k$.
  - What needed is the inverse multiplying a vector.
  - Approximate result maybe good enough in use.
- Quasi-Newton: use gradient vector to approximate Hessian
  - DFP and BFGS updating formula
  - SR1 (Symmetric Rank 1 update) method
  - Broyden class

The secant formula

- Suppose we have Hessian B_k at current point x_k
- Next point is $x_{k+1} = x_k + \alpha_k p_k$
- The model at x_{k+1} is $m_{k+1}(p) = f_{k+1} + g_{k+1}^T p + \frac{1}{2} p^T B_{k+1} p$
- Assume $\nabla m_{k+1}(-\alpha_k p_k) = g_{k+1} - \alpha_k B_{k+1} p_k = g_k$ and let $s_k = x_{k+1} - x_k = \alpha_k p_k$ and $y_k = g_{k+1} - g_k$
  - Then we have the *secant* formula

$$B_{k+1} s_k = y_k$$

  - Also require B_{k+1} to be spd $s_k^T B_{k+1} s_k = s_k^T y_k > 0$

DFP updating formula

- The problem of approximating B_{k+1} becomes

$$\min_B \|B - B_k\|$$

$$\text{subject to } B = B^T, Bs_k = y_k$$

- With some special norm, the solution is

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

$$\rho_k = 1 / y_k^T s_k$$

  - This is the DFP updating formula
    - Davidon, Fletcher and Powell

BFGS updating formula

- Let $H_k = B_k^{-1}$. Similar results can be obtained

$$\min_H \|H - H_k\|$$

$$\text{subject to } H = H^T, H y_k = s_k$$

- The solution is

$$H_{k+1} = (I - \rho_k s_k y_k^T) H_k (I - \rho_k y_k s_k^T) + \rho_k s_k s_k^T$$
$$\rho_k = 1 / y_k^T s_k$$

  - This is the BFGS updating formula
    - Broyden, Fletcher, Goldfarb, and Shanno

More on BFGS

- The initial H_0 need be constructed
  - Finite difference or automatic differentiation (chap 8)
  - Using identity matrix or $(y_k^T s_k / y_k^T y_k) I$
- Hessian B_k can be constructed via the Sherman-Morrison-Woodbury formula (rank 2 update)

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

- If H_k is positive definite, H_{k+1} is positive definite.

The SR1 method

- Desire a symmetric rank 1 update, $B_{k+1}=\mathrm{B}_k+\rho vv^{\mathrm{T}}$, that satisfies the secant constrain: $y_k=B_{k+1}s_k$.

$$y_k = (B_k + \rho v v^T)s_k = B_k s_k + (\rho v^T s_k)v$$

  - Vector v is parallel to $(y_k-B_ks_k)$. Let $v = \delta(y_k-B_ks_k)$
  - Substitute back to the secant constrain

$$\rho = \mathrm{sign}[s_k^T(y_k - B_k s_k)], \pm = \S\, [s_k^T(y_k - B_k s_k)]^{-1/2}$$

  - Let $z_k=y_k-B_ks_k$. The SR1 is $B_{k+1} = B_k + \dfrac{z_k z_k^T}{z_k^T s_k}$

More on the SR1

- Let $w_k = s_k - H_k y_k$. By Sherman-Morrison formula,

$$H_{k+1} = H_k + \frac{w_k w_k^T}{w_k^T y_k}$$

- If $y_k = B_k s_k$, $B_{k+1} = B_k$.
- If $y_k \neq B_k s_k$ but $(y_k - B_k s_k)^T s_k = 0$, there is no the SR1.
  - If $(y_k - B_k s_k)^T s_k \approx 0$, the SR1 is numerical instable.
  - Use $B_{k+1} = B_k$.
- In practice, the B_k generated by the SR1 satisfies

$$\lim_{k\to\infty} \|B_k - \nabla^2 f(x^*)\| = 0$$

The Broyden class

- The Broyden class is a family of quasi-Newton updating formulas that have the form

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

  - ϕ_k is a scalar function. $v_k = \left(\frac{y_k}{y_k^T s_k} - \frac{B_k s_k}{s_k^T B_k s_k} \right)$

- It is a linear combination of BFGS and DFP

$$B_{k+1}^{\text{Broyden}} = (1-\phi) B_{k+1}^{\text{BFGS}} + \phi B_{k+1}^{\text{DFP}}$$

  - It is a *restricted Broyden class* if $\phi \in [0,1]$

More on the Broyden class

- The SR1 is also belong to the Broyden class with

$$\phi_k = \frac{s_k^T y_k}{s_k^T y_k - s_k^T B_k s_k}$$

  - But may not be belong to the restricted Broyden class.
- If $\phi_{k+1} \geq 0$ and B_k is positive definite, then B_k is positive definite.

Convergence of BFGS

- Global convergence

  For any spd B_0 and any x_0, if f is twice continuously differentiable and the working set is convex with *properly bounded* Hessian, then the BFGS converges to the minimizer x^* of f.

- Convergent rate

  If f is twice continuously differentiable and $\nabla^2 f$ is Lipschitz continuous near x^*, and $\sum_{k=1}^{\infty} \|x_k - x^*\| < \infty$ then the BFGS converges superlinearly