CS5321
Numerical Optimization

07 Large-Scale Unconstrained Optimization

Large-scaled optimizations

- The problem size n may be thousands to millions.
- Storage of Hessian is n^2.
  - Even if it is sparse, its decompositions (LU, Cholesky…) and its approximations (BFGS, SR1…) are not.
- Methods
  - Inexact Newton methods
    - Line-search: Truncated Newton method
    - Trust region: Steihaug's algorithm
  - Limited memory BFGS, Sparse quasi-Newton updates
  - Algorithms for partially separable functions

Inexact Newton method

- Inexact Newton methods use iterative methods to solve the Newton's direction $p = -H^{-1}g$ inexactly.
  - The exactness is measured by the residual $r_k = H_k p_k + g_k$
  - It stops when $\|r_k\| \le \eta_k \|g_k\|$ for $0 < \eta_k < 1$.
- Convergence (Theorem 7.1, 7.2)

  If H is spd for x near x^*, and x_0 is close enough to x^*, the inexact Newton method converges to x^*. If $\eta_k \to 0$, the convergence is superlinear. In addition, if H is Lipschitz continuous for x near x^*, the convergence is quadratic.

Truncated Newton method

- CG + line search + termination conditions
- What CG need is matrix-vector multiplications
  - Use finite difference to approximate Hessian (multiplying a vector d.).

$$Hd = \nabla^2 f_k d = \frac{\nabla f(x_k + hd) - \nabla f(x_k)}{h}$$

  - The cost is the computation of $\nabla f\,(x_k+hd)$
- Additional termination conditions
  - When negative curvature is detected $p^T Hp<0$, return $-g$

Steihaug's algorithm

- CG + trust-region + termination conditions
  - Change terminations of CG: $x \to z$, $p \to d$. Index j
- Additional termination conditions
  1. When negative curvature is detected $p^T H p < 0$,
  2. When the step size $\|z_j\| \geq \Delta_k$,
     - It can be shown that $\|z_j\|$ increases monotonically
- When stops abnormally, it returns $z_j + \tau d_j$, where τ minimizes $m_k(z_j + \tau d_j)$ subject to $\| z_j + \tau d_j \| = \Delta_k$.

Limited memory BFGS

- Review of BFGS

$$x_{k+1} = x_k + \alpha_k p_k,\ s_k = x_{k+1} - x_k = \alpha_k p_k,\ y_k = g_{k+1} - g_k$$

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

- L-BFGS: Not form H_k explicitly
  - Store s_k and y_k for m iterations (m << n)

$$\begin{aligned} H_k &= (V_{k-1}^T \cdots V_{k-m}^T) H_k^0 (V_{k-m}^T \cdots V_{k-1}^T) \\ &\quad + \rho_{k-m} (V_{k-1}^T \cdots V_{k-m+1}^T) s_{k-m} s_{k-m}^T (V_{k-m+1}^T \cdots V_{k-1}^T) \\ &\quad + \rho_{k-m+1} (V_{k-1}^T \cdots V_{k-m+2}^T) s_{k-m} s_{k-m}^T (V_{k-m+2}^T \cdots V_{k-1}^T) \\ &\quad + \cdots + \rho_{k-1} s_{k-1} s_{k-1}^T \end{aligned}$$

Compact form of L-BFGS

$$B_k = B_0 - \begin{pmatrix} B_0 S_k & Y_k \end{pmatrix} \begin{pmatrix} S_k^T B_0 S_k & L_k \\ L_k^T & -D_k \end{pmatrix}^{-1} \begin{pmatrix} S_k^T B_0 \\ Y_k^T \end{pmatrix}$$

$$S_k = \begin{pmatrix} s_0 & \dots & s_{k-1} \end{pmatrix}$$

$$Y_k = \begin{pmatrix} y_0 & \dots & y_{k-1} \end{pmatrix}$$

$$(L_k)_{i,j} = \begin{cases} s_{i-1}^T y_{j-1} & \text{if } i > j, \\ 0 & \text{otherwise,} \end{cases}$$

$$D_k = \mathrm{diag}(s_0^T y_0, \dots, s_{k-1}^T y_{k-1})$$

- Only need to store S_k, Y_k, L_k, D_k.
  - S_k, Y_k are $n \times m$.
  - L_k is $m \times m$ upper triangular.
  - D_k is $m \times m$ diagonal.

Sparse quasi-Newton updates

- Compute B_{k+1} that is symmetric and
  - has the same sparsity as the exact Hessian.
  - satisfies secant direction. $B_{k+1}s_k=y_k$.
- Sparsity: $\Omega=\{(i,j) \mid [\nabla^2 f(x)]_{ij}\neq 0\}$

$$\min_{B} \|B - B_k\|_F^2 = \sum_{(i,j)\in\Omega} [B_{ij} - (B_k)_{ij}]^2$$

  - Subject to $B_{k+1}s_k=y_k$, $B=B^T$ and $B_{ij}=0$ for $(i,j)\in\Omega$
  - Constrained nonlinear least square problem
- The solution may fail to be positive definite.

Partially separable functions

- Separable object function: $f(x) = f_1(x_1,x_3)+f_2(x_2,x_4)$
- Partially separable object function $f(x)$:
  - $f(x)$ can be decomposed as a sum of element functions f_i, which depends only a few components of x.

$$f(x) = (x_1 - x_3^2)^2 + (x_2 - x_4^2)^2 + (x_3 - x_2^2)^2 + (x_4 - x_1^2)^2$$

- Gradient and Hessian are linear operators

$$f(x) = \sum_{i=1}^{\ell} f_i(x), \quad \nabla f(x) = \sum_{i=1}^{\ell} \nabla f_i(x), \quad \nabla^2 f(x) = \sum_{i=1}^{\ell} \nabla^2 f_i(x),$$

  - Compute Hessian of each element functions separately