Unconstrained Optimization

January 1, 2023

General Smooth Function $\psi(\mathbf{x})$

- Iterative methods to generate an improved sequence $\{\mathbf{x}_k\}$ converging to solution $\mathbf{x}^*$
- Each step, objective funtion approximated by Taylor expansion

$$\psi(\mathbf{x}_k + \Delta\mathbf{x}) \approx \psi(\mathbf{x}_k) + \mathbf{g}^T(\mathbf{x}_k)\Delta\mathbf{x} + \frac{1}{2}\Delta\mathbf{x}^T\mathbf{H}(\mathbf{x}_k)\Delta\mathbf{x}$$

step reduction

$$\Delta\psi(\mathbf{x}_k) \approx \mathbf{g}^T(\mathbf{x}_k)\Delta\mathbf{x} + \frac{1}{2}\Delta\mathbf{x}^T\mathbf{H}(\mathbf{x}_k)\Delta\mathbf{x}$$

- Gradient $\mathbf{g}(\mathbf{x}_k)$ is always required.
- Hessian $\mathbf{H}(\cdot)$, depends on the choice of methods.
- $\mathbf{x}^*$ satisfies stationary condition $\|\mathbf{g}(\mathbf{x}^*)\| = 0$ and curvature condition $\mathbf{H}(\mathbf{x}^*)$ being at least positive semidefinite.

Line Search Methods

Step Length

- $\Delta x = \alpha \mathbf{p}$ for step length α and direction $\mathbf{p}$
- acceptable α: not undershooting, not overshooting.
- Wolfe conditions:

$$\psi(\mathbf{x}^k + \alpha \mathbf{p}) \leq \psi(\mathbf{x}^k) + c_1 \alpha \mathbf{g}(\mathbf{x}^k)^T \mathbf{p},$$
$$\mathbf{g}(\mathbf{x}^k + \alpha \mathbf{p})^T \mathbf{p} \geq c_2 \mathbf{g}(\mathbf{x}^k)^T \mathbf{p}$$

, where $0 < c_1 < c_2 < 1$.

- Strong Wolfe conditions:

$$\psi(\mathbf{x}^k + \alpha \mathbf{p}) \leq \psi(\mathbf{x}^k) + c_1 \alpha \mathbf{g}(\mathbf{x}^k)^T \mathbf{p},$$
$$\|\mathbf{g}(\mathbf{x}^k + \alpha \mathbf{p})^T \mathbf{p}\| \leq -c_2 \mathbf{g}(\mathbf{x}^k)^T \mathbf{p}$$

, where $0 < c_1 < c_2 < 1$.

Step Length (continued)

- fix search direction **p** and write objective function as $\psi(\alpha)$
- search $\alpha \in (\alpha_{lo}, \alpha_{hi})$, where initially $\alpha_{lo} = 0$ and α_{hi} is a max step length
- generate trial sequence $\{\alpha_i\}$ by safeguarded quadratic or cubic interpolation of $\psi(\alpha_i)$
- reduce the search interval $(\alpha_{lo}, \alpha_{hi})$ by testing the Wolfe condition at each α_i.

Line Search Methods

Search Direction p: Steepest Descent

- $\Delta\psi(\mathbf{x}_k) \approx \alpha \mathbf{g}^T(\mathbf{x}_k)\mathbf{p} + \frac{\alpha^2}{2}\mathbf{p}^T\mathbf{H}(\mathbf{x}_k)\mathbf{p}$
- take $\mathbf{p} = -\mathbf{g}(\mathbf{x}_k)$, the first order term dominates for small α
- works well when $\psi(\mathbf{x})$ doesn't have strong curvature

Line Search Methods

Search Direction p: Modified Newton

- classic Newton direction is to solve $\mathbf{H}(\mathbf{x}_k)\mathbf{p} = -\mathbf{g}(\mathbf{x}_k)$
- compute modified Cholesky $\mathbf{B}_k = \mathbf{H}(\mathbf{x}_k) + \mathbf{E} = \mathbf{L}^T\mathbf{D}\mathbf{L}$ such that $\|\mathbf{E}\|_\infty$ is minimized and solve $\mathbf{L}^T\mathbf{D}\mathbf{L}\mathbf{p} = -\mathbf{g}(\mathbf{x}_k)$
- if $\mathbf{H}(\mathbf{x}_k)$ is positivesemi definite, $\mathbf{E} = 0$, $\mathbf{p}$ is the classic Newton direction.
- if $\mathbf{H}(\mathbf{x}_k)$ is indefinite, $\mathbf{B}_k$ is the "closest" modification. $\mathbf{p}$ is still a descent direction.
- if $\mathbf{x}_k$ is stationary but $\|\mathbf{E}\|_\infty > 0$, the algorithm renders a negative curvature direction $\mathbf{p}$.
- if $\mathbf{H}(\cdot)$ is not available, approximate it by finite difference.

Line Search Methods

Search Direction p: Quasi-newton

- solve $\mathbf{B}_k\mathbf{p} = -\mathbf{g}(\mathbf{x}_k)$, where $\mathbf{B}_k$ is a positive definite approximation of $\mathbf{H}(\mathbf{x}_k)$
- $\mathbf{B}_k = \mathbf{B}_{k-1} + \mathbf{U}_k$, where $\mathbf{U}_k$ is a rank-1 or rank-2 update matrix.
- BFGS:

$$\mathbf{U}_k = \frac{1}{\mathbf{g}(\mathbf{x}_k)^T\mathbf{p}}\mathbf{g}(\mathbf{x}_k)\mathbf{g}(\mathbf{x}_k)^T + \frac{1}{\alpha\mathbf{y}^T\mathbf{p}}\mathbf{y}\mathbf{y}^T, \mathbf{y} = \mathbf{g}(\mathbf{x}_k) - \mathbf{g}(\mathbf{x}_{k-1})$$

- given Cholesky factorization $\mathbf{B}_{k-1} = \mathbf{L}_{k-1}\mathbf{L}_{k-1}^T$, obtain $\mathbf{B}_k = \mathbf{L}_k\mathbf{L}_k^T$ by economy matrix update introduced by $\mathbf{U}_k$.

Search Direction p: Conjugate Gradient (CG)

- assume $\mathbf{H}(\cdot)$ is positive definite, the general CG step update is:

$$\begin{aligned}\mathbf{p}_0 &= -\mathbf{g}(\mathbf{x}_0)\\ \mathbf{p}_k &= -\mathbf{g}(\mathbf{x}_k) + \beta\mathbf{p}_{k-1}\end{aligned}$$

- the choice of β needs satisfy CG properties yet produce a descent direction $\mathbf{p}$
- Polak-Ribiere+ (PR+) method:

$$\beta = \max\left(\frac{\mathbf{g}(\mathbf{x}_k)^T(\mathbf{g}(\mathbf{x}_k) - \mathbf{g}(\mathbf{x}_{k-1}))}{\|\mathbf{g}(\mathbf{x}_{k-1})\|^2}, 0\right)$$

- together with a strong Wolfe condition with $0 < c_1 < c_2 < \frac{1}{2}$, PR+ satisfies all necessary properties.

Trust Region Methods

- General idea:

$$\arg\min_{\Delta\mathbf{x}} \Delta\psi(\mathbf{x}_k) = \mathbf{g}(\mathbf{x}_k)^T\Delta\mathbf{x} + \frac{1}{2}\Delta\mathbf{x}^T\mathbf{H}(\mathbf{x}_k)\Delta\mathbf{x}, \text{ s.t. } \|\Delta\mathbf{x}\| \leq \Delta_k \quad (1)$$

- Δ_k is an appropriately choosen trust region radius at each iteration
- $\rho_k = \frac{\psi(\mathbf{x}_k+\Delta\mathbf{x})-\psi(\mathbf{x}_k)}{\Delta\psi(\mathbf{x}_k)}$ measures the actual reduction relative to model reduction
- adjust Δ_k based on how good is ρ_k

Trust Region Methods

Nearly Exact Search Direction

- for moderate problem size, we can solve sub-problem (1) exact.
- global solution $\Delta\mathbf{x}^*$ to problem (1) exists iff,

$$(\mathbf{H}(\mathbf{x}_k) + \lambda\mathbf{I})\,\Delta\mathbf{x}^* = -\mathbf{g}(\mathbf{x}_k) \tag{2}$$

$$\lambda\,(\Delta_k - \|\Delta\mathbf{x}^*\|) = 0 \tag{3}$$

$$(\mathbf{H}(\mathbf{x}_k) + \lambda\mathbf{I}) \text{ is at least positive semidefinite} \tag{4}$$

- given λ, $\Delta\mathbf{x}^*$ can be computed from equation (2)
- trick is to solve λ
  - $\mathbf{H}(\cdot)$ is positive definite, $\lambda = 0$ or root finding.
  - $\mathbf{H}(\cdot)$ is semi-definite or indefinite, need explore eigen structure such that the modified matrix $(\mathbf{H}(\mathbf{x}_k) + \lambda\mathbf{I})$ is positive definite.
  - need choose appropriate matrix factorization in different situations for best performance.

Trust Region Methods

Conjugate Gradient-Steihaug Direction

- for large problem, sufficient to get an non-exact but descent direction at each iteration k.
- generate a sequence $\{(\alpha_i, \mathbf{d}_i)\}$ of step length α_i and CG directions $\mathbf{d}_i$ computed as usual, initially choose $\mathbf{d}_0 = -\mathbf{g}(\mathbf{x}_k)$.
- for each i, try $\alpha_i = 1$ and $\mathbf{p} = \sum_{j=0}^{i-1} \alpha_j \mathbf{d}_j + \alpha_i \mathbf{d}_i$. If $\|\mathbf{p}\| > \Delta_k$, scale down α_i such that $\|\mathbf{p}\| = \Delta_k$ and make a step move with $\Delta \mathbf{x} = \mathbf{p}$.

Trust Region Methods

Conjugate Gradient-Steihaug Direction (continued)

- to ensure CG propterties and descent direction,
  - test $\mathbf{d}_j^T \mathbf{H}(\mathbf{x}_k)\mathbf{d}_j > 0, \forall j < i$ and stop at the first i such that $\mathbf{d}_i^T \mathbf{H}(\mathbf{x}_k)\mathbf{d}_i \leq 0$, let $\mathbf{q} = \sum_{j=0}^{i-1} \alpha_j \mathbf{d}_j$
  - $\mathbf{q}$ is certainly an acceptable choice of $\Delta\mathbf{x}$
  - can obtain further reduction to choose $\Delta\mathbf{x} = \mathbf{q} + \tau\mathbf{d}_i$ for some τ

$$\Delta\psi(\mathbf{x}_k) = \underbrace{\mathbf{g}(\mathbf{x}_k)^T\mathbf{q} + \frac{1}{2}\mathbf{q}^T\mathbf{H}(\mathbf{x}_k)\mathbf{q}}_{\mathbf{q}\text{ reduction component}} + \underbrace{\tau\mathbf{g}^T(\mathbf{x}_k)\mathbf{d}_i + \tau^2\frac{1}{2}\mathbf{d}_i^T\mathbf{H}(\mathbf{x}_k)\mathbf{d}_i}_{\mathbf{d}_i\text{ reduction component}}$$

  - choose τ with correct sign and $\|\Delta\mathbf{x}\| = \Delta_k$

References I

[1] Jorge Nocedal and Stephen J. Wright: Numerical Optimization, Springer, 1999

[2] Philip E. Gill, Walter Murray and Margaret H. Wright: Practical Optimization, Academic Press, 1981

[3] Philip E. Gill and Walter Murray: Newton-Type Methods for Unconstrained and Linearly Constrained Optimization. Mathematical Programming 7 (1974), pp. 311-350

[4] Philip E. Gill, G. H. Golub, Walter Murray and Michael. A. Saunders: Methods for Modifying Matrix Factorizations, Mathematics of Computation, Volumn 28, Number 126, April 1974, pages 505-535

[5] Philip E. Gill, Walter Murray and Michael A. Saunders: Methods for computing and modifying the LDV factors of a matrix, Mathematics of Computation, Volumn 29, Number 132, October 1975, pages 1051-1077

References II

[6] Anders. Forsgren, Philip E. Gill and Walter Murray: Computing Modified Newton Directions Using a Partial Cholesky Factorization, SIAM J. SCI. COMPUT. Vol. 16, No. 1, pp. 139-150

[7] J.E. Dennis and Robert B. Schnabel: A New Derivation of Symmetric Positive Definite Secant Updates; CU-CS-185-80 August 1980 Computer Science Technical Reports, Summer 8-1-1980, University of Colorado at Boulder