Numerical Optimization

Graduate Course

Unconstrained Smooth Optimization

Part II: Line search-based and steepest descent methods

Wen Huang

School of Mathematical Sciences
Xiamen University

Compiled on February 14, 2022

Line Search-Based Methods

A descent method

Input: Initial iterate x_0;
  Initial descent direction p_0 and set $k \leftarrow 0$;
  while not accurate enough **do**
    Set $x_{k+1} \leftarrow x_k + \alpha_k p_k$ with an appropriate step size α_k;
    Find a descent direction p_{k+1};
    $k \leftarrow k + 1$;
  end while

Direction! and step size!

Conditions for Step Size

Basic idea

Define $h(\alpha) = f(x + \alpha p)$. The task is to find an appropriate step size α.

Is a step size such that $h(\alpha) < h(0)$ sufficient for global convergence?

Define $h(\alpha) = f(x + \alpha p)$. The task is to find an appropriate step size α.

Is a step size such that $h(\alpha) < h(0)$ sufficient for global convergence?

No! See an example:

Define $h(\alpha) = f(x + \alpha p)$. The task is to find an appropriate step size α.

Is a step size such that $h(\alpha) < h(0)$ sufficient for global convergence?

No! See an example:

- $f(x) = \frac{1}{2}x^2$, $x^* = 0$, and $x_0 = 1$
- Choose step size $\alpha_k = 2 - \frac{1}{|x_k|2^{k+2}}$
- Update $x_{k+1} = x_k - \alpha_k \nabla f(x_k) = -x_k \left(1 - \frac{1}{|x_k|2^{k+2}}\right)$
- $x_{2k} = \frac{1}{2} + \frac{1}{2^{2k+1}}$ and $x_{2k+1} = -\frac{1}{2} - \frac{1}{2^{2k+2}}$
- Therefore, $f(x_{k+1}) < f(x_k)$
- However, $x_{2k} \to \frac{1}{2}$ and $x_{2k+1} \to -\frac{1}{2}$

Conditions for Step Size

Basic idea

- Decrease is not sufficient. We need "sufficient" decrease.
- Sufficient descent condition on the objective function:

$$h(\alpha) \leq h(0) + c_1 \alpha h'(0),$$

where $c_1 \in (0, 1)$. This condition is also called Armijo condition.

Sufficient decrease condition alone is still not sufficient for global convergence

- $f(x) = \frac{1}{2}x^2$, $x^* = 0$, and $x_0 = 1$
- $\alpha_k = \frac{1}{x_k 2^{k+2}}$
- $x_{k+1} = x_k - \alpha \nabla f(x_k) = x_k - \frac{1}{2^{k+2}}$
- $x_{k+1} = \frac{1}{2} + \frac{1}{2^{k+2}}$
- The sufficient decrease condition is satisfied with $c_1 = 0.5$
- However $x_{k+1} = x_0 - \sum_{i=0}^{k} \frac{1}{2^{i+2}} \to \frac{1}{2} \neq x^*$ as $k \to 0$

Sufficient decrease condition alone is still not sufficient for global convergence

- $f(x) = \frac{1}{2}x^2$, $x^* = 0$, and $x_0 = 1$
- $\alpha_k = \frac{1}{x_k 2^{k+2}}$
- $x_{k+1} = x_k - \alpha \nabla f(x_k) = x_k - \frac{1}{2^{k+2}}$
- $x_{k+1} = \frac{1}{2} + \frac{1}{2^{k+2}}$
- The sufficient decrease condition is satisfied with $c_1 = 0.5$
- However $x_{k+1} = x_0 - \sum_{i=0}^{k} \frac{1}{2^{i+2}} \to \frac{1}{2} \neq x^*$ as $k \to 0$

Step size can not be too small. (Note that the step size α_k converges to zero in this example)

Definition 1 (Armijo-Goldstein condition)

The step size α satisfies

$$h(\alpha) \leq h(0) + c_1 \alpha h'(0),$$

where α is the largest value in the set

$$\{t^{(i)} : t^{(i)} \in [\tau_1 t^{(i-1)}, \tau_2 t^{(i-1)}], t^{(0)} = 1\},$$

for any $c_1 \in (0, 1)$ and $0 < \tau_1 \leq \tau_2 < 1$.

Note that if $\tau_1 = \tau_2$, then the step size α can be found by a simple backtracking algorithm.

Backtracking

Input: Function $h : \mathbb{R} \to \mathbb{R}$ with $h'(0) < 0$; initial step size $\alpha^{(0)}$; shrinking parameter $\rho \in (0, 1)$;

Set step size $\alpha \leftarrow \alpha^{(0)}$;

while $h(\alpha) > h(0) + c_1 \alpha h'(0)$ **do**

  $\alpha \leftarrow \rho\alpha$;

end while

Let α denote the step size satisfying the Armijo-Goldstein condition.

- Sufficient descent condition by its definition

$$h(\alpha) \leq h(0) + c_1 \alpha h'(0)$$

- Not too small
  - $\alpha = 1$ if the initial step size is accepted
  - Otherwise,

$$h\left(\frac{\alpha}{\tau}\right) > h(0) + c_1 \frac{\alpha}{\tau} h'(0)$$

  for $\tau \in [\tau_1, \tau_2]$.

Definition 2 (Weak Wolfe conditions)

The step size α satisfies

$$\begin{aligned} h(\alpha) &\leq h(0) + c_1 \alpha h'(0) \text{ (Armijo condition), and} \\ h'(\alpha) &\geq c_2 h'(0) \text{ (Curvature condition),} \end{aligned}$$

for any $0 < c_1 < c_2 < 1$.

Definition 3 (Strong Wolfe conditions)

The step size α satisfies

$$\begin{aligned} h(\alpha) &\leq h(0) + c_1 \alpha h'(0) \text{ (Armijo condition), and} \\ |h'(\alpha)| &\leq c_2 |h'(0)| \text{ (Curvature condition),} \end{aligned}$$

for any $0 < c_1 < c_2 < 1$.

Let α denote the step size satisfying either the weak Wolfe conditions or the strong Wolfe condition.

- Sufficient descent condition by its definition
- Curvature condition implies that the step size can not be too small.

Theorem 4 (Existence of step size satisfying the conditions)

Suppose $f \in C^1$. Let p_k be a descent direction at x_k, and assume f is bounded from below along the ray $\{x_k + \alpha p_k : \alpha > 0\}$. Then if $0 < c_1 < c_2 < 1$ and $0 < \tau_1 \leq \tau_2 < 1$, there exists a step length satisfying the Armijo-Goldstein condition, a step size satisfying the weak Wolfe conditions and a step size satisfying the strong Wolfe conditions.

Definition 5 (Byrd-Nocedal conditions in [BN89])

The step size α satisfies

$$h(\alpha) - h(0) \leq -\chi_1 \frac{h'(0)^2}{\|p\|^2}$$

or

$$h(\alpha) - h(0) \leq \chi_2 h'(0),$$

for some values of $\chi_1, \chi_2 \in (0, 1)$.

Theorem 6

Let $\mathcal{N}_0$ denote $\{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$. If the gradient of f is Lipschitz continuous, i.e., $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$ for all $x, y \in \mathcal{N}_0$, then Byrd-Nocedal conditions is implied by Armijo-Goldstein condition or Wolfe conditions.

Theorem 6

Let $\mathcal{N}_0$ denote $\{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$. If the gradient of f is Lipschitz continuous, i.e., $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$ for all $x, y \in \mathcal{N}_0$, then Byrd-Nocedal conditions is implied by Armijo-Goldstein condition or Wolfe conditions.

Note that χ_1 and χ_2 are typically chosen to be smaller than β/L, where β is a constant and L is the Lipschitz constant of $\nabla f(x)$.

Conditions for Step Size

Line search conditions summary

- The Armijo-Goldstein condition and weak/strong Wolfe conditions are easy to use
  - c_1, c_2, τ_1, and τ_2 can be any positive values satisfying $0 < c_1 < c_2 < 1$ and $0 < \tau_1 \leq \tau_2 < 1$;
- The Byrd-Nocedal conditions are useful in theorem but not in implementation
  - Not easy to use: χ_1 depends on the Lipschitz constant of ∇f
  - Zoutendijk's condition;

Theorem 7 (Zoutendijk's condition with slight modification)

Consider the line search algorithm below. Suppose $f \in C^1$ is bounded from below, α_k satisfies the Byrd-Nocedal conditions, and $\|p_k\| \geq \mu\|\nabla f(x_k)\|$ for all $k > 0$ and a constant $\mu > 0$. Then

$$\textstyle\sum_{k\geq 0} \cos^2\theta_k \|\nabla f(x_k)\|^2 < \infty.$$

Note: "$\|p_k\| \geq \mu\|\nabla f(x_k)\|$" is not required if the Wolfe conditions are used.

A line search algorithm

Input: Initial iterate x_0;

  Initial descent search direction p_0 at x_0 and set $k \leftarrow 0$;

  while not accurate enough **do**

    Set $x_{k+1} \leftarrow x_k + \alpha_k p_k$ with an appropriate step size α_k;

    Set p_{k+1} to be a descent direction at x_{k+1};

    $k \leftarrow k + 1$;

  end while

Step Size Selection Algorithms

Polynomial interpolation based backtracking

- Quadratic polynomial
  - Three conditions: $h(0)$, $h'(0)$, and $h(\alpha_1)$;
  - Minimizer of the quadratic polynomial:

$$\alpha_+ = \frac{-h'(0)\alpha_1^2}{2(h(\alpha_1) - h(0) - h'(0)\alpha_1)} \tag{1}$$

- Cubic polynomial
  - Four conditions: $h(0)$, $h'(0)$, $h(\alpha_1)$, and $h(\alpha_2)$;
  - Minimizer of the cubic polynomial:

$$\alpha_+ = \frac{-b + \sqrt{b^2 - 3ah'(0)}}{3a}, \tag{2}$$

  where

$$\begin{bmatrix} a \\ b \end{bmatrix} = \frac{1}{\alpha_1 - \alpha_2} \begin{bmatrix} \frac{1}{\alpha_1^2} & \frac{-1}{\alpha_2^2} \\ \frac{-\alpha_2}{\alpha_1^2} & \frac{\alpha_1}{\alpha_2^2} \end{bmatrix} \begin{bmatrix} h(\alpha_1) - h(0) - h'(0)\alpha_1 \\ h(\alpha_2) - h(0) - h'(0)\alpha_2 \end{bmatrix}.$$

- Finally, set $\alpha = \min(\max(\alpha_+, \tau_1\alpha_1), \tau_2\alpha_1)$

Polynomial interpolation based backtracking algorithm

Input: Initial step size $\alpha^{(0)}$; $0 < \tau_1 \leq \tau_2 < 1$; $h(0)$ and $h'(0)$;

Output: Step size α_* satisfying the Armijo-Goldstein condition

Set $i \leftarrow 0$;

loop

  if $h(\alpha^{(i)}) \leq h(0) + c_1\alpha^{(i)}h'(0)$ **then**

    $\alpha_* \leftarrow \alpha^{(i)}$ and return;

  end if

  if $h(\alpha^{(i)}) > h(0) + c_1\alpha^{(i)}h'(0)$ and $i = 0$ **then**

    Compute $\tilde{\alpha}$ by (1) with $\alpha_1 = \alpha^{(i)}$;

  end if

  if $h(\alpha^{(i)}) > h(0) + c_1\alpha^{(i)}h'(0)$ and $i > 0$ **then**

    Compute $\tilde{\alpha}$ by (2) with $\alpha_1 = \alpha^{(i)}$ and $\alpha_2 = \alpha^{(i-1)}$;

  end if

  $\alpha^{(i+1)} = \min(\max(\tilde{\alpha}, \tau_1\alpha^{(i)}), \tau_2\alpha^{(i)})$;

  $i \leftarrow i + 1$;

end loop

Summary for Step Size Selections

- Other conditions
  - the Curry-Altman condition:

$$\alpha = \min\{\alpha > 0 : h'(\alpha) = \mu h'(0)\}, \quad \mu \in [0, 1);$$

  - the Goldstein condition

$$h(0) + (1-c)\alpha h'(0) \leq h(\alpha) \leq h(0) + c\alpha h'(0), \quad c \in (0, 0.5);$$

  - Both imply the Byrd Nocedal conditions
  - etc
- Algorithms
  - Polynomial interpolation based algorithm [DS83, Algorithm A6.3.1mod] for the weak Wolfe conditions and [NW06] for the strong Wolfe conditions
  - etc
- The conditions and step size selection algorithms can be used for other line search based methods

Steepest Descent Methods

A steepest descent method

Input: Initial iterate x_0, initial step size $\alpha^{(0)}$;
Set $k \leftarrow 0$;
while not accurate enough **do**
Find a step size satisfying the Byrd-Nocedal conditions with initial step size $\alpha^{(0)}$;
Set $x_{k+1} \leftarrow x_k - \alpha_k \nabla f(x_k)$;
$k \leftarrow k + 1$;
end while

- Global convergence follows from the Zoutendijk's condition
- Convergence rate?
- Initial step size?

Local Convergence Rate Analysis of the Steepest Descent Method

Theorem 8 (R-linear local convergence rate analysis)

Let $\mathcal{N}_{x_0} = \{x : f(x) \le f(x_0)\}$. Suppose $f \in C^2$, $\mathcal{N}_{x_0}$ is convex, and there exists positive constants $0 < m \le M$ such that

$$m \le \lambda_{\min}(\nabla^2 f(x)) \le \lambda_{\max}(\nabla^2 f(x)) \le M$$

for all $x \in \mathcal{N}_{x_0}$, where $\lambda_{\min}(A)$, $\lambda_{\max}(A)$ denote the smallest and largest eigenvalues of A respectively. Let x^* denote the unique minimizer of f in $\mathcal{N}_{x_0}$ and $\{x_k\}$ denote the iterates generated by the steepest descent method with the Byrd Nocedal conditions. Then we have

$$f(x_k) - f(x^*) \le \left(1 - \beta \frac{m}{M}\right)^k (f(x_0) - f(x^*))$$

where β is a positive constant.

Local Convergence Rate Analysis of the Steepest Descent Method

Theorem 9 (Sublinear local convergence rate analysis)

Let $\mathcal{N}_{x_0} = \{x : f(x) \leq f(x_0)\}$. Suppose $f \in C^2$, $\mathcal{N}_{x_0}$ is convex, and there exists positive constants $M > 0$ such that

$$0 \leq \lambda_{\min}(\nabla^2 f(x)) \leq \lambda_{\max}(\nabla^2 f(x)) \leq M$$

for all $x \in \mathcal{N}_{x_0}$, where $\lambda_{\min}(A)$, $\lambda_{\max}(A)$ denote the smallest and largest eigenvalues of A respectively. Let x^* denote a minimizer of f in $\mathcal{N}_{x_0}$ and $\{x_k\}$ denote the iterates generated by the steepest descent method with the Armijo-Goldstein condition with $c_1 = 0.5$ or the Wolfe conditions with $c_1 = 0.5$. Then we have

$$f(x_k) - f(x^*) \leq \frac{\beta \|x_0 - x^*\|^2}{2k},$$

where β is a positive constant.

Initial Step Size Selection

Methods in [NW06]

- Assume that ratio of the consecutive step sizes is proportional to the ratio of the consecutive first order values, i.e.,

$$\frac{\alpha_k^{(0)}}{\alpha_{k-1}} = \frac{\nabla f(x_{k-1})^T p_{k-1}}{\nabla f(x_k)^T p_k} \implies \alpha_k^{(0)} = \alpha_{k-1} \frac{\nabla f(x_{k-1})^T p_{k-1}}{\nabla f(x_k)^T p_k}$$

- Assume that $\alpha_k^{(0)}$ is the minimizer of the quadratic polynomial q that interpolate $q(0) = h_{k-1}(0)$, $q'(0) = h'_{k-1}(0)$ and $q(\alpha_k^{(0)}) = h_k(0)$, i.e.,

$$\alpha_k^{(0)} = \frac{-h'_{k-1}(0)(\alpha_k^{(0)})^2}{2(h_k(0) - h_{k-1}(0) - h'_{k-1}(0)\alpha_k^{(0)})}$$

$\implies$

$$\alpha_k^{(0)} = \frac{2(h_k(0) - h_{k-1}(0))}{h'_{k-1}(0)}.$$

Consider

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}x^T Ax,$$

where A is a symmetric positive definite matrix. Let λ_i and v_i denote the eigenvalues and corresponding eigenvectors of A respectively.

- Gradient $\nabla f(x) = Ax$, let g_k denote $\nabla f(x_k)$
- $x_k = x_{k-1} - \alpha_{k-1}g_{k-1} \Rightarrow g_k = g_{k-1} - \alpha_{k-1}Ag_{k-1}$
- Let $g_k = \sum_{i=1}^n \mu_{k,i}v_i$
- $\mu_{k,i} = (1-\alpha_{k-1}\lambda_i)\mu_{k-1,i} = \ldots = \left(\prod_{j=0}^{k-1}(1-\alpha_j\lambda_i)\right)\mu_{0,i}$
- If $\alpha_j = 1/\lambda_i$, then $\mu_{k,i} = 0$ for all $k \geq j$
- If $\alpha_k < 1/\max_i(\lambda_i)$ for all k, then $\mu_{k,i}$ is decreasing as $k \to \infty$ for all i.

Initial Step Size Selection

Convex quadratic problems

- Ideally, *n*-steps terminates at the exact solution
- Goal: estimate reciprocal of eigenvalues to get step size
  - Step size by limited memory [Fle12]
  - Barzilai-Borwein (BB) step size [BB88]

Barzilai-Borwein (BB) step size

- Define $s_{k-1} = x_k - x_{k-1}$ and $y_{k-1} = g_k - g_{k-1} = As_{k-1}$
- BB1: $\alpha_k^{\mathrm{BB1}} = \frac{s_{k-1}^T s_{k-1}}{s_{k-1}^T A s_{k-1}} = \frac{s_{k-1}^T s_{k-1}}{s_{k-1}^T y_{k-1}}$
- BB1: exact step size at iteration $k-1$
- BB2: $\alpha_k^{\mathrm{BB2}} = \frac{s_{k-1}^T A s_{k-1}}{s_{k-1}^T A^2 s_{k-1}} = \frac{s_{k-1}^T y_{k-1}}{y_{k-1}^T y_{k-1}}$
- BB2: minimize gradient at iteration $k-1$
- $\frac{1}{\lambda_{\max}} \le \alpha_k^{\mathrm{BB2}} \le \alpha_k^{\mathrm{BB1}} \le \frac{1}{\lambda_{\min}}$
- Empirically, BB step sizes tend to sweeping the reciprocals of the spectrum of A.
- Adaptive BB variants, e.g., [ZGD06]

Initial Step Size Selection

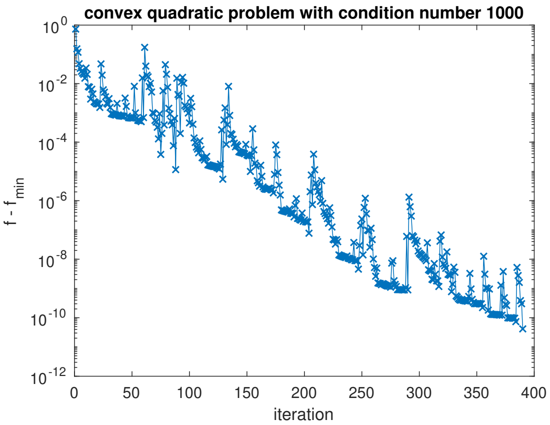
Nonlinear problems

- Step size by limited memory
- BB step size
  - Not always positive
  - Wolfe $\Rightarrow$ positivity
  - Safeguard with methods in [NW06]

Initial Step Size Selection

Preference for good initial step size

Convex quadratic problems



The function value is not monotonically descent but the iterates converges to the minimizer.

Algorithm 1 Nonmonotonic line search with BB step size

Input: Initial iterate x_0, a positive integer $m > 0$, $\rho \in (0, 1)$;

1: **for** $k = 0, 1, 2, \ldots$ **do**
2: Set step size $\alpha_k = \alpha_k^{(0)}$;
3: **while** $f(x_k - \alpha_k \nabla f(x_k)) > \max(f(x_k), f(x_{k-1}), \ldots, f(x_{k-m+1})) + c_1 \alpha_k h_k'(0)$ **do**
4: $\alpha_k \leftarrow \rho \alpha_k$;
5: **end while**
6: $x_{k+1} \leftarrow x_k - \alpha_k \nabla f(x_k)$;
7: $k \leftarrow k + 1$;
8: **end for**

References I

J. Barzilai and J. M. Borwein.
Two-Point Step Size Gradient Methods.
IMA Journal of Numerical Analysis, 8:141–148, 1988.

R. H. Byrd and J. Nocedal.
A tool for the analysis of quasi-Newton methods with application to unconstrained minimization.
SIAM Journal on Numerical Analysis, 26(3):727–739, 1989.

J. E. Dennis and R. B. Schnabel.
Numerical methods for unconstrained optimization and nonlinear equations.
Springer, New Jersey, 1983.

Roger Fletcher.
A limited memory steepest descent method.
Mathematical Programming, 135(1):413–436, Oct 2012.

J. Nocedal and S. J. Wright.
Numerical Optimization.
Springer, second edition, 2006.

Bin Zhou, Li Gao, and Yu-Hong Dai.
Gradient methods with adaptive step-sizes.
Computational Optimization and Applications, 35(1):69–86, Sep 2006.