Optimization Theory

Piotr Więcek,
Faculty of Pure and Applied Mathematics,
Wrocław University of Science and Technology

4.11.2020

Gradient methods: Part III

Comparison of the steepest descent and the Newton methods

Advantages of the Newton method:

- Much faster than the steepest descent, especially close to the stationary point.

Advantages of the steepest descent:

- Much weaker conditions necessary for it to be convergent.
- Newton's method requires computation of $\left(\nabla^2 f(x)\right)^{-1}$. We may not have the values of second-order derivatives. Even if we do, computation of the inverse of a matrix may be both very time consuming and inaccurate if their size is large.

Goal of alternative methods:

To obtain convergence almost as fast as that of Newton's method using much simpler computations.

Types of alternative methods

Quasi-Newton methods:

- We construct successive D_k matrices in a **recursive manner**;
- D_{k+1} depending only on D_k, $x_{k+1} - x_k$ and $\nabla f(x_{k+1}) - \nabla f(x_k)$, no matrix inversion required;
- The construction should result in D_k **close to** $\left(\nabla^2 f(x_k)\right)^{-1}$ **when** $x_{k+1} - x_k$ **is small** and **positive definite** at every step.

Conjugate directions methods:

- Instead of one d_k depending only on x_k, **we construct** d_k**s in sequences** – d_k depending on $\nabla f(x_k)$ and a given number of previous directions $d_{k-1}, \ldots, d_{k-m}$; Typically we take d_k in such a way that it is orthogonal to these previous directions;
- Each step in the sequence requireing simple computations;
- A given **fixed number of successive steps should mimic the behaviour of one step of Newton's method**.

Quasi-Newton methods I

General form of quasi-Newton methods:

$D := I$, $x := x_0$
$g := \nabla f(x)$
for $i = 0, 1, 2, \ldots$
  $d := -Dg$
  find α^* minimizing $f(x + \alpha d)$ using line search
  $p := \alpha^* d$, $x := x + p$
  $q := g$, $g := \nabla f(x)$, $q := g - q$
  compute C based on D, p and q
  $D := D + C$

Quasi-Newton methods II

Different quasi-Newton methods differ by the correction matrix C.

Requirements regarding matrix C:

- If D is positive definite $D + C$ also should.
- If f is a convex quadratic function (and thus $\nabla^2 f(x) \equiv Q$ for some symmetric positive definite matrix Q), then after n steps of the method D should become Q^{-1}.
- Moreover, the subsequent directions chosen by the method form a conjugate set with respect to the scalar product defined by matrix Q, that is $d_k^T Q d_l = 0$ for $k \neq l$.

Quasi-Newton methods III

Popular quasi-Newton methods:

1. Davidon-Fletcher-Powell method:

$$C = \frac{pp^T}{p^T q} - \frac{Dqq^T D}{q^T Dq}$$

2. Broyden-Fletcher-Goldfarb-Shanno method

$$C = \frac{pp^T}{p^T q} - \frac{Dqq^T D}{q^T Dq} + rvv^T$$

with $v = \frac{p}{p^T q} - \frac{Dq}{r}$ and $r = q^T Dq$

3. Hoshino method:

$$C = \Theta pp^T - \Psi(pq^T D + Dqp^T + Dqq^T D)$$

with $\Theta = \frac{q^T p + 2q^T Dq}{q^T p(q^T p + q^T Dq)}$ and $\Psi = \frac{1}{q^T p + q^T Dq}$

Quasi-Newton methods IV

4. Broyden method

$$C = \frac{pp^T}{p^T q} - \frac{Dqq^T D}{q^T Dq} + \Phi rvv^T$$

with $v = \frac{p}{p^T q} - \frac{Dq}{r}$ and $r = q^T Dq$,
where $\Phi \in [0, 1]$ is the parameter of the method.
Alternatively this can be written as
$C_B = (1 - \Phi)C_{DFP} + \Phi C_{BFGS}$.
All the methods presented above are in the Broyden family.

5. Fletcher (switch) method:
It is a specific version of Broyden method where Φ is switched between 0 and 1, depending on a test checking which of the DFP and BFGS formulas better approximates the direction given by the inverse Hessian.

Conjugate gradient method I

Suppose we want to find a minimum of a convex quadratic function $f(x) = \frac{1}{2}x^T Qx + b^T x$. Then we can use the algorithm:

The idea of the method:

- $d_0 = -\nabla f(x_0)$. The next directions d_1, d_2,..., d_{n-1} are chosen in such a way that they form the orthogonal basis for $\mathbb{R}^n$.
- For each chosen direction α is chosen using minimization rule.
- It can be proved that after at most n iterations we shall reach the minimum of the function.

Implementation:

$d := -(Qx + b)$, $g := d$

$\alpha := \frac{g^T g}{d^T Qd}$, $x := x + \alpha d$

for $k = 1, \ldots, n-1$

$g_0 := g$, $d_0 := d$, $g := -(Qx + b)$

$\beta := \frac{g^T g}{g_0^T g_0}$, $d := g + \beta d_0$

$\alpha := \frac{g^T g}{d^T Qd}$, $x := x + \alpha d$

Conjugate gradient method II

The idea how to generalize the method for any function comes from Fletcher and Reeves:

- Perform n steps as in the original conjugate gradient method. After that, repeat the same sequence of steps from the new point starting with one step of steepest descent. Repeat such sequences until convergence.

Implementation:

for $i = 0, 1, \ldots$

$d := -\nabla f(x_{in})$, $g := d$

choose α^* minimizing $f(x_{in} + \alpha d)$, $x_{in+1} := x_{in} + \alpha^* d$

for $k = 1, \ldots, n-1$

$g_0 := g$, $d_0 := d$, $g := -\nabla f(x_{in+k})$

$\beta := \frac{g^T g}{g_0^T g_0}$

$d := g + \beta d_0$

choose α^* minimizing $f(x_{in+k} + \alpha d)$, $x_{in+k+1} := x_{in+k} + \alpha^* d$

Other conjugate directions methods

Powell's method:

- At the beginning of the algorithm n line searches are performed successively along each axis in $\mathbb{R}^n$ (that is, with $d_1 = \pm e_1$, $d_2 = \pm e_2, \ldots$, $d_n = \pm e_n$). This gives point x_n and a new direction $d_{n+1} = x_n - x_0$.
- The same procedure is repeated from x_n using n directions $d_2, \ldots d_n$, d_{n+1}. This gives the point x_{2n} and the direction $d_{n+2} = x_{2n} - x_n$.
- The procedure is repeated – after each iteration one new direction replaces one old one.

Partan method:

- At the beginning of the algorithm two steps of the steepest descent with optimal stepsize are performed giving x_2 and $d_3 = x_2 - x_0$.
- In each subsequent iteration, the new direction is chosen based on last two iterations: $d_k = x_{k-1} - x_{k-3}$ and the stepsize is chosen optimally.

Recapitulation

Basic methods of both groups are widely used in optimization software to solve large-scale problems.

- The most commonly used are the Fletcher-Reeves conjugate gradient method and two basic quasi-Newton methods: Davidon-Fletcher-Powell and Broyden-Fletcher-Goldfarb-Shanno methods.
- They do not require matrix inversion nor the Hessian.
- It can be proved that their convergence is superlinear.
- Even though they typically require significantly more iterations than the Newton method to obtain the same quality of approximation, they outperform it in terms of time already for problems in $\mathbb{R}^n$ when n is around 10.