Numerical Optimisation: Quasi-Newton methods

Department of Computer Sc
Centre for Medical Image Com
Centre for Inverse Proble
University College London

Lecture 7 & 8

- First idea by William C. Davidon in mid 1950, who was frustrated by performance of coordinate descent.
- Quickly picked up by Fletcher and Powell who demonstrated that the new algorithm was much faster and more reliable then existing methods.

-

- Like steepest gradient, Quasi Newton me
the gradient of the objective function at eac
Measuring changes in gradient they build a
objective function which is good enough to p
superlinear convergence.

- As the Hessian is not required, Quasi-Newton methods can be more efficient than Newton methods which take a long time to evaluate the Hessian and solve for the Newton direction.

Quadratic model of the objective function at x_k:

$$m_k(p) = f_k + \nabla f_k^{\mathrm{T}} p + \frac{1}{2} p^{\mathrm{T}} B_k p,$$

where $B_k \in \mathbb{R}^{n \times n}$ symmetric positive definite which will be upda

The m

$$k$$

p_k is used as a search direction and the next iterate b

$$x_{k+1} = x_k + \alpha$$

The step length α_k is chosen to satisfy the Wolfe conditions.

The iteration is similar to the line search Newton with the key difference that the Hessian B_k is an approximation.

B_k update

Davidon proposed to update B_k in each iteration instead of computing it anew.

Question: When we computed the new iterate x_{k+1} and construct the new model

$$m_{k+1}(p) = f_{k+1} + \nabla f_{k+1}^{\mathrm{T}} p + \frac{1}{2} p^{\mathrm{T}} B_{k+1} p,$$

what r
kno

Require: gradient of m_{k+1} should matc
last two iterates x_k, x_{k+1}

i) At x_{k+1}: $p_{k+1} = 0$,
$\nabla m_{k+1}(0) = \nabla f_{k+1}$ is satisfied automatically.

ii) At $x_k = x_{k+1} - \alpha_k p_k$:

$$\nabla m_{k+1}(-\alpha_k p_k) = \nabla f_{k+1} - \alpha_k B_{k+1} p_k = \nabla f_k.$$

By rearranging ii) we obtain

$$B_{k+1}\alpha_k p_k = \nabla f_{k+1} - \nabla f_k.$$

Define vectors

$$s_k = x_{k+1} - x_k = \alpha_k p_k, \quad y_k = \nabla f_{k+1} - \nabla f_k,$$

ii) bec

As B_{k+1} is symmetric positive definite, this is o
curvature condition holds

$$s_k^{\mathrm{T}} y_k > 0,$$

which can be easily seen multiplying the secant equation by s_k^{T} from the left.

If f is strongly convex $s_k^{\mathrm{T}} y_k > 0$ is satisfied for any x_k, x_{k+1}. However, for nonconvex functions in general this condition will have to be enforced explicitly by imposing restrictions on the line search.

$s_k^{\mathrm{T}} y_k$

cond

Fro k $k+1$ 2 k k 1 $c_2 < 1$ it follows

$$s_k^{\mathrm{T}} y_k \geq (c_2 - 1)\alpha_k p$$

since $c_2 < 1$ and p_k is a descent direction, a condition holds.

Davidon Flecher Powell (DFP)

When $s_k^{\mathrm{T}} y_k > 0$, the secant equation always has a solution B_{k+1}.
In fact the secant equation is heavily underdetermined: a symmetric matrix has $n(n+1)/2$ dofs, secant equation: n conditions, positive definiteness: n inequalities.

Extra conditions to obtain unique solutions: we look for B_{k+1} close to B

DFP u

$$B_{k+1} = (I - \rho_k y_k s_k^{\mathrm{T}}) B_k (I - \rho_k s_k y_k^{\mathrm{T}}) + \rho_k y_k y_k^{\mathrm{T}} \qquad \text{(DFP B)}$$

with $\rho_k = 1/y_k^{\mathrm{T}} s_k$.

The inverse $H_k = B_k^{-1}$ can be obtained wit
Sherman-Morrison-Woodbury formula

$$H_{k+1} = H_k - \frac{H_k y_k y_k^{\mathrm{T}} H_k}{y_k^{\mathrm{T}} H_k y_k} + \frac{s_k s_k^{\mathrm{T}}}{y_k^{\mathrm{T}} s_k}. \qquad \text{(DFP H)}$$

Figure: Nocedal Wright (

Applying the same argument directly to the inverse of the Hessian H_k. The updated approximation H_{k+1} must be symmetric and positive definite and must satisfy the secant equation

$$H_{k+1} y_k = s_k.$$

BFG

$k+1$ k k k k k k k k k k (BFGS)

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

How to choose H_0? Depends on the situation
the problem e.g. start with an inverse of an approximated Hessian calculated by a finite difference at x_0. Otherwise, we can set H_0 to identity or diagonal matrix to reflect the scaling of the variables.

Given x_0, inverse Hessian approximation H_0, tolerance $\varepsilon > 0$
Set $k = 0$
while $\|\nabla f_k\| > \varepsilon$ **do**
  Compute search direction
  $x_{k+1} = x_k + \alpha_k p_k$ where α_k is computed with a line search procedure satisfying Wolfe condition
  Define $s_k = x_{k+1} - x_k$ and $y_k =$
  Compute H_{k+1} using (BFGS)
  $k = k + 1$
end while

- Complexity of each iteration is $\mathcal{O}(n^2)$ plus the cost of function and gradient evaluations.
- There are no $\mathcal{O}(n^3)$ operations such as linear system solves or matrix-matrix multiplications.
- The algorithm is robust and the rate of convergence is superlinear. In many cases it outperforms Netwon method, which while converging quadratically, has higher complexity
- [illegible] ther

Sherman-Morrison-Woodbury form

$$B_{k+1} = B_k - \frac{B_k s_k s_k^{\mathrm{T}} B_k}{s_k^{\mathrm{T}} B_k s_k} \text{[illegible]}$$

An $\mathcal{O}(n^2)$ implementation can be achieved based on updates of LDL^{T} factors of B_k (with possible diagonal modification for stability) but no computational advantage is observed on above algorithm using (BFGS) to update H_k.

- The positive definiteness of H_k is not explicitly forced, but if H_k is positive definite so will be H_{k+1}.
- What happens if at some iteration H_k becomes as poor approximation to the true inverse Hessian e.g. if $s_k^{\mathrm{T}} y_k$ is tiny (positive) than the elements of H_{k+1} get very large.
  It turns out that BFGS has effective self correcting properties,

  sampled at points which allow the model curvature information.
- On the other hand DFP method is less effectiv

  itself.
- DFP and BFGS are dual in the sense that they can be obtained by switching $s \leftrightarrow y, B \leftrightarrow H$.

- $\alpha_k = 1$ should always be tried first, because this step length

-

  line search.

- $c_1 = 10^{-4}$, $c_2 = 0.9$ are commonly use

Heuristic for scaling H_0

Choice $H_0 = \beta I$ is popular, but there is no good strategy for estimating β.

If β is too large, the first step $p_0 = -\beta g_0$ is too long and line search may require many iterations to find a suitable step length α_0.

Heur

H_0

upda

$H_0 = \frac{s_k^{\mathrm{T}} y_k}{y_k^{\mathrm{T}} y_k} I$. This scaling attempts to approxi

eigenvalue of the inverse Hessian: from Taylor th

$$y_k = \bar{G}_k \alpha_k p_k =$$

we have that the secant equation is satisfied for average Hessian

$$\bar{G}_k = \int_0^1 \nabla^2 f(x_k + \tau \alpha_k p_k) d\tau.$$

Symmetric rank-1 (SR-1) update

Both BFGS and DFP methods perform a rank-2 update while preserving symmetry and positive definiteness.

Question: Does a rank-1 update exist such that the secant equation is satisfied and the symmetry and definiteness are preserved?

Ran

and v is chosen such that B_{k+1} satisfi

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

Substituting the explicit rank-1 form into the secant equation

$$y_k = B_k s_k + \underbrace{\left(\sigma v^{\mathrm{T}} s_k\right)}_{:=\delta^{-1},\ \delta \neq 0} v$$

we see that v must be of the form $v = \delta(y_k - B_k s_k)$.

Substituting $v = \delta(y_k - B_k s_k)$ back into the secant equation we obtain

$$y_k - B_k s_k = \sigma\delta^2[s_k^{\mathrm{T}}(y_k - B_k s_k)](y_k - B_k s_k)$$

which is satisfied if and only if

$$\sigma = \mathrm{sign}[s_k^{\mathrm{T}}(y_k - B_k s_k)], \quad \delta = \pm|s_k^{\mathrm{T}}(y_k - B_k s_k)|^{-1/2}$$

Hence, the only symmetric rank-1 update satisfying the secant equa

(SR-1)

Applying the Sherman-Morrison-Woodbur

inverse Hessian update

$$H_{k+1} = H_k + \frac{(s_k - H_k y_k)(s}{(s_k - H_k y_k)^{\mathrm{T}} y_k}. \qquad \text{(SR-1)}$$

SR-1 update does not preserve the positive definiteness. It is a drawback for line search methods but could be an asset for trust region as it allows to generate indefinite Hessians.

SR-1 breakdown

The main drawback of SR-1 is that $(y_k - B_k s_k)^{\mathrm{T}} s_k$ (same for H_k) can become 0 even for a convex quadratic function i.e. there may be steps where the is no symmetric rank-1 update which satisfies the secant equation.

Three cases:

-
-
- $(y_k - B_k s_k)^{\mathrm{T}} s_k = 0$ and $y_k \neq B_k$
  rank-1 update satisfying secant equation

Remedy: Skipping i.e. apply update only if

$$|(y_k - B_k s_k)^{\mathrm{T}} s_k| \geq r \|s_k\| \|y_k - B_k s_k\|,$$

where $r \in (0, 1)$ is a small number (typically $r = 10^{-8}$), otherwise set $B_{k+1} = B_k$.

- This simple safeguard adequately prevents the breakdown. Recall: for BFGS update skipping is not recommended if the curvature condition $s_k^{\mathrm{T}} y_k > 0$ fails. Because it can occur often by e.g. taking too small step if the line search does not impose the Wolfe conditions. For SR-1 $s_k^{\mathrm{T}}(y_k - B_k s_k) \approx 0$

  s_k and

  $_k s_k$,

  k

  already correct.

- The Hessian approximations generated

  often better than those by BFGS.

- When the curvature condition $y_k^{\mathrm{T}} s_k > 0$ cannot be imposed e.g. constraint problems or partially separable functions, where indefinite Hessian approximations are desirable as they reflect the indefiniteness of the true Hessian.

Given x_0, B_0, Δ, $\eta \in (0, 10^{-3})$, $r \in (0,1)$ and $\varepsilon > 0$
Set $k = 0$
while $\|\nabla f_k\| > \varepsilon$ **do**
  $s_k = \arg\min_s s^{\mathrm{T}} \nabla f_k + \frac{1}{2} s^{\mathrm{T}} B_k s$, subject to $\|s\| \le \Delta_k$
  $y_k = \nabla f(x_k + s_k) - \nabla f_k$
  $\rho_k = (f_k - f(x_k + s_k)) / -(s_k^{\mathrm{T}} \nabla f_k + \frac{1}{2} s_k^{\mathrm{T}} B_k s_k)$
  if $\rho > \eta$ **then**

  end if
  Update Δ_k in dependence of ρ_k, $\|s$
  if $|(y_k - B_k s_k)^{\mathrm{T}} s_k| \ge r \|s_k\| \|y_k - B$
    Update B_{k+1} using (SR-1) (even if
    approximation along s_k)
  else
    $B_{k+1} = B_k$
  end if
  $k = k + 1$
end while

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a strongly quadratic function $f(x) = b^{\mathrm{T}}x + \frac{1}{2}x^{\mathrm{T}}Ax$ with A symmetric positive definite. For any starting point x_0 and any symmetric initial matrix H_0, the iterates

$$x_{k+1} = x_k + p_k, \quad p_k = -H_k \nabla f_k,$$

wher

most ter n

steps

$H_n = A^{-1}$.

Proof idea Show by induction that the secant e $= s_j$

is satisfied for all $j = 1, \dots, k-1$ i.e. $k-1$). Use that for such quadratic function it holds $y_i = As_i$.

For SR-1 $H_k y_j = s_j$, $j = 1, \dots, k-1$ holds regardless how the line search is performed. In contrast for BFGS, it can only be shown under the assumption that the line search is exact.

Theorem: Hessian approximation for general function

Let $f : \mathbb{R}^n \to \mathbb{R}$ twice continuously differentiable with the Hessian bounded and Lipschitz continuous in a neighbourhood of a point $x^\star \in \mathbb{R}^n$ and $\{x_k\}$ a sequence of iterates such that $x_k \to x^\star$.

Suppose that

hold

re

uniformly independent (steps do not tend to fall in a s

dimension less than n).

Then the matrices B_k generated by the up

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

Broyden class is a family of updates of the form

$$B_{k+1} = B_k - \frac{B_k s_k s_k^{\mathrm{T}} B_k}{s_k^{\mathrm{T}} B_k s_k^{\mathrm{T}}} + \frac{y_k y_k^{\mathrm{T}}}{y_k^{\mathrm{T}} s_k} + \tau_k (s_k^{\mathrm{T}} B_k s_k) v_k v_k^{\mathrm{T}}, \quad \text{(Broyden)}$$

where τ_k is a scalar parameter and

For

Hence we can write (Broyden) as a linear combina

$$B_{k+1} = (1 - \tau_k) B_{k+1}^{\mathrm{BFG}}$$

Since both BFGS and DFP satisfy secant equation so does the whole Broyden class.
Since BFSG and DFP updates preserve positive definiteness of the Hessian when $s_k^{\mathrm{T}} y_k > 0$, so does the **restricted Broyden class** which is obtained by restricting $0 \leq \tau_k \leq 1$.

Theorem: monotonicity of eigenvalue approximation

Let $f : \mathbb{R}^n \to \mathbb{R}$ is the strongly convex quadratic function $f(x) = b^{\mathrm{T}}x + \frac{1}{2}x^{\mathrm{T}}Ax$ with A symmetric positive definite. Let B_0 any symmetric positive matrix and x_0 be any starting point for the iteration

$$x_{k+1} = x_k + p_k, \quad p_k = -B_k^{-1}\nabla f_k,$$

wher

Den

Then for all k, we have

$$\min\{\lambda_i^k, 1\} \le \lambda_i^{k+1} \le \max\{\lambda_i^k,$$

The interlacing property does not hold if $_k$ /

Consequence: The eigenvalues λ_i^k converge monotonically (but not strictly monotonically) to 1, which are the eigenvalues when $B_k = A$. Significantly, the result holds even if the line search is not exact.

So do the best updates belong to the restricted Broyden class?

We recover SR-1 formula for

$$\tau_k = \frac{s_k^{\mathrm{T}} y_k}{s_k^{\mathrm{T}} y_k - s_k^{\mathrm{T}} B_k s_k},$$

which does not belong to the restricted Broyden class as τ_k may fall outside of $[0, 1]$.

It can b [illegible] k
$s_k^{\mathrm{T}} y_k$ [illegible] ain
symmetric and positive definite. Here

$$\tau_k^c = (1 - \mu_k)^{-1} \leq 0, \quad \mu_k = \frac{(y \ldots}{\ldots}$$

When the line search is exact all the methods in the Broyden class with $\tau_k \geq \tau_k^c$ generate the same sequence of iterates, even for nonlinear functions because the directions differ only by length and this is compensated by the exact line search.

Let $f : \mathbb{R}^n \to \mathbb{R}$ be the strongly convex quadratic function $f(x) = b^{\mathrm{T}} x + \frac{1}{2} x^{\mathrm{T}} A x$ with A symmetric positive definite. Let x_0 be any starting point and B_0 any symmetric positive definite matrix. Assume that α_k is the exact step length and $\tau_k \geq \tau_k^*$ for all k. Then it holds

(i) T

(ii) T

$$B_k s_j = y_j, \quad j = 1$$

(iii) If $B_0 = I$, then the sequence of iterates that generated by the conjugate gradient particular the search directions s_k are conjugate

$$s_i^{\mathrm{T}} A s_j = 0, \quad i \neq j.$$

(iv) If n iterations are performed, we have $B_n = A$.

Few comments ...

- The theorem can be slightly generalised to hold if the Hessian approximation remains nonsingular but not necessarily positive definite i.e. τ_k could be smaller than τ_k^c provided the chosen

-
  method with the preconditioner B .

- The theorem is mainly of theoretical intere
  line search used in practice significantly alt
  of the methods. This type of analysis howeve
  of the development in quasi-Newton methods.

For general nonlinear objective function, there is no global convergence result for quasi-Newton methods i.e. convergence to a stationary point from any starting point and any suitable Hessian approximation.

Theorem: [BFGS global convergence]

Let f be a
start $f(x_0)\}$
is con hat

$$m\|z\|^2 \leq z^{\mathrm{T}} \nabla^2 f(x) z \leq M\|z\|$$

Then for any symmetric positive definite matrix
$\{x_k\}$ generated by BFGS algorithm (with $\varepsilon = 0$) converges to the miminizer $x^\star$ of f.

This results can be generalised to the restricted Broyden class with $\tau_k \in [0, 1)$ i.e. except for DFP method.

Theorem: Superlinear local convergence of BFGS

Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable and the sequence of iterates generated by BFGS algorithm converge to $x^\star \in \mathbb{R}^n$ such that the Hessian $^2 f$ is Lipschitz continuous at $x^\star$

$$\|\nabla \quad < \infty,$$

and that it holds

$$\sum_{k=}^{\infty} \|x_k - x^\star\|$$

then x_k converges to $x^\star$ at a superlinear ra

Theorem: SR-1 trust region convergence

Let $\{x_k\}$ be the sequence of iterates generated by the SR-1 trust region method. Suppose the following conditions hold:

- the sequence $\{x_k\}$ does not terminate, but remains in a closed bounded convex set D on which f is twice continuously differentiable and in which f has a unique stationary point $x^\star$;
-
-
- $|(y_k - B_k s_k)^{\mathrm{T}} s_k| \geq r\|s_k\|\|y_k - B$

Then for the sequence $\{x_k\}$ we have lim

$$\lim_{k\to\infty} \frac{\|x_{k+n+1} - x^\star\|}{\|x_k - x^\star\|} = 0 \quad (n+1\text{-step superlinear rate}).$$

Remarks:

- SR-1 update does not maintain positive definiteness of B_k. In practice B_k can be indefinite at any iteration (trust region) but it

0

- The theorem does not require exact solutio region subproblem.