

“On Sizing and Shifting The BFGS Update Within The Sized-Broyden Family of Secant Updates”

Richard Tapia

(H. Yabe, H.J. Martinez, and R.A. Tapia
SIAM Journal on Optimization 2004)

Rice University



Department of Computational and Applied Mathematics
Center for Excellence and Equity in Education

Eighth US-Mexico Workshop on Optimization and its Applications

Huatulco, Oaxaca

January 11, 2007

Preliminaries

The Problem: $\min_x f(x)$ $f : \mathbb{R}^n \rightarrow \mathbb{R}$

“Equivalently”: $\nabla f(x) = 0$.

Time Honored Work-Horse Methods

(Cauchy 1847) Gradient Method (steepest descent)

$$x_+ = x - \alpha \nabla f(x) \quad \alpha > 0$$

(≈1660's) Newton's Method

$$x_+ = x + s$$

where

$$\nabla^2 f(x)s = -\nabla f(x) \quad (\text{Newton Equation})$$

Characteristics

□ Gradient Method:

- Inexpensive
- Good global properties
- Slow local convergence

□ Newton's Method

- Expensive $O(n^3)$ per iteration
- Poor global properties, excellent local properties
- Fast local convergence

The Middle Ground and The Algorithm of Interest

Secant Methods

$$x_+ = x + s$$

where

$$Bs = -\nabla f(x)$$

secant equation

$$B_+s = y \quad (B_+ \text{ new approximation})$$

where

$$y = \nabla f(x + s) - \nabla f(x)$$

Remark: We view B as an approximation of $\nabla^2 f(x)$.

Characteristics:

Similar properties as Newton's Method, but not as expensive,
 $O(n^2)$ per iteration

History/Chronology

- In one dimension ($n=1$) the secant equation uniquely gives

$$B_+ = \frac{f'(x_+) - f'(x)}{x_+ - x}$$

as an approximation to $f''(x_+)$. The 1-dimension 2-point secant method was probably discovered in the middle of the 18th century BC by Babylonians. It was called The Rule of Double False Position and was used to solve only the linear equation problems. It is extremely effective and efficient and has a convergence rate of $\frac{1+\sqrt{5}}{2}$ (Golden mean)

- Gauss formulated a 3-point secant method for two dimensions in 1847.
- There was considerable research activity on $(n+1)$ -point secant methods in the 1960's. While these methods had good theoretical properties they were numerical failures. The iterates tend to cluster in a lower dimensional manifold, and lead to linear systems that are ill-conditioned and nearly singular. These $(n+1)$ -point secant methods have been discarded.

The New Generation of Secant Methods (Variable Metric or Quasi-Newton Methods)

- DFP Davidon-Fletcher-Powell Davidon(1958)
Fletcher-Powell 1963
 - DFP was the work-horse secant method from 1963-1970 in spite of the serious numerical flaw that the diagonal of the approximating matrices approached zero (excessively small eigenvalues). This required restarts using the identity as a Hessian approximation.
- BFGS (1970) Broyden-Fletcher-Goldfarb-Shanno
 - A new secant update that does not generate excessively small eigenvalues
- BFGS has become the secant method of choice based on numerical performance
- In some cases BFGS is not effective and generates approximations with excessively large eigenvalues.

Broyden Family of Secant Updates (1970)

Write

$$BFGS(B, s, y) = B - \frac{Bss^T B}{s^T Bs} + \frac{yy^T}{y^T s}$$

Broyden Family

$$B_+ = BFGS(B, s, y) + \phi vv^T$$

where parameter $\phi \in \mathbb{R}$ and

$$v = \sqrt{s^T Bs} \left(\frac{y}{y^T s} - \frac{Bs}{s^T Bs} \right)$$

- 1963 DFP $\phi = 1$ promotes small eigenvalues
- 1970 BFGS $\phi = 0$ may promote large eigenvalues
- Convex class $\phi \in [0,1]$
- Preconvex class $\phi < 0$

Two Interesting Research Ideas That We Build On

□ John Dennis (1972)

Notion of least change secant update

Choose ϕ in the Broyden class so that B_+ is closest to B in a weighted Frobenius norm. In this case we can explain BFGS and DFP.

□ Oren-Luenberger (1974) (SSVM)

Size the matrix B before updating

$$B \rightarrow \gamma_{OL} B$$

where

$$\gamma_{OL} = \frac{y^T s}{s^T B s}$$

Terminology

□ Def:

- (i) A and B are said to be relatively sized if
$$\text{Spectrum}(A) \cap \text{Spectrum}(B) \neq \phi$$
- (ii) $\gamma \in \mathbb{R}$ sizes B relative to A if γB and A are relatively sized

□ Proposition: γ sizes B relative to A



$\exists u, v$ satisfying

$$\gamma = \frac{u^T A u}{u^T u} \frac{v^T v}{v^T B v}$$

- Corollary: For any u

$$\gamma = \frac{u^T Au}{u^T Bu}$$

sizes B relative to A

- Def: γ sizes B relative to the Hessian of f if there exists x such that γ sizes B relative to $\nabla^2 f(x)$.

Historical Background on Sizing

- 1974 Oren-Luenberger (SSVM)

size at each iteration with $\gamma_{OL} = \frac{y^T s}{s^T B s}$

Proposition: γ_{OL} sizes B relative to the Hessian of f

Proof:

$$y^T s = (\nabla f(x + s) - \nabla f(x))^T = s^T \nabla^2 f(x + \theta s) d$$

□ 1978 Shanno-Phua

- Observation: Secant equation $B_+s=y$ implies

$\frac{y^T s}{s^T B_+ s} = 1$. Therefore all secant updates are sized relative to the Hessian of f .

- Suggestion: Size only initial approximation in BFGS secant method and do so using γ_{OL} .

Question?

- Effectiveness of γ_{OL}
- Effective sizing strategy
 - Initial approximation only?
 - All approximations?
 - Selective approximations?

M. Contreras and R. Tapia (1993)

“Sizing The DFP and BFGS Updates: A Numerical Study”

- Propositions: If the sized-secant method converges q -superlinearly, then γ_{OL} converges to one.
- Selective sizing: size if
$$\varepsilon_1 < \gamma_{OL} < 1 - \varepsilon_2 \quad \varepsilon_1, \varepsilon_2 > 0$$

Contreras-Tapia Findings

- The DFP update loves to be sized by γ_{OL} . Sizing at every iteration is only slightly inferior to selective sizing. Without sizing DFP is vastly inferior to BFGS. With selective sizing competitive with a selectively sized BFGS.
- When sizing is working, γ_{OL} converges very nicely to one.
- Selective sizing for BFGS is best, sizing at each iteration is not good; it does not like to be sized.
- γ_{OL} is not a real good fit with BFGS. It tends to size too much especially for large dimensional problems.

New Research

Yabe-Martinez-Tapia (2004)

□ Premise:

For BFGS, especially in higher dimensions, B often has large eigenvalues (indeed by design) and this tends to give large Rayleigh quotients $s^T B s / s^T s$. Hence $\gamma_{OL} = y^T s / s^T B s$ is small and this in turn moves $\gamma_{OL} B$ in the direction of singularity.

□ Idea:

Follow sizing with γ_{OL} with shift within the Broyden class to compensate for near singularity.

Sized Broyden class

$$B_+ = BFGS(\gamma B, s, y) + \phi v v^T$$

set $\gamma = \gamma_{OL}$ and then find best ϕ

Byrd-Nocedal (1989)

A General Measure of Goodness

$$\omega(A) = TR(A) - \ln[\det(A)]$$

□ Proposition:

The measure ω is globally and uniquely minimized by $A = I$ over the class of symmetric positive definite matrices

□ Size and Shift Approach

Consider choices of the parameters γ and ϕ determined from the minimization problem

$$\min \omega\left(D^{-\frac{1}{2}} B_+ D^{-\frac{1}{2}}\right)$$

where $B_+ = BFGS(\gamma B, s, y) + \phi v v^T$
and D is a symmetric positive definite
weighting matrix.

Observe that $B_+ = D$ solves this problem; if B_+
is not restricted to the sized Broyden class

□ Obvious choices for D

- $D = I$ Obtain member of sized Broyden class closest to the identity – Gradient flavored
- $D = B$ Obtain member of sized Broyden class closest to D – least-change secant flavored
- $D = \nabla^2 f(x)$ Obtain member of sized Broyden class closest to the Hessian – Newton flavored

Three Optimization Problems

I. Given γ^* find ϕ^* as solution of

$$\min_{\phi} w\left(D^{-\frac{1}{2}}B_+D^{-\frac{1}{2}}\right)$$

II. Given ϕ^* find γ^* as solution of

$$\min_{\gamma} w\left(D^{-\frac{1}{2}}B_+D^{-\frac{1}{2}}\right)$$

III. Find γ^* and ϕ^* as solution of

$$\min_{\gamma, \phi} w\left(D^{-\frac{1}{2}}B_+D^{-\frac{1}{2}}\right)$$

Solutions

□ Problem I: Given γ $\phi^* = \frac{1}{v^T D^{-1} v} - \frac{\gamma}{\tau - 1}$

where $\tau = \frac{(s^T B s)(y^T B^{-1} y)}{(y^T s)^2}$

$$v = \sqrt{s^T B s} \left(\frac{y}{y^T s} - \frac{B s}{s^T B s} \right)$$

Observation: For $D = B$ $\phi^* = \frac{1 - \gamma}{\tau - 1}$.

Interpretation: In least change sense $\gamma = 1$ (no sizing) implies $\phi^* = 0$ (BFGS).

□ Problem II: Given ϕ

$$\gamma^* = \frac{1}{2\beta} \left[(n-1) - \beta\phi(\tau-1) + \sqrt{\delta} \right]$$

where

$$\beta = TR\left(D^{-\frac{1}{2}}B - D^{-\frac{1}{2}}\right) - \frac{s^T B D^{-1} B s}{s^T B s}$$

$$\tau = \frac{(s^T B s)(y^T B^{-1} y)}{(y^T s)^2}$$

$$\delta = \left(\beta\phi(\tau-1) - (n-1)^2 + r\beta\phi(\tau-1)(n-2) \right)$$

Observation: For $D = B$

$$\gamma^* = \frac{1}{2} \left(1 - \phi(\tau - 1) + \sqrt{\phi(\tau - 1) - 1)^2 + r\phi(\tau - 1) \left(\frac{n-2}{n-1} \right)} \right)$$

Hence $\phi = 0$ implies $\gamma^* = 1$

Interpretation: In least-change sense BFGS should not be sized.

- Problem III: Find both γ^* and ϕ^* from minimization problem

$$\gamma^* = \frac{n-2}{\beta - \frac{v^T D^{-1} v}{\tau-1}}$$

$$\phi^* = \frac{1}{v^T D^{-1} v} - \frac{\gamma^*}{\tau-1}$$

Observation: For $D = B$

$$\gamma^* = 1$$

$$\phi^* = 0$$

Interpretation: In least change sense BFGS with no sizing is best.

Numerical Experimentation

- Selectively size BFGS using γ_{OL}
- Shift using solution obtained with $D = I$ (gradient flavored), $D = B$ (least-change flavored), and $D = \nabla^2 f(x)$ (Newton flavored)

Surprise

The winner is $D = I$ (gradient flavored)

- Comment: There is consistency in this choice. Our sizing indicator has told us that we should size; hence
 - BFGS is probably not best and we should shift
 - Either B is bad, $\nabla^2 f(x)$ is bad, or there is a bad match between the two. Therefore least change $D = B$ may be dangerous and Newton $D = \nabla^2 f(x)$ may be dangerous. The choice $D = I$ prevents this faulty information from further contaminating the update; i.e. we use the member of the Broyden class which is closest to steepest descent.