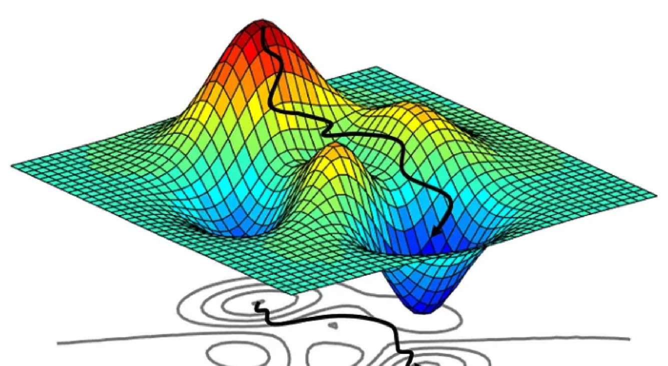
Multisecant Extensions of Quasi-Newton method

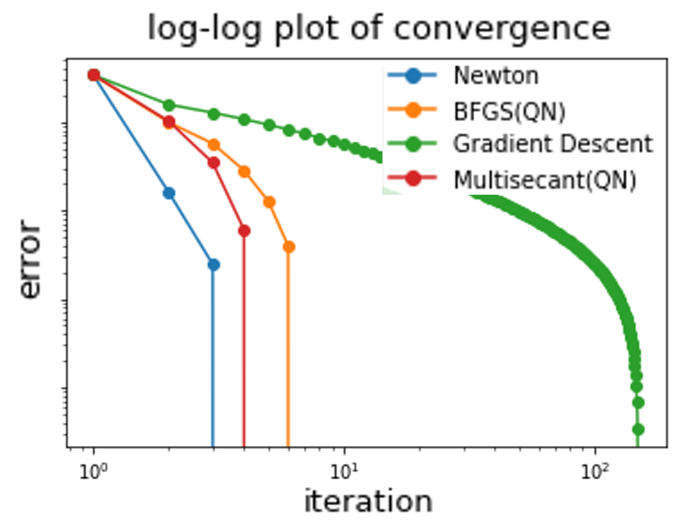
Mokhwa Lee, Yifan Sun
Stony Brook University

Abstract

When dealing with a large-scale optimization problem, classical second-order methods, such as Newton's method, are no longer practical because it requires iteratively solving a large-scale linear system of order n. For this reason, Quasi-Newton(QN) methods, like BFGS or Broyden's method, are introduced because they are more efficient than Newton's method. This project focuses on multi-secant extensions of the BFGS method, to improve its Hessian approximation properties. Unfortunately, doing so sacrifices the matrix estimate's positive semi-definiteness, and steps are no longer assured to be descent directions. Therefore, we apply a perturbation strategy to construct an almost-secant positive-definite Hessian estimate matrix. This strategy has a low computational cost, involving only low-rank updates with variable and gradient successive differences. We also explore several ways of improving this method, accepting and rejecting older updates according to several non-degeneracy metrics. Future goals include extending these techniques to limited memory versions.

Optimization method





• Main Problem : $\min_{x\in\mathbb{R}^n} f(x)$ where f is continuous and differentiable

Method	Gradient Descent	Newton	Quasi-Newton
Convergence rate	$O(C^{n})$	$O(C^{n^2})$	$O(C^{n^{1.618}})$
Memory	$O(n)$	$O(n^2)$	$O(n^2)$
Search direction	$-\nabla f(x_k)$	$-[\nabla^2 f(x_k)]^{-1}\nabla f(x_k)$	$-B_k^{-1}\nabla f(x_k)$
Per iteration cost	low	high	medium
Total iterations	high	low	low-medium

Hessian approximation

- Taylor expansion :

$$\nabla f(x_{k+1}) \approx \nabla f(x_k) + \nabla^2 f(x_k)(x_{k+1} - x_k)$$

- Hessian Approximation :

$$\nabla^2 f(x_{k+1}) \underbrace{(x_{k+1} - x_k)}_{s_k} \approx \underbrace{\nabla f(x_{k+1}) - \nabla f(x_k)}_{y_k}$$

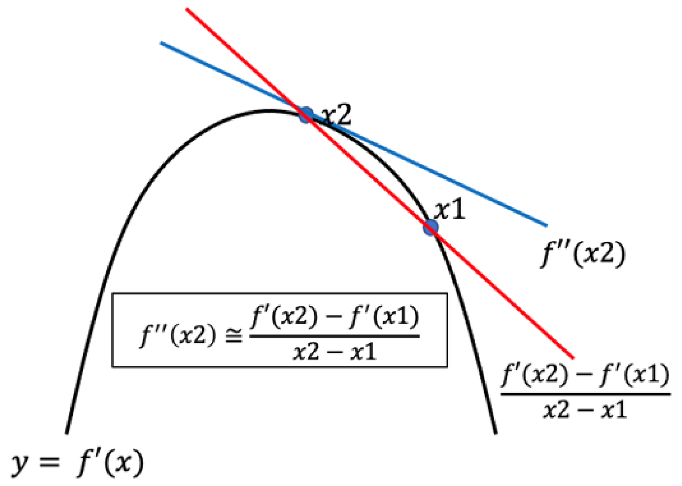
- Quasi-Newton's method :

$$x_{k+1} = x_k - \alpha B_k^{-1}\nabla f(x_k)$$

- Search Direction :

To guarantee descent direction, we need $-\nabla f_k^T B_k^{-1}\nabla f_k < 0$

$\Rightarrow$ Goal : Want to get positive definite approximate Hessian $B_{k+1} \approx \nabla^2 f(x_{k+1})$



Secant Conditions : single and multiple

Single-Secant Condition:

$$B_{k+1}s_k = y_k \quad \text{where} \quad B_{k+1} \in \mathbb{R}^{n\times n}$$

Multi-Secant (MS) Condition : With small $p \ll n$ (e.g. $p = 10$),

$$B_{k+1}s_i = y_i \quad \text{for} \quad i = k-p, \dots, k$$

or equivalently,

$$B_{k+1}S_k = Y_k$$

$$S = \begin{bmatrix} | & | & & | \\ s_{k-p} & s_{k-p+1} & \dots & s_k \\ | & | & & | \end{bmatrix} \quad and \quad Y = \begin{bmatrix} | & | & & | \\ y_{k-p} & y_{k-p+1} & \dots & y_k \\ | & | & & | \end{bmatrix}.$$

Symmetric matrix B_k has $\frac{n(n+1)}{2}$ degrees of freedom for the single-secant case:

Given the matrices S and Y, finding a $n \times n$ symmetric hessian matrix approximation B_k, which satisfies above single/multi-secant conditions, may have multiple free variables. This explains why there are many variations of QN methods.

Classic Multi-secant Quasi-Newton methods

For symmetric (PSD) hessian estimate update,

$$\exists B = B^T \text{ (and } B \succeq 0) \quad s.t \quad BS = Y \iff Y^TS \text{ is symmetric (and PSD)}$$

However, when $f(x)$ is not quadratic (or PSD), Y^TS is not symmetric (or PSD) in general.

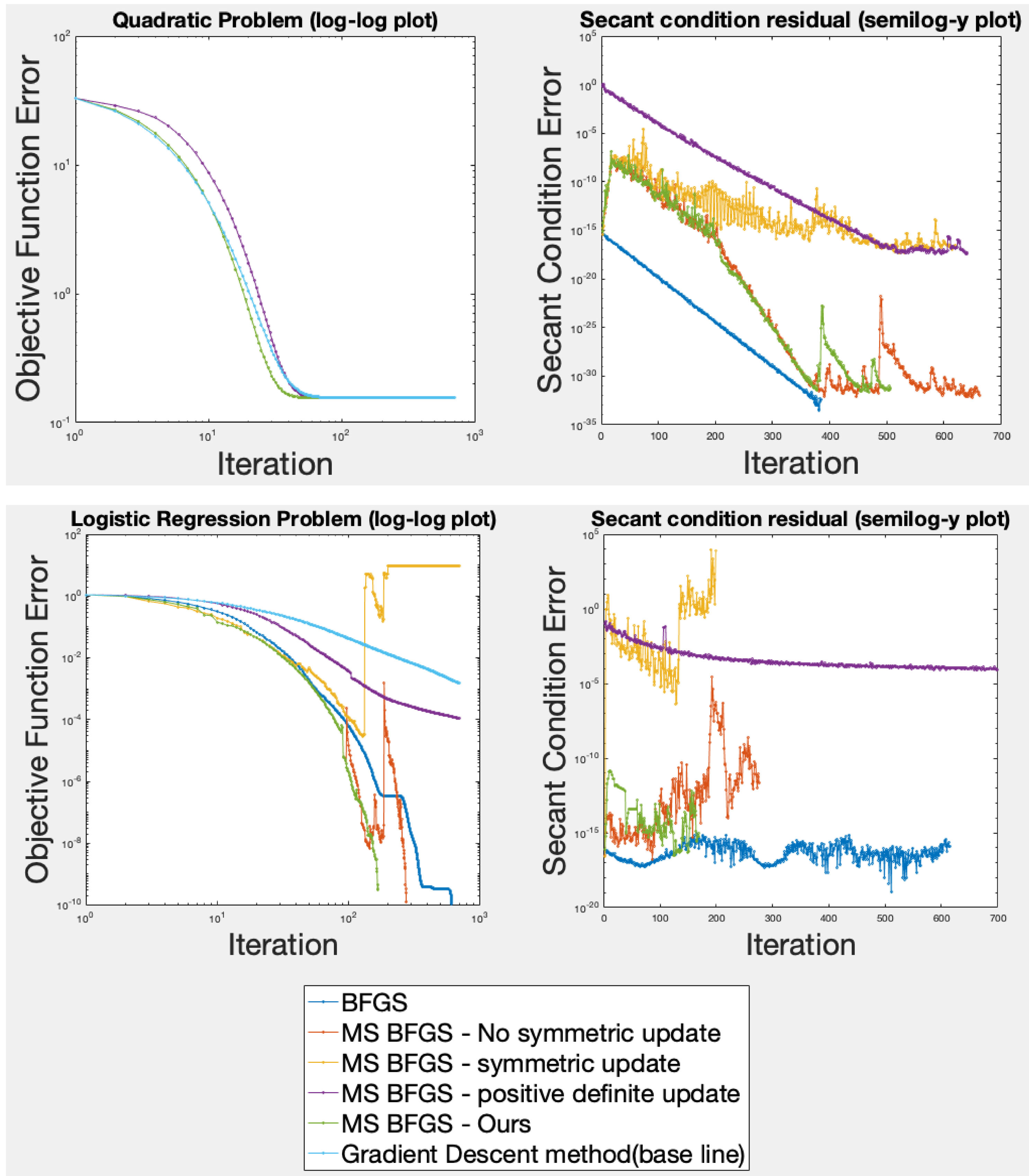
Method	Optimization Problem	Approximate Hessian matrix update
Broyden	$\min_B \|H - B\|_F \quad s.t \quad BS = Y$	$B = H + (Y - HS)(S^TS)^{-1}S^T$
PSB	$\min_B \|H - B\|_F \quad s.t \quad BS = Y, \quad B = B^T$	$B = H + (Y - HS)(S^TS)^{-1}S^T$ $+S^T(S^TS)^{-1}(Y - HS)^T$ $-S(S^TS)^{-1}(Y - HS)^TS(S^TS)^{-1}S^T$
DFP	$\min_B (W^{-T}(H - B)W^{-1})$ s.t $BS = Y,\ B = B^T,\ B \succeq 0$ where $W^TWS = Y$	$B = H + (Y - HS)(Y^TS)^{-1}Y^T$ $+Y(Y^TS)^{-1}(Y - HS)^T$ $-Y(Y^TS)^{-1}(Y - HS)^TS(Y^TS)^{-1}Y^T$
BFGS	$\min_B (W^T(H^{-1} - B^{-1})W^1)$ s.t $BS = Y,\ B^{-1} = B^{-T},\ B^{-1} \succeq 0$ where $W^TWS = Y$	$B = H + Y(Y^TS)^{-1}Y^T$ $-HS(S^THS)^{-1}S^TH$

Quasi-Newton methods comparison

Method	Broyden	PSB	DFP	BFGS	Multisecant	Ours
B is symmetric	X	X	✓	✓	X	✓
B is positive definite	X	X	✓	✓	X	✓
MS condition satisfied	X	X	X	X	✓	≈
Hessian update rank	1	2	2	2	$2p$	$2p$

$-B^{-1}\nabla f(x)$ is a guaranteed descent direction only if B is symmetric positive definite ($B \succeq 0$).

Simulation Results (Quadratic and Logistic Regression problems)

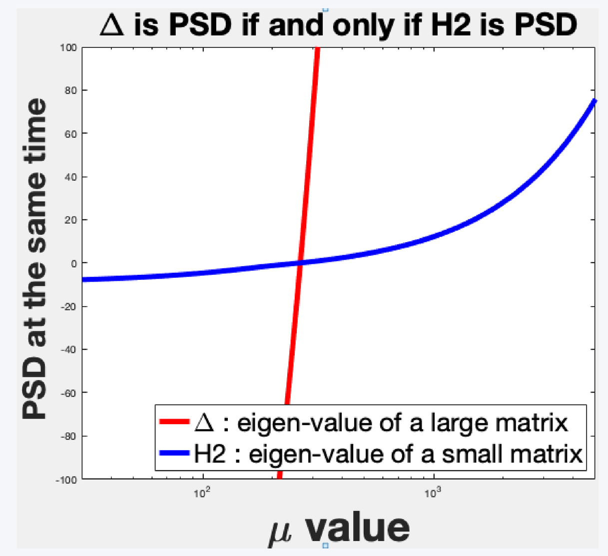


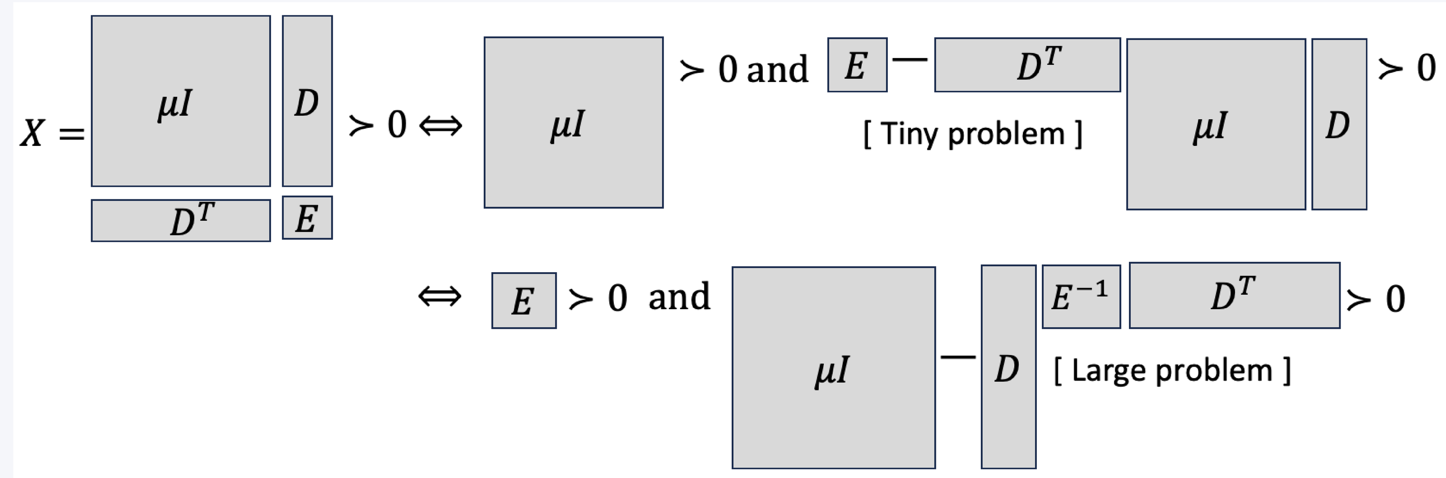
- Quadratic problem loss value monotonically decreasing because $B \succeq 0$
- Logistic regression problem (not quadratic) monotonically decrease only if B is positive definite.

Schur Complement - Symmetric positive semidefinite MS-BFGS

Summary

- Getting the smallest eigenvalue of B_{k+1} is expensive.
- We approximate via smallest eigenvalue of low-rank Δ where $B_{k+1} = B_k + \Delta$.
- This can be done by computing the smallest eigenvalue of $2p \times 2p$ matrix H_2 where $\mathbf{p} \ll \mathbf{n}$.





Scheme : pick smallest μ which satisfies the 'tiny problem' for **positive-definite** hessian estimate update.

- Almost secant multisecant BFGS method to get μ:

we want to guarantee the descent direction(positive-definite) by adding smallest μ

$$B_{k+1} = B - \underbrace{\begin{bmatrix} Y & BS \end{bmatrix}}_{D_1} \underbrace{\begin{bmatrix} \underbrace{-(Y^TS)^{-1}}_{\text{nonsym.}} & 0 \\ 0 & (S^TBS)^{-1} \end{bmatrix}}_{W^{-1}} \underbrace{\begin{bmatrix} Y^T \\ S^TB \end{bmatrix}}_{D_2} \qquad : \text{Multisecant BFGS}$$

$$\begin{aligned} B_{k+1} &= B - \frac{D_1W^{-1}D_2^T + (D_1W^{-1}D_2^T)^T}{2} + \mu I \\ &= B - \frac{1}{2}\begin{bmatrix} D_1 & D_2 \end{bmatrix} \begin{bmatrix} 0 & W_k^{-1} \\ W_k^{-T} & 0 \end{bmatrix} \begin{bmatrix} D_1^T \\ D_2^T \end{bmatrix} + \mu I \qquad : \text{Ours(symmetric + PSD)} \end{aligned}$$

where $B \in \mathbb{R}^{n\times n}$, $D_1, D_2 \in \mathbb{R}^{n\times 2p}$, $W^{-1} \in \mathbb{R}^{2p\times 2p}$ and $\mu \in \mathbb{R}$.

- Theorem

We symmetrize the update term in B_{k+1} and want it to be positive definite as below

$$B_{k+1} = B \underbrace{-\frac{1}{2}\begin{bmatrix} D_1 & D_2 \end{bmatrix} \begin{bmatrix} 0 & W_k^{-1} \\ W_k^{-T} & 0 \end{bmatrix} \begin{bmatrix} D_1^T \\ D_2^T \end{bmatrix} + \mu I}_{\Delta}.$$

Consider W a non-symmetric matrix. For any $c \in \mathbb{R}^+$, define

$$P = (cI - c^{-1}FF^T)^{-1} \quad \text{and} \quad Q = (cI - c^{-1}F^TF)^{-1}.$$

Pick $F = c_3USV^T$ where $W^{-1} = U\Sigma V^T$ is the full SVD of W^{-1} , and $c_3 = \frac{c\epsilon}{c+\|W\|_2}$.
Let S be a diagonal matrix satisfying

$$\Sigma = (S^2 - c^2I)^{-1}S.$$

Then Δ is PSD if and only if

$$H_2 = \begin{bmatrix} cI & F \\ F^T & cI \end{bmatrix} - \begin{bmatrix} D_1^T \\ D_2^T \end{bmatrix} (A + 2\mu I)^{-1} \begin{bmatrix} D_1 & D_2 \end{bmatrix} \in \mathbb{R}^{4p\times 4p}$$

is PSD, for

$$A = \begin{bmatrix} D_1 & D_2 \end{bmatrix} \begin{bmatrix} P & -(c^2I - F^TF)^{-1}F^T - W^{-1} \\ -F(c^2I - F^TF)^{-1} - W^{-T} & Q \end{bmatrix} \begin{bmatrix} D_1^T \\ D_2^T \end{bmatrix}.$$

- Reference

(1) Schnabel, Robert B. "Quasi-Newton methods using multiple secant equations." Computer Science Technical Reports 244.41 (1983): 06.
(2) Gay, David et al. "Solving systems of nonlinear equations by Broyden's method with projected updates(1978)".
(3) Jorge Nocedal et al. "On the limited memory BFGS method for large scale optimization (1989)."
(4) Byrd, Richard H., Jorge Nocedal, and Robert B. Schnabel. "Representations of quasi-Newton matrices and their use in limited memory methods." Mathematical Programming 63.1-3 (1994): 129-156.
(5) Hassan, Basim A., and Issam AR Moghrabi. "A modified secant equation quasi-Newton method for unconstrained optimization." Journal of Applied Mathematics and Computing 69.1 (2023): 451-464.

mokhwa.lee@stonybrook.edu