

Statistical Quasi-Newton: A New Look at Least Change

Chuanhai Liu and Scott Vander Wiel

Bell Laboratories, Lucent Technologies

July 23, 2004

Abstract

A new method for quasi-Newton minimization outperforms BFGS by combining a new least-change update of the Hessian with a step-size estimate obtained from a Wishart model of uncertainty. The Hessian update preserves accuracy from one iteration to the next. It is in the Broyden family but uses a negative parameter, outside the convex range that is usually regarded as the safe-zone for Broyden updates. Although full Newton steps based on this update tend to be too long, excellent performance is obtained with step sizes estimated from a Wishart model of Hessian uncertainty. In numerical comparisons to BFGS the new *Statistical quasi-Newton* algorithm typically converges with about 20% fewer iterations and gradient evaluations and 10% fewer function evaluations on a suite of standard test functions. Our statistical framework provides a simple way to understand differences among various Broyden updates such as BFGS and DFP and shows that these methods do not preserve Hessian accuracy while the new method does. In fact, BFGS, DFP and all other updates with non-negative Broyden parameters tend to inflate Hessian estimates and this accounts for their observed propensity to correct eigenvalues that are too small more readily than eigenvalues that are too large. Numerical results on three new test functions validate these conclusions.

Key Words: BFGS, DFP, negative Broyden family, Wishart model.

1 Introduction

Quasi-Newton methods for unconstrained optimization are important computational tools in many fields of scientific investigation and are a standard subject in text books on computation, including statistical computation (*e.g.*, Chambers, 1977). The BFGS method, proposed individually by Broyden (1970), Fletcher (1970), Goldfarb (1970), and Shanno (1970), is implemented in most optimization software and is widely recognized as efficient. In theoretical investigations BFGS is known as a special case of the Broyden class (Broyden, 1967). Some Broyden updates with negative Broyden parameters have been found to produce faster convergence than BFGS updates

(e.g., Zhang and Tewarson, 1988; Byrd, Liu, and Nocedal, 1992) but, for various reasons, have not been widely adopted. Indeed, Byrd et. al. conclude that “practical algorithms that preserve the excellent properties of the BFGS method are difficult to design.” Nocedal and Wright (1999) state that “the BFGS formula ... is presently considered to be the most effective of all quasi-Newton updating formulae.” In our opinion, BFGS remains the most popular front-runner because of two important unanswered questions: What is the “best” negative Broyden parameter and how should initial step lengths be scaled when using negative Broyden parameters? This paper answers these questions by formulating a least-change problem whose goal is to approximate Newton directions and by estimating step sizes through a statistical model of Hessian uncertainty. Originally we derived the new least-change update by incorporating a prior distribution into the model for Hessian uncertainty as discussed in (Section 7.1). Whereas both the Hessian update and the step size were obtained from *statistical* considerations, the new algorithm is called the *Statistical Quasi-Newton* (SQN) method.

1.1 Quasi-Newton methods

Quasi-Newton methods solve the unconstrained optimization problem

$$\min_x f(x), \quad x \in \mathcal{R}^n,$$

in which both the objective function $f(x)$ and its gradient $g(x) \equiv \nabla f(x)$ are easy to compute but Newton’s method is not applicable because direct evaluation of the Hessian matrix $G(x) \equiv \nabla^2 f(x)$ is practically infeasible. Quasi-Newton methods build up an approximate Hessian matrix using successive gradient evaluations. The general method iterates between a minimization (M) step consisting of a one-dimensional search for a good point along an approximate Newton direction and an estimation (E) step consisting of an update to the Hessian estimate. A more specific definition follows.

Generic quasi-Newton algorithm: Select a starting point $x_0 \in \mathcal{R}^n$ and a symmetric positive definite estimate, B_0 , of the Hessian matrix $G(x_0)$. Let $g_0 = g(x_0)$ and iterate over $k = 0, 1, 2, \dots$ the following two steps.

M-Step. Carry out a line search from the point x_k in the direction $-B_k^{-1}g_k$, to obtain a step $-s_k B_k^{-1}g_k$ (with $s_k > 0$) that satisfies *the Wolfe conditions* for sufficient decrease of the function and for curvature (see (2) and (3) below). The new evaluation point and gradient are

$$x_{k+1} = x_k - s_k B_k^{-1}g_k \quad \text{and} \quad g_{k+1} = g(x_{k+1}).$$

E-Step. Estimate the Hessian matrix at x_{k+1} using the quantities B_k, x_k, x_{k+1}, g_k , and g_{k+1} . The estimate, B_{k+1} , must be symmetric and positive definite and must satisfy

the *quasi-Newton condition*

$$B_{k+1}\delta_k = \gamma_k, \quad (1)$$

where $\delta_k \equiv x_{k+1} - x_k$ and $\gamma_k \equiv g_{k+1} - g_k$.

Condition (1) requires the vector of estimated second derivatives in the current step direction, $B_{k+1}\delta_k/s_k$, to agree with the corresponding numerical second derivatives γ_k/s_k . Different principles have been used to derive Hessian update formulae but the general goal has been to minimize the change from B_k to B_{k+1} in some sense. This paper derives an update that minimizes change in a canonical sense and provides a model-based estimate for the step size s_k .

The Wolfe conditions referenced in the M-step are two standard requirements to ensure that sufficient progress is made toward the optimum even when the line search is not required to find the exact minimum in the given search direction. The Wolfe *sufficient decrease condition*,

$$f(x_{k+1}) \leq f(x_k) - c_1 s_k g'_k B_k^{-1} g_k, \quad (c_1 \in (0, 1), \text{ say } c_1 = 10^{-4}), \quad (2)$$

requires a reduction in $f(x)$ that is at least a fraction c_1 of that predicted by the directional derivative $-g_k B_k^{-1} g_k$. The Wolfe *strong curvature condition*

$$|g'_{k+1}(B_k^{-1} g_k)| \leq c_2 g'_k(B_k^{-1} g_k), \quad (c_2 \in (c_1, 1), \text{ say } c_2 = 0.9), \quad (3)$$

requires at least a proportional decrease in the magnitude of the derivative in the search direction. Some algorithms impose a weaker curvature condition in which the absolute value is removed from the left side of (3). Nocedal and Wright (1999, p. 37-41) discuss the importance of the Wolfe conditions in assuring that sufficient progress is made on each iteration.

The best-known class of Hessian estimates used in the M-step are the rank-two Broyden updates (Broyden, 1967):

$$B_{k+1} = B_k - \frac{B_k \delta_k \delta'_k B_k}{\delta'_k B_k \delta_k} + \frac{\gamma_k \gamma'_k}{\delta'_k \gamma_k} + c_k \omega_k \omega'_k, \quad (4)$$

where

$$\omega_k \equiv \frac{\gamma_k}{\delta'_k \gamma_k} - \frac{B_k \delta_k}{\delta'_k B_k \delta_k} \quad (5)$$

and c_k is a scalar parameter to be determined. The usual parameterization takes $c_k = \phi_k (\delta'_k B_k \delta_k)$ where ϕ_k is known as the *Broyden parameter*. However, our exposition is more natural with the parameterization

$$c_k = (\lambda_k - 1) (\delta'_k \gamma_k) \quad (6)$$

where the parameter λ_k is shown in Section 3 to regulate the inflation of B_{k+1} relative to B_k . BFGS is the Broyden update with $\lambda_k = 1$ (*i.e.*, $\phi_k = c_k = 0$).

There is a critical value λ_k^c such that B_{k+1} is positive definite for any $\lambda_k > \lambda_k^c \equiv 1 - r_k^{-1}$ where

$$r_k \equiv \frac{\gamma'_k B_k^{-1} \gamma_k}{\gamma'_k \delta_k} - \frac{\delta'_k \gamma_k}{\delta'_k B_k \delta_k}.$$

It can be shown that $r_k \geq 0$ by making use of the curvature condition (3) and the Cauchy-Schwarz inequality. If $r_k = 0$ then λ_k^c is taken to be $-\infty$.

1.2 Preview of a statistical quasi-Newton method

The SQN method developed in this paper is remarkably simple and effective. This section briefly defines SQN and demonstrates its superiority to BFGS.

Statistical quasi-Newton (SQN) algorithm: Follow the generic quasi-Newton algorithm with the following additional specifications.

E-Step. Update B_k using a Broyden update (4)–(6) with

$$\lambda_k = \max\{0, 1 - (1 - \epsilon)r_k^{-1}\} \quad (7)$$

where ϵ is a small positive constant (e.g., $\epsilon = 10^{-6}$) to guarantee that B_{k+1} remains positive definite. The corresponding Broyden parameter $\phi_k = (\lambda_k - 1)(\delta'_k \gamma_k) / (\delta'_k B_k \delta_k)$ is negative because (3) implies $\delta'_k \gamma_k > 0$.

M-Step. Begin the line search from an initial evaluation point $-\hat{s}(\lambda_k)B_k^{-1}g_k$ where

$$\hat{s}(\lambda_k) = \frac{g'_{k+1} B_{k+1}^{-1} g_{k+1}}{g'_{k+1} B_{k+1}^{-1} g_{k+1} + (1 - \lambda_k)(\delta'_k \gamma_k)(g'_{k+1} B_{k+1}^{-1} \omega_k)^2} \leq 1. \quad (8)$$

The shortened initial step is crucial to improving the performance of Broyden updates with negative Broyden parameters. Zhang and Tewarson (1988) use $\hat{s} = 1$ and comment that their negative Broyden algorithm improves iteration counts but “less or no savings are achieved on the number of function evaluations” because initial steps are often too long to provide a sufficient decrease in the function value. SQN corrects this problem by effectively estimating the optimal step size for the given search direction.

Figure 1 compares the performance of SQN to BFGS on a standard set of small-dimensional test functions given by Moré, Garbow and Hillstom (1981). For each of twenty test functions, the plot shows the number of iterations until convergence of SQN relative to BFGS. Test functions are listed on the right axis with the dimension of x given in parenthesis. Each point in the plot shows the iteration ratio for minimization from a given starting point x_0 . Typically, ten starting points are shown for each function. The solid curve connects average iteration ratios. The overall average is 79%, indicating that SQN typically requires 21% fewer iterations than BFGS on these test cases. Specifications of the testing setup are given in Section 5 along with more detailed results on these

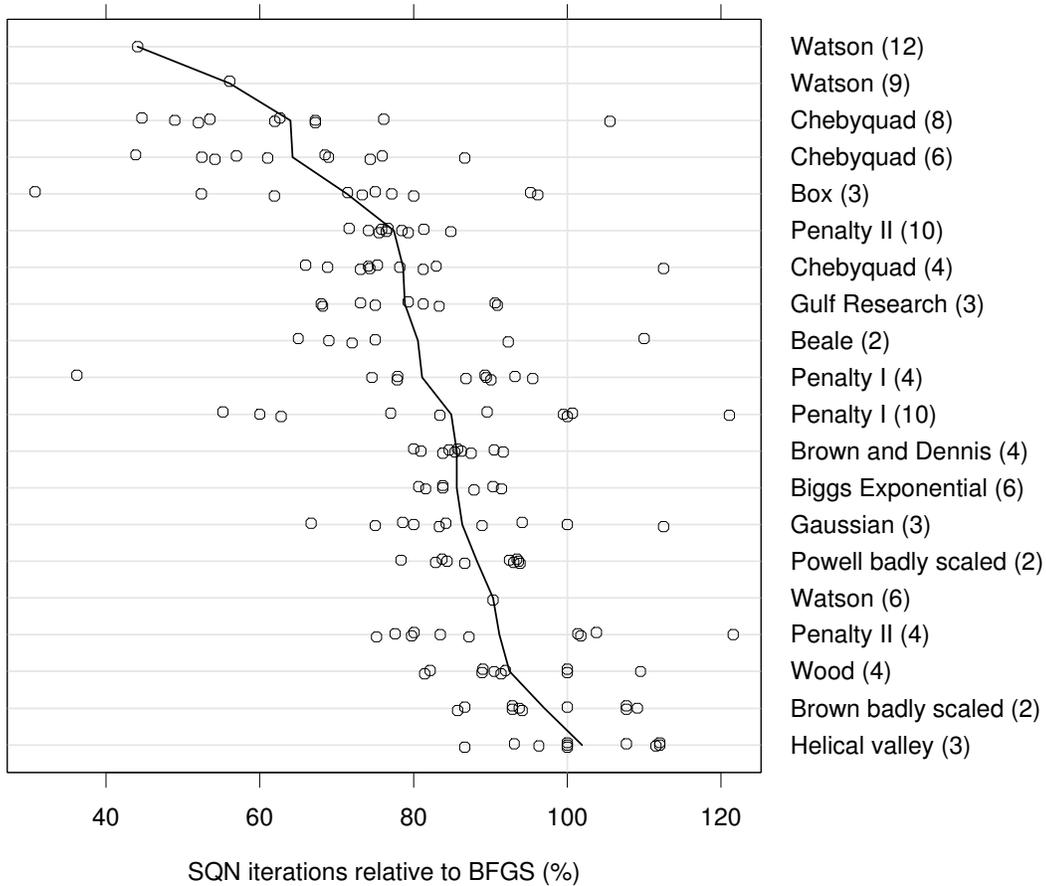


Figure 1: Iteration counts for SQN relative to BFGS on 20 problems of small dimension. On average SQN converges in 21% fewer iterations on these cases.

cases and additional numerical tests. The experiment underlying Figure 1 closely follows that of Zhang and Tewarson (1988) who also achieved 21% iteration improvement on this set of problems using their SDQN algorithm.

Figure 2 shows that SQN's efficiency relative to BFGS actually *improves* with the difficulty of the problem and the improvement holds not just for iteration counts but also for counts of function and gradient evaluations. Demonstrating improvement on function and gradient counts is important because otherwise iteration improvement could come at the price of worse efficiency overall. Each panel in the display has one point for each of the 44 Moré test problems listed in Tables 2 and 3 of Section 5. SQN appears to obtain a cumulative advantage over BFGS as additional iterations offer further opportunities for improvement. Performance on easy problems with few iterations is often

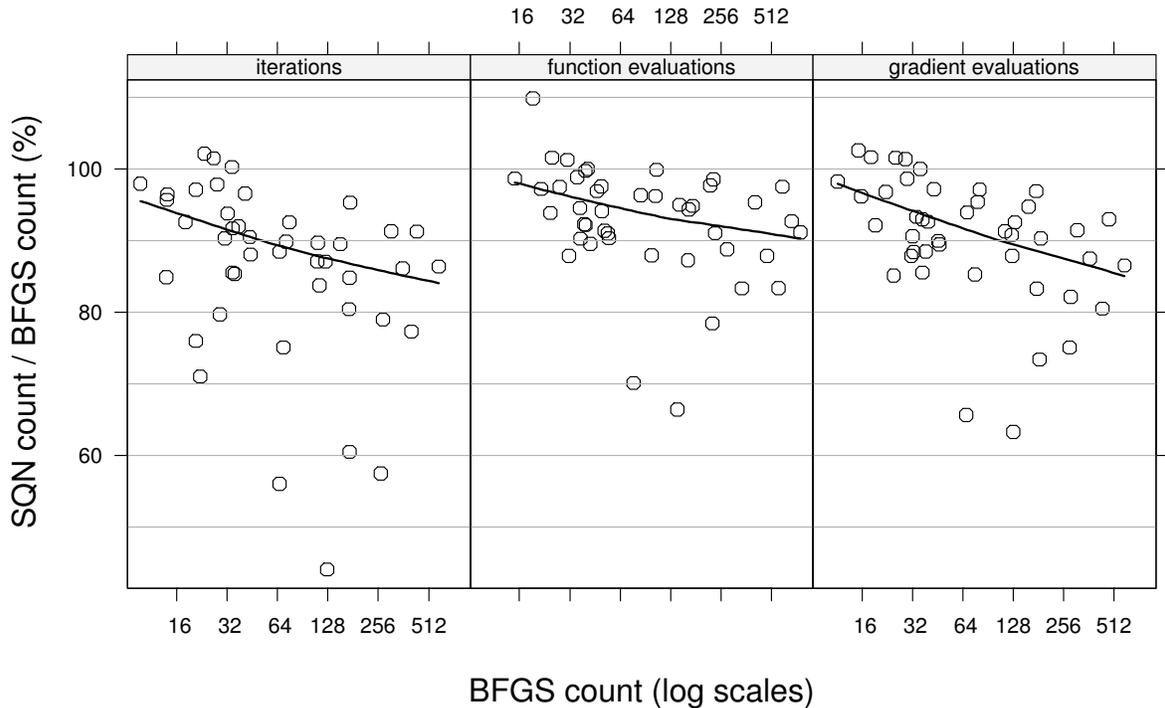


Figure 2: Improvement in SQN efficiency with problem difficulty for each of three counts—iterations, function evaluations and gradient evaluations. Each point represents the average performance of SQN and BFGS on a given problem from multiple starting points. The data come from Tables 2 and 3 in Section 5.

dominated by the first iteration in which a poor choice of B_0 produces a poor search vector for any quasi-Newton algorithm. In harder problems these startup effects wash out so that the advantage of SQN over BFGS becomes more apparent. The trend curves in Figure 2 highlight this tendency. The curves are robust local regression smoothes (Cleveland, 1979) that follow the trends without being unduly influenced by outlying points. The trend is strongest for the number of iterations to convergence.

The remainder of the article is arranged as follows. Section 2 gives a select history of ideas in quasi-Newton development with emphasis on the least change principle and argues for a particular scale-free matrix as the most appropriate measure of change in consecutive Hessian estimates. Section 3 introduces a transformation into canonical coordinates, derives (4)–(7) as the least-change update and shows that the update preserves Hessian accuracy from one iteration to the next. Section 4 introduces a Wishart model to describe Hessian uncertainty and uses it to derive (8) as an estimate of the optimal step size. Section 5 compares performance of SQN to BFGS on

the Moré test problems. Section 6 compares performance on three new test functions designed to verify our understanding of why SQN is better than other Broyden updates. Section 7 discusses a statistical formulation of the least-change principle used to derive SQN, explores connections to other least-change derivations and concludes with two ideas for future research.

2 Least-Change Updates

Fletcher’s (1994) overview of methods for unconstrained optimization is an excellent introduction to the huge literature on quasi-Newton methods. This section briefly reviews the historical ideas that led to the least-change principle on which the most influential quasi-Newton methods are based. A line of reasoning is then given to suggest a certain relative change matrix as being the most appropriate measure of change for the goal of approximating Newton search directions.

2.1 Historical developments

Crockett and Chernoff (1955) stated the idea of building up a Hessian estimate iteratively so as to approximate the Newton method:

..., it is possible to obtain, from the successive approximations, certain relevant information about terms of order higher than those actually computed, and to conveniently use this information to improve the rate of convergence.

The basic idea of Broyden (1965) as articulated in Broyden (2000) was that the Hessian update “Should therefore require, if possible, ..., no change to B_k in any direction orthogonal to δ_k .” Broyden was solving a system of differential equations and his mathematical formulation [$B_{k+1}\delta_k = \gamma_k$ and $(B_{k+1} - B_k)q = 0, \forall q : q'\delta_k = 0$] produces an asymmetric update that is not appropriate for the problem $\min f(x)$.

Taking a more mathematical approach, Broyden (1967) dropped the “orthogonality” part of his original intuition and sought instead a low-rank Hessian update. This led to the Broyden class (4) of symmetric rank-two updates. Subsequent researchers also focused on making small modifications to the Hessian without explicit concern for the space orthogonal to the search direction. Greenstadt (1970), for example, wrote,

Let us ask for the “best” correction in some sense. There are many possible choices to make, but a good one is to ask for the smallest correction, in the sense of some norm. To a certain extent, this would tend to keep the elements of $[B_k^{-1}]$ from growing too large, which might cause an undesirable instability.

The extensive review of Lukšan, Spedicato and Vlček (1999) emphasizes the importance of the *least change principle* in deriving many of the most effective quasi-Newton methods.

The important special case of a Broyden update with $\lambda_k = 1$ is called BFGS after the four authors who individually published the update formula in 1970. Goldfarb (1970) worked with the scaled difference of inverse Hessian estimates

$$E_W^* \equiv W^{1/2} (B_{k+1}^{-1} - B_k^{-1}) W^{1/2} \quad (9)$$

where the symmetric matrix W satisfies $W \delta_k = \gamma_k$. He derived the BFGS update by using Greenstadt's (1970) results to minimize the Frobenius norm $\|E_W^*\|_F \equiv [\text{tr}(E_W^* E_W^*)]^{1/2}$ over the class of symmetric matrices B_{k+1} that satisfy the Newton condition (1). Thus, BFGS is a least-change update. But the metric of change is important. For example, using the same W but minimizing the Frobenius norm of

$$E_W \equiv W^{-1/2} (B_{k+1} - B_k) W^{-1/2} \quad (10)$$

produces the Broyden update with $\lambda_k = 1 + \delta'_k B_k \delta_k / (\delta'_k \gamma_k)$. This is known as DFP after Davidon (1959) and Fletcher and Powell (1963) and is generally regarded as inferior to BFGS.

Fletcher (1970) advocated restricting attention to Broyden updates that are convex combinations of the BFGS and DFP updates because such updates satisfy a monotone eigenvalue property when used to minimize quadratic functions. Recently, however, various choices of negative Broyden parameters ($\phi_k < 0$ corresponding to $\lambda_k < 1$) have been studied. See, for example, Zhang and Tewarson (1988), Byrd, Liu, and Nocedal (1992), Lukšan (1992), Fletcher (1994) and Mifflin and Nazareth (1994). These authors report that negative Broyden parameters can reduce iteration counts, although in some cases this comes at the cost of increased numbers of function evaluations. The potential for improvement relative to BFGS seems to be best if the initial Hessian estimate is much too large. However, no clear and consistent principle has been articulated for selecting the best Broyden parameter and robust improvement over BFGS has been elusive. Indeed, Zhang and Tewarson (1988) concluded that such investigations have not shaken the position of BFGS as the most popular front-runner.

2.2 A new least-change metric

Minimizing the change from B_k to B_{k+1} is a generally accepted principle. Section 7.1 discusses why the principle is valid even in the context of a statistical model for Hessian uncertainty. There is not agreement, however, on how to measure the distance from B_k to B_{k+1} . Zhao (1992) derives ten different optimal updates by considering five possible matrix norms applied to two different matrices that measure change. The metric for measuring change is empirically important: BFGS outperforms DFP even though the two are least change duals derived from E_W^* and E_W respectively.

The proliferation of quasi-Newton updates suggests that a refinement of the generic least-change principle is needed. We offer the following line of reasoning as justification for minimizing change as measured by a particular scale-free matrix E_B , defined in (11) below.

1. The search directions $-B_{k+1}^{-1}g_{k+1}$ of a quasi-Newton method should closely approximate the Newton directions $-G^{-1}(x_{k+1})g_{k+1}$, and therefore Hessian accuracy is most appropriately measured on the inverse scale.
2. Suppose B_k^{-1} approximately preserves the curvature information obtained in recent iterations; that is,

$$B_k^{-1}\gamma_{k-i} \approx \delta_{k-i} \quad i = 1, 2, \dots$$

where the approximations tend to be better for smaller i . The new estimate B_{k+1} should maintain these approximations as closely as possible while incorporating the new curvature information as specified by the quasi-Newton condition (1). In particular, the update should be small along directions $\gamma_{k-1}, \gamma_{k-2}, \dots$; that is

$$(B_{k+1}^{-1} - B_k^{-1})\gamma_{k-i} \approx 0 \quad (i = 1, 2, \dots).$$

But, in the context a quasi-Newton algorithm, the $\{\gamma_{k-i}\}$ are not known so the goal must be to make $(B_{k+1}^{-1} - B_k^{-1})$ small in every direction. The update from B_0 to B_1 is an exception because B_0 is typically set arbitrarily. Thus, it seems reasonable to use pre-scaling (Shanno and Phua, 1978) or some other aggressive updating technique for the initial Hessian update.

3. The deviation from B_k^{-1} to B_{k+1}^{-1} should be normalized so that the problem of finding the best update becomes affine invariant. Greenstadt (2000) argues that this is desirable “in that it renders harmless the accidents of coordinate selection in a given problem.” Symmetric scaling gives $G(x_k)^{1/2} (B_{k+1}^{-1} - B_k^{-1}) G(x_k)^{1/2}$ as the best matrix to measure relative change. However, $G(x_k)$ is unknown so we substitute the best available estimate. The current estimate B_k could be used but B_{k+1} is a better choice because it will incorporate the most recently obtained curvature information. (This reasoning assumes that curvature differences from x_k to x_{k+1} are inconsequential relative to the error in approximating $G(x_k)$ by B_k .) Therefore, a close approximation to the Newton method is obtained by minimizing the “size” of the scale-free matrix

$$E = B_{k+1}^{1/2} (B_{k+1}^{-1} - B_k^{-1}) B_{k+1}^{1/2}.$$

4. A matrix norm defines the “size” of E . For many norms, E is equivalent to

$$E_B \equiv B_k^{-1/2} (B_{k+1} - B_k) B_k^{-1/2} \tag{11}$$

in the sense that $\|E\| = \|E_B\|$. This equivalence holds for any matrix norm that depends only on eigenvalues (*e.g.*, determinant). It also holds for the Frobenius norm. Thus, we regard E and E_B as equivalent affine invariant measures of the change from B_k^{-1} to B_{k+1}^{-1} .

There is a fascinating historical connection that ties the relative change matrix E_B to BFGS, DFP and Greenstadt's (1970) method from which Goldfarb (1970) derived BFGS. As reviewed in Section 2.1, E_W^* and E_W , defined in (9) and (10) are well-known duals that measure change on the inverse and nominal scales and lead to the BFGS and DFP methods respectively. In the same sense, the dual of E_B , is

$$E_I \equiv B_k^{1/2} (B_{k+1}^{-1} - B_k^{-1}) B_k^{1/2}$$

which is the matrix that Greenstadt minimized to derive the E_I method. Therefore the SQN update derived from E_B (in the next section) is the dual of Greenstadt's E_I method in the way that BFGS is the dual of the older DFP method. (One small difference is that Greenstadt did not constrain B_{k+1} to be positive definite, as SQN does.)

3 SQN: Least Relative Change

The form of E_B in (11) as a measure of change motivates transforming the coordinates of x by $B_k^{1/2}$ so that the problem of updating the Hessian estimate takes a simple form. This section uses Broyden's original idea of preserving the portion of B_k that is orthogonal to δ_k but applies it in the transformed coordinate system.

As the focus is on the k -th step of the quasi-newton algorithm, the notation is streamlined from this point forward by dropping subscripts k and replacing subscripts $k + 1$ by '+'.

3.1 Canonical coordinates

For conceptual convenience, at the k -th iteration transform x in such a way that the line search is along the first component direction and the current Hessian estimate B transforms to the identity matrix. This is accomplished by the linear transformation

$$\tilde{x} = U' B^{1/2} x, \tag{12}$$

where U is an orthonormal rotation matrix with first column equal to $B^{1/2}\delta/(\delta'B\delta)^{1/2}$. In the transformed space the current step is strictly along the first component direction:

$$\tilde{x}_+ - \tilde{x} = (\delta'B\delta)^{1/2}(1, 0, \dots, 0)'$$

and the objective function is

$$\tilde{f}(\tilde{x}) \equiv f(x)$$

with gradient

$$\tilde{g}(\tilde{x}) \equiv \nabla \tilde{f}(\tilde{x}) = U' B^{-1/2} g(x)$$

and Hessian

$$\tilde{G}(\tilde{x}) \equiv \nabla^2 \tilde{f}(\tilde{x}) = U' B^{-1/2} G(x) B^{-1/2} U. \quad (13)$$

Substituting the estimated Hessian B for $G(x)$ in (13) produces the transformed estimate $\tilde{B} = I_n$, the n -dimensional identity matrix.

3.2 Observed and missing information

Define second-order numerical derivatives of $\tilde{f}(\tilde{x})$ along the search direction as

$$\begin{bmatrix} a \\ b \end{bmatrix} \equiv \frac{\tilde{g}(\tilde{x}_+) - \tilde{g}(\tilde{x})}{(1, 0, \dots, 0)(\tilde{x}_+ - \tilde{x})} = \frac{U' B^{-1/2} \gamma}{(\delta' B \delta)^{1/2}}, \quad (14)$$

where the first element a is a scalar and b is an $(n-1)$ -dimensional vector. The curvature condition (3) implies that $a \geq (1 - c_2)/s > 0$. The quasi-Newton condition (1) is equivalent to the intuitive idea that the above numerical derivatives form the first column of the updated Hessian matrix. Since the Hessian is symmetric, the general form of Hessian update in transformed coordinates becomes

$$\tilde{B}_+ = \begin{bmatrix} a & b' \\ b & C \end{bmatrix}, \quad (15)$$

where C is to be determined subject only to the constraint $\tilde{B}_+ > 0$ which is equivalent to $C - a^{-1}bb' > 0$. (The notation $M > 0$ indicates that the matrix M is positive definite.) C represents curvature in the complimentary space—that is, the space canonically orthogonal to the current search direction.

Following Broyden's idea that no information is gained in directions orthogonal to δ suggests the updating scheme obtained by taking $C = I_{n-1}$ if doing so produces $\tilde{B}_+ > 0$ — *i.e.*, if $a > b'b$. But, what to do if $a \leq b'b$? The question itself implies that certain information on C is provided by the observed data (a, b) along with the assumption that the Hessian matrix is positive definite. In general, C should be a function of a and b .

The following theorem provides the least-change update based on the Frobenius norm of E_B .

Theorem 1 (SQN Update) *The quasi-Newton update that minimizes $\|E_B\|_F$ subject to (1) and $B_+ \geq 0$ has canonical form*

$$\tilde{B}_+ = \begin{bmatrix} a & b' \\ b & I_{n-1} + \lambda_{\text{SQN}} bb'/a \end{bmatrix}, \quad (16)$$

where for $r \equiv b'b/a$,

$$\lambda_{\text{SQN}} = \begin{cases} 0, & \text{if } r \leq 1 \\ 1 - r^{-1} & \text{otherwise.} \end{cases} \quad (17)$$

\tilde{B}_+ is singular for $r \geq 1$.

Proof.

$$\begin{aligned}
\|E_B\|_F &= \left\| B^{-1/2} (B_+ - B) B^{-1/2} \right\|_F = \left\| U' B^{-1/2} B_+ B^{-1/2} U - I_n \right\|_F = \left\| \tilde{B}_+ - I_n \right\|_F \\
&= \text{tr} \left(\left[\begin{pmatrix} a & b' \\ b & C \end{pmatrix} - I \right] \left[\begin{pmatrix} a & b' \\ b & C \end{pmatrix} - I \right] \right) \\
&= \text{tr} (\Phi^2) - 2\text{tr} (\Phi) + 2b'\Phi b/a
\end{aligned}$$

where $\Phi \equiv C - bb'/a$ and we have used $\tilde{B}_+ = U' B^{-1/2} B_+ B^{-1/2} U$ from (13). $B_+ \geq 0$ is equivalent to $\Phi \geq 0$ so minimizing $\|E_B\|_F$ over $B_+ \geq 0$ is equivalent to minimizing the final expression over $\Phi \geq 0$.

Denote the eigenvalues Φ by $0 \leq \eta_1 \leq \dots \leq \eta_n$. Then

$$\text{tr} (\Phi^2) - 2\text{tr} (\Phi) + 2b'\Phi b/a \geq \sum_i \eta_i^2 - 2 \sum_i \eta_i + 2\eta_1 b'b/a \tag{18}$$

with equality if and only if the first eigenvector of Φ is proportional to b . The right-hand side of (18) is minimized by $\eta_2 = \dots = \eta_n = 1$ and

$$\eta_1 = \max \{0, 1 - r\}.$$

Thus, the left-hand side of (18) is minimized by a matrix with the specified eigenvalues and first eigenvector equal to $b/\sqrt{b'b}$. The required matrix is

$$\Phi = I + \left[\frac{\max \{0, 1 - r\} - 1}{r} \right] \frac{bb'}{a}.$$

and this corresponds to the optimal C given in the theorem. \square

Theorem 1 demonstrates that making no change in the complementary space (*i.e.*, $C = I_{n-1}$) does, in fact, produce a least change update. The theorem also provides a larger estimate for C when needed to preserve nonnegative definiteness. Our implementation of SQN uses a safe-guarded choice of λ_{SQN} to prevent the Hessian estimate from becoming singular. See Equation (7).

Behind the intuition that one should make small alterations to the Hessian estimate in the complementary space lies a principle that accuracy obtained on previous iterations should be preserved as much as possible. The following proposition demonstrates that the SQN update achieves the goal of preserving Hessian accuracy in a certain sense. Section 7.1 explores this topic further in the context of a statistical model for C and shows that the SQN update is, in fact, a Bayesian estimate of the Hessian under a prior distribution designed to preserve accuracy built into the previous Hessian estimate.

Proposition 1 (SQN Accuracy Preservation) *If the true Hessian in canonical coordinates is positive definite and given by*

$$\tilde{B}_+ = \begin{bmatrix} a & b' \\ b & C_{\text{TRUE}} \end{bmatrix}, \quad (19)$$

then $C_{\text{SQN}} \equiv I_{n-1} + \lambda_{\text{SQN}} bb'/a$ is at least as accurate as I_{n-1} for estimating C_{TRUE} in any direction either parallel to b or orthogonal to b . That is,

$$\left| u' \left(C_{\text{SQN}} - C_{\text{TRUE}} \right) u \right| \leq \left| u' \left(I_{n-1} - C_{\text{TRUE}} \right) u \right| \quad (20)$$

for any u such that either $u'b = 0$ or $u \propto b$. Furthermore, this is not necessarily true for any larger estimate $\hat{C} = C_{\text{SQN}} + VV'$ where V is any non-zero matrix with $n - 1$ rows.

See Appendix A for a proof.

The following proposition provides the canonical form for the well-known Broyden family and shows that SQN updates are particular members.

Proposition 2 (Canonical Broyden Updates) *Under the canonical transform (12) the Broyden update (4) transforms to*

$$\tilde{B}_+ = \begin{bmatrix} a & b' \\ b & I_{n-1} + \lambda bb'/a \end{bmatrix} \quad (21)$$

where $\lambda = 1 + c/(\delta'\gamma)$. In particular, the usual Broyden parameter is $\phi = (\lambda - 1)a$ and important special cases are given as follows:

<i>Method</i>	λ	ϕ
<i>SQN</i>	$\max\{0, 1 - r^{-1}\}$	$\max\{-a, -ar^{-1}\}$
<i>BFGS</i>	1	0
<i>DFP</i>	$1 + a^{-1}$	1
<i>E_I</i>	$1 - a^{-1}(a + r)^{-1}$	$-(a + r)^{-1}$

where if $r = 0$ the max is taken to be the first argument.

See Appendix B for a proof. The proof also provides formulae for a , b and r in terms of the usual quantities δ , γ and B .

The E_I value for λ can fall below the critical point $\lambda^c = 1 - r^{-1}$ and thus produce an indefinite update. Minimizing $\|E_I\|_F$ over nonnegative definite updates has the effect of truncating Greenstadt's solution at the critical value.

Although BFGS minimizes several different metrics of change (see Fletcher, 1991), Proposition 2 indicates that BFGS *increases* the lower right block over its previous value of I_{n-1} whereas

SQN leaves it unchanged if possible, or adds a fraction of the BFGS correction in order to preserve nonnegative definiteness. The conclusion of Proposition 1 is that neither BFGS nor DFP preserves the accuracy of the previous Hessian estimate. It is interesting that both BFGS and DFP explode as a becomes small.

4 Step Size Estimation

In trial experiments with the SQN update, we carried out the quasi-Newton M-step using a line search in which the initial step size was unity; that is, the line search used an initial evaluation point of $x - sB^{-1}g$ with $s = 1$, which is the Newton step under the assumption that B is the actual Hessian. The experiments demonstrated that the SQN update tended to reduce the number of iterations to convergence compared to BFGS but did not consistently reduce the number of function evaluations required. Further investigation showed the reason: unit steps are often too long when the SQN update is used. Zhang and Tewarson’s (1988) steepest descent method (SDQN) also uses negative Broyden parameters and they state, “*SDQN tends to give steps longer than BFGS steps, and therefore is more likely to violate the [sufficient decrease] condition.*” When unit steps are used, fewer iterations seem to come with the price of more function evaluations per iteration. Some numerical results with unit step sizes are reported in Section 6.

Why do negative Broyden parameters produce steps that are too long? A rough explanation is that a negative Broyden parameter produces a smaller Hessian estimate than BFGS. Compare $\lambda < 1$ in Proposition 2 with $\lambda = 1$. A smaller B implies a longer unit step $-B^{-1}g$. Therefore, if unit steps are suitable for BFGS, then unit steps may well be too long for use with negative Broyden parameters. This reasoning is admittedly rough; it does not account for differences in the step *direction* and it does not provide guidance for selecting more appropriate step sizes. This section proposes a Wishart model to describe uncertainty of the unknown Hessian and then derives an estimate of the optimal step size as a function of the Broyden parameter used in updating the Hessian. The SQN initial step size (8) is a special case.

4.1 A Wishart model for the Hessian matrix

The unknown Hessian $\tilde{G}_+ = \tilde{G}(\tilde{x}_+)$ can be modeled as a random matrix whose probability distribution quantifies the plausibility of all possible canonical Hessians. The distribution of \tilde{G}_+ should naturally be centered at the previous estimate $\tilde{B} = I_n$. Although \tilde{G}_+ likely has greater accuracy in the directions of recent steps, it is reasonable to model \tilde{G}_+ with equal uncertainty in every direction because the previous step directions are not available in the quasi-Newton framework. The most basic statistical model for a symmetric positive-definite matrix with expected value I_n is the

Wishart model:

$$\nu \tilde{G}_+ \sim \text{Wishart}_n(I_n, \nu), \quad (22)$$

where $\nu \geq n + 1$ is the degrees of freedom parameter. The distribution of \tilde{G}_+ becomes more concentrated around I_n as ν increases. See, *e.g.*, Anderson (1984, p. 244–257) for the definition and properties of the Wishart family. The probability density function of \tilde{G}_+ is proportional to

$$|\tilde{G}_+|^{(\nu-n-1)/2} \exp \left\{ -\frac{\nu}{2} \text{tr}(\tilde{G}_+) \right\}. \quad (23)$$

Because (23) only involves \tilde{G}_+ through its determinant and trace, any orthogonal rotation, $R' \tilde{G}_+ R$ where $R'R = I_n$, is distributed identically to \tilde{G}_+ . This *directional symmetry* seems an appropriate requirement for modeling the Hessian in canonical coordinates.

In the quasi-Newton framework, the first row and column of \tilde{G}_+ are considered to be known from the numerical second derivatives (14). Therefore

$$\tilde{G}_+ = \begin{bmatrix} a & b' \\ b & C \end{bmatrix}, \quad (24)$$

where a and b are observed and C is not. Standard Wishart theory provides the conditional distribution $[C|a, b]$ through

$$\nu \left[C - \frac{bb'}{a} \middle| a, b \right] \sim \text{Wishart}_{n-1}(I_{n-1}, \nu - 1).$$

The conditional expectation and mode are

$$E(C|a, b) = \frac{\nu - 1}{\nu} I_{n-1} + \frac{bb'}{a}, \quad (25)$$

$$\text{Mode}(C|a, b) = \frac{\nu}{\nu - n - 1} I_{n-1} + \frac{bb'}{a}. \quad (26)$$

The two multipliers on I_{n-1} depend on the degrees of freedom, ν , and they differ because the Wishart model is skewed toward large positive definite matrices. But both coefficients approach unity as $\nu \rightarrow \infty$, which is called the *large-sample* limit.

Comparing (25) and (26) to (21) in Proposition 2 shows that the the large-sample conditional expectation and mode under a Wishart model are exactly equal to the BFGS update. Specifically, let $B_+(\lambda)$ denote the Broyden update (4)–(6) with parameter λ and let $\tilde{B}_+(\lambda)$ denote the corresponding canonical form given by (21). Then

$$\begin{aligned} \lim_{\nu \rightarrow \infty} E(G_+|a, b) &= B^{1/2} U \left[\lim_{\nu \rightarrow \infty} E(\tilde{G}_+|a, b) \right] U' B^{1/2} \\ &= B^{1/2} U \tilde{B}_+(1) U' B^{1/2} \\ &= B_+(1) \end{aligned} \quad (27)$$

which is the BFGS update. Section 7.1 discusses why (27), the simplest statistical estimate, is not used in the *Statistical* quasi-Newton method and then derives the SQN update from a generalized version of the Wishart model.

4.2 Optimal step size

An estimate of the optimal step size for any given Broyden update can be derived from the Wishart model. Let d_+ represent an arbitrary search direction to be taken in the M-step on iteration $k + 1$. A second order Taylor expansion of $f(\cdot)$ about the point x_+ gives the quadratic approximation

$$f(x_+ + sd_+) \approx f(x_+) + sd'_+g_+ + \frac{s^2}{2}d'_+G_+d_+ \quad (28)$$

with optimum step size

$$s^* = \frac{-d'_+g_+}{d'_+G_+d_+}. \quad (29)$$

obtained by differentiating (28) with respect to s and setting the result to zero. The denominator of (29) involves the unknown Hessian but an estimate of s^* can be obtained by replacing G_+ with its large-sample conditional expectation from (27):

$$\lim_{\nu \rightarrow \infty} E(G_+|a, b) = B_+(1) = B_+(\lambda) + (1 - \lambda)(\delta'\gamma)\omega\omega' \quad (30)$$

where (4) has been used with to express $B_+(1)$ in terms of a general Broyden update. The resulting optimum step size is obtained by plugging (30) into (29) and taking $d_+ = -B_+^{-1}(\lambda)g_+$, the next quasi-Newton step direction:

$$\hat{s}(\lambda) = \frac{g'_+B_+^{-1}(\lambda)g_+}{g'_+B_+^{-1}(\lambda)g_+ + (1 - \lambda)(\delta'\gamma)(g'_+B_+^{-1}(\lambda)\omega)^2}. \quad (31)$$

This is the step size formula (8) of the SQN algorithm. The estimated optimum for BFGS is $\hat{s}(1) = 1$ which suggests that unit steps may work better for BFGS than for any other Broyden update.

Results comparing BFGS to the SQN algorithm using (31) are shown in Figure 2 and demonstrate that SQN achieves consistent reduction in function evaluations, as well as iteration counts and gradient evaluations compared to BFGS. Additional comparisons to SQN using unit steps are shown for three new test functions in Section 6.

The inequality in (8) indicates that the estimated optimum step size is at most 1 for the SQN choice of λ . This is true for any $\lambda \leq 1$. Similarly, $\hat{s}(\lambda) \geq 1$ for any $\lambda \geq 1$. The following proposition, proved in Appendix C, gives tight bounds on $\hat{s}(\lambda)$.

Proposition 3 *For the step estimate $\hat{s}(\lambda)$ given by (31),*

$$\begin{aligned} \lambda \in (1 - r^{-1}, 1] &\Rightarrow \hat{s}(\lambda) \in [1 - (1 - \lambda)r, 1] \\ \lambda > 1 &\Rightarrow \hat{s}(\lambda) \in [1, 1 - (1 - \lambda)r] \end{aligned}$$

Furthermore, $\hat{s}(\lambda) = 1 - (1 - \lambda)r$ if and only if $\tilde{g}_+ \equiv U'B^{-1/2}g(x_+) \propto (0, b)'$.

The proposition implies that \hat{s} can get arbitrarily close to zero only if $(1 - \lambda)r$ is close to 1 — that is, only if λ is close to its critical value, $\lambda^c = 1 - r^{-1}$. Even in this case, the actual value of \hat{s} depends on how closely the new gradient \tilde{g}'_+ is aligned with $(0, b)'$.

Table 1: More, Garbow and Hillstrom unconstrained test functions

index	function	m	n tested
1	Helical valley	3	3
2	Biggs Exponential	13	6
3	Gaussian	15	3
4	Powell badly scaled	2	2
5	Box	10	3
6	Variably dimensioned	n+2	4, 8, 16, 32, 64, 128
7	Watson	31	6, 9, 12
8	Penalty I	n+1	4, 10
9	Penalty II	2n	4, 10
10	Brown badly scaled	3	2
11	Brown and Dennis	20	4
12	Gulf Research	100	3
13	Trigonometric	n	4, 8, 16, 32, 64, 128
14	Rosenbrock	n	4, 8, 16, 32, 64, 128
15	Powell singular	n	4, 8, 16, 32, 64, 128
16	Beale	3	2
17	Wood	6	4
18	Chebyquad	n	4, 6, 8

5 Performance Evaluation

Figures 1 and 2 demonstrate that SQN performs better than BFGS. This section describes the details of the optimization setup and the problems tested. The numerical study closely follows that of Zhang and Tewarson (1988) but not in every detail. Section 6 uses three new test functions to delineate more precisely the cases in which SQN is expected to outperform other Broyden updates.

5.1 Moré test problems and convergence criterion

Table 1 lists the eighteen functions of Moré, Garbow and Hillstrom (1981) for testing unconstrained optimization algorithms. Each function is a nonlinear sum of squares of the form

$$f(x) = \sum_{i=1}^m f_i^2(x)$$

where $x \in R^n$. The final two columns in Table 1 give the values of m and n used in the numerical comparisons. The choices for n follow Zhang and Tewarson. However, these authors did not report

their choices for m so in cases with variable m we used either values for which Moré et. al provide the minimum function value (functions 2, 11 and 18) or arbitrary values (functions 5 and 12).

Moré et. al provide standard starting points $x_S \in R^n$ for each problem and suggest setting $x_0 = x_S \times (\text{start factor})$ for various start factors. Nominally we used start factors of 1(1)10, but in some cases results are provided on only a subset of these values. On problem 7, $x_S = 0$ so the start factor is irrelevant. On problem 12 the point $10x_S$ is the minimum so only 1(1)9 are used. The only other exceptions are some start factors on functions 2, 13 and 16 for which BFGS and SQN converged to different solutions so these start factors are omitted as reported in Section 5.3.

The Moré test problems have diverse features that challenge optimization algorithms, such as multiple local minima (*e.g.*, function 12) or very flat regions near the solutions (*e.g.*, function 4). When comparing algorithms it is desirable that they converge not only to the same function value but also to the same point. For this reason we use the following two stage procedure to assess convergence. For each test problem and starting point after running both SQN and BFGS until either no further progress can be made or to a maximum of 2000 iterations, the best point x_* achieved by either algorithm is identified. The resulting iteration traces are then truncated at the first k for which

$$[f(x_k) - f(x_*)] + |(x_k - x_*)'g(x_*)| + |(x_k - x_*)'G(x_*)(x_k - x_*)| < 10^{-9} [1 + |f(x_*)|]. \quad (32)$$

The three terms on the left-hand side are the positive parts of a quadratic Taylor expansion around x_* . If (32) is not achieved by one of the algorithms, then the given starting point is omitted from our comparisons. This typically occurs when different algorithms converge to different local minima so the third term on the left, and possibly the first term, remain large. Criterion (32) is a generalization of Gill and Murray's (1979) assessment criterion; it requires the gradient to be small and both the minimum and minimizer to agree with those of the best algorithm. The absolute value is used in the third term because in rare cases the true Hessian $G(x_*)$ can be singular but numerically indefinite.

5.2 Line search and quasi-Newton details

We use Fletcher's (1987, pp. 33-38) line search algorithm with the tunable parameters set to the values that he suggests $(\bar{f}, \tau_1, \tau_2, \tau_3) = (0, 9, 0.1, 0.5)$ and with tolerances of $(c_1, c_2) = (10^{-4}, 0.9)$ in the Wolf conditions (2) and (3). The initial step size is given by (8) unless stated otherwise. The only modification made to Fletcher's algorithm is that steps $x_{k+1} - x_k$ are restricted to have a maximum Euclidean length of 10^6 , a precaution that Zhang and Tewarson (1988) found useful but is rarely imposed.

On rare occasions the line search along $-B_k^{-1}g_k$ fails to achieve a point that satisfies the Wolfe conditions because of finite numerical precision. In this case an additional line search is attempted in the steepest descent direction with initial step $-g_k \text{tr}(B_k^{-1})/n$. Over all the cases in the study

only seven steepest descent steps were used prior to (32) being satisfied, and these occurred only for BFGS on Chebyquad with $n = 8$.

For programming convenience and computational efficiency, Broyden updates of the inverse Hessian were implemented using the well-known dual form of equation (4). See, for example, Nocedal and Wright (1999) for details. The initial estimate is $B_0^{-1} = I_n$. A rare phenomenon due to finite precision is that the update can produce an inverse Hessian that is numerically indefinite and the new search direction can be non-decreasing. When this occurs we take

$$B_{k+1}^{-1} \leftarrow B_{k+1}^{-1} + \epsilon g_{k+1}' g_{k+1}'$$

where ϵ is chosen so that the new inverse Hessian satisfies

$$g_{k+1}' B_{k+1}^{-1} g_{k+1} = 10^{-4} g_{k+1}' g_{k+1}$$

to produce a new search direction that is slightly decreasing. Such modifications were required 11 times for BFGS on Chebyquad with $n = 8$ and a total of 26 times for SQN on Chebyquad with $n = 6$ and 8.

5.3 Results on Moré problems

Figure 1 plots SQN iteration counts relative to BFGS for test problems of small dimension. Table 2 gives further details on these runs. The start factors, listed in the third column, are those for which both SQN and BFGS converged to the same point as determined by (32). In most cases this is 1(1)10. The columns of BFGS counts give the averages (over start factors) of the numbers of iterations, function evaluations (f) and gradient evaluations (g) until convergence. The final three columns report the ratios of average counts for SQN divided by those for BFGS.

Table 3 compares SQN to BFGS using four of the Moré problems in which the dimension is taken to increase by powers of 2 from 4 to 128, as in Zhang and Tewarson (1988, Table 4). SQN generally performs better than BFGS but the efficiency on these four problems is not as good as on some of the problems with small dimension.

Although Table 3 does not show a strong dependency of SQN efficiency on problem size, the plot in Figure 2 of all data from Tables 2 and 3 demonstrates that the relative efficiency of SQN improves as problems get more difficult. Substantial variation occurs over different problems but the trend is clear and applies not only to iteration counts but also to function and gradient evaluations.

Table 4 shows that SQN compares favorably with results of other published studies that use negative Broyden parameters. Only rough comparisons are valid in the table because of some differences in the test functions used and in the manner of reporting results. The table lists overall average efficiencies of SQN alongside comparable quantities reported by other authors on the Moré problems. The most equivalent comparisons are to the SDQN and LCCB methods of Zhang and

Table 2: SQN performance relative to BFGS on problems of small dimension. BFGS counts are averages over the start factors. SQN/BFGS columns contain ratios of count averages to indicate SQN efficiency.

index	n	start factors	BFGS counts			SQN / BFGS		
			iter.	f	g	iter.	f	g
1	3	1(1)10	26.6	39.7	29.0	1.02	1.00	1.01
2	6	1,2,3,4,6,7,9	34.6	42.4	38.4	0.86	0.90	0.88
3	3	1(1)10	13.9	19.3	15.2	0.85	1.10	1.03
4	2	1(1)10	111.7	164.0	125.1	0.90	0.94	0.91
5	3	1(1)10	22.1	37.0	31.4	0.71	0.90	0.88
7	6	1	31.0	39.0	32.0	0.90	0.92	0.91
7	9	1	66.0	77.0	67.0	0.56	0.70	0.66
7	12	1	127.0	140.0	128.0	0.44	0.66	0.63
8	4	1(1)10	114.0	162.6	126.2	0.84	0.87	0.88
8	10	1(1)10	172.2	236.1	186.8	0.85	0.91	0.90
9	4	1(1)10	434.9	596.3	476.4	0.91	0.98	0.93
9	10	1(1)10	404.3	566.2	436.0	0.77	0.83	0.81
10	2	1(1)10	14.1	25.1	18.1	0.96	1.02	1.02
11	4	1(1)10	35.5	54.9	36.7	0.85	0.90	0.86
12	3	1(1)9	29.0	39.8	32.6	0.80	0.92	0.88
16	2	1,2,3,5,7,10	20.8	31.7	24.7	0.76	0.88	0.85
17	4	1(1)10	75.5	105.3	80.5	0.93	1.00	0.97
18	4	1(1)10	69.5	98.8	75.5	0.75	0.88	0.85
18	6	1(1)10	173.0	227.0	183.3	0.61	0.78	0.73
18	8	1(1)10	264.5	340.5	278.1	0.58	0.83	0.75
Average of 20:						0.79	0.90	0.87

Table 3: SQN performance relative to BFGS on four problems as dimension increases. The layout is the same as in Table 2

index	n	start factors	BFGS counts			SQN / BFGS		
			iter.	f	g	iter.	f	g
6	4	1(1)10	9.7	15.0	11.4	0.98	0.99	0.98
6	8	1(1)10	14.0	21.6	15.8	0.96	0.97	0.96
6	16	1(1)10	20.8	27.8	22.1	0.97	0.97	0.97
6	32	1(1)10	23.4	31.0	25.3	1.02	1.01	1.02
6	64	1(1)10	28.0	35.3	29.7	0.98	0.99	0.99
6	128	1(1)10	34.3	41.0	35.6	1.00	1.00	1.00
13	4	1(1)6	18.0	24.5	19.2	0.93	0.94	0.92
13	8	1,3,4,6,8,9,10	34.6	46.4	36.7	0.92	0.97	0.93
13	16	1,2,3,4,6,7,8	44.3	54.0	46.1	0.88	0.91	0.89
13	32	1,2,3,4,7,8,9	43.7	51.4	45.6	0.91	0.91	0.90
13	64	1,2,3,7,8	41.0	49.4	43.0	0.97	0.98	0.97
13	128	1,2,3,7	32.2	36.8	33.8	0.94	0.95	0.93
14	4	1(1)10	72.1	103.7	78.5	0.90	0.96	0.95
14	8	1(1)10	123.8	172.9	130.6	0.87	0.95	0.93
14	16	1(1)10	151.7	220.9	158.2	0.90	0.98	0.95
14	32	1(1)10	171.2	279.0	176.7	0.80	0.89	0.83
14	64	1(1)10	272.7	481.1	281.1	0.79	0.88	0.82
14	128	1(1)10	358.5	678.1	365.8	0.86	0.93	0.88
15	4	1(1)10	37.5	49.6	39.4	0.92	0.94	0.93
15	8	1(1)10	65.8	85.0	67.7	0.88	0.96	0.94
15	16	1(1)10	111.0	144.7	114.0	0.87	0.95	0.91
15	32	1(1)10	173.4	230.3	175.2	0.95	0.99	0.97
15	64	1(1)10	305.2	408.4	308.5	0.91	0.95	0.91
15	128	1(1)10	588.4	763.4	590.9	0.86	0.91	0.87
Average of 24:						0.92	0.95	0.93

Tewarson (1988) because our study was patterned after theirs. However, they do not report function and gradient evaluations separately but roll them into

$$\text{EFE} = (\text{function count}) + n \times (\text{gradient count}).$$

When we tested the SQN update using unit steps, its EFE efficiency was consistently better than BFGS but the function evaluation efficiency was not. Method I is that of Byrd, Liu, and Nocedal (1992, Table 5) which they offer as setting a performance standard for replacing BFGS with a competitor. However, Method I is not practical as a quasi-Newton update because it requires evaluation of the true Hessian matrix. A big difference in Byrd, et. al’s test setup is that they used only single starting points, so their results could be more affected by atypical performance on a single problem. Also their numbering of functions does not match the original Moré numbering, so the set of functions tested is not completely known.

Table 4: Published efficiencies of various negative Broyden updates relative to BFGS¹.

Count	Small Dimension			Method I	Increasing Dimension		
	SQN	SDQN	LCCB		SQN	SDQN	LCCB
iteration	0.79	0.79	0.93	0.82	0.92	0.87	0.94
function eval.	0.90	—	—	0.88	0.95	—	—
gradient eval.	0.87	—	—	—	0.93	—	—
EFE	0.87	0.85	0.94	—	0.93	0.89	0.95

1. Only rough comparisons are valid because of differences in the function sets, starting points, convergence criteria and other implementation details.

6 Three New Test Functions

The Moré test problems have become standard for comparing quasi-Newton algorithms but they are not particularly useful for empirically validating our claim that BFGS tends to inflate B_k and that SQN is more neutral. This section uses three new test functions for that purpose.

Byrd, Liu and Nocedal (1992) found that BFGS is lopsided: it can more readily increase Hessian estimates that are too small than shrink ones that are too large. This was surprising in light of the strong “self-correcting” property of the BFGS update that was established by Byrd, Nocedal and Yuan (1987): the relative error between the curvature predicted by B_k and the curvature observed in the current line search is transmitted exactly to the relative change of the determinant from $|B_k|$ to $|B_{k+1}|$. Proposition 2, on the other hand, shows that BFGS corrections actually inflate B_k in the space canonically orthogonal to the search direction whereas SQN corrections leave that part of the Hessian unchanged (subject to positive definiteness), and therefore, should cope equally well

with estimates that need to shrink as ones that need to grow. Furthermore, choosing λ_k to be less than 0 or greater than 1 should make these effects more pronounced.

To test this understanding, we employ three new test functions, f^{dec} , f^{inc} and f^{sin} with respective Hessians that decrease, increase and change sinusoidally as x_k moves toward the optimum value. We also implement a range of Broyden updates with

$$\lambda_k = \max \{ \lambda_{\text{NOM}}, 1 - (1 - \epsilon)r^{-1} \}$$

where $\epsilon = 10^{-6}$ and $\lambda_{\text{NOM}} = 0$ is SQN, $\lambda_{\text{NOM}} = 1$ is BFGS and more extreme values of λ_{NOM} should produce more extreme effects.

General forms of the new test functions are defined in Appendix D but this section uses specific versions with parameters $n = 4$, $U = I_4$, $(\eta_1, \eta_2, \eta_3, \eta_4) = (1, 2, 4, 8)$ and, for f^{sin} , $\rho = 1$. Each function is convex with a minimum value of 0 at $(0, 0, 0, 0)$. The Hessians are diagonal and the i -th diagonal elements, for $(i = 1, 2, 3, 4)$, are

$$\begin{aligned} G_{ii}^{\text{dec}}(x) &= 1 + (\eta_i x_i)^2 \\ G_{ii}^{\text{inc}}(x) &= [1 + (\eta_i x_i)^2]^{-1} \\ G_{ii}^{\text{sin}}(x) &= 1 + \sin(\eta_i x_i). \end{aligned}$$

At the optimum each Hessian is the identity. The values of η_i scale how quickly curvature changes in each of the coordinate directions.

6.1 Results for different λ_{NOM}

The rationale for testing with functions whose Hessians change monotonically (f^{dec} , f^{inc}) or unpredictably (f^{sin}) is to verify our claim that BFGS needlessly inflates the previous Hessian estimate whereas SQN treats it neutrally. With f^{inc} , for example, the most appropriate Hessian estimate in iteration $k + 1$ will tend to be larger than in iteration k . BFGS could have an advantage over SQN because it tends to inflate the Hessian beyond its previous value in the complementary space. In this case, the best choice of λ_{NOM} should be larger than 0 and possibly even larger than 1, the BFGS value. For f^{dec} , on the other hand, SQN should have the advantage over BFGS and the optimal λ_{NOM} should be negative. For f^{sin} , there is no consistent pattern for the Hessian on one step to be either larger or smaller than on the previous step so $\lambda_{\text{NOM}} = 0$ (*i.e.*, SQN) should be nearly optimal.

Figure 3 plots average counts to convergence as a function of λ_{NOM} with each panel representing one of the new test functions. Each plotted symbol represents an average count over 1000 random starting points. The vertical scales are set to support relative comparisons, the most obvious of which is that λ_{NOM} has the greatest effect for f^{dec} and the least for f^{sin} . Iterations, function evaluations and gradient evaluations are shown using different plotting symbols. Initial step sizes are given by (8). The true value is used for the starting Hessian estimate, $B_0 = G(x_0)$.

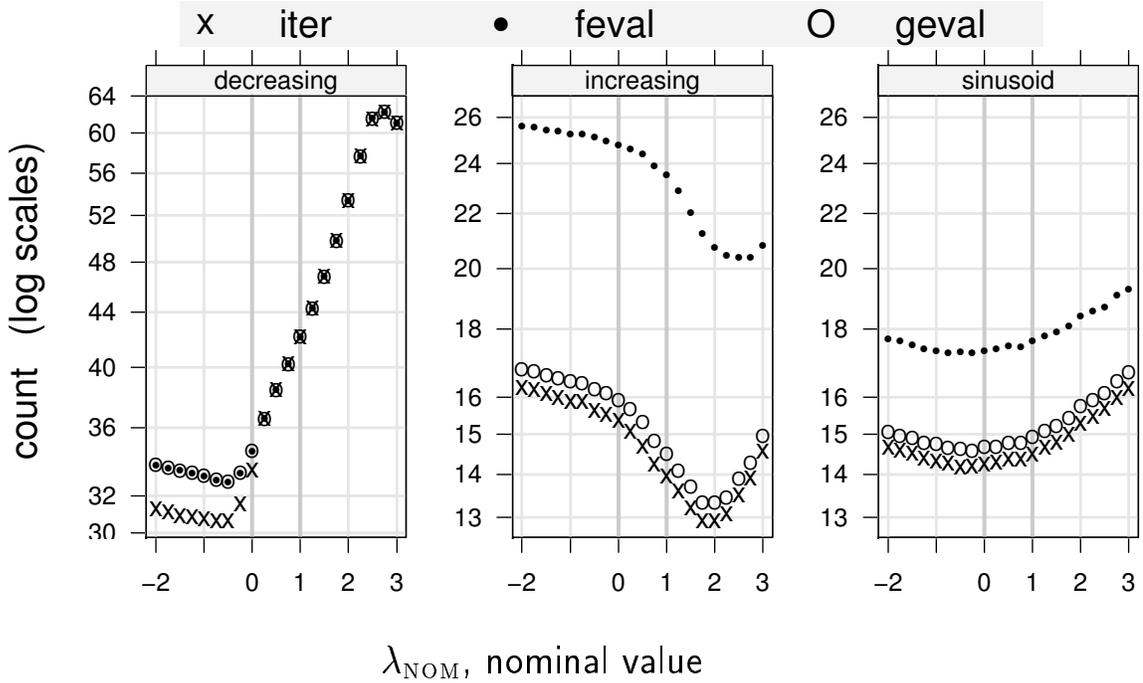


Figure 3: Performance counts versus λ_{NOM} on test functions with Hessians that are decreasing, increasing and sinusoidal as x_k moves toward the minimum. Different symbols are used for iterations, function evaluations and gradient evaluations. Initial steps are estimated using Equation (8). The special value $\lambda_{\text{NOM}} = 0$ is SQN and $\lambda_{\text{NOM}} = 1$ is BFGS.

The starting points x_0 were chosen at random in such a way that they tend to be oriented in the direction of $(\eta_1, \eta_2, \eta_3, \eta_4)$. Specifically, the i th component of x_0 was drawn randomly as

$$x_{0,i} = K\eta_i(1 + z_i/3)$$

where the z_i are independent $N(0, 1)$ random variables and the scale was set as $K = 200$ for f^{dec} , $K = 50$ for f^{inc} , and $K = 1000\|\eta\|$ for f^{sin} . These choices reflect a little experimentation aimed at producing differences between BFGS and SQN that are large enough to be interesting without requiring unwieldy numbers of iterations. As far as we know, other choices for η and the starting vectors produce similar results, though we have not studied this extensively. Convergence was declared when $f(x_k) < 10^{-10}$. The more elaborate assessment criterion (32) was not needed in this study because the three functions are well-behaved.

For f^{dec} , Figure 3 demonstrates that SQN is indeed better able to cope with a decreasing Hessian than BFGS and further improvement is obtained by using slightly negative values of λ_{NOM} .

The situation is reversed for f^{inc} . BFGS handles the increasing Hessian better than SQN and further improvement is obtained by taking λ_{NOM} as large as 2. Finally, for f^{sin} the Hessian changes arbitrarily and the SQN update ($\lambda_{\text{NOM}} = 0$) is nearly optimal.

Several additional comments on these results are worth noting. First, in each panel all three curves have nearly the same shape. But function evaluations are always equal to gradient evaluations on f^{dec} whereas they are substantially higher on f^{inc} and f^{sin} . This indicates that the step size estimate is better for f^{dec} than for the other two functions. Second, for any $\lambda_{\text{NOM}} < 1$ some values of λ_k will likely exceed λ_{NOM} because of the requirement that B_{k+1} remain positive definite. This produces an asymmetry in the results, so that the performance differences between $\lambda_{\text{NOM}} = -1$ and 0 are not as great as the differences between 0 and 1. In fact, our selection of starting points that are biased in the direction of η was made to enhance the effect of λ_{NOM} below 1 on f^{inc} and f^{sin} . The patterns in Figure 3 are smooth because they average across 1000 starting points. If counts from a single starting point were plotted, the patterns for f^{inc} and f^{sin} would be virtually impossible to discern because of noise in the data. In this case it would be meaningless to make comparisons on only a few test cases.

6.2 Results for different step sizes

Figure 4 demonstrates the importance of using estimated step sizes, especially with $\lambda_{\text{NOM}} < 1$. The experiment is the same as in Figure 3 except that the algorithm was also run with unit initial step sizes. The plots compare average function evaluation counts for unit steps against those for estimated steps. In each panel as λ_{NOM} decreases from 1 (BFGS) the unit step results become much worse than the results with estimated steps. The same appears to be true as λ_{NOM} becomes positive and large. The curves intersect at $\lambda_{\text{NOM}} = 1$ because the estimated step size is 1.

At $\lambda_{\text{NOM}} = 0$ (SQN) the results of Figure 4 are most revealing on f^{inc} . In this case the SQN Hessian estimate tends to be too small so that unit step sizes are too large. Estimated step sizes are smaller and perform much better although they may still be too large as indicated in Figure 3 by the gap between the number of function and gradient evaluations. The only case where unit steps perform substantially better than estimated ones is on f^{inc} with $1 < \lambda_{\text{NOM}} < 3$. These values of λ_{NOM} inflate the Hessian estimates more than BFGS. We suspect that the inflated Hessians are producing estimated steps that are too short. Significantly, estimated steps are *uniformly* better than unit steps on f^{sin} , for which Hessian changes are fairly unpredictable.

7 Discussion

This paper has investigated two estimation problems that arise in the design of quasi-Newton algorithms: (1) estimation of Newton directions by way of sequential updates to a Hessian estimate; and (2) estimation of the optimum along a given search direction. SQN solves the two problems

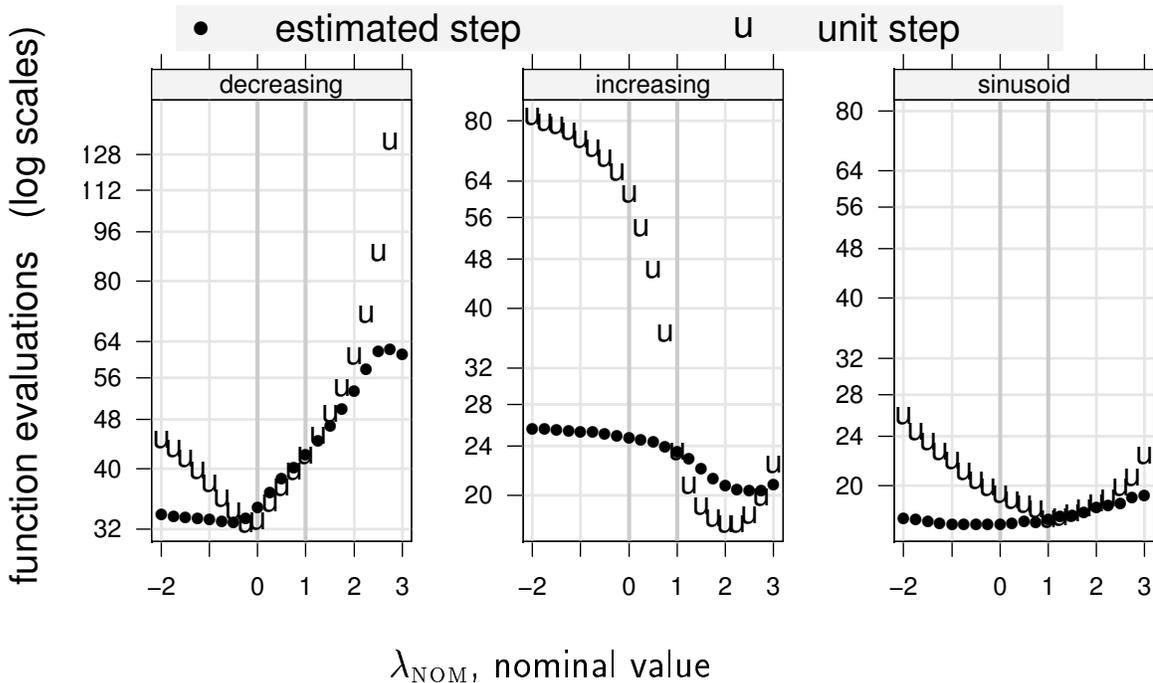


Figure 4: Function evaluation counts verses λ_{NOM} for three test functions. The plots compare performance with unit initial steps against estimated initial steps using Equation (8) The dots in this figure are the same as in Figure 3.

rather differently, using a least-change principle for the Hessian update and a statistical model for the step size. This raises the question of why the statistical model is not also used for the Hessian update.

In fact, the Wishart model leads to BFGS. This is seen in Equation (27) showing that the BFGS update is the large-sample conditional expectation of the Hessian given the most recently obtained curvature information. Another derivation of BFGS is obtained by taking the negative logarithm of the Wishart density (23), dividing by $\nu/2$ and taking the large-sample limit, $\nu \rightarrow \infty$. The result is the following metric:

$$\psi(\tilde{B}_+) \equiv \text{tr}(\tilde{B}_+) - \ln|\tilde{B}_+|$$

Fletcher (1991) demonstrated that BFGS minimizes $\psi(\tilde{B}_+)$. Thus, the BFGS update is both the conditional mean and mode estimate under a large-sample Wishart model.

7.1 Least change and statistical estimation

The reason for using a least-change principle rather than conditional expectation to derive the SQN update is that Hessian estimation proceeds sequentially. While, statistical theory says that the conditional expectation (BFGS) is the optimal estimator *within a given iteration* under a squared-error loss function, this theory is not directly applicable to sequential estimation of Hessians. A conservative strategy for estimating the Hessian is to preserve the accuracy obtained in previous iterations while incorporating new curvature information obtained in the current iteration—that is, to use a least change principle. *Preserving information from previous iterations enables future estimates of the Hessian to be more accurate.*

Estimating the optimal step size is a fundamentally different problem because the appropriate step size changes on each iteration and there is little reason to expect that an appropriate size in one iteration will be relevant to the next. Thus, it makes sense to estimate the optimal step size by using a conditional expectation as we did in deriving (31).

It is possible, however, to obtain the SQN update through a statistical model (as opposed to the least-change approach in Section 4. The statistical derivation begins with the following.

Statistical Least Change Principle:

For the purpose of updating a Hessian estimate, the uncertainty model for the Hessian in canonical form (15) should be augmented with a prior distribution that strongly concentrates C near I_{n-1} without modifying the distribution of (a, b)

The prior distribution called for by this principle does not have the usual Bayesian meaning of a model for ones prior belief about C ; rather, it penalizes departures of C from I_{n-1} as a mechanism to force the new Hessian estimate to be close to the old one.

Appendix E derives the SQN update by generalizing the Wishart model (22) in accord with the above Statistical Least Change Principle. This framework is then used to derive the SQN Hessian estimate from the conditional distribution $[C|a, b]$. Although this derivation of SQN uses different mathematics from the Frobenius norm derivation in Section 3, the fundamental reasoning is the same as enumerated in Section 2.2—namely, it is important to preserve accuracy that has been built into the Hessian estimate on previous iterations by minimizing the *relative change* from B to B_+ as measured by E_B .

The SQN formulation of the least-change problem is closely related to Broyden’s (1965) desire to make no-change in directions orthogonal to the search direction δ . SQN applies Broyden’s idea to the directions that are orthogonal in a canonical sense—that is, directions q such that $q'B^{-1}\delta = 0$.

Viewing Broyden updates under the canonical transformation sheds light on why both BFGS and DFP tend to *inflate* the Hessian estimate: they add a multiple of bb'/a in the complementary space as shown in Proposition 2. This interpretation is born out in the empirical results on new test problems reported in Section 6.

7.2 Possible extensions

Use of a statistical framework to design a quasi-Newton method motivates several interesting topics. The numerical results on three new test functions suggest that information on the bias of previous Hessian estimates could be captured and used to obtain a better update that uses either varying values of λ_{NOM} within the Broyden family or a self-scaling update outside of the Broyden family. Use of the Wishart model to estimate the optimal step size also suggests a more general class of quasi-Newton methods obtained by searching, not in the estimated Newton direction $-B^{-1}g$ but rather in an alternative direction determined from the conditional distribution $[-B^{-1}g|a, b]$. We have obtained promising results in some limited tests of these ideas.

Acknowledgment

We are grateful to a number of colleagues, including J. Chambers, D. Gay, D. Lambert, C. Mallows, and M. Wright, for helpful discussion.

Appendix: Proofs of Theoretical Results

A. Proof of Proposition 1

If $r \leq 1$ then $C_{\text{SQN}} = I_{n-1}$ and (20) holds as an equality for *all* u . Suppose $r > 1$ so that $C_{\text{SQN}} = I_{n-1} + (1 - r^{-1})a^{-1}bb'$. Then for any $u : u'b = 0$,

$$u'C_{\text{SQN}}u = u'I_{n-1}u$$

and thus (20) holds as an equality. Suppose $u = \rho b$ for some $\rho \neq 0$. Positive definiteness of the true Hessian implies $(C_{\text{TRUE}} - a^{-1}bb') > 0$ and thus

$$\begin{aligned} u'C_{\text{TRUE}}u &> a^{-1}u'bb'u &= \rho^2(b'b)r \\ &= u'C_{\text{SQN}}u &> \rho^2(b'b) = u'I_{n-1}u > 0. \end{aligned}$$

That is, in the direction of u , C_{SQN} is closer to C_{TRUE} than I_{n-1} is, and this implies that (20) holds as a strict inequality.

To prove the final statement, suppose that $C_{\text{TRUE}} = I_{n-1}$ so the right-hand side of (20) equals zero and consider two cases as follows. First suppose that $\|V'b\| > 0$ and take $u = \rho b$ with $\rho \neq 0$. Then

$$u'(\hat{C} - C_{\text{TRUE}})u = \rho^2b'(\lambda a^{-1}bb' + VV')b > 0$$

and (20) is violated. On the other hand, if $\|V'b\| = 0$ then assume, without loss of generality, that V has full column rank and take $u = V(V'V)^{-1}y$ for some vector $y \neq 0$. Then $u'b = y'(V'V)^{-1}V'b = 0$

but

$$u'(\hat{C} - C_{\text{TRUE}})u = y'(V'V)^{-1}V'(VV')V(V'V)^{-1}y = y'y > 0$$

which violates (20).

B. Proof of Proposition 2

Proof. Using (13), the relation between B_+ and \tilde{B}_+ is given by $B_+ = B^{1/2}U\tilde{B}_+U'B^{1/2}$. This can be expressed as follows:

$$\begin{aligned} B_+ &= B^{1/2}U \begin{bmatrix} a & b' \\ b & I_{n-1} + \lambda bb'/a \end{bmatrix} U'B^{1/2} \\ &= B + B^{1/2}U \begin{bmatrix} a-1 & b' \\ b & \lambda bb'/a \end{bmatrix} U'B^{1/2} \\ &= B + B^{1/2}U(D_1 + D_2 + D_3)U'B^{1/2}, \end{aligned} \tag{33}$$

where

$$D_1 = \begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix}, \quad D_2 = \begin{bmatrix} a^2/a & b' \\ b & bb'/a \end{bmatrix} \quad \text{and} \quad D_3 = \begin{bmatrix} 0 & 0 \\ 0 & a(\lambda-1)bb'/a^2 \end{bmatrix}.$$

Denote by $U[, 1]$ the first column of U . Then,

$$U[, 1] = \frac{B^{1/2}\delta}{(\delta'B\delta)^{1/2}}, \quad \begin{bmatrix} a \\ b \end{bmatrix} = \frac{U'B^{-1/2}\gamma}{(\delta'B\delta)^{1/2}},$$

$$a = \frac{\delta'\gamma}{\delta'B\delta}, \quad \text{and} \quad r \equiv \frac{b'b}{a} = \frac{\gamma'B^{-1}\gamma}{\delta'\gamma} - \frac{\delta'\gamma}{\delta'B\delta}$$

Simple algebraic operations lead to the following equalities

$$B^{1/2}UD_1U'B^{1/2} = -B^{1/2}U[, 1](U[, 1])'B^{1/2} = -\frac{B\delta\delta'B}{\delta'B\delta},$$

$$B^{1/2}UD_2U'B^{1/2} = \frac{1}{a}B^{1/2}U \begin{bmatrix} a \\ b \end{bmatrix} [a, b]U'B^{1/2} = \frac{\gamma\gamma'}{\delta'\gamma},$$

and

$$\begin{aligned} B^{1/2}UD_3U'B^{1/2} &= \frac{a(\lambda-1)}{a^2}B^{1/2}U \left(\begin{bmatrix} a \\ b \end{bmatrix} - \begin{bmatrix} a \\ 0 \end{bmatrix} \right) \left(\begin{bmatrix} a \\ b \end{bmatrix} - \begin{bmatrix} a \\ 0 \end{bmatrix} \right)' U'B^{1/2} \\ &= (\lambda-1)(\delta'\gamma) \left(\frac{\gamma}{\delta'\gamma} - \frac{B\delta}{\delta'B\delta} \right) \left(\frac{\gamma}{\delta'\gamma} - \frac{B\delta}{\delta'B\delta} \right)'. \end{aligned}$$

From these equalities and (33), we see that the expression for B_+ is identical to (4) with $c = (\lambda-1)(\delta'\gamma)$.

C. Proof of Proposition 3

Substituting $B_+(1)$ for G_+ in (29) and using $d_+ = -B_+^{-1}(\lambda)g_+$ gives

$$\hat{s}(\lambda) = \frac{-d'_+g_+}{d'_+B_+(1)d_+} = \frac{d'_+B_+(\lambda)d_+}{d'_+B_+(1)d_+}.$$

The denominator can be written as

$$d'_+B_+(1)d_+ = d_+B^{1/2}U\tilde{B}_+(1)U'B^{1/2}d_+ = y'[I_n + (1 - \lambda)R]y$$

with

$$y \equiv \tilde{B}_+^{1/2}(\lambda)U'B^{1/2}d_+ \quad \text{and} \quad R \equiv \tilde{B}_+^{-1/2}(\lambda) \begin{pmatrix} 0 & 0 \\ 0 & bb'/a \end{pmatrix} \tilde{B}_+^{-1/2}(\lambda).$$

Thus,

$$\hat{s}(\lambda) = \left[1 + (1 - \lambda) \frac{y'Ry}{y'y} \right]^{-1}.$$

R has a single (potentially) non-zero eigenvalue

$$\eta = \frac{r}{1 - (1 - \lambda)r}$$

with corresponding eigenvector

$$w \propto \tilde{B}_+^{1/2}(\lambda) \begin{pmatrix} -r \\ b \end{pmatrix}$$

as is seen by noting that the following are equivalent:

$$\begin{aligned} Rw &= \eta w \\ \tilde{B}_+^{1/2}(\lambda)Rw &= \eta\tilde{B}_+^{1/2}(\lambda)w \\ \begin{pmatrix} 0 & 0 \\ 0 & bb'/a \end{pmatrix} \begin{pmatrix} -r \\ b \end{pmatrix} &= \eta \begin{pmatrix} a & b' \\ b & I_{n-1} + \lambda bb'/a \end{pmatrix} \begin{pmatrix} -r \\ b \end{pmatrix}. \end{aligned}$$

Simple algebra establishes the final equality. Note that $\eta > 0$ for $\lambda > 1 - r^{-1}$.

The eigen-decomposition of R implies that $0 \leq y'Ry/(y'y) \leq \eta$ and this is equivalent to the bounds on $\hat{s}(\lambda)$ stated in the proposition. The extreme case $\hat{s}(\lambda) = 1 - (1 - \lambda)r$ occurs if and only if $y \propto w$ which is equivalent to

$$\begin{aligned} \tilde{B}_+^{1/2}(\lambda)w &\propto -\tilde{B}_+^{1/2}(\lambda)y \\ &= -\tilde{B}_+(\lambda)U'B^{1/2}d_+ \\ &= \tilde{B}_+(\lambda)U'B^{1/2}B_+^{-1}(\lambda)g_+ \\ &= U'B^{-1/2}B_+(\lambda)B^{-1/2}UU'B^{1/2}B_+^{-1}(\lambda)g_+ \\ &= U'B^{-1/2}g_+ \equiv \tilde{g}_+. \end{aligned}$$

Thus,

$$\tilde{g}_+ \propto \tilde{B}_+^{1/2}(\lambda)w = \tilde{B}_+(\lambda) \begin{pmatrix} -r \\ b \end{pmatrix} \propto \begin{pmatrix} 0 \\ b \end{pmatrix}.$$

D. Three test functions

Three convex functions are constructed. The first, denoted $f^{\text{dec}}(x)$, has a Hessian matrix that becomes smaller as x moves in the direction of the optimal point x_* . The second function, $f^{\text{inc}}(x)$, has a Hessian matrix that becomes larger as x moves in the direction of the optimal point x_* . The third function, $f^{\text{sin}}(x)$, has a Hessian matrix that behaves in a sinusoidal fashion as x moves to the optimal point x_* .

D.1 Decreasing-Hessian function

Define a quartic function as follows:

$$f^{\text{dec}}(x) = \frac{1}{2}x'x + \frac{1}{12} \sum_{i=1}^n \eta_i^2 (u_i'x)^4 \quad (x \in \mathcal{R}^n),$$

where $\eta_i \geq 0$ are known scalars and u_1, \dots, u_n are the n columns of a fixed n -dimensional unitary matrix $U = (u_1, \dots, u_n)$ so that $U'U = I_n$. The gradient of $f^{\text{dec}}(x)$ is

$$g^{\text{dec}}(x) = x + \frac{1}{3} \sum_{i=1}^n \eta_i^2 (u_i'x)^3 u_i,$$

for which the equation $g^{\text{dec}}(x) = 0$ has the unique solution $x_* = 0$. The Hessian matrix of $f^{\text{dec}}(x)$ at x is

$$G^{\text{dec}}(x) = \sum_{i=1}^n [1 + \eta_i^2 (u_i'x)^2] u_i u_i',$$

which satisfies both $G^{\text{dec}}(x) > 0$ and $G^{\text{dec}}(x) - G^{\text{dec}}(x_*) \geq 0$. So the quartic function $f^{\text{dec}}(x)$ is convex and its Hessian decreases as x moves to x_* .

D.2 Increasing-Hessian function

The following function is constructed so that its Hessian matrix at any x is the inverse of $G^{\text{dec}}(x)$. Thus, the constructed function is also convex but its Hessian matrix increases as x moves to x_* . The function is defined as

$$f^{\text{inc}}(x) = \sum_{i=1}^n f_i(x) \quad (x \in \mathcal{R}^n),$$

where

$$f_i(x) = \begin{cases} \frac{1}{2}(u'_i x)^2 & \text{if } \eta_i = 0, \\ \frac{1}{\eta_i^2} [(\eta_i u'_i x) \arctan(\eta_i u'_i x) - \frac{1}{2} \ln(1 + (\eta_i u'_i x)^2)] & \text{if } \eta_i > 0 \end{cases}$$

for $i = 1, \dots, n$. The gradient of $f^{\text{inc}}(x)$ is

$$g^{\text{inc}}(x) = \begin{cases} (u'_i x) u_i & \text{if } \eta_i = 0, \\ \sum_{i=1}^n \frac{1}{\eta_i} \arctan(\eta_i u'_i x) u_i & \text{if } \eta_i > 0. \end{cases}$$

The Hessian matrix of $f^{\text{inc}}(x)$ is

$$G^{\text{inc}}(x) = \sum_{i=1}^n \frac{1}{1 + \eta_i^2 (u'_i x)^2} u_i u'_i.$$

Note that all the three functions $f^{\text{inc}}(x)$, $g^{\text{inc}}(x)$, and $G^{\text{inc}}(x)$ are continuous with respect to η_i at the value $\eta_i = 0$

D.3 Sinusoidal Hessian function

Consider the function

$$f^{\text{sin}}(x) = \sum_{i=1}^n \left[\frac{1}{2} (u'_i x)^2 + \frac{\rho}{\eta_i^2} (\eta_i u'_i x - \sin(\eta_i u'_i x)) \right] \quad (x \in \mathcal{R}^n),$$

where $\rho \in (-1, 1)$ and $\eta_i \geq 0$ are given scalars, and u_1, \dots, u_n are the n columns of an n -dimensional unitary matrix $U = (u_1, \dots, u_n)$. The gradient is

$$g^{\text{sin}}(x) = x + \sum_{i=1}^n \frac{\rho [1 - \cos(\eta_i u'_i x)]}{\eta_i} u_i,$$

and the Hessian is

$$G^{\text{sin}}(x) = \sum_{i=1}^n [1 + \rho \sin(\eta_i u'_i x)] u_i u'_i.$$

G^{sin} has a sinusoidal behavior as x moves toward x_* .

E. Derivation of SQN from a Wishart Prior

The Wishart model (22) provides a joint distribution for the blocks $[a, b, C]$. The model can be described hierarchically in terms of the distributions $[a|b, C]$, $[b|C]$ and $[C]$. Standard Wishart theory provides the following:

$$\nu [(a - b' C^{-1} b) | b, C] \sim \chi_{\nu - n + 1}^2, \quad (34)$$

$$\left[\nu^{1/2} C^{-1/2} b \middle| C \right] \sim \text{Normal}_{n-1} (0, I_{n-1}), \quad (35)$$

$$\nu C \sim \text{Wishart}_{n-1} (I_{n-1}, \nu). \quad (36)$$

To implement the Statistical Least Change Principle, we replace (36) with

$$(\nu_0 + \nu)C \sim \text{Wishart}_{n-1}(I_{n-1}, \nu_0 + \nu) \quad (37)$$

where $\nu_0 \geq 0$ denotes the prior degrees of freedom. Taking ν_0 to be arbitrarily large drives $[C]$ to be arbitrarily concentrated around I_{n-1} . The posterior distribution $[C|a, b]$ will also be arbitrarily close to I_{n-1} if this does not contradict that \tilde{G}_+ is positive definite. But for the case where I_{n-1} is *not* a valid estimate of C , the generalized Wishart model (34), (35), and (37) provides the necessary framework to derive a posterior estimate of C that is as close to I_{n-1} as possible while maintaining a positive definite update formula. An important feature of (37) is that it retains directional symmetry in the space of C with the consequence that Hessian uncertainty is modeled as equal in every direction (canonically) orthogonal to the current step direction.

Although the model with $\nu_0 > 0$ is not standard the following proposition derives two estimates, \hat{C}_\pm , from the posterior distribution of $C - a^{-1}bb'$. In the limit as $\nu_0 \rightarrow \infty$ both estimates converge to the SQN update.

Proposition 4 *If the joint distribution $[a, b, C]$ is given by (34), (35) and (37) and if $b \neq 0$, then the modes of $[C - a^{-1}bb'|a, b]$ and of $[(C - a^{-1}bb')^{-1}|a, b]$ are unique and the corresponding estimates of C are*

$$\hat{C}_\pm = \left(1 \mp \frac{n+1}{\nu_0 + \nu}\right) I_{n-1} + \left(\frac{\mu_\pm}{r}\right) \frac{bb'}{a},$$

where

$$\mu_\pm \equiv \frac{1}{2} \left[r - 1 \pm \frac{n+1}{\nu_0 + \nu} + \sqrt{\left(1 + r \mp \frac{n+1}{\nu_0 + \nu}\right)^2 - \frac{4r\nu_0}{\nu_0 + \nu}} \right] \quad (38)$$

and $r \equiv b'b/a$. Furthermore,

$$\lim_{\nu_0 \rightarrow \infty} \hat{C}_\pm = I_{n-1} + \lambda_{\text{SQN}} \frac{bb'}{a}$$

where λ_{SQN} is given by (17).

Proof. From (34) and (35), the conditional density of $[C|a, b]$ is proportional to

$$(a - b'C^{-1}b)^{\frac{\nu-n-1}{2}} |C|^{-\frac{1}{2}} \propto \left| C - \frac{bb'}{a} \right|^{\frac{\nu-n-1}{2}} |C|^{-\frac{\nu-n}{2}} \quad (C > 0).$$

Multiplying this by the density for C corresponding to (37) produces the posterior density of $[C|a, b]$, which, written in terms of $\Phi \equiv C - a^{-1}bb'$, is proportional to

$$|\Phi|^{\frac{\nu-n-1}{2}} \left| \Phi + \frac{bb'}{a} \right|^{\frac{\nu_0}{2}} \exp \left\{ -\frac{\nu_0 + \nu}{2} \text{tr}(\Phi) \right\} \quad (\Phi > 0). \quad (39)$$

To determine the mode of (39) let $\mu_1 < \dots < \mu_{n-1}$ denote the eigenvalues of Φ , which are necessarily positive. Then

$$\left| \Phi + \frac{bb'}{a} \right| = |\Phi|(1 + a^{-1}b'\Phi^{-1}b) \leq |\Phi| \left(1 + \frac{r}{\mu_1} \right) = (\mu_1 + r) \prod_{i=2}^{n-1} \mu_i$$

and therefore (39) is bounded above by

$$\left[\mu_1^{\frac{\nu-n-1}{2}} (\mu_1 + r)^{\frac{\nu_0}{2}} e^{-\frac{\nu_0+\nu}{2}\mu_1} \right] \prod_{i=2}^{n-1} \left(\mu_i^{\frac{\nu_0+\nu-n-1}{2}} e^{-\frac{\nu_0+\nu}{2}\mu_i} \right). \quad (40)$$

Furthermore, the bound is achieved if and only if the first eigenvector (corresponding to μ_1) is proportional to b . Taking the log of (40) and setting derivatives to zero provides the maximizers $\mu_1 = \mu_+$ (defined in 38) and $\mu_i = (\nu_0 + \nu - n - 1)/(\nu_0 + \nu)$ for $i = 2, \dots, n - 1$. The unique matrix that has this set of eigenvalues, $\{\mu_i\}$, and a first eigenvector proportional to b is $\hat{\Phi}_+ \equiv \hat{C}_+ - a^{-1}bb'$. Therefore $\hat{\Phi}_+$ uniquely maximizes (39) and \hat{C}_+ is the corresponding estimate of C .

Since the Jacobian of the transformation $\Phi \rightarrow \Phi^{-1}$ is $J(\Phi \rightarrow \Phi^{-1}) = |\Phi|^{n+1}$, the posterior density of $(C - a^{-1}bb')^{-1} = \Phi^{-1}$ is proportional to

$$|\Phi|^{\frac{\nu+n+1}{2}} \left| \Phi + \frac{bb'}{a} \right|^{\frac{\nu_0}{2}} \exp \left\{ -\frac{\nu_0 + \nu}{2} \text{tr}(\Phi) \right\} \quad (\Phi > 0). \quad (41)$$

and \hat{C}_- can be derived using the same approach taken for \hat{C}_+ . Finding the limiting values of \hat{C}_\pm is straightforward when the cases $r \leq 1$ and $r > 1$ are taken separately.

References

1. ANDERSON, T. W. (1984). *An Introduction to Multivariate Statistical Analysis* (second edition). John Wiley & Sons, New York.
2. BROYDEN, C. G. (1965). A class of methods for solving nonlinear simultaneous equations. *Math. Comp.*, **19**, 577-593.
3. BROYDEN, C. G. (1967). Quasi-Newton methods and their applications to function minimization, *Maths. of Computation*, **21**, 368-381.
4. BROYDEN, C. G. (1970). The convergence of a class of double rank minimization algorithms: 2. The new algorithm. *J. Inst. Math. Appl.* **6** 222-231.
5. BROYDEN, C. G. (2000). On the discovery of the ‘‘good Broyden’’ method. *Math. Program. Ser. B* **87** 209-213.
6. BYRD, R. H., LIU, D. C., AND NOCEDAL, J. (1992). On the behavior of Broyden’s class of quasi-Newton methods. *SIAM J. Optimization* **2**, No. 4, 533-557.

7. BYRD, R. H. AND NOCEDAL, J. (1989). A tool for the analysis of quasi-Newton methods with application to unconstrained minimization. *SIAM J. Numer. Anal.*, **26**, 727-739.
8. BYRD, R. H., NOCEDAL, J. AND YUAN, Y. (1987). Global convergence of a class of quasi-Newton methods on convex problems. *SIAM J. of Numer. Anal.* **24**, 1171-1190.
9. CHAMBERS, J. M. (1977). *Computational Methods for Data Analysis*. John Wiley & Sons, New York.
10. CLEVELAND, W. S. (1979). Robust locally weighted regression and smoothing scatter plots. *Journal of American Statistical Association*, **74**, 829-836.
11. CROCKETT, J. B. AND CHERNOFF, H. (1955). Gradient method of maximization. *Pacific J. Math.* **5** 33-50.
12. DAVIDON, W. C. (1959). Variable metric methods for minimization. *A. E. C. Res. and Develop. Report ANL-5990* and in *SIAM J. Optimization* **1** (1991) 1-17.
13. DAVIS, P. J. (1968). Gamma function and related functions. In *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Seventh Printing, eds. Ailton Abramowitz and Irene, A. Stegun, U.S. Department of Commerce, National Bureau of Standards.
14. FLETCHER, R. (1970). A new approach to variable metric algorithms. *Computer J.* **13** 317-322.
15. FLETCHER, R. (1987). *Practical Methods of Optimization, 2nd Ed.*, John Wiley & Sons, New York.
16. FLETCHER, R. (1991). A new variational result for quasi-Newton formulae. *SIAM J. Optimization* **1** 18-21.
17. FLETCHER, R. (1994). An overview of unconstrained optimization. In *Algorithms for Continuous Optimization: The State of the art*, ed. Emilio Spedicato, 109-143, Kluwer Academic Publishers, Netherlands.
18. FLETCHER, R. AND POWELL, M. J. D. (1963). A rapidly convergent descent method for minimization. *Comput. J.* **6** 163-168.
19. GILL, P. E., AND MURRAY, W. (1979). Performance evaluation for nonlinear optimization. In *Performance Evaluation for Numerical Software* (FOSDICK, L. D., ED.), 221-234. North-Holland, Amsterdam.

20. GOLDFARB, D. (1970). A family of variable metric methods derived by variational means. *Math. Comp.* **24** 23-26.
21. GREENSTADT, J. (1970). Variations on variable metric methods, *Math. Comp.* **24** 1-22.
22. GREENSTADT, J. (2000). Reminiscences on the development of the variational approach to Davidon's variable-metric method, *Math. Program. Ser. B* **87** 265-280.
23. LUKŠAN, L. (1992). Computational Experience with Known Variable Metric Updates. Institute of Computer Science, Academy of Sciences of the Czech Republic, Prague, Czech Republic. (Technical Report No. V-534.)
24. LUKŠAN, L., SPEDICATO, E. AND VLČEK (1999). Variable metric methods for unconstrained optimization, Institute of Computer Science, Academy of Sciences of the Czech Republic, Prague, Czech Republic. (Technical Report No. 780.)
25. MIFFLIN, R. B. AND NAZARETH, J. L. (1994). The least prior deviation quasi-Newton update. *Mathematical Programming*, **65**, 247-261.
26. MORÉ, J. J., GARBOW, B. S. AND HILLSTROM, K. E. (1981). Testing Unconstrained Optimization Software. *ACM Transactions on Mathematical Software*, **7**, 17-41.
27. NOCEDAL, J. AND WRIGHT, S. J. (1999). *Numerical Optimization*, Springer, New York.
28. SHANNO, D. F. (1970). Conditioning of quasi-Newton methods for function minimization. *Math. Comp.* **24** 647-650.
29. ZHANG, Y. AND TEWARSON R. P. (1988). Quasi-Newton algorithms with updates from the preconvex part of Broyden family. *IMA Journal of Numerical Analysis* **8**, 487-509.