SIAM J. OPTIM.
Vol. 18, No. 4, pp. 1266–1285

STATISTICAL QUASI-NEWTON: A NEW LOOK AT LEAST CHANGE*

CHUANHAI LIU† AND SCOTT A. VANDER WIEL‡

Abstract. A new method for quasi-Newton minimization outperforms BFGS by combining least-change updates of the Hessian with step sizes estimated from a Wishart model of uncertainty. The Hessian update 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 shorter steps estimated from the Wishart model. In numerical comparisons to BFGS the new *statistical quasi-Newton (SQN)* algorithm typically converges with about 25% fewer iterations, functions, and gradient evaluations on the top 1/3 hardest unconstrained problems in the CUTE library. Typical improvement on the 1/3 easiest problems is about 5%. The framework used to derive SQN provides a simple way to understand differences among various Broyden updates such as BFGS and DFP and shows that these methods do not preserve accuracy of the Hessian, in a certain sense, while the new method does. In fact, BFGS, DFP, and all other updates with nonnegative 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

AMS subject classifications. 65K10, 90C53

DOI. 10.1137/040614700

1. Introduction.

Quasi-Newton methods for unconstrained optimization are important computational tools in many scientific fields and are a standard subject in textbooks on computation. The BFGS method, proposed individually in [6], [14], [20], and [30], is implemented in most optimization software and is widely recognized as efficient. Generalizations of BFGS are available for large problems with memory limitations, for problems with bound constraints, and for a parallel computing environment. In theoretical investigations BFGS is known as a special case of the Broyden class [5]. Some Broyden updates with negative Broyden parameters have been found to produce faster convergence than BFGS updates [31], [8] 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 [29] 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

*Received by the editors September 9, 2004; accepted for publication (in revised form) April 9, 2007; published electronically October 24, 2007. Most of this work was done while the authors were in the Statistics Research Department, Bell Labs, Lucent Technologies. This work was performed by an employee of the U.S. Government or under U.S. Government contract. The U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. Copyright is owned by SIAM to the extent not limited by these rights.

http://www.siam.org/journals/siopt/18-4/61470.html

†Statistics and Data Mining Research, Bell Labs, 700 Mountain Avenue, Murray Hill, NJ 07974, and Department of Statistics, Purdue University, 150 N. University Street, West Lafayette, IN 47907 (chuanhai@stat.purdue.edu).

‡Statistics and Data Mining Research, Bell Labs, 700 Mountain Avenue, Murray Hill, NJ 07974, and Statistical Sciences Group, MS F600, Los Alamos National Laboratory, Los Alamos, NM 87545 (scottv@lanl.gov).

questions: What is the "best" negative Broyden parameter? and What initial step sizes should be used with negative Broyden parameters? This paper answers these questions by solving a least-change problem to approximate Newton directions and by estimating step sizes through a statistical model of Hessian uncertainty. We call the new algorithm *statistical quasi-Newton* (SQN).

1.1. Quasi-Newton methods. Quasi-Newton methods solve the unconstrained optimization problem

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

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 $G(x_0)$. Evaluate $g_0 = g(x_0)$ and iterate over $k = 0, 1, 2, \ldots$ the following two steps.

M-Step. Search in the direction $-B_k^{-1} g_k$ for a step size $s_k > 0$ to obtain a new evaluation point and gradient,

$$x_{k+1} = x_k - s_k B_k^{-1} g_k, \qquad g_{k+1} = g(x_{k+1}),$$

that satisfy *the Wolfe conditions* for sufficient decrease of the function and for curvature (see (2) and (3) below).

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*

$$B_{k+1}\delta_k = \gamma_k, \tag{1}$$

where

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

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. Various 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*,

$$f(x_{k+1}) \le f(x_k) - \rho_1 s_k g_k' B_k^{-1} g_k \qquad (\rho_1 \in (0,1), \text{ say } \rho_1 = 10^{-4}), \tag{2}$$

requires a reduction in $f(x)$ that is at least a fraction ρ_1 of that predicted by the directional derivative $-g_k B_k^{-1} g_k$. The Wolfe *strong curvature condition*,

$$|g'_{k+1}(B_k^{-1}g_k)| \le \rho_2 g'_k(B_k^{-1}g_k) \qquad (\rho_2 \in (\rho_1, 1), \text{ say } \rho_2 = 0.9), \tag{3}$$

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-hand side of (3). Nocedal and Wright [29] discuss the importance of the Wolfe conditions in ensuring that sufficient progress is made on each iteration.

The best-known class of Hessian estimates used in the E-step are the rank-two Broyden updates [5]:

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

where

$$\omega_k \equiv \frac{\gamma_k}{\delta'_k \gamma_k} - \frac{B_k \delta_k}{\delta'_k B_k \delta_k} \tag{5}$$

and c_k is a scalar parameter to be specified. 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

$$c_k = (\lambda_k - 1)\,(\delta'_k \gamma_k)\,, \tag{6}$$

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

$$r_k \equiv \frac{\gamma'_k B_k^{-1} \gamma_k}{\gamma'_k \delta_k} - \frac{\delta'_k \gamma_k}{\delta'_k B_k \delta_k}. \tag{7}$$

It can be shown that $r_k \ge 0$ 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 SQN. The SQN method is remarkably simple and effective. This section briefly defines SQN and demonstrates its superiority to BFGS. Derivations and additional experimental results are provided in the following sections.

SQN algorithm. Follow the generic quasi-Newton algorithm with the following additional specifications. Initialize $\hat{s}_0 = 1$.
M-Step. Begin the line search from an initial evaluation point $x_k - \hat{s}_k B_k^{-1} g_k$.
E-Step. Estimate the Hessian using a Broyden update (4)–(6) with parameter

$$\lambda_k = \max\{0, 1 - (1-\epsilon) r_k^{-1}\}, \tag{8}$$

where ϵ is a small positive constant (e.g., $\epsilon = 10^{-6}$) and if $r_k = 0$, the max is taken to be 0. Estimate the next step size as

$$\hat{s}_{k+1} = \frac{g'_{k+1} B_{k+1}^{-1} g_{k+1}}{g'_{k+1} B_{k+1}^{-1} g_{k+1} + (1-\lambda_k)(\delta'_k \gamma_k)(g'_{k+1} B_{k+1}^{-1} \omega_k)^2} \;<\; 1. \tag{9}$$

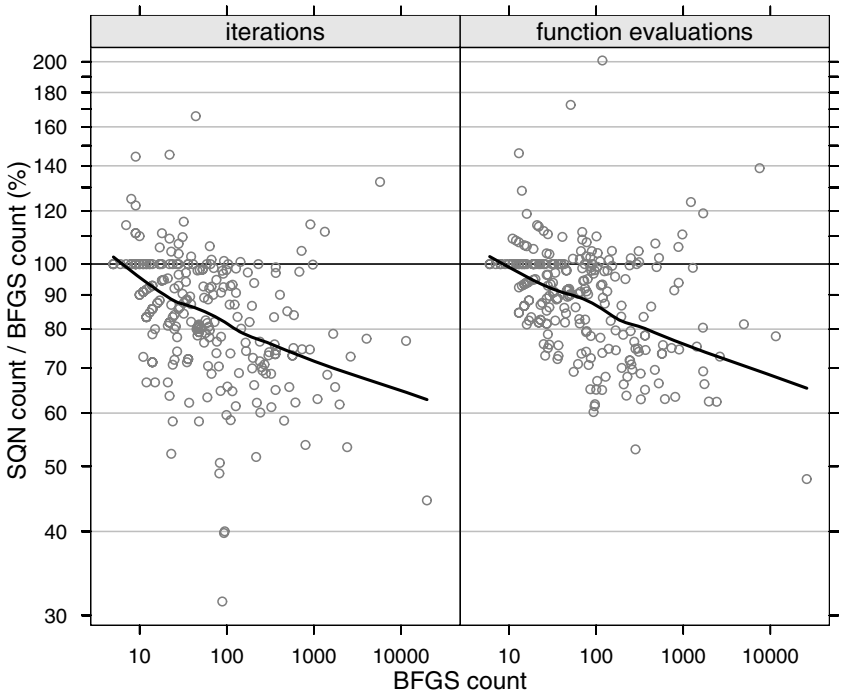

FIG. 1. *Improvement in SQN efficiency with problem difficulty for iterations, function evaluations, and gradient evaluations. Each point represents performance of SQN and BFGS on a given problem from the standard starting point.*

All the quantities needed to calculate $\hat{s}_{k+1}$ are readily available from the preceding M-step with no extra function or gradient evaluations required. Using $\epsilon > 0$ guarantees that B_{k+1} remains positive definite. The Broyden parameter corresponding to λ_k is $\phi_k = (\lambda_k - 1)\,(\delta_k'\gamma_k) \,/\, (\delta_k' B_k \delta_k)$, and this is negative because (3) implies $\delta_k'\gamma_k > 0$. Equation (9) is written in terms of the inverse Hessian estimate because Broyden updates are typically implemented on the inverse scale using the well-known dual form of (4). Similarly, $B_k\delta_k = -s_k g_k$ can be substituted into (7) for computational efficiency. See, for example, [29].

The shortened initial step size, $\hat{s}_k$, is crucial to improving the performance of Broyden updates with negative Broyden parameters. Zhang and Tewarson [31] use $\hat{s} = 1$ and comment that their negative Broyden algorithm improves iteration counts but that "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 shows that SQN typically converges with substantially fewer iterations and function evaluations than BFGS on 248 unconstrained optimization problems in the CUTE [3] suite. The left panel plots SQN iterations as a percent of BFGS iterations against BFGS iterations. The right panel shows the same information for function evaluations. SQN becomes more efficient relative to BFGS on the more difficult problems, as additional iterations offer additional opportunities for improvement. Performance on easy problems with few iterations is often dominated by the first iteration in which a poor choice of B_0 produces a poor search vector for any quasi-Newton

TABLE 1
Median percent improvement of SQN relative to BFGS by difficulty of problem.

	Easy	Medium	Hard
Iterations	8	14	26
Function evaluations	4	9	23

algorithm. In harder problems these start-up effects wash out so that the advantage of SQN over BFGS becomes more apparent. The trend curves in Figure 1 highlight this tendency. The smooth curves are robust local regressions [11] that follow the data without being unduly influenced by the low outlying points that would tend to make the trends even stronger.

Table 1 summarizes the improvement by splitting the test problems into three equal groups, *easy*, *medium*, and *hard*, according to the number of iterations for BFGS to converge. SQN's median improvement over BFGS is largest for the hardest 1/3 of the test problems, 25% in round numbers.

Our setup uses the line search [28] available from Argonne National Lab at ftp://info.mcs.anl.gov/pub/MINPACK-2/csrch in MINPACK-2. This line search evaluates the function value and gradient an equal number of times. The starting point x_0 is as given in the CUTE collection, and the initial Hessian estimate is $B_0 = c \cdot I_n$, where c is the geometric mean of the positive diagonal elements of the true Hessian at x_0. This is similar to the usual choice of $B_0 = I_n$, but scaling by c provides a more fair comparison because the true Hessian at x_0 tends to be much larger than I_n on the CUTE problems, and this gives an unfair advantage to BFGS, which has a bias toward inflating the Hessian estimate, as explained in section 3 below. For each test problem and starting point both SQN and BFGS are run until no valid step is found due to finite numerical precision. Then the best point x_* achieved by either algorithm is identified, and convergence is retrospectively declared at the first k for which

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

This generalization of the assessment criterion [19] ensures that both the optima and the optimizers match.

The comparisons reported in Figure 1 and Table 1 are based on 248 problems in the CUTE collection. The test set consists of all unconstrained problems with maximum dimension of 500 that have continuous analytic second derivatives and compile with the default "large" version of the CUTE software. Of the 306 that fit these criteria, 15 appear to start at the optimum, 23 converge to a better local minimum with SQN than with BFGS, and 20 converge to a better minimum with BFGS. Removing these 58 cases leaves 248 test problems that support clean comparisons between SQN and BFGS.

We also conducted an initial study of the SQN algorithm, patterned after [31] using 20 of the test problems [27], each with 10 starting points. The results were similar: about 20% fewer iterations and gradient evaluations and about 10% fewer function evaluations compared to BFGS. This initial study used Fletcher's line search algorithm [15] with the tunable parameters set as suggested and utilizing his "sensible" choices for trial step lengths based on minimizing interpolating polynomials.

SQN compares favorably to other studies that have used negative Broyden parameters. Zhang and Tewarson [31] report 21% and 13% fewer iterations for their SDQN method relative to BFGS on problems of small and "increasing" dimension, respectively. However, their improvements were smaller using the EFE metric that

incorporates the number of function evaluations. Byrd, Liu, and Nocedal [8] report improvements of 18% on iterations and 12% on function evaluations for a smaller set of tests using their Method I, which is not practical as a quasi-Newton update because it requires evaluation of $G(x)$.

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 the change between consecutive Hessian estimates. Section 3 introduces a transformation into canonical coordinates, derives (4)–(8) as the new least-change update, and shows that it preserves Hessian accuracy from one iteration to the next in a certain sense. Section 4 introduces a Wishart model to describe Hessian uncertainty and derives (9) as an estimate of the optimal step size. Section 5 compares performance on three new test functions designed to verify our understanding of why SQN is better than other Broyden methods. Section 6 explores connections to other least-change derivations and mentions ideas for future research.

2. Least-change updates. Fletcher's overview [17] 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. This leads to the SQN update that was introduced in section 1.2. Although the SQN update happens to be in the Broyden class, it is derived in section 3 by minimizing change over *all possible* quasi-Newton updates.

2.1. Historical developments. Crockett and Chernoff [12] 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 [4] as articulated in [7] 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 [5] 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 [21], 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 [25] 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 [20]

worked with the scaled difference of inverse Hessian estimates

$$E_W^* \equiv W^{1/2}\left(B_{k+1}^{-1} - B_k^{-1}\right)W^{1/2}, \tag{11}$$

where the symmetric matrix W satisfies $W\delta_k = \gamma_k$. He derived the BFGS update by using the results in [21] to minimize the Frobenius norm $\|E_W^*\|_F \equiv [\mathrm{tr}\,(E_W^* E_W^*)]^{1/2}$ 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

$$E_W \equiv W^{-1/2}\left(B_{k+1} - B_k\right)W^{-1/2} \tag{12}$$

produces the Broyden update with $\lambda_k = 1+\delta_k' B_k \delta_k/(\delta_k'\gamma_k)$. This is known as DFP [13], [18] and is generally regarded as inferior to BFGS.

Fletcher [14] 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, [31], [8], [23], [17], and [26]. 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. Robust improvement over BFGS has been elusive. Indeed, Zhang and Tewarson [31] concluded that such investigations have not shaken the position of BFGS as the most popular front-runner.

2.2. A new measure for least change. Minimizing the change from B_k to B_{k+1} is a generally accepted principle. There is no agreement, however, on how to measure that change. Zhao [32] derives 10 different optimal updates by considering five possible matrix norms applied to two different matrices that measure change. The function 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.

A sensible matrix measure of change that has received little attention in the literature is the difference $B_{k+1} - B_k$ scaled by the current estimate B_k, namely

$$E_B \equiv B_k^{-1/2}\left(B_{k+1} - B_k\right)B_k^{-1/2}. \tag{13}$$

Normalizing a difference is appropriate because, in every direction, E_B measures change of the Hessian estimate *relative to current nominal value*, and this produces a scale-free method. Greenstadt [22] states that such normalization "renders harmless the accidents of coordinate selection in a given problem." One possible danger in making E_B small is that B_{k+1} could become singular (or even indefinite if allowed), and this could produce unstable quasi-Newton search vectors, based on B_{k+1}^{-1}. However, applying no direct penalty to large differences on the inverse scale is more aggressive than BFGS, in the same spirit as employing negative values of the Broyden parameter. In fact, the next section will demonstrate that minimizing $\|E_B\|_F$ produces exactly a negative Broyden update.

Interestingly, minimizing E_B (with respect to commonly used scalar measures of matrices) is equivalent to minimizing

$$E_B^* \equiv B_{k+1}^{1/2}(B_k^{-1} - B_{k+1}^{-1})B_{k+1}^{1/2} \tag{14}$$

because E_B^* and E_B have the same eigenvalues, as shown in Appendix C. The matrix E_B^* scales the difference in inverse estimates by the still-to-be-determined update. Greenstadt [21] minimized a weighted change of the inverse estimates. In deriving BFGS, Goldfarb [20] writes, "If, instead, $[B_{k+1}]$ is substituted for [the weight matrix] in [Greenstadt's result], then [BFGS] is obtained." Although this sounds like minimizing E_B^*, Goldfarb in fact minimized E_W^* with a fixed weight matrix that satisfied the quasi-Newton condition required of B_{k+1}, namely $W\delta_k = \gamma_k$. SQN, on the other hand, can be viewed as directly using the unknown B_{k+1} as the weight matrix in Greenstadt's objective function.

3. SQN: Least relative change. The form of E_B in (13) 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 making no change to the portion of B_k that is orthogonal to δ_k but applies the idea in a transformed coordinate system.

As the focus is on the kth 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 kth 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

$$\tilde{x} = U'B^{1/2}x, \tag{15}$$

where U is an orthonormal rotation matrix with the 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:

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

The objective function and gradient become

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

and the transformed Hessian is

$$\tilde{G}(\tilde{x}) \equiv \nabla^2\tilde{f}(\tilde{x}) = U'B^{-1/2}G(x)B^{-1/2}U. \tag{16}$$

Substituting the estimated Hessian B for $G(x)$ in (16) 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

$$\begin{bmatrix} a \\ b \end{bmatrix} \equiv \frac{\tilde{g}(\tilde{x}_+) - \tilde{g}(\tilde{x})}{(1, 0, \ldots, 0)(\tilde{x}_+ - \tilde{x})} = \frac{U'B^{-1/2}\gamma}{(\delta'B\delta)^{1/2}}, \tag{17}$$

where the first element a is a scalar and b is an $(n-1)$-dimensional vector. The curvature condition (3) implies that $a \geq (1-\rho_2)/s > 0$. The quasi-Newton condition (1) is equivalent to the intuitive idea that the numerical derivatives (17) form the first

column of the updated Hessian matrix. Since the Hessian is symmetric, the general form of update in transformed coordinates becomes

$$\tilde{B}_+ = \begin{bmatrix} a & b' \\ b & C \end{bmatrix}, \tag{18}$$

where symmetric 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 does one 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* $\|E_B\|_F$ *(and hence also* $\|E_B^*\|_F$*) subject to* (1) *and* $B_+ \geq 0$ *has canonical form*

$$\tilde{B}_+ = \begin{bmatrix} a & b' \\ b & I_{n-1} + \lambda_{\text{SQN}}\, bb'/a \end{bmatrix}, \tag{19}$$

where, for $\tilde{r} \equiv b'b/a$,

$$\lambda_{\text{SQN}} = \begin{cases} 0 & \text{if } \tilde{r} \leq 1, \\ 1 - \tilde{r}^{-1} & \text{otherwise.} \end{cases} \tag{20}$$

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

Proof. Appendix C implies that $\|E_B\|_F = \|E_B^*\|_F$:

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

where $\Phi \equiv C - bb'/a$ and we have used $\tilde{B}_+ = U'B^{-1/2}B_+B^{-1/2}U$ from (16). $B_+ \geq 0$ is equivalent to $\Phi \geq 0$ so that minimizing $\|E_B\|_F$ over $B_+ \geq 0$ is equivalent to minimizing the first three terms of the final expression over $\Phi \geq 0$.

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

$$\operatorname{tr}\left(\Phi^2\right) - 2\operatorname{tr}\left(\Phi\right) + 2b'\Phi b/a \;\geq\; \sum_i \eta_i^2 - 2\sum_i \eta_i + 2\eta_1 b'b/a \tag{21}$$

with equality if and only if the first eigenvector of Φ is proportional to b. The right-hand side of (21) is minimized by $\eta_2 = \cdots = \eta_n = 1$ and

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

Thus, the left-hand side of (21) is minimized by a matrix with the specified eigenvalues and first eigenvector equal to $b/\sqrt{b'b}$. The required matrix is

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

and this corresponds to the optimal C given in the theorem. □

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 safeguarded choice of λ_{SQN} to prevent the Hessian estimate from becoming singular. See (8).

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.

PROPOSITION 1 (SQN accuracy preservation). *If the true Hessian in canonical coordinates is positive definite and given by*

$$\tilde{B}_+ = \begin{bmatrix} a & b' \\ b & C_{\text{TRUE}} \end{bmatrix}, \tag{22}$$

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,

$$\left|u'\left(C_{\text{SQN}} - C_{\text{TRUE}}\right)u\right| \;\le\; \left|u'\left(I_{n-1} - C_{\text{TRUE}}\right)u\right| \tag{23}$$

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 nonzero 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* (15) *the Broyden update* (4) *transforms to*

$$\tilde{B}_+ = \begin{bmatrix} a & b' \\ b & I_{n-1} + \lambda bb'/a \end{bmatrix}, \tag{24}$$

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:

Method	λ	ϕ
SQN	$\max\{0, 1-\tilde{r}^{-1}\}$	$\max\{-a, -a\tilde{r}^{-1}\}$,
BFGS	1	0
DFP	$1 + a^{-1}$	1

where if $\tilde{r} = 0$, the max *is taken to be the first argument.*

See Appendix B for a proof. The proof also provides formulae for a and b in terms of the usual quantities δ, γ, and B and shows that $\tilde{r} = r$, where $\tilde{r} = b'b/a$ is defined in Theorem 1 and r is given in (7).

Although BFGS minimizes several different measures of change [16], Proposition 2 indicates that BFGS *increases* the lower right block ($\lambda > 0$) 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 positive semidefiniteness. The conclusion of Proposition 1 is that neither BFGS nor DFP preserves accuracy of the previous Hessian estimate (in the canonical sense of (23)) over a large class of directions. It is interesting that DFP explodes 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 $x - \hat{s}B^{-1}g$ with $\hat{s} = 1$, 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. The steepest descent method (SDQN) [31] 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 on SQN and other Broyden updates are reported in section 5.

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 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 (9) 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. It is reasonable to use a probability model for this purpose because the true Hessian varies unpredictably from one quasi-Newton iteration to the next and from one objective function to another. Therefore $\tilde{G}_+$ is never completely known. Furthermore, modeling $\tilde{G}_+$ as a random matrix provides a means of incorporating new curvature information obtained in a line search and appropriately updating the distribution of the unknown Hessian. The updated distribution is the key to determining what length of step should be taken in any given direction.

Several properties are desirable for the distribution of $\tilde{G}_+$. It should

(i) be centered at the previous estimate $\tilde{B} = I_n$,

(ii) have probabilities that taper off toward zero for matrices far from I_n, and

(iii) describe equal uncertainty in every direction because, although $\tilde{G}_+$ is likely less uncertain in the directions of recent steps, these directions are not available for use within the quasi-Newton framework.

The simplest statistical model for symmetric positive definite matrices that has the above properties is the Wishart distribution with expectation I_n:

$$\nu\tilde{G}_+ \sim \text{Wishart}_n\left(I_n, \nu\right), \tag{25}$$

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., [2] for the definition and properties of the Wishart family. The probability density function of $\tilde{G}_+$ is proportional to

$$\left|\tilde{G}_+\right|^{(\nu-n-1)/2} \exp\left\{-\frac{\nu}{2}\text{tr}\left(\tilde{G}_+\right)\right\}. \tag{26}$$

Because (26) involves only $\tilde{G}_+$ 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 (17). Therefore

$$\tilde{G}_+ = \begin{bmatrix} a & b' \\ b & C \end{bmatrix}, \tag{27}$$

where a and b are observed and C is not. Standard Wishart theory (see, e.g., [2]) provides the conditional distribution $[C|a,b]$ through

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

The conditional expectation and mode are

$$\text{E}(C|a,b) = \frac{\nu-1}{\nu} I_{n-1} + \frac{bb'}{a}, \tag{28}$$

$$\text{Mode}(C|a,b) = \frac{\nu}{\nu-n-1} I_{n-1} + \frac{bb'}{a}. \tag{29}$$

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 $\nu \to \infty$, the *large-sample* limit. Although ν could be estimated from a and b, using the large-sample limit is an attractive simplification that corresponds to modeling the current Hessian estimate as arbitrarily accurate *before* observing a and b.

Comparing (28) and (29) to (24) in Proposition 2 shows that 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 (24). Then

$$\begin{aligned} \lim_{\nu\to\infty} E\left(G_+|a,b\right) &= B^{1/2}U\left[\lim_{\nu\to\infty} E\left(\tilde{G}_+\middle| a,b\right)\right]U'B^{1/2} \\ &= B^{1/2}U\tilde{B}_+(1)U'B^{1/2} \\ &= B_+(1), \end{aligned} \tag{30}$$

which is the BFGS update.

Although (30) is the simplest statistical estimate of the Hessian, the SQN update is a better choice for *sequential* Hessian estimation because it preserves accuracy obtained in previous iterations (section 2.2 and Proposition 1). When estimating the optimal step size, however, accuracy preservation is not a concern—an appropriate step size in one iteration may or may not be appropriate in the next. Thus, while the SQN Hessian update is derived to preserve accuracy, the SQN step size, derived in the next section, uses conditional expectation to estimate the optimal step, given the most recent curvature information.

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

$$f(x_+ + sd_+) \approx f(x_+) + sd'_+g_+ + \frac{s^2}{2}d'_+G_+d_+ \tag{31}$$

with optimum step size

$$s^* = \frac{-d'_+g_+}{d'_+G_+d_+} \tag{32}$$

obtained by differentiating (31) with respect to s and setting the result to zero. The denominator of (32) involves the unknown Hessian, but an estimate of s^* can be obtained by replacing G_+ with its large-sample conditional expectation from (30):

$$\lim_{\nu\to\infty} E\left(G_+|a,b\right) = B_+(1) = B_+(\lambda) + (1-\lambda)(\delta'\gamma)\omega\omega', \tag{33}$$

where (4) has been used to express $B_+(1)$ in terms of a general Broyden update. The resulting optimum step size is obtained by plugging (33) into (32) and taking $d_+ = -B_+^{-1}(\lambda)g_+$, the next quasi-Newton step direction:

$$\hat{s}(\lambda) = \frac{g'_+B_+^{-1}(\lambda)g_+}{g'_+B_+^{-1}(\lambda)g_+ + (1-\lambda)(\delta'\gamma)(g'_+B_+^{-1}(\lambda)\omega)^2}. \tag{34}$$

This is the step size formula (9) of the SQN algorithm. For BFGS ($\lambda = 1$), the estimated optimum 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 (34) are shown in Figure 1 and demonstrate that SQN achieves consistent reduction in function evaluations, as well as iteration counts, compared to BFGS. Additional comparisons to SQN with unit initial steps for three new test functions are reported next.

5. Results on three new test functions. The CUTE 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.

It was found that BFGS is lopsided [8]: 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 [10]: 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

TABLE 2
Three test functions $f(x) = \sum_1^4 f_i(x_i)$ with simple Hessians.

	$f_i(x_i)$	$G_{ii}(x)$	Anticipated best λ_{NOM}
f^{dec} :	$\frac{1}{2}x_i^2 + \frac{1}{12}\eta_i^2 x_i^4$	$1 + (\eta_i x_i)^2$	negative
f^{inc} :	$\eta_i^{-2}\left[\eta_i x_i \arctan(\eta_i x_i) - \frac{1}{2}\ln(1+\eta_i^2 x_i^2)\right]$	$[1 + (\eta_i x_i)^2]^{-1}$	positive
f^{sin} :	$\frac{1}{2}x_i^2 + \eta_i^{-2}\left[\eta_i x_i - \sin(\eta_i x_i)\right]$	$1 + \sin(\eta_i x_i)$	near zero

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 with 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 check this understanding, we employ three new test functions f^{dec}, f^{inc}, and f^{sin} with simple Hessians that respectively decrease, increase, and change sinusoidally as x_k moves in the direction of the optimum value. Each function has $n = 4$ dimensions and has the form $f(x) = \sum_1^4 f_i(x_i)$. The functions are defined in Table 2, where the values $(\eta_1, \eta_2, \eta_3, \eta_4) \equiv (1, 2, 4, 8)$ scale how quickly curvature changes in each coordinate direction. Each function is convex and has a diagonal Hessian with an ith diagonal element as listed in the table. In each case the minimizer is $x^* = (0, 0, 0, 0)$, $f(x^*) = 0$, and $G(x^*) = I_4$.

For these functions we implement a range of Broyden updates with

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

where λ_{NOM} is set between -2 and 3, $\epsilon = 10^{-6}$, and initial step sizes are given by (34). Special cases are $\lambda_{\text{NOM}} = 0$ and 1, which correspond to SQN and BFGS, respectively.

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 compared to the previous step so that $\lambda_{\text{NOM}} = 0$ (i.e., SQN) should be nearly optimal. In each case, more extreme values of λ_{NOM} should produce more extreme effects.

5.1. Results for different λ_{NOM}. Figure 2 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. Fletcher's line search [15], as discussed in section 1.2, was used in this study. The true value is used for the starting Hessian estimate, $B_0 = G(x_0)$.

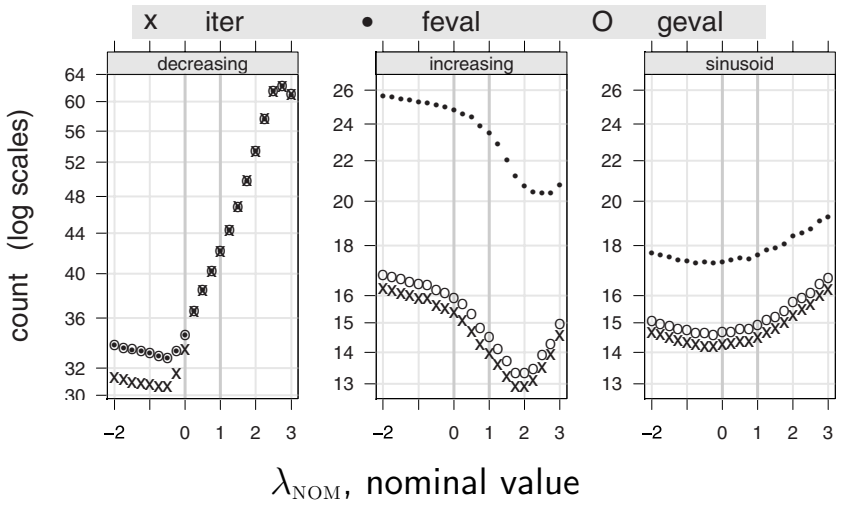

FIG. 2. *Performance counts versus* λ_{NOM} *on test functions with Hessians that are decreasing, increasing, and sinusoidal as* x_k *moves toward the minimum. Different symbols are used for iterations (*$\times$ *iter), function evaluations (*• *feval), and gradient evaluations (* O *geval). Initial steps are estimated using* (9)*. 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 ith component of x_0 was drawn randomly as

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

where the z_i are independent $\mathrm{N}(0,1)$ 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 produce similar results, though we have not studied this extensively. Convergence was declared when $f(x_k) < 10^{-10}$.

For f^{dec}, Figure 2 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, within each panel all three curves have nearly the same shape. But on f^{dec} function evaluations are always equal to gradient evaluations, whereas function evaluations are substantially higher on f^{inc} and f^{sin}. This indicates that the initial step size estimate is better for f^{dec} than for the other two functions because, with the Fletcher line search, if initial step sizes are too large to produce a sufficient decrease in the function value, then the function is reevaluated at additional trial steps with no gradient evaluations. 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

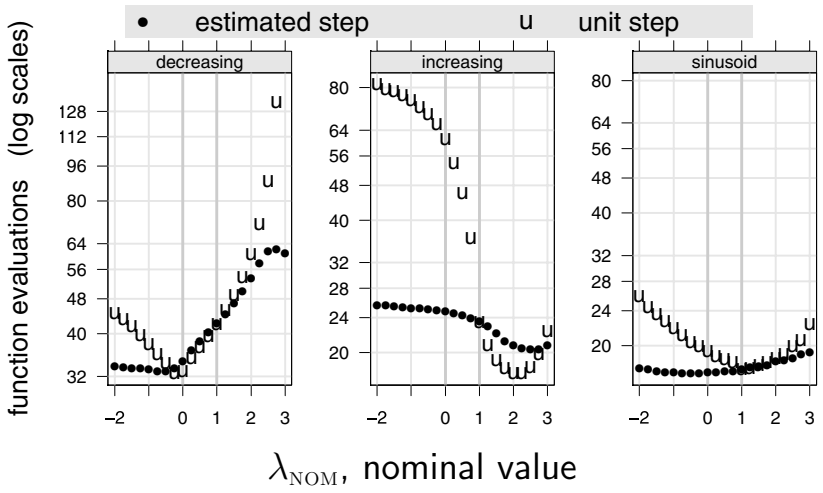

FIG. 3. *Function evaluation counts versus* λ_{NOM} *for three test functions. The plots compare performance with unit initial steps* (u) *against estimated initial steps* (•) *using* (9). *The dots in this figure are the same as in Figure* 2.

are biased in the direction of η was made to enhance the effect of λ_{NOM} below 1 on $f^{\mathrm{inc}}$and f^{sin}. The patterns in Figure 2 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. Thus, it would be meaningless to compare different choices of λ_{NOM} based on only a few test cases.

5.2. Results for different step sizes. Figure 3 demonstrates the importance of using estimated step sizes, especially with $\lambda_{\mathrm{NOM}} < 1$. The experiment is the same as in Figure 2, except that the algorithm was also run with unit initial step sizes. The plots compare average function evaluation counts for unit initial steps against those for estimated steps. In each panel, as λ_{NOM} decreases from 1 (BFGS), the unit initial step results eventually 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 $\lambda_{\mathrm{NOM}} = 1$ because the estimated step size is 1.

At $\lambda_{\mathrm{NOM}} = 0$ (SQN) the results of Figure 3 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 2, 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 $1 < \lambda_{\mathrm{NOM}} < 3$. 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.

6. 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 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.

Straightforward application of the Wishart model leads, in fact, to the BFGS update as is seen in (30). Another derivation of BFGS is obtained by taking the negative logarithm of the Wishart density (26), dividing by $\nu/2$, and taking $\nu \to \infty$. The result is the following function:

$$\psi(\tilde{B}_+) \equiv \operatorname{tr}\left(\tilde{B}_+\right) - \ln|\tilde{B}_+|.$$

Fletcher [16] demonstrated that BFGS minimizes $\psi(\tilde{B}_+) = \psi(E_B + I) = \psi(E_B^* + I)$, where E_B and E_B^* are defined in (13) and (14), respectively. Similarly DFP minimizes $\psi(\tilde{B}_+^{-1}) = \psi((E_B + I)^{-1}) = \psi((E_B^* + I)^{-1})$. Once again, the measure of change is influential.

We argue, however, that accuracy preservation (as measured by E_B^* and E_B) is more important than achieving the best one-step statistical estimate for the problem of *sequentially* estimating the Hessian matrix; this leads to the least-change formulation of Theorem 1. But the SQN update can also be derived from a statistical approach. We first obtained it by combining the Wishart model (25) with a prior distribution that strongly forced C toward the identity matrix. The prior was the statistical embodiment of the least-change principle. Details of this derivation are omitted to save space.

There is a fascinating historical connection that ties the relative change matrices E_B in (13) and E_B^* in (14) to BFGS, DFP, and the E_I method [21] from which Goldfarb [20] derived BFGS. E_W^* and E_W in (11) and (12) are well-known duals that measure change on the inverse and nominal scales and lead to the BFGS and DFP updates, respectively. In the same sense, the dual of E_B is

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

which is the matrix that Greenstadt minimized. Therefore the SQN update derived from E_B and E_B^* is the dual of the E_I update in the same sense that BFGS is the dual of the older DFP method. Although Greenstadt did not constrain B_{k+1} to be positive definite, minimizing $\|E_I\|_F$ over positive semidefinite updates results in truncating Greenstadt's solution at the critical value of the Broyden parameter. E_B^* was used in [1] to derive an optimally scaled BFGS update. Lukšan [24] generalized the technique to the Broyden family. Specializing Lukšan's result to the case of no scaling produces $\lambda = 0$ for $r < 1$.

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 alternate direction determined from the conditional distribution $[-G(x)^{-1}g|a, b]$. We have obtained promising results in some limited tests of these ideas.

Appendix A. Proof of Proposition 1.

Proof. If $r \leq 1$, then $C_{\mathrm{SQN}} = I_{n-1}$ and (23) holds as an equality for *all* u. Suppose $r > 1$ so that $C_{\mathrm{SQN}} = I_{n-1} + (1 - r^{-1})a^{-1}bb'$. Then for any $u : u'b = 0$,

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

and thus (23) holds as an equality. Suppose $u = \rho b$ for some $\rho \neq 0$. Positive definiteness of the true Hessian implies $(C_{\mathrm{TRUE}} - a^{-1}bb') > 0$, and thus

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

That is, in the direction of u, C_{SQN} is closer to C_{TRUE} than I_{n-1} is, and this implies that (23) holds as a strict inequality.

To prove the final statement, suppose that $C_{\mathrm{TRUE}} = I_{n-1}$ so that the right-hand side of (23) equals zero and consider two cases as follows. First, suppose that $\|V'b\| > 0$ and take $u = \rho b$ with $\rho \neq 0$. Then

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

and (23) 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

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

which violates (23). □

Appendix B. Proof of Proposition 2.

Proof. Using (16), 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:

$$\begin{aligned} B_+ &= B^{1/2}U \begin{bmatrix} a & b' \\ b & I_{n-1} + \lambda bb'/a \end{bmatrix} U'B^{1/2} \\ &= B + B^{1/2}U \begin{bmatrix} a-1 & b' \\ b & \lambda bb'/a \end{bmatrix} U'B^{1/2} \\ &= B + B^{1/2}U(D_1 + D_2 + D_3)U'B^{1/2}, \end{aligned} \tag{35}$$

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

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

From these equalities and (35), we see that the expression for B_+ is identical to (4) with $c = (\lambda-1)(\delta'\gamma)$. ∎

Appendix C. Equivalence of change matrices E_B and E_B^*. A scalar measure is required to define the "size" of the matrix E_B defined in (13). Many of the most common scalar measures depend only on eigenvalues—for example, trace, determinant, spectral norm, Frobenius norm, and the ψ-function [9], $\psi(E_B+I) = \operatorname{tr}(E_B+I) - \ln|E_B+I|$.

LEMMA 1. *The eigenvalues of E_B in* (13) *are identical to those of E_B^* in* (14). *Also, the eigenvalues of E_B+I are identical to those of E_B^*+I.*

Proof. Let $\simeq$ denote equality of eigenvalues and note that $PQ \simeq QP$ for square P and Q. Thus,

$$E_B^* \simeq B_{k+1}\left(B_k^{-1}-B_{k+1}^{-1}\right) = \left(B_{k+1}-B_k\right)B_k^{-1} \simeq E_B$$

and

$$E_B^*+I = B_{k+1}^{1/2}B_k^{-1}B_{k+1}^{1/2} \simeq B_k^{-1/2}B_{k+1}B_k^{-1/2} = E_B+I. \quad ∎$$

Acknowledgments. We are grateful to colleagues J. Chambers, D. Gay, D. Lambert, C. Mallows, and M. Wright for helpful discussions. We are also grateful to the reviewers whose careful comments resulted in a clearer presentation.

REFERENCES

[1] M. Al-Baali and R. Fletcher, *Variational methods for nonlinear least-squares*, J. Oper. Res. Soc., 36 (1985), pp. 405–421.
[2] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, 2nd ed., John Wiley and Sons, New York, 1984.
[3] I. Bongartz, A. R. Conn, N. I. M. Gould, and P. L. Toint, *CUTE: Constrained and unconstrained testing environments*, ACM Trans. Math. Software, 21 (1995), pp. 123–160.
[4] C. G. Broyden, *A class of methods for solving nonlinear simultaneous equations*, Math. Comp., 19 (1965), pp. 577–593.
[5] C. G. Broyden, *Quasi-Newton methods and their applications to function minimisation*, Math. Comp., 21 (1967), pp. 368–381.
[6] C. G. Broyden, *The convergence of a class of double rank minimization algorithms: 2. The new algorithm*, J. Inst. Math. Appl., 6 (1970), pp. 222–231.
[7] C. G. Broyden, *On the discovery of the "good Broyden" method*, Math. Program., 87 (2000), pp. 209–213.
[8] R. H. Byrd, D. C. Liu, and J. Nocedal, *On the behavior of Broyden's class of quasi-Newton methods*, SIAM J. Optim., 2 (1992), pp. 533–557.
[9] R. H. Byrd and J. Nocedal, *A tool for the analysis of quasi-Newton methods with application to unconstrained minimization*, SIAM J. Numer. Anal., 26 (1989), pp. 727–739.
[10] R. H. Byrd, J. Nocedal, and Y.-X. Yuan, *Global convergence of a class of quasi-Newton methods on convex problems*, SIAM J. Numer. Anal., 24 (1987), pp. 1171–1190.
[11] W. S. Cleveland, *Robust locally weighted regression and smoothing scatter plots*, J. Amer. Statist. Assoc., 74 (1979), pp. 829–836.

[12] J. B. Crockett and H. Chernoff, *Gradient method of maximization*, Pacific J. Math., 5 (1955), pp. 33–50.
[13] W. C. Davidon, *Variable metric method for minimization*, SIAM J. Optim., 1 (1991), pp. 1–17.
[14] R. Fletcher, *A new approach to variable metric algorithms*, Comput. J., 13 (1970), pp. 317–322.
[15] R. Fletcher, *Practical Methods of Optimization*, 2nd ed., John Wiley and Sons, New York, 1987.
[16] R. Fletcher, *A new variational result for quasi-Newton formulae*, SIAM J. Optim., 1 (1991), pp. 18–21.
[17] R. Fletcher, *An overview of unconstrained optimization*, in Algorithms for Continuous Optimization: The State of the Art, E. Spedicato, ed., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1994, pp. 109–143.
[18] R. Fletcher and M. J. D. Powell, *A rapidly convergent descent method for minimization*, Comput. J., 6 (1963), pp. 163–168.
[19] P. E. Gill and W. Murray, *Performance evaluation for nonlinear optimization*, in Performance Evaluation for Numerical Software, L. D. Fosdick, ed., North–Holland, Amsterdam, 1979, pp. 221–234.
[20] D. Goldfarb, *A family of variable metric methods derived by variational means*, Math. Comp., 24 (1970), pp. 23–26.
[21] J. Greenstadt, *Variations on variable metric methods*, Math. Comp., 24 (1970), pp. 1–22.
[22] J. Greenstadt, *Reminiscences on the development of the variational approach to Davidon's variable-metric method*, Math. Program., 87 (2000), pp. 265–280.
[23] L. Lukšan, *Computational experience with known variable metric updates*, J. Optim. Theory Appl., 83 (1994), pp. 27–47.
[24] L. Lukšan, *Variationally derived scaling and variable metric updates from the preconvex part of the Broyden family*, J. Optim. Theory Appl., 73 (1992), pp. 299–307.
[25] L. Lukšan and E. Spedicato, *Variable metric methods for unconstrained optimization and nonlinear least squares*, J. Comput. Appl. Math., 124 (2000), pp. 61–95.
[26] R. B. Mifflin and J. L. Nazareth, *The least prior deviation quasi-Newton update*, Math. Programming, 65 (1994), pp. 247–261.
[27] J. J. Moré, B. S. Garbow, and K. E. Hillstrom, *Testing unconstrained optimization software*, ACM Trans. Math. Software, 7 (1981), pp. 17–41.
[28] J. J. Moré and D. J. Thuente, *Line search algorithms with guaranteed sufficient decrease*, ACM Trans. Math. Software, 20 (1994), pp. 286–307.
[29] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, New York, 1999.
[30] D. F. Shanno, *Conditioning of quasi-Newton methods for function minimization*, Math. Comp., 24 (1970), pp. 647–650.
[31] Y. Zhang and R. P. Tewarson, *Quasi-Newton algorithms with updates from the preconvex part of Broyden family*, IMA J. Numer. Anal., 8 (1988), pp. 487–509.
[32] Q. Zhao, *Measures for Least Change Secant Methods*, Master's thesis, University of Waterloo, ON, Canada 1992.