BLOCK BFGS METHODS

WENBO GAO AND DONALD GOLDFARB[†]

Department of Industrial Engineering and Operations Research, Columbia University

Abstract. We introduce a quasi-Newton method with block updates called *Block BFGS*. We show that this method, performed with inexact Armijo-Wolfe line searches, converges globally and superlinearly under the same convexity assumptions as BFGS. We also show that Block BFGS is globally convergent to a stationary point when applied to non-convex functions with bounded Hessian, and discuss other modifications for non-convex minimization. Numerical experiments comparing Block BFGS, BFGS and gradient descent are presented.

1. Introduction

The classical BFGS method is perhaps the best known *quasi-Newton method* for minimizing an unconstrained function $f(x)$. These methods iteratively proceed along search directions $d_k = -B_k^{-1}\nabla f(x_k)$, where B_k is an approximation to the Hessian $\nabla^2 f(x_k)$ at the current iterate x_k. Quasi-Newton methods differ primarily in the manner in which they update the approximation B_k. The BFGS method constructs an update B_{k+1} which is the nearest matrix to B_k (in a variable metric) satisfying the *secant equation* $B_{k+1}(x_{k+1} - x_k) = \nabla f(x_{k+1}) - \nabla f(x_k)$. This can be interpreted as modifying B_k to act like $\nabla^2 f(x)$ along the direction $x_{k+1} - x_k$, so that successive updates induce B_k to resemble $\nabla^2 f(x)$ along the search directions.

A natural extension of the classical BFGS method is to incorporate information about $\nabla^2 f(x)$ along *multiple* directions in each update. Early work in this area includes the development by Schnabel [17] of quasi-Newton methods that satisfy multiple (say, q) secant equations $B_{k+1}s_k^{(i)} = \nabla f(x_{k+1}) - \nabla f(x_{k+1} - s_k^{(i)})$ for directions $s_k^{(1)}, \ldots, s_k^{(q)}$. This approach has the disadvantage that the resulting update is generally not symmetric, and considerable modifications are required to ensure B_k remains positive definite. Consequently, there appears to have been little interest in quasi-Newton methods with block updates in the years following Schnabel's initial report.

More recently, a stochastic quasi-Newton method with block updates was introduced by Gower, Goldfarb, and Richtárik [6]. Their approach constructs an update which satisfies *sketching equations* of the form

$$B_{k+1}s_k^{(i)} = \nabla^2 f(x_{k+1})s_k^{(i)}$$

for multiple directions $s_k^{(i)}$. By using sketching equations instead of secant equations, the update is guaranteed to remain symmetric, and in the case where $f(x)$ is convex, positive definite. The sketching equations can be thought of as 'tangent' equations that require B_{k+1} to incorporate information about the Hessian $\nabla^2 f(x_{k+1})$ at the most recent point x_{k+1}, as opposed to information about the average of $\nabla^2 f(x)$ between two points, i.e, along a secant.

E-mail address: `wg2279@columbia.edu, goldfarb@columbia.edu`.

Date: November 30, 2016.

2010 *Mathematics Subject Classification.* 90C53, 90C26.

[†]Research of this author was supported in part by NSF Grant CCF-1527809.

Experimental results from [6] show that their limited memory method *Stochastic Block L-BFGS* often outperforms other state-of-the-art methods when applied to a class of machine learning problems. This is promising, and provides evidence that quasi-Newton methods with block updates are a practical tool for unconstrained minimization.

In this paper, we introduce a deterministic quasi-Newton method *Block BFGS*. The key feature of Block BFGS is the inclusion of information about $\nabla^2 f(x)$ along multiple directions, by enforcing that B_{k+1} satisfies the sketching equations for a subset of previous search directions. We show that this method, performed with inexact Armijo-Wolfe line searches, has the same convergence properties as the classical BFGS method. Namely, if f is twice differentiable, convex, and bounded below, and the gradient of f is Lipschitz continuous, then Block BFGS converges. If, in addition, f is strongly convex and the Hessian of f is Lipschitz continuous, then Block BFGS achieves superlinear convergence.

Block BFGS can also be applied to non-convex functions. We show that if f has bounded Hessian, then Block BFGS converges to a stationary point of f. Modified forms of the classical BFGS method also have natural extensions to block updates, so modified block quasi-Newton methods are applicable in the non-convex setting.

The paper is organized as follows. Section 2 contains preliminaries and describes Armijo-Wolfe inexact line searches. In Section 3, we formally define the Block BFGS method and several variants. In Sections 4 and 5 respectively, we show that Block BFGS converges, and converges superlinearly, for f satisfying appropriate conditions. In Section 6, we show that Block BFGS converges for suitable non-convex functions, and describe several other modifications to adapt Block BFGS for non-convex optimization. In Section 7, we present the results of numerical experiments for several classes of convex and non-convex problems.

2. Preliminaries

The following notation will be used. The objective function of n variables is denoted by $f : \mathbb{R}^n \to \mathbb{R}$. We write $g(x)$ for the gradient $\nabla f(x)$ and $G(x)$ for the Hessian $\nabla^2 f(x)$. For a sequence $\{x_k\}$, $f_k = f(x_k)$ and $g_k = g(x_k)$. However, we deliberately use $G_k = G(x_{k+1})$ to simplify the update formula.

The norm $\|\cdot\|$ denotes the L_2 norm, or for matrices, the L_2 operator norm. The Frobenius norm will be explicitly indicated as $\|\cdot\|_F$. Angle brackets $\langle \cdot, \cdot \rangle$ denote the standard inner product $\langle x, y\rangle = y^T x$ and the trace inner product $\langle X, Y \rangle = \mathrm{Tr}(Y^T X)$. We use either notation $\langle x, y \rangle$ or $y^T x$ as is convenient. The symbol Σ^n denotes the space of $n \times n$ symmetric matrices, and $\preceq$ denotes the Löwner partial order; hence $A \succ 0$ means A is positive definite.

An $L\Sigma L^T$ decomposition is a factorization of a positive definite matrix into a product $L\Sigma L^T$, where L is lower triangular with ones on the diagonal, and $\Sigma = \mathrm{Diag}(\sigma_1^2, \ldots, \sigma_n^2)$. This is commonly called an LDL^T decomposition in the literature, but we write Σ in place of D as we use D to denote a matrix whose columns are previous search directions.

In the pseudocode for our algorithm, $\texttt{size}(A, 1)$ and $\texttt{size}(A, 2)$ denote the number of rows and columns of a matrix A respectively. The ij-entry of a matrix A will be denoted by A_{ij}. We use $\mathrm{Col}(A)$ to denote the linear space spanned by the columns of A. By convention, a sum over an empty index set is equal to 0.

Our inexact line search selects step sizes λ_k satisfying the *Armijo-Wolfe* conditions: for parameters α, β with $0 < \alpha < \frac{1}{2}$ and $\alpha < \beta < 1$, the step satisfies

$$f(x_k + \lambda_k d_k) \le f(x_k) + \alpha \lambda_k \langle g_k, d_k \rangle \tag{2.1}$$

and

$$\langle g(x_k + \lambda_k d_k), d_k \rangle \ge \beta \langle g_k, d_k \rangle \tag{2.2}$$

Furthermore, our line search always selects $\lambda_k = 1$ whenever this step size is admissible. This is important in the analysis of superlinear convergence in Section 5.

3. Block quasi-Newton Methods

In this section, we introduce *Block BFGS*, a quasi-Newton method with block updates, and several variants.

Algorithm 1 Block BFGS

input: $x_1^{(1)}, B_1, q$
for $k = 1, 2, 3 \dots$ **do**
  for $i = 1, \dots, q$ **do**
    $d_k^{(i)} \leftarrow -B_k^{-1} g_k^{(i)}$
    $\lambda_k^{(i)} \leftarrow \text{LINESEARCH}(x_k^{(i)}, d_k^{(i)})$
    $s_k^{(i)} \leftarrow \lambda_k^{(i)} d_k^{(i)}$
    $x_k^{(i+1)} \leftarrow x_k^{(i)} + s_k^{(i)}$
  end for
  $G_k \leftarrow G(x_k^{(q+1)})$
  $S_k \leftarrow [s_k^{(1)} \dots s_k^{(q)}]$
  $D_k \leftarrow \text{FILTERSTEPS}(S_k, G_k)$
  if D_k is not empty **then**
    $B_{k+1} \leftarrow B_k - B_k D_k (D_k^T B_k D_k)^{-1} D_k^T B_k + G_k D_k (D_k^T G_k D_k)^{-1} D_k^T G_k$
  else
    $B_{k+1} \leftarrow B_k$
  end if
  $x_{k+1}^{(1)} \leftarrow x_k^{(q+1)}$
end for

Algorithm 2 FILTERSTEPS

input: S_k, G_k **output**: D_k **parameters**: threshold $\tau > 0$
Initialize $D_k \leftarrow S_k, i \leftarrow 1$
while $i \leq \texttt{size}(D_k, 2)$ **do**
  $\sigma_i^2 \leftarrow [D_k^T G_k D_k]_{ii} - \sum_{j=1}^{i-1} L_{ij}^2 \Sigma_{jj}$
  $s_i \leftarrow$ column i of D_k
  if $\sigma_i^2 \geq \tau \|s_i\|^2$ **then**
    $\Sigma_{ii} \leftarrow \sigma_i^2$
    $L_{ii} \leftarrow 1$
    for $j = i+1, \dots, \texttt{size}(D_k, 2)$ **do**
      $L_{ji} \leftarrow \frac{1}{\Sigma_{ii}} ([D_k^T G_k D_k]_{ji} - \sum_{k=1}^{i-1} L_{ik} L_{jk} \Sigma_{kk})$
    end for
    $i \leftarrow i + 1$
  else
    Delete column i from D_k and row i from L
  end if
end while

3.1. **Block BFGS.** Block BFGS (Algorithm 1) takes q steps in each block, using a fixed Hessian approximation B_k. We may also take a varying number of steps, bounded above by q, but we assume every block contains q steps to simplify the presentation. We use a subscript k for the block index, and superscripts (i) for the steps within each block. The k-th block contains the iterates $x_k^{(1)}, \dots, x_k^{(q+1)}$, and $x_{k+1}^{(1)} = x_k^{(q+1)}$. At each point $x_k^{(i)}$, the step direction is $d_k^{(i)} = -B_k^{-1} g_k^{(i)}$, and line search is performed to obtain a step size $\lambda_k^{(i)}$. We take a step $s_k^{(i)} = \lambda_k^{(i)} d_k^{(i)}$. The angle between $s_k^{(i)}$ and $-g_k^{(i)}$ is denoted $\theta_k^{(i)}$. As B_k is positive definite, $\theta_k^{(i)} \in [0, \frac{\pi}{2})$.

After taking q steps, the matrix B_k is updated. Let $G_k = G(x_k^{(q+1)})$ denote the Hessian at the final iterate, and form the matrix $S_k = [s_k^{(1)} \dots s_k^{(q)}]$. We apply the FILTERSTEPS procedure (Algorithm 2) to S_k, which returns a subset D_k of the columns of S_k satisfying $\sigma_i^2 \geq \tau \|s_i\|^2$, where s_i is the i-th column of D_k and σ_i^2 is the i-th diagonal entry of the $L\Sigma L^T$ decomposition of $D_k^T G_k D_k$. $\tau > 0$ is a parameter which controls the strictness of the filtering; a small value of τ permits D_k to contain steps that are closer to being linearly dependent. In essence, FILTERSTEPS iteratively computes the $L\Sigma L^T$ decomposition of $S_k^T G_k S_k$ and discards columns of S_k corresponding to small diagonal entries, with the remaining columns forming D_k.

Define q_k to be the number of columns of D_k. If D_k is the empty matrix (all columns were removed), then no update is performed and $B_{k+1} = B_k$. If D_k is not empty, the matrix B_k is updated to have the same action as the Hessian G_k on the column space of D_k, or equivalently,

$$B_{k+1} D_k = G_k D_k \tag{3.1}$$

Let $D = D_k, G = G_k$. The formula for the update is given by

$$B_{k+1} = B_k - B_k D(D^T B_k D)^{-1} D^T B_k + GD(D^T GD)^{-1} D^T G \tag{3.2}$$

This formula is invariant under a change of basis of $\mathrm{Col}(D_k)$, so we can choose D_k to be any matrix with the same column space.

As is the case for standard quasi-Newton updates, there are many possible updates that satisfy equation (3.1). The specific Block BFGS update (3.2) is derived as follows. Let $H_k = B_k^{-1}$ be the approximation of the inverse Hessian. In contrast with the classical BFGS update, the update (3.2) is chosen so that H_{k+1} is the nearest matrix to H_k in a weighted norm, satisfying the system of sketching equations $H_{k+1} G_k D_k = D_k$ rather than a set of secant equations. That is, H_{k+1} is the solution to the minimization problem

$$\begin{aligned} \min_{\widetilde{H} \in \mathbb{R}^{n \times n}} \quad & \|\widetilde{H} - H_k\|_{G_k} \\ \text{s.t} \quad & \widetilde{H} = \widetilde{H}^T, \widetilde{H} G_k D_k = D_k \end{aligned} \tag{3.3}$$

where $\| \cdot \|_{G_k}$ is the norm $\|X\|_{G_k} = \mathrm{Tr}(X G_k X^T G_k)$, in analogy with the classical BFGS update. This norm is induced by an inner product, so H_{k+1} is an orthogonal projection onto the subspace $\{\widetilde{H} \in \Sigma^n : \widetilde{H} G_k D_k = D_k\}$. In Appendix A, it is shown that H_{k+1} has the explicit formula

$$H_{k+1} = D(D^T GD)^{-1} D^T + (I - D(D^T GD)^{-1} D^T G) H_k (I - GD(D^T GD)^{-1} D^T) \tag{3.4}$$

Taking the inverse yields formula (3.2). Moreover, as shown in [17], we have

Lemma 3.1. *If B_k (H_k) and $D_k^T G_k D_k$ are positive definite, then the Block BFGS update (3.2) for B_{k+1} ((3.4) for H_{k+1}) is positive definite.*

Proof. Our proof is adapted from Theorem 3.1 of [17]. Let $z \in \mathbb{R}^n$, and define $w = D_k^T z, v = z - G_k D_k (D_k^T G_k D_k)^{-1} w$. Using formula (3.4), we find that

$$z^T H_{k+1} z = w^T (D_k^T G_k D_k)^{-1} w + v^T H_k v$$

so $z^T H_{k+1} z \geq 0$. Furthermore, $z^T H_{k+1} z = 0$ only if both $w = 0$ and $v = 0$, in which case $z = 0$. Hence H_{k+1} is positive definite. □

In Section 4, we show that Block BFGS converges even if $B_k = B_{k+1} = \dots$ is stationary. In Section 5, we show that when f is strongly convex, the parameter τ can be chosen so an update is always performed, and the convergence is superlinear.

In practice, one may omit FILTERSTEPS. However, filtering may improve numerical stability, by removing nearly linearly dependent steps from D_k. Also, notice that $G_k D_k$ can be computed by performing q_k Hessian-vector products in parallel. It is often faster to compute Hessian-vector products than the full Hessian.

3.2. **Rolling Block BFGS.** Block BFGS uses the same matrix B_k throughout each block of q steps. We could also add information from these steps immediately, at the cost of doing far more updates. This variant, *Rolling Block BFGS*, performs a block update after every step, using a subset D_k of the previous q steps. D_k is formed by adding s_k as the first column of D_{k-1}, removing s_{k-q} if present, and filtering.

3.3. **Other Variants.** Block updates may be used within other quasi-Newton methods as well. For instance, the limited memory BFGS (L-BFGS) algorithm of Liu and Nocedal [12] is readily modified to use block updates. In [6], the authors tested a stochastic L-BFGS algorithm with block updates. Another possibility is to interleave standard BFGS updates with periodic block updates, to capture additional second-order information.

4. Convergence of Block BFGS

In this section we prove that Block BFGS with inexact Armijo-Wolfe line searches converges under the same conditions as does the classical BFGS method. These conditions are given in Assumption 1.

Assumption 1.

(1) f is convex, twice differentiable, and bounded below.
(2) For all x in the level set $\Omega = \{x \in \mathbb{R}^n : f(x) \leq f(x_1)\}$, the Hessian satisfies $G(x) \preceq MI$, or equivalently, $g(x)$ is Lipschitz continuous with Lipschitz constant M.

The main goal of this section is to prove the following theorem. The concept of our proof is similar to the analysis given by Powell [14] for the classical BFGS method.

Theorem 4.1. *Let f be a function satisfying Assumption 1, and let $\{x_k\}_{k=1}^{\infty}$ denote the sequence of all iterates produced by Block BFGS. Then $\liminf_k \|g_k\| = 0$.*

We begin by proving several lemmas. The first two are well known; see [2, 14].

Lemma 4.2. *$\sum_{k=1}^{\infty} \langle -g_k, s_k \rangle < \infty$, and therefore $\langle -g_k, s_k \rangle \to 0$.*

Proof. From the Armijo condition (2.1), $\langle -g_k, s_k \rangle = \lambda_k \langle -g_k, d_k \rangle \leq (1/\alpha)(f_k - f_{k+1})$. As f is bounded below,

$$\sum_{k=1}^{\infty} \langle -g_k, s_k \rangle \leq (1/\alpha) \sum_{k=1}^{\infty} (f_k - f_{k+1}) \leq (1/\alpha)(f_1 - \lim_{k\to\infty} f_k) < \infty$$

□

Lemma 4.3. *If the gradient $g(x)$ is Lipschitz continuous with constant M, then for $c_1 = \frac{1-\beta}{M}$, we have $\|s_k\| \geq c_1 \|g_k\| \cos\theta_k$.*

Proof. Let $y_k = g_{k+1} - g_k$. From the Wolfe condition (2.2),

$$\langle y_k, s_k \rangle = \langle g_{k+1}, s_k \rangle - \langle g_k, s_k \rangle \geq (1-\beta)\langle -g_k, s_k \rangle$$

By the Lipschitz continuity of the gradient, $\|y_k\| \leq M\|s_k\|$. Therefore

$$(1-\beta)\|g_k\|\|s_k\| \cos\theta_k = (1-\beta)\langle -g_k, s_k \rangle \leq \langle y_k, s_k \rangle \leq M\|s_k\|^2$$

yielding $\|s_k\| \geq c_1 \|g_k\| \cos\theta_k$. □

It is possible that D_k is empty for all $k \geq k_0$, and no further updates are made to B_{k_0}. This may occur, for example, if $G(x)$ has arbitrarily small eigenvalues and τ is large. We handle this case separately, as the theoretical properties of Block BFGS resemble gradient descent if this occurs.

Lemma 4.4. *Suppose that for some k_0, no further updates are made to B_k, so $B_k = B_{k_0}$ for all $k \geq k_0$. Then $\lim_k \|g_k\| = 0$.*

Proof. In the proof of Lemma 4.3, we obtained the inequality $\|s_k\|^2 \geq c_1\langle -g_k, s_k\rangle$. Multiplying by λ_k, we have $\lambda_k\|s_k\|^2 \geq c_1 s_k^T B_k s_k = c_1 s_k^T B_{k_0} s_k \geq c_1\lambda_{\min}(B_{k_0})\|s_k\|^2$, where $\lambda_{\min}(B_{k_0})$ is the smallest eigenvalue of B_{k_0}. Hence there exists a constant $\rho = c_1\lambda_{\min}(B_{k_0}) > 0$ with $\lambda_k \geq \rho$ for all $k \geq k_0$. We then have

$$\sum_{k=k_0}^{\infty} \frac{1}{\lambda_k}\langle -g_k, s_k\rangle = \sum_{k=k_0}^{\infty} g_k^T B_{k_0}^{-1} g_k \geq \frac{1}{\lambda_{\max}(B_{k_0})} \sum_{k=k_0}^{\infty} \|g_k\|^2$$

The left side is bounded above by $\sum_{k=k_0}^{\infty} \frac{1}{\rho}\langle -g_k, s_k\rangle < \infty$, so $\|g_k\| \to 0$. ☐

For the remainder of this section, we assume that there is an infinite sequence of updates. In fact, we may further assume that an update is made for every k, as one can verify that the propositions of this section continue to hold when we restrict our arguments to the subsequence of $\{B_k\}$ for which updates are made. This simplifies the notation. Note, however, that the same cannot simply be assumed in Section 5. The results in that section do *not* hold if updates are skipped. However, in Section 5 we are able to choose τ so as to guarantee that an update is made for every k.

Lemma 4.5. *Let $c_3 = \mathrm{Tr}(B_1) + qM$. Then for all k,*

$$\mathrm{Tr}(B_k) \leq c_3 k \quad \textit{and} \quad \sum_{j=1}^{k} \mathrm{Tr}(D_j^T B_j^2 D_j (D_j^T B_j D_j)^{-1}) \leq c_3 k$$

Proof. Clearly $\mathrm{Tr}(B_1) \leq c_3$. Define $E_j = G_j^{\frac{1}{2}} D_j$, and let $P_j = E_j(E_j^T E_j)^{-1}E_j^T$ be the orthogonal projection onto $\mathrm{Col}(E_j)$, so that $G_j D_j (D_j^T G_j D_j)^{-1} D_j^T G_j = G_j^{\frac{1}{2}} P_j G_j^{\frac{1}{2}}$. For $k \geq 1$, we expand $\mathrm{Tr}(B_{k+1})$ using Equation (3.2):

$$\begin{aligned} 0 < \mathrm{Tr}(B_{k+1}) &= \mathrm{Tr}(B_1) + \sum_{j=1}^{k} \mathrm{Tr}(G_j^{\frac{1}{2}} P_j G_j^{\frac{1}{2}}) - \sum_{j=1}^{k} \mathrm{Tr}(D_j^T B_j^2 D_j (D_j^T B_j D_j)^{-1}) \\ &\leq \mathrm{Tr}(B_1) + k(qM) - \sum_{j=1}^{k} \mathrm{Tr}(D_j^T B_j^2 D_j (D_j^T B_j D_j)^{-1}) \end{aligned}$$

where the first inequality follows from the positive definiteness of B_{k+1} (Lemma 3.1) and the second inequality follows since $\mathrm{rank}(P_j) \leq q$, and $\|G_j^{\frac{1}{2}} P_j G_j^{\frac{1}{2}}\| \leq \|G_j\|\|P_j\| \leq M$. This shows $\mathrm{Tr}(B_{k+1}) \leq c_3(k+1)$ and $\sum_{j=1}^{k} \mathrm{Tr}(D_j^T B_j^2 D_j (D_j^T B_j D_j)^{-1}) \leq c_3 k$. ☐

Lemma 4.6. *Let $s_k^{(i)}$ be a step included in D_k. Then*

$$\frac{\lambda_k^{(i)}\|g_k^{(i)}\|^2}{\langle -g_k^{(i)}, s_k^{(i)}\rangle} \leq \mathrm{Tr}(D_k^T B_k^2 D_k (D_k^T B_k D_k)^{-1})$$

Proof. By the Gram-Schmidt process applied to the columns of D_k, we can find a set of B_k-orthogonal vectors $\{v_1, \ldots, v_{q_k}\}$ spanning $\mathrm{Col}(D_k)$ with $v_1 = s_k^{(i)}$. Using the matrix $[v_1 \ldots v_{q_k}]$ for D_k, we have

$$D_k^T B_k D_k = \mathrm{Diag}(\langle s_k^{(i)}, -\lambda_k^{(i)} g_k^{(i)}\rangle, \langle v_2, B_k v_2\rangle, \ldots, \langle v_{q_k}, B_k v_{q_k}\rangle)$$

and therefore

$$\begin{aligned} \mathrm{Tr}(D_k^T B_k^2 D_k (D_k^T B_k D_k)^{-1}) &= \sum_{\ell=1}^{q_k} [D_k^T B_k^2 D_k]_{\ell\ell} [D_k^T B_k D_k]_{\ell\ell}^{-1} \\ &= \frac{(\lambda_k^{(i)}\|g_k^{(i)}\|)^2}{\lambda_k^{(i)}\langle -g_k^{(i)}, s_k^{(i)}\rangle} + \sum_{\ell=2}^{q_k} \frac{\|B_k v_\ell\|^2}{\langle v_\ell, B_k v_\ell\rangle} \geq \frac{\lambda_k^{(i)}\|g_k^{(i)}\|^2}{\langle -g_k^{(i)}, s_k^{(i)}\rangle} \end{aligned}$$

☐

We may assume without loss of generality that $D_k = [s_k^{(1)} \dots s_k^{(q_k)}]$.

Corollary 4.7.

$$\prod_{j=1}^{k}\prod_{i=1}^{q_j} \frac{\lambda_j^{(i)}\|g_j^{(i)}\|^2}{\langle -g_j^{(i)}, s_j^{(i)}\rangle} \leq (qc_3)^{qk}$$

Proof. Let $\widehat{q}_k = \sum_{j=1}^{k} q_j$, and note that $k \leq \widehat{q}_k \leq qk$. Hence, from Lemmas 4.5 and 4.6,

$$\frac{1}{\widehat{q}_k}\sum_{j=1}^{k}\sum_{i=1}^{q_j} \frac{\lambda_j^{(i)}\|g_j^{(i)}\|^2}{\langle -g_j^{(i)}, s_j^{(i)}\rangle} \leq \frac{qk}{\widehat{q}_k}c_3 \leq qc_3$$

Applying the arithmetic mean-geometric mean (AM-GM) inequality,

$$\left(\prod_{j=1}^{k}\prod_{i=1}^{q_j} \frac{\lambda_j^{(i)}\|g_j^{(i)}\|^2}{\langle -g_j^{(i)}, s_j^{(i)}\rangle}\right) \leq (qc_3)^{\widehat{q}_k} \leq (qc_3)^{qk}$$

□

Lemma 4.8. $\det(B_k) \leq \left(\frac{c_3 k}{n}\right)^n$ *for all* k.

Proof. By Lemma 4.5, $\mathrm{Tr}(B_k) \leq c_3 k$. Recall that the trace is equal to the sum of the eigenvalues, and the determinant to the product. Applying the AM-GM inequality to the eigenvalues of B_k, we obtain $\det(B_k) \leq \left(\frac{c_3 k}{n}\right)^n$. □

We will need the following two classical results from matrix theory; see [9].

Sylvester's Determinant Identity Let $A \in \mathbb{R}^{n\times m}, B \in \mathbb{R}^{m\times n}$. Then

$$\det(I_n + AB) = \det(I_m + BA)$$

Sherman-Morrison-Woodbury Formula Let $A \in \mathbb{R}^{n\times n}$ and $C \in \mathbb{R}^{k\times k}$ be invertible, and $U \in \mathbb{R}^{n\times k}, V \in \mathbb{R}^{k\times n}$. If $A + UCV$ and $C^{-1} + VA^{-1}U$ are invertible, then $(A+UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}$.

Lemma 4.9.

$$\det(B_{k+1}) = \frac{\det(D_k^T G_k D_k)}{\det(D_k^T B_k D_k)}\det(B_k)$$

Proof. Let $B = B_k, B^+ = B_{k+1}, D = D_k, G = G_k$. Then

$$\det(B^+) = \det(B)\det(I + B^{-\frac{1}{2}}GD(D^TGD)^{-1}D^TGB^{-\frac{1}{2}} - B^{\frac{1}{2}}D(D^TBD)^{-1}D^TB^{\frac{1}{2}})$$

Define $X = B^{-\frac{1}{2}}GD(D^TGD)^{-1}D^TGB^{-\frac{1}{2}}$ and $Y = D^TGD + D^TGB^{-1}GD$. Note that $I + X$ is invertible since $X \succeq 0$ and $I \succ 0$, and Y is invertible since $D^TGD \succ 0$. Thus, we can write

$$\det(B^+) = \det(B)\det(I+X)\det(I - (I+X)^{-1}B^{\frac{1}{2}}D(D^TBD)^{-1}D^TB^{\frac{1}{2}})$$

By Sylvester's determinant identity,

$$\det(I+X) = \det(I + (D^TGB^{-\frac{1}{2}})(B^{-\frac{1}{2}}GD(D^TGD)^{-1})) = \det(Y)\det(D^TGD)^{-1}$$

and

$$\det(I - (I+X)^{-1}B^{\frac{1}{2}}D(D^TBD)^{-1}D^TB^{\frac{1}{2}}) = \det(I - D^TB^{\frac{1}{2}}(I+X)^{-1}B^{\frac{1}{2}}D(D^TBD)^{-1})$$

Applying the Sherman-Morrison-Woodbury formula to $I+X$ with $U = B^{-\frac{1}{2}}GD, C = (D^TGD)^{-1}, V = D^TGB^{-\frac{1}{2}}$, we obtain $(I+X)^{-1} = I - B^{-\frac{1}{2}}GDY^{-1}D^TGB^{-\frac{1}{2}}$, and so

$$\det(I - (I+X)^{-1}B^{\frac{1}{2}}D(D^TBD)^{-1}D^TB^{\frac{1}{2}}) = \det(D^TGD)^2\det(Y)^{-1}\det(D^TBD)^{-1}$$

Thus $\det(B^+) = \det(B)\det(D^TGD)\det(D^TBD)^{-1}$ as desired. □

Lemma 4.10.

$$\det(B_{k+1}) \geq \left(\prod_{i=1}^{q_k} \frac{1}{\lambda_i}\right) (\tau c_1)^{q_k} \det(B_k)$$

Proof. Recall that the columns of D_k satisfy $\sigma_i^2 \geq \tau \|s_k^{(i)}\|^2$, where σ_i is the i-th diagonal element of the $L\Sigma L^T$ decomposition of $D_k^T G_k D_k$. We have $\det(D_k^T G_k D_k) = \prod_{i=1}^{q_k} \sigma_i^2$ and $\det(D_k^T B_k D_k) \leq \prod_{i=1}^{q_k} [D_k^T B_k D_k]_{ii} = \prod_{i=1}^{q_k} \langle s_k^{(i)}, -\lambda_k^{(i)} g_k^{(i)} \rangle$. By Lemma 4.9,

$$\begin{aligned}\det(B_{k+1}) &= \det(B_k) \frac{\det(D_k^T G_k D_k)}{\det(D_k^T B_k D_k)} \\ &\geq \det(B_k) \frac{\prod_{i=1}^{q_k} \tau \|s_k^{(i)}\|^2}{\prod_{i=1}^{q_k} \langle s_k^{(i)}, -\lambda_k^{(i)} g_k^{(i)} \rangle} \geq \det(B_k) \prod_{i=1}^{q_k} \frac{\tau}{\lambda_k^{(i)}} \frac{\|s_k^{(i)}\|}{\|g_k^{(i)}\| \cos \theta_k^{(i)}}\end{aligned}$$

By Lemma 4.3, $\frac{\|s_k^{(i)}\|}{\|g_k^{(i)}\| \cos \theta_k^{(i)}} \geq c_1$. Hence $\det(B_{k+1}) \geq \left(\prod_{i=1}^{q_k} \frac{1}{\lambda_k^{(i)}}\right) (\tau c_1)^{q_k} \det(B_k)$. □

Corollary 4.11.

$$\det(B_{k+1}) \geq (\tau c_1)^{qk} \det(B_1) \prod_{j=1}^{k} \prod_{i=1}^{q_j} \frac{1}{\lambda_j^{(i)}}$$

Corollary 4.12. *There exists a constant c_4 such that for all k,*

$$\prod_{j=1}^{k} \prod_{i=1}^{q_j} \frac{\|g_j^{(i)}\|^2}{\langle -g_j^{(i)}, s_j^{(i)} \rangle} \leq c_4^k$$

Proof. Multiplying the inequalities of Corollary 4.7 and Lemma 4.8, we obtain

$$\left(\prod_{j=1}^{k} \prod_{i=1}^{q_j} \frac{\lambda_j^{(i)} \|g_j^{(i)}\|^2}{\langle -g_j^{(i)}, s_j^{(i)} \rangle}\right) \left(\frac{\det(B_{k+1})}{\det(B_1)}\right) \leq (qc_3)^{qk} \left(\frac{(c_3(k+1)/n)^n}{\det(B_1)}\right) \leq \rho_1^k$$

for some constant ρ_1. Using the lower bound of Corollary 4.11, we also obtain

$$\begin{aligned}\left(\prod_{j=1}^{k} \prod_{i=1}^{q_j} \frac{\lambda_j^{(i)} \|g_j^{(i)}\|^2}{\langle -g_j^{(i)}, s_j^{(i)} \rangle}\right) \left(\frac{\det(B_{k+1})}{\det(B_1)}\right) &\geq \left(\prod_{j=1}^{k} \prod_{i=1}^{q_j} \frac{\lambda_j^{(i)} \|g_j^{(i)}\|^2}{\langle -g_j^{(i)}, s_j^{(i)} \rangle}\right) \cdot (\tau c_1)^{qk} \prod_{j=1}^{k} \prod_{i=1}^{q_j} \frac{1}{\lambda_j^{(i)}} \\ &= (\tau c_1)^{qk} \left(\prod_{j=1}^{k} \prod_{i=1}^{q_j} \frac{\|g_j^{(i)}\|^2}{\langle -g_j^{(i)}, s_j^{(i)} \rangle}\right)\end{aligned}$$

Take $c_4 = \frac{\rho_1}{(\tau c_1)^q}$, whence $\prod_{j=1}^{k} \prod_{i=1}^{q_j} \frac{\|g_j^{(i)}\|^2}{\langle -g_j^{(i)}, s_j^{(i)} \rangle} \leq c_4^k$. □

Finally, we can establish our main result.

Proof. (of Theorem 4.1) Assume to the contrary that $\|g_k^{(i)}\|$ is bounded away from zero. Lemma 4.2 implies that $\langle g_k^{(i)}, -s_k^{(i)} \rangle \to 0$. Thus, there exists k_0 such that for $k \geq k_0$, $\frac{\|g_k^{(i)}\|^2}{\langle g_k^{(i)}, -s_k^{(i)} \rangle} > c_4 + 1$. This contradicts Corollary 4.12, as $\prod_{j=1}^{k} \prod_{i=1}^{q_j} \frac{\|g_j^{(i)}\|^2}{\langle -g_j^{(i)}, s_j^{(i)} \rangle} \leq c_4^k$ for all k. We conclude that $\liminf_k \|g_k\| = 0$. □

A similar analysis shows that Rolling Block BFGS (Section 3.2) converges.

Theorem 4.13. *Assume f satisfies Assumption 1. Then the sequence $\{g_k\}_{k=1}^{\infty}$ produced by Rolling Block BFGS satisfies* $\liminf_k \|g_k\| = 0$.

Proof. By the calculations for Corollary 4.7, we have $\prod_{j=1}^{k} \frac{\lambda_j \|g_j\|^2}{\langle -g_j, s_j\rangle} \leq c_3^k$.

D_k is produced by adding column s_k to D_{k-1}, removing s_{k-q} if present, and then running Algorithm 2. Without loss of generality, assume that $D_k = [s_k \dots s_{k-q_k+1}]$. By definition, B_k satisfies $B_k D_{k-1} = G_{k-1} D_{k-1}$. Thus, we have

$$\det(D_k^T B_k D_k) \leq \prod_{i=0}^{q_k-1} \langle s_{k-i}, B_k s_{k-i}\rangle = \langle s_k, B_k s_k\rangle \prod_{i=1}^{q_k-1} \langle s_{k-i}, G_{k-1} s_{k-i}\rangle$$

which gives an analogue of Lemma 4.10:

$$\det(B_{k+1}) \geq \frac{\prod_{i=0}^{q_k-1} \tau \|s_{k-i}\|^2}{\langle s_k, -\lambda_k g_k\rangle \prod_{i=1}^{q_k-1} \langle s_{k-i}, G_{k-1} s_{k-i}\rangle} \det(B_k) \geq \frac{1}{\lambda_k} \frac{c_1 \tau^q}{M^{q-1}} \det(B_k)$$

Thus $\det(B_{k+1}) \geq \left(\frac{c_1\tau^q}{M^{q-1}}\right)^k \det(B_1) \prod_{j=1}^{k} \frac{1}{\lambda_k}$. The remainder of the proof follows exactly as in the proofs of Corollary 4.12 and Theorem 4.1. □

5. Superlinear Convergence of Block BFGS

In this section we show that Block BFGS converges superlinearly under the same conditions as does BFGS, namely, that f is strongly convex and its Hessian is Lipschitz continuous. We use the characterization of superlinear convergence given by Dennis and Moré [4], and employ an argument similar to the analysis used by Griewank and Toint [7] for partitioned quasi-Newton updates.

Assumption 2. The level set $\Omega = \{x \in \mathbb{R}^n : f(x) \leq f(x_1)\}$ is convex, and

(1) f is strongly convex on Ω, so there exist constants $m, M > 0$ such that for all $x \in \Omega$,

$$mI \preceq G(x) \preceq MI$$

Note that this implies f has a unique minimizer x_*, with value f_*.

(2) $G(x)$ is Lipschitz in a neighborhood of x_*, with Lipschitz constant μ.

For this section we assume $\tau \leq m$, where τ is the parameter in FILTERSTEPS. Since $\sigma_1^2 = [D_k^T G_k D_k]_{11} = \langle s_k^{(1)}, G_k s_k^{(1)}\rangle \geq m\|s_k^{(1)}\|^2$, the first column of D_k is never removed by FILTERSTEPS. This guarantees that an update is always performed.

Theorem 5.1. *Let f be a function satisfying Assumption 2. If the first step $s_k^{(1)}$ in each block is included in D_k, then Block BFGS converges superlinearly in the sense that*

$$\lim_{k\to\infty} \frac{\|x_k^{(i)} - x_*\|}{\|x_k^{(1)} - x_*\|} = 0 \qquad \textit{for } i = 2, \dots, q+1$$

We begin by showing that Block BFGS converges R-linearly. The first three lemmas are well known; see [2, 14].

Lemma 5.2. *For $c_1 = \frac{1-\beta}{M}$ and $c_2 = \frac{2(1-\alpha)}{m}$,*

$$c_1\|g_k\| \cos\theta_k \leq \|s_k\| \leq c_2\|g_k\| \cos\theta_k$$

Proof. By Taylor's theorem, there exists a point $\widetilde{x}$ on the line segment joining x_k, x_{k+1} such that $f(x_{k+1}) = f(x_k) + \langle g_k, s_k\rangle + \frac{1}{2} s_k^T G(\widetilde{x}) s_k$. From (2.1), $f(x_{k+1}) - f(x_k) \leq \alpha\langle g_k, s_k\rangle$, so $(1-\alpha)\langle -g_k, s_k\rangle \geq \frac{1}{2} s_k^T G(\widetilde{x}) s_k \geq \frac{1}{2} m\|s_k\|^2$. Rearranging yields $\|s_k\| \leq c_2\|g_k\| \cos\theta_k$. The lower bound was shown in Lemma 4.3. □

Lemma 5.3. *For any $x \in \Omega$, $\|g(x)\|^2 \geq 2m(f(x) - f_*)$.*

Proof. The result is immediate if $x = x_*$, so assume $x \neq x_*$. By Taylor's theorem, there exists a point $\widetilde{x}$ on the line segment joining x, x_* such that $f(x_*) = f(x) + g(x)^T(x_* - x) + \frac{1}{2}(x_* - x)^T G(\widetilde{x})(x_* - x)$, in which case

$$g(x)^T(x - x_*) = f(x) - f_* + \frac{1}{2}(x_* - x)^T G(\widetilde{x})(x_* - x) \geq f(x) - f_* + \frac{1}{2} m\|x - x_*\|^2$$

Using the Cauchy-Schwarz inequality, we find that $\|g(x)\|\|x - x_*\| \geq f(x) - f_* + \frac{1}{2}m\|x - x_*\|^2$. Applying the AM-GM inequality and squaring yields $\|g(x)\|^2 \geq 2m(f(x) - f_*)$. □

Lemma 5.4.

$$f_{k+1} - f_* \leq (1 - 2\alpha m c_1 \cos^2\theta_k)(f_k - f_*)$$

Proof. The Armijo condition (2.1) and Lemma 5.2 imply that

$$f_{k+1} - f_k \leq \alpha\langle g_k, s_k\rangle = -\alpha\|g_k\|\|s_k\|\cos\theta_k \leq -\alpha c_1\|g_k\|^2\cos^2\theta_k$$

By Lemma 5.3, $\|g_k\|^2 \geq 2m(f_k - f_*)$. Hence $f_{k+1} - f_* \leq \left(1 - 2\alpha m c_1 \cos^2\theta_k\right)(f_k - f_*)$. □

Define $r_k = \|x_k^{(q+1)} - x_*\|$. R-linear convergence implies that the errors r_k diminish to zero rapidly enough that $\sum_{k=1}^{\infty} r_k < \infty$, a key property.

Theorem 5.5. *There exists $\delta < 1$ such that $f(x_k^{(q+1)}) - f_* \leq \delta^k(f(x_1^{(1)}) - f_*)$, and thus $\sum_{k=1}^{\infty} r_k < \infty$.*

Proof. From Lemma 4.12, $\prod_{j=1}^{k}\prod_{i=1}^{q_j} \frac{\|g_j^{(i)}\|}{\|s_j^{(i)}\|\cos\theta_j^{(i)}} \leq c_4^k$. Lemma 5.2 gives the upper bound $\|s_j^{(i)}\| \leq c_2\|g_j^{(i)}\|\cos\theta_j^{(i)}$. Substituting, we find

$$\prod_{j=1}^{k}\prod_{i=1}^{q_j}\cos^2\theta_j^{(i)} \geq \left(\frac{1}{c_2^q c_4}\right)^k$$

From this, we see that at least $\frac{1}{2}k$ of the angles must satisfy $\cos^2\theta_j^{(i)} \geq \left(\frac{1}{c_2^q c_4}\right)^2$.

By Lemma 5.4, $f(x_k^{(i+1)}) - f_* \leq (1 - 2\alpha m c_1\cos^2\theta_k)(f(x_k^{(i)}) - f_*)$. Using our bound on the angles,

$$f(x_k^{(q+1)}) - f_* \leq \left(1 - 2\alpha m c_1\left(\frac{1}{c_2^q c_4}\right)^2\right)^{\frac{1}{2}k}(f(x_1^{(1)}) - f_*)$$

Hence, we may take $\delta = \left(1 - \frac{2\alpha m c_1}{c_2^{2q}c_4^2}\right)^{1/2}$. The strong convexity of f implies that $\frac{1}{2}m\|x - x_*\|^2 \leq f(x) - f_* \leq \frac{1}{2}M\|x - x_*\|^2$, so we have $r_k \leq (\sqrt{\delta})^k\sqrt{\frac{M}{m}}\|x_1^{(1)} - x_*\|$. Therefore $\sum_{k=1}^{\infty} r_k < \infty$. □

The classical BFGS method is invariant under a linear change of coordinates. It is easy to verify that Block BFGS also has this invariance, so we may assume without loss of generality that $G(x_*) = I$. This greatly simplifies the following calculations.

Lemma 5.6. *For any $v \in \mathbb{R}^n$, $\|(G_k - I)v\| \leq \mu r_k\|v\|$.*

Proof. Since $G(x_*) = I$,

$$\|(G_k - I)v\| \leq \|G(x_k^{(q+1)}) - G(x_*)\|\|v\| \leq \mu\|x_k^{(q+1)} - x_*\|\|v\| = \mu r_k\|v\|$$

□

The following notion is useful in our analysis. Define $\widetilde{B}_{k+1}$ to be the matrix obtained by performing a Block BFGS update on B_k with $G_k = G(x_*)$. Since we assumed $G(x_*) = I$, we have the explicit formula

$$\widetilde{B}_{k+1} = B_k - B_k D_k(D_k^T B_k D_k)^{-1}D_k^T B_k + D_k(D_k^T D_k)^{-1}D_k^T$$

and its inverse $\widetilde{H}_{k+1}$ is given by

$$\widetilde{H}_{k+1} = D_k(D_k^T D_k)^{-1}D_k^T + (I - D_k(D_k^T D_k)^{-1}D_k^T)H_k(I - D_k(D_k^T D_k)^{-1}D_k^T)$$

Lemma 5.7. *Let $B = B_k, \widetilde{B} = \widetilde{B}_{k+1}, D = D_k$. Define the following orthogonal projections:*

(1) $P = B^{\frac{1}{2}}D(D^T B D)^{-1}D^T B^{\frac{1}{2}}$, *the projection onto* $\mathrm{Col}(B^{\frac{1}{2}}D)$.

(2) $P_D = D(D^TD)^{-1}D^T$, *the projection onto* $\mathrm{Col}(D)$.
(3) $P_B = BD(D^TB^2D)^{-1}D^TB$, *the projection onto* $\mathrm{Col}(BD)$.

Then

$$\|B-I\|_F^2 - \|\widetilde{B}-I\|_F^2 = \|P_B - B^{\frac{1}{2}}PB^{\frac{1}{2}}\|_F^2 + 2\,\mathrm{Tr}(B(B^{\frac{1}{2}}PB^{\frac{1}{2}}) - (B^{\frac{1}{2}}PB^{\frac{1}{2}})^2)$$

Furthermore, $\mathrm{Tr}(B(B^{\frac{1}{2}}PB^{\frac{1}{2}}) - (B^{\frac{1}{2}}PB^{\frac{1}{2}})^2) \geq 0$, *and thus* $\|\widetilde{B}-I\|_F \leq \|B-I\|_F$.

Proof. Expand the Frobenius norm and use the identity $\mathrm{Tr}(BP_D) = \mathrm{Tr}(B^{\frac{1}{2}}PB^{\frac{1}{2}}P_D)$ to obtain

$$\begin{aligned}\|B-I\|_F^2 - \|\widetilde{B}-I\|_F^2 &= 2\,\mathrm{Tr}(B(B^{\frac{1}{2}}PB^{\frac{1}{2}})) - \mathrm{Tr}((B^{\frac{1}{2}}PB^{\frac{1}{2}})^2) - 2\,\mathrm{Tr}(B^{\frac{1}{2}}PB^{\frac{1}{2}}) \\ &\quad - \mathrm{Tr}(P_D^2) + 2\,\mathrm{Tr}(P_D) \\ &= 2\,\mathrm{Tr}(B(B^{\frac{1}{2}}PB^{\frac{1}{2}})) - 2\,\mathrm{Tr}((B^{\frac{1}{2}}PB^{\frac{1}{2}})^2) \\ &\quad + \mathrm{Tr}((B^{\frac{1}{2}}PB^{\frac{1}{2}})^2) - 2\,\mathrm{Tr}(B^{\frac{1}{2}}PB^{\frac{1}{2}}) + \mathrm{Tr}(I) \\ &\quad - \mathrm{Tr}(P_D^2) + 2\,\mathrm{Tr}(P_D) - \mathrm{Tr}(I)\end{aligned}$$

Factoring the above equation produces

$$\|B-I\|_F^2 - \|\widetilde{B}-I\|_F^2 = \|I - B^{\frac{1}{2}}PB^{\frac{1}{2}}\|_F^2 - \|I-P_D\|_F^2 + 2\,\mathrm{Tr}(B(B^{\frac{1}{2}}PB^{\frac{1}{2}}) - (B^{\frac{1}{2}}PB^{\frac{1}{2}})^2)$$

Let $P_B^{\perp}$ be the projection onto the orthogonal complement of $\mathrm{Col}(BD)$; hence $I = P_B + P_B^{\perp}$. Since $\langle P_B^{\perp}, B^{\frac{1}{2}}PB^{\frac{1}{2}}\rangle = \mathrm{Tr}(P_B^{\perp}BD(D^TBD)^{-1}D^TB) = 0$, we have $\|I - B^{\frac{1}{2}}PB^{\frac{1}{2}}\|_F^2 = \|P_B - B^{\frac{1}{2}}PB^{\frac{1}{2}}\|_F^2 + \|P_B^{\perp}\|_F^2$. The Frobenius norm of an orthogonal projection is equal to the square root of its rank, and thus

$$\|I - B^{\frac{1}{2}}PB^{\frac{1}{2}}\|_F^2 - \|I-P_D\|_F^2 = \|P_B - B^{\frac{1}{2}}PB^{\frac{1}{2}}\|_F^2 + \|P_B^{\perp}\|_F^2 - \|I-P_D\|_F^2 = \|P_B - B^{\frac{1}{2}}PB^{\frac{1}{2}}\|_F^2$$

This gives the desired equation. Now, observe that

$$\begin{aligned}\mathrm{Tr}(B(B^{\frac{1}{2}}PB^{\frac{1}{2}}) - (B^{\frac{1}{2}}PB^{\frac{1}{2}})^2) &= \mathrm{Tr}(BPB(I-P)) \\ &= \mathrm{Tr}((I-P)BPB(I-P)) \geq 0\end{aligned}$$

where in the second equality we have used that $I-P$ is the orthogonal projection onto $\mathrm{Col}(B^{\frac{1}{2}}D)^{\perp}$, and is therefore idempotent. This proves $\|\widetilde{B}-I\|_F \leq \|B-I\|_F$. □

We will later analyze the individual terms in Lemma 5.7. Let us define

$$\begin{aligned}\varphi_k &= \|P_{B_k} - B_k^{\frac{1}{2}}P_kB_k^{\frac{1}{2}}\|_F^2 \\ \psi_k &= \mathrm{Tr}(B_k(B_k^{\frac{1}{2}}P_kB_k^{\frac{1}{2}}) - (B_k^{\frac{1}{2}}P_kB_k^{\frac{1}{2}})^2)\end{aligned}$$

Intuitively, $\widetilde{B}_{k+1}$ and $\widetilde{H}_{k+1}$ should be closer approximations of I than B_k and H_k. This is made precise in the next lemma.

Lemma 5.8. $\|\widetilde{B}_{k+1}-I\|_F \leq \|B_k-I\|_F$ *and* $\|\widetilde{H}_{k+1}-I\|_F \leq \|H_k-I\|_F$.

Proof. That $\|\widetilde{B}_{k+1}-I\|_F \leq \|B_k-I\|_F$ was shown in Lemma 5.7. Clearly $\|\widetilde{H}_{k+1}-I\|_F \leq \|H_k-I\|_F$, as $\widetilde{H}_{k+1}$ is defined as the orthogonal projection of H_k onto the subspace of matrices $\{\widetilde{H} \in \Sigma^n : \widetilde{H}D_k = D_k\}$, which contains I (see (3.3)). □

Lemma 5.9. *There exists an index* k_0 *and constants* κ_1, κ_2 *such that* $\|B_{k+1} - \widetilde{B}_{k+1}\|_F \leq \kappa_1 r_k$ *and* $\|H_{k+1} - \widetilde{H}_{k+1}\|_F \leq (\|H_k - I\|_F + 1)\kappa_2 r_k$ *for all* $k \geq k_0$.

Proof. Let $\widetilde{B} = \widetilde{B}_{k+1}, \widetilde{H} = \widetilde{H}_{k+1}, H = H_k, D = D_k, G = G_k$, and define $\Delta = (G-I)D$. We may assume the columns of D are orthonormal, so $D^TD = I$. By Lemma 5.6, every column δ_i of Δ satisfies $\|\delta_i\| \leq \mu r_k$, which gives the useful bounds $\|\Delta\|, \|\Delta^T\| \leq \mu\sqrt{q}r_k$. This stems from the fact that a matrix A of rank q satisfies $\|A\| = \|A^T\| \leq \|A\|_F \leq \sqrt{q}\|A\|$, which we will use frequently.

To prove the first inequality, we write

$$\begin{aligned}\|B_{k+1} - \widetilde{B}\|_F &= \|GD(D^TGD)^{-1}D^TG - DD^T\|_F \\ &= \|GD(I + D^T\Delta)^{-1}D^TG - DD^T\|_F\end{aligned}$$

By the Sherman-Morrison-Woodbury formula, $(I + D^T\Delta)^{-1} = I - D^T(I + \Delta D^T)^{-1}\Delta$. Let $X = I + \Delta D^T$. Inserting this expression and using the triangle inequality, we have

$$\begin{aligned}\|GD(I + D^T\Delta)^{-1}D^TG - DD^T\|_F &= \|GDD^TG - DD^T - GDD^TX^{-1}\Delta D^TG\|_F \\ &\leq \|GDD^TG - DD^T\|_F + \|GDD^TX^{-1}\Delta D^TG\|_F\end{aligned}$$

By a routine calculation,

$$\|GDD^TG - DD^T\|_F = \|\Delta\Delta^T + \Delta D^T + D\Delta^T\|_F$$

whence $\|GDD^TG - DD^T\|_F \leq \rho_2 r_k$ for some constant ρ_2.

To bound the Frobenius norm of the other term, we bound its operator norm. Since $\Delta \to 0$ as $r_k \to 0$, there exists an index k_0 such that for $k \geq k_0$,

(1) $\|X - I\| \leq \frac{1}{2}$, so $\|X^{-1}\| \leq 2$, and
(2) $\|G - I\| \leq 1$, so $\|G\| \leq 2$

in which case $\|GDD^TX^{-1}\Delta D^TG\| \leq \rho_3 r_k$ for some ρ_3. Taking $\kappa_1 = \rho_2 + \sqrt{q}\rho_3$, we then have $\|B_{k+1} - \widetilde{B}\|_F \leq \kappa_1 r_k$ for all $k \geq k_0$.

A similar analysis applies to $\|H_{k+1} - \widetilde{H}\|_F$. Using the triangle inequality,

$$\begin{aligned}\|H_{k+1} - \widetilde{H}\|_F \leq\ & \|D(D^TGD)^{-1}D^T - DD^T\|_F \\ &+ \|(D(D^TGD)^{-1}D^TG - DD^T)H + H(GD(D^TGD)^{-1}D^T - DD^T)\|_F \\ &+ \|D(D^TGD)^{-1}D^TGHGD(D^TGD)^{-1}D^T - DD^THDD^T\|_F\end{aligned}$$

We bound each of the three terms. As before, $(D^TGD)^{-1} = I - D^TX^{-1}\Delta$, so we have $\|D(D^TGD)^{-1}D^T - DD^T\|_F = \|DD^TX^{-1}\Delta D^T\|_F$. For $k \geq k_0$, $\|X^{-1}\| \leq 2$, so $\|D(D^TGD)^{-1}D^T - DD^T\|_F \leq \rho_4 r_k$ for some ρ_4.

For the second term, observe that

$$GD(D^TGD)^{-1}D^T - DD^T = \Delta D^T - DD^TX^{-1}\Delta D^T - \Delta DX^{-1}\Delta D^T$$

Therefore the operator norm of the second term is bounded above by $\rho_5 r_k\|H\|$ for some ρ_5.

Finally, we bound the operator norm of the third term. Factoring out D and D^T on the left and right, we can write the inside term as

$$\begin{aligned}D^TGHGD - D^THD - (D^TX^{-1}\Delta D^TGHGD + D^TGHGDD^TX^{-1}\Delta) \\ + D^TX^{-1}\Delta D^TGHGDD^TX^{-1}\Delta\end{aligned}$$

Since $D^TGHGD - D^THD = \Delta^THD + D^TH\Delta + \Delta^TH\Delta$, the operator norm of the third term is bounded above by $\rho_6 r_k\|H\|$ for some ρ_6.

Adding the three terms, there is a constant κ_2 with $\|H_{k+1} - \widetilde{H}\|_F \leq (\|H_k - I\|_F + 1)\kappa_2 r_k$. □

Since superlinear convergence is an asymptotic property, we may assume $k_0 = 1$ in Lemma 5.9. We will also need the following technical result from [4].

Lemma 5.10 (3.3 of [4]). *Let $\{\phi_k\}$ and $\{\delta_k\}$ be sequences of non-negative numbers such that $\phi_{k+1} \leq (1 + \delta_k)\phi_k + \delta_k$ and $\sum_{k=1}^{\infty} \delta_k < \infty$. Then $\{\phi_k\}$ converges.*

Corollary 5.11. *$\{\|B_k - I\|_F\}_{k=1}^{\infty}$ and $\{\|H_k - I\|_F\}_{k=1}^{\infty}$ converge, and are therefore uniformly bounded.*

Proof. By Lemma 5.8 and Lemma 5.9, we have

$$\|H_{k+1} - I\|_F \leq \|H_{k+1} - \widetilde{H}_{k+1}\|_F + \|\widetilde{H}_{k+1} - I\|_F \leq (1 + \kappa_2 r_k)\|H_k - I\|_F + \kappa_2 r_k$$

Set $\phi_k = \|H_k - I\|_F$ and $\delta_k = \kappa_2 r_k$ in Lemma 5.10. Since $\sum_{k=1}^{\infty} r_k < \infty$, the sequence $\{\|H_k - I\|_F\}$ converges. The same reasoning applies to $\{\|B_k - I\|_F\}$. □

Corollary 5.12. *The condition numbers of* $\{B_k\}_{k=1}^{\infty}$ *are uniformly bounded.*

Lemma 5.13. *We have* $\lim_{k\to\infty} \varphi_k = 0$ *and* $\lim_{k\to\infty} \psi_k = 0$.

Proof. By Lemma 5.9 and Corollary 5.12, there exists a constant κ_3 such that

$$\|\widetilde{B}_{k+1} - I\|_F^2 \geq (\|B_{k+1} - I\|_F - \|B_{k+1} - \widetilde{B}_{k+1}\|_F)^2 \geq \|B_{k+1} - I\|_F^2 - \kappa_3 r_k$$

Hence

$$\begin{aligned}\sum_{k=1}^{\infty} \left(\|B_k - I\|_F^2 - \|\widetilde{B}_{k+1} - I\|_F^2\right) &\leq \sum_{k=1}^{\infty} \left(\|B_k - I\|_F^2 - \|B_{k+1} - I\|_F^2\right) + \kappa_3 r_{k+1} \\ &\leq \|B_1 - I\|_F^2 + \kappa_3 \sum_{k=1}^{\infty} r_{k+1} < \infty\end{aligned}$$

from which we deduce that $\|B_k - I\|_F^2 - \|\widetilde{B}_{k+1} - I\|_F^2 \to 0$. The desired limits then follow from Lemma 5.7, since $\|B_k - I\|_F^2 - \|\widetilde{B}_{k+1} - I\|_F^2 = \varphi_k + 2\psi_k$, and $\varphi_k, \psi_k \geq 0$. □

Lemma 5.14. *For any* $w_k \in \mathrm{Col}(D_k)$,

$$\left(1 - \frac{w_k^T B_k^2 w_k}{w_k^T B_k w_k}\right)^2 \leq \varphi_k \quad \text{and} \quad 0 \leq \frac{w_k^T B_k^3 w_k}{w_k^T B_k w_k} - \left(\frac{w_k^T B_k^2 w_k}{w_k^T B_k w_k}\right)^2 \leq \varphi_k + \psi_k$$

Consequently, for any sequence $\{w_k\}_{k=1}^{\infty}$ *with* $w_k \in \mathrm{Col}(D_k)$, *we have* $\lim_{k\to\infty} \frac{w_k^T B_k^2 w_k}{w_k^T B_k w_k} = 1$ *and* $\lim_{k\to\infty} \frac{w_k^T B_k^3 w_k}{w_k^T B_k w_k} = 1$.

Proof. For a fixed k, let $B = B_k, D = D_k$, and let $\Delta = (D^T B^2 D)^{-1} - (D^T B D)^{-1}$. Recall the definitions of P, P_B from Lemma 5.7. We can write

$$\begin{aligned}\varphi_k = \|P_B - B^{\frac{1}{2}} P B^{\frac{1}{2}}\|_F^2 = \mathrm{Tr}((BD\Delta D^T B)^2) &= \mathrm{Tr}(D^T B^2 D \Delta D^T B^2 D \Delta) \\ &= \mathrm{Tr}((I - D^T B^2 D (D^T B D)^{-1})^2)\end{aligned}$$

Take a B_k-orthogonal basis $\{v_1, \ldots, v_{q_k}\}$ for $\mathrm{Col}(D_k)$ with $v_1 = w_k$. The i-th diagonal entry of $(I - D^T B^2 D(D^T B D)^{-1})^2$ is then

$$\left(1 - \frac{v_i^T B^2 v_i}{v_i^T B v_i}\right)^2 + \sum_{j\neq i} \frac{(v_i^T B^2 v_j)^2}{v_i^T B v_i v_j^T B v_j}$$

Since every term is non-negative, we conclude that $\left(1 - \frac{w_k^T B^2 w_k}{w_k^T B w_k}\right)^2 \leq \varphi_k$, which proves the first statement. Also, notice that $\sum_{i=1}^{q_k} \sum_{j\neq i} \frac{(v_i^T B^2 v_j)^2}{v_i^T B v_i v_j^T B v_j} \leq \varphi_k$.

Next, write $\mathrm{Tr}(B(B^{\frac{1}{2}} P B^{\frac{1}{2}})) = \mathrm{Tr}(D^T B^3 D(D^T B D)^{-1})$ and $\mathrm{Tr}((B^{\frac{1}{2}} P B^{\frac{1}{2}})^2) = \mathrm{Tr}((D^T B^2 D (D^T B D)^{-1})^2)$. Again taking a B_k-orthogonal basis $\{v_1, \ldots, v_{q_k}\}$, we have

$$\begin{aligned}\mathrm{Tr}(D^T B^3 D(D^T B D)^{-1}) &= \sum_{i=1}^{q_k} \frac{v_i^T B^3 v_i}{v_i^T B v_i} \\ \mathrm{Tr}((D^T B^2 D (D^T B D)^{-1})^2) &= \sum_{i=1}^{q_k} \left(\frac{v_i^T B^2 v_i}{v_i^T B v_i}\right)^2 + \sum_{i=1}^{q_k} \sum_{j\neq i} \frac{(v_i^T B^2 v_j)^2}{v_i^T B v_i v_j^T B v_j}\end{aligned}$$

Thus

$$\begin{aligned}\operatorname{Tr}(B(B^{\frac{1}{2}}PB^{\frac{1}{2}})-(B^{\frac{1}{2}}PB^{\frac{1}{2}})^2) &= \sum_{i=1}^{q_k}\left(\frac{v_i^TB^3v_i}{v_i^TBv_i}-\left(\frac{v_i^TB^2v_i}{v_i^TBv_i}\right)^2\right)-\sum_{i=1}^{q_k}\sum_{j\neq i}\frac{(v_i^TB^2v_j)^2}{v_i^TBv_iv_j^TBv_j}\\ &\geq \sum_{i=1}^{q_k}\left(\frac{v_i^TB^3v_i}{v_i^TBv_i}-\left(\frac{v_i^TB^2v_i}{v_i^TBv_i}\right)^2\right)-\varphi_k\end{aligned}$$

By the Cauchy-Schwarz inequality applied to $v^TB^2v=\langle B^{\frac{1}{2}}v, B^{\frac{3}{2}}v\rangle$, we have $\frac{v^TB^3v}{v^TBv}\geq\left(\frac{v^TB^2v}{v^TBv}\right)^2$ for every $v\in\mathbb{R}^n$. Hence $0\leq\frac{w_k^TB^3w_k}{w_k^TBw_k}-\left(\frac{w_k^TB^2w_k}{w_k^TBw_k}\right)^2\leq\varphi_k+\psi_k$. The limits then follow from Lemma 5.13, since $\varphi_k,\psi_k\to 0$. ☐

Corollary 5.15. *Given any $w_k\in\operatorname{Col}(D_k)$,*

$$\frac{\|(B_k-I)w_k\|}{\|w_k\|}\leq\sqrt{2\varphi_k+\psi_k}$$

Consequently, for any sequence $\{w_k\}_{k=1}^{\infty}$ with $w_k\in\operatorname{Col}(D_k)$,

$$\lim_{k\to\infty}\frac{\|(B_k-I)w_k\|}{\|w_k\|}=0$$

Proof. By Lemma 5.14 and a routine calculation,

$$\begin{aligned}\frac{\|B_k^{\frac{1}{2}}(B_k-I)w_k\|}{\|B_k^{\frac{1}{2}}w_k\|} &= \sqrt{\frac{w_k^TB_k^3w_k}{w_k^TB_kw_k}-2\frac{w_k^TB_k^2w_k}{w_k^TB_kw_k}+1}\\ &= \sqrt{\frac{w_k^TB_k^3w_k}{w_k^TB_kw_k}-\left(\frac{w_k^TB_k^2w_k}{w_k^TB_kw_k}\right)^2+\left(1-\frac{w_k^TB_k^2w_k}{w_k^TB_kw_k}\right)^2}\\ &\leq\sqrt{2\varphi_k+\psi_k}\end{aligned}$$

Since the condition numbers of $\{B_k\}$ are uniformly bounded, the result follows. ☐

Lemma 5.16. *A step size of $\lambda_k=1$ is eventually admissible for steps d_k included in D_k.*

Proof. We check that $\lambda_k=1$ satisfies the Armijo-Wolfe conditions for all sufficiently large k. Let α and β be the Armijo-Wolfe parameters and choose a constant γ such that $0<\gamma<\frac{\frac{1}{2}-\alpha}{1-\alpha}$. By Corollary 5.15, for all sufficiently large k, the steps $d_k\in\operatorname{Col}(D_k)$ satisfy

$$\frac{\|(B_k-I)d_k\|}{\|d_k\|}\leq\gamma \tag{5.1}$$

in which case $\langle g_k,d_k\rangle=\langle g_k+d_k,d_k\rangle-\|d_k\|^2\leq-(1-\gamma)\|d_k\|^2$.

By Taylor's theorem, there exists a point $\widetilde{x}_k$ on the line segment joining x_k,x_k+d_k with $f(x_k+d_k)=f(x_k)+\langle g_k,d_k\rangle+\frac{1}{2}d_k^TG(\widetilde{x}_k)d_k$. Since $f(x_k)\leq f(x_{k-1}^{(q+1)})$, the strong convexity of f implies that $\|x_k-x_*\|\leq\sqrt{M/m}\,r_{k-1}$. Hence, taking $\rho_7=\mu\sqrt{M/m}$, we have $\|G(\widetilde{x}_k)-I\|\leq\mu\|\widetilde{x}_k-x_*\|\leq\rho_7(r_{k-1}+\|d_k\|)$. For the step size $\lambda_k=1$,

$$\begin{aligned}f(x_k+d_k)-f(x_k) &= \alpha\langle g_k,d_k\rangle+(1-\alpha)\langle g_k,d_k\rangle+\frac{1}{2}d_k^TG(\widetilde{x})d_k\\ &\leq\alpha\langle g_k,d_k\rangle-((1-\alpha)(1-\gamma)-1/2-(\rho_7/2)(r_{k-1}+\|d_k\|))\|d_k\|^2\end{aligned}$$

Since $(1-\alpha)(1-\gamma)-1/2>0$ and $r_{k-1}+\|d_k\|\to 0$, a step size of $\lambda_k=1$ satisfies the Armijo condition (2.1) for all sufficiently large k.

Next, apply Taylor's theorem to the function $t\mapsto\langle g(x_k+td_k),d_k\rangle$ to obtain a point $\widetilde{x}_k$ on the line segment joining x_k,x_k+d_k with $\langle g(x_k+d_k),d_k\rangle=\langle g_k,d_k\rangle+d_k^TG(\widetilde{x}_k)d_k$. Choosing $\gamma=\frac{\beta}{2-\beta}$

in (5.1), Corollary 5.15 implies that for sufficiently large k, $\langle -g_k, d_k\rangle = \langle g_k + d_k, -d_k\rangle + \|d_k\|^2 \leq (1 - \frac{1}{2}\beta)^{-1}\|d_k\|^2$. We can also take k large enough so that $1 - \rho_7(r_{k-1} + \|d_k\|) \geq 0$, and we then have

$$\begin{aligned}\langle g(x_k + d_k), d_k\rangle &\geq \langle g_k, d_k\rangle + (1 - \rho_7(r_{k-1} + \|d_k\|))\|d_k\|^2 \\ &\geq (\beta/2 + (1 - \beta/2)\rho_7(r_{k-1} + \|d_k\|))\langle g_k, d_k\rangle\end{aligned}$$

Thus, the Wolfe condition (2.2) is satisfied for all sufficiently large k. □

Lemma 5.16 applies only to steps d_k included in D_k. However, since Block BFGS does not prefer any particular step for inclusion in D_k, it is likely that eventually $\lambda_k = 1$ is admissible for *all* steps. This issue reveals a subtle artifact of the proof method, and we return to discuss it in the remark after the following proof of Theorem 5.1.

Proof. (of Theorem 5.1) Assume that the first step $s_k^{(1)}$ in each block is included in D_k. Let us write $x_k = x_k^{(1)}, d_k = d_k^{(1)}, g_k = g_k^{(1)}$. By Lemma 5.16, eventually $\lambda_k = 1$ is admissible for d_k, so $s_k = d_k$. From the triangle inequality, $\|d_k\| \leq \|x_k^{(1)} - x_*\| + \|x_k^{(2)} - x_*\|$, so

$$\frac{\|g_k^{(2)}\|}{\|d_k\|} \geq \frac{m\|x_k^{(2)} - x_*\|}{\|x_k^{(1)} - x_*\| + \|x_k^{(2)} - x_*\|} \tag{5.2}$$

Next, write

$$\begin{aligned}\frac{\|(B_k - I)d_k\|}{\|d_k\|} &= \frac{\|g(x_k + d_k) - g(x_k) - G(x_*)d_k - g(x_k + d_k)\|}{\|d_k\|} \\ &\geq \frac{\|g(x_k + d_k)\|}{\|d_k\|} - \frac{\|g(x_k + d_k) - g(x_k) - G(x_*)d_k\|}{\|d_k\|}\end{aligned}$$

By continuity of the Hessian $G(x)$, the second term converges to 0. Thus, Corollary 5.15 implies that $\frac{\|g_k^{(2)}\|}{\|d_k\|} = \frac{\|g(x_k+d_k)\|}{\|d_k\|} \to 0$. We deduce from (5.2) that

$$\frac{\|x_k^{(2)} - x_*\|}{\|x_k^{(1)} - x_*\|} \to 0$$

The strong convexity of f implies that $\frac{1}{2}m\|x - x_*\|^2 \leq f(x) - f(x_*) \leq \frac{1}{2}M\|x - x_*\|^2$. Since $f_k^{(i)} \leq f_k^{(2)}$ for $i \geq 2$, Block BFGS achieves the desired superlinear convergence:

$$\lim_{k\to\infty} \frac{\|x_k^{(i)} - x_*\|}{\|x_k^{(1)} - x_*\|} = 0 \qquad i = 2, \ldots, q+1$$

□

The same argument, with minimal alteration, applies to Rolling Block BFGS.

Remark. As we observed earlier, the choice to include $s_k^{(1)}$ in D_k is arbitrary. The proof of Theorem 5.1 holds with *any* selection rule for D_k as long as it guarantees $\sum_{k=1}^{\infty} r_k < \infty$. Therefore, it is likely that Theorem 5.1 and Lemma 5.16 apply to *all* steps. That is, eventually $\lambda_k = 1$ is admissible for all steps and $\frac{\|x_k^{(i+1)} - x_*\|}{\|x_k^{(i)} - x_*\|} \to 0$. In fact, by selecting D_k in a particular way, we can ensure that eventually $\lambda_k = 1$ is admissible for all steps.

Corollary 5.17. *Suppose that D_k is constructed to always contain a step for which $\lambda_k = 1$ is not admissible, whenever such a step exists in the k-th block. Then $\lambda_k = 1$ is eventually admissible for* all *steps.*

Proof. When executing the k-th update, we specifically set the first column of D_k to a step d_k from the k-th block for which $\lambda_k = 1$ is not admissible, if any such step exists. If we could find such a step d_k for infinitely many k, then this process would produce an infinite sequence of steps $d_k \in \mathrm{Col}(D_k)$ for which $\lambda_k = 1$ is never eventually admissible. This contradicts Lemma 5.16. □

However, Corollary 5.17 does *not* show that in general, $\lambda_k = 1$ is eventually admissible for all steps, as it only holds when we select steps in an adversarial manner. This example highlights an interesting dichotomy arising from our proof method. On one hand, Theorem 5.1 and Lemma 5.16 are retrospective and apply to any sequence $\{D_k\}$ that we select. This strongly suggests that they should hold for all steps. On the other hand, the method of proof (based on analyzing the convergence of $\|B_k - I\|_F^2 - \|\widetilde{B}_{k+1} - I\|_F^2$) makes use only of the steps in D_k, and thus can only prove things about the steps in D_k.

6. Modified Block BFGS for Non-Convex Optimization

Convergence theory for the classical BFGS method does not extend to non-convex functions. However, with minor modifications, BFGS performs well for non-convex optimization and can be shown to converge in some cases. Modifications that have been studied include:

(1) Cautious Updates (Li and Fukushima, [11])
A BFGS update is performed only if

$$\frac{y_k^T s_k}{\|s_k\|^2} \geq \epsilon \|g_k\|^\alpha$$

(2) Modified Updates (Li and Fukushima, [10])
The secant equation is modified to $B_{k+1}s_k = z_k$, where $z_k = y_k + r_k s_k$ and the parameter r_k is chosen so that $z_k^T s_k \geq \epsilon \|s_k\|^2$.

(3) Damped BFGS (Powell, [15])
The secant equation is modified to $B_{k+1}s_k = z_k$, where $z_k = \theta_k y_k + (1-\theta_k)B_k s_k$, and for $0 < \phi < 1$, the damping constant θ_k is determined by

$$\theta_k = \begin{cases} 1, & \text{if } y_k^T s_k \geq \phi s_k^T B_k s_k \\ \frac{(1-\phi)s_k^T B_k s_k}{s_k^T B_k s_k - y_k^T s_k}, & \text{otherwise} \end{cases}$$

This is perhaps the most widely used modified BFGS method. Unfortunately, no convergence proof is known for this method.

We show Block BFGS converges for non-convex functions, and describe analogous modifications for block updates. The next theorem provides a framework for proving convergence in the non-convex setting.

Theorem 6.1. *Assume f is twice differentiable and $-MI \preceq G(x) \preceq MI$ for all x in the convex hull of the level set $\{x \in \mathbb{R}^n : f(x) \leq f(x_1)\}$. Suppose that $\{\widetilde{G}_k\}_{k=1}^\infty$ is a sequence of symmetric matrices satisfying, for all k, the conditions*

(1) $-MI \preceq \widetilde{G}_k \preceq MI$
(2) *For some constant $\eta > 0$, the matrix D_k produced by* FILTERSTEPS$(S_k, \widetilde{G}_k)$ *satisfies* $D_k^T \widetilde{G}_k D_k \succeq \eta D_k^T D_k$

Then we may perform Block BFGS using the updates

$$B_{k+1} = B_k - B_k D_k (D_k^T B_k D_k)^{-1} D_k^T B_k + \widetilde{G}_k D_k (D_k^T \widetilde{G}_k D_k)^{-1} D_k^T \widetilde{G}_k$$

and Block BFGS converges in the sense that $\liminf_k \|g_k\| = 0$.

Proof. The proof follows that of Theorem 4.1, with several changes. First, note that Lemma 3.1 implies that B_{k+1} remains positive definite, since FILTERSTEPS ensures that $D_k^T \widetilde{G}_k D_k$ is positive definite. Observe that Lemma 4.3 continues to hold, as the condition $-MI \preceq G(x) \preceq MI$ for all

x in the convex hull of the level set implies that the gradient g is Lipschitz with constant M. In Lemma 4.5, take the constant c_3 to be $c_3 = \mathrm{Tr}(B_1) + \frac{qM^2}{\eta}$ and notice that

$$\mathrm{Tr}(\widetilde{G}_j D_j (D_j^T \widetilde{G}_j D_j)^{-1} D_j^T \widetilde{G}_j) \leq \frac{1}{\eta} \mathrm{Tr}(\widetilde{G}_j D_j (D_j^T D_j)^{-1} D_j^T \widetilde{G}_j) \leq \frac{qM^2}{\eta}$$

where the last inequality follows because $D_j(D_j^T D_j)^{-1} D_j^T$ is the orthogonal projection onto $\mathrm{Col}(D_j)$ and has rank $q_j \leq q$, and $\|\widetilde{G}_j D_j (D_j^T D_j)^{-1} D_j^T \widetilde{G}_j\| \leq \|\widetilde{G}_j\|^2 = M^2$.

The remainder of the proof is exactly as in Theorem 4.1. □

Using this result and the next lemma, we can show Block BFGS converges for non-convex functions.

Lemma 6.2. *Assume f is twice differentiable and $-MI \preceq G(x) \preceq MI$ for all x in the level set $\{x \in \mathbb{R}^n : f(x) \leq f(x_1)\}$. If $D_k^T G_k D_k$ satisfies $\sigma_i^2 \geq \tau \|s_i\|^2$, where σ_i is the i-th diagonal entry of the $L\Sigma L^T$ decomposition of $D_k^T G_k D_k$, then $D_k^T G_k D_k \succeq \eta D_k^T D_k$ for $\eta = \frac{\tau^q}{q^q M^{q-1}}$.*

Proof. Let $G = G_k, D = D_k$. Without loss of generality, we may assume the columns of D have norm 1, as otherwise we can normalize D by right-multiplying by a positive diagonal matrix. Then the diagonal entries σ_i^2 of the $L\Sigma L^T$ decomposition of $D^T G D$ satisfy $\sigma_i^2 \geq \tau$.

Order the eigenvalues of $D^T G D$ as $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_q > 0$. We have

$$\lambda_q = \frac{\det(D^T G D)}{\prod_{i=1}^{q-1} \lambda_i} \geq \frac{\tau^q}{(qM)^{q-1}}$$

Since every column of D has norm 1, the eigenvalues of $D^T D$ are bounded by $\mathrm{Tr}(D^T D) = q$. Hence $I \succeq \frac{1}{q} D^T D$ and so

$$D^T G D \succeq \frac{\tau^q}{(qM)^{q-1}} I \succeq \frac{\tau^q}{q^q M^{q-1}} D^T D$$

□

Block BFGS (Algorithm 1) satisfies the conditions of Lemma 6.2 when we take $\widetilde{G}_k = G_k$ and apply FILTERSTEPS (Algorithm 2). Thus Theorem 6.1 shows that Block BFGS converges for non-convex functions.

Performing updates with a filtered matrix is analogous to the cautious update (1). We can also modify G_k by adding a diagonal matrix Λ_k. This is analogous to the modified update (2).

Theorem 6.3. *Assume f is twice differentiable and $-MI \preceq G(x) \preceq MI$ for all x in the convex hull of the level set $\{x \in \mathbb{R}^n : f(x) \leq f(x_1)\}$. Let $\widetilde{G}_k = G_k + \Lambda_k$, where $\Lambda_k \preceq (M + \eta)I$ is a diagonal matrix satisfying $D_k^T (G_k + \Lambda_k) D_k \succeq \eta D_k^T D_k$. The modified Block BFGS using $\widetilde{G}_k$ converges.*

Proof. Observe that such a Λ_k always exists, and that $-MI \preceq \widetilde{G}_k \preceq (2M + \eta)I$. The conditions of Theorem 6.1 are satisfied, so this modified method converges. □

7. Numerical Experiments

We evaluate the performance of several block quasi-Newton methods by generating a *performance profile* [5], which can be described as follows. Given a set of algorithms $\mathcal{S}$ and a set of problems $\mathcal{P}$, let $t_{s,p}$ be the cost for algorithm s to complete problem p. For each problem p, let m_p be the minimum cost to solve p of any algorithm. A performance profile is a plot comparing the functions

$$\rho_s(r) = \frac{|\{p \in \mathcal{P} : t_{s,p}/m_p \leq r\}|}{|\mathcal{P}|}$$

for all $s \in \mathcal{S}$. Observe that $\rho_s(r)$ is the fraction of problems in $\mathcal{P}$ that algorithm s completed within a factor r of the cost of the best algorithm for problem p. As a reference point, we include the classical BFGS method as one of the algorithms.

For our inexact line search, we used the function `WolfeLineSearch` from *minFunc* [16], a mature and widely used Matlab library for unconstrained optimization. The line search parameters were $\alpha = 0.1$ and $\beta = 0.75$, and `WolfeLineSearch` was configured to use interpolation with an initial step size $\lambda = 1$ (options $\texttt{LS\_type} = 1, \texttt{LS\_init} = 0, \texttt{LS\_interp} = 1, \texttt{LS\_multi} = 0$).

From preliminary experiments, we found that large values of q tend to increase numerical errors, eventually leading to search directions d_k that are not descent directions. This effect is particularly pronounced when $q \geq \sqrt{n}$. In creating performance profiles, we opted for $q = \lfloor n^{1/3} \rfloor$.

7.1. **Convex Experiments.** We compared the methods listed below.

(1) *BFGS*
(2) *Block BFGS Variant 1*, or *B-BFGS1*
Block BFGS (Algorithm 1). We store the full inverse Hessian approximation H_k and compute $d_k = -H_k g_k$ by a matrix-vector product. We do not perform Algorithm 2, so the update (3.4) uses all steps.
(3) *Block BFGS Variant 2*, or *B-BFGS2*
Block BFGS (Algorithm 1), with Algorithm 2 and $\tau = 10^{-3}$. As in B-BFGS1, the full Hessian approximation H_k is stored. H_k is updated by (3.4) using the steps returned by Algorithm 2.
(4) *Block BFGS with* $q = 1$, or *B-BFGS-q1*
This compares the effect of using a single sketching equation as in Block BFGS updates versus using the standard secant equation of BFGS updates.
(5) *Rolling Block BFGS*, or *RB-BFGS*
See Section 3.2. We take a smaller value $q = \min\{3, \lfloor n^{1/3} \rfloor\}$ for this method, and omit filtering.
(6) *Gradient Descent*, or *GD*

Each algorithm is considered to have *completed* a problem when it reaches a solution with objective value less than some threshold f_{stop}. The thresholds f_{stop} are pre-computed for each problem p by minimizing p with minFunc to obtain a near-optimal solution f_*, and setting $f_{stop} = f_* + 0.01|f_*|$.

We measure the cost $t_{s,p}$ in two metrics: the number of steps, and the amount of CPU time, to completion.

7.1.1. *Logistic Regression Tests.* As in [6], we ran tests on *logistic regression* problems, a common classification technique in statistics. For our purposes, it suffices to describe the objective function. Given a set of m data points (y_i, x_i), where $y_i \in \{0, 1\}$ is the class, and $x_i \in \mathbb{R}^n$ is the vector of features of the i-th data point, we minimize, over all weights $w \in \mathbb{R}^n$, the loss function

$$L(w) = -\frac{1}{m}\sum_{i=1}^{m} \log \phi(y_i, x_i, w) + \frac{1}{2m} w^T Q w \tag{7.1}$$

$$\phi(y_i, x_i, w) = \begin{cases} \frac{1}{1+\exp(-x_i^T w)} & \text{if } y_i = 1 \\ 1 - \frac{1}{1+\exp(-x_i^T w)} & \text{if } y_i = 0 \end{cases}$$

where $Q \succ 0$ in the 'regularization' term. Figure 1 shows the performance profiles for this test. See Appendix B for a list of the data sets and our choices for Q.

In Figure 1, we see that the block methods B-BFGS1, B-BFGS2, and RB-BFGS all outperform BFGS in terms of the number of steps to completion. Considering the amount of CPU time used, B-BFGS1 is competitive with BFGS, while B-BFGS2 and RB-BFGS are more expensive than BFGS. This suggests that the additional curvature information added in block updates allows Block BFGS to find better search directions, but at the cost of the update operation being more expensive. B-BFGS-q1 and BFGS exhibit very similar performance when measured in steps, so there appears to be little difference between using a single sketching equation and a secant equation.

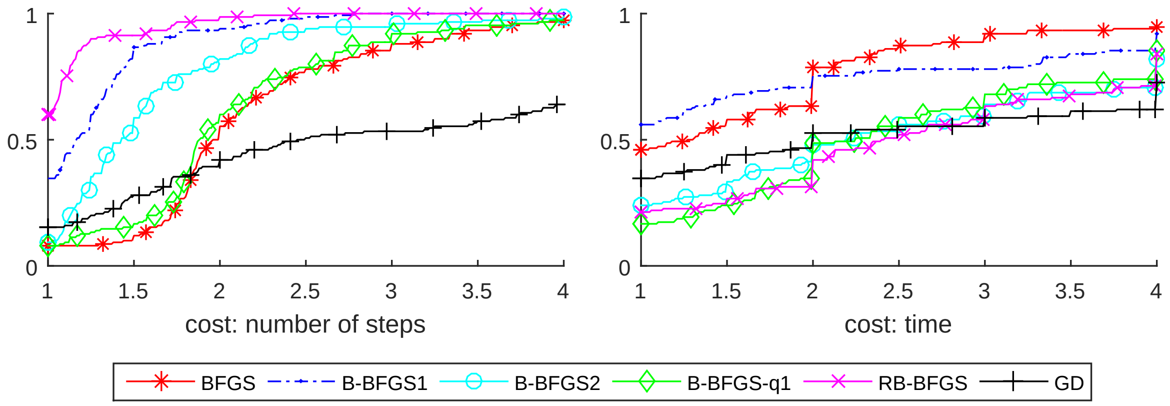

FIGURE 1. Logistic Regression profiles ($\rho_s(r)$)

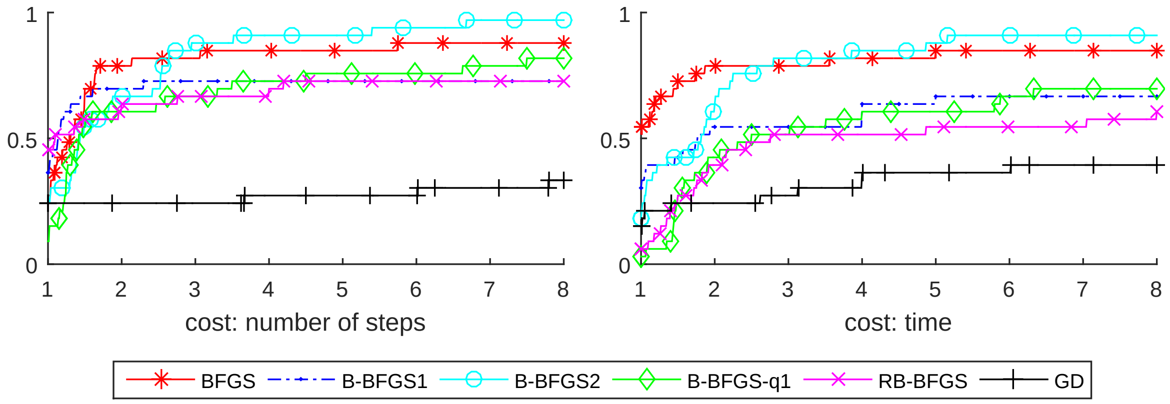

FIGURE 2. Log Barrier QP profiles ($\rho_s(r)$)

Interestingly, B-BFGS1 outperformed B-BFGS2, indicating that steps are being removed from the update, which would improve the search directions. The most likely explanation is that $\tau = 10^{-3}$ is excessively large relative to the eigenvalues of $G(x)$.

7.1.2. *Log Barrier QP Tests.* We tested problems of the form

$$\min_{y \in \mathbb{R}^s} \; F(y) = \frac{1}{2} y^T \overline{Q} y + \overline{c}^T y - 1000 \sum_{i=1}^{n} \log(\overline{b} - \overline{A} y)_i \tag{7.2}$$

where $\overline{Q} \succeq 0$, $\overline{c} \in \mathbb{R}^s$, $\overline{b} \in \mathbb{R}^n$, and $\overline{A} \in \mathbb{R}^{n \times s}$. Note that the objective value is $+\infty$ if y does not satisfy $\overline{A}y \leq \overline{b}$. In Appendix B, we explain how to derive a log barrier problem from a QP in standard form. See Figure 2 for the performance profile. Note that problems with a barrier structure are atypical in the context of unconstrained minimization, and are usually solved with specific interior point methods. However, they are somewhat interesting as they can be quite challenging to solve.

Since $\nabla^2 F(y) = Q + 1000\overline{A}^T S \overline{A}$, where S is diagonal with entries $(\overline{b} - \overline{A}y)_i^{-2}$, these problems are often extremely ill-conditioned. This leads to issues when using `WolfeLineSearch`, as the line search can require many backtracking iterations, or even fail completely, when the current iterate is near the boundary of the log barrier. This causes particular issues with block updates, as $\nabla^2 F(y)$ has small numerical rank when S has a small number of extremely large entries. Consequently, we removed problems from the test set which were ill-conditioned to the extent that even after performing step filtering, the line search failed at some step before reaching the optimal solution. Quasi-Newton methods, and those using block updates with large q in particular, are poorly suited for these ill-conditioned problems. However, we note that, although the standard BFGS method also can fail on these problems, it is more robust than the block methods.

7.2. **Non-Convex Experiments.** Since non-convex functions often have multiple stationary points, more complex behavior is possible than in the convex case. For instance, one algorithm may generally require more steps to converge, but may be taking advantage of additional information to help avoid spurious local minima.

Let f_p denote the best objective value obtained for problem p by any algorithm. To evaluate both the early and asymptotic performance of our algorithms, we generated performance profiles comparing the cost for each algorithm to reach a solution with objective value less than $f_p + \epsilon|f_p|$ for $\epsilon = 0.2$, $\epsilon = 0.1$, and $\epsilon = 0.01$. When $|f_p|$ is very small (for instance, $|f_p| < 10^{-10}$), we essentially have $f_p = 0$ and treat all solutions with objective value within 10^{-10} as being optimal.

We compared four different algorithms for non-convex minimization:

(1) *Damped BFGS*, or *D-BFGS*
  Damped BFGS with $\phi = 0.2$ (see Section 6).
(2) *Block BFGS*, or *B-BFGS*
  Block BFGS (Algorithm 1) with $q = \lfloor n^{1/3} \rfloor$ and $\tau = 10^{-5}$.
(3) *Block BFGS with* $q = 1$, or *B-BFGS-q1*
  Block BFGS (Algorithm 1) with $q = 1$ and $\tau = 10^{-5}$.
(4) *Gradient Descent*, or *GD*

7.2.1. *Hyperbolic Tangent Loss Tests.* This is also a classification technique; however, unlike the logistic regression problems in Section 7.1.1, these problems are generally non-convex. Given a set of m data points (y_i, x_i) where $y_i \in \{0, 1\}$ is the class, and $x_i \in \mathbb{R}^n$ the features, we seek to minimize over $w \in \mathbb{R}^n$ the loss function

$$L(w) = \frac{1}{m} \sum_{i=1}^{m} \big(1 - \tanh(y_i x_i^T w)\big) + \frac{1}{2m} \|w\|^2$$

Figure 3 presents performance profiles for $\epsilon = 0.2, 0.1, 0.01$, with cost measured in both steps and CPU time. See Appendix B for a list of the data sets.

B-BFGS and gradient descent perform well at first, making rapid progress to within $0.2|f_p|$ of f_p in the fewest number of steps. B-BFGS continues to converge quickly, generally requiring the fewest steps to reach $0.1|f_p|$ and $0.01|f_p|$ of f_p, while gradient descent is overtaken by BFGS and B-BFGS-q1.

Surprisingly, all four algorithms used nearly the same amount of CPU time, with each algorithm completing a majority of problems after using only 1% more time than the fastest algorithm.

7.2.2. *Standard Benchmark Tests.* This test used 19 functions from the test collection of Andrei [1], many of which originate from the CUTEst test set. The functions are listed below, with the number of variables n in parentheses:

`arwhead` (300), `bdqrtic` (200), `cube` (400), `diag1` (250), `dixonprice` (200), `edensch` (300), `eg2` (400), `explin2` (200), `fletchcr` (400), `genhumps` (250), `indef` (250), `mccormick` (400), `raydan1` (400), `rosenbrock` (300), `sine` (400), `sinquad` (400), `tointgss` (200), `trid` (200), `whiteholst` (300).

The gradients and Hessians were computed using the automatic differentiation program ADiGator [18].

For each of these functions, we generated 6 random starting points and tested the 4 algorithms using each starting point, for a total of 114 problems. Figure 4 presents performance profiles for $\epsilon = 0.2, 0.1, 0.01$, with cost measured in steps. We see from Figure 4 that D-BFGS consistently

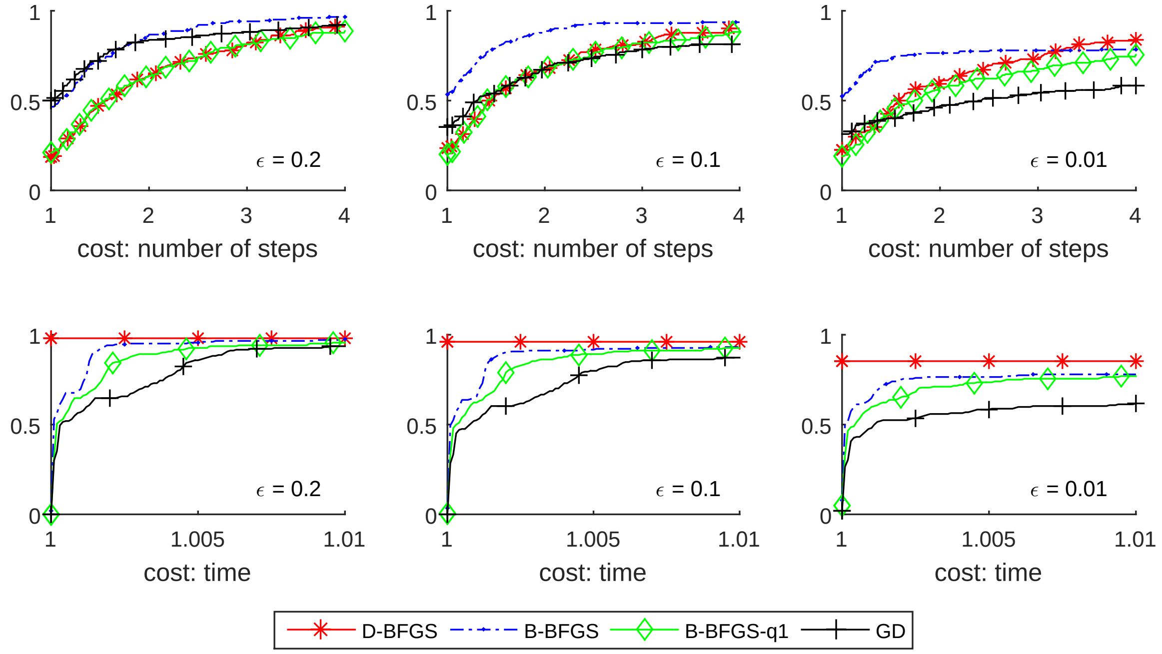

FIGURE 3. Hyperbolic Tangent Loss profiles ($\rho_s(r)$)

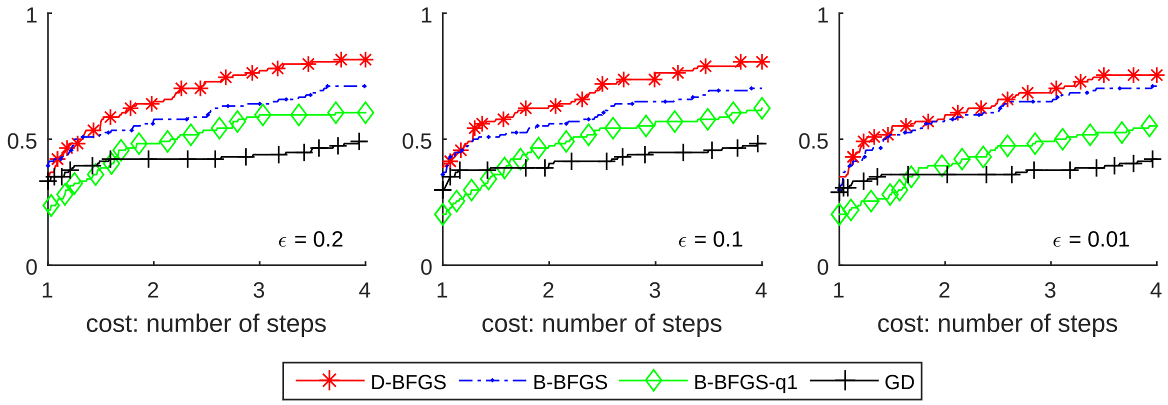

FIGURE 4. Standard Benchmark profiles ($\rho_s(r)$)

outperforms B-BFGS-q1, which suggests that Powell's damping method is superior to cautious updates.

8. Concluding Remarks

We have shown that Block BFGS provides the same theoretical rate of convergence as the classical BFGS method. Further investigation is needed to determine how Block BFGS performs on a wider range of real problems. In our experiments, we focused on a very basic implementation of Block BFGS, but many simple heuristics for improving performance and numerical stability are possible. In particular, it is important to select good values of q and τ based on insights from the problem domain. We also briefly investigated the effect of using the action of the Hessian on the previous step versus the change in gradient over the previous step (as in classical BFGS) in constructing the update. Further study of the benefits and drawbacks of such an approach would be of interest, as would study of parallel implementation. We hope that this work will serve as a useful foundation for future research on quasi-Newton methods using block updates.

References

[1] N. Andrei, *An unconstrained optimization test functions collection*, Adv. Model. Optim., 10 (2008), pp. 147–161.

[2] 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., (1987), pp. 1171–1190.

[3] C.-C. Chang and C.-J. Lin, *LIBSVM: A library for support vector machines*, ACM Trans. Intell. Syst. Technol., 2 (2011).

[4] J. E. Dennis Jr. and J. J. Moré, *Characterization of superlinear convergence and its application to quasi-Newton methods*, Math. Comp., 28 (1974), pp. 549–560.

[5] E. D. Dolan and J. J. Moré, *Benchmarking optimization software with performance profiles*, Math. Prog., 91 (2002), pp. 201–213.

[6] R. Gower, D. Goldfarb, and P. Richtárik, *Stochastic block BFGS : Squeezing more curvature out of data*, in JMLR: Workshop and Conference Proceedings, vol. 48, 2016.

[7] A. Griewank and P. L. Toint, *Local convergence analysis for partitioned quasi-Newton updates*, Numer. Math., 39 (1982), pp. 429–448.

[8] W. Hager, *A collection of optimization test problems*, `http://users.clas.ufl.edu/hager/coap/Pages/matlabpage.html`. Accessed 2016-06-13.

[9] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, 2nd ed., 2013.

[10] D.-H. Li and M. Fukushima, *A modified BFGS method and its global convergence in nonconvex minimization*, J. Comput. Appl. Math., 129 (2001), pp. 15–35.

[11] ———, *On the global convergence of the BFGS method for nonconvex unconstrained optimization problems*, Siam. J. Optim., 11 (2001), pp. 1054–1064.

[12] D. Liu and J. Nocedal, *On the limited memory BFGS method for large scale optimization*, Math. Prog., 45 (1989), pp. 503–528.

[13] I. Maros and C. Mészáros, *A repository of convex quadratic programming problems*, Optim. Methods Softw., 11 (1999), pp. 671–681.

[14] M. J. D. Powell, *Some global convergence properties of a variable metric algorithm for minimization without exact line searches*, in Nonlinear Programming, R. Cottle and C. Lemke, eds., vol. IX, SIAM-AMS Proceedings, 1976.

[15] ———, *Algorithms for nonlinear constraints that use Lagrangian functions*, Math. Prog., 14 (1978), pp. 224–248.

[16] M. Schmidt, *minFunc: unconstrained differentiable multivariate optimization in Matlab*, `http://www.cs.ubc.ca/~schmidtm/Software/minFunc.html`, (2005).

[17] R. Schnabel, *Quasi-Newton methods using multiple secant quotations*, Tech. Report CU-CS-247-83, University of Colorado at Boulder, June 1983.

[18] M. J. Weinstein and A. V. Rao, *Adigator, a toolbox for the algorithmic differentiation of mathematical functions in MATLAB using source transformation via operator overloading*, ACM Trans. Math. Soft., (submitted).

Appendix A. Derivation of the Block BFGS Update Formula

Let $\|X\|_{G_k}$ denote the matrix norm $\mathrm{Tr}(XG_kX^TG_k)$. We show that the unique solution of

$$(P) \quad \begin{cases} \min\limits_{\widetilde{H}\in\mathbb{R}^{n\times n}} & \|\widetilde{H}-H_k\|_{G_k} \\ \text{s.t} & \widetilde{H}=\widetilde{H}^T, \widetilde{H}G_kD_k=D_k \end{cases}$$

is given by formula (3.4). Introduce a new variable $E=\widetilde{H}-H_k$, and let $D=D_k$, $G=G_k$, $H=H_k, Y=G_kD_k, Z=D_k-H_kG_kD_k$. We rewrite the problem (P) in terms of E and express its Lagrangian as

$$\mathcal{L}(E,\Sigma,\Lambda)=\frac{1}{2}\,\mathrm{Tr}(EGE^TG)+\mathrm{Tr}(\Sigma(E-E^T))+\mathrm{Tr}(\Lambda^T(EY-Z)$$

Solving $\frac{\partial \mathcal{L}}{\partial E} = 0$ in terms of E, we obtain $E = -G^{-1}(Y\Lambda^T + \Sigma - \Sigma^T)G^{-1}$. Thus $E - E^T = G^{-1}(\Lambda Y^T - Y\Lambda^T + 2(\Sigma^T - \Sigma))G^{-1} = 0$, from which we obtain $\Sigma - \Sigma^T = \frac{1}{2}(\Lambda Y^T - Y\Lambda^T)$. Therefore $E = -\frac{1}{2}G^{-1}(Y\Lambda^T + \Lambda Y^T)G^{-1}$.
To solve for Λ, substituting this expression for E into the constraint $EY = Z$ yields

$$G^{-1}(Y\Lambda^T + \Lambda Y^T)G^{-1}Y + 2Z = 0 \tag{A.1}$$

Left multiplying by Y^T and using the definition $Y = GD$, we have

$$(D^TGD)(\Lambda^T D) + (D^T\Lambda)(D^TGD) + 2Y^TZ = 0$$

Now, it is easy to verify that $\Lambda^T D = -(D^TGD)^{-1}(Y^TZ)$ is the solution. Therefore, from (A.1), $\Lambda Y^T D = -Y\Lambda^T G^{-1}Y - 2GZ = Y(D^TGD)^{-1}Y^TZ - 2GZ$. Hence, $\Lambda = (Y(D^TGD)^{-1}Y^TZ - 2GZ)(D^TGD)^{-1}$. Substituting Λ into our expression for E and rearranging produces formula (3.4).

Appendix B. Details of Experiments

B.1. **Logistic Regression Tests (7.1.1).** The following 18 data sets from LIBSVM [3] were used: `a1a`, `a2a`, `a3a`, `a4a`, `australian`, `colon-cancer`, `covtype`, `diabetes`, `duke`, `ionosphere-scale`, `madelon`, `mushrooms`, `sonar-scale`, `splice`, `svmguide3`, `w1a`, `w2a`, `w3a`.
Each data set was partitioned into 3 disjoint subsets with at most 2000 points. For each subset, we have a problem of the form (7.1) with the standard L_2 regularizer $Q = I$, producing 54 standard problems. An additional 96 problems with $Q = I + Q'$ were produced by adding a randomly generated convex quadratic Q' to one of the standard problems. Two such problems were produced for each standard problem, except those from `duke` and `colon-cancer` (omitted for problem size).

B.2. **Log Barrier QP Tests (7.1.2).** Given a convex quadratic program $\min_{x\in\mathbb{R}^n}\{\frac{1}{2}x^TQx + c^Tx \mid Ax = b, x \geq 0\}$, we derive a log barrier QP problem as follows. Taking a basis N for the null space of A (of dimension s), and a solution $Ax_0 = b, x_0 \geq 0$, the given QP is equivalent to $\min_{y\in\mathbb{R}^s}\{\frac{1}{2}y^T\overline{Q}y + \overline{c}^Ty \mid \overline{A}y \leq \overline{b}\}$, where $\overline{Q} = N^TQN, \overline{c} = N^T(c + Qx_0), \overline{b} = x_0$ and $\overline{A} = -N$. Replacing the constraint by a log barrier $-\mu\sum_{i=1}^n \log(\overline{b} - \overline{A}y)_i$ (with $\mu = 1000$), we obtain problem (7.2).
This test included 43 problems in total. There were 35 log barrier problems derived from the QP test collection of Maros and Mészáros [13]:
`cvxqp1_m`, `cvxqp1_s`, `cvxqp2_m`, `cvxqp2_s`, `cvxqp3_m`, `cvxqp3_s`, `dual1`, `dual2`, `dual3`, `dual4`, `primal1`, `primal3`, `primal4`, `primalc1`, `primalc2`, `primalc5`, `primalc8`, `q25fv47`, `qbeaconf`, `qgrow15`, `qgrow22`, `qgrow7`, `qisrael`, `qscagr7`, `qscfxm1`, `qscfxm2`, `qscfxm3`, `qscorpio`, `qscrs8`, `qsctap1`, `qsctap3`, `qshare1b`, `qship08l`, `stadat1`, `stadat2`.
An additional 8 problems were derived from the following LP problems in the COAP collection [8]: `adlittle`, `agg`, `agg2`, `agg3`, `bnl1`, `brandy`, `fffff800`, `ganges`.

B.3. **Hyperbolic Tangent Loss Tests (7.2.1).** This test used the same data sets as the logistic regression test, with `duke` omitted because of large problem size ($n = 7130$). As in the logistic regression test, each data set was partitioned into 3 subsets with at most 2000 points, producing 51 loss functions. For each loss function, we tried 4 random starting points, for a total of 204 problems.