NONSMOOTH VARIANTS OF POWELL'S BFGS CONVERGENCE THEOREM

JIAYI GUO* AND A.S. LEWIS†

Abstract. The popular BFGS quasi-Newton minimization algorithm under reasonable conditions converges globally on smooth convex functions. This result was proved by Powell in a landmark 1976 paper: we consider its implications for functions that are *not* smooth. In particular, an analogous convergence result holds for functions (like the Euclidean norm) whose minimizers are isolated nonsmooth points.

Key words. convex; BFGS; quasi-Newton; Powell's theorem; nonsmooth.

AMS subject classifications. 90C30; 65K05.

1. Introduction. The BFGS (Broyden-Fletcher-Goldfarb-Shanno) method for minimizing a smooth function has been popular for decades [9]. Surprisingly, however, it can also be an effective general-purpose tool for nonsmooth optimization [5]. For twice continuously differentiable convex functions with compact level sets, Powell [10] proved global convergence of the algorithm in 1976. By contrast, in the nonsmooth case, despite substantial computational experience, the method is supported by little theory. Beyond one dimension, with the exception of some contrived model examples [6], the only previous convergence proof for the standard BFGS algorithm applied to a nonsmooth function seems to be the analysis of the two-dimensional Euclidean norm in [5].

As an illustration, consider the nonsmooth convex function $f\colon \mathbf{R}^2 \to \mathbf{R}$ defined by $f(u,v) = u^2 + |v|$. A routine implementation of the BFGS method, using a random initial point and an Armijo-Wolfe line search [5], apparently always converges to the unique optimizer at zero. Figure 1.1 plots function values for a thousand runs of BFGS for this function, against both iteration count and a count of the number of function-gradient evaluations, including those incurred in each line search. Precisely, the initial Hessian approximation is the identity, the Armijo-Wolfe line search (from [5]) uses Armijo parameter 10^{-4} and Wolfe parameter 0.9, and the initial function value is normalized to one. Although the output compellingly supports convergence, a general theoretical result, even for this very simple example, does not seem easy.

The computational results for this example even suggest a linear convergence rate. For comparison, the bold line overlaid on the first panel corresponds to one particular sequence of BFGS iterates $(2^{-k}, \frac{2}{5}(-1)^k 2^{-2k})$ generated by an exact line search [6]. The next section includes a more detailed description.

By contrast, analogous experiments with the steepest descent method — exactly the same algorithm and line search, and a random initial point, but using the steepest descent step $-\nabla f$ instead of the BFGS quasi-Newton step — produce very different results (Figure 1.2). For the function $f(u,v) = 3u^2 + |v|$, for example, steepest descent *essentially always* converges to a nonoptimal point $(u, 0)$ with $u \neq 0$. We explain this behavior in the Appendix.[1] The failure of the method of steepest descent for

*School of ORIE, Cornell University, Ithaca NY, USA; jg826@cornell.edu.

†School of ORIE, Cornell University, Ithaca NY, USA; people.orie.cornell.edu/aslewis/.
Research supported in part by National Science Foundation Grant DMS–1613996.

[1]For the original function, $f(u,v) = u^2 + |v|$, a misleading artifact of a bisection-based line search (see [5]) is convergence of steepest descent to the optimal point $(0,0)$, because the iterates accidentally land and remain on the axis $u = 0$.

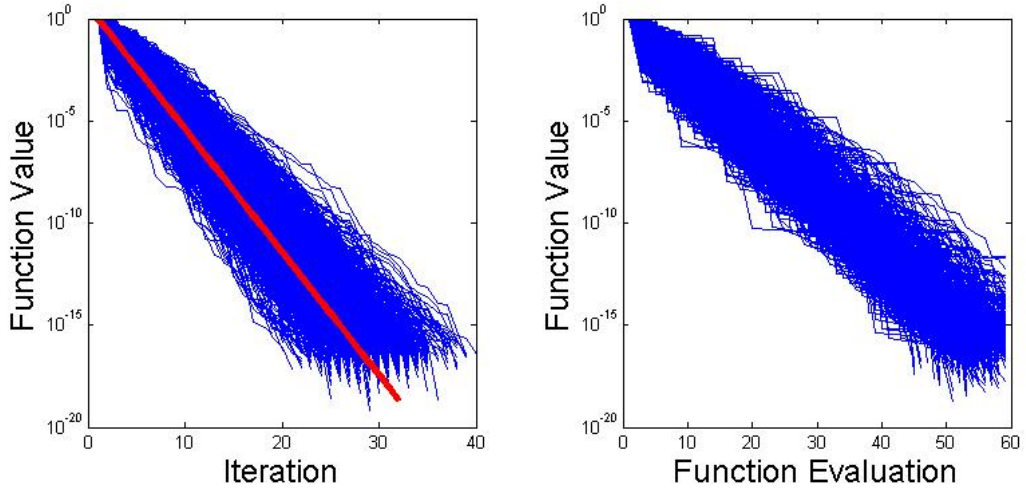

FIG. 1.1. *BFGS method for $f(u,v) = u^2 + |v|$. A thousand random starts, using inexact line search, and initial approximate Hessian I. Semilog plots of function value $f(u_k, v_k)$, initially normalized. Panel 1: against iteration count k. (Bold line plots 2^{-2k}.) Panel 2: against function evaluation count, including line search.*

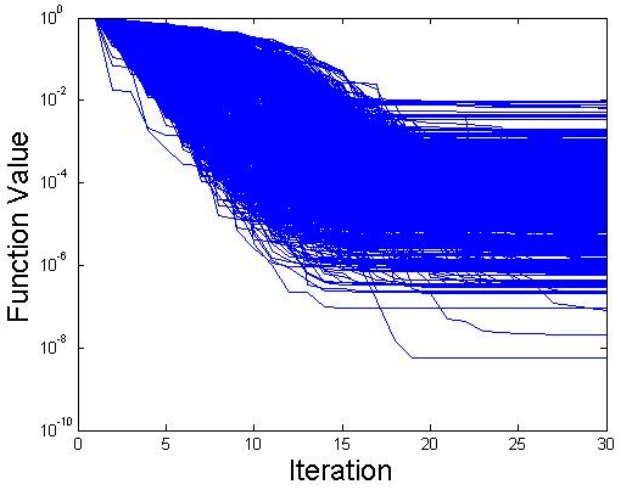

FIG. 1.2. *Steepest descent for $f(u,v) = 3u^2 + |v|$. A thousand random starts, using inexact line search. Semilog plots of function value, initially normalized, against iteration count.*

nonsmooth optimization is well-known: a simple example is [5, p. 136], and a famous example (stable to initial conditions) is [4, p. 363].

Nonetheless, Powell's theory does have consequences even in the nonsmooth case. Loosely speaking, we prove, at least under a strict-convexity-like assumption, that global convergence can only fail for the BFGS method if a subsequence of the iterates converges to a nonsmooth point — a point at which the function is not differentiable. A variation of the same technique shows, for example, for the function $f(u,v) = u^2 + |v|$, that BFGS iterates cannot remain a uniform distance away from the line

$v = 0$. While intuitive, results of this type are also reassuring, and in fact suffice to prove convergence on some interesting examples, such as the Euclidean norm on $\mathbf{R}^n$ (generalizing the result for $n = 2$ for an exact line search in [5]).

2. BFGS sequences. Given a set $U \subset \mathbf{R}^n$, we consider the BFGS method for minimizing a possibly nonsmooth function $f\colon U \to \mathbf{R}$. We call a sequence (x_k) in U "BFGS" if the BFGS method could generate it using a line search satisfying the Armijo and Wolfe conditions. More precisely, we make the following definition.

DEFINITION 2.1. *A sequence (x_k) is a* **BFGS sequence** *for the function f if f is differentiable at each iterate x_k with nonzero gradient $\nabla f(x_k)$, and there exist parameters $\mu < \nu$ in the interval $(0, 1)$ and an n-by-n positive definite matrix H_0 such that the vectors*

$$s_k = x_{k+1} - x_k \quad \textit{and} \quad y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$$

and the matrices defined recursively by

$$V_k = I - \frac{s_k y_k^T}{s_k^T y_k} \quad \textit{and} \quad H_{k+1} = V_k H_k V_k^T + \frac{s_k s_k^T}{s_k^T y_k} \tag{2.1}$$

satisfy

$$H_k \nabla f(x_k) \in -\mathbf{R}_+ s_k \tag{2.2}$$

$$f(x_{k+1}) \le f(x_k) + \mu \nabla f(x_k)^T s_k \tag{2.3}$$

$$\nabla f(x_{k+1})^T s_k \ge \nu \nabla f(x_k)^T s_k \tag{2.4}$$

for $k = 0, 1, 2, \ldots$.

Notice that this property is independent of any particular line search algorithm used to generate the sequence (x_k): it depends only on the sequences of functions values $f(x_k)$ and gradients $\nabla f(x_k)$. Conceptually, in the definition, the matrices H_k are approximate inverse Hessians for the function f at the iterate x_k: the equations (2.1) define the BFGS quasi-Newton update and the inclusion (2.2) expresses the fact that the step s_k is in the corresponding approximate Newton direction. The inequalities (2.3) and (2.4) are the Armijo and Wolfe line search conditions respectively, with parameters μ and ν respectively. By a simple and standard induction argument, they imply that the property $s_k^T y_k > 0$ then holds for all k, ensuring the matrices H_k are well-defined and positive definite, and that the function values $f(x_k)$ decrease strictly. Any implementation of the BFGS method for a convex function f using an Armijo-Wolfe line search will generate a BFGS sequence of iterates, assuming that those iterates stay in the set U and that the method never encounters a nonsmooth or critical point. (An implementation should terminate if it encounters a smooth critical point. Behavior in the event of encountering a nonsmooth point depends on the implementation, but in numerical examples like those described above, that event is rare, as we discuss later.)

Example: a simple nonsmooth function. Consider the function $f\colon \mathbf{R}^2 \to \mathbf{R}$ defined by $f(u, v) = u^2 + |v|$. (We abuse notation slightly and identify the vector $[u\ v]^T \in \mathbf{R}^2$ with the point (u, v).) Then the sequence in $\mathbf{R}^2$ defined by

$$\left(2^{-k}\ ,\ \frac{2}{5}(-1)^k 2^{-2k}\right) \quad (k = 0, 1, 2, \ldots)$$

is a BFGS sequence, as observed in [6, Prop 3.2]. Specifically, if we define a matrix

$$H_0 = \begin{bmatrix} \frac{1}{4} & 0 \\ 0 & \frac{1}{2} \end{bmatrix},$$

then the definition of a BFGS sequence holds for any parameter values $\mu \in (0, 0.7]$ and $\nu \in (\mu, 1)$. In this example, the "exact" line search property $\nabla f(x_{k+1})^T s_k = 0$ holds for all k, and the approximate inverse Hessians are

$$H_1 = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{4} \end{bmatrix}, \quad H_k = \frac{1}{6} \begin{bmatrix} 5 & (-1)^k 2^{1-k} \\ (-1)^k 2^{1-k} & 2^{3-2k} \end{bmatrix} \quad (k > 1).$$

Example: the Euclidean norm. Consider the function $f = \|\cdot\|$ on $\mathbf{R}^2$. Beginning with the initial vector $[1\ 0]^T$, generate a sequence of vectors by, at each iteration, rotating clockwise through an angle of $\frac{\pi}{3}$ and shrinking by a factor $\frac{1}{2}$. The result is a BFGS sequence for f, as observed in [5]. Specifically, if we define a matrix

$$H_0 = \begin{bmatrix} 3 & -\sqrt{3} \\ -\sqrt{3} & 3 \end{bmatrix},$$

then the definition of a BFGS sequence holds for any parameter values $\mu \in \left(0, \frac{2}{3}\right]$ and any $\nu \in (\mu, 1)$. Again, the exact line search property $\nabla f(x_{k+1})^T s_k = 0$ holds for all k. In this case the approximate inverse Hessians have eigenvalues behaving asymptotically like $2^{-k}(3 \pm \sqrt{3})$ (see [5]).

3. Main result. The following theorem captures a key global convergence property of the BFGS method.

THEOREM 3.1 (Powell, 1976). *Consider an open convex set $U \subset \mathbf{R}^n$ containing a BFGS sequence (x_k) for a convex function $f\colon U \to \mathbf{R}$. Assume that the level set $\{x \in U : f(x) \le f(x_0)\}$ is compact, and that*

$$\nabla^2 f \text{ is continuous throughout } U. \tag{3.1}$$

Then the sequence of function values $f(x_k)$ converges to $\min f$.

To better suit our basic technique, our statement of Powell's result differs slightly from the original, in which the function f was defined throughout $\mathbf{R}^n$, with Hessian continuous on the given level set. To see how our stated version follows from the original, denote the given level set by K', and the closed unit ball in $\mathbf{R}^n$ by B. By compactness, there exists a constant $\delta > 0$ such that the compact convex set $K = K' + \delta B$ is contained in the open set U. By convexity, f is L-Lipschitz on K for some constant $L > 0$. Hence there exists a convex Lipschitz function $\hat{f}\colon \mathbf{R}^n \to \mathbf{R}$ agreeing with f on K, specifically the Lipschitz regularization defined by

$$\hat{f}(y) \;=\; \min_{x \in K} \{f(x) + L\|y - x\|\} \quad (y \in \mathbf{R}^n). \tag{3.2}$$

We can now apply Powell's original result, with the function f replaced by $\hat{f}$, and Theorem 3.1 follows.

Powell's proof depends crucially on convexity. Among the assumptions, at least for dimension $n > 2$ (see [11]), convexity is central. Although the BFGS method works well in practice on general smooth functions [9], nonconvex counterexamples are known where convergence fails. Two particularly interesting examples appear in [2, 8]. Each present bounded but nonconvergent BFGS sequences, the first for a

polynomial $f\colon \mathbf{R}^4 \to \mathbf{R}$ (but with unbounded level sets), and the second for a $C^{(2)}$ smooth function with bounded level sets. In the general *convex* case, on the other hand, whether the smoothness assumption (3.1) can be weakened seems unclear.

We present here a result analogous to Powell's theorem. We modify the assumptions, strengthening the convexity assumption but weakening the smoothness requirement (3.1). Specifically, we only assume smoothness on an open set V containing the sequence and all its limit points, and over which the infimum of the objective is unchanged: notably, V might in theory exclude the minimizer. The proof of the result is a little involved, technically, but the essential idea is simple. We shrink the set V slightly and intersect it with a level set of f. On the resulting nonconvex compact set, the function f is smooth, and we can extend it a smooth convex function using a key tool from [14]. We then apply Powell's theorem to this new function, and the result follows.

Similar results to the one below hold for many common minimization algorithms possessing suitable global convergence properties in the smooth case. Such algorithms generate sequences of iterates x_k characterized by certain properties of the function values $f(x_k)$ and gradients $\nabla f(x_k)$ (for $k = 0, 1, 2, \ldots$), analogous to the definition of a BFGS sequence. Providing the algorithm generates function values $f(x_k)$ that must decrease to the minimum value $\min f$ for any convex function whose level sets are compact and whose Hessian is continuous and positive definite throughout those level sets, exactly the same proof technique applies. Examples of such algorithms include standard versions of steepest descent [9], coordinate descent (see for example [7]), and conjugate gradient methods (see for example [3]). Here we concentrate on BFGS because, in striking contrast to these methods, the BFGS method works well in practice on nonsmooth functions [5].

THEOREM 3.2. *Powell's Theorem also holds with the smoothness assumption (3.1) replaced by the following assumption:*

$$\left\{\begin{array}{l} \nabla^2 f \text{ is positive-definite and continuous throughout} \\ \text{an open set } V \subset U \text{ containing the set } \mathrm{cl}(x_k) \text{ and} \\ \text{satisfying } \inf_V f = \min f. \end{array}\right. \tag{3.3}$$

Proof. Recalling the hypotheses of Powell's theorem, we consider an open convex set $U \subset \mathbf{R}^n$ containing a BFGS sequence (x_k) for a convex function $f\colon U \to \mathbf{R}$, and we assume that the level set $\{x \in U : f(x) \le f(x_0)\}$ is compact. Under assumption (3.3), we aim to prove that the sequence of function values $f(x_k)$ converges to $\min f$.

Assume first that the theorem is true in the special case when $U = \mathbf{R}^n$ and the complement V^c is bounded. We then deduce the general case as follows. First, exactly as in equation (3.2), define a compact convex neighborhood $K \subset U$ of the level set, and a convex Lipschitz function $\hat{f}\colon \mathbf{R}^n \to \mathbf{R}$ agreeing with f on K. Now, for any sufficiently large $\beta \in \mathbf{R}$, the convex function $\tilde{f}\colon \mathbf{R}^n \to \mathbf{R}$ defined by

$$\tilde{f}(x) = \max\left\{\hat{f}(x), \frac{1}{2}\|x\|^2 - \beta\right\} \quad (x \in \mathbf{R}^n)$$

also agrees with f on K. The Hessian of $\tilde{f}$ is just the identity throughout the open set

$$W \;=\; \left\{x : \hat{f}(x) < \frac{1}{2}\|x\|^2 - \beta\right\}.$$

Furthermore, this set has bounded complement, and therefore so does the open set

$$\tilde{V} \;=\; W \;\cup\; (V \cap \operatorname{int} K).$$

Now notice that (x_k) is also a BFGS sequence for the function $\tilde{f}$, and all the assumptions of the theorem hold with f replaced by $\tilde{f}$, U replaced by $\mathbf{R}^n$, and V replaced by $\tilde{V}$. Applying the special case of the theorem, we deduce

$$f(x_k) = \tilde{f}(x_k) \to \min \tilde{f} = \min f,$$

as required.

We can therefore concentrate on the special case when $U = \mathbf{R}^n$ and the set $N = V^c$ is compact. We can assume N is nonempty, since otherwise the result follows immediately from Powell's Theorem. The convex function f is then continuous throughout $\mathbf{R}^n$. It is not constant, and hence is unbounded above. Furthermore, by the definition of a BFGS sequence, the initial point x_0 is not a minimizer, so all the level sets $\{x : f(x) \leq \alpha\}$ are compact. Since N is compact and f is continuous, we can fix a constant $\alpha > f(x_0)$ satisfying $\alpha > \max_N f$.

Since the values $f(x_k)$ are decreasing, the sequence (x_k) is bounded and hence the closure $\operatorname{cl}(x_k)$ is compact. For all sufficiently small $\epsilon > 0$, we then have

$$\operatorname{cl}(x_k) \cap (N + 2\epsilon B) = \emptyset \quad \text{and} \quad \max_{N+2\epsilon B} f < \alpha.$$

The distance function $d_N \colon \mathbf{R}^n \to \mathbf{R}$ defined by $d_N(x) = \min_N \|\cdot - x\|$ (for $x \in \mathbf{R}^n$) is continuous, so the set

$$\Omega_\epsilon \;=\; \{x : d_N(x) \geq 2\epsilon \text{ and } f(x) \leq \alpha\}$$

is compact, and is contained in the open set $\{x : d_N(x) > \epsilon\}$. On this nonconvex open set, the function f is convex, in the sense of [14], and $C^{(2)}$ with positive-definite Hessian. Hence, by [14, Theorem 3.2], there exists a $C^{(2)}$ convex function f_ϵ on a convex open neighborhood U_ϵ of the convex hull $\operatorname{conv}\Omega_\epsilon$ agreeing with f on Ω_ϵ. Our choice of ϵ ensures

$$\{x : f(x) = \alpha\} \;\subset\; \Omega_\epsilon \;\subset\; \{x : f(x) \leq \alpha\}. \tag{3.4}$$

Recall that f is a continuous convex function with compact level sets, whence

$$\operatorname{conv}\{x : f(x) = \alpha\} \;=\; \{x : f(x) \leq \alpha\}.$$

To see this, note that the right-hand side set C is convex, so it contains the left-hand side. On the other hand, C is also compact, so any line containing an element of C intersects it in a compact line segment with endpoints y and z. By continuity, $f(y) = \alpha = f(z)$, so in fact C is contained in the left-hand side.

We now return to the inclusions (3.4). Taking convex hulls implies $\operatorname{conv}\Omega_\epsilon = \{x : f(x) \leq \alpha\}$. (Although superfluous for this proof, [14, Theorem 3.2] even guarantees that f_ϵ has positive-definite Hessian on this compact convex set, and hence is strongly convex on it.)

We next observe that the level set $\{x \in U_\epsilon : f_\epsilon(x) \leq f_\epsilon(x_0)\}$ is compact, since it is contained in the compact level set $\{x : f(x) \leq \alpha\}$. Otherwise there would exist a point $x \in U_\epsilon$ satisfying $f_\epsilon(x) \leq f_\epsilon(x_0) = f(x_0) < \alpha$ and $f(x) > \alpha$. By continuity of f, there exists a point y on the line segment between x_0 and x satisfying $f(y) = \alpha$. But

then we must have $y \in \Omega_\epsilon$ and hence $f_\epsilon(y) = f(y) = \alpha$, contradicting the convexity of f_ϵ. Recall also that f_ϵ is continuous on $\{x : f(x) \leq \alpha\}$.

The values and gradients of the functions f and $f_\epsilon \colon U_\epsilon \to \mathbf{R}$ agree at each iterate x_k, so since those iterates comprise a BFGS sequence for f, they also do so for f_ϵ. We can therefore apply Theorem 3.1 to deduce

$$f(x_k) = f_\epsilon(x_k) \downarrow \min f_\epsilon \quad \text{as } k \to \infty.$$

By assumption, there exists a sequence of points $x^r \in V$ (for $r = 1, 2, 3, \ldots$) satisfying $\lim_r f(x^r) = \min f$. For any fixed index r, we know $x^r \in \Omega_\epsilon$ for all $\epsilon > 0$ sufficiently small, so we have

$$\min f \leq \lim_k f(x_k) = \min f_\epsilon \leq f_\epsilon(x^r) = f(x^r).$$

Taking the limit as $r \to \infty$ shows $\lim_k f(x_k) = \min f$, as required. □

The following consequence suggests simple examples.

COROLLARY 3.3. *Powell's Theorem also holds with smoothness assumption (3.1) replaced by the assumption that $\nabla^2 f$ is positive-definite and continuous throughout the set $\{x \in U : f(x) > \min f\}$.*

Proof. Suppose the result fails. The given set, which we denote V, must contain the set $\mathrm{cl}(x_k)$: otherwise there would exist a subsequence of (x_k) converging to a minimizer of f, and since the values $f(x_k)$ decrease monotonically, they would converge to $\min f$, a contradiction. Clearly we have $\inf_V f = \min f$. But now applying Theorem 3.2 gives a contradiction. □

COROLLARY 3.4. *Consider an open semi-algebraic convex set $U \subset \mathbf{R}^n$ containing a BFGS sequence for a semi-algebraic strongly convex function $f \colon U \to \mathbf{R}$ with compact level sets. Assume that the sequence and all its limit points lie in the interior of the set where f is twice differentiable. Then the sequence of function values converges to the minimum value of f.*

Proof. Denote the interior of the set where f is twice differentiable by V. Standard results in semi-algebraic geometry [13, p. 502] guarantee that V is dense in U, whence $\inf_V f = \min f$, and furthermore that the Hessian $\nabla^2 f$ is continuous throughout V, and hence positive-definite by strong convexity. The result now follows by Theorem 3.2. □

The open set V in the proof of Corollary 3.4, where the function f is smooth, has full measure in the underlying set U. Hence, if we initialize the algorithm in question with a starting point x_0 generated at random from a continuous probability distribution on U, and use a computationally realistic line search to generate each iterate x_k from its predecessor, then we would expect $(x_k) \subset V$ almost surely. Then, according to the result, exactly one of two cases hold.

(i) **Success**: $f(x_k) \to \min f$.
(ii) **Failure**: a subsequence of (x_k) converges to a point where f is neither smooth nor minimized.

Specifically, if case (i) fails, the result implies the existence of a nonsmooth limit point $\bar{x}$. The function values $f(x_k)$ decrease monotonically to some limit $l > \min f$, and by continuity, $f(\bar{x}) = l$.

Extensive computational experiments with BFGS suggest that case (i) holds almost surely [5]. However, as we saw in the introduction, this is not generally true for other algorithms (like steepest descent) for which analogous versions of Theorem 3.2 and Corollary 3.4 hold, and yet for which case (ii) is a real possibility. In the special situation described in Corollary 3.3, case (ii) is impossible, so analogous results

will hold for many common algorithms, like steepest descent, coordinate descent, or conjugate gradients.

4. Special constructions. Unlike Powell's original result, Theorem 3.2 requires the Hessian $\nabla^2 f$ to be positive-definite on an appropriate set, an assumption that fails for some simple but interesting examples like the Euclidean norm. We can sometimes circumvent this difficulty by a more direct construction, avoiding tools from [14]. The following result is a version of Corollary 3.3 under a more complicated but weaker assumption.

THEOREM 4.1. *Powell's Theorem also holds with the smoothness assumption (3.1) replaced by the following weaker condition:*

> *For all constants $\delta > 0$, there is a convex open neighborhood $U_\delta \subset U$ of the set $\{x \in U : f(x) \leq f(x_0)\}$, and a $C^{(2)}$ convex function $f_\delta \colon U_\delta \to \mathbf{R}$ satisfying $f_\delta(x) = f(x)$ whenever $f(x_0) \geq f(x) \geq \min f + \delta$.*

Proof. Clearly condition (3.1) implies the given condition, since we could choose $U_\delta = U$ and $f_\delta = f$. Assuming this new condition instead, suppose the conclusion of Powell's Theorem 3.1 fails, so there exists a number $\delta > 0$ such that $f(x_k) > \min f + 2\delta$ for all $k = 0, 1, 2, \ldots$. Consider the function f_δ guaranteed by our assumption. Since f is continuous, there exists a point $\bar{x} \in U$ satisfying $f(\bar{x}) = \min f + \delta$, and since $f_\delta(\bar{x}) = f(\bar{x})$, we deduce $\min f_\delta \leq \min f + \delta$.

Since (x_k) is a BFGS sequence for the function f, it is also a BFGS sequence for the function f_δ. Applying Theorem 3.1 with f replaced by f_δ shows the contradiction

$$\min f + 2\delta \leq f(x_k) = f_\delta(x_k) \downarrow \min f_\delta \leq \min f + \delta,$$

so the result follows. □

We can apply this result directly to the Euclidean norm.

COROLLARY 4.2. *Any BFGS sequence for the Euclidean norm on $\mathbf{R}^n$ converges to zero.*

Proof. For any $\delta > 0$, consider the function $g_\delta \colon \mathbf{R} \to \mathbf{R}$ defined by

$$g_\delta(t) = \begin{cases} \frac{\delta^3 + 3\delta t^2 - |t|^3}{3\delta^2} & (|t| \leq \delta) \\ \\ |t| & (|t| \geq \delta). \end{cases} \tag{4.1}$$

This function is $C^{(2)}$, convex and even. The function $f_\delta \colon \mathbf{R}^n \to \mathbf{R}$ defined by $f_\delta(x) = g_\delta(\|x\|)$ is also $C^{(2)}$ and convex, both as a consequence of [12] and via a straightforward direct calculation. The result now follows from Theorem 4.1. □

Analogously, the following result is a more direct version of Theorem 3.2.

THEOREM 4.3. *Powell's Theorem also holds with the smoothness assumption (3.1) replaced by the assumption that some open set $V \subset U$ containing the set $\mathrm{cl}(x_k)$ and satisfying $\inf_V f = \min f$ also satisfies the following condition:*

> *For all constants $\delta > 0$, there is a convex open neighborhood $U_\delta \subset U$ of the set $\{x \in U : f(x) \leq f(x_0)\}$, and a $C^{(2)}$ convex function $f_\delta \colon U_\delta \to \mathbf{R}$ satisfying $f_\delta(x) = f(x)$ for all points $x \in U_\delta$ such that $d_{V^c}(x) > \delta$.*

Proof. Denote the distance between the compact set $\mathrm{cl}(x_k)$ and the closed set V^c by $\bar{\delta}$, so we know $\bar{\delta} > 0$. For any constant $\delta \in (0, \bar{\delta})$, we have $d_{V^c}(x_k) > \delta$ for all indices $k = 0, 1, 2, \ldots$, and hence $f_\delta(x_k) = f(x_k)$.

The values and gradients of the functions f and f_δ agree at each iterate x_k, so since those iterates comprise a BFGS sequence for f, they also do so for f_δ. We can therefore apply Theorem 3.1 to deduce

$$f(x_k) = f_\delta(x_k) \downarrow \min f_\delta \quad \text{as } k \to \infty.$$

By assumption, there exists a sequence of points $x^r \in V$ (for $r = 1, 2, 3, \ldots$) satisfying $\lim_r f(x^r) = \min f$. For any fixed index r, we know $d_{V^c}(x^r) > \delta$ for all sufficiently small $\delta > 0$, so since $f_\delta(x^r) = f(x^r)$, we deduce $\min f_\delta \le f(x^r)$. The inequality $\lim_k f(x_k) \le f(x^r)$ follows, and letting $r \to \infty$ proves $\lim_k f(x_k) = \min f$ as required. □

We end by proving a claim from the introduction.

COROLLARY 4.4. *Any BFGS sequence for the function $f\colon \mathbf{R}^2 \to \mathbf{R}$ given by $f(u, v) = u^2 + |v|$ has a subsequence converging to a point on the line $v = 0$.*

Proof. Suppose the result fails, so some BFGS sequence $\big((u_k, v_k)\big)$ has its closure contained in the open set

$$V \;=\; \{(u, v) \in \mathbf{R}^2 : v \neq 0\}.$$

Clearly we have $\inf_V f = \min f$. For any constant $\delta > 0$, define a function $f_\delta\colon \mathbf{R}^2 \to \mathbf{R}$ by $f(u, v) = u^2 + g_\delta(v)$, where the function g_δ is given by equation (4.1). Then we have $f(u, v) = f_\delta(u, v)$ for any point (u, v) satisfying $|v| > \delta$, or equivalently $d_{V^c}(u, v) > \delta$. Hence the assumptions of Theorem 4.3 hold (using the set $U_\delta = \mathbf{R}^2$), so we deduce $f(u_k, v_k) \to 0$, and hence $(u_k, v_k) \to (0, 0)$. This contradiction completes the proof. □

As we remarked in the introduction, numerical evidence strongly supports a conjecture that all BFGS sequences for the function $f(u, v) = u^2 + |v|$ converge to zero. That conjecture remains open.

5. Appendix: failure of steepest descent. In the introduction we claimed that the method of steepest applied to the function $f\colon \mathbf{R}^2 \to \mathbf{R}$ defined by $f(u, w) = u^2 + |w|$ typically converges to a nonoptimal point with $u \neq 0$. We here explain that claim. The argument is motivated by an analogous earlier result [1] for the function $u + |w|$.

We consider any sequence of points $x_k = (u_k, w_k) \in \mathbf{R}^2$ (for $k = 0, 1, 2, \ldots$) associated with the method of steepest descent using the Armijo and Wolfe conditions. More precisely, we assume that Definition 2.1 holds, but with the approximate Hessian definition (2.1) replaced simply by $H_k = I$ for all iterations k. (In particular, by assumption, the component w_k is always nonzero.) We prove that *no* such sequence can converge to the minimizer at zero unless the component u_k is zero for some iteration k.

Suppose the initial value $f(x_0)$ is less than the positive constant

$$\delta \;=\; \min\Big\{\frac{1}{4}, \frac{1-\nu}{4\nu}, \frac{\mu}{8(1-\mu)}\Big\}.$$

Since $f(x_k)$ is decreasing, we deduce that both components u_k^2 and $|w_k|$ are less than δ for all iterations k. (Notice that any infinite sequence (x_k) converging to zero must eventually enter the region where f is less than δ.)

At each iteration $k = 0, 1, 2, \ldots$, the iterates are related by

$$(u_{k+1}, w_{k+1}) \;=\; \big(u_k(1 - 2t_k), w_k - \operatorname{sgn}(w_k)t_k\big),$$

for some scalar $t_k > 0$. To simplify notation, denote the current iterate (u_k, w_k) simply by (u, w), and t_k by t, and suppose for simplicity $w > 0$. The Wolfe condition (2.4) simplifies to

$$4u^2(1 - 2t - \nu) + \operatorname{sgn}(w - t) \;\leq\; \nu,$$

and as a consequence we have $t > w$: otherwise we deduce $t < w < \frac{1}{2}$, whence

$$4u^2(1 - 2t - \nu) + 1 \;\leq\; \nu,$$

giving $1 - 4u^2\nu \leq \nu$, contradicting our choice of δ.

Next, simplifying the Armijo condition (2.3) gives

$$t\big(1 + \mu + 4u^2(\mu - 1 + t)\big) \;\leq\; 2w.$$

Notice

$$t < \frac{2w}{1 + \mu/2},$$

since otherwise we deduce

$$1 + \mu + 4u^2(\mu - 1 + t) \;<\; 1 + \frac{\mu}{2},$$

whence

$$4u^2 \geq \frac{\mu}{2(1 - \mu)},$$

again contradicting our choice of δ. We deduce

$$w_k > 0 > w_{k+1} > -\gamma w_k,$$

where

$$\gamma = \frac{2 - \mu}{2 + \mu} \in (0, 1)$$

By induction, we now see that the component w_k changes sign at each iteration k, and satisfies

$$|w_k| \leq \gamma^k |w_0|, \quad \text{for all } k = 0, 1, 2, \ldots.$$

In particular we see $w_k \to 0$.

Without loss of generality, suppose $u_0 > 0$. Since

$$t_k < \frac{2|w_k|}{1 + \mu/2} < \frac{1}{2}$$

(because $|w_k| < 1/4$) and

$$u_{k+1} = u_k(1 - 2t_k),$$

we deduce by induction that the sequence (u_k) is strictly positive and decreasing. In fact we have

$$u_{k+1} \;>\; u_k\Big(1 - \frac{4|w_k|}{1 + \mu/2}\Big) \;\geq\; u_k\Big(1 - \frac{4|w_0|}{1 + \mu/2}\gamma^k\Big) \;\geq\; u_k(1 - \gamma^k)$$

so

$$u_k \;\geq\; u_0 \prod_{j=1}^{k} (1-\gamma^j).$$

Consequently we have

$$\log u_k \;\geq\; \log u_0 + \sum_{j=1}^{k} \log(1-\gamma^j).$$

Since the function $\tau \mapsto \log(1-\tau)$ is concave on the interval $[0,1)$, an easy argument shows that the function $\tau \mapsto \frac{1}{\tau}\log(1-\tau)$ is decreasing, so

$$\log(1-\tau) \geq \frac{\tau}{\gamma}\log(1-\gamma) \quad \text{whenever} \quad 0 < \tau < \gamma.$$

Hence we deduce

$$\log u_k \;\geq\; \log u_0 + \log(1-\gamma)\sum_{j=1}^{k} \gamma^{j-1} \;\geq\; \log u_0 + \frac{\log(1-\gamma)}{1-\gamma}$$

(summing the geometric series), and so $\lim_k u_k > 0$. We have shown that the sequence (u_k, w_k) converges to a nonzero point on the axis $w = 0$.

REFERENCES

[1] A. Asl and M.L. Overton. Analysis of the gradient method with an Armijo-Wolfe line search on a class of nonsmooth convex functions. `arXiv:1711.08517`, 2017.

[2] Y.-H. Dai. A perfect example for the BFGS method. *Math. Program.*, 138:501–530, 2013.

[3] J.C. Gilbert and J. Nocedal. Global convergence properties of conjugate gradient methods for optimization. *SIAM J. Optim.*, 2:21–42, 1992.

[4] J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I.* Springer-Verlag, Berlin, 1993.

[5] A.S. Lewis and M.L. Overton. Nonsmooth optimization via quasi-Newton methods. *Math. Program.*, 141:135–163, 2013.

[6] A.S. Lewis and S. Zhang. Nonsmoothness and a variable metric method. *J. Optim. Theory Appl.*, 165:151–171, 2015.

[7] Z.Q. Luo and P. Tseng. On the convergence of the coordinate descent method for convex differentiable minimization. *J. Optim. Theory Appl.*, 72(1):7–35, 1992.

[8] W.F. Mascarenhas. The BFGS method with exact line searches fails for non-convex objective functions. *Math. Program.*, 99:49–61, 2004.

[9] J. Nocedal and S.J. Wright. *Numerical Optimization.* Springer Series in Operations Research and Financial Engineering. Springer, New York, second edition, 2006.

[10] M.J.D. Powell. Some global convergence properties of a variable metric algorithm for minimization without exact line searches. In *Nonlinear Programming (Proc. Sympos., New York, 1975)*, pages 53–72. SIAM–AMS Proc., Vol. IX. Amer. Math. Soc., Providence, R. I., 1976.

[11] M.J.D. Powell. On the convergence of the DFP algorithm for unconstrained optimization when there are only two variables. *Math. Program.*, 87:281–301, 2000. Studies in algorithmic optimization.

[12] H.S. Sendov. Nonsmooth analysis of Lorentz invariant functions. *SIAM J. Optim.*, 18:1106–1127, 2007.

[13] L. van den Dries and C. Miller. Geometric categories and o-minimal structures. *Duke Math. J.*, 84:497–540, 1996.

[14] M. Yan. Extension of convex function. *J. Convex Anal.*, 21:965–987, 2014.