BFGS convergence to nonsmooth minimizers of convex functions

J. Guo [*]   A.S. Lewis [†]

March 19, 2017

Abstract

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

Key words: convex; BFGS; quasi-Newton; nonsmooth.

AMS 2000 Subject Classification: 90C30; 65K05.

1 Introduction

The BFGS (Broyden-Fletcher-Goldfarb-Shanno) method for minimizing a smooth function has been popular for decades [6]. Surprisingly, however, it can also be an effective general-purpose tool for nonsmooth optimization [3]. For twice continuously differentiable convex functions with compact level sets, Powell [7] 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 [4], 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 [3].

As a simple 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

---

[*]ORIE, Cornell University, Ithaca, NY 14853, U.S.A. `jg826@cornell.edu`.

[†]ORIE, Cornell University, Ithaca, NY 14853, U.S.A. `people.orie.cornell.edu/aslewis`.
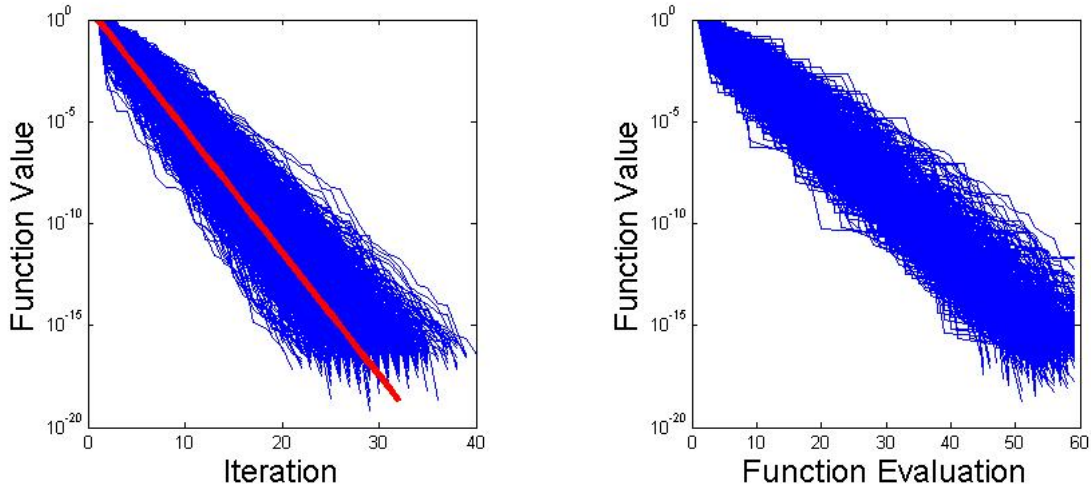
Figure 1: BFGS method for $f(u,v) = u^2 + |v|$. A thousand random starts, using inexact line search, and initial approximate Hessian I. Semilog plots of function value $f(u_k, v_k)$, initially normalized. Panel 1: against iteration count k. (Bold line plots 2^{-2k}.) Panel 2: against function evaluation count, including line search.

a random initial point and a standard backtracking line search, invariably converges to the unique optimizer at zero. Not surprisingly, the method of steepest descent, using the same line search, often converges to a nonoptimal point $(u, 0)$ with $u \neq 0$.

For example, Figure 1 plots function values for a thousand runs of BFGS 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 weak Wolfe line search uses Armijo parameter 10^{-4} and Wolfe parameter 0.9, and the initial function value is normalized to one.) The results compellingly support convergence, and indeed suggest a linear rate: the bold line overlaid on the first panel corresponds to the BFGS iterates $(2^{-k}, \frac{2}{5}(-1)^k 2^{-2k})$ generated by an exact line search [4]. However, even for this very simple example, a general convergence result does not seem easy.

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. For example, for the function $f(u, v) = u^2 + |v|$, BFGS iterates cannot remain a uniform distance away from the line $v = 0$. While intuitive — a successful smooth algorithm should somehow detect nonsmoothness — this result is also reassuring, and in fact suffices to prove convergence on some interesting examples. An analogous technique proves convergence for the Euclidean

2

norm on $\mathbf{R}^n$, generalizing the result for $n = 2$ in [3].

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 weak 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

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

and the matrices defined recursively by

(2.2) $$V_k = I - \frac{s_k y_k^T}{s_k^T y_k} \quad \text{and} \quad H_{k+1} = V_k H_k V_k^T + \frac{s_k s_k^T}{s_k^T y_k}$$

satisfy

$$
\begin{array}{rll}
(2.3) & H_k \nabla f(x_k) \in & -\mathbf{R}_+ s_k \\
(2.4) & f(x_{k+1}) \leq & f(x_k) + \mu \nabla f(x_k)^T s_k \\
(2.5) & \nabla f(x_{k+1})^T s_k \geq & \nu \nabla f(x_k)^T s_k
\end{array}
$$

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 (2.2) define the BFGS quasi-Newton update and the inclusion (2.3) expresses the fact that the step s_k is in the corresponding approximate Newton direction. The inequalities (2.4) and (2.5) are the Armijo and weak Wolf 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. An implementation of the BFGS method for a convex function f using a standard backtracking 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.

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 [4, 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 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 [3]. Specifically, if we define a matrix

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

then the 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 [3]).

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 bounded, and that*

(3.2) $$\nabla^2 f \text{ is continuous throughout } U.$$

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

4

Among the assumptions in Powell's theorem, at least for dimension $n > 2$ (see [8]), convexity is central. Although the BFGS method works well in practice on general smooth functions [6], nonconvex counterexamples are known where convergence fails: in particular, [1] presents a bounded but nonconvergent BFGS sequence for a polynomial $f \colon \mathbf{R}^4 \to \mathbf{R}$. In the general *convex* case, on the other hand, whether the smoothness assumption (3.2) 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 (3.2). 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 bounded 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 [6], coordinate descent (see for example [5]), and conjugate gradient methods (see for example [2]). Here we concentrate on BFGS because, in striking contrast to these methods, the BFGS method works well in practice on nonsmooth functions [3].

Theorem 3.3 *Powell's Theorem also holds with the smoothness assumption (3.2) replaced by the following assumption:*

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

Proof We consider an open convex set $U \subset \mathbf{R}^n$ containing a BFGS sequence (x_k) for a convex function $f \colon U \to \mathbf{R}$ satisfying assumption (3.4). We further assume that the level set $\{x \in U : f(x) \le f(x_0)\}$ is bounded, and our aim is 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. Note by assumption, that the function f is not constant, so by convexity there exists a point $\bar{x} \in U$ with $f(\bar{x}) > f(x_0)$. Convexity also ensures that f is L-Lipschitz on the nonempty compact convex set

$$
K \;=\; \{x \in U : f(x) \le f(\bar{x})\},
$$

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

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

5

Now, for any sufficiently large $\beta \in \mathbf{R}$, the convex function $\tilde{f} : \mathbf{R}^n \to \mathbf{R}$ 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

$$\tilde{V} = W \cup \{x \in V : f(x) < f(\bar{x})\}.$$

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 $\tilde{V}$. 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 assumption, the initial point x_0 is not a minimizer, so all the level sets $\{x : f(x) \le \alpha\}$ 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 $\mathrm{cl}(x_k)$ is compact. For all sufficiently small $\epsilon > 0$, we then have

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

where B denotes the closed unit ball in $\mathbf{R}^n$. The distance function $d_N : \mathbf{R}^n \to \mathbf{R}$ defined by $d_N(x) = \min_N \|\cdot - x\|$ (for $x \in \mathbf{R}^n$) is continuous, so the set

$$\Omega_\epsilon = \{x : d_N(x) \ge 2\epsilon \text{ and } f(x) \le \alpha\}$$

is compact, and is contained in the open set $\{x : d_N(x) > \epsilon\}$. On this open set, the function f is convex, in the sense of [11], and $C^{(2)}$ with positive-definite Hessian. Hence, by [11, Theorem 3.2], there exists a $C^{(2)}$ convex function f_ϵ on a convex open neighborhood U_ϵ of the convex hull $\mathrm{conv}\,\Omega_\epsilon$ agreeing with f on Ω_ϵ. Our choice of ϵ ensures

$$\{x : f(x) = \alpha\} \subset \Omega_\epsilon \subset \{x : f(x) \le \alpha\},$$

6

so in fact $\operatorname{conv} \Omega_\epsilon = \{x : f(x) \leq \alpha\}$. (Although superfluous for this proof, [11, 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 bounded, since it is contained in the 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.

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. $\qquad \square$

The following consequence suggests simple examples.

Corollary 3.5 *Powell's Theorem also holds with smoothness assumption (3.2) 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 $\operatorname{cl}(x_k)$: 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 3.3 gives a contradiction. $\qquad \square$

Corollary 3.6 *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 bounded 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 [10, 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 3.3. $\qquad\square$

The open set V in the proof of Corollary 3.6, 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, one (or both) of two cases hold.

(i) The algorithm succeeds: $f(x_k) \to \min f$.

(ii) A subsequence of the iterates converges to a point where f is nonsmooth.

Extensive computational experiments suggest case (i) holds almost surely [3].

Like Theorem 3.3, analogous versions of Corollary 3.6 hold for many other algorithms, in addition to the BFGS method. By contrast with BFGS, however, those algorithms often fail in general, due to the possibility of case (ii). In the special situation described in Corollary 3.5, case (ii) implies case (i), 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 3.3 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 [11]. The following result is a version of Corollary 3.5 under a more complicated but weaker assumption.

Theorem 4.1 *Powell's Theorem also holds with the smoothness assumption (3.2) 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 (3.2) 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

(4.3)
$$g_\delta(t) = \begin{cases} \frac{\delta^3 + 3\delta t^2 - |t|^3}{3\delta^2} & (|t| \leq \delta) \\ \\ |t| & (|t| \geq \delta). \end{cases}$$

This function is $C^{(2)}$ convex and symmetric. The function $f_\delta \colon \mathbf{R}^n \to \mathbf{R}$ defined by $f_\delta(x) = g_\delta(\|x\|)$ is also $C^{(2)}$ convex, either as a consequence of [9] or via a straightforward direct calculation. The result now follows from Theorem 4.1. □

Analogously, the following result is a more direct version of Theorem 3.3.

Theorem 4.4 *Powell's Theorem also holds with the smoothness assumption (3.2) 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

9

sufficiently small $\delta > 0$, so since $f_\delta(x^r) = f(x^r)$, we deduce $\min f_\delta \leq f(x^r)$. The inequality $\lim_k f(x_k) \leq f(x^r)$ follows, and letting $r \to \infty$ proves $\lim_k f(x_k) = \min f$ as required. $\qquad\square$

We end by proving a claim from the introduction.

Corollary 4.5 *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 (4.3). 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 4.4 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. $\qquad\square$

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.

References

[1] Y.-H. Dai. A perfect example for the BFGS method. *Mathematical Programming*, 138(1):501–530, 2013.

[2] J.C. Gilbert and J. Nocedal. Global convergence properties of conjugate gradient methods for optimization. *SIAM J. Optim.*, 2(1):21–42, 1992.

[3] A.S. Lewis and M.L. Overton. Nonsmooth optimization via quasi-Newton methods. *Math. Program.*, 141(1-2, Ser. A):135–163, 2013.

[4] A.S. Lewis and S. Zhang. Nonsmoothness and a variable metric method. *J. Optim. Theory Appl.*, 165(1):151–171, 2015.

[5] 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.

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

[7] 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.

[8] M.J.D. Powell. On the convergence of the DFP algorithm for unconstrained optimization when there are only two variables. *Math. Program.*, 87(2, Ser. B):281–301, 2000. Studies in algorithmic optimization.

[9] H.S. Sendov. Nonsmooth analysis of Lorentz invariant functions. *SIAM J. Optim.*, 18(3):1106–1127, 2007.

[10] L. van den Dries and C. Miller. Geometric categories and o-minimal structures. *Duke Math. J.*, 84(2):497–540, 1996.

[11] M. Yan. Extension of convex function. *J. Convex Anal.*, 21(4):965–987, 2014.