On Convergence of Minimization Methods: Attraction, Repulsion and Selection

Yin Zhang* Richard Tapia† Leticia Velázquez‡

CAAM Technical Report TR99-12
March, 1999 (Revised March, 2000)

Abstract

In this paper, we revisit the convergence properties of the iteration process

$$x^{k+1} = x^k - \alpha(x^k)B(x^k)^{-1}\nabla f(x^k)$$

for minimizing a function $f(x)$. After reviewing some classic results and introducing the notion of strong attraction, we give necessary and sufficient conditions for a stationary point of $f(x)$ to be a point of strong attraction for the iteration process. This result not only gives a new algorithmic interpretation to the classic Ostrowski theorem, but also provides insight into the interesting phenomenon called selective minimization. We also present illustrative numerical examples arising from nonlinear least squares problems.

Keywords: Attraction, repulsion, selective minimization.
AMS Classification: 65K05, 90C30

1 Introduction

We consider the unconstrained minimization problem

$$\min f(x), \tag{1}$$

*Department of Computational and Applied Mathematics, Rice University, Houston, Texas 77005, USA. (zhang@caam.rice.edu). This author was supported in part by DOE Grant DE-FG03-97ER25331, DOE/LANL Contract 03891-99-23 and NSF Grant DMS-9973339.

†Department of Computational and Applied Mathematics, Rice University, Houston, Texas 77005, USA. This author was supported in part by DE-FG03-93ER25178 and DOE/LANL Contract 03891-99-23.

‡Department of Computational and Applied Mathematics, Rice University, Houston, Texas 77005, USA. This author was supported in part by NSF RTG Grant BIR-94-13229 (the Keck Center on Computational Biology) and Sloan Foundation Grant 94-12-12.

where $f : \Re^n \to \Re$ is assumed to be twice Frechet differentiable. Many iterative methods for solving (1) can be represented by the general iteration:

$$x^{k+1} = x^k - \alpha^k (B^k)^{-1} \nabla f(x^k). \tag{2}$$

This iterative framework includes Newton's method, quasi-Newton methods and gradient methods with variant step-length control schemes. In the case of nonlinear least squares problems, it also includes the Gauss-Newton and the Levenberg-Marquardt methods.

The convergence properties of the iterative framework (1) have often been studied through the more general scheme of fixed-point iteration:

$$x^{k+1} = T(x^k) \tag{3}$$

for some function $T : \Re^n \to \Re^n$. For (2), $T(x)$ is defined as

$$T(x) = x - \alpha(x) B(x)^{-1} \nabla f(x). \tag{4}$$

It is clear that any stationary point x^* of f, where $\nabla f(x^*) = 0$, is a fixed point of the function (4) that satisfies the equation $x = T(x)$.

The iterative framework (2) has been studied extensively and many results are available for various choices of α^k and B^k that guarantee convergence, for example see the classic books by Ostrowski [7], Ortega and Rheinboldt [6], and Dennis and Schnabel [2] on this subject.

A less frequently asked question is the following. Given certain conditions on α^k and B^k, what type of stationary points of $f(x)$ are or are not points of attraction of the iteration (2)? In this paper, we present some observations in this aspect. A particular interesting observation is that for proper choices of B^k and α^k, it is possible to construct iteration (2) so that certain undesirable minimizers of $f(x)$ become points of repulsion, while more desirable minimizers remain points of attraction of iteration (2). We will call this phenomenon selective minimization.

In order to classify fixed points of (3), we will need the derivative of $T(x)$, $T'(x)$, at stationary points of $f(x)$. The following proposition shows that for $T'(x)$ to exist at a stationary point x^* of $f(x)$, the function $\alpha(x)B(x)^{-1}$ need not be differentiable at x^*; instead, continuity at x^* will suffice. This result is a special case of **10.2.1** in Ortega and Rheinboldt [6]. For completeness, we include a short proof.

Proposition 1 *Let x^* be a stationary point of $f(x)$. Assume that $\alpha(x)$ and $B(x)$ are continuous at x^* where $B(x)$ is also nonsingular. Then the derivative of $T(x)$ in (4) exists at x^*, and*

$$T'(x^*) = I - \alpha(x^*) B(x^*)^{-1} \nabla^2 f(x^*). \tag{5}$$

Proof. Let $H(x) \equiv \alpha(x)B(x)^{-1}$. It suffices to show that the derivative of $H(x)\nabla f(x)$ exists at x^* and is $H(x^*)\nabla^2 f(x^*)$. The continuity of $\alpha(x)$ and $B(x)$ at x^* plus the nonsingularity of $B(x^*)$ imply the continuity of $H(x)$ at x^*. Noting $\nabla f(x^*) = 0$, we consider

$$\begin{aligned} & [H(x^*+h)\nabla f(x^*+h) - H(x^*)\nabla f(x^*) - H(x^*)\nabla^2 f(x^*)h]/\|h\| \\ = \; & H(x^*+h)[\nabla f(x^*+h) - \nabla f(x^*) - \nabla^2 f(x^*)h]/\|h\| \\ & +[H(x^*+h) - H(x^*)](\nabla^2 f(x^*)h/\|h\|). \end{aligned}$$

By continuity of $H(x)$ and differentiability of $f(x)$ at x^*, both terms on the right-hand side vanish as $\|h\| \to 0$. This completes the proof. ∎

We mention that the continuity assumption on $B(x)$ and $\alpha(x)$ at x^* may not always be satisfied by some popular methods. The following example is from Wolfe [8]. When the steepest descent method with exact line search is applied to the function $f(x,y) = x^3/3 + y^2/2$ starting from $(0.5, 0.5)$, the iterates will converge to the stationary point $(0,0)$. However, the observed step size oscillates between 5 and 1.25 as the iterates approach the stationary point.

This paper is organized as follows. In Sections 2, we will introduce the definitions of points of attraction and repulsion, and state some classic results on attraction and repulsion for the iteration (3). In Section 3, we give necessary and sufficient conditions for stationary points of $f(x)$ to be points of attraction of iteration (2). The results of Section 3 are applied to nonlinear least squares problems in section 4. We devote Section 5 to the discussion of selective minimization. Finally, in Section 6 we present two numerical examples to illustrate that (i) convergence to a point of repulsion appears to be unlikely in general, and (ii) selective minimization may be useful for certain global optimization problems.

In this paper, we will use the following notation. The spectral radius of a matrix M is denoted by $\rho(M)$, and an eigenvalue by $\lambda_i(M)$. Moreover, $\lambda_{\max}(M)$ and $\lambda_{\min}(M)$ are, respectively, the maximum and minimum eigenvalues of a symmetric matrix M. We use the the usual partial ordering for symmetric matrices: $A \succeq B$ means $A - B$ is positive semidefinite; similarly for the relationships $\preceq$, $\succ$ and $\prec$. The norm $\|\cdot\|$ will be either the Euclidean norm for vectors or the norm it induces for matrices, unless otherwise specified.

2 Points of Attraction and Repulsion

In this section, we state the definitions of points of attraction and of repulsion for the iteration (3), first introduced by Ostrowski in [7], and then review some basic results in regard to attraction and repulsion.

Definition 1 (Attraction and Repulsion) *Let X^* be a fixed point of $T(x)$. It is said to be a point of attraction of the iteration (3) if there is a neighborhoods N of x^* such that for any point $x^0 \in N$, the iterates $\{x^k\}$ generated by (3) all lie in N and converge to x^*. Otherwise, it is a point of repulsion of iteration (3).*

The well-known Ostrowski Theorem (Theorem 22.1 in [7] and **10.1.3** in Ortega and Rheinboldt [6]) states that a sufficient condition for a stationary point x^* to be a point of attraction of the iteration (3), assuming that $T(x)$ is differentiable at x^*, is that the spectral radius of $T'(x^*)$ be less than one, i.e., $\rho(T'(x^*)) < 1$.

The above condition is sufficient but not necessary. There are indeed points of attraction x^* where $\rho(T'(x^*)) = 1$ (see **10.1.x** in [6], for example). For convenience of discussion, in this paper we will call a stationary point x^* satisfying the condition $\rho(T'(x^*)) < 1$ a point of strong attraction.

Definition 2 (Strong Attraction) *A fixed point x^* of $T(x)$ is said to be a point of strong attraction of the iteration (3) if $T(x)$ is differentiable at x^* and $\rho(T'(x^*)) < 1$.*

On the other hand, Ostrowski also established (Theorem 22.2 in [7]) that a sufficient condition for a stationary point x^* to be a point of repulsion of the iteration (3), again assuming that $T(x)$ is differentiable at x^*, is that the spectral radius of $T'(x^*)$ be greater than one, i.e., $\rho(T'(x^*)) > 1$. Again this condition is sufficient but not necessary because there are points of repulsion x^* that satisfy $\rho(T'(x^*)) = 1$.

At a fixed point x^* of $T(x)$, if all the eigenvalues of $T'(x^*)$ have magnitude greater than one, then it is not possible for a sequence $\{x^k\}$ generated by iteration (3) to converge to x^*. Such a fixed point x^* is termed a point of *definite repulsion* by Ostrowski (see 4.2 in [7]).

On the other hand, there exist mappings $T(x)$ so that iteration (3), when started from certain initial points, generates iterates that converge to a point of repulsion (see 22.8 in [7], for example). Hence, convergence to a point of repulsion (but not definite repulsion) remains a possibility. However, we feel safe to say that convergence to a point of repulsion appears to be highly unlikely in practice. In Section 6, we will present some numerical experiments to demonstrate this point. A rigorous, quantitative study on this subject seems worthwhile.

3 Necessary and Sufficient Conditions for Strong Attraction

We now present necessary and sufficient conditions for a stationary point of $f(x)$ to be a point of strong attraction of iteration (2). The result itself is a rather straightforward consequence of the condition $\rho(T'(x^*)) < 1$ in the context of $T(x)$ being defined by (4). However, it does provide an interesting new interpretation to the classic result and, more

importantly, leads to a few useful observations that have not been fully exploited in the literature.

Proposition 2 *Let x^* be a stationary point of $f(x)$ and $T(x)$ be defined by (4). Assume that*

(i) $B(x)$ and $\alpha(x)$ are continuous at x^*,

(ii) $B(x^*)$ is symmetric positive definite, and $\alpha(x^*) > 0$.

Then $\rho(T'(x^*)) \leq 1$ if and only if

$$0 \preceq \nabla^2 f(x^*) \preceq \frac{2B(x^*)}{\alpha(x^*)}. \tag{6}$$

Moreover, $\rho(T'(x^*)) < 1$ (i.e., x^* is a point of strong attraction) if and only if strict inequalities hold in (6).

Proof. We note that $T'(x^*)$ is similar to the symmetric matrix

$$M = I - \alpha(x^*)B(x^*)^{-1/2}\nabla^2 f(x^*)B(x^*)^{-1/2},$$

and $\rho(T'(x^*)) \leq 1$ is equivalent to $-1 \leq \lambda_i(M) \leq 1$, i.e.,

$$-I \preceq M \preceq I.$$

The inequality $M \preceq I$ is equivalent to

$$\alpha(x^*)B(x^*)^{-1/2}\nabla^2 f(x^*)B(x^*)^{-1/2} \succeq 0,$$

which is in turn equivalent to the left inequality of (6). On the other hand, the inequality $-I \preceq M$ is equivalent to

$$2I - \alpha(x^*)B(x^*)^{-1/2}\ \nabla^2 f(x^*)\ B(x^*)^{-1/2} \succeq 0,$$

or

$$B(x^*)^{-1/2}[2B(x^*) - \alpha(x^*)\nabla^2 f(x^*)]B(x^*)^{-1/2} \succeq 0,$$

which is in turn equivalent to the right inequality of (6). This proves (6). The proof of the second assertion is entirely parallel. ∎

The left inequality in (6) immediately implies the following fact.

Corollary 1 *Under the assumptions of Proposition 2, any stationary point of $f(x)$ where the Hessian matrix has a negative eigenvalue is a point of repulsion of the iteration (3).*

We recall that a stationary point of $f(x)$ is called a nondegenerate saddle point if the Hessian matrix at this point has both positive and negative eigenvalues. We also recall that a necessary condition for a stationary point of $f(x)$ to be a maximizer is that the Hessian matrix at this point be negative semidefinite. In view of Corollary 1, we have the following observation.

Remark 1 *In the iteration (3), if one keeps B^k positive definite, then under mild conditions all nondegenerate saddle points of $f(x)$ and all maximizers of $f(x)$ where the Hessian is not the zero matrix are points of repulsion; hence, none of these points can be a point of strong attraction of the iteration (3).*

The above fact about nondegenerate saddle points has been observed by Wolfe in [8], who conjectured that "steepest ascent will *almost* never converge to a stationary point at which the Hessian of f is nonsingular and not negative definite". Wolfe's conjecture is close to saying that convergence to a point of repulsion is highly improbable.

We call a stationary point x^* a *strong minimizer* of $f(x)$ if $\nabla^2 f(x^*) \succ 0$. Thus, the left inequality in (6) always holds at any strong minimizer. By considering the right inequality in (6) for some particular choices of $B(x^*)$, we immediately obtain the following known facts.

Remark 2 *Assume that $\alpha(x)$ is continuous at the points of interest.*
(1) For $B(x) = \nabla^2 f(x)$, any strong minimizer of $f(x)$ is a point of strong attraction if and only if $\alpha(x^*) \in (0, 2)$. In particular, any strong minimizer is a point of attraction of Newton's method where $\alpha(x) \equiv 1$.
(2) For $B(x) = I$ (gradient methods), any strong minimizer of $f(x)$ is a point of strong attraction if and only if $\alpha(x^*) < 2/\lambda_{\max}(\nabla^2 f(x^*))$.

As is mentioned at the end of Section 2, the continuity assumption on $\alpha(x)$ at x^* may not always be satisfied by some popular methods such as the steepest descent method with exact line search. Finally, the following observation will be useful later.

Remark 3 *Any minimizer x^* is a point of repulsion if the Hessian matrix at x^* is not majorized by $2B(x^*)/\alpha(x^*)$.*

4 Nonlinear Least squares Problem

For the nonlinear least squares problem, we have

$$f(x) = \frac{1}{2} R^T(x) R(x). \tag{7}$$

where $R : \Re^n \to \Re^m$, $m > n$, is twice continuously differentiable. The gradient and Hessian of $f(x)$ are, respectively,

$$\nabla f(x) = J(x)^T R(x) \quad \text{and} \quad \nabla^2 f(x) = J(x)^T J(x) + S(x), \tag{8}$$

where $J(x)$ is the Jacobian of $R(x)$ and

$$S(x) = \sum_{i=1}^{m} r_i(x) \nabla^2 r_i(x). \tag{9}$$

Consider the iteration (3) with $\alpha(x) = 1$ and

$$B(x) = J(x)^T J(x) + P(x).$$

In this case,

$$T(x) = x - (J(x)^T J(x) + P(x))^{-1} J(x)^T R(x), \tag{10}$$

and at any stationary point x^* of $f(x)$

$$T'(x^*) = (J(x^*)^T J(x^*) + P(x^*))^{-1} (P(x^*) - S(x^*)), \tag{11}$$

assuming continuity of $P(x)$ and nonsingularity of $J(x)^T J(x) + P(x)$ at x^*. Several well-known choices of $P(x)$ are the following:

1. Newton's method: $P(x) = S(x)$;
2. the Gauss-Newton method: $P(x) = 0$;
3. the Levenberg-Marquardt method: $P(x) = \mu(x) I$.

The Gauss-Newton method and the Levenberg-Marquardt method are popular choices for nonlinear least squares problems because they do not require second-order derivatives.

Now consider the iteration

$$x^{k+1} = T(x^k), \tag{12}$$

where $T(x)$ is defined in (10). The structure of least squares problem allows a simplification of Proposition 2.

Proposition 3 *Let x^* be a stationary point of $f(x) = \frac{1}{2} R(x)^T R(x)$ where $f(x)$ is twice differentiable. Assume $J(x)^T J(x) + P(x)$ is continuous and symmetric positive definite at x^*. Then $\rho(T'(x^*)) \leq 1$ if and only if*

$$-J(x^*)^T J(x^*) \preceq S(x^*) \preceq J(x^*)^T J(x^*) + 2P(x^*). \tag{13}$$

Moreover, $\rho(T'(x^*)) < 1$ if and only if strict inequalities hold in (13).

The right inequality in (13) says that the more "positive" $P(x^*)$ is, the more points of attraction the iteration may have. In view of this, we compare the Gauss-Newton method and the Levenberg-Marquardt method.

Proposition 4 *Let x^* be a stationary point of $f(x) = \frac{1}{2}R(x)^T R(x)$ where $f(x)$ is twice differentiable and $J(x)$ has full column rank.*

1. *If x^* is a point of repulsion of the Levenberg-Marquardt method, it is also a point of repulsion of the Gauss-Newton method.*

2. *If x^* is a point of strong attraction of the Gauss-Newton method, it is also a point of strong attraction of the Levenberg-Marquardt method.*

The converses are not necessarily true whenever $\mu(x^*) > 0$ in the Levenberg-Marquardt method.

Analogous to Corollary 1, we also have the following.

Corollary 2 *Any stationary points x^* of $f(x) = \frac{1}{2}R(x)^T R(x)$ where the Hessian matrix has a negative eigenvalue, including all nondegenerate saddle points and maximizers where the Hessian is not the zero matrix, are points of repulsion of the Gauss-Newton method whenever $J(x^*)$ has full column rank. The same statement holds for the Levenberg-Marquardt method if either $J(x^*)$ has full column rank or $\mu(x^*) > 0$.*

It is known that iterates are generally repelled from saddle points in the Gauss-Newton method (see Björck [1], for example). It appears to us that the same property for the Levenberg-Marquardt method is less known.

5 Selective Minimization

Proposition 2 implies that a strong minimizer can be a point of strong attraction of the iteration (3) only if the corresponding Hessian matrix is majorized above by the matrix $2B(x^*)/\alpha(x^*)$.

In most applications, one would ideally like to find a global minimizer. Short of that, one would prefer local minimizers with low objective values. The fact that a given iterative method may turn certain minimizers into points of repulsion could be a useful tool for constructing algorithms whose iterates are attracted to desirable minimizers, but repelled from some undesirable minimizers.

To demonstrate this, we consider applying the Gauss-Newton and Levenberg-Marquardt methods to minimization of nonlinear, nonconvex least squares problems where the global

minimum value of the objective functions is zero or very small. For this type of problems, under mild conditions the global minimizers are points of strong attraction while local minimizers of high objective values are less likely to be points of strong attraction, as is illustrated by the following two lemmas.

Lemma 1 *Let x^* be a strong minimizer of $f(x) = \frac{1}{2}R(x)^T R(x)$ where $f(x)$ is twice differentiable, $J(x)^T J(x) + P(x)$ is continuous and symmetric positive definite at x^*. Then x^* is a point of strong attraction of the iteration (12) if either $\nabla^2 r_i(x)$, $i = 1, 2, \cdots, m$, are not all zero and*

$$\|R(x^*)\| < \frac{\lambda_{\min}[J(x^*)^T J(x^*) + 2P(x^*)]}{\sum_{i=1}^m \|\nabla^2 r_i(x^*)\|}, \tag{14}$$

or $\|R(x^*)\| > 0$ and

$$\sum_{i=1}^m \|\nabla^2 r_i(x^*)\| < \frac{\lambda_{\min}[J(x^*)^T J(x^*) + 2P(x^*)]}{\|R(x^*)\|}. \tag{15}$$

Proof. It suffices to show that the strict inequalities hold in (13). Note that the left strict inequality in (13) , i.e., $-J(x^*)^T J(x^*) \prec S(x)$, holds at any strong minimizer. Since

$$|\lambda_{\max}(S(x^*))| \le \|S(x^*)\| \le \|R(x^*)\| \left(\sum_{i=1}^m \|\nabla^2 r_i(x^*)\| \right),$$

the right strict inequality in (13), i.e., $S(x) \prec J(x^*)^T J(x^*) + 2P(x^*)$, holds if

$$\|R(x^*)\| \left(\sum_{i=1}^m \|\nabla^2 r_i(x^*)\| \right) < \lambda_{\min}[J(x^*)^T J(x^*) + 2P(x^*)],$$

which, under the respective conditions, leads to (14) and (15). ∎

It is well-known that a strong minimizer x^* is a point of strong attraction of the Gauss-Newton method (or the Levenberg-Marquardt method) if either the residuals $r_i(x^*)$ or the Hessian matrices $\nabla^2 r_i(x^*)$, $i = 1, 2, \cdots, m$, are sufficiently small (see Dennis and Steihaug [3], for example). The above lemma is an extension to a slightly more general setting.

Now let us define

$$\theta_i = r_i(x^*)/\|R(x^*)\|_1, \quad i = 1, 2, \cdots, m,$$

and

$$C^* = \sum_{i=1}^m \theta_i \nabla^2 r_i(x^*). \tag{16}$$

Clearly, C^* is a linear combination of the Hessian matrices $\nabla^2 r_i(x^*)$, $i = 1, 2, \cdots, m$, where the coefficients θ_i satisfy $|\theta_i| \in [0, 1]$ and $\sum_{i=1}^m |\theta_i| = 1$. To prove repulsion, an assumption on C^* is needed.

Lemma 2 *Let x^* be a minimizer of $f(x) = \frac{1}{2}R(x)^T R(x)$ where $f(x)$ is twice differentiable and $J(x)^T J(x) + P(x)$ is continuous and symmetric positive definite. Assume further that $\lambda_{\max}(C^*) > 0$ where C^* is defined in (16). Then x^* is a point of repulsion of the iteration (12) if*

$$\|R(x^*)\|_1 > \frac{\lambda_{\max}(J(x^*)^T J(x^*) + 2P(x^*))}{\lambda_{\max}(C^*)}, \tag{17}$$

Proof. We first note that $S(x^*) = \|R(x^*)\|_1 C^*$. A sufficient condition for x^* to be a point of repulsion of the iteration (12) is that

$$\lambda_{\max}(S(x^*)) = \|R(x^*)\|_1 \lambda_{\max}(C^*) > \lambda_{\max}(J(x^*)^T J(x^*) + 2P(x^*)),$$

which violates the right inequality in (13). Clearly, the above inequality is equivalent to (17) whenever $\lambda_{\max}(C^*) > 0$. ∎

Remark 4 *Lemma 1 provides a guarantees that any strong minimizer with sufficiently small residual value is a point of strong attraction of the iteration (12). On the other hand, Lemma 2 shows that minimizers with sufficiently large residual values will become a point of repulsion of the iteration (12) under some circumstances.*

We have done some numerical experiments on applying the Gauss-Newton and the Levenberg-Marquardt methods to global minimization of least squares problems where the optimal residual value is either zero or very small. Our numerical results have shown that the algorithms do skip some local minimizers, and have greater chances of converging to a global minimizer than, say, Newton's method which is attracted to any stationary point under mild conditions.

For more general problems, it is also possible to construct minimization algorithms that skip minimizers of high objective values while targeting lower-valued minimizers. For example, the following is a simple scheme:

$$B(x) = \begin{cases} I, & f(x) \geq \xi, \\ \nabla^2 f(x) + D(x), & \text{otherwise}, \end{cases}$$

where $D(x)$ is a diagonal matrix chosen to ensure $B(x) \succ 0$, and

$$\alpha(x) = \begin{cases} 2/\eta, & f(x) \geq \xi, \\ 1, & \text{otherwise}, \end{cases}$$

where $\eta > 0$. With these choices, the iteration

$$x^{k+1} = x^k - \alpha(x^k) B(x^k)^{-1} \nabla f(x^k)$$

will have the properties:

1. Any minimizer x^* with $f(x^*) \geq \xi$ and $\lambda_{\max}(\nabla^2 f(x^*)) > \eta$ is a point of repulsion.

2. Any strong minimizer x^* with $f(x^*) < \xi$ is a point of strong attraction.

Although we do not claim that the above construction is of any practical value, we do hope that combined with some random sampling techniques such as simulated annealing [4], the selective minimization property may lead to improved global optimization algorithms. This topic merits further study, but is outside the scope of this short paper. Instead, in the next section, we present a simple example showing the phenomenon of selective minimization.

6 Numerical Examples

In this section, we provide a couple of simple examples to illustrate the following points: (i) if $\{B^k\}$ is uniformly positive definite and $\{\alpha^k\}$ uniformly positive, then convergence to a point of repulsion seems to be highly unlikely in general; (ii) selective minimization does occur for certain problems. All our numerical experiments were done using Matlab.

6.1 First Example: repulsion

We consider the following function $f : \Re^n \to \Re$ (Levy and Gómez [5]):

$$f(x) = sin^2\left(\frac{\pi}{4}(x_1 + 3)\right) + \sum_{i=1}^{n-1} \frac{(x_i - 1)^2}{16}\left[1 + sin^2\left(\frac{\pi}{4}(x_{i+1} + 3)\right)\right] + \frac{(x_n - 1)^2}{16}, \quad (18)$$

This function has many local minima but a unique global minimum at $x_i^* = 1, i = 1, 2, \ldots, n$, where $f(x^*) = 0$.

We use the gradient method to construct an iteration

$$x^{k+1} = T(x^k) \equiv x^k - \alpha \nabla f(x^k) \quad (19)$$

and always choose

$$\alpha > \frac{2}{\lambda_{\max}(\nabla^2 f(x^*))}$$

so that at least one of the eigenvalues of $T'(x^*) \equiv I - \alpha \nabla^2 f(x^*)$ has an absolute value greater than one. By this very construction, the global minimizer x^* is a point of repulsion of the iteration (19) since $|\lambda_{\max}(T'(x^*))| > 1$.

We applied iteration (19) to problem (18) for $n = 2, 3, 10, 50, 100$. The actual values of the steplength α vary with n and are not of interest here. For each n value, we selected 100 random starting points close to x^*, namely

$$x^1 = x^* + \epsilon(\texttt{rand}(n, 1) - 0.5),$$

where $\epsilon = 10^{-3}$ and `rand` is the Matlab command for generating a uniformly distributed random n-vector with components in [0,1]. The stopping criterion used in our experiments is that either $\|\nabla f(x^k)\| < 10^{-14}$ or the number of iterations reaches 100. In all of these numerical experiments, we did not observe a single case of convergence to the global minimizer x^*, which is a point of repulsion by our construction, although the starting points are all very close to x^*. These experiments give a rather strong indication that convergence to a point of repulsion may be improbable in general.

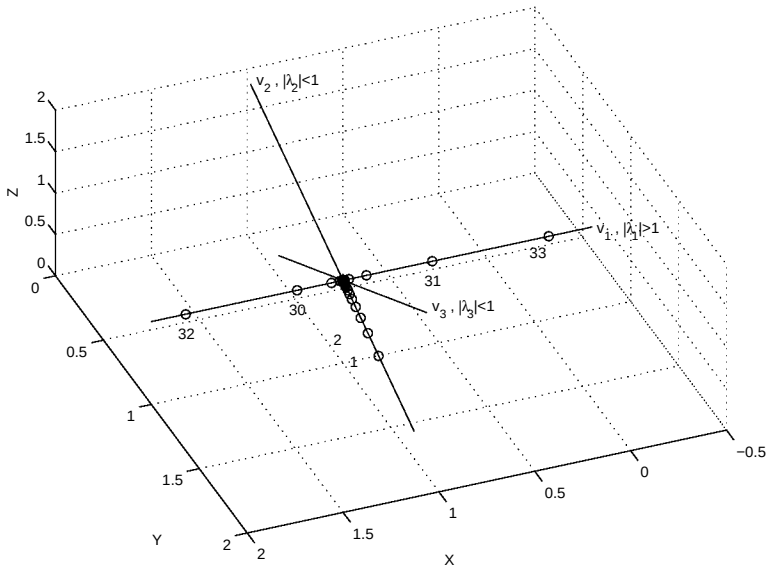

Figure 1: Non-convergence to a point of repulsion

In Figure 1, we present a specific example for $n = 3$ and

$$\alpha = \frac{3}{\lambda_{\max}(\nabla^2 f(x^*))} = 4.3061.$$

For this choice of α, the three eigenvalues of $T'(x^*)$ are

$$\lambda_1 = -2, \quad \lambda_2 = \lambda_3 = 0.7182.$$

In order to dramatize the situation, we choose a starting point $x^1 = (1, 1.3709, 0.6647)$ so that $x^1 - x^*$ is in the direction of v_2 — the eigenvector direction corresponding to the eigenvalue $\lambda_2 = 0.7182$. In the picture, the small circles represent the positions of the iterates, and the numbers beside the circles are the iteration numbers. As one can see, initially the iterates approach x^* along the direction of v_2. However, as the iterates get closer to x^* (with $\|\nabla f(x)\| \approx 10^{-4}$), unable to stay in the direction of v_2 they start to drift

away from x^* along the direction of v_1, which is the eigenvector direction corresponding to the eigenvalue $\lambda_1 = -2$.

6.2 Second Example: Selective Minimization

We now consider the following two-dimensional least squares problem

$$f(x,y) = \frac{1}{2} R(x,y)^T R(x,y), \tag{20}$$

where, for $\alpha = 1.2$ and $\beta = 6$,

$$R(x,y) = \begin{pmatrix} \alpha\, sin(\pi(1+x/4)) \\ \beta(x/4)[1+\alpha^2 sin^2(\pi(1+y/4))]^{1/2} \\ y/4 \\ \alpha\, sin(\pi(1+y/4)) \\ \beta(y/4)[1+\alpha^2 sin^2(\pi(1+x/4))]^{1/2} \\ x/4 \end{pmatrix}. \tag{21}$$

This function $f(x,y)$ is symmetric about both the x-axis and the y-axis, and has a unique global minimizer at the origin with zero-residual. We will concentrate our attention to the square: $-5 \le x, y \le 5$, which will be considered to be the area of our interest. In this square, the function has four local minimizers at

$$(x^*, y^*) \approx (\pm 3.64, \pm 3.64)$$

with relatively high residual value $f(x^*, y^*) \approx 34.09$. The function also has four saddle points in the square of interest at

$$(x^*, y^*) \approx (\pm 2.98, \pm 2.98)$$

with residual value $f(x^*, y^*) \approx 36.12$. See Figure 2 for a plot of $f(x,y)$ in the square of interest.

We apply the Gauss-newton method, i.e., the iteration (12) with $P(x) = 0$ in (10), to minimizing $f(x,y)$ defined in (20) and (21). In our experiments, we have found that the Gauss-newton method is always well defined in the square of interest.

From Lemma 1, we know that the global minimizer at the origin is a point of strong attraction for the Gauss-Newton iteration. On the other hand, our calculation shows that for the Gauss-Newton iteration, $\lambda_i(T'(x))$, $i = 1, 2$, are, respectively and approximately -16.86 and -2.29 at the four local minimizers. Therefore, they are points of definite repulsion.

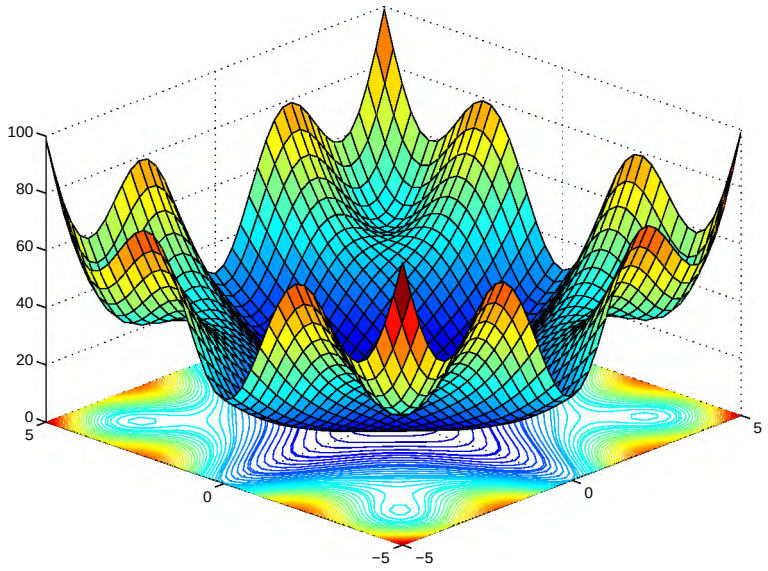

Figure 2: The 2-dimensional Test Function

The saddle points are nondegenerate and hence points of repulsion of the Gauss-Newton iteration. In fact,

$$\min_{i=1,\cdots,n} |\lambda_i(T'(x^*))| < 1 \tag{22}$$

holds at the saddle points which means that they are not points of definite repulsion.

For the purpose of comparison, we also apply the Levenberg-Marquardt method and the Newton method to the problem as well. For the Levenberg-Marquardt method, we choose $P(x) = 10I$ in (10). With this choice, all minimizers in the square, global or local, are points of strong attraction, and the saddle points remain points of repulsion where (22) holds. On the other hand, all the stationary points in the square are points of strong attraction of the Newton method.

We run the three methods starting from the following grid of initial points in the first quadrant:

$$(x_i, y_j) = (i, j)/4, \;\; 0 \le i, j \le 20.$$

Since the function is symmetric about both axes, we can duplicate the behavior of the methods in the first quadrant in the other three quadrants. For each method and each initial point, we record whether or not the iterates converge to the global minimizer at the origin, or to one of the other stationary points (some may be outside of the square of interest), or do not converge within a prescribed maximum number of iterations, which is set to 100 in our experiments. The convergence criterion is that the norm of the gradient

be less that 10^{-8}.

We summarize the numerical results for the three methods below.

1. **Gauss-Newton:** From all the starting points without exception, the Gauss-Newton method converged to the global minimizer at the origin. We note that never did any starting point lead to a point of repulsion (saddle point) no matter how close it was.

2. **Levenberg-Marquardt:** With $P(x) = 10I$ for the Levenberg-Marquardt method, all the starting points led to one of the five minimizers in the square, with around 75% to the global minimizer and the rest 25% to the local ones. Again, never did a starting points lead to a saddle point.

3. **Newton:** For the Newton method, about 50% of the starting points led to the global minimizer, and about 30% to other stationary points in the square. The rest of points either led to stationary points outside the square, or were such that the method did not terminate after 100 iterations.

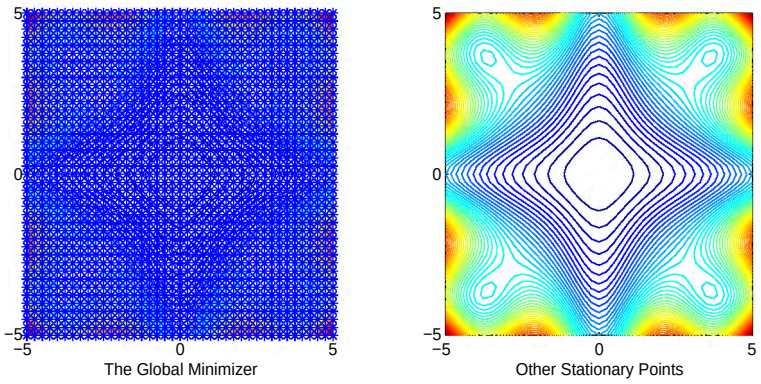

Figure 3: Estimated Regions of Attraction for the Gauss-Newton Method

In Figures 3-5, we plot the (estimated) region of attraction of the global minimizer and the combined region of attraction of all the other stationary points in the square for the three methods, respectively. The asterisks represent points from which a method converged to the global minimizer (in the pictures on the left) or to one of the other stationary points inside the square (in the pictures on the right). On the background, we also plot the contour of the test function.

In the picture on the right side of Figure 5, it appears that at each corner an area of attraction of the local minimizer is separated by a narrow band from an area of attraction of the nearby saddle point.

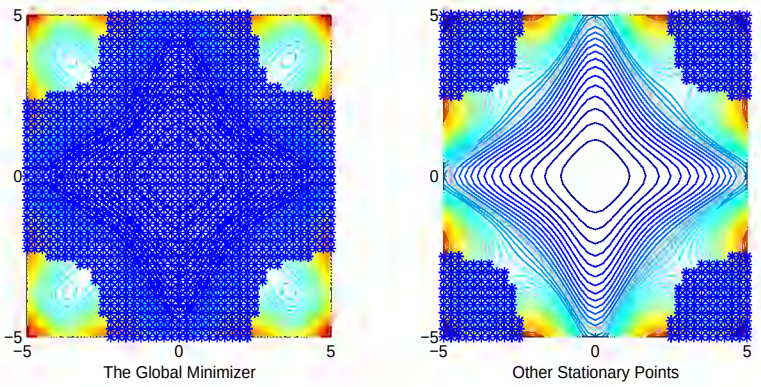

Figure 4: Estimated Regions of Attraction for the Levenberg-Marquardt Method

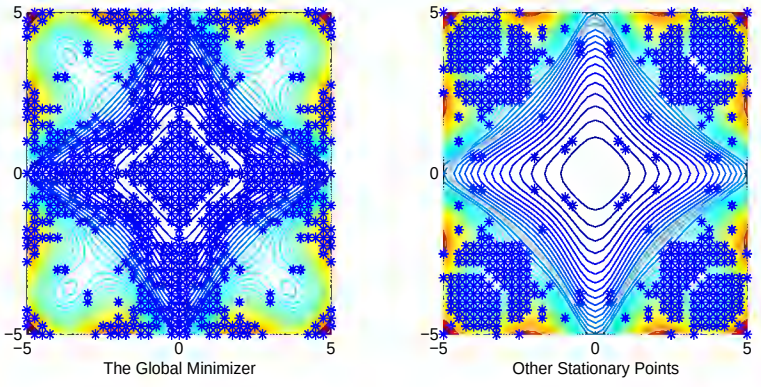

Figure 5: Estimated Regions of Attraction for the Newton Method

7 Final Remarks

Concerning the general iteration process

$$x^{k+1} = x^k - \alpha^k (B^k)^{-1} \nabla f(x^k),$$

Proposition 2 provides a new interpretation to the classic Ostrowski theorem and leads to some interesting observations. We consider the following two to be particularly worthwhile.

Firstly, as long as one keeps $\{B^k\}$ uniformly positive definite and $\{\alpha^k\}$ bounded away from zero, then the undesirable case of converging to a nondegenerate saddle point should not be of general concern. This statement is based on the premise that convergence to a point of repulsion is unlikely, which seems to be well supported by empirical evidence. Further quantification in this direction is certainly desirable.

Secondly, if one does not always enforce monotone descent, then under favorable conditions a method can actually generate iterates that skip undesirable local minimizers while still being attracted to desirable, global minimizers. It is not surprising that a minimization algorithm can fail to converge to a minimum if monotone decrease in the function value is not enforced. However, it is useful to know that under proper conditions, certain minimization algorithms will only escape from local minima, but never from a global minimum.

References

[1] A. Björck. Numerical Methods for Least Squares Problems. SIAM, Philadelphia, 1996.

[2] J. E. Dennis Jr. and R.B. Schnabel. Numerical Methods for Unconstrained Optimization and Nonlinear Equations. Prentice-Hall, Englewood Cliffs, NJ, 1983 and SIAM, Philadelphia, 1996.

[3] J. E. Dennis Jr. and T. Steihaug. On the successive projection approach to least squares problems. SIAM Journal on Numerical Analysis, 23:717-733, 1986.

[4] S. Kirkpatrick, C.D. Gelatt and M.P. Vecchi. Optimization by Simulated Annealing. Science, May 1983.

[5] A. V. Levy and S. Gómez. The Tunneling Method Applied to Global Optimization. Numerical Optimization, P.T. Boggs, R.H. Byrd and R.B. Schnabel, Editors, SIAM, pp. 213-244, 1985.

[6] J. M. Ortega and W. C. Rheinboldt. Iterative Solution of Nonlinear Equations in Several Variables. Academic Press, New York, NY, 1970.

[7] A. M. Ostrowski. Solution of Equations and Systems of Equations. Academic Press, New York, NY, 1966 (2nd Ed.).

[8] Ph. Wolfe. Convergence Conditions for Ascent Methods. II: Some Corrections. SIAM Review, vol.13, pp, 185-188, 1971.