Maximization algorithms with imperfect line searches

Shelby J. Haberman

Haberman Statistics

Abstract

Maximization algorithms with imperfect line searches are discussed for continuously differentiable real functions on a convex nonempty open subset of the finite-dimensional space R^p of p-dimensional vectors. Convergence conditions are provided under strict pseudoconcavity assumptions.

Keywords: strict pseudoconcavity, Newton-Raphson algorithm, gradient ascent, conjugate gradient algorithm

Introduction

Maximization of a continuously differentiable real function f on an open convex subset O of the space R^p of p-dimensional vectors often involves iterative algorithms with two stages. Let f have a finite supremum $\sup(f)$, and let $\mathbf{x}$ in O satisfy $f(\mathbf{x}) = \sup(f)$. The algorithms under study can be described as one-point algorithms. To begin, let $\mathbf{y}_0$ approximate $\mathbf{x}$. For each nonnegative integer t, the vector $\mathbf{y}_t$ in O is employed to find a new approximation $\mathbf{y}_{t+1}$ in O. The new approximation $\mathbf{y}_{t+1}$ for $\mathbf{x}$ begins with an initial direction equal to the p-dimensional vector $\mathbf{s}_t$. Given $\mathbf{s}_t$ and $\mathbf{y}_t$, a real multiplier $\alpha_t > 0$ is then chosen, and the new approximation

$$\mathbf{y}_{t+1} = \mathbf{y}_t + \alpha_t \mathbf{s}_t. \tag{1}$$

This approach is especially attractive if f has the strict pseudoconcavity (SPC) property described in section Strict Pseudoconcavity. Conditions on selection of $\mathbf{s}_t$ and α_t for $t \geq T$ are then provided in The General Algorithm under which convergence of $\mathbf{y}_t$ is guaranteed. To apply these conditions, section Methods to Select Directions

---

Shelby J. Haberman https://orcid.org/0000-0002-5490-0405

explores common approaches to selection of the $\mathbf{s}_t$, and section {namerefstep considers selection of the α_t. In section Limiting Convergence Rates, convergence rates are considered for the procedures introduced in sections Methods to Select Directions and Methods for Line Searching. The approach here is clearly related to that of Zangwill (1969).

Strict Pseudoconcavity

Consider algorithms for maximization of a continuously differentiable real function f on a convex nonempty open subset O of the space R^p such that, for some real a, the set A_0 of $\mathbf{y}$ in O such that $f(\mathbf{y}) > a$ is nonempty and the set A of $\mathbf{y}$ in O with $f(\mathbf{y}) \geq a$ is bounded and closed. This assumption on A certainly holds if $O = R^p$ and $f(\mathbf{y})$, $\mathbf{y}$ in R^p, approaches $-\infty$ whenever the vector norm $|\mathbf{y}| = (\mathbf{y}'\mathbf{y})^{1/2}$ approaches ∞. Because f is continuous, the assumption on A implies that the maximization problem has at least one solution, for some $\mathbf{x}$ in R^p exists such that $f(\mathbf{x})$ is the supremum $\sup(f)$ of f over R^p.

Use of a strict assumption of pseudoconcavity on A_0 guarantees that $\mathbf{x}$ is the unique solution of the maximization problem. To define strict pseudoconcavity on A_0, let f have gradient function ∇f, so that $\nabla f(\mathbf{x})$ is the p-dimensional vector $\mathbf{0}_p$ with all elements 0. The strict pseudoconcavity assumption (SPC) is that, for distinct $\mathbf{y}$ and $\mathbf{z}$ such that $f(\mathbf{y}) > a$, $f(\mathbf{z}) \geq f(\mathbf{y})$ implies that $(\mathbf{z} - \mathbf{y})'\nabla f(\mathbf{y}) > 0$ (Ponstein, 1967). This assumption is slightly unconventional because A_0 is not explicitly assumed to be convex; however, the requirement on f does imply convexity of A_0. To verify this claim, consider distinct $\mathbf{y}$ and $\mathbf{z}$ in A_0. Without loss of generality, let $f(\mathbf{z}) \geq f(\mathbf{y})$. Suppose that some positive real $u < 1$ exists such that $u\mathbf{y} + (1 - u)\mathbf{z}$ is not in A_0. Then the continuity of f implies that some positive real $v < 1$ exists such that $\mathbf{w} = v\mathbf{y} + (1 - v)\mathbf{z}$ is in A_0 and $f(\mathbf{w}) < f(\mathbf{y})$. Because $\mathbf{y} - \mathbf{w} = (1 - v)(\mathbf{z} - \mathbf{y})$ and $\mathbf{z} - \mathbf{w} = -v(\mathbf{z} - \mathbf{y})$, the SPC assumption implies that $(\mathbf{z} - \mathbf{y})'\nabla f(\mathbf{w})$ must be both positive and negative, an impossibility. Thus A_0 is indeed convex. The SPC assumption holds if f has the strict concavity (SC) property that $f(u\mathbf{y} + (1 - u)\mathbf{z}) > uf(\mathbf{y}) + (1 - u)f(\mathbf{z})$ if $\mathbf{y}$ and $\mathbf{z}$ are distinct members of A_0 and $0 < u < 1$. The assumption that SPC holds then implies that $\nabla f(\mathbf{y}) = \mathbf{0}_p$ for y in A_0 if, and only if, $\mathbf{y} = \mathbf{x}$.

In addition, for $\mathbf{y}$ in A_0 and $\mathbf{s}$ in R^p such that $\mathbf{s} \neq \mathbf{0}_p$, SPC also applies to maximization over the bounded line segment $B(\mathbf{y}, \mathbf{s})$ that consists of $\mathbf{z}$ in A_0 such that $\mathbf{z} - \mathbf{y} = u\mathbf{s}$ for some real number u. The function f has a unique maximum $\mathbf{b}(\mathbf{y}, \mathbf{s})$ on $B(\mathbf{y}, \mathbf{s})$ and $\mathbf{b}(\mathbf{y}, \mathbf{s})$ is the only member $\mathbf{z}$ of $B(\mathbf{y}, \mathbf{s})$ such that $\mathbf{s}'\nabla f(\mathbf{z}) = 0$. Equivalently, let $Q(\mathbf{y}, \mathbf{s})$ be the nonempty open and bounded interval of real u such that $\mathbf{y} + u\mathbf{s}$ is in A_0, and let $h(\mathbf{y}, \mathbf{s})$ be the strictly pseudoconcave real function on $Q(\mathbf{y}, \mathbf{s})$ with value $h(u; \mathbf{y}, \mathbf{s}) = f(\mathbf{y} + u\mathbf{s})$ at u in $Q(\mathbf{y}, \mathbf{s})$. The derivative $h_1(\mathbf{y}, \mathbf{s})$ of $h(\mathbf{y}, \mathbf{s})$ has value $h_1(u; \mathbf{y}, \mathbf{s}) = \mathbf{s}'\nabla f(\mathbf{y} + u\mathbf{s})$ at u in $Q(\mathbf{y}, \mathbf{s})$, so that $h(\mathbf{y}, \mathbf{s})$ achieves its maximum value at the unique $q(\mathbf{y}, \mathbf{s})$ in $Q(\mathbf{y}, \mathbf{s})$ such that $h_1(q(\mathbf{y}, \mathbf{s}); \mathbf{y}, \mathbf{s}) = 0$, and $\mathbf{b}(\mathbf{y}, \mathbf{s}) = \mathbf{y} + q(\mathbf{u}, \mathbf{s})\mathbf{s}$.

The General Algorithm

To specify the $\mathbf{s}_t$ and α_t for nonnegative integers t, two positive real numbers $\gamma_1 > 1$ and $\gamma_2 < 1$ are selected in advance. The constant γ_1 provides a standard for an adequate approximation to an optimal line search, and the constant γ_2 provides a standard for the direction of the step. The direction $\mathbf{s}_t$ satisfies the constraints that $\mathbf{s}_t = \mathbf{0}_p$ if, and only if, $\nabla f(\mathbf{y}_t) = \mathbf{0}_p$ and

$$\mathbf{s}_t'\nabla f(\mathbf{y}_t) \geq \gamma_2 |\mathbf{s}_t||\nabla f(\mathbf{y}_t)|. \tag{2}$$

If $\mathbf{s}_t = \mathbf{0}_p$, then the choice of α_t does not matter. By convention, α_t is then set to 1. If $\mathbf{s}_t \neq \mathbf{0}_p$, then the real number α_t is an approximation to $q(\mathbf{y}_t, \mathbf{s}_t)$ that satisfies the inequality

$$[f(\mathbf{y}_{t+1}) - f(\mathbf{y}_t)] \geq \gamma_1 \alpha_t |\mathbf{s}_t'\nabla f(\mathbf{y}_{t+1})| \tag{3}$$

(Haberman, 1974; Wolfe, 1969, 1971). The conditions on $\mathbf{s}_t$ and α_t imply that $\mathbf{y}_{t+1} = \mathbf{y}_t$ if, and only if, $\mathbf{y}_t = \mathbf{x}$, the inequality $f(\mathbf{y}_{t+1}) \geq f(\mathbf{y}_t)$ holds, and the stronger inequality $f(\mathbf{y}_{t+1}) > f(\mathbf{y}_t)$ holds whenever $\mathbf{y}_t \neq \mathbf{x}$. If $\mathbf{s}_u = \mathbf{0}_p$ for any integer $u \geq 0$, then $\mathbf{y}_t = \mathbf{y}_u = \mathbf{x}$ for all integers $t \geq u$.

The algorithm requirements can always be satisfied. For example, in gradient ascent (Cauchy, 1847), the vector $\mathbf{s}_t = \nabla f(\mathbf{y}_t)$ obviously satisfies Equation 2. Given any acceptable definition of $\mathbf{s}_t$, Equation 3 holds if either $\nabla f(\mathbf{y}_t) = \mathbf{0}_p$ or if $\nabla f(\mathbf{y}_t) \neq \mathbf{0}_p$ and $\alpha_t = q(\mathbf{y}_t, \mathbf{s}_t)$. The case of $\nabla f(\mathbf{y}_t)$ is trivial. If $\nabla f(\mathbf{y}_t) \neq \mathbf{0}_p$, then Equation 2 implies that $\mathbf{s}_t'\nabla f(\mathbf{y}_t) > 0$, so that $f(\mathbf{y}_t + \alpha \mathbf{s}_t)$ is greater than $f(\mathbf{y}_t)$ for some positive α and is less than $f(\mathbf{y}_t)$ for α negative. Thus $\alpha_t = q(\mathbf{y}_t, \mathbf{s}_t) > 0$ satisfies

$$\mathbf{s}_t'\nabla f(\mathbf{y}_t + \alpha_t \mathbf{s}_t) = 0, \tag{4}$$

and

$$f(\mathbf{y}_t + \alpha_t \mathbf{s}_t) > f(\mathbf{y}_t). \tag{5}$$

The following convergence theorem follows.

Theorem 1. *As t increases, $\mathbf{y}_t$ approaches $\mathbf{x}$.*

Proof. If $\nabla f(\mathbf{y}_u) = \mathbf{0}_p$ for some integer $u \geq 0$, then $\mathbf{y}_t = \mathbf{x}$ for all integers $t \geq u$, so that the result follows. It now suffices to consider the case in which $\nabla f(\mathbf{y}_t) \neq \mathbf{0}_p$ for all integers $t \geq 0$. In this case, $f(\mathbf{x}) \geq f(\mathbf{y}_{t+1}) > f(\mathbf{y}_t)$ for all nonnegative integers t, so that, as t approaches ∞, $f(\mathbf{y}_t)$ converges to some real number $g \leq f(\mathbf{x})$ and $f(\mathbf{y}_u) - f(\mathbf{y}_t)$ converges to 0 as the nonnegative integers t and u both approach ∞.

Because A is closed and bounded, a strictly increasing sequence $v(k)$, $k \geq 1$, of nonnegative integers exists such that, for some $\mathbf{z}_0$ and $\mathbf{z}_1$ in A and some $\mathbf{r}$ in R^p such that $|\mathbf{r}| = 1$, as k increases, $\mathbf{y}_{v(k)+u}$ converges to $\mathbf{z}_u$, u equal 0 or 1, and $|\mathbf{s}_{v(k)}|^{-1}\mathbf{s}_{v(k)}$ converges to $\mathbf{r}$. As k increases, $f(\mathbf{y}_{v(k)+1}) - f(\mathbf{y}_{v(k)})$ converges to 0 and $f(\mathbf{y}_{v(k)+u})$ converges to $f(\mathbf{z}_u) > a$ for u equal 0 or 1. Thus $\mathbf{z}_0$ and $\mathbf{z}_1$ are in A_0, and $\nabla f(\mathbf{y}_{v(k)+u})$

converges to $\nabla f(\mathbf{z}_u)$ for u equal 0 or 1. Because $f(\mathbf{y}_t)$, $t \geq 0$, is strictly increasing, $f(\mathbf{z}_0) = f(\mathbf{z}_1)$. Equation 3 in the case of integers $t = v(k)$, $k \geq 1$, and the limiting results as k approaches ∞ then imply that

$$(\mathbf{z}_1 - \mathbf{z}_0)'\nabla f(\mathbf{z}_1) = 0. \tag{6}$$

The SPC requirement implies that Equation 6 cannot hold if $\mathbf{z}_1 \neq \mathbf{z}_0$. Thus $\mathbf{z}_1 = \mathbf{z}_0$.

As k increases, $\mathbf{y}_{v(k)+1} - \mathbf{y}_{v(k)} = \alpha_{v(k)}\mathbf{s}_{v(k)}$ converges to $\mathbf{z}_1 - \mathbf{z}_0 = \mathbf{0}_p$. By Equation 2, $\mathbf{r}'\nabla f(\mathbf{z}_0) \geq \gamma_2|\nabla f(\mathbf{z}_0)|$. Division of both sides of Equation 3 by $\alpha_t|\mathbf{s}_t|$ for $t = v(k)$, $k \geq 1$, and use of the continuous differentiability of f then implies that $\mathbf{r}'\nabla f(\mathbf{z}_0) \geq \gamma_1\mathbf{r}'\nabla f(\mathbf{z}_0)$, an impossible result unless $\mathbf{r}'\nabla f(\mathbf{z}_0) = 0$ and $\nabla f(\mathbf{z}_0) = \mathbf{0}_p$. Thus $\mathbf{z}_0- = \mathbf{x}$. Because $v(k)$, $k \geq 1$, is arbitrary, $\mathbf{y}_t$ converges to $\mathbf{x}$ as t increases. $\square$

Methods to Select Directions

A large number of approaches to selection of directions are in common use. As already noted, gradient ascent is certainly available. Some alternatives include the Newton-Raphson algorithm, the approximate Newton-Raphson algorithm, the conjugate gradient algorithm, and the secant algorithm.

The Gradient Ascent Algorithm

As already noted, in gradient ascent, $T = 1$ and $\mathbf{s}_t = \nabla f(\mathbf{y}_t)$. In the other cases in this section, gradient ascent provides a backup method.

The Newton-Raphson Algorithm

In the case of the Newton-Raphson algorithm, $T = 1$ and f must be twice continuously differential. Let $\nabla^2 f$ be the p by p Hessian matrix of f, so that, for $\mathbf{y}$ in O and a p-dimensional vector $\mathbf{c}$ such that $\mathbf{y} + \mathbf{c}$ is in O and $\mathbf{c} \neq \mathbf{0}_p$,

$$|\mathbf{c}|^{-1}|\nabla f(\mathbf{y} + \mathbf{c}) - \nabla f(\mathbf{y}) - [\nabla^2 f(\mathbf{y})]\mathbf{c}| \tag{7}$$

converges to 0 as $|\mathbf{c}|$ converges to 0. If $-\nabla^2 f(\mathbf{y}_t)$ is positive-definite and if the direction

$$\mathbf{s}_{t0} = [-\nabla^2 f(\mathbf{y}_t)]^{-1}\nabla f(\mathbf{y}_t) \tag{8}$$

satisfies

$$\mathbf{s}_{t0}'\nabla f(\mathbf{y}_t) \geq \gamma_2|\mathbf{s}_{t0}||\nabla f(\mathbf{y}_t)|, \tag{9}$$

then $\mathbf{s}_t = \mathbf{s}_{t0}$. Otherwise, as in gradient ascent, $\mathbf{s}_t = \nabla f(\mathbf{y}_t)$. In traditional development of the Newton-Raphson algorithm, $\mathbf{s}_t$ is always $\mathbf{s}_{t0}$, an obvious problem if $-\nabla^2 f(\mathbf{y}_t)$ is singular. For $p = 1$, use of Equation 8 has a very lengthy history (Ypma, 1995). For $p > 1$, a historical discussion is provided by Yamamoto (2001).

Approximate Newton-Raphson Algorithms

A number of approximate Newton-Raphson algorithms exist. In these cases, $T = 1$ and f is only required to be continuously differentiable. Instead of $-\nabla^2 f(\mathbf{y}_t)$, a matrix $\mathbf{V}(\mathbf{y}_t)$ is employed, where $\mathbf{V}$ is a function from O to the space of p by p matrices. These applications generally involves a matrix function $\mathbf{V}$ that is easier to compute than is $-\nabla^2 f$. A number of such cases are encountered in statistics in estimation problems. Examples include the scoring algorithm (Fisher, 1925) and the Louis (Haberman, 2013; Louis, 1982) approximation to the scoring algorithm. Cases can also arise in which $\mathbf{V}$ is an approximation of $-\nabla^2 f$ obtained by numerical differentiation. For example, let $\boldsymbol{\delta}_j$, $1 \le j \le p$, be the p-dimensional vector with element j equal to 1 and all other elements equal to 0. Let d be a very small real number, and let c be a positive real function on O such that, for $\mathbf{y}$ in O, $c(\mathbf{y}) \le d$ and $\mathbf{y} + c(\mathbf{y})\mathbf{z}$ is in O if $\mathbf{z}$ is in R^p and $|\mathbf{z}| < c(\mathbf{y})$. Let

$$\mathbf{V}(\mathbf{y}) = [c(\mathbf{y})]^{-1} \sum_{j=1}^{p} [\nabla f(\mathbf{y} + c(\mathbf{y})\boldsymbol{\delta}_j) - \nabla f(\mathbf{y})] \otimes \boldsymbol{\delta}_j. \tag{10}$$

If $\mathbf{V}(\mathbf{y}_t)$ is positive-definite and nonsingular and if the direction

$$\mathbf{s}_{t0} = [\mathbf{V}(\mathbf{y}_t)]^{-1} \nabla f(\mathbf{y}_t) \tag{11}$$

satisfies Equation 9, then $\mathbf{s}_t = \mathbf{s}_{t0}$. Otherwise, as in gradient ascent, $\mathbf{s}_t = \nabla f(\mathbf{y}_t)$.

The Conjugate Gradient Algorithm

The conjugate gradient algorithm is a variation of gradient ascent (Cohen, 1972) in which $T = 1$. As in gradient ascent, $\mathbf{s}_0 = \nabla f(\mathbf{y}_0)$. For $t > 1$, let $\mathbf{s}_t = \mathbf{0}_p$ if $\nabla f(\mathbf{y}_t) = \mathbf{0}_p$. Otherwise, as in Polak and Ribière (1969), let

$$\tau_t = \frac{[\nabla f(\mathbf{y}_t) - \nabla f(\mathbf{y}_{t-1})]' \nabla f(\mathbf{y}_t)}{|\nabla f(\mathbf{y}_{t-1})|^2}, \tag{12}$$

and let

$$\mathbf{s}_{t0} = \nabla f(\mathbf{y}_t) + \tau_t \mathbf{s}_{t-1}. \tag{13}$$

If Equation 9 holds, then $\mathbf{s}_t = \mathbf{s}_{t0}$. Otherwise, $\mathbf{s}_t = \nabla f(\mathbf{y}_t)$. The use of $\mathbf{s}_t = \mathbf{s}_{t0}$ in the algorithm used here is more restricted than in an algorithm cited in Shewchuk (1994) that uses $\mathbf{s}_t = \mathbf{s}_{t0}$ whenever γ_t is nonnegative. An alternative selection of τ_t is

$$\tau_t = \frac{|\nabla f(\mathbf{y}_t)|^2}{|\nabla f(\mathbf{y}_{t-1})|^2}. \tag{14}$$

(Fletcher & Reeves, 1964).

Methods for Line Searching

Except in special cases, line searches involve iterative algorithms for maximization of a real function on a real interval. One approach involves the two-point quadratic approach (Haberman, 2013) studied in this section. In this section, use of this approach is explained, and its convergence properties are justified. A positive real constant $\eta < 1$ and a real constant $\kappa > 0$ are fixed in advance for all line searches. If $\mathbf{s}_t \neq \mathbf{0}_p$ for some integer $t \geq T$, then let $g_t = h(\mathbf{y}_t, \mathbf{s}_t)$ have derivative g_{t1} on $Q_t = Q(\mathbf{y}_t, \mathbf{s}_t)$. A sequence of real approximations α_{tv}, $v \geq 0$, is defined such that α_{t0} is 0 and α_{tv} is in Q_t for all nonnegative integers v. The sequence is generated so that $g_t(\alpha_{tv})$, $v \geq 0$, is nondecreasing and $g_t(\alpha_{tv}) = g_t(\alpha_{(t(v+1)})$ for a nonnegative integer v if, and only if, $\alpha_{tv} = q_t = q(\mathbf{y}_t, \mathbf{s}_t)$. Associated with α_{tv} are nonnegative bounds b_{Ltv} and $b_{Utv} = \infty$. It is always the case that $b_{Ltv} \leq \alpha_{tv} \leq b_{Utv}$. Here b_{Utv} can be finite or infinite, but b_{Ltv} is finite. Initially $b_{Lt0} = 0$ and $b_{Ut0} = \infty$. The sequence b_{Ltv}, $v \geq 0$, is nondecreasing, while b_{Utv}, $v \geq 0$, is nonincreasing. For any integer $v \geq 0$, if

$$g_t(\alpha_{tv}) - g_t(0) \geq \gamma_1 \alpha_{tv} |g_1(\alpha_{tv}; \mathbf{y}_t)|, \tag{15}$$

then the line search terminates with $\alpha_t = \alpha_{tv}$.

If Equation 15 does not hold, then $g_1(\alpha_{tv}) \neq 0$. In this case, to select $\alpha_{t(v+1)}$ requires several steps. Let β_{t00} be the minimum of 1 and $\kappa/|\mathbf{s}_t|$. If v is a positive integer, let $\beta_{tv0} = \alpha_{t(v-1)}$, let $b_{Ltv0} = b_{Ltv}$, and $b_{Utv0} = b_{Utv}$. If v is a nonnegative integer and w is a positive integer, let

$$\beta_{tvw*} = \alpha_{tv} + g_1(\alpha_{tv})/\max(\Gamma_{tv}, \Delta_{tvw}), \tag{16}$$

where

$$\Delta_{tvw} = \frac{2g_1(\alpha_{tv})}{\beta_{tv(w-1)} - \alpha_{tv})} - \frac{2[g(\beta_{tv(w-1)} - g(\alpha_{tv})]}{(\beta_{tv(w-1)} - \alpha_{tv})^2}. \tag{17}$$

and

$$\Gamma_{tv} = |g_1(\alpha_{tv})||\mathbf{s}_t|/\kappa. \tag{18}$$

If $\beta_{tvw*} > \alpha_{tv}$, let β_{tvw} be the minimum of β_{tvw*} and $\alpha_{tv} + \eta(b_{Uv} - \alpha_{tv})$. Thus $\beta_{tvw} = \beta_{tvw*}$ if $b_{Utv} = \infty$. If $\beta_{tvw*} < \alpha_{tv}$, let β_{tvw} be the maximum of β_{tvw*} and $\alpha_{tv} + \eta(b_{Ltv} - \alpha_{tv})$.

Let $m(t, v, w)$ be the smallest nonnegative iinteger such that

$$\beta_{tvw} = (1 - \eta^{m(t,v,w)})\alpha_{tv} + \eta^{m(t,v,w)}(\beta_{tv0} - \alpha_{tv}) \tag{19}$$

satisfies the constraint that $\mathbf{y}_t + \beta_{tvw}\mathbf{s}_t$ is in O. If $m(t, v, w) = 0$, let $b_{Ltvw} = b_{Ltv(w-1)}$ and $b_{Utvw} = b_{Utv(w-1)}$. If $m(t, v, w) > 0$ and $(\alpha_{tv} - \beta_{tvw})g_0(\alpha_{tv}) > 0$, let $b_{Ltvw} = b_{Ltv(w-1)}$ and

$$b_{Utvw} = (1 - \eta^{m(t,v,w)-1})\alpha_{tv} + \eta^{m(t,v,w)-1}(\beta_{tvw} - \alpha_{tv}). \tag{20}$$

If $m(t, v, w) > 0$ and $(\alpha_{tv} - \beta_{tvw})g_{t1}(\alpha_{tv}) < 0$, let $b_{Utvw} = b_{Utv(w-1)}$ and

$$b_{Ltvw} = (1 - \eta^{m(t,v,w)-1})\alpha_{tv} + \eta^{m(t,v,w)-1}(\beta_{tvw} - \alpha_{tv}). \tag{21}$$

With these definitions, $b_{Ltvw} \leq q_t \leq b_{Utvw}$. If $g_t(\beta_{tvw}) > g_t(\alpha_{tv})$, then let $\alpha_{t(v+1)} = \beta_{tvw}$. In this case, $b_{Lt(v+1)} = b_{Ltvw}$ and $b_{Ut(v+1)} = \alpha_{t(v+1)}$ if $g_1(\alpha_{t(v+1)}) < 0$, $b_{Lt(v+1)} = \alpha_{t(v+1)}$ and $b_{Ut(v+1)} = b_{Utv}$ if $g_1(\alpha_{t(v+1)}) > 0$, and $q_t = \alpha_{t(v+1)}$ if $g_1(\alpha_{t(v+1)}) = 0$.

The algorithm for a line search always results in a satisfactory α_t, as shown in the following theorem:

Theorem 2. *For any nonnegative integer t such that $\nabla f(\mathbf{y}_t) \neq \mathbf{0}_p$, a nonnegative integer v exists such that Equation 15 holds.*

Proof. It suffices to consider the case in which Equation 15 holds for no nonnegative integer v. Because $g_t(\alpha_{tv})$, $v \geq 0$, is strictly increasing and bounded above by $g_t(q_t)$, $g_t(\alpha_{tv})$, $v \geq 0$, converges to a real number h_t. Because Q_t is a bounded interval, real numbers ν_0 and ν_1 in the closure of Q_t and a strictly increasing nonnegative integer-valued function w on the positive integers exist such that $w(1) \geq 0$, $\alpha_{tw(i)}$, $i \geq 1$, converges to ν_0, and $\alpha_{t[w(i)+1]}$, $i \geq 1$, converges to ν_1. Because $w(i)$, $i \geq 1$, is unbounded, $h_t = g_t(\nu) = g_t\nu_1)$, and ν and ν_1 are in Q_t. In addition, $g_{t1}(\alpha_{tw(i)})$, $i \geq 1$, converges to $g_{t1}(\nu)$, and $g_{t1}(\alpha_{t[w(i)+1]})$, $i \geq 1$, converges to $g_{t1}(\nu_1)$.

If

$$g_t(\nu) - g_t(0)] > \gamma_1\nu|g_{t1}(\nu; \mathbf{y}_t)| \tag{22}$$

or

$$g_t(\nu_1) - g_t(0) > \gamma_1\nu_1|g_{t1}(\nu_1)|, \tag{23}$$

then, for some integer $i \geq 1$, Equation 15 holds for v equal to either $w(i)$ or $w(i) + 1$, a contradiction. It follows that neither ν nor ν_1 is q_t, and both $g_{t1}(\nu)$ and $g_{t1}(\nu_1)$ are nonzero.

If $\nu = \nu_1$, then $\alpha_{t[w(i)+1]} - \alpha_{tw(i)}$, $i \geq 1$, converges to 0. Either $\nu < q_t$ or $\nu > q_t$. If $\nu < q_t$, then a nonnegative integer i_0 exists such that $q_t > \alpha_{t[w(i)+1]} > \alpha_{tw(i)}$ for all integers $i > i_0$. The inequality $q_t < b_{Utv}$ for $v \geq 0$ and Equations 16, 17, 18, and ?? then imply that $\nu_1 - \nu$ is not 0, a contradiction. Essentially the same argument applies if $\nu > q_t$. Thus $\nu \neq \nu_1$.

The SPC condition implies that $\nu \neq \nu_1$ can only occur if $\nu - q_t$ and $q_t - \nu_1$ have the same sign, so that $g_{t1}(\nu)$ and $g_{t2}(\nu_1)$ are nonzero and have opposite signs. It follows that $g_t(\alpha_{t(w(i)+1)} - g_t(\alpha_{tw(i)})$, $i \geq 1$ has a positive limit, an impossible result. The only remaining possibility is that Equation 15 holds for some integer $v \geq 0$. $\square$

Limiting Convergence Rates

Limiting convergence rates for all cases are relatively well-known, although some attention must be paid to imperfect line searches. To compare cases, it will

be assumed throughout this section that f is thrice continuously differentiable on an open subset O_3 of O that includes $\mathbf{x}$ and $-\nabla^2 f(\mathbf{x})$ is positive-definite. Let ξ be the ratio of the largest and smallest eigenvalues of $-\nabla^2 f(\mathbf{x})$. It is assumed that $\gamma_2 < (\xi + 1)^2/(4\xi)$ (Cleveland, 1971; Haberman, 1975). For a p by p matrix $\mathbf{A}$, let $|\mathbf{A}|$ be the supremum of $|\mathbf{A}\mathbf{z}|/|\mathbf{z}|$ for $\mathbf{z}$ in R^p such that $\mathbf{z} \neq \mathbf{0}_p$. Let $\nabla^3 f(\mathbf{y})$, $\mathbf{y}$ in O_3, be the function from R^p to the set of p by p matrices with value $\nabla^3 f(\mathbf{z}; \mathbf{y})$ for $\mathbf{z}$ in R^p such that $|\nabla^2 f(\mathbf{y} + \mathbf{z}) - \nabla^2 f(\mathbf{z}) - \nabla^3 f(\mathbf{z}; \mathbf{y})|/|\mathbf{z}|$ approaches 0 as $\mathbf{z}$ in R^p, $\mathbf{z} \neq \mathbf{0}_p$ and $\mathbf{y} + \mathbf{z}$ in O, approaches $\mathbf{0}_p$. Let F be the supremum of $|[\nabla^2 f(\mathbf{x}]^{-1}\nabla^3 f(\mathbf{z}; \mathbf{x})|/|\mathbf{z}|$ for $\mathbf{z}$ in R^p, $\mathbf{z} \neq \mathbf{0}_p$. If f is a quadratic function, then $F = 0$.

Gradient Ascent

Here the direction is defined as in section The Gradient Ascent Algorithm and the step size is defined as in section Searches for Gradient Ascent and the Conjugate Gradient Method. In this case, it is easiest to apply the norm $||\mathbf{z}|| = (-\mathbf{z}'\nabla^2 f(\mathbf{x})\mathbf{z})^{1/2}$ for $\mathbf{z}$ in R^p. If $\mathbf{y}_t$ is not $\mathbf{x}$ for an integer $t \geq 0$, then the upper limit of $||\mathbf{y}_{t+1} - \mathbf{x}||/||\mathbf{y}_t - \mathbf{x}||$ does not exceed $(\xi - 1)/(\xi + 1)$. If f is quadratic, then $||\mathbf{y}_{t+1} - \mathbf{x}||/||\mathbf{y}_t - \mathbf{x}|| \leq (\xi - 1)/(\xi + 1)$ for t sufficiently large (Shewchuk, 1994).

Newton-Raphson

Consider both a direction defined as in section The Newton-Raphson Algorithm and a step size defined as in section The Newton-Raphson Search. If no $\mathbf{y}_t$, $t \geq 0$, is $\mathbf{x}$, then the quadratic convergence property holds that the upper limit of $|\mathbf{y}_{t+1} - \mathbf{x}|/|\mathbf{y}_t - \mathbf{x}|$, $t \geq 0$, does not exceed $F/2$ and $\alpha_t = 1$ for all t sufficiently large (Haberman, 1974, p. 48). If f is a quadratic function, then $\mathbf{y}_t = \mathbf{x}$ for some $t \geq 0$, with $\mathbf{y}_1 = \mathbf{x}$ if $|\mathbf{s}_0| \leq \kappa$.

Approximate Newton-Raphson

Consider both a direction defined as in section Approximate Newton-Raphson Algorithms and a step size defined as in section Approximate Newton-Raphson Searches. Assume that $\mathbf{y}_t \neq \mathbf{x}$ for all $t \geq 0$, and assume that $\mathbf{V}$ is continuous at $\mathbf{x}$ and $\mathbf{V}(\mathbf{x})$ is positive-definite. The most satisfactory results arise when $\mathbf{V}(\mathbf{x})$ is close to $-\nabla^2 f(\mathbf{x})$. Because $\nabla f(\mathbf{x}) = \mathbf{0}_p$, for any real $\epsilon > 0$ a real $\delta > 0$ exists such that $\mathbf{y}$ is in O,

$$|\nabla f(\mathbf{y}) - \nabla^2 f(\mathbf{x}(\mathbf{y} - \mathbf{x})| < \epsilon|\mathbf{y} - \mathbf{x}|, \tag{24}$$

and

$$|f(\mathbf{y} - f(\mathbf{x}) - \frac{1}{2}(\mathbf{y} - \mathbf{x})'\nabla^2 f(\mathbf{x}(\mathbf{y} - \mathbf{x})| < \epsilon|\mathbf{y} - \mathbf{x}|^2 \tag{25}$$

if $\mathbf{y}$ is in R^p and $|\mathbf{y} - \mathbf{x}| < \delta$. If $\mathbf{I}_p$ is the p by p identity matrix,

$$\omega = |\mathbf{I}_p + [\mathbf{V}(\mathbf{x})]^{-1}\nabla^2 f(\mathbf{x})| < 1, \tag{26}$$

$(1+2\gamma_1)\xi\omega^2 < 1$, and $\mathbf{y}_t \neq \mathbf{x}$ for all nonnegative integers t, then the linear convergence result holds that $|\mathbf{y}_{t+1} - \mathbf{x}|/|\mathbf{y}_t - \mathbf{x}|$, $t \geq 0$, does not exceed ω, and $\alpha_t = 1$ for all sufficiently large t. If $\mathbf{V}$ is defined as in Equation 10 and $\phi = dF/2 < 1$, then $\omega \leq \phi/(1 - \phi)$ (Wilkinson, 1971). If f is a quadratic function, then $\mathbf{y}_t = \mathbf{x}$ for some $t \geq 0$, with $\mathbf{y}_1 = \mathbf{x}$ if $|\mathbf{s}_0| \leq \kappa$. Although results such that d be small, it must be noted that some lower limits on d arise due to rounding errors involved in numerical differentiation. This issue is important because rounding errors for differences of nearly equal numbers depend on the magnitude of the numbers rather than on their differences.

Conjugate Gradients

Consider a direction defined as in section The Conjugate Gradient Algorithm and a step size defined as in section Searches for Gradient Ascent and the Conjugate Gradient Method. In principal, if f is quadratic, then $\mathbf{y}_p = \mathbf{x}$, although rounding errors and large initial values of $\mathbf{s}_t$ may cause problems with this result, especially for large p Shewchuk, 1994. In general, if no $\mathbf{y}_t = \mathbf{x}$ for a nonnegative integer t, then for any fixed positive integer $u \leq p$, a nonnegative integer v exists such that the upper limit of $||\mathbf{y}_{v+pi+u} - \mathbf{x}||/||\mathbf{y}_{v+pi} - \mathbf{x}||$, $i \geq 0$, does not exceed

$$\frac{2}{[(\xi^{1/2} - 1)/(\xi^{1/2} + 1)]^u + [(\xi^{1/2} + 1)/(\xi^{1/2} - 1)]^u}.$$

This result may be affected by rounding errors if p is large. It appears possible to modify the proof in Cohen (1972) to show, at least in principal, that the upper limit of $||\mathbf{y}_{v+p(i+1)} - \mathbf{x}||/||\mathbf{y}_{v+pi} - \mathbf{x}||^2$, $i \geq 0$, is finite.

Algorithm Selection

In the end, selection of algorithms depends on the balance between convergence rate and the computational labor required per iteration. This balance depends on the application. The Newton-Raphson algorithm generally requires the most computation per iteration but the fewest iterations. Relative to the Newton-Raphson algorithm, the approximations to the Newton-Raphson algorithm often require less computation per cycle and more iterations. The conjugate gradient approach typically involves much less computation per iteration but far more iterations. It is most likely to be appropriate for a high dimension p. Use of gradient ascent involves the minimum computation per iteration, but usually entails the most iterations.

References

Cauchy, A. (1847). Méthode générale pour la résolution des systèmes d'équations simultanées. *Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences*, *25*, 53–538.

Cleveland, W. S. (1971). Projection with the wrong inner product and its application to regression with correlated errors and linear filtering of time series. *Annals of Mathematical Statistics*, *42*, 616–624. https://doi.org/10.1214/aoms/1177693411

Cohen, A. I. (1972). Rate of convergence of several conjugate gradient algorithms. *SIAM Journal on Numerical Analysis*, *9*, 248–259. https://doi.org/10.2307/2156398

Fisher, R. A. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, *22*, 700–725. https://doi.org/10.1017/s0305004100009580

Fletcher, R., & Reeves, C. M. (1964). Function minimization by conjugate gradients. *The Computer Journal*, *7*, 149–154. https://doi.org/10.1093/comjnl/7.2.149

Haberman, S. J. (1974). *The analysis of frequency data*. University of Chicago Press.

Haberman, S. J. (1975). How much do Gauss-Markov and least square estimates differ? a coordinate-free approach. *Annals of Statistics*, *3*, 982–990. https://doi.org/10.1214/aos/1176343201

Haberman, S. J. (2013). *A general program for item-response analysis that employs the stabilized Newton-Raphson algorithm* (ETS Research Report RR-13-32). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2013.tb02339.x

Louis, T. (1982). Finding the observed information matrix when using the *em* algorithm. *Journal of the Royal Statistical Society, Ser. B*, *44*, 226–233. https://doi.org/10.1111/j.2517-6161.1982.tb01203.x

Polak, E., & Ribière, G. (1969). Note sur la convergence de méthodes de directions conjuguées. *Revue française d'informatique et de recherche opérationnelle. Série rouge*, *3*(16), 35–43. https://doi.org/10.1051/m2an/196903r100351

Ponstein, J. (1967). Seven kinds of convexity. *SIAM Review*, *9*, 115–119. https://doi.org/10.1137/1009007

Shewchuk, J. R. (1994). *An introduction to the conjugate gradient method without the agonizing pain* (Computer Science Technical Report CMU-CS-94-125). School of Computer Science, Carnegie-Mellon University.

Wilkinson, J. H. (1971). Modern error analysis. *SIAM Review*, *13*, 548–568. https://doi.org/10.2307/2029191

Wolfe, P. (1969). Convergence conditions for ascent methods. *SIAM Review*, *11*, 226–235. https://doi.org/10.1137/1011036

Wolfe, P. (1971). Convergence conditions for ascent methods. II: Some corrections. *SIAM Review*, *13*, 185–188. https://doi.org/10.1137/1013035

Yamamoto, T. (2001). Historical developments in convergence analysis for Newton's and Newton-like methods. *Numerical analysis: Historical developments in the 20th century* (pp. 241–263). Elsevier. https://doi.org/10.1016/b978-0-444-50617-7.50011-2

Ypma, T. J. (1995). Historical development of the Newton-Raphson method. *SIAM Review*, *37*, 531–551. https://doi.org/10.2307/2132904

Zangwill, W. I. (1969). *Nonlinear programming: A unified approach.* Prentice-Hall.