An improvement of the Goldstein line search

Supplementary material for Optimization Letters

Arnold Neumaier
Fakultät für Mathematik, Universität Wien
Oskar-Morgenstern-Platz 1, A-1090 Wien, Austria
email: Arnold.Neumaier@univie.ac.at
WWW: `http://www.mat.univie.ac.at/~neum`

Morteza Kimiaei
Fakultät für Mathematik, Universität Wien
Oskar-Morgenstern-Platz 1, A-1090 Wien, Austria
email: kimiaeim83@univie.ac.at
WWW: `http://www.mat.univie.ac.at/~kimiaei`

August 15, 2023

This paper provides the supplementary material for [2]. In the first section, we discuss two theoretical results: the first is used to obtain the complexity result for `CLS` and the second may be used to further improve `CLS` as discussed in the second section.

1 Variations on a theme by Goldstein

The goal of a line search is to find a value for the step size such that $f(x(\alpha))$ is sufficiently smaller than $f(x)$. Given $\beta \in]0, 1/4[$, this is measured by the **efficiency criterion**

$$(f(x) - f(x(\alpha)))\frac{\|p\|^2}{(g(x)^Tp)^2} \ge \frac{2\beta}{\overline{\gamma}} \tag{1}$$

of Warth & Werner [3]. A useful measure of progress of a line search is the **Goldstein quotient**

$$\mu(\alpha) := \frac{f(x+\alpha p) - f(x)}{\alpha g(x)^T p} \quad \text{for} \quad \alpha > 0 \tag{2}$$

first considered by Goldstein [1]. The Goldstein condition

$$f(x) + \alpha\mu'' g(x)^T p \le f(x+\alpha p) \le f(x) + \alpha\mu' g(x)^T p \quad \text{with fixed } 0 < \mu' < \mu'' < 1 \tag{3}$$

is equivalent to

$$\mu' \le \mu(\alpha) \le \mu''. \tag{4}$$

We define the **sufficient descent condition** (**SDC**)

$$\mu(\alpha)|\mu(\alpha) - 1| \ge \beta \tag{5}$$

with fixed $\beta \in]0, 1/4[$.

A practical, easily checkable condition can be given in terms of curvature information about the graph of

$$\psi(\alpha) := f(x + \alpha p).$$

Such curvature information is contained in the magnitude of the second order **divided differences**

$$\psi[\alpha_1, \alpha_2, \alpha_3] := \frac{\psi[\alpha_1, \alpha_2] - \psi[\alpha_1, \alpha_3]}{\alpha_2 - \alpha_3}, \tag{6}$$

where

$$\psi[\alpha_1, \alpha_2] := \frac{\psi(\alpha_2) - \psi(\alpha_1)}{\alpha_2 - \alpha_1} = \psi[\alpha_2, \alpha_1] \tag{7}$$

defines the **slopes** (first order divided differences) of ψ. Using $\psi[\alpha, \alpha] := \psi'(\alpha)$, the divided differences make also sense when two of the arguments coincide; clearly the above result remains valid in this limited case. In particular,

$$\psi[0, 0] = g(x)^T p, \quad \psi[0, \alpha] = \mu(\alpha) g(x)^T p, \tag{8}$$

$$\psi[0, \alpha, \alpha'] = \frac{\psi[0, \alpha] - \psi[0, \alpha']}{\alpha - \alpha'} = \frac{\mu(\alpha) g(x)^T p - \mu(\alpha') g(x)^T p}{\alpha - \alpha'} = \frac{\mu(\alpha) - \mu(\alpha')}{\alpha - \alpha'} g(x)^T p, \tag{9}$$

$$(\mu(\alpha) - 1) g(x)^T p = \psi[0, \alpha] - \psi[0, 0] = \alpha \psi[0, 0, \alpha]. \tag{10}$$

The following result is used in proving [2, Theorem 3].

Proposition 1 *For arbitrary $\alpha_1, \alpha_2, \alpha_3 \in \mathbb{R}$, we have*

$$|\psi[\alpha_1, \alpha_2, \alpha_3]| \le \frac{\overline{\gamma}}{2} \|p\|^2. \tag{11}$$

In particular, for a straight line search $x(\alpha) = x + \alpha p$,

$$f(x + \alpha p) = f(x) + \alpha g(x)^T p + \alpha^2 \psi[0, 0, \alpha], \quad |\psi[0, 0, \alpha]| \le \frac{\overline{\gamma}}{2} \|p\|^2 \tag{12}$$

holds and the Goldstein quotient is Lipschitz continuous, i.e.,

$$|\mu(\alpha) - \mu(\alpha')| \le \Gamma |\alpha - \alpha'| \ \ \text{for } \alpha, \alpha' > 0 \tag{13}$$

holds, and satisfies

$$|\mu(\alpha) - 1| \le \Gamma \alpha \ \ \text{for } \alpha > 0, \tag{14}$$

holds, where Γ is

$$\Gamma := \frac{\overline{\gamma} \|p\|^2}{2\nu}. \tag{15}$$

with the Lipschitz constant $\overline{\gamma} > 0$.

Proof. If the α_j are distinct, the functions ϕ_2 and ϕ_3 defined by

$$\phi_j(t) := \frac{\psi(\alpha_1 + t(\alpha_j - \alpha_1))}{\alpha_j - \alpha_1}$$

for $j = 2, 3$ satisfy

$$\phi_j'(t) = \psi'(\alpha_1 + t(\alpha_j - \alpha_1)) = \Big(g(x + (\alpha_1 + t(\alpha_j - \alpha_1))p)\Big)^T p.$$

By the generalized Cauchy-Schwarz inequality and (A1) for the gradient,

$$\begin{aligned} |\phi_2'(t) - \phi_3'(t)| &= \Big|\Big(g(x + (\alpha_1 + t(\alpha_2 - \alpha_1))p) - g(x + (\alpha_1 + t(\alpha_3 - \alpha_1))p)\Big)^T p\Big| \\ &\le \Big\|g(x + (\alpha_1 + t(\alpha_2 - \alpha_1))p) - g(x + (\alpha_1 + t(\alpha_3 - \alpha_1))p)\Big\|_* \|p\| \\ &\le \overline{\gamma}\|t(\alpha_2 - \alpha_3)p\|\|p\| = \overline{\gamma}|t(\alpha_2 - \alpha_3)|\ \|p\|^2. \end{aligned}$$

Therefore, the derivative of

$$\phi(t) := \frac{\phi_2(t) - \phi_3(t)}{\alpha_2 - \alpha_3}$$

is bounded by $|\phi'(t)| \le \overline{\gamma} t \|p\|_2^2$ for $t \ge 0$. This implies

$$|\psi[\alpha_1, \alpha_2, \alpha_3]| = |\phi(1) - \phi(0)| = \Big|\int_0^1 \phi'(t)dt\Big| \le \int_0^1 |\phi'(t)|dt \le \int_0^1 t\overline{\gamma}\|p\|_2^2 dt = \frac{\overline{\gamma}}{2}\|p\|_2^2$$

and proves (11) when the α_j are distinct. Taking limits, (11) follows generally. From (8) and (10) we conclude that

$$f(x + \alpha p) - f(x) = \alpha\mu(\alpha)g(x)^T p = \alpha g(x)^T p + \alpha^2\psi[0, 0, \alpha].$$

A comparison with Taylor's theorem now shows $|\psi[0, 0, \alpha]| \le \frac{\overline{\gamma}}{2}\|p\|_2^2$, and (12) follows.

(13) is obtained from (9) and (11),

$$|\mu(\alpha) - \mu(\alpha')| = |\psi[0, \alpha, \alpha']|\frac{|\alpha - \alpha'|}{|g(x)^T p|} \le \frac{\overline{\gamma}\|p\|_2^2}{2|g(x)^T p|}|\alpha - \alpha'| = \Gamma|\alpha - \alpha'|.$$

In particular, choosing $\alpha' = 0$ and using $\mu(0) = 1$ and (13), we find (14). □

The following result is used to derive the early stopping test

$$(f(x) - f(x + \alpha' p))\rho \ge \beta(g(x)^T p)^2 \tag{16}$$

for an improved version of `CLS` discussed below. Here ρ is the maximum over all $|\psi[\alpha_i, \alpha_j, \alpha_k]|$ computable from the information accumulated so far in the line search.

Theorem 1
(i) If α, α', α'' are distinct then

$$\psi[\alpha,\alpha'] = \frac{(\widetilde{\alpha}-\alpha)\psi[\alpha,\widetilde{\alpha}] + (\alpha'-\widetilde{\alpha})\psi[\alpha',\widetilde{\alpha}]}{\alpha'-\alpha}, \tag{17}$$

$$\psi[\alpha,\alpha',\alpha''] = \frac{(\widetilde{\alpha}-\alpha)\psi[\alpha,\alpha'',\widetilde{\alpha}] + (\alpha'-\widetilde{\alpha})\psi[\alpha',\alpha'',\widetilde{\alpha}]}{\alpha'-\alpha}. \tag{18}$$

(ii) Suppose that $\alpha_0 \leq \alpha_1 \leq \ldots \leq \alpha_m$. Then

$$\Big|\psi[\alpha_i,\alpha_j,\alpha_k]\Big| \leq \max_{l=1:m-1} \Big|\psi[\alpha_{l-1},\alpha_l,\alpha_{l+1}]\Big| \quad \textit{for } i,j,k = 0,\ldots,m. \tag{19}$$

holds.

Proof. (i) (17) is verified directly by expanding both sides using (7). By expanding the definition (6), we see that

$$\begin{aligned}
\psi[\alpha,\alpha',\alpha''] &= \frac{\psi[\alpha,\alpha'] - \psi[\alpha,\alpha'']}{\alpha'-\alpha''} = \frac{\dfrac{\psi(\alpha')-\psi(\alpha)}{\alpha'-\alpha} - \dfrac{\psi(\alpha'')-\psi(\alpha)}{\alpha''-\alpha}}{\alpha'-\alpha''} \\
&= \frac{(\alpha''-\alpha)(\psi(\alpha')-\psi(\alpha)) - (\alpha'-\alpha)(\psi(\alpha'')-\psi(\alpha))}{(\alpha'-\alpha)(\alpha''-\alpha)(\alpha'-\alpha'')} \\
&= \frac{(\alpha''-\alpha')\psi(\alpha) + (\alpha-\alpha')\psi(\alpha'') + (\alpha''-\alpha)\psi(\alpha')}{(\alpha'-\alpha)(\alpha''-\alpha)(\alpha'-\alpha'')} \\
&= \frac{\psi(\alpha)}{(\alpha-\alpha')(\alpha-\alpha'')} + \frac{\psi(\alpha')}{(\alpha'-\alpha)(\alpha'-\alpha'')} + \frac{\psi(\alpha'')}{(\alpha''-\alpha)(\alpha''-\alpha')}
\end{aligned}$$

is a symmetric function of α,α',α''. Now (18) follows for fixed α'' from (17) applied to $\phi(\alpha) := \psi[\alpha,\alpha'']$ in place of $\psi(\alpha)$ since

$$\psi[\alpha,\alpha',\alpha''] = \phi[\alpha,\alpha'].$$

(ii) We only need to prove the case where the α_i are distinct and $i < j < k$, since the general result then follows by symmetry and by taking confluent limits. Under this restriction we prove (19) by induction on $k-i$. Clearly, $k-i \geq 2$. If $k-i = 2$ then $i = j-1, k = j+1$, and (19) holds trivially. Thus assume that (19) holds when $k-i < d \in \{3,\ldots,m\}$. The case where $k-i = d$ can be reduced to the case $k-i < d$ by applying (18) with $\alpha = \alpha_i$, $\alpha' = \alpha_k$, $\alpha'' = \alpha_j$, and $\widetilde{\alpha} = \alpha_h$ with $i < h < k$ and $h \neq j$ and using the triangle inequality. Thus (19) holds generally. □

2 Further possible improvements on CLS

We first recall from [2] Theorem 2 and the `CLS` algorithm. Then we discuss two possible improvements on the `CLS` algorithm.

Theorem 2 *Suppose that the restriction of the search path to* $[0, \alpha^*]$ *is a ray.*

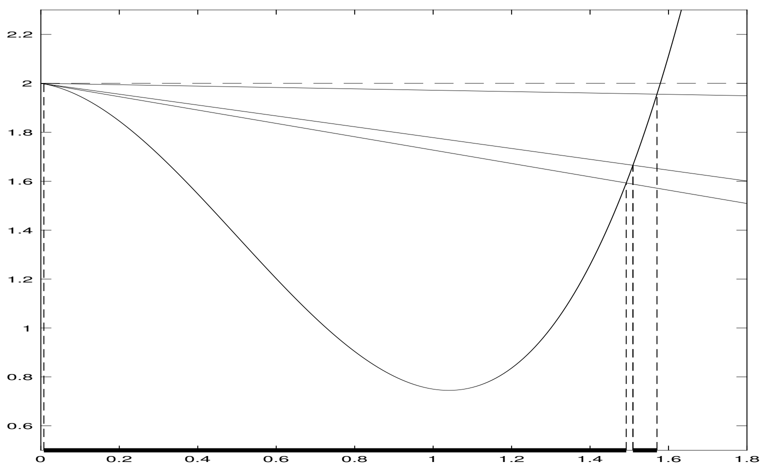
(i) If, for any function f *with Lipschitz continuous gradients, a line search procedure produces, if it terminates, a step satisfying*

$$(f(x) - f(x + \alpha p))\Big|\psi[\alpha_1, \alpha_2, \alpha_3]\Big| \geq \beta(g(x)^T p)^2 \tag{20}$$

for suitable $\alpha_1, \alpha_2, \alpha_3 \in [0, \alpha^*]$ *and* $\beta > 0$*, then the efficiency criterion* (1) *holds for any step size* $\alpha' \in [0, \alpha^*]$ *with* $f(x(\alpha')) \leq f(x(\alpha))$*. In particular, the line search procedure is efficient.*

(ii) The efficiency criterion (1) *also holds if* $\alpha \in [0, \alpha^*]$ *satisfies the sufficient descent condition* (5)*.*

Figure 1: The sufficient descent condition (5) with tuning parameter $\beta = 0.1$ for $f(x(\alpha)) = 2 - 0.25\alpha - 3\alpha^2 + 2\alpha^3$. Drawn are the lines with slopes zero (horizontal dashed lines), $\mu' g(x)^T p$, $\mu'' g(x)^T p$, $\mu''' g(x)^T p$ (solid lines), the resulting set of acceptable step sizes (fat lines), and their boundaries (vertical dash lines).



As one can see from Figure 1, where the minimizing $\widehat{\alpha}$ satisfies $\mu(\widehat{\alpha}) = 1$, the global minimizer does not necessarily satisfy the sufficient descent condition (5). Hence this condition is more demanding than just an efficient line search – it does not always allow to accept all points close to a minimizer of f along the search direction. For an efficient line search that does not suffer from this defect. We need to use the freedom to choose $\alpha_1, \alpha_2, \alpha_3$ different from the step size α actually employed. Thus we can satisfy (20) by setting $f(x + \alpha' p)$ to the best function value found so far, and verifying (16). Thus we can improve `CLS` by adding (16) as an early stopping test. In this case, α' is returned as the accepted step size. When (16) holds, $\rho = |\psi[\alpha_i, \alpha_j, \alpha_k]|$ for some i, j, k and for this i, j, k the efficiency criterion (1) is satisfied.

In an implementation, one initializes ρ with zero, sorts the step sizes already tried as (18) with $\alpha_0 = \alpha_1 = 0$. By Proposition 1, it is sufficient to compute the divided differences

Algorithm 1 `CLS`, **curved line search**

Purpose: `CLS` finds a step size α with $\mu(\alpha)|\mu(\alpha)-1| \geq \beta$

Input: $x(\alpha)$ (search path), $f_0 = f(x(0))$ (initial function value), $\nu = -g(x(0))^T x'(0)$ (minus directional derivative)

Tuning parameters: α_{init} (initial step size), $\alpha_{\max}$ (maximal step size), $\beta \in \,]0, \frac{1}{4}[$ (parameter for efficiency), $Q > 1$ (factor for extrapolation and interpolation), $0 < \kappa < \lambda < \infty$ (parameters for choosing α_{init} and $\alpha_{\max}$).

Requirements: $\nu > 0$, $\frac{\kappa\nu}{\|p\|^2} \leq \alpha_{\text{init}} \leq \alpha_{\max} \leq \frac{\lambda\nu}{\|p\|^2} < \infty$

Initialization: `first`=1; $\underline{\alpha} = 0$; $\overline{\alpha} = \infty$; $\alpha = \alpha_{\text{init}}$;

while 1 **do**
  compute the Goldstein quotient $\mu(\alpha) = (f_0 - f(x(\alpha)))/(\alpha\nu)$;
  if $\mu(\alpha)|\mu(\alpha)-1| \geq \beta$, **break**; **end** ▷ sufficient descent condition was satisfied
  if $\mu(\alpha) > \frac{1}{2}$, $\underline{\alpha} = \alpha$;
  elseif $\alpha = \alpha_{\max}$, **break**;
  else, set $\overline{\alpha} = \alpha$; ▷ linear decrease or more
  end
  if `first`, ▷ initially check whether function is almost quadratic or not
    `first` = 0;
    if $\mu(\alpha) < 1$, $\alpha = \frac{1}{2}\alpha/(1-\mu(\alpha))$; **else** $\alpha = \alpha Q$; **end**
  else
    if $\overline{\alpha} = \infty$, expand to $\alpha = \alpha Q$; ▷ extrapolation was done
    elseif $\underline{\alpha} = 0$, compute $\alpha = \frac{1}{2}\alpha/(1-\mu(\alpha))$; ▷ interpolation was done
    else, calculate $\alpha = \sqrt{\underline{\alpha}\,\overline{\alpha}}$; ▷ interval was found; geometric mean was computed
    end
  end
  restrict $\alpha = \min(\alpha, \alpha_{\max})$;
  end
end while
return α;

$\psi[\alpha_{l-1}, \alpha_l, \alpha_{l+1}]$ for $l = 1, \ldots, m-1$. Thus with each new function evaluation, one has to compute at most three new divided differences.

If second derivatives are available, one can also compute

$$\psi[0,0,0] = \frac{1}{2}\psi''(0) = \frac{1}{2}p^T G p$$

from the Hessian matrix, and initialize κ with $\frac{1}{2}|p^T G p|$. In this case, a quadratic model $\psi(\alpha) = f(x) + \alpha g(x)^T p + \dfrac{\alpha^2}{2} p^T G p$ suggests to begin with

$$\alpha_{\text{init}} = \min\left(-\frac{g(x)^T p}{p^T G p}, \alpha_{\max}\right)$$

and skip lines 13–16 of Algorithm 1.

References

[1] A. A. Goldstein. On steepest descent. *J. SIAM, Ser. A: Control* **3** (1965), 147–151.

[2] A. Neumaier, M. Kimiaei. `CLS`: An improvement of the Goldstein line search. Manuscript, (2023). `https://optimization-online.org/?p=21115`.

[3] W. Warth and J. Werner. Effiziente Schrittweitenfunktionen bei unrestringierten Optimierungsaufgaben. *Computing* **19** (1977), 59–72.