manuscript No.
(will be inserted by the editor)

An improvement of the Goldstein line search

Arnold Neumaier and Morteza Kimiaei

the date of receipt and acceptance should be inserted later

Abstract This paper introduces `CLS`, a new line search along an arbitrary smooth search path, that starts at the current iterate tangentially to a descent direction. Like the Goldstein line search and unlike the Wolfe line search, the new line search uses, beyond the gradient at the current iterate, only function values. Using this line search with search directions satisfying the bounded angle condition, global convergence to a stationary point is proved for continuously differentiable objective functions that are bounded below and have Lipschitz continuous gradients. The standard complexity bounds are proved under several natural assumptions.

Keywords Unconstrained optimization · gradient-free line search · complexity

2000 AMS Subject Classification: primary 90C56.

August 15, 2023

1 Introduction

The unconstrained optimization problem

$$\min f(x), \text{s.t. } x \in \mathbb{R}^n \tag{1}$$

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

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

has a very long history which we do not trace here; for the history and basic techniques see, instead, the books by Fletcher [5] or Nocedal & Wright [15]. Here we only discuss the state-of-the-art concerning line search conditions used in this context. Many of the current optimization methods employ the **Wolfe conditions** (Wolfe [19]) for line searches along descent directions. These line searches are called Wolfe line searches, which contain the **Armijo condition** (Armijo [1]) and the **curvature condition**. When the Armijo condition is satisfied, a step size is found that guarantees a sufficient reduction of f such that this reduction should be proportional to both the obtained step size and the directional derivative. Since the Armijo condition is satisfied for all sufficiently small step sizes, the curvature condition is required to guarantee that the directional derivative at the current accepted point is greater than the product of the directional derivative at the previous accepted point and a constant factor in $(0, 1)$. Checking the Armijo condition requires a function evaluation at each trial point, and checking the curvature condition requires an additional gradient evaluation at each trial point. The **Goldstein conditions** (Goldstein [6]) involve the two inequalities. The first inequality is used to avoid too small step sizes, while the second inequality is the Armijo condition. Although line searches based on the Armijo condition and backtracking or on satisfying the Goldstein conditions (Goldstein [6]) guarantee a sufficient reduction of f and can therefore be used to design globally convergent algorithms, they often behave poorly in strongly nonconvex regions. The Wolfe line searches partially overcome this weakness, but requires additional gradient evaluations before determining a new iterate.

Cartis et al. [3] proposed a non-monotone gradient-related descent algorithm using the Goldstein line search method. This line search enforces the Goldstein conditions by a construction of extrapolation, backtracking, or arithmetic mean bisection. If no step size satisfying the Goldstein conditions can be found by backtracking or extrapolation, a bracket is found instead whose lower bound is positive and whose upper bound is a positive finite number. This bracket contains an interval of points that satisfy the Goldstein conditions. Then, bisection is performed by taking the midpoint of the lower and upper bounds of the bracket while updating either the lower or upper bound of the bracket until the Goldstein conditions are violated.

Standard convergence theory requires that line search always finds a step size for which the production of the opposite reduction of f and the square of the norm of the direction over the directional derivative at the previous point from below remains bounded by a fixed positive number. This condition is called **efficiency criterion** by us since such a line search was called **efficient** by Warth & Werner [18]. For the basic convergence result see [14, Theorem 1.1], which is a result of Lemma 2.1 and Satz 2.2 of [18] and Theorem 4.2 of [11].

For a non-monotone gradient-related descent algorithm using the Goldstein line search condition, Cartis et al. [3, Theorem 3.4] proved the complexity of $\mathcal{O}(\varepsilon^{-2})$ iterations to reach a point x at which the gradient norm is below a given threshold ε. Royer & Wright [17] proved complexity results for second-order descent algorithms using a backtracking line search condition. Under the Polyak-Lojasiewicz condition, Karimi et al. [9, Theorem 1] obtained the $\mathcal{O}(\log \varepsilon^{-1})$ complexity result for gradient-type methods. For the history of complexity results see the recent book of Cartis et al. [2].

What is new. This paper is closely related to our unpublished preprint [11], which we split for publication into three papers (the present paper and [12,14]). In addition to the curved line search CLS and its discussion, taken from [11], the present paper discusses and proves complexity results (not yet in [11]). Compared to the traditional approaches, our paper has the following features:
• CLS is an efficient line search in the sense of Warth & Werner ([14, Theorem 1.1]), without requiring additional gradient evaluations, hence gives global convergence under a weak condition on the search directions.
• CLS uses a new sufficient descent condition, defined by (4) below, which is much easier to satisfy than the Goldstein conditions and the Wolfe conditions. In particular, CLS behaves well in strongly nonconvex regions (see, e.g., Figure 1), where it often accepts a much larger range of meaningful step sizes than Goldstein or Wolfe conditions.
• Like Cartis et al. [3], CLS satisfies a complexity bound on the number of function evaluations needed in CLS (Section 3), resulting for optimization methods using CLS in the same complexity results as Cartis et al. [3] and Karimi et al. [9].
• For strictly convex quadratic functions, termination after at most two iterations is guaranteed.
• The line search and its analysis are formulated for the case of searching along an arbitrary search path that need not be a ray but starts at the current iterate tangentially to a descent direction.

CLS was implemented in Matlab as part of the bound-constrained optimization software LMBOPT, whose excellent practical performance was documented in Kimiaei et al. [10]. However, in this paper, we perform CLS, the Wolfe line search algorithm, the Armijo line search algorithm, and the Goldstein line search algorithm along the standard BFGS direction and compare them. Numerical results show on the 130 unconstrained test problems with dimensions 2 to 10 from the CUTEst collection [8] that CLS competes with the Wolfe line search and is much more efficient and robust than the Armijo line search and Goldstein line search. Since CLS requires only the directional derivative as input and no gradient evaluation within the line search, CLS is recommended for solving real-world problems where computing gradients is more expensive than computing function values.

2 A curved line search

In this section, we introduce a curved line search algorithm (CLS) and define some important concepts, such as a new sufficient descent condition, the efficiency criterion of Warth & Werner [18] and the Goldstein quotient of Goldstein [7], provide the motivation for constructing an efficient line search method, and discuss requirements needed for complexity results.

Throughout of the paper we assume the following assumption:
(L) The function f is continuously differentiable on $\mathbb{R}^n$. Its gradient $g(x) = f'(x)^T$ is Lipschitz continuous with Lipschitz constant $\overline{\gamma} > 0$, i.e.,

$$\|g(x) - g(x')\|_* \leq \overline{\gamma}\|x - x'\| \quad \text{with } \overline{\gamma} > 0.$$

Here $\|\cdot\|$ is an arbitrary norm and $\|\cdot\|_*$ is its dual norm, satisfying the **generalized Cauchy–Schwarz inequality** $|y^T s| \leq \|y\|_* \|s\|$.

2.1 The `CLS` algorithm

A **line search** proceeds by searching points $x(\alpha)$ on a directionally differentiable curve of feasible points parameterized by a **step size** $\alpha \geq 0$ starting at the current point $x = x(0)$. If the gradient $g = g(x) = f'(x)$ of the function f at x is nonzero, the existence of an $\alpha > 0$ with $f(x(\alpha)) < f(x)$ is guaranteed if the tangent vector

$$p := x'(0) \tag{2}$$

is a **descent direction**, i.e.,

$$g(x)^T p < 0 \tag{3}$$

holds. In most line searches, **straight** line search paths $x(\alpha) = x + \alpha p$ in direction p are used, where $\alpha > 0$ is an accepted step length. However, application may require other search paths. For example, the bound-constrained solver `LMBOPT` [10] uses our line search for a piecewise linear search path $x(\alpha)$.

The goal of a line search is to find a value for the step size such that $f(x(\alpha))$ is sufficiently smaller than $f(x)$.

Our line search algorithm `CLS` uses a simple bisection procedure that updates a bracket $[\underline{\alpha}, \overline{\alpha}]$ containing $\widehat{\alpha}$ with $\mu(\widehat{\alpha}) = \frac{1}{2}$ until the **sufficient descent condition** (**SDC**)

$$\mu(\alpha)|\mu(\alpha) - 1| \geq \beta \tag{4}$$

holds for some fixed $\beta \in]0, 1/4[$. The restriction on β is needed since the left hand side of (4) is $\leq 1/4$ for $\mu(\alpha) \in [0, 1]$.

The Boolean variable `first` in the while loop ensures that the quadratic case will be optimally handled. In the first iteration we use a formula that for strictly quadratic objective functions leads to the minimizer; cf. (20) below. If the resulting next value for μ does not satisfy SDC, the function is far from quadratic and bounded, and we proceed with a simple bisection scheme: Until we know a bracket with $\underline{\alpha} > 0$ and $\overline{\alpha} < \infty$, we either interpolate (**interpolation step**), or we extrapolate with a constant factor $Q > 1$ (**extrapolation step**). The interpolation step after the first iteration guarantees, for sufficiently large $\alpha_{\max}$, that for nearly quadratic objective functions, the line search takes at most two function values.

Once a bracket $[\underline{\alpha}, \overline{\alpha}]$ with $0 < \underline{\alpha} < \overline{\alpha} < \infty$ is found, a **geometric mean step** is done by taking the geometric mean of the lower bound $\underline{\alpha}$ and the upper bound $\overline{\alpha}$ updating $\underline{\alpha}$ or $\overline{\alpha}$ in the next iteration. This is an improvement over the arithmetic mean bisection of Cartis et al. [3] whenever the bracket spans several orders of magnitude. This may happens when the objective function is nearly quadratic due to the interpolation step after the first iteration. We quit the line search once the stopping test is satisfied and return the final step size α.

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;

The argument in the proof of Theorem 1 below shows that `CLS` with $\alpha_{\max} = \infty$ either terminates, or produces an infinite sequence of α with $f(x + \alpha p) \to -\infty$. Because of Theorem 2 below, `CLS` defines for $\alpha_{\max} = \infty$ an efficient line search and achieves a well-defined minimal reduction in the function value. A finite but large bound on $\alpha_{\max}$ is needed in practice to account for the possibility that f is unbounded below.

To comply with our complexity analysis in Section 3 below, `CLS` restricts α_{init} and $\alpha_{\max}$ such that

$$\frac{\kappa\nu}{\|p\|^2} \leq \alpha_{\text{init}} \leq \alpha_{\max} \leq \frac{\lambda\nu}{\|p\|^2}, \quad \nu = |g(x)^T p| \tag{5}$$

holds for fixed tuning parameters $0 < \kappa < \lambda < \infty$. The bounds for the number of function evaluations for extrapolation, interpolation, and geometric mean phases, to be derived in Section 3, then depend on κ and λ but not on ν.

The best values for β and Q depend on the particular algorithm calling the line search, and must be determined by calibration on a set of test problems. In `LMBOPT` [10, Section 5.2], the default values used are $\beta = 0.02$ and $Q = 25$.

Sufficient progress in a line search algorithm is measured by the **efficiency criterion**

$$(f(x) - f(x(\alpha))\frac{\|p\|^2}{(g(x)^T p)^2} \geq \frac{2\beta}{\overline{\gamma}} \tag{6}$$

of Warth & Werner [18] for some $\beta > 0$ and the Lipschitz constant $\overline{\gamma} > 0$. Since the Lipschitz constant is generally unknown, a computationally more useful measure of progress of a line search is the **Goldstein quotient**

$$\mu(\alpha) := \frac{f(x + \alpha p) - f(x)}{\alpha g(x)^T p} \quad \text{for} \quad \alpha > 0 \tag{7}$$

first considered by Goldstein [7]. The descent condition (3) implies that for every $\alpha > 0$, we have $f(x(\alpha)) < f(x)$ iff $\mu(\alpha) > 0$. The Goldstein quotient can be extended to a continuous function $\mu : [0, \infty] \to \mathbb{R}^n$ by defining $\mu(0) := 1$ since, by l'Hôpital's rule and (2),

$$\lim_{\alpha \to 0} \mu(\alpha) = \lim_{\alpha \to 0} \frac{f'(x(\alpha))x'(\alpha)}{g(x)^T p} = \frac{f'(x)x'(0)}{f'(x)p} = 1.$$

More generally, we shall need the second order **divided differences**

$$\psi[\alpha_1, \alpha_2, \alpha_3] := \frac{\psi[\alpha_1, \alpha_2] - \psi[\alpha_1, \alpha_3]}{\alpha_2 - \alpha_3} \tag{8}$$

of $\psi(\alpha) := f(x + \alpha p)$, where

$$\psi[\alpha_1, \alpha_2] := \frac{\psi(\alpha_2) - \psi(\alpha_1)}{\alpha_2 - \alpha_1} = \psi[\alpha_2, \alpha_1] \tag{9}$$

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,

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

$$\psi[0, \alpha, \alpha'] = \frac{\psi[0, \alpha] - \psi[0, \alpha']}{\alpha - \alpha'} = \frac{\mu(\alpha) - \mu(\alpha')}{\alpha - \alpha'} g(x)^T p. \tag{11}$$

In a straight line search path $x(\alpha) = x + \alpha p$, the Armijo condition

$$f(x + \alpha p) \leq f(x) + \alpha \mu' g(x)^T p \quad \text{with } 0 < \mu' < 1 \tag{12}$$

is equivalent to $\mu(\alpha) \leq \mu'$, and the Goldstein conditions

$$f(x) + \alpha \mu'' g(x)^T p \leq f(x + \alpha p) \leq f(x) + \alpha \mu' g(x)^T p \text{ with fixed } 0 < \mu' < \mu'' < 1 \tag{13}$$

are equivalent to

$$\mu' \leq \mu(\alpha) \leq \mu''. \tag{14}$$

The curvature condition by Wolfe

$$g(x + \alpha p)^T p \geq \eta g(x)^T p \quad \text{with } \mu' \leq \eta < 1 \tag{15}$$

cannot be expressed in terms of the Goldstein quotient since it depends on an additional gradient at the trial point $x + \alpha p$.

2.3 Satisfaction of the sufficient descent condition

In the following, we write

$$\underline{f} = \inf_{\alpha \geq 0} f(x(\alpha)); \tag{16}$$

$\underline{f}$ is finite if f is bounded below.

Theorem 1 *Let $\beta \in]0, \frac{1}{4}[$, $g(x)^T p < 0$.*
(i) If the equation $\mu(\widehat{\alpha}) = \frac{1}{2}$ has a solution $\widehat{\alpha} > 0$ then α sufficiently close to $\widehat{\alpha}$ satisfies (4).
(ii) If the equation $\mu(\widehat{\alpha}) = \frac{1}{2}$ has no solution $\widehat{\alpha} > 0$, then $f(x + \alpha p)$ is unbounded below for $\alpha \to \infty$.

Proof (i) Since $\mu(\widehat{\alpha})|\mu(\widehat{\alpha}) - 1| = \frac{1}{4} > \beta$, (4) holds for all α sufficiently close to $\widehat{\alpha}$.

(ii) $\mu(0) = 1$ and the assumption that $\mu(\widehat{\alpha}) = \frac{1}{2}$ has no solution $\widehat{\alpha} > 0$ imply by continuity that $\mu_0 := \inf_{\alpha \geq 0} \mu(\alpha) \geq \frac{1}{2}$ and therefore for all $\alpha > 0$

$$\underline{f} - f(x) \leq f(x(\alpha)) - f(x) = \alpha g(x)^T p \mu(\alpha) \leq \alpha g(x)^T p \mu_0 \leq \frac{\alpha}{2} g(x)^T p \mu_0. \tag{17}$$

For $\alpha \to \infty$, we find $\underline{f} = -\infty$. □

The sufficient descent condition SDC requires $\mu(\alpha)$ to be both not too close to one, forbidding steps that are too short, and sufficiently positive, typically forbidding steps that are too long by forcing $f(x(\alpha)) < f(x)$. The condition is easier to satisfy than the Goldstein conditions (13). Indeed, (13) is equivalent to (14); hence (4) holds with $\beta = \mu'(1 - \mu'') > 0$. Conversely, with

$$\mu' = \frac{2\beta}{1 + \sqrt{1 - 4\beta}}, \quad \mu'' = \frac{1 + \sqrt{1 - 4\beta}}{2}, \tag{18}$$

The SDC implies that either (13) holds or the alternative **fast descent condition**

$$\mu(\alpha) \geq \mu''' \tag{19}$$

holds with $\mu''' = (1 + \sqrt{1 + 4\beta})/2$.

The Goldstein conditions (14) can be interpreted geometrically: In the graph of $f(x(\alpha))$, the cone defined by the two lines through $(0, f)$ with slopes $\mu' g(x)^T p$ and $\mu'' g(x)^T p$ cuts out a section of the graph, which defines the admissible step size parameters. Similarly, equality in (19) defines another line that determines the boundary of another section of the graph leading to admissible step size parameters. An illustrative example is given in the online supplement [13, Figure 1].

Satisfying The SDC guarantees by the preceding discussion a sensible decrease in the objective function.

The left hand side of (4) is largest for $\mu(\alpha) = \frac{1}{2}$, but the value $\mu(\alpha) = \frac{1}{2}$ has another significance: Near a local minimizer, twice continuously differentiable functions are bounded from below and, because of Taylor's theorem, almost quadratic. In the special case of a linear search path and a strictly convex quadratic function,

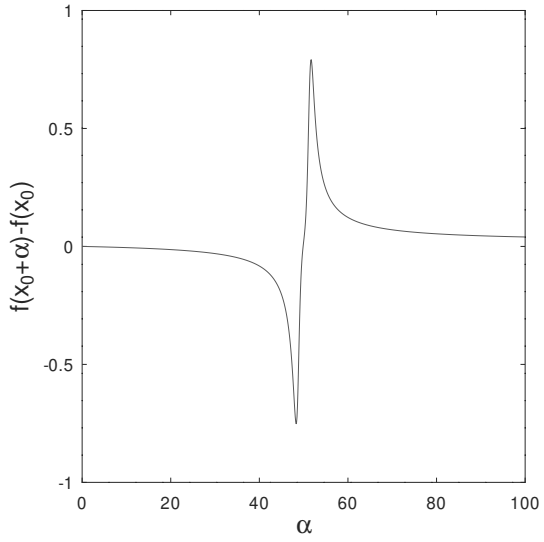
$$f(x + \alpha p) = f(x) + \alpha g(x)^T p + \frac{\alpha^2}{2} p^T G(x) p =: f + a\alpha + b\alpha^2 = f - \frac{a^2}{4b} + b(\alpha - \widehat{\alpha})^2$$

with $a < 0 < b$ and $\widehat{\alpha} = -a/2b > 0$. This implies that $\mu(\alpha) = 1 + b\alpha/a = 1 - \alpha/2\widehat{\alpha} < 1$ for $\alpha > 0$. In particular, $\mu(\widehat{\alpha}) = \frac{1}{2}$, and the minimizer

$$\widehat{\alpha} = \frac{\alpha}{2(1 - \mu(\alpha))} \tag{20}$$

along the search ray can be computed from any $\alpha > 0$. It is therefore natural to attempt to find a step size α with $\mu(\alpha) \approx \frac{1}{2}$. This can be done by a simple bisection procedure, updating a bracket $[\underline{\alpha}, \overline{\alpha}]$ containing $\widehat{\alpha}$ with $\mu(\widehat{\alpha}) = \frac{1}{2}$. This is our motivation for `CLS`.

Fig. 1: The function $f(x) := (x^3 + x)/((x^2 - 1)^2 + 5)$ along the path $x(\alpha) = x_0 + \alpha$ for $x_0 = -50$. Values $0 < \alpha < 49.87$ improve the function value. The SDC with our default choice $\beta = 0.02$ is satisfied for all $\alpha \in [1, 49.76]$.



Like, the Goldstein conditions, `CLS` is gradient-free once the search direction is given. But it avoids a defect of the Goldstein conditions and the Wolfe conditions: For the function drawn in Figure 1, the SDC allows all $\alpha \in [1, 49.76]$ since it includes the values where the fast descent condition (19) holds. On the other hand, one can easily see that both the Goldstein line search and the Wolfe line search require α to be a

tiny interval around 49.76. Qualitatively the same behavior is observed. In all cases where for small α, the graph of $f(x(\alpha))$ is – as Figure 1 – concave and fairly flat, while for large α $f(x(\alpha))$ is strongly increasing. Qualitatively the same behavior is observed.

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*

$$(f(x) - f(x + \alpha p))\Big|\psi[\alpha_1, \alpha_2, \alpha_3]\Big| \geq \beta(g(x)^T p)^2 \tag{21}$$

for suitable $\alpha_1, \alpha_2, \alpha_3 \in [0, \alpha^*]$ *and* $\beta > 0$*, then the efficiency criterion* (6) *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 (6) *also holds if* $\alpha \in [0, \alpha^*]$ *satisfies the sufficient descent condition* (4)*.*

Proof (i) By setting $(\alpha_1, \alpha_2, \alpha_3) = (0, 0, \alpha)$, we conclude from the definition of $\mu(\alpha)$, (4), and (11) that

$$\begin{aligned}(f(x) - f(x + \alpha p))\Big|\psi[\alpha_1, \alpha_2, \alpha_3]\Big| &= (f(x) - f(x + \alpha p))\Big|\psi[0, 0, \alpha]\Big| \\ &= \mu(\alpha)|g(x)^T p||(1 - \mu(\alpha))g(x)^T p| \\ &= \mu(\alpha)|1 - \mu(\alpha)|(g(x)^T p)^2 \geq \beta(g(x)^T p)^2.\end{aligned}$$

Hence (21) holds and (i) follows.

(ii) By [13, Proposition 1], we have for arbitrary $\alpha_1, \alpha_2, \alpha_3 \in \mathbb{R}$

$$|\psi[\alpha_1, \alpha_2, \alpha_3]| \leq \frac{\overline{\gamma}}{2}\|p\|^2. \tag{22}$$

Using (21) and (22), we obtain

$$\frac{(f(x) - f(x + \alpha p))\|p\|_2^2}{(g(x)^T p)^2} \geq \frac{2}{\overline{\gamma}}\frac{(f(x) - f(x + \alpha p))|\psi[\alpha_1, \alpha_2, \alpha_3]|}{(g(x)^T p)^2} \geq \frac{2\beta}{\overline{\gamma}}.$$

Since by assumption $x(\alpha) = x + \alpha p$ for $0 \leq \alpha \leq \alpha^*$, the left hand side is uniformly bounded away from zero and the efficiency criterion (6) holds. □

3 Complexity results for CLS

3.1 Line search complexity

Theorem 3 *Given* $0 < \kappa < \lambda < \infty$*,* $Q > 1$*,* $\beta \in]0, \frac{1}{4}[$*, and straight search paths* $x(\alpha) = x + \alpha p$*, where* $x, p \in \mathbb{R}^n$*,* $g(x)^T p < 0$*. If* α_{init} *and* $\alpha_{\max}$ *satisfy the condition*

(5) *then the number of function evaluations in* `CLS` *is bounded by a constant depending only on* $Q, \beta, \kappa, \lambda$, *and* $\overline{\gamma}$. *More precisely:*
(i) If $\mu(\alpha_{\text{init}}) > \mu''$, *then* `CLS` *ends after at most* $\overline{L}_E + \overline{M}_E$ *function evaluations, where*

$$\overline{L}_E = \left\lceil \log \frac{Q\lambda}{\kappa} \Big/ \log Q \right\rceil, \quad \overline{M}_E := \left\lceil \log_2 \frac{\overline{\gamma}\lambda \log Q}{2\mu'' - 2\mu'} \right\rceil. \tag{23}$$

(ii) If $\mu(\alpha_{\text{init}}) < \mu'$, *then* `CLS` *ends after at most* $\overline{L}_Q + \overline{M}_Q$ *function evaluations, where*

$$\overline{L}_Q := \left\lceil \log(\overline{\gamma}\lambda) \Big/ \log(2 - 2\mu') \right\rceil, \quad \overline{M}_Q := \left\lceil \log_2 \frac{\overline{\gamma}\lambda \log(\overline{\gamma}\lambda)}{2\mu'' - 2\mu'} \right\rceil. \tag{24}$$

(iii) Otherwise, `CLS` *ends after a single function evaluation.*

Proof As long as no step size satisfying (4) is found, either the extrapolation phase or the interpolation phase is performed until we find a bracket $[\underline{\alpha}, \overline{\alpha}]$ with $\underline{\alpha} > 0$ and $\overline{\alpha} < \infty$. Then, the geometric mean phase iteratively updates $\underline{\alpha}$ or $\overline{\alpha}$ by taking the geometric mean of $\underline{\alpha}$ and $\overline{\alpha}$, until we reach a step size α_{L_M} that satisfies (4), i.e.,

$$0 < \mu' \le \mu(\alpha_{L_M}) \le \mu'' < 1$$

with $0 < \mu' < \mu'' < 1$ from (18). Note that these are functions of β only.

Denote by $[\underline{\alpha}_\ell, \overline{\alpha}_\ell]$ the bracket $[\underline{\alpha}, \overline{\alpha}]$ after ℓ iterations of the geometric mean phase. By the updating rule,

$$\mu(\underline{\alpha}_\ell) > \mu'' > \mu' > \mu(\overline{\alpha}_\ell) \quad \text{for } \ell = 0, \dots, L_M. \tag{25}$$

By [13, Proposition 1], the Goldstein quotient is Lipschitz continuity

$$|\mu(\alpha) - \mu(\alpha')| \le \Gamma |\alpha - \alpha'| \ \text{ for } \alpha, \alpha' > 0, \tag{26}$$

where

$$\Gamma := \frac{\overline{\gamma}\|p\|^2}{2|g(x)^T p|}. \tag{27}$$

Using (26) and substituting $\ell = L_M$ into (25), we

$$\overline{\alpha}_{L_M} - \underline{\alpha}_{L_M} \ge \Gamma^{-1}(\mu(\underline{\alpha}_{L_M}) - \mu(\overline{\alpha}_{L_M})) > \Gamma^{-1}(\mu'' - \mu'). \tag{28}$$

By construction, $\alpha_{\ell+1} = \sqrt{\underline{\alpha}_{\ell+1}\overline{\alpha}_{\ell+1}}$ for $\ell \le L_M$ with either $\overline{\alpha}_{\ell+1} = \alpha_\ell$, $\underline{\alpha}_{\ell+1} = \underline{\alpha}_\ell$ or $\underline{\alpha}_{\ell+1} = \alpha_\ell$, $\overline{\alpha}_{\ell+1} = \overline{\alpha}_\ell$. Therefore

$$r_\ell := \log \frac{\overline{\alpha}_\ell}{\underline{\alpha}_\ell} > 0 \ \text{ for } \ell = 0, \dots, L_M,$$

satisfies

$$r_{\ell+1} = r_\ell / 2 \ \text{ for } \ell = 0, \dots, L_M - 1. \tag{29}$$

By induction and applying (28) to

$$r_{L_M} = \log \frac{\overline{\alpha}_{L_M}}{\underline{\alpha}_{L_M}} > 1 - \frac{\underline{\alpha}_{L_M}}{\overline{\alpha}_{L_M}} = \frac{\overline{\alpha}_{L_M} - \underline{\alpha}_{L_M}}{\overline{\alpha}_{L_M}},$$

we obtain

$$L_M = \log_2 \frac{r_0}{r_{L_M}} \leq \log_2 \frac{\overline{\alpha}_{L_M} r_0}{\overline{\alpha}_{L_M} - \underline{\alpha}_{L_M}} \leq \log_2 \frac{\overline{\alpha}_{L_M} r_0 \varGamma}{\mu'' - \mu'}. \tag{30}$$

We now distinguish the three cases: (A) $\mu(\alpha_{\text{init}}) > \mu'' > \frac{1}{2}$, (B) $\mu(\alpha_{\text{init}}) < \mu' < \frac{1}{2}$, and (C) $\mu'' \leq \mu(\alpha_{\text{init}}) \leq \mu'$.

Since case C is trivial, we need to consider only case A and case B.

Case A. Here `CLS` begins with the extrapolation phase. We first prove that to find a bracket, `CLS` takes at most $\overline{L}_E$ function evaluations. Suppose that the extrapolation phase takes L_E iterations. Then

$$\alpha_1 > \alpha_{\text{init}}, \quad \alpha_k = Q\alpha_{k-1} \quad \text{for } k = 2, \ldots, L_E, \tag{31}$$

$$\mu(\alpha_{L_E}) \leq \mu'' < \mu(\alpha_{L_E - 1}). \tag{32}$$

By (5) and (31), we have

$$Q^{L_E - 1}\alpha_{\text{init}} < Q^{L_E - 1}\alpha_1 = \alpha_{L_E} \leq \alpha_{\max} \leq \frac{\lambda\nu}{\|p\|^2} \leq \frac{\lambda\alpha_{\text{init}}}{\kappa}. \tag{33}$$

Therefore $Q^{L_E} \leq \dfrac{Q\lambda}{\kappa}$, giving

$$1 \leq L_E \leq \log \frac{Q\lambda}{\kappa} \Big/ \log Q \leq \overline{L}_E. \tag{34}$$

We now prove that once a bracket has been found, `CLS` uses at most $\overline{M}_E$ function evaluations. From

$$\alpha_{\text{init}} \leq \alpha_{L_E - 1} = \underline{\alpha}_0 \leq \underline{\alpha}_{L_M} \leq \alpha_{L_M} \leq \alpha_{\max}, \tag{35}$$

we conclude that

$$r_0 = \log \frac{\overline{\alpha}_0}{\underline{\alpha}_0} = \log \frac{\alpha_{L_E}}{\alpha_{L_E - 1}} = \log Q, \tag{36}$$

$$\overline{\alpha}_{L_M} \leq \alpha_{\max} \leq \lambda \frac{\nu}{\|p\|^2} \leq \frac{\overline{\gamma}\lambda}{2\varGamma}. \tag{37}$$

Here, we used (27), (33) and (35) to obtain (37). By substituting (27), (36), and (37) into (30), we conclude that

$$0 \leq L_M \leq \log_2 \frac{\overline{\alpha}_{L_M} r_0 \varGamma}{\mu'' - \mu'} \leq \log_2 \frac{\overline{\gamma}\lambda \log Q}{2\mu'' - 2\mu'} \leq \overline{M}_E.$$

Case B. Here `CLS` begins with the interpolation phase. We first prove that to find a bracket, `CLS` takes at most $\overline{L}_Q$ function evaluations. Suppose that the quadratic interpolation phase takes L_Q iterations. Then

$$\alpha_k = \alpha_{k-1}/(2 - 2\mu(\alpha_{k-1})) \quad \text{for } k = 1, \ldots, L_Q, \tag{38}$$

$$\mu(\alpha_{L_Q}) \geq \mu' > \mu(\alpha_{L_Q-1}). \tag{39}$$

From (38) and (39), we obtain

$$\alpha_k \leq \frac{\alpha_{k-1}}{2 - 2\mu'} \quad \text{for } k = 1, \ldots, L_Q, \tag{40}$$

inductively leading to

$$\alpha_{L_Q-1} \leq \frac{\alpha_{\text{init}}}{(2 - 2\mu')^{L_Q-1}}. \tag{41}$$

By [13, Proposition 1], the Goldstein quotient is bounded, i.e.,

$$|\mu(\alpha) - 1| \leq \Gamma\alpha \quad \text{for } \alpha > 0. \tag{42}$$

From (41) and (42), we conclude that

$$(1 - \mu') < |1 - \mu(\alpha_{L_Q-1})| \leq \Gamma\alpha_{L_Q-1} \leq \frac{\Gamma\alpha_{\text{init}}}{(2 - 2\mu')^{L_Q-1}}. \tag{43}$$

Hence from (5), (27), and (43) we obtain

$$(2 - 2\mu')^{L_Q} \leq 2\Gamma\alpha_{\text{init}} \leq \overline{\gamma}\lambda, \tag{44}$$

so that

$$1 \leq L_Q \leq \log(\overline{\gamma}\lambda) \Big/ \log(2 - 2\mu') \leq \overline{L}_Q.$$

We now prove that once a bracket has been found, `CLS` uses at most $\overline{M}_Q$ further function evaluations. From

$$\alpha_{L_Q} = \underline{\alpha}_0 \leq \underline{\alpha}_{L_M} \leq \alpha_{L_M} \leq \overline{\alpha}_{L_M} \leq \alpha_{L_Q-1} = \overline{\alpha}_0 \leq \alpha_{\text{init}}, \tag{45}$$

(27), and (5), we conclude that

$$r_0 = \log \frac{\alpha_{L_Q-1}}{\alpha_{L_Q}} = \log|2 - 2\mu(\alpha_{L_Q-1})| \leq \log(2\Gamma\alpha_{\text{init}}) \leq \log(\overline{\gamma}\lambda) \tag{46}$$

and

$$\overline{\alpha}_{L_M} \leq \alpha_{\text{init}} \leq \frac{\overline{\gamma}\lambda}{2\Gamma}. \tag{47}$$

From (30), (46), and (47), we conclude that

$$0 \leq L_M \leq \log_2 \frac{r_0\overline{\alpha}_{L_M}\Gamma}{\mu'' - \mu'} \leq \log_2 \frac{\overline{\gamma}\lambda \log(\overline{\gamma}\lambda)}{2\mu'' - 2\mu'} \leq \overline{M}_Q.$$

□

As a consequence of Theorem 3, we obtain a complexity result for descent methods that generate sequence $x^0, x^1, x^2, \ldots$ of feasible points and $\ell = 0, 1, 2, \ldots$, assuming that we perform straight line searches along descent directions p^ℓ. We write $f_\ell := f(x^\ell)$ and $g^\ell := g(x^\ell)$.

We call a point $\widehat{x} \in \mathbb{R}^n$ a **strong local minimizer** of f if f is twice continuously differentiable in a neighborhood of $\widehat{x}$, the gradient $g(\widehat{x})$ of f at $\widehat{x}$ vanishes, and the Hessian $G(\widehat{x})$ of f at $\widehat{x}$ is positive definite.

Theorem 4 *Given constants $0 < \kappa < \lambda < \infty$, suppose that the search directions satisfy the bounded angle condition*

$$\frac{(g^\ell)^T p^\ell}{\|g^\ell\|_* \|p^\ell\|} \leq -\delta < 0 \quad \textit{for } \ell = 1, 2, \ldots \tag{48}$$

for some $\delta > 0$ and the initial step sizes are chosen such that (5) *holds. Then:*

(i) The number of function values needed to reach a point x with

$$\|g(x)\|_* \leq \varepsilon. \tag{49}$$

is $\mathcal{O}(\varepsilon^{-2})$.

(ii) If the sublevel set $\{x \in \mathbb{R}^n \mid f(x) \leq f(x^0)\}$ is bounded then, starting with x^0, some subsequence of the points generated converges to a stationary point.

(iii) If f has a strong local minimizer $\widehat{x}$ and no other stationary point then the number of function values needed to reach a point x with (49) *is $\mathcal{O}(\log \varepsilon^{-1})$. In particular, this holds if the Polyak-Lojasiewicz (PL) condition*

$$\frac{1}{2}\|g(x)\|^2 \geq \omega(f(x) - f(\widehat{x}))$$

is satisfied for some $\omega > 0$ and all $x \in \mathbb{R}^n$.

Proof We write $f_{\ell+1} := f(x^\ell + \alpha_\ell p^\ell)$ and assume that the algorithm terminates at x^L; hence

$$\|g(x^L)\|_* < \varepsilon \leq \|g(x^\ell)\|_* \quad \text{for } \ell < L. \tag{50}$$

(i) Since the efficiency criterion (6) holds and the search direction p satisfies the bounded angle condition (48), we have

$$f_\ell - f_{\ell+1} \geq \frac{2\beta}{\overline{\gamma}} \delta^2 \|g(x^\ell)\|_*^2 \geq \frac{2\beta}{\overline{\gamma}} \delta^2 \varepsilon^2 \quad \text{for } \ell < L. \tag{51}$$

Summing all inequalities (51) and using (16) gives

$$f_0 - \underline{f} \geq f_0 - f_L = \sum_{\ell=0}^{L-1} (f_\ell - f_{\ell+1}) \geq \frac{2\beta}{\overline{\gamma}} \delta^2 \varepsilon^2 L,$$

leading to

$$L \leq C\varepsilon^{-2}.$$

where $C := \dfrac{\overline{\gamma}(f_0 - \underline{f})}{2\beta\delta^2}$. Together with Theorem 3, this implies that the number of function evaluations is $\mathcal{O}(\varepsilon^{-2})$. This proves (i).

(ii) This follows by a standard compactness argument since $\inf_{\ell \geq 0} \|g(x^\ell)\|_* = 0$ by (i).

(iii) Under this condition, Theorem 2 guarantees that [14, Theorem 1.1(ii)] can be applied. Thus $\|x^\ell - \widehat{x}\| \leq cq^\ell$ and $\|g^\ell\|_* \leq c'q^\ell$ hold for some $0 < q < 1$. It follows that $\|g(x^\ell)\|_* \leq \varepsilon$ if

$$\ell = \left\lceil \frac{\log c'\varepsilon^{-1}}{\log(1/q)} \right\rceil = \mathcal{O}(\log \varepsilon^{-1}),$$

where $c' > 0$ and $0 < q < 1$. Again by Theorem 3, the number of function evaluations is $\mathcal{O}(\log \varepsilon^{-1})$. Karimi [9] et al. observed that the PI condition implies $\widehat{x}$ is a strong local minimizer of f and the hypothesis of (iii) holds. □

Cartis et al. [3] proved a complexity of $\mathcal{O}(\varepsilon^{-2})$ for a method that uses search directions satisfying

$$g(x)^T p \leq -\kappa_1 \|g(x)\|_*^2, \quad \|p\| \leq \kappa_2 \|g(x)\|_* \quad \text{for } \kappa_1, \kappa_2 > 0.$$

Since these imply the bounded angle condition (48) with $\delta = \kappa_1/\kappa_2$, their complexity result is analogous to our result (i). Closer analysis of their proof shows that under the stronger assumption on f stated in (iii) our $\mathcal{O}(\log \varepsilon^{-1})$ complexity result also applies for their method, since the argument given above extends to their situation.

4 Numerical results

In this section, we compare `CLS` with the Armijo line search algorithm (`ALS`), the Goldstein line search algorithm (`GLS`) and the Wolfe line search algorithm (`WLS`). We run all four line searches with a standard optimization algorithm, whose search directions are the quasi-Newton directions $p = -B^{-1}g$, where B is updated by the BFGS formula [5].

The values of the tuning parameters of `CLS` are $\beta = 0.02$, $Q = 25$ (as in `LMBOPT` [10]), $\kappa = 10^{-3}$, $\lambda = 10^3$, $l_{\max} = 8$, and $\alpha_{\max} = \infty$. The values of the tuning parameters of `WLS` of Moré and Thuente [16] are default values. In this algorithm, $\mu' = 0.1$ in the Armijo condition (12) and $\eta = 0.9$ in the curvature condition (15) are chosen. `ALS`, like `WLS`, chooses $\mu' = 0.1$ and `GLS` chooses $\mu' = 0.1$ and $\mu'' = 0.9$ in the Goldstein conditions (13). For all algorithms, the initial step size was set to one, except for `CLS`, which projected one into $[\kappa\nu/\|p\|^2, \lambda\nu/\|p\|^2]$ to guarantee the complexity result for `CLS`, where $\nu = |g(x)^T p|$.

Note that in a computer implementation, this idealized CLS needs an extra stopping test to ensure that it ends after finitely many steps even when f is unbounded below along the search curve. In addition, one needs to take measures that make CLS robust in the presence of rounding errors by forbidding steps that are so small that the change in function value is dominated by rounding errors. Details were discussed in the companion paper Kimiaei et al. [10].

We compare the four line search algorithms on all 130 unconstrained test problems with dimensions 1 to 10 from the CUTEst collection [8] using the standard initial points. Figure 2 shows the performance profile of Dolan and Moré [4] whose cost measures are the number nf of function evaluations and the number ng of gradient evaluations. CLS has the lowest ng on 75% of the test problems compared to the other three algorithms, while WLS has the lowest nf on 65% of the test problems compared to the other three algorithms. Moreover, out of the 130 test problems, WLS and CLS solve 114 and 112 problems, respectively. Consequently, CLS is competitive with WLS and preferable for solving real-world problems where computing gradients is expensive.

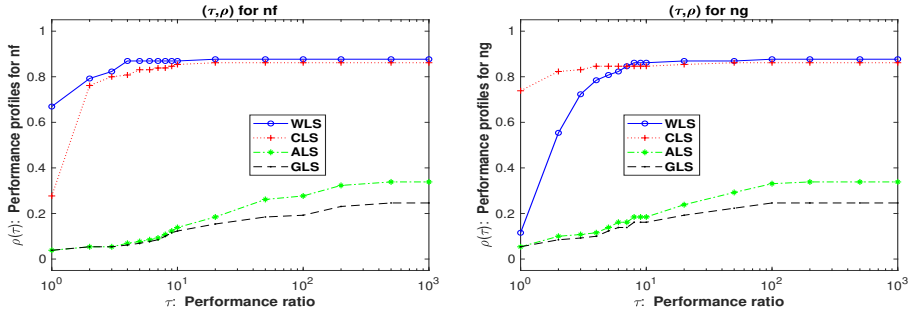


Fig. 2: Performance profile $\rho(\tau)$ independent of a bound τ on the performance ratio. Problems solved by no solver are ignored.

References

1. L. Armijo. Minimization of functions having Lipschitz continuous first partial derivatives. *Pacific J. Math.* **16**(1) (1966), 1–3.
2. C. Cartis, N. I. M. Gould, and Ph. L. Toint. *Evaluation Complexity of Algorithms for Nonconvex Optimization: Theory, Computation and Perspectives*, volume MO30 of MOS-SIAM Series on Optimization. SIAM, 2022.
3. C. Cartis, Ph. R. Sampaio, and Ph. L. Toint. Worst-case evaluation complexity of non-monotone gradient-related algorithms for unconstrained optimization. *Optimization* **64**(3) (2015), 1349–1361.
4. E. D. Dolan and J. J. Moré. Benchmarking optimization software with performance profiles. *Math. Program.* **91** (January 2002), 201–213.
5. R. Fletcher. *Practical Methods of Optimization.* John Wiley & Sons, Ltd (2000).
6. A. Goldstein and J. Price. An effective algorithm for minimization. *Numer. Math.* **10** (1967), 184–189.

7. A. A. Goldstein. On steepest descent. *J. SIAM, Ser. A: Control* **3** (1965), 147–151.
8. N. I. M. Gould, D. Orban, and Ph. L. Toint. CUTEst: a constrained and unconstrained testing environment with safe threads for mathematical optimization. *Comput. Optim. Appl.* **60** (2015), 545–557.
9. H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Lojasiewicz condition. In Machine Learning and Knowledge Discovery in Databases, pages 795–811, 2016.
10. M. Kimiaei, A. Neumaier, and B. Azmi. LMBOPT – A limited memory method for bound-constrained optimization. *Math. Program. Comput.* **14** (2022), 271–318.
11. A. Neumaier and B. Azmi. Line search and convergence in bound-constrained optimization. Unpublished manuscript, University of Vienna (2019). `http://www.optimization-online.org/DB_HTML/2019/03/7138.html`.
12. A. Neumaier, B. Azmi, and M. Kimiaei. An active set method for bound-constrained optimization. Manuscript, (2023). `https://optimization-online.org/?p=21354`.
13. A. Neumaier, M. Kimiaei. An improved of the Goldstein line search. Supplemental martial. (2023). `https://github.com/GS1400/CLS_supplement.git`.
14. A. Neumaier, M. Kimiaei, and B. Azmi. Globally linearly convergent nonlinear conjugate gradients without Wolfe line search. Manuscript, (2023). `https://optimization-online.org/?p=21354`.
15. J. Nocedal and S. Wright. *Numerical optimization*. Springer Science & Business Media (2006).
16. J. J. Moré, D. J. Thuente. Line search algorithms with guaranteed sufficient decrease. *(ACM) Trans. Math. Softw.* **20** (1994), 286–307.
17. C. W. Royer and S. J. Wright. Complexity analysis of second-order line-Search algorithms for smooth nonconvex optimization *SIAM J. Optim.* **28** (2018), 1448–1477.
18. W. Warth and J. Werner. Effiziente Schrittweitenfunktionen bei unrestringierten Optimierungsaufgaben. *Computing* **19** (1977), 59–72.
19. P. Wolfe. Convergence conditions for ascent methods. *SIAM Rev.* **11** (1969), 226–235.