A note on the optimal convergence rate of descent methods with fixed step sizes for smooth strongly convex functions

André Uschmajew* Bart Vandereycken†

Abstract

Based on a recent result by de Klerk, Glineur, and Taylor (*SIAM J. Optim.*, 30(3):2053–2082, 2020) on the attainable convergence rate of gradient descent for smooth and strongly convex functions in terms of function values, a convergence analysis for general descent methods with fixed step sizes is presented. It covers variable metric methods as well as gradient related search directions under angle and scaling conditions. An application to inexact gradient methods is also presented.

1 Introduction

An L-smooth and μ-strongly convex function $f\colon \mathbb{R}^n \to \mathbb{R}$ is characterized by the two properties

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x-y\|$$

and

$$f(x) \ge f(y) + \langle \nabla f(y), x-y\rangle + \frac{\mu}{2}\|x-y\|^2$$

for some constants $0 < \mu \le L$ and all $x, y \in \mathbb{R}^n$. Here, $\langle\ ,\ \rangle$ can be any inner product on $\mathbb{R}^n$ with corresponding norm $\|\cdot\|$, and ∇f denotes the gradient with respect to the this inner product. Note that the constants μ and L depend on the chosen inner product. The class of such functions plays a main role in the convergence theory of the gradient method and related descent methods for finding the unique global minimum x^* of a given f. The update rule of the gradient method is

$$x^+ = x - h\nabla f(x),$$

where $h > 0$ is a step size which may depend on the current point x. It is well known that the fixed step size

$$h = \frac{2}{L+\mu}$$

achieves the optimal error reduction

$$\|x^+ - x^*\|^2 \le \left(\frac{\kappa_f - 1}{\kappa_f + 1}\right)^2 \|x - x^*\|^2, \quad \kappa_f = \frac{L}{\mu}, \tag{1.1}$$

per step. We refer to [6] for details.

*Max Planck Institute for Mathematics in the Sciences, 04103 Leipzig, Germany
†Section of Mathematics, University of Geneva, 1211 Geneva, Switzerland

Building on a novel technique from [4], it has recently been shown by de Klerk, Glineur, and Taylor [3, Theorem 5.3 with $\varepsilon = 0$] that the same rate is also valid for the error in function value. Specifically, for any

$$0 \leq h \leq \frac{2}{L+\mu} \tag{1.2}$$

it holds that

$$f(x^+) - f(x^*) \leq (1 - h\mu)^2 (f(x) - f(x^*)). \tag{1.3}$$

The optimal choice in this estimate is again $h = 2/(L+\mu)$ and leads to

$$f(x^+) - f(x^*) \leq \left(\frac{\kappa_f - 1}{\kappa_f + 1}\right)^2 (f(x) - f(x^*)). \tag{1.4}$$

This estimate for one step of the method is highly nontrivial. Obviously, it implies the same rate for the gradient descent method with exact line search, which has been obtained earlier by the same authors in [2] and using similar techniques. Moreover, this rate is known to be optimal in the class of L-smooth and μ-strongly convex functions. In fact, it is already optimal for quadratic functions in that class; see, e.g., [2, Example 1.3].

Of course, in many applications the difference $f(x) - f(x^*)$ is a natural error measure by itself. For example, for strongly convex quadratic functions it is proportional to the square energy norm of the quadratic form. In general, for an L-smooth and μ-convex function we always have

$$\frac{\mu}{2}\|x - x^*\|^2 \leq f(x) - f(x^*) \leq \frac{L}{2}\|x - x^*\|^2,$$

which shows that both error measures will exhibit the same R-linear convergence rate. The novelty of the estimate (1.4) is that one also has an optimal Q-linear rate for the function values, both for fixed step sizes and exact line search. (We refer to [8] for the definitions of R- and Q-linear rate.) However, compared to (1.1) an estimate like (1.4) is "more intrinsic", because the chosen inner product in $\mathbb{R}^n$ enters only via the constants μ and L. In this short note, we illustrate this advantage by showing that (1.4) allows for a rather clean analysis of general variable metric methods, as well as gradient related methods subject to angle and scaling conditions. In addition, Theorem 3.2 below allows to improve in most cases upon a result in [2] for inexact gradient methods with fixed step sizes.

2 Variable metric method

We first consider the variable metric method. Here the update rule reads

$$x^+ = x - hA^{-1}\nabla f(x), \tag{2.1}$$

where A is a symmetric (with respect to the given inner product) and positive definite matrix.

Theorem 2.1. *Assume the eigenvalues of A are in the positive interval $[\lambda, \Lambda]$ and define*

$$\bar{h} = \frac{2}{L/\lambda + \mu/\Lambda}.$$

Then x^+ in (2.1) with $0 \leq h \leq \bar{h}$ satisfies

$$f(x^+) - f(x^*) \leq \left(1 - \frac{h\mu}{\Lambda}\right)^2 (f(x) - f(x^*)).$$

In particular, the step size $h = \bar{h}$ yields

$$f(x^+) - f(x^*) \leq \left(\frac{\kappa_{f,A} - 1}{\kappa_{f,A} + 1}\right)^2 (f(x) - f(x^*)), \quad \kappa_{f,A} = \frac{L}{\mu}\frac{\Lambda}{\lambda}. \tag{2.2}$$

Proof. The result is almost immediately obtained from (1.3) by noting that $\nabla_A f(x) = A^{-1}\nabla f(x)$ is the gradient of f with respect to the A-inner product $\langle x, y\rangle_A = \langle x, Ay\rangle$. We have

$$\langle \nabla_A f(x) - \nabla_A f(y), x - y\rangle_A \leq L\|x - y\|^2 \leq \frac{L}{\lambda}\|x - y\|_A^2$$

as well as

$$\langle \nabla_A f(x) - \nabla_A f(y), x - y\rangle_A \geq \mu\|x - y\|^2 \geq \frac{\mu}{\Lambda}\|x - y\|_A^2$$

for all x, y. These two conditions are equivalent to f being (L/λ)-smooth and (μ/Λ)-strongly convex in that A-inner product; see, e.g., [6, Theorems 2.1.5 & 2.1.9]. Thus in (1.2) and (1.3), we can replace μ with μ/Λ and L by L/λ, which is exactly the statement of the theorem. □

Observe that $\kappa_{f,A} = \kappa_f \cdot \kappa_A$ with $\kappa_A = \Lambda/\lambda \geq 1$ the condition number of A. The contraction factor in (2.2) will therefore always be worse than the original factor in (1.4), which corresponds to $A = I$. This might seem suboptimal since in Newton's method, and under additional regularity conditions, the contraction factor improves when choosing $A = \nabla^2 f(x)$. However, for the general class of methods (2.1), the result in Theorem 2.1 is optimal. This can already be seen for the function $f(x) = \frac{1}{2}\|x\|^2$, in which case (2.1) becomes the linear iteration $x^+ = (I - hA^{-1})x$. Its contraction factor as predicted by (2.2) is bounded by $(\kappa_A - 1)^2/(\kappa_A + 1)^2$, which is indeed a tight bound: As in [2, Example 1.3], take $A = \mathrm{diag}(\lambda, \ldots, \Lambda)$ and $x = (x_1, 0, \ldots, 0, x_n)$. Then an exact line search yields $x^+ = (\kappa_A - 1)/(\kappa_A + 1) \cdot (-x_1, 0, \ldots, 0, x_n)$, and clearly there cannot be a better contraction factor with fixed step size. Note that the step size $\bar{h}$ in Theorem 2.1 also leads to equality in (2.2) when x is an eigenvector corresponding to λ or Λ. For a less trivial example, consider $f(x) = \frac{1}{2}\langle x, A^{-1}x\rangle$. Then (2.1) becomes $x^+ = (I - hA^{-2})x$ and the same x from above now leads to a contraction with the factor $(\kappa_{A^2} - 1)^2/(\kappa_{A^2} + 1)^2$ where indeed $\kappa_{A^2} = \kappa_f\kappa_A$, as predicted by Theorem 2.1.

3 Gradient related methods

Next we provide error estimates for gradient related descent methods under angle and scaling conditions. Specifically, we consider the update rule

$$x^+ = x - hd, \tag{3.1}$$

where $-d$ is a descent direction, that is, d satisfies

$$\langle \nabla f(x), d\rangle = \cos\theta\|\nabla f(x)\|\|d\|, \quad \cos\theta > 0, \tag{3.2}$$

for some $\theta \in [0, \pi/2)$. For the case of exact line search, it has been shown in [2, Theorem 5.1] that

$$f(x^+) - f(x^*) \leq \left(\frac{\kappa_{f,\theta} - 1}{\kappa_{f,\theta} + 1}\right)^2 (f(x) - f(x^*)), \quad \kappa_{f,\theta} = \frac{L}{\mu}\left(\frac{1 + \sin\theta}{1 - \sin\theta}\right), \tag{3.3}$$

and that this Q-linear rate is optimal. For the case of quadratic functions this has been known before, see, e.g., [5]. We also mention the result of [1, Theorem 3.3], which identifies the rate in (3.3) as optimal R-linear rate for exact line search when f is twice continuously differentiable.

Here we aim to generalize this result to fixed step sizes. The extent to which this is possible depends on the available information about the quantities $\|\nabla f(x)\|$, $\|d\|$, and $\langle \nabla f(x), d\rangle$. The basic idea is to interpret (3.1) as a variable metric method in order to apply Theorem 2.1. For this we need to find a symmetric and positive definite matrix A satisfying

$$Ad = \nabla f(x)$$

and estimate its condition number. Such a matrix can be found explicitly using the following lemma, which originates from the SR1 update rule; see, e.g., [7].

Lemma 3.1. *Let $u, v \in \mathbb{R}^n$ such that $\|u\| = \|v\| = 1$ and $\langle u, v\rangle = \cos\theta$. Then the matrix*

$$B = \frac{1}{\alpha}\left(I - \frac{rr^*}{\langle r, u\rangle}\right), \quad r = u - \alpha v, \quad \alpha = \frac{1-\sin\theta}{\cos\theta} = \frac{\cos\theta}{1+\sin\theta}$$

is symmetric (for the given inner product), satisfies $Bu = v$ and has

$$\lambda_{\min}(B) = \frac{\cos\theta}{1+\sin\theta}, \quad \lambda_{\max}(B) = \frac{\cos\theta}{1-\sin\theta},$$

as its smallest and largest eigenvalues, respectively. Here, rr^* denotes the rank-one matrix satisfying $rr^*x = r\langle r, x\rangle$ for all $x \in \mathbb{R}^n$.

Proof. This is checked by a straightforward calculation. Obviously, the matrix $I - \frac{rr^*}{\langle r,u\rangle}$ equals the identity on the orthogonal complement of r. Its eigenvalue belonging to the eigenvector r is

$$1 - \frac{\|r\|^2}{\langle r, u\rangle} = 1 - \frac{1 - 2\alpha\cos\theta + \alpha^2}{1 - \alpha\cos\theta} = \frac{1 - \sin\theta - \alpha^2}{\sin\theta} = \alpha^2,$$

where one uses $1 - \alpha\cos\theta = \sin\theta$ and $\alpha^2 = (1-\sin\theta)/(1+\sin\theta)$. Therefore, the largest eigenvalue of B is $1/\alpha$ (with multiplicity $n-1$), and the smallest eigenvalue is α. □

With Lemma 3.1 and Theorem 2.1 at our disposal, we can state our main result.

Theorem 3.2. *Assume (3.2) and*

$$\|d\| = c\|\nabla f(x)\| \tag{3.4}$$

for some $c > 0$. Define

$$\bar{h} = \frac{2\cos\theta}{Lc(1+\sin\theta) + \mu c(1-\sin\theta)}.$$

Then x^+ in (3.1) with $0 \le h \le \bar{h}$ satisfies

$$f(x^+) - f(x^*) \le \left(1 - \frac{h\mu c(1-\sin\theta)}{\cos\theta}\right)^2 (f(x) - f(x^*)).$$

In particular, the step size $h = \bar{h}$ yields

$$f(x^+) - f(x^*) \le \left(\frac{\kappa_{f,\theta} - 1}{\kappa_{f,\theta} + 1}\right)^2 (f(x) - f(x^*)).$$

Proof. If $d = 0$, the assertion is trivially true. Let $d \neq 0$. By Lemma 3.1, there exists a symmetric and positive definite matrix of the form $A = \frac{\|\nabla f(x)\|}{\|d\|} B = \frac{1}{c} B$ such that $Ad = \nabla f(x)$ and

$$\lambda_{\min}(A) = \frac{1}{c}\left(\frac{\cos\theta}{1+\sin\theta}\right), \quad \lambda_{\max}(A) = \frac{1}{c}\left(\frac{\cos\theta}{1-\sin\theta}\right).$$

The assertion follows therefore directly from Theorem 2.1. □

Remark 3.3. The condition (3.4) can be replaced with equivalent conditions such as, e.g.,

$$\langle \nabla f(x), d\rangle = \sigma \|d\|^2$$

for some $\sigma > 0$. An equivalent version of Theorem 3.2 is obtained by observing that $\cos\theta = \sigma c$.

To achieve the optimal rate in Theorem 3.2, the exact values of θ and c need to be known in order to compute the optimal step size $\bar{h}$. In practice, this is almost never the case and only bounds are available. We therefore formulate another, more practical result under the following angle and scaling conditions: there exists $0 < c_1 \leq c_2$ and $\theta' \in [0, \pi/2)$ such that

$$\theta \leq \theta', \quad c_1 \|\nabla f(x)\| \leq \|d\| \leq c_2 \|\nabla f(x)\|. \tag{3.5}$$

Under these conditions, the eigenvalues of the matrix $A = \frac{\|\nabla f(x)\|}{\|d\|} B$ in the proof of Theorem 3.2 can be bounded as

$$\lambda_{\min}(A) \geq \frac{1}{c_2}\left(\frac{\cos\theta'}{1+\sin\theta'}\right), \quad \lambda_{\max}(A) \leq \frac{1}{c_1}\left(\frac{\cos\theta'}{1-\sin\theta'}\right),$$

since $\cos\theta/(1 \pm \sin\theta)$ is monotonically decreasing/increasing in $\theta \in [0, \pi/2)$. The following result is then again immediately obtained from Theorem 2.1.

Theorem 3.4. *Assume* (3.5) *and define*

$$\bar{h} = \frac{2\cos\theta'}{Lc_2(1+\sin\theta') + \mu c_1(1-\sin\theta')}.$$

Then x^+ *in* (3.1) *with* $0 \leq h \leq \bar{h}$ *satisfies*

$$f(x^+) - f(x^*) \leq \left(1 - \frac{h\mu c_1(1-\sin\theta')}{\cos\theta'}\right)^2 (f(x) - f(x^*)).$$

In particular, the step size $h = \bar{h}$ *yields*

$$f(x^+) - f(x^*) \leq \left(\frac{\kappa'-1}{\kappa'+1}\right)^2 (f(x) - f(x^*)), \quad \kappa' = \frac{L}{\mu}\frac{c_2}{c_1}\left(\frac{1+\sin\theta'}{1-\sin\theta'}\right).$$

We remark again that if $c_1 = c_2 = \|d\|/\|\nabla f(x)\|$ and $\theta' = \theta$ are known, the resulting statements from Theorem 3.4 coincide with those in Theorem 3.2.

As an application of Theorem 3.4, we discuss the case of an inexact gradient method, where instead of the angle and scaling conditions (3.5) it is only known that

$$\|d - \nabla f(x)\| \leq \varepsilon \|\nabla f(x)\|. \tag{3.6}$$

This important scenario is also considered in [3, Theorem 5.3] where for the particular step size $h = \frac{2\mu - \varepsilon(L+\mu)}{(1-\varepsilon)\mu(L+\mu)}$ the following rate

$$f(x^+) - f(x^*) \leq \left(\frac{\kappa_f - 1}{\kappa_f + 1} + \varepsilon\right)^2 (f(x) - f(x^*)) \tag{3.7}$$

is obtained under the condition that $\varepsilon \leq \frac{2\mu}{L+\mu} = 1 - \frac{\kappa_f - 1}{\kappa_f + 1}$. This imposes a considerable restriction on the size of ε if κ_f is large. On the other hand, using Theorem 3.4 (which is ultimately based on the special case (1.4) of (3.7)), one can obtain a result for all $\varepsilon \in [0, 1)$, which furthermore

in most (but not all) cases improves upon (3.7) in the range $\varepsilon \leq \frac{2\mu}{L+\mu}$. Indeed, note that (3.6) implies

$$\sin\theta \leq \varepsilon, \quad (1-\varepsilon)\|\nabla f(x)\| \leq \|d\| \leq (1+\varepsilon)\|\nabla f(x)\|.$$

Therefore, according to Theorem 3.4, using the step size

$$\bar{h}_\varepsilon = \frac{2\sqrt{1-\varepsilon^2}}{L(1+\varepsilon)^2 + \mu(1-\varepsilon)^2}$$

yields

$$f(x^+) - f(x^*) \leq \left(\frac{\kappa_f \left(\frac{1+\varepsilon}{1-\varepsilon}\right)^2 - 1}{\kappa_f \left(\frac{1+\varepsilon}{1-\varepsilon}\right)^2 + 1} \right)^2 (f(x) - f(x^*)) \tag{3.8}$$

for any $\varepsilon \in [0,1)$. An elementary calculation shows that this rate is strictly smaller than (3.7) if

$$\max(\sqrt{1-\rho^2} - \rho, 0) < \varepsilon < 1, \quad \rho = \frac{\kappa_f - 1}{\kappa_f + 1}.$$

In particular, if $\sqrt{1-\rho^2} - \rho \leq 0$, which is equivalent to

$$\kappa_f \geq (\sqrt{2}+1)^2 \approx 5.83,$$

then the rate (3.8) is better than (3.7) for all $\varepsilon \in (0,1)$. If $\kappa_f < (\sqrt{2}+1)^2$, then there is a range $0 < \varepsilon < \sqrt{1-\rho^2} - \rho$ (which ensures $\varepsilon \leq \frac{2\mu}{L+\mu} = 1-\rho$) where (3.7) is sharper than (3.8).

When $n \geq 2$, the ideal contraction factor using the model (3.6) would be $\left(\frac{\kappa_f \left(\frac{1-\varepsilon}{1+\varepsilon}\right) - 1}{\kappa_f \left(\frac{1-\varepsilon}{1+\varepsilon}\right) + 1} \right)^2$, which is neither achieved by (3.7) nor (3.8). This is natural since the information provided by (3.6) is not sufficient to design a fixed step size rule achieving this optimal rate. Of course, if additional information, such as, e.g., $\|d\| \leq \|\nabla f(x)\|$, is available, one can further improve (3.8) using Theorem 3.4. As shown in [2], the optimal rate is achieved when exact line search is used. This can also be derived from Theorem 3.4, since one can assume $c_1 = c_2$ then.

Remark 3.5. We conclude with a side remark. When just looking at the proofs of Theorems 3.2 or 3.4, it would be natural to ask if there exists a symmetric and positive definite matrix B (and thus A) with a smaller condition number than the one from Lemma 3.1. As for the SR1 update rule, when matrix $B = B_\alpha$ in the lemma is regarded as a function of $\alpha \neq 0$, then it is well known that the stated α is one of the minimizers for the condition number in the class of all positive definite B_α (another is $\cos\theta/(1-\sin\theta)$); see, e.g., [9]. And indeed, any B with a smaller condition number would lead to a faster rate in Theorem 3.2 (via Theorem 2.1), which is not possible since the rate is known to be optimal when exact line search is used. This reasoning therefore provides a (rather indirect) proof for the following general statement.

Theorem 3.6. *Let $u, v \in \mathbb{R}^n$ such that $\|u\| = \|v\| = 1$ and $\cos\theta = \langle u, v\rangle > 0$ with $\theta \in [0, \pi/2)$. Then $(1+\sin\theta)/(1-\sin\theta)$ is the minimum possible (spectral) condition number among all symmetric and positive definite matrices B satisfying $Bu = v$.*

While probably well known in the field, we did not find this fact explicitly stated in the literature.

References

[1] A. I. Cohen. Stepsize analysis for descent methods. *J. Optim. Theory Appl.*, 33(2):187–205, 1981.

[2] E. de Klerk, F. Glineur, and A. B. Taylor. On the worst-case complexity of the gradient method with exact line search for smooth strongly convex functions. *Optim. Lett.*, 11(7):1185–1199, 2017.

[3] E. de Klerk, F. Glineur, and A. B. Taylor. Worst-case convergence analysis of inexact gradient and Newton methods through semidefinite programming performance estimation. *SIAM J. Optim.*, 30(3):2053–2082, 2020.

[4] Y. Drori and M. Teboulle. Performance of first-order methods for smooth convex minimization: a novel approach. *Math. Program.*, 145(1-2, Ser. A):451–482, 2014.

[5] H. Munthe-Kaas. The convergence rate of inexact preconditioned steepest descent algorithm for solving linear systems. Technical Report NA-87-04, Stanford University, 1987.

[6] Y. Nesterov. *Introductory lectures on convex optimization*. Kluwer Academic Publishers, Boston, MA, 2004.

[7] J. Nocedal and S. J. Wright. *Numerical optimization*. Springer, New York, second edition, 2006.

[8] J. M. Ortega and W. C. Rheinboldt. *Iterative solution of nonlinear equations in several variables*. Academic Press, New York-London, 1970.

[9] H. Wolkowicz. Measures for symmetric rank-one updates. *Math. Oper. Res.*, 19(4):815–830, 1994.