# Max-Planck-Institut
# für Mathematik
# in den Naturwissenschaften
# Leipzig

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

(revised version: March 2022)

by

*André Uschmajew and Bart Vandereycken*

# 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 result by Taylor, Hendrickx, and Glineur (*J. Optim. Theory Appl.*, 178(2):455–476, 2018) on the attainable convergence rate of gradient descent for smooth and strongly convex functions in terms of function values, an elementary convergence analysis for general descent methods with fixed step sizes is presented. It covers general variable metric methods, gradient related search directions under angle and scaling conditions, as well as inexact gradient methods. In all cases, optimal rates are obtained.

## 1  Introduction

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

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

and

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

for some constants $0 < \mu \leq L$ 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 $\nabla f$ denotes the gradient with respect to this inner product. Note that the constants $\mu$ 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

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

per step, which inductively implies the convergence of the method to $x^*$. We refer to [6, Theorem 2.1.15] for details.

---

*Max Planck Institute for Mathematics in the Sciences, 04103 Leipzig, Germany.

†Section of Mathematics, University of Geneva, 1211 Geneva, Switzerland. This work was supported by the SNSF under research project 192129.

In a more general setting of proximal gradient methods, it has recently been shown by Taylor, Hendrickx, and Glineur [9, Theorem 3.3 with $h = 0$] that the same rate is also valid for the error in function value. Specifically, for any

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

it holds that

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

Moreover, for $\frac{2}{L+\mu} \le h < \frac{2}{L}$ one has $f(x^+) - f(x^*) \le (hL - 1)^2 (f(x) - f(x^*))$. This automatically follows from (1.2) and (1.3) by using a weaker strong convexity bound $0 < \mu' \le \mu$ satisfying $h = \frac{2}{L+\mu'}$ and noting that $1 - h\mu' = hL - 1$. The optimal choice in the estimates is $h = 2/(L + \mu)$ and leads to

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

This estimate for one step of the method is highly nontrivial. Obviously, it implies the same inequality for the gradient descent method with exact line search (when the left side is minimized over all $h$), which has been obtained earlier in [2]. Moreover, this estimate is known to be optimal in the class of $L$-smooth and $\mu$-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 squared energy norm of the quadratic form. In general, for an $L$-smooth and $\mu$-convex function we always have

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

which clearly shows that $f(x_\ell) - f(x^*) \to 0$ for an iterative method implies $\|x_\ell - x^*\| \to 0$ for $\ell \to \infty$. Moreover, 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 $\mu$ 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, in Theorem 4.2 below we show how (1.4) already implies the sharp rates for inexact gradient methods under relative error bounds with fixed step sizes, based on a suitable change of the metric, thereby improving and simplifying a similar result in [3].

## 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. It is well known that such an update step can also be interpreted as a gradient step with respect to a modified inner product. This leads to the following result that will be the basis for our further considerations.

**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 $0 \le h \le \bar{h}$ satisfies*

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

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

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

*Proof.* The result is 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

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

as well as

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

for all $x, y$. These two conditions are equivalent to $f$ being $(L/\lambda)$-smooth and $(\mu/\Lambda)$-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 $\mu$ with $\mu/\Lambda$ and $L$ by $L/\lambda$, which is exactly the statement of the theorem. $\qquad\square$

An alternative, and somewhat more direct proof of Theorem 2.1 that does not require changing the inner product, can be given by applying the result (1.3) directly to the function $g(y) = f(A^{-1/2}y)$ at $y = A^{1/2}x$.

Observe that $\kappa_{f,A} = \kappa_f \cdot \kappa_A$ with $\kappa_A = \Lambda/\lambda \ge 1$ 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 $A = \text{diag}(\lambda, \ldots, \Lambda)$ 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 $\lambda$ or $\Lambda$. 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 $\theta \in [0, \pi/2)$. This condition is very natural since it guarantees the convergence of (3.1); see, e.g., [7, Chapter 3.2]. In particular, 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 $\alpha$. $\square$

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 $d \ne 0$. 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

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

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 $\theta$ 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 of (3.2) under the following relaxed angle and scaling conditions: there exists $0 < c_1 \le c_2$ and $\theta' \in [0, \pi/2)$ such that

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

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

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

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 $0 \le h \le \bar{h}$ satisfies

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

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

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

5

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.

*Remark* 3.5. We conclude the section 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 $\alpha$ is one of the minimizers for the condition number in the class of all positive definite $B_\alpha$ (another is $\cos\theta/(1 - \sin\theta)$); see, e.g., [10]. 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. It is, of course, not very difficult to prove this result directly by an elementary calculation on $2 \times 2$ matrices.

# 4 Inexact gradient method

We now discuss the important case of an inexact gradient method, where instead of the angle and scaling conditions (3.5), it is assumed that

$$\|d - \nabla f(x)\| \leq \varepsilon \|\nabla f(x)\| \tag{4.1}$$

for some $\varepsilon \in [0, 1)$. This model is also considered in [2, 3, 4]. Our aim is again deriving convergence rates for a fixed step size rule from the variable metric approach. Since the matrix $A$ in the proof of Theorem 3.2 no longer provides the optimal rates in this case, we use a different construction.

**Lemma 4.1.** *Let $u, v \in \mathbb{R}^n$ such that $v \neq 0$ and $\|u - v\| < \|v\|$. There exists a positive definite matrix $A$ that satisfies $Au = v$ and has eigenvalues $\left(1 \pm \frac{\|u-v\|}{\|v\|}\right)^{-1}$.*

*Proof.* Define $A^{-1} = I + \frac{\|u-v\|}{\|v\|}Q$ with $Q = I - 2\frac{ww^*}{\|w\|^2}$ and $w = \frac{v}{\|v\|} - \frac{u-v}{\|u-v\|}$. Observe that $Q$ is the orthogonal reflection matrix that sends $\frac{v}{\|v\|}$ to $\frac{u-v}{\|u-v\|}$, which implies $A^{-1}v = u$. Since $Q$ is symmetric with eigenvalues $\pm 1$, the result follows. $\qquad \square$

Applying the lemma to $u = d$ and $v = \nabla f(x)$, the following theorem on the inexact gradient model (4.1) is an immediate consequence of Theorem 2.1.

**Theorem 4.2.** *Assume $\nabla f(x) \neq 0$ and (4.1) for some $\varepsilon \in [0, 1)$ and define*

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

*Then $x^+ = x - hd$ with $0 \leq h \leq \bar{h}$ satisfies*

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

$$\leq (1 - h\mu(1-\varepsilon))^2 (f(x) - f(x^*)).$$

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

$$f(x^+) - f(x^*) \le \left( \frac{\kappa_\varepsilon - 1}{\kappa_\varepsilon + 1} \right)^2 (f(x) - f(x^*)), \quad \kappa_\varepsilon = \frac{L}{\mu} \left( \frac{1 + \varepsilon}{1 - \varepsilon} \right). \tag{4.2}$$

The rate in (4.2) is optimal under the general assumption (4.1), in particular for quadratic $f$ and $d$ satisfying $\langle \nabla f(x), d \rangle = \cos \theta \|d\| \|\nabla f(x)\|$ with $\sin \theta = \varepsilon$. Trivially, for $f(x) = \frac{1}{2}\|x\|^2$ the estimate (4.2) is sharp for all $d$ satisfying (4.1).

The result in Theorem 4.2 is not new. In [4, Proposition 1.5] it has been shown that $\left( \frac{\kappa_\varepsilon - 1}{\kappa_\varepsilon + 1} \right)^2$ is an upper bound for the $R$-linear convergence rate of the inexact gradient method with fixed step size $\bar{h}$. According to [4, Remark 1.6], the estimate (4.2) per step is implicitly contained in the proof of [3, Theorem 5.3], which, however, is rather technical. In addition, the statement of [3, Theorem 5.3] itself covers the rate (4.2) only for a range $\varepsilon \in [0, \bar{\varepsilon}]$ with some $\bar{\varepsilon} < \frac{2\mu}{L+\mu}$. Our proof via Lemma 4.1 provides a simple alternative for obtaining the result for all $\varepsilon \in [0, 1)$ directly from the estimate (1.4) for the gradient method (which coincides with [3, Theorem 5.3] when $\varepsilon = 0$).

## 5   Conclusions

Based on the result (1.4) due to [9], we have derived optimal convergence rates for the function values in gradient related descent methods and inexact gradient methods with fixed step sizes for smooth and strongly convex functions. The results are obtained using an elementary variable metric approach, in which a single step is interpreted as a standard gradient step. This is possible since function values are a metric independent error measure. Compared to existing results, our proofs offer a more direct way for obtaining the convergence rate estimates of perturbed gradient methods given the rates of their exact counterpart.

## 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] O. Gannot. A frequency-domain analysis of inexact gradient methods. *Math. Program.*, 2021.

[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] A. B. Taylor, J. M. Hendrickx, and F. Glineur. Exact worst-case convergence rates of the proximal gradient method for composite convex minimization. *J. Optim. Theory Appl.*, 178(2):455–476, 2018.

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