Continuous-time Lower Bounds for Gradient-based Algorithms

Michael Muehlebach[1] **Michael I. Jordan**[1]

Abstract

This article derives lower bounds on the convergence rate of continuous-time gradient-based optimization algorithms. The algorithms are subjected to a time-normalization constraint that avoids a reparametrization of time in order to make the discussion of continuous-time convergence rates meaningful. We reduce the multi-dimensional problem to a single dimension, recover well-known lower bounds from the discrete-time setting, and provide insight into why these lower bounds occur. We present algorithms that achieve the proposed lower bounds, even when the function class under consideration includes certain nonconvex functions.

1. Introduction

Many problems in machine learning and statistics can be formulated as optimization problems. First-order optimization algorithms, such as gradient descent, are commonly used due to their simplicity and due to the fact that their complexity scales mildly in the number of decision variables. These algorithms are known to have natural limits on their convergence rate. We will examine these complexity lower bounds from a dynamical systems perspective. Unusually for the literature on lower bounds, we will work in continuous time. Our continuous-time perspective not only leads to insights into why complexity bounds arise, but also provides guidance for algorithm design.

A commonly used technique for deriving lower bounds is to construct a function that is difficult to optimize (see, e.g., Nesterov, 2004, p. 59). Such a function is typically obtained by ensuring that, at the jth iteration, all gradients that an algorithm has evaluated so far belong to a $(j+1)$-dimensional subspace which is far away from the optimum. Dimension-independent lower bounds then result from an unbounded increase in the problem dimension. This establishes, for example, that for the class of smooth and strongly convex functions at least $\sqrt{\kappa}/2\ \ln(c/\epsilon)$ iterations are needed to achieve an ϵ-distance to the optimizer (for large κ), where c is a constant and κ refers to the condition number (Nesterov, 2004, p. 68). The lower bound is achieved by accelerated gradient methods; for example, Scoy et al. (2018) provide an algorithm that attains the lower bound up to a factor of two. In the nonconvex setting, deriving tight lower bounds for smooth functions is an active area of research, where important recent contributions have been made by Carmon et al. (2019), for example.

We are motivated by a line of recent work that views optimization algorithms as continuous-time dynamical systems (see, e.g., Su et al., 2016; Krichene et al., 2015; Wibisono et al., 2016; Diakonikolas & Jordan, 2019; Muehlebach & Jordan, 2019). This work has provided significant insight into convergence rates of discrete-time algorithms via translating *upper bounds* from continuous time to discrete time. We ask the question whether it is also possible to obtain insight into *lower bounds* on gradient-based algorithms via a continuous-time analysis.

Instead of constructing a function that is difficult to optimize, we exploit invariance properties of the function class under consideration, which, combined with a local analysis about a (local) minimum, greatly simplifies the dynamics that need to be considered. This reduces the problem of determining the worst-case convergence rate over the given class of functions to the analysis of a parameter-dependent characteristic polynomial. We show that under certain circumstances the classical dimension-independent discrete-time lower bound for smooth and strongly convex functions can be recovered. We also derive continuous-time algorithms that achieve faster convergence rates. These algorithms include very fast dynamics and it is a matter of future research to investigate whether it is possible to derive practical discretizations of these dynamics.

A related—but discrete-time—perspective has been presented by Arjevani et al. (2016), where optimization algorithms are modeled by kth-order linear dynamical systems, and the complexity is shown to be lower bounded

[1] Division of Electrical Engineering and Computer Science, and Department of Statistics, University of California, Berkeley, Berkeley, USA. Correspondence to: Michael Muehlebach <michaelm@berkeley.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

by $\Omega(\kappa^{1/k} \ln(1/\epsilon))$.[1] In contrast, we model algorithms as continuous-time nonlinear dynamical systems and show how the lower bound $\Omega(\sqrt{\kappa} \ln(1/\epsilon))$ for the class of strongly convex quadratic functions can be recovered.

1.1. Notation and outline

We focus on optimizing real-valued functions $f : \mathbb{R}^n \rightarrow \mathbb{R}$, where $n > 0$ is an integer. Without loss of generality, we assume that the functions f have a local minimum at $x = 0$ with value $f(0) = 0$. The functions are assumed to have Lipschitz-continuous gradients. Our aim is to find a lower bound on the convergence rate that any continuous-time gradient-based algorithm can possibly achieve on a given class of functions. This class of functions will be denoted by $C_{\mu,L}$ and is required to satisfy the following assumptions:

(C1) Each $f \in C_{\mu,L}$ is twice continuously differentiable in a neighborhood of the origin.

(C2) For every $f \in C_{\mu,L}$, it holds that

$$\text{spec}(\Delta f(0)) \subset [\mu, L],$$

where $0 < \mu \leq L$ are fixed constants, spec denotes the spectrum, and $\Delta f(0)$ refers to the Hessian of the function f evaluated at the origin. Conversely, for any $\lambda_f \in [\mu, L]$, there exists a function $f \in C_{\mu,L}$ such that

$$\lambda_f \in \text{spec}(\Delta f(0)).$$

(C3) The class $C_{\mu,L}$ is invariant under orthogonal transformations. In other words, $f \in C_{\mu,L}$ implies that $f \circ T \in C_{\mu,L}$ for all $T \in O(n)$, where $O(n)$ denotes the set of orthogonal matrices of size $n \times n$.

Assumption (C1) imposes local smoothness and Assumption (C2) encodes prior information about the local curvature. It excludes degeneracies, which arise either due to a non-isolated minimum, or when the curvature about the minimum is arbitrarily small. Assumption (C3) implies that the function class $C_{\mu,L}$ is invariant under permutations and rotations. As we shall see in the sequel, this has important implications for algorithm design.

Given these assumptions, the lower bounds that we derive apply to smooth nonconvex functions with isolated nondegenerate critical points, convex quadratic functions, and smooth and strongly convex functions. The assumptions emphasize the importance of the local shape of the objective function about a local minimum. From a dynamical systems perspective imposing limits on the local instead of the global structure seems more natural. Even though our analysis includes results about certain nonconvex functions, we will not consider the impact of saddle points, for example, which greatly limits the convergence rate (Jin et al., 2019; Carmon et al., 2019).

The complexity of an algorithm can be characterized by the number of iterations required to achieve an ϵ-distance to the optimizer. In the following, it will be more convenient to characterize the convergence rate. We say that an algorithm converges with rate $\rho > 0$, if the distance to the local optimum decays at least with $\exp(-\rho t)$ for a certain set of initial conditions, where t refers to time. Both notions are equivalent. However, an upper bound on the convergence rate leads to a lower bound on the complexity, and vice versa. For example, if the convergence rate is upper bounded by $\mathcal{O}(1/\kappa)$, the complexity is lower bounded by $\Omega(\kappa \ln(1/\epsilon))$.

The article is structured as follows: Sec. 2 introduces the class of algorithms that are studied. The resulting lower bounds are presented in Sec. 3 and simulation results are provided in Sec. 4. The article concludes with a brief discussion in Sec. 5.

2. Continuous-time Gradient-based Optimization Algorithms

We model a gradient-based optimization algorithm as a dynamical system of the form

$$x^{(k)}(t) = g(x^{(k-1)}(t), \dots, \dot{x}(t), x(t), \\ \nabla f[h(x(t), \dot{x}(t), \dots, x^{(k-1)}(t))\,]\,), \quad (1)$$

where the functions $g : \mathbb{R}^{n\times k} \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ and $h : \mathbb{R}^{n\times k} \rightarrow \mathbb{R}^n$ are independent of f, where $k > 0$ is an integer, and where $x^{(p)}(t)$ denotes the pth derivative with respect to time. We say that (g, h) is a continuous-time gradient-based optimization algorithm, $(g, h) \in \mathcal{G}$, if (1) has the following properties (for all $f \in C_{\mu,L}$):

(G1) g and h are continuously differentiable in all arguments,

(G2) critical points of f correspond to equilibria of (1),

(G3) local minima of f correspond to asymptotically stable equilibria of (1) (in the sense of Lyapunov).

Assumption (G1) implies that (1) is a well-posed differential equation. Assumption (G2) and (G3) ensures that the algorithm locally converges to local minima. Assumption (G1)-(G3) are therefore minimal requirements for ensuring that (1) minimizes f. A graphical representation of the system (1) is shown in Fig. 1.

Remark: We model a gradient-based optimization algorithm as an autonomous dynamical system. On a fundamental level, introducing non-autonomous dynamics would lead to

[1] Here, kth order refers to the fact that the jth iterate depends only on the past k iterates. This should not be confused with accessing higher derivatives of the objective function.

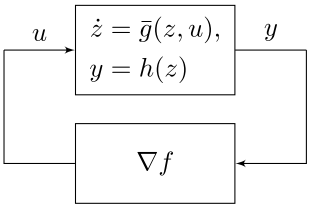

Figure 1. Graphical representation of the dynamics (1), where z denotes the internal state, and $\bar g$ is related to g. The feedback connection with the gradient ∇f can be viewed as an oracle query. The assumptions on the structure of the dynamical system (as given by (1)) are without loss of generality.

a time-varying vector field, which contrasts with the fact that the objective function f is fixed. In physics and engineering, non-autonomous dynamical systems typically arise in situations where only a sub-component of a system is studied. In that case, the interactions of the sub-component with the rest might lead to non-autonomous dynamics, even though the system as a whole is autonomous. Thus, non-autonomous dynamics, as introduced in Su et al. (2016), for example, might be useful to approximate (1) with a reduced-order model, whereby higher-order dynamics are captured by time-varying terms. In the following, we will characterize fundamental limits on the convergence rate for any integer $k > 0$; therefore there is no need to render g and h time varying.

2.1. Rescaling time

In contrast to the discrete-time setting, the time t has no absolute meaning in continuous time. The trajectory $\tilde x(t) := x(\alpha(t))$, where $\alpha : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is any diffeomorphism satisfies a differential equation similar to (1).

For $k = 2$, for example, the trajectory $\tilde x(t)$ evolves according to

$$\ddot{\tilde x} = g(\dot{\tilde x}/\dot\alpha, \tilde x, \nabla f[h(\tilde x, \dot{\tilde x}/\dot\alpha)])\dot\alpha^2 + \dot{\tilde x}\frac{\ddot\alpha}{\dot\alpha}, \qquad (2)$$

where the dependence on time has been omitted. In order for a discussion of convergence rates to be meaningful, such a rescaling needs to be avoided. This is done by adding the following requirement: $(g,h) \in \mathcal{G}'$,

$$\mathcal{G}' := \left\{ (g,h) \in \mathcal{G} \mid \det\left(\frac{\partial g(w,v)}{\partial v}\bigg|_{0,0} \frac{\partial h}{\partial x}\bigg|_{0} \right) = (-1/L)^n \right\}, \qquad (3)$$

where the variables $w \in \mathbb{R}^{n\times k}$ and $v \in \mathbb{R}^n$ are placeholders, and det denotes the determinant. In the example of $k = 2$, the right-hand-side of (2) satisfies the time-normalization constraint (3) if and only if

$$\left(\frac{-1}{L}\right)^n = \det\left(\frac{\partial g(w,v)}{\partial v}\bigg|_{0,0} \frac{\partial h}{\partial x}\bigg|_{0} \right) \dot\alpha(t)^{2n} = \left(\frac{-\dot\alpha(t)^2}{L}\right)^n,$$

where $(g,h) \in \mathcal{G}'$ has been used for the last equality. This implies $\dot\alpha(t)^2 = 1$ for all $t \in \mathbb{R}_{\geq 0}$, or equivalently, $\alpha(t) = t + \text{const}$. The same argument applies for $k = 1$ or $k > 2$ and implies that (3) fixes the time scale.

Choosing any other normalization in (3), such as the trace, the induced two-norm, a specific entry, or a constant different than $1/L$, fixes the time scale in a different way; however, the results derived in the remainder of the paper still apply. The normalization according to (3) is convenient, since, as a result, a convergence rate of unity is achieved for the class $C_{L,L}$, which consists of functions that behave locally like $L|x|^2/2$. The sign is imposed by the asymptotic stability requirement (G3). Additional context on the normalization (3) is provided in App. A.

2.2. First-order approximation

For deriving the lower bounds on the convergence rate it will be enough to consider initial conditions that are close to the origin. We therefore apply Taylor's theorem to the dynamics (1),

$$x^{(k)} = -\sum_{j=1}^{k-1} G_j x^{(j)} - G_0 \Delta f(0) \left(x + \sum_{j=1}^{k-1} H_j x^{(j)} \right) + r(x, \dot x, \dots, x^{(k-1)}), \qquad (4)$$

where the dependence on time is omitted, and the matrices $G_0, \dots, G_{k-1} \in \mathbb{R}^{n\times n}$ and $H_1, \dots H_{k-1} \in \mathbb{R}^{n\times n}$ represent the different partial derivatives of g and h evaluated at the origin. The remainder term is denoted by the function $r : \mathbb{R}^{n\times k} \to \mathbb{R}^n$, and captures the second-order terms. In deriving (4), we exploited the fact that the partial derivative of g with respect to x vanishes, when evaluated at the origin, due to Assumption (G2). Furthermore, we set the partial derivative of h with respect to x, evaluated at the origin, to the identity. This amounts to a normalization of the state x, which, according to Assumption (G2) can always be done and does not affect the convergence rate.

By introducing the state variable $z := (x, \dot x, \dots, x^{(k-1)})$, the dynamics (4) can be rewritten as

$$\dot z(t) = Az(t) + \tilde r(z(t)), \qquad (5)$$

where $A \in \mathbb{R}^{kn\times kn}$ is obtained by appropriately stacking the matrices $G_0, \dots, G_{k-1}$, $H_1, \dots, H_{k-1}$ and $\tilde r$ captures the remainder term. The following lemma relates the convergence rate of the nonlinear dynamics to the convergence rate of the linear dynamics.

Lemma 2.1 *Assume that there exists a neighborhood N of the origin such that any solution $z(t)$ of (5) with $z(0) \in N$ satisfies*

$$|z(t)| \leq c_1 |z(0)| \exp(-a_r t), \quad \forall t \in [0, \infty),$$

where $c_1 \geq 1$ and $a_r > 0$ are constants. Then, there exists a constant $c_2 \geq 1$, such that any solution of the corresponding linear equation $\Delta\dot{z}(t) = A\Delta z(t)$ satisfies the estimate

$$|\Delta z(t)| \leq c_2 |\Delta z(0)| \exp(-a_r t).$$

The lemma is a standard result from the theory of ordinary differential equations. We included a proof in App. B. A lower bound on the convergence rate of the linearized dynamics provides us therefore with a lower bound on the convergence of the nonlinear dynamics, as this would otherwise contradict the statement of Lemma 2.1. In order to find lower bounds on the convergence rate, we will therefore replace (1) with its first-order approximation (4), where we neglect the remainder term r.

2.3. Invariance under orthogonal transformations

Assumption (C3) requires that the class $C_{\mu,L}$ is invariant under orthogonal transformations. We will show next that, without loss of generality, the linearized dynamics (4) (when r is neglected) can be assumed to be invariant under orthogonal transformations, which will simplify our derivations.

Proposition 2.2 *Assume* (4) *converges with rate ρ (or faster) for all $f \in C_{\mu,L}$. Then, the rate ρ can be achieved for $G_0 = g_0 I, \dots, G_{k-1} = g_{k-1} I$, $H_1 = h_1 I, \dots, H_{k-1} = h_{k-1} I$, where $g_0, \dots, g_{k-1}$ and $h_1, \dots, h_{k-1}$ are scalars, and $I \in \mathbb{R}^{n\times n}$ is the identity.*

Proof The convergence rate of (4) (with $r = 0$) is determined by the roots of the characteristic polynomial

$$\det\left(s^k I + \sum_{j=1}^{k-1} G_j s^j + G_0 \Delta f(0)\left(I + \sum_{j=1}^{k-1} H_j s^j\right)\right).$$

We set $\Delta f(0) = \lambda I$ and analyze the characteristic polynomial for different $\lambda \in [\mu, L]$. The polynomial can be factorized into n factors, each having the form

$$s^k + \tilde{g}_{k-1}(\lambda) s^{k-1} + \dots + \tilde{g}_1(\lambda) s + \lambda \tilde{g}_0, \qquad (6)$$

where the coefficients $\tilde{g}_j(\lambda)$ continuously depend on λ and $\tilde{g}_0$ is given by an eigenvalue of G_0. The latter follows from evaluating the above determinant at $s = 0$. By assumption, each of these factors must have roots with real parts smaller than $-\rho$ for all $\lambda \in [\mu, L]$. This implies, by Kharitonov's theorem (Minnichelli et al., 1989), that the four Kharitonov polynomials, which are given by different combinations of the maxima and minima of $\tilde{g}_j(\lambda)$ over $\lambda \in [\mu, L]$, have real parts smaller than $-\rho$. We pick any of the n factors, (6), and choose $g_0, \dots, g_{k-1}$ and $h_1, \dots, h_{k-1}$ such that $g_0 = \tilde{g}_0$,

$$g_j + \mu h_j = \min_{\lambda\in[\mu,L]} \tilde{g}_j(\lambda), \quad g_j + L h_j = \max_{\lambda\in[\mu,L]} \tilde{g}_j(\lambda),$$

$j = 1, \dots, k-1$. In that case, for any $f \in C_{\mu,L}$, the characteristic polynomial factorizes into n equal factors, each having the form

$$s^k + (g_{k-1} + h_{k-1}\bar{\lambda}) s^{k-1} + \dots + \bar{\lambda} g_0, \qquad (7)$$

where $\bar{\lambda} \in [\mu, L]$ is an eigenvalue of $\Delta f(0)$. By construction, the Kharitonov polynomials of (6) and (7) agree, which guarantees a convergence rate of at least ρ. ∎

The fact that a square matrix is invariant under orthogonal transformations if and only if it is a scaled identity matrix implies that the dynamics (4) are invariant under orthogonal transformations of f if and only if the matrices $G_0, \dots, H_{k-1}$ are scaled identity matrices. A game-theoretic interpretation of Prop. 2.2 is included in App. C.

We therefore conclude that for any $(g, h) \in \mathcal{G}'$, a lower bound on the convergence rate is obtained by the first-order approximation of (1), which, due to the time-normalization constraint and the invariance under orthogonal transformations, takes the form

$$x^{(k)} = -\sum_{j=1}^{k-1} g_j x^{(j)} - \frac{1}{L}\Delta f(0)\Big(x + \sum_{j=1}^{k-1} h_j x^{(j)}\Big), \qquad (8)$$

where $g_1, \dots, g_{k-1}$, $h_1, \dots, h_{k-1} \in \mathbb{R}$ are scalars and the dependence on time has been omitted. The time-normalization constraint (3) implies $g_0 = 1/L$.

3. Lower Bounds on the Convergence Rate

3.1. Gradient flow ($k = 1$)

This section analyses the case $k = 1$, resulting in gradient-flow algorithms. According to (8), we obtain the first-order approximation

$$\dot{x}(t) = -\frac{1}{L}\Delta f(0) x(t). \qquad (9)$$

By choosing an appropriate coordinate system that diagonalizes $\Delta f(0)$, we conclude that each component x_j of x converges according to

$$x_j(t) = x_j(0)\exp(-\lambda_f t), \qquad (10)$$

where λ_f is an eigenvalue of $\Delta f(0)/L$. Due to the fact that $f \in C_{\mu,L}$, which implies $1/\kappa \leq \lambda_f \leq 1$, according to (C2), $x(t)$ converges in the worst-case with $c\exp(-t/\kappa)$, where $\kappa := L/\mu$ and c is constant. Hence, by virtue of Lemma 2.1 we conclude:

The convergence rate of any continuous-time optimization algorithm $(g, h) \in \mathcal{G}'$ (with $k = 1$) is upper bounded by $1/\kappa$ on functions $C_{\mu,L}$. Thus, the time required to achieve an ϵ-distance to the optimizer is lower bounded by κ ln(c/ϵ) on functions $C_{\mu,L}$, where $c > 0$ is constant.

3.2. Accelerated gradient flow ($k = 2$)

This section discusses the case $k = 2$. According to (8), the following first-order approximation is obtained:

$$\ddot{x}(t) = -\left(g_1 + \frac{h_1}{L}\Delta f(0)\right)\dot{x}(t) - \frac{1}{L}\Delta f(0)x(t). \quad (11)$$

Choosing an appropriate coordinate system that diagonalizes $\Delta f(0)$, results in the scalar second-order differential equation for each component x_j of x,

$$\ddot{x}_j(t) = -(g_1 + h_1\lambda_f)\dot{x}_j(t) - \lambda_f x_j(t), \quad (12)$$

where λ_f is an eigenvalue of $\Delta f(0)/L$. The convergence rate of a single component of x is therefore dictated by the real part of the roots of the following polynomial:

$$s^2 + (g_1 + h_1\lambda_f)s + \lambda_f = 0. \quad (13)$$

The roots are given by

$$-d_{\lambda_f}(g_1, h_1) \pm \sqrt{d_{\lambda_f}(g_1, h_1)^2 - \lambda_f}, \quad (14)$$

with $d_{\lambda_f}(g_1, h_1) := (g_1 + \lambda_f h_1)/2$. Due to the fact that the square-root term is either positive or imaginary, the convergence rate is limited by the first root (with the $+$ sign). In addition, $f \in C_{\mu,L}$, hence λ_f may vary between $1/\kappa$ and 1. As a result, the convergence rate is lower bounded by

$$\min_{g_1,h_1} \max_{\lambda_f \in [\frac{1}{\kappa},1]} \operatorname{Re}\left(-d_{\lambda_f} + \sqrt{d_{\lambda_f}^2 - \lambda_f}\right), \quad (15)$$

where the dependence of d_{λ_f} on g_1 and h_1 has been omitted. For a fixed $\lambda_f > 0$ the function achieves its minimum value $-\sqrt{\lambda_f}$ when g_1 and h_1 are chosen such that $d_{\lambda_f}(g_1, h_1) = \sqrt{\lambda_f}$. Thus, interchanging the min and the max concludes that (15) is lower bounded by $-1/\sqrt{\kappa}$. Choosing $g_1 = 2/\sqrt{\kappa}$, $h_1 = 0$, reveals that the lower bound is actually attained.[1] This implies

$$\min_{g_1,h_1} \max_{\lambda_f \in [\frac{1}{\kappa},1]} \operatorname{Re}\left(-d_{\lambda_f} + \sqrt{d_{\lambda_f}^2 - \lambda_f}\right) = -1/\sqrt{\kappa},$$

which, by virtue of Lemma 2.1, implies:

The convergence rate of any continuous-time optimization algorithm $(g, h) \in \mathcal{G}'$ (with $k = 2$) is upper bounded by $1/\sqrt{\kappa}$ on functions $C_{\mu,L}$. Thus, the time required to achieve an ϵ-distance to the optimizer is lower bounded by $\sqrt{\kappa}\, ln(c/\epsilon)$ on functions $C_{\mu,L}$, where $c > 0$ is constant.

3.3. Higher-order methods ($k > 2$)

We follow the reasoning of the previous sections and obtain the characteristic polynomial

$$s^k + (g_{k-1} + h_{k-1}\lambda_f)s^{k-1} + \dots + (g_1 + h_1\lambda_f)s + \lambda_f = 0, \quad (16)$$

whose roots determine the convergence rate of a single component of $x(t)$, where $x(t)$ satisfies (8) and $\lambda_f \in [1/\kappa, 1]$ is an eigenvalue of $\Delta f(0)/L$. Expressing (16) in terms of its roots $-\pi_1, \dots, -\pi_k \in \mathbb{C}$, where $\mathbb{C}$ denotes the set of complex numbers, results in

$$(s + \pi_1)(s + \pi_2)\dots(s + \pi_k) = 0. \quad (17)$$

The minus sign is introduced for notational convenience. Equating the coefficients of (16) and (17) yields

$$\lambda_f = \pi_1\pi_2\dots\pi_k. \quad (18)$$

In other words, no matter how the coefficient g_j and h_j are chosen, the product of the roots of (16) is always equal to λ_f. Due to the fact that the coefficients of the polynomial (16) are real, the roots $\pi_1, \dots, \pi_k$ are complex conjugated. As a consequence, the previous equation simplifies to

$$\lambda_f = |\pi_1|\,|\pi_2|\dots|\pi_k| \geq |\pi_{\text{min}}|^k, \quad (19)$$

where π_{min} denotes the the root with the smallest absolute value. Evaluating (19) for $\lambda_f = 1/\kappa$ therefore yields the upper bound $1/\kappa^{1/k}$ on the absolute value of the smallest root, which suggests that the convergence rate is limited by $1/\kappa^{1/k}$. We will show next that such a convergence rate can, in fact, be achieved.

Proposition 3.1 *The convergence rate of any continuous-time optimization algorithm $(g, h) \in \mathcal{G}'$ with $k \geq 1$ is upper bounded by $1/\kappa^{1/k}$ on functions $C_{\mu,L}$. The algorithm given by (8) with*

$$h_j = \kappa^{j/k}\binom{k-1}{j}, \quad g_j = \frac{1}{\kappa^{(k-j)/k}}\binom{k}{j} - \frac{h_j}{\kappa}, \quad (20)$$

locally achieves the upper bound.

Proof The previous discussion implies that $1/\kappa^{1/k}$ is indeed a lower bound. In case g_j and h_j are chosen according to (20), the characteristic polynomial (16) simplifies to

$$(s + 1/\kappa^{1/k})^k + \bar{\lambda}_f(s\kappa^{1/k} + 1)^{k-1} = 0,$$

where $\bar{\lambda}_f := \lambda_f - 1/\kappa$. This can be further factorized to

$$(s + 1/\kappa^{1/k})^{k-1}(s + 1/\kappa^{1/k} + \bar{\lambda}_f\kappa^{(k-1)/k}) = 0,$$

which shows that there are $k - 1$ real roots at $-1/\kappa^{1/k}$ and one real root at

$$-\frac{1}{\kappa^{1/k}} - \bar{\lambda}_f\kappa^{(k-1)/k}$$

that depends on $\bar{\lambda}_f$. For $\bar{\lambda}_f \in [0, 1 - 1/\kappa]$, this root takes values in $[-\kappa^{1-1/k}, -1/\kappa^{1/k}]$. ∎

[1] In that case, the real part of (14) evaluates to $-1/\sqrt{\kappa}$, whereas changing $\lambda_f \in [1/\kappa, 1]$ only affects the imaginary part.

The example above locally achieves a convergence rate of $1/\kappa^{1/k}$, but introduces a single real root that tends to $-\infty$ for large κ, when $\lambda_f > 1/\kappa$. Such a fast root poses a problem for any explicit discretization scheme (Nevanlinna & Sipilä, 1974). We show that the convergence rate of any continuous-time optimization algorithm cannot exceed $\mathcal{O}(1/\sqrt{\kappa})$ when all roots of (16) are required to remain bounded for all κ.

Proposition 3.2 *The convergence rate of any continuous-time optimization algorithm $(g, h) \in \mathcal{G}'$ ($k \geq 2$), whose characteristic polynomial* (16) *has bounded roots, cannot exceed $\mathcal{O}(1/\sqrt{\kappa})$.*

Proof We consider the case where the roots are constrained to the unit disk (unit bound); the same arguments also apply for a bound greater than one. For the subsequent analysis it will be beneficial to rescale g_j and λ_f by introducing $\bar{g}_j := g_j + h_j/\kappa$, $\bar{\lambda}_f = \lambda_f - 1/\kappa$ such that (16) takes the form

$$\begin{aligned} s^k + \bar{g}_{k-1}s^{k-1} + \dots + \bar{g}_1 s + 1/\kappa \\ + \bar{\lambda}_f\left(h_{k-1}s^{k-1} + \dots + h_1 s + 1\right) = 0, \end{aligned} \tag{21}$$

with $\bar{\lambda}_f \in [0, 1-1/\kappa]$. In order to study the dependence of the roots of (21) on $\bar{\lambda}_f$ we use the Nyquist criterion (see App. D), which provides a necessary and sufficient condition for (21) to have all roots in the left-half complex plane for all $\bar{\lambda}_f \in [0, 1-1/\kappa]$. The Nyquist criterion implies that if the roots of (21) are all in the left-half complex plane, there is no $\bar{\lambda}_f \in [0, 1-1/\kappa]$, such that the graph of the complex function $P: \mathbb{R} \to \mathbb{C}$,

$$P(\omega) := \bar{\lambda}_f \left.\frac{h_{k-1}s^{k-1} + \dots + h_1 s + 1}{s^k + \bar{g}_{k-1}s^{k-1} + \dots + \bar{g}_1 s + 1/\kappa}\right|_{s=i\omega}, \tag{22}$$

passes through the point -1, where $i := \sqrt{-1}$ denotes the imaginary unit. We will show that this condition cannot be fulfilled when the roots π_j of (21) are required to satisfy $|\pi_j| \in (1/\sqrt{\kappa}, 1]$ for all $\kappa \geq 1$. To that end, the function $P(\omega)$ is rewritten as

$$\frac{\bar{\lambda}_f}{1-1/\kappa}\left(\underbrace{\frac{(s+\bar{\pi}_1)\dots(s+\bar{\pi}_k)}{(s+\underline{\pi}_1)\dots(s+\underline{\pi}_k)}}_{:=H(s)} - 1\right)\Bigg|_{s=i\omega}, \tag{23}$$

where $\underline{\pi}_j$ are the roots of (21) for $\bar{\lambda}_f = 0$ and $\bar{\pi}_j$ are the roots of (21) for $\bar{\lambda}_f = 1 - 1/\kappa$. These therefore satisfy (cf. (19))

$$\frac{1}{\kappa} = |\underline{\pi}_1| \dots |\underline{\pi}_k|, \quad 1 = |\bar{\pi}_1| \dots |\bar{\pi}_k|. \tag{24}$$

Combined with the requirement $|\pi_j| \leq 1$, the latter implies $|\bar{\pi}_j| = 1$ for $j = 1, 2, \dots, k$, whereas the former implies that at least three roots, denoted by $\underline{\pi}_1, \underline{\pi}_2, \underline{\pi}_3$ tend to zero for $\kappa \to \infty$, since $|\underline{\pi}_j| > \mathcal{O}(1/\sqrt{\kappa})$. Next, we analyze the graph of $P(\omega)$, and start by considering the numerator and denominator of $H(i\omega)$ separately. The numerator of $H(i\omega)$ takes the value 1 for $\omega = 0$, is continuous in ω, and its graph spirals outwards for $\omega \geq 0$ (since $\text{Re}(\bar{\pi}_j) > 0$, the phase strictly increases for $\omega \geq 0$ until it reaches $90 \cdot k$ degrees). All roots $\bar{\pi}_j$ satisfy $|\bar{\pi}_j| = 1$. Hence, there exists a value $\omega_{\text{num}} > 0$ such that the numerator of $H(i\omega)$ has a phase below, say, 10 degrees for all $\omega \in [0, \omega_{\text{num}}]$, and all $\kappa \geq 1$. The denominator of $H(i\omega)$ is likewise continuous in ω, takes the value $1/\kappa$ for $\omega = 0$, and spirals outwards for $\omega \geq 0$. However, since at least three of the roots $\underline{\pi}_j$ tend to zero for $\kappa \to \infty$, the denominator approaches (for large κ)

$$s^3(s+\underline{\pi}_4)\dots(s+\underline{\pi}_k)|_{s=i\omega},$$

which has a phase of more than 270 degrees for small ω (due to the s^3 term). Thus, there exists a small value $0 < \omega_{\text{denum}} \leq \omega_{\text{num}}$, such that for sufficiently large κ, the denominator of $H(i\omega)$ reaches a phase of more than 190 degrees for $\omega \in [0, \omega_{\text{denum}}]$. The situation is illustrated in Fig. 2. The phase of $H(i\omega)$ is given by the difference between the phase of the numerator and the denominator, and therefore, for sufficiently large κ, there exists the value $\omega_{\text{c}} \in [0, \omega_{\text{denum}}]$, such that $H(i\omega_{\text{c}})$ has a phase of -180 degrees. Hence, according to (23), $P(i\omega_{\text{c}})$ has likewise a phase of -180 degrees and $P(i\omega_{\text{c}}) < -1$. Thus, there exists a $\bar{\lambda}_f$ such that $P(i\omega_{\text{c}}) = -1$, contradicting the condition established by the Nyquist criterion. ∎

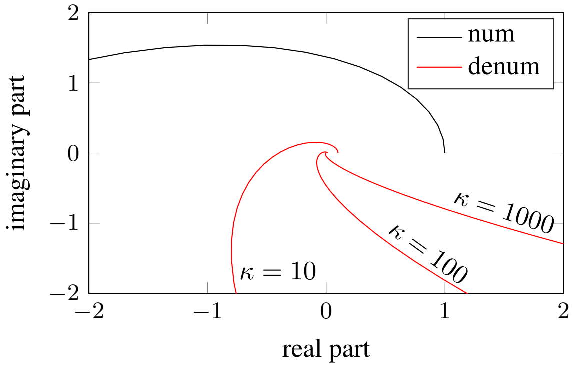

Figure 2. The graph illustrates the behavior of the numerator and denumerator of $H(i\omega)$ for $\omega \geq 0$ on the example $(i\omega+1)^4/((i\omega + 1/\kappa^{0.25})^4$. Both numerator and denumerator spiral outwards and their phase approaches 360 degrees for large ω. While the numerator (black) has a phase close to 0 for small ω, the phase of the denumerator (red) approaches 360 degrees for large κ.

3.4. Discussion

The analysis provides insights into the convergence limits of any gradient-based algorithm. It emphasizes the fact that these convergence limits result from limited curvature information, since by Assumption (C2), only upper and lower bounds on $\Delta f(0)$ are known (i.e. $\lambda_f \in [1/\kappa, 1]$). Compared to the discrete-time case, where the convergence rate is upper bounded by $\mathcal{O}(1/\sqrt{\kappa})$, faster convergence rates can be achieved with continuous-time algorithms. These algorithms, however, necessarily include arbitrarily fast converging dynamics as shown by Prop. 3.2 and cannot be discretized by explicit linear methods (Nevanlinna & Sipilä, 1974). This recovers the classical discrete-time result. Even though the worst-case complexity in discrete-time cannot be improved, discretizing the dynamics (20) with variable step-size schemes, or integration methods whose order increases with κ might still lead to new discrete-time algorithms that achieve fast convergence rates. Additional background on the relation to discrete time is included in App. E.

The proof of Prop. 3.2 provides a graphical interpretation of the fact that a convergence rate of $\mathcal{O}(1/\sqrt{\kappa})$ cannot be improved (under the assumption of bounded rates): A faster convergence rate requires that at least three roots of (16) become arbitrarily small for large κ. These three roots introduce a negative phase shift of more than 180 degrees (for $\kappa \to \infty$ each introduces -90 degrees) in the function $P(i\omega)$, cf. (22), which leads to instability for large κ. Clearly, these issues do not occur for low-order dynamics with $k \leq 2$, as in that case, the phase shift is limited from the outset to -180 degrees.

One might suspect that the algorithm provided in Prop. 3.2 is fragile, in the sense that it only achieves the rate of $1/\kappa^{1/k}$ on quadratic functions. The following proposition shows that this is not the case; the algorithm converges and achieves the rate $1/\kappa^{1/k}$ for all smooth and strongly convex functions and even for certain nonconvex functions.

Proposition 3.3 *Let $f \in C_{\mu,L}$ be such that*

$$(\nabla f(x)/L - \alpha_s x)^\mathsf{T}(\nabla f(x)/L - x) \leq 0, \qquad (25)$$

holds for all $x \in \mathbb{R}^n$, where $\alpha_s > 0$ is constant. Then, the origin is a globally asymptotically stable equilibrium for the dynamical system

$$\begin{aligned} x^{(k)}(t) &= -g_{k-1}x^{(k-1)}(t) - \dots - g_1\dot{x}(t) \\ &- \nabla f(x(t) + h_1\dot{x}(t) + \dots + h_{k-1}x^{(k-1)}(t))/L, \end{aligned} \qquad (26)$$

where g_j and h_j are defined in Prop. 3.1. The trajectories converge with rate $1/\kappa^{1/k}$ if (25) holds for $\alpha_s = 1/\kappa$.

The result can be proved by rewriting the dynamics as a Luré problem with $\nabla f(x)/L - x/\kappa$ as the nonlinear feedback term and applying the circle criterion (Khalil, 1996, p. 265). A full proof can be found in App. F. The assumption (25) ensures that $\nabla f(x)/L$ belongs to the sector $[\alpha_\mathrm{s}, 1]$, as illustrated on a scalar example in Fig. 3 (top left). The condition $\alpha_\mathrm{s} = 1/\kappa$ is fulfilled by smooth and strongly convex functions, which shows that the lower bound suggested by Prop. 3.1 is achieved. However, (25) does not require f to be convex.

	I_1	I_2-I_4	I_5	$I_6 - I_8$	I_9
Δf	$(0, 1]$	$[-1, 1]$	$[\mu, 1]$	$[-1, 1]$	$(0, 1]$

Table 1. The table summarizes the values that the Hessian can take. The interval I_5 ranges from $(-0.5, 0.5)$, thus, the lower bound $\mu > 0$ ensures that each f has a non-degenerate local minimum at the origin. The Hessian is restricted to be positive on the intervals I_1 and I_9 in order to exclude any stationary point except the origin.

4. Simulation Results

This section illustrates the results from Sec. 3 on a simulation example. We choose $C_{\mu,1}$ to be the set of all scalar functions that have a single non-degenerate minimum at the origin and whose Hessian is constant on the intervals $I_1 = (-\infty, -4.5), I_2 = (-4.5, -3.5), \dots, I_9 = (4.5, \infty)$. The Hessian may change on the interval boundaries and takes the values summarized in Table 1. Thus, the set $C_{\mu,1}$ includes certain nonconvex functions and satisfies the Assumptions (C1)-(C3). In addition, it is straightforward to generate random functions $f \in C_{\mu,1}$, by uniformly sampling potential values of the Hessian until the resulting function has a single local minimum.

We evaluate and compare the performance of two algorithms. The first algorithm is that provided by Prop. 3.1, for $k = 3$, which we refer to as Alg. 1. According to Prop. 3.3, Alg. 1 is guaranteed to converge on functions $f \in C_{\mu,1}$, and its convergence rate is lower bounded by $1/\kappa^{1/3}$. The second algorithm is given by a variant of Heavy Ball:

$$\ddot{x}(t) = -2\dot{x}(t)/\sqrt{\kappa} - \nabla f(x(t)). \qquad (27)$$

The algorithm is guaranteed to converge on functions $f \in C_{\mu,1}$, since it dissipates energy as long as $|\dot{x}(t)| > 0$. According to Sec. 3.2, its convergence rate is upper bounded by $1/\sqrt{\kappa}$.

The trajectories are simulated with the standard fourth-order Runge-Kutta method with a time step of 0.01. Both algorithms are evaluated on 50 randomly generated functions $f \in C_{\mu,1}$, and for each f, 50 simulations with randomized initial conditions are performed. The initial conditions are sampled from independent normal distributions with zero mean and standard deviation 4.5 (motivated by the interval boundary I_1 and I_9). Each simulation terminates once a tolerance of $|z_\mathrm{sim}(T_\mathrm{max})| \leq 10^{-8}$ is reached, where $z_\mathrm{sim}(t)$ denotes the simulated state trajectory that includes position, velocity, and, potentially, acceleration. The convergence rate ρ_sim is estimated by performing a least-squares fit:

$$\ln(|z_\mathrm{sim}(t)|/|z_\mathrm{sim}(0)|) \approx -\rho_\mathrm{sim}t + \ln(c),$$

with parameters $(\rho_\mathrm{sim}, \ln(c))$. Only times $t > 10$ are consid-

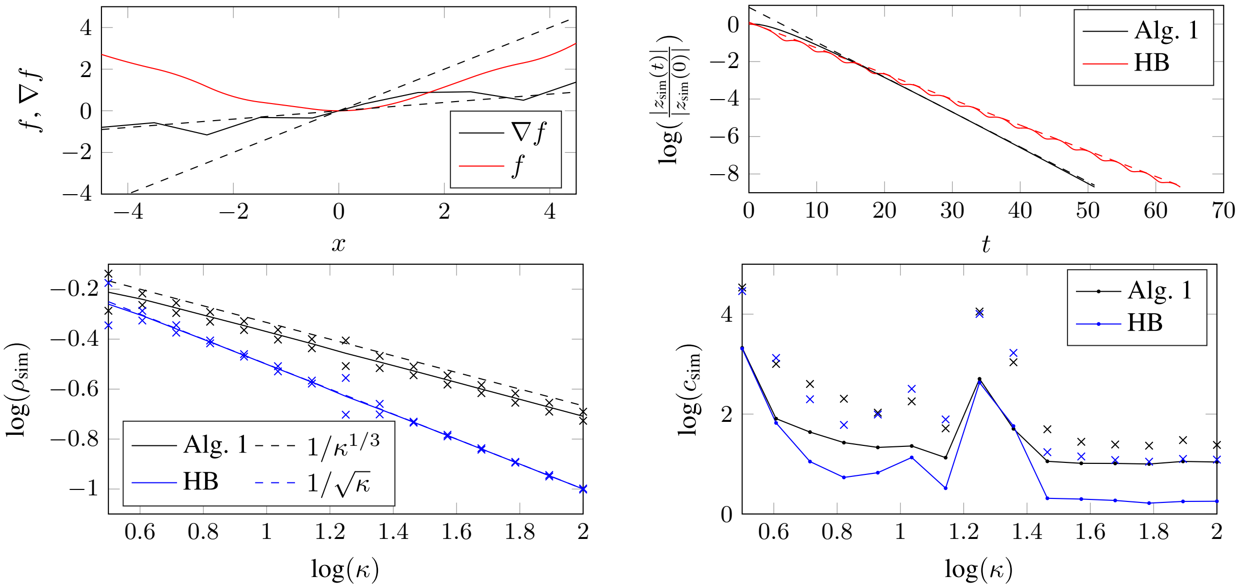

Figure 3. Top left: The graph shows a randomly generated nonconvex function $f \in C_{\mu,1}$ in red. Its gradient (black, solid) is shown with the sector bound $[1/\kappa, 1]$ (black, dashed). Thus, ∇f satisfies the sector bound $[\alpha_\text{s}, 1]$ for some $\alpha_\text{s} > 0$, but not the sector bound $[1/\kappa, 1]$. Top right: The graph compares two trajectories obtained with Alg. 1 and Heavy Ball, respectively. The resulting convergence estimates ρ_{sim} and c_{sim} are indicated with dashed lines. Bottom left: The graph shows the convergence rate ρ_{sim} when varying κ. Black relates to Alg. 1, blue to Heavy Ball. The solid lines show the mean across all initial conditions and all functions. The two-sigma bounds are marked with crosses. The dashed lines show the theoretical limits $1/\kappa^{1/3}$, respectively $1/\sqrt{\kappa}$. The mean for Heavy Ball lies almost exactly on the bound $1/\sqrt{\kappa}$. Bottom right: The graph shows the constant c_{sim} for different κ. The solid lines indicate the mean across all initial conditions and all functions (black for Prop. 3.1 and blue for Heavy Ball). The crosses indicate the upper two-sigma bound.

ered in order to avoid biases from fast-decaying transients. In a subsequent step, the smallest constant c_{sim} satisfying

$$|z_{\text{sim}}(t)| \leq c_{\text{sim}} |z_{\text{sim}}(0)| e^{-\rho_{\text{sim}} t}, \quad \forall t \in [0, T_{\max}]$$

is determined. The situation is illustrated on an example in Fig. 3 (top right).[1]

The resulting values ρ_{sim} and c_{sim} for different κ are summarized in Fig. 3 (bottom row). The results indicate that the convergence rate of Heavy Ball roughly scales with $-1/\sqrt{\kappa}$, whereas the convergence rate of Alg. 1 scales with $-1/\kappa^{1/3}$ as suggested by Prop. 3.1 and Prop. 3.3. Even though the convergence rate of Alg. 1 is superior, it tends to have higher constants c_{sim}. Also the variance in ρ_{sim} and c_{sim} seems higher. We observe that the obtained convergence rate estimates closely match the theoretical predictions.

5. Conclusions

We have shown that the convergence rate of first-order optimization algorithms is lower bounded, not only in discrete-time, but also in continuous time. The analysis shows that these limits are due to incomplete curvature information, since only upper and lower bounds on the local curvature of the objective function are assumed to be known. We found that the convergence rate of a kth-order algorithm is limited by $1/\kappa^{1/k}$ and provided an explicit algorithm that achieves this rate on smooth and strongly convex functions and even on certain nonconvex functions. We also note that this result is deceptive, since any algorithm whose convergence rate improves upon $\mathcal{O}(1/\sqrt{\kappa})$ necessarily includes dynamics that converge arbitrarily fast for large κ. Such an algorithm cannot be discretized with explicit linear methods. If such fast-converging dynamics are excluded, the analysis recovers the well-known asymptotic lower bound $\mathcal{O}(1/\sqrt{\kappa})$.

Numerical results with second and third-order dynamics indicate that the lower bound $1/\kappa^{1/3}$ is likely to be achieved even on nonconvex functions.

[1] The estimation of ρ_{sim} and c_{sim} over the finite-time interval $t \in [0, T_{\max}]$ is ill posed. For example, the constant c_{sim} can be increased in favor of a better convergence rate ρ_{sim}. Nevertheless, the suggested least-squares procedure typically provides reliable estimates as shown in Fig. 3 (top right). We also tested the procedure on quadratic functions, where the numerical results match the available closed-form expressions.

Acknowledgements

We thank the Branco Weiss Fellowship, administered by ETH Zurich, for the generous support and the Office of Naval Research under grant number N00014-18-1-2764.

References

Arjevani, Y., Shalev-Shwartz, S., and Shamir, O. On lower and upper bounds in smooth and strongly convex optimization. *Journal of Machine Learning Research*, 17 (126):1–51, 2016.

Åström, K. J. and Murray, R. M. *Feedback Systems*. Princeton University Press, second edition, 2008.

Bellman, R. *Stability Theory of Differential Equations*. McGraw-Hill, 1953.

Butcher, J. C. *Numerical Methods for Ordinary Differential Equations*. Wiley, third edition, 2016.

Callier, F. M. and Desoer, C. A. *Linear System Theory*. Springer Science and Business Media, 1991.

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. Lower bounds for finding stationary points II: first-order methods. *Mathematical Programming*, 2019. Preprint available online.

Diakonikolas, J. and Jordan, M. I. Generalized momentum-based methods: A Hamiltonian perspective. *arXiv:1906.00436 [math.OC]*, pp. 1–30, 2019.

Hahn, W. *Stability of Motion*. Springer, 1967.

Hairer, E., Nørsett, S. P., and Wanner, G. *Solving Ordinary Differential Equations*. Springer, second edition, 1993.

Jin, C., Netrapalli, P., Ge, R., Kakade, S. M., and Jordan, M. I. On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points. *arXiv:1902.04811 [cs.LG]*, pp. 1–31, 2019.

Khalil, H. K. *Nonlinear Systems*. Prentice-Hall, third edition, 1996.

Krichene, W., Bayen, A. M., and Bartlett, P. L. Accelerated mirror descent in continuous and discrete time. *Advances in Neural Information Processing Systems 28*, pp. 2845–2853, 2015.

Kunimatsu, S., Sang-Hoon, K., Fujii, T., and Ishitobi, M. On positive real lemma for non-minimal realization systems. *Proceedings of the 17th World Congress of the International Federation of Automatic Control*, pp. 5868–5873, 2008.

Minnichelli, R. J., Anagost, J. J., and Desoer, C. A. An elementary proof of Kharitonov's stability theorem with extensions. *IEEE Transactions on Automatic Control*, 34 (9):995–998, 1989.

Muehlebach, M. and Jordan, M. I. A dynamical systems perspective on Nesterov acceleration. *Proceedings of the International Conference on Machine Learning*, pp. 1–7, 2019.

Nesterov, Y. *Introductory Lectures on Convex Optimization*. Springer, 2004.

Nevanlinna, O. and Sipilä, A. H. A nonexistence theorem for explicit A-stable methods. *Mathematics of Computation*, 28(128):1053–1055, 1974.

Sastry, S. *Nonlinear Systems*. Springer, 1999.

Scoy, B. V., Freeman, R. A., and Lynch, K. M. The fastest known globally convergent first-order method for minimizing strongly convex functions. *IEEE Control Systems Letters*, 2(1):49–54, 2018.

Su, W., Boyd, S., and Candès, E. J. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016.

Wibisono, A., Wilson, A. C., and Jordan, M. I. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113(47): E7351–E7358, 2016.