Optimization with Momentum: Dynamical, Control-Theoretic, and Symplectic Perspectives

Michael Muehlebach   MICHAELM@BERKELEY.EDU
Department of Electrical Engineering and Computer Sciences
Department of Statistics
University of California
Berkeley, CA 94720-1776, USA

Michael I. Jordan   JORDAN@CS.BERKELEY.EDU
Department of Electrical Engineering and Computer Sciences
Department of Statistics
University of California
Berkeley, CA 94720-1776, USA

Editor: Ohad Shamir

Abstract

We analyze the convergence rate of various momentum-based optimization algorithms from a dynamical systems point of view. Our analysis exploits fundamental topological properties, such as the continuous dependence of iterates on their initial conditions, to provide a simple characterization of convergence rates. In many cases, closed-form expressions are obtained that relate algorithm parameters to the convergence rate. The analysis encompasses discrete time and continuous time, as well as time-invariant and time-variant formulations, and is not limited to a convex or Euclidean setting. In addition, the article rigorously establishes why symplectic discretization schemes are important for momentum-based optimization algorithms, and provides a characterization of algorithms that exhibit accelerated convergence.

Keywords: Gradient-based optimization, convergence rate analysis, Nesterov acceleration, symplectic integration, nonconvex optimization

1. Introduction

Optimization problems lie at the heart of many machine-learning formulations. As a result, a better understanding of optimization algorithms, combined with implementations that target distinctive properties of machine-learning problems, have contributed significantly to the recent progress in the field.

One of the most popular methods for large-scale optimization is the (stochastic) gradient method, due to its simplicity, wide applicability, and efficiency. However, in the deterministic setting, where gradients are evaluated exactly, it has been shown that better convergence rates can often be achieved by leveraging two successive gradients (Nesterov, 1983). Based on analogies to mechanical systems these methods are referred to *momentum-based optimization algorithms*, and can be viewed as particular discretizations of continuous-time harmonic oscillators, (Polyak, 1964). Yet, even for the class of strongly convex functions, most proofs that establish the superior convergence are algebraic

and provide little qualitative understanding. As a consequence, there is little guide to the generality or robustness of the acceleration phenomenon across instances of optimization problems.

This article characterizes the convergence rate of momentum-based optimization algorithms by taking fundamental topological properties into account. In many cases our analysis leads to closed-form expressions that relate the convergence rate to the different algorithm parameters, such as the step size and the damping. The analysis provides insight into the design of algorithms that, for example, require little tuning when applied to large-scale and ill-conditioned optimization problems. For simplicity of notation, we focus on the Euclidean setting, but we note that the scope of our analysis is not limited to Euclidean problems.

We will derive convergence rates of the form

$$|x(t) - x^*| \leq C_c |x(0) - x^*| \rho_c(t), \tag{1}$$

where $|x(t) - x^*|$ is a distance measure between the iterate $x(t)$ and the isolated local minimum x^*, t refers to time or the iteration number, C_c is a constant which does not depend on $x(0)$, and $\rho_c(t)$ is monotonically decreasing, satisfies $\rho_c(0) = 1$, and characterizes the convergence rate. An important aspect of the analysis is to characterize how the convergence rate ρ_c is affected by the shape of the objective function about an isolated local minimum. The local shape is summarized by the (condition) number κ, which is defined as the ratio between the maximum and minimum curvature about a local minimum. We call an isolated minimum degenerate if the minimum curvature vanishes in at least one direction. The main results and insights, which will be rigorously derived in the remainder of the article, are summarized as follows:

- The convergence rate, characterized by ρ_c in (1), is uniquely determined by the *local* shape of the objective function about an isolated minimum. The global shape of the objective function determines the constant C_c and the set of initial conditions $x(0)$ for which (1) holds. In other words, the global shape (for example described by convexity) determines the stability and region of attraction of a local minimum, whereas the local shape, i.e., the curvature at an isolated minimum, determines the convergence rate.

- An algorithm is called accelerated if the convergence rate scales favorably for ill-conditioned optimization problems, meaning that the convergence rate scales favorably for large κ. Acceleration is therefore a statement about the robustness of the convergence rate with respect to changes in the curvature, which is well defined for continuous-time as well as for discrete-time formulations.

- Accelerated convergence is generic to momentum-based optimization algorithms, provided that the damping scales with $1/\sqrt{\kappa}$ for large κ. Neither the evaluation of the gradient at a shifted position, nor a specifically engineered damping parameter, as for example proposed in Nesterov (2004, Sec. 2.2), are necessary.

- From a physics perspective, a momentum-based optimization algorithm can be thought of as a mass-spring-damper system, where the spring potential is given by the objective function. The algorithm design specifies the damping. The fact that the system has inertia (due to the second-order dynamics) implies that the convergence rate is robust to small changes in the spring, i.e., robust to small changes in the curvature of the objective function. The inertia gives the system the tendency to keep its velocity, which, provided that the damping is chosen

appropriately, implies that small changes in the spring will not slow down convergence. This captures the mechanism behind accelerated convergence.

- The convergence rate typically becomes arbitrarily small for large κ (if the dynamics are time-varying this statement applies for $t \to \infty$). The underlying dynamics are therefore close to conservative, which means that great care is required for the discretization. A discretization that introduces an artificial energy drift, such as the explicit forward Euler discretization, for example, might even lead to instability, which clearly makes a favorable scaling of the convergence rate with κ impossible. This motivates the use of symplectic discretization schemes, which preserve the Hamiltonian structure of the underlying conservative part of the dynamics.

- By introducing time-varying dynamics, which are obtained by adjusting the damping parameters with the number of iterations, the linear convergence rate is improved with an additional sublinearly converging term. If the local minimum is close to degenerate, the convergence is dominated by this term, as the rate of the linear convergence becomes arbitrarily small. In that way, the analysis relates the non-degenerate and degenerate cases.

Related work: The phenomenon of acceleration, which fundamentally motivates the use of momentum, has puzzled many researchers for almost 40 years. We do not attempt to give a full overview of the literature, but highlight some of the most recent work. Important contributions were made by Bubeck et al. (2015), who proposed an accelerated algorithm that has a geometric interpretation, by Allen-Zhu and Orecchia (2014), who show that coupling of gradient and mirror descent can lead to acceleration, and Lessard et al. (2016), who propose a general control-theoretic analysis framework. The framework has subsequently been extended and refined, for example by Michalowsky et al. (2019), who analyzed and quantified robustness and convergence trade-offs. Other work includes Diakonikolas and Orecchia (2018), who unify the analysis of first-order methods by imposing certain decay conditions, and Scieur et al. (2017), who interpret the accelerated gradient method as a multi-step discretization of gradient flow.

The work of Su et al. (2016) and Krichene et al. (2015) showed that the trajectories of the accelerated gradient method approach the solutions of a certain second-order ordinary differential equation. The resulting differential equation was analyzed in further detail in Attouch et al. (2018) and placed within a variational framework by Wibisono et al. (2016). This motivated further research on structure-preserving integration schemes for discretizing continuous-time optimization algorithms (Betancourt et al., 2018). While the continuous-time formulation of Su et al. (2016) is based on an accelerated optimization algorithm for smooth and convex objective functions (i.e., the case where the minimum could be degenerate), an alternative for strongly convex objective functions (non-degenerate case) was proposed in Dürr and Ebenbauer (2012). In França et al. (2019) and Muehlebach and Jordan (2019), important geometric properties of the underlying dynamics are highlighted and corresponding structure-preserving discretization schemes are analyzed. In addition, França et al. (2019) propose relativistic dynamics, as these naturally bound the rate of change of the position by the speed of light, which is supposed to prevent overshoot. The work of Maddison et al. (2018) suggests that convergence can be improved by a suitable choice of the kinetic energy (for example choosing the kinetic energy to be the convex conjugate of the objective function).

While this line of work suggests strong parallels between dynamical systems and optimization algorithms, a main part of the research focuses on engineering Lyapunov functions or estimate sequences that arise separately from the dynamical system. These typically arise from educated

guesses, or from the solutions of semidefinite programs, and enable the precise characterization of the convergence rate for a given algorithm. The aim of our work is different: We exploit fundamental topological properties of dynamical systems in order to provide an explicit characterization of the convergence rate up to constants, which, in most cases, can be evaluated in closed form. Hence, our analysis does not aim at precisely bounding the number of iterations needed to reach a local minimum with a certain tolerance, but provides a qualitative and quantitative characterization of the convergence rate up to constants. The analysis does not rely on convexity and applies both to continuous-time and discrete-time formulations. This provides a tool to understand how the convergence rate depends on various algorithm parameters and the shape of the objective function, and it characterizes the mechanism leading to acceleration. Our characterization of the convergence rate is both necessary and sufficient.

In the context of nonconvex optimization, the aim is typically to find a local minimum of a twice continuously differentiable function that satisfies certain non-degeneracy conditions. It has been shown that gradient descent converges to a local minimum from almost every initial condition, which is due to the fact that local maxima and saddle points are unstable equilibria (Lee et al., 2016). The same reasoning applies to gradient descent with momentum. However, even though gradient descent and gradient descent with momentum are guaranteed (in an almost everywhere sense) to ultimately reach a local minimum, it may take an arbitrarily long time to escape saddle points and local maxima. A major concern arises due to the fact that an objective function with n decision variables might have $n-1$ different isolated saddle points, which have to be traversed before reaching a local minimum. This implies that even for generic initialization strategies, the worst-case convergence rate depends on n (Du et al., 2017). This article is concerned with characterizing the convergence of momentum-based algorithms up to a constant factor. We will not analyze how this factor scales with n; the example from Du et al. (2017) suggests that the factor scales exponentially in n. However, the results from Ge et al. (2015) indicate that adding small random perturbations to gradient descent can improve the convergence rate and reduce the dimension-dependence of the convergence rate on n to polylog factors. A similar result applies likewise to gradient descent with momentum (Jin et al., 2017). A recent account of the state-of-the art is given in Jin et al. (2019), for example.

The poor scaling in the dimension n in the nonconvex case can be avoided by explicitly leveraging curvature information, as for example proposed in Nesterov and Polyak (2006) and Curtis et al. (2017). However, the computational cost per iteration of these methods is generally higher and increases with n. For simplicity, this article focuses on first-order algorithms with momentum, even though many ideas generalize in straightforward ways.

Outline: This article views optimization algorithms as dynamical systems. Two cases are distinguished: Section 2 assumes that the algorithm parameters are fixed, leading to time-invariant dynamics. The more general case, where the parameters are allowed to vary with the number of iterations, is discussed in Section 3.

Section 2 starts by introducing the notation and defining the scope of the analysis. A momentum-based optimization algorithm is understood as a second-order dynamical system,[1] where the local minima of the objective function correspond to asymptotically stable equilibria. Section 2.2 and Section 2.3 introduce two prototypical examples of momentum-based optimization algorithms. These are generalized versions of Nesterov's acclerated gradient scheme (Nesterov, 2004, Ch. 2.2),

---

1. In this context, the term "second order" has nothing to do with accessing higher derivatives of the objective function.

and also include Heavy-Ball methods (Polyak, 1964). By focusing on smooth dynamical systems based on smooth objective functions, our analysis excludes, for example, the treatment of constraints via (non-differentiable) indicator functions, as well as optimization algorithms with restart schemes. The subsequent result derived in Section 2.4 highlights our assertion that fundamental topological properties can be exploited for characterizing the convergence rate of momentum-based optimization algorithms. The results are applied to the analysis of the two prototypical examples in Section 2.6 and Section 2.7, leading to a broad characterization of the phenomenon of acceleration. In addition, Section 2.7 rigorously motivates the use of symplectic discretization schemes in the context of optimization: The symplectic discretization enables the computation of a modified energy function that can be used for stability analysis.

The structure of Section 3 is analogous to Section 2. The general result for characterizing the convergence rate in the time-varying case is presented in Section 3.1 and illustrated with two subsequent examples in Section 3.2 and Section 3.3. The results highlight the fact that time-varying damping parameters can speed up the convergence rate by an additional sublinearly converging factor. For ill-conditioned problems, this additional gain in the convergence rate becomes significant. The analysis connects the non-degenerate case with the degenerate case and motivates the update rules for the parameters of well-known accelerated gradient schemes, such as Nesterov (2004, p. 90, Constant Step Scheme II).

The article concludes with a summary and final remarks in Section 4.

2. The Time-Invariant Case

2.1 Introduction

Throughout the article we consider the problem of minimizing the function $f : \mathbb{R}^n \to \mathbb{R}$, which satisfies the following assumption:

Assumption 1 *The function f has a Lipschitz-continuous gradient. The critical points are non-degenerate and isolated.*[1]

The Lipschitz continuity of the gradient is important for ensuring that the resulting continuous-time trajectories exist and are unique. The fact that f has isolated non-degenerate critical points excludes pathological cases where the function has multiple connected local minima. The less pathological case, where the local minima are isolated but have vanishing curvature in certain directions, can be obtained via a limit argument, as will be discussed in Section 3.

Due to the Lipschitz continuity of the gradient, the Hessian exists almost everywhere and is essentially bounded. We will summarize the (essential) upper and lower bounds on the Hessian with the two constants $C_{\mathrm{f}} \geq 0$ and $\bar{C}_{\mathrm{f}} \geq 0$:

$$- C_{\mathrm{f}}|u|^2 \leq u^{\mathsf{T}} \frac{\partial^2 f}{\partial x^2} u \leq \bar{C}_{\mathrm{f}}|u|^2, \quad \forall u \in \mathbb{R}^n, \forall x \in \mathbb{R}^n \text{ (a.e.)}, \tag{2}$$

where $|\cdot|$ denotes the Euclidean norm. Thus for $C_{\mathrm{f}} = 0$ the function is convex, for $C_{\mathrm{f}} > 0$ it is nonconvex.

In order to simplify our exposition, we will consider functions that satisfy:

---

1. A critical point is non-degenerate if the Hessian $\mathrm{d}^2 f/\mathrm{d}x^2$ is nonsingular for all $x \in \mathbb{R}^n$ (almost everywhere) in a neighborhood of the critical point.

Assumption 2 *In addition to Assumption 1, f has a second derivative that is Lipschitz continuous.*

None of the subsequent results will explicitly depend on a Lipschitz constant related to the second derivative. Hence, under very mild conditions, all our results characterizing the convergence rate $\rho_c(t)$ in (1) apply when f satisfies Assumption 1 instead of Assumption 2.[1] This is further discussed in Appendix F.

Without loss of generality, we further assume that f has a local minimum at $x^* = 0$ and that $f(x^*) = f(0) = 0$. The convergence rate of a momentum-based optimization algorithm will depend on the local shape of the local minimum in question, which is determined by the constants

$$\mu := \min_{u \in \mathbb{R}^n, |u|^2 = 1} u^\mathsf{T} \frac{\partial^2 f}{\partial x^2}\bigg|_{x=x^*} u, \quad L := \max_{u \in \mathbb{R}^n, |u|^2 = 1} u^\mathsf{T} \frac{\partial^2 f}{\partial x^2}\bigg|_{x=x^*} u, \quad \kappa := \frac{L}{\mu}. \tag{3}$$

In case Assumption 2 is replaced with Assumption 1, the above constants are defined via the essential supremum and essential infimum of the Hessian in a neighborhood of x^*.

We model a momentum-based optimization algorithm either as a continuous-time or discrete-time dynamical system of the form

$$q^+(t) = g_\mathrm{q}(q(t), p(t)), \quad p^+(t) = g_\mathrm{p}(q(t), p(t)), \quad \forall t \in I, \tag{4}$$

$$q(0) = q_0, \quad p(0) = p_0, \tag{5}$$

where the superscript $+$ denotes either differentiation with respect to t (continuous-time setting), in which case $I = \mathbb{R}_{\geq 0}$, or a unit time-shift (i.e., $q^+(t) = q(t+1)$, discrete-time setting), in which case $I = \{0, 1, 2, \dots\}$. The nonnegative real numbers are denoted by $\mathbb{R}_{\geq 0}$, whereas the positive real numbers are denoted by $\mathbb{R}_{>0}$. The dynamics

$$g_\mathrm{q} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n, \quad g_\mathrm{p} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n, \tag{6}$$

are implicitly dependent on ∇f and are assumed to satisfy the following assumption.[2]

Assumption 3 *The dynamics g_q and g_p are continuously differentiable in both arguments and the derivatives are Lipschitz continuous.*

In the continuous-time case, Assumption 3 implies that the resulting trajectories exist and are unique for all times $t \in I$ (Arnol'd, 1992, p. 93, Corollary 3). Differentiability also implies that the dynamics can be linearized about an equilibrium, which typically provides a means to study the local behavior of the resulting trajectories. In addition, Assumption 3 implies that, over a finite time interval, trajectories are continuously dependent on their initial conditions (Arnol'd, 1992, p. 93, Corollary 4). These topological properties will be exploited in the following.

In order to simplify notation, we define

$$g : \mathbb{R}^{2n} \to \mathbb{R}^{2n}, \quad g := (g_\mathrm{q}, g_\mathrm{p}), \tag{7}$$

and introduce $z(t) := (q(t), p(t))$ for all $t \in I$. Moreover, the map $(q_0, p_0) \to (q(t), p(t))$ is denoted by $\varphi_t : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$, $t \in I$, and is referred to as the flow of the dynamical system (4) - (5).

Next, we provide a formal definition of a momentum-based optimization algorithm.

---

1. However, the smoothness of f might affect the constant C_c in (1). The important difference to the higher-order smoothness assumptions made in Nesterov and Polyak (2006) or Bubeck et al. (2019), for example, is that Assumption 2 does not impose *predefined* bounds on the higher derivatives of f, which would restrict the shape of f.

2. In the same way that the additional assumptions on the Hessian of f can be relaxed to mere Lipschitz continuity of ∇f, Assumption 3 can be relaxed to Lipschitz continuity of g_q and g_p; cf. Appendix F.

Definition 1 *We call the dynamical system* (4) - (5) *satisfying Assumption 3 a* momentum-based optimization algorithm *for the function f if $x^* = 0$ is an asymptotically stable equilibrium in the sense of Lyapunov.*

This means that $\varphi_t(0) = 0$, for all $t \in I$, φ_t is continuous at 0, uniformly in t, and $\lim_{t\to\infty} \varphi_t(z_0) = 0$ for any z_0 in a neighborhood of the origin.

Even though the dynamics g_q and g_p can capture gradient flow, as a special case (e.g., $g_\mathrm{p} = 0$, $g_\mathrm{q} = -\nabla f(q)$ in continuous time), we are interested in analyzing momentum methods, which arise from a nontrivial choice of g_q and g_p.

Throughout the article we will illustrate our ideas with two examples, which are prototypical versions of momentum-based optimization algorithms.

2.2 Example 1

The first example is based on the following continuous-time dynamics

$$\dot{q}(t) = p(t), \qquad \dot{p}(t) = -\nabla f(q(t)) + f_\mathrm{d}(q(t), p(t)), \qquad \forall t \in \mathbb{R}_{\geq 0}, \tag{8}$$

where ∇f denotes the gradient of f and f_d the dissipative forces. These dynamics can be viewed as a mass-spring-damper system, where f represents the spring potential and f_d the damping. The dissipative forces are assumed to take the form

$$f_\mathrm{d}(q, p) := -2dp - (\nabla f(q + \beta p) - \nabla f(q)), \tag{9}$$

where the parameters $d > 0, \beta \geq 0$ are constant. Ideally, the parameters d and β are designed to take into account additional information about the the function f; for example, upper and lower bounds on the curvature $\mathrm{d}^2 f/\mathrm{d}x^2$. If β is chosen to be zero, the dissipative forces reduce to $-2dp$. In that case, the dynamics (8) describe a continuous-time heavy ball method (Polyak, 1964). In case $\beta > 0$, the dynamics are related to Nesterov's accelerated gradient method (Muehlebach and Jordan, 2019).

An intuitive interpretation of the dissipative forces (9) can be given in the following way: For $\beta = 0$, (9) describes linear isotropic damping. For $\beta > 0$, (9) can be rewritten as

$$-2dp - \int_0^\beta \frac{\mathrm{d}^2 f}{\mathrm{d}x^2}\bigg|_{q+\tau p} \mathrm{d}\tau \, p, \tag{10}$$

which implies that (9) includes an additional damping term that averages the local curvature in the interval between q and $q + \beta p$. As a result, the damping increases if the local curvature is large, and reduces if the local curvature is small. As the velocity p is larger, the interval over which the average is taken is increased. The two forms of damping, linear isotropic and curvature dependent, are balanced by the coefficients d and β.

The equilibria of (8) are given by the critical points of f. Moreover, if a given critical point is a non-degenerate local minimum, the corresponding equilibrium is asymptotically stable. This follows by evaluating the total energy,

$$H(q, p) = \frac{1}{2}|p|^2 + f(q), \tag{11}$$

along the trajectories of (8),

$$\frac{\mathrm{d}}{\mathrm{d}t} H(q(t), p(t)) = f_{\mathrm{d}}(q(t), p(t))^{\mathsf{T}} p(t) = -2d|p|^2 - p^{\mathsf{T}} \int_0^\beta \frac{\mathrm{d}^2 f}{\mathrm{d}x^2}\bigg|_{q+\tau p} \mathrm{d}\tau\, p, \qquad (12)$$

which shows that the energy necessarily decreases in a neighborhood of the equilibrium. Combined with the fact that H is locally positive definite about a non-degenerate local minimum, this implies stability in the sense of Lyapunov. Asymptotic stability of the non-degenerate local minimum can then be concluded from La Salle's theorem (see, for example, Sastry, 1999, Ch. 5.4), which is based on examining the invariant sets satisfying $\mathrm{d}H/\mathrm{d}t = 0$. Thus, the dynamical system (8) is a momentum-based optimization algorithm for f according to Definition 1.

However, analyzing how the energy evolves along the trajectories of (8) also reveals global properties of the dynamics. For $\beta = 0, d > 0$ it follows that the set of initial conditions that do not converge to a local minimum is a set of measure zero, given by the critical points of f that are not local minima. The analysis holds without assuming convexity of f. However, if f happens to be convex and has a unique global minimum, then the corresponding equilibrium is globally asymptotically stable for any choice of parameters $\beta \geq 0, d > 0$.

The same reasoning applies in case f satisfies Assumption 1 instead of 2, or when f has degenerate local minima.

2.3 Example 2

The second example is obtained by discretizing (8) in the following way:

$$q_{k+1} = q_k + T p_{k+1}, \qquad p_{k+1} = p_k - T\nabla f(q_k) + T f_{\mathrm{d}}(q_k, p_k), \qquad \forall k \in \{0, 1, \dots\}, \qquad (13)$$

where $T > 0$ is the step size. The discretization consists of a forward Euler update of the momentum coordinates, and uses the newly computed momentum for the position update. As will be further discussed in the remainder of the article, the fact that the newly computed momentum coordinate is used for the position update makes the scheme symplectic for $f_{\mathrm{d}} = 0$. This means that the transformation (13) from $(q_k, p_k) \to (q_{k+1}, p_{k+1})$ preserves the symplectic form (for $f_{\mathrm{d}} = 0$), which has important consequences. One of these consequences concerns the spectrum of the linearization of (13). In case $f_{\mathrm{d}} = 0$, the corresponding eigenvalues are guaranteed to lie on the unit circle (for $T \leq 2/\sqrt{L}$). This is in sharp contrast to the standard explicit Euler discretization, which is not symplectic, and where the eigenvalues lie outside the unit circle even for arbitrarily small $T > 0$. Indeed, we will exploit the fact that the map (13) is symplectic (for $f_{\mathrm{d}} = 0$) to construct a modified energy function, which will be used for a stability analysis that extends beyond the linearization. Additional background information on symplectic integration can be found in Sanz-Serna (1992) or Hairer et al. (2002), for example.

For $\beta = 0$, the resulting algorithm is referred to as gradient descent with momentum (Polyak, 1964). For $\beta > 0$, Nesterov's accelerated gradient scheme (Nesterov, 2004, Constant step scheme III, p. 81) is obtained by choosing the parameters as follows:

$$T = \frac{1}{\sqrt{L}}, \quad d = \frac{\sqrt{L}}{\sqrt{\kappa}+1}, \quad \beta = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\frac{1}{\sqrt{L}}, \qquad (14)$$

where κ and L characterize the local shape of a local minimum, (Muehlebach and Jordan, 2019). Moreover, as pointed out in Muehlebach and Jordan (2019), the "Constant step scheme II" algorithm

of Nesterov (2004) is obtained by a particular choice of time-varying coefficients β and d. The generalization to time-varying coefficients will be discussed in Section 3.1.

The equilibria of (13) are again the stationary points of f. In order to determine the stability of an isolated local minimum, we linearize the dynamics,

$$\delta q_{k+1} = \delta q_k + T\delta p_{k+1}, \qquad \delta p_{k+1} = \delta p_k - 2dT\delta p_k - \beta T H_e \delta p_k - T H_e \delta q_k, \qquad (15)$$

where $H_e := \mathrm{d}f/\mathrm{d}x^2|_{x=0}$ (without loss of generality we consider $x^* = 0$). An eigenvalue analysis then reveals that the corresponding non-degenerate local minimum is asymptotically stable if

$$0 < T(2d + \beta h) < 2 - hT^2/2 \qquad (16)$$

holds for all eigenvalues h of H_e, i.e. $\mu \le h \le L$.[1] Asymptotic stability of the linearized dynamics implies that the same equilibrium is asymptotically stable under the nonlinear dynamics (13) (Sastry, 1999, p. 215).

For example, provided that the parameters d and β are chosen according to (14), we obtain that the given equilibrium is asymptotically stable if

$$0 < T < (\sqrt{5} - 1)/\sqrt{L}. \qquad (17)$$

We thus conclude that (13) is an optimization algorithm for f according to Definition 1 provided that the constants d, β, and T are chosen such that (16) is satisfied.

Unlike Example 1, our analysis for Example 2 is valid only in a neighborhood about the equilibrium, which corresponds to a local minimum. However, the specific structure of the discretization and the ideas from Example 1 can be exploited for obtaining a nonlinear analysis that is valid beyond a neighborhood of the equilibrium. This will be illustrated in Section 2.7.

2.4 Characterizing the convergence rate

Guaranteeing mere convergence is often not enough, as we are primarily interested in how quickly the trajectories of the nonlinear dynamics (4) and (5) converge to a local minimum. In the following, we argue that in most cases, a linear analysis characterizes the convergence rate up to constants. The linear analysis typically reduces to the computation of eigenvalues.

Our main proposition is based on the following assumption and definition.

Assumption 4 *Let the linearized dynamics*

$$\delta z^+(t) = \left.\frac{\partial g}{\partial z}\right|_{z=0} \delta z(t), \quad \forall t \in I, \quad \delta z(t_0) = z_0,$$

be such that there exists an estimate

$$|\delta z(t)| \le C_l |\delta z(t_0)| \exp(-\alpha(t - t_0)), \quad \forall z_0 \in \mathbb{R}^{2n}, \quad \forall t_0, t \in I, t \ge t_0,$$

where $C_l \ge 1$ and $\alpha > 0$ are constant.

---

1. The condition (16) is necessary and sufficient for the the linearized dynamics to be asymptotically stable for all h, with $\mu \le h \le L$.

Definition 2 *The region of attraction of the equilibrium at the origin (of the nonlinear dynamics (4)) is defined as the set*

$$\mathcal{R} := \{z_0 \in \mathbb{R}^{2n} \mid \lim_{t \to \infty} \varphi_t(z_0) = 0\}.$$

Proposition 3 *Let Assumption 4 be satisfied. Then, for any compact set $A \subset \mathcal{R}$ there exists a finite constant $\hat{C} \geq 1$ such that for all $z_0 \in A$*

$$|\varphi_t(z_0)| \leq \hat{C}|z_0| \exp(-\alpha t), \quad \forall t \in I. \tag{18}$$

Proof If A is empty the statement is trivial. We therefore assume that A is non-empty. Lemma 14 (see Appendix A) implies that there exists an open ball B_δ of radius $\delta > 0$, centered at the origin, such that any trajectory starting in B_δ converges with rate α. We make the following claim.

Claim: There exists a finite time $T_{\mathrm{m}} > 0$, $T_{\mathrm{m}} \in I$ such that for all $z_0 \in A$, $\varphi_t(z_0) \in B_\delta$, for all $t \in I, t \geq T_{\mathrm{m}}$.

Proof of the claim: The origin is a stable equilibrium. Hence there exists a constant $\epsilon > 0$ such that all trajectories starting in B_ϵ, the open ball of radius ϵ centered at the origin, remain in B_δ for all times. In addition, each $z_0 \in A$ satisfies $\lim_{t \to \infty} \varphi_t(z_0) = 0$, since $A \subset \mathcal{R}$. Thus, for each $z_0 \in A$ there exists a time $T(z_0)$ such that $\varphi_{T(z_0)}(z_0) < \epsilon/2$. The continuity assumptions on g imply that $\varphi_{T(z_0)}$ is continuous, which can be verified by the Grönwall inequality. Therefore, for each $z_0 \in A$ there exists an open ball $B(z_0)$ centered about z_0 such that for all $\bar{z}_0 \in B(z_0)$, $\varphi_{T(z_0)}(\bar{z}_0) < \epsilon$, which implies $\varphi_t(\bar{z}_0) \in B_\delta$ for all $t \geq T(z_0)$. The collection of all $B(z_0), z_0 \in A$ is an open cover for A. Due to the fact that A is compact, there exists a finite sub-cover (according to the Heine-Borel theorem), which we denote $B(\hat{z}_i), i = 1, 2, \ldots, N$, where N is finite. As a result, choosing $T_{\mathrm{m}} := \max_{i \in \{1,2,\ldots,N\}} T(\hat{z}_i)$ implies $\varphi_t(z_0) \in B_\delta$ for all $z_0 \in A, t \geq T_{\mathrm{m}}$, which proves the claim.

Moreover, the continuity of φ_t for all $t \in I, 0 \leq t \leq T_{\mathrm{m}}$, implies further that $\varphi_t(A)$ is bounded for any $t \in I, 0 \leq t \leq T_{\mathrm{m}}$. Combined with Lemma 14 (see Appendix A), this yields the following bound

$$|\varphi_t(z_0)| \leq \begin{cases} C_{\mathrm{A}} & 0 \leq t \leq T_{\mathrm{m}}, t \in I, \\ \delta\tilde{C} \exp(-\alpha(t - T)) & t > T_{\mathrm{m}}, t \in I, \end{cases} \tag{19}$$

for all $z_0 \in A$, where $C_{\mathrm{A}} \geq \delta$ and $\tilde{C} \geq 1$ are positive constants. We fix $z_0 \in A$, consider the trajectory $z(t) := \varphi_t(z_0), t \in I$, and apply the mean value theorem,

$$z^+(t) = \left(\left. \frac{\partial g}{\partial z} \right|_{z=0} + \left. \frac{\partial g}{\partial z} \right|_{z=\xi(t)} - \left. \frac{\partial g}{\partial z} \right|_{z=0} \right) z(t), \tag{20}$$

where $\xi(t)$ lies between $z(t)$ and the origin. Due to the fact that the dynamics are assumed to have Lipschitz-continuous derivatives, we obtain the following bound:

$$\left| \left. \frac{\partial g}{\partial z} \right|_{z=\xi(t)} - \left. \frac{\partial g}{\partial z} \right|_{z=0} \right| \leq \bar{C}_A |\xi(t)| \leq \bar{C}_A |z(t)|, \tag{21}$$

where $\bar{C}_A$ denotes a Lipschitz constant of $\partial g / \partial z$ on A. According to (19), the trajectory $|z(t)|$ is integrable (in continuous time) and absolutely summable (in discrete time). We obtain, by virtue of Lemma 13 (see Appendix A),

$$|z(t)| \leq C_1 \exp(C_1 \bar{C}_A C_{\mathrm{z}}) |z_0| \exp(-\alpha t), \quad \forall z_0 \in A, \quad \forall t \in I, \tag{22}$$

where C_z is constant. The constant C_z is related to an upper bound on the integral (in continuous time) or the sum (in discrete time) of $|z(t)|$ over $t \in I$, which according to (19), is guaranteed to be finite. ∎

Remarks:

- Proposition 3 characterizes the convergence rate and states that the number of iterations required to obtain an ϵ-accuracy approaches $\log(1/\epsilon)/\alpha$ for small ϵ. This does not provide a tight bound on the number of iterations required to achieve a certain accuracy, as the constant $\hat{C}$ might depend on α or grow rapidly with the size of A. Nevertheless, it enables a qualitative and quantitative discussion of the convergence rate α. In particular Proposition 3 highlights the fact that the convergence rate α is determined by the local properties of the dynamics, which depend on the local shape of the objective function f.

- The assumptions required for invoking Proposition 3 are often straightforward to verify, as Assumption 4 hinges on an eigenvalue analysis of $\partial g/\partial z$ at $z = 0$. Proposition 3 can also be generalized to the case where f satisfies Assumption 1 instead of Assumption 2, as shown in Appendix F.

- The proof highlights the following alternative statement of Proposition 3: There exists a finite time $T_m > 0$ after which all trajectories starting in A are guaranteed to be within a small neighborhood of the origin. Within this neighborhood the convergence is exponential with rate α.

- Given the smoothness properties of g, the assumption that the linearized dynamics converge with rate $\alpha > 0$ (Assumption 4) is necessary and sufficient for the convergence of the nonlinear dynamics with rate α, as can be shown with the arguments of Lemma 14 in Appendix A.

2.5 Implications for optimization algorithms

Proposition 3 enables a characterization of the convergence rate based on a linear analysis of the dynamics about an equilibrium. In the following, we will use Proposition 3 to discuss the convergence rate of the optimization algorithms given in Example 1 and Example 2. In particular, this provides conditions guaranteeing accelerated convergence.

2.6 Example 1

The total energy of the dynamical system is given by (11), and according to (12), energy is dissipated along trajectories. As we will show in the following, the energy function can therefore be used to characterize the region of attraction of the equilibrium $z^* = 0$, whereby the topology of the level sets of H and f will play an important role. This will be discussed next.

By assumption, f has a local non-degenerate minimum at $x^* = 0$, which implies that $f^{-1}([0, c])$ contains a compact, connected component containing the origin for sufficiently small $c > 0$. The set $f^{-1}([0, c])$ describes all values $x \in \mathbb{R}^n$ such that $0 \leq f(x) \leq c$. Morse theory (Milnor, 1963) concludes that the topology of the set $f^{-1}([0, c])$ is determined by the critical points of f. Thus, the set $f^{-1}([0, c])$ includes a compact, connected component that contains the origin (and no other
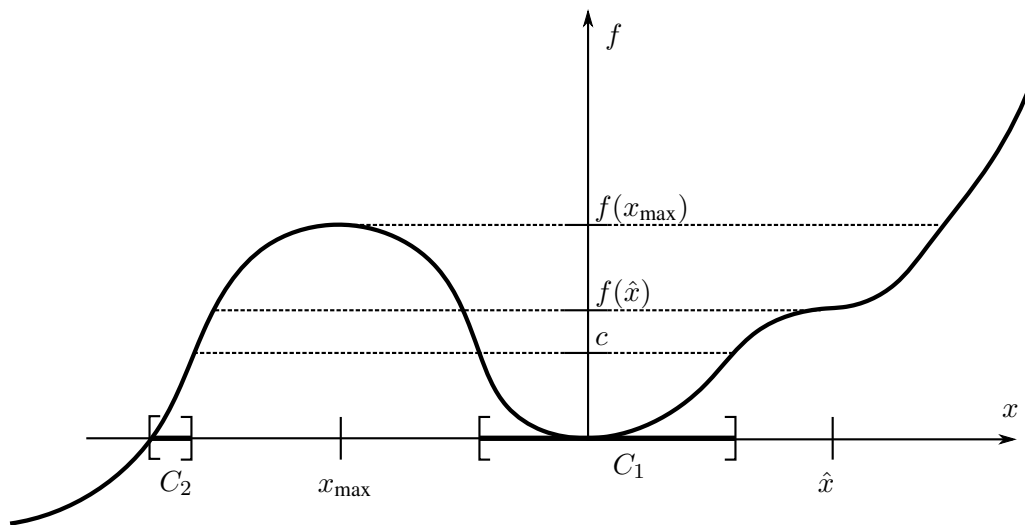
Figure 1: The figure illustrates the set $f^{-1}([0, c])$, with $c > 0$, when the function f is scalar. In this example, $f^{-1}([0, c])$ contains two connected components C_1 and C_2. The function f has a local minimum at the origin, a saddle point at $\hat{x}$, and a local maximum at $x_{\max}$. The origin is contained in the set C_1. Morse theory states that C_1 is guaranteed to be compact provided that $0 < c < f(\hat{x})$. Due to the fact that f is one-dimensional C_1 remains compact for $0 < c < f(x_{\max})$; this is, however, no longer true in higher dimensions.

critical point), as long as $c < \hat{f} = f(\hat{x})$, where $\hat{x}$ is any other critical point. The situation is illustrated with an example in Figure 1.

From the definition of H we infer that the critical points of H are all of the form $(q^*, 0)$, where q^* corresponds to a critical point of f. The above reasoning therefore also applies to the total energy H, and concludes that the set $H^{-1}([0, c])$ includes a compact, connected component that contains the origin (and no other critical point), as long as $c < \hat{f}$. This motivates the following definition, which will be used throughout the remainder of the article.

Definition 4 *The set $\mathcal{A}_f$ is defined as the connected component of $H^{-1}([0, \hat{f}))$ that contains the origin.*

Analyzing the rate of change of the energy along the trajectories of (8) (cf. (12)),

$$\frac{\mathrm{d}}{\mathrm{d}t} H(q(t), p(t)) = -2d|p(t)|^2 - p(t)^\mathsf{T} \int_0^\beta \frac{\mathrm{d}^2 f}{\mathrm{d}x^2}\bigg|_{q(t)+\tau p(t)} \mathrm{d}\tau \, p(t), \tag{23}$$

reveals that by a suitable choice of the parameters d and β, the energy necessarily decays. In particular, this is the case for $d > 0$, $0 \le \beta < 2d/C_\mathrm{f}$, where $C_\mathrm{f} \ge 0$ denotes a lower bound on the Hessian of f, as defined in (2). We therefore conclude as follows.

Proposition 5 *Provided that $d > 0$ and $0 \le \beta \le 2d/C_f$, the origin is an asymptotically stable equilibrium in the sense of Lyapunov. Its region of attraction contains the set $\mathcal{A}_f$.*

Proof We consider any initial condition $z(0) \in \mathcal{A}_f$ and define $H_0 := H(z(0))$. For $d > 0$ and $0 \le \beta \le 2d/C_\mathrm{f}$, the energy necessarily decays along the trajectories of (8), c.f. (23). The trajectory $z(t) = (q(t), p(t))$ is therefore confined to the connected component of $H^{-1}([0, H_0])$ that contains the origin, which, according to the above discussion, is necessarily compact. This implies that the origin is stable. In addition, the energy strictly decreases except when $p(t) = 0$. Hence, according to La Salle's theorem, (see, for example, Sastry, 1999, Ch. 5.4), $q(t)$ necessarily converges to a critical point of f, whereas $p(t)$ converges to zero. The origin is the only critical point contained in $\mathcal{A}_f$, which implies asymptotic stability of the origin. ∎

An eigenvalue analysis of the linearized dynamics (about the equilibrium) reveals that the eigenvalues are given by

$$- d - \frac{\beta h}{2} \pm \sqrt{\left(d + \frac{\beta h}{2}\right)^2 - h}, \tag{24}$$

where h is any eigenvalue of $\mathrm{d}^2 f/\mathrm{d}x^2|_{x=0}$. Thus, Proposition 3 asserts that the convergence rate for all initial conditions in $\mathcal{A}_f$ is directly determined by the real part of the eigenvalues, provided that $d > 0$ and $0 \le \beta < 2d/C_f$.

The eigenvalues h satisfy the upper and lower bounds $1/\kappa \le h/L \le 1$, cf. (3). An appropriate normalization of the constants d and β reduces the analysis to the case $L = 1$, which we consider in the following. Figure 2 shows how the eigenvalues vary as a function of d, β and h, according to the formula (24).

We consider first the case $\beta = 0$: For small d—i.e., $0 < d \le 1/\sqrt{\kappa}$—the eigenvalues are complex conjugates, have real part $-d$ (independent of h), and their imaginary part varies between $\sqrt{1/\kappa - d^2}$ and $\sqrt{1 - d^2}$ (as h changes). As d is increased above $1/\sqrt{\kappa}$ the eigenvalues can be
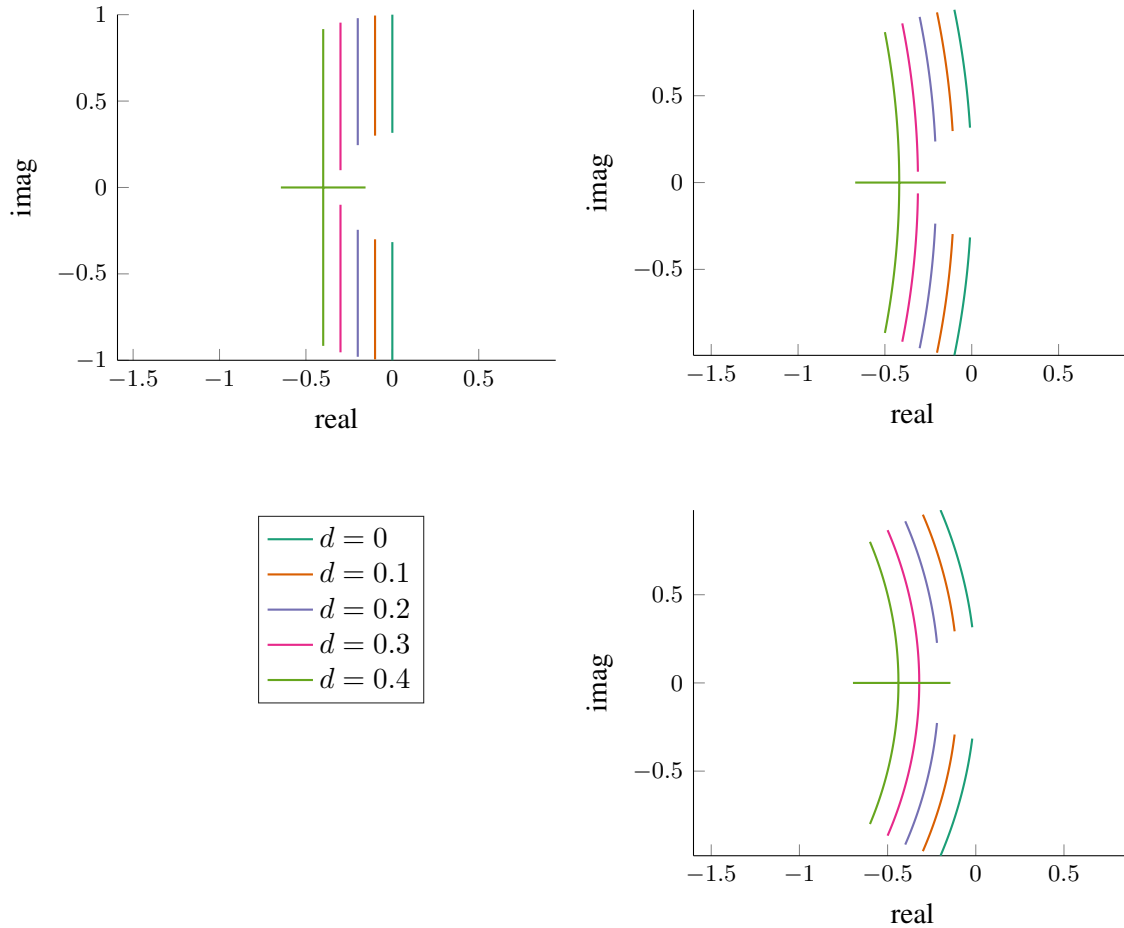
Figure 2: This figure shows how the eigenvalues of the linearization change as a function of the damping parameters d and β, and as a function of the curvature at the equilibrium. The top left plot shows the behavior of the eigenvalues for $\beta = 0$, as the curvature h is varied from 0 to 1 (the different colors represent the different values of d). The top right plot shows the behavior of the eigenvalues for $\beta = 0.2$ and the bottom right plot the behavior for $\beta = 0.4$, where the curvature h is varied from 0 to 1. The figure indicates that for $\beta = 0$, the eigenvalues are real for large values of d and small values of h. As h is increased the eigenvalues may become complex conjugated, in which case a further increase in h affects only the imaginary part. The additional parameter β has the effect of reducing the real part for larger values of h. This increases the convergence rate for large values of h.

both real or complex conjugates, depending on h. However, their worst-case real part is given by $-d + \sqrt{d^2 - 1/\kappa}$, which rapidly increases for larger d. The worst-case convergence rate (for a fixed d) is given by the maximum real part as h is varied between $1/\kappa$ and 1. Thus, in that sense, the optimal worst-case convergence rate is $1/\sqrt{\kappa}$ for $d = 1/\sqrt{\kappa}$.

Increasing β has the effect of increasing the convergence rate for larger values of h, since the parameter β introduces additional damping as h becomes large. Provided that $\beta \leq 1$, the qualitative behavior of the eigenvalues remains the same: For small d, i.e., $0 < d \leq 1/\sqrt{\kappa} - \beta/(2\kappa)$, the eigenvalues are complex conjugates with worst-case convergence rate $-d - \beta/(2\kappa)$. As d is increased above $1/\sqrt{\kappa} + \beta/(2\kappa)$, the eigenvalues can be both real or complex conjugated, depending on h. Their worst-case real part is again achieved for $h = 1/\kappa$ and is rapidly increasing for larger d. Again, the optimal worst-case convergence rate is $1/\sqrt{\kappa}$ obtained for $d = -\beta/(2\kappa) + 1/\sqrt{\kappa}$.

In continuous time, any desired convergence rate can be realized, by a simple reparametrization of time. A linear reparameterization, $\hat{t} = c_t t$, where $c_t > 0$ is constant, will simply scale the real parts and imaginary parts of the eigenvalues with c_t. A general, nonlinear, but diffeomorphic transformation will lead to time-varying dynamics, which will be discussed in Section 3. Thus, it might seem that any discussion of continuous-time convergence rates in the context of optimization is pointless. However, the analysis above tells us something different; it reveals how the convergence rate is affected by the condition number, which characterizes the shape of a local minimum. The analysis should be interpreted in the following way: provided that the time scale is fixed such that a convergence rate of 1 (or $1s^{-1}$ if the physical units are kept) is achieved for $\kappa = 1$, the analysis reveals how the convergence rate of any optimization algorithm of the type (8) deteriorates as κ increases.

For the following analysis we introduce the notation $u_1 \succ u_2$ if the function u_1 dominates the function u_2 for large arguments; that is, $\lim_{\kappa \to \infty} u_1(\kappa)/u_2(\kappa) = \infty$, where u_1 and u_2 are real-valued functions that are positive for large arguments. In the same way, the notation $u_1 \prec u_2$ implies that the function u_2 dominates u_1 for large arguments. We further use $u_1 \sim u_2$ to imply that neither u_1 nor u_2 are dominant for large arguments; that is,

$$\lim_{\kappa \to \infty} \frac{u_1(\kappa)}{u_2(\kappa)} < \infty, \quad \lim_{\kappa \to \infty} \frac{u_2(\kappa)}{u_1(\kappa)} < \infty. \tag{25}$$

We will analyze the performance of the algorithm given in Example 1 not only on a single function, but on a whole class of functions. In order to make the following statements precise, we will fix a compact set $A \subset \mathbb{R}^{2n}$ that contains the origin and introduce the following class of functions (parametrized by the constants $\bar{\kappa} \geq 1, C_f > 0, \bar{C}_f > 0$).

Definition 6 *Let $\mathcal{F}_{\bar{\kappa}, C_f, \bar{C}_f}$ denote the set of all functions such that each element $f \in \mathcal{F}_{\bar{\kappa}, C_f, \bar{C}_f}$ satisfies the following conditions: 1) f has an isolated local minimum at the origin with condition number $\kappa \leq \bar{\kappa}$, 2) f satisfies Assumption 2 and the bounds in (2), and 3) $\mathcal{A}_f \supset A$, where $\mathcal{A}_f$ is defined according to Definition 4.*

These conditions are motivated as follows. The first prescribes the local geometry about the local minimum at the origin, which influences the convergence rate in significant ways. The second ensures smoothness of the gradient, and the third guarantees that there are no critical points in A other than the origin.[1]

---

1. It is important to exclude additional critical points from the set A, since these are equilibria (a gradient-based algorithm cannot distinguish between saddle points, maxima, and minima). Clearly, a bound such as (18) cannot be satisfied for all $z_0 \in A$ if A contains equilibria other than the origin.

Acceleration is obtained whenever the convergence rate (again, relative to the convergence rate achieved for $\kappa = 1$) scales with $1/\sqrt{\kappa}$ for large values of κ. More precisely:

Definition 7 *A momentum-based algorithm is accelerated if there exists constants $\kappa_0 \geq 1$ and $c_a > 0$, such that for any $\kappa \geq \kappa_0$ and $f \in \mathcal{F}_{\kappa, C_f, \bar{C}_f}$, the following bound holds (for some constant $C_a > 0$):*

$$|\varphi_t(z_0)| \leq C_a |z_0| \exp(-c_a t/\sqrt{\kappa}), \quad \forall t \in I, \forall z_0 \in A.$$

We will now proceed to derive conditions on the parameters d and β of the algorithm given in Example 1 that guarantee accelerated convergence. We start with the case $\beta = 0$. For $d \succ 1/\sqrt{\kappa}$ it follows that the real part of the worst-case convergence rate scales with $-d + \sqrt{d^2 - 1/\kappa} \approx -1/(2d\kappa)$, which makes acceleration impossible. For $d \prec 1/\sqrt{\kappa}$ the damping is too small; i.e., the real part of (24) scales worse than $1/\sqrt{\kappa}$. Hence, acceleration is only achieved for $d \sim 1/\sqrt{\kappa}$. A similar argument applies to the case $0 < \beta$ and yields $d + \beta/(2\kappa) \sim 1/\sqrt{\kappa}$.

The above analysis is summarized with the following proposition.

Proposition 8 *In the nonconvex case ($C_f > 0$), the algorithm given in Example 1 is accelerated for the set of parameters $d \sim 1/\sqrt{\kappa}$, $0 < d$, and $0 \leq \beta < 2d/C_f$.*

In the convex case ($C_f = 0$), the algorithm given in Example 1 is accelerated for the set of parameters $d + \beta/(2\kappa) \sim 1/\sqrt{\kappa}$, $0 < d$, $0 \leq \beta$.

We would like to emphasize that the bound in Definition 7 is checked for each function $f \in \mathcal{F}_{\kappa, C_f, \bar{C}_f}$ individually, whereby the constant C_a may depend on the specific function f. Given the (potentially large) compact set A, Proposition 8 therefore answers the following question: What are the algorithm parameters ensuring that for any $f \in \mathcal{F}_{\kappa, C_f, \bar{C}_f}$ the time for reaching an ϵ accuracy scales with $\sqrt{\kappa}\ln(1/\epsilon)$ up to constants, i.e., for small ϵ? One important aspect of the analysis is that the convergence rate depends only on the local geometry of the objective function, as characterized by the constant κ.

The basic notion of local analysis, obtained from a derivation of eigenvalues, is well known (see, for example, Polyak (1987)). But Proposition 8 goes beyond classical local analysis in that, by virtue of Proposition 3, the local rate can be guaranteed for a large portion of the region of attraction of a given equilibrium. Key for this result is a nonlinear and global stability analysis. In the strongly convex setting, the results regarding the case $\beta = 0$ have been derived in Gadat et al. (2018). Similar results regarding the case $\beta \neq 0$ can be found in Attouch et al. (2018).

The bounds that we establish throughout the manuscript are restricted to initial conditions that are contained in a compact set within the region of attraction of a non-degenerate and isolated local minimum. This can be further motivated by considering the function shown in Figure 1. Provided that the algorithm dissipates energy, we infer from the above discussion that the region of attraction of the origin (an open set) will comprise points in the state space that are arbitrarily close to the saddle $(\hat{x}, 0)$ and the maximum $(x_{\max}, 0)$ (see Figure 1). Saddle points or maxima are equilibria, since any gradient-based algorithm cannot distinguish between them and minima. Hence, when initializing the algorithm close enough to $(\hat{x}, 0)$ and $(x_{\max}, 0)$, convergence of the algorithm can potentially take an arbitrarily long time (see Du et al. (2017) for a formal analysis). In that sense, the restriction of initial conditions to a compact set within the region of attraction of a given equilibrium appears to be natural.

2.7 Example 2

In discrete time, the stability analysis is not as straightforward, since the energy H is in general not dissipated along trajectories. However, the specific structure of the discretization can be exploited for obtaining a nonlinear stability analysis that is valid beyond a neighborhood of a local minimum.

As remarked in Muehlebach and Jordan (2019), the discretization (13) can be divided into two parts, an energy dissipation step and a symplectic Euler step (Hairer et al., 2002, p. 3),

$$\underbrace{\bar{z}_k = \Phi_{d,T}(z_k) = \begin{pmatrix} q_k \\ p_k - T f_{\mathrm{d}}(q_k, p_k) \end{pmatrix}}_{\text{energy dissipation}}, \quad \underbrace{z_{k+1} = \Phi_T(\bar{z}_k) = \begin{pmatrix} \bar{q}_k + T p_{k+1} \\ \bar{p}_k - T \nabla f(\bar{q}_k) \end{pmatrix}}_{\text{symplectic Euler}}, \quad (26)$$

with intermediate state $\bar{z}_k = (\bar{q}_k, \bar{p}_k)$. In order to simplify notation, we introduce the maps $\Phi_{d,T}$ and Φ_T to denote the energy dissipation and the symplectic Euler step, respectively. The symplectic Euler step is a well-known structure-preserving first-order integration scheme. Due to the symplectic integration, there exists a modified energy function that is nearly (up to exponentially small terms) conserved by the symplectic Euler step (Hairer et al., 2002, Chapter VI). The modified energy function can be computed by means of truncated Taylor-series expansions. In order to make the analysis rigorous, f is assumed to be analytic, which enables the estimation of higher-order derivatives using Cauchy's integral formula. It will be shown that for the subsequent stability analysis the assumption of f being analytic poses essentially no restriction, as any continuous function can be approximated arbitrarily closely by an analytic function on a compact domain (Stone-Weierstrass Theorem; Rudin, 1976, p. 159). The modified energy function is characterized by the following result.

Proposition 9 *Let f be analytic on B_r^c, the closed n-dimensional ball of radius r centered at the origin, and let L_H be a Lipschitz constant of ∇H on $B_r^c \times B_r^c$. Then there exists a perturbed Hamiltonian $\tilde{H} : B_r^c \times B_r^c \to \mathbb{R}$ such that*

$$|\tilde{H}(z_0) - \tilde{H}(\Phi_T(z_0))| \leq T C_{\Delta \tilde{H}} |\nabla H(z_0)|^2 e^{-T_0/T}, \quad (27)$$

for all $0 < T \leq T_0/3$ and for all $|z_0| \leq \frac{r}{2}(1 + 3.63 L_H T(1 + eT_0/3))^{-1}$, where

$$C_{\Delta \tilde{H}} := e(2.9 + 0.1 T_0)(1 + eT_0/3), \quad T_0 := \frac{2 \ln(2) - 1}{2 e L_H}$$

are constant. The perturbed Hamiltonian has the form

$$\tilde{H}(q, p) = H(q, p) - \frac{T}{2} \nabla f(q)^\mathsf{T} p + T^2 F(q, p),$$

where $F : B_r^c \times B_r^c \to \mathbb{R}$ is an analytic function. The perturbed Hamiltonian $\tilde{H}$ has the same critical points as H and satisfies

$$|H(z) - \tilde{H}(z)| \leq 15 L_H T |H(z)|, \quad \forall z \in \mathcal{A}_f \cap (B_{r/2}^c \times B_{r/2}^c),$$
$$|\nabla F(z)| \leq 356 L_H^2 |\nabla H(z)|, \quad \forall z \in B_{r/2}^c \times B_{r/2}^c. \quad (28)$$

Remarks:

- A more general version of Proposition 9 is stated and proved in Appendix C.

17

- The proof follows the reasoning of Hairer et al. (2002, p. 307), which enables the construction of $\tilde{H}$ by means of a truncated series expansion. Cauchy's integral formula is used to assert the convergence of the truncated series. The result from Hairer et al. (2002, p. 307) is extended by exploiting the specific structure of the underlying dynamics, which leads to the additional statements about the modified energy function $\tilde{H}$. These are crucial in the context of stability analysis.

- The constants T_0 and $C_{\Delta \tilde{H}}$ are determined as a function of the Lipschitz constant of ∇H, which can be regarded as the natural time constant for the dynamics governed by the Hamiltonian H. For $L_{\mathrm{H}} = 1$ we obtain the following values: $T_0 \approx 0.071$, $C_{\Delta \tilde{H}} \approx 8.4$,

$$|\tilde{H}(z_0) - \tilde{H}(\Phi_T(z_0))| \leq 8.4 T |\nabla H(z_0)|^2 e^{-0.071/T}, \tag{29}$$

  for all $0 < T \leq 0.023$ and all z_0 such that $|z_0| \leq 0.45r$. Choosing, for example, a time step $T = 0.001$ leads to the bound $1.2 \cdot 10^{-33}$ on the right-hand side of expression (29), indicating that $\tilde{H}$ is virtually exactly conserved by the symplectic Euler scheme.

- The estimate for the maximum time step $T_0/3$ is typically conservative. However, the proposition rigorously establishes that for a small enough time step, the perturbed Hamiltonian will be almost exactly conserved. The importance lies in the fact that the upper bound on the time step is only dependent on the Lipschitz constant of ∇H, which is directly related to the Lipschitz constant of ∇f. Due to the fact that Proposition 9 is a statement about the integration of the conservative part of the dynamics, the maximum time step $T_0/3$ is independent of the parameters d, β, and κ.

- The bounds (28) enable a nonlinear stability analysis, based on the modified Hamiltonian $\tilde{H}$. Due to the fact that $15 L_{\mathrm{H}} T < 1$ for all $T \leq T_0/3$, $\tilde{H}$ is necessarily positive in a neighborhood of the origin, cf. (28). Combined with the fact that the perturbed Hamiltonian has the same critical points as H, this concludes that the level sets of $\tilde{H}$ are compact in a region about the origin. The size of this region is determined by the critical points of H, as argued in the analysis of Example 1; see Section 2.6.

In the following, we will use the modified energy function $\tilde{H}$, whose existence is ensured by Proposition 9, for analyzing the stability of the dynamics (13) in the large. As pointed out, the dynamics (13) can be subdivided into a dissipation step and a symplectic Euler step.

In order to simplify the presentation we focus on the case where T is small and the map $\Phi_{d,T}$ is close to the identity (little damping). For stability analysis this is the most challenging setting, since, due to the almost vanishing damping, the convergence can be arbitrarily slow (the equilibrium is almost non-attractive). In case $\Phi_{d,T}$ is not close to the identity, which is, for example, obtained for a constant parameter $\beta > 0$, independent of κ, stability can be analyzed by means of the unperturbed energy function H, as shown in Appendix E. We will thus concentrate on the following result.

Proposition 10 *Let the dissipative forces $f_d(q_k, p_k)$ be such that $-d_2|p_k|^2 \leq p_k^\mathsf{T} f_d(q_k, p_k) \leq -d_1|p_k|^2$ with $0 < d_1 \leq d_2$.*

Then, there exists a maximum time step $T_{max} > 0$, such that the origin is an asymptotically stable equilibrium of the dynamics (13) for all $T \leq T_{max}$, with domain of attraction at least $\mathcal{A}_f \cap A$, where A is any compact set. Up to exponentially small terms due to (27), the maximum time step T_{max} depends on an upper bound on d_2, d_1/d_2, and L_H, the Lipschitz constant of ∇H.

Remarks:

- The proof of Proposition 10, which is included in Appendix D, is based on the Lyapunov function

$$V(q,p) = \tilde{H}(q,p) + \frac{Td_1}{2}\nabla f(q)^{\mathsf{T}}p,$$

  where $\tilde{H}$ denotes the perturbed energy function introduced in Proposition 9. The function V is motivated by the following observation. As the damping parameter d_1 decreases and $\Phi_{d,T}$ approaches the identity, V reduces to the perturbed energy function $\tilde{H}$, which (up to exponentially small terms) is known to be conserved by Φ_T. The correction $Td_1p^{\mathsf{T}}\nabla f(q)/2$ is required due to the fact that $\Phi_{d,T}$ is a pure contraction in the momentum variable, and as a result $\tilde{H}$ is not necessarily decreasing through the application of $\Phi_{d,T}$. A similar correction has been previously used in the literature on dissipative systems; see, for example, Hale (1988) or Haraux (1991), and in the context of the heavy-ball method in Gadat et al. (2018).

- Proposition 10 ensures that the region of attraction is the same as the continuous-time counterpart, and that the requirements on the time step for guaranteeing asymptotic stability is independent of the minimum damping d_1 (up to the exponentially small terms). This will be important in the following, where we will analyze how the convergence rate scales with κ.

- Provided that $\beta > 0$ and the function f is convex, the stability analysis simplifies substantially. In particular, a sufficient condition for stability is obtained by analyzing the total energy H along the trajectories of (13), as shown in Appendix E.

Next we will analyze the specific implications for the dynamics (13). We recall that the contraction step $\Phi_{d,T}$ is given by

$$p_{k+1} = p_k + Tf_{\mathrm{d}}(q_k,p_k) = (1-2dT)p_k - T\int_0^{\beta} \left.\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}\right|_{q_k+\tau p_k} \mathrm{d}\tau\, p_k,$$

which concludes that the upper and lower bounds d_2 and d_1 of Proposition 10 are given by $d_2 = 2d+\bar{C}_{\mathrm{f}}\beta$ and $d_1 = 2d-C_{\mathrm{f}}\beta$ (assuming $\beta \geq 0$). The upper and lower bounds d_2 and d_1 are therefore functions of κ.

Hence, according to Proposition 10, provided that $d_1 > 0$ and d_2/d_1 and d_2 are bounded with respect to κ, there exists a time step T, *independent* of κ, such that the origin is asymptotically stable, where the region of attraction includes any compact subset of $\mathcal{A}_f$. Then, according to Proposition 3, the convergence rate of any trajectory starting in $A \subset \mathcal{A}_f$, A compact, is determined by the magnitude of the eigenvalues of the linearized dynamics. The eigenvalues are given by

$$\lambda_{1,2} = 1 - T\left(d + \frac{\beta h}{2} + \frac{Th}{2} \pm \sqrt{(d + \frac{\beta h}{2} + \frac{Th}{2})^2 - h}\right),$$

enabling the calculation of the convergence rate for a given choice of T, d, β and h.

For the following discussion, we again assume that d and β are normalized such that $1/\kappa \leq h \leq 1$. We fix d, β, and T and analyze how the eigenvalues change as a function of h, as done in Figure 3. The worst-case convergence rate is determined by the maximum magnitude of the eigenvalues (over $1/\kappa \leq h \leq 1$). For very small values of h, the eigenvalues are real, and one

eigenvalue is very close to 1. As h is increased, the eigenvalues become complex conjugates, where for $\beta = 0$, the eigenvalues are located along circles centered at the origin, whereas for $\beta > 0$ their magnitude slightly decreases. If h is increased further, the eigenvalues become real again and their magnitude increases. The worst-case convergence rate, i.e., the largest magnitude of $|\lambda_{1,2}|$ is therefore either achieved for $h = 1/\kappa$ or $h = 1$.

We are interested in determining the conditions on d and β such that the convergence rate scales with $1/\sqrt{\kappa}$. We consider first the case $h = 1/\kappa$ and $\beta = 0$, and assume that d is large enough such that the eigenvalues are real. Then, provided that $d \sim 1/\sqrt{\kappa}$ it follows that $\lambda_{1,2} \sim 1 - T/\sqrt{\kappa}$. However, if d is chosen to be larger, i.e., $d \succ 1/\sqrt{\kappa}$, it follows that for large κ,

$$|\lambda_{1,2}| \approx 1 - Td + Td\sqrt{1 - \frac{1/\kappa}{d^2}} \approx 1 - \frac{T}{2d\kappa}, \tag{30}$$

that is, the convergence rate scales worse than $1/\sqrt{\kappa}$. In case d is small enough, such that the eigenvalues are complex conjugates, their magnitude is given by

$$|\lambda_{1,2}| = \sqrt{1 - 2dT}, \tag{31}$$

which indicates that a scaling of the convergence rate with $1/\sqrt{\kappa}$ is only achieved for $d \sim 1/\sqrt{\kappa}$.

For $h = 1$ and $\beta = 0$ it follows from $d \sim 1/\sqrt{\kappa}$ that $\lambda_{1,2}$ approach

$$\lambda_{1,2} \to 1 - \frac{T^2}{2} \mp T\sqrt{\frac{T^2}{4} - 1}, \tag{32}$$

for large κ. Thus, for $T \leq 1$, for example, the eigenvalues are complex conjugates and the worst-case convergence rate scales with $1/\sqrt{\kappa}$. The condition $d \sim 1/\sqrt{\kappa}$ is therefore necessary and sufficient for ensuring that the worst-case magnitude of $|\lambda_{1,2}|$ scales with $1/\sqrt{\kappa}$. A similar analysis applies to the case where $\beta > 0$, as shown below.

As in the discussion of Section 2.6 we say that the optimization algorithm (13) is accelerated if the convergence rate scales with $1/\sqrt{\kappa}$ (see Definition 7), where exponentially small terms resulting from the application of Proposition 9 are neglected. As a consequence, the above discussion allows us to translate the results from Proposition 8 almost verbatim to the discrete-time setting. This results in a broad characterization of the parameters d and β leading to acceleration.

Proposition 11 *In the nonconvex case, ($C_f > 0$), there exists a time step $T > 0$, such that the algorithm given in Example 2 is accelerated provided that $d \sim 1/\sqrt{\kappa}$, $0 < d$, $0 \leq \beta < 2d/C_f$.*

In the convex case, ($C_f = 0$), there exists a time step $T > 0$, such that the algorithm given in Example 2 is accelerated provided that $d + \beta/(2\kappa) \sim 1/\sqrt{\kappa}$, $0 < d$, $0 \leq \beta$ and either $\beta \sim 1$, $\beta \sim 1/\sqrt{\kappa}$, or $\beta \prec 1/\sqrt{\kappa}$.

Proof The requirements on β are needed for guaranteeing asymptotic stability of the origin for a time step T that is small enough, but independent of κ, as implied by Proposition 10 and Appendix E. It therefore remains to analyze the behavior of the eigenvalues $\lambda_{1,2}$ as a function of h for $1/\kappa \leq h \leq 1$.

For small h the eigenvalues are real. As h is increased, they may become complex conjugates, and if h is increased further they become real again. In case the eigenvalues are complex conjugates, their magnitude is given by

$$|\lambda_{1,2}| = \sqrt{1 - 2dT - \beta hT}. \tag{33}$$

We consider first the case $h = 1/\kappa$: The eigenvalues are real provided that $d + \beta/(2\kappa) + T/(2\kappa) \geq 1/\sqrt{\kappa}$, which, given the assumptions on β, implies that $d \geq 1/\sqrt{\kappa}$ for large κ. In case $d \sim 1/\sqrt{\kappa}$ it follows that $\lambda_{1,2} \sim 1 - T/\sqrt{\kappa}$, which therefore yields accelerated convergence. In case $d \succ 1/\sqrt{\kappa}$ it follows that

$$(d + \frac{\beta}{2\kappa} + \frac{T}{2\kappa})\sqrt{1 - \frac{1/\kappa}{(d + \frac{\beta}{2\kappa} + \frac{T}{2\kappa})^2}} \approx (d + \frac{\beta}{2\kappa} + \frac{T}{2\kappa}) - \frac{1}{2d\kappa + \beta + T}, \quad (34)$$

due to the fact that $1/(d^2\kappa) \to 0$ for large κ. Thus, accelerated convergence is impossible for $d \succ 1/\sqrt{\kappa}$. In case the eigenvalues are complex conjugates for $h = 1/\kappa$, it follows from (33) that $d \sim 1/\sqrt{\kappa}$ is necessary for accelerated convergence.

We therefore proceed to the case $h = 1$, where due to the fact that d vanishes for large κ, the two eigenvalues approach constant values,

$$\lambda_{1,2} \to 1 - T\left(\frac{\beta^\infty}{2} + \frac{T}{2} \pm \sqrt{(\frac{\beta^\infty}{2} + \frac{T}{2})^2 - 1}\right), \quad (35)$$

for $\beta^\infty := \lim_{\kappa \to \infty} \beta$, which is either constant or zero. This results in a convergence rate that is independent of κ, since T is chosen small enough to guarantee convergence. $\blacksquare$

In particular, this leads to the conclusion that, for example, the following heavy-ball scheme,

$$q_{k+1} = q_k + Tp_{k+1} \qquad p_{k+1} = p_k - (2T/\sqrt{\kappa})p_k - T\nabla f(q_k) \quad (36)$$

leads to accelerated convergence for a sufficiently small value of T. In case the Lipschitz constant of ∇f is bounded by one, the bound $T \leq 0.001$ for guaranteeing acceleration is obtained by following the proof of Proposition 10.[1] Note that compared to Nesterov's original scheme, a gradient evaluation at $q_k + \beta p_k$ is not required.

We would like to emphasize that our analysis captures the computational complexity up to constants. For a given objective function, choosing $\beta \neq 0$ typically allows larger steps T, which is likely to reduce the number of iterations needed for convergence. On strongly convex functions, for example, the "Constant step scheme III" presented in Nesterov (2004, p. 81), seems to provide a good compromise between a straightforward implementation and fast convergence (considering constants).

Local results akin to Proposition 11 which are based on a quadratic objective function are well known. A summary can be found for example in Lessard et al. (2016). In addition to unifying the continuous-time and discrete-time analysis, an important aspect of Proposition 11 is to establish a rate that is valid for a large portion of the region of attraction of a given equilibrium. Furthermore, the fact that the heavy-ball scheme leads to accelerated convergence for a sufficiently small value of T appears to be new. If T is chosen to be too large, however, for example by tuning T on quadratic functions, this may lead to convergence issues even on strongly convex functions, as has been reported in Lessard et al. (2016).

---

1. Strictly speaking, due to the exponential terms in (27), the bound $T \leq 0.001$ is valid for $\kappa \leq 4.1 \cdot 10^{52}$.
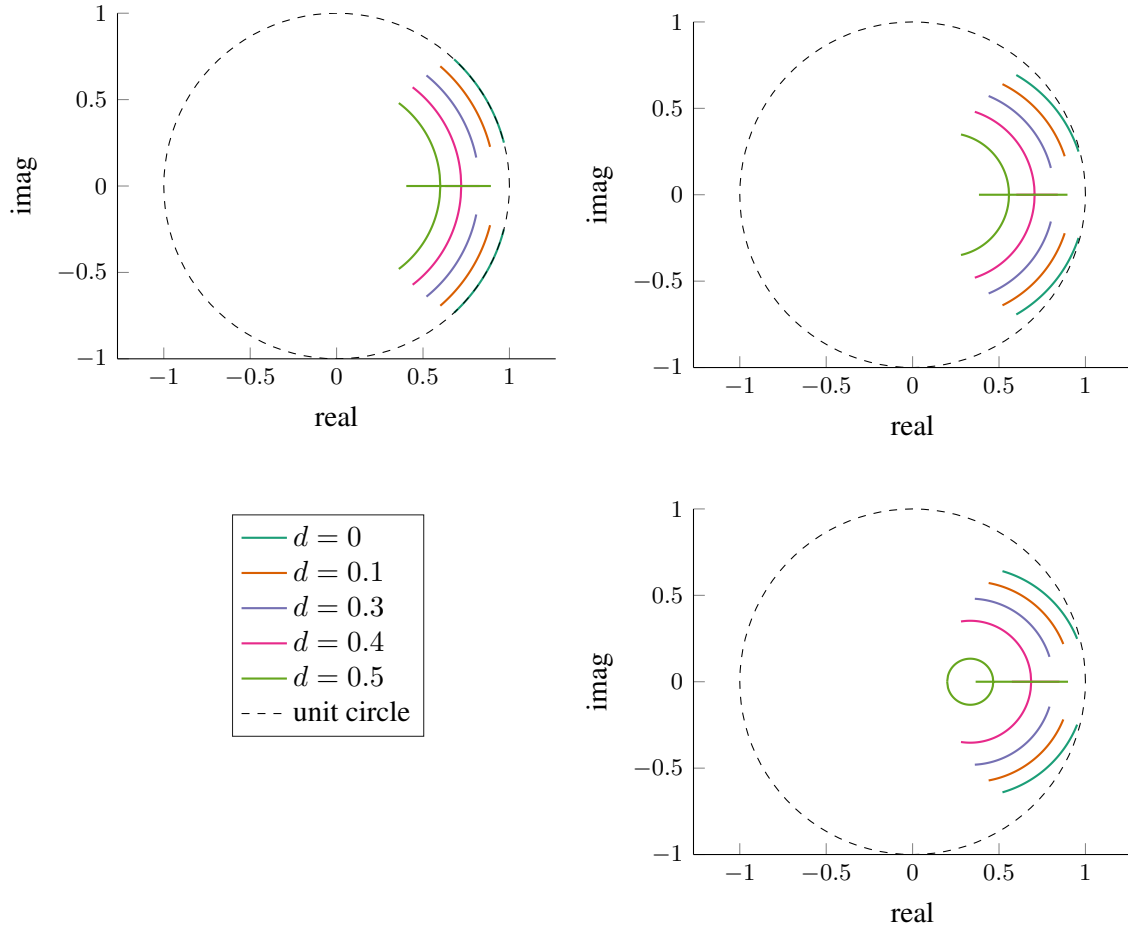
Figure 3: This figure shows how the eigenvalues of the linearization change as a function of the damping parameters d and β, and as a function of the curvature at the equilibrium, h. The top left plot shows the behavior of the eigenvalues for $\beta = 0$, as the curvature h is varied from 0.1 to 1 (the different colors represent the different values of d). The top right plot shows the behavior of the eigenvalues for $\beta = 0.2$ and the bottom right plot the behavior for $\beta = 0.4$, as the curvature h is varied from 0.1 to 1. For all the plots the time step is set to $T = 0.8$. The figure indicates that for $\beta = 0$, the eigenvalues move first along the real axis and then along concentric circles (as h changes). The additional parameter β has the effect of reducing the magnitude of the eigenvalues for larger values of h. While this can increase the convergence rate, it bears also the risk of instability, when d, β, and T are chosen to be too large.

3. The Time-Varying Case

Next we will consider the case where the dynamics g are time dependent. The main motivation for introducing time-dependent dynamics lies in improving the convergence rate for functions that have an almost vanishing curvature at a local minimum.

We slightly abuse notation and reuse the variables introduced in Section 2.4. It will be clear from context whether we refer to the time-varying or time-invariant case.

Following Section 2, we model a momentum-based optimization algorithm as a continuous-time or discrete-time dynamical system of the form

$$q^+(t) = g_q(t, q(t), p(t)), \quad p^+(t) = g_p(t, q(t), p(t)), \quad \forall t \in I, t \geq t_0, \tag{37}$$

$$q(t_0) = q_0, \quad p(t_0) = p_0, \quad t_0 \in I, \tag{38}$$

where the dynamics satisfy the following assumptions:[1]

Assumption 5 *The dynamics g_q and g_p are continuously differentiable in all their arguments. The derivative in the second and third arguments is Lipschitz continuous, uniformly in t.*

We define the map from $(q_0, p_0, t_0) \to (q(t), p(t))$ as $\varphi_t : \mathbb{R}^{2n} \times I \to \mathbb{R}^{2n}$, for $t \in I$, $t \geq t_0$, and define the dynamical system (37)-(38) to be a *momentum-based optimization algorithm* for the function f if $x^* = 0$ is an asymptotically stable equilibrium in the sense of Lyapunov (uniformly in the initial time t_0). Due to the continuity assumptions on the dynamics g_q and g_p, the continuous-time existence and uniqueness results, as well as the continuity of trajectories with respect to their initial conditions continues to hold (Arnol'd, 1992, p. 93, Corollaries 3 and 4).

The examples presented in Section 2.2 and Section 2.3 are modified by allowing the constants $d > 0$ and $\beta \geq 0$ to be time varying. We restrict ourselves to the case where d and β converge for large time, which leads to asymptotically time-invariant dynamics. This has the advantage that the asymptotic convergence rate is independent of the initial time t_0, and, under mild continuity conditions, the (optimal) convergence rates from Section 2.4 will be recovered for large t. Thus, (8) and (13) are extended by replacing the constants d and β with the functions $d : \mathbb{R}_{\geq 0} \to \mathbb{R}_{>0}$ and $\beta : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, which are assumed to be continuously differentiable and convergent,

$$d_\infty := \lim_{t \to \infty} d(t) > 0, \quad \beta_\infty := \lim_{t \to \infty} \beta(t) \geq 0. \tag{39}$$

The stability analysis of Example 1 (Section 2.2) translates to the time-varying case.[2] This implies that the algorithm (8) with time-varying parameters d and β is a momentum-based optimization algorithm according to the above definition.

The stability analysis of Example 2 is extended to the time-varying case by rewriting the dynamics (37) in the following way:

$$z_{t+1} = \left(\underbrace{\left.\frac{\partial g}{\partial z}\right|_{t \to \infty, 0}}_{\text{linear time-invariant part}} + \underbrace{\left(\left.\frac{\partial g}{\partial z}\right|_{t,0} - \left.\frac{\partial g}{\partial z}\right|_{t \to \infty, 0} \right)}_{\text{linear time-variant part}} \right) z_t + \underbrace{r_{\text{NL}}(t, z_t)}_{\text{nonlinear part}}, \tag{40}$$

---

1. The assumptions can be relaxed to mere Lipschitz continuity of the second and third argument as discussed in Appendix F.

2. La Salle's invariance principle must be extended (as discussed in, e.g., Sastry, 1999, p. 205).

where the linear time-varying part vanishes for $t \to \infty$. The remainder r_{NL} is of second order in z_t (uniformly in $t \in I$). Thus, if the linear time-invariant part has all eigenvalues strictly within the unit circle, the dynamics (40) are asymptotically stable, uniformly in t_0.[1] Consequently, the analysis from Section 2.3 yields the following conditions, cf. (16):

$$0 < T(2d_\infty + \beta_\infty h) \leq 2 - hT^2/2, \quad d_\infty > 0, \quad \beta_\infty \geq 0, \tag{41}$$

for all eigenvalues h of $H_{\mathrm{e}} = \mathrm{d}^2 f/\mathrm{d}x^2 \big|_{x=0}$.

3.1 Characterizing the convergence rate

As in Section 2.4 we argue that a linear analysis can be used to characterize the convergence rate of the nonlinear dynamics up to constants. The results are derived only for non-degenerate isolated local minima. We believe, however, that the discussion also provides insights into the case where the curvature vanishes at a local minimum; see Section 3.2 for further discussion of this point.

A similar analysis of the role of choosing a time-varying damping parameter can be found in Attouch et al. (2018). The emphasis of the following section lies in providing intuition through a unified treatment between the degenerate and non-degenerate case.

We make the following assumption.

Assumption 6 *Let the linearized dynamics*

$$\delta z^+(t) = \frac{\partial g}{\partial z}\bigg|_{t,z=0} \delta z(t), \quad \forall t \in I, \quad t \geq \tau, \quad \delta z(\tau) = z_0,$$

be such that there is an estimate

$$|\delta z(t)| \leq C_l |\delta z(\tau)| \frac{\rho(t)}{\rho(\tau)} \exp(-\alpha(t - \tau)), \quad \forall z_0 \in \mathbb{R}^{2n}, \quad \forall \tau, t \in I, t \geq \tau,$$

where $C_l \geq 1$ and $\alpha > 0$ are constant, and $\rho : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is continuous, and monotonically decreasing.

Proposition 12 *Let Assumption 6 be satisfied and let the region of attraction of the equilibrium at the origin of the nonlinear dynamics (37) be denoted by $\mathcal{R} \subset \mathbb{R}^{2n}$. Then, for any compact set $A \subset \mathcal{R}$ and any initial time $t_0 \in I$, there exists a finite constant $\hat{C} \geq 1$ such that for all $z_0 \in A$,*

$$|\varphi_t(z_0, t_0)| \leq \hat{C} |z_0| \frac{\rho(t)}{\rho(t_0)} \exp(-\alpha(t - t_0)), \quad \forall t \in I, \quad t \geq t_0.$$

Proof The set A is compact, nonempty, and contains the origin. Lemma 16 (see Appendix B) implies that there is a ball B_δ of radius $\delta > 0$, centered at the origin, such that any trajectory starting in B_δ at time $\tau \in I$ converges with rate $(\rho(t)/\rho(\tau)) \exp(-\alpha(t - \tau))$. The following claim can be verified with the arguments of Proposition 3.
Claim: There exists a finite time $T_{\mathrm{m}} > t_0, T_{\mathrm{m}} \in I$ such that for all $z_0 \in A$, $\varphi_{T_{\mathrm{m}}}(z_0, t_0) \in B_\delta$.

---

1. The linear part then converges at an exponential rate, as can be shown by Lemma 15 (see Appendix B). An argument similar to Bellman (1969, Thm. 1') can be applied to argue that the nonlinear dynamics are uniformly asymptotically stable in a neighborhood of the origin.

Moreover, the uniform continuity of φ_t for all $t \in I$, $t_0 \leq t \leq T_{\mathrm{m}}$, implies further that $\varphi_t(A, t_0)$ is bounded for any $t \in I$, $t_0 \leq t \leq T_{\mathrm{m}}$. Combined with Lemma 16, this yields the following bound

$$|\varphi_t(z_0, t_0)| \leq \begin{cases} C_{\mathrm{A}} & t_0 \leq t \leq T_{\mathrm{m}}, t \in I, \\ \delta \tilde{C}(\rho(t)/\rho(T_{\mathrm{m}})) \exp(-\alpha(t-T)) & t > T_{\mathrm{m}}, t \in I, \end{cases} \tag{42}$$

for all $z_0 \in A$, where $C_{\mathrm{A}} \geq \delta$ and $\tilde{C} \geq 1$ are positive constants. We fix $z_0 \in A$, consider the trajectory $z(t) := \varphi_t(z_0, t_0)$, $t \in I$, and apply the mean value theorem,

$$z^+(t) = \left(\left.\frac{\partial g}{\partial z}\right|_{t,z=0} + \left.\frac{\partial g}{\partial z}\right|_{t,z=\xi(t)} - \left.\frac{\partial g}{\partial z}\right|_{t,z=0} \right) z(t), \tag{43}$$

where $\xi(t)$ lies between $z(t)$ and the origin. Due to the fact that the dynamics are assumed to have Lipschitz continuous derivatives (uniformly in t), we obtain the following bound

$$\left| \left.\frac{\partial g}{\partial z}\right|_{t,z=\xi(t)} - \left.\frac{\partial g}{\partial z}\right|_{t,z=0} \right| \leq \bar{C}_{\mathrm{A}}|z(t)|, \tag{44}$$

where $\bar{C}_{\mathrm{A}}$ denotes the Lipschitz constant of $\partial g/\partial z$ on A (uniformly in t). According to (42), the trajectory $|z(t)|$ is integrable (in continuous time) and absolutely summable (in discrete time). We obtain, by virtue of Lemma 15,

$$|z(t)| \leq C_1 \exp(C_1 \bar{C}_{\mathrm{A}} C_{\mathrm{z}})|z_0| \frac{\rho(t)}{\rho(t_0)} \exp(-\alpha(t-t_0)), \quad \forall z_0 \in A, \tag{45}$$

where C_{z} is constant. The constant C_{z} is related to an upper bound on the integral (in continuous time) or the sum (in discrete time) of $|z(t)|$ over $t \in I$, which according to (42), is guaranteed to be finite. ∎

3.2 Example 1

The energy function, as defined in (11), is not explicitly dependent on time, even if the parameters d and β are chosen to be time-varying. Following the discussion of Section 2.6, we conclude that $\mathcal{A}_f$ is contained in the region of attraction of the equilibrium at the origin, as long as

$$0 < d(t), \quad 0 \leq \beta(t) < 2d(t)/C_{\mathrm{f}}, \quad \forall t \in I,$$
$$0 < d_\infty, \quad 0 \leq \beta_\infty < 2d_\infty/C_{\mathrm{f}}.$$

As a result, Proposition 12 implies that the convergence rate of the trajectories starting from $A \subset \mathcal{A}_f$, A compact, is given by the linearization of (8) about the origin.

After a change of coordinates that diagonalizes H_{e}, the linearized dynamics of a single coordinate are given by

$$\delta \dot{q}(t) = \delta p(t), \qquad \delta \dot{p}(t) = (-2d(t) - \beta(t)h)\delta p(t) - h\delta q(t), \tag{46}$$
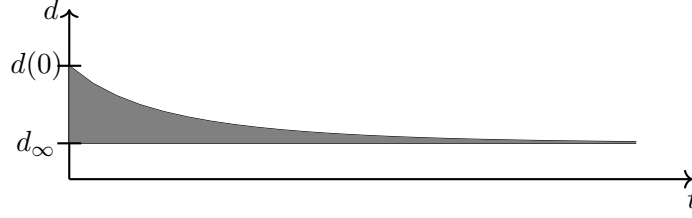
Figure 4: The figure shows $d(t)$ defined according to (49) with $d(0) = 1$ and $d_\infty = 1/\sqrt{10}$ as an example. Compared to the time-invariant dynamics (obtained in the asymptotic limit), which converge with rate d_∞, the time-varying terms improve the convergence rate by the shaded area, as indicated with (47).

where $\delta q(t) \in \mathbb{R}$, $\delta p(t) \in \mathbb{R}$, and h denotes the corresponding eigenvalue of H_e. As in Section 2.6, upper and lower bounds on the eigenvalues of H_e are given by $1/\kappa \le h \le 1$. The analysis simplifies by expressing the dynamics using the coordinates $\delta\tilde{q}$, which are defined by

$$\exp\left(-\int_{t_0}^t \bar{d}(\tau)\mathrm{d}\tau\right)\delta\tilde{q}(t) := \delta q(t), \tag{47}$$

with $\bar{d}(t) := d(t) + \beta(t)h/2$, and yields

$$\delta\ddot{\tilde{q}}(t) = -(-\dot{\bar{d}}(t) - \bar{d}(t)^2 + h)\delta\tilde{q}(t). \tag{48}$$

The dynamics (48) represent an undamped linear harmonic oscillator whose frequency varies with time. This time-varying frequency can be interpreted as the time-varying generalization of the square-root term in (24). The non-square-root term in (24) is captured by the coordinate change (47).

In the following, the choice

$$\dot{d}(t) = -d(t)^2 + d_\infty^2, \quad \beta(t) = 0, \tag{49}$$

is discussed. This leads to a time-invariant right-hand side of (48), and simplifies the subsequent analysis. The more general case (no restriction on $\bar{d}$) can be analyzed numerically, or by approximating $\bar{d}$ with a rational function and applying tools from asymptotic analysis (White, 2010). The solutions of (49) monotonically approach d_∞ as t increases, whereby the time-invariant case discussed in Section 2.6 is recovered with $d(0) = d_\infty$ ($d(0) = d_\infty$ implies $d(t) = d_\infty$ for all $t \ge 0$). The coordinate transformation (47) therefore implies that, compared to the time-invariant case, the convergence rate can be improved by choosing $d(0) > d_\infty$, as this increases the area underneath the curve $d(t)$. Figure 4 illustrates the situation.

This intuition can be made quantitative by noticing that the solutions of (49) are given by

$$d(t) = d_\infty \tanh(d_\infty(t + C_\mathrm{d})), \quad t \ge t_0, \tag{50}$$

26

where C_d is constant and related to $d(t_0)$. This implies

$$\exp\left(-\int_{t_0}^{t} \bar{d}(\tau)\mathrm{d}\tau\right) = \frac{\cosh(d_\infty(t_0 + C_d))}{\cosh(d_\infty(t + C_d))} = \frac{e^{-2d_\infty(C_d+t_0)} + 1}{e^{-2d_\infty(C_d+t)} + 1}e^{-d_\infty(t-t_0)}$$

$$= \frac{d_\infty + d(t)}{d_\infty + d(t_0)}e^{-d_\infty(t-t_0)} =: \frac{\rho_{tv}(t)}{\rho_{tv}(t_0)}e^{-d_\infty(t-t_0)}. \tag{51}$$

Due to the fact that (48) reduces to a time-invariant harmonic oscillator, the solutions to (46) can be expressed in closed form with elementary functions. The fundamental solutions to (46) are therefore given by[1]

$$\frac{\rho_{tv}(t)}{\rho_{tv}(t_0)}e^{(-d_\infty \pm \sqrt{d_\infty^2 - h})(t-t_0)}. \tag{52}$$

The rate of the exponential decay behaves in the same way as the real part of the expression (24) (d_∞ corresponds to d in (24)) and as a result, the discussion for the time-invariant case applies here in the same way. This concludes that accelerated convergence occurs as long as $d_\infty \sim 1/\sqrt{\kappa}$. Compared to the time-invariant case, however, the convergence is improved by the factor $\rho_{tv}(t)/\rho_{tv}(t_0)$. For large values of d_∞ the difference is not substantial (at most a constant factor of size $\lim_{t\to\infty} \rho_{tv}(t)/\rho_{tv}(t_0) = 2d_\infty/(d_\infty + d(t_0))$ can be gained). However, the improvement becomes crucial for small values of d_∞, since $\rho_{tv}(t)/\rho_{tv}(t_0)$ behaves as

$$\frac{\rho_{tv}(t)}{\rho_{tv}(t_0)} \approx \frac{1}{1 + d(t_0)(t - t_0)},$$

for small t. Thus choosing $d_\infty \sim 1/\sqrt{\kappa}$, ensures convergence roughly with $1/(1+d(t_0)(t-t_0))$ for small t, and exponential convergence with rate $1/\sqrt{\kappa}$ for large t; the transition occurs for $(t - t_0) \approx 1/d_\infty$. The situation is shown in Figure 5.

Summarizing, we therefore conclude according to Proposition 12 that choosing $d_\infty \sim 1/\sqrt{\kappa}$ implies that the trajectories of the time-varying generalization of (8) satisfy the estimate

$$|(q(t), p(t))| \leq C_{tv}|(q(0), p(0))|\frac{\rho_{tv}(t)}{\rho_{tv}(0)}\begin{cases} e^{-d_\infty t}, & 0 \leq d_\infty \leq 1/\sqrt{\kappa}, \\ e^{(-d_\infty + \sqrt{d_\infty^2 - 1/\kappa})t}, & d_\infty > 1/\sqrt{\kappa}, \end{cases} \tag{53}$$

for some constant C_{tv} (that might strongly depend on κ) and for all $(q(0), p(0)) \in A \subset \mathcal{A}_f$, where A is compact. For any $\kappa \geq 1$, the right-hand side can be bounded by

$$|(q(t), p(t))| \leq \frac{C_{tv}|(q(0), p(0))|}{1 + d(0)t}, \tag{54}$$

which implies that the convergence rate is at least $\mathcal{O}(1/t)$ for any fixed function f that has non-degenerate isolated local minima (the minima can have an arbitrarily small curvature). We believe that this provides useful insights for the case where the isolated minimum at the origin is degenerate. For example, given a function with an isolated minimum that is degenerate, we can add a regularization of the type $\epsilon|x|^2/2$, where $\epsilon > 0$ is small. The bound (54) applies to the resulting regularized function, and implies a convergence rate of the order $\mathcal{O}(1/t)$.

---

1. We assume $h \neq d_\infty^2$. The case $h = d_\infty^2$ is obtained by considering the solutions of (46) either as the limit $h \downarrow d_\infty^2$ or the limit $h \uparrow d_\infty^2$.
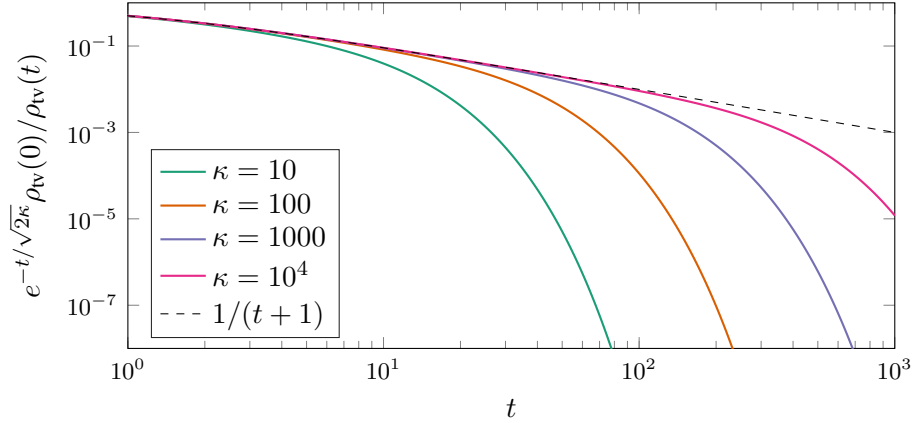
Figure 5: The plot shows the convergence rate of the fundamental solutions (52) as a function of κ for the choice $d_\infty = 1/\sqrt{2\kappa}$, $t_0 = 0$, $d(0) = 1$. For small values of t, the convergence rate is roughly $1/(1+t)$, whereas for larger values of t the convergence rate is approximately $\exp(-t/\sqrt{2\kappa})$.

It is important to notice that the results in this section are stated in terms of the distance between the current iterate and the optimizer, i.e., $|(q(t) - x^*, p(t))|$. In the case of smooth and convex functions, this yield the following bound on the function value:

$$f(q(t)) - f(x^*) \leq L/2|q(t) - x^*|^2 \leq \frac{C_{\text{tv}}^2}{(1+t)^2}|q(0) - x^*|^2,$$

provided that the algorithm is initialized with $p(0) = 0$. Thus, our analysis recovers the well-known optimal rate $\mathcal{O}(1/t^2)$ (in terms of function value) for smooth and convex functions.

3.3 Example 2

We consider the time-varying generalization of (13), where the parameters $d > 0$ and $\beta \geq 0$ are assumed to be time varying. We focus on the case where f is convex and β_k is chosen to be $\beta_k = T(1 - 2d_k T)$, as this simplifies the stability analysis and enables explicit closed-form evaluation of certain expressions. The stability analysis carried out in Appendix E applies likewise to the time-varying case; it suffices to replace d with d_k and β with β_k. This concludes that the origin is asymptotically stable as long as $0 < d_k T < 1$, $0 < T \leq 1/\sqrt{L}$.

After a change of coordinates that diagonalizes H_e, the linearized dynamics of a single coordinate are given by

$$\delta q(k+1) = \delta q(k) + T\delta p(k+1), \qquad \delta p(k+1) = \delta p(k) - T(2d_k + \beta_k h)\delta p(k) - Th\delta q(k), \quad (55)$$

where $\delta q(k) \in \mathbb{R}$, $\delta p(k) \in \mathbb{R}$, and h denotes the corresponding eigenvalue of H_e with upper and lower bounds $1/\kappa \leq h \leq 1$. In analogy to the transformation (47) in the continuous-time case, we define

$$\left(-\frac{1}{2}\right)^{k_0+1}(1 - hT^2)^{k-k_0-2}\left(\prod_{j=k_0+1}^{k-1}(1 - d_j T)\right)\delta\tilde{q}(k) := \delta q(k), \quad (56)$$

which transforms (55) into

$$\delta\tilde{q}(k+2) - 2\delta\tilde{q}(k+1) + \frac{4\beta_{k+1}/T}{(1-hT^2)(1+\beta_{k+1}/T)(1+\beta_k/T)}\delta\tilde{q}(k) = 0. \tag{57}$$

This transformation is well known in the literature on linear difference equations (see, for example, Elaydi, 2005, p. 369). As in the continuous-time case, the analysis is considerably simplified if β_k is chosen in such a way that the coefficient multiplying $\delta\tilde{q}(k)$ in (57) remains constant. This can be achieved, for example, with

$$4\beta_{k+1}/T = \alpha_{\text{tv}}(1+\beta_{k+1}/T)(1+\beta_k/T), \tag{58}$$

where α_{tv} is constant and independent of h. This equation defines a recurrence relation for β_k, and through $\beta_k = T(1 - 2d_kT)$ a relation for d_k. The choice (58) is motivated by the fact that 1) the recurrence relation is independent of h and 2) the difference equations (57) is time-invariant, which enables closed-form solutions. Thus, any trajectory $\delta q(k)$ satisfying (55) is a linear combination of the two fundamental solutions

$$(1 - hT^2 \pm \sqrt{(1-hT^2)^2 - \alpha(1-hT^2)})^{k-k_0} \prod_{j=k_0+1}^{k-1} (1 - d_jT). \tag{59}$$

This explicitly characterizes the convergence rate of (55) and, by virtue of Proposition 12, the convergence rate of (13), (for time-varying d_k and β_k). Similar to the continuous-time discussion in Section 3.2, the convergence rate can be improved by choosing d_k to be close to one for small k. For the specific choice $T = 0.5$, $\beta_0 = 0$, $d_0 = 1$, $\lim_{k\to\infty} d_k = d_\infty = 1/\sqrt{2\kappa}$, $L = 1$, the dependence of the convergence rate on κ is illustrated in Figure 6. As in the continuous-time case, the introduction of the time-varying parameter d_k ensures that the fundamental solutions converge approximately with $1/(1 + d_0Tk)$ for small k, whereas for large k the convergence is linear with rate $1 - T/\sqrt{2\kappa}$. The transition occurs for $k \sim \sqrt{2\kappa}/T$. This indicates that even for very large κ, the fundamental solutions converge roughly with $1/(1 + d_0Tk)$.

4. Conclusions

We have presented a convergence-rate analysis of momentum-based optimization algorithms from a dynamical systems point of view. Our analysis emphasizes the importance of the curvature properties about an isolated local minimum, which, up to a multiplicative constant, dictate the convergence rate. In addition, we find that reducing the damping over time improves the convergence rate by sublinear terms, which is important for objective functions that have minima with almost vanishing curvature. The use of momentum ensures robustness of the convergence rate against small changes in the curvature and leads, in many cases, to acceleration.

We also provided a rigorous motivation for the use of symplectic discretization schemes, showing that they enable the computation of a modified energy function that is (almost exactly) preserved by the conservative parts of the dynamics. The modified energy function provides a means for stability analysis, which implies, for example, that certain heavy-ball-type methods are in fact accelerated. Thus, an evaluation of the gradient at a shifted position is not necessary for achieving convergence rates that scale with $1/\sqrt{\kappa}$ for large κ.
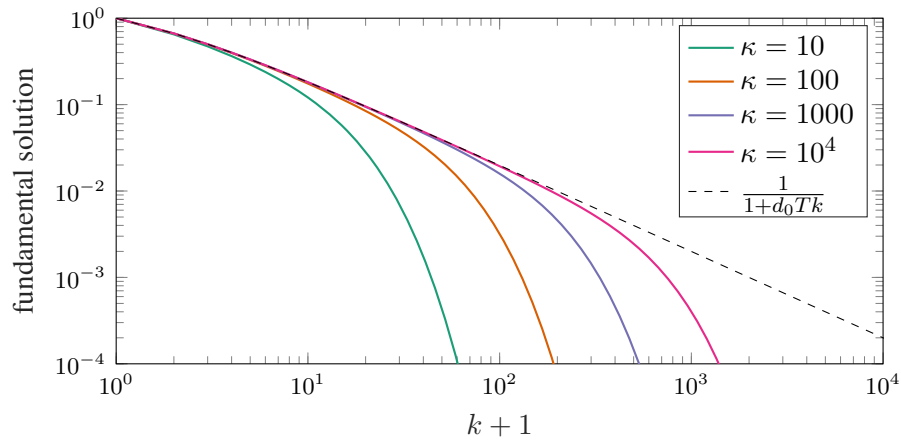
Figure 6: The plot shows the convergence rate of the fundamental solutions (59) as a function of κ. For small values of k, the convergence rate is roughly $1/(1 + d_0 T k)$, whereas for larger values of k the convergence is linear with rate $1 - T/\sqrt{2\kappa}$.

Our analysis emphasizes fundamental similarities between convergence rate analysis in continuous time and discrete time, and provides intuitive and rigorous explanations for various phenomena encountered in optimization, without resorting to convexity.

Acknowledgments

Appendix A. Proof of Proposition 3

For proving the proposition we rely on the following lemmas.

Lemma 13 *Consider the trajectories of a time-varying dynamical system*

$$x^+(t) = (A + B(t))x(t), \quad \forall t \in I, \tag{60}$$

with $A \in \mathbb{R}^{2n \times 2n}$ and $B : I \to \mathbb{R}^{2n \times 2n}$, where B is either continuous (in the continuous-time case) or $A + B(t)$ is invertible (in the discrete-time case). Let A be such that the trajectories $w^+(t) = Aw(t)$ converge exponentially with rate α, i.e., $|w(t)| \leq C_w|w(0)|\exp(-\alpha t)$, for all $w(0) \in \mathbb{R}^{2n}$, $t \in I$, and for some constant $C_w \geq 1$. Then, $x(t)$ satisfies the estimate

$$|x(t)| \leq C_w|x(0)|\exp\left(-\alpha t + C_w \int_0^t |B(\tau)|\mathrm{d}\tau\right), \qquad resp. \tag{61}$$

$$|x(t)| \leq C_w|x(0)|\exp\left(-\alpha t + C_w \exp(\alpha) \sum_{\tau=0}^{t-1} |B(\tau)|\right). \tag{62}$$

Proof We consider the continuous-time case first. Due to the continuity assumptions on $B(t)$, the dynamics are guaranteed to have a unique solution, $x(t)$, satisfying

$$x(t) = \exp(At)x(0) + \int_0^t \exp(A(t-\tau))B(\tau)x(\tau)\mathrm{d}\tau, \tag{63}$$

where $\exp$ denotes the matrix exponential. The two-norm of $\exp(A(t-\tau))$ is, by assumption, upper bounded by $C_w \exp(-\alpha(t-\tau))$, for all $t, \tau \in I$, $t \geq \tau$, yielding the following estimate:

$$\exp(\alpha t)|x(t)| \leq C_w|x(0)| + C_w \int_0^t |B(\tau)|\exp(\alpha\tau)|x(\tau)|\mathrm{d}\tau. \tag{64}$$

Applying the Grönwall inequality to (64) yields the desired result.

The discrete-time case is analogous. Due to the assumption of $A + B(t)$ being full rank, the dynamics are guaranteed to have a unique solution $x(t)$, which satisfies

$$x(t) = A^t x(0) + \sum_{\tau=1}^{t} A^{t-\tau}B(\tau-1)x(\tau-1); \tag{65}$$

see, for example, Agarwal (2000, p. 59). By assumption, the two-norm of $A^{t-\tau}$ is bounded by $C_w \exp(-\alpha(t-\tau))$, for all $t, \tau \in I$, $t \geq \tau$. As a result, the following estimate is obtained:

$$\exp(\alpha t)|x(t)| \leq C_w|x(0)| + C_w \sum_{\tau=1}^{t} \exp(\alpha\tau)|B(\tau-1)||x(\tau-1)| \tag{66}$$

$$= C_w|x(0)| + C_w \exp(\alpha) \sum_{\tau=0}^{t-1} |B(\tau)|\exp(\alpha\tau)|x(\tau)|. \tag{67}$$

Applying the Grönwall inequality (Agarwal, 2000, p. 186), yields

$$|x(t)| \leq C_{\mathrm{w}}|x(0)| \exp(-\alpha t) \prod_{\tau=0}^{t-1} (1 + C_{\mathrm{w}} \exp(\alpha)|B(\tau)|) \tag{68}$$

$$\leq C_{\mathrm{w}}|x(0)| \exp\left(-\alpha t + C_{\mathrm{w}} \exp(\alpha) \sum_{\tau=0}^{t-1} |B(\tau)|\right). \tag{69}$$

∎

Lemma 14 *Let Assumption 4 be fulfilled. Then there exists an open ball B_δ (with radius $\delta > 0$) centered at the origin and a constant $\tilde{C} \geq 1$, such that for all $z_0 \in B_\delta$,*

$$|\varphi_t(z_0)| \leq \tilde{C}|z_0| \exp(-\alpha t), \quad \forall t \in I. \tag{70}$$

Proof We consider the continuous-time case first. Assumption 4 implies that the dynamics (4) and (5) are (asymptotically) stable. This implies that for any $\varepsilon > 0$ there exists a constant $\delta(\varepsilon) > 0$ such that $|\varphi_t(z_0)| < \varepsilon$ for all $z_0 \in \mathbb{R}^{2n}$ with $|z_0| < \delta(\varepsilon)$. For any $\varepsilon > 0$, we consider the trajectory starting at $z_0 \in \mathbb{R}^{2n}$, with $|z_0| < \delta(\varepsilon)$ and reformulate the dynamics by applying the mean value theorem:

$$z^+(t) = \left.\frac{\partial g}{\partial z}\right|_{z=\xi(t)} z(t) = \left(\left.\frac{\partial g}{\partial z}\right|_{z=0} + \left.\frac{\partial g}{\partial z}\right|_{z=\xi(t)} - \left.\frac{\partial g}{\partial z}\right|_{z=0}\right) z(t), \tag{71}$$

where $\xi(t) \in \mathbb{R}^{2n}$ lies between the origin and $z(t)$. The continuity assumption on the dynamics leads to the following estimate:

$$\left|\left.\frac{\partial g}{\partial z}\right|_{z=\xi(t)} - \left.\frac{\partial g}{\partial z}\right|_{z=0}\right| \leq C_{\mathrm{g}}|\xi(t)| \leq C_{\mathrm{g}}|z(t)| < C_{\mathrm{g}}\varepsilon, \tag{72}$$

for all $t \in I$ and where C_{g} is a Lipschitz constant of $\partial g/\partial z$ in a neighborhood of the origin.

Applying Lemma 13 then yields

$$|z(t)| \leq C_l|z_0| \exp(-\alpha t + C_l C_{\mathrm{g}} \varepsilon t), \quad \forall t \in I. \tag{73}$$

We fix $\varepsilon > 0$ such that $C_l C_{\mathrm{g}} \varepsilon < \alpha/2$. This leads, in turn, to a refinement of the estimate (72),

$$\left|\left.\frac{\partial g}{\partial z}\right|_{z=\xi(t)} - \left.\frac{\partial g}{\partial z}\right|_{z=0}\right| \leq C_l C_{\mathrm{g}} \delta(\varepsilon) \exp(-\alpha t/2) < \frac{\alpha}{2} \exp(-\alpha t/2), \quad \forall t \in I, \tag{74}$$

where we have used the fact that $\delta(\varepsilon) \leq \varepsilon$. This results in

$$\int_0^t \left|\left.\frac{\partial g}{\partial z}\right|_{z=\xi(\tau)} - \left.\frac{\partial g}{\partial z}\right|_{z=0}\right| \mathrm{d}\tau < \frac{\alpha}{2} \int_0^\infty \exp(-\alpha\tau/2)\mathrm{d}\tau = 1, \quad \forall t \in I. \tag{75}$$

Applying Lemma 13 once more therefore implies

$$|z(t)| \leq \underbrace{C_l e^{C_l}}_{:=\tilde{C}} |z_0| \exp(-\alpha t), \quad \forall t \in I, \tag{76}$$

for all $z_0 \in \mathbb{R}^{2n}$ with $|z_0| < \delta(\varepsilon)$, which leads to the desired result.

The same arguments (with slightly modified constants) apply to the discrete-time case. ∎

Appendix B. Proof of Proposition 12

For proving the proposition we rely on the following lemmas.

Lemma 15 *Consider the trajectories of a time-varying dynamical system*

$$x^+(t) = (A(t) + B(t))x(t), \quad \forall t \in I, \tag{77}$$

with $A : I \to \mathbb{R}^{2n \times 2n}$ and $B : I \to \mathbb{R}^{2n \times 2n}$, where A and B are either continuous (in the continuous-time case) or $A(t) + B(t)$ is invertible for all $t \in I$ (in the discrete-time case). Let $\phi(t, \tau) \in \mathbb{R}^{2n \times 2n}$, $t, \tau \in I$, $t \geq \tau$, be defined by $\phi(t, \tau)^+ = A(t)\phi(t, \tau)$ (for a fixed τ) and $\phi(\tau, \tau) = I$, and assume $\phi(t, \tau)$ satisfies

$$|\phi(t, \tau)| \leq C_\phi \frac{\rho(t)}{\rho(\tau)},$$

where $C_\phi \geq 1$ is constant, $\rho : I \to \mathbb{R}_{\geq 0}$ is continuous, monotonically decreasing, and $\lim_{t \to \infty} \rho(t) = 0$. Then, $x(t)$ satisfies the estimate

$$|x(t)| \leq C_\phi |x(\tau)| \frac{\rho(t)}{\rho(\tau)} \exp\left(C_\phi \int_\tau^t |B(s)| \mathrm{d}s \right), \qquad resp. \tag{78}$$

$$|x(t)| \leq C_\phi |x(\tau)| \frac{\rho(t)}{\rho(\tau)} \exp\left(C_\phi \sum_{j=\tau}^{t-1} \frac{\rho(j)}{\rho(j+1)} |B(j)| \right), \quad \forall t, \tau \in I, t \geq \tau. \tag{79}$$

Proof We consider the continuous-time case first. Due to the continuity assumptions on $A(t)$ and $B(t)$, the dynamics are guaranteed to have a unique solution $x(t)$, satisfying

$$x(t) = \phi(t, \tau)x(\tau) + \int_\tau^t \phi(t, s)B(s)x(s)\mathrm{d}s. \tag{80}$$

The bound on the fundamental solution matrix $\phi(t, \tau)$ yields the following estimate:

$$\frac{1}{\rho(t)}|x(t)| \leq C_\phi \frac{|x(\tau)|}{\rho(\tau)} + C_\phi \int_\tau^t |B(s)| \frac{|x(s)|}{\rho(s)} \mathrm{d}s. \tag{81}$$

Applying the Grönwall inequality to (81) yields the desired result.

The discrete-time case is analogous. Due to the assumption of $A(t) + B(t)$ being full rank, the dynamics are guaranteed to have a unique solution $x(t)$, which satisfies

$$x(t) = \phi(t, \tau)x(\tau) + \sum_{j=\tau+1}^{t} \phi(t, j)B(j-1)x(j-1); \tag{82}$$

see, for example, (Agarwal, 2000, p. 59). As a result, the bound on $\phi(t, \tau)$ leads to the following estimate

$$\frac{1}{\rho(t)}|x(t)| \leq C_\phi \frac{|x(\tau)|}{\rho(\tau)} + C_\phi \sum_{j=\tau+1}^{t} \frac{\rho(j-1)}{\rho(j)} |B(j-1)| \frac{|x(j-1)|}{\rho(j-1)}. \tag{83}$$

Applying the Grönwall inequality (Agarwal, 2000, p. 186), yields

$$|x(t)| \le C_\phi |x(\tau)| \frac{\rho(t)}{\rho(\tau)} \prod_{j=\tau}^{t-1} (1 + C_\phi \frac{\rho(j)}{\rho(j+1)} |B(j)|) \tag{84}$$

$$\le C_\phi |x(\tau)| \frac{\rho(t)}{\rho(\tau)} \exp \left(C_\phi \sum_{j=\tau}^{t-1} \frac{\rho(j)}{\rho(j+1)} |B(j)| \right), \tag{85}$$

where the fact that $1 + x \le e^x$, for all $x \ge 0$ has been used. ∎

Lemma 16 *Let Assumption 6 be fulfilled. Then there exists an open ball B_δ (with radius $\delta > 0$) centered at the origin and a constant $\tilde{C} \ge 1$, such that for all $z_0 \in B_\delta$, and all $\tau \in I$,*

$$|\varphi_t(z_0, \tau)| \le \tilde{C} |z_0| \frac{\rho(t)}{\rho(\tau)} e^{-\alpha(t-\tau)}, \quad \forall t \in I, t \ge \tau. \tag{86}$$

Proof We consider the continuous-time case first and fix the initial time $\tau \in I$. Assumption 6 implies that the dynamics (37) are (asymptotically) stable, uniformly in τ. This implies that for any $\varepsilon > 0$ there exists a constant $\delta(\varepsilon) > 0$ (independent of τ) such that $|\varphi_t(z_0, \tau)| < \varepsilon$ for all $t \ge \tau$ and for all $z_0 \in \mathbb{R}^{2n}$ with $|z_0| < \delta(\varepsilon)$. For any $\varepsilon > 0$, we consider the trajectory starting at $z_0 \in \mathbb{R}^{2n}$ at time $\tau \in I$, with $|z_0| < \delta(\varepsilon)$ and reformulate the dynamics by applying the mean value theorem:

$$z^+(t) = \left. \frac{\partial g}{\partial z} \right|_{t, z = \xi(t)} z(t) = \left(\left. \frac{\partial g}{\partial z} \right|_{t, z = 0} + \left. \frac{\partial g}{\partial z} \right|_{t, z = \xi(t)} - \left. \frac{\partial g}{\partial z} \right|_{t, z = 0} \right) z(t), \tag{87}$$

where $\xi(t) \in \mathbb{R}^{2n}$ lies between the origin and $z(t)$. The assumptions on the dynamics, leads to the following estimate:

$$\left| \left. \frac{\partial g}{\partial z} \right|_{t, z = \xi(t)} - \left. \frac{\partial g}{\partial z} \right|_{t, z = 0} \right| \le C_{\mathrm{g}} |\xi(t)| \le C_{\mathrm{g}} |z(t)| < C_{\mathrm{g}} \varepsilon, \tag{88}$$

for all $t \in I$, $t \ge \tau$, and where C_{g} is a time-independent Lipschitz constant of $\partial g / \partial z$ in a neighborhood of the origin.

By assumption, the linear dynamics decay at least with $\exp(-\alpha(t - \tau))$. Thus, applying Lemma 15 yields

$$|z(t)| \le C_l |z(\tau)| \exp(-\alpha(t - \tau) + C_l C_{\mathrm{g}} \varepsilon (t - \tau)), \quad \forall t \in I, \quad t \ge \tau. \tag{89}$$

We fix $\varepsilon > 0$ such that $C_l C_{\mathrm{g}} \varepsilon < \alpha/2$. This leads, in turn, to a refinement of the estimate (88),

$$\left| \left. \frac{\partial g}{\partial z} \right|_{t, z = \xi(t)} - \left. \frac{\partial g}{\partial z} \right|_{t, z = 0} \right| \le C_l C_{\mathrm{g}} \delta(\varepsilon) \exp(-\alpha(t - \tau)/2) < \frac{\alpha}{2} \exp(-\alpha(t - \tau)/2), \tag{90}$$

$\forall t \in I, t \ge \tau$, where we have used the fact that $\delta(\varepsilon) \le \varepsilon$. This results in

$$\int_\tau^t \left| \left. \frac{\partial g}{\partial z} \right|_{t = \hat{t}, z = \xi(\hat{t})} - \left. \frac{\partial g}{\partial z} \right|_{t = \hat{t}, z = 0} \right| \mathrm{d}\hat{t} < \frac{\alpha}{2} \int_\tau^\infty \exp(-\alpha(\hat{t} - \tau)/2) \mathrm{d}\hat{t} = 1, \quad \forall t \in I. \tag{91}$$

34

Applying Lemma 15 once more therefore implies

$$|z(t)| \leq \underbrace{C_1 e^{C_1}}_{:= \tilde{C}} |z(\tau)| \frac{\rho(t)}{\rho(\tau)} e^{-\alpha(t-\tau)}, \quad \forall t \in I, \quad t \geq \tau, \tag{92}$$

for all $z_0 \in \mathbb{R}^{2n}$ with $|z_0| < \delta(\varepsilon)$, which leads to the desired result.

The same arguments (with slightly modified constants) apply to the discrete-time case. ∎

Appendix C. Proof of Proposition 9

We prove the following generalization of Proposition 9:

Proposition 17 *Let f be analytic on B_r^c, the closed ball of radius r about the origin, and let L_H be a Lipschitz constant of ∇H, i.e., $|\nabla H(z)| \leq L_H|z|$ for all $z \in B_r^c \times B_r^c$. Then there exists a perturbed Hamiltonian $\tilde{H} : B_r^c \times B_r^c \to \mathbb{R}$, whose trajectories*

$$\dot{\tilde{q}}(t) = \frac{\partial \tilde{H}}{\partial p}^{\mathsf{T}}, \quad \dot{\tilde{p}}(t) = -\frac{\partial \tilde{H}}{\partial q}^{\mathsf{T}}, \quad (\tilde{q}(0), \tilde{p}(0)) = z_0, \tag{93}$$

are such that

$$|(\tilde{q}(T), \tilde{p}(T)) - \Phi_T(z_0)| \leq TC_E|z_0|e^{-T_0/T}, \tag{94}$$

for all $0 < T \leq T_0/3$ and all z_0 such that

$$|z_0| \leq \frac{r}{2}e^{-10.2L_HT}, \tag{95}$$

where T_0 and C_E are constants given by

$$T_0 := \frac{2ln(2) - 1}{2eL_H}, \quad C_E := L_H(4.36e + e^2 T_0/3).$$

The perturbed Hamiltonian has the form

$$\tilde{H}(q, p) = H(q, p) - \frac{T}{2}\nabla f(q)^{\mathsf{T}}p + T^2 F(q, p), \tag{96}$$

where $\tilde{H}$ and $F : B_r^c \times B_r^c \to \mathbb{R}$ satisfy

$$|\nabla H(z) - \nabla \tilde{H}(z)| \leq 15L_HT|\nabla H(z)|, \quad |\nabla F(z)| \leq C_F|\nabla H(z)|, \quad \forall z \in B_{r/2}^c \times B_{r/2}^c, \tag{97}$$

with $C_F := 356L_H^2$. Moreover, $\tilde{H}$ has the same critical points as H and satisfies

$$|H(z) - \tilde{H}(z)| \leq 15L_HT|H(z)|, \quad |F(z)| \leq C_F|H(z)|, \quad \forall z \in \mathcal{A}_f \cap (B_{r/2}^c \times B_{r/2}^c). \tag{98}$$

Finally, at time T, the difference of the perturbed energy is bounded by

$$|\tilde{H}(\tilde{q}(T), \tilde{p}(T)) - \tilde{H}(\Phi_T(z_0))| \leq TC_{\Delta\tilde{H}}|\nabla H(z_0)|^2 e^{-T_0/T}, \tag{99}$$

where

$$C_{\Delta\tilde{H}} := e(2.9 + 0.1T_0)(1 + eT_0/3), \quad |z_0| \leq \frac{r}{2}(1 + 3.63L_HT(1 + eT_0/3))^{-1}.$$

Proof We will complexify the position and momentum variables; i.e., we take $q \in \mathbb{C}^n$, $p \in \mathbb{C}^n$. Due to the fact that f is analytic, it can be interpreted as a mapping from $\mathbb{C}^n$ to $\mathbb{C}$; a similar reasoning applies to H. Let $B_\rho^c \subset \mathbb{C}^{2n}$ be the closed ball of radius $\rho > 0$, centered about the origin, and let the norm $|| \cdot ||_\rho$ be defined as

$$||\bar{g}||_\rho := \sup_{z \in B_\rho^c} |\bar{g}(z)|, \tag{100}$$

for any analytic function $\bar{g} : \mathbb{C}^{2n} \to \mathbb{C}^{2n}$.

We define $f_1(q, p) := \Omega \partial H / \partial z = (p, -\nabla f(q))$, where $\Omega \in \mathbb{R}^{2n \times 2n}$ denotes the standard symplectic matrix, and introduce the following Ansatz for the differential equation (93):

$$\dot{\tilde{z}}(t) = f_1(\tilde{z}(t)) + f_2(\tilde{z}(t))T + f_3(\tilde{z}(t))T^2 + \dots, \tag{101}$$

where $\tilde{z}(t) := (\tilde{q}(t), \tilde{p}(t))$ is a trajectory satisfying (93), and f_i are analytic functions, $i = 1, 2, \dots$. Performing a Taylor expansion of $\tilde{z}(t)$ about 0 and evaluating at $t = T$ yields, after rearranging terms,

$$\tilde{z}(T) = \tilde{z}(0) + f_1(\tilde{z}(0))T + (f_2(\tilde{z}(0)) + \frac{1}{2}D_1 f_1(\tilde{z}(0)))T^2 \tag{102}$$

$$+ (f_3(\tilde{z}(0)) + \frac{1}{3!}D_1^2 f_1(\tilde{z}(0)) + \frac{1}{2!}(D_2 f_1(\tilde{z}(0)) + D_1 f_2(\tilde{z}(0))))T^3 + \dots, \tag{103}$$

where the Lie derivative $D_i : C^\infty \to C^\infty$ is defined by

$$D_i \bar{g}(z) = \left. \frac{\partial \bar{g}}{\partial z} \right|_z f_i(z), \quad \forall \bar{g} \in C^\infty, \tag{104}$$

where C^∞ denotes the set of infinitely differentiable functions mapping from $\mathbb{C}^{2n}$ to $\mathbb{C}^{2n}$ (see also Hairer et al., 2002, Ch. IX). The notation on the left-hand side of the previous equation should be read in the following order: The operator D_i is applied to $\bar{g}$, yielding the function $D_i \bar{g}$, which is then evaluated at z. Applying the operator D_i to the function $D_i \bar{g}$ is denoted by $D_i^2 \bar{g}$.

We require $\tilde{z}(T) = (q_{k+1}, p_{k+1})$, which leads due to (26) to

$$\tilde{z}(T) = \tilde{z}(0) + f_1(\tilde{z}(0))T + \begin{pmatrix} -\nabla f(\tilde{q}(0)) \\ 0 \end{pmatrix} T^2. \tag{105}$$

Equating equal powers of T in (103) and (105) yields the following recursive scheme for computing the functions f_i:

$$f_2(z) = (-\nabla f(q), 0)^\mathsf{T} - \frac{1}{2!}D_1 f_1(z), \tag{106}$$

$$f_j(z) = -\sum_{i=2}^{j} \frac{1}{i!} \sum_{k_1 + \dots + k_i = j} D_{k_1} D_{k_2} \dots D_{k_{i-1}} f_{k_i}(z), \quad j > 2, \tag{107}$$

where the last sum ranges over all strictly positive integers $k_1, \dots, k_i$ that sum up to j. We will construct the perturbed Hamiltonian $\tilde{H}$ by appropriately truncating the series $\sum f_i(z)T^i$, and will show that the resulting truncated series has the desired properties.

We start by explicitly computing the function f_2:

$$f_2(z) = \begin{pmatrix} -\nabla f(q) \\ 0 \end{pmatrix} - \frac{1}{2} \begin{pmatrix} 0 & I \\ -\frac{d^2 f}{dx^2} & 0 \end{pmatrix} \begin{pmatrix} p \\ -\nabla f(q) \end{pmatrix} = \underbrace{\frac{1}{2} \begin{pmatrix} 0 & I \\ \frac{d^2 f}{dx^2} & 0 \end{pmatrix}}_{:=A_2(z)} f_1(z), \qquad (108)$$

and note that it is of the form $A_2(z)f_1(z)$, where $A_2(z) \in \mathbb{C}^{2n \times 2n}$. We further note that $f_2(z)$ is in fact a Hamiltonian vector field with corresponding Hamiltonian $-p^\mathsf{T} \nabla f(q)/2$. From an induction argument relying on (107) and the definition of the Lie derivative, it follows that each f_j can be written as $f_j(z) = A_j(z)f_1(z)$, where $A_1(z)$ is the identity, $A_2(z)$ is defined in (108), and

$$A_j(z) := -\sum_{i=2}^{j} \frac{1}{i!} \sum_{k_1 + \cdots + k_i = j} \frac{\partial}{\partial z} (D_{k_2} \dots D_{k_{i-1}} f_{k_i}) \Big|_z A_{k_1}(z), \quad j > 2. \qquad (109)$$

Moreover, as shown in Hairer et al. (2002, p. 295,Theorem 3.2), the vector fields f_j are guaranteed to be Hamiltonian for all $j \geq 1$.

We will rely on the assumption that f is analytic for bounding the functions A_j using Cauchy's integral formula. More precisely, Cauchy's integral formula provides us with the following bound on $\partial \bar{g}/\partial z$ for any function $\bar{g}$ analytic in B_r^{c}:

$$\left| \frac{\partial \bar{g}}{\partial z} \right|_z \leq \frac{||\bar{g}||_\rho}{\delta}, \quad \forall z \in B_{\rho-\delta}^{\mathrm{c}}, \qquad \left\| \frac{\partial \bar{g}}{\partial z} \right\|_{\rho-\delta} \leq \frac{||\bar{g}||_\rho}{\delta}, \qquad (110)$$

for any two constants ρ, δ such that $0 \leq \delta < \rho \leq r$. A similar argument yields the following bound on the Lie-derivative of a function $\bar{g}$ analytic in B_r^{c} (Hairer et al., 2002, p. 308),

$$||D_i \bar{g}||_\sigma \leq \frac{1}{\rho - \sigma} ||f_i||_\sigma ||\bar{g}||_\rho, \qquad (111)$$

where $0 \leq \sigma < \rho \leq r$. In the following these two inequalities are used to bound the right-hand side of (109).

Claim: The function A_j is bounded above by

$$||A_j||_{r/2} \leq \frac{\ln(2)}{2\ln(2) - 1} \left(\frac{2L_{\mathrm{H}}(j-1)}{2\ln(2) - 1} \right)^{j-1}, \qquad (112)$$

for all $j \geq 1$.

Proof of the claim: Explicit calculations show that the claim holds for $j = 1$. We thus fix the index $J > 1$ and estimate $||A_j||_{r-(j-1)\delta}$ for all $1 \leq j \leq J$, where $\delta = r/(2(J-1))$ is fixed, which, for $j = J$, yields the desired bound on A_J. To simplify notation we denote $|| \cdot ||_{r-(j-1)\delta}$ as $|| \cdot ||_j$. The right-hand side of (109) is upper bounded by

$$\left\| \frac{\partial}{\partial z} (D_{k_2} \dots D_{k_{i-1}} f_{k_i}) \Big|_z A_{k_1} \right\|_j \leq \frac{1}{\delta} ||D_{k_2} \dots D_{k_{i-1}} f_{k_i}||_{j-1} ||A_{k_1}||_j$$

$$\leq \frac{1}{\delta^2} ||D_{k_3} \dots D_{k_{i-1}} f_{k_i}||_{j-2} ||A_{k_2}||_{j-1} ||f_1||_{j-1} ||A_{k_1}||_j$$

$$\leq \frac{1}{\delta^{i-1}} ||A_{k_i}||_{j-(i-1)} ||f_1||_{j-(i-1)} \dots ||A_{k_2}||_{j-1} ||f_1||_{j-1} ||A_{k_1}||_j.$$

Due to the fact that $k_1 + \cdots + k_i = j$ and $k_1 > 0, \ldots k_i > 0$, it follows $k_i \leq j - (i - 1)$, which implies $||\bar{g}||_{j-(i-1)} \leq ||\bar{g}||_{k_i}$ for all $i \leq j$ and for any function $\bar{g}$ analytic on B_r^c. Therefore, the above bound reduces to

$$||\frac{\partial}{\partial z}(D_{k_2} \ldots D_{k_{i-1}} f_{k_i})\Big|_z A_{k_1}||_j \leq \frac{1}{\delta^{i-1}}||A_{k_i}||_{k_i}||f_1||_{k_i} \ldots ||A_{k_2}||_{k_2}||f_1||_{k_2}||A_{k_1}||_{k_1} \quad (113)$$

$$\leq \left(\frac{||f_1||_r}{\delta}\right)^{i-1} ||A_{k_i}||_{k_i} \ldots ||A_{k_1}||_{k_1}. \quad (114)$$

By introducing the constant $\hat{\delta} := \delta/r$, which, by definition of δ satisfies $\hat{\delta} \leq 1/2$, we can reformulate the above bound as

$$||\frac{\partial}{\partial z}(D_{k_2} \ldots D_{k_{i-1}} f_{k_i})\Big|_z A_{k_1}||_j \leq \left(\frac{L_H}{\hat{\delta}}\right)^{i-1} ||A_{k_i}||_{k_i} \ldots ||A_{k_1}||_{k_1}. \quad (115)$$

By exploiting the structure of $A_2(z)$ it can be verified that

$$||A_2||_2 = \frac{1}{2}\max\{1, ||d^2 f/dx^2||_2\} = \frac{1}{2}L_H \leq \frac{L_H}{2\hat{\delta}}, \quad (116)$$

which will simplify the following development. Following Hairer et al. (2002, p. 309), we introduce the constants b_j for all $j \geq 1$ in the following way

$$b_1 = \frac{L_H}{\hat{\delta}}, \quad b_2 = \frac{L_H^2}{\hat{\delta}^2}, \quad b_j = \sum_{i=2}^{j} \frac{1}{i!} \sum_{k_1 + \cdots + k_i = j} b_{k_1} b_{k_2} \ldots b_{k_i}, \quad j > 2.$$

It can be verified with an induction argument taking (115) and (116) into account, that

$$||A_j||_j \frac{L_H}{\hat{\delta}} \leq b_j, \quad 1 \leq j \leq J. \quad (117)$$

The variables b_j are well-defined for any $j \geq 1$ and not just $1 \leq j \leq J$. In order to bound the constants b_j, we formally introduce the generating function $b(\zeta)$, $b : \mathbb{C} \to \mathbb{C}$, $\zeta \in \mathbb{C}$, and note that b satisfies

$$b(\zeta) := \sum_{j=1}^{\infty} b_j \zeta^j = b_1 \zeta + \sum_{j=2}^{\infty} \zeta^j \sum_{i=2}^{j} \frac{1}{i!} \sum_{k_1 + \cdots + k_i = j} b_{k_1} \ldots b_{k_i} \quad (118)$$

$$= b_1 \zeta + \sum_{i=2}^{\infty} \frac{b(\zeta)^i}{i!} \quad (119)$$

$$= b_1 \zeta + e^{b(\zeta)} - 1 - b(\zeta), \quad (120)$$

where the order of summation has been interchanged to arrive at (118). It can be verified by means of the implicit function theorem that, given any $v \in \mathbb{C}$, the equation $2b - \exp(b) + 1 = v$ can be uniquely solved for b as long as $\exp(b) \neq 2$. This implies that $b(\zeta)$ is well-defined and analytic for $|b_1\zeta| \leq 2\ln(2) - 1$, or equivalently, $|\zeta| \leq (2\ln(2) - 1)/b_1$. Furthermore, explicit calculations show

that $|b(\zeta)|$ is bounded by $\ln(2)$ for $|\zeta| \leq (2\ln(2) - 1)/b_1$ (see Hairer et al., 2002, p. 310). Cauchy's inequality therefore implies

$$|b_j| \leq \frac{\ln(2)}{((2\ln(2) - 1)/b_1)^j} = \ln(2) \left(\frac{L_{\mathrm{H}}}{\hat{\delta}(2\ln(2) - 1)} \right)^j, \quad \forall j \geq 1. \tag{121}$$

Evaluating the above bound for $j = J$ yields

$$\|A_J\|_J \leq \frac{\ln(2)}{2\ln(2) - 1} \left(\frac{2L_{\mathrm{H}}(J-1)}{2\ln(2) - 1} \right)^{J-1}, \tag{122}$$

and proves the claim.

We define

$$\Omega \nabla \tilde{H}(z) := \sum_{j=1}^{N} f_j(z) T^{j-1}, \quad z \in B_{r/2}^{\mathrm{c}}, \tag{123}$$

where N is chosen to be the largest integer such that $NT \leq T_0$, with $T_0 := (2\ln(2) - 1)/(2eL_{\mathrm{H}})$. The choice for the integer N is motivated by the following observation: Due to the bound on $\|A_j\|_{r/2}$ the terms in the above sum behave as $(Tj/(eT_0))^j$, which attains a minimum for $j \approx T_0/T$. The fact that all f_j, $j \geq 1$, are Hamiltonian vector fields proves (96).

As a result, we obtain for $z \in B_{r/2}^{\mathrm{c}}$ and $T \leq T_0$ (implying $T/T_0 \leq 1/N$ by our choice of N)

$$|\nabla \tilde{H}(z) - \nabla H(z)| \leq \sum_{j=2}^{N} |f_j(z)| T^{j-1} \leq \frac{\ln(2)}{2\ln(2) - 1} |\nabla H(z)| \sum_{j=2}^{N} \left(\frac{T(j-1)}{eT_0} \right)^{j-1}$$

$$\leq \frac{\ln(2)}{2\ln(2) - 1} \frac{T}{T_0} |\nabla H(z)| \sum_{j=2}^{N} \left(\frac{j-1}{e} \right)^{j-1} \left(\frac{T}{T_0} \right)^{j-2}$$

$$\leq \frac{\ln(2)}{2\ln(2) - 1} \frac{T}{T_0} |\nabla H(z)| \underbrace{\sum_{j=1}^{N-1} (j/e)^j (1/N)^{j-1}}_{\leq 0.588}$$

$$\leq 15 L_{\mathrm{H}} T |\nabla H(z)|, \tag{124}$$

which proves the first bound in (97). The second bound is proved analogously; that is,

$$|\nabla F(z)| \leq \frac{1}{T^2} \sum_{j=3}^{N} |f_j(z)| T^{j-1} \tag{125}$$

$$\leq \frac{\ln(2)}{(2\ln(2) - 1)T_0^2} |\nabla H(z)| \sum_{j=3}^{N} \left(\frac{j-1}{e} \right)^{j-1} \left(\frac{T}{T_0} \right)^{j-3} \tag{126}$$

$$\leq \frac{\ln(2)}{(2\ln(2) - 1)T_0^2} |\nabla H(z)| \underbrace{\sum_{j=2}^{N-1} \left(\frac{j}{e} \right)^j \left(\frac{1}{N} \right)^{j-2}}_{\leq 1} \tag{127}$$

$$\leq 356 L_{\mathrm{H}}^2 |\nabla H(z)|. \tag{128}$$

The fact that $\tilde{H}$ has the same critical points is derived from the following observation. First, all critical points of H are necessarily critical points of $\tilde{H}$, which follows from (124) with $\nabla H(z) = 0$, and second, $\nabla\tilde{H}(\hat{z}) = 0$ implies, according to (124), that $|\nabla H(\hat{z})| \leq 15 L_\text{H} T |\nabla H(\hat{z})| \leq 0.36 |\nabla H(\hat{z})|$ (for $T \leq T_0/3$), concluding $\nabla H(\hat{z}) = 0$.

The first bound in (98) is derived by noticing that

$$\tilde{H}(z) - H(z) = \int_0^1 (\nabla\tilde{H}(\gamma(t)) - \nabla H(\gamma(t)))^\mathsf{T}\dot{\gamma}(t)\mathrm{d}t, \tag{129}$$

for any path $\gamma : [0,1] \to \mathbb{C}^{2n}$ that connects the origin with z; i.e., $\gamma(0) = 0$, $\gamma(1) = z$. We choose $\gamma(t)$ such that $\nabla H(\gamma(t))^\mathsf{T}\dot{\gamma}(t) = |\nabla H(\gamma(t))||\dot{\gamma}(t)|$. This can be done by considering the gradient flow $\tilde{\gamma}(t)$,

$$\dot{\tilde{\gamma}}(t) = -\nabla H(\tilde{\gamma}(t)), \quad \tilde{\gamma}(0) = z, \tag{130}$$

which is guaranteed to converge to the origin for any $z \in \mathcal{A}_f \cap (B_{r/2}^\text{c} \times B_{r/2}^\text{c})$. As a result, choosing $\gamma(t) = \tilde{\gamma}(1/t - 1)$ leads to

$$\begin{aligned}
|\tilde{H}(z) - H(z)| &\leq \int_0^1 |\nabla\tilde{H}(\gamma(t)) - \nabla H(\gamma(t))||\dot{\gamma}(t)|\mathrm{d}t \\
&\leq 15 L_\text{H} T \int_0^1 |\nabla H(\gamma(t))||\dot{\gamma}(t)|\mathrm{d}t \\
&= 15 L_\text{H} T \int_0^1 \nabla H(\gamma(t))^\mathsf{T}\dot{\gamma}(t)\mathrm{d}t = 15 L_\text{H} T H(z).
\end{aligned} \tag{131}$$

The same argument (with the same path) applies to $|\nabla F(z)|$, which yields the second bound in (98).

We show next that the solution of (93) is exponentially close to the iterate (26). We fix T, $z_0 \in B_{r/2}^\text{c}$ such that

$$|z_0| \leq re^{-10.2 L_\text{H} T}/2 \tag{132}$$

(the choice becomes evident), and as above, N to be the largest integer such that $NT \leq T_0$. For small enough $\zeta \in \mathbb{C}$, we denote the map from $\tilde{z}(0) = z_0$ to $\tilde{z}(t = \zeta)$, where $\tilde{z}(t)$ satisfies $\dot{\tilde{z}}(t) = \sum_{j=1}^N f_j(\tilde{z}(t))\zeta^j$, by $\tilde{\varphi}_\zeta : \mathbb{C}^{2n} \to \mathbb{C}^{2n}$. The symplectic Euler step $\Phi_\zeta(z_0)$ and $\tilde{\varphi}_\zeta(z_0)$ are both analytic in ζ and by the above construction, their Taylor expansion in ζ about $\zeta = 0$ agrees for the first N terms (z_0 is fixed). Thus, the function $\tilde{g}(\zeta)/\zeta^{N+1}$, where $\tilde{g}(\zeta) := \Phi_\zeta(z_0) - \tilde{\varphi}_\zeta(z_0)$ is analytic. Applying the maximum modulus principle implies

$$\frac{|\tilde{g}(T)|}{T^{N+1}} \leq \max_{|\zeta|\leq\epsilon} \frac{|\tilde{g}(\zeta)|}{|\epsilon|^{N+1}} = \max_{|\zeta|=\epsilon} \frac{|\tilde{g}(\zeta)|}{\epsilon^{N+1}} = \frac{1}{\epsilon^{N+1}} \max_{|\zeta|=\epsilon} |\tilde{g}(\zeta)|, \tag{133}$$

where ϵ is chosen as $\epsilon = eT_0/N$ (again, the choice becomes evident subsequently). It remains to bound $|\tilde{g}(\zeta)|$ for $|\zeta| = \epsilon$, which will be done by estimating $|\Phi_\zeta(z_0) - z_0|$ and $|\tilde{\varphi}_\zeta(z_0) - z_0|$ separately. For the first term we obtain (by definition of (105))

$$|\Phi_\zeta(z_0) - z_0| \leq |\zeta||f_1(z_0)|(1 + |\zeta|) \leq \epsilon|f_1(z_0)|(1 + \epsilon) \leq \epsilon(1 + eT_0/3)L_\text{H}|z_0|, \tag{134}$$

where $N \geq 3$ has been used for the last inequality. For bounding the second term we first derive a bound on $\nabla \tilde{H}(z)$, for all $z \in B_{r/2}^c$, similar to (124) (unlike (124) we have $\epsilon = eT_0/N$). We obtain

$$
\begin{aligned}
|\nabla \tilde{H}(z)| &\leq |\nabla H(z)| \left(1 + \sum_{j=2}^{N} ||A_j||_{r/2} \epsilon^{j-1} \right) \\
&\leq |\nabla H(z)| \left(1 + \frac{\ln(2)}{2\ln(2)-1} \sum_{j=2}^{N} \left(\frac{2L_{\mathrm{H}}\epsilon(j-1)}{2\ln(2)-1} \right)^{j-1} \right) \\
&= |\nabla H(z)| \left(1 + \frac{\ln(2)}{2\ln(2)-1} \sum_{j=1}^{N-1} (j/N)^j \right) \leq 2.8|\nabla H(z)|,
\end{aligned}
\tag{135}
$$

for all $z \in B_{r/2}^c$. The gradient of $\nabla \tilde{H}$ has therefore a Lipschitz constant of less than $2.8L_{\mathrm{H}}$ for $z \in B_{r/2}^c$, which implies that

$$
|\tilde{\varphi}_\epsilon(z_0)| \leq |z_0|e^{2.8L_{\mathrm{H}}\epsilon} \leq |z_0|e^{10.2L_{\mathrm{H}}T},
\tag{136}
$$

as long as the corresponding trajectory stays within $B_{r/2}^c$. Note that the last inequality stems from the fact that $T_0 < T(N+1)$ and thus $\epsilon < eT(N+1)/N \leq 4eT/3$. Due to the condition (132), the trajectory $\tilde{\varphi}_\epsilon(z_0)$ necessarily stays within $B_{r/2}^c$. Moreover, from $\epsilon \leq eT_0/3$ ($N \geq 3$ since $T \leq T_0/3$), it can be inferred that

$$
|\tilde{\varphi}_\epsilon(z_0)| \leq |z_0|e^{2.8eT_0L_{\mathrm{H}}/3} \leq 1.2|z_0|,
\tag{137}
$$

which implies that $|\tilde{\varphi}_\epsilon(z_0) - z_0| \leq 3.36\epsilon L_{\mathrm{H}}|z_0|$. Combined with (134), this leads to $|\tilde{g}(\zeta)| \leq \epsilon \tilde{C}_{\mathrm{g}}$ for $|\zeta| = \epsilon$ and $\tilde{C}_{\mathrm{g}} := (4.36 + eT_0/3)L_{\mathrm{H}}|z_0|$, which, according to (133), implies

$$
\begin{aligned}
|\tilde{g}(T)| &\leq \epsilon \tilde{C}_{\mathrm{g}} \left(\frac{T}{\epsilon} \right)^{N+1} \leq T\tilde{C}_{\mathrm{g}} \left(\frac{T}{\epsilon} \right)^{N} \leq T\tilde{C}_{\mathrm{g}} \left(\frac{NT}{eT_0} \right)^{N} \leq T\tilde{C}_{\mathrm{g}}e^{-N} \\
&= T\tilde{C}_{\mathrm{g}}ee^{-(N+1)} \leq T\tilde{C}_{\mathrm{g}}ee^{-T_0/T}.
\end{aligned}
\tag{138}
$$

The fact that $N \leq T_0/T < N+1$ has been used to obtain the last inequality. The above result justifies the choice $\epsilon = eT_0/N$.

We derive the bound on $\tilde{H}(\tilde{\varphi}_T(z_0)) - \tilde{H}(\Phi_T(z_0))$ in a similar way. Consider the analytic function $\hat{g}(\zeta) = \tilde{H}(\tilde{\varphi}_\zeta(z_0)) - \tilde{H}(\Phi_\zeta(z_0))$, where $z_0 \in B_{r/2}^c$, T, and N are fixed. The modified Hamiltonian $\tilde{H}$ is dependent on ζ and has the form

$$
\tilde{H}(z) = \sum_{j=1}^{N} H_j(z)\zeta^{j-1},
\tag{139}
$$

where $H_1 = H$, and $\Omega \nabla H_j(z) = f_j(z)$ for all j with $1 \leq j \leq N$. The function $\hat{g}(\zeta)/(\zeta^{N+1})$ is analytic, as can be seen, for example, by a Taylor expansion of $\tilde{H}$ in z about $z = \Phi_\zeta(z_0)$, where $\tilde{\varphi}_\zeta(z_0)$ and $\Phi_\zeta(z_0)$ agree up to Nth order. We invoke the maximum principle, similar to (133), which implies,

$$
\frac{|\hat{g}(T)|}{T^{N+1}} \leq \frac{1}{\epsilon^{N+1}} \max_{|\zeta|=\epsilon} |\hat{g}(\zeta)|,
\tag{140}
$$

and where ϵ is again set to $\epsilon = eT_0/N$. A bound for $\hat{g}(\zeta)$ is obtained in the following way (applying Taylor's theorem):

$$|\tilde{H}(\tilde{\varphi}_\zeta(z_0)) - \tilde{H}(\Phi_\zeta(z_0))| = |\tilde{H}(z_0) - \tilde{H}(\Phi_\zeta(z_0))| \tag{141}$$

$$\leq |\nabla\tilde{H}(z_0)||\Phi_\zeta(z_0) - z_0| + \frac{1}{2}\left|\frac{\partial^2\tilde{H}}{\partial z^2}\Big|_\eta\right||\Phi_\zeta(z_0) - z_0|^2, \tag{142}$$

where η lies between $\Phi_\zeta(z_0)$ and z_0. The gradient of $\tilde{H}$ is necessarily bounded by $|\nabla\tilde{H}(z)| \leq 2.8L_{\mathrm{H}}|z|$, which therefore implies that the Hessian of $\tilde{H}$ in the expression is bounded by $2.8L_{\mathrm{H}}$ provided that η is in $B^{\mathrm{c}}_{r/2}$. The bound (134) implies that this is the case for

$$|z_0| \leq \frac{r}{2}(1 + 4L_{\mathrm{H}}eT(1 + eT_0/3)/3)^{-1}, \tag{143}$$

where $\epsilon \leq eT_0/3$ and $\epsilon < 4eT/3$ have been used. Combining (142) with the bound from (134) yields

$$|\tilde{H}(\tilde{\varphi}_\zeta(z_0)) - \tilde{H}(\Phi_\zeta(z_0))| \leq 2.8\epsilon(1 + eT_0/3)|\nabla H(z_0)|^2(1 + L_{\mathrm{H}}\epsilon/2(1 + eT_0/3)) \tag{144}$$

$$\leq \epsilon|\nabla H(z_0)|^2(2.8 + 0.1(1 + eT_0/3))(1 + eT_0/3) \tag{145}$$

$$\leq \epsilon|\nabla H(z_0)|^2(2.9 + 0.1T_0)(1 + eT_0/3). \tag{146}$$

Applying the same chain of arguments as in (138) allows us to conclude:

$$|\hat{g}(T)| \leq (2.9 + 0.1T_0)(1 + eT_0/3)eTe^{-T_0/T}|\nabla H(z_0)|^2. \tag{147}$$

∎

Appendix D. Proof of Proposition 10

Proof We choose $r > 0$ such that $A \subset B^{\mathrm{c}}_r \times B^{\mathrm{c}}_r$, where B^{c}_r is the closed n-dimensional ball of radius r centered at the origin. We use the Stone-Weierstrass theorem (Rudin, 1976, p. 159) to approximate $\mathrm{d}^2 f/\mathrm{d}x^2$ on B^{c}_{2r} by $\mathrm{d}^2\tilde{f}/\mathrm{d}x^2$ such that $|\mathrm{d}^2\tilde{f}/\mathrm{d}x^2 - \mathrm{d}^2 f/\mathrm{d}^2 x| \leq \epsilon/(2r)$ for all $x \in B^{\mathrm{c}}_{2r}$ and

$$\frac{\mathrm{d}^2\tilde{f}}{\mathrm{d}x^2}\Big|_{x=0} = \frac{\mathrm{d}^2 f}{\mathrm{d}x^2}\Big|_{x=0},$$

where $\epsilon > 0$ will be chosen subsequently. We integrate $\mathrm{d}^2\tilde{f}/\mathrm{d}x^2$ and impose $\nabla\tilde{f}(0) = 0$, which yields the function $\nabla\tilde{f}$ on B^{c}_{2r} that is ϵ-close to ∇f. We integrate once more and impose $\tilde{f}(0) = 0$. Moreover, by choosing ϵ small enough, the origin is guaranteed to be the only critical point of $H(z)$ (with f replaced with $\tilde{f}$) for all $z \in A$, since $A \subset \mathcal{A}_f$ and $A \subset B^{\mathrm{c}}_r \times B^{\mathrm{c}}_r$. Due to the fact that $\tilde{f}$ is a polynomial and therefore analytic, we can invoke Proposition 9 with f replaced by $\tilde{f}$. Next, we will show that the origin is an asymptotically stable equilibrium of the dynamics (13), where f is replaced by $\tilde{f}$, with region of attraction at least A.

42

In order to simplify notation, we reformulate the contraction step $\Phi_{d,T}$ (with f replaced by $\tilde{f}$) in the following way:

$$\Phi_{d,T}(q_k, p_k) = \begin{pmatrix} q_k \\ (I - T\Lambda(q_k, p_k))p_k \end{pmatrix}, \tag{148}$$

where by assumption $\Lambda(q_k, p_k)$ is symmetric and positive definite. Due to the assumption on the dissipative forces, the singular values of $\Lambda(q_k, p_k)$ are upper and lower bounded by d_2 and d_1, respectively. Provided that the time step T is small enough, $T \leq T_{\max}$, we conclude from Proposition 9 that there exists a modified energy function $\tilde{H}$ that has the same critical points as H, and such that the bound (27) holds for all $z_0 \in B_r^c \times B_r^c$. Morse theory implies that $\tilde{H}(z)$ is locally positive definite and has compact level sets for $z \in A$. The same applies therefore to the function

$$V(q_k, p_k) := \tilde{H}(q_k, p_k) + \frac{Td_1}{2}\nabla\tilde{f}(q_k)^\mathsf{T}p_k, \tag{149}$$

provided that $T \leq T_{\max}$, where $T_{\max}$ is chosen to be small enough (independently of f_d; this can be done since the positive definiteness of $\tilde{H}$ is not affected by the dissipative forces f_d).

The stability analysis simplifies when analyzing $\Phi_{d,T} \circ \Phi_T$ instead of $\Phi_T \circ \Phi_{d,T}$, due to the simple structure of $\Phi_{d,T}$. In order to simplify the notation we hide all constants that may depend on d_2/d_1, an upper bound on d_2, L_H, or higher orders of T with the notation $\mathcal{O}(\cdot)$. It will be shown that $V(q_k, p_k)$ necessarily decays along the step $\Phi_{d,T} \circ \Phi_T$. To that extent, we define $z_2 = \Phi_{d,T}(z_1)$ and $z_1 = \Phi_T(z_0)$ and consider $V(z_2) - V(z_0)$, which, due to the fact that $q_2 = q_1$, is bounded by

$$
\begin{aligned}
V(q_2, p_2) &= \frac{1}{2}|p_2|^2 - \frac{T}{2}\nabla\tilde{f}(q_2)^\mathsf{T}(I - d_1 I)p_2 + \tilde{f}(q_2) + T^2 F(q_2, p_2) \\
&\leq \frac{1}{2}|p_1|^2 - Tp_1^\mathsf{T}\Lambda p_1 - \frac{T}{2}\nabla\tilde{f}(q_1)^\mathsf{T}(I - d_1 I)p_1 + \tilde{f}(q_1) + T^2 F(q_1, p_2) \\
&\quad + \frac{T^2}{2}\nabla\tilde{f}(q_1)^\mathsf{T}(I - d_1 I)\Lambda p_1 + \mathcal{O}(d_1^2 T^2)|p_1|^2, \\
&\leq \tilde{H}(q_1, p_1) - Tp_1^\mathsf{T}\Lambda p_1 + \frac{Td_1}{2}\nabla\tilde{f}(q_1)^\mathsf{T}p_1 \\
&\quad + \mathcal{O}(d_1 T^2)|p_1|^2 + \mathcal{O}(d_1 T^2)|\nabla\tilde{f}(q_1)||p_1|,
\end{aligned}
$$

where the arguments of Λ have been dropped and the results from Proposition 9 have been used, which enable the following bound:

$$
\begin{aligned}
F(q_2, p_2) - F(q_1, p_1) &\leq \nabla F(q_1, p_1)^\mathsf{T}(p_2 - p_1) + \frac{1}{2}|p_2 - p_1|^2 \sup_{z \in B_r}|\frac{\mathrm{d}^2 F}{\mathrm{d}z^2}| \\
&\leq 356 L_H^2 T|\nabla H(q_1, p_1)||\Lambda p_1| + 178 L_H^3 T^2|\Lambda p_1|^2 \\
&\leq \mathcal{O}(d_1 T)(|p_1|^2 + |p_1||\nabla\tilde{f}(q_1)|) + \mathcal{O}(T^2 d_1^2)|p_1|^2.
\end{aligned}
$$

Due to Proposition 9, it follows that $\tilde{H}(\Phi_T(z_0)) - \tilde{H}(z_0)$ is bounded by a term of the order $|\nabla H(z_0)|^2 Te^{-T_0/T}$. This yields

$$
\begin{aligned}
V(q_2, p_2) &\leq \tilde{H}(q_0, p_0) - Tp_1^\mathsf{T}\Lambda p_1 + \frac{Td_1}{2}\nabla\tilde{f}(q_1)^\mathsf{T}p_1 + |\nabla H(z_0)|^2\mathcal{O}(Te^{-T_0/T}) \\
&\quad + \mathcal{O}(d_1 T^2)|p_1|^2 + \mathcal{O}(d_1 T^2)|\nabla\tilde{f}(q_1)||p_1| \\
&\leq V(q_0, p_0) - \frac{Td_1}{2}\nabla\tilde{f}(q_0)^\mathsf{T}p_0 - Tp_1^\mathsf{T}\Lambda p_1 + \frac{Td_1}{2}\nabla\tilde{f}(q_1)^\mathsf{T}p_1 + |\nabla H(z_0)|^2\mathcal{O}(Te^{-T_0/T}) \\
&\quad + \mathcal{O}(d_1 T^2)|p_1|^2 + \mathcal{O}(d_1 T^2)|\nabla\tilde{f}(q_1)||p_1|.
\end{aligned}
$$

Due to the continuity of $\nabla\tilde{f}$ we can bound $|\nabla\tilde{f}(q_1) - \nabla\tilde{f}(q_0)|$ by $T(L+\epsilon)|p_1|$, where L is the Lipschitz constant of ∇f. Moreover, L_H is an upper bound on L. This implies

$$
\begin{aligned}
V(q_2, p_2) \leq{}& V(q_0, p_0) - Tp_1^\mathsf{T}\Lambda p_1 - \frac{T^2 d_1}{2}|\nabla\tilde{f}(q_0)|^2 + |\nabla H(z_0)|^2\mathcal{O}(Te^{-T_0/T}) + \mathcal{O}(d_1 T^2)|p_1|^2 \\
&+ \mathcal{O}(d_1 T^2)|\nabla\tilde{f}(q_1)||p_1|, \\
\leq{}& V(q_0, p_0) - \frac{T}{2}p_1^\mathsf{T}\Lambda p_1 - \frac{T^2 d_1}{4}|\nabla\tilde{f}(q_0)|^2 + |\nabla H(z_0)|^2\mathcal{O}(Te^{-T_0/T}) \\
&+ \mathcal{O}(d_1 T^2)|\nabla\tilde{f}(q_0)||p_0|,
\end{aligned}
$$

provided that T is chosen small enough such that $-p_1^\mathsf{T}\Lambda p_1 T/2$ dominates the $\mathcal{O}(d_1 T^2)|p_1|^2$ terms and $-T^2 d_1|\nabla\tilde{f}(q_0)|^2/4$ dominates the $\mathcal{O}(d_1 T^3)|\nabla\tilde{f}(q_0)|^2$ terms. Expanding the term $-Tp_1^\mathsf{T}\Lambda p_1/2$ and applying Young's inequality,

$$
C_Y d_1 T^2 p_0^\mathsf{T}\nabla\tilde{f}(q_0) \leq \left(\frac{\sqrt{T}}{\sqrt{2}}\sqrt{d_1}|p_0|\right)\left(C_Y\sqrt{T}^3\sqrt{2}\sqrt{d_1}|\nabla\tilde{f}(q_0)|\right) \leq \frac{T}{4}d_1|p_0|^2 + C_Y^2 T^3 d_1|\nabla\tilde{f}(q_0)|^2,
$$

where $C_Y > 0$ is an arbitrary constant, results in

$$
\begin{aligned}
V(q_2, p_2) - V(q_0, p_0) \leq{}& -\frac{T}{4}d_1|p_0|^2 - \frac{T^2 d_1}{8}|\nabla\tilde{f}(q_0)|^2 + |\nabla H(z_0)|^2\mathcal{O}(Te^{-T_0/T}). \\
\leq{}& -\frac{d_1 T^2}{8}|\nabla H(z_0)|^2 + |\nabla H(z_0)|^2\mathcal{O}(Te^{-T_0/T}),
\end{aligned}
$$

where for the last inequality $T \leq 1$ has been used. Note that the term $Te^{-T_0/T}$ decays quickly for small T. Thus for $d_1 \geq \mathcal{O}(e^{-T_0/T}/T)$ the function V strictly decreases along the trajectories of $\Phi_{d,T} \circ \Phi_T$. The lower bound on d_1 vanishes exponentially for small T. This concludes that the origin is asymptotically stable under the dynamics (13) (with $\tilde{f}$ instead of f) provided that T is chosen small enough, i.e. $T \leq T_{\max}$, where up to the exponential terms in d_1, $T_{\max}$ only depends on the ratio d_2/d_1, and L_H. Due to the fact that the first-order approximations of the dynamics resulting from f and $\tilde{f}$ about the equilibrium are ϵ-close (by construction of $\tilde{f}$), the origin is likewise asymptotically stable for the dynamics resulting from f, and its region of attraction contains at least a small neighborhood of the origin. Consequently, due to the continuity of V, we may choose ϵ small enough such that $V(\tilde{q}_2, \tilde{p}_2) - V(q_0, p_0) < 0$ for all (q_0, p_0) outside a neighborhood of the origin, where $\tilde{q}_2, \tilde{p}_2$ denotes the trajectory resulting from (13) (with f instead of $\tilde{f}$). This shows that A is likewise contained in the region of attraction of the origin under the dynamics (13) with f instead of $\tilde{f}$. $\blacksquare$

Appendix E. Discrete-time stability (convex case)

By using the change of variables $(q_k, p_k) \to (\hat{q}_k, \hat{p}_k)$,

$$
\hat{q}_k = q_k + \left(\frac{\beta}{1 - 2dT} - T\right)p_k, \quad \hat{p}_k = p_k, \tag{150}
$$

the discrete-time algorithm (13) can be reformulated as

$$\hat{q}_{k+1} = y_k - T\tau\nabla f(y_k), \tag{151}$$
$$\hat{p}_{k+1} = (1 - 2dT)\hat{p}_k - T\nabla f(y_k), \tag{152}$$

where $\tau := \beta/(1 - 2dT)$, and

$$y_k = q_k + \beta p_k \tag{153}$$
$$= \hat{q}_k + T(1 - 2d\tau)\hat{p}_k. \tag{154}$$

The change of variables is motivated by the fact that the convexity and smoothness of the objective function implies $f(\hat{q}_{k+1}) \leq f(y_k)$,

$$f(\hat{q}_{k+1}) \leq f(y_k) - |\nabla f(y_k)|^2 \left(T\tau - \frac{T^2\tau^2 L}{2} \right), \tag{155}$$

where L is the Lipschitz constant of the gradient of f. In the following it will be shown that the energy H decreases along $\hat{q}_k$ and $\hat{p}_k$ provided that certain conditions on the parameters T, d, and β are met. We evaluate $H(\hat{q}_{k+1}, \hat{p}_{k+1})$,

$$H(\hat{q}_{k+1}, \hat{p}_{k+1}) = \frac{1}{2}(1 - 2dT)^2|\hat{p}_k|^2 - T(1 - 2dT)\hat{p}_k^\mathsf{T}\nabla f(y_k) + \frac{T^2}{2}|\nabla f(y_k)|^2 + f(\hat{q}_{k+1}) \tag{156}$$

$$\leq \frac{1}{2}|\hat{p}_k|^2 - (2dT - 2d^2T^2)|\hat{p}_k|^2 - T(1 - 2dT)\hat{p}_k^\mathsf{T}\nabla f(y_k)$$
$$- |\nabla f(y_k)|^2 \left(T\tau - \frac{T^2}{2}(\tau^2 L + 1) \right) + f(y_k). \tag{157}$$

Due to the fact that f is convex it holds that

$$f(y_k) - f(\hat{q}_k) \leq T(1 - 2d\tau)\nabla f(y_k)^\mathsf{T}\hat{p}_k, \tag{158}$$

and as a result

$$H(\hat{q}_{k+1}, \hat{p}_{k+1}) - H(\hat{q}_k, \hat{p}_k) \leq -(2dT - 2d^2T^2)|\hat{p}_k|^2 + 2dT(T - \tau)\hat{p}_k^\mathsf{T}\nabla f(y_k)$$
$$- |\nabla f(y_k)|^2(T\tau - \frac{T^2}{2}(\tau^2 L + 1)). \tag{159}$$

The right-hand side of the above expression is guaranteed to decrease provided that the following matrix

$$\begin{pmatrix} 2dT - 2d^2T^2 & -dT(T - \tau) \\ -dT(T - \tau) & T\tau - T^2(\tau^2 L + 1)/2 \end{pmatrix} \tag{160}$$

is positive definite. Thus, the minimum of f is an asymptotically stable equilibrium of the dynamics (13) provided that (160) is positive semidefinite. In particular, choosing $T = \tau$, i.e., $\beta = T(1 - 2dT)$, $0 < T \leq 1/\sqrt{L}$, $0 < dT < 1$, ensures asymptotic stability of the minimum. Moreover, for any fixed $\beta > 0$ the above matrix can be made negative definite by choosing T and d sufficiently small.

Appendix F. Smoothness assumptions on f

This section discusses the implications of Assumption 1 compared to Assumption 2. We therefore consider any function f that satisfies Assumption 1 and construct a sequence of functions f_j, where $j > 0$ is an integer, in the following way: Let $f_j : \mathbb{R}^n \to \mathbb{R}$ be such that $f_j(0) = 0$, $\nabla f_j(0) = 0$, and

$$\left. \frac{\partial^2 f_j}{\partial x^2} \right|_x := \int_{\mathbb{R}^n} \left. \frac{\partial^2 f}{\partial x^2} \right|_{\bar{x}} s_j(x - \bar{x}) \mathrm{d}\bar{x}, \tag{161}$$

where $s_j : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ is an infinitely differentiable function that has support on $B^c_{1/(2j)}$ and satisfies $\int_{\mathbb{R}^n} s_j(\bar{x}) \mathrm{d}\bar{x} = 1$. As a consequence, it holds that

$$|\nabla f(x) - \nabla f_j(x)| \leq 2\bar{C}_f/j, \quad |f(x) - f_j(x)| \leq 2\bar{C}_f|x|/j,$$

for all $x \in \mathbb{R}^n$ and $j > 0$ (assuming $\bar{C}_f > C_f$). In other words, ∇f_j converges uniformly to ∇f; the same holds for $f_j \to f$ on any compact subset of $\mathbb{R}^n$. As a result of (161), any essential upper or lower bound on the curvature of f are translated to the functions f_j. This implies, for example, that for large enough j the origin is an isolated non-degenerate critical point of f_j. In addition, each f_j, $j > 0$, is infinitely differentiable and satisfies Assumption 2. We therefore consider the dynamical system (37), where f is replaced with f_j:

$$z_j^+(t) = g_s(t, z_j(t), \nabla f_j(q_j(t))), \quad \forall t \in I, t \geq t_0, \qquad z_j(t_0) = z_0, \tag{162}$$

where $z_j(t) := (q_j(t), p_j(t))$. Compared to (37), the dependence on ∇f_j is made explicit. Let $z(t)$ be the trajectory satisfying (162), where f_j is replaced with f.

In the continuous-time case, (162) implies

$$|z_j(t) - z(t)| \leq \int_{t_0}^t |g_s(\hat{t}, z_j(\hat{t}), \nabla f_j(q_j(\hat{t}))) - g_s(\hat{t}, z(\hat{t}), \nabla f(q(\hat{t})))| \mathrm{d}\hat{t}$$

$$\leq C_{g1} \int_{t_0}^t |z_j(\hat{t}) - z(\hat{t})| \mathrm{d}\hat{t} + 2C_{g2}(t - t_0)\bar{C}_f/j, \tag{163}$$

where $C_{g1} > 0$ and $C_{g2} > 0$ are Lipschitz constants related to g_s. By virtue of the Grönwall inequality, (163) readily implies $z_j(t) \to z(t)$ pointwise for any $t \in I$. A similar argument applies to the discrete-time case.

We show next that, in case the equilibrium at the origin is uniformly stable (uniformity with respect to j and t_0), the convergence $z_j(t) \to z(t)$ is in fact uniform in $t \in I$. As a consequence, the results from Proposition 12 and similarly Proposition 3 generalize to the case where f satisfies merely Assumption 1 instead of Assumption 2, as shown below.

Proposition 18 *Let the origin be a stable equilibrium for* (162)*, uniformly in j and t_0 (for $j > 0$ sufficiently large). If $z(t)$ converges to the origin for $t \to \infty$, $z_j(t)$ converges to $z(t)$, uniformly in $t \in I$.*

Proof We pick any $\epsilon > 0$ and show that there exists an integer N, independent of t, such that $|z_j(t) - z(t)| < \epsilon$ for all $j > N$ and all $t \geq t_0$. Due to the uniform stability of the origin, there exists a $\delta > 0$ (independent of $t_0 \in I$ and j) such that $|z_j(t_0)| < \delta$ implies $|z_j(t)| < \epsilon/2$ for all

$t \geq t_0$ and $j > 0$ sufficiently large. The trajectory $z(t)$ converges to the origin. Hence, there exists a finite time $T \in I$ such that $|z(T)| < \delta/2$. Applying the Grönwall inequality to (163) yields

$$|z_j(t) - z(t)| \leq 2C_{\mathrm{g2}}(T - t_0)\bar{C}_{\mathrm{f}}e^{C_{\mathrm{g1}}(T-t_0)}/j, \quad \forall t \in I, t_0 \leq t \leq T; \tag{164}$$

a similar conclusion holds in the discrete-time case. Thus, choosing j large enough guarantees that $|z_j(t) - z(t)| < \delta/2 \leq \epsilon/2$ for all $t \in I, t_0 \leq t \leq T$. Therefore $|z_j(T)| < \delta$, which readily implies $|z_j(t)| < \epsilon/2$ for all $t \geq T$. Combined with $|z(t)| < \epsilon/2$ for all $t \geq T$, this yields $|z_j(t) - z(t)| < \epsilon$ for all $t \geq t_0$. ∎

Verifying that the origin is stable (uniformly in j) is typically straightforward. In order to illustrate the ideas we consider the discussion of Section 2.6. The total energy H_j (where f is replaced with f_j) is well-defined for any $j > 0$. For large enough j, the curvature of f_j is a localized average of the curvature of f. The origin is a non-degenerate isolated critical point of f, and therefore the same applies to f_j for large enough j. This concludes that in a neighborhood of the origin the total energy H_j is upper and lower bounded:

$$\frac{1}{2}|p|^2 + \frac{\mu}{4}|q|^2 \leq H_j(q, p) \leq \frac{1}{2}|p|^2 + L|q|^2,$$

for all $j > 0$ sufficiently large. These upper and lower bounds are independent of j, which, combined with the fact that H_j is non-increasing along trajectories, readily implies stability of the origin (uniformly in j and t_0) (see, for example, Sastry, 1999, p. 189). The same argument applies in the discrete-time case, where, by assumption, f and f_j are locally strongly convex. Thus, uniform stability follows from Appendix D.

The fact that the convergence $z_j \to z$ is uniform in t implies that the convergence estimates form Proposition 12 and Proposition 3 apply for $j \to \infty$. More precisely, we have

Proposition 19 *Let the origin be a stable equilibrium for* (38) *uniformly in j and t_0 (for $j > 0$ sufficiently large). Let $A \subset \mathbb{R}^{2n}$ be compact and such that $z_0 \in A$ implies $z(t) \to 0$ for $t \to \infty$. Then, provided that for sufficiently large $j > 0$ the trajectories $z_j(t)$ satisfy the estimates*

$$|z_j(t)| \leq \hat{C}_j|z_0|\frac{\rho(t)}{\rho(t_0)}\exp(-\alpha(t - t_0)), \quad \forall t \in I, t \geq t_0, \quad \forall z_0 \in A, \tag{165}$$

where $\hat{C}_j > 0$ and $\alpha > 0$ are constant, and $\rho : \mathbb{R}_{\geq 0} \to \mathbb{R}_{>0}$ is continuous and monotonically decreasing, there exists a constant $\hat{C}$ such that $z(t)$ satisfies the estimate

$$|z(t)| \leq \hat{C}|z_0|\frac{\rho(t)}{\rho(t_0)}\exp(-\bar{\alpha}(t - t_0)), \quad \forall t \in I, t \geq t_0, \quad \forall z_0 \in A, \tag{166}$$

for any constant $\bar{\alpha} < \alpha$.

Proof For any $\epsilon > 0$, (165) implies that for any $j > 0$ sufficiently large,

$$\lim_{t \to \infty} |z_j(t)|\frac{\rho(t_0)}{\rho(t)}\exp((\alpha - \epsilon)(t - t_0)) = 0. \tag{167}$$

Due to the fact that z_j converges uniformly to z, it follows that

$$\lim_{t\to\infty} |z(t)|\frac{\rho(t_0)}{\rho(t)} \exp((\alpha-\epsilon)(t-t_0)) = \lim_{j\to\infty} \lim_{t\to\infty} |z_j(t)|\frac{\rho(t_0)}{\rho(t)} \exp((\alpha-\epsilon)(t-t_0)) = 0, \quad (168)$$

which yields the desired result. ■

References

Ravi P. Agarwal. *Difference Equations and Inequalities*. Marcel Dekker, Inc., second edition, 2000.

Zeyuan Allen-Zhu and Lorenzo Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. `arXiv:1407.1537 [cs.DS]`, 2014.

Vladimir I. Arnol'd. *Ordinary Differential Equations*. Springer, third edition, 1992.

Hedy Attouch, Zaki Chbani, Juan Peypouquet, and Patrick Redont. Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity. *Mathematical Programming, Series B*, 168(1-2):123–175, 2018.

Richard Bellman. *Stability Theory of Differential Equations*. Dover, 1969.

Michael Betancourt, Michael I. Jordan, and Ashia C. Wilson. On symplectic optimization. *arXiv:1802.03653v2*, pages 1–20, 2018.

Sébastien Bubeck, Yin Tat Lee, and Mohit Singh. A geometric alternative to Nesterov's accelerated gradient descent. `arXiv:1506.08187 [math.OC]`, 2015.

Sébastien Bubeck, Qijia Jiang, Yin Tat Lee, Yuanzhi Li, and Aaron Sidford. Near-optimal method for highly smooth convex optimization. *Proceedings of Machine Learning Research*, 99:492–507, 2019.

Frank E. Curtis, Daniel P. Robinson, and Mohammadreza Samadi. A trust region algorithm with a worst-case iteration complexity of $\mathcal{O}(\epsilon^{-3/2})$ for nonconvex optimization. *Mathematical Programming, Series A*, 162(1-2):1–32, 2017.

Jelena Diakonikolas and Lorenzo Orecchia. The approximate duality gap technique: A unified theory of first-order methods. *SIAM Journal on Optimization*, 29(1):660–689, 2018.

Simon S. Du, Chi Jin, Jason D. Lee, Michael I. Jordan, Barnabás Póczos, and Aarti Singh. Gradient descent can take exponential time to escape saddle points. *Advances in Neural Information Processing Systems 30*, pages 1067–1077, 2017.

Hans-Bernd Dürr and Christian Ebenbauer. On a class of smooth optimization algorithms with applications in control. *Proceedings of the IFAC Nonlinear Model Predictive Control Conference*, pages 291–298, 2012.

Saber Elaydi. *An Introduction to Difference Equations*. Springer, third edition, 2005.

Guilherme França, Jeremias Sulam, Daniel Robinso, and René Vidal. Conformal symplectic and relativistic optimization. *arXiv:1903.04100 [math.OC]*, pages 1–27, 2019.

Sébastien Gadat, Fabien Panloup, and Sofiane Saadane. Stochastic heavy ball. *Electronic Journal of Statistics*, 12(1):461–529, 2018.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points - online stochastic gradient for tensor decomposition. *Proceedings of Machine Learning Research*, 40:797–842, 2015.

Ernst Hairer, Gerhard Wanner, and Christian Lubich. *Geometric Numerical Integration*. Springer, 2002.

Jack K. Hale. *Asymptotic Behavior of Dissipative Systems*. American Mathematical Society, 1988.

Alain Haraux. *Systèmes Dynamiques Dissipatifs et Applications*. Masson, 1991.

Chi Jin, Praneeth Netrapalli, and Michael I. Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. *arXiv:1711.10456*, pages 1–43, 2017.

Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M. Kakade, and Michael I. Jordan. On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points. *arXiv:1902.04811*, pages 1–31, 2019.

Walid Krichene, Alexandre M. Bayen, and Peter L. Bartlett. Accelerated mirror descent in continuous and discrete time. *Advances in Neural Information Processing Systems 28*, pages 2845–2853, 2015.

Jason D. Lee, Max Simchowitz, Michael I. Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. *Proceedings of Machine Learning Research*, 49(1):1246–1257, 2016.

Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

Chris J. Maddison, Daniel Paulin, Yee Whye Teh, and Brendan O'Donoghue. Hamiltonian descent methods. *arXiv:1809.05042 [math.OC]*, pages 1–72, 2018.

Simon Michalowsky, Carsten Scherer, and Christian Ebenbauer. Robust and structure exploiting optimization algorithms: An integral quadratic constraint approach. *arXiv:1905.00279*, pages 1–30, 2019.

J. Milnor. *Morse Theory*. Princeton University Press, 1963.

Michael Muehlebach and Michael I. Jordan. A dynamical systems perspective on Nesterov acceleration. *Proceedings of Machine Learning Research*, 97:4656–4662, 2019.

Yuri E. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

Yuri E. Nesterov and Boris T. Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming, Series A*, 108:177–205, 2006.

Yurii Nesterov. *Introductory Lectures on Convex Optimization - A Basic Course*. Springer Science+Business Media, LLC, 2004.

Boris T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

Boris T. Polyak. *Introduction to Optimization*. Optimization Software, Inc., 1987.

Walter Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, third edition, 1976.

J. M. Sanz-Serna. Symplectic integrators for Hamiltonian problems: an overview. *Acta Numerica*, pages 243–286, 1992.

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

Damien Scieur, Vincent Roulet, Francis Bach, and Alexandre d'Aspremont. Integration methods and optimization algorithms. *Advances in Neural Information Processing Systems 30*, pages 1109–1118, 2017.

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

Roscoe B White. *Asymptotic Analysis of Differential Equations*. Imperial College Press, 2010.

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