A Dynamical Systems Perspective on Nesterov Acceleration

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

Abstract

We present a dynamical system framework for understanding Nesterov's accelerated gradient method. In contrast to earlier work, our derivation does not rely on a vanishing step size argument. We show that Nesterov acceleration arises from discretizing an ordinary differential equation with a semi-implicit Euler integration scheme. We analyze both the underlying differential equation as well as the discretization to obtain insights into the phenomenon of acceleration. The analysis suggests that a curvature-dependent damping term lies at the heart of the phenomenon. We further establish connections between the discretized and the continuous-time dynamics.

1. Introduction

Many tasks in machine learning and statistics can be formulated as optimization problems. In its basic form, an optimization problem consists of finding a value x^* that minimizes the function f, that is $f(x^*) \leq f(x)$ for all x, provided such a value exists. In the following, we consider the subclass of problems where the function f is real-valued, convex, and has a Lipschitz-continuous gradient. These assumptions are often met, at least locally, in practical problem instances.

One of the most popular algorithms for computing x^* up to a desired accuracy is the gradient method, due to its simplicity and due to the fact that it scales well to large problem sizes. It is also possible to obtain faster convergence rates within the family of gradient-based methods by considering accelerated gradient methods, which make use of two successive gradients (Nesterov, 1983).

Given the important role that acceleration has played not only in generating useful new algorithms but also in understanding natural limits on convergence rate, there have

been many attempts to understand and characterize the phenomenon. Bubeck et al. (2015) suggest a modification of the accelerated gradient method that achieves the same convergence rate, but has a geometric interpretation. Allen-Zhu & Orecchia (2014) present an algorithm that couples gradient and mirror descent, while achieving an accelerated convergence rate and Lessard et al. (2016) propose a general analysis framework, inspired by control theory, for comparing different first-order optimization algorithms including the accelerated gradient method. Other work includes Diakonikolas & Orecchia (2018), who unify the analysis of first-order methods by imposing certain decay conditions, or Scieur et al. (2017), where the accelerated gradient method is interpreted as a multi-step discretization of the gradient flow.

An important step was taken by Su et al. (2016) and Krichene et al. (2015), who showed that for a vanishing step size the trajectories of the accelerated gradient scheme approach the solutions of a certain second-order ordinary differential equation (ODE). The resulting ODE was analyzed in further detail by Attouch et al. (2018). It was placed within a variational framework by Wibisono et al. (2016), which led to the discovery of higher-order schemes that achieve even faster convergence rates.

In the current paper, we provide a different interpretation of the accelerated gradient method, one that is not based on beginning with the difference equation and deriving an underlying ODE via a vanishing step size argument. We go in the other direction, starting with a certain second-order ODE that has a clear interpretation as a mass-spring-damper system, and showing that Nesterov acceleration can be derived by discretizing the ODE with a semi-implicit Euler method. In contrast to the continuous-time limit of the heavy-ball method (Polyak, 1964), our model includes a damping term that locally averages the curvature. This additional damping seems instrumental to the acceleration phenomenon. In addition to providing intuition, we believe that the model can be used to translate properties of the continuous-time system, which are, as we show, often easier to derive, to the resulting discrete-time algorithm. In particular, we show that under some regularity conditions fundamental geometric properties (phase-space area contraction and time-reversibility) are preserved by the discretization.

---

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

Notation and outline: We focus our presentation on n-dimensional optimization problems over the real numbers, where the integer n is finite. It will be clear that most derivations generalize in a straightforward way to non-Euclidean inner product spaces. We also focus, again for simplicity of presentation, on strongly convex functions; that is, functions for which there exists a constant $\kappa \geq 1$ (the condition number) such that for any $\bar{x} \in \mathbb{R}^n$

$$f(x) \geq f(\bar{x}) + \nabla f(\bar{x})(x - \bar{x}) + \frac{L}{2\kappa}|x - \bar{x}|^2, \forall x \in \mathbb{R}^n,$$

where L is the Lipschitz constant of the gradient and $|\cdot|$ denotes the Euclidean norm. We also treat the non-strongly convex setting, however. The gradient and the Hessian (if it exists) evaluated at $x \in \mathbb{R}^n$ are denoted by $\nabla f(x) \in \mathbb{R}^n$ and $\Delta f(x) \in \mathbb{R}^{n \times n}$ respectively. We assume throughout the article that the function f is smooth; i.e., that f has a Lipschitz continuous gradient. We further assume that f attains its minimum value at $x^* = 0$ and that $f(0) = 0$, which, in the case of a strongly convex function is without loss of generality.

The article is structured in the following way: Sec. 2 introduces the dynamical system model that serves as the foundation for our analysis. The dynamical system is shown to yield Nesterov's accelerated gradient method when discretized with the semi-implicit Euler scheme with a certain non-vanishing step size. In Sec. 3, the continuous-time dynamics are analyzed and an intuitive interpretation as a mass-spring-damper system is provided. The dynamics are shown to have a curvature-dependent damping term, responsible for the acceleration. Sec. 4 analyzes and motivates the discretization. It is shown that for a step size $T_s \in (0, 1)$, the discrete-time dynamics preserve geometric properties of the continuous-time dynamics. Sec. 5 presents a simulation example that illustrates the properties of the continuous-time dynamics and the discretization. The article concludes with a discussion in Sec. 6.

2. Dynamical systems model

In this section we show that the accelerated gradient method presented in Nesterov (2004) (p. 81) results from a semi-implicit Euler discretization of the following ODE:

$$\ddot{x}(t) + 2d\dot{x}(t) + \frac{1}{L\gamma^2}\nabla f(x(t) + \beta\dot{x}(t)) = 0, \quad (1)$$

where γ is a constant and

$$d := \frac{1}{\sqrt{\kappa} + 1} \frac{1}{\gamma}, \quad \beta := \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \gamma. \quad (2)$$

Changing the constant γ amounts to a rescaling of the solutions of (1) in time. We set the constant γ to one for ease of notation.

By the semi-implicit Euler scheme, we mean the following integration algorithm:

$$q_{k+1} = q_k + T_s \nabla T(p_{k+1}), \quad (3)$$

$$p_{k+1} = p_k + T_s(-\frac{1}{L}\nabla f(q_k) + f_{\text{NP}}(q_k, p_k)), \quad (4)$$

where the corresponding continuous-time dynamics evolve according to

$$\dot{q}(t) = \nabla T(p(t)), \quad (5)$$

$$\dot{p}(t) = -\frac{1}{L}\nabla f(q(t)) + f_{\text{NP}}(q(t), p(t)), \quad (6)$$

and where the real-valued, continuously differentiable function $T(p)$ represents the kinetic energy, the continuous, real-valued function $f_{\text{NP}}(q, p)$ the non-potential forces, and the real number $T_s > 0$ the step size. The total energy of the dynamic system described by (5) and (6) is given by

$$H(q, p) := T(p) + \frac{1}{L}f(q). \quad (7)$$

The dynamic system (1) can be brought to the form (5) and (6) by introducing

$$T(p) := \frac{1}{2}|p|^2, \quad (8)$$

$$f_{\text{NP}}(q, p) := -2dp - \frac{1}{L}(\nabla f(q + \beta p) - \nabla f(q)), \quad (9)$$

and identifying $q(t)$ with $x(t)$ and $p(t)$ with $\dot{x}(t)$. The discretization according to (3) and (4) with $T_s = 1$ then yields

$$p_{k+1} = p_k - 2dp_k - \frac{1}{L}\nabla f(q_k + \beta p_k)$$

$$= \beta p_k - \frac{1}{L}\nabla f(q_k + \beta p_k), \quad (10)$$

$$q_{k+1} = q_k + p_{k+1}$$

$$= q_k + \beta p_k - \frac{1}{L}\nabla f(q_k + \beta p_k), \quad (11)$$

where the identity $2d + \beta = 1$ has been used in (10). By defining $y_k := q_k + \beta p_k$ and $x_k := q_k$, we obtain

$$x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k) \quad (12)$$

$$y_{k+1} = x_{k+1} + \beta(x_{k+1} - x_k), \quad (13)$$

which corresponds to the accelerated gradient scheme with constant step size (Nesterov, 2004) (p. 81).

A very similar derivation can be performed to obtain a formulation of the accelerated gradient method for smooth but non-strongly-convex functions. In particular, replacing d and β in (1) with

$$\bar{d}(t) := \frac{3}{2(t + 2)}, \quad \bar{\beta}(t) := \frac{t - 1}{t + 2}, \quad (14)$$

and applying the above discretization with time-step $T_{\mathrm{s}} = 1$ yields the optimization algorithm that is used as a starting point in Su et al. (2016). The constants d, β, as well as the variables $\bar{d}(t)$, $\bar{\beta}(t)$ satisfy $2d + \beta = 1$, respectively $2\bar{d}(t) + \bar{\beta}(t) = 1$ for all $t \geq 0$.

3. Interpretation of the model

In this section we discuss the interpretation of the ODE (1) as a mass-spring-damper system with a curvature-dependent damping term. In addition, we show that the trajectories of (1) converge exponentially with rate at least $1/(2\sqrt{\kappa}) - 1/(4\kappa)$, matching the well-known rate of the accelerated gradient method.

We start by reformulating the non-potential forces in (9) in the following way:[1]

$$f_{\mathrm{NP}}(q, p) = -2dp - \frac{1}{L}\int_0^\beta \Delta f(q + \tau p)\mathrm{d}\tau\, p. \qquad (15)$$

Thus, the dynamics (1) can be rewritten as

$$\ddot{x}(t) + (2dI + D_{x\dot{x}})\,\dot{x}(t) + \frac{1}{L}\nabla f(x(t)) = 0, \qquad (16)$$

where $I \in \mathbb{R}^{n\times n}$ denotes identity matrix and where the damping $D_{x\dot{x}}$ is defined as

$$D_{x\dot{x}} := \frac{1}{L}\int_0^\beta \Delta f(x(t) + \tau\dot{x}(t))\,\mathrm{d}\tau. \qquad (17)$$

To lighten notation the dependence of $D_{x\dot{x}}$ on $x(t)$ and $\dot{x}(t)$ is indicated with subscripts. Direct calculations show that

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}H(x(t), \dot{x}(t)) &= f_{\mathrm{NP}}(x(t), \dot{x}(t))^{\mathsf{T}}\dot{x}(t) \\
&= -\dot{x}(t)^{\mathsf{T}}(2dI + D_{x\dot{x}})\dot{x}(t), \qquad (18)
\end{aligned}$$

which bounds the energy dissipation by

$$-|\dot{x}(t)|^2 \leq \frac{\mathrm{d}}{\mathrm{d}t}H(x(t), \dot{x}(t)) \leq -(2d + \beta/\kappa)|\dot{x}(t)|^2. \qquad (19)$$

Note that the identity $2d + \beta = 1$, as well as the assumption of f being smooth and strongly convex have been used for obtaining these inequalities. In the non-strongly convex case the bounds reduce to

$$-|\dot{x}(t)|^2 \leq \frac{\mathrm{d}}{\mathrm{d}t}H(x(t), \dot{x}(t)) \leq -2\bar{d}(t)|\dot{x}(t)|^2, \qquad (20)$$

for $t \geq 1$, respectively

$$-2\bar{d}(t)|\dot{x}(t)|^2 \leq \frac{\mathrm{d}}{\mathrm{d}t}H(x(t), \dot{x}(t)) \leq -|\dot{x}(t)|^2, \qquad (21)$$

---

[1]By assumption, the function f has a Lipschitz-continuous gradient, hence the (Lebesgue) integral appearing in (15) and (17) is well-defined even if the Hessian might not exist everywhere.
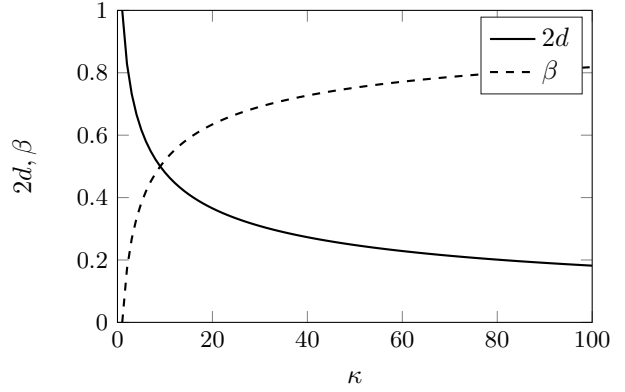

Figure 1. The figure shows the constants $2d$ and β as a function of the maximum curvature κ. The constant $2d$ decreases with $1/\sqrt{\kappa}$, whereas β increases with $1/\sqrt{\kappa}$ (for large values of κ).

for $t \in (0, 1)$.

Summarizing, it may therefore be concluded (the conclusions in the non-strongly-convex case are analogous):

- In the absence of damping, i.e., $d = D_{x\dot{x}} \equiv 0$, the total energy $H(x(t), \dot{x}(t))$ is conserved. The trajectories are thus confined to the level sets of $H(x(t), \dot{x}(t))$.

- In the presence of damping, energy is dissipated. As a result, La Salle's theorem (see, for example, Sastry, 1999, Ch. 5.4) implies that the trajectories of (1) converge asymptotically to the origin. Similar conclusions can be drawn in the non-convex case.

- The damping coefficient $D_{x\dot{x}}$ can be interpreted as a local average of the curvature near $x(t)$. The amount of averaging depends on the maximum curvature of the objective function and on the current velocity. More averaging is performed at higher velocities (larger $\dot{x}(t)$) and for a larger maximum curvature (large value of κ). The latter is due to the fact that the coefficient β increases monotonically with κ.

- The total damping is a linear combination of constant damping and local curvature-dependent damping. The two sources of damping are balanced by the coefficient $2d$ (constant damping) that decreases with larger κ and the coefficient β (local curvature-dependent damping) that increases with larger κ. The total amount of damping remains constant in the sense that $2d + \beta = 1$ for any maximum curvature κ. The constants $2d$ and β are shown in Fig. 1 for various values of κ.

- Energy is dissipated even for $\kappa \to \infty$ due to the local curvature-dependent damping.

The following two propositions provide an estimate of the

convergence rate of the differential equation (1), the proof of which can be found in the appendix, App. A.1 and App. A.2.

Proposition 3.1 *(Strongly convex case) The trajectories of (1) decay exponentially to the origin with rate at least* $1/(2\sqrt{\kappa}) - 1/(4\kappa)$.

Proposition 3.2 *(Non-strongly-convex case) When replacing d with $\bar{d}(t)$ and β with $\bar{\beta}(t)$ in (1), $f(x(t))$ decays at least with $\mathcal{O}(1/t^2)$.*

We conclude the section with the following remarks about Prop. 3.1 and Prop. 3.2:

- Prop. 3.1 and Prop. 3.2 are based on Lyapunov functions of the type

$$\frac{1}{2}|aq + p|^2 + \frac{1}{L}f(q), \qquad (22)$$

  where a is related to the rate of convergence and is either constant, in the strongly convex case, or time-varying, in the non-strongly-convex case.

- In both cases, the choice of a is motivated by convenience, such that the resulting cross-term $q^\mathsf{T}p$ vanishes for the given parameters d and β (or $\bar{d}$ and $\bar{\beta}$ in the non-strongly-convex case). However, the proof can be extended to derive more general sufficiency conditions on the parameters d and β (or $\bar{d}$ and $\bar{\beta}$) that provide an accelerated rate.

- The asymptotic convergence rate of $\mathcal{O}(1/\sqrt{\kappa})$ in Prop. 3.1 can also be established using a Lyapunov function that is inspired by the Popov criterion; see for example (Sastry, 1999) (Ch. 6.2.2). The Popov criterion gives a sufficient condition for the closed-loop stability of a linear time-invariant system subject to a static nonlinear feedback path, and is applicable to the ODE (1).

4. The discretization

The following section highlights that fundamental geometric properties, such as the phase-space area contraction rate and the time-reversibility of the dynamics (1) are preserved through the discretization. In addition to providing insights, both properties will be used to bound the convergence rate in a certain worst-case sense. The section concludes with a proposition stating that the discretized dynamics indeed converge at the given accelerated rate. Unfortunately, the proof relies on a Lyapunov function, similar to (22), and does not seem to provide additional insights into the discretization process. For simplifying the presentation, the result concerning the phase-space area contraction is stated for a one-dimensional objective function, while

the more general case can be found in App. A.4. The map $(q_k, p_k) \to (q_{k+1}, p_{k+1})$ given by the discretized dynamics (3) and (4) is denoted by ψ.

The discretization (3) and (4) can be divided into two parts, a non-conservative step that involves an update of the momentum based on the non-potential forces f_{NP},

$$\bar{q}_{k+1} = q_k, \quad \bar{p}_{k+1} = p_k + T_\mathrm{s}(f_{\mathrm{NP}}(q_k, p_k)), \qquad (23)$$

followed by a symplectic Euler step based on the conservative part of the system,

$$q_{k+1} = \bar{q}_{k+1} + T_\mathrm{s}\nabla T(p_{k+1}), \qquad (24)$$

$$p_{k+1} = \bar{p}_{k+1} + T_\mathrm{s}(-\frac{1}{L}\nabla f(\bar{q}_{k+1})). \qquad (25)$$

The symplectic Euler scheme is a well-known integration scheme for Hamiltonian systems (see, for example, Hairer et al., 2002, Ch. VI.3). It is one of the simplest symplectic integration schemes and is known to be energy consistent (nearly energy conserving) over exponentially long time intervals (Hairer et al., 2002).

The combination of (23) with (24) and (25) leads to a phase-space area contraction, which we quantify with the following proposition. (Here formulated for $n = 1$.)

Proposition 4.1 *Let $\partial\Gamma_k$ be a simple closed contour in $\mathbb{R}^2$ and let the signed area corresponding to $\partial\Gamma_k$ be defined as*

$$A_k := \int_{\Gamma_k} dq_k \wedge dp_k, \qquad (26)$$

where Γ_k describes all points enclosed by $\partial\Gamma_k$.[1] The integration scheme (3) and (4) maps $\partial\Gamma_k$ to $\partial\Gamma_{k+1}$ such that

$$A_{k+1} - A_k = -T_s \int_{\partial\Gamma_k} f_{NP}(q_k, p_k)dq_k \qquad (27)$$

$$= -T_s \int_{\Gamma_k} 2d + \frac{\beta}{L}\Delta f(q_k + \beta p_k)\, dq_k \wedge dp_k. \qquad (28)$$

The setting of the proposition is illustrated by Fig. 2 and its proof can be found in App. A.3.

Provided that the area A_k has a positive sign (for a negative sign the inequalities are reversed), the integral on the right-hand side of (28) is bounded by

$$\left(2d+\frac{\beta}{\kappa}\right)A_k \leq \int_{\Gamma_k} 2d+\frac{\beta}{L}\Delta f(q_k+\beta p_k)\mathrm{d}q_k\mathrm{d}p_k \leq A_k, \qquad (29)$$

---

[1] The sign is inferred from the contour, that is, a counter-clockwise direction yields a positive sign: $\mathrm{d}q_k \wedge \mathrm{d}p_k = \mathrm{d}q_k\mathrm{d}p_k$, a clockwise direction yields a negative sign, i.e. $\mathrm{d}q_k \wedge \mathrm{d}p_k = -\mathrm{d}q_k\mathrm{d}p_k$, where $\mathrm{d}q_k\mathrm{d}p_k$ refers to the standard area measure.
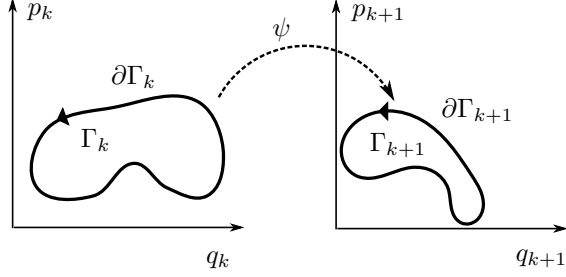
Figure 2. The figure illustrates the setting of Prop. 4.1. The direction of the contours is indicated by the two arrows, and as a result, both A_k, enclosed by $\partial\Gamma_k$, and A_{k+1}, enclosed by $\partial\Gamma_{k+1}$, have a positive sign. The proposition states that the areas A_k, $k = 0, 1, \ldots$, contract with a rate matching the continuous-time dynamics for $T_s \in (0, 1)$.

which follows from the fact that f is strongly convex (yielding the lower bound) and smooth (yielding the upper bound), and the identity $2d + \beta = 1$. This concludes that the phase-space area contracts for $T_s \in (0, 2)$. In case $T_s \in (0, 1]$ the area does not change sign; i.e., the contour is guaranteed to preserve its original orientation. For $T_s = 1$, which yields the accelerated gradient method, the contour might contract to a single point.

With very similar arguments it can be shown that the phase-space area contraction rate of the continuous-time dynamics (1) is given by

$$\dot{A}(t) = -\int_{\partial\Gamma(t)} f_{\mathrm{NP}}(q, p)\mathrm{d}q \tag{30}$$

$$= -\int_{\Gamma(t)} 2d + \frac{\beta}{L}\Delta f(q + \beta p)\,\mathrm{d}q \wedge \mathrm{d}p, \tag{31}$$

where $\partial\Gamma(t)$ denotes a simple closed contour in $\mathbb{R}^2$ that evolves according to the dynamics (1), and $A(t)$ is the signed area enclosed by the contour. It can therefore be concluded that the phase-space area of the discretized dynamics essentially contracts at the same rate as the continuous-time counterpart. More precisely, we have:

Proposition 4.2 *The contraction of the signed phase-space area enclosed by an energy level-set with the discretized dynamics is larger than the contraction with the continuous-time dynamics over the length of one time step.*

The proof relies on the mean-value theorem and can be found in App. A.4.

We show next that the discrete integration describes a homeomorphism (a continuous bijection that has a continuous inverse) for $T_s \in (0, 1)$. This implies that the discrete-time dynamics are time reversible, which parallels the continuous-time counterpart.

Proposition 4.3 *The discrete integration given by* (3) *and* (4) *describes a homeomorphism provided that* $T_s \in (0, 1)$.

The proof can be found in App. A.5.

Combining the above proposition with the phase-space area contraction yields the following result:

Proposition 4.4 *For every* $T_s \in (0, 1)$ *there exists at least one trajectory that converges linearly with rate* $1 - T_s\mathcal{O}(1/\sqrt{\kappa})$.

Proof We consider the closed contour $\partial\Gamma_0$, given by a circle with radius $R > 0$ centered at the origin (with positive orientation). The map ψ is a homeomorphism, and therefore, as the contour evolves, the origin will always be contained in its interior. The area A_0 enclosed by the contour at time zero is $R^2\pi$. We claim that at time $k > 0$, there is at least one trajectory that is at a distance at least

$$R_k := R(1 - T_s(2d + \beta/\kappa))^{k/2} \tag{32}$$

from the origin. For the sake of contradiction we assume that the claim is false. This implies, however, that all trajectories starting from $\partial\Gamma_0$ remain outside a closed ball of radius R_k centered at the origin, and hence $A_k > R_k^2\pi$. This is contradicting the fact that A_k decays at least with $(1 - T_s(2d + \beta/\kappa))^k$, according to Prop. 4.1 and (29), which concludes the proof. ∎

We show in the appendix (App. A.6) that all trajectories indeed converge at an accelerated rate, as suggested by Prop. 4.4:

Proposition 4.5 *For every* $T_s \in (0, 1]$ *the trajectories of the discrete-time dynamics* (3) *and* (4) *converge linearly with rate at least* $1 - T_s\mathcal{O}(1/\sqrt{\kappa})$.

5. Simulation example

This section presents a numerical example that illustrates certain properties of the discrete integration.

We choose a function f with the following gradient

$$\nabla f(x) = \begin{cases} \kappa x, & x < 1, \\ \kappa - 1 + x, & 1 \le x < 2, \\ 1 - \kappa + \kappa x, & x \ge 2, \end{cases} \tag{33}$$

where the condition number κ is set to 5. The function and its gradient are shown in Fig. 5, and is inspired by the counterexample provided in Lessard et al. (2016). The integration algorithm given by (3) and (4) is applied to the initial conditions $(q_0, 0)$, where q_0 is varied from -2 to 5 in steps of 0.2. The step size T_s is successively increased from 0.1 to 1.2. The evolution of the corresponding trajectories
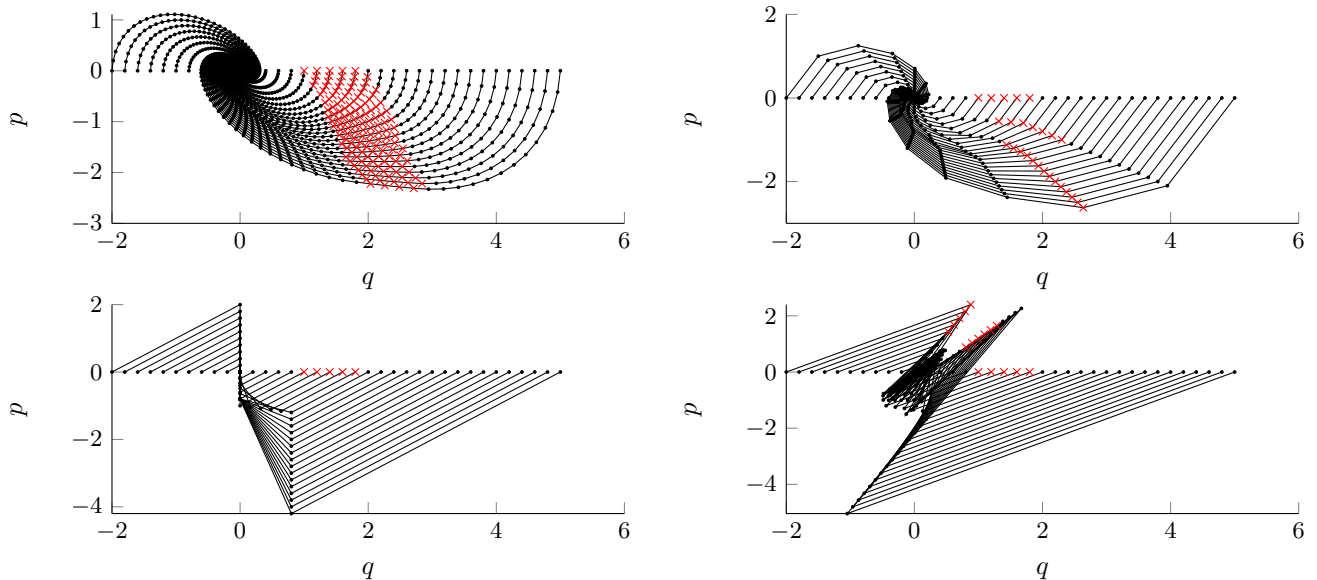
Figure 3. The plot shows the evolution of different trajectories in the phase space. The trajectories are obtained by applying the discrete integration algorithm (3) and (4) with the time steps $T_s = 0.1, 0.5, 1, 1.2$ (top left, top right, bottom left, bottom right). The points (q, p) with $q + \beta p \in [1, 2)$ are marked with red crosses.
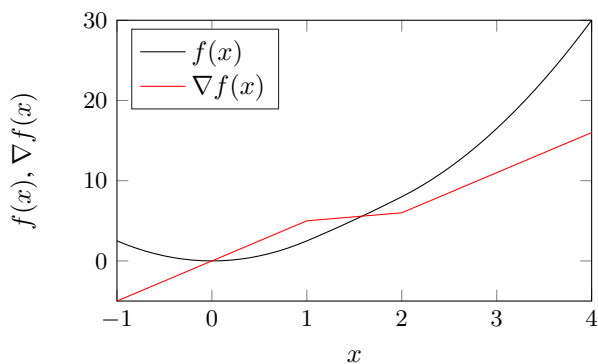

Figure 4. The plot shows the function f and its gradient ∇f used for the numerical example.

in the phase space is shown in Fig. 3. For a time step of $T_s = 0.1$ the trajectories resemble the solutions of (1). As the time step is increased the trajectories are more and more "folded" in clockwise direction. For the time step $T_s = 1$, which corresponds to the accelerated gradient method, multiple initial conditions are mapped to the same point (for example the point $q = 0$, $p = -0.8$) due to the course of the integration. The trajectories starting from an initial condition $q_0 < 1$, $p_0 = 0$ converge exactly in two steps. If the time step is increased above $T_s = 1$ the discrete integration map ψ is no longer orientation preserving, and the resulting motion is much less structured, as shown in Fig. 3 (bottom right). For a time step of $T_s = 1.3$ trajectories starting from $(q_0, 0)$ with $q_0 \geq 4.4$ are found to diverge.

6. Conclusions

We have shown that Nesterov's accelerated gradient method can be interpreted as the discretization of a certain ODE with the semi-implicit Euler scheme. The differential equation describes a mass-spring-damper system with a curvature-dependent damping term, which seems essential for the acceleration. The analysis suggests that geometric properties (time reversibility and phase-space area contraction) are conserved by the discretization.

The discretization can be divided into two steps, a contraction step based on the non-conservative part and a symplectic Euler step based on the conservative part of the dynamics. The fact that the symplectic Euler scheme is long-term energy consistent makes the discretization appear very natural. However, as the steps do not commute, a straightforward application of backwards error analysis, which would rigorously justify this intuition, seems difficult.

Even though we characterized and related the phase-space area contraction property of the discretized dynamics to their continuous-time counterpart, the area contraction property alone does not seem to be enough for deriving a convergence rate of $f(q)$.

Acknowledgements

References

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

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

Bubeck, S., Lee, Y. T., and Singh, M. A geometric alternative to Nesterov's accelerated gradient descent. `arXiv:1506.08187 [math.OC]`, 2015.

Diakonikolas, J. and Orecchia, L. The approximate duality gap technique: A unified theory of first-order methods. `arXiv:arXiv:1712.02485v2 [math.OC]`, 2018.

Hairer, E., Lubich, C., and Wanner, G. *Geometric Numerical Integration*. Springer, 2002.

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

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

Munkres, J. R. *Topology*. Prentice Hall, second edition, 2000.

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

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

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

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

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

Scieur, D., Roulet, V., Bach, F., and d'Aspermont, A. Integration methods and optimization algorithms. *Advances in Neural Information Processing Systems 30*, pp. 1109–1118, 2017.

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

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

A. Appendix

A.1. Proof of Prop. 3.1

We claim that the following function is a Lyapunov function for the differential equation (1) (with $q(t) = x(t)$, $p(t) = \dot{x}(t)$):

$$V(q, p) := \frac{1}{2}|aq + p|^2 + \frac{1}{L}f(q), \qquad (34)$$

where

$$a := d + \frac{\beta}{2\kappa} = \frac{1}{\sqrt{\kappa}} - \frac{1}{2\kappa}. \qquad (35)$$

The value of the constant a is motivated below. It can be verified that V is positive definite. We will omit the dependence on time to simplify notation. Evaluating the time derivative of V along the trajectories of (1) yields

$$\dot{V} = -\frac{a}{L}q^{\mathsf{T}}\nabla f(y) + a(a - 2d)q^{\mathsf{T}}p + (a - 2d)|p|^2$$
$$- \frac{1}{L}p^{\mathsf{T}}(\nabla f(y) - \nabla f(q)), \quad (36)$$

with $y := q + \beta p$. Due to the fact that f is smooth and strongly convex, the following two inequalities hold:

$$-y^{\mathsf{T}}\nabla f(y) \leq -f(y) - \frac{L}{2\kappa}|y|^2, \qquad (37)$$

$$f(y) \geq f(q) + \nabla f(q)^{\mathsf{T}}\beta p + \frac{L}{2\kappa}\beta^2|p|^2. \qquad (38)$$

These can be combined to

$$-y^{\mathsf{T}}\nabla f(y) \leq -f(q) - \nabla f(q)^{\mathsf{T}}\beta p - \frac{L}{2\kappa}(|y|^2 + \beta^2|p|^2),$$

which implies, by definition of y, that

$$-q^{\mathsf{T}}\nabla f(y) \leq -f(q) - \frac{L}{2\kappa}(|q|^2 + 2\beta q^{\mathsf{T}}p + 2\beta^2|p|^2)$$
$$+ \beta p^{\mathsf{T}}(\nabla f(y) - \nabla f(q)). \quad (39)$$

Substituting the above bound on $-q^{\mathsf{T}}\nabla f(y)$ in (36) and rearranging terms yields

$$\dot{V} = -aV + \frac{a}{2}\left(a^2 - \frac{1}{\kappa}\right)|q|^2 + \left(\frac{3a}{2} - 2d - \frac{a\beta^2}{\kappa}\right)|p|^2$$
$$+ a\left(2a - 2d - \frac{\beta}{\kappa}\right)q^{\mathsf{T}}p$$
$$- \frac{1}{L}(1 - \beta a)p^{\mathsf{T}}(\nabla f(y) - \nabla f(q)). \quad (40)$$

The constant a is deliberately chosen such that the cross-term $q^{\mathsf{T}}p$ in (40) vanishes. Due to the fact that $1 - \beta a \geq 0$ and $\beta \geq 0$, the last term of (40) can be bounded by

$$-\frac{1}{L}(1 - \beta a)p^{\mathsf{T}}(\nabla f(y) - \nabla f(q)) \leq -(1 - \beta a)\frac{\beta}{\kappa}|p|^2,$$

which implies,

$$\dot{V} \leq -aV + \frac{a}{2}\left(a^2 - \frac{1}{\kappa}\right)|q|^2 + \left(\frac{3a}{2} - 2d - \frac{\beta}{\kappa}\right)|p|^2. \qquad (41)$$

It can be verified that $a^2 - 1/\kappa \leq 0$ and $(3a/2 - 2d - \beta/\kappa) \leq 0$ for all $\kappa \geq 1$. Hence, $\dot{V} \leq -aV$, which, by applying Grönwall's inequality, establishes the proposition. ∎

A.2. Proof of Prop. 3.2

The proof is analogous to the proof of Prop. 3.1, and hinges on the following Lyapunov-like function

$$\bar{V}(t) := \frac{1}{2}|\bar{a}(t)q(t) + p(t)|^2 + \frac{1}{L}f(q(t)), \qquad (42)$$

where $\bar{a}(t) := 2/(t + 2)$, and $q(t) = x(t), p(t) = \dot{x}(t)$ are the solutions of (1), when d is replaced with $\bar{d}(t)$ and β with $\bar{\beta}(t)$. The dependence on time will be omitted to simplify notation. Evaluating the time derivative of $\bar{V}$, applying inequalities (37) and (38) (with $\kappa \to \infty$) as in App. A.1, and rearranging terms yields

$$\dot{\bar{V}} \leq -\bar{a}\bar{V} + \left(\frac{\bar{a}^3}{2} + \bar{a}\dot{\bar{a}}\right)|q|^2 + \left(\frac{3\bar{a}}{2} - 2\bar{d}\right)|p|^2$$
$$+ (2\bar{a}^2 - 2\bar{a}\bar{d} + \dot{\bar{a}})q^{\mathsf{T}}p$$
$$- \frac{1}{L}(1 - \bar{\beta}\bar{a})p^{\mathsf{T}}(\nabla f(y) - \nabla f(q)). \quad (43)$$

As in the strongly-convex case, the variable $\bar{a}(t)$ is deliberately chosen such that the cross term $q^{\mathsf{T}}p$ vanishes. Direct calculations show that $\bar{a}\dot{\bar{a}} + \bar{a}^3/2$ and $3\bar{a}/2 - 2\bar{d}$ vanish, which yields

$$\dot{\bar{V}} \leq -\bar{a}\bar{V} - \frac{1}{L}(1 - \bar{\beta}\bar{a})p^{\mathsf{T}}(\nabla f(y) - \nabla f(q)).$$

The smoothness and the convexity of f enables us to bound the last term, resulting in

$$\dot{\bar{V}} \leq \begin{cases} -\bar{a}\bar{V} - (1 - \bar{\beta}\bar{a})\bar{\beta}|p|^2, & t \in [0, 1) \\ -\bar{a}\bar{V}, & t \geq 1. \end{cases} \qquad (44)$$

The case analysis is due to the fact that $\bar{\beta} < 0$ for $t \in [0, 1)$ and $\bar{\beta} \geq 0$ for $t \geq 1$.

The dynamics are uniformly Lipschitz continuous, and hence the rate of growth of the trajectories is bounded. As a result, there exists a constant $\bar{C} > 0$ such that $\bar{V}(1) \leq \bar{C}(|q(0)|^2 + |p(0)|^2)$. Applying Grönwall's inequality yields

$$\bar{V}(t) \leq \frac{9}{(t + 2)^2}\bar{V}(1), \quad t \geq 1,$$

and therefore,

$$\frac{1}{L}f(q(t)) \leq \bar{V}(t) \leq \frac{9\bar{C}}{(t+2)^2}(|q(0)|^2 + |p(0)|^2),$$

for all $t \geq 1$, which establishes the proposition. It can be shown that $\bar{C} \leq 5/6$. ∎

A.3. Generalization of Prop. 4.1 to dimensions $n \geq 1$

We follow the notation and definitions of Rudin (1976) (Chapter 10), and for simplicity, we assume that the function f is sufficiently smooth. We say that $E \subset \mathbb{R}^{2n}$ is a smooth two-dimensional surface with boundary if there exists a smooth map $\Phi : \mathbb{R}^2 \to \mathbb{R}^{2n}$ such that $\Phi(S) = E$, where S is the unit simplex in $\mathbb{R}^2$.

Proposition A.1 *Let Γ_k be a smooth two-dimensional surface with boundary and define*

$$A_k := \int_{\Gamma_k} dq_k \wedge dp_k, \tag{45}$$

where $dq_k \wedge dp_k$ refers to the canonical symplectic form in $\mathbb{R}^{2n}$. The integration scheme (24) and (25) maps Γ_k to Γ_{k+1} such that

$$A_{k+1} - A_k = -T_s \int_{\partial\Gamma_k} f_{NP}(q_k, p_k)dq_k \tag{46}$$

$$= -T_s \int_{\Gamma_k} 2d + \frac{\beta}{L}\Delta f(q_k + \beta p_k) \, dq_k \wedge dp_k. \tag{47}$$

Proof Let $\psi : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ denote the map induced by (3) and (4), and let $\Phi' : \mathbb{R}^2 \to \mathbb{R}^{2n}$ be a smooth map such that $\Phi'(S) = \Gamma_k$, where S is the unit simplex in $\mathbb{R}^2$. Then $\Gamma_{k+1} = \psi(\Gamma_k) = \psi \circ \Phi'(S)$ is likewise a smooth two-dimensional surface with boundary.

The integration scheme (3) and (4) can be rewritten as a variational equality:

$$\delta q_k^{\mathsf{T}}(p_{k+1} - p_k) - \delta p_{k+1}^{\mathsf{T}}(q_{k+1} - q_k) + T_s\delta q_k^{\mathsf{T}}\nabla f(q_k) \\ + T_s\delta p_{k+1}^{\mathsf{T}}\nabla T(p_{k+1}) - T_s\delta q_k^{\mathsf{T}}f_{NP}(q_k, p_k) = 0, \tag{48}$$

for all $\delta q_k \in \mathbb{R}^n$ and $\delta p_{k+1} \in \mathbb{R}^n$. Let the points (q_k, p_k) along the boundary $\partial\Gamma_k = \Phi'(\partial S)$ be parametrized by $s \in [0, 1]$, that is $(q_k(s), p_k(s)) \in \partial\Gamma_k$ for all $s \in [0, 1]$ and $q_k(1) = q_k(0)$, $p_k(1) = p_k(0)$. For each $s \in [0, 1]$, every $(q_k(s), p_k(s))$ will be mapped to a pair $(q_{k+1}(s), p_{k+1}(s))$, which forms (by continuity of the integration scheme) the boundary of $\partial\Gamma_{k+1}$ (see, for example, (Rudin, 1976) (p. 93)). We choose $\delta q_k = dq_k/ds \, ds$ and $\delta p_{k+1} = dp_{k+1}/ds \, ds$ in (48) and integrate over $s \in [0, 1]$. This

yields, after rearranging terms,

$$\int_0^1 \left(-p_k^{\mathsf{T}}\frac{dq_k}{ds} - q_{k+1}^{\mathsf{T}}\frac{dp_{k+1}}{ds} - T_s f_{NP}^{\mathsf{T}}(q_k, p_k)\frac{dq_k}{ds} \right) \, ds \\ + \int_0^1 d\left(q_k^{\mathsf{T}}p_{k+1} + T_s H(q_k, p_{k+1})\right) = 0, \tag{49}$$

where the argument s has been omitted to shorten notation. The second part is a total differential and vanishes due to the fact that $\partial\Gamma_k$ and $\partial\Gamma_{k+1}$ are closed curves. This implies that

$$\int_{\partial\Gamma_k} p_k dq_k + \int_{\partial\Gamma_{k+1}} q_{k+1}dp_{k+1} = -T_s \int_{\partial\Gamma_k} f_{NP}(q_k, p_k)dq_k. \tag{50}$$

Applying Stokes' theorem (Rudin, 1976)(p. 272) to the left-hand side yields (46). The formula (47) results from applying Stokes' theorem to the right-hand side. ∎

A.4. Proof of Prop. 4.2

Let $\partial\Gamma(0)$ be the positively oriented contour along the boundary of an energy level set, and consider its evolution through the continuous-time dynamics, $\partial\Gamma(t)$, $t \in [0, T_s]$, and the discretized dynamics, $\partial\Gamma_1$. By the mean value theorem it follows from (31) that there exists a time instant $t_m \in (0, T_s)$ such that

$$A(T_s) = A(0) - T_s \int_{\Gamma(t_m)} 2d + \frac{\beta}{L}\Delta f(q + \beta p) \, dq dp, \tag{51}$$

where $A(t)$ refers to the area enclosed by $\partial\Gamma(t)$. Due to the fact that energy is dissipated as the continuous-time system evolves, it holds that $\Gamma(t) \subset \Gamma(0)$ for all $t > 0$. Therefore it follows that

$$A(T_s) \geq A(0) - T_s \int_{\Gamma(0)} 2d + \frac{\beta}{L}\Delta f(q + \beta p) \, dq dp \tag{52}$$

$$= A_1, \tag{53}$$

where A_1 denotes the (signed) phase-space area enclosed by $\partial\Gamma_1$. ∎

A.5. Proof of Prop. 4.3

The map ψ, induced by (3) and (4) is Lipschitz continuous due to the fact that ∇f is Lipschitz continuous. As mentioned earlier, ψ can be divided into two parts, a contraction step and a semi-implicit Euler step based on the conservative part of the system. We show that both steps are bijective, the result then follows by the invariance of domain theorem (Munkres, 2000).

We start with the contraction step

$$\bar{p}_{k+1} = p_k + T_s(-2dp_k - \frac{1}{L}(\nabla f(q_k + \beta p_k) - \nabla f(q_k))), \tag{54}$$

and apply fixed-point iteration to solve for p_k, given $q_k = \bar{q}_{k+1}$ and $\bar{p}_{k+1}$. Let p_k^j be the jth iterate, which is updated according to

$$p_k^{j+1} = \bar{p}_{k+1} + T_s(2dp_k^j + \frac{1}{L}(\nabla f(q_k + \beta p_k^j) - \nabla f(q_k)))$$
$$=: F_{\bar{p}_{k+1}, q_k}(p_k^j),$$

with $p_k^0 = 0$. Note that $F_{\bar{p}_{k+1}, q_k}$ is Lipschitz continuous, as for any $p_1, p_2 \in \mathbb{R}^n$,

$$|F_{\bar{p}_{k+1}, q_k}(p_1) - F_{\bar{p}_{k+1}, q_k}(p_2)| = |T_s(2d(p_1 - p_2) +$$
$$\frac{1}{L}(\nabla f(q_k + \beta p_1) - \nabla f(q_k + \beta p_2)))|$$
$$\leq T_s(2d + \beta)|p_1 - p_2|.$$

Due to the fact that $2d + \beta = 1$, the map F is a contraction for any $T_s \in (0, 1)$ and the fixed-point iteration is guaranteed to converge. Thus, $p_k = \lim_{j \to \infty} p_k^j$ exists and is unique, which shows that the contraction step is a bijection.

The semi-implicit Euler step according to (24) and (25) can be inverted in the following way:

$$\bar{q}_{k+1} = q_{k+1} - T_s \nabla T(p_{k+1}) \tag{55}$$

$$\bar{p}_{k+1} = p_{k+1} + T_s \frac{1}{L} \nabla f(\bar{q}_{k+1})$$
$$= p_{k+1} + T_s \frac{1}{L} \nabla f(q_{k+1} - T_s \nabla T(p_{k+1})), \tag{56}$$

which maps (q_{k+1}, p_{k+1}) to $(\bar{q}_{k+1}, \bar{p}_{k+1})$. ∎

A.6. Proof of Prop. 4.5

The proof is inspired by Nesterov (2004)(Section 2.2.1). Without loss of generality we set the Lipschitz constant L to one and introduce the following change of coordinates $(q_k, p_k) \to (\hat{q}_k, \hat{p}_k)$:

$$\hat{q}_k := q_k + \left(\frac{\beta}{1 - 2dT_s} - T_s \right) p_k, \quad \hat{p}_k := p_k. \tag{57}$$

The discrete-time dynamics expressed in the $(\hat{q}_k, \hat{p}_k)$-coordinates read as

$$\hat{q}_{k+1} = y_k - T_s \tau \nabla f(y_k), \tag{58}$$
$$\hat{p}_{k+1} = (1 - 2dT_s)\hat{p}_k - T_s \nabla f(y_k), \tag{59}$$

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

$$y_k = q_k + \beta p_k \tag{60}$$
$$= \hat{q}_k + \underbrace{\left(T_s - \frac{2d\beta T_s}{1 - 2dT_s} \right)}_{=:\hat{\beta}} \hat{p}_k. \tag{61}$$

The coordinate transformation is motivated by the fact that the smoothness and convexity of the objective function

can be used to conclude $f(\hat{q}_{k+1}) \leq f(y_k)$ (see also Nesterov (2004)(p. 76, General scheme of optimal method, step 1.c).)),

$$f(\hat{q}_{k+1}) - f(y_k) \leq -|\nabla f(y_k)|^2 \underbrace{\left(T_s \tau - \frac{T_s^2 \tau^2}{2} \right)}_{\geq 0, \forall T_s \in (0,1], \kappa \geq 1}. \tag{62}$$

We claim that the following function is a Lyapunov function

$$\hat{V}(\hat{q}_k, \hat{p}_k) = \frac{1}{2} \left| d\hat{q}_k + (1 - d\tau)\hat{p}_k \right|^2 + f(\hat{q}_k). \tag{63}$$

It can be checked that the function is indeed positive definite for all $T_s \in (0, 1]$ and all $\kappa \geq 1$. Moreover, the evolution of $\hat{V}$ along the discrete-time trajectories is given by

$$\hat{V}_{k+1} - \hat{V}_k \leq -dT_s \hat{q}_k^\mathsf{T} \nabla f(y_k) - \hat{\beta}\hat{p}_k^\mathsf{T} \nabla f(y_k)$$
$$+ C_{pn}\hat{p}_k^\mathsf{T} \nabla f(y_k) + C_{pp}|\hat{p}_k|^2 + C_{pq}\hat{p}_k^\mathsf{T}\hat{q}_k$$
$$+ C_{nn}|\nabla f(y_k)|^2 + f(y_k) - f(\hat{q}_k),$$

where (62) has been used to relate $f(\hat{q}_{k+1})$ to $f(y_k)$, $C_{pp}, C_{pn}, C_{pq}, C_{nn}$ are constants (dependent on κ and T_s), $\hat{V}_k$ denotes $\hat{V}(\hat{q}_k, \hat{p}_k)$, and $\hat{V}_{k+1}$ denotes $\hat{V}(\hat{q}_{k+1}, \hat{p}_{k+1})$. It follows from the strong convexity and the smoothness of f that

$$-y_k \nabla f(y_k) + \frac{1}{2\kappa}|y_k|^2 + f(y_k) \leq 0,$$
$$f(y_k) - \hat{\beta}\hat{p}_k^\mathsf{T} \nabla f(y_k) + \frac{1}{2\kappa}\hat{\beta}^2|\hat{p}_k|^2 \leq f(\hat{q}_k),$$

which can be used to conclude

$$- dT_s \hat{q}_k^\mathsf{T} \nabla f(y_k) - \hat{\beta}\hat{p}_k^\mathsf{T} \nabla f(y_k) + f(y_k) - f(\hat{q}_k)$$
$$\leq -dT_s f(\hat{q}_k) - \frac{dT_s}{2\kappa}|\hat{q}_k|^2 - \frac{d\hat{\beta}T_s}{\kappa}\hat{q}_k^\mathsf{T}\hat{p}_k - \frac{1}{2\kappa}\hat{\beta}^2|\hat{p}_k|^2.$$

As a result, the difference $\hat{V}_{k+1} - \hat{V}_k$ is upper bounded by

$$\hat{V}_{k+1} - \hat{V}_k \leq -dT_s \hat{V}_k + C_{pn} \nabla f(y_k)^\mathsf{T} \hat{p}_k$$
$$+ \left(C_{pp} - \frac{\hat{\beta}^2}{2\kappa} + \frac{dT_s}{2}(1 - d\tau)^2 \right) |\hat{p}_k|^2$$
$$+ \left(C_{pq} - \frac{dT_s\hat{\beta}}{\kappa} + d^2 T_s(1 - d\tau) \right) \hat{p}_k^\mathsf{T}\hat{q}_k$$
$$+ dT_s \left(\frac{d^2}{2} - \frac{1}{2\kappa} \right) |\hat{q}_k|^2 + C_{nn}|\nabla f(y_k)|^2.$$

Rearranging terms we obtain

$$\hat{V}_{k+1} - \hat{V}_k \leq -dT_s \hat{V}_k +$$
$$(\hat{p}_k^\mathsf{T}, \nabla f(y_k)^\mathsf{T}, \hat{q}_k^\mathsf{T})(M \otimes I)(\hat{p}_k^\mathsf{T}, \nabla f(y_k)^\mathsf{T}, \hat{q}_k^\mathsf{T})^\mathsf{T}, \tag{64}$$

where the matrix $M \in \mathbb{R}^{n \times n}$ is constant (depends on κ and T_{s}), $\otimes$ denotes the Kronecker product, and $I \in \mathbb{R}^{n \times n}$ the identity. It can be shown, for instance by applying Sylvester's criterion combined with Taylor expansions, that for every $T_{\mathrm{s}} \in (0,1)$, the matrix M is negative definite for $\kappa \to \infty$. For $T_{\mathrm{s}} = 1$, the matrix M is negative semi-definite for all $\kappa \geq 1$.[1] Thus, for every $T_{\mathrm{s}} \in (0,1]$, we obtain, for sufficiently large κ,

$$\hat{V}_{k+1} - \hat{V}_k \leq -dT_{\mathrm{s}}\hat{V}_k, \qquad (65)$$

yielding the desired result. ∎

---

[1]In fact, numerical evaluations indicate that M is negative semi-definite for all $\kappa \geq 3$ and all $T_{\mathrm{s}} \in (0,1]$.