Optimal local linear convergence of Nesterov's accelerated gradient method for C^2 functions under the Polyak–Łojasiewicz inequality

Anonymous authors
Paper under double-blind review

Abstract

In this work, we establish that Nesterov's accelerated gradient method, applied to C^2 functions satisfying the Polyak–Łojasiewicz (PL) inequality around local minimizers, achieves the optimal local linear convergence rate $\rho = \frac{\sqrt{3L+\mu}-2\sqrt{\mu}}{\sqrt{3L+\mu}} + \varepsilon$, where ε is an arbitrarily small constant. Our analysis is distinctive in requiring only C^2 smoothness of the objective function, avoiding stronger assumptions regarding higher-order derivatives or the geometry of the local minimizer submanifold. Notably, this requirement is only a slight departure from the standard $C^{1,1}$ assumption commonly adopted in the linear convergence analysis of first-order methods. Moreover, under identical conditions (C^2 smoothness and the PL inequality), the same analytical framework allows us to recover the optimal local exponential convergence rate $\sqrt{\mu}$ for the continuous-time Heavy Ball dynamics. Finally, a representative numerical experiment corroborates our theoretical findings.

1 Introduction

We consider the following unconstrained minimization problem

$$\text{Find} \quad x_* \in \arg\min_{x\in\mathbb{R}^d} f(x), \tag{1.1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is lower-bounded and twice continuously differentiable. For large-scale optimization, due to the relatively low computational cost per iteration, first-order algorithms are widely used . Among them Nesterov's accelerated gradient (NAG) methods Nesterov (2018) have attracted considerable attention for their potential to accelerate convergence toward local minimizers. Specifically, given two initial points $x^0, x^1 \in \mathbb{R}^d$, the NAG iterates are defined for $n \geq 1$ by

$$\begin{cases} y^n = x^n + \beta v_n,\ v_n := x^n - x^{n-1} \\ x^{n+1} = y^n - \alpha\nabla f(y^n) \end{cases}, \tag{1.2}$$

where $\alpha > 0$ is the stepsize and $\beta \in [0, 1)$ is the momentum parameter.

In the locally strongly convex setting, i.e., the Hessian of f at a local minimizer x_* satisfies the second-order sufficient condition ($\nabla^2 f(x_*) \succ 0$), it is well known that NAG can achieve an accelerated local convergence rate compared to standard gradient descent method (GD). Indeed, building on the analysis in Lessard et al. (2016), one can show that for optimally tuned hyperparameters α and β, and under the additional assumption that the iterates are initialized sufficiently close to x_*, the method converges linearly with asymptotic rate

$$\rho_{\text{NAG}} = \frac{\sqrt{3L+\mu}-2\sqrt{\mu}}{\sqrt{3L+\mu}} + \varepsilon,$$

where L and μ denote the largest and smallest eigenvalues of $\nabla^2 f(x_*)$, respectively, and $\varepsilon > 0$ is an arbitrarily small constant. In contrast, GD with optimal stepsize achieves only

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the optimal local linear rate

$$\rho_{\mathrm{GD}} = \frac{L-\mu}{L+\mu} + \varepsilon.$$

As shown in the inequality

$$\frac{\sqrt{3L+\mu} - 2\sqrt{\mu}}{\sqrt{3L+\mu}} \leq \frac{L-\mu}{L+\mu}, \quad \text{for all } L \geq \mu > 0,$$

with strict inequality when $L > \mu$, the NAG indeed provides a provably faster local convergence rate in this structured.

In many practical applications, the strong convexity assumption is frequently violated. A wide range of optimization problems of interest, such as those arising in machine learning, signal processing, and inverse problems, often exhibit non-isolated minimizers and may even admit infinitely many minimizers. This has motivated decades of research into the convergence behavior of momentum-based algorithms, particularly the Heavy Ball method and NAG, under assumptions weaker than strong convexity. To this end, several classical relaxed conditions have been proposed. Among the most prominent is the Polyak–Łojasiewicz (PL) inequality, independently introduced by Polyak Polyak (1963) and Łojasiewicz Łojasiewicz (1963). Owing to its simplicity, broad applicability, and validity even for certain non-convex functions, the PL inequality has become a widely used tool for establishing linear convergence of gradient-based algorithms without requiring strong convexity. Beyond the PL condition, other relaxations of strong convexity have been studied, including the error bound property Luo & Tseng (1993), quadratic growth Anitescu (2000), the restricted secant inequality Zhang & Yin (2013), and essential strong convexity Liu et al. (2015). In Necoara et al. (2019), the authors further discuss a family of closely related conditions, such as quasi-(weak) strong convexity, quadratic functional growth, quadratic gradient growth, and quadratic under-approximation. Under the assumption that f has a Lipschitz-continuous gradient, these conditions often imply one another (see (Necoara et al., 2019, Theorem 4) and (Karimi et al., 2016, Theorem 2)). Moreover, if one further assumes that $f \in C^2$, then according to the recent results in Rebjock & Boumal (2025) and (Kassing & Weissmann, 2024, Lemma 3.2), all these conditions are equivalent in a sufficiently small neighborhood of minimizers.

A central question regarding the aforementioned weakened strong convexity conditions is whether momentum-based algorithms, such as the Heavy Ball method and NAG, can achieve acceleration over standard gradient descent. Here, by "acceleration" we mean that the linear convergence rate ρ improves from depending on $\frac{\mu}{L}$ to depending on $\sqrt{\frac{\mu}{L}}$. To the best of our knowledge, under only the assumptions that f satisfies one of the aforementioned conditions and that ∇f is L-Lipschitz continuous or $f \in C^{1,1}$, no general theoretical result guarantees such acceleration for momentum methods. In fact, negative results exist: Yue et al. (2023) shows that for general PL functions with L-Lipschitz gradients, momentum-based algorithms cannot achieve global acceleration. Nevertheless, acceleration can be recovered under additional structural assumptions on f. All known positive results assume that at least ∇f is L-Lipschitz continuous or that $f \in C^{1,1}$. For the Heavy Ball method, Wang et al. (2022) establishes provable acceleration for a class of non-convex PL functions whose non-convexity is "averaged out" in a suitable sense. More recently, Kassing & Weissmann (2024) shows that if $f \in C^4$ satisfies the PL inequality around its minimizer set, then the Heavy Ball method recovers the optimal asymptotic local convergence rate originally derived by Polyak Polyak (1964) for C^2 strongly convex functions, thereby achieving local acceleration. For NAG, Necoara et al. (2019) proves acceleration of NAG for quasi-strongly convex functions, provided that the auxiliary iterates share the same projection onto the set of minimizers. Aujol et al. (2024) demonstrates acceleration of NAG for convex functions satisfying the quadratic growth condition. Hermant et al. (2024) establishes acceleration of NAG for γ-strongly quasar-convex functions (which reduce to quasi-strongly convex when $\gamma = 1$) under an additional curvature condition. Gupta & Wojtowytsch (2025) proves acceleration of NAG for quasi-strongly convex functions under the assumption that the set of minimizers is a C^2 embedded submanifold of $\mathbb{R}^d$ and that the projection mapping onto this set satisfies a mild regularity condition. More recently, Hermant (2026) clarifies provable acceleration by momentum for PL functions under an aiming condition.

Our contribution. In this work, we prove that NAG achieves local acceleration under the assumption that f satisfies the PL inequality in a neighborhood of its minimizer set and that $f \in C^2$. Specially, we recover the optimal asymptotic local convergence rate

$$\rho = \frac{\sqrt{3L+\mu} - 2\sqrt{\mu}}{\sqrt{3L+\mu}} + \varepsilon,$$

where $\varepsilon > 0$ is an arbitrarily small constant, which matches the sharp rate established by Lessard et al. (2016). To the best of our knowledge, this is the first result establishing such a sharp local acceleration rate for NAG under merely C^2 smoothness and a local PL condition. Our analysis, distinct from existing methods that require high-order smoothness (e.g., $f \in C^4$), relies on a novel use of the a priori linear convergence rate. This approach enables a refined spectral characterization under merely $f \in C^2$, without imposing additional geometric regularity on the local minimizer submanifold. Moreover, leveraging this idea, we also establish the optimal local exponential decay rate $\sqrt{\mu}$ for the continuous-time Heavy Ball accelerated flow trajectory associated with NAG. Notably, this result holds under C^2 smoothness assumption, thereby relaxing the stringent C^4 regularity requirement imposed in Kassing & Weissmann (2024).

The remainder of this paper is organized as follows. Section 2 introduces the necessary notation and assumptions, recalls key properties of the objective function under these assumptions, and presents the main results. In Section 3, we validate the theoretical findings through numerical experiments. Section 4 concludes the paper. Complete proofs of the main results are deferred to Appendix A.

2 Preliminaries and main results

In this section, we first introduce problem settings, basic notations, and some important properties of the problem. Then, the main results are presented.

2.1 Problem settings, notations, and properties

The goal of this paper is to analyze the local convergence rate of iterative algorithms near minimizers. To this end, in addition to assuming that the objective function f in equation 1.1 is bounded from below and of class C^2, we introduce the following local PL condition.

Definition 2.1. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a C^2 function, and let $x_* \in \mathbb{R}^d$ be a minimizer of f. We say that f satisfies the local PL condition at x_* if there exist a sufficiently small open neighborhood $U \subset \mathbb{R}^d$ of x_*, and a constant $\nu > 0$ such that

$$\|\nabla f(x)\|^2 \geq 2\nu\big(f(x) - f(x_*)\big), \quad \text{for all } \ x \in U,$$

where $\|\cdot\|$ denotes the Euclidean norm.

Note that the L-Lipschitz continuity of the gradient is not assumed explicitly here, as it is automatically satisfied on the bounded set U for a C^2 function.

In particular, it is worth noting that for a C^2 objective function, if it is coercive or its set of minimizers is bounded, then recent results Criscitiello et al. (2025); Nejma (2025) show that the global PL condition guarantees uniqueness of the minimizer. Thus, the global PL condition is, in some sense, a rather strong assumption.

Next, we introduce the following set and associated function value level constructed from a local minimizer x_*:

$$\mathcal{S} := \left\{ x \in \mathbb{R}^d \,\middle|\, x \text{ is a local minimizer of } f \text{ and } f(x) = f(x_*) \right\}. \tag{2.3}$$

In the work of Rebjock & Boumal (2025), it was shown that the local PL condition defined in equation 2.1 is equivalent to the following local Morse–Bott condition.

Definition 2.2. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a C^2 function. We say that f satisfies the local Morse–Bott condition near a local minimizer x_* if there exists a sufficiently small open neighborhood $U \subset \mathbb{R}^d$ of x_* such that the set

$$\mathcal{S}_U := \mathcal{S} \cap U$$

is a C^1 embedded submanifold of $\mathbb{R}^d$, and for every $x'_* \in \mathcal{S}_U$,

$$K_{x'_*} := \ker\left(\nabla^2 f(x'_*)\right) = \left\{ v \in \mathbb{R}^d \,\middle|\, \nabla^2 f(x'_*)[v] = 0 \right\} = T_{x'_*}\mathcal{S}_U.$$

Here, the tangent space $T_{x'_*}\mathcal{S}_U$ is defined by

$$T_{x'_*}\mathcal{S}_U := \left\{ \gamma'(0) \in \mathbb{R}^d \,\middle|\, \gamma : (-\varepsilon, \varepsilon) \to \mathcal{S}_U \text{ is a } C^1 \text{ curve with } \gamma(0) = x'_* \right\}.$$

Remark 2.1. If $\dim T_{x'_*}\mathcal{S}_U = d$ for some (hence all) $x'_* \in \mathcal{S}_U$, then $\mathcal{S}_U$ is an open subset of $\mathbb{R}^d$. In this case, the condition $f(x'_*) = f(x_*)$ for all $x'_* \in \mathcal{S}_U$ implies that f is constant on a neighborhood of x_*. We exclude this trivial scenario throughout the paper, as it leads to a completely flat landscape. Consequently, we always assume that $\mathcal{S}_U$ is a proper submanifold of $\mathbb{R}^d$, i.e., $\dim T_{x'_*}\mathcal{S}_U < d$.

In addition, we introduce the normal space

$$N_{x'_*}\mathcal{S}_U := \left(T_{x'_*}\mathcal{S}_U\right)^{\perp},$$

and let $P_{x'_*} : \mathbb{R}^d \to N_{x'_*}\mathcal{S}_U$ denote the orthogonal projection onto this normal space. In the nontrivial case, i.e., when $\mathcal{S}_U$ is not an open subset of $\mathbb{R}^d$, the local Morse–Bott condition implies that for every $x'_* \in \mathcal{S}_U$, the Hessian $\nabla^2 f(x'_*)$ is nondegenerate on $N_{x'_*}\mathcal{S}_U$. Consequently, its eigenvalues restricted to $N_{x'_*}\mathcal{S}_U$ satisfy

$$0 < \mu(x'_*) := \lambda_{\min}\left(\nabla^2 f(x'_*)|_{N_{x'_*}\mathcal{S}_U}\right) \le \lambda_{\max}\left(\nabla^2 f(x'_*)|_{N_{x'_*}\mathcal{S}_U}\right) =: L(x'_*).$$

Note that $\lambda_{\max}\left(\nabla^2 f(x'_*)|_{N_{x'_*}\mathcal{S}_U}\right)$ is precisely the largest eigenvalue of the full Hessian $\nabla^2 f(x'_*)$ on $\mathbb{R}^d$. Indeed, under the local Morse–Bott condition at a local minimizer $x'_* \in \mathcal{S}_U$, the Hessian vanishes on the tangent space $T_{x'_*}\mathcal{S}_U$, and $\mathbb{R}^d$ decomposes orthogonally as $T_{x'_*}\mathcal{S}_U \oplus N_{x'_*}\mathcal{S}_U$. Hence all nonzero eigenvalues of $\nabla^2 f(x'_*)$ are contained in the normal subspace, and the maximum eigenvalue is attained there.

We also emphasize that the local constants $\mu(x'_*)$ and $L(x'_*)$ generally depend on the base point $x'_* \in \mathcal{S}_U$. However, our analysis is purely local. Since f is of class C^2 and $\mathcal{S}_U$ is a C^1 embedded submanifold, the tangent spaces $T_{x'_*}\mathcal{S}_U$, and hence the normal spaces $N_{x'_*}\mathcal{S}_U$, vary continuously with x'_*. It follows that the orthogonal projections $P_{x'_*}$ depend continuously on x'_*, and so do the extremal eigenvalues $\mu(x'_*), L(x'_*)$ by the continuity of the map $x'_* \mapsto P_{x'_*}\nabla^2 f(x'_*)P_{x'_*}$. Therefore, on any sufficiently small neighborhood of the reference point x_*, the quantities $\mu(x'_*)$ and $L(x'_*)$ remain uniformly bounded away from zero and infinity. This mild dependence on the base point can be absorbed into the small parameter $\varepsilon > 0$, which governs both the stepsize restriction and the asymptotic convergence rate. For notational simplicity, we henceforth suppress the explicit dependence on x'_* and denote these uniform local bounds simply by μ and L, which, up to the negligible perturbation governed by ε, coincide with the standard PL constant μ and Lipschitz gradient constant L in the usual sense.

Finally, for notational convenience, we adopt the following conventions throughout the paper:

- The symbol $C > 0$ denotes a generic positive constant that is independent of the iterate x and may vary from line to line.
- We write $a \approx b$ for two nonnegative quantities a, b if there exist constants $C_1, C_2 > 0$, independent of x, such that

$$C_1 b \le a \le C_2 b.$$

2.2 Main results

In this subsection, under the local PL condition, we establish a sharp local linear convergence rate for NAG around local minimizers. Despite the failure of the second-order sufficient conditions due to degeneracy of the Hessian, we show that NAG achieves the optimal local contraction factor that matches the optimal one for strongly convex quadratic problems (see Lessard et al. (2016)) up to an arbitrarily small $\varepsilon > 0$. This result highlights the robustness of momentum-based acceleration in nonconvex optimization when the local PL condition is properly exploited.

We consider the local dynamics of NAG near local minimizers x_*. To this end, we introduce the following Lyapunov function : for all $x, v \in \mathbb{R}^d$,

$$\mathcal{L}(x, v) := f(x) - f(x_*) + \frac{\beta}{2\alpha}\|v\|^2 - \frac{\beta^2}{2(1+\beta)}\nabla^2 f(x_*)[v, v], \tag{2.4}$$

where $\nabla^2 f(x_*)[v, v] := \langle \nabla^2 f(x_*)[v], v\rangle$. The following result establishes local Lyapunov stability for this sequence.

Theorem 2.1. Let $f \in C^2$ satisfy the local PL condition at the local minimizer $x_* \in \mathcal{S}$. Then, for every sufficiently small $\varepsilon > 0$, there exists a sufficiently small open neighborhood $U \subset \mathbb{R}^d$ of x_* such that for any $x^0, x^1 \in U$, the discrete Lyapunov sequence generated by NAG satisfies

$$0 \le \mathcal{L}(x^{n+1}, v_{n+1}) \le \rho_{\varepsilon,\alpha,\beta}\, \mathcal{L}(x^n, v_n) \quad \text{for all } n \ge 1,$$

provided that the step size α satisfies

$$\alpha \in \left(0, \frac{2(\beta+1)}{(L+\varepsilon)(2\beta+1)}\right), \quad \text{for all } \beta \in [0, 1),$$

where $\rho_{\varepsilon,\alpha,\beta} \in (0, 1)$ is a contraction factor depending on ε, α, and β.

The theorem above implies the local linear convergence of NAG, yet the contraction factor cannot be precisely quantified. Next, for NAG, the optimal local convergence rate is obtained in the following theorem.

Theorem 2.2. Let $f \in C^2$ satisfy the local PL condition at the local minimizer $x_* \in \mathcal{S}$. Then, for every sufficiently small $\varepsilon > 0$, there exists a sufficiently small open neighborhood $U \subset \mathbb{R}^d$ of x_* such that for any $x^0, x^1 \in U$, the sequence $\{x^n\}_{n\in\mathbb{N}}$ generated by NAG converges linearly to a local minimizer $x'_* \in \mathcal{S}$, i.e.,

$$\|x^n - x'_*\| \le C_\varepsilon\big(\|x^1 - x'_*\| + \|x^0 - x'_*\|\big)(\rho_{\alpha,\beta} + \varepsilon)^n, \quad \forall\, n \ge 1,$$

for the step size α satisfying

$$\alpha \in \left(0, \frac{2(\beta+1)}{(L+\varepsilon)(2\beta+1)}\right), \quad \text{for all } \beta \in [0, 1),$$

where $\rho_{\alpha,\beta} \in (0, 1)$ is a contraction factor depending on α and β, and $C_\varepsilon > 0$ is a constant depending on ε. In particular, with the optimal parameter choice

$$\alpha_{\text{opt}} = \frac{4}{3L+\mu} \quad \text{and} \quad \beta_{\text{opt}} = \frac{\sqrt{3L+\mu} - 2\sqrt{\mu}}{\sqrt{3L+\mu} + 2\sqrt{\mu}},$$

NAG achieves the optimal linear convergence rate

$$\|x^n - x'_*\| \le C_\varepsilon\big(\|x^1 - x'_*\| + \|x^0 - x'_*\|\big)\left(\frac{\sqrt{3L+\mu} - 2\sqrt{\mu}}{\sqrt{3L+\mu}} + \varepsilon\right)^n, \quad \forall\, n \ge 1. \tag{2.5}$$

In essence, the second theorem refines the first by providing a precise quantification of the contraction factor, building directly on the local linear convergence established therein. Notably, this approach extends seamlessly to the Heavy Ball ordinary differential equation, namely

$$\ddot{x}(t) + \gamma\dot{x}(t) + \nabla f(x(t)) = 0, \tag{2.6}$$

yielding the optimal exponential local convergence rate $\sqrt{\mu}$ with $\gamma = 2\sqrt{\mu}$ for its trajectories under the assumption that $f \in C^2$ and satisfies the local PL condition.

Theorem 2.3. Let $f \in C^2$ satisfy the local PL condition at the local minimizer $x_* \in \mathcal{S}$. Then, for every sufficiently small $\varepsilon > 0$, there exists a sufficiently small open neighborhood $\mathcal{U} \subset \mathbb{R}^d \times \mathbb{R}^d$ of $(x_*, 0)$ such that for any initial condition $(x(0), \dot{x}(0)) \in \mathcal{U}$, the trajectories $x(t)$ generated by equation 2.6 converges exponentially to a local minimizer $x'_* \in \mathcal{S}$, i.e.,

$$\|x(t) - x'_*\| \le C_\varepsilon \big(\|x(0) - x'_*\| + \|\dot{x}(0)\|\big) e^{-(\rho_\gamma - \varepsilon)t}, \quad \forall\, t \ge 0,$$

where $\rho_\gamma = \frac{\gamma - \sqrt{\max\{0, \gamma^2 - 4\mu\}}}{2}$ is a convergence rate depending on $\gamma > 0$, and $C_\varepsilon > 0$ is a constant depending on ε. In particular, with the optimal damping parameter

$$\gamma = 2\sqrt{\mu},$$

the flow equation 2.6 achieves the optimal cxponential convergence rate

$$\rho_{\text{opt}} = \sqrt{\mu}. \tag{2.7}$$

3 Numerical experiments

In this section, we will introduce a series of numerical experiments to verify our theoretical results.

3.1 Test function and its theoretical properties

To numerically validate the optimal local linear convergence of NAG established in Theorem 2.2, we consider a class of nonconvex objective functions that are smooth, possess a manifold of minimizers, and satisfy the local PL condition near the minimizer set by the local Morse–Bott condition.

Specifically, let $u \in \mathbb{R}^k$, $v \in \mathbb{R}^m$, and define

$$f(u, v) := \frac{1}{2} u^\top \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_k) u + \frac{1}{8} \lambda_{k+1} \big(\|v\|^2 - 1\big)^2,$$

where the parameters satisfy

$$0 < \lambda_1 \le \lambda_2 \le \cdots \lambda_i \le \cdots \lambda_k \le \lambda_{k+1} \quad \text{for all } i = 1, \ldots, k+1.$$

We now verify the key properties of f.

(i) Nonconvexity.

The Hessian of f is given by

$$\nabla^2 f(u, v) = \begin{pmatrix} \operatorname{diag}(\lambda_1, \ldots, \lambda_k) & 0 \\ 0 & \lambda_{n+1} \left(\frac{1}{2}(\|v\|^2 - 1) I_m + v v^\top\right) \end{pmatrix}.$$

At the point $(u, v) = (0, 0)$, the lower-right block reduces to $-\frac{\lambda_{k+1}}{2} I_m$, which is negative definite. Hence $\nabla^2 f(0, 0)$ has negative eigenvalues, and f is nonconvex.

(ii) Local non-strong convexity and Morse–Bott structure.

The global minimizer set of f is the smooth submanifold

$$\mathcal{M} := \left\{ (u, v) \in \mathbb{R}^k \times \mathbb{R}^m : u = 0,\ \|v\|^2 = 1 \right\}.$$

The gradient vanishes precisely on $\mathcal{M}$ and at the isolated saddle point $(0, 0)$. Thus, f has no other local minima, and it is not strongly convex in any neighborhood of a minimizer since the minimizers are not isolated.

At any $(0, v_*) \in \mathcal{M}$ (so $\|v_*\| = 1$), the Hessian simplifies to

$$\nabla^2 f(0, v_*) = \begin{pmatrix} \operatorname{diag}(\lambda_1, \ldots, \lambda_k) & 0 \\ 0 & \lambda_{k+1} v_* v_*^\top \end{pmatrix}.$$

Its spectrum consists of:

- positive eigenvalues $\lambda_1, \ldots, \lambda_k$ from the u-block;
- one positive eigenvalue λ_{k+1} in the normal direction to the sphere $\|v\| = 1$;
- zero eigenvalues (multiplicity $m-1$) along the tangent directions of the sphere.

Hence, $\nabla^2 f(0, v_*)$ is positive semidefinite with constant rank, and its kernel coincides with the tangent space of $\mathcal{M}$. This confirms that f is a Morse–Bott function. Moreover, since the local PL condition is equivalent to the local Morse–Bott condition for smooth functions, f naturally satisfies the local PL condition near $\mathcal{M}$.

Finally, the smallest positive eigenvalue and the largest eigenvalue of $\nabla^2 f$ on $\mathcal{M}$ are

$$\mu := \min_{i=1,2,\cdots,k+1} \{\lambda_i\} = \lambda_1, \qquad L := \max_{i=1,2,\cdots,k+1} \{\lambda_i\} = \lambda_{k+1}. \tag{3.8}$$

To sum up, f satisfies all structural assumptions required by our theory.

3.2 Numerical results

We now instantiate the test function f with the following concrete parameters:

$$k = 10^2, 10^3, 10^4, 10^5, \quad m = 10, \quad \lambda_i = i \quad \text{for } i = 1, 2, \ldots, k+1.$$

Consequently, the curvature constants on the minimizer manifold $\mathcal{M}$ are

$$\mu = \lambda_1 = 1, \qquad L = \lambda_{k+1} = k+1,$$

yielding a condition number $\kappa = L/\mu = 1/(k+1)$.

The NAG iterates are initialized at a point (u^0, v^0) in a neighborhood of $\mathcal{M}$:

$$u^0 = \varepsilon_u \cdot \xi_u, \quad v^0 = (1 + \varepsilon_v) \cdot \frac{\xi_v}{\|\xi_v\|},$$

where $\xi_u \in \mathbb{R}^k$ and $\xi_v \in \mathbb{R}^m$ are standard Gaussian random vectors, and we set $\varepsilon_u = \varepsilon_v = 10^{-4}$ to ensure the initial point lies sufficiently close to $\mathcal{M}$ for local convergence to manifest.

We apply NAG with the optimal parameters, namely

$$\alpha_{\text{opt}} = \frac{4}{3L + \mu}, \qquad \beta_{\text{opt}} = \frac{\sqrt{3L+\mu} - 2\sqrt{\mu}}{\sqrt{3L+\mu} + 2\sqrt{\mu}},$$

which yield the asymptotically optimal accelerated convergence rate

$$\rho_{\text{opt}} = \frac{\sqrt{3L+\mu} - 2\sqrt{\mu}}{\sqrt{3L+\mu}}.$$

These choices are consistent with the local analysis in our theorem. The algorithm is terminated when

$$\left| f(u^n, v^n) \right| < 10^{-20}$$

is reached.

Figure 1 displays the evolution of the momentum norm $\|x^{n+1} - x^n\|$ versus the iteration count n, plotted on a logarithmic scale. We emphasize that, the momentum $\{x^{n+1} - x^n\}_{n \in \mathbb{N}}$ converges to zero at a ρ-linear rate if and only if the underlying iterates $\{x^n\}_{n \in \mathbb{N}}$ themselves converge at the same ρ-linear rate. Consequently, monitoring the decay of $\|x^{n+1} - x^n\|$ provides a reliable proxy for assessing the convergence behavior of the sequence.

As shown in the figure, the blue dashed curves—representing the empirical error $\|x^{n+1} - x^n\|$ across different problem scales ($k = 10^2, 10^3, 10^4, 10^5$)—asymptotically align with the red solid line, which corresponds to the theoretical optimal linear rate ρ_{opt}. This alignment becomes increasingly pronounced in the later iterations, indicating that the algorithm enters its local linear convergence regime. In accordance with Theorem 2.2, both the theoretical prediction and numerical observation confirm a local ρ_{opt}-linear decay, where the asymptotic

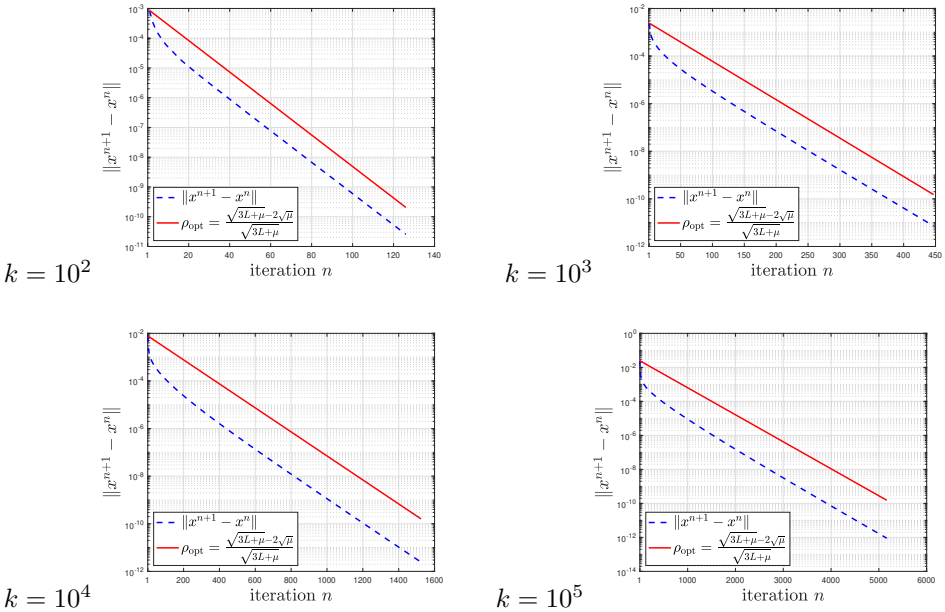


Figure 1: Convergence behavior of NAG for $k = 10^2$, 10^3, 10^4, and 10^5. In each subplot, the red solid line represents the theoretical optimal convergence rate ρ_{opt}, while the blue dashed line shows the actual error $\|x^{n+1} - x^n\|$.

higher-order terms (e.g., ε-level perturbations) have been omitted for clarity. Notably, the observed slope in the semi-log plot closely matches the analytically derived rate

$$\rho_{\text{opt}} = \frac{\sqrt{3L+\mu} - 2\sqrt{\mu}}{\sqrt{3L+\mu}},$$

which is known to be the optimal local convergence factor for NAG in strongly convex settings.

It is worth emphasizing that our experimental setting deliberately excludes the classical cases in which the local PL condition follows from (local) strong convexity. Instead, we consider genuinely nonconvex scenarios that are not locally strongly convex, yet still satisfy a local PL condition near the solution. This provides strong numerical evidence that NAG retains its optimal local acceleration rate beyond the traditional strongly convex framework—specifically, in regions where the PL inequality holds without requiring strong convexity. Consequently, the results corroborate both the robustness and sharpness of the guarantees established in Theorem 2.2.

4 Conclusion

In this paper, we establish the optimal local linear convergence rate of Nesterov's Accelerated Gradient method (NAG) for solving locally non-strongly convex unconstrained optimization problems, where the objective function is C^2 and satisfies the local Polyak–Łojasiewicz (PL) condition in a neighborhood of a local minimizer. This result naturally demonstrates that NAG retains its acceleration capability even in the absence of strong convexity within a local nonconvex regime. Notably, under the assumption of the local PL condition, our analysis requires only a mild strengthening of the standard regularity assumption: we replace the $C^{1,1}$ condition (i.e., locally Lipschitz continuous gradient) commonly used with C^2 smoothness. This modest requirement renders our framework relatively mild and widely applicable. Finally, our numerical experiments are conducted on instances that deliberately exclude trivial cases such as locally strongly convex objectives. Specifically, we consider genuinely nonconvex and non-locally-strongly-convex functions that still satisfy the local

PL condition near the solution. The close match between theory and practice in these representative scenarios strongly corroborates the validity, generality, and sharpness of our convergence guarantees.

AI use statement

In this work, we have not used generative AI tools .

References

M. Anitescu. Degenerate nonlinear programming with a quadratic growth condition. SIAM J. Optim., 10:1116–1135, 2000.

J.-F. Aujol, C. Dossal, H. Labariére, and A. Rondepierre. Heavy ball momentum for non-strongly convex optimization. arXiv preprint arXiv:2403.06930, 2024.

C. Criscitiello, Q. Rebjock, and N. Boumal. If a smooth function is globally Polyak–Łojasiewicz and coercive, then it has a unique minimizer. Race to the Bottom, 2025. URL https://www.racetothebottom.xyz/posts/PL-smooth-unique/.

K. Gupta and S. Wojtowytsch. Nesterov acceleration in benignly non-convex landscapes. In The International Conference on Learning Representations, 2025.

J. Hermant. Acceleration for Polyak–Łojasiewicz functions with a gradient aiming condition. arXiv preprint arXiv:2602.10022, 2026.

J. Hermant, J.-F. Aujol, C. Dossal, and A. Rondepierre. Study of the behaviour of Nesterov accelerated gradient in a non-convex setting: the strongly quasar convex case. arXiv preprint arXiv:2405.19809, 2024.

H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak–Łojasiewicz condition. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 795–811. Springer, 2016.

S. Kassing and S. Weissmann. Polyak's heavy ball method achieves accelerated local rate of convergence under polyak–lojasiewicz inequality. arXiv preprint arXiv:2410.16849, 2024.

L. Lessard, B. Recht, and A. Packard. Analysis and design of optimization algorithms via integral quadratic constraints. SIAM J. Optim., 26:57–95, 2016.

J. Liu, S. Wright, C. Re, V. Bittorf, and S. Sridhar. An asynchronous parallel stochastic coordinate descent algorithm. J. Mach. Learn. Res., 16:285–322, 2015.

S. Łojasiewicz. A topological property of real analytic subsets. In Coll. du CNRS, Les équations aux dérivées partielles, volume 117, pp. 87–89, 1963.

Z.-Q. Luo and P. Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. Ann. Oper. Res., 46:157–178, 1993.

I. Necoara, Y. Nesterov, and F. Glineur. Linear convergence of first order methods for non-strongly convex optimization. Math. Program., 175:69–107, 2019.

A. B. Nejma. Polyak–Łojasiewicz inequality is essentially no more general than strong convexity for functions. arXiv preprint arXiv:2512.05285, 2025.

Y. Nesterov. Lectures on Convex Optimization. Springer, Cham, 2nd edition, 2018.

B. T. Polyak. Gradient methods for the minimisation of functionals. USSR Comput. Math. and Math. Phys., 3:864–878, 1963.

B. T. Polyak. Some methods of speeding up the convergence of iteration methods. USSR Comput. Math. and Math. Phys., 4:1–17, 1964.

B. T. Polyak and P. Shcherbakov. Lyapunov functions: An optimization theory perspective. IFACPapersOnLine, 50:7456–7461, 2017.

Q. Rebjock and N. Boumal. Fast convergence to non-isolated minima: four equivalent conditions for C^2 functions. Math. Program., 213:151–199, 2025.

J.-K. Wang, C.-H. Lin, A. Wibisono, and B. Hu. Provable acceleration of heavy ball beyond quadratics for a class of Polyak–Łojasiewicz functions when the non-convexity is averaged out. In Proceedings of the 39th International Conference on Machine Learning, pp. 22839–22864. PMLR, 2022.

P. Yue, C. Fang, and Z. Lin. On the lower bound of minimizing Polyak–Łojasiewicz functions. In The Thirty Sixth Annual Conference on Learning Theory, pp. 2948–2968. PMLR, 2023.

H. Zhang and W. Yin. Gradient methods for convex minimization: better rates under weaker conditions. arXiv preprint arXiv:1303.4645, 2013.

A Appendix

In this appendix, we provide complete proofs of all the preceding results.

A.1 Technical lemmas

Before presenting the proof, we introduce several key lemmas that will be instrumental in establishing various aspects of our results. Firstly, under the local PL condition, we provide the upper and lower bounds for this Lyapunov function equation 2.4.

Lemma A.1. Let $f \in C^2$ satisfy the local PL condition at the local minimizer $x_* \in \mathcal{S}$. Then, there exists a sufficiently small open neighborhood $U \subset \mathbb{R}^d$ of x_* such that for all $x \in U$, the Lyapunov function $\mathcal{L}(x, v)$ defined in equation 2.4 satisfies

$$\mathcal{L}(x,v) \approx \left(\left\|\nabla f(x)\right\|^2 + \|v\|^2\right), \quad \forall\, \alpha \in \left(0, \frac{1+\beta}{(L+\varepsilon)\beta}\right).$$

Here, $A \approx B$ means that there exist constants $c_1, c_2 > 0$ such that

$$c_1 B \le A \le c_2 B.$$

Proof. First, since f satisfies the local PL condition at the local minimizer x_*, there exists $\sigma_1 > 0$ such that for all $x \in \mathcal{B}_{\sigma_1}(x_*) := \{y \in \mathbb{R}^d : \|y - x_*\| < \sigma_1\}$, the PL inequality holds. Together with the non-negativity of the Hessian at x_*, i.e., $\nabla^2 f(x_*)[v, v] \ge 0$, we immediately obtain the upper bound

$$\mathcal{L}(x,v) \le \frac{1}{2\nu}\left\|\nabla f(x)\right\|^2 + \frac{\beta}{2\alpha}\|v\|^2 \le C\left(\left\|\nabla f(x)\right\|^2 + \|v\|^2\right).$$

Next, by the boundedness of the Hessian at x_*, i.e., $\nabla^2 f(x_*)[v, v] \le L\|v\|^2$, it follows that

$$\left(\frac{\beta}{2\alpha} - \frac{\beta^2 L}{2(1+\beta)}\right)\|v\|^2 \ge C\|v\|^2, \quad \text{whenever} \quad \alpha \in \left(0, \frac{1+\beta}{L\beta}\right).$$

Finally, by (Rebjock & Boumal, 2025, Lemma 1.4, Proposition 2.2), there exists $\sigma_2 > 0$ such that for all $x \in \mathcal{B}_{\sigma_2}(x_*)$, the projection of x onto the set $\mathcal{S}$, denoted $P_{\mathcal{S}}(x)$, is nonempty and the quadratic growth condition holds:

$$f(x) - f(x_*) = f(x) - f(P_{\mathcal{S}}(x)) \ge C\|x - P_{\mathcal{S}}(x)\|^2 = C\,\mathrm{dist}(x, \mathcal{S})^2.$$

Noting that $\|x - x_*\| \ge \|x - P_{\mathcal{S}}(x)\|$ and combining this with the local Lipschitz continuity of $\nabla f(x)$, there exists a sufficiently small constant $\sigma \in (0, \min\{\sigma_1, \sigma_2\}]$ such that

$$f(x) - f(x_*) \ge C\left\|\nabla f(x)\right\|^2, \quad \forall\, x \in \mathcal{B}_{\sigma}(x_*).$$

Combining the above estimates, we conclude the lower bound

$$\mathcal{L}(x,v) \geq C\left(\left\|\nabla f(x)\right\|^2 + \|v\|^2\right),$$

which together with the upper bound yields the desired equivalence. □

To prove Theorem 2.2, we adopt the same decomposition strategy as in Gupta & Wojtowytsch (2025); Kassing & Weissmann (2024), splitting the full space into tangent and normal components to analyze the convergence rate. We now introduce the matrix $\mathcal{G}_\alpha(x'_*) : N_{x'_*}\mathcal{S}_U \to N_{x'_*}\mathcal{S}_U$ defined by

$$\mathcal{G}_\alpha(x'_*) := \big(I - \alpha\nabla^2 f(x'_*)\big)P_{x'_*}.$$

We also define the block matrix $G_{\alpha,\beta}(x'_*)$ acting on the invariant subspace $\mathcal{N}_{x'_*} := N_{x'_*}\mathcal{S}_U \times N_{x'_*}\mathcal{S}_U \times T_{x'_*}\mathcal{S}_U$ as

$$G_{\alpha,\beta}(x'_*) := \begin{pmatrix} (1+\beta)\,\mathcal{G}_\alpha(x'_*) & -\beta\,\mathcal{G}_\alpha(x'_*) & \mathbf{0} \\ I & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \beta\, I \end{pmatrix}. \tag{1.9}$$

The spectral radius of the matrix $G_{\alpha,\beta}(x'_*)|_{\mathcal{N}_{x'_*}}$ is characterized as follows.

Lemma A.2. Let $f \in C^2$ satisfy the local PL condition at the local minimizer x_*. Then the spectral radius of the matrix $G_{\alpha,\beta}(x'_*)|_{\mathcal{N}_{x'_*}}$ is strictly less than 1 if

$$\alpha \in \left(0, \frac{2(\beta+1)}{L(2\beta+1)}\right) \quad \text{for all } \beta \in [0,1).$$

Moreover, the spectral radius is minimized over

$$(\alpha,\beta) \in \left(0, \frac{2(\beta+1)}{L(2\beta+1)}\right) \times [0,1)$$

by the optimal choice

$$\alpha_{\text{opt}} = \frac{4}{3L+\mu}, \qquad \beta_{\text{opt}} = \frac{\sqrt{3L+\mu} - 2\sqrt{\mu}}{\sqrt{3L+\mu} + 2\sqrt{\mu}},$$

and the minimal value is given by

$$\rho_{\text{opt}}\big(G_{\alpha,\beta}(x'_*)\big) = \frac{\sqrt{3L+\mu} - 2\sqrt{\mu}}{\sqrt{3L+\mu}}.$$

Proof. It is clear that the spectral radius of $G_{\alpha,\beta}(x'_*)|_{\mathcal{N}_{x'_*}}$ is determined by the β and the upper-left 2×2 block matrix

$$\widetilde{G}_{\alpha,\beta}(x'_*) := \begin{pmatrix} (1+\beta)\mathcal{G}_\alpha(x'_*) & -\beta\mathcal{G}_\alpha(x'_*) \\ I & \mathbf{0} \end{pmatrix}.$$

Note that each $\lambda \in \text{Spec}\big(\widetilde{G}_{\alpha,\beta}(x'_*)\big)$ satisfies the quadratic equation

$$\lambda^2 - (1+\beta)\mu_\alpha\lambda + \mu_\alpha\beta = 0, \tag{1.10}$$

for some $\mu_\alpha \in \text{Spec}\big(\mathcal{G}_\alpha(x'_*)\big)$.

Below, we can discuss what choices of parameter α ensure that for all $\beta \in [0,1)$, all eigenvalues λ lie inside the unit circle. This depends solely on $|\lambda|$. Using the quadratic formula, we obtain the following expression

$$|\lambda| = \begin{cases} \dfrac{1}{2}(1+\beta)|\mu_\alpha| + \dfrac{1}{2}\sqrt{D} & \text{if} \quad D \geq 0, \\ \sqrt{\beta|\mu_\alpha|} & \text{otherwise,} \end{cases}$$

where $D = (1+\beta)^2\mu_\alpha^2 - 4\beta\mu_\alpha$. To ensure that all eigenvalues λ of the linearized iteration operator satisfy $|\lambda| < 1$ for every $\mu_\alpha \in \mathrm{Spec}(\mathcal{G}_\alpha(x'_*)) \subseteq [\,1-\alpha L,\ 1-\alpha\mu\,]$ and for all $\beta \in [0,1)$, the parameter $\alpha > 0$ must satisfy the classical Schur conditions:

$$1 - \mu_\alpha > 0, \quad 1 + (1+2\beta)\mu_\alpha > 0, \quad \text{and} \quad \beta|\mu_\alpha| < 1.$$

These inequalities are guaranteed uniformly when

$$\alpha \in \left(0, \frac{2(\beta+1)}{L(2\beta+1)}\right) \quad \text{for all } \beta \in [0,1).$$

Finally, according to (Lessard et al., 2016, Proposition 1), the optimal parameters

$$\alpha_{\mathrm{opt}} = \frac{4}{3L+\mu}, \qquad \beta_{\mathrm{opt}} = \frac{\sqrt{3L+\mu} - 2\sqrt{\mu}}{\sqrt{3L+\mu} + 2\sqrt{\mu}},$$

minimize the spectral radius of $\widetilde{G}_{\alpha,\beta}(x'_*)$ with respect to α, β, yielding

$$\rho_{\mathrm{opt}}\big(\widetilde{G}_{\alpha,\beta}(x'_*)\big) = \frac{\sqrt{3L+\mu} - 2\sqrt{\mu}}{\sqrt{3L+\mu}}.$$

A direct computation shows that $\beta_{\mathrm{opt}} \le \rho_{\mathrm{opt}}\big(\widetilde{G}_{\alpha,\beta}(x'_*)\big)$ with equality only when $\mu = L$. Since

$$\mathrm{Spec}\big(G_{\alpha,\beta}(x'_*)|_{\mathcal{N}_{x'_*}}\big) = \mathrm{Spec}\big(\widetilde{G}_{\alpha,\beta}(x'_*)\big) \cup \{\beta\},$$

the spectral radius of $\mathrm{Spec}\big(G_{\alpha,\beta}(x'_*)|_{\mathcal{N}_{x'_*}}\big)$ is governed by the larger of the two quantities. Hence,

$$\rho_{\mathrm{opt}}\big(G_{\alpha,\beta}(x'_*)|_{\mathcal{N}_{x'_*}}\big) = \frac{\sqrt{3L+\mu} - 2\sqrt{\mu}}{\sqrt{3L+\mu}}.$$

□

The following is a classical result in functional analysis that relates the local convergence rate of a nonlinear iteration to the spectral radius of its linear part. It is commonly used in the stability analysis of ordinary differential equations and dynamical systems.

Lemma A.3. Suppose that the linear operator T on a Hilbert space X satisfies the condition $\rho(T) = \rho < 1$, and the sequence $\{v^n\}_{n\in\mathbb{N}} \subset X$ satisfies:

$$v^{n+1} = Tv^n + Y(v^n) \quad \textit{and} \quad \lim_{\|v\|_X \to 0} \frac{\|Y(v)\|_X}{\|v\|_X} = 0.$$

Then, for all sufficiently small ε, there exists σ such that for all $\|v^0\|_X \le \sigma$,

$$\|v^n\|_X \le C_\varepsilon \|v^0\|_X (\rho+\varepsilon)^n.$$

Proof. Based on the discrete Gronwall inequality, the result is standard. Since

$$\lim_{n\to\infty} \big\|T^n\big\|^{\frac{1}{n}} = \rho < 1,$$

then for any sufficiently small $\varepsilon > 0$, there exists a constant C_ε depending on ε such that

$$\big\|T^n\big\| \le C_\varepsilon(\rho + \varepsilon/3)^n, \quad \text{for all } n \in \mathbb{N}.$$

The condition $\lim_{\|v\|_X\to 0} \big\|Y(v)\big\|_X / \|v\|_X = 0$ indicates that for any sufficiently small ε, there exists a small enough σ_1 such that for all $\|v\|_X \le \sigma_1$, there holds

$$\big\|Y(v)\big\|_X \le \frac{\varepsilon}{3C_\varepsilon}\big\|v\big\|_X.$$

Let $\sigma \le \frac{\sigma_1}{(1+C_\varepsilon)}$, we use mathematical induction to prove

$$\|v^n\|_X \le \sigma_1, \quad \text{for all } n \ge 0.$$

Obviously, $n = 0$ is true. Then let us assume $\|v^k\|_X \le \sigma_1$ for all $k \le n-1$ ($n \ge 2$). Hence, the following inequality holds for $k = n$

$$\begin{aligned}
\|v^n\|_X &= \|Tv^{n-1} + Y(v^{n-1})\|_X \\
&= \|T^2 v^{n-2} + TY(v^{n-2}) + Y(v^{n-1})\|_X = \left\| T^n v^0 + \sum_{k=0}^{n-1} T^{n-1-k} Y(v^k) \right\|_X \\
&\le \|T^n v^0\|_X + \sum_{k=0}^{n-1} \|T^{n-1-k}\| \|Y(v^k)\|_X \\
&\le C_\varepsilon \|v^0\|_X (\rho + \varepsilon/3)^n + \sum_{k=0}^{n-1} (\rho + \varepsilon/3)^{n-1-k} \frac{\varepsilon}{3} \|v^k\|_X \\
\implies \quad & (\rho + \varepsilon/3)^{-n} \|v^n\|_X \le C_\varepsilon \|v^0\|_X + \sum_{k=0}^{n-1} \frac{\varepsilon}{3\rho + \varepsilon} (\rho + \varepsilon/3)^{-k} \|v^k\|_X.
\end{aligned}$$

Applying the classical discrete Gronwall inequality, we derive

$$(\rho + \varepsilon/3)^{-n} \|v^n\|_X \le C_\varepsilon \|v^0\|_X \left(1 + \frac{\varepsilon}{3\rho + \varepsilon}\right)^n \implies \|v^n\|_X \le C_\varepsilon \|v^0\|_X (\rho + \varepsilon)^n \le \sigma_1.$$

This not only completes the induction but also proves the conclusion. □

With this, we are ready to prove the theorems.

A.2 Proof of Theorem 2.1

Proof of Theorem 2.1. Define

$$e_n := |v_{n+1} - v_n| + |v_{n+1}|.$$

Recall that in NAG the extrapolation point is given by

$$y^n = x^n + \beta v_n, \quad x^{n+1} = y^n - \alpha \nabla f(y^n),$$

so that

$$x^{n+1} - y^n = v_{n+1} - \beta v_n, \qquad x^n - y^n = -\beta v_n.$$

It follows that both displacements are controlled by e_n:

$$\begin{aligned}
\|x^{n+1} - y^n\| &= \|v_{n+1} - \beta v_n\| = \|(1-\beta) v_{n+1} + \beta (v_{n+1} - v_n)\| \le C \|e_n\|, \\
\|x^n - y^n\| &= \|\beta v_n\| = \beta \|v_n - v_{n+1} + v_{n+1}\| \le C \|e_n\|.
\end{aligned}$$

Under the C^2 regularity of f, we expand both $f(x^{n+1})$ and $f(x^n)$ around the intermediate point y^n near x_*. Using the continuity of $\nabla^2 f$ at x_*, we obtain

$$\begin{aligned}
f(x^{n+1}) &= f(y^n) + \langle \nabla f(y^n),\, x^{n+1} - y^n \rangle + \frac{1}{2} \nabla^2 f(x_*)[x^{n+1} - y^n,\, x^{n+1} - y^n] + o(\|e_n\|^2), \\
f(x^n) &= f(y^n) + \langle \nabla f(y^n),\, x^n - y^n \rangle + \frac{1}{2} \nabla^2 f(x_*)[x^n - y^n,\, x^n - y^n] + o(\|e_n\|^2).
\end{aligned}$$

Subtracting these two expansions and using the identities

$$\begin{aligned}
\langle \nabla f(y^n),\, x^{n+1} - x^n \rangle &= -\frac{1}{\alpha} \langle v_{n+1} - \beta v_n,\, v_{n+1} \rangle \\
&= -\frac{1}{\alpha} \|v_{n+1}\|^2 + \frac{\beta}{\alpha} \langle v_n, v_{n+1} \rangle \\
&= -\frac{1-\beta}{\alpha} \|v_{n+1}\|^2 - \frac{\beta}{2\alpha} \|v_{n+1}\|^2 + \frac{\beta}{2\alpha} \|v_n\|^2 - \frac{\beta}{2\alpha} \|v_{n+1} - v_n\|^2
\end{aligned}$$

and

$$\begin{aligned}
&\frac{1}{2}\nabla^2 f(x_*)[x^{n+1}-y^n, x^{n+1}-y^n] - \frac{1}{2}\nabla^2 f(x_*)[x^n-y^n, x^n-y^n] \\
=&\frac{1}{2}\nabla^2 f(x_*)[v_{n+1}, v_{n+1}-2\beta v_n] \\
=&\frac{1-\beta}{2}\nabla^2 f(x_*)[v_{n+1}, v_{n+1}] - \frac{\beta}{2}\nabla^2 f(x_*)[v_n, v_n] + \frac{\beta}{2}\nabla^2 f(x_*)[v_{n+1}-v_n, v_{n+1}-v_n] \\
=&\frac{\beta^2}{2(1+\beta)}\nabla^2 f(x_*)[v_{n+1}, v_{n+1}] - \frac{\beta^2}{2(1+\beta)}\nabla^2 f(x_*)[v_n, v_n] + \frac{1+\beta-2\beta^2}{2(1+\beta)}\nabla^2 f(x_*)[v_{n+1}, v_{n+1}] \\
&\quad - \frac{\beta}{4(1+\beta)}\nabla^2 f(x_*)[v_{n+1}+v_n, v_{n+1}+v_n] + \frac{2\beta^2+\beta}{4(1+\beta)}\nabla^2 f(x_*)[v_{n+1}-v_n, v_{n+1}-v_n],
\end{aligned}$$

we arrive at

$$\begin{aligned}
\mathcal{L}(x^{n+1}, v_{n+1}) = \mathcal{L}(x^n, v_n) &- \frac{1-\beta}{\alpha}\|v_{n+1}\|^2 - \frac{\beta}{2\alpha}\|v_{n+1}-v_n\|^2 \\
&- \frac{\beta}{4(1+\beta)}\nabla^2 f(x_*)[v_{n+1}+v_n, v_{n+1}+v_n] + \frac{1+\beta-2\beta^2}{2(1+\beta)}\nabla^2 f(x_*)[v_{n+1}, v_{n+1}] \\
&+ \frac{2\beta^2+\beta}{4(1+\beta)}\nabla^2 f(x_*)[v_{n+1}-v_n, v_{n+1}-v_n] + o(\|e_n\|^2).
\end{aligned}$$

Since $\nabla^2 f(x_*) \preceq LI$, we have $\nabla^2 f(x_*)[v,v] \le L\|v\|^2$ for all $v \in \mathbb{R}^d$. Consequently, we further get

$$\begin{aligned}
\mathcal{L}(x^{n+1}, v_{n+1}) \le \mathcal{L}(x^n, v_n) &- \left(\frac{1-\beta}{\alpha} - \frac{1+\beta-2\beta^2}{2(1+\beta)}L\right)\|v_{n+1}\|^2 \\
&- \left(\frac{\beta}{2\alpha} - \frac{2\beta^2+\beta}{4(1+\beta)}L\right)\|v_{n+1}-v_n\|^2 + o(\|e_n\|^2).
\end{aligned}$$

Assume that $\{x^n\}_{n\in\mathbb{N}} \subset \mathcal{B}_{\sigma_1}(x_*)$ for some sufficiently small $\sigma_1 > 0$. By the triangle inequality, we have

$$\|v_n\| = \|x^n - x^{n-1}\| \le \|x^n - x_*\| + \|x^{n-1} - x_*\| \le 2\sigma_1.$$

For sufficiently small $\sigma_1 > 0$, the C^2 regularity of f ensures that all higher-order remainder terms are dominated by $\varepsilon\|e_n\|^2$ for a sufficiently small $\varepsilon > 0$. Consequently, the Lyapunov function satisfies

$$\begin{aligned}
\mathcal{L}(x^{n+1}, v_{n+1}) \le \mathcal{L}(x^n, v_n) &- \left(\frac{1-\beta}{\alpha} - \frac{1+\beta-2\beta^2}{2(1+\beta)}(L+\varepsilon)\right)\|v_{n+1}\|^2 \\
&- \left(\frac{\beta}{2\alpha} - \frac{2\beta^2+\beta}{4(1+\beta)}(L+\varepsilon)\right)\|v_{n+1}-v_n\|^2.
\end{aligned}$$

The coefficients of $\|v_{n+1}\|^2$ and $\|v_{n+1}-v_n\|^2$ are positive provided that

$$\frac{1-\beta}{\alpha} > \frac{1+\beta-2\beta^2}{2(1+\beta)}(L+\varepsilon), \quad \frac{\beta}{2\alpha} > \frac{2\beta^2+\beta}{4(1+\beta)}(L+\varepsilon),$$

which are equivalent to

$$\alpha \in \left(0, \frac{2(\beta+1)}{(L+\varepsilon)(2\beta+1)}\right), \quad \text{for all } \beta \in [0,1).$$

Under this condition, there exists a constant $C_{\varepsilon,\alpha,\beta} > 0$ such that

$$\mathcal{L}(x^{n+1}, v_{n+1}) \le \mathcal{L}(x^n, v_n) - C_{\varepsilon,\alpha,\beta}\big(\|v_{n+1}\|^2 + \|v_{n+1}-v_n\|^2\big).$$

Moreover, it can be verified that

$$\frac{2(\beta+1)}{(L+\varepsilon)(2\beta+1)} \le \frac{\beta+1}{L\beta},$$

which shows that the chosen α is compatible with the condition in Lemma A.1 when σ_1 is sufficiently small. Next, we invoke the higher-order expansion of NAG:

$$v_{n+1} = \beta v_n - \alpha \nabla f(x^n) - \alpha\beta \nabla^2 f(x^n)[v_n] + o(\|v_n\|),$$

which implies that, near x_*,

$$\mathcal{L}(x^n, v_n) \approx \|\nabla f(x^n)\|^2 + \|v_n\|^2 \le C\big(\|v_{n+1}\|^2 + \|v_{n+1} - v_n\|^2\big).$$

Consequently, there exists a constant $C'_{\varepsilon,\alpha,\beta} > 0$, depending on ε, α, and β, such that

$$\mathcal{L}(x^{n+1}, v_{n+1}) \le \mathcal{L}(x^n, v_n) - C'_{\varepsilon,\alpha,\beta}\, \mathcal{L}(x^n, v_n).$$

Since $\mathcal{L}(x, v) \ge 0$ and vanishes only on $\mathcal{S} \times \{0\}$, the above inequality implies $C'_{\varepsilon,\alpha,\beta} \in (0, 1)$. Defining the contraction factor as

$$\rho_{\varepsilon,\alpha,\beta} := 1 - C'_{\varepsilon,\alpha,\beta} \in (0, 1),$$

we obtain

$$0 \le \mathcal{L}(x^{n+1}, v_{n+1}) \le \rho_{\varepsilon,\alpha,\beta}\, \mathcal{L}(x^n, v_n).$$

It remains to verify that the entire sequence $\{x^n\}_{n\in\mathbb{N}}$ remains in $\mathcal{B}_{\sigma_1}(x_*)$, provided x^0, x^1 are sufficiently close to x_*. To this end, choose $\sigma \in (0, \sigma_1)$ such that $x^0, x^1 \in \mathcal{B}_\sigma(x_*)$. Then the claim holds for $n = 0, 1$.

We now proceed by induction. Assume that $x^k \in \mathcal{B}_{\sigma_1}(x_*)$ for all $k \le n$ with $n \ge 1$. Then all local estimates derived above—including the Lyapunov decay estimate—are valid up to step n, yielding

$$\mathcal{L}(x^{n+1}, v_{n+1}) \le \rho\, \mathcal{L}(x^n, v_n) \le \cdots \le \rho^n \mathcal{L}(x^1, v_1),$$

for some $\rho \in (0, 1)$. Using this geometric decay and the local Lipschitz continuity of the gradient, we bound the deviation of x^{n+1} from x_* as follows:

$$\begin{aligned}
\|x^{n+1} - x_*\| &\le \|x^0 - x_*\| + \sum_{j=0}^{n} \|x^{j+1} - x^j\| \\
&= \|x^0 - x_*\| + \sum_{j=0}^{n} \|v_{j+1}\| \\
&\le \|x^0 - x_*\| + C \sum_{j=0}^{n} \sqrt{\mathcal{L}(x^{j+1}, v_{j+1})} \\
&\le \sigma + C\sqrt{\mathcal{L}(x^1, v_1)} \sum_{j=0}^{n} \rho^{j/2} \\
&\le \sigma + C\sqrt{\mathcal{L}(x^1, v_1)} \cdot \frac{1}{1 - \sqrt{\rho}} \\
&\le \sigma + C\big(\|\nabla f(x^1)\| + \|x^1 - x_*\| + \|x^0 - x_*\|\big) \\
&\le \sigma + C\sigma \\
&\le C\sigma,
\end{aligned}$$

where $C > 1$ is a uniform constant independent of n.

Choosing $\sigma < \sigma_1/C$, we obtain $x^{n+1} \in \mathcal{B}_{\sigma_1}(x_*)$. This completes the induction step and confirms that the entire sequence $\{x^n\}_{n\in\mathbb{N}} \subseteq \mathcal{B}_{\sigma_1}(x_*)$, thereby validating all local estimates used above. □

A.3 Proof of Theorem 2.2

Proof of Theorem 2.2. By Lemma A.1 and Theorem 2.1, for every sufficiently small $\varepsilon > 0$, there exists $\sigma \in (0, \sigma_1)$ such that, for any $x^0, x^1 \in \mathcal{B}_\sigma(x_*)$, the NAG iterates satisfy

$\{x^n\}_{n\in\mathbb{N}} \subset \mathcal{B}_{\sigma_1}(x_*)$ and hence the difference sequence $\{v_n\}_{n\in\mathbb{N}}$ converge linearly to 0, provided (α, β) satisfies

$$\alpha \in \left(0, \frac{2(1+\beta)}{(2\beta+1)(L+\varepsilon)}\right), \quad \beta \in (0,1).$$

Specifically, there exists a constant $C_\varepsilon > 0$ such that

$$\|x^{n+1} - x^n\| \le C\sqrt{\mathcal{L}(x^1, v_1)}\rho^{n/2},$$

where $\rho \in (0,1)$ depends on ε, α, and β.

This geometric decay of successive differences implies that the iterates $\{x^n\}_{n\in\mathbb{N}}$ themselves converge linearly to the local minimizer x'_*. Indeed, for any integers $m > n \ge 1$, we have

$$\begin{aligned}\|x^m - x^n\| &\le \sum_{k=n+1}^{m} \|x^k - x^{k-1}\| \le \sum_{k=n+1}^{\infty} \|x^k - x^{k-1}\| \\ &\le C\sqrt{\mathcal{L}(x^1, v_1)} \sum_{k=n+1}^{\infty} \rho^{k/2} = C\sqrt{\mathcal{L}(x^1, v_1)} \cdot \frac{\sqrt{\rho}^{n+1}}{1-\sqrt{\rho}}.\end{aligned}$$

Hence $\{x^n\}_{n\in\mathbb{N}}$ is a Cauchy sequence in $\mathbb{R}^d$, and therefore converges to some limit point $x'_* \in \mathcal{B}_{\sigma_1}(x_*)$. Moreover, passing to the limit in the above estimate yields the linear convergence rate

$$\|x^n - x'_*\| \le C\sqrt{\mathcal{L}(x^1, v_1)}\rho^{n/2} \le C\big(\|x^1 - x'_*\| + \|x^0 - x'_*\|\big)\rho^{n/2}.$$

The optimal local linear convergence rate is established below. A local linearization of the NAG iteration at x'_* yields

$$\begin{aligned}x^{n+1} - x'_* &= x^n - x'_* + \beta(x^n - x^{n-1}) - \alpha\nabla^2 f(x'_*)[x^n - x'_* + \beta(x^n - x^{n-1})] \\ &\quad + o(\|x^n - x'_*\| + \|x^n - x^{n-1}\|) \\ &= (I - P_{x'_*})(x^n - x'_* + \beta(x^n - x^{n-1})) + (1+\beta)\mathcal{G}_\alpha(x'_*)P_{x'_*}(x^n - x'_*) \\ &\quad - \beta\mathcal{G}_\alpha(x'_*)P_{x'_*}(x^{n-1} - x'_*) + o(\|x^n - x'_*\| + \|x^n - x^{n-1}\|).\end{aligned}$$

From this expansion, the following decoupled asymptotic relations follow:

$$\begin{aligned}(I - P_{x'_*})(x^{n+1} - x^n) &= \beta(I - P_{x'_*})(x^n - x^{n-1}) + o(\|x^n - x'_*\| + \|x^n - x^{n-1}\|), \\ P_{x'_*}(x^{n+1} - x'_*) &= (1+\beta)\mathcal{G}_\alpha(x'_*)P_{x'_*}(x^n - x'_*) - \beta\mathcal{G}_\alpha(x'_*)P_{x'_*}(x^{n-1} - x'_*) \\ &\quad + o(\|x^n - x'_*\| + \|x^n - x^{n-1}\|).\end{aligned}$$

Telescopic summation of the first relation gives

$$(I - P_{x'_*})(x^n - x'_*) - \beta(I - P_{x'_*})(x^{n-1} - x'_*) = \sum_{k=n}^{\infty}(I - P_{x'_*})(x^k - x^{k+1} - \beta(x^{k-1} - x^k)) \tag{1.11}$$

with $\big\|(I - P_{x'_*})(x^k - x_*^{k+1}) - \beta(I - P_{x'_*})(x^{k-1} - x_*^k)\big\| \to 0$ as $(\|x^k - x'_*\| + \|x^k - x^{k-1}\|) \to 0$. In particular, for every $\sigma > 0$ there exists $\delta_\sigma \to 0^+$ as $\sigma \to 0$ such that

$$\big\|(I - P_{x'_*})(x^k - x^{k+1} - \beta(x^{k-1} - x^k))\big\| \le \delta_\sigma(\|x^k - x'_*\| + \|x^k - x^{k-1}\|)$$

whenever

$$\|x^k - x'_*\| + \|x^k - x^{k-1}\| \le \sigma.$$

Hence, from the telescoping identity equation 1.11 and for n large enough so that

$$\|x^k - x'_*\| + \|x^k - x^{k-1}\| \le \sigma, \quad \text{for all } k \ge n,$$

it follows that

$$\big\|(I - P_{x'_*})(x^n - x'_*) - \beta(I - P_{x'_*})(x^{n-1} - x'_*)\big\| \le \delta_\sigma \sum_{k=n}^{\infty}\big(\|x^k - x'_*\| + \|x^k - x^{k-1}\|\big). \tag{1.12}$$

Given the linear convergence of $\{x^n\}_{n\in\mathbb{N}}$ and $\{x^n - x^{n-1}\}_{n\in\mathbb{N}}$, there exists $\rho \in (0,1)$ such that

$$\|x^k - x'_*\| + \|x^k - x^{k-1}\| \le C\rho^{k-n}\left(\|x^n - x'_*\| + \|x^n - x^{n-1}\|\right) \quad \text{for all } k \ge n,$$

Consequently,

$$\begin{aligned}\left\|(I - P_{x'_*})(x^n - x'_*) - \beta(I - P_{x'_*})(x^{n-1} - x'_*)\right\| &\overset{equation\ 1.12}{\le} \delta_\sigma \sum_{k=n}^{\infty}\left(\|x^k - x'_*\| + \|x^k - x^{k-1}\|\right)\\ &\le C\delta_\sigma\left(\|x^n - x'_*\| + \|x^n - x^{n-1}\|\right),\end{aligned}$$

where $\delta_\sigma \to 0^+$ as $\sigma \to 0^+$. It follows that

$$\begin{aligned}&(I - P_{x'_*})(x^n - x'_*) - \beta(I - P_{x'_*})(x^{n-1} - x'_*) = o(\|x^n - x'_*\| + \|x^n - x^{n-1}\|),\\ \Longrightarrow\ &(I - P_{x'_*})(x^n - x'_*) = \frac{\beta}{1-\beta}(I - P_{x'_*})(x^{n-1} - x^n) + o(\|x^n - x'_*\| + \|x^n - x^{n-1}\|).\end{aligned}$$

Using the identities

$$\begin{aligned}x^n - x'_* &= P_{x'_*}(x^n - x'_*) + (I - P_{x'_*})(x^n - x'_*),\\ x^n - x^{n-1} &= P_{x'_*}(x^n - x'_*) - P_{x'_*}(x^{n-1} - x'_*) + (I - P_{x'_*})(x^n - x^{n-1}),\end{aligned}$$

and applying the triangle inequality, we obtain

$$\begin{aligned}&o\left(\|x^n - x'_*\| + \|x^n - x^{n-1}\|\right)\\ =&o\left(\left\|P_{x'_*}(x^n - x'_*)\right\| + \left\|P_{x'_*}(x^{n-1} - x'_*)\right\| + \left\|(I - P_{x'_*})(x^n - x^{n-1})\right\|\right).\end{aligned}$$

Define the augmented state vector

$$X^n := \begin{pmatrix} P_{x'_*}(x^n - x'_*)\\ P_{x'_*}(x^{n-1} - x'_*)\\ (I - P_{x'_*})(x^n - x^{n-1})\end{pmatrix}.$$

We can then rewrite the iteration using the matrix $G_{\alpha,\beta}(x'_*)$ from equation 1.9 as

$$X^{n+1} = G_{\alpha,\beta}(x'_*)X^n + o\big(\|X^n\|\big),$$

where

$$\|X^n\| = \sqrt{\|P_{x'_*}(x^n - x'_*)\|^2 + \|P_{x'_*}(x^{n-1} - x'_*)\|^2 + \|(I - P_{x'_*})(x^n - x^{n-1})\|^2}.$$

By Lemma A.2 and Lemma A.3, we immediately obtain the following local linear convergence rate:

$$\left\|X^{n+1}\right\| \le C_\varepsilon \|X^n\| \left(\rho\big(G_{\alpha,\beta}(x'_*)|_{\mathcal{N}_{x'_*}}\big) + \varepsilon\right)^n, \quad \forall\, n \ge 1,$$

Together with the inequality

$$\begin{aligned}\|x^n - x^{n-1}\| &\le \left\|P_{x'_*}(x^n - x'_*)\right\| + \left\|P_{x'_*}(x^{n-1} - x'_*)\right\| + \left\|(I - P_{x'_*})(x^n - x^{n-1})\right\|\\ &= \|X^n\|,\end{aligned}$$

and the geometric decay estimate of $\{X^n\}_{n\in\mathbb{N}}$, we deduce that for all parameters (α,β) satisfying

$$\alpha \in \left(0, \frac{2(\beta+1)}{(L+\varepsilon)(2\beta+1)}\right),\ \beta \in [0,1),$$

the following locally linear convergence rate holds

$$\|x^n - x^{n-1}\| \le C_\varepsilon\big(\|x^1 - x'_*\| + \|x^0 - x'_*\|\big)\left(\rho\big(G_{\alpha,\beta}(x'_*)|_{\mathcal{N}_{x'_*}}\big) + \varepsilon\right)^n, \quad \forall\, n \ge 1.$$

This suggests that

$$\|x^n - x'_*\| = \sum_{k=n+1}^{\infty} \|x^k - x^{k-1}\| \\ \leq C_\varepsilon \big(\|x^1 - x'_*\| + \|x^0 - x'_*\|\big) \Big(\rho\big(G_{\alpha,\beta}(x'_*)|_{\mathcal{N}_{x'_*}}\big) + \varepsilon\Big)^n, \quad \forall\, n \geq 1.$$

In particular, the optimal contraction factor is attained when

$$\alpha_{\text{opt}} = \frac{4}{3L+\mu}, \qquad \beta_{\text{opt}} = \frac{\sqrt{3L+\mu} - 2\sqrt{\mu}}{\sqrt{3L+\mu} + 2\sqrt{\mu}},$$

and the minimal value is given by

$$\rho_{\text{opt}}\big(G_{\alpha,\beta}(x'_*)|_{\mathcal{N}_{x'_*}}\big) = \frac{\sqrt{3L+\mu} - 2\sqrt{\mu}}{\sqrt{3L+\mu}}.$$

□

A.4 Proof of Theorem 2.3

Proof of Theorem 2.3. The proof follows the same strategy as that of Theorem 2.2: we first establish local exponential convergence of the trajectories via a Lyapunov function argument. Specifically, for any sufficiently small $\varepsilon > 0$, we employ the Lyapunov function constructed in (Polyak & Shcherbakov, 2017, Theorem 4), namely

$$\widetilde{\mathcal{L}}(x, v) := f(x) - f(x_*) + \frac{\gamma}{\gamma^2 + 2(L+\varepsilon)} \langle \nabla f(x), v\rangle + \frac{L+\varepsilon}{\gamma^2 + 2(L+\varepsilon)} \|v\|^2.$$

Since $f \in C^2$, for any given $\varepsilon > 0$, there exists a sufficiently small open neighborhood U_1 of x_* such that

$$\nabla^2 f(x)[v, v] \leq (L+\varepsilon)\|v\|^2, \quad \text{for all } x \in U_1 \text{ and all } v \in \mathbb{R}^d.$$

Assuming that the trajectory remains in U_1 for all $t \geq 0$, the argument in (Polyak & Shcherbakov, 2017, Theorem 4) directly yields the exponential decay estimate

$$\big\|\nabla f(x(t))\big\|^2 + \|\dot{x}(t)\|^2 \approx \widetilde{\mathcal{L}}(x(t), \dot{x}(t)) \leq e^{-\rho t}\, \widetilde{\mathcal{L}}(x(0), \dot{x}(0)), \quad \forall\, t \geq 0,$$

for some constant $\rho > 0$ depending on ε, γ. Finally, by choosing the initial state $(x(0), \dot{x}(0))$ in a sufficiently small open neighborhood $\mathcal{U} \subset \mathbb{R}^d \times \mathbb{R}^d$ of $(x_*, 0)$, one can guarantee that $x(t) \in U_1$ for all $t \geq 0$. This step follows verbatim from the proof of Theorem 2.1 and is therefore omitted here. This implies that for all $t' \geq t \geq 0$,

$$\|x(t') - x(t)\| = \left\| \int_t^{t'} \dot{x}(s)\, ds \right\| \leq \int_t^{t'} \|\dot{x}(s)\|\, ds \\ \leq \int_t^{t'} Ce^{-\rho s} \sqrt{\widetilde{\mathcal{L}}(x(0), \dot{x}(0))}\, ds \leq Ce^{-\rho t} \sqrt{\widetilde{\mathcal{L}}(x(0), \dot{x}(0))}, \tag{1.13}$$

where we used the exponential decay of $\|\dot{x}(t)\|$. By the Cauchy criterion, $\{x(t)\}_{t\geq 0}$ converges to some limit point $x'_* \in U_1$ as $t \to \infty$, and the convergence is locally exponential, i.e.,

$$\|x(t) - x'_*\| \leq Ce^{-\rho t} \sqrt{\widetilde{\mathcal{L}}(x(0), \dot{x}(0))} \leq Ce^{-\rho t} \big(\|x(0) - x'_*\| + \|\dot{x}(0)\|\big)$$

Next, we linearize the dynamics equation 2.6 around the limit point x'_*. Since $f \in C^2$, we have

$$\ddot{x}(t) + \gamma \dot{x}(t) = -\nabla^2 f(x'_*)(x(t) - x'_*) + o(\|x(t) - x'_*\|), \quad \text{as } t \to \infty.$$

Decomposing the above equation yields the asymptotic relations:

$$(I - P_{x'_*})\ddot{x}(t) = -\gamma (I - P_{x'_*})\dot{x}(t) + o(\|x(t) - x'_*\|), \\ P_{x'_*}\ddot{x}(t) + \gamma P_{x'_*}\dot{x}(t) = -\nabla^2 f(x'_*) P_{x'_*}(x(t) - x'_*) + o(\|x(t) - x'_*\|).$$

Denote the remainder term by $R(t) := o(\|x(t) - x'_*\|)$ as $t \to \infty$. Integrating the first equation of equation 2.6 from t to t' yields

$$(I - P_{x'_*})(\dot{x}(t') - \dot{x}(t)) = -\gamma(I - P_{x'_*})(x(t') - x(t)) + \int_t^{t'} R(s)\, ds.$$

Taking the limit as $t' \to \infty$, and using the facts that

$$\lim_{t'\to\infty} x(t') = x'_*, \qquad \lim_{t'\to\infty} \dot{x}(t') = 0,$$

we obtain

$$(I - P_{x'_*})\dot{x}(t) = -\gamma(I - P_{x'_*})(x(t) - x'_*) - \int_t^{\infty} R(s)\, ds.$$

Moreover, since $\|R(s)\| \le \sigma_s \|x(s) - x'_*\|$ with $\sigma_s \to 0$ as $s \to \infty$, and

$$\|x(s) - x'_*\| \le Ce^{-\rho(s-t)}\big(\|x(t) - x'_*\| + \|\dot{x}(t)\|\big) \quad \text{for all } s \ge t \text{ with } t \text{ sufficiently large,}$$

it follows that

$$\begin{aligned}
\left\| \int_t^{\infty} R(s)\, ds \right\| &\le \int_t^{\infty} \|R(s)\|\, ds \le \int_t^{\infty} \sigma_s \|x(s) - x'_*\|\, ds \\
&\le C\big(\|x(t) - x'_*\| + \|\dot{x}(t)\|\big) \left(\sup_{s\ge t} \sigma_s\right) \int_t^{\infty} e^{-\rho(s-t)}\, ds \\
&= \frac{C}{\rho}\big(\|x(t) - x'_*\| + \|\dot{x}(t)\|\big) \sup_{s\ge t} \sigma_s \\
&= o\big(\|x(t) - x'_*\| + \|\dot{x}(t)\|\big) \quad \text{as } t \to \infty.
\end{aligned}$$

Hence,

$$(I - P_{x'_*})(x(t) - x'_*) = -\frac{1}{\gamma}(I - P_{x'_*})\dot{x}(t) + o\big(\|x(t) - x'_*\| + \|\dot{x}(t)\|\big). \tag{1.14}$$

Now decompose the state using the orthogonal projection $P_{x'_*}$. We have

$$\begin{aligned}
x(t) - x'_* &= P_{x'_*}(x(t) - x'_*) + (I - P_{x'_*})(x(t) - x'_*), \\
\dot{x}(t) &= P_{x'_*}\dot{x}(t) + (I - P_{x'_*})\dot{x}(t).
\end{aligned}$$

Combining these with equation 1.14 and applying the triangle inequality, the asymptotic relation

$$o\big(\|x(t) - x'_*\| + \|\dot{x}(t)\|\big) = o\left(\big\|P_{x'_*}(x(t) - x'_*)\big\| + \big\|P_{x'_*}\dot{x}(t)\big\| + \big\|(I - P_{x'_*})\dot{x}(t)\big\|\right)$$

holds as $t \to \infty$.

Define the augmented state vector

$$X(t) := \begin{pmatrix} P_{x'_*}(x(t) - x'_*) \\ P_{x'_*}\dot{x}(t) \\ (I - P_{x'_*})\dot{x}(t) \end{pmatrix}.$$

Then, by decomposing the accelerated flow equation 2.6 around the equilibrium $(x_*, 0)$, the dynamics of $X(t)$ are governed asymptotically by the linear system

$$\frac{\mathrm{d}}{\mathrm{d}t} X(t) = AX(t) + o(\|X(t)\|)$$

with

$$A := \begin{pmatrix} \mathbf{0} & I & \mathbf{0} \\ -\nabla^2 f(x'_*) & -\gamma I & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -\gamma I \end{pmatrix}$$

Observe that the spectrum of the matrix A restricted to the invariant subspace

$$\mathcal{N}_{x'_*} = N_{x'_*}\mathcal{S}_U \times N_{x'_*}\mathcal{S}_U \times T_{x'_*}\mathcal{S}_U \subset \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d$$

is given by

$$\mathrm{Spec}(A|_{\mathcal{N}_{x'_*}}) = \left\{ \frac{-\gamma \pm \sqrt{\gamma^2 - 4\gamma\lambda_i}}{2} \ : \ \lambda_i \in [\mu, L] \right\} \cup \{-\gamma\}.$$

Consequently, the decay rate is determined by the spectral bound

$$-\rho_\gamma = -\frac{\gamma - \sqrt{\max\{0,\, \gamma^2 - 4\mu\}}}{2}.$$

Standard results (see, e.g., (Polyak, 1964, Lemma 7) or (Kassing & Weissmann, 2024, Lemma 4.5)) imply that for any $\varepsilon > 0$, there exists a constant $C_\varepsilon > 0$ such that

$$\|X(t)\| \leq C_\varepsilon e^{-(\rho_\gamma - \varepsilon)t} \|X(0)\|, \quad \forall\, t \geq 0.$$

Combining this exponential decay estimate with the inequality

$$\|\dot{x}(t)\| \leq \left\|P_{x'_*} \dot{x}(t)\right\| + \left\|(I - P_{x'_*})\dot{x}(t)\right\| \leq \|X(t)\|,$$

and repeating the argument used in equation 1.13, we obtain the following quantitative exponential convergence rate:

$$\|x(t) - x'_*\| \leq C_\varepsilon e^{-(\rho_\gamma - \varepsilon)t} \big(\|x(0) - x'_*\| + \|\dot{x}(0)\|\big), \quad \forall\, t \geq 0.$$

In particular, when the damping parameter is chosen as $\gamma = 2\sqrt{\mu}$, the optimal exponential decay rate is achieved:

$$\rho_{\mathrm{opt}} = \sqrt{\mu}.$$

□