Adaptive Acceleration Without Strong Convexity Priors Or Restarts

Joao V. Cavalcanti[1], Laurent Lessard[2] & Ashia C. Wilson[1]
[1] MIT, [2] Northeastern University
{caval,ashia}@mit.edu, l.lessard@northeastern.edu

Abstract

Accelerated optimization methods are typically parametrized by the Lipschitz smoothness and strong convexity parameters of a given problem, which are rarely known in practice. While the Lipschitz smoothness parameter L can be easily estimated by backtracking, directly estimating (i.e., without restarts) the strong convexity parameter m remains an open problem. This paper addresses this challenge by proposing NAG-free, an optimization method that achieves acceleration while estimating m online *without priors or restarts*. Our method couples Nesterov's accelerated gradient (NAG) method with an inexpensive estimator that does not require additional computation, only storage of one more iterate and gradient. We prove that in the canonical setting of the open problem, NAG-free converges globally at least as fast as gradient descent and locally at an accelerated rate, without resorting to m priors or restarts. We also introduce a complementary estimator for L, symmetric to the m estimator. We provide extensive empirical evidence that this NAG-free heuristic with L and m estimators often achieves acceleration in practice, outperforming restart schemes and traditional accelerated methods parametrized by offline estimates of L and m. Our experiments also suggest that NAG-free might be effective even for nonconvex problems when combined with restarts.

1 Introduction

Accelerated methods are notable for achieving optimal convergence rates among first-order optimization algorithms on key problem classes [Nesterov, 2018]. However, to attain these optimal rates in practice, they require knowledge of problem-specific parameters. Consider the class of Lipschitz-smooth, strongly convex functions, characterized by the smoothness parameter L and the strong convexity parameter m. To apply accelerated methods effectively in this setting, both L and m must be known; yet, as noted by Boyd and Vandenberghe [2004, p.463], these parameters "are known only in rare cases." While L can often be estimated via backtracking [Tseng, 2008, Beck and Teboulle, 2009], estimating m is considerably more difficult. As Su et al. [2016, p.21] put it: "while it is relatively easy to bound the Lipschitz constant L by the use of backtracking, estimating the strong convexity parameter m, if not impossible, is very challenging." Similarly, O'Donoghue and Candès [2015, p.3] emphasize that "estimating the strong convexity parameter is much more challenging." In light of this, restart schemes have emerged as the dominant approach to handling unknown m [d'Aspremont et al., 2021, Sec. 6]. These methods restart accelerated algorithms (e.g., Nesterov's method) based on heuristic or adaptive criteria. Some approaches sweep over a grid of candidate m values [Roulet and d'Aspremont, 2017], while others yield empirical success without formal convergence guarantees, relying instead on heuristic arguments grounded in continuous-time analysis [O'Donoghue and Candès, 2015, Su et al., 2016]. Despite these advances, it remains an:

Preprint. Under review.

> *"open question whether strong convexity parameters can be efficiently estimated while maintaining reasonable worst-case guarantees and without resorting to restart schemes."* [d'Aspremont et al., 2021, p. 96].

Contributions This paper directly addresses that open question, with two main contributions:

1. **An efficient method for estimating the strong convexity parameter without restarts.** We introduce **NAG-free**, an optimization algorithm that estimates the strong convexity parameter m online without priors and without relying on restart schemes. NAG-free couples Nesterov's accelerated gradient (NAG) method with a lightweight estimator that requires only the storage of one additional iterate and gradient, both already computed in standard NAG. We prove that NAG-free converges globally at least as fast as gradient descent, and achieves local acceleration.
2. **An efficient parameter-free heuristic with good empirical performance:** Although our focus in this paper is the open question above, which only concerns estimating m, we also consider the problem of simultaneously estimating both L and m. Namely, we equip NAG-free with an L estimator symmetric to the m estimator that requires no further computations, resulting in a parameter-free method. We demonstrate that this NAG-free heuristic often achieves acceleration in practice, outperforming restart schemes and traditional accelerated methods that rely on problem-specific parameter bounds. Moreover, we present promising results on the combination of NAG-free with a restart scheme on nonconvex problems.

2 Problem statement

Consider the task of finding $x^\star(f)$, the unique minimum of the problem

$$\min_x f(x), \tag{1}$$

where $f \in \mathcal{F}(L, m)$, the set of functions that satisfy the following definition.

Definition 2.1 (Lipschitz-Smooth and Strongly Convex Functions)**.** We define $\mathcal{F}(L, m)$ to be the set of all differentiable convex functions $f : \mathbb{R}^d \to \mathbb{R}$ that satisfy both:

- **Lipschitz-smoothness:** There exists $L > 0$ such that for all $x, y \in \mathbb{R}^d$,

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + (L/2)\|y - x\|^2. \tag{2}$$

- **Strong convexity:** There exists $m > 0$ such that for all $x, y \in \mathbb{R}^d$,

$$f(x) + \langle \nabla f(x), y - x \rangle + (m/2)\|y - x\|^2 \leq f(y). \tag{3}$$

That is, $\mathcal{F}(L, m)$ contains functions that are both L-smooth and m-strongly convex.[1]

Having defined $\mathcal{F}(L, m)$, we now precisely state the problem addressed in this paper:

> **Problem statement.** Given $f \in \mathcal{F}(L, m)$, we address the open question *as stated* in [d'Aspremont et al., 2021, p. 96]: whether m can be efficiently estimated while maintaining reasonable worst-case guarantees and without resorting to restart schemes. Accordingly, to focus on the open question of estimating m, we assume that an upper bound on L is known, which is common in practice. Moreover, since the question presumes strong convexity, we assume that inequality (3) holds for some $m > 0$.

[1] More generally, f is L-smooth if $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$ for all x, y. Under convexity, this is equivalent to (2); see Nesterov [2018, Thm. 2.1.5].

3 The NAG-free algorithm

For any $f \in \mathcal{F}(L, m)$ and for all $x \neq y$, the following inequality holds:

$$m \leq c(x, y) := \frac{\|\nabla f(x) - \nabla f(y)\|}{\|x - y\|} \leq L. \tag{4}$$

This inequality follows from standard results on smooth and strongly convex functions [Nesterov, 2018, Thm. 2.1.5, 2.1.10]. The quantity $c(x, y)$ captures a local notion of curvature between two points and lies in the interval $[m, L]$. Given iterates x_{t+1} and x_t produced by Nesterov's accelerated gradient (NAG) method, we propose to estimate m online via the recurrence

$$m_{t+1} := \min(m_t, c_{t+1}), \quad \text{where } c_{t+1} = c(x_{t+1}, x_t). \tag{5}$$

This estimate is then used to parametrize Nesterov's method to produce a new iterate, which in turn is used to compute a new effective curvature and update m_{t+1}. The resulting procedure is computationally lightweight: it reuses gradients already computed by NAG and only requires storing one additional iterate and gradient. To initialize m_0, we use a single evaluation of $c(x_0, y)$, where x_0 is the initial point and y is sampled uniformly from a small neighborhood around x_0. This initialization guarantees that $m_0 \in [m, L]$, making it a safe and inexpensive starting point. The complete algorithm, which we call **NAG-free**, is presented below.

Algorithm 1 NAG-free: an accelerated method that estimates m without restarts.

Input: $T > 0, x_0 = y_0, \bar{L} > L$
Output: x_T, y_T
$y \sim x_0 + U[0, 10^{-6}]^d$
$m_0 \leftarrow \|\nabla f(x_0) - \nabla f(y)\| / \|x_0 - y\|$
for $t = 0, \ldots, T-1$ **do**
  $y_{t+1} \leftarrow x_t - \nabla f(x_t)/\bar{L}$
  $x_{t+1} \leftarrow y_{t+1} + (y_{t+1} - y_t)(\sqrt{\bar{L}} - \sqrt{m_t})/(\sqrt{\bar{L}} + \sqrt{m_t})$
  $c_{t+1} \leftarrow \|\nabla f(x_{t+1}) - \nabla f(x_t)\| / \|x_{t+1} - x_t\|$
  $m_{t+1} \leftarrow \min(m_t, c_{t+1})$
end for

Convergence intuition. NAG-free exhibits two key features that underlie its convergence behavior:

1. **Adaptive interpolation between GD and NAG.** The update rule for x_{t+1} interpolates between gradient descent and NAG. If $m_t \to L$, then the momentum term becomes small NAG-free approximates gradient descent. If $m_t \to m$, then the update becomes equivalent to NAG with correct parameterization. Thus, NAG-free converges globally at least as fast as gradient descent.
2. **Power iteration-like behavior near the optimum.** Near the optimum, the curvature estimate c_t evolves similarly to a power method applied to the Hessian, with some additional dynamics. As a result, the iterate x_t rapidly concentrates in the eigenspace corresponding to the least eigenvalue of the Hessian, m, which translates into c_t quickly approaching m, accelerating NAG-free.

4 Convergence guarantees

In this section, we present convergence results for Algorithm 1 and proof sketches. Full details can be found in the supplementary materials.

In the following, we assume that Algorithm 1 receives an upper bound $\bar{L}$ on L, which is often the case in practice, as the experiments in Section 5 illustrate. While we also believe other approximation strategies, such as backtracking line search, could also be incorporated into our framework, we make this mild assumption for simplicity and clarity of analysis, and to focus on the actual open problem which is to estimate m, not to estimate both m *and* L.

4.1 Global convergence

The main global result is that Algorithm 1 converges at least as fast as gradient descent.

Theorem 4.1. *Let $f \in \mathcal{F}(L, m)$, $\bar{L} \geq L$ and $\bar{\kappa} = \bar{L}/m$. Then, the iterates of Algorithm 1 satisfy*

$$f(y_t) - f(x^\star) \leq 2\bar{L}\left(\frac{\bar{\kappa}-1}{\bar{\kappa}}\right)^t \|x_0 - x^\star\|^2.$$

Proof sketch. The key idea is that NAG-free iterates can be viewed as convex combinations of gradient descent (GD) and NAG iterates, initialized appropriately.

Let $s_t = (x_t, y_t)$, and consider the Lyapunov function

$$W(s_t) = f(y_t) - f(x^\star) + \frac{m}{2}\left\|x_t + \sqrt{\bar{\kappa}}(x_t - y_t) - x^\star\right\|^2,$$

adapted from Bansal and Gupta [2019], which we show to be a Lyapunov function for both GD and NAG. Then, since W is convex, using the fact that NAG-free iterates are convex combinations of GD and NAG iterations, we show that

$$\begin{aligned}(1+\delta^{\mathrm{GD}})W(s_{t+1}) - W(s_t) \leq{}& (1-\alpha_t)\left[(1+\delta^{\mathrm{GD}})W(s_{t+1}^{\mathrm{GD}}) - W(s_t)\right] \\ &+ \alpha_t\left[(1+\delta^{\mathrm{NAG}})W(s_{t+1}^{\mathrm{NAG}}) - W(s_t)\right],\end{aligned}$$

where $s_t^{\mathrm{GD}} = (y_t, y_t)$, $s_t^{\mathrm{NAG}} = s_t$, $\delta^{\mathrm{GD}} = 1/(\bar{\kappa}-1)$, $\delta^{\mathrm{NAG}} = 1/(\sqrt{\bar{\kappa}}-1)$, and $\alpha_t \in [0,1]$ determines the interpolation between GD and NAG. Now, $(1+\delta^{\mathrm{NAG}})W(s_{t+1}^{\mathrm{NAG}}) - W(s_t) \leq 0$, but $(1+\delta^{\mathrm{GD}})W(s_{t+1}^{\mathrm{GD}}) - W(s_t)$ is indefinite. However, we know that $(1+\delta^{\mathrm{GD}})W(s_{t+1}^{\mathrm{GD}}) - W(s_t^{\mathrm{GD}}) \leq 0$, since W is a Lyapunov function centered at $s_t^{\mathrm{GD}} = (y_t, y_t)$. To bridge this mismatch, we augment W with an auxiliary term. Namely, we define

$$V_{t+1}(s_t) = W(s_t) + \frac{1-\alpha_t}{\sqrt{\bar{\kappa}}}U(s_t), \qquad \text{where} \qquad U(s_t) = f(y_t) - f(x^\star) + \frac{\bar{L}}{2}\|y_t - x^\star\|^2.$$

Then, we show $(1+\delta^{\mathrm{GD}})V_{t+1}(s_{t+1}) - V_{t+1}(s_t) \leq 0$ and bound $V_{t+1}(s_t)$ in terms of $V_t(s_t)$. Together, these establish that V_t is a Lyapunov function for NAG-free that decays at a rate of at least $1+\delta^{\mathrm{GD}}$, matching the worst-case global convergence rate of GD. □

4.2 Local acceleration

We now analyze Algorithm 1 around $x^\star$, and prove that it achieves accelerated convergence up to a constant factor. To formalize this result, we introduce the following set of assumptions, along with a succinct description of their significance and how they are used in the acceleration analysis.

Assumption 4.2. The estimates m_{t+1} of Algorithm 1 decay by a factor of at least $\gamma \geq 2$ every time they decrease and are always greater than m/γ: if $m_{t+1} < m_t$, then $m/\gamma \leq m_{t+1} \leq m_t/\gamma$.

On Assumption 4.2. This assumption can be enforced in practice with a simple backtracking procedure: if $c_{t+1} < m_t$, then $m_{t+1} \leftarrow \min\{c_{t+1}, m_t/\gamma\}$; else $m_{t+1} \leftarrow m_t$. Since $c_{t+1} \geq m$, m_{t+1} can only be adjusted *to a new value* if $m_t > m$, in which case m_{t+1} is reduced by a factor of at least $\gamma \geq 2$ such that $m_{t+1} \geq m/\gamma$. This implies that m_t takes finitely many values, which is essentially why we make this assumption. Although $\gamma > 1$ would suffice, we assume $\gamma \geq 2$ to simplify some results and derivations. Importantly, this assumption alone does not imply that m_t converges to m, it only ensures that $m_{t+1} \in [m/\gamma, L]$. The next assumption is a standard condition used in optimization theory and satisfied in many practical settings.

Assumption 4.3. The Hessian of f is locally Lipschitz-smooth at $x^\star$: there are $L_H > 0$ and $\epsilon_H > 0$ such that if $\|x - x^\star\| \leq \epsilon_H$, then $\|\nabla^2 f(x) - \nabla^2 f(x^\star)\| \leq L_H\|x - x^\star\|$.

On Assumption 4.3. The condition of Lipschitz continuity for the Hessian is mild. For example, all twice continuously differentiable functions satisfy this condition, which ensures that the Hessian does not exhibit erratic behavior within a local neighborhood around the optimal point $x^\star$.

To present the remaining assumptions, we introduce the following notation.

Notation. Let (λ_i, v_i) denote the d eigenvalues λ_i and associated eigenvectors v_i of $\nabla^2 f(x^\star)$. If $f \in \mathcal{F}(L, m)$, then v_i can be chosen to form an orthonormal basis for $\mathbb{R}^d$. Hence, $x_0 - x^\star$ uniquely decomposes into $x_0 - x^\star = \sum_{i=1}^d x_{i,0}v_i$. Moreover, $\lambda_i \in [m, L]$. In the following, without loss of generality we assume λ_i ordered by their indices, as in $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_d$. Thus, $x_{1,0}$ denotes the coordinate of $x_0 - x^\star$ in the eigenspace associated with λ_1, the least eigenvalue of $\nabla^2 f(x^\star)$.

Assumption 4.4. There exists some $\delta_\lambda \in (0,1)$ such that $|m_t - \lambda_i| > \delta_\lambda L$ for every $\lambda_i > m$, where $m = \lambda_1 \leq \ldots \leq \lambda_d \leq L$ denote the eigenvalues of $\nabla^2 f(x^\star)$.

On Assumption 4.4. This assumption ensures that the value of m_t remains sufficiently separated from the eigenvalues λ_i of the Hessian, where m is the smallest eigenvalue. Although it simplifies the proof by avoiding complications arising from closely spaced eigenvalues, this assumption is not strictly necessary. Dropping it would require a more involved analysis, using Schur decompositions instead of diagonal matrices, and would introduce additional constant factors. In practice, this assumption holds with high probability due to numerical quantization errors.

Assumption 4.5. There exists some $\omega > 0$ such that $\omega x_{1,0}^2 \geq \|x_0 - x^\star\|^2$.

On Assumption 4.5. This assumption, which is mild and holds in most practical scenarios, prevents pathological cases in which $x_{1,0}$, the component of $x_0 - x^\star$ along the eigenspace associated with m, is arbitrarily small compared to the other components of $x_0 - x^\star$.

Theorem 4.6. *Let $f \in \mathcal{F}(L, m)$, $\bar{L} > L$ and $\bar{\kappa} = \bar{L}/m$. Suppose that $\bar{\kappa} > \kappa = L/m \geq 4$ and that Assumptions 4.2 to 4.5 hold. Then, there is some $\epsilon > 0$ such that if $\|x_0 - x^\star\| \leq \epsilon$, then the iterates x_t produced by Algorithm 1 satisfy*

$$\|x_{t+1} - x^\star\| \leq C r_{\text{acc}}(\sigma\bar{\kappa})^t \|x_0 - x^\star\|,$$

where σ depends on γ, C depends on $\bar{\kappa}$ and ω, with γ and ω given by Assumptions 4.2 to 4.5.

Proof sketch. In the proof, we analyze the two regimes that characterize the behavior of Algorithm 1 around $x^\star$. In the initial regime, m_t approaches m at an accelerated rate. In the final regime, which begins once m_t becomes sufficiently accurate, $x_t \to x^\star$ at a different, but also accelerated rate.

The fundamental mechanisms behind both regimes are captured by the case in which $f \in \mathcal{F}(L, m)$ is quadratic. Having shown that $x_t \to x^\star$ globally at a linear rate, we use Assumptions 4.3 to 4.5 to show that the general case consists in a perturbation of the quadratic case. We therefore convey the main ideas behind the proof by focusing our exposition on the quadratic case.

Let $H \in \mathbb{R}^{d \times d}$ denote $\nabla^2 f(x^\star)$, the Hessian of $f \in \mathcal{F}(L, m)$ at $x^\star$. Since $\nabla^2 f$ is locally Lipschitz at $x^\star$, it is also continuous at $x^\star$, which implies that H is real symmetric. Hence, its eigenvectors v_i can be chosen to form an orthonormal eigenbasis for $\mathbb{R}^d$. With that in mind, we consider $x_t - x^\star = \sum_{i=1}^d x_{i,t} v_i$. Assumption 4.2 implies that m_t can only take finitely many values, and using this fact we can show that each $x_{i,t}$ evolves as sequence of finitely many second-order linear time-invariant (LTI) systems. We analyze the dynamics of these d LTI systems, and show that $x_{i,t}$ for which $\lambda_i > m_t$ decrease much faster than $x_{i,t}$ such that $\lambda_i \leq m_t$ while m_t is not close to m. Then, we note that $c_{t+1} = \|H(x_{t+1} - x_t)\|/\|x_{t+1} - x_t\|$, which amounts to a power iteration with some additional dynamics due to the internal dynamics of $x_{t+1} - x_t$. It follows that c_t approaches m at an accelerated rate $r_{\text{acc}}(\sigma_\phi \bar{\kappa})$, where σ_ϕ is a somewhat complicated suboptimality factor. This can be seen by noticing that $\|H(x_{t+1} - x_t)\| = \sum_{i=1}^d \lambda_i^2 (x_{i,t+1} - x_{i,t})^2$, which is a consequence of v_i being orthonormal. We then analyze the final regime. Since m_t is nonincreasing, once m_t reaches some neighborhood $[m/\gamma, (1+\delta_m)m]$, m_t never leaves it. Each neighborhood translates into different suboptimality factors. The lower half of the neighborhood, $[m/\gamma, m]$, entails a suboptimality factor of γ, whereas the upper half, $[m, (1+\delta_m)m]$, entails a factor of $\sigma_m = 1 + 2\delta_m + 2\sqrt{\delta_m(1+\delta_m)}$. The final convergence rate $r_{\text{acc}}(\sigma\bar{\kappa})$ accounts for $\sigma_\phi, \gamma, \sigma_m$, the approximation error in the general case and an additional factor due to the Lyapunov argument that we use to analyze the final regime. □

5 Numerical Experiments

In this section, we present Algorithm 2, a heuristic to solve (1) based on Algorithm 1 and benchmark it against several standard methods described below. Algorithm 2 extends Algorithm 1 by incorporating an analogous estimator for L, resulting in an entirely parameter-free method. Since in this section we mostly present experiments involving Algorithm 2, we refer to it as NAG-free.

Since L_0 and m_0 are random variables, for each experiment we run NAG-free five times with different seeds, take the mean, maximum and minimum of the suboptimality gaps at each iteration, and then plot the mean with a thicker line inside a shaded region between the maximum and minimum.

The code for experiments below was heavily based on two sources: `github.com/ymalitsky/adaptive_GD` and `github.com/konstmish/opt_methods`. See Appendix C for more details.

Algorithm 2 NAG-free: a parameter-free heuristic extension of Algorithm 1 that estimates L and m.

Input: $T > 0, x_0 = y_0$
Output: x_T
$y \sim x_0 + U[0, 10^{-6}]^d$
$L_0, m_0 \leftarrow \|\nabla f(x_0) - \nabla f(y)\| / \|x_0 - y\|$
for $t = 0, \dots, T-1$ **do**
  $y_{t+1} \leftarrow x_t - (1/L_t)\nabla f(x_t)$
  $x_{t+1} \leftarrow y_{t+1} + \dfrac{\sqrt{L_t} - \sqrt{m_t}}{\sqrt{L_t} + \sqrt{m_t}}(y_{t+1} - y_t)$
  $c_{t+1} \leftarrow \|\nabla f(x_{t+1}) - \nabla f(x_t)\| / \|x_{t+1} - x_t\|$
  $L_{t+1} \leftarrow \max(L_t, c_{t+1})$
  $m_{t+1} \leftarrow \min(m_t, c_{t+1})$
end for

5.1 Methods

As baselines, we take GD, NAG and the Triple Momentum Method [Van Scoy et al., 2017, TMM], replacing L and m with problem-specific bounds. As competing methods, we consider two restart schemes based on [O'Donoghue and Candès, 2015]: one where L is replaced with a problem-specific bound, that we refer to as NAG+R, and another where L is estimated online with backtracking, that we refer to as NAG+RB. For NAG+RB, every time the L estimate L_t fails to produce enough descent, it is increased to $1.01L_t$ and tested again. This choice of adjustment factor produces less conservative estimates at the expense of more function evaluations, which largely favors NAG+RB since in all experiments we plot the suboptimality gap $f(x_t) - f(x^\star)$ versus iterations. As additional methods for comparison, we consider the adaptive gradient method AdGD [Malitsky and Mishchenko, 2024] and its accelerated heuristic, AdGD-accel2, both using $\gamma = 1/\sqrt{2}$, as well as a previous variant of the accelerated heuristic [Malitsky and Mishchenko, 2020], which we denote by AdGD-accel.

5.2 Smoothed and regularized log-sum-exp

First, we solve the smoothed, regularized log-sum-exp problem defined by the objective function

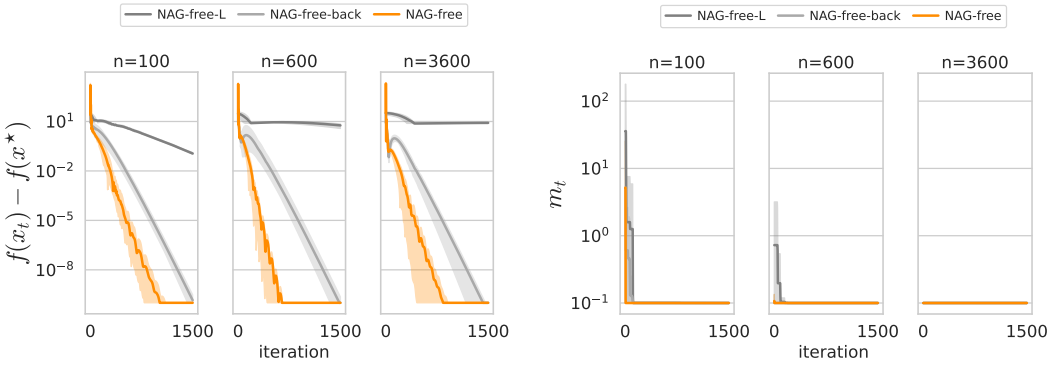
$$f(x) = \theta \log\left(\sum_{i=1}^{n} \exp\left(\frac{A_i^\mathsf{T} x - b_i}{\theta}\right)\right) + (\eta/2)\|x\|^2, \tag{6}$$

where $(A_i, b_i) \in \mathbb{R}^d \times \mathbb{R}$ are n observations, $\eta > 0$ and $\theta > 0$ is a smoothing parameter that controls how well f approximates the $\max$ function (see Appendix C for details). For this problem, we have $L \leq \bar{L} = (1 + 1/\theta)\sigma_{\max}(A)^2 + \eta$ and $m \geq \eta$, where A and $\sigma_{\max}(A)$ denote the matrix with rows given by A_i^T and the largest singular value of A. We use an implementation of (6) and corresponding initialization settings available at `https://github.com/konstmish/opt_methods`.

We start by assessing how well different approaches to estimate L work together with our proposed m estimator. Specifically, we consider three approaches: upper bounding L, backtracking, and directly estimating L. Figure 1a shows the suboptimality gap for log-sum-exp ($d = 600, \eta = 0.1, \theta = 0.1$) solved by Algorithms 1, 2 and 3, representing the three approaches, which we denote by NAG-free-L, NAG-free-back and NAG-free, respectively. We see that Algorithm 1 outperforms the other methods even though Algorithm 3 uses an adjustment factor of 1.01, which leads to minimally conservative bounds at the expense of an unreasonable number of backtracking adjustments.

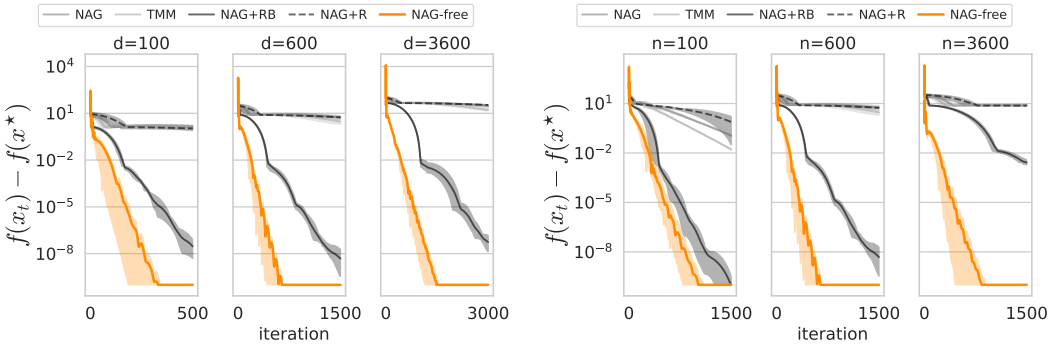
As a sanity check, we also confirm whether the m estimator continues to produce accurate estimates when working along different L estimators. In Figure 1b we see that m_t converges to 0.01 for all variants of NAG-free, which is the value of the regularization parameter η, indicating that the Hessian of the log-sum-exp component of (6) is singular at the minimum.

Having validated that Algorithm 2 is a reasonable heuristic for this problem, we consider several variations of (6) by changing problem parameters. To assess sensitivity to problem dimensionality, we fix $n = 600$ and $\theta = \eta = 0.1$, and then vary $d \in \{100, 600, 3600\}$. Likewise, we then vary n, θ and η, fixing the other problem parameters. The results in Figures 2 and 8 (see Appendix C) show that NAG-free again outperforms the other methods.



(a) Suboptimality gap. (b) m_t estimates.

Figure 1: Suboptimality gap and m_t estimates for log-sum-exp ($d = 600, \eta = 0.1, \theta = 0.1$).



(a) $n = 600$. (b) $d = 600$.

Figure 2: Suboptimality gap for log-sum-exp ($d = 600, \eta = 0.1$) using methods from Section 5.1.

5.3 Regularized logistic regression

Next, we consider the regularized logistic regression objective

$$f(x) = -(1/n)\sum_{i=1}^{n} \log\big(1 + \exp(-b_i A_i^\mathsf{T} x)\big) + (\eta/2)\|x\|^2, \tag{7}$$

where $\eta > 0$ and $(A_i, b_i) \in \mathbb{R}^d \times \{0, 1\}$ are n observations from a given dataset, which we take from the LIBSVM library [Chang and Lin, 2011]. See Table 1 in Appendix C for dataset details. Together with η, the datapoints determine the unknown parameters of $f \in \mathcal{F}(L, m)$, bounded by $L \leq \bar{L} = (1/4n)\lambda_{\max}(A^\mathsf{T}A) + \eta$ and $m \geq \eta$, where A denotes the matrix with rows A_i^T and $\lambda_{\max}(A^\mathsf{T}A)$ denotes the top eigenvalue of $A^\mathsf{T}A$.

Figure 3 shows results for six datasets, when $x_0 = 0$. We see that NAG-free is competitive with the other methods, while enjoying significant advantages over them. For example, NAG-free iterations are simpler and computationally less expensive than those of restarting schemes, which translates into clock time savings. Moreover, both NAG and TMM require pre-computing conservative parameter bounds to be used as surrogates for the true parameters. If the bounds are tight, then TMM slightly outperforms NAG-free, which is expected since it enjoys the fastest known provable convergence rate for this kind of algorithm. But in general, there are no guarantees of how tight the bounds are, which undermines the theoretical advantage that TMM has over NAG-free.

5.4 Cubic regularization

The third problem we consider is cubic regularization, defined by

$$f(x) = g^\mathsf{T}x + (1/2)x^\mathsf{T}Hx + (\eta/6)\|x\|^3, \tag{8}$$

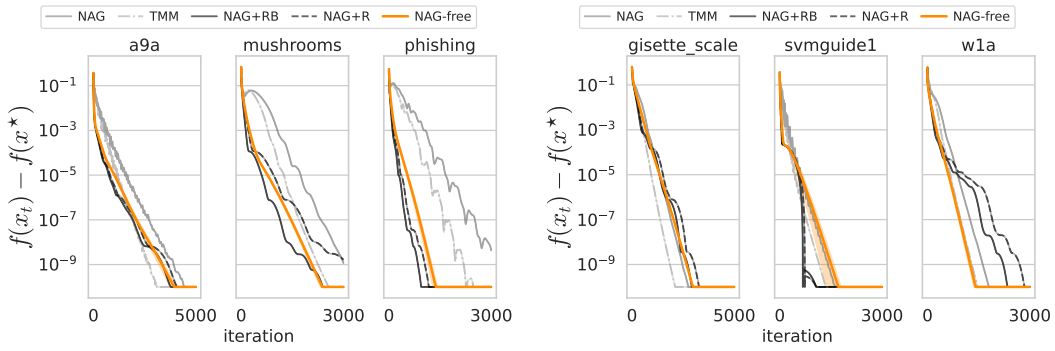


Figure 3: Suboptimality gap for logistic regression on six datasets using the methods from Section 5.1.

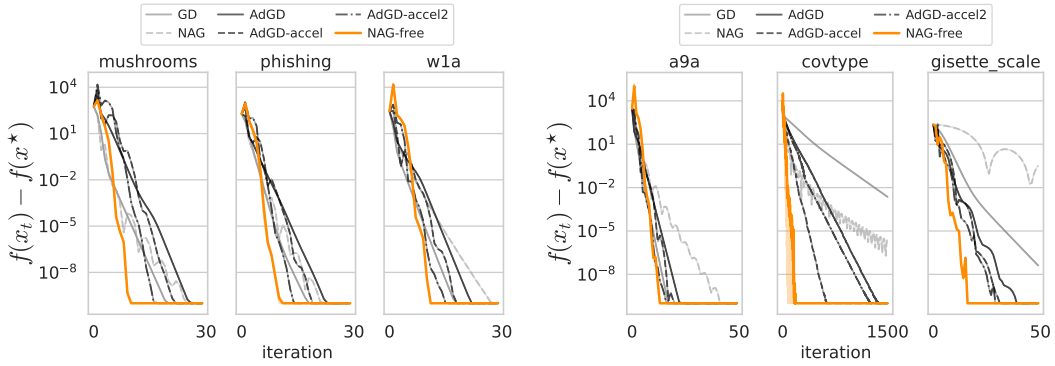


Figure 4: Suboptimality gap for cubic regularization on six datasets using methods from Section 5.1.

where $g \in \mathbb{R}^d$, $H \in \mathbb{R}^{d \times d}$ and $\eta > 0$. Following https://github.com/konstmish/opt_methods, from which we borrowed the code for this experiment, g and H are taken as the gradient and the Hessian at $x = 0$ of (7) initialized with LIBSVM datasets [Chang and Lin, 2011], without regularization—that is, $\eta = 0$ on (7), but $\eta > 0$ on (8). Also, as in the original source code, the regularizer parameter is set to $\eta = 10\bar{L}n$ (see Section 5.3.) Since (7) is convex, the eigenvalues of H are nonnegative, but not necessarily positive because H is taken as the Hessian of (7) with $\eta = 0$. Hence, the problem is not strongly convex because the regularizer in (8) is cubic. For the same reason, (8) is only locally smooth. Accordingly, the NAG method for weakly convex problems is used, with step sizes found through a grid search, which is also how the step sizes of GD are determined.

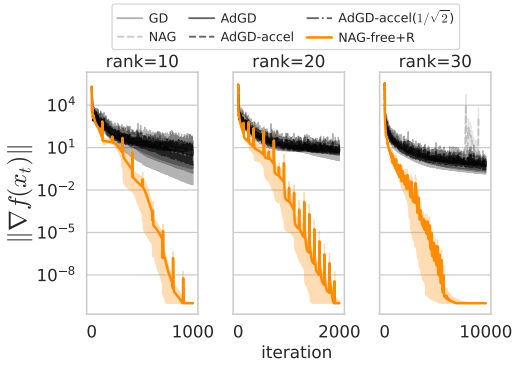
Because this problem is only locally smooth, we expect its curvature to change substantially locally. Hence, to keep only relevant curvature information, we restart NAG-free after some fixed number of iterations, resetting L_t and m_t to the last value taken by c_t and resetting $y_t = x_t$. In this experiment, NAG-free is restarted every iteration. Figure 4 shows that NAG-free consistently outperforms all other methods substantially. In particular, exploiting local information with NAG-free leads to better results than finding step sizes for GD and NAG through grid search.

5.5 Matrix factorization

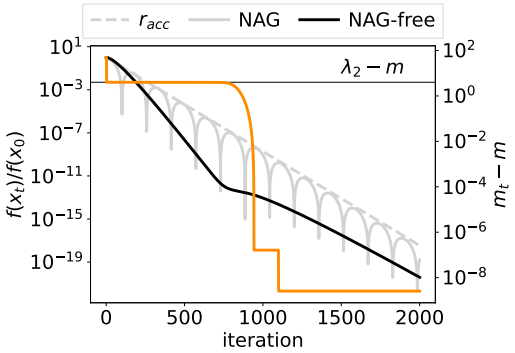
Our last experiment is matrix factorization, defined by the objective

$$f(U, V) = \tfrac{1}{2}\|UV^\mathsf{T} - A\|_F^2, \tag{9}$$

where $U \in \mathbb{R}^{m \times r}$, $V \in \mathbb{R}^{n \times r}$, $r < \min\{m, n\}$, and $\|\cdot\|_F$ denotes the Frobenius norm. In this experiment, the matrix A is taken from the MovieLens 100K dataset [Harper and Konstan, 2015], so that $m = 943$ and $n = 1682$. Because (9) is nonconvex, we restart NAG-free every 100 iterations to retain only relevant curvature information. As Figure 5a shows, after every restart there is a



(a) Gradient norm for matrix factorization with three different ranks using Section 5.1 methods.



(b) Normalized performance for $f(x) = (1/2)x^\mathsf{T}Ax$, with $A = \mathrm{diag}(1, 5, 10^4)$ and $x_0 = [1, 10^5, 1]^\mathsf{T}$.

Figure 5: Matrix factorization and an adversarial toy problem.

momentary spike in the norm of the gradient—which could be due to overly aggressive step sizes. But the spikes plummet as local information is updated, and NAG-free outperforms the other methods.

5.6 Further experiments

In Appendix C.6, we present further experiments targeting edge cases of the theoretical results from Section 4.2. For example, on Figure 5b, we show the normalized performance of NAG-free on a quadratic problem on the left-hand y-axis and the estimate gap $m_t - m$ with the golden line on right-hand y-axis. The Hessian spectrum is specifically designed to stress test NAG-free when the initial conditions are concentrated on the second smallest eigenvalue. We see that instead of slowing down, NAG-free in fact accelerates, and behaves as if the effective condition number had increased until the relative 2-norm of x_t on the m-eigenspace becomes comparable to the total 2-norm.

6 Conclusion

In this paper, we introduced NAG-free, an optimization method designed for Lipschitz-smooth, strongly convex problems. NAG-free efficiently estimates the strong convexity parameter m, while maintaining reasonable worst-case guarantees without restart schemes. The method couples Nesterov's accelerated gradient (NAG) method with a lightweight estimator that requires only the storage of one more iterate and gradient already computed by NAG. We proved that NAG-free converges globally as fast as gradient descent (GD) and achieves acceleration locally near the minimum. In addition, we proposed a heuristic with good empirical performance, which estimates both m and the Lipschitz-smoothness parameter L online. Importantly, the L-estimator does not require additional storage or computation, and the two estimators make the heuristic free of hyperparameters.

Despite these advances, many questions remain for future work. An important one is whether methods with stronger convergence guarantees than NAG, such as TMM, can be effectively coupled with parameter estimators. Another avenue for exploration is whether replacing the curvature term used by NAG-free with similar terms that lie in $[m, L]$ can further improve practical performance, such as

$$\langle \nabla f(x_{t+1}) - \nabla f(x_t), x_{t+1} - x_t \rangle / \|x_{t+1} - x_t\|^2.$$

A key limitation of this work is that, while we address the open question of efficiently estimating m online without restarts, our experiments demonstrate that both m and L can be efficiently estimated in practice. Theoretically analyzing the L-estimator will be an interesting future challenge. Moreover, we conjecture that our L-estimator has promising applications in domains beyond the strong convexity setting, and exploring this broader applicability will be another important area for future investigation.

References

N. Bansal and A. Gupta. Potential-function proofs for gradient methods. *Theory of Computing*, 15 (1):1–32, 2019.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

S. Bodine and D. A. Lutz. *Asymptotic Integration of Differential and Difference Equations*. Springer Cham, 2015.

S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(27):1–27, 2011.

J. B. Conway. *A course in functional analysis*. Springer, 2019.

A. d'Aspremont, D. Scieur, and A. Taylor. Acceleration methods. *Foundations and trends in optimization*, 5(1-2):1–245, 2021.

G. H. Golub and C. F. Van Loan. *Matrix computations*. Johns Hopkins University Press, 4th edition, 2013.

M.F. Harper and J.A. Konstan. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5(4), 2015.

J. P. Hespanha. *Linear systems theory*. Princeton Univ. Press, 2nd edition, 2009. ISBN 9780691140216.

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

Y. Malitsky and K. Mishchenko. Adaptive gradient descent without descent. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

Y. Malitsky and K. Mishchenko. Adaptive proximal gradient method for convex optimization. In *NeurIPS*, 2024.

Y. Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

J. Nocedal and S. Wright. *Numerical Optimization*. Springer Science & Business Media, 2006.

B. O'Donoghue and E. J. Candès. Adaptive restart for accelerated gradient schemes. *Foundations of computational mathematics*, 15(3):715–732, 2015.

V. Roulet and A. d'Aspremont. Sharpness, restart and acceleration. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

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

P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. 2008. URL http://www.mit.edu/~dimitrib/PTseng/papers.html.

B. Van Scoy, R. A. Freeman, and K. M. Lynch. The fastest known globally convergent first-order method for minimizing strongly convex functions. *IEEE Control System Letters*, 2(1):49–54, 2017.

A Global convergence

In the following two theorems we prove that, up to a constant factor, Algorithm 1 converges globally at least as fast as gradient descent (GD). The first theorem concerns the case where $m_t \geq m$, which always holds because $c_t \geq m$, by design. The second theorem further allows for the case where $m_t < m$. We consider this case for the sake of completeness because it can occur under Assumption 4.2, which we use to prove local acceleration. But since Assumption 4.2 only serves a theoretical purpose, we omit the second theorem in the main paper.

Theorem (4.1)**.** Let $f \in \mathcal{F}(L, m)$, $\bar{L} \geq L$ and $\bar{\kappa} = \bar{L}/m$. Then, the iterates of Algorithm 1 satisfy

$$f(y_t) - f(x^\star) \leq 2\bar{L}\left(\frac{\bar{\kappa} - 1}{\bar{\kappa}}\right)^t \|x_0 - x^\star\|^2.$$

Theorem A.1. *Let $f \in \mathcal{F}(L, m)$, $\bar{L} \geq L$ and $\bar{\kappa} = \bar{L}/m$. Also, suppose that Assumption 4.2 holds for some $\gamma > 1$. Then the iterates y_t of Algorithm 1 satisfy*

$$f(y_t) - f(x^\star) \leq 2\gamma\bar{L}\bar{\kappa}^2\left(\frac{\gamma\bar{\kappa} - 1}{\gamma\bar{\kappa}}\right)^t \|x_0 - x^\star\|^2. \tag{10}$$

Algorithm 1 takes some estimate $\bar{L}$ of L such that $\bar{L} \geq L$ as an input. If $f \in \mathcal{F}(L, m)$, then since $(\bar{L}/2)\|y - x\|^2 \geq (L/2)\|y - x\|^2$ for any x and y, we have that $f \in \mathcal{F}(\bar{L}, m)$:

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + (\bar{L}/2)\|y - x\|^2. \tag{11}$$

A.1 Case 1: $m_t \geq m$

The iterations of Algorithm 1 in which $m_t \geq m$ can be expressed as a convex combination of appropriate GD and NAG iterations. In this subsection, we exploit this property to prove that if Algorithm 1 receives some $\bar{L} \geq L$ as input, then it converges at least as fast as GD. The argument is built upon a Lyapunov function that we denote by V_t^{GD}. The superscript "GD" indicates that V_t^{GD} decreases at a gradient-descent type of rate along iterations in which $m_t \geq m$. In the same vein, in the next subsection we introduce V_t^{NAG} and show that it decreases at an accelerated rate along iterations in which $m_t < m$.

The Lyapunov function V_t^{GD} is the sum of two components W and U scaled. First, we show W is a common Lyapunov function for GD and NAG, then we analyze U and finally combine all results to give Algorithm 1 the same type of convergence guarantees of GD. To analyze GD and NAG through a common Lyapunov function, we add a trivial momentum step x_{t+1}^{GD} to GD, as in

$$y_{t+1}^{\text{GD}} = x_t^{\text{GD}} - (1/\bar{L})\nabla f(x_t^{\text{GD}}), \tag{12}$$
$$x_{t+1}^{\text{GD}} = y_{t+1}^{\text{GD}}, \tag{13}$$

conforming GD to the algorithmic structure of NAG:

$$y_{t+1}^{\text{NAG}} = x_t^{\text{NAG}} - (1/\bar{L})\nabla f(x_t^{\text{NAG}}), \tag{14}$$
$$x_{t+1}^{\text{NAG}} = y_{t+1}^{\text{NAG}} + \theta(y_{t+1}^{\text{NAG}} - y_t^{\text{NAG}}), \tag{15}$$

where the coefficient θ defining the affine combination in (15) is given by

$$\theta = (\sqrt{\bar{\kappa}} - 1)/(\sqrt{\bar{\kappa}} + 1), \tag{16}$$

a function of an upper bound on the *condition number* of $f \in \mathcal{F}(L, m)$:

$$\bar{\kappa} = \bar{L}/m \geq L/m = \bar{\kappa}. \tag{17}$$

The coefficient θ is defined in terms of $\bar{\kappa}$ because we consider the NAG implementation (14)–(15) where m is known. In contrast, Algorithm 1 is implemented with an estimate m_t of m, so that

$$y_{t+1} = x_t - (1/\bar{L})\nabla f(x_t), \tag{18}$$
$$x_{t+1} = y_{t+1} + \beta_t(y_{t+1} - y_t), \tag{19}$$

where now the coefficient

$$\beta_t = (\sqrt{\kappa_t} - 1)/(\sqrt{\kappa_t} + 1) \tag{20}$$

is a function of the *estimated condition number*

$$\bar{\kappa}_t = \bar{L}/m_t. \tag{21}$$

The fact that we are able to write NAG-free as a convex combination of GD and NAG does not imply that m is known and is only possible for iterations in which $m_t \geq m$.

Expressed in the common structure of (12) to (15), GD and NAG can be analyzed with a common Lyapunov function that can be found in [Bansal and Gupta, 2019, 5.5], and given by

$$W(s_t) = \tilde{f}(y_t) + (m/2)\|z_t^\star\|^2, \tag{22}$$

where s_t groups the descent and momentum steps into a single pair, as in

$$s_t = (x_t, y_t), \qquad s_t^{\text{GD}} = (x_t^{\text{GD}}, y_t^{\text{GD}}), \qquad \text{and} \qquad s_t^{\text{NAG}} = (x_t^{\text{NAG}}, y_t^{\text{NAG}}), \tag{23}$$

$\tilde{f}$ denotes the objective function normalized to 0, meaning that

$$\tilde{f} = f - f(x^\star), \tag{24}$$

and $z_t^\star$ is the pseudo-state defined as

$$z_t^\star = z_t - x^\star, \qquad z_t = x_t + \sqrt{\bar{\kappa}}(x_t - y_t) \tag{25}$$

For the sake of exposition, we refer to the gradient at iteration t as $g_t = \nabla f(x_t)$. Also, we define

$$x_t^\star = x_t - x^\star \qquad \text{and} \qquad x_t^y = x_t - y_t. \tag{26}$$

Remark A.2. Superscripts carry over from (12) to (15) to the notation above in the natural way. For example, by $g_t^{\text{NAG}} = \nabla f(x_t^{\text{NAG}})$ and $x_t^{\text{GD},\star} = x_t^{\text{GD}} - x^\star$.

Gradient descent is slower than Nesterov's method, so even though W serves as a Lyapunov function for both, we should expect W to decrease at a faster rate along NAG iterations than along GD iterations. We now show that W decreases at the expected rate for each of the two methods, namely $(1 + \delta(\bar{\kappa}))^{-1}$ for GD and $(1 + \delta(\sqrt{\bar{\kappa}}))^{-1}$ for NAG, where the rate increment δ is defined by

$$\delta(p) = 1/(p-1). \tag{27}$$

For the sake of presentation, we also define

$$\delta^{\text{GD}} = \delta(\bar{\kappa}) = 1/(\bar{\kappa} - 1) \qquad \text{and} \qquad \delta^{\text{NAG}} = \delta(\sqrt{\bar{\kappa}}) = 1/(\sqrt{\bar{\kappa}} - 1). \tag{28}$$

To prove global convergence results for Algorithm 1, we work with Lyapunov functions that fit the same template of W: a sum of the difference $f(y_t) - f(x^\star)$ and an appropriately weighted 2-norm of a pseudo-state, $z_t^\star$ in the case of W. We characterize two types of changes for W and subsequent Lyapunov functions: the decrease in a fixed W from one iteration s_t to the next s_{t+1} and the increase from W to W for the same iteration s_t. Considering the second type of change is necessary because W varies with t, following $\sqrt{\bar{\kappa}}$ on (25) as it varies with t. The first type of change always looks like an inequality such as (29). To prove (29) and similar inequalities, the procedure is always the same: express the difference $f(y_t) - f(x^\star)$ as the sum of two or three other differences, parse them using properties of $\mathcal{F}(L, m)$, expand the pseudo-states inside the 2-norms and then put everything together. The proof of the next result illustrates this procedure.

Lemma A.3. *Let $f \in \mathcal{F}(L, m)$, $\bar{L} \geq L$, $y_t^{\text{GD}} = x_t^{\text{GD}}$. If $y_{t+1}^{\text{GD}}, x_{t+1}^{\text{GD}}$ are given by (12) and (13), then*

$$(1 + \delta^{\text{GD}})W(s_{t+1}^{\text{GD}}) - W(s_t^{\text{GD}}) \leq -(1/2\bar{L})\|g_t^{\text{GD}}\|^2. \tag{29}$$

Proof. Let $f \in \mathcal{F}(L, m)$ and $\bar{L} \geq L$. If y_{t+1}^{GD} and x_{t+1}^{GD} are given by (12) and (13), then plugging $y = y_{t+1}^{\text{GD}}$ and $x = x_t^{\text{GD}}$ back into (11), it follows that

$$f(y_{t+1}^{\text{GD}}) - f(x_t^{\text{GD}}) \leq -(1/2\bar{L})\|g_t^{\text{GD}}\|^2. \tag{30}$$

Following the procedure described above, we start by expressing $f(y_t) - f(x^\star)$ as the sum of two differences

$$(1 + \delta^{\text{GD}})\tilde{f}(y_{t+1}^{\text{GD}}) - \tilde{f}(y_t^{\text{GD}}) = \underbrace{(1 + \delta^{\text{GD}})(f(y_{t+1}^{\text{GD}}) - f(y_t^{\text{GD}}))}_{\#1} + \underbrace{\delta^{\text{GD}}(f(y_t^{\text{GD}}) - f(x^\star))}_{\#2}.$$

In the gradient descent implementation above, step (13) simply propagates step (12), i.e., $x_t^{\mathrm{GD}} = y_t^{\mathrm{GD}}$ and in turn $f(y_t^{\mathrm{GD}}) = f(x_t^{\mathrm{GD}})$. Hence, since (30) holds, then the first difference is bounded as

$$(1+\delta^{\mathrm{GD}})(f(y_{t+1}^{\mathrm{GD}}) - f(y_t^{\mathrm{GD}})) \leq -\frac{1+\delta^{\mathrm{GD}}}{2\bar{L}}\|g_t^{\mathrm{GD}}\|^2. \tag{31}$$

Moreover, applying (3) with $x = y_t^{\mathrm{GD}} = x_t^{\mathrm{GD}}$ and $y = x^\star$, we bound the second difference as

$$\delta^{\mathrm{GD}}(f(y_t^{\mathrm{GD}}) - f(x^\star)) \leq \delta^{\mathrm{GD}}\langle g_t^{\mathrm{GD}}, x_t^{\mathrm{GD},\star}\rangle - \delta^{\mathrm{GD}}(m/2)\|x_t^{\mathrm{GD},\star}\|^2. \tag{32}$$

Also from $x_t^{\mathrm{GD}} = y_t^{\mathrm{GD}}$, it follows that $z_t^{\mathrm{GD}} = x_t^{\mathrm{GD}}$ and, likewise, $z_{t+1}^{\mathrm{GD}} = y_{t+1}^{\mathrm{GD}}$. Therefore

$$\begin{aligned}(1+\delta^{\mathrm{GD}})\|z_{t+1}^{\mathrm{GD},\star}\|^2 - \|z_t^{\mathrm{GD},\star}\|^2 &= (1+\delta^{\mathrm{GD}})\|y_{t+1}^{\mathrm{GD},\star}\|^2 - \|x_t^{\mathrm{GD},\star}\|^2 \\ &= (1+\delta^{\mathrm{GD}})\Big(\frac{\|g_t^{\mathrm{GD}}\|^2}{\bar{L}^2} - \frac{2\langle g_t^{\mathrm{GD}}, x_t^{\mathrm{GD},\star}\rangle}{\bar{L}}\Big) + \delta^{\mathrm{GD}}\|x_t^{\mathrm{GD},\star}\|^2. \end{aligned} \tag{33}$$

To simplify the above and conclude the proof, we use the identities

$$(1+\delta^{\mathrm{GD}})\Big(1-\frac{1}{\bar{\kappa}}\Big) = \frac{\bar{\kappa}}{\bar{\kappa}-1}\frac{\bar{\kappa}-1}{\bar{\kappa}} = 1 \qquad \text{and} \qquad \frac{1+\delta^{\mathrm{GD}}}{\bar{\kappa}} = \frac{\bar{\kappa}/(\bar{\kappa}-1)}{\bar{\kappa}} = \delta^{\mathrm{GD}}.$$

Multiplying (33) by $m/2$, summing the result with (31) and (32), then using the above, we obtain

$$\begin{aligned}(1+\delta^{\mathrm{GD}})W(s_{t+1}^{\mathrm{GD}}) - W(s_t^{\mathrm{GD}}) \leq\ & -(1+\delta^{\mathrm{GD}})(1-1/\bar{\kappa})(1/2\bar{L})\|g_t^{\mathrm{GD}}\|^2 \\ & -(\delta^{\mathrm{GD}} - (1+\delta^{\mathrm{GD}})/\bar{\kappa})\langle g_t^{\mathrm{GD}}, x_t^{\mathrm{GD},\star}\rangle \\ \leq\ & -(1/2\bar{L})\|g_t^{\mathrm{GD}}\|^2,\end{aligned}$$

proving (29). □

Lemma A.4. *Let $f \in \mathcal{F}(L,m)$ and $\bar{L} \geq L$. Given $y_t^{\mathrm{NAG}}, x_t^{\mathrm{NAG}}$, if $y_{t+1}^{\mathrm{NAG}}, x_{t+1}^{\mathrm{NAG}}$ are generated by (14) and (15), then*

$$(1+\delta^{\mathrm{NAG}})W(s_{t+1}^{\mathrm{NAG}}) - W(s_t^{\mathrm{NAG}}) \leq 0. \tag{34}$$

Proof. Let $f \in \mathcal{F}(L,m)$ and $\bar{L}$. If y_{t+1}^{NAG} and x_{t+1}^{NAG} are generated by (14) and (15), then plugging $y = y_{t+1}^{\mathrm{NAG}}$ and $x = x_t^{\mathrm{NAG}}$ back into (11), it follows that

$$f(y_{t+1}^{\mathrm{NAG}}) - f(x_t^{\mathrm{NAG}}) \leq -(1/2\bar{L})\|g_t^{\mathrm{NAG}}\|^2, \tag{35}$$

To prove (34), our first move is to express a difference as the sum of three further differences:

$$\begin{aligned}(1+\delta^{\mathrm{NAG}})\tilde{f}(y_{t+1}^{\mathrm{NAG}}) - \tilde{f}(y_t^{\mathrm{NAG}}) = & \underbrace{(1+\delta^{\mathrm{NAG}})(f(y_{t+1}^{\mathrm{NAG}}) - f(x_t^{\mathrm{NAG}}))}_{\#1} \\ & + \underbrace{f(x_t^{\mathrm{NAG}}) - f(y_t^{\mathrm{NAG}})}_{\#2} \\ & + \underbrace{\delta^{\mathrm{NAG}}(f(x_t^{\mathrm{NAG}}) - f(x^\star))}_{\#3}.\end{aligned}$$

Since (35) holds, then we bound the first difference as

$$(1+\delta^{\mathrm{NAG}})(f(y_{t+1}^{\mathrm{NAG}}) - f(x_t^{\mathrm{NAG}})) \leq -(1+\delta^{\mathrm{NAG}})(1/2\bar{L})\|g_t^{\mathrm{NAG}}\|^2.$$

Using that f is convex and applying (3) with $x = x_t^{\mathrm{NAG}}$ and $y = x^\star$, we bound #2 and #3 as

$$\begin{aligned}f(x_t^{\mathrm{NAG}}) - f(y_t^{\mathrm{NAG}}) &\leq \langle g_t^{\mathrm{NAG}}, x_t^{\mathrm{NAG},y}\rangle, \\ f(x_t^{\mathrm{NAG}}) - f(x^\star) &\leq \langle g_t^{\mathrm{NAG}}, x_t^{\mathrm{NAG},\star}\rangle - (m/2)\|x_t^{\mathrm{NAG},\star}\|^2.\end{aligned}$$

To address the rest of $(1+\delta^{\mathrm{NAG}})W(s_{t+1}^{\mathrm{NAG}}) - W(s_t^{\mathrm{NAG}})$, we expand $z_{t+1}^{\mathrm{NAG},\star}$, as in

$$\begin{aligned}z_{t+1}^{\mathrm{NAG},\star} &= x_{t+1}^{\mathrm{NAG}} + \sqrt{\bar{\kappa}}(x_{t+1}^{\mathrm{NAG}} - y_{t+1}^{\mathrm{NAG}}) - x^\star \\ &= y_{t+1}^{\mathrm{NAG}} + \theta(y_{t+1}^{\mathrm{NAG}} - y_t^{\mathrm{NAG}}) + \sqrt{\bar{\kappa}}\theta(y_{t+1}^{\mathrm{NAG}} - y_t^{\mathrm{NAG}}) - x^\star \\ &= -(1+\theta(1+\sqrt{\bar{\kappa}}))g_t^{\mathrm{NAG}}/\bar{L} + \theta(1+\sqrt{\bar{\kappa}})x_t^{\mathrm{NAG},y} + x_t^{\mathrm{NAG},\star} \\ &= -\sqrt{\bar{\kappa}}g_t^{\mathrm{NAG}}/\bar{L} + (\sqrt{\bar{\kappa}}-1)x_t^{\mathrm{NAG},y} + x_t^{\mathrm{NAG},\star},\end{aligned}$$

and then note that the 2-norm term in $W(s_t^{\text{NAG}})$ is $(m/2)\|x_t^{\text{NAG},\star} + \sqrt{\bar{\kappa}}x_t^{\text{NAG},\star}\|^2$, because

$$W(s_t^{\text{NAG}}) = \tilde{f}(y_t^{\text{NAG}}) + (m/2)\|x_t^{\text{NAG},\star} + \sqrt{\bar{\kappa}}x_t^{\text{NAG},\star}\|^2.$$

Also, we use the following identities after colons to simplify the coefficients of the terms before colons:

$$\begin{aligned}
\langle g_t^{\text{NAG}}, x_t^{\text{NAG},y}\rangle &: && (1+\delta^{\text{NAG}})\sqrt{\bar{\kappa}}(\sqrt{\bar{\kappa}}-1)/\bar{\kappa} = 1,\\
\langle g_t^{\text{NAG}}, x_t^{\text{NAG},\star}\rangle &: && (1+\delta^{\text{NAG}})\sqrt{\bar{\kappa}}/\bar{\kappa} = \delta^{\text{NAG}},\\
\|x_t^{\text{NAG},y}\|^2 &: && (1+\delta^{\text{NAG}})(\sqrt{\bar{\kappa}}-1)^2 = \sqrt{\bar{\kappa}}(\sqrt{\bar{\kappa}}-1),\\
\langle x_t^{\text{NAG},y}, x_t^{\text{NAG},\star}\rangle &: && (1+\delta^{\text{NAG}})(\sqrt{\bar{\kappa}}-1) = \sqrt{\bar{\kappa}},
\end{aligned}$$

Then, we write the 2-norm difference in $(1+\delta^{\text{NAG}})W(s_{t+1}^{\text{NAG}}) - W(s_t^{\text{NAG}})$ as

$$\begin{aligned}
(1+\delta^{\text{NAG}})(m/2)\|z_{t+1}^{\text{NAG},\star}\|^2 - (m/2)\|x_t^{\text{NAG},\star} + \sqrt{\bar{\kappa}}x_t^{\text{NAG},\star}\|^2 = \frac{1+\delta^{\text{NAG}}}{2\bar{L}}\|g_t^{\text{NAG}}\|^2 \\
- \langle g_t^{\text{NAG}}, x_t^{\text{NAG},y}\rangle \\
- \delta^{\text{NAG}}\langle g_t^{\text{NAG}}, x_t^{\text{NAG},\star}\rangle \\
- (m/2)\sqrt{\bar{\kappa}}\|x_t^{\text{NAG},y}\|^2 \\
+ \delta^{\text{NAG}}(m/2)\|x_t^{\text{NAG},\star}\|^2.
\end{aligned}$$

Finally, we put everything together and then cancel several terms to get

$$(1+\delta^{\text{NAG}})W(s_{t+1}^{\text{NAG}}) - W(s_t^{\text{NAG}}) \leq -(m/2)\sqrt{\bar{\kappa}}\|x_t^{\text{NAG},y}\|^2 \leq 0,$$

proving (34). □

Now that we have shown that W is a common Lyapunov function for GD and NAG, we introduce the second piece of V_t^{GD}, the function U defined by

$$U(s_t) = \tilde{f}(y_t) + (\bar{L}/2)\|y_t^\star\|^2 \tag{36}$$

where $\tilde{f} = f - f(x^\star)$ and, in the same spirit of $x_t^\star$ and $z_t^\star$, $y_t^\star$ is a pseudo-state defined by

$$y_t^\star = y_t - x^\star. \tag{37}$$

Lemma A.5. *Let $f \in \mathcal{F}(L,m)$ and $\bar{L} \geq L$. If Algorithm 1 receives $\bar{L}$ as input, then it generates iterates s_t such that*

$$(1+\delta^{\text{GD}})U(s_{t+1}) - U(s_t) \leq \bar{L}\langle x_t^y, x_t^\star\rangle - (\bar{L}/2)\|x_t^y\|^2. \tag{38}$$

Proof. First, we address the difference $(1+\delta^{\text{GD}})\tilde{f}(y_{t+1}) - \tilde{f}(y_t)$. By definition, $f(x^\star) \leq f(y_t)$, thus $-\tilde{f}(y_t) \leq 0$. Hence, adding $\pm(1+\delta^{\text{GD}})f(x_t)$ to $(1+\delta^{\text{GD}})\tilde{f}(y_{t+1}) - \tilde{f}(y_t)$ and discarding $-\tilde{f}(y_t)$, we get

$$(1+\delta^{\text{GD}})\tilde{f}(y_{t+1}) - \tilde{f}(y_t) \leq \underbrace{(1+\delta^{\text{GD}})(f(y_{t+1}) - f(x_t))}_{\#1} + \underbrace{(1+\delta^{\text{GD}})(f(x_t) - f(x^\star))}_{\#2}.$$

If $\bar{L}$, then plugging $y = y_{t+1}$ and $x = x_t$ back into (11), it follows that #1 is bounded as

$$(1+\delta^{\text{GD}})(f(y_{t+1}) - f(x_t)) \leq -(1+\delta^{\text{GD}})(1/2\bar{L})\|g_t\|^2.$$

To address the second difference, #2, we apply (3) with $x = x_t$ and $y = x^\star$, obtaining

$$(1+\delta^{\text{GD}})(f(x_t) - f(x^\star)) \leq (1+\delta^{\text{GD}})\Big(\langle g_t, x_t^\star\rangle - (m/2)\|x_t^\star\|^2\Big).$$

Then, we put the two bounds together to get

$$(1+\delta^{\text{GD}})\tilde{f}(y_{t+1}) - \tilde{f}(y_t) \leq -\frac{1+\delta^{\text{GD}}}{2\bar{L}}\|g_t\|^2 + (1+\delta^{\text{GD}})\langle g_t, x_t^\star\rangle - \delta^{\text{GD}}(\bar{L}/2)\|x_t^\star\|^2, \tag{39}$$

where the coefficient multiplying $\|x_t^\star\|^2$ on the right-hand side above stems from the identity

$$(1+\delta^{\mathrm{GD}})(m/2) = \frac{\bar{\kappa}}{\bar{\kappa}-1}(m/2) = \delta^{\mathrm{GD}}(\bar{L}/2).$$

To address the 2-norm difference in $(1+\delta^{\mathrm{GD}})U(s_{t+1}) - U(s_t)$, we expand pseudo-states inside 2-norms as:

$$(1+\delta^{\mathrm{GD}})\|y_{t+1}^\star\|^2 = (1+\delta^{\mathrm{GD}})\Big((1/\bar{L}^2)\|g_t\|^2 - (2/\bar{L})\langle g_t, x_t^\star\rangle + \|x_t^\star\|^2\Big),$$
$$\|y_t^\star\|^2 = \|x_t^y\|^2 - 2\langle x_t^y, x_t^\star\rangle + \|x_t^\star\|^2.$$

Expanding $\|y_{t+1}^\star\|^2$ and $\|y_t^\star\|^2$ as above, we get

$$\begin{aligned}(\bar{L}/2)((1+\delta^{\mathrm{GD}})\|y_{t+1}^\star\|^2 - \|y_t^\star\|^2) = (1+\delta^{\mathrm{GD}})\Big((1/2\bar{L})\|g_t\|^2 - \langle g_t, x_t^\star\rangle\Big)\\ + (\bar{L}/2)(\|x_t^y\|^2 + 2\langle x_t^y, x_t^\star\rangle + \|x_t^\star\|^2).\end{aligned} \tag{40}$$

Finally, combining (39) and (40), several terms cancel each other and we are left with

$$(1+\delta^{\mathrm{GD}})U(s_{t+1}) - U(s_t) \leq \bar{L}\langle x_t^y, x_t^\star\rangle - (\bar{L}/2)\|x_t^y\|^2,$$

proving (38). □

With Lemmas A.3 to A.5, we have sufficiently characterized W and U to prove the main result for iterations of Algorithm 1 in which $m_t \geq m$, using the Lyapunov function V_t^{GD} given by

$$V_t^{\mathrm{GD}} = W + \frac{\bar{\alpha}_{t-1}}{\sqrt{\bar{\kappa}}}U, \tag{41}$$

with

$$\bar{\alpha}_{t-1} = \begin{cases} 1-\alpha_0, & t-1 \leq 0, \\ 1-\alpha_{t-1}, & t-1 > 0, \end{cases} \qquad \alpha_t = \frac{\beta_t}{\theta}. \tag{42}$$

Before using V_t^{GD}, we show that $V_t^{\mathrm{GD}} \geq 0$ for iterations in which $m_t \geq m$.

Lemma A.6. *If $m_t \geq m$ and m_t is nonincreasing, then $V_{t'}^{\mathrm{GD}} \geq 0$ for all $0 \leq t' \leq t$. Moreover, $\bar{\alpha}_t$ is nonincreasing.*

Proof. The assumptions that $m_t \geq m$ and that m_t is nonincreasing imply $m_t \geq m$ for all $t' \leq t$. Moreover, $m_t \geq m$ implies that $\bar{\kappa}_t = \bar{L}/m_t < \bar{L}/m = \bar{\kappa}$, therefore

$$\beta_t = \frac{\sqrt{\bar{\kappa}_t}-1}{\sqrt{\bar{\kappa}_t}+1} < \frac{\sqrt{\bar{\kappa}}-1}{\sqrt{\bar{\kappa}}+1} = \theta.$$

Hence, $\beta_{t'} < \theta$ for all $t' \leq t$. Therefore, $\alpha_{t'}, \bar{\alpha}_{t'} \in [0,1]$ and, in turn, $V_{t'}^{\mathrm{GD}} \geq 0$ for all $t' \leq t$.

Moreover, since m_t is nonincreasing, β_t given by (20) is nondecreasing, hence so is α_t, by (42). In turn, by (42), it follows that $\bar{\alpha}_t$ is nonincreasing. □

Theorem A.7. *Let $f \in \mathcal{F}(L,m)$ and $\bar{L} \geq L$. If Algorithm 1 receives $\bar{L}$ as input, then it generates iterates s_t such that*

$$V_{t+1}^{\mathrm{GD}}(s_{t+1}) \leq 2\bar{L}(1+\delta^{\mathrm{GD}})^{-t}\|x_0^\star\|^2. \tag{43}$$

Proof. We establish (43) in three steps. First, we bound $(1+\delta^{\mathrm{GD}})V_{t+1}^{\mathrm{GD}}(s_{t+1})$ in terms of $V_{t+1}^{\mathrm{GD}}(s_t)$. Second, we bound $V_{t+1}^{\mathrm{GD}}(s_t)$ in terms of $V_t^{\mathrm{GD}}(s_t)$. Third, we use the bounds in an inductive argument.

To bound $(1+\delta^{\mathrm{GD}})V_{t+1}^{\mathrm{GD}}(s_{t+1})$ in terms of $V_{t+1}^{\mathrm{GD}}(s_t)$, we analyze their difference, which is the sum of one difference involving W and another one involving U. We address the one involving W first. To this end, we use the assumption that $m_t \geq m$ to show Algorithm 1 iterates can be expressed as a convex combination of appropriate GD and NAG iterates and then we exploit the fact that W is convex to bound the corresponding difference.

To show Algorithm 1 iterates are a convex combination of GD and NAG iterations, we consider fictitious "one-shot" GD and NAG iterations initialized at the given iterate of Algorithm 1. We let $y_t^{\text{GD}} = x_t^{\text{GD}} = x_t^{\text{NAG}} = x_t$ and $y_t^{\text{NAG}} = y_t$. We initialize x_t^{GD} and y_t^{GD} "backwards" from x_t to conform them to the GD iteration constraint that $y_t^{\text{GD}} = x_t^{\text{GD}}$. On the other hand, since NAG works with arbitrary initial points, we initialize NAG at the t-th iteration of Algorithm 1 exactly. With these initial points in mind, let y_{t+1}^{GD}, x_{t+1}^{GD}, y_{t+1}^{NAG} and x_{t+1}^{NAG} be the GD and NAG iterations produced by (12) to (15). Then, GD, NAG and Algorithm 1 produce the same descent step

$$y_{t+1}^{\text{GD}} = x_t^{\text{GD}} - \nabla f(x_t^{\text{GD}})/\bar{L} = \underbrace{x_t - \nabla f(x_t)/\bar{L}}_{y_{t+1}} = x_t^{\text{NAG}} - \nabla f(x_t^{\text{NAG}})/\bar{L} = y_{t+1}^{\text{NAG}}.$$

In turn, x_{t+1}^{NAG} reduces to an affine combination of descent steps y_{t+1} and y_t of Algorithm 1:

$$x_{t+1}^{\text{NAG}} = (1+\theta_t)y_{t+1}^{\text{NAG}} - \theta_t y_t^{\text{NAG}} = (1+\theta_t)y_{t+1} - \theta_t y_t.$$

Hence, for all $t \geq 0$ such that $m_t \geq m$, x_{t+1} is a convex combination of $x_{t+1}^{\text{GD}} = y_{t+1}^{\text{GD}} = y_{t+1}$ and x_{t+1}^{NAG}, as in

$$\begin{aligned}
x_{t+1} = (1+\beta_t)y_{t+1} - \beta_t y_t &= \Big(1 + \theta_t\frac{\beta_t}{\theta_t} \pm \frac{\beta_t}{\theta_t}\Big)y_{t+1} - \theta_t\frac{\beta_t}{\theta_t}y_t \\
&= \Big(1 - \frac{\beta_t}{\theta_t}\Big)y_{t+1} + \frac{\beta_t}{\theta_t}((1+\theta_t)y_{t+1} - \theta_t y_t) \\
&= \Big(1 - \frac{\beta_t}{\theta_t}\Big)y_{t+1}^{\text{GD}} + \frac{\beta_t}{\theta_t}((1+\theta_t)y_{t+1}^{\text{NAG}} - \theta_t y_t^{\text{NAG}}) \\
&= \bar{\alpha}_t x_{t+1}^{\text{GD}} + \alpha_t x_{t+1}^{\text{NAG}},
\end{aligned}$$

where, as defined in (41), the coefficients defining the convex combination are given by

$$\alpha_t = \beta_t/\theta \in [0,1], \qquad \bar{\alpha}_t = 1 - \alpha_t \in [0,1].$$

Likewise, $y_{t+1}^{\text{GD}} = y_{t+1}^{\text{NAG}} = y_{t+1}$ trivially implies $y_{t+1} = \bar{\alpha}_t y_{t+1}^{\text{GD}} + \alpha_t y_{t+1}^{\text{NAG}}$ so, in fact, the entire iterate of Algorithm 1 can be expressed as a convex combination of GD and NAG iterates, as in

$$s_{t+1} = \bar{\alpha}_t s_{t+1}^{\text{GD}} + \alpha_t s_{t+1}^{\text{NAG}},$$

where s_{t+1}, s_{t+1}^{GD} and s_{t+1}^{NAG} comprise the iterates of Algorithm 1, GD and NAG:

$$\begin{aligned}
s_{t+1} &= (x_{t+1}, y_{t+1}), \\
s_{t+1}^{\text{GD}} &= (x_{t+1}^{\text{GD}}, y_{t+1}^{\text{GD}}) = (y_{t+1}, y_{t+1}) \\
s_{t+1}^{\text{NAG}} &= (x_{t+1}^{\text{NAG}}, y_{t+1}^{\text{NAG}}) = (x_{t+1}^{\text{NAG}}, y_{t+1}).
\end{aligned}$$

Hence, since W is convex[2], we can bound $W(s_{t+1})$ solely in terms of GD and NAG iterations:

$$W(s_{t+1}) \leq \bar{\alpha}_t W(s_{t+1}^{\text{GD}}) + \alpha_t W(s_{t+1}^{\text{NAG}}).$$

In turn, since $W(s_t) = \bar{\alpha}_t W(s_t) + \alpha_t W(s_t)$ holds trivially, it follows that

$$\begin{aligned}
(1+\delta^{\text{GD}})W(s_{t+1}) - W(s_t) = \bar{\alpha}_t((1+\delta^{\text{GD}})W(s_{t+1}^{\text{GD}}) - W(s_t)) \\
+ \alpha_t\Big((1+\delta^{\text{NAG}})W(s_{t+1}^{\text{NAG}}) - W(s_t)\Big) \\
+ \alpha_t(\delta^{\text{GD}} - \delta^{\text{NAG}})W(s_{t+1}^{\text{NAG}}).
\end{aligned}$$

Now, since $y_{t+1}^{\text{NAG}} = y_{t+1}$ and $x_t^{\text{NAG}} = x_t$, then applying (11) with $y = y_{t+1}$ and $x = x_t$ we get

$$f(y_{t+1}^{\text{NAG}}) - f(x_t^{\text{NAG}}) = f(y_{t+1}) - f(x_t) \leq -\|g_t\|^2/2\bar{L} = -\|g_t^{\text{NAG}}\|^2/\bar{L}.$$

Moreover, by assumption, m_0 is initialized in $[m, L]$ and $\bar{L} \geq m_0$. Hence, the assumption that m_{t+1} is computed by Algorithm 1 from m_t implies that

$$\bar{L} \geq m_0 \geq \cdots \geq m_t \geq m_{t+1} > 0. \tag{44}$$

[2]By assumption $f \in \mathcal{F}(L,m)$ is convex, thus so is $\tilde{f}$. Moreover, the affine transformation that defines $z_t^\star$ composed with the 2-norm yields a convex function. That is, V_t^{GD} is the positive sum of convex functions and is therefore convex.

That is, m_t is positive nonincreasing. Therefore, Lemma A.4 applies because Lemma A.4 imposes no restrictions on neither y_t^{NAG} nor x_t^{NAG}. So, letting

$$s_t^{\text{NAG}} = (x_t^{\text{NAG}}, y_t^{\text{NAG}}) = (x_t, y_t) =: s_t,$$

then Lemma A.4 combined with both the fact that $\delta^{\text{NAG}} \geq \delta^{\text{GD}}$ and that $\alpha_t > 0$, imply

$$\alpha_t \Big((1+\delta^{\text{NAG}}) W(s_{t+1}^{\text{NAG}}) - W(s_t) + (\delta^{\text{GD}} - \delta^{\text{NAG}}) W(s_{t+1}^{\text{NAG}}) \Big) \leq 0. \tag{45}$$

The natural next move would be to address $(1+\delta^{\text{GD}})W(s_{t+1}^{\text{GD}}) - W(s_t)$ in an analogous way. The caveat, however, is that although Lemma A.4 applies to NAG iterations with arbitrary x_t^{NAG} and y_t^{NAG}, the same is not true of Lemma A.3. That is, Lemma A.3 applies to consecutive GD iterations, requiring that $y_t^{\text{GD}} = x_t^{\text{GD}}$. Hence, to be able to apply Lemma A.3, we add $\mp W(s_t^{\text{GD}})$ to the difference involving W, using a GD iteration s_t^{GD} such that $y_t^{\text{GD}} = x_t^{\text{GD}}$. That is, we define a fictitious GD iteration s_t^{GD} "backwards" from x_t using the points that we already defined as $y_t^{\text{GD}} = x_t^{\text{GD}}$, as in

$$s_t^{\text{GD}} = (x_t^{\text{GD}}, y_t^{\text{GD}}) = (x_t, x_t).$$

Although s_t^{GD} need not equal s_t, we do have that $y_{t+1}^{\text{GD}} = y_{t+1}$ and $x_t^{\text{GD}} = x_t$, thus

$$f(y_{t+1}^{\text{GD}}) - f(x_t^{\text{GD}}) = f(y_{t+1}) - f(x_t) \leq -\|g_t\|^2/2\bar{L} = -\|g_t^{\text{GD}}\|^2/\bar{L}.$$

Therefore, since $\bar{L} > 0$ by (44), Lemma A.3 applies with the newly defined GD iteration s_t^{GD}, and implies that

$$(1+\delta^{\text{GD}})W(s_{t+1}^{\text{GD}}) - W(s_t^{\text{GD}}) \leq -\|g_t^{\text{GD}}\|^2/2\bar{L} = -\|g_t\|^2/2\bar{L}. \tag{46}$$

Moreover, $y_t^{\text{GD}} = x_t^{\text{GD}} = x_t$ implies that $x_t^{\text{GD},y} = 0$, thus

$$z_t^{\text{GD},\star} = x_t^{\text{GD},\star} + \sqrt{\bar{\kappa}} x_t^{\text{GD},y} = x_t^{\text{GD},\star} = x_t^\star \qquad \text{and} \qquad f(y_t^{\text{GD}}) = f(x_t),$$

and it follows that

$$W(s_t^{\text{GD}}) - W(s_t) = f(x_t) - f(y_t) + (m/2)(\|x_t^\star\|^2 - \|x_t^\star + \sqrt{\bar{\kappa}} x_t^y\|^2). \tag{47}$$

Applying (3) with $x = x_t$ and $y = y_t$, then using the simple fact that a cross product can be bounded by a sum of quadratics, as in $2\langle g_t, x_t^y\rangle \leq \bar{L}^{-1}\|g_t\|^2 + \bar{L}\|x_t^y\|^2$, we obtain

$$f(x_t) - f(y_t) \leq \langle g_t, x_t^y\rangle - (m/2)\|x_t^y\|^2 \leq (1/2\bar{L})\|g_t\|^2 + \frac{\bar{L}-m}{2}\|x_t^y\|^2.$$

Hence, expanding $\|x_t^\star + \sqrt{\bar{\kappa}} x_t^y\|^2$ on (47) and then using the above inequality, we get

$$\begin{aligned} W(s_t^{\text{GD}}) - W(s_t) &\leq (1/2\bar{L})\|g_t\|^2 + \frac{\bar{L}-m}{2}\|x_t^y\|^2 + (m/2)(-2\sqrt{\bar{\kappa}}\langle x_t^y, x_t^\star\rangle - \bar{\kappa}\|x_t^y\|^2) \\ &= (1/2\bar{L})\|g_t\|^2 - (m/2)\|x_t^y\|^2 - \sqrt{\bar{L}m}\langle x_t^y, x_t^\star\rangle, \end{aligned} \tag{48}$$

where $m\sqrt{\bar{\kappa}} = \sqrt{\bar{L}m}$ follows directly from (17). In turn, if we put (46) and (48) together, then

$$(1+\delta^{\text{GD}})W(s_{t+1}^{\text{GD}}) \mp W(s_t^{\text{GD}}) - W(s_t) \leq -\sqrt{\bar{L}m}\langle x_t^y, x_t^\star\rangle. \tag{49}$$

Therefore, since $\delta^{\text{GD}} \leq \delta^{\text{NAG}}$, combining (45) and (49), we obtain

$$\begin{aligned} (1+\delta^{\text{GD}})W(s_{t+1}) - W(s_t) &\leq \bar{\alpha}_t((1+\delta^{\text{GD}})W(s_{t+1}^{\text{GD}}) \mp W(s_t^{\text{GD}}) - W(s_t)) \\ &\quad + \alpha_t((1+\delta^{\text{NAG}})W(s_{t+1}^{\text{NAG}}) - W(s_t)) \\ &\quad + \alpha_t(\delta^{\text{GD}} - \delta^{\text{NAG}})W(s_{t+1}^{\text{NAG}}) \\ &\leq -\bar{\alpha}_t\sqrt{\bar{L}m}\langle x_t^y, x_t^\star\rangle. \end{aligned} \tag{50}$$

Next, we address the U term in $(1+\delta^{\text{GD}})V_{t+1}^{\text{GD}}(s_{t+1}) - V_{t+1}^{\text{GD}}(s_t)$. By Lemma A.5, we have that

$$\frac{\bar{\alpha}_t}{\sqrt{\bar{\kappa}}}((1+\delta^{\text{GD}})U(s_{t+1}) - U(s_t)) \leq \frac{\bar{\alpha}_t}{\sqrt{\bar{\kappa}}}\bar{L}\langle x_t^y, x_t^\star\rangle = \bar{\alpha}_t\sqrt{\bar{L}m}\langle x_t^y, x_t^\star\rangle, \tag{51}$$

where the identity on the right-hand side follows from (17). Combining (50) with (51), we get

$$(1+\delta^{\mathrm{GD}})V_{t+1}^{\mathrm{GD}}(s_{t+1}) \le V_{t+1}^{\mathrm{GD}}(s_t). \tag{52}$$

To continue the proof, we bound $V_{t+1}^{\mathrm{GD}}(s_t)$ in terms of $V_t^{\mathrm{GD}}(s_t)$. By (44), m_t are monotonic nonincreasing and so is $\bar{\alpha}_t$, by Lemma A.6. Thus, $\bar{\alpha}_t/\sqrt{\bar{\kappa}}$ is nonincreasing. Hence, using the definition of V_t^{GD}, we obtain

$$V_{t+1}^{\mathrm{GD}} = W + \frac{\bar{\alpha}_t}{\sqrt{\bar{\kappa}}}U \le W + \frac{\bar{\alpha}_{t-1}}{\sqrt{\bar{\kappa}}}U = V_t^{\mathrm{GD}}. \tag{53}$$

In turn, combining (52) and (53), we get

$$V_{t+1}^{\mathrm{GD}}(s_{t+1}) \le (1+\delta^{\mathrm{GD}})^{-1}V_t^{\mathrm{GD}}(s_t). \tag{54}$$

To conclude the proof, we employ (54) in an inductive argument. To establish the base case of the inductive argument, we apply (11) with $y = y_0$ and $x = x^\star$, obtaining $\tilde{f}(y_0) \le (\bar{L}/2)\|y_0^\star\|^2$. Then, using the assumptions that $y_0 = x_0$ and that $\bar{L}$, we have $z_0^\star = x_0^\star = y_0^\star$, and it follows that

$$\begin{aligned} W(s_0) &= \tilde{f}(y_0) + (m/2)\|x_0^\star\|^2 \le \frac{\bar{L}+m}{2}\|x_0^\star\|^2 \le \bar{L}\|x_0^\star\|^2, \\ U(s_0) &= \tilde{f}(y_0) + (\bar{L}/2)\|y_0^\star\|^2 \le \bar{L}\|x_0^\star\|^2. \end{aligned}$$

Since $\bar{\alpha}_0/\sqrt{\bar{\kappa}} \in [0,1)$, the above inequalities imply

$$V_0^{\mathrm{GD}}(s_0) = W(s_0) + \frac{\bar{\alpha}_0}{\sqrt{\bar{\kappa}}}U(s_0) \le 2\bar{L}\|x_0^\star\|^2. \tag{55}$$

Moreover, $V_1^{\mathrm{GD}} = V_0^{\mathrm{GD}}$. Hence, if $m_1 > m$, then combining (52) with (55) yields

$$V_1^{\mathrm{GD}}(s_1) \le 2\bar{L}(1+\delta^{\mathrm{GD}})^{-1}\|x_0^\star\|^2. \tag{56}$$

Having established the base case (56), suppose that

$$V_{t'+1}^{\mathrm{GD}}(s_{t'+1}) \le 2\bar{L}(1+\delta^{\mathrm{GD}})^{-t'}\|x_0^\star\|^2, \tag{57}$$

holds for all $t' \le t-1$ such that $m_{t'} \ge m$. Then, suppose $m_t \ge m$. By (44), we have that $m_{t'} \ge m$ for all $t' \le t$. Hence, plugging the induction hypothesis (57) with $t' = t-1$ into (54), we obtain

$$V_{t+1}^{\mathrm{GD}}(s_{t+1}) \le 2\bar{L}(1+\delta^{\mathrm{GD}})^{-t}\|x_0^\star\|^2,$$

recovering (57) with $t' = t$. Therefore, by induction, (57) holds for all $t \ge 0$, proving (43). □

Theorem 4.1 follows immediately from Theorem A.7.

Proof of Theorem 4.1. From (41) and (22), we have that

$$f(y_{t+1}) - f(x^\star) \le W(s_{t+1}) \le V_{t+1}^{\mathrm{GD}}(s_{t+1}). \tag{58}$$

Hence, plugging (43) back into the right-hand side of the above inequality, we get

$$f(y_{t+1}) - f(x^\star) \le 2\bar{L}(1+\delta^{\mathrm{GD}})^{-t}\|x_0^\star\|^2,$$

which proves Theorem 4.1. □

A.2 Case 2: $m_t < m$

In the previous section, we analyzed iterations in which $m_t \ge m$. Now, we analyze iterations in which $m_t < m$ and also the transition iteration t such that $m_t \ge m$ and $m_{t+1} < m$. By design of Algorithm 1, m_t is nonincreasing, therefore there is at most one such t.

Iterations in which $m_t < m$

In Theorem A.7, we proved that up to a constant factor, $\tilde{f}(y_t)$ decreases along iterates of Algorithm 1 in which $m_t \geq m$ at least as fast as it does along GD iterates. Now, we prove that up to a constant factor, $\tilde{f}(y_t)$ decreases along iterates of Algorithm 1 in which $m_t < m$ at an accelerated rate. Specifically, if $m_t < m$, then Algorithm 1 achieves the accelerated rate $(1+\delta(\sqrt{\bar{\kappa}_t}))^{-1}$, where

$$\hat{\delta}_t^{\text{NAG}} = \begin{cases} \dfrac{1}{\sqrt{\bar{\kappa}_0}-1}, & t=0, \\ \dfrac{1}{\sqrt{\bar{\kappa}_{t-1}}-1}, & t \geq 1, \end{cases} \tag{59}$$

to work with more concise notation while maintaining it consistent.

Our proof once again consists in an inductive argument built upon descending and ascending bounds on a Lyapunov function. The function we work with this time is V_t^{NAG}, given by

$$V_t^{\text{NAG}}(s_t) = \begin{cases} \tilde{f}(y_0) + (m_0/2)\|w_0^\star\|^2, & t=0, \\ \tilde{f}(y_t) + (m_{t-1}/2)\|w_t^\star\|^2, & t \geq 1, \end{cases} \tag{60}$$

where the pseudo-state $w_t^\star$, analogous to $z_t^\star$, is given by

$$w_t^\star = w_t - x^\star, \qquad w_t = \begin{cases} x_0 + \sqrt{\bar{\kappa}_0}(x_0 - y_0), & t=0, \\ x_t + \sqrt{\bar{\kappa}_{t-1}}(x_t - y_t), & t \geq 1. \end{cases} \tag{61}$$

Once again, we first prove a descending bound and then immediately after prove an ascending one.

Lemma A.8. *Let $f \in \mathcal{F}(L,m)$ and $\bar{L} \geq L$. Given x_t, y_t and $m_t \leq m$, if s_{t+1} is generated by Algorithm 1, then*

$$(1+\hat{\delta}_{t+1}^{\text{NAG}})V_{t+1}^{\text{NAG}}(s_{t+1}) - V_{t+1}^{\text{NAG}}(s_t) \leq 0. \tag{62}$$

Proof. The difference $(1+\hat{\delta}_{t+1}^{\text{NAG}})V_{t+1}^{\text{NAG}}(s_{t+1}) - V_{t+1}^{\text{NAG}}(s_t)$ is the sum of two differences, one involving $\tilde{f}$ and another involving 2-norms. We start with the one involving $\tilde{f}$, splitting it into three further differences:

$$\begin{aligned}(1+\hat{\delta}_{t+1}^{\text{NAG}})\tilde{f}(y_{t+1}) - \tilde{f}(y_t) = &\underbrace{(1+\hat{\delta}_{t+1}^{\text{NAG}})(f(y_{t+1}) - f(x_t))}_{\#1} \\ &+ \underbrace{\hat{\delta}_{t+1}^{\text{NAG}}(f(x_t) - f(x^\star))}_{\#2} \\ &+ \underbrace{f(x_t) - f(y_t)}_{\#3}.\end{aligned}$$

If s_{t+1} is generated by Algorithm 1, then applying (11) with $y = y_{t+1}$ and $x = x_t$, we get

$$(1+\hat{\delta}_{t+1}^{\text{NAG}})(f(y_{t+1}) - f(x_t)) \leq -(1+\hat{\delta}_{t+1}^{\text{NAG}})(1/2\bar{L})\|g_t\|^2. \tag{63}$$

At the same time, the assumption that $m_t < m$ combined with (3) imply that for all x and y

$$f(x) + \langle \nabla f(x), y - x\rangle + (m_t/2)\|x-y\|^2 \leq f(y). \tag{64}$$

Hence, plugging $x = x_t$ and $y = x^\star$ in (64) and using the fact that f is convex, we bound #2 and #3 above as

$$\hat{\delta}_{t+1}^{\text{NAG}}(f(x_t) - f(x^\star)) \leq \hat{\delta}_{t+1}^{\text{NAG}}\langle g_t, x_t^\star\rangle - \hat{\delta}_{t+1}^{\text{NAG}}(m_t/2)\|x_t^\star\|^2, \tag{65}$$

$$f(x_t) - f(y_t) \leq \langle g_t, x_t^y\rangle. \tag{66}$$

Next, we address the 2-norm difference in $(1+\hat{\delta}_{t+1}^{\text{NAG}})V_{t+1}^{\text{NAG}}(s_{t+1}) - V_{t+1}^{\text{NAG}}(s_t)$, by expanding the pseudo-states inside 2-norms. One pseudo-state is $w_{t+1}^\star$ which, using (20) and (61), we express as

$$\begin{aligned} w_{t+1}^\star &= x_{t+1} + \sqrt{\bar{\kappa}_t}(x_{t+1} - y_{t+1}) - x^\star \\ &= y_{t+1} + \beta_t(y_{t+1} - y_t) + \sqrt{\bar{\kappa}_t}\beta_t(y_{t+1} - y_t) - x^\star \\ &= -(1+\beta_t(1+\sqrt{\bar{\kappa}_t}))g_t/\bar{L} + \beta_t(1+\sqrt{\bar{\kappa}_t})x_t^y + x_t^\star \\ &= -\sqrt{\bar{\kappa}_t}g_t/\bar{L} + (\sqrt{\bar{\kappa}_t}-1)x_t^y + x_t^\star. \end{aligned}$$

After expanding $w_{t+1}^\star$ inside the 2-norm, we use the following identities after colons to simplify the coefficients of terms before colons:

$$\begin{aligned}
\|g_t\|^2 &: && (\bar\kappa_t/\bar L^2)(m_t/2) = 1/2\bar L,\\
\langle g_t, x_t^y\rangle &: && m_t(1+\hat\delta_{t+1}^{\text{NAG}})\sqrt{\bar\kappa_t}(\sqrt{\bar\kappa_t}-1)/\bar L = 1,\\
\langle g_t, x_t^\star\rangle &: && m_t(1+\hat\delta_{t+1}^{\text{NAG}})\sqrt{\bar\kappa_t}/\bar L = \hat\delta_{t+1}^{\text{NAG}},\\
\|x_t^y\|^2 &: && (1+\hat\delta_{t+1}^{\text{NAG}})(\sqrt{\bar\kappa_t}-1)^2 = \sqrt{\bar\kappa_t}(\sqrt{\bar\kappa_t}-1),\\
\langle x_t^y, x_t^\star\rangle &: && (1+\hat\delta_{t+1}^{\text{NAG}})(\sqrt{\bar\kappa_t}-1) = \sqrt{\bar\kappa_t}.
\end{aligned}$$

Thus, the 2-norm difference in $(1+\hat\delta_{t+1}^{\text{NAG}})V_{t+1}^{\text{NAG}}(s_{t+1}) - V_{t+1}^{\text{NAG}}(s_t)$ reduces to

$$\begin{aligned}
&(1+\hat\delta_{t+1}^{\text{NAG}})(m_t/2)\|w_{t+1}^\star\|^2 - (m_t/2)\|x_t^\star + \sqrt{\bar\kappa_t}x_t^y\|^2\\
&= \frac{1+\hat\delta_{t+1}^{\text{NAG}}}{2\bar L}\|g_t\|^2 - \langle g_t, x_t^y\rangle - \hat\delta_{t+1}^{\text{NAG}}\langle g_t, x_t^\star\rangle + (m_t/2)\sqrt{\bar\kappa_t}(\sqrt{\bar\kappa_t}-1)\|x_t^y\|^2\\
&\quad + (m_t/2)(2\sqrt{\bar\kappa_t}\langle x_t^y, x_t^\star\rangle + (1+\hat\delta_{t+1}^{\text{NAG}})\|x_t^\star\|^2) - (m_t/2)(\bar\kappa_t\|x_t^y\|^2 + 2\sqrt{\bar\kappa_t}\langle x_t^y, x_t^\star\rangle + \|x_t^\star\|^2)\\
&= \frac{1+\hat\delta_{t+1}^{\text{NAG}}}{2\bar L}\|g_t\|^2 - \langle g_t, x_t^y\rangle - \hat\delta_{t+1}^{\text{NAG}}\langle g_t, x_t^\star\rangle - (m_t/2)\sqrt{\bar\kappa_t}\|x_t^y\|^2 + \hat\delta_{t+1}^{\text{NAG}}(m_t/2)\|x_t^\star\|^2. \qquad (67)
\end{aligned}$$

Finally, combining (63) and (65) to (67), cancelling terms and then using the assumption that $m_t > 0$, we obtain

$$(1+\hat\delta_{t+1}^{\text{NAG}})V_{t+1}^{\text{NAG}}(s_{t+1}) - V_{t+1}^{\text{NAG}}(s_t) \le -(m_t/2)\sqrt{\bar\kappa_t}\|x_t^y\|^2 \le 0.$$

□

Lemma A.9. *Let $f \in \mathcal{F}(L,m)$ and $\bar L \ge L$. Given s_t, if $m_{t-1} \ge m_t$ and $m_t < m$, then*

$$V_{t+1}^{\text{NAG}}(s_t) \le \frac{\bar\kappa_t^2}{\bar\kappa_{t-1}^2}V_t^{\text{NAG}}(s_t). \qquad (68)$$

Proof. We divide the analysis in two cases, each representing a possible sign of $\langle x_t^y, x_t^\star\rangle$. Building upon the sign underpinning each case, we bound

$$\|x_t^\star + \sqrt{\bar\kappa_t}x_t^y\|^2 - \|w_t^\star\|^2 = 2(\sqrt{\bar\kappa_t} - \sqrt{\bar\kappa_{t-1}})\langle x_t^\star, x_t^y\rangle + (\bar\kappa_t - \bar\kappa_{t-1})\|x_t^y\|^2. \qquad (69)$$

Assuming $m_t \ge m_{t-1}$, bounds on (69) translate into bounds on $V_{t+1}^{\text{NAG}}(s_t) - V_t^{\text{NAG}}(s_t)$, since

$$\begin{aligned}
V_{t+1}^{\text{NAG}}(s_t) - V_t^{\text{NAG}}(s_t) &= (m_t/2)\|x_t^\star + \sqrt{\bar\kappa_t}x_t^y\|^2 - (m_{t-1}/2)\|w_t^\star\|^2\\
&\le (m_t/2)(\|x_t^\star + \sqrt{\bar\kappa_t}x_t^y\|^2 - \|w_t^\star\|^2). \qquad (70)
\end{aligned}$$

Then, to prove (68), we express bounds on (70) in terms of V_{t+1}^{NAG} and V_t^{NAG}.

First, suppose $\langle x_t^y, x_t^\star\rangle \ge 0$. Since $m_{t-1} \ge m_t$, then $\sqrt{\bar\kappa_{t-1}}/\sqrt{\bar\kappa_t} \le 1$, so that

$$\sqrt{\bar\kappa_t} - \sqrt{\bar\kappa_{t-1}} \le \frac{\bar\kappa_t}{\sqrt{\bar\kappa_t}} - \sqrt{\bar\kappa_{t-1}}\frac{\sqrt{\bar\kappa_{t-1}}}{\sqrt{\bar\kappa_t}} = \frac{\bar\kappa_t - \bar\kappa_{t-1}}{\sqrt{\bar\kappa_t}}.$$

Hence, applying the above inequality to (69) and then adding a nonnegative $\|x_t^\star\|^2$ term to it, we get

$$\begin{aligned}
\|x_t^\star + \sqrt{\bar\kappa_t}x_t^y\|^2 - \|w_t^\star\|^2 &\le 2\frac{\bar\kappa_t - \bar\kappa_{t-1}}{\sqrt{\bar\kappa_t}}\langle x_t^\star, x_t^y\rangle + (\bar\kappa_t - \bar\kappa_{t-1})\|x_t^y\|^2 + \frac{\bar\kappa_t - \bar\kappa_{t-1}}{\bar\kappa_t}\|x_t^\star\|^2\\
&= \frac{\bar\kappa_t - \bar\kappa_{t-1}}{\bar\kappa_t}\|x_t^\star + \sqrt{\bar\kappa_t}x_t^y\|^2. \qquad (71)
\end{aligned}$$

Plugging (71) back into (70) yields

$$V_{t+1}^{\text{NAG}}(s_t) - V_t^{\text{NAG}}(s_t) \le \frac{\bar\kappa_t - \bar\kappa_{t-1}}{\bar\kappa_t}\frac{m_t}{2}\|x_t^\star + \sqrt{\bar\kappa_t}x_t^y\|^2 \le \frac{\bar\kappa_t - \bar\kappa_{t-1}}{\bar\kappa_t}V_{t+1}^{\text{NAG}}(s_t), \qquad (72)$$

where the last inequality follows from the definition of V_t^{NAG}, (60), as $\tilde{f} \geq 0$ implies

$$V_{t+1}^{\text{NAG}}(s_t) = \tilde{f}(y_t) + (m_t/2)\|x_t^\star + \sqrt{\bar{\kappa}_t} x_t^y\|^2 \geq (m_t/2)\|x_t^\star + \sqrt{\bar{\kappa}_t} x_t^y\|^2. \tag{73}$$

Thus, rearranging terms in (72) and then multiplying both sides by $\bar{\kappa}_t/\bar{\kappa}_{t-1}$, we obtain

$$V_{t+1}^{\text{NAG}}(s_t) \leq \frac{\bar{\kappa}_t}{\bar{\kappa}_{t-1}} V_t^{\text{NAG}}(s_t) \leq \frac{\bar{\kappa}_t^2}{\bar{\kappa}_{t-1}^2} V_t^{\text{NAG}}(s_t),$$

where the second inequality holds because $\bar{\kappa}_t/\bar{\kappa}_{t-1} \geq 1$.

Now, suppose $\langle x_t^y, x_t^\star \rangle < 0$. We start by bounding the gap (69). But given the negative sign of $\langle x_t^y, x_t^\star \rangle$ term, we bound the $\|x_t^y\|^2$ term instead. To this end, we first use the assumption that $\langle x_t^y, x_t^\star \rangle < 0$ to establish that

$$\|y_t^\star\|^2 = \|y_t^\star \mp x_t^\star\|^2 = \|x_t^\star - x_t^y\|^2 = \|x_t^\star\|^2 - 2\langle x_t^\star, x_t^y \rangle + \|x_t^y\|^2 \geq \|x_t^\star\|^2. \tag{74}$$

To use the above inequality on (69), first we rewrite it more conveniently as

$$\begin{aligned}
\|x_t^\star + \sqrt{\bar{\kappa}_t} x_t^y\|^2 - \|w_t^\star\|^2 &= 2\frac{\sqrt{\bar{\kappa}_t} - \sqrt{\bar{\kappa}_{t-1}}}{\sqrt{\bar{\kappa}_t}} \langle x_t^\star, \sqrt{\bar{\kappa}_t} x_t^y \rangle + \sqrt{\bar{\kappa}_t}(\sqrt{\bar{\kappa}_t} - \sqrt{\bar{\kappa}_{t-1}})\|x_t^y\|^2 \\
&\quad + \sqrt{\bar{\kappa}_{t-1}}(\sqrt{\bar{\kappa}_t} - \sqrt{\bar{\kappa}_{t-1}})\|x_t^y\|^2 \pm \frac{\sqrt{\bar{\kappa}_t} - \sqrt{\bar{\kappa}_{t-1}}}{\sqrt{\bar{\kappa}_t}}\|x_t^\star\|^2 \\
&= \frac{\sqrt{\bar{\kappa}_t} - \sqrt{\bar{\kappa}_{t-1}}}{\sqrt{\bar{\kappa}_t}}\|x_t^\star + \sqrt{\bar{\kappa}_t} x_t^y\|^2 + \sqrt{\bar{\kappa}_{t-1}}(\sqrt{\bar{\kappa}_t} - \sqrt{\bar{\kappa}_{t-1}})\|x_t^y\|^2 \\
&\quad - \frac{\sqrt{\bar{\kappa}_t} - \sqrt{\bar{\kappa}_{t-1}}}{\sqrt{\bar{\kappa}_t}}\|x_t^\star\|^2.
\end{aligned} \tag{75}$$

Then, we establish an elementary inequality for 2-norms. If $a, b \in \mathbb{R}^d$ and $c \neq 0$, then

$$(1/c^2)\|a\|^2 + 2\langle a, b \rangle + c^2\|b\|^2 = \|a/c + bc\|^2 \geq 0,$$

so that $-2\langle a, b \rangle \leq (1/c^2)\|a\|^2 + c^2\|b\|^2$, which implies

$$\|a - b\|^2 = \|a\|^2 - 2\langle a, b \rangle + \|b\|^2 \leq (1 + 1/c^2)\|a\|^2 + (1 + c^2)\|b\|^2. \tag{76}$$

Applying (76) with $a = w_t^\star$, $b = x_t^\star$ and $c^2 = \sqrt{\bar{\kappa}_{t-1}}/\sqrt{\bar{\kappa}_t}$ to bound the $\|x_t^y\|^2$ term on (75), we obtain

$$\begin{aligned}
\sqrt{\bar{\kappa}_{t-1}}(\sqrt{\bar{\kappa}_t} - \sqrt{\bar{\kappa}_{t-1}})\|x_t^y\|^2 &= \sqrt{\bar{\kappa}_{t-1}}(\sqrt{\bar{\kappa}_t} - \sqrt{\bar{\kappa}_{t-1}})\|x_t^y \pm x_t^\star/\sqrt{\bar{\kappa}_{t-1}}\|^2 \\
&= \frac{\sqrt{\bar{\kappa}_t} - \sqrt{\bar{\kappa}_{t-1}}}{\sqrt{\bar{\kappa}_{t-1}}}\|w_t^\star - x_t^\star\|^2 \\
&\leq \frac{\sqrt{\bar{\kappa}_t} - \sqrt{\bar{\kappa}_{t-1}}}{\sqrt{\bar{\kappa}_{t-1}}}\Big(1 + \frac{\sqrt{\bar{\kappa}_t}}{\sqrt{\bar{\kappa}_{t-1}}}\Big)\|w_t^\star\|^2 \\
&\quad + \frac{\sqrt{\bar{\kappa}_t} - \sqrt{\bar{\kappa}_{t-1}}}{\sqrt{\bar{\kappa}_{t-1}}}\Big(1 + \frac{\sqrt{\bar{\kappa}_{t-1}}}{\sqrt{\bar{\kappa}_t}}\Big)\|x_t^\star\|^2 \\
&= \frac{\bar{\kappa}_t - \bar{\kappa}_{t-1}}{\bar{\kappa}_{t-1}}\|w_t^\star\|^2 + \frac{\sqrt{\bar{\kappa}_t} - \sqrt{\bar{\kappa}_{t-1}}}{\sqrt{\bar{\kappa}_t}}\frac{\sqrt{\bar{\kappa}_t} + \sqrt{\bar{\kappa}_{t-1}}}{\sqrt{\bar{\kappa}_{t-1}}}\|x_t^\star\|^2.
\end{aligned} \tag{77}$$

Plugging (77) back into (75) and then using (74), we get

$$\begin{aligned}
\|x_t^\star + \sqrt{\bar{\kappa}_t} x_t^y\|^2 - \|w_t^\star\|^2 &\leq \frac{\sqrt{\bar{\kappa}_t} - \sqrt{\bar{\kappa}_{t-1}}}{\sqrt{\bar{\kappa}_t}}\|x_t^\star + \sqrt{\bar{\kappa}_t} x_t^y\|^2 + \frac{\bar{\kappa}_t - \bar{\kappa}_{t-1}}{\bar{\kappa}_{t-1}}\|w_t^\star\|^2 \\
&\quad + \frac{\sqrt{\bar{\kappa}_t} - \sqrt{\bar{\kappa}_{t-1}}}{\sqrt{\bar{\kappa}_t}}\Big(\frac{\sqrt{\bar{\kappa}_t} + \sqrt{\bar{\kappa}_{t-1}}}{\sqrt{\bar{\kappa}_{t-1}}} - 1\Big)\|x_t^\star\|^2 \\
&\leq \frac{\sqrt{\bar{\kappa}_t} - \sqrt{\bar{\kappa}_{t-1}}}{\sqrt{\bar{\kappa}_t}}\|x_t^\star + \sqrt{\bar{\kappa}_t} x_t^y\|^2 + \frac{\bar{\kappa}_t - \bar{\kappa}_{t-1}}{\bar{\kappa}_{t-1}}\|w_t^\star\|^2 \\
&\quad + \frac{\sqrt{\bar{\kappa}_t} - \sqrt{\bar{\kappa}_{t-1}}}{\sqrt{\bar{\kappa}_{t-1}}}\|y_t^\star\|^2.
\end{aligned} \tag{78}$$

Then, plugging (78) back into (70) and using the assumptions that $m_{t-1} \geq m_t$ and $m_t < m$ yields

$$\begin{aligned} V_{t+1}^{\text{NAG}}(s_t) - V_t^{\text{NAG}}(s_t) &\leq \frac{m_t}{2}\frac{\sqrt{\bar{\kappa}_t} - \sqrt{\bar{\kappa}_{t-1}}}{\sqrt{\bar{\kappa}_t}}\|x_t^\star + \sqrt{\bar{\kappa}_t}x_t^y\|^2 + \frac{m_t}{2}\frac{\bar{\kappa}_t - \bar{\kappa}_{t-1}}{\bar{\kappa}_{t-1}}\|w_t^\star\|^2 \\ &\quad + \frac{m_t}{2}\frac{\sqrt{\bar{\kappa}_t} - \sqrt{\bar{\kappa}_{t-1}}}{\sqrt{\bar{\kappa}_{t-1}}}\|y_t^\star\|^2 \\ &\leq \frac{\sqrt{\bar{\kappa}_t} - \sqrt{\bar{\kappa}_{t-1}}}{\sqrt{\bar{\kappa}_t}}\frac{m_t}{2}\|x_t^\star + \sqrt{\bar{\kappa}_t}x_t^y\|^2 + \frac{\bar{\kappa}_t - \bar{\kappa}_{t-1}}{\bar{\kappa}_{t-1}}\frac{m_{t-1}}{2}\|w_t^\star\|^2 \\ &\quad + \frac{\sqrt{\bar{\kappa}_t} - \sqrt{\bar{\kappa}_{t-1}}}{\sqrt{\bar{\kappa}_{t-1}}}\frac{m}{2}\|y_t^\star\|^2. \end{aligned} \tag{79}$$

Now, as in (73), applying $\tilde{f} \geq 0$ to the definition of V_t^{NAG}, we get

$$V_t^{\text{NAG}}(s_t) = \tilde{f}(y_t) + (m_{t-1}/2)\|w_t^\star\|^2 \geq (m_{t-1}/2)\|w_t^\star\|^2. \tag{80}$$

In the same vein, applying (3) with $x = x^\star$ and $y = y_t$ to the definition of V_t^{NAG}, we obtain

$$V_t^{\text{NAG}}(s_t) = \tilde{f}(y_t) + (m_{t-1}/2)\|w_t^\star\|^2 \geq (m/2)\|y_t^\star\|^2. \tag{81}$$

Plugging in (73), (80) and (81) back into (79), and then moving all $V_{t+1}^{\text{acc}}(s_t)$ terms to the left-hand side and all $V_t^{\text{NAG}}(s_t)$ to the right-hand side, we obtain

$$\frac{\sqrt{\bar{\kappa}_{t-1}}}{\sqrt{\bar{\kappa}_t}}V_{t+1}^{\text{NAG}}(s_t) \leq \Big(\frac{\bar{\kappa}_t}{\bar{\kappa}_{t-1}} + \frac{\sqrt{\bar{\kappa}_t} - \sqrt{\bar{\kappa}_{t-1}}}{\sqrt{\bar{\kappa}_{t-1}}}\Big)V_t^{\text{NAG}}(s_t) \tag{82}$$

Multiplying both sides of (82) by $\sqrt{\bar{\kappa}_t}/\sqrt{\bar{\kappa}_{t-1}}$, and then using the fact that $\sqrt{\bar{\kappa}_t} \geq \sqrt{\bar{\kappa}_{t-1}}$ yields

$$V_{t+1}^{\text{NAG}}(s_t) \leq \frac{\sqrt{\bar{\kappa}_t}}{\sqrt{\bar{\kappa}_{t-1}}}\Big(\frac{\bar{\kappa}_t}{\bar{\kappa}_{t-1}} + \frac{\sqrt{\bar{\kappa}_t} - \sqrt{\bar{\kappa}_{t-1}}}{\sqrt{\bar{\kappa}_{t-1}}}\Big)V_t^{\text{NAG}}(s_t) \leq \frac{\bar{\kappa}_t^2}{\bar{\kappa}_{t-1}^2}V_t^{\text{NAG}}(s_t),$$

where the last inequality above holds because $\bar{\kappa}_t \geq \bar{\kappa}_{t-1}$ implies the following equivalences hold:

$$\begin{aligned} \frac{\bar{\kappa}_t}{\bar{\kappa}_{t-1}} + \frac{\sqrt{\bar{\kappa}_t} - \sqrt{\bar{\kappa}_{t-1}}}{\sqrt{\bar{\kappa}_{t-1}}} \leq \frac{\bar{\kappa}_t^{3/2}}{\bar{\kappa}_{t-1}^{3/2}} &\iff \sqrt{\bar{\kappa}_{t-1}}\bar{\kappa}_t + \bar{\kappa}_{t-1}(\sqrt{\bar{\kappa}_t} - \sqrt{\bar{\kappa}_{t-1}}) \leq \bar{\kappa}_t^{3/2}, \\ &\iff \bar{\kappa}_{t-1}(\sqrt{\bar{\kappa}_t} - \sqrt{\bar{\kappa}_{t-1}}) \leq \bar{\kappa}_t(\sqrt{\bar{\kappa}_t} - \sqrt{\bar{\kappa}_{t-1}}). \end{aligned}$$

Therefore, both when $\langle x_t^\star, x_t^y\rangle \geq 0$ and when $\langle x_t^\star, x_t^y\rangle < 0$, the inequality

$$V_{t+1}^{\text{NAG}}(s_t) \leq \frac{\bar{\kappa}_t^2}{\bar{\kappa}_{t-1}^2}V_t^{\text{NAG}}(s_t)$$

holds generically for all s_t, proving (68). □

Theorem A.10. *Let $f \in \mathcal{F}(L, m)$ and $\bar{L} \geq L$. If Algorithm 1 receives $\bar{L}$ as input and $m_T \leq m$ for some T, then for all $t > T$ the iterates s_t of Algorithm 1 satisfy*

$$f(y_{t+1}) - f(x^\star) \leq \frac{\bar{\kappa}_t^2}{\bar{\kappa}_{T-1}^2}\left(\prod_{i=T}^{t+1}(1 + \hat{\delta}_i^{\text{NAG}})^{-1}\right)V_T^{\text{NAG}}(s_T). \tag{83}$$

Proof. Since Algorithm 1 generates m_t, they are nonincreasing. Hence, $m_t \leq L$ for all $t \geq 0$ since $m_0 \leq L$ by assumption. Moreover, if $m_T \leq m$, then $m_t < m$ for all $t \geq T$. Therefore, if $m_T \leq m$, then Lemmas A.8 and A.9 hold for all $t \geq T$. So, plugging (68) into (62), we have that for all $t \geq T$

$$(1 + \hat{\delta}_i^{\text{NAG}})V_{t+1}^{\text{NAG}}(s_{t+1}) - \frac{\bar{\kappa}_t^2}{\bar{\kappa}_{t-1}^2}V_t^{\text{NAG}}(s_t) \leq (1 + \hat{\delta}_i^{\text{NAG}})V_{t+1}^{\text{NAG}}(s_{t+1}) - V_{t+1}^{\text{NAG}}(s_t) \leq 0. \tag{84}$$

Rearranging terms, we obtain

$$V_{t+1}^{\text{NAG}}(s_{t+1}) \leq (1 + \hat{\delta}_i^{\text{NAG}})^{-1}\frac{\bar{\kappa}_t^2}{\bar{\kappa}_{t-1}^2}V_t^{\text{NAG}}(s_t). \tag{85}$$

Applying the above inductively, we conclude that for all $t > T$

$$f(y_{t+1}) - f(x^\star) \le V_{t+1}^{\text{NAG}}(s_{t+1}) \le \frac{\bar{\kappa}_t^2}{\bar{\kappa}_{T-1}^2}\left(\prod_{i=T}^{t+1}(1+\hat{\delta}_i^{\text{NAG}})^{-1}\right)V_T^{\text{NAG}}(s_T),$$

where the first inequality follows directly from (60), the definition of V_t^{NAG}, since

$$V_{t+1}^{\text{NAG}}(s_{t+1}) = \tilde{f}(y_{t+1}) + (m_t/2)\|w_{t+1}^\star\|^2 \ge \tilde{f}(y_{t+1}).$$

□

We now have completed the analysis of iterations in which $m_t < m$. To prove Theorem A.1, it remains to analyze the transition to this kind of iteration from iterations in which $m_t \ge m$.

From iterations in which $m_t \ge m$ to iterations in which $m_t < m$

In Appendix A.1 and so far in this subsection, we have separately analyzed iterations in which $m_t \ge m$ and iterations in which $m_t < m$. To prove Theorem A.1, we must join the analyses, allowing for a transition from the first kind of iteration to the second kind. Since m_t is nonincreasing by the design of Algorithm 1, there can be at most t iterations such that $m_t \ge m$ and $m_{t+1} < m$. Prior to this transition iteration, we analyze Algorithm 1 with V_t^{GD} and after it, we analyze Algorithm 1 with V_t^{NAG}. Therefore, to join the analyses, we bound V_t^{NAG} in terms of V_t^{GD}.

Lemma A.11. *Let $f \in \mathcal{F}(L,m)$. If $\bar{L} \ge L$ and $m_{t-1} \ge m$, then for all s_t*

$$V_t^{\text{NAG}}(s_t) \le \frac{m_{t-1}}{m}V_t^{\text{GD}}(s_t). \tag{86}$$

Proof. To prove (86), we split the analysis according to the sign of $\langle x_t^\star, x_t^y\rangle$, bounding the gap

$$\begin{aligned} V_t^{\text{NAG}}(s_t) - V_t^{\text{GD}}(s_t) &= \frac{m_{t-1}}{2}\|x_t^\star + \sqrt{\bar{\kappa}_{t-1}}x_t^y\|^2 - \frac{m}{2}\|x_t^\star + \sqrt{\bar{\kappa}}x_t^y\|^2 \\ &= \frac{m_{t-1}-m}{2}\|x_t^\star\|^2 + 2\frac{\sqrt{\bar{L}}(\sqrt{m_{t-1}}-\sqrt{m})}{2}\langle x_t^\star, x_t^y\rangle \end{aligned} \tag{87}$$

in terms of V_t^{GD} or V_t^{NAG}. We consider the case $\langle x_t^\star, x_t^y\rangle \ge 0$ first.

Multiplying the coefficient of $\langle x_t^\star, x_t^y\rangle$ on (87) by $(\sqrt{m_{t-1}}+\sqrt{m})/\sqrt{m_{t-1}} \ge 1$, we obtain

$$\sqrt{\bar{L}}(\sqrt{m_{t-1}}-\sqrt{m}) \le \sqrt{\bar{\kappa}_{t-1}}(m_{t-1}-m). \tag{88}$$

Hence, if $\langle x_t^\star, x_t^y\rangle \ge 0$, then plugging (88) into (87), adding a nonnegative $\|x_t^y\|^2$ term, completing a square to form a $\|w_t^\star\|^2$ term and then applying (80), we get

$$\begin{aligned} V_t^{\text{NAG}}(s_t) - V_t^{\text{GD}}(s_t) &\le \frac{m_{t-1}-m}{m_{t-1}}\frac{m_{t-1}}{2}(\|x_t^\star\|^2 + 2\sqrt{\bar{\kappa}_{t-1}}\langle x_t^\star, x_t^y\rangle + \bar{\kappa}_{t-1}\|x_t^y\|^2) \\ &\le \frac{m_{t-1}-m}{m_{t-1}}V_t^{\text{NAG}}(s_t). \end{aligned}$$

Moving terms around and then multiplying both sides by $m_{t-1}/m \ge 1$, we get

$$V_t^{\text{NAG}}(s_t) \le \frac{m_{t-1}}{m}V_t^{\text{GD}}(s_t).$$

Now, suppose $\langle x_t^\star, x_t^y\rangle < 0$. In this case, we cannot increase the $\langle x_t^\star, x_t^y\rangle$ coefficient to complete a square as we did before. Instead, we complete a square with the given $\langle x_t^\star, x_t^y\rangle$ coefficient by splitting the $\|x_t^\star\|^2$ term using the following identity and then handling an extra $\|x_t^\star\|^2$ term,

$$\frac{m_{t-1}-m}{m} = \frac{\sqrt{m_{t-1}}-\sqrt{m}}{\sqrt{m}}\frac{\sqrt{m_{t-1}}+\sqrt{m}}{\sqrt{m}} = \frac{\sqrt{m_{t-1}}-\sqrt{m}}{\sqrt{m}}\left(1+\frac{\sqrt{m_{t-1}}}{\sqrt{m}}\right). \tag{89}$$

To handle the $\|x_t^\star\|^2$ term that stays out of the square, we use that $\langle x_t^\star, x_t^y\rangle < 0$ implies

$$\|y_t^\star\|^2 = \|y_t^\star \pm x_t^\star\|^2 = \|x_t^\star - x_t^y\|^2 = \|x_t^\star\|^2 - 2\langle x_t^\star, x_t^y\rangle + \|x_t^y\|^2 \ge \|x_t^\star\|^2. \tag{90}$$

By the definition of $\bar{\alpha}_{t-1}$, the assumption that $m_{t-1} \geq m$ yields $\bar{\alpha}_{t-1} \geq 0$, thus $\bar{\alpha}_{t-1}/\sqrt{\bar{\kappa}} \geq 0$. Moreover, $U \geq 0$ by its definition, (36). Since $\tilde{f}$ is also nonnegative, from (41) we obtain

$$V_t^{\mathrm{GD}}(s_t) \geq W(s_t) = \tilde{f}(y_t) + (m/2)\|z_t^\star\|^2 \geq (m/2)\max\left\{\|z_t^\star\|^2, \|y_t^\star\|^2\right\}, \tag{91}$$

where the last inequality on the right-hand side follows from applying (3) with $x = x^\star$ and $y = y_t$. Splitting the coefficient of $\|x_t^\star\|^2$ on (87) according to (89), we get

$$\frac{m_{t-1}-m}{2}\|x_t^\star\|^2 = \frac{m_{t-1}-m}{m}\frac{m}{2}\|x_t^\star\|^2 = \frac{\sqrt{m_{t-1}}-\sqrt{m}}{\sqrt{m}}\frac{m}{2}\Big(1+\frac{\sqrt{m_{t-1}}}{\sqrt{m}}\Big)\|x_t^\star\|^2.$$

In the same vein, we rewrite the $\langle x_t^\star, x_t^y\rangle$ term on (87) as

$$\begin{aligned}2\frac{\sqrt{\bar{L}}(\sqrt{m_{t-1}}-\sqrt{m})}{2}\langle x_t^\star, x_t^y\rangle &= 2\frac{\sqrt{\bar{L}}(\sqrt{m_{t-1}}-\sqrt{m})}{2}\frac{m}{m}\langle x_t^\star, x_t^y\rangle \\ &= 2\frac{\sqrt{m_{t-1}}-\sqrt{m}}{\sqrt{m}}\frac{m}{2}\sqrt{\bar{\kappa}}\langle x_t^\star, x_t^y\rangle.\end{aligned}$$

Plugging the above back into (87), adding a positive $\|x_t^y\|^2$ term to form a $\|z_t^\star\|^2$ term, applying (90) and then using (91), we obtain

$$\begin{aligned}V_t^{\mathrm{NAG}}(s_t) - V_t^{\mathrm{GD}}(s_t) &= \frac{\sqrt{m_{t-1}}-\sqrt{m}}{\sqrt{m}}\frac{m}{2}\Big(\Big(1+\frac{\sqrt{m_{t-1}}}{\sqrt{m}}\Big)\|x_t^\star\|^2 + 2\sqrt{\bar{\kappa}}\langle x_t^\star, x_t^y\rangle\Big) \\ &\leq \frac{\sqrt{m_{t-1}}-\sqrt{m}}{\sqrt{m}}\frac{m}{2}\|z_t^\star\|^2 + \frac{\sqrt{m_{t-1}}-\sqrt{m}}{\sqrt{m}}\frac{\sqrt{m_{t-1}}}{\sqrt{m}}\frac{m}{2}\|y_t^\star\|^2 \\ &\leq \frac{m_{t-1}-m}{m}V_t^{\mathrm{GD}}(s_t).\end{aligned}$$

Finally, moving terms around, we get

$$V_t^{\mathrm{NAG}}(s_t) \leq \frac{m_{t-1}}{m}V_t^{\mathrm{GD}}(s_t).$$

Therefore, both when $\langle x_t^\star, x_t^y\rangle \geq 0$ and when $\langle x_t^\star, x_t^y\rangle < 0$, the inequality

$$V_t^{\mathrm{NAG}}(s_t) \leq \frac{m_{t-1}}{m}V_t^{\mathrm{GD}}(s_t)$$

holds generically for all s_t, establishing (86). □

Now, we are ready to prove Theorem A.1, which combines Theorems A.7 and A.10 into a single result that holds for all iterations of Algorithm 1 under Assumption 4.2. To this end, we replace δ_t^{GD} and $\hat{\delta}_t^{\mathrm{NAG}}$ with a common rate increment for all iterations. If Assumption 4.2 holds, then

$$\bar{\kappa}_t = \frac{\bar{L}}{m_t} \leq \frac{\bar{L}}{m/\gamma} = \gamma\bar{\kappa}. \tag{92}$$

Plugging (92) into the definitions of δ_t^{GD} and $\hat{\delta}_t^{\mathrm{NAG}}$, it follows that

$$\delta_{t+1}^{\mathrm{GD}} = \frac{1}{\bar{\kappa}_t - 1} \geq \frac{1}{\gamma\bar{\kappa}-1} = \delta(\gamma\bar{\kappa}), \tag{93}$$

$$\hat{\delta}_{t+1}^{\mathrm{NAG}} = \frac{1}{\sqrt{\bar{\kappa}_t}-1} \geq \frac{1}{\sqrt{\gamma\bar{\kappa}}-1} \geq \frac{1}{\gamma\bar{\kappa}-1} = \delta(\gamma\bar{\kappa}). \tag{94}$$

Moreover, by (4), $m_t \leq L$ holds for all m_t generated by Algorithm 1. Furthermore, if Assumption 4.2 holds, then $m_t \geq m/\gamma$. Therefore, given t and t' such that $t' > t$, we have that

$$\frac{\bar{\kappa}_{t'}^2}{\bar{\kappa}_t\bar{\kappa}} = \frac{m_t^2}{m_{t'}m} \leq \frac{\bar{L}^2}{m^2/\gamma} \leq \frac{\bar{L}^2}{m^2/\gamma} = \gamma\bar{\kappa}^2. \tag{95}$$

Proof of Theorem A.1. If $m_t \geq m$ for all $t \geq 0$, then by Theorem 4.1

$$f(y_{t+1}) - f(x^\star) \leq 2\bar{L}(1+\delta^{\text{GD}})^{-t}\|x_0 - x^\star\|^2 \leq 2\gamma\bar{L}\bar{\kappa}^2\left(\frac{\gamma\bar{\kappa}-1}{\gamma\bar{\kappa}}\right)^t \|x_0 - x^\star\|^2.$$

Otherwise, let $T+1$ be the first iteration for which $m_{T+1} < m$. Then, combining (86) with (83), plugging (95) in and then using (43), we obtain

$$\begin{aligned}
f(y_{t+1}) - f(x^\star) &\leq \frac{\bar{\kappa}_t^2}{\bar{\kappa}_{T-1}^2}\prod_{i=T}^{t+1}(1+\hat{\delta}_i^{\text{NAG}})^{-1} V_T^{\text{NAG}}(s_T) \\
&\leq \frac{\bar{\kappa}_t^2}{\bar{\kappa}_{T-1}\bar{\kappa}}\prod_{i=T}^{t+1}(1+\hat{\delta}_i^{\text{NAG}})^{-1} V_T^{\text{GD}}(s_T) \\
&\leq \gamma\bar{\kappa}^2(1+\delta(\gamma\bar{\kappa}))^{-(t+1-T)} V_T^{\text{GD}}(s_T) \\
&\leq 2\gamma\bar{L}\bar{\kappa}^2(1+\delta(\gamma\bar{\kappa}))^{-(t+1)}\|x_0 - x^\star\|^2 \\
&\leq 2\gamma\bar{L}\bar{\kappa}^2\left(\frac{\gamma\bar{\kappa}-1}{\gamma\bar{\kappa}}\right)^{t+1}\|x_0 - x^\star\|^2.
\end{aligned}$$

Therefore, (10) holds for all $t \geq 0$. □

B Local acceleration

In this section, we prove that Algorithm 1 converges at an accelerated rate to $x^\star$, the minimum of the objective function $f \in \mathcal{F}(L, m)$, when the iterates of Algorithm 1 get sufficiently close to $x^\star$. By accelerated rate, we mean $r_{\text{acc}}(\sigma\bar{\kappa})$, expressed in terms of a suboptimality factor $\sigma > 1$ multiplying the condition number $\bar{\kappa} = \bar{L}/m > L/m = \kappa$, and r_{acc} defined over $[1, +\infty)$ as

$$r_{\text{acc}}(z) = \frac{\sqrt{z} - 1}{\sqrt{z}}, \tag{96}$$

where $\bar{L} > L$ is the upper bound on L used in Algorithm 1.

Before proceeding, some notation remarks are in order.

Deriving most results assuming $\bar{L} = L$. We derive most of the results in this section using $\bar{L} = L$ to avoid working with the cluttered notation $\bar{L}$. This does not incur any loss of generality as long as we replace $\bar{L}$ and $\bar{\kappa}$ for L and κ in the final results. In fact, the consequence of allowing $\bar{L} = L$ is that the eigenvalues of $\nabla^2 f(x^\star)$ can actually take the value of $\bar{L}$, which cannot occur when $\bar{L} > L$. That is, we derive more general results that would also apply to the case where $\bar{L} > L$.

Hessian eigenpairs and eigendecomposition of $x_0 - x^\star$. Let (λ_i, v_i) denote the d eigenvalues λ_i and associated eigenvectors v_i of $\nabla^2 f(x^\star)$. If $f \in \mathcal{F}(L, m)$, then v_i can be chosen to form an orthonormal basis for $\mathbb{R}^d$. Hence, $x_0 - x^\star$ uniquely decomposes into $x_0 - x^\star = \sum_{i=1}^d x_{i,0} v_i$. Moreover, $\lambda_i \in [m, L]$. In the following, without loss of generality we assume λ_i ordered by their indices, as in $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_d$. Thus, $x_{1,0}$ denotes the coordinate of $x_0 - x^\star$ in the eigenspace associated with λ_1, the least eigenvalue of $\nabla^2 f(x^\star)$.

Assumption (4.2). The estimates m_{t+1} of Algorithm 1 decay by a factor of at least $\gamma \geq 2$ every time they decrease and are always greater than m/γ: if $m_{t+1} < m_t$, then $m/\gamma \leq m_{t+1} \leq m_t/\gamma$.

Assumption (4.3). The Hessian of f is locally Lipschitz-smooth at $x^\star$: there are $L_H > 0$ and $\epsilon_H > 0$ such that if $\|x - x^\star\| \leq \epsilon_H$, then $\|\nabla^2 f(x) - \nabla^2 f(x^\star)\| \leq L_H \|x - x^\star\|$.

Assumption (4.4). There exists some $\delta_\lambda \in (0, 1)$ such that $|m_t - \lambda_i| > \delta_\lambda L$ for every $\lambda_i > m$, where $m = \lambda_1 \leq \ldots \leq \lambda_d \leq L$ denote the eigenvalues of $\nabla^2 f(x^\star)$.

Assumption (4.5). There exists some $\omega > 0$ such that $\omega x_{1,0}^2 \geq \|x_0 - x^\star\|^2$.

To prove Theorem 4.6, we first consider the simplified case where f is quadratic, and then analyze the general case by showing it is a perturbation of the quadratic case. To this end, we use the fact established by Theorem 4.1 that the iterates of Algorithm 1 converge to $x^\star$ at a rate no worse than that of gradient descent, regardless of the initial point x_0. By that we mean

$$f(y_t) - f(x^\star) \leq 2\bar{L} r_{\text{GD}}(\bar{\kappa})^t \|x_0 - x^\star\|^2,$$

where r_{GD} is defined over $[1, +\infty)$ as

$$r_{\text{GD}}(z) = \frac{z - 1}{z}. \tag{97}$$

B.1 Quadratic case

First, we assume the objective function is quadratic and given by $f(x) = (1/2)(x - x^\star)^\mathsf{T} H (x - x^\star)$, with $H \in \mathbb{R}^{d \times d}$. Every quadratic function $(1/2)x^\mathsf{T} H x + x^\mathsf{T} g + f(0)$ can be expressed[3] in the form $(1/2)(x - x^\star)^\mathsf{T} H (x - x^\star) + f(x^\star)$, and minimizing the latter is equivalent to minimizing $(1/2)(x - x^\star)^\mathsf{T} H (x - x^\star)$. Thus, $\nabla f(x) = H(x - x^\star)$. Moreover, since $f \in \mathcal{F}(L, m)$, H must be positive definite with all d eigenvalues λ_i inside $[m, L]$. Hence, assuming λ_i ordered by their indices, we have that

$$m = \lambda_1 \leq \cdots \leq \lambda_d = L.$$

[3] Since H is strongly convex, H is invertible and the first-order condition $Hx^\star + g = 0$ admits a unique solution $x^\star$. Plugging $x = x^\star$ back into $f(x)$, and solving for $f(0)$, we get that $f(0) = -\frac{1}{2} x^{\star,\mathsf{T}} H x^\star$. Then, plugging $f(0)$ back into $f(x)$ and replacing the inner-product $g^\mathsf{T} x$ with $g^\mathsf{T} x = -x^{\star,\mathsf{T}} H x = -\frac{1}{2} x^{\star,\mathsf{T}} H x - \frac{1}{2} x^\mathsf{T} H x^\star$ yields the desired form of $f(x)$.

Since $\nabla^2 f$ is locally Lipschitz at $x^\star$, it is also continuous at $x^\star$. Hence, $H = \nabla^2 f(x^\star)$ is real symmetric in general, not only in the case where f is quadratic. Therefore, by the spectral theorem [Conway, 2019] we can pick eigenvectors v_i associated with λ_i such that $\{v_i\}_{i=1}^d$ form an orthonormal basis for $\mathbb{R}^d$. Then, $x_t - x^\star$ and $y_{t+1} - x^\star$ can be uniquely decomposed in this eigenbasis as

$$x_t - x^\star = \sum_{i=1}^{d} x_{i,t} v_i, \tag{98}$$

$$y_{t+1} - x^\star = x_t - \frac{1}{L}\nabla f(x_t) - x^\star = \sum_{i=1}^{d} \Big(1 - \frac{\lambda_i}{L}\Big) x_{i,t} v_i. \tag{99}$$

Substituting (99) for the descent steps yields

$$\begin{aligned}
\sum_{i=1}^{d} x_{i,t+1} v_i &= x_{t+1} - x^\star \\
&= (1+\beta_t) y_{t+1} - \beta_t y_t - x^\star \mp \beta_t x^\star \\
&= (1+\beta_t)(y_{t+1} - x^\star) - \beta_t (y_t - x^\star) \\
&= \sum_{i=1}^{d} \left[(1+\beta_t)\Big(1 - \frac{\lambda_i}{L}\Big) x_{i,t} - \beta_t \Big(1 - \frac{\lambda_i}{L}\Big) x_{i,t-1} \right] v_i,
\end{aligned} \tag{100}$$

where $\beta_t = \beta(m_t)$ is a particular value taken by the function $\beta : (0, L] \to [0, 1)$ defined by

$$\beta(m_t) = \frac{\sqrt{L} - \sqrt{m_t}}{\sqrt{L} + \sqrt{m_t}}. \tag{101}$$

That is, each component $x_{i,t}$ of $x_t - x^\star$ behaves as an LTV system Hespanha [2009]. But under Assumption 4.2, m_t decreases by a factor of at least $\gamma > 1$ every time it is updated, which implies m_t only changes finitely many times. Hence, each $x_{i,t}$ behaves as a sequence of linear time-invariant (LTI) systems described by

$$X_{i,t+1} = G_i(m_t) X_{i,t}, \tag{102}$$

where $X_{i,t}$ denote the vectors of current and past coordinates stacked together as in

$$X_{i,t} = \begin{cases} [x_{i,0} \quad x_{i,0}]^\mathsf{T}, & t = 0, \\ [x_{i,t-1} \quad x_{i,t}]^\mathsf{T}, & t > 0, \end{cases} \tag{103}$$

and $G_i : (0, L] \to \mathbb{R}^{2\times 2}$ map estimates m_t to system matrices given by

$$G_i(m_t) = \begin{bmatrix} 0 & 1 \\ -\beta(m_t)\Big(1 - \frac{\lambda_i}{L}\Big) & (1+\beta(m_t))\Big(1 - \frac{\lambda_i}{L}\Big) \end{bmatrix}. \tag{104}$$

Hence, the dynamics of (102) is determined by the eigenvalues of $G_i(m_t)$, which are given by

$$\lambda(G_i(m_t)) = \frac{1+\beta(m_t)}{2}\Big(1 - \frac{\lambda_i}{L}\Big) \pm \sqrt{\frac{(1+\beta(m_t))^2}{4}\Big(1 - \frac{\lambda_i}{L}\Big)^2 - \beta(m_t)\Big(1 - \frac{\lambda_i}{L}\Big)}. \tag{105}$$

The greatest between the two eigenvalues given by (105) defines the so-called spectral radius [Golub and Van Loan, 2013] of G_i, captured by the function $\rho : (0, L] \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ defined by

$$\rho(s, \ell) = \max \left| \frac{1+\beta(m_t)}{2}\Big(1 - \frac{\ell}{L}\Big) \pm \sqrt{\frac{(1+\beta(m_t))^2}{4}\Big(1 - \frac{\ell}{L}\Big)^2 - \beta(m_t)\Big(1 - \frac{\ell}{L}\Big)} \right|. \tag{106}$$

We also define a function $\varrho : (0, L] \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ for the least of the two eigenvalues:

$$\varrho(s, \ell) = \min \left| \frac{1+\beta(m_t)}{2}\Big(1 - \frac{\ell}{L}\Big) \pm \sqrt{\frac{(1+\beta(m_t))^2}{4}\Big(1 - \frac{\ell}{L}\Big)^2 - \beta(m_t)\Big(1 - \frac{\ell}{L}\Big)} \right|. \tag{107}$$

Note that ρ and ϱ take an argument "ℓ" that need not be an actual eigenvalue λ_i of H, which will be convenient later on. Before that, in the next section, we prove some auxiliary results on ρ and ϱ.

B.1.1 Properties of the spectral radius ρ

Lemma B.1. *Let $s, \ell \in (0, L]$. The two numbers*

$$\frac{1+\beta(s)}{2}\Big(1-\frac{\ell}{L}\Big) \pm \sqrt{\frac{(1+\beta(s))^2}{4}\Big(1-\frac{\ell}{L}\Big)^2 - \beta(s)\Big(1-\frac{\ell}{L}\Big)} \tag{108}$$

have nonzero imaginary part if and only if $s < \ell < L$. If (108) have zero imaginary part, then

$$\rho(s,\ell) = \frac{1+\beta(s)}{2}\Big(1-\frac{\ell}{L}\Big) + \sqrt{\frac{(1+\beta(s))^2}{4}\Big(1-\frac{\ell}{L}\Big)^2 - \beta(s)\Big(1-\frac{\ell}{L}\Big)}, \tag{109}$$

otherwise, if (108) have nonzero imaginary part, then

$$\rho(s,\ell) = \sqrt{\beta(s)\Big(1-\frac{\ell}{L}\Big)}. \tag{110}$$

Proof. Let r_+ be defined by

$$r_+ = \frac{1+\beta(s)}{2}\Big(1-\frac{\ell}{L}\Big) + \sqrt{\frac{(1+\beta(s))^2}{4}\Big(1-\frac{\ell}{L}\Big)^2 - \beta(s)\Big(1-\frac{\ell}{L}\Big)},$$

and let r_- be defined by

$$r_- = \frac{1+\beta(s)}{2}\Big(1-\frac{\ell}{L}\Big) - \sqrt{\frac{(1+\beta(s))^2}{4}\Big(1-\frac{\ell}{L}\Big)^2 - \beta(s)\Big(1-\frac{\ell}{L}\Big)}.$$

Also, let Δ be defined by

$$\Delta(s,\ell) = \frac{(1+\beta(s))^2}{4}\Big(1-\frac{\ell}{L}\Big)^2 - \beta(s)\Big(1-\frac{\ell}{L}\Big). \tag{111}$$

If $\ell = 0$, then $\ell \le s$ since $s \ge 0$, because $s \in (0, L]$. Moreover, plugging $\ell = 0$ into (111), we obtain

$$\Delta(s,\ell) = \frac{(1+\beta(s))^2}{4} - \beta(s) < 0 \iff (1-\beta(s))^2 = (1+\beta(s))^2 - 4\beta(s) < 0.$$

Hence, $\Delta(s,\ell) \ge 0$ because $(1-\beta)^2 \ge 0$. Furthermore, $1 - \ell/L = 1$ and $\rho(s,\ell)$ trivially reduces to form (109).

Now, suppose $\ell > 0$. If $\ell = L$, then $1 - \ell/L = 0$ and $\rho(s,\ell) = 0$ trivially has zero imaginary part and takes the form (109). Otherwise, if $\ell < L$, then $1 - \ell/L > 0$ and $\Delta < 0$ if and only if

$$(1+\beta)^2\Big(1-\frac{\ell}{L}\Big) - 4\beta < 0 \iff (1-\beta)^2 L < (1+\beta)^2 \ell \iff \frac{L}{\ell} < \left(\frac{1+\beta}{1-\beta}\right)^2, \tag{112}$$

where L/ℓ is well-defined since $\ell > 0$, by assumption, while $(1-\beta)^{-1}$ is well-defined because $0 \le \beta(s) < 1$ for all $s \in (0, L]$. Plugging (101) into β, the squared factor on the right-hand side of (112) turns into

$$\frac{1+\beta}{1-\beta} = \frac{2\sqrt{L}/(\sqrt{L}+\sqrt{s})}{2\sqrt{s}/(\sqrt{L}+\sqrt{s})} = \sqrt{L/s}. \tag{113}$$

Thus, by (112), $\Delta(s,\ell)$ is negative if and only if $s < \ell$. Hence, if $s \ge \ell$, then $\Delta \ge 0$, which combined with the the assumption that $L > \ell$ implies

$$1 - \frac{\ell}{L} = \left|1 - \frac{\ell}{L}\right| > 0,$$

so that

$$\frac{1+\beta}{2}\Big(1-\frac{\ell}{L}\Big) \ge \sqrt{\frac{(1+\beta)^2}{4}\Big(1-\frac{\ell}{L}\Big)^2 - \beta\Big(1-\frac{\ell}{L}\Big)} = \sqrt{\Delta}.$$

Plugging the above inequality back into r_+ and r_-, we obtain

$$\begin{aligned}|r_+| &= r_+ \\ &= \frac{1+\beta}{2}\Big(1-\frac{\ell}{L}\Big)+\sqrt{\Delta} \\ &\geq \frac{1+\beta}{2}\Big(1-\frac{\ell}{L}\Big)-\sqrt{\Delta} \\ &= \left|\frac{1+\beta}{2}\Big(1-\frac{\ell}{L}\Big)-\sqrt{\Delta}\right| \\ &= |r_-|.\end{aligned}$$

That is, $\rho(s,\ell)$ takes the form (109).

Finally, if $s<\ell$, then $\Delta(s,\ell)$ is negative, so r_+ and r_- are complex conjugates with the same norm given by

$$|r_+| = \sqrt{\frac{1+\beta(s)^2}{4}\Big(1-\frac{\ell}{L}\Big)^2+\beta(s)\Big(1-\frac{\ell}{L}\Big)-\frac{(1+\beta(s))^2}{4}\Big(1-\frac{\ell}{L}\Big)^2} = \sqrt{\beta(s)\Big(1-\frac{\ell}{L}\Big)}.$$

Therefore, $\rho(s,\ell)$ takes the form (110). □

Corollary B.2. *If $m_t \in (0,L]$, then the eigenvalues of $G_i(m_t)$ have nonzero imaginary part if and only if $m_t<\lambda_i<L$. Moreover, if $\lambda_i<L$, then the eigenvalues of $G_i(m_t)$ coincide if and only if $m_t=\lambda_i$. Furthermore, if $\lambda_i<m_t$, then the eigenvalues of $G_i(m_t)$ are positive and distinct.*

Proof. Plugging $s=m_t$ and $\ell=\lambda_i$ into (108), we recover the two eigenvalues of $G_i(m_t)$ which, by Lemma B.1, have nonzero imaginary part if and only if $m_t<\lambda_i<L$.

Moreover, the eigenvalues of $G_i(m_t)$ coincide if and only if the discriminant (111) is zero for $\ell=\lambda_i$ and $s=m_t$. In turn, by (112) and (113), the discriminant (111) is zero for $\ell=\lambda_i$ and $s=m_t$ if and only if $m_t=\lambda_i$.

Furthermore, if $\lambda_i<m_t$, then for all $\lambda_i\in(0,L]$, we have that

$$\frac{1+\beta}{2}\Big(1-\frac{\lambda_i}{L}\Big) \geq \sqrt{\Delta(m_t,\lambda_i)} > 0.$$

Therefore, the eigenvalues of $G_i(m_t)$ are positive and distinct. □

Lemma B.3. *Given a and b such that $0\leq a<b\leq L$, then $\rho(s,b)<\rho(s,a)$ for all $s\in(0,L]$. In particular, if $b\in(m,L]$, then $\rho(s,b)<\rho(s,m)$ for all $s\in(0,L]$.*

Proof. Consider the following two cases:

case 1 ($b\leq s$). By assumption, $s\in(0,L]$, hence $s\leq L$ and if $b\leq s$, then $1-b/L\geq 0$. Moreover, $a<b\leq s$, so Lemma B.1 implies $\rho(s,a)$ and $\rho(s,b)$ both take form (109). If, in addition $s=L$, then $\beta=0$, which when substituted back into (109) yields

$$\rho(s,b) = 1-b/L < 1-a/L = \rho(s,a).$$

Otherwise, if $s<L$, then $\beta>0$. Moreover, $a<b\leq s$, so that $b-a>0$, therefore

$$\begin{aligned}\Delta(s,b)<\Delta(s,a) \iff & (1+\beta)^2\frac{b^2-a^2}{L} < 2((1+\beta)^2-2\beta)(b-a) \\ \iff & (1+\beta)^2\frac{b+a}{L} < 2(1+\beta^2) \\ \iff & \frac{4(L/s)}{(\sqrt{L/s}+1)^2}\frac{b+a}{L} < 2(1+\beta^2),\end{aligned}$$

where the last equivalence follows at once from (101). Furthermore, $a<b\leq s<L$, thus $\sqrt{L/s}+1>2$ and

$$\frac{4L/s}{(\sqrt{L/s}+1)^2}\frac{b+a}{L} = \frac{4}{(\sqrt{L/s}+1)^2}\frac{b+a}{s} \leq \frac{8}{(\sqrt{L/s}+1)^2} < 2(1+\beta^2).$$

Thus, $\Delta(s,b) < \Delta(s,a)$. Hence, since $\rho(s,a)$ and $\rho(s,b)$ are given by (109) and $1-b/L < 1-a/L$, it follows that

$$\rho(s,b) = \frac{1+\beta}{2}\Big(1-\frac{b}{L}\Big) + \sqrt{\Delta(s,b)} < \frac{1+\beta}{2}\Big(1-\frac{a}{L}\Big) + \sqrt{\Delta(s,a)} = \rho(s,a).$$

case 2 ($s < b$)**.** By assumption $b \le L$, so $a < b \le L$ and it follows that

$$\frac{(1+\beta)^2}{4}\Big(1-\frac{a}{L}\Big)^2 - \beta\Big(1-\frac{b}{L}\Big) > \frac{(1+\beta)^2}{4}\Big(1-\frac{a}{L}\Big)^2 - \beta\Big(1-\frac{a}{L}\Big) \ge 0,$$

that is

$$0 \le \beta\Big(1-\frac{b}{L}\Big) < \frac{(1+\beta)^2}{4}\Big(1-\frac{a}{L}\Big)^2.$$

If, in addition $b = L$, then $\rho(s,b) = 0$ the above inequality implies $\rho(s,b) < \rho(s,a)$. Otherwise, it must be that $s < b$, in which case $\rho(s,b)$ takes the form (110) by Lemma B.1 and the above inequality yields

$$\rho(s,b) = \sqrt{\beta\Big(1-\frac{b}{L}\Big)} < \frac{1+\beta}{2}\Big(1-\frac{a}{L}\Big) \le \rho(s,a).$$

□

Lemma B.4. *For every $s \in (0, L]$ and every $\ell \in [m, L]$, $\rho(s,\ell) \le r_{\mathrm{GD}}(\kappa) < 1$.*

Proof. Let $s \in (0, L]$ and $\ell \in [m, L]$. By Lemma B.3, $\rho(s,\ell) \le \rho(s,m)$, so it suffices to show $\rho(s,m) \le r_{\mathrm{GD}}(\kappa)$. If $m < m_t$, then by Lemma B.1, the eigenvalues of $G_1(s)$ have zero imaginary part and, omitting the argument s in $\beta = \beta(s)$, $\rho(s,m)$ is given by

$$\rho(s,m) = \frac{1+\beta}{2}\Big(1-\frac{m}{L}\Big) + \sqrt{\frac{(1+\beta)^2}{4}\Big(1-\frac{m}{L}\Big)^2 - \beta\Big(1-\frac{m}{L}\Big)}.$$

Hence, after simple manipulations, we obtain the equivalences

$$\begin{aligned}\rho(s,m) \le r_{\mathrm{GD}}(\kappa) \iff & \sqrt{\frac{(1+\beta)^2}{4}\Big(1-\frac{1}{\kappa}\Big)^2 - \beta\Big(1-\frac{1}{\kappa}\Big)} \le \frac{1-\beta}{2}\frac{\kappa-1}{\kappa} \\ \iff & \frac{(1+\beta)^2}{4}\Big(\frac{\kappa-1}{\kappa}\Big)^2 \le \frac{(1-\beta)^2}{4}\Big(\frac{\kappa-1}{\kappa}\Big)^2 + \beta\frac{\kappa-1}{\kappa}.\end{aligned}$$

Since $(1+\beta)^2 = (1-\beta)^2 + 4\beta$, $\beta \ge 0$ and $(\kappa-1) < \kappa$, it follows that

$$\frac{(1+\beta)^2}{4}\Big(\frac{\kappa-1}{\kappa}\Big)^2 = \frac{(1-\beta)^2+4\beta}{4}\Big(\frac{\kappa-1}{\kappa}\Big)^2 \le \frac{(1-\beta)^2}{4}\Big(\frac{\kappa-1}{\kappa}\Big)^2 + \beta\frac{\kappa-1}{\kappa}.$$

Therefore, $\rho(s,m) \le r_{\mathrm{GD}}(\kappa)$. Otherwise, if $s \le m$, then by Lemma B.1 the eigenvalues of $G_1(s)$ are complex, so that

$$\rho(s,m) = \sqrt{\beta\Big(1-\frac{m}{L}\Big)}.$$

Hence, after simple manipulations, we obtain the equivalences

$$\rho(s,m) \le r_{\mathrm{GD}}(\kappa) \iff \beta\frac{\kappa-1}{\kappa} \le \Big(\frac{\kappa-1}{\kappa}\Big)^2 \iff \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \le \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}}\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}}.$$

Since the right-hand side inequality above holds, so does $\rho(s,m) \le r_{\mathrm{GD}}(\kappa)$ and we are done. □

Lemma B.5. *If Assumption 4.4 holds, then $|\zeta_i - \xi_i| \ge \sqrt{\delta_L \delta_\lambda}$, where $\zeta_i = \zeta_i(m_t)$ and $\xi_i = \xi_i(m_t)$ denote the eigenvalues of $G_i(m_t)$ and $\delta_L = (\bar{L} - L)/\bar{L}$.*

Proof. If Assumption 4.4 holds, then there exists some $\delta_\lambda > 0$ such that $|m_t - \lambda_i| \ge \delta_\lambda$. Moreover, since $\bar{L} > L$, we have that $\delta_L = (\bar{L} - L)/\bar{L}$. Moreover, whether $\zeta_i = \zeta_i(m_t)$ and $\xi_i = \xi_i(m_t)$ are complex or real, we have that

$$|\zeta_i - \xi_i| = 2\left|\frac{(1+\beta)^2}{4}\frac{(\bar{L}-\lambda_i)^2}{\bar{L}^2} - \beta\frac{\bar{L}-\lambda_i}{\bar{L}}\right|^{1/2} = \frac{1+\beta}{\bar{L}}|(\bar{L}-\lambda_i)(m_t-\lambda_i)|^{1/2} \ge \sqrt{\delta_L\delta_\lambda},$$

where in the last equality we have replaced L with $\bar{L}$ in the identity

$$\frac{4\beta L}{(1+\beta)^2} = 4\frac{\sqrt{L}-\sqrt{s}}{\sqrt{L}+\sqrt{s}}\frac{(\sqrt{L}+\sqrt{s})^2}{4L}L = L - s. \tag{114}$$

□

B.1.2 Sufficiently accurate m_t estimates

In this section, we determine how good the estimate m_t must be for x_t to converge to $x^\star$ at an accelerated rate. From Lemma B.3, it follows that $\rho(m_t, m)$ dominates the convergence of x_t, therefore our goal is to characterize $\sigma = \sigma(m_t)$ such that $\rho(m_t, m) \leq r_{\text{acc}}(\sigma\kappa)$, which may represent a suboptimality factor relative to the optimal convergence rate of $r_{\text{acc}}(\kappa)$.

By the arguments in Appendix A.2, and Theorem A.10 in particular, if $m_t < m$, then the iterates converge at an accelerated rate. So, in this section, we focus on $m_t \in [m, (1+\delta_m)m]$, where $\delta_m > 0$ is a small number. We proceed in two steps. First, we bound $\rho(m_t, m)$ for $m_t \in [m, (1+\delta)m]$ in terms of a rate r_δ that depends on the relative precision $\delta > 0$ and the condition number κ. Second, given some $\sigma > 0$, we characterize δ_σ for which $r_\delta(\kappa) \leq r_{\text{acc}}(\sigma\kappa)$ holds for all $\delta \in (0, \delta_\sigma]$ and $\kappa \geq 1 + \delta$. The rate r_δ is parameterized by $\delta \in (0, 1)$ and defined over $z \geq 1 + \delta$ as

$$r_\delta(z) = \frac{1+\beta_\delta(z)}{2}\frac{z-1}{z} + \sqrt{\frac{(1+\beta_\delta(z))^2}{4}\left(\frac{z-1}{z}\right)^2 - \beta_\delta(z)\frac{z-1}{z}}, \tag{115}$$

where β_δ is also defined over $z \geq 1 + \delta$ as

$$\beta_\delta(z) = \frac{\sqrt{z} - \sqrt{1+\delta}}{\sqrt{z} + \sqrt{1+\delta}}. \tag{116}$$

Lemma B.6. *If $m \leq s \leq (1+\delta)m \leq L$, then $\rho(s, m) \leq r_\delta(\kappa)$ for all $\kappa \geq 1 + \delta$.*

Proof. Let $m \leq s \leq (1+\delta)m \leq L$. Since $m \leq s \leq L$, then by Lemma B.1

$$\rho(s, m) = \frac{1+\beta(L,s)}{2}\left(1 - \frac{m}{L}\right) + \sqrt{\frac{(1+\beta(L,s))^2}{4}\left(1-\frac{m}{L}\right)^2 - \beta(L,s)\left(1-\frac{m}{L}\right)}.$$

Omitting the arguments in $\beta = \beta(L, s)$ and using the identity (114), the discriminant above can be expressed as

$$\frac{(1+\beta)^2}{4}\left(1-\frac{m}{L}\right)^2 - \beta\left(1-\frac{m}{L}\right) = \frac{4L(L-m)(s-m)}{4L^2(\sqrt{L}+\sqrt{s})^2} = \frac{(L-m)(s-m)}{L(\sqrt{L}+\sqrt{s})^2}.$$

Plugging the above expression back into $\rho(s, m)$, we obtain

$$\rho(s,m) = \frac{\sqrt{L}}{\sqrt{L}+\sqrt{s}}\frac{L-m}{L} + \frac{\sqrt{L-m}}{\sqrt{L}}\frac{\sqrt{s-m}}{\sqrt{L}+\sqrt{s}} = \frac{\sqrt{L-m}}{\sqrt{L}}\frac{\sqrt{L-m}+\sqrt{s-m}}{\sqrt{L}+\sqrt{s}}.$$

The right-hand side above is increasing in $s \geq m$ since $Ls > (L-m)(s-m)$, which implies that

$$\begin{aligned}
\frac{\partial}{\partial s}\frac{\sqrt{L-m}+\sqrt{s-m}}{\sqrt{L}+\sqrt{s}} &= \frac{1}{2\sqrt{s-m}}\frac{1}{\sqrt{L}+\sqrt{s}} - \frac{\sqrt{L-m}+\sqrt{s-m}}{2\sqrt{s}(\sqrt{L}+\sqrt{s})^2} \\
&= \frac{m + \sqrt{Ls} - \sqrt{(L-m)(s-m)}}{2\sqrt{s}\sqrt{s-m}(\sqrt{L}+\sqrt{s})^2} \\
&> 0.
\end{aligned}$$

Therefore, for all $m \leq s \leq (1+\delta)m$ and $\kappa \geq 1 + \delta$, we have that

$$\rho(s, m) \leq \rho((1+\delta)m, m) = r_\delta(\kappa).$$

□

Next, we bound r_δ in terms of r_{acc}. We start with an identity involving $\beta_\delta(\kappa)$, analogous to (114):

$$4\frac{\beta_\delta(\kappa)}{(1+\beta_\delta(\kappa))^2} = 4\frac{\sqrt{\kappa}-\sqrt{1+\delta}}{\sqrt{\kappa}+\sqrt{1+\delta}}\frac{(\sqrt{\kappa}+\sqrt{1+\delta})^2}{4\kappa} = \frac{\kappa-(1+\delta)}{\kappa}.$$

Plugging the above identity into the discriminant of $r_\delta(\kappa)$ yields

$$\begin{aligned}
\frac{(1+\beta_\delta(\kappa))^2}{4}\left(\frac{\kappa-1}{\kappa}\right)^2 - \beta_\delta(\kappa)\frac{\kappa-1}{\kappa} &= \frac{\kappa}{(\sqrt{\kappa}+\sqrt{1+\delta})^2}\frac{\kappa-1}{\kappa}\left(\frac{\kappa-1}{\kappa} - \frac{\kappa-(1+\delta)}{\kappa}\right) \\
&= \frac{\kappa-1}{(\sqrt{\kappa}+\sqrt{1+\delta})^2}\frac{\delta}{\kappa}.
\end{aligned}$$

In turn, plugging the above expression for the discriminant back into $r_\delta(\kappa)$, we obtain an alternative expression for $r_\delta(\kappa)$:

$$r_\delta(\kappa) = \frac{\sqrt{\kappa}}{\sqrt{\kappa}+\sqrt{1+\delta}}\frac{\kappa-1}{\kappa} + \frac{\sqrt{\kappa-1}}{\sqrt{\kappa}+\sqrt{1+\delta}}\frac{\sqrt{\delta}}{\sqrt{\kappa}} = \frac{\sqrt{\kappa-1}}{\sqrt{\kappa}}\frac{\sqrt{\kappa-1}+\sqrt{\delta}}{\sqrt{\kappa}+\sqrt{1+\delta}}. \tag{117}$$

Using this alternative expression, we obtain the following.

Lemma B.7. *Given $\sigma > 1$, then $r_\delta(\kappa) \le r_{\text{acc}}(\sigma'\kappa)$ for all $\delta \in (0, \delta_\sigma]$, $\sigma' \ge \sigma$ and $\kappa \ge 1+\delta$, where $\delta_\sigma = (\sigma-1)^2/4\sigma$. Conversely, given $\delta > 0$, then $r_{\delta'}(\kappa) \le r_{\text{acc}}(\sigma\kappa)$ for all $\delta' \in (0, \delta]$, $\sigma \ge \sigma_\delta$ and $\kappa \ge 1+\delta'$, where $\sigma_\delta = 1 + 2\delta + 2\sqrt{\delta(1+\delta)}$.*

Proof. Let $\sigma > 1$. From (117) and (96), it follows that the condition that $r_\delta(\kappa) \le r_{\text{acc}}(\sigma\kappa)$ for some $\delta > 0$ and $\kappa \ge 1+\delta$ is equivalent to

$$\frac{\sqrt{\kappa-1}}{\sqrt{\kappa}}\frac{\sqrt{\kappa-1}+\sqrt{\delta}}{\sqrt{\kappa}+\sqrt{1+\delta}} \le \frac{\sqrt{\sigma\kappa}-1}{\sqrt{\sigma\kappa}}. \tag{118}$$

By successively manipulating (118), it follows that

$$\begin{aligned} r_\delta(\kappa) \le r_{\text{acc}}(\sigma\kappa) &\iff \sqrt{\kappa-1}(\sqrt{\kappa-1}+\sqrt{\delta})\sqrt{\sigma} \le (\sqrt{\sigma\kappa}-1)(\sqrt{\kappa}+\sqrt{1+\delta}) \\ &\iff \sqrt{\kappa}+\sqrt{1+\delta} \le (1+\sqrt{(1+\delta)\kappa}-\sqrt{\delta(\kappa-1)})\sqrt{\sigma} \\ &\iff \frac{\sqrt{\kappa}+\sqrt{1+\delta}}{1+\sqrt{(1+\delta)\kappa}-\sqrt{\delta(\kappa-1)}} \le \sqrt{\sigma}. \end{aligned} \tag{119}$$

Taking the derivative of the left-hand side of the (119) with respect to κ, we obtain

$$\begin{aligned} \frac{\partial}{\partial\kappa}\frac{\sqrt{\kappa}+\sqrt{1+\delta}}{1+\sqrt{(1+\delta)\kappa}-\sqrt{\delta(\kappa-1)}} &= \frac{\delta\sqrt{\kappa}+\delta\kappa\sqrt{(1+\delta)}-\delta\sqrt{\delta(\kappa-1)\kappa}}{2\kappa\sqrt{\delta(\kappa-1)}(1+\sqrt{(1+\delta)\kappa}-\sqrt{\delta(\kappa-1)})^2} \\ &= \frac{\delta}{2\sqrt{\delta\kappa(\kappa-1)}(1+\sqrt{(1+\delta)\kappa}-\sqrt{\delta(\kappa-1)})} > 0. \end{aligned}$$

That is, the left-hand side of (119) is increasing in $\kappa \ge 1+\delta$ and it follows that

$$\frac{\sqrt{\kappa}+\sqrt{1+\delta}}{1+\sqrt{(1+\delta)\kappa}-\sqrt{\delta(\kappa-1)}} \le \lim_{k\to+\infty}\frac{\sqrt{\kappa}+\sqrt{1+\delta}}{1+\sqrt{(1+\delta)\kappa}-\sqrt{\delta(\kappa-1)}} = \frac{1}{\sqrt{1+\delta}-\sqrt{\delta}}.$$

Moreover, $1/(\sqrt{1+\delta}-\sqrt{\delta})$ is increasing in $\delta > 0$. Therefore, if $\delta_\sigma = (\sigma-1)^2/4\sigma$, then for all $\delta \in (0, \delta_\sigma]$, $\kappa \ge 1+\delta$ and $\sigma' \ge \sigma$, we have that

$$\begin{aligned} \frac{\sqrt{\kappa}+\sqrt{1+\delta}}{1+\sqrt{(1+\delta)\kappa}-\sqrt{\delta(\kappa-1)}} &\le \frac{1}{\sqrt{1+\delta}-\sqrt{\delta}} \\ &\le \frac{1}{\sqrt{1+\delta_\sigma}-\sqrt{\delta_\sigma}} \\ &= \frac{2\sqrt{\sigma}}{\sqrt{(1+\sigma)^2}-\sqrt{(\sigma-1)^2}} \\ &= \sqrt{\sigma} \\ &\le \sqrt{\sigma'}. \end{aligned}$$

Conversely, given $\delta > 0$, if $\delta' \in (0, \delta]$ and $\sigma \ge \sigma_\delta$, where $\sigma_\delta = 1 + 2\delta + 2\sqrt{\delta(1+\delta)}$, then

$$\frac{1}{\sqrt{1+\delta'}-\sqrt{\delta'}} \le \frac{1}{\sqrt{1+\delta}-\sqrt{\delta}} = \sqrt{\sigma_\delta} \le \sqrt{\sigma}.$$

Therefore, $r_{\delta'}(\kappa) \le r_{\text{acc}}(\sigma\kappa)$ for all $\delta' \le \delta$, $\kappa \ge 1+\delta'$ and $\sigma \ge \sigma_\delta$. □

Corollary B.8. *Given $\sigma > 1$, then $\rho(s,m) \le r_{\text{acc}}(\sigma'\kappa)$ for all $s \in [m, (1+\delta)m]$, $\delta \in (0, \delta_\sigma]$, $\sigma' \ge \sigma$ and $\kappa \ge 1+\delta$, where $\delta_\sigma = (\sigma-1)^2/4\sigma$. Conversely, given $\delta > 0$, then $\rho(s,m) \le r_{\text{acc}}(\sigma\kappa)$ for all $s \in [m, (1+\delta)m]$, $\delta' \in (0, \delta]$, $\sigma \ge \sigma_\delta$ and $\kappa \ge 1+\delta'$, where $\sigma_\delta = 1 + 2\delta + 2\sqrt{\delta(1+\delta)}$.*

Proof. The corollary follows by combining Lemmas B.6 and B.7. □

B.1.3 Iterate dynamics between m_t updates

Through $G_i(m_t)$, the dynamics of $X_{i,t}$ are determined by m_t, which is updated by Algorithm 1 *after* the t-th iterate is computed. Moreover, under Assumption 4.2, m_t is updated at most $\log_\gamma \kappa + 1$ times. So, suppose the estimates m_t take $M + 1 \leq \log_\gamma \kappa + 1$ values. Then, let t_j denote the iteration in which m_t is adjusted to its j-th value μ_j, $j = 0, \dots, M$. Since NAG-free computes the iterate x_t and then adjusts m_t in iteration t, this means that $t_j + 1$ is the first iteration in which the estimate μ_j takes effect, and Algorithm 1 computes iterates for $t \in (t_j, t_{j+1}]$ using $m_t = \mu_j$. For example, $t_0 = 0$ and $m_t = \mu_0 = m_0$ for all $t < t_1$. Therefore, given t and t' such that $t_j < t' \leq t_{j+1} \leq t_J < t \leq t_{J+1}$,

$$X_{i,t} = \prod_{k=0}^{t-1} G_i(\mu_k) X_{i,0} = G_i(\mu_J)^{t-t_J} \left(\prod_{k=j+1}^{J-1} G_i(\mu_k)^{t_{k+1}-t_k} \right) G_i(\mu_j)^{t_{j+1}-t'} X_{i,t'}. \tag{120}$$

Now, if $m_t > m$, then under Assumption 4.4, Corollary B.2 implies that the eigenvalues of $G_i(m_t)$ are distinct. So, letting $\zeta_i = \zeta_i(m_t)$ and $\xi_i = \xi_i(m_t)$ denote the eigenvalues of $G_i(m_t)$, we define

$$T_i(m_t) = \begin{bmatrix} 1 & 1 \\ \zeta_i & \xi_i \end{bmatrix}. \tag{121}$$

It can be checked that the columns of $T_i(m_t)$ are eigenvectors of $G_i(m_t)$, therefore $T_i(m_t)$ diagonalizes $G_i(m_t)$:

$$G_i(m_t) = T_i(m_t) D_i(m_t) T_i(m_t)^{-1}. \tag{122}$$

That is, $D_i(m_t)$ is a diagonal matrix whose diagonal entries are the eigenvalues of $G_i(m_t)$:

$$D_i(m_t) = \begin{bmatrix} \zeta_i & 0 \\ 0 & \xi_i \end{bmatrix}. \tag{123}$$

Combining (120), (122) and (123), then applying Lemma B.3 it follows that for every $t_j < t \leq t_{j+1}$

$$\|X_{i,t}\|^2 \leq \overline{C}_i \rho(\mu_j, \lambda_i)^{2(t-t_j)} \left(\prod_{k=0}^{j-1} \rho(\mu_k, \lambda_i)^{2(t_{k+1}-t_k)} \right) x_{i,0}^2, \tag{124}$$

where the constant $\overline{C}_i$ that is uniformly bounded, since

$$\begin{aligned}
\|X_{i,t}\|^2 &= \left\| T_i(\mu_j) D_i(\mu_j)^{t-t_j} T_i(\mu_j)^{-1} \left(\prod_{k=0}^{j-1} T_i(\mu_k) D_i(\mu_k)^{t_{k+1}-t_k} T_i(\mu_k)^{-1} \right) X_{i,0} \right\|^2 \\
&\leq \|T_i(\mu_j) D_i(\mu_j)^{t-t_j} T_i(\mu_j)^{-1}\|^2 \left(\prod_{k=0}^{j-1} \|T_i(\mu_k) D_i(\mu_k)^{t_{k+1}-t_k} T_i(\mu_k)^{-1}\|^2 \right) x_{i,0}^2 \\
&\leq \left(\prod_{k=0}^{M} \|T_i(\mu_k)\|^2 \|T_i(\mu_k)^{-1}\|^2 \right) \|D_i(\mu_j)^{t-t_j}\|^2 \left(\prod_{k=0}^{j-1} \|D_i(\mu_k)^{t_{k+1}-t_k}\|^2 \right) 2 x_{i,0}^2 \\
&\leq \left(2 \prod_{k=0}^{\log_\gamma \kappa + 1} \|T_i(\mu_k)\|^2 \|T_i(\mu_k)^{-1}\|^2 \right) \rho(\mu_j, \lambda_i)^{2(t-t_j)} \left(\prod_{k=0}^{j-1} \rho(\mu_k, \lambda_i)^{2(t_{k+1}-t_k)} \right) x_{i.0}^2
\end{aligned}$$

and, by applying Lemmas B.4 and B.5 to (121), for all μ_k we obtain

$$\|T_i(\mu_k)\|^2 \leq 4, \qquad \|T_i(\mu_k)^{-1}\|^2 = \frac{1}{|\zeta_i - \xi_i|^2} \left\| \begin{bmatrix} \xi_i & -1 \\ -\zeta_i & 1 \end{bmatrix} \right\|^2 \leq \frac{4}{\sqrt{\delta_\lambda \delta_L}},$$

where $\delta_L = (\bar{L} - L)/\bar{L}$ and δ_λ is given by Assumption 4.4. Furthermore, omitting the m_t arguments, for $t \in (t_j, t_{j+1}]$, we have that

$$\begin{aligned}
X_{i,t} &= G_i^{t-t_j} X_{i,t_j} \\
&= T_i D_i^{t-t_j} T_i^{-1} X_{i,t_j} \\
&= \begin{bmatrix} 1 & 1 \\ \zeta_i & \xi_i \end{bmatrix} \begin{bmatrix} \zeta_i^{t-t_j} & 0 \\ 0 & \xi_i^{t-t_j} \end{bmatrix} \frac{1}{\xi_i - \zeta_i} \begin{bmatrix} \xi_i & -1 \\ -\zeta_i & 1 \end{bmatrix} \begin{bmatrix} x_{i,t_j-1} \\ x_{i,t_j} \end{bmatrix} \\
&= \frac{1}{\xi_i - \zeta_i} \begin{bmatrix} \zeta_i^{t-t_j} & \xi_i^{t-t_j} \\ \zeta_i^{t+1-t_j} & \xi_i^{t+1-t_j} \end{bmatrix} \begin{bmatrix} \xi_i & -1 \\ -\zeta_i & 1 \end{bmatrix} \begin{bmatrix} x_{i,t_j-1} \\ x_{i,t_j} \end{bmatrix} \\
&= \frac{1}{\xi_i - \zeta_i} \begin{bmatrix} \xi_i \zeta_i^{t-t_j} - \zeta_i \xi_i^{t-t_j} & \xi_i^{t-t_j} - \zeta_i^{t-t_j} \\ \xi_i \zeta_i^{t+1-t_j} - \zeta_i \xi_i^{t+1-t_j} & \xi_i^{t+1-t_j} - \zeta_i^{t+1-t_j} \end{bmatrix} \begin{bmatrix} x_{i,t_j-1} \\ x_{i,t_j} \end{bmatrix} \\
&= \frac{1}{\xi_i - \zeta_i} \begin{bmatrix} (\xi_i x_{i,t_j-1} - x_{i,t_j}) \zeta_i^{t-t_j} + (x_{i,t_j} - \zeta_i x_{i,t_j-1}) \xi_i^{t-t_j} \\ (\xi_i x_{i,t_j-1} - x_{i,t_j}) \zeta_i^{t+1-t_j} + (x_{i,t_j} - \zeta_i x_{i,t_j-1}) \xi_i^{t+1-t_j} \end{bmatrix}.
\end{aligned}$$

Therefore, $X_{i,t}$ can be decomposed into two modes:

$$X_{i,t} = A_{i,t_j} \zeta_i^{t-t_j} + B_{i,t_j} \xi^{t-t_j}, \tag{125}$$

where A_i and B_i are two-dimensional vectors given by

$$A_{i,t_j} = \frac{x_{i,t_j} - \xi_i x_{i,t_j-1}}{\zeta_i - \xi_i} \begin{bmatrix} 1 \\ \zeta_i \end{bmatrix} \qquad \text{and} \qquad B_{i,t_j} = \frac{\zeta_i x_{i,t_j-1} - x_{i,t_j}}{\zeta_i - \xi_i} \begin{bmatrix} 1 \\ \xi_i \end{bmatrix}, \tag{126}$$

which are well-defined, by Lemma B.5. In particular, for $t_0 < t \leq t_1$, we have that

$$X_{i,t} = \frac{(1-\xi_i) x_{i,0}}{\zeta_i - \xi_i} \begin{bmatrix} 1 \\ \zeta_i \end{bmatrix} \zeta_i^t + \frac{(\zeta_i - 1) x_{i,0}}{\zeta_i - \xi_i} \begin{bmatrix} 1 \\ \xi_i \end{bmatrix} \xi_i^t.$$

In turn, if without loss of generality we assume $x_{1,0} > 0$, then

$$x_{1,t} - x_{1,t-1} = \frac{(1-\xi_1)(\zeta_1 - 1)\zeta_1^t x_{1,0} + (\zeta_1 - 1)(\xi_1 - 1)\xi_1^t x_{1,0}}{\zeta_1 - \xi_1} \leq \kappa^{-1} \zeta_1^{t-1} x_{1,0} < 0,$$

where in first inequality above we used the fact that $0 < \xi_1 < \zeta_1$ and the identity

$$(1-\zeta_i)(1-\xi_i) = \left(1 - \frac{1+\beta}{2}\left(1 - \frac{\lambda_i}{L}\right)\right)^2 - \frac{(1+\beta)^2}{4}\left(1 - \frac{\lambda_i}{L}\right)^2 + \beta\left(1 - \frac{\lambda_i}{L}\right) = \frac{\lambda_i}{L}.$$

Moreover, for $t_0 < t \leq t_1$ we also have that

$$x_{1,t} - \xi_1 x_{1,t-1} = \frac{(1-\xi_1)(\zeta_1 - \xi_1)\zeta_1^t x_{1,0} + (\zeta_1 - 1)(\xi_1 - \xi_1)\xi_1^t x_{1,0}}{\zeta_1 - \xi_1} = (1-\xi_1)\zeta_1^t x_{1,0} < 0,$$

$$\zeta_1 x_{1,t-1} - x_{1,t} = \frac{(1-\xi_1)(\zeta_1 - \zeta_1)\zeta_1^t x_{1,0} + (\zeta_1 - 1)(\zeta_1 - \xi_1)\xi_1^t x_{1,0}}{\zeta_1 - \xi_1} = (\zeta_1 - 1)\xi_1^t x_{1,0} > 0.$$

It follows that, for $t_1 < t \leq t_2$

$$\begin{aligned}
x_{1,t} - x_{1,t-1} &= \frac{(x_{1,t_j} - \xi_1 x_{1,t_1-1})(\zeta_1 - 1)\zeta_1^{t-t_1} + (\zeta_1 x_{1,t_1-1} - x_{1,t_1})(\xi_1 - 1)\xi_1^{t-t_1}}{\zeta_1 - \xi_1} \\
&\leq \zeta_1^{t-t_1} \frac{(x_{1,t_j} - \xi_1 x_{1,t_1-1})(\zeta_1 - 1) + (\zeta_1 x_{1,t_1-1} - x_{1,t_1})(\xi_1 - 1)}{\zeta_1 - \xi_1} \\
&= \zeta_1^{t-t_1}(x_{1,t_1} - x_{1,t_1-1}) \\
&\leq \kappa^{-1} \zeta_1(\mu_1)^{t-t_1} \zeta_1(\mu_0)^{t_1-1} x_{1,0} \\
&< 0,
\end{aligned}$$

since $0 < \xi_1(m_1) < \zeta_1(m_1)$, and moreover

$$x_{1,t} - \xi_1 x_{1,t-1} = \zeta_1^{t-t_1}(x_{1,t_1} - \xi_1 x_{1,t_1-1}) = \zeta_1(\mu_1)^{t-t_1}(1 - \xi_1(\mu_0))\zeta_1(\mu_0)^t x_{1,0} < 0,$$

$$\zeta_1 x_{1,t-1} - x_{1,t} = \xi_1^{t-t_1}(\zeta_1 x_{1,t_1-1} - x_{1,t_1}) = \xi_1(\mu_1)^{t-t_1}(\zeta_1(\mu_0) - 1)\xi_1(\mu_0)^t x_{1,0} > 0.$$

Therefore, using the fact that $\zeta_1(m_t) = \rho(m_t, m)$, it follows by induction that for $t_j < t \leq t_{j+1}$

$$(x_{1,t+1} - x_{1,t})^2 \geq \underline{C}_1 \rho(\mu_j, m)^{2t-t_j} \left(\prod_{k=0}^{j-1} \rho(\mu_k, m)^{2(t_{k+1}-t_k)} \right) x_{1,0}^2 \geq 0, \tag{127}$$

for some $\underline{C}_1 \geq \kappa^{-2}$.

B.1.4 The dynamics of c_t

Lemma B.7 bounds the suboptimality factor in the convergence rate of x_t when $m_t \in [m, (1+\delta)m]$, for a given $\delta > 0$. Now, we determine how long m_t takes to reach the interval $[m, (1+\delta)m]$. Our starting point is to determine the dynamics of c_{t+1}. To this end, we plug (98) and (100) into (4), obtaining[4]

$$c_{t+1}^2 = \left\| \frac{\nabla f(x_{t+1}) - \nabla f(x_t)}{x_{t+1} - x_t} \right\|^2 = \left\| \frac{\sum_{i=1}^d (x_{i,t+1} - x_{i,t})\lambda_i v_i}{\sum_{i=1}^d (x_{i,t+1} - x_{i,t}) v_i} \right\|^2 = \frac{\sum_{i=1}^d (x_{i,t+1} - x_{i,t})^2 \lambda_i^2}{\sum_{i=1}^d (x_{i,t+1} - x_{i,t})^2}. \tag{128}$$

The identity (128) reveals that c_{t+1}^2 can be expressed as an average of the squared eigenvalues λ_i^2 weighted by $(x_{i,t+1} - x_{i,t})^2$. Since the weights are a static map of $x_{i,t}$, the dynamics of $x_{i,t}$ determine the dynamics of the estimated effective curvature c_{t+1}. In particular, $x_{i,t}$ determine if one weight can outweigh the others, in which case c_{t+1} tends to λ_i.

By Lemma B.3, $\rho(s, \lambda_i) < \rho(s, m)$ for all $\lambda_i \in (m, L]$. Hence, from (124) and (127), we conclude that the weight associated with m eventually dominates the other weights, so that c_{t+1} converges to m. In the following, we show that this happens at an accelerated rate. To this end, we define[5] $\phi : \mathcal{D} \to [0, 1]$ as

$$\phi(s, a, b) = \begin{cases} \min\left\{1, \dfrac{\rho(s,a)}{\rho(s,b)}\right\}, & \rho(s,b) > 0, \\ 1, & \rho(s,b) = 0, \end{cases} \tag{129}$$

where the domain $\mathcal{D}$ is given by

$$\mathcal{D} = (0, L] \times \left\{(a, b) \in \mathbb{R}_{>0}^2 : a \neq b\right\}, \tag{130}$$

$\mathbb{R}_{>0}$ being the set of positive real numbers. With ϕ, we can bound how fast c_{t+1} takes to decrease below $(1+\delta)\ell$ for a given $\ell \in [m, L]$, not necessarily an eigenvalue of H, where $\delta > 0$ represents some estimate precision relative to ℓ. To this end, we characterize $\phi((1+\delta)\ell, \ell, m)^2$, first showing that it is decreasing in ℓ.

Lemma B.9. *If $\delta \in (0, 1]$ and $\kappa \geq 2$, then $\phi((1+\delta)\ell, \ell, m)$ is decreasing in $\ell \geq m > 0$.*

Proof. Let $L > m > 0$. Given ℓ and $\delta > 0$ such that $m \leq \ell < (1+\delta)\ell \leq L$, by (129) and Lemma B.1, we have that

$$\phi((1+\delta)\ell, \ell, m) = \frac{L - \ell + \sqrt{(L-\ell)\delta\ell}}{L - m + \sqrt{(L-m)((1+\delta)\ell - m)}}.$$

Letting ϕ_ℓ the derivative of $\phi((1+\delta)\ell, \ell, m)$ with respect to ℓ, we obtain

$$\begin{aligned}
\phi_\ell =& \frac{-(L-m)\delta^2\ell - (L-m)\sqrt{(L-\ell)\delta\ell}(L + \ell + \sqrt{(L-m)((1+\delta)\ell - m)} - 2m)}{2\sqrt{(L-\ell)\delta\ell}\sqrt{(L-m)((1+\delta)\ell - m)}(L + \sqrt{(L-m)((1+\delta)\ell - m)} - m)^2} \\
&- \frac{(L-m)\delta(\ell^2 + \ell(\sqrt{(L-\ell)\delta\ell} + 2\sqrt{(L-m)((1+\delta)\ell - m)} - 2m))}{2\sqrt{(L-\ell)\delta\ell}\sqrt{(L-m)((1+\delta)\ell - m)}(L + \sqrt{(L-m)((1+\delta)\ell - m)} - m)^2} \\
&- \frac{(L-m)\delta L(\sqrt{(L-\ell)\delta\ell} - \sqrt{(L-m)((1+\delta)\ell - m)} + m)}{2\sqrt{(L-\ell)\delta\ell}\sqrt{(L-m)((1+\delta)\ell - m)}(L + \sqrt{(L-m)((1+\delta)\ell - m)} - m)^2} \\
\leq& - \frac{(L-m)((L-m)\sqrt{(L-\ell)\delta\ell} - \delta(L-2\ell)\sqrt{(L-m)((1+\delta)\ell - m)})}{2\sqrt{(L-\ell)\delta\ell}\sqrt{(L-m)((1+\delta)\ell - m)}(L + \sqrt{(L-m)((1+\delta)\ell - m)} - m)^2}.
\end{aligned}$$

So, to show $\phi((1+\delta)\ell, \ell, m)$ is decreasing in ℓ, it suffices to show the numerator above is positive. To this end, since $L > m$, it suffices to show that the second factor is positive:

$$\begin{aligned}
&(L-m)\sqrt{(L-\ell)\delta\ell} + \sqrt{(L-\ell)\delta\ell}\sqrt{(L-m)((1+\delta)\ell - m)} \\
&- \delta(L-2\ell)\sqrt{(L-m)((1+\delta)\ell - m)} > 0.
\end{aligned} \tag{131}$$

[4] Note that $x_{t+1} - x_t = (x_{t+1} - x^\star) - (x_t - x^\star) = \sum_{i=1}^d (x_{i,t+1} - x_{i,t}) v_i$.

[5] Note that $\rho < 0$ cannot occur by the definition of ρ, (106).

The negative term on the left-hand side above is maximized at the critical point characterized by

$$\begin{aligned}\frac{\partial}{\partial \ell}(L-2\ell)\sqrt{((1+\delta)\ell-m)} &= -2\sqrt{(1+\delta)\ell-m}+\frac{(L-2\ell)(1+\delta)}{2\sqrt{(1+\delta)\ell-m}}\\ &= \frac{(1+\delta)(L-2\ell)-4((1+\delta)\ell-m)}{2\sqrt{(1+\delta)\ell-m}}\\ &= 0.\end{aligned}$$

Taking ℓ at this critical point, $\ell = \frac{1}{6}L + \frac{2}{3(1+\delta)}m$, and using the assumptions that $\kappa \geq 2$ and $\delta \leq 1$, it follows that

$$(L-\ell)\ell \geq \frac{5L-4m}{6}\frac{(1+\delta)L+4m}{6(1+\delta)} = \frac{5(1+\delta)L^2+4(5-(1+\delta))Lm-16m^2}{36(1+\delta)} \geq \frac{5}{36}L^2.$$

Hence, plugging $\ell = \frac{1}{6}L + \frac{2}{3(1+\delta)}m$ back into (131) and using the assumptions that $\delta \leq 1$ and $\kappa \geq 2$ yields

$$\begin{aligned}&(\delta(L-2\ell)-\sqrt{(L-\ell)\delta\ell})\sqrt{(L-m)((1+\delta)\ell-m)}\\ &\leq \delta\frac{(4-\sqrt{5})L}{6}\sqrt{(L-m)\frac{1+\delta}{6}\left(L-\frac{2m}{1+\delta}\right)}\\ &\leq \frac{2\delta\sqrt{1+\delta}}{6\sqrt{6}}L(L-m),\end{aligned}$$

and, similarly

$$\sqrt{(L-\ell)\delta\ell}(L-m) \geq \sqrt{\delta\frac{5L-4m}{6}\frac{L}{6}}(L-m) \geq \frac{\sqrt{\delta}}{3\sqrt{2}}L(L-m).$$

Hence, canceling the common factor $\sqrt{\delta}L(L-m)$ above and then rearranging, we conclude that (131) holds if

$$\sqrt{\delta}\sqrt{1+\delta} \leq \sqrt{3},$$

which is true since $\sqrt{\delta} \leq 1$. □

In fact, $\phi((1+\delta)\ell, \ell, m)$ is decreasing for any $\delta > 0$, which can be seen in its graph, but the case where $\delta \in (0,1]$ suffices for the upcoming results. Namely, given $\delta_\ell > 0$ and $\delta_u \in (0,1]$, by Lemma B.9 we have that for every $\ell \in [(1+\delta_\ell)m, L]$

$$\begin{aligned}\phi((1+\delta_u)\ell,\ell,m) =& \frac{L-\ell+\sqrt{(L-\ell)\delta_u\ell}}{L-m+\sqrt{(L-m)((1+\delta_u)\ell-m)}}\\ \leq& \frac{L-(1+\delta_\ell)m+\sqrt{(L-(1+\delta_\ell)m)\delta_u(1+\delta_\ell)m}}{L-m+\sqrt{(L-m)((1+\delta_u)(1+\delta_\ell)m-m)}}\\ =& \frac{\kappa-(1+\delta_\ell)+\sqrt{(\kappa-(1+\delta_\ell))\delta_u(1+\delta_\ell)}}{\kappa-1+\sqrt{(\kappa-1)(\delta_u+\delta_\ell+\delta_u\delta_\ell)}}\\ =:& r_\phi(\delta_u,\delta_\ell,\kappa).\end{aligned} \tag{132}$$

Hence, to show $\phi((1+\delta_u)\ell,\ell,m)^2$ is an accelerated rate, suffices to show that $r_\phi(\delta_u,\delta_\ell,\kappa)^2$ is an accelerated rate for appropriate κ, δ_u and δ_ℓ, which we do in the next result. The function r_ϕ is well-defined for $\delta_\ell > 0$, $\delta_u > 0$ and $k \geq 1+\delta_\ell$ and, by simple inspection, it follows that $r_\phi(\delta_u,\delta_\ell,\kappa) \in (0,1)$.

Lemma B.10. *Given $\delta_u > 0$, $\delta_\ell > 0$ and $\kappa \geq 1+\delta_\ell$, there is a $\sigma_\phi = \sigma_\phi(\delta_u,\delta_\ell,\kappa)$ such that*

$$r_\phi(\delta_u,\delta_\ell,\kappa)^2 \leq r_{\text{acc}}(\sigma_\phi\kappa).$$

Moreover, the function $\kappa \mapsto \sigma_\phi(\delta_u,\delta_\ell,\kappa)$ is bounded and satisfies

$$\lim_{\kappa\to+\infty}\sigma_\phi(\delta_u,\delta_\ell,\kappa) = \frac{1}{4(\sqrt{\delta_u+\delta_\ell+\delta_u\delta_\ell}-\sqrt{\delta_u(1+\delta_\ell)})^2}.$$

Proof. Let $\delta_u > 0$, $\delta_\ell > 0$ and $\kappa \geq 1 + \delta_\ell$. By direct algebraic manipulation, we obtain

$$r_\phi(\delta_u, \delta_\ell, \kappa)^2 \leq r_{\text{acc}}(\sigma_\phi \kappa) = \frac{\sqrt{\sigma_\phi \kappa} - 1}{\sqrt{\sigma_\phi \kappa}} \iff \frac{1}{(1 - r_\phi(\delta_u, \delta_\ell, \kappa)^2)^2 \kappa} \leq \sigma_\phi. \tag{133}$$

For such δ_u, δ_ℓ and κ, we have $r_\phi(\delta_u, \delta_\ell, \kappa) \in (0, 1)$, so that $1 - r_\phi^2 > 0$. Therefore, the lower bound of the inequality on the right-hand side of (133) is well-defined. So, let σ_ϕ be defined such that (133) holds with equality:

$$\sigma_\phi(\delta_u, \delta_\ell, \kappa) = \frac{1}{(1 - r_\phi(\delta_u, \delta_\ell, \kappa)^2)^2 \kappa}.$$

For fixed $\delta_u > 0$ and $\delta_\ell > 0$, the map $r_\phi(\delta_u, \delta_\ell, \kappa)$ is continuous in $\kappa > 1 + \delta_\ell$ and right-continuous at $\kappa = 1 + \delta_\ell$, hence so is $(1 - r_\phi(\delta_u, \delta_\ell, \kappa)^2)\sqrt{\kappa}$. Moreover, $(1 - r_\phi(\delta_u, \delta_\ell, \kappa)^2)\sqrt{\kappa} > 0$ for $\kappa \geq 1 + \delta_\ell$. Therefore, $1/((1 - r_\phi(\delta_u, \delta_\ell, \kappa)^2)^2 \kappa)$ is continuous in $\kappa > 1 + \delta_\ell$ and right-continuous at $\kappa = 1 + \delta_\ell$. Furthermore, $\lim_{\kappa \to +\infty} 1 + r_\phi(\delta_u, \delta_\ell, \kappa) = 2$ and

$$\begin{aligned}\lim_{\kappa \to +\infty} (1 - r_\phi(\delta_u, \delta_\ell, \kappa))\sqrt{\kappa} &= \lim_{\kappa \to +\infty} \frac{\delta_\ell \sqrt{\kappa} + \sqrt{\kappa(\kappa - 1)\delta_s} - \sqrt{\kappa(\kappa - (1 + \delta_\ell))\delta_u(1 + \delta_\ell)}}{\kappa - 1 + \sqrt{(\kappa - 1)\delta_s}} \\ &= \sqrt{\delta_s} - \sqrt{\delta_u(1 + \delta_\ell)},\end{aligned}$$

where $\delta_s = \delta_\ell + \delta_u + \delta_u \delta_\ell$. It follows that

$$\lim_{\kappa \to +\infty} \sigma_\phi(\delta_u, \delta_\ell, \kappa) = \lim_{\kappa \to +\infty} \frac{1}{((1 - r_\phi(\delta_u, \delta_\ell, \kappa)^2)^2 \kappa} = \frac{1}{4(\sqrt{\delta_s} - \sqrt{\delta_u(1 + \delta_\ell)})^2}.$$

Hence, $\kappa \mapsto \sigma_\phi(\delta_u, \delta_\ell, \kappa)$ attains a maximum on $[1 + \delta_\ell, \infty)$ and is bounded.

□

Figure 6 shows a plot of the map $\kappa \mapsto 1/((1 - r_\phi(\kappa)^2)^2 \kappa)$ for $\kappa = 10, \ldots, 10^9$ and the asymptotic value of σ_ϕ,

$$\lim_{\kappa \to +\infty} \sigma_\phi(\delta_u, \delta_\ell, \kappa) = \frac{1}{4(\sqrt{\delta_s} - \sqrt{\delta_u(1 + \delta_\ell)})^2} \approx 2.31,$$

for $\delta_u = 0.01$ and $\delta_\ell = 0.18$. We see that the asymptotic value of σ_ϕ is slightly less than the peak value of σ_ϕ, but the first still provides a good approximation to the second.

Building upon the two lemmas above, we now establish that $\phi((1 + \delta)\ell, \ell, m)$ is actually much faster than $r_{\text{acc}}(\sigma_\phi \kappa)$ for most values of ℓ.

Lemma B.11. *Given $\delta_u \in (0, 1]$, $\delta_\ell \in (0, 1]$ and $\kappa \geq 2$, there exist $\sigma_\phi = \sigma_\phi(\delta_u, \delta_\ell, \kappa) > 0$ and $\alpha_\phi = \alpha_\phi(\delta_u, \delta_\ell, m) > 0$ such that for all $\ell \in [(1 + \delta_\ell)m, L/(1 + \delta_u)]$*

$$\phi((1 + \delta_u)\ell, \ell, m)^2 \leq r_{\text{acc}}(\sigma_\phi \kappa)^{1 + \alpha_\phi(\ell - (1 + \delta_\ell)m)}, \tag{134}$$

where the function $\kappa \mapsto \sigma_\phi(\delta_u, \delta_\ell, \kappa)$ is bounded and satisfies

$$\lim_{\kappa \to +\infty} \sigma_\phi(\delta_u, \delta_\ell, \kappa) = \frac{1}{4(\sqrt{\delta_u + \delta_\ell + \delta_u \delta_\ell} - \sqrt{\delta_u(1 + \delta_\ell)})^2}.$$

Proof. Combining Lemmas B.9 and B.10, we have that

$$\phi((1 + \delta_u)\ell, \ell, m)^2 \leq r_{\text{acc}}(\sigma_\phi \kappa)$$

for all $\ell \in [(1 + \delta_\ell)m, L/(1 + \delta_u)]$. Moreover, $\phi((1 + \delta_u)\ell, \ell, m)$ is decreasing and continuously differentiable with respect to ℓ. So, consider the maximum slope of $\phi((1 + \delta_u \ell, \ell, m))$ over the interval $[(1 + \delta_\ell)m, L/(1 + \delta_u)]$:

$$s = \max_{\ell \in [(1 + \delta_\ell)m, L/(1 + \delta_u)]} \frac{\partial}{\partial \ell} \phi((1 + \delta_u)\ell, \ell, m) < 0.$$

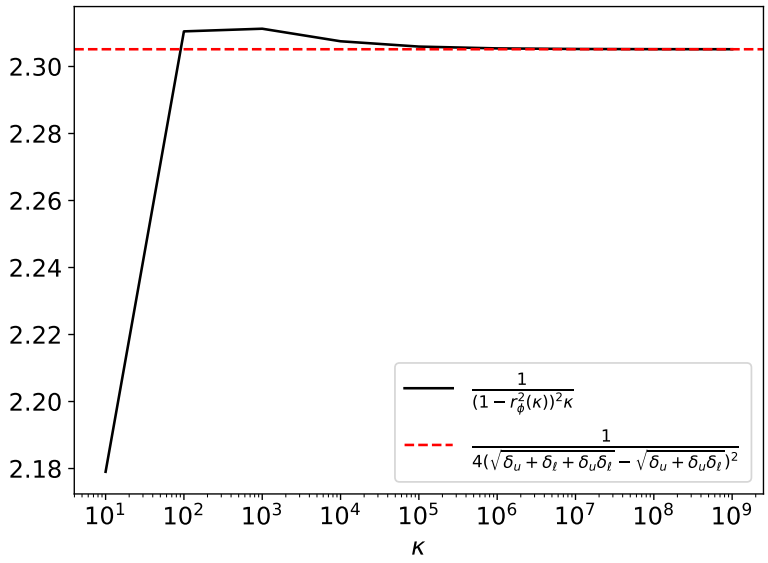


Figure 6: Numerical (black solid line) lower bound on and asymptotic value (dashed red line) of σ_ϕ such that $r_\phi^2 \leq r_{\text{acc}}(\sigma_\phi\kappa)$ holds for all $\kappa \geq 1 + 1/\delta_s$, with $\delta_u = 0.01$ and $\delta_\ell = 0.18$.

Then, it follows that for all $[(1+\delta_\ell)m, L/(1+\delta_u)]$

$$\frac{\partial}{\partial \ell}\phi((1+\delta_u\ell, \ell, m)) \leq s \leq a\sqrt{r_{\text{acc}}(\sigma_\phi)} \leq a\phi((1+\delta_u\ell, \ell, m)) < 0,$$

where $a = s/r_{\text{acc}}(\sigma_\phi\kappa) < 0$. Hence, Grönwall's inequality implies that

$$\phi((1+\delta_u\ell, \ell, m))^2 \leq \exp(2a(\ell - (1+\delta_\ell)m))r_{\text{acc}}(\sigma_\phi\kappa) = r_{\text{acc}}(\sigma_\phi\kappa)^{1+\alpha_\phi(\ell-(1+\delta_\ell)m)}, \quad (135)$$

where, since $a < 0$ and $\log_{r_{\text{acc}}(\sigma_\phi\kappa)} e < 0$, α_ϕ is a positive constant given by

$$\alpha_\phi = 2a \log_{r_{\text{acc}}(\sigma_\phi\kappa)} e > 0,$$

which proves the claim. □

Figure 7 illustrates Lemma B.11 numerically with $L = 10, m = 1, \delta_u = 1, \sigma_\phi = 4$ and $\alpha_\phi = 10$. We see that $\phi((1+\delta_u)\ell, \ell, m)^2$ becomes significantly smaller than $r_{\text{acc}}(\sigma_\phi\kappa)$ as ℓ approaches L. In fact, $\phi(L, L, m) = 0$, therefore the estimate adjustments take place extremely fast when the estimate is large and gradually slow down as the estimate improves, but always at an accelerated rate. As we now show, this implies drastic estimate convergence speed-up.

Lemma B.12. *Let $f \in \mathcal{F}(L, m)$ be a quadratic function, let $\bar{L} > L$, and suppose that Assumptions 4.2 and 4.5 hold. Also, let δ_m and δ_u be positive numbers such that $\delta_m > \delta_u > 0$ and let $r' = r'(\delta_u, \delta_\ell, \kappa)$ be a function such that $r_{\text{acc}}(\sigma_\phi\kappa) \leq r' < 1$ for all $\kappa \geq 1 + \delta_\ell$, where $\delta_\ell = ((1+\delta_m)/(1+\delta_u)) - 1 > 0$ and $\sigma_\phi = \sigma_\phi(\delta_u, \delta_\ell, \kappa)$ is given by Lemma B.10. Then, the estimates m_t of Algorithm 1 reach $[m/\gamma, (1+\delta_m)m]$ after no more than τ iterations, where*

$$\tau = \frac{-\log(4\kappa^2 M_1\omega/\delta_u)}{\log r'} \sum_{j=0}^{\log_\gamma(\kappa/(1+\delta_m))} \frac{1}{1+\alpha_\phi m(1+\delta_\ell)(\gamma^j - 1)}, \quad (136)$$

$M_1 = \max_i \overline{C}_i/\underline{C}_1$, with ω and γ given by Assumptions 4.2 and 4.5.

Proof. Suppose that the last value of m_t before reaching the interval $[m/\gamma, (1+\delta_m)m)$ is $(1+\delta_m)m$. Then, suppose that the value before last is $\gamma(1+\delta_m)m$, and so on, up to $\gamma^K(1+\delta_m)m$ for some K such that $\gamma^K(1+\delta_m)m \leq L < \gamma^{K+1}(1+\delta_m)m$. Using this m_t schedule, we bound the number of

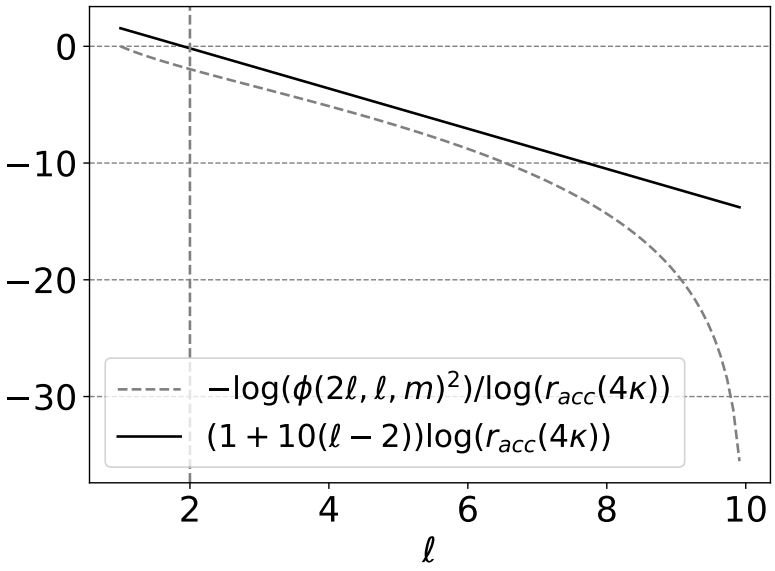


Figure 7: Numerical illustration of Lemma B.11 with $L = 10, m = 1, \delta_\ell = 1, \sigma_\phi = 4$ and $\alpha_\phi = 10$.

iterations that m_t takes to reach the interval $[m/\gamma, (1+\delta_m)m]$, and then we argue that no other m_t schedule can lead to a worse bound.

Let $\ell_j = \gamma^j(1+\delta_m)m/(1+\delta_u)$. Then, we have that $\ell_j \geq (1+\delta_\ell)m$ for $\delta_\ell = ((1+\delta_m)/(1+\delta_u))-1$. Since $\delta_m > \delta_u$, then $\delta_\ell > 0$, and Lemma B.11 applies. Now, let $I_j = \min\{i : \lambda_i \geq \ell_j\}$. That is, $\lambda_i \geq \ell_j$ if and only if $i \geq I_j$. Then, using this fact and separating the terms indexed by $i < I_0$ from those indexed by $i \geq I_0$ in (128) into two sums yields

$$c_{t+1}^2 < \ell_0^2 \frac{\sum_{i=1}^{I_0-1}(x_{i,t+1}-x_{i,t})^2}{\sum_{i=1}^{d}(x_{i,t+1}-x_{i,t})^2} + \frac{\sum_{i=I_0}^{d}\lambda_i^2(x_{i,t+1}-x_{i,t})^2}{\sum_{i=1}^{d}(x_{i,t+1}-x_{i,t})^2}.$$

In turn, plugging the above inequality into the identity $(c_{t+1}+\ell_0)(c_{t+1}-\ell_0) = c_{t+1}^2 - \ell_0^2$, and then using the fact that $\lambda_i \leq L$ and $\ell_0 \geq m$, we obtain

$$c_{t+1} - \ell_0 = \frac{c_{t+1}^2 - \ell_0^2}{c_{t+1}+\ell_0} < \frac{\sum_{i=I_0}^{d}(\lambda_i^2-\ell_0^2)(x_{i,t+1}-x_{i,t})^2}{\ell_0\sum_{i=1}^{d}(x_{i,t+1}-x_{i,t})^2} \leq \ell_0\kappa^2 \sum_{i=I_0}^{d}\frac{(x_{i,t+1}-x_{i,t})^2}{(x_{1,t+1}-x_{1,t})^2}. \quad (137)$$

Moreover, using (124), we have that

$$(x_{i,t+1}-x_{i,t})^2 = ([-1 \quad 1]\,X_{i,t+1})^2 \leq 2\|X_{i,t+1}\|^2 \leq 2\overline{C}_i\rho(\mu_0,\lambda_i)^{2t}x_{i,0}^2. \quad (138)$$

To address the terms in the sum in (137), we combine (138) and (127), assuming $t_K \leq t < t_{K+1}$. That is, we consider the last adjustment before m_t reaches the interval $[m/\gamma, (1+\delta_m)m)$. Then, we apply Lemma B.3 twice, to get $\rho(m_k,\lambda_i) \leq \rho(m_k,\ell_0)$ and $\rho(m_k,\ell_0) < \rho(m_k,m)$ for all $i \geq I_0$, which gives

$$\begin{aligned}\sum_{i=I_0}^{d}\frac{(x_{i,t+1}-x_{i,t})^2}{(x_{1,t+1}-x_{1,t})^2} &\leq 2M_1\sum_{i=I_0}^{d}\frac{x_{i,0}^2}{x_{1,0}^2}\prod_{k=t_K}^{t}\frac{\rho(m_k,\lambda_i)^2}{\rho(m_k,m)^2}\prod_{k=1}^{\tau_K}\frac{\rho(m_k,\lambda_i)^2}{\rho(m_k,m)^2}\\ &\leq 2M_1\omega\phi((1+\delta_u)\ell_0,\ell_0,m)^{2(t-t_K+1)}. \quad (139)\end{aligned}$$

where $M_1 = 2\max_i \overline{C}_i/\underline{C}_1$. Next, we put (137) and (139) together, and since $\ell_0 \geq m$, we get

$$c_{t+1} - \ell_0 < 2\kappa^2 M_1\omega\phi((1+\delta_m)m,\ell_0,m)^{2(t-t_K+1)}\ell_0 \leq \delta_u\ell_0/2,$$

for all $t \geq t_K + \Delta t_0$, where

$$\Delta t_0 = -\frac{\log(4\kappa^2 M_1\omega/\delta_u)}{\log r_{\text{acc}}(\sigma_\phi\kappa)}.$$

Therefore, $c_{t+1} < (1+\delta_u)\ell_0 = (1+\delta_m)m$ for $t \geq t_K + \Delta t_0$ or, equivalently,

$$t_{K+1} - t_K \leq \Delta t_0.$$

Note that for every m_t schedule, if μ_K denotes the last value of m_t before reaching $[m/\gamma, (1+\delta_m)m]$, then $\mu_K \geq (1+\delta_m)m$, by definition. Hence, letting $\ell_0' = \mu_K/(1+\delta_u)$ and $I_0' = \{i : \lambda_i \geq \ell_0'\}$, then $I_0' \geq I_0$, and it follows that

$$\begin{aligned}
\sum_{i=I_0'}^{d} \frac{(x_{i,t+1}-x_{i,t})^2}{(x_{1,t+1}-x_{1,t})^2} &\leq 2M_1 \sum_{i=I_0'}^{d} \frac{x_{i,0}^2}{x_{1,0}^2} \prod_{k=t_K}^{t} \frac{\rho(m_k,\lambda_i)^2}{\rho(m_k,m)^2} \prod_{k=1}^{\tau_K} \frac{\rho(m_k,\lambda_i)^2}{\rho(m_k,m)^2} \\
&\leq 2M_1 \sum_{i=I_0}^{d} \frac{x_{i,0}^2}{x_{1,0}^2} \prod_{k=t_K}^{t} \frac{\rho(m_k,\lambda_i)^2}{\rho(m_k,m)^2} \prod_{k=1}^{\tau_K} \frac{\rho(m_k,\lambda_i)^2}{\rho(m_k,m)^2} \\
&\leq 2M_1 \omega \phi((1+\delta_u)\ell_0, \ell_0, m)^{2(t-t_K+1)}.
\end{aligned}$$

Therefore, the last adjustment cannot take more than Δt_0 iterations for any m_t schedule.

Then, let Δt_j be quantities analogous to Δt_0, defined for $j = 0, \ldots, K$ as

$$\Delta t_j = -\frac{\log(4\kappa^2 M_1 \omega/\delta_u)}{(1+\alpha_\phi(\ell_j - (1+\delta_\ell)m))\log r_{\text{acc}}(\sigma_\phi \kappa)}.$$

By Assumption 4.2, m_t decreases by a factor of at least γ every time it is adjusted to a new value. Hence, $\mu_{K-1} \geq \gamma \mu_K$ for every m_t schedule, which implies that $\ell_1 \leq \mu_{K-1}/(1+\delta_u)$ for every m_t schedule. Hence, by the same rationale above, it cannot take more than Δt_1 for m_t to be adjusted to its second last value before reaching the interval $[m/\gamma, (1+\delta_m)m]$. It follows by induction that it cannot take more than Δt_j for m_t to be adjust to its $K-j$-th to last value before reaching the interval $[m/\gamma, (1+\delta_m)m]$. Moreover, since by design $m_t \leq L$, it cannot more than $K \leq \log_\gamma(\kappa/(1+\delta_m))$ adjustments before m_t reaches the interval $[m/\gamma, (1+\delta_m)m]$. Therefore, we conclude that $m_{t+1} \leq (1+\delta_m)m$ for all $t \geq \tau$, where

$$\begin{aligned}
\tau &= \frac{-\log(4\kappa^2 M_1 \omega/\delta_u)}{\log r'} \sum_{j=0}^{\log_\gamma(\kappa/(1+\delta_m))} \frac{1}{1+\alpha_\phi m(1+\delta_\ell)(\gamma^j - 1)} \\
&\geq \frac{-\log(4\kappa^2 M_1 \omega/\delta_u)}{\log r_{\text{acc}}(\sigma_\phi \kappa)} \sum_{j=0}^{K} \frac{1}{1+\alpha_\phi m(1+\delta_\ell)(\gamma^j - 1)},
\end{aligned}$$

because log is monotone and $r_{\text{acc}}(\sigma_\phi \kappa) \leq r' < 1$, and $K \leq \log_\gamma(\kappa/(1+\delta_m))$. □

B.1.5 Main result in the quadratic case

We now prove the main local convergence result for NAG-free when the objective function is quadratic. There is no difference between local and global convergence in this case, but it will be the foundation to derive the main local convergence in the general case later. To this end, we first establish that for every $G_i(m_t)$, there is a quadratic Lyapunov function certifying convergence of $X_{i,t}$ at rate $\rho(m_t, \lambda_i)$ up to arbitrary precision, at the expense of worse condition numbers.

Lemma B.13. *Let $m_t \in [m/\gamma, L]$ for some $\gamma > 1$, and let $\rho(G_i(m_t))$ denote the spectral radius of $G_i(m_t)$. Then, given $r \in [\rho(G_i(m_t)), 1)$ and $\delta > 0$ such that $(1+\delta)r < 1$, there is some $P = P(G_i(m_t), r, \delta) \in \mathbb{R}^{d\times d}$ such that $G_i(m_t)^\mathsf{T} P G_i(m_t) \prec (1+\delta)^2 r^2 P$ and $P \succeq I$. Moreover, letting $\lambda_{\min}(P)$ and $\lambda_{\max}(P)$ denote the least and the greatest eigenvalues of P, then*

$$\max_{m_t \in [m/\gamma, L]} \|P(G_i(m_t), r, \delta)\| < \frac{1+(1+\delta)^{-2}}{1-(1+\delta)^{-2}} + \frac{2M_2^2}{(1+\delta)^2 r^2} \frac{1+(1+\delta)^{-2}}{(1-(1+\delta)^{-2})^3}, \tag{140}$$

where M_2 is an appropriate constant that does not depend on neither δ nor r.

Proof. By Lemma B.4, $\rho(m_t, \lambda_i) < 1$ for all $m_t \in (0, L]$ and $i = 1, \ldots, d$. Thus, $\rho(G_i(m_t)) < 1$, where $\rho(G_i(m_t))$ denotes the spectral radius of $G_i(m_t)$. Therefore, the interval $[\rho(G_i(m_t)), 1)$ is nonempty. So, let r and δ be two positive numbers such that $r \in [\rho(G_i(m_t)), 1)$ and $(1+\delta)r < 1$.

Then, take $\nu = (1+\delta)r$ and $P = \sum_{k=0}^{+\infty} (G_i(m_t)^\mathsf{T}/\nu)^k (G_i(m_t)/\nu)^k$. The matrix P is well-defined because $\rho(G_i(m_t)/\nu) \leq 1/(1+\delta) < 1$, and $P \succeq I$, by construction. Moreover, it satisfies

$$(G_i(m_t))/\nu)^\mathsf{T} P (G_i(m_t)/\nu) = \sum_{k=1}^{+\infty} (G_i(m_t))^\mathsf{T}/\nu)^k (G_i(m_t))/\nu)^k = P - I.$$

Therefore, $G_i(m_t)^\mathsf{T} P G_i(m_t) \prec (1+\delta)^2 r^2 P$, which proves the first claim.

To prove the second claim, we first express $G_i(m_t)$ in Schur form [Golub and Van Loan, 2013, 7.1.3]. To this end, we construct a two-by-two orthogonal matrix $Q_i(m_t)$, whose first column is a unit eigenvector $q_{i,1}$ associated with $\zeta_i = \zeta_i(m_t)$, the top eigenvalue of $G_i(m_t)$, as in

$$q_{i,1} = \frac{1}{\sqrt{1+|\zeta_i|^2}} \begin{bmatrix} 1 \\ \zeta_i \end{bmatrix}.$$

To determine the second column of $Q_i(m_t)$, we apply the Gram-Schmidt orthogonalization procedure [Golub and Van Loan, 2013, 5.2.7] to obtain from e_1 a vector orthonormal to $q_{i,1}$:

$$e_1 - \frac{\langle e_1, q_{i,1} \rangle}{\langle q_{i,1}, q_{i,1} \rangle} q_{i,1} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} - \frac{1}{1+|\zeta_i|^2} \begin{bmatrix} 1 \\ \zeta_i \end{bmatrix} = \frac{1}{1+|\zeta_i|^2} \begin{bmatrix} |\zeta_i|^2 \\ -\zeta_i \end{bmatrix}.$$

Normalizing the vector above, we obtain

$$q_{i,2} = \frac{1}{\sqrt{1+|\zeta_i|^2}} \frac{1}{|\zeta_i|} \begin{bmatrix} |\zeta_i|^2 \\ -\zeta_i \end{bmatrix}.$$

So, letting $Q_i(m_t)$ be the orthogonal matrix given by

$$Q_i(m_t) = \begin{bmatrix} q_{i,1} & q_{i,2} \end{bmatrix}, \tag{141}$$

and letting $T_i(m_t)$ be the matrix given by

$$T_i(m_t) = Q_i(m_t)^\mathsf{H} G_i(m_t) Q_i(m_t), \tag{142}$$

where $Q_i(m_t)^\mathsf{H}$ denotes the conjugate-transpose of $Q_i(m_t)$, it follows that

$$\begin{aligned} T_i(m_t) &= \begin{bmatrix} q_{i,1}(m_t)^\mathsf{H} G_i(m_t) q_{i,1}(m_t) & q_{i,1}(m_t)^\mathsf{H} G_i(m_t) q_{i,2}(m_t) \\ q_{i,2}(m_t)^\mathsf{H} G_i(m_t) q_{i,1}(m_t) & q_{i,2}(m_t)^\mathsf{H} G_i(m_t) q_{i,2}(m_t) \end{bmatrix} \\ &= \begin{bmatrix} \zeta_i & q_{i,1}(m_t)^\mathsf{H} G_i(m_t) q_{i,2}(m_t) \\ \zeta_i q_{i,2}(m_t)^\mathsf{H} q_{i,1}(m_t) & q_{i,2}(m_t)^\mathsf{H} G_i(m_t) q_{i,2}(m_t) \end{bmatrix} \\ &= \begin{bmatrix} \zeta_i & q_{i,1}(m_t)^\mathsf{H} G_i(m_t) q_{i,2}(m_t) \\ 0 & q_{i,2}(m_t)^\mathsf{H} G_i(m_t) q_{i,2}(m_t) \end{bmatrix} \end{aligned}$$

because $q_{i,1}(m_t)$ is a unit eigenvector of $G_i(m_t)$ associated with ζ_i, and is orthogonal to $q_{i,2}(m_t)$. Moreover, the product $Q_i(m_t)^\mathsf{H} G_i(m_t) Q_i(m_t)$ preserves the eigenvalues of $G_i(m_t)$ because $Q_i(m_t)$ is orthogonal, therefore

$$T_i(m_t) = \begin{bmatrix} \zeta_i & q_{i,1}(m_t)^\mathsf{H} G_i(m_t) q_{i,2}(m_t) \\ 0 & \xi_i \end{bmatrix}, \tag{143}$$

where ξ_i denotes the other eigenvalue of $G_i(m_t)$. Now, $G_i(m_t)$ and, therefore, $Q_i(m_t)$ are continuous functions of m_t, thus

$$M_2 = \max_{m_t \in [m/\gamma, L]} q_{i,1}(m_t)^\mathsf{H} G_i(m_t) q_{i,2}(m_t) < +\infty$$

is well-defined. Moreover, left-multiplying and right-multiplying (143) by $Q_i(m_t)$ and $Q_i(m_t)^\mathsf{H}$, respectively, yields

$$Q_i(m_t) T_i(m_t) Q_i(m_t)^\mathsf{H} = T_i(m_t).$$

Substituting the above for $G_i(m_t)$, using submultiplicativity of the Euclidean norm, the fact that $Q_i(m_t)$ are orthogonal, and the fact that $\rho(m_t, \lambda_i)/\nu \leq 1/(1+\delta)$, where $\nu = (1+\delta)r$, we get

$$\begin{aligned}
\|P\| &\leq \sum_{k=0}^{+\infty} \|((G_i(m_t)/\nu)^k)^{\mathsf{T}}(G_i(m_t^k)/\nu)\| \\
&\leq \sum_{k=0}^{+\infty} \|Q_i(m_t)(T_i(m_t)/\nu)^k Q_i(m_t)^{\mathsf{H}}\|^2 \\
&\leq 1 + \sum_{k=0}^{+\infty} \|Q_i(m_t)^{\mathsf{H}}\|^2 \|T_i(m_t)/\nu\|^{2k} \|Q_i(m_t)\|^2 \\
&\leq 1 + \sum_{k=0}^{+\infty} \nu^{-2(k+1)} \left\| \begin{bmatrix} \rho(m_t, \lambda_i)^{k+1} & (k+1)M_2\rho(m_t, \lambda_i)^k \\ 0 & \rho(m_t, \lambda_i)^{k+1} \end{bmatrix} \right\|^2 \\
&\leq 1 + \sum_{k=0}^{+\infty} \nu^{-2(k+1)} \left(\rho(m_t, \lambda_i)^{k+1} + (k+1)M_2\rho(m_t, \lambda_i)^k \right)^2 \\
&= 1 + \sum_{k=0}^{+\infty} \left((1+\delta)^{-(k+1)} + \frac{M_2}{\nu}(k+1)(1+\delta)^{-k} \right)^2.
\end{aligned}$$

Then, using the fact that $(a+b)^2 \leq 2a^2 + 2b^2$ for any a and b, yields

$$\begin{aligned}
\|P\| &\leq 1 + \sum_{k=0}^{+\infty} \left(2(1+\delta)^{-2(k+1)} + \frac{2M_2^2}{\nu^2}(k+1)^2(1+\delta)^{-2k} \right) \\
&\leq 1 + \frac{2(1+\delta)^{-2}}{1-(1+\delta)^{-2}} + \frac{2M_2^2}{\nu^2} \frac{1+(1+\delta)^{-2}}{(1-(1+\delta)^{-2})^3} \\
&= \frac{1+(1+\delta)^{-2}}{1-(1+\delta)^{-2}} + \frac{2M_2^2}{\nu^2} \frac{1+(1+\delta)^{-2}}{(1-(1+\delta)^{-2})^3},
\end{aligned}$$

where the second inequality follows by noting that for any α such that $0 < \alpha < 1$, we have that

$$\sum_{k=0}^{+\infty} (k+1)^2 \alpha^k = \frac{1}{\alpha} \sum_{k=1}^{+\infty} k^2 \alpha^k = \frac{1}{\alpha} \frac{\alpha(1+\alpha)}{(1-\alpha)^3} = \frac{1+\alpha}{(1-\alpha)^3},$$

and then plugging $(1+\delta)^{-2}$ into α. □

Proposition B.14. *Let $f \in \mathcal{F}(L, m)$ be a quadratic function with $\kappa = L/m \geq 4$, and let $\bar{L} > L$. Suppose that Assumption 4.2 holds for some $\gamma \geq 2$, Assumption 4.5 holds for some $\omega > 0$, and that Assumption 4.4 holds as well. Also, let δ_m and δ_u be positive numbers such that $\delta_u < \min\{\delta_m, 1/2\}$ and $\delta_m \leq \gamma - 1$. If Algorithm 1 receives $\bar{L}$ as input, then its iterates x_t satisfy*

$$\|x_{t+1} - x^\star\| \leq C\bar{\kappa}^{2(1+\nu)} r_{\text{acc}}(2\sigma\bar{\kappa})^t \|x_0 - x^\star\|, \tag{144}$$

where $\bar{\kappa} = \bar{L}/m > \kappa$, $\sigma = \max\{\gamma, \sigma_m, \sigma_\phi\}$, $\sigma_m = 1 + 2\delta_m + 2\sqrt{\delta_m(1+\delta_m)}$, $\sigma_\phi = \sigma_\phi(\delta_u, \delta_\ell, \bar{\kappa})$ is a function that depends on $\delta_\ell = (1+\delta_m)/(1+\delta_u) - 1$ and is bounded in $\bar{\kappa} \geq 1 + \delta_\ell$, such that

$$\lim_{\bar{\kappa} \to +\infty} \sigma_\phi(\delta_u, \delta_\ell, \bar{\kappa}) = \frac{1}{4(\sqrt{\delta_u(1+\delta_\ell) + \delta_\ell} - \sqrt{\delta_u(1+\delta_\ell)})^2},$$

and C is a constant factor that depends on γ, δ_u, σ and ω.

Proof. Let $r' = r'(\delta_u, \delta_\ell, \kappa)$ be a function such that $r_{\text{acc}}(\sigma_\phi\kappa) \leq r' < 1$ for all $\kappa \geq 1 + \delta_\ell$, where $\sigma_\phi = \sigma_\phi(\delta_u, \delta_\ell, \kappa)$ is given by Lemma B.10. Then, by Lemma B.12 we have that $m_t \leq (1+\delta_m)m$ for all $t \geq \tau$, where

$$\tau = \frac{-\log(4\kappa^2 M_1 \omega/\delta_u)}{\log r'} \sum_{j=0}^{\log_\gamma(\kappa/(1+\delta_m))} \frac{1}{1 + \alpha_\phi m(1+\delta_\ell)(\gamma^j - 1)}, \tag{145}$$

$M_1 = \max_i \overline{C}_i/\underline{C}_1$.

Assumption 4.2 implies that then $m_t \geq m/\gamma$. If $m_t < m$, then by Lemma B.1, and unsing the fact that $(\sqrt{L/m_t}-1)/(\sqrt{L/m_t}+1)$ is decreasing in m_t and $(\kappa-1)/\kappa$ is increasing in κ, we get

$$\rho(m_t, m) = \sqrt{\frac{\sqrt{L/m_t}-1}{\sqrt{L/m_t}+1}\frac{\kappa-1}{\kappa}} \leq \sqrt{\frac{\sqrt{\gamma\kappa}-1}{\sqrt{\gamma\kappa}+1}\frac{\gamma\kappa-1}{\gamma\kappa}} = r_{\text{acc}}(\gamma\kappa).$$

Otherwise, if $m_t \in [m, (1+\delta_m)m]$, then by Corollary B.8, we have that $\rho(m_t, m) \leq r_{\text{acc}}(\sigma_m\kappa)$ for all $\kappa \geq 1+\delta_m$, where $\sigma_m = \sigma_m(\delta_m) = 1 + 2\delta_m + 2\sqrt{\delta_m(1+\delta_m)}$. Hence, $\rho(m_t, m) \leq r_{\text{acc}}(\sigma_1\kappa)$ for all $t \geq \tau$, where $\sigma_1 = \max\{\gamma, \sigma_m\}$.

Now, by Lemma B.3, we have that $\rho(m_t, \lambda_i) \leq r_{\text{acc}}(\sigma_1\kappa)$ for all λ_i. Hence, given δ_σ such that $(1+\delta_\sigma)r_{\text{acc}}(\sigma_1\kappa) < 1$, by Lemma B.13 there is a $P_i(m_t) = P_i(m_t, \delta_\sigma) \succeq I$ for each λ_i such that $G_i(m_t)^\mathsf{T} P_i(m_t) G_i(m_t) \preceq (1+\delta_\sigma)^2 r_{\text{acc}}(\sigma_1\kappa)^2 P_i(m_t)$. Hence, if $t_j \leq t < t_{j+1}$ and $t \geq \tau$, then

$$\begin{aligned} X_{i,t+1}^\mathsf{T} P_i(m_t) X_{i,t+1} &= X_{i,t}^\mathsf{T} G_i(\mu_j)^\mathsf{T} P_i(\mu_j) G_i(\mu_j) X_{i,t} \\ &\leq (1+\delta_\sigma)^2 r_{\text{acc}}(\sigma_1\kappa)^2 X_{i,t}^\mathsf{T} P_i(\mu_j) X_{i,t}, \end{aligned}$$

since $m_t = \mu_j$. Consecutively applying this inequality, we obtain

$$\begin{aligned} \lambda_{\min}(P_i(m_t))\|X_{i,t}\|^2 \leq X_{i,t}^\mathsf{T} P_i(m_t) X_{i,t} &\leq ((1+\delta_\sigma)r_{\text{acc}}(\sigma_1\kappa))^{2(t-t_j)} X_{i,t_j}^\mathsf{T} P_i(\mu_j) X_{i,t_j} \\ &\leq \lambda_{\max}(P_i(\mu_j))((1+\delta_\sigma)r_{\text{acc}}(\sigma_1\kappa))^{2(t-t_j)}\|X_{i,t_j}\|^2. \end{aligned}$$

Rearranging the above yields

$$\|X_{i,t}\|^2 \leq \frac{\lambda_{\max}(P_i(\mu_j))}{\lambda_{\min}(P_i(\mu_j))}((1+\delta_\sigma)r_{\text{acc}}(\sigma_1\kappa))^{2(t-t_j)}\|X_{i,t_j}\|^2.$$

Moreover, since by assumption $1+\delta_m < \gamma$, m_t is adjusted at most once if $m_t \leq (1+\delta_m)m$, therefore denoting by μ_{-1} and μ_{-2} respectively the last and before last values taken by m_t, for all $t \geq \tau$ we have that

$$\|X_{i,t}\|^2 \leq \frac{\lambda_{\max}(P_i(\mu_{-1}))}{\lambda_{\min}(P_i(\mu_{-1}))}\frac{\lambda_{\max}(P_i(\mu_{-2}))}{\lambda_{\min}(P_i(\mu_{-2}))}((1+\delta_\sigma)r_{\text{acc}}(\sigma_1\kappa))^{2(t-\lceil\tau\rceil)}\|X_{i,\lceil\tau\rceil}\|^2,$$

In turn, the above bound yields

$$\begin{aligned} \|X_t\|^2 &= \sum_{i=1}^d \|X_{i,t}\|^2 \\ &\leq \sum_{i=1}^d \frac{\lambda_{\max}(P_i(\mu_{-1}))}{\lambda_{\min}(P_i(\mu_{-1}))}\frac{\lambda_{\max}(P_i(\mu_{-2}))}{\lambda_{\min}(P_i(\mu_{-2}))}((1+\delta_\sigma)r_{\text{acc}}(\sigma_1\kappa))^{2(t-\lceil\tau\rceil)}\|X_{i,\lceil\tau\rceil}\|^2. \end{aligned}$$

Since r_{acc} is monotone and $\sum_{i=1}^d \|X_{i,t}\|^2 = \|X_t\|^2$, defining $\sigma = \max\{\gamma, \sigma_m, \sigma_\phi\}$, it follows that

$$\|X_t\|^2 \leq M_3^2((1+\delta_\sigma)r_{\text{acc}}(\sigma\kappa))^{2(t-\lceil\tau\rceil)}\|X_{\lceil\tau\rceil}\|^2, \tag{146}$$

where M_3^2 is given by the product of the worst condition numbers of all $P_i(\mu_{-1})$ and $P_i(\mu_{-2})$:

$$M_3 = \sqrt{\max_{i=1,\dots,d}\frac{\lambda_{\max}(P_i(\mu_{-1}))}{\lambda_{\min}(P_i(\mu_{-1}))}\max_{i=1,\dots,d}\frac{\lambda_{\max}(P_i(\mu_{-2}))}{\lambda_{\min}(P_i(\mu_{-2}))}}.$$

Plugging $r' = (1+\delta_\sigma)r_{\text{acc}}(\sigma\kappa)$ in (145), it follows that $m_t \in [m/\gamma, (1+\delta_m)m]$ for all $t \geq \tau$, where

$$\tau = \frac{-\log(4\kappa^2 M_1\omega/\delta_u)}{\log(1+\delta_\sigma)r_{\text{acc}}(\sigma\kappa)} \sum_{j=0}^{\log_\gamma(\kappa/(1+\delta_m))} \frac{1}{1+\alpha_\phi m(1+\delta_\ell)(\gamma^j-1)}.$$

By assumption, $\gamma \geq 2$, which implies that $\gamma^j - 1 \geq \gamma^{j-1}$ for all $j \geq 1$, so that

$$\begin{aligned}\tau &\leq \frac{-\log(4\kappa^2 M_1 \omega/\delta_u)}{\log(1+\delta_\sigma) r_{\text{acc}}(\sigma\kappa)} \left(1 + \frac{1}{\alpha_\phi m (1+\delta_\ell)} \sum_{j=1}^{\log_\gamma(\kappa/(1+\delta_m))} \frac{1}{\gamma^{j-1}}\right) \\ &\leq \frac{-\log(4\kappa^2 M_1 \omega/\delta_u)}{\log(1+\delta_\sigma) r_{\text{acc}}(\sigma\kappa)} \left(1 + \frac{1}{\alpha_\phi m} \frac{\gamma}{\gamma - 1}\right)\end{aligned}$$

Therefore, since $r_{\text{acc}}(\kappa) \geq 1/2$, $r_{\text{acc}}(\kappa) \leq r_{\text{acc}}(\sigma\kappa) \in (0,1)$ and $\lceil \tau \rceil \leq \tau + 1$, it follows that

$$((1+\delta_\sigma) r_{\text{acc}}(\sigma\kappa))^{-\lceil \tau \rceil} \leq M_4, \tag{147}$$

for a constant M_4 given by

$$M_4 = (4\kappa^2 M_1 \omega/\delta_u)^\nu, \quad \text{where } \nu = 1 + \gamma/(\alpha_\phi m(\gamma - 1)).$$

Then, plugging (147) into (146), we obtain

$$\|x_t\| \leq \|X_t\| \leq M_3 M_4 ((1+\delta_\sigma) r_{\text{acc}}(\sigma\kappa))^t \|X_{\lceil \tau \rceil}\|.$$

To establish (144), it remains to bound $\|X_{\lceil \tau \rceil}\|$. To this end, we plug $x = x^\star$ and $y = y_{t+1}$ into (3), and then use the global convergence bound from (4.1) to get

$$\|y_{t+1} - x^\star\|^2 \leq \frac{2}{m}(f(y_{t+1}) - f(x^\star)) \leq 4\kappa r_{\text{GD}}(\kappa)^t \|x_0 - x^\star\|^2.$$

Then, substituting x_{t+1} with its definition from Algorithm 1, summing $\pm\beta_t x^\star = 0$, using the above bound and then the fact that $\beta_t \in [0,1)$ and that $r_{\text{GD}} \in (0,1)$, we obtain

$$\begin{aligned}\|x_{t+1} - x^\star\|^2 = \|(1+\beta_t) y_{t+1} - \beta_t y_t - x^\star \pm \beta_t x^\star\|^2 &= \|(1+\beta_t)(y_{t+1} - x^\star) - \beta_t (y_t - x^\star)\|^2 \\ &\leq (2\|y_{t+1} - x^\star\| + \|y_t - x^\star\|)^2 \\ &\leq 36\kappa r_{\text{GD}}(\kappa)^{t-1} \|x_0 - x^\star\|^2, \end{aligned} \tag{148}$$

which implies that

$$\|X_{\lceil \tau \rceil}\| \leq \|x_{\lceil \tau \rceil}\| + \|x_{\lceil \tau \rceil - 1}\| \leq 12\sqrt{\kappa} r_{\text{GD}}(\kappa)^{(\lceil \tau \rceil - 3)/2} \|x_0 - x^\star\| \leq 12\sqrt{\kappa} \|x_0 - x^\star\|.$$

If $\lceil \tau \rceil \geq 3$, then

$$\|X_{\lceil \tau \rceil}\| \leq 12\sqrt{\kappa} \|x_0 - x^\star\|.$$

Otherwise, if $\lceil \tau \rceil < 3$, then using the fact that $r_{\text{acc}}(\sigma\kappa) \geq r_{\text{acc}}(\kappa) = 1/2$, which follows from the assumption that $\kappa \geq 4$, and the assumption that $\gamma \geq 2$, we obtain

$$((1+\delta_\sigma) r_{\text{acc}}(\sigma\kappa))^{-\lceil \tau \rceil} \leq r_{\text{acc}}(\sigma\kappa)^{-3} \leq \frac{\gamma^2}{r_{\text{acc}}(\sigma\kappa)}.$$

Moreover, since the following equivalences hold for $\kappa \geq 4$

$$r_{\text{GD}}(\kappa)^2 \geq r_{\text{acc}}(\kappa) \iff \frac{(\kappa-1)^2}{\kappa^2} \geq \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa}} \iff \kappa^2 - 2\kappa\sqrt{\kappa} + \sqrt{\kappa} \geq 0,$$

we have that

$$r_{\text{GD}}(\kappa)^{-3/2} \leq r_{\text{acc}}(\kappa)^{-1} \leq \delta_u^{-1}.$$

Combining the two bounds above yields

$$((1+\delta_\sigma) r_{\text{acc}}(\sigma\kappa))^{-\lceil \tau \rceil} r_{\text{GD}}(\kappa)^{-3/2} \leq M_4.$$

Therefore, for all values of $\lceil \tau \rceil$, we have that

$$\|x_{t+1} - x^\star\| \leq M_3 M_4 \sqrt{\kappa} ((1+\delta_\sigma) r_{\text{acc}}(\sigma\kappa))^t \|x_0 - x^\star\|. \tag{149}$$

Our next step is to express the rate $(1+\delta_\sigma) r_{\text{acc}}(\sigma\kappa)$ in terms of $r_{\text{acc}}(\sigma_2\kappa)$ for some σ_2, as in

$$(1+\delta_\sigma) r_{\text{acc}}(\sigma\kappa) = (1+\delta_\sigma) \frac{\sqrt{\sigma\kappa} - 1}{\sqrt{\sigma\kappa}} = \frac{\sqrt{\sigma_2\kappa} - 1}{\sqrt{\sigma_2\kappa}}.$$

Solving the above identity for σ_2, we obtain

$$\sigma_2 = \frac{\sigma}{(1+\delta_\sigma - \delta_\sigma\sqrt{\sigma\kappa})^2} \leq \frac{\sigma}{(1-\delta_\sigma\sqrt{\sigma\kappa})^2}.$$

That is, $\sigma_2 = (1+\delta)\sigma$, where

$$\delta = \frac{\delta_\sigma\sqrt{\sigma\kappa}(2-\delta_\sigma\sqrt{\sigma\kappa})}{(1-\delta_\sigma\sqrt{\sigma\kappa})^2}.$$

So, if $\delta_\sigma = 1/(4\sqrt{\sigma\kappa})$, then $\delta \leq 7/9$, which implies that $1+\delta \leq 2$ and

$$(1+\delta_\sigma) r_{\text{acc}}(\sigma\kappa) \leq r_{\text{acc}}(2\sigma\kappa).$$

Moreover, since $\sigma \geq \gamma \geq 2$ and $\kappa \geq 4$, it follows that $\delta_\sigma = 1/(4\sqrt{\sigma\kappa}) \leq 1/11$ and

$$\begin{aligned}
\frac{1}{1-(1+\delta_\sigma)^{-2}} &= \frac{(1+\delta_\sigma)^2}{\delta_\sigma(2+\delta_\sigma)} \leq \frac{(1+1/11)^2}{2}\frac{1}{\delta_\sigma} = \frac{12^2}{2\cdot 11^2}\frac{1}{\delta_\sigma}, \\
1-(1+\delta_\sigma)^{-2} &= \frac{\delta_\sigma(2+\delta_\sigma)}{(1+\delta_\sigma)^2} \leq \delta_\sigma(2+\sigma_\delta) = \frac{1}{4\sqrt{\sigma\kappa}}(2+1/(4\sqrt{\sigma\kappa})) \leq \frac{9}{11^2}, \\
1+(1+\delta_\sigma)^{-2} &= \frac{2+\delta_\sigma(2+\delta_\sigma)}{1+\delta_\sigma} \leq 2(1+\delta_\sigma).
\end{aligned}$$

In the same vein, using the fact that $\lambda_{\min}(P_i(\mu_{-1})) \geq 1$ and that $\lambda_{\max}(P_i(\mu_{-1})) \leq \|P_i(\mu_{-1})\|$, plugging $\delta_\sigma = 1/(4\sqrt{\sigma\kappa})$ into (140) and using the fact that $r_{\text{acc}}(\sigma\kappa) \geq r_{\text{acc}}(8)$ yields

$$\begin{aligned}
\max_{i=1,\dots,d} \frac{\lambda_{\max}(P_i(\mu_{-1}))}{\lambda_{\min}(P_i(\mu_{-1}))} &\leq \max_{i=1,\dots,d} \|P_i(\mu_{-1})\| \\
&< 1 + 2\frac{(1+\delta_\sigma)^{-2}}{1-(1+\delta_\sigma)^{-2}} + 2\frac{M_2^2}{(1+\delta_\sigma)^2 r_{\text{acc}}(\sigma\kappa)^2}\frac{1+(1+\delta_\sigma)^{-2}}{(1-(1+\delta_\sigma)^{-2})^3} \\
&< 1 + \frac{12^2}{2\cdot 11^2}\frac{(1+\delta_\sigma)^{-2}}{\delta_\sigma} + 4\frac{M_2^2}{(1+\delta_\sigma)^2 r_{\text{acc}}(\sigma\kappa)^2}\frac{1+\delta_\sigma}{\delta_\sigma^3} \\
&< \left(\frac{1}{11^3} + \frac{12^2}{2\cdot 11^4} + \frac{4}{r_{\text{acc}}(\sigma\kappa)^2}\right)\frac{M_2^2}{\delta_\sigma^3} \\
&< 7\frac{M_2^2}{\delta_\sigma^3}.
\end{aligned}$$

Using the above bound twice yields

$$M_3 = \sqrt{\max_{i=1,\dots,d} \frac{\lambda_{\max}(P_i(\mu_{-1}))}{\lambda_{\min}(P_i(\mu_{-1}))} \max_{i=1,\dots,d} \frac{\lambda_{\max}(P_i(\mu_{-2}))}{\lambda_{\min}(P_i(\mu_{-2}))}} < 7\frac{M_2^2}{\delta_\sigma^3} = 7\cdot 4^3 M_2^2 \sigma^{3/2}\kappa^{3/2}. \quad (150)$$

Finally, we prove (144) by plugging (150) into (149), and then replacing κ with $\bar{\kappa}$, so that

$$\|x_{t+1} - x^\star\| \leq C\bar{\kappa}^{2(1+\nu)} r_{\text{acc}}(2\sigma\bar{\kappa})^t \|x_0 - x^\star\|,$$

where the constant C is given by

$$C = 84\cdot 4^3 M_2 (4M_1\omega/\delta_u)^\nu \sigma^{3/2}.$$

□

B.2 General case

We now build on the quadratic case to prove that the iterates x_t of Algorithm 1 also converge to the optimum $x^\star$ at an accelerated rate when the objective function f is not necessarily quadratic. Our approach is to show that if x_t is sufficiently close to $x^\star$, then $x_t - x^\star$ consists of a perturbation of the iterate when the objective is given by local quadratic approximation of f at $x^\star$.

B.2.1 Iterate dynamics in the general case

Under Assumption 4.3, it follows that f is twice continuously differentiable at $x^\star$. Hence, by Taylor's theorem [Nocedal and Wright, 2006, theorem 2.1], the gradient at x_t can be expressed as

$$\begin{aligned}\nabla f(x_t) =&\nabla f(x^\star) + \int_0^1 \nabla^2 f(x^\star + s(x_t - x^\star))(x_t - x^\star)ds \\ =&\nabla^2 f(x^\star)x_t + \int_0^1 (\nabla^2 f(x^\star + s(x_t - x^\star)) - \nabla^2 f(x^\star))(x_t - x^\star)ds \\ =&(H + \tilde{H}_t)(x_t - x^\star),\end{aligned} \tag{151}$$

where the Hessian error term $\tilde{H}_t = \tilde{H}_t(x_t)$ is given by

$$\tilde{H}_t = \int_0^1 (\nabla^2 f(x^\star + s(x_t - x^\star)) - \nabla^2 f(x^\star))ds. \tag{152}$$

Moreover, by (148), we have that $\|x_{t+1} - x^\star\| \leq \sqrt{36\kappa r_{\text{GD}}(\kappa)^{t-1}}\|x_0 - x^\star\|$. Since $r_{\text{GD}}(\kappa) \in (0, 1)$, if $\|x_0 - x^\star\| \leq \epsilon\sqrt{r_{\text{GD}}(\kappa)/36\kappa}$ and $\epsilon \leq \delta_H$, then for all $t \geq 0$, we have that $\|x_t - x^\star\| \leq \delta_H$, and

$$\begin{aligned}\|\tilde{H}_t\| &\leq \int_0^1 \|\nabla^2 f(x^\star + s(x_t - x^\star)) - \nabla^2 f(x^\star))\|ds \\ &\leq L_H \int_0^1 s\|x_t - x^\star\|ds \\ &\leq \epsilon L_H r_{\text{GD}}(\kappa)^{t/2}.\end{aligned} \tag{153}$$

Since v_j form an eigenbasis for $\mathbb{R}^d$, $\tilde{H}_t v_j$ can be expressed in v_j-coordinates, $\tilde{h}_{i,j,t}$, as

$$\tilde{H}_t v_j = \sum_{i=1}^d \tilde{h}_{i,j,t} v_i, \qquad\qquad j = 1, \ldots, d. \tag{154}$$

Then, using (154) and the decomposition $x_t - x^\star = \sum_{j=1}^d x_{j,t}v_j$ yields

$$\tilde{H}_t(x_t - x^\star) = \tilde{H}_t \sum_{j=1}^d x_{j,t}v_j = \sum_{j=1}^d x_{j,t}\tilde{H}_t v_j = \sum_{j=1}^d x_{j,t} \sum_{i=1}^d \tilde{h}_{i,j,t}v_i = \sum_{i=1}^d \sum_{j=1}^d \tilde{h}_{i,j,t}x_{j,t}v_i. \tag{155}$$

In turn, combining the decomposition $x_t - x^\star = \sum_{j=1}^d x_{j,t}v_j$ with (151) and (155), we obtain

$$\begin{aligned}y_{t+1} - x^\star &= x_t - (1/L)\nabla f(x_t) - x^\star \\ &= (I - H/L - \tilde{H}_t/L)(x_t - x^\star) \\ &= \sum_{i=1}^d \Big[(1 - \lambda_i/L)x_{i,t} + \sum_{j=1}^d (\tilde{h}_{i,j,t}/L)x_{j,t}\Big]v_i,\end{aligned}$$

from which it follows that

$$\begin{aligned}\sum_{j=1}^d x_{j,t+1}v_j &= x_{t+1} - x^\star \\ &= (1 + \beta_t)y_{t+1} - \beta_t y_t - x^\star \mp \beta_t x^\star \\ &= \sum_{i=1}^d \Big[(1 + \beta_t)\Big(1 - \frac{\lambda_i}{L}\Big)x_{i,t} - \beta_t\Big(1 - \frac{\lambda_i}{L}\Big)x_{i,t-1} \\ &\quad + \sum_{j=1}^d \Big((1 + \beta_t)\frac{\tilde{h}_{i,j,t}}{L}x_{j,t} - \beta_t\frac{\tilde{h}_{i,j,t}}{L}x_{j,t-1}\Big)\Big]v_i.\end{aligned}$$

Therefore, we have that

$$X_{t+1} = (G(m_t) + \tilde{G}_t)X_t, \tag{156}$$

where X_t is the vector with "stacked" $X_{i,t}$, as in

$$X_t = \begin{bmatrix} X_{1,t} \\ \vdots \\ X_{d,t} \end{bmatrix}, \tag{157}$$

while $G(m_t)$ and $\tilde{G}_t$ are matrices given by

$$G(m_t) = \operatorname{diag}(G_1(m_t), \dots, G_d(m_t)), \tag{158}$$

$$\tilde{G}_t = \frac{1}{L}\begin{bmatrix} 0 & 0 & \dots & 0 & 0 \\ -\beta_t\tilde{h}_{1,1,t-1} & (1+\beta_t)\tilde{h}_{1,1,t} & \dots & -\beta_t\tilde{h}_{1,d,t-1} & (1+\beta_t)\tilde{h}_{1,d,t} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 0 \\ -\beta_t\tilde{h}_{d,1,t-1} & (1+\beta_t)\tilde{h}_{d,1,t} & \dots & -\beta_t\tilde{h}_{d,d,t-1} & (1+\beta_t)\tilde{h}_{d,d,t} \end{bmatrix}, \tag{159}$$

where $G_i(m_t)$, defined by (104), are the system matrices governing the dynamics of each $X_{i,t}$ in the quadratic case where $f(x) = (x - x^\star)^\mathsf{T} H_t (x - x^\star)$. Using the fact that $\tilde{h}_{i,j,t} = v_i^\mathsf{T} \tilde{H}_t v_j$, the matrix $\tilde{G}_t$ given by (159) can be expressed as

$$\tilde{G}_t = \frac{1+\beta_t}{L} W_1^\mathsf{T} V^\mathsf{T} \tilde{H}_t V W_1 - \frac{\beta_t}{L} W_1^\mathsf{T} V^\mathsf{T} \tilde{H}_t V W_2, \tag{160}$$

where the matrices $V \in \mathbb{R}^{d\times d}$, $W_1, W_2 \in \mathbb{R}^{d\times 2d}$ are given by

$$V = \begin{bmatrix} v_1^\mathsf{T} \\ \vdots \\ v_d^\mathsf{T} \end{bmatrix}^\mathsf{T}, \quad W_1 = \begin{bmatrix} 0 & 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 \end{bmatrix}, \quad W_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 1 & 0 \end{bmatrix}.$$

Since v_i are orthonormal, so is V. Thus, V has unitary norm, as do W_1 and W_2. Therefore, applying the triangle inequality and norm submultiplicativity to (160), then using the fact that $\beta_t \in [0, 1)$ and lastly plugging in (153), we obtain

$$\begin{aligned} \|\tilde{G}_t\| &\le \frac{2}{L}\|W_1^\mathsf{T}\|\|V^\mathsf{T}\|\|\tilde{H}_t\|\|V\|\|W_1\| + \frac{1}{L}\|W_1^\mathsf{T}\|\|V^\mathsf{T}\|\|\tilde{H}_t\|\|V\|\|W_2\| \\ &= \frac{3}{L}\|\tilde{H}_t\| \\ &\le \epsilon\frac{L_H}{L} r_{\text{GD}}(\kappa)^{t/2}. \end{aligned} \tag{161}$$

We continue by noting that if Assumption 4.4 holds for some $\delta_\lambda > 0$, then it also holds for every $\delta'_\lambda < \delta_\lambda$. So, without loss of generality, suppose that Assumption 4.4 holds for some $\delta_\lambda \le \delta_m m$. Then, while $m_t > (1+\delta_m)m$, we have that $|m_t - \lambda_i| \ge \delta_\lambda$ for all $i = 1, \dots, d$. Hence, noting that for all λ_i we have that $\lambda_i < \bar{L}$, then from Corollary B.2, it follows that the two eigenvalues $\zeta_i(m_t)$ and $\xi_i(m_t)$ of each $G_i(m_t)$ are distinct. Therefore, because $\zeta_i(m_t)$ and $\xi_i(m_t)$ are continuous in m_t, we have that

$$\delta_T = \min_{m_t\in\mathcal{S}} \min_{i=1,\dots,d} |\zeta_i(m_t) - \xi_i(m_t)| > 0, \tag{162}$$

where $\mathcal{S} = \mathcal{S}(\delta_\lambda)$ is a compact set defined in terms

$$\mathcal{S} = [(1+\delta_m)m, L] \setminus \cup_{i=1}^d B(\lambda_i, \delta_\lambda),$$

and $B(\lambda_i, \delta_\lambda) = \{x : |x - \lambda_i| < \delta_\lambda\}$ is the open ball of radius δ_λ centered at λ_i. In the same vein, since T_i defined by (121) are continuous in ζ_i and ξ_i, and $\|\cdot\|$ is continuous, it follows that

$$\max_{m_t\in\mathcal{S}} \max_{i=1,\dots,d} \|T_i(m_t)\| < \infty.$$

Furthermore, explicitly computing the inverse of T_i for $m_t \in \mathcal{S}$ yields

$$\|T_i(m_t)^{-1}\| = \frac{1}{|\zeta_i - \xi_i|}\left\|\begin{bmatrix} \xi_i & -1 \\ -\zeta_i & 1 \end{bmatrix}\right\| \leq \frac{1}{\delta_T}\left\|\begin{bmatrix} \xi_i & -1 \\ -\zeta_i & 1 \end{bmatrix}\right\|. \tag{163}$$

Hence, since both sides of (163) are continuous in m_t, it follows that

$$\max_{m_t \in \mathcal{S}} \max_{i=1,\dots,d} \|T_i(m_t)^{-1}\| < \infty.$$

Therefore, we have that

$$M_T = \max_{m_t \in \mathcal{S}} \max_{i=1,\dots,d} \|T_i(m_t)\|\|T_i(m_t)^{-1}\| < +\infty. \tag{164}$$

Then, let T denote the coordinate transformation given by

$$T(m_t) = \operatorname{diag}(T_1(m_t), \dots, T_d(m_t)). \tag{165}$$

The block-diagonal structure of T combined with (164) implies that

$$\max_{m_t \in \mathcal{S}} \|T(m_t)\|\|T(m_t)^{-1}\| \leq M_T. \tag{166}$$

Furthermore, $T(m_t)$ diagonalizes $G(m_t)$, as in

$$G(m_t) = T(m_t)D(m_t)T(m_t)^{-1}, \tag{167}$$

where $D(m_t)$ is the block-diagonal matrix defined as $D(m_t) = \operatorname{diag}(D_1(m_t), \dots, D_d(m_t))$ and $D_i(m_t)$ are the diagonal matrices given by (122). So, defining the state $Z_t = T^{-1}(\mu_0)X_t$ for $t \in [t_0, t_1]$ and plugging Z_t and (167) into (156), since $m_t \equiv \mu_0$ for $t \in [t_0, t_1)$, it follows that

$$\begin{aligned} Z_{t+1} &= T^{-1}(\mu_0)X_{t+1} \\ &= T^{-1}(\mu_0)(G(m_t) + \tilde{G}_t)X_t \\ &= T^{-1}(G(\mu_0) + \tilde{G}_t)T(\mu_0)Z_t \\ &= (D(\mu_0) + \tilde{D}_t)Z_t, \end{aligned} \tag{168}$$

where $\tilde{D}_t$ is a perturbation matrix given by

$$\tilde{D}_t = T^{-1}(m_t)\tilde{G}_t T(m_t). \tag{169}$$

Using submultiplicativity and then combining (161) with (166), yields

$$\|\tilde{D}_t\| \leq \|T^{-1}(m_t)\|\|\tilde{G}_t\|\|T(m_t)\| \leq \epsilon M_T \frac{L_H}{L} r_{\text{GD}}(\kappa)^{t/2}. \tag{170}$$

Then, summing (170), we obtain

$$\sum_{t=0}^{+\infty} \|\tilde{D}_t\| \leq \epsilon M_T \frac{L_H}{L} \sum_{t=0}^{+\infty} r_{\text{GD}}(\kappa)^{t/2} \leq \epsilon M_T \frac{L_H}{L} \frac{1}{1 - \sqrt{r_{\text{GD}}(\kappa)}}. \tag{171}$$

Moreover, since $\lambda_i \leq L < \bar{L}$, it follows that $G(m_t)$ are nonsingular. This fact combined with (171) allows us to use results from the theory of asymptotic integration of difference equations [Bodine and Lutz, 2015] to establish that the solutions to (168) are perturbed solutions of the particular case when $\tilde{D}_t \equiv 0$. Namely, by [Bodine and Lutz, 2015, Thm 3.4, p.73], for $t \in [t_0, t_1)$ we have that

$$Z_{t+1} = [I + O(\epsilon)]D(\mu_0)^t Z_0,$$

which implies that for $t \in [t_0, t_1)$, we have that

$$X_{t+1} = T(\mu_0)[I + O(\epsilon)]D(\mu_0)^t Z_0 = T(\mu_0)[I + O(\epsilon)]D(\mu_0)^t T(\mu_0)^{-1} X_0,$$

which can also be written as a perturbation of the solution of the quadratic case $G(\mu_0)^t X_0$:

$$X_{t+1} = G(\mu_0)^t X_0 + T(\mu_0)D_0^t O(\epsilon)T(\mu_0)^{-1} X_0.$$

By repeatedly following the above procedure, we conclude that

$$\begin{aligned}
X_{t+1} =& T(\mu_J)[I + O(\epsilon)]D(\mu_J)^{t-t_J}T(\mu_J)^{-1}\left(\prod_{j=0}^{J-1} T(\mu_j)[I + O(\epsilon)]D(\mu_j)^{t_{j+1}-t_j}T(\mu_j)^{-1}\right)X_0 \\
=& T(\mu_J)D(\mu_J)^{t-t_J}T(\mu_J)^{-1}\left(\prod_{j=0}^{J-1} T(\mu_j)D(\mu_j)^{t_{j+1}-t_j}T(\mu_j)^{-1}\right)X_0 \\
&+ T(\mu_J)O(\epsilon)D(\mu_J)^{t-t_J}T(\mu_J)^{-1}\left(\prod_{j=0}^{J-1} T(\mu_j)D(\mu_j)^{t_{j+1}-t_j}T(\mu_j)^{-1}\right)X_0 + \dots \\
&+ T(\mu_J)O(\epsilon)D(\mu_J)^{t-t_J}T(\mu_J)^{-1}\left(\prod_{j=0}^{J-1} T(\mu_j)O(\epsilon)D(\mu_j)^{t_{j+1}-t_j}T(\mu_j)^{-1}\right)X_0, \quad (172)
\end{aligned}$$

where $t \in [t_J, t_{J+1})$.

B.2.2 The dynamics of c_t in the general case

Having established that the components $X_{i,t}$ in the general case behave like perturbed components of the quadratic case, we can also derive the dynamics of c_t in the general case. To this end, we establish bounds on the differences $x_{i,t+1} - x_{i,t}$. First, we notice that

$$X_{i,t+1} = [0 \quad \dots \quad 0 \quad I \quad 0 \quad \dots \quad 0]\, X_{t+1},$$

where $[0 \quad \dots \quad 0 \quad I \quad 0 \quad \dots \quad 0] \in \mathbb{R}^{2\times 2d}$ is a matrix made of a row of two-by-two blocks, where the i-th block is $I \in \mathbb{R}^{2\times 2}$ and all other blocks are $0 \in \mathbb{R}^{2\times 2}$. Also, we have that

$$\begin{aligned}
&[0 \quad \dots \quad 0 \quad I \quad 0 \quad \dots \quad 0]\, T(\mu_J)D(\mu_J)^{t-t_J}T(\mu_J)^{-1}\left(\prod_{j=0}^{J-1} T(\mu_j)D(\mu_j)^{t_{j+1}-t_j}T(\mu_j)^{-1}\right)X_0 \\
&= [0 \quad \dots \quad 0 \quad I \quad 0 \quad \dots \quad 0]\, G(\mu_J)^{t-t_J}\prod_{j=0}^{J-1} G(\mu_j)^{t_{j+1}-t_j}X_0 \\
&= \left[0 \quad \dots \quad 0 \quad G_i^{t-t_J}\textstyle\prod_{j=0}^{J-1} G_i(\mu_j)^{t_{j+1}-t_j} \quad 0 \quad \dots \quad 0\right] X_0 \\
&= G_i^{t-t_J}\prod_{j=0}^{J-1} G_i(\mu_j)^{t_{j+1}-t_j}X_{i,0},
\end{aligned}$$

since $G_i(\mu_j) = T_i(\mu_j)D_i(\mu_j)T_i(\mu_j)^{-1}$, by (122), and G, T and D are block-diagonal matrices with blocks given by G_i, T_i and D_i, respectively. Then, we notice that all but the first term in (172) are $O(\epsilon)$, and $\rho(\mu_j, \lambda_i) \leq \rho(\mu_j, m)$ for all the eigenvalues $\rho(\mu_j, \lambda_i)$ of $D(\mu_j)$, by Lemma B.3. Therefore, combining the above remarks with Assumption 4.5, it follows that for $t \in [t_j, t_{j+1})$

$$\begin{aligned}
\|X_{i,t+1}\|^2 \leq& \overline{C}_i\rho(\mu_j, \lambda_i)^{2(t-t_j)}\left(\prod_{k=0}^{j-1}\rho(\mu_k, \lambda_i)^{2(t_{k+1}-t_k)}\right)x_{i,0}^2 \\
&+ O(\epsilon)\rho(\mu_j, m)^{2(t-t_j)}\left(\prod_{k=0}^{j-1}\rho(\mu_k, m)^{2(t_{k+1}-t_k)}\right)x_{1,0}^2,
\end{aligned}$$

for some $\overline{C}_i$. The above bound is analogous to (124), but with an additional $O(\epsilon)$ term accounting for the perturbation of the quadratic solution. Thus, for $t \in [t_j, t_{j+1})$, we have that

$$\begin{aligned}
(x_{i,t+1} - x_{i,t})^2 =& ([-1 \quad 1]\, X_{i,t+1})^2 \\
&\leq 2\overline{C}_i\rho(\mu_j, \lambda_i)^{2(t-t_j)}\left(\prod_{k=0}^{j-1}\rho(\mu_k, \lambda_i)^{2(t_{k+1}-t_k)}\right)x_{i,0}^2 \\
&\quad + O(\epsilon)\rho(\mu_j, m)^{2(t-t_j)}\left(\prod_{k=0}^{j-1}\rho(\mu_k, m)^{2(t_{k+1}-t_k)}\right)x_{1,0}^2. \quad (173)
\end{aligned}$$

In the same vein, combining (172) with Lemma B.3, we have that for $t \in [t_j, t_{j+1})$

$$\|X_{t+1}\|^2 \leq 2(1+O(\epsilon))\overline{C}\rho(\mu_j, m)^{2(t-t_j)}\left(\prod_{k=0}^{j-1}\rho(\mu_k, m)^{2(t_{k+1}-t_k)}\right)\|x_0\|^2, \tag{174}$$

where $\overline{C} = \max_{i=1,\ldots,d}\overline{C}_i$. Similarly, combining the derivation of (127) with (172) and Assumption 4.5, for $t \in [t_j, t_{j+1})$ we have that

$$(x_{1,t+1} - x_{1,t})^2 \geq (1-O(\epsilon))\underline{C}_1\rho(\mu_j, m)^{2(t-t_j)}\left(\prod_{k=0}^{j-1}\rho(\mu_k, m)^{2(t_{k+1}-t_k)}\right)x_{1,0}^2, \tag{175}$$

for some $\underline{C}_1$.

Our next step is to also express $\nabla f(x_{t+1}) - \nabla f(x_t)$ as a perturbation of the quadratic case. To this end, we substitute (151) for $\nabla f(x_{t+1})$ and $\nabla f(x_t)$, and obtain

$$\begin{aligned}\nabla f(x_{t+1}) - \nabla f(x_t) &= (H + \tilde{H}_{t+1})(x_{t+1} - x^\star) - (H + \tilde{H}_t)(x_t - x^\star)\\ &= H(x_{t+1} - x_t) + \tilde{H}_{t+1}(x_{t+1} - x^\star) - \tilde{H}_t(x_t - x^\star).\end{aligned}$$

Using $x_t - x^\star = \sum_{j=1}^d x_{j,t}v_j$, the terms of $\nabla f(x_{t+1}) - \nabla f(x_t)$ above can be written as

$$\begin{aligned}H(x_{t+1} - x_t) &= \sum_{i=1}^d (x_{i,t+1} - x_{i,t})\lambda_i v_i,\\ \tilde{H}_{t+1}(x_{t+1} - x^\star) - \tilde{H}_t(x_t - x^\star) &= \sum_{i=1}^d\left(\sum_{j=1}^d(\tilde{h}_{i,j,t+1}x_{j,t+1} - \tilde{h}_{i,j,t}x_{j,t})\right)v_i.\end{aligned}$$

In turn, using the above expressions, it follows that

$$\begin{aligned}\|\nabla f(x_{t+1}) - \nabla f(x_t)\|^2 = &\sum_{i=1}^d \lambda_i^2(x_{i,t+1} - x_{i,t})^2\\ &+ 2\sum_{i=1}^d \lambda_i(x_{i,t+1} - x_{i,t})\sum_{j=1}^d(\tilde{h}_{i,j,t+1}x_{j,t+1} - \tilde{h}_{i,j,t}x_{j,t})\\ &+ \sum_{i=1}^d\left(\sum_{j=1}^d(\tilde{h}_{i,j,t+1}x_{j,t+1} - \tilde{h}_{i,j,t}x_{j,t})\right)^2.\end{aligned}$$

Our next step is to bound the third term above in $\|\nabla f(x_{t+1}) - \nabla f(x_t)\|^2$. Combining the identities $\|\tilde{H}_{t+1}\|_F^2 + \|\tilde{H}_t\|_F^2 = \sum_{i=1}^d\sum_{j=1}^d(\tilde{h}_{i,j,t+1}^2 + \tilde{h}_{i,j,t}^2)$ and $\|X_{t+1}\|^2 = \sum_{j=1}^d(x_{j,t+1}^2 + x_{j,t}^2)$ with the bound (153), it follows that

$$\begin{aligned}&\sum_{i=1}^d\left(\sum_{j=1}^d(\tilde{h}_{i,j,t+1}x_{j,t+1} - \tilde{h}_{i,j,t}x_{j,t})\right)^2\\ &\leq \sum_{i=1}^d\left(\sum_{j=1}^d(|\tilde{h}_{i,j,t+1}| + |\tilde{h}_{i,j,t}|)(|x_{j,t+1}| + |x_{j,t}|)\right)^2\\ &\leq \sum_{i=1}^d\left(\sum_{j=1}^d(|\tilde{h}_{i,j,t+1}| + |\tilde{h}_{i,j,t}|)^2\right)\left(\sum_{j=1}^d(|x_{j,t+1}| + |x_{j,t}|)^2\right)\\ &\leq \sum_{i=1}^d\left(2\sum_{j=1}^d(\tilde{h}_{i,j,t+1}^2 + \tilde{h}_{i,j,t}^2)\right)\left(2\sum_{j=1}^d(x_{j,t+1}^2 + x_{j,t}^2)\right)\\ &= 4(\|\tilde{H}_{t+1}\|_F^2 + \|\tilde{H}_t\|_F^2)\|X_{t+1}\|^2\\ &\leq 4d(\|\tilde{H}_{t+1}\|^2 + \|\tilde{H}_t\|^2)\|X_{t+1}\|^2\\ &\leq 4\epsilon^2 L_H^2 d\|X_{t+1}\|^2.\end{aligned} \tag{176}$$

In turn, we address the second term of $\|\nabla f(x_{t+1}) - \nabla f(x_t)\|^2$ using (176), which gives

$$\begin{aligned}
&\sum_{i=1}^{d}(x_{i,t+1}-x_{i,t})\sum_{j=1}^{d}(\tilde{h}_{i,j,t+1}x_{j,t+1}-\tilde{h}_{i,j,t}x_{j,t}) \\
&\leq \sqrt{\sum_{i=1}^{d}(x_{i,t+1}-x_{i,t})^2}\sqrt{\sum_{i=1}^{d}\left(\sum_{j=1}^{d}(\tilde{h}_{i,j,t+1}x_{j,t+1}-\tilde{h}_{i,j,t}x_{j,t})\right)^2} \\
&\leq \sqrt{2\sum_{i=1}^{d}(x_{i,t+1}^2+x_{i,t}^2)}\sqrt{4\epsilon^2 L_H^2 d\|X_{t+1}\|^2} \\
&\leq 2\epsilon L_H\sqrt{2d}\|X_{t+1}\|^2. \qquad (177)
\end{aligned}$$

Then, combining (174), (176) and (177) we obtain

$$\begin{aligned}
\|\nabla f(x_{t+1}) - \nabla f(x_t)\|^2 &\leq \sum_{i=1}^{d}\lambda_i^2(x_{i,t+1}-x_{i,t})^2 \\
&\quad + O(\epsilon)\rho(\mu_j,m)^{2(t-t_j)}\left(\prod_{k=0}^{j-1}\rho(\mu_k,m)^{2(t_{k+1}-t_k)}\right)x_{1,0}^2. \qquad (178)
\end{aligned}$$

In turn, plugging (178) into (4), and then using (175), it follows that

$$c_{t+1}^2 = c(x_{t+1},x_t)^2 \leq \frac{\sum_{i=1}^{d}\lambda_i^2(x_{i,t+1}-x_{i,t})^2}{\sum_{i=1}^{d}(x_{i,t+1}-x_{i,t})^2} + O(\epsilon). \qquad (179)$$

Lemma B.15. *Let $f \in \mathcal{F}(L,m)$ and suppose that Assumptions 4.2 to 4.5 hold. Also, let δ_m and δ_u be positive numbers such that $\delta_m > \delta_u > 0$ and let $r' = r'(\delta_u,\delta_\ell,\kappa)$ be a function such that $r_{\text{acc}}(\sigma_\phi\kappa) \leq r' < 1$ for all $\kappa \geq 1+\delta_\ell$, where $\delta_\ell = ((1+\delta_m)/(1+\delta_u)) - 1 > 0$ and $\sigma_\phi = \sigma_\phi(\delta_u,\delta_\ell,\kappa)$ is given by Lemma B.10. Then, there exists some $\epsilon > 0$ such that if $\|x_0 - x^\star\| \leq \epsilon$, then the estimates m_t of Algorithm 1 reach $[m/\gamma,(1+\delta_m)m]$ after no more than τ iterations, where*

$$\tau = \frac{-\log(8\kappa^2 M_1\omega/\delta_u)}{\log r'}\sum_{j=0}^{\log_\gamma(\kappa/(1+\delta_m))}\frac{1}{1+\alpha_\phi m(1+\delta_\ell)(\gamma^j-1)}, \qquad (180)$$

$M_1 = \max_i \overline{C}_i/\underline{C}_1$, with ω and γ given by Assumptions 4.2 to 4.5.

Proof. Suppose that the last value of m_t before reaching the interval $[m/\gamma,(1+\delta_m)m)$ is $(1+\delta_m)m$. Then, suppose that the value before last is $\gamma(1+\delta_m)m$, and so on, up to $\gamma^K(1+\delta_m)m$ for some K such that $\gamma^K(1+\delta_m)m \leq L < \gamma^{K+1}(1+\delta_m)m$. Using this m_t schedule, we bound the number of iterations that m_t takes to reach the interval $[m/\gamma,(1+\delta_m)m]$, and then we argue that no other m_t schedule can lead to a worse bound.

Let $\ell_j = \gamma^j(1+\delta_m)m/(1+\delta_u)$. Then, we have that $\ell_j \geq (1+\delta_\ell)m$ for $\delta_\ell = ((1+\delta_m)/(1+\delta_u))-1$. Since $\delta_m > \delta_u$, then $\delta_\ell > 0$, and Lemma B.11 applies. Now, let $I_j = \min\{i : \lambda_i \geq \ell_j\}$. That is, $\lambda_i \geq \ell_j$ if and only if $i \geq I_j$. Then, using this fact and separating the terms indexed by $i < I_0$ from those indexed by $i \geq I_0$ in (179) into two sums yields

$$c_{t+1}^2 < \ell_0^2\frac{\sum_{i=1}^{I_0-1}(x_{i,t+1}-x_{i,t})^2}{\sum_{i=1}^{d}(x_{i,t+1}-x_{i,t})^2} + \frac{\sum_{i=I_0}^{d}\lambda_i^2(x_{i,t+1}-x_{i,t})^2}{\sum_{i=1}^{d}(x_{i,t+1}-x_{i,t})^2} + O(\epsilon).$$

In turn, plugging the above inequality into the identity $(c_{t+1}+\ell_0)(c_{t+1}-\ell_0) = c_{t+1}^2 - \ell_0^2$, and then using the fact that $\lambda_i \leq L$ and $\ell_0 \geq m$, we obtain

$$\begin{aligned}
c_{t+1}-\ell_0 = \frac{c_{t+1}^2-\ell_0^2}{c_{t+1}+\ell_0} &< \frac{\sum_{i=I_0}^{d}(\lambda_i^2-\ell_0^2)(x_{i,t+1}-x_{i,t})^2}{\ell_0\sum_{i=1}^{d}(x_{i,t+1}-x_{i,t})^2} + O(\epsilon) \\
&\leq \ell_0\kappa^2\sum_{i=I_0}^{d}\frac{(x_{i,t+1}-x_{i,t})^2}{(x_{1,t+1}-x_{1,t})^2} + O(\epsilon). \qquad (181)
\end{aligned}$$

To address the terms in the sum in (181), we combine (173) and (175), assuming $t_K \le t < t_{K+1}$. That is, we consider the last adjustment before m_t reaches the interval $[m/\gamma, (1+\delta_m)m)$. Then, we apply Lemma B.3 twice, to get $\rho(m_k, \lambda_i) \le \rho(m_k, \ell_0)$ and $\rho(m_k, \ell_0) < \rho(m_k, m)$ for all $i \ge I_0$, which gives

$$\begin{aligned}\sum_{i=I_0}^{d} \frac{(x_{i,t+1}-x_{i,t})^2}{(x_{1,t+1}-x_{1,t})^2} &\le 2(1+O(\epsilon))M_1 \sum_{i=I_0}^{d} \frac{x_{i,0}^2}{x_{1,0}^2} \prod_{k=t_K}^{t} \frac{\rho(m_k,\lambda_i)^2}{\rho(m_k,m)^2} \prod_{k=1}^{\tau_K} \frac{\rho(m_k,\lambda_i)^2}{\rho(m_k,m)^2} + O(\epsilon) \\ &\le 2(1+O(\epsilon))M_1\omega\phi((1+\delta_u)\ell_0,\ell_0,m)^{2(t-t_K+1)} + O(\epsilon). \end{aligned} \tag{182}$$

where $M_1 = 2\max_i \overline{C}_i/\underline{C}_1$. Next, we put (181) and (182) together, and since $\ell_0 \ge m$, by choosing ϵ sufficiently small, we get

$$c_{t+1} - \ell_0 \le 2(1+O(\epsilon))\kappa^2 M_1\omega\phi((1+\delta_m)m,\ell_0,m)^{2(t-t_K+1)}\ell_0 + O(\epsilon) \le \delta_u\ell_0/2,$$

for all $t \ge t_K + \Delta t_0$, where

$$\Delta t_0 = -\frac{\log(8\kappa^2 M_1\omega/\delta_u)}{\log r_{\text{acc}}(\sigma_\phi\kappa)}.$$

Therefore, $c_{t+1} < (1+\delta_u)\ell_0 = (1+\delta_m)m$ for $t \ge t_K + \Delta t_0$ or, equivalently,

$$t_{K+1} - t_K \le \Delta t_0.$$

Note that for every m_t schedule, if μ_K denotes the last value of m_t before reaching $[m/\gamma, (1+\delta_m)m]$, then $\mu_K \ge (1+\delta_m)m$, by definition. Hence, letting $\ell_0' = \mu_K/(1+\delta_u)$ and $I_0' = \{i : \lambda_i \ge \ell_0'\}$, then $I_0' \ge I_0$, and by applying Lemma B.3 before, it follows that

$$\begin{aligned}\sum_{i=I_0'}^{d} \frac{(x_{i,t+1}-x_{i,t})^2}{(x_{1,t+1}-x_{1,t})^2} &\le 2(1+O(\epsilon))M_1 \sum_{i=I_0'}^{d} \frac{x_{i,0}^2}{x_{1,0}^2} \prod_{k=t_K}^{t} \frac{\rho(m_k,\lambda_i)^2}{\rho(m_k,m)^2} \prod_{k=1}^{\tau_K} \frac{\rho(m_k,\lambda_i)^2}{\rho(m_k,m)^2} + O(\epsilon) \\ &\le 2(1+O(\epsilon))M_1 \sum_{i=I_0}^{d} \frac{x_{i,0}^2}{x_{1,0}^2} \prod_{k=t_K}^{t} \frac{\rho(m_k,\lambda_i)^2}{\rho(m_k,m)^2} \prod_{k=1}^{\tau_K} \frac{\rho(m_k,\lambda_i)^2}{\rho(m_k,m)^2} + O(\epsilon) \\ &\le 2(1+O(\epsilon))M_1\omega\phi((1+\delta_u)\ell_0,\ell_0,m)^{2(t-t_K+1)} + O(\epsilon).\end{aligned}$$

Therefore, the last adjustment cannot take more than Δt_0 iterations for any m_t schedule.

Then, let Δt_j be quantities analogous to Δt_0, defined for $j = 0, \dots, K$ as

$$\Delta t_j = -\frac{\log(8\kappa^2 M_1\omega/\delta_u)}{(1+\alpha_\phi(\ell_j - (1+\delta_\ell)m))\log r_{\text{acc}}(\sigma_\phi\kappa)}.$$

By Assumption 4.2, m_t decreases by a factor of at least γ every time it is adjusted to a new value. Hence, $\mu_{K-1} \ge \gamma\mu_K$ for every m_t schedule, which implies that $\ell_1 \le \mu_{K-1}/(1+\delta_u)$ for every m_t schedule. Hence, by the same rationale above, it cannot take more than Δt_1 for m_t to be adjusted to its second last value before reaching the interval $[m/\gamma, (1+\delta_m)m]$. It follows by induction that it cannot take more than Δt_j for m_t to be adjust to its $K-j$-th to last value before reaching the interval $[m/\gamma, (1+\delta_m)m]$. Moreover, since by design $m_t \le L$, it cannot more than $K \le \log_\gamma(\kappa/(1+\delta_m))$ adjustments before m_t reaches the interval $[m/\gamma, (1+\delta_m)m]$. Therefore, we conclude that $m_{t+1} \le (1+\delta_m)m$ for all $t \ge \tau$, where

$$\begin{aligned}\tau &= \frac{-\log(8\kappa^2 M_1\omega/\delta_u)}{\log r'} \sum_{j=0}^{\log_\gamma(\kappa/(1+\delta_m))} \frac{1}{1+\alpha_\phi m(1+\delta_\ell)(\gamma^j-1)} \\ &\ge \frac{-\log(8\kappa^2 M_1\omega/\delta_u)}{\log r_{\text{acc}}(\sigma_\phi\kappa)} \sum_{j=0}^{K} \frac{1}{1+\alpha_\phi m(1+\delta_\ell)(\gamma^j-1)},\end{aligned}$$

because log is monotone and $r_{\text{acc}}(\sigma_\phi\kappa) \le r' < 1$, and $K \le \log_\gamma(\kappa/(1+\delta_m))$. □

B.3 Main result

At last, we are ready to prove Theorem 4.6, which we restate below, establishing that Algorithm 1 achieves acceleration around the minimum.

Theorem (4.6)**.** Let $f \in \mathcal{F}(L, m)$, $\bar{L} > L$ and $\bar{\kappa} = \bar{L}/m$. Suppose that $\bar{\kappa} > \kappa = L/m \geq 4$ and that Assumptions 4.2 to 4.5 hold. Then, there is some $\epsilon > 0$ such that if $\|x_0 - x^\star\| \leq \epsilon$, then the iterates x_t produced by Algorithm 1 satisfy

$$\|x_{t+1} - x^\star\| \leq C r_{\text{acc}}(\sigma\bar{\kappa})^t \|x_0 - x^\star\|,$$

where σ depends on γ, C depends on $\bar{\kappa}$ and ω, with γ and ω given by Assumptions 4.2 to 4.5.

Proof. Let δ_m and δ_u be positive numbers such that $\delta_u < \min\{\delta_m, 1/2\}$ and $\delta_m \leq \gamma - 1$. Then, define $\delta_\ell = (1+\delta_m)/(1+\delta_u) - 1$, and let $r' = r'(\delta_u, \delta_\ell, \kappa)$ be a function such that $r_{\text{acc}}(\sigma_\phi\kappa) \leq r' < 1$ for all $\kappa \geq 1 + \delta_\ell$, where $\sigma_\phi = \sigma_\phi(\delta_u, \delta_\ell, \kappa)$ is given by Lemma B.10. By Lemma B.15, it follows that $m_t \leq (1 + \delta_m)m$ for all $t \geq \tau$, where

$$\tau = \frac{-\log(8\kappa^2 M_1 \omega/\delta_u)}{\log r'} \sum_{j=0}^{\log_\gamma(\kappa/(1+\delta_m))} \frac{1}{1 + \alpha_\phi m(1+\delta_\ell)(\gamma^j - 1)}, \tag{183}$$

$M_1 = \max_i \overline{C}_i/\underline{C}_1$. In turn, as in the proof of Proposition B.14, Corollary B.8 then implies that $\rho(m_t, m) \leq r_{\text{acc}}(\sigma_1\kappa)$ for all $t \geq \tau$, where $\sigma_1 = \max\{\gamma, \sigma_m\}$, and $\sigma_m = 1 + 2\delta_m + 2\sqrt{\delta_m(1+\delta_m)}$.

Now, by Lemma B.3, we have that $\rho(m_t, \lambda_i) \leq r_{\text{acc}}(\sigma_1\kappa)$ for all λ_i. Hence, given δ_σ such that $(1 + \delta_\sigma) r_{\text{acc}}(\sigma_1\kappa) < 1$, by Lemma B.13 there is a $P_i(m_t) = P_i(m_t, \delta_\sigma) \succeq I$ for each λ_i such that $G_i(m_t)^\mathsf{T} P_i(m_t) G_i(m_t) \preceq (1 + \delta_\sigma)^2 r_{\text{acc}}(\sigma_1\kappa)^2 P_i(m_t)$. Using $P_i(m_t)$ as diagonal blocks, we define the matrix $P(m_t) = P(m_t, \delta_\sigma) = \operatorname{diag}(P_1(m_t), \ldots, P_d(m_t))$. The block diagonal structure of P and G implies that $P(m_t) \succeq I$, and that $G(m_t)^\mathsf{T} P(m_t) G(m_t) \preceq (1 + \delta_\sigma)^2 r_{\text{acc}}(\sigma_1\kappa)^2 P(m_t)$. Hence, if $t_j \leq t < t_{j+1}$ and $t \geq \tau$, then (156) yields

$$\begin{aligned} X_{t+1}^\mathsf{T} P(m_t) X_{t+1} &= X_t^\mathsf{T} (G(\mu_j) + \tilde{G}_t)^\mathsf{T} P(\mu_j)(G(\mu_j) + \tilde{G}_t) X_t \\ &\leq (1 + \delta_\sigma)^2 r_{\text{acc}}(\sigma_1\kappa)^2 X_t^\mathsf{T} P(\mu_j) X_t + X_t^\mathsf{T} \tilde{P}_t X_t, \end{aligned} \tag{184}$$

since $m_t = \mu_j$, where

$$\tilde{P}_t = \tilde{G}_t^\mathsf{T} P(m_t) G(m_t) + G(m_t)^\mathsf{T} P(m_t) \tilde{G}_t + \tilde{G}_t^\mathsf{T} P(m_t) \tilde{G}_t.$$

By Equation (104), we have that

$$\begin{aligned} \|G_i(m_t)\| &\leq \left\| \begin{bmatrix} 0 & 1 \\ -\beta(m_t)\left(1 - \frac{\lambda_i}{L}\right) & 0 \end{bmatrix} \right\| + \left\| \begin{bmatrix} 0 & 0 \\ 0 & (1 + \beta(m_t))\left(1 - \frac{\lambda_i}{L}\right) \end{bmatrix} \right\| \\ &= \max\left\{1, \beta(m_t)\left(1 - \frac{\lambda_i}{L}\right)\right\} + (1 + \beta(m_t))\left(1 - \frac{\lambda_i}{L}\right) \\ &\leq 3. \end{aligned} \tag{185}$$

Furthermore, since $\delta_\sigma > 0$ and $r_{\text{acc}}(\sigma_1\kappa) \geq r_{\text{acc}}(4) = 1/2$, because by assumption $\kappa \geq 4$, the block diagonal structure of P combined with (140) yields

$$\|P(m_t)\| \leq \frac{2}{1 - (1 + \delta_\sigma)^{-2}} + \frac{16 M_2^2}{(1 - (1 + \delta_\sigma)^{-2})^3} = M_\delta. \tag{186}$$

Therefore, combining (161), (185) and (186), and taking ϵ such that $\epsilon L_H/L < 1$, we obtain

$$\|\tilde{P}_t\| \leq (2\|G(m_t)\| + \|\tilde{G}_t\|)\|\tilde{G}_t\| \leq 7\epsilon M_\delta L_H/L. \tag{187}$$

Then, since $P(m_t) \succeq I$, from (184) and (187) it follows that

$$X_{t+1}^T P X_{t+1} \leq ((1 + \delta)^2 r^2 + 7\epsilon M_\delta L_H/L) X_t^T P X_t = \tilde{r}^2 X_t^T P X_t, \tag{188}$$

where we conveniently use a perturbed rate $\tilde{r}$, given by

$$\tilde{r} = \sqrt{(1 + \delta_\sigma)^2 r^2 + 7\epsilon M_\delta L_H/L}. \tag{189}$$

Consecutively applying (188) and reproducing the steps in the proof of Proposition B.14, we get

$$\|x_{t+1} - x^\star\| \leq C' \bar{\kappa}^{2(1+\nu)} \tilde{r}^t \|x_0 - x^\star\|, \tag{190}$$

where the constant C is given by

$$C' = 84 \cdot 4^3 M_2 (8M_1\omega/\delta_u)^\nu \sigma_2^{3/2},$$

with $\nu = 1 + \gamma/(\alpha_\phi m(\gamma - 1))$, $\sigma_2 = \max\{\gamma, \sigma_m, \sigma_\phi\}$, and $\delta_\sigma = 1/(4\sqrt{\sigma_2}\kappa)$ is chosen such that

$$(1 + \delta_\sigma) r_{\text{acc}}(\sigma_2\kappa) < r_{\text{acc}}(2\sigma_2\kappa).$$

Hence, choosing ϵ sufficiently small such that

$$\sqrt{(1 + \delta_\sigma)^2 r^2 + 7\epsilon M_\delta L_H / L} \leq r_{\text{acc}}(2\sigma_2\kappa),$$

and then plugging this choice of ϵ back into (190), we obtain

$$\|x_{t+1} - x^\star\| \leq C r_{\text{acc}}(\sigma\bar{\kappa})^t \|x_0 - x^\star\|,$$

where $\sigma = 2\sigma_2$ and $C = C'\bar{\kappa}^{2(1+\nu)}$, which concludes the proof. □

Remarks To conclude this section, we make a few remarks on the constant C and convergence rate σ in Theorem 4.6. One of the factors of C involves a power $\nu = 1 + \gamma/(\alpha_\phi m(\gamma - 1))$. Although this factor is an artifact of our conservative analysis, ν is nonetheless small. For example, we find numerically that $\alpha_\phi m \approx 10$ when the base rate is $r_{\text{acc}}(4\kappa)$, as Figure 7 illustrates. Moreover, $\gamma \geq 2$ by Assumption 4.2, therefore $\nu \leq 1.2$. In the proof of Theorem 4.6, we work with $\sigma = 2\sigma_2$, where $\sigma_2 = \max\{2, \sigma_m, \sigma_\phi\}$, which is a function of two further suboptimality factors given by $\sigma_m = 1 + 2\delta_m + 2\sqrt{\delta_m(1 + \delta_m)}$ and $\sigma_\phi \lesssim 1/4(\sqrt{\delta_u + \delta_\ell(1 + \delta_u)} - \sqrt{\delta_u(1 + \delta_\ell)})^2$. The factor 2 in $\sigma = 2\sigma_2$ is a result of a compromise to obtain a reasonable condition number for the matrix P in the Lyapunov analysis of the final regime of NAG-free, when m_t is sufficiently accurate for accelerated convergence. In reality, this compromise is an artifact of *any* Lyapunov analysis of linear systems, whose solutions are linear combinations of some $t^k r^t$ terms, rather than purely exponential solutions r^t. Hence, this compromise is typically ignored, e.g. as in Lessard et al. [2016], in which case the convergence rate is $\max\{2, \sigma_m, \sigma_\phi\}$. Then, for example, letting $\gamma = 2, \delta_m = 0.2, \delta_u = 0.01$ and $\delta_\ell = (1 + \delta_m)/(1 + \delta_u) - 1$, we obtain $\max\{\gamma, \sigma_m, \sigma_\phi\} \leq 2.4$. Therefore, the convergence rate in this case would not be worse than $r_{\text{acc}}(2.4\kappa)$. The γ factor is an artifact of our analysis, but there is a real tension between the other suboptimality factors σ_m and σ_ϕ. That is because as m_t approaches m, the estimation convergence rate slows down, whereas the actual convergence rate improves, since the estimate gets more accurate. This translates into reduced δ_m and δ_ℓ, which reduces σ_m, but increases σ_ϕ. The bound on these suboptimality rates is conservative, however. In the next section, we present experiments in which we see that in practice the suboptimality factor is much closer to 1.

C Numerical experiments

In this section, we present further experiments and provide more details of those already presented in Section 5. We start by recalling which methods are used in the experiments.

C.1 Methods

As baselines, we take GD, NAG and the Triple Momentum Method [Van Scoy et al., 2017, TMM], replacing L and m with problem-specific bounds. As competing methods, we consider two restart schemes based on [O'Donoghue and Candès, 2015]: one where L is replaced with a problem-specific bound, that we refer to as NAG+R, and another where L is estimated online with backtracking, that we refer to as NAG+RB. For NAG+RB, every time the L estimate L_t fails to produce enough descent, it is increased to $1.01L_t$ and tested again. This choice of adjustment factor produces less conservative estimates at the expense of more function evaluations, which largely favors NAG+RB since in all experiments we plot the suboptimality gap $f(x_t) - f(x^\star)$ versus iterations. As additional methods for comparison, we consider the adaptive gradient method AdGD [Malitsky and Mishchenko, 2024] and its accelerated heuristic, AdGD-accel2, both using $\gamma = 1/\sqrt{2}$, as well as a previous variant of the accelerated heuristic [Malitsky and Mishchenko, 2020], which we denote by AdGD-accel.

C.2 Smoothed and regularized log-sum-exp

We start by clarifying the remark made in Appendix C.2 that log-sum-exp approximates the max function, by which we mean that

$$\max_{i=1,\dots,n} p_i \le \theta \log\left(\sum_{i=1}^{n} \exp\left(\frac{p_i}{\theta}\right)\right) \le \max_{i=1,\dots,n} p_i + \theta \log n.$$

Thus, as θ decreases, the approximation gets tighter. In the first experiment, we also compare some variants of Algorithm 1 that employ different strategies to estimate L. One of them is Algorithm 3, which estimates L by backtracking line search.

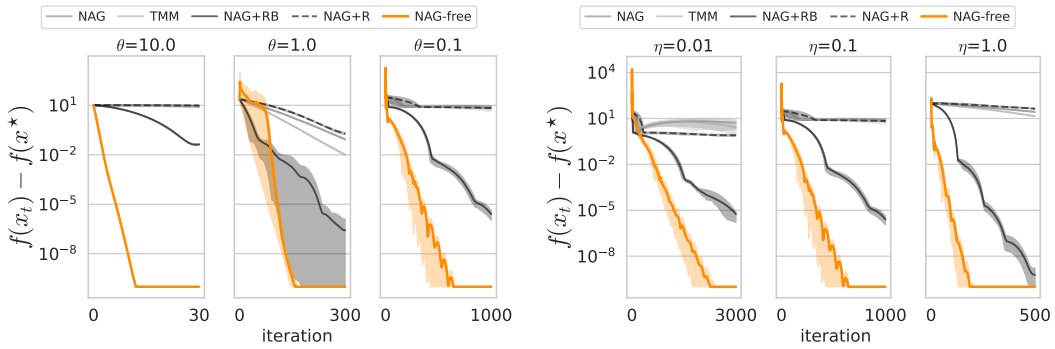
Algorithm 3 NAG-free-back: a variant of NAG-free that estimates L via backtracking.

Input: $T > 0, x_0 = y_0$
Output: x_T
$y \sim x_0 + U[0, 10^{-6}]^d$
$L_0, m_0 \leftarrow \|\nabla f(x_0) - \nabla f(y)\| / \|x_0 - y\|$
for $t = 0, \dots, T-1$ **do**
  $y_{t+1} \leftarrow x_t - (1/L_t)\nabla f(x_t)$
  while $f(y_{t+1}) - f(x_t) > -(1/2L_t)\|\nabla f(x_t)\|^2$ **do**
    $L_t \leftarrow 1.01 L_t$
    $y_{t+1} \leftarrow x_t - (1/L_t)\nabla f(x_t)$
  end while
  $x_{t+1} \leftarrow y_{t+1} + \dfrac{\sqrt{L_t} - \sqrt{m_t}}{\sqrt{L_t} + \sqrt{m_t}}(y_{t+1} - y_t)$
  $c_{t+1} \leftarrow \|\nabla f(x_{t+1}) - \nabla f(x_t)\| / \|x_{t+1} - x_t\|$
  $m_{t+1} \leftarrow \min(m_t, c_{t+1})$
end for

In Appendix C.2, we also mention different variants of Equation (6), in which we fix $n = d = 600$ and vary the regularization and smoothing parameters, η and θ, respectively. Figure 8 shows the results of those experiments. In particular, Figure 8a shows the results for $\theta \in \{0.1, 1, 10\}$, when $\eta = 0.1$. Similarly, Figure 8b shows the results for $\eta \in \{0.01, 0.1, 1\}$, when $\theta = 0.1$.

C.3 Regularized logistic regression

Table 1 contains details of the datasets used in the experiments described in Appendix C.3, all of which were taken from the LIBSVM dataset Chang and Lin [2011].



(a) $\eta = 0.1$. (b) $\theta = 0.1$.

Figure 8: Suboptimality gap for log-sum-exp ($n = 600, d = 600$) using methods from Appendix C.1.

Table 1: Details of datasets and method precisions used in the logistic regression problem.

dataset	datapoints	dimensions
a9a	32,561	123
covtype	581,012	54
gisette_scale	6000	5000
mushrooms	8,124	112
phishing	11,055	68
svmguide1	3,089	4
web-1	2,477	300

C.4 Cubic regularization

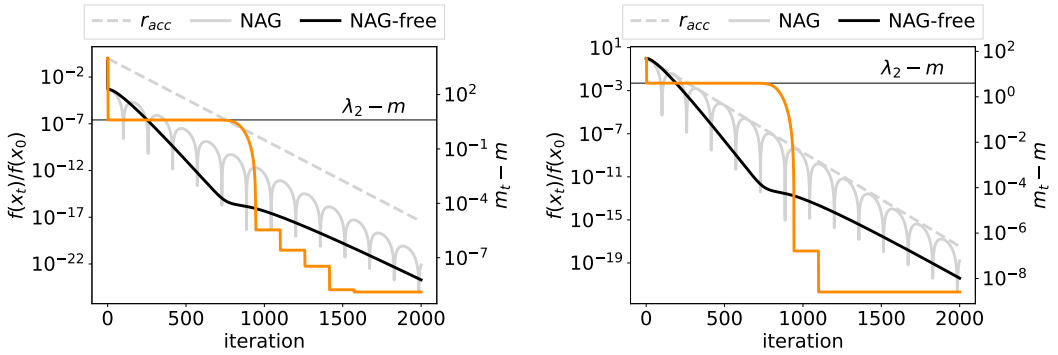
Section 5.4 describes experiments that refer to a "restarting" variant of NAG-free, which is summarized by Algorithm 4. In those experiments, $t_{test} = 100$.

Algorithm 4 NAG-free with restarts.

Input: $T > 0, x_0 = y_0, t_{rest} > 0$
Initialize:
$y \sim x_0 + U[0, 10^{-6}]^d$
$L_0, m_0 \leftarrow \|\nabla f(x_0) - \nabla f(y)\| / \|x_0 - y\|$
for $t = 0, \ldots, T$ **do**
  $y_{t+1} \leftarrow x_t - (1/L_t)\nabla f(x_t)$
  $x_{t+1} \leftarrow y_{t+1} + \dfrac{\sqrt{L_t} - \sqrt{m_t}}{\sqrt{L_t} + \sqrt{m_t}}(y_{t+1} - y_t)$
  $c_{t+1} \leftarrow \|\nabla f(x_{t+1}) - \nabla f(x_t)\| / \|x_{t+1} - x_t\|$
  if $t + 1 \bmod t_{rest} = 0$ **then**
    $L_{t+1} \leftarrow c_{t+1}$
    $m_{t+1} \leftarrow c_{t+1}$
    $y_{t+1} \leftarrow x_{t+1}$
  else
    $L_{t+1} \leftarrow \max(L_t, c_{t+1})$
    $m_{t+1} \leftarrow \min(m_t, c_{t+1})$
  end if
end for

C.5 Matrix factorization

Appendix C.5, like Appendix C.4, also refers to a "restarting" variant of NAG-free, which is summarized by Algorithm 4. In those experiments, $t_{test} = 1$.



(a) $x_0 = [10^{-5}, 1, 1]^\mathsf{T}$.

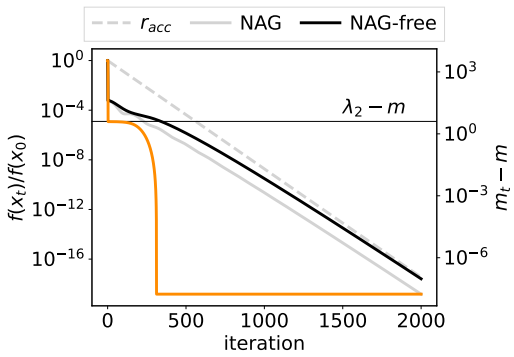
(b) $x_0 = [1, 10^5, 1]^\mathsf{T}$.

Figure 9: Normalized suboptimality gaps for $f(x) = (1/2)x^\mathsf{T}Ax$, with $A = \mathrm{diag}(1, 5, 10^4)$.

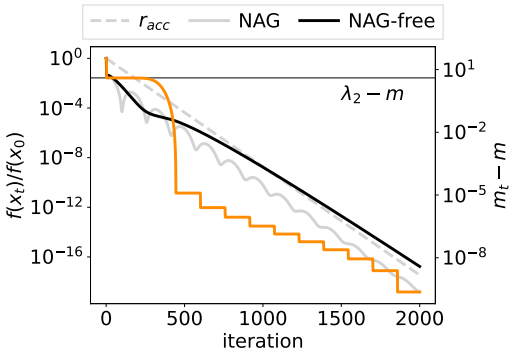
C.6 Further experiments

Finally, we discuss how to interpret some edge cases of the theoretical results derived in Appendix B. Specifically, we are interested in how initial conditions, dimensionality, and distribution of Hessian eigenvalues affect the performance of NAG-free. To study these features, we consider a stylized quadratic objective given by $f(x) = (1/2)x^\mathsf{T}Ax$, where A is a diagonal matrix. We choose this simplified model because it captures the essential features of the local convergence setting of NAG-free in which the theoretical results in Appendix B were derived.

First, we consider the issue of initial conditions. Assumption 4.5 requires the existence of some constant ω such that $\omega x_{1,0}^2 \geq \|x_0\|^2$. In reality, this requirement only serves the technical purpose of allowing us to define a region of local acceleration. Otherwise, if $x_{1,0}^2$ could be arbitrarily small relative to $\|x_0\|^2$, then we would not be able to lower bound the difference $(x_{1,t+1} - x_{1,t})^2$ only in terms of $x_{1,0}$ and $\rho(m_t, m)$, since the perturbation components of $x_{1,t}$ due to other $x_{i,t}$ could dominate the entire solution of $x_{1,t}$, if $x_{1,0}^2$ were sufficiently small relative to $x_{i,0}^2$. With that in mind, we let $A = \mathrm{diag}(1, 5, 10^4)$, and initialize x_0 with $x_0 = [10^{-5}, 1, 1]^\mathsf{T}$, which corresponds to a situation where $\omega \ll 1$, since $\omega = \|x_0\|^2/x_{1,0}^2 \approx \sqrt{2} \cdot 10^{-5}$. Moreover, we choose $L = 10^4$ such that the condition number is $\kappa = L/m = 10^4$, which is among typical values found in practice. Figure 9a shows the normalized suboptimality gaps for NAG initialized with $L = 10^4$ and $m = 1$, for Algorithm 1 initialized with $\bar{L} = L = 10^4$, and the normalized suboptimality gap $r_{\text{acc}}(\kappa)^t$ based on nominal convergence rate for this problem, $r_{\text{acc}}(\kappa) = (\sqrt{\kappa} - 1)/\sqrt{\kappa}$. By normalized suboptimality gap, we mean that the suboptimality gap, which is the same as function values, is divided by the initial suboptimality gap, such that the initial normalized suboptimality gap is unity for all methods. In addition, we plot in gold the estimate gap $m_t - m$, measured on the right-hand y axis. As we can see in Figure 9a, m_t almost immediately reaches the value of $m_t = \lambda_2 = 5$, as depicted by the solid black horizontal line. Indeed, the effect of small $x_{1,0}$ is that, initially, the effective strong convexity parameter *increases* to $\lambda_2 = 5$, which makes the problem better conditioned. Hence, during this stage, Algorithm 1 converges faster at a rate faster than the nominal rate. Once $x_{1,t}^2$ becomes comparable with $\|x_t\|^2$, a second stage begins. In this second stage, m_t approaches the actual strong convexity parameter $m = \lambda_1 = 1$, and Algorithm 1 slows down to the true nominal convergence rate for the problem. In particular, we see that m_t only has to be sufficiently accurate for Algorithm 1 to achieve acceleration. Accordingly, the rate at which the estimate gap $m_t - m$ converges to 0 decelerates as m_t becomes more accurate, as expected. Figure 9b shows a similar result, where instead of reducing $x_{1,0}^2$ relative to $\|x_0\|^2$, we equivalently increase $x_{2,0}^2$, therefore $\|x_0\|^2$, relative to $x_{1,0}^2$. Therefore, as ω becomes large, NAG-free performs *better* initially, and Theorem 4.6 should be applied to each stage with a different effective strong convexity parameter. Conversely, small ω represents the worst-case scenario for NAG-free, since it cannot take advantage initially of increased effective strong convexity parameters. Figure 10a illustrates this remark, where x_0 is initialized with $x_0 = [1, 1, 1]^\mathsf{T}$. We see that m_t approaches $\lambda_1 = m$ much faster than in the previous two edge cases, and NAG-free only matches the nominal convergence rate for this problem.

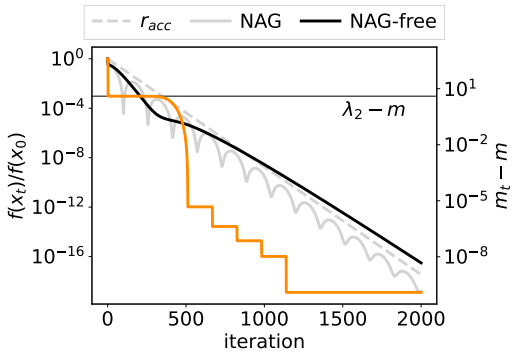


(a) $A = \mathrm{diag}(1, 5, 10^4), x_0 = [1, 1, 1]^\mathsf{T}$.

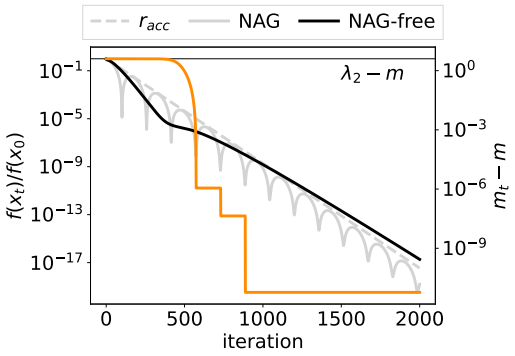


(b) $A = \mathrm{diag}(1, 5I, 10^4), I \in \mathbb{R}^{10^2 \times 10^2}, x_0 = 1$.

Figure 10: Normalized suboptimality gaps for $f(x) = (1/2)x^\mathsf{T}Ax$.



(a) $A = \mathrm{diag}(1, 5I, 10^4), I \in \mathbb{R}^{10^3 \times 10^3}, x_0 = 1$.

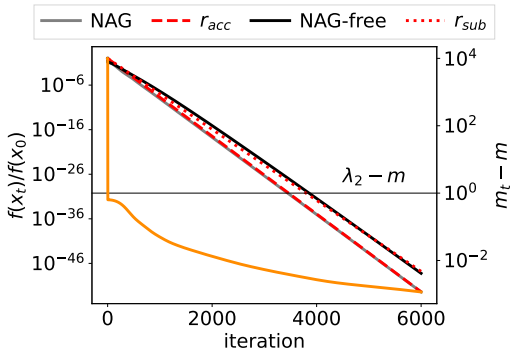


(b) $A = \mathrm{diag}(1, 5I, 10^4), I \in \mathbb{R}^{10^4 \times 10^4}, x_0 = 1$.

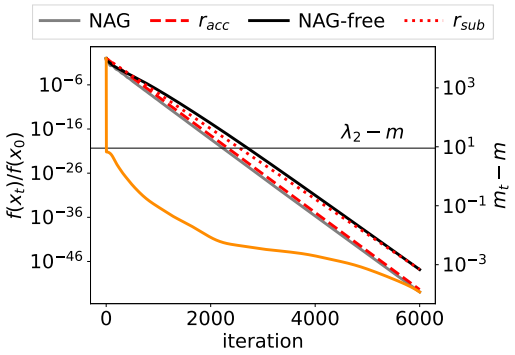
Figure 11: Normalized suboptimality gaps for $f(x) = (1/2)x^\mathsf{T}Ax$.

Next, we consider the issue of dimensionality. Like initial conditions, dimensionality also influences ω under most reasonable initializations of x_0. For instance, if x_0 is initialized uniformly, then on average $x_0 = 1$, a vector of appropriate dimensions whose entries are all 1's. To assess the impact of dimensionality, we continue with a diagonal quadratic objective where $\lambda_1 = 1$ and $\lambda_d = 10^4$, but increase the number of eigenvalues λ_i such that $\lambda_i = \lambda_2 = 5$. That is, $f(x) = (1/2)x^\mathsf{T}Ax$, with $A = \mathrm{diag}(1, 5I, 10^4)$, where $I \in \mathbb{R}^{(d-2)\times(d-2)}$ is the identity matrix with dimension $d - 2$. We consider $d \in \{10^2, 10^3, 10^4\}$. The results, shown in Figures 10b, 11a and 11b, demonstrate that dimensionality has a somewhat similar effect of initial conditions. That is, as the number of eigenvalues λ_i such that $\lambda_i = \lambda_2$ increases, the effective condition number of the problem increases, and NAG-free initially enjoys improved convergence rates. Then, as the relative norm of the coordinates associated with these eigenvalues become comparable with that of $x_{1,t}$, then NAG-free enters a second stage in which it matches the nominal convergence rate of the problem. The influence of dimensionality is slightly more subtle than that of initial conditions but the overall effect is similar. That is because each coordinate $x_{i,t}$ associated with $\lambda_i = \lambda_2$ decays in parallel, independent of d, but as d increases, but before m_t can start to approach m, $\sum_{j=2}^{d-1} x_{j,t}^2 \approx d x_{2,t}^2$ must become comparable with $x_{1,t}^2$. In any case, we see that dimensionality cannot degrade the performance of NAG-free beyond that of nominal acceleration.

The last issue that we consider is the distribution of Hessian eigenvalues. We start by experimenting with uniform distributions. Specifically, we fix $\lambda_1 = m = 1$ and $\lambda_d = L = 10^4$, and consider the remaining eigenvalues uniformly distributed in the interval $[1, \bar{\lambda}_{d-1}]$, for $\bar{\lambda}_{d-1} \in \{2, 10, 100, 10^4\}$. In all experiments, $x_0 = 1 \in \mathbb{R}^d$ and $d = 10^3$. Also, we fix $m_0 = L$. On Figures 12 and 13, in addition to NAG, NAG-free and the nominal convergence curve $r_{\text{acc}}(\kappa)^{2t}$, we also plot a suboptimal convergence curve r_{sub}^{2t}, where $r_{sub} = r_{\text{acc}}(1.2\kappa)$. As we can see, although the performance of

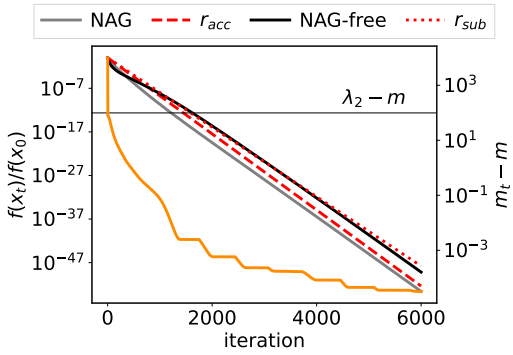


(a) $\bar{\lambda}_{d-1} = 2$.

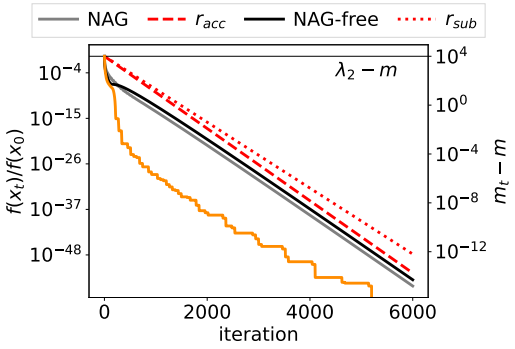


(b) $\bar{\lambda}_{d-1} = 10$.

Figure 12: Normalized suboptimality gaps for $f(x) = (1/2)x^\mathsf{T}Ax$, with $A = \mathrm{diag}(1, D, 10^4)$ and D a diagonal matrix with eigenvalues uniformly distributed in $[1, \bar{\lambda}_{d-1}]$. The dashed and dotted red lines represent $r_{\mathrm{acc}}(\kappa)^{2t}$ and r_{sub}^{2t}, respectively, where $r_{sub} = r_{\mathrm{acc}}(1.2\kappa)$.



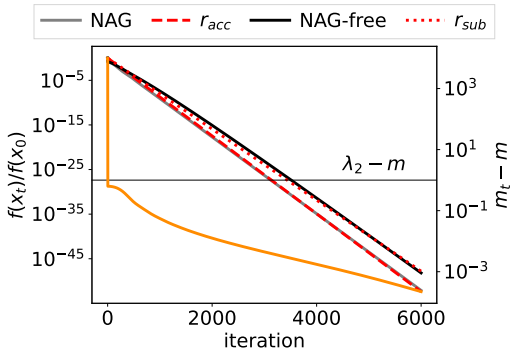
(a) $\bar{\lambda}_{d-1} = 100$.



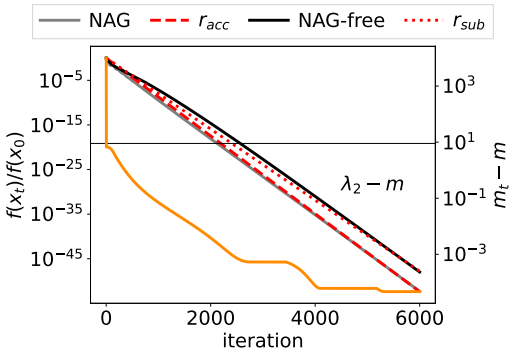
(b) $\bar{\lambda}_{d-1} = 10^4$.

Figure 13: Normalized suboptimality gaps for $f(x) = (1/2)x^\mathsf{T}Ax$, with $A = \mathrm{diag}(1, D, 10^4)$ and D a diagonal matrix with eigenvalues uniformly distributed in $[1, \bar{\lambda}_{d-1}]$. The dashed and dotted red lines represent $r_{\mathrm{acc}}(\kappa)^{2t}$ and r_{sub}^{2t}, respectively, where $r_{sub} = r_{\mathrm{acc}}(1.2\kappa)$.

NAG-free can degrade when the eigenvalues are uniformly distributed, the impact is mild, and the convergence rate is never worse than r_{sub}, which represents a suboptimality factor of only 20% in the condition number. To conclude, we consider the situation where the eigenvalues are clustered. We study three levels of clustering, where there are roughly $20, 40$ and 60% of the total number of eigenvalues. In each experiment, the clusters are evenly distributed in $[m, \bar{\lambda}_{d-1}]$, for $\bar{\lambda}_{d-1} \in \{2, 10, 100, 10^4\}$. In Figures 14 to 19, we see that the results are similar to the case where the eigenvalues are uniformly distributed, and again the rate of convergence of NAG-free is never does worse than r_{sub}, which represents a suboptimality factor of only 20% in the condition number.

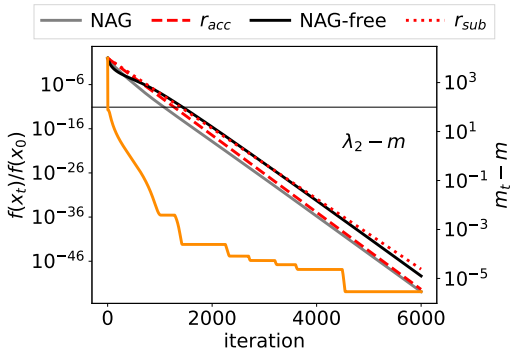


(a) $\bar{\lambda}_{d-1} = 2$.

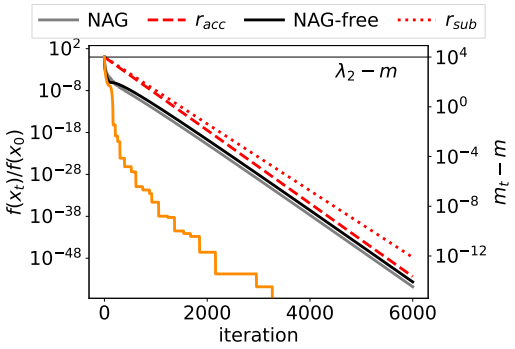


(b) $\bar{\lambda}_{d-1} = 2$.

Figure 14: Normalized suboptimality gaps for $f(x) = (1/2)x^\mathsf{T}Ax$, with $A = \mathrm{diag}(1, C, 10^4)$ and D a diagonal matrix of roughly $0.2d$ clusters of eigenvalues evenly distributed in $[1, \bar{\lambda}_{d-1}]$. The dashed and dotted red lines represent $r_{\mathrm{acc}}(\kappa)^{2t}$ and r_{sub}^{2t}, respectively, where $r_{sub} = r_{\mathrm{acc}}(1.2\kappa)$.

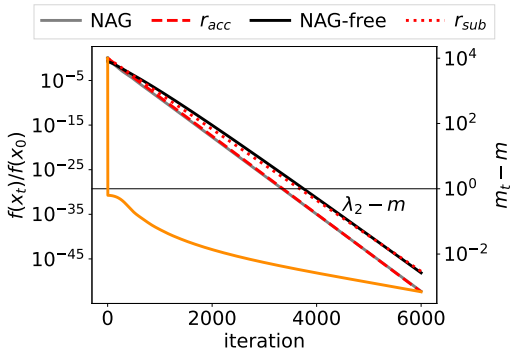


(a) $\bar{\lambda}_{d-1} = 100$.

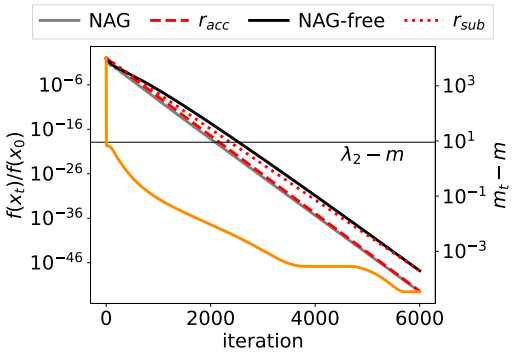


(b) $\bar{\lambda}_{d-1} = 10000$.

Figure 15: Normalized suboptimality gaps for $f(x) = (1/2)x^\mathsf{T}Ax$, with $A = \mathrm{diag}(1, C, 10^4)$ and D a diagonal matrix of roughly $0.2d$ clusters of eigenvalues evenly distributed in $[1, \bar{\lambda}_{d-1}]$. The dashed and dotted red lines represent $r_{\mathrm{acc}}(\kappa)^{2t}$ and r_{sub}^{2t}, respectively, where $r_{sub} = r_{\mathrm{acc}}(1.2\kappa)$.

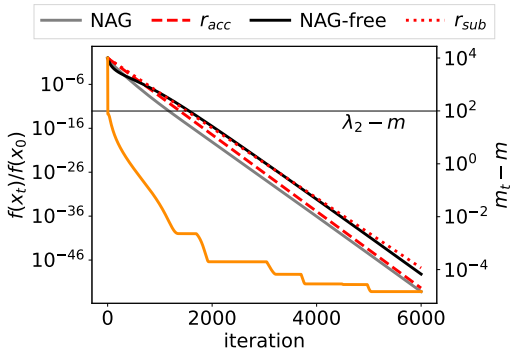


(a) $\bar{\lambda}_{d-1} = 2$.

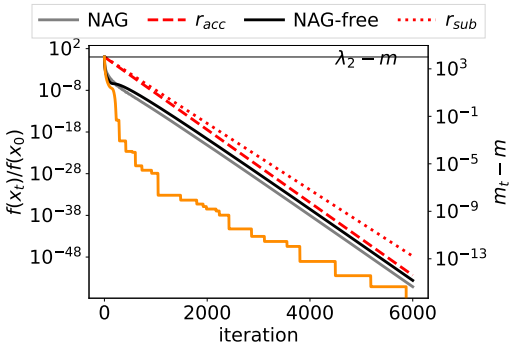


(b) $\bar{\lambda}_{d-1} = 2$.

Figure 16: Normalized suboptimality gaps for $f(x) = (1/2)x^\mathsf{T}Ax$, with $A = \mathrm{diag}(1, C, 10^4)$ and D a diagonal matrix of roughly $0.4d$ clusters of eigenvalues evenly distributed in $[1, \bar{\lambda}_{d-1}]$. The dashed and dotted red lines represent $r_{\mathrm{acc}}(\kappa)^{2t}$ and r_{sub}^{2t}, respectively, where $r_{sub} = r_{\mathrm{acc}}(1.2\kappa)$.

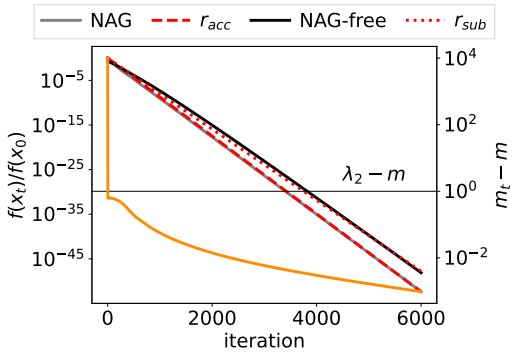


(a) $\bar{\lambda}_{d-1} = 100$.

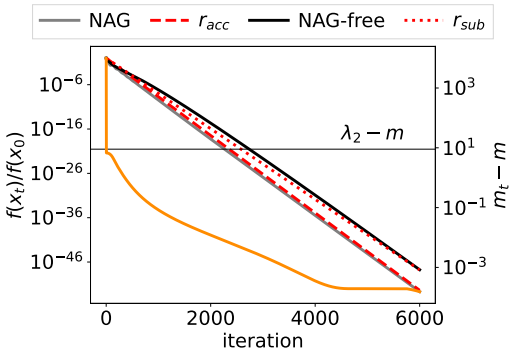


(b) $\bar{\lambda}_{d-1} = 10000$.

Figure 17: Normalized suboptimality gaps for $f(x) = (1/2)x^\mathsf{T}Ax$, with $A = \mathrm{diag}(1, C, 10^4)$ and D a diagonal matrix of roughly $0.4d$ clusters of eigenvalues evenly distributed in $[1, \bar{\lambda}_{d-1}]$. The dashed and dotted red lines represent $r_{\mathrm{acc}}(\kappa)^{2t}$ and r_{sub}^{2t}, respectively, where $r_{sub} = r_{\mathrm{acc}}(1.2\kappa)$.

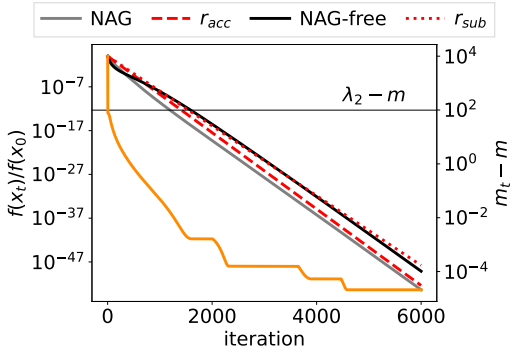


(a) $\bar{\lambda}_{d-1} = 2$.

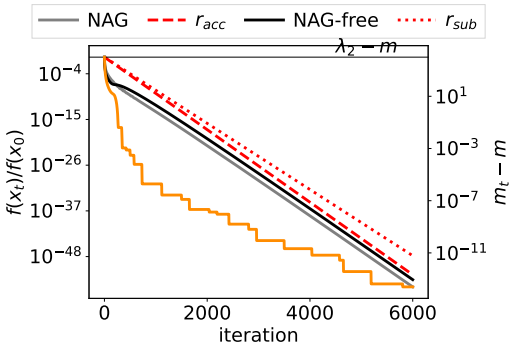


(b) $\bar{\lambda}_{d-1} = 2$.

Figure 18: Normalized suboptimality gaps for $f(x) = (1/2)x^\mathsf{T}Ax$, with $A = \mathrm{diag}(1, C, 10^4)$ and D a diagonal matrix of roughly $0.6d$ clusters of eigenvalues evenly distributed in $[1, \bar{\lambda}_{d-1}]$. The dashed and dotted red lines represent $r_{\mathrm{acc}}(\kappa)^{2t}$ and r_{sub}^{2t}, respectively, where $r_{sub} = r_{\mathrm{acc}}(1.2\kappa)$.



(a) $\bar{\lambda}_{d-1} = 100$.



(b) $\bar{\lambda}_{d-1} = 10000$.

Figure 19: Normalized suboptimality gaps for $f(x) = (1/2)x^\mathsf{T}Ax$, with $A = \mathrm{diag}(1, C, 10^4)$ and D a diagonal matrix of roughly $0.6d$ clusters of eigenvalues evenly distributed in $[1, \bar{\lambda}_{d-1}]$. The dashed and dotted red lines represent $r_{\mathrm{acc}}(\kappa)^{2t}$ and r_{sub}^{2t}, respectively, where $r_{sub} = r_{\mathrm{acc}}(1.2\kappa)$.