Adaptive acceleration without strong convexity priors or restarts

Joao V. Cavalcanti
MIT

Laurent Lessard
Northeastern University

Ashia C. Wilson
MIT

Abstract

In this paper, we propose a parameter-free algorithm for smooth and strongly convex objective problems called NAG-free. To our knowledge, NAG-free is the first adaptive algorithm capable of directly estimating the strong convexity parameter without priors or resorting to restart schemes. We prove that NAG-free converges globally at least as fast gradient descent, and achieves accelerated convergence locally around the minimum if the Hessian is locally smooth at the minimum and other mild additional assumptions hold. We present real-world experiments in which NAG-free is competitive with restart schemes and adapts to better local curvature conditions.

1. Introduction

Accelerated methods are special for achieving optimal convergence rates among first-order optimization algorithms on key problem classes [18]. A notable example is $\mathcal{F}(L, m)$, the class of Lipschitz-smooth, strongly convex functions characterized by the smoothness parameter L and the strong convexity parameter m, which finds applications in signal processing [8], imaging [6] and machine learning [23]. To apply accelerated methods effectively in this setting, both L and m must be known; yet, as noted by Boyd and Vandenberghe [5, p.463], these parameters "are known only in rare cases." While L can be bounded via backtracking [3, 26], "estimating the strong convexity parameter is much more challenging" O'Donoghue and Candès [20, p.3]. As Su et al. [24, 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." In this light, restart schemes have emerged as the only viable approach to handling unknown m [10, Sec. 6]. These methods restart accelerated algorithms (e.g., Nesterov's method) based on adaptive criteria, which can be predetermined [1, 22] or checked at runtime [13, 15–17, 20, 21, 25].

Contributions

In this paper, we propose NAG-free, an adaptive method based on Nesterov's accelerated gradient method (NAG) that, to our knowledge, is the first that estimates Lipschitz-smoothness and strong convexity parameters L and m without restarting. We prove that NAG-free converges globally at least as fast as gradient descent (GD) for problems in $\mathcal{F}(L, m)$. A byproduct of this analysis, and a secondary contribution, is that NAG converges globally at least as fast as GD even if it is parameterized with an overestimate of m in $[m, L]$. Also, we prove that NAG-free achieves acceleration around the minimum $x^\star(f)$ if the Hessian is locally Hölder-smooth at $x^\star(f)$ and some mild additional assumptions hold. We present real-world experiments in which NAG-free is competitive with restart schemes and adapts to better local curvature conditions.

2. Preliminaries

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 Lipschitz-smooth strongly convex functions, defined below.

Definition 1 (Lipschitz-Smooth and Strongly Convex Functions.) *We say that a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ belongs to $\mathcal{F}(L, m)$, the set of Lipschitz-smooth and strongly convex functions, if there exist $L > 0$ and $m > 0$ such that for all $x, y \in \mathbb{R}^d$*

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

and

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

The following will be useful for analyzing NAG-free locally.

Definition 2 (Locally Hölder-smooth Hessian) *Let $f \in \mathcal{F}(L, m)$ be twice differentiable at $x^\star = x^*(f)$. Then, $\nabla^2 f$ is called locally Hölder-smooth at $x^\star$ if there are δ_H, L_H and α_H such that*

$$\|\nabla^2 f(x) - \nabla^2 f(x^\star)\| \leq L_H \|x - x^\star\|^{\alpha_H}, \qquad \forall \|x - x^\star\| \leq \delta_H. \tag{4}$$

3. The NAG-free Algorithm

If $f \in \mathcal{F}(L, m)$, then for all $x \neq y$

$$m \leq c(x, y) := \|\nabla f(x) - \nabla f(y)\| / \|x - y\| \leq L, \tag{5}$$

which follows from standard results on smooth and strongly convex functions [18, theorems 2.1.5 and 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 and x_{t-1} produced by Nesterov's accelerated gradient (NAG) method and letting $c_t = c(x_t, x_{t-1})$ and $\gamma > 1$, we propose to estimate m online via the following update: if $c_t < m_{t-1}$, then $m_t = \min(m_{t-1}/\gamma, c_t)$, otherwise $m_t = m_{t-1}$. This guarantees substantial improvement from one estimate to the next. Moreover, since $c_t \geq m$, it follows that m_t can only take finitely many values, which is important for theoretical reasons that will become clearer later. Then, m_t parametrizes NAG to produce a new iterate, which in turn feeds c_{t+1}, to 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 in $[m, L]$, 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. Algorithm 1 summarizes the complete procedure, which we call **NAG-free**. Although our focus is on estimating m, Algorithm 1 also estimates the Lipschitz constant L through regular backtracking. As x_t converges to the optimum, the descent condition checked by the backtracking subroutine in Algorithm 1 can run into numerical issues. Therefore, in practice we enforce that $f(y_{t+1}) \leq (1 + 10^{-6})(f(x_t) - (1/2L_t)\|\nabla f(x_t)\|^2)$.

Algorithm 1: NAG-free, an algorithm that estimates the strong convexity parameter.

Data: $T > 0, x_0 = y_0, L_0 > 0, \gamma > 1, \gamma_L > 1$
Result: x_T, y_T
$y \sim x_0 + U[0, 10^{-6}]^d$; // initialization
$m_0 \leftarrow \|\nabla f(x_0) - \nabla f(y)\| / \|x_0 - y\|$
$L_0 \leftarrow \max(L_0, m_0)$
for $t = 0, 1, \ldots, T-1$ **do**
  $y_{t+1} \leftarrow x_t - (1/L_t)\nabla f(x_t)$; // NAG #1
  while $f(y_{t+1}) - f(x_t) > -(1/2L_t)\|\nabla f(x_t)\|^2$ **do** ; // BLS
    $L_t \leftarrow \gamma_L L_t$
    $y_{t+1} \leftarrow x_t - (1/L_t)\nabla f(x_t)$;
  $L_{t+1} \leftarrow L_t$
  $x_{t+1} \leftarrow y_{t+1} + \frac{\sqrt{L_t} - \sqrt{m_t}}{\sqrt{L_t} + \sqrt{m_t}}(y_{t+1} - y_t)$; // NAG #2
  $c_{t+1} \leftarrow \|\nabla f(x_{t+1}) - \nabla f(x_t)\| / \|x_{t+1} - x_t\|$; // estimate m
  if $c_{t+1} < m_t$ **then**
    $m_{t+1} \leftarrow \min(m_t/\gamma, c_{t+1})$
  else
    $m_{t+1} \leftarrow m_t$
  end
end

Convergence intuition

Two key features underlie the convergence of NAG-free:

1. **Adaptive interpolation between GD and NAG.** The update rule for x_{t+1} interpolates between GD and NAG. If $m_t \to L$, then the momentum term becomes zero and the update becomes equivalent to GD. Otherwise, if $m_t \to m$, then the update becomes equivalent to NAG. Backtracking preserves the convergence guarantees of both GD and NAG, up to a suboptimality factor of γ_L. 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. Summary of Convergence Guarantees

In this section, we summarize the most important convergence results for NAG-free. The full derivation of global convergence and local acceleration can be found in Appendices A and B, respectively.

4.1. Global Convergence

Theorem 3 *Let $f \in \mathcal{F}(L,m)$ and suppose that $\kappa = L/m \geq 2$. If y_t are iterates generated by Algorithm 1 with some L_0 and $\gamma = 2$, then letting $\bar{\kappa} = \max(L_0, 2L)/m$, we have that*

$$f(y_{t+1}) - f(x^\star) \leq \left(1 - \frac{1}{\bar{\kappa}}\right)^t 8\max(L_0, L)\bar{\kappa}^3\|x_0^\star\|^2. \tag{6}$$

The proof of Theorem 3 can be found in Appendix A. A byproduct of this proof is that even if NAG uses an overestimate of m, it still converges at least as fast as GD.

Corollary 4 *Let $f \in \mathcal{F}(L,\underline{m})$, $m \in [\underline{m}, L]$ and $\bar{\kappa} = L/\underline{m}$. If y_t are generated by NAG with $\underline{m}$ replacing m, then*

$$f(y_{t+1}) - f(x^\star) \leq \left(1 - \frac{1}{\bar{\kappa}}\right)^t 2L\bar{\kappa}^2\|x_0 - x^\star\|^2. \tag{7}$$

4.2. Local Acceleration

To prove that NAG-free achieves acceleration locally, we make the following assumptions.

Assumption 4.1 *The Hessian of f is locally Hölder-smooth at $x^\star$: there are positive numbers δ_H, L_H and α_H such that if $\|x - x^\star\| \leq \delta_H$, then $\|\nabla^2 f(x) - \nabla^2 f(x^\star)\| \leq L_H\|x - x^\star\|^{\alpha_H}$.*

Assumption 4.2 *Given $f \in \mathcal{F}(L,m)$, there is some $L_0 > L$ that can be used by Algorithm 1.*

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 $\nabla^2 f(x^\star)$. Under Assumption 4.1, $\nabla^2 f(x^\star)$ is real symmetric, therefore v_i can be chosen to form an orthonormal basis for $\mathbb{R}^d$. Hence, $x_0 - x^\star$ uniquely decomposes into d unique coordinates $x_{i,0}$ such that $x_0 - x^\star = \sum_{i=1}^d x_{i,0}v_i$. Moreover, if $f \in \mathcal{F}(L,m)$, then $\lambda_i \in [m, L]$. In the following, without loss of generality we assume λ_i ordered by their indices, as in $m = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_d$. Thus, $x_{1,0}$ denotes the m-coordinate of $x_0 - x^\star$, where m is the least eigenvalue of $\nabla^2 f(x^\star)$.

Assumption 4.3 *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 4.4 *There exists some $\omega > 0$ such that $\omega x_{1,0}^2 \geq \|x_0 - x^\star\|^2$.*

Assumption 4.3 simplifies the analysis and is not strictly necessary. Assumption 4.4 prevents pathological cases in which $x_{1,0}$, the m-coordinate of $x_0 - x^\star$, is arbitrarily small compared with the other coordinates. We shall see in Appendix C that violations of this assumption actually improve the performance of NAG-free.

Theorem 5 *Let $f \in \mathcal{F}(L,m)$, suppose that Assumptions 4.1 to 4.4 hold and $\bar{\kappa} = L_0/m > L/m \geq 4$. There is $\epsilon > 0$ such that if $\|x_0 - x^\star\| \leq \epsilon$, then the iterates x_t produced by NAG-free satisfy*

$$\|x_{t+1} - x^\star\| \leq C\left(1 - \frac{1}{\sqrt{\sigma\bar{\kappa}}}\right)^t \|x_0 - x^\star\|,$$

where σ depends on γ, C depends on $\bar{\kappa}$ and ω, with ω given by Assumption 4.4.

The proof of Theorem 5 and a discussion of C and σ can be found in Appendix B. For now, we mention that the suboptimality factor σ is similar to those of restart schemes, where [10, page 167] "the convergence rate is slowed down by roughly a factor four."

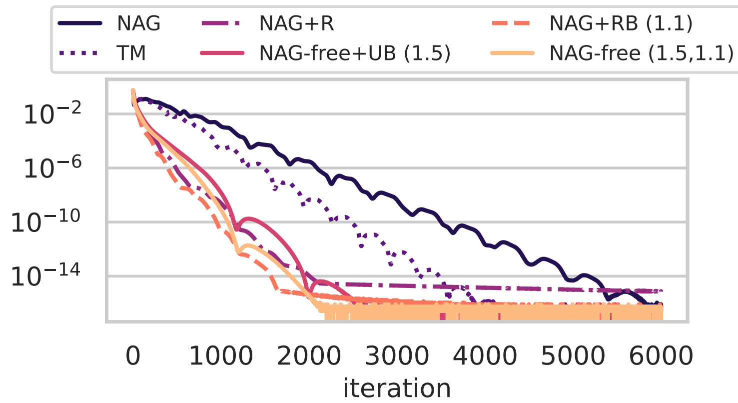

Figure 1: Suboptimality gap $f(x_t) - f(x^\star)$ for logistic regression on PHISHING dataset with $x_0 = 0$. For NAG-free variants, $\gamma = 1.5$ is used. The backtracking factor is 1.1 for NAG+RB and NAG-free. The NAG-free+UB is initialized with $L_0 = \bar{L} \geq L$.

5. Numerical Experiments

We validate NAG-free on regularized logistic regression over several datasets from LIBSVM [7]. Letting $\bar{L}$ denote an upper bound on L, we consider two initializations for NAG-free: $L_0 = \bar{L}$ and $L_0 = \bar{L}/100$. Letting $\eta \leq m$ denote the regularization parameter, for comparison we also consider the following methods: NAG with $L = \bar{L}$ and $m = \eta$; triple momentum method [27, TM] with $L = \bar{L}$ and $m = \eta$; NAG+R, a restart scheme based on [20] with $L = \bar{L}$; NAG+RB, a restart scheme based on [20], where L is found via backtracking.

Appendix C presents several results. In general, NAG-free and restarting methods perform similarly well and outperform TM when the estimates $\bar{L}$ and η are loose, as is often the case, despite TM having better theoretical convergence rates. Figure 1 shows results for the PHISHING dataset. Backtracking only marginally improves performance, suggesting the gains come from better m, not L, estimates. In Appendix C, we confirm this hunch, demonstrating that NAG-free can adapt to better local conditioning and achieve faster convergence than methods with constant parameters.

6. Conclusion

In this paper, we propose NAG-free, a parameter-free algorithm for smooth and strongly convex objective problems. To our knowledge, NAG-free is the first adaptive algorithm capable of directly estimating the strong convexity parameter without priors or resorting to restart schemes. We prove that NAG-free converges globally at least as fast gradient descent, and achieves accelerated convergence locally around the minimum if the Hessian is locally smooth and other mild additional assumptions hold. We present real-world experiments in which NAG-free performs comparably well with restart schemes, demonstrating that it can adapt to better local curvature conditions.

Interesting avenues for future work include coupling parameter estimators with base methods faster than Nesterov's method, such as triple momentum, experimenting with different curvature terms c_t, and estimating L with a similar approach used to estimate m.

References

[1] J.-F. Aujol, L. Calatroni, C. Dossal, H. Labarrière, and A. Rondepierre. Parameter-free fista by adaptive restart and backtracking. *SIAM Journal on Optimization*, 34(4):3259–3285, 2024.

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

[3] 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.

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

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

[6] A. Chambolle and T. Pock. An introduction to continuous optimization for imaging. *Acta Numerica*, 25:161–319, 2016.

[7] 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.

[8] P.L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation*, 4:1168–1200, 2005.

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

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

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

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

[13] G. Lan, Y. Ouyang, and Z. Zhang. Optimal and parameter-free gradient minimization methods for convex and nonconvex optimization. 2023. doi: 10.48550/ARXIV.2310.12139.

[14] 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.

[15] J. Liang, T. Luo, and C.-B. Schonlieb. Improving "fast iterative shrinkage-thresholding algorithm": Faster, smarter, and greedier. *SIAM Journal on Scientific Computing*, 44(3), 2022.

[16] Q. Lin and L. Xiao. An adaptive accelerated proximal gradient methodand its homotopy continuation for sparse optimization. *Comput. Optim. Appl.*, 60:633–674, 2015.

[17] Y. Nesterov. Gradient methods for minimizing composite functions. *Math. Program. Ser. B*, 140:125–161, 2013.

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

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

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

[21] J. Renegar and B. Grimmer. A simple nearly optimal restart scheme for speeding up first-order methods. *Foundations of computational mathematics*, 22(1):211–256, 2022.

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

[23] S. Sra, S. Nowozin, and S. J. Wright. *Optimization for Machine Learning*. The MIT Press, 2011.

[24] 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.

[25] A. Sujanani and R.D.C. Monteiro. Efficient parameter-free restarted accelerated gradient-methods for convex and strongly convex optimization. *Journal of Optimization Theory and Application*, 2025.

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

[27] 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.

Appendix A. Global Convergence

In this section, we prove Theorem 3, which establishes that Algorithm 1 converges globally at least as fast as gradient descent (GD). To this end, we first analyze iterations in which $m_t \geq m$. Then, we analyze iterations in which $m_t < m$, and the transition from the first kind of iteration to the second.

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

The iterations in which $m_t \geq m$ can be expressed as a convex combination of appropriate GD and NAG iterations. We exploit this property to prove that Algorithm 1 converges at least as fast as GD. We use an argument based on 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 $m_t \geq m$.

The Lyapunov function V_t^{GD} is the sum of two functions W_t and U_t. First, we show W_t is a common Lyapunov function for GD and NAG, then we analyze U_t 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

$$y_{t+1}^{\text{GD}} = x_t^{\text{GD}} - (1/L_t)\nabla f(x_t^{\text{GD}}), \tag{8}$$
$$x_{t+1}^{\text{GD}} = y_{t+1}^{\text{GD}}, \tag{9}$$

conforming GD to the algorithmic structure of NAG:

$$y_{t+1}^{\text{NAG}} = x_t^{\text{NAG}} - (1/L_t)\nabla f(x_t^{\text{NAG}}), \tag{10}$$
$$x_{t+1}^{\text{NAG}} = y_{t+1}^{\text{NAG}} + \theta_t(y_{t+1}^{\text{NAG}} - y_t^{\text{NAG}}), \tag{11}$$

where the coefficient θ defining the momentum step in (11) is given by

$$\theta_t = (\sqrt{p_t} - 1)/(\sqrt{p_t} + 1), \qquad p_t = (L_t/m). \tag{12}$$

Similarly, the iterates of Algorithm 1 are given by

$$y_{t+1} = x_t - (1/L_t)\nabla f(x_t), \tag{13}$$
$$x_{t+1} = y_{t+1} + \beta_t(y_{t+1} - y_t), \tag{14}$$

where y_{t+1} and L_t are such that

$$f(y_{t+1}) - f(x_t) \leq -(1/2L_t)\|\nabla f(x_t)\|^2, \tag{15}$$

and β_t is the affine coefficient given by

$$\beta_t = (\sqrt{q_t} - 1)/(\sqrt{q_t} + 1), \qquad q_t = (L_t/m_t), \tag{16}$$

Expressed in the common structure of (8) to (11), GD and NAG can be analyzed with a common Lyapunov function very similar to the one used in [2, section 5.5], and given by

$$W_t(s_t) = \tilde{f}(y_t) + (m/2)\|z_t^\star\|^2, \tag{17}$$

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

$$s_t = (x_t, y_t), \qquad s_t^{\text{GD}} = (x_t^{\text{GD}}, y_t^{\text{GD}}), \qquad \text{and} \qquad s_t^{\text{NAG}} = (x_t^{\text{NAG}}, y_t^{\text{NAG}}), \tag{18}$$

$\tilde{f}$ denotes the objective function with minimum shifted to 0, meaning that

$$\tilde{f} = f - f(x^\star), \tag{19}$$

and $z_s^\star = z_s^\star(x_t, y_t)$ is the pseudo-state defined as

$$z_s^\star = z_s - x^\star, \qquad\qquad z_s = z_s(x_t, y_t) = \begin{cases} x_0 + \sqrt{p_0}(x_0 - y_0), & s = 0, \\ x_t + \sqrt{p_{s-1}}(x_t - y_t), & s \geq 1. \end{cases} \tag{20}$$

Remark 6 *In the definition of W_t, we note that the subscript t determines the subscript of p_0 or p_{t-1} in z_t independently of the subscript of x_t and y_t. So, for example, we have that*

$$\begin{aligned} W_{t+1}(s_t) &= \tilde{f}(y_t) + (m/2)\|x_t + \sqrt{p_t}(x_t - y_t)\|^2 \\ &\neq \tilde{f}(y_{t+1}) + (m/2)\|x_{t+1} + \sqrt{p_t}(x_{t+1} - y_{t+1})\|^2 = W_{t+1}(s_{t+1}). \end{aligned}$$

Remark 7 *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. Thus, V_t^{GD} is the sum of convex functions and is therefore convex.*

In the following, we often use $g_t = \nabla f(x_t)$. For brevity, we also define

$$x_t^\star = x_t - x^\star \qquad\qquad \text{and} \qquad\qquad x_t^y = x_t - y_t. \tag{21}$$

Remark 8 *Superscripts carry over from (8) to (11) to the notation above in the natural way. For example, by g_t^{NAG} we mean $\nabla f(x_t^{\text{NAG}})$ and by $x_t^{\text{GD},\star}$ we mean $x_t^{\text{GD}} - x^\star$.*

Although W_t serves as a Lyapunov function for GD and NAG, we should expect W_t to decrease at a faster rate along NAG iterations than along GD iterations. We now show that W_t decreases at the expected rate for each of the two methods, namely $(1 + \delta(p_t))^{-1}$ for GD and $(1 + \delta(\sqrt{p_t}))^{-1}$ for NAG, where the rate increment δ is defined by

$$\delta(p) = 1/(p - 1). \tag{22}$$

The following rate increments will also be convenient:

$$\delta_t^{\text{GD}} = \delta(p_{t-1}) = 1/(p_{t-1} - 1) \qquad \text{and} \qquad \delta_t^{\text{ACC}} = \delta(\sqrt{p_{t-1}}) = 1/(\sqrt{p_{t-1}} - 1). \tag{23}$$

Lemma 9 *Let $f \in \mathcal{F}(L, m)$ and $y_t^{\text{GD}} = x_t^{\text{GD}} \in \mathbb{R}^d$. If y_{t+1}^{GD} given by (8) and $L_t > 0$ are such that*

$$f(y_{t+1}^{\text{GD}}) - f(x_t^{\text{GD}}) \leq -(1/2L_t)\|g_t^{\text{GD}}\|^2, \tag{24}$$

and x_{t+1}^{GD} is given by (9), then

$$(1 + \delta_{t+1}^{\text{GD}})W_{t+1}(s_{t+1}^{\text{GD}}) - W_t(s_t^{\text{GD}}) \leq -(1/2L_t)\|g_t^{\text{GD}}\|^2. \tag{25}$$

Proof Let $f \in \mathcal{F}(L, m)$ and $\bar{L} \geq L$. Following the procedure described above, we start by expressing $(1+\delta_{t+1}^{\mathrm{GD}})\tilde{f}(y_{t+1}^{\mathrm{GD}}) - \tilde{f}(y_t^{\mathrm{GD}})$ as the sum of two differences:

$$(1+\delta_{t+1}^{\mathrm{GD}})\tilde{f}(y_{t+1}^{\mathrm{GD}}) - \tilde{f}(y_t^{\mathrm{GD}}) = (1+\delta_{t+1}^{\mathrm{GD}})(f(y_{t+1}^{\mathrm{GD}}) - f(y_t^{\mathrm{GD}})) + \delta_{t+1}^{\mathrm{GD}}(f(y_t^{\mathrm{GD}}) - f(x^\star)).$$

If $y_t^{\mathrm{GD}} = x_t^{\mathrm{GD}}$, y_{t+1}^{GD} is given by (8) and (24) holds, then the first difference is bounded as

$$(1+\delta_{t+1}^{\mathrm{GD}})(f(y_{t+1}^{\mathrm{GD}}) - f(y_t^{\mathrm{GD}})) \leq -(1+\delta_{t+1}^{\mathrm{GD}})(1/2L_t)\|g_t^{\mathrm{GD}}\|^2. \tag{26}$$

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

$$\delta_{t+1}^{\mathrm{GD}}(f(y_t^{\mathrm{GD}}) - f(x^\star)) \leq \delta_{t+1}^{\mathrm{GD}}\langle g_t^{\mathrm{GD}}, x_t^{\mathrm{GD},\star}\rangle - \delta_{t+1}^{\mathrm{GD}}(m/2)\|x_t^{\mathrm{GD},\star}\|^2. \tag{27}$$

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

$$\begin{aligned}(1+\delta_{t+1}^{\mathrm{GD}})\|z_{t+1}^{\mathrm{GD},\star}\|^2 - \|z_t^{\mathrm{GD},\star}\|^2 &= (1+\delta_{t+1}^{\mathrm{GD}})\|y_{t+1}^{\mathrm{GD},\star}\|^2 - \|x_t^{\mathrm{GD},\star}\|^2 \\ &= (1+\delta_{t+1}^{\mathrm{GD}})\Big(\frac{\|g_t^{\mathrm{GD}}\|^2}{L_t^2} - \frac{2\langle g_t^{\mathrm{GD}}, x_t^{\mathrm{GD},\star}\rangle}{L_t}\Big) + \delta_{t+1}^{\mathrm{GD}}\|x_t^{\mathrm{GD},\star}\|^2. \end{aligned} \tag{28}$$

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

$$(1+\delta_{t+1}^{\mathrm{GD}})\Big(1 - \frac{1}{p_t}\Big) = \frac{p_t}{p_t - 1}\frac{p_t - 1}{p_t} = 1 \qquad \text{and} \qquad \frac{1+\delta_{t+1}^{\mathrm{GD}}}{p_t} = \frac{p_t/(p_t-1)}{p_t} = \delta_{t+1}^{\mathrm{GD}}.$$

Multiplying (28) by $m/2$, summing the result with (26) and (27), then using the identities above, we obtain

$$\begin{aligned}(1+\delta_t^{\mathrm{GD}})W_{t+1}(s_{t+1}^{\mathrm{GD}}) - W_t(s_t^{\mathrm{GD}}) \leq & -(1+\delta_{t+1}^{\mathrm{GD}})\Big(1 - \frac{1}{p_t}\Big)\frac{1}{2L_t}\|g_t^{\mathrm{GD}}\|^2 \\ & - \Big(\delta_{t+1}^{\mathrm{GD}} - \frac{1+\delta_{t+1}^{\mathrm{GD}}}{p_t}\Big)\langle g_t^{\mathrm{GD}}, x_t^{\mathrm{GD},\star}\rangle \\ \leq & -(1/2L_t)\|g_t^{\mathrm{GD}}\|^2,\end{aligned}$$

proving (25). ■

The analysis of Lyapunov functions like W_t is challenging because it varies with t, so we consider two types of changes for W_t and subsequent Lyapunov functions: the decrease in a fixed W_{t+1} from one iteration s_t to the next s_{t+1} and the increase from W_t to W_{t+1} for the same iteration s_t. For GD, the steps y_t^{GD} and x_t^{GD} coincide, hence both also coincide with z_t^{GD}. Therefore, W_t is effectively the same for all $t \geq 0$ when evaluated at s_t^{GD}, that is

$$W_t(s_t^{\mathrm{GD}}) = \tilde{f}(y_t^{\mathrm{GD}}) + (m/2)\|x_t^{\mathrm{GD},\star}\|^2 = W_{t+1}(s_t^{\mathrm{GD}}).$$

In contrast, the steps y_t^{NAG} and x_t^{NAG} need not coincide. Therefore, as $\sqrt{p_t}$ change with t, each $z_t^{\mathrm{NAG},\star}$ turns into a different affine combination of $x_t^{\mathrm{NAG},\star}$ and $y_t^{\mathrm{NAG},\star}$. That is, for $t \geq 1$

$$\begin{aligned} z_{t+1}^{\mathrm{NAG}}(x_t, y_t) &= x_t^{\mathrm{NAG}} + \sqrt{p_t}(x_t^{\mathrm{NAG}} - y_t^{\mathrm{NAG}}) \\ &\neq x_t^{\mathrm{NAG}} + \sqrt{p_{t-1}}(x_t^{\mathrm{NAG}} - y_t^{\mathrm{NAG}}) = z_t^{\mathrm{NAG}}(x_t, y_t)\end{aligned}$$

due to mismatching $\sqrt{p_t}$ and $\sqrt{p_{t-1}}$. To handle this mismatch, instead of analyzing the difference $(1+\delta_{t+1}^{\mathrm{ACC}})W_{t+1}(s_{t+1}^{\mathrm{NAG}}) - W_t(s_t^{\mathrm{NAG}})$, in the next two results we analyze the difference $(1+\delta_{t+1}^{\mathrm{ACC}})W_{t+1}(s_{t+1}^{\mathrm{NAG}}) - W_{t+1}(s_t^{\mathrm{NAG}})$, and then bound $W_{t+1}(s_{t+1}^{\mathrm{NAG}})$ in terms of $W_t(s_t^{\mathrm{NAG}})$.

Lemma 10 *Let $f \in \mathcal{F}(L, m)$. If y_{t+1}^{NAG} given by (10) and $L_t > 0$ are such that*

$$f(y_{t+1}^{\text{NAG}}) - f(x_t^{\text{NAG}}) \leq -(1/2L_t)\|g_t^{\text{NAG}}\|^2, \tag{29}$$

and x_{t+1}^{NAG} is given by (11), then

$$(1 + \delta_{t+1}^{\text{ACC}})W_{t+1}(s_{t+1}^{\text{NAG}}) - W_{t+1}(s_t^{\text{NAG}}) \leq 0. \tag{30}$$

Proof To prove (30), we start by expressing $(1 + \delta_{t+1}^{\text{ACC}})\tilde{f}(y_{t+1}^{\text{NAG}}) - \tilde{f}(y_t^{\text{NAG}})$ as the sum of three further differences:

$$\begin{aligned}(1 + \delta_{t+1}^{\text{ACC}})\tilde{f}(y_{t+1}^{\text{NAG}}) - \tilde{f}(y_t^{\text{NAG}}) = (1 &+ \delta_{t+1}^{\text{ACC}})(f(y_{t+1}^{\text{NAG}}) - f(x_t^{\text{NAG}})) \\ &+ f(x_t^{\text{NAG}}) - f(y_t^{\text{NAG}}) \\ &+ \delta_{t+1}^{\text{ACC}}(f(x_t^{\text{NAG}}) - f(x^\star)).\end{aligned}$$

If (29) holds, then we bound the first difference as

$$(1 + \delta_{t+1}^{\text{ACC}})(f(y_{t+1}^{\text{NAG}}) - f(x_t^{\text{NAG}})) \leq -(1 + \delta_{t+1}^{\text{ACC}})(1/2L_t)\|g_t^{\text{NAG}}\|^2.$$

Using convexity and applying (3) with $x = x_t^{\text{NAG}}$ and $y = x^\star$, we bound the second and third differences as

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

To address the rest of $(1 + \delta_{t+1}^{\text{ACC}})W_{t+1}(s_{t+1}^{\text{NAG}}) - W_{t+1}(s_t^{\text{NAG}})$, we expand $z_{t+1}^{\text{NAG},\star}$, and then use the definition of θ_t to simplify the resulting expression, as in

$$\begin{aligned}z_{t+1}^{\text{NAG},\star} &= x_{t+1}^{\text{NAG}} + \sqrt{p_t}(x_{t+1}^{\text{NAG}} - y_{t+1}^{\text{NAG}}) - x^\star \\ &= y_{t+1}^{\text{NAG}} + \theta_t(y_{t+1}^{\text{NAG}} - y_t^{\text{NAG}}) + \sqrt{p_t}\theta_t(y_{t+1}^{\text{NAG}} - y_t^{\text{NAG}}) - x^\star \\ &= -(1/L_t)(1 + \theta_t(1 + \sqrt{p_t}))g_t^{\text{NAG}} + \theta_t(1 + \sqrt{p_t})x_t^{\text{NAG},y} + x_t^{\text{NAG},\star} \\ &= -(1/L_t)\sqrt{p_t}g_t^{\text{NAG}} + (\sqrt{p_t} - 1)x_t^{\text{NAG},y} + x_t^{\text{NAG},\star}.\end{aligned}$$

Next, we note that the 2-norm term that goes into $W_{t+1}(s_t^{\text{NAG}})$ is $(m/2)\|x_t^{\text{NAG},\star} + \sqrt{p_t}x_t^{\text{NAG},\star}\|^2$, aand then we write the 2-norm difference in $(1 + \delta_{t+1}^{\text{ACC}})W_{t+1}(s_{t+1}^{\text{NAG}}) - W_{t+1}(s_t^{\text{NAG}})$ as

$$\begin{aligned}(1 + \delta_{t+1}^{\text{ACC}})\frac{m}{2}\|z_{t+1}^{\text{NAG},\star}\|^2 - \frac{m}{2}\|x_t^{\text{NAG},\star} + \sqrt{p_t}x_t^{\text{NAG},\star}\|^2 = &\ \frac{1 + \delta_{t+1}^{\text{ACC}}}{2L_t}\|g_t^{\text{NAG}}\|^2 \\ &- \langle g_t^{\text{NAG}}, x_t^{\text{NAG},y}\rangle \\ &- \delta_{t+1}^{\text{ACC}}\langle g_t^{\text{NAG}}, x_t^{\text{NAG},\star}\rangle \\ &- \frac{m}{2}\sqrt{p_t}\|x_t^{\text{NAG},y}\|^2 \\ &+ \delta_{t+1}^{\text{ACC}}\frac{m}{2}\|x_t^{\text{NAG},\star}\|^2,\end{aligned}$$

where we used the following identities after colons to simplify the coefficients of the terms before colons:

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

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

$$(1+\delta_{t+1}^{\text{ACC}})W_{t+1}(s_{t+1}^{\text{NAG}}) - W_{t+1}(s_t^{\text{NAG}}) \le -(m/2)\sqrt{p_t}\|x_t^{\text{NAG},y}\|^2 \le 0,$$

proving (30). ■

By pairing $W_{t+1}(s_{t+1}^{\text{NAG}})$ with $W_{t+1}(s_t^{\text{NAG}})$, we deferred the problem of mismatching $\sqrt{p_t}$ and $\sqrt{p_{t-1}}$ to obtain (30). We now handle the mismatch problem by bounding $W_{t+1}(s_t^{\text{NAG}})$ in terms of $W_t(s_t^{\text{NAG}})$. In contrast with (30), the inequality we prove next holds for arbitrary x_t and y_t, not necessarily generated by NAG. We therefore drop the "*NAG*" superscript.

Lemma 11 *Let $f \in \mathcal{F}(L,m)$, $y_t, x_t \in \mathbb{R}^d$. If $L_t \ge L_{t-1} \ge m$, then*

$$W_{t+1} \le \frac{p_t^2}{p_{t-1}^2} W_t. \tag{31}$$

Proof The key to prove (31) is to analyze the difference between the mismatching terms

$$\|x_t^\star + \sqrt{p_t}x_t^y\|^2 - \|z_t^\star\|^2 = 2(\sqrt{p_t}-\sqrt{p_{t-1}})\langle x_t^\star, x_t^y\rangle + (p_t - p_{t-1})\|x_t^y\|^2. \tag{32}$$

We split the analysis in two cases, according to the sign of $\langle x_t^\star, x_t^y\rangle$. First, we consider the case $\langle x_t^\star, x_t^y\rangle \ge 0$. Assuming $L_t \ge L_{t-1}$, then $p_t \ge p_{t-1}$, which in turn implies

$$\sqrt{p_t} - \sqrt{p_{t-1}} \le \frac{p_t}{\sqrt{p_t}} - \sqrt{p_{t-1}}\frac{\sqrt{p_{t-1}}}{\sqrt{p_t}} = \frac{p_t - p_{t-1}}{\sqrt{p_t}}. \tag{33}$$

Hence, adding $(p_t - p_{t-1})p_t^{-1}\|x_t^\star\|^2 \ge 0$ to (32) and then using (33), we obtain

$$\begin{aligned}
\|x_t^\star + \sqrt{p_t}x_t^y\|^2 - \|z_t^\star\|^2 &\le 2\frac{p_t - p_{t-1}}{\sqrt{p_t}}\langle x_t^\star, x_t^y\rangle + (p_t - p_{t-1})\|x_t^y\|^2 + \frac{p_t - p_{t-1}}{p_t}\|x_t^\star\|^2\\
&= \frac{p_t - p_{t-1}}{p_t}\|x_t^\star + \sqrt{p_t}x_t^y\|^2.
\end{aligned} \tag{34}$$

In turn, multiplying right and left-hand side of (34) by $m/2$, it follows from the definition (17) that

$$W_{t+1}(s_t) - W_t(s_t) = \frac{m}{2}(\|x_t^\star + \sqrt{p_t}x_t^y\|^2 - \|z_t^\star\|^2) \le \frac{p_t - p_{t-1}}{p_t}W_{t+1}(s_t).$$

Moving terms around then multiplying both sides by p_t/p_{t-1}, we obtain

$$W_{t+1}(s_t) \le \frac{p_t}{p_{t-1}}W_t(s_t) \le \frac{p_t^2}{p_{t-1}^2}W_t(s_t),$$

where the second inequality follows from the fact that $p_t \geq p_{t-1}$.

Now, suppose $\langle x_t^\star, x_t^y \rangle < 0$. Expressing $(p_t - p_{t-1})\|x_t^y\|^2$ in (32) as

$$(p_t - p_{t-1})\|x_t^y\|^2 = (\sqrt{p_t}(\sqrt{p_t} - \sqrt{p_{t-1}}) + \sqrt{p_{t-1}}(\sqrt{p_t} - \sqrt{p_{t-1}}))\|x_t^y\|^2$$

and then adding $\pm(\sqrt{p_t} - \sqrt{p_{t-1}})\|x_t^\star\|^2/\sqrt{p_t}$ to (32) to complete a square, we obtain

$$\begin{aligned}
\|x_t^\star + \sqrt{p_t}x_t^y\|^2 - \|z_t^\star\|^2 &= 2\frac{\sqrt{p_t} - \sqrt{p_{t-1}}}{\sqrt{p_t}}\langle x_t^\star, \sqrt{p_t}x_t^y\rangle + \sqrt{p_t}(\sqrt{p_t} - \sqrt{p_{t-1}})\|x_t^y\|^2 \\
&\quad + \sqrt{p_{t-1}}(\sqrt{p_t} - \sqrt{p_{t-1}})\|x_t^y\|^2 \pm \frac{\sqrt{p_t} - \sqrt{p_{t-1}}}{\sqrt{p_t}}\|x_t^\star\|^2 \\
&= \frac{\sqrt{p_t} - \sqrt{p_{t-1}}}{\sqrt{p_t}}\|x_t^\star + \sqrt{p_t}x_t^y\|^2 + \sqrt{p_{t-1}}(\sqrt{p_t} - \sqrt{p_{t-1}})\|x_t^y\|^2 \\
&\quad - \frac{\sqrt{p_t} - \sqrt{p_{t-1}}}{\sqrt{p_t}}\|x_t^\star\|^2.
\end{aligned} \tag{35}$$

Next, we bound the $\|x_t^y\|^2$ term on (35) using $\|z_t^\star\|^2$ and $\|x_t^\star\|^2$ terms. To this end, we use an elementary inequality for 2-norms. If $a, b \in \mathbb{R}^d$ and $c \in \mathbb{R} \setminus \{0\}$, 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

$$\|a - b\|^2 = \|a\|^2 - 2\langle a, b\rangle + \|b\|^2 \leq (1 + 1/c^2)\|a\|^2 + (1 + c^2)\|b\|^2. \tag{36}$$

Applying (36) with $a = z_t^\star, b = x_t^\star$ and some $c \neq 0$, we obtain

$$\|x_t^y\|^2 = \|x_t^y \pm x_t^\star/\sqrt{p_{t-1}}\|^2 = \frac{1}{p_{t-1}}\|z_t^\star - x_t^\star\|^2 \leq \frac{1 + 1/c^2}{p_{t-1}}\|x_t^\star\|^2 + \frac{1 + c^2}{p_{t-1}}\|z_t^\star\|^2. \tag{37}$$

We then choose c such that

$$\frac{\sqrt{p_t} - \sqrt{p_{t-1}}}{\sqrt{p_{t-1}}}(1 + c^2) = \frac{p_t - p_{t-1}}{p_{t-1}} = \frac{\sqrt{p_t} - \sqrt{p_{t-1}}}{\sqrt{p_{t-1}}}\frac{\sqrt{p_t} + \sqrt{p_{t-1}}}{\sqrt{p_{t-1}}}. \tag{38}$$

Cancelling $(\sqrt{p_t} - \sqrt{p_{t-1}})/\sqrt{p_{t-1}}$ on both sides of (38) yields

$$c^2 = \frac{\sqrt{p_t} + \sqrt{p_{t-1}}}{\sqrt{p_{t-1}}} - 1 = \frac{\sqrt{p_t}}{\sqrt{p_{t-1}}}.$$

Having fixed c as above, it follows that the coefficient multiplying $\|x_t^\star\|^2$ in (37) is

$$\frac{1 + 1/c^2}{p_{t-1}} = \frac{1}{p_{t-1}}\left(1 + \frac{\sqrt{p_{t-1}}}{\sqrt{p_t}}\right) = \frac{\sqrt{p_t} + \sqrt{p_{t-1}}}{p_{t-1}\sqrt{p_t}}. \tag{39}$$

Plugging (38) and (39) back into (37), we obtain

$$\sqrt{p_{t-1}}(\sqrt{p_t} - \sqrt{p_{t-1}})\|x_t^y\|^2 \leq \frac{p_t - p_{t-1}}{\sqrt{p_t}\sqrt{p_{t-1}}}\|x_t^\star\|^2 + \frac{p_t - p_{t-1}}{p_{t-1}}\|z_t^\star\|^2. \tag{40}$$

In turn, plugging (40) back into (35) yields

$$\|x_t^\star + \sqrt{p_t}x_t^y\|^2 - \|z_t^\star\|^2 \le \frac{\sqrt{p_t}-\sqrt{p_{t-1}}}{\sqrt{p_t}}\|x_t^\star + \sqrt{p_t}x_t^y\|^2 + \frac{p_t - p_{t-1}}{p_{t-1}}\|z_t^\star\|^2 + \frac{\sqrt{p_t}-\sqrt{p_{t-1}}}{\sqrt{p_{t-1}}}\|x_t^\star\|^2, \tag{41}$$

where the coefficient multiplying $\|x_t^\star\|^2$ is the result of summing that in (35) and the one in (40)

$$\frac{p_t - p_{t-1}}{\sqrt{p_t}\sqrt{p_{t-1}}} - \frac{\sqrt{p_t}-\sqrt{p_{t-1}}}{\sqrt{p_t}} = \frac{\sqrt{p_t}-\sqrt{p_{t-1}}}{\sqrt{p_t}}\Big(\frac{\sqrt{p_t}+\sqrt{p_{t-1}}}{\sqrt{p_{t-1}}} - 1\Big) = \frac{\sqrt{p_t}-\sqrt{p_{t-1}}}{\sqrt{p_{t-1}}}.$$

Multiplying both sides of (41) by $m/2$ and using the definition (17), we obtain

$$W_{t+1}(s_t) - W_t(s_t) \le \frac{\sqrt{p_t}-\sqrt{p_{t-1}}}{\sqrt{p_t}}W_{t+1}(s_t) + \frac{p_t - p_{t-1}}{p_{t-1}}W_t(s_t) + \frac{\sqrt{p_t}-\sqrt{p_{t-1}}}{\sqrt{p_{t-1}}}\frac{m}{2}\|x_t^\star\|^2.$$

Moving all $W_{t+1}(s_t)$ terms above to the left-hand side, all $W_t(s_t)$ above to the right-hand side and then multiplying both sides by $\sqrt{p_t}/\sqrt{p_{t-1}}$, we get

$$\begin{aligned} W_{t+1}(s_t) &\le \frac{p_t}{p_{t-1}}\frac{\sqrt{p_t}}{\sqrt{p_{t-1}}}W_t(s_t) + \frac{\sqrt{p_t}}{\sqrt{p_{t-1}}}\frac{\sqrt{p_t}-\sqrt{p_{t-1}}}{\sqrt{p_{t-1}}}\frac{m}{2}\|x_t^\star\|^2 \\ &\le \frac{p_t}{p_{t-1}}\frac{\sqrt{p_t}}{\sqrt{p_{t-1}}}W_t(s_t) + \frac{p_t}{p_{t-1}}\frac{\sqrt{p_t}-\sqrt{p_{t-1}}}{\sqrt{p_{t-1}}}\frac{m}{2}\|y_t^\star\|^2, \end{aligned} \tag{42}$$

where the second inequality follows from $\sqrt{p_t} \ge \sqrt{p_{t-1}}$ and $\langle x_t^\star, x_t^y\rangle < 0$, the assumption underpinning the case we are analyzing, which implies

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

Finally, bounding $(m/2)\|y_t^\star\|^2$ by $\tilde{f}(y_t)$ on (42) using (3) with $x = x^\star$ and $y = y_t$, then bounding $\tilde{f}(y_t)$ by W_t directly from the definition of W_t, we obtain

$$W_{t+1}(s_t) \le \frac{p_t}{p_{t-1}}\Big(\frac{\sqrt{p_t}}{\sqrt{p_{t-1}}} + \frac{\sqrt{p_t}-\sqrt{p_{t-1}}}{\sqrt{p_{t-1}}}\Big)W_t(s_t) \le \frac{p_t^2}{p_{t-1}^2}W_t(s_t),$$

where the last inequality follows from

$$\frac{\sqrt{p_t}}{\sqrt{p_{t-1}}} + \frac{\sqrt{p_t}-\sqrt{p_{t-1}}}{\sqrt{p_{t-1}}} \le \frac{p_t}{p_{t-1}},$$

which holds because $\sqrt{p_t} \ge \sqrt{p_{t-1}}$ implies

$$\sqrt{p_{t-1}}(\sqrt{p_t} - \sqrt{p_{t-1}}) \le \sqrt{p_t}(\sqrt{p_t} - \sqrt{p_{t-1}}).$$

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

$$W_{t+1}(s_t) \leq \frac{p_t^2}{p_{t-1}^2} W_t(s_t)$$

holds generically for all s_t, which proves (31). ■

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

$$U_t(s_t) = \begin{cases} \tilde{f}(y_0) + (L_0/2)\|y_0^\star\|^2, & t = 0, \\ \tilde{f}(y_t) + (L_{t-1}/2)\|y_t^\star\|^2, & t \geq 1, \end{cases} \tag{43}$$

where $\tilde{f} = f - f(x^\star)$ and $y_t^\star$ is a pseudo-state defined by

$$y_t^\star = y_t - x^\star. \tag{44}$$

Remark 12 *The subscript t of U_t determines the subscript of L_0 or L_{t-1} independently of the argument of U_t. So, for example, if $t \geq 1$, then*

$$\begin{aligned} U_{t+1}(s_t) &= \tilde{f}(y_t) + (L_t/2)\|y_t^\star\|^2 \\ &\neq \tilde{f}(y_t) + (L_{t-1}/2)\|y_t^\star\|^2 = U_t(s_t). \end{aligned}$$

Analogously to $W_{t+1}(s_{t+1})$ and $W_t(s_t)$, $U_{t+1}(s_{t+1})$ and $U_t(s_t)$ have a mismatch in the coefficients of their 2-norm terms, in this case $(L_t/2)\|y_{t+1}^\star\|^2$ and $(L_{t-1}/2)\|y_t^\star\|^2$. Hence, as with W_t, we pair $U_{t+1}(s_{t+1})$ with $U_{t+1}(s_t)$ instead of $U_t(s_t)$ to avoid the mismatch and then address the mismatch problem immediately after. In contrast with the first piece, however, we analyze U_t along NEST iterations explicitly.

Lemma 13 *Let $f \in \mathcal{F}(L, m)$. If $s_{t+1} = (x_{t+1}, y_{t+1})$ is given by (13) and (14), then*

$$(1 + \delta_{t+1}^{\mathrm{GD}})U_{t+1}(s_{t+1}) - U_{t+1}(s_t) \leq L_t\langle x_t^y, x_t^\star\rangle - (L_t/2)\|x_t^y\|^2. \tag{45}$$

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

$$(1 + \delta_{t+1}^{\mathrm{GD}})\tilde{f}(y_{t+1}) - \tilde{f}(y_t) \leq (1 + \delta_{t+1}^{\mathrm{GD}})(f(y_{t+1}) - f(x_t)) + (1 + \delta_{t+1}^{\mathrm{GD}})(f(x_t) - f(x^\star)).$$

By assumption y_{t+1} is given by (13), with y_{t+1} and L_t such that (15) holds. Therefore

$$(1 + \delta_{t+1}^{\mathrm{GD}})(f(y_{t+1}) - f(x_t)) \leq -(1 + \delta_{t+1}^{\mathrm{GD}})(1/2L_t)\|g_t\|^2.$$

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

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

Then, we put the two bounds together to get

$$(1+\delta_{t+1}^{\mathrm{GD}})\tilde{f}(y_{t+1}) - \tilde{f}(y_t) \leq -\frac{1+\delta_{t+1}^{\mathrm{GD}}}{2L_t}\|g_t\|^2 + (1+\delta_{t+1}^{\mathrm{GD}})\langle g_t, x_t^\star\rangle - \delta_{t+1}^{\mathrm{GD}}\frac{L_t}{2}\|x_t^\star\|^2, \tag{46}$$

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

$$(1+\delta_{t+1}^{\mathrm{GD}})\frac{m}{2} = \frac{p_t}{p_t-1}\frac{m}{2} = \delta_{t+1}^{\mathrm{GD}}\frac{L_t}{2}.$$

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

$$(1+\delta_{t+1}^{\mathrm{GD}})\|y_{t+1}^\star\|^2 = (1+\delta_{t+1}^{\mathrm{GD}})\Big(\frac{1}{L_t^2}\|g_t\|^2 - \frac{2}{L_t}\langle g_t, x_t^\star\rangle + \|x_t^\star\|^2\Big),$$
$$\|y_t^\star\|^2 = \|x_t^y\|^2 - 2\langle x_t^y, x_t^\star\rangle + \|x_t^\star\|^2.$$

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

$$\begin{aligned}\frac{L_t}{2}((1+\delta_{t+1}^{\mathrm{GD}})\|y_{t+1}^\star\|^2 - \|y_t^\star\|^2) &= (1+\delta_{t+1}^{\mathrm{GD}})\Big(\frac{\|g_t\|^2}{2L_t} - \langle g_t, x_t^\star\rangle\Big)\\ &\quad + \frac{L_t}{2}(-\|x_t^y\|^2 + 2\langle x_t^y, x_t^\star\rangle + \delta_{t+1}^{\mathrm{GD}}\|x_t^\star\|^2).\end{aligned} \tag{47}$$

Finally, combining (46) and (47), several terms cancel each other and we are left with

$$(1+\delta_{t+1}^{\mathrm{GD}})U_{t+1}(s_{t+1}) - U_t(s_t) \leq L_t\langle x_t^y, x_t^\star\rangle - \frac{L_t}{2}\|x_t^y\|^2,$$

proving (45). ■

Lemma 14 *Let $f \in \mathcal{F}(L, m)$. If $L_t \geq L_{t-1}$, then*

$$U_{t+1} \leq \frac{p_t}{p_{t-1}}U_t. \tag{48}$$

Proof Expanding $U_{t+1}(s_t) - U_t(s_t)$, multiplying the result by L_{t-1}/L_{t-1}, using that $\frac{1}{2}L_{t-1}\|y_t^\star\|^2 \leq U_t(s_t)$ and assuming $L_t \geq L_{t-1}$, we obtain

$$U_{t+1}(s_t) - U_t(s_t) = \frac{L_t - L_{t-1}}{2}\|y_t^\star\|^2 = \frac{L_t - L_{t-1}}{L_{t-1}}\frac{L_{t-1}}{2}\|y_t^\star\|^2 \leq \frac{L_t - L_{t-1}}{L_{t-1}}U_t(s_t).$$

Multiplying the right-hand side by m/m to substitute p_t and p_{t-1} for L_t and L_{t-1}, and then moving $-U_t(s_t)$ to the right-hand side, we get

$$U_{t+1}(s_t) \leq \frac{p_t}{p_{t-1}}U_t(s_t).$$

Since s_t is arbitrary, (48) follows. ■

With Lemmas 9 to 11, 13 and 14, we can prove the main result for NEST iterations in which $m_t > m$, using the Lyapunov function V_t^{GD} given by

$$V_t^{\mathrm{GD}} = \begin{cases} W_0 + (\bar{\alpha}_0/\sqrt{p_0})U_0, & t = 0,\\ W_t + (\bar{\alpha}_{t-1}/\sqrt{p_{t-1}})U_t, & t \geq 1,\end{cases} \quad \text{with} \quad \bar{\alpha}_t = 1 - \alpha_t \quad \text{and} \quad \alpha_t = \beta_t/\theta_t. \tag{49}$$

Remark 15 *The subscript t on V_t^{GD} determines the subscripts on $W_0 + (\bar{\alpha}_0/\sqrt{p_0})U_0$ or $W_t + (\bar{\alpha}_{t-1}/\sqrt{p_{t-1}})U_t$ independently of the argument.*

First, we show that $V_t^{\mathrm{GD}} \geq 0$ for iterations in which $m_t > m$. For future reference, we also show that $\bar{\alpha}_j$ is nonincreasing for all $0 \leq j \leq t$.

Lemma 16 *If $m_t \geq m$ and, in addition, L_t and m_t are respectively nondecreasing and nonincreasing, then $V_j^{\mathrm{GD}} \geq 0$ and $\bar{\alpha}_j$ is nonincreasing for all $0 \leq j \leq t$.*

Proof The assumptions that $m_t \geq m$ and that m_t is nonincreasing imply that $m_j \geq m$ for all $0 \leq j \leq t$. Moreover, $m_t \geq m$ implies that $q_t = L_t/m_t \leq L_t/m = p_t$, therefore

$$\beta_t = \frac{\sqrt{q_t} - 1}{\sqrt{q_t} + 1} \leq \frac{\sqrt{q_t} - 1}{\sqrt{q_t} + 1} = \theta_t.$$

Hence, $\beta_j \leq \theta_j$ for all $0 \leq j \leq t$. Therefore, $\alpha_j, \bar{\alpha}_j \in [0, 1]$ and, in turn, $V_j^{\mathrm{GD}} \geq 0$ for all $0 \leq j \leq t$. Then, expanding β_t and θ_t in α_t, we obtain

$$\frac{\beta_t}{\theta_t} = \frac{\sqrt{q_t} - 1}{\sqrt{q_t} + 1}\frac{\sqrt{p_t} + 1}{\sqrt{p_t} - 1} = \frac{\sqrt{L_t} - \sqrt{m_t}}{\sqrt{L_t} + \sqrt{m_t}}\frac{\sqrt{L_t} + \sqrt{m}}{\sqrt{L_t} - \sqrt{m}} = \frac{L_t - (\sqrt{m}_t - \sqrt{m})\sqrt{L_t} - \sqrt{m_t m}}{L_t + (\sqrt{m}_t - \sqrt{m})\sqrt{L_t} - \sqrt{m_t m}}. \tag{50}$$

Letting $l = L_t$, $d = m_t - m \geq 0$ and $a = \sqrt{m_t m}$, then after simplifying several terms, we obtain

$$\begin{aligned}\frac{\partial}{\partial l}(50) = \frac{\partial}{\partial l}\frac{l - d\sqrt{l} - a}{l + d\sqrt{l} - a} &= \frac{(1 - d/2\sqrt{l})(l + d\sqrt{l} - a) - (1 + d/2\sqrt{l})(l - d\sqrt{l} - a)}{(l + d\sqrt{l} - a)^2} \\ &= \frac{d\sqrt{l} + ad/\sqrt{l}}{(l + d\sqrt{l} - a)^2} \geq 0.\end{aligned}$$

That is, α_t is nondecreasing in L_t while α_t is decreasing in m_t, because β_t is decreasing in m_t and θ_t is not a function of m_t. By assumption L_t and m_t are respectively nondecreasing and nonincreasing, therefore α_j is nondecreasing, so that $\bar{\alpha}_j$ is nonincreasing for all $0 \leq j \leq t$. ■

Lemma 17 *Let $f \in \mathcal{F}(L, m)$, $L_t \geq m_t$ and let s_{t+1} denote the iterate generated by Algorithm 1 from s_t. If $m_t \geq m$, then*

$$(1 + \delta_{t+1}^{\mathrm{GD}})V_{t+1}^{\mathrm{GD}}(s_{t+1}) \leq V_{t+1}^{\mathrm{GD}}(s_t). \tag{51}$$

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

To show that Algorithm 1 iterations are a convex combination of GD and NAG iterations, we consider fictitious "one-shot" GD and NAG iterations taking the value of T into account and appropriately initialized at a given iteration of Algorithm 1. We let $y_t^{\mathrm{GD}} = x_t^{\mathrm{GD}} = x_t^{\mathrm{NAG}} = x_t$ 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 NEST iteration 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 (8) to (11). Then, GD, NAG and Algorithm 1 produce the same descent step:

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

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

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

It follows that, 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 (49), the coefficients defining the convex combination are given by

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

Likewise, $y_{t+1}^{\text{GD}} = y_{t+1}^{\text{NAG}} = y_{t+1}$ implies $y_{t+1} = \bar{\alpha}_t y_{t+1}^{\text{GD}} + \alpha_t y_{t+1}^{\text{NAG}}$ so, in fact, the entire iteration of Algorithm 1 can be expressed as a convex combination of GD and NAG iterations, 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 iterations 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_t is convex (see Theorem 7), we can bound $W_{t+1}(s_{t+1})$ in terms of GD and NAG iterations, as in

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

Then, it follows from $W_{t+1}(s_t) = \bar{\alpha}_t W_{t+1}(s_t) + \alpha_t W_{t+1}(s_t)$ that

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

Since $y_{t+1}^{\text{NAG}} = y_{t+1}$ and $x_t^{\text{NAG}} = x_t$, then the fact that y_{t+1} and L_t satisfy (15) implies that

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

Moreover, $L_t > 0$ is nondecreasing and $m_t > 0$ is nonincreasing. Therefore, Lemma 10 applies because Lemma 10 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 10 combined with both the fact that $\delta_{t+1}^{\text{ACC}} \geq \delta_{t+1}^{\text{GD}}$ and that $\alpha_t > 0$, imply

$$\alpha_t\Big((1 + \delta_{t+1}^{\text{ACC}})W_{t+1}(s_{t+1}^{\text{NAG}}) - W_{t+1}(s_t) + (\delta_{t+1}^{\text{GD}} - \delta_{t+1}^{\text{ACC}})W_{t+1}(s_{t+1}^{\text{NAG}})\Big) \leq 0. \tag{52}$$

The natural next move would be to address $(1 + \delta_{t+1}^{\text{GD}})W_{t+1}(s_{t+1}^{\text{GD}}) - W_{t+1}(s_t)$ in an analogous way. The caveat, however, is that although Lemma 10 applies to NAG iterations with arbitrary x_t^{NAG} and y_t^{NAG}, the same is not true of Lemma 9. That is, Lemma 9 applies to consecutive GD iterations, requiring that $y_t^{\text{GD}} = x_t^{\text{GD}}$. Hence, to be able to apply Lemma 9, we add $\mp W_{t+1}(s_t^{\text{GD}})$ to the difference involving W_{t+1}, 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

$$s_t^{\text{GD}} = (x_t^{\text{GD}}, y_t^{\text{GD}}) = (x_t, x_t). \tag{53}$$

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

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

Therefore, since $L_t > 0$, Lemma 9 applies with s_t^{GD} given by (53), and implies that

$$(1 + \delta_{t+1}^{\text{GD}})W_{t+1}(s_{t+1}^{\text{GD}}) - W_{t+1}(s_t^{\text{GD}}) \leq -(1/2L_t)\|g_t^{\text{GD}}\|^2 = -(1/2L_t)\|g_t\|^2. \tag{54}$$

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

$$z_t^{\text{GD},\star} = x_t^{\text{GD},\star} + \sqrt{p_{t-1}}x_t^{\text{GD},y} = x_t^{\text{GD},\star} = x_t^\star \qquad \text{and} \qquad f(y_t^{\text{GD}}) = f(x_t),$$

and it follows that

$$W_{t+1}(s_t^{\text{GD}}) - W_{t+1}(s_t) = f(x_t) - f(y_t) + (m/2)(\|x_t^\star\|^2 - \|x_t^\star + \sqrt{p_t}x_t^y\|^2). \tag{55}$$

Applying (3) with $x = x_t$ and $y = y_t$, then using the fact that $2\langle g_t, x_t^y\rangle \leq (1/L_t)\|g_t\|^2 + L_t\|x_t^y\|^2$, we obtain

$$f(x_t) - f(y_t) \leq \langle g_t, x_t^y\rangle - (m/2)\|x_t^y\|^2 \leq (1/2L_t)\|g_t\|^2 + ((L_t - m)/2)\|x_t^y\|^2.$$

Hence, expanding $\|x_t^\star + \sqrt{p_t}x_t^y\|^2$ on (55) and then using the above inequality, we get

$$\begin{aligned} W_{t+1}(s_t^{\text{GD}}) - W_{t+1}(s_t) &\leq \frac{1}{L_t}\|g_t\|^2 + \frac{L_t - m}{2}\|x_t^y\|^2 + \frac{m}{2}(-2\sqrt{p_t}\langle x_t^y, x_t^\star\rangle - p_t\|x_t^y\|^2) \\ &= \frac{1}{2L_t}\|g_t\|^2 - \frac{m}{2}\|x_t^y\|^2 - \sqrt{L_t m}\langle x_t^y, x_t^\star\rangle, \end{aligned} \tag{56}$$

where $m\sqrt{p_t} = \sqrt{L_t m}$ follows directly (12). Then, combining (54) and (56) yields

$$(1 + \delta_{t+1}^{\text{GD}})W_{t+1}(s_{t+1}^{\text{GD}}) \mp W_{t+1}(s_t^{\text{GD}}) - W_{t+1}(s_t) \le -\sqrt{L_t m}\langle x_t^y, x_t^\star\rangle. \tag{57}$$

Therefore, since $\delta_{t+1}^{\text{GD}} \le \delta_{t+1}^{\text{ACC}}$, combining (52) and (57), we obtain

$$\begin{aligned}(1 + \delta_{t+1}^{\text{GD}})W_{t+1}(s_{t+1}) - W_{t+1}(s_t) \le\ & \bar{\alpha}_t((1 + \delta_{t+1}^{\text{GD}})W_{t+1}(s_{t+1}^{\text{GD}}) \mp W_{t+1}(s_t^{\text{GD}}) - W_{t+1}(s_t)) \\ & + \alpha_t\Big((1 + \delta_{t+1}^{\text{ACC}})W_{t+1}(s_{t+1}^{\text{NAG}}) - W_{t+1}(s_t)\Big) \\ & + \alpha_t(\delta_{t+1}^{\text{GD}} - \delta_{t+1}^{\text{ACC}})W_{t+1}(s_{t+1}^{\text{NAG}}) \\ \le\ & -\bar{\alpha}_t\sqrt{L_t m}\langle x_t^y, x_t^\star\rangle. \end{aligned} \tag{58}$$

Next, we address the difference on $(1 + \delta_{t+1}^{\text{GD}})V_{t+1}^{\text{GD}}(s_{t+1}) - V_{t+1}^{\text{GD}}(s_t)$ involving U_t. Lemma 13 implies that

$$(\bar{\alpha}_t/\sqrt{p_t})((1 + \delta_{t+1}^{\text{GD}})U_{t+1}(s_{t+1}) - U_{t+1}(s_t)) \le (\bar{\alpha}_t/\sqrt{p_t})L_t\langle x_t^y, x_t^\star\rangle = \bar{\alpha}_t\sqrt{L_t m}\langle x_t^y, x_t^\star\rangle, \tag{59}$$

since $L_t/\sqrt{p_t} = \sqrt{L_t m}$. Then, combining (58) with (59) yields

$$(1 + \delta_{t+1}^{\text{GD}})V_{t+1}^{\text{GD}}(s_{t+1}) \le V_{t+1}^{\text{GD}}(s_t),$$

proving (51). ■

Lemma 18 *Let $f \in \mathcal{F}(, m)$. If $L_t \ge L_{t-1} \ge m$ and $m_t \le m_{t-1} \le L$, then*

$$V_{t+1}^{\text{GD}} \le \frac{p_t^2}{p_{t-1}^2}V_t^{\text{GD}}. \tag{60}$$

Proof If $L_t \ge L_{t-1} \ge m$ and $m_t \le m_{t-1} \le L$, then Lemma 11 and Lemma 14 apply. It also follows that $\sqrt{p_t} \ge \sqrt{p_{t-1}}$ and, by Lemma 16, that $\bar{\alpha}_t \le \bar{\alpha}_{t-1}$ Thus, $\bar{\alpha}_t/\sqrt{p_t} \le \bar{\alpha}_{t-1}/\sqrt{p_{t-1}}$. Hence, combining Lemma 11 and Lemma 14 and then using the definition of V_t^{GD}, we obtain

$$V_{t+1}^{\text{GD}} = W_{t+1} + \frac{\bar{\alpha}_t}{\sqrt{p}_t}U_{t+1} \le \frac{p_t^2}{p_{t-1}^2}W_t + \frac{\bar{\alpha}_{t-1}}{\sqrt{p}_{t-1}}\frac{p_t}{p_{t-1}}U_t \le \frac{p_t^2}{p_{t-1}^2}\Big(W_t + \frac{\bar{\alpha}_{t-1}}{\sqrt{p}_{t-1}}U_t\Big) = \frac{p_t^2}{p_{t-1}^2}V_t^{\text{GD}},$$

proving (60). ■

Theorem 19 *Let $f \in \mathcal{F}(L, m)$ and let s_t denote the iterates generated by Algorithm 1. If $m_t \ge m$, then*

$$V_{t+1}^{\text{GD}}(s_{t+1}) \le 2\max(L, L_0)\frac{p_t^2}{p_0^2}\|x_0 - x^\star\|^2\prod_{i=1}^{t+1}(1 + \delta_i^{\text{GD}})^{-1}. \tag{61}$$

Proof Under the above assumptions, Theorem 17 and Theorem 18 hold. Hence, combining (51) and (60), for all s_{t+1} and s_t such that $m_t \geq m$ we have that

$$V_{t+1}^{\mathrm{GD}}(s_{t+1}) \leq (1+\delta_{t+1}^{\mathrm{GD}})^{-1} \frac{p_t^2}{p_{t-1}^2} V_t^{\mathrm{GD}}(s_t). \tag{62}$$

We proceed with an inductive argument based on (62). To establish the base case, we apply (2) with $y = y_0$ and $x = x^\star$, obtaining $\tilde{f}(y_0) \leq (L/2)\|y_0^\star\|^2$. Then, since $y_0 = x_0$, it follows that $z_0^\star = x_0^\star = y_0^\star$, and

$$\begin{aligned} W_0(s_0) &= \tilde{f}(y_0) + (m/2)\|x_0^\star\|^2 \leq ((L+m)/2)\|x_0^\star\|^2 \leq L\|x_0^\star\|^2, \\ U_0(s_0) &= \tilde{f}(y_0) + (L_0/2)\|y_0^\star\|^2 \leq \max(L, L_0)\|x_0^\star\|^2. \end{aligned}$$

Since $\bar{\alpha}_0/\sqrt{p_0} \in [0,1]$, the above inequalities imply

$$V_0^{\mathrm{GD}}(s_0) = W_0(s_0) + (\bar{\alpha}_0/\sqrt{p_0})U_0(s_0) \leq 2\max(L, L_0)\|x_0^\star\|^2. \tag{63}$$

Moreover, $W_1 = W_0$ and $U_1 = U_0$, so that $V_1^{\mathrm{GD}} = V_0^{\mathrm{GD}}$. Hence, if $m_1 \geq m$, then combining (51) with (63), we obtain

$$V_1^{\mathrm{GD}}(s_1) \leq 2\max(L, L_0)\|x_0^\star\|^2(1+\delta_1^{\mathrm{GD}})^{-1}. \tag{64}$$

Having established the base case (64), suppose that

$$V_{j+1}^{\mathrm{GD}}(s_{j+1}) \leq 2\max(L, L_0)\frac{p_j^2}{p_0^2}\|x_0^\star\|^2 \prod_{i=1}^{j+1}(1+\delta_i^{\mathrm{GD}})^{-1}, \tag{65}$$

holds for all $0 \leq j \leq t-1$ such that $m_j \geq m$. Then, suppose $m_t \geq m$. Since the m estimates are nonincreasing, it follows that $m_j \geq m$ for all $0 \leq j \leq t$. Hence, plugging the induction hypothesis (65) with $j = t-1$ into (62), the p_{t-1}^2 term on the numerator of (65) and on the denominator of (62) cancel each other and we obtain

$$V_{t+1}^{\mathrm{GD}}(s_{t+1}) \leq 2\max(L, L_0)\frac{p_t^2}{p_0^2}\|x_0^\star\|^2 \prod_{i=1}^{t+1}(1+\delta_i^{\mathrm{GD}})^{-1}.$$

Therefore, we conclude by induction that (65) holds for all $j \geq 0$, proving (61). ■

The same arguments above directly imply Theorem 4.

Proof [Proof of Theorem 4] The proof of Theorem 3 does not use of the particular dynamics of m_t induced by NAG-free (Algorithm 1), and the initialization of m_0 is also not important as long as $m_0 \in [m, L]$. Hence, the same arguments also apply to the analysis of NAG (10) and (11). In this case, $L_t \equiv L$ and $m_t \equiv m$ for all t. Therefore, denoting by $s_t = (x_t, y_t)$ the iterates of NAG (10) and (11) and using the definition of V_t^{GD}, (49), it follows that

$$f(y_{t+1}) - f(x^\star) \leq V_{t+1}^{\mathrm{GD}}(s_{t+1}) \leq \left(1 - \frac{1}{\kappa}\right)^t 2L\kappa^2\|x_0^\star\|^2.$$

■

A.2. Case 2: $m_t < m$

In the previous section, we analyzed iterations in which $m_t \geq m$. Now, we analyze iterations in which $m_t < m$ and also the iteration t in which $m_t \geq m$ and $m_{t+1} < m$. Since m_t are nonincreasing, there is a most one such transition iteration.

In Theorem 19, we proved that if $m_t \geq m$, then Algorithm 1 converges at least as fast as GD. Now, we prove that if $m_t \geq m$, then Algorithm 1 converges evvn faster. Specifically, if $m_t < m$, then Algorithm 1 achieves the accelerated rate $(1 + \delta(\sqrt{\bar{\kappa}_t}))^{-1}$, where

$$\hat{\delta}_t^{\text{ACC}} = \begin{cases} 1/(\sqrt{q_0} - 1), & t = 0, \\ 1/(\sqrt{q_{t-1}} - 1), & t \geq 1. \end{cases} \tag{66}$$

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

$$V_t^{\text{ACC}}(s_t) = \begin{cases} \tilde{f}(y_0) + (m_0/2)\|w_0^\star\|^2, & t = 0, \\ \tilde{f}(y_t) + (m_{t-1}/2)\|w_t^\star\|^2, & t \geq 1, \end{cases} \tag{67}$$

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

$$w_t^\star = w_t - x^\star, \qquad w_t = \begin{cases} x_0 + \sqrt{q_0}(x_0 - y_0), & t = 0, \\ x_t + \sqrt{q_{t-1}}(x_t - y_t), & t \geq 1. \end{cases} \tag{68}$$

We first prove the descending bound and then prove the ascending one.

Lemma 20 *Let $f \in \mathcal{F}(L, m)$, and let s_{t+1} be generated by Algorithm 1. If $m_t \leq m$, then*

$$(1 + \hat{\delta}_{t+1}^{\text{ACC}})V_{t+1}^{\text{ACC}}(s_{t+1}) - V_{t+1}^{\text{ACC}}(s_t) \leq 0. \tag{69}$$

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

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

Since y_{t+1} produced by Algorithm 1 satisfies (15), we have that

$$(1 + \hat{\delta}_{t+1}^{\text{ACC}})(f(y_{t+1}) - f(x_t)) \leq -(1 + \hat{\delta}_{t+1}^{\text{ACC}})(1/2L_t)\|g_t\|^2. \tag{70}$$

Moreover, if $m_t \leq m$, then (3) implies that for all x and y

$$f(x) + \langle \nabla f(x), y - x \rangle + (m_t/2)\|x - y\|^2 \leq f(y). \tag{71}$$

Hence, plugging $x = x_t$ and $y = x^\star$ in (71) and using the fact that f is convex, we obtain

$$\hat{\delta}_{t+1}^{\text{ACC}}(f(x_t) - f(x^\star)) \leq \hat{\delta}_{t+1}^{\text{ACC}} \langle g_t, x_t^\star \rangle - \hat{\delta}_{t+1}^{\text{ACC}}(m_t/2)\|x_t^\star\|^2, \tag{72}$$

$$f(x_t) - f(y_t) \leq \langle g_t, x_t^y \rangle. \tag{73}$$

Next, we address the 2-norm difference in $(1+\hat{\delta}_{t+1}^{\mathrm{ACC}})V_{t+1}^{\mathrm{ACC}}(s_{t+1}) - V_{t+1}^{\mathrm{ACC}}(s_t)$ by expanding the pseudo-states inside 2-norms. One pseudo-state is $w_{t+1}^\star$ which, using (16) and (68), we express as

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

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:

$$\begin{aligned}
\|g_t\|^2 :& \qquad (q_t/L_t^2)(m_t/2) = 1/2L_t, \\
\langle g_t, x_t^y\rangle :& \qquad m_t(1+\hat{\delta}_{t+1}^{\mathrm{ACC}})\sqrt{q_t}(\sqrt{q_t}-1)/L_t = 1, \\
\langle g_t, x_t^\star\rangle :& \qquad m_t(1+\hat{\delta}_{t+1}^{\mathrm{ACC}})\sqrt{q_t}/L_t = \delta_t^{\mathrm{ACC}}, \\
\|x_t^y\|^2 :& \qquad (1+\hat{\delta}_{t+1}^{\mathrm{ACC}})(\sqrt{q_t}-1)^2 = \sqrt{q_t}(\sqrt{q_t}-1), \\
\langle x_t^y, x_t^\star\rangle :& \qquad (1+\hat{\delta}_{t+1}^{\mathrm{ACC}})(\sqrt{q_t}-1) = \sqrt{q_t}.
\end{aligned}$$

Thus, the 2-norm difference in $(1+\hat{\delta}_{t+1}^{\mathrm{ACC}})V_{t+1}^{\mathrm{ACC}}(s_{t+1}) - V_{t+1}^{\mathrm{ACC}}(s_t)$ reduces to

$$\begin{aligned}
&(1+\hat{\delta}_{t+1}^{\mathrm{ACC}})\frac{m_t}{2}\|w_{t+1}^\star\|^2 - \frac{m_t}{2}\|x_t^\star + \sqrt{q_t}x_t^y\|^2 \\
=&\frac{1+\hat{\delta}_{t+1}^{\mathrm{ACC}}}{2L_t}\|g_t\|^2 - \langle g_t, x_t^y\rangle - \delta_t^{\mathrm{ACC}}\langle g_t, x_t^\star\rangle + \frac{m_t}{2}\sqrt{q_t}(\sqrt{q_t}-1)\|x_t^y\|^2 \\
&+ \frac{m_t}{2}(2\sqrt{q_t}\langle x_t^y, x_t^\star\rangle + (1+\hat{\delta}_{t+1}^{\mathrm{ACC}})\|x_t^\star\|^2) - \frac{m_t}{2}(q_t\|x_t^y\|^2 + 2\sqrt{q_t}\langle x_t^y, x_t^\star\rangle + \|x_t^\star\|^2) \\
=&\frac{1+\hat{\delta}_{t+1}^{\mathrm{ACC}}}{2L_t}\|g_t\|^2 - \langle g_t, x_t^y\rangle - \delta_t^{\mathrm{ACC}}\langle g_t, x_t^\star\rangle - \frac{m_t}{2}\sqrt{q_t}\|x_t^y\|^2 + \delta_t^{\mathrm{ACC}}\frac{m_t}{2}\|x_t^\star\|^2. \qquad (74)
\end{aligned}$$

Finally, combining (70) and (72) to (74), cancelling terms and then using the assumption that $m_t > 0$, we obtain

$$(1+\hat{\delta}_{t+1}^{\mathrm{ACC}})V_{t+1}^{\mathrm{ACC}}(s_{t+1}) - V_{t+1}^{\mathrm{ACC}}(s_t) \le -(m_t/2)\sqrt{q_t}\|x_t^y\|^2 \le 0.$$

■

Lemma 21 *Let $f \in \mathcal{F}(L,m)$. If $L_t \ge L_{t-1} \ge m_{t-1} \ge m_t$ and $m_t \le m$, then*

$$V_{t+1}^{\mathrm{ACC}} \le \frac{q_t^2}{q_{t-1}^2}V_t^{\mathrm{ACC}}. \qquad (75)$$

Proof If $m_t \le m_{t-1}$, then

$$\begin{aligned}
V_{t+1}^{\mathrm{ACC}}(s_t) - V_t^{\mathrm{ACC}}(s_t) &= \frac{m_t}{2}\|x_t^\star + \sqrt{q_t}x_t^y\|^2 - \frac{m_{t-1}}{2}\|w_t^\star\|^2 \\
&\le \frac{m_t}{2}(\|x_t^\star + \sqrt{q_t}x_t^y\|^2 - \|w_t^\star\|^2). \qquad (76)
\end{aligned}$$

Hence, to prove (75), we express bounds on (76) in terms of V_{t+1}^{ACC} and V_t^{ACC}. To this end, we first note that the term in parenthesis on the right-hand side of (76) can be expressed as

$$\|x_t^\star + \sqrt{q_t} x_t^y\|^2 - \|w_t^\star\|^2 = 2(\sqrt{q_t} - \sqrt{q_{t-1}})\langle x_t^\star, x_t^y\rangle + (q_t - q_{t-1})\|x_t^y\|^2. \tag{77}$$

We consider two cases, each representing a possible sign of $\langle x_t^y, x_t^\star\rangle$.

First, suppose $\langle x_t^y, x_t^\star\rangle \geq 0$. If $L_t \geq L_{t-1}$, then $\sqrt{q_{t-1}}/\sqrt{q_t} \leq 1$, so that

$$\sqrt{q_t} - \sqrt{q_{t-1}} \leq q_t/\sqrt{q_t} - \sqrt{q_{t-1}}(\sqrt{q_{t-1}}/\sqrt{q_t}) = (q_t - q_{t-1})/\sqrt{q_t}. \tag{78}$$

Plugging (84) into (77) and then adding a nonnegative $\|x_t^\star\|^2$ term, we get

$$\begin{aligned}\|x_t^\star + \sqrt{q_t} x_t^y\|^2 - \|w_t^\star\|^2 &\leq 2\frac{q_t - q_{t-1}}{\sqrt{q_t}}\langle x_t^\star, x_t^y\rangle + (q_t - q_{t-1})\|x_t^y\|^2 + \frac{q_t - q_{t-1}}{q_t}\|x_t^\star\|^2 \\ &= \frac{q_t - q_{t-1}}{q_t}\|x_t^\star + \sqrt{q_t} x_t^y\|^2. \end{aligned}\tag{79}$$

Then, plugging (79) back into (76) yields

$$V_{t+1}^{\mathrm{ACC}}(s_t) - V_t^{\mathrm{ACC}}(s_t) \leq \frac{q_t - q_{t-1}}{q_t}\frac{m_t}{2}\|x_t^\star + \sqrt{q_t} x_t^y\|^2 \leq \frac{q_t - q_{t-1}}{q_t} V_{t+1}^{\mathrm{ACC}}(s_t), \tag{80}$$

where the last inequality follows from the definition of V_t^{ACC}, (67), as $\tilde{f} \geq 0$ implies

$$V_{t+1}^{\mathrm{ACC}}(s_t) = \tilde{f}(y_t) + \frac{m_t}{2}\|x_t^\star + \sqrt{q_t} x_t^y\|^2 \geq \frac{m_t}{2}\|x_t^\star + \sqrt{q_t} x_t^y\|^2. \tag{81}$$

Thus, rearranging terms in (80) and then multiplying both sides by q_t/q_{t-1}, we obtain

$$V_{t+1}^{\mathrm{ACC}}(s_t) \leq \frac{q_t}{q_{t-1}} V_t^{\mathrm{ACC}}(s_t) \leq \frac{q_t^2}{q_{t-1}^2} V_t^{\mathrm{ACC}}(s_t),$$

where the second inequality holds because $q_t/q_{t-1} \geq 1$.

Now, suppose $\langle x_t^y, x_t^\star\rangle < 0$. As in the previous case, we start by bounding the gap (77). 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

$$\|y_t^\star\|^2 = \|y_t^\star \mp x_t^\star\|^2 = \|x_t^\star - x_t^y\|^2 = \|x_t^\star\|^2 - 2\langle x_t^\star, x_t^y\rangle + \|x_t^y\|^2 \geq \|x_t^\star\|^2. \tag{82}$$

To use the above inequality on (77), first we rewrite (77) as

$$\begin{aligned}\|x_t^\star + \sqrt{q_t} x_t^y\|^2 - \|w_t^\star\|^2 &= 2\frac{\sqrt{q_t} - \sqrt{q_{t-1}}}{\sqrt{q_t}}\langle x_t^\star, \sqrt{q_t} x_t^y\rangle + \sqrt{q_t}(\sqrt{q_t} - \sqrt{q_{t-1}})\|x_t^y\|^2 \\ &\quad + \sqrt{q_{t-1}}(\sqrt{q_t} - \sqrt{q_{t-1}})\|x_t^y\|^2 \pm \frac{\sqrt{q_t} - \sqrt{q_{t-1}}}{\sqrt{q_t}}\|x_t^\star\|^2 \\ &= \frac{\sqrt{q_t} - \sqrt{q_{t-1}}}{\sqrt{q_t}}\|x_t^\star + \sqrt{q_t} x_t^y\|^2 + \sqrt{q_{t-1}}(\sqrt{q_t} - \sqrt{q_{t-1}})\|x_t^y\|^2 \\ &\quad - \frac{\sqrt{q_t} - \sqrt{q_{t-1}}}{\sqrt{q_t}}\|x_t^\star\|^2. \end{aligned}\tag{83}$$

Then, we apply (36) with $a = w_t^\star$, $b = x_t^\star$ and $c^2 = \sqrt{q_{t-1}}/\sqrt{q_t}$ to bound the $\|x_t^y\|^2$ term on (83), as in

$$\begin{aligned}\sqrt{q_{t-1}}(\sqrt{q_t} - \sqrt{q_{t-1}})\|x_t^y\|^2 &= \sqrt{q_{t-1}}(\sqrt{q_t} - \sqrt{q_{t-1}})\|x_t^y \pm x_t^\star/\sqrt{q_{t-1}}\|^2 \\ &= \frac{\sqrt{q_t} - \sqrt{q_{t-1}}}{\sqrt{q_{t-1}}}\|w_t^\star - x_t^\star\|^2 \\ &\leq \frac{\sqrt{q_t} - \sqrt{q_{t-1}}}{\sqrt{q_{t-1}}}\Big(1 + \frac{\sqrt{q_t}}{\sqrt{q_{t-1}}}\Big)\|w_t^\star\|^2 \\ &\quad + \frac{\sqrt{q_t} - \sqrt{q_{t-1}}}{\sqrt{q_{t-1}}}\Big(1 + \frac{\sqrt{q_{t-1}}}{\sqrt{q_t}}\Big)\|x_t^\star\|^2 \\ &= \frac{q_t - q_{t-1}}{q_{t-1}}\|w_t^\star\|^2 + \frac{\sqrt{q_t} - \sqrt{q_{t-1}}}{\sqrt{q_t}}\frac{\sqrt{q_t} + \sqrt{q_{t-1}}}{\sqrt{q_{t-1}}}\|x_t^\star\|^2. \end{aligned} \tag{84}$$

Plugging (84) back into (83) and then using (82), we get

$$\begin{aligned}\|x_t^\star + \sqrt{q_t}x_t^y\|^2 - \|w_t^\star\|^2 &\leq \frac{\sqrt{q_t} - \sqrt{q_{t-1}}}{\sqrt{q_t}}\|x_t^\star + \sqrt{q_t}x_t^y\|^2 + \frac{q_t - q_{t-1}}{q_{t-1}}\|w_t^\star\|^2 \\ &\quad + \frac{\sqrt{q_t} - \sqrt{q_{t-1}}}{\sqrt{q_t}}\Big(\frac{\sqrt{q_t} + \sqrt{q_{t-1}}}{\sqrt{q_{t-1}}} - 1\Big)\|x_t^\star\|^2 \\ &\leq \frac{\sqrt{q_t} - \sqrt{q_{t-1}}}{\sqrt{q_t}}\|x_t^\star + \sqrt{q_t}x_t^y\|^2 + \frac{q_t - q_{t-1}}{q_{t-1}}\|w_t^\star\|^2 \\ &\quad + \frac{\sqrt{q_t} - \sqrt{q_{t-1}}}{\sqrt{q_{t-1}}}\|y_t^\star\|^2. \end{aligned} \tag{85}$$

In turn, since $m_{t-1} \geq m_t$ and $m_t \leq m$, plugging (85) back into (76) we obtain

$$\begin{aligned}V_{t+1}^{\text{ACC}}(s_t) - V_t^{\text{ACC}}(s_t) &\leq \frac{m_t}{2}\frac{\sqrt{q_t} - \sqrt{q_{t-1}}}{\sqrt{q_t}}\|x_t^\star + \sqrt{q_t}x_t^y\|^2 + \frac{m_t}{2}\frac{q_t - q_{t-1}}{q_{t-1}}\|w_t^\star\|^2 \\ &\quad + \frac{m_t}{2}\frac{\sqrt{q_t} - \sqrt{q_{t-1}}}{\sqrt{q_{t-1}}}\|y_t^\star\|^2 \\ &\leq \frac{\sqrt{q_t} - \sqrt{q_{t-1}}}{\sqrt{q_t}}\frac{m_t}{2}\|x_t^\star + \sqrt{q_t}x_t^y\|^2 + \frac{q_t - q_{t-1}}{q_{t-1}}\frac{m_{t-1}}{2}\|w_t^\star\|^2 \\ &\quad + \frac{\sqrt{q_t} - \sqrt{q_{t-1}}}{\sqrt{q_{t-1}}}\frac{m}{2}\|y_t^\star\|^2. \end{aligned} \tag{86}$$

Now, as in (81), applying $\tilde{f} \geq 0$ to the definition of V_t^{ACC}, we get

$$V_t^{\text{ACC}}(s_t) = \tilde{f}(y_t) + \frac{m_{t-1}}{2}\|w_t^\star\|^2 \geq \frac{m_{t-1}}{2}\|w_t^\star\|^2. \tag{87}$$

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

$$V_t^{\text{ACC}}(s_t) = \tilde{f}(y_t) + \frac{m_{t-1}}{2}\|w_t^\star\|^2 \geq \frac{m}{2}\|y_t^\star\|^2. \tag{88}$$

Plugging in (81), (87) and (88) back into (86), and then moving all $V_{t+1}^{acc}(s_t)$ terms to the left-hand side and all $V_t^{\mathrm{ACC}}(s_t)$ to the right-hand side, we obtain

$$\frac{\sqrt{q_{t-1}}}{\sqrt{q_t}} V_{t+1}^{\mathrm{ACC}}(s_t) \leq \Big(\frac{q_t}{q_{t-1}} + \frac{\sqrt{q_t} - \sqrt{q_{t-1}}}{\sqrt{q_{t-1}}}\Big) V_t^{\mathrm{ACC}}(s_t) \tag{89}$$

Multiplying both sides of (89) by $\sqrt{q_t}/\sqrt{q_{t-1}}$, and then using the fact that $\sqrt{q_t} \geq \sqrt{q_{t-1}}$ yields

$$V_{t+1}^{\mathrm{ACC}}(s_t) \leq \frac{\sqrt{q}_t}{\sqrt{q_{t-1}}} \Big(\frac{q_t}{q_{t-1}} + \frac{\sqrt{q_t} - \sqrt{q_{t-1}}}{\sqrt{q_{t-1}}}\Big) V_t^{\mathrm{ACC}}(s_t) \leq \frac{q_t^2}{q_{t-1}^2} V_t^{\mathrm{ACC}}(s_t),$$

where the last inequality above is a consequence of the following equivalences:

$$\begin{aligned} \frac{q_t}{q_{t-1}} + \frac{\sqrt{q_t} - \sqrt{q_{t-1}}}{\sqrt{q_{t-1}}} \leq \frac{q_t^{3/2}}{q_{t-1}^{3/2}} &\iff \sqrt{q_{t-1}} q_t + q_{t-1}(\sqrt{q_t} - \sqrt{q_{t-1}}) \leq q_t^{3/2}, \\ &\iff q_{t-1}(\sqrt{q_t} - \sqrt{q_{t-1}}) \leq q_t(\sqrt{q_t} - \sqrt{q_{t-1}}), \end{aligned}$$

which hold since $q_t \geq q_{t-1}$. Therefore, whether $\langle x_t^\star, x_t^y \rangle \geq 0$ or $\langle x_t^\star, x_t^y \rangle < 0$, we have that

$$V_{t+1}^{\mathrm{ACC}}(s_t) \leq \frac{q_t^2}{q_{t-1}^2} V_t^{\mathrm{ACC}}(s_t)$$

for all s_t, proving (75). ■

Theorem 22 *Let $f \in \mathcal{F}(L, m)$ and let s_{t+1} be generated by Algorithm 1. If $m_N \leq m$, then for all $t \geq N$ we have that*

$$f(y_{t+1}) - f(x^\star) \leq \frac{q_t^2}{q_N^2} \prod_{i=N}^{t+1} (1 + \hat{\delta}_i^{\mathrm{ACC}})^{-1} V_N^{\mathrm{ACC}}(s_N). \tag{90}$$

Proof Since m_t is nonincreasing, if $m_N \leq m$, then $m_t \leq m$ for all $t \geq N$. Therefore, if $m_N \leq m$, then Lemmas 20 and 21 hold for all $t \geq N$. So, plugging (75) into (69) and proceeding with a simple inductive argument, it follows that

$$f(y_{t+1}) - f(x^\star) \leq V_{t+1}^{\mathrm{ACC}}(s_{t+1}) \leq \frac{q_t^2}{q_N^2} \prod_{i=N}^{t+1} (1 + \hat{\delta}_i^{\mathrm{ACC}})^{-1} V_N^{\mathrm{ACC}}(s_N),$$

where the first inequality follows directly from (67), the definition of V_t^{ACC}, since

$$V_{t+1}^{\mathrm{ACC}}(s_{t+1}) = \tilde{f}(y_{t+1}) + (m_t/2)\|w_{t+1}^\star\|^2 \geq \tilde{f}(y_{t+1}).$$

■

The transition iteration from $m_t \geq m$ to $m_t < m$

We now have analyzed iterations of Algorithm 1 in which $m_t \geq m$ and iterations in which $m_t < m$. To prove Theorem 3, it remains to join the two analyses, considering the transition from the first kind of iteration to the second kind. Since m_t is nonincreasing, there can be at most one such transition. We start by bounding V_t^{ACC} in terms of V_t^{GD}.

Lemma 23 *If $f \in \mathcal{F}(L, m)$ and $L_{t-1} \geq m_{t-1} \geq m > 0$, then*

$$V_t^{\text{ACC}} \leq (m_{t-1}/m) V_t^{\text{GD}}. \tag{91}$$

Proof To prove (91), we split the analysis according to the sign of $\langle x_t^\star, x_t^y \rangle$ in

$$\begin{aligned} V_t^{\text{ACC}}(s_t) - V_t^{\text{GD}}(s_t) &= \frac{m_{t-1}}{2} \|x_t^\star + \sqrt{q_{t-1}} x_t^y\|^2 - \frac{m}{2} \|x_t^\star + \sqrt{p_{t-1}} x_t^y\|^2 \\ &= \frac{m_{t-1} - m}{2} \|x_t^\star\|^2 + 2 \frac{\sqrt{L_{t-1}}(\sqrt{m_{t-1}} - \sqrt{m})}{2} \langle x_t^\star, x_t^y \rangle \end{aligned} \tag{92}$$

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

Multiplying the $\langle x_t^\star, x_t^y \rangle$ coefficient on (92) by $(\sqrt{m_{t-1}} + \sqrt{m})/\sqrt{m_{t-1}} \geq 1$, we obtain

$$\sqrt{L_{t-1}}(\sqrt{m_{t-1}} - \sqrt{m}) \leq \sqrt{q_{t-1}}(m_{t-1} - m). \tag{93}$$

Hence, if $\langle x_t^\star, x_t^y \rangle \geq 0$, then plugging (93) into (92), adding a nonnegative $\|x_t^y\|^2$ term, completing a square to form a $\|w_t^\star\|^2$ term and then applying (87), we get

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

Moving terms around and then multiplying both sides by $m_{t-1}/m > 0$, we get

$$V_t^{\text{ACC}}(s_t) \leq (m_{t-1}/m) V_t^{\text{GD}}(s_t).$$

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:

$$\frac{m_{t-1} - m}{m} = \frac{\sqrt{m_{t-1}} - \sqrt{m}}{\sqrt{m}} \frac{\sqrt{m_{t-1}} + \sqrt{m}}{\sqrt{m}} = \frac{\sqrt{m_{t-1}} - \sqrt{m}}{\sqrt{m}} \Big(1 + \frac{\sqrt{m_{t-1}}}{\sqrt{m}}\Big). \tag{94}$$

To handle the $\|x_t^\star\|^2$ term that stays out of the square, we use the fact that

$$\|y_t^\star\|^2 = \|y_t^\star \pm x_t^\star\|^2 = \|x_t^\star - x_t^y\|^2 = \|x_t^\star\|^2 - 2\langle x_t^\star, x_t^y \rangle + \|x_t^y\|^2 \geq \|x_t^\star\|^2, \tag{95}$$

which follows since $\langle x_t^\star, x_t^y \rangle < 0$. By the definition of $\bar{\alpha}_{t-1}$, the assumption that $m_{t-1} \geq m$ yields $\bar{\alpha}_{t-1} \geq 0$, thus $\bar{\alpha}_{t-1}/\sqrt{p_{t-1}} \geq 0$. Moreover, $U_t \geq 0$ because of the assumption that $L_{t-1} > 0$. Since $\tilde{f}$ is also nonnegative, from (49) we obtain

$$V_t^{\text{GD}}(s_t) \geq W_t(s_t) = \tilde{f}(y_t) + (m/2)\|z_t^\star\|^2 \geq (m/2) \max(\|z_t^\star\|^2, \|y_t^\star\|^2), \tag{96}$$

where the right-hand side follows from applying (3) with $x = x^\star$ and $y = y_t$.

Hence, splitting the coefficient of $\|x_t^\star\|^2$ on (92) according to (94), adding a positive $\|x_t^y\|^2$ term to form a $\|z_t^\star\|^2$ term, applying (95) and then using (96), we obtain

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

Finally, moving terms around, we get

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

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

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

holds generically for all s_t, and we recover (91). ■

Now, we are ready to prove Theorem 3, which combines Theorems 19 and 22 into a single result that holds for all iterations. First, we note that by design $m_0 \in [m, L]$. Hence, since $c_t \in [m, L]$, it follows that $m_t \geq m/\gamma$ for all $t \geq 0$. If $L_0 \geq L$, then since (2) holds for all $\bar{L} \geq L$, it follows that $L_t = L_0$ for all t. Otherwise, if $L_0 \in [m, L]$, then since L_t is adjusted by a factor of 2 every time (2) is violated, and (2) holds for all $\bar{L} \geq L$, it follows that $L_t \leq 2L$ for all t. With that in mind, let $\bar{L} = \max(L_0, 2L)$ and $\bar{\kappa} = \bar{L}/m$. Then, we have that

$$p_t = (L_t/m) \leq \bar{\kappa} \qquad \text{and} \qquad q_t = (L_t/m_t) \leq \gamma\bar{\kappa}. \tag{97}$$

Plugging (97) into the definitions of δ_t^{GD} and $\hat{\delta}_t^{\mathrm{ACC}}$, we obtain

$$\delta_{t+1}^{\mathrm{GD}} = 1/(p_t - 1) \geq 1/(\bar{\kappa} - 1) = \delta(\bar{\kappa}), \tag{98}$$

$$\hat{\delta}_{t+1}^{\mathrm{ACC}} = 1/(\sqrt{q_t} - 1) \geq 1/(\sqrt{\ } - 1) = \delta(\sqrt{\gamma\bar{\kappa}}). \tag{99}$$

Moreover, by design $m_0 \geq L_0$ and m_t is nonincreasing, therefore $m_t \leq L_0$ for all t, and it follows from (97) that

$$\frac{q_t^2}{q_j^2}\frac{p_j^2}{p_0^2} = q_t^2 \frac{m_j^2}{L_j^2}\frac{L_j^2}{L_0^2} = q_t^2 \frac{m_j^2}{L_0^2} \leq q_t^2 \leq \gamma^2\bar{\kappa}^2. \tag{100}$$

Proof [Proof of Theorem 3] By (98), we have that

$$(1 + \delta_{t+1}^{\mathrm{GD}})^{-1} \leq (1 + \delta(\bar{\kappa}))^{-1} = (\bar{\kappa} - 1)/\bar{\kappa}$$

for all t such that $m_t \geq m$. Hence, by Theorem 19 and (97), it follows that

$$f(y_t) - f(x^\star) \leq \left(1 - \frac{1}{\bar{\kappa}}\right)^t 2\max(L_0, L)\bar{\kappa}^2\|x_0^\star\|^2$$

for all t such that $m_t \geq m$. If $m_t \geq m$ for all t, then (6) holds for all t. Otherwise, let N be the first iteration for which $m_{N+1} \leq m$. By simple manipulations, we have that

$$(\sqrt{\gamma\bar{\kappa}} - 1)/\sqrt{\gamma\bar{\kappa}} = (1 + \delta(\sqrt{\gamma\bar{\kappa}}))^{-1} \leq (1 + \delta(\bar{\kappa}))^{1} = (\bar{\kappa} - 1)/\bar{\kappa}$$

if and only if $\gamma \leq \bar{\kappa}$. Hence, since by assumption $\bar{\kappa} \geq \kappa \geq 2 = \gamma$, from (97) it follows that

$$(1 + \delta_{t+1}^{\text{ACC}})^{-1} \leq (1 + \delta(\sqrt{\gamma\bar{\kappa}}))^{-1} \leq (1 + \delta(\bar{\kappa}))^{-1}$$

for all $t \geq N$. Combining Lemma 23 with Theorem 22, then applying Theorem 19 and plugging in (100) with $\gamma = 2$ and $m_N \leq L$ into the result, it follows that for all $t \geq N$,

$$\begin{aligned} f(y_{t+1}) - f(x^\star) &\leq \frac{q_t^2}{q_N^2}\left(1 - \frac{1}{\bar{\kappa}}\right)^{t-N} \frac{m_N}{m} 2\max(L_0, L)\frac{p_N^2}{p_0^2}\left(1 - \frac{1}{\bar{\kappa}}\right)^N \|x_0^\star\|^2 \\ &\leq \left(1 - \frac{1}{\bar{\kappa}}\right)^t 8\max(L_0, L)\bar{\kappa}^3\|x_0^\star\|^2. \end{aligned}$$

Therefore, (6) holds for all $t \geq 0$. ■

Appendix B. Local Acceleration

In this section, we prove Theorem 5, establishing that NAG-free (Algorithm 1) converges at an accelerated rate to $x^\star$, the minimum of $f \in \mathcal{F}(L, m)$, when its iterates get sufficiently close to $x^\star$.

$$r_{\text{NAG}}(z) = \frac{\sqrt{z} - 1}{\sqrt{z}}. \tag{101}$$

Remark 24 (Deriving most results assuming $L_0 = L$.) *To prove Theorem 5, we assume that some $L_0 > L$ is known and can be used to initialize Algorithm 1. If $L_0 > L$, then L_0 also satisfies (2). In turn, if Algorithm 1 is initialized with L_0, then the descent lemma condition $f(y_{t+1}) - f(x_t) \leq -(1/2L_t)\|\nabla f(x_t)\|^2$ is always satisfied, therefore $L_t \equiv L_0$ for all t. However, we derive most of the results in this section using L to avoid working with the cluttered notation L_0, since essentially all of the results below hold for any $L_0 \geq L$. Once we prove the local acceleration results, we plug $L_0 > L$ back in.*

To prove Theorem 5, we first consider the simplified case where f is quadratic, and then analyze the general case as a perturbation of the quadratic case. To this end, we use the fact established by Theorem 3 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

$$f(y_t) - f(x^\star) \leq r_{\text{GD}}(\bar{\kappa})^t 8 \max(L_0, L)\bar{\kappa}^3 \|x_0^\star\|^2,$$

where $\bar{\kappa} = \max(L_0, 2L)/m$ and r_{GD} is defined over $[1, +\infty)$ as

$$r_{\text{GD}}(z) = \frac{z - 1}{z}. \tag{102}$$

B.1. Quadratic Case

First, we assume the objective function is 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 $(1/2)x^\mathsf{T} Hx + x^\mathsf{T} g + f(0)$ can be expressed[1] 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

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

Since $\nabla^2 f$ is locally smooth 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 [9] 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

$$x_t - x^\star = \sum_{i=1}^d x_{i,t} v_i, \tag{103}$$

$$y_{t+1} - x^\star = x_t - \frac{1}{L}\nabla f(x_t) - x^\star = \sum_{i=1}^d \Big(1 - \frac{\lambda_i}{L}\Big) x_{i,t} v_i. \tag{104}$$

1. 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 $f(0) = -\frac{1}{2}x^{\star,\mathsf{T}} Hx^\star$. Then, plugging $f(0)$ back into $f(x)$ and replacing the inner-product $g^\mathsf{T} x$ with $g^\mathsf{T} x = -x^{\star,\mathsf{T}} Hx = -\frac{1}{2}x^{\star,\mathsf{T}} Hx - \frac{1}{2}x^\mathsf{T} Hx^\star$ yields the desired form of $f(x)$.

Substituting (104) for the descent steps yields

$$\begin{aligned}\sum_{i=1}^{d} x_{i,t+1}v_i &= x_{t+1} - x^\star \\ &= (1+\beta_t)y_{t+1} - \beta_t y_t - x^\star \mp \beta_t x^\star \\ &= (1+\beta_t)(y_{t+1} - x^\star) - \beta_t(y_t - x^\star) \\ &= \sum_{i=1}^{d} \left[(1+\beta_t)\Big(1 - \frac{\lambda_i}{L}\Big)x_{i,t} - \beta_t\Big(1 - \frac{\lambda_i}{L}\Big)x_{i,t-1}\right] v_i, \end{aligned} \tag{105}$$

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

$$\beta(m_t) = \frac{\sqrt{L} - \sqrt{m_t}}{\sqrt{L} + \sqrt{m_t}}. \tag{106}$$

That is, each component $x_{i,t}$ of $x_t - x^\star$ behaves as an LTV system [12]. But if $\gamma > 1$, then by design m_t decreases by a factor of at least γ 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

$$X_{i,t+1} = G_i(m_t)X_{i,t}, \tag{107}$$

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

$$X_{i,t} = \begin{cases} \begin{bmatrix} x_{i,0} & x_{i,0} \end{bmatrix}^\mathsf{T}, & t = 0, \\ \begin{bmatrix} x_{i,t-1} & x_{i,t} \end{bmatrix}^\mathsf{T}, & t > 0, \end{cases} \tag{108}$$

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

$$G_i(m_t) = \begin{bmatrix} 0 & 1 \\ -\beta(m_t)\Big(1 - \frac{\lambda_i}{L}\Big) & (1+\beta(m_t))\Big(1 - \frac{\lambda_i}{L}\Big) \end{bmatrix}. \tag{109}$$

Hence, the dynamics of (107) is determined by the eigenvalues of $G_i(m_t)$, which are given by

$$\lambda(G_i(m_t)) = \frac{1+\beta(m_t)}{2}\Big(1 - \frac{\lambda_i}{L}\Big) \pm \sqrt{\frac{(1+\beta(m_t))^2}{4}\Big(1 - \frac{\lambda_i}{L}\Big)^2 - \beta(m_t)\Big(1 - \frac{\lambda_i}{L}\Big)}. \tag{110}$$

The greatest between the two eigenvalues given by (110) defines the so-called spectral radius [11] of G_i, captured by the function $\rho : (0, L] \times \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ defined by

$$\rho(s, \ell) = \max \left| \frac{1+\beta(m_t)}{2}\Big(1 - \frac{\ell}{L}\Big) \pm \sqrt{\frac{(1+\beta(m_t))^2}{4}\Big(1 - \frac{\ell}{L}\Big)^2 - \beta(m_t)\Big(1 - \frac{\ell}{L}\Big)} \right|. \tag{111}$$

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:

$$\varrho(s, \ell) = \min \left| \frac{1+\beta(m_t)}{2}\Big(1 - \frac{\ell}{L}\Big) \pm \sqrt{\frac{(1+\beta(m_t))^2}{4}\Big(1 - \frac{\ell}{L}\Big)^2 - \beta(m_t)\Big(1 - \frac{\ell}{L}\Big)} \right|. \tag{112}$$

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. Next, we derive several auxiliary results on ρ and ϱ that will be useful later on.

Properties of the Spectral Radius ρ

Lemma 25 *Let $s, \ell \in (0, L]$. The two numbers*

$$\frac{1+\beta(s)}{2}\Big(1-\frac{\ell}{L}\Big) \pm \sqrt{\frac{(1+\beta(s))^2}{4}\Big(1-\frac{\ell}{L}\Big)^2 - \beta(s)\Big(1-\frac{\ell}{L}\Big)} \tag{113}$$

have nonzero imaginary part if and only if $s < \ell < L$. If (113) have zero imaginary part, then

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

otherwise, if (113) have nonzero imaginary part, then

$$\rho(s,\ell) = \sqrt{\beta(s)\Big(1-\frac{\ell}{L}\Big)}. \tag{115}$$

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

$$\Delta(s,\ell) = \frac{(1+\beta(s))^2}{4}\Big(1-\frac{\ell}{L}\Big)^2 - \beta(s)\Big(1-\frac{\ell}{L}\Big). \tag{116}$$

If $\ell = 0$, then $\ell \leq s$ since $s \geq 0$, because $s \in (0, L]$. Moreover, plugging $\ell = 0$ into (116), 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, $\Delta(s,\ell) \geq 0$ because $(1-\beta)^2 \geq 0$. Furthermore, $1 - \ell/L = 1$ and $\rho(s,\ell)$ trivially reduces to form (114).

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 (114). Otherwise, if $\ell < L$, then $1 - \ell/L > 0$ and $\Delta < 0$ if and only if

$$(1+\beta)^2\Big(1-\frac{\ell}{L}\Big) - 4\beta < 0 \iff (1-\beta)^2 L < (1+\beta)^2 \ell \iff \frac{L}{\ell} < \Big(\frac{1+\beta}{1-\beta}\Big)^2, \tag{117}$$

where L/ℓ is well-defined since $\ell > 0$, by assumption, while $(1-\beta)^{-1}$ is well-defined because $0 \leq \beta(s) < 1$ for all $s \in (0, L]$. Plugging (106) into β, the squared factor on the right-hand side of (117) turns into

$$\frac{1+\beta}{1-\beta} = \frac{2\sqrt{L}/(\sqrt{L}+\sqrt{s})}{2\sqrt{s}/(\sqrt{L}+\sqrt{s})} = \sqrt{L/s}. \tag{118}$$

Thus, by (117), $\Delta(s,\ell)$ is negative if and only if $s < \ell$. Hence, if $s \geq \ell$, then $\Delta \geq 0$, which combined with the the assumption that $L > \ell$ implies

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

so that

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

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 (114).

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 (115). ■

Corollary 26 *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 (113), we recover the two eigenvalues of $G_i(m_t)$ which, by Theorem 25, 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 (116) is zero for $\ell = \lambda_i$ and $s = m_t$. In turn, by (117) and (118), the discriminant (116) 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 27 *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 Theorem 25 implies $\rho(s,a)$ and $\rho(s,b)$ both take form (114). If, in addition $s = L$, then $\beta = 0$, which when substituted back into (114) 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

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

where the last equivalence follows at once from (106). 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 (114) 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 $b \leq L$, so $a < b \leq L$ and it follows that

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

that is

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

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 (115) by Theorem 25 and the above inequality yields

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

■

Lemma 28 *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 Theorem 27, $\rho(s, \ell) \le \rho(s, m)$, so it suffices to show $\rho(s, m) \le r_{\mathrm{GD}}(\kappa)$. If $m < m_t$, then by Theorem 25, 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

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

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 Theorem 25 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 29 *If Assumption 4.3 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 $\delta_L = (L_0 - L)/L_0$.*

Proof If Assumptions 4.2 and 4.3 hold, then there exists some $\delta_\lambda > 0$ such that $|m_t - \lambda_i| \ge \delta_\lambda$. Moreover, since $L_0 > L$, we have that $\delta_L = (L_0 - L)/L_0$. Moreover, whether $\zeta_i = \zeta_i(m_t)$ and $\xi_i = \xi_i(m_t)$ are complex or real, we have that

$$|\zeta_i - \xi_i| = 2\left|\frac{(1+\beta)^2}{4}\frac{(L_0-\lambda_i)^2}{L_0^2} - \beta\frac{L_0-\lambda_i}{L_0}\right|^{1/2} = \frac{1+\beta}{L_0}|(L_0-\lambda_i)(m_t-\lambda_i)|^{1/2} \ge \sqrt{\delta_L\delta_\lambda},$$

where in the last equality we have replaced L with L_0 in the identity

$$\frac{4\beta L}{(1+\beta)^2} = 4\frac{\sqrt{L}-\sqrt{s}}{\sqrt{L}+\sqrt{s}}\frac{(\sqrt{L}+\sqrt{s})^2}{4L}L = L - s. \tag{119}$$

■

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 Theorem 27, 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 $\rho(m_t, m) \leq r_{\text{NAG}}(\sigma\kappa)$, where σ represents a suboptimality factor relative to the optimal convergence rate of $r_{\text{NAG}}(\kappa)$.

By Theorem 22, 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 $r_\delta(\kappa) \leq r_{\text{NAG}}(\sigma\kappa)$ 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

$$r_\delta(z) = \frac{1 + \beta_\delta(z)}{2}\frac{z-1}{z} + \sqrt{\frac{(1 + \beta_\delta(z))^2}{4}\Big(\frac{z-1}{z}\Big)^2 - \beta_\delta(z)\frac{z-1}{z}}, \tag{120}$$

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

$$\beta_\delta(z) = \frac{\sqrt{z} - \sqrt{1+\delta}}{\sqrt{z} + \sqrt{1+\delta}}. \tag{121}$$

Lemma 30 *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 Theorem 25

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

Omitting the arguments in $\beta = \beta(L, s)$ and using the identity (119), the discriminant above can be expressed as

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

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 $m \le s \le (1+\delta)m$ and $\kappa \ge 1+\delta$, we have that

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

■

Next, we bound r_δ in terms of r_{NAG}. We start with an identity involving $\beta_\delta(\kappa)$, analogous to (119):

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

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

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

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

$$r_\delta(\kappa) = \frac{\sqrt{\kappa}}{\sqrt{\kappa}+\sqrt{1+\delta}}\frac{\kappa-1}{\kappa} + \frac{\sqrt{\kappa-1}}{\sqrt{\kappa}+\sqrt{1+\delta}}\frac{\sqrt{\delta}}{\sqrt{\kappa}} = \frac{\sqrt{\kappa-1}}{\sqrt{\kappa}}\frac{\sqrt{\kappa-1}+\sqrt{\delta}}{\sqrt{\kappa}+\sqrt{1+\delta}}. \tag{122}$$

Using this alternative expression, we obtain the following.

Lemma 31 *Given $\sigma > 1$, then $r_\delta(\kappa) \le r_{\text{NAG}}(\sigma'\kappa)$ 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 $r_{\delta'}(\kappa) \le r_{\text{NAG}}(\sigma\kappa)$ 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 (122) and (101), it follows that the condition that $r_\delta(\kappa) \le r_{\text{NAG}}(\sigma\kappa)$ for some $\delta > 0$ and $\kappa \ge 1+\delta$ is equivalent to

$$\frac{\sqrt{\kappa-1}}{\sqrt{\kappa}}\frac{\sqrt{\kappa-1}+\sqrt{\delta}}{\sqrt{\kappa}+\sqrt{1+\delta}} \le \frac{\sqrt{\sigma\kappa}-1}{\sqrt{\sigma\kappa}}. \tag{123}$$

By successively manipulating (123), it follows that

$$\begin{aligned} r_\delta(\kappa) \le r_{\text{NAG}}(\sigma\kappa) &\iff \sqrt{\kappa-1}(\sqrt{\kappa-1}+\sqrt{\delta})\sqrt{\sigma} \le (\sqrt{\sigma\kappa}-1)(\sqrt{\kappa}+\sqrt{1+\delta})\\ &\iff \sqrt{\kappa}+\sqrt{1+\delta} \le (1+\sqrt{(1+\delta)\kappa}-\sqrt{\delta(\kappa-1)})\sqrt{\sigma}\\ &\iff \frac{\sqrt{\kappa}+\sqrt{1+\delta}}{1+\sqrt{(1+\delta)\kappa}-\sqrt{\delta(\kappa-1)}} \le \sqrt{\sigma}. \end{aligned} \tag{124}$$

Taking the derivative of the left-hand side of the (124) 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 (124) is increasing in $\kappa \geq 1+\delta$ and it follows that

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

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

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

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

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

Therefore, $r_{\delta'}(\kappa) \leq r_{\text{NAG}}(\sigma\kappa)$ for all $\delta' \leq \delta$, $\kappa \geq 1+\delta'$ and $\sigma \geq \sigma_\delta$. ■

Corollary 32 *Given $\sigma > 1$, then $\rho(s, m) \leq r_{\text{NAG}}(\sigma'\kappa)$ for all $s \in [m, (1+\delta)m]$, $\delta \in (0, \delta_\sigma]$, $\sigma' \geq \sigma$ and $\kappa \geq 1+\delta$, where $\delta_\sigma = (\sigma-1)^2/4\sigma$. Conversely, given $\delta > 0$, then $\rho(s, m) \leq r_{\text{NAG}}(\sigma\kappa)$ for all $s \in [m, (1+\delta)m]$, $\delta' \in (0, \delta]$, $\sigma \geq \sigma_\delta$ and $\kappa \geq 1+\delta'$, where $\sigma_\delta = 1+2\delta+2\sqrt{\delta(1+\delta)}$.*

Proof The corollary follows by combining Theorems 30 and 31. ■

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, if $\gamma > 1$, then 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 $t_j < t' \leq t_{j+1} \leq t_J < t \leq t_{J+1}$,

$$X_{i,t} = \prod_{k=0}^{t-1} G_i(\mu_k) X_{i,0} = G_i(\mu_J)^{t-t_J} \left(\prod_{k=j+1}^{J-1} G_i(\mu_k)^{t_{k+1}-t_k} \right) G_i(\mu_j)^{t_{j+1}-t'} X_{i,t'}. \tag{125}$$

Now, if $m_t > m$, then under Assumption 4.3, Corollary 26 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

$$T_i(m_t) = \begin{bmatrix} 1 & 1 \\ \zeta_i & \xi_i \end{bmatrix}. \tag{126}$$

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 $G_i(m_t)$:

$$G_i(m_t) = T_i(m_t) D_i(m_t) T_i(m_t)^{-1}. \tag{127}$$

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

$$D_i(m_t) = \begin{bmatrix} \zeta_i & 0 \\ 0 & \xi_i \end{bmatrix}. \tag{128}$$

Combining (125), (127) and (128), then applying Theorem 27 it follows that for every $t_j < t \leq t_{j+1}$

$$\|X_{i,t}\|^2 \leq \overline{C}_i \rho(\mu_j, \lambda_i)^{2(t-t_j)} \left(\prod_{k=0}^{j-1} \rho(\mu_k, \lambda_i)^{2(t_{k+1}-t_k)} \right) x_{i,0}^2, \tag{129}$$

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

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

and, by applying Theorems 28 and 29 to (126), 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 $\delta_L = (L_0 - L)/L_0$ and δ_λ is given by Assumption 4.3. 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:

$$X_{i,t} = A_{i,t_j}\zeta_i^{t-t_j} + B_{i,t_j}\xi^{t-t_j}, \tag{130}$$

where A_i and B_i are two-dimensional vectors given by

$$A_{i,t_j} = \frac{x_{i,t_j} - \xi_i x_{i,t_j-1}}{\zeta_i - \xi_i} \begin{bmatrix} 1 \\ \zeta_i \end{bmatrix} \qquad \text{and} \qquad B_{i,t_j} = \frac{\zeta_i x_{i,t_j-1} - x_{i,t_j}}{\zeta_i - \xi_i} \begin{bmatrix} 1 \\ \xi_i \end{bmatrix}, \tag{131}$$

which are well-defined, by Theorem 29. In particular, for $t_0 < t \le t_1$, 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

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

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

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

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

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

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

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

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

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

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

$$(x_{1,t+1} - x_{1,t})^2 \ge \underline{C}_1 \rho(\mu_j, m)^{2t-t_j} \left(\prod_{k=0}^{j-1} \rho(\mu_k, m)^{2(t_{k+1}-t_k)} \right) x_{1,0}^2 \ge 0, \tag{132}$$

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

The Dynamics of c_t

Theorem 31 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 (103) and (105) into (5), obtaining[2]

$$c_{t+1}^2 = \left\| \frac{\nabla f(x_{t+1}) - \nabla f(x_t)}{x_{t+1} - x_t} \right\|^2 = \left\| \frac{\sum_{i=1}^d (x_{i,t+1} - x_{i,t})\lambda_i v_i}{\sum_{i=1}^d (x_{i,t+1} - x_{i,t}) v_i} \right\|^2 = \frac{\sum_{i=1}^d (x_{i,t+1} - x_{i,t})^2 \lambda_i^2}{\sum_{i=1}^d (x_{i,t+1} - x_{i,t})^2}. \tag{133}$$

The identity (133) 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 Theorem 27, $\rho(s, \lambda_i) < \rho(s, m)$ for all $\lambda_i \in (m, L]$. Hence, from (129) and (132), 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[3] $\phi : \mathcal{D} \to [0, 1]$ as

$$\phi(s, a, b) = \begin{cases} \min\left\{1, \dfrac{\rho(s, a)}{\rho(s, b)}\right\}, & \rho(s, b) > 0, \\ 1, & \rho(s, b) = 0, \end{cases} \tag{134}$$

2. Note that $x_{t+1} - x_t = (x_{t+1} - x^\star) - (x_t - x^\star) = \sum_{i=1}^d (x_{i,t+1} - x_{i,t}) v_i$.
3. Note that $\rho < 0$ cannot occur by the definition of ρ, (111).

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

$$\mathcal{D} = (0, L] \times \left\{(a, b) \in \mathbb{R}^2_{>0} : a \neq b\right\}, \tag{135}$$

$\mathbb{R}_{>0}$ 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 33 *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 (134) and Theorem 25, 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:

$$\begin{aligned}
&(L-m)\sqrt{(L-\ell)\delta\ell} + \sqrt{(L-\ell)\delta\ell}\sqrt{(L-m)((1+\delta)\ell - m)} \\
&- \delta(L - 2\ell)\sqrt{(L-m)((1+\delta)\ell - m)} > 0.
\end{aligned} \tag{136}$$

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 (136) 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 (136) 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 Theorem 33 we have that for every $\ell \in [(1+\delta_\ell)m, L]$

$$
\begin{aligned}
\phi((1+\delta_u)\ell, \ell, m) =& \frac{L - \ell + \sqrt{(L-\ell)\delta_u\ell}}{L - m + \sqrt{(L-m)((1+\delta_u)\ell - m)}} \\
\leq& \frac{L - (1+\delta_\ell)m + \sqrt{(L-(1+\delta_\ell)m)\delta_u(1+\delta_\ell)m}}{L - m + \sqrt{(L-m)((1+\delta_u)(1+\delta_\ell)m - m)}} \\
=& \frac{\kappa - (1+\delta_\ell) + \sqrt{(\kappa-(1+\delta_\ell))\delta_u(1+\delta_\ell)}}{\kappa - 1 + \sqrt{(\kappa-1)(\delta_u + \delta_\ell + \delta_u\delta_\ell)}} \\
=:& r_\phi(\delta_u, \delta_\ell, \kappa).
\end{aligned}
\tag{137}
$$

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 34 *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*

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

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

$$r_\phi(\delta_u, \delta_\ell, \kappa)^2 \leq r_{\text{NAG}}(\sigma_\phi \kappa) = \frac{\sqrt{\sigma_\phi \kappa} - 1}{\sqrt{\sigma_\phi \kappa}} \iff \frac{1}{(1 - r_\phi(\delta_u, \delta_\ell, \kappa)^2)^2 \kappa} \leq \sigma_\phi. \tag{138}$$

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 (138) is well-defined. So, let σ_ϕ be defined such that (138) 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 2 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 $r_{\text{NAG}}(\sigma_\phi \kappa)$ for most values of ℓ.

Lemma 35 *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 $\ell \in [(1 + \delta_\ell)m, L/(1 + \delta_u)]$*

$$\phi((1 + \delta_u)\ell, \ell, m)^2 \leq r_{\text{NAG}}(\sigma_\phi \kappa)^{1 + \alpha_\phi(\ell - (1 + \delta_\ell)m)}, \tag{139}$$

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

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

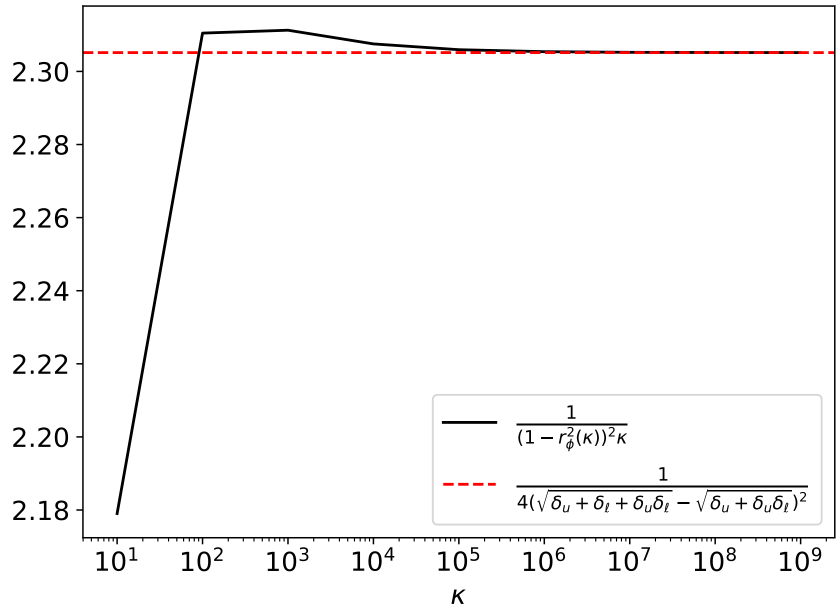

Figure 2: Numerical (black solid line) lower bound on and asymptotic value (dashed red line) of σ_ϕ such that $r_\phi^2 \leq r_{\text{NAG}}(\sigma_\phi\kappa)$ holds for all $\kappa \geq 1 + 1/\delta_s$, with $\delta_u = 0.01$ and $\delta_\ell = 0.18$.

Proof Combining Theorems 33 and 34, we have that

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

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 $[(1+\delta_\ell)m, L/(1+\delta_u)]$:

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

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

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

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

$$\phi((1+\delta_u\ell, \ell, m))^2 \leq \exp(2a(\ell - (1+\delta_\ell)m))r_{\text{NAG}}(\sigma_\phi\kappa) = r_{\text{NAG}}(\sigma_\phi\kappa)^{1+\alpha_\phi(\ell-(1+\delta_\ell)m)}, \quad (140)$$

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

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

which proves the claim. ■

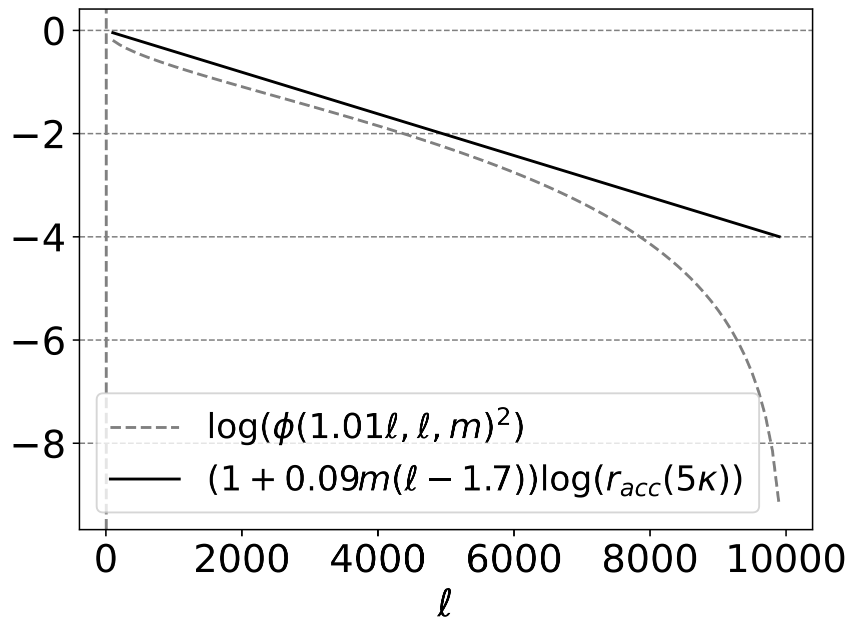

Figure 3: Numerical illustration of Theorem 35 with $L = 10^4, m = 1, \delta_\ell = 0.01, \sigma_\phi = 5$ and $\alpha_\phi = 0.09$.

Figure 3 illustrates Theorem 35 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 $r_{\text{NAG}}(\sigma_\phi \kappa)$ 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 36 *Let $f \in \mathcal{F}(L, m)$ be a quadratic function, suppose that Assumption 4.2 holds for some $L_0 > L$ and let $\kappa = L_0/m$. 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 $r_{\text{NAG}}(\sigma_\phi \kappa) \leq r' < 1$ 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 Theorem 34. Then, there exists some $\nu \geq 1$ such that the estimates m_t of Algorithm 1 reach $[m/\gamma, (1+\delta_m)m]$ after no more than τ iterations, where*

$$\tau = \nu \frac{-\log(4\kappa^2 M_1 \omega/\delta_u)}{\log r'} \tag{141}$$

and $M_1 = \max_i \overline{C}_i / \underline{C}_1$, with ω given by Assumption 4.4.

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 Theorem 35 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 (133) 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

$$c_{t+1} - \ell_0 = \frac{c_{t+1}^2 - \ell_0^2}{c_{t+1}+\ell_0} < \frac{\sum_{i=I_0}^{d}(\lambda_i^2 - \ell_0^2)(x_{i,t+1}-x_{i,t})^2}{\ell_0 \sum_{i=1}^{d}(x_{i,t+1}-x_{i,t})^2} \leq \ell_0\kappa^2 \sum_{i=I_0}^{d} \frac{(x_{i,t+1}-x_{i,t})^2}{(x_{1,t+1}-x_{1,t})^2}. \tag{142}$$

Moreover, using (129), we have that

$$(x_{i,t+1}-x_{i,t})^2 = (\begin{bmatrix}-1 & 1\end{bmatrix} X_{i,t+1})^2 \leq 2\|X_{i,t+1}\|^2 \leq 2\overline{C}_i\rho(\mu_0,\lambda_i)^{2t}x_{i,0}^2. \tag{143}$$

To address the terms in the sum in (142), we combine (143) and (132), 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 Theorem 27 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

$$\begin{aligned}\sum_{i=I_0}^{d} \frac{(x_{i,t+1}-x_{i,t})^2}{(x_{1,t+1}-x_{1,t})^2} &\leq 2M_1 \sum_{i=I_0}^{d} \frac{x_{i,0}^2}{x_{1,0}^2} \prod_{k=t_K}^{t} \frac{\rho(m_k,\lambda_i)^2}{\rho(m_k,m)^2} \prod_{k=1}^{\tau_K} \frac{\rho(m_k,\lambda_i)^2}{\rho(m_k,m)^2} \\ &\leq 2M_1\omega\phi((1+\delta_u)\ell_0,\ell_0,m)^{2(t-t_K+1)}.\end{aligned} \tag{144}$$

where $M_1 = 2\max_i \overline{C}_i/\underline{C}_1$. Next, we put (142) and (144) 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

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

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 $j = 0, \dots, K$ as

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

If $\gamma > 1$, then 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, letting

$$\nu = \sum_{j=0}^{+\infty} \frac{1}{1 + \alpha_\phi m(1+\delta_\ell)(\gamma^j - 1)}, \tag{145}$$

we conclude that $m_{t+1} \leq (1+\delta_m)m$ for all $t \geq \tau$, where

$$\tau = \nu \frac{-\log(4\kappa^2 M_1\omega/\delta_u)}{\log r'} \geq \frac{-\log(4\kappa^2 M_1\omega/\delta_u)}{\log r_{\mathrm{NAG}}(\sigma_\phi\kappa)} \sum_{j=0}^{K} \frac{1}{1 + \alpha_\phi m(1+\delta_\ell)(\gamma^j - 1)},$$

because log is monotone and $r_{\mathrm{NAG}}(\sigma_\phi\kappa) \leq r' < 1$, and $1 + \alpha_\phi m(1+\delta_\ell)(\gamma^j - 1) > 0$. ■

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 37 *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*

$$\max_{m_t \in [m/\gamma, L]} \|P(G_i(m_t), r, \delta)\| < \frac{1 + (1+\delta)^{-2}}{1 - (1+\delta)^{-2}} + \frac{2M_2^2}{(1+\delta)^2 r^2} \frac{1 + (1+\delta)^{-2}}{(1 - (1+\delta)^{-2})^3}, \tag{146}$$

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

Proof By Theorem 28, $\rho(m_t, \lambda_i) < 1$ for all $m_t \in (0, L]$ and $i = 1, \dots, d$. 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 $r_\delta = (1+\delta)r$ and $P = \sum_{k=0}^{+\infty}(G_i(m_t)^\mathsf{T}/r_\delta)^k(G_i(m_t)/r_\delta)^k$. The matrix P is well-defined because $\rho(G_i(m_t)/r_\delta) \leq 1/(1+\delta) < 1$, and $P \succeq I$, by construction. Moreover, it satisfies

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

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 [11, section 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 [11, section 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

$$Q_i(m_t) = \begin{bmatrix}q_{i,1} & q_{i,2}\end{bmatrix}, \tag{147}$$

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

$$T_i(m_t) = Q_i(m_t)^\mathsf{H} G_i(m_t) Q_i(m_t), \tag{148}$$

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

$$T_i(m_t) = \begin{bmatrix} \zeta_i & q_{i,1}(m_t)^\mathsf{H} G_i(m_t) q_{i,2}(m_t)\\ 0 & \xi_i\end{bmatrix}, \tag{149}$$

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 (149) 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 $\rho(m_t, \lambda_i)/r_\delta \le 1/(1+\delta)$, where $r_\delta = (1+\delta)r$, we get

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

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

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

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 38 *Let $f \in \mathcal{F}(L,m)$ be a quadratic function with $\kappa = L/m \geq 4$, and let $L_0 > L$. Suppose that Assumption 4.4 holds for some $\omega > 0$, and that Assumption 4.3 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 is initialized with L_0 as input, then its iterates x_t satisfy*

$$\|x_{t+1} - x^\star\| \leq C\bar{\kappa}^{7/2+2\nu} r_{\text{NAG}}(2\sigma\bar{\kappa})^t \|x_0 - x^\star\|, \tag{150}$$

where $\bar{\kappa} = L_0/m > \kappa$, $\sigma = \max\{\gamma, \sigma_m, \sigma_\phi\}$, $\sigma_m = 1+2\delta_m+2\sqrt{\delta_m(1+\delta_m)}$, $\sigma_\phi = \sigma_\phi(\delta_u, \delta_\ell, \bar{\kappa})$ is a function of $\delta_\ell = (1+\delta_m)/(1+\delta_u) - 1$ 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 and ν are constants that depend on γ, δ_u, σ and ω.

Proof Let $r' = r'(\delta_u, \delta_\ell, \kappa)$ be a function such that $r_{\text{NAG}}(\sigma_\phi \kappa) \leq r' < 1$ for all $\kappa \geq 1 + \delta_\ell$, where $\sigma_\phi = \sigma_\phi(\delta_u, \delta_\ell, \kappa)$ is given by Theorem 34. Then, by Theorem 36 we have that $m_t \leq (1+\delta_m)m$ for all $t \geq \tau$, where

$$\tau = \frac{-\log(4\kappa^2 M_1 \omega/\delta_u)}{\log r'} \sum_{j=0}^{\log_\gamma(\kappa/(1+\delta_m))} \frac{1}{1+\alpha_\phi m(1+\delta_\ell)(\gamma^j - 1)}, \tag{151}$$

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

By design, we have that $m_t \geq m/\gamma$. If $m_t < m$, then by Theorem 25, 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

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

Otherwise, if $m_t \in [m, (1+\delta_m)m]$, then by Theorem 32, we have that $\rho(m_t, m) \leq r_{\text{NAG}}(\sigma_m\kappa)$ 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, $\rho(m_t, m) \leq r_{\text{NAG}}(\sigma_1\kappa)$ for all $t \geq \tau$, where $\sigma_1 = \max\{\gamma, \sigma_m\}$.

Now, by Theorem 27, we have that $\rho(m_t, \lambda_i) \leq r_{\text{NAG}}(\sigma_1\kappa)$ for all λ_i. Hence, given δ_σ such that $(1+\delta_\sigma)r_{\text{NAG}}(\sigma_1\kappa) < 1$, by Theorem 37 there is a $P_i(m_t) = P_i(m_t, \delta_\sigma) \succeq I$ for each λ_i such that $G_i(m_t)^\mathsf{T} P_i(m_t) G_i(m_t) \preceq (1+\delta_\sigma)^2 r_{\text{NAG}}(\sigma_1\kappa)^2 P_i(m_t)$. Hence, if $t_j \leq t < t_{j+1}$ and $t \geq \tau$, then

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

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

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

Rearranging the above yields

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

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

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

In turn, the above bound yields

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

Since r_{NAG} 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

$$\|X_t\|^2 \le M_3^2 ((1 + \delta_\sigma) r_{\text{NAG}}(\sigma \kappa))^{2(t - \lceil \tau \rceil)} \|X_{\lceil \tau \rceil}\|^2, \tag{152}$$

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 $r' = (1 + \delta_\sigma) r_{\text{NAG}}(\sigma\kappa)$ in (151), it follows that $m_t \in [m/\gamma, (1 + \delta_m)m]$ for all $t \ge \tau$, where

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

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

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

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

$$((1 + \delta_\sigma) r_{\text{NAG}}(\sigma\kappa))^{-\lceil \tau \rceil} \le M_4, \tag{153}$$

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 (153) into (152), we obtain

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

To establish (150), 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 Theorem 3 to get

$$\|y_{t+1} - x^\star\|^2 \leq \frac{2}{m}(f(y_{t+1}) - f(x^\star)) \leq r_{\text{GD}}(\kappa)^t 16\kappa^4 \|x_0 - x^\star\|^2.$$

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

$$\begin{aligned} \|x_{t+1} - x^\star\|^2 = \|(1+\beta_t)y_{t+1} - \beta_t y_t - x^\star \pm \beta_t x^\star\|^2 &= \|(1+\beta_t)(y_{t+1} - x^\star) - \beta_t(y_t - x^\star)\|^2 \\ &\leq (2\|y_{t+1} - x^\star\| + \|y_t - x^\star\|)^2 \\ &\leq r_{\text{GD}}(\kappa)^{t-1} 144\kappa^4 \|x_0 - x^\star\|^2, \end{aligned} \tag{154}$$

which implies that

$$\|X_{\lceil\tau\rceil}\| \leq \|x_{\lceil\tau\rceil}\| + \|x_{\lceil\tau\rceil-1}\| \leq r_{\text{GD}}(\kappa)^{(\lceil\tau\rceil-3)/2} 24\kappa^2 \|x_0 - x^\star\| \leq 24\kappa^2 \|x_0 - x^\star\|.$$

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

$$\|X_{\lceil\tau\rceil}\| \leq 24\kappa^2 \|x_0 - x^\star\|.$$

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

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

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

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

we have that

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

Combining the two bounds above yields

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

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

$$\|x_{t+1} - x^\star\| \leq M_3 M_4 \sqrt{\kappa}((1+\delta_\sigma) r_{\text{NAG}}(\sigma\kappa))^t \|x_0 - x^\star\|. \tag{155}$$

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

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

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

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

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 (146) and using the fact that $r_{\text{NAG}}(\sigma\kappa) \geq r_{\text{NAG}}(8)$ yields

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

Using the above bound twice yields

$$M_3 = \sqrt{\max_{i=1,\dots,d}\frac{\lambda_{\max}(P_i(\mu_{-1}))}{\lambda_{\min}(P_i(\mu_{-1}))}\max_{i=1,\dots,d}\frac{\lambda_{\max}(P_i(\mu_{-2}))}{\lambda_{\min}(P_i(\mu_{-2}))}} < 7\frac{M_2^2}{\delta_\sigma^3} = 7\cdot 4^3 M_2^2\sigma^{3/2}\kappa^{3/2}. \tag{156}$$

Finally, we prove (150) by plugging (156) into (155), and then replacing κ with $\bar{\kappa}$, so that

$$\|x_{t+1}-x^\star\| \leq C\bar{\kappa}^{7/2+2\nu} r_{\text{NAG}}(2\sigma\bar{\kappa})^t\|x_0-x^\star\|,$$

where the constant C is given by

$$C = 42\cdot 4^4 M_2(4M_1\omega/\delta_u)^\nu\sigma^{3/2}.$$

■

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 the local quadratic approximation of f at $x^\star$.

Iterate Dynamics in the General Case

Under Assumption 4.1, it follows that f is twice continuously differentiable at $x^\star$. Hence, by Taylor's theorem [19, theorem 2.1], the gradient at x_t can be expressed as

$$\begin{aligned}\nabla f(x_t) =&\nabla f(x^\star) + \int_0^1 \nabla^2 f(x^\star + s(x_t - x^\star))(x_t - x^\star)ds \\ =&\nabla^2 f(x^\star)x_t + \int_0^1 (\nabla^2 f(x^\star + s(x_t - x^\star)) - \nabla^2 f(x^\star))(x_t - x^\star)ds \\ =&(H + \tilde{H}_t)(x_t - x^\star),\end{aligned} \tag{157}$$

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

$$\tilde{H}_t = \int_0^1 (\nabla^2 f(x^\star + s(x_t - x^\star)) - \nabla^2 f(x^\star))ds. \tag{158}$$

Moreover, by (154), we have that $\|x_{t+1} - x^\star\| \le \sqrt{144\bar{\kappa}^4 r_{\text{GD}}(\kappa)^{t-1}}\|x_0 - x^\star\|$. Hence, since $r_{\text{GD}}(\kappa) \in (0, 1)$, if $\|x_0 - x^\star\| \le \epsilon^{\max(1,1/\alpha_H)}\sqrt{r_{\text{GD}}(\kappa)/144\bar{\kappa}^4}$ and $\epsilon \le \delta_H$, then for all $t \ge 0$, we have that $\|x_t - x^\star\| \le \delta_H$, and it follows from Assumption 4.1 that

$$\begin{aligned}\|\tilde{H}_t\| &\le \int_0^1 \|\nabla^2 f(x^\star + s(x_t - x^\star)) - \nabla^2 f(x^\star))\|ds \\ &\le L_H \int_0^1 s\|x_t - x^\star\|ds \\ &\le \epsilon L_H r_{\text{GD}}(\kappa)^{t\alpha_H/2}.\end{aligned} \tag{159}$$

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

$$\tilde{H}_t v_j = \sum_{i=1}^d \tilde{h}_{i,j,t} v_i, \qquad\qquad j = 1, \ldots, d. \tag{160}$$

Then, using (160) and the decomposition $x_t - x^\star = \sum_{j=1}^d x_{j,t} v_j$ yields

$$\tilde{H}_t(x_t - x^\star) = \tilde{H}_t \sum_{j=1}^d x_{j,t} v_j = \sum_{j=1}^d x_{j,t} \tilde{H}_t v_j = \sum_{j=1}^d x_{j,t} \sum_{i=1}^d \tilde{h}_{i,j,t} v_i = \sum_{i=1}^d \sum_{j=1}^d \tilde{h}_{i,j,t} x_{j,t} v_i. \tag{161}$$

In turn, combining the decomposition $x_t - x^\star = \sum_{j=1}^d x_{j,t} v_j$ with (157) and (161), 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

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

Therefore, we have that

$$X_{t+1} = (G(m_t) + \tilde{G}_t)X_t, \tag{162}$$

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

$$X_t = \begin{bmatrix} X_{1,t} \\ \vdots \\ X_{d,t} \end{bmatrix}, \tag{163}$$

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

$$G(m_t) = \operatorname{diag}(G_1(m_t), \ldots, G_d(m_t)), \tag{164}$$

$$\tilde{G}_t = \frac{1}{L}\begin{bmatrix} 0 & 0 & \ldots & 0 & 0 \\ -\beta_t \tilde{h}_{1,1,t-1} & (1+\beta_t)\tilde{h}_{1,1,t} & \ldots & -\beta_t \tilde{h}_{1,d,t-1} & (1+\beta_t)\tilde{h}_{1,d,t} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & 0 & 0 \\ -\beta_t \tilde{h}_{d,1,t-1} & (1+\beta_t)\tilde{h}_{d,1,t} & \ldots & -\beta_t \tilde{h}_{d,d,t-1} & (1+\beta_t)\tilde{h}_{d,d,t} \end{bmatrix}, \tag{165}$$

where $G_i(m_t)$, defined by (109), are the system matrices governing the dynamics of each $X_{i,t}$ in the quadratic case where $f(x) = (x - x^\star)^\top H_t (x - x^\star)$. Using the fact that $\tilde{h}_{i,j,t} = v_i^\top \tilde{H}_t v_j$, the matrix $\tilde{G}_t$ given by (165) can be expressed as

$$\tilde{G}_t = \frac{1+\beta_t}{L} W_1^\top V^\top \tilde{H}_t V W_1 - \frac{\beta_t}{L} W_1^\top V^\top \tilde{H}_t V W_2, \tag{166}$$

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 (166), then using the fact that $\beta_t \in [0, 1)$ and lastly plugging in (159), we obtain

$$\begin{aligned} \|\tilde{G}_t\| &\le \frac{2}{L}\|W_1^\mathsf{T}\|\|V^\mathsf{T}\|\|\tilde{H}_t\|\|V\|\|W_1\| + \frac{1}{L}\|W_1^\mathsf{T}\|\|V^\mathsf{T}\|\|\tilde{H}_t\|\|V\|\|W_2\| \\ &= \frac{3}{L}\|\tilde{H}_t\| \\ &\le \epsilon\frac{L_H}{L} r_{\mathrm{GD}}(\kappa)^{t\alpha_H/2}. \end{aligned} \tag{167}$$

We continue by noting that if Assumption 4.3 holds for some $\delta_\lambda > 0$, then it also holds for every $\delta'_\lambda < \delta_\lambda$. So, without loss of generality, suppose that Assumption 4.3 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 $\lambda_i < L_0$, then from Corollary 26, 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

$$\delta_T = \min_{m_t\in\mathcal{S}} \min_{i=1,\dots,d} |\zeta_i(m_t) - \xi_i(m_t)| > 0, \tag{168}$$

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 (126) 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

$$\|T_i(m_t)^{-1}\| = \frac{1}{|\zeta_i - \xi_i|}\left\| \begin{bmatrix} \xi_i & -1 \\ -\zeta_i & 1 \end{bmatrix} \right\| \le \frac{1}{\delta_T}\left\| \begin{bmatrix} \xi_i & -1 \\ -\zeta_i & 1 \end{bmatrix} \right\|. \tag{169}$$

Hence, since both sides of (169) 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

$$M_T = \max_{m_t\in\mathcal{S}} \max_{i=1,\dots,d} \|T_i(m_t)\|\|T_i(m_t)^{-1}\| < +\infty. \tag{170}$$

Then, let T denote the coordinate transformation given by

$$T(m_t) = \mathrm{diag}(T_1(m_t), \dots, T_d(m_t)). \tag{171}$$

The block-diagonal structure of T combined with (170) implies that

$$\max_{m_t \in \mathcal{S}} \|T(m_t)\| \|T(m_t)^{-1}\| \leq M_T. \tag{172}$$

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

$$G(m_t) = T(m_t) D(m_t) T(m_t)^{-1}, \tag{173}$$

where $D(m_t)$ is the block-diagonal matrix defined as $D(m_t) = \mathrm{diag}(D_1(m_t), \dots, D_d(m_t))$ and $D_i(m_t)$ are the diagonal matrices given by (127). So, defining the state $Z_t = T^{-1}(\mu_0) X_t$ for $t \in [t_0, t_1]$ and plugging Z_t and (173) into (162), since $m_t \equiv \mu_0$ for $t \in [t_0, t_1)$, it follows that

$$\begin{aligned} Z_{t+1} &= T^{-1}(\mu_0) X_{t+1} \\ &= T^{-1}(\mu_0)(G(m_t) + \tilde{G}_t) X_t \\ &= T^{-1}(G(\mu_0) + \tilde{G}_t) T(\mu_0) Z_t \\ &= (D(\mu_0) + \tilde{D}_t) Z_t, \end{aligned} \tag{174}$$

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

$$\tilde{D}_t = T^{-1}(m_t) \tilde{G}_t T(m_t). \tag{175}$$

Using submultiplicativity and then combining (167) with (172), yields

$$\|\tilde{D}_t\| \leq \|T^{-1}(m_t)\| \|\tilde{G}_t\| \|T(m_t)\| \leq \epsilon M_T \frac{L_H}{L} r_{\mathrm{GD}}(\kappa)^{t\alpha_H/2}. \tag{176}$$

Then, summing (176), we obtain

$$\sum_{t=0}^{+\infty} \|\tilde{D}_t\| \leq \epsilon M_T \frac{L_H}{L} \sum_{t=0}^{+\infty} r_{\mathrm{GD}}(\kappa)^{t\alpha_H/2} \leq \epsilon M_T \frac{L_H}{L} \frac{1}{1 - r_{\mathrm{GD}}(\kappa)^{\alpha_H/2}}. \tag{177}$$

Moreover, since $\lambda_i \leq L < L_0$, it follows that $G(m_t)$ are nonsingular. This fact combined with (177) allows us to use results from the theory of asymptotic integration of difference equations [4] to establish that the solutions to (174) are perturbed solutions of the particular case when $\tilde{D}_t \equiv 0$. Namely, by [4, theorem 3.4], 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

$$
\begin{aligned}
X_{t+1} =& T(\mu_J)[I+O(\epsilon)]D(\mu_J)^{t-t_J}T(\mu_J)^{-1}\left(\prod_{j=0}^{J-1} T(\mu_j)[I+O(\epsilon)]D(\mu_j)^{t_{j+1}-t_j}T(\mu_j)^{-1}\right)X_0 \\
=& T(\mu_J)D(\mu_J)^{t-t_J}T(\mu_J)^{-1}\left(\prod_{j=0}^{J-1} T(\mu_j)D(\mu_j)^{t_{j+1}-t_j}T(\mu_j)^{-1}\right)X_0 \\
&+ T(\mu_J)O(\epsilon)D(\mu_J)^{t-t_J}T(\mu_J)^{-1}\left(\prod_{j=0}^{J-1} T(\mu_j)D(\mu_j)^{t_{j+1}-t_j}T(\mu_j)^{-1}\right)X_0 + \dots \\
&+ T(\mu_J)O(\epsilon)D(\mu_J)^{t-t_J}T(\mu_J)^{-1}\left(\prod_{j=0}^{J-1} T(\mu_j)O(\epsilon)D(\mu_j)^{t_{j+1}-t_j}T(\mu_j)^{-1}\right)X_0, \qquad (178)
\end{aligned}
$$

where $t \in [t_J, t_{J+1})$.

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

$$X_{i,t+1} = \begin{bmatrix} 0 & \dots & 0 & I & 0 & \dots & 0 \end{bmatrix} X_{t+1},$$

where $\begin{bmatrix} 0 & \dots & 0 & I & 0 & \dots & 0 \end{bmatrix} \in \mathbb{R}^{2\times 2d}$ 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

$$
\begin{aligned}
&\begin{bmatrix} 0 & \dots & 0 & I & 0 & \dots & 0 \end{bmatrix} T(\mu_J)D(\mu_J)^{t-t_J}T(\mu_J)^{-1}\left(\prod_{j=0}^{J-1} T(\mu_j)D(\mu_j)^{t_{j+1}-t_j}T(\mu_j)^{-1}\right)X_0 \\
&= \begin{bmatrix} 0 & \dots & 0 & I & 0 & \dots & 0 \end{bmatrix} G(\mu_J)^{t-t_J}\prod_{j=0}^{J-1} G(\mu_j)^{t_{j+1}-t_j}X_0 \\
&= \begin{bmatrix} 0 & \dots & 0 & G_i^{t-t_J}\prod_{j=0}^{J-1} G_i(\mu_j)^{t_{j+1}-t_j} & 0 & \dots & 0 \end{bmatrix} X_0 \\
&= G_i^{t-t_J}\prod_{j=0}^{J-1} G_i(\mu_j)^{t_{j+1}-t_j}X_{i,0},
\end{aligned}
$$

since $G_i(\mu_j) = T_i(\mu_j)D_i(\mu_j)T_i(\mu_j)^{-1}$, by (127), 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 (178) are $O(\epsilon)$, and $\rho(\mu_j, \lambda_i) \le \rho(\mu_j, m)$ for all the eigenvalues $\rho(\mu_j, \lambda_i)$ of $D(\mu_j)$, by Theorem 27. Therefore, combining the above remarks with Assumption 4.4, it follows that for $t \in [t_j, t_{j+1})$

$$
\begin{aligned}
\|X_{i,t+1}\|^2 \le& \overline{C}_i \rho(\mu_j, \lambda_i)^{2(t-t_j)}\left(\prod_{k=0}^{j-1}\rho(\mu_k, \lambda_i)^{2(t_{k+1}-t_k)}\right)x_{i,0}^2 \\
&+ O(\epsilon)\rho(\mu_j, m)^{2(t-t_j)}\left(\prod_{k=0}^{j-1}\rho(\mu_k, m)^{2(t_{k+1}-t_k)}\right)x_{1,0}^2,
\end{aligned}
$$

for some $\overline{C}_i$. The above bound is analogous to (129), 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

$$\begin{aligned}(x_{i,t+1} - x_{i,t})^2 &= (\begin{bmatrix}-1 & 1\end{bmatrix} X_{i,t+1})^2 \\ &\le 2\overline{C}_i \rho(\mu_j, \lambda_i)^{2(t-t_j)} \left(\prod_{k=0}^{j-1} \rho(\mu_k, \lambda_i)^{2(t_{k+1}-t_k)}\right) x_{i,0}^2 \\ &\quad + O(\epsilon)\rho(\mu_j, m)^{2(t-t_j)} \left(\prod_{k=0}^{j-1} \rho(\mu_k, m)^{2(t_{k+1}-t_k)}\right) x_{1,0}^2. \end{aligned} \tag{179}$$

In the same vein, combining (178) with Theorem 27, we have that for $t \in [t_j, t_{j+1})$

$$\|X_{t+1}\|^2 \le 2(1 + O(\epsilon))\overline{C}\rho(\mu_j, m)^{2(t-t_j)} \left(\prod_{k=0}^{j-1} \rho(\mu_k, m)^{2(t_{k+1}-t_k)}\right) \|x_0\|^2, \tag{180}$$

where $\overline{C} = \max_{i=1,\ldots,d} \overline{C}_i$. Similarly, combining the derivation of (132) with (178) and Assumption 4.4, for $t \in [t_j, t_{j+1})$ we have that

$$(x_{1,t+1} - x_{1,t})^2 \ge (1 - O(\epsilon))\underline{C}_1 \rho(\mu_j, m)^{2(t-t_j)} \left(\prod_{k=0}^{j-1} \rho(\mu_k, m)^{2(t_{k+1}-t_k)}\right) x_{1,0}^2, \tag{181}$$

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 (157) 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 (159), it follows that

$$
\begin{aligned}
&\sum_{i=1}^d \left(\sum_{j=1}^d (\tilde{h}_{i,j,t+1} x_{j,t+1} - \tilde{h}_{i,j,t} x_{j,t}) \right)^2 \\
&\leq \sum_{i=1}^d \left(\sum_{j=1}^d (|\tilde{h}_{i,j,t+1}| + |\tilde{h}_{i,j,t}|)(|x_{j,t+1}| + |x_{j,t}|) \right)^2 \\
&\leq \sum_{i=1}^d \left(\sum_{j=1}^d (|\tilde{h}_{i,j,t+1}| + |\tilde{h}_{i,j,t}|)^2 \right) \left(\sum_{j=1}^d (|x_{j,t+1}| + |x_{j,t}|)^2 \right) \\
&\leq \sum_{i=1}^d \left(2 \sum_{j=1}^d (\tilde{h}_{i,j,t+1}^2 + \tilde{h}_{i,j,t}^2) \right) \left(2 \sum_{j=1}^d (x_{j,t+1}^2 + x_{j,t}^2) \right) \\
&= 4(\|\tilde{H}_{t+1}\|_F^2 + \|\tilde{H}_t\|_F^2)\|X_{t+1}\|^2 \\
&\leq 4d(\|\tilde{H}_{t+1}\|^2 + \|\tilde{H}_t\|^2)\|X_{t+1}\|^2 \\
&\leq 4\epsilon^2 L_H^2 d \|X_{t+1}\|^2.
\end{aligned} \tag{182}
$$

In turn, we address the second term of $\|\nabla f(x_{t+1}) - \nabla f(x_t)\|^2$ using (182), which gives

$$
\begin{aligned}
&\sum_{i=1}^d (x_{i,t+1} - x_{i,t}) \sum_{j=1}^d (\tilde{h}_{i,j,t+1} x_{j,t+1} - \tilde{h}_{i,j,t} x_{j,t}) \\
&\leq \sqrt{\sum_{i=1}^d (x_{i,t+1} - x_{i,t})^2} \sqrt{\sum_{i=1}^d \left(\sum_{j=1}^d (\tilde{h}_{i,j,t+1} x_{j,t+1} - \tilde{h}_{i,j,t} x_{j,t}) \right)^2} \\
&\leq \sqrt{2 \sum_{i=1}^d (x_{i,t+1}^2 + x_{i,t}^2)} \sqrt{4\epsilon^2 L_H^2 d \|X_{t+1}\|^2} \\
&\leq 2\epsilon L_H \sqrt{2d} \|X_{t+1}\|^2.
\end{aligned} \tag{183}
$$

Then, combining (180), (182) and (183) we obtain

$$
\begin{aligned}
\|\nabla f(x_{t+1}) - \nabla f(x_t)\|^2 \leq &\sum_{i=1}^d \lambda_i^2 (x_{i,t+1} - x_{i,t})^2 \\
&+ O(\epsilon) \rho(\mu_j, m)^{2(t-t_j)} \left(\prod_{k=0}^{j-1} \rho(\mu_k, m)^{2(t_{k+1} - t_k)} \right) x_{1,0}^2.
\end{aligned} \tag{184}
$$

In turn, plugging (184) into (5), and then using (181), it follows that

$$
c_{t+1}^2 = c(x_{t+1}, x_t)^2 \leq \frac{\sum_{i=1}^d \lambda_i^2 (x_{i,t+1} - x_{i,t})^2}{\sum_{i=1}^d (x_{i,t+1} - x_{i,t})^2} + O(\epsilon). \tag{185}
$$

Lemma 39 *Let $f \in \mathcal{F}(L, m)$, suppose that Assumptions 4.1 to 4.4 hold and let $\kappa = L_0/m$. 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 $r_{\text{NAG}}(\sigma_\phi \kappa) \leq r' < 1$ 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 Theorem 34. Then, there exist ν and $\epsilon > 0$ 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*

$$\tau = \nu \frac{-\log(8\kappa^2 M_1 \omega/\delta_u)}{\log r'} \tag{186}$$

$M_1 = \max_i \overline{C}_i/\underline{C}_1$, with ω given by Assumptions 4.1, 4.3 and 4.4.

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 Theorem 35 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 (185) 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

$$\begin{aligned} c_{t+1} - \ell_0 = \frac{c_{t+1}^2 - \ell_0^2}{c_{t+1} + \ell_0} &< \frac{\sum_{i=I_0}^{d}(\lambda_i^2 - \ell_0^2)(x_{i,t+1} - x_{i,t})^2}{\ell_0 \sum_{i=1}^{d}(x_{i,t+1} - x_{i,t})^2} + O(\epsilon) \\ &\leq \ell_0 \kappa^2 \sum_{i=I_0}^{d} \frac{(x_{i,t+1} - x_{i,t})^2}{(x_{1,t+1} - x_{1,t})^2} + O(\epsilon). \end{aligned} \tag{187}$$

To address the terms in the sum in (187), we combine (179) and (181), 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 Theorem 27 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

$$\begin{aligned} \sum_{i=I_0}^{d} \frac{(x_{i,t+1} - x_{i,t})^2}{(x_{1,t+1} - x_{1,t})^2} &\leq 2(1 + O(\epsilon))M_1 \sum_{i=I_0}^{d} \frac{x_{i,0}^2}{x_{1,0}^2} \prod_{k=t_K}^{t} \frac{\rho(m_k, \lambda_i)^2}{\rho(m_k, m)^2} \prod_{k=1}^{\tau_K} \frac{\rho(m_k, \lambda_i)^2}{\rho(m_k, m)^2} + O(\epsilon) \\ &\leq 2(1 + O(\epsilon))M_1 \omega \phi((1+\delta_u)\ell_0, \ell_0, m)^{2(t - t_K + 1)} + O(\epsilon). \end{aligned} \tag{188}$$

where $M_1 = 2\max_i \overline{C}_i/\underline{C}_1$. Next, we put (187) and (188) together, and since $\ell_0 \geq m$, by choosing ϵ sufficiently small, we get

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

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

$$\Delta t_0 = -\frac{\log(8\kappa^2 M_1 \omega/\delta_u)}{\log r_{\text{NAG}}(\sigma_\phi \kappa)}.$$

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 by applying Theorem 27 before, it follows that

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

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 $j = 0, \ldots, K$ as

$$\Delta t_j = \frac{-\log(8\kappa^2 M_1 \omega/\delta_u)}{\log r_{\text{NAG}}(\sigma_\phi \kappa)} \frac{1}{1 + \alpha_\phi(\ell_j - (1+\delta_\ell)m)}.$$

If $\gamma > 1$, then 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, letting ν be given by (145), as

$$\nu = \sum_{j=0}^{+\infty} \frac{1}{1 + \alpha_\phi m (1+\delta_\ell)(\gamma^j - 1)},$$

we conclude that $m_{t+1} \leq (1+\delta_m)m$ for all $t \geq \tau$, where

$$\tau = \nu \frac{-\log(8\kappa^2 M_1 \omega/\delta_u)}{\log r'} \geq \frac{-\log(8\kappa^2 M_1 \omega/\delta_u)}{\log r_{\text{NAG}}(\sigma_\phi \kappa)} \sum_{j=0}^{K} \frac{1}{1 + \alpha_\phi m (1+\delta_\ell)(\gamma^j - 1)},$$

because $\log$ is monotone and $r_{\text{NAG}}(\sigma_\phi \kappa) \leq r' < 1$, and $1 + \alpha_\phi m (1+\delta_\ell)(\gamma^j - 1) > 0$. ■

B.3. Main Result

At last, we are ready to prove Theorem 5, establishing that Algorithm 1 achieves acceleration around the minimum.

Proof [Proof of Theorem 5] 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 $r_{\text{NAG}}(\sigma_\phi \kappa) \leq r' < 1$ for all $\kappa \geq 1 + \delta_\ell$, where $\sigma_\phi = \sigma_\phi(\delta_u, \delta_\ell, \kappa)$ is given by Theorem 34. By Theorem 39, there is some ν such that $m_t \leq (1+\delta_m)m$ for all $t \geq \tau$, where

$$\tau = \nu \frac{-\log(8\kappa^2 M_1 \omega/\delta_u)}{\log r'} \tag{189}$$

$M_1 = \max_i \overline{C}_i/\underline{C}_1$. In turn, as in the proof of Theorem 38, Theorem 32 then implies that $\rho(m_t, m) \leq r_{\text{NAG}}(\sigma_1 \kappa)$ 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 Theorem 27, we have that $\rho(m_t, \lambda_i) \leq r_{\text{NAG}}(\sigma_1\kappa)$ for all λ_i. Hence, given δ_σ such that $(1+\delta_\sigma) r_{\text{NAG}}(\sigma_1\kappa) < 1$, by Theorem 37 there is a $P_i(m_t) = P_i(m_t, \delta_\sigma) \succeq I$ for each λ_i such that $G_i(m_t)^\top P_i(m_t) G_i(m_t) \preceq (1+\delta_\sigma)^2 r_{\text{NAG}}(\sigma_1\kappa)^2 P_i(m_t)$. 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 $G(m_t)^\top P(m_t) G(m_t) \preceq (1+\delta_\sigma)^2 r_{\text{NAG}}(\sigma_1\kappa)^2 P(m_t)$. Hence, if $t_j \leq t < t_{j+1}$ and $t \geq \tau$, then (162) yields

$$\begin{aligned} X_{t+1}^\top P(m_t) X_{t+1} &= X_t^\top (G(\mu_j) + \tilde{G}_t)^\top P(\mu_j)(G(\mu_j) + \tilde{G}_t) X_t \\ &\leq (1+\delta_\sigma)^2 r_{\text{NAG}}(\sigma_1\kappa)^2 X_t^\top P(\mu_j) X_t + X_t^\top \tilde{P}_t X_t, \end{aligned} \tag{190}$$

since $m_t = \mu_j$, where

$$\tilde{P}_t = \tilde{G}_t^\top P(m_t) G(m_t) + G(m_t)^\top P(m_t) \tilde{G}_t + \tilde{G}_t^\top P(m_t) \tilde{G}_t.$$

By Equation (109), we have that

$$\begin{aligned} \|G_i(m_t)\| &\leq \left\| \begin{bmatrix} 0 & 1 \\ -\beta(m_t)\left(1 - \frac{\lambda_i}{L}\right) & 0 \end{bmatrix} \right\| + \left\| \begin{bmatrix} 0 & 0 \\ 0 & (1+\beta(m_t))\left(1 - \frac{\lambda_i}{L}\right) \end{bmatrix} \right\| \\ &= \max\left\{1, \beta(m_t)\left(1 - \frac{\lambda_i}{L}\right)\right\} + (1+\beta(m_t))\left(1 - \frac{\lambda_i}{L}\right) \\ &\leq 3. \end{aligned} \tag{191}$$

Furthermore, since $\delta_\sigma > 0$ and $r_{\text{NAG}}(\sigma_1\kappa) \geq r_{\text{NAG}}(4) = 1/2$, because by assumption $\kappa \geq 4$, the block diagonal structure of P combined with (146) yields

$$\|P(m_t)\| \leq \frac{2}{1 - (1+\delta_\sigma)^{-2}} + \frac{16 M_2^2}{(1 - (1+\delta_\sigma)^{-2})^3} = M_\delta. \tag{192}$$

Therefore, combining (167), (191) and (192), and taking ϵ such that $\epsilon L_H/L < 1$, we obtain

$$\|\tilde{P}_t\| \leq (2\|G(m_t)\| + \|\tilde{G}_t\|)\|\tilde{G}_t\| \leq 7\epsilon M_\delta L_H/L. \tag{193}$$

Then, since $P(m_t) \succeq I$, from (190) and (193) it follows that

$$X_{t+1}^T P X_{t+1} \leq ((1+\delta)^2 r^2 + 7\epsilon M_\delta L_H/L) X_t^T P X_t = \tilde{r}^2 X_t^T P X_t, \tag{194}$$

where we conveniently use a perturbed rate $\tilde{r}$, given by

$$\tilde{r} = \sqrt{(1+\delta_\sigma)^2 r^2 + 7\epsilon M_\delta L_H/L}. \tag{195}$$

Consecutively applying (194) and reproducing the steps in the proof of Theorem 38, we get

$$\|x_{t+1} - x^\star\| \leq C' \bar{\kappa}^{7/2+2\nu} \tilde{r}^t \|x_0 - x^\star\|, \tag{196}$$

where the constant C is given by

$$C' = 42 \cdot 4^4 M_2 (8 M_1 \omega/\delta_u)^\nu \sigma_2^{3/2},$$

with ν, $\sigma_2 = \max\{\gamma, \sigma_m, \sigma_\phi\}$, and $\delta_\sigma = 1/(4\sqrt{\sigma_2}\kappa)$ is chosen such that

$$(1+\delta_\sigma) r_{\text{NAG}}(\sigma_2 \kappa) < r_{\text{NAG}}(2\sigma_2\kappa).$$

Hence, choosing ϵ sufficiently small such that

$$\sqrt{(1+\delta_\sigma)^2 r^2 + 7\epsilon M_\delta L_H/L} \leq r_{\text{NAG}}(2\sigma_2\kappa),$$

and then plugging this choice of ϵ back into (196), we obtain

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

where $\sigma = 2\sigma_2$ and $C = C'\bar{\kappa}^{7/2+2\nu}$, which concludes the proof. ■

To conclude this section, we make a few remarks on C and σ in Theorem 5.

One of the factors of C involves a power ν, which is defined by (145) and implicitly involves some quantities that are arbitrarily set in the local analysis such as δ_ℓ. Figure 3 illustrates a numerical example for a particular choice of these quantities, in which case $\nu \approx 1.9$. In reality, ν is an artifact of a conservative analysis and does not play a role in practical performance. Indeed, in Theorems 36 and 39 we bound the number of iterates that m_t takes to be updated to a new value disregarding that the λ_i-coordinates for $\lambda_i \geq m_t$ have already been reduced substantially relative to the others in the previous update, which is reflected in the value of c_t. In other words, we analyze the convergence of m_t as if it was starting from the same initial conditions every time for every update.

Now consider the suboptimality factor σ. In the proof of Theorem 5, we work with $\sigma = 2\sigma_2$, where $\sigma_2 = \max\{\gamma, \sigma_m, \sigma_\phi\}$, which is a function of $\sigma_m = 1 + 2\delta_m + 2\sqrt{\delta_m(1+\delta_m)}$ and $\sigma_\phi \lesssim 1/4(\sqrt{\delta_u + \delta_\ell(1+\delta_u)} - \sqrt{\delta_u(1+\delta_\ell)})^2$. The suboptimality factor σ_m is decreasing in δ_m, which determines the gap in the upper bound of $[m/\gamma, (1+\delta_m)m]$, the interval that contains m_t in the final convergence regime of NAG-free around $x^\star$. Intuitively, the smaller δ_m is, the better the estimate m_t is in the final regime, therefore a smaller suboptimality rate. On the other hand, if δ_m is small, then so is $\delta_\ell < \delta_m$, therefore σ_ϕ increases. Intuitively, σ_ϕ represents that the time that m_t takes to become sufficiently accurate. Thus, a smaller δ_m means that m_t takes longer to reach the interval $[m/\gamma, (1+\delta_m)m]$. More importantly, as m_t approaches m, the rate at which m_t converges to m slows down. 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 [14], in which case the convergence rate is $\max\{\gamma, \sigma_m, \sigma_\phi\}$. 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 $r_{\text{NAG}}(2.4\kappa)$, which is competitive with restart schemes, where "the convergence rate is slowed down by roughly a factor four" [10, page 167]. Notwithstanding, 2.4 is still conservative and, in the next section, we present experiments in which we see that the suboptimality factor is much closer to 1.

Appendix C. Numerical Experiments

In this section, we validate NAG-free (Algorithm 1) on a classical machine learning problem and examine the practical implications of violating one of the technical assumptions made to prove local acceleration.

We consider the regularized logistic regression objective, given by

$$f(x) = -(1/n) \sum_{i=1}^{n} \log\bigl(1 + \exp(-b_i A_i^\top x)\bigr) + (\eta/2)\|x\|^2, \tag{197}$$

where $\eta > 0$ is a regularization parameter 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 [7] and whose details are summarized on Table 1. Together with η, the datapoints determine the unknown parameters of $f \in \mathcal{F}(L,m)$, bounded by $L \leq \bar{L} = (1/4n)\lambda_{\max}(A^\top A) + \eta$ and $m \geq \eta$, where A denotes the matrix with rows $A_i^\top$ and $\lambda_{\max}(A^\top A)$ denotes the top eigenvalue of $A^\top A$. Following the sources (`github.com/ymalitsky/adaptive_GD` and `github.com/konstmish/opt_methods` from which we borrowed the base code for this experiment, we set $\eta = \bar{L}/10n$ and $x_0 = 0$.

Table 1: Details of datasets from LIBSVM [7] used in the logistic regression experiment.

dataset	**datapoints**	**dimensions**
gisette_scale	6,000	5,000
madelon	2,000	500
mushrooms	8,124	112
phishing	11,055	68
svmguide1	3,089	4
w1a	2,477	300

We start with a sanity-check of the estimates of the strong convexity parameter produced by NAG-free when $\gamma = 1.5$ and $\gamma_L = 1.1$. To this end, we compare m_{-1}, the estimate m_t held by NAG-free after 10,000 iterations, with the regularization parameter η. We consider two variants of NAG-free: one where $L_0 = \bar{L}$ and another where $L_0 = \bar{L}/100$. The variant initialized with $L_0 = \bar{L}$ satisfies the assumptions of Theorem 5, therefore we expect its estimate m_{-1} to be accurate and that it achieve acceleration. In contrast, initializing Algorithm 1 with $L_0 = \bar{L}/100$ should activate backtracking, violating the assumptions of Theorem 5. Table 2 presents the values of η, η/γ and the final estimates of each NAG-free variant. We see that for most datasets, the final estimates m_{-1} fall within the interval given by $[\eta/\gamma, \eta]$. The exception is the PHISHING dataset, on which m_{-1} for the two NAG-free variants are roughly five and a half times greater than η. We investigate this discrepancy in more detail below. Thus, for most datasets above, the strong convexity parameter m reduces to the regularization parameter η. To interpret these results, we compare the performance of the two NAG-free variants above with that of the following methods:

- NAG parameterized with $L = \bar{L}$ and $m = \eta$;
- TM, the triple momentum method [27] parameterized with $L = \bar{L}$ and $m = \eta$;
- NAG+R, the (function value) restart scheme [20] parameterized with $L = \bar{L}$;

Table 2: Regularization parameter η, η/γ and the estimates m_{-1} held by NAG-free variants with $\gamma = 1.5$ after 10,000 iterations solving the logistic regression problem. For most datasets, m_{-1} fall within the interval $[\eta/\gamma, \eta]$, except for PHISHING, which is highlighted in gray..

Dataset	η/γ	$m_{-1}(\bar{L})$	$m_{-1}(\bar{L}/100)$	η
gisette_scale	9.37e-3	1.26e-2	1.07e-2	1.40e-2
madelon	9.93e2	1.39e3	1.20e3	1.49e3
mushrooms	2.12e-5	2.24e-5	2.62e-5	3.18e-5
phishing	9.80e-7	7.95e-6	8.35e-6	1.47e-6
svmguide1	1.90e-1	2.47e-1	2.01e-1	2.85e-1
w1a	1.67e-5	2.48e-5	2.13e-5	2.51e-5

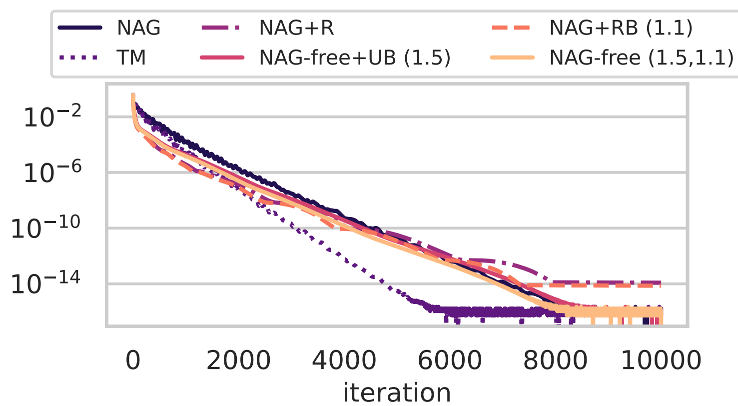

Figure 4: Suboptimality gap $f(x_t) - f(x^\star)$ for logistic regression on the A9A dataset. For NAG-free variants, $\gamma = 1.5$ is used. The backtracking factor is 1.1 for NAG+RB and NAG-free. The NAG-free+UB is initialized with $L_0 = \bar{L} \geq L$.

- NAG+RB, the (function value) restart scheme [20] where L is found via backtracking;

Figures 4 and 5 show the progression suboptimality gap on the A9A and MUSHROOMS datasets, where NAG-free and NAG-free+UB denote the variants of NAG-free initialized with $L_0 = \bar{L}/100$ and $L_0 = \bar{L}$, respectively. We see that TM performs best on A9A, and NAG-free performs best on MUSHROOMS. To explain these results, we compare $\bar{L}$ with the estimate L_{-1} held by NAG-free at the last iteration. On A9A, $\bar{L} = 1.57$ and $L_{-1} = 1.39$, which implies that $\bar{L}$ is a good estimate of the true value of L. Since η also seems to be a good estimate of m, we expect TM to outperform the other methods, since it has the best theoretical convergence rates among the six methods above. On the other hand, $\bar{L} = 2.59$ and $L_{-1} = 0.80$ on MUSHROOMS, suggesting that $\bar{L}$ is a somewhat loose estimate of the true value of L. Similarly, for NAG+RB, $L_{-1} = 1.06$, which is slightly worse than the L estimate produced by NAG-free, even though both methods use the same geometric factor of 1.5 for backtracking.

Now, consider the results obtained from the PHISHING dataset, where the estimates m_{-1} were considerably greater than the regularization parameter. As Figure 6 shows, the NAG-free and restart-

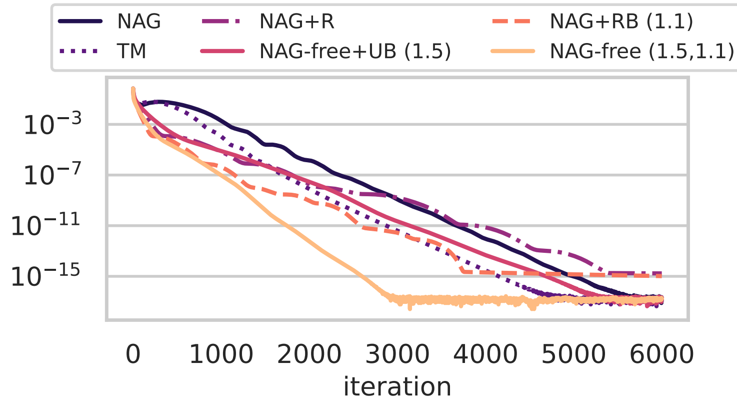

Figure 5: Suboptimality gap $f(x_t) - f(x^\star)$ for logistic regression on the MUSHROOMS dataset. For NAG-free variants, $\gamma = 1.5$ is used. The backtracking factor is 1.1 for NAG+RB and NAG-free. The NAG-free+UB is initialized with $L_0 = \bar{L} \geq L$.

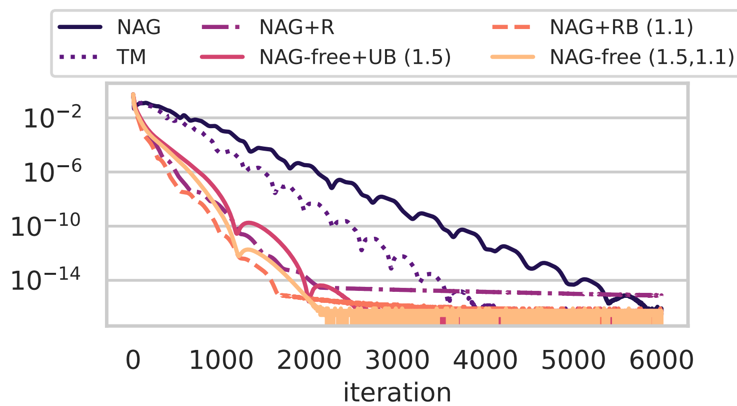

Figure 6: Suboptimality gap $f(x_t) - f(x^\star)$ for logistic regression on PHISHING dataset with $x_0 = 0$. For NAG-free variants, $\gamma = 1.5$ is used. The backtracking factor is 1.1 for NAG+RB and NAG-free. The NAG-free+UB is initialized with $L_0 = \bar{L} \geq L$.

ing methods outperform both NAG and TM. Crucially, backtracking only marginally improves the performance of NAG-free and the restarting scheme. In other words, the NAG-free and restarting methods better NAG and TM thanks to better estimates of the strong convexity parameter. Thus, to assess whether η is a loose estimate of the true strong convexity parameter, we compute the least eigenvalue of $\nabla^2 f(x^\star)$. We find that they approximately match, however. That is, η is actually a good approximation of the true strong convexity parameter. At first, this seems to be at odds with the theory presented above, as we expect that at least the NAG-free+UB to correctly estimate m. To investigate this conundrum, we inspect the estimates produced by this NAG-free variant.

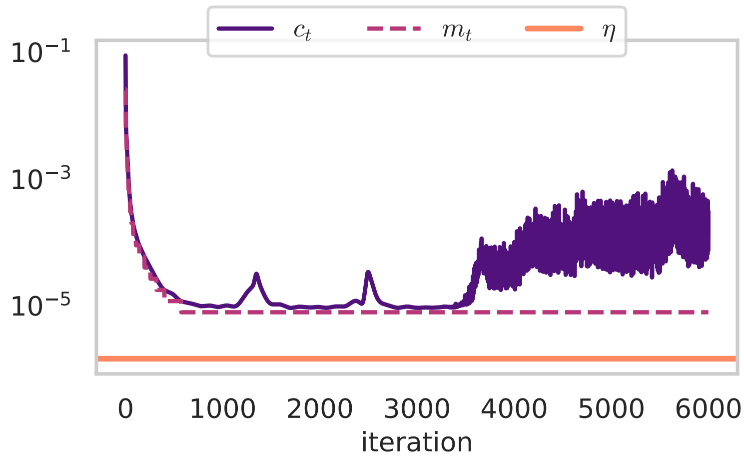

Figure 7: Estimates of m for logistic regression on PHISHING dataset with $x_0 = 0$. For NAG-free variants, $\gamma = 1.5$ is used. The backtracking factor is 1.1 for NAG+RB and NAG-free. The NAG-free+UB is initialized with $L_0 = \bar{L} \geq L$.

Figure 7 shows the NAG-free+UB estimates c_t and m_t, along with η. We see that in the first half of the iterations, the estimates converge exponentially to the final value m_{-1}, which would be in accordance with theory, except $m_{-1} \neq m$. In the second half, c_t begins to jitter arbitrarily, and we notice that at this point the suboptimality gap has essentially reached machine precision. Based on these pattern, we hypothesize the following: the coordinates of x_t on the eigenspace associated with m are initially very small, and by the time they would become significant enough for m_t to converge to m, numerical errors corrupt the estimate m_t. To test this hypothesis, we slightly perturb the initial point, sampling $x_0 \sim 10^{-6} \times \mathcal{U}[0,1)^d$. Figure 8 shows the corresponding m estimates. We see that, indeed, after reaching an initial plateau, c_t and m_t eventually reach the correct value of m, just before c_t starts to jitter. Figure 10 and Figure 11 show that NAG-free estimates of m behave similarly to the NAG-free+UB estimates when $x_0 = 0$ and $x_0 \sim 10^{-6} \times \mathcal{U}[0,1)^d$. Therefore, we conclude that around the origin, the coordinates of x_0 on the eigenspace associated with m are really small. Effectively, this improves the condition number of the problem, which NAG-free and restart methods take advantage of to achieve superior rates of convergence. In summary, NAG-free and restart methods adapt to and benefit from better local conditioning, opposite to methods with constant parametrization.

To further substantiate the local adaption phenomenon, we elaborate a stylized problem intended to capture the essence of the phenomenon. Namely, we consider a simple three-dimensional quadratic objective $f(x) = (1/2)x^\mathsf{T} Hx$, where $H = \mathrm{diag}(1, 5, 10^4)$, and then fix $x_0 = (1, 10^3, 1)^\mathsf{T}$. The left-hand y-axis on Figure 9 shows the suboptimality gap obtained by NAG and NAG-free+UB solving this quadratics problem, while the right-hand y-axis shows the NAG-free+UB estimates m_t. For reference, the dashed lines r_1 and r_5 represent the nominal performance from methods respectively converging at rates $r_{\text{ACC}}(L/m)$ and $r_{\text{ACC}}(L/\lambda_2)$, where $m = 1$, $\lambda_2 = 5$ and $L = 10^4$. Likewise, the dashed horizontal black lines mark the values of λ_2 and m. We see that initially, NAG-free+UB converges roughly at the rate $r_{\text{ACC}}(L/\lambda_2)$, and then eventually settles

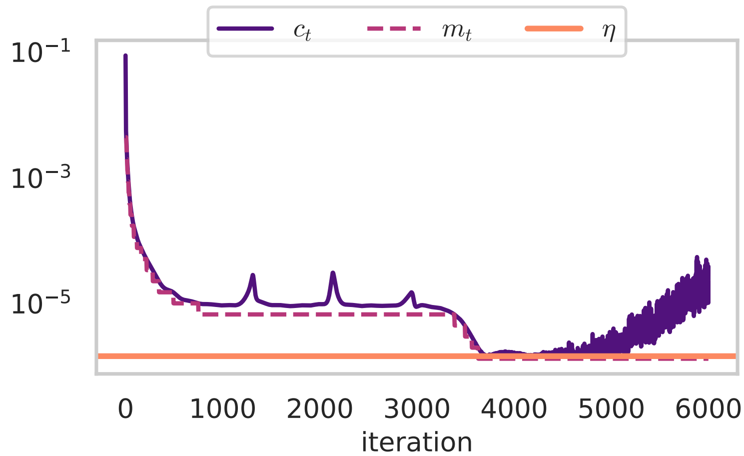

Figure 8: NAG-free+UB estimates of m for logistic regression on PHISHING dataset, with $x_0 \sim 10^6 \times \mathcal{U}[0,1)^d$.

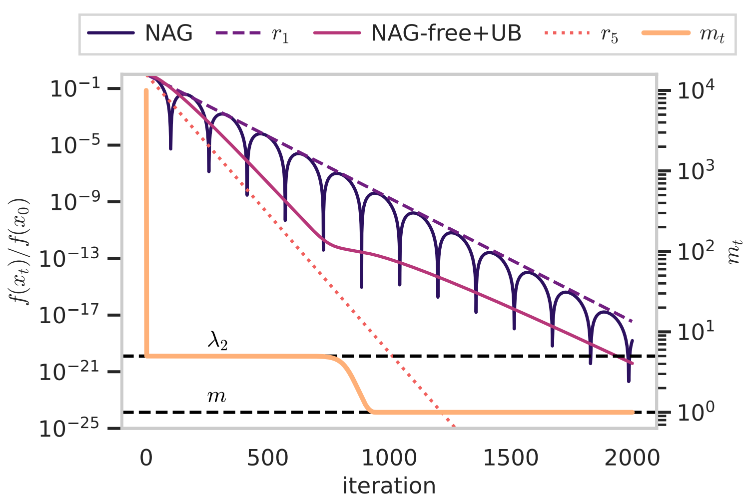

Figure 9: Suboptimality gap for a quadratics problem superimposed by NAG-free+UB estimates of m and by accelerated rates when $m = 1$ and $m = 5$, r_1 and r_5.

at the nominal rate for this problem, $r_{\text{ACC}}(L/m)$. As the plot of m_t show, the estimates determine in which regime NAG-free+UB operates. Therefore, NAG-free+UB is able to adapt to the beneficial initial distribution of the x_0 coordinates, which improves the effective condition number of the problem by a factor of 5, As a result, NAG-free+UB converges substantially faster than NAG until the m-coordinates of x_t become non-negligible relative to the λ_2-coordinates.

The PHISHING and stylized quadratic examples above illustrate that in practice Assumption 4.4 need not hold for $x_{1,0}$, but it must hold for some other $x_{i,0}$ with $i \geq 1$. That is, it may be that

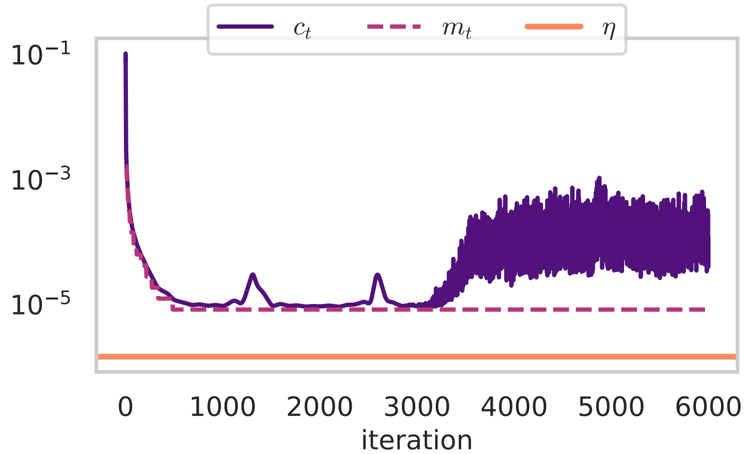

Figure 10: NAG-free estimates of m for logistic regression on the PHISHING dataset with $x_0 = 0$.

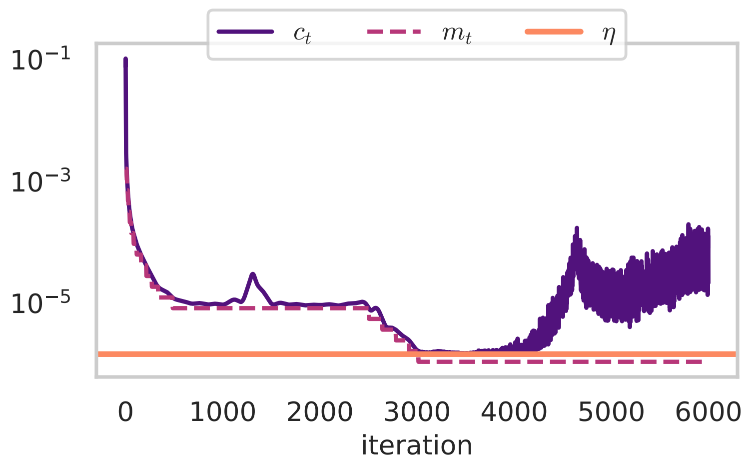

Figure 11: NAG-free estimates of m for logistic regression on the PHISHING dataset with $x_0 \sim 10^{-6} \times \mathcal{U}[0,1)^d$.

the m-coordinates of x_0 are sufficiently small for the effective strong convexity parameter to be some other eigenvalue $\lambda_i \geq m$ of $\nabla^2 f(x^\star)$. In turn, the effective condition number of the problem improves to $L/\lambda_i \leq L/m$. NAG-free is able to adapt to this improved condition number, achieving better convergence rates.