A APPENDIX

A.1 CONTINUOUS-TIME ANALYSIS OF QUADRATICS

Consider the μ-strongly convex function $f : \mathbb{R}^d \to \mathbb{R}$ to be a quadratic of the form $f(X) = \frac{1}{2}X^T \boldsymbol{A} X$ where $\boldsymbol{A}$ is a diagonal matrix. Then the second-order ODE (11) can be rewritten as

$$\begin{bmatrix} \dot{X} \\ \dot{V} \end{bmatrix} = \begin{bmatrix} \mathbf{0}_d & \boldsymbol{I}_d \\ -(np+mq)\boldsymbol{A} & -(n+q)\boldsymbol{I}_d - m\boldsymbol{A} \end{bmatrix} \begin{bmatrix} X \\ V \end{bmatrix}, \tag{16}$$

where $\mathbf{0}_d$ is $d \times d$ zero matrix, and $\boldsymbol{I}_d$ is $d \times d$ identity matrix. Without loss of generality, we assume that $\boldsymbol{A}$ is a diagonal matrix and its entries are sorted decreasingly. The equation (16) is solved by taking $Z = [X^T \quad V^T]^T$ and

$$\boldsymbol{M} = -\begin{bmatrix} \mathbf{0}_d & -\boldsymbol{I}_d \\ (np+mq)\boldsymbol{A} & (n+q)\boldsymbol{I}_d + m\boldsymbol{A} \end{bmatrix},$$

so that $\dot{Z} = -\boldsymbol{M}Z$ which has the famous solution $Z = e^{-\boldsymbol{M}t}Z_0$ for an initialization Z_0. Note that we have

$$\begin{aligned} \|Z\|_2 \leq \|e^{-\boldsymbol{M}t}\|_{2\to 2}\|Z_0\|_2 \leq \sum_{k=0}^{\infty} \frac{\|(-\boldsymbol{M})^k\| t^k}{k!}\|Z_0\|_2 \leq \sum_{k=0}^{\infty} \frac{(\rho(-\boldsymbol{M}) + o(1))^k t^k}{k!}\|Z_0\|_2 \\ = e^{(\rho(-\boldsymbol{M}+o(1)))t}\|Z_0\|_2 \end{aligned}$$

where $\rho(\boldsymbol{M})$ is the spectral radius of $\boldsymbol{M}$, $\|.\|_{2\to 2}$ denotes the spectral norm, and the last inequality is true asymptotically as $k \to \infty$ since ($\|A^k\| \leq (\rho(A) + o(1))^k$). To find the convergence rate we need maximum eigenvalue of $-\boldsymbol{M}$ (minimum eigenvalue of $\boldsymbol{M}$) which corresponds to the largest spectral radius of ($-\boldsymbol{M}$). Matrix $\boldsymbol{M}$ is

$$\begin{bmatrix} 0 & & & -1 & & \\ & \ddots & & & \ddots & \\ & & 0 & & & -1 \\ (np+mq)a_{11} & & & (n+q)+ma_{11} & & \\ & \ddots & & & \ddots & \\ & & (np+mq)a_{dd} & & & (n+q)+ma_{dd} \end{bmatrix},$$

which after permuting its rows and columns becomes

$$\begin{bmatrix} 0 & -1 & \cdots & 0 & 0 \\ (np+mq)a_{11} & (n+q)+ma_{11} & & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ 0 \quad 0 & & & 0 & -1 \\ 0 \quad 0 & & \cdots & (np+mq)a_{dd} & (n+q)+ma_{dd} \end{bmatrix},$$

such that $a_{11} \geq a_{22} \geq \ldots \geq a_{dd}$. Due to the μ-strong convexity of f, we have $a_{dd} = \mu$. Next, the eigenvalues of each 2×2 matrix in the block matrix $\boldsymbol{M}$ will lead to the eigenvalues of the whole matrix. The matrix $\boldsymbol{M}$ has d blocks and each block has 2 eigenvalues. The eigenvalues of i-th block are noted with $\lambda^i_{1,2}$ for $i \in \{1, .., d\}$. This will lead to

$$\lambda^i_{1,2} = \tfrac{1}{2}\left(ma_{ii} + (n+q) \pm \sqrt{((ma_{ii}) + (n+q))^2 - 4a_{ii}(np+mq)}\right) \qquad \forall i \in \{1, .., d\}. \tag{17}$$

Now, taking $n = q, p = \frac{q}{a_{ii}} + \frac{(ma_{ii})^2}{4qa_{ii}}$ results in the critical damping setting i.e.

$$\sqrt{((ma_{ii}) + (n+q))^2 - 4a_{ii}(np+mq)} = 0 \qquad \forall i \in \{1, .., d\}.$$

Note that under this setting, all the eigenvalues are real and nonnegative. Since $m \geq 0$, choosing $a_{dd} = \mu$ will lead to the smallest eigenvalue (slowest one in convergence) which is

$$\lambda^d_{1,2} = \tfrac{1}{2}\left(m\mu + (2q)\right) = \tfrac{m\mu}{2} + q. \tag{18}$$

The analysis above gives us

$$\|Z_t\|_2 \leq e^{-(\frac{m\mu}{2}+q)t}\|Z_0\|_2, \tag{19}$$

and

$$f(X_t) \leq \tfrac{\|A\|_{2\to 2}}{2}\|X_t\|_2^2 \leq \tfrac{\|A\|_{2\to 2}}{2}\|Z_t\|_2^2 \leq \tfrac{\|A\|_{2\to 2}}{2}e^{-2(\frac{m\mu}{2}+q)t}\|Z_0\|_2^2.$$

In particular, note that increasing q and m can lead to arbitrary fast convergence with the rate of $e^{-(m\mu+2q)t}$ under the conditions mentioned.

Remark A.0.1 (Comparison with quadratics)**.** *The rates found for μ-strongly convex functions are compared with their quadratic counterparts. Specifically, for the quadratic function $f(X_t) = \frac{1}{2}X_t^T\boldsymbol{A}X_t$ we showed $|f(X_t)| \leq C_q e^{-(m\mu+2q)t}$. Deeper analysis on the auxiliary variable $Z_t = [X_t^T \quad V_t^T]^T$ shows*

$$\|Z_t\|^2 = \|X_t\|^2 + \|V_t\|^2 = X_t^T\boldsymbol{I}_d X_t + \|\frac{\dot{X}_t}{n} + X_t + \frac{m}{n}\boldsymbol{A}X_t\|^2,$$

where the second equality is due to (GM^2-ODE) and the Lyapunov function for $f(X_t) = \frac{1}{2}X_t^T\boldsymbol{A}X_t$ is

$$\varepsilon(t) = \frac{1}{2}X_t^T\boldsymbol{A}X_t + \frac{n}{2p}\|\frac{\dot{X}_t}{n} + X_t + \frac{m}{n}\boldsymbol{A}X_t\|^2 = \frac{1}{2}X_t^T\boldsymbol{A}X_t + \frac{n}{2p}\|V_t\|^2 + \leq \max\{\frac{\|\boldsymbol{A}\|_{2\to 2}}{2}, \frac{n}{2p}\}\|Z_t\|^2$$

$$\overset{(19)}{\leq} C_q \max\{\frac{\|\boldsymbol{A}\|_{2\to 2}}{2}, \frac{n}{2p}\}e^{-(m\mu+2q)t}\|Z_0\|^2$$

which is twice faster than the rate found in Theorem 3.2. One can notice the existence of coefficient 2 instead of 1 for q and m in the convergence rate of convex quadratics.

A.2 DISCRETE-TIME ANALYSIS OF THE QUADRATICS

We consider discretizing (GM^2-ODE) and investigate the convergence behaviour of it for μ-strongly convex L-smooth quadratic function of the form $f(X) = \frac{1}{2}X^T\boldsymbol{A}X$. Applying the SIE discretization on (GM^2-ODE) we get

$$\begin{cases} x_{k+1} - x_k = & -m\sqrt{s}\boldsymbol{A}x_k - n\sqrt{s}(x_{k+1} - v_k), \\ v_{k+1} - v_k = & -p\sqrt{s}\boldsymbol{A}x_{k+1} - q\sqrt{s}(v_k - x_{k+1}). \end{cases} \tag{20}$$

Without loss of generality we can assume that $\boldsymbol{A}$ is a diagonal matrix in which case the diagonal elements of $\boldsymbol{A}$ are its eigenvalues. The one line representation of (20) is

$$x_{k+1}^i = \left(1 + \frac{(1 - q\sqrt{s} - npsa_{ii} - m\sqrt{s}a_{ii})}{1 + n\sqrt{s}}\right)x_k^i + \left(\frac{m\sqrt{s}a_{ii} - 1 + q\sqrt{s}(1 - m\sqrt{s}a_{ii})}{1 + n\sqrt{s}}\right)x_{k-1}^i, \tag{21}$$

where upper index i denotes the i'th element and a_{ii} is the i'th element of $\boldsymbol{A}$'s diagonal elements for $i = 1, \ldots, d$. For comparison, the one line representation of the NAG algorithm for quadratic function $f(X)$ is

$$x_{k+1}^i = (\frac{2}{1+\sqrt{\mu s}}(1 - sa_{ii}))x_k^i + \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}(sa_{ii} - 1)x_{k-1}^i \qquad \forall i \in \{1, \ldots d\},$$

which can be derived from (21) by setting

$$n = \sqrt{\mu}, q = \sqrt{\mu}, p = \frac{1}{\sqrt{\mu}}, m = \sqrt{s}.$$

To study the convergence rate of (20), we reformulate (21) as

$$\begin{aligned} y_k = \begin{bmatrix} x_{k+1} \\ x_k \end{bmatrix} &= \begin{bmatrix} \left((1 + \frac{(1-q\sqrt{s})}{1+n\sqrt{s}})\boldsymbol{I}_d - \frac{(nps+m\sqrt{s})}{1+n\sqrt{s}}\boldsymbol{A}\right) & \left(\frac{(1-q\sqrt{s})(\boldsymbol{A}m\sqrt{s}-\boldsymbol{I}_d)}{1+n\sqrt{s}}\right) \\ \boldsymbol{I}_d & \boldsymbol{0}_d \end{bmatrix} \begin{bmatrix} x_k \\ x_{k-1} \end{bmatrix} \\ &= \boldsymbol{T}y_{k-1} \end{aligned} \tag{22}$$

with $\mathbf{0}_d$ and $\boldsymbol{I}_d$ as a $d \times d$ zero matrix and d-dimension identity matrix. Next, we have

$$\|y_k\|_2 = \|\boldsymbol{T}y_{k-1}\|_2 = \|\boldsymbol{T}^k y_0\|_2 \leq \|\boldsymbol{T}^k\|_2 \|y_0\|_2 \leq (\rho(\boldsymbol{T}) + o(1))^k \|y_0\|_2,$$

where $\rho(\boldsymbol{T})$ is the spectral radius of $\boldsymbol{T}$ and the last inequality is true asymptotically as $k \to \infty$ through Gelfand's formula (Horn & Johnson, 2012). To define convergence rate we need the largest eigenvalue of $\boldsymbol{T}$ which corresponds to the largest spectral radius. Note that $\boldsymbol{T}$ is the block diagonal matrix

$$\boldsymbol{T} = \begin{bmatrix} \boldsymbol{T}_1 & 0 & \dots & 0 \\ 0 & \boldsymbol{T}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \boldsymbol{T}_d \end{bmatrix}, \quad \boldsymbol{T}_i = \begin{bmatrix} \left((1 + \frac{(1-q\sqrt{s})}{1+n\sqrt{s}}) - \frac{(nps+m\sqrt{s})}{1+n\sqrt{s}} a_{ii}\right) & \left(\frac{(1-q\sqrt{s})(a_{ii}m\sqrt{s}-1)}{1+n\sqrt{s}}\right) \\ 1 & 0 \end{bmatrix}, \tag{23}$$

for $i \in \{1, \dots, d\}$. Hence, the eigenvalues of $\boldsymbol{T}$ are the union of the eigenvalues of $\boldsymbol{T}_i$'s. For each $\boldsymbol{T}_i$ there exist two eigenvalues as the solutions of

$$r^2 - \left(1 + \frac{1-q\sqrt{s}}{1+n\sqrt{s}} - \frac{(nps + m\sqrt{s}a_{ii})}{1+n\sqrt{s}}\right) r - \frac{(1-q\sqrt{s})(a_{ii}m\sqrt{s}-1)}{1+n\sqrt{s}} = 0,$$

with

$$\Delta = \left(1 + \frac{1-q\sqrt{s}}{1+n\sqrt{s}} - \frac{(nps + m\sqrt{s}a_{ii})}{1+n\sqrt{s}}\right)^2 - 4\frac{(1-q\sqrt{s})(a_{ii}m\sqrt{s}-1)}{1+n\sqrt{s}}.$$

Taking $n = q = \sqrt{a_{ii}}, np = 1, m = \sqrt{s}$ leads to $\Delta = 0$ and $r_{1,2} = 1 - \sqrt{sa_{ii}}$. With this choice of parameters, the convergence rate of (20) for μ-strongly convex and L-Lipschitz quadratics of the form $f(X) = \frac{1}{2}X^T \boldsymbol{A} X$ will be

$$f(x_k) - f(x^*) = \frac{1}{2}\sum_{i=1}^{d} a_{ii}(x_k^i)^2 \leq \max_i a_{ii}\|x_k\|^2 \leq C \max_i (1 - \sqrt{sa_{ii}})^{2k},$$

for $\sqrt{s} \leq \frac{1}{\sqrt{L}}$ (due to $1 - \sqrt{sa_{ii}} \geq 0$). The worst case scenario happens for $a_{ii} = \mu$ (closest possible rate to 1) which leads to the rate $O((1 - \sqrt{\frac{1}{\kappa}})^{2k})$. The $np = 1$ condition does not exist in the continuous case. This observation is used for our analysis in the general case.

A.3 CONVERGENCE OF OPTIMIZATION ALGORITHMS THROUGH DYNAMICAL SYSTEMS

In the state space, dynamical systems are usually presented in the form of

$$\dot{\xi}(t) = A\xi(t) + Bu(t), \quad y(t) = C\xi(t), \quad u(t) = \nabla f(y(t)) \quad \forall t \geq 0, \tag{24}$$

where $\xi \in \mathbb{R}^n$ is the state, $y(t) \in \mathbb{R}^d (d \leq n)$ is the output, and $u(t)$ is the continuous feedback input. Here, we would have $u^* = 0$ and the fixed point of (24) is

$$A\xi^* = 0, \quad y^* = C\xi^*.$$

Consider the nonnegative function

$$\varepsilon(t) = e^{\lambda t}\left(f(y(t)) - f(y^*) + (\xi(t) - \xi^*)^T P(\xi(t) - \xi^*)\right),$$

with $\lambda > 0$, $y^* = x^*$ and $P \succeq 0$ where $A \succeq B$ denotes that $A - B$ is positive semi-definite. If when $\xi \to \xi^*$ we have $\frac{d}{dt}\varepsilon(t) \leq 0$, then $\varepsilon(t) \leq \varepsilon(0)$. This results in

$$f(y(t)) - f(y^*) \leq e^{-\lambda t}\varepsilon(0).$$

The following result from (Fazlyab et al., 2018) proposes a Linear Matrix Inequality (LMI) that guarantees the existence of a Lyapunov function through which we can show that $f(x)$ converges exponentially fast. For simplicity, we adopt the presentation of (Sanz Serna & Zygalakis, 2021).

Theorem A.1 (Theorem 6.4 in (Fazlyab et al., 2018)). *Suppose that for (24) there exists* $\lambda > 0, P \succeq 0$ *and* $\sigma \geq 0$ *such that* $\boldsymbol{T} = M^{(0)} + M^{(1)} + \lambda M^{(2)} + \sigma M^{(3)} \preceq 0$ *where*

$$M^{(0)} = \begin{bmatrix} PA + A^TP + \lambda P & PB \\ B^TP & 0 \end{bmatrix},$$

$$M^{(1)} = \frac{1}{2}\begin{bmatrix} 0 & (CA)^T \\ CA & CB + (CB)^T \end{bmatrix},$$

$$M^{(2)} = \begin{bmatrix} C^T & 0 \\ 0 & I_d \end{bmatrix}\begin{bmatrix} -\frac{\mu}{2}I_d & \frac{1}{2}I_d \\ \frac{1}{2}I_d & 0 \end{bmatrix}\begin{bmatrix} C & 0 \\ 0 & I_d \end{bmatrix},$$

$$M^{(3)} = \begin{bmatrix} C^T & 0 \\ 0 & I_d \end{bmatrix}\begin{bmatrix} -\frac{\mu L}{\mu+L}I_d & \frac{1}{2}I_d \\ \frac{1}{2}I_d & \frac{-1}{\mu+L} \end{bmatrix}\begin{bmatrix} C & 0 \\ 0 & I_d \end{bmatrix},$$

and $(.)^T$ *denotes the transpose operator and* I_d *is the identity matrix of size* d*. Then for* $f \in \mathcal{F}_{\mu,L}$ *we have*

$$f(y(t)) - f(y^*) \leq e^{-\lambda t}\varepsilon(0).$$

A.4 ACCELERATION OF THE EXPLICIT EULER DISCRETIZATION

We would like to show the correspondence between EE and SIE discretizations of (GM^2-ODE). The following lemma shows how to update the coefficients of EE method such that SIE is derived.

Lemma 1. *Consider the parameters of EE discretization of (GM^2-ODE) as* n_{EE}, m_{EE}, q, p *and the parameters of SIE discretization of (GM^2-ODE) as* n_{SIE}, m_{SIE}, q, p*. Then by taking*

$$\begin{cases} n_{EE} = \frac{n_{SIE} - qn_{SIE}\sqrt{s}}{1+n_{SIE}\sqrt{s}}, \\ m_{EE} = \frac{m_{SIE} + n_{SIE}p\sqrt{s}}{1+n_{SIE}\sqrt{s}}, \end{cases} \tag{25}$$

SIE discretization of (GM^2-ODE) will be the same as its EE discretization with step-size $\sqrt{s}$*.*

For proving the result, note that the EE discretization of (GM^2-ODE) is

$$\begin{cases} x_{k+1} - x_k = & -m\sqrt{s}\nabla f(x_k) - n\sqrt{s}(x_k - v_k), \\ v_{k+1} - v_k = & -p\sqrt{s}\nabla f(x_{k_k}) - \sqrt{s}q(v_k - x_k), \end{cases} \tag{26}$$

which can be written in one line format

$$x_{k+1} = x_k - m\sqrt{s}\nabla f(x_k) + (1 - q\sqrt{s} - n\sqrt{s})(x_k - x_{k-1}) + (m\sqrt{s}(1 - q\sqrt{s}) - nps)\nabla f(x_{k-1}), \tag{27}$$

replacing the coefficient updates from (25) in above gives the SIE one line update of (6) which is

$$x_{k+1} = x_k - \frac{m\sqrt{s}+nps}{1+n\sqrt{s}}\nabla f(x_k) + \frac{1-q\sqrt{s}}{1+n\sqrt{s}}(x_k - x_{k-1}) + \frac{m\sqrt{s}(1-q\sqrt{s})}{1+n\sqrt{s}}\nabla f(x_{k-1}), \tag{28}$$

With Lemma 1 and Theorem 4.1, we establish the convergence result for (26) as follows.

Corollary A.1.1 (Convergence of (26)). *For* μ*-strongly convex* L*-smooth function* f *with* $0 \leq \mu < L$ *and parameters* m, n, p, q *such that*

$$q/p \leq \mu, 0 \leq qps \leq m\sqrt{s}(1 + q\sqrt{s}) - qps \leq \frac{1}{L}, n = q\frac{1 - q\sqrt{s}}{1 + q\sqrt{s}}, 0 \leq q\sqrt{s} < 1, p > 0,$$

the sequence x_k *in (26) will satisfy*

$$f(x_k) - f(x^*) \leq LC'_{GM}(1 - q\sqrt{s})^k,$$

for constant $C'_{GM} > 0$ *and any* $x_0, v_0 = x_0 - (\frac{m}{n} - \frac{2qps}{n\sqrt{s}(1+q\sqrt{s})})\nabla f(x_0)$*.*

The proof is simply done by using (25) in Theorem 4.1. Note that for the initial condition we need to get the same result as in Theorem 4.1 with the new coefficients and the new update rule (26).

Corollary A.1.1 suggests that choosing

$$q = \sqrt{\mu}, p = \frac{1}{\sqrt{\mu}}, m = \frac{2s}{1 + \sqrt{\mu s}}, n = \sqrt{\mu}\frac{1 - \sqrt{\mu s}}{1 + \sqrt{\mu s}},$$

in (26) will recover the NAG algorithm.

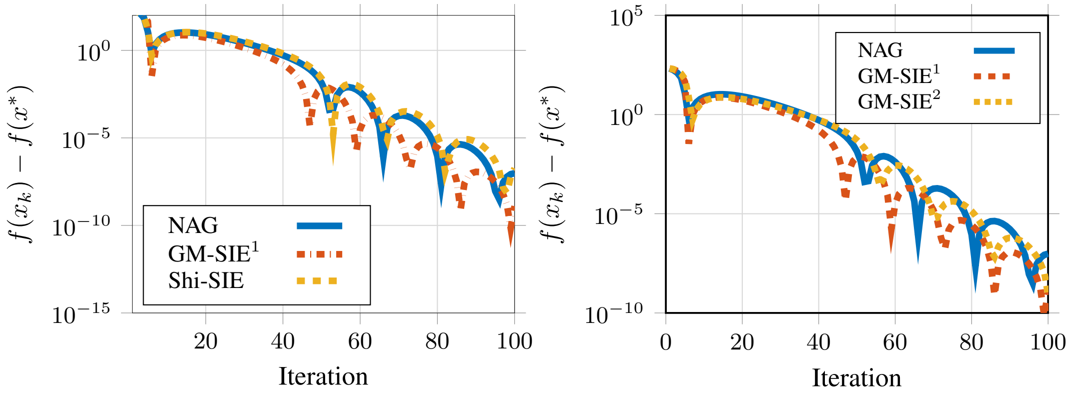

Figure 3: Comparison between the NAG, (Shi-SIE) as Shi-SIE, SIE discretization of (GM-ODE) as GM-SIE with superscript 1 and 2 when $n' = 1, m' = \sqrt{s}, q' = 2\sqrt{\mu}$ and $n' = 1 - 2\sqrt{\mu s}, m' = \sqrt{s}, q' = 2\sqrt{\mu}$ respectively. The simulation function was $f(x) = 4(L-\mu)\log(1+e^{-x}) + \frac{\mu}{2}x^2$ with $L = 1, \mu = 0.01$. (a) The effect of the approximations $(1/(1-\sqrt{\mu s}) \approx 1$ in Shi-SIE and coefficient deviation in GM-SIE^1) in the ODE trajectories, (b) different coefficients used for discretizing (GM-ODE). GM-SIE^1 is the SIE discretization of GM-ODE^1 (the recovered high-resolution NAG ODE from (GM-ODE)) and GM-SIE^2 is the SIE discretization of GM-ODE^2 (the ODE used to recover the NAG algorithm).

A.5 NUMERICAL RESULTS

In this section, numerical experiments are designed for further illustration of the previous findings. An important note is that (GM-ODE) in (Zhang et al., 2021) uses different parameters to recover (NAG-ODE) ($m' = \sqrt{s}, q' = 2\sqrt{\mu}, n' = 1$) and the NAG algorithm after discretization ($m' = \sqrt{s}, q' = 2\sqrt{\mu}, n' = 1 - 2\sqrt{\mu s}$). In Figure 3 we have considered two SIE discretizations of (GM-ODE) and

$$\begin{cases} q_{k+1} - q_k = & j_k\sqrt{s}, \\ j_{k+1} - j_k = & -2\sqrt{\mu s}j_{k+1} - \sqrt{s}(1+\sqrt{\mu s})\nabla f(q_{k+1}) - \sqrt{s}(\nabla f(q_{k+1}) - \nabla f(q_k)), \end{cases} \tag{Shi-SIE}$$

which is the SIE discretization of (11) used in (Shi et al., 2021). The discretizations of (GM-ODE) are shown with GM-SIE^1 and GM-SIE^2. The aim of Figure 3 is to highlight two things; First, the effect of coefficient inconsistency before and after discretization of (GM-ODE) and second, to depict the approximation $1/(1-\sqrt{\mu s}) \approx 1$ made in (Shi et al., 2021). The step-size was $s = 1/L$ in all simulations. All algorithms are simulated with the parameters they use to recover (NAG-ODE) except for GM-SIE^2 which uses $n' = 1 - 2\sqrt{\mu s}$ for the sake of comparison with GM-SIE^1. GM-SIE^2 does not fall exactly on the NAG algorithm due to different initializations. We did not simulate (6) due to its exact match with the NAG method.

Table 2: Parameters used for comparing different algorithms in Figure 4.

Nesterov	(13)	Shi-SIE	GM-SIE^2
$s = \frac{1}{L}$	$s = \frac{1}{L}$	$s = \frac{4}{9L}$	$s = \frac{1}{4L}$
$\alpha = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$	$m = \sqrt{s}$	-	$m' = \sqrt{s}$
-	$q = n = \sqrt{\mu}$	-	$q' = 2\sqrt{\mu}$
-	$p = 1/\sqrt{\mu}$	-	$n' = 1 - 2\sqrt{\mu s}$
$x_0 = y_0 \sim N(0,1)$	$x_0, v_0 \sim N(0,1)$	$x_0 \sim N(0,1), v_0 = \frac{-2\nabla f(x_0)}{1+\sqrt{\mu s}}$	$x_0 = v_0 \sim N(0,1)$

Next, we provide an example of the performance of the NAG method (as Nesterov) with (13), GM-SIE^2, and Shi-SIE under the conditions they prove their convergence results (see table 2). For

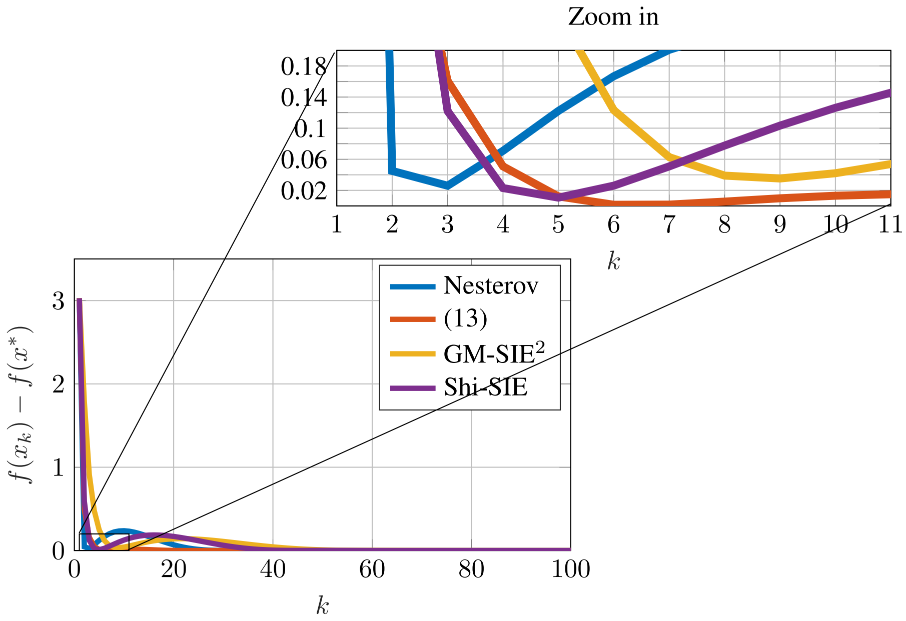

Figure 4: Comparison between the NAG, (13), GM-SIE^2, and Shi-SIE. All algorithms are simulated under the best performance conditions (see table 2). The simulation function was $f(x) = 4(L - \mu)\log(1 + e^{-x}) + \frac{\mu}{2}x^2$ with $L = 1, \mu = 0.01$.

the sake of visualization, we simulated (13) with random v_0 with 100 monte-carlo simulations instead of $v_0 = x_0 - \frac{m}{n}\nabla f(x_0)$. The result is shown in Figure 4.

B PROOFS

B.1 PROOF OF PROPOSITION 5.1

By comparing the one-line presentation of (GM-ODE)

$$\ddot{U}_t + (q' + m'\nabla^2 f(U_t))\dot{U}_t + (n' + m'q')\nabla f(U_t) = 0, \tag{29}$$

and the one-line presentation of (GM^2-ODE)

$$\ddot{X} + ((n + q) + m\nabla^2 f(X))\dot{X} + (np + mq)\nabla f(X) = 0, \tag{30}$$

we can see that if the parameters are chosen as in (10), we get the equivalence.

B.2 PROOF OF THEOREM 3.1

Here, Theorem A.1 is used to find both a Lyapunov function and a convergence rate for (GM^2-ODE). In order to apply the theorem we need to present (GM^2-ODE) in the form of (24). Taking $X, V \in \mathbb{R}^d$,

$$\xi = [X, V]^T, \quad A = \begin{bmatrix} -nI_d & nI_d \\ qI_d & -qI_d \end{bmatrix}, \quad B = \begin{bmatrix} -mI_d \\ -pI_d \end{bmatrix}, \quad C = \begin{bmatrix} I_d \\ 0_d \end{bmatrix}^T, \tag{31}$$

will present the state space of (GM^2-ODE). For simplicity, we set $\sigma = 0$, i.e. we remove any sign of L from our formulation. Thus, the result holds for μ-strongly convex functions. We need to find $P \succeq 0, \lambda > 0$ such that $T \preceq 0$. Consider

$$P = \hat{P} \otimes I_d, \qquad \hat{P} = \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix}, \tag{32}$$

$$T = \hat{T} \otimes I_d, \qquad \hat{T} = \begin{bmatrix} t_{11} & t_{12} & t_{13} \\ t_{12} & t_{22} & t_{23} \\ t_{13} & t_{23} & t_{33} \end{bmatrix}. \tag{33}$$

Using the structure in Theorem A.1, one can find the elements t_{ij} as

$$\begin{aligned} t_{11} &= -(2n-\lambda)p_{11} + 2qp_{12} - \frac{\lambda\mu}{2}, \\ t_{12} &= np_{11} + qp_{22} - (n+q-\lambda)p_{12}, \\ t_{13} &= -mp_{11} - pp_{12} + \frac{\lambda-n}{2}, \\ t_{22} &= 2(np_{12} - qp_{22}) + \lambda p_{22}, \\ t_{23} &= -mp_{12} - pp_{22} + \frac{n}{2}, \\ t_{33} &= -m, \end{aligned}$$

Next step is to ensure that $\hat{P} \succeq 0$. If we take $\det(\hat{P}) = 0$ and find one of the diagonal elements such that it would be positive, then one of the eigenvalues of $\hat{P}$ is zero and the other one is positive which results in our favor. Before doing so, we will find p_{11} and p_{22} as a function of p_{12}. The latter is done by setting $t_{13} = 0$ and $t_{23} = 0$. Then

$$\begin{aligned} t_{13} = 0 \quad &\Longrightarrow \quad p_{11} = \frac{-pp_{12} + \frac{\lambda-n}{2}}{m}, \\ t_{23} = 0 \quad &\Longrightarrow \quad p_{22} = \frac{-mp_{12} + \frac{n}{2}}{p}. \end{aligned} \tag{34}$$

These choices will lead to a block diagonal $\hat{T}$ which is easier to handle later. Now, we will find p_{12} such that $\det(\hat{P}) = 0$.

$$\det(\hat{P}) = 0 \rightarrow p_{11}p_{22} - p_{12}^2 = 0 \rightarrow p_{12} = \frac{n}{4}\left(\frac{n-\lambda}{\frac{m(n-\lambda)-np}{2}}\right) \tag{35}$$

From the quadratics analysis, we expect the fastest convergence rate to relate to q with condition $n = q$. Therefore, we set $\lambda = q$ and $n = q$. These two conditions lead to

$$p_{12} = 0, \quad p_{11} = 0, \quad p_{22} = \frac{n}{2p}, \tag{36}$$

and since $\frac{n}{2p} \geq 0$ we get $\hat{P} \succeq 0$. Also, we have

$$t_{11} = -\frac{q\mu}{2}, \quad t_{12} = \frac{q^2}{2p}, \quad t_{13} = 0, \quad t_{22} = \frac{-q^2}{2p}, \quad t_{23} = 0, \quad t_{33} = -m. \tag{37}$$

Now, to establish $\hat{T} \preceq 0$, consider

$$\hat{T}_1 = \begin{bmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{bmatrix}, \quad \hat{T}_2 = [t_{33}],$$

as the blocks in the block diagonal matrix $\hat{T}$. We know that $t_{33} \leq 0$. Also if $\mathrm{Tr}(.)$ denotes the trace operator, $\mathrm{Tr}(A)$ is equal to the sum of the eigenvalues of the matrix A and the determinant of A is equal to the multiplication of its eigenvalues. Therefore, if the determinant of the first 2×2 block matrix, $\hat{T}_1$, is positive and t_{11}, t_{22} are negative, we ensure that $\mathrm{Tr}(\hat{T}_1) = \Gamma_1 + \Gamma_2 \leq 0$, $\Gamma_1\Gamma_2 \geq 0$, and $\Gamma_3 \leq 0$ where $\Gamma_i, i = 1, 2, 3$ are the eigenvalues of $\hat{T}$. Hence, $\Gamma_i \leq 0$ for $i = 1, 2, 3$ and this means $\hat{T} \preceq 0$. One can formulate the above arguments as

$$t_{11}t_{22} - t_{12}^2 \geq 0 \rightarrow (\frac{q\mu}{2})(\frac{q^2}{2p}) - (\frac{q^2}{2p})^2 \geq 0 \rightarrow q \leq \mu p,$$

which indeed holds for the quadratic case as well.

According to Theorem A.1 we have

$$f(x(t)) - f(x^*) \leq e^{-qt}(f(x(0)) - f(x^*) + (\xi(0) - \xi^*)^T P(\xi(0) - \xi^*)),$$

which is

$$f(x(t)) - f(x^*) \leq e^{-qt}(f(x(0)) - f(x^*) + \frac{q}{2p}\|V - x^*\|^2). \tag{38}$$

Note that due to the form of (GM^2-ODE), $x^* = v^*$ (see Figure 1). Therefore, the claim of the theorem is proved.

B.3 PROOF OF THEOREM 3.2

Recall the form of Lyapunov function as

$$\varepsilon(t) = f(X_t) - f(x^*) + \frac{n}{2p}\|V_t - x^*\|^2.$$

In order to prove, we will take derivative of $e^{\lambda t}\varepsilon(t)$ with respect to time and show the result is negative.

$$\frac{de^{\lambda t}\varepsilon(t)}{dt} = \lambda e^{\lambda t}[f(X_t) - f(x^*) + \frac{n}{2p}\|V_t - x^*\|^2] \tag*{\textbf{(I)}}$$

$$+ e^{\lambda t}[\langle \dot{X}_t, \nabla f(X_t)\rangle + \frac{n}{p}\underbrace{\langle \frac{\ddot{X}_t}{n} + \dot{X}_t + \frac{m}{n}\nabla^2 f(X_t)\dot{X}_t}_{\textbf{(A)}}, \frac{\dot{X}_t}{n} + X_t - x^* + \frac{m}{n}\nabla f(X_t)\rangle]. \tag*{\textbf{(II)}}$$

Using (30) we have

$$\begin{aligned}\textbf{(A)} = \frac{\ddot{X}_t}{n} + \dot{X}_t + \frac{m}{n}\nabla^2 f(X_t)\dot{X}_t &= -\dot{X}_t - \frac{q}{n}\dot{X}_t - \frac{m}{n}\nabla^2 f(X_t)\dot{X}_t - (p + \frac{mq}{n})\nabla f(X_t) \\ + \dot{X}_t + \frac{m}{n}\nabla^2 f(X_t)\dot{X}_t &= -\frac{q}{n}\dot{X}_t - (p + \frac{mq}{n}\nabla f(X_t)),\end{aligned} \tag{39}$$

Replacing (39) in **(II)** gives

$$\begin{aligned}&\langle \dot{X}_t, \nabla f(X_t)\rangle + \frac{n}{p}\langle \frac{\ddot{X}_t}{n} + \dot{X}_t + \frac{m}{n}\nabla^2 f(X_t)\dot{X}_t, \frac{\dot{X}_t}{n} + X_t - x^* + \frac{m}{n}\nabla f(X_t)\rangle \\ &= \langle \dot{X}_t, \nabla f(X_t)\rangle + \frac{n}{p}\langle -\frac{q}{n}\dot{X}_t - (p + \frac{mq}{n}\nabla f(X_t)), \frac{\dot{X}_t}{n} + X_t - x^* + \frac{m}{n}\nabla f(X_t)\rangle \\ &= \langle \dot{X}_t, \nabla f(X_t)\rangle + \frac{n}{p}\Big[-\frac{qm}{n^2}\langle \dot{X}_t, \nabla f(X_t)\rangle - (\frac{p}{n} + \frac{mq}{n^2})\langle \dot{X}_t, \nabla f(X_t)\rangle - \frac{q}{n^2}\|\dot{X}_t\|^2 - \frac{q}{n}\langle \dot{X}_t, X_t - x^*\rangle \\ &-(p + \frac{mq}{n})\langle \nabla f(X_t), X_t - x^*\rangle - \frac{m}{n}(p + \frac{mq}{n})\|\nabla f(X_t)\|^2\Big] \\ &= -2\frac{qm}{np}\langle \dot{X}_t, \nabla f(X_t)\rangle - \frac{q}{np}\|\dot{X}_t\|^2 - \frac{qn}{p}\langle \frac{\dot{X}_t}{n}, X_t - x^*\rangle - (n + \frac{mq}{p})\langle \nabla f(X_t), X_t - x^*\rangle \\ &- m(1 + \frac{mq}{np})\|\nabla f(X_t)\|^2.\end{aligned} \tag{40}$$

Now, **(I)** gives

$$\begin{aligned}&\lambda e^{\lambda t}\Big[f(X_t) - f(x^*) + \frac{n}{2p}\|V_t - x^*\|^2\Big] = \lambda e^{\lambda t}\Big[f(X_t) - f(x^*) \\ &+ \frac{n}{2p}\Big(\|\frac{\dot{X}_t}{n}\|^2 + \|X_t - x^*\|^2 + 2\langle \frac{\dot{X}_t}{n}, X_t - x^*\rangle + \frac{m^2}{n^2}\|\nabla f(X_t)\|^2 + 2\langle \frac{\dot{X}_t}{n}, \frac{m}{n}\nabla f(X_t)\rangle \\ &+ 2\frac{m}{n}\langle \nabla f(X_t), X_t - x^*\rangle\Big)\Big].\end{aligned} \tag{41}$$

Now, using (41) and (equation 40) in **(I)** and **(II)** respectively, gives

$$\begin{aligned}\frac{de^{\lambda t}\varepsilon(t)}{dt} &= e^{\lambda t}[\lambda(f(X_t) - f(x^*)) \\ &\quad + (\frac{m\lambda - 2mq}{np})\langle \dot{X}_t, \nabla f(X_t)\rangle + (\frac{\lambda m}{p} - n - \frac{mq}{p})\langle \nabla f(X_t), X_t - x^*\rangle + (\frac{\lambda n}{2p} - \frac{qn}{p})\|\frac{\dot{X}_t}{n}\|^2 \\ &\quad + (\frac{\lambda n}{p} - \frac{qn}{p})\langle \frac{\dot{X}_t}{n}, X_t - x^*\rangle + (\frac{\lambda m^2}{2np} - m - \frac{qm^2}{np})\|\nabla f(X_t)\|^2 + \frac{\lambda n}{2p}\|X_t - x^*\|^2].\end{aligned} \tag{42}$$

Due to strong convexity of $f(X_t)$ we have

$$\langle \nabla f(X_t), x^* - X\rangle \leq -(f(X_t) - f(x^*) + \frac{\mu}{2}\|X_t - x^*\|^2).$$

Using the above inequality in (42) we get

$$\frac{de^{\lambda t}\varepsilon(t)}{dt} \leq e^{\lambda t}[\left(\lambda - (n + \frac{mq}{p} - \frac{\lambda m}{p})\right)(f(X_t) - f(x^*)) + \mu(\frac{\lambda m}{2p} - \frac{mq}{2p})\|X_t - x^*\|^2 \tag{III}$$

$$+ \left(\frac{\lambda n}{2p} - \frac{\mu n}{2}\right)\|X_t - x^*\|^2 + \left(\frac{\lambda n - qn}{2p}\right)\left(2\langle \frac{\dot{X}_t}{n}, X_t - x^*\rangle + \|\frac{\dot{X}_t}{n}\|^2\right) \tag{IV}$$

$$- \frac{q}{2np}\|\dot{X}_t\|^2 + 2(\frac{\lambda}{2np} - \frac{q}{np})\langle \dot{X}_t, m\nabla f(X_t)\rangle + \left(\frac{\lambda}{2np} - \frac{1}{m} - \frac{q}{np}\right)\|m\nabla f(X_t)\|^2. \tag{V}$$

Now, we need to find conditions such that (III), (IV) and (V) are negative. For (III) to be negative we need that

$$\lambda - (n + \frac{mq}{p} - \frac{\lambda m}{p}) \leq 0 \text{ and } (\frac{m}{2p}(\lambda - q)) \leq 0,$$

which are satisfied as long as

$$\lambda - (n + \frac{mq}{p} - \frac{\lambda m}{p}) \leq (1 + \frac{m}{p})(\lambda - \min\{n, q\}) \to \lambda \leq \min\{n, q\}.$$

Next, we can upper bound (IV) with a negative term with coefficient if

$$\frac{q}{p} \leq \mu \text{ and } \lambda \leq q.$$

To see this, if we have $\lambda \leq q$ and $\frac{q}{p} \leq \mu$ then $\frac{\lambda}{p} \leq \mu$ and therefore,

$$\left(\frac{\lambda n}{2p} - \frac{\mu n}{2}\right)\|X_t - x^*\|^2 \leq -\left(\frac{qn - \lambda n}{2p}\right)\|X_t - x^*\|^2.$$

Replacing in (IV) we get

$$\text{(IV)} \leq -\left(\frac{qn - \lambda n}{2p}\right)\|\frac{\dot{X}_t}{n} + X_t - x^*\|^2 \leq 0.$$

Lastly, one needs to have

$$2q \geq \lambda \geq q \text{ and } n, m, p \neq 0,$$

so that

$$\left(\frac{\lambda}{2np} - \frac{1}{m} - \frac{q}{np}\right)\|m\nabla f(X_t)\|^2 \leq (\frac{\lambda}{2np} - \frac{q}{np})\|m\nabla f(X_t)\|^2 \leq 0,$$

and

$$-\frac{q}{2np}\|\dot{X}_t\|^2 \leq (\frac{\lambda}{2np} - \frac{q}{np})\|\dot{X}_t\|^2.$$

Then, replacing in (V) results in

$$\text{(V)} \leq (\frac{\lambda - 2q}{2np})\|\dot{X}_t + m\nabla f(X_t)\|^2 \leq 0.$$

Putting all the above conditions together we conclude that

$$\frac{de^{\lambda t}\varepsilon(t)}{dt} \leq 0,$$

if $n, m, p \neq 0$ and $q = \lambda$ and $n \geq q$ and $q/p \leq \mu$ Therefore, $e^{qt}\varepsilon(t) \leq \varepsilon(0)$ for $t \geq 0$ which concludes the proof.

B.4 PROOF OF THEOREM 4.1

To show the claim of the theorem, we will bound the difference $\varepsilon(k+1) - \varepsilon(k)$ such that

$$\varepsilon(k+1) \leq (1 - q\sqrt{s})^{(k+1)}\varepsilon(0),$$

holds with $\varepsilon(k)$ as the Lyapunov function

$$\varepsilon(k) = f(x_k) - f(x^*) + \frac{B}{2}\|v_k - x^*\|_2^2 - \frac{Bp^2 s}{2}\|\nabla f(x_k)\|^2, \tag{43}$$

with B as a positive constant to be found. Using (43) we have

$$\begin{aligned}\varepsilon(k+1) - \varepsilon(k) = f(x_{k+1}) - f(x_k) + \underbrace{\frac{B}{2}(\|v_{k+1} - v_k\|^2 + 2\langle v_{k+1} - v_k, v_k - x^*\rangle)}_{\textbf{(I)}} \\ - \frac{Bp^2 s}{2}\|\nabla f(x_{k+1})\|^2 + \frac{Bp^2 s}{2}\|\nabla f(x_k)\|^2. \end{aligned} \tag{44}$$

Note that for **(I)** we have

$$\begin{aligned}\textbf{(I)} \overset{6}{=} \frac{B}{2}(p^2 s\|\nabla f(x_{k+1})\|^2 + q^2 s\|v_k - x_{k+1}\|^2 + 2pqs\langle \nabla f(x_{k+1}), v_k - x_{k+1}\rangle \\ - 2p\sqrt{s}\langle \nabla f(x_{k+1}), v_k - x_{k+1}\rangle - 2p\sqrt{s}\langle \nabla f(x_{k+1}), x_{k+1} - x^*\rangle - 2q\sqrt{s}\langle v_k - x_{k+1}, v_k - x^*\rangle),\end{aligned}$$

where we have added and subtracted x_{k+1} in $\langle \nabla f(x_{k+1}), v_k - x^*\rangle$. Next, using

$$\langle a - b, a - c\rangle = \frac{1}{2}\|a - b\|^2 + \frac{1}{2}\|a - c\|^2 - \frac{1}{2}\|b - c\|^2,$$

we get

$$\begin{aligned}\textbf{(I)} = \frac{B}{2}(p^2 s\|\nabla f(x_{k+1})\|^2 + q^2 s\|v_k - x_{k+1}\|^2 + 2pqs\langle \nabla f(x_{k+1}), v_k - x_{k+1}\rangle \\ - 2p\sqrt{s}\langle \nabla f(x_{k+1}), v_k - x_{k+1}\rangle - 2p\sqrt{s}\langle \nabla f(x_{k+1}), x_{k+1} - x^*\rangle \\ - q\sqrt{s}\|v_k - x_{k+1}\|^2 - q\sqrt{s}\|v_k - x^*\|^2 + q\sqrt{s}\|x_{k+1} - x^*\|^2).\end{aligned} \tag{45}$$

Utilizing strong convexity of $f(x)$ we have

$$\langle \nabla f(x_{k+1}), x_{k+1} - x^*\rangle \geq f(x_{k+1}) - f(x^*) + \frac{\mu}{2}\|x_{k+1} - x^*\|^2, \tag{S.C}$$

thus, we can upper bound **(I)** as

$$\begin{aligned}\textbf{(I)} \overset{\textbf{(S.C)}}{\leq} \frac{B}{2}(p^2 s\|\nabla f(x_{k+1})\|^2 + q^2 s\|v_k - x_{k+1}\|^2 + 2pqs\langle \nabla f(x_{k+1}), v_k - x_{k+1}\rangle \\ - 2p\sqrt{s}\langle \nabla f(x_{k+1}), v_k - x_{k+1}\rangle - 2p\sqrt{s}(f(x_{k+1}) - f(x^*)) - \mu p\sqrt{s}\|x_{k+1} - x^*\|^2 \\ - q\sqrt{s}\|v_k - x_{k+1}\|^2 - q\sqrt{s}\|v_k - x^*\|^2 + q\sqrt{s}\|x_{k+1} - x^*\|^2) \\ \overset{\pm f(x_k)}{=} \frac{B}{2}(p^2 s\|\nabla f(x_{k+1})\|^2 + q^2 s\|v_k - x_{k+1}\|^2 + 2pqs\langle \nabla f(x_{k+1}), v_k - x_{k+1}\rangle \\ - 2p\sqrt{s}\langle \nabla f(x_{k+1}), v_k - x_{k+1}\rangle - 2p\sqrt{s}(f(x_{k+1}) - f(x_k)) - 2p\sqrt{s}(f(x_k) - f(x^*)) \\ - \mu p\sqrt{s}\|x_{k+1} - x^*\|^2 - q\sqrt{s}\|v_k - x_{k+1}\|^2 - q\sqrt{s}\|v_k - x^*\|^2 + q\sqrt{s}\|x_{k+1} - x^*\|^2).\end{aligned} \tag{46}$$

Replacing (46) in (44) we have

$$\begin{aligned}\varepsilon(k+1) - \varepsilon(k) \overset{46}{\leq} f(x_{k+1}) - f(x_k) + \frac{Bp^2 s}{2}\|\nabla f(x_{k+1})\|^2 + \frac{Bq^2 s}{2}\|v_k - x_{k+1}\|^2 \\ + Bpqs\langle \nabla f(x_{k+1}), v_k - x_{k+1}\rangle - Bp\sqrt{s}\langle \nabla f(x_{k+1}), v_k - x_{k+1}\rangle \\ - Bp\sqrt{s}(f(x_{k+1}) - f(x_k)) - Bp\sqrt{s}(f(x_k) - f(x^*)) \\ - \frac{B\mu p\sqrt{s}}{2}\|x_{k+1} - x^*\|^2 - \frac{Bq\sqrt{s}}{2}\|v_k - x_{k+1}\|^2 - \frac{Bq\sqrt{s}}{2}\|v_k - x^*\|^2 \\ + \frac{Bq\sqrt{s}}{2}\|x_{k+1} - x^*\|^2 - \frac{Bp^2 s}{2}\|\nabla f(x_{k+1})\|^2 + \frac{Bp^2 s}{2}\|\nabla f(x_k)\|^2.\end{aligned} \tag{47}$$

By setting $q/p \leq \mu$ we get

$$\frac{Bq\sqrt{s}}{2}\|x_{k+1}-x^*\|^2 \leq \frac{B\mu p\sqrt{s}}{2}\|x_{k+1}-x^*\|^2,$$

and thus simplifying (47) results in

$$\begin{aligned}\varepsilon(k+1)-\varepsilon(k) &\overset{47}{\leq} (1-Bp\sqrt{s})(f(x_{k+1})-f(x_k)) + Bp\sqrt{s}(q\sqrt{s}-1)\langle\nabla f(x_{k+1}), v_k - x_{k+1}\rangle \\ &- Bp\sqrt{s}(f(x_k)-f(x^*)) - \frac{Bq\sqrt{s}}{2}(1-q\sqrt{s})\|v_k - x_{k+1}\|^2 \\ &- \frac{Bq\sqrt{s}}{2}\|v_k - x^*\|^2 + \frac{Bp^2 s}{2}\|\nabla f(x_k)\|^2. \qquad (48)\end{aligned}$$

Next, using smoothness of $f(x)$ we get

$$\begin{aligned}f(x_{k+1})-f(x_k) &\leq \langle\nabla f(x_{k+1}), x_{k+1}-x_k\rangle - \frac{1}{2L}\|\nabla f(x_{k+1})-\nabla f(x_k)\|^2 \\ &\overset{6}{=} -m\sqrt{s}\langle\nabla f(x_{k+1}),\nabla f(x_k)\rangle - n\sqrt{s}\langle\nabla f(x_{k+1}), x_{k+1}-v_k\rangle \\ &- \frac{1}{2L}\|\nabla f(x_{k+1})-\nabla f(x_k)\|^2. \qquad (\textbf{S.L})\end{aligned}$$

Upper-bounding (48) using (**S.L**) and considering $q\sqrt{s} \leq 1$ leads to

$$\begin{aligned}\varepsilon(k+1)-\varepsilon(k) &\overset{\textbf{S.L}}{\leq} -m\sqrt{s}(1-Bp\sqrt{s})\langle\nabla f(x_{k+1}),\nabla f(x_k)\rangle - \frac{(1-Bp\sqrt{s})}{2L}\|\nabla f(x_{k+1})-\nabla f(x_k)\|^2 \\ &+ [Bp\sqrt{s}(q\sqrt{s}-1) + n\sqrt{s}(1-Bp\sqrt{s})]\langle\nabla f(x_{k+1}), v_k - x_{k+1}\rangle \\ &- Bp\sqrt{s}(f(x_k)-f(x^*)) - \frac{Bq\sqrt{s}}{2}\|v_k - x^*\|^2 + \frac{Bp^2 s}{2}\|\nabla f(x_k)\|^2. \qquad (49)\end{aligned}$$

Next, by setting $B = \frac{n}{p}$, $n = q$, and $q\sqrt{s} < 1$ we have

$$\begin{aligned}\varepsilon(k+1)-\varepsilon(k) &\overset{49}{\leq} -m\sqrt{s}(1-Bp\sqrt{s})\langle\nabla f(x_{k+1}),\nabla f(x_k)\rangle - \frac{(1-Bp\sqrt{s})}{2L}\|\nabla f(x_{k+1})-\nabla f(x_k)\|^2 \\ &- n\sqrt{s}(f(x_k)-f(x^*)) - \frac{Bq\sqrt{s}}{2}\|v_k - x^*\|^2 + \frac{Bp^2 s}{2}\|\nabla f(x_k)\|^2 \\ &= -m\sqrt{s}(1-Bp\sqrt{s})\langle\nabla f(x_{k+1}),\nabla f(x_k)\rangle - \frac{(1-Bp\sqrt{s})}{2L}\|\nabla f(x_{k+1})-\nabla f(x_k)\|^2 \\ &- n\sqrt{s}(f(x_k)-f(x^*)) - \frac{Bq\sqrt{s}}{2}\|v_k - x^*\|^2 \\ &+ \frac{Bp^2 s}{2}(1-q\sqrt{s})\|\nabla f(x_k)\|^2 + \frac{Bp^2 s}{2}(q\sqrt{s})\|\nabla f(x_k)\|^2. \qquad (50)\end{aligned}$$

Now, adding $\frac{Bp^2 s(1-q\sqrt{s})}{2}\|\nabla f(x_{k+1})\|^2$ to (50) results in

$$\begin{aligned}\varepsilon(k+1)-\varepsilon(k) &\overset{50}{\leq} -m\sqrt{s}(1-Bp\sqrt{s})\langle\nabla f(x_{k+1}),\nabla f(x_k)\rangle - \frac{(1-Bp\sqrt{s})}{2L}\|\nabla f(x_{k+1})-\nabla f(x_k)\|^2 \\ &- n\sqrt{s}(f(x_k)-f(x^*)) - \frac{Bq\sqrt{s}}{2}\|v_k - x^*\|^2 + \frac{Bp^2 s}{2}(1-q\sqrt{s})\|\nabla f(x_{k+1})\|^2 \\ &+ \frac{Bp^2 s}{2}(1-q\sqrt{s})\|\nabla f(x_k)\|^2 + \frac{Bp^2 s}{2}(q\sqrt{s})\|\nabla f(x_k)\|^2. \qquad (51)\end{aligned}$$

Next, setting $nps \leq m\sqrt{s}$ and noting that $n = q$, we get

$$\begin{aligned}&- m\sqrt{s}(1-Bp\sqrt{s})\langle\nabla f(x_{k+1}),\nabla f(x_k)\rangle + \tfrac{Bp^2 s(1-q\sqrt{s})}{2}\|\nabla f(x_{k+1})\|^2 + \tfrac{Bp^2 s(1-q\sqrt{s})}{2}\|\nabla f(x_k)\|^2 \\ &= (1-q\sqrt{s})\left[-m\sqrt{s}\langle\nabla f(x_{k+1}),\nabla f(x_k)\rangle + \frac{nps}{2}\|\nabla f(x_{k+1})\|^2 + \frac{nps}{2}\|\nabla f(x_k)\|^2\right] \\ &\leq \frac{(1-q\sqrt{s})m\sqrt{s}}{2}\|\nabla f(x_{k+1})-\nabla f(x_k)\|^2. \qquad (52)\end{aligned}$$

Using (52) in (51) we get

$$\varepsilon(k+1)-\varepsilon(k) \overset{52}{\leq} (1-q\sqrt{s})(\frac{m\sqrt{s}}{2}-\frac{1}{2L})\|\nabla f(x_{k+1})-\nabla f(x_k)\|^2 - n\sqrt{s}(f(x_k)-f(x^*))$$
$$-\frac{Bq\sqrt{s}}{2}\|v_k-x^*\|^2+\frac{Bp^2 s}{2}(q\sqrt{s})\|\nabla f(x_k)\|^2.$$

Setting $m\sqrt{s} \leq \frac{1}{L}$ leads to

$$\varepsilon(k+1)-\varepsilon(k) \leq -n\sqrt{s}(f(x_k)-f(x^*)) - \frac{Bq\sqrt{s}}{2}\|v_k-x^*\|^2+\frac{Bp^2 s}{2}(q\sqrt{s})\|\nabla f(x_k)\|^2,$$

which can be expressed in a more favourable way

$$\begin{aligned}\varepsilon(k+1)-\varepsilon(k) &\leq -(q\sqrt{s})\left[f(x_k)-f(x^*)+\frac{B}{2}\|v_k-x^*\|^2-\frac{Bp^2 s}{2}\|\nabla f(x_k)\|^2\right]\\ &= -(q\sqrt{s})\varepsilon(k). \end{aligned} \tag{53}$$

which gives

$$\varepsilon(k+1) \leq (1-q\sqrt{s})\varepsilon(k). \tag{54}$$

Therefore,

$$\varepsilon(k+1) \leq (1-q\sqrt{s})^{k+1}\varepsilon(0). \tag{55}$$

Using the form of $\varepsilon(k)$ in (43) and the inequality

$$f(x_k)-f(x^*) \geq \frac{1}{2L}\|\nabla f(x_k)\|^2,$$

which is true for any L-smooth function with x^* such that $\nabla f(x^*)=0$, we get

$$\begin{aligned}\varepsilon(k) &= (f(x_k)-f(x^*))+\frac{B}{2}\|v_k-x^*\|_2^2-\frac{Bp^2 s}{2}\|\nabla f(x_k)\|^2\\ &\geq (1-Bp^2 sL)(f(x_k)-f(x^*))+\frac{B}{2}\|v_k-x^*\|^2. \end{aligned} \tag{56}$$

Note that

$$1-Bp^2 sL = 1-npsL,$$

and under the conditions in Theorem 4.1 we have $npsL \leq 1$ and thus,

$$1-Bp^2 sL = 1-npsL \geq 0.$$

Hence, $(1-Bp^2 sL)(f(x_k)-f(x^*)) \geq 0$ and (56) leads to

$$\varepsilon(k) \geq \frac{B}{2}\|v_k-x^*\|_2^2. \tag{57}$$

From (55) we have

$$\begin{aligned}
&\varepsilon(k) \leq (1-q\sqrt{s})^k \varepsilon(0) \\
&\quad= (1-q\sqrt{s})^k \left[f(x_0) - f(x^*) + \frac{B}{2}\|v_0 - x^*\|_2^2 - \frac{Bp^2 s}{2}\|\nabla f(x_0)\|^2 \right] \\
&\quad\overset{v_0 = x_0 - \frac{m}{n}\nabla f(x_0)}{=} (1-q\sqrt{s})^k \left[f(x_0) - f(x^*) + \frac{B}{2}\|x_0 - \tfrac{m}{n}\nabla f(x_0) - x^*\|_2^2 \right. \\
&\quad \left. - \frac{Bp^2 s}{2}\|\nabla f(x_0)\|^2 \right] \\
&\quad= (1-q\sqrt{s})^k \left[f(x_0) - f(x^*) + \frac{B}{2}\|x_0 - x^*\|^2 + (\frac{Bm^2}{2n^2} - \frac{Bp^2 s}{2})\|\nabla f(x_0)\|^2 \right. \\
&\quad \left. - \frac{Bm}{n}\langle \nabla f(x_0), x_0 - x^* \rangle \right] \\
&\quad\overset{\textbf{S.C}}{\leq} (1-q\sqrt{s})^k \left[f(x_0) - f(x^*) + \frac{B}{2}\|x_0 - x^*\|^2 + (\frac{Bm^2}{2n^2} - \frac{Bp^2 s}{2})\|\nabla f(x_0)\|^2 \right. \\
&\quad \left. - \frac{Bm}{n}(f(x_0) - f(x^*)) - \frac{B\mu m}{2n}\|x_0 - x^*\|^2 \right] \\
&\quad= (1-q\sqrt{s})^k \left[(1 - B(\frac{m}{n}))(f(x_0) - f(x^*)) + \frac{B}{2}(1 - \frac{\mu m}{2n})\|x_0 - x^*\|^2 \right. \\
&\quad \left. + (\frac{Bm^2}{2n^2} - \frac{Bp^2 s}{2})\|\nabla f(x_0)\|^2 \right] \\
&\quad\overset{\textbf{S.L}}{\leq} (1-q\sqrt{s})^k \left[\frac{L}{2}(1 - B(\frac{m}{n})) + \frac{B}{2}(1 - \frac{\mu m}{2n}) + (\frac{Bm^2}{2n^2} - \frac{Bp^2 s}{2})L^2) \right] \|x_0 - x^*\|^2. \qquad (58)
\end{aligned}$$

Therefore, from (57) and (58) we get

$$\begin{aligned}
\frac{B}{2}\|v_k - x^*\|_2^2 \leq (1-q\sqrt{s})^k &\left[\frac{L}{2}(1 - B(\frac{m}{n})) + \right. \\
&\left. \frac{B}{2}(1 - \frac{\mu m}{2n}) + (\frac{Bm^2}{2n^2} - \frac{Bp^2 s}{2})L^2) \right] \|x_0 - x^*\|^2. \qquad (59)
\end{aligned}$$

The positivity of the coefficient

$$\left[\frac{L}{2}(1 - B(\frac{m}{n})) + \frac{B}{2}(1 - \frac{\mu m}{2n}) + (\frac{Bm^2}{2n^2} - \frac{Bp^2 s}{2})L^2) \right] \|x_0 - x^*\|^2$$

is guaranteed under the conditions of Theorem 4.1. Multiplying both sides by $2/B$ gives

$$\|v_k - x^*\|_2^2 \leq M' \|x_0 - x^*\|^2$$

for $M' = \left[\frac{L}{B}(1 - B(\frac{m}{n})) + (1 - \frac{\mu m}{2n}) + (\frac{m^2}{n^2} - p^2 s)L^2) \right]$. Now, we know that $\lim_{k\to\infty} v_k = x^*$. By representing algorithm (6) in one-line format of sequence v_k we get

$$v_{k+1} - v_k = -p\sqrt{s}\nabla f(x_{k+1}) + \frac{1-q\sqrt{s}}{1+n\sqrt{s}}(v_k - v_{k-1}) + \frac{q\sqrt{s}}{1+n\sqrt{s}}(\frac{p}{q} - m\sqrt{s})\nabla f(x_k). \qquad (60)$$

Analyzing (60) in limit and Taking $\iota = \frac{1}{1+n\sqrt{s}}(1 - \frac{qm\sqrt{s}}{p})$ we have

$$\lim_{k\to\infty} (\nabla f(x_{k+1}) - \iota \nabla f(x_k)) = 0. \qquad (61)$$

for $\iota < 1$. Thus,

$$\forall \epsilon \quad \exists k_0 \quad \text{s.t. } \|\nabla f(x_{k+1}) - \iota \nabla f(x_k)\| \leq \epsilon \qquad k \geq k_0,$$

From the above argument we get

$$\|\nabla f(x_{k+1})\| - \iota \|\nabla f(x_k)\| \leq \epsilon \to \|\nabla f(x_{k+1})\| \leq \iota \|\nabla f(x_k)\| + \epsilon.$$

Unrolling the above inequality results in

$$\|\nabla f(x_{k+1})\| \leq \iota^{k-k_0+1}\|\nabla f(x_{k_0})\| + \epsilon(1+\iota+\iota^2+\ldots+\iota^{k-k_0}),$$

$$\rightarrow \|\nabla f(x_{k+1})\| \leq \iota^{k-k_0+1}\|\nabla f(x_{k_0})\| + \frac{\epsilon}{1-\iota}.$$

Taking $\epsilon = \frac{(1-\iota)\delta}{2}$ and noting that

$$\forall\delta \quad \exists k_0 \leq k : \iota^{k-k_0+1}\|\nabla f(x_{k_0})\| \leq \delta/2,$$

we get

$$\|\nabla f(x_{k+1})\| \leq \delta,$$

and therefore,

$$\lim_{k\to\infty} \|\nabla f(x_{k+1})\| = 0.$$

Thus, the limit of $\nabla f(x_k)$ exists and since f is strongly convex $x_k \to x^*$ as $k \to \infty$. We therefore proved that sequence x_k converges to x^*.

B.5 PROOF OF COROLLARY 3.2.1

By comparing the one line representation (11) with (HR-TM) we can easily check the validity of parameters stated in the corollary. All that remains is to show that these parameters satisfy the conditions of Theorem 3.2. Due to positiveness of the parameters $(s, \beta, \gamma, \delta)$ in (TM-Method) and $n \geq q$ for $\xi \leq 2/3$ the first two conditions in Theorem 3.2 are satisfied. The last condition is $q/p \leq \mu$. Using the values for q and p we have

$$\begin{aligned} \frac{q}{p} &= \frac{\xi(2-\xi)M}{(1-\xi\gamma\sqrt{sM})(1+\sqrt{Ms})} \leq \mu \\ &\Rightarrow \frac{(1-\xi\gamma\sqrt{sM})(1+\sqrt{Ms})}{\xi(2-\xi)} - \frac{M}{\mu} \geq 0 \end{aligned} \tag{62}$$

Replacing TM method parameters and defining function $G(\xi)$ we get

$$G(\xi,\kappa) = \frac{\left(1-\xi(\frac{\rho^2}{(1+\rho)(2-\rho)})\sqrt{(1+\rho)\frac{M}{L}}\right)\left(1+\sqrt{(1+\rho)\frac{M}{L}}\right)}{\xi(2-\xi)} - \frac{M}{\mu}.$$

Now, using $M = \left(\frac{1-\beta}{\sqrt{s}(1+\beta)}\right)^2$ and equation (8) in (Sun et al., 2020) we have

$$\begin{aligned} \frac{M}{\mu} &= \frac{9\kappa^3\sqrt{\kappa} - 6\kappa^3 + \kappa^2\sqrt{\kappa}}{8\kappa^3\sqrt{\kappa} - 12\kappa^3 + 14\kappa^2\sqrt{\kappa} - 9\kappa^2 + 4\kappa\sqrt{\kappa - \kappa}}, \\ \frac{M}{L} &= \frac{9\kappa^2\sqrt{\kappa} - 6\kappa^2 + \kappa\sqrt{\kappa}}{8\kappa^3\sqrt{\kappa} - 12\kappa^3 + 14\kappa^2\sqrt{\kappa} - 9\kappa^2 + 4\kappa\sqrt{\kappa - \kappa}}. \end{aligned} \tag{63}$$

With $\rho = 1 - \frac{1}{\sqrt{\kappa}}$, all the terms in (63) depend on κ and we need $G(\xi,\kappa) \geq 0$ for (62) to hold. By analyzing $\lim_{\kappa\to\infty} G(\xi,\kappa)$ we see that for $\xi = \{\frac{2}{3}, \frac{4}{3}\}$ we get $\lim_{\kappa\to\infty} G(\xi,\kappa) = 0$. However, the solution $\xi = \frac{4}{3}$ is not acceptable since it leads to $q \geq n$ which is not the case in Theorem 3.2. Also, for $\kappa = 1$ we have

$$G(\xi,1) = \frac{3}{\xi(2-\xi)} - 3$$

which remains non-negative for $0 \leq \xi \leq \frac{2}{3}$. Therefore, invoking Theorem 3.2 results in (5) and concludes the proof.

B.6 PROOF OF COROLLARY 4.1.1

The proof is based on the QHM representation in (Zhang et al., 2021). It is possible to rewrite (QHM) in one-line format

$$x_{k+1} - x_k = b(x_k - x_{k-1}) - s\nabla f(x_k) + sb(1-a)\nabla f(x_{k-1}).$$

On the other hand, (6) has the one-line format

$$x_{k+1} - x_k = \frac{1 - q\sqrt{s}}{1 + n\sqrt{s}}(x_k - x_{k-1}) - (\frac{m\sqrt{s} + nps}{1 + n\sqrt{s}})\nabla f(x_k) + \frac{m\sqrt{s}(1 - q\sqrt{s})}{1 + n\sqrt{s}}\nabla f(x_{k-1}).$$

Taking $b = \frac{1-q\sqrt{s}}{1+n\sqrt{s}}$, $m = (1-a)\sqrt{s}$, $n = q$, $p = \frac{a}{n} + \sqrt{s}$ result in

$$p > m, nps = s(a + n\sqrt{s}) \leq (1-a)s = m\sqrt{s} \to a \leq \frac{1 - q\sqrt{s}}{2},$$

Also, since $n\sqrt{s} \leq 1/2$ we have $(1 - n\sqrt{s})/2 \geq 1/4$ and therefore $a \leq 1/4$. For $q/p \leq \mu$ to hold one needs

$$\frac{q}{\frac{a}{q} + \sqrt{s}} \leq \frac{q^2}{a} \leq \mu \to q \leq \sqrt{a\mu} \to q = \sqrt{a\mu},$$

and for $m\sqrt{s} = (1-a)s \leq 1/L$ we need

$$s \leq \frac{1}{L(1-a)} \overset{\frac{1}{1-a} \leq \frac{4}{3}}{\Longrightarrow} s \leq \frac{4}{3L}.$$

Since all the conditions of Theorem 4.1 are satisfied, for (QHM) we have

$$f(x_k) - f(x^*) \leq C(1 - \sqrt{a\mu s})^k.$$