A Appendix

A.1 Proof of Theorem 2.1

Consider the Lyapunov function

$$\varepsilon(t) = \frac{1}{2}\|X_t + e^{-\alpha_t}\dot{X}_t - x^* + \sqrt{s}e^{-\alpha_t}\nabla f(X_t)\|^2 + e^{\beta_t}(f(X_t) - f(x^*)). \tag{34}$$

Taking derivative with respect to t gives

$$\begin{aligned}\frac{d\varepsilon}{dt} =& \langle \frac{d}{dt}(X_t + e^{-\alpha_t}\dot{X}_t - x^* + \sqrt{s}e^{-\alpha_t}\nabla f(X_t)), X_t + e^{-\alpha_t}\dot{X}_t - x^* + \sqrt{s}e^{-\alpha_t}\nabla f(X_t)\rangle \\ &+ \dot{\beta}_t e^{\beta_t}(f(X_t) - f(x^*)) + e^{\beta_t}\langle \nabla f(X_t), \dot{X}_t\rangle. \end{aligned}\tag{35}$$

Note that (7) can be represented as

$$\frac{d}{dt}\left[X_t + e^{-\alpha_t}\dot{X}_t + \sqrt{s}e^{-\alpha_t}\nabla f(X_t)\right] = -e^{\alpha_t+\beta_t}\nabla f(X_t). \tag{36}$$

Using (36) in (35) we have

$$\begin{aligned}\frac{d\varepsilon}{dt} =& \langle -e^{\alpha_t+\beta_t}\nabla f(X_t), X_t + e^{-\alpha_t}\dot{X}_t - x^* + \sqrt{s}e^{-\alpha_t}\nabla f(X_t)\rangle \\ &+ \dot{\beta}_t e^{\beta_t}(f(X_t) - f(x^*)) + e^{\beta_t}\dot{X}_t\nabla f(X_t) \\ &= -e^{\alpha_t+\beta_t}\langle \nabla f(X_t), X_t - x^*\rangle - e^{\beta_t}\langle \nabla f(X_t), \dot{X}_t\rangle - \sqrt{s}e^{\beta_t}\|\nabla f(X_t)\|^2 \\ &+ \dot{\beta}_t e^{\beta_t}(f(X_t) - f(x^*)) + e^{\beta_t}\langle \nabla f(X_t), \dot{X}_t\rangle \\ &\overset{\text{(convexity)}}{\leq} -e^{\alpha_t+\beta_t}(f(X_t) - f(x^*)) + \dot{\beta}_t e^{\beta_t}(f(X_t) - f(x^*)) \\ &= -e^{\beta_t}\left[(e^{\alpha_t} - \dot{\beta}_t)(f(X_t) - f(x^*))\right].\end{aligned}$$

Utilizing the ideal scaling condition $\dot{\beta}_t \leq e^{\alpha_t}$ we have

$$\frac{d\varepsilon}{dt} \leq 0.$$

Thus, for the initialization point t_0 we have

$$e^{\beta_t}(f(X_t) - f(x^*)) \leq \varepsilon(t) \leq \varepsilon(t_0),$$

and the proof is complete.

A.2 Proof of Theorem 2.2

Consider the Lyapunov function

$$\varepsilon(t) = \frac{1}{2}\|X_t + e^{-\alpha_t}\dot{X}_t - x^* + \sqrt{s}e^{-\alpha_t}\nabla f(X_t)\|^2 + (e^{\beta_t} + \sqrt{s}e^{-2\alpha_t}\dot{\beta}_t)(f(X_t) - f(x^*)). \tag{37}$$

Taking derivative with respect to t gives

$$\begin{aligned}\frac{d\varepsilon}{dt} =& \langle \frac{d}{dt}(X_t + e^{-\alpha_t}\dot{X}_t - x^* + \sqrt{s}e^{-\alpha_t}\nabla f(X_t)), X_t + e^{-\alpha_t}\dot{X}_t - x^* + \sqrt{s}e^{-\alpha_t}\nabla f(X_t)\rangle \\ &+ (\dot{\beta}_t e^{\beta_t} - \sqrt{s}(2\dot{\alpha}_t)e^{-2\alpha_t}\dot{\beta}_t + \sqrt{s}e^{-\alpha_t}\ddot{\beta}_t)(f(X_t) - f(x^*)) \\ &+ (e^{\beta_t} + \sqrt{s}e^{-2\alpha_t}\dot{\beta}_t)\dot{X}_t\nabla f(X_t).\end{aligned}\tag{38}$$

Note that (7) can be represented as

$$\frac{d}{dt}\left[X_t + e^{-\alpha_t}\dot{X}_t + \sqrt{s}e^{-\alpha_t}\nabla f(X_t)\right] = -\left(e^{\alpha_t+\beta_t} + \sqrt{s}e^{-\alpha_t}\dot{\beta}_t\right)\nabla f(X_t). \tag{39}$$

Using (39) in (38) we have

$$
\begin{aligned}
\frac{d\varepsilon}{dt} =& \langle -\left(e^{\alpha_t+\beta_t} + \sqrt{s}e^{-\alpha_t}\dot{\beta}_t\right)\nabla f(X_t), X_t + e^{-\alpha_t}\dot{X}_t - x^* + \sqrt{s}e^{-\alpha_t}\nabla f(X_t)\rangle \\
&+ (\dot{\beta}_t e^{\beta_t} - \sqrt{s}(2\dot{\alpha}_t)e^{-2\alpha_t}\dot{\beta}_t + \sqrt{s}e^{-2\alpha_t}\ddot{\beta}_t)(f(X_t) - f(x^*)) \\
&+ (e^{\beta_t} + \sqrt{s}e^{-2\alpha_t}\dot{\beta}_t)\langle \nabla f(X_t), \dot{X}_t\rangle, \\
&= -\left(e^{\alpha_t+\beta_t} + \sqrt{s}e^{-\alpha_t}\dot{\beta}_t\right)\langle \nabla f(X_t), X_t - x^*\rangle - \left(e^{\beta_t} + \sqrt{s}e^{-2\alpha_t}\dot{\beta}_t\right)\langle \nabla f(X_t), \dot{X}_t\rangle \\
&- \sqrt{s}\left(e^{\beta_t} + \sqrt{s}e^{-2\alpha_t}\dot{\beta}_t\right)\|\nabla f(X_t)\|^2 \\
&+ (\dot{\beta}_t e^{\beta_t} - \sqrt{s}(2\dot{\alpha}_t)e^{-2\alpha_t}\dot{\beta}_t + \sqrt{s}e^{-2\alpha_t}\ddot{\beta}_t)(f(X_t) - f(x^*)) \\
&+ (e^{\beta_t} + \sqrt{s}e^{-2\alpha_t}\dot{\beta}_t)\langle \nabla f(X_t), \dot{X}_t\rangle, \\
&\overset{\text{(convexity)}}{\leq} -\left(e^{\alpha_t+\beta_t} + \sqrt{s}e^{-\alpha_t}\dot{\beta}_t\right)(f(X_t) - f(x^*)) \\
&+ (\dot{\beta}_t e^{\beta_t} - \sqrt{s}(2\dot{\alpha}_t)e^{-2\alpha_t}\dot{\beta}_t + \sqrt{s}e^{-2\alpha_t}\ddot{\beta}_t)(f(X_t) - f(x^*)), \\
&= -\left[e^{\beta_t}(e^{\alpha_t} - \dot{\beta}_t) + \sqrt{s}e^{-\alpha_t}(\dot{\beta}_t + 2\dot{\alpha}_t e^{-\alpha_t}\dot{\beta}_t - e^{-\alpha_t}\ddot{\beta}_t)\right](f(X_t) - f(x^*)).
\end{aligned}
$$

Utilizing the modified ideal scaling conditions $\dot{\beta}_t \leq e^{\alpha_t}$ and $\ddot{\beta}_t \leq e^{\alpha_t}\dot{\beta}_t + 2\dot{\alpha}_t\dot{\beta}_t$ we have

$$
\frac{d\varepsilon}{dt} \leq 0.
$$

Thus, for the initialization point t_0 we have

$$
(e^{\beta_t} + \sqrt{s}e^{-2\alpha_t}\dot{\beta}_t)(f(X_t) - f(x^*)) \leq \varepsilon(t) \leq \varepsilon(t_0),
$$

and the proof is complete.

A.3 Proof of Theorem 2.3

Consider the Lyapunov function

$$
\varepsilon(t) = e^{\beta_t}\left(\frac{\mu}{2}\|X_t - x^* + e^{-\alpha_t}\dot{X}_t + \frac{\sqrt{s}e^{\alpha_t}}{\mu}\nabla f(X_t)\|^2 + f(X_t) - f(x^*)\right). \tag{40}
$$

Taking derivative w.r.t. time gives

$$
\begin{aligned}
\frac{d\varepsilon(t)}{dt} &= \dot{\beta}e^{\beta_t}\left(\frac{\mu}{2}\|X_t - x^* + e^{-\alpha_t}\dot{X}_t + \frac{\sqrt{s}e^{\alpha_t}}{\mu}\nabla f(X_t)\|^2 + f(X_t) - f(x^*)\right) \\
&+ \mu e^{\beta_t}\left\langle \dot{X}_t - \dot{\alpha}_t e^{-\alpha_t}\dot{X}_t + e^{-\alpha_t}\ddot{X}_t + \frac{\sqrt{s}}{\mu}\dot{\alpha}_t e^{\alpha_t}\nabla f(X_t) + \frac{\sqrt{s}}{\mu}e^{\alpha_t}\nabla^2 f(X_t)\dot{X}_t \right. \\
&\left. , X_t - x^* + e^{-\alpha_t}\dot{X}_t + \frac{\sqrt{s}e^{\alpha_t}}{\mu}\nabla f(X_t)\right\rangle + e^{\beta_t}\langle \nabla f(X_t), \dot{X}_t\rangle.
\end{aligned} \tag{41}
$$

Next, we will use (17) in (41)

$$
\begin{aligned}
\frac{d\varepsilon(t)}{dt} &= \dot{\beta}e^{\beta_t}\left(\frac{\mu}{2}\|X_t - x^* + e^{-\alpha_t}\dot{X}_t + \frac{\sqrt{s}e^{\alpha_t}}{\mu}\nabla f(X_t)\|^2 + f(X_t) - f(x^*)\right) \\
&+ \mu e^{\beta_t}\left\langle \dot{X}_t - e^{-\alpha_t}(\dot{\gamma}_t + \dot{\beta}_t)\dot{X}_t + \frac{e^{\alpha_t}}{\mu}(\sqrt{s}(\dot{\alpha}_t - \dot{\beta}_t) - 1)\nabla f(X_t)\right. \\
&\left., X - x^* + e^{-\alpha_t}\dot{X}_t + \frac{\sqrt{s}e^{\alpha_t}}{\mu}\nabla f(X_t)\right\rangle + e^{\beta_t}\langle \nabla f(X_t), \dot{X}_t\rangle, \\
&= \dot{\beta}_t e^{\beta_t}\left(\frac{\mu}{2}\left[\|X_t - x^*\|^2 + e^{-2\alpha_t}\|\dot{X}_t\|^2 + \|\frac{\sqrt{s}e^{\alpha_t}}{\mu}\nabla f(X_t)\|^2 + 2e^{-\alpha_t}\langle X_t - x^*, \dot{X}_t\rangle\right.\right. \\
&\left.\left.+\frac{2\sqrt{s}e^{\alpha_t}}{\mu}\langle X_t - x^*, \nabla f(X_t)\rangle + \frac{2\sqrt{s}}{\mu}\langle \nabla f(X_t), \dot{X}_t\rangle\right]\right) + \dot{\beta}_t e^{\beta_t}(f(X_t) - f(x^*)) \\
&+ \mu e^{\beta_t}\left[(1 - e^{-\alpha_t}(\dot{\gamma}_t + \dot{\beta}_t))\langle \dot{X}_t, X_t - x^*\rangle + e^{-\alpha_t}(1 - e^{-\alpha_t}(\dot{\gamma}_t + \dot{\beta}_t))\|\dot{X}_t\|^2\right. \\
&+ \frac{\sqrt{s}e^{\alpha_t}}{\mu}(1 - e^{-\alpha_t}(\dot{\gamma}_t + \dot{\beta}_t))\langle \dot{X}_t, \nabla f(X_t)\rangle + \frac{e^{\alpha_t}}{\mu}(\sqrt{s}(\dot{\alpha}_t - \dot{\beta}_t) - 1)\langle \nabla f(X_t), X_t - x^*\rangle \\
&\left.+\frac{(\sqrt{s}(\dot{\alpha}_t - \dot{\beta}_t) - 1)}{\mu}\langle \nabla f(X_t), \dot{X}_t\rangle + \frac{\sqrt{s}e^{2\alpha_t}}{\mu^2}(\sqrt{s}(\dot{\alpha}_t - \dot{\beta}_t) - 1)\|\nabla f(X_t)\|^2\right] \\
&+ e^{\beta_t}\langle \nabla f(X_t), \dot{X}_t\rangle.
\end{aligned} \tag{42}
$$

Now, using strong convexity of f and applying $\alpha_t = \alpha$, $\dot{\beta} \geq 0$, $\dot{\gamma}_t = e^{\alpha_t}$, and $\dot{\beta}_t \leq e^{\alpha_t}$ gives

$$
\begin{aligned}
\frac{d\varepsilon(t)}{dt} &\leq -\dot{\beta}_t e^{\beta_t}\|\sqrt{\frac{\mu}{2}}e^{-\alpha_t}\dot{X}_t\|^2 - \sqrt{s}e^{\beta_t}\dot{\beta}_t\langle \dot{X}_t, \nabla f(X_t)\rangle - \dot{\beta}_t e^{\beta_t}\|\frac{\sqrt{s}e^{\alpha_t}}{\sqrt{2\mu}}\nabla f(X_t)\|^2 \\
&= -\dot{\beta}_t e^{\beta_t}\|\sqrt{\frac{\mu}{2}}e^{-\alpha_t}\dot{X}_t + \frac{\sqrt{s}e^{\alpha_t}}{\sqrt{2\mu}}\nabla f(X_t)\|^2 \leq 0,
\end{aligned} \tag{43}
$$

and therefore,

$$
e^{\beta_t}(f(X_t) - f(x^*)) \leq \varepsilon(t) \leq \varepsilon(0)
$$

and the proof is complete.

A.4 Proof of Theorem 3.1

Take the Lyapunov function

$$
\varepsilon(k) = \frac{s(k+2)k}{4}(f(x_k) - f(x^*)) + \frac{1}{2}\|x_{k+1} - x^* + \frac{k}{2}(x_{k+1} - x_k) + \frac{ks}{2}\nabla f(x_k)\|^2. \tag{44}
$$

The choice of Lyapunov function is the same as [Shi et al., 2021]. Note that the second term is equivalent to $\frac{1}{2}\|v_k - x^*\|^2$ through the first line of the update rule (21). Next, we will show that

$$
\varepsilon(k+1) - \varepsilon(k) \leq -\frac{s^2 k(k+2)}{8}\|\nabla f(x_k)\|^2. \tag{45}
$$

Using (44) we have

$$
\begin{aligned}
\varepsilon(k+1)-\varepsilon(k) &= \frac{s(k+3)(k+1)}{4}(f(x_{k+1})-f(x^*)) + \frac{1}{2}\|v_{k+1}-x^*\|^2 \\
&- \frac{s(k+2)(k)}{4}(f(x_k)-f(x^*)) + \frac{1}{2}\|v_k - x^*\|^2 \\
&= \frac{s(k+2)k}{4}(f(x_{k+1})-f(x_k)) + \frac{s(2k+3)}{4}(f(x_{k+1})-f(x^*)) \\
&+ \frac{1}{2}(2\langle v_{k+1}-v_k, v_k - x^*\rangle + \|v_{k+1}-v_k\|^2) \\
&= \frac{s(k+2)k}{4}(f(x_{k+1})-f(x_k)) + \frac{s(2k+3)}{4}(f(x_{k+1})-f(x^*)) \\
&+ \frac{1}{2}(2\langle -s(\frac{k+2}{2})\nabla f(x_{k+1}), x_{k+1}-x^* + \frac{k}{2}(x_{k+1}-x_k) + \frac{ks}{2}\nabla f(x_k)\rangle \\
&+ \|s(\frac{k+2}{2})\nabla f(x_{k+1})\|^2) \\
&= \frac{s(k+2)k}{4}(f(x_{k+1})-f(x_k)) + \frac{s(2k+3)}{4}(f(x_{k+1})-f(x^*)) \\
&- s(\frac{k+2}{2})\langle \nabla f(x_{k+1}), x_{k+1}-x^*\rangle - s\frac{k(k+2)}{4}\langle \nabla f(x_{k+1}), x_{k+1}-x_k\rangle \\
&- s^2(\frac{k(k+2)}{4})\langle \nabla f(x_{k+1}), \nabla f(x_k)\rangle + \frac{(s(k+2))^2}{8}\|\nabla f(x_{k+1})\|^2
\end{aligned} \tag{46}
$$

Now, from convexity and smoothness of the function f we have

$$
f(x_{k+1})-f(x_k) \leq \langle \nabla f(x_{k+1}), x_{k+1}-x_k\rangle - \frac{1}{2L}\|\nabla f(x_{k+1})-\nabla f(x_k)\|^2. \tag{47}
$$

Applying (47) in (46) we get

$$
\begin{aligned}
\varepsilon(k+1)-\varepsilon(k) &\leq \frac{s(k+2)k}{4}\left[\langle \nabla f(x_{k+1}), x_{k+1}-x_k\rangle - \frac{1}{2L}\|\nabla f(x_{k+1})-\nabla f(x_k)\|^2\right] \\
&+ \frac{s(2k+3)}{4}\left[\langle \nabla f(x_{k+1}), x_{k+1}-x^*\rangle - \frac{1}{2L}\|\nabla f(x_{k+1})\|^2\right] \\
&- s(\frac{k(k+2)}{4})\langle \nabla f(x_{k+1}), x_{k+1}-x_k\rangle - s(\frac{k+2}{2})\langle \nabla f(x_{k+1}), x_{k+1}-x^*\rangle \\
&- s^2(\frac{k(k+2)}{4})\langle \nabla f(x_{k+1}), \nabla f(x_k)\rangle + \frac{(s(k+2))^2}{8}\|\nabla f(x_{k+1})\|^2 \\
&\leq -\frac{s(k+2)k}{8L}\|\nabla f(x_{k+1})-\nabla f(x_k)\|^2 - \frac{s(2k+4)}{8L}\|\nabla f(x_{k+1})\|^2 \\
&- 2s^2(\frac{k(k+2)}{8})\langle \nabla f(x_{k+1}), \nabla f(x_k)\rangle + \frac{s^2k(k+2)}{8}\|\nabla f(x_{k+1})\|^2 \\
&+ \frac{s^2(k+2)}{4}\|\nabla f(x_{k+1})\|^2 \\
&= -\frac{s(k+2)k}{8}(\frac{1}{L}-s)\|\nabla f(x_{k+1})\|^2 - \frac{s(2k+4)}{8L}\|\nabla f(x_{k+1})\|^2 \\
&+ 2s(\frac{k(k+2)}{8})(\frac{1}{L}-s)\langle \nabla f(x_{k+1}), \nabla f(x_k)\rangle \\
&- \frac{s(k+2)k}{8}(\frac{1}{L}-s)\|\nabla f(x_k)\|^2 - s^2(\frac{k(k+2)}{8})\|\nabla f(x_k)\|^2 \\
&= -\frac{s(k+2)k}{8}(\frac{1}{L}-s)\|\nabla f(x_{k+1})-\nabla f(x_k)\|^2 - \frac{s(2k+4)}{8L}\|\nabla f(x_{k+1})\|^2 \\
&- s^2(\frac{k(k+2)}{8})\|\nabla f(x_k)\|^2 \leq -s^2(\frac{k(k+2)}{8})\|\nabla f(x_k)\|^2,
\end{aligned} \tag{48}
$$

where in the second inequality we used $-\langle \nabla f(x_{k+1}), x_{k+1}-x^*\rangle \leq -\frac{1}{2L}\|\nabla f(x_{k+1})\|^2$ and the last inequality holds as long as $s \leq 1/L$.

With (45) at hand, we can make sum both sides from $i = 0$ till $i = k-1$ and get

$$\begin{aligned}\varepsilon(k) - \varepsilon(0) &\leq -\frac{s^2}{8}\sum_{i=0}^{k} i(i+2)\|\nabla f(x_i)\|^2 \\ &\leq -\frac{s^2}{8}\min_{0\leq i\leq k}\|\nabla f(x_i)\|^2 \sum_{i=0}^{k} i(i+2) \\ &= -\frac{s^2}{8}\min_{0\leq i\leq k}\|\nabla f(x_i)\|^2 \sum_{i=1}^{k} i(i+2) \\ &= -\frac{s^2}{8}\min_{0\leq i\leq k}\|\nabla f(x_i)\|^2 \left[\frac{k(k+1)(2k+1)}{6} + k(k+1)\right] \\ &\leq -\frac{s^2}{8}\min_{0\leq i\leq k}\|\nabla f(x_i)\|^2 \left[\frac{k(k+1)(2k+1)}{6}\right] \\ &\leq -\frac{k^3 s^2}{24}\min_{0\leq i\leq k}\|\nabla f(x_i)\|^2. \end{aligned} \tag{49}$$

Note that $\varepsilon(k) \geq 0$. Therefore, we have

$$\begin{aligned} -\varepsilon(0) &\leq -\frac{k^3 s^2}{24}\min_{0\leq i\leq k}\|\nabla f(x_i)\|^2 \\ \rightarrow \varepsilon(0) &\geq \frac{k^3 s^2}{24}\min_{0\leq i\leq k}\|\nabla f(x_i)\|^2. \end{aligned} \tag{50}$$

Next, not ethat for $k = 0$, Lyapunov function (44) is equivalent to $1/2\|v_0 - x^*\|^2$. With initialization $v_0 = x_0$ we get

$$\frac{1}{2}\|x_0 - x^*\|^2 \geq \frac{k^3 s^2}{24}\min_{0\leq i\leq k}\|\nabla f(x_i)\|^2, \tag{51}$$

and therefore,

$$\min_{0\leq i\leq k}\|\nabla f(x_i)\|^2 \leq \frac{12}{k^3 s^2}\|x_0 - x^*\|^2, \tag{52}$$

for $0 < s \leq 1/L$ and $k \geq 1$. Also, from (49) we have $\varepsilon(k) \leq \varepsilon(0) = 1/2\|v_0 - x^*\|^2$. Thus,

$$f(x_k) - f(x^*) \leq \frac{2}{sk(k+2)}\|x_0 - x^*\|^2, \tag{53}$$

since $x_0 = v_0$. This completes the proof.

A.5 Proof of Proposition 4.1

From the update rule (24) we have

$$z_k = x_{k+1} + \frac{k}{2}(x_{k+1} - y_k) = x_{k+1} + \frac{k}{2}(x_{k+1} - x_k + s\nabla f(x_k)). \tag{54}$$

Replacing (54) in the update rule of z_k in (24), we get

$$x_{k+1} - x_k + \frac{k}{2}(x_{k+1} - x_k + s\nabla f(x_k)) - \frac{k-1}{2}(x_k - x_{k-1} + s\nabla f(x_{k-1})) = -\frac{sk}{2}\nabla f(y_k). \tag{55}$$

By rearranging we have

$$\begin{aligned} &x_{k+1} - x_k + \frac{1}{2}(x_k - x_{k-1}) + \frac{s}{2}\nabla f(x_{k-1}) + \frac{k}{2}(x_{k+1} + x_{k-1} - 2x_k) \\ &\quad + \frac{ks}{2}(\nabla f(x_k - \nabla f(x_{k-1})) = -\frac{sk}{2}\nabla f(y_k), \\ \rightarrow &\frac{2}{k\sqrt{s}}(\frac{x_{k+1} - x_k}{\sqrt{s}}) + \frac{1}{k\sqrt{s}}(\frac{x_k - x_{k-1}}{\sqrt{s}}) + \frac{1}{k}\nabla f(x_{k-1}) \\ &\quad + \frac{x_{k+1} - 2x_k + x_{k-1}}{s} + \nabla f(x_k) - \nabla f(x_{k-1}) = -\nabla f(y_k). \end{aligned} \tag{56}$$

Using approximations

$$\frac{2}{k\sqrt{s}}(\frac{x_{k+1}-x_k}{\sqrt{s}})+\frac{1}{k\sqrt{s}}(\frac{x_k-x_{k-1}}{\sqrt{s}})\approx\frac{1}{t_k}(2\dot{X}(t_k)) \tag{57}$$

$$\frac{x_{k+1}-2x_k+x_{k-1}}{s}\approx\ddot{X}(t_k) \tag{58}$$

$$\nabla f(x_k)-\nabla f(x_{k-1})\approx\sqrt{s}\nabla^2 f(X(t_k))\dot{X}(t_k) \tag{59}$$

$$\nabla f(y_k)=\nabla f(x_k-s\nabla f(x_k))\approx\nabla f(x_k)-s\nabla^2 f(x_k)\nabla f(x_k)\approx\nabla f(x_k) \tag{60}$$

$$X(t)\approx X(t_k) \qquad \dot{X}(t)\approx\dot{X}(t_k) \qquad \ddot{X}(t)\approx\ddot{X}(t_k) \qquad Y(t_k)\approx Y(t) \tag{61}$$

and $t_k=k\sqrt{s}$ in (56), we get

$$\ddot{X}(t)+(\frac{3}{t}+\sqrt{s}\nabla f(X(t)))\dot{X}(t)+(1+\frac{\sqrt{s}}{t})\nabla f(X(t))=0. \tag{62}$$

This, concludes the proof.

A.6 Proof of Theorem 5.2

Take the Lyapunov function

$$\varepsilon(k)=(\frac{t_k^2}{4}+\frac{s_k t_k}{2})(f(x_k)-f(x^*))+\frac{1}{2}\|v_k-x^*\|^2. \tag{63}$$

Next, we will bound the difference $\varepsilon(k+1)-\varepsilon(k)$. Using (63) we have

$$\begin{aligned}\varepsilon(k+1)-\varepsilon(k)&=(\frac{t_{k+1}^2}{4}+\frac{s_{k+1}t_{k+1}}{2})(f(x_{k+1})-f(x^*))\\&\quad-(\frac{t_k^2}{4}+\frac{s_k t_k}{2})(f(x_k)-f(x^*))+\frac{1}{2}(\|v_{k+1}-x^*\|^2-\|v_k-x^*\|^2),\\&=(\frac{t_{k+1}^2-t_k^2}{4}+\frac{s_{k+1}t_{k+1}-t_k s_k}{2})(f(x_{k+1})-f(x^*))\\&\quad+(\frac{t_k^2}{4}+\frac{s_k t_k}{2})(f(x_{k+1})-f(x_k))+\frac{1}{2}(\|v_{k+1}-x^*\|^2-\|v_k-x^*\|^2),\\&=(\frac{t_{k+1}^2-t_k^2}{4}+\frac{s_{k+1}t_{k+1}-t_k s_k}{2})(f(x_{k+1})-f(x^*))\\&\quad+(\frac{t_k^2}{4}+\frac{s_k t_k}{2})(f(x_{k+1})-f(x_k))+\frac{1}{2}(\|v_{k+1}-v_k\|^2+2\langle v_{k+1}-v_k,v_k-x^*\rangle),\end{aligned} \tag{64}$$

where in the last equality we used

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

Next, from the update (31) we have

$$\begin{cases} v_k-x^* = & \frac{t_k}{2s_k}(x_{k+1}-x_k)+x_{k+1}-x^*+\frac{t_k}{2\sqrt{L}}(\nabla f(x_k)+e_k),\\ v_{k+1}-v_k = & -\frac{1}{2}(t_k+2s_k)s_k(\nabla f(x_{k+1})+e_{k+1}).\end{cases} \tag{65}$$

Using (65) in (64) we have

$$\begin{aligned}\varepsilon(k+1)-\varepsilon(k)&=(\frac{t_{k+1}^2-t_k^2}{4}+\frac{s_{k+1}t_{k+1}-t_k s_k}{2})(f(x_{k+1})-f(x^*))\\&\quad+(\frac{t_k^2}{4}+\frac{s_k t_k}{2})(f(x_{k+1})-f(x_k))+\frac{1}{8}((t_k+2s_k)s_k)^2\|\nabla f(x_{k+1})+e_{k+1}\|^2\\&\quad-\frac{1}{2}\langle(t_k+2s_k)s_k(\nabla f(x_{k+1})+e_{k+1}),\frac{t_k}{2s_k}(x_{k+1}-x_k)+x_{k+1}-x^*\\&\quad+\frac{t_k}{2\sqrt{L}}(\nabla f(x_k)+e_k)\rangle.\end{aligned} \tag{66}$$

Now, using (47) in (66) we get

$$
\begin{aligned}
\varepsilon(k+1)-\varepsilon(k) \leq & \left(\frac{t_{k+1}^2-t_k^2}{4}+\frac{s_{k+1}t_{k+1}-t_ks_k}{2}\right)\left(\langle \nabla f(x_{k+1}), x_{k+1}-x^*\rangle - \frac{1}{2L}\|\nabla f(x_{k+1})\|^2\right) \\
&+\left(\frac{t_k^2}{4}+\frac{s_kt_k}{2}\right)\left(\langle \nabla f(x_{k+1}), x_{k+1}-x_k\rangle - \frac{1}{2L}\|\nabla f(x_{k+1})-\nabla f(x_k)\|^2\right) \\
&+\frac{1}{8}((t_k+2s_k)s_k)^2(\|\nabla f(x_{k+1})\|^2+\|e_{k+1}\|^2+2\langle \nabla f(x_{k+1}), e_k\rangle) \\
&-\left(\frac{t_k^2}{4}+\frac{s_kt_k}{2}\right)\langle \nabla f(x_{k+1})+e_{k+1}, x_{k+1}-x_k\rangle \\
&-\frac{(t_k+2s_k)s_k}{2}\langle \nabla f(x_{k+1})+e_{k+1}, x_{k+1}-x^*\rangle \\
&-\frac{t_k(t_k+2s_k)s_k}{4\sqrt{L}}\langle \nabla f(x_{k+1})+e_{k+1}, \nabla f(x_k)+e_k\rangle.
\end{aligned} \tag{67}
$$

Now, note that in (67) the terms containing $\langle \nabla f(x_{k+1}), x_{k+1}-x_k\rangle$ disappear. Due to $t_k=\sum_{i=1}^k s_k$, we get $t_{k+1}^2=t_k^2+2t_ks_{k+1}+s_{k+1}^2$ and $s_{k+1}t_{k+1}=s_{k+1}t_k+s_{k+1}^2$. Also, note that by definition $s_{k+1}\leq sk$ as long as $0<\alpha<1$. Thus,

$$
\left(\frac{t_{k+1}^2-t_k^2}{4}+\frac{s_{k+1}t_{k+1}-t_ks_k}{2}-\frac{(t_k+2s_k)s_k}{2}\right)\leq 0.
$$

Then, due to convexity and smoothness of f we get

$$
\begin{aligned}
&\left(\frac{t_{k+1}^2-t_k^2}{4}+\frac{s_{k+1}t_{k+1}-t_ks_k}{2}-\frac{(t_k+2s_k)s_k}{2}\right)\langle \nabla f(x_{k+1}), x_{k+1}-x^*\rangle \\
&\qquad\leq \frac{\left(\frac{t_{k+1}^2-t_k^2}{4}+\frac{s_{k+1}t_{k+1}-t_ks_k}{2}-\frac{(t_k+2s_k)s_k}{2}\right)}{2L}\|\nabla f(x_{k+1})\|^2.
\end{aligned} \tag{68}
$$

Replacing (68) in (67) and simplification gives

$$
\begin{aligned}
\varepsilon(k+1)-\varepsilon(k) \leq & -\frac{(t_k+2s_k)s_k}{4L}\|\nabla f(x_{k+1})\|^2-\frac{\left(\frac{t_k^2}{4}+\frac{s_kt_k}{2}\right)}{2L}\|\nabla f(x_{k+1})-\nabla f(x_k)\|^2 \\
&+\frac{1}{8}((t_k+2s_k)s_k)^2(\|\nabla f(x_{k+1})\|^2+\|e_{k+1}\|^2+2\langle \nabla f(x_{k+1}), e_k\rangle) \\
&-\left(\frac{t_k^2}{4}+\frac{s_kt_k}{2}\right)\langle e_{k+1}, x_{k+1}-x_k\rangle-\frac{(t_k+2s_k)s_k}{2}\langle e_{k+1}, x_{k+1}-x^*\rangle \\
&-\frac{t_k(t_k+2s_k)s_k}{4\sqrt{L}}\langle \nabla f(x_{k+1})+e_{k+1}, \nabla f(x_k)+e_k\rangle, \\
=&\ \frac{1}{2}\left(\frac{t_k}{2}+s_k\right)^2\left(s_k^2-\frac{1}{L}\right)\|\nabla f(x_{k+1})\|^2 \\
&+\frac{1}{2}\left(\frac{t_k^2}{4}+\frac{s_kt_k}{2}\right)\left(\frac{1}{L}-\frac{s_k}{\sqrt{L}}\right)2\langle \nabla f(x_{k+1}), \nabla f(x_k)\rangle \\
&-\frac{\left(\frac{t_k^2}{4}+\frac{s_kt_k}{2}\right)}{2L}\|\nabla f(x_k)\|^2+\frac{1}{8}((t_k+2s_k)s_k)^2(\|e_{k+1}\|^2+2\langle \nabla f(x_{k+1}), e_k\rangle) \\
&-\left(\frac{t_k^2}{4}+\frac{s_kt_k}{2}\right)\langle e_{k+1}, x_{k+1}-x_k\rangle-\frac{(t_k+2s_k)s_k}{2}\langle e_{k+1}, x_{k+1}-x^*\rangle \\
&-\frac{t_k(t_k+2s_k)s_k}{4\sqrt{L}}\left(\langle \nabla f(x_{k+1}), e_k\rangle+\langle \nabla f(x_k), e_{k+1}\rangle+\langle e_{k+1}, e_k\rangle\right).
\end{aligned} \tag{69}
$$

Here, the proof divides in 2 sections for each of the results in Theorem 5.2. First, we prove the rate for $\mathbb{E}\left[f(x_k)\right]-f(x^*)$.

a) Taking $c \leq 1/\sqrt{L}$ in $s_k = c/k^\alpha$, we get $s_k \leq 1/\sqrt{L}$. This implies

$$
\begin{aligned}
-\frac{1}{2}\left(\frac{t_k}{2}+s_k\right)^2\left(\frac{1}{L}-s_k^2\right) &\leq -\frac{1}{2}\left(\frac{t_k^2}{4}+\frac{s_k t_k}{2}\right)\left(\frac{1}{L}-\frac{s_k}{\sqrt{L}}\right), \\
-\frac{1}{2L}\left(\frac{t_k^2}{4}+\frac{s_k t_k}{2}\right) &\leq -\frac{1}{2}\left(\frac{t_k^2}{4}+\frac{s_k t_k}{2}\right)\left(\frac{1}{L}-\frac{s_k}{\sqrt{L}}\right).
\end{aligned}
\tag{70}
$$

Utilizing (70) in (69) gives

$$
\begin{aligned}
\varepsilon(k+1)-\varepsilon(k) \leq& -\frac{1}{2}\left(\frac{t_k^2}{4}+\frac{s_k t_k}{2}\right)\left(\frac{1}{L}-\frac{s_k}{\sqrt{L}}\right)\|\nabla f(x_{k+1})-\nabla f(x_k)\|^2 \\
&\frac{1}{8}((t_k+2s_k)s_k)^2(\|e_{k+1}\|^2)+\frac{1}{4}((t_k+2s_k)s_k)^2\langle \nabla f(x_{k+1}), e_k\rangle) \\
&-\left(\frac{t_k^2}{4}+\frac{s_k t_k}{2}\right)\langle e_{k+1}, x_{k+1}-x_k\rangle - \frac{(t_k+2s_k)s_k}{2}\langle e_{k+1}, x_{k+1}-x^*\rangle \\
&-\frac{t_k(t_k+2s_k)s_k}{4\sqrt{L}}\left(\langle \nabla f(x_{k+1}), e_k\rangle + \langle \nabla f(x_k), e_{k+1}\rangle + \langle e_{k+1}, e_k\rangle\right), \\
&\leq \frac{1}{8}((t_k+2s_k)s_k)^2\|e_{k+1}\|^2+\frac{1}{4}((t_k+2s_k)s_k)^2\langle \nabla f(x_{k+1}), e_k\rangle) \\
&-\left(\frac{t_k^2}{4}+\frac{s_k t_k}{2}\right)\langle e_{k+1}, x_{k+1}-x_k\rangle - \frac{(t_k+2s_k)s_k}{2}\langle e_{k+1}, x_{k+1}-x^*\rangle \\
&-\frac{t_k(t_k+2s_k)s_k}{4\sqrt{L}}\left(\langle \nabla f(x_{k+1}), e_k\rangle + \langle \nabla f(x_k), e_{k+1}\rangle + \langle e_{k+1}, e_k\rangle\right), \\
&\leq \frac{s_k(t_k+2s_k)}{2}\left(\frac{(t_k+2s_k)s_k}{4}\|e_{k+1}\|^2+\frac{s_k(t_k+2s_k)}{2}+\langle \nabla f(x_{k+1}), e_k\rangle\right. \\
&-\frac{t_k}{s_k}\langle e_{k+1}, x_{k+1}-x_k\rangle - \langle e_{k+1}, x_{k+1}-x^*\rangle \\
&\left.-\frac{t_k}{2\sqrt{L}}\left(\langle \nabla f(x_{k+1}), e_k\rangle + \langle \nabla f(x_k), e_{k+1}\rangle + \langle e_{k+1}, e_k\rangle\right)\right) \\
&= \frac{s_k(t_k+2s_k)}{2} g_k.
\end{aligned}
\tag{71}
$$

Next, note that $\mathbb{E}[g_k] = \frac{s_k(t_k+2s_k)}{4}\sigma^2$ and therefore, $\mathbb{E}[\varepsilon(k+1)] - \mathbb{E}[\varepsilon(k)] \leq \frac{s_k^2(t_k+2s_k)^2}{8}\sigma^2$. Also, by the form of $\varepsilon(k)$ in (63) we have

$$
\left(\frac{t_k^2}{4}+\frac{t_k s_k}{2}\right)(\mathbb{E}[f(x_k)] - f(x^*)) \leq \mathbb{E}[\varepsilon(k)].
$$

Thus, forming a telescope summation leads to

$$
\left(\frac{t_k^2}{4}+\frac{t_k s_k}{2}\right)(\mathbb{E}[f(x_k)] - f(x^*)) \leq \mathbb{E}[\varepsilon(k)] \leq \mathbb{E}[\varepsilon(0)] + \sum_{i=1}^{k-1}\frac{s_i^2(t_i+2s_i)^2}{8}\sigma^2,
\tag{72}
$$

with $s_0 = t_0 = 0$. From (72) one can get

$$
\mathbb{E}[f(x_k)] - f(x^*) \leq \frac{\mathbb{E}[\varepsilon(0)] + \sum_{i=1}^{k-1}\frac{s_i^2(t_i+2s_i)^2}{8}\sigma^2}{\left(\frac{t_k^2}{4}+\frac{t_k s_k}{2}\right)}.
\tag{73}
$$

Now, we should bound $\sum_{i=1}^{k-1} s_i^2(t_i+2s_i)^2$. Note that

$$
\sum_{i=1}^{k-1} s_i^2(t_i+2s_i)^2 = \sum_{i=1}^{k-1} s_i^2 t_i^2 + 4t_i s_i^3 + 4s_i^4,
$$

and

$$
t_i = \left(\sum_{j=1}^{i}\frac{c}{j^\alpha}\right) \leq \left(\int_0^i \frac{c}{t^\alpha}dt\right) = \frac{c i^{1-\alpha}}{(1-\alpha)},
$$

$$t_i^2 = \left(\sum_{j=1}^{i} \frac{c}{j^\alpha} \right)^2 \leq \left(\int_0^i \frac{c}{t^\alpha} dt \right)^2 = \frac{c^2 i^{2-2\alpha}}{(1-\alpha)^2}.$$

Therefore, for the first term we have

$$\begin{aligned} \sum_{i=1}^{k-1} s_i^2 t_i^2 &\leq \frac{c^4}{(1-\alpha)^2} \sum_{i=1}^{k-1} \frac{1}{i^{4\alpha-2}} \leq \frac{c^4}{(1-\alpha)^2}(1 + \int_1^k \frac{1}{t^{4\alpha-2}} dt) \leq \frac{c^4}{(1-\alpha)^2}(1 + \frac{1}{4\alpha-3}) \\ &= \frac{c^4(4\alpha-2)}{(1-\alpha)^2(4\alpha-3)}, \end{aligned} \tag{74}$$

when $\alpha > 3/4$ and if $\alpha = 3/4$ we have

$$\sum_{i=1}^{k-1} s_i^2 t_i^2 \leq 16c^4 \sum_{i=1}^{k-1} \frac{1}{i} \leq 16c^4(1 + \log(k)). \tag{75}$$

For the second term we have

$$\sum_{i=1}^{k-1} s_i^3 t_i \leq \frac{c^4}{(1-\alpha)} \sum_{i=1}^{k-1} \frac{1}{i^{4\alpha-1}} \leq \frac{c^4}{(1-\alpha)}(1 + \int_1^k \frac{1}{t^{4\alpha-1}} dt) \leq \frac{c^4(4\alpha-1)}{(1-\alpha)(4\alpha-2)} \tag{76}$$

for $\alpha > 1/2$. The third term gives

$$\sum_{i=1}^{k-1} s_i^4 \leq c^4 \sum_{i=1}^{k-1} \frac{1}{i^{4\alpha}} \leq c^4(1 + \int_1^k \frac{1}{t^{4\alpha}} dt) \leq \frac{c^4(4\alpha)}{(4\alpha-1)}, \tag{77}$$

for $\alpha > 1/4$.

For the terms in denominator of (73) we use lower bounds as

$$\begin{aligned} \frac{t_k^2}{4} &\geq \frac{c^2}{4(1-\alpha)^2}(k^{1-\alpha} - 1)^2, \\ \frac{t_k s_k}{2} &\geq \frac{c^2 k^{-\alpha}}{2(1-\alpha)}(k^{1-\alpha} - 1). \end{aligned} \tag{78}$$

Using(74,75,76,77,78) in (73) leads to

$$\mathbb{E}[f(x_k)] - f(x^*) \leq \begin{cases} \frac{\mathbb{E}[\varepsilon(0)] + \frac{c^4\sigma^2}{8}[16(1+\log(k))+32+6]}{2c^2\left[2(k^{\frac{1}{4}}-1)^2 + k^{-\frac{3}{4}}(k^{\frac{1}{4}}-1)\right]} & \alpha = \frac{3}{4} \\ \frac{\mathbb{E}[\varepsilon(0)] + \frac{c^4\sigma^2}{8}\left[\frac{(4\alpha-2)}{(1-\alpha)^2(4\alpha-3)} + \frac{4(4\alpha-1)}{(1-\alpha)(4\alpha-2)} + \frac{4(4\alpha)}{(4\alpha-1)}\right]}{\frac{c^2}{2(1-\alpha)}\left[\frac{(k^{1-\alpha}-1)^2}{2(1-\alpha)} + k^{-\alpha}(k^{(1-\alpha)}-1)\right]} & 1 > \alpha > \frac{3}{4} \end{cases} \tag{79}$$

with $\mathbb{E}[\varepsilon(0)] = \frac{1}{2}\|v_0 - x^*\|^2$.

b) On the transition from (69) to (71), one can write

$$\begin{aligned} \varepsilon(k+1) - \varepsilon(k) &\leq \frac{s_k(t_k + 2s_k)}{2} g_k - \frac{1}{2}\left(\frac{t_k^2}{4} + \frac{s_k t_k}{2}\right)(\frac{1}{L} - \frac{s_k}{\sqrt{L}})\|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2 \\ &\quad - \frac{1}{2}\left(\frac{t_k^2}{4} + \frac{s_k t_k}{2}\right)(\frac{s_k}{\sqrt{L}})\|\nabla f(x_k)\|^2 \\ &\leq \frac{s_k(t_k + 2s_k)}{2} g_k - \frac{1}{2}\left(\frac{t_k^2}{4} + \frac{s_k t_k}{2}\right)(\frac{s_k}{\sqrt{L}})\|\nabla f(x_k)\|^2 \end{aligned} \tag{80}$$

Recursively summing (80) from 0 to k gives

$$0 \leq \varepsilon(k) \leq \varepsilon(0) + \sum_{i=0}^{k-1} \frac{s_i(t_i + 2s_i)}{2} g_i - \frac{1}{2}\sum_{i=0}^{k-1}\left(\frac{t_i^2}{4} + \frac{s_i t_i}{2}\right)(\frac{s_i}{\sqrt{L}})\|\nabla f(x_i)\|^2, \tag{81}$$

which results in

$$\frac{1}{2}\sum_{i=0}^{k-1}\left(\frac{t_i^2}{4}+\frac{s_it_i}{2}\right)\left(\frac{s_i}{\sqrt{L}}\right)\|\nabla f(x_i)\|^2 \le \varepsilon(0)+\sum_{i=0}^{k-1}\frac{s_i(t_i+2s_i)}{2}g_i, \tag{82}$$

and therefore,

$$\min_{0\le i\le k-1}\|\nabla f(x_i)\|^2 \le \frac{\varepsilon(0)+\sum_{i=0}^{k-1}\frac{s_i(t_i+2s_i)}{2}g_i}{\frac{1}{2\sqrt{L}}\sum_{i=0}^{k-1}\left(\frac{t_i^2}{4}+\frac{s_it_i}{2}\right)s_i}, \tag{83}$$

Evaluating the expectation of both hand sides gives

$$\begin{aligned}
\mathbb{E}\left[\min_{0\le i\le k-1}\|\nabla f(x_i)\|^2\right] &\le \frac{\mathbb{E}\left[\varepsilon(0)\right]+\sum_{i=0}^{k-1}\frac{s_i(t_i+2s_i)}{2}\mathbb{E}\left[g_i\right]}{\frac{1}{2\sqrt{L}}\sum_{i=0}^{k-1}\left(\frac{t_i^2}{4}+\frac{s_it_i}{2}\right)s_i},\\
&= \frac{\mathbb{E}\left[\varepsilon(0)\right]+\sum_{i=0}^{k-1}\frac{s_i^2(t_i+2s_i)^2}{8}\sigma^2}{\frac{1}{2\sqrt{L}}\sum_{i=0}^{k-1}\left(\frac{t_i^2}{4}+\frac{s_it_i}{2}\right)s_i},
\end{aligned}$$

and the last equality is due to $\mathbb{E}[g_i]=\frac{s_i(t_i+2s_i)}{4}\sigma^2$. Next, we will bound the numerator from above and lower bound the denominator as

$$\sum_{i=0}^{k-1}\frac{s_i^2(t_i+2s_i)^2}{8}\sigma^2 \overset{(75,76,77)}{\le} 2\sigma^2c^4(1+\log(k))+\frac{\sigma^2c^4(4\alpha-1)}{2(1-\alpha)(4\alpha-2)}+\frac{\sigma^2c^4(4\alpha)}{2(4\alpha-1)} \tag{84}$$

for $\alpha=3/4$. For bounding the denominator one can use (78) as

$$\begin{aligned}
\frac{s_it_i^2}{4} &\ge \frac{c^3i^{-\alpha}}{4(1-\alpha)^2}(i^{1-\alpha}-1)^2,\\
\frac{t_is_i^2}{2} &\ge \frac{c^3i^{-2\alpha}}{2(1-\alpha)}(i^{1-\alpha}-1).
\end{aligned} \tag{85}$$

Applying the summation to (85) gives

$$\begin{aligned}
\sum_{i=0}^{k-1}\left(\frac{s_it_i^2}{4}+\frac{s_i^2t_i}{2}\right) &\ge \sum_{i=0}^{k-1}\frac{c^3}{4(1-\alpha)^2}(i^{2-3\alpha}-2i^{1-2\alpha}+i^{-\alpha})\\
&+\sum_{i=0}^{k-1}\frac{c^3}{2(1-\alpha)}(i^{1-3\alpha}-i^{-2\alpha})\\
&\ge \frac{c^3}{4(1-\alpha)^2}\left(\frac{k^{3(1-\alpha)}-1}{3-3\alpha}+\frac{(k^{1-\alpha}-1)}{1-\alpha}+\frac{1-2\alpha+k^{2-2\alpha}}{1-\alpha}\right)\\
&+\frac{c^3}{2(1-\alpha)}\left(\frac{k^{2-3\alpha}-1}{2-3\alpha}+\frac{(2\alpha-k^{1-2\alpha})}{1-2\alpha}\right)\\
&\overset{\alpha=\frac{3}{4}}{=} 4c^3\left(\frac{k^{3/4}-1}{3/4}+\frac{(k^{1/4}-1)}{1/4}+\frac{-\frac{1}{2}+k^{1/2}}{1/4}\right)\\
&+2c^3\left(-4k^{-1/4}+4-3+2k^{-1/2}\right)\\
&\ge 4c^3\left(\frac{k^{3/4}-1}{3/4}+\frac{(k^{1/4}-1)}{1/4}+\frac{-\frac{1}{2}+k^{1/2}}{1/4}\right)
\end{aligned} \tag{86}$$

Combining (86) and (84) gives

$$\begin{aligned}
2\sqrt{L}\frac{\sum_{i=0}^{k-1}\frac{s_i^2(t_i+2s_i)^2}{8}\sigma^2}{\sum_{i=0}^{k-1}\left(\frac{s_it_i^2}{4}+\frac{s_i^2t_i}{2}\right)} &\le 2\sqrt{L}\frac{2\sigma^2c^4(1+\log(k))+\frac{\sigma^2c^4(4\alpha-1)}{2(1-\alpha)(4\alpha-2)}+\frac{\sigma^2c^4(4\alpha)}{2(4\alpha-1)}}{4c^3\left(\frac{k^{3/4}-1}{3/4}+\frac{(k^{1/4}-1)}{1/4}+\frac{-\frac{1}{2}+k^{1/2}}{1/4}\right)}\\
&\overset{\alpha=\frac{3}{4}}{=} 2c\sigma^2\sqrt{L}\frac{2\log(k)+6+\frac{3}{4}}{16\left(\frac{k^{3/4}-1}{3}+k^{1/4}-\frac{3}{2}+k^{1/2}\right)},
\end{aligned} \tag{87}$$

and therefore,

$$\mathbb{E}\left[\min_{0\le i\le k-1}\|\nabla f(x_i)\|^2\right] \le \frac{2\sqrt{L}\mathbb{E}[\varepsilon(0)]}{16c^3\left(\frac{k^{3/4}-1}{3}+k^{1/4}-\frac{3}{2}+k^{1/2}\right)} + 2c\sigma^2\sqrt{L}\frac{2\log(k)+6+\frac{3}{4}}{16\left(\frac{k^{3/4}-1}{3}+k^{1/4}-\frac{3}{2}+k^{1/2}\right)}, \tag{88}$$

with $\mathbb{E}[\varepsilon(0)] = \frac{1}{2}\|x_0-x^*\|^2$.

A.7 Proof of Theorem 5.1

Take the Lyapunov function

$$\varepsilon(k) = (\frac{t_k^2}{4}+\frac{t_k\beta}{2\sqrt{L}})(f(x_k)-f(x^*)) + \frac{1}{2}\|v_k-x^*\|^2. \tag{89}$$

Using (89) we have

$$\begin{aligned}
\varepsilon(k+1)-\varepsilon(k) &= (\frac{t_{k+1}^2}{4}+\frac{\beta t_{k+1}}{2\sqrt{L}})(f(x_{k+1})-f(x^*)) \\
&- (\frac{t_k^2}{4}+\frac{\beta t_k}{2\sqrt{L}})(f(x_k)-f(x^*)) + \frac{1}{2}(\|v_{k+1}-x^*\|^2-\|v_k-x^*\|^2), \\
&= (\frac{t_{k+1}^2-t_k^2}{4}+\frac{\beta t_{k+1}-\beta t_k}{2\sqrt{L}})(f(x_{k+1})-f(x^*)) \\
&+ (\frac{t_k^2}{4}+\frac{\beta t_k}{2\sqrt{L}})(f(x_{k+1})-f(x_k)) + \frac{1}{2}(\|v_{k+1}-x^*\|^2-\|v_k-x^*\|^2), \\
&= (\frac{t_{k+1}^2-t_k^2}{4}+\frac{\beta(t_{k+1}-t_k)}{2\sqrt{L}})(f(x_{k+1})-f(x^*)) \\
&+ (\frac{t_k^2}{4}+\frac{\beta t_k}{2\sqrt{L}})(f(x_{k+1})-f(x_k)) + \frac{1}{2}(\|v_{k+1}-v_k\|^2+2\langle v_{k+1}-v_k, v_k-x^*\rangle),
\end{aligned} \tag{90}$$

where in the last equality we used

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

Next, from the update (29) we have

$$\begin{cases} v_k - x^* = & \frac{t_k}{2s_k}(x_{k+1}-x_k)+x_{k+1}-x^*+\frac{t_k\beta}{2\sqrt{L}}(\nabla f(x_k)+e_k), \\ v_{k+1}-v_k = & -\frac{1}{2}(t_ks_k+\frac{2s_k\beta}{\sqrt{L}})(\nabla f(x_{k+1})+e_{k+1}). \end{cases} \tag{91}$$

Using (91) in (90) we have

$$\begin{aligned}
\varepsilon(k+1)-\varepsilon(k) &= (\frac{t_{k+1}^2-t_k^2}{4}+\frac{\beta(t_{k+1}-t_k)}{2\sqrt{L}})(f(x_{k+1})-f(x^*)) \\
&+ (\frac{t_k^2}{4}+\frac{\beta t_k}{2\sqrt{L}})(f(x_{k+1})-f(x_k)) + \frac{1}{8}((t_k+\frac{2\beta}{\sqrt{L}})s_k)^2\|\nabla f(x_{k+1})+e_{k+1}\|^2 \\
&- \frac{1}{2}\langle(t_k+\frac{2\beta}{\sqrt{L}})s_k(\nabla f(x_{k+1})+e_{k+1}), \frac{t_k}{2s_k}(x_{k+1}-x_k)+x_{k+1}-x^* \\
&+ \frac{t_k\beta}{2\sqrt{L}}(\nabla f(x_k)+e_k)\rangle.
\end{aligned} \tag{92}$$

Now, using (47) in (92) we get

$$\begin{aligned}
\varepsilon(k+1)-\varepsilon(k) \leq & \, (\frac{t_{k+1}^2-t_k^2}{4}+\frac{\beta(t_{k+1}-t_k)}{2\sqrt{L}})(\langle\nabla f(x_{k+1}), x_{k+1}-x^*\rangle-\frac{1}{2L}\|\nabla f(x_{k+1})\|^2) \\
& + (\frac{t_k^2}{4}+\frac{\beta t_k}{2\sqrt{L}})(\langle\nabla f(x_{k+1}), x_{k+1}-x_k\rangle-\frac{1}{2L}\|\nabla f(x_{k+1})-\nabla f(x_k)\|^2) \\
& + \frac{1}{8}((t_k+\frac{2\beta}{\sqrt{L}})s_k)^2(\|\nabla f(x_{k+1})\|^2+\|e_{k+1}\|^2+2\langle\nabla f(x_{k+1}), e_k\rangle) \\
& - (\frac{t_k^2}{4}+\frac{\beta t_k}{2\sqrt{L}})\langle\nabla f(x_{k+1})+e_{k+1}, x_{k+1}-x_k\rangle \\
& - \frac{(t_k+\frac{2\beta}{\sqrt{L}})s_k}{2}\langle\nabla f(x_{k+1})+e_{k+1}, x_{k+1}-x^*\rangle \\
& - \frac{\beta t_k(t_k+\frac{2\beta}{\sqrt{L}})s_k}{4\sqrt{L}}\langle\nabla f(x_{k+1})+e_{k+1}, \nabla f(x_k)+e_k\rangle.
\end{aligned} \tag{93}$$

Due to $t_k=\sum_{i=1}^k s_k$, we get $t_{k+1}^2=t_k^2+2t_k s_{k+1}+s_{k+1}^2$ and $t_{k+1}=t_k+s_{k+1}$. Also, note that by definition $s_{k+1}\leq sk$ as long as $0<\alpha<1$. Thus,

$$\left(\frac{t_{k+1}^2-t_k^2}{4}+\frac{\beta(t_{k+1}-t_k)}{2\sqrt{L}}-\frac{(t_k+2\frac{\beta}{\sqrt{L}})s_k}{2}\right)\leq 0,$$

for $\beta\geq 1/(2k^\alpha)$. Then, due to convexity and smoothness of f we get

$$\begin{aligned}
& \left(\frac{t_{k+1}^2-t_k^2}{4}+\frac{\beta(t_{k+1}-t_k)}{2\sqrt{L}}-\frac{(t_k+2\frac{\beta}{\sqrt{L}})s_k}{2}\right)\langle\nabla f(x_{k+1}), x_{k+1}-x^*\rangle \\
& \qquad \leq \frac{\left(\frac{t_{k+1}^2-t_k^2}{4}+\frac{\beta(t_{k+1}-t_k)}{2\sqrt{L}}-\frac{(t_k+2\frac{\beta}{\sqrt{L}})s_k}{2}\right)}{2L}\|\nabla f(x_{k+1})\|^2.
\end{aligned} \tag{94}$$

Replacing (94) in (93) and simplification gives

$$\begin{aligned}
\varepsilon(k+1)-\varepsilon(k) \leq & -\frac{(t_k+2\frac{\beta}{\sqrt{L}})s_k}{4L}\|\nabla f(x_{k+1})\|^2-\frac{(\frac{t_k^2}{4}+\frac{\beta t_k}{2\sqrt{L}})}{2L}\|\nabla f(x_{k+1})-\nabla f(x_k)\|^2 \\
& + \frac{1}{8}((t_k+\frac{2\beta}{\sqrt{L}})s_k)^2(\|\nabla f(x_{k+1})\|^2+\|e_{k+1}\|^2+2\langle\nabla f(x_{k+1}), e_k\rangle) \\
& - (\frac{t_k^2}{4}+\frac{\beta t_k}{2\sqrt{L}})\langle e_{k+1}, x_{k+1}-x_k\rangle-\frac{(t_k+\frac{2\beta}{\sqrt{L}})s_k}{2}\langle e_{k+1}, x_{k+1}-x^*\rangle \\
& - \frac{\beta t_k(t_k+\frac{2\beta}{\sqrt{L}})s_k}{4\sqrt{L}}\langle\nabla f(x_{k+1})+e_{k+1}, \nabla f(x_k)+e_k\rangle, \\
= & \left(\frac{(t_k+\frac{2\beta}{\sqrt{L}})s_k)^2}{8}-\frac{(\frac{t_k^2}{4}+\frac{\beta t_k}{2\sqrt{L}})}{2L}-\frac{(t_k+2\frac{\beta}{\sqrt{L}})s_k}{4L}\right)\|\nabla f(x_{k+1})\|^2 \\
& + \frac{1}{2}\left(\frac{t_k^2}{4}+\frac{\beta t_k}{2\sqrt{L}}\right)(\frac{1}{L}-\frac{\beta s_k}{\sqrt{L}})2\langle\nabla f(x_{k+1}), \nabla f(x_k)\rangle \\
& - \frac{(\frac{t_k^2}{4}+\frac{\beta t_k}{2\sqrt{L}})}{2L}\|\nabla f(x_k)\|^2+\frac{((t_k+\frac{2\beta}{\sqrt{L}})s_k)^2}{8}(\|e_{k+1}\|^2+2\langle\nabla f(x_{k+1}), e_k\rangle) \\
& - (\frac{t_k^2}{4}+\frac{\beta t_k}{2\sqrt{L}})\langle e_{k+1}, x_{k+1}-x_k\rangle-\frac{(t_k+\frac{2\beta}{\sqrt{L}})s_k}{2}\langle e_{k+1}, x_{k+1}-x^*\rangle \\
& - \frac{\beta t_k(t_k+\frac{2\beta}{\sqrt{L}})s_k}{4\sqrt{L}}\left(\langle\nabla f(x_{k+1}), e_k\rangle+\langle\nabla f(x_k), e_{k+1}\rangle+\langle e_{k+1}, e_k\rangle\right).
\end{aligned} \tag{95}$$

Lemma A.1. *Consider* $t_k = \sum_{i=0}^{k} s_i$ *and* $s_i = \frac{c}{i^\alpha}$ *with* $0 < \alpha < 1$ *and* $c \leq 1/\sqrt{L}$*. For any* $\beta \geq 1$ *there exists* k_0 *such that*

$$\left(\frac{((t_k + \frac{2\beta}{\sqrt{L}})s_k)^2}{8} - \frac{(\frac{t_k^2}{4} + \frac{\beta t_k}{2\sqrt{L}})}{2L} - \frac{(t_k + 2\frac{\beta}{\sqrt{L}})s_k}{4L} \right) \leq -\left| \frac{1}{2}\left(\frac{t_k^2}{4} + \frac{\beta t_k}{2\sqrt{L}}\right)(\frac{1}{L} - \frac{\beta s_k}{\sqrt{L}}) \right| \tag{96}$$

Proof. Without loss of generality, consider $\frac{1}{L} - \frac{\beta s_k}{\sqrt{L}} \geq 0$. Then, after simplification of (96) we have

$$\frac{((t_k + \frac{2\beta}{\sqrt{L}})s_k)^2}{8} - \frac{(t_k + 2\frac{\beta}{\sqrt{L}})s_k}{4L} - \frac{1}{2}\left(\frac{t_k^2}{4} + \frac{\beta t_k}{2\sqrt{L}}\right)(\frac{\beta s_k}{\sqrt{L}}) \leq 0$$

$$\frac{t_k^2 s_k}{8}(s_k - \frac{\beta}{\sqrt{L}}) + \frac{t_k \beta s_k}{2\sqrt{L}}(s_k - \frac{\beta}{2\sqrt{L}}) + \frac{\beta s_k}{2L}(\beta s_k - \frac{1}{\sqrt{L}}) - \frac{t_k s_k}{4L} \leq 0 \tag{97}$$

which holds as long as $\frac{2}{k^\alpha} \leq \beta \leq k^\alpha$. Note that for any choice of β there exists k_0 such that $\beta \leq k_0^\alpha$. Note that if $c = 1/\sqrt{L}$, then we can improve this bound (in the sense that smaller k_0 is needed) to $\beta \leq k_0^\alpha(\frac{1}{8}(\sum_{i=1}^{k}\frac{1}{i^\alpha})^2 + 1)$. This is done by considering $\frac{t_k^2 s_k}{8}(-\frac{\beta}{2\sqrt{L}})$ from the first term when showing the third term is negative. □

Next, using Theorem A.1 in (95) gives

$$\begin{aligned}
\varepsilon(k+1) - \varepsilon(k) \leq{}& \frac{((t_k + \frac{2\beta}{\sqrt{L}})s_k)^2}{8}(\|e_{k+1}\|^2 + 2\langle \nabla f(x_{k+1}), e_k\rangle) \\
&- (\frac{t_k^2}{4} + \frac{\beta t_k}{2\sqrt{L}})\langle e_{k+1}, x_{k+1} - x_k\rangle - \frac{(t_k + \frac{2\beta}{\sqrt{L}})s_k}{2}\langle e_{k+1}, x_{k+1} - x^*\rangle \\
&- \frac{\beta t_k (t_k + \frac{2\beta}{\sqrt{L}})s_k}{4\sqrt{L}}\left(\langle \nabla f(x_{k+1}), e_k\rangle + \langle \nabla f(x_k), e_{k+1}\rangle + \langle e_{k+1}, e_k\rangle\right).
\end{aligned} \tag{98}$$

By evaluating the expectation of (98) we have

$$\mathbb{E}\left[\varepsilon(k) - \varepsilon(k-1)\right] \leq \frac{((t_k + \frac{2\beta}{\sqrt{L}})s_k)^2}{8}\sigma^2. \tag{99}$$

Summing both sides from k_0 to k gives

$$\mathbb{E}[\varepsilon(k)] \leq \mathbb{E}[\varepsilon(k_0)] + \sum_{i=k_0+1}^{k} \frac{((t_i + \frac{2\beta}{\sqrt{L}})s_i)^2}{8}\sigma^2 \tag{100}$$

As in (75,76,77), one can bound (100) as

$$\begin{aligned}
\mathbb{E}[\varepsilon(k)] \leq{}& \mathbb{E}[\varepsilon(k_0)] + \frac{\sigma^2 c^4}{(1-\alpha)^2}\left[k_0^{3-4\alpha} - k^{3-4\alpha}\right] + \frac{\sigma^2 c^3 \beta}{2\sqrt{L}(1-\alpha)(3\alpha-2)}\left[k_0^{2-3\alpha} - k^{2-3\alpha}\right] \\
&+ \frac{\beta^2 c^2 \sigma^2}{2L(2\alpha-1)}\left[k_0^{1-2\alpha} - k^{1-2\alpha}\right]
\end{aligned} \tag{101}$$

for $\alpha > 3/4$ and

$$\mathbb{E}[\varepsilon(k)] \leq \mathbb{E}[\varepsilon(k_0)] + 2\sigma^2 c^4\left[\log(\frac{k}{k_0})\right] + \frac{8\sigma^2 c^3 \beta}{\sqrt{L}}\left[k_0^{-\frac{1}{4}} - k^{-\frac{1}{4}}\right] + \frac{\beta^2 c^2 \sigma^2}{L}\left[k_0^{-\frac{1}{2}} - k^{-\frac{1}{2}}\right] \tag{102}$$

for $\alpha = 3/4$. Therefore,

$$\mathbb{E}[f(x_k)] - f(x^*) \leq \frac{\mathbb{E}[\varepsilon(k_0)] + \frac{\sigma^2 c^4}{(1-\alpha)^2}\left[k_0^{3-4\alpha} - k^{3-4\alpha}\right] + \frac{\sigma^2 c^3 \beta}{2\sqrt{L}(1-\alpha)(3\alpha-2)}\left[k_0^{2-3\alpha} - k^{2-3\alpha}\right] + \frac{\beta^2 c^2 \sigma^2}{2L(2\alpha-1)}\left[k_0^{1-2\alpha} - k^{1-2\alpha}\right]}{\frac{c^2}{4(1-\alpha)^2}((k^{1-\alpha}-1)^2) + \frac{c\beta}{2\sqrt{L}(1-\alpha)}\left(k^{(1-\alpha)}-1\right)} \tag{103}$$

for $\alpha > 3/4$ and

$$\mathbb{E}[f(x_k)] - f(x^*) \leq \frac{\mathbb{E}[\varepsilon(k_0)] + 2\sigma^2 c^4\left[\log(\frac{k}{k_0})\right] + \frac{8\sigma^2 c^3 \beta}{\sqrt{L}}\left[k_0^{-1/4} - k^{-1/4}\right] + \frac{\beta^2 c^2 \sigma^2}{L}\left[k_0^{-1/2} - k^{-1/2}\right]}{4c^2\left((k^{1/4}-1)^2\right) + \frac{2c\beta}{\sqrt{L}}\left(k^{1/4}-1\right)} \tag{104}$$

for $\alpha = 3/4$.