# Supplementary material

## A. Explaining the Model of Stochastic Delayed Oracle

In Section 2, we model the delayed gradient oracle as follows,

$$\mathbf{g}_{t-\tau_t} = \nabla f(\mathbf{x}_{t-\tau_t}) + \boldsymbol{\xi}_t \ ,$$

where $\mathbb{E}[\boldsymbol{\xi}_t|\mathbf{x}_t] = 0$, as well as $\mathbb{E}[\boldsymbol{\xi}_t|\mathbf{w}_t] = 0$

Recall that standard stochastic optimization problems in ML can be described as follows,

$$\min_{\mathbf{x}\in\mathcal{K}} f(\mathbf{x}) := \mathbb{E}_{z\sim\mathcal{D}}[f(\mathbf{x}; z)] \ ,$$

when we assume we have an access to i.i.d. samples $z_1, z_2 \dots z_T \sim \mathcal{D}$, which can be used to compute gradient estimates. Thus, using this formulation, in the delayed setting we can assume that,

$$\mathbf{g}_{t-\tau_t} = \nabla f(\mathbf{x}_{t-\tau_t}; z_t)$$

where $z_t$ is the random sample that is used by the (possibly stale) machine that provides $\mathbf{g}_{t-\tau_t}$, which is the gradient estimate that we use during our $t$'th update.

Since $z_1, z_2 \dots z_t$ are i.i.d., and since $\mathbf{w}_t$ and $\mathbf{x}_t$ depend only on $\mathbf{w}_1, z_1, z_2 \dots z_{t-1}$, which are independent of $z_t$, we have,

$$\begin{aligned}\mathbb{E}[\mathbf{g}_{t-\tau_t}|\mathbf{x}_t] &= \mathbb{E}\left[\mathbb{E}_{z_t}\left(\nabla f(\mathbf{x}_{t-\tau_t}; z_t)|\mathbf{x}_t, \mathbf{x}_{t-\tau_t}\right)\Big|\mathbf{x}_t\right] \\ &= \mathbb{E}\left[\nabla f(\mathbf{x}_{t-\tau_t})|\mathbf{x}_t\right] \ , \end{aligned} \tag{8}$$

where we have used the law of total expectation.

Similarly,

$$\begin{aligned}\mathbb{E}[\mathbf{g}_{t-\tau_t}|\mathbf{w}_t] &= \mathbb{E}\left[\mathbb{E}_{z_t}\left(\nabla f(\mathbf{x}_{t-\tau_t}; z_t)|\mathbf{w}_t, \mathbf{x}_{t-\tau_t}\right)\Big|\mathbf{w}_t\right] \\ &= \mathbb{E}\left[\nabla f(\mathbf{x}_{t-\tau_t})|\mathbf{w}_t\right] \ . \end{aligned} \tag{9}$$

Thus, the above Equations immediately imply $\mathbb{E}[\boldsymbol{\xi}_t|\mathbf{x}_t] = 0$, as well as $\mathbb{E}[\boldsymbol{\xi}_t|\mathbf{w}_t] = 0$.

## B. Proof of Theorem 3.1

*Proof.* First, note that from the definition of $\mathbf{x}_t$,

$$\alpha_t(\mathbf{x}_t - \mathbf{w}_t) = \alpha_{1:t-1}(\mathbf{x}_{t-1} - \mathbf{x}_t) \ ,$$

where we define $\alpha_{1:0} = 0$ and $\mathbf{x}_0$ is an arbitrary element in $\mathcal{K}$. Now, we use the standard gradient inequality to obtain:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t(f(\mathbf{x}_t)-f(\mathbf{w}^*))\right] &\leq \mathbb{E}\left[\sum_{t=1}^{T}\alpha_t\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t-\mathbf{w}^*)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\alpha_t\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t-\mathbf{w}_t)+\sum_{t=1}^{T}\alpha_t\nabla f(\mathbf{x}_t)^\top(\mathbf{w}_t-\mathbf{w}^*)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\alpha_{1:t-1}\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_{t-1}-\mathbf{x}_t)+\sum_{t=1}^{T}\alpha_t\nabla f(\mathbf{x}_{t-\tau_t})^\top(\mathbf{w}_t-\mathbf{w}^*)\right] \\
&\quad + \mathbb{E}\left[\alpha_t(\nabla f(\mathbf{x}_t)-\nabla f(\mathbf{x}_{t-\tau_t}))^\top(\mathbf{w}_t-\mathbf{w}^*)\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T}\alpha_{1:t-1}\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_{t-1}-\mathbf{x}_t)+\sum_{t=1}^{T}\alpha_t\mathbf{g}_{t-\tau_t}^\top(\mathbf{w}_t-\mathbf{w}^*)\right] \\
&\quad + \mathbb{E}\left[\sum_{t=1}^{T}\alpha_t\|\nabla f(\mathbf{x}_t)-\nabla f(\mathbf{x}_{t-\tau_t})\|\|\mathbf{w}_t-\mathbf{w}^*\|\right], \qquad (10)
\end{aligned}$$

where the third line uses the note above $\alpha_t(\mathbf{x}_t - \mathbf{w}_t) = \alpha_{1:t-1}(\mathbf{x}_{t-1} - \mathbf{x}_t)$, and the last is due to Cauchy-Schwarzt inequality and the definition of $\mathbf{g}_{t-\tau_t}$.

The second term of Equation (10) is bounded by the OCO algorithm regret. Focusing on the last term, we wish to apply smoothness assumption to bound $\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-\tau_t})\|$. For this purpose, we will examine the difference in iterate average $\|\mathbf{x}_t - \mathbf{x}_{t-\tau_t}\|$, and show that $\|\mathbf{x}_t - \mathbf{x}_{t-\tau_t}\| \leq O(\tau_t/t)$.

**Bounding $\|\mathbf{x}_t - \mathbf{x}_{t-\tau_t}\|$ :** Let $\mathbf{z}$ be the tail average of $\mathbf{x}_t$ after $t - \tau_t$, i.e.

$$\mathbf{z} := \frac{1}{\alpha_{t-\tau_t+1:t}}\sum_{i=t-\tau_t+1}^{t}\alpha_i\mathbf{w}_i\ .$$

Clearly $\mathbf{z} \in \mathcal{K}$ and the following holds,

$$\begin{aligned}
\alpha_{1:t}\mathbf{x}_t &= \sum_{i=1}^{t}\alpha_i\mathbf{w}_i \\
&= \sum_{i=1}^{t-\tau_t}\alpha_i\mathbf{w}_i + \sum_{i=t-\tau_t+1}^{t}\alpha_i\mathbf{w}_i \\
&= \alpha_{1:t-\tau_t}\mathbf{x}_{t-\tau_t} + \alpha_{t-\tau_t+1:t}\mathbf{z}\ .
\end{aligned}$$

Therefore, $\alpha_{1:t-\tau_t}(\mathbf{x}_t - \mathbf{x}_{t-\tau_t}) = \alpha_{t-\tau_t+1:t}(\mathbf{z} - \mathbf{x}_t)$, and by taking $\alpha_t = t$, we obtain:

$$\begin{aligned}
\|\mathbf{x}_t - \mathbf{x}_{t-\tau_t}\| &= \frac{\alpha_{t-\tau_t+1:t}}{\alpha_{1:t-\tau_t}}\|\mathbf{z} - \mathbf{x}_t\| \\
&= \frac{\tau_t(t-\tau_t+1+t)}{(t-\tau_t)(t-\tau_t+1)}\|\mathbf{z} - \mathbf{x}_t\| \\
&= \left[\frac{2\tau_t t}{(t-\tau_t)(t-\tau_t+1)}\right]\|\mathbf{z} - \mathbf{x}_t\| + \left[\frac{\tau_t(1-\tau_t)}{(t-\tau_t)(t-\tau_t+1)}\right]\|\mathbf{z} - \mathbf{x}_t\| \\
&\leq \left[\frac{2\tau_t t}{(t-\tau_t)^2}\right]\|\mathbf{z} - \mathbf{x}_t\|\ .
\end{aligned}$$

If $t \geq 2\tau_t$ we have:

$$\|\mathbf{x}_t - \mathbf{x}_{t-\tau_t}\| \leq \frac{8\tau_t D}{t} .$$

Now Recall that the domain is bounded and therefore $\|\mathbf{x}_t - \mathbf{x}_{t-\tau_t}\| \leq D \; ; \forall t$. In addition, for $t < 2\tau_t$, we have $D < \frac{8\tau_t D}{t}$. Combining this with the above equation we conclude that,

$$\|\mathbf{x}_t - \mathbf{x}_{t-\tau_t}\| \leq \frac{8\tau_t D}{t} \qquad \forall t, \tau_t \leq t \; . \tag{11}$$

Using the property of smooth functions,

$$\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-\tau_t})\| \leq L\|\mathbf{x}_t - \mathbf{x}_{t-\tau_t}\| \leq \frac{8\tau_t LD}{t} \; . \tag{12}$$

**Final Bound :** Combining Equations (10), (12) together with $\alpha_t = t$ and $\|w_t - w^*\| \leq D$ yields:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \alpha_t (f(\mathbf{x}_t) - f(\mathbf{w}^*))\right] \leq & \mathbb{E}\left[\sum_{t=1}^{T} \alpha_{1:t-1} \nabla f(\mathbf{x}_t)^\top (\mathbf{x}_{t-1} - \mathbf{x}_t) + \sum_{t=1}^{T} \alpha_t \mathbf{g}_{t-\tau_t}^\top (\mathbf{w}_t - \mathbf{w}^*)\right] \\
& + \mathbb{E}\left[\sum_{t=1}^{T} \alpha_t \|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-\tau_t})\| \|\mathbf{w}_t - \mathbf{w}^*\|\right] \\
= & \mathbb{E}\left[\sum_{t=1}^{T} \alpha_{1:t-1} \nabla f(\mathbf{x}_t)^\top (\mathbf{x}_{t-1} - \mathbf{x}_t)\right] + \mathbb{E}[\mathrm{Reg}_T(\mathbf{w}^*)] + \sum_{t=1}^{T} 8\tau_t LD^2 \\
= & \mathbb{E}\left[\sum_{t=1}^{T} \alpha_{1:t-1} \nabla f(\mathbf{x}_t)^\top (\mathbf{x}_{t-1} - \mathbf{x}_t)\right] + \mathbb{E}[\mathrm{Reg}_T(\mathbf{w}^*)] + 8LD^2 T\mu_\tau \; .
\end{aligned}$$

where $\mu_\tau = \frac{\sum_{t=1}^{T} \tau_t}{T}$ is the average delay.

Next, we follow similar steps as in the proof of Theorem 1 of (Cutkosky, 2019). Using gradient inequality we have, $\mathbb{E}\left[\nabla f(x_t)^\top (\mathbf{x}_{t-1} - \mathbf{x}_t)\right] \leq \mathbb{E}\left[f(\mathbf{x}_{t-1}) - f(\mathbf{x}_t)\right]$. Hence,

$$\mathbb{E}\left[\sum_{t=1}^{T} \alpha_t (f(\mathbf{x}_t) - f(\mathbf{w}^*))\right] \leq \mathbb{E}\left[\sum_{t=1}^{T} \alpha_{1:t-1} (f(\mathbf{x}_{t-1}) - f(\mathbf{x}_t))\right] + \mathbb{E}[\mathrm{Reg}_T(\mathbf{w}^*)] + 8LD^2 T\mu_\tau \; .$$

By subtracting $\mathbb{E}\left[\sum_{t=1}^{T} \alpha_t f(\mathbf{x}_t)\right]$ from both sides of the equation we obtain,

$$-\alpha_{1:T}\mathbb{E}\left[f(\mathbf{w}^*))\right] \leq \mathbb{E}\left[\sum_{t=1}^{T} \alpha_{1:t-1} f(\mathbf{x}_{t-1}) - \alpha_{1:t} f(\mathbf{x}_t)\right] + \mathbb{E}[\mathrm{Reg}_T(\mathbf{w}^*)] + 8LD^2 T\mu_\tau \; .$$

Telescoping the above sum and dividing by $\alpha_{1:T} = \frac{T(T+1)}{2}$ conveys,

$$\mathbb{E}\left[f(\mathbf{x}_T) - f(\mathbf{w}^*)\right] \leq \mathbb{E}\left[\frac{2\mathrm{Reg}_T(\mathbf{w}^*)}{T^2}\right] + \frac{16LD^2\mu_\tau}{T+1} \; , \tag{13}$$

as desired. □

## C. SGD for Delayed Setting

**Lemma C.1.** *Assume that $f : \mathcal{K} \mapsto \mathbb{R}$ is L-smooth. Let the online learning algorithm, $\mathcal{A}$, be SGD algorithm, with update rule*

$$\mathbf{w}_{t+1} = \Pi_{\mathcal{K}}(\mathbf{w}_t - \eta_t \alpha_t \mathbf{g}_{t-\tau_t}) \ .$$

*Then, for $\alpha_t = 1$, $\eta_t = \frac{D}{\sqrt{t(2G^2+2\sigma^2)}}$ we obtain,*

$$\mathbb{E}\left[\sum_{t=1}^{T} \alpha_t \mathbf{g}_{t-\tau_t}^\top (\mathbf{w}_t - \mathbf{w}^*)\right] \leq \frac{3D\sqrt{T(2G^2+2\sigma^2)}}{2} \ ,$$

*while for $\alpha_t = t$, $\eta_t = \frac{D}{\sqrt{t^3(2G^2+2\sigma^2)}}$ we obtain,*

$$\mathbb{E}\left[\sum_{t=1}^{T} \alpha_t \mathbf{g}_{t-\tau_t}^\top (\mathbf{w}_t - \mathbf{w}^*)\right] \leq DT^{3/2}\sqrt{2G^2+2\sigma^2} \ .$$

*The above bounds yield,*

$$O\left(\frac{Reg_T(\mathbf{w}^*)}{\alpha_{1:T}}\right) = O\left(D\sqrt{\frac{G+\sigma}{T}}\right) \ .$$

**Remark:** Note that in both cases if we denote the effective learning rate by $\tilde{\eta}_t = \alpha_t \eta_t$, implies $\tilde{\eta}_t = \theta(1/\sqrt{t})$.

*Proof.* Using the Pythagorean Theorem we obtain,

$$\begin{aligned}\|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2 &\leq \|\mathbf{w}_t - \eta_t \alpha_t \mathbf{g}_{t-\tau_t} - \mathbf{w}^*\|^2 \\ &= \|\mathbf{w}_t - \mathbf{w}^*\|^2 - 2\eta_t \alpha_t \mathbf{g}_{t-\tau_t}^\top (\mathbf{w}_t - \mathbf{w}^*) + \eta_t^2 \alpha_t^2 \|\mathbf{g}_{t-\tau_t}\|^2 \ .\end{aligned}$$

Re-arranging,

$$\begin{aligned}2\alpha_t \mathbf{g}_{t-\tau_t}^\top (\mathbf{w}_t - \mathbf{w}^*) &\leq \frac{\|\mathbf{w}_t - \mathbf{w}^*\|^2 - \|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2}{\eta_t} + \eta_t \alpha_t^2 \|\mathbf{g}_{t-\tau_t}\|^2 \\ &\leq \frac{\|\mathbf{w}_t - \mathbf{w}^*\|^2 - \|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2}{\eta_t} + \eta_t \alpha_t^2 (2G^2 + 2\sigma^2) \ .\end{aligned}$$

Summing from $t = 1$ to $T$ in expectation and applying $\alpha_t = 1$, $\eta_t = \frac{D}{\sqrt{t(2G^2+2\sigma^2)}}$ gives,

$$\begin{aligned}2\mathbb{E}\left[\sum_{t=1}^{T} \mathbf{g}_{t-\tau_t}^\top (\mathbf{w}_t - \mathbf{w}^*)\right] &\leq D^2 \sum_{t=1}^{T} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + (2G^2 + 2\sigma^2) \sum_{t=1}^{T} \eta_t \\ &\leq \frac{D^2}{\eta_T} + (2G^2 + 2\sigma^2) \sum_{t=1}^{T} \eta_t \\ &\leq 3D\sqrt{T(2G^2 + 2\sigma^2)} \ ,\end{aligned}$$

where we define $\frac{1}{\eta_0} = 0$. In the last line we used $\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq 2\sqrt{T}$.

Summing from $t = 1$ to $T$ in expectation and applying $\alpha_t = t$, $\eta_t = \frac{D}{\sqrt{t^3(2G^2+2\sigma^2)}}$ gives,

$$\begin{aligned}
2\mathbb{E}\left[\sum_{t=1}^{T} t\mathbf{g}_{t-\tau_t}^{\top}(\mathbf{w}_t - \mathbf{w}^*)\right] &\leq D^2 \sum_{t=1}^{T}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + (2G^2 + 2\sigma^2)\sum_{t=1}^{T} t^2\eta_t \\
&\leq \frac{D^2}{\eta_T} + (2G^2 + 2\sigma^2)\sum_{t=1}^{T} t^2\eta_t \\
&\leq 2DT^{3/2}\sqrt{2G^2 + 2\sigma^2}\,,
\end{aligned}$$

where we again define $\frac{1}{\eta_0} = 0$. In the last line we used $\sum_{t=1}^{T}\sqrt{t} \leq T^{3/2}$.

□

## D. Proof of theorem 3.3

*Proof.* As stated in Theorem 3.3, under optimistic OCO algorithm, we assume,

$$\text{Reg}_T(\mathbf{w}^*) := \sum_{t=1}^{T} \alpha_t \mathbf{g}_{t-\tau_t}^\top (\mathbf{w}_t - \mathbf{w}^*) \leq O\left( D\sqrt{\sum_{T=1}^{T} \alpha_t^2 \|\mathbf{M}_{t-\tau_t} - \mathbf{g}_{t-\tau_t}\|^2} \right) \tag{14}$$

To bound the regret, we examine one summand, which depends on the difference between the hint $\mathbf{M}_{t-\tau_t} = \mathbf{g}_{t-1-\tau_{t\ 1}}$, and the received gradient $\mathbf{g}_{t-\tau_t}$.

$$\begin{aligned}
\|\mathbf{M}_{t-\tau_t} - \mathbf{g}_{t-\tau_t}\| &= \|\mathbf{g}_{t-\tau_t} - \mathbf{g}_{t-1-\tau_{t\ 1}}\| \\
&\leq \|\nabla f(\mathbf{x}_{t-\tau_t}) - \nabla f(\mathbf{x}_{t-1-\tau_{t\ 1}})\| + \|\boldsymbol{\xi}_t - \boldsymbol{\xi}_{t-1}\| \\
&\leq L\|\mathbf{x}_{t-\tau_t} - \mathbf{x}_{t-1-\tau_{t\ 1}}\| + \|\boldsymbol{\xi}_t - \boldsymbol{\xi}_{t-1}\| \\
&\leq L\|\mathbf{x}_t - \mathbf{x}_{t-1}\| + L\|\mathbf{x}_{t-1} - \mathbf{x}_{t-1-\tau_{t\ 1}}\| + L\|\mathbf{x}_t - \mathbf{x}_{t-\tau_t}\| + \|\boldsymbol{\xi}_t\| + \|\boldsymbol{\xi}_{t-1}\| \\
&\leq \frac{2LD}{t-1} + \frac{8LD\tau_{t-1}}{t-1} + \frac{8LD\tau_t}{t} + \|\boldsymbol{\xi}_t\| + \|\boldsymbol{\xi}_{t-1}\| \,,
\end{aligned}$$

where the first inequality is achieved by plugging in the equation for the gradients and triangle inequality and the second uses smoothness. The third line uses again triangle inequality, and in the forth line we plugged in Eq. (11).

Using the Root Mean Square and Arithmetic Mean inequality, i.e. $\frac{\sum_{i=1}^{n} a_i}{n} \leq \sqrt{\frac{\sum_{i=1}^{n} a_i^2}{n}}$, we obtain:

$$\begin{aligned}
\mathbb{E}[\|\mathbf{M}_{t-\tau_t} - \mathbf{g}_{t-\tau_t}\|^2] &= \left( \frac{2LD}{t-1} + \frac{8LD\tau_{t-1}}{t-1} + \frac{8LD\tau_t}{t} + \|\boldsymbol{\xi}_t\| + \|\boldsymbol{\xi}_{t-1}\| \right)^2 \\
&\leq 5\left( \frac{4L^2D^2}{(t-1)^2} + \frac{64L^2D^2\tau_{t-1}^2}{(t-1)^2} + \frac{64L^2D^2\tau_t^2}{t^2} + 2\sigma^2 \right) .
\end{aligned}$$

Therefore, for $\alpha_t = t$,

$$\begin{aligned}
\mathbb{E}[\text{Reg}_T(\mathbf{w}^*)] &\leq \mathbb{E}\left[ D\sqrt{2\sum_{t=1}^{T} \alpha_t^2 \|\mathbf{M}_{t-\tau_t} - \mathbf{g}_{t-\tau_t}\|^2} \right] \\
&\leq D\sqrt{\sum_{t=1}^{T} 160L^2D^2 + \sum_{t=1}^{T} 2560L^2D^2\tau_{t-1}^2 + 640L^2D^2\sum_{t=1}^{T}\tau_t^2 + 20\sigma^2\sum_{t=1}^{T} t^2} \\
&\leq D\sqrt{160L^2D^2T + 640L^2D^2\sum_{t=1}^{T}(4\tau_{t-1}^2 + \tau_t^2) + 20\sigma^2\sum_{t=1}^{T} t^2} \\
&\leq 13LD^2\sqrt{T} + D\sqrt{3200L^2D^2\sum_{t=1}^{T}\tau_t^2} + D\sqrt{20\sigma^2\sum_{t=1}^{T} t^2} \\
&\leq 13LD^2\sqrt{T} + 57LD^2\sqrt{T(\sigma_\tau^2 + \mu_\tau^2)} + 5D\sigma(T+1)^{3/2} \,,
\end{aligned}$$

where $\mu_\tau$ is the delay average and $\sigma_\tau^2$ is the delay variance. The third line uses $\tau_0 = 0$ to combine the sums, the forth line exploits the known inequality $\sqrt{\sum_{i=1}^{N} a_i} \leq \sum_{i=1}^{N}\sqrt{a_i}$, and in the last we plugged in $\sum_{t=1}^{T} t^2 \leq \frac{3(T+1)^3}{2}$.

Adding this to Equation (13), concludes the proof:

$$\mathbb{E}\left[f(\mathbf{x}_T) - f(\mathbf{w}^*)\right] = O\left( \frac{LD^2(1+\sqrt{\sigma_\tau^2 + \mu_\tau^2})}{T^{3/2}} + \frac{\sigma D}{\sqrt{T}} + \frac{LD^2\mu_\tau}{T} \right)$$

□

## E. Proof of Theorem 4.2

*Proof.* First we state a technical lemma that will be used throughout the proof of Theorem 4.2. Its proof is given in Section E.1.

**Lemma E.1.** *For any $t \geq 1$ let $\alpha_t = t^2$, and $\eta_t = C\frac{1}{\alpha_{1:t}}$ where $C$ is some constant. Also define $\alpha_0 := \alpha_1$ and $\eta_0 := \eta_1$. Then, the following holds $\forall 0 \leq t \leq s$,*

$$\alpha_t\eta_t \geq \alpha_s\eta_s \; ; \quad \& \quad \alpha_t^2\eta_t \leq 4\alpha_{t-1}^2\eta_{t-1}$$

In addition, we require the following Lemma (see proof in Section E.3),

**Lemma E.2.** *Under the same conditions of Theorem 4.2, the following holds,*

$$\sum_{t=1}^{T}\alpha_t(\mathbf{x}_t - \mathbf{w}^*)^\top\mathbf{g}_t \leq \underbrace{\sum_{t=1}^{T}\frac{\alpha_t^2\eta_t}{2}\|\mathbf{g}_t - \mathbf{M}_t\|^2}_{(i)} + 2\sum_{t=1}^{T-1}\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)\|\mathbf{x}_{t+1} - \mathbf{w}^*\|^2 + 2\sum_{t=1}^{T-1}\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)\|\mathbf{x}_{t+1} - \mathbf{y}_t\|^2 + \frac{1}{\eta_1}D^2 \tag{15}$$

Next, we relate term (i) to $\sum_{t=1}^{T}\alpha_t(\mathbf{x}_t - \mathbf{w}^*)^\top\nabla f(\mathbf{x}_t)$. To do so, we require the following lemma,

**Lemma E.3.** *Let $f : \mathcal{K} \mapsto \mathbb{R}$ be an $L$-smooth function, then,*

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2 \leq L(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^\top(\mathbf{x} - \mathbf{y})$$

*Concretely, if $\mathbf{y} \in \arg\min_{x\in\mathcal{K}} f(\mathbf{x})$ then for any $\mathbf{x} \in \mathcal{K}$ we have*

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2 \leq L\nabla f(\mathbf{x})^\top(\mathbf{x} - \mathbf{y})$$

*Proof of Lemma E.3.* The first part is proven in (Needell et al., 2013). For the second part, notice that if $\mathbf{y} \in \arg\min_{\mathbf{x}\in\mathcal{K}} f(\mathbf{x})$ then optimality condition imply that $\nabla f(\mathbf{y})^\top(\mathbf{x} - \mathbf{y}) \geq 0 \; ; \; \forall \mathbf{x} \in \mathcal{K}$. Combining this with the first part of the lemma establishes the second part. □

Using Lemma E.3 we obtain,

$$\begin{aligned}
\|\mathbf{g}_t - \mathbf{M}_t\|^2 =& \|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1}) + \boldsymbol{\xi}_t - \boldsymbol{\xi}_{t-1}\|^2 \\
=& \|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{w}^*) + \nabla f(\mathbf{w}^*) - \nabla f(\mathbf{x}_{t-1}) + \boldsymbol{\xi}_t - \boldsymbol{\xi}_{t-1}\|^2 \\
\leq& 4\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{w}^*)\|^2 + 4\|\nabla f(\mathbf{x}_{t-1}) - \nabla f(\mathbf{w}^*)\|^2 + 4\|\boldsymbol{\xi}_t\|^2 + 4\|\boldsymbol{\xi}_{t-1}\|^2 \\
\leq& 4L\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{w}^*) + 4L\nabla f(\mathbf{x}_{t-1})^\top(\mathbf{x}_{t-1} - \mathbf{w}^*) + 4\|\boldsymbol{\xi}_t\|^2 + 4\|\boldsymbol{\xi}_{t-1}\|^2
\end{aligned} \tag{16}$$

where we used $\frac{\sum_{i=1}^{n} a_i}{n} \leq \sqrt{\frac{\sum_{i=1}^{n} a_i^2}{n}}$ in the third line.

Thus, we can bound (i) as follows,

$$\begin{aligned}
\text{(i)} :=& \sum_{t=1}^{T}\frac{\alpha_t^2\eta_t}{2}\|\mathbf{g}_t - \mathbf{M}_t\|^2 \\
\leq& 2L\sum_{t=1}^{T}\alpha_t^2\eta_t\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{w}^*) + 2L\sum_{t=1}^{T}\alpha_t^2\eta_t\nabla f(\mathbf{x}_{t-1})^\top(\mathbf{x}_{t-1} - \mathbf{w}^*) + 2\sum_{t=1}^{T}\alpha_t^2\eta_t(\|\boldsymbol{\xi}_t\|^2 + \|\boldsymbol{\xi}_{t-1}\|^2) \\
\leq& 2L\sum_{t=1}^{T}\alpha_t^2\eta_t\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{w}^*) + 8L\sum_{t=1}^{T}\alpha_{t-1}^2\eta_{t-1}\nabla f(\mathbf{x}_{t-1})^\top(\mathbf{x}_{t-1} - \mathbf{w}^*) + 2\sum_{t=1}^{T}\|\boldsymbol{\xi}_t\|^2(\alpha_t^2\eta_t + \alpha_{t+1}^2\eta_{t+1}) \\
\leq& 10L\sum_{t=0}^{T}\alpha_t^2\eta_t\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{w}^*) + 10\sum_{t=1}^{T}\alpha_t^2\eta_t\|\boldsymbol{\xi}_t\|^2
\end{aligned} \tag{17}$$

where we define $\boldsymbol{\xi}_0 = 0$. The third and forth line use $\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{w}^*) \geq 0$, which holds due to convexity of $f$ and optimality of $\mathbf{w}^*$, together with Lemma E.1.

Now, lets define, $t^* := \min\{t : 10L\alpha_t^2\eta_t \leq \frac{1}{2}\alpha_t\}$. Note that according to Lemma E.1 $\alpha_t\eta_t$ is monotonic decreasing and therefore $\forall t \geq t^*;\ 10L\alpha_t^2\eta_t \leq \frac{1}{2}\alpha_t$. Using this together with Eq. (17), as well as using the convexity of $f$ that implies $\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{w}^*) \geq 0\,; \forall t$ gives,

$$\begin{aligned}
\text{(i)} &\leq 10L\sum_{t=0}^{T}\alpha_t^2\eta_t\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - x^*) + 10\sum_{t=1}^{T}\alpha_t^2\eta_t\|\boldsymbol{\xi}_t\|^2 \\
&= 10L\sum_{t=0}^{t^*-1}\alpha_t^2\eta_t\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{w}^*) + 10L\sum_{t=t^*}^{T}\alpha_t^2\eta_t\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{w}^*) + 10\sum_{t=1}^{T}\alpha_t^2\eta_t\|\boldsymbol{\xi}_t\|^2 \\
&\leq 10L\sum_{t=0}^{t^*-1}\alpha_t^2\eta_t\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{w}^*) + \frac{1}{2}\sum_{t=t^*}^{T}\alpha_t\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{w}^*) + 10\sum_{t=1}^{T}\alpha_t^2\eta_t\|\boldsymbol{\xi}_t\|^2 \\
&\leq 10L\sum_{t=0}^{t^*-1}\alpha_t^2\eta_t\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{w}^*) + \frac{1}{2}\sum_{t=1}^{T}\alpha_t\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{w}^*) + 10\sum_{t=1}^{T}\alpha_t^2\eta_t\|\boldsymbol{\xi}_t\|^2\,, \qquad (18)
\end{aligned}$$

where in the last line we use again the fact that $\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{w}^*) \geq 0\,; \forall t$.

Plugging the above into Eq. (15) and re-arranging we obtain,

$$\begin{aligned}
\sum_{t=1}^{T}\alpha_t(\mathbf{x}_t - \mathbf{w}^*)^\top\mathbf{g}_t &= \sum_{t=1}^{T}\alpha_t(\mathbf{x}_t - \mathbf{w}^*)^\top\nabla f(\mathbf{x}_t) + \sum_{t=1}^{T}\alpha_t(\mathbf{x}_t - \mathbf{w}^*)^\top\boldsymbol{\xi}_t \\
&\leq 10L\sum_{t=0}^{t^*-1}\alpha_t^2\eta_t\nabla f(x_t)^\top(\mathbf{x}_t - \mathbf{w}^*) + 2\sum_{t=1}^{T-1}\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)\|\mathbf{x}_{t+1} - \mathbf{w}^*\|^2 + \frac{1}{\eta_1}D^2 \\
&\quad + 2\sum_{t=1}^{T-1}\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)\|\mathbf{x}_{t+1} - \mathbf{y}_t\|^2 + 10\sum_{t=1}^{T}\alpha_t^2\eta_t\|\boldsymbol{\xi}_t\|^2 + \frac{1}{2}\sum_{t=1}^{T}\alpha_t(\mathbf{x}_t - \mathbf{w}^*)^\top\nabla f(\mathbf{x}_t)\,.
\end{aligned}$$

which implies,

$$\begin{aligned}
\frac{1}{2}\sum_{t=1}^{T}\alpha_t(\mathbf{x}_t - \mathbf{w}^*)^\top\nabla f(\mathbf{x}_t) &\leq 10L\sum_{t=0}^{t^*-1}\alpha_t^2\eta_t\nabla f(x_t)^\top(\mathbf{x}_t - \mathbf{w}^*) + 2\sum_{t=1}^{T-1}\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)\|\mathbf{x}_{t+1} - \mathbf{w}^*\|^2 + \frac{1}{\eta_1}D^2 \\
&\quad + 2\sum_{t=1}^{T-1}\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)\|\mathbf{x}_{t+1} - \mathbf{y}_t\|^2 + 10\sum_{t=1}^{T}\alpha_t^2\eta_t\|\boldsymbol{\xi}_t\|^2 - \sum_{t=1}^{T}\alpha_t(\mathbf{x}_t - \mathbf{w}^*)^\top\boldsymbol{\xi}_t\,.
\end{aligned}$$

Combining the above with the strong-convexity of $f(\cdot)$ implies

$$\begin{aligned}
\sum_{t=1}^{T}\alpha_t(f(\mathbf{x}_t) - f(\mathbf{w}^*)) &\leq \sum_{t=1}^{T}\left(\alpha_t(\mathbf{x}_t - \mathbf{w}^*)^\top\nabla f(\mathbf{x}_t) - \frac{\alpha_t H}{2}\|\mathbf{x}_t - \mathbf{w}^*\|^2\right) \\
&\leq \underbrace{20L\sum_{t=0}^{t^*-1}\alpha_t^2\eta_t\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{w}^*)}_{\text{(A)}} + \underbrace{4\sum_{t=1}^{T-1}\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} - \frac{H}{8}\alpha_{t+1}\right)\|\mathbf{x}_{t+1} - \mathbf{w}^*\|^2}_{\text{(B)}} \\
&\quad + \underbrace{4\sum_{t=1}^{T-1}\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)\|\mathbf{x}_{t+1} - \mathbf{y}_t\|^2}_{\text{(C)}} + \frac{2}{\eta_1}D^2 + 20\sum_{t=1}^{T}\alpha_t^2\eta_t\|\boldsymbol{\xi}_t\|^2 - 2\sum_{t=1}^{T}\alpha_t(\mathbf{x}_t - \mathbf{w}^*)^\top\boldsymbol{\xi}_t\,.
\end{aligned} \qquad (19)$$

Next we bound the above three terms.

**Bounding (A)** Using the expression for $\eta_t$ and assumption 1,3 we have,

$$\begin{aligned}
\text{(A)} &:= 10L \sum_{t=0}^{t\ -1} \alpha_t^2 \eta_t \nabla f(\mathbf{x}_t)^\top (\mathbf{x}_t - \mathbf{w}^*) \\
&\leq \frac{80LGD}{H} \sum_{t=0}^{t\ -1} \frac{\alpha_t^2}{\alpha_{1:t}} \\
&\leq \frac{480LGD}{H} \sum_{t=0}^{t\ -1} (t+1) \\
&\leq \frac{240LGD}{H} (t^*)^2
\end{aligned}$$

where in the two last lines we have used $\alpha_t = t^2$. Now, using the weights together with the definition of $t^* := \min\{t : 10L\alpha_t^2\eta_t \leq \frac{1}{2}\alpha_t\}$ and the expression of $\eta_t$ implies,

$$t^* \leq 160 \frac{L}{H}$$

Plugging this back to the above bound on (A) we finally obtain,

$$\text{(A)} \leq 2 \cdot 10^6 GD \left(\frac{L}{H}\right)^3 \tag{20}$$

**Bounding (B)** Recalling that $\eta_{t+1} := \frac{8}{H\alpha_{1:t+1}}$ immediately implies,

$$\text{(B)} := \sum_{t=1}^{T-1} \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} - \frac{H}{8}\alpha_{t+1}\right) \|\mathbf{x}_{t+1} - \mathbf{w}^*\|^2 = 0 \tag{21}$$

**Bounding (C)** To bound term (C) we use Remark 4.1, in conjunction with the contraction property of the projection operator to obtain in expectation,

$$\mathbb{E}\left[\|\mathbf{x}_{t+1} - \mathbf{y}_t\|\right] \leq \tilde{\eta}_{t+1} \mathbb{E}\left[\|\mathbf{M}_{t+1}\|\right] \leq \tilde{\eta}_{t+1} \sqrt{2G^2 + 2\sigma^2}\ .$$

The above enables to bound term (C)

$$\begin{aligned}
\mathbb{E}\left[\text{(C)}\right] &:= \mathbb{E}\left[\sum_{t=1}^{T-1} \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) \|\mathbf{x}_{t+1} - \mathbf{y}_t\|^2\right] \\
&= \frac{H}{4} \mathbb{E}\left[\sum_{t=1}^{T-1} \alpha_{t+1} \|\mathbf{x}_{t+1} - \mathbf{y}_t\|^2\right] \\
&\leq \frac{H(G^2+\sigma^2)}{2} \sum_{t=1}^{T-1} \alpha_{t+1} \tilde{\eta}_{t+1}^2 \\
&= \frac{32(G^2+\sigma^2)}{H} \sum_{t=1}^{T-1} \alpha_{t+1} \left(\frac{\alpha_{t+1}}{\alpha_{1:t+1}}\right)^2 \\
&\leq \frac{1200(G^2+\sigma^2)}{H} \sum_{t=1}^{T-1} \frac{(t+1)^6}{(t(t+1)(2t+3))^2} \\
&\leq \frac{1200(G^2+\sigma^2)}{H} T
\end{aligned} \tag{22}$$

where we used $\alpha_t = t^2$.

**Final Bound** Combining the bounds in Equations (20) (21) and (22) and assumption 2 into Eq. (19) and taking expectation implies,

$$\mathbb{E}\left[\sum_{t=1}^{T}\alpha_t(f(\mathbf{x}_t)-f(\mathbf{w}^*))\right]\leq O\left(GD\left(L/H\right)^3+(G^2+\sigma^2)/H)T+HD^2+(\sigma^2/H)T^2\right) \tag{23}$$

Recalling $\bar{\mathbf{x}}_T \propto \sum_{t=1}^{T}\alpha_t\mathbf{x}_t$ and using Jensen's inequality established the theorem. □

### E.1. Proof of Lemma E.1

*Proof of Lemma E.1.* When $s > t$,

$$\begin{gathered}
\alpha_t\eta_t \geq \alpha_s\eta_s \\
\Updownarrow \\
\frac{1}{\alpha_t\eta_t} \leq \frac{1}{\alpha_s\eta_s} \\
\Updownarrow \\
2t+3+\frac{1}{t} \leq 2s+3+\frac{1}{s} \\
\Updownarrow \\
2(t-s) \leq \frac{t-s}{ts} \Updownarrow \\
2 \geq \frac{1}{ts}
\end{gathered}$$

which is true $\forall t, s \geq 1$. For $s = t$, $\alpha_t\eta_t \geq \alpha_s\eta_s$ is trivially true.

The second part follows from

$$\begin{gathered}
\alpha_t^2\eta_t \leq 4\alpha_{t-1}^2\eta_{t-1} \\
\Updownarrow \\
\frac{t^3}{(t+1)(2t+1)} \leq \frac{4(t-1)^3}{t(2t-1)}
\end{gathered}$$

which is true for $t \geq 2$. For $t = 1$, the inequality is true by definition of $\alpha_1, \eta_1, \alpha_0, \eta_0$.

□

### E.2. Proof of Remark 4.1

*Proof.* According to the update rule for $\mathbf{x}_t$ as stated in Alg. 3,

$$\begin{aligned}
\mathbf{x}_t &= \arg\min_{\mathbf{x}\in\mathcal{K}}\left[\alpha_t\mathbf{x}^\top\mathbf{M}_t+\frac{1}{2\eta_t}\|\mathbf{x}-\mathbf{y}_{t-1}\|^2\right] \\
&= \arg\min_{\mathbf{x}\in\mathcal{K}}\left[\frac{1}{2\eta_t}\left(\|\mathbf{x}\|^2-2\mathbf{x}^\top(\mathbf{y}_{t-1}-\alpha_t\eta_t\mathbf{M}_t)+\|\mathbf{y}_{t-1}\|^2\right)\right] \\
&= \arg\min_{\mathbf{x}\in\mathcal{K}}\left[\left\|\mathbf{x}-(\mathbf{y}_{t-1}-\alpha_t\eta_t\mathbf{M}_t)\right\|^2\right] \\
&= \Pi_{\mathcal{K}}\left(\mathbf{y}_{t-1}-\alpha_t\eta_t\mathbf{M}_t\right)
\end{aligned}$$

Plugging in $\eta_t = \frac{8}{H\alpha_{1:t}}$ concludes the first part of the Remark. The proof of the second part is equivalent to the above, for update rule $\mathbf{y}_t = \arg\min_{\mathbf{y}\in\mathcal{K}}\left[\alpha_t\mathbf{y}^\top\mathbf{g}_t+\frac{1}{2\eta_t}\|\mathbf{y}-\mathbf{y}_{t-1}\|^2\right]$. □

### E.3. Proof of Lemma E.2

*Proof.*

$$\sum_{t=1}^{T}\alpha_t(\mathbf{x}_t-\mathbf{w}^*)^\top\mathbf{g}_t=\sum_{t=1}^{T}\underbrace{\alpha_t(\mathbf{x}_t-\mathbf{y}_t)^\top(\mathbf{g}_t-\mathbf{M}_t)}_{\text{(A)}}+\underbrace{\alpha_t(\mathbf{x}_t-\mathbf{y}_t)^\top\mathbf{M}_t}_{\text{(B)}}+\underbrace{\alpha_t(\mathbf{y}_t-\mathbf{w}^*)^\top\mathbf{g}_t}_{\text{(C)}} \tag{24}$$

For simplicity, we will denote $D_{\mathcal{R}}(\mathbf{x},\mathbf{y}) := \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|^2$. Note that since $\frac{1}{2}\|\mathbf{x}-\mathbf{y}\|^2$ is the Bregman Divergence of $\mathcal{R}(\mathbf{x})=\frac{1}{2}\|\mathbf{x}\|^2$, inequalities of Bregman Divergence hold true. Specifically, we make use of the three point property,

$$D_{\mathcal{R}}(\mathbf{x},\mathbf{y})+D_{\mathcal{R}}(\mathbf{y},\mathbf{z})=D_{\mathcal{R}}(\mathbf{x},\mathbf{z})+\nabla_{\mathbf{z}}D_{\mathcal{R}}(\mathbf{z},\mathbf{y})^\top(\mathbf{x}-\mathbf{y})\,.$$

**Bounding (A)**

$$\begin{aligned}\sum_{t=1}^{T}\alpha_t(\mathbf{x}_t-\mathbf{y}_t)^\top(\mathbf{g}_t-\mathbf{M}_t)&\le\sum_{t=1}^{T}\alpha_t\|\mathbf{g}_t-\mathbf{M}_t\|\|\mathbf{x}_t-\mathbf{y}_t\| \quad \text{(Cauchy Schwartz Inequality)}\\&\le\sum_{t=1}^{T}\frac{\rho}{2}\|\mathbf{g}_t-\mathbf{M}_t\|^2+\frac{\alpha_t^2}{2\rho}\|\mathbf{x}_t-\mathbf{y}_t\|^2\end{aligned}$$

where the last line is due to Young's Inequality, $ab\le\inf_{\rho>0}\left(\rho a^2/2+b^2/(2\rho)\right)$.

By setting $\rho=\alpha_t^2\eta_t$, we get the following upper bound for term (A)

$$\sum_{t=1}^{T}\alpha_t(\mathbf{x}_t-\mathbf{y}_t)^\top(\mathbf{g}_t-\mathbf{M}_t)\le\sum_{t=1}^{T}\frac{\alpha_t^2\eta_t}{2}\|\mathbf{g}_t-\mathbf{M}_t\|^2+\frac{1}{2\eta_t}\|\mathbf{x}_t-\mathbf{y}_t\|^2$$

**Bounding (B)**

$$\begin{aligned}\sum_{t=1}^{T}\alpha_t(\mathbf{x}_t-\mathbf{y}_t)^\top\mathbf{M}_t&\le\sum_{t=1}^{T}\frac{1}{\eta_t}\nabla_x D_{\mathcal{R}}(\mathbf{x}_t,\mathbf{y}_{t-1})^T(\mathbf{y}_t-\mathbf{x}_t) \quad \text{(Optimality for } \mathbf{x}_t\text{)} &(25)\\&=\sum_{t=1}^{T}\frac{1}{\eta_t}\left(D_{\mathcal{R}}(\mathbf{y}_t,\mathbf{y}_{t-1})-D_{\mathcal{R}}(\mathbf{x}_t,\mathbf{y}_{t-1})-D_{\mathcal{R}}(\mathbf{y}_t,\mathbf{x}_t)\right) \quad \text{(Bregman Divergence property )} &(26)\end{aligned}$$

**Bounding (C)**

$$\begin{aligned}\sum_{t=1}^{T}\alpha_t(\mathbf{y}_t-\mathbf{w}^*)^\top\mathbf{g}_t&\le\sum_{t=1}^{T}\frac{1}{\eta_t}\nabla_x D_{\mathcal{R}}(\mathbf{y}_t,\mathbf{y}_{t-1})^T(\mathbf{w}^*-\mathbf{y}_t) \quad \text{(Optimality for } \mathbf{y}_t\text{)} &(27)\\&=\sum_{t=1}^{T}\frac{1}{\eta_t}\left(D_{\mathcal{R}}(\mathbf{w}^*,\mathbf{y}_{t-1})-D_{\mathcal{R}}(\mathbf{y}_t,\mathbf{y}_{t-1})-D_{\mathcal{R}}(\mathbf{w}^*,\mathbf{y}_t)\right) \quad \text{(Bregman Divergence property )} &(28)\end{aligned}$$

**Final Bound** Combining the bounds in Equations (25) (26) and (28) into Eq. (24) implies,

$$\begin{aligned}
\sum_{t=1}^{T} \alpha_t (\mathbf{x}_t - \mathbf{w}^*)^\top \mathbf{g}_t &\leq \sum_{t=1}^{T} \frac{\alpha_t^2 \eta_t}{2} \|\mathbf{g}_t - \mathbf{M}_t\|^2 + \frac{1}{2\eta_t} \|\mathbf{x}_t - \mathbf{y}_t\|^2 \\
&\quad + \frac{1}{\eta_t} \left( D_{\mathcal{R}}(\mathbf{w}^*, \mathbf{y}_{t-1}) - D_{\mathcal{R}}(\mathbf{w}^*, \mathbf{y}_t) - D_{\mathcal{R}}(\mathbf{x}_t, \mathbf{y}_{t-1}) - D_{\mathcal{R}}(\mathbf{y}_t, \mathbf{x}_t) \right) \\
&= \sum_{t=1}^{T} \frac{\alpha_t^2 \eta_t}{2} \|\mathbf{g}_t - \mathbf{M}_t\|^2 + \frac{1}{2\eta_t} \|\mathbf{x}_t - \mathbf{y}_t\|^2 \\
&\quad + \frac{1}{\eta_t} \left( D_{\mathcal{R}}(\mathbf{w}^*, \mathbf{y}_{t-1}) - D_{\mathcal{R}}(\mathbf{w}^*, \mathbf{y}_t) - \frac{1}{2} \left( \|\mathbf{x}_t - \mathbf{y}_t\|^2 + \|\mathbf{x}_t - \mathbf{y}_{t-1}\|^2 \right) \right) \\
&\leq \sum_{t=1}^{T} \frac{\alpha_t^2 \eta_t}{2} \|\mathbf{g}_t - \mathbf{M}_t\|^2 + \sum_{t=1}^{T-1} \left( \frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) D_{\mathcal{R}}(\mathbf{w}^*, \mathbf{y}_t) + \frac{1}{\eta_1} D^2 \\
&= \sum_{t=1}^{T} \frac{\alpha_t^2 \eta_t}{2} \|\mathbf{g}_t - \mathbf{M}_t\|^2 + \sum_{t=1}^{T-1} \left( \frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \|\mathbf{y}_t - \mathbf{w}^*\|^2 + \frac{1}{\eta_1} D^2 \\
&\leq \sum_{t=1}^{T} \frac{\alpha_t^2 \eta_t}{2} \|\mathbf{g}_t - \mathbf{M}_t\|^2 + 2 \sum_{t=1}^{T-1} \left( \frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \|\mathbf{x}_{t+1} - \mathbf{w}^*\|^2 \\
&\quad + 2 \sum_{t=1}^{T-1} \left( \frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \|\mathbf{x}_{t+1} - \mathbf{y}_t\|^2 + \frac{1}{\eta_1} D^2
\end{aligned} \tag{29}$$

where in the third line we used $D_{\mathcal{R}}(\mathbf{x}, \mathbf{y}) := \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|^2$.

This establishes the lemma. □

## F. Proof of Theorem 4.3

Recall that under assumptions 1,2 from Section 2, we can show the following bound on the expected norm of the gradients, $\mathbb{E}\|\mathbf{g}_t\| \leq \tilde{G} = \sqrt{2G^2 + 2\sigma^2}$. Nevertheless, working with this in-expectation bound makes the proof a bit cumbersome. Therefore, to simplify the analysis, from now on we will assume that $\|\mathbf{g}\| \leq \tilde{G}$ with probability 1. Both assumptions lead to exactly the same in expectation guarantees as those we state in Theorem 4.3.

*Proof.* We first Note that Lemma E.2 is true for the delayed setting as well,

$$\sum_{t=1}^{T} \alpha_t (\mathbf{x}_t - \mathbf{w}^*)^\top \mathbf{g}_{t-\tau_t} \leq \underbrace{\sum_{t=1}^{T} \frac{\alpha_t^2 \eta_t}{2} \|\mathbf{g}_{t-\tau_t} - \mathbf{M}_{t-\tau_t}\|^2}_{\text{(i)}} + 2\sum_{t=1}^{T-1} \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) \|\mathbf{x}_{t+1} - \mathbf{w}^*\|^2 + 2\sum_{t=1}^{T-1} \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) \|\mathbf{x}_{t+1} - \mathbf{y}_t\|^2 + \frac{1}{\eta_1} D^2 \tag{30}$$

Next, we relate term (i) to $\sum_{t=1}^{T} \alpha_t (\mathbf{x}_t - \mathbf{w}^*)^\top \nabla f(\mathbf{x}_t)$.

$$\begin{aligned}
\|\mathbf{g}_{t-\tau_t} - \mathbf{M}_{t-\tau_t}\|^2 &= \|\nabla f(\mathbf{x}_{t-\tau_t}) - \nabla f(\mathbf{x}_{t-1-\tau_{t\ 1}}) + \boldsymbol{\xi}_{t-\tau_t} - \boldsymbol{\xi}_{t-1-\tau_{t\ 1}}\|^2 \\
&= \|\nabla f(\mathbf{x}_{t-\tau_t}) - \nabla f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t) + \nabla f(\mathbf{x}_{t-1}) - \nabla f(\mathbf{x}_{t-1}) - \nabla f(\mathbf{x}_{t-1-\tau_{t\ 1}}) + \boldsymbol{\xi}_{t-\tau_t} - \boldsymbol{\xi}_{t-1-\tau_{t\ 1}}\|^2 \\
&\leq 5\|\nabla f(\mathbf{x}_{t-\tau_t}) - \nabla f(\mathbf{x}_t)\|^2 + 5\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|^2 + 5\|\nabla f(\mathbf{x}_{t-1}) - \nabla f(\mathbf{x}_{t-1-\tau_{t\ 1}})\|^2 \\
&\quad + 5(\|\boldsymbol{\xi}_{t-\tau_t}\|^2 + \|\boldsymbol{\xi}_{t-1-\tau_{t\ 1}}\|^2) \\
&\leq 5\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|^2 + 5L^2\|\mathbf{x}_t - \mathbf{x}_{t-\tau_t}\|^2 + 5L^2\|\mathbf{x}_{t-1} - \mathbf{x}_{t-1-\tau_{t\ 1}}\|^2 + 5(\|\boldsymbol{\xi}_{t-\tau_t}\|^2 + \|\boldsymbol{\xi}_{t-1-\tau_{t\ 1}}\|^2) \\
&\leq 10\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{w}^*)\|^2 + 10\|\nabla f(\mathbf{x}_{t-1}) - \nabla f(\mathbf{w}^*)\|^2 + 5L^2\|\mathbf{x}_t - \mathbf{x}_{t-\tau_t}\|^2 + 5L^2\|\mathbf{x}_{t-1} - \mathbf{x}_{t-1-\tau_{t\ 1}}\|^2 \\
&\quad + 5(\|\boldsymbol{\xi}_{t-\tau_t}\|^2 + \|\boldsymbol{\xi}_{t-1-\tau_{t\ 1}}\|^2) \\
&\leq 10L\nabla f(\mathbf{x}_t)^\top (\mathbf{x}_t - \mathbf{w}^*) + 10L\nabla f(\mathbf{x}_{t-1})^\top (\mathbf{x}_{t-1} - \mathbf{w}^*) + 5L^2\|\mathbf{x}_t - \mathbf{x}_{t-\tau_t}\|^2 + 5L^2\|\mathbf{x}_{t-1} - \mathbf{x}_{t-1-\tau_{t\ 1}}\|^2 \\
&\quad + 5(\|\boldsymbol{\xi}_{t-\tau_t}\|^2 + \|\boldsymbol{\xi}_{t-1-\tau_{t\ 1}}\|^2)
\end{aligned} \tag{31}$$

where we used $\frac{\sum_{i=1}^{n} a_i}{n} \leq \sqrt{\frac{\sum_{i=1}^{n} a_i^2}{n}}$ and smoothness. The last line is due to Lemma E.3.

Next, we wish to bound $\|\mathbf{x}_t - \mathbf{x}_{t-\tau_t}\|$. Using Remark 4.1 with the contraction property of the projection operator, we obtain for $\alpha_t = t^2$,

$$\begin{aligned}
\|\mathbf{y}_t - \mathbf{y}_{t-1}\| &\leq \tilde{\eta}_t \|\mathbf{g}_t\| \ \leq \frac{24\tilde{G}}{Ht}\,, \\
\|\mathbf{x}_t - \mathbf{y}_t\| &\leq \tilde{\eta}_t \|\mathbf{M}_t - \mathbf{g}_t\| \leq \frac{48\tilde{G}}{Ht}\,, \\
\|\mathbf{x}_{t-\tau_t} - \mathbf{y}_{t-\tau_t}\| &\leq \tilde{\eta}_{t-\tau_t} \|\mathbf{M}_{t-\tau_t} - \mathbf{g}_{t-\tau_t}\| \leq \frac{48\tilde{G}}{H(t-\tau_t)}\,.
\end{aligned}$$

Combining all of the above, we obtain

$$\begin{aligned}\|\mathbf{x}_t - \mathbf{x}_{t-\tau_t}\| &\leq \|\mathbf{x}_t - \mathbf{y}_t\| + \|\mathbf{y}_{t-\tau_t} - \mathbf{x}_{t-\tau_t}\| + \sum_{s=t-\tau_t}^{t} \|\mathbf{y}_s - \mathbf{y}_{s-1}\| \\ &\leq \frac{48\tilde{G}}{Ht} + \frac{48\tilde{G}}{H(t-\tau_t)} + \frac{24\tilde{G}}{H}\sum_{s=t-\tau_t}^{t}\frac{1}{s} \\ &\leq \frac{48\tilde{G}}{Ht} + \frac{24\tilde{G}(\tau_t+2)}{H(t-\tau_t)} \\ &\leq \frac{48\tilde{G}}{Ht} + \frac{48\tilde{G}(\tau_t+1)}{Ht} \\ &\leq \frac{48\tilde{G}(\tau_t+2)}{Ht}\end{aligned}$$

where in the forth line we assume $2\tau_t + 2 \leq t$. When $t < 2\tau_t + 2$, we have,

$$\frac{2G}{H} \leq \frac{24\tilde{G}}{H} \leq \frac{48\tilde{G}(\tau_t+2)}{Ht}\,.$$

Recalling that for strongly convex functions we have,

$$\|\mathbf{x} - \mathbf{y}\| \leq \frac{\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|}{H} \leq \frac{2G}{H}\,, \tag{32}$$

which we also prove in Subsection F.1 for completeness, we obtain,

$$\|\mathbf{x}_t - \mathbf{x}_{t-\tau_t}\| \leq \frac{48\tilde{G}(\tau_t+2)}{Ht} \quad \forall t\,. \tag{33}$$

Combining Equations (31) and (33), we obtain

$$\begin{aligned}\text{(i)} &:= \sum_{t=1}^{T}\frac{\alpha_t^2\eta_t}{2}\|\mathbf{g}_{t-\tau_t} - \mathbf{M}_{t-\tau_t}\|^2 \\ &\leq 5L\sum_{t=1}^{T}\alpha_t^2\eta_t\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{w}^*) + 5L\sum_{t=1}^{T}\alpha_t^2\eta_t\nabla f(\mathbf{x}_{t-1})^\top(\mathbf{x}_{t-1} - \mathbf{w}^*) \\ &\quad + 3L^2\sum_{t=1}^{T}\alpha_t^2\eta_t\|\mathbf{x}_t - \mathbf{x}_{t-\tau_t}\|^2 + 3L^2\sum_{t=1}^{T}\alpha_t^2\eta_t\|\mathbf{x}_{t-1} - \mathbf{x}_{t-1-\tau_{t\ 1}}\|^2 + 3\sum_{t=1}^{T}\alpha_t^2\eta_t(\|\boldsymbol{\xi}_{t-\tau_t}\|^2 + \|\boldsymbol{\xi}_{t-1-\tau_{t\ 1}}\|^2) \\ &\leq 5L\sum_{t=1}^{T}\alpha_t^2\eta_t\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{w}^*) + 20L\sum_{t=1}^{T}\alpha_{t-1}^2\eta_{t-1}\nabla f(\mathbf{x}_{t-1})^\top(\mathbf{x}_{t-1} - \mathbf{w}^*) \\ &\quad + 12L^2\sum_{t=1}^{T}\frac{2\cdot 24^3\tilde{G}^2(\tau_t+2)^2}{H^3 t} + 3L^2\eta_1 D + 12L^2\sum_{t=2}^{T}\frac{2\cdot 24^3\tilde{G}^2(\tau_{t-1}+2)^2}{H^3(t-1)} \\ &\quad + 3\sum_{t=1}^{T}\alpha_t^2\eta_t(\|\boldsymbol{\xi}_{t-\tau_t}\|^2 + \|\boldsymbol{\xi}_{t-1-\tau_{t\ 1}}\|^2) \\ &\leq 25L\sum_{t=1}^{T}\alpha_t^2\eta_t\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{w}^*) + \frac{2\cdot 24^4L^2\tilde{G}^2T(\sigma_\tau^2 + \mu_\tau^2 + 4\mu_\tau + 4)}{H^3} + 3L^2\eta_1 D \\ &\quad + 3\sum_{t=1}^{T}\alpha_t^2\eta_t(\|\boldsymbol{\xi}_{t-\tau_t}\|^2 + \|\boldsymbol{\xi}_{t-1-\tau_{t\ 1}}\|^2)\end{aligned} \tag{34}$$

where in the second inequality we used Lemma E.1 with $\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{w}^*) \geq 0$; the third inequality uses again the bound $\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{w}^*) \geq 0$.

Now, lets define, $t^* := \min\{t : 25L\alpha_t^2\eta_t \leq \frac{1}{2}\alpha_t\}$. Similarly to the proof of Theorem 4.2, since $\alpha_t\eta_t$ is monotonic decreasing, $\forall t \geq t^*;\ 25L\alpha_t^2\eta_t \leq \frac{1}{2}\alpha_t$. Using this in (34) together with the convexity of $f$ that implies $\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{w}^*) \geq 0\ ; \forall t$ gives,

$$
\begin{aligned}
\text{(i)} \leq\ & 25L\sum_{t=1}^{T}\alpha_t^2\eta_t\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{w}^*) + 3L^2\eta_1 D + \frac{2\cdot 24^4 L^2\tilde{G}^2T(\sigma_\tau^2 + \mu_\tau^2 + 4\mu_\tau + 4)}{H^3} \\
& + 3\sum_{t=1}^{T}\alpha_t^2\eta_t(\|\boldsymbol{\xi}_{t-\tau_t}\|^2 + \|\boldsymbol{\xi}_{t-1-\tau_t\ \ 1}\|^2) \\
=\ & 25L\sum_{t=1}^{t\ -1}\alpha_t^2\eta_t\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{w}^*) + 25L\sum_{t=t}^{T}\alpha_t^2\eta_t\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{w}^*) + 3L^2\eta_1 D \\
& + \frac{2\cdot 24^4 L^2\tilde{G}^2T(\sigma_\tau^2 + \mu_\tau^2 + 4\mu_\tau + 4)}{H^3} + 3\sum_{t=1}^{T}\alpha_t^2\eta_t(\|\boldsymbol{\xi}_{t-\tau_t}\|^2 + \|\boldsymbol{\xi}_{t-1-\tau_t\ \ 1}\|^2) \\
\leq & 25L\sum_{t=1}^{t\ -1}\alpha_t^2\eta_t\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{w}^*) + \frac{1}{2}\sum_{t=t}^{T}\alpha_t\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{w}^*) + 3L^2\eta_1 D \\
& + \frac{2\cdot 24^4 L^2\tilde{G}^2T(\sigma_\tau^2 + \mu_\tau^2 + 4\mu_\tau + 4)}{H^3} + 3\sum_{t=1}^{T}\alpha_t^2\eta_t(\|\boldsymbol{\xi}_{t-\tau_t}\|^2 + \|\boldsymbol{\xi}_{t-1-\tau_t\ \ 1}\|^2) \\
\leq & 25L\sum_{t=1}^{t\ -1}\alpha_t^2\eta_t\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{w}^*) + \frac{1}{2}\sum_{t=1}^{T}\alpha_t\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{w}^*) + 3L^2\eta_1 D \\
& + \frac{2\cdot 24^4 L^2\tilde{G}^2T(\sigma_\tau^2 + \mu_\tau^2 + 4\mu_\tau + 4)}{H^3} + 3\sum_{t=1}^{T}\alpha_t^2\eta_t(\|\boldsymbol{\xi}_{t-\tau_t}\|^2 + \|\boldsymbol{\xi}_{t-1-\tau_t\ \ 1}\|^2) \\
= & 25L\sum_{t=1}^{t\ -1}\alpha_t^2\eta_t\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{w}^*) + \frac{1}{2}\sum_{t=1}^{T}\alpha_t\nabla f(\mathbf{x}_{t-\tau_t})^\top(\mathbf{x}_t - \mathbf{w}^*) \\
& + \frac{1}{2}\sum_{t=1}^{T}\alpha_t(\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-\tau_t}))^\top(\mathbf{x}_t - \mathbf{w}^*) + 3L^2\eta_1 D \\
& + \frac{2\cdot 24^4 L^2\tilde{G}^2T(\sigma_\tau^2 + \mu_\tau^2 + 4\mu_\tau + 4)}{H^3} + 3\sum_{t=1}^{T}\alpha_t^2\eta_t(\|\boldsymbol{\xi}_{t-\tau_t}\|^2 + \|\boldsymbol{\xi}_{t-1-\tau_t\ \ 1}\|^2)\,, \qquad (35)
\end{aligned}
$$

where in the third inequality we use again the fact that $\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{w}^*) \geq 0\ ; \forall t$.

Plugging the above into Eq. (30) and re-arranging we obtain,

$$\begin{aligned}\sum_{t=1}^{T}\alpha_t(\mathbf{x}_t-\mathbf{w}^*)^\top\mathbf{g}_{t-\tau_t} &= \sum_{t=1}^{T}\alpha_t(\mathbf{x}_t-\mathbf{w}^*)^\top\nabla f(\mathbf{x}_{t-\tau_t})+\sum_{t=1}^{T}\alpha_t(\mathbf{x}_t-\mathbf{w}^*)^\top\boldsymbol{\xi}_t\\ &\le 25L\sum_{t=0}^{t-1}\alpha_t^2\eta_t\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t-\mathbf{w}^*)+2\sum_{t=1}^{T-1}\left(\frac{1}{\eta_{t+1}}-\frac{1}{\eta_t}\right)\|\mathbf{x}_{t+1}-\mathbf{w}^*\|^2\\ &\quad+2\sum_{t=1}^{T-1}\left(\frac{1}{\eta_{t+1}}-\frac{1}{\eta_t}\right)\|\mathbf{x}_{t+1}-\mathbf{y}_t\|^2+\frac{1}{2}\sum_{t=1}^{T}\alpha_t\nabla f(\mathbf{x}_{t-\tau_t})^\top(\mathbf{x}_t-\mathbf{w}^*)\\ &\quad+\frac{1}{2}\sum_{t=1}^{T}\alpha_t(\nabla f(\mathbf{x}_t)-\nabla f(\mathbf{x}_{t-\tau_t}))^\top(\mathbf{x}_t-\mathbf{w}^*)+3L^2\eta_1D+\frac{1}{\eta_1}D^2\\ &\quad+\frac{2\cdot 24^4L^2\tilde{G}^2T(\sigma_\tau^2+\mu_\tau^2+4\mu_\tau+4)}{H^3}+3\sum_{t=1}^{T}\alpha_t^2\eta_t(\|\boldsymbol{\xi}_{t-\tau_t}\|^2+\|\boldsymbol{\xi}_{t-1-\tau_t\ \ 1}\|^2)\,,\end{aligned}$$

which implies,

$$\begin{aligned}\frac{1}{2}\sum_{t=1}^{T}\alpha_t(\mathbf{x}_t-\mathbf{w}^*)^\top\nabla f(\mathbf{x}_{t-\tau_t}) &\le 25L\sum_{t=0}^{t-1}\alpha_t^2\eta_t\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t-\mathbf{w}^*)+2\sum_{t=1}^{T-1}\left(\frac{1}{\eta_{t+1}}-\frac{1}{\eta_t}\right)\|\mathbf{x}_{t+1}-\mathbf{w}^*\|^2\\ &\quad+2\sum_{t=1}^{T-1}\left(\frac{1}{\eta_{t+1}}-\frac{1}{\eta_t}\right)\|\mathbf{x}_{t+1}-\mathbf{y}_t\|^2+\frac{1}{2}\sum_{t=1}^{T}\alpha_t\nabla f(\mathbf{x}_{t-\tau_t})^\top(\mathbf{x}_t-\mathbf{w}^*)\\ &\quad+\frac{1}{2}\sum_{t=1}^{T}\alpha_t(\nabla f(\mathbf{x}_t)-\nabla f(\mathbf{x}_{t-\tau_t}))^\top(\mathbf{x}_t-\mathbf{w}^*)+3L^2\eta_1D+\frac{1}{\eta_1}D^2\\ &\quad+\frac{2\cdot 24^4L^2\tilde{G}^2T(\sigma_\tau^2+\mu_\tau^2+4\mu_\tau+4)}{H^3}+3\sum_{t=1}^{T}\alpha_t^2\eta_t(\|\boldsymbol{\xi}_{t-\tau_t}\|^2+\|\boldsymbol{\xi}_{t-1-\tau_t\ \ 1}\|^2)\\ &\quad-\sum_{t=1}^{T}\alpha_t(\mathbf{x}_t-\mathbf{w}^*)^\top\boldsymbol{\xi}_t\,.\end{aligned}$$

Combining the above with the strong-convexity of $f\cdot)$ implies,

$$\begin{aligned}\sum_{t=1}^{T}\alpha_t(f(\mathbf{x}_t)-f(\mathbf{w}^*)) &\le \sum_{t=1}^{T}\left(\alpha_t(\mathbf{x}_t-\mathbf{w}^*)^\top\nabla f(\mathbf{x}_t)-\frac{\alpha_tH}{2}\|\mathbf{x}_t-\mathbf{w}^*\|^2\right)\\ &= \sum_{t=1}^{T}\left(\alpha_t(\mathbf{x}_t-\mathbf{w}^*)^\top\nabla f(\mathbf{x}_{t-\tau_t})+\alpha_t(\nabla f(\mathbf{x}_t)-\nabla f(\mathbf{x}_{t-\tau_t}))^\top(\mathbf{x}_t-\mathbf{w}^*)-\frac{\alpha_tH}{2}\|\mathbf{x}_t-\mathbf{w}^*\|^2\right)\\ &\le \underbrace{50L\sum_{t=0}^{t-1}\alpha_t^2\eta_t\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t-\mathbf{w}^*)}_{\text{(A)}}+\underbrace{4\sum_{t=1}^{T-1}\left(\frac{1}{\eta_{t+1}}-\frac{1}{\eta_t}-\frac{H}{8}\alpha_{t+1}\right)\|\mathbf{x}_{t+1}-\mathbf{w}^*\|^2}_{\text{(B)}}\\ &\quad+\underbrace{4\sum_{t=1}^{T-1}\left(\frac{1}{\eta_{t+1}}-\frac{1}{\eta_t}\right)\|\mathbf{x}_{t+1}-\mathbf{y}_t\|^2}_{\text{(C)}}+\underbrace{2\sum_{t=1}^{T}\alpha_t(\nabla f(\mathbf{x}_t)-\nabla f(\mathbf{x}_{t-\tau_t}))^\top(\mathbf{x}_t-\mathbf{w}^*)}_{\text{(D)}}\\ &\quad+\frac{4\cdot 24^4L^2\tilde{G}^2T(\sigma_\tau^2+\mu_\tau^2+4\mu_\tau+4)}{H^3}+\frac{2}{\eta_1}D^2+6L^2\eta_1D-2\sum_{t=1}^{T}\alpha_t(\mathbf{x}_t-\mathbf{w}^*)^\top\boldsymbol{\xi}_t\\ &\quad+6\sum_{t=1}^{T}\alpha_t^2\eta_t(\|\boldsymbol{\xi}_{t-\tau_t}\|^2+\|\boldsymbol{\xi}_{t-1-\tau_t\ \ 1}\|^2)\end{aligned}\tag{36}$$

Terms (A) − (C) can be bounded exactly as in Theorem 4.2 proof, with $t^* \leq 320\frac{L}{H}$, i.e.:

$$\begin{aligned}
&\text{(A)} \leq O\left(GD\left(\frac{L}{H}\right)^3\right) , \\
&\text{(B)} = 0 , \\
&\text{(C)} = O\left(\frac{G^2T}{H}\right) .
\end{aligned}$$

So, we are left with bounding term (D).

**Bounding (D)**

$$\begin{aligned}
\text{(D)} :=& \sum_{t=1}^{T} \alpha_t (\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-\tau_t}))^\top (\mathbf{x}_t - \mathbf{w}^*) \\
\leq& \sum_{t=1}^{T} \alpha_t \left\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-\tau_t})\right\| \|\mathbf{x}_t - \mathbf{w}^*\| \\
\leq& \sum_{t=1}^{T} \frac{\rho}{2} \|\mathbf{x}_t - \mathbf{w}^*\|^2 + \frac{\alpha_t^2}{2\rho} \left\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-\tau_t})\right\|^2 \\
\leq& \sum_{t=1}^{T} \frac{\alpha_t H}{8} \|\mathbf{x}_t - \mathbf{w}^*\|^2 + \frac{2\alpha_t}{H} \left\|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-\tau_t})\right\|^2 \\
\leq& \sum_{t=1}^{T} \frac{\alpha_t}{4} (f(\mathbf{x}_t) - f(\mathbf{w}^*)) + \frac{2\alpha_t L}{H} \|\mathbf{x}_t - \mathbf{x}_{t-\tau_t}\|^2 \\
\leq& \sum_{t=1}^{T} \frac{\alpha_t}{4} (f(\mathbf{x}_t) - f(\mathbf{w}^*)) + \frac{10^3 L^2 \tilde{G}^2 T(\sigma_\tau^2 + \mu_\tau^2 + 4\mu_\tau + 4)}{H^3} ,
\end{aligned} \tag{37}$$

where the second line is due to Cauchy-Schwartz Inequality, the third line is due to Young's Inequality, $ab \leq \inf_{\rho>0}\left(\rho a^2/2 + b^2/(2\rho)\right)$, when in the fourth line we took $\rho = \frac{\alpha_t H}{4}$. The fifth line utilizes smoothness and strong convexity which implies $\frac{H}{2}\|\mathbf{x} - \mathbf{w}^*\|^2 \leq f(\mathbf{x}) - f(\mathbf{w}^*)$. The last line uses (33) and Root Mean Square with Arithmetic Mean inequality, i.e. $\frac{\sum_{i=1}^{n} a_i}{n} \leq \sqrt{\frac{\sum_{i=1}^{n} a_i^2}{n}}$.

**Final Bound** As we mentioned at the beginning of this proof, under assumption 1,2 from Section 2, in expectation we have $\tilde{G} = \sqrt{2G^2 + 2\sigma^2}$. Combining the bounds in Equations (20) (21) (22), (37) and assumption 2 into Eq. (36) and taking expectation implies,

$$\mathbb{E}\left[\sum_{t=1}^{T} \alpha_t (f(\mathbf{x}_t) - f(\mathbf{w}^*))\right] \leq O\left(\frac{GDL^3}{H^3} + \frac{(G^2+\sigma^2)T}{H} + HD^2 + \frac{L^2D^2}{H} + \frac{L^2(G^2+\sigma^2)T(\sigma_\tau^2+\mu_\tau^2)}{H^3} + \frac{\sigma^2T^2}{H}\right) \tag{38}$$

Recalling $\bar{\mathbf{x}}_T \propto \sum_{t=1}^{T} \alpha_t \mathbf{x}_t$ and using Jensen's inequality established the theorem. □

### F.1. Proof of Inequality (32)

*Proof.* Let us define $F(\mathbf{x}) \triangleq f(\mathbf{x}) - \frac{H}{2}\|\mathbf{x}\|^2$. Note that $F(\mathbf{x})$ is convex, which follows from,

$$\begin{aligned} F(\mathbf{y}) - F(\mathbf{x}) &= f(\mathbf{y}) - f(\mathbf{x}) - \frac{H}{2}(\|\mathbf{y}\|^2 - \|\mathbf{x}\|^2) \\ &\geq \nabla f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x}) + \frac{H}{2}\|\mathbf{x}-\mathbf{y}\|^2 - \frac{H}{2}(\|\mathbf{y}\|^2 - \|\mathbf{x}\|^2) \\ &= \nabla f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x}) - H\mathbf{x}^\top(\mathbf{y}-\mathbf{x}) \\ &= \nabla F(\mathbf{x})^\top(\mathbf{y}-\mathbf{x})\,. \end{aligned}$$

From the monotone gradient condition for convexity of $F(\mathbf{x})$ we obtain,

$$(\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}))^\top(\mathbf{y}-\mathbf{x}) = (\nabla F(\mathbf{y}) - \nabla F(\mathbf{x}))^\top(\mathbf{y}-\mathbf{x}) + H(\mathbf{y}-\mathbf{x})^\top(\mathbf{y}-\mathbf{x}) \geq H\|\mathbf{y}-\mathbf{x}\|^2$$

,

where the second line uses strong convexity of $f(\cdot)$. Using Cauchy-Schwartz on the above inequality gives,

$$\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\|\|\mathbf{y}-\mathbf{x}\| \geq (\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}))^\top(\mathbf{y}-\mathbf{x}) \geq H\|\mathbf{y}-\mathbf{x}\|^2\,.$$

Dividing both sides of the inequality above by $\|\mathbf{y}-\mathbf{x}\|$ concludes the proof. □

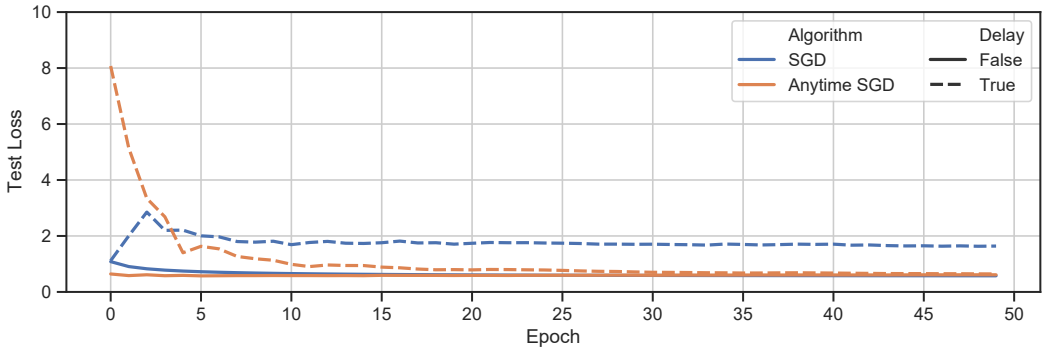


*Figure 3.* Expected excess loss as function of epochs when $_t$ = 500, with learning rate optimized for each of the algorithms for zero delay regime.

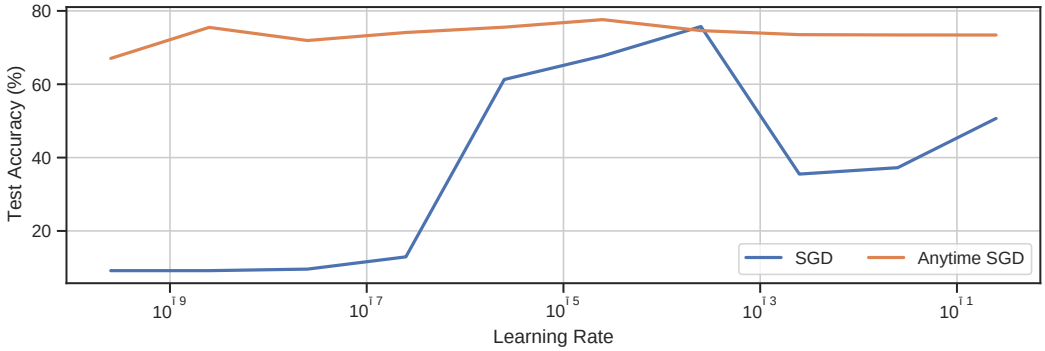


*Figure 4.* Accuracy as a function of learning rate when $_t$ Lognormal(7; 0:4$^2$)

## G. Further Experiments

As was mentioned in Section 5, Figure 1 demonstrates the final test accuracy for different delay regimes. Figure 3 expands on the regime of $\tau_t = 500$, and compares between the expected excess loss of our algorithm and that of SGD as was suggested by (Stich & Karimireddy, 2020). While the addition of the delay affects the convergence, as evident from the theoretical bounds as well, in anytime SGD the expected loss approaches the optimal one, while that of SGD does not.

While Figure 2 demonstrates the performance of the algorithms on a wide range of learning rates when $\tau_t = 500$, in Figure 4 the delay is distributed Lognormal$(7, 0.4^2)$, a heavy-tail distribution. This shows that anytime SGD performs better when a high maximal delay, but reasonable mean, is present.