

A Technical Results

A.1 Stochastic Gradient Descent Proofs

Proof of Theorem 1. Due to H -smoothness of convex function L we have that,

$$\begin{aligned} L(\mathbf{w}_{i+1}) &\leq L(\mathbf{w}_i) + \langle \nabla L(\mathbf{w}_i), \mathbf{w}_{i+1} - \mathbf{w}_i \rangle + \frac{H}{2} \|\mathbf{w}_{i+1} - \mathbf{w}_i\|^2 \\ &= L(\mathbf{w}_i) + \langle \nabla L(\mathbf{w}_i) - \nabla \ell_i(\mathbf{w}_i), \mathbf{w}_{i+1} - \mathbf{w}_i \rangle + \frac{H}{2} \|\mathbf{w}_{i+1} - \mathbf{w}_i\|^2 + \langle \nabla \ell_i(\mathbf{w}_i), \mathbf{w}_{i+1} - \mathbf{w}_i \rangle \end{aligned}$$

by Cauchy-Schwartz inequality we get,

$$\leq L(\mathbf{w}_i) + \|\nabla L(\mathbf{w}_i) - \nabla \ell_i(\mathbf{w}_i)\| \|\mathbf{w}_{i+1} - \mathbf{w}_i\| + \frac{H}{2} \|\mathbf{w}_{i+1} - \mathbf{w}_i\|^2 + \langle \nabla \ell_i(\mathbf{w}_i), \mathbf{w}_{i+1} - \mathbf{w}_i \rangle$$

since for any $\alpha > 0$, $ab \leq \frac{a^2}{2\alpha} + \frac{\alpha b^2}{2}$,

$$\begin{aligned} &\leq L(\mathbf{w}_i) + \frac{\|\nabla L(\mathbf{w}_i) - \nabla \ell_i(\mathbf{w}_i)\|^2}{2(1/\eta - H)} + \frac{(1/\eta - H)}{2} \|\mathbf{w}_{i+1} - \mathbf{w}_i\|^2 + \frac{H}{2} \|\mathbf{w}_{i+1} - \mathbf{w}_i\|^2 + \langle \nabla \ell_i(\mathbf{w}_i), \mathbf{w}_{i+1} - \mathbf{w}_i \rangle \\ &= L(\mathbf{w}_i) + \frac{\|\nabla L(\mathbf{w}_i) - \nabla \ell_i(\mathbf{w}_i)\|^2}{2(1/\eta - H)} + \frac{\|\mathbf{w}_{i+1} - \mathbf{w}_i\|^2}{2\eta} + \langle \nabla \ell_i(\mathbf{w}_i), \mathbf{w}_{i+1} - \mathbf{w}_i \rangle \end{aligned}$$

We now note that the update step can be written equivalently as

$$\mathbf{w}_{i+1} = \underset{\mathbf{w}: \|\mathbf{w}\| \leq D}{\operatorname{argmin}} \left\{ \eta \langle \nabla \ell_i(\mathbf{w}_i), \mathbf{w} - \mathbf{w}_i \rangle + \frac{\|\mathbf{w} - \mathbf{w}_i\|^2}{2} \right\}.$$

It can be shown that (see for instance Lemma 1 of [5])

$$\eta \langle \nabla \ell_i(\mathbf{w}_i), \mathbf{w}_{i+1} - \mathbf{w}_i \rangle \leq \eta \langle \nabla \ell_i(\mathbf{w}_i), \mathbf{w}^* - \mathbf{w}_i \rangle + \frac{\|\mathbf{w}^* - \mathbf{w}_i\|^2 - \|\mathbf{w}^* - \mathbf{w}_{i+1}\|^2 - \|\mathbf{w}_i - \mathbf{w}_{i+1}\|^2}{2}$$

Plugging this we get that,

$$\begin{aligned} L(\mathbf{w}_{i+1}) &\leq L(\mathbf{w}_i) + \frac{\|\nabla L(\mathbf{w}_i) - \nabla \ell_i(\mathbf{w}_i)\|^2}{2(1/\eta - H)} + \frac{\|\mathbf{w}_i - \mathbf{w}_{i+1}\|^2}{2\eta} + \langle \nabla \ell_i(\mathbf{w}_i), \mathbf{w}^* - \mathbf{w}_i \rangle \\ &\quad + \frac{1}{2\eta} \left(\|\mathbf{w}_i - \mathbf{w}^*\|^2 - \|\mathbf{w}'_{i+1} - \mathbf{w}^*\|^2 - \|\mathbf{w}_i - \mathbf{w}'_{i+1}\|^2 \right) \\ &= L(\mathbf{w}_i) + \frac{\|\nabla L(\mathbf{w}_i) - \nabla \ell_i(\mathbf{w}_i)\|^2}{2(1/\eta - H)} + \frac{\|\mathbf{w}_i - \mathbf{w}_{i+1}\|^2}{2\eta} + \langle \nabla \ell_i(\mathbf{w}_i) - \nabla L(\mathbf{w}_i), \mathbf{w}^* - \mathbf{w}_i \rangle + \langle \nabla L(\mathbf{w}_i), \mathbf{w}^* - \mathbf{w}_i \rangle \\ &\quad + \frac{1}{2\eta} \left(\|\mathbf{w}_i - \mathbf{w}^*\|^2 - \|\mathbf{w}'_{i+1} - \mathbf{w}^*\|^2 - \|\mathbf{w}_i - \mathbf{w}'_{i+1}\|^2 \right) \\ &\geq L(\mathbf{w}_i) + \frac{\|\nabla L(\mathbf{w}_i) - \nabla \ell_i(\mathbf{w}_i)\|^2}{2(1/\eta - H)} + \frac{\|\mathbf{w}_i - \mathbf{w}_{i+1}\|^2}{2\eta} + \langle \nabla \ell_i(\mathbf{w}_i) - \nabla L(\mathbf{w}_i), \mathbf{w}^* - \mathbf{w}_i \rangle - \langle \nabla L(\mathbf{w}_i), \mathbf{w}_i - \mathbf{w}^* \rangle \\ &\quad + \frac{1}{2\eta} \left(\|\mathbf{w}_i - \mathbf{w}^*\|^2 - \|\mathbf{w}_{i+1} - \mathbf{w}^*\|^2 - \|\mathbf{w}_i - \mathbf{w}_{i+1}\|^2 \right) \\ &= L(\mathbf{w}_i) + \frac{\|\nabla L(\mathbf{w}_i) - \nabla \ell_i(\mathbf{w}_i)\|^2}{2(1/\eta - H)} + \langle \nabla \ell_i(\mathbf{w}_i) - \nabla L(\mathbf{w}_i), \mathbf{w}^* - \mathbf{w}_i \rangle - \langle \nabla L(\mathbf{w}_i), \mathbf{w}_i - \mathbf{w}^* \rangle \\ &\quad + \frac{1}{2\eta} \left(\|\mathbf{w}_i - \mathbf{w}^*\|^2 - \|\mathbf{w}_{i+1} - \mathbf{w}^*\|^2 \right) \end{aligned}$$

since $\eta \leq \frac{1}{2H}$,

$$\begin{aligned} &\leq L(\mathbf{w}_i) + \eta \|\nabla L(\mathbf{w}_i) - \nabla \ell_i(\mathbf{w}_i)\|^2 + \langle \nabla \ell_i(\mathbf{w}_i) - \nabla L(\mathbf{w}_i), \mathbf{w}^* - \mathbf{w}_i \rangle - \langle \nabla L(\mathbf{w}_i), \mathbf{w}_i - \mathbf{w}^* \rangle \\ &\quad + \frac{1}{2\eta} \left(\|\mathbf{w}_i - \mathbf{w}^*\|^2 - \|\mathbf{w}_{i+1} - \mathbf{w}^*\|^2 \right) \end{aligned}$$

by convexity, $L(\mathbf{w}_i) - \langle \nabla L(\mathbf{w}_i), \mathbf{w}_i - \mathbf{w}^\star \rangle \leq L(\mathbf{w}^\star)$ and so

$$\begin{aligned} &\leq L(\mathbf{w}^\star) + \eta \|\nabla L(\mathbf{w}_i) - \nabla \ell_i(\mathbf{w}_i)\|^2 + \langle \nabla \ell_i(\mathbf{w}_i) - \nabla L(\mathbf{w}_i), \mathbf{w}^\star - \mathbf{w}_i \rangle \\ &\quad + \frac{1}{2\eta} \left(\|\mathbf{w}_i - \mathbf{w}^\star\|^2 - \|\mathbf{w}_{i+1} - \mathbf{w}^\star\|^2 \right) \end{aligned}$$

Hence we conclude that :

$$\begin{aligned} \frac{1}{n-1} \sum_{i=1}^{n-1} L(\mathbf{w}_{i+1}) - L(\mathbf{w}^\star) &\leq \frac{\eta}{(n-1)} \sum_{i=1}^{n-1} \|\nabla L(\mathbf{w}_i) - \nabla \ell_i(\mathbf{w}_i)\|^2 + \frac{1}{n-1} \sum_{i=1}^{n-1} \langle \nabla \ell_i(\mathbf{w}_i) - \nabla L(\mathbf{w}_i), \mathbf{w}^\star - \mathbf{w}_i \rangle \\ &\quad + \frac{1}{n-1} \sum_{i=1}^{n-1} \frac{\|\mathbf{w}_i - \mathbf{w}^\star\|^2 - \|\mathbf{w}_{i+1} - \mathbf{w}^\star\|^2}{2\eta} \\ &= \frac{\eta}{(n-1)} \sum_{i=1}^{n-1} \|\nabla L(\mathbf{w}_i) - \nabla \ell_i(\mathbf{w}_i)\|^2 + \frac{1}{n-1} \sum_{i=1}^{n-1} \langle \nabla \ell_i(\mathbf{w}_i) - \nabla L(\mathbf{w}_i), \mathbf{w}^\star - \mathbf{w}_i \rangle \\ &\quad + \frac{1}{n-1} \sum_{i=1}^{n-1} \frac{\|\mathbf{w}_1 - \mathbf{w}^\star\|^2 - \|\mathbf{w}_n - \mathbf{w}^\star\|^2}{2\eta} \\ &\leq \frac{\eta}{(n-1)} \sum_{i=1}^{n-1} \|\nabla L(\mathbf{w}_i) - \nabla \ell_i(\mathbf{w}_i)\|^2 + \frac{1}{n-1} \sum_{i=1}^{n-1} \langle \nabla \ell_i(\mathbf{w}_i) - \nabla L(\mathbf{w}_i), \mathbf{w}^\star - \mathbf{w}_i \rangle \\ &\quad + \frac{\|\mathbf{w}^\star\|^2}{2\eta(n-1)} \\ &\leq \frac{\eta}{(n-1)} \sum_{i=1}^{n-1} \|\nabla L(\mathbf{w}_i) - \nabla \ell_i(\mathbf{w}_i)\|^2 + \frac{1}{n-1} \sum_{i=1}^{n-1} \langle \nabla \ell_i(\mathbf{w}_i) - \nabla L(\mathbf{w}_i), \mathbf{w}^\star - \mathbf{w}_i \rangle \\ &\quad + \frac{D^2}{2\eta(n-1)} \end{aligned}$$

Taking expectation with respect to sample on both sides and noticing that $\mathbb{E}[\langle \nabla \ell_i(\mathbf{w}_i) - \nabla L(\mathbf{w}_i), \mathbf{w}^\star - \mathbf{w}_i \rangle] = 0$, we get that,

$$\mathbb{E} \left[\frac{1}{n-1} \sum_{i=1}^{n-1} L(\mathbf{w}_{i+1}) - L(\mathbf{w}^\star) \right] \leq \frac{\eta}{(n-1)} \sum_{i=1}^{n-1} \mathbb{E} [\|\nabla L(\mathbf{w}_i) - \nabla \ell_i(\mathbf{w}_i)\|^2] + \frac{D^2}{2\eta(n-1)}$$

Now note that

$$\nabla L(\mathbf{w}_i) - \nabla \ell_i(\mathbf{w}_i) = \frac{1}{b} \sum_{t=(i-1)b+1}^{bi} (\nabla L(\mathbf{w}_i) - \nabla \ell(\mathbf{w}_i, z_t))$$

and that $(\nabla L(\mathbf{w}_i) - \nabla \ell(\mathbf{w}_i, z_t))$ is a mean zero vector drawn i.i.d. Also note that \mathbf{w}_i only depends on the first $(i-1)b$ examples and so when we consider expectation w.r.t. $z_{(i-1)b+1}, \dots, z_{ib}$ alone, \mathbf{w}_i is fixed. Hence we have that

$$\begin{aligned} \mathbb{E} [\|\nabla L(\mathbf{w}_i) - \nabla \ell_i(\mathbf{w}_i)\|^2] &= \frac{1}{b^2} \mathbb{E} \left[\left\| \sum_{t=(i-1)b+1}^{bi} (\nabla L(\mathbf{w}_i) - \nabla \ell(\mathbf{w}_i, z_t)) \right\|^2 \right] \\ &= \frac{1}{b^2} \sum_{t=(i-1)b+1}^{bi} \mathbb{E} [\|(\nabla L(\mathbf{w}_i) - \nabla \ell(\mathbf{w}_i, z_t))\|^2] \end{aligned}$$

Plugging this back we get that

$$\begin{aligned}\mathbb{E} \left[\frac{1}{n-1} \sum_{i=1}^{n-1} L(\mathbf{w}_{i+1}) - L(\mathbf{w}^*) \right] &\leq \frac{\eta}{b^2(n-1)} \sum_{i=1}^{n-1} \sum_{t=(i-1)b+1}^{ib} \mathbb{E} \left[\|\nabla L(\mathbf{w}_i) - \nabla \ell(\mathbf{w}_i, z_t)\|^2 \right] + \frac{D^2}{2\eta(n-1)} \\ &\leq \frac{2\eta}{b^2(n-1)} \sum_{i=1}^{n-1} \sum_{t=(i-1)b+1}^{ib} \mathbb{E} \left[\|\nabla L(\mathbf{w}_i)\|^2 + \|\nabla \ell(\mathbf{w}_i, z_t)\|^2 \right] + \frac{D^2}{2\eta(n-1)}\end{aligned}$$

for any non-negative H -smooth convex function f , we have the self-bounding property that $\|\nabla f(\mathbf{w})\| \leq \sqrt{4Hf(\mathbf{w})}$. Using this,

$$\begin{aligned}&\leq \frac{8H\eta}{b^2(n-1)} \sum_{i=1}^{n-1} \sum_{t=(i-1)b+1}^{ib} \mathbb{E} [L(\mathbf{w}_i) + \ell(\mathbf{w}_i, z_t)] + \frac{D^2}{2\eta(n-1)} \\ &= \frac{16\eta H}{b} \mathbb{E} \left[\frac{1}{n-1} \sum_{i=1}^{n-1} L(\mathbf{w}_i) \right] + \frac{D^2}{2\eta(n-1)}\end{aligned}$$

Adding $\frac{1}{n-1}L(\mathbf{w}_1)$ on both sides and removing $L(\mathbf{w}_n)$ on the left we conclude that

$$\mathbb{E} \left[\frac{1}{n-1} \sum_{i=1}^{n-1} L(\mathbf{w}_i) \right] - L(\mathbf{w}^*) \leq \frac{16\eta H}{b} \mathbb{E} \left[\frac{1}{n-1} \sum_{i=1}^{n-1} L(\mathbf{w}_i) \right] + \frac{D^2}{2\eta(n-1)} + \frac{L(\mathbf{w}_1)}{n-1}$$

Hence we conclude that

$$\begin{aligned}\mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n L(\mathbf{w}_i) \right] - L(\mathbf{w}^*) &\leq \frac{1}{\left(1 - \frac{16\eta H}{b}\right)} \left(\frac{16\eta H}{b} L(\mathbf{w}^*) + \frac{L(\mathbf{w}_1)}{n} + \frac{D^2}{2\eta n} \right) \\ &= \left(\frac{1}{1 - \frac{16\eta H}{b}} - 1 \right) L(\mathbf{w}^*) + \frac{1}{1 - \frac{16\eta H}{b}} \left(\frac{L(\mathbf{w}_1)}{n} + \frac{D^2}{2\eta n} \right) \\ &= \left(\frac{1}{1 - \frac{16\eta H}{b}} - 1 \right) L(\mathbf{w}^*) + \left(\frac{1}{1 - \frac{16\eta H}{b}} \right) \frac{L(\mathbf{w}_1)}{n} + \left(\frac{1}{1 - \frac{16\eta H}{b}} \right) \frac{b}{16\eta H} \frac{8HD^2}{bn}\end{aligned}$$

Writing $\alpha = \frac{1}{1 - \frac{16\eta H}{b}} - 1$, so that $\eta = \frac{b}{16H} \left(1 - \frac{1}{\alpha+1}\right)$ we get,

$$\begin{aligned}\mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n L(\mathbf{w}_i) \right] - L(\mathbf{w}^*) &\leq \alpha L(\mathbf{w}^*) + \frac{(\alpha+1)L(\mathbf{w}_1)}{n} + \frac{16H(\alpha+1)^2}{\alpha} \frac{D^2}{2bn} \\ &\leq \alpha L(\mathbf{w}^*) + \frac{(\alpha+1)L(\mathbf{w}_1)}{n} + \left(\alpha + \frac{1}{\alpha} \right) \frac{8HD^2}{bn}\end{aligned}$$

Now we shall always pick $\eta \leq \frac{b}{32H}$ so that $\alpha \leq 1$ and so

$$\mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n L(\mathbf{w}_i) \right] - L(\mathbf{w}^*) \leq \alpha L(\mathbf{w}^*) + \frac{HD^2}{8\alpha bn} + \frac{2L(\mathbf{w}_1)}{n} + \frac{8HD^2}{bn}$$

Picking

$$\eta = \min \left\{ \frac{1}{2H}, \frac{b}{32H}, \frac{\sqrt{\frac{bD^2}{8L(\mathbf{w}^*)Hn}}}{16 \left(1 + \sqrt{\frac{HD^2}{8L(\mathbf{w}^*)bn}}\right)} \right\},$$

or equivalently $\alpha = \min \left\{ 1, \sqrt{\frac{HD^2}{8L(\mathbf{w}^*)bn}} \right\}$ we get,

$$\mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n L(\mathbf{w}_i) \right] - L(\mathbf{w}^*) \leq \sqrt{\frac{HD^2 L(\mathbf{w}^*)}{2bn}} + \frac{2L(\mathbf{w}_1)}{n} + \frac{9HD^2}{bn}$$

Finally note that by smoothness, $L(\mathbf{w}_1) = L(0) \leq HD^2$. Hence we conclude that

$$\mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n L(\mathbf{w}_i) \right] - L(\mathbf{w}^*) \leq \sqrt{\frac{HD^2 L(\mathbf{w}^*)}{2bn}} + \frac{2HD^2}{n} + \frac{9HD^2}{bn}$$

Using Jensen's inequality concludes the proof. \square

A.2 Accelerated Gradient Proofs

Lemma A.1. *For the accelerated update rule, if the step sizes $\beta_i \in [1, \infty)$ and $\gamma_i \in (0, \infty)$ are chosen such that $\beta_1 = 1$ and for all $i \in [n]$*

$$0 < \gamma_{i+1}(\beta_{i+1} - 1) \leq \beta_i \gamma_i \quad \text{and} \quad 2H\gamma_i \leq \beta_i$$

then we have that

$$\mathbb{E}[L(\mathbf{w}_n^{\text{ag}})] - L(\mathbf{w}^*) \leq \frac{\gamma_1(\beta_1 - 1)}{\gamma_n(\beta_n - 1)} L(\mathbf{w}_1^{\text{ag}}) + \frac{32H}{b\gamma_n(\beta_n - 1)} \sum_{i=1}^{n-1} \gamma_i^2 \mathbb{E}[L(\mathbf{w}_i^{\text{ag}})] + \frac{D^2}{2\gamma_n(\beta_n - 1)} + \frac{16H^2 D^2}{b\gamma_n(\beta_n - 1)} \sum_{i=1}^{n-1} \frac{\gamma_i^2}{\beta_i^2}$$

Proof. First note that for any i ,

$$\begin{aligned} \mathbf{w}_{i+1}^{\text{ag}} - \mathbf{w}_i^{\text{md}} &= \beta_i^{-1} \mathbf{w}_{i+1} + (1 - \beta_i^{-1}) \mathbf{w}_i^{\text{ag}} - \mathbf{w}_i^{\text{md}} \\ &= \beta_i^{-1} \mathbf{w}_{i+1} + (1 - \beta_i^{-1}) \mathbf{w}_i^{\text{ag}} - \beta_i^{-1} \mathbf{w}_i - (1 - \beta_i^{-1}) \mathbf{w}_i^{\text{ag}} \\ &= \beta_i^{-1} (\mathbf{w}_{i+1} - \mathbf{w}_i) \end{aligned} \tag{9}$$

Now by smoothness we have that

$$\begin{aligned} L(\mathbf{w}_{i+1}^{\text{ag}}) &\leq L(\mathbf{w}_i^{\text{md}}) + \langle \nabla L(\mathbf{w}_i^{\text{md}}), \mathbf{w}_{i+1}^{\text{ag}} - \mathbf{w}_i^{\text{md}} \rangle + \frac{H}{2} \|\mathbf{w}_{i+1}^{\text{ag}} - \mathbf{w}_i^{\text{md}}\|^2 \\ &= L(\mathbf{w}_i^{\text{md}}) + \langle \nabla L(\mathbf{w}_i^{\text{md}}), \mathbf{w}_{i+1}^{\text{ag}} - \mathbf{w}_i^{\text{md}} \rangle + \frac{H}{2\beta_i^2} \|\mathbf{w}_{i+1} - \mathbf{w}_i\|^2 \\ &= L(\mathbf{w}_i^{\text{md}}) + \langle \nabla L(\mathbf{w}_i^{\text{md}}), \mathbf{w}_{i+1}^{\text{ag}} - \mathbf{w}_i^{\text{md}} \rangle + \frac{1}{2\beta_i \gamma_i} \|\mathbf{w}_{i+1} - \mathbf{w}_i\|^2 - \frac{\beta_i/\gamma_i - H}{2\beta_i^2} \|\mathbf{w}_{i+1} - \mathbf{w}_i\|_*^2 \end{aligned}$$

since $\mathbf{w}_{i+1}^{\text{ag}} = \beta_i^{-1} \mathbf{w}_{i+1} + (1 - \beta_i^{-1}) \mathbf{w}_i^{\text{ag}}$,

$$\begin{aligned} &= L(\mathbf{w}_i^{\text{md}}) + \langle \nabla L(\mathbf{w}_i^{\text{md}}), \beta_i^{-1} \mathbf{w}_{i+1} + (1 - \beta_i^{-1}) \mathbf{w}_i^{\text{ag}} - \mathbf{w}_i^{\text{md}} \rangle + \frac{\|\mathbf{w}_i - \mathbf{w}_{i+1}\|^2}{2\beta_i \gamma_i} - \frac{\beta_i/\gamma_i - H}{2\beta_i^2} \|\mathbf{w}_{i+1} - \mathbf{w}_i\|^2 \\ &= L(\mathbf{w}_i^{\text{md}}) + (1 - \beta_i^{-1}) \langle \nabla L(\mathbf{w}_i^{\text{md}}), \mathbf{w}_i^{\text{ag}} - \mathbf{w}_i^{\text{md}} \rangle + \frac{\langle \nabla L(\mathbf{w}_i^{\text{md}}), \mathbf{w}_{i+1} - \mathbf{w}_i^{\text{md}} \rangle}{\beta_i} + \frac{\|\mathbf{w}_i - \mathbf{w}_{i+1}\|^2}{2\beta_i \gamma_i} \\ &\quad - \frac{\beta_i/\gamma_i - H}{2\beta_i^2} \|\mathbf{w}_{i+1} - \mathbf{w}_i\|^2 \\ &= (1 - \beta_i^{-1}) (L(\mathbf{w}_i^{\text{md}}) + \langle \nabla L(\mathbf{w}_i^{\text{md}}), \mathbf{w}_i^{\text{ag}} - \mathbf{w}_i^{\text{md}} \rangle) + \frac{L(\mathbf{w}_i^{\text{md}}) + \langle \nabla L(\mathbf{w}_i^{\text{md}}), \mathbf{w}_{i+1} - \mathbf{w}_i^{\text{md}} \rangle}{\beta_i} + \frac{\|\mathbf{w}_i - \mathbf{w}_{i+1}\|^2}{2\beta_i \gamma_i} \\ &\quad - \frac{\beta_i/\gamma_i - H}{2\beta_i^2} \|\mathbf{w}_{i+1} - \mathbf{w}_i\|^2 \\ &= (1 - \beta_i^{-1}) L(\mathbf{w}_i^{\text{ag}}) + \frac{L(\mathbf{w}_i^{\text{md}}) + \langle \nabla L(\mathbf{w}_i^{\text{md}}), \mathbf{w}_{i+1} - \mathbf{w}_i^{\text{md}} \rangle}{\beta_i} + \frac{\|\mathbf{w}_i - \mathbf{w}_{i+1}\|^2}{2\beta_i \gamma_i} - \frac{\beta_i/\gamma_i - H}{2\beta_i^2} \|\mathbf{w}_{i+1} - \mathbf{w}_i\|^2 \\ &= (1 - \beta_i^{-1}) L(\mathbf{w}_i^{\text{ag}}) - \frac{\beta_i/\gamma_i - H}{2\beta_i^2} \|\mathbf{w}_{i+1} - \mathbf{w}_i\|^2 + \frac{\|\mathbf{w}_i - \mathbf{w}_{i+1}\|^2}{2\beta_i \gamma_i} \\ &\quad + \frac{L(\mathbf{w}_i^{\text{md}}) + \langle \nabla \ell_i(\mathbf{w}_i^{\text{md}}), \mathbf{w}_{i+1} - \mathbf{w}_i^{\text{md}} \rangle + \langle \nabla L(\mathbf{w}_i^{\text{md}}) - \nabla \ell_i(\mathbf{w}_i^{\text{md}}), \mathbf{w}_{i+1} - \mathbf{w}_i^{\text{md}} \rangle}{\beta_i} \\ &= (1 - \beta_i^{-1}) L(\mathbf{w}_i^{\text{ag}}) - \frac{\beta_i/\gamma_i - H}{2\beta_i^2} \|\mathbf{w}_{i+1} - \mathbf{w}_i\|^2 + \frac{\|\mathbf{w}_i - \mathbf{w}_{i+1}\|^2}{2\beta_i \gamma_i} + \frac{\langle \nabla L(\mathbf{w}_i^{\text{md}}) - \nabla \ell_i(\mathbf{w}_i^{\text{md}}), \mathbf{w}_{i+1} - \mathbf{w}_i \rangle}{\beta_i} \\ &\quad + \frac{L(\mathbf{w}_i^{\text{md}}) + \langle \nabla \ell_i(\mathbf{w}_i^{\text{md}}), \mathbf{w}_{i+1} - \mathbf{w}_i^{\text{md}} \rangle + \langle \nabla L(\mathbf{w}_i^{\text{md}}) - \nabla \ell_i(\mathbf{w}_i^{\text{md}}), \mathbf{w}_i - \mathbf{w}_i^{\text{md}} \rangle}{\beta_i} \end{aligned}$$

by Cauchy-Schwartz inequality,

$$\begin{aligned} &\leq (1 - \beta_i^{-1})L(\mathbf{w}_i^{\text{ag}}) - \frac{\beta_i/\gamma_i - H}{2\beta_i^2} \|\mathbf{w}_{i+1} - \mathbf{w}_i\|^2 + \frac{\|\mathbf{w}_i - \mathbf{w}_{i+1}\|^2}{2\beta_i\gamma_i} + \frac{\|\nabla L(\mathbf{w}_i^{\text{md}}) - \nabla \ell_i(\mathbf{w}_i^{\text{md}})\| \|\mathbf{w}_{i+1} - \mathbf{w}_i\|}{\beta_i} \\ &\quad + \frac{L(\mathbf{w}_i^{\text{md}}) + \langle \nabla \ell_i(\mathbf{w}_i^{\text{md}}), \mathbf{w}_{i+1} - \mathbf{w}_i^{\text{md}} \rangle + \langle \nabla L(\mathbf{w}_i^{\text{md}}) - \nabla \ell_i(\mathbf{w}_i^{\text{md}}), \mathbf{w}_i - \mathbf{w}_i^{\text{md}} \rangle}{\beta_i} \end{aligned}$$

since for any a, b and $\alpha > 0$, $ab \leq \frac{a^2}{2\alpha} + \frac{\alpha b^2}{2}$

$$\begin{aligned} &\leq (1 - \beta_i^{-1})L(\mathbf{w}_i^{\text{ag}}) + \frac{\|\nabla L(\mathbf{w}_i^{\text{md}}) - \nabla \ell_i(\mathbf{w}_i^{\text{md}})\|^2}{2(\beta_i/\gamma_i - H)} + \frac{\|\mathbf{w}_i - \mathbf{w}_{i+1}\|^2}{2\beta_i\gamma_i} \\ &\quad + \frac{L(\mathbf{w}_i^{\text{md}}) + \langle \nabla \ell_i(\mathbf{w}_i^{\text{md}}), \mathbf{w}_{i+1} - \mathbf{w}_i^{\text{md}} \rangle + \langle \nabla L(\mathbf{w}_i^{\text{md}}) - \nabla \ell_i(\mathbf{w}_i^{\text{md}}), \mathbf{w}_i - \mathbf{w}_i^{\text{md}} \rangle}{\beta_i} \end{aligned}$$

We now note that the update step 2 of accelerated gradient can be written equivalently as

$$\mathbf{w}_{i+1} = \underset{\mathbf{w}: \|\mathbf{w}\| \leq D}{\operatorname{argmin}} \left\{ \gamma_i \langle \nabla \ell_i(\mathbf{w}_i^{\text{md}}), \mathbf{w} - \mathbf{w}_i^{\text{md}} \rangle + \frac{\|\mathbf{w} - \mathbf{w}_i^{\text{md}}\|^2}{2} \right\}.$$

It can be shown that (see for instance Lemma 1 of [5])

$$\gamma_i \langle \nabla \ell_i(\mathbf{w}_i), \mathbf{w}_{i+1} - \mathbf{w}_i^{\text{md}} \rangle \leq \gamma_i \langle \nabla \ell_i(\mathbf{w}_i^{\text{md}}), \mathbf{w}^* - \mathbf{w}_i^{\text{md}} \rangle + \frac{\|\mathbf{w}^* - \mathbf{w}_i\|^2 - \|\mathbf{w}^* - \mathbf{w}_{i+1}\|^2 - \|\mathbf{w}_i - \mathbf{w}_{i+1}\|^2}{2}$$

Plugging this we get that,

$$\begin{aligned} L(\mathbf{w}_{i+1}^{\text{ag}}) &\leq (1 - \beta_i^{-1})L(\mathbf{w}_i^{\text{ag}}) + \frac{\|\nabla L(\mathbf{w}_i^{\text{md}}) - \nabla \ell_i(\mathbf{w}_i^{\text{md}})\|^2}{2(\beta_i/\gamma_i - H)} + \frac{\|\mathbf{w}_i - \mathbf{w}_{i+1}\|^2}{2\beta_i\gamma_i} + \frac{\langle \nabla \ell_i(\mathbf{w}_i^{\text{md}}), \mathbf{w}^* - \mathbf{w}_i^{\text{md}} \rangle}{\beta_i} \\ &\quad + \frac{L(\mathbf{w}_i^{\text{md}}) + \langle \nabla L(\mathbf{w}_i^{\text{md}}) - \nabla \ell_i(\mathbf{w}_i^{\text{md}}), \mathbf{w}_i - \mathbf{w}_i^{\text{md}} \rangle}{\beta_i} + \frac{\|\mathbf{w}^* - \mathbf{w}_i\|^2 - \|\mathbf{w}^* - \mathbf{w}_{i+1}\|^2 - \|\mathbf{w}_i - \mathbf{w}_{i+1}\|^2}{2\gamma_i\beta_i} \\ &= (1 - \beta_i^{-1})L(\mathbf{w}_i^{\text{ag}}) + \frac{\|\nabla L(\mathbf{w}_i^{\text{md}}) - \nabla \ell_i(\mathbf{w}_i^{\text{md}})\|^2}{2(\beta_i/\gamma_i - H)} + \frac{\langle \nabla \ell_i(\mathbf{w}_i^{\text{md}}), \mathbf{w}^* - \mathbf{w}_i^{\text{md}} \rangle}{\beta_i} \\ &\quad + \frac{L(\mathbf{w}_i^{\text{md}}) + \langle \nabla L(\mathbf{w}_i^{\text{md}}) - \nabla \ell_i(\mathbf{w}_i^{\text{md}}), \mathbf{w}_i - \mathbf{w}_i^{\text{md}} \rangle}{\beta_i} + \frac{\|\mathbf{w}^* - \mathbf{w}_i\|^2 - \|\mathbf{w}^* - \mathbf{w}_{i+1}\|^2}{2\gamma_i\beta_i} \\ &= (1 - \beta_i^{-1})L(\mathbf{w}_i^{\text{ag}}) + \frac{\|\nabla L(\mathbf{w}_i^{\text{md}}) - \nabla \ell_i(\mathbf{w}_i^{\text{md}})\|^2}{2(\beta_i/\gamma_i - H)} + \frac{\langle \nabla \ell_i(\mathbf{w}_i^{\text{md}}) - \nabla L(\mathbf{w}_i^{\text{md}}), \mathbf{w}^* - \mathbf{w}_i^{\text{md}} \rangle}{\beta_i} \\ &\quad + \frac{\langle \nabla L(\mathbf{w}_i^{\text{md}}) - \nabla \ell_i(\mathbf{w}_i^{\text{md}}), \mathbf{w}_i - \mathbf{w}_i^{\text{md}} \rangle}{\beta_i} + \frac{\|\mathbf{w}^* - \mathbf{w}_i\|^2 - \|\mathbf{w}^* - \mathbf{w}_{i+1}\|^2}{2\gamma_i\beta_i} \\ &\quad + \frac{L(\mathbf{w}_i^{\text{md}}) + \langle \nabla L(\mathbf{w}_i^{\text{md}}), \mathbf{w}^* - \mathbf{w}_i^{\text{md}} \rangle}{\beta_i} \end{aligned}$$

by convexity, $L(\mathbf{w}^*) \geq L(\mathbf{w}_i^{\text{md}}) + \langle \nabla L(\mathbf{w}_i^{\text{md}}), \mathbf{w}^* - \mathbf{w}_i^{\text{md}} \rangle$, hence

$$\begin{aligned}
&\leq (1 - \beta_i^{-1})L(\mathbf{w}_i^{\text{ag}}) + \frac{\|\nabla L(\mathbf{w}_i^{\text{md}}) - \nabla \ell_i(\mathbf{w}_i^{\text{md}})\|^2}{2(\beta_i/\gamma_i - H)} + \frac{\langle \nabla \ell_i(\mathbf{w}_i^{\text{md}}) - \nabla L(\mathbf{w}_i^{\text{md}}), \mathbf{w}^* - \mathbf{w}_i^{\text{md}} \rangle}{\beta_i} \\
&\quad + \frac{\langle \nabla L(\mathbf{w}_i^{\text{md}}) - \nabla \ell_i(\mathbf{w}_i^{\text{md}}), \mathbf{w}_i - \mathbf{w}_i^{\text{md}} \rangle}{\beta_i} + \frac{\|\mathbf{w}^* - \mathbf{w}_i\|^2 - \|\mathbf{w}^* - \mathbf{w}_{i+1}\|^2}{2\gamma_i\beta_i} + \frac{L(\mathbf{w}^*)}{\beta_i} \\
&= (1 - \beta_i^{-1})L(\mathbf{w}_i^{\text{ag}}) + \frac{\|\nabla L(\mathbf{w}_i^{\text{md}}) - \nabla \ell_i(\mathbf{w}_i^{\text{md}})\|^2}{2(\beta_i/\gamma_i - H)} + \frac{\langle \nabla L(\mathbf{w}_i^{\text{md}}) - \nabla \ell_i(\mathbf{w}_i^{\text{md}}), \mathbf{w}_i - \mathbf{w}^* \rangle}{\beta_i} \\
&\quad + \frac{\|\mathbf{w}^* - \mathbf{w}_i\|^2 - \|\mathbf{w}^* - \mathbf{w}_{i+1}\|^2}{2\gamma_i\beta_i} + \beta_i^{-1}L(\mathbf{w}^*) \\
&= L(\mathbf{w}^*) + (1 - \beta_i^{-1})(L(\mathbf{w}_i^{\text{ag}}) - L(\mathbf{w}^*)) + \frac{\|\nabla L(\mathbf{w}_i^{\text{md}}) - \nabla \ell_i(\mathbf{w}_i^{\text{md}})\|^2}{2(\beta_i/\gamma_i - H)} + \frac{\langle \nabla L(\mathbf{w}_i^{\text{md}}) - \nabla \ell_i(\mathbf{w}_i^{\text{md}}), \mathbf{w}_i - \mathbf{w}^* \rangle}{\beta_i} \\
&\quad + \frac{\|\mathbf{w}^* - \mathbf{w}_i\|^2 - \|\mathbf{w}^* - \mathbf{w}_{i+1}\|^2}{2\gamma_i\beta_i}
\end{aligned}$$

Thus we conclude that

$$\begin{aligned}
L(\mathbf{w}_{i+1}^{\text{ag}}) - L(\mathbf{w}^*) &\leq (1 - \beta_i^{-1})(L(\mathbf{w}_i^{\text{ag}}) - L(\mathbf{w}^*)) + \frac{\|\nabla L(\mathbf{w}_i^{\text{md}}) - \nabla \ell_i(\mathbf{w}_i^{\text{md}})\|^2}{2(\beta_i/\gamma_i - H)} + \frac{\langle \nabla L(\mathbf{w}_i^{\text{md}}) - \nabla \ell_i(\mathbf{w}_i^{\text{md}}), \mathbf{w}_i - \mathbf{w}^* \rangle}{\beta_i} \\
&\quad + \frac{\|\mathbf{w}_i - \mathbf{w}^*\|^2 - \|\mathbf{w}_{i+1} - \mathbf{w}^*\|^2}{2\beta_i\gamma_i}
\end{aligned}$$

Multiplying throughout by $\beta_i\gamma_i$ we get

$$\begin{aligned}
\gamma_i\beta_i(L(\mathbf{w}_{i+1}^{\text{ag}}) - L(\mathbf{w}^*)) &\leq \gamma_i(\beta_i - 1)(L(\mathbf{w}_i^{\text{ag}}) - L(\mathbf{w}^*)) + \frac{\gamma_i\beta_i\|\nabla L(\mathbf{w}_i^{\text{md}}) - \nabla \ell_i(\mathbf{w}_i^{\text{md}})\|^2}{2(\beta_i/\gamma_i - H)} \\
&\quad + \frac{\|\mathbf{w}_i - \mathbf{w}^*\|^2 - \|\mathbf{w}_{i+1} - \mathbf{w}^*\|^2}{2} + \gamma_i\langle \nabla L(\mathbf{w}_i^{\text{md}}) - \nabla \ell_i(\mathbf{w}_i^{\text{md}}), \mathbf{w}_i - \mathbf{w}^* \rangle
\end{aligned}$$

Owing to the condition that $\gamma_{i+1}(\beta_{i+1} - 1) \leq \gamma_i\beta_i$ we have that

$$\begin{aligned}
\gamma_{i+1}(\beta_{i+1} - 1)(L(\mathbf{w}_{i+1}^{\text{ag}}) - L(\mathbf{w}^*)) &\leq \gamma_i(\beta_i - 1)(L(\mathbf{w}_i^{\text{ag}}) - L(\mathbf{w}^*)) + \frac{\gamma_i\beta_i\|\nabla L(\mathbf{w}_i^{\text{md}}) - \nabla \ell_i(\mathbf{w}_i^{\text{md}})\|^2}{2(\beta_i/\gamma_i - H)} \\
&\quad + \frac{\|\mathbf{w}_i - \mathbf{w}^*\|^2 - \|\mathbf{w}_{i+1} - \mathbf{w}^*\|^2}{2} + \gamma_i\langle \nabla L(\mathbf{w}_i^{\text{md}}) - \nabla \ell_i(\mathbf{w}_i^{\text{md}}), \mathbf{w}_i - \mathbf{w}^* \rangle
\end{aligned}$$

Using the above inequality repeatedly we conclude that

$$\begin{aligned}
\gamma_n(\beta_n - 1)(L(\mathbf{w}_n^{\text{ag}}) - L(\mathbf{w}^*)) &\leq \gamma_1(\beta_1 - 1)(L(\mathbf{w}_1^{\text{ag}}) - L(\mathbf{w}^*)) + \sum_{i=1}^{n-1} \frac{\gamma_i\beta_i\|\nabla L(\mathbf{w}_i^{\text{md}}) - \nabla \ell_i(\mathbf{w}_i^{\text{md}})\|^2}{2(\beta_i/\gamma_i - H)} \\
&\quad + \frac{\|\mathbf{w}_1 - \mathbf{w}^*\|^2 - \|\mathbf{w}_{n-1} - \mathbf{w}^*\|^2}{2} + \sum_{i=1}^{n-1} \gamma_i\langle \nabla L(\mathbf{w}_i^{\text{md}}) - \nabla \ell_i(\mathbf{w}_i^{\text{md}}), \mathbf{w}_i - \mathbf{w}^* \rangle \\
&\leq \gamma_1(\beta_1 - 1)(L(\mathbf{w}_1^{\text{ag}}) - L(\mathbf{w}^*)) + \sum_{i=1}^{n-1} \frac{\gamma_i\beta_i\|\nabla L(\mathbf{w}_i^{\text{md}}) - \nabla \ell_i(\mathbf{w}_i^{\text{md}})\|^2}{2(\beta_i/\gamma_i - H)} \\
&\quad + \frac{\|\mathbf{w}^*\|^2}{2} + \sum_{i=1}^{n-1} \gamma_i\langle \nabla L(\mathbf{w}_i^{\text{md}}) - \nabla \ell_i(\mathbf{w}_i^{\text{md}}), \mathbf{w}_i - \mathbf{w}^* \rangle \\
&= \gamma_1(\beta_1 - 1)(L(\mathbf{w}_1^{\text{ag}}) - L(\mathbf{w}^*)) + \sum_{i=1}^{n-1} \frac{\gamma_i\beta_i\|\nabla L(\mathbf{w}_i^{\text{md}}) - \nabla \ell_i(\mathbf{w}_i^{\text{md}})\|^2}{2(\beta_i/\gamma_i - H)} \\
&\quad + \frac{D^2}{2} + \sum_{i=1}^{n-1} \gamma_i\langle \nabla L(\mathbf{w}_i^{\text{md}}) - \nabla \ell_i(\mathbf{w}_i^{\text{md}}), \mathbf{w}_i - \mathbf{w}^* \rangle
\end{aligned}$$

since $2H\gamma_i \leq \beta_i$,

$$\begin{aligned}
&\leq \gamma_1(\beta_1 - 1)(L(\mathbf{w}_1^{\text{ag}}) - L(\mathbf{w}^*)) + \sum_{i=1}^{n-1} \gamma_i^2 \|\nabla L(\mathbf{w}_i^{\text{md}}) - \nabla \ell_i(\mathbf{w}_i^{\text{md}})\|^2 + \frac{D^2}{2} \\
&\quad + \sum_{i=1}^{n-1} \gamma_i \langle \nabla L(\mathbf{w}_i^{\text{md}}) - \nabla \ell_i(\mathbf{w}_i^{\text{md}}), \mathbf{w}_i - \mathbf{w}^* \rangle \\
&\leq \gamma_1(\beta_1 - 1)L(\mathbf{w}_1^{\text{ag}}) + \sum_{i=1}^{n-1} 2\gamma_i^2 \|\nabla L(\mathbf{w}_i^{\text{ag}}) - \nabla \ell_i(\mathbf{w}_i^{\text{ag}})\|^2 + \frac{D^2}{2} \\
&\quad + \sum_{i=1}^{n-1} \gamma_i \langle \nabla L(\mathbf{w}_i^{\text{md}}) - \nabla \ell_i(\mathbf{w}_i^{\text{md}}), \mathbf{w}_i - \mathbf{w}^* \rangle \\
&\quad + \sum_{i=1}^{n-1} 2\gamma_i^2 \|\nabla L(\mathbf{w}_i^{\text{md}}) - \nabla \ell_i(\mathbf{w}_i^{\text{md}}) - \nabla L(\mathbf{w}_i^{\text{ag}}) + \nabla \ell_i(\mathbf{w}_i^{\text{ag}})\|^2
\end{aligned}$$

Taking expectation we get that

$$\begin{aligned}
\gamma_n(\beta_n - 1)(\mathbb{E}[L(\mathbf{w}_n^{\text{ag}})] - L(\mathbf{w}^*)) &\leq \gamma_1(\beta_1 - 1)L(\mathbf{w}_1^{\text{ag}}) + \sum_{i=1}^{n-1} 2\gamma_i^2 \mathbb{E}[\|\nabla L(\mathbf{w}_i^{\text{ag}}) - \nabla \ell_i(\mathbf{w}_i^{\text{ag}})\|^2] + \frac{D^2}{2} \\
&\quad + \sum_{i=1}^{n-1} 2\gamma_i^2 \mathbb{E}[\|\nabla L(\mathbf{w}_i^{\text{md}}) - \nabla \ell_i(\mathbf{w}_i^{\text{md}}) - \nabla L(\mathbf{w}_i^{\text{ag}}) + \nabla \ell_i(\mathbf{w}_i^{\text{ag}})\|^2]
\end{aligned} \tag{10}$$

Now note that

$$\nabla L(\mathbf{w}_i^{\text{ag}}) - \nabla \ell_i(\mathbf{w}_i^{\text{ag}}) = \frac{1}{b} \sum_{t=(i-1)b+1}^{bi} (\nabla L(\mathbf{w}_i^{\text{ag}}) - \nabla \ell(\mathbf{w}_i^{\text{ag}}, z_t)) \quad \text{and}$$

$$\nabla L(\mathbf{w}_i^{\text{ag}}) - \nabla \ell_i(\mathbf{w}_i^{\text{ag}}) - \nabla L(\mathbf{w}_i^{\text{md}}) + \nabla \ell_i(\mathbf{w}_i^{\text{md}}) = \frac{1}{b} \sum_{t=(i-1)b+1}^{bi} (\nabla L(\mathbf{w}_i^{\text{ag}}) - \nabla \ell(\mathbf{w}_i^{\text{ag}}, z_t) - \nabla L(\mathbf{w}_i^{\text{md}}) + \nabla \ell(\mathbf{w}_i^{\text{md}}, z_t))$$

Further $(\nabla L(\mathbf{w}_i) - \nabla \ell(\mathbf{w}_i, z_t))$ and $(\nabla L(\mathbf{w}_i^{\text{ag}}) - \nabla \ell(\mathbf{w}_i^{\text{ag}}, z_t) - \nabla L(\mathbf{w}_i^{\text{md}}) + \nabla \ell(\mathbf{w}_i^{\text{md}}, z_t))$ are mean zero vectors drawn i.i.d. Also note that \mathbf{w}_i^{ag} only depends on the first $(i-1)b$ examples and so when we consider expectation w.r.t. $z_{(i-1)b+1}, \dots, z_{ib}$, \mathbf{w}_i is fixed. Hence we have that

$$\begin{aligned}
\mathbb{E}[\|\nabla L(\mathbf{w}_i^{\text{ag}}) - \nabla \ell_i(\mathbf{w}_i^{\text{ag}})\|^2] &= \frac{1}{b^2} \mathbb{E} \left[\left\| \sum_{t=(i-1)b+1}^{bi} (\nabla L(\mathbf{w}_i^{\text{ag}}) - \nabla \ell(\mathbf{w}_i^{\text{ag}}, z_t)) \right\|^2 \right] \\
&\leq \frac{1}{b^2} \sum_{t=(i-1)b+1}^{bi} \mathbb{E}[\|\nabla L(\mathbf{w}_i^{\text{ag}}) - \nabla \ell(\mathbf{w}_i^{\text{ag}}, z_t)\|^2]
\end{aligned}$$

and similarly

$$\begin{aligned}
&\mathbb{E}[\|\nabla L(\mathbf{w}_i^{\text{ag}}) - \nabla \ell_i(\mathbf{w}_i^{\text{ag}}) - \nabla L(\mathbf{w}_i^{\text{md}}) + \nabla \ell_i(\mathbf{w}_i^{\text{md}})\|^2] \\
&\leq \frac{1}{b^2} \sum_{t=(i+1)b+1}^{bi} \mathbb{E}[\|\nabla L(\mathbf{w}_i^{\text{ag}}) - \nabla \ell(\mathbf{w}_i^{\text{ag}}, z_t) - \nabla L(\mathbf{w}_i^{\text{md}}) + \nabla \ell(\mathbf{w}_i^{\text{md}}, z_t)\|^2]
\end{aligned}$$

Plugging these back in Equation 10 we get :

$$\begin{aligned}
\gamma_n(\beta_n - 1) (\mathbb{E}[L(\mathbf{w}_n^{\text{ag}})] - L(\mathbf{w}^*)) &\leq \gamma_1(\beta_1 - 1)L(\mathbf{w}_1^{\text{ag}}) + \sum_{i=1}^{n-1} \frac{2\gamma_i^2}{b^2} \sum_{t=(i-1)b+1}^{bi} \mathbb{E} \left[\|\nabla L(\mathbf{w}_i^{\text{ag}}) - \nabla \ell(\mathbf{w}_i^{\text{ag}}, z_t)\|^2 \right] + \frac{D^2}{2} \\
&\quad + \sum_{i=1}^{n-1} \frac{2\gamma_i^2}{b} \sum_{t=(i+1)b+1}^{bi} \mathbb{E} \left[\left\| \nabla L(\mathbf{w}_i^{\text{ag}}) - \nabla \ell(\mathbf{w}_i^{\text{ag}}, z_t) - \nabla L(\mathbf{w}_i^{\text{md}}) + \nabla \ell(\mathbf{w}_i^{\text{md}}, z_t) \right\|^2 \right] \\
&\leq \gamma_1(\beta_1 - 1)L(\mathbf{w}_1^{\text{ag}}) + \sum_{i=1}^{n-1} \frac{4\gamma_i^2}{b^2} \sum_{t=(i-1)b+1}^{bi} \mathbb{E} \left[\|\nabla L(\mathbf{w}_i^{\text{ag}})\|^2 + \|\nabla \ell(\mathbf{w}_i^{\text{ag}}, z_t)\|^2 \right] + \frac{D^2}{2} \\
&\quad + \sum_{i=1}^{n-1} \frac{4\gamma_i^2}{b} \sum_{t=(i+1)b+1}^{bi} \mathbb{E} \left[\left\| \nabla L(\mathbf{w}_i^{\text{ag}}) - \nabla L(\mathbf{w}_i^{\text{md}}) \right\|^2 + \left\| \nabla \ell(\mathbf{w}_i^{\text{md}}, z_t) - \nabla \ell(\mathbf{w}_i^{\text{ag}}, z_t) \right\|^2 \right]
\end{aligned}$$

for any non-negative H -smooth convex function f , we have the self-bounding property that $\|\nabla f(\mathbf{w})\| \leq \sqrt{4Hf(\mathbf{w})}$. Using this,

$$\begin{aligned}
&\leq \gamma_1(\beta_1 - 1)L(\mathbf{w}_1^{\text{ag}}) + \sum_{i=1}^{n-1} \frac{16H\gamma_i^2}{b^2} \sum_{t=(i-1)b+1}^{bi} \mathbb{E} [L(\mathbf{w}_i^{\text{ag}}) + \ell(\mathbf{w}_i^{\text{ag}}, z_t)] + \frac{D^2}{2} \\
&\quad + \sum_{i=1}^{n-1} \frac{4\gamma_i^2}{b^2} \sum_{t=(i+1)b+1}^{bi} \mathbb{E} \left[\left\| \nabla L(\mathbf{w}_i^{\text{ag}}) - \nabla L(\mathbf{w}_i^{\text{md}}) \right\|^2 + \left\| \nabla \ell(\mathbf{w}_i^{\text{md}}, z_t) - \nabla \ell(\mathbf{w}_i^{\text{ag}}, z_t) \right\|^2 \right] \\
&= \gamma_1(\beta_1 - 1)L(\mathbf{w}_1^{\text{ag}}) + \sum_{i=1}^{n-1} \frac{32H\gamma_i^2}{b} \mathbb{E} [L(\mathbf{w}_i^{\text{ag}})] + \frac{D^2}{2} \\
&\quad + \sum_{i=1}^{n-1} \frac{4\gamma_i^2}{b^2} \sum_{t=(i+1)b+1}^{bi} \mathbb{E} \left[\left\| \nabla L(\mathbf{w}_i^{\text{ag}}) - \nabla L(\mathbf{w}_i^{\text{md}}) \right\|^2 + \left\| \nabla \ell(\mathbf{w}_i^{\text{md}}, z_t) - \nabla \ell(\mathbf{w}_i^{\text{ag}}, z_t) \right\|^2 \right]
\end{aligned}$$

by H -smoothness of L and ℓ we have that $\|\nabla L(\mathbf{w}_i^{\text{ag}}) - \nabla L(\mathbf{w}_i^{\text{md}})\| \leq H \|\mathbf{w}_i^{\text{ag}} - \mathbf{w}_i^{\text{md}}\|$. Similarly we also have that $\|\nabla \ell(\mathbf{w}_i^{\text{ag}}, z_t) - \nabla \ell(\mathbf{w}_i^{\text{md}}, z_t)\| \leq H \|\mathbf{w}_i^{\text{ag}} - \mathbf{w}_i^{\text{md}}\|$. Hence,

$$\begin{aligned}
&\leq \gamma_1(\beta_1 - 1)L(\mathbf{w}_1^{\text{ag}}) + \sum_{i=1}^{n-1} \frac{32H\gamma_i^2}{b} \mathbb{E} [L(\mathbf{w}_i^{\text{ag}})] + \frac{D^2}{2} \\
&\quad + \sum_{i=1}^{n-1} \frac{8H^2\gamma_i^2}{b} \mathbb{E} \left[\left\| \mathbf{w}_i^{\text{ag}} - \mathbf{w}_i^{\text{md}} \right\|^2 \right]
\end{aligned}$$

However, $\mathbf{w}_i^{\text{md}} \leftarrow \beta_i^{-1} \mathbf{w}_i + (1 - \beta_i^{-1}) \mathbf{w}_i^{\text{ag}}$. Hence $\|\mathbf{w}_i^{\text{ag}} - \mathbf{w}_i^{\text{md}}\|^2 \leq \frac{\|\mathbf{w}_i - \mathbf{w}_i^{\text{ag}}\|^2}{\beta_i^2} \leq \frac{2D^2}{\beta_i^2}$. Hence,

$$\leq \gamma_1(\beta_1 - 1)L(\mathbf{w}_1^{\text{ag}}) + \sum_{i=1}^{n-1} \frac{32H\gamma_i^2}{b} \mathbb{E} [L(\mathbf{w}_i^{\text{ag}})] + \frac{D^2}{2} + \frac{16H^2D^2}{b} \sum_{i=1}^{n-1} \frac{\gamma_i^2}{\beta_i^2}$$

Dividing throughout by $\gamma_n(\beta_n - 1)$ concludes the proof. \square

Proof of Theorem 2. First note that the for any i ,

$$2H\gamma_i = 2H\gamma_i^p \leq \frac{i^p}{2} \leq \beta_i$$

Also note that since $p \in [0, 1]$,

$$\gamma_{i+1}(\beta_{i+1} - 1) = \gamma \frac{i(i+1)^p}{2} \leq \gamma \frac{i^p(i+1)}{2} = \gamma_i \beta_i$$

Thus we have verified that the step sizes satisfy the conditions required by previous lemma. From the previous lemma we have that

$$\begin{aligned}
\mathbb{E}[L(\mathbf{w}_n^{\text{ag}})] - L(\mathbf{w}^*) &\leq \frac{\gamma_1(\beta_1 - 1)}{\gamma_n(\beta_n - 1)} L(\mathbf{w}_1^{\text{ag}}) + \frac{32H}{b\gamma_n(\beta_n - 1)} \sum_{i=1}^{n-1} \gamma_i^2 \mathbb{E}[L(\mathbf{w}_i^{\text{ag}})] + \frac{D^2}{2\gamma_n(\beta_n - 1)} + \frac{16H^2 D^2}{b\gamma_n(\beta_n - 1)} \sum_{i=1}^{n-1} \frac{\gamma_i^2}{\beta_i^2} \\
&= \frac{64H\gamma}{bn^p(n-1)} \sum_{i=1}^{n-1} i^{2p} \mathbb{E}[L(\mathbf{w}_i^{\text{ag}})] + \frac{D^2}{\gamma n^p(n-1)} + \frac{128H^2 D^2 \gamma}{bn^p(n-1)} \sum_{i=1}^{n-1} \frac{i^{2p}}{(i+1)^2} \\
&\leq \frac{64H\gamma(n-1)^{2p}}{bn^p(n-1)} \sum_{i=1}^{n-1} \mathbb{E}[L(\mathbf{w}_i^{\text{ag}})] + \frac{D^2}{\gamma(n-1)^{p+1}} + \frac{128H^2 D^2 \gamma}{b(n-1)^{p+1}} \sum_{i=1}^{n-1} \frac{1}{i^{2(1-p)}} \\
&\leq \frac{64H\gamma}{b(n-1)^{1-p}} \sum_{i=1}^{n-1} \mathbb{E}[L(\mathbf{w}_i^{\text{ag}})] + \frac{D^2}{\gamma(n-1)^{p+1}} + \frac{128H^2 D^2 \gamma}{b(n-1)^{p+1}} \sum_{i=1}^{n-1} \frac{1}{i^{2(1-p)}} \\
&\leq \frac{64H\gamma}{b(n-1)^{1-p}} \sum_{i=1}^{n-1} \mathbb{E}[L(\mathbf{w}_i^{\text{ag}})] + \frac{D^2}{\gamma(n-1)^{p+1}} + \frac{128H^2 D^2 \gamma}{b(n-1)} \\
&\leq \frac{64H\gamma}{b(n-1)^{1-p}} \sum_{i=1}^{n-1} (\mathbb{E}[L(\mathbf{w}_i^{\text{ag}})] - L(\mathbf{w}^*)) + \frac{64H\gamma L(\mathbf{w}^*)(n-1)^p}{b} + \frac{D^2}{\gamma(n-1)^{p+1}} + \frac{128H^2 D^2 \gamma}{b(n-1)}
\end{aligned}$$

since $\gamma \leq 1/4H$,

$$\leq \frac{64H\gamma}{b(n-1)^{1-p}} \sum_{i=1}^{n-1} (\mathbb{E}[L(\mathbf{w}_i^{\text{ag}})] - L(\mathbf{w}^*)) + \frac{64H\gamma L(\mathbf{w}^*)(n-1)^p}{b} + \frac{D^2}{\gamma(n-1)^{p+1}} + \frac{32HD^2}{b(n-1)}$$

Thus we have shown that

$$\mathbb{E}[L(\mathbf{w}_n^{\text{ag}})] - L(\mathbf{w}^*) \leq \frac{64H\gamma}{b(n-1)^{1-p}} \sum_{i=1}^{n-1} (\mathbb{E}[L(\mathbf{w}_i^{\text{ag}})] - L(\mathbf{w}^*)) + \frac{64H\gamma L(\mathbf{w}^*)(n-1)^p}{b} + \frac{D^2}{\gamma(n-1)^{p+1}} + \frac{32HD^2}{b(n-1)}$$

Now if we use the notation $a_i = \mathbb{E}[L(\mathbf{w}_i^{\text{ag}})] - L(\mathbf{w}^*)$, $A(i) = \frac{64H\gamma}{b(i-1)^{1-p}}$ and

$$B(i) = \frac{64H\gamma L(\mathbf{w}^*)(i-1)^p}{b} + \frac{D^2}{\gamma(i-1)^{p+1}} + \frac{32HD^2}{b(i-1)}$$

then first note that for any i by smoothness,

$$a_i \leq 4HD^2 + \sqrt{4HD^2 L(\mathbf{w}^*)}$$

Also notice that

$$\sum_{i=n-M-1}^n A(i) = \frac{64H\gamma}{b} \sum_{i=n-M-1}^n \frac{1}{(i-1)^{1-p}} \leq \frac{64H\gamma n^p}{b}$$

Hence as long as

$$\gamma \leq \frac{b}{64Hn^p}, \tag{11}$$

$\sum_{i=n-M-1}^n A(i) \leq 1$. We shall ensure that the γ we choose will satisfy the above condition. Now applying lemma A.2 we get that for any M ,

$$a_n \leq eA(n) \left(a_0(n-M) + \sum_{i=n-M-1}^n B(i) \right) + B(n) \tag{12}$$

Now notice that

$$\begin{aligned}
\sum_{i=n-M-1}^n B(i) &= \frac{64H\gamma L(\mathbf{w}^*)}{b} \sum_{i=n-M-1}^n \frac{1}{(i-1)^p} + \frac{D^2}{\gamma} \sum_{i=n-M-1}^n \frac{1}{(i-1)^{p+1}} + \frac{32HD^2}{b} \sum_{i=n-M-1}^n \frac{1}{(i-1)} \\
&\leq \frac{64H\gamma L(\mathbf{w}^*)(n-M-2)^p}{b} + \frac{D^2}{\gamma(n-M-2)^{p+1}} + \frac{32HD^2}{b(n-M-2)} + \frac{64H\gamma L(\mathbf{w}^*)(n-1)^{p+1}}{b} \\
&\quad + \frac{D^2}{\gamma(n-M-2)^p} + \frac{32HD^2 \log n}{b}
\end{aligned}$$

Plugging this back in Equation 12 we conclude that

$$\begin{aligned}
a_n &\leq \frac{64eH\gamma}{b(n-1)^{1-p}} \left(\left(4HD^2 + \sqrt{4HD^2L(\mathbf{w}^*)} \right) (n-M) + \frac{64H\gamma L(\mathbf{w}^*)(n-M-2)^p}{b} + \frac{D^2}{\gamma(n-M-2)^{p+1}} \right. \\
&\quad \left. + \frac{32HD^2}{b(n-M-2)} + \frac{64H\gamma L(\mathbf{w}^*)(n-1)^{p+1}}{b} + \frac{D^2}{\gamma(n-M-2)^p} + \frac{32HD^2 \log n}{b} \right) \\
&\quad + \frac{64H\gamma L(\mathbf{w}^*)(n-1)^p}{b} + \frac{D^2}{\gamma(n-1)^{p+1}} + \frac{32HD^2}{b(n-1)} \\
&\leq \frac{64eH\gamma}{b(n-1)^{1-p}} \left(\left(4HD^2 + \sqrt{4HD^2L(\mathbf{w}^*)} \right) (n-M-2) + \frac{32HD^2}{b(n-M-2)} + \frac{2D^2}{\gamma(n-M-2)^p} \right. \\
&\quad \left. + \frac{128H\gamma L(\mathbf{w}^*)(n-1)^{p+1}}{b} + \frac{32HD^2 \log n}{b} \right) + \frac{64H\gamma L(\mathbf{w}^*)(n-1)^p}{b} + \frac{D^2}{\gamma(n-1)^{p+1}} + \frac{32HD^2}{b(n-1)}
\end{aligned}$$

since $\gamma \leq \frac{b}{64Hn^p}$, $\frac{32HD^2}{b(n-M-2)} \leq \frac{2D^2}{\gamma(n-M-2)^p}$. Hence

$$\begin{aligned}
&\leq \frac{64eH\gamma}{b(n-1)^{1-p}} \left(\left(4HD^2 + \sqrt{4HD^2L(\mathbf{w}^*)} \right) (n-M-2) + \frac{3D^2}{\gamma(n-M-2)^p} \right. \\
&\quad \left. + \frac{128H\gamma L(\mathbf{w}^*)(n-1)^{p+1}}{b} + \frac{32HD^2 \log n}{b} \right) + \frac{64H\gamma L(\mathbf{w}^*)(n-1)^p}{b} + \frac{D^2}{\gamma(n-1)^{p+1}} + \frac{32HD^2}{b(n-1)}
\end{aligned}$$

We now optimize over the choice of M above by using

$$(n-M-2) = \left(\frac{3D^2}{\gamma \left(4HD^2 + \sqrt{4HD^2L(\mathbf{w}^*)} \right)} \right)^{\frac{1}{p+1}}$$

Ofcourse for the choice of M to be valid we need that $n-M-2 \leq n$ which gives our second condition on γ which is

$$\gamma \geq \frac{3D^2}{n^{p+1} \left(4HD^2 + \sqrt{4HD^2L(\mathbf{w}^*)} \right)} \quad (13)$$

Plugging in this M we get,

$$\begin{aligned}
a_n &\leq \frac{64eH\gamma}{b(n-1)^{1-p}} \left(2 \left(4HD^2 + \sqrt{4HD^2L(\mathbf{w}^*)} \right)^{\frac{p}{p+1}} \left(\frac{3D^2}{\gamma} \right)^{\frac{1}{p+1}} + \frac{128H\gamma L(\mathbf{w}^*)(n-1)^{p+1}}{b} + \frac{32HD^2 \log n}{b} \right) \\
&\quad + \frac{64H\gamma L(\mathbf{w}^*)(n-1)^p}{b} + \frac{D^2}{\gamma(n-1)^{p+1}} + \frac{32HD^2}{b(n-1)} \\
&= \frac{128eH\gamma^{\frac{p}{p+1}} \left(4HD^2 + \sqrt{4HD^2L(\mathbf{w}^*)} \right)^{\frac{p}{p+1}} (3D^2)^{\frac{1}{p+1}}}{b(n-1)^{1-p}} + \frac{2e(64H\gamma)^2 L(\mathbf{w}^*)(n-1)^{2p}}{b^2} + \frac{2e(32H)^2 D^2 \gamma \log n}{b^2(n-1)^{1-p}} \\
&\quad + \frac{64H\gamma L(\mathbf{w}^*)(n-1)^p}{b} + \frac{D^2}{\gamma(n-1)^{p+1}} + \frac{32HD^2}{b(n-1)}
\end{aligned}$$

however by condition in Equation 11, $\gamma \leq \frac{b}{64Hn^p}$, hence

$$\begin{aligned}
&\leq \frac{348H\gamma^{\frac{p}{p+1}} \left(4HD^2 + \sqrt{4HD^2L(\mathbf{w}^*)} \right)^{\frac{p}{p+1}} (3D^2)^{\frac{1}{p+1}}}{b(n-1)^{1-p}} + \frac{2e(32H)^2 D^2 \gamma \log n}{b^2(n-1)^{1-p}} \\
&\quad + \frac{412H\gamma L(\mathbf{w}^*)(n-1)^p}{b} + \frac{D^2}{\gamma(n-1)^{p+1}} + \frac{32HD^2}{b(n-1)} \quad (14)
\end{aligned}$$

We shall try to now optimize the above bound w.r.t. γ , To this end set

$$\gamma = \min \left\{ \frac{1}{4H}, \sqrt{\frac{bD^2}{412HL(\mathbf{w}^*)(n-1)^{2p+1}}}, \left(\frac{b}{1044H(n-1)^{2p}} \right)^{\frac{p+1}{2p+1}} \left(\frac{D^2}{4HD^2 + \sqrt{4HD^2L(\mathbf{w}^*)}} \right)^{\frac{p}{2p+1}} \right\} \quad (15)$$

We first need to verify that this choice of γ satisfies the conditions in Equation 11 and 13. To this end, note that as for the condition in Equation 11,

$$\gamma \leq \left(\frac{b}{1044H(n-1)^{2p}} \right)^{\frac{p+1}{2p+1}} \left(\frac{D^2}{4HD^2 + \sqrt{4HD^2L(\mathbf{w}^*)}} \right)^{\frac{p}{2p+1}}$$

and hence it can be easily verified that for $n \geq 3$, $\gamma \leq \frac{b}{64Hn^p}$. On the other hand to verify the condition in Equation 13, we need to show that

$$\begin{aligned} \gamma &= \min \left\{ \frac{1}{4H}, \sqrt{\frac{bD^2}{412HL(\mathbf{w}^*)(n-1)^{2p+1}}}, \left(\frac{b}{1044H(n-1)^{2p}} \right)^{\frac{p+1}{2p+1}} \left(\frac{D^2}{4HD^2 + \sqrt{4HD^2L(\mathbf{w}^*)}} \right)^{\frac{p}{2p+1}} \right\} \\ &\geq \frac{3D^2}{n^{p+1} \left(4HD^2 + \sqrt{4HD^2L(\mathbf{w}^*)} \right)} \end{aligned}$$

It can be verified that this condition is satisfied as long as,

$$n \geq \max \left\{ 3, \frac{103L(\mathbf{w}^*)}{b}, \frac{904}{b} \right\}$$

So in effect as long as $n \geq 3$ and sample size $nb \geq \max\{904, \frac{103L(\mathbf{w}^*)}{HD^2}\}$ the conditions are satisfied. Now plugging in this choice of γ into the bound in Equation 14, we get

$$\begin{aligned} a_n &\leq \sqrt{\frac{1648HD^2L(\mathbf{w}^*)}{b(n-1)}} + 2 \left(\frac{1044HD^2 \left(4HD^2 + \sqrt{4HD^2L(\mathbf{w}^*)} \right)^{\frac{p}{2p+1}}}{b(n-1)} \right)^{\frac{p+1}{2p+1}} + \frac{32HD^2}{b(n-1)} + \frac{4HD^2}{(n-1)^{p+1}} \\ &\quad + \frac{\log(n)}{5} \left(\frac{75HD^2}{b(n-1)} \right)^{\frac{3p+1}{2p+1}} \left(\frac{1}{4HD^2 + \sqrt{4HD^2L(\mathbf{w}^*)}} \right)^{\frac{p}{2p+1}} \\ &\leq \sqrt{\frac{1648HD^2L(\mathbf{w}^*)}{b(n-1)}} + 2 \left(4HD^2 + \sqrt{4HD^2L(\mathbf{w}^*)} \right) \left(\frac{174}{b(n-1)} \right)^{\frac{p+1}{2p+1}} + \frac{32HD^2}{b(n-1)} + \frac{4HD^2}{(n-1)^{p+1}} \\ &\quad + \frac{6HD^2 \log(n)}{5} \left(\frac{75}{6b(n-1)} \right)^{\frac{3p+1}{2p+1}} \\ &\leq \sqrt{\frac{1648HD^2L(\mathbf{w}^*)}{b(n-1)}} + 348 \left(4HD^2 + \sqrt{4HD^2L(\mathbf{w}^*)} \right) \left(\frac{1}{b(n-1)} \right)^{\frac{p+1}{2p+1}} + \frac{32HD^2}{b(n-1)} + \frac{4HD^2}{(n-1)^{p+1}} \\ &\quad + \frac{36HD^2}{b(n-1)} \frac{\log(n)}{(b(n-1))^{\frac{p}{2p+1}}} \\ &= \sqrt{\frac{1648HD^2L(\mathbf{w}^*)}{b(n-1)}} + \frac{348 \left(4HD^2 + \sqrt{4HD^2L(\mathbf{w}^*)} \right)}{b(n-1)} (b(n-1))^{\frac{p}{2p+1}} + \frac{32HD^2}{b(n-1)} + \frac{4HD^2}{(n-1)^{p+1}} \\ &\quad + \frac{36HD^2}{b(n-1)} \frac{\log(n)}{(b(n-1))^{\frac{p}{2p+1}}} \end{aligned}$$

Picking

$$p = \min \left\{ \max \left\{ \frac{\log(b)}{2 \log(n-1)}, \frac{\log \log(n)}{2 (\log(b(n-1)) - \log \log(n))} \right\}, 1 \right\}$$

we get the bound,

$$a_n \leq 358 \sqrt{\frac{HD^2L(\mathbf{w}^*)}{b(n-1)}} + \frac{1545HD^2}{\sqrt{b(n-1)}} + \frac{1428HD^2\sqrt{\log n}}{b(n-1)} + \frac{4HD^2}{(n-1)^2}$$

This concludes the proof. \square

Lemma A.2. Consider a sequence of non-negative number $a_1, \dots, a_n \in [0, a_0]$ that satisfy

$$a_n \leq A(n) \sum_{i=1}^{n-1} a_i + B(n)$$

where A is decreasing in n . For such a sequence, for any $m \in [n]$, as long as $A(i) \leq 1/2$ for any $i \geq n - m - 1$ and $\sum_{i=n-m-1}^n A(i) \leq 1$ then

$$a_n \leq eA(n) \left(a_0(n-m) + \sum_{i=n-m-1}^n B(i) \right) + B(n)$$

Proof. We shall unroll this recursion. Note that

$$\begin{aligned} a_n &\leq A(n) \sum_{i=1}^{n-1} a_i + B(n) \\ &= A(n) \left(\sum_{i=1}^{n-2} a_i + a_{n-1} \right) + B(n) \\ &\leq A(n) \left(\sum_{i=1}^{n-2} a_i + A(n-1) \sum_{i=1}^{n-2} a_i + B(n-1) \right) + B(n) \\ &= A(n)(1 + A(n-1)) \sum_{i=1}^{n-2} a_i + B(n) + A(n)B(n-1) \\ &\leq A(n)(1 + A(n-1)) \left(\sum_{i=1}^{n-3} a_i + A(n-2) \sum_{i=1}^{n-3} a_i + B(n-2) \right) + B(n) + A(n)B(n-1) \\ &= A(n)(1 + A(n-1))(1 + A(n-2)) \sum_{i=1}^{n-3} a_i + B(n) + A(n)B(n-1) + A(n)(1 + A(n-1))B(n-2) \end{aligned}$$

Continuing so upto m steps we get

$$a_n \leq A(n) \left(\prod_{i=1}^{m-1} (1 + A(n-i)) \right) \sum_{i=1}^{n-m} a_i + B(n) + A(n) \left(\sum_{i=1}^{m-1} \left(\prod_{j=1}^{i-1} (1 + A(n-j)) \right) B(n-i) \right) \quad (16)$$

We would now like to bound in general the term $\prod_{i=1}^{m-1} (1 + A(n-i))$. To this extant note that,

$$\prod_{i=1}^{m-1} (1 + A(n-i)) = \exp \left(\sum_{i=1}^{m-1} \log(1 + A(n-i)) \right)$$

Now assume $A(i) \leq 1/2$ for all $i \geq n - m - 1$ so that $\log(1 + A(n-i)) \leq A(n-i)$. We get

$$\prod_{i=1}^{m-1} (1 + A(n-i)) \leq \exp \left(\sum_{i=1}^{m-1} A(n-i) \right)$$

Now if $\sum_{i=n-m-1}^n A(i) \leq 1$ then we can conclude that

$$\prod_{i=1}^{m-1} (1 + A(n-i)) \leq e$$

Plugging this in Equation A.2 we get

$$\begin{aligned} a_n &\leq eA(n) \left(\sum_{i=1}^{n-m} a_i + \sum_{i=1}^{m-1} B(n-i) \right) + B(n) \\ &= eA(n) \left(\sum_{i=1}^{n-m} a_i + \sum_{i=n-m-1}^n B(i) \right) + B(n) \end{aligned}$$

Now if for each $i \leq n$, $a_i \leq a_0$ then we see that

$$a_n \leq eA(n) \left(a_0(n-m) + \sum_{i=n-m-1}^n B(i) \right) + B(n)$$

Hence we conclude that as long as $\sum_{i=n-m-1}^n A(i) \leq 1$

$$a_n \leq eA(n) \left(a_0(n-m) + \sum_{i=n-m-1}^n B(i) \right) + B(n)$$

□