

## Appendix A. Appendix

### A.1. Definition of Mini-batch Normalized Stochastic Heavy Ball and Stochastic Heavy Ball

---

**Algorithm 2** Mini-batch Stochastic Heavy Ball (SHB)
 

---

**Require:**  $\mathbf{x}_0 \in \mathbb{R}^d$  (initial point),  $\hat{\alpha}_k \in [0, +\infty)$  (learning rate),  $\hat{\beta}_k \in [0, +\infty)$  (momentum parameter),  $b_k \in \mathbb{N}$  (batch size),  $K \in \mathbb{N}$  (steps)

**Ensure:**  $\mathbf{x}_K$

- 1: **for**  $k = 0, 1, \dots, K - 1$  **do**
  - 2:    $\nabla f_{B_k}(\mathbf{x}_k) := \frac{1}{b_k} \sum_{i \in [b_k]} \nabla f_{\xi_{k,i}}(\mathbf{x}_k)$
  - 3:    $\mathbf{m}_k := \nabla f_{B_k}(\mathbf{x}_k) + \hat{\beta}_k \mathbf{m}_{k-1}$
  - 4:    $\mathbf{x}_{k+1} := \mathbf{x}_k - \hat{\alpha}_k \mathbf{m}_k$
  - 5: **end for**
- 

---

**Algorithm 3** Mini-batch Normalized Stochastic Heavy Ball (NSHB)
 

---

**Require:**  $\mathbf{x}_0 \in \mathbb{R}^d$  (initial point),  $\tilde{\alpha}_k \in [0, +\infty)$  (learning rate),  $\tilde{\beta}_k \in [0, 1]$  (momentum parameter),  $b_k \in \mathbb{N}$  (batch size),  $K \in \mathbb{N}$  (steps)

**Ensure:**  $\mathbf{x}_K$

- 1: **for**  $k = 0, 1, \dots, K - 1$  **do**
  - 2:    $\nabla f_{B_k}(\mathbf{x}_k) := \frac{1}{b_k} \sum_{i \in [b_k]} \nabla f_{\xi_{k,i}}(\mathbf{x}_k)$
  - 3:    $\mathbf{m}_k := (1 - \tilde{\beta}_k) \nabla f_{B_k}(\mathbf{x}_k) + \tilde{\beta}_k \mathbf{m}_{k-1}$
  - 4:    $\mathbf{x}_{k+1} := \mathbf{x}_k - \tilde{\alpha}_k \mathbf{m}_k$
  - 5: **end for**
- 

### A.2. Relationship between Mini-batch SHB and NSHB

**Proposition 6** (i) *When we use mini-batch SHB, mini-batch SHB becomes mini-batch NSHB by setting*

$$\hat{\alpha}_k = \tilde{\alpha}_k(1 - \tilde{\beta}_k), \quad \hat{\beta}_k = \frac{\tilde{\beta}_k}{1 - \tilde{\beta}_k}.$$

(ii) *When we use mini-batch NSHB, mini-batch NSHB becomes mini-batch SHB by setting*

$$\tilde{\alpha}_k = \hat{\alpha}_k(1 + \hat{\beta}_k), \quad \tilde{\beta}_k = \frac{\hat{\beta}_k}{1 + \hat{\beta}_k}.$$

**Proof** Mini-batch SHB and NSHB can be expressed as follows from their update rules:

$$\text{mini-batch SHB: } \mathbf{x}_{k+1} := \mathbf{x}_k - \hat{\alpha}_k \nabla f_{B_k}(\mathbf{x}_k) - \hat{\alpha}_k \hat{\beta}_k \mathbf{m}_{k-1},$$

$$\text{mini-batch NSHB: } \mathbf{x}_{k+1} := \mathbf{x}_k - \tilde{\alpha}_k (1 - \tilde{\beta}_k) \nabla f_{B_k}(\mathbf{x}_k) - \tilde{\alpha}_k \tilde{\beta}_k \mathbf{m}_{k-1}.$$

Therefore, the fact that mini-batch SHB is equal to mini-batch NSHB, it is sufficient that  $\hat{\alpha}_k = \tilde{\alpha}_k(1 - \tilde{\beta}_k)$  and  $\hat{\alpha}_k\hat{\beta}_k = \tilde{\alpha}_k\tilde{\beta}_k$  hold. These equations can be solved separately for  $\hat{\alpha}_k$  and  $\hat{\beta}_k$ , and for  $\tilde{\alpha}_k$  and  $\tilde{\beta}_k$ . These proofs have thus been completed.  $\blacksquare$

### A.3. Lemma

Next, we will prove the following lemma.

**Lemma 7** *for all  $k \in \mathbb{N}_0$ ,*

$$\|\nabla f_{B_k}(\mathbf{x}_k)\| \leq G, \quad \|\mathbf{d}_k\| \leq G, \quad \|\mathbf{m}_k\| \leq G.$$

**Proof** First, we will prove that, for all  $k \in \mathbb{N}_0$ ,

$$\|\nabla f_{B_k}(\mathbf{x}_k)\| \leq G. \tag{13}$$

From Condition (C3) and the triangle inequality, we find that, for all  $k \in \mathbb{N}_0$ ,

$$\|\nabla f_{B_k}(\mathbf{x}_k)\| \leq \frac{1}{b_k} \sum_{i=1}^{b_k} \|\nabla f_{\xi_{k,i}}(\mathbf{x}_k)\|. \tag{14}$$

Since  $\xi_{k,i}$  is a random variable sampled from the index set  $[n]$ , we find that, for all  $k \in \mathbb{N}_0$  and  $i \in [n]$ , there exists  $j \in [n]$  such that

$$\|\nabla f_{\xi_{k,i}}(\mathbf{x}_k)\| = \|\nabla f_j(\mathbf{x}_k)\| \leq G. \tag{15}$$

Substituting (15) for (14) implies that (13) holds.

Second, we will prove that, for all  $k \in \mathbb{N}_0$ ,

$$\|\mathbf{d}_k\| \leq G. \tag{16}$$

In the case of  $k = 0$ , the update rule of mini-batch QHM and the triangle inequality together yield

$$\|\mathbf{d}_0\| \leq (1 - \beta_0)\|\nabla f_{B_0}(\mathbf{x}_0)\| + \beta_0\|\mathbf{d}_{-1}\| \leq (1 - \beta_0)G \leq G, \tag{17}$$

where  $\mathbf{d}_{-1} = \mathbf{0}$ . In the case of  $k \in \mathbb{N}$ , we assume (16). Then, as above, we obtain

$$\begin{aligned} \|\mathbf{d}_{k+1}\| &\leq (1 - \beta_k)\|\nabla f_{B_k}(\mathbf{x}_k)\| + \beta_k\|\mathbf{d}_k\| \\ &\leq (1 - \beta_k)G + \beta_kG = G. \end{aligned}$$

Therefore, (16) holds by mathematical induction.  $\|\mathbf{m}_k\| \leq G$  can also be proved by mathematical induction. These proofs have thus been completed.  $\blacksquare$

Next, we will prove the following lemma using the proof of Theorem 4.

**Lemma 8** *The sequence  $\{\mathbf{x}_k\}_{k \in \mathbb{N}_0}$  generated by Algorithm 1 under (C1) – (C3) satisfies*

$$\mathbb{E} \left[ \|\mathbf{d}_k - \nabla f(\mathbf{x}_{k+1})\|^2 | \mathbf{x}_k \right] \leq \beta_k \|\mathbf{d}_{k-1} - \nabla f(\mathbf{x}_k)\|^2 + 2(1 - \beta_k)^2 \frac{\sigma^2}{b_k} + \alpha_k^2 L^2 \left( 2 + \frac{1}{1 - \beta_k} \right) \mathbb{E} [\|\mathbf{m}_k\|^2 | \mathbf{x}_k],$$

where  $\mathbb{E}[\cdot | \mathbf{x}_k]$  is the expectation conditioned on  $\mathbf{x}_k$ .

**Proof** The update rule of mini-batch QHM ensures that, for all  $k \in \mathbb{N}_0$ ,

$$\begin{aligned} \mathbf{d}_k - \nabla f(\mathbf{x}_{k+1}) &= (1 - \beta_k) \nabla f_{B_k}(\mathbf{x}_k) + \beta_k \mathbf{d}_{k-1} - \nabla f(\mathbf{x}_{k+1}) + \nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}_k) \\ &= \beta_k (\mathbf{d}_{k-1} - \nabla f(\mathbf{x}_k)) + (1 - \beta_k) (\nabla f_{B_k}(\mathbf{x}_k) - \nabla f(\mathbf{x}_k)) + (\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}_{k+1})). \end{aligned}$$

Therefore, for all  $k \in \mathbb{N}_0$ ,

$$\begin{aligned} \|\mathbf{d}_k - \nabla f(\mathbf{x}_{k+1})\|^2 &= \beta_k^2 \|\mathbf{d}_{k-1} - \nabla f(\mathbf{x}_k)\|^2 + \|(1 - \beta_k) (\nabla f_{B_k}(\mathbf{x}_k) - \nabla f(\mathbf{x}_k)) + (\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}_{k+1}))\|^2 \\ &\quad + 2\beta_k \langle \mathbf{d}_{k-1} - \nabla f(\mathbf{x}_k), (1 - \beta_k) (\nabla f_{B_k}(\mathbf{x}_k) - \nabla f(\mathbf{x}_k)) + (\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}_{k+1})) \rangle \\ &\leq \beta_k^2 \|\mathbf{d}_{k-1} - \nabla f(\mathbf{x}_k)\|^2 + 2(1 - \beta_k)^2 \|\nabla f_{B_k}(\mathbf{x}_k) - \nabla f(\mathbf{x}_k)\|^2 \\ &\quad + 2\|\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}_{k+1})\|^2 + 2\beta_k \langle \mathbf{d}_{k-1} - \nabla f(\mathbf{x}_k), (1 - \beta_k) (\nabla f_{B_k}(\mathbf{x}_k) - \nabla f(\mathbf{x}_k)) \rangle \\ &\quad + 2\beta_k \langle \mathbf{d}_{k-1} - \nabla f(\mathbf{x}_k), \nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}_{k+1}) \rangle \\ &\leq \beta_k^2 \|\mathbf{d}_{k-1} - \nabla f(\mathbf{x}_k)\|^2 + 2(1 - \beta_k)^2 \|\nabla f_{B_k}(\mathbf{x}_k) - \nabla f(\mathbf{x}_k)\|^2 \\ &\quad + 2\|\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}_{k+1})\|^2 + 2\beta_k \langle \mathbf{d}_{k-1} - \nabla f(\mathbf{x}_k), (1 - \beta_k) (\nabla f_{B_k}(\mathbf{x}_k) - \nabla f(\mathbf{x}_k)) \rangle \\ &\quad + \beta_k \left( \frac{\epsilon}{2} \|\mathbf{d}_{k-1} - \nabla f(\mathbf{x}_k)\|^2 + \frac{1}{2\epsilon} \|\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}_{k+1})\|^2 \right), \end{aligned} \tag{18}$$

where  $\epsilon > 0$ , and in the first and second inequality, we used Young's inequality. From the condition (C1) ( $f$  is  $L$ -smooth), we have

$$\|\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}_{k+1})\|^2 \leq L^2 \|\mathbf{x}_k - \mathbf{x}_{k+1}\|^2 = \alpha_k^2 L^2 \|\mathbf{m}_k\|^2. \tag{19}$$

Combining (19) with (18) ensures that

$$\begin{aligned} \|\mathbf{d}_k - \nabla f(\mathbf{x}_{k+1})\|^2 &\leq \left( \beta_k^2 + \frac{\beta_k \epsilon}{2} \right) \|\mathbf{d}_{k-1} - \nabla f(\mathbf{x}_k)\|^2 + 2(1 - \beta_k)^2 \|\nabla f_{B_k}(\mathbf{x}_k) - \nabla f(\mathbf{x}_k)\|^2 \\ &\quad + 2\alpha_k^2 L^2 \|\mathbf{m}_k\|^2 + 2\beta_k \langle \mathbf{d}_{k-1} - \nabla f(\mathbf{x}_k), (1 - \beta_k) (\nabla f_{B_k}(\mathbf{x}_k) - \nabla f(\mathbf{x}_k)) \rangle \\ &\quad + \frac{\beta_k \alpha_k^2 L^2}{2\epsilon} \|\mathbf{m}_k\|^2 \\ &\leq \left( \beta_k^2 + \frac{\beta_k \epsilon}{2} \right) \|\mathbf{d}_{k-1} - \nabla f(\mathbf{x}_k)\|^2 + 2(1 - \beta_k)^2 \|\nabla f_{B_k}(\mathbf{x}_k) - \nabla f(\mathbf{x}_k)\|^2 \\ &\quad + 2\beta_k \langle \mathbf{d}_{k-1} - \nabla f(\mathbf{x}_k), (1 - \beta_k) (\nabla f_{B_k}(\mathbf{x}_k) - \nabla f(\mathbf{x}_k)) \rangle \\ &\quad + \alpha_k^2 L^2 \left( 2 + \frac{1}{2\epsilon} \right) \|\mathbf{m}_k\|^2, \end{aligned}$$

where, in the last inequality, we used  $\beta_k < 1$ . Since  $\beta_k < 1$ , we can choose  $\epsilon = 2(1 - \beta_k) > 0$ . So, we have

$$\begin{aligned} \|\mathbf{d}_k - \nabla f(\mathbf{x}_{k+1})\|^2 &\leq \beta_k \|\mathbf{d}_{k-1} - \nabla f(\mathbf{x}_k)\|^2 + 2(1 - \beta_k)^2 \|\nabla f_{B_k}(\mathbf{x}_k) - \nabla f(\mathbf{x}_k)\|^2 \\ &\quad + 2\beta_k \langle \mathbf{d}_{k-1} - \nabla f(\mathbf{x}_k), (1 - \beta_k)(\nabla f_{B_k}(\mathbf{x}_k) - \nabla f(\mathbf{x}_k)) \rangle \\ &\quad + \alpha_k^2 L^2 \left(2 + \frac{1}{1 - \beta_k}\right) \|\mathbf{m}_k\|^2. \end{aligned} \quad (20)$$

Condition (C2) guarantees that

$$\mathbb{E}_{\xi_k} [\nabla f_{B_k}(\mathbf{x}_k) | \mathbf{x}_k] = \nabla f(\mathbf{x}_k), \quad \mathbb{V}_{\xi_k} [\nabla f_{B_k}(\mathbf{x}_k) | \mathbf{x}_k] \leq \frac{\sigma^2}{b_k}. \quad (21)$$

Taking the expectation conditioned on  $\mathbf{x}_k$  on both sides of (20) together with (21), guarantees that, for all  $k \in \mathbb{N}_0$ ,

$$\begin{aligned} \mathbb{E} \left[ \|\mathbf{d}_k - \nabla f(\mathbf{x}_{k+1})\|^2 | \mathbf{x}_k \right] &\leq \beta_k \|\mathbf{d}_{k-1} - \nabla f(\mathbf{x}_k)\|^2 + 2(1 - \beta_k)^2 \frac{\sigma^2}{b_k} \\ &\quad + 2\beta_k \mathbb{E} [\langle \mathbf{d}_{k-1} - \nabla f(\mathbf{x}_k), (1 - \beta_k)(\nabla f_{B_k}(\mathbf{x}_k) - \nabla f(\mathbf{x}_k)) \rangle | \mathbf{x}_k] \\ &\quad + \alpha_k^2 \left(2 + \frac{1}{1 - \beta_k}\right) \mathbb{E} [\|\mathbf{m}_k\|^2 | \mathbf{x}_k] \\ &\leq \beta_k \|\mathbf{d}_{k-1} - \nabla f(\mathbf{x}_k)\|^2 + 2(1 - \beta_k)^2 \frac{\sigma^2}{b_k} \\ &\quad + \alpha_k^2 L^2 \left(2 + \frac{1}{1 - \beta_k}\right) \mathbb{E} [\|\mathbf{m}_k\|^2 | \mathbf{x}_k]. \end{aligned} \quad (22)$$

■

#### A.4. Proof of Theorem 2

**Proof** Condition (C1) ( $f$  is  $L$ -smooth) implies that the descent lemma holds (see e.g. Lemma 5.7 in (Beck, 2017)); i.e. for all  $k \in \mathbb{N}_0$ ,

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle + \frac{L}{2} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2, \quad (23)$$

which, together with the update rule of mini-batch QHM, implies that

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \alpha_k \langle \nabla f(\mathbf{x}_k), \mathbf{m}_k \rangle + \frac{\alpha_k^2 L}{2} \|\mathbf{m}_k\|^2. \quad (24)$$

Furthermore, the update rule of mini-batch QHM also ensures that

$$\begin{aligned} \mathbf{m}_k &= (1 - \gamma_k) \nabla f_{B_k}(\mathbf{x}_k) + \gamma_k \mathbf{d}_k \\ &= (1 - \gamma_k) \nabla f_{B_k}(\mathbf{x}_k) + \gamma_k ((1 - \beta_k) \nabla f_{B_k}(\mathbf{x}_k) + \beta_k \mathbf{d}_{k-1}) \\ &= (1 - \gamma_k \beta_k) \nabla f_{B_k}(\mathbf{x}_k) + \gamma_k \beta_k \mathbf{d}_{k-1}. \end{aligned} \quad (25)$$

Substituting (25) for (24) yields

$$\begin{aligned}
 f(\mathbf{x}_{k+1}) &\leq f(\mathbf{x}_k) - \alpha_k(1 - \gamma_k\beta_k) \langle \nabla f(\mathbf{x}_k), \nabla f_{B_k}(\mathbf{x}_k) \rangle - \alpha_k\gamma_k\beta_k \langle \nabla f(\mathbf{x}_k), \mathbf{d}_{k-1} \rangle \\
 &\quad + \frac{\alpha_k^2 L}{2} \|(1 - \gamma_k\beta_k)\nabla f_{B_k}(\mathbf{x}_k) + \gamma_k\beta_k\mathbf{d}_{k-1}\|^2 \\
 &= f(\mathbf{x}_k) - \alpha_k(1 - \gamma_k\beta_k) \langle \nabla f(\mathbf{x}_k), \nabla f_{B_k}(\mathbf{x}_k) \rangle - \alpha_k\gamma_k\beta_k \langle \nabla f(\mathbf{x}_k), \mathbf{d}_{k-1} \rangle \\
 &\quad + \frac{\alpha_k^2 L}{2} ((1 - \gamma_k\beta_k)^2 \|\nabla f_{B_k}(\mathbf{x}_k)\|^2 + 2(1 - \gamma_k\beta_k)\gamma_k\beta_k \langle \nabla f_{B_k}(\mathbf{x}_k), \mathbf{d}_{k-1} \rangle + \gamma_k^2\beta_k^2 \|\mathbf{d}_{k-1}\|^2) \\
 &\leq f(\mathbf{x}_k) - \alpha_k(1 - \gamma_k\beta_k) \langle \nabla f(\mathbf{x}_k), \nabla f_{B_k}(\mathbf{x}_k) \rangle + \alpha_k\gamma_k\beta_k \|\nabla f(\mathbf{x}_k)\| \|\mathbf{d}_{k-1}\| \\
 &\quad + \frac{\alpha_k^2 L}{2} ((1 - \gamma_k\beta_k)^2 \|\nabla f_{B_k}(\mathbf{x}_k)\|^2 + 2(1 - \gamma_k\beta_k)\gamma_k\beta_k \|\nabla f_{B_k}(\mathbf{x}_k)\| \|\mathbf{d}_{k-1}\| + \gamma_k^2\beta_k^2 \|\mathbf{d}_{k-1}\|^2),
 \end{aligned}$$

where, in the last inequality, we used the Cauchy-Schwarz inequality. Using Lemma 7, we have

$$\begin{aligned}
 f(\mathbf{x}_{k+1}) &\leq f(\mathbf{x}_k) - \alpha_k(1 - \gamma_k\beta_k) \langle \nabla f(\mathbf{x}_k), \nabla f_{B_k}(\mathbf{x}_k) \rangle + \alpha_k\gamma_k\beta_k G^2 \\
 &\quad + \frac{\alpha_k^2 L}{2} ((1 - \gamma_k\beta_k)^2 \|\nabla f_{B_k}(\mathbf{x}_k)\|^2 + 2(1 - \gamma_k\beta_k)\gamma_k\beta_k G^2 + \gamma_k^2\beta_k^2 G^2) \\
 &= f(\mathbf{x}_k) - \alpha_k(1 - \gamma_k\beta_k) \langle \nabla f(\mathbf{x}_k), \nabla f_{B_k}(\mathbf{x}_k) \rangle + \alpha_k\gamma_k\beta_k G^2 \\
 &\quad + \frac{\alpha_k^2 L}{2} ((1 - \gamma_k\beta_k)^2 \|\nabla f_{B_k}(\mathbf{x}_k)\|^2 + (2 - \gamma_k\beta_k)\gamma_k\beta_k G^2)
 \end{aligned} \tag{26}$$

From (21), we have

$$\begin{aligned}
 \mathbb{E}_{\xi_k} [\|\nabla f_{B_k}(\mathbf{x}_k)\|^2 | \mathbf{x}_k] &= \|\nabla f(\mathbf{x}_k)\|^2 + \mathbb{V}_{\xi_k} [\nabla f_{B_k}(\mathbf{x}_k) | \mathbf{x}_k] \\
 &\leq \|\nabla f(\mathbf{x}_k)\|^2 + \frac{\sigma^2}{b_k}.
 \end{aligned} \tag{27}$$

Taking the expectation conditioned on  $\mathbf{x}_k$  on both sides of (26), together with (21) and (27), guarantees that, for all  $k \in \mathbb{N}_0$ ,

$$\begin{aligned}
 \mathbb{E}[f(\mathbf{x}_{k+1}) | \mathbf{x}_k] &\leq f(\mathbf{x}_k) - \alpha_k(1 - \gamma_k\beta_k) \|\nabla f(\mathbf{x}_k)\|^2 + \alpha_k\gamma_k\beta_k G^2 \\
 &\quad + \frac{\alpha_k^2 L}{2} \left( (1 - \gamma_k\beta_k)^2 \left( \frac{\sigma^2}{b_k} + \|\nabla f(\mathbf{x}_k)\|^2 \right) + (2 - \gamma_k\beta_k)\gamma_k\beta_k G^2 \right) \\
 &\leq f(\mathbf{x}_k) - \frac{1}{2} \alpha_k(1 - \gamma_k\beta_k)(2 - \alpha_k L(1 - \gamma_k\beta_k)) \|\nabla f(\mathbf{x}_k)\|^2 + \frac{L\sigma^2}{2} \frac{\alpha_k^2(1 - \gamma_k\beta_k)^2}{b_k} \\
 &\quad + \frac{1}{2} \alpha_k\gamma_k\beta_k(2 + \alpha_k L(2 - \gamma_k\beta_k)) G^2.
 \end{aligned}$$

Therefore, taking the total expectation on both sides of the above inequality ensures that, for all  $k \in \mathbb{N}_0$ ,

$$\begin{aligned}
 \frac{1}{2} \alpha_k(1 - \gamma_k\beta_k)(2 - \alpha_k L(1 - \gamma_k\beta_k)) \mathbb{E} [\|\nabla f(\mathbf{x}_k)\|^2] &\leq \mathbb{E}[f(\mathbf{x}_k) - f(\mathbf{x}_{k+1})] + \frac{L\sigma^2}{2} \frac{\alpha_k^2(1 - \gamma_k\beta_k)^2}{b_k} \\
 &\quad + \frac{G^2}{2} \alpha_k\gamma_k\beta_k(2 + \alpha_k L(2 - \gamma_k\beta_k)).
 \end{aligned}$$

Let  $K \in \mathbb{N}$ . Summing the above inequality from  $k = 0$  to  $k = K - 1$  ensures that

$$\begin{aligned} \frac{1}{2} \sum_{k=0}^{K-1} \alpha_k (1 - \gamma_k \beta_k) (2 - \alpha_k L (1 - \gamma_k \beta_k)) \mathbb{E} [\|\nabla f(\mathbf{x}_k)\|^2] &\leq \mathbb{E} [f(\mathbf{x}_0) - f(\mathbf{x}_K)] + \frac{L\sigma^2}{2} \sum_{k=0}^{K-1} \frac{\alpha_k^2 (1 - \gamma_k \beta_k)^2}{b_k} \\ &\quad + \frac{G^2}{2} \sum_{k=0}^{K-1} \alpha_k \gamma_k \beta_k (2 + \alpha_k L (2 - \gamma_k \beta_k)). \end{aligned}$$

From (C1) on the lower bound  $f_\star$  of  $f$ ,  $\alpha_k \in [\alpha_{\min}, \alpha_{\max}]$ , and  $0 \leq \gamma_k \beta_k \leq \overline{\gamma\beta} < 1$ , we have that

$$\begin{aligned} \frac{1}{2} (1 - \overline{\gamma\beta}) (2 - \alpha_{\max} L) \sum_{k=0}^{K-1} \alpha_k \mathbb{E} [\|\nabla f(\mathbf{x}_k)\|^2] &\leq f(\mathbf{x}_0) - f_\star + \frac{L\sigma^2}{2} \sum_{k=0}^{K-1} \frac{\alpha_k^2}{b_k} \\ &\quad + \alpha_{\max} (1 + \alpha_{\max} L) G^2 \sum_{k=0}^{K-1} \gamma_k \beta_k. \end{aligned}$$

Since,  $\alpha_{\max} \leq \frac{2}{L}$ , we obtain

$$\begin{aligned} \min_{k \in [0:K-1]} \mathbb{E} [\|\nabla f(\mathbf{x}_k)\|^2] &\leq \frac{2(f(\mathbf{x}_0) - f_\star)}{(1 - \overline{\gamma\beta})(2 - \alpha_{\max} L)} \underbrace{\frac{1}{\sum_{k=0}^{K-1} \alpha_k}}_{A_K} \\ &\quad + \frac{L\sigma^2}{(1 - \overline{\gamma\beta})(2 - \alpha_{\max} L)} \underbrace{\frac{\sum_{k=0}^{K-1} \alpha_k^2 / b_k}{\sum_{k=0}^{K-1} \alpha_k}}_{B_K} \\ &\quad + \frac{2\alpha_{\max}(1 + \alpha_{\max} L)G^2}{(1 - \overline{\gamma\beta})(2 - \alpha_{\max} L)} \underbrace{\frac{\sum_{k=0}^{K-1} \gamma_k \beta_k}{\sum_{k=0}^{K-1} \alpha_k}}_{C_k}. \end{aligned}$$

This proves Theorem 2(i). Second, we assume the condition of Theorem 2(ii).  $\sum_{k=0}^{+\infty} \alpha_k = +\infty$ ,  $\sum_{k=0}^{+\infty} \gamma_k \beta_k < +\infty$ , and  $\sum_{k=0}^{+\infty} \frac{\alpha_k^2}{b_k} < +\infty$  imply that

$$\lim_{K \rightarrow \infty} A_K = \lim_{K \rightarrow \infty} B_K = \lim_{K \rightarrow \infty} C_K = 0.$$

This proves Theorem 2(ii). ■

### A.5. Proof of Corollary 3

**Proof** [Constant LR]: We have that

$$A_K = \frac{1}{\sum_{k=0}^{K-1} \alpha_k} = \frac{1}{\alpha_{\max} K}.$$

In the case of using [Constant BS (1)], we obtain

$$B_K = \frac{\sum_{k=0}^{K-1} \alpha_k^2}{b \sum_{k=0}^{K-1} \alpha_k} = \frac{\alpha_{\max}}{b}.$$

Also, in the case of using [Exponential BS (2)], since  $b'_m$  is monotone increasing,  $T_m = \left\lceil \frac{n}{b'_m} \right\rceil$  is monotone decreasing. Hence, we have that

$$\sum_{k=0}^{K-1} \frac{1}{b_k} = \sum_{m=0}^{M-1} \sum_{t=0}^{T_m-1} \frac{1}{b'_m} \leq T_0 \sum_{m=0}^{M-1} \frac{1}{b'_m} = \frac{T_0 E}{b_0} \sum_{m'=0}^{M'} \frac{1}{\delta^{m'}} \leq \frac{T_0 E \delta}{b_0(\delta - 1)}.$$

Therefore,

$$B_K = \frac{\sum_{k=0}^{K-1} \alpha_k^2 / b_k}{\sum_{k=0}^{K-1} \alpha_k} \leq \frac{\alpha_{\max} T_0 E \delta}{b_0(\delta - 1) K}.$$

In the case of [Exponential BS (2)], when using [Step Decay Beta (9)] and [Step Decay Gamma (12)], we find that

$$\sum_{k=0}^{K-1} \gamma_k \beta_k = \sum_{m=0}^{M-1} \sum_{t=0}^{T_m-1} \gamma'_m \beta'_m \leq \gamma_{\max} \beta_{\max} T_0 E \sum_{m'=0}^{M'} (\lambda \zeta)^{m'} \leq \frac{\gamma_{\max} \beta_{\max} T_0 E}{1 - \lambda \zeta}.$$

This result shows that, by setting  $T_m = T_0 = \left\lceil \frac{n}{b} \right\rceil$ , we can obtain the same outcome even when using [Constant BS (1)].

[Sqrt-Decaying LR]: We have that

$$\alpha_k = \frac{\alpha_{\max}}{\sqrt{\left\lfloor \frac{k}{T_m} \right\rfloor + 1}} \geq \frac{\alpha_{\max}}{\sqrt{k+1}}.$$

Therefore,

$$\sum_{k=0}^{K-1} \alpha_k \geq \alpha_{\max} \int_0^K \frac{dk}{\sqrt{k+1}} = 2\alpha_{\max}(\sqrt{K+1} - 1).$$

Hence,

$$A_K = \frac{1}{\sum_{k=0}^{K-1} \alpha_k} \leq \frac{1}{2\alpha_{\max}(\sqrt{K+1} - 1)},$$

$$C_K = \frac{\sum_{k=0}^{K-1} \gamma_k \beta_k}{\sum_{k=0}^{K-1} \alpha_k} \leq \frac{\gamma_{\max} \beta_{\max} T_0 E}{2\alpha_{\max}(1 - \lambda \zeta)(\sqrt{K+1} - 1)}.$$

We also have

$$\sum_{m=0}^{M-1} \alpha'_m \leq \alpha_{\max}^2 \left( 1 + \int_0^M \frac{dm}{m+1} \right) = \alpha_{\max}^2 (1 + \log(M+1)) \leq \alpha_{\max}^2 (1 + \log(K+1)),$$

which implies that

$$\sum_{k=0}^{K-1} \alpha_k^2 = T \sum_{m=0}^{M-1} \alpha'_k{}^2 \leq \alpha_{\max}^2 T (1 + \log(K+1)).$$

Therefore, in the case of [Constant BS (1)],

$$B_K = \frac{\sum_{k=0}^{K-1} \alpha_k^2}{b \sum_{k=0}^{K-1} \alpha_k} \leq \frac{\alpha_{\max} T (1 + \log(K+1))}{2b(\sqrt{K+1} - 1)}.$$

Similarly, in the case of [Exponential BS (2)],

$$B_K = \frac{\sum_{k=0}^{K-1} \alpha_k^2 / b_k}{\sum_{k=0}^{K-1} \alpha_k} \leq \frac{\alpha_{\max} T_0 E \delta}{2b_0(\delta - 1)(\sqrt{K+1} - 1)}.$$

[Cosine LR]: If  $M = 1$ , [Cosine LR] corresponds to [Constant LR]. So, let  $M \geq 2$  and let  $T_{\min}$  satisfy  $0 \leq T_{\min} \leq T_m$ . Consequently, we have

$$\begin{aligned} \sum_{k=0}^{K-1} \alpha_k &= \alpha_{\min} K + \frac{\alpha_{\max} - \alpha_{\min}}{2} K + \frac{\alpha_{\max} - \alpha_{\min}}{2} \sum_{m=0}^{M-1} T_m \cos \frac{m\pi}{M-1} \\ &\geq \alpha_{\min} K + \frac{\alpha_{\max} - \alpha_{\min}}{2} K + \frac{\alpha_{\max} - \alpha_{\min}}{2} T_{\min} \sum_{m=0}^{M-1} \cos \frac{m\pi}{M-1}. \end{aligned} \quad (28)$$

Here,  $\sum_{m=0}^{M-1} \cos \frac{m\pi}{M-1} = 0$  holds. Actually, in the case of  $M = 2F$  ( $F \in \mathbb{N}$ ), we have that

$$\sum_{m=0}^{2F-1} \cos \frac{m\pi}{M-1} = \cos 0 + \cos \frac{\pi}{M-1} + \cdots + \cos \frac{F-1}{M-1} \pi + \cos \frac{F}{M-1} \pi + \cdots + \cos \frac{M-2}{M-1} \pi + \cos \pi. \quad (29)$$

From  $\cos(\pi - x) = -\cos x$ , (29) satisfies

$$\cos 0 + \cos \pi = \cos \frac{\pi}{M-1} + \cos \frac{M-2}{M-1} \pi = \cdots = \cos \frac{k-1}{M-1} \pi + \cos \frac{k}{M-1} \pi = 0.$$

Therefore,

$$\sum_{m=0}^{M-1} \cos \frac{m\pi}{M-1} = 0.$$

Next, in the case of  $M = 2F + 1$ , we have

$$\sum_{m=0}^{2F} \cos \frac{m\pi}{M-1} = \cos 0 + \cos \frac{\pi}{M-1} + \cdots + \cos \frac{F-1}{M-1} \pi + \cos \frac{F}{M-1} \pi + \cos \frac{F+1}{M-1} \pi + \cdots + \cos \pi.$$



Since  $\cos \frac{G}{M-1}\pi = \cos \frac{k}{2k}\pi = 0$  and  $\cos(\pi - x) = -\cos x$ , the same holds as in the case above. Hence,  $\sum_{m=0}^{M-1} \cos \frac{m\pi}{M-1} = 0$  holds. So, (28) satisfies

$$\sum_{k=0}^{K-1} \alpha_k \geq \frac{\alpha_{\max} + \alpha_{\min}}{2} K.$$

Therefore,

$$A_K = \frac{1}{\sum_{k=0}^{K-1} \alpha_k} \leq \frac{2}{(\alpha_{\max} + \alpha_{\min})K},$$

$$C_K = \frac{\sum_{k=0}^{K-1} \gamma_k \beta_k}{\sum_{k=0}^{K-1} \alpha_k} \leq \frac{2\gamma_{\max} \beta_{\max} T_0 E}{(1 - \lambda \zeta)(\alpha_{\max} + \alpha_{\min})K}.$$

In the case of using [Constant BS (1)], we have

$$\begin{aligned} \sum_{k=0}^{K-1} \alpha_k^2 &= \alpha_{\min}^2 K + \alpha_{\min}(\alpha_{\max} - \alpha_{\min})K + \alpha_{\min}(\alpha_{\max} - \alpha_{\min})T_0 \sum_{m=0}^{M-1} \cos \frac{m\pi}{M-1} \\ &\quad + \frac{(\alpha_{\max} - \alpha_{\min})^2}{4} K + \frac{(\alpha_{\max} - \alpha_{\min})^2}{2} T_0 \sum_{m=0}^{M-1} \cos \frac{m\pi}{M-1} + \frac{(\alpha_{\max} - \alpha_{\min})^2}{4} T_0 \sum_{m=0}^{M-1} \cos^2 \frac{m\pi}{M-1} \end{aligned}$$

Here,

$$\begin{aligned} \sum_{m=0}^{M-1} \cos^2 \frac{m\pi}{M-1} &\leq 1 + \int_0^{M-1} \cos^2 \frac{m\pi}{M-1} dm \\ &= 1 + \frac{1}{2} \int_0^{M-1} 1 + \cos 2 \frac{m\pi}{M-1} dm \\ &= \frac{M+1}{2}. \end{aligned}$$

From this and  $\sum_{m=0}^{M-1} \cos \frac{m\pi}{M-1} = 0$ , we obtain

$$\begin{aligned} \sum_{k=0}^{K-1} \alpha_k^2 &\leq \alpha_{\min} \alpha_{\max} K + \frac{\alpha_{\max}^2}{4} K - \frac{\alpha_{\max} \alpha_{\min}}{2} K + \frac{\alpha_{\min}^2}{4} K + \frac{(M+1)(\alpha_{\max} - \alpha_{\min})^2}{4} T_0 \\ &= \left( \frac{\alpha_{\max} + \alpha_{\min}}{2} \right)^2 K + \frac{(M+1)(\alpha_{\max} - \alpha_{\min})^2}{4} T_0 \\ &= \left( \left( \frac{\alpha_{\max} + \alpha_{\min}}{2} \right)^2 + \frac{(\alpha_{\max} - \alpha_{\min})^2}{4} \right) K + \frac{(\alpha_{\max} - \alpha_{\min})^2}{4} T_0 \\ &= \frac{\alpha_{\max}^2 + \alpha_{\min}^2}{2} K + \frac{(\alpha_{\max} - \alpha_{\min})^2}{4} T_0 \\ &\leq \frac{(\alpha_{\max} + \alpha_{\min})^2}{2} K + \frac{(\alpha_{\max} - \alpha_{\min})^2}{4} T_0, \end{aligned}$$

which implies that

$$B_K = \frac{\sum_{k=0}^{K-1} \alpha_k^2}{b \sum_{k=0}^{K-1} \alpha_k} \leq \frac{\alpha_{\max} + \alpha_{\min}}{b} + \frac{T_0(\alpha_{\max} - \alpha_{\min})^2}{2b(\alpha_{\max} + \alpha_{\min})K}.$$

On the other hand, in the case of using [Exponential BS (2)], we obtain

$$B_K = \frac{\sum_{k=0}^{K-1} \alpha_k^2 / b_k}{\sum_{k=0}^{K-1} \alpha_k} \leq \frac{\alpha_{\max}^2 \sum_{k=0}^{K-1} 1/b_k}{\sum_{k=0}^{K-1} \alpha_k} \leq \frac{2\alpha_{\max}^2 T_0 E \delta}{b_0(\delta - 1)(\alpha_{\max} + \alpha_{\min})K}.$$

[Polynomial LR]: Since  $f(x) = (1 - x)^p$  is monotone decreasing for  $x \in [0, 1]$ , we have that

$$\int_0^1 (1 - x)^p dx \leq \frac{1}{M} \sum_{m=0}^{M-1} \left(1 - \frac{m}{M}\right)^p,$$

which implies that

$$M \int_0^1 (1 - x)^p dx \leq \sum_{m=0}^{M-1} \left(1 - \frac{m}{M-1}\right)^p. \quad (30)$$

Since  $\int_0^1 (1 - x)^p dx = \frac{1}{p+1}$ , (30) satisfies

$$\sum_{m=0}^{M-1} \left(1 - \frac{m}{M-1}\right)^p \geq \frac{M}{p+1}.$$

Therefore, in the case of using [Constant BS (1)],

$$\begin{aligned} \sum_{k=0}^{K-1} \alpha_k &= \alpha_{\min} K + (\alpha_{\max} - \alpha_{\min}) T_0 \sum_{m=0}^{M-1} \left(1 - \frac{m}{M}\right)^p \\ &\geq \alpha_{\min} K + (\alpha_{\max} - \alpha_{\min}) \frac{T_0 M}{p+1} \\ &= \frac{\alpha_{\max} + p\alpha_{\min}}{p+1} K, \end{aligned}$$

which implies that

$$\begin{aligned} A_K &= \frac{1}{\sum_{k=0}^{K-1} \alpha_k} \leq \frac{p+1}{(\alpha_{\max} + p\alpha_{\min})K}, \\ C_K &= \frac{\sum_{k=0}^{K-1} \gamma_k \beta_k}{\sum_{k=0}^{K-1} \alpha_k} \leq \frac{\gamma_{\max} \beta_{\max} (p+1) T_0 E}{(\alpha_{\max} + p\alpha_{\min})(1 - \lambda \zeta) K}. \end{aligned}$$

On the other hand, since  $f(x) = (1 - x)^p$  and  $g(x) = (1 - x)^{2p}$  are monotone decreasing for  $x \in [0, 1]$ , we have

$$\frac{1}{M} \sum_{m=0}^{M-1} \left(1 - \frac{m}{M}\right)^p \leq \frac{1}{M} + \int_0^1 (1 - x)^p dx, \quad \frac{1}{M} \sum_{m=0}^{M-1} \left(1 - \frac{m}{M}\right)^{2p} \leq \frac{1}{M} + \int_0^1 (1 - x)^{2p} dx,$$

which implies that

$$\sum_{m=0}^{M-1} \left(1 - \frac{m}{M}\right)^p \leq 1 + M \int_0^1 (1-x)^p dx, \quad \sum_{m=0}^{M-1} \left(1 - \frac{m}{M}\right)^{2p} \leq 1 + M \int_0^1 (1-x)^{2p} dx. \quad (31)$$

Since  $\int_0^1 (1-x)^p dx = \frac{1}{p+1}$  and  $\int_0^1 (1-x)^{2p} dx = \frac{1}{2p+1}$ , (31) satisfies

$$\sum_{m=0}^{M-1} \left(1 - \frac{m}{M}\right)^p \leq 1 + \frac{M}{p+1}, \quad \sum_{m=0}^{M-1} \left(1 - \frac{m}{M}\right)^{2p} \leq 1 + \frac{M}{2p+1}.$$

Hence,

$$\begin{aligned} \sum_{k=0}^{K-1} \alpha_k^2 &= \alpha_{\min}^2 K + 2\alpha_{\min}(\alpha_{\max} - \alpha_{\min})T_0 \sum_{m=0}^{M-1} \left(1 - \frac{m}{M}\right)^p + (\alpha_{\max} - \alpha_{\min})^2 T_0 \sum_{m=0}^{M-1} \left(1 - \frac{m}{M}\right)^{2p} \\ &\leq \alpha_{\min}^2 K + 2\alpha_{\min}(\alpha_{\max} - \alpha_{\min})T_0 \left(1 + \frac{M}{p+1}\right) + (\alpha_{\max} - \alpha_{\min})^2 T_0 \left(1 + \frac{M}{2p+1}\right) \\ &= \frac{(p+1)\alpha_{\max}^2 + 2p\alpha_{\max}\alpha_{\min} + 2p^2\alpha_{\min}^2}{(p+1)(2p+1)} K + (\alpha_{\max}^2 - \alpha_{\min}^2)T_0, \end{aligned}$$

which implies that

$$B_K = \frac{\sum_{k=0}^{K-1} \alpha_k^2}{b \sum_{k=0}^{K-1} \alpha_k} \leq \frac{(p+1)\alpha_{\max}^2 + 2p\alpha_{\max}\alpha_{\min} + 2p^2\alpha_{\min}^2}{b(2p+1)(\alpha_{\max} + p\alpha_{\min})} + \frac{(p+1)(\alpha_{\max}^2 - \alpha_{\min}^2)T_0}{(\alpha_{\max} + p\alpha_{\min})K}.$$

In the case of using [Exponential BS (2)], the following inequality holds for all  $m \in [0 : M-1]$  and  $k \in [0 : K-1]$ :

$$\frac{\lfloor \frac{k}{T_m} \rfloor}{M} \leq \frac{k}{K}. \quad (32)$$

Here, let  $S_m = \sum_{m'=0}^m T_{m'}$ . Since  $\lfloor \frac{k}{T_m} \rfloor = m$  for all  $k \in [S_m : S_{m+1} - 1]$ , showing that the following inequality holds for all  $m \in [0 : M-1]$  is sufficient to establish (32):

$$\frac{m}{M} \leq \frac{S_m}{K}.$$

From the definition of  $S_m$  and  $K$ ,

$$\begin{aligned} M \sum_{m'=0}^m T_{m'} - m \sum_{m'=0}^{M-1} T_{m'} &= (M-m) \sum_{m'=0}^m T_{m'} - m \sum_{m'=m+1}^{M-1} T_{m'} \\ &\geq ((M-m)(m+1) - m(M-1-(m+1)+1))T_m \\ &= MT_m \geq 0, \end{aligned}$$

where the first inequality holds because  $T_m$  is monotone decreasing. Since  $f(x)$  is monotone decreasing, from (32), we obtain

$$\alpha_k = \alpha_{\min} + (\alpha_{\max} - \alpha_{\min}) \left(1 - \frac{\lfloor \frac{k}{T_m} \rfloor}{M}\right)^p$$

$$\geq \alpha_{\min} + (\alpha_{\max} - \alpha_{\min}) \left(1 - \frac{k}{K}\right)^p,$$

which implies that

$$\begin{aligned} \sum_{k=0}^{K-1} \alpha_k &\geq \alpha_{\min} K + (\alpha_{\max} - \alpha_{\min}) \sum_{k=0}^{K-1} \left(1 - \frac{k}{K}\right)^p \\ &= \alpha_{\min} K + \frac{\alpha_{\max} - \alpha_{\min}}{p+1} K \\ &= \frac{\alpha_{\max} - p\alpha_{\min}}{p+1} K. \end{aligned}$$

As a result, we find that

$$\begin{aligned} A_K &= \frac{1}{\sum_{k=0}^{K-1} \alpha_k} \leq \frac{p+1}{(\alpha_{\max} + p\alpha_{\min})K}, \\ B_K &= \frac{\sum_{k=0}^{K-1} \alpha_k^2/b_k}{\sum_{k=0}^{K-1} \alpha_k} \leq \frac{\alpha_{\max}^2 \sum_{k=0}^{K-1} 1/b_k}{\sum_{k=0}^{K-1} \alpha_k} \leq \frac{\alpha_{\max}^2 (p+1)T_0 E \delta}{b_0 (\alpha_{\max} + p\alpha_{\min}) (\delta - 1) K}. \end{aligned}$$

■

#### A.6. Proof of Theorem 4

**Proof** From (24), we have that, for all  $k \in \mathbb{N}_0$ ,

$$\begin{aligned} f(\mathbf{x}_{k+1}) &\leq f(\mathbf{x}_k) - \alpha_k \langle \nabla f(\mathbf{x}_k), \mathbf{m}_k \rangle + \frac{\alpha_k^2 L}{2} \|\mathbf{m}_k\|^2 \\ &= f(\mathbf{x}_k) - \alpha_k \|\nabla f(\mathbf{x}_k)\|^2 - \alpha_k \langle \nabla f(\mathbf{x}_k), \mathbf{m}_k - \nabla f(\mathbf{x}_k) \rangle + \frac{\alpha_k^2 L}{2} \|\mathbf{m}_k\|^2 \end{aligned} \tag{33}$$

Here, from (25), we find that

$$\begin{aligned} \mathbf{m}_k - \nabla f(\mathbf{x}_k) &= (1 - \gamma_k \beta_k) \nabla f_{B_k}(\mathbf{x}_k) + \gamma_k \beta_k \mathbf{d}_{k-1} - \nabla f(\mathbf{x}_k) \\ &= (1 - \gamma_k \beta_k) (\nabla f_{B_k}(\mathbf{x}_k) - \nabla f(\mathbf{x}_k)) + \gamma_k \beta_k (\mathbf{d}_{k-1} - \nabla f(\mathbf{x}_k)). \end{aligned} \tag{34}$$

Substituting (34) for (33) yields

$$\begin{aligned} f(\mathbf{x}_{k+1}) &\leq f(\mathbf{x}_k) - \alpha_k \|\nabla f(\mathbf{x}_k)\|^2 - \alpha_k \gamma_k \beta_k \langle \nabla f(\mathbf{x}_k), \mathbf{d}_{k-1} - \nabla f(\mathbf{x}_k) \rangle \\ &\quad - \alpha_k (1 - \gamma_k \beta_k) \langle \nabla f(\mathbf{x}_k), \nabla f_{B_k}(\mathbf{x}_k) - \nabla f(\mathbf{x}_k) \rangle + \frac{\alpha_k^2 L}{2} \|\mathbf{m}_k\|^2 \\ &\leq f(\mathbf{x}_k) - \alpha_k \|\nabla f(\mathbf{x}_k)\|^2 + \alpha_k \gamma_k \beta_k \left( \frac{1}{2} \|\nabla f(\mathbf{x}_k)\|^2 + \frac{1}{2} \|\mathbf{d}_{k-1} - \nabla f(\mathbf{x}_k)\|^2 - \frac{1}{2} \|\mathbf{d}_{k-1}\|^2 \right) \\ &\quad - \alpha_k (1 - \gamma_k \beta_k) \langle \nabla f(\mathbf{x}_k), \nabla f_{B_k}(\mathbf{x}_k) - \nabla f(\mathbf{x}_k) \rangle + \frac{\alpha_k^2 L}{2} \|\mathbf{m}_k\|^2 \\ &\leq f(\mathbf{x}_k) - \frac{1}{2} \alpha_k \|\nabla f(\mathbf{x}_k)\|^2 + \frac{1}{2} \alpha_k \gamma_k \beta_k \|\mathbf{d}_{k-1} - \nabla f(\mathbf{x}_k)\|^2 \\ &\quad - \alpha_k (1 - \gamma_k \beta_k) \langle \nabla f(\mathbf{x}_k), \nabla f_{B_k}(\mathbf{x}_k) - \nabla f(\mathbf{x}_k) \rangle + \frac{\alpha_k^2 L}{2} \|\mathbf{m}_k\|^2. \end{aligned} \tag{35}$$

where, in the second inequality, we used  $2\langle \mathbf{x}, \mathbf{y} \rangle = \|\mathbf{x} - \mathbf{y}\|^2 - \|\mathbf{x}\|^2 - \|\mathbf{y}\|^2$  for all  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ , and in the last inequality, we used  $0 \leq \gamma_k \beta_k \leq 1$ . Taking the expectation conditioned on  $\mathbf{x}_k$  on both sides of (35) together with (21), guarantees that, for all  $k \in \mathbb{N}_0$ ,

$$\begin{aligned} \mathbb{E}[f(\mathbf{x}_{k+1})|\mathbf{x}_k] &\leq f(\mathbf{x}_k) - \frac{1}{2}\alpha_k \|\nabla f(\mathbf{x}_k)\|^2 + \frac{1}{2}\alpha_k \gamma_k \beta_k \|\mathbf{d}_{k-1} - \nabla f(\mathbf{x}_k)\|^2 \\ &\quad - \alpha_k(1 - \gamma_k \beta_k) \mathbb{E}[\langle \nabla f(\mathbf{x}_k), \nabla f_{B_k}(\mathbf{x}_k) - \nabla f(\mathbf{x}_k) \rangle | \mathbf{x}_k] + \frac{\alpha_k^2 L}{2} \mathbb{E}[\|\mathbf{m}_k\|^2 | \mathbf{x}_k] \\ &= f(\mathbf{x}_k) - \frac{1}{2}\alpha_k \|\nabla f(\mathbf{x}_k)\|^2 + \frac{1}{2}\alpha_k \gamma_k \beta_k \|\mathbf{d}_{k-1} - \nabla f(\mathbf{x}_k)\|^2 + \frac{\alpha_k^2 L}{2} \mathbb{E}[\|\mathbf{m}_k\|^2 | \mathbf{x}_k]. \end{aligned} \quad (36)$$

Adding both sides of (36) and Lemma 8, we obtain, from  $L \leq L^2$ , that

$$\begin{aligned} \mathbb{E}\left[f(\mathbf{x}_{k+1}) + \|\mathbf{d}_k - \nabla f(\mathbf{x}_{k+1})\|^2 | \mathbf{x}_k\right] &\leq f(\mathbf{x}_k) + \beta_k \left(\frac{\alpha_k \gamma_k}{2} + 1\right) \|\mathbf{d}_{k-1} - \nabla f(\mathbf{x}_k)\|^2 - \frac{1}{2}\alpha_k \|\nabla f(\mathbf{x}_k)\|^2 \\ &\quad + 2(1 - \beta_k) \frac{\sigma^2}{b_k} + \alpha_k^2 L^2 \left(\frac{5}{2} + \frac{1}{1 - \beta_k}\right) \mathbb{E}_k[\|\mathbf{m}_k\|^2 | \mathbf{x}_k]. \end{aligned} \quad (37)$$

Additionally, from Lemma (7) and  $\beta_k(\frac{\alpha_k}{2} + 1) \leq 1$  for all  $k \in \mathbb{N}_0$ , we obtain

$$\begin{aligned} \mathbb{E}_k\left[f(\mathbf{x}_{k+1}) + \|\mathbf{d}_k - \nabla f(\mathbf{x}_{k+1})\|^2\right] &\leq f(\mathbf{x}_k) + \|\mathbf{d}_{k-1} - \nabla f(\mathbf{x}_k)\|^2 - \frac{1}{2}\alpha_k \|\nabla f(\mathbf{x}_k)\|^2 \\ &\quad + 2(1 - \beta_k) \frac{\sigma^2}{b_k} + \alpha_k^2 L^2 \left(\frac{5}{2} + \frac{1}{1 - \beta_k}\right) G^2. \end{aligned} \quad (38)$$

Therefore, taking the total expectation on both sides of the above inequality ensures that, for all  $k \in \mathbb{N}_0$ ,

$$\begin{aligned} \frac{1}{2}\alpha_k \mathbb{E}[\|\nabla f(\mathbf{x}_k)\|^2] &\leq \mathbb{E}\left[f(\mathbf{x}_k) + \|\mathbf{d}_{k-1} - \nabla f(\mathbf{x}_k)\|^2 - f(\mathbf{x}_{k+1}) + \|\mathbf{d}_k - \nabla f(\mathbf{x}_{k+1})\|^2\right] \\ &\quad + 2(1 - \beta_k) \frac{\sigma^2}{b_k} + \alpha_k^2 L^2 \left(\frac{5}{2} + \frac{1}{1 - \beta_k}\right) G^2 \end{aligned}$$

Moreover, summing from  $k = 0$  to  $k = K - 1$  ensures that

$$\begin{aligned} \frac{1}{2} \sum_{k=0}^{K-1} \alpha_k \mathbb{E}[\|\nabla f(\mathbf{x}_k)\|^2] &\leq \mathbb{E}\left[f(\mathbf{x}_0) + \|\mathbf{d}_{-1} - \nabla f(\mathbf{x}_0)\|^2 - f(\mathbf{x}_K) + \|\mathbf{d}_{K-1} - \nabla f(\mathbf{x}_K)\|^2\right] \\ &\quad + 2\sigma^2 \sum_{k=0}^{K-1} \frac{1 - \beta_k}{b_k} + \frac{5}{2} L^2 G^2 \sum_{k=0}^{K-1} \alpha_k^2 + L^2 G^2 \sum_{k=0}^{K-1} \frac{\alpha_k^2}{1 - \beta_k}. \end{aligned}$$

By setting  $\mathbf{d}_{-1} = \|\nabla f(\mathbf{x}_0)\|^2$ , from the condition (C1)(the lower bound  $f_\star$  of  $f$ ), we find that

$$\frac{1}{2} \sum_{k=0}^{K-1} \alpha_k \mathbb{E}[\|\nabla f(\mathbf{x}_k)\|^2] \leq f(\mathbf{x}_0) - f_\star + 2\sigma^2 \sum_{k=0}^{K-1} \frac{1 - \beta_k}{b_k}$$

$$+ \frac{5}{2} L^2 G^2 \sum_{k=0}^{K-1} \alpha_k^2 + L^2 G^2 \sum_{k=0}^{K-1} \frac{\alpha_k^2}{1 - \beta_k}.$$

Therefore,

$$\begin{aligned} \min_{k \in [0:K-1]} \mathbb{E} \left[ \|\nabla f(\mathbf{x}_k)\|^2 \right] &\leq 2(f(\mathbf{x}_0) - f_\star) \underbrace{\frac{1}{\sum_{k=0}^{K-1} \alpha_k}}_{A_K} + 4\sigma^2 \underbrace{\sum_{k=0}^{K-1} \frac{1 - \beta_k}{b_k} \frac{1}{\sum_{k=0}^{K-1} \alpha_k}}_{B'_K} \\ &\quad + 5L^2 G^2 \underbrace{\frac{\sum_{k=0}^{K-1} \alpha_k^2}{\sum_{k=0}^{K-1} \alpha_k}}_{C'_K} + 2L^2 G^2 \underbrace{\sum_{k=0}^{K-1} \frac{\alpha_k^2}{1 - \beta_k} \frac{1}{\sum_{k=0}^{K-1} \alpha_k}}_{D'_K}. \end{aligned}$$

This proves Theorem 4(i). Second, we assume the condition of Theorem 4(ii). From  $\sum_{k=0}^{+\infty} \alpha_k = +\infty$ ,  $\sum_{k=0}^{\infty} \frac{1 - \beta_k}{b_k} < +\infty$ , and  $\sum_{k=0}^{+\infty} \frac{\alpha_k^2}{1 - \beta_k} < +\infty$ , we find that

$$\lim_{K \rightarrow \infty} A_K = \lim_{K \rightarrow \infty} B'_K = \lim_{K \rightarrow \infty} C'_K = \lim_{K \rightarrow \infty} D'_K = 0.$$

This proves Theorem 4(ii). ■

### A.7. Proof of Corollary 5

**Proof** [Sqrt-Decaying LR]  $\wedge$  [Constant Beta]: The upper bounds of  $A'_K$  and  $C'_K$  are provided in Corollary 3. So, here, we will only provide upper bounds of  $B'_K$  and  $D'_K$ . In particular, we have that

$$\sum_{k=0}^{K-1} (1 - \beta_k) = K(1 - \beta_{\max}).$$

Therefore, in the case of using [Constant BS (1)],

$$B'_K = \frac{\sum_{k=0}^{K-1} (1 - \beta_k)}{b \sum_{k=0}^{K-1} \alpha_k} = \frac{K(1 - \beta_{\max})}{b(\sqrt{K+1} - 1)}.$$

On the other hand, in the case of using [Exponential BS (2)], we have that

$$B'_K = \sum_{k=0}^{K-1} \frac{1 - \beta_k}{b_k} \frac{1}{\sum_{k=0}^{K-1} \alpha_k} \leq \frac{(1 - \beta_{\max}) T_0 E \delta}{2b_0 \alpha_{\max} (\delta - 1) (\sqrt{K+1} - 1)}.$$

We also have

$$D'_K = \sum_{k=0}^{K-1} \frac{\alpha_k^2}{1 - \beta_k} \frac{1}{\sum_{k=0}^{K-1} \alpha_k} = \frac{C'_K}{1 - \beta_{\max}} \leq \frac{\alpha_{\max} T (1 + \log(K+1))}{2(1 - \beta_{\max}) (\sqrt{K+1} - 1)}.$$

[Decaying LR]  $\wedge$  [Increasing Beta]: We have that

$$\alpha_k = \frac{\alpha_{\max}}{\lfloor \frac{k}{T} \rfloor + 1} \geq \frac{\alpha_{\max}}{k+1}.$$

Therefore,

$$\sum_{k=0}^{K-1} \alpha_k \geq \alpha_{\max} \int_0^K \frac{dk}{k+1} = \alpha_{\max} \log(K+1).$$

Hence,

$$A'_K = \frac{1}{\sum_{k=0}^{K-1} \alpha_k} \leq \frac{1}{\alpha_{\max} \log(K+1)}.$$

We also have

$$\sum_{m=0}^{M-1} \alpha'_m \leq 1 + \int_0^M \frac{dm}{(m+1)^2} = 1 + \left( -\frac{1}{M+1} + 1 \right) \leq 2,$$

which implies that

$$\sum_{k=0}^{K-1} \alpha_k \leq T_0 \sum_{m=0}^{M-1} \alpha'_m \leq 2\alpha_{\max}^2 T_0.$$

Therefore,

$$C'_K = \frac{\sum_{k=0}^{K-1} \alpha_k^2}{\sum_{k=0}^{K-1} \alpha_k} \leq \frac{2\alpha_{\max} T_0}{\log(K+1)}.$$

On the other hand, we obtain

$$\begin{aligned} \sum_{k=0}^{K-1} (1 - \beta_k) &= T_0 \sum_{m=0}^{M-1} \frac{1 - \beta_{\min}}{(m+1)^{3/4}} \leq T_0 \left( 1 + \int_0^M \frac{1 - \beta_{\min}}{(m+1)^{3/4}} dm \right) = T_0 \left( 1 + \frac{4}{3} (1 - \beta_{\min}) \left( (M+1)^{1/4} - 1 \right) \right) \\ &\leq T_0 \left( 1 + 4(1 - \beta_{\min})(K+1)^{1/4} \right). \end{aligned}$$

Therefore, in the case of using [Constant BS (1)],

$$B'_K = \frac{\sum_{k=0}^{K-1} (1 - \beta_k)}{b \sum_{k=0}^{K-1} \alpha_k} \leq \frac{T_0 (1 + 4(1 - \beta_{\min})(K+1)^{1/4})}{b \alpha_{\max} \log(K+1)}.$$

On the other hand, in the case of using [Exponential BS (2)], we have that

$$B'_K = \sum_{k=0}^{K-1} \frac{1 - \beta_k}{b_k} \frac{1}{\sum_{k=0}^{K-1} \alpha_k} \leq \frac{(1 - \beta_{\min}) T_0 E \delta}{b_0 \alpha_{\max} (\delta - 1) \log(K+1)}.$$

We also have

$$\begin{aligned} \sum_{k=0}^{K-1} \frac{\alpha_k^2}{1-\beta_k} &\leq \frac{\alpha_{\max}^2 T_0}{1-\beta_{\min}} \sum_{m=0}^{M-1} \frac{(m+1)^{3/4}}{(m+1)^2} = \frac{\alpha_{\max}^2 T_0}{1-\beta_{\min}} \sum_{m=0}^{M-1} \frac{1}{(m+1)^{5/4}} \leq \frac{\alpha_{\max}^2 T_0}{1-\beta_{\min}} \left(1 + \int_0^M \frac{dm}{(m+1)^{5/4}}\right) \\ &= \frac{\alpha_{\max}^2 T_0}{1-\beta_{\min}} \left(1 - \frac{1}{(M+1)^{1/4}} + 1\right) \leq \frac{2\alpha_{\max}^2 T_0}{1-\beta_{\min}}. \end{aligned}$$

Hence,

$$D'_K = \sum_{k=0}^{K-1} \frac{\alpha_k^2}{1-\beta_k} \frac{1}{\sum_{k=0}^{K-1} \alpha_k} \leq \frac{2\alpha_{\max} T_0}{(1-\beta_{\min}) \log(K+1)}.$$

■

### A.8. Discussion on the boundedness of the gradient

In this section, we present the condition for the boundedness of the gradient (Assumption 1) and demonstrate that it is satisfied in our experiments.

**Proposition 9** *Let the loss function  $f_i$  be  $L_i$ -smooth, and let it satisfy that there exist  $f_i^*$  and  $f_{i\star}$  such that, for any  $\mathbf{x} \in \mathbb{R}^d$ ,  $f_{i\star} \leq f_i(\mathbf{x}) \leq f_i^*$ . Then, for any  $i \in [n]$  and any  $\mathbf{x} \in \mathbb{R}^d$ ,*

$$\|\nabla f_i(\mathbf{x})\| \leq \sqrt{2L_i(f_i^* - f_{i\star})} =: G_i \leq \max_{i \in [n]} G_i =: G.$$

**Proof** Let  $\bar{\mathbf{x}} := \mathbf{x} - \frac{1}{L_i} \|\nabla f_i(\mathbf{x})\|$ . From the descent lemma, we have

$$\begin{aligned} f(\bar{\mathbf{x}}) &\leq f(\mathbf{x}) + \langle \nabla f_i(\mathbf{x}), \bar{\mathbf{x}} - \mathbf{x} \rangle + \frac{L_i}{2} \|\bar{\mathbf{x}} - \mathbf{x}\|^2 \\ &= f(\mathbf{x}) - \frac{1}{L_i} \|\nabla f_i(\mathbf{x})\|^2 + \frac{1}{2L_i} \|\nabla f_i(\mathbf{x})\|^2 \\ &= f(\mathbf{x}) - \frac{1}{2L_i} \|\nabla f_i(\mathbf{x})\|^2. \end{aligned}$$

Therefore, we also have

$$\frac{1}{2L_i} \|\nabla f(\mathbf{x})\|^2 \leq f(\mathbf{x}) - f(\bar{\mathbf{x}}) \leq f_i^* - f_{i\star}.$$

The proof follows from a straightforward algebraic manipulation of the above. ■

In the context of machine learning, we have  $f_{i\star} = 0$  and that the maximum training loss typically occurs around the initial point and it rarely grows without bound after that. For example, when training on CIFAR-100 with the cross-entropy loss as the objective function, suppose that, at the initial point, each class is equally likely to be output. In this case,

$$f_i(\mathbf{x}_0) = -\log p_i = -\log \frac{1}{100} \approx 4.40517 \dots$$



Therefore, the assumption that the loss has both upper and lower bounds appears to be a natural one. Figure 5 presents the empirical loss values at each epoch using various learning rate schedulers, obtained under the setup described in this paper. Figure 5 shows  $f(\mathbf{x}_0) = f^* \approx 4.07 \dots$  and  $f(\mathbf{x}_k)$  decreasing for epochs. Accordingly, the above discussion, together with (C1)( $f$  is  $L$ -smooth), leads to the finding that, for all  $k$ ,

$$\|\nabla f(\mathbf{x}_k)\| \leq \sqrt{2L(f^* - f_*)} = \sqrt{2Lf(\mathbf{x}_0)}.$$

Since  $f$  is defined by the sum of loss functions  $f_i$ , there exists  $G_i \geq 0$  such that, for all  $k$ ,  $\|\nabla f_i(\mathbf{x}_k)\| \leq G_i$ . That is, Assumption 1 holds.

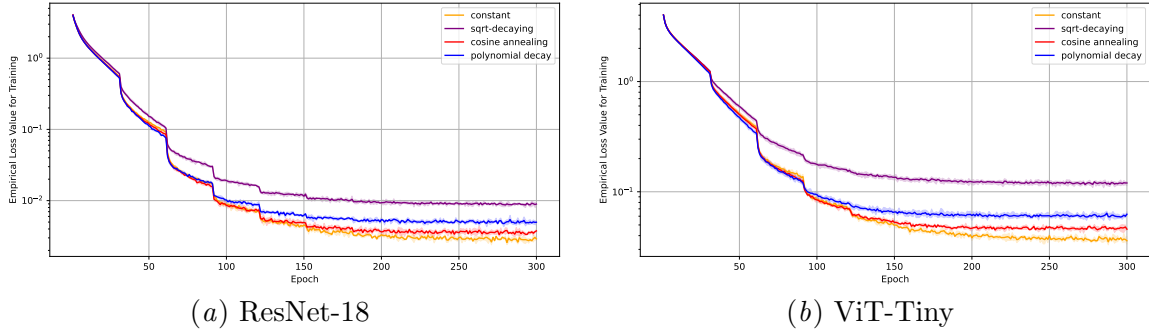


Figure 5: Empirical loss using mini-batch QHM with Step Decay Beta and Gamma to train (a) ResNet-18 and (b) ViT-Tiny on the CIFAR-100 dataset.

### A.9. Performance comparison across different batch sizes using stochastic first-order oracle (SFO)

In this section, for a fair comparison of different batch size schedulers, we use a stochastic first-order oracle (SFO). SFO represents the total number of data samples fed into the model during  $K$  training steps, and it is defined as follows:

$$\text{SFO} := \sum_{m=0}^{M-1} b'_m T_m.$$

Since  $\sum_{m=0}^{M-1} b'_m T_m = \sum_{m=0}^{M-1} b'_m \left\lceil \frac{n}{b'_m} \right\rceil \approx nM$ , SFO plays a role similar to that of the number of epochs.

Figures 6 and 7 illustrate the relationship between the performance of mini-batch QHM (test accuracy and the minimum of the full gradient norm) and SFO. Note that, as discussed above, the overall shapes of these graphs do not differ from those of Figures 1 and 2. Figures 6 and 7 plot the minimum SFO required to achieve at least 70% test accuracy and less than 0.5 in the full gradient norm, respectively. From this, we can see that, in most cases, the sharp performance improvement brought by increasing the batch size successfully reduces the SFO needed to reach the thresholds. However, it should be emphasized that depending on the choice of threshold values, one could make exactly the opposite claim.

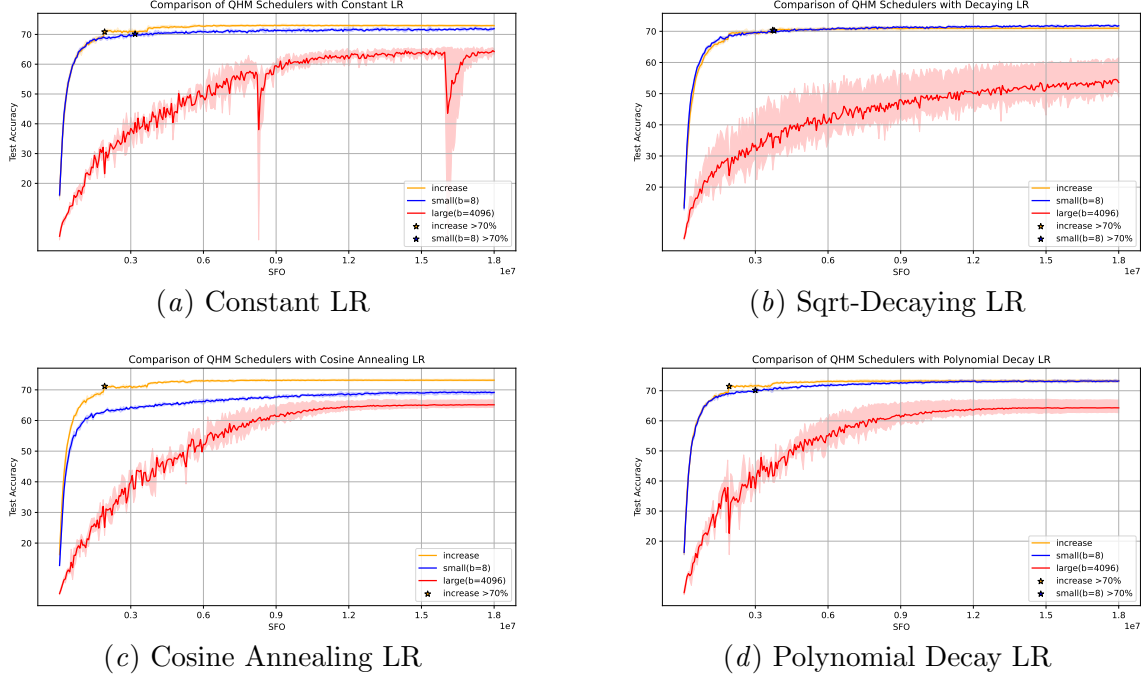


Figure 6: Test accuracy score versus SFO for comparison of increasing and constant batch sizes in using mini-batch QHM with various lr-schedulers and step-decay momentum weights to train ResNet-18 on the CIFAR-100 dataset.

#### A.10. Impact of Hyperparameters on Performance Robustness

In this section, we discuss how robust the performance is to changes in the hyperparameters, learning rate, momentum weights, and batch size from both theoretical and empirical perspectives. In the numerical experiments, we trained ResNet-18 on CIFAR-100 using mini-batch QHM with a cosine learning rate and step-decay momentum weights schedulers. We will denote the hyperparameters explicitly as  $\alpha_0 := \alpha_{\max}$ ,  $\beta_0 := \beta_{\max}$ , and  $\gamma_0 := \gamma_{\max}$ .

From a theoretical perspective, Theorems 2 and 4 impose the condition of  $\alpha_0 \leq \frac{2}{L}$ .  $L$  is determined by the model and dataset, and in these analyses, which guarantee convergence for general non-convex smooth objective functions. Hence, the initial value of the learning rate  $\alpha_0$  lacks robustness. On the other hand, for the batch size and momentum weight, the only required condition is whether the corresponding infinite series diverges or converges, and thus no assumptions are made regarding the initial values ( $b_0$ ,  $\gamma_0$ , and  $\beta_0$ ). Therefore, we believe that robustness is ensured in this aspect.

From an empirical perspective, Figures 8 and 9 present heatmaps that show the best performance (test accuracy and minimum of the full gradient norm of empirical loss) achieved during training, plotted in the  $\beta_0$ - $\gamma_0$  plane using various values of  $\alpha_0$ . Here, the batch size was increased every 30 epochs from  $2^5$  to  $2^9$ , in a training lasting a total of 150 epochs. Figures 8 and 9 indicate that changing the value of  $\alpha_0$  results in up to about a 5% difference in test accuracy and about a 0.04 difference in the minimum full gradient norm. However, even when the values of  $\beta_0$  or  $\gamma_0$  are changed, there is no significant change in performance,

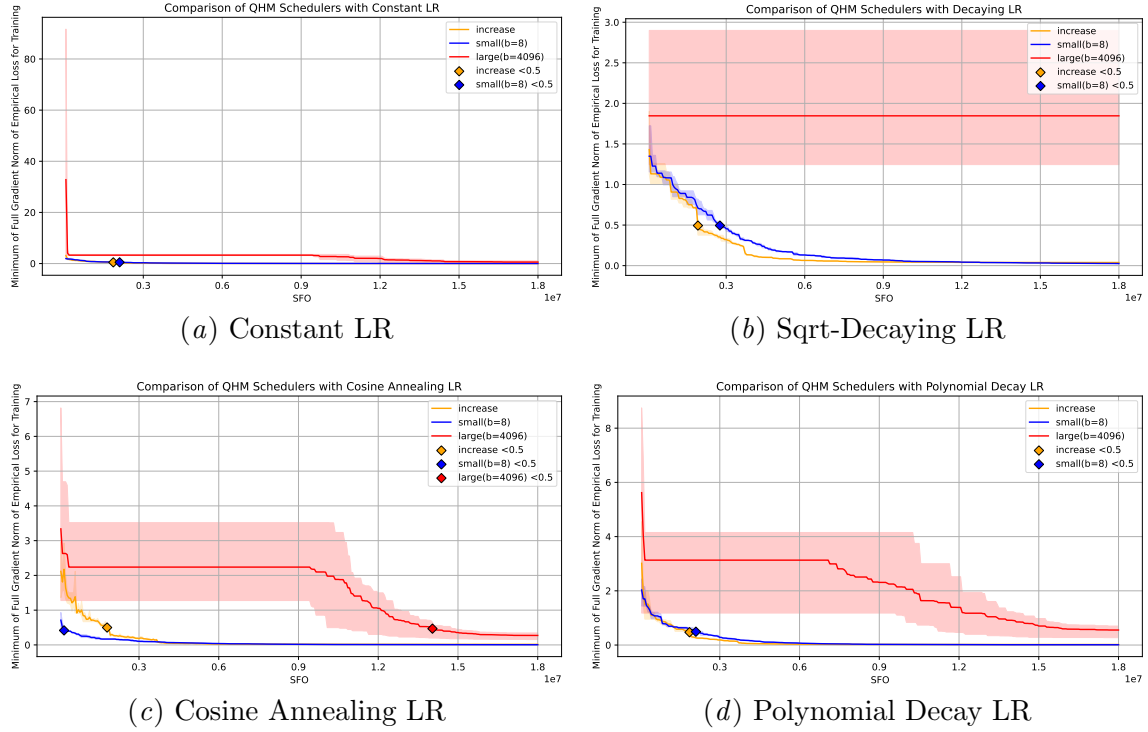


Figure 7: Minimum of full gradient norm of empirical loss versus SFO for comparison of increasing and constant batch sizes in using mini-batch QHM with various lr-schedulers and step-decay momentum weights to train ResNet-18 on the CIFAR-100 dataset.

indicating robustness. When  $\alpha_0 = 0.5$  is set to a large value, the test accuracy becomes sensitive to the values of  $\beta_0$  and  $\gamma_0$ . However, it should be noted that this is also due to the learning rate being set to a relatively large value.

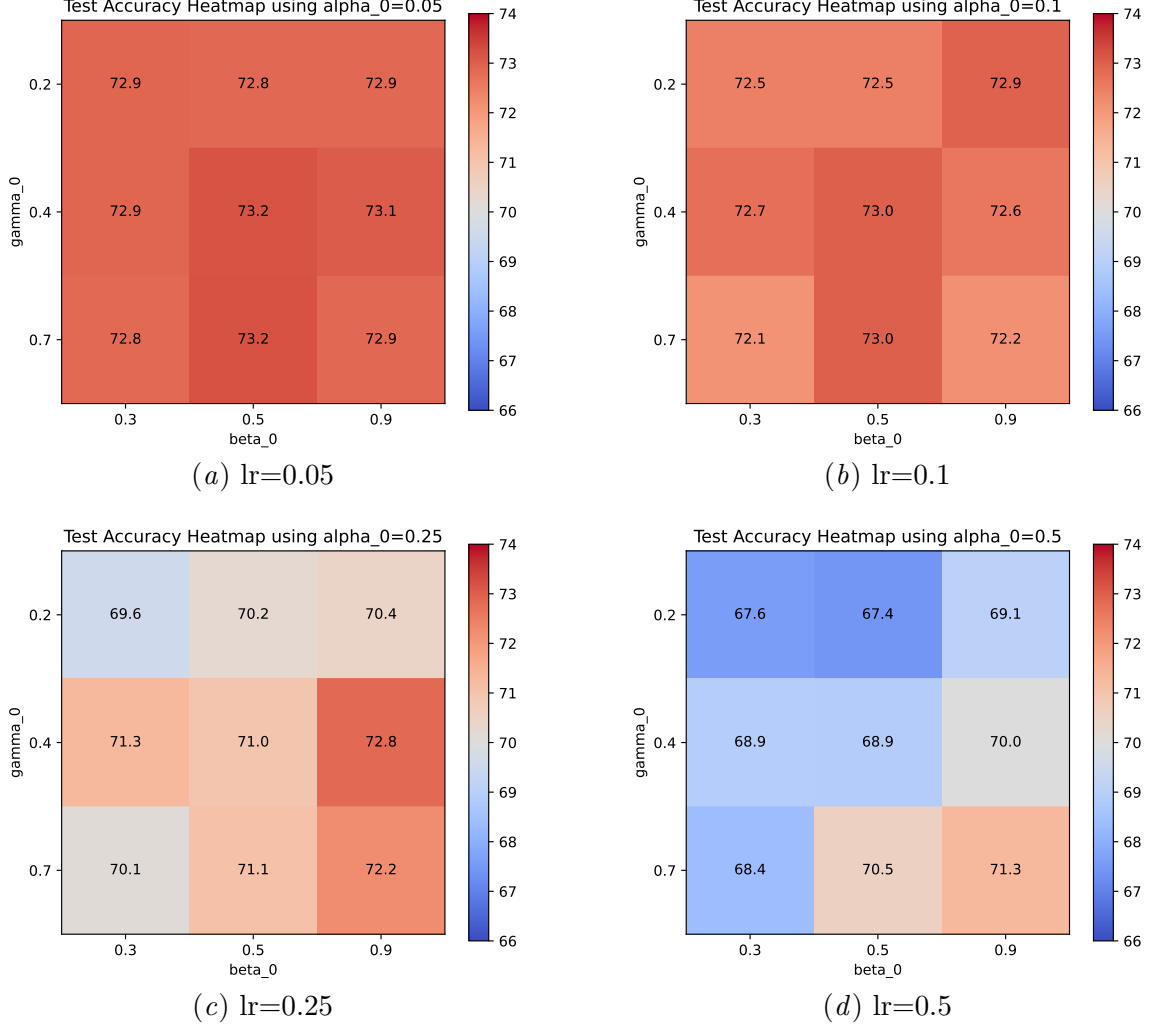


Figure 8: Test accuracy heatmaps comparing various values of  $\alpha_0$  in using mini-batch QHM with a cosine lr-scheduler and step-decay momentum weights to train ResNet-18 on the CIFAR-100 dataset.

#### A.11. Comparison with adaptive batch size methods

In this section, we compare the exponentially increasing batch size scheduler, as in (De et al., 2017), with the linearly increasing batch size scheduler. The linearly increasing scheduler increases the batch size by  $\delta > 0$  at each epoch, i.e., for all  $m \in [0 : M - 1]$ , we have

$$b'_{m+1} = b'_m + \delta.$$

# QHM USING INCREASING BATCH SIZE

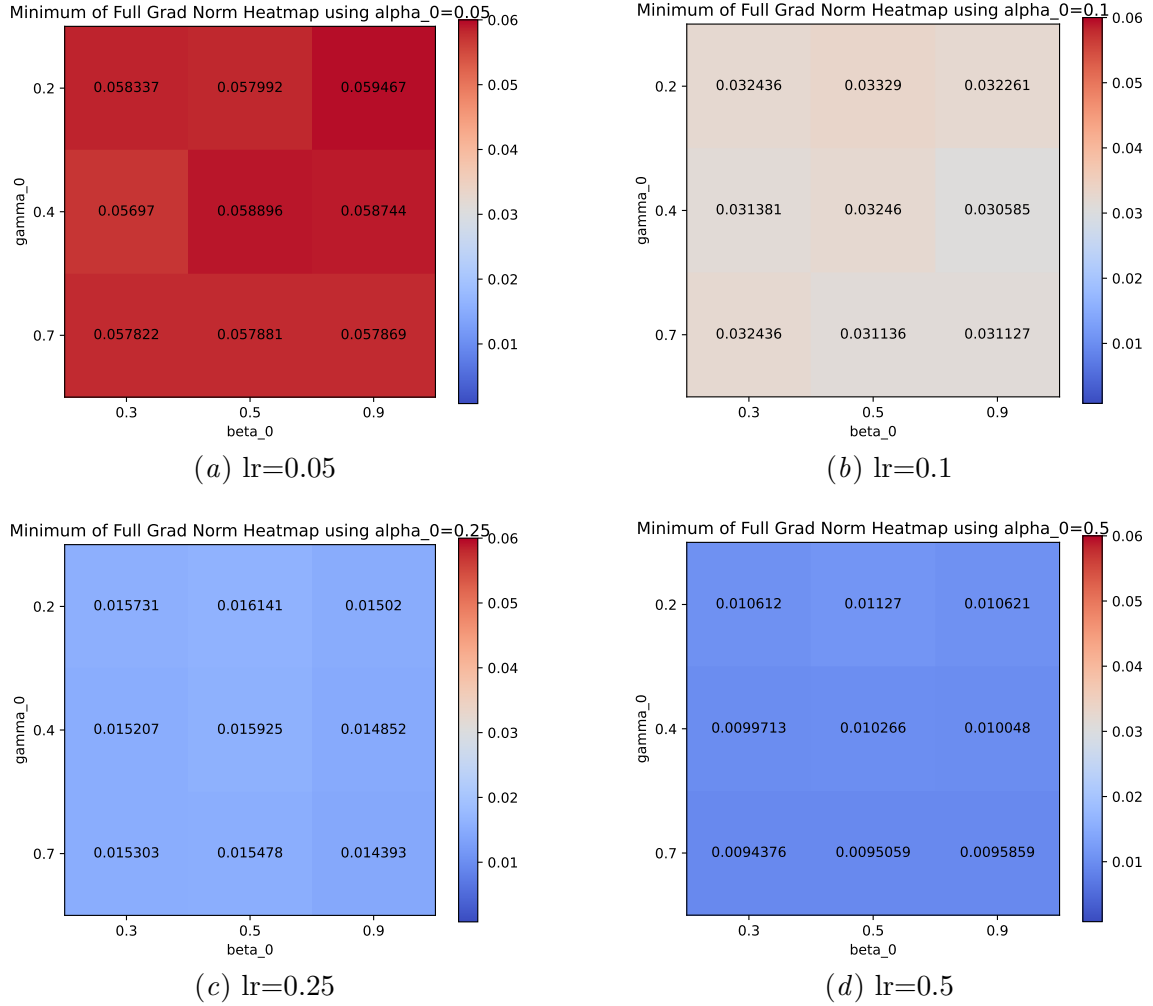


Figure 9: Minimum of full gradient norm of empirical loss heatmaps comparing various values of  $\alpha_0$  in using mini-batch QHM with a cosine lr-scheduler and step-decay momentum weights to train ResNet-18 on the CIFAR-100 dataset.

Regardless of which learning rate scheduler is used, the exponentially increasing scheduler consistently outperforms the linear one. This result accords with Theorem 2 in our paper. For mini-batch QHM using step-decay momentum weights to achieve asymptotic convergence, the following condition must be satisfied:

$$\sum_{k=0}^{K-1} \frac{\alpha_k^2}{b_k} < +\infty.$$

In the case of using the linear scheduler, by using the following inequality:

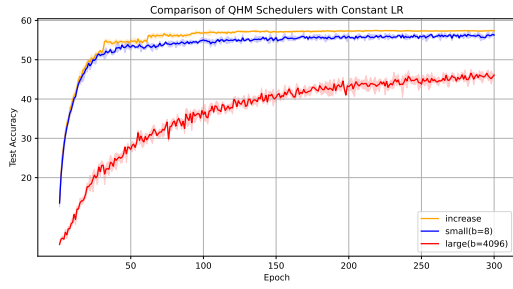
$$\begin{aligned} \sum_{k=0}^{K-1} \frac{1}{b_k} &= T_{\max} \sum_{m=0}^{M-1} \frac{1}{b_0 + m\delta} \leq T_{\max} \left( b_0 + \int_0^M \frac{dx}{b_0 + m\delta} \right) \\ &= T_{\max}(b_0 + \log(b_0 + (M-1)\delta) - \log b_0) \\ &\leq T_{\max}(b_0 + \log(b_0 + (K-1)\delta) - \log b_0), \end{aligned}$$

we obtain

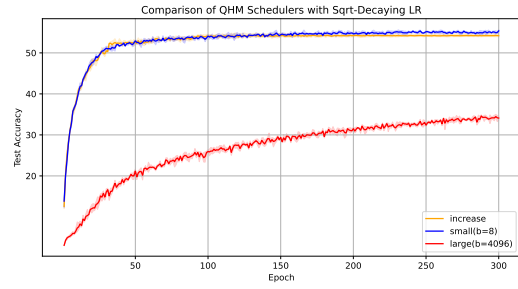
$$\sum_{k=0}^{K-1} \frac{\alpha_k^2}{b_k} \leq \alpha_{\max}^2 T_{\max} \sum_{m=0}^{M-1} \frac{1}{b_0 + m\delta} \leq \alpha_{\max}^2 T_{\max}(b_0 + \log(b_0 + (K-1)\delta) - \log b_0).$$

This implies that  $\sum_{k=0}^{K-1} \frac{\alpha_k^2}{b_k} < +\infty$  is not satisfied. In this case, only  $B_K$  in Corollary 3 is replaced by the expression proven above. As the result, the convergence rate when using a constant LR enjoys  $\mathcal{O}\left(\frac{\log K}{K}\right)$ , but is worse than  $\mathcal{O}\left(\frac{1}{K}\right)$  in the case of an exponentially increasing scheduler.

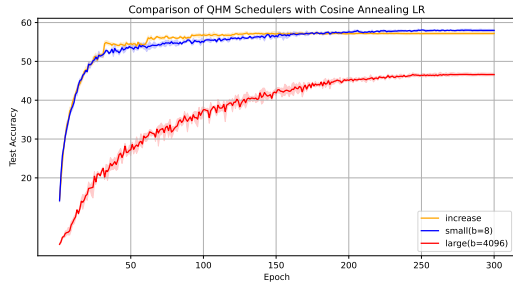
#### A.12. Training Vision Transformer-Tiny on the CIFAR-100 dataset



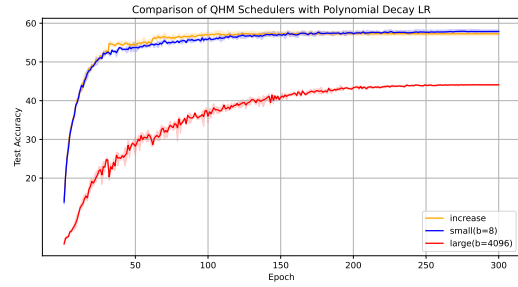
(a) Constant LR



(b) Sqrt-Decaying LR



(c) Cosine Annealing LR



(d) Polynomial Decay LR

Figure 10: Test accuracy score versus number of epochs for comparison of increasing and constant batch sizes in using mini-batch QHM with various lr-schedulers and step-decay momentum weights to train ViT-Tiny on the CIFAR-100 dataset.

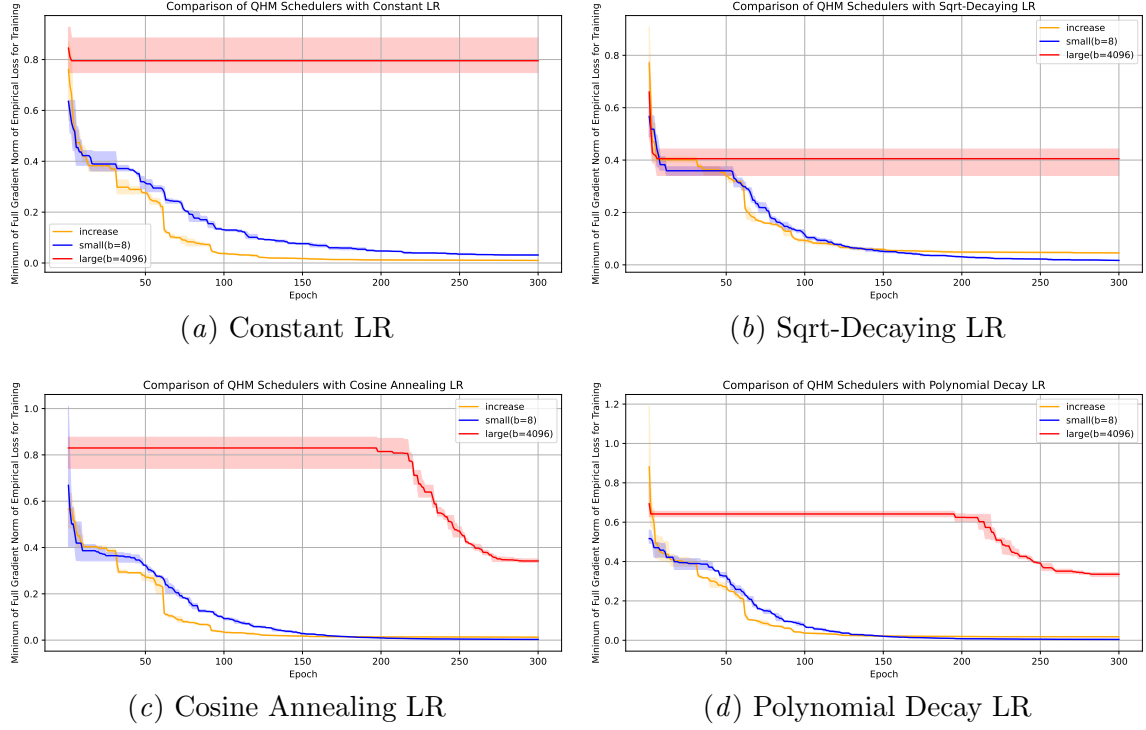


Figure 11: Minimum of full gradient norm of empirical loss versus number of epochs for comparison of increasing and constant batch sizes in using mini-batch QHM with various lr-schedulers and step-decay momentum weights to train ViT-Tiny on the CIFAR-100 dataset.

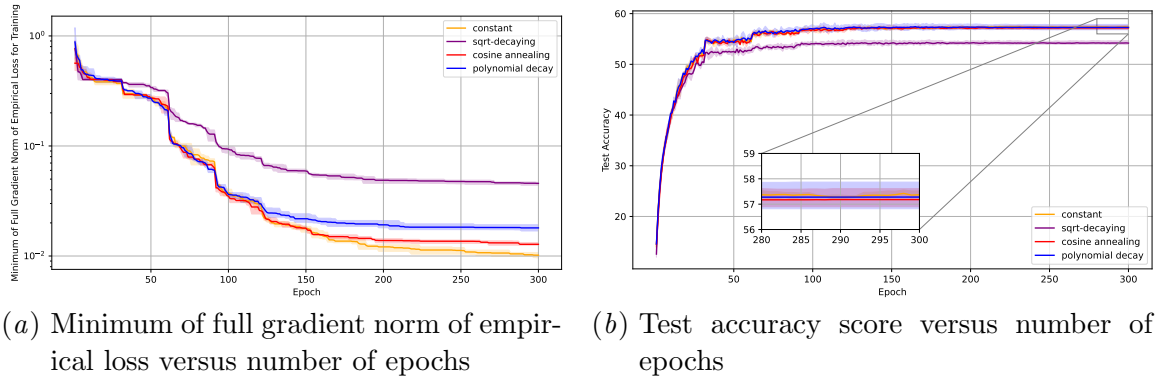


Figure 12: Comparison of lr-scheduler using mini-batch QHM with Step Decay Beta and Gamma to train ViT-Tiny on the CIFAR-100 dataset.



