

## Appendix A. Calculation of SFO Complexity

### A.1. SFO Complexity with Constant BS and Increasing BS

We compute the SFO complexity required to achieve  $\|G_T\|^2 \leq \epsilon^2$  ( $\epsilon > 0$ ), as described in Section 4.3.

**[Constant BS]**

From Theorem 5, the number of iterations  $T$  required to achieve  $\|G_T\|^2 \leq \epsilon^2$  is

$$T = \frac{\tilde{Q}_1}{\epsilon^2 - \tilde{Q}_2 \sigma^2 b^{-1}}.$$

Because the SFO complexity of constant BS can be represented by  $\text{SFO}^{\text{const}}(b) = bT$ , the  $\text{SFO}^{\text{const}}(b)$  required to achieve  $\|G_T\|^2 \leq \epsilon^2$  is given by

$$\text{SFO}_\epsilon^{\text{const}}(b) = bT = b \frac{\tilde{Q}_1}{\epsilon^2 - \tilde{Q}_2 \sigma^2 b^{-1}} = \frac{b^2 \tilde{Q}_1}{b\epsilon^2 - \tilde{Q}_2 \sigma^2}.$$

Because

$$\frac{d}{db} \text{SFO}_\epsilon^{\text{const}}(b) = \frac{2b\tilde{Q}_1(b\epsilon^2 - \tilde{Q}_2 \sigma^2) - b^2 \tilde{Q}_1 \epsilon^2}{(b\epsilon^2 - \tilde{Q}_2 \sigma^2)^2} = \frac{b\tilde{Q}_1(b\epsilon^2 - 2\tilde{Q}_2 \sigma^2)}{(b\epsilon^2 - \tilde{Q}_2 \sigma^2)^2}$$

holds, from

$$b \leq \frac{2\tilde{Q}_2 \sigma^2}{\epsilon^2} \Rightarrow \frac{d}{db} \text{SFO}_\epsilon^{\text{const}}(b) \leq 0 \quad \text{and} \quad b \geq \frac{2\tilde{Q}_2 \sigma^2}{\epsilon^2} \Rightarrow \frac{d}{db} \text{SFO}_\epsilon^{\text{const}}(b) \geq 0,$$

we have the critical BS of constant BS:

$$b^\star := \frac{2\tilde{Q}_2 \sigma^2}{\epsilon^2} = \underset{b>0}{\operatorname{argmin}} \text{SFO}_\epsilon^{\text{const}}(b),$$

which implies

$$\text{SFO}_\epsilon^{\text{const}}(b^\star) = \frac{4\tilde{Q}_1 \tilde{Q}_2^2 \sigma^4 \epsilon^{-4}}{\tilde{Q}_2 \sigma^2} = 4\tilde{Q}_1 \tilde{Q}_2 \sigma^2 \epsilon^{-4} = O(\epsilon^{-4}).$$

**[Increasing BS]**

From Theorem 6, the number of iterations  $T$  required to achieve  $\|G_T\|^2 \leq \epsilon^2$  is given by

$$T = \frac{\tilde{Q}_1 + \tilde{Q}_2 \sigma^2 b_0^{-1}}{\epsilon^2}.$$

We assume a setting in which  $T = MK$ , meaning that by the  $T$ -th iteration, the BS has been increased  $M$  times, and RSGD has been updated for  $K$  steps at each BS. If the exact value of  $T$  takes the form  $T = MK + k$  ( $1 \leq k \leq K-1$ ), we can set  $T$  to  $M(K+1)$  to ensure that  $\|G_T\|^2 \leq \epsilon^2$  still holds. Similarly, if  $T$  takes the form  $T = MK - k$  ( $1 \leq k \leq K-1$ ), we can set  $T$  to  $MK$  to satisfy the same condition. Hence, our assumption that  $T = MK$  is

reasonable.

SFO complexity for increasing BS (i.e., exponential growth BS) can be represented as

$$\text{SFO}^{\text{incr}} = \left( \sum_{m=0}^{M-1} b_0 \gamma^m \right) K = \frac{b_0}{\gamma - 1} K (\gamma^M - 1).$$

Under our theoretical assumption that  $K$  is fixed and  $M$  is dynamic, and with  $M = \frac{T}{K}$ , the SFO complexity of increasing BS for achieving  $\|G_T\|^2 \leq \epsilon^2$  is

$$\text{SFO}_\epsilon^{\text{incr}} = \frac{b_0 K}{\gamma - 1} (\gamma^{\frac{\tilde{Q}_1 + \tilde{Q}_2 \sigma^2 b_0^{-1}}{K \epsilon^2}} - 1) = O(\gamma^{\epsilon^{-2}}).$$

To calculate SFO complexity under our experimental setting, we must reconsider the proof [D.3](#) because it relies on letting  $M \rightarrow \infty$ , which is not permitted in our experimental setting. This issue can, however, be resolved by the following discussion. The key is to replace the evaluation at the penultimate inequality of [\(9\)](#) with

$$\begin{aligned} \sum_{t=0}^{T-1} \frac{1}{b_t} &= \sum_{m=0}^{M-1} \frac{K}{b_0 \gamma^m} + \frac{T - KM}{b_0 \gamma^M} \leq \sum_{m=0}^M \frac{K}{b_0 \gamma^m} = \frac{K(1 - \gamma^{-M-1})}{b_0(1 - \gamma^{-1})} \\ &= \frac{K(\gamma^M - \gamma^{-1})}{b_0 \gamma^{M-1}(\gamma - 1)} \leq \frac{K \gamma^M}{b_0 \gamma^{M-1}(\gamma - 1)} = \frac{K \gamma}{b_0(\gamma - 1)}. \end{aligned}$$

Thus,

$$\sum_{t=0}^{T-1} \frac{\eta_{\max}^2}{b_t} \leq \frac{\eta_{\max}^2 K \gamma}{b_0(\gamma - 1)}$$

implies the same result with [Theorem 6](#):

$$\|G_T\|^2 \leq \frac{\tilde{Q}_1 + \tilde{Q}_2 \sigma^2 b_0^{-1}}{T}.$$

Because  $\tilde{Q}_2$  contains  $K$ , as shown by the formal statement of [Theorem 6](#) (with a constant LR and an exponential growth BS), we rearrange  $\tilde{Q}_2$  for calculating SFO complexity under our experimental setting as

$$\tilde{Q}_2 = K \tilde{Q}_3,$$

where  $\tilde{Q}_3$  is defined through this substitution. Then, we obtain

$$MK = T = \frac{\tilde{Q}_1 + \tilde{Q}_2 \sigma^2 b_0^{-1}}{\epsilon^2} = \frac{\tilde{Q}_1 + \tilde{Q}_3 K \sigma^2 b_0^{-1}}{\epsilon^2},$$

which implies

$$K = \frac{\tilde{Q}_1}{M \epsilon^2 - \tilde{Q}_3 \sigma^2 b_0^{-1}}.$$

Therefore, under our experimental setting—where  $K$  is determined by fixed  $M$ —and with  $K = \frac{T}{M}$ , the SFO complexity of increasing BS for achieving  $\|G_T\|^2 \leq \epsilon^2$  becomes

$$\text{SFO}_\epsilon^{\text{incr}} = \frac{b_0(\gamma^M - 1)}{\gamma - 1} \frac{\tilde{Q}_1}{M\epsilon^2 - \tilde{Q}_3\sigma^2b_0^{-1}} = O(\epsilon^{-2}).$$

The calculations in this section are for the case of a constant LR, but it is immediately clear from Theorems 5 and 6 that exactly the same results hold for a cosine annealing LR, which was also used in our experiments.

## A.2. Objective Function Value versus SFO Complexity

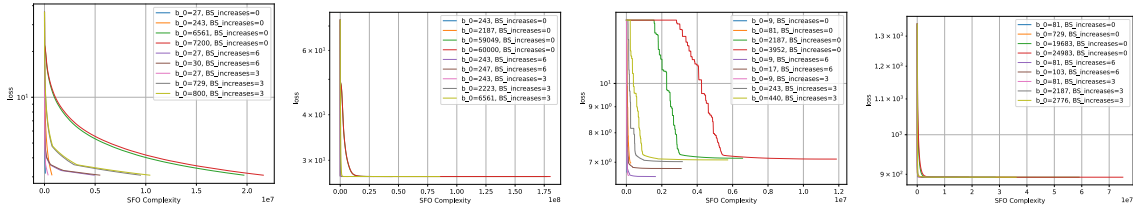


Figure 7: Objective function value (loss) versus SFO complexity on COIL100 (PCA), MNIST (PCA), MovieLens-1M (LRMC), and Jester (LRMC) datasets in order from left to right.

For the COIL100 (PCA) and MovieLens-1M (LRMC), performance was better with an increasing BS than with a constant BS. Although the differences in the objective function values for the MNIST (PCA) and Jester (LRMC) are small (one possible reason for this is that the objective function may be flat around the optimal solution), the performance with an increasing BS was equal to or better than that with a constant BS. A more detailed discussion of this hypothesis is presented in Section 4.3.

## Appendix B. Trade-offs for Each Hyperparameter of Exponential Growth BS Scheduler

We calculate the trade-offs for each hyperparameter of the exponential growth BS.

$[\gamma]$ : From Theorem 6,  $\|G_T\|^2 = O\left(1 + \frac{\gamma}{\gamma-1}\right)$  and  $\text{SFO}_\epsilon^{\text{incr}} = O(\gamma^M)$  hold, and  $\gamma^M$  is clearly monotonically increasing with respect to  $\gamma$ . Let  $f(\gamma) := 1 + \frac{\gamma}{\gamma-1}$ . Because  $f'(\gamma) = \frac{-1}{(\gamma-1)^2} < 0$ ,  $f$  is monotonically decreasing. Thus, for a smaller gradient norm,  $\gamma$  should be set to a larger value, while for a smaller SFO complexity,  $\gamma$  should be set to a smaller value.

$[b_0]$ : From Theorem 6,  $\|G_T\|^2 = O(1 + b_0^{-1})$  and  $\text{SFO}_\epsilon^{\text{incr}} = O\left(\frac{b_0^3}{b_0^2 - 1}\right)$  hold.  $1 + b_0^{-1}$  is clearly monotonically decreasing. Let  $g(x) := \frac{x^3}{x^2 - 1}$  ( $x > 1$ ). Because  $g'(x) = \frac{x^2(x - \sqrt{3})(x + \sqrt{3})}{(x^2 - 1)^2}$ ,  $g$  is monotonically decreasing for  $1 < x \leq \sqrt{3}$ , and monotonically increasing for  $x \geq \sqrt{3}$ . Now, because  $x > 1$  and  $b_0 \in \mathbb{N}$ , it suffices to consider only the case  $x \geq 2$ , in which  $g$  is monotonically increasing. Summarizing the above and considering  $b_0 \in \text{dom } g$ ,

we find that, for a smaller gradient norm,  $b_0$  should be set to a larger value, while for a smaller SFO complexity,  $b_0$  should be set to a smaller value.

[M]: Recall  $T = MK$ . From Theorem 6,  $\|G_T\|^2 = O\left(\frac{1}{T} + \frac{K}{T}\right) = O\left(\frac{1}{MK} + \frac{K}{MK}\right) = O(M^{-1})$ , and  $\text{SFO}_\epsilon^{\text{incr}} = O\left(\frac{\gamma^M}{M}\right)$  hold.  $M^{-1}$  is clearly monotonically decreasing. Let  $h(x) := \frac{\gamma^x}{x}$  ( $x \geq 1$ ). Because  $h'(x) = \frac{\gamma^x(x \ln \gamma - 1)}{x^2}$ ,  $h$  is monotonically increasing for  $x \geq \frac{1}{\ln \gamma}$  when  $\gamma \geq e$ . Thus, when we set  $\gamma \geq e$  ( $\gamma$  was set to 3 in our experiments except for the far right plot in Figure 6), we find that, for a smaller gradient norm,  $M$  should be set to a larger value, while for a smaller SFO complexity,  $M$  should be set to a smaller value.

Finally, we provide a remark on the setting of  $b_0$  in our experiments. Although the initial BS ( $b_0$ ) appears different for each dataset, the values were similarly determined using the same method to enable a comparison of the effectiveness between a small constant BS, a large constant BS, and an increasing BS. The  $N$  was set to the total number of data points in each dataset. For a large initial BS, it was set to  $3^k$ , where  $k$  is the largest integer such that  $3^k < N$ . For a medium-size initial BS, it was set to  $3^{k-3}$ . For the smallest initial BS, it was set to  $3^{k-5}$ . For example, because  $N = 7200$  for the COIL100 dataset, the large, the medium-size small, and the smallest initial BSs were set to  $3^8 = 6561$ ,  $3^5 = 243$ ,  $3^3 = 27$ , respectively.

## Appendix C. Our Proposed Optimization Problem

Our proposed optimization problem:

$$\begin{aligned} \text{minimize} \quad & f(w) := \frac{1}{N} \sum_{j=1}^N \sqrt{|\langle x_j, w \rangle|}, \\ \text{subject to} \quad & w \in S^{n-1} := \{w \in \mathbb{R}^n \mid \|w\| = 1\}. \end{aligned}$$

$\{x_j\}_{j=1}^N$  is a dataset uniformly sampled from  $S^{n-1}$ .  $N$  is a total number of data points in the dataset, and  $n$  is a dimension of each one. The objective function defined above has unbounded gradient norms, which allows us to experimentally verify that our theoretical results are indeed applicable to functions with unbounded gradient norms. In our experiment,  $R_x(d) := \frac{x+d}{\|x+d\|}$  was used as a retraction for  $S^{n-1}$ .

**Proposition 11** *The gradient norm of the objective function for this problem  $f$  is not bounded.*

**Proof** When  $\langle x_j, w \rangle \neq 0$  ( $\forall j \in \{1, \dots, N\}$ ),

$$\nabla f(w) = \frac{1}{N} \sum_{j=1}^N \text{Sign}(\langle x_j, w \rangle) |\langle x_j, w \rangle|^{-\frac{1}{2}} x_j$$

is obtained by

$$\begin{aligned} \frac{\partial f}{\partial w_i}(w) &= \frac{1}{N} \sum_{j=1}^N \frac{\partial}{\partial w_i} \sqrt{|\langle x_j, w \rangle|} = \frac{1}{N} \sum_{j=1}^N |\langle x_j, w \rangle|^{-\frac{1}{2}} \text{Sign}(\langle x_j, w \rangle) \frac{\partial}{\partial w_i} \sum_{k=1}^N x_j^k w_k \\ &= \frac{1}{N} \sum_{j=1}^N \text{Sign}(\langle x_j, w \rangle) |\langle x_j, w \rangle|^{-\frac{1}{2}} x_j^i, \end{aligned}$$

where  $w = (w_1, \dots, w_n)^\top$ ,  $x_j = (x_j^1, \dots, x_j^n)^\top \in S^{n-1} \subset \mathbb{R}^n$ . Thus, the gradient on  $S^{n-1}$  is

$$\text{grad}f(w) = (I - ww^\top)\nabla f(w) = \frac{1}{N} \sum_{j=1}^N \text{Sign}(\langle x_j, w \rangle) |\langle x_j, w \rangle|^{-\frac{1}{2}} (I - ww^\top)x_j.$$

Fix one data  $x_i$  such that  $x_i \neq 0$ . Let  $u \in S^{n-1}$  be a vector obtained by orthonormalizing  $x_i$  via the Gram-Schmidt process. We consider a sequence  $(w_t)_{t \in \mathbb{N}}$  such that  $w_t := \frac{x_i}{t\|x_i\|} + \sqrt{1 - \frac{1}{t^2}}u$ , which lies in  $S^{n-1}$  because

$$\|w_t\|^2 = \frac{1}{t^2\|x_i\|^2}\|x_i\|^2 + \left(1 - \frac{1}{t^2}\right)\|u\|^2 + \frac{2}{t\|x_i\|}\sqrt{1 - \frac{1}{t^2}}\langle x_i, u \rangle = 1.$$

By simple calculations,

$$\begin{aligned} w_t w_t^\top &= \left( \frac{x_i}{t\|x_i\|} + \sqrt{1 - \frac{1}{t^2}}u \right) \left( \frac{x_i^\top}{t\|x_i\|} + \sqrt{1 - \frac{1}{t^2}}u^\top \right) \\ &= \frac{x_i x_i^\top}{t^2\|x_i\|^2} + \frac{1}{t^2\|x_i\|^2} \sqrt{1 - \frac{1}{t^2}} (x_i u^\top + u x_i^\top) + \left(1 - \frac{1}{t^2}\right) u u^\top \\ &\xrightarrow{t \rightarrow \infty} u u^\top \end{aligned}$$

and

$$\langle x_j, w_t \rangle = \left\langle x_j, \frac{x_i}{t\|x_i\|} + \sqrt{1 - \frac{1}{t^2}}u \right\rangle = \frac{\langle x_j, x_i \rangle}{t\|x_i\|} + \sqrt{1 - \frac{1}{t^2}}\langle x_j, u \rangle$$

hold. Now, we define a set  $\Lambda$  satisfying  $\{x_j\}_{j \in \Lambda} \subset \{x_j\}_{j=1}^N$  such that  $\langle x_j, u \rangle = 0$ . We find  $\langle x_j, w_t \rangle = \frac{\langle x_j, x_i \rangle}{t\|x_i\|}$  ( $\forall j \in \Lambda$ ), and  $\Lambda \neq \emptyset$  because of  $i \in \Lambda$ . Note that, for all  $j \notin \Lambda$ ,  $\langle x_j, u \rangle \neq 0$

holds, which yields  $\langle x_j, w_t \rangle \neq 0$ . From these observations, if  $\langle x_j, x_i \rangle \neq 0$  ( $\forall j \in \Lambda$ ),

$$\begin{aligned}
\|\text{grad}f(w_t)\| &= \left\| \frac{1}{N} \sum_{j=1}^N \text{Sign}(\langle x_j, w_t \rangle) |\langle x_j, w_t \rangle|^{-\frac{1}{2}} (I - w_t w_t^\top) x_j \right\| \\
&= \frac{1}{N} \left\| \sum_{j=1}^N |\langle x_j, w_t \rangle|^{-\frac{1}{2}} (I - w_t w_t^\top) x_j \right\| \\
&\geq \frac{1}{N} \left\| \sum_{j \in \Lambda} |\langle x_j, w_t \rangle|^{-\frac{1}{2}} (I - w_t w_t^\top) x_j \right\| - \frac{1}{N} \left\| \sum_{j \notin \Lambda} |\langle x_j, w_t \rangle|^{-\frac{1}{2}} (I - w_t w_t^\top) x_j \right\| \\
&= \frac{1}{N} \sqrt{\frac{t \|x_i\|}{\langle x_j, x_i \rangle}} \left\| \sum_{j \in \Lambda} (I - w_t w_t^\top) x_j \right\| - \frac{1}{N} \left\| \sum_{j \notin \Lambda} \left| \frac{\langle x_j, x_i \rangle}{t \|x_i\|} + \sqrt{1 - \frac{1}{t^2}} \langle x_j, u \rangle \right|^{-\frac{1}{2}} (I - w_t w_t^\top) x_j \right\| \\
&\geq \frac{1}{N} \sqrt{\frac{t \|x_i\|}{\langle x_j, x_i \rangle}} \left\| \sum_{j \in \Lambda} (I - w_t w_t^\top) x_j \right\| - \frac{1}{N} \sum_{j \notin \Lambda} \left| \frac{\langle x_j, x_i \rangle}{t \|x_i\|} + \sqrt{1 - \frac{1}{t^2}} \langle x_j, u \rangle \right|^{-\frac{1}{2}} \left\| (I - w_t w_t^\top) x_j \right\| \\
&\xrightarrow{t \rightarrow \infty} \left( \infty - \frac{1}{N} \sum_{j \notin \Lambda} |\langle x_j, u \rangle|^{-\frac{1}{2}} \left\| (I - u u^\top) x_j \right\| \right) = \infty
\end{aligned}$$

holds; even if  $\langle x_j, x_i \rangle = 0$  ( $\exists j \in \Lambda$ ), for all  $t \in \mathbb{N}$ ,

$$\begin{aligned}
\|\text{grad}f(w_t)\| &\geq \frac{1}{N} \left\| \sum_{j \in \Lambda} |\langle x_j, w_t \rangle|^{-\frac{1}{2}} (I - w_t w_t^\top) x_j \right\| - \frac{1}{N} \left\| \sum_{j \notin \Lambda} |\langle x_j, w_t \rangle|^{-\frac{1}{2}} (I - w_t w_t^\top) x_j \right\| \\
&= \left( \infty - \frac{1}{N} \left\| \sum_{j \notin \Lambda} |\langle x_j, w_t \rangle|^{-\frac{1}{2}} (I - w_t w_t^\top) x_j \right\| \right) = \infty
\end{aligned}$$

holds. These complete the proof. ■

### C.1. Gradient Norm versus SFO Complexity for Our Proposed] Optimization Problem

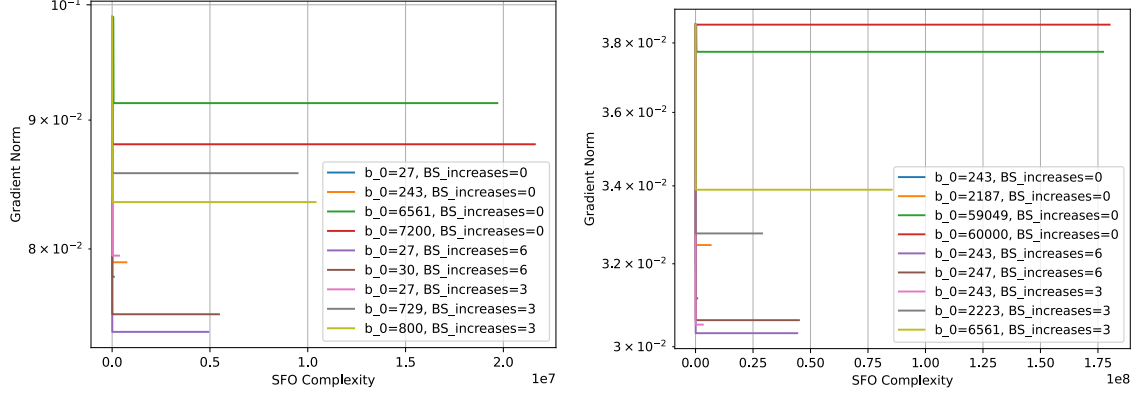


Figure 8: Norm of objective function gradient versus SFO complexity. Plot on left (resp. right) shows norm of objective function gradient for our proposed problem when  $(N, n) = (7200, 1024)$  (resp. when  $(N, n) = (60000, 1024)$ ).

As shown in the plot of the norm of objective function gradient versus SFO complexity 8 for our proposed problem, performance was better with an increasing BS than with either a small or large constant BS. This result can be explained by Theorem 9, which shows that an increasing BS reduces SFO complexity compared with a constant BS.

### Appendix D. Proofs of our Lemma and Theorems

Recall that  $\Pi$  is a  $\{1, \dots, N\}$ -valued probability distribution. Thus, we consider the probability space  $(\{1, \dots, N\}, \mathfrak{P}(\{1, \dots, N\}), \Pi)$ . For a random variable  $\xi$  distributed as  $\Pi$  and a function  $g : \{1, \dots, N\} \times \mathcal{M} \ni (\xi, x) \mapsto g_\xi(x) \in \mathbb{R}$ , we define  $\mathbb{E}_\xi[g_\xi(x)] := \mathbb{E}_{\xi \sim \Pi}[g_\xi(x)] := \int_{\{1, \dots, N\}} g_\xi(x) \Pi(d\xi) = \sum_{j=1}^N g_j(x) \Pi(\{j\})$ . Hence, the variance is defined as  $\mathbb{V}_{\xi \sim \Pi}(g_\xi(x)) := \mathbb{E}_{\xi \sim \Pi}[\|g_\xi(x) - \mathbb{E}_{\xi \sim \Pi}[g_\xi(x)]\|^2]$ . Let  $(\xi_{i,t})_{i=1}^N$  be a sequence of i.i.d. random variables distributed as  $\Pi$ , let  $\boldsymbol{\xi}_t := (\xi_{1,t}, \dots, \xi_{b_t,t})^\top \in \{1, \dots, N\}^{b_t}$ , let  $g : \{1, \dots, N\}^{b_t} \times \mathcal{M} \ni (\boldsymbol{\xi}, x) \mapsto g_{\boldsymbol{\xi}}(x) \in \mathbb{R}$  be a function, and let  $x_t$  be a point at the  $t$ -th iteration generated by the updating rule of RSGD. We can introduce a natural extension of the expectation notation to a multivariate random variable:

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\xi}_t}[g_{\boldsymbol{\xi}_t}(x_t)] &:= \mathbb{E}_{\boldsymbol{\xi}_t \sim \Pi^{b_t}}[g_{\boldsymbol{\xi}_t}(x_t)] := \int_{\{1, \dots, N\}^{b_t}} g_{\boldsymbol{\xi}_t}(x_t) \Pi^{b_t}(d\boldsymbol{\xi}_t) \\ &= \sum_{j_1=1}^N \cdots \sum_{j_{b_t}=1}^N g_{(j_1, \dots, j_{b_t})^\top}(x_t) \Pi(\{j_1\}) \cdots \Pi(\{j_{b_t}\}) \\ &= \mathbb{E}_{\xi_{1,t}} \mathbb{E}_{\xi_{2,t}} \cdots \mathbb{E}_{\xi_{b_t,t}}[g_{\boldsymbol{\xi}_t}(x_t)]. \end{aligned}$$

For  $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_t$ , a total expectation is defined as  $\mathbb{E} := \mathbb{E}_{\boldsymbol{\xi}_1} \cdots \mathbb{E}_{\boldsymbol{\xi}_t}$ . Note that, from Assumption 3,  $\mathbb{E}_\xi[\|\text{grad} f_\xi(x) - \text{grad} f(x)\|_x^2] = \mathbb{V}_\xi(\text{grad} f_\xi(x)) \leq \sigma^2$  holds, which implies

$\mathbb{E}_{\xi} [\|\text{grad} f_{\xi}(x) - \text{grad} f(x)\|_x^2] \leq \frac{\sigma^2}{b_t}$ . The following result serves as a preliminary for Lemma 4.

**Lemma 12 (Descent Lemma)** *Let  $(x_t)_t$  be a sequence generated by RSGD and  $(\eta_t)_t$  be a positive-valued sequence. Then, under Assumptions 2 and 3, we obtain*

$$\mathbb{E}[f(x_{t+1})] \leq \mathbb{E}[f(x_t)] + \frac{L_r \sigma^2 \eta_t^2}{2b_t} - \eta_t \left(1 - \frac{L_r \eta_t}{2}\right) \mathbb{E}[\|\text{grad} f(x_t)\|_{x_t}^2].$$

**Proof** Under Assumption 2, we start with

$$f(x_{t+1}) \leq f(x_t) - \eta_t \langle \text{grad} f(x_t), \text{grad} f_{B_t}(x_t) \rangle_{x_t} + \frac{\eta_t^2}{2} L_r \|\text{grad} f_{B_t}(x_t)\|_{x_t}^2,$$

$$\begin{aligned} \mathbb{E}_{\xi_t} [\|\text{grad} f_{B_t}(x_t)\|_{x_t}^2] &= \mathbb{E}_{\xi_t} [\|\text{grad} f_{B_t}(x_t) - \text{grad} f(x_t) + \text{grad} f(x_t)\|_{x_t}^2] \\ &\leq \mathbb{E}_{\xi_t} [\|\text{grad} f_{B_t}(x_t) - \text{grad} f(x_t)\|_{x_t}^2] + \mathbb{E}_{\xi_t} [\|\text{grad} f(x_t)\|_{x_t}^2] \\ &\quad + 2\mathbb{E}_{\xi_t} [\langle \text{grad} f_{B_t}(x_t) - \text{grad} f(x_t), \text{grad} f(x_t) \rangle_{x_t}] \\ &= \mathbb{E}_{\xi_t} [\|\text{grad} f_{B_t}(x_t) - \text{grad} f(x_t)\|_{x_t}^2] + \mathbb{E}_{\xi_t} [\|\text{grad} f(x_t)\|_{x_t}^2] \\ &\quad + 2\langle \mathbb{E}_{\xi_t} [\text{grad} f_{B_t}(x_t)] - \text{grad} f(x_t), \text{grad} f(x_t) \rangle_{x_t} \\ &\leq \frac{\sigma^2}{b_t} + \mathbb{E}_{\xi_t} [\|\text{grad} f(x_t)\|_{x_t}^2], \end{aligned}$$

and

$$\mathbb{E}_{\xi_t} [\langle \text{grad} f(x_t), \text{grad} f_{B_t}(x_t) \rangle_{x_t}] = \langle \text{grad} f(x_t), \mathbb{E}_{\xi_t} [\text{grad} f_{B_t}(x_t)] \rangle_{x_t} = \|\text{grad} f(x_t)\|_{x_t}^2.$$

Taking the total expectation of the above equations, we have

$$\begin{aligned} \mathbb{E}[f(x_{t+1})] &\leq \mathbb{E}[f(x_t)] - \eta_t \mathbb{E}[\langle \text{grad} f(x_t), \text{grad} f_{B_t}(x_t) \rangle_{x_t}] + \frac{\eta_t^2}{2} L_r \mathbb{E}[\|\text{grad} f_{B_t}(x_t)\|_{x_t}^2] \\ &= \mathbb{E}[f(x_t)] - \eta_t \mathbb{E}[\mathbb{E}_{\xi_t} [\langle \text{grad} f(x_t), \text{grad} f_{B_t}(x_t) \rangle_{x_t}]] + \frac{\eta_t^2}{2} L_r \mathbb{E}[\mathbb{E}_{\xi_t} [\|\text{grad} f_{B_t}(x_t)\|_{x_t}^2]] \\ &\leq \mathbb{E}[f(x_t)] - \eta_t \mathbb{E}[\|\text{grad} f(x_t)\|_{x_t}^2] + \frac{\eta_t^2 L_r}{2} \left( \frac{\sigma^2}{b_t} + \mathbb{E}[\|\text{grad} f(x_t)\|_{x_t}^2] \right) \\ &= \mathbb{E}[f(x_t)] + \frac{L_r \sigma^2 \eta_t^2}{2b_t} - \eta_t \left(1 - \frac{L_r \eta_t}{2}\right) \mathbb{E}[\|\text{grad} f(x_t)\|_{x_t}^2]. \end{aligned}$$

■

#### D.1. Proof of Lemma 4

On the basis of Lemma 12, Lemma 4 is proven as follows.

**Proof** Taking  $\eta_{\max} < \frac{2}{L_r}$  into consideration, we start with

$$\sum_{t=0}^{T-1} \eta_t \left(1 - \frac{L_r \eta_t}{2}\right) \mathbb{E}[\|\text{grad} f(x_t)\|_{x_t}^2] \geq \left(1 - \frac{L_r \eta_{\max}}{2}\right) \min_{t \in \{0, \dots, T-1\}} \mathbb{E}[\|\text{grad} f(x_t)\|_{x_t}^2] \sum_{t=0}^{T-1} \eta_t.$$



By taking the summation on both sides of the inequality for Lemma 12 and evaluating it using the above inequality, we obtain

$$\begin{aligned}
 & \left(1 - \frac{L_r \eta_{\max}}{2}\right) \min_{t \in \{0, \dots, T-1\}} \mathbb{E}[\|\text{grad} f(x_t)\|_{x_t}^2] \sum_{t=0}^{T-1} \eta_t \\
 & \leq \sum_{t=0}^{T-1} \eta_t \left(1 - \frac{L_r \eta_t}{2}\right) \mathbb{E}[\|\text{grad} f(x_t)\|_{x_t}^2] \\
 & \leq \sum_{t=0}^{T-1} (\mathbb{E}[f(x_t)] - \mathbb{E}[f(x_{t+1})]) + \sum_{t=0}^{T-1} \frac{L_r \sigma^2 \eta_t^2}{2b_t} = \mathbb{E}[f(x_0) - f(x_T)] + \sum_{t=0}^{T-1} \frac{L_r \sigma^2 \eta_t^2}{2b_t}.
 \end{aligned}$$

$(\mathbb{E}[f(x_t)])_{t \in \{0, \dots, T\}}$  being a decreasing sequence implies

$$\mathbb{E}[f(x_0) - f(x_T)] \leq \mathbb{E}[f(x_0) - f^*] = f(x_0) - f^*.$$

Therefore,

$$\min_{t \in \{0, \dots, T-1\}} \mathbb{E}[\|\text{grad} f(x_t)\|_{x_t}^2] \leq \frac{2(f(x_0) - f^*)}{2 - L_r \eta_{\max}} \frac{1}{\sum_{t=0}^{T-1} \eta_t} + \frac{L_r \sigma^2}{2 - L_r \eta_{\max}} \frac{1}{\sum_{t=0}^{T-1} \eta_t} \sum_{t=0}^{T-1} \frac{\eta_t^2}{b_t}.$$

■

## D.2. Proof of Theorem 5

**Proof** We set  $b_t = b$  in Lemma 4.

**[Constant LR (1)]**

From Lemma 4, we have

$$\sum_{t=0}^{T-1} \eta_t = \sum_{t=0}^{T-1} \eta_{\max} = \eta_{\max} T, \quad \sum_{t=0}^{T-1} \frac{\eta_t^2}{b_t} = \frac{1}{\sum_{t=0}^{T-1} \eta_{\max}} \sum_{t=0}^{T-1} \frac{\eta_{\max}^2}{b} = \frac{\eta_{\max}^2}{b} T,$$

which yields

$$\min_{t \in \{0, \dots, T-1\}} \mathbb{E}[\|\text{grad} f(x_t)\|_{x_t}^2] \leq \frac{2(f(x_0) - f^*)}{(2 - L_r \eta_{\max}) \eta_{\max} T} + \frac{L_r \eta_{\max}}{2 - L_r \eta_{\max}} \frac{\sigma^2}{b}.$$

**[Diminishing LR (2)]**

From Lemma 4, we have

$$\sum_{t=0}^{T-1} \eta_t = \eta_{\max} \sum_{t=0}^{T-1} \frac{1}{\sqrt{t+1}} \geq \eta_{\max} \sum_{t=0}^{T-1} \frac{1}{\sqrt{T}} = \eta_{\max} \frac{T}{\sqrt{T}} = \eta_{\max} \sqrt{T}$$

and

$$\sum_{t=0}^{T-1} \frac{\eta_t^2}{b_t} = \frac{\eta_{\max}^2}{b} \sum_{t=0}^{T-1} \frac{1}{t+1} \leq \frac{\eta_{\max}^2}{b} \left(1 + \int_1^T \frac{dt}{t}\right) = \frac{\eta_{\max}^2}{b} + \frac{\eta_{\max}^2}{b} \log T,$$

which yields

$$\min_{t \in \{0, \dots, T-1\}} \mathbb{E}[\|\text{grad} f(x_t)\|_{x_t}^2] \leq \frac{f(x_0) - f^*}{(2 - L_r \eta_{\max}) \eta_{\max}} \frac{1}{\sqrt{T}} + \frac{L_r \eta_{\max}}{2 - L_r \eta_{\max}} \frac{\sigma^2}{b} \frac{1 + \log T}{\sqrt{T}}.$$

**[Cosine Annealing LR (3)]**

From Lemma 4, we have

$$\begin{aligned} \sum_{t=0}^{T-1} \eta_t &\geq \int_0^T \eta_t dt = \frac{\eta_{\max} + \eta_{\min}}{2} T + \frac{\eta_{\max} - \eta_{\min}}{2} \int_0^T \cos\left(\frac{t}{T} \pi\right) dt \\ &= \frac{\eta_{\max} + \eta_{\min}}{2} T \end{aligned}$$

and

$$\begin{aligned} \sum_{t=0}^{T-1} \eta_t^2 &= \frac{(\eta_{\max} + \eta_{\min})^2}{4} T + \frac{\eta_{\max}^2 + \eta_{\min}^2}{2} \sum_{t=0}^{T-1} \cos^2 \frac{t}{T} \pi + \frac{(\eta_{\max} - \eta_{\min})^2}{4} \sum_{t=0}^{T-1} \cos^2 \frac{t}{T} \pi \\ &\leq \frac{(\eta_{\max} + \eta_{\min})^2}{4} T + \frac{\eta_{\max}^2 - \eta_{\min}^2}{2} T + \frac{(\eta_{\max} - \eta_{\min})^2}{4} T \\ &= \eta_{\max}^2 T, \end{aligned}$$

which yields

$$\min_{t \in \{0, \dots, T-1\}} \mathbb{E}[\|\text{grad} f(x_t)\|_{x_t}^2] \leq \frac{4(f(x_0) - f^*)}{(2 - L_r \eta_{\max})(\eta_{\max} + \eta_{\min})} \frac{1}{T} + \frac{2L_r \eta_{\max}^2}{(2 - L_r \eta_{\max})(\eta_{\max} + \eta_{\min})} \frac{\sigma^2}{b}.$$

**[Polynomial Decay LR (4)]**

We start by recalling the Riemann integral of  $g(t) := (1 - t)^p \geq 0$  ( $0 \leq t \leq 1$ ).  $U(P_T) := \sum_{t=0}^{T-1} \frac{1}{T} (1 - \frac{t}{T})^p$  is an upper Riemann sum for the function  $g$  over the partition  $P_T := \{0, \frac{1}{T}, \dots, \frac{T-1}{T}, 1\}$  due to  $g$  being monotonically decreasing. From the definition of a Riemann integral, if  $\int_0^1 g(t) dt$  exists, then

$$\inf_{P: \text{Partitions of } [0,1]} U(P) = \int_0^1 g(t) dt.$$

In fact, the integral value exists. Thus, we obtain

$$\sum_{t=0}^{T-1} \frac{1}{T} \left(1 - \frac{t}{T}\right)^p = U(P_T) \geq \inf_{P: \text{Partitions of } [0,1]} U(P) = \int_0^1 (1 - t)^p dt = \frac{1}{p+1}.$$

Similarly, if we define  $L(P_T)$  as  $\sum_{t=1}^T \frac{1}{T} (1 - \frac{t}{T})^p$ , then  $L(P_T)$  is a lower Riemann sum for the function  $g$  over the partition  $P_T$ , and the supremum of  $L(P_T)$  equals  $\int_0^1 g(t) dt$ . Considering

$$\sum_{t=0}^{T-1} \frac{1}{T} \left(1 - \frac{t}{T}\right)^p = L(P_T) + \frac{1}{T},$$

we have

$$\sum_{t=0}^{T-1} \frac{1}{T} \left(1 - \frac{t}{T}\right)^p = L(P_T) + \frac{1}{T} \leq \sup_{P: \text{Partitions of } [0,1]} L(P) + \frac{1}{T} = \int_0^1 (1-t)^p dt + \frac{1}{T} = \frac{1}{p+1} + \frac{1}{T}.$$

Hence,

$$\frac{T}{p+1} \leq \sum_{t=0}^{T-1} \left(1 - \frac{t}{T}\right)^p \leq \frac{T}{p+1} + 1$$

holds. By replacing  $g$  with  $(1+t)^{2p}(0 \leq t \leq 1)$  and using the same logic, we obtain

$$\sum_{t=0}^{T-1} \frac{1}{T} \left(1 - \frac{t}{T}\right)^{2p} \leq \int_0^1 (1-t)^{2p} dt + \frac{1}{T} = \frac{1}{2p+1} + \frac{1}{T}.$$

Therefore, considering Lemma 4,

$$\begin{aligned} \sum_{t=0}^{T-1} \eta_t &= \eta_{\min} T + (\eta_{\max} - \eta_{\min}) \sum_{t=0}^{T-1} \left(1 - \frac{t}{T}\right)^p \\ &\geq \eta_{\min} T + (\eta_{\max} - \eta_{\min}) \frac{T}{p+1} = \frac{\eta_{\max} + p\eta_{\min}}{p+1} T \end{aligned}$$

and

$$\begin{aligned} \sum_{t=0}^{T-1} \eta_t^2 &= \eta_{\min} T + 2\eta_{\min}(\eta_{\max} - \eta_{\min}) \sum_{t=0}^{T-1} \left(1 - \frac{t}{T}\right)^p + (\eta_{\max} - \eta_{\min})^2 \sum_{t=0}^{T-1} \left(1 - \frac{t}{T}\right)^{2p} \\ &\leq \eta_{\min} T + 2\eta_{\min}(\eta_{\max} - \eta_{\min}) \left(1 + \frac{T}{p+1}\right) + (\eta_{\max} - \eta_{\min})^2 \left(1 + \frac{T}{2p+1}\right) \\ &= \eta_{\max}^2 - \eta_{\min}^2 + \left(\eta_{\min} + \frac{2\eta_{\min}(\eta_{\max} - \eta_{\min})}{p+1} + \frac{(\eta_{\max} - \eta_{\min})^2}{2p+1}\right) T \end{aligned}$$

hold, which yields

$$\begin{aligned} &\min_{t \in \{0, \dots, T-1\}} \mathbb{E}[\|\text{grad} f(x_t)\|_{x_t}^2] \\ &\leq \frac{p+1}{(2 - L_r \eta_{\max})(\eta_{\max} + p\eta_{\min})} \left\{ 2(f(x_0) - f^*) + \frac{L_r(\eta_{\max}^2 - \eta_{\min}^2)\sigma^2}{b} \right\} \frac{1}{T} \\ &\quad + \frac{L_r(p+1)(\eta_{\max}^2 - \eta_{\min}^2)}{(2 - L_r \eta_{\max})(\eta_{\max} + p\eta_{\min})} \left( \eta_{\min} + \frac{2\eta_{\min}(\eta_{\max} - \eta_{\min})}{p+1} + \frac{(\eta_{\max} - \eta_{\min})^2}{2p+1} \right) \frac{\sigma^2}{b}. \end{aligned}$$

■

### D.3. Proof of Theorem 6

**Proof** The evaluations in all the following cases are based on Lemma 4 and use estimations of  $\sum_{t=0}^{T-1} \eta_t$  in the proof of Theorem 5.

**[Exponential Growth BS (5) and Constant LR (1)]**

From the fact that sums of positive term series are the supremum of their finite sums, we have

$$\sum_{t=0}^{T-1} \frac{1}{b_t} = \sum_{m=0}^{M-1} \frac{K}{b_0 \gamma^m} + \frac{T-KM}{b_0 \gamma^M} \leq \sum_{m=0}^M \frac{K}{b_0 \gamma^m} \leq \frac{K}{b_0} \sum_{m=0}^{\infty} \frac{1}{\gamma^m} = \frac{K\gamma}{b_0(\gamma-1)}, \quad (9)$$

which implies

$$\sum_{t=0}^{T-1} \frac{\eta_{\max}^2}{b_t} = \eta_{\max}^2 \sum_{t=0}^{T-1} \frac{1}{b_t} \leq \frac{\eta_{\max}^2 K\gamma}{b_0(\gamma-1)}.$$

Therefore,

$$\min_{t \in \{0, \dots, T-1\}} \mathbb{E}[\|\text{grad} f(x_t)\|_{x_t}^2] \leq \frac{1}{(2 - L_r \eta_{\max}) \eta_{\max}} \left\{ 2(f(x_0) - f^*) + \frac{L_r \eta_{\max}^2 K\gamma \sigma^2}{\gamma - 1} \frac{1}{b_0} \right\} \frac{1}{T}.$$

**[Exponential Growth BS (5) and Diminishing LR (2)]**

Using (9), we have

$$\sum_{t=0}^{T-1} \frac{\eta_t^2}{b_t} = \eta_{\max}^2 \sum_{t=0}^{T-1} \frac{1}{(t+1)b_t} \leq \eta_{\max}^2 \sum_{t=0}^{T-1} \frac{1}{b_t} \leq \frac{\eta_{\max}^2 K\gamma}{b_0(\gamma-1)},$$

which yields

$$\min_{t \in \{0, \dots, T-1\}} \mathbb{E}[\|\text{grad} f(x_t)\|_{x_t}^2] \leq \frac{1}{(2 - L_r \eta_{\max}) \eta_{\max}} \left\{ 2(f(x_0) - f^*) + \frac{L_r \eta_{\max}^2 K\gamma \sigma^2}{\gamma - 1} \frac{1}{b_0} \right\} \frac{1}{\sqrt{T}}.$$

**[Exponential Growth BS (5) and Cosine Annealing LR (3)]**

From (9) and  $|\cos x| \leq 1$ , we have

$$\begin{aligned} \sum_{t=0}^{T-1} \frac{\eta_t^2}{b_t} &= \sum_{t=0}^{T-1} \frac{1}{b_t} \left( \frac{(\eta_{\max} + \eta_{\min})^2}{4} + \frac{(\eta_{\max} - \eta_{\min})^2}{4} \cos^2 \frac{t}{T} \pi + \frac{\eta_{\max}^2 - \eta_{\min}^2}{2} \cos \frac{t}{T} \pi \right) \\ &\leq \eta_{\max}^2 \sum_{t=0}^{T-1} \frac{1}{b_t} \leq \frac{\eta_{\max}^2 K\gamma}{b_0(\gamma-1)}, \end{aligned}$$

which yields

$$\min_{t \in \{0, \dots, T-1\}} \mathbb{E}[\|\text{grad} f(x_t)\|_{x_t}^2] \leq \frac{2}{(2 - L_r \eta_{\max})(\eta_{\max} + \eta_{\min})} \left\{ 2(f(x_0) - f^*) + \frac{L_r \eta_{\max}^2 K\gamma \sigma^2}{\gamma - 1} \frac{1}{b_0} \right\} \frac{1}{T}.$$

**[Exponential Growth BS (5) and Polynomial Decay LR (4)]**

From (9) and  $t \leq T$ , we have

$$\begin{aligned} \sum_{t=0}^{T-1} \frac{\eta_t^2}{b_t} &= \sum_{t=0}^{T-1} \frac{1}{b_t} \left\{ \eta_{\min}^2 + (\eta_{\max} - \eta_{\min})^2 \left(1 - \frac{t}{T}\right)^{2p} + 2\eta_{\min}(\eta_{\max} - \eta_{\min}) \left(1 - \frac{t}{T}\right)^p \right\} \\ &\leq \eta_{\max}^2 \sum_{t=0}^{T-1} \frac{1}{b_t} \leq \frac{\eta_{\max}^2 K \gamma}{b_0(\gamma - 1)}, \end{aligned}$$

which yields

$$\min_{t \in \{0, \dots, T-1\}} \mathbb{E}[\|\text{grad} f(x_t)\|_{x_t}^2] \leq \frac{p+1}{(2 - L_r \eta_{\max})(\eta_{\max} + p\eta_{\min})} \left\{ 2(f(x_0) - f^*) + \frac{L_r \eta_{\max}^2 K \gamma \sigma^2}{\gamma - 1} \frac{1}{b_0} \right\} \frac{1}{T}.$$

**[Polynomial Growth BS (6) and Constant LR (1)]**

We set  $\underline{a} := a \wedge b_0$ . Considering  $c > 1$ , we have

$$\begin{aligned} \sum_{t=0}^{T-1} \frac{1}{b_t} &= \sum_{m=0}^{M-1} \frac{K}{(am + b_0)^c} + \frac{T - KM}{(aM + b_0)^c} \leq \sum_{m=0}^M \frac{K}{(am + b_0)^c} \leq \frac{K}{\underline{a}^{[c]}} \sum_{m=0}^M \frac{1}{(m+1)^c} \\ &= \frac{K}{\underline{a}^{[c]}} \sum_{m=1}^{\infty} \frac{1}{m^c} = \frac{K\zeta(c)}{\underline{a}^{[c]}}, \end{aligned} \tag{10}$$

which yields

$$\sum_{t=0}^{T-1} \frac{\eta_t^2}{b_t} = \eta_{\max}^2 \sum_{t=0}^{T-1} \frac{1}{b_t} \leq \frac{\eta_{\max}^2 K \zeta(c)}{\underline{a}^{[c]}}.$$

Therefore,

$$\min_{t \in \{0, \dots, T-1\}} \mathbb{E}[\|\text{grad} f(x_t)\|_{x_t}^2] \leq \frac{1}{(2 - L_r \eta_{\max})\eta_{\max}} \left( 2(f(x_0) - f^*) + \frac{L_r \eta_{\max}^2 K \zeta(c) \sigma^2}{\underline{a}^{[c]}} \right) \frac{1}{T},$$

where, for  $c > 1$ , Riemann zeta function  $\zeta(c) < \infty$ ;  $\zeta(c)$  is monotonically decreasing on  $c \in (1, \infty)$  and  $\lim_{c \rightarrow \infty} \zeta(c) = 1$ .

**[Polynomial Growth BS (6) and Diminishing LR (2)]**

From (10), we have

$$\sum_{t=0}^{T-1} \frac{\eta_t^2}{b_t} = \eta_{\max}^2 \sum_{t=0}^{T-1} \frac{1}{(t+1)b_t} \leq \eta_{\max}^2 \sum_{t=0}^{T-1} \frac{1}{b_t} \leq \frac{\eta_{\max}^2 K \zeta(c)}{\underline{a}^{[c]}},$$

which yields

$$\min_{t \in \{0, \dots, T-1\}} \mathbb{E}[\|\text{grad} f(x_t)\|_{x_t}^2] \leq \frac{1}{(2 - L_r \eta_{\max})\eta_{\max}} \left( 2(f(x_0) - f^*) + \frac{L_r \eta_{\max}^2 K \zeta(c) \sigma^2}{\underline{a}^{[c]}} \right) \frac{1}{\sqrt{T}}.$$

**[Polynomial Growth BS (6) and Cosine Annealing LR (3)]**

From (10) and the previous analysis for Cosine Annealing LR (3), we have

$$\sum_{t=0}^{T-1} \frac{\eta_t^2}{b_t} \leq \eta_{\max}^2 \sum_{t=0}^{T-1} \frac{1}{b_t} \leq \frac{\eta_{\max}^2 K \zeta(c)}{\underline{a}^{[c]}},$$

which yields

$$\min_{t \in \{0, \dots, T-1\}} \mathbb{E}[\|\text{grad} f(x_t)\|_{x_t}^2] \leq \frac{2}{(2 - L_r \eta_{\max})(\eta_{\max} + \eta_{\min})} \left( 2(f(x_0) - f^*) + \frac{L_r \eta_{\max}^2 K \zeta(c) \sigma^2}{\underline{a}^{[c]}} \right) \frac{1}{T}.$$

**[Polynomial Growth BS (6) and Polynomial Decay LR (4)]**

From (10) and the previous analysis for Polynomial Decay LR (4), we have

$$\sum_{t=0}^{T-1} \frac{\eta_t^2}{b_t} \leq \eta_{\max}^2 \sum_{t=0}^{T-1} \frac{1}{b_t} \leq \frac{\eta_{\max}^2 K \zeta(c)}{\underline{a}^{[c]}},$$

which yields

$$\min_{t \in \{0, \dots, T-1\}} \mathbb{E}[\|\text{grad} f(x_t)\|_{x_t}^2] \leq \frac{p+1}{(2 - L_r \eta_{\max})(\eta_{\max} + p\eta_{\min})} \left( 2(f(x_0) - f^*) + \frac{L_r \eta_{\max}^2 K \zeta(c) \sigma^2}{\underline{a}^{[c]}} \right) \frac{1}{T}.$$

■

**D.4. Proof of Theorem 7**

**Theorem 13 (Detailed Version of Theorem 7)** *We consider BSs (5) and (6) and warm-up LR (7) and (8) with decay parts given by (1), (2), (3), or (4) under the assumptions of Lemma 4. Then, the following holds for both constant and increasing BSs.*

- *Decay part: Diminishing (2)*

$$\min_{t \in \{T_w, \dots, T-1\}} \mathbb{E}[\|\text{grad} f(x_t)\|_{x_t}^2] \leq \frac{Q_1 + Q_2 \sigma^2 b_0^{-1}}{\sqrt{T+1} - \sqrt{T_w+1}},$$

- *Decay part: Otherwise (1), (3), (4)*

$$\min_{t \in \{T_w, \dots, T-1\}} \mathbb{E}[\|\text{grad} f(x_t)\|_{x_t}^2] \leq \frac{\tilde{Q}_1 + \tilde{Q}_2 \sigma^2 b_0^{-1}}{T - T_w},$$

where  $Q_1, Q_2, \tilde{Q}_1$ , and  $\tilde{Q}_2$  are constants that do not depend on  $T$ .

**Proof** The evaluations in all these cases are based on Lemma 4. We start with an exponential growth BS and a warm-up LR. From  $l, l_w \in \mathbb{N}$ , and  $l_w \geq l$ , there exist  $\alpha, \beta \in \mathbb{N} \cup \{0\}$  such that  $l_w = \alpha l + \beta$ . Note that we define summations from 0 to  $-1$  as 0. Furthermore, from  $\forall u, v > 0 : 0 \leq u^2 + v^2 + uv$  holds.

**[LR Decay Part: Constant (1)]**

Considering  $T_w = l_w K' = \alpha(lK') + \beta K'$ , we have

$$\sum_{t=T_w}^{T-1} \eta_t = \sum_{t=T_w}^{T-1} \eta_{\max} = (T - T_w) \eta_{\max}$$

and

$$\sum_{t=T_w}^{T-1} \frac{\eta_t^2}{b_t} = \sum_{t=l_w K'}^{(\alpha+1)l-1} \frac{\eta_{\max}^2}{b_t} + \sum_{t=(\alpha+1)l}^{T-1} \frac{\eta_{\max}^2}{b_t} = \sum_{t=T_w}^{T-1} \frac{\eta_{\max}^2}{b_t} \leq \frac{\eta_{\max}^2 K}{b_0} \sum_{m=0}^M \frac{1}{\gamma^m} \leq \frac{\eta_{\max}^2 K}{b_0} \sum_{m=0}^{\infty} \frac{1}{\gamma^m} = \frac{\eta_{\max}^2 \gamma K}{(\gamma - 1) b_0},$$

which yields

$$\min_{t \in \{T_w, \dots, T-1\}} \mathbb{E}[\|\text{grad} f(x_t)\|_{x_t}^2] \leq \frac{1}{(2 - L_r \eta_{\max}) \eta_{\max}} \left( 2(f(x_0) - f^*) + \frac{L_r \eta_{\max}^2 K \gamma \sigma^2}{\gamma - 1} \frac{1}{b_0} \right) \frac{1}{T - T_w}.$$

**[LR Decay Part: Diminishing (2)]**

Similarly, we have

$$\sum_{t=T_w}^{T-1} \eta_t = \sum_{t=T_w}^{T-1} \frac{\eta_{\max}}{\sqrt{t+1}} \geq \eta_{\max} \int_{T_w}^T \frac{dt}{\sqrt{t+1}} = 2\eta_{\max}(\sqrt{T+1} - \sqrt{T_w+1})$$

and

$$\sum_{t=0}^{T-1} \frac{\eta_t^2}{b_t} \leq \sum_{t=T_w}^{T-1} \frac{\eta_{\max}^2}{b_t(t+1)} \leq \sum_{t=T_w}^{T-1} \frac{\eta_{\max}^2}{b_t} \leq \frac{\eta_{\max}^2 \gamma K}{(\gamma - 1) b_0},$$

which yields

$$\min_{t \in \{T_w, \dots, T-1\}} \mathbb{E}[\|\text{grad} f(x_t)\|_{x_t}^2] \leq \frac{1}{2(2 - L_r \eta_{\max}) \eta_{\max}} \left( 2(f(x_0) - f^*) + \frac{L_r \eta_{\max}^2 K \gamma \sigma^2}{\gamma - 1} \frac{1}{b_0} \right) \frac{1}{\sqrt{T+1} - \sqrt{T_w+1}}.$$

**[LR Decay Part: Cosine Annealing (3)]**

We have

$$\begin{aligned} \sum_{t=T_w}^{T-1} \eta_t &= \sum_{t=T_w}^{T-1} \left( \frac{\eta_{\max} + \eta_{\min}}{2} + \frac{\eta_{\max} - \eta_{\min}}{2} \cos \frac{t - T_w}{T - T_w} \pi \right) \\ &= \frac{\eta_{\max} + \eta_{\min}}{2} (T - T_w) + \frac{\eta_{\max} - \eta_{\min}}{2} \sum_{t=T_w}^{T-1} \cos \frac{t - T_w}{T - T_w} \pi \\ &\geq \frac{\eta_{\max} + \eta_{\min}}{2} (T - T_w) + \frac{\eta_{\max} - \eta_{\min}}{2} \int_{T_w}^T \cos \left( \frac{t - T_w}{T - T_w} \pi \right) dt = \frac{\eta_{\max} + \eta_{\min}}{2} (T - T_w) \end{aligned}$$

and

$$\begin{aligned} \sum_{t=T_w}^{T-1} \frac{\eta_t^2}{b_t} &\leq \sum_{t=T_w}^{T-1} \frac{1}{b_t} \left( \frac{(\eta_{\max} + \eta_{\min})^2}{4} + \frac{\eta_{\max}^2 - \eta_{\min}^2}{2} \cos \frac{t - T_w}{T - T_w} \pi + \frac{(\eta_{\max} - \eta_{\min})^2}{4} \cos^2 \frac{t - T_w}{T - T_w} \pi \right) \\ &\leq \sum_{t=0}^{T-1} \frac{1}{b_t} \left( \frac{(\eta_{\max} + \eta_{\min})^2}{4} + \frac{\eta_{\max}^2 - \eta_{\min}^2}{2} + \frac{(\eta_{\max} - \eta_{\min})^2}{4} \right) = \sum_{t=0}^{T-1} \frac{\eta_{\max}^2}{b_t} \leq \frac{\eta_{\max}^2 \gamma K}{(\gamma - 1) b_0}, \end{aligned}$$

which yields

$$\min_{t \in \{T_w, \dots, T-1\}} \mathbb{E}[\|\text{grad} f(x_t)\|_{x_t}^2] \leq \frac{2}{(2 - L_r \eta_{\max})(\eta_{\max} + \eta_{\min})} \left( 2(f(x_0) - f^*) + \frac{L_r \eta_{\max}^2 K \gamma \sigma^2}{\gamma - 1} \frac{1}{b_0} \right) \frac{1}{T - T_w}.$$

**[LR Decay Part: Polynomial Decay (4)]**

As with the proof for a polynomial decay LR (4) in Theorem 5, we have

$$\begin{aligned} \sum_{t=T_w}^{T-1} \eta_t &= \sum_{t=T_w}^{T-1} \left\{ \eta_{\min} + (\eta_{\max} - \eta_{\min}) \left( 1 - \frac{t - T_w}{T - T_w} \right)^p \right\} \\ &\geq \eta_{\min}(T - T_w) + (\eta_{\max} - \eta_{\min}) \frac{T - T_w}{p + 1} = \frac{\eta_{\max} + p\eta_{\min}}{p + 1} (T - T_w) \end{aligned}$$

and

$$\begin{aligned} \sum_{t=T_w}^{T-1} \frac{\eta_t^2}{b_t} &\leq \sum_{t=T_w}^{T-1} \frac{1}{b_t} \left( \eta_{\min}^2 + (\eta_{\max} - \eta_{\min})^2 \left( 1 - \frac{t - T_w}{T - T_w} \right)^{2p} + 2\eta_{\min}(\eta_{\max} - \eta_{\min}) \left( 1 - \frac{t - T_w}{T - T_w} \right)^p \right) \\ &\leq \sum_{t=T_w}^{T-1} \frac{\eta_{\max}^2}{b_t} \leq \frac{\eta_{\max}^2 \gamma K}{(\gamma - 1)b_0}, \end{aligned}$$

which yields

$$\min_{t \in \{T_w, \dots, T-1\}} \mathbb{E}[\|\text{grad} f(x_t)\|_{x_t}^2] \leq \frac{p + 1}{(2 - L_r \eta_{\max})(\eta_{\max} + p\eta_{\min})} \left( 2(f(x_0) - f^*) + \frac{L_r \eta_{\max}^2 K \gamma \sigma^2}{\gamma - 1} \frac{1}{b_0} \right) \frac{1}{T - T_w}.$$

Next we consider a polynomial growth BS and a warm-up LR.

**[LR Decay Part: Constant (1)]**

We have

$$\sum_{t=T_w}^{T-1} \eta_t = \sum_{t=T_w}^{T-1} \eta_{\max} = \eta_{\max}(T - T_w)$$

and

$$\sum_{t=T_w}^{T-1} \frac{\eta_t^2}{b_t} = \sum_{t=l_w K'}^{(\alpha+1)l-1} \frac{\eta_{\max}^2}{b_t} + \sum_{(\alpha+1)l}^{T-1} \frac{\eta_{\max}^2}{b_t} = \sum_{t=T_w}^{T-1} \frac{\eta_{\max}^2}{b_t} \leq \frac{\eta_{\max}^2 K}{\underline{a}^{[c]}} \sum_{m=0}^{M-1} \frac{1}{(m+1)^c} \leq \frac{\eta_{\max}^2 K}{\underline{a}^{[c]}} \sum_{m=1}^{\infty} \frac{1}{m^c} = \frac{\eta_{\max}^2 K \zeta(c)}{\underline{a}^{[c]}},$$

which yields

$$\min_{t \in \{T_w, \dots, T-1\}} \mathbb{E}[\|\text{grad} f(x_t)\|_{x_t}^2] \leq \frac{1}{(2 - L_r \eta_{\max})\eta_{\max}} \left( 2(f(x_0) - f^*) + \frac{L_r \eta_{\max}^2 K \zeta(c) \sigma^2}{\underline{a}^{[c]}} \right) \frac{1}{T - T_w}.$$

**[LR Decay Part: Diminishing (2)]**

Similarly, we have

$$\sum_{t=T_w}^{T-1} \eta_t = \sum_{t=T_w}^{T-1} \frac{\eta_{\max}}{\sqrt{t+1}} \geq \eta_{\max} \int_{T_w}^T \frac{dt}{\sqrt{t+1}} = 2\eta_{\max}(\sqrt{T+1} - \sqrt{T_w+1})$$



and

$$\sum_{t=T_w}^{T-1} \frac{\eta_t^2}{b_t} = \sum_{t=T_w}^{T-1} \frac{\eta_{\max}^2}{b_t(t+1)} \leq \sum_{t=T_w}^{T-1} \frac{\eta_{\max}^2}{b_t} \leq \frac{\eta_{\max}^2 K\zeta(c)}{\underline{a}^{[c]}},$$

which yields

$$\min_{t \in \{T_w, \dots, T-1\}} \mathbb{E}[\|\text{grad} f(x_t)\|_{x_t}^2] \leq \frac{1}{2(2 - L_r \eta_{\max}) \eta_{\max}} \left( 2(f(x_0) - f^*) + \frac{L_r \eta_{\max}^2 K\zeta(c) \sigma^2}{\underline{a}^{[c]}} \right) \frac{1}{\sqrt{T+1} - \sqrt{T_w+1}}.$$

**[LR Decay Part: Cosine Annealing (3)]**

We have

$$\begin{aligned} \sum_{t=T_w}^{T-1} \eta_t &= \sum_{t=T_w}^{T-1} \left( \frac{\eta_{\max} + \eta_{\min}}{2} + \frac{\eta_{\max} - \eta_{\min}}{2} \cos \frac{t - T_w}{T - T_w} \pi \right) \\ &\geq \frac{\eta_{\max} + \eta_{\min}}{2} (T - T_w) + \frac{\eta_{\max} - \eta_{\min}}{2} \int_{T_w}^T \cos \left( \frac{t - T_w}{T - T_w} \pi \right) dt = \frac{\eta_{\max} + \eta_{\min}}{2} (T - T_w) \end{aligned}$$

and

$$\begin{aligned} \sum_{t=T_w}^{T-1} \frac{\eta_t^2}{b_t} &= \sum_{t=T_w}^{T-1} \frac{1}{b_t} \left( \frac{(\eta_{\max} + \eta_{\min})^2}{4} + \frac{\eta_{\max}^2 - \eta_{\min}^2}{2} \cos \frac{t - T_w}{T - T_w} \pi + \frac{(\eta_{\max} - \eta_{\min})^2}{4} \cos^2 \frac{t - T_w}{T - T_w} \pi \right) \\ &\leq \sum_{t=T_w}^{T-1} \frac{1}{b_t} \left( \frac{(\eta_{\max} + \eta_{\min})^2}{4} + \frac{\eta_{\max}^2 - \eta_{\min}^2}{2} + \frac{(\eta_{\max} - \eta_{\min})^2}{4} \right) = \sum_{t=T_w}^{T-1} \frac{\eta_{\max}^2}{b_t} \leq \frac{\eta_{\max}^2 K\zeta(c)}{\underline{a}^{[c]}}, \end{aligned}$$

which yields

$$\min_{t \in \{T_w, \dots, T-1\}} \mathbb{E}[\|\text{grad} f(x_t)\|_{x_t}^2] \leq \frac{2}{(2 - L_r \eta_{\max})(\eta_{\max} + \eta_{\min})} \left( 2(f(x_0) - f^*) + \frac{L_r \eta_{\max}^2 K\zeta(c) \sigma^2}{\underline{a}^{[c]}} \right) \frac{1}{T - T_w}.$$

**[LR Decay Part: Polynomial Decay (4)]**

Considering the proof of the polynomial decay LR (4) in Theorem 5, we have

$$\begin{aligned} \sum_{t=T_w}^{T-1} \eta_t &= \sum_{t=T_w}^{T-1} \left( \eta_{\min} + (\eta_{\max} - \eta_{\min}) \left( 1 - \frac{t - T_w}{T - T_w} \right)^p \right) \geq \eta_{\min} (T - T_w) + (\eta_{\max} - \eta_{\min}) \frac{T - T_w}{p + 1} \\ &= \frac{\eta_{\max} + p \eta_{\min}}{p + 1} (T - T_w) \end{aligned}$$

and

$$\begin{aligned} \sum_{t=T_w}^{T-1} \frac{\eta_t^2}{b_t} &= \sum_{t=T_w}^{T-1} \frac{1}{b_t} \left( \eta_{\min}^2 + (\eta_{\max} - \eta_{\min})^2 \left( 1 - \frac{t - T_w}{T - T_w} \right)^{2p} + 2\eta_{\min}(\eta_{\max} - \eta_{\min}) \left( 1 - \frac{t - T_w}{T - T_w} \right)^p \right) \\ &\leq \sum_{t=T_w}^{T-1} \frac{\eta_{\max}^2}{b_t} \leq \frac{\eta_{\max}^2 K\zeta(c)}{\underline{a}^{[c]}}, \end{aligned}$$

which yields

$$\min_{t \in \{T_w, \dots, T-1\}} \mathbb{E}[\|\text{grad} f(x_t)\|_{x_t}^2] \leq \frac{p + 1}{(2 - L_r \eta_{\max})(\eta_{\max} + p \eta_{\min})} \left( 2(f(x_0) - f^*) + \frac{L_r \eta_{\max}^2 K\zeta(c) \sigma^2}{\underline{a}^{[c]}} \right) \frac{1}{T - T_w}.$$

■

### D.5. Proof of Theorem 8

**Theorem 14 (Detailed Version of Theorem 8)** *We consider a constant BS ( $b_t = b > 0$ ) and warm-up LR (7) and (8) with decay parts given by (1), (2), (3), or (4) under the assumptions of Lemma 4. Then, we obtain*

- *Decay part: Diminishing (2)*

$$\min_{t \in \{T_w, \dots, T-1\}} \mathbb{E}[\|\text{grad} f(x_t)\|_{x_t}^2] \leq (Q_1 + \frac{Q_2 \sigma^2}{b} \log \frac{T}{T_w}) \frac{1}{\sqrt{T+1} - \sqrt{T_w+1}},$$

- *Decay part: Otherwise (1), (3), (4)*

$$\min_{t \in \{T_w, \dots, T-1\}} \mathbb{E}[\|\text{grad} f(x_t)\|_{x_t}^2] \leq \frac{\tilde{Q}_1}{T - T_w} + \frac{\tilde{Q}_2 \sigma^2}{b},$$

where  $Q_1, Q_2, \tilde{Q}_1$ , and  $\tilde{Q}_2$  are constants that do not depend on  $T$ .

#### Proof [LR Decay Part: Constant (1)]

Considering D.4, we have

$$\sum_{t=T_w}^{T-1} \frac{\eta_t^2}{b_t} = \frac{\eta_{\max}^2}{b} (T - T_w),$$

which yields

$$\min_{t \in \{T_w, \dots, T-1\}} \mathbb{E}[\|\text{grad} f(x_t)\|_{x_t}^2] \leq \frac{2(f(x_0) - f^*)}{(2 - L_r \eta_{\max}) \eta_{\max}} \frac{1}{T - T_w} + \frac{L_r \eta_{\max}}{2 - L_r \eta_{\max}} \frac{\sigma^2}{b}.$$

#### [LR Decay Part: Diminishing (2)]

Similarly, we have

$$\sum_{t=T_w}^{T-1} \frac{\eta_t^2}{b_t} = \frac{\eta_{\max}^2}{b} \sum_{t=T_w}^{T-1} \frac{1}{t+1} \leq \frac{\eta_{\max}^2}{b} \left(1 + \int_{T_w}^T \frac{dt}{t}\right) = \frac{\eta_{\max}^2}{b} + \frac{\eta_{\max}^2}{b} \log \frac{T}{T_w},$$

which yields

$$\min_{t \in \{T_w, \dots, T-1\}} \mathbb{E}[\|\text{grad} f(x_t)\|_{x_t}^2] \leq \frac{f(x_0) - f^*}{(2 - L_r \eta_{\max}) \eta_{\max}} \frac{1}{\sqrt{T+1} - \sqrt{T_w+1}} + \frac{L_r \eta_{\max}}{2(2 - L_r \eta_{\max})} \frac{\sigma^2}{b} \frac{1 + \log \frac{T}{T_w}}{\sqrt{T+1} - \sqrt{T_w+1}}.$$

#### [LR Decay Part: Cosine Annealing (3)]

Doing the same with D.4, we have

$$\begin{aligned} \sum_{t=T_w}^{T-1} \frac{\eta_t^2}{b_t} &= \frac{1}{b} \sum_{t=T_w}^{T-1} \left( \frac{(\eta_{\max} + \eta_{\min})^2}{4} + \frac{\eta_{\max}^2 - \eta_{\min}^2}{2} \cos \frac{t - T_w}{T - T_w} \pi + \frac{(\eta_{\max} - \eta_{\min})^2}{4} \cos^2 \frac{t - T_w}{T - T_w} \pi \right) \\ &\leq \frac{1}{b} \sum_{t=T_w}^{T-1} \left( \frac{(\eta_{\max} + \eta_{\min})^2}{4} + \frac{\eta_{\max}^2 - \eta_{\min}^2}{2} + \frac{(\eta_{\max} - \eta_{\min})^2}{4} \right) = \frac{\eta_{\max}^2}{b} (T - T_w), \end{aligned}$$

which yields

$$\min_{t \in \{T_w, \dots, T-1\}} \mathbb{E}[\|\text{grad} f(x_t)\|_{x_t}^2] \leq \frac{4(f(x_0) - f^*)}{(2 - L_r \eta_{\max})(\eta_{\max} + \eta_{\min})} \frac{1}{T - T_w} + \frac{2L_r \eta_{\max}^2}{(2 - L_r \eta_{\max})(\eta_{\max} + \eta_{\min})} \frac{\sigma^2}{b}.$$

**[LR Decay Part: Polynomial Decay (4)]**

Similarly, we have

$$\begin{aligned} \sum_{t=T_w}^{T-1} \frac{\eta_t^2}{b_t} &= \frac{1}{b} \sum_{t=T_w}^{T-1} \left( \eta_{\min}^2 + (\eta_{\max}^2 - \eta_{\min}^2) \left(1 - \frac{t - T_w}{T - T_w}\right)^{2p} + 2\eta_{\min}(\eta_{\max} - \eta_{\min}) \left(1 - \frac{t - T_w}{T - T_w}\right)^p \right) \\ &\leq \sum_{t=T_w}^{T-1} \frac{\eta_{\max}^2}{b_t} = \frac{\eta_{\max}^2}{b} (T - T_w), \end{aligned}$$

which yields

$$\min_{t \in \{T_w, \dots, T-1\}} \mathbb{E}[\|\text{grad} f(x_t)\|_{x_t}^2] \leq \frac{2(f(x_0) - f^*)(p+1)}{(2 - L_r \eta_{\max})(\eta_{\max} + p\eta_{\min})} \frac{1}{T - T_w} + \frac{2L_r(p+1)\eta_{\max}^2}{(2 - L_r \eta_{\max})(\eta_{\max} + p\eta_{\min})} \frac{\sigma^2}{b}.$$

■

## Appendix E. Objective Function Values in Sections 4.1 and 4.2

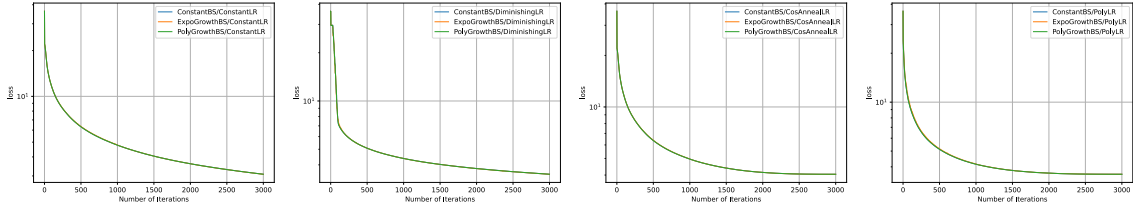


Figure 9: Objective function value (loss) versus number of iterations for LR (1), (2), (3), and (4) in order from left to right on COIL100 dataset (PCA).

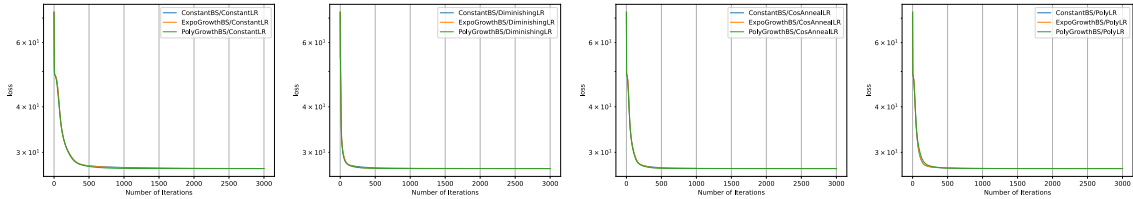


Figure 10: Objective function value (loss) versus number of iterations for LR (1), (2), (3), and (4) in order from left to right on MNIST dataset (PCA).

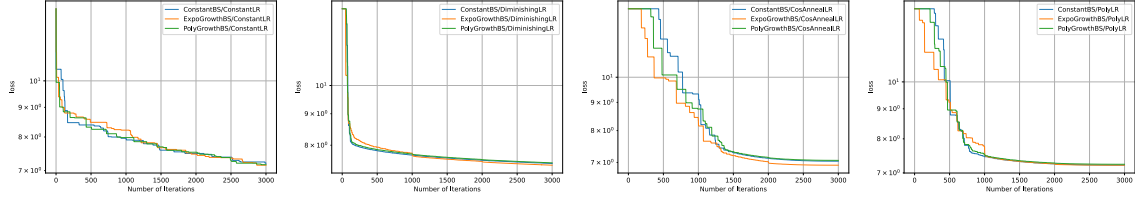


Figure 11: Objective function value (loss) versus number of iterations for LR (1), (2), (3), and (4) in order from left to right on MovieLens-1M dataset (LRMC).

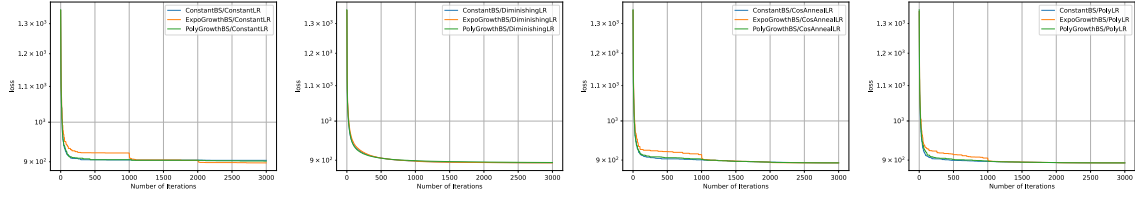


Figure 12: Objective function value (loss) versus number of iterations for LR (1), (2), (3), and (4) in order from left to right on Jester dataset (LRMC).

The performance in terms of the objective function value versus the number of iterations for LR (1), (2), (3), and (4) on the COIL100, MNIST, MovieLens-1M, and Jester datasets are shown in Figures 9, 10, 11, and 12, respectively. Although the differences in the objective function values are small (one possible reason for this is that the objective function may be flat around the optimal solution), the performance with an increasing BS was equal to or better than that with a constant BS. A more detailed discussion of this hypothesis is provided in Section 4.3.

## Appendix F. Convergence Criteria in Table 1

In previous studies and in our work, three types of convergence criteria were used:

1. the objective function value  $\mathbb{E}[f(x_T) - f^*]$ ,
2. the gradient value  $\lim_{t \rightarrow \infty} \mathbb{E}[\text{grad} f(x_t)] = 0$  or  $\|G_T\|^2$ ,
3. (extended) Riemannian distance between  $x_T$  and an optimal solution  $x^*$ :  $\mathbb{E}[\|\Delta_T\|] := \mathbb{E}[\|R_{x^*}^{-1}(x_T)\|]$ .

Criterion (3) applies because  $\mathbb{E}[\|\Delta_T\|] = \mathbb{E}[d(x_T, x^*)]$  holds when  $R = \text{Exp}$ .

## Appendix G. Additional Numerical Results

This section presents the numerical results for a warm-up LR. We used a constant BS, an exponential growth BS (5), a polynomial growth BS (6), and a warm-up LR with an

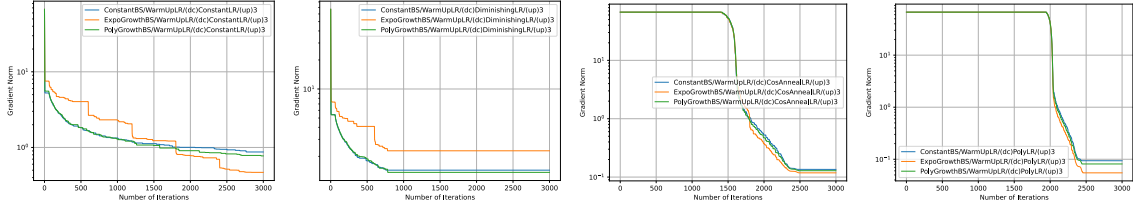


Figure 13: Norm of the gradient of the objective function versus number of iterations for warm-up LR that have an increasing part with three increments on COIL100 dataset (PCA).

increasing part (exponential growth LR) and a decaying part (either a constant LR (1), a diminishing LR (2), a cosine annealing LR (3), or a polynomial decay LR (4)). We set  $K' = 200$ . When we used an exponential growth LR, a polynomial growth BS was not required to satisfy  $\delta^{2l} < \gamma$ , whereas an exponential growth BS was required to satisfy it (see Section 3.3). For comparison, even when using a polynomial growth BS, we adopted a setting that satisfies this condition. Furthermore, we chose  $\eta_{\max}$  (the initial value of decaying part) from  $\{0.5, 0.05, 0.005\}$ . When using an exponential BS, we set the initial BS  $b_0 := 3^4$  in **Cases A, B** and  $b_0 := 3^3$  in **Cases C, D**. When using the other BSs, we set the same initial batch size as in Section 4.

From the definition of an exponential growth LR (7), we can represent the part after warm-up as  $\eta_{T_w-1} = \eta_0 \delta^{l_w}$ . Consequently, we chose hyperparameters  $(l, l_w, \gamma, \eta_{\max}, \eta_0)$  satisfying  $\eta_0 \delta^{l_w} = \eta_{\max}$  and  $\delta^{2l} < \gamma$ ; i.e.,  $\delta = (\frac{\eta_{\max}}{\eta_0})^{\frac{1}{l_w}} < \gamma^{\frac{1}{2l}}$ . Hence, when  $l = l_w = 3$  and  $\gamma = 3.0$ , we set  $\eta_{\max} = 0.5, 0.05, 0.005$  and  $\eta_0 = \frac{5}{17}, \frac{5}{170}, \frac{5}{1700}$ , respectively. In this setting,  $\delta = \sqrt[3]{\frac{\eta_{\max}}{\eta_0}} = \sqrt[3]{1.7} < \sqrt[6]{3} = \sqrt[6]{\gamma} = \gamma^{\frac{1}{2l}}$  holds. Similarly, when  $l = 3, l_w = 8, \gamma = 3$ , we set  $\eta_{\max} = 0.5, 0.05, 0.005$ ,  $\eta_0 = \frac{1}{8}, \frac{1}{80}, \frac{1}{800}$ , respectively. In this setting,  $\delta = \sqrt[8]{\frac{\eta_{\max}}{\eta_0}} = \sqrt[8]{4} < \sqrt[6]{3} = \sqrt[6]{\gamma} = \gamma^{\frac{1}{2l}}$  holds. Because we terminated RSGD after the 3000th iteration and set  $K' = 200$ , we used  $l = l_w = 3$  (resp.  $l = 3, l_w = 8$ ). In the first setting, the batch size increases five times and the learning rate three times; in the second, the batch size increases five times and the learning rate eight times.

### G.1. Principle Components Analysis

**[Case A]**  $l = l_w = 3$

Figures 13 and 15 plot performance in terms of the gradient norm of the objective function versus the number of iterations for a warm-up LR with decay parts given by (1), (2), (3), and (4) on the COIL100 and MNIST datasets, respectively. Figures 14 and 16 plot performance in terms of the objective function value versus the number of iterations for a warm-up LR with decay parts given by (1), (2), (3), and (4) for the COIL100 and MNIST datasets, respectively.

**[Case B]**  $l = 3, l_w = 8$

Figures 17 and 19 plot performance in terms of the gradient norm of the objective function versus the number of iterations for a warm-up LR with decay parts given by (1), (2), (3), and

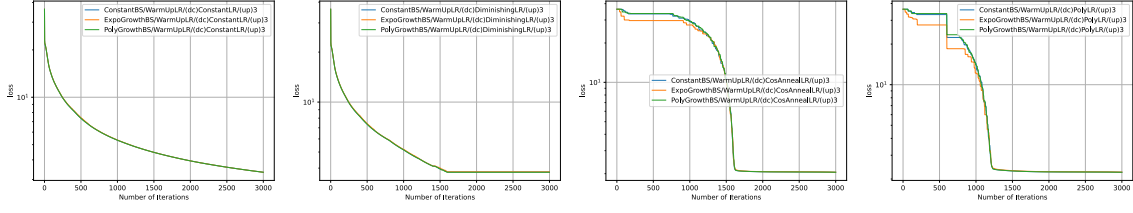


Figure 14: Objective function value (loss) versus number of iterations for warm-up LR schedules that have an increasing part with three increments on COIL100 dataset (PCA).

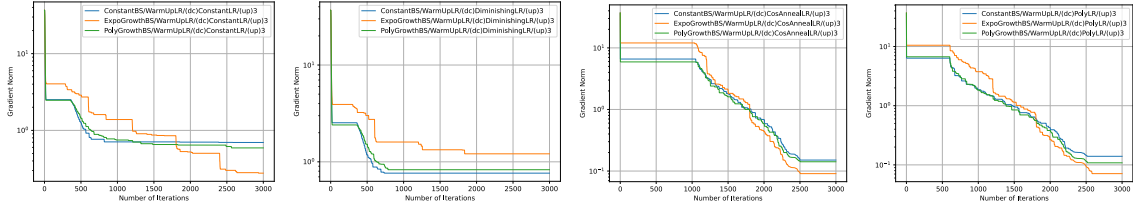


Figure 15: Norm of the gradient of the objective function versus number of iterations for warm-up LR schedules that have an increasing part with three increments on MNIST dataset (PCA).

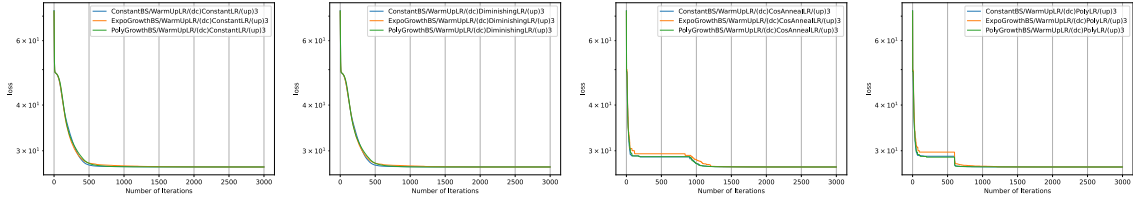


Figure 16: Objective function value (loss) versus number of iterations for warm-up LR schedules that have an increasing part with three increments on MNIST dataset (PCA).

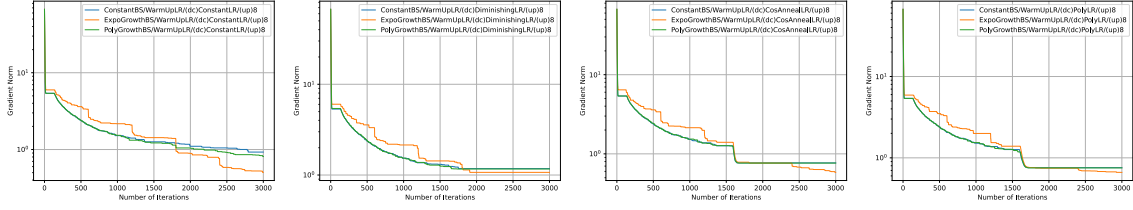


Figure 17: Norm of the gradient of the objective function versus number of iterations for warm-up LR schedules that have an increasing part with eight increments on COIL100 dataset (PCA).

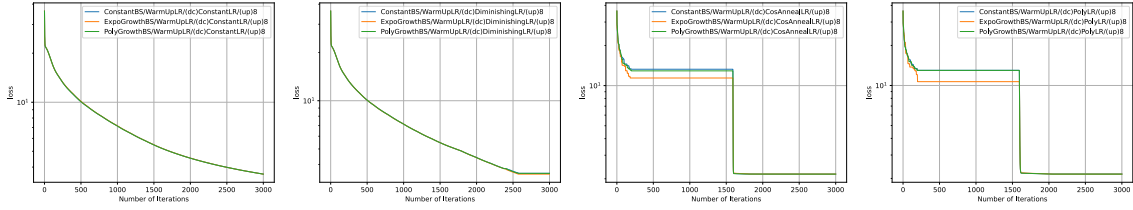


Figure 18: Objective function value (loss) versus number of iterations for warm-up LR schedules that have an increasing part with eight increments on COIL100 dataset (PCA).

(4) for the COIL100 and MNIST datasets, respectively. Figures 18 and 20 plot performance in terms of the objective function value versus the number of iterations for a warm-up LR with decay parts given by (1), (2), (3), and (4) for the COIL100 and MNIST datasets, respectively.

## G.2. Low-rank Matrix Completion

[Case C]  $l = l_w = 3$

Figures 21 and 23 plot performance in terms of the gradient norm of the objective function versus the number of iterations for a warm-up LR with decay parts given by (1), (2), (3),

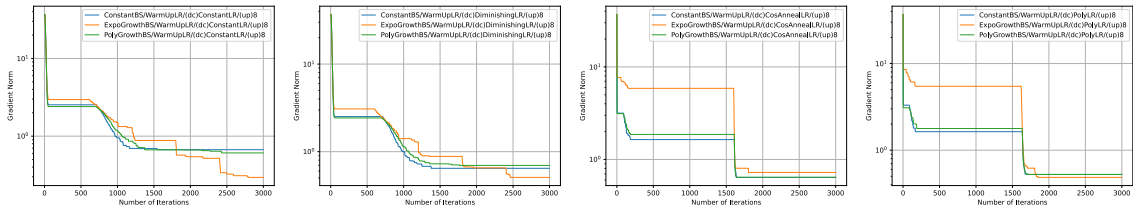


Figure 19: Norm of the gradient of the objective function versus number of iterations for warm-up LR schedules that have an increasing part with eight increments on MNIST dataset (PCA).

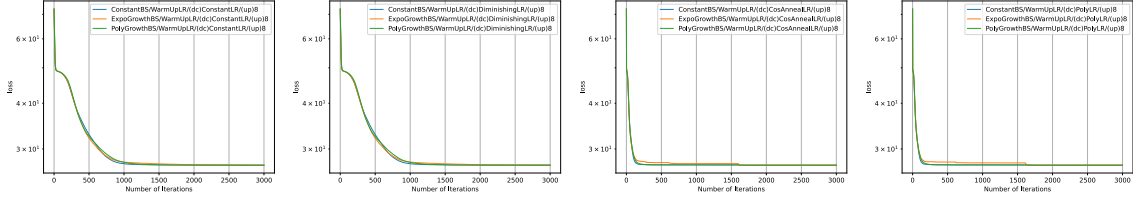


Figure 20: Objective function value (loss) versus number of iterations for warm-up LR schedules that have an increasing part with eight increments on MNIST dataset (PCA).

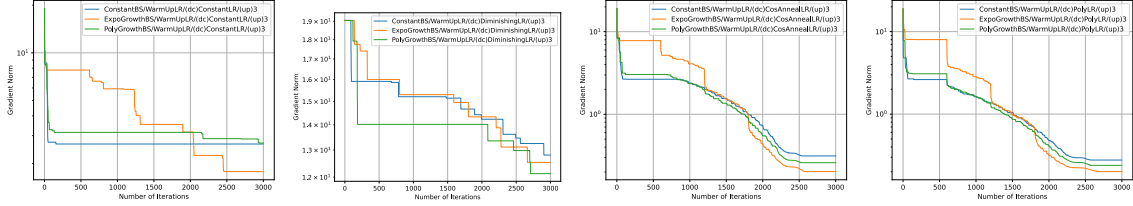


Figure 21: Norm of the gradient of the objective function versus number of iterations for warm-up LR schedules that have an increasing part with three increments on MovieLens-1M dataset (LRMC).

and (4) on the MovieLens-1M and Jester datasets, respectively. Figures 21 and 24 plot performance in terms of the objective function value versus the number of iterations for a warm-up LR with decay parts given by (1), (2), (3), and (4) on the MovieLens-1M and Jester datasets, respectively.

**[Case D]**  $l = 3, l_w = 8$

Figures 25 and 27 plot performance in terms of the gradient norm of the objective function versus the number of iterations for a warm-up LR with decay parts given by (1), (2), (3), and (4) on the MovieLens-1M and Jester datasets, respectively. Figures 25 and 28 plot performance in terms of the objective function value versus the number of iterations for

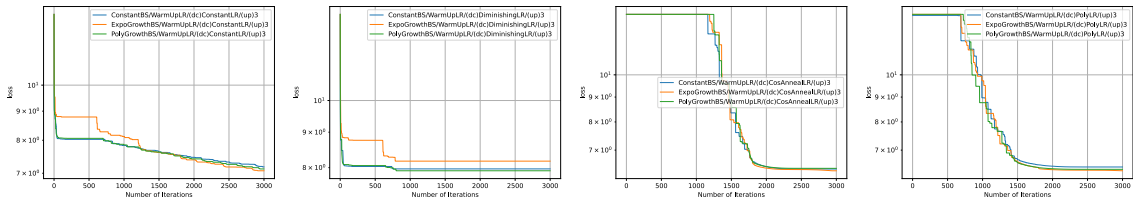


Figure 22: Objective function value (loss) versus number of iterations for warm-up LR schedules that have an increasing part with three increments on MovieLens-1M dataset (LRMC).



## FASTER CONVERGENCE OF RSGD WITH INCREASING BS

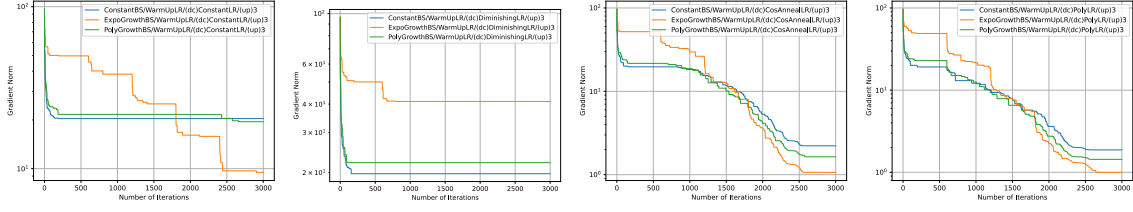


Figure 23: Norm of the gradient of the objective function versus number of iterations for warm-up LR schedules that have an increasing part with three increments on Jester dataset (LRMC).

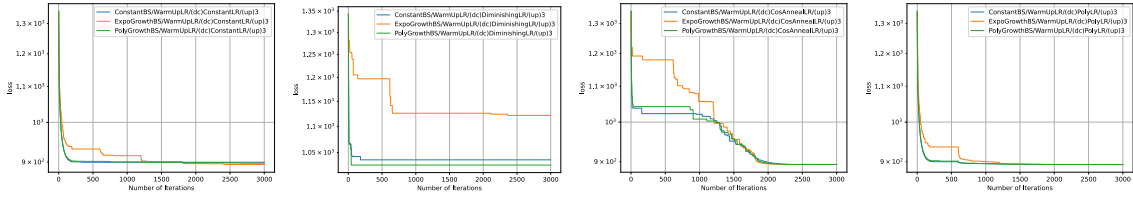


Figure 24: Objective function value (loss) versus number of iterations for warm-up LR schedules that have an increasing part with three increments on Jester dataset (LRMC).

a warm-up LR with decay parts given by (1), (2), (3), and (4) on the MovieLens-1M and Jester datasets, respectively.

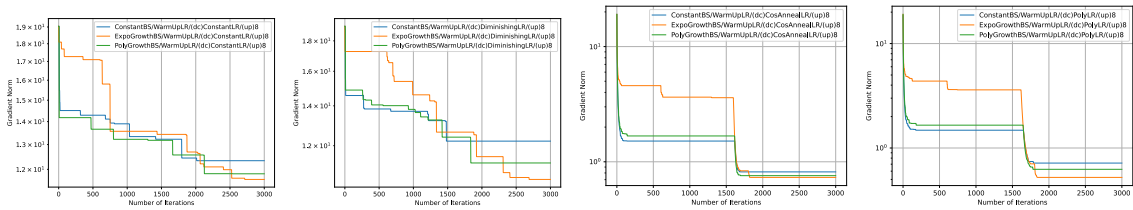


Figure 25: Norm of the gradient of the objective function versus number of iterations for warm-up LR schedules that have an increasing part with eight increments on MovieLens-1M dataset (LRMC).

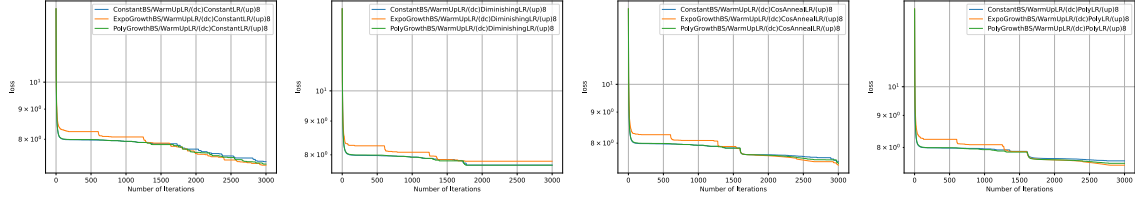


Figure 26: Objective function value (loss) versus number of iterations for warm-up LR schedules that have an increasing part with eight increments on MovieLens-1M dataset (LRMC).

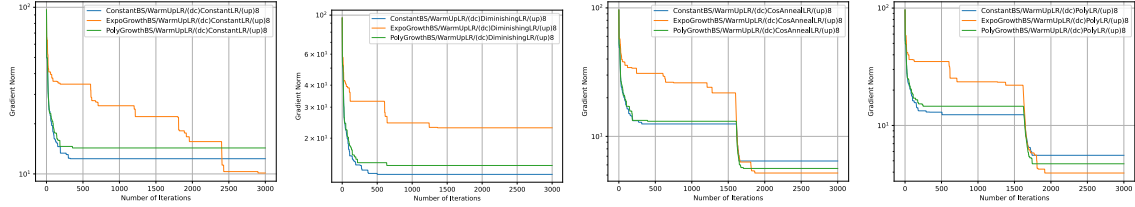


Figure 27: Norm of the gradient of the objective function versus number of iterations for warm-up LR schedules that have an increasing part with eight increments on Jester dataset (LRMC).

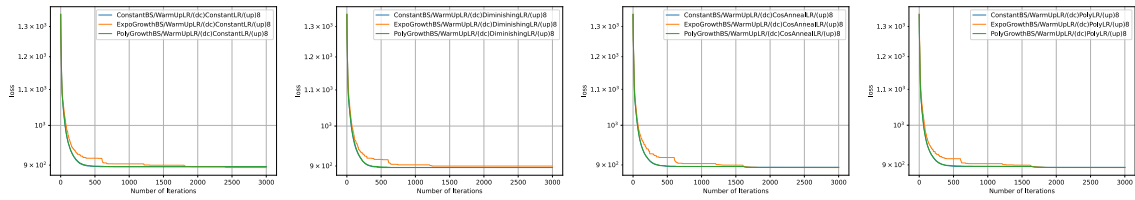


Figure 28: Objective function value (loss) versus number of iterations for warm-up LR schedules that have an increasing part with eight increments on Jester dataset (LRMC).