Convergence of Adafactor under Non-Convex Smooth Stochastic Optimization

Anonymous Author(s)
Affiliation
Address
email

Abstract

As model sizes in deep learning continue to expand, memory-efficient optimizers are increasingly critical to manage the substantial memory demands of popular algorithms like Adam and AdamW. Among these, Adafactor has emerged as one of the widely adopted choices for training deep learning tasks, particularly large language models. However, despite its practical success, there is limited theoretical analysis on Adafactor's convergence. This paper presents a comprehensive analysis on Adafactor in a non-convex smooth setting, demonstrating its convergence to find a stationary point at a rate of $\tilde{\mathcal{O}}(1/\sqrt{T})$. We find that the default hyper-parameter setting results in a sub-optimal rate in our framework, and propose an alternative setting that could theoretically achieve optimal convergence rate. This finding is further supported by some experimental results. We also prove that Adafactor with a suitable time-varying clipping threshold could also converge, achieving performance in experiments comparable to that of the standard constant setting.

1 Introduction

The adaptive gradient-based methods, such as the well-known AdaGrad [9, 29], RMSProp [30], Adadelta [35], Adam [15] and AdamW [22], have become the preferred approaches in solving the following unconstrained stochastic optimization problem in deep learning fields:

$$\min_{\boldsymbol{X}\in\mathbb{R}^{n\times m}} f(\boldsymbol{X}) = \mathbb{E}_{\boldsymbol{Z}\in\mathcal{P}}[l(\boldsymbol{X};\boldsymbol{Z})], \tag{1}$$

where the object function f is non-convex and $\mathcal{P}$ denotes a probability distribution. During the training process, these adaptive methods require to store the historical gradients' information so as to adaptively tune their step-sizes. For example, both Adam and AdamW maintain the exponential average of gradients and squared gradients, and AdaGrad stores the cumulative of squared gradients. Despite their effectiveness, these algorithms pose substantial memory challenges for GPUs to save these additional gradients' information, especially when training large language models (LLMs), such as GPT-3 [5], which contains over 175 billion parameters.

To address memory constraints, several memory-efficient optimization algorithms have been developed, e.g., [26, 1, 23, 17]. One of the most popular optimizers is Adafactor [26] which employs a rank-1 matrix factorization to approximate the second moment matrix in Adam. For an $n \times m$ weight matrices, this technique reduces memory usage from $\mathcal{O}(mn)$ to $\mathcal{O}(m+n)$ by only tracking the moving averages of the row and column sums of the squared gradients matrix. Additionally, Adafactor eliminates the first-order momentum used in Adam and incorporates update clipping to enhance training stability.

The empirical results reveal that Adafactor achieves comparable performance to Adam in training Transformer models [26] . In real applications, several LLMs including PaLM [6] and T5 [24] have

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

applied Adafactor as their main optimizers [38]. In spite of Adafactor's widely usage, there is still limited understanding on its convergence in theory, especially the effect of the matrix approximation and update clipping, and the explanation for its hyper-parameter setting in experiments.

In this paper, we take a closer look on Adafactor's convergence under non-convex smooth optimization problems, considering the typical bounded gradient setting as those for AdaGrad [19, 32] and Adam [34]. We aim to provide a convergence rate for Adafactor and explain the influence of the hyper-parameters for the convergence speed. We also prove in theory why the default parameter setting is effective in practical scenarios. The analysis to Adafactor is non-trivial compared to other adaptive methods such as AdaGrad and Adam due to the unique matrix factorization and update clipping mechanisms. Based on a new proxy step-size construction and some new compositions as well as estimations, we analyze the additional error terms in the Descent Lemma introduced by the matrix approximation and update clipping. Our main contributions are summarized as follows.

Contributions

- We provide a convergence analysis for the full-batch Adafactor considering bounded gradients and a broader range of parameter setting which covers the default one in [26]. The result shows that Adafactor could converge to find a stationary point with a rate of $\tilde{\mathcal{O}}(1/\sqrt{T})$ where T denotes the total iteration number.
- We further investigate the more realistic stochastic Adafactor. It's found that a simple variant of Adafactor, which drops the update clipping, could attain the best convergence rate of $\tilde{\mathcal{O}}(1/\sqrt{T})$ when the second moment decay rate is $1 - 1/k$. We also verify that the default decay rate $1 - 1/k^{0.8}$ could lead to a sub-optimal convergence rate in our framework. To illustrate this finding, we provide some empirical results, showing that the potential best hyper-parameter setting in theory could perform better than the default one used in experiments.
- We extend our study to include a time-varying clipping threshold. Our analysis implies that with proper selections of clipping threshold and hyper-parameters, Adafactor could also achieve the best convergence rate of $\tilde{\mathcal{O}}(1/\sqrt{T})$. We also do some experiments to show that the new clipping threshold scheme achieves comparable performance and training stability to the original constant threshold setting.

The rest of the paper is organized as follows. The next section provides some most relevant works. Section 3 presents some necessary notations definitions and problem setup. Section 4 reviews Adafactor and introduces its essential mechanism. In Section 5 and Section 6, we separately provide convergence bounds for full-batch Adafactor and stochastic Adafactor without update clipping. We further discuss the hyper-parameters' dependency. In Section 7, we investigate Adafactor using a time-increasing update clipping threshold. Section 8 provides experimental results to support our theory. All the detailed proof could be found in the appendix.

2 Related work

In this paper, we mainly investigate the theoretical convergence of Adafactor. Although there is limited works on Adafactor in theory, it's necessary to briefly discuss related works on the convergence of other adaptive methods, particularly on non-convex smooth optimization. Here, we briefly list some of the most related works.

Convergence of adaptive methods Several studies address the convergence of AdaGrad in non-convex settings. For example, [19] considered a simple variant with delayed step-size, while [32] and [39] assumed bounded stochastic gradients. Other works [14, 10, 21, 3, 31, 27, 33] derived convergence bounds under more relaxed assumptions. Another line of research has investigated the convergence of Adam. For instance, [34, 7, 39, 11, 8] assumed bounded gradients. [28, 36, 31] considered more relaxed noise assumptions without relying on bounded gradients. Additionally, [18] derived convergence bounds for Adam under generalized smooth conditions.

Overall, the convergence analysis of optimizers typically starts with standard assumptions, such as bounded gradients and smooth objective functions. In subsequent studies, these assumptions are gradually relaxed to investigate the convergence properties of the optimizers under less stringent conditions.

Memory efficient algorithms As large models are increasingly used in deep learning, memory constraints have become a central issue during training. Consequently, several memory-efficient optimizers have been developed to address this challenge.

One approach to save memory involves applying matrix factorization to oeptimization algorithms. For instance, [25] used matrix factorization in the second moment estimator of gradients in Adam, similar to the concept behind Adafactor. [23] introduced CAME, a variant of Adafactor, which incorporates a confidence-guided strategy to mitigate instability caused by erroneous updates. [37] proposed Adapprox, leveraging randomized low-rank matrix approximation for Adam's second moment estimator, demonstrating superior performance and reduced memory usage compared to AdamW.

There are some other techniques to save the memory. For example, [12] relied on a "Shampoo" technique to reduce the storage requirement of full-matrix preconditioning methods. Notably, their method could be further extended to the more realistic tensor case. [1] presented a memory-saved version of AdaGrad, called SM3, by maintaining k sets gradient accumulator. They proved the convergence guarantee of SM3 on online convex optimization and the effectiveness in experiments. Recently, [17] built a 4-bit Adam using quantization techniques to compress the first and second moment estimators in Adam, also reducing memory usage.

In summary, many existing optimizers, particularly adaptive methods like AdaGrad and Adam, face memory overhead. In response, the discussed works have designed memory-efficient optimizers that aim to achieve comparable performance to these existing methods while achieving memory benefits.

3 Problem setup

To start with, we introduce some necessary notations.

Notations For any two matrices $\boldsymbol{X} = (x_{ij})_{ij}, \boldsymbol{Y} = (y_{ij})_{ij} \in \mathbb{R}^{n\times m}$, we define $\langle \boldsymbol{X}, \boldsymbol{Y}\rangle = \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}y_{ij}$. $\boldsymbol{X} \odot \boldsymbol{Y}$, $\boldsymbol{X}/\boldsymbol{Y}$ and $\sqrt{\boldsymbol{X}}$ denote the coordinate-wise product, quotient and squared root respectively. $\mathbf{0}_n$ and $\mathbf{1}_n$ denote the zero and one n-dimensional vector respectively, and $\mathbf{1}_{n\times m}$ denotes the one $n \times m$-dimensional matrix. The index set $[n]$ denotes $\{1, 2, \cdots, n\}$. $\|\cdot\|_F$ denotes the Frobenius norm. For a positive sequence $\{\alpha_i\}_{i\geq 1}$, we define $\sum_{i=a}^{b}\alpha_i = 0$ and $\prod_{i=a}^{b}\alpha_i = 1$ if $a > b$. The operator RMS$(\cdot)$ denotes

$$\text{RMS}(\boldsymbol{X}) = \sqrt{\frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}^2}.$$

We consider unconstrained stochastic optimization (1) over $\mathbb{R}^{n\times m}$ with the Frobenius norm. The objective function $f : \mathbb{R}^{n\times m} \to \mathbb{R}$ is differentiable. Given an $n \times m$ matrix $\boldsymbol{X}$, we assume a gradient oracle that returns a random matrix $g(\boldsymbol{X}, \boldsymbol{Z}) \in \mathbb{R}^{n\times m}$ dependent by the random sample $\boldsymbol{Z}$. The deterministic gradient of f at $\boldsymbol{X}$ is denoted by $\nabla f(\boldsymbol{X}) \in \mathbb{R}^{n\times m}$.

Assumptions We make the following standard assumptions throughout the paper.

- (**A1**) L-smoothness: For any $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{n\times m}$, $\|\nabla f(\boldsymbol{Y}) - \nabla f(\boldsymbol{X})\|_F \leq L\|\boldsymbol{Y} - \boldsymbol{X}\|_F$;
- (**A2**) Bounded below: There exists $f^* > -\infty$ such that $f(\boldsymbol{X}) \geq f^*, \forall \boldsymbol{X} \in \mathbb{R}^{n\times m}$;
- (**A3**) Unbiased estimator: The gradient oracle provides an unbiased estimator of $\nabla f(\boldsymbol{X})$, i.e., $\mathbb{E}_{\boldsymbol{Z}}[g(\boldsymbol{X}, \boldsymbol{Z})] = \nabla f(\boldsymbol{X}), \forall \boldsymbol{X} \in \mathbb{R}^{n\times m}$;
- (**A4**) Almost surely bounded stochastic gradient: for any $\boldsymbol{X} \in \mathbb{R}^{n\times m}$, $\|g(\boldsymbol{X}, \boldsymbol{Z})\|_F \leq G$, a.s..

Combining (A3) and (A4), it's easy to verify that $\|\nabla f(\boldsymbol{X})\| \leq G, \forall \boldsymbol{X} \in \mathbb{R}^{n\times m}$. Assumptions (A1)-(A3) are standard in the non-convex smooth convergence analysis. Although Assumption (A4) is a bit strong since it requires an almost surely bounded stochastic gradients instead of an expected one, it's still commonly used to derive the high probability convergence bound, see e.g., [32, 14], which is a stronger result than an expected convergence. In coordinate-wise algorithm, another standard assumption is l_∞-bounded gradient where $\|g(\boldsymbol{X}, \boldsymbol{Z})\|_\infty \leq G_\infty$, see e.g., [8]. These two types of assumption are equivalent up to dimension factors.

4 Review of Adafactor

In this section, we briefly discuss Adafactor based on the reference [26]. The pseudocode for Adafactor is presented in Algorithm 1.

Algorithm 1 Adafactor

Input: Initialization point $\boldsymbol{X}_1 \in \mathbb{R}^{n\times m}$, $\boldsymbol{R}_0 = \boldsymbol{0}_m$, $\boldsymbol{C}_0 = \boldsymbol{0}_n^\top$, relative step-sizes $\{\rho_k\}_{k\geq 1}$, decay rate $\{\beta_{2,k}\}_{k\geq 1} \in [0,1)$, regularization constants $\epsilon_1, \epsilon_2 > 0$, clipping threshold d.
for $k = 1, \cdots, T$ **do**
$\boldsymbol{G}_k = g(\boldsymbol{X}_k, \boldsymbol{Z}_k)$;
$\boldsymbol{R}_k = \beta_{2,k}\boldsymbol{R}_{k-1} + (1-\beta_{2,k})(\boldsymbol{G}_k \odot \boldsymbol{G}_k + \epsilon_1 \boldsymbol{1}_n \boldsymbol{1}_m^\top)\boldsymbol{1}_m$;
$\boldsymbol{C}_k = \beta_{2,k}\boldsymbol{C}_{k-1} + (1-\beta_{2,k})\boldsymbol{1}_n^\top(\boldsymbol{G}_k \odot \boldsymbol{G}_k + \epsilon_1 \boldsymbol{1}_n \boldsymbol{1}_m^\top)$;
$\boldsymbol{W}_k = (\boldsymbol{R}_k \boldsymbol{C}_k)/\boldsymbol{1}_n^\top \boldsymbol{R}_k$;
$\boldsymbol{U}_k = \boldsymbol{G}_k/\sqrt{\boldsymbol{W}_k}$;
$\eta_k = \max\{\epsilon_2, \text{RMS}(\boldsymbol{X}_k)\}\rho_k / \max\{1, \text{RMS}(\boldsymbol{U}_k)/d\}$;
$\boldsymbol{X}_{k+1} = \boldsymbol{X}_k - \eta_k \cdot \boldsymbol{G}_k/\sqrt{\boldsymbol{W}_k}$;
end for

Matrix factorization Adafactor could be severed as a saved-memory version of Adam. Throughout the training process, Adam maintain two $n \times m$ matrices $\boldsymbol{M}_k$ and $\boldsymbol{V}_k$ using exponential moving average update,

$$\boldsymbol{M}_k = \beta_{1,k}\boldsymbol{M}_{k-1} + (1-\beta_{1,k})\boldsymbol{G}_k, \quad \boldsymbol{V}_k = \beta_{2,k}\boldsymbol{V}_{k-1} + (1-\beta_{2,k})\boldsymbol{G}_k \odot \boldsymbol{G}_k, \tag{2}$$

where $\beta_{1,k}, \beta_{2,k} \in (0,1)$, thereby tripling the memory usage. The innovation in Adafactor lies in its method of approximating $\boldsymbol{V}_k$ by factoring it into two rank-1 matrices, specifically the row sums and column sums of $\boldsymbol{V}_k$. This approximation is guided by maintaining a minimal general Kullback-Leibler (KL) divergence as follows,

$$\min_{\boldsymbol{X}\in\mathbb{R}^n, \boldsymbol{Y}\in\mathbb{R}^{1\times m}} \sum_{i=1}^{n}\sum_{j=1}^{m} d\left((\boldsymbol{V}_k)_{ij}, (\boldsymbol{X}\boldsymbol{Y})_{ij}\right), \quad \text{s.t.} \quad (\boldsymbol{X})_i \geq 0, (\boldsymbol{Y})_j \geq 0, \forall i \in [n], j \in [m],$$

where $d(p,q) = p\log(p/q) - p + q$. The choice of KL-divergence over the more typical Frobenius norm allows for an analytical solution to be derived, specifically given by

$$\boldsymbol{X} = \boldsymbol{V}_k \boldsymbol{1}_m, \quad \boldsymbol{Y} = \boldsymbol{1}_n^\top \boldsymbol{V}_k / \left(\boldsymbol{1}_n^\top \boldsymbol{V}_k \boldsymbol{1}_m\right).$$

Therefore, Adafactor only requires to maintain two vectors $\boldsymbol{R}_k = \boldsymbol{V}_k \boldsymbol{1}_m$, $\boldsymbol{C}_k = \boldsymbol{1}_n^\top \boldsymbol{V}_k$, sufficiently reducing the memory from $2mn$ to $m+n$. Although this factorization sacrifices some information of the squared gradients, Adafactor still delivers performance comparable to Adam in many real application tasks, making it a practical choice where memory is a constraint.

Increasing decay rate In Adam, corrective terms are introduced into $\boldsymbol{M}_k$ and $\boldsymbol{V}_k$, resulting in two increasing-to-one decay rates. Theoretically, it has been demonstrated that a value closed to one for $\beta_{2,k}$ would ensure the convergence, e.g., [8, 39, 36]. Inspired by this observation, Adafactor used an increasing second moment decay rate $\beta_{2,k} = 1 - 1/k^c, c \geq 0$, and the empirical default setting is $c = 0.8$. As pointed out by [26], this setting allows for enjoying the stability of a low $\beta_{2,k}$ at the early stage of training and the insurance of convergence from a high $\beta_{2,k}$ as the run progresses. Moreover, it also leverages the bias correction.

Update clipping Adafactor modifies the update process by discarding the first-order moment $\boldsymbol{M}_k$ and instead applies an update clipping technique inside the step-size η_k. This involves dividing the root-mean-square of the update $\boldsymbol{U}_k$, denoted as $\text{RMS}(\boldsymbol{U}_k)$, when it exceeds a threshold d. This mechanism helps to calibrate the second moment estimator $\boldsymbol{W}_k$ when it's larger-than-desired $\boldsymbol{G}_k \odot \boldsymbol{G}_k$. Empirical findings in [26] indicated that implementing update clipping leads to significant performance improvements when the warm-up technique is not used.

Relative step-sizes Adafactor incorporates a step-size proportional to scale of $\boldsymbol{X}_k$, denoted by $\text{RMS}(\boldsymbol{X}_k)$, which is shown in experiments more resilient to the more naive parameter initialization and scaling schemes [26].

5 Convergence result for full-batch Adafactor

We first provide the convergence bound for full-batch Adafactor. At each iteration, full-batch Adafactor obtains the deterministic gradient $\nabla f(\boldsymbol{X}_k)$ and then updates $\boldsymbol{R}_k, \boldsymbol{C}_k$ using $\nabla f(\boldsymbol{X}_k)$ instead of $\boldsymbol{G}_k$ in Algorithm 1.

Theorem 5.1. *Let $\{\boldsymbol{X}_k\}_{k\geq 1}$ be generated by Algorithm 1 with $g(\boldsymbol{X}_k, \boldsymbol{Z}_k) = \nabla f(\boldsymbol{X}_k), \forall k \geq 1$. If Assumptions (A1) and (A2) hold, $\|\nabla f(\boldsymbol{X}_k)\|_F \leq G, \forall k \geq 1$, $\beta_{2,1} = \frac{1}{2}$ and*

$$\rho_k = \rho_0/\sqrt{k}, \quad 0 < \beta_{2,k} < 1, \quad \forall k \geq 1, \tag{3}$$

for some positive constant ρ_0, then for any $T \geq 1$,

$$\min_{k\in[T]} \|\nabla f(\boldsymbol{X}_k)\|_F^2 \leq \mathcal{O}\left(\frac{\log T}{\sqrt{T}}\right). \tag{4}$$

The result indicates that full-batch Adafactor could find a stationary point at a rate of $\mathcal{O}(\log T/\sqrt{T})$ under the non-convex smooth case, which is similar to gradient descent but with a sub-optimal rate compared to $\mathcal{O}(1/T)$ [4]. The hyper-parameter setting in (3) only requires $\beta_{2,k} \in (0, 1)$, denoting a much wider range including the default one which requires $\beta_{2,k}$ to increase to one. The detailed version for the above result can be found in Theorem A.1 from the appendix.

6 Stochastic Adafactor without update clipping

In the stochastic case, we start from the simple scenario of

$$\eta_k = \max\{\epsilon_2, \text{RMS}(\boldsymbol{X}_k)\}\rho_k \tag{5}$$

without considering the update clipping $1/\max\{1, \text{RMS}(\boldsymbol{U}_k)/d)\}$ in Algorithm 1, where the main reasons are as follows.

- As pointed out in the experiments from [26], Adafactor's performance shows little difference with and without update clipping when implementing learning rate warm-up. Since the warm-up technique is a popular method in deep learning [38], it's reasonable to drop the update clipping.
- In stochastic Adafactor, the correlation between $\boldsymbol{G}_k$ and η_k would be more complex if the update clipping is involved. The proof would be simpler when dropping the update clipping, which could help to better understand the analysis for Adafactor.

We now present the probabilistic convergence bound for Adafactor without update clipping as follows, where we summarize different convergence rate with respect to the factor c from $\beta_{2,k} = 1 - 1/k^c, c \in [1/2, 1]$.

Theorem 6.1. *Let $\{\boldsymbol{X}_k\}_{k\geq 1}$ be generated by Algorithm 1 without update clipping where η_k is given by (5) for each $k \geq 1$. If Assumptions (A1)-(A4) hold, and*

$$\begin{aligned} &\beta_{2,1} = 1/2, \quad \rho_1 = \rho_0, \\ &\beta_{2,k} = 1 - 1/k^c, \quad \rho_k = \rho_0/\sqrt{k}, \quad \forall k \geq 2, \end{aligned} \tag{6}$$

for some constants $1/2 \leq c \leq 1, \rho_0 > 0$, then for any $T \geq 1, \delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\min_{k\in[T]} \|\nabla f(\boldsymbol{X}_k)\|_F^2 \leq \mathcal{O}\left(\frac{1}{T^{c-1/2}} \log\left(\frac{T}{\delta}\right)\right).$$

The above result indicates that with appropriate hyper-parameters, Adafactor without update clipping could approximately find a stationary point. When the decay rate $\beta_{2,k}$ is $1 - 1/k$, the convergence rate could attain to $\mathcal{O}(\log T/\sqrt{T})$, matching the rate of stochastic gradient descent [4] and the lower rate in [2] up to only a logarithm factor. The hyper-parameter setting in (6) covers the experimental default setting where $c = 0.8$. The result shows a sub-optimal rate of $\mathcal{O}(\log T/T^{0.3})$ under the default setting. This finding is further complemented by the coming numerical experiments in Section 8. The detailed version of the above results can be found in Theorem B.1 from the appendix.

6.1 Discussion of the hyper-parameter dependency

In this section, we discuss the dependency of several important hyper-parameters in Theorem 6.1 and the detailed version in Theorem B.1 in the appendix. It's worthy to mention that the dominated order in our convergence bound is determined by the total iteration number T, whereas other hyper-parameters could be regarded as constants. However, we hope to improve the dependency of these hyper-parameters as much as possible to make the convergence bound tight.

Discussion of c and the optimal rate The convergence bound in Theorem 6.1 reveals that when $c = 1, \beta_{2,k} = 1 - 1/k$ and $\rho_k = \rho_0/\sqrt{k}$, the convergence rate attains the optimal rate matching the lower bound. In addition, when c is closed to $1/2$, the convergence rate deteriorates. This phenomenon somehow explains that a small decay rate $\beta_{2,k}$ (c is low) may harm the convergence speed, as $\beta_{2,k}$ should be closed enough to 1 to ensure the convergence, which has been pointed out similarly for Adam in [8, 39, 36].

The theoretical best parameter setting remains a small gap to the default one of $c = 0.8$. To verify our theoretical finding, we provide some empirical evidence in Section 8, showing that $\beta_{2,k} = 1 - 1/k$ performs even better than the default one and the performance would be better when c increases from $1/2$ to 1.

Dependency to mn It's clear to see that the convergence bounds in Theorem A.1 and Theorem B.1 are free of the curse of the dimension factor mn as mn only appears on the denominator in each coefficient. We think that solving the curse of dimension is vital since the applied range for Adafactor includes many deep learning tasks where mn are comparable large to T.

Dependency to ϵ_1, ϵ_2 The convergence bounds in (37) and (39) from Theorem B.1 has a dependency of $\mathcal{O}(\epsilon_1^{-1}\log(1/\epsilon_1))$ on ϵ_1.[1] Although the polynomial dependency to ϵ_1 is a bit worse since ϵ_1 ususally takes a small value in experiments, e.g., the default setting is 10^{-30}, it's still common in some theoretical convergence results, e.g., [34, 18]. We also perform some experiments to show that a relatively large ϵ_1, roughly 10^{-3}, makes no observable effect on the performance. Thereby, ϵ_1 could be regarded as a constant in comparison to T and the influence brought by $1/\epsilon_1$ could be somehow acceptable.

Since the default value of ϵ_2 is 10^{-3} in experiments, it could also be regarded as a constant compared to T. Therefore, the dependency $\mathcal{O}(1/\epsilon_2)$ on ϵ_2 shows little effect on convergence bounds given the sufficiently large T.

Dependency to the scale of parameters. The convergence bounds in Theorem B.1 contain a $\mathcal{O}(\Theta_{\max})$ factor where $\Theta_{\max}$ denotes the maximum values of $\|\boldsymbol{X}_k\|_\infty$ along the training process. It's reasonable to assume that $\Theta_{\max} \le G_0$ for a comparable large constant G_0 in practice.

7 Convergence of Adafactor with update clipping

In this section, we take a closer look on the comprehensive Adafactor with both matrix factorization and update clipping. We slightly change the update clipping threshold d in Algorithm 1 to a time-varying threshold d_k. The step-size η_k then becomes

$$\eta_k = \frac{\max\{\epsilon_2, \text{RMS}(\boldsymbol{X}_k)\}\rho_k}{\max\{1, \text{RMS}(\boldsymbol{U}_k)/d_k\}}. \tag{7}$$

Then, we present the following convergence bound.

Theorem 7.1. *Let $\{\boldsymbol{X}_k\}_{k\ge1}$ be generated by Algorithm 1 with η_k given by* (7) *for each $k \ge 1$. If Assumptions (A1)-(A4) hold, and*

$$\begin{aligned} &d_1 = 1, \quad \beta_{2,1} = 1/2, \quad \rho_1 = \rho_0, \\ &d_k = k^{\frac{c}{2(\alpha-1)}}, \quad \beta_{2,k} = 1 - 1/k^c, \quad \rho_k = \rho_0/\sqrt{k}, \quad \forall k \ge 2, \end{aligned} \tag{8}$$

[1] The detailed calculation could be found in (45) and (46) in the appendix.

for some constants $\alpha > 1, 1/2 \leq c \leq 1, \rho_0 > 0$*, then for any* $T \geq 1, \delta \in (0,1)$*, with probability at least* $1-\delta$*,*

$$\min_{k\in[T]} \|\nabla f(\boldsymbol{X}_k)\|_F^2 \leq \mathcal{O}\left(\frac{1}{T^{c-1/2}}\log\left(\frac{T}{\delta}\right)\right).$$

Discussion of Theorem 7.1 The convergence result indicates that with a proper selection of the clipping threshold, along with the commonly used step-size ρ_k and decay rate $\beta_{2,k}$, Adafactor can find a stationary point when T is sufficiently large. The dependency of convergence bound on c remains consistent with Theorem 6.1, achieving the optimal order when $c = 1$. In addition, the convergence bound can still avoid the curse of dimension, which is shown in the detailed version Theorem C.1 from the appendix.

The additional hyper-parameter α primarily influences the dependency on ϵ_1, specifically as $\mathcal{O}\left(\epsilon_1^{-\alpha}\log(1/\epsilon_1)\right)$. Thus, our convergence bound may deteriorate as α increases, possibly due to the limitation of our proof framework. This dependency could be potentially improved to $\mathcal{O}\left(\epsilon_1^{-1}\log(1/\epsilon_1)\right)$ when mn is comparable to $1/\epsilon_1$, which is practical when implementing a large-size model.[2] In our experiments, we tested different values of α and found that suitably small values, such as $\alpha = 4, 6, 7, 8$ can lead to performance and training stability comparable to the default setting, even without implementing the warm-up technique. This finding suggests that our new threshold setting plays a similar role in enhancing training stability as the default one, which is also the main motivation of update clipping. Since ϵ_1 can be set to a relatively large value, e.g., 10^{-3}, a dependency like $\mathcal{O}(\epsilon_1^{-4}\log(1/\epsilon_1))$ is somewhat acceptable for sufficiently large T.

The time-increasing d_k provides the following intuition: As shown in [26, Figure 1], during the early stages of training, a high decay rate $\beta_{2,k}$ can cause larger-than-desired updates and training instability. Therefore, we set a low threshold d_k to ensure that the update clipping mechanism effectively calibrates these larger-than-desired updates. As training progresses, the sequences and updates become more stable, and the second moment estimator $\boldsymbol{W}_k$ becomes more accurate in estimating the squared gradients, which is also shown in [26, Figure 1]. Consequently, there is less need for update clipping, corresponding to a relatively large d_k. We have also verified through experiments that our setting can achieve performance comparable to the default setting of $d = 1$.

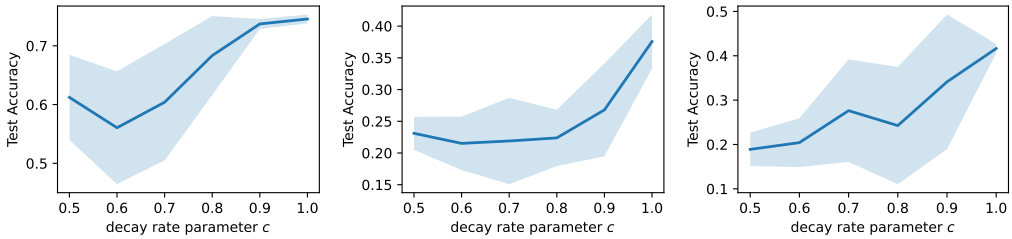
8 Experiments

In this section, we will report our experimental results based on the insights obtained in our theory. We will mainly provide the following three experiments:

- We test Adafactor without update clipping under different decay rate parameters c, aiming to demonstrate performance improvement as c increases from 0.5 to 1 with optimal performance at $c = 1$, as indicated in Theorem 6.1 and Theorem 7.1.
- We evaluate the sensitivity of Adafactor to different values of ϵ_1, particularly showing that a relatively large ϵ_1 does not significantly impact performance.
- We assess the performance of Adafactor with a time-increasing d_k setting, as described in Theorem 7.1, and compare it to the default constant setting.

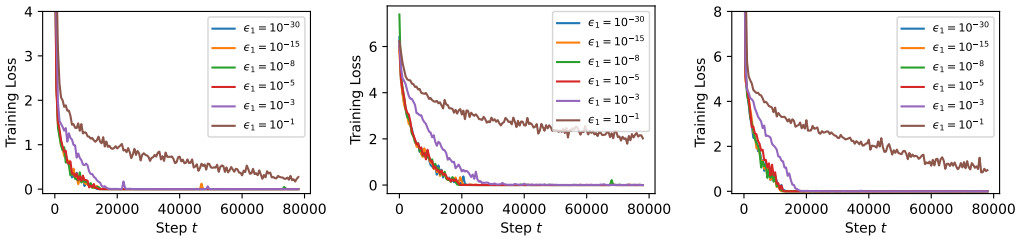
8.1 Experiment setup

In all experiments, the initialization is $\boldsymbol{R}_0 = \boldsymbol{0}_m$ and $\boldsymbol{C}_0 = \boldsymbol{0}_n^\top$. We use a learning rate with the warm-up technique as described in [26], specifically $\rho_k = \min\{10^{-6}\cdot k, 1/\sqrt{k}\}$ for all experiments unless otherwise specified. The batch size is set to 256, and the total number of epochs is 400 by default. Our models are ResNet-20 and ResNet-110 [13], and we use the CIFAR-10 and CIFAR-100 datasets [16] without any data augmentation. The experiments are conducted using the PyTorch implementation of Adafactor on a single NVIDIA GeForce RTX 4090 GPU.



(a) ResNet-20 on CIFAR-10 (b) ResNet-20 on CIFAR-100 (c) ResNet-110 on CIFAR-100

Figure 1: Average test accuracy and standard deviation (shallow blue region) under different decay rate parameters c.



(a) ResNet-20 on CIFAR-10 (b) ResNet-20 on CIFAR-100 (c) ResNet-110 on CIFAR-100

Figure 2: Training loss vs. steps using Adafactor without update clipping under different ϵ_1. The step-size η_t, decay rate $\beta_{2,k}$, and learning rate warm-up are set by default.

8.2 Report on Experiment 1

We test Adafactor without update clipping using decay rate parameter c ranging from 0.5 to 1.0 in increments of 0.05, while keeping other hyper-parameters at their default values. Each experiment is run 10 times with 100 epochs, and we plot the average test accuracy and standard deviation (shallow blue region) in Figure 1. The results indicate that $c = 1.0$ yields better performance and stability compared to $c < 1.0$ on different models and datasets, corresponding to the highest test accuracy and thinner shallow blue band. These performances show a noticeable improving trend as c increases from 0.5 to 1.0, aligning roughly with the results in Theorem 6.1.

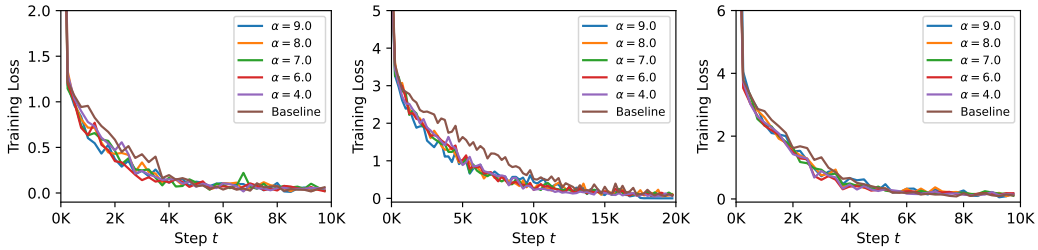
8.3 Report on Experiment 2

In the second experiment, we test Adafactor without update clipping under different ϵ_1 values. We plot the training loss against the step t on different models and datasets in Figure 2. The performance for $\epsilon_1 = 10^{-8}$ and $\epsilon_1 = 10^{-5}$ is nearly identical to that for $\epsilon_1 = 10^{-30}$. Moreover, even a larger value of 10^{-3} achieves comparable training performance, though with a slower decrease in loss. Notably, $\epsilon_1 = 10^{-3}$ requires approximately the same number of steps ($t \approx 20000$) as $\epsilon_1 = 10^{-30}$ to achieve near-zero training loss. We conclude that Adafactor is not sensitive to the choice of ϵ_1, and a relatively large ϵ_1 can still lead to convergence, making the polynomial dependency $\mathcal{O}(1/\epsilon_1)$ in our convergence bounds acceptable.

8.4 Report on Experiment 3

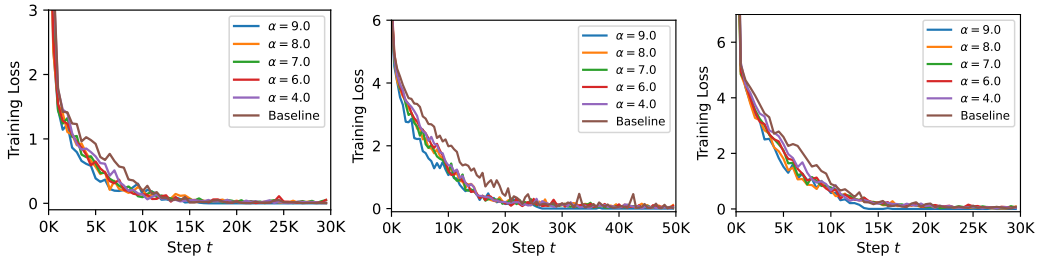
In this experiment, we explore the appropriate values of α in Theorem 7.1 to achieve performance comparable to the default setting of $d = 1$. As indicated by Theorem 7.1, a relatively small α is desirable for better dependency on ϵ_1. We train models with α set to 4, 6, 7, 8, and 9, keeping other hyper-parameters at their default values. We also train models with the default $d = 1$ setting as the baseline. We plot the training loss against the steps in Figures 3 without step-size warm-up and 4 with step-size warm-up.

[2]The detailed calculation could be found in (95) from the appendix.



(a) ResNet-20 on CIFAR-10 (b) ResNet-20 on CIFAR-100 (c) ResNet-110 on CIFAR-100

Figure 3: Training loss vs. steps on different models and datasets. We use step-size without warm-up technique and test under different α.



(a) ResNet-20 on CIFAR-10 (b) ResNet-20 on CIFAR-100 (c) ResNet-110 on CIFAR-100

Figure 4: Training loss vs. steps on different models and datasets. We use step-size with warm-up technique by default and test under different α.

The results indicate that, for these values of α, Adafactor achieves comparable or even better convergence speed compared to the default threshold (represented by "Baseline"). The comparable results to the "Baseline" in Figure 3 further suggest that the time-increasing d_k in Theorem 7.1 plays a role similar to that of the default setting, enhancing training stability even when the step-size warm-up is turned off.

9 Conclusions

In this paper, we investigate the convergence behavior of Adafactor on non-convex smooth landscapes, considering bounded stochastic gradients. We introduce a new proxy step-size to decouple the stochastic gradients from the unique adaptive step-size. Additionally, we use new estimations to control the errors introduced by matrix factorization and update clipping in Adafactor.

Our findings reveal that full-batch Adafactor is capable of finding a stationary point, requiring only a step-size $\eta_k \sim \mathcal{O}(1/\sqrt{k})$ and a second moment decay rate $\beta_{2,k} \in (0, 1)$, denoting a wide range including the default setup. In the case of stochastic Adafactor without update clipping, the convergence rate can achieve the optimal order $\tilde{\mathcal{O}}(1/\sqrt{T})$ when $\beta_{2,k} = 1 - 1/k^c, c = 1$. However, performance deteriorates as c decreases. This finding is supported by experimental results. We also explore Adafactor with a time-increasing clipping threshold and derive similar convergence results. The empirical results demonstrate that the new clipping threshold provides performance comparable to the default constant setting.

Limitations There are several limitations in our work that warrant further investigation. First, the polynomial dependency on ϵ_1 in our convergence bounds may be further improved to a better dependency, such as $\log(1/\epsilon_1)$. Second, although we provide convergence results for several variants of Adafactor and demonstrate comparable performance to the original one in experiments, the convergence bound for stochastic vanilla Adafactor remains unknown. Finally, our experimental results primarily focus on traditional deep learning tasks due to our GPU limitations. It would be beneficial to test the scalability of our theoretical results, e.g., on large language models.

References

[1] Rohan Anil, Vineet Gupta, Tomer Koren, and Yoram Singer. Memory efficient adaptive optimization. In *Advances in Neural Information Processing Systems*, 2019.

[2] Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1-2):165–214, 2023.

[3] Amit Attia and Tomer Koren. SGD with AdaGrad stepsizes: full adaptivity with high probability to unknown parameters, unbounded gradients and affine variance. In *International Conference on Machine Learning*, 2023.

[4] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[6] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

[7] Soham De, Anirbit Mukherjee, and Enayat Ullah. Convergence guarantees for RMSProp and Adam in non-convex optimization and an empirical comparison to Nesterov acceleration. *arXiv preprint arXiv:1807.06766*, 2018.

[8] Alexandre Défossez, Leon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of Adam and Adagrad. *Transactions on Machine Learning Research*, 2022.

[9] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7):2121–2159, 2011.

[10] Matthew Faw, Isidoros Tziotis, Constantine Caramanis, Aryan Mokhtari, Sanjay Shakkottai, and Rachel Ward. The power of adaptivity in SGD: self-tuning step sizes with unbounded gradients and affine variance. In *Conference on Learning Theory*, 2022.

[11] Zhishuai Guo, Yi Xu, Wotao Yin, Rong Jin, and Tianbao Yang. A novel convergence analysis for algorithms of the Adam family. In *Annual Workshop on Optimization for Machine Learning*, 2021.

[12] Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, pages 1842–1850. PMLR, 2018.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

[14] Ali Kavis, Kfir Yehuda Levy, and Volkan Cevher. High probability bounds for a class of nonconvex algorithms with AdaGrad stepsize. In *International Conference on Learning Representations*, 2022.

[15] Diederik P Kingma and Jimmy Ba. Adam: a method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[16] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[17] Bingrui Li, Jianfei Chen, and Jun Zhu. Memory efficient optimizers with 4-bit states. *Advances in Neural Information Processing Systems*, 36, 2024.

[18] Haochuan Li, Ali Jadbabaie, and Alexander Rakhlin. Convergence of Adam under relaxed assumptions. In *Advances in Neural Information Processing Systems*, 2023.

[19] Xiaoyu Li and Francesco Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. In *International Conference on Artificial Intelligence and Statistics*, 2019.

[20] Xiaoyu Li and Francesco Orabona. A high probability analysis of adaptive SGD with momentum. In *Workshop on International Conference on Machine Learning*, 2020.

[21] Zijian Liu, Ta Duy Nguyen, Thien Hang Nguyen, Alina Ene, and Huy Nguyen. High probability convergence of stochastic gradient methods. In *International Conference on Machine Learning*, 2023.

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[23] Yang Luo, Xiaozhe Ren, Zangwei Zheng, Zhuo Jiang, Xin Jiang, and Yang You. CAME: Confidence-guided adaptive memory efficient optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023.

[24] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[25] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

[26] Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, 2018.

[27] Li Shen, Congliang Chen, Fangyu Zou, Zequn Jie, Ju Sun, and Wei Liu. A unified analysis of AdaGrad with weighted aggregation and momentum acceleration. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[28] Naichen Shi, Dawei Li, Mingyi Hong, and Ruoyu Sun. RMSProp converges with proper hyper-parameter. In *International Conference on Learning Representations*, 2020.

[29] Matthew Streeter and H Brendan McMahan. Less regret via online conditioning. *arXiv preprint arXiv:1002.4862*, 2010.

[30] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-RMSProp: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 2012.

[31] Bohan Wang, Jingwen Fu, Huishuai Zhang, Nanning Zheng, and Wei Chen. Closing the gap between the upper bound and lower bound of Adam's iteration complexity. In *Advances in Neural Information Processing Systems*, 2023.

[32] Rachel Ward, Xiaoxia Wu, and Leon Bottou. Adagrad stepsizes: sharp convergence over nonconvex landscapes. *Journal of Machine Learning Research*, 21(1):9047–9076, 2020.

[33] Junchi Yang, Xiang Li, and Niao He. Nest your adaptive algorithm for parameter-agnostic nonconvex minimax optimization. In *Advances in Neural Information Processing Systems*, 2022.

[34] Manzil Zaheer, Sashank Reddi, Devendra Sachan, Satyen Kale, and Sanjiv Kumar. Adaptive methods for nonconvex optimization. In *Advances in Neural Information Processing Systems*, 2018.

[35] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[36] Yushun Zhang, Congliang Chen, Naichen Shi, Ruoyu Sun, and Zhi-Quan Luo. Adam can converge without any modification on update rules. In *Advances in Neural Information Processing Systems*, 2022.

[37] Pengxiang Zhao, Ping Li, Yingjie Gu, Yi Zheng, Stephan Ludger Kölker, Zhefeng Wang, and Xiaoming Yuan. Adapprox: Adaptive approximation in adam optimization via randomized low-rank matrices. *arXiv preprint arXiv:2403.14958*, 2024.

[38] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

[39] Fangyu Zou, Li Shen, Zequn Jie, Weizhong Zhang, and Wei Liu. A sufficient condition for convergences of Adam and RMSProp. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

A Proof detail for full-batch case

We first provide the full-batch Adafactor as follows. The only difference to Algorithm (1) is the replacement of stochastic gradient by deterministic gradient $\nabla f(\boldsymbol{X}_k)$ at each iteration.

Algorithm 2 Full-batch Adafactor

Input: Initialization point $\boldsymbol{X}_1 \in \mathbb{R}^{n\times m}$, $\boldsymbol{R}_0 = \boldsymbol{0}_n, \boldsymbol{C}_0 = \boldsymbol{0}_m^\top$, relative step-sizes $\{\rho_k\}_{k\geq 1}$, decay rate $\{\beta_{2,k}\}_{k\geq 1} \in [0,1)$, regularization constants $\epsilon_1, \epsilon_2 > 0$, clipping threshold d.
for $k = 1, \cdots, T$ **do**
$\bar{\boldsymbol{G}}_k = \nabla f(\boldsymbol{X}_k)$;
$\bar{\boldsymbol{R}}_k = \beta_{2,k}\bar{\boldsymbol{R}}_{k-1} + (1-\beta_{2,k})(\bar{\boldsymbol{G}}_k \odot \bar{\boldsymbol{G}}_k + \epsilon_1 \boldsymbol{1}_n \boldsymbol{1}_m^\top)\boldsymbol{1}_m$;
$\bar{\boldsymbol{C}}_k = \beta_{2,k}\bar{\boldsymbol{C}}_{k-1} + (1-\beta_{2,k})\boldsymbol{1}_n^\top(\bar{\boldsymbol{G}}_k \odot \bar{\boldsymbol{G}}_k + \epsilon_1 \boldsymbol{1}_n \boldsymbol{1}_m^\top)$;
$\bar{\boldsymbol{W}}_k = (\bar{\boldsymbol{R}}_k\bar{\boldsymbol{C}}_k)/\boldsymbol{1}_n^\top\bar{\boldsymbol{R}}_k$;
$\bar{\boldsymbol{U}}_k = \bar{\boldsymbol{G}}_k/\sqrt{\bar{\boldsymbol{W}}_k}$;
$\hat{\eta}_k = \max\{\epsilon_2, \mathrm{RMS}(\boldsymbol{X}_k)\}\rho_k/\max\{1, \mathrm{RMS}(\bar{\boldsymbol{U}}_k)/d\}$;
$\boldsymbol{X}_{k+1} = \boldsymbol{X}_k - \hat{\eta}_k \cdot \bar{\boldsymbol{G}}_k/\sqrt{\bar{\boldsymbol{W}}_k}$;
end for

Then, we provide the detailed version of Theorem 5.1 as follows.

Theorem A.1. *Let $\{\boldsymbol{X}_k\}_{k\geq 1}$ be generated by Algorithm 2. If Assumptions (A1), (A2) hold, $\|\nabla f(\boldsymbol{X}_k)\|_F \leq G, \forall k \geq 1$ and*

$$\rho_k = \rho_0/\sqrt{k}, \quad 0 < \beta_{2,k} < 1, \quad \forall k \geq 1,$$

for some positive constant ρ_0, then for any $T \geq 1$,

$$\begin{aligned}
\min_{k\in[T]} \|\nabla f(\boldsymbol{X}_k)\|_F^2 &\leq \frac{A_0 A_1 (f(\boldsymbol{X}_1) - f^* + \Delta_0^2 \log T + \Delta_0^2)}{\sqrt{T}}, \\
\min_{k\in[T]} \|\nabla f(\boldsymbol{X}_k)\|_F^2 &\leq \frac{A_0 A_1' (f(\boldsymbol{X}_1) - f^* + \tilde{\Delta}_0^2 \log T + \Delta_0^2)}{\sqrt{T}},
\end{aligned} \tag{9}$$

where we define

$$\Theta_{\min} = \min_{k\in[T]} \|\boldsymbol{X}_k\|_\infty, \quad \Theta_{\max} = \max_{k\in[T]} \|\boldsymbol{X}_k\|_\infty, \quad \mathcal{G} = G^2 + mn\epsilon_1, \tag{10}$$

and the other constant parameters are given by

$$\begin{aligned}
&\Delta_0^2 = \frac{Ld^2 mn(\epsilon_2 + \Theta_{\max})^2 \rho_0^2}{2}, \quad \tilde{\Delta}_0^2 = \frac{LG^2\mathcal{G}(\epsilon_2 + \Theta_{\max})^2\rho_0^2}{2mn\epsilon_1^2(1-\beta_{2,1})^2}, \\
&A_0 = \frac{\max\left\{1, \frac{G\sqrt{\mathcal{G}}}{d\epsilon_1 mn(1-\beta_{2,1})}\right\}}{\rho_0 \max\{\epsilon_2, \Theta_{\min}\}}, A_1 = \sqrt{G^4 + G^2(m+n)\epsilon_1 + mn\epsilon_1^2}, \\
&A_1' = \sqrt{2\left(\frac{G^4}{mn\epsilon_1} + G^2 + \epsilon_1\right)}.
\end{aligned} \tag{11}$$

A.1 Preliminary

We first denote the auxiliary matrix $\bar{\boldsymbol{G}}_{k,\epsilon_1}^2 = \bar{\boldsymbol{G}}_k \odot \bar{\boldsymbol{G}}_k + \epsilon_1 \boldsymbol{1}_n \boldsymbol{1}_m^\top$. In addition, we define $\bar{\boldsymbol{V}}_k = \left(\bar{v}_{ij}^{(k)}\right)_{ij}$ as follows,

$$\bar{\boldsymbol{V}}_0 = \boldsymbol{0}_{n\times m}, \quad \bar{\boldsymbol{V}}_k = \beta_{2,k}\bar{\boldsymbol{V}}_{k-1} + (1-\beta_{2,k})\bar{\boldsymbol{G}}_{k,\epsilon_1}^2, \quad k \geq 1. \tag{12}$$

To simplify the notation, we let $\bar{\boldsymbol{G}}_k = \left(\bar{g}_{ij}^{(k)}\right)_{ij}$, $R_{\bar{\boldsymbol{V}}_k}^{(i)}$, $C_{\bar{\boldsymbol{V}}_k}^{(j)}$ and $S_{\bar{\boldsymbol{V}}_k}$ be the i-th row sum, j-th column sum and the coordinate sum of $\bar{\boldsymbol{V}}_k$ respectively. The same definition principal is applied to the notation $R_{\bar{\boldsymbol{G}}_{k,\epsilon_1}^2}^{(i)}$ and $C_{\bar{\boldsymbol{G}}_{k,\epsilon_1}^2}^{(j)}$. We also use $\bar{w}_{ij}^{(k)}, \bar{v}_{ij}^{(k)}, \bar{u}_{ij}^{(k)}$ to denote the coordinates of $\bar{\boldsymbol{W}}_k, \bar{\boldsymbol{V}}_k, \bar{\boldsymbol{U}}_k$ in Algorithm 2 respectively. We also define values $\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}$ as follows:

$$\mathcal{G}_1 = G^2 + m\epsilon_1, \quad \mathcal{G}_2 = G^2 + n\epsilon_1, \quad \mathcal{G} = G^2 + mn\epsilon_1. \tag{13}$$

A.2 Technical lemmas

Following the descent lemma for a L-smooth objective function f, we derive that

$$f(\boldsymbol{Y}) \leq f(\boldsymbol{X}) + \langle \nabla f(\boldsymbol{X}), \boldsymbol{Y} - \boldsymbol{X} \rangle + \frac{L}{2}\|\boldsymbol{Y} - \boldsymbol{X}\|_F^2, \quad \forall \boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{n\times m}. \tag{14}$$

In the following, we will provide some necessary technical lemmas.

Lemma A.1. *Let $\beta_{2,k} \in (0,1)$ and Γ_k be defined by*

$$\Gamma_0 = 0, \quad \Gamma_k = \beta_{2,k}\Gamma_{k-1} + (1-\beta_{2,k}), \quad \forall k \geq 1.$$

Then, $(1-\beta_{2,1}) \leq \Gamma_k \leq 1, \forall k \geq 1$.

Proof. We could prove the result by induction. Since $\Gamma_0 = 0$, it's easy to derive that $(1-\beta_{2,1}) = \Gamma_1 \leq 1$. Suppose that for any $j \in [k-1]$, $(1-\beta_{2,1}) \leq \Gamma_j \leq 1$. Then

$$\Gamma_k \geq \beta_{2,k}(1-\beta_{2,1}) + (1-\beta_{2,k}) \geq 1 - \beta_{2,1}, \quad \Gamma_k \leq \beta_{2,k} + (1-\beta_{2,k}) \leq 1.$$

The induction is then complete. □

Lemma A.2. *Let $\bar{\boldsymbol{V}}_k$ be defined in* (12)*. For any $k \geq 0$, it holds that*

$$\bar{\boldsymbol{R}}_k = \bar{\boldsymbol{V}}_k \mathbf{1}_m, \quad \bar{\boldsymbol{C}}_k = \mathbf{1}_n^\top \bar{\boldsymbol{V}}_k, \quad S_{\bar{\boldsymbol{V}}_k} = \mathbf{1}_n^\top \bar{\boldsymbol{R}}_k = \mathbf{1}_n^\top \bar{\boldsymbol{V}}_k \mathbf{1}_m.$$

As a consequence,

$$R^{(i)}_{\bar{\boldsymbol{V}}_k} = \beta_{2,k} R^{(i)}_{\bar{\boldsymbol{V}}_{k-1}} + (1-\beta_{2,k}) R^{(i)}_{\bar{\boldsymbol{G}}^2_{k,\epsilon_1}}, \quad C^{(j)}_{\bar{\boldsymbol{V}}_k} = \beta_{2,k} C^{(j)}_{\bar{\boldsymbol{V}}_{k-1}} + (1-\beta_{2,k}) C^{(j)}_{\bar{\boldsymbol{G}}^2_{k,\epsilon_1}}. \tag{15}$$

Proof. Note that $\bar{\boldsymbol{R}}_0 = \bar{\boldsymbol{V}}_0 \mathbf{1}_m = \mathbf{0}_n$ and $\bar{\boldsymbol{C}}_0 = \mathbf{1}_n^\top \bar{\boldsymbol{V}}_0 = \mathbf{0}_m^\top$. Suppose that for any $j \leq k-1$, $\bar{\boldsymbol{R}}_j = \bar{\boldsymbol{V}}_j \mathbf{1}_m, \bar{\boldsymbol{C}}_j = \mathbf{1}_n^\top \bar{\boldsymbol{V}}_j$. Then using the updated rule in Algorithm 2 and (12),

$$\begin{aligned}
\bar{\boldsymbol{R}}_k &= \beta_{2,k}\bar{\boldsymbol{R}}_{k-1} + (1-\beta_{2,k})\bar{\boldsymbol{G}}^2_{k,\epsilon_1}\mathbf{1}_m = \left(\beta_{2,k}\bar{\boldsymbol{V}}_{k-1} + (1-\beta_{2,k})\bar{\boldsymbol{G}}^2_{k,\epsilon_1}\right)\mathbf{1}_m = \bar{\boldsymbol{V}}_k\mathbf{1}_m, \\
\bar{\boldsymbol{C}}_k &= \beta_{2,k}\bar{\boldsymbol{C}}_{k-1} + (1-\beta_{2,k})\mathbf{1}_n^\top\bar{\boldsymbol{G}}^2_{k,\epsilon_1} = \mathbf{1}_n^\top\left(\beta_{2,k}\bar{\boldsymbol{V}}_{k-1} + (1-\beta_{2,k})\bar{\boldsymbol{G}}^2_{k,\epsilon_1}\right) = \mathbf{1}_n^\top\bar{\boldsymbol{V}}_k.
\end{aligned} \tag{16}$$

Since $S_{\bar{\boldsymbol{V}}_k}$ represents the coordinate sum of $\bar{\boldsymbol{V}}_k$, we could derive that

$$S_{\bar{\boldsymbol{V}}_k} = \sum_{i=1}^n \sum_{j=1}^m \bar{v}_{ij}^{(k)} = \mathbf{1}_n^\top \bar{\boldsymbol{R}}_k = \mathbf{1}_n^\top \bar{\boldsymbol{V}}_k \mathbf{1}_m. \tag{17}$$

Since $R^{(i)}_{\bar{\boldsymbol{V}}_k}$ denotes the i-th row sum of $\bar{\boldsymbol{V}}_k$, it's the i-th coordinate of $\bar{\boldsymbol{R}}_k$. Hence, for each coordinate of $\bar{\boldsymbol{R}}_k$, using (16),

$$R^{(i)}_{\bar{\boldsymbol{V}}_k} = \beta_{2,k} R^{(i)}_{\bar{\boldsymbol{V}}_{k-1}} + (1-\beta_{2,k}) R^{(i)}_{\bar{\boldsymbol{G}}^2_{k,\epsilon_1}}.$$

Similarly, we could derive the results related to $C^{(j)}_{\bar{\boldsymbol{V}}_k}$. □

Lemma A.3. *Following the parameter setting in* (3)*, for any $i \in [n], j \in [m], k \geq 1$, it holds that*

$$R^{(i)}_{\bar{\boldsymbol{V}}_k} \in [m\epsilon_1(1-\beta_{2,1}), \mathcal{G}_1], \quad C^{(j)}_{\bar{\boldsymbol{V}}_k} \in [n\epsilon_1(1-\beta_{2,1}), \mathcal{G}_2], \quad S_{\bar{\boldsymbol{V}}_k} \in [mn\epsilon_1(1-\beta_{2,1}), \mathcal{G}].$$

Proof. Recalling the definition of $\bar{\boldsymbol{V}}_k$ in (12) and $\|\nabla f(\boldsymbol{X}_k)\|_F \leq G, \forall k \geq 1$, we derive that

$$\begin{aligned}
S_{\bar{\boldsymbol{V}}_k} &= \sum_{i=1}^n \sum_{j=1}^m \bar{v}_{ij}^{(k)} = \sum_{i=1}^n \sum_{j=1}^m \sum_{p=1}^k (1-\beta_{2,p}) \left(\left(\bar{g}_{ij}^{(p)}\right)^2 + \epsilon_1 \right) \left(\prod_{l=p+1}^k \beta_{2,l} \right) \\
&\leq \sum_{p=1}^k (1-\beta_{2,p}) \left(\prod_{l=p+1}^k \beta_{2,l} \right) \|\bar{\boldsymbol{G}}_p\|_F^2 + \Gamma_k mn\epsilon_1 \leq G^2\Gamma_k + mn\epsilon_1 \leq \mathcal{G},
\end{aligned} \tag{18}$$

where the last inequality comes from Lemma A.1. Following (18) and Lemma A.1, we also derive that

$$S_{\bar{\boldsymbol{V}}_k} \geq mn\epsilon_1\Gamma_k \geq mn\epsilon_1(1-\beta_{2,1}).$$

We also derive the upper bounds for $R^{(i)}_{\bar{\boldsymbol{V}}_k}$ and $C^{(j)}_{\bar{\boldsymbol{V}}_k}$ as follows,

$$R^{(i)}_{\bar{\boldsymbol{V}}_k} = \sum_{j=1}^{m} \bar{v}^{(k)}_{ij} \leq \sum_{p=1}^{k}(1-\beta_{2,p})\left(\prod_{l=p+1}^{k}\beta_{2,l}\right)\|\bar{\boldsymbol{G}}_p\|_F^2 + \Gamma_k m\epsilon_1 \leq G^2\Gamma_k + m\epsilon_1 \leq \mathcal{G}_1,$$

$$C^{(j)}_{\bar{\boldsymbol{V}}_k} = \sum_{i=1}^{n} \bar{v}^{(k)}_{ij} \leq \sum_{p=1}^{k}(1-\beta_{2,p})\left(\prod_{l=p+1}^{k}\beta_{2,l}\right)\|\bar{\boldsymbol{G}}_p\|_F^2 + \Gamma_k n\epsilon_1 \leq G^2\Gamma_k + n\epsilon_1 \leq \mathcal{G}_2. \quad (19)$$

Similarly, the lower bound could be derived by

$$R^{(i)}_{\bar{\boldsymbol{V}}_k} \geq m\epsilon_1\Gamma_k \geq m\epsilon_1(1-\beta_{2,1}), \quad C^{(j)}_{\bar{\boldsymbol{V}}_k} \geq n\epsilon_1\Gamma_k \geq n\epsilon_1(1-\beta_{2,1}).$$

□

A.3 Proof of Theorem A.1

Now we move to prove the main result. Using (14) and the updated rule in Algorithm 2,

$$\begin{aligned} f(\boldsymbol{X}_{k+1}) &\leq f(\boldsymbol{X}_k) + \langle \bar{\boldsymbol{G}}_k, \boldsymbol{X}_{k+1} - \boldsymbol{X}_k\rangle + \frac{L}{2}\|\boldsymbol{X}_{k+1} - \boldsymbol{X}_k\|_F^2 \\ &= f(\boldsymbol{X}_k) - \hat{\eta}_k \left\langle \bar{\boldsymbol{G}}_k, \frac{\bar{\boldsymbol{G}}_k}{\sqrt{\bar{\boldsymbol{W}}_k}} \right\rangle + \frac{L\hat{\eta}_k^2}{2}\left\| \frac{\bar{\boldsymbol{G}}_k}{\sqrt{\bar{\boldsymbol{W}}_k}} \right\|_F^2. \end{aligned}$$

We then re-arrange the order, sum up both sides over $k \in [t]$ and apply $f(\boldsymbol{X}_{t+1}) \geq f^*$ from Assumption (A2) to get,

$$\underbrace{\sum_{k=1}^{t}\hat{\eta}_k\left\| \frac{\bar{\boldsymbol{G}}_k}{\sqrt[4]{\bar{\boldsymbol{W}}_k}} \right\|_F^2}_{(\mathbf{a})} \leq f(\boldsymbol{X}_1) - f^* + \underbrace{\frac{L}{2}\sum_{k=1}^{t}\hat{\eta}_k^2\left\| \frac{\bar{\boldsymbol{G}}_k}{\sqrt{\bar{\boldsymbol{W}}_k}} \right\|_F^2}_{(\mathbf{b})}. \quad (20)$$

Since $\Theta_{\min} \leq \|\boldsymbol{X}_k\|_\infty \leq \Theta_{\max}$, we have $\Theta_{\min} \leq \text{RMS}(\boldsymbol{X}_k) \leq \Theta_{\max}$ for any $k \geq 1$. Hence, using $\hat{\eta}_k$ defined in Algorithm 2,

$$\hat{\eta}_k = \frac{\max\{\epsilon_2, \text{RMS}(\boldsymbol{X}_k)\}\rho_k}{\max\left\{1, \|\bar{\boldsymbol{U}}_k\|_F/(d\sqrt{mn})\right\}} \leq (\epsilon_2 + \Theta_{\max})\rho_k \min\left\{1, \frac{d\sqrt{mn}}{\|\bar{\boldsymbol{U}}_k\|_F}\right\}. \quad (21)$$

Using (21), $\bar{\boldsymbol{U}}_k = \bar{\boldsymbol{G}}_k/\sqrt{\bar{\boldsymbol{W}}_k}$, Δ_0 in (11) and $\rho_k = \rho_0/\sqrt{k}$, we thus derive that

$$(\mathbf{b}) \leq \frac{Ld^2mn(\epsilon_2 + \Theta_{\max})^2}{2}\sum_{k=1}^{t}\rho_k^2 \cdot \frac{\|\bar{\boldsymbol{U}}_k\|_F^2}{\|\bar{\boldsymbol{U}}_k\|_F^2} = \Delta_0^2\sum_{k=1}^{t}\frac{1}{k}. \quad (22)$$

To lower bound **(a)**, we first discuss the maximum operator inside $\hat{\eta}_k$. Let

$$E_1 = \left\{k \in [t] \mid \|\bar{\boldsymbol{U}}_k\|_F \geq d\sqrt{mn}\right\}, \quad E_2 = \left\{k \in [t] \mid \|\bar{\boldsymbol{U}}_k\|_F \leq d\sqrt{mn}\right\}.$$

When $k \in E_1$, since $\|\boldsymbol{X}_k\|_\infty \geq \Theta_{\min}$, it derives that

$$\hat{\eta}_k \geq \frac{d\sqrt{mn}\max\{\epsilon_2, \Theta_{\min}\}\rho_k}{\|\bar{\boldsymbol{U}}_k\|_F}. \quad (23)$$

Using Lemma A.2, we first derive that $\bar{w}^{(k)}_{ij} = (R^{(i)}_{\bar{\boldsymbol{V}}_k}C^{(j)}_{\bar{\boldsymbol{V}}_k})/S_{\bar{\boldsymbol{V}}_k}$. Then, applying Lemma A.3 and $\|\nabla f(\boldsymbol{X}_k)\|_F \leq G$, we could upper bound $\|\bar{\boldsymbol{U}}_k\|_F^2$ as follows,

$$\|\bar{\boldsymbol{U}}_k\|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\left(\bar{g}^{(k)}_{ij}\right)^2 S_{\bar{\boldsymbol{V}}_k}}{R^{(i)}_{\bar{\boldsymbol{V}}_k}C^{(j)}_{\bar{\boldsymbol{V}}_k}} \leq \frac{\|\bar{\boldsymbol{G}}_k\|_F^2\mathcal{G}}{mn\epsilon_1^2(1-\beta_{2,1})^2} \leq \frac{G^2\mathcal{G}}{mn\epsilon_1^2(1-\beta_{2,1})^2}. \quad (24)$$

Hence, combining with (23) and (24), we have

$$\sum_{k\in E_1}\hat{\eta}_k\left\|\frac{\bar{\boldsymbol{G}}_k}{\sqrt[4]{\bar{\boldsymbol{W}}_k}}\right\|_F^2 \geq d\sqrt{mn}\max\{\epsilon_2,\Theta_{\min}\}\sum_{k\in E_1}\frac{\rho_k}{\|\bar{\boldsymbol{U}}_k\|_F}\left\|\frac{\bar{\boldsymbol{G}}_k}{\sqrt[4]{\bar{\boldsymbol{W}}_k}}\right\|_F^2$$
$$\geq \frac{d\epsilon_1 mn(1-\beta_{2,1})\max\{\epsilon_2,\Theta_{\min}\}}{G\sqrt{\mathcal{G}}}\sum_{k\in E_1}\rho_k\left\|\frac{\bar{\boldsymbol{G}}_k}{\sqrt[4]{\bar{\boldsymbol{W}}_k}}\right\|_F^2. \tag{25}$$

When $k\in E_2$, we obtain that $\hat{\eta}_k=\max\{\epsilon_2,\text{RMS}(\boldsymbol{X}_k)\}\rho_k\geq\max\{\epsilon_2,\Theta_{\min}\}\rho_k$ and thus

$$\sum_{k\in E_2}\hat{\eta}_k\left\|\frac{\bar{\boldsymbol{G}}_k}{\sqrt[4]{\bar{\boldsymbol{W}}_k}}\right\|_F^2 \geq \max\{\epsilon_2,\Theta_{\min}\}\sum_{k\in E_2}\rho_k\left\|\frac{\bar{\boldsymbol{G}}_k}{\sqrt[4]{\bar{\boldsymbol{W}}_k}}\right\|_F^2. \tag{26}$$

Combining with (25) and (26), we derive that

$$(\mathbf{a})\geq\max\{\epsilon_2,\Theta_{\min}\}\min\left\{1,\frac{d\epsilon_1 mn(1-\beta_{2,1})}{G\sqrt{\mathcal{G}}}\right\}\sum_{k=1}^t\rho_k\left\|\frac{\bar{\boldsymbol{G}}_k}{\sqrt[4]{\bar{\boldsymbol{W}}_k}}\right\|_F^2. \tag{27}$$

We also derive from Lemma A.2 and Lemma A.3 that for any $i\in[n],j\in[m]$,

$$\bar{w}_{ij}^{(k)}=\frac{R_{\bar{\boldsymbol{V}}_k}^{(i)}C_{\bar{\boldsymbol{V}}_k}^{(j)}}{S_{\bar{\boldsymbol{V}}_k}}\leq\frac{R_{\bar{\boldsymbol{V}}_k}^{(i)}C_{\bar{\boldsymbol{V}}_k}^{(j)}}{\sqrt{R_{\bar{\boldsymbol{V}}_k}^{(i)}C_{\bar{\boldsymbol{V}}_k}^{(j)}}}\leq\sqrt{R_{\bar{\boldsymbol{V}}_k}^{(i)}C_{\bar{\boldsymbol{V}}_k}^{(j)}}\leq\sqrt{\mathcal{G}_1\mathcal{G}_2}. \tag{28}$$

Using (28), we have

$$\left\|\frac{\bar{\boldsymbol{G}}_k}{\sqrt[4]{\bar{\boldsymbol{W}}_k}}\right\|_F^2=\sum_{i=1}^n\sum_{j=1}^m\frac{\left(\bar{g}_{ij}^{(k)}\right)^2}{\sqrt{\bar{w}_{ij}^{(k)}}}\geq\frac{\|\bar{\boldsymbol{G}}_k\|_F^2}{\sqrt{\mathcal{G}_1\mathcal{G}_2}}=\frac{\|\bar{\boldsymbol{G}}_k\|_F^2}{A_1}, \tag{29}$$

where A_1 has been defined in (11). Plugging (29) into (27), we derive that

$$(\mathbf{a})\geq\frac{\max\{\epsilon_2,\Theta_{\min}\}}{A_1}\min\left\{1,\frac{d\epsilon_1 mn(1-\beta_{2,1})}{G\sqrt{\mathcal{G}}}\right\}\sum_{k=1}^t\rho_k\|\bar{\boldsymbol{G}}_k\|_F^2. \tag{30}$$

Plugging (22) and (30) into (20), and using $\rho_k=\rho_0/\sqrt{k}$, we thus derive that

$$\min_{k\in[t]}\|\bar{\boldsymbol{G}}_k\|_F^2\sum_{k=1}^t\frac{1}{\sqrt{k}}\leq\sum_{k=1}^t\frac{\rho_k\|\bar{\boldsymbol{G}}_k\|_F^2}{\rho_0}\leq A_0A_1\left(f(\boldsymbol{X}_1)-f^*+\Delta_0^2\sum_{k=1}^t\frac{1}{k}\right),$$

where A_0 is given in (11). Moreover, we have the following results,

$$\sum_{k=1}^t\frac{1}{k}\leq 1+\int_1^t\frac{1}{x}dx=1+\log t,\quad \sum_{k=1}^t\frac{1}{\sqrt{k}}\geq\sqrt{t}. \tag{31}$$

We thus derive the first desired result in (9) as follows,

$$\min_{k\in[t]}\|\bar{\boldsymbol{G}}_k\|_F^2\leq\frac{A_0A_1}{\sqrt{t}}\left(f(\boldsymbol{X}_1)-f^*+\Delta_0^2+\Delta_0^2\log t\right). \tag{32}$$

Avoiding the curse of dimension To derive a free-dimension numerator bound, we first derive from (21) and (24) with $\rho_k=\rho_0/\sqrt{k}$ that

$$(\mathbf{b})\leq\frac{L(\epsilon_2+\Theta_{\max})^2}{2}\sum_{k=1}^t\rho_k^2\|\bar{\boldsymbol{U}}_k\|_F^2\leq\frac{LG^2\mathcal{G}(\epsilon_2+\Theta_{\max})^2}{2mn\epsilon_1^2(1-\beta_{2,1})^2}\sum_{k=1}^t\rho_k^2=\tilde{\Delta}_0^2\sum_{k=1}^t\frac{1}{k}, \tag{33}$$

where $\tilde{\Delta}_0$ has been defined in (11). In addition, we derive from Lemma A.2, Lemma A.3 and (13) that

$$\bar{w}_{ij}^{(k)}=\frac{R_{\bar{\boldsymbol{V}}_k}^{(i)}C_{\bar{\boldsymbol{V}}_k}^{(j)}}{S_{\bar{\boldsymbol{V}}_k}}\leq\frac{2\mathcal{G}_1\mathcal{G}_2}{mn\epsilon_1}\leq 2\left(\frac{G^4}{mn\epsilon_1}+G^2+\epsilon_1\right)=(A_1')^2, \tag{34}$$

where we use $m+n \leq mn$ and A_1' in (11). Thereby, we have

$$\left\| \frac{\bar{\boldsymbol{G}}_k}{\sqrt[4]{\bar{\boldsymbol{W}}_k}} \right\|_F^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\left(\bar{g}_{ij}^{(k)}\right)^2}{\sqrt{\bar{w}_{ij}^{(k)}}} \geq \frac{\|\bar{\boldsymbol{G}}_k\|_F^2}{A_1'}.$$

Combining with (27), we thus derive that

$$(\mathbf{a}) \geq \frac{\max\{\epsilon_2, \Theta_{\min}\}}{A_1'} \min\left\{1, \frac{d\epsilon_1 mn(1-\beta_{2,1})}{G\sqrt{\mathcal{G}}}\right\} \sum_{k=1}^{t} \rho_k \|\bar{\boldsymbol{G}}_k\|_F^2 \tag{35}$$

Plugging (33) and (35) into (20), and using $\rho_k = \rho_0/\sqrt{k}$, we derive that

$$\min_{k\in[t]} \|\bar{\boldsymbol{G}}_k\|_F^2 \sum_{k=1}^{t} \frac{1}{\sqrt{k}} \leq \sum_{k=1}^{t} \frac{\rho_k \|\bar{\boldsymbol{G}}_k\|_F^2}{\rho_0} \leq A_0 A_1' \left(f(\boldsymbol{X}_1) - f^* + \tilde{\Delta}_0^2 \sum_{k=1}^{t} \frac{1}{k} \right),$$

where A_0 has been defined in (11). Using (31), we derive the second desired result in (9).

$$\min_{k\in[t]} \|\bar{\boldsymbol{G}}_k\|_F^2 \leq \frac{A_0 A_1'}{\sqrt{t}} \left(f(\boldsymbol{X}_1) - f^* + \tilde{\Delta}_0^2 + \tilde{\Delta}_0^2 \log t \right). \tag{36}$$

B Proof detail for stochastic Adafactor without update clipping

We first provide the detailed version of Theorem 6.1.

Theorem B.1 (*Formal statement of Theorem 6.1*)**.** *Let $\{\boldsymbol{X}_k\}_{k\geq 1}$ be generated by Algorithm 1 without update clipping where η_k is given by* (5) *for each $k \geq 1$. If Assumptions (A1)-(A4) hold, and*

$$\beta_{2,1} = 1/2, \quad \rho_1 = \rho_0,$$
$$\beta_{2,k} = 1 - 1/k^c, \quad \rho_k = \rho_0/\sqrt{k}, \quad \forall k \geq 2,$$

for some constants $1/2 \leq c \leq 1, \rho_0 > 0$, then for any $T \geq 1, \delta \in (0,1)$, we have the following results.

When $c = 1$, with probability at least $1-\delta$,

$$\min_{k\in[T]} \|\bar{\boldsymbol{G}}_k\|_F^2 \leq \frac{C_0}{\sqrt{T}} \left(C_1 \log\left(\frac{T}{\delta}\right) + C_2 \log T + C_2 + C_3 \right), \tag{37}$$

$$\min_{k\in[T]} \|\bar{\boldsymbol{G}}_k\|_F^2 \leq \frac{C_0'}{\sqrt{T}} \left(C_1 \log\left(\frac{T}{\delta}\right) + (C_2' + C_3') \log T + C_2' + C_3' \right). \tag{38}$$

When $1/2 \leq c < 1$, with probability at least $1-\delta$,

$$\min_{k\in[T]} \|\bar{\boldsymbol{G}}_k\|_F^2 \leq \frac{C_0}{\sqrt{T}} \left(C_1 \log\left(\frac{T}{\delta}\right) + \frac{C_2}{1-c} \cdot T^{1-c} + C_2 + C_3 \right), \tag{39}$$

$$\min_{k\in[T]} \|\bar{\boldsymbol{G}}_k\|_F^2 \leq \frac{C_0}{\sqrt{T}} \left(C_1 \log\left(\frac{T}{\delta}\right) + \frac{2C_2'}{1-c} \cdot T^{1-c} + C_3' \log T + C_2' + C_3' \right). \tag{40}$$

Here, $\Theta_{\min}, \Theta_{\max}$ and $\mathcal{G}$ are as in (10)*, and*

$$C_1 = f(\boldsymbol{X}_1) - f^* + \frac{24G^2(\epsilon_2 + \Theta_{\max})\rho_0}{\sqrt{\epsilon_1}}. \tag{41}$$

The C_0, C_2, C_3 are constants defined as

$$C_0 = \frac{2\sqrt{2\mathcal{G}}}{\rho_0 \max\{\epsilon_2, \Theta_{\min}\}}, \quad C_3 = \frac{C_2}{4} \log\left(2 + \frac{2G^2}{\epsilon_1}\right),$$
$$C_2 = \frac{32mn\mathcal{G}^{\frac{3}{2}}(\epsilon_2 + \Theta_{\max})\rho_0}{\max\{m,n\}\epsilon_1} + \frac{4Lmn\mathcal{G}(\epsilon_2 + \Theta_{\max})^2\rho_0^2}{\max\{m,n\}\epsilon_1}. \tag{42}$$

The C_0', C_2', C_3' are positive constants (that could be further upper bounded by constants independent from m, n), defined by

$$C_0' = \frac{2\sqrt{2\left(\frac{G^2}{mn\epsilon_1} + G + \epsilon_1\right)}}{\rho_0 \max\{\epsilon_2, \Theta_{\min}\}}, C_2' = 4G_3(G_1+G_2)(\epsilon_2+\Theta_{\max})\rho_0, C_3' = \frac{LG_3(\epsilon_2+\Theta_{\max})^2\rho_0^2}{2}, \tag{43}$$

and G_1, G_2, G_3 are given by

$$G_1 = \sqrt{6\left(\frac{G^4}{mn\epsilon_1} + G^2 + \epsilon_1\right)}, \quad G_3 = \frac{4(G^4 + G^2 mn\epsilon_1)}{mn\epsilon_1^2},$$
$$G_2 = 2\left(\frac{G^3}{mn\epsilon_1} + \frac{2G^2}{\sqrt{mn\epsilon_1}} + \frac{G}{\sqrt{mn}} + G + \sqrt{\epsilon_1}\right). \tag{44}$$

Calculation of hyper-parameter dependency To derive a free dimension bound, we shall use the convergence bounds in (38) and (40). From (43), it's easy to show that m, n could only exist in the denominator of C_0', C_2', C_3', which could avoid the curse of dimension.

To calculate the dependency of ϵ_1, we first show that its dependency in coefficients C_0, C_1, C_2, C_3 as follows, based on the assumption that $0 < \epsilon_1 < 1$,

$$C_0 \sim \mathcal{O}(1), \quad C_1 \sim \mathcal{O}(1/\sqrt{\epsilon_1}), \quad C_2 \sim \mathcal{O}(1/\epsilon_1), \quad C_3 \sim \mathcal{O}(C_2 \log(1/\epsilon_1)). \tag{45}$$

Thereby, with the convergence bounds in (37) and (39), it's easy to show that

$$\min_{k\in[T]} \|\bar{\boldsymbol{G}}_k\|_F^2 \leq \mathcal{O}\left(\epsilon_1^{-1}\log(1/\epsilon_1)\right). \tag{46}$$

Proposition B.1. *Following the same assumptions and settings in Theorem 6.1, then with probability at least $1-\delta$,*

$$\min_{k\in[T]} \|\bar{\boldsymbol{G}}_k\|_F^2 \leq \frac{C_0}{\sqrt{T}}\left(C_1 \log\left(\frac{T}{\delta}\right) + C_2 \sum_{k=1}^{T} \frac{1}{k^c} + C_3\right),$$

and with probability at least $1-\delta$,

$$\min_{k\in[T]} \|\bar{\boldsymbol{G}}_k\|_F^2 \leq \frac{C_0'}{\sqrt{T}}\left(C_1 \log\left(\frac{T}{\delta}\right) + C_2' \sum_{k=1}^{T} \frac{1}{k^{c/2+1/2}} + C_3' \sum_{k=1}^{T} \frac{1}{k}\right),$$

where all constants are given as in Theorem B.1.

B.1 Preliminary

We first follow the notations of $\bar{\boldsymbol{G}}_k = \left(\bar{g}_{ij}^{(k)}\right)_{ij}$ and $\mathcal{G}, \mathcal{G}_1, \mathcal{G}_2$ in (13). Let $\boldsymbol{G}_k = \left(g_{ij}^{(k)}\right)_{ij}$ and $\boldsymbol{\xi}_k = \boldsymbol{G}_k - \bar{\boldsymbol{G}}_k$. We also define $\boldsymbol{G}_{k,\epsilon_1}^2 = \boldsymbol{G}_k \odot \boldsymbol{G}_k + \epsilon_1 \mathbf{1}_n \mathbf{1}_m^\top$ and $\boldsymbol{V}_k = \left(v_{ij}^{(k)}\right)_{ij}$ as follows,

$$\boldsymbol{V}_0 = \mathbf{0}_{n\times m}, \quad \boldsymbol{V}_k = \beta_{2,k}\boldsymbol{V}_{k-1} + (1-\beta_{2,k})\boldsymbol{G}_{k,\epsilon_1}^2, \quad k \geq 1. \tag{47}$$

We also define $R_{\boldsymbol{V}_k}^{(i)}, C_{\boldsymbol{V}_k}^{(j)}$ and $S_{\boldsymbol{V}_k}$ as the i-th row sum, j-th column sum and coordinate sum of $\boldsymbol{V}_k$ respectively. $R_{\boldsymbol{G}_{k,\epsilon_1}^2}^{(i)}$ and $C_{\boldsymbol{G}_{k,\epsilon_1}^2}^{(j)}$ represent the same definitions with respect to $\boldsymbol{G}_{k,\epsilon_1}^2$. Then, using a similar deduction in Lemma A.2, we also obtain that for all $k \geq 1$,

$$R_{\boldsymbol{V}_k}^{(i)} = \beta_{2,k} R_{\boldsymbol{V}_{k-1}}^{(i)} + (1-\beta_{2,k})\boldsymbol{G}_{k,\epsilon_1}^2 \mathbf{1}_m, \quad C_{\boldsymbol{V}_k}^{(j)} = \beta_{2,k} C_{\boldsymbol{V}_{k-1}}^{(j)} + (1-\beta_{2,k})\mathbf{1}_n^\top \boldsymbol{G}_{k,\epsilon_1}^2. \tag{48}$$

As a consequence of (48), each coordinate of $\boldsymbol{W}_k$ satisfies that

$$w_{ij}^{(k)} = \frac{R_{\boldsymbol{V}_k}^{(i)} C_{\boldsymbol{V}_k}^{(j)}}{S_{\boldsymbol{V}_k}} = \frac{\left(\beta_{2,k} R_{\boldsymbol{V}_{k-1}}^{(i)} + (1-\beta_{2,k}) R_{\boldsymbol{G}_{k,\epsilon_1}^2}^{(i)}\right)\left(\beta_{2,k} C_{\boldsymbol{V}_{k-1}}^{(j)} + (1-\beta_{2,k}) C_{\boldsymbol{G}_{k,\epsilon_1}^2}^{(j)}\right)}{\beta_{2,k} S_{\boldsymbol{V}_{k-1}} + (1-\beta_{2,k}) S_{\boldsymbol{G}_{k,\epsilon_1}^2}}. \tag{49}$$

Next, we introduce a proxy step-size matrix $\boldsymbol{A}_k = \left(a_{ij}^{(k)}\right)_{ij}$ such that

$$a_{ij}^{(k)} = \frac{\left(\beta_{2,k} R^{(i)}_{\boldsymbol{V}_{k-1}} + (1-\beta_{2,k})\mathcal{G}_1\right)\left(\beta_{2,k} C^{(j)}_{\boldsymbol{V}_{k-1}} + (1-\beta_{2,k})\mathcal{G}_2\right)}{\beta_{2,k} S_{\boldsymbol{V}_{k-1}} + (1-\beta_{2,k})\mathcal{G}}. \tag{50}$$

The proxy step-size technique is a standard way in the convergence analysis of adaptive methods, e.g., [32, 8]. We provide a new proxy step-size in (50) to handle the matrix factorization in Adafactor. This construction satisfies two properties. First, it's independent from $\boldsymbol{Z}_k$ in order to disrupt the correlation of stochastic gradients and adaptive step-sizes. Second, it needs to remain sufficiently close to the original adaptive step-size $w_{ij}^{(k)}$ to avoid generating divergent terms.

B.2 Technical lemmas

In the following, we first provide some more necessary technical lemmas. We introduce a concentration inequality for the martingale difference sequence, see [20] for a proof.

Lemma B.1. *Suppose that $\{Z_s\}_{s\in[T]}$ is a martingale difference sequence with respect to $\zeta_1, \cdots, \zeta_T$. Assume that for each $s \in [T]$, σ_s is a random variable dependent on $\zeta_1, \cdots, \zeta_{s-1}$ and satisfies that*

$$\mathbb{E}\left[\exp\left(\frac{Z_s^2}{\sigma_s^2}\right) \mid \zeta_1, \cdots, \zeta_{s-1}\right] \leq \mathrm{e}.$$

Then for any $\lambda > 0$, and for any $\delta \in (0,1)$, it holds that

$$\mathbb{P}\left(\sum_{s=1}^{T} Z_s > \frac{1}{\lambda}\log\left(\frac{1}{\delta}\right) + \frac{3}{4}\lambda\sum_{s=1}^{T}\sigma_s^2\right) \leq \delta.$$

Lemma B.2. *Following the parameter setting in (6), for any $i \in [n], j \in [m], k \geq 1$, it holds that*

$$R^{(i)}_{\boldsymbol{G}^2_{k,\epsilon_1}}, R^{(i)}_{\boldsymbol{V}_k} \in [m\epsilon_1/2, \mathcal{G}_1], \quad C^{(j)}_{\boldsymbol{G}^2_{k,\epsilon_1}}, C^{(j)}_{\boldsymbol{V}_k} \in [n\epsilon_1/2, \mathcal{G}_2], \quad S_{\boldsymbol{G}^2_{k,\epsilon_1}}, S_{\boldsymbol{V}_k} \in [mn\epsilon_1/2, \mathcal{G}].$$

Proof. First, using Assumption (A4), we derive that

$$\begin{aligned}
mn\epsilon_1/2 \leq S_{\boldsymbol{G}^2_{k,\epsilon_1}} &= \sum_{i=1}^{n}\sum_{j=1}^{m}\left(\left(g_{ij}^{(k)}\right)^2 + \epsilon_1\right) = \|\boldsymbol{G}_k\|_F^2 + mn\epsilon_1 \leq \mathcal{G}, \\
m\epsilon_1/2 \leq R^{(i)}_{\boldsymbol{G}^2_{k,\epsilon_1}} &= \sum_{j=1}^{m}\left(\left(g_{ij}^{(k)}\right)^2 + \epsilon_1\right) \leq \|\boldsymbol{G}_k\|_F^2 + m\epsilon_1 \leq \mathcal{G}_1, \\
n\epsilon_1/2 \leq C^{(j)}_{\boldsymbol{G}^2_{k,\epsilon_1}} &= \sum_{i=1}^{n}\left(\left(g_{ij}^{(k)}\right)^2 + \epsilon_1\right) \leq \|\boldsymbol{G}_k\|_F^2 + n\epsilon_1 \leq \mathcal{G}_2.
\end{aligned}$$

Using the similar deduction for Lemma A.3, we could show that $m\epsilon_1(1-\beta_{2,1}) \leq R^{(i)}_{\boldsymbol{V}_k} \leq \mathcal{G}_1$. Since $\beta_{2,1} = 1/2$ from (6), we then obtain the desired result. The bounds for $C^{(j)}_{\boldsymbol{V}_k}, S_{\boldsymbol{V}_k}$ could be also derived by using similar arguments. □

We have the following lemma to upper bound each coordinate of the proxy step-size matrix $\boldsymbol{A}_k$ defined in (50) .

Lemma B.3. *For any $k \geq 1$, it holds that*

$$\beta_{2,k}(1-\beta_{2,k})\epsilon_1 \leq a_{ij}^{(k)} \leq 2\min\left\{\mathcal{G}, \frac{G^2}{mn\epsilon_1} + G + \epsilon_1\right\}, \quad \forall i \in [n], j \in [m].$$

Proof. We first have

$$\frac{\beta_{2,k} R^{(i)}_{\boldsymbol{V}_{k-1}} + (1-\beta_{2,k})\mathcal{G}_1}{\beta_{2,k} S_{\boldsymbol{V}_{k-1}} + (1-\beta_{2,k})\mathcal{G}} \leq \frac{\beta_{2,k} R^{(i)}_{\boldsymbol{V}_{k-1}}}{\beta_{2,k} S_{\boldsymbol{V}_{k-1}}} + \frac{(1-\beta_{2,k})\mathcal{G}_1}{(1-\beta_{2,k})\mathcal{G}} \leq 2. \tag{51}$$

Then, recalling the definition of $a_{ij}^{(k)}$ in (50) and Lemma B.2, it derives that $C_{\boldsymbol{V}_{k-1}}^{(j)} \leq \mathcal{G}_2$ and thereby $\beta_{2,k} C_{\boldsymbol{V}_{k-1}}^{(j)} + (1-\beta_{2,k})\mathcal{G}_2 \leq \mathcal{G}_2 \leq \mathcal{G}$. Then combining with (51), we derive $a_{ij}^{(k)} \leq 2\mathcal{G}$. We also derive a free dimension bound from Lemma B.2 for $a_{ij}^{(k)}$ as follows,

$$a_{ij}^{(k)} \leq \frac{2\mathcal{G}_1\mathcal{G}_2}{mn\epsilon_1} = \frac{2(G^2 + G(m+n)\epsilon_1 + mn\epsilon_1^2)}{mn\epsilon_1} \leq 2\left(\frac{G^2}{mn\epsilon_1} + G + \epsilon_1\right),$$

where we use $m+n \leq mn$ when $m, n \geq 2$ and $\beta_{2,k} S_{\boldsymbol{V}_{k-1}} + (1-\beta_{2,k})\mathcal{G} \geq mn\epsilon_1/2$. To lower bound $a_{ij}^{(k)}$, we derive from Lemma B.2 that $\beta_{2,k} S_{\boldsymbol{V}_{k-1}} + (1-\beta_{2,k})\mathcal{G} \leq \mathcal{G}$. Thereby,

$$\begin{aligned} a_{ij}^{(k)} &\geq \frac{\beta_{2,k}(1-\beta_{2,k})\left(R_{\boldsymbol{V}_{k-1}}^{(i)}\mathcal{G}_2 + C_{\boldsymbol{V}_{k-1}}^{(j)}\mathcal{G}_1\right)}{\mathcal{G}} \geq \beta_{2,k}(1-\beta_{2,k}) \cdot \frac{(m\mathcal{G}_2 + n\mathcal{G}_1)\epsilon_1}{2\mathcal{G}} \\ &= \beta_{2,k}(1-\beta_{2,k}) \cdot \frac{[(m+n)G^2 + 2mn\epsilon_1]\epsilon_1}{2(G^2 + mn\epsilon_1)} \geq \beta_{2,k}(1-\beta_{2,k})\epsilon_1. \end{aligned}$$

□

Lemma B.4. *Let $\boldsymbol{W}_k$ and $\boldsymbol{V}_k$ be defined in Algorithm 1 without update clipping where η_k is given by* (5) *and* (47) *respectively. For any $k \geq 1$, it holds that*

$$\left\| \frac{\boldsymbol{G}_k}{\sqrt{\boldsymbol{W}_k}} \right\|_F^2 \leq \frac{2\mathcal{G}}{\max\{m,n\}\epsilon_1} \left\| \frac{\boldsymbol{G}_k}{\sqrt{\boldsymbol{V}_k}} \right\|_F^2.$$

Proof. Recalling (49), $v_{ij}^{(k)} \leq R_{\boldsymbol{V}_k}^{(i)}$,$v_{ij}^{(k)} \leq C_{\boldsymbol{V}_k}^{(j)}$ and Lemma B.2, one could verify that

$$\frac{\left(g_{ij}^{(k)}\right)^2}{w_{ij}^{(k)}} = \frac{\left(g_{ij}^{(k)}\right)^2 S_{\boldsymbol{V}_k}}{R_{\boldsymbol{V}_k}^{(i)} C_{\boldsymbol{V}_k}^{(j)}} \leq \frac{2\left(g_{ij}^{(k)}\right)^2 \mathcal{G}}{n\epsilon_1 v_{ij}^{(k)}}, \quad \frac{\left(g_{ij}^{(k)}\right)^2}{w_{ij}^{(k)}} = \frac{\left(g_{ij}^{(k)}\right)^2 S_{\boldsymbol{V}_k}}{R_{\boldsymbol{V}_k}^{(i)} C_{\boldsymbol{V}_k}^{(j)}} \leq \frac{2\left(g_{ij}^{(k)}\right)^2 \mathcal{G}}{m\epsilon_1 v_{ij}^{(k)}},$$

which leads to the desired result that

$$\|\boldsymbol{U}_k\|_F^2 = \left\| \frac{\boldsymbol{G}_k}{\sqrt{\boldsymbol{W}_k}} \right\|_F^2 \leq \frac{2\mathcal{G}}{\max\{m,n\}\epsilon_1} \left\| \frac{\boldsymbol{G}_k}{\sqrt{\boldsymbol{V}_k}} \right\|_F^2.$$

□

The following lemma is inspired by [8, Lemma 5.2] where they considered a constant $\beta_{2,k}$. Here, we generalize the result to the case of time-varying $\beta_{2,k}$ and provide the proof detail.

Lemma B.5. *For any $t \geq 1$, if $\beta_{2,k}$ are as in* (6)*, then it holds that*

$$\sum_{k=1}^{t}(1-\beta_{2,k}) \left\| \frac{\boldsymbol{G}_k}{\sqrt{\boldsymbol{V}_k}} \right\|_F^2 \leq mn \log\left(\frac{2(G^2+\epsilon_1)}{\epsilon_1}\right) + 4mn\sum_{k=1}^{t}(1-\beta_{2,k}).$$

Proof. Recalling the definition of $\boldsymbol{V}_k$ and since $\boldsymbol{V}_0 = \boldsymbol{0}_{n\times m}$, we have that for any $k \geq 1$,

$$\begin{aligned} v_{ij}^{(k)} &= \beta_{2,k} v_{ij}^{(k-1)} + (1-\beta_{2,k})\left[\left(g_{ij}^{(k)}\right)^2 + \epsilon_1\right] \\ &= \sum_{p=1}^{k}(1-\beta_{2,p})\left[\left(g_{ij}^{(p)}\right)^2 + \epsilon_1\right]\left(\prod_{l=p+1}^{k}\beta_{2,l}\right). \end{aligned}$$

Then, we have

$$(1-\beta_{2,k}) \cdot \frac{\left(g_{ij}^{(k)}\right)^2}{v_{ij}^{(k)}} = \frac{x_k}{y_k + \theta_k}, \tag{52}$$

where we set $y_0 = 0, \theta_0 = 0$ and

$$x_k = (1-\beta_{2,k})\left(g_{ij}^{(k)}\right)^2, \quad y_k = \sum_{p=1}^{k}(1-\beta_{2,p})\left(g_{ij}^{(p)}\right)^2\left(\prod_{l=p+1}^{k}\beta_{2,l}\right),$$
$$\theta_k = \epsilon_1\sum_{p=1}^{k}(1-\beta_{2,p})\left(\prod_{l=p+1}^{k}\beta_{2,l}\right), \quad \forall k \geq 1.$$

Then we have $y_k - x_k = \beta_{2,k}y_{k-1}, \forall k \geq 1$. Moreover, since $y_k \geq x_k$, we could use $\log x \geq 1 - 1/x, \forall x \geq 1$ to derive that

$$\begin{aligned}\frac{x_k}{y_k+\theta_k} &\leq \log(y_k+\theta_k) - \log(y_k+\theta_k-x_k) = \log(y_k+\theta_k) - \log(\beta_{2,k}y_{k-1}+\theta_k)\\ &= \log\left(\frac{y_k+\theta_k}{y_{k-1}+\theta_{k-1}}\right) + \log\left(\frac{y_{k-1}+\theta_{k-1}}{\beta_{2,k}y_{k-1}+\theta_k}\right).\end{aligned}$$

Noting that $\theta_k = \beta_{2,k}\theta_{k-1} + (1-\beta_{2,k})\epsilon_1$, which leads to $\beta_{2,k}\theta_{k-1} \leq \theta_k$. Hence, we further have

$$\frac{x_k}{y_k+\theta_k} \leq \log\left(\frac{y_k+\theta_k}{y_{k-1}+\theta_{k-1}}\right) + \log\left(\frac{y_{k-1}+\theta_{k-1}}{\beta_{2,k}(y_{k-1}+\theta_{k-1})}\right) = \log\left(\frac{y_k+\theta_k}{y_{k-1}+\theta_{k-1}}\right) - \log\beta_{2,k}. \tag{53}$$

Hence, summing up on both sides of (52) and (53) over $k \in [t]$, and noting that $x_1 = y_1$, we obtain that

$$\begin{aligned}\sum_{k=1}^{t}(1-\beta_{2,k})\cdot\frac{\left(g_{ij}^{(k)}\right)^2}{v_{ij}^{(k)}} &= \frac{x_1}{y_1+\theta_1} + \sum_{k=2}^{t}\frac{x_k}{y_k+\epsilon_k}\\ \leq 1 + \log\left(\frac{y_t+\theta_t}{y_1+\theta_1}\right) &- \sum_{k=2}^{t}\log\beta_{2,k}.\end{aligned} \tag{54}$$

Note that $y_1 + \theta_1 \geq (1-\beta_{2,1})\epsilon_1 = \epsilon_1/2$. Moreover, using Lemma A.1 and Assumption (A4), we have $\theta_t = \Gamma_t\epsilon_1 \leq \epsilon_1$ and $y_t \leq \Gamma_t G^2 \leq G^2$. We then derive that

$$\frac{y_t+\theta_t}{y_1+\theta_1} \leq \frac{2(G^2+\epsilon_1)}{\epsilon_1}. \tag{55}$$

Noting that for $k \geq 2$, $c \in [1/2, 1]$, $\beta_{2,k} \geq \beta_{2,2} = 1 - 1/2^c \geq 1 - 1/\sqrt{2}$, we then derive that

$$-\log\beta_{2,k} \leq \frac{1-\beta_{2,k}}{\beta_{2,k}} \leq \frac{\sqrt{2}(1-\beta_{2,k})}{\sqrt{2}-1} \leq 4(1-\beta_{2,k}). \tag{56}$$

Finally, plugging (55), (56) into (54), and then summing (54) up over $i \in [n], j \in [m]$, we obtain the desired result. □

Next, we have the following probabilistic result relying on the property of the martingale difference sequence which is commonly used in the analysis of adaptive methods.

Lemma B.6. *Following the parameter setting in* (6)*, for any $T \geq 1$ and $\lambda > 0$, with probability at least $1-\delta$, $\forall t \in [T]$,*

$$-\sum_{k=1}^{t}\eta_k\left\langle\bar{\boldsymbol{G}}_k, \frac{\boldsymbol{\xi}_k}{\sqrt{\boldsymbol{A}_k}}\right\rangle \leq \frac{1}{4}\sum_{k=1}^{t}\eta_k\left\|\frac{\bar{\boldsymbol{G}}_k}{\sqrt[4]{\boldsymbol{A}_k}}\right\|_F^2 + \frac{24G^2(\epsilon_2+\Theta_{\max})\rho_0}{\sqrt{\epsilon_1}}\log\left(\frac{T}{\delta}\right).$$

Proof. Let $\zeta_k = -\eta_k\left\langle\bar{\boldsymbol{G}}_k, \frac{\boldsymbol{\xi}_k}{\sqrt{\boldsymbol{A}_k}}\right\rangle$ and the filtration $\mathcal{F}_k = \sigma\left(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_k\right)$ where $\sigma(\cdot)$ denotes the σ-algebra. Note that η_k, $\bar{\boldsymbol{G}}_k$ and $\boldsymbol{A}_k$ are dependent by $\{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_{k-1}\}$ and thereby $\mathcal{F}_{k-1}$. Since $\boldsymbol{\xi}_k$ is dependent by $\mathcal{F}_k$, we could prove that $\{\zeta_k\}_{k\geq 1}$ is a martingale difference sequence since

$$\mathbb{E}\left[\zeta_k \mid \mathcal{F}_{k-1}\right] = -\eta_k\left\langle\bar{\boldsymbol{G}}_k, \frac{\mathbb{E}\left[\boldsymbol{\xi}_k \mid \mathcal{F}_{k-1}\right]}{\sqrt{\boldsymbol{A}_k}}\right\rangle = 0,$$

where we apply that $\mathbb{E}\left[\boldsymbol{\xi}_k \mid \mathcal{F}_{k-1}\right] = \mathbb{E}_{\boldsymbol{Z}_k}[\boldsymbol{\xi}_k] = 0$ from Assumption (A3). Then, using Assumption (A3) and Assumption (A4), we have

$$\|\bar{\boldsymbol{G}}_k\|_F = \|\mathbb{E}_{\boldsymbol{Z}_k}[\boldsymbol{G}_k]\|_F \le \mathbb{E}_{\boldsymbol{Z}_k}\|\boldsymbol{G}_k\|_F \le G, \quad \|\boldsymbol{\xi}_k\|_F = \|\boldsymbol{G}_k - \bar{\boldsymbol{G}}_k\|_F \le 2G.$$

Let $\omega_k = 2G\eta_k \left\| \frac{\bar{\boldsymbol{G}}_k}{\sqrt{\boldsymbol{A}_k}} \right\|_F$. We thus derive from the Cauchy-Schwarz inequality that

$$\mathbb{E}\left[\exp\left(\frac{\zeta_k^2}{\omega_k^2}\right) \mid \mathcal{F}_{k-1}\right] \le \mathbb{E}\left[\exp\left(\frac{\left\| \frac{\bar{\boldsymbol{G}}_k}{\sqrt{\boldsymbol{A}_k}} \right\|_F^2 \|\boldsymbol{\xi}_k\|_F^2}{4G^2 \left\| \frac{\bar{\boldsymbol{G}}_k}{\sqrt{\boldsymbol{A}_k}} \right\|_F^2}\right) \mid \mathcal{F}_{k-1}\right] \le \exp(1).$$

Then, using Lemma B.1, it leads to that for any $\lambda > 0$, with probability at least $1-\delta$,

$$\begin{aligned} -\sum_{k=1}^{t} \eta_k \left\langle \bar{\boldsymbol{G}}_k, \frac{\boldsymbol{\xi}_k}{\sqrt{\boldsymbol{A}_k}} \right\rangle &\le 3\lambda G^2 \sum_{k=1}^{t} \eta_k^2 \left\| \frac{\bar{\boldsymbol{G}}_k}{\sqrt{\boldsymbol{A}_k}} \right\|_F^2 + \frac{1}{\lambda} \log\left(\frac{1}{\delta}\right) \\ &= 3\lambda G^2 \sum_{k=1}^{t} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\eta_k}{\sqrt{a_{ij}^{(k)}}} \cdot \eta_k \frac{\left(\bar{g}_{ij}^{(k)}\right)^2}{\sqrt{a_{ij}^{(k)}}} + \frac{1}{\lambda} \log\left(\frac{1}{\delta}\right). \end{aligned} \tag{57}$$

Meanwhile, when $\Theta_{\min} \le \|\boldsymbol{X}_k\|_\infty \le \Theta_{\max}$, $\rho_k = \rho_0/\sqrt{k}$, we have

$$\Theta_{\min} \le \mathrm{RMS}(\boldsymbol{X}_k) \le \Theta_{\max}, \quad \frac{\max\{\epsilon_2, \Theta_{\min}\}\rho_0}{\sqrt{k}} \le \eta_k \le \frac{(\epsilon_2 + \Theta_{\max})\rho_0}{\sqrt{k}}. \tag{58}$$

Combining with Lemma B.3, we derive that

$$\frac{\eta_k}{\sqrt{a_{ij}^{(k)}}} \le \frac{\eta_k}{\sqrt{\beta_{2,k}(1-\beta_{2,k})\epsilon_1}} \le \frac{(\epsilon_2 + \Theta_{\max})\rho_0}{\sqrt{\beta_{2,k}\epsilon_1}} \cdot \frac{k^{c/2}}{\sqrt{k}} \tag{59}$$

$$\le \frac{(\epsilon_2 + \Theta_{\max})\rho_0}{\sqrt{\min\{\beta_{2,1}, \beta_{2,2}\}\epsilon_1}} \le \frac{2(\epsilon_2 + \Theta_{\max})\rho_0}{\sqrt{\epsilon_1}}, \tag{60}$$

where we use $\beta_{2,1} = 1/2, \beta_{2,2} = 1 - 1/2^c \ge 1 - 1/\sqrt{2}, c \in [1/2, 1]$ from (6) in the last inequality. Hence, plugging (60) into (57) and then re-scaling the δ, we found that with probability at least $1-\delta$, for all $t \in [T]$,

$$-\sum_{k=1}^{t} \eta_k \left\langle \bar{\boldsymbol{G}}_k, \frac{\boldsymbol{\xi}_k}{\sqrt{\boldsymbol{A}_k}} \right\rangle \le \frac{6\lambda G^2(\epsilon_2 + \Theta_{\max})\rho_0}{\sqrt{\epsilon_1}} \sum_{k=1}^{t} \eta_k \left\| \frac{\bar{\boldsymbol{G}}_k}{\sqrt[4]{\boldsymbol{A}_k}} \right\|_F^2 + \frac{1}{\lambda} \log\left(\frac{T}{\delta}\right).$$

Setting $\lambda = \sqrt{\epsilon_1}/(24G^2(\epsilon_2 + \Theta_{\max})\rho_0)$, we derive the desired result. □

The following key lemma provides an upper bound for the error brought by the proxy step-size $a_{ij}^{(k)}$, illustrating the error is controllable.

Lemma B.7. *For any $k \ge 1, i \in [n], j \in [m]$, it holds that*

$$\frac{\left| w_{ij}^{(k)} - a_{ij}^{(k)} \right|}{\sqrt{a_{ij}^{(k)}}} \le \sqrt{1-\beta_{2,k}} \min\{4\sqrt{\mathcal{G}}, G_1 + G_2\}, \tag{61}$$

where $\mathcal{G}$ is as in (13) *and G_1, G_2 are as in* (44)*.*

Proof. To simplify the notation, we let

$$\begin{aligned} X &= \beta_{2,k} R_{\boldsymbol{V}_{k-1}}^{(i)} + (1-\beta_{2,k}) R_{\boldsymbol{G}_{k,\epsilon_1}^2}^{(i)}, \quad \Delta X = (1-\beta_{2,k})(\mathcal{G}_1 - R_{\boldsymbol{G}_{k,\epsilon_1}^2}^{(i)}), \\ Y &= \beta_{2,k} C_{\boldsymbol{V}_{k-1}}^{(j)} + (1-\beta_{2,k}) C_{\boldsymbol{G}_{k,\epsilon_1}^2}^{(j)}, \quad \Delta Y = (1-\beta_{2,k})(\mathcal{G}_2 - C_{\boldsymbol{G}_{k,\epsilon_1}^2}^{(j)}), \\ Z &= \beta_{2,k} S_{\boldsymbol{V}_{k-1}} + (1-\beta_{2,k}) S_{\boldsymbol{G}_{k,\epsilon_1}^2}, \quad \Delta Z = (1-\beta_{2,k})(\mathcal{G} - S_{\boldsymbol{G}_{k,\epsilon_1}^2}). \end{aligned} \tag{62}$$

Then we have

$$\left|w_{ij}^{(k)} - a_{ij}^{(k)}\right| = \left|\frac{XY}{Z} - \frac{(X+\Delta X)(Y+\Delta Y)}{Z+\Delta Z}\right| = \left|\frac{XY\Delta Z - XZ\Delta Y - YZ\Delta X - Z(\Delta X\Delta Y)}{Z(Z+\Delta Z)}\right|.$$

Applying Lemma B.2, we could verify that $X, Y, Z \geq 0$ and

$$0 \leq \Delta X \leq (1-\beta_{2,k})\mathcal{G}_1, \quad 0 \leq \Delta Y \leq (1-\beta_{2,k})\mathcal{G}_2, \quad 0 \leq \Delta Z \leq (1-\beta_{2,k})\mathcal{G}. \tag{63}$$

Hence, we derive that

$$\begin{aligned}\frac{\left|w_{ij}^{(k)} - a_{ij}^{(k)}\right|}{\sqrt{a_{ij}^{(k)}}} &= \frac{|XY\Delta Z - XZ\Delta Y - YZ\Delta X - Z(\Delta X\Delta Y)|}{Z\sqrt{(X+\Delta X)(Y+\Delta Y)(Z+\Delta Z)}} \\ &\leq \underbrace{\frac{|X\Delta Y + Y\Delta X + (\Delta X\Delta Y)|}{\sqrt{(X+\Delta X)(Y+\Delta Y)(Z+\Delta Z)}}}_{(\mathbf{I})} + \underbrace{\frac{XY\Delta Z}{Z\sqrt{(X+\Delta X)(Y+\Delta Y)(Z+\Delta Z)}}}_{(\mathbf{II})}.\end{aligned} \tag{64}$$

Since $XY \geq 0$ from (62), Term **(I)** could be bounded as

$$(\mathbf{I}) \leq \frac{|X\Delta Y + Y\Delta X + (\Delta X\Delta Y)|}{\sqrt{(X\Delta Y + Y\Delta X + (\Delta X\Delta Y))(Z+\Delta Z)}} \leq \sqrt{\frac{X\Delta Y + Y\Delta X + (\Delta X\Delta Y)}{Z+\Delta Z}}. \tag{65}$$

Recalling the definition, we have $R_{\boldsymbol{V}_{k-1}}^{(i)} \leq S_{\boldsymbol{V}_{k-1}}$, $C_{\boldsymbol{V}_{k-1}}^{(j)} \leq S_{\boldsymbol{V}_{k-1}}$ for any $i \in [n], j \in [m]$. Further, applying Lemma B.2 and (63), we derive that

$$\begin{aligned}\frac{X\Delta Y}{Z+\Delta Z} &\leq \left(\frac{R_{\boldsymbol{V}_{k-1}}^{(i)}}{S_{\boldsymbol{V}_{k-1}}} + \frac{R_{\boldsymbol{G}_{k,\epsilon_1}^2}^{(i)}}{\mathcal{G}}\right)\Delta Y \leq 2(1-\beta_{2,k})\mathcal{G}_2. \\ \frac{Y\Delta X}{Z+\Delta Z} &\leq \left(\frac{C_{\boldsymbol{V}_{k-1}}^{(j)}}{S_{\boldsymbol{V}_{k-1}}} + \frac{C_{\boldsymbol{G}_{k,\epsilon_1}^2}^{(j)}}{\mathcal{G}}\right)\Delta X \leq 2(1-\beta_{2,k})\mathcal{G}_1, \\ \frac{\Delta X\Delta Y}{Z+\Delta Z} &\leq \frac{\Delta X(1-\beta_{2,k})\mathcal{G}}{(1-\beta_{2,k})\mathcal{G}} \leq (1-\beta_{2,k})\mathcal{G}_1.\end{aligned}$$

We then derive from (65), $\mathcal{G}_1 \leq \mathcal{G}$ and $\mathcal{G}_2 \leq \mathcal{G}$ that

$$(\mathbf{I}) \leq \sqrt{5(1-\beta_{2,k})\mathcal{G}}. \tag{66}$$

To derive a free dimension bound, we could obtain from Lemma B.2, (63) and $\mathcal{G} \geq mn\epsilon_1/2$ that $Z + \Delta Z \geq mn\epsilon_1/2$. Hence,

$$\frac{X\Delta Y}{Z+\Delta Z} \leq \frac{2(1-\beta_{2,k})\mathcal{G}_1\mathcal{G}_2}{mn\epsilon_1}, \quad \frac{Y\Delta X}{Z+\Delta Z} \leq \frac{2(1-\beta_{2,k})\mathcal{G}_1\mathcal{G}_2}{mn\epsilon_1}, \quad \frac{\Delta X\Delta Y}{Z+\Delta Z} \leq \frac{2(1-\beta_{2,k})\mathcal{G}_1\mathcal{G}_2}{mn\epsilon_1}.$$

We then derive that

$$(\mathbf{I}) \leq \sqrt{\frac{6(1-\beta_{2,k})\mathcal{G}_1\mathcal{G}_2}{mn\epsilon_1}} = \sqrt{\frac{6(1-\beta_{2,k})(G^4 + G^2\epsilon_1(m+n) + mn\epsilon_1^2)}{mn\epsilon_1}} \leq G_1\sqrt{1-\beta_{2,k}}, \tag{67}$$

where we used $m + n \leq mn$, and G_1 is defined in (44). Then, combining with (66) and (67), we have

$$(\mathbf{I}) \leq \sqrt{1-\beta_{2,k}}\min\{\sqrt{5\mathcal{G}}, G_1\}, \tag{68}$$

where we applied that $m + n \leq mn$ when $m, n \geq 2$. Then we move to bound **(II)**. Recalling the definitions in (62), we have $X \leq Z, Y \leq Z$. Applying (63), we have

$$(\mathbf{II}) \leq \frac{XY\Delta Z}{Z\sqrt{XY\Delta Z}} \leq \frac{\sqrt{XY\Delta Z}}{Z} \leq \sqrt{\Delta Z} \leq \sqrt{(1-\beta_{2,k})\mathcal{G}}.$$

Similarly, we derive from Lemma B.2 that $Z \geq mn\epsilon_1/2$, $X \leq \mathcal{G}_1, Y \leq \mathcal{G}_2$. Hence,

$$\begin{aligned}(\mathbf{II}) &\leq \frac{\sqrt{XY\Delta Z}}{Z} \leq \frac{2\sqrt{(1-\beta_{2,k})\mathcal{G}_1\mathcal{G}_2\mathcal{G}}}{mn\epsilon_1} \\ &\leq 2\sqrt{1-\beta_{2,k}}\left(\frac{G^3}{mn\epsilon_1} + \frac{2G^2}{\sqrt{mn\epsilon_1}} + G + \frac{G}{\sqrt{mn}} + \sqrt{\epsilon_1}\right) \leq G_2\sqrt{1-\beta_{2,k}},\end{aligned}$$

where G_2 has been defined in (44). We thus derive that

$$(\mathbf{II}) \leq \sqrt{1-\beta_{2,k}}\min\{\sqrt{\mathcal{G}}, G_2\}. \tag{69}$$

Combining (68) with (69), we then derive the desired result. □

B.3 Proof of Proposition B.1

Using the inequality in (14), we have

$$\begin{aligned}f(\boldsymbol{X}_{k+1}) &\leq f(\boldsymbol{X}_k) + \langle \bar{\boldsymbol{G}}_k, \boldsymbol{X}_{k+1} - \boldsymbol{X}_k\rangle + \frac{L}{2}\|\boldsymbol{X}_{k+1} - \boldsymbol{X}_k\|_F^2 \\ &\leq f(\boldsymbol{X}_k) - \eta_k\left\langle \bar{\boldsymbol{G}}_k, \frac{\boldsymbol{G}_k}{\sqrt{\boldsymbol{W}_k}}\right\rangle + \frac{L\eta_k^2}{2}\left\|\frac{\boldsymbol{G}_k}{\sqrt{\boldsymbol{W}_k}}\right\|_F^2.\end{aligned}$$

Introducing the proxy step-size matrix $\boldsymbol{A}_k$ in (50) and then summing up both sides over $k \in [t]$, we derive that

$$\begin{aligned}f(\boldsymbol{X}_{t+1}) \leq f(\boldsymbol{X}_1) &\underbrace{- \sum_{k=1}^t \eta_k \left\langle \bar{\boldsymbol{G}}_k, \frac{\boldsymbol{G}_k}{\sqrt{\boldsymbol{A}_k}}\right\rangle}_{\mathbf{A}} \\ &\underbrace{+ \sum_{k=1}^t \eta_k \left\langle \bar{\boldsymbol{G}}_k, \boldsymbol{G}_k \odot \left(\frac{1}{\sqrt{\boldsymbol{A}_k}} - \frac{1}{\sqrt{\boldsymbol{W}_k}}\right)\right\rangle}_{\mathbf{B}} + \underbrace{\sum_{k=1}^t \frac{L\eta_k^2}{2}\left\|\frac{\boldsymbol{G}_k}{\sqrt{\boldsymbol{W}_k}}\right\|_F^2}_{\mathbf{C}}.\end{aligned} \tag{70}$$

Estimation for A We first introduce $\boldsymbol{\xi}_k$ into **A**,

$$\mathbf{A} = -\sum_{k=1}^t \eta_k \left\|\frac{\bar{\boldsymbol{G}}_k}{\sqrt[4]{\boldsymbol{A}_k}}\right\|_F^2 - \sum_{k=1}^t \eta_k \left\langle \bar{\boldsymbol{G}}_k, \frac{\boldsymbol{\xi}_k}{\sqrt{\boldsymbol{A}_k}}\right\rangle. \tag{71}$$

Then, using Lemma B.6, with probability at least $1-\delta$, for all $t \in [T]$,

$$\mathbf{A} = -\frac{3}{4}\sum_{k=1}^t \eta_k \left\|\frac{\bar{\boldsymbol{G}}_k}{\sqrt[4]{\boldsymbol{A}_k}}\right\|_F^2 + \frac{24G^2(\epsilon_2 + \Theta_{\max})\rho_0}{\sqrt{\epsilon_1}}\log\left(\frac{T}{\delta}\right). \tag{72}$$

Estimation for B Term **B** is essentially the error brought by the proxy step-size $\boldsymbol{A}_k$. We will first calculate the gap of $1/\sqrt{w_{ij}^{(k)}}$ and $1/\sqrt{a_{ij}^{(k)}}$ as follows,

$$\left|\frac{1}{\sqrt{w_{ij}^{(k)}}} - \frac{1}{\sqrt{a_{ij}^{(k)}}}\right| = \frac{1}{\sqrt{w_{ij}^{(k)}}\sqrt{a_{ij}^{(k)}}}\left|\sqrt{w_{ij}^{(k)}} - \sqrt{a_{ij}^{(k)}}\right| \leq \frac{1}{\sqrt{w_{ij}^{(k)}}\sqrt{a_{ij}^{(k)}}}\sqrt{\left|w_{ij}^{(k)} - a_{ij}^{(k)}\right|}. \tag{73}$$

We then apply (73) and Young's inequality,

$$\begin{aligned}\mathbf{B} &\leq \sum_{k=1}^t\sum_{i=1}^n\sum_{j=1}^m \eta_k \left|\bar{g}_{ij}^{(k)} g_{ij}^{(k)}\right|\left|\frac{1}{\sqrt{w_{ij}^{(k)}}} - \frac{1}{\sqrt{a_{ij}^{(k)}}}\right| \\ &\leq \sum_{k=1}^t\sum_{i=1}^n\sum_{j=1}^m \eta_k \frac{\left|\bar{g}_{ij}^{(k)} g_{ij}^{(k)}\right|}{\sqrt{w_{ij}^{(k)}}\sqrt{a_{ij}^{(k)}}}\sqrt{\left|w_{ij}^{(k)} - a_{ij}^{(k)}\right|} \\ &\leq \frac{1}{4}\sum_{k=1}^t\sum_{i=1}^n\sum_{j=1}^m \eta_k \cdot \frac{\left(\bar{g}_{ij}^{(k)}\right)^2}{\sqrt{a_{ij}^{(k)}}} + 4\sum_{k=1}^t\sum_{i=1}^n\sum_{j=1}^m \eta_k \cdot \frac{\left|w_{ij}^{(k)} - a_{ij}^{(k)}\right|}{\sqrt{a_{ij}^{(k)}}} \cdot \left(\frac{g_{ij}^{(k)}}{\sqrt{w_{ij}^{(k)}}}\right)^2.\end{aligned} \tag{74}$$

Thus, plugging (61) in Lemma B.7 into (74), we derive that

$$\begin{aligned}\mathbf{B} &\leq \frac{1}{4}\sum_{k=1}^{t}\eta_k\left\|\frac{\bar{\boldsymbol{G}}_k}{\sqrt[4]{\boldsymbol{A}_k}}\right\|_F^2 + 4\sqrt{\mathcal{G}}\sum_{k=1}^{t}\eta_k\sqrt{1-\beta_{2,k}}\left\|\frac{\boldsymbol{G}_k}{\sqrt{\boldsymbol{W}_k}}\right\|_F^2 \\ &\leq \frac{1}{4}\sum_{k=1}^{t}\eta_k\left\|\frac{\bar{\boldsymbol{G}}_k}{\sqrt[4]{\boldsymbol{A}_k}}\right\|_F^2 + 4\sqrt{\mathcal{G}}\sum_{k=1}^{t}\frac{(\epsilon_2+\Theta_{\max})\rho_0}{\sqrt{k}}\sqrt{1-\beta_{2,k}}\left\|\frac{\boldsymbol{G}_k}{\sqrt{\boldsymbol{W}_k}}\right\|_F^2 \\ &\leq \frac{1}{4}\sum_{k=1}^{t}\eta_k\left\|\frac{\bar{\boldsymbol{G}}_k}{\sqrt[4]{\boldsymbol{A}_k}}\right\|_F^2 + 4\sqrt{\mathcal{G}}\sum_{k=1}^{t}(\epsilon_2+\Theta_{\max})\rho_0(1-\beta_{2,k})\left\|\frac{\boldsymbol{G}_k}{\sqrt{\boldsymbol{W}_k}}\right\|_F^2, \end{aligned} \tag{75}$$

where we used (58) in the second inequality and $1/\sqrt{k} \leq 1/k^{c/2}, c \in [1/2, 1]$. Furthermore, using Lemma B.4 and Lemma B.5, we derive that

$$\mathbf{B} \leq \frac{1}{4}\sum_{k=1}^{t}\eta_k\left\|\frac{\bar{\boldsymbol{G}}_k}{\sqrt[4]{\boldsymbol{A}_k}}\right\|_F^2 + \frac{8mn\mathcal{G}^{\frac{3}{2}}(\epsilon_2+\Theta_{\max})\rho_0}{\max\{m,n\}\epsilon_1}\left[\log\left(2+\frac{2G^2}{\epsilon_1}\right)+4\sum_{k=1}^{t}(1-\beta_{2,k})\right]. \tag{76}$$

Estimating C Using the similar deduction in (75) and (76), we derive that

$$\mathbf{C} \leq \frac{Lmn\mathcal{G}(\epsilon_2+\Theta_{\max})^2\rho_0^2}{\max\{m,n\}\epsilon_1}\left[\log\left(2+\frac{2G^2}{\epsilon_1}\right)+4\sum_{k=1}^{t}(1-\beta_{2,k})\right]. \tag{77}$$

Putting together We first re-arrange the order in (70) and use $f(\boldsymbol{X}_{t+1}) \geq f^*$ in Assumption (A2) to derive that

$$0 \leq f(\boldsymbol{X}_1) - f^* + \mathbf{A} + \mathbf{B} + \mathbf{C}. \tag{78}$$

We then plug (72), (76), (77) into (78) and set $t = T$, which leads to that with probability at least $1-\delta$,

$$\frac{1}{2}\sum_{k=1}^{T}\eta_k\left\|\frac{\bar{\boldsymbol{G}}_k}{\sqrt[4]{\boldsymbol{A}_k}}\right\|_F^2 \leq C_1\log\left(\frac{T}{\delta}\right) + C_2\sum_{k=1}^{T}(1-\beta_{2,k}) + C_3, \tag{79}$$

where C_1, C_2, C_3 are as in Theorem B.1. Moreover, using Lemma B.3 and (58), we have

$$\frac{1}{2}\sum_{k=1}^{T}\eta_k\left\|\frac{\bar{\boldsymbol{G}}_k}{\sqrt[4]{\boldsymbol{A}_k}}\right\|_F^2 \geq \sum_{k=1}^{T}\frac{\eta_k\left\|\bar{\boldsymbol{G}}_k\right\|_F^2}{2\max_{i,j}\sqrt{a_{ij}^{(k)}}} \geq \frac{\rho_0\max\{\epsilon_2,\Theta_{\min}\}}{2\sqrt{2\mathcal{G}}}\sum_{k=1}^{T}\frac{\left\|\bar{\boldsymbol{G}}_k\right\|_F^2}{\sqrt{k}}. \tag{80}$$

Combining with (80) and (79), and using $\sum_{k=1}^{T}1/\sqrt{k} \geq \sqrt{T}$, we derive that

$$\min_{k\in[T]}\|\bar{\boldsymbol{G}}_k\|^2 \leq \frac{C_0}{\sqrt{T}}\left(C_1\log\left(\frac{T}{\delta}\right) + C_2\sum_{k=1}^{T}(1-\beta_{2,k}) + C_3\right), \tag{81}$$

where C_0 has already been defined in (42). We then derive the first desired result that

$$\min_{k\in[T]}\|\bar{\boldsymbol{G}}_k\|^2 \leq \frac{C_0}{\sqrt{T}}\left(C_1\log\left(\frac{T}{\delta}\right) + C_2\sum_{k=1}^{T}\frac{1}{k^c} + C_3\right).$$

Free dimension bound We follow the similar deduction in (75) and use Lemma B.7 to derive that

$$\mathbf{B} \leq \frac{1}{4}\sum_{k=1}^{t}\eta_k\left\|\frac{\bar{\boldsymbol{G}}_k}{\sqrt[4]{\boldsymbol{A}_k}}\right\|_F^2 + 4(G_1+G_2)(\epsilon_2+\Theta_{\max})\rho_0\sum_{k=1}^{t}\frac{1}{k^{c/2+1/2}}\left\|\frac{\boldsymbol{G}_k}{\sqrt{\boldsymbol{W}_k}}\right\|_F^2. \tag{82}$$

Recalling the definition of $w_{ij}^{(k)}$ in (49) and Lemma B.2, we derive that

$$w_{ij}^{(k)} = \frac{R_{\boldsymbol{V}_k}^{(i)}C_{\boldsymbol{V}_k}^{(j)}}{S_{\boldsymbol{V}_k}} \geq \frac{mn\epsilon_1^2}{4\mathcal{G}}, \quad \left\|\frac{\boldsymbol{G}_k}{\sqrt{\boldsymbol{W}_k}}\right\|_F^2 \leq \frac{\|\boldsymbol{G}_k\|_F^2}{\min_{i,j}w_{ij}^{(k)}} \leq \frac{4G^2\mathcal{G}}{mn\epsilon_1^2} \leq G_3, \tag{83}$$

where G_3 is as in (44). We thus derive from (82) and (83) that

$$\mathbf{B} \leq \frac{1}{4}\sum_{k=1}^{t}\eta_k \left\| \frac{\bar{\boldsymbol{G}}_k}{\sqrt[4]{\boldsymbol{A}_k}} \right\|_F^2 + 4G_3(G_1+G_2)(\epsilon_2+\Theta_{\max})\rho_0 \sum_{k=1}^{t}\frac{1}{k^{c/2+1/2}}. \tag{84}$$

Using (58) and (83), we derive that

$$\mathbf{C} = \sum_{k=1}^{t}\frac{L\eta_k^2}{2}\left\| \frac{\boldsymbol{G}_k}{\sqrt{\boldsymbol{W}_k}} \right\|_F^2 \leq \frac{LG_3(\epsilon_2+\Theta_{\max})^2\rho_0^2}{2}\sum_{k=1}^{t}\frac{1}{k}. \tag{85}$$

Plugging the unchanged estimation for $\mathbf{A}$ in (72), (84) and (85) into (70), we have that with probability at least $1-\delta$, for all $t \in [T]$,

$$\frac{1}{2}\sum_{k=1}^{t}\eta_k \left\| \frac{\bar{\boldsymbol{G}}_k}{\sqrt[4]{\boldsymbol{A}_k}} \right\|_F^2 \leq C_1 \log\left(\frac{T}{\delta}\right) + C_2' \sum_{k=1}^{t}\frac{1}{k^{c/2+1/2}} + C_3' \sum_{k=1}^{t}\frac{1}{k}, \tag{86}$$

where C_2', C_3' are given as in (43) and C_1 is as in (41). Further, using Lemma B.3 and the similar deduction for (80),

$$\frac{1}{2}\sum_{k=1}^{t}\eta_k \left\| \frac{\bar{\boldsymbol{G}}_k}{\sqrt[4]{\boldsymbol{A}_k}} \right\|_F^2 \geq \sum_{k=1}^{t}\frac{\eta_k \left\|\bar{\boldsymbol{G}}_k\right\|_F^2}{2\max_{i,j}\sqrt{a_{ij}^{(k)}}} \geq \frac{1}{C_0'}\sum_{k=1}^{t}\frac{\left\|\bar{\boldsymbol{G}}_k\right\|_F^2}{\sqrt{k}}, \tag{87}$$

where C_0' is as in (43). Combining with (86) and (87), and setting $t = T$, we derive the second desired result in Proposition B.1 that

$$\min_{k\in[T]} \|\bar{\boldsymbol{G}}_k\|^2 \leq \frac{C_0'}{\sqrt{T}}\left(C_1 \log\left(\frac{T}{\delta}\right) + C_2' \sum_{k=1}^{T}\frac{1}{k^{c/2+1/2}} + C_3' \sum_{k=1}^{T}\frac{1}{k} \right).$$

B.4 Proof of Theorem B.1

Now based on the result in Proposition B.1, we could further derive the final convergence rate. Noting that when $c = 1$, we could bound that

$$\sum_{k=1}^{T}\frac{1}{k} \leq 1 + \int_1^T \frac{1}{x}dx \leq 1 + \log T. \tag{88}$$

Then, we obtain that

$$\begin{aligned}
\min_{k\in[T]} \|\bar{\boldsymbol{G}}_k\|_F^2 &\leq \frac{C_0}{\sqrt{T}}\left(C_1 \log\left(\frac{T}{\delta}\right) + C_2 \log T + C_2 + C_3 \right), \\
\min_{k\in[T]} \|\bar{\boldsymbol{G}}_k\|_F^2 &\leq \frac{C_0'}{\sqrt{T}}\left(C_1 \log\left(\frac{T}{\delta}\right) + (C_2'+C_3') \log T + C_2' + C_3' \right).
\end{aligned}$$

When $1/2 \leq c < 1$, we have

$$\begin{aligned}
\sum_{k=1}^{T}\frac{1}{k^c} &\leq 1 + \int_1^T \frac{1}{x^c}dx \leq 1 + \frac{T^{1-c}}{1-c}, \\
\sum_{k=1}^{T}\frac{1}{k^{c/2+1/2}} &\leq 1 + \int_1^T \frac{1}{x^{c/2+1/2}}dx \leq 1 + \frac{2T^{(1-c)/2}}{1-c}.
\end{aligned} \tag{89}$$

Then, we obtain that

$$\begin{aligned}
\min_{k\in[T]} \|\bar{\boldsymbol{G}}_k\|_F^2 &\leq \frac{C_0}{\sqrt{T}}\left(C_1 \log\left(\frac{T}{\delta}\right) + \frac{C_2}{1-c}\cdot T^{1-c} + C_2 + C_3 \right), \\
\min_{k\in[T]} \|\bar{\boldsymbol{G}}_k\|_F^2 &\leq \frac{C_0}{\sqrt{T}}\left(C_1 \log\left(\frac{T}{\delta}\right) + \frac{2C_2'}{1-c}\cdot T^{1-c} + C_3' \log T + C_2' + C_3' \right).
\end{aligned}$$

C Proof detail for stochastic Adafactor with update clipping

We first provide the detailed version of Theorem 7.1 as follows.

Theorem C.1. *Let $\{\boldsymbol{X}_k\}_{k\geq 1}$ be the sequence generated by Algorithm 1 with* (7)*. If Assumptions (A1) -(A4) hold, and*

$$\rho_k = \rho_0/\sqrt{k}, \quad d_k = k^{\frac{c}{2(\alpha-1)}}, \quad \forall k \geq 1,$$
$$\beta_{2,1} = 1/2, \quad \beta_{2,k} = 1 - 1/k^c, \forall k \geq 2.$$

When $c = 1$, with probability at least $1-\delta$,

$$\min_{k\in[T]} \|\bar{\boldsymbol{G}}_k\|_F^2 \leq \frac{D_0}{\sqrt{T}}\left(C_1 \log\left(\frac{T}{\delta}\right) + (C_2 + D_1(\alpha))\log T + C_2 + D_1(\alpha) + C_3\right), \tag{90}$$

$$\min_{k\in[T]} \|\bar{\boldsymbol{G}}_k\|_F^2 \leq \frac{D_0}{\sqrt{T}}\left(C_1 \log\left(\frac{T}{\delta}\right) + (C_2' + C_3' + D_1(\alpha))\log T + C_2' + C_3' + D_1(\alpha)\right). \tag{91}$$

When $1/2 \leq c < 1$, with probability at least $1-\delta$,

$$\min_{k\in[T]} \|\bar{\boldsymbol{G}}_k\|_F^2 \leq \frac{D_0}{\sqrt{T}}\left(C_1 \log\left(\frac{T}{\delta}\right) + \frac{C_2 + D_1(\alpha)}{1-c}\cdot T^{1-c} + C_2 + D_1(\alpha) + C_3\right), \tag{92}$$

$$\min_{k\in[T]} \|\bar{\boldsymbol{G}}_k\|_F^2 \leq \frac{D_0}{\sqrt{T}}\left(C_1 \log\left(\frac{T}{\delta}\right) + C_3' \log T + \frac{2(C_2' + D_1(\alpha))}{1-c}\cdot T^{\frac{1-c}{2}} + C_2' + C_3' + D_1(\alpha)\right), \tag{93}$$

where $C_1, C_2, C_3, C_2', C_3'$ are as in Theorem B.1 and

$$D_0 = \min\{C_0, C_0'\}, \quad D_1(\alpha) = \frac{G^{1+\alpha}G_4^{1-\alpha}\sqrt{\mathcal{G}}(\epsilon_2 + \Theta_{\max})\rho_0}{\sqrt{mn}\epsilon_1}, \quad G_4 = \frac{mn\epsilon_1}{2\sqrt{\mathcal{G}}}. \tag{94}$$

Calculation of hyper-parameters' dependency We first calculate the dependency on m, n, ϵ_1, α in the additional coefficient $D_1(\alpha)$ as follows,

$$D_1(\alpha) \sim \mathcal{O}\left(\left(\frac{\sqrt{1+mn\epsilon_1}}{mn\epsilon_1}\right)^{\alpha-1}\sqrt{\frac{1}{mn\epsilon_1^2} + \frac{1}{\epsilon_1}}\right), \tag{95}$$

which is free of the curse of dimension since mn exists in the denominator. Recalling the definitions of C_0', C_1, C_2', C_3' in (41) and (43), it's easy to verify that these coefficients are also free of the curse of dimension factor m, n since m, n exist in the denominator. Thereby, we also derive a free dimension bound selecting (91) and (93).

To calculate the dependency on ϵ_1, we could combine with (45) and (95) to derive that

$$C_0 D_1(\alpha) \sim \mathcal{O}\left(\epsilon_1^{-\alpha}\right), \quad C_0 C_1 \sim \mathcal{O}\left(1/\epsilon_1^{-1/2}\right), \quad C_0 C_3 \sim \mathcal{O}\left(\epsilon_1^{-1}\log(1/\epsilon_1)\right).$$

Thereby, selecting the bounds in (90) and (92) and noting that $\alpha > 1$, we derive that the order on ϵ_1 is

$$\mathcal{O}\left(\frac{1}{\epsilon_1^{\alpha}}\log\left(\frac{1}{\epsilon_1}\right)\right). \tag{96}$$

Moreover, it's clear to reveal that there exist mn in denominator, which could improve the dependency on ϵ_1. If we suppose that mn is comparable to ϵ_1, then we derive that $C_0 D_1(\alpha) \sim \mathcal{O}(\epsilon_1^{-1/2})$ and the order on ϵ_1 is

$$\mathcal{O}\left(\frac{1}{\epsilon_1}\log\left(\frac{1}{\epsilon_1}\right)\right). \tag{97}$$

C.1 Proof of Theorem C.1

We define

$$\tilde{\boldsymbol{G}}_k = \frac{\boldsymbol{G}_k}{\max\{1, \|\boldsymbol{U}_k\|_F/(d_k\sqrt{mn})\}}, \quad \hat{\rho}_k = \max\{\epsilon_2, \text{RMS}(\boldsymbol{X}_k)\}\rho_k. \tag{98}$$

Since $\text{RMS}(\boldsymbol{U}_k) = \|\boldsymbol{U}_k\|_F/\sqrt{mn}$, $\Theta_{\min} \le \text{RMS}(\boldsymbol{X}_k) \le \Theta_{\max}$, we derive that

$$\begin{aligned}
&\boldsymbol{X}_{k+1} = \boldsymbol{X}_k - \hat{\rho}_k \frac{\tilde{\boldsymbol{G}}_k}{\sqrt{\boldsymbol{W}_k}}, \\
&\frac{\max\{\epsilon_2, \Theta_{\min}\}\rho_0}{\sqrt{k}} \le \hat{\rho}_k \le \frac{(\epsilon_2 + \Theta_{\max})\rho_0}{\sqrt{k}} \le (\epsilon_2 + \Theta_{\max})\rho_0\sqrt{1-\beta_{2,k}},
\end{aligned} \tag{99}$$

where we applied that $1/\sqrt{k} \le 1/k^{c/2}$, $c \in [1/2, 1]$ and $\beta_{2,k} = 1 - 1/k^c$ in the last inequality. Using the inequalities in (14) and (99), we have

$$\begin{aligned}
f(\boldsymbol{X}_{k+1}) &\le f(\boldsymbol{X}_k) + \langle \bar{\boldsymbol{G}}_k, \boldsymbol{X}_{k+1} - \boldsymbol{X}_k \rangle + \frac{L}{2}\|\boldsymbol{X}_{k+1} - \boldsymbol{X}_k\|_F^2 \\
&\le f(\boldsymbol{X}_k) - \hat{\rho}_k \left\langle \bar{\boldsymbol{G}}_k, \frac{\tilde{\boldsymbol{G}}_k}{\sqrt{\boldsymbol{W}_k}} \right\rangle + \frac{L\hat{\rho}_k^2}{2} \left\| \frac{\tilde{\boldsymbol{G}}_k}{\sqrt{\boldsymbol{W}_k}} \right\|_F^2.
\end{aligned}$$

Summing up both sides over $k \in [t]$ and using $f(\boldsymbol{X}_{t+1}) \ge f^*$ from Assumption (A2), we derive that

$$0 \le f(\boldsymbol{X}_1) - f^* + \underbrace{\sum_{k=1}^t -\hat{\rho}_k \left\langle \bar{\boldsymbol{G}}_k, \frac{\tilde{\boldsymbol{G}}_k}{\sqrt{\boldsymbol{W}_k}} \right\rangle}_{\mathbf{D}} + \underbrace{\sum_{k=1}^t \frac{L\hat{\rho}_k^2}{2} \left\| \frac{\tilde{\boldsymbol{G}}_k}{\sqrt{\boldsymbol{W}_k}} \right\|_F^2}_{\mathbf{E}}. \tag{100}$$

Introducing $\boldsymbol{A}_k$ in (50), we further have the following decomposition,

$$\begin{aligned}
\mathbf{D} &= -\sum_{k=1}^t \hat{\rho}_k \left\langle \bar{\boldsymbol{G}}_k, \frac{\tilde{\boldsymbol{G}}_k}{\sqrt{\boldsymbol{A}_k}} \right\rangle + \underbrace{\sum_{k=1}^t \hat{\rho}_k \left\langle \bar{\boldsymbol{G}}_k, \left(\frac{1}{\sqrt{\boldsymbol{A}_k}} - \frac{1}{\sqrt{\boldsymbol{W}_k}} \right) \odot \tilde{\boldsymbol{G}}_k \right\rangle}_{\mathbf{D.1}} \\
&= -\sum_{k=1}^t \hat{\rho}_k \left\| \frac{\bar{\boldsymbol{G}}_k}{\sqrt[4]{\boldsymbol{A}_k}} \right\|_F^2 + \mathbf{D.1} \\
&\quad \underbrace{-\sum_{k=1}^t \hat{\rho}_k \left\langle \bar{\boldsymbol{G}}_k, \frac{\tilde{\boldsymbol{G}}_k}{\sqrt{\boldsymbol{A}_k}} - \mathbb{E}_{\boldsymbol{Z}_k}\left[\frac{\tilde{\boldsymbol{G}}_k}{\sqrt{\boldsymbol{A}_k}} \right] \right\rangle}_{\mathbf{D.2}} + \underbrace{\sum_{k=1}^t \hat{\rho}_k \left\langle \bar{\boldsymbol{G}}_k, \frac{\bar{\boldsymbol{G}}_k}{\sqrt{\boldsymbol{A}_k}} - \mathbb{E}_{\boldsymbol{Z}_k}\left[\frac{\tilde{\boldsymbol{G}}_k}{\sqrt{\boldsymbol{A}_k}} \right] \right\rangle}_{\mathbf{D.3}}.
\end{aligned} \tag{101}$$

Estimating E Hence, using (98), (99), Lemma B.4 and Lemma B.5, we derive that

$$\begin{aligned}
\mathbf{E} &\le \frac{L}{2}\sum_{k=1}^t \hat{\rho}_k^2 \left\| \frac{\boldsymbol{G}_k}{\sqrt{\boldsymbol{W}_k}} \right\|_F^2 \le \frac{L(\epsilon_2 + \Theta_{\max})^2\rho_0^2}{2} \sum_{k=1}^t (1-\beta_{2,k}) \left\| \frac{\boldsymbol{G}_k}{\sqrt{\boldsymbol{W}_k}} \right\|_F^2 \\
&\le \frac{Lmn\mathcal{G}(\epsilon_2 + \Theta_{\max})^2\rho_0^2}{\max\{m,n\}\epsilon_1} \left[\log\left(2 + \frac{2G^2}{\epsilon_1}\right) + 4\sum_{k=1}^t (1-\beta_{2,k}) \right].
\end{aligned} \tag{102}$$

To avoid the curse of dimension, we drive from (98) and (83) that

$$\left\| \frac{\tilde{\boldsymbol{G}}_k}{\sqrt{\boldsymbol{W}_k}} \right\|_F^2 = \frac{1}{\left(\max\{1, \|\boldsymbol{U}_k\|_F/(d_k\sqrt{mn})\}\right)^2} \left\| \frac{\boldsymbol{G}_k}{\sqrt{\boldsymbol{W}_k}} \right\|_F^2 \le \left\| \frac{\boldsymbol{G}_k}{\sqrt{\boldsymbol{W}_k}} \right\|_F^2 \le G_3. \tag{103}$$

Then, using (99) and (103), we derive that

$$\mathbf{E} \le \frac{LG_3(\epsilon_2 + \Theta_{\max})^2\rho_0^2}{2} \sum_{k=1}^t \frac{1}{k}. \tag{104}$$

Estimating D.1 We could follow the similar deduction in (73) and (74) to derive that

$$\begin{aligned}\mathbf{D.1} &\leq \sum_{k=1}^{t}\sum_{i=1}^{n}\sum_{j=1}^{m}\hat{\rho}_k|\bar{g}_{ij}^{(k)}\tilde{g}_{ij}^{(k)}|\left|\frac{1}{\sqrt{w_{ij}^{(k)}}}-\frac{1}{\sqrt{a_{ij}^{(k)}}}\right| \\ &\leq \sum_{k=1}^{t}\sum_{i=1}^{n}\sum_{j=1}^{m}\hat{\rho}_k\frac{|\bar{g}_{ij}^{(k)}\tilde{g}_{ij}^{(k)}|}{\sqrt{w_{ij}^{(k)}}\sqrt{a_{ij}^{(k)}}}\sqrt{\left|w_{ij}^{(k)}-a_{ij}^{(k)}\right|} \\ &\leq \frac{1}{4}\sum_{k=1}^{t}\sum_{i=1}^{n}\sum_{j=1}^{m}\hat{\rho}_k\cdot\frac{\left(\bar{g}_{ij}^{(k)}\right)^2}{\sqrt{a_{ij}^{(k)}}}+4\sum_{k=1}^{t}\sum_{i=1}^{n}\sum_{j=1}^{m}\hat{\rho}_k\cdot\frac{\left|w_{ij}^{(k)}-a_{ij}^{(k)}\right|}{\sqrt{a_{ij}^{(k)}}}\cdot\left(\frac{\tilde{g}_{ij}^{(k)}}{\sqrt{w_{ij}^{(k)}}}\right)^2. \end{aligned} \tag{105}$$

Using Lemma B.7 and (105), we further derive that

$$\begin{aligned}\mathbf{D.1} &\leq \frac{1}{4}\sum_{k=1}^{t}\hat{\rho}_k\left\|\frac{\bar{\boldsymbol{G}}_k}{\sqrt[4]{\boldsymbol{A}_k}}\right\|_F^2+4\sqrt{\mathcal{G}}\sum_{k=1}^{t}\hat{\rho}_k\sqrt{1-\beta_{2,k}}\left\|\frac{\tilde{\boldsymbol{G}}_k}{\sqrt{\boldsymbol{W}_k}}\right\|_F^2 \\ &\leq \frac{1}{4}\sum_{k=1}^{t}\hat{\rho}_k\left\|\frac{\bar{\boldsymbol{G}}_k}{\sqrt[4]{\boldsymbol{A}_k}}\right\|_F^2+4\sqrt{\mathcal{G}}\sum_{k=1}^{t}\hat{\rho}_k\sqrt{1-\beta_{2,k}}\left\|\frac{\boldsymbol{G}_k}{\sqrt{\boldsymbol{W}_k}}\right\|_F^2.\end{aligned}$$

Using (99), Lemma B.4 and Lemma B.5, we further have

$$\begin{aligned}\mathbf{D.1} &\leq \frac{1}{4}\sum_{k=1}^{t}\hat{\rho}_k\left\|\frac{\bar{\boldsymbol{G}}_k}{\sqrt[4]{\boldsymbol{A}_k}}\right\|_F^2+4\sqrt{\mathcal{G}}(\epsilon_2+\Theta_{\max})\rho_0\sum_{k=1}^{t}(1-\beta_{2,k})\left\|\frac{\boldsymbol{G}_k}{\sqrt{\boldsymbol{W}_k}}\right\|_F^2 \\ &\leq \frac{1}{4}\sum_{k=1}^{t}\hat{\rho}_k\left\|\frac{\bar{\boldsymbol{G}}_k}{\sqrt[4]{\boldsymbol{A}_k}}\right\|_F^2+\frac{8mn\mathcal{G}^{\frac{3}{2}}(\epsilon_2+\Theta_{\max})\rho_0}{\max\{m,n\}\epsilon_1}\left[\log\left(2+\frac{2G^2}{\epsilon_1}\right)+4\sum_{k=1}^{t}(1-\beta_{2,k})\right].\end{aligned} \tag{106}$$

To avoid the curse of dimension, we apply Lemma B.7, (99) and (83) to derive that

$$\begin{aligned}\mathbf{D.1} &\leq \frac{1}{4}\sum_{k=1}^{t}\hat{\rho}_k\left\|\frac{\bar{\boldsymbol{G}}_k}{\sqrt[4]{\boldsymbol{A}_k}}\right\|_F^2+4(G_1+G_2)\sum_{k=1}^{t}\hat{\rho}_k\sqrt{1-\beta_{2,k}}\left\|\frac{\boldsymbol{G}_k}{\sqrt{\boldsymbol{W}_k}}\right\|_F^2 \\ &\leq \frac{1}{4}\sum_{k=1}^{t}\hat{\rho}_k\left\|\frac{\bar{\boldsymbol{G}}_k}{\sqrt[4]{\boldsymbol{A}_k}}\right\|_F^2+4(G_1+G_2)(\epsilon_2+\Theta_{\max})\rho_0\sum_{k=1}^{t}\frac{1}{k^{c/2+1/2}}\left\|\frac{\boldsymbol{G}_k}{\sqrt{\boldsymbol{W}_k}}\right\|_F^2 \\ &\leq \frac{1}{4}\sum_{k=1}^{t}\hat{\rho}_k\left\|\frac{\bar{\boldsymbol{G}}_k}{\sqrt[4]{\boldsymbol{A}_k}}\right\|_F^2+4G_3(G_1+G_2)(\epsilon_2+\Theta_{\max})\rho_0\sum_{k=1}^{t}\frac{1}{k^{c/2+1/2}}.\end{aligned} \tag{107}$$

Estimating D.2 Since $\boldsymbol{A}_k$ is independent from $\boldsymbol{Z}_k$, it further leads to

$$\mathbf{D.2}=-\sum_{k=1}^{t}\hat{\rho}_k\left\langle\frac{\bar{\boldsymbol{G}}_k}{\sqrt{\boldsymbol{A}_k}},\tilde{\boldsymbol{G}}_k-\mathbb{E}_{\boldsymbol{Z}_k}\left[\tilde{\boldsymbol{G}}_k\right]\right\rangle.$$

Then, the deduction for estimating **D.2** follows the similar idea as in Lemma B.6, relying on a martingale difference sequence.

Let us set $\varphi_k=-\hat{\rho}_k\left\langle\frac{\bar{\boldsymbol{G}}_k}{\sqrt{\boldsymbol{A}_k}},\tilde{\boldsymbol{G}}_k-\mathbb{E}_{\boldsymbol{Z}_k}\left[\tilde{\boldsymbol{G}}_k\right]\right\rangle$ and the filtration $\mathcal{F}_k=\sigma\left(\boldsymbol{Z}_1,\cdots,\boldsymbol{Z}_k\right)$. Noting that $\hat{\rho}_k$, $\bar{\boldsymbol{G}}_k$ and $\boldsymbol{A}_k$ are dependent by $\mathcal{F}_{k-1}$. Since $\boldsymbol{\xi}_k$ is dependent by $\mathcal{F}_k$, we could prove that $\{\varphi_k\}_{k\geq 1}$ is a martingale difference sequence by showing that

$$\mathbb{E}\left[\varphi_k\mid\mathcal{F}_{k-1}\right]=-\hat{\rho}_k\left\langle\frac{\bar{\boldsymbol{G}}_k}{\sqrt{\boldsymbol{A}_k}},\mathbb{E}_{\boldsymbol{Z}_k}\left[\tilde{\boldsymbol{G}}_k-\mathbb{E}_{\boldsymbol{Z}_k}[\tilde{\boldsymbol{G}}_k]\right]\right\rangle=0.$$

In addition, using Assumptions (A3), (A4) and Jensen's inequality, we have

$$\|\tilde{\boldsymbol{G}}_k\|_F=\frac{\|\boldsymbol{G}_k\|_F}{\max\{1,\|\boldsymbol{U}_k\|/(d_k\sqrt{mn})\}}\leq\|\boldsymbol{G}_k\|_F\leq G,\quad\|\mathbb{E}_{\boldsymbol{Z}_k}[\tilde{\boldsymbol{G}}_k]\|_F\leq\mathbb{E}_{\boldsymbol{Z}_k}\|\tilde{\boldsymbol{G}}_k\|_F\leq G.$$

Therefore, we derive that

$$\|\tilde{\boldsymbol{G}}_k - \mathbb{E}_{\boldsymbol{Z}_k}[\tilde{\boldsymbol{G}}_k]\|_F \leq \|\tilde{\boldsymbol{G}}_k\|_F + \|\mathbb{E}_{\boldsymbol{Z}_k}[\tilde{\boldsymbol{G}}_k]\|_F \leq 2G. \tag{108}$$

Let $\omega_k' = 2G\hat{\rho}_k \left\| \frac{\bar{\boldsymbol{G}}_k}{\sqrt{\boldsymbol{A}_k}} \right\|_F$. We thus derive from the Cauchy-Schwarz inequality and (108) that

$$\mathbb{E}\left[\exp\left(\frac{\varphi_k^2}{(\omega_k')^2}\right) \mid \mathcal{F}_{k-1}\right] \leq \mathbb{E}\left[\exp\left(\frac{\left\|\frac{\bar{\boldsymbol{G}}_k}{\sqrt{\boldsymbol{A}_k}}\right\|_F^2 \|\tilde{\boldsymbol{G}}_k - \mathbb{E}_{\boldsymbol{Z}_k}[\tilde{\boldsymbol{G}}_k]\|_F^2}{4G^2\left\|\frac{\bar{\boldsymbol{G}}_k}{\sqrt{\boldsymbol{A}_k}}\right\|_F^2}\right) \mid \mathcal{F}_{k-1}\right] \leq \exp(1).$$

Then, using Lemma B.1, it leads to that for any $\lambda > 0$, with probability at least $1-\delta$,

$$\begin{aligned}
\mathbf{D.2} &= \sum_{k=1}^{t} \varphi_k \leq 3\lambda G^2 \sum_{k=1}^{t} \hat{\rho}_k^2 \left\| \frac{\bar{\boldsymbol{G}}_k}{\sqrt{\boldsymbol{A}_k}} \right\|_F^2 + \frac{1}{\lambda}\log\left(\frac{1}{\delta}\right) \\
&= 3\lambda G^2 \sum_{k=1}^{t}\sum_{i=1}^{n}\sum_{j=1}^{m} \frac{\hat{\rho}_k}{\sqrt{a_{ij}^{(k)}}} \cdot \hat{\rho}_k \frac{\left(\bar{g}_{ij}^{(k)}\right)^2}{\sqrt{a_{ij}^{(k)}}} + \frac{1}{\lambda}\log\left(\frac{1}{\delta}\right).
\end{aligned}$$

Since $\{\beta_{2,k}\}_{k\geq 2}$ is non-decreasing, we could apply Lemma B.3 to derive that

$$\frac{1}{\sqrt{a_{ij}^{(k)}}} \leq \sqrt{\frac{1}{\beta_{2,k}(1-\beta_{2,k})\epsilon_1}} \leq \sqrt{\frac{1}{\min\{\beta_{2,1},\beta_{2,2}\}(1-\beta_{2,k})\epsilon_1}} \leq \frac{2}{\sqrt{(1-\beta_{2,k})\epsilon_1}}.$$

Then, we apply (99), and re-scale δ to obtain that for any $\lambda > 0$, with probability at least $1-\delta$, for all $t \in [T]$,

$$\mathbf{D.2} \leq \frac{6\lambda G^2 \rho_0(\epsilon_2 + \Theta_{\max})}{\sqrt{\epsilon_1}} \sum_{k=1}^{t} \hat{\rho}_k \left\| \frac{\bar{\boldsymbol{G}}_k}{\sqrt[4]{\boldsymbol{A}_k}} \right\|_F^2 + \frac{1}{\lambda}\log\left(\frac{T}{\delta}\right).$$

Setting $\lambda = \sqrt{\epsilon_1}/(24G^2\rho_0(\epsilon_2 + \Theta_{\max}))$, we derive that

$$\mathbf{D.2} \leq \frac{1}{4}\sum_{k=1}^{t} \hat{\rho}_k \left\| \frac{\bar{\boldsymbol{G}}_k}{\sqrt[4]{\boldsymbol{A}_k}} \right\|_F^2 + \frac{24G^2\rho_0(\epsilon_2 + \Theta_{\max})}{\sqrt{\epsilon_1}}\log\left(\frac{T}{\delta}\right). \tag{109}$$

Estimating D.3 First, since $\boldsymbol{A}_k$ is independent from $\boldsymbol{Z}_k$ and $\mathbb{E}_{\boldsymbol{Z}_k}[\boldsymbol{G}_k] = \bar{\boldsymbol{G}}_k$, we have

$$\begin{aligned}
\mathbf{D.3} &= \sum_{k=1}^{t} \hat{\rho}_k \left\langle \bar{\boldsymbol{G}}_k, \frac{\mathbb{E}_{\boldsymbol{Z}_k}[\boldsymbol{G}_k]}{\sqrt{\boldsymbol{A}_k}} - \frac{\mathbb{E}_{\boldsymbol{Z}_k}[\tilde{\boldsymbol{G}}_k]}{\sqrt{\boldsymbol{A}_k}} \right\rangle \\
&\leq \sum_{k=1}^{t} \hat{\rho}_k \left\| \frac{\bar{\boldsymbol{G}}_k}{\sqrt{\boldsymbol{A}_k}} \right\|_F \cdot \left\| \mathbb{E}_{\boldsymbol{Z}_k} \underbrace{\left[\boldsymbol{G}_k - \frac{\boldsymbol{G}_k}{\max\{1, \|\boldsymbol{U}_k\|_F/(d_k\sqrt{mn})\}}\right]}_{\Omega_k} \right\|_F.
\end{aligned} \tag{110}$$

We define the random variable $S_k^{(1)}$, $S_k^{(2)}$ and $\tilde{S}_k^{(1)}$ using the indicator function χ and G_4 in (94) as follows,

$$S_k^{(1)} = \chi_{\{\|\boldsymbol{U}_k\|_F > d_k\sqrt{mn}\}}, \quad S_k^{(2)} = \chi_{\{\|\boldsymbol{U}_k\|_F \leq d_k\sqrt{mn}\}}, \quad \tilde{S}_k^{(1)} = \chi_{\{\|\boldsymbol{G}_k\|_F \geq d_k G_4\}}.$$

From (83), we derive that

$$\|\boldsymbol{U}_k\|_F \leq \|\boldsymbol{G}_k\|_F \cdot \frac{2\sqrt{\mathcal{G}}}{\sqrt{mn}\epsilon_1}.$$

Hence, $S_k^{(1)} \leq \tilde{S}_k^{(1)}, \forall k \geq 1$. Note that when $S_k^{(2)} = 1$, it's equivalent to $\Omega_k = 0$. Then, we derive that

$$\begin{aligned}
&\|\mathbb{E}_{\boldsymbol{Z}_k}[\Omega_k]\|_F = \left\|\mathbb{E}_{\boldsymbol{Z}_k}[\Omega_k S_k^{(1)}] + \mathbb{E}_{\boldsymbol{Z}_k}[\Omega_k S_k^{(2)}]\right\|_F = \left\|\mathbb{E}_{\boldsymbol{Z}_k}[\Omega_k S_k^{(1)}]\right\|_F \\
&\leq \mathbb{E}_{\boldsymbol{Z}_k}\left\|\Omega_k S_k^{(1)}\right\|_F \leq \mathbb{E}_{\boldsymbol{Z}_k}\left\|\Omega_k \tilde{S}_k^{(1)}\right\|_F \leq \mathbb{E}_{\boldsymbol{Z}_k}\left\|\boldsymbol{G}_k \tilde{S}_k^{(1)}\right\|_F \leq G^{\alpha}\left(d_k G_4\right)^{1-\alpha},
\end{aligned} \tag{111}$$

Furthermore, we use Assumption (A4) and Lemma B.2 to derive a lower bound for $a_{ij}^{(k)}$ where

$$a_{ij}^{(k)} \geq \frac{mn\epsilon_1^2}{4\mathcal{G}}, \quad \left\| \frac{\bar{\boldsymbol{G}}_k}{\sqrt{\boldsymbol{A}_k}} \right\|_F \leq \frac{\|\bar{\boldsymbol{G}}_k\|_F}{\min_{i,j} \sqrt{a_{ij}^{(k)}}} \leq \frac{2G\sqrt{\mathcal{G}}}{\sqrt{mn}\epsilon_1}. \tag{112}$$

Combining with (99), (110), (111) and (112), we thus derive that

$$\mathbf{D.3} \leq \frac{2G^{1+\alpha}G_4^{1-\alpha}\sqrt{\mathcal{G}}(\epsilon_2 + \Theta_{\max})\rho_0}{\sqrt{mn}\epsilon_1} \sum_{k=1}^{t} \frac{1}{d_k^{\alpha-1}\sqrt{k}}. \tag{113}$$

Putting together Both **E** and **D.1** are bounded with two estimations, one of which owns a better dependency to $1/\epsilon_1$ and the other avoids the curse of the dimension. We thereby derive two results. Plugging (106), (109) and (113) into (101) and then combining with (102) and (100), we then derive that with probability at least $1-\delta$, for all $t \in [T]$,

$$\frac{1}{2}\sum_{k=1}^{t} \hat{\rho}_k \left\| \frac{\bar{\boldsymbol{G}}_k}{\sqrt[4]{\boldsymbol{A}_k}} \right\|_F^2 \leq C_1 \log\left(\frac{T}{\delta}\right) + C_2 \sum_{k=1}^{t}(1-\beta_{2,k}) + C_3 + D_1(\alpha)\sum_{k=1}^{t} \frac{1}{d_k^{\alpha-1}\sqrt{k}}, \tag{114}$$

where C_1, C_2, C_3 are as in Theorem B.1 and $D_1(\alpha)$ is as in (94). Plugging (107), (109) and (113) into (101), then combining with (104) and (100), we then derive that with probability at least $1-\delta$, for all $t \in [T]$,

$$\frac{1}{2}\sum_{k=1}^{t} \hat{\rho}_k \left\| \frac{\bar{\boldsymbol{G}}_k}{\sqrt[4]{\boldsymbol{A}_k}} \right\|_F^2 \leq C_1 \log\left(\frac{T}{\delta}\right) + C_2' \sum_{k=1}^{t} \frac{1}{k^{c/2+1/2}} + C_3' \sum_{k=1}^{t} \frac{1}{k} + D_1(\alpha)\sum_{k=1}^{t} \frac{1}{d_k^{\alpha-1}\sqrt{k}}. \tag{115}$$

where C_2', C_3' are as in Theorem B.1. Moreover, using (99), we reveal that the lower bound for $\hat{\rho}_k$ is the same the one for η_k in (58). Thereby, following the same deduction in (80) and (86), we derive that

$$\frac{1}{2}\sum_{k=1}^{T} \hat{\rho}_k \left\| \frac{\bar{\boldsymbol{G}}_k}{\sqrt[4]{\boldsymbol{A}_k}} \right\|_F^2 \geq \sum_{k=1}^{T} \frac{\hat{\rho}_k}{2} \frac{\left\|\bar{\boldsymbol{G}}_k\right\|_F^2}{\max_{i,j}\sqrt{a_{ij}^{(k)}}} \geq \frac{1}{D_0}\sum_{k=1}^{T} \frac{1}{\sqrt{k}} \left\|\bar{\boldsymbol{G}}_k\right\|_F^2, \tag{116}$$

where $D_0 = \min\{C_0, C_0'\}$ that has been defined in (94). Setting $t = T$ on (114) and (115), and then using (116), we then derive that

$$\min_{t\in[T]} \left\|\bar{\boldsymbol{G}}_k\right\|_F^2 \leq \frac{D_0}{\sum_{t=1}^{T} 1/\sqrt{k}} \left(C_1 \log\left(\frac{T}{\delta}\right) + C_2 \sum_{k=1}^{t}(1-\beta_{2,k}) + C_3 + D_1(\alpha)\sum_{k=1}^{t} \frac{1}{d_k^{\alpha-1}\sqrt{k}} \right),$$

$$\min_{t\in[T]} \left\|\bar{\boldsymbol{G}}_k\right\|_F^2 \leq \frac{D_0}{\sum_{t=1}^{T} 1/\sqrt{k}} \left(C_1 \log\left(\frac{T}{\delta}\right) + C_2' \sum_{k=1}^{t} \frac{1}{k^{(c+1)/2}} + C_3' \sum_{k=1}^{t} \frac{1}{k} + D_1(\alpha)\sum_{k=1}^{t} \frac{1}{d_k^{\alpha-1}\sqrt{k}} \right).$$

Then, using the results in (88) and (89), we could derive the desired result in Theorem C.1.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: [NA]

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Our code is based on Pytorch package which is standard. In addition, we have clarified the detailed experimental setup in our paper and the experiments are easy to reproduce.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.