Memory Efficient Stochastic Adaptive Optimization via Subset-Norm

Thien Hang Nguyen NGUYEN.THIEN@NORTHEASTERN.EDU
Huy Le Nguyen HU.NGUYEN@NORTHEASTERN.EDU
Northeastern University, Boston, MA

Abstract

As deep neural networks grow larger, memory efficiency becomes crucial, with optimizer states of popular algorithms like Adam consuming substantial memory. This paper generalizes existing high-probability convergence analysis for AdaGrad and AdaGrad-Norm to arbitrary parameter partitions, encompassing both algorithms. We reveal a trade-off between coordinate-noise density and the convergence rate's dimensional dependency, suggesting an optimal grouping between the full coordinate version (AdaGrad) and the scalar version (AdaGrad-Norm). This insight leads to a principled compression approach called *Subset-Norm*, targeting coordinate-wise second moment term in AdaGrad, RMSProp, and Adam. We demonstrate the empirical effectiveness of subset-norm step sizes in LLM pre-training tasks on LLaMA models, showing competitive performance to baselines like Adam while significantly reducing memory usage for the optimizer's state from $O(d)$ to $O(\sqrt{d})$ while introducing no additional hyperparameter.

1. Introduction

Training modern deep neural networks, such as large language models (LLMs) and large vision models (LVMs), is costly. Consequently, theoretical interest in the convergence analysis of adaptive methods extends beyond asymptotic considerations. It now encompasses not only assumptions about the objective function (e.g., convexity, smoothness) and stochastic gradients (e.g., noise distribution), but also non-asymptotic dependencies on the total number of iterations, parameter count, and failure probability. As deep neural networks continue to grow in the era of LLMs and LVMs, concerns that were previously overlooked, such as the memory consumption of optimizer states, have become an active area of research. Indeed, numerous methods have recently emerged to reduce the memory footprint of optimizer states (e.g. Adam's momentum and second moment terms) with approaches ranging from quantization [6, 7, 17], low-rank decomposition [14, 19, 28, 36] , sketching-based dimensionality reduction [12, 22], etc.

Algorithm 1: AdaGrad-Norm

Input: x_1, $\eta > 0$
for $t = 1$ **to** T **do**
  $b_t = \sqrt{b_0^2 + \sum_{i=1}^{t} \|\widehat{\nabla} f(x_i)\|^2}$;
  $x_{t+1} = x_t - \frac{\eta}{b_t} \widehat{\nabla} f(x_t)$;
end

Algorithm 2: AdaGrad-Coordinate

Input: x_1, $b_0 \in \mathbb{R}^d$, $\eta \in \mathbb{R}$
for $t = 1$ **to** T **do**
  $b_{t,i} = \sqrt{b_{0,i}^2 + \sum_{j=1}^{t} \widehat{\nabla}_i f(x_j)^2}$, $i \in [d]$;
  $x_{t+1,i} = x_{t,i} - \frac{\eta}{b_{t,i}} \widehat{\nabla}_i f(x_t)$, $i \in [d]$;
end

Algorithm 3: AdaGrad-Subset-Norm

Input: $x_1 \in \mathbb{R}^d$ and step size $\eta > 0$
Data: Partition coordinates into c subsets: $[d] = \bigcup_{i=0}^{c-1} \Psi_i$ where $\Psi_k \cap \Psi_j = \emptyset$ if $k \neq j$.
for $t = 1$ *to* T **do**

$$b_{t,i}^2 = b_{t-1,i}^2 + \left\| \widehat{\nabla}_{\Psi_i} f(x_t) \right\|^2, \text{ for } i = 0, \ldots, c-1;$$

$$x_{t+1,k} = x_{t,k} - \frac{\eta}{b_{t,i}} \widehat{\nabla} f(x_t), \quad \text{where } k \in \Psi_i \text{ for } i = 0, \ldots, c-1;$$

end

Convergence analysis for adaptive methods like Adam [16] and AdaGrad [9] remains an active area of research. Recently, [21] presented high-probability noise-adapted and optimal (in terms of total iterations T) convergence analyses for AdaGrad-Norm [33] (Algorithm 1) and the standard AdaGrad [9] algorithm (Algorithm 2) under relaxed conditions, including sub-Gaussian noise and unbounded gradients. While the per-coordinate version of AdaGrad (Algorithm 2) is primarily used in practice, certain technical difficulties (as discussed in [21]) have led researchers to focus theoretical analysis on the normed version (Algorithm 1) as a proxy [3, 15, 20, 21, 33]. While AdaGrad-Norm has primarily served as a theoretical proxy for the original AdaGrad (and by extension, for Adam), its practical performance and comparisons against other methods have been underexplored. Theoretical results for coordinate-wise AdaGrad were limited until recent developments [5, 13, 21], resulting in a scarcity of comparisons between AdaGrad-Norm and coordinate-wise AdaGrad. There are several advantages of AdaGrad-Norm over AdaGrad-Coordinate that the present results (Theorem 4.5 and 4.6 of [21]) suggest. First, the current results suggest that AdaGrad-Norm converges with no dependency on the dimension of the problem i.e. the parameter count. Second, the memory cost of AdaGrad-Norm is constant whereas AdaGrad-Coordinate needs to maintain the adaptive step-size state $b_t \in \mathbb{R}^d$. Hence, current theory suggests that AdaGrad-Norm should be the superior optimizer. However, we perform pre-training experiments to compare the practical performance between AdaGrad-Coordinate and AdaGrad-Norm: there is a wide gap between AdaGrad-Norm and AdaGrad-Coordinate, as shown in Figure 1. Indeed, the theoretical comparison previously discussed is not fair: the noise models and step size dependency on the dimension do not entirely align as the dimensions could be hidden.

Hence, in this paper, we provide an answer to this discrepancy by unifying the analysis of AdaGrad-Norm over AdaGrad-Coordinate under both the coordinate-wise sub-Gaussian noise model and provide a proof that generalizes the adaptive step sizes of the algorithms to be able to use arbitrary partitions of the model parameters (Algorithm 3).

Our Contributions.

- We unify and generalize high-probability non-convex convergence proofs for AdaGrad-Norm and AdaGrad-Coordinate under a general adaptive step size using subset-norm (Algorithm 3) for coordinate-wise sub-Gaussian noise (Section 3.1).
- We analyze the interactions between coordinate noise sparsity, subset size, and convergence rate's dimensional dependency, showing that the optimal subset size lies between AdaGrad-Norm and AdaGrad-Coordinate, depending on noisy coordinate density (Section 3.2).

- We demonstrate the effectiveness of the subset-norm adaptive step size in LLM pre-training by incorporating it into Adam [16] and RMSProp [29], replacing the second moment term with an exponential moving average subset-norm adaptive step size (Algorithms 4 and 5). This approach outperforms baselines for LLaMA 60M and 130M while using significantly less memory ($\sqrt{d}$ instead of d) and requiring minimal additional tuning (Section 4).

2. Preliminaries

We consider the unconstrained non-convex stochastic optimization problem $\min_{x\in\mathbb{R}^d} f(x)$ where $f:\mathbb{R}^d\to\mathbb{R}$ is the objective function. We assume access to an history independent, non-biased stochastic gradient $\widehat{\nabla} f(x)$ for any $x\in\mathcal{X}$, that is $\mathbb{E}\left[\widehat{\nabla} f(x)\mid x\right]=\nabla f(x)$. Furthermore, we assume that f is an L-smooth function: $\|\nabla f(x)-\nabla f(y)\|\le L\|x-y\|$ for all $x,y\in\mathbb{R}^d$. Smoothness implies the following quadratic upperbound that we will utilize: for all $x,y\in\mathbb{R}^d$ we have $f(y)-f(x)\le\langle\nabla f(x),y-x\rangle+\frac{L}{2}\|y-x\|^2$. Before discussing the assumption on the stochastic gradient noise, let us first define some notations.

Notations. We let v_i denote the i-th coordinate of a vector $v\in\mathbb{R}^d$. If a vector like x_t is already indexed as part of a sequence of vectors (where x_t denotes the t-th update) then we use $x_{t,i}$ to denote x_t's i-th coordinate and $x_{t,\Psi}\in\mathbb{R}^k$ to denote the indexing with respect to an ordered subset $\Psi\subseteq[d]$ of size k where $(x_{t,\Psi})_k=x_{t,\Psi^{(k)}}$ where $\Psi^{(k)}$ is the k-th element of Ψ. For gradients, we let $\nabla_i f(x):=\frac{\partial f}{\partial x_i}$ denote the partial derivative with respect to the i-th coordinate. Similarly, for stochastic gradients $\widehat{\nabla} f(x)$, we let $\widehat{\nabla}_i f(x)$ denotes its i-th coordinate. If $a,b\in\mathbb{R}^d$, then ab and a/b denotes coordinate-wise multiplication and division, respectively: $(ab)_i=a_ib_i$ and $(a/b)_i=a_i/b_i$.

Coordinate sub-Gaussian noise assumptions. If we denote the stochastic gradient noise as $\xi_t:=\widehat{\nabla} f(x_t)-\nabla f(x_t)$ and $\xi_{t,i}$ as the i-th coordinate of ξ_t, then we assume the noise is per-coordinate subgaussian i.e. there exists $\sigma_i>0$ for $i\in[d]$ such that ξ_t satisfies

$$\mathbb{E}\left[\exp\left(\lambda^2\xi_{t,i}^2\right)\right]\le\exp\left(\lambda^2\sigma_i^2\right),\forall\,|\lambda|\le\frac{1}{\sigma_i},\forall i\in[d]\,. \tag{1}$$

Note that $\|\xi_t\|$ being σ-subgaussian implies that each $\xi_{t,i}$ is also σ-subgaussian, so our assumption is more general than the subgaussian noise assumption. Furthermore, when $\|\cdot\|$ is used without explicitly specifying the norm, one can assume it is the ℓ_2 norm $\|\cdot\|_2$. We also use 0-indexing convention i.e. $[n]:=\{0,1,\ldots,n-1\}$ for integer $n\in\mathbb{N}$.

3. AdaGrad-Subset-Norm: Better Convergence, Less Memory

We partition the parameters' coordinates $[d]$ into disjoint subsets $[d]=\bigcup_{i=0}^{c-1}\Psi_i$ with $\Psi_i\cap\Psi_j=\emptyset$, if $i\neq j$ (e.g. $\Psi_i=\{ik+1,ik+2,\ldots,ik+k\}$ for some subset size $k\in\mathbb{N}$ so that $kc=d$). Given a stochastic gradient $\widehat{\nabla} f(x_t)\in\mathbb{R}^d$ at time t for parameter x_t, we denote $\widehat{\nabla}_{\Psi_i} f(x_t)\in\mathbb{R}^k$ to be the subset of the coordinates of the stochastic gradient with respect to the subset Ψ_i (e.g. $\left(\widehat{\nabla}_{\Psi_i} f(x_t)\right)_j=\widehat{\nabla}_{ik+j-1} f(x_t)$). Similarly, we can define $\nabla_{\Psi_i} f(x_t)$ to be $\frac{\partial f(x_t)}{\partial x_{\Psi_i}}$. We define the

"subset-norm adaptive step size" $b_{t,i}$ for subset Ψ_i and the update rule for x_{t+1}:

$$b_{t,i}^2 = b_{t-1,i}^2 + \left\|\widehat{\nabla}_{\Psi_i} f(x_t)\right\|^2 = b_0^2 + \sum_{j=1}^{t} \left\|\widehat{\nabla}_{\Psi_i} f(x_t)\right\|^2, \ i = 0, 1, \ldots, c-1$$
$$x_{t+1,k} = x_{t,k} - \frac{\eta}{b_{t,i}} \widehat{\nabla}_k f(x_t), \quad \text{where } k \in \Psi_i, \text{ for all } i \in [c]. \tag{2}$$

The algorithm is also presented in Algorithm 3. Note that choosing $c = d$ and $c = 1$ recovers AdaGrad-Coordinate and AdaGrad-Norm, respectively.

3.1. High-probability convergence of AdaGrad-Subset-Norm for non-convex objectives

We present the following high-probability convergence result for AdaGrad-subset-norm from (2):

Theorem 1 *Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and lower bounded by $f_* \in \mathbb{R}$. Assume access to unbiased stochastic gradients $\widehat{\nabla} f(x_t)$ with stochastic gradient noise $\xi_t := \widehat{\nabla} f(x_t) - \nabla f(x_t)$ being σ_i-per-coordinate subgaussian for $i \in [d]$. For partitions of the parameters into disjoint subsets $[d] = \bigcup_{i=0}^{c-1} \Psi_i$ with $\Psi_i \cap \Psi_j = \emptyset$ if $i \neq j$, the iterates x_t given by (2) satisfy the following with probability at least $1 - \delta$ (for failure probability $\delta > 0$):*

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(x_t)\|_2^2 \leq G(\delta) \cdot \tilde{O}\left(\frac{\sum_{i=0}^{c-1} \|\sigma_{\Psi_i}\|}{\sqrt{T}} + \frac{\|\sigma\|_2^2 + \sum_{i=0}^{c-1} \|\sigma_{\Psi_i}\| + Lc}{T}\right), \text{ where}$$
$$G(\delta) := \tilde{O}\left(\sum_{i=0}^{c-1} \|\sigma_{\Psi_i}\|^4 + \sigma_{\max} \|\sigma\|_2^2 + cL + c^{3/2}\sigma_{\max}\right),$$

and $\|\sigma\|_2^2 = \sum_{i=1}^{d} \sigma_i^2$ and $\|\sigma_{\Psi_i}\|^2 = \sum_{j \in \Psi_i} \sigma_j^2$.

Polylog terms are hidden in Theorem 1 for simplicity. The full theorem (Theorem 2) and proofs are deferred to Appendix D. Theorem 1 provides guarantee for all partitions of the parameters into arbitrary disjoint subsets and generalizes AdaGrad-Norm ($c = 1$) and AdaGrad-Coordinate ($c = d$) results. The result is noise-adapted: if $\sum_{i=0}^{c-1} \|\sigma_{\Psi_i}\|$ is small enough, the rate becomes the optimal deterministic rate of $O(\frac{1}{T})$. The next section explores implications of Theorem 1.

3.2. Coordinate-noise sparsity and dimension dependency

Theorem 1 presents trade-offs between the number of subsets c, and stochastic gradient noise. Intuitively, if few coordinates contribute to the total noise, the scalar version is more useful as $\|\sigma_{\Psi_i}\|^2$ is small for most subsets. However, when many coordinates contribute to the noise, $\|\sigma_{\Psi_i}\|^2$ can be large for many subsets and become the dominating term.

Coordinate-noise sparsity d^β. To make the intuition above concrete, consider the scenario with various coordinate-noise sparsity rate: for rate $\beta \in [0, 1]$, some d^β coordinates have noise $\alpha > 0$ while the rest are 0. When $\beta = 0$, we only have 1 coordinate with noise. When $\beta = 1$, all coordinates have noise. The rate β controls the density of coordinate noise. Furthermore, α upper bounds all coordinate noise, i.e. $\|\sigma\|_\infty \leq \alpha$, which is common in coordinate-wise analysis [5].

Table 1: Dimension dependency versus convergence rate under various coordinate-noise sparsity. Given a sparsity rate $\beta \in [0, 1]$, convergence rates are highlighted in red and green to denote the worst and best dependency on the dimension d, respectively. Note that memory usage of AdaGrad-Coordinate and AdaGrad-Norm is $O(d)$ and $O(1)$ while AdaGrad-Subset-Norm (with the equal partition strategy) is $O(d/k)$, where $k = d^{1.4\beta-0.6}$ is the noise dependent subset size.

Sparsity	AdaGrad-Coordinate	AdaGrad-Norm	AdaGrad-Subset-Norm (equal partition)
$\beta \in [0, 1]$	$\tilde{O}\left(d^{1.5+\beta}/\sqrt{T} + d^{2.5}/T\right)$	$\tilde{O}\left(d^{2.5\beta}/\sqrt{T} + d^{3\beta}/T\right)$	$\tilde{O}\left(d^{0.3+1.8\beta}/\sqrt{T} + d^{\beta+1}/T\right)$ if $\beta \in [0, 2/3]$ $\tilde{O}\left(d^{0.3+1.8\beta}/\sqrt{T} + d^{1.6\beta+0.6}/T\right)$ if $\beta \in [2/3, 1]$
$\beta = 0$	$\tilde{O}\left(d^{1.5}/\sqrt{T} + d^{2.5}/T\right)$	$\tilde{O}\left(1/\sqrt{T} + 1/T\right)$	$\tilde{O}\left(d^{0.3}/\sqrt{T} + d/T\right)$
$\beta = 0.5$	$\tilde{O}\left(d^{2}/\sqrt{T} + d^{2.5}/T\right)$	$\tilde{O}\left(d^{1.25}/\sqrt{T} + d^{1.5}/T\right)$	$\tilde{O}\left(d^{1.2}/\sqrt{T} + d^{1.5}/T\right)$
$\beta = 1$	$\tilde{O}\left(d^{2.5}/\sqrt{T} + d^{2.5}/T\right)$	$\tilde{O}\left(d^{2.5}/\sqrt{T} + d^{3}/T\right)$	$\tilde{O}\left(d^{2.1}/\sqrt{T} + d^{2.2}/T\right)$

Derivation of convergence rate given coordinate noise sparsity d^β. Given $\beta \in [0, 1]$, we can obtain a concrete expression for the convergence rates of various methods (different subset size) from Theorem 1. For AdaGrad-Subset-Norm, we consider an *equal partition strategy*, where we divide the coordinates into $c = d^{1-\beta}k$ subsets of size d^β/k each with the d^β noisy coordinates into just k subsets so that the rest of the $c - k$ subsets have no noisy coordinate. We defer the derivation details to Appendix C and summarize the results in the first row of Table 1.

Discussions. In Table 1, the equal subset-size partition strategy for AdaGrad-Subset-Norm has much better dependency on the dimension when the noise is not completely sparse i.e. $\beta = 0$. Hence, if we expect the actual noise sparsity β to be around 0.75[1], then compressing with a subset size of around $d^{0.45}$ is optimal. The dependency on d is important for modern neural network due to the number of parameters d being much greater than the total number of iterations T.

4. Experiments

We perform LLMs pre-training experiments on *Adam-Subset-Norm* (AdamSN) and *RMSProp-Subset-Norm* (RMSPropSN), where we replace the second moment term of Adam [16] and RMSProp[2] [29] with the subset-norm (SN) adaptive step size (Algorithms 4 and 5 in Appendix B.3). We use a simple subset partitioning scheme with no additional hyperparameter: for $p \in \mathbb{R}^{m\times n}$, the adaptive step size state is set to $\max(m, n)$. This compression scheme maintains the norm of the larger dimension and aims for the rough $d^{0.45}$ subset size as discussed in Section 3.2.

Setup. We test our method on the task of pre-training LLaMA models [8, 30] on the C4 dataset [26]. All of our experiments are conducted on NVIDIA RTX4090/3090 GPUs. We follow the experimental setup as in GaLore [36]. Hyperparameter details are presented in Appendix B.1.

Results. Table 2 contains the main results on LLaMA 60M and LLaMA 130M, where we compare against Adam [16], and memory efficient methods like RMSProp [29], and GaLore [36].

1. This is a prior implicitly imposed when selecting a subset size, where one should empirically estimate the actual noise sparsity rate.
2. Note that we consider the version of RMSProp *with* a bias correction term, which is equivalent to Adam with $\beta_1 = 0$.

Table 2: Final validation perplexity for various optimizers for pre-training LLaMA.

Method	60M	130M	350M	Opt-size[a]
Adam	30.45	24.59	18.67	$2mn$
AdamSN	**29.75**	**22.90**	**17.49**	$mn + m$
RMSProp	35.51	25.94	20.01	mn
RMSPropSN	34.57	25.67	18.72	m
GaLore	34.73	25.31	18.95	$2mk$[b]

a. Opt-size = optimizer state memory for parameter of size $m \times n$ with $m \geq n$.
b. k is typically $n/4$.

Table 3: Peak GPU Memory Usage (Gb) for various model sizes, obtained with batch size 1 and activation checkpointing to measure the optimizer state footprint.

Model Size	Adam	AdamSN	RMSPropSN	GaLore
60M	2.26	2.14	2.03	2.27
130M	2.98	2.62	2.35	2.78
350M	5.40	4.13	3.37	4.09
1B	15.37	10.36	7.55	9.41
3B	OOM[a]	18.25	12.68	16.01

a. Max memory of RTX4090/3090 is 24Gb.

Discussions. In Table 2, for both 60M and 130M models, AdamSN performs the best while using less memory for the second moment state than Adam. For memory efficient methods, RMSPropSN's performance is competitive to other memory efficient methods like GaLore (fourth row), despite using much less memory (due to no momentum state). See Table 3 for memory footprint of the methods considered across different model sizes with batch size 1.

Finally, recent memory efficient methods for pre-training LLMs like GaLore [36], FLORA [12], and GRASS [22] require additional hyperparameter tuning (e.g. GaLore has 3 additional parameters including the scaling, rank k, and update gap) and computation (e.g. GaLore requires an expensive SVD computation $O(mn^2)$ every 200 steps). In contrast, the subset-norm step size only requires an additional norm computation, which can be fused to existing kernels for more efficiency.

5. Conclusion and Future Works

Our unified high-probability analysis for AdaGrad-Coordinate and AdaGrad-Norm, generalized to arbitrary subsets, yields a convergence rate with improved dimensional dependency and a smaller memory footprint. The proposed subset-norm adaptive step size shows promise in LLM pretraining, offering a memory-efficient alternative to traditional adaptive optimizers without performance loss or excessive additional tuning.

Future works. It is important to extend experiments to larger models to assess scalability. Some future directions include combining subset-norm adaptive step size with momentum compression techniques [12, 22, 36] for enhanced efficiency/performance and experimenting with different subset sizes. It will be interesting to obtain principled parameter-sharing scheme similarly to subset-norm but for the momentum term. Obtaining convergence results for other optimizers like Adam, and/or under affine smoothness [3, 32], affine noise [11, 13], heavy-tailed noise [24, 25, 34, 35] are also of great interest. These extensions will broaden the applicability of our findings and potentially lead to more robust, efficient training algorithms across diverse machine learning applications.

References

[1] Rohan Anil, Vineet Gupta, Tomer Koren, and Yoram Singer. Memory efficient adaptive optimization. *Advances in Neural Information Processing Systems*, 32, 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):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*, pages 1147–1171. PMLR, 2023.

[4] Xiangyi Chen, Sijia Liu, Ruoyu Sun, and Mingyi Hong. On the convergence of a class of adam-type algorithms for non-convex optimization. In *International Conference on Learning Representations*, 2018.

[5] Alexandre Défossez, Léon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of adam and adagrad. *Transactions on Machine Learning Research*, 2022.

[6] Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 8-bit optimizers via block-wise quantization. *arXiv preprint arXiv:2110.02861*, 2021.

[7] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36, 2024.

[8] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[9] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[10] Alina Ene and Huy L Nguyen. Adaptive and universal algorithms for variational inequalities with optimal convergence s. *arXiv preprint arXiv:2010.07799*, 2021.

[11] 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. *arXiv preprint arXiv:2202.05791*, 2022.

[12] Yongchang Hao, Yanshuai Cao, and Lili Mou. Flora: Low-rank adapters are secretly gradient compressors. *arXiv preprint arXiv:2402.03293*, 2024.

[13] Yusu Hong and Junhong Lin. Revisiting convergence of adagrad with relaxed assumptions. *arXiv preprint arXiv:2402.13794*, 2024.

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[15] 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*, 2021.

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[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, Alexander Rakhlin, and Ali Jadbabaie. Convergence of adam under relaxed assumptions. *Advances in Neural Information Processing Systems*, 36, 2024.

[19] Vladislav Lialin, Sherin Muckatira, Namrata Shivagunde, and Anna Rumshisky. Relora: High-rank training through low-rank updates. In *The Twelfth International Conference on Learning Representations*, 2023.

[20] Zijian Liu, Ta Duy Nguyen, Alina Ene, and Huy Nguyen. On the convergence of adagrad(norm) on $\mathbb{R}^d$: Beyond convexity, non-asymptotic rate and acceleration. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ULnHxczCBaE.

[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*, pages 21884–21914. PMLR, 2023.

[22] Aashiq Muhamed, Oscar Li, David Woodruff, Mona Diab, and Virginia Smith. Grass: Compute efficient low-memory llm training with structured sparse gradients. *arXiv preprint arXiv:2406.17660*, 2024.

[23] Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence o (1/kˆ 2). In *Doklady an ussr*, volume 269, pages 543–547, 1983.

[24] Ta Duy Nguyen, Thien H Nguyen, Alina Ene, and Huy Nguyen. Improved convergence in high probability of clipped gradient methods with heavy tailed noise. *Advances in Neural Information Processing Systems*, 36:24191–24222, 2023.

[25] Ta Duy Nguyen, Thien Hang Nguyen, Alina Ene, and Huy Le Nguyen. High probability convergence of clipped-sgd under heavy-tailed noise. *arXiv preprint arXiv:2302.05437*, 2023.

[26] 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, 2023. URL https://arxiv.org/abs/1910.10683.

[27] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018.

[28] Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR, 2018.

[29] 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*, 4(2): 26–31, 2012.

[30] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[31] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[32] Bohan Wang, Huishuai Zhang, Zhiming Ma, and Wei Chen. Convergence of adagrad for non-convex objectives: Simple proofs and relaxed assumptions. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 161–190. PMLR, 2023.

[33] Rachel Ward, Xiaoxia Wu, and Leon Bottou. Adagrad stepsizes: Sharp convergence over nonconvex landscapes. In *International Conference on Machine Learning*, pages 6677–6686. PMLR, 2019.

[34] Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. *arXiv preprint arXiv:1905.11881*, 2019.

[35] Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems (NeurIPS)*, 33:15383–15393, 2020.

[36] Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. Galore: Memory-efficient llm training by gradient low-rank projection. *arXiv preprint arXiv:2403.03507*, 2024.

[37] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 11127–11135, 2019.

Table 4: Learning rate obtained from grid search. We find the best learning for 60M and found that it works similarly well for the 130M model (as compared to experiments from [36] that have tuned the learning rate for larger models)

	Adam	AdamSN	GaLore	RMSPropSN	RMSProp
Learning rate 60M	0.005	0.05	0.01	0.01	0.001
Learning rate 130M	0.005	0.05	0.01	0.01	0.001

Appendix A. Related Works

Convergence analysis of non-convex optimization methods has seen significant progress, with recent works providing convergence proofs for adaptive algorithms like Adam [5, 16, 18]. Numerous studies have explored convergence properties of various adaptive and stochastic gradient methods [4, 5, 10, 20, 21, 23, 27, 33, 37], while lower bound analyses [2] have illuminated fundamental limitations in non-convex optimization.

As model sizes grow, memory-efficient training techniques have become crucial. Following up on AdaFactor [28], low-rank decomposition methods like Galore [36], LoRA [12], and ReLORA [19] approximate large weight matrices with lower-rank representations. Projection-based approaches, such as GRASS [22] and Flora [12], compress gradients or combine low-rank ideas with projections to reduce memory requirements. A related but different method from ours is SM3 [1] where subset (cover) statistics are used to show convergence in the context of online learning. These techniques align with our work's goal of enhancing memory efficiency in adaptive optimization methods for large-scale machine learning models.

Appendix B. Additional Experimental Details

B.1. Hyperparameter details

In Table 2, we run all experiments on BF16 format, weight decay of 0, gradient clipping of 1.0, cosine learning rate decay to 10% of the max learning rate with 10% linear warmup steps, and batch size of 512. We only tune for the learning rate across a grid of $\{0.1, 0.05, 0.01, 0.005, 0.001\}$. We train for 10,000 steps and 20,000 steps for the 60M and 130M models, respectively. Table 4 shows the learning rate obtained for each method which is used across both the 60M and 130M model's experiments.

B.2. AdaGrad, AdaGrad-Norm, and AdaGrad-Subset-Norm

We examine the subset-norm step size for AdaGrad in Figure 1. We again see that subset-norm is slightly better than the full coordinate version while using a lot less memory. This is consistent with our observations for Adam and RMSProp when we replace the standard coordinate-wise step size with the subset-norm adaptive step size.

B.3. Adam-Subset-Norm Implementation

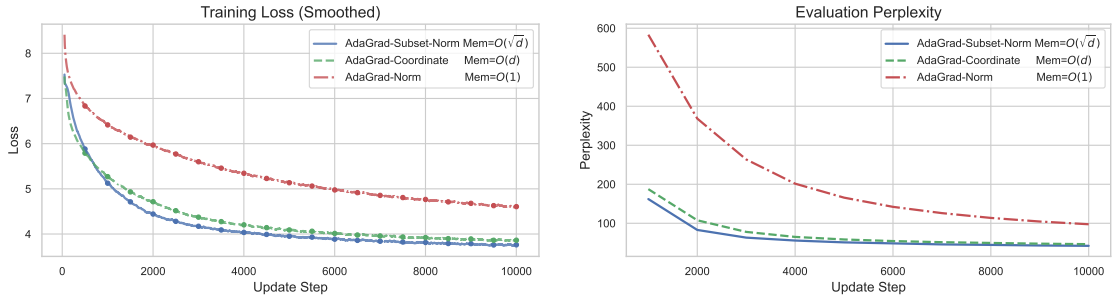
Algorithm 4 presents the pseudocode for Adam-Subset-Norm as mentioned in Section 4.

Algorithm 4: Adam-Subset-Norm with a simple partitioning scheme

Input: Learning rate η, EMA parameters β_1 and β_2, $\epsilon > 0$, optional weight decay $wd \geq 0$
Output: Updated parameters
for $p \in \mathbb{R}^{m \times n}$ *in params* **do**
  grad $\leftarrow p$.grad;
  $r \leftarrow 0$ if $m \geq n$ else 1;
  $k \leftarrow p$.shape[r] ; // where $k = m$ if $r = 0$ else $k = n$
  gradN $\leftarrow$ grad.norm(dim=$1 - r$) $\in \mathbb{R}^k$; // subset norm
  $m \leftarrow \beta_1 m + (1 - \beta_1)\cdot$ grad $\in \mathbb{R}^{m \times n}$;
  $v \leftarrow \beta_2 v + (1 - \beta_2) \cdot$ gradN$^2 \in \mathbb{R}^k$; // omitting bias correction terms
  $p \leftarrow p + \eta \frac{m}{\sqrt{v}+\epsilon}$; // broadcast division
  $p \leftarrow p - \eta \cdot wd$; // weight decay
end

Algorithm 5: RMSProp-Subset-Norm with a simple partitioning scheme

Input: Learning rate η, EMA parameter β, $\epsilon > 0$, optional weight decay $\kappa \geq 0$
Output: Updated parameters
for $p \in \mathbb{R}^{m \times n}$ *in params* **do**
  grad $\leftarrow p$.grad;
  $r \leftarrow 0$ if $m \geq n$ else 1;
  $k \leftarrow p$.shape[r] ; // where $k = m$ if $r = 0$ else $k = n$
  gradN $\leftarrow$ grad.norm(dim=$1 - r$) $\in \mathbb{R}^k$; // subset norm
  $v \leftarrow \beta \cdot v + (1 - \beta) \cdot$ gradN$^2 \in \mathbb{R}^k$;
  $p \leftarrow p + \eta \frac{\text{grad}}{\sqrt{v}+\epsilon}$; // broadcast division
  $p \leftarrow p - \eta \cdot \kappa$; // weight decay
end



(*a*) Training loss (10,000 steps) (*b*) Validation perplexity (10,000 steps)

Figure 1: Pretraining LLaMA 60M on the C4 dataset for AdaGrad variants. Memory consumption estimate as a function of parameter count d is shown in the legend.

Algorithm 6: Generic Subset-Norm Adaptive Step Size Update Rule (PyTorch-y notation)

Input: Parameter $P \in \mathbb{R}^{m \times n}$, step size $\eta > 0$, β, and $\epsilon > 0$, and partition size k such that k divides mn

$R \leftarrow (\nabla P).\texttt{reshape}(m \times n/k, k)$; `// Reshape gradient into shape` $\frac{mn}{k} \times k$

$V \leftarrow \beta V + (1-\beta) \cdot (R.\texttt{sum(dim=1)})$; `// Update state` V `via subset norm reduction on dim 1`

$U \leftarrow \frac{R}{\sqrt{V}+\epsilon} \in \mathbb{R}^{\frac{mn}{k} \times k}$; `// Broadcast addition and division for update step`

$P \leftarrow P - \eta \cdot U.\texttt{view}(m, n)$; `// Reshape` U `back to` $\mathbb{R}^{m \times n}$ `and update` P

B.4. Generic Subset-Norm Adaptive Step Size Implementation

The implementations above is simple and does not require any tuning. To modify existing algorithms to work with arbitrary subsets, one could utilize reshape as in Algorithm 6 for RMSProp as an example.

Appendix C. Coordinate-noise sparsity convergence rate derivation

AdaGrad-Coordinate. For $c = d$ (AdaGrad-Coordinate), we get $\sum_{i=0}^{c-1} \|\sigma_{\Psi_i}\| = \alpha d^\beta$, $\|\sigma\|_2^2 = \alpha^2 d^\beta$, and $\sum_{i=0}^{c-1} \|\sigma_{\Psi_i}\|^4 = \alpha^4 d^\beta$, so our bound is

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla_t\|_2^2 \leq \tilde{O}\left(\alpha^4 d^\beta + \alpha^3 d^\beta + dL + d^{1.5}\alpha\right) \cdot \tilde{O}\left(\frac{\alpha d^\beta}{\sqrt{T}} + \frac{\alpha^2 d^\beta + \alpha d^\beta + Ld}{T}\right).$$

The dependency on d for the slow term $O(1/\sqrt{T})$ is $d^{1.5} d^\beta = d^{1.5+\beta}$. The dependency on d for the fast term $O(1/T)$ is $d^{1.5} d = d^{2.5}$. Note that there is an inherent $d^{1.5}$ dependency for the slow term that does not reduce as the coordinate-noise density decrease.

AdaGrad-Norm For $c = 1$ (AdaGrad-Norm), we get $\|\sigma\|_2^2 = \sum_{i=0}^{d} \|\sigma_i\|^2 = \alpha^2 d^\beta$, $\|\sigma\|_2 = \alpha d^{\beta/2}$, and $\|\sigma\|^4 = \alpha^4 d^{2\beta}$. This means that our bound is

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla_t\|_2^2 \leq \tilde{O}\left(\alpha^4 d^{2\beta} + \alpha^3 d^\beta + L + \alpha\right) \cdot \tilde{O}\left(\frac{\alpha d^{\beta/2}}{\sqrt{T}} + \frac{\alpha^2 d^\beta + \alpha d^{\beta/2} + L}{T}\right).$$

The dependency on d for the slow term $O(1/\sqrt{T})$ is $d^{2\beta} \cdot d^{\beta/2} = d^{2.5\beta}$. The dependency on d for the fast term $O(1/T)$ is $d^{2\beta} \cdot d^\beta = d^{3\beta}$. Note that when $\beta = 0$, or when all the noise is on a single coordinate, we recover the dimension-free results of previous works.

AdaGrad-Subset-Norm. Now, consider the following partition strategy, where we divide the coordinates into $c = d^{1-\beta}k$ subsets of size d^β/k each with the d^β noisy coordinates into just k subsets so that the rest of the $c - k$ subsets do not contain any noisy coordinate. We have $\left\|\sigma_{\Psi_j}\right\|_2^2 = \alpha^2 d^\beta/k \implies \left\|\sigma_{\Psi_j}\right\|_2 = \alpha d^{\beta/2}/k^{0.5}$ if j is a noisy subset. We can compute $\sum_{i=0}^{c-1} \|\sigma_{\Psi_i}\| = \alpha d^{\beta/2} k^{0.5}$, $\|\sigma\|_2^2 = \sum_{i=0}^{c-1} \|\sigma_{\Psi_i}\|_2^2 = \alpha^2 d^\beta$, and $\sum_{i=0}^{c-1} \|\sigma_{\Psi_i}\|^4 = \alpha^4 d^{2\beta}/k$. We get a bound of

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla_t\|_2^2 \leq \tilde{O}\left(\alpha^4 d^{2\beta}/k + \alpha^3 d^\beta + d^{1-\beta}kL + \left(d^{1-\beta}k\right)^{3/2}\alpha\right) \cdot \tilde{O}\left(\frac{\alpha d^{\beta/2}k^{0.5}}{\sqrt{T}} + \frac{\alpha^2 d^\beta + \alpha d^{\beta/2}k^{0.5} + Ld^{1-\beta}k}{T}\right).$$

Set $k = d^{7\beta/5-3/5}$ so that $\left(d^{1-\beta}k\right)^{3/2} = d^{2\beta}/k = d^{3\beta/5+3/5}$. Then we can simplify

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla_t\|_2^2 \leq \tilde{O}\left(\alpha^4 d^{3(\beta+1)/5} + \alpha^3 d^\beta + d^{2(\beta+1)/5}L + d^{3(\beta+1)/5}\alpha\right) \cdot \tilde{O}\left(\frac{\alpha d^{(12\beta-3)/10}}{\sqrt{T}} + \frac{\alpha^2 d^\beta + \alpha d^{(12\beta-3)/10} + Ld^{2(\beta+1)/5}}{T}\right).$$

The dependency on d for the slow term $O(1/\sqrt{T})$ is $d^{3(\beta+1)/5} \cdot d^{(12\beta-3)/10} = d^{3(1+6\beta)/10} = d^{0.3+1.8\beta}$. The dependency on d for the fast term $O(1/T)$ is a bit more complicated: For $\beta \in [0, \frac{2}{3}]$, we have the dependency on d is $d^{3(\beta+1)/5} \cdot d^{2(\beta+1)/5} = d^{\beta+1}$. For $\beta \in [\frac{2}{3}, 1]$, we have the dependency on d is $d^{3(\beta+1)/5} \cdot d^\beta = d^{3(\beta+1)/5+\beta} = d^{1.6\beta+0.6}$. Note that this is only a possible partition strategy where the subset sizes are of equal size (which is probably the most natural and easiest to implement). There, the optimal subset size is $k = d^{1.4\beta-0.6}$, for which if we plug in $\beta \in [0, 1]$ we get a range from 1 to $d^{0.8}$.

Appendix D. Full Theorem and Proof

We show the full result in Theorem 2 with all the polylog terms omitted from Theorem 1.

Theorem 2 *Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and lower bounded by f_*. Given unbiased stochastic gradients $\widehat{\nabla} f(x_t)$ with stochastic gradient noise $\xi_t := \widehat{\nabla} f(x_t) - \nabla f(x_t)$ being σ_i-per-coordinate subgaussian for $i \in [d]$. For partitions of the parameters into disjoint subsets $[d] =$*

$\bigcup_{i=0}^{c-1} \Psi_i$ *with* $\Psi_i \cap \Psi_j = \emptyset$, *if* $i \neq j$, *the iterates* x_t *given by (2) satisfies the following inequality with probability at least* $1 - 6c\delta$ *(for failure probability* $\delta > 0$*):*

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla_t\|_2^2 \leq G(\delta) \cdot \left(\frac{4\sum_{i=0}^{c-1} \|\sigma_{\Psi_i}\|}{\sqrt{T}} + \frac{I(\delta)}{T} \right), \text{ where } G(\delta) \text{ and } I(\delta) \text{ are polylog terms:}$$

$$G(\delta) := \frac{\Delta_1}{\eta} + H(\delta) + \left(\ln T/\delta \, \|\sigma\|_2^2 + c\eta L + 4c^{3/2}\sigma_{\max}\sqrt{\log\frac{1}{\delta}} \right) \log\left(\frac{4\sqrt{T}\sum_{i=0}^{c-1} \|\sigma_{\Psi_i}\| + I(\delta)}{b_{0,\min}} \right)$$

$$I(\delta) := \|b_0\|_1 + \frac{2\Delta_1}{\eta} + \frac{8\log\frac{1}{\delta}}{b_{0,\min}} \|\sigma\|_2^2 + \sqrt{\log\frac{1}{\delta}} \sum_{i=0}^{c-1} \|\sigma_{\Psi_i}\| + 8\eta L c \log\frac{4\eta L}{b_{0,\min}}$$

$$H(\delta) := \sum_{i=0}^{c-1} \left(\ln(T/\delta) \, \|\sigma_{\Psi_i}\|^2 + 2\alpha \right) \left(\frac{8\|\sigma_{\Psi_i}\|^2 \log\frac{1}{\delta}}{b_{0,i}^2} + 2\log\left(1 + \|\sigma_{\Psi_i}\|^2 T + \|\sigma_{\Psi_i}\|^2 \log\frac{1}{\delta} \right) \right).$$

where $\|\sigma\|_2^2 = \sum_{i=1}^{d} \sigma_i^2$, $\|\sigma_{\Psi_i}\|^2 = \sum_{j\in\Psi_i} \sigma_j^2$, $\sigma_{\max} = \max_{i\in[d]} \sigma_i$, $\Delta_1 = f(x_1) - f_*$, $b_{0,\min} = \min_{i\in[d]} b_{0,i} > 0$.

D.1. Proof of Theorem 2

For simplicity, in our analysis, we will use $\widehat{\nabla}_{t,i} := \widehat{\nabla}_i f(x_t)$ and $\nabla_{t,i} := \nabla_i f(x_t)$ to denote the i-th coordinate of the stochastic gradients and gradients at iterate t, respectively. The proof utilizes techniques and follows the strategies [21], where the main effort is to adapt the techniques for handling subsets from the AdaGrad-Norm and AdaGrad-Coordinate proofs in [21].

Proof We write $\frac{\widehat{\nabla}_t}{b_t}$ to denote $\left(\frac{\widehat{\nabla}_t}{b_t}\right)_k = \frac{\widehat{\nabla}_k f(x_t)}{b_{t,i}}$ for $k \in \Psi_i$ (we will use this notation briefly to show some steps and will not be crucial in the main analysis). We start with the smoothness of f and $\Delta_t := f(x_t) - f_*$.

$$\begin{aligned}
\Delta_{t+1} - \Delta_t &\leq \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2 \\
&= -\eta \left\langle \nabla_t, \frac{\widehat{\nabla}_t}{b_t} \right\rangle + \frac{\eta^2 L}{2} \left\| \frac{\widehat{\nabla}_t}{b_t} \right\|^2 \qquad (3) \\
&= -\eta \sum_{i=0}^{c-1} \sum_{j\in\Psi_i} \frac{\nabla_{t,j}\widehat{\nabla}_{t,j}}{b_{t,i}} + \frac{\eta^2 L}{2} \sum_{i=0}^{c-1} \sum_{j\in\Psi_i} \frac{\widehat{\nabla}_{t,j}^2}{b_{t,i}^2} \\
&= -\eta \sum_{i=0}^{c-1} \sum_{j\in\Psi_i} \frac{\nabla_{t,j}(\xi_{t,j} + \nabla_{t,j})}{b_{t,i}} + \frac{\eta^2 L}{2} \sum_{i=0}^{c-1} \sum_{j\in\Psi_i} \frac{\widehat{\nabla}_{t,j}^2}{b_{t,i}^2} \qquad (\xi_{t,i} = \widehat{\nabla}_{t,i} - \nabla_{t,i}) \\
&= -\eta \sum_{i=0}^{c-1} \sum_{j\in\Psi_i} \frac{\nabla_{t,j}^2}{b_{t,i}} - \eta \sum_{i=0}^{c-1} \sum_{j\in\Psi_i} \frac{\nabla_{t,j}\xi_{t,j}}{b_{t,i}} + \frac{\eta^2 L}{2} \sum_{i=0}^{c-1} \sum_{j\in\Psi_i} \frac{\widehat{\nabla}_{t,j}^2}{b_{t,i}^2} \\
&= -\eta \sum_{i=0}^{c-1} \sum_{j\in\Psi_i} \frac{\nabla_{t,j}^2}{b_{t,i}} - \eta \sum_{i=0}^{c-1} \sum_{j\in\Psi_i} \frac{\nabla_{t,j}\xi_{t,j}}{a_{t,i}} + \eta \sum_{i=0}^{c-1} \sum_{j\in\Psi_i} \left(\frac{1}{a_{t,i}} - \frac{1}{b_{t,i}} \right) \nabla_{t,j}\xi_{t,j} + \frac{\eta^2 L}{2} \sum_{i=0}^{c-1} \sum_{j\in\Psi_i} \frac{\widehat{\nabla}_{t,j}^2}{b_{t,i}^2}. \qquad (4)
\end{aligned}$$

Now, we analyze $\frac{1}{a_{t,i}} - \frac{1}{b_{t,i}}$ for $i = 0, 1, \ldots, c-1$:

$$\begin{aligned}
\left|\frac{1}{a_{t,i}} - \frac{1}{b_{t,i}}\right| &= \left|\frac{b_{t,i} - a_{t,i}}{a_{t,i} b_{t,i}}\right| \\
&= \left|\frac{b_{t,i}^2 - a_{t,i}^2}{a_{t,i} b_{t,i} \left(b_{t,i} + a_{t,i}\right)}\right| \\
&= \left|\frac{b_{t-1,i}^2 + \left\|\widehat{\nabla}_{\Psi_i} f(x_t)\right\|^2 - b_{t-1,i}^2 - \left\|\nabla_{\Psi_i} f(x_t)\right\|^2}{a_{t,i} b_{t,i} \left(b_{t,i} + a_{t,i}\right)}\right| \\
&= \left|\frac{\left\|\widehat{\nabla}_{\Psi_i} f(x_t)\right\|^2 - \left\|\nabla_{\Psi_i} f(x_t)\right\|^2}{a_{t,i} b_{t,i} \left(b_{t,i} + a_{t,i}\right)}\right| \\
&= \left|\frac{\left(\left\|\widehat{\nabla}_{\Psi_i} f(x_t)\right\| - \left\|\nabla_{\Psi_i} f(x_t)\right\|\right)\left(\left\|\widehat{\nabla}_{\Psi_i} f(x_t)\right\| + \left\|\nabla_{\Psi_i} f(x_t)\right\|\right)}{a_{t,i} b_{t,i} \left(b_{t,i} + a_{t,i}\right)}\right|.
\end{aligned}$$

Since $b_{t,i} = \sqrt{b_{t-1,i}^2 + \left\|\widehat{\nabla}_{\Psi_i} f(x_t)\right\|^2} \geq \left\|\widehat{\nabla}_{\Psi_i} f(x_t)\right\|$ and $a_{t,i} = \sqrt{b_{t-1,i}^2 + \left\|\nabla_{\Psi_i} f(x_t)\right\|^2} \geq \left\|\nabla_{\Psi_i} f(x_t)\right\|$, we have

$$\begin{aligned}
\left|\frac{1}{a_{t,i}} - \frac{1}{b_{t,i}}\right| &\leq \left|\frac{\left(\left\|\widehat{\nabla}_{\Psi_i} f(x_t)\right\| - \left\|\nabla_{\Psi_i} f(x_t)\right\|\right)\left(\left\|\widehat{\nabla}_{\Psi_i} f(x_t)\right\| + \left\|\nabla_{\Psi_i} f(x_t)\right\|\right)}{a_{t,i} b_{t,i} \left(\left\|\widehat{\nabla}_{\Psi_i} f(x_t)\right\| + \left\|\nabla_{\Psi_i} f(x_t)\right\|\right)}\right| \\
&\leq \left|\frac{\left\|\widehat{\nabla}_{\Psi_i} f(x_t)\right\| - \left\|\nabla_{\Psi_i} f(x_t)\right\|}{a_{t,i} b_{t,i}}\right| \\
&\leq \frac{\left\|\widehat{\nabla}_{\Psi_i} f(x_t) - \nabla_{\Psi_i} f(x_t)\right\|}{a_{t,i} b_{t,i}} \\
&= \frac{\left\|\xi_{t,\Psi_i}\right\|}{a_{t,i} b_{t,i}}.
\end{aligned}$$

Hence, we have

$$\left|\frac{1}{a_{t,i}} - \frac{1}{b_{t,i}}\right| \leq \frac{\left\|\xi_{t,\Psi_i}\right\|}{a_{t,i} b_{t,i}}.$$

Then from 4, taking the absolute value of $\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\left(\frac{1}{a_{t,i}}-\frac{1}{b_{t,i}}\right)\nabla_{t,j}\xi_{t,j}$, we can bound:

$$\begin{aligned}
\Delta_{t+1}-\Delta_t &\le -\eta\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{\nabla_{t,j}^2}{b_{t,i}} - \eta\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{\nabla_{t,j}\xi_{t,j}}{a_{t,i}} + \eta\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\left|\frac{1}{a_{t,i}}-\frac{1}{b_{t,i}}\right||\nabla_{t,j}\xi_{t,j}| + \frac{\eta^2L}{2}\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{\widehat{\nabla}_{t,j}^2}{b_{t,i}^2}\\
&\le -\eta\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{\nabla_{t,j}^2}{b_{t,i}} - \eta\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{\nabla_{t,j}\xi_{t,j}}{a_{t,i}} + \eta\sum_{i=0}^{c-1}\frac{\|\xi_{t,\Psi_i}\|}{a_{t,i}b_{t,i}}\sum_{j\in\Psi_i}|\nabla_{t,j}\xi_{t,j}| + \frac{\eta^2L}{2}\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{\widehat{\nabla}_{t,j}^2}{b_{t,i}^2}\\
&\overset{(1)}{\le} -\eta\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{\nabla_{t,j}^2}{b_{t,i}} - \eta\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{\nabla_{t,j}\xi_{t,j}}{a_{t,i}} + \eta\sum_{i=0}^{c-1}\frac{\|\xi_{t,\Psi_i}\|}{a_{t,i}b_{t,i}}\|\nabla_{t,\Psi_i}\|\,\|\xi_{t,\Psi_i}\| + \frac{\eta^2L}{2}\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{\widehat{\nabla}_{t,j}^2}{b_{t,i}^2}\\
&\le -\eta\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{\nabla_{t,j}^2}{b_{t,i}} - \eta\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{\nabla_{t,j}\xi_{t,j}}{a_{t,i}}\\
&\quad + \eta\sum_{i=0}^{c-1}\|\xi_{t,\Psi_i}\|\left(\frac{\|\xi_{t,\Psi_i}\|^2}{2b_{t,i}^2}+\frac{\|\nabla_{t,\Psi_i}\|^2}{2a_{t,i}^2}\right) + \frac{\eta^2L}{2}\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{\widehat{\nabla}_{t,j}^2}{b_{t,i}^2},
\end{aligned}$$

where (1) is due to $\sum_{j\in\Psi_i}|\nabla_{t,j}\xi_{t,j}| = \langle|\nabla_{t,\Psi_i}|,|\xi_{t,\Psi_i}|\rangle \le \|\nabla_{t,\Psi_i}\|\,\|\xi_{t,\Psi_i}\|$ and $|\cdot|$ denotes coordinate-wise absolute value when we apply to vectors. The last inequality is due to $2ab \le a^2+b^2$. Now, we can sum both sides for $t=1,\dots,T$ to telescope the LHS:

$$\begin{aligned}
\Delta_{T+1}-\Delta_1 \le \sum_{t=1}^{T}\Big(&-\eta\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{\nabla_{t,j}^2}{b_{t,i}} - \eta\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{\nabla_{t,j}\xi_{t,j}}{a_{t,i}}\\
&+ \eta\sum_{i=0}^{c-1}\|\xi_{t,\Psi_i}\|\left(\frac{\|\xi_{t,\Psi_i}\|^2}{2b_{t,i}^2}+\frac{\|\nabla_{t,\Psi_i}\|^2}{2a_{t,i}^2}\right) + \frac{\eta^2L}{2}\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{\widehat{\nabla}_{t,j}^2}{b_{t,i}^2}\Big).
\end{aligned}$$

Rearranging gives

$$\begin{aligned}
\sum_{t=1}^{T}\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{\nabla_{t,j}^2}{b_{t,i}} \le \frac{\Delta_1-\Delta_{T+1}}{\eta} &- \underbrace{\sum_{t=1}^{T}\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{\nabla_{t,j}\xi_{t,j}}{a_{t,i}}}_{A}\\
&+ \underbrace{\sum_{t=1}^{T}\sum_{i=0}^{c-1}\|\xi_{t,\Psi_i}\|\left(\frac{\|\xi_{t,\Psi_i}\|^2}{2b_{t,i}^2}+\frac{\|\nabla_{t,\Psi_i}\|^2}{2a_{t,i}^2}\right)}_{B} + \underbrace{\frac{\eta L}{2}\sum_{t=1}^{T}\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{\widehat{\nabla}_{t,j}^2}{b_{t,i}^2}}_{C}.
\end{aligned}$$

On the LHS, we note that

$$\sum_{t=1}^{T}\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{\nabla_{t,j}^2}{b_{t,i}} = \sum_{t=1}^{T}\sum_{i=0}^{c-1}\frac{\|\nabla_{t,\Psi_i}\|^2}{b_{t,i}}.$$

We now bound each term separately. It's easiest to bound C: $\sum_{t=1}^{T}\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{\widehat{\nabla}_{t,j}^2}{b_{t,i}^2}$:

$$\begin{aligned}
\sum_{t=1}^{T}\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{\widehat{\nabla}_{t,j}^2}{b_{t,i}^2} &= \sum_{i=0}^{c-1}\sum_{t=1}^{T}\sum_{j\in\Psi_i}\frac{\widehat{\nabla}_{t,j}^2}{b_{t,i}^2} = \sum_{i=1}^{d}\sum_{t=1}^{T}\frac{b_{t,i}^2-b_{t-1,i}^2}{b_{t,i}^2} \le \sum_{i=1}^{d}2\log\frac{b_{T,i}}{b_{0,i}}.\\
&= \sum_{i=0}^{c-1}\sum_{t=1}^{T}\frac{\left\|\widehat{\nabla}_{t,\Psi_i}\right\|^2}{b_{t,i}^2}\\
&= \sum_{i=0}^{c-1}\sum_{t=1}^{T}\frac{b_{t,i}^2-b_{t-1,i}^2}{b_{t,i}^2}\\
&= \sum_{i=0}^{c-1}\sum_{t=1}^{T}1-\frac{b_{t-1,i}^2}{b_{t,i}^2}\\
&\le \sum_{i=0}^{c-1}\sum_{t=1}^{T}\log\frac{b_{t,i}^2}{b_{t-1,i}^2}\\
&= 2\sum_{i=0}^{c-1}\log\prod_{t=1}^{T}\frac{b_{t,i}}{b_{t-1,i}}\\
&= 2\sum_{i=0}^{c-1}\log\frac{b_{T,i}}{b_{0,i}}.
\end{aligned}$$

We now have a useful inequality

$$\sum_{t=1}^{T}\frac{\left\|\widehat{\nabla}_{t,\Psi_i}\right\|^2}{b_{t,i}^2} \le 2\log\frac{b_{T,i}}{b_{0,i}},\ \forall i=0,\dots,c-1. \tag{5}$$

Next, we deal with $-\sum_{t=1}^{T}\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{\nabla_{t,j}\xi_{t,j}}{a_{t,i}}$ via a martingale argument. Let $\mathcal{F}_t := \sigma\left(\xi_1,\dots,\xi_{t-1}\right)$ denote the natural filtration. Note that x_t is $\mathcal{F}_t$-measurable. For any $w>0$, we have for each $i\in[c]$:

$$\begin{aligned}
&\mathbb{E}\left[\exp\left(-w\sum_{j\in\Psi_i}\frac{\nabla_{t,j}\xi_{t,j}}{a_{t,i}}-2w^2\sum_{j\in\Psi_i}\frac{\sigma_j^2\nabla_{t,j}^2}{a_{t,i}^2}\right)\mid\mathcal{F}_t\right]\\
&=\exp\left(-2w^2\sum_{j\in\Psi_i}\frac{\sigma_j^2\nabla_{t,j}^2}{a_{t,i}^2}\right)\mathbb{E}\left[\exp\left(-w\sum_{j\in\Psi_i}\frac{\nabla_{t,j}\xi_{t,j}}{a_{t,i}}\right)\mid\mathcal{F}_t\right]\\
&\le 1.
\end{aligned}$$

Then a simple inductive argument and using Markov's inequality gives with probability at least $1-\delta$:

$$-w\sum_{t=1}^{T}\sum_{j\in\Psi_i}\frac{\nabla_{t,j}\xi_{t,j}}{a_{t,i}} \le 2w^2\sum_{t=1}^{T}\sum_{j\in\Psi_i}\frac{\sigma_j^2\nabla_{t,j}^2}{a_{t,i}^2}+\log\frac{1}{\delta}.$$

By a union bound across all c subsets, we have w.p. at least $1 - c\delta$:

$$-\sum_{t=1}^{T}\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{\nabla_{t,j}\xi_{t,j}}{a_{t,i}} \le \sum_{t=1}^{T}\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{w\sigma_j^2\nabla_{t,j}^2}{a_{t,i}^2} + \frac{c}{w}\log\frac{1}{\delta}. \tag{6}$$

Let's call the event that (6) happens E_1. Now, consider $\sum_{t=1}^{T}\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{\nabla_{t,j}^2}{a_{t,i}^2}$. We have

$$\begin{aligned}
\sum_{j\in\Psi_i}\frac{\nabla_{t,j}^2}{a_{t,i}^2} &= \frac{\|\nabla_{t,\Psi_i}\|^2}{a_{t,i}^2} = \frac{\|\nabla_{t,\Psi_i}\|^2}{b_{t-1,i}^2 + \|\nabla_{t,\Psi_i}\|^2} \\
&\overset{(*)}{\le} \frac{2\left\|\widehat{\nabla}_{t,\Psi_i}\right\|^2 + 2\|\xi_{t,\Psi_i}\|^2}{b_{t-1,i}^2 + 2\left\|\widehat{\nabla}_{t,\Psi_i}\right\|^2 + 2\|\xi_{t,\Psi_i}\|^2} \\
\frac{\|\nabla_{t,\Psi_i}\|^2}{a_{t,i}^2} &\le 2\frac{\left\|\widehat{\nabla}_{t,\Psi_i}\right\|^2}{b_{t,i}^2} + 2\frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2}.
\end{aligned}$$

For $(*)$ we use the fact that $\frac{x}{c+x}$ is an increasing function and $\|\nabla_{t,\Psi_i}\|^2 = \left\|\widehat{\nabla}_{t,\Psi_i} + \xi_{t,\Psi_i}\right\|^2 \le 2\left\|\widehat{\nabla}_{t,\Psi_i}\right\|^2 + 2\|\xi_{t,\Psi_i}\|^2$. Let $\sigma_{\max} := \max_{i\in[d]}\sigma_i$, then under event E_1, we have with probability at least $1 - c\delta$:

$$\begin{aligned}
-\sum_{t=1}^{T}\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{\nabla_{t,j}\xi_{t,j}}{a_{t,i}} &\le \sum_{t=1}^{T}\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{w\sigma_j^2\nabla_{t,j}^2}{a_{t,i}^2} + \frac{c}{w}\log\frac{1}{\delta} \\
&\le w\sigma_{\max}^2\sum_{t=1}^{T}\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{\nabla_{t,j}^2}{a_{t,i}^2} + \frac{c}{w}\log\frac{1}{\delta} \\
&\le w\sigma_{\max}^2\sum_{t=1}^{T}\sum_{i=0}^{c-1}\left(2\frac{\left\|\widehat{\nabla}_{t,\Psi_i}\right\|^2}{b_{t,i}^2} + 2\frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2}\right) + \frac{c}{w}\log\frac{1}{\delta} \\
&= \underbrace{\sigma_{\max}\sqrt{c\log\frac{1}{\delta}}}_{=:\alpha}\sum_{t=1}^{T}\sum_{i=0}^{c-1}\left(2\frac{\left\|\widehat{\nabla}_{t,\Psi_i}\right\|^2}{b_{t,i}^2} + 2\frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2}\right) + \sigma_{\max}\sqrt{c\log\frac{1}{\delta}} \\
&\qquad\qquad \left(\text{set } w := \frac{\sqrt{c\log\frac{1}{\delta}}}{\sigma_{\max}}\right) \\
&= 2\alpha\sum_{t=1}^{T}\sum_{i=0}^{c-1}\left(\frac{\left\|\widehat{\nabla}_{t,\Psi_i}\right\|^2}{b_{t,i}^2} + \frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2}\right) + \alpha.
\end{aligned}$$

where the second to last equality is due to choosing $w = \frac{\sqrt{c\log\frac{1}{\delta}}}{\sigma_{\max}}$ and the last equality is letting $\alpha := \sigma_{\max}\sqrt{c\log\frac{1}{\delta}}$ for readability.

Let $M_{T,i} = \max_{t \le T} |\xi_{t,i}|$. Using our notation, we can define $M_{T,\Psi_i} := \max_{t \le T} \|\xi_{t,\Psi_i}\|$. Under event E_1 (and our new bound for C), we have that with probability at least $1 - c\delta$:

$$\begin{aligned}
\sum_{t=1}^{T}\sum_{i=0}^{c-1}\frac{\|\nabla_{t,\Psi_i}\|^2}{b_{t,i}} &\overset{(C)}{\le} \frac{\Delta_1}{\eta} - \sum_{t=1}^{T}\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{\nabla_{t,j}\xi_{t,j}}{a_{t,i}} + \sum_{t=1}^{T}\sum_{i=0}^{c-1}\|\xi_{t,\Psi_i}\|\left(\frac{\|\xi_{t,\Psi_i}\|^2}{2b_{t,i}^2} + \frac{\|\nabla_{t,\Psi_i}\|^2}{2a_{t,i}^2}\right) + \eta L\sum_{i=0}^{c-1}\log\frac{b_{T,i}}{b_{0,i}} \\
&\le \frac{\Delta_1}{\eta} - \sum_{t=1}^{T}\sum_{i=0}^{c-1}\sum_{j\in\Psi_i}\frac{\nabla_{t,j}\xi_{t,j}}{a_{t,i}} \qquad (7) \\
&\quad + \sum_{t=1}^{T}\sum_{i=0}^{c-1}M_{T,\Psi_i}\left(\frac{\|\xi_{t,\Psi_i}\|^2}{2b_{t,i}^2} + \frac{\|\nabla_{t,\Psi_i}\|^2}{2a_{t,i}^2}\right) + \eta L\sum_{i=0}^{c-1}\log\frac{b_{T,i}}{b_{0,i}} \\
&\qquad\qquad \text{(def of } M_{T,\Psi_i}\text{)} \\
&\overset{(E_1)}{\le} \frac{\Delta_1}{\eta} + 2\alpha\sum_{t=1}^{T}\sum_{i=0}^{c-1}\left(\underbrace{\frac{\left\|\widehat{\nabla}_{t,\Psi_i}\right\|^2}{b_{t,i}^2}}_{\text{bound with (C)}} + \frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2}\right) + \alpha + \\
&\quad \sum_{t=1}^{T}\sum_{i=0}^{c-1}M_{T,\Psi_i}\left(\frac{\|\xi_{t,\Psi_i}\|^2}{2b_{t,i}^2} + \frac{\|\nabla_{t,\Psi_i}\|^2}{2a_{t,i}^2}\right) + \eta L\sum_{i=0}^{c-1}\log\frac{b_{T,i}}{b_{0,i}} \qquad (8) \\
&\overset{(C)}{\le} \frac{\Delta_1}{\eta} + 2\alpha\sum_{t=1}^{T}\sum_{i=0}^{c-1}\frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2} + \alpha + \\
&\quad \sum_{t=1}^{T}\sum_{i=0}^{c-1}M_{T,\Psi_i}\left(\frac{\|\xi_{t,\Psi_i}\|^2}{2b_{t,i}^2} + \frac{\|\nabla_{t,\Psi_i}\|^2}{2a_{t,i}^2}\right) + (\eta L + 4\alpha)\sum_{i=0}^{c-1}\log\frac{b_{T,i}}{b_{0,i}} \qquad (9) \\
&\le \frac{\Delta_1}{\eta} + 2\alpha\sum_{t=1}^{T}\sum_{i=0}^{c-1}\frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2} + \alpha + \qquad (10) \\
&\quad \sum_{t=1}^{T}\sum_{i=0}^{c-1}M_{T,\Psi_i}\frac{\|\xi_{t,\Psi_i}\|^2}{2b_{t,i}^2} + \sum_{t=1}^{T}\sum_{i=0}^{c-1}M_{T,\Psi_i}\frac{\|\nabla_{t,\Psi_i}\|^2}{2a_{t,i}^2} + (\eta L + 4\alpha)\sum_{i=0}^{c-1}\log\frac{b_{T,i}}{b_{0,i}}. \qquad (11)
\end{aligned}$$

Let us turn our attention to $M_{T,\Psi_i} := \max_{t \le T} \|\xi_{t,\Psi_i}\|$. Note that

$$\begin{aligned}
\Pr\left[\max_{t\in[T]}\|\xi_{t,\Psi_i}\|^2 \ge A\right] &= \Pr\left[\exp\left(\frac{\max_{t\in[T]}\|\xi_{t,\Psi_i}\|^2}{w}\right) \ge \exp\left(\frac{A}{w}\right)\right] \qquad \text{(for } w > 0\text{)} \\
&\le \exp\left(-\frac{A}{w}\right)\mathbb{E}\left[\exp\left(\frac{\max_{t\in[T]}\|\xi_{t,\Psi_i}\|^2}{w}\right)\right] \qquad \text{(Markov)} \\
&= \exp\left(-\frac{A}{w}\right)\mathbb{E}\left[\max_{t\in[T]}\exp\left(\frac{\|\xi_{t,\Psi_i}\|^2}{w}\right)\right] \\
&\le \exp\left(-\frac{A}{w}\right)\sum_{t\in[T]}\mathbb{E}\left[\exp\left(\frac{\|\xi_{t,\Psi_i}\|^2}{w}\right)\right].
\end{aligned}$$

We have

$$\begin{aligned}\mathbb{E}\left[\exp\left(\frac{\|\xi_{t,\Psi_i}\|^2}{w}\right)\right] &= \mathbb{E}\left[\exp\left(\frac{\sum_{j\in\Psi_i}\xi_{t,j}^2}{w}\right)\right] \\ &= \mathbb{E}\left[\exp\left(\frac{\sum_{j\in\Psi_i}\xi_{t,j}^2}{w}\right)\right] \\ &= \mathbb{E}\left[\prod_{j\in\Psi_i}\exp\left(\frac{\xi_{t,j}^2}{w}\right)\right] \\ &= \prod_{j\in\Psi_i}\mathbb{E}\left[\exp\left(\frac{\xi_{t,j}^2}{w}\right)\right]. \qquad \text{(independence)}\end{aligned}$$

Since sub-gaussianity give us

$$\mathbb{E}\left[\exp\left(\lambda^2\xi_{t,i}^2\right)\right] \le \exp\left(\lambda^2\sigma_i^2\right), \forall\,|\lambda| \le \frac{1}{\sigma_i}, \forall i \in [d],$$

we have $\mathbb{E}\left[\exp\left(\frac{\xi_{t,j}^2}{w}\right)\right] \le \exp\left(\frac{\sigma_j^2}{w}\right)$ if $\sqrt{\frac{1}{w}} \le \frac{1}{\sigma_j}$. We pick $w := \|\sigma_{\Psi_i}\|^2 = \sum_{j\in\Psi_i}\sigma_j^2 \ge \sigma_j^2,\ \forall j \in \Psi_i$. Hence, we have

$$\begin{aligned}\mathbb{E}\left[\exp\left(\frac{\|\xi_{t,\Psi_i}\|^2}{\|\sigma_{\Psi_i}\|^2}\right)\right] &\le \prod_{j\in\Psi_i}\exp\left(\frac{\sigma_j^2}{\|\sigma_{\Psi_i}\|^2}\right) \\ &= \exp\left(\frac{\|\sigma_{\Psi_i}\|^2}{\|\sigma_{\Psi_i}\|^2}\right) = 1. \qquad (12)\end{aligned}$$

We have actually shown that ξ_{t,Ψ_i} is a $\|\sigma_{\Psi_i}\|^2$-subgaussian random variable in $\mathbb{R}^k$ (see Proposition 2.5.2 in [31]). This fact will come in handy later. Now, we have

$$\begin{aligned}\Pr\left[\max_{t\in[T]}\|\xi_{t,\Psi_i}\|^2 \ge A\right] &\le \exp\left(-\frac{A}{\|\sigma_{\Psi_i}\|^2}\right)\sum_{t\in[T]}\mathbb{E}\left[\exp\left(\frac{\|\xi_{t,\Psi_i}\|^2}{\|\sigma_{\Psi_i}\|^2}\right)\right] \\ &= \exp\left(-\frac{A}{\|\sigma_{\Psi_i}\|^2}\right)T.\end{aligned}$$

Setting $\exp\left(-\frac{A}{\|\sigma_{\Psi_i}\|^2}\right)T = \delta$ gives $A = \|\sigma_{\Psi_i}\|^2 \ln T/\delta$. Hence, we have with probability at least $1-\delta$,

$$M_{T,\Psi_i} = \max_{t\in[T]}\|\xi_{t,\Psi_i}\|^2 \le \|\sigma_{\Psi_i}\|^2 \ln T/\delta. \qquad (13)$$

Union bounding across all $i = 0, 1, \ldots, c-1$, we have that with probability at least $1 - c\delta$,

$$M_{T,\Psi_i} \le \|\sigma_{\Psi_i}\|^2 \ln T/\delta,\ \forall i = 0, 1, \ldots, c-1. \qquad (14)$$

Let us denote the event in (14) by E_2. Combining it with event E_1 and starting from 10, we have that with probability $1 - c\delta$:

$$\begin{aligned}
\sum_{t=1}^{T}\sum_{i=0}^{c-1}\frac{\|\nabla_{t,\Psi_i}\|^2}{b_{t,i}} &\le \frac{\Delta_1}{\eta} + 2\alpha\sum_{t=1}^{T}\sum_{i=0}^{c-1}\frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2} + \alpha + \sum_{t=1}^{T}\sum_{i=0}^{c-1}M_{T,\Psi_i}\frac{\|\xi_{t,\Psi_i}\|^2}{2b_{t,i}^2} + \\
&\quad \sum_{t=1}^{T}\sum_{i=0}^{c-1}M_{T,\Psi_i}\frac{\|\nabla_{t,\Psi_i}\|^2}{2a_{t,i}^2} + (\eta L + 4\alpha)\sum_{i=0}^{c-1}\log\frac{b_{T,i}}{b_{0,i}} \\
&\le \frac{\Delta_1}{\eta} + 2\alpha\sum_{t=1}^{T}\sum_{i=0}^{c-1}\frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2} + \ln T/\delta\sum_{t=1}^{T}\sum_{i=0}^{c-1}\|\sigma_{\Psi_i}\|^2\frac{\|\xi_{t,\Psi_i}\|^2}{2b_{t,i}^2} + \alpha + \\
&\quad \ln T/\delta\sum_{t=1}^{T}\sum_{i=0}^{c-1}\|\sigma_{\Psi_i}\|^2\frac{\|\nabla_{t,\Psi_i}\|^2}{2a_{t,i}^2} + (\eta L + 4\alpha)\sum_{i=0}^{c-1}\log\frac{b_{T,i}}{b_{0,i}} \\
&= \frac{\Delta_1}{\eta} + \sum_{i=0}^{c-1}\left(\ln T/\delta\frac{\|\sigma_{\Psi_i}\|^2}{2} + 2\alpha\right)\sum_{t=1}^{T}\frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2} + \alpha + \\
&\quad \ln T/\delta\sum_{i=0}^{c-1}\frac{\|\sigma_{\Psi_i}\|^2}{2}\sum_{t=1}^{T}\frac{\|\nabla_{t,\Psi_i}\|^2}{a_{t,i}^2} + (\eta L + 4\alpha)\sum_{i=0}^{c-1}\log\frac{b_{T,i}}{b_{0,i}}.
\end{aligned}$$

Recall that $\frac{\|\nabla_{t,\Psi_i}\|^2}{a_{t,i}^2} \le 2\frac{\|\widehat{\nabla}_{t,\Psi_i}\|^2}{b_{t,i}^2} + 2\frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2}$, we then have

$$\begin{aligned}
\ln T/\delta\sum_{i=0}^{c-1}\frac{\|\sigma_{\Psi_i}\|^2}{2}\sum_{t=1}^{T}\frac{\|\nabla_{t,\Psi_i}\|^2}{a_{t,i}^2} &\le \ln T/\delta\sum_{i=0}^{c-1}\frac{\|\sigma_{\Psi_i}\|^2}{2}\sum_{t=1}^{T}\left(2\frac{\left\|\widehat{\nabla}_{t,\Psi_i}\right\|^2}{b_{t,i}^2} + 2\frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2}\right) \\
&= \ln T/\delta\sum_{i=0}^{c-1}\|\sigma_{\Psi_i}\|^2\sum_{t=1}^{T}\frac{\left\|\widehat{\nabla}_{t,\Psi_i}\right\|^2}{b_{t,i}^2} + \ln T/\delta\sum_{i=0}^{c-1}\|\sigma_{\Psi_i}\|^2\sum_{t=1}^{T}\frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2} \\
&\le \ln T/\delta\sum_{i=0}^{c-1}\|\sigma_{\Psi_i}\|^2\log\frac{b_{T,i}}{b_{0,i}} + \ln T/\delta\sum_{i=0}^{c-1}\|\sigma_{\Psi_i}\|^2\sum_{t=1}^{T}\frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2}. \qquad \text{(from 5)}
\end{aligned}$$

Hence, we have with probability at least $1 - 2c\delta$:

$$\begin{aligned}
\sum_{t=1}^{T}\sum_{i=0}^{c-1}\frac{\|\nabla_{t,\Psi_i}\|^2}{b_{t,i}} &\le \frac{\Delta_1}{\eta} + \sum_{i=0}^{c-1}\left(\ln T/\delta\,\|\sigma_{\Psi_i}\|^2 + 2\alpha\right)\sum_{t=1}^{T}\frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2} \qquad (15) \\
&\quad + \alpha + \sum_{i=0}^{c-1}\ln T/\delta\,\|\sigma_{\Psi_i}\|^2\log\frac{b_{T,i}}{b_{0,i}} + \sum_{i=0}^{c-1}(\eta L + 4\alpha)\log\frac{b_{T,i}}{b_{0,i}} \\
&= \frac{\Delta_1}{\eta} + \sum_{i=0}^{c-1}\left(\ln T/\delta\,\|\sigma_{\Psi_i}\|^2 + 2\alpha\right)\sum_{t=1}^{T}\frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2} \qquad (16) \\
&\quad + \alpha + \sum_{i=0}^{c-1}\left(\ln T/\delta\,\|\sigma_{\Psi_i}\|^2 + \eta L + 4\alpha\right)\log\frac{b_{T,i}}{b_{0,i}}. \qquad (17)
\end{aligned}$$

Now, we bound $\sum_{t=1}^{T} \frac{\left\|\xi_{t,\Psi_i}\right\|^2}{b_{t,i}^2}$ and $\log \frac{b_{T,i}}{b_{0,i}}$. We need to first lower bound $\sum_{s=1}^{t} \left\|\widehat{\nabla}_{t,\Psi_i}\right\|^2$. We proceed by noting that

$$\begin{aligned}
\|\widehat{\nabla}_{t,\Psi_i}\|^2 &= \|\nabla_{t,\Psi_i} + \xi_{t,\Psi_i}\|^2 \\
&= \|\nabla_{t,\Psi_i}\|^2 + 2\langle \xi_{t,\Psi_i}, \nabla_{t,\Psi_i}\rangle + \|\xi_{t,\Psi_i}\|^2 \\
\Rightarrow \|\nabla_{t,\Psi_i}\| - \|\widehat{\nabla}_{t,\Psi_i}\|^2 + \|\xi_{t,\Psi_i}\|^2 &= 2\langle \xi_{t,\Psi_i}, \nabla_{t,\Psi_i}\rangle.
\end{aligned}$$

Define for $t \in \{0, 1, \cdots, T\}$ and some constant v_s to be specified later:

$$\begin{aligned}
U_{t+1} &= \exp\left(\sum_{s=1}^{t} w_s \left(\|\nabla_{s,\Psi_i}\| - \|\widehat{\nabla}_{s,\Psi_i}\|^2 + \|\xi_{s,\Psi_i}\|^2\right) - v_s\|\nabla_{s,\Psi_i}\|^2\right) \\
&= U_t \cdot \exp\left(w_t \left(\|\nabla_{t,\Psi_i}\| - \|\widehat{\nabla}_{t,\Psi_i}\|^2 + \|\xi_{t,\Psi_i}\|^2\right) - v_t\|\nabla_{t,\Psi_i}\|^2\right) \\
&= U_t \cdot \exp\left(w_t \left(2\langle \xi_{t,\Psi_i}, \nabla_{t,\Psi_i}\rangle\right) - v_t\|\nabla_{t,\Psi_i}\|^2\right).
\end{aligned}$$

First, note that $U_t \in \mathcal{F}_t$. We show that U_t is a supermartingale

$$\begin{aligned}
\mathbb{E}\left[U_{t+1} \mid \mathcal{F}_t\right] &= \mathbb{E}\left[U_t \cdot \exp\left(w_t \left(2\langle \xi_{t,\Psi_i}, \nabla_{t,\Psi_i}\rangle\right) - v_t\|\nabla_{t,\Psi_i}\|^2\right) \mid \mathcal{F}_t\right] \\
&= U_t \exp\left(-v_t\|\nabla_{t,\Psi_i}\|^2\right) \mathbb{E}\left[\exp\left(2w_t\langle \xi_{t,\Psi_i}, \nabla_{t,\Psi_i}\rangle\right) \mid \mathcal{F}_t\right] \\
&\overset{(*)}{\leq} U_t \exp\left(-v_t\|\nabla_{t,\Psi_i}\|^2\right) \mathbb{E}\left[\exp\left(4w_t^2 \left\|\sigma_{\Psi_i}\right\|^2 \|\nabla_{t,\Psi_i}\|^2\right) \mid \mathcal{F}_t\right] \\
&= U_t, \qquad\qquad (v_t{=}4w_t^2 \left\|\sigma_{\Psi_i}\right\|^2)
\end{aligned}$$

where $(*)$ is due to Lemma 2.2 of [21] and the fact that ξ_{t,Ψ_i} is $\left\|\sigma_{\Psi_i}\right\|^2$-subgaussian from (12). Hence, by Ville's supermartingale inequality, we have

$$\Pr\left[\max_{t\in[T+1]} U_t \geq \delta^{-1}\right] \leq \delta\mathbb{E}\left[U_1\right] = \delta.$$

This implies w.p. $\geq 1-\delta$, $\forall 0 \leq t \leq T$:

$$\begin{aligned}
&\sum_{s=1}^{t} w_s \left(\|\nabla_{s,\Psi_i}\| - \|\widehat{\nabla}_{s,\Psi_i}\|^2 + \|\xi_{s,\Psi_i}\|^2\right) - v_s\|\nabla_{s,\Psi_i}\|^2 \leq \log\frac{1}{\delta} \\
\Longrightarrow &\sum_{s=1}^{t} \left(w_s - 4w_s^2 \left\|\sigma_{\Psi_i}\right\|^2\right) \|\nabla_{s,\Psi_i}\|^2 + \sum_{s=1}^{t} w_s\|\xi_{s,\Psi_i}\|^2 \leq \sum_{s=1}^{t} w_s\|\widehat{\nabla}_{s,\Psi_i}\|^2 + \log\frac{1}{\delta} \\
\Longleftrightarrow &\sum_{s=1}^{t} \left(1 - 4w_s \left\|\sigma_{\Psi_i}\right\|^2\right) \|\nabla_{s,\Psi_i}\|^2 + \sum_{s=1}^{t} \|\xi_{s,\Psi_i}\|^2 \leq \sum_{s=1}^{t} \|\widehat{\nabla}_{s,\Psi_i}\|^2 + \frac{1}{w_s}\log\frac{1}{\delta}.
\end{aligned}$$

Set $w_s = \frac{1}{4\left\|\sigma_{\Psi_i}\right\|^2}$ to get

$$\sum_{s=1}^{t} \|\xi_{s,\Psi_i}\|^2 \leq \sum_{s=1}^{t} \|\widehat{\nabla}_{s,\Psi_i}\|^2 + 4\left\|\sigma_{\Psi_i}\right\|^2 \log\frac{1}{\delta},\ \forall t \leq T. \tag{18}$$

We are now ready to bound $\sum_{t=1}^{T} \frac{\left\|\xi_{t,\Psi_i}\right\|^2}{b_{t,i}^2}$. Starting by applying (18), we have that with probability at least $1-\delta$

$$\sum_{t=1}^{T} \frac{\left\|\xi_{t,\Psi_i}\right\|^2}{b_{t,i}^2} = \sum_{t=1}^{T} \frac{\left\|\xi_{t,\Psi_i}\right\|^2}{b_{0,i}^2 + \sum_{s=1}^{t} \left\|\widehat{\nabla}_{t,\Psi_i}\right\|^2} \leq \sum_{t=1}^{T} \frac{\left\|\xi_{t,\Psi_i}\right\|^2}{b_{0,i}^2 + \left(\sum_{s=1}^{t} \left\|\xi_{s,\Psi_i}\right\|^2 - 4\left\|\sigma_{\Psi_i}\right\|^2 \log\frac{1}{\delta}\right)^+}$$

where $(x)^+ = \max\{x, 0\}$. Let $\tau = \max\left(\{0\} \cup \left\{t \in \mathbb{N}_{\leq T} \mid \sum_{s=1}^{t} \left\|\xi_{s,\Psi_i}\right\|^2 \leq 2C\right\}\right)$ for some $C \geq 0$. We have

$$\begin{aligned}
\sum_{t=1}^{T} \frac{\left\|\xi_{t,\Psi_i}\right\|^2}{b_{t,i}^2} &= \sum_{t=1}^{\tau} \frac{\left\|\xi_{t,\Psi_i}\right\|^2}{b_{t,i}^2} + \sum_{t=\tau+1}^{T} \frac{\left\|\xi_{t,\Psi_i}\right\|^2}{b_{0,i}^2 + \sum_{s=1}^{t} \left\|\widehat{\nabla}_{t,\Psi_i}\right\|^2} \\
&\leq \frac{1}{b_{0,i}^2} \sum_{t=1}^{\tau} \left\|\xi_{t,\Psi_i}\right\|^2 + \sum_{t=\tau+1}^{T} \frac{\left\|\xi_{t,\Psi_i}\right\|^2}{b_{0,i}^2 + \sum_{s=1}^{t} \left\|\xi_{s,\Psi_i}\right\|^2 - 4\left\|\sigma_{\Psi_i}\right\|^2 \log\frac{1}{\delta}} \\
&\leq \frac{2C}{b_{0,i}^2} + \sum_{t=\tau+1}^{T} \frac{\left\|\xi_{t,\Psi_i}\right\|^2}{b_{0,i}^2 + \sum_{s=1}^{t} \left\|\xi_{s,\Psi_i}\right\|^2 - 4\left\|\sigma_{\Psi_i}\right\|^2 \log\frac{1}{\delta}}.
\end{aligned}$$

Now, since $\frac{\sum_{s=1}^{t}\left\|\xi_{s,\Psi_i}\right\|^2}{2} \geq C$ for $t > \tau$, we have $b_{0,i}^2 + \sum_{s=1}^{t} \left\|\xi_{s,\Psi_i}\right\|^2 - 4\left\|\sigma_{\Psi_i}\right\|^2 \log\frac{1}{\delta} \geq b_{0,i}^2 - 4\left\|\sigma_{\Psi_i}\right\|^2 \log\frac{1}{\delta} + C + \frac{1}{2}\sum_{s=1}^{t} \left\|\xi_{s,\Psi_i}\right\|^2$. If $b_{0,i}^2 - 4\left\|\sigma_{\Psi_i}\right\|^2 \log\frac{1}{\delta} \geq 0$, then we pick $C = 0$ and $b_{0,i}^2 - 4\left\|\sigma_{\Psi_i}\right\|^2 \log\frac{1}{\delta} + C + \frac{1}{2}\sum_{s=1}^{t} \left\|\xi_{s,\Psi_i}\right\|^2 \geq \frac{1}{2}\sum_{s=1}^{t} \left\|\xi_{s,\Psi_i}\right\|^2$. If $b_{0,i}^2 - 4\left\|\sigma_{\Psi_i}\right\|^2 \log\frac{1}{\delta} < 0$, we pick $C = 4\left\|\sigma_{\Psi_i}\right\|^2 \log\frac{1}{\delta} - b_{0,i}^2 > 0$, which gives $b_{0,i}^2 - 4\left\|\sigma_{\Psi_i}\right\|^2 \log\frac{1}{\delta} + C + \frac{1}{2}\sum_{s=1}^{t} \left\|\xi_{s,\Psi_i}\right\|^2 \geq \frac{1}{2}\sum_{s=1}^{t} \left\|\xi_{s,\Psi_i}\right\|^2$. In either case, we have $b_{0,i}^2 - 4\left\|\sigma_{\Psi_i}\right\|^2 \log\frac{1}{\delta} + C + \frac{1}{2}\sum_{s=1}^{t} \left\|\xi_{s,\Psi_i}\right\|^2 \geq \frac{1}{2}\sum_{s=1}^{t} \left\|\xi_{s,\Psi_i}\right\|^2$. Hence, letting $C = \max\left(0, 4\left\|\sigma_{\Psi_i}\right\|^2 \log\frac{1}{\delta} - b_{0,i}^2\right) \leq 4\left\|\sigma_{\Psi_i}\right\|^2 \log\frac{1}{\delta}$, we have with probability at least $1-\delta$:

$$\begin{aligned}
\sum_{t=1}^{T} \frac{\left\|\xi_{t,\Psi_i}\right\|^2}{b_{t,i}^2} &\leq \frac{2C}{b_{0,i}^2} + 2\sum_{t=\tau+1}^{T} \frac{\left\|\xi_{t,\Psi_i}\right\|^2}{\sum_{s=1}^{t} \left\|\xi_{s,\Psi_i}\right\|^2} \\
&\leq \frac{2C}{b_{0,i}^2} + 2\sum_{t=1}^{T} \frac{\left\|\xi_{t,\Psi_i}\right\|^2}{\sum_{s=1}^{t} \left\|\xi_{s,\Psi_i}\right\|^2} \\
&\leq \frac{8\left\|\sigma_{\Psi_i}\right\|^2 \log\frac{1}{\delta}}{b_{0,i}^2} + 2\sum_{t=1}^{T} \frac{\left\|\xi_{t,\Psi_i}\right\|^2}{\sum_{s=1}^{t} \left\|\xi_{s,\Psi_i}\right\|^2}.
\end{aligned}$$

Let $X_t = 1 + \sum_{s=1}^{t} \|\xi_{s,\Psi_i}\|^2 = X_{t-1} + \|\xi_{t,\Psi_i}\|^2$, where $X_0 = 1$. Then,

$$\begin{aligned}\sum_{t=1}^{T} \frac{\|\xi_{t,\Psi_i}\|^2}{\sum_{s=1}^{t} \|\xi_{s,\Psi_i}\|^2} &= \sum_{t=1}^{T} \frac{X_t - X_{t-1}}{X_t} = \sum_{t=1}^{T} 1 - \frac{X_{t-1}}{X_t} \\ &\leq \sum_{t=1}^{T} \log\left(\frac{X_t}{X_{t-1}}\right) \\ &= \log\left(\prod_{t=1}^{T} \frac{X_t}{X_{t-1}}\right) \\ &= \log\left(\frac{X_T}{X_0}\right) = \log\left(1 + \sum_{t=1}^{T} \|\xi_{s,\Psi_i}\|^2\right).\end{aligned}$$

Hence, with probability at least $1 - \delta$:

$$\sum_{t=1}^{T} \frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2} \leq \frac{8\|\sigma_{\Psi_i}\|^2 \log\frac{1}{\delta}}{b_{0,i}^2} + 2\log\left(1 + \sum_{t=1}^{T} \|\xi_{s,\Psi_i}\|^2\right). \tag{19}$$

It remains to bound $\sum_{t=1}^{T} \|\xi_{s,\Psi_i}\|^2$. Note that

$$\begin{aligned}\Pr\left[\sum_{t=1}^{T} \|\xi_{s,\Psi_i}\|^2 \geq u\right] &= \Pr\left[\exp\left(\sum_{t=1}^{T} \frac{\|\xi_{s,\Psi_i}\|^2}{\|\sigma_{\Psi_i}\|^2}\right) \geq \exp\left(\frac{u}{\|\sigma_{\Psi_i}\|^2}\right)\right] \\ &\leq \frac{\mathbb{E}\left[\exp\left(\sum_{t=1}^{T} \frac{\|\xi_{s,\Psi_i}\|^2}{\|\sigma_{\Psi_i}\|^2}\right)\right]}{\exp\left(\frac{u}{\|\sigma_{\Psi_i}\|^2}\right)} \\ &\leq \frac{\exp(T)}{\exp\left(\frac{u}{\|\sigma_{\Psi_i}\|^2}\right)} \qquad (\xi_{s,\Psi_i} \text{is } \|\sigma_{\Psi_i}\|^2\text{-subgaussian})\end{aligned}$$

Choosing $u = \|\sigma_{\Psi_i}\|^2 T + \|\sigma_{\Psi_i}\|^2 \log\frac{1}{\delta}$ gives that with probability at least $1 - \delta$, we have

$$\sum_{t=1}^{T} \|\xi_{s,\Psi_i}\|^2 \leq \|\sigma_{\Psi_i}\|^2 T + \|\sigma_{\Psi_i}\|^2 \log\frac{1}{\delta}. \tag{20}$$

Having a high probability bound on the sum of the stochastic error of the subset-norm, we can combine both events from (19) and (20) to get that with probability at least $1 - 2\delta$:

$$\sum_{t=1}^{T} \frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2} \leq \frac{8\|\sigma_{\Psi_i}\|^2 \log\frac{1}{\delta}}{b_{0,i}^2} + 2\log\left(1 + \|\sigma_{\Psi_i}\|^2 T + \|\sigma_{\Psi_i}\|^2 \log\frac{1}{\delta}\right). \tag{21}$$

Then we can also condition on the event that (21) happens and combine it with the event in (17) to get that with probability at least $1 - 2c\delta$ (assuming $c \geq 2$), we have

$$\sum_{t=1}^{T}\sum_{i=0}^{c-1}\frac{\|\nabla_{t,\Psi_i}\|_2^2}{b_{t,i}} \leq \frac{\Delta_1}{\eta} + \sum_{i=0}^{c-1}\left(\ln T/\delta\,\|\sigma_{\Psi_i}\|^2 + 2\alpha\right)\sum_{t=1}^{T}\frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}^2} \tag{22}$$

$$+\alpha + \sum_{i=0}^{c-1}\left(\ln T/\delta\,\|\sigma_{\Psi_i}\|^2 + \eta L + 4\alpha\right)\log\frac{b_{T,i}}{b_{0,i}} \tag{23}$$

$$\leq \frac{\Delta_1}{\eta} + \underbrace{\sum_{i=0}^{c-1}\left(\ln T/\delta\,\|\sigma_{\Psi_i}\|^2 + 2\alpha\right)\left(\frac{8\,\|\sigma_{\Psi_i}\|^2\log\frac{1}{\delta}}{b_{0,i}^2} + 2\log\left(1 + \|\sigma_{\Psi_i}\|^2 T + \|\sigma_{\Psi_i}\|^2\log\frac{1}{\delta}\right)\right)}_{=:H(\delta)} \tag{24}$$

$$+\alpha + \sum_{i=0}^{c-1}\left(\ln T/\delta\,\|\sigma_{\Psi_i}\|^2 + \eta L + 4\alpha\right)\log\frac{b_{T,i}}{b_{0,i}}$$

$$= \frac{\Delta_1}{\eta} + H(\delta) + \alpha + \sum_{i=0}^{c-1}\left(\ln T/\delta\,\|\sigma_{\Psi_i}\|^2 + \eta L + 4\alpha\right)\log\frac{b_{T,i}}{b_{0,i}}. \tag{25}$$

First, note that $b_{T,i} \leq \|b_T\|_1 = \sum_{i=0}^{c-1} b_{T,i}$. Letting $b_{0,\min} := \min_i b_{0,i}$, we then have

$$\sum_{i=0}^{c-1}\left(\ln T/\delta\,\|\sigma_{\Psi_i}\|^2 + \eta L + 4\alpha\right)\log\frac{b_{T,i}}{b_{0,i}} \leq \log\frac{\|b_T\|_1}{b_{0,\min}}\sum_{i=0}^{c-1}\left(\ln T/\delta\,\|\sigma_{\Psi_i}\|^2 + \eta L + 4\alpha\right)$$

$$= \log\frac{\|b_T\|_1}{b_{0,\min}}\left(\ln T/\delta\,\|\sigma\|_2^2 + c\eta L + 4c\alpha\right).$$

Now, note the LHS term $\sum_{t=1}^{T}\sum_{i=0}^{c-1}\frac{\|\nabla_{t,\Psi_i}\|_2^2}{b_{t,i}}$ of (23):

$$\left(\sum_{i=0}^{c-1}\frac{\|\nabla_{t,\Psi_i}\|_2^2}{b_{t,i}}\right)\left(\sum_{i=0}^{c-1} b_{t,i}\right) \geq \left(\sum_{i=0}^{c-1}\|\nabla_{t,\Psi_i}\|_2\right)^2 \geq \sum_{i=0}^{c-1}\|\nabla_{t,\Psi_i}\|_2^2 = \|\nabla_t\|_2^2$$

$$\implies \frac{\|\nabla_t\|_2^2}{\left(\sum_{i=0}^{c-1} b_{t,i}\right)} \leq \sum_{i=0}^{c-1}\frac{\|\nabla_{t,\Psi_i}\|_2^2}{b_{t,i}}.$$

Now, $\sum_{i=0}^{c-1} b_{t,i} = \sum_{i=0}^{c-1} |b_{t,i}| = \|b_t\|_1$, so with probability $1 - 2c\delta$:

$$\sum_{t=1}^{T} \frac{\|\nabla_t\|_2^2}{\|b_T\|_1} \leq \sum_{t=1}^{T} \frac{\|\nabla_t\|_2^2}{\|b_t\|_1} \leq \sum_{t=1}^{T} \sum_{i=0}^{c-1} \frac{\|\nabla_{t,\Psi_i}\|_2^2}{b_{t,i}}$$

$$\implies \sum_{t=1}^{T} \|\nabla_t\|_2^2 \leq \|b_T\|_1 \sum_{t=1}^{T} \sum_{i=0}^{c-1} \frac{\|\nabla_{t,\Psi_i}\|_2^2}{b_{t,i}}$$

$$\leq \|b_T\|_1 \left(\frac{\Delta_1}{\eta} + cH(\delta) + \left(\ln T/\delta \, \|\sigma\|_2^2 + c\eta L + 4c\alpha \right) \log \frac{\|b_T\|_1}{b_{0,\min}} \right) \tag{26}$$

$$\leq \|b_T\|_1 \left(\frac{\Delta_1}{\eta} + cH(\delta) + \left(\ln T/\delta \, \|\sigma\|_2^2 + c\eta L + 4c\alpha \right) \log \frac{\|b_T\|_1}{b_{0,\min}} \right). \tag{27}$$

It remains to bound $\|b_T\|_1$. We start again from smoothness of f:

$$\begin{aligned}
\Delta_{t+1} - \Delta_t &\leq \langle \nabla_t, x_{t+1} - x_t \rangle + \frac{L}{2} \|x_{t+1} - x_t\|^2 \\
&= -\eta \left\langle \nabla_t, \frac{\widehat{\nabla}_t}{b_t} \right\rangle + \frac{\eta^2 L}{2} \left\| \frac{\widehat{\nabla}_t}{b_t} \right\|^2 \\
&= -\eta \left\langle \widehat{\nabla}_t - \xi_t, \frac{\widehat{\nabla}_t}{b_t} \right\rangle + \frac{\eta^2 L}{2} \sum_{i=0}^{c-1} \sum_{j \in \Psi_i} \frac{\widehat{\nabla}_{t,\Psi_j}^2}{b_{t,i}^2} \\
&= -\eta \left\langle \widehat{\nabla}_t, \frac{\widehat{\nabla}_t}{b_t} \right\rangle + \eta \left\langle \xi_t, \frac{\widehat{\nabla}_t}{b_t} \right\rangle + \frac{\eta^2 L}{2} \sum_{i=0}^{c-1} \frac{\left\| \widehat{\nabla}_{t,\Psi_i} \right\|^2}{b_{t,i}^2} \\
&= -\eta \sum_{i=0}^{c-1} \sum_{j \in \Psi_i} \frac{\widehat{\nabla}_{t,j}^2}{b_{t,i}} + \eta \sum_{i=0}^{c-1} \sum_{j \in \Psi_i} \frac{\xi_{t,j} \widehat{\nabla}_{t,j}}{b_{t,i}} + \frac{\eta^2 L}{2} \sum_{i=0}^{c-1} \frac{\left\| \widehat{\nabla}_{t,\Psi_i} \right\|^2}{b_{t,i}^2} \\
&= -\eta \sum_{i=0}^{c-1} \frac{\left\| \widehat{\nabla}_{t,\Psi_i} \right\|^2}{b_{t,i}} + \frac{\eta^2 L}{2} \sum_{i=0}^{c-1} \frac{\left\| \widehat{\nabla}_{t,\Psi_i} \right\|^2}{b_{t,i}^2} + \eta \sum_{i=0}^{c-1} \sum_{j \in \Psi_i} \frac{\xi_{t,j} \widehat{\nabla}_{t,j}}{b_{t,i}}.
\end{aligned} \tag{28}$$

Note that

$$\begin{aligned}
\sum_{i=0}^{c-1} \sum_{j \in \Psi_i} \frac{\xi_{t,j} \widehat{\nabla}_{t,j}}{b_{t,i}} &\leq \frac{1}{2} \sum_{i=0}^{c-1} \sum_{j \in \Psi_i} \frac{\xi_{t,j}^2}{b_{t,i}} + \frac{1}{2} \sum_{i=0}^{c-1} \sum_{j \in \Psi_i} \frac{\widehat{\nabla}_{t,j}^2}{b_{t,i}} \\
&= \frac{1}{2} \sum_{i=0}^{c-1} \sum_{j \in \Psi_i} \frac{\xi_{t,j}^2}{b_{t,i}} + \frac{1}{2} \sum_{i=0}^{c-1} \frac{\left\| \widehat{\nabla}_{t,\Psi_i} \right\|^2}{b_{t,i}}.
\end{aligned}$$

Plugging back in, we have

$$\Delta_{t+1} - \Delta_t \leq -\frac{\eta}{2} \sum_{i=0}^{c-1} \frac{\left\| \widehat{\nabla}_{t,\Psi_i} \right\|^2}{b_{t,i}} + \eta^2 L \sum_{i=0}^{c-1} \frac{\left\| \widehat{\nabla}_{t,\Psi_i} \right\|^2}{b_{t,i}^2} + \frac{\eta}{2} \sum_{i=0}^{c-1} \frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}}.$$

Summing over T and rearranging, we get

$$\sum_{t=1}^{T}\sum_{i=0}^{c-1}\frac{\left\|\widehat{\nabla}_{t,\Psi_i}\right\|^2}{b_{t,i}} \le \frac{2\Delta_1}{\eta} + \sum_{t=1}^{T}\sum_{i=0}^{c-1}\frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}} + 2\eta L\sum_{t=1}^{T}\sum_{i=0}^{c-1}\frac{\left\|\widehat{\nabla}_{t,\Psi_i}\right\|^2}{b_{t,i}^2}$$

$$\implies \sum_{t=1}^{T}\sum_{i=0}^{c-1}\frac{\left\|\widehat{\nabla}_{t,\Psi_i}\right\|^2}{b_{t,i}} \le \frac{4\Delta_1}{\eta} + 2\sum_{t=1}^{T}\sum_{i=0}^{c-1}\frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}} + \sum_{t=1}^{T}\sum_{i=0}^{c-1}\left(\frac{4\eta L}{b_{t,i}^2} - \frac{1}{b_{t,i}}\right)\left\|\widehat{\nabla}_{t,\Psi_i}\right\|^2.$$

We can bound $\sum_{t=1}^{T}\sum_{i=0}^{c-1}\left(\frac{4\eta L}{b_{t,i}^2} - \frac{1}{b_{t,i}}\right)\left\|\widehat{\nabla}_{t,\Psi_i}\right\|^2$ as follows. Consider $i \in [c]$. Let $\tau_i = \max\{t \le T \mid b_{t,i} \le 4\eta L\}$ so that $t \ge \tau_i$ implies $b_{t,i} > 4\eta L \iff \frac{4\eta L}{b_{t,i}^2} < \frac{1}{b_{t,i}}$:

$$\begin{aligned}
\sum_{t=1}^{T}\left(\frac{4\eta L}{b_{t,i}^2} - \frac{1}{b_{t,i}}\right)\left\|\widehat{\nabla}_{t,\Psi_i}\right\|^2 &= \sum_{t=1}^{\tau_i}\left(\frac{4\eta L}{b_{t,i}^2} - \frac{1}{b_{t,i}}\right)\left\|\widehat{\nabla}_{t,\Psi_i}\right\|^2 + \sum_{t=\tau_i+1}^{T}\left(\underbrace{\frac{4\eta L}{b_{t,i}^2} - \frac{1}{b_{t,i}}}_{<0}\right)\left\|\widehat{\nabla}_{t,\Psi_i}\right\|^2 \\
&\le \sum_{t=1}^{\tau_i}\left(\frac{4\eta L}{b_{t,i}^2} - \frac{1}{b_{t,i}}\right)\left\|\widehat{\nabla}_{t,\Psi_i}\right\|^2 \\
&\le 4\eta L\sum_{t=1}^{\tau_i}\frac{\left\|\widehat{\nabla}_{t,\Psi_i}\right\|^2}{b_{t,i}^2} \\
&\le 8\eta L\log\frac{b_{\tau_i,i}}{b_{0,i}} \le 8\eta L\log\frac{4\eta L}{b_{0,i}}.
\end{aligned}$$

Hence, we have

$$\sum_{t=1}^{T}\sum_{i=0}^{c-1}\frac{\left\|\widehat{\nabla}_{t,\Psi_i}\right\|^2}{b_{t,i}} \le \frac{4\Delta_1}{\eta} + 2\sum_{t=1}^{T}\sum_{i=0}^{c-1}\frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}} + 8\eta L\sum_{i=0}^{c-1}\log\frac{4\eta L}{b_{0,i}}.$$

Consider the LHS

$$\begin{aligned}
\sum_{t=1}^{T}\sum_{i=0}^{c-1}\frac{\left\|\widehat{\nabla}_{t,\Psi_i}\right\|^2}{b_{t,i}} &= \sum_{t=1}^{T}\sum_{i=0}^{c-1}\frac{b_{t,i}^2 - b_{t-1,i}^2}{b_{t,i}} = \sum_{t=1}^{T}\sum_{i=0}^{c-1} b_{t,i} - \frac{b_{t-1,i}^2}{b_{t,i}} \\
&\ge \sum_{t=1}^{T}\sum_{i=0}^{c-1} b_{t,i} - \frac{b_{t-1,i}^2}{b_{t-1,i}} = \sum_{t=1}^{T}\sum_{i=0}^{c-1} b_{t,i} - b_{t-1,i} \\
&= \sum_{i=0}^{c-1}\sum_{t=1}^{T} b_{t,i} - b_{t-1,i} = \sum_{i=0}^{c-1} b_{T,i} - b_{0,i} \\
&= \|b_T\|_1 - \|b_0\|_1 .
\end{aligned}$$

Hence, we have

$$\|b_T\|_1 \le \|b_0\|_1 + \frac{2\Delta_1}{\eta} + \sum_{i=0}^{c-1}\sum_{t=1}^{T}\frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}} + 8\eta Lc\log\frac{4\eta L}{b_{0,\min}}.$$

It remains to bound $\sum_{t=1}^{T} \frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}}$ for each $i \in [c]$. Recall from (21), with probability at least $1-\delta$

$$\sum_{s=1}^{t} \|\xi_{t,\Psi_i}\|^2 \le \sum_{s=1}^{t} \|\widehat{\nabla}_{t,\Psi_i}\|^2 + 4\|\sigma_{\Psi_i}\|^2 \log\frac{1}{\delta},\ \forall t \le T.$$

We have with probability at least $1-2c\delta$,

$$\begin{aligned}
\sum_{t=1}^{T} \frac{\|\xi_{t,\Psi_i}\|^2}{b_{t,i}} &= \sum_{t=1}^{T} \frac{\|\xi_{t,\Psi_i}\|^2}{\sqrt{b_{0,i}^2 + \sum_{s=1}^{t} \|\widehat{\nabla}_{s,\Psi_i}\|^2}} \\
&\overset{(1)}{\le} \sum_{t=1}^{T} \frac{\xi_{t,i}^2}{\sqrt{b_{0,i}^2 + \left(\sum_{s=1}^{t} \|\xi_{s,\Psi_i}\|^2 - 4\|\sigma_{\Psi_i}\|^2 \log\frac{1}{\delta}\right)^+}} \\
&\le \frac{8\|\sigma_{\Psi_i}\|^2 \log\frac{1}{\delta}}{b_{0,i}} + 2\sqrt{2}\sqrt{\sum_{s=1}^{T} \|\xi_{s,\Psi_i}\|^2} \\
&\overset{(2)}{\le} \frac{8\|\sigma_{\Psi_i}\|^2 \log\frac{1}{\delta}}{b_{0,i}} + 4\sqrt{\|\sigma_{\Psi_i}\|^2 T + \|\sigma_{\Psi_i}\|^2 \log\frac{1}{\delta}},
\end{aligned}$$

where (1) is due to (18) and (2) is due to Lemma (20). Hence, we have that with probability at least $1-2c\delta$,

$$\begin{aligned}
\|b_T\|_1 &\le \|b_0\|_1 + \frac{2\Delta_1}{\eta} + \sum_{i=0}^{c-1} \frac{8\|\sigma_{\Psi_i}\|^2 \log\frac{1}{\delta}}{b_{0,i}} + \sum_{i=0}^{c-1} 4\sqrt{\|\sigma_{\Psi_i}\|^2 T + \|\sigma_{\Psi_i}\|^2 \log\frac{1}{\delta}} + 8\eta L c \log\frac{4\eta L}{b_{0,\min}} \\
&\le \|b_0\|_1 + \frac{2\Delta_1}{\eta} + \frac{8\log\frac{1}{\delta}}{b_{0,\min}} \sum_{i=0}^{c-1} \|\sigma_{\Psi_i}\|^2 + 4\sqrt{T}\sum_{i=0}^{c-1} \|\sigma_{\Psi_i}\| + \sqrt{\log\frac{1}{\delta}} \sum_{i=0}^{c-1} \|\sigma_{\Psi_i}\| + 8\eta L c \log\frac{4\eta L}{b_{0,\min}} \\
&= 4\sqrt{T}\sum_{i=0}^{c-1} \|\sigma_{\Psi_i}\| + \underbrace{\|b_0\|_1 + \frac{2\Delta_1}{\eta} + \frac{8\log\frac{1}{\delta}}{b_{0,\min}} \|\sigma\|_2^2 + \sqrt{\log\frac{1}{\delta}} \sum_{i=0}^{c-1} \|\sigma_{\Psi_i}\| + 8\eta L c \log\frac{4\eta L}{b_{0,\min}}}_{=:I(\delta)}.
\end{aligned}$$

Hence, we can combine (27) with the bound for $\|b_T\|_1$ to get that with probability $1-6c\delta$:

$$\begin{aligned}
\sum_{t=1}^{T} \|\nabla_t\|_2^2 &\le \|b_T\|_1 \left(\frac{\Delta_1}{\eta} + H(\delta) + \left(\ln T/\delta \|\sigma\|_2^2 + c\eta L + 4c\sigma_{\max}\sqrt{c\log\frac{1}{\delta}}\right) \log\frac{\|b_T\|_1}{b_{0,\min}}\right) \\
&\le \left(4\sqrt{T}\sum_{i=0}^{c-1} \|\sigma_{\Psi_i}\| + I(\delta)\right) \cdot \\
&\quad \left(\frac{\Delta_1}{\eta} + H(\delta) + \left(\ln T/\delta \|\sigma\|_2^2 + c\eta L + 4c^{3/2}\sigma_{\max}\sqrt{\log\frac{1}{\delta}}\right) \log\left(\frac{4\sqrt{T}\sum_{i=0}^{c-1} \|\sigma_{\Psi_i}\| + I(\delta)}{b_{0,\min}}\right)\right).
\end{aligned}$$

Dividing both sides by T, we get the theorem that with probability $1 - 6c\delta$:

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla_t\|_2^2 \le G(\delta)\cdot\left(\frac{4\sum_{i=0}^{c-1}\|\sigma_{\Psi_i}\|}{\sqrt{T}}+\frac{I(\delta)}{T}\right), \text{ where } G(\delta) \text{ and } I(\delta) \text{ are polylog terms:}$$

$$G(\delta) := \frac{\Delta_1}{\eta} + H(\delta) + \left(\ln T/\delta\,\|\sigma\|_2^2 + c\eta L + 4c^{3/2}\sigma_{\max}\sqrt{\log\frac{1}{\delta}}\right)\log\left(\frac{4\sqrt{T}\sum_{i=0}^{c-1}\|\sigma_{\Psi_i}\| + I(\delta)}{b_{0,\min}}\right)$$

$$I(\delta) := \|b_0\|_1 + \frac{2\Delta_1}{\eta} + \frac{8\log\frac{1}{\delta}}{b_{0,\min}}\|\sigma\|_2^2 + \sqrt{\log\frac{1}{\delta}}\sum_{i=0}^{c-1}\|\sigma_{\Psi_i}\| + 8\eta Lc\log\frac{4\eta L}{b_{0,\min}}$$

$$H(\delta) := \sum_{i=0}^{c-1}\left(\ln(T/\delta)\,\|\sigma_{\Psi_i}\|^2 + 2\alpha\right)\left(\frac{8\,\|\sigma_{\Psi_i}\|^2\log\frac{1}{\delta}}{b_{0,i}^2} + 2\log\left(1 + \|\sigma_{\Psi_i}\|^2 T + \|\sigma_{\Psi_i}\|^2\log\frac{1}{\delta}\right)\right).$$

■