Adam Can Converge with Proper β_1 and β_2

Anonymous Author(s)
Affiliation
Address
email

Abstract

Adam is one of the most popular algorithms in deep learning. Despite its practical popularity, the convergence of Adam is far from being understood. In this work, we prove that when the 2nd-order momentum parameter β_2 is large and 1st-order momentum parameter $\beta_1 < \sqrt{\beta_2} < 1$, randomly shuffled Adam converges to the neighborhood of stationary points. The size of the neighborhood is propositional to the noise variance of stochastic gradients. With extra strong growth condition, we prove that Adam can converge to stationary points. As β_2 increases, our convergence result can cover any $\beta_1 \in [0, 1)$ including the default setting $\beta_1 = 0.9$. To our best knowledge, we are the first to prove that Adam with any large β_1 can converge *without* any form of bounded gradient assumption.

We also investigate the divergence behavior of Adam. When β_2 is small, we point out a large region of (β_1, β_2) combinations where Adam can diverge to infinity. We emphasize that both our results above consider a quite different setting from the divergence result in Reddi et al. (2018): they consider picking (β_1, β_2) before picking the sample size n; while we consider a more practical scenario with n fixed in advance. Our analysis shows that Adam's behavior changes significantly with hyperparameters. These results may provide new guidance towards better hyperparameter tuning of Adam.

1 Introduction

Modern machine learning tasks often aim to solve the following finite-sum problem.

$$\min_{x \in \mathbb{R}^d} f(x) = \sum_{i=0}^{n-1} f_i(x), \tag{1}$$

where n often stands for the number of samples or mini-batches and x denotes the trainable parameters. When combined with deep neural nets, Adam (Kingma & Ba, 2014) is one of the most dominant algorithms for solving (1). In the past five years, Adam has reached tremendous success in different machine learning applications including natural language processing (NLP) (Vaswani et al., 2017; Brown et al., 2020; Devlin et al., 2018), generative adversarial networks (GANs) (Isola et al., 2017; Zhu et al., 2017; Radford et al., 2015) and computer vision (CV) (Dosovitskiy et al., 2021), etc.. Despite its prevalence, Reddi et al. (2018) point out that Adam can diverge under a wide range of hyperparameters. A main result in (Reddi et al., 2018) conveys the message that [1]:

For any β_1, β_2 s.t. $\beta_1 < \sqrt{\beta_2} < 1$, there exists a problem that Adam will diverge.

Here, β_1 is the hyperparameter to control Adam's 1st-order momentum and β_2 control the weighted average of squared past gradients (a.k.a. 2nd-order momentum). More description can be seen in Algorithm 1 (presented later in Section 2.1). Despite the divergence issue, Adam works well in practice after proper tuning. Even more mysteriously, we find that the commonly reported hyperparameters actually satisfy the divergence condition above. For instance, Kingma & Ba (2014)

[1] For reades who are not familiar with (Reddi et al., 2018), we formally re-state their results in Appendix C.2.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

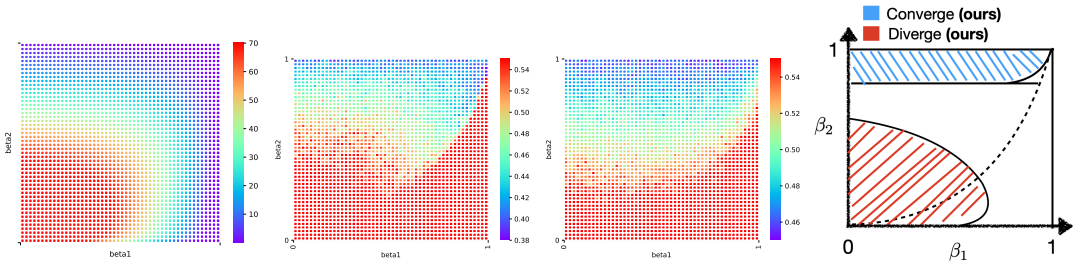
claim that $(\beta_1, \beta_2) = (0.9, 0.999)$ is a "good choice for the tested machine learning problems" and it is indeed the default setting in practice. In recent super-large models GPT-3 and Megatron (Brown et al., 2020; Smith et al., 2022), $(\beta_1, \beta_2) = (0.9, 0.95)$ is chosen as their final hyperparameters. For GAN researchers (e.g. (Isola et al., 2017)), $(\beta_1, \beta_2) = (0.5, 0.999)$ is often preferred. All these hyperparameters satisfy $\beta_1 < \sqrt{\beta_2} < 1$, and thus live in the "divergence region" by (Reddi et al., 2018). However, they all reach good performances and the divergence issue is not observed.

Why does Adam work well despite its divergence issue? One might argue that neural nets training might be somewhat "simpler". We would like to point out it is *not* the crucial reason. Instead, we notice a non-negligible gap between practical applications and the above divergence result: practical applications often fix the number of samples or mini-batches n in advance; while the divergence result picks n *after* picking (β_1, β_2). This difference turns out to be crucial. We explain as follows.

To construct the divergence example, one needs to change n for different (β_1, β_2). This is shown in the proof of (Reddi et al., 2018) and is also restated in Appendix C.2. For instance, for $(\beta_1, \beta_2) = (0, 0.99)$, we need one n to construct the divergent example; for $(\beta_1, \beta_2) = (0, 0.9999)$, we need another n to construct another divergent example. However, in practical applications, sample size n is often given *before* picking (β_1, β_2) so the divergence result cannot be applied. Considering the different settings, the above divergence result might not provide proper guidance for Adam users.

So how would Adam behave when the sample size n is fixed? Can Adam possibly converge? If so, what (β_1, β_2) pairs can ensure convergence? With these questions in mind, we run experiments on different tasks. First, we run Adam under different choices of (β_1, β_2) on a convex function (2) with fixed n (see the definition in Section 3.2). Then, we run Adam on real dataset MNIST and CIFAR-10 with fixed batchsize. We observe some interesting phenomena in Figure 1 (a), (b) and (c).

First, when β_2 is large, the optimization error is small for almost all values of β_1. Second, when $\beta_1,\ \beta_2$ are both small, there is a red region with relatively large error. On MNIST, CIFAR-10, the error in the red region is around 0.4 times higher than that in the blue region. The situation is a lot worse on function (2): the error in the red region is 70 times higher than that in the blue region.



(a) Function (2) (b) MNIST (c) CIFAR-10 (d) Our contribution

Figure 1: **(a), (b), (c)**: The performance of Adam on different tasks. For each task, we show the results with β_1 and β_2 in grids $\{(k_1/50, k_2/50) | k_1 = 0, \cdots, 49, k_2 = 0, \cdots, 49\}$. **(a)**: the optimality gap $x - x^*$ on the convex function (2). **(b)**: The training loss on MNIST. **(c)**: the training loss on CIFAR-10. All experimental details are shown in Appendix D. **(d)**: An illustration of our contribution in (β_1, β_2) phase diagram. The shape of the region follows the solution to our analytic conditions. The size of the region depends on n. The dotted curve satisfies $\beta_1 = \sqrt{\beta_2}$. All figures here consider the setting where n is fixed.

While Adam's performances seem unstable in the red region, we find that *Adam always performs well in the top blue region in Figure 1.* We are led to the following question:

When n is fixed ahead, does Adam provably converge in the top region of Figure 1?

In this work, we theoretically explore this question. When n is fixed in advance, we prove the results below. Our contributions are also visualized in Figure 1 (d).

- When β_2 is large and $\beta_1 < \sqrt{\beta_2}$, we prove that Adam converges to the neighborhood of stationary points. The size of the neighborhood is propositional to noise variance of stochastic gradients. With extra strong growth condition, we prove that Adam can converge to stationary points. As β_2 increases, these results can cover any momentum parameter $\beta_1 \in [0, 1)$ including the default setting $\beta_1 = 0.9$. In particular, our analysis does *not* require bounded gradient assumption.
- We point out the algorithmic divergence in the red region in Figure 1 (d). We prove that: there exists a function such that, Adam diverges to infinity when (β_1, β_2) is picked in the red region

in Figure 1 (d). The size of the red region depends on n. The shape of the region follows the solution to our analytic conditions. We emphasize that our divergence result consider "fix n *before* choosing (β_1, β_2)". This setting is quite different from (Reddi et al., 2018) mentioned in Section 1 which considers a reverse ordering.

We believe our results can boost new understandings for Adam. While Reddi et al. (2018) reveal that "Adam can diverge", our results show the other side of the coin: when n is fixed, Adam can still work if hyperparameters are chosen properly. Since "fixing n" is closer to practical scenarios, our convergence and divergence results may provide better guidance for hyperparameter tuning.

We emphasize two features in this work. First, our convergence results can cover any $\beta_1 \in [0, 1)$, which can involve massive momentum. It turns out that large-momentum Adam is not easy to analyze. Even with stronger assumptions like bounded gradient ($\|\nabla f(x)\| < \mathrm{C}, \forall x$), its convergence is not well understood (see related works in Section 2.2). To our best knowledge, this is the first result that proves Adam with any large β_1 can converge *without* any form of bounded gradient assumption. We will highlight our technical novelty in Section 5.

Second, when n is fixed, we point out Adam's behavior significantly changes under different (β_1, β_2): it converges in certain region but it may diverge in others. Removing boundedness condition is important to characterize such qualitative behavior. With assumed boundedness on gradient or iterates, Adam never diverges, yet we have theoretical and numerical evidence showing that it can.

2 Preliminaries

2.1 Problem Settings

We consider finite-sum problem (1). We use x to denote the optimization variable. We denote ∇f_j as the gradient of f_j and let $\circ$ be the component-wise product. The division and square-root operator are component-wise as well. We present randamly shuffled Adam in Algorithm 1.

Algorithm 1 Adam

Initialize $x_{1,0} = x_0$, $m_{1,-1} = \nabla f(x_0)$ and $v_{1,-1} = \nabla f(x_0) \circ \nabla f(x_0)$.
for $k = 1 \to \infty$ **do**
  Sample $\{\tau_{k,0}, \tau_{k,1}, \cdots, \tau_{k,n-1}\}$ as a random permutation of $\{0, 1, 2, \cdots, n-1\}$
  for $i = 0 \to n-1$ **do**
    $m_{k,i} = \beta_1 m_{k,i-1} + (1-\beta_1) \nabla f_{\tau_{k,i}}(x_{k,i})$
    $v_{k,i} = \beta_2 v_{k,i-1} + (1-\beta_2) \nabla f_{\tau_{k,i}}(x_{k,i}) \circ \nabla f_{\tau_{k,i}}(x_{k,i})$
    $x_{k,i+1} = x_{k,i} - \frac{\eta_k}{\sqrt{v_{k,i}}+\epsilon} \circ m_{k,i}$
  end for
  $x_{k+1,0} = x_{k,n}$
  $v_{k+1,-1} = v_{k,n-1}$
  $m_{k+1,-1} = m_{k,n-1}$
  Break if certain exit condition is satisfied.
end for

In Algorithm 1, m denotes the 1st-order momentum and v denotes the 2nd-order momentum. they are weighted averaged by hyperparameter β_1, β_2, respectively. Larger β_1, β_2 will adopt more history information. We denote $x_{k,i}, m_{k,i}, v_{k,i} \in \mathbb{R}^d$ as the value of x, m, v at the k-th outer loop (epoch) and i-th inner loop (batch), respectively. We choose $\eta_k = \frac{\eta_1}{\sqrt{nk}}$ as the stepsize. In practice, ϵ is adopted for numerical stability and it is often chosen to be 10^{-8}. In our theory, we allow ϵ to be an arbitrary non-negative constant including 0.

In the vanilla version of Adam, it has an additional "bias correction" step (see (Kingma & Ba, 2014)). This "bias correction" step can be implemented by changing the stepsize η_k into $\hat{\eta_k} = \frac{\sqrt{1-\beta_2^k}}{1-\beta_1^k}\eta_k$ and using zero initialization. In Algorithm 1, the "bias correction" step is replaced by a special initialization, which also corrects the bias. Note that $\hat{\eta_k} \in [\sqrt{1-\beta_2}\eta_k, \frac{1}{1-\beta_1}\eta_k]$ is well-bounded near η_k, so η_k and $\hat{\eta_k}$ brings the same convergence rate. In addition, as the effect of initialization becomes negligible when the training progresses, Adam with zero & our initialization will have the same asymptotic behavior. We follow the form of Algorithm 1, which makes results cleaner. In our analysis, we make the assumptions below.

Assumption 2.1. we consider $x \in \mathbb{R}^d$ and $f_i(x)$ satisfies gradient Lipschitz continuous with constant L. We assume $f(x)$ is lower bounded by a finite constant f^*.

Assumption 2.2. $f_i(x)$ and $f(x)$ satisfy: $\sum_{i=0}^{n-1} \|\nabla f_i(x)\|_2^2 \leq D_1 \|\nabla f(x)\|_2^2 + D_0, \forall x \in \mathbb{R}^d$.

Assumption 2.2 is quite general, it covers some stronger conditions. When $D_1 = n$, it becomes the "constant variance" with constant D_0/n. "constant variance" condition is commonly used in both SGD and Adam analysis (e.g. (Ghadimi et al., 2016; Zaheer et al., 2018; Huang et al., 2021)). Assumption 2.2 allows more flexible choices of $D_1 \neq n$ and thus it is weaker than "constant variance".

When $D_0 > 0$, the problem instance is sometimes called "non-realizable" (Shi et al., 2020). In this case, adaptive gradient methods are not guaranteed to reach the exact stationary point. Instead, they only converge to a bounded region (near stationary points) (Zaheer et al., 2018; Shi et al., 2020). This phenomenon indeed occurs for Adam in practice, even with diminishing stepsize (see Appendix D.2). The behavior of SGD is similar: constant stepsize SGD converges to a bounded region with its size propositional to the noise level D_0 (Yan et al., 2018; Yu et al., 2019; Liu et al., 2020b).

When $D_0 = 0$, Assumption 2.2 is often called "strong growth condition" (SGC) (Vaswani et al., 2019). When $\|\nabla f(x)\| = 0$, under SGC we have $\|\nabla f_j(x)\| = 0$ for all j. SGC is sometimes used in SGD analysis e.g. (Schmidt & Roux, 2013; Vaswani et al., 2019). This condition is reasonable in the overparameterized regime where neural networks are capable of interpolating all data points (Vaswani et al., 2019). We will show that Adam can converge stationary points if SGC holds.

When n, f^*, L, D_0, D_1 are fixed a priori, we use $\mathcal{F}_{L,D_0,D_1}^{n,f^*}(\mathbb{R}^d)$ to denote the (sub) function class containing $f(x)$ satisfying Assumption 2.1 and 2.2 with constant n, f^*, etc.. We introduce this notation to clearly present the divergence result in Proposition 3.3. Without this pre-defined function class, the claim of divergence might be misleading or confusing.

2.2 Related Works

Ever since Reddi et al. (2018) pointed out the divergence issue, there are many attempts on designing new variants of Adam. For instance, Zou et al. (2019) and Chen et al. (2018) replace the constant (β_1, β_2) by iterate-dependent ones e.g. β_{1t} or β_{2t}. These methods can address the divergence of the original Adam. Many variants also achieve satisfactory performance in practice. Since we focus on understanding Adam without modification, we introduce more variants later in Appendix C.1.

There are several works trying to understand Adam's convergence behavior, but they often require extra conditions. For instance, the analysis in Zaheer et al. (2018) reveals relation between mini-batch sizes and (non-)convergence of Adam. Gadat & Gavra (2020) point out that Adam only converges to minimizers. However, these two works require $\beta_1 = 0$ and Adam is reduced to RMSProp (Hinton et al., 2012). De et al. (2018) further analyze Adam with non-zero β_1, but they assume the sign of all stochastic gradients to keep the same. It seems unclear how to check this condition a priori. Additionally, they require β_1 inversely related to the upper bound of gradient, which forces β_1 to be small. Défossez et al. (2020) analyze Adam $\beta_1 < \beta_2$ and provide some insights on the momentum mechanisms. However, they require ϵ (the hyperparameter for numerical stability) to be large, in which case ϵ will be much larger than $\sqrt{v}$. This is against the nature of adaptive gradient methods because $\sqrt{v}$ no longer dominates in the choice of stepsize. Two very recent works (Huang et al., 2021) and (Guo et al., 2021) propose novel and simple frameworks to analyze Adam-family with large β_1. Yet, they require the effective stepsize of Adam will be bounded in certain interval, i.e., $\frac{1}{\sqrt{v_t}+\epsilon} \in [C_l, C_u]$ [2]. This boundedness condition changes Adam into AdaBound (Luo et al., 2019) and thus they cannot explain the observations on original Adam in Section 1. Additionally, all of these works (including those for new variants) require bounded gradient assumptions.

Among all these works, (Shi et al., 2020) is the first work that proves RMSProp (a non-momentum version of Adam with $\beta_1 = 0$) can converge without bounded gradient assumption. When the 2nd-order parameter β_2 is large, they show that RMSProp converges to a bounded region in general, and to stationary points under SGC. They also provide a convergence result for Adam with β_1 close enough to 0. However, a simple calculation by Zhang et al. (2022) shows that they require $\beta_1 < 10^{-7}$ on real datasets. Thus their result does not provide much extra information other than RMSProp. It is still unclear how β_1 affects the (non-)convergence of Adam. Proving convergence under large β_1 is non-trivial. We point out there are at least two issues. First, it seems unclear how to control

[2] For completeness, we explain why they require this condition in Appendix C.1.

the massive momentum m_t. Second, the momentum m_t in Adam is multiplied by $1/\sqrt{v_t}$, causing non-linear perturbation. It turns out that we requires much different techniques to resolve these two issues. We highlight our technical novelties in Section 5.

2.3 The Importance and Difficulties of Removing Bounded Gradient Assumptions

Here, we emphasize the importance to remove bounded gradient assumption. First, unlike the assumptions in Section 2.1, bounded gradient is *not* common in SGD analysis. So it is of theoretical interests to remove this condition for Adam. Second, bounded gradient condition rules out the chances of gradient divergence a priori. However, there are numerical evidences showing that Adam's gradient can diverge (see Section 6 and Appendix D.2). Removing the boundedness assumption helps us point out the divergence and convergence behavior of Adam in the (β_1, β_2) diagram.

However, it is often difficult to remove the boundedness condition, especially for stochastic algorithms with momentum. For instance, An early paper Bertsekas & Tsitsiklis (2000) analyzed SGD-type methods without any boundedness condition. But it is not until recently that Yu et al. (2019); Liu et al. (2020b) prove SGDM (SGD with momentum) converges without bounded gradient assumption. [3] Such attempts of removing boundedness assumptions are often appreciated for general optimization problems where "bounded-assumption-free" is considered as a major contribution.

For Adam, the role of momentum is even more intricate since it is multiplied by $1/\sqrt{v_t}$. Combined with v_t, the impact of previous signals not only affect the update direction, but also change the stepsize for each component. As such, the role of momentum in Adam could be much different from that in SGDM or GDM. Even with boundedness conditions, the convergence of large-β_1 Adam is still not well understood (see related works in Section 2.2). In this work, we propose new techniques to handle Adam's momentum under any large β_1, regardless of the gradient magnitude. These techniques are not revealed in any existing works. We introduce our technical contribution in Section 5.

3 Main Results

3.1 Convergence Results

Here, we give the convergence results under large β_2.

Theorem 3.1. *For any $f(x) \in \mathcal{F}^{n,f^*}_{L,D_0,D_1}(\mathbb{R}^d)$, when the hyperparameters in Algorithm 1 satisfies: $\beta_1 < \sqrt{\beta_2} < 1$; β_2 is greater or equal to a threshold $\gamma_1(n)$; and $\eta_k = \frac{\eta_1}{\sqrt{nk}}$, we will have:*

$$\min_{k\in[k_m,T]} \mathbb{E}\left\{\min\left[\sqrt{\frac{2D_1 d}{D_0}}\|\nabla f(x_{k,0})\|_2^2, \|\nabla f(x_{k,0})\|_2\right]\right\} = \mathcal{O}\left(\frac{\log T}{\sqrt{T}}\right) + \mathcal{O}(\sqrt{D_0}),$$

where $k_m \in \mathbb{N}$ satisfies $k_m \geq 4$ and $\beta_1^{(k_m-1)n} \leq \frac{\beta_1^n}{\sqrt{k_m-1}}$. [4]

Remark 1: the choice of β_2. The threshold $\gamma_1(n)$ satisfies $\beta_2 \geq 1 - \mathcal{O}\left(\frac{1-\beta_1^n}{n^2\rho}\right)$ (see inequality (32) and Remark G.8), where ρ is a constant that depends on the training trajectory. In worst cases, ρ is upper bounded by $n^{2.5}$, but we find the practical ρ to be much smaller. In Appendix D.2, we estimate ρ on MNIST and CIFAR-10. In practical training process, we empirically observe that $\rho \approx \mathcal{O}(n)$, thus the required $\gamma_1(n) \approx 1 - \mathcal{O}\left(n^{-3}\right)$. Note that our threshold of β_2 is a sufficient condition for convergence, so there may be a gap between the practical choices and the theoretical bound of β_2. Closing the gap will be an interesting future direction.

We further remark that $\gamma_1(n)$ increases with n. This condition suggests that larger β_2 should be used when the batchsize is small. This phenomenon is also verified by our experiments in Figure 4 (d), Section 6. Intuitively, this phenomenon is understandable since larger n brings larger noise in the stochastic gradient, and thus larger β_2 is required.

Remark 2: the choice of β_1. $\beta_1 < \sqrt{\beta_2}$ is also a requirement in (Kingma & Ba, 2014), but their convergence analysis requires bounded gradient and Reddi et al. (2018) later pointed out that the proof has flaws. Theorem 3.1 is the first result that proves the convergence of Adam with the same choice of β_1, but with less assumptions. Since β_2 is suggested to be large, our convergence result can cover flexible choice of $\beta_1 \in [0, 1)$. For instance, $\beta_2 = 0.999$ brings the threshold of $\beta_1 < 0.9995$, which covers and basically all practical choices of β_1 reported in the literature (see Section 1), including the default setting $\beta_1 = 0.9$.

[3] A recent work (Jin et al., 2022) further extend the results to SGDM with adaptive stepsize.

[4] When $\beta_1 = 0.9$, $k_m = 15$ for any $n \geq 1$.

Remark 3: convergence to a bounded region. When $D_0 > 0$, Adam converges to a bounded region near stationary points. As discussed in Section 2.1, converging to bounded region is common for stochastic methods including constant-stepsize SGD (Yan et al., 2018; Yu et al., 2019; Liu et al., 2020b) and diminishing-stepsize RMSProp (Zaheer et al., 2018; Shi et al., 2020). This phenomenon is also observed in practice: even for convex quadratic function with $D_0 > 0$, Adam with diminishing stepsize *cannot* reach exactly zero gradient (see Appendix D.2). The good news is that, the constant $\mathcal{O}(\sqrt{D_0})$ vanishes to 0 as β_2 goes to 1. This can be seen in Remark G.17 in Appendix G.9.

As a corollary of Theorem 3.1, we have the following result under SGC (i.e., $D_0 = 0$).

Corollary 3.2. *Under the setting in Theorem 3.1. When $D_0 = 0$ for Assumption 2.2, we have*

$$\min_{k\in[k_m,T]} \mathbb{E}\|\nabla f(x_{k,0})\|_2 = \mathcal{O}\left(\frac{\log T}{\sqrt{T}}\right),$$

where $k_m \in \mathbb{N}$ satisfies $k_m \geq 4$ and $\beta_1^{(k_m-1)n} \leq \frac{\beta_1^n}{\sqrt{k_m-1}}$.

Under SGC (i.e. $D_0 = 0$), Corollary 3.2 states that Adam can converge to stationary points. This is indeed the case in practice. For instance, function (2) satisfies SGC and we observe 0 gradient norm after Adam converges (see Appendix D.2). The convergence rate in Corollary 3.2 is comparable to that of SGD under the same condition in (Vaswani et al., 2019).

3.2 Divergence Results

Theorem 3.1 shows that when β_2 is large, any $\beta_1 < \sqrt{\beta_2}$ ensures convergence. Now we consider the case where β_2 is small. We will show that in this case, a wide range of β_1 is facing the risk of diverging to infinity. We construct a counter-example in $\mathcal{F}^{n,f^*}_{L,D_0,D_1}(\mathbb{R}^d)$. Consider $f(x) = \sum_{i=0}^{n-1} f_i(x)$ for $x \in \mathbb{R}$, we define $f_i(x)$ as:

$$\begin{aligned} f_i(x) &= \begin{cases} nx, & x \geq -1 \\ \frac{n}{2}(x+2)^2 - \frac{3n}{2}, & x < -1 \end{cases} \quad \text{for } i = 0, \\ f_i(x) &= \begin{cases} -x, & x \geq -1 \\ -\frac{1}{2}(x+2)^2 - \frac{3}{2}, & x < -1 \end{cases} \quad \text{for } i > 0. \end{aligned} \tag{2}$$

Summing up all the $f_i(x)$, we can see that

$$f(x) = \begin{cases} x, & x \geq -1 \\ \frac{1}{2}(x+2)^2 - \frac{3}{2}, & x < -1 \end{cases}$$

is a lower bounded convex smooth function with optimal solution $x^* = -2$. Function (2) allows both iterates and gradients to diverge to infinity.

As shown in Figure 1 (a), when running Adam on (2), there exists a red large-error region. This shows the sign of divergence. We further theoretically verify the conjecture.

Proposition 3.3. *For any function class $\mathcal{F}^{n,f^*}_{L,D_0,D_1}(\mathbb{R}^d)$, there exists a $f(x) \in \mathcal{F}^{n,f^*}_{L,D_0,D_1}(\mathbb{R}^d)$, s.t. when (β_1, β_2) satisfies analytic condition* (9), (10), (11) *in Appendix E, Adam's iterates and function values diverge to infinity. By solving these conditions in* `NumPy`*, we plot the orange region in 2. The size of the region depends on n and it expands to the whole region when n goes to infinity.*

We present the proof in Appendix E. We find the "divergence region" always stays below the "convergence threshold" $\gamma_1(n)$ in Theorem 3.1, so the two results are self-consistent (see the remark in Appendix E). Proposition 3.3 states the divergence of iterates and function values. Our experiments also show the divergence of gradient (see Section 6 and Appendix D.2). These results characterize Adam's divergence behavior both numerically and theoretically.

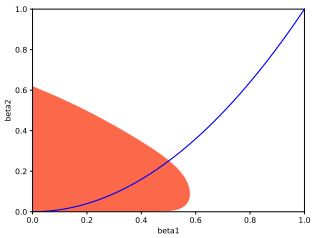


Figure 2: On function (2) with $n = 20$, Adam diverges in the colored region. The region is plotted by solving condition (9), (10), (11) in `NumPy`. The blue curve satisfies $\beta_1 = \sqrt{\beta_2}$.

We emphasize that the orange region is *not* discussed in (Reddi et al., 2018) because we consider n fixed while they allow n changing. When n is allowed to increase, our orange region will expand to the whole region and thus we can derive a similar result as (Reddi et al., 2018). We provide more explanation in Section 4. Combining Theorem 3.1 and Proposition 3.3, we establish a clearer image on the relation between (β_1, β_2) and qualitative behavior of Adam.

4 Reconciling Our Results with (Reddi et al., 2018)

We discuss more on the relation between (Reddi et al., 2018) and our results. Recall Reddi et al. (2018) show that "for any $\beta_1 < \sqrt{\beta_2} < 1$, Adam can diverge". This does not contradict with our convergence results in Theorem 3.1. Further, it is different from our divergence result in Proposition 3.3. As discussed in Section 1, the key difference lies in whether (β_1, β_2) is picked *before or after* picking n, which often stands for sample size or the number of mini-batch.

Since n is fixed when the (sub-)function class $\mathcal{F}^{n,f^*}_{L,D_0,D_1}(\mathbb{R}^d)$ is given, this key difference can also be expressed as whether (β_1, β_2) is picked *before or after* picking $\mathcal{F}^{n,f^*}_{L,D_0,D_1}(\mathbb{R}^d)$. Now we discuss the following two cases.

Case I: When (β_1, β_2) is picked before picking $\mathcal{F}^{n,f^*}_{L,D_0,D_1}(\mathbb{R}^d)$. As discussed in Section 1, the divergence result requires different n for different (β_1, β_2). In this sense, the considered function class is constantly changing. It does *not* contradict with our Theorem 3.1 which considers a fixed function class with fixed n. For **Case I**, we illustrate Adam's behavior in Figure 3. The red region with is proved by (Reddi et al., 2018). For completeness, we remove the condition "$\beta_1 < \sqrt{\beta_2}$" and further prove that Adam will diverge to infinity for *any* $(\beta_1, \beta_2) \in [0, 1)^2$. The result is shown in the following Corollary 4.1.

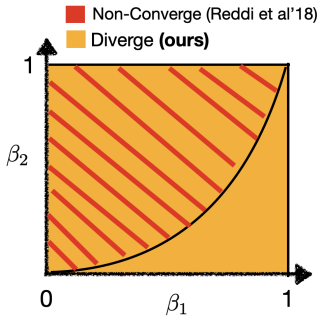


Figure 3: Adam's behavior when (β_1, β_2) in **Case I**. "Non-converge" means that the iterates will stay in a huge region, e.g. the whole feasible set for constrained problems. "Diverge" means diverge to infinity.

Corollary 4.1. *For any $(\beta_1, \beta_2) \in [0, 1)^2$, there exists a function satisfying Assumption 2.1 and 2.2 that the Adam's iterates and function values diverge to infinity.*

Rigorously speaking, Reddi et al. (2018) consider constrained problems. So their claimed "divergence" actually means the iterates will stay in a huge region with the size of whole feasible set. We call it "non-convergence" to distinguish from our result of "diverge to infinity" in Figure 3.

Proof of Corollary 4.1 can be seen in the final paragraph in Appendix E. In the proof, we also require different n to cause divergence for different (β_1, β_2). So the function class is constantly changing. As a result, in **Case I**, we cannot prove any convergence result.

Table 1: Possible algorithmic behaviors of Adam in **Case II**.

Setting	Hyperparameters	Adam's behavior
$\forall f \in \mathcal{F}^{n,f^*}_{L,D_0,D_1}(\mathbb{R}^d)$ with $D_0 = 0$	β_2 is large and $\beta_1 < \sqrt{\beta_2}$	converge to stationary points **(Ours)**
$\forall f \in \mathcal{F}^{n,f^*}_{L,D_0,D_1}(\mathbb{R}^d)$ with $D_0 \neq 0$	β_2 is large and $\beta_1 < \sqrt{\beta_2}$	converge to a bounded region with size $\mathcal{O}(D_0)$ **(Ours)**
$\exists f \in \mathcal{F}^{n,f^*}_{L,D_0,D_1}(\mathbb{R}^d)$	The orange region in Figure 2	diverge to infinity **(Ours)**

Case II: When (β_1, β_2) is picked after picking $\mathcal{F}^{n,f^*}_{L,D_0,D_1}(\mathbb{R}^d)$. When the function class is picked in advance, sample size n will also be fixed. This case is closer to most practical applications. In this case, we find that Adam's behavior changes significantly in the different region of Figure 3. First, $\forall f(x) \in \mathcal{F}^{n,f^*}_{L,D_0,D_1}(\mathbb{R}^d)$ will converge when $\beta_1 < \sqrt{\beta_2}$ and β_2 is large. Second, $\exists f(x) \in \mathcal{F}^{n,f^*}_{L,D_0,D_1}(\mathbb{R}^d)$ will diverge to infinity when (β_1, β_2) are in the orange region in Figure 2. Since **Case II** is closer to practical scenarios, these results can provide better guidance for hyperparameter tuning for Adam users. We provide some suggestions in Appendix B.

For **Case II**, we summarize the possible behaviors of Adam in Table 1. We also illustrate our convergence and divergence results in Figure 1 (d). Note that there are some blanket areas where Adam's behavior remains unknown, this part will be left as interesting future work.

5 Proof Ideas for the Convergence Result

We now briefly introduce our proof ideas for the convergence result in Theorem 3.1. Simply put, we need to control the update direction $m_{k,i}/\sqrt{v_{k,i}}$ near the negative gradient direction. Namely, we

want to show

$$\mathbb{E}\langle \nabla f(x_{k,0}), \sum_{i=0}^{n-1} \frac{m_{k,i}}{\sqrt{v_{k,i}}} \rangle \approx \mathbb{E}\langle \nabla f(x_{k,0}), \nabla f(x_{k,0}) \rangle. \tag{3}$$

However, directly proving (3) could be difficult because both $m_{k,i}$ and $v_{k,i}$ distort the trajectory. To start with, we try to control the movement of $v_{k,i}$. We use a standard approach for Adam-type algorithms: use large or increasing β_2 (e.g. (Shi et al., 2020; Zou et al., 2019; Chen et al., 2018)). Recall $v_{k,i} = (1-\beta_2)\sum_{j=1}^{i} \beta_2^{i-j} \nabla f_{\tau_{k,j}}(x_{k,j}) \circ \nabla f_{\tau_{k,j}}(x_{k,j}) + \beta_2^i v_{k,0}$, we have $v_{k,i} \approx v_{k,0}$ when β_2 is large. In this case, we have:

$$\mathbb{E}\left\langle \nabla f(x_{k,0}), \sum_{i=0}^{n-1} \frac{m_{k,i}}{\sqrt{v_{k,i}}} \right\rangle \approx \mathbb{E}\left\langle \frac{\nabla f(x_{k,0})}{\sqrt{v_{k,0}}}, \sum_{i=0}^{n-1} m_{k,i} \right\rangle.$$

Now we need to show:

$$\mathbb{E}\left\langle \frac{\nabla f(x_{k,0})}{\sqrt{v_{k,0}}}, \left(\sum_{i=0}^{n-1} m_{k,i}\right) - \nabla f(x_{k,0}) \right\rangle \overset{(*)}{=} \mathbb{E}\left(\sum_{l=1}^{d}\sum_{i=0}^{n-1} \frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,0}}} \left(m_{l,k,i} - \partial_l f_{\tau_{k,i}}(x_{k,0})\right)\right) \approx 0, \tag{4}$$

where $\partial_l f(x_{k,0})$ is the l-th component of $\nabla f(x_{k,0})$, similarly for $m_{l,k,0}$ and $v_{l,k,0}$. $(*)$ is due to the finite-sum structure. However, it is not easy to prove (4). We point out some technical issues below.

Issue I: massive momentum. Directly proving (4) is still not easy. We need to first consider a simplified problem: for every $l \in [d]$, assume we treat $\partial_l f(x_{k,0})/\sqrt{v_{l,k,0}}$ as a constant, how to bound $\mathbb{E}\sum_{i=0}^{n-1}\left(m_{l,k,i} - \partial_l f_{\tau_{k,i}}(x_{k,0})\right)$?

It turns out that this simplified problem is still non-trivial. When β_1 is large, $m_{l,k,i}$ contains heavy historical signals which significantly distort the trajectory from gradient direction. Existing literatures (Zaheer et al., 2018; De et al., 2018; Shi et al., 2020) take a naive approach: they set $\beta_1 \approx 0$ so that $m_{l,k,i} \approx \partial_l f_{\tau_{k,i}}(x_{k,i})$. Then we get (4) ≈ 0. However, this method cannot be applied here since we are interested in practical cases where β_1 is large in $[0, 1)$.

Issue II: non-linear dynamics. Even if we solve **Issue I**, it is still unclear how to prove (4). This is because: for every $l \in [d]$, $\partial_l f(x_{k,0})/\sqrt{v_{l,k,0}}$ is a r.v. instead of a constant. With this term involved, we are facing with a stochastic non-linear dynamics, which could be difficult to analyze.

Unfortunately, even with additional assumption like bounded gradient, there is no general approach to tackle the above issues. In this work, we propose a solution regardless of gradient magnitude.

Step 1. We prove the following Lemma to resolve **Issue I**.

Lemma 5.1. *(Informal) Consider Algorithm 1. For every $l \in [d]$ and any $\beta_1 \in [0, 1)$, we have the following result under Assumption 2.1.*

$$\delta(\beta_1) := \mathbb{E}\sum_{i=0}^{n-1}\left(m_{l,k,i} - \partial_l f_{\tau_{k,i}}(x_{k,0})\right) = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right),$$

where $\partial_l f(x_{k,0})$ is the l-th component of $\nabla f(x_{k,0})$; $m_{l,k,i} = (1-\beta_1)\partial_l f(x_{k,i}) + \beta_1 m_{l,k,i-1}$.

We present the proof idea in Appendix A. Simply put, we construct a simple toy example called "color-ball" model, which shows a special property of $\delta(\beta_1)$. We find out: for Algorithm 1, error terms from successive epochs can be canceled, which keeps the momentum roughly in the descent direction. This important property is not revealed in any existing work. In Appendix A, we demonstrate the "color-ball" model in simple language where only basic probability knowledge is required.

Remark 4: When assuming bounded gradient $\|\nabla f(x)\| \leq G$, a naive upper bound would be $\delta(\beta_1) = \mathcal{O}(G)$. However, such constant upper bound does not imply $\delta(\beta_1)$ is close to 0. It will not help prove the convergence. This might be partially the reason why large-β_1 Adam is hard to analyze even under bounded gradient (see related works in Section 2.2). We emphasize Lemma 5.1 holds true regardless of gradient norm, so it could be deployed in both bounded or unbounded gradient analysis.

Step 2. We try to show (4) by adopting Lemma 5.1. However, the direct application cannot work since $\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,0}}}$ is not constant. Despite its randomness, we find out that when β_2 is large,

$\left\{\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,0}}}\right\}_{k=1}^{\infty}$ behaves like Cauchy sequence which has unique limit point. As such, although $\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,0}}}$ brings extra perturbation, the quantity in (4) share the similar asymptotic behavior as $\delta(\beta_1)$. Therefore, with the help of **Step 1**, we prove the following Lemma 5.2.

Lemma 5.2. *(Informal) Under Assumption 2.1 and 2.2, consider Algorithm 1 with large β_2 and $\beta_1 < \sqrt{\beta_2}$. For those l with gradient component larger than certain threshold, we have:*

$$\left|\frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}} - \frac{\partial_l f(x_{k-1,0})}{\sqrt{v_{k-1,0}}}\right| = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right);$$

$$\mathbb{E}\left(\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,0}}}\sum_{i=0}^{n-1}(m_{l,k,i} - \partial_l f_{\tau_{k,i}}(x_{k,0}))\right) = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right).$$

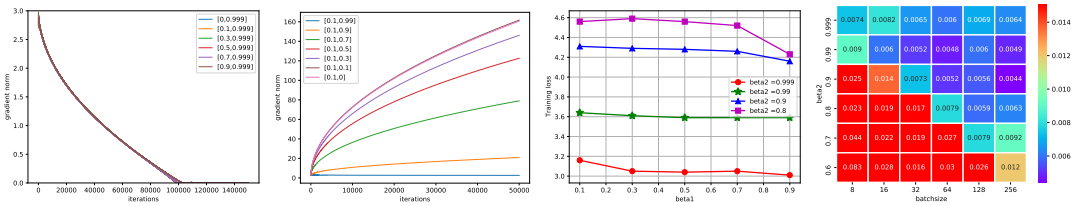
We conclude the proof of (4) by some additional analysis on "those l with small gradient component". This case is a bit easier since it reduce to bounded gradient case. Since the whole proof is quite long, we provide a proof sketch in Appendix G.1. The whole proof is presented in Appendix G.

6 Experiments

To support our theory, we provide more simulations and real-data experiments. All the experimental settings and hyperparameters are presented in Appendix D. We aim to show:

(I). When β_2 is large, a large range of β_1 gives good performance, including all $\beta_1 < \sqrt{\beta_2}$.

(II). When β_2 is small, a large range of β_1 performs relatively badly.



(a) Function (2) with large β_2 (b) Function (2) with small β_2 (c) WikiText-103 (d) MNIST (batchsize v.s. β_2)

Figure 4: The performance of Adam on different tasks. **(a) (b):** The final gradient norm on (2) under different (β_1, β_2). We use $n = 20$ and initialization $x = -5$. **(c):** The training loss on WikiText-103. **(d):** The training loss on MNIST under different batchsize and β_2. Here, β_1 is fixed to be 0.9.

Counter-example (2)**.** We run more experiments on (2) with initialization $x = -5$ and $n = 5, 10, 15, 20$. We show the result of $n = 20$ in Figure 4 (a), (b); the rest are shown in Appendix D.2. We find that: first, when β_2 is small, gradient norms increase dramatically. This emphasizes the importance of discarding bounded gradient assumptions. Second, when β_2 is large, Adam converges to stationary points for $\beta_1 < \sqrt{\beta_2}$. This matches our theory since (2) satisfies $D_0 = 0$. For functions with $D_0 > 0$, we further show that Adam can only converge to a bounded region in Appendix D.2.

MNIST and CIFAR-10. As shown in Figure 1 (b)& (c) in Section 1, the training results match both claim **(I)** and **(II).** In addition, there is a convex-shaped boundary on the transition from low loss to higher loss, this boundary roughly matches the condition in Theorem 3.1.

NLP. We use Adam to train Transformer XL (Dai et al., 2019) on the WikiText-103 dataset (Merity et al., 2016). This architecture and dataset is widely used in NLP tasks (e.g. (Howard & Ruder, 2018; See et al., 2017)). As shown in Figure 4 (c), the training results match both claim **(I)** and **(II).**

Batchsize and β_2. As shown in Figure 4 (d): on MNIST, smaller batchsize requires larger β_2 to reach small loss. This matches the message by Theorem 3.1: the threshold of β_2 increases with n.

7 Conclusions

In this work, we explore the (non-)convergence of Adam. When β_2 is large, we prove that Adam can converge with any $\beta_1 < \sqrt{\beta_2}$. When β_2 is small, we further show that Adam might diverge to infinity for a wide range of β_1. One interesting question is to verify the advantage of Adam over SGD. In this work, we focus on the fundamental issue of convergence. Proving faster convergence of Adam would be our future work.

References

Bertsekas, D. P. and Tsitsiklis, J. N. Gradient convergence in gradient methods with errors. SIAM Journal on Optimization, 10(3):627–642, 2000.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. arXiv preprint arXiv:2005.14165, 2020.

Chen, C., Shen, L., Zou, F., and Liu, W. Towards practical adam: Non-convexity, convergence theory, and mini-batch acceleration. arXiv preprint arXiv:2101.05471, 2021.

Chen, X., Liu, S., Sun, R., and Hong, M. On the convergence of a class of adam-type algorithms for non-convex optimization. arXiv preprint arXiv:1808.02941, 2018.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., and Salakhutdinov, R. Transformer-xl: Attentive language models beyond a fixed-length context. arXiv preprint arXiv:1901.02860, 2019.

De, S., Mukherjee, A., and Ullah, E. Convergence guarantees for rmsprop and adam in non-convex optimization and an empirical comparison to nesterov acceleration. arXiv preprint arXiv:1807.06766, 2018.

Défossez, A., Bottou, L., Bach, F., and Usunier, N. On the convergence of adam and adagrad. arXiv e-prints, pp. arXiv–2003, 2020.

Deng, L. The mnist database of handwritten digit images for machine learning research. IEEE Signal Processing Magazine, 29(6):141–142, 2012.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

Dozat, T. Incorporating nesterov momentum into adam. 2016.

Gadat, S. and Gavra, I. Asymptotic study of stochastic adaptive algorithm in non-convex landscape. arXiv preprint arXiv:2012.05640, 2020.

Ghadimi, S., Lan, G., and Zhang, H. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. Mathematical Programming, 155(1):267–305, 2016.

Guo, Z., Xu, Y., Yin, W., Jin, R., and Yang, T. A novel convergence analysis for algorithms of the adam family. arXiv preprint arXiv:2112.03459, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778, 2016.

Hinton, G., Srivastava, N., and Swersky, K. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. Cited on, 14(8):2, 2012.

Howard, J. and Ruder, S. Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146, 2018.

Huang, F., Li, J., and Huang, H. Super-adam: Faster and universal framework of adaptive gradients. arXiv preprint arXiv:2106.08208, 2021.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1125–1134, 2017.

Jin, R., Xing, Y., and He, X. On the convergence of msgd and adagrad for stochastic optimization. arXiv preprint arXiv:2201.11204, 2022.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., and Han, J. On the variance of the adaptive learning rate and beyond. In International Conference on Learning Representations, 2020a. URL https://openreview.net/forum?id=rkgz2aEKDr.

Liu, Y., Gao, Y., and Yin, W. An improved analysis of stochastic gradient descent with momentum. arXiv preprint arXiv:2007.07989, 2020b.

Luo, L., Xiong, Y., and Liu, Y. Adaptive gradient methods with dynamic bound of learning rate. In International Conference on Learning Representations, 2019. URL https://openreview.net/forum?id=Bkg3g2R9FX.

Merity, S., Xiong, C., Bradbury, J., and Socher, R. Pointer sentinel mixture models. arXiv preprint arXiv:1609.07843, 2016.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434, 2015.

Reddi, S. J., Kale, S., and Kumar, S. On the convergence of adam and beyond. In International Conference on Learning Representations, 2018. URL https://openreview.net/forum?id=ryQu7f-RZ.

Schmidt, M. and Roux, N. L. Fast convergence of stochastic gradient descent under a strong growth condition. arXiv preprint arXiv:1308.6370, 2013.

See, A., Liu, P. J., and Manning, C. D. Get to the point: Summarization with pointer-generator networks. arXiv preprint arXiv:1704.04368, 2017.

Shi, N., Li, D., Hong, M., and Sun, R. Rmsprop converges with proper hyper-parameter. In International Conference on Learning Representations, 2020.

Smith, S., Patwary, M., Norick, B., LeGresley, P., Rajbhandari, S., Casper, J., Liu, Z., Prabhumoye, S., Zerveas, G., Korthikanti, V., et al. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. arXiv preprint arXiv:2201.11990, 2022.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In Advances in neural information processing systems, pp. 5998–6008, 2017.

Vaswani, S., Bach, F., and Schmidt, M. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In The 22nd International Conference on Artificial Intelligence and Statistics, pp. 1195–1204. PMLR, 2019.

Yan, Y., Yang, T., Li, Z., Lin, Q., and Yang, Y. A unified analysis of stochastic momentum methods for deep learning. arXiv preprint arXiv:1808.10396, 2018.

Yu, H., Jin, R., and Yang, S. On the linear speedup analysis of communication efficient momentum sgd for distributed non-convex optimization. In International Conference on Machine Learning, pp. 7184–7193. PMLR, 2019.

Zaheer, M., Reddi, S., Sachan, D., Kale, S., and Kumar, S. Adaptive methods for nonconvex optimization. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), Advances in Neural Information Processing Systems, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/90365351ccc7437a1309dc64e4db32a3-Paper.pdf.

Zhang, Y., Chen, C., and Luo, Z.-Q. Does adam converge and when? In ICLR Blog Track, 2022. URL https://iclr-blog-track.github.io/2022/03/25/does-adam/. https://iclr-blog-track.github.io/2022/03/25/does-adam/.

Zhou, Z., Zhang, Q., Lu, G., Wang, H., Zhang, W., and Yu, Y. Adashift: Decorrelation and convergence of adaptive learning rate methods. arXiv preprint arXiv:1810.00143, 2018.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE international conference on computer vision, pp. 2223–2232, 2017.

Zou, F., Shen, L., Jie, Z., Zhang, W., and Liu, W. A sufficient condition for convergences of adam and rmsprop. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11127–11135, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Negative Social Impact

This script may provide better guidance for deep neural network training. It would have certain negative social impact if neural nets are deployed for illegal usage.

Appendix Organization

The Appendix is organized as follows.

- Appendix A discusses the ideas of "color-ball" model and its application in proving Lemma 5.1. This part is important for proving Theorem 3.1.
- Appendix B provide some suggestions for hyperparameter tuning of Adam.
- Appendix C.1 provide more discussions on some recent related works. Appendix C.2 re-state the non-convergence results in (Reddi et al., 2018).
- Appendix D provides more detailed setting of the experiments, as well as some additional experiments mentioned in the full script.
- Appendix E provides detailed proof for Proposition 3.3.
- Appendix F provides more some notations and technical lemmas that serve for the proof of Theorem 3.1.
- Appendix G provides detailed proof for Theorem 3.1. Especially, Appendix G.1 provide a proof roadmap.

A An Introduction to the "Color-Ball" Model

We now introduce ideas of the "color-ball" model and its application in bounding $\delta(\beta_1)$ in Lemma 5.1. This part is important for proving Theorem 3.1.

As discussed in Section 5, we wish to show that $\delta(\beta_1)$ vanishes with k. Formally, we wish to get the following equation (5) for every $l \in [d]$.

$$\delta(\beta_1) = \left| \mathbb{E}\left[\sum_{i=0}^{n-1} \left(m_{l,k,i} - \partial_l f_i\left(x_{k,0}\right) \right) \right] \right| = \mathcal{O}(\beta_1^k), \quad \forall \beta_1 \in [0,1) \tag{5}$$

When k is large, $\mathcal{O}(\beta_1^k)$ vanishes faster than $\mathcal{O}(\frac{1}{\sqrt{k}})$ and thus Lemma 5.1 can be proved. In the following context, we will carefully quantify the mismatch between $m_{l,k,i}$ and the $\partial_l f(x_{k,0})$. We find out that the error terms from successive epochs can be cancelled, which keeps the momentum roughly in the descent direction. To help readers understand our idea, we consider a toy model called the "color-ball " model as follows.

The "color-ball " model. Consider a box containing two balls labeled with constant $c_0, c_1 \in \mathbb{R}$, respectively. In each round (epoch), we randomly sample balls from the box without replacement, then we put them back. We denote the 1st sampled label in the k-th epoch as a_k and the 2nd sampled one as b_k; $a_k, b_k \in \{c_0, c_1\}$. We define two random variables m_0 and m_1 as follows (assume $\beta \in [0,1)$):

$$m_1 = \underbrace{b_k + \beta a_k}_{m_{1,k}} + \underbrace{\beta^2 b_{k-1} + \beta^3 a_{k-1}}_{m_{1,k-1}} + \cdots + \underbrace{\beta^{2(k-1)} b_1 + \beta^{2(k-1)+1} a_1}_{m_{1,1}};$$

$$m_0 = \underbrace{\qquad a_k}_{m_{0,k}} + \underbrace{\beta^1 b_{k-1} + \beta^2 a_{k-1}}_{m_{0,k-1}} + \cdots + \underbrace{\beta^{2(k-1)-1} b_1 + \beta^{2(k-1)} a_1}_{m_{1,1}};$$

where $m_{0,k}$ denotes the summand of m_0 in k-th epoch, similarly for $m_{1,k}$. Note that in each epoch, m_0 and m_1 share the same sample order. Further, we introduce the following deterministic constants.

$$f_1 = c_1(\underbrace{1+\beta}_{f_{1,k}} + \underbrace{\beta^2+\beta^3}_{f_{1,k-1}} \cdots + \underbrace{\beta^{2(k-1)}+\beta^{2(k-1)+1}}_{f_{1,1}});$$

$$f_0 = c_0(\underbrace{1+\beta}_{f_{0,k}} + \underbrace{\beta^2+\beta^3}_{f_{0,k-1}} \cdots + \underbrace{\beta^{2(k-1)}+\beta^{2(k-1)+1}}_{f_{0,1}});$$

where $f_{0,k}$ denotes the summand of f_0 in k-th epoch, similarly for $f_{1,k}$. Now we prove the following Lemma A.1.

Lemma A.1. *In this color-ball example, we have*

$$\left|\mathbb{E}\left[\sum_{i=0}^{1} m_i - \sum_{i=0}^{1} f_i\right]\right| = \beta^{2(k-1)+1}\left(\frac{c_0}{2}+\frac{c_1}{2}\right),$$

where the expectation is taken on all the possible draws. For the color-ball example with $n \geq 2$ balls, we have

$$\left|\mathbb{E}\left[\sum_{i=0}^{n-1} m_i - \sum_{i=0}^{n-1} f_i\right]\right| = \beta^{n(k-1)}\sum_{i=0}^{n-1} c_i(\frac{1}{n}\beta^1 \cdots + \frac{n-1}{n}\beta^{n-1}).$$

How is Lemma A.1 related to (5)**?** In this color-ball toy example, m_i mimics the possible realization of momentum up to the i-th inner loop in k-th epoch. f_i mimics the stochastic gradient $\nabla f_i\left(x_{k,0}\right)$. This is because we can expand $\nabla f_i\left(x_{k,0}\right)$ into an infinite-sum sequence $\nabla f_i\left(x_{k,0}\right) = (1-\beta_1)\nabla f_i\left(x_{k,0}\right)\sum_{j=0}^{\infty}\beta_1^j$, which shares a similar structure as f_i. In this sense, Lemma A.1 may provide ideas to prove (5). Nevertheless, there are still gap between these two, we will explain the gap later.

Proof. We use $\mathbb{E}_k\left[\cdot\right]$ to denote the conditional expectation given all the history up to the beginning of k-th epoch. Since $\mathbb{E}\left[\cdot\right] = \mathbb{E}\left[\mathbb{E}_k\left[\cdot\right]\right]$, we first calculate $\mathbb{E}_k\left[\sum_{i=0}^{1} m_i\right]$. Since all the history before k-th epoch is fixed, we relegate this part to later discussion and first focus on the expectation of $\sum_{i=0}^{1} m_{i,k}$. As shown in Figure 5 (upper part), there are 2 possible realization.

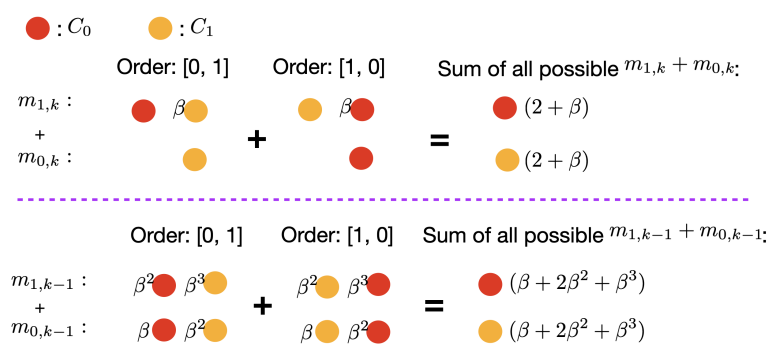


Figure 5: All possible realization of $\sum_{i=0}^{1} m_{i,k}$ and $\sum_{i=0}^{1} m_{i,k-1}$.

With the help of Figure 5 (upper part), we have the following result.

$$\begin{aligned}\mathbb{E}_k\left[\sum_{i=0}^{1} m_{i,k} - \sum_{i=0}^{1} f_{i,k}\right] &= \mathbb{E}_k\left[\sum_{i=0}^{1} m_{i,k}\right] - (1+\beta)\left(c_0+c_1\right) \\ &= -\frac{\beta}{2}(c_0+c_1) \qquad (6)\end{aligned}$$

Now we move one step further to calculate $\mathbb{E}_{k-1}\mathbb{E}_k\left[\sum_{i=0}^{1} m_i - \sum_{i=0}^{1} f_i\right]$. Using the similar strategy as (6), we have the following result. The calculation is illustrated in Figure 5 (lower part) and Figure 6.

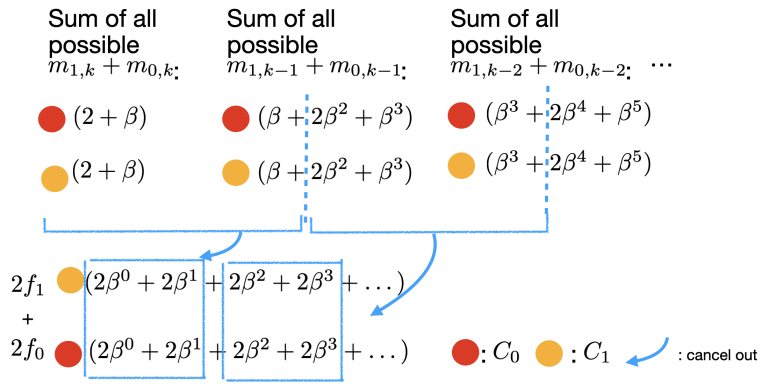


Figure 6: For every k, $\mathbb{E}_k\left[\sum_{i=0}^{1} m_{i,k} - \sum_{i=0}^{1} f_{i,k}\right]$ will create residues, while these residues will be canceled out in the $(k-1)$-th epoch.

$$
\begin{aligned}
& \mathbb{E}_{k-1}\left\{\mathbb{E}_k\left[\sum_{i=0}^{1} m_{i,k} - \sum_{i=0}^{1} f_{i,k} + \sum_{i=0}^{1} m_{i,k-1} - \sum_{i=0}^{1} f_{i,k-1}\right]\right\} \\
\overset{(6)}{=} \quad & -\frac{\beta}{2}(c_0 + c_1) + \mathbb{E}_{k-1}\left\{\sum_{i=0}^{1} m_{i,k-1}\right\} - \sum_{i=0}^{1} f_{i,k-1} \\
\overset{\text{Figure 5 and 6}}{=} \quad & -\frac{\beta}{2}(c_0 + c_1) + \frac{1}{2}(\beta + 2\beta^2 + \beta^3)(c_0 + c_1) - (\beta^2 + \beta^3)(c_0 + c_1) \\
= \quad & \frac{\beta^3}{2}(c_0 + c_1)
\end{aligned}
$$

We observe that *only the highest order term remains in the calculation.* Repeat this process until $k = 1$, we will get the results in Lemma A.1. The above analysis also holds for general $n \geq 2$. For more evidence, please refer to Appendix G.1 where we further provide a derivation for $n = 3$.

□

The gap between Lemma A.1 and Equation (5)**.** Lemma A.1 shows the key idea of proving equation (5). However, due to its idealized setting, the color-ball toy example is still far from our real goal (5). We list some of the gaps here.

(i) In each possible trajectory: $x_{k,i}$ is changing with k and i, while the balls are fixed in the color-ball example.

(ii) When taking expectation: x is changing in different trajectory, while the balls is fixed in the color-ball example.

(iii) x is a vector in $\mathbb{R}^d$ while the label of the balls are constant in $\mathbb{R}$.

It requires extra technical lemmas to handle these differences. Briefly speaking, the error caused by (i) and (ii) can be neutralized by diminishing stepsize. Issue (iii) can be handled by component-wise analysis of x. For more details, please refer to the complete proof in Appendix G.

B Some Potential Implications for Practical Users

Theorem 3.1 and Proposition 3.3 establish a clearer image on the relation between (β_1, β_2) and qualitative behavior of Adam, which may have certain implication for practical users. Many practitioners

are still doing grid or random search over (β_1, β_2), which could be costly. The following advice may cut down a large portion of search space. Suppose we start with some random point (β_1^*, β_2^*), we provide the following suggestions for tuning β_1 and β_2.

Case 1: If (β_1^*, β_2^*) lies above the blue curve in Figure 2. We point out two sub-cases. First, if Adam with (β_1^*, β_2^*) diverges, then any points with "$\beta_1 < \sqrt{\beta_2}$" and "$\beta_2 < \beta_2^*$" shall not be tried due to the risk of divergence. Second, if Adam with (β_1^*, β_2^*) converges, then any points above this point will converge. You may either fix $\beta_1 = \beta_1^*$ and increase β_2, or fix $\beta_2 = \beta_2^*$ and try smaller β_1. Both ways have convergence guarantee.

Case 2: If (β_1^*, β_2^*) lies below the blue curve in Figure 2. We still discuss two sub-cases. First, if you observe divergence at (β_1^*, β_2^*), then do not further try any point on the left. We do not suggest exploring the points in below either, since the majority of them will still face the risk of divergence. Instead, we suggest fix $\beta_1 = \beta_1^*$ and increase β_2. Since our Theorem 3.1 applies for any $\beta_1 < \sqrt{\beta_2}$, algorithm will converge when β_2 is large enough.

Second, if you observe convergence at (β_1^*, β_2^*) (which is also possible according to Figure 1 (a)), then we suggest: (i) either fix $\beta_1 = \beta_1^*$ and increase β_2, or (ii) fix $\beta_2 = \beta_2^*$ and try smaller β_1. Both ways push (β_1, β_2) into the region of Theorem 3.1 with convergence guarantee.

C More Discussion on Related Works.

C.1 More Related Works on the Convergence Analysis of Adam Family

New variants of Adam. Ever since Reddi et al. (2018) pointed out the non-convergence issue of Adam, one active line of work has tried to design new variants of Adam that can be proved to converge. For instance, Zou et al. (2019); Gadat & Gavra (2020); Chen et al. (2018, 2021) replace the constant hyperparameters by iterate-dependent ones e.g. β_{1t} or β_{2t}. AMSGrad (Reddi et al., 2018) and AdaFom (Chen et al., 2018) modify $\{v_t\}$ to be an non-decreasing sequence. Similarly, AdaBound (Luo et al., 2019) impose lower and upper bounds on $\{v_t\}$ to prevent the effective stepsize from vanishing or exploding. Zhou et al. (2018) also adopt a new estimate of v_t to correct the bias. There are also attempts to combine Adam with Nesterov momentum (Dozat, 2016) as well as warm-up techniques (Liu et al., 2020a). As reported in the literature, these variants often reach good performance in many applications.

Some more discussions on (Guo et al., 2021) and (Huang et al., 2021). Here, we discuss more on two recent works Guo et al. (2021) and Huang et al. (2021). As mentioned in Section 2.2, they both requires some extra conditions. First, both Guo et al. (2021) and Huang et al. (2021) requires bounded gradient assumption. This can be seen in Assumption 2 in (Guo et al., 2021). In (Huang et al., 2021), they require bounded iterates (their Theorem 1) or change Adam into AdaBound (Luo et al., 2019) by clipping (their Remark 2, Corollary 1). Both settings inherent boundedness on gradient.

Besides bounded gradient, both (Huang et al., 2021) and (Guo et al., 2021) requires $1/(\sqrt{v_t}+\epsilon) \leq C_u$. This condition is stated in Assumption 2 in (Guo et al., 2021) and Assumption 3 in (Huang et al., 2021) (they presented it as $H_t \succeq \rho I_d \succ 0$, where matrix $H_t = \text{diag}(\sqrt{v_t} + \epsilon)$). Combining these two conditions, the effective stepsize of Adam will be bounded in certain interval $\frac{1}{\sqrt{v_t}+\epsilon} \in [C_l, C_u]$. Such boundedness condition changes Adam into AdaBound (Luo et al., 2019) and thus cannot explain the observations of Adam in Section 1.

C.2 An Brief Introduction to (Reddi et al., 2018)

Here, we re-state two counter examples by (Reddi et al., 2018). For the consistence of notation, we will re-state their results under our notation in the full script. They consider the convex problem ((Reddi et al., 2018)): $\min \sum_{i=0}^{n-1} f_i(x)$ where $x \in [-1, 1]$, $n \geq 3$:

$$f_i(x) = \begin{cases} nx, & \text{for } i = 0 \\ -x, & \text{otherwise,} \end{cases} \tag{7}$$

Since it is a constrained problem with feasible set $x \in [-1, 1]$, the optimal solution is $x^* = -1$. In (Reddi et al., 2018), they use notation C rather than n.

They consider sampling f_i in the cyclic order: f_0, f_1, f_2. In (Reddi et al., 2018), Function (7) is presented as an online optimization problem. We choose to use the form of (7) since it is more consistent with our notation in Algorithm 1. We re-state their results as follows.

Theorem C.1 (Theorem 2 in (Reddi et al., 2018))**.** *For any fixed* (β_1, β_2) *s.t.* $\beta_1 < \sqrt{\beta_2}$*, there exists function* (7) *with large enough* n*, s.t. Adam will converge to highly sub-optimal solution* $x = 1$.

We briefly re-state the non-convergent condition for this Theorem. As stated in Equation (7), Appendix B in (Reddi et al., 2018), for every fixed (β_1, β_2), they need a "constant n that depends on β_1 and β_2". As such, they requires different n to cause non-convergence on different (β_1, β_2). So the considered function class is constantly changing.

For completeness, we further re-state Theorem 1 in (Reddi et al., 2018).

Theorem C.2 (Theorem 1 in (Reddi et al., 2018))**.** *For function* (7)*, when* $\beta_1 = 0$ *and* $\beta_2 = 1/(n^2+1)$*, Adam will converge to highly sub-optimal solution* $x = 1$.

This theorem considers choosing (β_1, β_2) after n. However, this result only show divergence on *a single point* $(\beta_1, \beta_2) = (0, 1/(n^2 + 1))$. This point lies somewhere on the left boundary of Figure 3. It seems unclear how Adam's behavior would change as we change the (β_1, β_2) to anywhere else.

D Experimental Settings and More Experiments

D.1 Experimental Settings

Here, we introduce our experimental settings that are omitted in the full script.

- **Experiments on function** (2)**.** We use Algorithm 1 with cyclic order f_0, f_1, f_2 and so on. We report the optimality gap $x - x^*$ after 50k iteration, or equivalently $50000/n$ epochs. We use $\epsilon = 10^{-8}$ for numerical stability. We use diminishing stepsize $\eta_k = 0.1/\sqrt{k}$, where k is the index of epoch. Unless otherwise stated, this setting applies to all the other experiments on function (2).

  In Figure 1 (a), we use $n = 10$ and initialization $x = 1$. We will report more results with different n and different initialization in Appendix D.2.
- **MNIST (Deng, 2012).** We use one-hidden-layer neural network with width =16. We set batchsize =1, weight decay =0, stepsize =0.0001 and train for 20 epochs. We use $\epsilon = 10^{-8}$ for numerical stability.
- **CIFAR-10 (Krizhevsky et al., 2009).** We use ResNet-18 (He et al., 2016) as the architecture. We choose batchsize =16, weight decay =5e-4 , initial stepsize=1e-3. We use a stage-wise constant learning rate scheduling with a multiplicative factor of 0.1 on epoch 30, 60 and 90. We use $\epsilon = 10^{-8}$ for numerical stability.

  For MNIST and CIFAR-10, larger batchsize will bring similar pattern as that in Figure 1, but the phase transition will occur at some smaller β_2.
- **NLP.** The WikiText-103 dataset is a collection of over 100 million tokens extracted from the set of verified 'Good' and 'Featured' articles on Wikipedia. The base model of Transformer XL contains 16 self-attention layers. In each self-attention layer, there are 10 heads. The encoding dimension of each head is set to be 41. We set batchsize = 60, number of iteration = 200k, and initial stepsize = 0.00025. We use cosine learning rate scheduler, which is a popular choice for training Transformer. We use $\epsilon = 10^{-8}$ for numerical stability.

D.2 More experiments

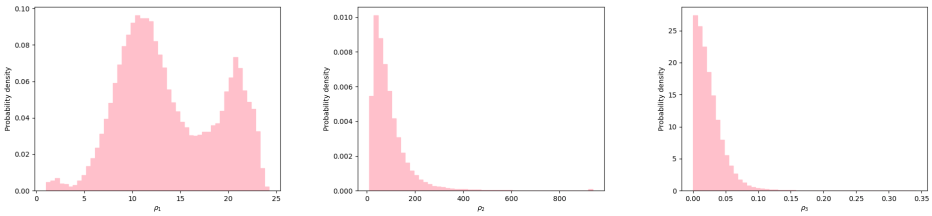
Estimation on ρ in Theorem 3.1. To ensure convergence, Theorem 3.1 requires $\beta_2 \geq \gamma_1(n) = 1 - \mathcal{O}((1 - \beta_1^n)/n^2\rho)$. Now we estimate the constant ρ. According to our definition in Remark G.8 in Appendix G.1, $\rho = \rho_1\rho_2\rho_3$, where ρ_1, ρ_2, ρ_3 as defined as follows.

$$\rho_1 \geq \frac{\sum_{i=1}^n |\partial_l f_i(x_{k,0})|}{\sqrt{\sum_{i=1}^n |\partial_l f_i(x_{k,0})|^2}};$$

$$\rho_2 \geq \frac{|\max_i \partial_l f_i(x_{k,0})|^2}{\frac{1}{n}\sum_{i=1}^n |\partial_l f_i(x_{k,0})|^2};$$

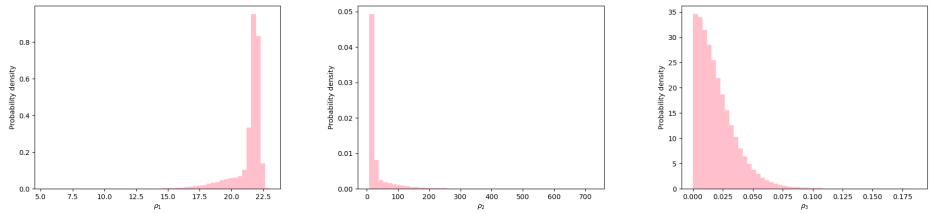
$$\rho_3 \geq \frac{|\sum_{i=1}^n \partial_l f_i(x_{k,0})|}{\sqrt{\frac{1}{n}\sum_{i=1}^n |\partial_l f_i(x_{k,0})|^2}}.$$

These definitions can also be seen in Appendix F.1. In worst case, we have $0 \leq \rho_3 \leq \sqrt{n}\rho_1 \leq n$. However, ρ is highly dependent on the problem instance $f(x)$ and training process. We now estimate how ρ changes with Adam's trajectory on MNIST and CIFAR-10. We use $\beta_1 = 0.9$, $\beta_2 = 0.99$. On both datasets, we set batchsize to be 64, which brings $n = 937$ on MNIST and $n = 781$ on CIFAR-10. We collect ρ_1, ρ_2, ρ_3 along the training process and estimate their distribution density. The results are shown in Figure 7 and 8.



(a) Histogram of ρ_1 (b) Histogram of ρ_2 (c) Histogram of ρ_3

Figure 7: Histograms of ρ_1, ρ_2, ρ_3 along the training process on MNIST.



(a) Histogram of ρ_1 (b) Histogram of ρ_2 (c) Histogram of ρ_3

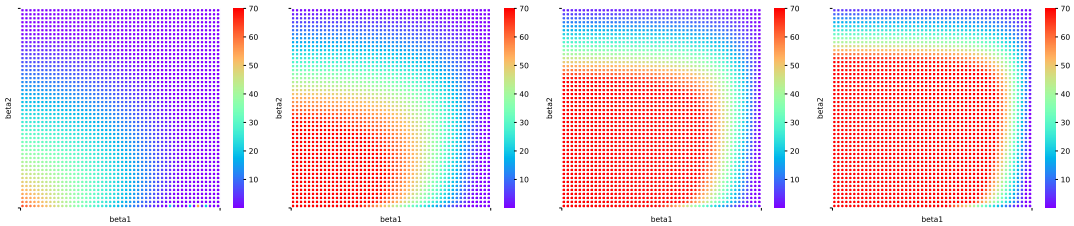
Figure 8: Histograms of ρ_1, ρ_2, ρ_3 along the training process on CIFAR-10.

On both CIFAR-10 and MNIST, we observe that the maximal $\rho_1 < 25 \approx \mathcal{O}(\sqrt{n})$, $\rho_2 < 400 \approx \mathcal{O}(n)$, $\rho_3 < 0.1 \approx \mathcal{O}(1/\sqrt{n})$. Therefore, $\rho = \rho_1\rho_2\rho_3 \approx \mathcal{O}(n)$.

More experiments on function (2)**.** Figure 1 (a) shows the optimality gap after 50k iterations when $n = 10$ and initialization $x = 1$. Here, we provide more relevant experiments. First, when initialized at $x = 1$, we run experiments with $n = 5, 10, 15, 20$. The results are shown in Figure 9. We observe that the blue region shrinks as n increases. This matches our conclusion in Theorem 3.1: when n increases, the convergent threshold of β_2 increases, which means we need larger β_2 to ensure convergence. It also matches the conclusion in Theorem 3.3: when n increases, the divergence region will expand (more evidence can be seen in Appendix E).

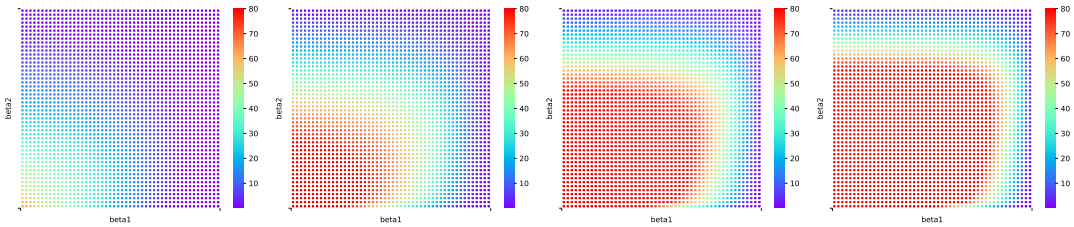
When initializing at $x = -5$, we further demonstrate that the gradient norm of $f(x)$ can dramatically increase. All the setting is the same as that in Figure 9 except for the change of initialization. Starting at $x = -5$, the algorithm can touch the "quadratic" side of function (2) where the gradient is unbounded. The results are shown in Figure 10. We observe an similar pattern as that in Figure 9. As a result, the gradient norm of $f(x)$ is large in the left bottom corner.

We further plot the change of gradient norm along the iterations. We pick $\beta_1 = 0.1$ and $\beta_2 = 0, 0.1, 0.3, 0.5, 0.7, 0.9, 0.99$ to see the phase transition when increasing β_2. The result is shown in Figure 11. When β_2 is small, the gradient norm of $f(x)$ increases rapidly along the iteration. Most of them are even much larger than the upper bound of color bar in Figure 10. As a result, there is a phase transition from diverge to converge when increasing β_2 from 0 to 1.



(a) $n = 5$ (b) $n = 10$ (c) $n = 15$ (d) $n = 20$

Figure 9: The optimality gap $x - x^*$ after running 50k iterations of Adam on function (2). We use initialization $x = 1$.



(a) $n = 5$ (b) $n = 10$ (c) $n = 15$ (d) $n = 20$

Figure 10: The gradient norm of $f(x)$ after running 50k iterations of Adam on function (2). We use initialization $x = -5$.

Adam converges to a exact stationary point under SGC where $D_0 = 0$. Since function (2) satisfies SGC, we further provide empirical evidence that Adam converges to a exact stationary point under SGC. We use function (2) $n = 20$ and initialization $x = -5$. We choose some large enough β_2 to ensure the convergence. As shown in Figure 12, We observe 0 gradient norm after Adam converges. All the hyperparameter setting is the same as before.

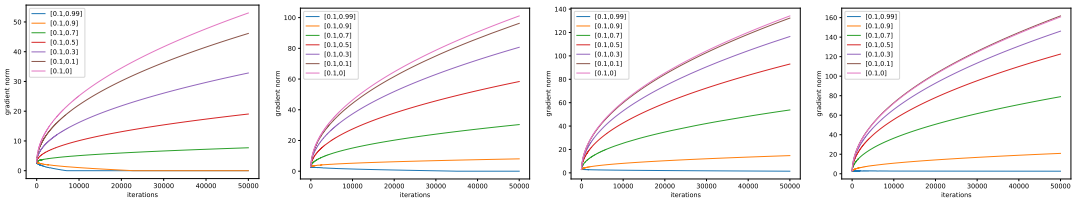
Adam converges to a bounded region when $D_0 > 0$. Now we show that Adam may not converge to an exact stationary point when $D_0 > 0$. Instead, it converge to a bounded region near the stationary point. For this part, we re-state the example from (Shi et al., 2020). Consider the following function.

$$f_j(x) = \begin{cases} (x-a)^2 \text{ if } j = 0 \\ -0.1\left(x - \frac{10}{9}a\right)^2 \text{ if } 1 \leq j \leq 9 \end{cases}$$

Summing up $f_j(x)$ we get

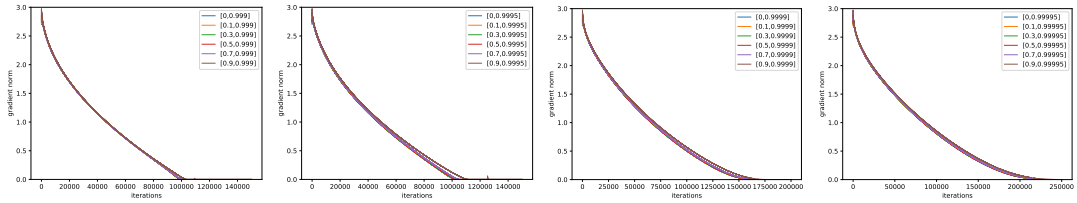
$$f(x) = \sum_{j=0}^{9} f_j(x) = \frac{1}{10}x^2 - \frac{1}{9}a^2.$$

This $f(x)$ is lower bounded by $\frac{-a^2}{9}$ with optimal solution $x^* = 0$. In (Shi et al., 2020), they try SGD, RMSProp and AMSGrad on this function, the result is shown in Figure 13. They observe that



(a) $n = 5$ (b) $n = 10$ (c) $n = 15$ (d) $n = 20$

Figure 11: The change of gradient norm along the iterations of Adam on function (2). We use initialization $x = 1$. We use $[\beta_1, \beta_2]$ to label the curves trained with corresponding hyperparameters.



(a) $\beta_2 = 0.999$ (b) $\beta_2 = 0.9995$ (c) $\beta_2 = 0.9999$ (d) $\beta_2 = 0.99995$

Figure 12: Under SGC, we observe 0 gradient norm after Adam converges. We use function (2) with $n = 20$ and initialization $x = -5$. We use $[\beta_1, \beta_2]$ to label the curves trained with corresponding hyperparameters.

RMSProp with diminishing stepsize may not converge to the optimal solution when $D_0 > 0$, even if they set the initialization at $x = x^* = 0$. See more description in (Shi et al., 2020, Appendix A.4)

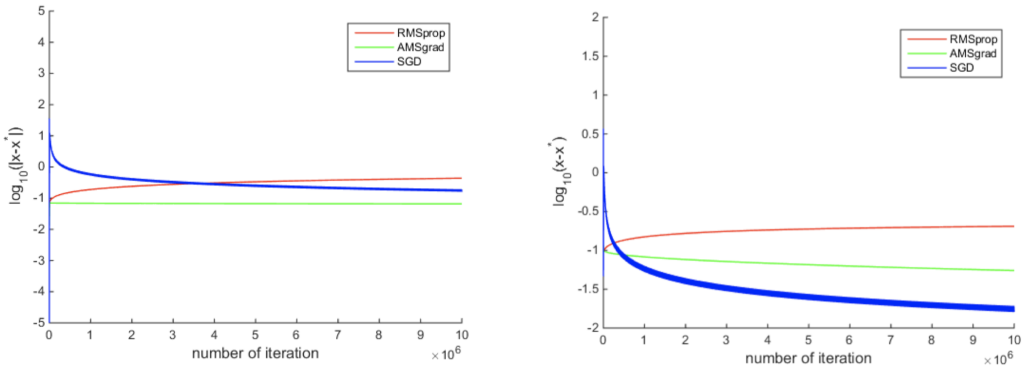


Figure 13: Figure A.4 in (Shi et al., 2020): RMSProp with diminishing stepsize does not converge to the optimal solution when $D_0 > 0$. Left figure uses initialization at optimal $x = 0$. Right figure uses initialization $x = -0.1$.

E Proof of Proposition 3.3

We restate our counter-example here. Consider $f(x) = \sum_{i=0}^{n-1} f_i(x)$ for $x \in \mathbb{R}$, we define $f_i(x)$ as follows:

$$
\begin{aligned}
f_i(x) &= \begin{cases} nx, & x \geq -1 \\ \frac{n}{2}(x+2)^2 - \frac{3n}{2}, & x < -1 \end{cases} \quad \text{for } i = 0, \\
f_i(x) &= \begin{cases} -x, & x \geq -1 \\ -\frac{1}{2}(x+2)^2 - \frac{3}{2}, & x < -1 \end{cases} \quad \text{for } i > 0.
\end{aligned} \tag{8}
$$

Summing up all the $f_i(x)$, we can see that

$$
f(x) = \begin{cases} x, & x \geq -1 \\ \frac{1}{2}(x+2)^2 - \frac{3}{2}, & x < -1 \end{cases}
$$

is a convex smooth function with optimal solution $x^* = -2$ and optimal value $f(x^*) = -3/2$.

However, we are going to show that, for any fixed $n > 2$, there exists an orange region shown in Figure 2, s.t., Adam with any β_1-β_2 combination in the yellow region converges to $x = \infty$ rather than the optimal solution $x = -2$, causing the divergence. Now we introduce the formal statement of Proposition 3.3.

Proposition E.1. (Formal statement of Proposition 3.3.) *Consider the convex function* (8) *for a fixed n. Starting at the initialization $x = 1$ and initial stepsize η_1, the iterates of Adam diverge to infinity if the following holds:*

$$(\mathbf{C1}): \left(n-1-\min\left\{n-1, \log_{\beta_2}(\frac{1}{10n^2})\right\}\right) \frac{1-\beta_1^{\min\left\{n-1, \log_{\beta_2}(\frac{1}{10n^2})\right\}}}{\sqrt{1+\max\left\{0.1, \beta_2^{n-1}n^2\right\}}} \geq \frac{1-\beta_1}{\sqrt{1-\beta_2}} + \frac{\beta_1}{\sqrt{1-\beta_2}}n; \tag{9}$$

$$(\mathbf{C2}): (1-\beta_1^{n-1}) > (1-\beta_1)\beta_1^{n-1}n. \tag{10}$$

$$(\mathbf{C3}): \eta_1 \leq 2\sqrt{(1-\beta_2)\beta_2^n}. \tag{11}$$

The analysis is motivated by that in (Reddi et al., 2018, Theorem 1). However, (Reddi et al., 2018, Theorem 1) considers a simplified case with $\beta_1 = 0$. Here, we consider non-zero β_1, especially for those $\beta_1 > \sqrt{\beta_2}$. To show the divergence, we aim to prove the following claim: (we denote $x_{k,i}$ as the value of x at the k-th outer loop and i-th inner loop)

Claim: for any fixed $n > 2$, there exists an orange region shown in Figure 2 s.t., Adam with any β_1-β_2 combination in the orange region gives $x_{k+1,0} > 1$ as long as $x_{k,0} = 1$.

Since the gradient stays constant when $x > 1$, so x will go to infinity if the claim holds. To prove this claim, we only need to analyze the trajectory of Adam within one particular outer loop, e.g., the k-th outer loop. We will show that $x_{k+1,0} > 1$ if this outer loop is initialized with $x_{k,0} = 1$. Similarly as (Reddi et al., 2018), we assume $f_i(x)$ are sampled in the order of $f_0(x), f_1(x), \cdots, f_{n-1}(x)$ within the k-th outer loop.

Now let us prove the claim. For function (8), the update rule of Adam is shown as follows.

$$x_{k,1} = (x_{k,0} + \delta_{k,0}), \quad \delta_{k,0} = -\frac{\eta_1}{\sqrt{k}}\left(\frac{n(1-\beta_1) + \beta_1 m_{k-1,n-1}}{\sqrt{(1-\beta_2)n^2 + \beta_2 v_{k-1,n-1}}}\right) \tag{12}$$

$$x_{k,i+1} = (x_{k,i} + \delta_{k,i}), \quad i = 1, \cdots, n-1; \tag{13}$$

where $\delta_{k,i} = -\frac{\eta_1}{\sqrt{k}}\left(\frac{(1-\beta_1)\sum_{j=0}^{i-1}(-1)\beta_1^j + (1-\beta_1)\beta_1^i n + \beta_1^{i+1} m_{k-1,n-1}}{\sqrt{(1-\beta_2)+\beta_2 v_{k,i-1}}}\right)$.

We decompose the total movement $\sum_{i=0}^{n-1}\delta_{k,i}$ into three terms as follows.

$$\begin{aligned}
\sum_{i=0}^{n-1}\delta_{k,i} \quad = \quad & \frac{\eta_1}{\sqrt{k}}\underbrace{\left(-\frac{\beta_1 m_{k-1,n-1}}{\sqrt{(1-\beta_2)n^2+\beta_2 v_{k-1,n-1}}} - \frac{\beta_1^2 m_{k-1,n-1}}{\sqrt{(1-\beta_2)+\beta_2 v_{k,0}}} - \cdots - \frac{\beta_1^n m_{k-1,n-1}}{\sqrt{(1-\beta_2)+\beta_2 v_{k,n-2}}}\right)}_{(I)} \\
& + \frac{\eta_1}{\sqrt{k}}\underbrace{\left(\frac{1-\beta_1}{\sqrt{(1-\beta_2)+\beta_2 v_{k,0}}} + \frac{(1-\beta_1)+\beta_1(1-\beta_1)}{\sqrt{(1-\beta_2)+\beta_2 v_{k,1}}} + \cdots + \frac{(1-\beta_1)\sum_{j=0}^{n-2}\beta_1^j}{\sqrt{(1-\beta_2)+\beta_2 v_{k,n-2}}}\right)}_{(II)} \\
& + \frac{\eta_1}{\sqrt{k}}\underbrace{\left(-\frac{n(1-\beta_2)}{\sqrt{(1-\beta_2)n^2+\beta_2 v_{k-1,n-1}}} - \frac{n(1-\beta_2)\beta_2}{\sqrt{(1-\beta_2)+\beta_2 v_{k,0}}} - \cdots - \frac{n(1-\beta_2)\beta_2^{n-1}}{\sqrt{(1-\beta_2)+\beta_2 v_{k,n-2}}}\right)}_{(III)}.
\end{aligned}$$

We will show that for some β_1 and β_2: (I), $(II) > 0$ and $(III) < 0$. However, (I) and (II) outweigh (III), causing the divergence.

First, we show that $m_{k-1,n-1} < 0$ when β_1 is small.

$$
\begin{aligned}
-m_{k-1,n-1} &= (1-\beta_1)\sum_{j=1}^{n-2}\beta_1^j - (1-\beta_1)\beta_1^{n-1}n - \beta_1^n m_{k-2,n-1} \\
&= (1-\beta_1^{n-1}) - (1-\beta_1)\beta_1^{n-1}n - \beta_1^n m_{k-2,n-1} \\
&= \left[(1-\beta_1^{n-1}) - (1-\beta_1)\beta_1^{n-1}n\right]\sum_{j=0}^{k}(\beta_1^n)^j .
\end{aligned}
$$

when β_1 is small,we have $(1-\beta_1^{n-1}) > (1-\beta_1)\beta_1^{n-1}n$, which implies $-m_{k-1,n-1} > 0$. For these choices of β_1, we have $(I) > 0$. Now we derive a lower bound for (II).

$$
\begin{aligned}
(II) &\geq \frac{1-\beta_1}{\sqrt{1+\beta_2 n^2}} + \frac{(1-\beta_1)+\beta_1(1-\beta_1)}{\sqrt{1+\beta_2^2 n^2}} + \cdots + \frac{(1-\beta_1)\sum_{j=0}^{n-2}\beta_1^j}{\sqrt{1+\beta_2^{n-1}n^2}} \\
&= \frac{1-\beta_1}{\sqrt{1+\beta_2 n^2}} + \frac{1-\beta_1^2}{\sqrt{1+\beta_2^2 n^2}} + \cdots + \frac{1-\beta_1^{n-1}}{\sqrt{1+\beta_2^{n-1}n^2}}.
\end{aligned}
$$

The inequality is due to the fact that $v_{k,0} \leq n^2$. Since $\beta_2^j n^2$ is small when β_2 is small and j is close to n, there exists some small β_2 such that $\beta_2^j n^2 \leq 0.1$ for at least one $j < n$. For these small enough β_2, we keep the summand with $j \geq \log_{\beta_2}(0.1/n^2)$ and drop the rest. We have the following lower bound for (II).

$$
(II) \geq \left(n-1-\log_{\beta_2}(\frac{1}{10n^2})\right)\frac{1-\beta_1^{\log_{\beta_2}(\frac{1}{10n^2})}}{\sqrt{1+0.1}}
$$

However, these lower bound only holds for the small β_2. With simple modification, we derive an universal lower bound of (II) for any $\beta_2 \in (0,1)$.

$$
(II) \geq \left(n-1-\min\left\{n-1, \log_{\beta_2}(\frac{1}{10n^2})\right\}\right)\frac{1-\beta_1^{\min\left\{n-1,\log_{\beta_2}(\frac{1}{10n^2})\right\}}}{\sqrt{1+\max\left\{0.1, \beta_2^{n-1}n^2\right\}}}
$$

Now we derive a lower bound for (III).

$$
\begin{aligned}
(III) &= -\frac{n(1-\beta_2)}{\sqrt{(1-\beta_2)n^2+\beta_2 v_{k-1,n-1}}} - \frac{n(1-\beta_2)\beta_2}{\sqrt{(1-\beta_2)+\beta_2 v_{k,0}}} - \cdots - \frac{n(1-\beta_2)\beta_2^{n-1}}{\sqrt{(1-\beta_2)+\beta_2 v_{k,n-2}}} \\
&\geq -\frac{1-\beta_1}{\sqrt{1-\beta_2}}\left(1+n(\sum_{j=1}^{n-1}\beta_1^j)\right) \\
&= -\frac{1-\beta_1}{\sqrt{1-\beta_2}} - \frac{\beta_1(1-\beta_1^{n-1})}{\sqrt{1-\beta_2}}n \\
&\geq -\frac{1-\beta_1}{\sqrt{1-\beta_2}} - \frac{\beta_1}{\sqrt{1-\beta_2}}n.
\end{aligned}
$$

The remaining step is to show for small enough step size η_1, the iterates will stay in the linear region, thus the above gradient holds for all iterates in the trajectory.

As we initial x as $x_0 = 1$, if for all $m < n$ and $k > 0$, $\sum_{i=0}^{m}\delta_{i,k} \geq -2$, we can conclude all iterates in the trajectory stay in the linear region.

Because it holds that $m_{k-1,n-1} \leq 0$ and $v_{k-1,n-1} \geq 0$, we have

$$\delta_{i,k} \geq -\eta_1 \frac{n(1-\beta_1)\beta_1^i}{\sqrt{n^2(1-\beta_2)\beta_2^i}} \geq -\eta_1 \frac{(1-\beta_1)\beta_1^i}{\sqrt{(1-\beta_2)\beta_2^n}}.$$

Therefore, to make $\sum_{i=0}^{m} \delta_{i,k} \geq \sum_{i=0}^{m} -\eta_1 \frac{(1-\beta_1)\beta_1^i}{\sqrt{(1-\beta_2)\beta_2^n}} \geq -2$, we have

$$\eta_1 \leq 2\sqrt{(1-\beta_2)\beta_2^n}.$$

To show the divergence, we want to show that there exists some β_1 and β_2 s.t. the both of the following conditions hold:

$$(\mathbf{C1}) : \left(n-1-\min\left\{n-1, \log_{\beta_2}(\frac{1}{10n^2})\right\}\right) \frac{1-\beta_1^{\min\left\{n-1, \log_{\beta_2}(\frac{1}{10n^2})\right\}}}{\sqrt{1+\max\left\{0.1, \beta_2^{n-1}n^2\right\}}} \geq \frac{1-\beta_1}{\sqrt{1-\beta_2}} + \frac{\beta_1}{\sqrt{1-\beta_2}}n;$$

$$(\mathbf{C2}) : \beta_1 \text{ is small s.t. } (1-\beta_1^{n-1}) > (1-\beta_1)\beta_1^{n-1}n.$$

$$(\mathbf{C3}) : \eta_1 \leq 2\sqrt{(1-\beta_2)\beta_2^n}.$$

The proof of Theorem 3.3 is completed. With the help of Python, we visualize the region where $(\mathbf{C1})$ and $(\mathbf{C2})$ hold. The results are shown in Figure 14. We use orange color to indicate the region where $(\mathbf{C1})$ holds. White color is used for the counter part. As for $(\mathbf{C2})$, we use the gray vertical line to indicate the line where $(1-\beta_1^{n-1}) = (1-\beta_1)\beta_1^{n-1}n$. Note that there are two solutions to this equation: one solution is $\beta_1 = 1$ and the other solution lies in $0 < \beta_1 < 1$, this is why there are two vertical lines in the figure. $(\mathbf{C2})$ holds on the left hand side of the left gray vertical line.

The intersection of two regions will be the region where Adam diverges, which is actually the orange region in Figure 14. As n increases, $(\mathbf{C2})$ holds for a wide range of β_1, so the grey vertical lines move towards $\beta_1 = 1$ and finally get overlapped. In addition, the size of divergence region increases with n

Relation with $\gamma_1(n)$ in Theorem 3.1 According to Theorem 3.1, $\gamma_1(n)$ is at least in the order of $1 - \mathcal{O}(n^{-2})$. Combining with Figure 14. It is not hard to see that $\gamma_1(n)$ is always larger than the upper boundary of the orange region, so there is no contradiction.

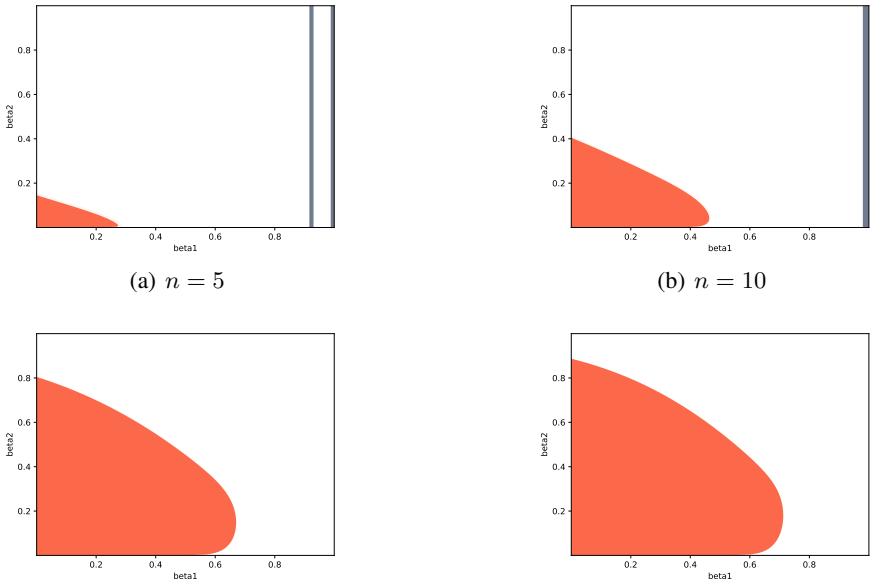
Proof of Corollary 4.1. For any $(\beta_1, \beta_2) \in [0,1)^2$, condition **(C1), (C2)** and **(C3)** can be satisfied by some sufficiently large n. Therefore, Adam will diverge and the proof is concluded.

F Some More Notations and Useful Lemmas for Convergence Analysis

F.1 More notations

- We denote $x_{k,i}, m_{k,i}, v_{k,i} \in \mathbb{R}^d$ as the value of x, m, v at the k-th outer loop and i-th inner loop. Further, we denote $x_{l,k,i}, m_{l,k,i}, v_{l,k,i} \in \mathbb{R}$ as the l-th component of $x_{k,i}, m_{k,i}, v_{k,i}$.
- We denote η_k as the stepsize at the k-th epoch (outer loop). We will focus mainly on diminishing step size, especially $\eta_k = \frac{\eta_1}{\sqrt{nk}}$, where n is the number of batches (inner loop).
- We denote $\partial_l f(x) = \frac{\partial}{\partial x_l} f(x)$, $\partial_l f_j(x) = \frac{\partial}{\partial x_l} f_j(x)$. Further, we will use $\tau_{k,i}$ to index the i-th randomly chosen batch in the k-th epoch. In this sense, we denote $\partial_l f_{\tau_{k,i}}(x)$ as $\frac{\partial}{\partial x_l} f_{\tau_{k,i}}(x)$.
- Given an epoch k, we denote α as the index of the coordinate with the greatest gradient:

$$\alpha = \arg\max_{l=1,2,\cdots,d} \left|\partial_l f\left(x_{k,0}\right)\right|.$$



(a) $n = 5$ (b) $n = 10$

(c) $n = 50$ (d) $n = 100$

Figure 14: This figure illustrates the region where both $(\mathbf{C1})$ and $(\mathbf{C2})$ in Proposition E.1 hold. The orange color indicates the region where $(\mathbf{C1})$ holds. White color is used for the counter part. The gray vertical lines are used to indicate the boundary of $(\mathbf{C2})$. Note that there are two solutions to the equation in $(\mathbf{C2})$: one solution is $\beta_1 = 1$ and the other solution lies in $0 < \beta_1 < 1$, this is why there are two vertical lines in the figure. $(\mathbf{C2})$ holds on the left hand side of the left gray vertical line. This figure is visualized in Python.

- We define ρ_1, ρ_2, ρ_3 as constants satisfying the following condition for any $l = 1, \cdots, d$:

$$\rho_1 \geq \frac{\sum_{i=1}^n |\partial_l f_i(x_{k,0})|}{\sqrt{\sum_{i=1}^n |\partial_l f_i(x_{k,0})|^2}};$$

$$\rho_2 \geq \frac{|\max_i \partial_l f_i(x_{k,0})|^2}{\frac{1}{n}\sum_{i=1}^n |\partial_l f_i(x_{k,0})|^2};$$

$$\rho_3 \geq \frac{|\sum_{i=1}^n \partial_l f_i(x_{k,0})|}{\sqrt{\frac{1}{n}\sum_{i=1}^n |\partial_l f_i(x_{k,0})|^2}}.$$

  ρ_1, ρ_2, ρ_3 are problem-dependent constants. In worst case, we have $0 \leq \rho_3 \leq \sqrt{n}\rho_1 \leq n$. ρ_3 is larger when $\partial_l f_i(x_{k,0})$ are more aligned.

- We define the constant $\triangle_x := \frac{\eta_0}{\sqrt{x}} \frac{L\sqrt{d}}{\sqrt{1-\beta_2}} \frac{1-\beta_1}{1-\frac{\beta_1}{\sqrt{\beta_2}}}$. This constant will be used repeatedly, especially in Lemma F.3.

- We define $Q_k := \triangle_1 \frac{n\sqrt{n}}{\sqrt{k}} \frac{32\sqrt{2}}{(1-\beta_2)^n \beta_2^n}$. This constant will be used repeatedly.

- We define

$$\delta_1 = (1-\beta_2)\frac{\left(-1 + \frac{4n\rho_2}{\beta_2^n}\right)}{2} + \left(\frac{1}{\sqrt{\beta_2^n}} - 1\right). \tag{14}$$

  This constant will be used repeatedly. Note that $\delta_1 \to 0$ when $\beta_2 \to 1$.

- $\mathbb{E}(\cdot)$ means taking expectation over the whole trajectory.

- We will use $\mathbb{E}_k[\cdot]$ as a shorthand for $\mathbb{E}[\cdot \mid x_{k,0}, x_{k-1,n-1}, \cdots, x_{1,0}]$, i.e. the conditional expectation given all the history up to $x_{k,0}$. Similarly, $\mathbb{E}_{k-1}$ stands for the conditional expectation given all the history up to $x_{k-1,0}$.

We further introduce some 'index' notation of the stochastic gradients, these notations will be useful in the proof. For all the stochastic gradient, we have

- **True index:** every stochastic gradient has its own true index: $f_1(\cdot), f_2(\cdot), \cdots, f_{n-1}(\cdot)$. All these $\{f_i(\cdot)\}_{i=0}^{n-1}$ are fixed once the optimization problem is formulated.
  Note that for a fixed x, we have the following relationship between f_i and f:

$$\sum_{i=0}^{n-1} \partial_l f_i(x) = \partial_l f(x). \tag{15}$$

- **Random-shuffle index:** At the k-th epoch, all the stochastic gradients are sampled in the order of $f_{\tau_{k,0}}(\cdot), f_{\tau_{k,1}}(\cdot), \cdots, f_{\tau_{k,n-1}}(\cdot)$.
  Since Algorithm 1 is sampling without replacement, there is an implicit bijective mapping between $\{f_i(\cdot)\}_{i=0}^{n-1}$ and $\{f_{\tau_{k,i}}(\cdot)\}_{i=0}^{n-1}$. Further, for a fixed x, we have the following useful property:

$$\sum_{i=0}^{n-1} \partial_l f_{\tau_{k,i}}(x) = \partial_l f(x). \tag{16}$$

- **Index for m and v:** As shown in Algorithm 1, we denote :

$$m_{l,k,i} = (1-\beta_1)\{\partial_l f_{\tau_{k,i}}(x_{k,i}) + \cdots + \beta_1^i \partial_l f_{\tau_{k,0}}(x_{k,0})\} + \beta_1^{i+1} m_{l,k,0},$$
$$v_{l,k,i} = (1-\beta_2)\{\partial_l f_{\tau_{k,i}}(x_{k,i})^2 + \cdots + \beta_1^i \partial_l f_{\tau_{k,0}}(x_{k,0})^2\} + \beta_1^{i+1} v_{l,k,0}.$$

F.2 Some useful lemmas

We begin with proving some useful lemmas.

Lemma F.1. *Let M and k be 2 integers with $k > M$, we have*

$$\sum_{p=1}^{M} \frac{1}{\sqrt{k-p}} \in \left[\frac{M}{\sqrt{k}}, 2\frac{M}{\sqrt{k-1}}\right]$$

Proof. For the lower bound,

$$\sum_{p=1}^{M} \frac{1}{\sqrt{k-p}} \geq \int_{k-M}^{k} \frac{dt}{\sqrt{t}} = \frac{2M}{\sqrt{k}+\sqrt{k-M}} \geq \frac{M}{\sqrt{k}}$$

For the upper bound,

$$\sum_{p=1}^{M} \frac{1}{\sqrt{k-p}} \leq \int_{k-M-1}^{k-1} \frac{dt}{\sqrt{t}} = \frac{2M}{\sqrt{k-M-1}+\sqrt{k-1}} \leq \frac{2M}{\sqrt{k-1}}$$

□

Lemma F.2. *For any $\beta \neq 1$, we have*

$$(1-\beta)\sum_{j=1}^{\infty} \beta^{j-1} = 1,$$
$$(1-\beta)\sum_{j=1}^{\infty} j\beta^{j-1} = \frac{1}{1-\beta},$$
$$(1-\beta)\sum_{j=1}^{\infty} j^2\beta^{j-1} = \frac{1+\beta}{(1-\beta)^2}.$$

Proof. Proof only involves basic calculation, we omit the proof here. □

Lemma F.3. *Under Assumption 2.1, if $\beta_1 < \sqrt{\beta_2}$, we have*

$$\left|\partial_l f_{\tau_{k,i+1}}(x_{k,i}) - \partial_l f_{\tau_{k,i}}(x_{k,i})\right| \leq \frac{\eta_0}{\sqrt{nk}} \frac{L\sqrt{d}}{\sqrt{1-\beta_2}} \frac{1-\beta_1}{1-\frac{\beta_1}{\sqrt{\beta_2}}} = \triangle_{nk}. \tag{17}$$

Proof. We start with bounding $|x_{l,k,i+1} - x_{l,k,i}|$. By the update formula of Adam, we have

$$\begin{aligned}
|x_{l,k,i+1} - x_{l,k,i}| &= \eta_k \frac{|m_{l,k,i}|}{\sqrt{v_{l,k,i}}} \\
&\leq \eta_k (1-\beta_1) \left\{ \beta_1^0 \frac{|\partial_l f_{\tau_{k,i}}(x_{k,i})|}{\sqrt{v_{l,k,i}}} + \beta_1 \frac{|\partial_l f_{\tau_{k,i-1}}(x_{k,i-1})|}{\sqrt{v_{l,k,i}}} + \dots \right\} \\
&\leq \eta_k (1-\beta_1) \left\{ \beta_1^0 \frac{|\partial_l f_{\tau_{k,i}}(x_{k,i})|}{\sqrt{(1-\beta_2)\beta_2^0}|\partial_l f_{\tau_{k,i}}(x_{k,i})|} + \beta_1 \frac{|\partial_l f_{\tau_{k,i-1}}(x_{k,i-1})|}{\sqrt{(1-\beta_2)\beta_2}|\partial_l f_{\tau_{k,i-1}}(x_{k,i-1})|} + \dots \right\} \\
&\leq \eta_k \frac{(1-\beta_1)}{\sqrt{1-\beta_2}} \left((\frac{\beta_1}{\sqrt{\beta_2}})^0 + (\frac{\beta_1}{\sqrt{\beta_2}})^1 + (\frac{\beta_1}{\sqrt{\beta_2}})^2 + \dots \right) \\
&= \eta_k \frac{(1-\beta_1)}{\sqrt{1-\beta_2}} \frac{1}{1-\frac{\beta_1}{\sqrt{\beta_2}}}.
\end{aligned} \tag{18}$$

Now, by the Lipschitz assumption of $\partial_l f_{\tau_{k,i}}(x_{k,i})$ (Assumption 2.1), we have

$$\begin{aligned}
\left|\partial_l f_{\tau_{k,i}}(x_{k,i+1}) - \partial_l f_{\tau_{k,i}}(x_{k,i})\right| &\leq \|\nabla f_{\tau_{k,i}}(x_{k,i+1}) - \nabla f_{\tau_{k,i}}(x_{k,i})\|_2 \\
&\leq L\|x_{k,i+1} - x_{k,i}\|_2 \\
&\leq L\sqrt{d} \max_l |x_{l,k,i+1} - x_{l,k,i}| \\
&\leq \eta_k \frac{L\sqrt{d}}{\sqrt{1-\beta_2}} \frac{1-\beta_1}{1-\frac{\beta_1}{\sqrt{\beta_2}}} \\
&= \frac{\eta_0}{\sqrt{nk}} \frac{L\sqrt{d}}{\sqrt{1-\beta_2}} \frac{1-\beta_1}{1-\frac{\beta_1}{\sqrt{\beta_2}}}
\end{aligned} \tag{19}$$

Proof is completed.

□

Lemma F.4. *Under Assumption 2.1 and 2.2, for any integer $i \in [0, n-1]$, we have the following two results,*

$$|\partial_l f_{\tau_{k,i}}(x_{k,i})| \leq i\triangle_{nk} + \sqrt{D_1}\rho_1 d \left(|\partial_\alpha f(x_{k,0})| + \sqrt{\frac{D_0}{D_1 d}} \right), \tag{20}$$

$$|\partial_l f_{\tau_{k,i}}(x_{k,0})| \leq \sum_{i=0}^{n-1} |\partial_l f_{\tau_{k,i}}(x_{k,0})| \leq \sum_{l=1}^{d} \sum_{i=0}^{n-1} |\partial_l f_{\tau_{k,i}}(x_{k,0})| \leq \sqrt{D_1}\rho_1 d \left(|\partial_\alpha f(x_{k,0})| + \sqrt{\frac{D_0}{D_1 d}} \right), \tag{21}$$

where $|\partial_\alpha f(x_{k,0})| = \max_{l\in[1,d]} |\partial_l f(x_{k,0})|$.

Proof.

$$
\begin{aligned}
|\partial_l f_{\tau_{k,j}}(x_{k,j})| &\overset{\text{Lemma } F.3}{\leq} j\triangle_{nk} + |\partial_l f_{\tau_{k,j}}(x_{k,0})| \\
&\leq j\triangle_{nk} + \sum_{l=1}^{d}\sum_{j=0}^{n-1} |\partial_l f_{\tau_{k,j}}(x_{k,0})| \\
&\overset{(*)}{\leq} j\triangle_{nk} + \rho_1 \sum_{l=1}^{d} \sqrt{\sum_{i=0}^{n-1} |\partial_l f_i(x_{k,0})|^2} \\
&\overset{\text{Cauchy-Swartz inequality}}{\leq} j\triangle_{nk} + \rho_1\sqrt{d}\sqrt{\sum_{l=1}^{d}\sum_{i=0}^{n-1} |\partial_l f_i(x_{k,0})|^2} \\
&\overset{\text{Assumption 2.2}}{\leq} j\triangle_{nk} + \rho_1\sqrt{d}\sqrt{D_1\|\nabla f(x_{k,0})\|_2^2 + D_0} \\
&\overset{(**)}{\leq} j\triangle_{nk} + \sqrt{D_1}\rho_1 d\sqrt{|\partial_\alpha f\left(x_{k,0}\right)|^2 + \frac{D_0}{D_1 d}} \\
&\overset{(***)}{\leq} j\triangle_{nk} + \sqrt{D_1}\rho_1 d\left(|\partial_\alpha f\left(x_{k,0}\right)| + \sqrt{\frac{D_0}{D_1 d}}\right)
\end{aligned}
$$

where $(*)$ is because of the definition (see Appendix F.1): ρ_1 is a constant satisfying $\rho_1 \geq \frac{\sum_{i=1}^n |\partial_l f_i(x_{k,0})|}{\sqrt{\sum_{i=1}^n |\partial_l f_i(x_{k,0})|^2}}$; $(**)$ is due to $\|\nabla f(x_{k,0})\|_2^2 \leq d\,|\partial_\alpha f\left(x_{k,0}\right)|^2$; $(**)$ is because $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$, for any $x, y \geq 0$. The proof of (21) follows the same procedure as above. □

G Proof of Theorem 3.1

G.1 A roadmap of the proof

Our proof is based on the Descent Lemma:

$$
\mathbb{E}f(x_{k+1,0}) - \mathbb{E}f(x_{k,0}) \quad \leq \quad \mathbb{E}\langle \nabla f(x_{k,0}), x_{k+1,0} - x_{k,0}\rangle + \frac{L}{2}\mathbb{E}\|x_{k+1,0} - x_{k,0}\|_2^2
$$

The expectation $\mathbb{E}(\cdot)$ is taken on the whole trajectory. Summing both sides from the initialization $k = t_0$ to $k = T$, we have the following re-arranged inequality: (usually we set $t_0 = 1$.)

$$
\mathbb{E}\sum_{k=t_0}^{T} \langle \nabla f(x_{k,0}), x_{k,0} - x_{k+1,0}\rangle \quad \leq \quad \frac{L}{2}\sum_{k=t_0}^{T} \mathbb{E}\|x_{k+1,0} - x_{k,0}\|_2^2 + \mathbb{E}f(x_{t_0,0}) - \mathbb{E}f(x_{T+1,0}). \quad (22)
$$

To prove the convergence, we need an upper bound for $\frac{L}{2}\mathbb{E}\|x_{k+1,0} - x_{k,0}\|_2^2$, as well as a lower bound for $\mathbb{E}\langle \nabla f(x_{k,0}), x_{k,0} - x_{k+1,0}\rangle$ (and such a lower bound should be in the order of $\frac{1}{\sqrt{k}}\mathbb{E}\|\nabla f(x_{k,0})\|$).

The upper bound for $\frac{L}{2}\mathbb{E}\|x_{k+1,0} - x_{k,0}\|_2^2$ is relatively easy to get: according to the update rule of Adam, this term is in the order of $\mathcal{O}(\eta_k^2 m_{l,k,i}/v_{l,k,i}) = \mathcal{O}(\eta_k^2\|\nabla f(x_{k,0})\|/\|\nabla f(x_{k,0})\|) = \mathcal{O}(\frac{1}{k})$. Further, recall $\sum_{k=t_0}^{T}\frac{1}{k} \leq \log\frac{T+1}{t_0}$, so $\frac{L}{2}\mathbb{E}\|x_{k+1,0} - x_{k,0}\|_2^2$ contributes to the $\log$ term in Theorem 3.1. The proof is shown in Lemma F.3.

However, the lower bound for $\mathbb{E}\langle \nabla f(x_{k,0}), x_{k+1,0} - x_{k,0}\rangle = \mathbb{E}\left[\sum_{l=1}^{d}\sum_{i=0}^{n-1}\partial_l f(x_{k,0})\frac{m_{l,k,i}}{\sqrt{v_{l,k,i}}}\right]$ requires sophisticated derivation, we explain as follows. Before taking the expectation, we first work on every possible realization of $\sum_{l=1}^{d}\sum_{i=0}^{n-1}\partial_l f(x_{k,0})\frac{m_{l,k,i}}{\sqrt{v_{l,k,i}}}$. We perform the following decomposition for every $l \in [1, d]$.

$$\begin{aligned}\sum_{i=0}^{n-1}\partial_l f(x_{k,0})\frac{m_{l,k,i}}{\sqrt{v_{l,k,i}}} &= \sum_{i=0}^{n-1}\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,i}}}\left(\partial_l f_i(x_{k,0})+m_{l,k,i}-\partial_l f_i(x_{k,0})\right)\\ &= \underbrace{\left[\sum_{i=0}^{n-1}\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,i}}}\partial_l f_i(x_{k,0})\right]}_{(a)}+\underbrace{\left[\partial_l f(x_{k,0})\sum_{i=0}^{n-1}\frac{1}{\sqrt{v_{l,k,i}}}(m_{l,k,i}-\partial_l f_i(x_{k,0}))\right]}_{(b)}.\end{aligned} \tag{23}$$

First of all, we introduce the following Lemma G.1 to further decompose (a) and (b). The intuition of this decomposition is as follows: by increasing β_2, we can control the movement of the moving average factor $v_{l,k,i}$, so the trajectory of Adam is easier to control.

Lemma G.1. *Under Assumption 2.1, for those l satisfying $\max_i|\partial_l f_i(x_{k,0})| \geq Q_k := \triangle_1\frac{n\sqrt{n}}{\sqrt{k}}\frac{32\sqrt{2}}{(1-\beta_2)^n\beta_2^n}$, we have the following lower bound for (a) in (23):*

$$\sum_{i=0}^{n-1}\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,i}}}\partial_l f_i(x_{k,0}) \geq \frac{\partial_l f\left(x_{k,0}\right)^2}{\sqrt{v_{l,k,0}}}-\delta_1\left|\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,0}}}\right|\sum_i\left|\partial_l f_i\left(x_{k,0}\right)\right|, \tag{24}$$

where $\delta_1 = (1-\beta_2)\frac{\left(-1+\frac{4n\rho_2}{\beta_2^n}\right)}{2}+\left(\frac{1}{\sqrt{\beta_2^n}}-1\right)$. ρ_2,ρ_3 are constants satisfying $\rho_2 \geq \frac{|\max_i\partial_l f_i(x_{k,0})|^2}{\frac{1}{n}\sum_{i=1}^n|\partial_l f_i(x_{k,0})|^2}$, $\rho_3 \geq \frac{\left|\sum_{i=1}^n\partial_l f_i(x_{k,0})\right|}{\sqrt{\frac{1}{n}\sum_{i=1}^n|\partial_l f_i(x_{k,0})|^2}}$. Note that $\delta_1\to 0$ when $\beta_2\to 1$.

Under the same condition, we also have a lower bound for (b) in (23):

$$\begin{aligned}&\sum_{i=0}^{n-1}\frac{\partial_l f\left(x_{k,0}\right)}{\sqrt{v_{l,k,i}}}(m_{l,k,i}-\partial_l f_i(x_{k,0}))\\ \geq\ &\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,0}}}\sum_{i=0}^{n-1}(m_{l,k,i}-\partial_l f_i(x_{k,0}))-\delta_1\left|\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,0}}}\right|\sum_{i=0}^{n-1}|m_{l,k,i}-\partial_l f_i(x_{k,0})|.\end{aligned} \tag{25}$$

The proof can be seen in Appendix G.2. Lemma G.1 decomposes both (a) and (b): in the denominator, we approximate $v_{l,k,i}$ (which is changing with i) by $v_{l,k,0}$ (which is fixed). Accordingly, the approximation error can be controlled by increasing β_2 (since $\delta_1\to 0$ when $\beta_2\to 1$). However, Lemma G.1 can only be applied to those l with $\max_i|\partial_l f_i(x_{k,0})|\geq Q_k$, so we need to discuss the following two cases to complete the whole picture.

Case 1: unbounded gradient. Given $x_{k,0}$, consider those l with $\max_i|\partial_l f_i(x_{k,0})|\geq Q_k$, we have the following decomposition:

$$
\begin{aligned}
\sum_{i=0}^{n-1} \partial_l f(x_{k,0}) \frac{m_{l,k,i}}{\sqrt{v_{l,k,i}}} \quad & \overset{(23)}{\geq} \quad \underbrace{\left[\sum_{i=0}^{n-1} \frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,i}}} \partial_l f_i(x_{k,0})\right]}_{(a)} + \underbrace{\left[\partial_l f(x_{k,0}) \sum_{i=0}^{n-1} \frac{1}{\sqrt{v_{l,k,i}}} (m_{l,k,i} - \partial_l f_i(x_{k,0}))\right]}_{(b)}. \\
& \overset{\text{Lemma G.1}}{\geq} \quad \frac{\partial_l f\left(x_{k,0}\right)^2}{\sqrt{v_{l,k,0}}} - \delta_1 \left|\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,0}}}\right| \sum_{i=0}^{n-1} \left|\partial_l f_i\left(x_{k,0}\right)\right| \\
& \quad + \frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,0}}} \sum_{i=0}^{n-1} (m_{l,k,i} - \partial_l f_i(x_{k,0})) - \delta_1 \left|\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,0}}}\right| \sum_{i=0}^{n-1} |m_{l,k,i} - \partial_l f_i(x_{k,0})| \\
& \overset{\text{Lemma G.12}}{\geq} \quad \underbrace{\left[\frac{\partial_l f\left(x_{k,0}\right)^2}{\sqrt{v_{l,k,0}}}\right]}_{(a_1)} - \underbrace{\left[\delta_1 \sqrt{\frac{2\rho_3^2}{\beta_2^n}} \sum_{i=0}^{n-1} \left|\partial_l f_i\left(x_{k,0}\right)\right|\right]}_{(a_2)} \\
& \quad + \underbrace{\left[\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,0}}} \sum_{i=0}^{n-1} (m_{l,k,i} - \partial_l f_i(x_{k,0}))\right]}_{(b_1)} - \underbrace{\left[\delta_1 \sqrt{\frac{2\rho_3^2}{\beta_2^n}} \sum_{i=0}^{n-1} |m_{l,k,i} - \partial_l f_i(x_{k,0})|\right]}_{(b_2)},
\end{aligned}
\tag{26}
$$

Case 2: bounded gradient. Given $x_{k,0}$, consider those l with $\max_i |\partial_l f_i(x_{k,0})| \leq Q_k$, the analysis degenerates to the "bounded gradient" scenario, we have the following lower bound:

$$
\begin{aligned}
\sum_{i=0}^{n-1} \partial_l f(x_{k,0}) \frac{m_{l,k,i}}{\sqrt{v_{l,k,i}}} \quad & \geq \quad -\sum_{i=0}^{n-1} |\partial_l f(x_{k,0})| \left|\frac{m_{l,k,i}}{\sqrt{v_{l,k,i}}}\right| \\
& \overset{(18)}{\geq} \quad -|\partial_l f(x_{k,0})| \frac{1-\beta_1}{1-\frac{\beta_1}{\sqrt{\beta_2}}} n \\
& \overset{(*)}{\geq} \quad -nQ_k \frac{1-\beta_1}{1-\frac{\beta_1}{\sqrt{\beta_2}}} n \\
& = \quad -\triangle_1 \frac{n^2\sqrt{n}}{\sqrt{k}} \frac{32\sqrt{2}}{(1-\beta_2)^n \beta_2^n} \frac{1-\beta_1}{1-\frac{\beta_1}{\sqrt{\beta_2}}} n \\
& := \quad -F_1 \frac{1}{\sqrt{k}},
\end{aligned}
$$

where $F_1 := \triangle_1 n^2 \sqrt{n} \frac{32\sqrt{2}}{(1-\beta_2)^n \beta_2^n} \frac{1-\beta_1}{1-\frac{\beta_1}{\sqrt{\beta_2}}} n$, $(*)$ is due to the fact that $|\partial_l f(x_{k,0})| \leq n \max_i |\partial_l f_i(x_{k,0})| \leq nQ_k$.

Combining **Case 1 & 2** together, we have the following result (note that $(a_1), (a_2), (b_1), (b_2)$ vary with l. A more precise notation will be $(a_1)_l$, etc. We drop the subscript for brevity).

$$
\begin{aligned}
\sum_{l=1}^{d}\sum_{i=0}^{n-1}\partial_l f(x_{k,0})\frac{m_{l,k,i}}{\sqrt{v_{l,k,i}}} &\geq \underbrace{\sum_{l \text{ large}}\{(a_1)-(a_2)+(b_1)-(b_2)\}}_{\text{unbounded gradient}} - \underbrace{\sum_{l \text{ small}} F_1\frac{1}{\sqrt{k}}}_{\text{bounded gradient}} \\
&\geq \underbrace{\sum_{l \text{ large}}\{(a_1)-(a_2)+(b_1)-(b_2)\}}_{\text{unbounded gradient}} - \underbrace{\sum_{l=1}^{d}\left\{F_1\frac{1}{\sqrt{k}}\right\}}_{\text{bounded gradient}} \\
&\geq \left\{\sum_{l \text{ large}}(a_1)\right\} + \left\{\sum_{l \text{ large}}(b_1)\right\} - \left\{\sum_{l \text{ large}}\{(a_2)+(b_2)\}\right\} - dF_1\frac{1}{\sqrt{k}},
\end{aligned} \tag{27}
$$

where "l large" stands for the gradient component in **Case 1** and "l small" indicates those in **Case 2**. We assume "l large" is not an empty set, otherwise the analysis degenerates to the proof with bounded gradient assumption.

Now we take expectation on (27). The expectation is taken on all the possible trajectories up to the i-th iteration in k-th epoch.

$$
\mathbb{E}\left[\sum_{l=1}^{d}\sum_{i=0}^{n-1}\partial_l f(x_{k,0})\frac{m_{l,k,i}}{\sqrt{v_{l,k,i}}}\right] \overset{(27)}{\geq} \mathbb{E}\left\{\sum_{l \text{ large}}(a_1) + \sum_{l \text{ large}}(b_1) - \sum_{l \text{ large}}\{(a_2)+(b_2)\}\right\} - dF_1\frac{1}{\sqrt{k}}, \tag{28}
$$

In the following context, we will discuss how to bound all terms in (28), respectively. First and foremost, we derive a lower bound for $\mathbb{E}\left[\sum_{l \textbf{ large}}(b_1)\right]$. Since $\mathbb{E}\left[\sum_{l \textbf{ large}}(b_1)\right]$ contains all the historical gradient information, it involves great effort to handle it. We will show that the lower bound of $\mathbb{E}\left[\sum_{l \textbf{ large}}(b_1)\right]$ will vanish when β_2 is large and k goes to infinity.

Recall $\mathbb{E}\left[\sum_{l \text{ large}}(b_1)\right] = \mathbb{E}\left[\sum_{l \text{ large}}\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,0}}}\sum_{i=0}^{n-1}(m_{l,k,i}-\partial_l f_i(x_{k,0}))\right]$. When β_1 is large, $m_{l,k,i}$ contains heavy historical signals. It seems unclear how large $\mathbb{E}\left[\sum_{l \text{ large}}(b_1)\right]$ would be when β_1 goes to 1. Existing literatures (Zaheer et al., 2018; De et al., 2018; Shi et al., 2020) take a naive approach: they set $\beta_1 \approx 0$ so that $m_{k,i} \approx \nabla f_{\tau_{k,i}}(x_{k,i})$. Then we get $\delta(\beta_1) \approx 0$. However, this naive method cannot be applied here since we are interested in practical cases where β_1 is large in $[0,1)$. We emphasize the following technical difficulties in bounding $\mathbb{E}\left[\sum_{l \text{ large}}(b_1)\right]$ for any $\beta_1 \in [0,1)$:

Issue (i) In order to bound $\mathbb{E}\left[\sum_{l \text{ large}}(b_1)\right]$, we need to first know how to bound its simpler version: $\mathbb{E}\left[\sum_{i=0}^{n-1}(m_{l,k,i}-\partial_l f_i(x_{k,0}))\right]$. This term measures the difference between the current gradient and weighted historical gradients. It seems unclear that how large this term could be when β_1 goes to 1.

Issue (ii) Even if we can bound $\mathbb{E}\left[\sum_{i=0}^{n-1}(m_{l,k,i}-\partial_l f_i(x_{k,0}))\right]$, it is still unclear how to bound $\mathbb{E}\left[\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,0}}}\sum_{i=0}^{n-1}(m_{l,k,i}-\partial_l f_i(x_{k,0}))\right]$, which is further multiplied by a random variable $\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,0}}}$.

Issue (iii) Even if we can bound $\mathbb{E}\left[\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,0}}}\sum_{i=0}^{n-1}(m_{l,k,i}-\partial_l f_i(x_{k,0}))\right]$, it is still different from $\mathbb{E}\left[\sum_{l \text{ large}}(b_1)\right]$ which contains additional operation "$\sum_{l \text{ large}}$" inside the expectation.

Note that the set "l large" is a random variable which changes along different trajectories, so there is still non-negligible gap between bounding $\mathbb{E}\left[\sum_{l \text{ large}}(b_1)\right]$ and $\mathbb{E}\left[\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,0}}}\sum_{i=0}^{n-1}\left(m_{l,k,i}-\partial_l f_i(x_{k,0})\right)\right]$.

To our best knowledge, there is no general approach to tackle the above issues. In the following content, we will overcome difficulties (i), (ii) and (iii) in Step 1, 2 and 3 respectively. In Step 1, we will discuss how to bound $\mathbb{E}\left[\sum_{i=0}^{n-1}\left(m_{l,k,i}-\partial_l f_i(x_{k,0})\right)\right]$, which is a simplified version of $\mathbb{E}\left[(b_1)\right]$. Bounding this term will shed light on bounding the whole term $\mathbb{E}\left[\sum_{l \text{ large}}(b_1)\right]$. Then in Step 2 and 3, we will prove several technical lemmas to handle the effect of $\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,0}}}$ and "$\sum_{l \text{ large}}$", by which we can tackle **(ii)** and **(iii)**. Combining all together we can bound $\mathbb{E}\left[\sum_{l \text{ large}}(b_1)\right]$.

Bounding $\mathbb{E}\left[\sum_{l \textbf{ large}}(b_1)\right]$: Step 1. We now introduce the key idea of bounding $\mathbb{E}\left[\sum_{i=0}^{n-1}\left(m_{l,k,i}-\partial_l f_i(x_{k,0})\right)\right]$. As discussed in Appendix A, we analyze $\mathbb{E}\left[\sum_{i=0}^{n-1}\left(m_{l,k,i}-\partial_l f_i(x_{k,0})\right)\right]$ for every epoch and find out some pattern across the successive epochs. In Appendix A, we distill our idea into a toy example with two balls in a box. Here, to better convey our message, we further describe a similar case with three balls.

The "color-ball " model with three balls. Consider a box containing 3 balls labeled $a_0, a_1, a_2 \in \mathbb{R}$ (could be negative). In each round (epoch), we randomly sample all the balls from the box without replacement, then we put them all back. We denote the 1st sampled label in the k-th epoch as $a_{\tau_{k,0}}$, and similarly for $a_{\tau_{k,1}}, a_{\tau_{k,2}}$. We define the following random variables: ($\beta \in [0,1]$)

$$m_2 = \underbrace{a_{\tau_{k,2}} + \beta a_{\tau_{k,1}} + \beta^2 a_{\tau_{k,0}}}_{k\text{-th epoch}} + \underbrace{\beta^3 a_{\tau_{k-1,2}} + \beta^4 a_{\tau_{k-1,1}} + \beta^5 a_{\tau_{k-1,0}}}_{(k-1)\text{-th epoch}} + \cdots + \beta^{3(k-1)} a_{\tau_{1,0}},$$

$$m_1 = \underbrace{a_{\tau_{k,1}} + \beta a_{\tau_{k,0}}}_{k\text{-th epoch}} + \underbrace{\beta^2 a_{\tau_{k-1,2}} + \beta^3 a_{\tau_{k-1,1}} + \beta^4 a_{\tau_{k-1,0}}}_{(k-1)\text{-th epoch}} + \cdots + \beta^{3(k-1)} a_{\tau_{1,0}},$$

$$m_0 = \underbrace{a_{\tau_{k,0}}}_{k\text{-th epoch}} + \underbrace{\beta^1 a_{\tau_{k-1,2}} + \beta^2 a_{\tau_{k-1,1}} + \beta^3 a_{\tau_{k-1,0}}}_{(k-1)\text{-th epoch}} + \cdots + \beta^{3(k-1)} a_{\tau_{1,0}}.$$

Further, we introduce the following deterministic constants.

$$f_2 = a_2(\underbrace{1+\beta+\beta^2}_{k\text{-th epoch}} + \underbrace{\beta^3+\beta^4+\beta^5}_{(k-1)\text{-th epoch}} \cdots + \beta^{3(k-1)}),$$

$$f_1 = a_1(\underbrace{1+\beta+\beta^2}_{k\text{-th epoch}} + \underbrace{\beta^3+\beta^4+\beta^5}_{(k-1)\text{-th epoch}} \cdots + \beta^{3(k-1)}),$$

$$f_0 = a_0(\underbrace{1+\beta+\beta^2}_{k\text{-th epoch}} + \underbrace{\beta^3+\beta^4+\beta^5}_{(k-1)\text{-th epoch}} \cdots + \beta^{3(k-1)}).$$

For the simplicity of analysis, we denote

$$\sum_{i=0}^{2} m_i = \left(\sum_{i=0}^{2} m_i\right)_k + \left(\sum_{i=0}^{2} m_i\right)_{k-1} + \cdots + \left(\sum_{i=0}^{2} m_i\right)_1,$$

$$\sum_{i=0}^{2} f_i = \left(\sum_{i=0}^{2} f_i\right)_k + \left(\sum_{i=0}^{2} f_i\right)_{k-1} + \cdots + \left(\sum_{i=0}^{2} f_i\right)_1,$$

where $\left(\sum_{i=0}^{2} m_i\right)_k$ and $\left(\sum_{i=0}^{2} f_i\right)_k$ contain the summand in the k-th epoch, so on so for. **Now, we try to calculate** $\mathbb{E}\left[\sum_{i=0}^{2} m_i - \sum_{i=0}^{2} f_i\right]$, where the expectation is taken on all the possible draws until the end of k-th round (epoch).

Lemma G.2. *In the colorb-ball example above, we have*

$$\mathbb{E}\left[\sum_{i=0}^{2} m_i - \sum_{i=0}^{2} f_i\right] = \sum_{i=0}^{2} a_i \beta^{(k-1)n}(-\frac{2}{6}\beta^1 - \frac{4}{6}\beta^2).$$

For general $n = 1, 2, 3, \cdots$, we have

$$\mathbb{E}\left[\sum_{i=0}^{n-1} m_i - \sum_{i=0}^{n-1} f_i\right] = \sum_{i=0}^{n-1} a_i \beta^{(k-1)n}(-\frac{1}{n}\beta^1 \cdots - \frac{n-1}{n}\beta^{n-1}).$$

Proof. Since $\mathbb{E}[\cdot] = \mathbb{E}[\mathbb{E}_k[\cdot]]$, we first calculate $\mathbb{E}_k\left[\sum_{i=0}^{2} m_i - \sum_{i=0}^{2} f_i\right]$. Since all the history before k-th epoch is fixed, we delegate the history part to later analysis. Now, we focus on $\left(\sum_{i=0}^{2} m_i\right)_k - \left(\sum_{i=0}^{2} f_i\right)_k$. As shown in Figure 15, there are $(3)! = 6$ possible order in the k-th epoch.

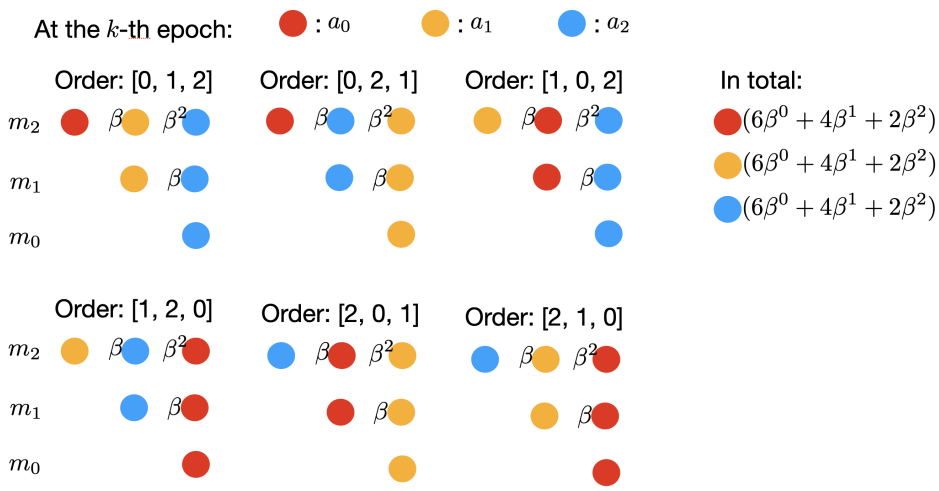


Figure 15: All possible order in the k-th epoch, i.e., $\sum_s \left(\sum_{i=0}^{2} m_i\right)_k$.

To calculate $\mathbb{E}_k(\cdot)$, we need to sum up all 6 possible shuffle order. Figure 15 tells us: when all the history sampling orders are fixed, we have: (we use the shorthand $\sum_s := \sum_{\text{all 6 possible shuffle}}$)

$$\begin{aligned}\sum_s \left\{\left(\sum_{i=0}^{2} m_i\right)_k - \left(\sum_{i=0}^{2} f_i\right)_k\right\} &= \sum_{i=0}^{2} a_i(6\beta^0 + 4\beta^1 + 2\beta^2) - \sum_{i=0}^{2} a_i(6\beta^0 + 6\beta^1 + 6\beta^2) \\ &= \sum_{i=0}^{2} a_i(-2\beta^1 - 4\beta^2).\end{aligned}$$

Since all the sampling order occur with the same probability $1/6$, we have:

$$\mathbb{E}_k\left\{\left(\sum_{i=0}^{2} m_i\right)_k - \left(\sum_{i=0}^{2} f_i\right)_k\right\} = \sum_{i=0}^{2} a_i(-\frac{2}{6}\beta^1 - \frac{4}{6}\beta^2).$$

Now we move one step further to calculate $\mathbb{E}_{k-1}\mathbb{E}_k\left[\sum_{i=0}^{2} m_i - \sum_{i=0}^{2} f_i\right]$. Since all the history before $(k-1)$-th epoch is fixed, we do not discuss them for now. In particular, we need to calculate

$$
\begin{aligned}
& \mathbb{E}_{k-1}\left\{\mathbb{E}_k\left[\left(\sum_{i=0}^{2} m_i\right)_k - \left(\sum_{i=0}^{2} f_i\right)_k + \left(\sum_{i=0}^{2} m_i\right)_{k-1} - \left(\sum_{i=0}^{2} f_i\right)_{k-1}\right]\right\} \\
= \; & \mathbb{E}_{k-1}\left\{\sum_{i=0}^{2} a_i(-\frac{2}{6}\beta^1 - \frac{4}{6}\beta^2) + \left(\sum_{i=0}^{2} m_i\right)_{k-1} - \left(\sum_{i=0}^{2} f_i\right)_{k-1}\right\}.
\end{aligned}
$$

We highlight the blue term since it is "the residue from the k-th epoch", this term will be canceled out later. Similarly as Figure 15, we sum up all possible draws $\sum_s \left(\sum_{i=0}^{2} m_i\right)_{k-1}$ in Figure 16.

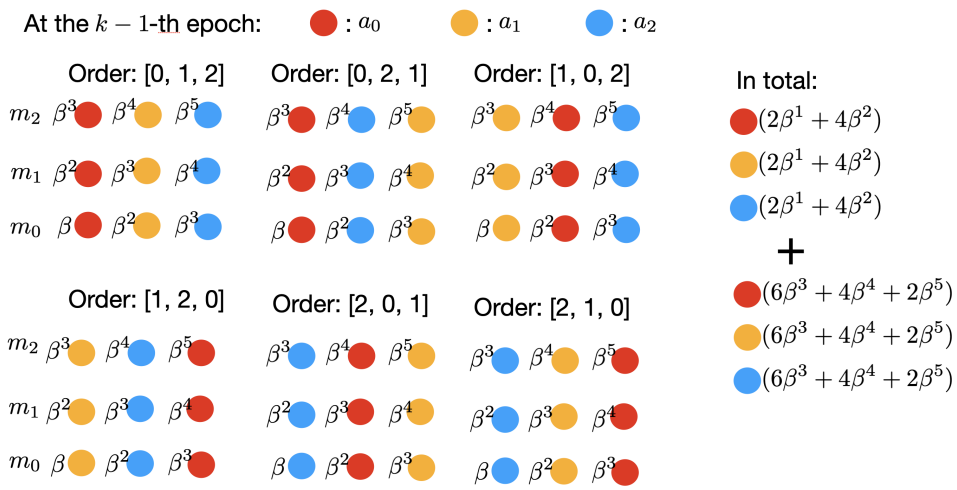


Figure 16: All possible order in the $k-1$-th epoch, i.e., $\sum_s \left(\sum_{i=0}^{2} m_i\right)_{k-1}$.

As shown in Figure 16, we have

$$
\begin{aligned}
& \sum_s \left\{\sum_{i=0}^{2} a_i(-\frac{2}{6}\beta^1 - \frac{4}{6}\beta^2) + \left(\sum_{i=0}^{2} m_i\right)_{k-1} - \left(\sum_{i=0}^{2} f_i\right)_{k-1}\right\} \\
= \; & \underbrace{\sum_{i=0}^{2} a_i(-2\beta^1 - 4\beta^2) + \sum_{i=0}^{2} a_i(2\beta^1 + 4\beta^2)}_{\text{cancel out the } k\text{-th residue}} + \underbrace{\sum_{i=0}^{2} a_i(6\beta^3 + 4\beta^4 + 2\beta^5) - \sum_{i=0}^{2} a_i(6\beta^3 + 6\beta^4 + 6\beta^5)}_{\text{create the } (k-1)\text{-th residue}} \\
= \; & \sum_{i=0}^{2} a_i(-2\beta^4 - 4\beta^5). \qquad (29)
\end{aligned}
$$

Equivalently,

$$
\mathbb{E}_{k-1}\left\{\mathbb{E}_k\left[\left(\sum_{i=0}^{2} m_i\right)_k - \left(\sum_{i=0}^{2} f_i\right)_k + \left(\sum_{i=0}^{2} m_i\right)_{k-1} - \left(\sum_{i=0}^{2} f_i\right)_{k-1}\right]\right\} \overset{\text{Figure 16}}{=} \sum_{i=0}^{2} a_i(-\frac{2}{6}\beta^4 - \frac{4}{6}\beta^5)
$$

From (29) we find out that: the residue from the k-th epoch will be cancelled out by part of the red terms in the $(k-1)$-th epoch; meanwhile, new residue term will be created. Moreover, we observe the following pattern in the residue : 1) the order of β involves the largest & second largest order that possibly occur within the epoch. 2) the coefficients are fixed to be $-\frac{2}{6}$ and $-\frac{4}{6}$. This is actually an universal pattern for every epoch from 1 to k. We illustrate this pattern in the following Figure 17.

Combining the above analysis and Figure 17, we observe the following pattern: when iteratively plugging in $\mathbb{E}_{k-j}$, the residue term remaining in the $(k-j+1)$-th epoch will be erased (i.e. $\sum_{i=0}^{2} a_i\beta^{(j-1)}(-2\beta^4 - 4\beta^5)$), whereas a new residue term will be produced (i.e. $\sum_{i=0}^{2} a_i\beta^j(-2\beta^4 -$

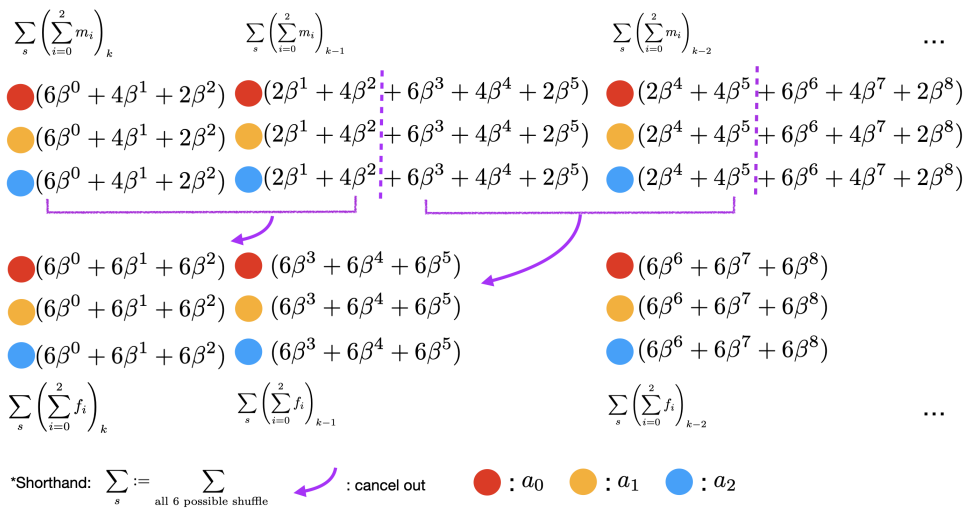


Figure 17: For every k, $\sum_s \left(\sum_{i=0}^{2} m_i\right)_k - \sum_s \left(\sum_{i=0}^{2} f_i\right)_k$ will create residues, while these residues will be canceled out in the $(k-1)$-th epoch.

$4\beta^5)$). Since $\mathbb{E}[\cdot] = \mathbb{E}[\mathbb{E}_1(\mathbb{E}_{k-1}(\cdots \mathbb{E}_k(\cdot)))]$, we derive the following result by repeatedly applying the steps in (29):

$$\mathbb{E}\left[\sum_{i=0}^{2} m_i - \sum_{i=0}^{2} f_i\right] = \sum_{i=0}^{2} a_i \beta^{(k-1)n}(-\frac{2}{6}\beta^1 - \frac{4}{6}\beta^2).$$

The above analysis also holds for general $n = 1, 2, 3, \cdots$, we just need to change the constants in the following form:

$$\mathbb{E}\left[\sum_{i=0}^{n-1} m_i - \sum_{i=0}^{n-1} f_i\right] = \sum_{i=0}^{n-1} a_i \beta^{(k-1)n}(-\frac{1}{n}\beta^1 \cdots - \frac{n-1}{n}\beta^{n-1}).$$

Since the analysis for general "n balls" is exactly the same as that of $n = 3$, we omit the proof for brevity.

□

The color-ball toy example above can provide insights in bounding $\mathbb{E}\left[\sum_{i=0}^{n-1}(m_{l,k,i} - \partial_l f_i(x_{k,0}))\right]$. However, there is still some gap between $\mathbb{E}\left[\sum_{i=0}^{n-1}(m_{l,k,i} - \partial_l f_i(x_{k,0}))\right]$ and the quantities in Lemma G.2.

We elaborate as follows:

- To mimic the color-ball example, we need to expand $\sum_{i=0}^{n-1} \partial_l f_i(x_{k,0})$ into an *infinite* sum sequence: $\sum_{i=0}^{n-1} \partial_l f_i(x_{k,0})(1-\beta_1)(1+\beta_1+\cdots\beta_1^{\infty})$. However, $\sum_{i=0}^{n-1} m_{l,k,i}$ is a *finite* sum sequence up to the order of β^{kn}. In contrast, both sequences in the color-ball example are "finite sum". As such, there is an error term caused by "finite sum v.s. infinite sum".
- When taking the expectation, the variable x in each possible trajectory is different. In contrast, in the color-ball example, $\{a_i\}_{i=0}^{2}$ are fixed in all shuffling order (so it is much easier to calculate the expectation by summing them up).

  To mimic the color-ball example, we repeatedly take conditional expectation at the beginning of each k-th epoch. In this way, $x_{k,0}$ will be fix. Despite $x_{k,i}$ is still changing across the trajectory, we can transform $x_{k,i}$ into $x_{k,0}$ by using Lipschitz property.

- In each trajectory of $\sum_{i=0}^{n-1} m_{l,k,i}$, the variable x in the summand of $m_{l,k,i}$ varies with k and i; while the variable in $\partial_l f_i(x_{k,0})$ is fixed to be $x_{k,0}$. In contrast, in the color-ball example, $\{a_i\}_{i=0}^2$ are the same across the epoch.

  To mimic the final step in the color-ball example (Figure 17), at each step of conditional expectation, we need to simultaneously move the variables in $m_{l,k,i}$ and $\partial_l f_i(x_{k,0})$ (using Lipschitz property) so that they can match and cancel out with each other. This operation will introduce new error terms and it is our duty to put them under control.

We omit the proof for bounding $\mathbb{E}\left[\sum_{i=0}^{n-1}\left(m_{l,k,i}-\partial_l f_i(x_{k,0})\right)\right]$ since this is not our actual goal. Instead, we will directly use the above ideas to bound $\mathbb{E}\left[(b_1)\right]$. To do so, we need to further tackle the issue **(ii)** and **(iii)** mentioned before. We explain as follows.

Bounding $\mathbb{E}\left[\sum_{l \text{ large}}(b_1)\right]$: Step 2. We now resolve (ii), i.e., handle the effect of $\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,0}}}$. The key idea is as follows: when we calculate $\mathbb{E}\left[(b_1)\right]$, we sequentially take conditional expectation $\mathbb{E}_k(\cdot)$, $\mathbb{E}_{k-1}(\cdot)$, etc.. When taking $\mathbb{E}_k(\cdot)$, we will fix all the historical information up to k-th epoch, so $\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,0}}}$ can be regarded as a constant. In this sense, $\mathbb{E}_k\left[\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,0}}}\sum_{i=0}^{n-1}\left(m_{l,k,i}-\partial_l f_i(x_{k,0})\right)\right]$ can be calculated following the same idea as color-ball toy example.

However, when taking $\mathbb{E}_{k-1}(\cdot)$, $\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,0}}}$ will become a random variable which changes with different trajectories. In this case, the color-ball method *cannot* be applied. To fix this issue, we introduce the following lemma to change $\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,0}}}$ into $\frac{\partial_l f(x_{k-1,0})}{\sqrt{v_{k-1,0}}}$, which can again be regarded as a fixed constant when taking $\mathbb{E}_{k-1}(\cdot)$.

Lemma G.3. *Suppose Assumption 2.1 holds and $\beta_1 < \sqrt{\beta_2}$. For any integer $j \in [0,k]$, if $\max_i |\partial_l f_i(x_{k,0})| \geq Q_k$, $\max_i |\partial_l f_i(x_{k-1,0})| \geq Q_{k-1}$, $\cdots$, $\max_i |\partial_l f_i(x_{k-j,0})| \geq Q_{k-j}$ (where $Q_k := \triangle_1 \frac{n\sqrt{n}}{\sqrt{k}} \frac{32\sqrt{2}}{(1-\beta_2)^n \beta_2^n}$), then we have the following result:*

$$\left|\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,0}}}-\frac{\partial_l f(x_{k-j,0})}{\sqrt{v_{l,k-j,0}}}\right| \leq \frac{1}{1-\frac{1}{\sqrt{\beta_2^n}}}\frac{n\triangle_{n(k-j)}}{\sqrt{v_{l,k-j,0}}}+j\sqrt{\frac{2\rho_3^2}{\beta_2^n}}\frac{1}{\left(1-\frac{1-\beta_2}{2}\left(-1+\frac{4\rho_2 n}{\beta_2^n}\right)\right)}\delta_1.$$

The proof of Lemma G.3 can be seen in Appendix G.3.

Bounding $\mathbb{E}\left[\sum_{l \text{ large}}(b_1)\right]$: Step 3. Now we shift gear to tackle (iii): handling the random variable " l large". We rewrite " l large" into the indicator function as follows:

$$\mathbb{E}\left[\sum_{l \text{ large}}(b_1)\right]=\mathbb{E}\left[\sum_{l \text{ large}}\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,0}}}\sum_{i=0}^{n-1}\left(m_{l,k,i}-\partial_l f_i(x_{k,0})\right)\right]=\mathbb{E}\left[\sum_{l=1}^{d}\mathbb{I}_k\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,0}}}\sum_{i=0}^{n-1}\left(m_{l,k,i}-\partial_l f_i(x_{k,0})\right)\right],$$

where $\mathbb{I}_k := \mathbb{I}\left(\max_i |\partial_l f_i(x_{k,0})| \geq Q_k := \triangle_1 \frac{n\sqrt{n}}{\sqrt{k}} \frac{32\sqrt{2}}{(1-\beta_2)^n \beta_2^n}\right)$ is the indicator function ($\mathbb{I}(A)=1$ when event A holds and $\mathbb{I}(A)=0$ otherwise.) Similarly as before, when taking $\mathbb{E}_k(\cdot)$, " $\mathbb{I}_k$" can be regarded as a constant index. Therefore, $\mathbb{E}_k\left[\sum_{l=1}^{d}\mathbb{I}_k\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,0}}}\sum_{i=0}^{n-1}\left(m_{l,k,i}-\partial_l f_i(x_{k,0})\right)\right]$ can be calculated following the same idea as color-ball toy example. However, when taking $\mathbb{E}_{k-1}(\cdot)$, $\mathbb{I}_k$ will become a random variable which changes with different trajectories. In this case, the color-ball method *cannot* be applied. Similarly as before, we introduce the following lemma to change $\mathbb{I}_k$ into $\mathbb{I}_{k-1}$ (defined later), which can again be regarded as a fixed when taking $\mathbb{E}_{k-1}(\cdot)$.

Lemma G.4. *Suppose Assumption 2.1 holds and $\beta_1 < \sqrt{\beta_2}$. For $0 \leq j \leq k$, we define $\mathbb{I}_{k-j} := \mathbb{I}\left(\max_i |\partial_l f_i(x_{k-j,0})| \geq \sum_{p=k-j}^{k} Q_p\right)$, where $Q_k := \triangle_1 \frac{n\sqrt{n}}{\sqrt{k}} \frac{32\sqrt{2}}{(1-\beta_2)^n \beta_2^n}$, then we have the following results.*

$$\begin{aligned}\mathbb{I}_k &= \mathbb{I}\left(\max_i |\partial_l f_i(x_{k,0})| \geq Q_k \textit{ and } \max_i |\partial_l f_i(x_{k-j,0})| \geq \sum_{p=k-j}^{k} Q_p\right) \\ &+ \mathbb{I}\left(\max_i |\partial_l f_i(x_{k,0})| \geq Q_k \textit{ and } \max_i |\partial_l f_i(x_{k-j,0})| \leq \sum_{p=k-j}^{k} Q_p\right), \end{aligned} \tag{30}$$

$$\mathbb{I}\left(\max_i |\partial_l f_i(x_{k,0})| \geq Q_k \textit{ and } \max_i |\partial_l f_i(x_{k-j,0})| \geq \sum_{p=k-j}^{k} Q_p\right) = \mathbb{I}\left(\max_i |\partial_l f_i(x_{k-j,0})| \geq \sum_{p=k-j}^{k} Q_p\right) = \mathbb{I}_{k-j}. \tag{31}$$

Proof. Equation (30) is straightforward, we only prove (31) here. Under Assumption 2.1, we have $|\partial_l f_i(x_{k,0}) - \partial_l f_i(x_{k-j,0})| \leq n\triangle_{n(k-1)} + \cdots + n\triangle_{n(k-j)} \leq Q_{k-1} + \cdots + Q_{k-j}$. To show the second inequality, it requires comparing the value between Q_k and $n\triangle_{n(k)}$, which are both problem-dependent constants. Here, the inequality holds when $n\triangle_{n(k)} \leq Q_k$. If otherwise, we can always define $\tilde{Q}_k := \max\{Q_k, n\triangle_{n(k)}\}$ and the inequality still holds by changing all the Q_k into $\tilde{Q}_k$. We temporarily omit this step for now.

Since $|\partial_l f_i(x_{k,0}) - \partial_l f_i(x_{k-j,0})| \leq Q_{k-1} + \cdots + Q_{k-j}$, the event $\left\{\max_i |\partial_l f_i(x_{k-j,0})| \geq \sum_{p=k-j}^{k} Q_p\right\}$ implies the event $\{\max_i |\partial_l f_i(x_{k,0})| \geq Q_k\}$, so the proof is completed. □

Now we are ready to bound $\mathbb{E}\left[\sum_{l \text{ large}}(b_1)\right]$. Combining Step 1, 2 and 3 together, we prove the following Lemma G.5.

Lemma G.5. *Under Assumption 2.1, consider $\beta_1 < \sqrt{\beta_2}$, when k is large such that: $k \geq 4$; $\beta_1^{(k-1)n} \leq \frac{\beta_1^n}{\sqrt{k-1}}$, we have the following result:*

$$\mathbb{E}\left[\sum_{l \textit{ large}}(b_1)\right] \geq -\mathcal{O}\left(\frac{1}{\sqrt{k}}\right) - \mathcal{O}\left(\delta_1 \mathbb{E}\left[\sum_{l=1}^{d}\sum_{i=0}^{n-1} |\partial_l f_i(x_{k,0})|\right]\right),$$

where $\delta_1 = (1-\beta_2)\frac{\left(-1+\frac{4n\rho_2}{\beta_2^n}\right)}{2} + \left(\frac{1}{\sqrt{\beta_2^n}} - 1\right)$, which goes to 0 when β_2 goes to 1.

Proof can be seen in Appendix G.4. Now we bound the error terms: $\mathbb{E}\left[\sum_{l \text{ large}}\{(a_2) + (b_2)\}\right]$. Since they are multiplied by δ_1, these two terms vanish when $\beta_2 \to 1$.

Bounding $\mathbb{E}\left[\sum_{l \textbf{ large}}\{(a_2) + (b_2)\}\right]$. We bound these two terms in the following Lemma G.6.

Lemma G.6. *Given all the history up to $x_{k,0}$, we denote α as the index of the coordinate with the greatest gradient: $\alpha = \arg\max_{l=1,2,\cdots,d} |\partial_l f(x_{k,0})|$. Under Assumption 2.1 and 2.2, consider $\beta_1 < \sqrt{\beta_2}$, when k is large such that: $k \geq 4$; $\beta_1^{(k-1)n} \leq \frac{\beta_1^n}{\sqrt{k-1}}$, we have the following results:*

$$\mathbb{E}\left[\sum_{l \textit{ large}}(a_2)\right] := \mathbb{E}\left[\sum_{l \textit{ large}} \delta_1 \sqrt{\frac{2\rho_3^2}{\beta_2^n}} \sum_{i=0}^{n-1} |\partial_l f_i(x_{k,0})|\right] \overset{(21)}{\leq} \mathcal{O}\left(\delta_1 |\partial_\alpha f(x_{k,0})|\right) + \mathcal{O}\left(\delta_1\sqrt{D_0}\right),$$

$$\mathbb{E}\left[\sum_{l \textit{ large}}(b_2)\right] := \mathbb{E}\left[\sum_{l \textit{ large}} \delta_1 \sqrt{\frac{2\rho_3^2}{\beta_2^n}} \sum_{i=0}^{n-1} |m_{l,k,i} - \partial_l f_i(x_{k,0})|\right] \leq \mathcal{O}\left(\delta_1 |\partial_\alpha f(x_{k,0})|\right) + \mathcal{O}\left(\delta_1\sqrt{D_0}\right) + \mathcal{O}\left(\frac{1}{\sqrt{k}}\right),$$

where δ_1 is defined in (14)*, D_0 is defined in Assumption 2.2.*

Detailed proof of Lemma G.6 is shown in Appendix G.8.

Bounding $\mathbb{E}\left\{\sum_{l\ \textbf{large}}(a_1)+\sum_{l\ \textbf{large}}(b_1)-\sum_{l\ \textbf{large}}\{(a_2)+(b_2)\}\right\}$. With the help of Lemma G.5 and G.6, we can bound $\mathbb{E}\left\{\sum_{l\ \text{large}}(a_1)+\sum_{l\ \text{large}}(b_1)-\sum_{l\ \text{large}}\{(a_2)+(b_2)\}\right\}$. The intuition is as follows:

- (a_1) is in the order of $|\partial_l f(x_{k,0})|$, multiplied by some positive constant;
- $\sum_{l\ \text{large}}\{(a_2)+(b_2)\}$ vanishes when $\beta_2 \to 1$.
- $\mathbb{E}\left\{\sum_{l\ \text{large}}(b_1)\right\}$ vanishes when $\beta_2 \to 1$ and k goes to infinity.

Therefore, $\mathbb{E}\left\{\sum_{l\ \text{large}}(a_1)+\sum_{l\ \text{large}}(b_1)-\sum_{l\ \text{large}}\{(a_2)+(b_2)\}\right\}$ is still in the order of $|\partial_l f(x_{k,0})|$ (multiplied by a positive constant when β_2 is large). More formal results are shown in Lemma G.7.

Lemma G.7. *Under Assumption 2.1 and 2.2, when the hyperparameters satisfy: i) $\beta_1 < \sqrt{\beta_2} < 1$, ii) β_2 is large enough such that $A(\beta_2)$ is small enough to satisfy* (32)*, where $A(\beta_2)$ is a non-negative constant that approaches 0 when β_2 approaches* 1*. More Specifically, $A(\beta_2)$ needs to satisfy (ρ_1, ρ_2 and ρ_3 are defined in Appendix F.1).*

$$
\begin{aligned}
A(\beta_2) \quad &:= \quad \left\{(1-\beta_2)\frac{\left(-1+\frac{4n\rho_2}{\beta_2^n}\right)}{2}+\left(\frac{1}{\sqrt{\beta_2^n}}-1\right)\right\}\left(\sqrt{\frac{2\rho_3^2}{\beta_2^n}}4n+\left(\sqrt{\frac{2\rho_3^2}{\beta_2^n}}\frac{3n}{\left(1-\frac{1-\beta_2}{2}\left(-1+\frac{4\rho_2 n}{\beta_2^n}\right)\right)}\right)\frac{1}{(1-\beta_1^n)}\right)\sqrt{D_1}\rho_1 d \\
&\leq \quad \frac{1}{\sqrt{10D_1 d}}, \qquad (32)
\end{aligned}
$$

Then, we have the following result when k is large enough such that $\beta_1^{(k-1)n} \leq \frac{\beta_1^n}{\sqrt{k-1}}$ and $k \geq 4$:

$$
\begin{aligned}
& \mathbb{E}\left\{\sum_{l\ large}(a_1)+\sum_{l\ large}(b_1)-\sum_{l\ large}\{(a_2)+(b_2)\}\right\} \\
:= \quad & \mathbb{E}\left[\sum_{l\ large}\frac{\partial_l f\left(x_{k,0}\right)^2}{\sqrt{v_{l,k,0}}}\right]+\mathbb{E}\left[\sum_{l\ large}\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,0}}}\sum_{i=0}^{n-1}\left(m_{l,k,i}-\partial_l f_i(x_{k,0})\right)\right] \\
& -d\delta_1\sqrt{\frac{2\rho_3^2}{\beta_2^n}}\mathbb{E}\left[\sum_{i=0}^{n-1}\left|\partial_l f_i\left(x_{k,0}\right)\right|\right]-d\delta_1\sqrt{\frac{2\rho_3^2}{\beta_2^n}}\mathbb{E}\left[\sum_{i=0}^{n-1}\left|m_{l,k,i}-\partial_l f_i(x_{k,0})\right|\right] \\
\geq \quad & \frac{1}{d\sqrt{10D_1 d}}\mathbb{E}\min\left\{\sqrt{\frac{2D_1 d}{D_0}}\|\nabla f(x_{k,0})\|_2^2, \|\nabla f(x_{k,0})\|_1\right\} \\
& -\mathcal{O}(\frac{1}{\sqrt{k}})-\mathcal{O}(\sqrt{D_0}).
\end{aligned}
$$

Proof of Lemma G.7 can be seen in Appendix G.9.

Remark G.8. Condition (32) specifies the smallest threshold of β_2 to ensure the convergence. This condition can be translated into the threshold funtion $\gamma_1(n)$ mentioned in Theorem 3.1. As a rough estimate, Lemma G.7 requires $\beta_2 \geq \gamma_1(n) = 1-\mathcal{O}\left((1-\beta_1^n)/(n^2\rho)\right)$, where $\rho = \rho_1\rho_2\rho_3$. As discussed in Appendix F.1, we have $0 \leq \rho_3 \leq \sqrt{n}\rho_1 \leq n$. When ρ_1, ρ_2, ρ_3 achieve their upper bound at the same time, we get the worst case bound $\beta_2 \geq \gamma_1(n) = 1-\mathcal{O}\left((1-\beta_1^n)/n^{4.5}\right)$. However, ρ is highly dependent on the problem instance $f(x)$ and training process. Our experiments in Appendix D.2 shows that ρ is often much smaller than its worst case bound, making the threshold of β_2 lower than it appears to be.

We further remark that $\gamma_1(n) \propto (1-\beta_1^n)$. That is to say, as β_1 increases, the threshold of β_2 also slowly increases (to ensure the convergence). However, such phenomenon is not observed in our experiments. We conjecture $\gamma_1(n)$ should be independent of β_1, removing the dependence $(1-\beta_1^n)$ is considered as an interesting future direction.

Remark G.9. We emphasize that the constant term $\mathcal{O}(\sqrt{D_0})$ will vanish to 0 as β_2 goes to 1. This can be seen in the proof in Appendix G.9 (the definition of F_4) and Remark G.17.

Based on Lemma G.7, we can further rewrite (28) as follows.

$$\begin{aligned}
\mathbb{E}\left[\sum_{l=1}^{d}\sum_{i=0}^{n-1}\partial_l f(x_{k,0})\frac{m_{l,k,i}}{\sqrt{v_{l,k,i}}}\right] &\overset{(28)}{\geq} \mathbb{E}\left\{\sum_{l \text{ large}}(a_1)+\sum_{l \text{ large}}(b_1)-\sum_{l \text{ large}}\{(a_2)+(b_2)\}\right\}-dF_1\frac{1}{\sqrt{k}}, \\
&\overset{\text{Lemma G.7}}{\geq} \frac{1}{d\sqrt{10D_1 d}}\mathbb{E}\min\left\{\sqrt{\frac{2D_1 d}{D_0}}\|\nabla f(x_{k,0})\|_2^2, \|\nabla f(x_{k,0})\|_1\right\} \\
&\quad -\mathcal{O}(\frac{1}{\sqrt{k}})-\mathcal{O}(\sqrt{D_0})-d\frac{F_1}{\sqrt{k}} \qquad (33)
\end{aligned}$$

The proof of Theorem 3.1 is concluded by plugging (33) into Descent Lemma (22) and then taking telescope some from $k = t_0$ to $k = T$. These steps are quite standard in non-convex optimization. We finish these calculation in Lemma G.10

Lemma G.10. *When inequality* (33) *holds, we have the following results based on Descent Lemma* (22)*:*

$$\begin{aligned}
&\min_{k\in[1,T]}\mathbb{E}\left[\min\left\{\sqrt{\frac{2D_1 d}{D_0}}\|\nabla f(x_{k,0})\|_2^2, \|\nabla f(x_{k,0})\|_1\right\}\right] \\
&= \mathcal{O}\left(\frac{\log T}{\sqrt{T}}\right)+\mathcal{O}(\sqrt{D_0}).
\end{aligned}$$

Proof of Lemma G.10 can be seen in Appendix G.10. Now, the whole proof of Theorem 3.1 is completed.

G.2 Proof of Lemma G.1

To prove Lemma G.1, we first show the following result.

Lemma G.11. *Under Assumption 2.1, if the l-th component of $\nabla f_i(x_{k,0})$ satisfies $\max_i |\partial_l f_i(x_{k,0})| \geq Q_k := \triangle_1 \frac{n\sqrt{n}}{\sqrt{k}}\frac{32\sqrt{2}}{(1-\beta_2)^n\beta_2^n}$ we have*

$$\frac{1}{\sqrt{v_{l,k,i}}} \geq \frac{1}{\sqrt{v_{l,k,0}}}\left(1-\frac{1-\beta_2}{2}\left(-1+\frac{4\rho_2 i}{\beta_2^n}\right)\right).$$

The proof idea of Lemma G.11 is similar as Lemma F.3 in (Shi et al., 2020), but with the following differences:

1. The condition of Lemma F.3 in (Shi et al., 2020) requires the full batch gradient $|\partial_l f(x_{k,0})| \geq nQ_k$, which is different from our condition. Here, we choose the condition of Lemma G.11 because it meets the need of our later decomposition strategy in (28) .
2. The constant Q_k is different. The difference is due to the different update formula of RMSProp and general Adam.

Anyhow, Lemma G.11 can be easily proved following most of the steps in Lemma F.3 (Shi et al., 2020). We omit the proof for brevity.

To prove Lemma G.1, we only prove (25). The proof for (24) can be analogized from the proof of (25). We discuss the following two cases:

Case 1: When $\partial_l f\left(x_{k,0}\right)\left(m_{l,k,i}-\partial_l f_i(x_{k,0})\right) \geq 0$, we have

$$\frac{\partial_l f\left(x_{k,0}\right)\left(m_{l,k,i}-\partial_l f_i(x_{k,0})\right)}{\sqrt{v_{l,k,i}}} \overset{\text{Lemma G.11}}{\geq} \frac{\partial_l f\left(x_{k,0}\right)\left(m_{l,k,i}-\partial_l f_i(x_{k,0})\right)}{\sqrt{v_{l,k,0}}}\left(1-\frac{1-\beta_2}{2}\left(-1+\frac{4\rho_2 i}{\beta_2^n}\right)\right), \qquad (34)$$

Case 2: When $\partial_l f\left(x_{k,0}\right)\left(m_{l,k,i}-\partial_l f_i(x_{k,0})\right) \leq 0$, we have

$$\frac{\partial_l f\left(x_{k,0}\right)\left(m_{l,k,i}-\partial_l f_i(x_{k,0})\right)}{\sqrt{v_{l,k,i}}} \overset{(a)}{\geq} \frac{\partial_l f(x_{k,0})(m_{l,k,i}-\partial_l f_i(x_{k,0}))}{\sqrt{v_{l,k,0}}} \frac{1}{\sqrt{\beta_2^i}},$$

where (a) is because of $\sqrt{v_{l,k,i}} \geq \sqrt{\beta_2^i}\sqrt{v_{l,k,0}}$.

Combining these two cases, we have

$$\begin{aligned}
\sum \frac{\partial_l f\left(x_{k,0}\right)\left(m_{l,k,i}-\partial_l f_i(x_{k,0})\right)}{\sqrt{v_{l,k,i}}} &\geq \sum_{i-} \frac{\partial_l f\left(x_{k,0}\right)\left(m_{l,k,i}-\partial_l f_i(x_{k,0})\right)}{\sqrt{v_{l,k,0}}} \frac{1}{\sqrt{\beta_2^i}} \\
&\quad + \sum_{i+} \frac{\partial_l f\left(x_{k,0}\right)\left(m_{l,k,i}-\partial_l f_i(x_{k,0})\right)}{\sqrt{v_{l,k,0}}} \left(1-\frac{1-\beta_2}{2}\left(-1+\frac{4\rho_2 i}{\beta_2^n}\right)\right) \\
&= \sum_{i=0}^{n-1} \frac{\partial_l f\left(x_{k,0}\right)\left(m_{l,k,i}-\partial_l f_i(x_{k,0})\right)}{\sqrt{v_{l,k,0}}} \\
&\quad + \sum_{i-} \frac{\partial_l f\left(x_{k,0}\right)\left(m_{l,k,i}-\partial_l f_i(x_{k,0})\right)}{\sqrt{v_{l,k,0}}} \left(\frac{1}{\sqrt{\beta_2^i}}-1\right) \\
&\quad + \sum_{i+} \frac{\partial_l f\left(x_{k,0}\right)\left(m_{l,k,i}-\partial_l f_i(x_{k,0})\right)}{\sqrt{v_{l,k,0}}} \left(-\frac{1-\beta_2}{2}\left(-1+\frac{4\rho_2 i}{\beta_2^n}\right)\right) \\
&\geq \sum_{i=0}^{n-1} \frac{\partial_l f\left(x_{k,0}\right)\left(m_{l,k,i}-\partial_l f_i(x_{k,0})\right)}{\sqrt{v_{l,k,0}}} \\
&\quad -\delta_1 \left|\frac{\partial_l f\left(x_{k,0}\right)}{\sqrt{v_{l,k,0}}}\right| \sum_{i=0}^{n-1} \left|m_{l,k,i}-\partial_l f_i\left(x_{k,0}\right)\right|,
\end{aligned}$$

where $\delta_1 = (1-\beta_2)\frac{\left(-1+\frac{4n\rho_2}{\beta_2^n}\right)}{2} + \left(\frac{1}{\sqrt{\beta_2^n}}-1\right)$. Proof is completed.

G.3 Proof of Lemma G.3

To prove Lemma G.3, we first show the following result.

Lemma G.12. *Under Assumption 2.1, if the l-th component of $\nabla f_i(x_{k,0})$ satisfies $\max_i |\partial_l f_i(x_{k,0})| \geq Q_k := \triangle_1 \frac{n\sqrt{n}}{\sqrt{k}} \frac{32\sqrt{2}}{(1-\beta_2)^n \beta_2^n}$, we have*

$$\frac{v_{l,k,0}}{\frac{1}{n}\sum_i \partial_l f_{\tau_{k,i}}(x_{k,0})^2} \geq \frac{\beta_2^n}{2},$$

$$\frac{v_{l,k,0}}{\left(\partial_l f\left(x_{k,0}\right)\right)^2} \geq \frac{\beta_2^n}{2\rho_3^2}.$$

Proof. The proof idea of Lemma G.12 is silimar with Lemma F.1 in (Shi et al., 2020), but with the following differences:

1. The condition of Lemma F.1 in (Shi et al., 2020) requires the full batch gradient $|\partial_l f(x_{k,0})| \geq nQ_k$, which is different from our condition. Here, we choose the condition of Lemma G.11 because it meets the need of our later decomposition strategy in (28).

2. The constant Q_k is different. The difference is due to the different update formula of RMSProp and general Adam.

Lemma G.12 can be proved following most of the steps of Lemma F.1 in (Shi et al., 2020). We omit the proof for brevity. □

Now we prove Lemma G.3. We first prove the result for $j = 1$. We discuss the following two cases.

Case 1: when $\frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}} \geq \frac{\partial_l f(x_{k-1,0})}{\sqrt{v_{k-1,0}}}$: when $\partial_l f(x_{k,0}) \leq 0$, we have

$$\frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}} \overset{\text{Lemma G.11}}{\leq} \frac{\partial_l f(x_{k,0})}{\sqrt{v_{k-1,0}}} \left(1 - \frac{1-\beta_2}{2}\left(-1 + \frac{4\rho_2 n}{\beta_2^n}\right)\right).$$

When $\partial_l f(x_{k,0}) > 0$, we have

$$\frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}} \leq \frac{\partial_l f(x_{k,0})}{\sqrt{v_{k-1,0}}} \frac{1}{\sqrt{\beta_2^n}}$$

.

In conclusion, we have:

$$\begin{aligned}
\frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}} &\leq \max\left\{\frac{\partial_l f(x_{k,0})}{\sqrt{v_{k-1,0}}}\left(1 - \frac{1-\beta_2}{2}\left(-1 + \frac{4\rho_2 n}{\beta_2^n}\right)\right), \frac{\partial_l f(x_{k,0})}{\sqrt{v_{k-1,0}}}\frac{1}{\sqrt{\beta_2^n}}\right\} \\
&\leq \frac{\partial_l f(x_{k,0})}{\sqrt{v_{k-1,0}}} + \max\left\{\frac{\partial_l f(x_{k,0})}{\sqrt{v_{k-1,0}}}\left(-\frac{1-\beta_2}{2}\left(-1 + \frac{4\rho_2 n}{\beta_2^n}\right)\right), \frac{\partial_l f(x_{k,0})}{\sqrt{v_{k-1,0}}}\left(\frac{1}{\sqrt{\beta_2^n}} - 1\right)\right\} \\
&\leq \frac{\partial_l f(x_{k,0})}{\sqrt{v_{k-1,0}}} + \frac{|\partial_l f(x_{k,0})|}{\sqrt{v_{k-1,0}}}\left(\frac{1}{\sqrt{\beta_2^n}} - 1 + \frac{1-\beta_2}{2}\left(-1 + \frac{4\rho_2 n}{\beta_2^n}\right)\right) \\
&\overset{\text{Lemma G.11}}{\leq} \frac{\partial_l f(x_{k,0})}{\sqrt{v_{k-1,0}}} + \frac{|\partial_l f(x_{k,0})|}{\sqrt{v_{k,0}}}\left(\frac{1}{\sqrt{\beta_2^n}} - 1 + \frac{1-\beta_2}{2}\left(-1 + \frac{4\rho_2 n}{\beta_2^n}\right)\right)\frac{1}{\left(1 - \frac{1-\beta_2}{2}\left(-1 + \frac{4\rho_2 n}{\beta_2^n}\right)\right)} \\
&\overset{\text{Lemma G.12}}{\leq} \frac{\partial_l f(x_{k,0})}{\sqrt{v_{k-1,0}}} + \sqrt{\frac{2\rho_3^2}{\beta_2^n}}\left(\frac{1}{\sqrt{\beta_2^n}} - 1 + \frac{1-\beta_2}{2}\left(-1 + \frac{4\rho_2 n}{\beta_2^n}\right)\right)\frac{1}{\left(1 - \frac{1-\beta_2}{2}\left(-1 + \frac{4\rho_2 n}{\beta_2^n}\right)\right)} \\
&= \frac{\partial_l f(x_{k,0})}{\sqrt{v_{k-1,0}}} + \sqrt{\frac{2\rho_3^2}{\beta_2^n}}\frac{1}{\left(1 - \frac{1-\beta_2}{2}\left(-1 + \frac{4\rho_2 n}{\beta_2^n}\right)\right)}\delta_1,
\end{aligned}$$

where $\delta_1 = \left(\frac{1}{\sqrt{\beta_2^n}} - 1 + \frac{1-\beta_2}{2}\left(-1 + \frac{4\rho_2 n}{\beta_2^n}\right)\right)$ is a constant that goes to 0 when β_2 goes to 1. Therefore, we have

$$\begin{aligned}
\frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}} - \frac{\partial_l f(x_{k-1,0})}{\sqrt{v_{k-1,0}}} &\leq \frac{\partial_l f(x_{k,0}) - \partial_l f(x_{k-1,0})}{\sqrt{v_{k-1,0}}} + \sqrt{\frac{2\rho_3^2}{\beta_2^n}}\frac{1}{\left(1 - \frac{1-\beta_2}{2}\left(-1 + \frac{4\rho_2 n}{\beta_2^n}\right)\right)}\delta_1 \\
&\overset{\text{Lemma F.3}}{\leq} \frac{n\triangle_{n(k-1)}}{\sqrt{v_{k-1,0}}} + \sqrt{\frac{2\rho_3^2}{\beta_2^n}}\frac{1}{\left(1 - \frac{1-\beta_2}{2}\left(-1 + \frac{4\rho_2 n}{\beta_2^n}\right)\right)}\delta_1
\end{aligned}$$

Case 2: when $\frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}} \leq \frac{\partial_l f(x_{k-1,0})}{\sqrt{v_{k-1,0}}}$: when $\partial_l f(x_{k,0}) \geq 0$, we have

$$\frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}} \overset{\text{Lemma G.11}}{\geq} \frac{\partial_l f(x_{k,0})}{\sqrt{v_{k-1,0}}} \left(1 - \frac{1-\beta_2}{2}\left(-1 + \frac{4\rho_2 n}{\beta_2^n}\right)\right).$$

When $\partial_l f(x_{k,0}) < 0$, we have

$$\frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}} \geq \frac{\partial_l f(x_{k,0})}{\sqrt{v_{k-1,0}}} \frac{1}{\sqrt{\beta_2^n}}$$

.

Following the same strategy as in Case 1, we can show that

$$\frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}} \geq \frac{\partial_l f(x_{k,0})}{\sqrt{v_{k-1,0}}} - \sqrt{\frac{2\rho_3^2}{\beta_2^n}} \frac{1}{\left(1 - \frac{1-\beta_2}{2}\left(-1 + \frac{4\rho_2 n}{\beta_2^n}\right)\right)} \delta_1,$$

which further implies

$$\begin{aligned}
\frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}} - \frac{\partial_l f(x_{k-1,0})}{\sqrt{v_{k-1,0}}} &\geq \frac{\partial_l f(x_{k,0}) - \partial_l f(x_{k-1,0})}{\sqrt{v_{k-1,0}}} - \sqrt{\frac{2\rho_3^2}{\beta_2^n}} \frac{1}{\left(1 - \frac{1-\beta_2}{2}\left(-1 + \frac{4\rho_2 n}{\beta_2^n}\right)\right)} \delta_1 \\
&\overset{\text{Lemma F.3}}{\geq} -\frac{n\triangle_{n(k-1)}}{\sqrt{v_{k-1,0}}} - \sqrt{\frac{2\rho_3^2}{\beta_2^n}} \frac{1}{\left(1 - \frac{1-\beta_2}{2}\left(-1 + \frac{4\rho_2 n}{\beta_2^n}\right)\right)} \delta_1.
\end{aligned}$$

Case 1 and Case 2 together, we have

$$\left| \frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}} - \frac{\partial_l f(x_{k-1,0})}{\sqrt{v_{k-1,0}}} \right| \leq \frac{n\triangle_{n(k-1)}}{\sqrt{v_{k-1,0}}} + \sqrt{\frac{2\rho_3^2}{\beta_2^n}} \frac{1}{\left(1 - \frac{1-\beta_2}{2}\left(-1 + \frac{4\rho_2 n}{\beta_2^n}\right)\right)} \delta_1.$$

Now we consider the case when $j > 1$. Based on the above inequality, we have

$$\begin{aligned}
\left| \frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}} - \frac{\partial_l f(x_{k-j,0})}{\sqrt{v_{k-j,0}}} \right| &\leq \left(\frac{n\triangle_{n(k-1)}}{\sqrt{v_{k-1,0}}} + \frac{n\triangle_{n(k-2)}}{\sqrt{v_{k-2,0}}} + \cdots + \frac{n\triangle_{n(k-j)}}{\sqrt{v_{k-j,0}}} \right) + j\sqrt{\frac{2\rho_3^2}{\beta_2^n}} \frac{1}{\left(1 - \frac{1-\beta_2}{2}\left(-1 + \frac{4\rho_2 n}{\beta_2^n}\right)\right)} \delta_1 \\
&\leq \left(\frac{1}{\sqrt{\beta_2^{jn}}} + \frac{1}{\sqrt{\beta_2^{(j-1)n}}} + \cdots + 1 \right) \frac{n\triangle_{n(k-j)}}{\sqrt{v_{k-j,0}}} + j\sqrt{\frac{2\rho_3^2}{\beta_2^n}} \frac{1}{\left(1 - \frac{1-\beta_2}{2}\left(-1 + \frac{4\rho_2 n}{\beta_2^n}\right)\right)} \delta_1 \\
&\leq \frac{1}{1 - \frac{1}{\sqrt{\beta_2^n}}} \frac{n\triangle_{n(k-j)}}{\sqrt{v_{k-j,0}}} + j\sqrt{\frac{2\rho_3^2}{\beta_2^n}} \frac{1}{\left(1 - \frac{1-\beta_2}{2}\left(-1 + \frac{4\rho_2 n}{\beta_2^n}\right)\right)} \delta_1.
\end{aligned}$$

The proof is completed.

G.4 Proof of Lemma G.5

To prove Lemma G.5, we need to further decompose $\mathbb{E}[\sum_{l \text{ large}}(b_1)]$. First and foremost, we write $\sum_{i=0}^{n-1} m_{l,k,i}$ in an explicit form.

$$\begin{aligned}
m_{l,k,i} = \; & (1-\beta_1)\{\partial_l f_{\tau_{k,i}}(x_{k,i}) + \cdots + \beta_1^i \partial_l f_{\tau_{k,0}}(x_{k,0}) \\
& + \beta_1^{i+1} f_{\tau_{k-1,n-1}}(x_{k-1,n-1}) + \cdots + \beta_1^{i+n} f_{\tau_{k-1,0}}(x_{k-1,0}) \\
& + \\
& \vdots \\
& + \\
& + \beta_1^{(k-2)n+i+1} \partial_l f_{\tau_{1,n-1}}(x_{1,n-1}) + \cdots + \beta_1^{(k-1)n+i} \partial_l f_{\tau_{1,0}}(x_{1,0})\} \\
& + \beta_1^{(k-1)n+i+1} \partial_l f(x_{1,0})
\end{aligned} \tag{35}$$

Since $\partial_l f(x_{1,0}) = \partial_l f(x_{1,0})(1-\beta_1)(1+\beta_1+\cdots+\beta_1^{\infty})$ and $\partial_l f(x_{1,0}) = \sum_{i=0}^{n-1} \partial_l f_i(x_{1,0})$, we have

$$
\begin{aligned}
\beta_1^{(k-1)n+i+1}\partial_l f(x_{1,0}) \quad = \quad & (1-\beta_1)\{\beta_1^{(k-1)n+i+1}\partial_l f_0(x_{1,0}) + \cdots + \beta_1^{(k-1)n+i+1}\partial_l f_{n-1}(x_{1,0}) \\
& +\beta_1^{(k-1)n+i+2}\partial_l f_0(x_{1,0}) + \cdots + \beta_1^{(k-1)n+i+2}\partial_l f_{n-1}(x_{1,0}) \\
& + \\
& \vdots \\
& + \\
& +\beta_1^{\infty}\partial_l f_0(x_{1,0}) + \cdots + \beta_1^{\infty}\partial_l f_{n-1}(x_{1,0})\}
\end{aligned} \tag{36}
$$

Plugging (36) into (35), we have

$$
\begin{aligned}
m_{l,k,i} \quad = \quad & (1-\beta_1)\{\partial_l f_{\tau_{k,i}}(x_{k,i}) + \cdots + \beta_1^{i}\partial_l f_{\tau_{k,0}}(x_{k,0}) \\
& +\beta_1^{i+1} f_{\tau_{k-1,n-1}}(x_{k-1,n-1}) + \cdots + \beta_1^{i+n} f_{\tau_{k-1,0}}(x_{k-1,0}) \\
& + \cdots \\
& +\beta_1^{(k-2)n+i+1}\partial_l f_{\tau_{1,n-1}}(x_{1,n-1}) + \cdots + \beta_1^{(k-1)n+i}\partial_l f_{\tau_{1,0}}(x_{1,0}) \\
& +\beta_1^{(k-1)n+i+1}\partial_l f_0(x_{1,0}) + \cdots + \beta_1^{(k-1)n+i+1}\partial_l f_{n-1}(x_{1,0}) \\
& +\beta_1^{(k-1)n+i+2}\partial_l f_0(x_{1,0}) + \cdots + \beta_1^{(k-1)n+i+2}\partial_l f_{n-1}(x_{1,0}) \\
& + \cdots \\
& +\beta_1^{\infty}\partial_l f_0(x_{1,0}) + \cdots + \beta_1^{\infty}\partial_l f_{n-1}(x_{1,0})\}
\end{aligned} \tag{37}
$$

Using (37), we have (For each 'epoch', we suggest readers to read from bottom to the top.)

$$
\begin{aligned}
M_{l,k} := \sum_{i=0}^{n-1} m_{l,k,i} &= m_{l,k,n-1} + \cdots + m_{l,k,0} \\
&= (1-\beta_1)\{ \\
&\quad \partial_l f_{\tau_{k,n-1}}(x_{k,n-1}) + \cdots + \beta_1^{n-1} \partial_l f_{\tau_{k,0}}(x_{k,0}) \\
&\quad + \partial_l f_{\tau_{k,n-2}}(x_{k,n-2}) + \cdots + \beta_1^{n-2} \partial_l f_{\tau_{k,0}}(x_{k,0}) \\
&\quad + \cdots \\
&\quad \underbrace{+ \partial_l f_{\tau_{k,0}}(x_{k,0})}_{k\text{-th epoch}} \\
&\quad + \beta_1^{n} \partial_l f_{\tau_{k-1,n-1}}(x_{k-1,n-1}) + \cdots + \beta_1^{n+n-1} \partial_l f_{\tau_{k-1,0}}(x_{k-1,0}) \\
&\quad + \beta_1^{n-1} \partial_l f_{\tau_{k-1,n-1}}(x_{k-1,n-1}) + \cdots + \beta_1^{n+n-2} \partial_l f_{\tau_{k-1,0}}(x_{k-1,0}) \\
&\quad + \cdots \\
&\quad \underbrace{+ \beta_1 \partial_l f_{\tau_{k-1,n-1}}(x_{k-1,n-1}) + \cdots + \beta_1^{n} \partial_l f_{\tau_{k-1,0}}(x_{k-1,0})}_{k-1\text{-th epoch}} \\
&\quad + \\
&\quad \vdots \\
&\quad + \\
&\quad + \beta_1^{(k-1)n} \partial_l f_{\tau_{1,n-1}}(x_{1,n-1}) + \cdots + \beta_1^{(k-1)n+n-1} \partial_l f_{\tau_{1,0}}(x_{1,0}) \\
&\quad + \beta_1^{(k-2)n+n-1} \partial_l f_{\tau_{1,n-1}}(x_{1,n-1}) + \cdots + \beta_1^{(k-1)n+n-2} \partial_l f_{\tau_{1,0}}(x_{1,0}) \\
&\quad + \cdots \\
&\quad \underbrace{+ \beta_1^{(k-2)n+1} \partial_l f_{\tau_{1,n-1}}(x_{1,n-1}) + \cdots + \beta_1^{(k-1)n} \partial_l f_{\tau_{1,0}}(x_{1,0})}_{1\text{-th epoch}} \\
&\quad + \beta_1^{(k-1)n+n} \partial_l f_0(x_{1,0}) + \cdots + \beta_1^{(k-1)n+n} \partial_l f_{n-1}(x_{1,0}) \\
&\quad + \beta_1^{(k-1)n+n-1} \partial_l f_0(x_{1,0}) + \cdots + \beta_1^{(k-1)n+n-1} \partial_l f_{n-1}(x_{1,0}) \\
&\quad + \cdots \\
&\quad \underbrace{+ \beta_1^{(k-1)n+1} \partial_l f_0(x_{1,0}) + \cdots + \beta_1^{(k-1)n+1} \partial_l f_{n-1}(x_{1,0}) +}_{0\text{-th epoch}} \\
&\quad \vdots \}
\end{aligned} \tag{38}
$$

Note that in (38), the power of β_1 grows slower in the blue part (when $k = 0$), such a transition will cause trouble in bounding the $\sum_{i=0}^{n-1}(m_{l,k,i} - \partial_l f_i(x_{k,0}))$. Therefore, we need to define an auxillary sequence $M'_{l,k}$ which does not involve such a phase transition. We define $M'_{l,k}$ as follows. (For each 'epoch', we suggest readers to read from bottom to the top.)

$$
\begin{aligned}
M'_{l,k} \quad := \quad & (1-\beta_1)\{ \\
& \partial_l f_{\tau_{k,n-1}}(x_{k,n-1}) + \cdots + \beta_1^{n-1}\partial_l f_{\tau_{k,0}}(x_{k,0}) \\
& +\partial_l f_{\tau_{k,n-2}}(x_{k,n-2}) + \cdots + \beta_1^{n-2}\partial_l f_{\tau_{k,0}}(x_{k,0}) \\
& +\cdots \\
& \underbrace{+\partial_l f_{\tau_{k,0}}(x_{k,0})}_{k\text{-th epoch}} \\
& +\beta_1^{n}\partial_l f_{\tau_{k-1,n-1}}(x_{k-1,n-1}) + \cdots + \beta_1^{n+n-1}\partial_l f_{\tau_{k-1,0}}(x_{k-1,0}) \\
& +\beta_1^{n-1}\partial_l f_{\tau_{k-1,n-1}}(x_{k-1,n-1}) + \cdots + \beta_1^{n+n-2}\partial_l f_{\tau_{k-1,0}}(x_{k-1,0}) \\
& +\cdots \\
& \underbrace{+\beta_1\partial_l f_{\tau_{k-1,n-1}}(x_{k-1,n-1}) + \cdots + \beta_1^{n}\partial_l f_{\tau_{k-1,0}}(x_{k-1,0})}_{k-1\text{-th epoch}} \\
& + \\
& \vdots \\
& + \\
& +\beta_1^{(k-1)n}\partial_l f_{\tau_{1,n-1}}(x_{1,n-1}) + \cdots + \beta_1^{(k-1)n+n-1}\partial_l f_{\tau_{1,0}}(x_{1,0}) \\
& +\beta_1^{(k-2)n+n-1}\partial_l f_{\tau_{1,n-1}}(x_{1,n-1}) + \cdots + \beta_1^{(k-1)n+n-2}\partial_l f_{\tau_{1,0}}(x_{1,0}) \\
& +\cdots \\
& \underbrace{+\beta_1^{(k-2)n+1}\partial_l f_{\tau_{1,n-1}}(x_{1,n-1}) + \cdots + \beta_1^{(k-1)n}\partial_l f_{\tau_{1,0}}(x_{1,0})}_{1\text{-th epoch}} \\
& +\beta_1^{(k-1)n+n}\partial_l f_0(x_{1,0}) + \cdots + \beta_1^{(k-1)n+n+n-1}\partial_l f_{n-1}(x_{1,0}) \\
& +\beta_1^{(k-1)n+n-1}\partial_l f_0(x_{1,0}) + \cdots + \beta_1^{(k-1)n+2(n-1)}\partial_l f_{n-1}(x_{1,0}) \\
& +\cdots \\
& \underbrace{+\beta_1^{(k-1)n+1}\partial_l f_0(x_{1,0}) + \cdots + \beta_1^{(k-1)n+n}\partial_l f_{n-1}(x_{1,0})+}_{0\text{-th epoch}} \\
& \vdots\} \qquad (39)
\end{aligned}
$$

Now we bound $\sum_{i=0}^{n-1}(m_{l,k,i} - \partial_l f_i(x_{k,0}))$ with the help of $M'_{l,k}$. Denoting $F_{l,k} := \sum_{i=0}^{n-1}\partial_l f_i(x_{k,0})$, we have:

$$
\begin{aligned}
\sum_{i=0}^{n-1}(m_{l,k,i} - \partial_l f_i(x_{k,0})) \quad &= \quad \sum_{i=0}^{n-1} m_{l,k,i} - \sum_{i=0}^{n-1}\partial_l f_i(x_{k,0}) \\
&:= \quad M_{l,k} - F_{l,k} \\
&\geq \quad -|M_{l,k} - M'_{l,k}| + M'_{l,k} - F_{l,k} \qquad (40)
\end{aligned}
$$

To proceed, we use the relation $1 = (1-\beta_1)(1+\beta_1+\beta_1^2+\cdots)$ to rewrite $F_{l,k}$.

$$
\begin{aligned}
F_{l,k} := \sum_{i=0}^{n-1} \partial_l f_i(x_{k,0}) \quad = \quad & (1-\beta_1)\{ \\
& \partial_l f_{n-1}(x_{k,0}) + \cdots + \beta_1^{n-1} \partial_l f_{n-1}(x_{k,0}) \\
& + \partial_l f_{n-2}(x_{k,0}) + \cdots + \beta_1^{n-1} \partial_l f_{n-2}(x_{k,0}) \\
& + \cdots \\
& \underbrace{+ \partial_l f_0(x_{k,0}) + \cdots + \beta_1^{n-1} \partial_l f_0(x_{k,0})}_{k\text{-th epoch}} \\
& + \\
& \vdots \\
& + \\
& + \beta_1^{(k-1)n} \partial_l f_{n-1}(x_{k,0}) + \cdots + \beta_1^{(k-1)n+n-1} \partial_l f_{n-1}(x_{k,0}) \\
& + \beta_1^{(k-1)n} \partial_l f_{n-2}(x_{k,0}) + \cdots + \beta_1^{(k-1)n+n-1} \partial_l f_{n-2}(x_{k,0}) \\
& + \cdots \\
& \underbrace{+ \beta_1^{(k-1)n} \partial_l f_0(x_{k,0}) + \cdots + \beta_1^{(k-1)n+n-1} \partial_l f_0(x_{k,0})}_{\text{1-th epoch}} \\
& + \beta_1^{kn} \partial_l f_{n-1}(x_{k,0}) + \cdots + \beta_1^{kn+n-1} \partial_l f_{n-1}(x_{k,0}) \\
& + \beta_1^{kn} \partial_l f_{n-2}(x_{k,0}) + \cdots + \beta_1^{kn+n-1} \partial_l f_{n-2}(x_{k,0}) \\
& + \cdots \\
& \underbrace{+ \beta_1^{kn} \partial_l f_0(x_{k,0}) + \cdots + \beta_1^{kn+n-1} \partial_l f_0(x_{k,0})}_{\text{0-th epoch}} \\
& \vdots \} \qquad (41)
\end{aligned}
$$

For the sake of better presentation, we rewrite $F_{l,k}$ and $M'_{l,k}$ as follows:

$$
F_{l,k} := (F_{l,k})_k + \cdots + (F_{l,k})_1 + (F_{l,k})_0 + (F_{l,k})_{-1} + \cdots ,
$$

$$
M'_{l,k} := \left(M'_{l,k}\right)_k + \cdots + \left(M'_{l,k}\right)_1 + \left(M'_{l,k}\right)_0 + \left(M'_{l,k}\right)_{-1} + \cdots ,
$$

where $(F_{l,k})_j$ contains the summand of $F_{l,k}$ in the j-th epoch, $j = k, k-1, \cdots - \infty$. Similarly for $M'_{l,k}$. Now, we separate $M'_{l,k} - F_{l,k}$ into two parts.

$$
\begin{aligned}
M'_{l,k} - F_{l,k} \quad = \quad & \left[\left(M'_{l,k}\right)_k - (F_{l,k})_k + \cdots \left(M'_{l,k}\right)_1 - (F_{l,k})_1\right] \\
& + \left[\left(M'_{l,k}\right)_0 - (F_{l,k})_0 + \cdots \left(M'_{l,k}\right)_{-\infty} - (F_{l,k})_{-\infty}\right]
\end{aligned}
$$

Now, we rewrite $\mathbb{E}[\sum_{l \text{ large}}(b_1)]$:

$$
\begin{aligned}
& \mathbb{E}\left[\sum_{l \text{ large}} \frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}} \sum_{i=0}^{n-1}(m_{l,k,i} - \partial_l f_i(x_{k,0}))\right] \\
= \quad & \mathbb{E}\left[\sum_{l \text{ large}} \frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}} (M_{l,k} - F_{l,k})\right] \\
\geq \quad & -\mathbb{E}\left[\sum_{l \text{ large}} \frac{|\partial_l f(x_{k,0})|}{\sqrt{v_{k,0}}} |M_{l,k} - M'_{l,k}|\right] + \mathbb{E}\left[\sum_{l \text{ large}} \frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}} (M'_{l,k} - F_{l,k})\right] \\
\geq \quad & \underbrace{-\mathbb{E}\left[\sum_{l \text{ large}} \frac{|\partial_l f(x_{k,0})|}{\sqrt{v_{k,0}}} |M_{l,k} - M'_{l,k}|\right]}_{(1)} \\
& \underbrace{+\mathbb{E}\left[\sum_{l \text{ large}} \frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}} \left(\left(M'_{l,k}\right)_k - (F_{l,k})_k + \cdots + \left(M'_{l,k}\right)_1 - (F_{l,k})_1\right)\right]}_{(2)} \\
& \underbrace{-\mathbb{E}\left[\sum_{l \text{ large}} \frac{|\partial_l f(x_{k,0})|}{\sqrt{v_{k,0}}} \left|\left(M'_{l,k}\right)_0 - (F_{l,k})_0 + \cdots \left(M'_{l,k}\right)_{-\infty} - (F_{l,k})_{-\infty}\right|\right]}_{(3)}
\end{aligned}
\tag{42}
$$

We bound (1), (2) and (3) respectively. We bound (1) and (3) in the following Lemma G.13 and Lemma G.14. Since the difference of $M_{l,k} - M'_{l,k}$ only occurs in the high order terms of the infinite sum sequence, (1) is expected to vanish as k grows. Similarly for (3).

Lemma G.13. *When k is large enough such that: $\beta_1^{(k-1)n} \leq \frac{\beta_1^n}{\sqrt{k-1}}$ and $k \geq 4$, then we have*

$$
\mathbb{E}\left[\sum_{l \text{ large}} \frac{|\partial_l f(x_{k,0})|}{\sqrt{v_{k,0}}} \left|M_{l,k} - M'_{l,k}\right|\right] \leq G_4 \frac{1}{\sqrt{k}}, \tag{43}
$$

where $G_4 = d\sqrt{\frac{2\rho_3^2}{\beta_2^n}}\beta_1^n \sqrt{2} n(n-1) \sum_{i=0}^{n-1} \mathbb{E}\left(|\partial_\alpha f_i(x_{1,0})|\right)$.

Lemma G.14. *When k is large enough such that: $\beta_1^{(k-1)n} \leq \frac{\beta_1^n}{\sqrt{k-1}}$ and $k \geq 4$, then we have*

$$
\mathbb{E}\left[\sum_{l \text{ large}} \frac{|\partial_l f(x_{k,0})|}{\sqrt{v_{k,0}}} \left|\left(M'_{l,k}\right)_0 - (F_{l,k})_0 + \cdots \left(M'_{l,k}\right)_{-\infty} - (F_{l,k})_{-\infty}\right|\right] \leq G_5 \frac{1}{\sqrt{k}}, \tag{44}
$$

where $G_5 = d\sqrt{\frac{2\rho_3^2}{\beta_2^n}} \left(\frac{\beta_1^{2n} 2n^3 \triangle_1 \sqrt{2}}{\sqrt{n}} \frac{1-\beta_1}{(1-\beta_1^n)^2} + n\left(1 - \beta_1^{n-1}\right) \sum_{i=0}^{n-1} \mathbb{E}|\partial_\alpha f_i(x_{1,0})| \frac{(1-\beta_1)\beta_1^n \sqrt{2}}{1-\beta_1^n}\right)$.

We relegate the proof of Lemma G.13 and G.14 to Appendix G.5 and G.6.

Now we bound (2). This part involves the difficulties (i), (ii) and (iii) mentioned in Appendix G.1. We bound (2) in Lemma G.15.

Lemma G.15. *When k is large enough such that: $\beta_1^{(k-1)n} \leq \frac{\beta_1^n}{\sqrt{k-1}}$ and $k \geq 2$, then we have*

$$
\begin{aligned}
& \mathbb{E}\left[\sum_{l \text{ large}} \frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}}\left(\left(M'_{l,k}\right)_k - \left(F_{l,k}\right)_k + \cdots + \left(M'_{l,k}\right)_1 - \left(F_{l,k}\right)_1\right)\right] \\
\geq \quad & -\frac{G_6}{\sqrt{k}} - G_7\delta_1\mathbb{E}\left(\sum_{l=1}^{d}\sum_{i=0}^{n-1}|\partial_l f_i(x_{k,0})|\right).
\end{aligned}
$$

The constant terms G_6 and G_7 are specified in Appendix G.7.

Proof is shown in Appendix G.7.

Combining Lemma G.13, G.14 and G.15 together, we conclude the proof.

$$
\mathbb{E}[\sum_{l \text{ large}}(b_1)] \geq \frac{1}{\sqrt{k}}\left(G_4 + G_5 + G_6\right) - G_7\delta_1\mathbb{E}\left(\sum_{l=1}^{d}\sum_{i=0}^{n-1}|\partial_l f_i(x_{k,0})|\right).
$$

G.5 Proof of Lemma G.13

By definition of $M_{l,k}$ and $M'_{l,k}$ (see (38) and (39)), they only differ when $k \leq 0$. More specifically, We have (For each 'epoch', we suggest readers to read from bottom to the top.)

$$
\begin{aligned}
\frac{M_{l,k} - M'_{l,k}}{1-\beta_1} \quad = \quad & \left(\beta_1^{(k-1)n+n} - \beta_1^{(k-1)n+n+1}\right)\partial_l f_1(x_{1,0}) + \cdots + \left(\beta_1^{(k-1)n+n} - \beta_1^{(k-1)n+n+n-1}\right)\partial_l f_{n-1}(x_{1,0}) \\
& + \cdots \\
& \underbrace{\begin{aligned}
& + \left(\beta_1^{(k-1)n+2} - \beta_1^{(k-1)n+3}\right)\partial_l f_1(x_{1,0}) + \cdots + \left(\beta_1^{(k-1)n+2} - \beta_1^{(k-1)n+n+1}\right)\partial_l f_{n-1}(x_{1,0}) \\
& + \left(\beta_1^{(k-1)n+1} - \beta_1^{(k-1)n+2}\right)\partial_l f_1(x_{1,0}) + \cdots + \left(\beta_1^{(k-1)n+1} - \beta_1^{(k-1)n+n}\right)\partial_l f_{n-1}(x_{1,0})
\end{aligned}}_{\text{0-th epoch}} \\
& + \left(\beta_1^{(k-1)n+n+1} - \beta_1^{kn+n}\right)\partial_l f_0(x_{1,0}) + \cdots + \left(\beta_1^{(k-1)n+n+1} - \beta_1^{kn+n+n-1}\right)\partial_l f_{n-1}(x_{1,0}) \\
& + \cdots \\
& \underbrace{\begin{aligned}
& + \left(\beta_1^{(k-1)n+3} - \beta_1^{kn+2}\right)\partial_l f_0(x_{1,0}) + \cdots + \left(\beta_1^{(k-1)n+3} - \beta_1^{kn+n+1}\right)\partial_l f_{n-1}(x_{1,0}) \\
& + \left(\beta_1^{(k-1)n+2} - \beta_1^{kn+1}\right)\partial_l f_0(x_{1,0}) + \cdots + \left(\beta_1^{(k-1)n+2} - \beta_1^{kn+n}\right)\partial_l f_{n-1}(x_{1,0})
\end{aligned}}_{\text{-1-th epoch}} \\
& + \left(\beta_1^{(k-1)n+n+2} - \beta_1^{(k+1)n+n}\right)\partial_l f_0(x_{1,0}) + \cdots + \left(\beta_1^{(k-1)n+n+2} - \beta_1^{(k+1)n+n+n-1}\right)\partial_l f_{n-1}(x_{1,0}) \\
& + \cdots \\
& \underbrace{\begin{aligned}
& + \left(\beta_1^{(k-1)n+4} - \beta_1^{(k+1)n+2}\right)\partial_l f_0(x_{1,0}) + \cdots + \left(\beta_1^{(k-1)n+4} - \beta_1^{(k+1)n+n+1}\right)\partial_l f_{n-1}(x_{1,0}) \\
& + \left(\beta_1^{(k-1)n+3} - \beta_1^{(k+1)n+1}\right)\partial_l f_0(x_{1,0}) + \cdots + \left(\beta_1^{(k-1)n+3} - \beta_1^{(k+1)n+n}\right)\partial_l f_{n-1}(x_{1,0})
\end{aligned}}_{\text{-2-th epoch}} \\
& + \cdots
\end{aligned} \tag{45}
$$

We start with the 1st column in the '0-th epoch' (from the bottom to the top).

$$
\begin{aligned}
\text{The 1st column in the '0-th epoch'} \quad = \quad & \beta_1^{(k-1)n+1}(1-\beta_1)|\partial_l f_1(x_{1,0})| \\
& + \beta_1^{(k-1)n+2}(1-\beta_1)|\partial_l f_1(x_{1,0})| \\
& + \cdots \\
& + \beta_1^{(k-1)n+n}(1-\beta_1)|\partial_l f_1(x_{1,0})| \\
\leq \quad & \beta_1^{(k-1)n+1} n(1-\beta_1)|\partial_l f_1(x_{1,0})|
\end{aligned} \tag{46}
$$

Similarly, we can bound every column in the in the '0-th epoch', e.g. last ($n-1$-th) column can be bounded as follows.

$$\begin{aligned}
\text{The last column in the '0-th epoch'} &= \beta_1^{(k-1)n+1}(1-\beta_1^{n-1})|\partial_l f_{n-1}(x_{1,0})| \\
&\quad +\beta_1^{(k-1)n+2}(1-\beta_1^{n-1})|\partial_l f_{n-1}(x_{1,0})| \\
&\quad +\cdots \\
&\quad +\beta_1^{(k-1)n+n}(1-\beta_1^{n-1})|\partial_l f_{n-1}(x_{1,0})| \\
&\leq \beta_1^{(k-1)n+1} n(n-1)(1-\beta_1)|\partial_l f_{n-1}(x_{1,0})| \qquad (47)
\end{aligned}$$

Summing up all the columns, we have

$$\begin{aligned}
\text{The '0-th epoch'} &\leq \beta_1^{(k-1)n+1} n(n-1)(1-\beta_1)\sum_{i=1}^{n-1}|\partial_l f_i(x_{1,0})| \\
&\leq \beta_1^{(k-1)n+1} n(n-1)(1-\beta_1)\sum_{i=0}^{n-1}|\partial_l f_i(x_{1,0})| \qquad (48)
\end{aligned}$$

Using the same technique, it can be shown that

$$\text{The '-1-th epoch'} \leq \beta_1^{(k-1)n+2} 2n(n-1)(1-\beta_1)\sum_{i=0}^{n-1}|\partial_l f_i(x_{1,0})|, \qquad (49)$$

$$\text{The '-2-th epoch'} \leq \beta_1^{(k-1)n+3} 3n(n-1)(1-\beta_1)\sum_{i=0}^{n-1}|\partial_l f_i(x_{1,0})| \qquad (50)$$

$$\cdots$$

Plugging all these results in (45), we have

$$\begin{aligned}
\frac{\left|M_{l,k}-M'_{l,k}\right|}{1-\beta_1} &\leq \beta_1^{(k-1)n+1} n(n-1)(1-\beta_1)\sum_{i=0}^{n-1}|\partial_l f_i(x_{1,0})|\left(1+2\beta_1+3\beta_1^2+\cdots\right) \\
&= \beta_1^{(k-1)n+1} n(n-1)\sum_{i=0}^{n-1}|\partial_l f_i(x_{1,0})|\frac{1}{1-\beta_1}. \\
&\leq \beta_1^{(k-1)n} n(n-1)\sum_{i=0}^{n-1}|\partial_l f_i(x_{1,0})|\frac{1}{1-\beta_1}.
\end{aligned}$$

That is to say, when k is large enough such that $\beta_1^{(k-1)n} \leq \frac{\beta_1^n}{\sqrt{k-1}}$ we have

$$\left|M_{l,k}-M'_{l,k}\right| \leq \frac{\beta_1^n}{\sqrt{k-1}} n(n-1)\sum_{i=0}^{n-1}|\partial_l f_i(x_{1,0})| \qquad (51)$$

$$\overset{\text{When } k\geq 2}{\leq} \frac{\beta_1^n\sqrt{2}}{\sqrt{k}} n(n-1)\sum_{i=0}^{n-1}|\partial_l f_i(x_{1,0})|. \qquad (52)$$

Combining with Lemma G.12, the proof is completed.

G.6 Proof of Lemma G.14

We start with $\left(M'_{l,k}\right)_0 - (F_{l,k})_0$.

$$
\begin{aligned}
\frac{\left(M'_{l,k}\right)_0 - (F_{l,k})_0}{1-\beta_1} &= \beta_1^{(k-1)n+n}\partial_l f_0(x_{1,0}) + \cdots + \beta_1^{(k-1)n+n+n-1}\partial_l f_{n-1}(x_{1,0}) \\
&\quad + \beta_1^{(k-1)n+n-1}\partial_l f_0(x_{1,0}) + \cdots + \beta_1^{(k-1)n+2(n-1)}\partial_l f_{n-1}(x_{1,0}) \\
&\quad + \cdots \\
&\quad \underbrace{+\beta_1^{(k-1)n+1}\partial_l f_0(x_{1,0}) + \cdots + \beta_1^{(k-1)n+n}\partial_l f_{n-1}(x_{1,0}) +}_{\left(M'_{l,k}\right)_0} \\
&\quad - \Big\{ \beta_1^{kn}\partial_l f_{n-1}(x_{k,0}) + \cdots + \beta_1^{kn+n-1}\partial_l f_{n-1}(x_{k,0}) \\
&\quad + \beta_1^{kn}\partial_l f_{n-2}(x_{k,0}) + \cdots + \beta_1^{kn+n-1}\partial_l f_{n-2}(x_{k,0}) \\
&\quad + \cdots \\
&\quad \underbrace{+\beta_1^{kn}\partial_l f_0(x_{k,0}) + \cdots + \beta_1^{kn+n-1}\partial_l f_0(x_{k,0})}_{(F_{l,k})_0} \Big\} \\
&:= \sum_{i=0}^{n-1} \delta_i,
\end{aligned}
$$

where

$$
\begin{aligned}
\delta_i &= \sum_{j=1}^{n} \left(\beta_1^{(k-1)n+j}\partial_l f_i(x_{1,0}) - \beta_1^{kn+j-1}\partial_l f_i(x_{k,0}) \right) \\
&= \underbrace{\sum_{j=1}^{n} \left(\beta_1^{(k-1)n+j} - \beta_1^{kn+j-1} \right) \partial_l f_i(x_{1,0})}_{\delta_i^a} + \underbrace{\sum_{j=1}^{n} \beta_1^{kn+j-1} \left(\partial_l f_i(x_{1,0}) - \partial_l f_i(x_{k,0}) \right)}_{\delta_i^b}.
\end{aligned}
$$

We further have

$$
\begin{aligned}
\delta_i^a &= \sum_{j=1}^{n} \beta_1^{(k-1)n+j} \left(1-\beta_1^{n-1}\right) \partial_l f_i(x_{1,0}) \\
&\le n\beta_1^{(k-1)n+1} \left(1-\beta_1^{n-1}\right) |\partial_l f_i(x_{1,0})|
\end{aligned} \tag{53}
$$

$$
\begin{aligned}
\delta_i^b &\le \sum_{j=1}^{n} \beta_1^{kn+j-1} \frac{\triangle_1 (k-1)2n}{\sqrt{n(k-1)-1}} \\
&\le \beta_1^{kn} \frac{\triangle_1 (k-1)2n^2}{\sqrt{n(k-1)-1}}.
\end{aligned} \tag{54}
$$

Therefore,

$$
\begin{aligned}
\frac{\left|\left(M'_{l,k}\right)_0 - (F_{l,k})_0\right|}{1-\beta_1} &= \left| \sum_{i=0}^{n-1} \delta_i \right| \\
&\le \sum_{i=0}^{n-1} \left(n\beta_1^{(k-1)n+1} \left(1-\beta_1^{n-1}\right) |\partial_l f_i(x_{1,0})| + \beta_1^{kn} \frac{\triangle_1 (k-1)2n^2}{\sqrt{n(k-1)-1}} \right) \\
&= \beta_1^{kn} \frac{\triangle_1 (k-1)2n^3}{\sqrt{n(k-1)-1}} + n\beta_1^{(k-1)n+1} \left(1-\beta_1^{n-1}\right) \sum_{i=0}^{n-1} |\partial_l f_i(x_{1,0})|
\end{aligned} \tag{55}
$$

Using the same calculation, we can get the following result:

$$\frac{\left|\left(M'_{l,k}\right)_{-1} - \left(F_{l,k}\right)_{-1}\right|}{1-\beta_1} \quad \leq \quad \beta_1^{(k+1)n}\frac{\triangle_1 k 2n^3}{\sqrt{n(k-1)-1}} + n\beta_1^{kn+1}\left(1-\beta_1^{n-1}\right)\sum_{i=0}^{n-1}|\partial_l f_i(x_{1,0})|.$$

Repeat this calculation, we have

$$\begin{aligned}
& \frac{1}{1-\beta_1}\left[\left(M'_{l,k}\right)_0 - \left(F_{l,k}\right)_0 + \cdots \left(M'_{l,k}\right)_{-\infty} - \left(F_{l,k}\right)_{-\infty}\right] \\
\leq \quad & \frac{\beta_1^{2n}2n^3\triangle_1}{\sqrt{n(k-1)-1}}\left(\beta_1^{(k-2)n}(k-1) + \beta_1^{(k-1)n}k + \cdots\right) \\
& + n\left(1-\beta_1^{n-1}\right)\sum_{i=0}^{n-1}|\partial_l f_i(x_{1,0})|\left(\beta_1^{(k-1)n+1} + \beta_1^{kn+1} + \cdots\right) \\
\leq \quad & \frac{\beta_1^{2n}2n^3\triangle_1}{\sqrt{n(k-1)-1}}\left(1 + \beta_1^n 2 + \cdots\right) \\
& + n\left(1-\beta_1^{n-1}\right)\sum_{i=0}^{n-1}|\partial_l f_i(x_{1,0})|\left(\beta_1^{(k-1)n+1} + \beta_1^{kn+1} + \cdots\right) \\
\overset{\text{Lemma F.2}}{\leq} \quad & \frac{\beta_1^{2n}2n^3\triangle_1}{\sqrt{n(k-1)-1}}\frac{1}{(1-\beta_1^n)^2} \\
& + n\left(1-\beta_1^{n-1}\right)\sum_{i=0}^{n-1}|\partial_l f_i(x_{1,0})|\frac{\beta_1^{(k-1)n}}{1-\beta_1^n}
\end{aligned}$$

When $\beta_1^{(k-1)n} \leq \frac{\beta_1^n}{\sqrt{k-1}}$, we further have:

$$\begin{aligned}
& \frac{1}{1-\beta_1}\left[\left(M'_{l,k}\right)_0 - \left(F_{l,k}\right)_0 + \cdots \left(M'_{l,k}\right)_{-\infty} - \left(F_{l,k}\right)_{-\infty}\right] \\
\leq \quad & \frac{\beta_1^{2n}2n^3\triangle_1}{\sqrt{n(k-1)-1}}\frac{1}{(1-\beta_1^n)^2} \\
& + n\left(1-\beta_1^{n-1}\right)\sum_{i=0}^{n-1}|\partial_l f_i(x_{1,0})|\frac{\beta_1^n}{1-\beta_1^n}\frac{1}{\sqrt{k-1}} \\
\overset{\frac{1}{\sqrt{k-1}}\leq\sqrt{\frac{2}{k}}}{\leq} \quad & \frac{\beta_1^{2n}2n^3\triangle_1}{\sqrt{n(k-1)-1}}\frac{1}{(1-\beta_1^n)^2} \\
& + n\left(1-\beta_1^{n-1}\right)\sum_{i=0}^{n-1}|\partial_l f_i(x_{1,0})|\frac{\beta_1^n}{1-\beta_1^n}\sqrt{\frac{2}{k}} \\
\overset{(*)}{\leq} \quad & \frac{\beta_1^{2n}2n^3\triangle_1\sqrt{2}}{\sqrt{nk}}\frac{1}{(1-\beta_1^n)^2} \\
& + n\left(1-\beta_1^{n-1}\right)\sum_{i=0}^{n-1}|\partial_l f_i(x_{1,0})|\frac{\beta_1^n}{1-\beta_1^n}\sqrt{\frac{2}{k}},
\end{aligned}$$

where $(*)$: $\frac{1}{\sqrt{n(k-1)-1}} \leq \sqrt{\frac{2}{nk}}$ when $k \geq 4$. Therefore, we have

$$\left|\left(M'_{l,k}\right)_0 - \left(F_{l,k}\right)_0 + \cdots \left(M'_{l,k}\right)_{-\infty} - \left(F_{l,k}\right)_{-\infty}\right| \leq \frac{1}{\sqrt{k}}\tilde{G}_5$$

,

$\tilde{G}_5 = \frac{\beta_1^{2n}2n^3\triangle_1\sqrt{2}}{\sqrt{n}}\frac{1-\beta_1}{(1-\beta_1^n)^2} + n\left(1-\beta_1^{n-1}\right)\sum_{i=0}^{n-1}|\partial_l f_i(x_{1,0})|\frac{(1-\beta_1)\beta_1^n\sqrt{2}}{1-\beta_1^n}$.

Combining with Lemma G.12, the proof is completed with constant G_5 defined in Lemma G.14.

G.7 Proof of Lemma G.15

In this section, we derive a lower bound for $\mathbb{E}\left[\sum_{l \text{ large}} \frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}}\left(\left(M'_{l,k}\right)_k - (F_{l,k})_k + \cdots + \left(M'_{l,k}\right)_1 - (F_{l,k})_1\right)\right]$. We will use the ideas mentioned in Appendix G.1 (i.e., Step 1,2 and 3). We first rewrite "l large" into indicator function as follows:

$$\mathbb{E}\left[\sum_{l \text{ large}} \frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}}\left(\left(M'_{l,k}\right)_k - (F_{l,k})_k + \cdots + \left(M'_{l,k}\right)_1 - (F_{l,k})_1\right)\right] = \mathbb{E}\left[\sum_{l=1}^{d} \mathbb{I}_k \frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}}\left(\left(M'_{l,k}\right)_k - (F_{l,k})_k + \cdots + \left(M'_{l,k}\right)_1 - (F_{l,k})_1\right)\right],$$

where $\mathbb{I}_k := \mathbb{I}\left(\max_i |\partial_l f_i(x_{k,0})| \geq Q_k := \triangle_1 \frac{n\sqrt{n}}{\sqrt{k}} \frac{32\sqrt{2}}{(1-\beta_2)^n \beta_2^n}\right)$. We also define $\mathbb{I}_{k-j} := \mathbb{I}\left(\max_i |\partial_l f_i(x_{k-j,0})| \geq \sum_{p=k-j}^{k} Q_p\right)$, it will be used later.

We take the conditional expectation over the history before $x_{k,0}$. We first focus on $\mathbb{E}_k\left[\sum_{l=1}^{d} \mathbb{I}_k \frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}}\left(\left(M'_{l,k}\right)_k - (F_{l,k})_k\right)\right]$ and we delegate the history part for later analysis.

For each possible $\left(M'_{l,k}\right)_k - (F_{l,k})_k$, we first convert all $x_{k,i}$ into $x_{k,0}$ using Lemma F.3. Since $\left(M'_{l,k}\right)_k$ contains $(n-1) + \cdots + 1 = \frac{n(n-1)}{2}$ terms of $x_{k,i}$ (with $i \neq 0$), we have

$$\begin{aligned}
\left(M'_{l,k}\right)_k \overset{\text{Lemma F.3}}{\geq} & -\frac{(1-\beta_1)n^2(n-1)\triangle_{nk}}{2} + (1-\beta_1)\{ \\
& \partial_l f_{\tau_{k,n-1}}(x_{k,0}) + \cdots + \beta_1^{n-1}\partial_l f_{\tau_{k,0}}(x_{k,0}) \\
& + \partial_l f_{\tau_{k,n-2}}(x_{k,0}) + \cdots + \beta_1^{n-2}\partial_l f_{\tau_{k,0}}(x_{k,0}) \\
& + \cdots \\
& \underbrace{+\partial_l f_{\tau_{k,0}}(x_{k,0})}_{k\text{-th epoch}}\} \\
:= & -\frac{(1-\beta_1)n^2(n-1)\triangle_{nk}}{2} + \mathscr{M}_{l,k}.
\end{aligned}$$

Using the same strategy for analyzing the color-ball toy example (Figure 15), we have

$$\begin{aligned}
\frac{\mathbb{E}_k\left[\mathscr{M}_{l,k} - (F_{l,k})_k\right]}{1-\beta_1} &= \frac{1}{n!}\left(-(n-1)!\beta_1 - 2(n-1)!\beta_1^2 - \cdots - (n-1)(n-1)!\beta_1^{n-1}\right)\sum_{i=0}^{n-1}\partial_l f_i(x_{k,0}) \\
&= \left(-\frac{1}{n}\beta_1 - \frac{2}{n}\beta_1^2 - \cdots - \frac{n-1}{n}\beta_1^{n-1}\right)\sum_{i=0}^{n-1}\partial_l f_i(x_{k,0})
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_k\left[\sum_{l=1}^{d} \mathbb{I}_k \frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}}\left(\left(M'_{l,k}\right)_k - (F_{l,k})_k\right)\right] &\geq -d\sqrt{\frac{2\rho_3^2}{\beta_2^n}}\frac{(1-\beta_1)n^2(n-1)\triangle_{nk}}{2} + \mathbb{E}_k\left[\sum_{l=1}^{d} \mathbb{I}_k \frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}}\left(\mathscr{M}_{l,k} - (F_{l,k})_k\right)\right] \\
&= -d\sqrt{\frac{2\rho_3^2}{\beta_2^n}}\frac{(1-\beta_1)n^2(n-1)\triangle_{nk}}{2} \\
&\quad + \sum_{l=1}^{d} \mathbb{I}_k(1-\beta_1)\left(-\frac{1}{n}\beta_1 - \frac{2}{n}\beta_1^2 - \cdots - \frac{n-1}{n}\beta_1^{n-1}\right)\sum_{i=0}^{n-1}\partial_l f_i(x_{k,0}) \\
&:= -d\sqrt{\frac{2\rho_3^2}{\beta_2^n}}\frac{(1-\beta_1)n^2(n-1)\triangle_{nk}}{2} + \sum_{l=1}^{d} \mathbb{I}_k \frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}}(1-\beta_1)J_1\sum_{i=0}^{n-1}\partial_l f_i(x_{k,0})
\end{aligned}$$

We denote $J_1 := \left(-\frac{1}{n}\beta_1 - \frac{2}{n}\beta_1^2 - \cdots - \frac{n-1}{n}\beta_1^{n-1}\right)$, it will be used repeatedly in the following derivation.

Now we move one step further exert $\mathbb{E}_{k-1}(\cdot)$. In particular, we need to calculate

$$
\begin{aligned}
& \mathbb{E}_{k-1}\left\{\mathbb{E}_k\left[\sum_{l=1}^{d}\mathbb{I}_k\frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}}\left(\left(M'_{l,k}\right)_k-\left(F_{l,k}\right)_k\right)+\sum_{l=1}^{d}\mathbb{I}_k\frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}}\left(\left(M'_{l,k}\right)_{k-1}-\left(F_{l,k}\right)_{k-1}\right)\right]\right\} \\
\geq\ & \mathbb{E}_{k-1}\left\{\underbrace{\sum_{l=1}^{d}\mathbb{I}_k\frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}}(1-\beta_1)J_1\sum_{i=0}^{n-1}\partial_l f_i(x_{k,0})}_{(a)}+\underbrace{\sum_{l=1}^{d}\mathbb{I}_k\frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}}\left(\left(M'_{l,k}\right)_{k-1}-\left(F_{l,k}\right)_{k-1}\right)}_{(b)}\right\} \\
& -d\sqrt{\frac{2\rho_3^2}{\beta_2^n}}\frac{(1-\beta_1)n^2(n-1)\triangle_{nk}}{2}.
\end{aligned}
\tag{56}
$$

The blue term is the residue from the k-th epoch. The red term is the main component in the $(k-1)$-th epoch. Before calculating $\mathbb{E}_{k-1}(\cdot)$, we need to convert all the variable x into $x_{k-1,0}$ using Lipschitz property. First of all, we work on (a).

$$
\begin{aligned}
(a) \quad &= \sum_{l=1}^{d}\mathbb{I}_k\frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}}(1-\beta_1)J_1\sum_{i=0}^{n-1}\partial_l f_i(x_{k,0}) \\
&\overset{\text{Lemma G.4}}{=} \sum_{l=1}^{d}\mathbb{I}_{k,k-1}\frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}}(1-\beta_1)J_1\sum_{i=0}^{n-1}\partial_l f_i(x_{k,0}) \\
&\quad +\sum_{l=1}^{d}\tilde{\mathbb{I}}_{k,k-1}\frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}}(1-\beta_1)J_1\sum_{i=0}^{n-1}\partial_l f_i(x_{k,0}) \\
&\overset{\text{Lemma G.12}}{\geq} \sum_{l=1}^{d}\mathbb{I}_{k,k-1}\frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}}(1-\beta_1)J_1\sum_{i=0}^{n-1}\partial_l f_i(x_{k,0})-(1-\beta_1)d\sqrt{\frac{2\rho_3^2}{\beta_2^n}}|J_1|nQ_k \\
&\overset{\text{Lemma G.12 and F.3}}{\geq} \sum_{l=1}^{d}\mathbb{I}_{k,k-1}\frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}}(1-\beta_1)J_1\sum_{i=0}^{n-1}\partial_l f_i(x_{k-1,0}) \\
&\quad -(1-\beta_1)d\sqrt{\frac{2\rho_3^2}{\beta_2^n}}|J_1|n^2\triangle_{n(k-1)}-(1-\beta_1)d\sqrt{\frac{2\rho_3^2}{\beta_2^n}}|J_1|nQ_k \\
&\geq \sum_{l=1}^{d}\mathbb{I}_{k,k-1}\frac{\partial_l f(x_{k-1,0})}{\sqrt{v_{k-1,0}}}(1-\beta_1)J_1\sum_{i=0}^{n-1}\partial_l f_i(x_{k-1,0}) \\
&\quad -\sum_{l=1}^{d}\mathbb{I}_{k,k-1}\left|\frac{\partial_l f(x_{k-1,0})}{\sqrt{v_{k-1,0}}}-\frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}}\right|(1-\beta_1)|J_1|\left|\sum_{i=0}^{n-1}\partial_l f_i(x_{k-1,0})\right| \\
&\quad -(1-\beta_1)d\sqrt{\frac{2\rho_3^2}{\beta_2^n}}|J_1|n^2\triangle_{n(k-1)}-(1-\beta_1)d\sqrt{\frac{2\rho_3^2}{\beta_2^n}}|J_1|nQ_k \\
&\overset{\text{Lemma G.3}}{\geq} \sum_{l=1}^{d}\mathbb{I}_{k,k-1}\frac{\partial_l f(x_{k-1,0})}{\sqrt{v_{k-1,0}}}(1-\beta_1)J_1\sum_{i=0}^{n-1}\partial_l f_i(x_{k-1,0}) \\
&\quad -\sum_{l=1}^{d}\mathbb{I}_{k,k-1}(1-\beta_1)|J_1|\left|\sum_{i=0}^{n-1}\partial_l f_i(x_{k-1,0})\right|\left(\frac{1}{1-\frac{1}{\sqrt{\beta_2^n}}}\frac{n\triangle_{n(k-1)}}{\sqrt{v_{k-1,0}}}+\sqrt{\frac{2\rho_3^2}{\beta_2^n}}\frac{1}{\left(1-\frac{1-\beta_2}{2}\left(-1+\frac{4\rho_2 n}{\beta_2^n}\right)\right)}\delta_1\right) \\
&\quad -(1-\beta_1)d\sqrt{\frac{2\rho_3^2}{\beta_2^n}}|J_1|n^2\triangle_{n(k-1)}-(1-\beta_1)d\sqrt{\frac{2\rho_3^2}{\beta_2^n}}|J_1|nQ_k,
\end{aligned}
$$

where $\mathbb{I}_{k,k-1} := \mathbb{I}\left(\max_i|\partial_l f_i(x_{k,0})|\geq Q_k \text{ and } \max_i|\partial_l f_i(x_{k-j,0})|\geq\sum_{p=k-j}^{k}Q_p\right)$ and $\tilde{\mathbb{I}}_{k,k-1} := \mathbb{I}\left(\max_i|\partial_l f_i(x_{k,0})|\geq Q_k \text{ and } \max_i|\partial_l f_i(x_{k-j,0})|\leq\sum_{p=k-j}^{k}Q_p\right)$. By Lemma G.4, we know $\mathbb{I}_{k,k-1}=\mathbb{I}_{k-1}$, so we have:

$$
\begin{aligned}
(a) \quad \overset{\text{Lemma G.12}}{\geq} \quad & \sum_{l=1}^{d} \mathbb{I}_{k-1} \frac{\partial_l f(x_{k-1,0})}{\sqrt{v_{k-1,0}}} (1-\beta_1) J_1 \sum_{i=0}^{n-1} \partial_l f_i(x_{k-1,0}) \\
& -d(1-\beta_1)|J_1| \frac{1}{1-\frac{1}{\sqrt{\beta_2^n}}} n \triangle_{n(k-1)} \sqrt{\frac{n\beta_2^n}{2}} \\
& -\sum_{l=1}^{d} (1-\beta_1)|J_1| \left| \sum_{i=0}^{n-1} \partial_l f_i(x_{k-1,0}) \right| \left(\sqrt{\frac{2\rho_3^2}{\beta_2^n}} \frac{1}{\left(1-\frac{1-\beta_2}{2}\left(-1+\frac{4\rho_2 n}{\beta_2^n}\right)\right)} \delta_1 \right) \\
& -(1-\beta_1) d \sqrt{\frac{2\rho_3^2}{\beta_2^n}} |J_1| n^2 \triangle_{n(k-1)} - (1-\beta_1) d \sqrt{\frac{2\rho_3^2}{\beta_2^n}} |J_1| n Q_k
\end{aligned}
$$

$$
\begin{aligned}
\overset{\text{Lemma F.3}}{\geq} \quad & \sum_{l=1}^{d} \mathbb{I}_{k-1} \frac{\partial_l f(x_{k-1,0})}{\sqrt{v_{k-1,0}}} (1-\beta_1) J_1 \sum_{i=0}^{n-1} \partial_l f_i(x_{k-1,0}) \\
& -d(1-\beta_1)|J_1| \frac{1}{1-\frac{1}{\sqrt{\beta_2^n}}} n \triangle_{n(k-1)} \sqrt{\frac{n\beta_2^n}{2}} \\
& -\sum_{l=1}^{d} (1-\beta_1)|J_1| \left(\left| \sum_{i=0}^{n-1} \partial_l f_i(x_{k,0}) \right| + n^2 \triangle_{n(k-1)} \right) \left(\sqrt{\frac{2\rho_3^2}{\beta_2^n}} \frac{1}{\left(1-\frac{1-\beta_2}{2}\left(-1+\frac{4\rho_2 n}{\beta_2^n}\right)\right)} \delta_1 \right) \\
& -(1-\beta_1) d \sqrt{\frac{2\rho_3^2}{\beta_2^n}} |J_1| n^2 \triangle_{n(k-1)} - (1-\beta_1) d \sqrt{\frac{2\rho_3^2}{\beta_2^n}} |J_1| n Q_k.
\end{aligned}
$$

The blue term will be handled using color-ball method when taking conditional expectation $\mathbb{E}_{k-1}(\cdot)$. Now we derive a lower bound for (b). Similarly as before, we rewrite $\left(M'_{l,k}\right)_{k-1}$ and $(F_{l,k})_{k-1}$ as follows.

$$
\left(M'_{l,k}\right)_{k-1} \overset{\text{Lemma F.3}}{\geq} \mathscr{M}_{l,k-1} - (1-\beta_1)\beta_1 n^3 \triangle_{n(k-1)};
$$

$$
(F_{l,k})_{k-1} \overset{\text{Lemma F.3}}{\geq} (F_{l,k-1})_{k-1} - (1-\beta_1)\beta_1^n n^3 \triangle_{n(k-1)};
$$

where

$$
\begin{aligned}
\mathscr{M}_{l,k-1} \quad := \quad & (1-\beta_1)\{ \\
& +\beta_1^n \partial_l f_{\tau_{k-1,n-1}}(x_{k-1,0}) + \cdots + \beta_1^{n+n-1} \partial_l f_{\tau_{k-1,0}}(x_{k-1,0}) \\
& +\beta_1^{n-1} \partial_l f_{\tau_{k-1,n-1}}(x_{k-1,0}) + \cdots + \beta_1^{n+n-2} \partial_l f_{\tau_{k-1,0}}(x_{k-1,0}) \\
& + \cdots \\
& \underbrace{+\beta_1 \partial_l f_{\tau_{k-1,n-1}}(x_{k-1,0}) + \cdots + \beta_1^n \partial_l f_{\tau_{k-1,0}}(x_{k-1,0})}_{k-1\text{-th epoch}}\}
\end{aligned}
$$

$$
\begin{aligned}
(F_{l,k})_{k-1} := \sum_{i=0}^{n-1} \partial_l f_i(x_{k,0}) \quad = \quad & (1-\beta_1)\beta_1^n \{ \\
& \partial_l f_{n-1}(x_{k-1,0}) + \cdots + \beta_1^{n-1} \partial_l f_{n-1}(x_{k-1,0}) \\
& +\partial_l f_{n-2}(x_{k-1,0}) + \cdots + \beta_1^{n-1} \partial_l f_{n-2}(x_{k-1,0}) \\
& + \cdots \\
& \underbrace{+\partial_l f_0(x_{k-1,0}) + \cdots + \beta_1^{n-1} \partial_l f_0(x_{k-1,0})}_{(k-1)\text{-th epoch}}\}
\end{aligned}
$$

We now calculate (b). Using the same idea as in (a), we have

$$
\begin{aligned}
(b) \quad &= \sum_{l=1}^{d} \mathbb{I}_k \frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}} \left(\left(M'_{l,k}\right)_{k-1} - \left(F_{l,k}\right)_{k-1} \right) \\
&\overset{\text{Lemma G.12 and F.3}}{\geq} \sum_{l=1}^{d} \mathbb{I}_k \frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}} \left(\mathscr{M}_{l,k-1} - \left(F_{l,k-1}\right)_{k-1} \right) - d\sqrt{\frac{2\rho_3^2}{\beta_2^n}}(1-\beta_1)\left(\beta_1+\beta_1^n\right) n^3 \triangle_{n(k-1)} \\
&\overset{\text{Lemma G.4}}{=} \sum_{l=1}^{d} \mathbb{I}_{k,k-1} \frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}} \left(\mathscr{M}_{l,k-1} - \left(F_{l,k-1}\right)_{k-1} \right) + \sum_{l=1}^{d} \tilde{\mathbb{I}}_{k,k-1} \frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}} \left(\mathscr{M}_{l,k-1} - \left(F_{l,k-1}\right)_{k-1} \right) \\
&\quad -d\sqrt{\frac{2\rho_3^2}{\beta_2^n}}(1-\beta_1)\left(\beta_1+\beta_1^n\right) n^3 \triangle_{n(k-1)} \\
&\overset{\text{Lemma G.12}}{\geq} \sum_{l=1}^{d} \mathbb{I}_{k,k-1} \frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}} \left(\mathscr{M}_{l,k-1} - \left(F_{l,k-1}\right)_{k-1} \right) - \sum_{l=1}^{d} \tilde{\mathbb{I}}_{k,k-1} \sqrt{\frac{2\rho_3^2}{\beta_2^n}} \left| \mathscr{M}_{l,k-1} - \left(F_{l,k-1}\right)_{k-1} \right| \\
&\quad -d\sqrt{\frac{2\rho_3^2}{\beta_2^n}}(1-\beta_1)\left(\beta_1+\beta_1^n\right) n^3 \triangle_{n(k-1)} \\
&\overset{\text{Def of } \tilde{\mathbb{I}}_{k,k-1}}{\geq} \sum_{l=1}^{d} \mathbb{I}_{k,k-1} \frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}} \left(\mathscr{M}_{l,k-1} - \left(F_{l,k-1}\right)_{k-1} \right) - d\sqrt{\frac{2\rho_3^2}{\beta_2^n}}(1-\beta_1)(\beta_1+\beta_1^n) n^2 \left(Q_k + Q_{k-1}\right) \\
&\quad -d\sqrt{\frac{2\rho_3^2}{\beta_2^n}}(1-\beta_1)\left(\beta_1+\beta_1^n\right) n^3 \triangle_{n(k-1)} \\
&\overset{\text{Lemma G.3}}{\geq} \sum_{l=1}^{d} \mathbb{I}_{k,k-1} \frac{\partial_l f(x_{k-1,0})}{\sqrt{v_{k-1,0}}} \left(\mathscr{M}_{l,k-1} - \left(F_{l,k-1}\right)_{k-1} \right) \\
&\quad - \sum_{l=1}^{d} \left(\frac{1}{1-\frac{1}{\sqrt{\beta_2^n}}} \frac{n\triangle_{n(k-1)}}{\sqrt{v_{k-1,0}}} + \sqrt{\frac{2\rho_3^2}{\beta_2^n}} \frac{1}{\left(1-\frac{1-\beta_2}{2}\left(-1+\frac{4\rho_2 n}{\beta_2^n}\right)\right)} \delta_1 \right) \left| \mathscr{M}_{l,k-1} - \left(F_{l,k-1}\right)_{k-1} \right| \\
&\quad -d\sqrt{\frac{2\rho_3^2}{\beta_2^n}}(1-\beta_1)(\beta_1+\beta_1^n) n^2 \left(Q_k + Q_{k-1}\right) - d\sqrt{\frac{2\rho_3^2}{\beta_2^n}}(1-\beta_1)\left(\beta_1+\beta_1^n\right) n^3 \triangle_{n(k-1)}
\end{aligned}
$$

To proceed, we derive an upper bound for $\left| \mathscr{M}_{l,k-1} - \left(F_{l,k-1}\right)_{k-1} \right|$.

$$
\begin{aligned}
\left| \mathscr{M}_{l,k-1} - \left(F_{l,k-1}\right)_{k-1} \right| &\leq \left| \mathscr{M}_{l,k-1} \right| + \left| \left(F_{l,k-1}\right)_{k-1} \right| \\
&\leq (1-\beta_1)\beta_1 n \sum_{i=0}^{n-1} \left| \partial_l f_i(x_{k-1,0}) \right| + (1-\beta_1)\beta_1^n n \sum_{i=0}^{n-1} \left| \partial_l f_i(x_{k-1,0}) \right|
\end{aligned}
$$

Therefore, we have:

$$
\begin{aligned}
(b) \quad \geq \quad & \sum_{l=1}^{d} \mathbb{I}_{k,k-1} \frac{\partial_l f(x_{k-1,0})}{\sqrt{v_{k-1,0}}} \left(\mathscr{M}_{l,k-1} - \left(F_{l,k-1}\right)_{k-1} \right) \\
& - \sum_{l=1}^{d} \left(\frac{1}{1-\frac{1}{\sqrt{\beta_2^n}}} \frac{n\triangle_{n(k-1)}}{\sqrt{v_{k-1,0}}} + \sqrt{\frac{2\rho_3^2}{\beta_2^n}} \frac{1}{\left(1-\frac{1-\beta_2}{2}\left(-1+\frac{4\rho_2 n}{\beta_2^n}\right)\right)} \delta_1 \right) (1-\beta_1)(\beta_1+\beta_1^n) n \sum_{i=0}^{n-1} \left| \partial_l f_i(x_{k-1,0}) \right| \\
& -d\sqrt{\frac{2\rho_3^2}{\beta_2^n}}(1-\beta_1)(\beta_1+\beta_1^n) n^2 \left(Q_k + Q_{k-1}\right) - d\sqrt{\frac{2\rho_3^2}{\beta_2^n}}(1-\beta_1)\left(\beta_1+\beta_1^n\right) n^3 \triangle_{n(k-1)}
\end{aligned}
$$

$$
\begin{aligned}
\overset{\text{Lemma G.12 and G.4}}{\geq} \quad & \sum_{l=1}^{d} \mathbb{I}_{k-1} \frac{\partial_l f(x_{k-1,0})}{\sqrt{v_{k-1,0}}} \left(\mathscr{M}_{l,k-1} - (F_{l,k-1})_{k-1} \right) \\
& -d \left(\frac{1}{1-\frac{1}{\sqrt{\beta_2^n}}} \sqrt{\frac{n\beta_2^n}{2}} \right) (1-\beta_1)(\beta_1+\beta_1^n) n^2 \triangle_{n(k-1)} \\
& -\sum_{l=1}^{d} \sqrt{\frac{2\rho_3^2}{\beta_2^n}} \frac{1}{\left(1-\frac{1-\beta_2}{2}\left(-1+\frac{4\rho_2 n}{\beta_2^n}\right)\right)} \delta_1 (1-\beta_1)(\beta_1+\beta_1^n) n \sum_{i=0}^{n-1} \left(|\partial_l f_i(x_{k-1,0})|\right) \\
& -d\sqrt{\frac{2\rho_3^2}{\beta_2^n}} (1-\beta_1)(\beta_1+\beta_1^n) n^2 \left(Q_k + Q_{k-1}\right) - d\sqrt{\frac{2\rho_3^2}{\beta_2^n}} (1-\beta_1)\left(\beta_1+\beta_1^n\right) n^3 \triangle_{n(k-1)}
\end{aligned}
$$

The red term will be handled using color-ball method when taking conditional expectation $Ex_{k-1}(\cdot)$. Now we have derived lower bounds for both (a) and (b). Combining together, we have:

$$
\begin{aligned}
& \mathbb{E}_{k-1} \left\{ \mathbb{E}_k \left[\sum_{l=1}^{d} \mathbb{I}_k \frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}} \left(\left(M'_{l,k}\right)_k - \left(F_{l,k}\right)_k \right) + \sum_{l=1}^{d} \mathbb{I}_k \frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}} \left(\left(M'_{l,k}\right)_{k-1} - \left(F_{l,k}\right)_{k-1} \right) \right] \right\} \\
\overset{(56)}{\geq} \quad & \mathbb{E}_{k-1} \left\{ \underbrace{\sum_{l=1}^{d} \mathbb{I}_k \frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}} (1-\beta_1) J_1 \sum_{i=0}^{n-1} \partial_l f_i(x_{k,0})}_{(a)} + \underbrace{\sum_{l=1}^{d} \mathbb{I}_k \frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}} \left(\left(M'_{l,k}\right)_{k-1} - \left(F_{l,k}\right)_{k-1} \right)}_{(b)} \right\} \\
& -d\sqrt{\frac{2\rho_3^2}{\beta_2^n}} \frac{(1-\beta_1) n^2 (n-1) \triangle_{nk}}{2} \\
\geq \quad & \mathbb{E}_{k-1} \left\{ \sum_{l=1}^{d} \mathbb{I}_{k-1} \frac{\partial_l f(x_{k-1,0})}{\sqrt{v_{k-1,0}}} (1-\beta_1) J_1 \sum_{i=0}^{n-1} \partial_l f_i(x_{k-1,0}) + \sum_{l=1}^{d} \mathbb{I}_{k-1} \frac{\partial_l f(x_{k-1,0})}{\sqrt{v_{k-1,0}}} \left(\mathscr{M}_{l,k-1} - (F_{l,k-1})_{k-1} \right) \right\} \\
& -d(1-\beta_1)|J_1| \frac{1}{1-\frac{1}{\sqrt{\beta_2^n}}} n \triangle_{n(k-1)} \sqrt{\frac{n\beta_2^n}{2}} \\
& -\sum_{l=1}^{d} (1-\beta_1)|J_1| \left(\left| \sum_{i=0}^{n-1} \partial_l f_i(x_{k,0}) \right| + n^2 \triangle_{n(k-1)} \right) \left(\sqrt{\frac{2\rho_3^2}{\beta_2^n}} \frac{1}{\left(1-\frac{1-\beta_2}{2}\left(-1+\frac{4\rho_2 n}{\beta_2^n}\right)\right)} \delta_1 \right) \\
& -(1-\beta_1) d \sqrt{\frac{2\rho_3^2}{\beta_2^n}} |J_1| n^2 \triangle_{n(k-1)} - (1-\beta_1) d \sqrt{\frac{2\rho_3^2}{\beta_2^n}} |J_1| n Q_k \\
& -d \left(\frac{1}{1-\frac{1}{\sqrt{\beta_2^n}}} \sqrt{\frac{n\beta_2^n}{2}} \right) (1-\beta_1)(\beta_1+\beta_1^n) n^2 \triangle_{n(k-1)} \\
& -\sum_{l=1}^{d} \sqrt{\frac{2\rho_3^2}{\beta_2^n}} \frac{1}{\left(1-\frac{1-\beta_2}{2}\left(-1+\frac{4\rho_2 n}{\beta_2^n}\right)\right)} \delta_1 (1-\beta_1)(\beta_1+\beta_1^n) n \left(\sum_{i=0}^{n-1} |\partial_l f_i(x_{k,0})| + n^2 \triangle_{n(k-1)} \right) \\
& -d\sqrt{\frac{2\rho_3^2}{\beta_2^n}} (1-\beta_1)(\beta_1+\beta_1^n) n^2 \left(Q_k + Q_{k-1}\right) - d\sqrt{\frac{2\rho_3^2}{\beta_2^n}} (1-\beta_1)\left(\beta_1+\beta_1^n\right) n^3 \triangle_{n(k-1)} \\
& -d\sqrt{\frac{2\rho_3^2}{\beta_2^n}} \frac{(1-\beta_1) n^2 (n-1) \triangle_{nk}}{2}
\end{aligned}
$$

For the first term in the above inequality, we can calculate it using the idea in the color-ball toy example:

$$
\begin{aligned}
& \mathbb{E}_{k-1}\left\{\sum_{l=1}^{d}\mathbb{I}_{k-1}\frac{\partial_l f(x_{k-1,0})}{\sqrt{v_{k-1,0}}}(1-\beta_1)J_1\sum_{i=0}^{n-1}\partial_l f_i(x_{k-1,0})+\sum_{l=1}^{d}\mathbb{I}_{k-1}\frac{\partial_l f(x_{k-1,0})}{\sqrt{v_{k-1,0}}}\left(\mathscr{M}_{l,k-1}-(F_{l,k-1})_{k-1}\right)\right\} \\
= \quad & \sum_{l=1}^{d}\mathbb{I}_{k-1}\frac{\partial_l f(x_{k-1,0})}{\sqrt{v_{k-1,0}}}(1-\beta_1)\beta_1^n J_1\sum_{i=0}^{n-1}\partial_l f_i(x_{k-1,0}).
\end{aligned}
$$

To proceed, we further take $\mathbb{E}_{k-2}(\cdot)$ and bound

$$
\mathbb{E}_{k-2}\left\{\underbrace{\sum_{l=1}^{d}\mathbb{I}_{k-1}\frac{\partial_l f(x_{k-1,0})}{\sqrt{v_{k-1,0}}}(1-\beta_1)\beta_1^n J_1\sum_{i=0}^{n-1}\partial_l f_i(x_{k-1,0})}_{(a)}+\underbrace{\sum_{l=1}^{d}\mathbb{I}_{k}\frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}}\left(\left(M'_{l,k}\right)_{k-2}-(F_{l,k})_{k-2}\right)}_{(b)}\right\}.
$$

Repeat this process until $k=1$, we have:

$$
\begin{aligned}
& \mathbb{E}\left[\sum_{l \text{ large}}\frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}}\left(\left(M'_{l,k}\right)_k-(F_{l,k})_k+\cdots+\left(M'_{l,k}\right)_1-(F_{l,k})_1\right)\right] \\
\geq \quad & (1-\beta_1)\beta_1^{kn}J_1\mathbb{E}\left[\sum_{l=1}^{d}\mathbb{I}_1\frac{\partial_l f(x_{1,0})}{\sqrt{v_{1,0}}}\sum_{i=0}^{n-1}\partial_l f_i(x_{1,0})\right] \\
& +\text{Error}_1+\text{Error}_2+\text{Error}_3 \\
\overset{(*)}{\geq} \quad & -\frac{\beta_1^n}{\sqrt{k-1}}(1-\beta_1)\left|J_1\mathbb{E}\left[\sum_{l=1}^{d}\mathbb{I}_1\frac{\partial_l f(x_{1,0})}{\sqrt{v_{1,0}}}\sum_{i=0}^{n-1}\partial_l f_i(x_{1,0})\right]\right| \\
& +\text{Error}_1+\text{Error}_2+\text{Error}_3
\end{aligned}
$$

where $(*)$ holds for large k such that $\beta^{(k-1)n}\leq\frac{\beta_1^n}{\sqrt{k-1}}$. We specifiy Error_1, Error_2, Error_3 as follows.

$$
\begin{aligned}
\text{Error}_1 \quad = \quad & -d(1-\beta_1)|J_1|\frac{1}{1-\frac{1}{\sqrt{\beta_2^n}}}n\sqrt{\frac{n\beta_2^n}{2}}\left(\sum_{j=0}^{\infty}\beta_1^{jn}\triangle_{n(k-j)}\right) \\
& -\sum_{l=1}^{d}(1-\beta_1)|J_1|\left(\left|\sum_{i=0}^{n-1}\partial_l f_i(x_{k,0})\right|\right)\left(\sqrt{\frac{2\rho_3^2}{\beta_2^n}}\frac{1}{\left(1-\frac{1-\beta_2}{2}\left(-1+\frac{4\rho_2 n}{\beta_2^n}\right)\right)}\delta_1\right)\left(\sum_{j=0}^{\infty}\beta_1^{jn}(j+1)\right) \\
& -d(1-\beta_1)\triangle_{n(k-1)}|J_1|n^2\left(\sqrt{\frac{2\rho_3^2}{\beta_2^n}}\frac{1}{\left(1-\frac{1-\beta_2}{2}\left(-1+\frac{4\rho_2 n}{\beta_2^n}\right)\right)}\delta_1\right)\left(\sum_{j=0}^{\infty}\beta_1^{jn}(j+1)^2\right) \\
& -(1-\beta_1)d\sqrt{\frac{2\rho_3^2}{\beta_2^n}}|J_1|n^2\left(\sum_{j=0}^{\infty}\beta_1^{jn}\triangle_{n(k-1-j)}\right) \\
& -(1-\beta_1)d\sqrt{\frac{2\rho_3^2}{\beta_2^n}}|J_1|n\left[Q_k+\beta_1^n(Q_k+Q_{k-1})+\beta_1^{2n}(Q_k+Q_{k-1}+Q_{k-2})+\cdots\right]
\end{aligned}
$$

Since $\delta_2=\lim_{k\to\infty}\sum_{j=1}^{k-1}(\beta_1^n)^j\sqrt{\frac{k}{k-j}}$ is a finite constant, we have $\sum_{j=0}^{\infty}\beta_1^{jn}\triangle_{n(k-j)}=\frac{\triangle_1\delta_2}{\sqrt{nk}}$. In addition, we have

$$
\begin{aligned}
\left[Q_k+\beta_1^n(Q_k+Q_{k-1})+\beta_1^{2n}(Q_k+Q_{k-1}+Q_{k-2})+\cdots\right] \quad & \overset{\text{Lemma F.1}}{\leq} \quad Q_k+2\left(\beta_1^n 2Q_k+\beta_1^{2n}3Q_k+\cdots\right) \\
& \overset{\text{Lemma F.2}}{\leq} \quad 2Q_k\frac{1}{(1-\beta_1^n)^2}
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
\text{Error}_1 \quad \geq \quad & -d(1-\beta_1)|J_1|\frac{1}{1-\frac{1}{\sqrt{\beta_2^n}}}n\sqrt{\frac{n\beta_2^n}{2}}\frac{\triangle_1\delta_2}{\sqrt{nk}} \\
& -\sum_{l=1}^{d}(1-\beta_1)|J_1|\left(\left|\sum_{i=0}^{n-1}\partial_l f_i(x_{k,0})\right|\right)\left(\sqrt{\frac{2\rho_3^2}{\beta_2^n}}\frac{1}{\left(1-\frac{1-\beta_2}{2}\left(-1+\frac{4\rho_2 n}{\beta_2^n}\right)\right)}\delta_1\right)\frac{1}{(1-\beta_1^n)^2} \\
& -d(1-\beta_1)|J_1|n^2\left(\sqrt{\frac{2\rho_3^2}{\beta_2^n}}\frac{1}{\left(1-\frac{1-\beta_2}{2}\left(-1+\frac{4\rho_2 n}{\beta_2^n}\right)\right)}\delta_1\right)\frac{1+\beta_1^n}{(1-\beta_1^n)^3}\frac{\triangle_1}{\sqrt{n(k-1)}} \\
& -(1-\beta_1)d\sqrt{\frac{2\rho_3^2}{\beta_2^n}}|J_1|n^2\frac{\triangle_1\delta_2}{\sqrt{n(k-1)}}-(1-\beta_1)d\sqrt{\frac{2\rho_3^2}{\beta_2^n}}|J_1|n2Q_k\frac{1}{(1-\beta_1^n)^2}
\end{aligned}
$$

$$
\begin{aligned}
\text{Error}_2 \quad = \quad & -d\left(\frac{1}{1-\frac{1}{\sqrt{\beta_2^n}}}\sqrt{\frac{n\beta_2^n}{2}}\right)(1-\beta_1)(\beta_1+\beta_1^n)n^2\left(\sum_{j=0}^{\infty}\beta_1^{jn}\triangle_{n(k-1-j)}\right) \\
& -\sum_{l=1}^{d}\sqrt{\frac{2\rho_3^2}{\beta_2^n}}\frac{1}{\left(1-\frac{1-\beta_2}{2}\left(-1+\frac{4\rho_2 n}{\beta_2^n}\right)\right)}\delta_1(1-\beta_1)(\beta_1+\beta_1^n)n\left(\sum_{i=0}^{n-1}|\partial_l f_i(x_{k,0})|\right)\left(\sum_{j=0}^{\infty}(j+1)\beta_1^{nj}\right) \\
& -d\sqrt{\frac{2\rho_3^2}{\beta_2^n}}\frac{1}{\left(1-\frac{1-\beta_2}{2}\left(-1+\frac{4\rho_2 n}{\beta_2^n}\right)\right)}\delta_1(1-\beta_1)(\beta_1+\beta_1^n)n^3\triangle_{n(k-1)}\left(\sum_{j=0}^{\infty}(j+1)^2\beta_1^{nj}\right) \\
& -d\sqrt{\frac{2\rho_3^2}{\beta_2^n}}(1-\beta_1)(\beta_1+\beta_1^n)n^2\left((Q_k+Q_{k-1})+\beta_1^n(Q_k+Q_{k-1}+Q_{k-2})+\cdots\right) \\
& -d\sqrt{\frac{2\rho_3^2}{\beta_2^n}}(1-\beta_1)\left(\beta_1+\beta_1^n\right)n^3\left(\sum_{j=0}^{\infty}\beta_1^{jn}\triangle_{n(k-1-j)}\right)
\end{aligned}
$$

Based on the calculation in Lemma F.2, we have

$$
\begin{aligned}
\text{Error}_2 \quad \geq \quad & -d\left(\frac{1}{1-\frac{1}{\sqrt{\beta_2^n}}}\sqrt{\frac{n\beta_2^n}{2}}\right)(1-\beta_1)(\beta_1+\beta_1^n)n^2\frac{\triangle_1}{\sqrt{n}}\frac{\delta_2}{\sqrt{k}} \\
& -\sum_{l=1}^{d}\sqrt{\frac{2\rho_3^2}{\beta_2^n}}\frac{1}{\left(1-\frac{1-\beta_2}{2}\left(-1+\frac{4\rho_2 n}{\beta_2^n}\right)\right)}\delta_1(1-\beta_1)(\beta_1+\beta_1^n)n\left(\sum_{i=0}^{n-1}|\partial_l f_i(x_{k,0})|\right)\frac{1}{(1-\beta_1^n)^2} \\
& -d\sqrt{\frac{2\rho_3^2}{\beta_2^n}}\frac{1}{\left(1-\frac{1-\beta_2}{2}\left(-1+\frac{4\rho_2 n}{\beta_2^n}\right)\right)}\delta_1(1-\beta_1)(\beta_1+\beta_1^n)n^3\frac{1+\beta_1^n}{(1-\beta_1^n)^3}\frac{\triangle_1\delta_2}{\sqrt{n(k-1)}} \\
& -d\sqrt{\frac{2\rho_3^2}{\beta_2^n}}(1-\beta_1)(\beta_1+\beta_1^n)n^2 2Q_k\frac{1}{\beta_1^n(1-\beta_1^n)^2} \\
& -d\sqrt{\frac{2\rho_3^2}{\beta_2^n}}(1-\beta_1)\left(\beta_1+\beta_1^n\right)n^3\frac{\triangle_1}{\sqrt{n}}\frac{\delta_2}{\sqrt{k}}.
\end{aligned}
$$

$$
\text{Error}_3 = -d\sqrt{\frac{2\rho_3^2}{\beta_2^n}}(1-\beta_1)n^3\left(\sum_{j=0}^{\infty}\beta_1^{jn}\triangle_{n(k-j)}\right) = -\frac{d\sqrt{\frac{2\rho_3^2}{\beta_2^n}}(1-\beta_1)n^3}{\sqrt{k}}\frac{\triangle_1}{\sqrt{n}}\delta_2,
$$

Since $\frac{1-\beta_1}{1-\beta_1^n} \leq 1$ and $|J_1| \leq n$, we have

$$
\begin{aligned}
& |J_1| \left(\sqrt{\frac{2\rho_3^2}{\beta_2^n}} \frac{1}{\left(1 - \frac{1-\beta_2}{2}\left(-1 + \frac{4\rho_2 n}{\beta_2^n}\right)\right)} \right) \frac{1-\beta_1}{(1-\beta_1^n)^2} + \sqrt{\frac{2\rho_3^2}{\beta_2^n}} \frac{1}{\left(1 - \frac{1-\beta_2}{2}\left(-1 + \frac{4\rho_2 n}{\beta_2^n}\right)\right)} \frac{2n(1-\beta_1)}{(1-\beta_1^n)^2} \\
\leq \quad & \left(\sqrt{\frac{2\rho_3^2}{\beta_2^n}} \frac{3n}{\left(1 - \frac{1-\beta_2}{2}\left(-1 + \frac{4\rho_2 n}{\beta_2^n}\right)\right)} \right) \frac{1}{(1-\beta_1^n)}.
\end{aligned}
$$

Using the fact that $\frac{1}{\sqrt{k-1}} \leq \frac{\sqrt{2}}{\sqrt{k}}$ (for $k \geq 2$), we have:

$$
\begin{aligned}
& \mathbb{E}\left[\sum_{l \text{ large}} \frac{\partial_l f(x_{k,0})}{\sqrt{v_{k,0}}} \left(\left(M'_{l,k}\right)_k - \left(F_{l,k}\right)_k + \cdots + \left(M'_{l,k}\right)_1 - \left(F_{l,k}\right)_1 \right) \right] \\
\geq \quad & -\frac{\beta_1^n}{\sqrt{k-1}}(1-\beta_1) \left| J_1 \mathbb{E}\left[\sum_{l=1}^{d} \mathbb{I}_1 \frac{\partial_l f(x_{1,0})}{\sqrt{v_{1,0}}} \sum_{i=0}^{n-1} \partial_l f_i(x_{1,0}) \right] \right| \\
& + \text{Error}_1 + \text{Error}_2 + \text{Error}_3 \\
\geq \quad & -\frac{G_6}{\sqrt{k}} - G_7 \delta_1 \mathbb{E}\left(\sum_{l=1}^{d} \sum_{i=0}^{n-1} |\partial_l f_i(x_{k,0})| \right),
\end{aligned}
$$

where

$$
\begin{aligned}
G_6 \quad := \quad & \beta_1^n (1-\beta_1)\sqrt{2}\,|J_1| \sqrt{\frac{2\rho_3^2}{\beta_2^n}} \mathbb{E}\left[\sum_{l=1}^{d} \left| \mathbb{I}_1 \sum_{i=0}^{n-1} \partial_l f_i(x_{1,0}) \right| \right] + \frac{\triangle_1}{\sqrt{n}} \delta_2 d \sqrt{\frac{2\rho_3^2}{\beta_2^n}} (1-\beta_1) n^3 \\
& + d(1-\beta_1)|J_1| \frac{1}{1 - \frac{1}{\sqrt{\beta_2^n}}} n \sqrt{\frac{n\beta_2^n}{2}} \frac{\triangle_1 \delta_2}{\sqrt{n}} \\
& + d(1-\beta_1)|J_1| n^2 \left(\sqrt{\frac{2\rho_3^2}{\beta_2^n}} \frac{1}{\left(1 - \frac{1-\beta_2}{2}\left(-1 + \frac{4\rho_2 n}{\beta_2^n}\right)\right)} \delta_1 \right) \frac{1+\beta_1^n}{(1-\beta_1^n)^3} \frac{\triangle_1 \sqrt{2}}{\sqrt{n}} \\
& + (1-\beta_1) d \sqrt{\frac{2\rho_3^2}{\beta_2^n}} |J_1| n^2 \frac{\triangle_1 \delta_2 \sqrt{2}}{\sqrt{n}} - 2(1-\beta_1) d \sqrt{\frac{2\rho_3^2}{\beta_2^n}} |J_1| n \frac{1}{(1-\beta_1^n)^2} \Delta_1 n \sqrt{n} \frac{32\sqrt{2}}{(1-\beta_2)^n \beta_2^n} \\
& + d \left(\frac{1}{1 - \frac{1}{\sqrt{\beta_2^n}}} \sqrt{\frac{n\beta_2^n}{2}} \right) (1-\beta_1)(\beta_1 + \beta_1^n) n^2 \frac{\triangle_1}{\sqrt{n}} \delta_2 \\
& + d \sqrt{\frac{2\rho_3^2}{\beta_2^n}} \frac{1}{\left(1 - \frac{1-\beta_2}{2}\left(-1 + \frac{4\rho_2 n}{\beta_2^n}\right)\right)} \delta_1 (1-\beta_1)(\beta_1 + \beta_1^n) n^3 \frac{1+\beta_1^n}{(1-\beta_1^n)^3} \frac{\triangle_1 \delta_2 \sqrt{2}}{\sqrt{n}} \\
& + 2d \sqrt{\frac{2\rho_3^2}{\beta_2^n}} (1-\beta_1)(\beta_1 + \beta_1^n) n^2 \frac{1}{\beta_1^n (1-\beta_1^n)^2} \Delta_1 n \sqrt{n} \frac{32\sqrt{2}}{(1-\beta_2)^n \beta_2^n} \\
& + d \sqrt{\frac{2\rho_3^2}{\beta_2^n}} (1-\beta_1)\left(\beta_1 + \beta_1^n\right) n^3 \frac{\triangle_1}{\sqrt{n}} \delta_2,
\end{aligned}
$$

$$
G_7 \quad := \quad \left(\sqrt{\frac{2\rho_3^2}{\beta_2^n}} \frac{3n}{\left(1 - \frac{1-\beta_2}{2}\left(-1 + \frac{4\rho_2 n}{\beta_2^n}\right)\right)} \right) \frac{1}{(1-\beta_1^n)}
$$

where $J_1 = \left(-\frac{1}{n}\beta_1 - \frac{2}{n}\beta_1^2 - \cdots - \frac{n-1}{n}\beta_1^{n-1}\right)$, $\delta_2 = \lim_{k\to\infty}\sum_{j=1}^{k-1}(\beta_1^n)^j\sqrt{\frac{k}{k-j}}$ (if needed, we can further bound J_1 by n for simplicity). This conclude the proof.

G.8 Proof of Lemma G.6

We now derive upper bounds for $\sum_{l \text{ large}}(a_2)$ and $\sum_{l \text{ large}}(b_2)$. The upper bound for $\sum_{l \text{ large}}(a_2)$ is very straightforward using inequality (21).

$$\sum_{l \text{ large}}(a_2) = \delta_1\sqrt{\frac{2\rho_3^2}{\beta_2^n}}\sum_{l \text{ large}}\sum_{i=0}^{n-1}|\partial_l f_i(x_{k,0})| \overset{(21)}{\leq} \delta_1\sqrt{\frac{2\rho_3^2}{\beta_2^n}}\sqrt{D_1}\rho_1 d\left(|\partial_\alpha f(x_{k,0})| + \sqrt{\frac{D_0}{D_1 d}}\right).$$

Now we shift gear to $\sum_{l \text{ large}}(b_2) := \delta_1\sqrt{\frac{2\rho_3^2}{\beta_2^n}}\sum_{l \text{ large}}\sum_{i=0}^{n-1}|m_{l,k,i} - \partial_l f_i(x_{k,0})|$. To proceed, we need an upper bound for $\sum_{l \text{ large}}\sum_{i=0}^{n-1}|m_{l,k,i} - \partial_l f_i(x_{k,0})|$. For each i, we perform the following decomposition.

$$\begin{aligned}\sum_{l \text{ large}}|m_{l,k,i} - \partial_l f_i(x_{k,0})| \quad \leq \quad & \underbrace{\sum_{l=1}^{d}|(1-\beta_1)\left[\beta_1^i\partial_l f_{\tau_{k,0}}(x_{k,0}) + \cdots + \partial_l f_{\tau_{k,i}}(x_{k,i})\right] - \partial_l f_i(x_{k,0})|}_{(d_1)} \\ & + \underbrace{\sum_{l=1}^{d}\left|\beta_1^{i+1}m_{l,k-1,n-1}\right|}_{(d_2)}. \end{aligned} \tag{57}$$

To start, we bound (d_1).

$$\begin{aligned}(d_1) \quad &\leq \quad \sum_{l=1}^{d}\left\{|\partial_l f_{\tau_{k,0}}(x_{k,0})| + |\partial_l f_{\tau_{k,1}}(x_{k,1})| + \cdots + |\partial_l f_{\tau_{k,i}}(x_{k,i})| + |\partial_l f_i(x_{k,0})|\right\} \\ &\leq \quad \sum_{l=1}^{d}\left\{|\partial_l f_{\tau_{k,0}}(x_{k,0})| + |\partial_l f_{\tau_{k,1}}(x_{k,1})| + \cdots + |\partial_l f_{\tau_{k,n-1}}(x_{k,n-1})| + |\partial_l f_i(x_{k,0})|\right\} \\ &\leq \quad \sum_{l=1}^{d}\sum_{i=0}^{n-1}\left\{\ |\partial_l f_i(x_{k,0})|\ \right\} + \sum_{l=1}^{d}|\partial_l f_i(x_{k,0})| + d\triangle_{nk} + \cdots + nd\triangle_{nk} \\ &\overset{(21)}{\leq} \quad 2\sqrt{D_1}\rho_1 d\left(|\partial_\alpha f(x_{k,0})| + \sqrt{\frac{D_0}{D_1 d}}\right) + \frac{n(n+1)d}{2}\triangle_{nk} \\ &= \quad 2\sqrt{D_1}\rho_1 d\left(|\partial_\alpha f(x_{k,0})| + \sqrt{\frac{D_0}{D_1 d}}\right) + \frac{(n+1)\sqrt{n}d}{2\sqrt{k}}\triangle_1. \end{aligned} \tag{58}$$

Now, we bound (d_2). Recall $\sum_{i=0}^{n-1}\partial_l f_i(x_{1,0}) = \partial_l f(x_{1,0})$, we have

$$\begin{aligned}|m_{l,k-1,n-1}| \quad \leq \quad & (1-\beta_1)\left[|\partial_l f_{\tau_{k-1,n-1}}(x_{k-1,n-1})| + \beta_1|\partial_l f_{\tau_{k-1,n-2}}(x_{k-1,n-2})| + \cdots\right] \\ & + \beta_1^{(k-1)n}\sum_{i=0}^{n-1}|\partial_l f_i(x_{1,0})| \end{aligned} \tag{59}$$

Note that for any $i \in [0, n-1]$, $j \in [0, n-1]$, $t \in [1, k-1]$, we have the following result.

$$
\begin{aligned}
|\partial_l f_i(x_{k-t,j})| &\leq |\partial_l f_i(x_{k,0})| + |\partial_l f_i(x_{k,0}) - \partial_l f_i(x_{k-t,j})| \\
&\overset{\text{Lemma } F.3}{\leq} |\partial_l f_i(x_{k,0})| + (n-j)\triangle_{(k-t)n} + n\triangle_{(k-t+1)n} + \cdots + n\triangle_{(k-1)n} \\
&\leq |\partial_l f_i(x_{k,0})| + \triangle_{(k-1)n} + \triangle_{(k-1)n-1} + \cdots + \triangle_{(k-1)n-[(n-j)+(t-1)n-1]} \\
&\leq |\partial_l f_i(x_{k,0})| + \triangle_{(k-1)n-1} + \triangle_{(k-1)n-2} + \cdots + \triangle_{(k-1)n-[(n-j)+(t-1)n]} \\
&\overset{\text{Lemma F.1}}{\leq} |\partial_l f_i(x_{k,0})| + \frac{2[(n-j)+(t-1)n]\triangle_1}{\sqrt{n(k-1)-1}}. \qquad (60)
\end{aligned}
$$

Plugging (60) into (59) and re-arranging the index, we have

$$
\begin{aligned}
\sum_{l=1}^{d} |m_{l,k-1,n-1}| &\overset{(59)}{\leq} (1-\beta_1)\sum_{l=1}^{d}\left[|\partial_l f_{\tau_{k-1,n-1}}(x_{k-1,n-1})| + \beta_1|\partial_l f_{\tau_{k-1,n-2}}(x_{k-1,n-2})| + \cdots\right] \\
&\quad + \beta_1^{(k-1)n}\sum_{l=1}^{d}\sum_{i=0}^{n-1}|\partial_l f_i(x_{1,0})| \\
&\overset{(21),(60)}{\leq} (1-\beta_1)\sum_{q=1}^{(k-1)n}\beta_1^{q-1}\left[\sqrt{D_1}\rho_1 d\left(|\partial_\alpha f(x_{k,0})| + \sqrt{\frac{D_0}{D_1 d}}\right) + \frac{2qd\triangle_1}{\sqrt{n(k-1)-1}}\right] \\
&\quad + \beta_1^{(k-1)n}\sum_{l=1}^{d}\sum_{i=0}^{n-1}|\partial_l f_i(x_{1,0})| \\
&\leq (1-\beta_1)\sum_{q=1}^{\infty}\beta_1^{q-1}\left[\sqrt{D_1}\rho_1 d\left(|\partial_\alpha f(x_{k,0})| + \sqrt{\frac{D_0}{D_1 d}}\right) + \frac{2qd\triangle_1}{\sqrt{n(k-1)-1}}\right] \\
&\quad + \beta_1^{(k-1)n}\sum_{l=1}^{d}\sum_{i=0}^{n-1}|\partial_l f_i(x_{1,0})| \\
&\overset{\text{Lemma } F.2}{\leq} \sqrt{D_1}\rho_1 d\left(|\partial_\alpha f(x_{k,0})| + \sqrt{\frac{D_0}{D_1 d}}\right) + \frac{1}{1-\beta_1}\frac{2d\triangle_1}{\sqrt{n(k-1)-1}} \\
&\quad + \beta_1^{(k-1)n}\sum_{l=1}^{d}\sum_{i=0}^{n-1}|\partial_l f_i(x_{1,0})| \\
&\overset{\text{When } k\geq 4}{\leq} \sqrt{D_1}\rho_1 d\left(|\partial_\alpha f(x_{k,0})| + \sqrt{\frac{D_0}{D_1 d}}\right) + \frac{1}{1-\beta_1}\frac{2\sqrt{2}d\triangle_1}{\sqrt{nk}} \\
&\quad + \beta_1^{(k-1)n}\sum_{l=1}^{d}\sum_{i=0}^{n-1}|\partial_l f_i(x_{1,0})| \\
&\overset{(i)}{\leq} \sqrt{D_1}\rho_1 d\left(|\partial_\alpha f(x_{k,0})| + \sqrt{\frac{D_0}{D_1 d}}\right) + \frac{1}{1-\beta_1}\frac{2\sqrt{2}d\triangle_1}{\sqrt{nk}} \\
&\quad + \frac{\sqrt{2}\beta_1^n}{\sqrt{k}}\sum_{l=1}^{d}\sum_{i=0}^{n-1}|\partial_l f_i(x_{1,0})| \qquad (61)
\end{aligned}
$$

where the second last inequality holds because: when $k \geq 4$, $\frac{1}{\sqrt{n(k-1)-1}} \leq \frac{\sqrt{2}}{\sqrt{nk}}$. (i) holds when k is large enough such that $\beta_1^{(k-1)n} \leq \frac{\beta_1^n}{\sqrt{k-1}}$. Further, $\frac{\beta_1^n}{\sqrt{k-1}} \leq \frac{\sqrt{2}\beta_1^n}{\sqrt{k}}$ when $k \geq 2$.

Now, we have derived upper bounds for (d_1) and (d_2). Plugging (61) and (58) into (57), we conclude the proof.

$$
\begin{aligned}
\sum_{l=1}^{d}\sum_{i=0}^{n-1}|m_{l,k,i}-\partial_l f_{\tau_{k,i}}(x_{k,0})| &\overset{(58),(61)}{\leq} 3n\sqrt{D_1}\rho_1 d\left(|\partial_\alpha f\left(x_{k,0}\right)|+\sqrt{\frac{D_0}{D_1 d}}\right)+\frac{(n+1)nd\sqrt{n}}{2\sqrt{k}}\Delta_1 \\
&\quad +\frac{nd}{1-\beta_1}\frac{2\sqrt{2}\triangle_1}{\sqrt{nk}}+\frac{\sqrt{2}n\beta_1^n}{\sqrt{k}}\sum_{l=1}^{d}\sum_{i=0}^{n-1}|\partial_l f_i(x_{1,0})| \\
&= \frac{1}{\sqrt{k}}\left[\frac{d(n+1)n^{\frac{3}{2}}}{2}\Delta_1+\frac{d2\sqrt{2}\sqrt{n}\triangle_1}{1-\beta_1}+\sqrt{2}n\beta_1^n\sum_{i=0}^{n-1}\|\nabla f_i(x_{1,0})\|_1\right] \\
&\quad +3n\sqrt{D_1}\rho_1 d\left(|\partial_\alpha f\left(x_{k,0}\right)|+\sqrt{\frac{D_0}{D_1 d}}\right). \qquad (62)
\end{aligned}
$$

$$
\begin{aligned}
\sum_{l \text{ large}}(b_2) &= \delta_1\sqrt{\frac{2\rho_3^2}{\beta_2^n}}\sum_{l \text{ large}}\sum_{i=0}^{n-1}|m_{l,k,i}-\partial_l f_i(x_{k,0})| \\
&\leq \delta_1\sqrt{\frac{2\rho_3^2}{\beta_2^n}}\left\{\frac{1}{\sqrt{k}}\left[\frac{d(n+1)n^{\frac{3}{2}}}{2}\Delta_1+\frac{d2\sqrt{2}\sqrt{n}\triangle_1}{1-\beta_1}+\sqrt{2}n\beta_1^n\sum_{i=0}^{n-1}\|\nabla f_i(x_{1,0})\|_1\right]\right. \\
&\quad \left.+3n\sqrt{D_1}\rho_1 d\left(|\partial_\alpha f\left(x_{k,0}\right)|+\sqrt{\frac{D_0}{D_1 d}}\right)\right\}.
\end{aligned}
$$

The proof is concluded by adding expectation on both sides of the inequality.

G.9 Proof of Lemma G.7

$$
\begin{aligned}
&\mathbb{E}\left\{\sum_{l \text{ large}}(a_1)+\sum_{l \text{ large}}(b_1)-\sum_{l \text{ large}}\{(a_2)+(b_2)\}\right\} \\
= &\mathbb{E}\left[\sum_{l \text{ large}}\frac{\partial_l f\left(x_{k,0}\right)^2}{\sqrt{v_{l,k,0}}}\right]+\mathbb{E}\left[\sum_{l \text{ large}}\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,0}}}\sum_{i=0}^{n-1}\left(m_{l,k,i}-\partial_l f_i(x_{k,0})\right)\right] \\
&-d\delta_1\sqrt{\frac{2\rho_3^2}{\beta_2^n}}\mathbb{E}\left[\sum_{i=0}^{n-1}|\partial_l f_i\left(x_{k,0}\right)|\right]-d\delta_1\sqrt{\frac{2\rho_3^2}{\beta_2^n}}\mathbb{E}\left[\sum_{i=0}^{n-1}|m_{l,k,i}-\partial_l f_i(x_{k,0})|\right] \\
\overset{\text{Lemma G.5 and G.6 and F.4}}{\geq} &\mathbb{E}\left[\sum_{l \text{ large}}\frac{\partial_l f\left(x_{k,0}\right)^2}{\sqrt{v_{l,k,0}}}\right]-\frac{1}{\sqrt{k}}\left(G_4+G_5+G_6\right)-G_7\delta_1\sqrt{D_1}\rho_1 d\left(\mathbb{E}\left|\partial_\alpha f\left(x_{k,0}\right)\right|+\sqrt{\frac{D_0}{D_1 d}}\right) \\
&-\delta_1\sqrt{\frac{2\rho_3^2}{\beta_2^n}}\sqrt{D_1}\rho_1 d\left(\mathbb{E}\left|\partial_\alpha f\left(x_{k,0}\right)\right|+\sqrt{\frac{D_0}{D_1 d}}\right) \\
&-\delta_1\sqrt{\frac{2\rho_3^2}{\beta_2^n}}\left[\frac{d(n+1)n^{\frac{3}{2}}}{2}\Delta_1+\frac{d2\sqrt{2}\sqrt{n}\triangle_1}{1-\beta_1}+\sqrt{2}n\beta_1^n\sum_{i=0}^{n-1}\mathbb{E}\|\nabla f_i(x_{1,0})\|_1\right]\frac{1}{\sqrt{k}} \\
&-\delta_1\sqrt{\frac{2\rho_3^2}{\beta_2^n}}3n\sqrt{D_1}\rho_1 d\left(\mathbb{E}\left|\partial_\alpha f\left(x_{k,0}\right)\right|+\sqrt{\frac{D_0}{D_1 d}}\right),
\end{aligned}
$$

where constant terms G_4, G_5, G_6, G_7 can be seen at the end of Appendix G.7.

Since $\alpha=\arg\max_{l=1,2,\cdots,d}|\partial_l f\left(x_{k,0}\right)|$ and $\sum_{l \text{ large}}\frac{\partial_l f(x_{k,0})^2}{\sqrt{v_{l,k,0}}}\geq\frac{\partial_\alpha f(x_{k,0})^2}{\sqrt{v_{\alpha,k,0}}}$, we have

$$
\begin{aligned}
& \mathbb{E}\left\{\sum_{l \text{ large}}(a_1)+\sum_{l \text{ large}}(b_1)-\sum_{l \text{ large}}\{(a_2)+(b_2)\}\right\} \\
\overset{(*)}{\geq} \quad & \mathbb{E}\left\{\frac{\partial_\alpha f\left(x_{k,0}\right)^2}{\sqrt{v_{\alpha,k,0}}}\right\} \\
& -\left[G_4+G_5+G_6+\delta_1\sqrt{\frac{2\rho_3^2}{\beta_2^n}}\left(\frac{d(n+1)n^{\frac{3}{2}}}{2}\Delta_1+\frac{d2\sqrt{2}\sqrt{n}\triangle_1}{1-\beta_1}+\sqrt{2}n\beta_1^n\sum_{i=0}^{n-1}\mathbb{E}\|\nabla f_i(x_{1,0})\|_1\right)\right]\frac{1}{\sqrt{k}} \\
& -\delta_1\left(\sqrt{\frac{2\rho_3^2}{\beta_2^n}}4n+G_7\right)\sqrt{D_1}\rho_1 d\left(\mathbb{E}\left|\partial_\alpha f\left(x_{k,0}\right)\right|+\sqrt{\frac{D_0}{D_1 d}}\right) \\
= \quad & \mathbb{E}\left\{\frac{\partial_\alpha f\left(x_{k,0}\right)^2}{\sqrt{v_{\alpha,k,0}}}\right\}-F_2\frac{1}{\sqrt{k}}-F_3\mathbb{E}\left|\partial_\alpha f\left(x_{k,0}\right)\right|-F_4, \qquad (63)
\end{aligned}
$$

where $F_2 := \delta_1\sqrt{\frac{2\rho_3^2}{\beta_2^n}}\left[\frac{d(n+1)n^{\frac{3}{2}}}{2}\Delta_1+\frac{d2\sqrt{2}\sqrt{n}\triangle_1}{1-\beta_1}+\sqrt{2}n\beta_1^n\sum_{i=0}^{n-1}\|\nabla f_i(x_{1,0})\|_1\right]+G_4+G_5+G_6$;

$F_3 := \delta_1\left(\sqrt{\frac{2\rho_3^2}{\beta_2^n}}4n+G_7\right)\sqrt{D_1}\rho_1 d$;

$F_4 := \delta_1\left(\sqrt{\frac{2\rho_3^2}{\beta_2^n}}4n+G_7\right)\sqrt{D_1}\rho_1 d\sqrt{\frac{D_0}{D_1 d}}$.

To proceed, we introduce the following Lemma G.16, which is Lemma F.2 in (Shi et al., 2020).

Lemma G.16. *(Lemma F.2 in (Shi et al., 2020)) Given k, we set α as the index of the coordinate with the greatest gradient:*

$$\alpha = \arg\max_{l=1,2,\cdots,d}\left|\partial_l f\left(x_{k,0}\right)\right|$$

Under Assumption 2.1 and 2.2, if $k \geq 4$ and $\sqrt{\left|\partial_\alpha f\left(x_{k,0}\right)\right|^2+\frac{D_0}{D_1 d}} \geq 4\sqrt{2}\frac{\Delta_1}{(1-\beta_2)\sqrt{D_1 nkd}}$, the following holds:

$$v_{\alpha,k,0} \leq \frac{5}{2}D_1 d\left(\left|\partial_\alpha f\left(x_{k,0}\right)\right|^2+\frac{D_0}{D_1 d}\right).$$

The proof of Lemma G.16 can be seen in (Shi et al., 2020), Lemma F.2, Page 26. Now, we discuss two cases.

Case (a): when $\left|\partial_\alpha f\left(x_{k,0}\right)\right| \geq 4\sqrt{2}\frac{\Delta_1}{(1-\beta_2)\sqrt{D_1 nkd}}$. In this case, the condition of Lemma G.16 is satisfied. We have we have $v_{\alpha,k,0} \leq \frac{5}{2}D_1 d\left(\left|\partial_\alpha f\left(x_{k,0}\right)\right|^2+\frac{D_0}{D_1 d}\right)$. Therefore, $\frac{\partial_\alpha f\left(x_{k,0}\right)^2}{\sqrt{v_{\alpha,k,0}}} \geq \frac{\partial_\alpha f(x_{k,0})^2}{\sqrt{\frac{5}{2}D_1 d\left(|\partial_\alpha f(x_{k,0})|^2+\frac{D_0}{D_1 d}\right)}}$. To proceed, we consider the following two sub-cases.

- When $\partial_\alpha f(x_{k,0})^2 \leq \frac{D_0}{D_1 d}$:

$$
\begin{aligned}
& \frac{\partial_\alpha f\left(x_{k,0}\right)^2}{\sqrt{v_{\alpha,k,0}}}-F_2\frac{1}{\sqrt{k}}-F_3\left|\partial_\alpha f\left(x_{k,0}\right)\right|-F_4 \\
\geq & \frac{\partial_\alpha f\left(x_{k,0}\right)^2}{\sqrt{5D_0}}-F_2\frac{1}{\sqrt{k}}-F_3\sqrt{\frac{D_0}{D_1 d}}-F_4
\end{aligned}
$$

- When $\partial_\alpha f(x_{k,0})^2 \geq \frac{D_0}{D_1 d}$:

$$\begin{aligned}
&\frac{\partial_\alpha f\left(x_{k,0}\right)^2}{\sqrt{v_{\alpha,k,0}}} - F_2\frac{1}{\sqrt{k}} - F_3\left|\partial_\alpha f\left(x_{k,0}\right)\right| - F_4 \\
&\geq \frac{|\partial_\alpha f(x_{k,0})|}{\sqrt{5D_1 d}} - F_2\frac{1}{\sqrt{k}} - F_3\left|\partial_\alpha f\left(x_{k,0}\right)\right| - F_4 \\
&= \left|\partial_\alpha f\left(x_{k,0}\right)\right|\left(\frac{1}{\sqrt{5D_1 d}} - F_3\right) - F_2\frac{1}{\sqrt{k}} - F_4 \\
&\geq \left|\partial_\alpha f\left(x_{k,0}\right)\right|\left(\frac{1}{\sqrt{5D_1 d}} - F_3\right) - F_2\frac{1}{\sqrt{k}} - F_3\sqrt{\frac{D_0}{D_1 d}} - F_4
\end{aligned}$$

Combining together, we have the following results for **Case (a)**.

$$\left\{\frac{\partial_\alpha f\left(x_{k,0}\right)^2}{\sqrt{v_{\alpha,k,0}}}\right\} - F_2\frac{1}{\sqrt{k}} - F_3\left|\partial_\alpha f\left(x_{k,0}\right)\right| - F_4 \geq \min\left\{\frac{\partial_\alpha f(x_{k,0})^2}{\sqrt{5D_0}}, \left|\partial_\alpha f\left(x_{k,0}\right)\right|\left(\frac{1}{\sqrt{5D_1 d}} - F_3\right)\right\} - F_2\frac{1}{\sqrt{k}} - F_3\sqrt{\frac{D_0}{D_1 d}} - F_4.$$

Case (b): When $\left|\partial_\alpha f\left(x_{k,0}\right)\right| < 4\sqrt{2}\frac{\Delta_1}{(1-\beta_2)\sqrt{D_1 nkd}}$, we have

$$\begin{aligned}
\left\{\frac{\partial_\alpha f\left(x_{k,0}\right)^2}{\sqrt{v_{\alpha,k,0}}}\right\} - F_2\frac{1}{\sqrt{k}} - F_3\left|\partial_\alpha f\left(x_{k,0}\right)\right| - F_4 &\geq -F_2\frac{1}{\sqrt{k}} - F_3\left|\partial_\alpha f\left(x_{k,0}\right)\right| - F_4 \\
&\geq -F_2\frac{1}{\sqrt{k}} - F_3 4\sqrt{2}\frac{\Delta_1}{(1-\beta_2)\sqrt{D_1 nkd}} - F_4 \\
&= -\frac{1}{\sqrt{k}}G_1 - F_4, \qquad (64)
\end{aligned}$$

where $G_1 := F_2 + F_3 4\sqrt{2}\frac{\Delta_1}{(1-\beta_2)\sqrt{D_1 nd}}$. Now, the following two claims are both true.

- Claim 1:

$$(64) \quad \geq \quad \frac{\partial_\alpha f\left(x_{k,0}\right)^2}{\sqrt{5D_0}} - \frac{1}{k}\frac{\left(4\sqrt{2}\frac{\Delta_1}{(1-\beta_2)\sqrt{D_1 nd}}\right)^2}{\sqrt{5D_0}} - \frac{1}{\sqrt{k}}G_1 - F_4.$$

- Claim 2:

$$\begin{aligned}
(64) \quad \geq \quad & \left|\partial_\alpha f\left(x_{k,0}\right)\right|\left(\frac{1}{\sqrt{5D_1 d}} - F_3\right) - \frac{1}{\sqrt{k}}4\sqrt{2}\frac{\Delta_1}{(1-\beta_2)\sqrt{D_1 nd}}\left(\frac{1}{\sqrt{5D_1 d}} - F_3\right) \\
& -\frac{1}{\sqrt{k}}G_1 - F_4.
\end{aligned}$$

Combining Claim 1 and Claim 2, we have

$$\begin{aligned}
[(64)] \quad \geq \quad & \min\left\{\frac{\partial_\alpha f\left(x_{k,0}\right)^2}{\sqrt{5D_0}}, \left|\partial_\alpha f\left(x_{k,0}\right)\right|\left(\frac{1}{\sqrt{5D_1 d}} - F_3\right)\right\} - F_4 \\
& -\frac{1}{\sqrt{k}}\left[\max\left\{\frac{\left(4\sqrt{2}\frac{\Delta_1}{(1-\beta_2)\sqrt{D_1 nd}}\right)^2}{\sqrt{5D_0}}, \frac{4\sqrt{2}\Delta_1}{(1-\beta_2)\sqrt{D_1 nd}}\left(\frac{1}{\sqrt{5D_1 d}} - F_3\right)\right\} + G_1\right] \\
:= \quad & \min\left\{\frac{\partial_\alpha f\left(x_{k,0}\right)^2}{\sqrt{5D_0}}, \left|\partial_\alpha f\left(x_{k,0}\right)\right|\left(\frac{1}{\sqrt{5D_1 d}} - F_3\right)\right\} - \frac{1}{\sqrt{k}}G_2 - F_4, \qquad (65)
\end{aligned}$$

where $G_2 := \max\left\{\frac{\left(4\sqrt{2}\frac{\Delta_1}{(1-\beta_2)\sqrt{D_1 nd}}\right)^2}{\sqrt{5D_0}}, \frac{4\sqrt{2}\Delta_1}{(1-\beta_2)\sqrt{D_1 nd}}\left(\frac{1}{\sqrt{5D_1 d}} - F_3\right)\right\} + G_1.$

Combining **Case (a)** and **Case (b)** together, we have

$$\begin{aligned}
\mathbb{E}\left\{\sum_{l \text{ large}}(a_1)+\sum_{l \text{ large}}(b_1)-\sum_{l \text{ large}}\{(a_2)+(b_2)\}\right\} &\geq \mathbb{E}\min\left\{\frac{\partial_\alpha f\left(x_{k,0}\right)^2}{\sqrt{5D_0}}, \left|\partial_\alpha f\left(x_{k,0}\right)\right|\left(\frac{1}{\sqrt{5D_1 d}}-F_3\right)\right\} \\
&\quad -\frac{1}{\sqrt{k}}G_2-F_4 \\
&\geq \mathbb{E}\min\left\{\frac{\|\nabla f(x_{k,0})\|_2^2}{d\sqrt{5D_0}}, \|\nabla f(x_{k,0})\|_1\left(\frac{1}{d\sqrt{5D_1 d}}-\frac{F_3}{d}\right)\right\} \\
&\quad -\frac{1}{\sqrt{k}}G_2-F_4,
\end{aligned}$$

where the last inequality is because of $\|\nabla f(x_{k,0})\|_2^2 \leq d\partial_\alpha f\left(x_{k,0}\right)^2$, $\|\nabla f(x_{k,0})\|_1 \leq d|\partial_\alpha f\left(x_{k,0}\right)|$. Recall $F_3 \to 0$ when $\beta_2 \to 1$, so there exists an interval $(1-\epsilon, 1]$, such that $\frac{1}{\sqrt{5D_1 d}}-F_3 \geq \frac{1}{\sqrt{10D_1 d}}$, or equivalently $F_3 \leq \frac{1}{\sqrt{10D_1 d}}$ (note that F_3 is the same as "$A(\beta_2)$" stated in the condition of Lemma G.7). With such a choice of β_2, we have the following results by changing all the F_3 into $\frac{1}{\sqrt{10D_1 d}}$:

$$\begin{aligned}
\mathbb{E}\left\{\sum_{l \text{ large}}(a_1)+\sum_{l \text{ large}}(b_1)-\sum_{l \text{ large}}\{(a_2)+(b_2)\}\right\} &\geq \mathbb{E}\min\left\{\frac{\|\nabla f(x_{k,0})\|_2^2}{d\sqrt{5D_0}}, \|\nabla f(x_{k,0})\|_1\frac{1}{d\sqrt{10D_1 d}}\right\} \\
&\quad -\frac{1}{\sqrt{k}}G_2-F_4. \\
&\overset{(*)}{\geq} \frac{1}{d\sqrt{10D_1 d}}\mathbb{E}\min\left\{\sqrt{\frac{2D_1 d}{D_0}}\|\nabla f(x_{k,0})\|_2^2, \|\nabla f(x_{k,0})\|_1\right\} \\
&\quad -\frac{1}{\sqrt{k}}G_2-F_5.
\end{aligned}$$

In inequality $(*)$, we change F_4 into $F_5 := \sqrt{\frac{D_0}{D_1 d}}\frac{1}{\sqrt{10D_1 d}}$. This is because: first, $F_4 = F_3\sqrt{\frac{D_0}{D_1 d}}$; second, $F_3 \leq \frac{1}{\sqrt{10D_1 d}}$.

The proof of Lemma G.7 is completed.

We restate all the constants as follows (G_4, G_5, G_6 are specified in Appendix G.7):

$$G_2 := \max\left\{\frac{\left(4\sqrt{2}\frac{\Delta_1}{(1-\beta_2)\sqrt{D_1 nd}}\right)^2}{\sqrt{5D_0}}, \frac{4\sqrt{2}\Delta_1}{(1-\beta_2)\sqrt{D_1 nd}\sqrt{10D_1 d}}\right\}+G_1;$$

$$G_1 := F_2 + 4\sqrt{2}\frac{\Delta_1}{(1-\beta_2)\sqrt{D_1 nd}\sqrt{10D_1 d}};$$

$$F_2 := \delta_1\sqrt{\frac{2\rho_3^2}{\beta_2^n}}\left[\frac{d(n+1)n^{\frac{3}{2}}}{2}\Delta_1+\frac{d2\sqrt{2}\sqrt{n}\triangle_1}{1-\beta_1}+\sqrt{2}n\beta_1^n\sum_{i=0}^{n-1}\|\nabla f_i(x_{1,0})\|_1\right]+G_4+G_5+G_6;$$

$$F_5 := \sqrt{\frac{D_0}{D_1 d}}\frac{1}{\sqrt{10D_1 d}}.$$

Remark G.17. We comment that we can always replace F_5 in the final result by F_4, i.e., we can choose not to apply the last inequality $(*)$ in the proof. The benefit of using F_4 is that F_4 monotonously decrease to 0 when increasing β_2 to 1. This monotone property is erased in the notation of F_5. Nevertheless, we choose to use F_5 since it is a much cleaner constant.

G.10 Proof of Lemma G.10

Based on Descent Lemma, we have

$$
\begin{aligned}
\sum_{k=t_0}^{T} \frac{\eta_0}{\sqrt{nk}}(\text{r.h.s. of (33)}) &\leq \sum_{k=t_0}^{T} \mathbb{E}\left\langle \nabla f\left(x_{k,0}\right), x_{k,0}-x_{k+1,0}\right\rangle \\
&\leq \mathbb{E} f\left(x_{t_0,0}\right)-\mathbb{E} f\left(x_{T+1,0}\right)+\sum_{k=t_0}^{T} \frac{L}{2} \mathbb{E}\left\|x_{k+1,0}-x_{k,0}\right\|_2^2 \\
&\overset{(18)}{\leq} \mathbb{E} f\left(x_{t_0,0}\right)-f^*+\sum_{k=t_0}^{T} \frac{L d}{2}\left(\frac{n \eta_0}{\sqrt{n k}} \frac{(1-\beta_1)}{\sqrt{1-\beta_2}} \frac{1}{1-\frac{\beta_1}{\sqrt{\beta_2}}}\right)^2 \\
&= \mathbb{E} f\left(x_{t_0,0}\right)-f^*+\sum_{k=t_0}^{T} \frac{L d n}{2 k}\left(\eta_0 \frac{(1-\beta_1)}{\sqrt{1-\beta_2}} \frac{1}{1-\frac{\beta_1}{\sqrt{\beta_2}}}\right)^2
\end{aligned}
$$

Plugging in the r.h.s. of (33), we have the following relation after rearranging.

$$
\begin{aligned}
&\sum_{k=t_0}^{T} \frac{\eta_0}{\sqrt{nk}}\left[\frac{1}{d \sqrt{10 D_1 d}} \mathbb{E} \min \left\{\sqrt{\frac{2 D_1 d}{D_0}}\|\nabla f(x_{k,0})\|_2^2, \|\nabla f(x_{k,0})\|_1\right\}-\mathbb{E}\left[\mathcal{O}(\frac{1}{\sqrt{k}})\right]-\mathcal{O}(\sqrt{D_0})\right] \\
&\leq \mathbb{E} f\left(x_{t_0,0}\right)-f^*+\sum_{k=t_0}^{T} \frac{1}{k}\left[\frac{L d n}{2}\left(\eta_0 \frac{(1-\beta_1)}{\sqrt{1-\beta_2}} \frac{1}{1-\frac{\beta_1}{\sqrt{\beta_2}}}\right)^2\right].
\end{aligned}
$$

Recall in Lemma F.1, we have $2\left(\sqrt{T}-\sqrt{t_0-1}\right) \leq \sum_{k=t_0}^{T} \frac{1}{\sqrt{k}}$, $\sum_{k=t_0}^{T} \frac{1}{k} \leq \log \frac{T+1}{t_0}$. Further, since $\|\cdot\|_1 \geq\|\cdot\|_2$, we get the following relation when $t_0=1$.

$$
\begin{aligned}
&\min_{k \in[1, T]} \mathbb{E}\left[\min \left\{\sqrt{\frac{2 D_1 d}{D_0}}\|\nabla f(x_{k,0})\|_2^2, \|\nabla f(x_{k,0})\|_2\right\}\right] \\
&\leq \frac{1}{\sqrt{T}} \frac{d \sqrt{10 D_1 d n}}{2 \eta_0}\left[\mathbb{E} f\left(x_{1,0}\right)-f^*+\log (T+1)\left(\frac{L d n}{2}\left(\eta_0 \frac{(1-\beta_1)}{\sqrt{1-\beta_2}} \frac{1}{1-\frac{\beta_1}{\sqrt{\beta_2}}}\right)^2+H_2\right)\right]+F_5,
\end{aligned}
\tag{66}
$$

We specify all the constant as follows. For all the following constants, We keep the same notation as their appearance in their corresponding lemmas.

$F_5 := \sqrt{\frac{D_0}{D_1 d}} \frac{1}{\sqrt{10 D_1 d}}$;

$H_2 := \frac{\eta_0(G_2+d F_1)}{\sqrt{n}}$;

$F_1 := \triangle_1 n^2 \sqrt{n} \frac{32 \sqrt{2}}{(1-\beta_2)^n \beta_2^n} \frac{1-\beta_1}{1-\frac{\beta_1}{\sqrt{\beta_2}}} n$;

$G_2 := \max \left\{\frac{\left(4 \sqrt{2} \frac{\triangle_1}{(1-\beta_2) \sqrt{D_1 n d}}\right)^2}{\sqrt{5 D_0}}, \frac{4 \sqrt{2} \triangle_1}{(1-\beta_2) \sqrt{D_1 n d} \sqrt{10 D_1 d}}\right\}+G_1$;

$G_1 := F_2+4 \sqrt{2} \frac{\triangle_1}{(1-\beta_2) \sqrt{D_1 n d} \sqrt{10 D_1 d}}$;

$F_2 := \delta_1 \sqrt{\frac{2 \rho_3^2}{\beta_2^n}}\left[\frac{d(n+1) n^{\frac{3}{2}}}{2} \triangle_1+\frac{d 2 \sqrt{2} \sqrt{n} \triangle_1}{1-\beta_1}+\sqrt{2} n \beta_1^n \sum_{i=0}^{n-1}\|\nabla f_i(x_{1,0})\|_1\right]+G_4+G_5+G_6$;

$G_5=d \sqrt{\frac{2 \rho_3^2}{\beta_2^n}}\left(\frac{\beta_1^{2 n} 2 n^3 \triangle_1 \sqrt{2}}{\sqrt{n}} \frac{1-\beta_1}{(1-\beta_1^n)^2}+n\left(1-\beta_1^{n-1}\right) \sum_{i=0}^{n-1} \mathbb{E}|\partial_\alpha f_i(x_{1,0})| \frac{(1-\beta_1) \beta_1^n \sqrt{2}}{1-\beta_1^n}\right)$;

$G_4 = d\sqrt{\frac{2\rho_3^2}{\beta_2^n}}\beta_1^n\sqrt{2}n(n-1)\sum_{i=0}^{n-1}\mathbb{E}\left(|\partial_\alpha f_i(x_{1,0})|\right);$

$$
\begin{aligned}
G_6 \quad := \quad & \beta_1^n(1-\beta_1)\sqrt{2}\,|J_1|\sqrt{\frac{2\rho_3^2}{\beta_2^n}}\mathbb{E}\left[\sum_{l=1}^{d}\left|\mathbb{I}_1\sum_{i=0}^{n-1}\partial_l f_i(x_{1,0})\right|\right] + \frac{\triangle_1}{\sqrt{n}}\delta_2 d\sqrt{\frac{2\rho_3^2}{\beta_2^n}}(1-\beta_1)n^3 \\
& +d(1-\beta_1)|J_1|\frac{1}{1-\frac{1}{\sqrt{\beta_2^n}}}n\sqrt{\frac{n\beta_2^n}{2}}\frac{\triangle_1\delta_2}{\sqrt{n}} \\
& +d(1-\beta_1)|J_1|n^2\left(\sqrt{\frac{2\rho_3^2}{\beta_2^n}}\frac{1}{\left(1-\frac{1-\beta_2}{2}\left(-1+\frac{4\rho_2 n}{\beta_2^n}\right)\right)}\delta_1\right)\frac{1+\beta_1^n}{(1-\beta_1^n)^3}\frac{\triangle_1\sqrt{2}}{\sqrt{n}} \\
& +(1-\beta_1)d\sqrt{\frac{2\rho_3^2}{\beta_2^n}}|J_1|n^2\frac{\triangle_1\delta_2\sqrt{2}}{\sqrt{n}} - 2(1-\beta_1)d\sqrt{\frac{2\rho_3^2}{\beta_2^n}}|J_1|n\frac{1}{(1-\beta_1^n)^2}\Delta_1 n\sqrt{n}\frac{32\sqrt{2}}{(1-\beta_2)^n\beta_2^n} \\
& +d\left(\frac{1}{1-\frac{1}{\sqrt{\beta_2^n}}}\sqrt{\frac{n\beta_2^n}{2}}\right)(1-\beta_1)(\beta_1+\beta_1^n)n^2\frac{\triangle_1}{\sqrt{n}}\delta_2 \\
& +d\sqrt{\frac{2\rho_3^2}{\beta_2^n}}\frac{1}{\left(1-\frac{1-\beta_2}{2}\left(-1+\frac{4\rho_2 n}{\beta_2^n}\right)\right)}\delta_1(1-\beta_1)(\beta_1+\beta_1^n)n^3\frac{1+\beta_1^n}{(1-\beta_1^n)^3}\frac{\triangle_1\delta_2\sqrt{2}}{\sqrt{n}} \\
& +2d\sqrt{\frac{2\rho_3^2}{\beta_2^n}}(1-\beta_1)(\beta_1+\beta_1^n)n^2\frac{1}{\beta_1^n(1-\beta_1^n)^2}\Delta_1 n\sqrt{n}\frac{32\sqrt{2}}{(1-\beta_2)^n\beta_2^n} \\
& +d\sqrt{\frac{2\rho_3^2}{\beta_2^n}}(1-\beta_1)\left(\beta_1+\beta_1^n\right)n^3\frac{\triangle_1}{\sqrt{n}}\delta_2,
\end{aligned}
$$

$\Delta_1 := \eta_0\frac{L\sqrt{d}}{\sqrt{1-\beta_2}}\frac{1-\beta_1}{1-\frac{\beta_1}{\sqrt{\beta_2}}};$

$\delta_1 := (1-\beta_2)\frac{\left(-1+\frac{4n\rho_2}{\beta_2^2}\right)}{2} + \left(\frac{1}{\sqrt{\beta_2^n}}-1\right);$

where $J_1 = \left(-\frac{1}{n}\beta_1 - \frac{2}{n}\beta_1^2 - \cdots - \frac{n-1}{n}\beta_1^{n-1}\right)$, $\delta_2 = \lim_{k\to\infty}\sum_{j=1}^{k-1}(\beta_1^n)^j\sqrt{\frac{k}{k-j}} = \frac{\beta_1}{1-\beta_1}$. If needed, we can further bound J_1 by n for simplicity.

Further, ρ_1, ρ_2, ρ_3 are constants satisfying the following conditions for $\forall l = 1, \cdots, d$. Usually, these constants can be different for different problems. In the worst case, we have $0 \le \rho_3 \le \sqrt{n}\rho_1 \le n$. ρ_3 is larger when $\partial_l f_i(x_{k,0})$ are more aligned.

$\rho_1 \ge \frac{\sum_{i=1}^n|\partial_l f_i(x_{k,0})|}{\sqrt{\sum_{i=1}^n|\partial_l f_i(x_{k,0})|^2}};$

$\rho_2 \ge \frac{|\max_i \partial_l f_i(x_{k,0})|^2}{\frac{1}{n}\sum_{i=1}^n|\partial_l f_i(x_{k,0})|^2};$

$\rho_3 \ge \frac{\left|\sum_{i=1}^n \partial_l f_i(x_{k,0})\right|}{\sqrt{\frac{1}{n}\sum_{i=1}^n|\partial_l f_i(x_{k,0})|^2}}.$

The proof of Lemma G.10 is now completed. This also concludes the whole proof of Theorem 3.1.