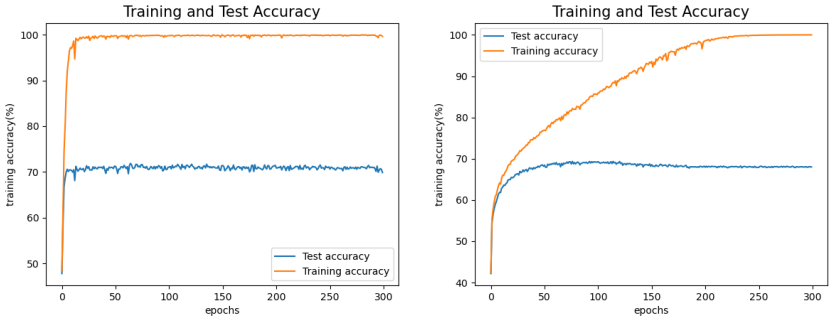
CONTENTS

A ADDITIONAL EXPERIMENTS

A.1 COMPARISON OF RMSPROP AND SGD TRAINING PERFORMANCE

In this section, we provide additional figures comparing the performance of RMSprop and SGD in the experiment mentioned in Section 5. Specifically, we plot the training and test accuracy at each epoch for both algorithms. We also run 300 epochs to guarantee the convergence of SGD. Figure A1 demonstrates typical difference of behaviors of RMSprop and SGD during training. In this case,

the batch size is set to be 32, and β_2 is set to be 0.99 for RMSprop. The result shows that although SGD manages to achieve 100% training accuracy finally, it takes substantially more time for SGD to converge. The test accuracy of SGD is not as high as RMSprop as well in this case.



(a) RMSprop with $\beta_2 = 0.99$ (b) SGD

Figure A1: Performance of RMSprop and SGD with ResNet-18 on CIFAR-10. We choose the batch size to be 32 and run 300 epochs for each experiment. For the RMSprop experiment, we choose $\beta_2 = 0.99$.

A.2 RMSprop Experiments on GANs

We also explored the performance of RMSprop with different β_2's on GANs. In the experiment, we choose CIFAR-10 as the dataset, DCGAN (Radford et al., 2016) as the architecture and Jensen-Shannon distance as the loss metric. In particular, a 5-layer CNN and a 7-layer CNN are used as the generator and the discriminator, respectively. We fix the learning rate as 0.0002 for both the generator and the discriminator and take 100K iterations on all experiments. For each β_2, 3 repetitions are conducted. The quantitative results, measured by FID scores, are shown in Table 4. The performance of RMSprop indeed benefits from larger choice of β_2.

Table 4: FID scores of the generator for RMSprop with different β_2's

β_2	FID score
0.9	41.57±1.00
0.95	39.92±1.10
0.99	38.94±0.77
0.995	39.09±0.57
0.999	38.26±1.16

A.3 A Simple Intuitive Counter example

For a simple example, consider the following problem:

$$\min_x f(x) = x^2 = 10x^2 + \sum_{j=1}^{9}(-x^2) \tag{5}$$

which corresponds to $f_0 = 10x^2$ and $f_1 = f_2 = \cdots = f_9 = -x^2$ in our setting. There is a strong positive gradient signal and many weak negative signals. Similar counter examples are also proposed by other researchers (Chen et al., 2019).

Weighing gradient square in the past and at present, β_2 controls the level of distortion of the gradient signal. Like what Figure A2 illustrated, small β_2 distorts the gradient more. While small β_2 may help v_t approach closer to $\mathbb{E}\left[g_t^2\right]$, thus facilitating optimization, it can also cause divergence.

We also explored the performance of RMSprop with different β_2. Results are listed in Figure A3, with all experiments starting from $x_1 = 1$, $v_0 = 0$. RMSprop behaves very differently when we increase β_2 from 0.9 to 0.99, suggesting a phase transition between the two values.

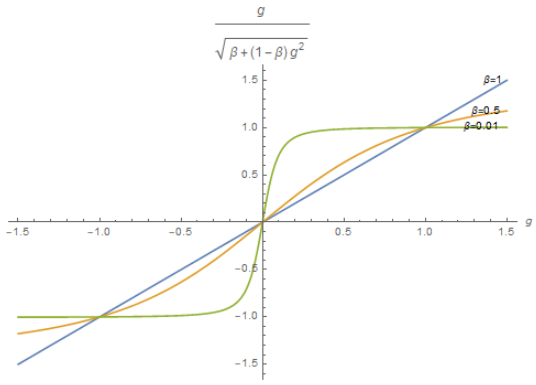
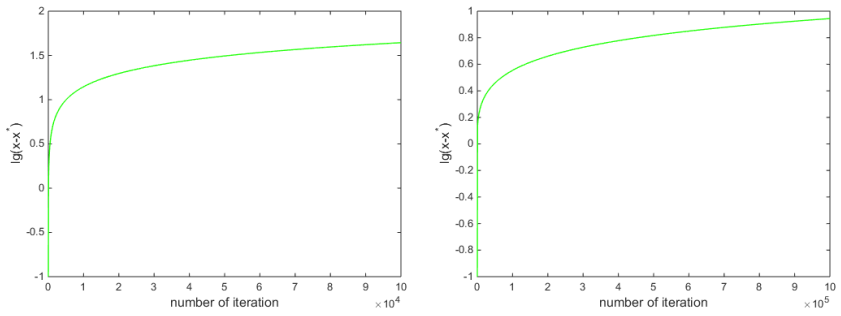
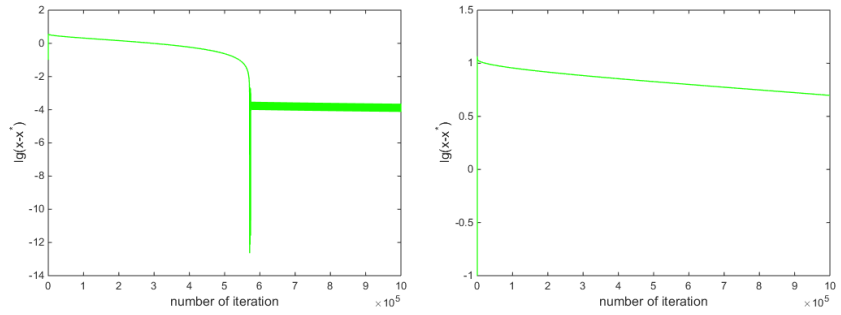


Figure A2: A figure showing how updating direction $\frac{g}{\sqrt{\beta_2+(1-\beta_2)g^2}}$ depends on gradient g with different β_2. When $\beta_2 = 1$, the update is linearly dependent on gradient. The smaller β_2 is, the more it deviates from linearity. After taking expectation, $\mathbb{E}\left[\frac{g}{\sqrt{\beta_2+(1-\beta_2)g^2}}\right]$ can be far away from $\mathbb{E}[g]$



(a) $\beta_2 = 0.1$ (b) $\beta_2 = 0.9$



(c) $\beta_2 = 0.99$ (d) $\beta_2 = 0.999$

Figure A3: Convergence of RMSprop with different β_2. It diverges for small β_2 while converges for big β_2. The phase transition point is between 0.9 and 0.99 for toy problem (5).

A.4 Non-realizable Examples

A non-realizable toy problem is:

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

Then,

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

is lower bounded by $-\frac{1}{9}a^2$. We apply diminishing step size RMSprop, AMSGrad, and SGD to this problem with $a = 10$. The results are shown in Figure A4.

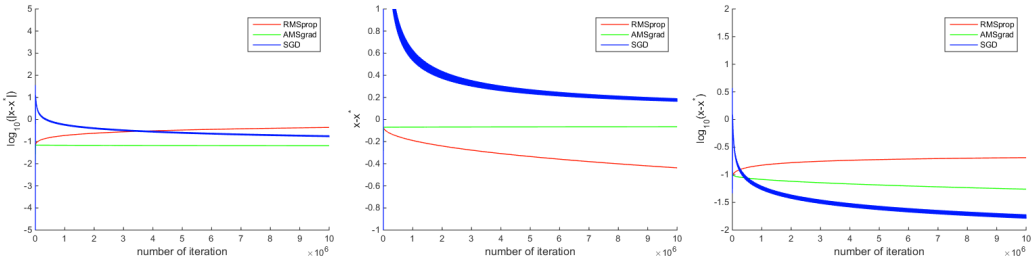
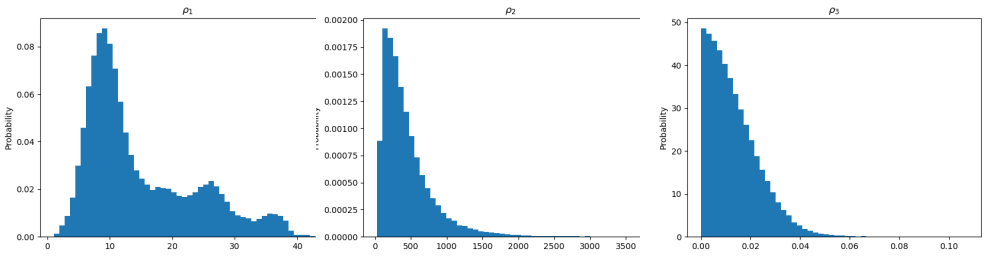


Figure A4: Convergence of 3 popular algorithms on the toy model. They all use diminishing step size $\eta_t = \frac{\eta_1}{\sqrt{t}}$ with $\eta_1 = 0.1$. The left figure plots $\log_{10}|x - x^\star|$ over iterations with the initialization point of $x_0 = 0$; the middle figure plots $x - x^\star$ over iterations with the initialization point of $x_0 = 0$; while the right figure plots $\log_{10}|x - x^\star|$ over iterations with the initialization point of $x_0 = -0.1$.

As Figure A4 shows, SGD converges, which is not surprising since diminishing step size SGD converges even for non-realizable problems. AMSGrad also converges slowly to $x^\star$ since the logarithm of distance to the optimum point decreases steadily. This is also in accordance with the result in Chen et al. (2019). In comparison, RMSprop fails to converge to the global minimum in 1×10^7 iterations: from the plot of $x - x^\star$ it seems that x moves away from the global optimum even if we set the initialization to be $x_0 = x^\star$. Our further experiments show that RMSprop cannot generally converge to the optimal point in this problem: x stays at some distance dependent on β_2 away from the optimum. Moreover, when β_2 becomes closer to zero, the distance is smaller.

A.5 Histogram of ρ_1, ρ_2, and ρ_3

In Appendix F, we introduce three parameters ρ_1, ρ_2, and ρ_3 that will effect the threshold of β_2. It will be helpful to estimate their size in practice. We study a typical image classification problem on MNIST using RMSProp. The batch size is set to 16 (thus $n = 3750$) and β_2 is 0.99, which falls in the convergence regime. Then we calculate the three quantities for each coordinate (each parameter in the neural network) in the beginning epochs of training, and estimate their distribution density.



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

Figure A5: Histograms of three parameters. Probability is normalized.

These results show that on this specific image classification problem, the maximal ρ_1 is upper bounded by $O(\sqrt{n})$ (which is close to the worst-case estimate), the maximal ρ_2 is close to $O(n)$, while ρ_3 is upper bounded by $O(1/\sqrt{n})$. Combining these three practical bounds, the product $\rho_1\rho_2\rho_3 \approx O(n)$.

Thus our bound for β_2 is approximately $1-\mathcal{O}\left(n^{-2}\right)$. Note that our bound $1-\mathcal{O}(n\rho_1\rho_2\rho_3)$ is just the worst-case bound; for practical purposes, we shall consider the average values of $\rho_i's$. The bulk of the distribution of ρ_1 lies at 10, and the build of the distributions of ρ_2 and ρ_3 lie at 200 and 0.01 respectively, thus the product $\rho_1\rho_2\rho_3$ is roughly $20<\sqrt{n}$, thus the suggested $\beta_2 \approx 1-\mathcal{O}\left(n^{-1.5}\right)$. This suggested β_2 is still larger than the practically used β_2, and we leave it as a future work to close this gap.

B THEORETICAL RESULTS OF ORIGINAL RMSPROP/ADAM

Adam introduced in (Kingma & Ba, 2015) starts from zero initialization and uses a bias correction step to rectify the early step sizes:

Algorithm 2 Randomly shuffled Adam with zero initialization and bias correction

Initialize $m_{1,-1}=0$ and $v_{1,-1}=0$
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**
    $m_{k,i}=\beta_1 m_{k,i-1}+(1-\beta_1)\nabla f_{\tau_{k,i}}$
    $v_{k,i}=\beta_2 v_{k,i-1}+(1-\beta_2)\nabla f_{\tau_{k,i}}\circ\nabla f_{\tau_{k,i}}$
    $x_{k,i+1}=x_{k,i}-\frac{\eta_{k*n}}{\sqrt{\frac{v_{k,i}}{1-\beta_2^{nk+i}}}+\epsilon}\circ\frac{m_{l,k,i}}{1-\beta_1^{nk+i}}$
    Break if certain exit condition is satisfied.
  end for
  $x_{k+1,0}=x_{k,n}$
  $v_{k+1,-1}=v_{k,n-1}$
  $m_{k+1,-1}=m_{k,n-1}$
end for
return x

We will refer to Algorithm 2 as the bias corrected version. The bias corrected version differs from Algorithm 1 only in earlier stage of training, thus it should have the same convergence pattern. Indeed, the following theorem shows that most of our results still apply, with some minor modifications in the beginning stages.

Theorem B.1. *(convergence of full-batch Adam with zero initialization and bias correction) For optimization problem (3) with $n=1$, assume that f is gradient Lipschitz continuous with constant L and lower bounded by f^*. Then, for full-batch Adam with diminishing step size $\eta_t=\frac{\eta_1}{\sqrt{t}}$ and any $\beta_1<\sqrt{\beta_2}<1$, we have:*

$$\min_{t\in(t_{init},T]}\|\nabla f_t\|_1\leq\mathcal{O}\left(\frac{\log T}{\sqrt{T}}\right)$$

where $t_{init}=\max\{1,\lceil\log_{\beta_1}\frac{1}{4}\rceil\}$.

Theorem B.2. *(convergence of small-β_1 Adam with zero initialization and bias correction) For optimization problem (3), we assume that f is lower-bounded by f^* and f_j is gradient Lipschitz continuous with constant L for all j. Furthermore, we assume that f_j satisfies assumption* (2)*. Then, for randomly shuffled RMSprop with diminishing step size $\eta_t=\frac{\eta_1}{\sqrt{t}}$ and β_1,β_2 satisfying*

$$T_1(\beta_1,\beta_2)+T_2(\beta_2)<1-\frac{1}{\sqrt{2}} \tag{6}$$

we have

$$\min_{t\in[t_{init},T]}\|\nabla f_{nt}\|_1\leq\mathcal{O}\left(\frac{\log T}{\sqrt{T}}\right)+\mathcal{O}\left(Q_{3,5}\sqrt{D_0}\right)$$

where $Q_{3,5}$ is a constant that approaches 0 in the limit as $\beta_2\to 1$, T_2 is defined in (4)*, and T_1 is defined in* (59)

It's worth mentioning that although t_{init} is nonzero, it is finite, thus not affecting the convergence results.

We interleave the proof of these two results with those of main theorems in the subsequent sections.

C NOTATIONS AND SOME LEMMAS

We are going to use multiple subscripts in the proof. The notations are explained here:

- Vector x is in parameter space $\mathbb{R}^d$. In the full-batch Adam, i.e. $n = 1$, we denote x_t as the value of x at the t-th epoch and $x_{l,t}$ is the l-th component of x_t. If $n > 1$, we denote $x_{k,i}$ as the value of x at the k-th outer loop and i-th inner loop, and $x_{l,k,i}$ as the l-th component of $x_{k,i}$.
- v_t, $v_{l,t}$, $v_{k,i}$, $v_{l,k,i}$, and m_t, $m_{l,t}$, $m_{k,i}$, $m_{l,k,i}$ are defined in a similar way as x.
- We denote η_t as the step-size. We will focus mainly on diminishing step size, especially $\eta_t = \frac{\eta_1}{\sqrt{t}}$. Nevertheless, for each epoch, the step size stays the same.
- Further, to simplify the notation, in the proof of the convergence of randomly shuffled Adam, we define $g_{l,k,i,j} \triangleq \frac{\partial}{\partial x_l} f_j(x)\big|_{x=x_{k,i}}$. We sometimes use f_t as a short-handed notation of $f(x_t)$, ∂_l as a short-handed notation of $\frac{\partial}{\partial x_l}$, and $\nabla f_{l,k,i}$ as a short-handed notation for $\frac{\partial}{\partial x_l} f(x)\big|_{x=x_{k,i}}$.

Then we will discuss some basic properties of Adam and RMSprop.

Lemma C.1. *For any version of RMSprop, in each iteration,*

$$|x_{i,t+1} - x_{i,t}| \leq \frac{\eta_t}{\sqrt{1-\beta_2}}$$

This is obvious since $\sqrt{\beta_2 v_{i,t-1} + (1-\beta_2)(g_{i,t})^2} \geq \sqrt{1-\beta_2}\,|g_{i,t}|$. Since this is an upper bound on the stepsize, the magnitude of ϵ does not matter here. If we use bias correction, the stepsize is contracted by a factor of $\frac{1}{1-\beta_2^t}$: $|x_{i,t+1} - x_{i,t}| \leq \frac{\eta_t\left(1-\beta_2^t\right)}{\sqrt{1-\beta_2}} \leq \frac{\eta_t}{\sqrt{1-\beta_2}}$, hence this lemma still holds.

Lemma C.2. *For any version of Adam, if $\beta_1 < \sqrt{\beta_2}$, in each iteration,*

$$|x_{i,t+1} - x_{i,t}| \leq \eta_t \frac{1}{\sqrt{1-\beta_2}} \frac{1-\beta_1 * bc}{1-\frac{\beta_1}{\sqrt{\beta_2}}}$$

where bc is an index which equals to 0 for the bias corrected version and 1 for the specially initialized version.

Proof. By the update rule of specially initialized Adam,

$$|x_{i,t+1} - x_{i,t}| = \eta_t \frac{|m_{i,t}|}{v_{i,t}}.$$

Since

$$m_{i,t} = (1-\beta_1) \sum_{s=0}^{t-1} g_{i,t-s} \beta_1^s,$$

We have

$$|m_{i,t}| \leq (1-\beta_1) \sum_{s=0}^{t-1} |g_{i,t-s}|\, \beta_1^s$$

Moreover, since $v_{i,t} = \beta_2 v_{i,t-1} + (1-\beta_2) g_{i,t-1}^2$, by recursively expand $v_{i,t}$, we have

$$v_{i,t} = (1-\beta_2) \sum_{s=0}^{t-1} g_{i,t-s}^2 \beta_2^s \geq g_{i,t-s}^2 \beta_2^s$$

for each $s \in \{1, \cdots, t-1\}$. Therefore:

$$\begin{aligned}
|x_{i,t+1} - x_{i,t}| &\leq \eta_t \sum_{s=0}^{t-1} \frac{(1-\beta_1)\,|g_{i,t-s}|}{\sqrt{v_{i,t}}} \beta_1^s \\
&\leq \eta_t \sum_{s=0}^{t-1} \frac{(1-\beta_1)\,|g_{i,t-s}|}{\sqrt{(1-\beta_2)\,\beta_2^s g_{i,t-s}^2}} \beta_1^s \\
&= \eta_t \sum_{s=0}^{t-1} \frac{(1-\beta_1)}{\sqrt{(1-\beta_2)}} \left(\frac{\beta_1}{\sqrt{\beta_2}}\right)^s \\
&\leq \eta_t \sum_{s=0}^{+\infty} \frac{(1-\beta_1)}{\sqrt{(1-\beta_2)}} \left(\frac{\beta_1}{\sqrt{\beta_2}}\right)^s \\
&= \eta_t \frac{1}{\sqrt{(1-\beta_2)}} \frac{1-\beta_1}{1-\frac{\beta_1}{\sqrt{\beta_2}}}.
\end{aligned}$$

If we use bias correction for both m and v, the result simply changes to the step size is multiplied by a factor of $\frac{\sqrt{1-\beta_2^t}}{1-\beta_1^t} \leq \frac{1}{1-\beta_1}$. Therefore we only need to multiply $\eta_t \frac{1}{\sqrt{1-\beta_2}} \frac{1-\beta_1 * bc}{1-\frac{\beta_1}{\sqrt{\beta_2}}}$ with $\frac{1}{1-\beta_1}$ to obtain the bound for bias corrected version. □

Lemma C.3. *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 C.4. *Let $x \in \mathbb{R}$, then for $b \geq 0$, $\min_y\{(x-y)^2 \,|\, |y| \leq b\} \geq x^2 - 2\,|x|\,b$*

Proof. The proof is very straight forward.

If $b \leq |x|$, $\min_{|y|\leq b} (x-y)^2 \geq (|x|-b)^2 = x^2 - 2\,|x|\,b + b^2 \geq x^2 - 2\,|x|\,b$.

If $b \geq |x|$, $\min_{|y|\leq b} (x-y)^2 = 0 \geq x^2 - 2\,|x|\,b$ □

D PROOF OF THEOREM 4.1

We start by determining the upper bound of $v_{t-1,i}$. In the following lemma, we use $\partial_i f_t$ as a simple notation for $\frac{\partial}{\partial x_i} f(x_t)$.

Lemma D.1. *Define $\Delta_t = \frac{\eta_1 L\sqrt{d}}{\sqrt{1-\beta_2}\sqrt{t}}$, where d is the dimension of parameter space. For any coordinate i, if $|\partial_i f_t| \geq \frac{4\sqrt{2}\Delta_t}{1-\beta_2}$, the following holds for RMSprop:*

$$v_{i,t-1} \leq \frac{5}{2} |\partial_i f_t|^2.$$

Remark: if we change the definition of Δ_t to $\Delta_t = \frac{\eta_1 L\sqrt{d}}{\sqrt{1-\beta_2}\sqrt{t}} \frac{1-\beta_1 * br}{1-\frac{\beta_1}{\sqrt{\beta_2}}}$, this lemma holds for Adam.

Proof. From Lemma C.1, the effective step-size of RMSprop is uniformly upper bounded. Therefore, if $\partial_i f_t$ is given, by the per-sample gradient Lipschitz continuity condition, the previous gradients cannot be too large. We thus have the following (under the initial condition $v_{i,0} = (1-\beta_2)^{-1} g_{i,0}^2$):

$$
\begin{aligned}
v_{i,t-1} &= (1-\beta_2)\sum_{j=1}^{t-1}(g_{i,t-j})^2\,\beta_2^{j-1} + \beta_2^{t-1}(g_{i,0})^2 \\
&= (1-\beta_2)\sum_{j=1}^{t-1}|\partial_i f_{t-j}|^2\,\beta_2^{j-1} + \beta_2^{t-1}(g_{i,0})^2 \\
&\le (1-\beta_2)\sum_{j=1}^{t-1}\left(|\partial_i f_t| + \sum_{k=1}^{j}\Delta_{t-k}\right)^2\beta_2^{j-1} + \left(|\partial_i f_t| + \sum_{k=1}^{t-1}\Delta_{t-k}\right)^2\beta_2^{t-1} \\
&\le (1-\beta_2)\sum_{j=1}^{t-1}\left(|\partial_i f_t| + \frac{2j}{\sqrt{t-1}+\sqrt{t-j-1}}\frac{\eta_1 L\sqrt{d}}{\sqrt{1-\beta_2}}\right)^2\beta_2^{j-1} \\
&+ \left(|\partial_i f_t| + \frac{2(t-1)}{\sqrt{t-1}}\frac{\eta_1 L\sqrt{d}}{\sqrt{1-\beta_2}}\right)^2\beta_2^{t-1} \\
&\le (1-\beta_2)\sum_{j=1}^{t-1}\left(|\partial_i f_t| + \frac{2j}{\sqrt{t-1}}\frac{\eta_1 L\sqrt{d}}{\sqrt{1-\beta_2}}\right)^2\beta_2^{j-1} + \left(|\partial_i f_t| + \frac{2(t-1)}{\sqrt{t-1}}\frac{\eta_1 L\sqrt{d}}{\sqrt{1-\beta_2}}\right)^2\beta_2^{t-1} \\
&\le (1-\beta_2)\left(\sum_{j=1}^{t-1}\left(|\partial_i f_t| + \frac{2\sqrt{2}j}{\sqrt{t}}\frac{\eta_1 L\sqrt{d}}{\sqrt{1-\beta_2}}\right)^2\beta_2^{j-1} + \frac{\beta_2^{t-1}}{1-\beta_2}\left(|\partial_i f_t| + \frac{2\sqrt{2}(t-1)}{\sqrt{t}}\frac{\eta_1 L\sqrt{d}}{\sqrt{1-\beta_2}}\right)^2\right) \\
&\le (1-\beta_2)\sum_{j=1}^{\infty}\left(|\partial_i f_t| + \frac{2\sqrt{2}j}{\sqrt{t}}\frac{\eta_1 L\sqrt{d}}{\sqrt{1-\beta_2}}\right)^2\beta_2^{j-1} \\
&\le (1-\beta_2)\sum_{j=1}^{\infty}\left(|\partial_i f_t|^2\,\beta_2^{j-1} + \frac{4\sqrt{2}j}{\sqrt{t}}\frac{\eta_1\sqrt{d}L}{\sqrt{1-\beta_2}}|\partial_i f_t|\,\beta_2^{j-1} + \frac{8j^2\eta_1^2 dL^2}{t(1-\beta_2)}\beta_2^{j-1}\right) \\
&= |\partial_i f_t|^2 + \frac{4\sqrt{2}L\eta_1\sqrt{d}L}{\sqrt{t}}|\partial_i f_t|\frac{1}{(1-\beta_2)^{\frac{3}{2}}} + \frac{8\eta_1^2 dL^2(1+\beta_2)}{t(1-\beta_2)^3} \\
&\le |\partial_i f_t|^2 + |\partial_i f_t|\frac{4\sqrt{2}\Delta_t}{(1-\beta_2)} + \frac{16\Delta_t^2}{(1-\beta_2)^2}
\end{aligned}
\tag{7}
$$

where the first inequality comes from the gradient Lipschitz continuity condition: each iteration changes $x_{i,t-1}$ by at most $\eta_t \frac{1}{\sqrt{1-\beta_2}}$, so the gradient changes by at most Δ_t by Lipschitz continuity; the second inequality comes from Lemma C.3; and the fourth inequality is because $t > 1$. For the fifth inequality, we extend the upper limit of the summation from t to infinity by the relation $\frac{1}{1-\beta_2} = 1 + \beta_@ + \beta_2^2 + \cdots$, which is feasible since we just add some non-negative terms on the right hand side of the inequality. The last equality comes from the following calculation:

$$
\begin{aligned}
(1-\beta_2)\sum_{j=1}^{\infty}\beta_2^{j-1} &= 1 \\
(1-\beta_2)\sum_{j=1}^{\infty}j\beta_2^{j-1} &= \frac{1}{1-\beta_2} \\
(1-\beta_2)\sum_{j=1}^{\infty}j^2\beta_2^{j-1} &= \frac{1+\beta_2}{(1-\beta_2)^2}.
\end{aligned}
$$

Therefore, if $|\partial_i f_t| \geq 4\sqrt{2}\frac{\Delta_t}{1-\beta_2}$, we have:

$$v_{i,t-1} \leq \frac{5}{2}\left|\partial_i f_t\right|^2$$

□

With some minor modifications in the proof, the above lemma also holds for RMSprop with bias correction.

Lemma D.2. *For RMSprop, the following holds for all ∇f_t:*

$$\left\langle \nabla f_t, \eta_t \frac{-\nabla f_t}{\sqrt{\beta_2 v_{t-1} + (1-\beta_2)\nabla f_t^2}} \right\rangle \leq -\eta_t \frac{\|\nabla f_t\|_1}{\sqrt{1+\frac{3}{2}\beta_2}} + \eta_t \frac{4\sqrt{2}\Delta_t}{1-\beta_2}\frac{1}{\sqrt{1+\frac{3}{2}\beta_2}}$$

where $\|\nabla f_t\|_1$ is the 1-norm of vector ∇f_t.

Proof. For any coordinate i, if $|\partial_i f_t| \geq \frac{4\sqrt{2}\Delta_t}{1-\beta_2}$, from Lemma D.1 we can see:

$$\begin{aligned}
\partial_i f_t \frac{\partial_i f_t}{\sqrt{\beta_2 v_{i,t-1} + (1-\beta_2)\partial_i f_t^2}} &\geq \frac{\left|\partial_i f_t\right|^2}{\sqrt{\left(1+\frac{3}{2}\beta_2\right)\left|\partial_i f_t\right|^2}} \\
&= \frac{\left|\partial_i f_t\right|}{\sqrt{1+\frac{3}{2}\beta_2}} \\
&\geq \frac{\left|\partial_i f_t\right|}{\sqrt{1+\frac{3}{2}\beta_2}} - \frac{4\sqrt{2}\Delta_t}{1-\beta_2}\frac{1}{\sqrt{1+\frac{3}{2}\beta_2}}.
\end{aligned} \tag{8}$$

When $|\partial_i f_t| \leq \frac{4\sqrt{2}\Delta_t}{1-\beta_2}$, we have :

$$\begin{aligned}
\partial_i f_t \frac{\partial_i f_t}{\sqrt{\beta_2 v_i^{t-1} + (1-\beta_2)\partial_i f_t^2}} &\geq 0 \\
&\geq \frac{\left|\partial_i f_t\right|}{\sqrt{1+\frac{3}{2}\beta_2}} - \frac{4\sqrt{2}\Delta_t}{1-\beta_2}\frac{1}{\sqrt{1+\frac{3}{2}\beta_2}}.
\end{aligned} \tag{9}$$

Next, we sum up both sides of the inequality by subscript i and multiply it by $-\eta_t$, obtaining:

$$\left\langle \nabla f_t, \eta_t \frac{-\nabla f_t}{\sqrt{\beta_2 v_{t-1} + (1-\beta_2)\nabla f_t^2}} \right\rangle \leq -\eta_t \frac{\|\nabla f_t\|_1}{\sqrt{1+\frac{3}{2}\beta_2}} + \eta_t \frac{4\sqrt{2}d\Delta_t}{1-\beta_2}\frac{1}{\sqrt{1+\frac{3}{2}\beta_2}}.$$

□

Then, to the proof of Theorem 4.1:

Proof. Since f_t is L-Lipschitz, by descent lemma and Lemma C.1,

$$\begin{aligned}
f\left(x_{t+1}\right) - f\left(x_t\right) &\leq \left\langle \nabla f\left(x_t\right), x_{t+1} - x_t \right\rangle + \frac{L}{2}\left\|x_{t+1} - x_t\right\|^2 \\
&\leq -\eta_t \left\langle \nabla f_t, \frac{\nabla f_t}{\sqrt{\beta_2 v_{t-1} + (1-\beta_2)\nabla f_t^2}} \right\rangle + \frac{L}{2}\eta_t^2 \frac{d}{1-\beta_2}.
\end{aligned} \tag{10}$$

We sum both sides of the inequality from t_{init} to T:

$$f\left(x_{T+1}\right) - f\left(x_{t_{init}}\right) \leq -\sum_{t=t_{init}}^{T} \frac{\eta_1}{\sqrt{t}} \left\langle \nabla f_t, \frac{\nabla f_t}{\sqrt{\beta_2 v_{t-1} + (1-\beta_2)\nabla f_t^2}} \right\rangle + \sum_{t=t_{init}}^{T} \eta_1^2 \frac{1}{t} \frac{Ld}{2\left(1-\beta_2\right)}$$

As $f\left(x_{T+1}\right) \geq f^{\star}$, we have:

$$\sum_{t=t_{init}}^{T} \frac{\eta_1}{\sqrt{t}} \left\langle \nabla f_t, \frac{\nabla f_t}{\sqrt{\beta_2 v_{t-1} + (1-\beta_2)\nabla f_t^2}} \right\rangle \leq \sum_{t=t_{init}}^{T} \eta_1^2 \frac{1}{t} \frac{Ld}{2(1-\beta_2)} + f\left(x_{t_{init}}\right) - f^{\star}$$

Next we apply the result from Lemma D.2 to further simplify it as:

$$\begin{aligned} &\sum_{t=t_{init}}^{T} \frac{\eta_1}{\sqrt{t}\sqrt{1+\frac{3}{2}\beta_2}} \left\|\nabla f_t\right\| \\ &\leq \sum_{t=t_{init}}^{T} \eta_1^2 \frac{1}{t} \left(\frac{Ld}{2(1-\beta_2)} + \frac{4\sqrt{2d}Ld}{(1-\beta_2)^{\frac{3}{2}}} \frac{1}{\sqrt{1+\frac{3}{2}\beta_2}} \right) + f\left(x_{t_{init}}\right) - f^{\star} \end{aligned} \tag{11}$$

On the right hand side, we have

$$\sum_{t=t_{init}}^{T} \frac{1}{t} \leq \log \frac{T+1}{t_{init}}$$

On the left hand side, we have

$$\sum_{t=t_{init}}^{T} \frac{1}{\sqrt{t}} \geq 2\left(\sqrt{T} - \sqrt{t_{init}-1}\right)$$

Hence, setting $t_{init} = 1$, we have:

$$\min_{t\in[t_{init},T]} \left\|\nabla f_t\right\|_1 \leq \frac{1}{\sqrt{T}} \left(Q_{1,1} + Q_{2,1}\log\left(T+1\right)\right),$$

where the constants are:

$$Q_{1,1} = \frac{f\left(x_1\right) - f^{*}}{2\eta_1}\sqrt{1+\frac{3}{2}\beta_2} \tag{12}$$

and

$$Q_{2,1} = \frac{\eta_1}{2}\left(\frac{Ld\sqrt{1+\frac{3}{2}\beta_2}}{2(1-\beta_2)} + \frac{4\sqrt{2d}Ld}{(1-\beta_2)^{\frac{3}{2}}} \right) \tag{13}$$

□

E PROOF OF THEOREM 4.2

The proof procedure of Theorem 4.2 is similar to that of Theorem 4.1. As mentioned in Appendix D, if we set $\Delta_t = \frac{\eta_1 L\sqrt{d}(1-bc*\beta_1)}{\sqrt{1-\beta_2}\left(1-\frac{\beta_1}{\sqrt{\beta_2}}\right)\sqrt{t}}$, where bc is set to zero for bias corrected version, and to one for specially initialized version, and keep other notations unchanged, Lemma D.1 holds for Adam. Then it suffices to find a sufficient decrease condition for Adam.

Lemma E.1. *For $t > 1$ and $\beta_1 < \sqrt{\beta_2}$, the following holds:*

$$\left\langle \nabla f_t, \eta_t \frac{-m_t}{\sqrt{v_t}} \right\rangle \leq -\eta_t \frac{\left\|\nabla f_t\right\|_1}{\sqrt{10}} + \eta_t 4\sqrt{2}\frac{\Delta_1 d}{\sqrt{t}} \left(\frac{1}{1-\beta_2} + \frac{2\beta_1}{1-\beta_1}\right)\left(\frac{1}{\sqrt{10}} + \frac{1-\beta_1}{\sqrt{1-\beta_2}\left(1-\frac{\beta_1}{\sqrt{\beta_2}}\right)}\right)$$

Proof. From Lipschitz continuity and Lemma C.2, we know that

$$\left|g_{i,t} - g_{i,t-s}\right| \leq \Delta_{t-1} + \cdots + \Delta_{t-s} \leq \frac{2s}{\sqrt{t-1}}\Delta_1,$$

where the last inequality comes from Lemma C.3. As a result,

$$g_{i,t}g_{i,t-s} \geq g_{i,t}^2 - |g_{i,t}| \frac{2s}{\sqrt{t-1}}\Delta_1.$$

The momentum $m_{i,t}$ is a discounted sum of $g_{i,t}$. Therefore,

$$\begin{aligned}
m_{i,t}g_{i,t} &= (1-\beta_1)\sum_{s=0}^{t-1} g_{i,t}g_{i,t-s}\beta_1^s + g_{i,t}g_{i,0}\beta_1^{t-1} \\
&\geq (1-\beta_1)\sum_{s=0}^{t-1}\left(g_{i,t}^2 - |g_{i,t}|\frac{2s\Delta_1}{\sqrt{t-1}}\right)\beta_1^s + \left(g_{i,t}^2 - |g_{i,t}|\frac{2(t-1)\Delta_1}{\sqrt{t-1}}\right)\beta_1^t \\
&\geq (1-\beta_1)\sum_{s=0}^{\infty}\left(g_{i,t}^2 - |g_{i,t}|\frac{2s\Delta_1}{\sqrt{t-1}}\right)\beta_1^s \\
&= g_{i,t}^2\left(1 - \frac{2\Delta_1}{\sqrt{t-1}\,|g_{i,t}|}\frac{\beta_1}{1-\beta_1}\right).
\end{aligned}$$

If $t > 1$ and $|g_{i,t}| > \frac{8\sqrt{2}\Delta_1}{\sqrt{t}}\frac{\beta_1}{1-\beta_1}$, it reduces to:

$$m_{i,t}g_{i,t} \geq \frac{3g_{i,t}^2}{4}.$$

When we use zero initialization and bias correction, the inner product should be changed to:

$$\begin{aligned}
m_{i,t}g_{i,t} &= (1-\beta_1)\sum_{s=0}^{t-1} g_{i,t}g_{i,t-s}\beta_1^s \\
&\geq (1-\beta_1)\sum_{s=0}^{t-1}\left(g_{i,t}^2 - |g_{i,t}|\frac{2s\Delta_1}{\sqrt{t-1}}\right)\beta_1^s \\
&= g_{i,t}^2\left(1 - \beta_1^t - \frac{2\Delta_1}{\sqrt{t-1}\,|g_{i,t}|}\left(\frac{\beta_1}{1-\beta_1} - \beta_1^t t - \frac{\beta_1^{t+1}}{1-\beta_1}\right)\right) \\
&> g_{i,t}^2\left(1 - \beta_1^t - \frac{2\Delta_1}{\sqrt{t-1}\,|g_{i,t}|}\frac{\beta_1}{1-\beta_1}\right).
\end{aligned}$$

Thus if $t > \max\{\log_{\beta_1}\frac{1}{4}, 1\}$ and $|g_{i,t}| > \frac{8\sqrt{2}\Delta_1}{\sqrt{t}}\frac{\beta_1}{1-\beta_1}$, it reduces to:

$$m_{i,t}g_{i,t} \geq \frac{g_{i,t}^2}{2}.$$

To accommodate the results under 2 settings, we will use the looser bound $\frac{g_{i,t}^2}{2}$ in the following derivations. Combining this bound and the result from the Adam version of Lemma D.1, we come to the conclusion that if $|g_{i,t}| > 4\sqrt{2}\frac{\Delta_1}{\sqrt{t}}\left(\frac{1}{1-\beta_2} + \frac{2\beta_1}{1-\beta_1}\right)$, the following holds:

$$m_{i,t}\frac{g_{i,t}}{\sqrt{v_{i,t}}} \geq \frac{g_{i,t}^2}{2\sqrt{v_{i,t}}} \geq \frac{g_{i,t}^2}{2\sqrt{\frac{5}{2}g_{i,t}^2}} = \frac{|g_{i,t}|}{\sqrt{10}}$$

Thus,

$$g_{i,t}\frac{m_{i,t}}{\sqrt{v_{i,t}}} \geq \frac{|g_{i,t}|}{\sqrt{10}} - 4\sqrt{2}\frac{\Delta_1}{\sqrt{t}}\left(\frac{1}{1-\beta_2} + \frac{2\beta_1}{1-\beta_1}\right)\left(\frac{1}{\sqrt{10}} + \frac{1-\beta_1}{\sqrt{1-\beta_2}\left(1-\frac{\beta_1}{\sqrt{\beta_2}}\right)}\right).$$

Otherwise, if $|g_{i,t}| \leq 4\sqrt{2}\frac{\Delta_1}{\sqrt{t}}\left(\frac{1}{1-\beta_2}+\frac{2\beta_1}{1-\beta_1}\right)$, from Lemma C.2:

$$\begin{aligned} g_{i,t}\frac{m_{i,t}}{\sqrt{v_{i,t}}} &\geq -|g_{i,t}|\frac{1-\beta_1}{\sqrt{1-\beta_2}\left(1-\frac{\beta_1}{\sqrt{\beta_2}}\right)} \\ &= \frac{|g_{i,t}|}{\sqrt{10}} - |g_{i,t}|\left(\frac{1}{\sqrt{10}}+\frac{1-\beta_1}{\sqrt{1-\beta_2}\left(1-\frac{\beta_1}{\sqrt{\beta_2}}\right)}\right) \\ &\geq \frac{|g_{i,t}|}{\sqrt{10}} - 4\sqrt{2}\frac{\Delta_1}{\sqrt{t}}\left(\frac{1}{1-\beta_2}+\frac{2\beta_1}{1-\beta_1}\right)\left(\frac{1}{\sqrt{10}}+\frac{1-\beta_1}{\sqrt{1-\beta_2}\left(1-\frac{\beta_1}{\sqrt{\beta_2}}\right)}\right) \end{aligned}$$

Therefore the inequality holds for every $g_{i,t}$, we can sum up all indices:

$$\sum_{i=1}^{d} g_{i,t}\frac{m_{i,t}}{\sqrt{v_{i,t}}} \geq \frac{\|\nabla f_t\|_1}{\sqrt{10}} - 4\sqrt{2}\frac{\Delta_1 d}{\sqrt{t}}\left(\frac{1}{1-\beta_2}+\frac{2\beta_1}{1-\beta_1}\right)\left(\frac{1}{\sqrt{10}}+\frac{1-\beta_1}{\sqrt{1-\beta_2}\left(1-\frac{\beta_1}{\sqrt{\beta_2}}\right)}\right)$$

This completes the proof. □

Finally, after repeating almost the same procedures at the end of Appendix D, we can prove

$$\min_{t\in[t_{init},T]}\|\nabla f_t\|_1 \leq \frac{1}{\sqrt{T}}\left(Q_{1,2}+Q_{2,2}\log\left(T+1\right)\right),$$

with the following constants:

$$Q_{1,2} = \frac{f\left(x_{t_{init}}\right)-f^*-Q_2^{'}\eta_1^2 Ld\log t_{init}}{2\eta_1}\sqrt{10}$$

and

$$Q_{2,2} = \frac{\eta_1}{2}LdQ_2^{'}$$

where

$$\begin{aligned} Q_2^{'} = \frac{1}{\sqrt{1-\beta_2}}&\left(\frac{1-\beta_1*bc}{1-\frac{\beta_1}{\sqrt{\beta_2}}}\right)\Bigg[\\ &\frac{1}{2\sqrt{1-\beta_2}}\left(\frac{1-\beta_1*bc}{1-\frac{\beta_1}{\sqrt{\beta_2}}}\right)+4\sqrt{2d}\left(\frac{1}{1-\beta_2}+\frac{2\beta_1}{1-\beta_1}\right)\left(\frac{1}{\sqrt{10}}+\frac{1-\beta_1*bc}{\sqrt{1-\beta_2}\left(1-\frac{\beta_1}{\sqrt{\beta_2}}\right)}\right)\Bigg] \end{aligned}$$

F PROOF OF THEOREM 4.3

It takes even more complicated calculations to prove the convergence of the randomly shuffled version of RMSprop, but the rationale is still to track the magnitude of v and calculate the diminishing speed of $\|\nabla f\|_1$. The derivations in this section can be considered as a template: following similar guidelines, the convergence of Adam(with small β_1) can also be proved. We will prove Theorem 4.4 in the later section.

At the begging of this section, we introduce several new notations.

We define $g_{l,k}^b$ as the largest coordinate of the gradient in the beginning of the k-th epoch:

$$\begin{aligned} b_{l,k} &= \arg\max_{i\in\{0,\cdots,n-1\}}\left|g_{l,k,0,i}\right|, \\ g_{l,k}^b &= g_{l,k,0,b_{l,k}}. \end{aligned}$$

We next introduce three constants to characterize the distribution of gradient norms among different batches i.

ρ_1 is a constant that measures similarity of gradient norms:

$$\rho_1 \geq \frac{\sum_{i=1}^{n} |g_{l,k,0,i}|}{\sqrt{\sum_{i=1}^{n} |g_{l,k,0,i}|^2}} \tag{14}$$

for all l and k. It's easy to verify that $1 \leq \rho_1 \leq \sqrt{n}$: the lower bound can be derived from Cauchy inequality, and the upper bound from Cauchy-Schwartz inequality that $\frac{u^T v}{\sqrt{\|u\|^2 \|v\|^2}} \leq 1$, where we set $u = (|g_{l,k,0,0}|, ... |g_{l,k,0,n}|)^T$ and $v = (1, 1, ...1)^T$

ρ_2 is a constant that represents the ratio of the largest gradient norm to average gradient norm:

$$\rho_2 \geq \frac{\left|g_{l,k}^b\right|^2}{\frac{1}{n}\sum_{i=1}^{n} |g_{l,k,0,i}|^2} \tag{15}$$

We can see $1 \leq \rho_2 \leq n$, and ρ_2 is lower when the gradients norm are more homogeneous.

ρ_3 is a constant that represents the the ratio of gradient norm to noisy gradient root mean square:

$$\rho_3 \geq \frac{|\sum_{i=1}^{n} g_{l,k,0,i}|}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} |g_{l,k,0,i}|^2}} \tag{16}$$

it's easy to see that $0 \leq \rho_3 \leq \sqrt{n}\rho_1 \leq n$. ρ_3 is alrger when $g_{l,k,0,i}$'s are more aligned.

Lemma F.1. *If the l-th component of the gradient ∇f satisfies*

$$|\partial_l f(x_{k,0})| \geq \frac{\eta_1 L \sqrt{dn} n^2}{\sqrt{k}\sqrt{1-\beta_2}} \left(\frac{32\sqrt{2}}{(1-\beta_2^n)\beta_2^n}\right)$$

we have

$$\frac{v_{l,k,0}}{\frac{1}{n}\sum_i g_{l,k,0,i}^2} \geq \frac{\beta_2^n}{2}$$

This lemma gives us a lower bound of v when the gradient norm is large enough.

Proof. We still define

$$\Delta_t = \frac{L d^{\frac{1}{2}}}{\sqrt{1-\beta_2}} \frac{\eta_1}{\sqrt{t}}.$$

Assume $M (M < k)$ is the largest integer satisfying

$$\sum_{j=1}^{M} \sqrt{n} \Delta_{k-j} \leq \left|g_{l,k}^b\right|.$$

M is greater than 1 and smaller than k, so such M must exist. By definition and Lemma C.3, $\left|g_{l,k}^b\right|$ is lower bounded by:

$$\left|g_{l,k}^b\right| \geq \frac{\sqrt{d} L \eta_1 \sqrt{n} n}{\sqrt{1-\beta_2}} \frac{M}{\sqrt{k}} \tag{17}$$

Since $v_{l,k,0}$ could be considered as exponential averaging of $g_{l,k,i}^2$:

$$v_{l,k,0} = (1-\beta_2)\left(g_{l,k,0,\tau_{k,0}}^2 + g_{l,k-1,n-1,\tau_{k-1,n-1}}^2 \beta_2 + g_{l,k-1,n-2,\tau_{k-1,n-2}}^2 \beta_2^2 + \cdots\right)$$

to estimate a lower bound of $v_{l,k,0}$ we have to find a lower bound for gradient norm in the summand. For a very loose estimate, we use Lemma C.4 to derive such lower bound . As $|g_{l,k}^b|$ is assumed to be sufficiently large, continuity restricts the range of gradient norm: in each iteration, all coordinates change by at most $\frac{\eta_{nk}}{\sqrt{1-\beta_2}}$, so change in the gradient norm is also bounded due to Lipschitz condition. Since one epoch contains n iterations, each coordinate shifts by at most $\frac{\eta_1 n}{\sqrt{nk}}$ in one epoch.

Recall that we explicitly required M<k, which contains two cases: $M < k-1$ and $M = k-1$, we will discuss them separately.

Case 1: When M<k-1, the definition of M indicates that $\sum_{j=1}^{M+1} \sqrt{n}\Delta_{k-j} > \left|g_{l,k}^b\right|$. This implies (by Lemma C.3):

$$\left|g_{l,k}^b\right| \leq \frac{\Delta_1 2\left(M+2\right)\sqrt{n}}{\sqrt{k+1}} < \frac{\Delta_1 4M\sqrt{n}}{\sqrt{k}}$$

Since we assumed $\left|\partial_l f\left(x_{k,0}\right)\right| \geq \frac{\eta_1 Ln\sqrt{dn}}{\sqrt{k}\sqrt{1-\beta_2}}\left(\frac{32\sqrt{2}}{1-\beta_2^n}\right)$, the largest coordinate of the gradient is also lower bounded by $\left|g_{l,k}^b\right| \geq \frac{\eta_1 L\sqrt{dn}}{\sqrt{k}\sqrt{1-\beta_2}}\left(\frac{32\sqrt{2}}{1-\beta_2^n}\right)$ Therefore

$$M \geq \frac{8\sqrt{2}}{1-\beta_2^n}$$

For each i, if $\tau_{k-j,r} = i$, by Lipschitz continuity we know that:

$$\left|\left|g_{l,k-j,r,\tau_{k-j,r}}\right| - \left|g_{l,k,0,i}\right|\right| \leq n\sum_{p=1}^{j}\Delta_{(k-p)n}$$

Then a natural lower bound on $\left|g_{l,k-j,r,\tau_{k-j,r}}\right|$ is:

$$\left|g_{l,k,0,\tau_{k-j,r}}\right| \geq \begin{cases} \left|g_{l,k,0,i}\right| - n\sum_{p=1}^{j}\Delta_{(k-p)n} \geq 0 \text{ if } n\sum_{p=1}^{j}\Delta_{(k-p)n} \leq \left|g_{l,k,0,i}\right| \\ 0 \text{ if } n\sum_{p=1}^{j}\Delta_{(k-p)n} \geq \left|g_{l,k,0,i}\right| \end{cases} \tag{18}$$

Combining this with Lemma C.4, we have:

$$\left|g_{l,k-j,r,\tau_{k-j,r}}\right|^2 \geq \begin{cases} g_{l,k,0,i}^2 - 2n\left|g_{l,k,0,i}\right|\sum_{p=1}^{j}\Delta_{(k-p)n} \text{ under all circumstances} \\ 0 \text{ if } n\sum_{p=1}^{j}\Delta_{(k-p)n} \geq \left|g_{l,k,0,i}\right| \end{cases} \tag{19}$$

We use the first bound in (19) for $j \leq M$ and the second bound for $j > M$, thus the lower bound of v is given by (we omitted the initialization of v in this case):

$$\begin{aligned} v_{l,k,0} \geq & \left(1-\beta_2\right)\sum_{i=1}^{n}\sum_{j=1}^{M}\left(\left|g_{l,k,0,i}\right|^2 - 2\left|g_{l,k,0,i}\right| n\sum_{p=1}^{j}\Delta_{(k-p)n}\right)\beta_2^{nj} \\ \geq & \left(1-\beta_2\right)\sum_{i=1}^{n}\sum_{j=1}^{M}\left(\left|g_{l,k,0,i}\right|^2 - 2\left|g_{l,k,0,i}\right|\Delta_1\frac{2j\sqrt{n}}{\sqrt{k-1}}\right)\beta_2^{nj} \\ = & \left(1-\beta_2\right)\beta_2^n\left[\sum_{i=1}^{n}\left|g_{l,k,0,i}\right|^2\frac{1-\beta_2^{nM}}{1-\beta_2^n} - 4\Delta_1\frac{\sum_{i=1}^{n}\left|g_{l,k,0,i}\right|\sqrt{n}}{\sqrt{k-1}}\frac{1-\beta_2^{nM}-M\beta_2^{nM}\left(1-\beta_2^n\right)}{\left(1-\beta_2^n\right)^2}\right] \\ = & \frac{1-\beta_2}{1-\beta_2^n}\beta_2^n\left[\sum_{i=1}^{n}\left|g_{l,k,0,i}\right|^2\left(1-\beta_2^{nM}\right) - 4\Delta_1\frac{\sum_{i=1}^{n}\left|g_{l,k,0,i}\right|\sqrt{n}}{\sqrt{k-1}}\left(\frac{1-\beta_2^{nM}}{1-\beta_2^n} - M\beta_2^{nM}\right)\right] \end{aligned} \tag{20}$$

where we applied Lemma C.3 for the third inequality. The second last equality used the relation:

$$\sum_{j=1}^{M} j\beta_2^{nj} = \beta_2^n\frac{1-\beta_2^{nM}-M\beta_2^{nM}\left(1-\beta_2^n\right)}{\left(1-\beta_2^n\right)^2}$$

Then, from (17) and the definition of ρ_1 in (14), $\frac{\sum_{i=1}^n |g_{l,k,0,i}|^2}{\sum_{i=1}^n |g_{l,k,0,i}|} \geq \frac{1}{\rho_1^2} \sum_{i=1}^n |g_{l,k,0,i}| \geq \frac{1}{n} \left|g_{l,k}^b\right| \geq \frac{\sqrt{d}L\eta_1\sqrt{n}}{\sqrt{1-\beta_2}} \frac{M}{\sqrt{k}} = \sqrt{n}M\Delta_1/\sqrt{k}$, thus:

$$\begin{aligned} v_{l,k,0} \geq & \frac{1-\beta_2}{1-\beta_2^n} \sum_{i=1}^n |g_{l,k,0,i}|^2 \beta_2^n \left[\left(1-\beta_2^{nM}\right) - 4\frac{1}{M}\sqrt{\frac{k}{k-1}} \left(\frac{1-\beta_2^{nM}}{1-\beta_2^n} - M\beta_2^{nM} \right) \right] \\ \geq & \frac{1-\beta_2}{1-\beta_2^n} \sum_{i=1}^n |g_{l,k,0,i}|^2 \beta_2^n \left[1 - \frac{4\sqrt{2}}{M} \frac{1-\beta_2^{nM}}{1-\beta_2^n} \right] \\ \geq & \frac{1}{n} \sum_{i=1}^n |g_{l,k,0,i}|^2 \beta_2^n \left(1 - \frac{4\sqrt{2}}{M} \frac{1}{1-\beta_2^n} \right) \end{aligned} \tag{21}$$

We already showed $M \geq \frac{8\sqrt{2}}{1-\beta_2^n}$, hence $\frac{v_{l,k,0}}{\frac{1}{n}\sum_{i=1}^n |g_{l,k,0,i}|^2} \geq \frac{\beta_2^n}{2}$.

Case 2: On the other hand, if $M = k-1$, this means

$$\left|g_{l,k}^b\right| \geq \frac{\sqrt{d}L\eta_1\sqrt{n}}{\sqrt{1-\beta_2}} \frac{k-1}{\sqrt{k}}$$

we can use Lemma C.4 to rederive equation (20) as:

$$\begin{aligned} v_{l,k,0} \geq & (1-\beta_2) \sum_{i=1}^n \sum_{j=1}^{k-1} \left(|g_{l,k,0,i}|^2 - 2|g_{l,k,0,i}| \, n \sum_{p=1}^j \Delta_{(k-p)n} \right) \beta_2^{nj} \\ & + \frac{1}{n} \sum_{i=1}^n \left(|g_{l,k,0,i}|^2 - 2|g_{l,k,0,i}| \, n \sum_{p=1}^{k-1} \Delta_{(k-p)n} \right) \beta_2^{n(k-1)} \\ \geq & (1-\beta_2) \sum_{i=1}^n \sum_{j=1}^{k-1} \left(|g_{l,k,0,i}|^2 - 2|g_{l,k,0,i}| \, \Delta_1 \frac{2j\sqrt{n}}{\sqrt{k-1}} \right) \beta_2^{nj} \\ & + \frac{1}{n} \sum_{i=1}^n \left(|g_{l,k,0,i}|^2 - 2|g_{l,k,0,i}| \, \Delta_1 \frac{2(k-1)\sqrt{n}}{\sqrt{k-1}} \right) \beta_2^{n(k-1)} \end{aligned} \tag{22}$$

where the initialization term is taken into consideration. We estimate the summations separately.

$$\begin{aligned} & (1-\beta_2) \sum_{j=1}^{k-1} |g_{l,k,0,i}|^2 \beta_2^{nj} + \frac{1}{n} \beta_2^{n(k-1)} |g_{l,k,0,i}|^2 \\ = & (1-\beta_2) \left(\sum_{j=1}^{k-1} |g_{l,k,0,i}|^2 \beta_2^{nj} + \frac{\beta_2^{n(k-1)}}{n} \left(1+\beta_2+\beta_2^2+...\right) |g_{l,k,0,i}|^2 \right) \\ \geq & (1-\beta_2) \left(\sum_{j=1}^{k-1} |g_{l,k,0,i}|^2 \beta_2^{nj} + \beta_2^{n(k-1)} \left(\beta_2^n+\beta_2^{2n}+...\right) |g_{l,k,0,i}|^2 \right) \\ = & \frac{1-\beta_2}{1-\beta_2^n} \beta_2^n |g_{l,k,0,i}|^2 \end{aligned}$$

and also

$$
\begin{aligned}
&4\left|g_{l,k,0,i}\right|\Delta_1\frac{\sqrt{n}}{\sqrt{k-1}}\left((1-\beta_2)\sum_{j=1}^{k-1}j\beta_2^{nj}+\frac{1}{n}\beta_2^{n(k-1)}(k-1)\right)\\
&\leq 4\left|g_{l,k,0,i}\right|\Delta_1\frac{\sqrt{n}}{\sqrt{k-1}}(1-\beta_2)\left(\sum_{j=1}^{k-1}j\beta_2^{nj}+\frac{1}{n}\beta_2^{n(k-1)}k\left(1+\beta_2+\beta_2^2+...\right)\right)\\
&\leq 4\left|g_{l,k,0,i}\right|\Delta_1\frac{\sqrt{n}}{\sqrt{k-1}}(1-\beta_2)\beta_2^{-n}\left(\sum_{j=1}^{k-1}j\beta_2^{nj}+\beta_2^{nk}k\left(1+\beta_2^n+\beta_2^{2n}+...\right)\right)\\
&\leq 4\left|g_{l,k,0,i}\right|\Delta_1\frac{\sqrt{n}}{\sqrt{\infty}}(1-\beta_2)\beta_2^{-n}\left(\sum_{j=1}^{k-1}j\beta_2^{nj}+\beta_2^{nk}k\left(1+\beta_2^n+\beta_2^{2n}+...\right)\right)\\
&=4\left|g_{l,k,0,i}\right|\Delta_1\frac{\sqrt{n}}{\sqrt{k-1}}(1-\beta_2)\beta_2^{-n}\beta_2^n\frac{1}{\left(1-\beta_2^n\right)^2}
\end{aligned}
$$

As a result:

$$
\begin{aligned}
v_{l,k,0}\geq&\frac{1-\beta_2}{1-\beta_2^n}\beta_2^n\left[\sum_{i=1}^{n}\left|g_{l,k,0,i}\right|^2-4\Delta_1\frac{\sum_{i=1}^{n}\left|g_{l,k,0,i}\right|\sqrt{n}}{\sqrt{k-1}}\frac{\beta_2^{-n}}{1-\beta_2^n}\right]\\
=&\frac{1}{n}\sum_{i=1}^{n}\left|g_{l,k,0,i}\right|^2\beta_2^n\left(1-\frac{4\Delta_1}{\sum_{i=1}^{n}\left|g_{l,k,0,i}\right|\sqrt{k-1}}\frac{\sqrt{n}\beta_2^{-n}\rho_1^2}{1-\beta_2^n}\right)
\end{aligned}
$$

Since we have $\frac{1}{\rho_1^2}\sum_{i=1}^{n}\left|g_{l,k,0,i}\right| \geq \frac{1}{n}\left|g_{l,k}^b\right| \geq \frac{\eta_1 L\sqrt{dn}\beta_2^{-n}}{\sqrt{k}\sqrt{1-\beta_2}}\left(\frac{32\sqrt{2}}{1-\beta_2^n}\right)$, the inequality $\frac{v_{l,k,0}}{\frac{1}{n}\sum_{i=1}^{n}\left|g_{l,k,0,i}\right|^2}\geq\frac{\beta_2^n}{2}$ holds this case as well. The proof is complete.

With the definition of ρ_3 in equation (16), we can then extend our result to $\frac{v_{l,k,0}}{\left(\partial_l f(x_{k,0})\right)^2}\geq\frac{\beta_2^n}{2\rho_3^2}$. For the bias corrected version, we need one more constraint $k>\frac{8\sqrt{2}}{1-\beta_2^n}+1$ to reach the same conclusion. The derivation is very similar: the only difference is that we don't have to consider case 2. □

Next we will try to find an upper bound for v.

Lemma F.2. *Assume that*

$$
\sum_{j=0}^{n-1}\left\|\nabla f_j\right\|_2^2\leq D_1\left\|\nabla f\right\|_2^2+D_0.
$$

Given k, we set α as the index of the coordinate with the greatest gradient:

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

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:

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

Proof. The rationale of this proof is similar to Lemma D.1. Recall that

$$
v_{\alpha,k,0}=(1-\beta_2)\left(g_{\alpha,k,0,\tau_{k,0}}^2+g_{\alpha,k-1,n-1,\tau_{k-1,n-1}}^2\beta_2+g_{\alpha,k-1,n-2,\tau_{k-1,n-2}}^2\beta_2^2+\cdots\right)
$$

As α is the index of the greatest gradient component, we have:

$$
\left\|\nabla f\right\|_2^2\leq d\left|\frac{\partial}{\partial x_\alpha}f\left(x_{k,0}\right)\right|^2.
$$

Thus our assumption $\sum_{j=0}^{n-1} \|\nabla f_j\|_2^2 \leq D_1 \|\nabla f\|_2^2 + D_0$ leads to

$$\sum_{j=0}^{n-1} \|g_{k,0,j}\|_2^2 = \sum_{j=0}^{n-1} \|\nabla f_j\left(x_{k,0}\right)\|_2^2 \leq D_1 d \left|\frac{\partial}{\partial x_\alpha} f\left(x_{k,0}\right)\right|^2 + D_0.$$

Specifically, for all $j \in \{0, 1, 2, \cdots, n-1\}$,

$$\left|g_{\alpha,k,0,j}\right|^2 \leq \|g_{k,0,j}\|_2^2 \leq D_1 d \left|\frac{\partial}{\partial x_\alpha} f\left(x_{k,0}\right)\right|^2 + D_0$$

To estimate an upper bound for $v_{\alpha,k,0}$, we will first determine an upper bound for each $g_{\alpha,k-j,i,\tau_{k-j,i}}$ with $k-1 \geq j > 0$:

$$\begin{aligned}\left|g_{\alpha,k-j,i,\tau_{k-j,i}}\right| &\leq \left|g_{\alpha,k,0,\tau_{k-j,i}}\right| + \sum_{q=0}^{j} \min\{n, jn-i-nq\}\Delta_{n(k-q)} \\ &\leq \sqrt{\left|\frac{\partial}{\partial x_\alpha} f\left(x_{k,0}\right)\right|^2 + \frac{D_0}{D_1 d}}\sqrt{D_1 d} + \sum_{q=1}^{j} \min\{n, jn-i-nq\}\Delta_{n(k-q)}\end{aligned} \tag{23}$$

The inequality is a result of Lipschitz continuity and Lemma C.1. Summing them up and combining these inequalities, we have

$$
\begin{aligned}
v_{\alpha,k,0} =& (1-\beta_2)\left(\left(g_{\alpha,k,0,\tau_{k,0}}\right)^2+\sum_{j=1}^{k}\sum_{i=0}^{n-1}\left(g_{\alpha,k-j,n-1-i,\tau_{k-j,n-1-i}}\right)^2\beta_2^{n(j-1)+i+1}\right)+\beta_2^{n(k-1)}\left(g_{\alpha,1,-1,J}\right)^2\\
\leq& (1-\beta_2)\left(\left(g_{\alpha,k,0,\tau_{k,0}}\right)^2+\beta_2\left(\left|g_{\alpha,k,0,\tau_{k-1,n-1}}\right|+\frac{\Delta_1}{\sqrt{n\left(k-1\right)}}\right)^2\right.\\
&\left.+\beta_2^2\left(\left|g_{\alpha,k,0,\tau_{k-1,n-2}}\right|+\frac{2\Delta_1}{\sqrt{n\left(k-1\right)}}+\right)^2+\cdots\right)+\beta_2^{n(k-1)}\left(\left|g_{\alpha,1,-1,J}\right|+\sum_{t=1}^{n(k-1)-1}\Delta_{n(k-1)-t}\right)^2\\
\leq& (1-\beta_2)\sum_{j=0}^{n(k-1)-1}\left(\sqrt{\left|\partial_\alpha f\left(x_{k,0}\right)\right|^2+\frac{D_0}{D_1 d}}\sqrt{D_1 d}+\sum_{t=1}^{j}\Delta_{n(k-1)-t}\right)^2\beta_2^j\\
&+\left(\sqrt{\left|\partial_\alpha f\left(x_{k,0}\right)\right|^2+\frac{D_0}{D_1 d}}\sqrt{D_1 d}+\sum_{t=1}^{n(k-1)-1}\Delta_{n(k-1)-t}\right)^2\beta_2^{n(k-1)}\\
\leq& (1-\beta_2)\sum_{j=0}^{n(k-1)-1}\left(\sqrt{\left|\partial_\alpha f\left(x_{k,0}\right)\right|^2+\frac{D_0}{D_1 d}}\sqrt{D_1 d}+\frac{2j}{\sqrt{n(k-1)-1}+\sqrt{n(k-1)-j-1}}\frac{\eta_1\sqrt{d}L}{\sqrt{1-\beta_2}}\right)^2\beta_2^j\\
&+\left(\sqrt{\left|\partial_\alpha f\left(x_{k,0}\right)\right|^2+\frac{D_0}{D_1 d}}\sqrt{D_1 d}+\frac{2(n(k-1)-1)}{\sqrt{n(k-1)-1}+\sqrt{n(k-1)-j-1}}\frac{\eta_1\sqrt{d}L}{\sqrt{1-\beta_2}}\right)^2\beta_2^{n(k-1)}\\
\leq& (1-\beta_2)\sum_{j=0}^{n(k-1)-1}\left(\sqrt{\left|\partial_\alpha f\left(x_{k,0}\right)\right|^2+\frac{D_0}{D_1 d}}\sqrt{D_1 d}+\frac{2j}{\sqrt{n(k-1)-1}}\frac{\eta_1\sqrt{d}L}{\sqrt{1-\beta_2}}\right)^2\beta_2^j\\
&+\left(\sqrt{\left|\partial_\alpha f\left(x_{k,0}\right)\right|^2+\frac{D_0}{D_1 d}}\sqrt{D_1 d}+\frac{2(n(k-1)-1)}{\sqrt{n(k-1)-1}}\frac{\eta_1\sqrt{d}L}{\sqrt{1-\beta_2}}\right)^2\beta_2^{n(k-1)}\\
\leq& (1-\beta_2)\sum_{j=0}^{n(k-1)-1}\left(\sqrt{\left|\partial_\alpha f\left(x_{k,0}\right)\right|^2+\frac{D_0}{D_1 d}}\sqrt{D_1 d}+\frac{2\sqrt{2}j}{\sqrt{nk}}\frac{\eta_1\sqrt{d}L}{\sqrt{1-\beta_2}}\right)^2\beta_2^j\\
&+\left(\sqrt{\left|\partial_\alpha f\left(x_{k,0}\right)\right|^2+\frac{D_0}{D_1 d}}\sqrt{D_1 d}+\frac{2\sqrt{2}(n(k-1)-1)}{\sqrt{nk}}\frac{\eta_1\sqrt{d}L}{\sqrt{1-\beta_2}}\right)^2\beta_2^{n(k-1)}\left(1-\beta_2\right)\left(1+\beta_2+\beta_2^2+...\right)\\
\leq& (1-\beta_2)\sum_{j=0}^{\infty}\left(\left(\left|\partial_\alpha f\left(x_{k,0}\right)\right|^2+\frac{D_0}{D_1 d}\right)D_1 d\beta_2^j+\frac{4\sqrt{2}j}{\sqrt{nk}}\frac{\eta_1 dL}{\sqrt{1-\beta_2}}\sqrt{\left|\partial_\alpha f\left(x_{k,0}\right)\right|^2+\frac{D_0}{D_1 d}}\sqrt{D_1}\beta_2^j\right.\\
&\left.+\frac{8j^2\eta_1^2 dL^2}{nk\left(1-\beta_2\right)}\beta_2^j\right)\\
=& D_1 d\left(\left(\left|\partial_\alpha f\left(x_{k,0}\right)\right|^2+\frac{D_0}{D_1 d}\right)+4\sqrt{2}L\eta_1\frac{\sqrt{\left|\partial_\alpha f\left(x_{k,0}\right)\right|^2+\frac{D_0}{D_1 d}}}{\sqrt{D_1 nk}}\frac{\beta_2}{\left(1-\beta_2\right)^{\frac{3}{2}}}+\frac{8\eta_1^2L^2\left(1+\beta_2\right)\beta_2}{nkD_1\left(1-\beta_2\right)^3}\right)\\
\leq& D_1 d\left(\left(\left|\partial_\alpha f\left(x_{k,0}\right)\right|^2+\frac{D_0}{D_1 d}\right)+4\sqrt{2}L\eta_1\frac{\sqrt{\left|\partial_\alpha f\left(x_{k,0}\right)\right|^2+\frac{D_0}{D_1 d}}}{\sqrt{D_1 nk}}\frac{1}{\left(1-\beta_2\right)^{\frac{3}{2}}}+\frac{16\eta_1^2L^2}{nkD_1\left(1-\beta_2\right)^3}\right)
\end{aligned}
\tag{24}
$$

where the first inequality is due to Lipschitz continuity, the second holds because of relation (23) and the fact that

$$\frac{i}{\sqrt{k\left(n-j\right)}} \leq \sum_{l=1}^{i} \frac{1}{\sqrt{k\left(n-j\right)-l}},$$

the third comes from Lemma C.3, the fourth comes from $\frac{\sqrt{nk}}{\sqrt{n(k-1)-1}} \leq \sqrt{2}$. Therefore, if $\sqrt{\left|\partial_\alpha f\left(x_{k,0}\right)\right|^2 + \frac{D_0}{D_1 d}} \geq 4\sqrt{2}\frac{\Delta_1}{(1-\beta_2)\sqrt{D_1 nkd}}$, we have

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

In the bias corrected version, the main result of this lemma still holds, with some minor modifications during the proof. □

Lemma F.3. *Under the same condition of Lemma F.1, there exists a lower bound of $\frac{1}{\sqrt{v_{l,k,i}}}$ given below:*

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

Remark: ρ_2 can be replaced by its upper bound n.

Proof. From the convexity of function $\frac{1}{\sqrt{1+x}}$, we have

$$\frac{1}{\sqrt{1+x}} \geq 1 - \frac{x}{2}$$

for $x > -1$. Applying this to $\frac{1}{\sqrt{v_{l,k,i}}}$ yields

$$\frac{1}{\sqrt{v_{l,k,i}}} = \frac{1}{\sqrt{v_{l,k,0} + \left(v_{l,k,i} - v_{l,k,0}\right)}} \geq \frac{1}{\sqrt{v_{l,k,0}}}\left(1 - \frac{v_{l,k,i} - v_{l,k,0}}{2v_{l,k,0}}\right) \geq \frac{1}{\sqrt{v_{l,k,0}}}\left(1 - \frac{\left|v_{l,k,i} - v_{l,k,0}\right|}{2v_{l,k,0}}\right).$$

Note that

$$v_{l,k,i} = v_{l,k,0}\beta_2^i + \left(g_{l,k,1,\tau_{k,1}}\right)^2 \beta_2^{i-1}\left(1-\beta_2\right) + \cdots + \left(g_{l,k,i,\tau_{k,i}}\right)^2\left(1-\beta_2\right),$$

the difference of $v_{l,k,i}$ and $v_{l,k,0}$ is given by

$$\begin{aligned} v_{l,k,i} - v_{l,k,0} = &\left(1-\beta_2\right)\Big[\left(g_{l,k,i,\tau_{k,i}}\right)^2 - v_{l,k,0} + \beta_2\left(\left(g_{l,k,i-1,\tau_{k,i-1}}\right)^2 - v_{l,k,0}\right) + \cdots \\ &+ \beta_2^{i-1}\left(\left(g_{l,k,1,\tau_{k,1}}\right)^2 - v_{l,k,0}\right)\Big] \end{aligned}$$

where we have applied the relation $v_{l,k,0} - \beta_2^i v_{l,k,0} = \left(1-\beta_2\right)\left(1+\beta_2+\cdots+\beta_2^{i-1}\right)v_{l,k,0}$. By the definition of $g_{l,k}^b$ and Lipschitz continuity, the following inequality holds:

$$\begin{aligned} &\frac{1}{v_{l,k,0}}\left(\left(g_{l,k,i,\tau_{k,i}}\right)^2 + \beta_2\left(g_{l,k,i-1,\tau_{k,i-1}}\right)^2 + \cdots + \beta_2^{i-1}\left(g_{l,k,1,\tau_{k,1}}\right)^2\right) \\ \leq &\frac{1}{v_{l,k,0}}\Bigg(\left(g_{l,k,0,\tau_{k,i}} + i\frac{\Delta_1}{\sqrt{nk}}\right)^2 + \beta_2\left(g_{l,k,0,\tau_{k,i-1}} + (i-1)\frac{\Delta_1}{\sqrt{nk}}\right)^2 + \cdots \\ &+ \beta_2^{i-1}\left(g_{l,k,0,\tau_{k,1}} + \frac{\Delta_1}{\sqrt{nk}}\right)^2\Bigg) \\ \leq &\frac{1}{v_{l,k,0}}\left(\left(\left|g_{l,k}^b\right| + i\frac{\Delta_1}{\sqrt{nk}}\right)^2 + \beta_2\left(\left|g_{l,k}^b\right| + (i-1)\frac{\Delta_1}{\sqrt{nk}}\right)^2 + \cdots + \beta_2^{i-1}\left(\left|g_{l,k}^b\right| + \frac{\Delta_1}{\sqrt{nk}}\right)^2\right) \\ \leq &i\frac{\left(\left|g_{l,k}^b\right| + i\frac{\Delta_1}{\sqrt{nk}}\right)^2}{v_{l,k,0}}. \end{aligned} \tag{25}$$

As we assumed $|\partial_l f(x_{k,0})| \geq \frac{\eta_1 L n\sqrt{dn}}{\sqrt{k}\sqrt{1-\beta_2}}\left(\frac{32\sqrt{2}}{1-\beta_2^n}\right)$, we have

$$|g_{l,k}^b| \geq \frac{\eta_1 L\sqrt{dn}}{\sqrt{k}\sqrt{1-\beta_2}}\left(\frac{32\sqrt{2}}{1-\beta_2^n}\right) = \frac{\Delta_1\sqrt{n}}{\sqrt{k}}\frac{32\sqrt{2}}{1-\beta_2^n}.$$

Therefore, we can further simplify the inequality as

$$\begin{aligned}
&\frac{1}{v_{l,k,0}}\left(\left(g_{l,k,i,\tau_{k,i}}\right)^2 + \beta_2\left(g_{l,k,i-1,\tau_{k,i-1}}\right)^2 + \cdots + \beta_2^{i-1}\left(g_{l,k,1,\tau_{k,1}}\right)^2\right)\\
\leq& i\frac{\left|g_{l,k}^b\right|^2\left(1+\frac{1-\beta_2^n}{32\sqrt{2}}\right)^2}{v_{l,k,0}} \leq 2i\frac{\left|g_{l,k}^b\right|^2}{v_{l,k,0}}.
\end{aligned}$$

Recall that the lower bound of $v_{l,k,0}$ is given by

$$\frac{v_{l,k,0}}{\left(g_{l,k}^b\right)^2} \geq \frac{v_{l,k,0}}{\frac{1}{n}\sum_{i=1}^n |g_{l,k,0,i}|^2\rho_2} \geq \frac{\beta_2^n}{2\rho_2}.$$

Thus,

$$\frac{1}{v_{l,k,0}}\left(\left(g_{l,k,i,\tau_{k,i}}\right)^2 + \beta_2\left(g_{l,k,i-1,\tau_{k,i-1}}\right)^2 + \cdots + \beta_2^{i-1}\left(g_{l,k,1,\tau_{k,1}}\right)^2\right) \leq \frac{4i\rho_2}{\beta_2^n}$$

As a result,

$$\begin{aligned}
\frac{|v_{l,k,i}-v_{l,k,0}|}{v_{l,k,0}} =& \frac{(1-\beta_2)}{v_{l,k,0}}\Bigg|\left(g_{l,k,i,\tau_{k,i}}\right)^2 - v_{l,k,0} + \beta_2\left(\left(g_{l,k,i-1,\tau_{k,i-1}}\right)^2 - v_{l,k,0}\right) + \cdots\\
&+ \beta_2^{i-1}\left(\left(g_{l,k,1,\tau_{k,1}}\right)^2 - v_{l,k,0}\right)\Bigg|\\
=& \frac{(1-\beta_2)}{v_{l,k,0}}\left(\left(g_{l,k,i,\tau_{k,i}}\right)^2 + \beta_2\left(g_{l,k,i-1,\tau_{k,i-1}}\right)^2 + \cdots + \beta_2^{i-1}\left(g_{l,k,1,\tau_{k,1}}\right)^2\right)\\
&- \frac{(1-\beta_2)}{v_{l,k,0}}\left(v_{l,k,0} + \beta_2 v_{l,k,0} + \cdots + \beta_2^{i-1}v_{l,k,0}\right)\\
\leq& (1-\beta_2)\left(\frac{4i\rho_2}{\beta_2^n} - 1\right)
\end{aligned}$$

where the definition of ρ_2 is in (15) and we finally have

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

This completes our proof of the lemma. □

The next lemma is about the inner product between the gradient and all iterations in one epoch.

Lemma F.4. *Under assumptions in Theorem 4.3, if the largest component α satisfies: (i) $|\partial_\alpha f(x_{k,0})| \geq 32\sqrt{2}n^2\frac{\Delta_1}{\left(1-\beta_2^n\right)\beta_2^n\sqrt{nk}}$; (ii) $\sqrt{|\partial_\alpha f(x_{k,0})|^2 + \frac{D_0}{D_1 d}} \geq 4\sqrt{2}\frac{\Delta_1}{(1-\beta_2)\sqrt{D_1 nkd}}$, we have*

$$\begin{aligned}
&-\left\langle \nabla f_{k,0}, \sum_{i=0}^{n-1}\frac{g_{k,i,\tau_{k,i}}}{\sqrt{v_{k,i}}}\right\rangle \leq\\
&-\frac{1}{\sqrt{\frac{5}{2}D_1 d}}\min\left\{(1-T_2(\beta_2))\,|\partial_\alpha f(x_{k,0})|, |\partial_\alpha f(x_{k,0})|^2\frac{1}{\sqrt{\frac{D_0}{D_1 d}}}\right\} + T_2(\beta_2)\sqrt{\frac{8D_0}{5D_1^2 d^2}} + \frac{\Delta_1}{\sqrt{nk}}C_3
\end{aligned}$$

with T_2 defined in (37).

Proof. **Case I**: We first consider those gradient component large enough, i.e. $|\partial_l f(x_{k,0})|$ greater than $\frac{\Delta_1}{\sqrt{nk}}\left(\frac{32\sqrt{2}}{(1-\beta_2^n)\beta_2^n}\right)n^2$. By Lemma C.1 and Lipschitz continuity,

$$g_{l,k,0,\tau_{k,i}} - \frac{i\eta_1 L\sqrt{d}}{\sqrt{kn}\sqrt{1-\beta_2}} \leq g_{l,k,i,\tau_{k,i}} \leq g_{l,k,0,\tau_{k,i}} + \frac{i\eta_1 L\sqrt{d}}{\sqrt{kn}\sqrt{1-\beta_2}}. \tag{26}$$

Therefore,

$$\partial_l f(x_{k,0})\, g_{l,k,0,\tau_{k,i}} - \frac{i\eta_1 L\sqrt{d}\,|\partial_l f(x_{k,0})|}{\sqrt{kn}\sqrt{1-\beta_2}} \leq \partial_l f(x_{k,0})\, g_{l,k,i,\tau_{k,i}} \leq \partial_l f(x_{k,0})\, g_{l,k,0,\tau_{k,i}} + \frac{i\eta_1 L\sqrt{d}\,|\partial_l f(x_{k,0})|}{\sqrt{kn}\sqrt{1-\beta_2}}. \tag{27}$$

As the signs of $g_{l,k,0,\tau_{k,i}}$ and $\partial_l f(x_{k,0})$ can be same or different, we have to treat the 2 cases respectively.

When $\partial_l f(x_{k,0})$ and $g_{l,k,0,\tau_{k,i}}$ share the same sign, their product is positive. Then from Lemma F.3,

$$\begin{aligned}
&\partial_l f(x_{k,0})\,\frac{g_{l,k,i,\tau_{k,i}}}{\sqrt{v_{l,k,i}}} \\
&\geq \partial_l f(x_{k,0})\,\frac{g_{l,k,0,\tau_{k,i}}}{\sqrt{v_{l,k,0}}}\left(1-\frac{1-\beta_2}{2}\left(-1+\frac{4\rho_2 i}{\beta_2^n}\right)\right) - \frac{i\eta_1 L\sqrt{d}\,|\partial_l f(x_{k,0})|}{\sqrt{kn}\sqrt{1-\beta_2}\sqrt{v_{l,k,i}}} \\
&\geq \partial_l f(x_{k,0})\,\frac{g_{l,k,0,\tau_{k,i}}}{\sqrt{v_{l,k,0}}}\left(1-\frac{1-\beta_2}{2}\left(-1+\frac{4\rho_2 i}{\beta_2^n}\right)\right) - \frac{i\eta_1 L\sqrt{d}\,|\partial_l f(x_{k,0})|}{\sqrt{kn}\sqrt{1-\beta_2}\sqrt{v_{l,k,0}\beta_2^i}}.
\end{aligned} \tag{28}$$

On the other hand, if they have different signs, we simply have

$$\begin{aligned}
&\partial_l f(x_{k,0})\,\frac{g_{l,k,i,\tau_{k,i}}}{\sqrt{v_{l,k,i}}} \\
&\geq \partial_l f(x_{k,0})\,\frac{g_{l,k,0,\tau_{k,i}}}{\sqrt{v_{l,k,0}}}\frac{1}{\sqrt{\beta_2^i}} - \frac{|\partial_l f(x_{k,0})|}{\sqrt{v_{l,k,0}}}\frac{i\eta_1 L\sqrt{d}}{\sqrt{kn}\sqrt{1-\beta_2}\sqrt{\beta_2^i}}.
\end{aligned} \tag{29}$$

Combining these two inequalities yields

$$\begin{aligned}
&\partial_l f(x_{k,0})\sum_{i=0}^{n-1}\frac{g_{l,k,i,\tau_{k,i}}}{\sqrt{v_{l,k,i}}} \\
\geq&\frac{\partial_l f(x_{k,0})}{\sqrt{v_{l,k,0}}}\left(\sum_{i\in i+} g_{l,k,0,\tau_{k,i}}\left(1-\frac{1-\beta_2}{2}\left(-1+\frac{4n\rho_2}{\beta_2^n}\right)\right)+\sum_{i\in i-}\frac{g_{l,k,i,\tau_{k,i}}}{\sqrt{\beta_2^n}}\right) - \\
&\frac{|\partial_l f(x_{k,0})|}{\sqrt{v_{l,k,0}}}\frac{\eta_1 L\sqrt{d}}{\sqrt{kn}\sqrt{1-\beta_2}}\left(\sum_{i=0}^{n-1}\frac{i}{\sqrt{\beta_2^i}}\right)
\end{aligned} \tag{30}$$

where $i+$ means the set of the indices of the components with the same sign of $\partial_l f(x_{k,0})$ and $i-$ means the set of the indices of the components with opposite sign. Note that we have added 2 non-positive terms on the right hand side. For simplicity, define

$$C_2 \triangleq \frac{n(n-1)}{2\sqrt{\beta_2^n}}.$$

Since

$$\sum_{i\in i+} g_{l,k,0,\tau_{k,i}} + \sum_{i\in i-} g_{l,k,0,\tau_{k,i}} = \partial_l f(x_{k,0}),$$

we have

$$
\begin{aligned}
&\partial_l f\left(x_{k,0}\right) \sum_{i=0}^{n-1} \frac{g_{l,k,i,\tau_{k,i}}}{\sqrt{v_{l,k,i}}} \\
\geq & \frac{\partial_l f\left(x_{k,0}\right)}{\sqrt{v_{l,k,0}}}\left(\sum_{i \in i+} g_{l,k,0,\tau_{k,i}}\left(1-\frac{1-\beta_2}{2}\left(-1+\frac{4 n \rho_2}{\beta_2^n}\right)\right)+\sum_{i \in i-} g_{l,k,i,\tau_{k,i}}\left(1+\frac{1}{\sqrt{\beta_2^n}}-1\right)\right) \\
&-\frac{\left|\partial_l f\left(x_{k,0}\right)\right|}{\sqrt{v_{l,k,0}}} \frac{\eta_1 L \sqrt{d}}{\sqrt{k n} \sqrt{1-\beta_2}} C_2 \\
\geq & \frac{\partial_l f\left(x_{k,0}\right)^2}{\sqrt{v_{l,k,0}}}-\frac{\left|\partial_l f\left(x_{k,0}\right)\right|}{\sqrt{v_{l,k,0}}}\left(\sum_{i \in i+}\left|g_{l,k,0,i}\right|\left(1-\beta_2\right) \frac{\left(-1+4 n \rho_2 \beta_2^{-n}\right)}{2}+\sum_{i \in i-}\left|g_{l,k,0,i}\right|\left(\frac{1}{\sqrt{\beta_2^n}}-1\right)\right) \\
&-\frac{\left|\partial_l f\left(x_{k,0}\right)\right|}{\sqrt{v_{l,k,0}}} \frac{\eta_1 L \sqrt{d}}{\sqrt{k n} \sqrt{1-\beta_2}} C_2 \\
\geq & \frac{\partial_l f\left(x_{k,0}\right)^2}{\sqrt{v_{l,k,0}}}-\frac{\left|\partial_l f\left(x_{k,0}\right)\right|}{\sqrt{v_{l,k,0}}}\left(\sum_{i \in i+}\left|g_{l,k,0,i}\right|\left(1-\beta_2\right) \frac{\left(-1+4 n \rho_2 \beta_2^{-n}\right)}{2}+\sum_{i \in i-}\left|g_{l,k,0,i}\right|\left(\frac{1}{\sqrt{\beta_2^n}}-1\right)\right) \\
&-\frac{\eta_1 L \sqrt{d 2} n}{\sqrt{k} \sqrt{1-\beta_2} \beta_2^{n / 2}} C_2 .
\end{aligned}
\tag{31}
$$

The last inequality holds due to Lemma F.1 and the fact that $\left|\partial_l f\left(x_{k,0}\right)\right|$ is greater than $\frac{\Delta_1}{\sqrt{nk}}\left(\frac{32\sqrt{2}}{(1-\beta_2^n)\beta_2^n}\right) n^2$. It can further reduce to

$$
\begin{aligned}
&\partial_l f\left(x_{k,0}\right) \sum_{i=0}^{n-1} \frac{g_{l,k,i,\tau_{k,i}}}{\sqrt{v_{l,k,i}}} \\
\geq & \frac{\partial_l f\left(x_{k,0}\right)^2}{\sqrt{v_{l,k,0}}}-\frac{\left|\partial_l f\left(x_{k,0}\right)\right|}{\sqrt{v_{l,k,0}}} \sum_i\left|g_{l,k,0,i}\right| \delta_1-\frac{\eta_1 L \sqrt{d 2} n}{\sqrt{k} \sqrt{1-\beta_2} \beta_2^{n / 2}} C_2
\end{aligned}
\tag{32}
$$

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

Case II: For those gradient components smaller than $\frac{\Delta_1}{\sqrt{nk}}\left(\frac{32\sqrt{2}}{(1-\beta_2^n)\beta_2^n}\right) n^2$, the inequality is simply

$$
\partial_l f\left(x_{k,0}\right) \sum_{i=0}^{n-1} \frac{g_{l,k,i,\tau_{k,i}}}{\sqrt{v_{l,k,i}}} \geq-\frac{\Delta_1}{\sqrt{nk}}\left(\frac{32 \sqrt{2}}{\left(1-\beta_2^n\right) \beta_2^n}\right) n^2 \frac{n}{\sqrt{1-\beta_2}}
$$

because of Lemma C.1.

We denote the gradient components in case I by "l large" (large in the sense that $\left|\partial_l f\left(x_{k,0}\right)\right| \geq \frac{\Delta_1}{\sqrt{nk}}\left(\frac{32\sqrt{2}}{(1-\beta_2^n)\beta_2^n}\right) n^2$) and the rest components of the gradient by "l small". Summing up all of them, we have

$$
\begin{aligned}
&\sum_{l=1}^d \partial_l f\left(x_{k,0}\right) \sum_{i=0}^{n-1} \frac{g_{l,k,i,\tau_{k,i}}}{\sqrt{v_{l,k,i}}} \\
\geq & \sum_{l \text{ large}}\left(\frac{\partial_l f\left(x_{k,0}\right)^2}{\sqrt{v_{l,k,0}}}-\frac{\left|\partial_l f\left(x_{k,0}\right)\right|}{\sqrt{v_{l,k,0}}} \sum_i\left|g_{l,k,0,i}\right| \delta_1-\frac{\eta_1 L \sqrt{d 2} n}{\sqrt{k} \sqrt{1-\beta_2} \beta_2^{n / 2}} C_2\right) \\
&+\sum_{l \text{ small}}-\frac{\Delta_1}{\sqrt{nk}}\left(\frac{32 \sqrt{2}}{\left(1-\beta_2^n\right) \beta_2^n}\right) \frac{n^3}{\sqrt{1-\beta_2}} .
\end{aligned}
\tag{33}
$$

We can further simplify the inequality to

$$\begin{aligned}&\sum_{l=1}^{d}\partial_l f\left(x_{k,0}\right)\sum_{i=0}^{n-1}\frac{g_{l,k,i,\tau_{k,i}}}{\sqrt{v_{l,k,i}}}\\ \geq&\sum_{l\text{ large}}\frac{\partial_l f\left(x_{k,0}\right)^2}{\sqrt{v_{l,k,0}}}-\sum_{l\text{ large}}\frac{\left|\partial_l f\left(x_{k,0}\right)\right|}{\sqrt{v_{l,k,0}}}\sum_i\left|g_{l,k,0,i}\right|\delta_1-\frac{\Delta_1}{\sqrt{nk}}C_3\end{aligned}\tag{34}$$

where

$$\begin{aligned}C_3&=\frac{\sqrt{2n}nd}{\beta_2^{n/2}}C_2+\left(\frac{32\sqrt{2}}{(1-\beta_2^n)\beta_2^n}\right)\frac{n^3d}{\sqrt{1-\beta_2}}\\&=\frac{\sqrt{2n}n^2(n-1)d}{2\beta_2^n}+\frac{32\sqrt{2}}{1-\beta_2^n}\frac{n^3d}{\sqrt{(1-\beta_2)}\beta_2^n}\end{aligned}$$

We have assumed that $\left|\partial_\alpha f\left(x_{k,0}\right)\right|>\frac{\Delta_1}{\sqrt{nk}}\frac{32\sqrt{2}}{(1-\beta_2^n)\beta 2^n}n^2$ in assumption (i), thus $\alpha\in l$ large. Furthermore, as a very loose estimate, we keep only the α component in the first term:

$$\begin{aligned}&\sum_{l=1}^{d}\partial_l f\left(x_{k,0}\right)\sum_{i=0}^{n-1}\frac{g_{l,k,i,\tau_{k,i}}}{\sqrt{v_{l,k,i}}}\\ \geq&\frac{\partial_\alpha f\left(x_{k,0}\right)^2}{\sqrt{v_{\alpha,k,0}}}-\sum_{l\text{ large}}\frac{\left|\partial_l f\left(x_{k,0}\right)\right|}{\sqrt{v_{l,k,0}}}\sum_i\left|g_{l,k,0,\tau_{k,i}}\right|\delta_1-\frac{\Delta_1}{\sqrt{nk}}C_3\end{aligned}\tag{35}$$

We know from Lemma F.1 that if $\left|\partial_l f\left(x_{k,0}\right)\right|\geq\frac{\eta_1\sqrt{dn}n^2}{\sqrt{k}\sqrt{(1-\beta_2)\beta_2^n}}\left(\frac{32\sqrt{2}}{1-\beta_2^n}\right)$:

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

where the definition of ρ_3 is in (16). Then by the assumption $\sum_{j=0}^{n-1}\left\|\nabla f_j\right\|_2^2\leq D_1\left\|\nabla f\right\|_2^2+D_0$, we have

$$\begin{aligned}&\sum_{l\text{ large}}\frac{\left|\partial_l f\left(x_{k,0}\right)\right|}{\sqrt{v_{l,k,0}}}\sum_{i=0}^{n-1}\left|g_{l,k,0,\tau_{k,i}}\right|\\ \leq&\sum_{l\text{ large}}\sqrt{\frac{2\rho_3^2}{\beta_2^n}}\sum_{i=0}^{n-1}\left|g_{l,k,0,\tau_{k,i}}\right|\\ \leq&\sqrt{\frac{2\rho_3^2}{\beta_2^n}}\sum_{l=1}^{d}\sum_{i=0}^{n-1}\left|g_{l,k,0,\tau_{k,i}}\right|\\ \leq&\sqrt{\frac{2\rho_3^2}{\beta_2^n}}\sqrt{d}\rho_1\sqrt{D_1\left\|\nabla f\right\|_2^2+D_0}\\ \leq&\sqrt{\frac{2\rho_3^2}{\beta_2^n}}\sqrt{D_1}\rho_1 d\sqrt{\left|\partial_\alpha f\left(x_{k,0}\right)\right|^2+\frac{D_0}{D_1 d}}\\ \leq&\sqrt{\frac{2\rho_3^2}{\beta_2^n}}\sqrt{D_1}\rho_1 d\left(\left|\partial_\alpha f\left(x_{k,0}\right)\right|+\sqrt{\frac{D_0}{D_1 d}}\right)\end{aligned}\tag{36}$$

where the third inequality comes from the fact that under the constraints (i) $\sum_{i=0}^{n-1}\left|g_{l,k,0,\tau_{k,i}}\right|\leq\rho_1\sqrt{\sum_{i=0}^{n-1}\left|g_{l,k,0,\tau_{k,i}}\right|^2}$ for each i and (ii)$\sum_{i=0}^{n-1}\sum_{l=1}^{d}\left|g_{l,k,0,\tau_{k,i}}\right|^2\leq D_1\left\|\nabla f\right\|_2^2+D_0$, $\sum_{l=1}^{d}\sum_{i=0}^{n-1}\left|g_{l,k,0,\tau_{k,i}}\right|$ is upper bounded by $\sqrt{d}\rho_1\sqrt{D_1\left\|\nabla f\right\|_2^2+D_0}$, the fourth is because $\left\|\nabla f\right\|_2^2\leq d\left|\partial_\alpha f\left(x_{k,0}\right)\right|^2$, and the last is because $\sqrt{x+y}\leq\sqrt{x}+\sqrt{y}$, for any $x,y\geq 0$.

Therefore, inequality (35) reduces to

$$\begin{aligned}
&\sum_{l=1}^{d}\partial_l f\left(x_{k,0}\right)\sum_{i=0}^{n-1}\frac{g_{l,k,i,\tau_{k,i}}}{\sqrt{v_{l,k,i}}}\\
\geq&\frac{\partial_\alpha f\left(x_{k,0}\right)^2}{\sqrt{v_{\alpha,k,0}}}-\sqrt{\frac{2}{\beta_2^n}D_1}\rho_1\rho_3 d\left(\left|\partial_\alpha f\left(x_{k,0}\right)\right|+\sqrt{\frac{D_0}{D_1 d}}\right)\delta_1-\frac{\Delta_1}{\sqrt{nk}}C_3.
\end{aligned}$$

Taking in the result from Lemma F.2 that $v_{\alpha,k,0}\leq\frac{5}{2}D_1 d\left(\left|\partial_\alpha f\left(x_{k,0}\right)\right|^2+\frac{D_0}{D_1 d}\right)$, we have

$$\begin{aligned}
&\sum_{l=1}^{d}\partial_l f\left(x_{k,0}\right)\sum_{i=0}^{n-1}\frac{g_{l,k,i,\tau_{k,i}}}{\sqrt{v_{l,k,i}}}\\
\geq&\frac{\partial_\alpha f\left(x_{k,0}\right)^2}{\sqrt{\frac{5}{2}D_1 d\left(\left|\partial_\alpha f\left(x_{k,0}\right)\right|^2+\frac{D_0}{D_1 d}\right)}}-\sqrt{\frac{2}{\beta_2^n}D_1}\rho_1\rho_3 d\left(\left|\partial_\alpha f\left(x_{k,0}\right)\right|+\sqrt{\frac{D_0}{D_1 d}}\right)\delta_1-\frac{\Delta_1}{\sqrt{nk}}C_3\\
\geq&\frac{1}{\sqrt{5D_1 d}}\min\{\left|\partial_\alpha f\left(x_{k,0}\right)\right|,\left|\partial_\alpha f\left(x_{k,0}\right)\right|^2\frac{1}{\sqrt{\frac{D_0}{D_1 d}}}\}\\
&-\sqrt{\frac{2}{\beta_2^n}D_1}\rho_1\rho_3 d\left(\left|\partial_\alpha f\left(x_{k,0}\right)\right|+\sqrt{\frac{D_0}{D_1 d}}\right)\delta_1-\frac{\Delta_1}{\sqrt{nk}}C_3\\
\geq&\frac{1}{\sqrt{5D_1 d}}\left(\min\{\left|\partial_\alpha f\left(x_{k,0}\right)\right|,\left|\partial_\alpha f\left(x_{k,0}\right)\right|^2\frac{1}{\sqrt{\frac{D_0}{D_1 d}}}\}-T_2(\beta_2)\left(\left|\partial_\alpha f\left(x_{k,0}\right)\right|+\sqrt{\frac{D_0}{D_1 d}}\right)\right)-\frac{\Delta_1}{\sqrt{nk}}C_3\\
\geq&\frac{1}{\sqrt{5D_1 d}}\left(\min\{\left(1-T_2(\beta_2)\right)\left|\partial_\alpha f\left(x_{k,0}\right)\right|,\left|\partial_\alpha f\left(x_{k,0}\right)\right|^2\frac{1}{\sqrt{\frac{D_0}{D_1 d}}}\}-2T_2(\beta_2)\sqrt{\frac{D_0}{D_1 d}}\right)-\frac{\Delta_1}{\sqrt{nk}}C_3\\
\geq&\frac{1}{\sqrt{5D_1 d}}\min\{\left(1-T_2(\beta_2)\right)\left|\partial_\alpha f\left(x_{k,0}\right)\right|,\left|\partial_\alpha f\left(x_{k,0}\right)\right|^2\frac{1}{\sqrt{\frac{D_0}{D_1 d}}}\}-T_2(\beta_2)\sqrt{\frac{8D_0}{5D_1^2 d^2}}-\frac{\Delta_1}{\sqrt{nk}}C_3
\end{aligned}$$

where T_2 is defined as

$$T_2\left(\beta_2\right)=\sqrt{\frac{10d}{\beta_2^n}}d\rho_1\rho_3 D_1\delta_1=\sqrt{\frac{10d}{\beta_2^n}}d\rho_1\rho_3 D_1\left(\left(1-\beta_2\right)\frac{\left(\frac{4n\rho_2}{\beta_2^n}-1\right)}{2}+\left(\frac{1}{\sqrt{\beta_2^n}}-1\right)\right). \tag{37}$$

Note that in the fourth inequality, we used the following inequality:

$$\begin{aligned}
&\min\{x,\frac{x^2}{\sqrt{\frac{D_0}{D_1 d}}}\}-T_2 x\\
&=\min\{(1-T_2)x,\frac{x^2}{\sqrt{\frac{D_0}{D_1 d}}}-T_2 x\}\\
&\geq\min\{(1-T_2)x,\frac{x^2}{\sqrt{\frac{D_0}{D_1 d}}}-T_2\sqrt{\frac{D_0}{D_1 d}}\}\\
&\geq\min\{(1-T_2)x,\frac{x^2}{\sqrt{\frac{D_0}{D_1 d}}}\}-T_2\sqrt{\frac{D_0}{D_1 d}}
\end{aligned}$$

with $x=|\partial_\alpha f(x_{k,0})|$. □

Remark:when β_2 is very close to 1, for the first order Taylor expansion:

$$T_2(\beta_2) \sim \mathcal{O}\left((1-\beta_2)\, n\rho_1\rho_2\rho_3\right)$$

Lemma F.5. *Under assumptions in Theorem 4.3, if we choose β_2 to be a constant satisfying the constraint $T_2(\beta_2) \leq 1 - \frac{1}{\sqrt{2}}$, we have the following for all $\nabla f_{k,0}$:*

$$-\left\langle \nabla f_{k,0}, \sum_{i=0}^{n-1} \frac{g_{k,i,\tau_{k,i}}}{\sqrt{v_{k,i}}} \right\rangle \leq -\frac{1}{\sqrt{10D_1 d}} \min\left\{ \frac{\|\nabla f_{k,0}\|_1}{d}, \frac{\|\nabla f_{k,0}\|_2^2}{\sqrt{\frac{D_0 d}{D_1}}} \right\} + \sqrt{D_0} C_5 + \frac{\Delta_1}{\sqrt{nk}} C_4$$

where $\|\nabla f_{k,0}\|_1$ is the 1-norm of vector $\nabla f_{k,0}$, C_4 and C_5 are defined in (38) *and* (39)*.*

Note that in the zero initialization and bias corrected version, we also need the condition $k > \frac{8\sqrt{2}}{1-\beta_2^n} + 1$.

Proof. We will discuss the cases where conditions in Lemma F.4 hold or become violated.

Case 1 If we have $|\partial_\alpha f(x_{k,0})| \geq 32\sqrt{2}n^2 \frac{\Delta_1}{(1-\beta_2^n)\beta_2^n\sqrt{nk}}$ and $|\partial_\alpha f(x_{k,0})| \geq 4\sqrt{2}\frac{\Delta_1}{(1-\beta_2)\sqrt{D_1 nkd}}$, then we can apply Lemma F.4:

$$\begin{aligned}
&\left\langle \nabla f_{k,0}, \sum_{i=0}^{n-1} \frac{g_{k,i,\tau_{k,i}}}{\sqrt{v_{k,i}}} \right\rangle \geq \\
&\frac{1}{\sqrt{5D_1 d}} \min\left\{ (1 - T_2(\beta_2))\, |\partial_\alpha f(x_{k,0})|, |\partial_\alpha f(x_{k,0})|^2 \frac{1}{\sqrt{\frac{D_0}{D_1 d}}} \right\} - T_2(\beta_2)\sqrt{\frac{8D_0}{5D_1^2 d^2}} - \frac{\Delta_1}{\sqrt{nk}} C_3 \\
&\geq \frac{1}{\sqrt{10D_1 d}} \min\left\{ |\partial_\alpha f(x_{k,0})|, \frac{|\partial_\alpha f(x_{k,0})|^2}{\sqrt{\frac{D_0}{D_1 d}}} \right\} - T_2(\beta_2)\sqrt{\frac{8D_0}{5D_1^2 d^2}} - \frac{\Delta_1}{\sqrt{nk}} C_3 \\
&\geq \frac{1}{\sqrt{10D_1 d}} \min\left\{ |\partial_\alpha f(x_{k,0})|, \frac{|\partial_\alpha f(x_{k,0})|^2}{\sqrt{\frac{D_0}{D_1 d}}} \right\} - T_2(\beta_2)\sqrt{\frac{8D_0}{5D_1^2 d^2}} - \frac{\Delta_1}{\sqrt{nk}} C_4
\end{aligned}$$

C_4 is a constant defined as

$$C_4 = C_3 + \frac{1}{(1-\beta_2)} \max\{32\sqrt{2}n, \frac{4\sqrt{2}}{\sqrt{D_1 d}}\} \left(\frac{1}{\sqrt{10D_1 d}} \min\{1, \frac{\Delta_1\sqrt{D_1 d}}{(1-\beta_2)\sqrt{nD_0}} \max\{32\sqrt{2}n^2\beta_2^{-n}, \frac{4\sqrt{2}}{\sqrt{D_1 d}}\}\} + \frac{dn}{\sqrt{1-\beta_2}} \right) \tag{38}$$

Case 2 Else-wise, $|\partial_\alpha f(x_{k,0})| \leq \frac{\Delta_1}{(1-\beta_2)\sqrt{nk}} \max\{32\sqrt{2}n^2\beta_2^{-n}, \frac{4\sqrt{2}}{\sqrt{D_1 d}}\}$. As a result,

$$\begin{aligned}
&\left\langle \nabla f_{k,0}, \sum_{i=0}^{n-1} \frac{g_{k,i,\tau_{k,i}}}{\sqrt{v_{k,i}}} \right\rangle \\
\geq& -d\frac{\Delta_1}{(1-\beta_2)\sqrt{nk}} \max\left\{ 32\sqrt{2}n^2\beta_2^{-n}, \frac{4\sqrt{2}}{\sqrt{D_1 d}} \right\} \frac{n}{\sqrt{1-\beta_2}} \\
\geq& \frac{1}{\sqrt{10D_1 d}} \min\left\{ |\partial_\alpha f(x_{k,0})|, \frac{|\partial_\alpha f(x_{k,0})|^2}{\sqrt{\frac{D_0}{D_1 d}}} \right\} - \frac{\Delta_1}{\sqrt{nk}} \left(C_3 + \frac{1}{(1-\beta_2)} \max\left\{ 32\sqrt{2}n^2\beta_2^{-n}, \frac{4\sqrt{2}}{\sqrt{D_1 d}} \right\} \right. \\
&\left. \cdot \left(\frac{1}{\sqrt{10D_1 d}} \min\left\{ 1, \frac{\Delta_1\sqrt{D_1 d}}{(1-\beta_2)\sqrt{nD_0}} \max\left\{ 32\sqrt{2}n^2\beta_2^{-n}, \frac{4\sqrt{2}}{\sqrt{D_1 d}} \right\} \right\} \right) + \frac{dn}{\sqrt{1-\beta_2}} \right) - T_2(\beta_2)\sqrt{\frac{8D_0}{5D_1^2 d^2}} \\
=& \frac{1}{\sqrt{10D_1 d}} \min\left\{ |\partial_\alpha f(x_{k,0})|, \frac{|\partial_\alpha f(x_{k,0})|^2}{\sqrt{\frac{D_0}{D_1 d}}} \right\} - \frac{\Delta_1}{\sqrt{nk}} C_4 - T_2(\beta_2)\sqrt{\frac{8D_0}{5D_1^2 d^2}}.
\end{aligned}$$

This finishes the proof, with

$$C_5 = \frac{2\sqrt{10}T_2(\beta_2)}{5D_1 d} \tag{39}$$

□

Proof of Theorem 4.3.

Since f is L-Lipschitz, by descent lemma and Lemma C.1,

$$\begin{aligned} f\left(x_{k+1,0}\right) - f\left(x_{k,0}\right) &\leq \left\langle \nabla f\left(x_{k,0}\right), x_{k+1,0} - x_{k,0} \right\rangle + \frac{L}{2}\left\| x_{k+1,0} - x_{k,0} \right\|^2 \\ &\leq -\frac{\eta_0}{\sqrt{kn}} \left\langle \nabla f\left(x_{k,0}\right), \sum_{i=0}^{n-1} \frac{g_{k,i,\tau_{k,i}}}{\sqrt{v_{k,i}}} \right\rangle + \frac{L}{2}\frac{\eta_1^2 nd}{\left(1-\beta_2\right)k} \end{aligned} \tag{40}$$

Summing both sides of the inequality ranging k from t_{init} to T,

$$f\left(x_{T+1,0}\right) - f\left(x_{t_{init},0}\right) \leq -\sum_{k=t_{init}}^{T} \frac{\eta_1}{\sqrt{nk}} \left\langle \nabla f_t, \sum_{i=0}^{n-1} \frac{g_{k,i,\tau_{k,i}}}{\sqrt{v_{k,i}}} \right\rangle + \sum_{k=t_{init}}^{T} \frac{L}{2}\frac{\eta_1^2 nd}{\left(1-\beta_2\right)k}.$$

Since $f\left(x_{T+10}\right) \geq f^\star$, we have

$$\sum_{k=t_{init}}^{T} \frac{\eta_1}{\sqrt{nk}} \left\langle \nabla f_t, \sum_{i=0}^{n-1} \frac{g_{k,i,\tau_{k,i}}}{\sqrt{v_{k,i}}} \right\rangle \leq \sum_{k=t_{init}}^{T} \frac{L}{2}\frac{\eta_1^2 nd}{\left(1-\beta_2\right)k} + f\left(x_{t_{init}0}\right) - f^\star$$

.

Let $t_{init} = 4$ for specially initialized version, and $1 + \lceil \frac{8\sqrt{2}}{1-\beta_2^n} \rceil$for zero initialized version, and apply the result from Lemma F.5, for all $k \geq t_{init}$,

$$\left\langle \nabla f_{k,0}, \sum_{i=0}^{n-1} \frac{g_{k,i,\tau_{k,i}}}{\sqrt{v_{k,i}}} \right\rangle \geq \frac{1}{\sqrt{10D_1 d}} \min\{\frac{\left\|\nabla f_{k,0}\right\|_1}{d}, \frac{\left\|\nabla f_{k,0}\right\|_2^2}{\sqrt{\frac{D_0 d}{D_1}}}\} - \sqrt{D_0}C_5 - \frac{\Delta_1}{\sqrt{nk}}C_4.$$

We can further simplify it as

$$\begin{aligned} &\sum_{k=t_{init}}^{T} \frac{\eta_1}{\sqrt{nk}} \left(\frac{1}{\sqrt{10D_1 d}} \min\left\{ \frac{\left\|\nabla f_{k,0}\right\|_1}{d}, \frac{\left\|\nabla f_{k,0}\right\|_2^2}{\sqrt{\frac{D_0 d}{D_1}}} \right\} - \sqrt{D_0}C_5 \right) \\ \leq &\sum_{k=t_{init}}^{T} \frac{C_6}{k} + f\left(x_{t_{init}0}\right) - f^\star \end{aligned} \tag{41}$$

where

$$C_6 = L\eta_1^2 \left(\frac{nd}{2\left(1-\beta_2\right)} + \frac{C_4\sqrt{d}}{n\sqrt{1-\beta_2}} \right) \tag{42}$$

On the right hand side, we have a summation proportional to

$$\sum_{k=t_{init}}^{T} \frac{1}{k} \leq \log \frac{T+1}{t_{init}}.$$

On the left hand side, we have a summation proportional to:

$$\sum_{k=t_{init}}^{T} \frac{1}{\sqrt{k}} \geq 2\left(\sqrt{T} - \sqrt{t_{init}-1}\right).$$

For Algorithm 1, we can set $t_{init} = 4$, while for Algorithm 2, we should set $t_{init} = \max\{\log_{\beta_1} \frac{1}{4}, 1 + \frac{8\sqrt{2}}{1-\beta_2^n}\}$. Hence, we have

$$\min_{k\in[t_{init},T]} \min\{\|\nabla f_{k,0}\|_1, \|\nabla f_{k,0}\|_2^2 \sqrt{\frac{D_1 d}{D_0}}\} \leq \frac{1}{\sqrt{T} - \sqrt{t_{init}-1}} \left(Q_{1,3} + Q_{2,3}\log(T+1)\right) + \sqrt{D_0} Q_{3,3}$$

where

$$Q_{1,3} = \frac{f(x_{t_{init},0}) - f^* - C_6 \log t_{init}}{2\eta_1} \sqrt{10 n D_1 d} d, \tag{43}$$

$$Q_{2,3} = \frac{C_6 \sqrt{10 n D_1 d} d}{2\eta_1}, \tag{44}$$

$$Q_{3,3} = C_5 \sqrt{10 D_1 d n} d, \tag{45}$$

C_6 is defined in (42), and C_5 is defined in (39).

Note that C_5 proportional to $\sqrt{T_2}$ and so is Q_3. Seemingly, C_6 depends on β_2^{-n}. Thus, it increases exponentially with the number of batch samples n. However, our choice of β_2 can prevent this: to keep T_2 which also contains β_2^{-n} terms small enough, we implicitly add an upper bound on β_2^{-n}.

G Proof of Theorem 4.4

Similar to the full batch version Adam, if we set $\Delta_t = \frac{\eta_1 L \sqrt{d}(1-\beta_1 * bc)}{\sqrt{1-\beta_2}\left(1-\frac{\beta_1}{\sqrt{\beta_2}}\right)\sqrt{t}}$, results of Lemma F.1, F.2, and F.3 still hold without further modifications. We will begin by finding an upper bound of the difference between $m_{l,k,i}$ and $g_{l,k,i,\tau_{k,i}}$, followed by a replacement for Lemma F.4.

Lemma G.1. *For $k > 1$, we have*

$$\left|m_{l,k,i} - g_{l,k,0,\tau_{k,i}}\right| \leq \beta_1 \sum_{q=0}^{n-1} \left|g_{l,k,0,\tau_{k,q}}\right| \left(\frac{1-\beta_1}{1-\beta_1^n} + 1\right) + \frac{\Delta_1}{\sqrt{kn}} \left(n + \beta_1 n^2 \frac{1-\beta_1}{(1-\beta_1^n)^2}\left(1 + 2\sqrt{2}\right)\right).$$

Proof. First, we upper bound $m_{l,k-1,n-1}$ by

$$\begin{aligned}
&|m_{l,k-1,n-1}| = \left|(1-\beta_1)\left(g_{l,k-1,n-1,\tau_{k-1,n-1}} + \beta_1 g_{l,k-1,n-2,\tau_{k-1,n-2}} \cdots\right) + \beta_1^{n(k-1)} \sum_{p=0}^{n-1} g_{l,1,-1,i}\right| \\
\leq &(1-\beta_1)\Big(\left(\left|g_{l,k-1,n-1,\tau_{k-1,n-1}}\right| + \left|g_{l,k-1,n-2,\tau_{k-1,n-2}}\right| + \cdots + \left|g_{l,k-1,0,\tau_{k-1,0}}\right|\right) + \\
&\beta_1^n \left(\left|g_{l,k-2,n-1,\tau_{k-2,n-1}}\right| + \left|g_{l,k-2,n-2,\tau_{k-2,n-2}}\right| + \cdots + \left|g_{l,k-2,0,\tau_{k-2,0}}\right|\right) + \cdots\Big) + \beta_1^{n(k-1)} \sum_{p=0}^{n-1} |g_{l,1,-1,i}| \\
= &(1-\beta_1) \sum_{p=1}^{k-1} \sum_{q=0}^{n-1} \left|g_{l,k-p,m-q,\tau_{k-p,m-q}}\right| \beta_1^{(p-1)n} + \beta_1^{n(k-1)} \sum_{p=0}^{n-1} |g_{l,1,-1,i}| \\
\leq &(1-\beta_1) \sum_{p=1}^{k-1} \sum_{q=0}^{n-1} \left(\left|g_{l,k,0,\tau_{k-p,q}}\right| + \sum_{t=1}^{p} n\Delta_{n(k-t)}\right) \beta_1^{(p-1)n} + \sum_{q=0}^{n-1} \left(\left|g_{l,k,0,\tau_{k-p,q}}\right| + \sum_{t=1}^{k-1} n\Delta_{n(k-t)}\right) \beta_1^{(k-1)n} \\
\leq &(1-\beta_1) \sum_{p=1}^{k-1} \sum_{q=0}^{n-1} \left(\left|g_{l,k,0,\tau_{k-p,q}}\right| + \frac{2\Delta_1 np}{\sqrt{n(k-1)}}\right) \beta_1^{(p-1)n} + \left(\left|g_{l,k,0,\tau_{k-p,q}}\right| + \frac{2\Delta_1 n(k-1)}{\sqrt{n(k-1)}}\right) \beta_1^{(k-1)n} \\
\leq &(1-\beta_1) \sum_{p=1}^{+\infty} \sum_{q=0}^{n-1} \left(\left|g_{l,k,0,\tau_{k-p,q}}\right| + \frac{2\Delta_1 np}{\sqrt{n(k-1)}}\right) \beta_1^{(p-1)n} \\
= &\sum_{q=0}^{n-1} \left|g_{l,k,0,\tau_{k,q}}\right| \frac{1-\beta_1}{1-\beta_1^n} + \frac{2\Delta_1 n^2}{\sqrt{n(k-1)}} \frac{1-\beta_1}{(1-\beta_1^n)^2}
\end{aligned}$$

where the first inequality is because $\beta_1 < 1$, the second comes from Lipschitz inequality, the third applies the result of Lemma C.3, the fourth combines two terms by the relation $1 = (1-\beta_1)(1+\beta_1+\beta_1^2+...)$, and the last equality follows the same calculation in Lemma D.1. This holds for the bias corrected version.

Therefore,

$$\begin{aligned}
&\left|m_{l,k,i} - g_{l,k,i,\tau_{k,i}}\right| \\
&= \left|\beta_1^{i+1} m_{l,k-1,n-1} + (1-\beta_1)\beta_1^i g_{l,k,0,\tau_{k,0}} + (1-\beta_1)\beta_1^{i-1} g_{l,k,1,\tau_{k,1}} + \cdots \right. \\
&\left. + (1-\beta_1)\beta_1 g_{l,k,i-1,\tau_{k,i-1}} - \beta_1 g_{l,k,i,\tau_{k,i}}\right| \\
&\leq \beta_1\left(\beta_1^i\left|m_{l,k-1,n-1}\right| + (1-\beta_1)\beta_1^{i-1}\left|g_{l,k,0,\tau_{k,0}}\right| + \cdots + (1-\beta_1)\left|g_{l,k,i-1,\tau_{k,i-1}}\right| + \left|g_{l,k,i,\tau_{k,i}}\right|\right) \\
&\leq \beta_1\left(\left|m_{l,k-1,n-1}\right| + \left|g_{l,k,0,\tau_{k,0}}\right| + \cdots + \left|g_{l,k,i-1,\tau_{k,i-1}}\right| + \left|g_{l,k,i,\tau_{k,i}}\right|\right)
\end{aligned}$$

By taking in the upper bound of $|m_{l,k-1,n-1}|$ and applying the Lipschitz gradient continuous condition, we have

$$\begin{aligned}
&\left|m_{l,k,i} - g_{l,k,i,\tau_{k,i}}\right| \\
&\leq \beta_1\left(\sum_{q=0}^{n-1}\left|g_{l,k,0,\tau_{k,q}}\right|\frac{1-\beta_1}{1-\beta_1^n} + \frac{2\Delta_1 n^2}{\sqrt{n(k-1)}}\frac{1-\beta_1}{\left(1-\beta_1^n\right)^2} + \sum_{q=0}^{i}\left(\left|g_{l,k,0,\tau_{k,q}}\right| + \frac{q\Delta_1}{\sqrt{kn}}\right)\right) \\
&\leq \beta_1\sum_{q=0}^{n-1}\left|g_{l,k,0,\tau_{k,q}}\right|\left(\frac{1-\beta_1}{1-\beta_1^n}+1\right) + \beta_1\frac{n^2\Delta_1}{\sqrt{kn}}\frac{1-\beta_1}{\left(1-\beta_1^n\right)^2}\left(1+2\sqrt{2}\right)
\end{aligned}$$

and

$$\begin{aligned}
&\left|m_{l,k,i} - g_{l,k,0,\tau_{k,i}}\right| \leq \left|m_{l,k,i} - g_{l,k,i,\tau_{k,i}}\right| + \left|g_{l,k,0,\tau_{k,i}} - g_{l,k,i,\tau_{k,i}}\right| \\
&\leq \left|m_{l,k,i} - g_{l,k,i,\tau_{k,i}}\right| + \frac{i\Delta_1}{\sqrt{kn}} \\
&\leq \beta_1\sum_{q=0}^{n-1}\left|g_{l,k,0,\tau_{k,q}}\right|\left(\frac{1-\beta_1}{1-\beta_1^n}+1\right) + \frac{\Delta_1}{\sqrt{kn}}\left(n + \beta_1 n^2\frac{1-\beta_1}{\left(1-\beta_1^n\right)^2}\left(1+2\sqrt{2}\right)\right)
\end{aligned}$$

where we have applied Lipschitz continuity in the first inequality. This completes the proof. □

Lemma G.2. *Under assumptions in Theorem 4.4, assume that the largest component α satisfies(i) $|\partial_\alpha f(x_{k,0})| \geq 32\sqrt{2}n^2\frac{\Delta_1}{\left(1-\beta_2^n\right)\beta_2^n\sqrt{nk}}$; (ii) $\sqrt{|\partial_\alpha f(x_{k,0})|^2 + \frac{D_0}{D_1 d}} \geq 4\sqrt{2}\frac{\Delta_1}{(1-\beta_2)\sqrt{D_1 nkd}}$. We have:*

$$\left\langle \nabla f_{k,0}, \sum_{i=0}^{n-1}\frac{m_{ki}}{\sqrt{v_{k,i}}}\right\rangle \geq \frac{|\partial_\alpha f(x_{k,0})|}{\sqrt{\frac{5}{2}D_1 d}}\left(1 - T_1(\beta_1) - T_2(\beta_2) - (T_1+T_2)\frac{D_0}{D_1 d|\partial_\alpha f|^2}\right) - \frac{\Delta_1}{\sqrt{nk}}C_8$$

with T_2 defined in (37)*.*

Proof. Similar to Lemma F.4, we first consider those gradient components large enough, i.e. $|\partial_l f(x_{k,0})|$ greater than $\frac{\Delta_1}{\sqrt{nk}}\left(\frac{32\sqrt{2}}{(1-\beta_2^n)\beta_2^n}\right)n^2$. By Lemma G.1,

$$\left|m_{l,k,i} - g_{l,k,0,\tau_{k,i}}\right| \leq \beta_1\left(\frac{1-\beta_1}{1-\beta_1^n}+1\right)\sum_{q=0}^{n-1}\left|g_{l,k,0,\tau_{k,q}}\right| + \frac{\Delta_1}{\sqrt{kn}}\left(n + \beta_1 n^2\frac{1-\beta_1}{\left(1-\beta_1^n\right)^2}\left(1+2\sqrt{2}\right)\right). \tag{46}$$

Therefore,

$$\begin{aligned}
&\partial_l f(x_{k,0})\, m_{l,k,i} \geq \partial_l f(x_{k,0})\, g_{l,k,0,\tau_{k,i}} \\
&- \left|\partial_l f(x_{k,0})\right|\left(\beta_1\left(\frac{1-\beta_1}{1-\beta_1^n}+1\right)\sum_{q=0}^{n-1}\left|g_{l,k,0,\tau_{k,q}}\right| + \frac{\Delta_1}{\sqrt{kn}}\left(n + \beta_1 n^2\frac{1-\beta_1}{\left(1-\beta_1^n\right)^2}\left(1+2\sqrt{2}\right)\right)\right).
\end{aligned} \tag{47}$$

As the sign of $g_{l,k,0,\tau_{k,i}}$ can be the same or different to $\partial_l f\left(x_{k,0}\right)$, we again have to treat 2 cases respectively.

When $\partial_l f\left(x_{k,0}\right)$ and $g_{l,k,0,\tau_{k,i}}$ share the same sign, their product is positive. Then from Lemma F.3,

$$\begin{aligned}
&\partial_l f\left(x_{k,0}\right) \frac{m_{l,k,i}}{\sqrt{v_{l,k,i}}} \\
&\geq \partial_l f\left(x_{k,0}\right) \frac{g_{l,k,0,\tau_{k,i}}}{\sqrt{v_{l,k,0}}}\left(1-\frac{1-\beta_2}{2}\left(-1+\frac{4n\rho_2}{\beta_2^n}\right)\right) \\
&-\frac{\left|\partial_l f\left(x_{k,0}\right)\right|}{\sqrt{v_{l,k,i}}}\left(\beta_1\left(\frac{1-\beta_1}{1-\beta_1^n}+1\right)\sum_{q=0}^{n-1}\left|g_{l,k,0,\tau_{k,q}}\right|+\frac{\Delta_1}{\sqrt{kn}}\left(n+\beta_1 n^2\frac{1-\beta_1}{\left(1-\beta_1^n\right)^2}\left(1+2\sqrt{2}\right)\right)\right) \\
&\geq \partial_l f\left(x_{k,0}\right) \frac{g_{l,k,0,\tau_{k,i}}}{\sqrt{v_{l,k,0}}}\left(1-\frac{1-\beta_2}{2}\left(-1+\frac{4n\rho_2}{\beta_2^n}\right)\right) \\
&-\frac{\left|\partial_l f\left(x_{k,0}\right)\right|}{\sqrt{\beta_2^i v_{l,k,0}}}\left(\beta_1\left(\frac{1-\beta_1}{1-\beta_1^n}+1\right)\sum_{q=0}^{n-1}\left|g_{l,k,0,\tau_{k,q}}\right|+\frac{\Delta_1}{\sqrt{kn}}\left(n+\beta_1 n^2\frac{1-\beta_1}{\left(1-\beta_1^n\right)^2}\left(1+2\sqrt{2}\right)\right)\right).
\end{aligned} \tag{48}$$

On the other hand, if they have different signs, we simply have

$$\begin{aligned}
&\partial_l f\left(x_{k,0}\right) \frac{m_{l,k,i}}{\sqrt{v_{l,k,i}}} \\
&\geq \partial_l f\left(x_{k,0}\right) \frac{g_{l,k,0,\tau_{k,i}}}{\sqrt{v_{l,k,0}}}\frac{1}{\sqrt{\beta_2^i}}- \\
&-\frac{\left|\partial_l f\left(x_{k,0}\right)\right|}{\sqrt{\beta_2^i v_{l,k,0}}}\left(\beta_1\left(\frac{1-\beta_1}{1-\beta_1^n}+1\right)\sum_{q=0}^{n-1}\left|g_{l,k,0,\tau_{k,q}}\right|+\frac{\Delta_1}{\sqrt{kn}}\left(n+\beta_1 n^2\frac{1-\beta_1}{\left(1-\beta_1^n\right)^2}\left(1+2\sqrt{2}\right)\right)\right).
\end{aligned} \tag{49}$$

Combining these two inequalities yields

$$\begin{aligned}
&\partial_l f\left(x_{k,0}\right)\sum_{i=0}^{n-1}\frac{m_{l,k,i}}{\sqrt{v_{l,k,i}}} \\
\geq&\frac{\partial_l f\left(x_{k,0}\right)}{\sqrt{v_{l,k,0}}}\left(\sum_{i\in i+} g_{l,k,0,\tau_{k,i}}\left(1-\frac{1-\beta_2}{2}\left(-1+\frac{4n\rho_2}{\beta_2^n}\right)\right)+\sum_{i\in i-}\frac{g_{l,k,i,\tau_{k,i}}}{\sqrt{\beta_2^n}}\right)- \\
&-\frac{\left|\partial_l f\left(x_{k,0}\right)\right|}{\sqrt{\beta_2^i v_{l,k,0}}}\left(\beta_1\left(\frac{1-\beta_1}{1-\beta_1^n}+1\right)n\sum_{q=0}^{n-1}\left|g_{l,k,0,\tau_{k,q}}\right|+\frac{\Delta_1 n}{\sqrt{kn}}\left(n+\beta_1 n^2\frac{1-\beta_1}{\left(1-\beta_1^n\right)^2}\left(1+2\sqrt{2}\right)\right)\right).
\end{aligned} \tag{50}$$

where $i+$ means the set of the indices of the components with the same sign of $\partial_l f\left(x_{k,0}\right)$ and $i-$ means the set of the indices of the components with opposite sign. Note that we have added 2 non-positive terms on the right hand side. For simplicity, define

$$C_7 \triangleq \frac{n^2}{\beta_2^{n/2}}\left(1+\beta_1 n\frac{1-\beta_1}{\left(1-\beta_1^n\right)^2}\left(1+2\sqrt{2}\right)\right).$$

Since

$$\sum_{i\in i+} g_{l,k,0,\tau_{k,i}}+\sum_{i\in i-} g_{l,k,0,\tau_{k,i}}=\partial_l f\left(x_{k,0}\right),$$

we have

$$
\begin{aligned}
&\partial_l f\left(x_{k,0}\right) \sum_{i=0}^{n-1} \frac{m_{l,k,i}}{\sqrt{v_{l,k,i}}} \\
\geq & \frac{\partial_l f\left(x_{k,0}\right)}{\sqrt{v_{l,k,0}}}\Big(\sum_{i \in i+} g_{l,k,0,\tau_{k,i}} \left(1-\frac{1-\beta_2}{2}\left(-1+\frac{4n\rho_2}{\beta_2^n}\right)\right) + \sum_{i \in i-} g_{l,k,i,\tau_{k,i}} \left(1+\frac{1}{\sqrt{\beta_2^n}}-1\right) + \\
& \frac{\beta_1}{\beta_2^n} \sum_i \left|g_{l,k,0,i}\right| \left(\frac{1-\beta_1}{1-\beta_1^n}+1\right) n\Big) - \frac{\left|\partial_l f\left(x_{k,0}\right)\right|}{\sqrt{v_{l,k,0}}} \frac{\Delta_1}{\sqrt{kn}} C_7 \\
\geq & \frac{\partial_l f\left(x_{k,0}\right)^2}{\sqrt{v_{l,k,0}}} - \frac{\left|\partial_l f\left(x_{k,0}\right)\right|}{\sqrt{v_{l,k,0}}} \frac{\Delta_1}{\sqrt{kn}} C_7 \\
& - \frac{\left|\partial_l f\left(x_{k,0}\right)\right|}{\sqrt{v_{l,k,0}}} \Bigg(\sum_{i \in i+} \left|g_{l,k,0,i}\right| \left(1-\beta_2\right) \frac{\left(-1+4n\rho_2\beta_2^{-n}\right)}{2} + \sum_{i \in i-} \left|g_{l,k,0,i}\right| \left(\frac{1}{\sqrt{\beta_2^n}}-1\right) \\
& + \frac{\beta_1}{\beta_2^n} \sum_i \left|g_{l,k,0,i}\right| \left(\frac{1-\beta_1}{1-\beta_1^n}+1\right) n \Bigg) \\
\geq & \frac{\partial_l f\left(x_{k,0}\right)^2}{\sqrt{v_{l,k,0}}} - \frac{n\sqrt{2n}}{\sqrt{\beta_2^n}} \frac{\Delta_1}{\sqrt{kn}} C_7 \\
& - \frac{\left|\partial_l f\left(x_{k,0}\right)\right|}{\sqrt{v_{l,k,0}}} \Bigg(\sum_{i \in i+} \left|g_{l,k,0,i}\right| \left(1-\beta_2\right) \frac{\left(-1+4n\rho_2\beta_2^{-n}\right)}{2} + \sum_{i \in i-} \left|g_{l,k,0,i}\right| \left(\frac{1}{\sqrt{\beta_2^n}}-1\right) \\
& + \frac{\beta_1}{\beta_2^n} \sum_i \left|g_{l,k,0,i}\right| \left(\frac{1-\beta_1}{1-\beta_1^n}+1\right) n \Bigg).
\end{aligned}
\tag{51}
$$

The last inequality holds due to Lemma F.1 and the fact that $\left|\partial_l f\left(x_{k,0}\right)\right| \leq n \left|g_{l,k}^b\right|$. It can further reduce to

$$
\begin{aligned}
&\partial_l f\left(x_{k,0}\right) \sum_{i=0}^{n-1} \frac{m_{l,k,i}}{\sqrt{v_{l,k,i}}} \\
\geq & \frac{\partial_l f\left(x_{k,0}\right)^2}{\sqrt{v_{l,k,0}}} - \frac{\left|\partial_l f\left(x_{k,0}\right)\right|}{\sqrt{v_{l,k,0}}} \sum_i \left|g_{l,k,0,i}\right| \left(\delta_1+\delta_2\right) - \frac{n\sqrt{2n}}{\sqrt{\beta_2^n}} \frac{\Delta_1}{\sqrt{kn}} C_7
\end{aligned}
\tag{52}
$$

where $\delta_1 = \left(1-\beta_2\right) \frac{\left(-1+\frac{4n\rho_2}{\beta_2^n}\right)}{2} + \left(\frac{1}{\sqrt{\beta_2^n}}-1\right)$ and $\delta_2 = \frac{\beta_1}{\beta_2^n}\left(\frac{1-\beta_1}{1-\beta_1^n}+1\right) n$.

When $\left|\partial_l f\left(x_{k,0}\right)\right|$ is smaller than $\frac{\Delta_1}{\sqrt{nk}}\left(\frac{32\sqrt{2}}{(1-\beta_2^n)\beta_2^n} n^2\right)$, the inequality is simply:

$$
\partial_l f\left(x_{k,0}\right) \sum_{i=0}^{n-1} \frac{m_{l,k,i}}{\sqrt{v_{l,k,i}}} \geq -\frac{\Delta_1}{\sqrt{nk}} \left(\frac{32\sqrt{2}}{(1-\beta_2^n)\beta_2^n}\right) n^2 \frac{n}{\sqrt{1-\beta_2}} \frac{1-\beta_1}{1-\frac{\beta_1}{\sqrt{\beta_2}}}
$$

because of Lemma C.2. We denote the large gradient components in the first case by "l large" and the rest components of the gradient by "l small". Summing up all of them, we have

$$
\begin{aligned}
&\sum_{l=1}^{d} \partial_l f\left(x_{k,0}\right) \sum_{i=0}^{n-1} \frac{m_{l,k,i}}{\sqrt{v_{l,k,i}}} \\
\geq & \sum_{l \text{ large}} \left(\frac{\partial_l f\left(x_{k,0}\right)^2}{\sqrt{v_{l,k,0}}} - \frac{\left|\partial_l f\left(x_{k,0}\right)\right|}{\sqrt{v_{l,k,0}}} \sum_i \left|g_{l,k,0,i}\right| \left(\delta_1+\delta_2\right) - \frac{m\sqrt{2n}}{\sqrt{\beta_2^n}} \frac{\Delta_1}{\sqrt{kn}} C_7 \right) \\
& + \sum_{l \text{ small}} -\frac{\Delta_1}{\sqrt{nk}} \left(\frac{32\sqrt{2}}{1-\beta_2^n}\right) d \frac{n}{\sqrt{1-\beta_2}} \frac{1-\beta_1}{1-\frac{\beta_1}{\sqrt{\beta_2}}}.
\end{aligned}
\tag{53}
$$

Since we have assumed that the largest component of the gradient is sufficiently large, we can further simplify the inequality to

$$\begin{aligned}&\sum_{l=1}^{d}\partial_l f\left(x_{k,0}\right)\sum_{i=0}^{n-1}\frac{m_{l,k,i}}{\sqrt{v_{l,k,i}}}\\ \geq&\sum_{l\text{ large}}\frac{\partial_l f\left(x_{k,0}\right)^2}{\sqrt{v_{l,k,0}}}-\sum_{l\text{ large}}\frac{\left|\partial_l f\left(x_{k,0}\right)\right|}{\sqrt{v_{l,k,0}}}\sum_i\left|g_{l,k,0,i}\right|\left(\delta_1+\delta_2\right)-\frac{\Delta_1}{\sqrt{k}}C_8,\end{aligned}\tag{54}$$

where $C_8=\frac{\sqrt{6}nd}{\beta_2^{n/2}}C_2+\left(\frac{32\sqrt{2}}{(1-\beta_2^n)\beta_2^n}\right)n^2\frac{\sqrt{n}}{\sqrt{1-\beta_2^n}}\frac{1-\beta_1}{1-\frac{\beta_1}{\sqrt{\beta_2}}}$.

Since $\left|\partial_\alpha f\left(x_{k,0}\right)\right|>\frac{\Delta_1}{\sqrt{nk}}\frac{32\sqrt{2}}{(1-\beta_2^n)\beta_2^n}n^2$, we have $\left|\partial_\alpha f\left(x_{k,0}\right)\right|>\frac{\Delta_1}{\sqrt{nk}}\frac{32\sqrt{2}}{(1-\beta_2^n)\beta_2^n}n^2$. Thus $\alpha\in l$ large. Furthermore, we keep only the α component in the first term, yielding

$$\begin{aligned}&\sum_{l=1}^{d}\partial_l f\left(x_{k,0}\right)\sum_{i=0}^{n-1}\frac{g_{l,k,i,\tau_{k,i}}}{\sqrt{v_{l,k,i}}}\\ \geq&\frac{\partial_\alpha f\left(x_{k,0}\right)^2}{\sqrt{v_{\alpha,k,0}}}-\sum_{l\text{ large}}\frac{\left|\partial_l f\left(x_{k,0}\right)\right|}{\sqrt{v_{l,k,0}}}\sum_i\left|g_{l,k,0,\tau_{k,i}}\right|\left(\delta_1+\delta_2\right)-\frac{\Delta_1}{\sqrt{nk}}C_8\\ =&\frac{\partial_\alpha f\left(x_{k,0}\right)^2}{\sqrt{v_{\alpha,k,0}}}\left(1-\frac{\sum_{l\text{ large}}\frac{\left|\partial_l f\left(x_{k,0}\right)\right|}{\sqrt{v_{l,k,0}}}\sum_i\left|g_{l,k,0,\tau_{k,i}}\right|\left(\delta_1+\delta_2\right)}{\frac{\partial_\alpha f\left(x_{k,0}\right)^2}{\sqrt{v_{\alpha,k,0}}}}\right)-\frac{\Delta_1}{\sqrt{nk}}C_8.\end{aligned}\tag{55}$$

We know from Lemma F.1 that for large l,

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

By our assumption:

$$\sum_{l=1}^{d}\sum_{i=0}^{n-1}g_{l,k,0,i}^2\leq D_1 d\left|\partial_\alpha f\left(x_{k,0}\right)\right|^2+D_0,$$

we have

$$\begin{aligned}&\sum_{l\text{ large}}\frac{\left|\partial_l f\left(x_{k,0}\right)\right|}{\sqrt{v_{l,k,0}}}\sum_{i=0}^{n-1}\left|g_{l,k,0,\tau_{k,i}}\right|\\ \leq&\sqrt{\frac{2\rho_3^2}{\beta_2^n}}\sum_{l=1}^{d}\sum_{i=0}^{n-1}\left|g_{l,k,0,\tau_{k,i}}\right|\\ \leq&\sqrt{\frac{2\rho_3^2}{\beta_2^n}}\sqrt{D_1 n}d\sqrt{\left|\partial_\alpha f\left(x_{k,0}\right)\right|^2+\frac{D_0}{D_1 d}}.\end{aligned}\tag{56}$$

The last inequality can be derived from Cauchy-Schwartz inequality. As a result,

$$\begin{aligned}&\sum_{l=1}^{d}\partial_l f\left(x_{k,0}\right)\sum_{i=0}^{n-1}\frac{m_{l,k,i}}{\sqrt{v_{l,k,i}}}\\ \geq&\frac{\partial_\alpha f\left(x_{k,0}\right)^2}{\sqrt{v_{\alpha,k,0}}}\left(1-\frac{\sqrt{\frac{2\rho_3^2}{\beta_2^n}}d\sqrt{D_1 n}\sqrt{\left|\partial_\alpha f\left(x_{k,0}\right)\right|^2+\frac{D_0}{D_1 d}}\left(\delta_1+\delta_2\right)}{\frac{\partial_\alpha f\left(x_{k,0}\right)^2}{\sqrt{v_{\alpha,k,0}}}}\right)-\frac{\Delta_1}{\sqrt{nk}}C_8\\ \geq&\frac{\partial_\alpha f\left(x_{k,0}\right)^2}{\sqrt{v_{\alpha,k,0}}}\left(1-T_1\left(\beta_1,\beta_2\right)-T_2\left(\beta_2\right)-\left(T_1+T_2\right)\frac{D_0}{\left|\partial_\alpha f\left(x_{k,0}\right)\right|^2 D_1 d}\right)-\frac{\Delta_1}{\sqrt{nk}}C_8\\ \geq&\frac{\left|\partial_\alpha f\left(x_{k,0}\right)\right|}{\sqrt{\frac{5}{2}D_1 d}}\left(1-T_1\left(\beta_1,\beta_2\right)-T_2\left(\beta_2\right)-\left(T_1+T_2\right)\frac{D_0}{\left|\partial_\alpha f\left(x_{k,0}\right)\right|^2 D_1 d}\right)-\frac{\Delta_1}{\sqrt{nk}}C_8\end{aligned}\tag{57}$$

where we have used Lemma F.2 since $k > 4$. T_2 is still defined as

$$T_2\left(\beta_2\right) = \sqrt{\frac{5d}{\beta_2^n}} d\rho_1\rho_3 D_1\delta_1 = \sqrt{\frac{5d}{\beta_2^n}} d\rho_1\rho_3 D_1 \left(\left(1-\beta_2\right) \frac{\left(\frac{4n\rho_2}{\beta_2^n} - 1\right)}{2} + \left(\frac{1}{\sqrt{\beta_2^n}} - 1 \right) \right), \quad (58)$$

and T_1 is defined as

$$T_1\left(\beta_1, \beta_2\right) = \sqrt{\frac{5d}{\beta_2^n}} d\rho_1\rho_3 D_1\delta_2 = \sqrt{\frac{5d}{\beta_2^n}} d\rho_1\rho_3 n D_1 \frac{\beta_1}{\beta_2^n} \left(\frac{1-\beta_1}{1-\beta_1^n} + 1 \right). \quad (59)$$

They approach zero when β_2 approaches one and β_1 approaches zero. This completes the proof.

If we use the bias corrected version, we need one additional constraint $1 + \frac{8\sqrt{2}}{1-\beta_2^n}$. □

Lemma G.2 is the Adam counterpart of Lemma F.4. Further, if we replace C_3 in Lemma F.4 with C_8 just defined and replace T_2 by $T_1 + T_2$, we can repeat the rest of the proof in Appendix F to prove Theorem 4.4. We omit the derivation and present the constants below:

$$\min_{k\in[t_{init},T]} \min\{\left\|\nabla f_{k,0}\right\|_1, \left\|\nabla f_{k,0}\right\|_2^2 \sqrt{\frac{D_1 d}{D_0}}\} \leq \frac{1}{\sqrt{T} - \sqrt{t_{init}-1}} \left(Q_{1,5} + Q_{2,5}\log\left(T+1\right)\right) + \sqrt{D_0} Q_{3,5}$$

where the constants are given by:

$$Q_{1,5} = \frac{f\left(x_{t_{init},0}\right) - f^* - C_9 \log t_{init}}{2\eta_1} \sqrt{5nD_1 d} d \quad (60)$$

$$Q_{2,5} = \frac{C_9\sqrt{5nD_1 d} d}{2\eta_1} \quad (61)$$

$$Q_{3,5} = C_{10}\sqrt{10D_1 dn} d \quad (62)$$

where C_9 and C_{10} defined as

$$\begin{aligned} C_9 \triangleq C_8 + \frac{1}{1-\beta_2} \max\{32\sqrt{2}n^2\beta_2^{-n}, \frac{4\sqrt{2}}{D_1 d}\}\big(\\ \max\{1, \frac{\Delta_1}{\left(1-\beta_2\right)\sqrt{n}} \max\{32\sqrt{2}n^2\beta_2^{-n}, \frac{4\sqrt{2}}{D_1 d}\}\}\big) + \frac{dn}{\sqrt{1-\beta_2}} \end{aligned} \quad (63)$$

$$C_{10} = \sqrt{\left(T_2 + T_1\right)\frac{\sqrt{2}-1}{2\sqrt{2}}\frac{1}{D_1 d}} \left(\frac{1}{\sqrt{10D_1 d}} \max\left\{1, \sqrt{\left(T_2+T_1\right)\frac{\sqrt{2}-1}{2\sqrt{2}}\frac{D_0}{D_1 d}}\right\} + \frac{dn}{\sqrt{1-\beta_2}} \right)$$