OPT2020: 12th Annual Workshop on Optimization for Machine Learning

On the Convergence of Adaptive Gradient Methods for Nonconvex Optimization

Dongruo Zhou* DRZHOU@CS.UCLA.EDU
Jinghui Chen* JINGHUIC@UCLA.EDU
Yuan Cao* YUANCAO@CS.UCLA.EDU
University of California, Los Angeles
Yiqi Tang TANG.1466@OSU.EDU
The Ohio State University
Ziyan Yang ZY3CX@VIRGINIA.EDU
University of Virginia
Quanquan Gu QGU@CS.UCLA.EDU
University of California, Los Angeles

Abstract

Adaptive gradient methods are workhorses in deep learning. However, the convergence guarantees of adaptive gradient methods for nonconvex optimization have not been thoroughly studied. In this paper, we provide a fine-grained convergence analysis for a general class of adaptive gradient methods including AMSGrad, RMSProp and AdaGrad. For smooth nonconvex functions, we prove that adaptive gradient methods in expectation converge to a first-order stationary point. Our convergence rate is better than existing results for adaptive gradient methods in terms of dimension, and is strictly faster than stochastic gradient decent (SGD) when the stochastic gradients are sparse. To the best of our knowledge, this is the first result showing the advantage of adaptive gradient methods over SGD in nonconvex setting. In addition, we also prove high probability bounds on the convergence rates of AMSGrad, RMSProp as well as AdaGrad, which have not been established before. Our analyses shed light on better understanding the mechanism behind adaptive gradient methods in optimizing nonconvex objectives.

1. Introduction

Stochastic gradient descent (SGD) [27] and its variants have been widely used in training deep neural networks. Among those variants, adaptive gradient methods (AdaGrad) [11, 21], which scale each coordinate of the gradient by a function of past gradients, can achieve better performance than vanilla SGD in practice when the gradients are sparse. An intuitive explanation for the success of AdaGrad is that it automatically adjusts the learning rate for each feature based on the partial gradient, which accelerates the convergence. However, AdaGrad was later found to demonstrate degraded performance especially in cases where the loss function is nonconvex or the gradient is dense, due to rapid decay of learning rate. This problem is especially exacerbated in deep learning due to the huge number of optimization variables. To overcome this issue, RMSProp [29] was proposed to use exponential moving average rather than the arithmetic average to scale the gradient, which mitigates the rapid decay of the learning rate. Kingma and Ba [16] proposed an adaptive momentum estimation method (Adam), which incorporates the idea of momentum [24, 28] into

RMSProp. Other related algorithms include AdaDelta [33] and Nadam [10], which combine the idea of exponential moving average of the historical gradients, Polyak's heavy ball [24] and Nesterov's accelerated gradient descent [23]. Recently, by revisiting the original convergence analysis of Adam, Reddi et al. [26] found that for some handcrafted simple convex optimization problem, Adam does not even converge to the global minimizer. In order to address this convergence issue of Adam, Reddi et al. [26] proposed a new variant of the Adam algorithm named AMSGrad, which has guaranteed convergence in the convex setting. The update rule of AMSGrad is as follows[1]:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \alpha_t \frac{\mathbf{m}_t}{\sqrt{\widehat{\mathbf{v}}_t + \epsilon}}, \text{ with } \widehat{\mathbf{v}}_t = \max(\widehat{\mathbf{v}}_{t-1}, \mathbf{v}_t), \tag{1}$$

where $\alpha_t > 0$ is the step size, ϵ is a small number to ensure numerical stability, $\mathbf{x}_t \in \mathbb{R}^d$ is the iterate in the t-th iteration, and $\mathbf{m}_t, \mathbf{v}_t \in \mathbb{R}^d$ are the exponential moving averages of the gradient and the squared gradient at the t-th iteration respectively[2]:

$$\mathbf{m}_t = \beta_1 \mathbf{m}_{t-1} + (1-\beta_1)\mathbf{g}_t, \quad \mathbf{v}_t = \beta_2 \mathbf{v}_{t-1} + (1-\beta_2)\mathbf{g}_t^2. \tag{2}$$

Here $\beta_1, \beta_2 \in [0, 1]$ are algorithm hyperparameters, and $\mathbf{g}_t$ is the stochastic gradient at $\mathbf{x}_t$.

Despite the successes of adaptive gradient methods for training deep neural networks, the convergence guarantees for these algorithms are mostly restricted to online convex optimization [11, 16, 26]. Therefore, there is a huge gap between existing online convex optimization guarantees for adaptive gradient methods and the empirical successes of adaptive gradient methods in nonconvex optimization. In order to bridge this gap, there are a few recent attempts to prove the nonconvex optimization guarantees for adaptive gradient methods. More specifically, Basu et al. [5] proved the convergence rate of RMSProp and Adam when using deterministic gradient rather than stochastic gradient. Li and Orabona [18] proved the convergence rate of AdaGrad, assuming the gradient is L-Lipschitz continuous. Ward et al. [30] proved the convergence rate of AdaGrad-Norm where the moving average of the norms of the gradient vectors is used to adjust the gradient vector in both deterministic and stochastic settings for smooth nonconvex functions. Nevertheless, the convergence guarantees in Basu et al. [5], Ward et al. [30] are still limited to simplified algorithms. Another attempt to obtain the convergence rate under stochastic setting is prompted recently by Zou and Shen [35], in which they only focus on the condition when the momentum vanishes. Chen et al. [7] studies the convergence properties of adaptive gradient methods in the nonconvex setting, however, its convergence rate has a quadratic dependency on the problem dimension d. Défossez et al. [9] proves the convergence of Adam and Adagrad in nonconvex smooth optimization under the assumption of almost sure uniform bound on the L_∞ norm of the gradients. In this paper, we provide a fine-grained convergence analysis of the adaptive gradient methods. In particular, we analyze several representative adaptive gradient methods, i.e., AMSGrad [26], which fixed the non-convergence issue in Adam and the RMSProp (fixed version via Reddi et al. [26]), and prove its convergence rate for smooth nonconvex objective functions in the stochastic optimization setting. Moreover, existing theoretical guarantees for adaptive gradient methods are mostly bounds in expectation over the randomness of stochastic gradients, and are therefore only on-average convergence guarantees. However, in practice the optimization algorithm is usually only run once, and therefore the performance cannot be

1. With slight abuse of notation, here we denote by $\sqrt{\mathbf{v}_t}$ the element-wise square root of the vector $\mathbf{v}_t$, $\mathbf{m}_t/\sqrt{\mathbf{v}_t}$ the element-wise division between $\mathbf{m}_t$ and $\sqrt{\mathbf{v}_t}$, and $\max(\widehat{\mathbf{v}}_{t-1}, \mathbf{v}_t)$ the element-wise maximum between $\widehat{\mathbf{v}}_{t-1}$ and $\mathbf{v}_t$.
2. Here we denote by $\mathbf{g}_t^2$ the element-wise square of the vector $\mathbf{g}_t$.

guaranteed by the in-expectation bounds. To deal with this problem, we also provide high probability convergence rates for AMSGrad and RMSProp, which can characterize the performance of the algorithms on single runs.

1.1. Our Contributions

The main contributions of our work are summarized as follows:

- We prove that the convergence rate of AMSGrad to a stationary point for stochastic nonconvex optimization is $O(d^{1/2}/T^{3/4-s/2} + d/T)$ when $\|\mathbf{g}_{1:T,i}\|_2 \leq G_\infty T^s$. Here $\mathbf{g}_{1:T,i} = [g_{1,i}, g_{2,i}, \ldots, g_{T,i}]^\top$ with $\{\mathbf{g}_t\}_{t=1}^T$ being the stochastic gradients satisfying $\|\mathbf{g}_t\|_\infty \leq G_\infty$, and $s \in [0, 1/2]$ is a parameter that characterizes the growth rate of the cumulative stochastic gradient $\mathbf{g}_{1:T,i}$. When the stochastic gradients are sparse, i.e., $s < 1/2$, AMSGrad achieves strictly faster convergence rate than that of vanilla SGD [13] in terms of iteration number T.
- Our result implies that the worst case (i.e., $s = 1/2$) convergence rate for AMSGrad is $O(\sqrt{d/T} + d/T)$, which has a better dependence on the dimension d and T than the convergence rate proved in Chen et al. [7], i.e., $O((\log T + d^2)/\sqrt{T})$.
- We also establish high probability bounds for adaptive gradient methods. To the best of our knowledge, it is the first high probability convergence guarantees for AMSGrad and RMSProp in nonconvex stochastic optimization setting.

2. Convergence of Adaptive Gradient Methods in Nonconvex Optimization

In this section we present our main theoretical results on the convergence of AMSGrad, RMSProp as well as AdaGrad. We study the following stochastic nonconvex optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) := \mathbb{E}_\xi\big[f(\mathbf{x};\xi)\big],$$

where ξ is a random variable satisfying certain distribution, $f(\mathbf{x};\xi) : \mathbb{R}^d \to \mathbb{R}$ is a L-smooth nonconvex function. In the stochastic setting, one cannot directly access the full gradient of $f(\mathbf{x})$. Instead, one can only get unbiased estimators of the gradient of $f(\mathbf{x})$, which is $\nabla f(\mathbf{x};\xi)$. This setting has been studied in Ghadimi and Lan [12, 13].

Assumption 1 (Bounded Gradient) *$f(\mathbf{x}) = \mathbb{E}_\xi f(\mathbf{x};\xi)$ has G_∞-bounded stochastic gradient. That is, for any ξ, we assume that $\|\nabla f(\mathbf{x};\xi)\|_\infty \leq G_\infty$.*

It is worth mentioning that Assumption 1 is slightly weaker than the ℓ_2-boundedness assumption $\|\nabla f(\mathbf{x};\xi)\|_2 \leq G_2$ used in [7, 25]. Since $\|\nabla f(\mathbf{x};\xi)\|_\infty \leq \|\nabla f(\mathbf{x};\xi)\|_2 \leq \sqrt{d}\|\nabla f(\mathbf{x};\xi)\|_\infty$, the ℓ_2-boundedness assumption implies Assumption 1 with $G_\infty = G_2$. Meanwhile, G_∞ will be tighter than G_2 by a factor of $\sqrt{d}$ when each coordinate of $\nabla f(\mathbf{x};\xi)$ almost equals to each other.

Assumption 2 (L-smooth) *$f(\mathbf{x}) = \mathbb{E}_\xi f(\mathbf{x};\xi)$ is L-smooth: for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we have*

$$\big|f(\mathbf{x}) - f(\mathbf{y}) - \langle\nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle\big| \leq \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|_2^2.$$

Assumption 2 is a standard assumption in the analysis of gradient-based algorithms. It is equivalent to the L-gradient Lipschitz condition, which is often written as $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2$.

We are now ready to present our main result.

Theorem 3 (AMSGrad) *Suppose $\beta_1 < \beta_2^{1/2}$, $\alpha_t = \alpha$ and $\|\mathbf{g}_{1:T,i}\|_2 \leq G_\infty T^s$ for $t = 1, \ldots, T, 0 \leq s \leq 1/2$. Then under Assumptions 1 and 2, the iterates $\mathbf{x}_t$ of AMSGrad satisfy that*

$$\frac{1}{T-1}\sum_{t=2}^{T}\mathbb{E}\big[\|\nabla f(\mathbf{x}_t)\|_2^2\big] \leq \frac{M_1}{T\alpha} + \frac{M_2 d}{T} + \frac{\alpha M_3 d}{T^{1/2-s}}, \tag{3}$$

where $\{M_i\}_{i=1}^3$ are defined as follows:

$$M_1 = 2G_\infty\Delta, M_2 = \frac{2G_\infty^3\epsilon^{-1/2}}{1-\beta_1} + 2G_\infty^2, M_3 = \frac{2LG_\infty^2}{\epsilon^{1/2}(1-\beta_2)^{1/2}(1-\beta_1/\beta_2^{1/2})}\left(1 + \frac{2\beta_1^2}{1-\beta_1}\right),$$

and $\Delta = f(\mathbf{x}_1) - \inf_{\mathbf{x}} f(\mathbf{x})$.

Remark 4 *Note that in Theorem 3 we have a condition that $\|\mathbf{g}_{1:T,i}\|_2 \leq G_\infty T^s$. Here s characterizes the growth rate condition [20] of $\mathbf{g}_{1:T,i}$, i.e., the cumulative stochastic gradient. In the worse case where the stochastic gradients are not sparse at all, $s = 1/2$, while in practice when the stochastic gradients are sparse, we have $s < 1/2$.*

Remark 5 *If we choose $\alpha = \Theta\big(d^{1/2}T^{1/4+s/2}\big)^{-1}$, then (3) implies that AMSGrad achieves an $O(d^{1/2}/T^{3/4-s/2} + d/T)$ convergence rate. In cases where the stochastic gradients are sparse, i.e., $s < 1/2$, we can see that the convergence rate of AMSGrad is strictly better than that of nonconvex SGD [13], i.e., $O(\sqrt{d/T} + d/T)$. In the worst case when $s = 1/2$, this result matches the convergence rate of nonconvex SGD [13]. Note that Chen et al. [7] also provided similar bound for AMSGrad that*

$$\frac{1}{T-1}\sum_{t=2}^{T}\mathbb{E}\big[\|\nabla f(\mathbf{x}_t)\|_2^2\big] = O\left(\frac{\log T + d^2}{\sqrt{T}}\right).$$

It can be seen that the dependence of d in their bound is quadratic, which is worse than the linear dependence implied by (3). A very recent work [9] discussed the convergence issue of Adam by showing that the bound consists of a constant term and does not converge to zero. In comparison, our result for AMSGrad does not have such a constant term and converges to zero in a rate $O(d^{1/2}T^{3/4-s/2})$. This demonstrates that the convergence issue of Adam is indeed fixed in AMSGrad.

Corollary 6 (corrected version of RMSProp) *Under the same conditions of Theorem 3, if $\alpha_t = \alpha$ and $\|\mathbf{g}_{1:T,i}\|_2 \leq G_\infty T^s$ for $t = 1, \ldots, T, 0 \leq s \leq 1/2$, then the the iterates $\mathbf{x}_t$ of RMSProp satisfy that*

$$\frac{1}{T-1}\sum_{t=2}^{T}\mathbb{E}\big[\|\nabla f(\mathbf{x}_t)\|_2^2\big] \leq \frac{M_1}{T\alpha} + \frac{M_2 d}{T} + \frac{\alpha M_3 d}{T^{1/2-s}},$$

where $\{M_i\}_{i=1}^3$ are defined as follows:

$$M_1 = 2G_\infty\Delta f,\ M_2 = 2G_\infty^3\epsilon^{-1/2} + 2G_\infty^2,\ M_3 = \frac{6LG_\infty^2}{\epsilon^{1/2}(1-\beta)^{1/2}},$$

Corollary 7 (AdaGrad) *Under the same conditions of Theorem 3, if* $\alpha_t = \alpha$ *and* $\|\mathbf{g}_{1:T,i}\|_2 \leq G_\infty T^s$ *for* $t = 1, \ldots, T, 0 \leq s \leq 1/2$*, then the the iterates* $\mathbf{x}_t$ *of AdaGrad satisfy that*

$$\frac{1}{T-1}\sum_{t=2}^{T}\mathbb{E}\big[\|\nabla f(\mathbf{x}_t)\|_2^2\big] \leq \frac{M_1}{T\alpha} + \frac{M_2 d}{T} + \frac{\alpha M_3 d}{T^{1/2-s}},$$

where $\{M_i\}_{i=1}^3$ *are defined as follows:*

$$M_1 = 2G_\infty \Delta f,\ M_2 = 2G_\infty^3 \epsilon^{-1/2} + 2G_\infty^2,\ M_3 = 6LG_\infty^2\epsilon^{-1/2},$$

Remark 8 *Corollaries 6, 7 imply that RMSProp and AdaGrad algorithm achieve the same rate of convergence as AMSGrad. In worst case where* $s = 1/2$*, both algorithms again, achieves* $O(\sqrt{d/T} + d/T)$ *convergence rate, which matches the convergences rate of nonconvex SGD given by Ghadimi and Lan [13].*

Remark 9 *Défossez et al. [9] gave a bound* $O(\alpha^{-1}T^{-1/2} + (1+\alpha)dT^{-1/2})$ *for Adagrad, which gives the an* $O(1/\sqrt{T} + d/\sqrt{T})$ *rate with optimal* α*. For the ease of comparison, we calculate the iteration complexity of both results. To converge to a* ϵ_{target}*-approximate first order stationary point, the result of Défossez et al. [9] suggests that AdaGrad requires* $\Omega(\epsilon_{target}^{-2}d^2)$ *iterations. In sharp contrast, our result suggests that AdaGrad only requires* $\Omega(\epsilon_{target}^{-2}d)$ *iterations. Evidently, our iteration complexity is better than theirs by a factor of* d*.*

2.1. High Probability Bounds

Theorem 3, Corollaries 6 and 7 bound the expectation of full gradients over the randomness of stochastic gradients. In other words, these bounds can only guarantee the average performance of a large number of trials of the algorithm, but cannot rule out extremely bad solutions. What's more, for practical applications such as training deep neural networks, usually we only perform one single run of the algorithm since the training time can be fairly large. Hence, it is essential to get high probability bounds which guarantee the performance of the algorithm on single runs. To overcome this limitation, in this section, we further establish high probability bounds of the convergence rate for AMSGrad and RMSProp as well as AdaGrad. We make the following additional assumption.

Assumption 10 *The stochastic gradients are sub-Gaussian random vectors [14]:*

$$\mathbb{E}_\xi[\exp(\langle \mathbf{v}, \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}, \xi)\rangle)] \leq \exp(\|\mathbf{v}\|_2^2\sigma^2/2)$$

for all $\mathbf{v} \in \mathbb{R}^d$ *and all* $\mathbf{x}$*.*

Remark 11 *Sub-Gaussian gradient assumptions are commonly considered when studying high probability bounds [19]. Note that our Assumption 10 is weaker than Assumption B2 in Li and Orabona [19]: for the case when* $\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}, \xi)$ *is a standard Gaussian vector,* σ^2 *defined in Li and Orabona [19] is of order* $O(d)$*, while* $\sigma^2 = O(1)$ *in our definition.*

Theorem 12 (AMSGrad) *Suppose* $\beta_1 < \beta_2^{1/2}$*,* $\alpha_t = \alpha \leq 1/2$ *and* $\|\mathbf{g}_{1:T,i}\|_2 \leq G_\infty T^s$ *for* $t = 1, \ldots, T, 0 \leq s \leq 1/2$*. Then for any* $\delta > 0$*, under Assumptions 1, 2 and 10, with probability at least* $1-\delta$*, the iterates* $\mathbf{x}_t$ *of AMSGrad satisfy that*

$$\frac{1}{T-1}\sum_{t=2}^{T}\|\nabla f(\mathbf{x}_t)\|_2^2 \leq \frac{M_1}{T\alpha} + \frac{M_2 d}{T} + \frac{\alpha M_3 d}{T^{1/2-s}}, \tag{4}$$

where $\{M_i\}_{i=1}^3$ *are defined as follows:*

$$M_1 = 4G_\infty \Delta f + C' G_\infty \epsilon^{-1} \sigma^2 G_\infty \log(2/\delta), M_2 = \frac{4G_\infty^3 \epsilon^{-1/2}}{1-\beta_1} + 4G_\infty^2,$$
$$M_3 = \frac{4LG_\infty^2}{\epsilon^{1/2}(1-\beta_2)^{1/2}(1-\beta_1/\beta_2^{1/2})}\left(1 + \frac{2\beta_1^2}{1-\beta_1}\right),$$

and $\Delta f = f(\mathbf{x}_1) - \inf_{\mathbf{x}} f(\mathbf{x})$.

Remark 13 *Similar to the discussion in Remark 5, we can choose* $\alpha = \Theta\big(d^{1/2}T^{1/4+s/2}\big)^{-1}$, *to achieve an* $O(d^{1/2}/T^{3/4-s/2} + d/T)$ *convergence rate. When* $s < 1/2$, *this rate of AMSGrad is strictly better than that of nonconvex SGD [13].*

We also have the following corollaries characterizing the high probability bounds for RMSProp and AdaGrad.

Corollary 14 (corrected version of RMSProp) *Under the same conditions of Theorem 12, if* $\alpha_t = \alpha \leq 1/2$ *and* $\|\mathbf{g}_{1:T,i}\|_2 \leq G_\infty T^s$ *for* $t = 1, \ldots, T, 0 \leq s \leq 1/2$, *then for any* $\delta > 0$, *with probability at least* $1-\delta$, *the iterates* $\mathbf{x}_t$ *of RMSProf satisfy that*

$$\frac{1}{T-1}\sum_{t=2}^{T} \|\nabla f(\mathbf{x}_t)\|_2^2 \leq \frac{M_1}{T\alpha} + \frac{M_2 d}{T} + \frac{\alpha M_3 d}{T^{1/2-s}}, \tag{5}$$

where $\{M_i\}_{i=1}^3$ *are defined as follows:*

$$M_1 = 4G_\infty \Delta + C' G_\infty \epsilon^{-1} \sigma^2 G_\infty \log(2/\delta), M_2 = 4G_\infty^3 \epsilon^{-1/2} + 4G_\infty^2, \ M_3 = \frac{4LG_\infty^2}{\epsilon^{1/2}(1-\beta)^{1/2}},$$

and $\Delta = f(\mathbf{x}_1) - \inf_{\mathbf{x}} f(\mathbf{x})$.

Corollary 15 (AdaGrad) *Under the same conditions of Theorem 12, if* $\alpha_t = \alpha \leq 1/2$ *and* $\|\mathbf{g}_{1:T,i}\|_2 \leq G_\infty T^s$ *for* $t = 1, \ldots, T, 0 \leq s \leq 1/2$, *then for any* $\delta > 0$, *with probability at least* $1-\delta$, *the iterates* $\mathbf{x}_t$ *of AdaGrad satisfy that*

$$\frac{1}{T-1}\sum_{t=2}^{T} \|\nabla f(\mathbf{x}_t)\|_2^2 \leq \frac{M_1}{T\alpha} + \frac{M_2 d}{T} + \frac{\alpha M_3 d}{T^{1/2-s}}, \tag{6}$$

where $\{M_i\}_{i=1}^3$ *are defined as follows:*

$$M_1 = 4G_\infty \Delta + C' G_\infty \epsilon^{-1} \sigma^2 G_\infty \log(2/\delta), M_2 = 4G_\infty^3 \epsilon^{-1/2} + 4G_\infty^2, \ M_3 = \frac{4LG_\infty^2}{\epsilon^{1/2}},$$

and $\Delta = f(\mathbf{x}_1) - \inf_{\mathbf{x}} f(\mathbf{x})$.

3. Conclusions

In this paper, we provided a fine-grained analysis of a general class of adaptive gradient methods, and proved their convergence rates for smooth nonconvex optimization. Our results provide faster convergence rates of AMSGrad and the corrected version of RMSProp as well as AdaGrad for smooth nonconvex optimization compared with previous works. In addition, we also prove high probability bounds on the convergence rates of AMSGrad and RMSProp as well as AdaGrad, which have not been established before.

References

[1] Zeyuan Allen-Zhu. Natasha: Faster non-convex stochastic optimization via strongly non-convex parameter. In *International Conference on Machine Learning*, pages 89–97, 2017.

[2] Zeyuan Allen-Zhu. Natasha 2: Faster non-convex optimization than sgd. *arXiv preprint arXiv:1708.08694*, 2017.

[3] Zeyuan Allen-Zhu. How to make the gradients small stochastically. *arXiv preprint arXiv:1801.02982*, 2018.

[4] Zeyuan Allen-Zhu and Elad Hazan. Variance reduction for faster non-convex optimization. In *International Conference on Machine Learning*, pages 699–707, 2016.

[5] Amitabh Basu, Soham De, Anirbit Mukherjee, and Enayat Ullah. Convergence guarantees for rmsprop and adam in non-convex optimization and their comparison to nesterov acceleration on autoencoders. *arXiv preprint arXiv:1807.06766*, 2018.

[6] Jinghui Chen, Dongruo Zhou, Yiqi Tang, Ziyan Yang, Yuan Cao, and Quanquan Gu. Closing the generalization gap of adaptive gradient methods in training deep neural networks. In *International Joint Conferences on Artificial Intelligence*, 2020.

[7] Xiangyi Chen, Sijia Liu, Ruoyu Sun, and Mingyi Hong. On the convergence of a class of adam-type algorithms for nonconvex optimization. *arXiv preprint arXiv:1808.02941*, 2018.

[8] Zaiyi Chen, Yi Xu, Enhong Chen, and Tianbao Yang. Sadagrad: Strongly adaptive stochastic gradient methods. In *International Conference on Machine Learning*, pages 913–921, 2018.

[9] Alexandre Défossez, Léon Bottou, Francis Bach, and Nicolas Usunier. On the convergence of adam and adagrad. *arXiv preprint arXiv:2003.02395*, 2020.

[10] Timothy Dozat. Incorporating nesterov momentum into adam. 2016.

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

[12] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[13] Saeed Ghadimi and Guanghui Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1-2):59–99, 2016.

[14] Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M Kakade, and Michael I Jordan. A short note on concentration inequalities for random vectors with subgaussian norm. *arXiv preprint arXiv:1902.03736*, 2019.

[15] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323, 2013.

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

[17] Lihua Lei, Cheng Ju, Jianbo Chen, and Michael I Jordan. Non-convex finite-sum optimization via scsg methods. In *Advances in Neural Information Processing Systems*, pages 2345–2355, 2017.

[18] Xiaoyu Li and Francesco Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. *arXiv preprint arXiv:1805.08114*, 2018.

[19] Xiaoyu Li and Francesco Orabona. A high probability analysis of adaptive sgd with momentum. *arXiv preprint arXiv:2007.14294*, 2020.

[20] Mingrui Liu, Youssef Mroueh, Jerret Ross, Wei Zhang, Xiaodong Cui, Payel Das, and Tianbao Yang. Towards better understanding of adaptive gradient algorithms in generative adversarial nets. *arXiv preprint arXiv:1912.11940*, 2019.

[21] H Brendan McMahan and Matthew Streeter. Adaptive bound optimization for online convex optimization. *arXiv preprint arXiv:1002.4908*, 2010.

[22] Mahesh Chandra Mukkamala and Matthias Hein. Variants of rmsprop and adagrad with logarithmic regret bounds. In *ICML*, 2017.

[23] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[24] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

[25] Sashank J. Reddi, Ahmed Hefny, Suvrit Sra, Barnabas Poczos, and Alex Smola. Stochastic variance reduction for nonconvex optimization. pages 314–323, 2016.

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

[27] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[28] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147, 2013.

[29] T. Tieleman and G. Hinton. Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning, 2012.

[30] Rachel Ward, Xiaoxia Wu, and Leon Bottou. Adagrad stepsizes: Sharp convergence over nonconvex landscapes, from any initialization. *arXiv preprint arXiv:1806.01811*, 2018.

[31] Tianbao Yang, Qihang Lin, and Zhe Li. Unified convergence analysis of stochastic momentum methods for convex and non-convex optimization. *arXiv preprint arXiv:1604.03257*, 2016.

[32] Manzil Zaheer, Sashank Reddi, Devendra Sachan, Satyen Kale, and Sanjiv Kumar. Adaptive methods for nonconvex optimization. In *Advances in neural information processing systems*, pages 9793–9803, 2018.

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

[34] Dongruo Zhou, Pan Xu, and Quanquan Gu. Stochastic nested variance reduction for nonconvex optimization. *arXiv preprint arXiv:1806.07811*, 2018.

[35] Fangyu Zou and Li Shen. On the convergence of adagrad with momentum for training deep neural networks. *arXiv preprint arXiv:1808.03408*, 2018.

Appendix A. Additional Related Work

Here we briefly review other related work that is not covered before. Mukkamala and Hein [22] proposed SC-Adagrad / SC-RMSprop, which derives logarithmic regret bounds for strongly convex functions. Chen et al. [8] proposed SADAGRAD for solving stochastic strongly convex optimization and more generally stochastic convex optimization that satisfies the second order growth condition. Zaheer et al. [32] studied the effect of adaptive denominator constant ϵ and minibatch size in the convergence of adaptive gradient methods. Chen et al. [6] proposed a partially adaptive gradient method for closing the generalization gap between SGD and adaptive gradient method and proved its convergence in nonconvex settings.

We also review other related work on nonconvex stochastic optimization. Ghadimi and Lan [12] proposed a randomized stochastic gradient (RSG) method, and proved its $O(1/\sqrt{T})$ convergence rate to a stationary point. Ghadimi and Lan [13] proposed an randomized stochastic accelerated gradient (RSAG) method, which achieves $O(1/T + \sigma^2/\sqrt{T})$ convergence rate, where σ^2 is an upper bound on the variance of the stochastic gradient. Motivated by the success of stochastic momentum methods in deep learning [28], Yang et al. [31] provided a unified convergence analysis for both stochastic heavy-ball method and the stochastic variant of Nesterov's accelerated gradient method, and proved $O(1/\sqrt{T})$ convergence rate to a stationary point for smooth nonconvex functions. Allen-Zhu and Hazan [4], Reddi et al. [25] proposed variants of stochastic variance-reduced gradient (SVRG) method [15] that is provably faster than gradient descent in the nonconvex finite-sum setting. Lei et al. [17] proposed a stochastically controlled stochastic gradient (SCSG), which further improves convergence rate of SVRG for finite-sum smooth nonconvex optimization. Recently, Zhou et al. [34] proposed a new algorithm called stochastic nested variance-reduced gradient (SNVRG), which achieves strictly better gradient complexity than both SVRG and SCSG for finite-sum and stochastic smooth nonconvex optimization.

There is another line of research in stochastic smooth nonconvex optimization, which makes use of the λ-nonconvexity of a nonconvex function f (i.e., $\nabla^2 f \succeq -\lambda \mathbf{I}$). More specifically, Natasha 1 [1] and Natasha 1.5 [2] have been proposed, which solve a modified regularized problem and achieve faster convergence rate to first-order stationary points than SVRG and SCSG in the finite-sum and stochastic settings respectively. In addition, Allen-Zhu [3] proposed an SGD4 algorithm, which optimizes a series of regularized problems, and is able to achieve a faster convergence rate than SGD.

Appendix B. Notation

Scalars are denoted by lower case letters, vectors by lower case bold face letters, and matrices by upper case bold face letters. For a vector $\mathbf{x} = [x_i] \in \mathbb{R}^d$, we denote the ℓ_p norm ($p \geq 1$) of $\mathbf{x}$ by $\|\mathbf{x}\|_p = \big(\sum_{i=1}^d |x_i|^p\big)^{1/p}$, the ℓ_∞ norm of $\mathbf{x}$ by $\|\mathbf{x}\|_\infty = \max_{i=1}^d |x_i|$. For a sequence of vectors $\{\mathbf{g}_j\}_{j=1}^t$, we denote by $g_{j,i}$ the i-th element in $\mathbf{g}_j$. We also denote $\mathbf{g}_{1:t,i} = [g_{1,i}, g_{2,i}, \ldots, g_{t,i}]^\top$. With slightly abuse of notation, for any two vectors $\mathbf{a}$ and $\mathbf{b}$, we denote $\mathbf{a}^2$ as the element-wise square, $\mathbf{a}^p$ as the element-wise power operation, $\mathbf{a}/\mathbf{b}$ as the element-wise division and $\max(\mathbf{a}, \mathbf{b})$ as the element-wise maximum. For a matrix $\mathbf{A} = [A_{ij}] \in \mathbb{R}^{d\times d}$, we define $\|\mathbf{A}\|_{1,1} = \sum_{i,j=1}^d |A_{ij}|$. Given two sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n = O(b_n)$ if there exists a constant $0 < C < +\infty$ such that $a_n \leq C\, b_n$. We use notation $\widetilde{O}(\cdot)$ to hide logarithmic factors.

Appendix C. Algorithms

We mainly consider the following three algorithms: AMSGrad [26], a corrected version of RMSProp [26, 29], and AdaGrad [11].

The AMSGrad algorithm is originally proposed by Reddi et al. [26] to fix the non-convergence issue in the original Adam optimizer [16]. Specifically, in Algorithm 1, the effective learning rate of AMSGrad is $\alpha_t \widehat{\mathbf{V}}_t^{-1/2}$ where $\widehat{\mathbf{V}}_t = \mathrm{diag}(\widehat{\mathbf{v}}_t)$, while in original Adam, the effective learning rate is $\alpha_t \mathbf{V}_t^{-1/2}$ where $\mathbf{V}_t = \mathrm{diag}(\mathbf{v}_t)$. This choice of effective learning rate guarantees that it is non-increasing and thus fix the possible convergence issue. In Algorithm 2 we present the corrected version of RMSProp [29] where the effective learning rate is also set as $\alpha_t \widehat{\mathbf{V}}_t^{-1/2}$.

In Algorithm 3 we further present the AdaGrad algorithm [11], which adopts the summation of past stochastic gradient squares instead of the running average of them to compute the effective learning rate.

Algorithm 1 AMSGrad [26]

Input: initial point $\mathbf{x}_1$, step size $\{\alpha_t\}_{t=1}^T$, $\beta_1, \beta_2, \epsilon$.

$\mathbf{m}_0 \leftarrow 0$, $\widehat{\mathbf{v}}_0 \leftarrow 0$, $\mathbf{v}_0 \leftarrow 0$
for $t = 1$ to T **do**
  $\mathbf{g}_t = \nabla f(\mathbf{x}_t, \xi_t)$
  $\mathbf{m}_t = \beta_1 \mathbf{m}_{t-1} + (1-\beta_1)\mathbf{g}_t$
  $\mathbf{v}_t = \beta_2 \mathbf{v}_{t-1} + (1-\beta_2)\mathbf{g}_t^2$
  $\widehat{\mathbf{v}}_t = \max(\widehat{\mathbf{v}}_{t-1}, \mathbf{v}_t)$
  $\mathbf{x}_{t+1} = \mathbf{x}_t - \alpha_t \widehat{\mathbf{V}}_t^{-1/2} \mathbf{m}_t$ with $\widehat{\mathbf{V}}_t = \mathrm{diag}(\widehat{\mathbf{v}}_t + \epsilon)$
end for

Output: Choose $\mathbf{x}_{\text{out}}$ from $\{\mathbf{x}_t\}, 2 \le t \le T$ with probability $\alpha_{t-1} / \sum_{i=1}^{T-1} \alpha_i$.

Algorithm 2 RMSProp [29] (corrected version by Reddi et al. [26])

Input: initial point $\mathbf{x}_1$, step size $\{\alpha_t\}_{t=1}^T$, β, ϵ.

$\widehat{\mathbf{v}}_0 \leftarrow 0$, $\mathbf{v}_0 \leftarrow 0$
for $t = 1$ to T **do**
  $\mathbf{g}_t = \nabla f(\mathbf{x}_t, \xi_t)$
  $\mathbf{v}_t = \beta \mathbf{v}_{t-1} + (1-\beta)\mathbf{g}_t^2$
  $\widehat{\mathbf{v}}_t = \max(\widehat{\mathbf{v}}_{t-1}, \mathbf{v}_t)$
  $\mathbf{x}_{t+1} = \mathbf{x}_t - \alpha_t \widehat{\mathbf{V}}_t^{-1/2} \mathbf{g}_t$ with $\widehat{\mathbf{V}}_t = \mathrm{diag}(\widehat{\mathbf{v}}_t + \epsilon)$
end for

Output: Choose $\mathbf{x}_{\text{out}}$ from $\{\mathbf{x}_t\}, 2 \le t \le T$ with probability $\alpha_{t-1} / \sum_{i=1}^{T-1} \alpha_i$.

Appendix D. Detailed Proof of the Main Theory

Here we provide the detailed proof of the main theorem.

Algorithm 3 AdaGrad [11]

Input: initial point $\mathbf{x}_1$, step size $\{\alpha_t\}_{t=1}^T$, ϵ.

1: $\widehat{\mathbf{v}}_0 \leftarrow 0$
2: **for** $t = 1$ to T **do**
3: $\quad \mathbf{g}_t = \nabla f(\mathbf{x}_t, \xi_t)$
4: $\quad \widehat{\mathbf{v}}_t = \widehat{\mathbf{v}}_{t-1} + \mathbf{g}_t^2$
5: $\quad \mathbf{x}_{t+1} = \mathbf{x}_t - \alpha_t \widehat{\mathbf{V}}_t^{-1/2} \mathbf{g}_t$ with $\widehat{\mathbf{V}}_t = \mathrm{diag}(\widehat{\mathbf{v}}_t + \epsilon)$
6: **end for**

Output: Choose $\mathbf{x}_{\text{out}}$ from $\{\mathbf{x}_t\}, 2 \le t \le T$ with probability $\alpha_{t-1} / \sum_{i=1}^{T-1} \alpha_i$.

D.1. Proof of Theorem 3

Let $\mathbf{x}_0 = \mathbf{x}_1$. To prove Theorem 3, we need the following lemmas:

Lemma 16 *Let $\widehat{\mathbf{v}}_t$ and $\mathbf{m}_t$ be as defined in Algorithm 1. Then under Assumption 1, we have $\|\nabla f(\mathbf{x})\|_\infty \le G_\infty$, $\|\widehat{\mathbf{v}}_t\|_\infty \le G_\infty^2$ and $\|\mathbf{m}_t\|_\infty \le G_\infty$.*

Lemma 17 *Let $\beta_1, \beta_2, \beta_1', \beta_2'$ be the weight parameters such that*

$$\mathbf{m}_t = \beta_1 \mathbf{m}_{t-1} + (1 - \beta_1') \mathbf{g}_t,$$
$$\mathbf{v}_t = \beta_2 \mathbf{v}_{t-1} + (1 - \beta_2') \mathbf{g}_t^2,$$

α_t, $t = 1, \ldots, T$ be the step sizes. We denote $\gamma = \beta_1 / \beta_2^{1/2}$. Suppose that $\alpha_t = \alpha$ and $\gamma \le 1$, then under Assumption 1, we have the following two results:

$$\sum_{t=1}^T \alpha_t^2 \big\| \widehat{\mathbf{V}}_t^{-1/2} \mathbf{m}_t \big\|_2^2 \le \frac{T^{1/2} \alpha_t^2 (1 - \beta_1')}{2 \epsilon^{1/2} (1 - \beta_2')^{1/2} (1 - \gamma)} \sum_{i=1}^d \|\mathbf{g}_{1:T,i}\|_2,$$

and

$$\sum_{t=1}^T \alpha_t^2 \big\| \widehat{\mathbf{V}}_t^{-1/2} \mathbf{g}_t \big\|_2^2 \le \frac{T^{1/2} \alpha_t^2}{2 \epsilon^{1/2} (1 - \beta_2')^{1/2} (1 - \gamma)} \sum_{i=1}^d \|\mathbf{g}_{1:T,i}\|_2.$$

Note that Lemma 17 is general and applicable to various algorithms. Specifically, set $\beta_1' = \beta_1$ and $\beta_2' = \beta_2$, we recover the case in Algorithm 1. Further set $\beta_1 = 0$ we recover the case in Algorithm 2. Set $\beta_1' = \beta_1 = 0$ and $\beta_2 = 1, \beta_2' = 0$ we recover the case in Algorithm 3.

To deal with stochastic momentum $\mathbf{m}_t$ and stochastic weight $\widehat{\mathbf{V}}_t^{-1/2}$, following Yang et al. [31], we define an auxiliary sequence $\mathbf{z}_t$ as follows: let $\mathbf{x}_0 = \mathbf{x}_1$, and for each $t \ge 1$,

$$\mathbf{z}_t = \mathbf{x}_t + \frac{\beta_1}{1 - \beta_1} (\mathbf{x}_t - \mathbf{x}_{t-1}) = \frac{1}{1 - \beta_1} \mathbf{x}_t - \frac{\beta_1}{1 - \beta_1} \mathbf{x}_{t-1}. \tag{7}$$

Lemma 18 shows that $\mathbf{z}_{t+1} - \mathbf{z}_t$ can be represented in two different ways.

Lemma 18 *Let $\mathbf{z}_t$ be defined in (7). For $t \geq 2$, we have*

$$\mathbf{z}_{t+1} - \mathbf{z}_t = \frac{\beta_1}{1-\beta_1}\Big[\mathbf{I} - \big(\alpha_t \widehat{\mathbf{V}}_t^{-1/2}\big)\big(\alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2}\big)^{-1}\Big](\mathbf{x}_{t-1} - \mathbf{x}_t) - \alpha_t \widehat{\mathbf{V}}_t^{-1/2}\mathbf{g}_t. \tag{8}$$

and

$$\mathbf{z}_{t+1} - \mathbf{z}_t = \frac{\beta_1}{1-\beta_1}\big(\alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2} - \alpha_t \widehat{\mathbf{V}}_t^{-1/2}\big)\mathbf{m}_{t-1} - \alpha_t \widehat{\mathbf{V}}_t^{-1/2}\mathbf{g}_t. \tag{9}$$

For $t = 1$, we have

$$\mathbf{z}_2 - \mathbf{z}_1 = -\alpha_1 \widehat{\mathbf{V}}_1^{-1/2}\mathbf{g}_1. \tag{10}$$

By Lemma 18, we connect $\mathbf{z}_{t+1} - \mathbf{z}_t$ with $\mathbf{x}_{t+1} - \mathbf{x}_t$ and $\alpha_t \widehat{\mathbf{V}}_t^{-1/2}\mathbf{g}_t$

The following two lemmas give bounds on $\|\mathbf{z}_{t+1} - \mathbf{z}_t\|_2$ and $\|\nabla f(\mathbf{z}_t) - \nabla f(\mathbf{x}_t)\|_2$, which play important roles in our proof.

Lemma 19 *Let $\mathbf{z}_t$ be defined in (7). For $t \geq 2$, we have*

$$\|\mathbf{z}_{t+1} - \mathbf{z}_t\|_2 \leq \big\|\alpha \widehat{\mathbf{V}}_t^{-1/2}\mathbf{g}_t\big\|_2 + \frac{\beta_1}{1-\beta_1}\|\mathbf{x}_{t-1} - \mathbf{x}_t\|_2.$$

Lemma 20 *Let $\mathbf{z}_t$ be defined in (7). For $t \geq 2$, we have*

$$\|\nabla f(\mathbf{z}_t) - \nabla f(\mathbf{x}_t)\|_2 \leq L\Big(\frac{\beta_1}{1-\beta_1}\Big) \cdot \|\mathbf{x}_t - \mathbf{x}_{t-1}\|_2.$$

Now we are ready to prove Theorem 3.

Proof [Proof of Theorem 3] Since f is L-smooth, we have:

$$\begin{aligned} f(\mathbf{z}_{t+1}) &\leq f(\mathbf{z}_t) + \nabla f(\mathbf{z}_t)^\top(\mathbf{z}_{t+1} - \mathbf{z}_t) + \frac{L}{2}\|\mathbf{z}_{t+1} - \mathbf{z}_t\|_2^2 \\ &= f(\mathbf{z}_t) + \underbrace{\nabla f(\mathbf{x}_t)^\top(\mathbf{z}_{t+1} - \mathbf{z}_t)}_{I_1} + \underbrace{(\nabla f(\mathbf{z}_t) - \nabla f(\mathbf{x}_t))^\top(\mathbf{z}_{t+1} - \mathbf{z}_t)}_{I_2} + \underbrace{\frac{L}{2}\|\mathbf{z}_{t+1} - \mathbf{z}_t\|_2^2}_{I_3} \end{aligned} \tag{11}$$

In the following, we bound I_1, I_2 and I_3 separately.

Bounding term I_1: When $t = 1$, we have

$$\nabla f(\mathbf{x}_1)^\top(\mathbf{z}_2 - \mathbf{z}_1) = -\nabla f(\mathbf{x}_1)^\top \alpha_1 \widehat{\mathbf{V}}_t^{-1/2}\mathbf{g}_1. \tag{12}$$

For $t \geq 2$, we have

$$\begin{aligned} &\nabla f(\mathbf{x}_t)^\top(\mathbf{z}_{t+1} - \mathbf{z}_t) \\ &= \nabla f(\mathbf{x}_t)^\top\left[\frac{\beta_1}{1-\beta_1}\big(\alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2} - \alpha_t \widehat{\mathbf{V}}_t^{-1/2}\big)\mathbf{m}_{t-1} - \alpha_t \widehat{\mathbf{V}}_t^{-1/2}\mathbf{g}_t\right] \\ &= \frac{\beta_1}{1-\beta_1}\nabla f(\mathbf{x}_t)^\top\big(\alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2} - \alpha_t \widehat{\mathbf{V}}_t^{-1/2}\big)\mathbf{m}_{t-1} - \nabla f(\mathbf{x}_t)^\top \alpha_t \widehat{\mathbf{V}}_t^{-1/2}\mathbf{g}_t, \end{aligned} \tag{13}$$

where the first equality holds due to (9) in Lemma 18. For $\nabla f(\mathbf{x}_t)^\top(\alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2} - \alpha_t\widehat{\mathbf{V}}_t^{-1/2})\mathbf{m}_{t-1}$ in (13), we have

$$\begin{aligned}\nabla f(\mathbf{x}_t)^\top(\alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2} - \alpha_t\widehat{\mathbf{V}}_t^{-1/2})\mathbf{m}_{t-1} &\leq \|\nabla f(\mathbf{x}_t)\|_\infty \cdot \big\|\alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2} - \alpha_t\widehat{\mathbf{V}}_t^{-1/2}\big\|_{1,1} \cdot \|\mathbf{m}_{t-1}\|_\infty \\ &\leq G_\infty^2\Big[\big\|\alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2}\big\|_{1,1} - \big\|\alpha_t\widehat{\mathbf{V}}_t^{-1/2}\big\|_{1,1}\Big]. \end{aligned} \tag{14}$$

The first inequality holds because for a positive diagonal matrix $\mathbf{A}$, we have $\mathbf{x}^\top\mathbf{A}\mathbf{y} \leq \|\mathbf{x}\|_\infty \cdot \|\mathbf{A}\|_{1,1} \cdot \|\mathbf{y}\|_\infty$. The second inequality holds due to $\alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2} \succeq \alpha_t\widehat{\mathbf{V}}_t^{-1/2} \succeq 0$. Next we bound $-\nabla f(\mathbf{x}_t)^\top\alpha_t\widehat{\mathbf{V}}_t^{-1/2}\mathbf{g}_t$. We have

$$\begin{aligned}&-\nabla f(\mathbf{x}_t)^\top\alpha_t\widehat{\mathbf{V}}_t^{-1/2}\mathbf{g}_t \\ &= -\nabla f(\mathbf{x}_t)^\top\alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2}\mathbf{g}_t - \nabla f(\mathbf{x}_t)^\top\big(\alpha_t\widehat{\mathbf{V}}_t^{-1/2} - \alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2}\big)\mathbf{g}_t \\ &\leq -\nabla f(\mathbf{x}_t)^\top\alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2}\mathbf{g}_t + \|\nabla f(\mathbf{x}_t)\|_\infty \cdot \big\|\alpha_t\widehat{\mathbf{V}}_t^{-1/2} - \alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2}\big\|_{1,1} \cdot \|\mathbf{g}_t\|_\infty \\ &\leq -\nabla f(\mathbf{x}_t)^\top\alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2}\mathbf{g}_t + G_\infty^2\Big(\big\|\alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2}\big\|_{1,1} - \big\|\alpha_t\widehat{\mathbf{V}}_t^{-1/2}\big\|_{1,1}\Big). \end{aligned} \tag{15}$$

The first inequality holds because for a positive diagonal matrix $\mathbf{A}$, we have $\mathbf{x}^\top\mathbf{A}\mathbf{y} \leq \|\mathbf{x}\|_\infty \cdot \|\mathbf{A}\|_{1,1} \cdot \|\mathbf{y}\|_\infty$. The second inequality holds due to $\alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2} \succeq \alpha_t\widehat{\mathbf{V}}_t^{-1/2} \succeq 0$. Substituting (14) and (15) into (13), we have

$$\nabla f(\mathbf{x}_t)^\top(\mathbf{z}_{t+1} - \mathbf{z}_t) \leq -\nabla f(\mathbf{x}_t)^\top\alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2}\mathbf{g}_t + \frac{1}{1-\beta_1}G_\infty^2\Big(\big\|\alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2}\big\|_{1,1} - \big\|\alpha_t\widehat{\mathbf{V}}_t^{-1/2}\big\|_{1,1}\Big). \tag{16}$$

Bounding term I_2: For $t \geq 1$, we have

$$\begin{aligned}&\big(\nabla f(\mathbf{z}_t) - \nabla f(\mathbf{x}_t)\big)^\top(\mathbf{z}_{t+1} - \mathbf{z}_t) \\ &\leq \big\|\nabla f(\mathbf{z}_t) - \nabla f(\mathbf{x}_t)\big\|_2 \cdot \|\mathbf{z}_{t+1} - \mathbf{z}_t\|_2 \\ &\leq \Big(\big\|\alpha_t\widehat{\mathbf{V}}_t^{-1/2}\mathbf{g}_t\big\|_2 + \frac{\beta_1}{1-\beta_1}\|\mathbf{x}_{t-1} - \mathbf{x}_t\|_2\Big) \cdot \frac{\beta_1}{1-\beta_1} \cdot L\|\mathbf{x}_t - \mathbf{x}_{t-1}\|_2 \\ &= L\frac{\beta_1}{1-\beta_1}\big\|\alpha_t\widehat{\mathbf{V}}_t^{-1/2}\mathbf{g}_t\big\|_2 \cdot \|\mathbf{x}_t - \mathbf{x}_{t-1}\|_2 + L\bigg(\frac{\beta_1}{1-\beta_1}\bigg)^2\|\mathbf{x}_t - \mathbf{x}_{t-1}\|_2^2 \\ &\leq L\big\|\alpha_t\widehat{\mathbf{V}}_t^{-1/2}\mathbf{g}_t\big\|_2^2 + 2L\bigg(\frac{\beta_1}{1-\beta_1}\bigg)^2\|\mathbf{x}_t - \mathbf{x}_{t-1}\|_2^2, \end{aligned} \tag{17}$$

where the second inequality holds because of Lemma 18 and Lemma 19, the last inequality holds due to Young's inequality.

Bounding term I_3: For $t \geq 1$, we have

$$\begin{aligned}\frac{L}{2}\|\mathbf{z}_{t+1} - \mathbf{z}_t\|_2^2 &\leq \frac{L}{2}\bigg[\big\|\alpha_t\widehat{\mathbf{V}}_t^{-1/2}\mathbf{g}_t\big\|_2 + \frac{\beta_1}{1-\beta_1}\|\mathbf{x}_{t-1} - \mathbf{x}_t\|_2\bigg]^2 \\ &\leq L\big\|\alpha_t\widehat{\mathbf{V}}_t^{-1/2}\mathbf{g}_t\big\|_2^2 + 2L\bigg(\frac{\beta_1}{1-\beta_1}\bigg)^2\|\mathbf{x}_{t-1} - \mathbf{x}_t\|_2^2. \end{aligned} \tag{18}$$

The first inequality is obtained by introducing Lemma 18.

For $t = 1$, substituting (12), (17) and (18) into (11), taking expectation and rearranging terms, we have

$$\begin{aligned}
&\mathbb{E}[f(\mathbf{z}_2) - f(\mathbf{z}_1)] \\
&\leq \mathbb{E}\bigg[- \nabla f(\mathbf{x}_1)^\top \alpha_1 \widehat{\mathbf{V}}_1^{-1/2} \mathbf{g}_1 + 2L \big\| \alpha_1 \widehat{\mathbf{V}}_1^{-1/2} \mathbf{g}_1 \big\|_2^2 + 4L \bigg(\frac{\beta_1}{1-\beta_1} \bigg)^2 \|\mathbf{x}_1 - \mathbf{x}_0\|_2^2 \bigg] \\
&= \mathbb{E}[- \nabla f(\mathbf{x}_1)^\top \alpha_1 \widehat{\mathbf{V}}_1^{-1/2} \mathbf{g}_1 + 2L \big\| \alpha_1 \widehat{\mathbf{V}}_1^{-1/2} \mathbf{g}_1 \big\|_2^2] \\
&\leq \mathbb{E}[d \alpha_1 G_\infty + 2L \big\| \alpha_1 \widehat{\mathbf{V}}_1^{-1/2} \mathbf{g}_1 \big\|_2^2], \qquad (19)
\end{aligned}$$

where the last inequality holds because

$$- \nabla f(\mathbf{x}_1)^\top \widehat{\mathbf{V}}_1^{-1/2} \mathbf{g}_1 \leq d \cdot \|\nabla f(\mathbf{x}_1)\|_\infty \cdot \|\widehat{\mathbf{V}}_1^{-1/2} \mathbf{g}_1\|_\infty \leq d G_\infty .$$

For $t \geq 2$, substituting (16), (17) and (18) into (11), taking expectation and rearranging terms, we have

$$\begin{aligned}
&\mathbb{E}\bigg[f(\mathbf{z}_{t+1}) + \frac{G_\infty^2 \big\| \alpha_t \widehat{\mathbf{V}}_t^{-1/2} \big\|_{1,1}}{1-\beta_1} - \bigg(f(\mathbf{z}_t) + \frac{G_\infty^2 \big\| \alpha_{t-1} \widehat{\mathbf{V}}_{t-1}^{-1/2} \big\|_{1,1}}{1-\beta_1} \bigg) \bigg] \\
&\leq \mathbb{E}\bigg[- \nabla f(\mathbf{x}_t)^\top \alpha_{t-1} \widehat{\mathbf{V}}_{t-1}^{-1/2} \mathbf{g}_t + 2L \big\| \alpha_t \widehat{\mathbf{V}}_t^{-1/2} \mathbf{g}_t \big\|_2^2 + 4L \bigg(\frac{\beta_1}{1-\beta_1} \bigg)^2 \|\mathbf{x}_t - \mathbf{x}_{t-1}\|_2^2 \bigg] \\
&= \mathbb{E}\bigg[- \nabla f(\mathbf{x}_t)^\top \alpha_{t-1} \widehat{\mathbf{V}}_{t-1}^{-1/2} \nabla f(\mathbf{x}_t) + 2L \big\| \alpha_t \widehat{\mathbf{V}}_t^{-1/2} \mathbf{g}_t \big\|_2^2 + 4L \bigg(\frac{\beta_1}{1-\beta_1} \bigg)^2 \big\| \alpha_{t-1} \widehat{\mathbf{V}}_{t-1}^{-1/2} \mathbf{m}_{t-1} \big\|_2^2 \bigg] \\
&\leq \mathbb{E}\bigg[- \alpha_{t-1} \big\| \nabla f(\mathbf{x}_t) \big\|_2^2 G_\infty^{-1} + 2L \big\| \alpha_t \widehat{\mathbf{V}}_t^{-1/2} \mathbf{g}_t \big\|_2^2 + 4L \bigg(\frac{\beta_1}{1-\beta_1} \bigg)^2 \big\| \alpha_{t-1} \widehat{\mathbf{V}}_{t-1}^{-1/2} \mathbf{m}_{t-1} \big\|_2^2 \bigg], \qquad (20)
\end{aligned}$$

where the equality holds because $\mathbb{E}[\mathbf{g}_t] = \nabla f(\mathbf{x}_t)$ conditioned on $\nabla f(\mathbf{x}_t)$ and $\widehat{\mathbf{V}}_{t-1}^{-1/2}$, the second inequality holds because of Lemma 16. Telescoping (20) for $t = 2$ to T and adding with (19), we have

$$\begin{aligned}
&G_\infty^{-1} \sum_{t=2}^T \alpha_{t-1} \mathbb{E} \big\| \nabla f(\mathbf{x}_t) \big\|_2^2 \\
&\leq \mathbb{E}\bigg[f(\mathbf{z}_1) + \frac{G_\infty^2 \big\| \alpha_1 \widehat{\mathbf{V}}_1^{-1/2} \big\|_{1,1}}{1-\beta_1} + d \alpha_1 G_\infty - \bigg(f(\mathbf{z}_{T+1}) + \frac{G_\infty^2 \big\| \alpha_T \widehat{\mathbf{v}}_T^{-1/2} \big\|_1}{1-\beta_1} \bigg) \bigg] \\
&\quad + 2L \sum_{t=1}^T \mathbb{E} \big\| \alpha_t \widehat{\mathbf{V}}_t^{-1/2} \mathbf{g}_t \big\|_2^2 + 4L \bigg(\frac{\beta_1}{1-\beta_1} \bigg)^2 \sum_{t=2}^T \mathbb{E}\Big[\big\| \alpha_{t-1} \widehat{\mathbf{V}}_{t-1}^{-1/2} \mathbf{m}_{t-1} \big\|_2^2 \Big] \\
&\leq \mathbb{E}\bigg[\Delta f + \frac{G_\infty^2 \alpha_1 \epsilon^{-1/2} d}{1-\beta_1} + d \alpha_1 G_\infty \bigg] + 2L \sum_{t=1}^T \mathbb{E} \big\| \alpha_t \widehat{\mathbf{V}}_t^{-1/2} \mathbf{g}_t \big\|_2^2 \\
&\quad + 4L \bigg(\frac{\beta_1}{1-\beta_1} \bigg)^2 \sum_{t=1}^T \mathbb{E}\Big[\big\| \alpha_t \widehat{\mathbf{V}}_t^{-1/2} \mathbf{m}_t \big\|_2^2 \Big]. \qquad (21)
\end{aligned}$$

By Lemma 17, we have

$$\sum_{t=1}^{T}\alpha_t^2\mathbb{E}\Big[\|\widehat{\mathbf{V}}_t^{-1/2}\mathbf{m}_t\|_2^2\Big] \le \frac{T^{1/2}\alpha_t^2(1-\beta_1)}{2\epsilon^{1/2}(1-\beta_2)^{1/2}(1-\gamma)}\mathbb{E}\bigg(\sum_{i=1}^{d}\|\mathbf{g}_{1:T,i}\|_2\bigg), \tag{22}$$

where $\gamma = \beta_1/\beta_2^{1/2}$. We also have

$$\sum_{t=1}^{T}\alpha_t^2\mathbb{E}\Big[\|\widehat{\mathbf{V}}_t^{-1/2}\mathbf{g}_t\|_2^2\Big] \le \frac{T^{1/2}\alpha_t^2}{2\epsilon^{1/2}(1-\beta_2)^{1/2}(1-\gamma)}\mathbb{E}\bigg(\sum_{i=1}^{d}\|\mathbf{g}_{1:T,i}\|_2\bigg). \tag{23}$$

Substituting (22) and (23) into (21), and rearranging (21), we have

$$\begin{aligned}
&\mathbb{E}\|\nabla f(\mathbf{x}_{\text{out}})\|_2^2 \\
&= \frac{1}{\sum_{t=2}^{T}\alpha_{t-1}}\sum_{t=2}^{T}\alpha_{t-1}\mathbb{E}\big\|\nabla f(\mathbf{x}_t)\big\|_2^2 \\
&\le \frac{G_\infty}{\sum_{t=2}^{T}\alpha_{t-1}}\mathbb{E}\bigg[\Delta f + \frac{G_\infty^2\alpha_1\epsilon^{-1/2}d}{1-\beta_1} + d\alpha_1 G_\infty\bigg] \\
&\quad + \frac{2LG_\infty}{\sum_{t=2}^{T}\alpha_{t-1}}\cdot\frac{T^{1/2}\alpha_t^2}{2\epsilon^{1/2}(1-\beta_2)^{1/2}(1-\gamma)}\mathbb{E}\bigg(\sum_{i=1}^{d}\|\mathbf{g}_{1:T,i}\|_2\bigg)^{1-q} \\
&\quad + \frac{4LG_\infty}{\sum_{t=2}^{T}\alpha_{t-1}}\bigg(\frac{\beta_1}{1-\beta_1}\bigg)^2\frac{T^{1/2}\alpha_t^2(1-\beta_1)}{2\epsilon^{1/2}(1-\beta_2)^{1/2}(1-\gamma)}\mathbb{E}\bigg(\sum_{i=1}^{d}\|\mathbf{g}_{1:T,i}\|_2\bigg)^{1-q} \\
&\le \frac{1}{T\alpha}2G_\infty\Delta f + \frac{2}{T}\bigg(\frac{G_\infty^3\epsilon^{-1/2}d}{1-\beta_1} + dG_\infty^2\bigg) \\
&\quad + \frac{2G_\infty L\alpha}{T^{1/2}\epsilon^{1/2}(1-\gamma)(1-\beta_2)^{1/2}}\mathbb{E}\bigg(\sum_{i=1}^{d}\|\mathbf{g}_{1:T,i}\|_2\bigg)\bigg(1+2(1-\beta_1)\bigg(\frac{\beta_1}{1-\beta_1}\bigg)^2\bigg),
\end{aligned} \tag{24}$$

where the second inequality holds because $\alpha_t = \alpha$. Rearranging (24), and note that in the theorem condition we have $\|\mathbf{g}_{1:T,i}\|_2 \le G_\infty T^s$, we obtain

$$\mathbb{E}\|\nabla f(\mathbf{x}_{\text{out}})\|_2^2 \le \frac{M_1}{T\alpha} + \frac{M_2 d}{T} + \frac{\alpha M_3 d}{T^{1/2-s}},$$

where $\{M_i\}_{i=1}^3$ are defined in Theorem 3. This completes the proof. ■

D.2. Proof of Corollary 6

Proof [Proof of Corollary 6] Following the proof for Theorem 3, setting $\beta_1' = \beta_1 = 0$ and $\beta_2' = \beta_2 = \beta$ in Lemma 17 we get the conclusion. ■

D.3. Proof of Corollary 7

Proof [Proof of Corollary 7] Following the proof for Theorem 3, setting $\beta_1' = \beta_1 = 0$, $\beta_2 = 1$ and $\beta_2' = 0$ in Lemma 17 we get the conclusion. ■

D.4. Proof of Theorem 12

Proof [Proof of Theorem 12] Since f is L-smooth, we have:

$$\begin{aligned} f(\mathbf{z}_{t+1}) &\leq f(\mathbf{z}_t) + \nabla f(\mathbf{z}_t)^\top(\mathbf{z}_{t+1} - \mathbf{z}_t) + \frac{L}{2}\|\mathbf{z}_{t+1} - \mathbf{z}_t\|_2^2 \\ &= f(\mathbf{z}_t) + \underbrace{\nabla f(\mathbf{x}_t)^\top(\mathbf{z}_{t+1} - \mathbf{z}_t)}_{I_1} + \underbrace{(\nabla f(\mathbf{z}_t) - \nabla f(\mathbf{x}_t))^\top(\mathbf{z}_{t+1} - \mathbf{z}_t)}_{I_2} + \underbrace{\frac{L}{2}\|\mathbf{z}_{t+1} - \mathbf{z}_t\|_2^2}_{I_3} \end{aligned} \tag{25}$$

In the following, we bound I_1, I_2 and I_3 separately.
Bounding term I_1: When $t = 1$, we have

$$\nabla f(\mathbf{x}_1)^\top(\mathbf{z}_2 - \mathbf{z}_1) = -\nabla f(\mathbf{x}_1)^\top \alpha_1 \widehat{\mathbf{V}}_t^{-1/2}\mathbf{g}_1. \tag{26}$$

For $t \geq 2$, we have

$$\begin{aligned} &\nabla f(\mathbf{x}_t)^\top(\mathbf{z}_{t+1} - \mathbf{z}_t) \\ &= \nabla f(\mathbf{x}_t)^\top\bigg[\frac{\beta_1}{1-\beta_1}\big(\alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2} - \alpha_t\widehat{\mathbf{V}}_t^{-1/2}\big)\mathbf{m}_{t-1} - \alpha_t\widehat{\mathbf{V}}_t^{-1/2}\mathbf{g}_t\bigg] \\ &= \frac{\beta_1}{1-\beta_1}\nabla f(\mathbf{x}_t)^\top\big(\alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2} - \alpha_t\widehat{\mathbf{V}}_t^{-1/2}\big)\mathbf{m}_{t-1} - \nabla f(\mathbf{x}_t)^\top\alpha_t\widehat{\mathbf{V}}_t^{-1/2}\mathbf{g}_t, \end{aligned} \tag{27}$$

where the first equality holds due to (9) in Lemma 18. For $\nabla f(\mathbf{x}_t)^\top(\alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2} - \alpha_t\widehat{\mathbf{V}}_t^{-1/2})\mathbf{m}_{t-1}$ in (27), we have

$$\begin{aligned} \nabla f(\mathbf{x}_t)^\top(\alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2} - \alpha_t\widehat{\mathbf{V}}_t^{-1/2})\mathbf{m}_{t-1} &\leq \|\nabla f(\mathbf{x}_t)\|_\infty \cdot \big\|\alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2} - \alpha_t\widehat{\mathbf{V}}_t^{-1/2}\big\|_{1,1} \cdot \|\mathbf{m}_{t-1}\|_\infty \\ &\leq G_\infty^2\Big[\big\|\alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2}\big\|_{1,1} - \big\|\alpha_t\widehat{\mathbf{V}}_t^{-1/2}\big\|_{1,1}\Big]. \end{aligned} \tag{28}$$

The first inequality holds because for a positive diagonal matrix $\mathbf{A}$, we have $\mathbf{x}^\top\mathbf{A}\mathbf{y} \leq \|\mathbf{x}\|_\infty \cdot \|\mathbf{A}\|_{1,1} \cdot \|\mathbf{y}\|_\infty$. The second inequality holds due to $\alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2} \succeq \alpha_t\widehat{\mathbf{V}}_t^{-1/2} \succeq 0$. Next we bound $-\nabla f(\mathbf{x}_t)^\top\alpha_t\widehat{\mathbf{V}}_t^{-1/2}\mathbf{g}_t$. We have

$$\begin{aligned} &-\nabla f(\mathbf{x}_t)^\top\alpha_t\widehat{\mathbf{V}}_t^{-1/2}\mathbf{g}_t \\ &= -\nabla f(\mathbf{x}_t)^\top\alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2}\mathbf{g}_t - \nabla f(\mathbf{x}_t)^\top\big(\alpha_t\widehat{\mathbf{V}}_t^{-1/2} - \alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2}\big)\mathbf{g}_t \\ &\leq -\nabla f(\mathbf{x}_t)^\top\alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2}\mathbf{g}_t + \|\nabla f(\mathbf{x}_t)\|_\infty \cdot \big\|\alpha_t\widehat{\mathbf{V}}_t^{-1/2} - \alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2}\big\|_{1,1} \cdot \|\mathbf{g}_t\|_\infty \\ &\leq -\nabla f(\mathbf{x}_t)^\top\alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2}\mathbf{g}_t + G_\infty^2\Big(\big\|\alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2}\big\|_{1,1} - \big\|\alpha_t\widehat{\mathbf{V}}_t^{-1/2}\big\|_{1,1}\Big). \end{aligned} \tag{29}$$

The first inequality holds because for a positive diagonal matrix $\mathbf{A}$, we have $\mathbf{x}^\top\mathbf{A}\mathbf{y} \leq \|\mathbf{x}\|_\infty \cdot \|\mathbf{A}\|_{1,1} \cdot \|\mathbf{y}\|_\infty$. The second inequality holds due to $\alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2} \succeq \alpha_t\widehat{\mathbf{V}}_t^{-1/2} \succeq 0$. Substituting (28) and (29) into (27), we have

$$\nabla f(\mathbf{x}_t)^\top(\mathbf{z}_{t+1} - \mathbf{z}_t) \leq -\nabla f(\mathbf{x}_t)^\top\alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2}\mathbf{g}_t + \frac{1}{1-\beta_1}G_\infty^2\Big(\big\|\alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2}\big\|_{1,1} - \big\|\alpha_t\widehat{\mathbf{V}}_t^{-1/2}\big\|_{1,1}\Big). \tag{30}$$

Bounding term I_2: For $t \geq 1$, we have

$$\begin{aligned}
&\big(\nabla f(\mathbf{z}_t) - \nabla f(\mathbf{x}_t)\big)^\top (\mathbf{z}_{t+1} - \mathbf{z}_t) \\
&\leq \big\|\nabla f(\mathbf{z}_t) - \nabla f(\mathbf{x}_t)\big\|_2 \cdot \|\mathbf{z}_{t+1} - \mathbf{z}_t\|_2 \\
&\leq \bigg(\big\|\alpha_t \widehat{\mathbf{V}}_t^{-1/2} \mathbf{g}_t\big\|_2 + \frac{\beta_1}{1-\beta_1}\|\mathbf{x}_{t-1} - \mathbf{x}_t\|_2\bigg) \cdot \frac{\beta_1}{1-\beta_1} \cdot L\|\mathbf{x}_t - \mathbf{x}_{t-1}\|_2 \\
&= L\frac{\beta_1}{1-\beta_1}\big\|\alpha_t \widehat{\mathbf{V}}_t^{-1/2} \mathbf{g}_t\big\|_2 \cdot \|\mathbf{x}_t - \mathbf{x}_{t-1}\|_2 + L\bigg(\frac{\beta_1}{1-\beta_1}\bigg)^2 \|\mathbf{x}_t - \mathbf{x}_{t-1}\|_2^2 \\
&\leq L\big\|\alpha_t \widehat{\mathbf{V}}_t^{-1/2} \mathbf{g}_t\big\|_2^2 + 2L\bigg(\frac{\beta_1}{1-\beta_1}\bigg)^2 \|\mathbf{x}_t - \mathbf{x}_{t-1}\|_2^2, \qquad (31)
\end{aligned}$$

where the second inequality holds because of Lemma 18 and Lemma 19, the last inequality holds due to Young's inequality.

Bounding term I_3: For $t \geq 1$, we have

$$\begin{aligned}
\frac{L}{2}\|\mathbf{z}_{t+1} - \mathbf{z}_t\|_2^2 &\leq \frac{L}{2}\bigg[\big\|\alpha_t \widehat{\mathbf{V}}_t^{-1/2} \mathbf{g}_t\big\|_2 + \frac{\beta_1}{1-\beta_1}\|\mathbf{x}_{t-1} - \mathbf{x}_t\|_2\bigg]^2 \\
&\leq L\big\|\alpha_t \widehat{\mathbf{V}}_t^{-1/2} \mathbf{g}_t\big\|_2^2 + 2L\bigg(\frac{\beta_1}{1-\beta_1}\bigg)^2 \|\mathbf{x}_{t-1} - \mathbf{x}_t\|_2^2. \qquad (32)
\end{aligned}$$

The first inequality is obtained by introducing Lemma 18.

For $t = 1$, substituting (26), (31) and (32) into (25) and rearranging terms, we have

$$\begin{aligned}
f(\mathbf{z}_2) - f(\mathbf{z}_1) &\leq -\nabla f(\mathbf{x}_1)^\top \alpha_1 \widehat{\mathbf{V}}_1^{-1/2} \mathbf{g}_1 + 2L\big\|\alpha_1 \widehat{\mathbf{V}}_1^{-1/2} \mathbf{g}_1\big\|_2^2 + 4L\bigg(\frac{\beta_1}{1-\beta_1}\bigg)^2 \|\mathbf{x}_1 - \mathbf{x}_0\|_2^2 \\
&= -\nabla f(\mathbf{x}_1)^\top \alpha_1 \widehat{\mathbf{V}}_1^{-1/2} \mathbf{g}_1 + 2L\big\|\alpha_1 \widehat{\mathbf{V}}_1^{-1/2} \mathbf{g}_1\big\|_2^2 \\
&\leq d\alpha_1 G_\infty + 2L\big\|\alpha_1 \widehat{\mathbf{V}}_1^{-1/2} \mathbf{g}_1\big\|_2^2, \qquad (33)
\end{aligned}$$

where the last inequality holds because

$$-\nabla f(\mathbf{x}_1)^\top \widehat{\mathbf{V}}_1^{-1/2} \mathbf{g}_1 \leq d \cdot \|\nabla f(\mathbf{x}_1)\|_\infty \cdot \|\widehat{\mathbf{V}}_1^{-1/2} \mathbf{g}_1\|_\infty \leq dG_\infty.$$

For $t \geq 2$, substituting (30), (31) and (32) into (25) and rearranging terms, we have

$$\begin{aligned}
&f(\mathbf{z}_{t+1}) + \frac{G_\infty^2 \big\|\alpha_t \widehat{\mathbf{V}}_t^{-1/2}\big\|_{1,1}}{1-\beta_1} - \bigg(f(\mathbf{z}_t) + \frac{G_\infty^2 \big\|\alpha_{t-1} \widehat{\mathbf{V}}_{t-1}^{-1/2}\big\|_{1,1}}{1-\beta_1}\bigg) \\
&\leq -\nabla f(\mathbf{x}_t)^\top \alpha_{t-1} \widehat{\mathbf{V}}_{t-1}^{-1/2} \mathbf{g}_t + 2L\big\|\alpha_t \widehat{\mathbf{V}}_t^{-1/2} \mathbf{g}_t\big\|_2^2 + 4L\bigg(\frac{\beta_1}{1-\beta_1}\bigg)^2 \|\mathbf{x}_t - \mathbf{x}_{t-1}\|_2^2 \\
&= -\nabla f(\mathbf{x}_t)^\top \alpha_{t-1} \widehat{\mathbf{V}}_{t-1}^{-1/2} \mathbf{g}_t + 2L\big\|\alpha_t \widehat{\mathbf{V}}_t^{-1/2} \mathbf{g}_t\big\|_2^2 + 4L\bigg(\frac{\beta_1}{1-\beta_1}\bigg)^2 \big\|\alpha_{t-1} \widehat{\mathbf{V}}_{t-1}^{-1/2} \mathbf{m}_{t-1}\big\|_2^2. \qquad (34)
\end{aligned}$$

Telescoping (34) for $t = 2$ to T and adding (33), we have

$$\begin{aligned}\sum_{t=2}^{T} \alpha_{t-1} \nabla f(\mathbf{x}_t)^\top \widehat{\mathbf{V}}_{t-1}^{-1/2} \mathbf{g}_t \leq f(\mathbf{z}_1) + \frac{G_\infty^2 \big\| \alpha_1 \widehat{\mathbf{V}}_1^{-1/2} \big\|_{1,1}}{1-\beta_1} + d\alpha_1 G_\infty - \left(f(\mathbf{z}_{T+1}) + \frac{G_\infty^2 \big\| \alpha_T \widehat{\mathbf{v}}_T^{-1/2} \big\|_1}{1-\beta_1} \right) \\ + 2L \sum_{t=1}^{T} \big\| \alpha_t \widehat{\mathbf{V}}_t^{-1/2} \mathbf{g}_t \big\|_2^2 + 4L \left(\frac{\beta_1}{1-\beta_1} \right)^2 \sum_{t=2}^{T} \big\| \alpha_{t-1} \widehat{\mathbf{V}}_{t-1}^{-1/2} \mathbf{m}_{t-1} \big\|_2^2 \\ \leq \Delta f + \frac{G_\infty^2 \alpha_1 \epsilon^{-1/2} d}{1-\beta_1} + d\alpha_1 G_\infty + 2L \sum_{t=1}^{T} \big\| \alpha_t \widehat{\mathbf{V}}_t^{-1/2} \mathbf{g}_t \big\|_2^2 \\ + 4L \left(\frac{\beta_1}{1-\beta_1} \right)^2 \sum_{t=1}^{T} \big\| \alpha_t \widehat{\mathbf{V}}_t^{-1/2} \mathbf{m}_t \big\|_2^2. \end{aligned} \tag{35}$$

By Lemma 17, we have

$$\sum_{t=1}^{T} \alpha_t^2 [\| \widehat{\mathbf{V}}_t^{-1/2} \mathbf{m}_t \|_2^2 \leq \frac{T^{1/2} \alpha_t^2 (1-\beta_1)}{2\epsilon^{1/2} (1-\beta_2)^{1/2} (1-\gamma)} \sum_{i=1}^{d} \| \mathbf{g}_{1:T,i} \|_2, \tag{36}$$

where $\gamma = \beta_1 / \beta_2^{1/2}$. We also have

$$\sum_{t=1}^{T} \alpha_t^2 \| \widehat{\mathbf{V}}_t^{-1/2} \mathbf{g}_t \|_2^2 \leq \frac{T^{1/2} \alpha_t^2}{2\epsilon^{1/2} (1-\beta_2)^{1/2} (1-\gamma)} \sum_{i=1}^{d} \| \mathbf{g}_{1:T,i} \|_2. \tag{37}$$

Moreover, consider the filtration $\mathcal{F}_t = \sigma(\xi_1, \ldots, \xi_t)$. Since $\mathbf{x}_t$ and $\widehat{\mathbf{V}}_{t-1}^{-1/2}$ only depend on $\xi_1, \ldots, \xi_{t-1}$. For any $\tau, \lambda > 0$, by Assumption 10 with $\mathbf{v} = \lambda \cdot \alpha_{t-1} \widehat{\mathbf{V}}_{t-1}^{-1/2} \nabla f(\mathbf{x}_t)$, we have

$$\mathbb{E} \Big\{ \exp \left[\lambda \alpha_{t-1} \nabla f(\mathbf{x}_t)^\top \widehat{\mathbf{V}}_{t-1}^{-1/2} (\mathbf{g}_t - \nabla f(\mathbf{x}_t)) \right] \Big| \mathcal{F}_{t-1} \Big\} \leq \exp(\sigma^2 \alpha_{t-1}^2 \lambda^2 \| \widehat{\mathbf{V}}_{t-1}^{-1/2} \nabla f(\mathbf{x}_t) \|_2^2 / 2).$$

Denote $Z_t = \alpha_{t-1} \nabla f(\mathbf{x}_t)^\top \widehat{\mathbf{V}}_{t-1}^{-1/2} (\mathbf{g}_t - \nabla f(\mathbf{x}_t))$. Then we have

$$\begin{aligned}\mathbb{P}(Z_t \geq \tau | \mathcal{F}_{t-1}) &= \mathbb{P}[\exp(\lambda Z_t) \geq \exp(\lambda \tau) | \mathcal{F}_{t-1}] \\ &= \mathbb{E}[\mathbb{1}\{\exp(-\lambda\tau + \lambda Z_t) \geq 1\} | \mathcal{F}_{t-1}] \\ &\leq \exp(-\lambda\tau) \cdot \mathbb{E}[\exp(\lambda Z_t) | \mathcal{F}_{t-1}] \\ &\leq \exp(-\lambda\tau) \cdot \exp(\sigma^2 \alpha_{t-1}^2 \lambda^2 \| \widehat{\mathbf{V}}_{t-1}^{-1/2} \nabla f(\mathbf{x}_t) \|_2^2 / 2) \\ &= \exp(-\lambda\tau + \sigma^2 \alpha_{t-1}^2 \lambda^2 \| \widehat{\mathbf{V}}_{t-1}^{-1/2} \nabla f(\mathbf{x}_t) \|_2^2 / 2). \end{aligned}$$

With exactly the same proof, we also have

$$\mathbb{P}(Z_t \leq -\tau | \mathcal{F}_{t-1}) \leq \exp(-\lambda\tau + \sigma^2 \alpha_{t-1}^2 \lambda^2 \| \widehat{\mathbf{V}}_{t-1}^{-1/2} \nabla f(\mathbf{x}_t) \|_2^2 / 2),$$

and therefore

$$\mathbb{P}(|Z_t| \geq \tau | \mathcal{F}_{t-1}) \leq 2 \exp(-\lambda\tau + \sigma^2 \alpha_{t-1}^2 \lambda^2 \| \widehat{\mathbf{V}}_{t-1}^{-1/2} \nabla f(\mathbf{x}_t) \|_2^2 / 2).$$

Choosing $\lambda = [\sigma^2\alpha_{t-1}^2\|\widehat{\mathbf{V}}_{t-1}^{-1/2}\nabla f(\mathbf{x}_t)\|_2^2]^{-1}\tau$, we finally obtain

$$\mathbb{P}(|Z_t| \geq \tau|\mathcal{F}_{t-1}) \leq 2\exp(-\tau^2/(2\sigma_t^2)) \tag{38}$$

for all $\tau > 0$, where $\sigma_t = \sigma\alpha_{t-1}\|\widehat{\mathbf{V}}_{t-1}^{-1/2}\nabla f(\mathbf{x}_t)\|_2$. The tail bound (38) enables the application of Lemma 6 in Jin et al. [14], which gives that with probability at least $1-\delta$,

$$\left|\sum_{t=2}^{T} Z_t\right| \leq (\epsilon\sigma^{-2}G_\infty^{-1}) \cdot \sum_{t=2}^{T}\sigma_t^2 + C(\epsilon\sigma^{-2}G_\infty^{-1})^{-1}\cdot\log(2/\epsilon),$$

where C is an absolute constant. Plugging in the definitions of Z_t and σ_t, we obtain

$$\begin{aligned}
&\left|\sum_{t=2}^{T}\alpha_{t-1}\nabla f(\mathbf{x}_t)^\top\widehat{\mathbf{V}}_{t-1}^{-1/2}\mathbf{g}_t - \sum_{t=2}^{T}\alpha_{t-1}\nabla f(\mathbf{x}_t)^\top\widehat{\mathbf{V}}_{t-1}^{-1/2}\nabla f(\mathbf{x}_t)\right| \\
&\qquad\leq (\epsilon\sigma^{-2}G_\infty^{-1})\cdot\sum_{t=2}^{T}\sigma^2\alpha_{t-1}^2\|\widehat{\mathbf{V}}_{t-1}^{-1/2}\nabla f(\mathbf{x}_t)\|_2^2 + C(\epsilon\sigma^{-2}G_\infty^{-1})^{-1}\log(2/\delta) \\
&\qquad\leq G_\infty^{-1}\sum_{t=2}^{T}\alpha_{t-1}^2\|\nabla f(\mathbf{x}_t)\|_2^2 + C\epsilon^{-1}\sigma^2 G_\infty\log(2/\delta),
\end{aligned} \tag{39}$$

where the second inequality is by the fact that the diagonal entries of $\widehat{\mathbf{V}}_{t-1}$ are all loewr bounded by ϵ. Substituting (36), (37) and (39) into (35), we have

$$\begin{aligned}
&\sum_{t=2}^{T}\alpha_{t-1}\nabla f(\mathbf{x}_t)^\top\widehat{\mathbf{V}}_{t-1}^{-1/2}\nabla f(\mathbf{x}_t) \\
&\qquad\leq \Delta f + \frac{G_\infty^2\alpha_1\epsilon^{-1/2}d}{1-\beta_1} + d\alpha_1 G_\infty + \frac{LT^{1/2}\alpha_t^2}{\epsilon^{1/2}(1-\beta_2)^{1/2}(1-\gamma)}\sum_{i=1}^{d}\|\mathbf{g}_{1:T,i}\|_2 \\
&\qquad\quad + \left(\frac{\beta_1}{1-\beta_1}\right)^2\frac{2LT^{1/2}\alpha_t^2(1-\beta_1)}{\epsilon^{1/2}(1-\beta_2)^{1/2}(1-\gamma)}\sum_{i=1}^{d}\|\mathbf{g}_{1:T,i}\|_2 + G_\infty^{-1}\sum_{t=2}^{T}\alpha_{t-1}^2\|\nabla f(\mathbf{x}_t)\|_2^2 \\
&\qquad\quad + C\epsilon^{-1}\sigma^2 G_\infty\log(2/\delta).
\end{aligned}$$

Moreover, by Lemma 16, we have $\nabla f(\mathbf{x}_t)^\top\widehat{\mathbf{V}}_{t-1}^{-1/2}\nabla f(\mathbf{x}_t) \geq G_\infty^{-1}\|\nabla f(\mathbf{x}_t)\|_2^2$, and therefore by choosing $\alpha_t = \alpha$ and rearranging terms, we have

$$\begin{aligned}
G_\infty^{-1}\sum_{t=2}^{T}\alpha(1-\alpha)\|\nabla f(\mathbf{x}_t)\|_2^2 &\leq \Delta f + \frac{G_\infty^2\alpha\epsilon^{-1/2}d}{1-\beta_1} + d\alpha G_\infty + \frac{LT^{1/2}\alpha^2}{\epsilon^{1/2}(1-\beta_2)^{1/2}(1-\gamma)}\sum_{i=1}^{d}\|\mathbf{g}_{1:T,i}\|_2 \\
&\quad + \left(\frac{\beta_1}{1-\beta_1}\right)^2\frac{2LT^{1/2}\alpha^2(1-\beta_1)}{\epsilon^{1/2}(1-\beta_2)^{1/2}(1-\gamma)}\sum_{i=1}^{d}\|\mathbf{g}_{1:T,i}\|_2 \\
&\quad + C\epsilon^{-1}\sigma^2 G_\infty\log(2/\delta).
\end{aligned}$$

Therefore when $\alpha < 1/2$, we have

$$\begin{aligned}\frac{1}{T-1}\sum_{t=2}^{T}\|\nabla f(\mathbf{x}_t)\|_2^2 \le \frac{4G_\infty}{T\alpha}\cdot\Delta f + \frac{4G_\infty^3\epsilon^{-1/2}}{1-\beta_1}\cdot\frac{d}{T} + 4G_\infty^2\cdot\frac{d}{T} \\ + \frac{4G_\infty L\alpha}{\epsilon^{1/2}(1-\beta_2)^{1/2}(1-\gamma)T^{1/2}}\sum_{i=1}^{d}\|\mathbf{g}_{1:T,i}\|_2 \\ + \left(\frac{\beta_1}{1-\beta_1}\right)^2\frac{8G_\infty L\alpha(1-\beta_1)}{\epsilon^{1/2}(1-\beta_2)^{1/2}(1-\gamma)T^{1/2}}\sum_{i=1}^{d}\|\mathbf{g}_{1:T,i}\|_2 \\ + \frac{C'G_\infty\epsilon^{-1}\sigma^2 G_\infty\log(2/\delta)}{T\alpha},\end{aligned}$$

where C' is an absolute constant.

Now by the theorem condition $\|\mathbf{g}_{1:T,i}\|_2 \le G_\infty T^s$, we have

$$\frac{1}{T-1}\sum_{t=2}^{T}\|\nabla f(\mathbf{x}_t)\|_2^2 \le \frac{M_1}{T\alpha} + \frac{M_2 d}{T} + \frac{\alpha M_3 d}{T^{1/2-s}},$$

where

$$\begin{aligned}M_1 &= 4G_\infty\Delta f + C'G_\infty\epsilon^{-1}\sigma^2 G_\infty\log(2/\delta),\\ M_2 &= \frac{4G_\infty^3\epsilon^{-1/2}}{1-\beta_1} + 4G_\infty^2,\\ M_3 &= \frac{4LG_\infty^2}{\epsilon^{1/2}(1-\beta_2)^{1/2}(1-\beta_1/\beta_2^{1/2})}\left(1+\frac{2\beta_1^2}{1-\beta_1}\right)\end{aligned}$$

where $\{M_i\}_{i=1}^3$ are defined in Theorem 12. This completes the proof. ■

D.5. Proof of Corollary 14

Proof [Proof of Corollary 14] Following the proof for Theorem 12, setting $\beta_1' = \beta_1 = 0$ and $\beta_2' = \beta_2 = \beta$ in Lemma 17 we get the conclusion. ■

D.6. Proof of Corollary 15

Proof [Proof of Corollary 15] Following the proof for Theorem 12, setting $\beta_1' = \beta_1 = 0$, $\beta_2 = 1$ and $\beta_2' = 0$ in Lemma 17 we get the conclusion. ■

Appendix E. Proof of Technical Lemmas

E.1. Proof of Lemma 16

Proof [Proof of Lemma 16] Since f has G_∞-bounded stochastic gradient, for any $\mathbf{x}$ and ξ, $\|\nabla f(\mathbf{x};\xi)\|_\infty \leq G_\infty$. Thus, we have

$$\|\nabla f(\mathbf{x})\|_\infty = \|\mathbb{E}_\xi \nabla f(\mathbf{x};\xi)\|_\infty \leq \mathbb{E}_\xi \|\nabla f(\mathbf{x};\xi)\|_\infty \leq G_\infty.$$

Next we bound $\|\mathbf{m}_t\|_\infty$. We have $\|\mathbf{m}_0\|_\infty = 0 \leq G_\infty$. Suppose that $\|\mathbf{m}_t\|_\infty \leq G_\infty$, then for $\mathbf{m}_{t+1}$, we have

$$\begin{aligned}\|\mathbf{m}_{t+1}\|_\infty &= \|\beta_1 \mathbf{m}_t + (1-\beta_1)\mathbf{g}_{t+1}\|_\infty \\ &\leq \beta_1 \|\mathbf{m}_t\|_\infty + (1-\beta_1)\|\mathbf{g}_{t+1}\|_\infty \\ &\leq \beta_1 G_\infty + (1-\beta_1) G_\infty \\ &= G_\infty.\end{aligned}$$

Thus, for any $t \geq 0$, we have $\|\mathbf{m}_t\|_\infty \leq G_\infty$. Finally we bound $\|\widehat{\mathbf{v}}_t\|_\infty$. First we have $\|\mathbf{v}_0\|_\infty = \|\widehat{\mathbf{v}}_0\|_\infty = 0 \leq G_\infty^2$. Suppose that $\|\widehat{\mathbf{v}}_t\|_\infty \leq G_\infty^2$ and $\|\mathbf{v}_t\|_\infty \leq G_\infty^2$. Note that we have

$$\begin{aligned}\|\mathbf{v}_{t+1}\|_\infty &= \|\beta_2 \mathbf{v}_t + (1-\beta_2)\mathbf{g}_{t+1}^2\|_\infty \\ &\leq \beta_2 \|\mathbf{v}_t\|_\infty + (1-\beta_2)\|\mathbf{g}_{t+1}^2\|_\infty \\ &\leq \beta_2 G_\infty^2 + (1-\beta_2) G_\infty^2 \\ &= G_\infty^2,\end{aligned}$$

and by definition, we have $\|\widehat{\mathbf{v}}_{t+1}\|_\infty = \max\{\|\widehat{\mathbf{v}}_t\|_\infty, \|\mathbf{v}_{t+1}\|_\infty\} \leq G_\infty^2$. Thus, for any $t \geq 0$, we have $\|\widehat{\mathbf{v}}_t\|_\infty \leq G_\infty^2$. ■

E.2. Proof of Lemma 17

Proof Recall that $\widehat{v}_{t,j}, m_{t,j}, g_{t,j}$ denote the j-th coordinate of $\widehat{\mathbf{v}}_t, \mathbf{m}_t$ and $\mathbf{g}_t$. We have

$$\begin{aligned}\alpha_t^2 \|\widehat{\mathbf{V}}_t^{-1/2}\mathbf{m}_t\|_2^2 &= \alpha_t^2 \sum_{i=1}^d \frac{m_{t,i}^2}{\widehat{v}_{t,i}^{1/2}} \cdot \frac{\widehat{v}_{t,i}^{1/2}}{\widehat{v}_{t,i} + \epsilon} \\ &\leq \alpha_t^2 \sum_{i=1}^d \frac{m_{t,i}^2}{\widehat{v}_{t,i}^{1/2}} \cdot \frac{\widehat{v}_{t,i}^{1/2}}{2\widehat{v}_{t,i}^{1/2}\epsilon^{1/2}} \\ &\leq \frac{\alpha_t^2}{2\epsilon^{1/2}} \sum_{i=1}^d \frac{m_{t,i}^2}{v_{t,i}^{1/2}} \\ &= \frac{\alpha_t^2}{2\epsilon^{1/2}} \sum_{i=1}^d \frac{(\sum_{j=1}^t (1-\beta_1')\beta_1^{t-j} g_{j,i})^2}{(\sum_{j=1}^t (1-\beta_2')\beta_2^{t-j} g_{j,i}^2)^{1/2}},\end{aligned} \tag{40}$$

where the first inequality holds since $a + b \geq 2\sqrt{ab}$ and the second inequality holds because $\widehat{v}_{t,i} \geq v_{t,i}$. Next we have

$$\frac{\alpha_t^2}{2\epsilon^{1/2}} \sum_{i=1}^{d} \frac{(\sum_{j=1}^{t}(1-\beta_1')\beta_1^{t-j} g_{j,i})^2}{(\sum_{j=1}^{t}(1-\beta_2')\beta_2^{t-j} g_{j,i}^2)^{1/2}} \leq \frac{\alpha_t^2(1-\beta_1')^2}{2\epsilon^{1/2}(1-\beta_2')^{1/2}} \sum_{i=1}^{d} \frac{(\sum_{j=1}^{t}\beta_1^{t-j})(\sum_{j=1}^{t}\beta_1^{t-j}|g_{j,i}|^2)}{(\sum_{j=1}^{t}\beta_2^{t-j} g_{j,i}^2)^{1/2}}$$
$$\leq \frac{\alpha_t^2(1-\beta_1')}{2\epsilon^{1/2}(1-\beta_2')^{1/2}} \sum_{i=1}^{d} \frac{\sum_{j=1}^{t}\beta_1^{t-j}|g_{j,i}|^2}{(\sum_{j=1}^{t}\beta_2^{t-j} g_{j,i}^2)^{1/2}}, \tag{41}$$

where the first inequality holds due to Cauchy inequality, and the last inequality holds because $\sum_{j=1}^{t}\beta_1^{t-j} \leq (1-\beta_1)^{-1}$. Note that

$$\sum_{i=1}^{d} \frac{\sum_{j=1}^{t}\beta_1^{t-j}|g_{j,i}|^2}{(\sum_{j=1}^{t}\beta_2^{t-j} g_{j,i}^2)^{1/2}} \leq \sum_{i=1}^{d}\sum_{j=1}^{t} \frac{\beta_1^{t-j}|g_{j,i}|^2}{(\beta_2^{t-j} g_{j,i}^2)^{1/2}} = \sum_{i=1}^{d}\sum_{j=1}^{t} \gamma^{t-j}|g_{j,i}|, \tag{42}$$

where the equality holds due to the definition of γ. Substituting (41) and (42) into (40), we have

$$\alpha_t^2 \|\widehat{\mathbf{V}}_t^{-1/2}\mathbf{m}_t\|_2^2 \leq \frac{\alpha_t^2(1-\beta_1')}{2\epsilon^{1/2}(1-\beta_2')^{1/2}} \sum_{i=1}^{d}\sum_{j=1}^{t} \gamma^{t-j}|g_{j,i}|. \tag{43}$$

Telescoping (43) for $t = 1$ to T, we have

$$\sum_{t=1}^{T} \alpha_t^2 \|\widehat{\mathbf{V}}_t^{-1/2}\mathbf{m}_t\|_2^2 \leq \frac{\alpha_t^2(1-\beta_1')}{2\epsilon^{1/2}(1-\beta_2')^{1/2}} \sum_{t=1}^{T}\sum_{i=1}^{d}\sum_{j=1}^{t} \gamma^{t-j}|g_{j,i}|$$
$$= \frac{\alpha_t^2(1-\beta_1')}{2\epsilon^{1/2}(1-\beta_2')^{1/2}} \sum_{i=1}^{d}\sum_{j=1}^{T} |g_{j,i}| \sum_{t=j}^{T} \gamma^{t-j}$$
$$\leq \frac{\alpha_t^2(1-\beta_1')}{2\epsilon^{1/2}(1-\beta_2')^{1/2}(1-\gamma)} \sum_{i=1}^{d}\sum_{j=1}^{T} |g_{j,i}|. \tag{44}$$

Finally, we have

$$\sum_{i=1}^{d}\sum_{j=1}^{T} |g_{j,i}| \leq \sum_{i=1}^{d} \Big(\sum_{j=1}^{T} g_{j,i}^2\Big)^{1/2} \cdot T^{1/2} = T^{1/2} \sum_{i=1}^{d} \|\mathbf{g}_{1:T,i}\|_2, \tag{45}$$

where the inequality holds due to Hölder's inequality. Substituting (45) into (44), we have

$$\sum_{t=1}^{T} \alpha_t^2 \|\widehat{\mathbf{V}}_t^{-1/2}\mathbf{m}_t\|_2^2 \leq \frac{T^{1/2}\alpha_t^2(1-\beta_1')}{2\epsilon^{1/2}(1-\beta_2')^{1/2}(1-\gamma)} \sum_{i=1}^{d} \|\mathbf{g}_{1:T,i}\|_2.$$

Specifically, taking $\beta_1 = 0$, we have $\mathbf{m}_t = \mathbf{g}_t$, then

$$\sum_{t=1}^{T} \alpha_t^2 \|\widehat{\mathbf{V}}_t^{-1/2}\mathbf{g}_t\|_2^2 \leq \frac{T^{1/2}\alpha_t^2}{2\epsilon^{1/2}(1-\beta_2')^{1/2}(1-\gamma)} \sum_{i=1}^{d} \|\mathbf{g}_{1:T,i}\|_2.$$

■

E.3. Proof of Lemma 18

Proof By definition, we have

$$\mathbf{z}_{t+1} = \mathbf{x}_{t+1} + \frac{\beta_1}{1-\beta_1}(\mathbf{x}_{t+1} - \mathbf{x}_t) = \frac{1}{1-\beta_1}\mathbf{x}_{t+1} - \frac{\beta_1}{1-\beta_1}\mathbf{x}_t.$$

Then we have

$$\begin{aligned}\mathbf{z}_{t+1} - \mathbf{z}_t &= \frac{1}{1-\beta_1}(\mathbf{x}_{t+1} - \mathbf{x}_t) - \frac{\beta_1}{1-\beta_1}(\mathbf{x}_t - \mathbf{x}_{t-1}) \\ &= \frac{1}{1-\beta_1}\big(-\alpha_t \widehat{\mathbf{V}}_t^{-1/2}\mathbf{m}_t\big) + \frac{\beta_1}{1-\beta_1}\alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2}\mathbf{m}_{t-1}.\end{aligned}$$

The equities above are based on definition. Then we have

$$\begin{aligned}\mathbf{z}_{t+1} - \mathbf{z}_t &= \frac{-\alpha_t \widehat{\mathbf{V}}_t^{-1/2}}{1-\beta_1}\Big[\beta_1\mathbf{m}_{t-1} + (1-\beta_1)\mathbf{g}_t\Big] + \frac{\beta_1}{1-\beta_1}\alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2}\mathbf{m}_{t-1} \\ &= \frac{\beta_1}{1-\beta_1}\mathbf{m}_{t-1}\big(\alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2} - \alpha_t\widehat{\mathbf{V}}_t^{-1/2}\big) - \alpha_t\widehat{\mathbf{V}}_t^{-1/2}\mathbf{g}_t \\ &= \frac{\beta_1}{1-\beta_1}\alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2}\mathbf{m}_{t-1}\Big[\mathbf{I} - \big(\alpha_t\widehat{\mathbf{V}}_t^{-1/2}\big)\big(\alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2}\big)^{-1}\Big] - \alpha_t\widehat{\mathbf{V}}_t^{-1/2}\mathbf{g}_t \\ &= \frac{\beta_1}{1-\beta_1}\Big[\mathbf{I} - \big(\alpha_t\widehat{\mathbf{V}}_t^{-1/2}\big)\big(\alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2}\big)^{-1}\Big](\mathbf{x}_{t-1} - \mathbf{x}_t) - \alpha_t\widehat{\mathbf{V}}_t^{-1/2}\mathbf{g}_t.\end{aligned}$$

The equalities above follow by combining the like terms. ■

E.4. Proof of Lemma 19

Proof By Lemma 18, we have

$$\begin{aligned}\|\mathbf{z}_{t+1} - \mathbf{z}_t\|_2 &= \left\|\frac{\beta_1}{1-\beta_1}\Big[\mathbf{I} - (\alpha_t\widehat{\mathbf{V}}_t^{-1/2})(\alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2})^{-1}\Big](\mathbf{x}_{t-1} - \mathbf{x}_t) - \alpha_t\widehat{\mathbf{V}}_t^{-1/2}\mathbf{g}_t\right\|_2 \\ &\le \frac{\beta_1}{1-\beta_1}\left\|\mathbf{I} - (\alpha_t\widehat{\mathbf{V}}_t^{-1/2})(\alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2})^{-1}\right\|_{\infty,\infty}\cdot\|\mathbf{x}_{t-1} - \mathbf{x}_t\|_2 + \left\|\alpha\widehat{\mathbf{V}}_t^{-1/2}\mathbf{g}_t\right\|_2,\end{aligned}$$

where the inequality holds because the term $\beta_1/(1-\beta_1)$ is positive, and triangle inequality. Considering that $\alpha_t\widehat{\mathbf{v}}_{t,j}^{-1/2} \le \alpha_{t-1}\widehat{\mathbf{v}}_{t-1,j}^{-1/2}$, when $p > 0$, we have $\left\|\mathbf{I} - (\alpha_t\widehat{\mathbf{V}}_t^{-1/2})(\alpha_{t-1}\widehat{\mathbf{V}}_{t-1}^{-1/2})^{-1}\right\|_{\infty,\infty} \le 1$. With that fact, the term above can be bound as:

$$\|\mathbf{z}_{t+1} - \mathbf{z}_t\|_2 \le \left\|\alpha\widehat{\mathbf{V}}_t^{-1/2}\mathbf{g}_t\right\|_2 + \frac{\beta_1}{1-\beta_1}\|\mathbf{x}_{t-1} - \mathbf{x}_t\|_2.$$

This completes the proof. ■

E.5. Proof of Lemma 20

Proof For term $\|\nabla f(\mathbf{z}_t) - \nabla f(\mathbf{x}_t)\|_2$, we have:

$$\begin{aligned}\|\nabla f(\mathbf{z}_t) - \nabla f(\mathbf{x}_t)\|_2 &\leq L\|\mathbf{z}_t - \mathbf{x}_t\|_2 \\ &\leq L\Big\|\frac{\beta_1}{1-\beta_1}(\mathbf{x}_t - \mathbf{x}_{t-1})\Big\|_2 \\ &\leq L\Big(\frac{\beta_1}{1-\beta_1}\Big)\cdot\|\mathbf{x}_t - \mathbf{x}_{t-1}\|_2,\end{aligned}$$

where the last inequality holds because the term $\beta_1/(1-\beta_1)$ is positive. ■