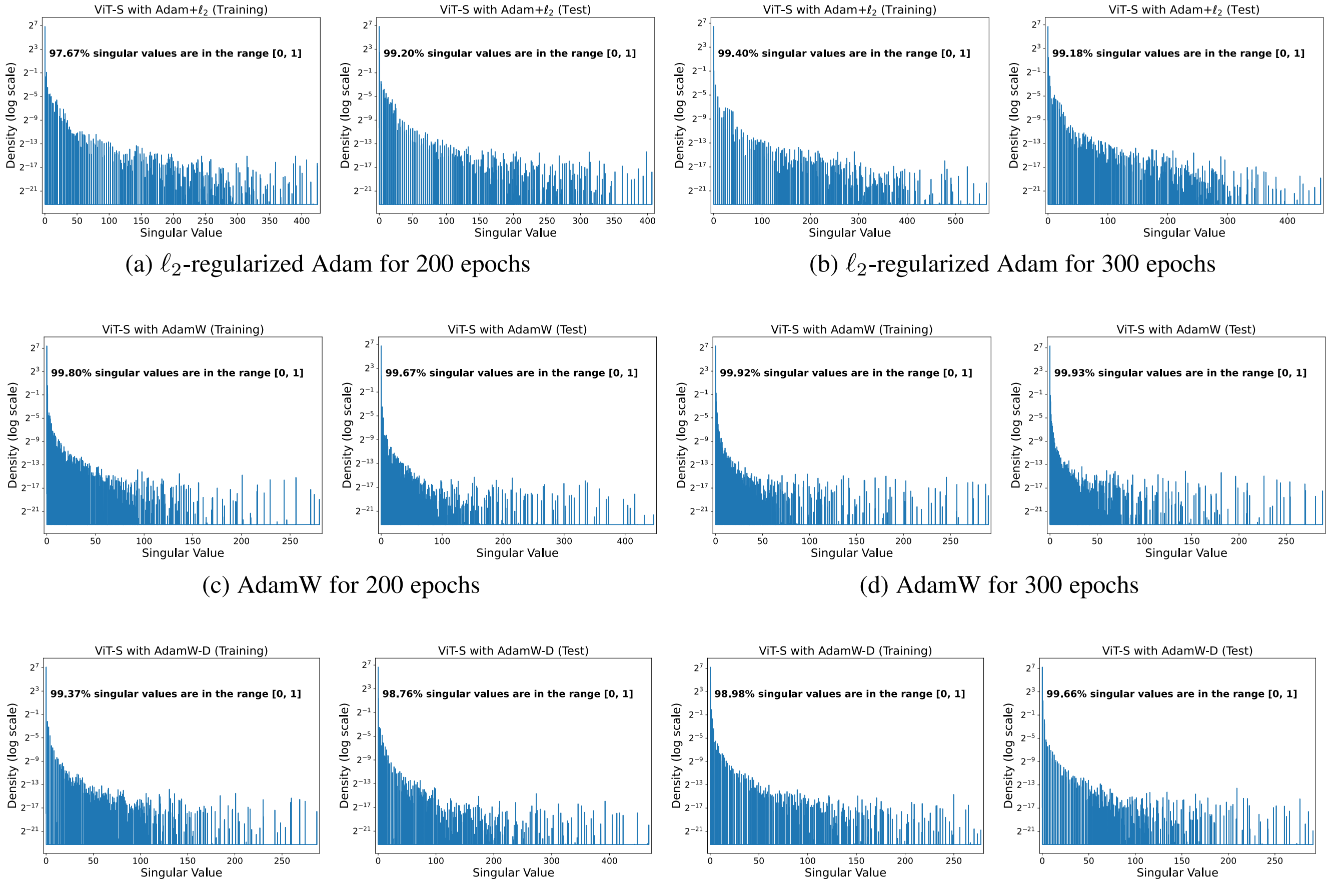
Towards Understanding Convergence and Generalization of AdamW (Supplementary Material)

Pan Zhou, Xingyu Xie, Zhouchen Lin, *Fellow, IEEE,* Shuicheng Yan, *Fellow, IEEE*

This supplementary document contains the technical proofs of convergence results and some additional experimental results of the paper entitled "Towards Understanding Convergence and Generalization of AdamW". It is structured as follows. Appendix A presents more experimental results. In Appendix B, we first give the detailed algorithmic frameworks of AdamW and its stagewise variant in Algorithms 1 and 2. Then Appendix F intuitively discusses the generalization benefits of coordinate-adaptive regularization in AdamW. Next, Appendix G introduces the main proof technique differences between this work and other works. Appendix E provides the theoretical justification for the approximation $\boldsymbol{n}_t' \approx \boldsymbol{F}_{\boldsymbol{x}_t} \approx \boldsymbol{H}_{\boldsymbol{x}_t}$ in Assumption 4. Appendix F provides some auxiliary lemmas throughout this document. Then Appendix G presents the proof of the convergence results in Sec. 4, *i.e.*, the proof of Theorems 2 $\sim$ 4. Next, in Appendix H, we introduce the proof of generalization results in Sec. 5, including Lemma 5 and Theorems 6 and 7. Finally, Appendix I provides the proofs of some auxiliary lemmas in Appendix F.



(a) ℓ_2-regularized Adam for 200 epochs

(b) ℓ_2-regularized Adam for 300 epochs

(c) AdamW for 200 epochs

(d) AdamW for 300 epochs

(e) AdamW-D for 200 epochs

(f) AdamW-D for 300 epochs

Fig. 3: Visualization of singular values in ViT-small trained by ℓ_2-regularized Adam, AdamW and AdamW-D for 200 and 300 epochs.

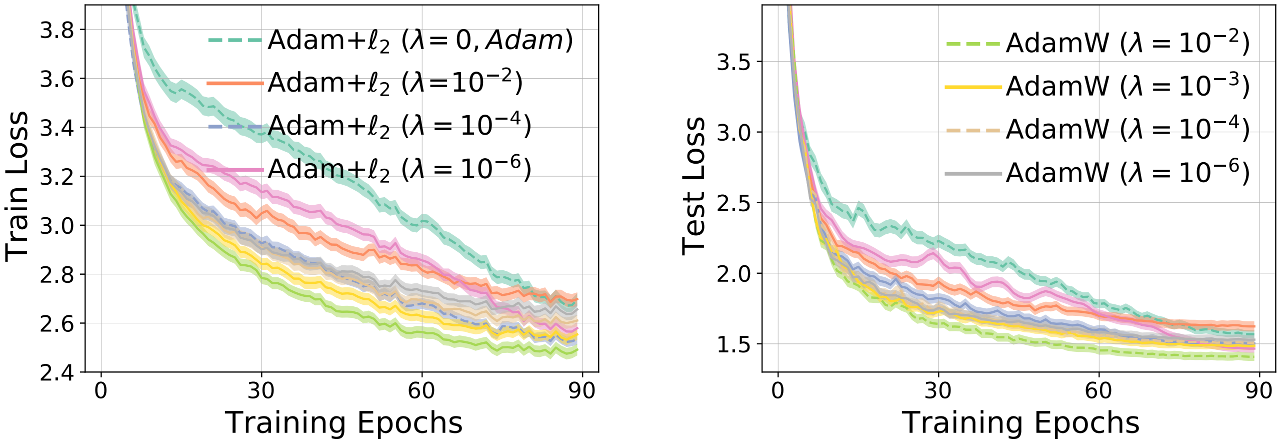


Fig. 4: Training and test curves comparison on ImageNet. We independently test AdamW on ResNet18 by using three different seeds, and plot the average and variance. Similarly, we evaluate ℓ_2-regularized Adam (ℓ_2-Adam) with three different seeds.

Algorithm 1: AdamW [1]

```
Input: initialization $x_0$, step size $\{\eta_k\}_{k=0}^T$, hyper-parameters $\{\beta_{1k}\}_{k=0}^T$ and $\{\beta_{2k}\}_{k=0}^T$ for first- and second-order moments $\{m_k\}_{k=0}^T$ and $\{n_k\}_{k=0}^T$.
Output: some average of $\{x_k\}_{k=0}^T$.
while $k < T$ do
    estimate stochastic gradient $g_k = \frac{1}{b}\sum_{i=1}^b \nabla f(x_k; \xi_i)$;
    estimate first-order moment $m_k = (1 - \beta_{1k}) m_k + \beta_{1k} g_k$;
    estimate second-order moment $n_k = (1 - \beta_{2k}) n_k + \beta_{2k} g_k^2$;
    update parameter $x_{k+1} = (1 - \lambda_k \eta_k) x_k - \eta_k m_k / \sqrt{n_k + \delta}$;
end while
```

Appendix A
More Experimental Results

Here we give more experimental investigation on singular values of Hessian in deep networks. In the manuscript, we provide investigation by training ResNet50 [2] and vision transformer small (ViT-small) [3] for both 100 epochs. Here we provide more visualization results of ResNet50 [2] and vision transformer small (ViT-small) [3] trained by 200 and 300 epochs. Similarly, we adopt the singular value estimation method in [4] to estimate the singular values of these two trained networks. Fig. 3 plots the spectral density of these singular values, and shows that there are more than 99% singular values that are in the range $[0, 1]$ and indeed are much smaller than one. All these results also accords with the observations on ResNet50 and ViT-small trained by 100 epochs. All these observations support the results in Sec. 5.2.

For multiple trials of the experiments, we independently test AdamW on ResNet18 by using three different seeds, and plot the average and variance in Fig. 4. Similarly, we evaluate ℓ_2-regularized Adam with three different seeds. From Fig. 4, one can observe that the performance of these algorithms are stable and consistent.

Algorithm 2: Stagewise AdamW

```
Input: initialization $x_0$, optimization accuracy $\{\epsilon_k\}_{k=1}^K$ .
Output: some average of $\{x_k\}_{k=0}^T$.
while $k < K$ do
    optimize the loss objective by AdamW (algorithm 1) to accuracy $\epsilon_k$, and output solution $x_k$;
end while
```

Appendix B
Details of AdamW and its Stagewise Variant

Due to space limitation, in the manuscript, we do not provide the detailed AdamW. Here we give algorithmic framework of AdamW in Algorithm 1 to help understand. Since in Sec. 4.4 we further propose the stagewise AdamW algorithm to solve PŁ-conditioned nonconvex problems, here we also provide the algorithmic framework of stagewise AdamW in Algorithm 2.

Appendix C
Generalization Benefits of Coordinate-Adaptive Regularization In AdamW

Now we intuitively discuss the generalization benefits of coordinate-adaptive regularization in AdamW. Due to the high nonconvexity, a deep network often contains many sharp minima and also flat ones, where sharp minimum often refers to the minimum around which loss landscape has sharp directions, *i.e.*, large gradient magnitude [5]. Assume current solution $\boldsymbol{x}_k$ is around a local sharp basin with a sharp minimum $\boldsymbol{x}_*$. Then the sharp directions indexed by $\mathcal{I}$ would have large gradients and thus large $\boldsymbol{v}_{k,i}$ $(i \in \mathcal{I})$. So for sharp directions $\mathcal{I}$, AdamW would have much stronger regularization and prevent $\boldsymbol{x}_k$ to fast approach $\boldsymbol{x}_*$; for flat directions $\mathcal{I}^c$, AdamW would still allow fast update due to small $\boldsymbol{v}_{k,i}$ $(i \in \mathcal{I}^c)$. This helps $\boldsymbol{x}_k$ escape from the local sharp basin in the subsequent training iterations, since a) the stochastic gradient brings perturbations and possibly brings $\boldsymbol{x}_k$ from the sharp basin as proved and also observed in many works, *e.g.* [6], [7]; b) $\boldsymbol{x}_k$ is at the bottleneck instead of the bottom of the basin due to the slow update on sharp directions $\mathcal{I}$ which largely increases the escaping probability. In contrast, for sharp directions $\mathcal{I}$, ℓ_2-regularized Adam would not penalize as stronger as AdamW, since it needs to trade-off the convergence speed and regularization: stronger regularization benefits the generalization due to its slow update on sharp directions $\mathcal{I}$, but impairs convergence speed on flat directions $\mathcal{I}^c$. Accordingly, the solution $\boldsymbol{x}_k$ in ℓ_2-regularized Adam could faster approach the bottom of the sharp basin which greatly increases the difficulty of escaping. Consider that flat minima are observed or proved to enjoy better generalization in many works, *e.g.*, the aforementioned three works, AdamW can better trade-off the generalization and convergence than ℓ_2-regularized Adam thanks to its coordinate-adaptive regularization.

Appendix D
Discussion on Our Proof Technique

For proof techniques, the most related work is [8]. Our convergence analysis and [8] share some similar overall proof roadmap. This is because we both analyze nonconvex problem under almost the same conditions which actually restricts the proof frameworks, *e.g.*, first using smoothness condition and bounded gradient to establish the relation of current loss $F_{k+1}(\boldsymbol{x}_{k+1})$ and previous loss$F_k(\boldsymbol{x}_k)$, and then accumulating this loss relation to bound the gradient (desired results). For this roadmap, most nonconvex optimization works, *e.g.*, [9]–[11], actually follow it to achieve their desire results but need to elaborate each proof pieces in the overall proof roadmap according to their algorithms.

Our convergence analysis also inherits the above overall proof roadmap, but is indeed more elaborated and simpler than the one in [8] which analyzes their proposed Adan instead of AdamW here. Specifically, both [8] and this work uses smoothness condition and bounded gradient to establish the relation between current loss $F_{k+1}(\boldsymbol{x}_{k+1})$ and previous loss $F_k(\boldsymbol{x}_k)$. Despite the algorithm differences, we apply the bounding technique in [8] to AdamW, and establish

$$F_{k+1}(\boldsymbol{x}_{k+1}) \leq F_k(\boldsymbol{x}_k) + \frac{\eta}{2c_1}\|\nabla F(\boldsymbol{x}_k) - \boldsymbol{m}_k\|_2^2 - \frac{\eta}{4c_2}\|\boldsymbol{u}_k\|_2^2, \tag{9}$$

where $\boldsymbol{u}_k = \boldsymbol{m}_k + \lambda_k \boldsymbol{x}_k \odot \boldsymbol{v}_k$, while we prove a tighter one by using different bounding strategy:

$$F_{k+1}(\boldsymbol{x}_{k+1}) \leq F_k(\boldsymbol{x}_k) + \frac{\eta}{2c_1}\|\nabla F(\boldsymbol{x}_k) - \boldsymbol{m}_k\|_2^2 - \frac{\eta}{2c_2}\|\nabla F(\boldsymbol{x}_k) + \lambda_k \boldsymbol{x}_k \odot \boldsymbol{v}_k\|_2^2 - \frac{\eta}{4c_2}\|\boldsymbol{u}_k\|_2^2. \tag{10}$$

By comparison, our Eqn. (10) is stronger than Eqn. (9) in [8] because of the term $(-\frac{\eta}{2c_2}\|\nabla F(\boldsymbol{x}_k) + \lambda_k \boldsymbol{x}_k \odot \boldsymbol{v}_k\|_2^2)$ which can help cancel many terms related to $\|\nabla F(\boldsymbol{x}_k) + \lambda_k \boldsymbol{x}_k \odot \boldsymbol{v}_k\|_2^2$ and greatly simplify the proof as discussed below. See the details and mathematical derivations of Eqn. (10) and Eqn. (9) in Appendix I.3.

Then Xie et al. accumulate their Eqn. (9) and also uses other techniques to sequentially upper bound

$$\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left[\|\boldsymbol{m}_k + \lambda_k \boldsymbol{x}_k \odot \boldsymbol{v}_k\|_2^2\right] \leq \epsilon^2, \qquad \frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left[\|\boldsymbol{m}_k - \nabla F(\boldsymbol{x}_k)\|_2^2\right] \leq \frac{1}{4}\epsilon^2$$

and then use them to prove the desired results

$$\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left[\|\nabla F_k(\boldsymbol{x}_k)\|_2^2\right] = \frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left[\|\nabla F(\boldsymbol{x}_k) + \lambda_k \boldsymbol{x}_k \odot \boldsymbol{v}_k\|_2^2\right] \leq \mathcal{O}\left(\epsilon^2\right).$$

In contrast, we can directly prove a stronger desired result in one step without need to prove the temporal bounds on $\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left[\|\boldsymbol{m}_k + \lambda_k \boldsymbol{x}_k \odot \boldsymbol{v}_k\|_2^2\right]$ and $\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left[\|\boldsymbol{m}_k - \nabla F(\boldsymbol{x}_k)\|_2^2\right]$:

$$\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left[\|\nabla F_k(\boldsymbol{x}_k)\|_2^2 + \frac{1}{4}\|\boldsymbol{u}_k\|_2^2\right] = \frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left[\|\nabla F(\boldsymbol{x}_k) + \lambda_k \boldsymbol{x}_k \odot \boldsymbol{v}_k\|_2^2\right] \leq \mathcal{O}\left(\epsilon^2\right).$$

As a result, our proof is much more straightforward and simpler. In our proof, we can directly prove the desired result $\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\left[\|\nabla F_k(\boldsymbol{x}_k)\|_2^2 + \frac{1}{4}\|\boldsymbol{u}_k\|_2^2\right]$, since a) our Eqn. (10) is tighter than Eqn. (9) in Adan which helps us cancel many terms related to $\|\boldsymbol{u}_k\|_2^2$, and b) in the proof, we always consider more elaborated and straightforward steps to prove the desired results. Moreover, we analyze the problem under the decayed learning rate and the PŁ-conditioned problem which is missing in [8].

Appendix E
Justification on The Approximation $\boldsymbol{n}'_t \approx \boldsymbol{F}_{\boldsymbol{x}_t} \approx \boldsymbol{H}_{\boldsymbol{x}_t}$

Staib et al. [12] proved that the moving average $\boldsymbol{n}'_t = (1-\beta_2)\boldsymbol{n}'_{t-1} + \beta_2 \boldsymbol{g}_t^\top \boldsymbol{g}_t$ is a very good estimation ot the Fisher information matrix $\boldsymbol{F}_{\boldsymbol{x}_t} = \frac{1}{n}\sum_{i=1}^n \nabla F(\boldsymbol{x}_t;\boldsymbol{\xi}_i)\nabla F(\boldsymbol{x}_t;\boldsymbol{\xi}_i)^\top$. More specifically, they proved

$$\Phi = \|\boldsymbol{n}'_t - \boldsymbol{F}_{\boldsymbol{x}_t}\| \leq \mathcal{O}\left(\eta L^{1/3}\right),$$

when the iteration number $T \geq \mathcal{O}\left(\eta_{-2/3}\right)$. Please refer to their Theorem 4.1, and Proposition 4.1 and 4.2. In our theories, e.g. Theorem 2, we use the learning rate $\eta = \mathcal{O}\left(\epsilon^2\right)$ which is very small. So the term $\mathcal{O}\left(\eta L^{1/3}\right)$ is indeed very small, and thus guarantees

$$\boldsymbol{n}'_t \approx \boldsymbol{F}_{\boldsymbol{x}_t}. \tag{11}$$

Then we follow the notation in [13] and [14] to show $\boldsymbol{F}_{\boldsymbol{x}_t}$ is a good estimation to Hessian $\boldsymbol{H}_{\boldsymbol{x}_t}$. For completeness, we quote the proof of [13] to here. Please find the same proof in the appendix of [13]. Assume each training sample $\boldsymbol{x}_i = (\boldsymbol{a}_i, \boldsymbol{b}_i)$ contains a sample $\boldsymbol{a}_i$ with a target $\boldsymbol{b}_i$. Let $F(\boldsymbol{x}_t;\boldsymbol{\xi}_i)$ is composed of a prediction function $\boldsymbol{c}_i = f(\boldsymbol{x}_t;\boldsymbol{a}_i)$ and a loss $\ell(\boldsymbol{b}_i;\boldsymbol{c}_i)$, namely, $F(\boldsymbol{x}_t;\boldsymbol{\xi}_i) = \ell(\boldsymbol{b}_i; f(\boldsymbol{x}_t;\boldsymbol{a}_i))$, where $\boldsymbol{c}_i = f(\boldsymbol{x}_t;\boldsymbol{a}_i)$ maps the neural network's input $\boldsymbol{a}_i$ to the output $\boldsymbol{c}_i$, and $\ell(\boldsymbol{b}_i;\boldsymbol{c}_i)$ measures the difference between $\boldsymbol{c}_i$ and $\boldsymbol{b}_i$. Let $P_{\boldsymbol{a},\boldsymbol{b}}(\boldsymbol{x})$ be the model distribution, and let $R_{\boldsymbol{b}|\boldsymbol{c}}$ be the predictive distribution used at the network output so that $R_{\boldsymbol{b}|\boldsymbol{c}} = P_{\boldsymbol{b}|f(\boldsymbol{x};\boldsymbol{a})}$. Next, let $P_{\boldsymbol{x}}(\boldsymbol{b}|\boldsymbol{a})$ be the associated probability density. Since many probabilistic models can be formulated as

$$\ell(\boldsymbol{b}_i; f(\boldsymbol{x}_t;\boldsymbol{a}_i)) = -\log P_{\boldsymbol{x}}(\boldsymbol{b}|\boldsymbol{a}),$$

we can formulate

$$\boldsymbol{F}_{\boldsymbol{x}_t} = \frac{1}{n}\sum_{i=1}^n \nabla F(\boldsymbol{x}_t;\boldsymbol{\xi}_i)\nabla F(\boldsymbol{x}_t;\boldsymbol{\xi}_i)^\top = \frac{1}{n}\sum_{i=1}^n \frac{\partial \log P_{\boldsymbol{x}}(\boldsymbol{b}|\boldsymbol{a})}{\partial \boldsymbol{x}}\frac{\partial \log P_{\boldsymbol{x}}(\boldsymbol{b}|\boldsymbol{a})}{\partial \boldsymbol{x}^\top},$$

For Hessian of this model, we can write it as follows:

$$\boldsymbol{H}_{\boldsymbol{x}_t} = \frac{1}{n}\sum_{i=1}^n \frac{\partial \log P_{\boldsymbol{x}}(\boldsymbol{b}|\boldsymbol{a})}{\partial \boldsymbol{x}}\frac{\partial \log P_{\boldsymbol{x}}(\boldsymbol{b}|\boldsymbol{a})}{\partial \boldsymbol{x}^\top} - \frac{1}{P_{\boldsymbol{x}}(\boldsymbol{b}|\boldsymbol{a})}\frac{\partial^2 \log P_{\boldsymbol{x}}(\boldsymbol{b}|\boldsymbol{a})}{\partial \boldsymbol{x}\partial \boldsymbol{x}^\top}.$$

One can observe that $\boldsymbol{H}_{\boldsymbol{x}_t}$ has an extra term $-\frac{1}{P_{\boldsymbol{x}}(\boldsymbol{b}|\boldsymbol{a})}\frac{\partial^2 \log P_{\boldsymbol{x}}(\boldsymbol{b}|\boldsymbol{a})}{\partial \boldsymbol{x}\partial \boldsymbol{x}^\top}$. This extra term can be negligible in the case where the model is realizable, namely the model's conditional distribution coincides with the training data's conditional distribution. Mathematically, when the parameter is close to an optimum, $P_{\boldsymbol{x}}(\boldsymbol{b}|\boldsymbol{a})$ is very close to $P(\boldsymbol{b}|\boldsymbol{a})$. Under this condition, the model has realized the data distribution and the extra term is a sample estimator of the following zero quantity:

$$\begin{aligned}\mathbb{E}_{(\boldsymbol{a},\boldsymbol{b})\sim P(\boldsymbol{b}|\boldsymbol{a})}\left[\frac{1}{P_{\boldsymbol{x}}(\boldsymbol{b}|\boldsymbol{a})}\frac{\partial^2 \log P_{\boldsymbol{x}}(\boldsymbol{b}|\boldsymbol{a})}{\partial \boldsymbol{x}\partial \boldsymbol{x}^\top}\right] &= \int d\boldsymbol{a}d\boldsymbol{b}P(\boldsymbol{a})\frac{\partial^2 \log P_{\boldsymbol{x}}(\boldsymbol{b}|\boldsymbol{a})}{\partial \boldsymbol{x}\partial \boldsymbol{x}^\top}\\ &= \int d\boldsymbol{a}P(\boldsymbol{a})\frac{\partial^2}{\partial \boldsymbol{x}\partial \boldsymbol{x}^\top}\left[\int d\boldsymbol{b}\log P_{\boldsymbol{x}}(\boldsymbol{b}|\boldsymbol{a})\right]\\ &= \int d\boldsymbol{a}P(\boldsymbol{a})\frac{\partial^2}{\partial \boldsymbol{x}\partial \boldsymbol{x}^\top}[1] = 0,\end{aligned}$$

with the estimator becoming more accurate with larger sample number n. Thus, when the parameter is close to an optimum, we have $\boldsymbol{F}_{\boldsymbol{x}_t} \approx \boldsymbol{H}_{\boldsymbol{x}_t}$.

Finally, combing the result in Eqn. (11), we have

$$\boldsymbol{n}'_t \approx \boldsymbol{F}_{\boldsymbol{x}_t} \approx \boldsymbol{H}_{\boldsymbol{x}_t}.$$

when the model parameter $\boldsymbol{x}_t$ is close to an optimum. It should be mentioned that some works on generalization analysis also directly use $\boldsymbol{F}_{\boldsymbol{x}_t} \approx \boldsymbol{H}_{\boldsymbol{x}_t}$, such as [13] (see its Assumption 2), and [15] (see its Eqn. (5)). Moreover, to approximate the loss function by a quadratic loss to simply the analysis challenges while providing theory insights, most works analyze the generalization performance of an algorithm around a local minimum, such as the references [16] (see its Assumption 4), [17] (see its discussion below Eq. (11)), [7] (see its section 4), [18] (see its Assumption 4), [19] (see its discussion above Eqn. (7)), [20] (see its Theorem 4.4) in the manuscript. This local assumption also indicates $\boldsymbol{F}_{\boldsymbol{x}_t} \approx \boldsymbol{H}_{\boldsymbol{x}_t}$ which further leads to $\boldsymbol{n}'_t \approx \boldsymbol{H}_{\boldsymbol{x}_t}$. This work also follows this conventional setting, and thus uses $\boldsymbol{n}'_t \approx \boldsymbol{H}_{\boldsymbol{x}_t}$ in Assumption 4.

Appendix F
Auxiliary Lemmas

Before giving our analysis, we first provide some important lemmas.

Lemma 1. *Assume $c_{s,\infty} \leq \|g_k\|_\infty \leq c_\infty$, then we have*

$$\|\boldsymbol{m}_k\|_\infty \leq c_\infty, \quad \|\boldsymbol{n}_i + \delta\|_\infty \leq c_\infty^2 + \delta, \quad \left\|\frac{(\boldsymbol{n}_k+\delta)^p}{(\boldsymbol{n}_{k+1}+\delta)^p}\right\|_\infty \in [1-\mu, 1+\mu] \ (\forall p \in [0,1]),$$

where $\mu = \frac{\beta_2 c_\infty^2}{c_{s,\infty}^2+\delta}$.

See its proof in Appendix I.1.

Lemma 2. *[8] The sequence* $\{\boldsymbol{x}_k\}_{k=0}^T$ *generated by AdamW in Eqn. (2) satisfies*

$$\mathbb{E}\left[\|\boldsymbol{m}_k - \nabla F(\boldsymbol{x}_k)\|^2\right] \leq (1-\beta_1)\mathbb{E}\left[\|\boldsymbol{m}_{k-1} - \nabla F(\boldsymbol{x}_{k-1})\|^2\right] + \frac{(1-\beta_1)^2 L^2}{\beta_1}\mathbb{E}\left[\|\boldsymbol{x}_k - \boldsymbol{x}_{k-1}\|^2\right] + \frac{\beta_1^2\sigma^2}{b}.$$

Lemma 3. *For any* $x \in (0, \frac{1}{4})$*, then there exists* $\alpha > 0$ *such that* $(1-x)^{\frac{3}{2}} \leq 1 - x^{1-\alpha}$.

See its proof in Appendix I.2.

Appendix G
Proof of The Main Results in Section 4

G.1 Proof of Theorem 1

Proof. Here we first use a specific least square problem to analyze the different convergence performance of AdamW and ℓ_2-Adam:

$$\min_{\boldsymbol{x}\in\mathbb{R}} F(\boldsymbol{x}) := \mathbb{E}_{\boldsymbol{\xi}\sim\mathcal{N}(0,1)}\frac{1}{2}\|a\boldsymbol{x} - \boldsymbol{\xi}\|_2^2,$$

where $a \neq 0$ is a constant. In the following we analyze AdamW and ℓ_2-regularized Adam in turn.
Step 1. Analysis of AdamW. For the above problem, AdamW has the following updating rule:

$$\boldsymbol{g}_k = a(a\boldsymbol{x}_k - \boldsymbol{\xi}), \quad \boldsymbol{m}_k = (1-\beta_1)\boldsymbol{m}_{k-1} + \beta_1\boldsymbol{g}_k, \quad \boldsymbol{n}_k = (1-\beta_2)\boldsymbol{n}_{k-1} + \beta_2\boldsymbol{g}_k^2,$$

where $\boldsymbol{m}_0 = 0$ and $\boldsymbol{n}_0 = 0$. In this way, by setting $\boldsymbol{\gamma}_k = 1/\sqrt{\boldsymbol{n}_k + \delta}$ for notation simplicity, the formulation of AdamW can be written as

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta_k\boldsymbol{\gamma}_k\boldsymbol{m}_k - \eta_k\lambda_k\boldsymbol{x}_k = (1-\eta_k\lambda_k)\boldsymbol{x}_k - \eta_k\boldsymbol{\gamma}_k\boldsymbol{m}_k.$$

Since $\boldsymbol{x}_* = 0$ is the optimum solution, we have

$$\boldsymbol{x}_k - \boldsymbol{x}_* = \left[\prod_{i=1}^k(1-\eta_i\lambda_i)\right](\boldsymbol{x}_0 - \boldsymbol{x}_*) - \sum_{i=1}^k \eta_i\boldsymbol{\gamma}_i\boldsymbol{m}_i\left[\prod_{j=i+1}^k(1-\eta_j\lambda_j)\right].$$

So we have

$$\mathbb{E}\|\boldsymbol{x}_k - \boldsymbol{x}_*\| = \left[\prod_{i=1}^k(1-\eta_i\lambda_i)\right]\mathbb{E}\|\boldsymbol{x}_0 - \boldsymbol{x}_*\| + \sum_{i=1}^k \eta_i\boldsymbol{\gamma}_i\mathbb{E}\|\boldsymbol{m}_i\|\left[\prod_{j=i+1}^k(1-\eta_j\lambda_j)\right]$$

Then by setting $\lambda_k = \lambda, \eta_k = \eta$, we have

$$\begin{aligned}\mathbb{E}\|\boldsymbol{x}_k - \boldsymbol{x}_*\| &= \left[\prod_{i=1}^k(1-\eta_i\lambda_i)\right]\mathbb{E}\|\boldsymbol{x}_0 - \boldsymbol{x}_*\| + \sum_{i=1}^k \eta_i\boldsymbol{\gamma}_i\mathbb{E}\|\boldsymbol{m}_i\|\left[\prod_{j=i+1}^k(1-\eta_j\lambda_j)\right]\\ &\leq (1-\eta\lambda)^k\Delta + \sum_{i=1}^k \eta\boldsymbol{\gamma}_i\tau(1-\eta\lambda)^{k-i}\\ &\leq (1-\eta\lambda)^k\Delta + \frac{\eta\tau}{\delta^{\frac{1}{2}}}\sum_{i=1}^k(1-\eta\lambda)^{k-i}\\ &\leq (1-\eta\lambda)^k\Delta + \frac{\tau}{\lambda\delta^{\frac{1}{2}}},\end{aligned}$$

where in the first inequality, we use $\mathbb{E}\|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2 \leq \Delta$, $\mathbb{E}[\|\boldsymbol{g}_k\|_2] \leq \tau$ which yields

$$\|\boldsymbol{m}_{k+1}\|_2 = \|(1-\beta_1)\boldsymbol{m}_k + \beta_1\boldsymbol{g}_k\|_2 \leq (1-\beta_1)\|\boldsymbol{m}_k\|_2 + \beta_1\|\boldsymbol{g}_k\|_2 \leq \tau.$$

Finally, by setting $\lambda = \frac{1}{\delta^{\frac{1}{2}}}k^{\frac{1}{2}+\alpha}$ and $\eta = \frac{3}{2}\delta^{\frac{1}{2}}k^{-\alpha-1}$, we have

$$\mathbb{E}\|\boldsymbol{x}_k - \boldsymbol{x}_*\| \leq \left(1 - \frac{3}{2}k^{-1/2}\right)^k\Delta + \frac{\tau}{k^{\frac{1}{2}+\alpha}} \leq \left(1 - \frac{3}{2}k^{-1/2}\right)^k\Lambda + \frac{\tau}{k^{\frac{1}{2}+\alpha}}, \tag{12}$$

where $\Lambda = \Delta + \eta_0$. This proves the desired result. In Theorem 1, we use the hyper-parameter setting in this proof framework.

Then, we give another solution to prove Eqn. (12). But in Theorem 1, we do not use the hyper-parameter setting in this proof framework, and just provide another analysis framework. To begin with, AdamW has the following updating rule:

$$\boldsymbol{g}_k = a(a\boldsymbol{x}_k - \boldsymbol{\xi}), \quad \boldsymbol{m}_k = (1-\beta_1)\boldsymbol{m}_{k-1} + \beta_1 \boldsymbol{g}_k, \quad \boldsymbol{n}_k = (1-\beta_2)\boldsymbol{n}_{k-1} + \beta_2 \boldsymbol{g}_k^2,$$

where $\boldsymbol{m}_0 = 0$ and $\boldsymbol{n}_0 = 0$. In this way, by setting $\boldsymbol{\gamma}_k = 1/\sqrt{\boldsymbol{n}_k + \delta}$ for notation simplicity, the formulation of AdamW can be written as

$$\begin{aligned}
\boldsymbol{x}_{k+1} =& \boldsymbol{x}_k - \eta_k \boldsymbol{\gamma}_k \boldsymbol{m}_k - \eta_k \lambda_k \boldsymbol{x}_k = (1-\eta_k\lambda_k)\boldsymbol{x}_k - \eta_k\boldsymbol{\gamma}_k\boldsymbol{m}_k = (1-\eta_k\lambda_k)\boldsymbol{x}_k - \eta_k\boldsymbol{\gamma}_k((1-\beta_1)\boldsymbol{m}_{k-1} + \beta_1\boldsymbol{g}_k) \\
=&(1-\eta_k\lambda_k)\boldsymbol{x}_k - \eta_k\boldsymbol{\gamma}_k(1-\beta_1)\frac{(1-\eta_{k-1}\lambda_{k-1})\boldsymbol{x}_{k-1} - \boldsymbol{x}_k}{\eta_{k-1}\boldsymbol{\gamma}_{k-1}} - \eta_k\boldsymbol{\gamma}_k\beta_1\boldsymbol{g}_k \\
=& \left(1 - \eta_k\lambda_k + \frac{\eta_k\boldsymbol{\gamma}_k}{\eta_{k-1}\boldsymbol{\gamma}_{k-1}}(1-\beta_1)\right)\boldsymbol{x}_k - (1-\beta_1)(1-\eta_{k-1}\lambda_{k-1})\frac{\eta_k\boldsymbol{\gamma}_k}{\eta_{k-1}\boldsymbol{\gamma}_{k-1}}\boldsymbol{x}_{k-1} - \eta_k\boldsymbol{\gamma}_k\beta_1\boldsymbol{g}_k.
\end{aligned}$$

Since $\boldsymbol{x}_* = 0$ is the optimum solution, we have

$$\begin{aligned}
\begin{bmatrix} \boldsymbol{x}_{k+1} - \boldsymbol{x}_* \\ \boldsymbol{x}_k - \boldsymbol{x}_* \end{bmatrix} =& \begin{bmatrix} 1 - \eta_k\lambda_k + \frac{\eta_k\boldsymbol{\gamma}_k}{\eta_{k-1}\boldsymbol{\gamma}_{k-1}}(1-\beta_1) & -(1-\beta_1)(1-\eta_{k-1}\lambda_{k-1})\frac{\eta_k\boldsymbol{\gamma}_k}{\eta_{k-1}\boldsymbol{\gamma}_{k-1}} \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_k - \boldsymbol{x}_* \\ \boldsymbol{x}_{k-1} - \boldsymbol{x}_* \end{bmatrix} - \begin{bmatrix} \eta_k\boldsymbol{\gamma}_k\beta_1\boldsymbol{g}_k \\ 0 \end{bmatrix} \\
=& \boldsymbol{A}_k \begin{bmatrix} \boldsymbol{x}_k - \boldsymbol{x}_* \\ \boldsymbol{x}_{k-1} - \boldsymbol{x}_* \end{bmatrix} - \begin{bmatrix} \eta_k\boldsymbol{\gamma}_k\beta_1\boldsymbol{g}_k \\ 0 \end{bmatrix} = \left[\prod_{i=1}^{k} \boldsymbol{A}_i\right] \begin{bmatrix} \boldsymbol{x}_1 - \boldsymbol{x}_* \\ \boldsymbol{x}_0 - \boldsymbol{x}_* \end{bmatrix} - \sum_{i=1}^{k} \left[\prod_{j=i+1}^{k} \boldsymbol{A}_j\right] \begin{bmatrix} \eta_i\boldsymbol{\gamma}_i\beta_1\boldsymbol{g}_i \\ 0 \end{bmatrix},
\end{aligned}$$

For matrix $\boldsymbol{A}_k$, we can compute its eigenvalues as

$$\frac{c_k \pm \sqrt{c_k^2 - 4b_k}}{2},$$

where $c_k = 1 - \eta_k\lambda_k + \frac{\eta_k\boldsymbol{\gamma}_k}{\eta_{k-1}\boldsymbol{\gamma}_{k-1}}(1-\beta_1)$ and $b_k = (1-\beta_1)(1-\eta_{k-1}\lambda_{k-1})\frac{\eta_k\boldsymbol{\gamma}_k}{\eta_{k-1}\boldsymbol{\gamma}_{k-1}}$. By setting

$$c_k^2 - 4b_k \leq 0, \tag{13}$$

then the two eigenvalues are complex, and in particular they must be complex conjugates of each other. So they must have the same absolute value (because a complex number and its conjugate have the same absolute value) and the square of their absolute value must be equal to their product (because a complex number's absolute value is the square root of itself times its conjugate). Explicitly, if we call the eigenvalues d_1 and d_2:

$$d_1^* = d_2, \qquad |d_1^2| = |d_2|^2 = d_1 d_2^* = d_1 d_2 = b_k,$$

which means that

$$d_1 = d_2 = \sqrt{b_k}.$$

So we have

$$\begin{aligned}
\mathbb{E}\left\|\begin{bmatrix} \boldsymbol{x}_{k+1} - \boldsymbol{x}_* \\ \boldsymbol{x}_k - \boldsymbol{x}_* \end{bmatrix}\right\| \leq& \mathbb{E}\left\|\left[\prod_{i=1}^{k} \boldsymbol{A}_i\right]\right\| \left\|\begin{bmatrix} \boldsymbol{x}_1 - \boldsymbol{x}_* \\ \boldsymbol{x}_0 - \boldsymbol{x}_* \end{bmatrix}\right\| + \mathbb{E}\sum_{i=1}^{k} \left\|\left[\prod_{j=i+1}^{k} \boldsymbol{A}_j\right]\right\| \left\|\begin{bmatrix} \eta_i\boldsymbol{\gamma}_i\beta_1\boldsymbol{g}_i \\ 0 \end{bmatrix}\right\| \\
\leq& \mathbb{E}\prod_{i=1}^{k} b_i^{\frac{1}{2}} \left\|\begin{bmatrix} \boldsymbol{x}_1 - \boldsymbol{x}_* \\ \boldsymbol{x}_0 - \boldsymbol{x}_* \end{bmatrix}\right\| + \mathbb{E}\sum_{i=1}^{k}\prod_{j=i+1}^{k} b_j^{\frac{1}{2}} \eta_i\boldsymbol{\gamma}_i\beta_1\|\boldsymbol{g}_i\| \\
\overset{①}{\leq}& (1-\beta_1)^{\frac{k}{2}}\frac{\eta_k\boldsymbol{\gamma}_k}{\eta_0\boldsymbol{\gamma}_0}\prod_{j=0}^{k-1}(1-\eta_j\lambda_j)\mathbb{E}\left\|\begin{bmatrix} \boldsymbol{x}_1 - \boldsymbol{x}_* \\ \boldsymbol{x}_0 - \boldsymbol{x}_* \end{bmatrix}\right\| + \tau\beta_1\eta_k\boldsymbol{\gamma}_k\sum_{i=1}^{k}(1-\beta_1)^{\frac{k-i}{2}}\prod_{j=i}^{k-1}(1-\eta_j\lambda_j) \\
\overset{②}{\leq}& (1-\beta_1)^{\frac{3k}{2}}\frac{\eta_k^2\boldsymbol{\gamma}_k^2}{\eta_0^2\boldsymbol{\gamma}_0^2}\mathbb{E}\left\|\begin{bmatrix} \boldsymbol{x}_1 - \boldsymbol{x}_* \\ \boldsymbol{x}_0 - \boldsymbol{x}_* \end{bmatrix}\right\| + \tau\beta_1\eta_k\boldsymbol{\gamma}_k\sum_{i=1}^{k}(1-\beta_1)^{\frac{3(k-i)}{2}}\frac{\eta_k\boldsymbol{\gamma}_k}{\eta_i\boldsymbol{\gamma}_i},
\end{aligned}$$

where ① holds since $\prod_{j=i+1}^{k} b_i^{\frac{1}{2}} = (1-\beta_1)^{\frac{k-i}{2}}\frac{\eta_k\boldsymbol{\gamma}_k}{\eta_i\boldsymbol{\gamma}_i}\prod_{j=i}^{k-1}(1-\eta_j\lambda_j)$ and $\mathbb{E}[\|\boldsymbol{g}_k\|_2] \leq \tau$; ② holds because of Eqn. (13). Then by setting $\lambda_k = \boldsymbol{\gamma}_k, \eta_k = \beta_1/\boldsymbol{\gamma}_k$ and $\beta_1 = 1/\sqrt{k}$, then the condition (13) is satisfied. Assume that $\mathbb{E}\|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2 \leq \Delta$, then we have

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{x}_1 - \boldsymbol{x}_*\| =& \mathbb{E}\|\boldsymbol{x}_0 - \eta_0\boldsymbol{\gamma}_0\boldsymbol{m}_0 - \eta_0\lambda_0\boldsymbol{x}_0 - \boldsymbol{x}_*\| = \mathbb{E}\|\boldsymbol{x}_0 - \eta_0\boldsymbol{g}_0/\sqrt{\|\boldsymbol{g}_0\|^2 + \delta} - \eta_0\lambda_0\boldsymbol{x}_0 - \boldsymbol{x}_*\| \\
\leq& (1-\eta_0\lambda_0)\mathbb{E}\|\boldsymbol{x}_0 - \boldsymbol{x}_*\| + \eta_0 \leq (1-\eta_0\lambda_0)\Delta + \eta_0,
\end{aligned}$$

where we use $\boldsymbol{x}_* = 0$. In this way, by setting $\Lambda = \Delta + \eta_0$, we have

$$\begin{aligned}
\mathbb{E}\left\|\begin{bmatrix} \boldsymbol{x}_{k+1} - \boldsymbol{x}_* \\ \boldsymbol{x}_k - \boldsymbol{x}_* \end{bmatrix}\right\| \leq& \left(1 - \frac{1}{\sqrt{k}}\right)^{\frac{3k}{2}}\frac{\eta_k^2\boldsymbol{\gamma}_k^2}{\eta_0^2\boldsymbol{\gamma}_0^2}\Lambda + \frac{\tau\eta_k\boldsymbol{\gamma}_k}{\sqrt{k}}\sum_{i=1}^{k}\left(1 - \frac{1}{\sqrt{k}}\right)^{\frac{3(k-i)}{2}}\frac{\eta_k\boldsymbol{\gamma}_k}{\eta_i\boldsymbol{\gamma}_i} \\
=& \left(1 - \frac{1}{\sqrt{k}}\right)^{\frac{3k}{2}}\Lambda + \frac{\tau}{k\left(1 - \left(1 - \frac{1}{\sqrt{k}}\right)^{\frac{3}{2}}\right)} \overset{①}{\leq} \left(1 - \frac{1}{\sqrt{k}}\right)^{\frac{3k}{2}}\Lambda + \frac{\tau}{k^{\frac{1}{2}+\alpha}},
\end{aligned}$$

where ① holds since from Lemma 3, we have $\frac{1}{k\left(1-\left(1-\frac{1}{\sqrt{k}}\right)^{\frac{3}{2}}\right)} \leq \frac{1}{k\left(1-\left(1-k^{-\frac{1}{2}+\alpha}\right)\right)} = \frac{1}{k^{\frac{1}{2}+\alpha}}$.

Step 2. Analysis of ℓ_2-Adam. By comparison, the ℓ_2-regularized Adam can be formulated as

$$\boldsymbol{g}_k = a(a\boldsymbol{x}_k - \boldsymbol{\xi}), \quad \boldsymbol{m}_k = (1-\beta_1)\boldsymbol{m}_{k-1} + \beta_1(\boldsymbol{g}_k + \lambda_k \boldsymbol{x}_k), \quad \boldsymbol{n}_k = (1-\beta_2)\boldsymbol{n}_{k-1} + \beta_2(\boldsymbol{g}_k + \lambda_k \boldsymbol{x}_k)^2,$$

where $\boldsymbol{m}_0 = 0$ and $\boldsymbol{n}_0 = 0$. In this way, by setting $\boldsymbol{\gamma}_k = 1/\sqrt{\boldsymbol{n}_k + \delta}$ for notation simplicity, the formulation of ℓ_2-Adam can be written as

$$\begin{aligned}
\boldsymbol{x}_{k+1} =& \boldsymbol{x}_k - \eta_k \boldsymbol{\gamma}_k \boldsymbol{m}_k = (1-\eta_k\lambda_k)\boldsymbol{x}_k - \eta_k\boldsymbol{\gamma}_k((1-\beta_1)\boldsymbol{m}_{k-1} + \beta_1(\boldsymbol{g}_k + \lambda_k\boldsymbol{x}_k)) \\
=& (1-\eta_k\boldsymbol{\gamma}_k\lambda_k\beta_1)\boldsymbol{x}_k - \eta_k\boldsymbol{\gamma}_k(1-\beta_1)\frac{\boldsymbol{x}_{k-1}-\boldsymbol{x}_k}{\eta_{k-1}\boldsymbol{\gamma}_{k-1}} - \eta_k\boldsymbol{\gamma}_k\beta_1\boldsymbol{g}_k \\
=& \left(1-\eta_k\boldsymbol{\gamma}_k\lambda_k\beta_1 + \frac{\eta_k\boldsymbol{\gamma}_k}{\eta_{k-1}\boldsymbol{\gamma}_{k-1}}(1-\beta_1)\right)\boldsymbol{x}_k - (1-\beta_1)\frac{\eta_k\boldsymbol{\gamma}_k}{\eta_{k-1}\boldsymbol{\gamma}_{k-1}}\boldsymbol{x}_{k-1} - \eta_k\boldsymbol{\gamma}_k\beta_1\boldsymbol{g}_k.
\end{aligned}$$

Since $\boldsymbol{x}_* = 0$ is the optimum solution, we have

$$\begin{aligned}
\begin{bmatrix}\boldsymbol{x}_{k+1}-\boldsymbol{x}_* \\ \boldsymbol{x}_k - \boldsymbol{x}_*\end{bmatrix} =& \begin{bmatrix}1-\eta_k\boldsymbol{\gamma}_k\lambda_k\beta_1 + \frac{\eta_k\boldsymbol{\gamma}_k}{\eta_{k-1}\boldsymbol{\gamma}_{k-1}}(1-\beta_1) & -(1-\beta_1)\frac{\eta_k\boldsymbol{\gamma}_k}{\eta_{k-1}\boldsymbol{\gamma}_{k-1}} \\ 1 & 0\end{bmatrix}\begin{bmatrix}\boldsymbol{x}_k-\boldsymbol{x}_* \\ \boldsymbol{x}_{k-1}-\boldsymbol{x}_*\end{bmatrix} - \begin{bmatrix}\eta_k\boldsymbol{\gamma}_k\beta_1\boldsymbol{g}_k \\ 0\end{bmatrix} \\
=& \boldsymbol{A}_k\begin{bmatrix}\boldsymbol{x}_k-\boldsymbol{x}_* \\ \boldsymbol{x}_{k-1}-\boldsymbol{x}_*\end{bmatrix} - \begin{bmatrix}\eta_k\boldsymbol{\gamma}_k\beta_1\boldsymbol{g}_k \\ 0\end{bmatrix} = \left[\prod_{i=1}^k \boldsymbol{A}_i\right]\begin{bmatrix}\boldsymbol{x}_1-\boldsymbol{x}_* \\ \boldsymbol{x}_0-\boldsymbol{x}_*\end{bmatrix} - \sum_{i=1}^k\left[\prod_{j=i+1}^k \boldsymbol{A}_j\right]\begin{bmatrix}\eta_i\boldsymbol{\gamma}_i\beta_1\boldsymbol{g}_i \\ 0\end{bmatrix},
\end{aligned}$$

For matrix $\boldsymbol{A}_k$, we can compute its eigenvalues as

$$\frac{c_k \pm \sqrt{c_k^2 - 4b_k}}{2},$$

where $c_k = 1 - \eta_k\boldsymbol{\gamma}_k\lambda_k\beta_1 + \frac{\eta_k\boldsymbol{\gamma}_k}{\eta_{k-1}\boldsymbol{\gamma}_{k-1}}(1-\beta_1)$ and $b_k = (1-\beta_1)\frac{\eta_k\boldsymbol{\gamma}_k}{\eta_{k-1}\boldsymbol{\gamma}_{k-1}}$. By setting

$$c_k^2 - 4b_k \leq 0, \tag{14}$$

then the two eigenvalues are complex, and in particular they must be complex conjugates of each other. So they must have the same absolute value (because a complex number and its conjugate have the same absolute value) and the square of their absolute value must be equal to their product (because a complex number's absolute value is the square root of itself times its conjugate). Explicitly, if we call the eigenvalues d_1 and d_2:

$$d_1^* = d_2, \qquad |d_1^2| = |d_2|^2 = d_1 d_2^* = d_1 d_2 = b_k,$$

which means that

$$d_1 = d_2 = \sqrt{b_k}.$$

So we have

$$\begin{aligned}
\mathbb{E}\left\|\begin{bmatrix}\boldsymbol{x}_{k+1}-\boldsymbol{x}_* \\ \boldsymbol{x}_k-\boldsymbol{x}_*\end{bmatrix}\right\| \leq& \mathbb{E}\left\|\left[\prod_{i=1}^k \boldsymbol{A}_i\right]\right\|\left\|\begin{bmatrix}\boldsymbol{x}_1-\boldsymbol{x}_* \\ \boldsymbol{x}_0-\boldsymbol{x}_*\end{bmatrix}\right\| + \mathbb{E}\sum_{i=1}^k\left\|\left[\prod_{j=i+1}^k \boldsymbol{A}_j\right]\right\|\left\|\begin{bmatrix}\eta_i\boldsymbol{\gamma}_i\beta_1\boldsymbol{g}_i \\ 0\end{bmatrix}\right\| \\
\leq& \mathbb{E}\prod_{i=1}^k b_i^{\frac{1}{2}}\left\|\begin{bmatrix}\boldsymbol{x}_1-\boldsymbol{x}_* \\ \boldsymbol{x}_0-\boldsymbol{x}_*\end{bmatrix}\right\| + \mathbb{E}\sum_{i=1}^k\prod_{j=i+1}^k b_j^{\frac{1}{2}}\eta_i\boldsymbol{\gamma}_i\beta_1\|\boldsymbol{g}_i\| \\
\overset{①}{\leq}& (1-\beta_1)^{\frac{k}{2}}\frac{\eta_k\boldsymbol{\gamma}_k}{\eta_0\boldsymbol{\gamma}_0}\mathbb{E}\left\|\begin{bmatrix}\boldsymbol{x}_1-\boldsymbol{x}_* \\ \boldsymbol{x}_0-\boldsymbol{x}_*\end{bmatrix}\right\| + \tau\beta_1\eta_k\boldsymbol{\gamma}_k\sum_{i=1}^k(1-\beta_1)^{\frac{k-i}{2}},
\end{aligned}$$

where ① holds since $\prod_{j=i+1}^k b_i^{\frac{1}{2}} = (1-\beta_1)^{\frac{k-i}{2}}\frac{\eta_k\boldsymbol{\gamma}_k}{\eta_i\boldsymbol{\gamma}_i}$ and $\mathbb{E}[\|\boldsymbol{g}_k\|_2] \leq \tau$; ② holds because of Eqn. (14).

Then by setting $\eta_k = \beta_1/\boldsymbol{\gamma}_k$ and $\beta_1 = 1/\sqrt{k}$, then the condition (14) becomes:

$$c_k^2 - 4b_k \leq 0 \qquad \Rightarrow \qquad \lambda_k \in [\frac{1}{\beta_1}\left(2-\beta_1-2\sqrt{1-\beta_1}\right), \frac{1}{\beta_1}\left(2-\beta_1+2\sqrt{1-\beta_1}\right)].$$

So we can set

$$\lambda_k = \lambda = \mathcal{O}\left(\sqrt{k}\right)$$

Assume that $\mathbb{E}\|\boldsymbol{x}_0 - \boldsymbol{x}_*\|_2 \leq \Delta$, then we have

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{x}_1-\boldsymbol{x}_*\| =& \mathbb{E}\|\boldsymbol{x}_0 - \eta_0\boldsymbol{\gamma}_0\boldsymbol{m}_0 - \boldsymbol{x}_*\| = \mathbb{E}\|\boldsymbol{x}_0 - \eta_0(\boldsymbol{g}_0+\lambda_0\boldsymbol{x}_0)/\sqrt{\|\boldsymbol{g}_0+\lambda_0\boldsymbol{x}_0\|^2+\delta} - \boldsymbol{x}_*\| \\
\leq& \mathbb{E}\|\boldsymbol{x}_0-\boldsymbol{x}_*\| + \eta_0 \leq \Delta + \eta_0,
\end{aligned}$$

where we use $\boldsymbol{x}_* = 0$. In this way, by setting $\Lambda = \Delta + \eta_0$, we have

$$\begin{aligned}\mathbb{E}\left\|\begin{bmatrix}\boldsymbol{x}_{k+1}-\boldsymbol{x}_*\\ \boldsymbol{x}_k-\boldsymbol{x}_*\end{bmatrix}\right\| &\leq \left(1-\frac{1}{\sqrt{k}}\right)^{\frac{k}{2}}\frac{\eta_k\gamma_k}{\eta_0\gamma_0}\Lambda + \frac{\tau\eta_k\gamma_k}{\sqrt{k}}\sum_{i=1}^{k}\left(1-\frac{1}{\sqrt{k}}\right)^{\frac{(k-i)}{2}}\\ &=\left(1-\frac{1}{\sqrt{k}}\right)^{\frac{k}{2}}\Lambda + \frac{\tau}{k\left(1-\left(1-\frac{1}{\sqrt{k}}\right)^{\frac{1}{2}}\right)} \overset{①}{\leq} \left(1-\frac{1}{\sqrt{k}}\right)^{\frac{k}{2}}\Lambda + \frac{2\tau}{k^{\frac{1}{2}}},\end{aligned}$$

where ① holds since from Lemma 3, we have $\left(1-\frac{1}{\sqrt{k}}\right)^{\frac{1}{2}} \leq 1-\frac{1}{2k^{\frac{1}{2}}}$. The proof is completed. □

G.2 Proof of Theorem 2

Proof. For brevity, we let

$$\boldsymbol{v}_k = \sqrt{\boldsymbol{n}_k+\delta}.$$

When $\|\boldsymbol{g}_i\|_\infty \leq c_\infty$, we have $\|\boldsymbol{m}_k\|_\infty \leq c_\infty$ and $\delta \leq \|\boldsymbol{n}_i+\delta\|_\infty \leq c_\infty^2+\delta$ in Lemma 1. For brevity, let

$$c_1 := \delta^p \leq \|\boldsymbol{v}_k\|_\infty \leq c_2 := (c_\infty^2+\delta)^p. \tag{15}$$

Also we define

$$\boldsymbol{u}_k := \boldsymbol{m}_k + \lambda\boldsymbol{x}_k\odot\boldsymbol{v}_k, \qquad \boldsymbol{x}_{k+1}-\boldsymbol{x}_k = -\eta\frac{\boldsymbol{m}_k+\lambda\boldsymbol{x}_k\odot\boldsymbol{v}_k}{\boldsymbol{v}_k} = -\eta\frac{\boldsymbol{u}_k}{\boldsymbol{v}_k}.$$

Moreover, we also define $F_k(\boldsymbol{x}_k)$ as follows:

$$F_k(\boldsymbol{x}_k) = F(\boldsymbol{x}) + \frac{\lambda_k}{2}\|\boldsymbol{x}\|_{\boldsymbol{v}_k}^2 = \mathbb{E}_{\boldsymbol{\xi}}[f(\boldsymbol{x};\boldsymbol{\xi})] + \frac{\lambda_k}{2}\|\boldsymbol{x}\|_{\boldsymbol{v}_k}^2,$$

where $\lambda_k = \lambda(1-\mu)^k$ in which $\mu = \frac{\beta_2 c_\infty^2}{\delta}$.

Then by using the smoothness of $f(\boldsymbol{x};\boldsymbol{\zeta})$, we can obtain

$$\begin{aligned}&F_{k+1}(\boldsymbol{x}_{k+1})\\ \leq& F(\boldsymbol{x}_k) + \langle\nabla F(\boldsymbol{x}_k), \boldsymbol{x}_{k+1}-\boldsymbol{x}_k\rangle + \frac{L}{2}\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\|^2 + \frac{\lambda_{k+1}}{2}\|\boldsymbol{x}_{k+1}\|_{\boldsymbol{v}_{k+1}}^2\\ \overset{①}{\leq}& F(\boldsymbol{x}_k) + \langle\nabla F(\boldsymbol{x}_k), \boldsymbol{x}_{k+1}-\boldsymbol{x}_k\rangle + \frac{L}{2}\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\|^2 + \frac{\lambda_{k+1}}{2(1-\mu)}\|\boldsymbol{x}_{k+1}\|_{\boldsymbol{v}_k}^2\\ \overset{②}{\leq}& F(\boldsymbol{x}_k) + \frac{\lambda_k}{2}\|\boldsymbol{x}_k\|_{\boldsymbol{v}_k}^2 + \langle\nabla F(\boldsymbol{x}_k)+\lambda\boldsymbol{x}_k\odot\boldsymbol{v}_k, \boldsymbol{x}_{k+1}-\boldsymbol{x}_k\rangle + \frac{L}{2}\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\|^2 + \frac{\lambda_k}{2}\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\|_{\boldsymbol{v}_k}^2\\ =& F_k(\boldsymbol{x}_k) + \langle\nabla F(\boldsymbol{x}_k)+\lambda\boldsymbol{x}_k\odot\boldsymbol{v}_k, \boldsymbol{x}_{k+1}-\boldsymbol{x}_k\rangle + \frac{L}{2}\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\|^2 + \frac{\lambda_k}{2}\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\|_{\boldsymbol{v}_k}^2\\ =& F_k(\boldsymbol{x}_k) - \eta\left\langle\nabla F(\boldsymbol{x}_k)+\lambda\boldsymbol{x}_k\odot\boldsymbol{v}_k, \frac{\boldsymbol{u}_k}{\boldsymbol{v}_k}\right\rangle + \frac{L\eta^2}{2}\left\|\frac{\boldsymbol{u}_k}{\boldsymbol{v}_k}\right\|^2 + \frac{\lambda_k\eta^2}{2}\left\|\frac{\boldsymbol{u}_k}{\boldsymbol{v}_k}\right\|_{\boldsymbol{v}_k}^2\\ =& F_k(\boldsymbol{x}_k) + \frac{1}{2}\left\|\sqrt{\frac{\eta}{\boldsymbol{v}_k}}(\nabla F(\boldsymbol{x}_k)+\lambda\boldsymbol{x}_k\odot\boldsymbol{v}_k-\boldsymbol{u}_k)\right\|^2 - \frac{1}{2}\left\|\sqrt{\frac{\eta}{\boldsymbol{v}_k}}(\nabla F(\boldsymbol{x}_k)+\lambda_k\boldsymbol{x}_k\odot\boldsymbol{v}_k)\right\|^2\\ &-\frac{1}{2}\left\|\sqrt{\frac{\eta}{\boldsymbol{v}_k}}\boldsymbol{u}_k\right\|^2 + \frac{L\eta^2}{2}\left\|\frac{\boldsymbol{u}_k}{\boldsymbol{v}_k}\right\|^2 + \frac{\lambda_k\eta^2}{2}\left\|\frac{\boldsymbol{u}_k}{\boldsymbol{v}_k}\right\|_{\boldsymbol{v}_k}^2\\ \leq& F_k(\boldsymbol{x}_k) + \frac{\eta}{2c_1}\|\nabla F(\boldsymbol{x}_k)-\boldsymbol{m}_k\|^2 - \frac{\eta}{2c_2}\|\nabla F(\boldsymbol{x}_k)+\lambda\boldsymbol{x}_k\odot\boldsymbol{v}_k\|^2 - \left[\frac{\eta}{2c_2}-\frac{L\eta^2}{2c_1^2}-\frac{\lambda_k\eta^2}{2c_1}\right]\|\boldsymbol{u}_k\|^2\\ \overset{③}{\leq}& F_k(\boldsymbol{x}_k) + \frac{\eta}{2c_1}\|\nabla F(\boldsymbol{x}_k)-\boldsymbol{m}_k\|^2 - \frac{\eta}{2c_2}\|\nabla F(\boldsymbol{x}_k)+\lambda\boldsymbol{x}_k\odot\boldsymbol{v}_k\|^2 - \frac{\eta}{4c_2}\|\boldsymbol{u}_k\|^2\end{aligned} \tag{16}$$

where ① holds since Lemma 1 proves $\left\|\frac{(\boldsymbol{n}_k+\delta)^{\frac{1}{2}}}{(\boldsymbol{n}_{k+1}+\delta)^{\frac{1}{2}}}\right\|_\infty \in [1-\mu, 1+\mu]$ $(\forall p \in [0,1])$ in which $\mu = \frac{\beta_2 c_\infty^2}{\delta}$; ② holds because $\lambda_{k+1} = \frac{\lambda_{k+1}}{1-\mu}$ and

$$\|\boldsymbol{x}_{k+1}\|_{\boldsymbol{v}_k}^2 = \|\boldsymbol{x}_k\|_{\boldsymbol{v}_k}^2 + 2\langle\boldsymbol{x}_{k+1}-\boldsymbol{x}_k, \boldsymbol{x}_k\rangle_{\boldsymbol{v}_k} + \|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\|_{\boldsymbol{v}_k}^2;$$

③ holds, since we set $\eta \leq \frac{c_1^2}{2c_2(L+\lambda c_1)}$ such that $\frac{\eta}{4c_2} \geq \frac{L\eta^2}{2c_1^2} + \frac{\lambda_k\eta^2}{2c_1}$.

From Lemma 2, we have

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{m}_k - \nabla F(\boldsymbol{x}_k)\|^2\right] \le&(1-\beta_1)\mathbb{E}\left[\|\boldsymbol{m}_{k-1} - \nabla F(\boldsymbol{x}_{k-1})\|^2\right] + \frac{(1-\beta_1)^2 L^2}{\beta_1}\mathbb{E}\left[\|\boldsymbol{x}_k - \boldsymbol{x}_{k-1}\|^2\right] + \frac{\beta_1^2\sigma^2}{b} \\ \le&(1-\beta_1)\mathbb{E}\left[\|\boldsymbol{m}_{k-1} - \nabla F(\boldsymbol{x}_{k-1})\|^2\right] + \frac{(1-\beta_1)^2 L^2\eta^2}{\beta_1 c_1^2}\mathbb{E}\left[\|\boldsymbol{u}_k\|^2\right] + \frac{\beta_1^2\sigma^2}{b}\end{aligned} \tag{17}$$

where we use $\boldsymbol{x}_k - \boldsymbol{x}_{k-1} = \eta\frac{\boldsymbol{u}_k}{\boldsymbol{v}_k}$.

Then we add Eqn. (16) and $\alpha\times$ (17) as follows:

$$\begin{aligned}&F_{k+1}(\boldsymbol{x}_{k+1}) + \alpha\mathbb{E}\left[\|\boldsymbol{m}_{k+1} - \nabla F(\boldsymbol{x}_{k+1})\|^2\right] \\ \le&F_k(\boldsymbol{x}_k) - \frac{\eta}{2c_2}\|\nabla F(\boldsymbol{x}_k) + \lambda_k\boldsymbol{x}_k \odot \boldsymbol{v}_k\|^2 + \left[(1-\beta_1)\alpha + \frac{\eta}{2c_1}\right]\mathbb{E}\left[\|\boldsymbol{m}_{k-1} - \nabla F(\boldsymbol{x}_{k-1})\|^2\right] \\ &-\left[\frac{\eta}{4c_2} - \frac{\alpha(1-\beta_1)^2L^2\eta^2}{\beta_1 c_1^2}\right]\mathbb{E}\left[\|\boldsymbol{u}_k\|^2\right] + \frac{\alpha\beta_1^2\sigma^2}{b}.\end{aligned}$$

Then by setting $\alpha = \frac{\eta}{2c_1\beta_1}$ and $G(\boldsymbol{x}_{k+1}) = F_{k+1}(\boldsymbol{x}_{k+1}) + \frac{\eta}{2c_1\beta_1}\mathbb{E}\left[\|\boldsymbol{m}_{k+1} - \nabla F(\boldsymbol{x}_{k+1})\|^2\right]$, we can obtain

$$\begin{aligned}G(\boldsymbol{x}_{k+1}) \le& G(\boldsymbol{x}_k) - \frac{\eta}{2c_2}\mathbb{E}\|\nabla F(\boldsymbol{x}_k) + \lambda_k\boldsymbol{x}_k \odot \boldsymbol{v}_k\|^2 - \frac{\eta}{4c_2}\left[1 - \frac{2c_2(1-\beta_1)^2L^2\eta^2}{\beta_1^2c_1^3}\right]\mathbb{E}\left[\|\boldsymbol{u}_k\|^2\right] + \frac{\eta\beta_1\sigma^2}{2c_1 b} \\ \overset{①}{\le}& G(\boldsymbol{x}_k) - \frac{\eta}{2c_2}\mathbb{E}\left[\|\nabla F(\boldsymbol{x}_k) + \lambda_k\boldsymbol{x}_k \odot \boldsymbol{v}_k\|^2\right] - \frac{\eta}{8c_2}\mathbb{E}\left[\|\boldsymbol{u}_k\|^2\right] + \frac{\eta\beta_1\sigma^2}{2c_1 b},\end{aligned}$$

where ① holds since set $\eta \le \frac{\beta_1 c_1}{2(1-\beta_1)L}\sqrt{\frac{c_1}{c_2}}$ such that $\frac{2c_2(1-\beta_1)^2L^2\eta^2}{\beta_1^2c_1^3} \le \frac{1}{2}$.

Then summing the above inequality from $k=0$ to $k=T-1$ gives

$$\begin{aligned}\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}\left[\|\nabla F(\boldsymbol{x}_k) + \lambda_k\boldsymbol{x}_k \odot \boldsymbol{v}_k\|^2 + \frac{1}{4}\|\boldsymbol{u}_k\|^2\right] \le& \frac{2c_2}{\eta T}\left[G(\boldsymbol{x}_0) - G(\boldsymbol{x}_T)\right] + \frac{c_2\beta_1\sigma^2}{c_1 b} \\ \le& \frac{2c_2\Delta}{\eta T} + \frac{c_2\sigma^2}{c_1\beta_1 bT} + \frac{c_2\beta_1\sigma^2}{c_1 b} \\ \le& \epsilon^2,\end{aligned} \tag{18}$$

where we set $T \ge \max\left(\frac{6c_2\Delta}{\eta\epsilon^2}, \frac{3c_2\sigma^2}{c_1\beta_1 b\epsilon^2}\right)$ and $\beta_1 \le \frac{c_1 b\epsilon^2}{3c_2\beta_1\sigma^2}$, in which

$$\begin{aligned}&G(\boldsymbol{x}_0) - G(\boldsymbol{x}_T) \\ =&F_0(\boldsymbol{x}_0) + \frac{\eta}{2c_1\beta_1}\mathbb{E}\left[\|\boldsymbol{m}_0 - \nabla F(\boldsymbol{x}_0)\|^2\right] - F_T(\boldsymbol{x}_T) - \frac{\eta}{2c_1\beta_1}\mathbb{E}\left[\|\boldsymbol{m}_T - \nabla F(\boldsymbol{x}_T)\|^2\right] \\ =&F(\boldsymbol{x}_0) + \frac{\eta}{2c_1\beta_1}\mathbb{E}\left[\|\boldsymbol{m}_0 - \nabla F(\boldsymbol{x}_0)\|^2\right] - F(\boldsymbol{x}_T) - \lambda_T\|\boldsymbol{x}_T\|_{\boldsymbol{v}_T} - \frac{\eta}{2c_1\beta_1}\mathbb{E}\left[\|\boldsymbol{m}_T - \nabla F(\boldsymbol{x}_T)\|^2\right] \\ \le&F(\boldsymbol{x}_0) + \frac{\eta}{2c_1\beta_1}\mathbb{E}\left[\|\boldsymbol{m}_0 - \nabla F(\boldsymbol{x}_0)\|^2\right] - F(\boldsymbol{x}_T) \\ \le&\Delta + \frac{\eta}{2c_1\beta_1}\mathbb{E}\left[\|\boldsymbol{m}_0 - \nabla F(\boldsymbol{x}_0)\|^2\right] \\ \le&\Delta + \frac{\eta\sigma^2}{2c_1\beta_1 b},\end{aligned}$$

where $\Delta = F(\boldsymbol{x}_0) - F(\boldsymbol{x}_*)$. This result directly bounds

$$\frac{1}{T}\sum_{k=0}^{T-1}\|\boldsymbol{v}_k \odot (\boldsymbol{x}_k - \boldsymbol{x}_{k+1})\|^2 = \frac{\eta^2}{T}\sum_{k=0}^{T-1}\|\boldsymbol{m}_k + \lambda_k\boldsymbol{x}_k \odot \boldsymbol{v}_k\|^2 \le \frac{\eta^2}{T}\sum_{k=0}^{T-1}\|\boldsymbol{u}_k\|^2 \le 4\eta^2\epsilon^2.$$

and

$$\frac{1}{T}\sum_{k=0}^{T-1}\|\boldsymbol{x}_k - \boldsymbol{x}_{k+1}\|^2 \le \frac{4\eta^2\epsilon^2}{c_1^2}.$$

Besides, we have

$$\begin{aligned}\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}\left[\|\boldsymbol{m}_k-\nabla F(\boldsymbol{x}_k)\|^2\right] \leq & \frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}\left[\|\boldsymbol{m}_k+\lambda_k\boldsymbol{x}_k\odot\boldsymbol{v}_k-\nabla F(\boldsymbol{x}_k)-\lambda_k\boldsymbol{x}_k\odot\boldsymbol{v}_k\|^2\right]\\ \leq & \frac{2}{T}\sum_{k=0}^{T-1}\mathbb{E}\left[\|\boldsymbol{m}_k+\lambda_k\boldsymbol{x}_k\odot\boldsymbol{v}_k\|^2+\|\nabla F(\boldsymbol{x}_k)-\lambda_k\boldsymbol{x}_k\odot\boldsymbol{v}_k\|^2\right]\\ = & \frac{2}{T}\sum_{k=0}^{T-1}\mathbb{E}\left[\|\boldsymbol{m}_k+\lambda_k\boldsymbol{x}_k\odot\boldsymbol{v}_k\|^2+\|\boldsymbol{u}_k\|^2\right]\\ \leq & 2\left[\epsilon^2+\frac{3}{4}\times 4\epsilon^2\right]\leq 8\epsilon^2.\end{aligned}$$

For all hyper-parameters, we put their constrains together:

$$\beta_1 \leq \frac{c_1 b\epsilon^2}{3c_2\sigma^2},$$

where $c_1=\delta^p\leq\|\boldsymbol{v}_k\|_\infty\leq\left(c_\infty^2+\delta\right)^p=c_2=\mathcal{O}\left(c_\infty^{2p}\right)$. For η, it should satisfy

$$\eta \leq \frac{\beta_1 c_1}{2(1-\beta_1)L}\sqrt{\frac{c_1}{c_2}}\leq\frac{c_1 b\epsilon^2}{3c_2\sigma^2}\frac{c_1}{2L}\sqrt{\frac{c_1}{c_2}}=\frac{c_1^2 b\epsilon^2}{6c_2\sigma^2 L}\sqrt{\frac{c_1}{c_2}}.$$

where δ is often much smaller than one, and β_1 is very small. For T, we have

$$\begin{aligned}T\geq & \max\left(\frac{6c_2\Delta}{\eta\epsilon^2},\frac{3c_2\sigma^2}{c_1\beta_1 b\epsilon^2}\right)=\mathcal{O}\left(\max\left(\frac{6c_2\Delta}{\epsilon^2}\frac{6c_2\sigma^2 L}{c_1^2 b\epsilon^2}\sqrt{\frac{c_2}{c_1}},\frac{3c_2\sigma^2}{c_1 b\epsilon^2}\frac{3c_2\sigma^2}{c_1 b\epsilon^2}\right)\right)\\ = & \mathcal{O}\left(\max\left(\frac{36c_2^{2.5}\Delta\sigma^2 L}{c_1^{2.5}b\epsilon^4},\frac{9c_2^2\sigma^4}{c_1^2 b^2\epsilon^4}\right)\right)=\mathcal{O}\left(\max\left(\frac{36c_\infty^{2.5}\Delta\sigma^2 L}{\delta^{1.25}b\epsilon^4},\frac{9c_\infty^2\sigma^4}{\delta b^2\epsilon^4}\right)\right).\end{aligned}$$

Now we compute the stochastic gradient complexity. For T iterations, the complexity is

$$\mathcal{O}\left(Tb\right)=\mathcal{O}\left(\max\left(\frac{36c_2^{2.5}\Delta\sigma^2 L}{c_1^{2.5}\epsilon^4},\frac{9c_2^2\sigma^4}{c_1^2 b\epsilon^4}\right)\right)=\mathcal{O}\left(\max\left(\frac{36c_\infty^{2.5}\Delta\sigma^2 L}{\delta^{1.25}\epsilon^4},\frac{9c_\infty^2\sigma^4}{\delta b\epsilon^4}\right)\right).$$

The proof is completed. □

G.3 Proof of Corollary 1

Proof. First, we have

$$\|\nabla F(\boldsymbol{x}_k)\|_2=\|\nabla F_k(\boldsymbol{x}_k)-\lambda_k\boldsymbol{v}_k\odot\boldsymbol{x}_k\|_2\leq\|\nabla F_k(\boldsymbol{x}_k)\|_2+\lambda_k\|\boldsymbol{v}_k\odot\boldsymbol{x}_k\|_2\leq\|\nabla F_k(\boldsymbol{x}_k)\|_2+\lambda_k\rho'\|\boldsymbol{x}_k\|_\infty\cdot\|\nabla F(\boldsymbol{x}_k)\|_2.$$

Then we can obtain

$$\|\nabla F(\boldsymbol{x}_k)\|_2\leq\frac{1}{1-\lambda_k\rho'\|\boldsymbol{x}_k\|_\infty}\|\nabla F_k(\boldsymbol{x}_k)\|_2.$$

This completes the proof. □

G.4 Proof of Corollary 2

Proof. For Adam and ℓ_2-Adam, since our Theorem 2 still holds for the cases where 1) $\lambda_k=0$ or 2) the loss $F(\boldsymbol{x})$ is a combination of the loss and an ℓ_2-regularization, they also enjoy the complexity $\mathcal{O}\left(c_\infty^{2.5}\epsilon^{-4}\right)$. When the loss $F(\boldsymbol{x})$ is a combination of the loss and an ℓ_2-regularization, one can follow the proof of Theorem 2 to prove the results on ℓ_2-Adam. This completes the proof. □

G.5 Proof of Theorem 3

Proof. For brevity, we let $\boldsymbol{v}_k=\sqrt{\boldsymbol{n}_k+\delta}$. Since we have $\|\boldsymbol{m}_k\|_\infty\leq c_\infty$ and $\delta\leq\|\boldsymbol{n}_i+\delta\|_\infty\leq c_\infty^2+\delta$ in Lemma 1, for brevity, let

$$c_1:=\delta^{0.5}\leq\|\boldsymbol{v}_k\|_\infty\leq c_2:=(c_\infty^2+\delta)^{0.5}.$$

Also we define

$$\boldsymbol{u}_k:=\boldsymbol{m}_k+\lambda\boldsymbol{x}_k\odot\boldsymbol{v}_k,\qquad \boldsymbol{x}_{k+1}-\boldsymbol{x}_k=-\eta_k\frac{\boldsymbol{m}_k+\lambda\boldsymbol{x}_k\odot\boldsymbol{v}_k}{\boldsymbol{v}_k}=-\eta_k\frac{\boldsymbol{u}_k}{\boldsymbol{v}_k}.$$

Then by using the smoothness of $f(\boldsymbol{x};\boldsymbol{\zeta})$, we can obtain

$$
\begin{aligned}
&F_{k+1}(\boldsymbol{x}_{k+1})\\
\leq &F(\boldsymbol{x}_k)+\langle\nabla F(\boldsymbol{x}_k),\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\rangle+\frac{L}{2}\left\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\right\|^2+\frac{\lambda_{k+1}}{2}\left\|\boldsymbol{x}_{k+1}\right\|_{\boldsymbol{v}_{k+1}}^2\\
\overset{\textcircled{1}}{\leq} &F(\boldsymbol{x}_k)+\langle\nabla F(\boldsymbol{x}_k),\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\rangle+\frac{L}{2}\left\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\right\|^2+\frac{\lambda_{k+1}}{2(1-\mu)}\left\|\boldsymbol{x}_{k+1}\right\|_{\boldsymbol{v}_{k}}^2\\
\overset{\textcircled{2}}{\leq} &F(\boldsymbol{x}_k)+\frac{\lambda_k}{2}\left\|\boldsymbol{x}_k\right\|_{\boldsymbol{v}_k}^2+\langle\nabla F(\boldsymbol{x}_k)+\lambda\boldsymbol{x}_k\odot\boldsymbol{v}_k,\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\rangle+\frac{L}{2}\left\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\right\|^2+\frac{\lambda_k}{2}\left\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\right\|_{\boldsymbol{v}_k}^2\\
= &F_k(\boldsymbol{x}_k)+\langle\nabla F(\boldsymbol{x}_k)+\lambda\boldsymbol{x}_k\odot\boldsymbol{v}_k,\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\rangle+\frac{L}{2}\left\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\right\|^2+\frac{\lambda_k}{2}\left\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\right\|_{\boldsymbol{v}_k}^2\\
= &F_k(\boldsymbol{x}_k)-\eta_k\left\langle\nabla F(\boldsymbol{x}_k)+\lambda_k\boldsymbol{x}_k\odot\boldsymbol{v}_k,\frac{\boldsymbol{u}_k}{\boldsymbol{v}_k}\right\rangle+\frac{L\eta_k^2}{2}\left\|\frac{\boldsymbol{u}_k}{\boldsymbol{v}_k}\right\|^2+\frac{\lambda_k\eta_k^2}{2}\left\|\frac{\boldsymbol{u}_k}{\boldsymbol{v}_k}\right\|_{\boldsymbol{v}_k}^2\\
= &F_k(\boldsymbol{x}_k)+\frac{1}{2}\left\|\sqrt{\frac{\eta_k}{\boldsymbol{v}_k}}\left(\nabla F(\boldsymbol{x}_k)+\lambda_k\boldsymbol{x}_k\odot\boldsymbol{v}_k-\boldsymbol{u}_k\right)\right\|^2-\frac{1}{2}\left\|\sqrt{\frac{\eta_k}{\boldsymbol{v}_k}}\left(\nabla F(\boldsymbol{x}_k)+\lambda_k\boldsymbol{x}_k\odot\boldsymbol{v}_k\right)\right\|^2\\
&-\frac{1}{2}\left\|\sqrt{\frac{\eta_k}{\boldsymbol{v}_k}}\boldsymbol{u}_k\right\|^2+\frac{L\eta_k^2}{2}\left\|\frac{\boldsymbol{u}_k}{\boldsymbol{v}_k}\right\|^2+\frac{\lambda_k\eta_k^2}{2}\left\|\frac{\boldsymbol{u}_k}{\boldsymbol{v}_k}\right\|_{\boldsymbol{v}_k}^2\\
\leq &F_k(\boldsymbol{x}_k)+\frac{\eta_k}{2c_1}\left\|\nabla F(\boldsymbol{x}_k)-\boldsymbol{m}_k\right\|^2-\frac{\eta_k}{2c_2}\left\|\nabla F(\boldsymbol{x}_k)+\lambda\boldsymbol{x}_k\odot\boldsymbol{v}_k\right\|^2-\left[\frac{\eta_k}{2c_2}-\frac{L\eta_k^2}{2c_1^2}-\frac{\lambda_k\eta_k^2}{2c_1}\right]\left\|\boldsymbol{u}_k\right\|^2\\
\overset{\textcircled{3}}{\leq} &F_k(\boldsymbol{x}_k)+\frac{\eta_k}{2c_1}\left\|\nabla F(\boldsymbol{x}_k)-\boldsymbol{m}_k\right\|^2-\frac{\eta_k}{2c_2}\left\|\nabla F(\boldsymbol{x}_k)+\lambda_k\boldsymbol{x}_k\odot\boldsymbol{v}_k\right\|^2-\frac{\eta_k}{4c_2}\left\|\boldsymbol{u}_k\right\|^2
\end{aligned}
\tag{19}
$$

where ① holds since Lemma 1 proves $\left\|\frac{(\boldsymbol{n}_k+\delta)^{0.5}}{(\boldsymbol{n}_{k+1}+\delta)^{0.5}}\right\|_\infty\in[1-\mu,1+\mu]$ $(\forall p\in[0,1])$ in which $\mu=\frac{\beta_2 c_\infty^2}{\delta}$; ② holds because $\lambda_{k+1}=\frac{\lambda_{k+1}}{1-\mu}$ and

$$
\left\|\boldsymbol{x}_{k+1}\right\|_{\boldsymbol{v}_k}^2=\left\|\boldsymbol{x}_{k}\right\|_{\boldsymbol{v}_k}^2+2\left\langle\boldsymbol{x}_{k+1}-\boldsymbol{x}_k,\boldsymbol{x}_k\right\rangle_{\boldsymbol{v}_k}+\left\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\right\|_{\boldsymbol{v}_k}^2;
$$

③ holds, since we set $\eta_k\leq\frac{c_1^2}{2c_2(L+\lambda c_1)}$ such that $\frac{\eta_k}{4c_2}\geq\frac{L\eta_k^2}{2c_1^2}+\frac{\lambda\eta_k^2}{2c_1}$.

From Lemma 2, we have

$$
\begin{aligned}
\mathbb{E}\left[\left\|\boldsymbol{m}_k-\nabla F(\boldsymbol{x}_k)\right\|^2\right]\leq&(1-\beta_{1,k})\mathbb{E}\left[\left\|\boldsymbol{m}_{k-1}-\nabla F(\boldsymbol{x}_{k-1})\right\|^2\right]+\frac{(1-\beta_{1,k})^2L^2}{\beta_{1,k}}\mathbb{E}\left[\left\|\boldsymbol{x}_k-\boldsymbol{x}_{k-1}\right\|^2\right]+\frac{\beta_{1,k}^2\sigma^2}{b}\\
\leq&(1-\beta_{1,k})\mathbb{E}\left[\left\|\boldsymbol{m}_{k-1}-\nabla F(\boldsymbol{x}_{k-1})\right\|^2\right]+\frac{(1-\beta_{1,k})^2L^2\eta_k^2}{\beta_{1,k}c_1^2}\mathbb{E}\left[\left\|\boldsymbol{u}_k\right\|^2\right]+\frac{\beta_{1,k}^2\sigma^2}{b}
\end{aligned}
\tag{20}
$$

where we use $\boldsymbol{x}_k-\boldsymbol{x}_{k-1}=\eta_k\frac{\boldsymbol{u}_k}{\boldsymbol{v}_k}$.

Then we add Eqn. (19) and $\alpha\times$ (20) as follows:

$$
\begin{aligned}
&F_{k+1}(\boldsymbol{x}_{k+1})+\alpha\mathbb{E}\left[\left\|\boldsymbol{m}_{k+1}-\nabla F(\boldsymbol{x}_{k+1})\right\|^2\right]\\
\leq&F_k(\boldsymbol{x}_k)-\frac{\eta_k}{2c_2}\left\|\nabla F(\boldsymbol{x}_k)+\lambda_k\boldsymbol{x}_k\odot\boldsymbol{v}_k\right\|^2+\left[(1-\beta_{1,k})\alpha+\frac{\eta_k}{2c_1}\right]\mathbb{E}\left[\left\|\boldsymbol{m}_{k-1}-\nabla F(\boldsymbol{x}_{k-1})\right\|^2\right]\\
&-\left[\frac{\eta_k}{4c_2}-\frac{\alpha(1-\beta_{1,k})^2L^2\eta_k^2}{\beta_{1,k}c_1^2}\right]\mathbb{E}\left[\left\|\boldsymbol{u}_k\right\|^2\right]+\frac{\alpha\beta_{1,k}^2\sigma^2}{b}.
\end{aligned}
$$

Then by setting $\alpha=\frac{\eta_k}{2c_1\beta_{1,k}}$ and $G(\boldsymbol{x}_{k+1})=F_{k+1}(\boldsymbol{x}_{k+1})+\frac{\eta_k}{2c_1\beta_{1,k}}\mathbb{E}\left[\left\|\boldsymbol{m}_{k+1}-\nabla F(\boldsymbol{x}_{k+1})\right\|^2\right]$, we can obtain

$$
\begin{aligned}
&G(\boldsymbol{x}_{k+1})\\
\leq&G(\boldsymbol{x}_k)-\frac{\eta_k}{2c_2}\mathbb{E}\left\|\nabla F(\boldsymbol{x}_k)+\lambda_k\boldsymbol{x}_k\odot\boldsymbol{v}_k\right\|^2-\frac{\eta_k}{4c_2}\left[1-\frac{2c_2(1-\beta_{1,k})^2L^2\eta_k^2}{\beta_{1,k}^2c_1^3}\right]\mathbb{E}\left[\left\|\boldsymbol{u}_k\right\|^2\right]+\frac{\eta_k\beta_{1,k}\sigma^2}{2c_1b}\\
\overset{\textcircled{1}}{\leq}&G(\boldsymbol{x}_k)-\frac{\eta_k}{2c_2}\mathbb{E}\left[\left\|\nabla F(\boldsymbol{x}_k)+\lambda_k\boldsymbol{x}_k\odot\boldsymbol{v}_k\right\|^2\right]-\frac{\eta_k}{8c_2}\mathbb{E}\left[\left\|\boldsymbol{u}_k\right\|^2\right]+\frac{\eta_k\beta_{1,k}\sigma^2}{2c_1b},
\end{aligned}
$$

where ① holds since we set $\eta_k\leq\frac{\beta_{1,k}c_1}{2(1-\beta_{1,k})L}\sqrt{\frac{c_1}{c_2}}$ such that $\frac{2c_2(1-\beta_{1,k})^2L^2\eta_k^2}{\beta_{1,k}^2c_1^3}\leq\frac{1}{2}$.

Then summing the above inequality from $k=0$ to $k=T-1$ gives

$$\begin{aligned}\sum_{k=0}^{T-1}\frac{\eta_k}{\sum_{k=0}^{T-1}\eta_k}\mathbb{E}\left[\left\|\nabla F(\boldsymbol{x}_k)+\lambda_k\boldsymbol{x}_k\odot\boldsymbol{v}_k\right\|^2+\frac{1}{4}\left\|\boldsymbol{u}_k\right\|^2\right]\leq&\frac{2c_2}{\sum_{k=0}^{T-1}\eta_k}\left[G(\boldsymbol{x}_0)-G(\boldsymbol{x}_T)\right]+\frac{c_2\sum_{k=0}^{T-1}\eta_k\beta_{1,k}\sigma^2}{c_1 b\sum_{k=0}^{T-1}\eta_k}\\ \leq&\frac{2c_2\Delta}{\sum_{k=0}^{T-1}\eta_k}+\frac{c_2\eta_0\sigma^2}{c_1\beta_{1,0}b\sum_{k=0}^{T-1}\eta_k}+\frac{c_2\sigma^2\sum_{k=0}^{T-1}\eta_k\beta_{1,k}}{c_1 b\sum_{k=0}^{T-1}\eta_k},\end{aligned}\tag{21}$$

where

$$\begin{aligned}&G(\boldsymbol{x}_0)-G(\boldsymbol{x}_T)\\ =&F_0(\boldsymbol{x}_0)+\frac{\eta_0}{2c_1\beta_{1,0}}\mathbb{E}\left[\left\|\boldsymbol{m}_0-\nabla F(\boldsymbol{x}_0)\right\|^2\right]-F_T(\boldsymbol{x}_T)-\ \frac{\eta_0}{2c_1\beta_{1,0}}\mathbb{E}\left[\left\|\boldsymbol{m}_T-\nabla F(\boldsymbol{x}_T)\right\|^2\right]\\ =&F(\boldsymbol{x}_0)+\frac{\eta_0}{2c_1\beta_{1,0}}\mathbb{E}\left[\left\|\boldsymbol{m}_0-\nabla F(\boldsymbol{x}_0)\right\|^2\right]-F(\boldsymbol{x}_T)-\lambda_T\left\|\boldsymbol{x}_T\right\|_{\boldsymbol{v}_T}-\frac{\eta_0}{2c_1\beta_{1,0}}\mathbb{E}\left[\left\|\boldsymbol{m}_T-\nabla F(\boldsymbol{x}_T)\right\|^2\right]\\ \leq&F(\boldsymbol{x}_0)+\frac{\eta_0}{2c_1\beta_{1,0}}\mathbb{E}\left[\left\|\boldsymbol{m}_0-\nabla F(\boldsymbol{x}_0)\right\|^2\right]-F(\boldsymbol{x}_T)\\ \leq&\Delta+\frac{\eta_0}{2c_1\beta_{1,0}}\mathbb{E}\left[\left\|\boldsymbol{m}_0-\nabla F(\boldsymbol{x}_0)\right\|^2\right]\\ \leq&\Delta+\frac{\eta_0\sigma^2}{2c_1\beta_{1,0}b},\end{aligned}$$

where $\Delta=F(\boldsymbol{x}_0)-F(\boldsymbol{x}_*)$. Then by setting $\beta_{1,k}=\frac{\gamma_1}{\sqrt{k+1}}$ and $\eta_k=\gamma_2\beta_{1,k}$ where $\gamma_2=\frac{c_1^{1.5}}{2c_2^{0.5}L}\gamma_3$ and $\gamma_3=1$ to satisfy $\eta_k\leq\frac{\beta_{1,k}c_1}{2(1-\beta_{1,k})L}\sqrt{\frac{c_1}{c_2}}$, we have

$$\begin{aligned}&\sum_{k=0}^{T-1}\frac{\eta_k}{\sum_{k=0}^{T-1}\eta_k}\mathbb{E}\left[\left\|\nabla F(\boldsymbol{x}_k)+\lambda\boldsymbol{x}_k\odot\boldsymbol{v}_k\right\|^2+\frac{1}{4}\left\|\boldsymbol{u}_k\right\|^2\right]\\ \leq&\frac{2c_2\Delta}{\sum_{k=0}^{T-1}\eta_k}+\frac{c_2\eta_0\sigma^2}{c_1\beta_{1,0}b\sum_{k=0}^{T-1}\eta_k}+\frac{c_2\sigma^2\sum_{k=0}^{T-1}\eta_k\beta_{1,k}}{c_1 b\sum_{k=0}^{T-1}\eta_k}\\ \overset{①}{\leq}&\frac{c_2\Delta}{\gamma_1\gamma_2(\sqrt{T+1}-2)}+\frac{c_2\sigma^2}{2c_1b\gamma_1(\sqrt{T+1}-2)}+\frac{c_2\gamma_1\sigma^2\log(T)}{2c_1b(\sqrt{T+1}-2)}\\ =&\frac{2c_2^{1.5}\Delta L}{c_1^{1.5}\gamma_1\gamma_3(\sqrt{T+1}-2)}+\frac{c_2\sigma^2}{2c_1b\gamma_1(\sqrt{T+1}-2)}+\frac{c_2\gamma_1\sigma^2\log(T)}{2c_1b(\sqrt{T+1}-2)}\\ \overset{②}{\leq}&\frac{2c_2}{c_1\gamma_1(\sqrt{T+1}-2)}\left(\frac{c_2^{0.5}L\Delta}{c_1^{0.5}}+\sigma^2\right)+\frac{c_2\gamma_1\sigma^2\log(T)}{2c_1b(\sqrt{T+1}-2)}\\ \leq&\epsilon^2,\end{aligned}$$

where ① uses $\sum_{k=0}^{T-1}\beta_{1,k}\geq\int_2^{T+1}\frac{\gamma_1}{\sqrt{x}}dx=2\gamma_1(\sqrt{T+1}-2)$ and $\sum_{k=0}^{T-1}\eta_k\beta_{1,k}\leq\gamma_1^2\gamma_2\int_1^T\frac{1}{x}dx=\gamma_1^2\gamma_2\log(T)$, and ② holds by setting

$$\begin{aligned}T=&\mathcal{O}\left(\max\left(\frac{4c_2}{c_1\gamma_1\epsilon^4}\left(\frac{c_2^{0.5}L\Delta}{c_1^{0.5}}+\sigma^2\right),\frac{c_2\gamma_1\sigma^2\log\left(\frac{1}{\epsilon}\right)}{2c_1b\epsilon^4}\right)\right)\\ =&\mathcal{O}\left(\max\left(\frac{c_2}{c_1\gamma_1\epsilon^4}\left(\frac{c_2^{0.5}L\Delta}{c_1^{0.5}}+\sigma^2\right),\frac{c_2\gamma_1\sigma^2\log\left(\frac{1}{\epsilon}\right)}{c_1b\epsilon^4}\right)\right)\\ =&\mathcal{O}\left(\max\left(\frac{c_2}{c_1\epsilon^4\max\left(1,\frac{c_2^{0.25}L^{0.5}\Delta^{0.5}}{c_1^{0.25}\sigma}\right)}\left(\frac{c_2^{0.5}L\Delta}{c_1^{0.5}}+\sigma^2\right),\frac{c_2\sigma^2\log\left(\frac{1}{\epsilon}\right)\max\left(1,\frac{c_2^{0.25}L^{0.5}\Delta^{0.5}}{c_1^{0.25}\sigma}\right)}{c_1b\epsilon^4}\right)\right)\\ =&\mathcal{O}\left(\max\left(\frac{c_2\sigma^2}{c_1b\epsilon^4}\log\left(\frac{1}{\epsilon}\right),\frac{c_2^{1.25}L^{0.5}\Delta^{0.5}\sigma}{c_1^{1.25}b\epsilon^4}\log\left(\frac{1}{\epsilon}\right)\right)\right)\end{aligned}$$

where we set $\gamma_1=\max\left(1,\frac{c_2^{0.25}L^{0.5}\Delta^{0.5}}{c_1^{0.25}\sigma}\right)$.

For all hyper-parameters, we put their constrains together:

$$\beta_{1,k}=\frac{\gamma}{\sqrt{k+1}},\quad\eta_k=\frac{c_1^{1.5}}{2c_2^{0.5}L}\beta_{1,k}=\frac{\gamma c_1^{1.5}}{2c_2^{0.5}L\sqrt{k+1}}=\frac{\gamma\delta^{0.75}}{2(c_\infty^2+\delta)^{0.25}L\sqrt{k+1}},$$

where $\gamma = \max\left(1, \frac{c_2^{0.25} L^{0.5} \Delta^{0.5}}{c_1^{0.25}\sigma}\right)$, $c_1 = \delta^{0.5} \leq \|\boldsymbol{v}_k\|_\infty \leq \left(c_\infty^2 + \delta\right)^{0.5} = c_2$. Then by setting minibatch size as one, one can easily compute the stochastic gradient complexity

$$\begin{aligned}\mathcal{O}\left(Tb\right) =& \mathcal{O}\left(\max\left(\frac{c_2\sigma^2}{c_1\epsilon^4}\log\left(\frac{1}{\epsilon}\right), \frac{c_2^{1.25}L^{0.5}\Delta^{0.5}\sigma}{c_1^{1.25}\epsilon^4}\log\left(\frac{1}{\epsilon}\right)\right)\right)\\ =& \mathcal{O}\left(\max\left(\frac{c_\infty\sigma^2}{\delta^{0.5}\epsilon^4}\log\left(\frac{1}{\epsilon}\right), \frac{c_\infty^{1.25}L^{0.5}\Delta^{0.5}\sigma}{\delta^{0.625}\epsilon^4}\log\left(\frac{1}{\epsilon}\right)\right)\right).\end{aligned}$$

The above result directly bounds

$$\begin{aligned}\sum_{k=0}^{T-1}\frac{\eta_k}{\sum_{k=0}^{T-1}\eta_k}\|\boldsymbol{v}_k\odot(\boldsymbol{x}_k-\boldsymbol{x}_{k+1})\|^2 =& \sum_{k=0}^{T-1}\frac{\eta_k^3}{\sum_{k=0}^{T-1}\eta_k}\|\boldsymbol{m}_k+\lambda_k\boldsymbol{x}_k\odot\boldsymbol{v}_k\|^2\\ =& \max_k \eta_k^2\left(\sum_{k=0}^{T-1}\frac{\eta_k}{\sum_{k=0}^{T-1}\eta_k}\|\boldsymbol{m}_k+\lambda_k\boldsymbol{x}_k\odot\boldsymbol{v}_k\|^2\right)\\ \leq& \eta_1^2\sum_{k=0}^{T-1}\frac{\eta_k}{\sum_{k=0}^{T-1}\eta_k}\|\boldsymbol{u}_k\|^2\\ \leq& 4\eta_1^2\epsilon^2.\end{aligned}$$

Besides, we have

$$\begin{aligned}\sum_{k=0}^{T-1}\frac{\eta_k}{\sum_{k=0}^{T-1}\eta_k}\mathbb{E}\left[\|\boldsymbol{m}_k-\nabla F(\boldsymbol{x}_k)\|^2\right] \leq& \sum_{k=0}^{T-1}\frac{\eta_k}{\sum_{k=0}^{T-1}\eta_k}\mathbb{E}\left[\|\boldsymbol{m}_k+\lambda_k\boldsymbol{x}_k\odot\boldsymbol{v}_k-\nabla F(\boldsymbol{x}_k)-\lambda_k\boldsymbol{x}_k\odot\boldsymbol{v}_k\|^2\right]\\ \leq& 2\sum_{k=0}^{T-1}\frac{\eta_k}{\sum_{k=0}^{T-1}\eta_k}\mathbb{E}\left[\|\boldsymbol{m}_k+\lambda_k\boldsymbol{x}_k\odot\boldsymbol{v}_k\|^2+\|\nabla F(\boldsymbol{x}_k)-\lambda_k\boldsymbol{x}_k\odot\boldsymbol{v}_k\|^2\right]\\ =& 2\sum_{k=0}^{T-1}\frac{\eta_k}{\sum_{k=0}^{T-1}\eta_k}\mathbb{E}\left[\|\boldsymbol{m}_k+\lambda_k\boldsymbol{x}_k\odot\boldsymbol{v}_k\|^2+\|\boldsymbol{u}_k\|^2\right]\\ \leq& 2\left[\epsilon^2+\frac{3}{4}\times 4\epsilon^2\right]\leq 8\epsilon^2.\end{aligned}$$

The proof is completed. □

G.6 Proof of Theorem 4

Proof. **Step 1. Results under constant learning rate.** Here we first consider the conventional one stage training. Firstly, we borrow the results in Eqn. (18) in Appendix G.2 (proofs for Theorem 2), if $\eta \leq \frac{\beta_1 c_1}{2(1-\beta_1)L}\sqrt{\frac{c_1}{c_2}}$, we have

$$\begin{aligned}\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}\left[\|\nabla F_k(\boldsymbol{x}_k)\|^2+\frac{1}{4}\|\boldsymbol{u}_k\|^2\right] =& \frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}\left[\|\nabla F(\boldsymbol{x}_k)+\lambda_k\boldsymbol{x}_k\odot\boldsymbol{v}_k\|^2+\frac{1}{4}\|\boldsymbol{u}_k\|^2\right]\\ \leq& \frac{2c_2}{\eta T}\left[G(\boldsymbol{x}_0)-G(\boldsymbol{x}_T)\right]+\frac{c_2\beta_1\sigma^2}{c_1 b}\\ \leq& \frac{2c_2\Delta}{\eta T}+\frac{c_2}{c_1\beta_1 T}\mathbb{E}\left[\|\boldsymbol{m}_0-\nabla F(\boldsymbol{x}_0)\|^2\right]+\frac{c_2\beta_1\sigma^2}{c_1 b},\end{aligned}\tag{22}$$

where $\Delta = F(\boldsymbol{x}_0) - F(\boldsymbol{x}_*)$. Then assume at the $(k-1)$-th stage, we already have

$$\mathbb{E}\left[F_{k-1}(\boldsymbol{x}_{k-1})-F_{k-1}(\boldsymbol{x}_*)\right]\leq\epsilon_{k-1},\qquad \mathbb{E}\left[\|\boldsymbol{m}_{k-1}-\nabla F(\boldsymbol{x}_{k-1})\|^2\right]\leq\mu\epsilon_{k-1}.$$

Then at the k-th stage with T_k iteration, by using Eqn. (25), we have

$$\frac{1}{T_k}\sum_{k=0}^{T_k-1}\mathbb{E}\left[\|\nabla F_k(\boldsymbol{x}_k)\|^2+\frac{1}{4}\|\boldsymbol{u}_k\|^2\right]\leq\frac{2c_2\epsilon_{k-1}}{\eta T}+\frac{c_2\mu\epsilon_{k-1}}{c_1\beta_1 T}+\frac{c_2\beta_1\sigma^2}{c_1 b}\leq\frac{\mu\epsilon_k}{8},\tag{23}$$

where we set $\beta_k \leq \frac{\mu c_1 b\epsilon_k}{24c_2\sigma^2}$ and $T_k \geq \max\left(\frac{16c_2\epsilon_{k-1}}{\mu\eta_k\epsilon_k}, \frac{8c_2\epsilon_{k-1}}{c_1\beta_1\epsilon_k}\right)$. Considering $\eta_k \leq \frac{\beta_1 c_1}{2(1-\beta_1)L}\sqrt{\frac{c_1}{c_2}}$, then we have

$$\begin{aligned}&\beta_1\leq\frac{c_1\mu b\epsilon_k}{24c_2\sigma^2},\qquad \eta\leq\frac{\beta_1 c_1}{2(1-\beta_1)L}\sqrt{\frac{c_1}{c_2}}=\mathcal{O}\left(\frac{c_1\mu b\epsilon_k}{24c_2\sigma^2}\cdot\frac{c_1}{2L}\sqrt{\frac{c_1}{c_2}}\right)=\mathcal{O}\left(\frac{\mu c_1^{2.5}b\epsilon_k}{48c_2^{1.5}L\sigma^2}\right),\\ &T_k\geq\max\left(\frac{16c_2\epsilon_{k-1}}{\mu\eta_k\epsilon_k},\frac{8c_2\epsilon_{k-1}}{c_1\beta_1\epsilon_k}\right)=\mathcal{O}\left(\max\left(\frac{c_2^{2.5}L\sigma^2\epsilon_{k-1}}{\mu^2c_1^{2.5}b\epsilon_k^2},\frac{c_2^2\sigma^2\epsilon_{k-1}}{\mu c_1^2 b\epsilon_k^2}\right)\right)=\mathcal{O}\left(\max\left(\frac{c_2^{2.5}L\sigma^2}{\mu^2c_1^{2.5}b\epsilon_k},\frac{c_2^2\sigma^2}{\mu c_1^2 b\epsilon_k}\right)\right),\end{aligned}$$

where the last inequality uses $\epsilon_k = \frac{\epsilon_0}{2^k} = \frac{1}{2}\epsilon_{k-1}$. Then by using the PŁcondition, we have

$$\mathbb{E}\left[F_k(\boldsymbol{x}_k) - F_k(\boldsymbol{x}_*)\right] \leq \frac{1}{T_k}\sum_{i=0}^{T_k-1}\mathbb{E}\left[F_i(\boldsymbol{x}_i) - F_i(\boldsymbol{x}_*)\right] \leq \frac{1}{2\mu T_k}\sum_{i=0}^{T_k-1}\mathbb{E}\left[\|\nabla F_i(\boldsymbol{x}_i)\|^2\right] \leq \epsilon_k,$$

and

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{m}_k - \nabla F(\boldsymbol{x}_k)\|^2\right] = \frac{1}{T_k}\sum_{i=0}^{T_k-1}\mathbb{E}\left[\|\boldsymbol{m}_i - \nabla F(\boldsymbol{x}_i)\|^2\right] =& \frac{1}{T_k}\sum_{i=0}^{T_k-1}\mathbb{E}\left[\|\boldsymbol{m}_i + \lambda_i\boldsymbol{x}_i\odot\boldsymbol{v}_i - \nabla F(\boldsymbol{x}_i) - \lambda_i\boldsymbol{x}_i\odot\boldsymbol{v}_i\|^2\right]\\ =&\frac{2}{T_k}\sum_{i=0}^{T_k-1}\mathbb{E}\left[\|\boldsymbol{m}_i + \lambda_i\boldsymbol{x}_i\odot\boldsymbol{v}_i\|^2 + \|\nabla F(\boldsymbol{x}_i) + \lambda_i\boldsymbol{x}_i\odot\boldsymbol{v}_i\|^2\right]\\ =&\frac{2}{T_k}\sum_{i=0}^{T_k-1}\mathbb{E}\left[\|\boldsymbol{m}_i + \lambda_i\boldsymbol{x}_i\odot\boldsymbol{v}_i\|^2 + \|\boldsymbol{u}_i\|^2\right]\\ \leq& 2\mu\left[\frac{1}{8}\epsilon_k + \frac{3}{4}\times 4\times\frac{1}{8}\epsilon_k\right] \leq \mu\epsilon_k.\end{aligned}$$

This means that we only need the stochastic gradient complexity for the k-th stage:

$$\mathcal{O}\left(T_k b\right) \leq \mathcal{O}\left(\max\left(\frac{c_2^{2.5}L\sigma^2}{\mu^2 c_1^{2.5}\epsilon_k}, \frac{c_2^2\sigma^2}{\mu c_1^2\epsilon_k}\right)\right).$$

Finally, to achieve ϵ-accuracy solution, we only need to run at most K stages which should satisfy

$$\epsilon_K = \frac{\epsilon_0}{2^K} \leq \epsilon,$$

where $\epsilon_0 = \Delta$. So it means that K should obey

$$K \geq \log_2\left(\frac{1}{\epsilon}\right).$$

In this way, we can compute the total computational complexity as follows:

$$\begin{aligned}\sum_{k=1}^{K}\mathbb{E}\left[T_k b\right] =&\mathbb{E}\left[\sum_{k=1}^{K}\mathcal{O}\left(\max\left(\frac{c_2^{2.5}L\sigma^2}{\mu^2 c_1^{2.5}\epsilon_k}, \frac{c_2^2\sigma^2}{\mu c_1^2\epsilon_k}\right)\right)\right] = \mathcal{O}\left(\max\left(\frac{c_2^{2.5}L\sigma^2}{\mu^2 c_1^{2.5}}, \frac{c_2^2\sigma^2}{\mu c_1^2}\right)\mathbb{E}\left[\sum_{k=1}^{K}\frac{1}{\epsilon_k}\right]\right)\\ =&\mathcal{O}\left(\max\left(\frac{c_2^{2.5}L\sigma^2}{\mu^2 c_1^{2.5}\epsilon}, \frac{c_2^2\sigma^2}{\mu c_1^2\epsilon}\right)\right) = \mathcal{O}\big(\max\big(\frac{c_\infty^{2.5}L\sigma^2}{\mu^2\delta^{1.25}\epsilon}, \frac{c_\infty^2\sigma^2}{\mu\delta\epsilon}\big)\big).\end{aligned}$$

Step 2. Results under decaying learning rate. Firstly, we borrow the results in Eqn. (21) in Appendix G.5 (proofs for Theorem 3), we have

$$\begin{aligned}\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}\left[\|\nabla F_k(\boldsymbol{x}_k)\|^2 + \frac{1}{4}\|\boldsymbol{u}_k\|^2\right] =& \sum_{k=0}^{T-1}\frac{\eta_k}{\sum_{k=0}^{T-1}\eta_k}\mathbb{E}\left[\|\nabla F(\boldsymbol{x}_k) + \lambda_k\boldsymbol{x}_k\odot\boldsymbol{v}_k\|^2 + \frac{1}{4}\|\boldsymbol{u}_k\|^2\right]\\ \leq&\frac{2c_2}{\sum_{k=0}^{T-1}\eta_k}\left[G(\boldsymbol{x}_0) - G(\boldsymbol{x}_T)\right] + \frac{c_2\sum_{k=0}^{T-1}\eta_k\beta_{1,k}\sigma^2}{c_1 b\sum_{k=0}^{T-1}\eta_k}\\ \leq&\frac{2c_2}{\sum_{k=0}^{T-1}\eta_k}\left[\Delta + \frac{\eta_0}{2c_1\beta_{1,0}}\mathbb{E}\left[\|\boldsymbol{m}_0 - \nabla F(\boldsymbol{x}_0)\|^2\right]\right] + \frac{c_2\sigma^2\sum_{k=0}^{T-1}\eta_k\beta_{1,k}}{c_1 b\sum_{k=0}^{T-1}\eta_k}\\ \leq&\frac{2c_2\Delta}{\sum_{k=0}^{T-1}\eta_k} + \frac{\eta_0 c_2}{c_1\beta_{1,0}\sum_{k=0}^{T-1}\eta_k}\mathbb{E}\left[\|\boldsymbol{m}_0 - \nabla F(\boldsymbol{x}_0)\|^2\right] + \frac{c_2\sigma^2\sum_{k=0}^{T-1}\eta_k\beta_{1,k}}{c_1 b\sum_{k=0}^{T-1}\eta_k}\end{aligned} \tag{24}$$

where $\Delta = F(\boldsymbol{x}_0) - F(\boldsymbol{x}_*)$. Then assume at the $(k-1)$-th stage, we already have

$$\mathbb{E}\left[F_{k-1}(\boldsymbol{x}_{k-1}) - F_{k-1}(\boldsymbol{x}_*)\right] \leq \epsilon_{k-1}, \qquad \mathbb{E}\left[\|\boldsymbol{m}_{k-1} - \nabla F(\boldsymbol{x}_{k-1})\|^2\right] \leq \mu\epsilon_{k-1}.$$

Then at the k-th stage with T_k iteration, by using Eqn. (24), we have

$$\frac{1}{T_k}\sum_{k=0}^{T_k-1}\mathbb{E}\left[\|\nabla F_k(\boldsymbol{x}_k)\|^2 + \frac{1}{4}\|\boldsymbol{u}_k\|^2\right] \leq \frac{2c_2\epsilon_{k-1}}{\sum_{k=0}^{T-1}\eta_k} + \frac{\mu\eta_0 c_2\epsilon_{k-1}}{c_1\beta_{1,0}\sum_{k=0}^{T-1}\eta_k} + \frac{c_2\sigma^2\sum_{k=0}^{T-1}\eta_k\beta_{1,k}}{c_1 b\sum_{k=0}^{T-1}\eta_k}. \tag{25}$$

Then, following the proof in Appendix G.5, we set $\beta_{1,k} = \frac{\gamma_1}{\sqrt{k+1}}$ and $\eta_k = \gamma_2 \beta_{1,k}$ where $\gamma_2 = \frac{c_1^{1.5}}{2c_2^{0.5}L}\gamma_3$ and $\gamma_3 = 1$ to satisfy $\eta_k \leq \frac{\beta_{1,k}c_1}{2(1-\beta_{1,k})L}\sqrt{\frac{c_1}{c_2}}$, then we have

$$
\begin{aligned}
\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}\left[\left\|\nabla F_k(\boldsymbol{x}_k)\right\|^2 + \frac{1}{4}\left\|\boldsymbol{u}_k\right\|^2\right] \leq & \frac{2c_2\epsilon_{k-1}}{\gamma_2\sum_{k=0}^{T-1}\beta_{1,k}} + \frac{\mu\eta_0 c_2\epsilon_{k-1}}{c_1\beta_{1,0}\gamma_2\sum_{k=0}^{T-1}\beta_{1,k}} + \frac{c_2\sigma^2\sum_{k=0}^{T-1}\eta_k\beta_{1,k}}{c_1 b\gamma_2\sum_{k=0}^{T-1}\beta_{1,k}} \\
\overset{\text{①}}{\leq} & \frac{c_2\epsilon_{k-1}}{\gamma_2\gamma_1(\sqrt{T+1}-2)} + \frac{\mu\eta_0 c_2\epsilon_{k-1}}{2c_1\beta_{1,0}\gamma_2\gamma_1(\sqrt{T+1}-2)} + \frac{c_2\sigma^2\gamma_1\log(T)}{2c_1 b(\sqrt{T+1}-2)} \\
= & \frac{2c_2^{1.5}L\epsilon_{k-1}}{\gamma_1 c_1^{1.5}(\sqrt{T+1}-2)} + \frac{\mu c_2\epsilon_{k-1}}{2c_1\gamma_1(\sqrt{T+1}-2)} + \frac{c_2\sigma^2\gamma_1\log(T)}{2c_1 b(\sqrt{T+1}-2)} \\
\overset{\text{②}}{\leq} & \frac{\mu}{8}\epsilon_k,
\end{aligned}
\tag{26}
$$

where ① uses $\sum_{k=0}^{T-1}\beta_{1,k} \geq \int_2^{T+1}\frac{\gamma_1}{\sqrt{x}}dx = 2\gamma_1(\sqrt{T+1}-2)$ and $\sum_{k=0}^{T-1}\eta_k\beta_{1,k} \leq \gamma_1^2\gamma_2\int_1^T\frac{1}{x}dx = \gamma_1^2\gamma_2\log(T)$, and ② holds by setting

$$
\gamma_1 = \frac{c_2^{0.25}L^{0.5}b^{0.5}\epsilon_k^{0.5}}{c_1^{0.25}\sigma}, \quad T_k = \mathcal{O}\left(\max\left(\frac{c_2^3L^2}{c_1^3\gamma_1^2\mu^2}, \frac{c_2^2\gamma_1^2\sigma^4\log(\frac{1}{\epsilon_k})}{\mu^2c_1^2b^2\epsilon_k^2}\right)\right) = \mathcal{O}\left(\frac{c_2^{2.5}L\sigma^2\log\left(\frac{1}{\epsilon_k}\right)}{\mu^2c_1^{2.5}b\epsilon_k}\right).
$$

This means that by setting

$$
\gamma_1 = \frac{c_2^{0.25}L^{0.5}b^{0.5}\epsilon_k^{0.5}}{c_1^{0.25}\sigma}, \quad \beta_{1,k} = \frac{\gamma_1}{\sqrt{k+1}}, \quad \eta_k = \frac{c_1^{1.5}}{2c_2^{0.5}L}\beta_{1,k}, \quad T_k = \mathcal{O}\left(\frac{c_2^{2.5}L\sigma^2\log\left(\frac{1}{\epsilon_k}\right)}{\mu^2c_1^{2.5}b\epsilon_k}\right),
$$

we have

$$
\frac{1}{T}\sum_{k=0}^{T-1}\mathbb{E}\left[\left\|\nabla F_k(\boldsymbol{x}_k)\right\|^2 + \frac{1}{4}\left\|\boldsymbol{u}_k\right\|^2\right] \leq \frac{\mu}{8}\epsilon_k. \tag{27}
$$

By using PŁcondition, we have

$$
\mathbb{E}\left[F_k(\boldsymbol{x}_k) - F_k(\boldsymbol{x}_*)\right] \leq \frac{1}{T_k}\sum_{i=0}^{T_k-1}\mathbb{E}\left[F_i(\boldsymbol{x}_i) - F_i(\boldsymbol{x}_*)\right] \leq \frac{1}{2\mu T_k}\sum_{i=0}^{T_k-1}\mathbb{E}\left[\|\nabla F_i(\boldsymbol{x}_i)\|^2\right] \leq \epsilon_k,
$$

and

$$
\begin{aligned}
\mathbb{E}\left[\|\boldsymbol{m}_k - \nabla F(\boldsymbol{x}_k)\|^2\right] = \frac{1}{T_k}\sum_{i=0}^{T_k-1}\mathbb{E}\left[\|\boldsymbol{m}_i - \nabla F(\boldsymbol{x}_i)\|^2\right] = & \frac{1}{T_k}\sum_{i=0}^{T_k-1}\mathbb{E}\left[\|\boldsymbol{m}_i + \lambda_i\boldsymbol{x}_i\odot\boldsymbol{v}_i - \nabla F(\boldsymbol{x}_i) - \lambda_i\boldsymbol{x}_i\odot\boldsymbol{v}_i\|^2\right] \\
= & \frac{2}{T_k}\sum_{i=0}^{T_k-1}\mathbb{E}\left[\|\boldsymbol{m}_i + \lambda_i\boldsymbol{x}_i\odot\boldsymbol{v}_i\|^2 + \|\nabla F(\boldsymbol{x}_i) + \lambda_i\boldsymbol{x}_i\odot\boldsymbol{v}_i\|^2\right] \\
= & \frac{2}{T_k}\sum_{i=0}^{T_k-1}\mathbb{E}\left[\|\boldsymbol{m}_i + \lambda_i\boldsymbol{x}_i\odot\boldsymbol{v}_i\|^2 + \|\boldsymbol{u}_i\|^2\right] \\
\leq & 2\mu\left[\frac{1}{8}\epsilon_k + \frac{3}{4}\times 4\times\frac{1}{8}\epsilon_k\right] \leq \mu\epsilon_k.
\end{aligned}
$$

This means that we only need the stochastic gradient complexity for the k-th stage:

$$
\mathcal{O}\left(T_k b\right) \leq \mathcal{O}\left(\frac{c_2^{2.5}L\sigma^2\log\left(\frac{1}{\epsilon_k}\right)}{\mu^2c_1^{2.5}\epsilon_k}\right).
$$

Finally, to achieve ϵ-accuracy solution, we only need to run at most K stages which should satisfy

$$
\epsilon_K = \frac{\epsilon_0}{2^K} \leq \epsilon,
$$

where $\epsilon_0 = \Delta$. So it means that K should obey

$$
K \geq \log_2\left(\frac{1}{\epsilon}\right).
$$

In this way, we can compute the total computational complexity as follows:

$$\sum_{k=1}^{K}\mathbb{E}\left[T_k b\right]=\mathbb{E}\left[\sum_{k=1}^{K}\mathcal{O}\left(\frac{c_2^{2.5}L\sigma^2\log\left(\frac{1}{\epsilon_k}\right)}{\mu^2 c_1^{2.5}\epsilon_k}\right)\right]=\mathcal{O}\left(\frac{c_2^{2.5}L\sigma^2}{\mu^2 c_1^{2.5}}\mathbb{E}\left[\sum_{k=1}^{K}\frac{\log\left(\frac{1}{\epsilon_k}\right)}{\epsilon_k}\right]\right)=\mathcal{O}\left(\frac{c_2^{2.5}L\sigma^2}{\mu^2 c_1^{2.5}\epsilon}\right)=\mathcal{O}(\frac{c_\infty^{2.5}L\sigma^2}{\mu^2\delta^{1.25}\epsilon}).$$

where

$$\mathbb{E}\left[\sum_{k=1}^{K}\frac{1}{\epsilon_k}\log\left(\frac{1}{\epsilon_k}\right)\right]\overset{①}{=}\mathbb{E}\left[\sum_{k=1}^{K}\frac{2^k}{\epsilon_0}\log\left(\frac{2^k}{\epsilon_0}\right)\right]=\mathcal{O}\left(\mathbb{E}\left[\sum_{k=1}^{K}\frac{k\cdot 2^k}{\epsilon_0}\right]\right)=\mathcal{O}\left(\mathbb{E}\left[S_K\right]\right)\overset{②}{=}\mathcal{O}\left(\frac{1}{\epsilon}\right)$$

where we use $\epsilon_k=\frac{\epsilon_0}{2^k}$ in ①. For ②, we can compute

$$S_K-2S_{K-1}=\sum_{k=1}^{K}\frac{k\cdot 2^k}{\epsilon_0}-2\sum_{k=1}^{K-1}\frac{k\cdot 2^k}{\epsilon_0}=\frac{2}{\epsilon_0}.$$

Consider $S_1=\frac{2}{\epsilon_0^2}$, then we have

$$S_K+\frac{2}{\epsilon_0}=2\left(S_{K-1}+\frac{2}{\epsilon_0}\right)=2^{K-1}\left(S_1+\frac{2}{\epsilon_0}\right)=\frac{2^{K+2}}{\epsilon_0}=\frac{4}{\epsilon},$$

where we use $\frac{\epsilon_0}{2^K}=\epsilon$. The proof is completed. □

Appendix H
Proof of Results in Sec. 5

To begin with, we first give one useful lemma to prove our generalization error bound.

Lemma 4. *(PAC-Bayesian generalization bound) [21] For any $\tau\in(0,1)$, the expected risk for the posterior hypothesis of an algorithm over a training dataset $\mathcal{D}_{tr}\sim\mathcal{D}$ with n samples holds with at least probability $1-\tau$:*

$$\mathbb{E}_{\boldsymbol{\xi}\sim\mathcal{D},\boldsymbol{x}\sim\mathcal{P}}[f(\boldsymbol{x},\boldsymbol{\xi})]-\mathbb{E}_{\boldsymbol{\xi}\in\mathcal{D}_{tr},\boldsymbol{x}\sim\mathcal{P}}[f(\boldsymbol{x},\boldsymbol{\xi})]\leq 4\sqrt{\frac{1}{n}\left(KL(\mathcal{P}\|\mathcal{P}_{pre})+\ln\left(\frac{2n}{\tau}\right)\right)},$$

where $KL(\mathcal{P}\|\mathcal{P}_{pre})$ denotes the Kullback-Leibler divergence from prior $\mathcal{P}_{pre}$ to posterior $\mathcal{P}$.

H.1 Proof of Lemma 5

Proof. Based on the assumptions in Lemma 5, we can write the SDE equations as follows:

$$\begin{aligned}\mathsf{d}\boldsymbol{x}_t=&-\boldsymbol{Q}_t\nabla F(\boldsymbol{x}_t)\mathsf{d}t-\lambda\boldsymbol{x}_t\mathsf{d}t+\boldsymbol{Q}_t\left(2\boldsymbol{\Sigma}_t\right)^{\frac{1}{2}}\mathsf{d}\boldsymbol{\zeta}_t\\=&-\boldsymbol{Q}_t\boldsymbol{H}_*\boldsymbol{x}_t\mathsf{d}t-\lambda\boldsymbol{x}_t\mathsf{d}t+\boldsymbol{Q}_t\left(2\boldsymbol{\Sigma}_t\right)^{\frac{1}{2}}\mathsf{d}\boldsymbol{\zeta}_t\\=&-(\boldsymbol{Q}\boldsymbol{H}_*+\lambda\boldsymbol{I})\boldsymbol{x}_t\mathsf{d}t+\sqrt{\frac{\eta}{b}}Q\boldsymbol{H}_*^{\frac{1}{2}}\mathsf{d}\boldsymbol{\zeta}_t,\end{aligned}$$

where $\mathsf{d}\boldsymbol{\zeta}_t\sim\mathcal{N}(0,\boldsymbol{I}\mathsf{d}t)$, $\boldsymbol{\Sigma}_t\approx\frac{\eta}{2B}\boldsymbol{H}_*$; $\boldsymbol{Q}_t=\boldsymbol{Q}:=\operatorname{diag}\left([\boldsymbol{H}_{*(11)}^{-\frac{1}{2}},\boldsymbol{H}_{*(22)}^{-\frac{1}{2}},\cdots,\boldsymbol{H}_{*(dd)}^{-\frac{1}{2}}]\right)$. Then for this Ornstein–Uhlenbeck process, we can compute its closed form solution as follows:

$$\boldsymbol{x}_t=\exp\left(-(\boldsymbol{Q}\boldsymbol{H}_*+\lambda\boldsymbol{I})t\right)\boldsymbol{x}_0+\sqrt{\frac{\eta}{b}}\int_0^t\exp\left(-(\boldsymbol{Q}\boldsymbol{H}_*+\lambda\boldsymbol{I})(t-t')\right)\boldsymbol{Q}\boldsymbol{H}_*^{\frac{1}{2}}\mathsf{d}\boldsymbol{\zeta}_{t'}.$$

Let $\boldsymbol{M}=\mathbb{E}\left[\boldsymbol{x}_t\boldsymbol{x}_t^\top\right]$. In this way, we follow [19] (see their Appendix B) and can further compute the algebraic relation for the stationary covariance of the multivariate Ornstein–Uhlenbeck process as follows:

$$\begin{aligned}&(\boldsymbol{Q}\boldsymbol{H}_*+\lambda\boldsymbol{I})\boldsymbol{M}+\boldsymbol{M}^\top(\boldsymbol{Q}\boldsymbol{H}_*+\lambda\boldsymbol{I})^\top\\=&\frac{\eta}{b}\int_{-\infty}^{t}(\boldsymbol{Q}\boldsymbol{H}_*+\lambda\boldsymbol{I})\exp\left(-(\boldsymbol{Q}\boldsymbol{H}_*+\lambda\boldsymbol{I})(t-t')\right)\boldsymbol{Q}\boldsymbol{H}_*^{\frac{1}{2}}(\boldsymbol{Q}\boldsymbol{H}_*^{\frac{1}{2}})^\top\exp\left(-(\boldsymbol{Q}\boldsymbol{H}_*+\lambda\boldsymbol{I})^\top(t-t')\right)\mathsf{d}t'\\&+\frac{\eta}{b}\int_{-\infty}^{t}\exp\left(-(\boldsymbol{Q}\boldsymbol{H}_*+\lambda\boldsymbol{I})(t-t')\right)\boldsymbol{Q}\boldsymbol{H}_*^{\frac{1}{2}}(\boldsymbol{Q}\boldsymbol{H}_*^{\frac{1}{2}})^\top\exp\left(-(\boldsymbol{Q}\boldsymbol{H}_*+\lambda\boldsymbol{I})^\top(t-t')\right)\mathsf{d}t'(\boldsymbol{Q}\boldsymbol{H}_*+\lambda\boldsymbol{I})\\=&\frac{\eta}{b}\int_{-\infty}^{t}\frac{\mathsf{d}}{\mathsf{d}t'}\left(\exp\left(-(\boldsymbol{Q}\boldsymbol{H}_*+\lambda\boldsymbol{I})(t-t')\right)\boldsymbol{Q}\boldsymbol{H}_*^{\frac{1}{2}}(\boldsymbol{Q}\boldsymbol{H}_*^{\frac{1}{2}})^\top\exp\left(-(\boldsymbol{Q}\boldsymbol{H}_*+\lambda\boldsymbol{I})^\top(t-t')\right)\right)\\=&\frac{\eta}{b}\boldsymbol{Q}\boldsymbol{H}_*^{\frac{1}{2}}(\boldsymbol{Q}\boldsymbol{H}_*^{\frac{1}{2}})^\top=\frac{\eta}{b}\boldsymbol{Q}\boldsymbol{H}_*\boldsymbol{Q},\end{aligned}$$

where we use the lower limits of the integral vanishes by the positivity of the eigenvalues of $\boldsymbol{Q}\boldsymbol{H}_*+\lambda\boldsymbol{I}$. Therefore, we know

$$\boldsymbol{M}_{\text{AdamW}}=\frac{\eta}{2b}(\boldsymbol{Q}\boldsymbol{H}_*+\lambda\boldsymbol{I})^{-1}\boldsymbol{Q}\boldsymbol{H}_*\boldsymbol{Q}.$$

The proof is completed. □

H.2 Proof of Theorem 6

Proof. According to the assumption in Theorem 6, we know that for AdamW, its prior and posterior distributions are both Gaussian distribution, namely $\mathcal{P}_{\text{pre}} \sim \mathcal{N}(0, \rho \boldsymbol{I})$ and $\mathcal{P} \sim \mathcal{N}(\boldsymbol{x}_*, \boldsymbol{M}_{\text{AdamW}})$ where

$$\boldsymbol{M}_{\text{AdamW}} = \frac{\eta}{2b}(\boldsymbol{Q}\boldsymbol{H}_* + \lambda \boldsymbol{I})^{-1}\boldsymbol{Q}\boldsymbol{H}_*\boldsymbol{Q}.$$

On the other hand, for KL between two Gaussian distributions $\boldsymbol{W}_1 \sim (\boldsymbol{u}_1, \boldsymbol{\Sigma}_1)$ and $\boldsymbol{W}_2 \sim (\boldsymbol{u}_2, \boldsymbol{\Sigma}_2)$, we can follow [22] and compute it as follows:

$$\mathrm{KL}(\boldsymbol{W}_2 \| \boldsymbol{W}_1) = \frac{1}{2}\left[\log \frac{\det(\boldsymbol{\Sigma}_1)}{\det(\boldsymbol{\Sigma}_2)} + \mathrm{Tr}\left(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2\right)\right] + \frac{1}{2}(\boldsymbol{u}_1 - \boldsymbol{u}_2)^\top \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{u}_1 - \boldsymbol{u}_2) - \frac{d}{2}.$$

Accordingly, for AdamW, we can compute

$$\begin{aligned}\mathrm{KL}(\mathcal{P}\|\mathcal{P}_{\text{pre}}) =& \frac{1}{2}\left[\log \frac{\rho^d}{\left(\frac{\eta}{2b}\right)^d \det(\boldsymbol{M}_{\text{AdamW}})} + \frac{\eta}{2\rho b}\mathrm{Tr}(\boldsymbol{M}_{\text{AdamW}}) + \frac{1}{2\rho}\|\boldsymbol{x}_*\|^2 - \frac{d}{2}\right] \\ =& \frac{1}{2}\left[-\log\det(\boldsymbol{M}_{\text{AdamW}}) + \frac{\eta}{2\rho b}\mathrm{Tr}(\boldsymbol{M}_{\text{AdamW}}) + d\log\frac{2b\rho}{\eta} + \frac{1}{2\rho}\|\boldsymbol{x}_*\|^2 - \frac{d}{2}\right].\end{aligned}$$

Then by using Lemma 4, it further yields the generalization bound of AdamW as follows:

$$\mathbb{E}_{\boldsymbol{\xi}\sim\mathcal{D},\boldsymbol{x}\sim\mathcal{P}}[f(\boldsymbol{x},\boldsymbol{\xi})] - \mathbb{E}_{\boldsymbol{\xi}\in\mathcal{D}_{\text{tr}},\boldsymbol{x}\sim\mathcal{P}}[f(\boldsymbol{x},\boldsymbol{\xi})] \leq \sqrt{\frac{8}{n}\left(-\log\det(\boldsymbol{M}_{\text{AdamW}}) + \frac{\eta}{2\rho b}\mathrm{Tr}(\boldsymbol{M}_{\text{AdamW}}) + d\log\frac{2b\rho}{\eta} + c_0\right)},$$

where $c_0 = \frac{1}{2\rho}\|\boldsymbol{x}_*\|^2 - \frac{d}{2} + 2\ln\left(\frac{2n}{\tau}\right)$. The proof is completed. □

H.3 Proof of Theorem 7

Proof. **Step 1. Posterior Analysis on Adam+ℓ_2-Regularization.** Here we borrow the same idea in Lemma 5 and Theorem 6 to analyze the covariance matrix $\boldsymbol{M} = \mathbb{E}\left[\boldsymbol{x}_t\boldsymbol{x}_t^\top\right]$. To begin with, we simplify the SDE of Adam+ℓ_2-Regularization. Based on the assumptions in Theorem 7, we can write the SDE equations as follows:

$$\begin{aligned}\mathsf{d}\boldsymbol{x}_t =& -\boldsymbol{Q}_t\nabla F(\boldsymbol{x}_t)\mathsf{d}t - \lambda\boldsymbol{Q}_t\boldsymbol{x}_t\mathsf{d}t + \boldsymbol{Q}_t\left(2\boldsymbol{\Sigma}_t\right)^{\frac{1}{2}}\mathsf{d}\boldsymbol{\zeta}_t \\ =& -\boldsymbol{Q}_t\boldsymbol{H}_*\boldsymbol{x}_t\mathsf{d}t - \lambda\boldsymbol{Q}_t\boldsymbol{x}_t\mathsf{d}t + \boldsymbol{Q}_t\left(2\boldsymbol{\Sigma}_t\right)^{\frac{1}{2}}\mathsf{d}\boldsymbol{\zeta}_t \\ =& -(\boldsymbol{Q}\boldsymbol{H}_* + \lambda\boldsymbol{Q})\boldsymbol{x}_t\mathsf{d}t + \sqrt{\frac{\eta}{b}}\boldsymbol{Q}\boldsymbol{H}_*^{\frac{1}{2}}\mathsf{d}\boldsymbol{\zeta}_t,\end{aligned}$$

where $\mathsf{d}\boldsymbol{\zeta}_t \sim \mathcal{N}(0, \boldsymbol{I}\mathsf{d}t)$, $\boldsymbol{\Sigma}_t \approx \frac{\eta}{2B}\boldsymbol{H}_*$; $\boldsymbol{Q}_t = \boldsymbol{Q} := \mathrm{diag}\left([\boldsymbol{H}_{*(11)}^{-\frac{1}{2}}, \boldsymbol{H}_{*(22)}^{-\frac{1}{2}}, \cdots, \boldsymbol{H}_{*(dd)}^{-\frac{1}{2}}]\right)$. Then for this Ornstein–Uhlenbeck process, we can compute its closed form solution as follows:

$$\boldsymbol{x}_t = \exp\left(-(\boldsymbol{Q}\boldsymbol{H}_* + \lambda\boldsymbol{Q})t\right)\boldsymbol{x}_0 + \sqrt{\frac{\eta}{b}}\boldsymbol{Q}\boldsymbol{H}_*^{\frac{1}{2}}\int_0^t \exp\left(-(\boldsymbol{Q}\boldsymbol{H}_* + \lambda\boldsymbol{Q})(t-t')\right)\mathsf{d}\boldsymbol{\zeta}_{t'}.$$

Let $\boldsymbol{M} = \mathbb{E}\left[\boldsymbol{x}_t\boldsymbol{x}_t^\top\right]$. In this way, we follow [19] (see their Appendix b) and can further compute the algebraic relation for the stationary covariance of the multivariate Ornstein–Uhlenbeck process as follows:

$$\begin{aligned}&(\boldsymbol{Q}\boldsymbol{H}_* + \lambda\boldsymbol{Q})\boldsymbol{M} + \boldsymbol{M}^\top(\boldsymbol{Q}\boldsymbol{H}_* + \lambda\boldsymbol{Q})^\top \\ =&\frac{\eta}{b}\int_{-\infty}^t (\boldsymbol{Q}\boldsymbol{H}_* + \lambda\boldsymbol{Q})\exp\left(-(\boldsymbol{H}_*^{\frac{1}{2}} + \lambda\boldsymbol{H}_*^{-\frac{1}{2}})(t-t')\right)\boldsymbol{Q}\boldsymbol{H}_*^{\frac{1}{2}}(\boldsymbol{Q}\boldsymbol{H}_*^{\frac{1}{2}})^\top\exp\left(-(\boldsymbol{Q}\boldsymbol{H}_* + \lambda\boldsymbol{Q})^\top(t-t')\right)\mathsf{d}t' \\ &+ \frac{\eta}{b}\int_{-\infty}^t \exp\left(-(\boldsymbol{Q}\boldsymbol{H}_* + \lambda\boldsymbol{Q})(t-t')\right)\boldsymbol{Q}\boldsymbol{H}_*^{\frac{1}{2}}(\boldsymbol{Q}\boldsymbol{H}_*^{\frac{1}{2}})^\top\exp\left(-(\boldsymbol{Q}\boldsymbol{H}_* + \lambda\boldsymbol{Q})^\top(t-t')\right)\mathsf{d}t'(\boldsymbol{Q}\boldsymbol{H}_* + \lambda\boldsymbol{Q}) \\ =&\frac{\eta}{b}\int_{-\infty}^t \frac{\mathsf{d}}{\mathsf{d}t'}\left(\exp\left(-(\boldsymbol{Q}\boldsymbol{H}_* + \lambda\boldsymbol{Q})(t-t')\right)\boldsymbol{Q}\boldsymbol{H}_*^{\frac{1}{2}}(\boldsymbol{Q}\boldsymbol{H}_*^{\frac{1}{2}})^\top\exp\left(-(\boldsymbol{Q}\boldsymbol{H}_* + \lambda\boldsymbol{Q})^\top(t-t')\right)\right) \\ =&\frac{\eta}{b}\boldsymbol{Q}\boldsymbol{H}_*^{\frac{1}{2}}(\boldsymbol{Q}\boldsymbol{H}_*^{\frac{1}{2}})^\top = \frac{\eta}{b}\boldsymbol{Q}\boldsymbol{H}_*\boldsymbol{Q},\end{aligned}$$

where we use the lower limits of the integral vanishes by the positivity of the eigenvalues of $\boldsymbol{Q}\boldsymbol{H}_* + \lambda\boldsymbol{Q}$. Then we have

$$\boldsymbol{M}_{\text{Adam+}\ell_2} = \frac{\eta}{2b}(\boldsymbol{Q}\boldsymbol{H}_* + \lambda\boldsymbol{Q})^{-1}\boldsymbol{Q}\boldsymbol{H}_*\boldsymbol{Q}.$$

Step 2. Generalization Analysis. According to the assumption in Theorem 7, we know that for Adam + ℓ_2 regularization, its prior and posterior distributions are both Gaussian distribution, namely $\mathcal{P}_{\text{pre}} \sim \mathcal{N}(0, \rho\boldsymbol{I})$ and $\mathcal{P} \sim \mathcal{N}(\boldsymbol{x}_*, \boldsymbol{M}_{\text{Adam+}\ell_2\text{-Reg.}})$ where

$$\boldsymbol{M}_{\text{Adam+}\ell_2} = \frac{\eta}{2b}(\boldsymbol{Q}\boldsymbol{H}_* + \lambda\boldsymbol{Q})^{-1}\boldsymbol{Q}\boldsymbol{H}_*\boldsymbol{Q}.$$

On the other hand, for KL between two Gaussian distributions $\boldsymbol{W}_1 \sim (\boldsymbol{u}_1, \boldsymbol{\Sigma}_1)$ and $\boldsymbol{W}_2 \sim (\boldsymbol{u}_2, \boldsymbol{\Sigma}_2)$, we can follow [22] and can compute

$$\mathbb{E}_{\boldsymbol{\xi}\sim\mathcal{D},\boldsymbol{x}\sim\mathcal{P}}[f(\boldsymbol{x},\boldsymbol{\xi})] - \mathbb{E}_{\boldsymbol{\xi}\in\mathcal{D}_{\text{tr}},\boldsymbol{x}\sim\mathcal{P}}[f(\boldsymbol{x},\boldsymbol{\xi})] \le \sqrt{\frac{8}{n}\left(-\log\det(\boldsymbol{M}_{\text{Adam}+\ell_2}) + \frac{\eta}{2\rho b}\text{Tr}(\boldsymbol{M}_{\text{Adam}+\ell_2}) + d\log\frac{2b\rho}{\eta} + c_0\right)},$$

where $c_0 = \frac{1}{2\rho}\|\boldsymbol{x}_*\|^2 - \frac{d}{2} + 2\ln\left(\frac{2n}{\tau}\right)$. The proof is completed. □

H.4 Proof of Corollary 3

Proof. Let $\boldsymbol{U}\boldsymbol{S}\boldsymbol{U}^\top$ is the SVD of $\boldsymbol{H}_*$, where $\boldsymbol{S} = \text{diag}\,(\sigma_1, \sigma_2, \cdots, \sigma_d)$. When we approximate $\boldsymbol{Q} \approx \boldsymbol{H}_*^{-\frac{1}{2}}$, then $\boldsymbol{M}_{\text{AdamW}} = \frac{\eta}{2b}(\boldsymbol{Q}\boldsymbol{H}_* + \lambda\boldsymbol{I})^{-1}\boldsymbol{Q}\boldsymbol{H}_*\boldsymbol{Q}$ can be written as

$$\boldsymbol{M}_{\text{AdamW}} = \frac{\eta}{2b}\boldsymbol{U}(\boldsymbol{S}^{\frac{1}{2}} + \lambda\boldsymbol{I})^{-1}\boldsymbol{U}^\top.$$

Similarly, we can write $\boldsymbol{M}_{\text{Adam}+\ell_2} = \frac{\eta}{2b}(\boldsymbol{Q}\boldsymbol{H}_* + \lambda\boldsymbol{Q})^{-1}\boldsymbol{Q}\boldsymbol{H}_*\boldsymbol{Q}$ as

$$\boldsymbol{M}_{\text{Adam}+\ell_2} = \frac{\eta}{2b}\boldsymbol{U}\boldsymbol{S}^{\frac{1}{2}}(\boldsymbol{S} + \lambda\boldsymbol{I})^{-1}\boldsymbol{U}^\top.$$

Accordingly, we can compute

$$\begin{aligned}\Phi_{\text{AdamW}} =& \frac{\sqrt{8}}{\sqrt{n}}\left(-\log\det(\boldsymbol{M}_{\text{AdamW}}) + \frac{\eta}{2\rho b}\text{Tr}(\boldsymbol{M}_{\text{AdamW}}) + d\log\frac{2b\rho}{\eta} + c_0\right)^{\frac{1}{2}}\\ =& 4\sqrt{\frac{1}{2n}\left(\sum_{i=1}^{d}\log\frac{2\rho b(\sigma_i^{\frac{1}{2}} + \lambda)}{\eta} + \frac{\eta}{2\rho b}\sum_{i=1}^{d}\frac{1}{\sigma_i^{\frac{1}{2}} + \lambda} + c_0\right)}\\ =& \frac{\sqrt{8}}{\sqrt{n}}(\mathsf{err}_{\text{adamw}} + c_0)^{\frac{1}{2}},\end{aligned}$$

where $c_0 = \frac{1}{2\rho}\|\boldsymbol{x}_*\|^2 - \frac{d}{2} + 2\ln\left(\frac{2n}{\tau}\right)$, $\mathsf{err}_{\text{adamw}} = \sum_{i=1}^{d} h(x_{\text{AdamW}}^{(i)})$ with $x_{\text{AdamW}}^{(i)} = 2\eta^{-1}\rho b(\sigma_i^{\frac{1}{2}} + \lambda)$ and $h(x) = \log x + \frac{1}{x}$. Similarly, we can obtain

$$\begin{aligned}\Phi_{\text{Adam }+\ell_2} =& \frac{\sqrt{8}}{\sqrt{n}}\left(-\log\det(\boldsymbol{M}_{\text{Adam }+\ell_2}) + \frac{\eta}{2\rho b}\text{Tr}(\boldsymbol{M}_{\text{Adam }+\ell_2}) + d\log\frac{2b\rho}{\eta} + c_0\right)^{\frac{1}{2}}\\ =& 4\sqrt{\frac{1}{2n}\left(\sum_{i=1}^{d}\log\frac{2\rho b(\sigma_i + \lambda)}{\eta\sigma_i^{\frac{1}{2}}} + \frac{\eta}{2\rho b}\sum_{i=1}^{d}\frac{\sigma_i^{\frac{1}{2}}}{\sigma_i + \lambda} + c_0\right)}\\ =& \frac{\sqrt{8}}{\sqrt{n}}(\mathsf{err}_{\text{adam}+\ell_2} + c_0)^{\frac{1}{2}},\end{aligned}$$

where $\mathsf{err}_{\text{adam}+\ell_2} = \sum_{i=1}^{d} h(x_{\text{Adam }+\ell_2}^{(i)})$ with $x_{\text{Adam }+\ell_2}^{(i)} = 2\eta^{-1}\rho b(\sigma_i^{\frac{1}{2}} + \lambda\sigma_i^{-\frac{1}{2}})$. The proof is completed. □

Appendix I
Proofs of Auxiliary Lemmas

I.1 Proof of Lemma 1

Proof. Here we use mathematical induction to prove the first two results. Assume for $t \le k$, we have $\|\boldsymbol{m}_t\|_\infty \le c_\infty$ and $\|\boldsymbol{n}_t + \delta\|_\infty \le c_\infty + \delta$. Then for $k+1$, we have

$$\begin{aligned}\|\boldsymbol{m}_{k+1}\|_\infty &= \|(1-\beta_1)\boldsymbol{m}_k + \beta_1\boldsymbol{g}_k\|_\infty \le (1-\beta_1)\|\boldsymbol{m}_k\|_\infty + \beta_1\|\boldsymbol{g}_k\|_\infty \le c_\infty,\\ \|\boldsymbol{n}_{k+1}\|_\infty &= \left\|(1-\beta_2)\boldsymbol{n}_k + \beta_2\boldsymbol{g}_k^2\right\|_\infty \le (1-\beta_2)\|\boldsymbol{n}_k\|_\infty + \beta_2\left\|\boldsymbol{g}_k^2\right\|_\infty \le c_\infty^2,\end{aligned}$$

where $\boldsymbol{g}_k = \frac{1}{b}\sum_{i=1}^{b}\nabla f(\boldsymbol{x}_k; \boldsymbol{\zeta}_i)$. On the other hand, we have

$$\left\|\frac{\boldsymbol{n}_k + \delta}{\boldsymbol{n}_{k+1} + \delta}\right\|_\infty = \left\|1 + \frac{\boldsymbol{n}_k - \boldsymbol{n}_{k+1}}{\boldsymbol{n}_{k+1} + \delta}\right\|_\infty = \left\|1 + \frac{\beta_2(\boldsymbol{n}_k - \boldsymbol{g}_k^2)}{\boldsymbol{n}_{k+1} + \delta}\right\|_\infty \in \left[1 - \frac{\beta_2 c_\infty^2}{c_{s,\infty}^2 + \delta}, 1 + \frac{\beta_2 c_\infty^2}{c_{s,\infty}^2 + \delta}\right]$$

where $\boldsymbol{n}_{k+1} = (1-\beta_2)\boldsymbol{n}_k + \beta_2\boldsymbol{g}_k^2$. Therefore, for any $1 \ge p \ge 0$, we can easily obtain

$$\left\|\frac{(\boldsymbol{n}_k + \delta)^p}{(\boldsymbol{n}_{k+1} + \delta)^p}\right\|_\infty \in \left[\left(1 - \frac{\beta_2 c_\infty^2}{c_{s,\infty}^2 + \delta}\right)^p, \left(1 + \frac{\beta_2 c_\infty^2}{c_{s,\infty}^2 + \delta}\right)^p\right] \in \left[1 - \frac{\beta_2 c_\infty^2}{c_{s,\infty}^2 + \delta}, 1 + \frac{\beta_2 c_\infty^2}{c_{s,\infty}^2 + \delta}\right].$$

where $\boldsymbol{n}_{k+1} = (1-\beta_2)\boldsymbol{n}_k + \beta_2\boldsymbol{g}_k^2$. The proof is completed. □

I.2 Proof of Lemma 3

Proof. To prove $(1-x)^{\frac{3}{2}} \leq 1-x^{1-\alpha}$, we only need to prove $3x+\frac{2}{x^{\alpha}} \leq 3+x^2+x^{1-2\alpha}$. Since $x \in (0, \frac{1}{4})$, we have $3x+\frac{2}{x^{\alpha}} \leq \frac{3}{4}+\frac{2}{x^{\alpha}}$. In this way, we only need to prove $\frac{2}{x^{\alpha}} \leq \frac{9}{4}+x^2+x^{1-2\alpha}$. This means that if we prove $\frac{2}{x^{\alpha}} \leq \frac{9}{4}$, then we can obtain the desired result, since $x^2+x^{1-2\alpha}>0$. For $\frac{2}{x^{\alpha}} \leq \frac{9}{4}$, we can transfer it into its equivalent formulation: $\frac{8}{9} \leq x^{\alpha}$. Since $x \in (0, \frac{1}{4})$, we can always find a very small $\alpha>0$ so that $\frac{8}{9} \leq x^{\alpha}$. This completes our proof. □

I.3 Proof of Eqn. (10) and Eqn. (9) in Appendix D

Proof. The improvement of Eqn. (10) over Eqn. (9) in Appendix comes from their different techniques. Our Eqn. (10) is derived from Eqn. (16) in the Appendix G.2, while Eqn. (9) in [8] is derived by applying the bounding technique in [8], namely, the technique in the equation below their Eqn. (11). By comparison, our Eqn. (16) uses more tighter bound to prove the desired results.

Specifically, we can extend the bounding technique in [8] to AdamW, to derive the results. In this way, Xie et al. and we can both obtain the following same inequality:

$$F_{k+1}(\boldsymbol{x}_{k+1}) \leq F_k(\boldsymbol{x}_k) + \langle \nabla F(\boldsymbol{x}_k) + \lambda_k \boldsymbol{x}_k \odot \boldsymbol{v}_k, \boldsymbol{x}_{k+1}-\boldsymbol{x}_k \rangle + \frac{L}{2}\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\|_2^2 + \frac{\lambda_k}{2}\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\|_{\boldsymbol{v}_k}^2. \tag{28}$$

For Eqn. (28), one can refer to the derivation in Eqn. (16) in our Appendix G.2. Then, we follow the bounding technique in [8], and can prove the following results on AdamW:

$$\begin{aligned}
F_{k+1}(\boldsymbol{x}_{k+1}) \overset{①}{\leq} & F_k(\boldsymbol{x}_k) + \langle \nabla F(\boldsymbol{x}_k) - \boldsymbol{m}_k + \boldsymbol{m}_k + \lambda_k \boldsymbol{x}_k \odot \boldsymbol{v}_k, \boldsymbol{x}_{k+1}-\boldsymbol{x}_k \rangle + \left(\frac{L}{2c_1}+\frac{\lambda_k}{2}\right)\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\|_{\boldsymbol{v}_k}^2 \\
\overset{②}{\leq} & F_k(\boldsymbol{x}_k) + \langle \nabla F(\boldsymbol{x}_k) - \boldsymbol{m}_k, \boldsymbol{x}_{k+1}-\boldsymbol{x}_k \rangle - \left(\frac{1}{\eta}-\frac{L}{2c_1}-\frac{\lambda_k}{2}\right)\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\|_{\boldsymbol{v}_k}^2 \\
\overset{③}{\leq} & F_k(\boldsymbol{x}_k) + \frac{\eta}{2}\|\nabla F(\boldsymbol{x}_k) - \boldsymbol{m}_k\|_{1/\sqrt{\boldsymbol{v}_k}}^2 - \left(\frac{1}{\eta}-\frac{L}{2c_1}-\frac{\lambda_k}{2}\right)\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\|_{\boldsymbol{v}_k}^2 \\
\overset{④}{\leq} & F_k(\boldsymbol{x}_k) + \frac{\eta}{2c_1}\|\nabla F(\boldsymbol{x}_k) - \boldsymbol{m}_k\|_2^2 - \frac{\eta}{4c_2}\|\boldsymbol{m}_k + \lambda_k \boldsymbol{x}_k \odot \boldsymbol{v}_k\|_2^2,
\end{aligned} \tag{29}$$

where ① holds since $\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\|_2^2 \leq \frac{1}{c_1}\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\|_{\boldsymbol{v}_k}^2$ because of $c_1 := \delta^{0.5} \leq \|\boldsymbol{v}_k\|_\infty$. ① holds since $\boldsymbol{x}_{k+1}-\boldsymbol{x}_k = -\eta\left(\frac{\boldsymbol{m}_k}{\boldsymbol{v}_k}+\lambda_k \boldsymbol{x}_k\right)$ and $c_1 \leq \|\boldsymbol{v}_k\|_\infty \leq c_2 := (c_\infty^2+\delta)^{0.5}$ which together yield

$$\begin{aligned}
\langle \boldsymbol{m}_k + \lambda_k \boldsymbol{x}_k \odot \boldsymbol{v}_k, \boldsymbol{x}_{k+1}-\boldsymbol{x}_k \rangle = & \langle \boldsymbol{v}_k\left(\frac{\boldsymbol{m}_k}{\boldsymbol{v}_k + \lambda_k \boldsymbol{x}_k}\right), \boldsymbol{x}_{k+1}-\boldsymbol{x}_k \rangle = -\frac{1}{\eta}\langle (\boldsymbol{x}_{k+1}-\boldsymbol{x}_k) \odot \boldsymbol{v}_k, \boldsymbol{x}_{k+1}-\boldsymbol{x}_k \rangle \\
= & -\frac{1}{\eta}\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\|_{\boldsymbol{v}_k}^2.
\end{aligned} \tag{30}$$

③ holds by using $\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \langle \boldsymbol{a}\sqrt{\boldsymbol{v}_k}, \boldsymbol{b}/\sqrt{\boldsymbol{v}_k} \rangle \leq \frac{1}{2\eta}\|\boldsymbol{a}\|_{\boldsymbol{v}_k}^2 + \frac{\eta}{2}\|\boldsymbol{b}\|_{1/\boldsymbol{v}_k}^2$, and ④ holds since $\eta \leq \frac{c_1}{2(L+\lambda c_1)}$ so that a) $\frac{1}{2\eta}-\frac{L}{2c_1}-\frac{\lambda_k}{2} \geq \frac{1}{2\eta}-\frac{L}{2c_1}-\frac{\lambda}{2} \geq \frac{1}{4\eta}$ and b) $\|\boldsymbol{x}_{k+1}-\boldsymbol{x}_k\|_{\boldsymbol{v}_k}^2 = \eta^2\left\|\frac{\boldsymbol{m}_k}{\boldsymbol{v}_k}+\lambda_k \boldsymbol{x}_k\right\|_{\boldsymbol{v}_k}^2 = \eta^2\left\|\boldsymbol{m}_k + \lambda_k \boldsymbol{x}_k \odot \boldsymbol{v}_k\right\|_{1/\boldsymbol{v}_k}^2 \geq \frac{\eta^2}{c_2}\left\|\boldsymbol{m}_k + \lambda_k \boldsymbol{x}_k \cdot \boldsymbol{v}_k\right\|_2^2$.

In contrast, based on Eqn. (28) and

$$\boldsymbol{u}_k = \boldsymbol{m}_k + \lambda_k \boldsymbol{x}_k \odot \boldsymbol{v}_k, \qquad \boldsymbol{x}_{k+1}-\boldsymbol{x}_k = -\eta\frac{\boldsymbol{m}_k + \lambda_k \boldsymbol{x}_k \odot \boldsymbol{v}_k}{\boldsymbol{v}_k} = -\eta\frac{\boldsymbol{u}_k}{\boldsymbol{v}_k}, \tag{31}$$

in this work, we use a different bounding technique and prove a tight bound as

$$\begin{aligned}
F_{k+1}(\boldsymbol{x}_{k+1}) \leq & F_k(\boldsymbol{x}_k) - \eta\left\langle \nabla F(\boldsymbol{x}_k) + \lambda_k \boldsymbol{x}_k \odot \boldsymbol{v}_k, \frac{\boldsymbol{u}_k}{\boldsymbol{v}_k} \right\rangle + \frac{L\eta^2}{2}\left\|\frac{\boldsymbol{u}_k}{\boldsymbol{v}_k}\right\|_2^2 + \frac{\lambda_k\eta^2}{2}\left\|\frac{\boldsymbol{u}_k}{\boldsymbol{v}_k}\right\|_{\boldsymbol{v}_k}^2 \\
\leq & F_k(\boldsymbol{x}_k) - \left\langle \sqrt{\frac{\eta}{\boldsymbol{v}_k}}\left(\nabla F(\boldsymbol{x}_k) + \lambda_k \boldsymbol{x}_k \odot \boldsymbol{v}_k\right), \sqrt{\frac{\eta}{\boldsymbol{v}_k}}\boldsymbol{u}_k \right\rangle + \frac{L\eta^2}{2}\left\|\frac{\boldsymbol{u}_k}{\boldsymbol{v}_k}\right\|_2^2 + \frac{\lambda_k\eta^2}{2}\left\|\frac{\boldsymbol{u}_k}{\boldsymbol{v}_k}\right\|_{\boldsymbol{v}_k}^2 \\
\overset{①}{\leq} & F_k(\boldsymbol{x}_k) + \frac{1}{2}\left\|\sqrt{\frac{\eta}{\boldsymbol{v}_k}}\left(\nabla F(\boldsymbol{x}_k) + \lambda_k \boldsymbol{x}_k \odot \boldsymbol{v}_k - \boldsymbol{u}_k\right)\right\|_2^2 - \frac{1}{2}\left\|\sqrt{\frac{\eta}{\boldsymbol{v}_k}}\left(\nabla F(\boldsymbol{x}_k) + \lambda_k \boldsymbol{x}_k \odot \boldsymbol{v}_k\right)\right\|_2^2 \\
& - \frac{1}{2}\left\|\sqrt{\frac{\eta}{\boldsymbol{v}_k}}\boldsymbol{u}_k\right\| + \frac{L\eta^2}{2}\left\|\frac{\boldsymbol{u}_k}{\boldsymbol{v}_k}\right\|_2^2 + \frac{\lambda_k\eta^2}{2}\left\|\frac{\boldsymbol{u}_k}{\boldsymbol{v}_k}\right\|_{\boldsymbol{v}_k}^2 \\
\overset{②}{\leq} & F_k(\boldsymbol{x}_k) + \frac{\eta}{2c_1}\left\|\nabla F(\boldsymbol{x}_k) - \boldsymbol{m}_k\right\|_2^2 - \left[\frac{\eta}{2c_2} - \frac{L\eta^2}{2c_1^2} - \frac{\lambda_k\eta^2}{2c_1}\right]\left\|\boldsymbol{u}_k\right\|_2^2 - \frac{\eta}{2c_2}\left\|\nabla F(\boldsymbol{x}_k) + \lambda_k \boldsymbol{x}_k \odot \boldsymbol{v}_k\right\|_2^2 \\
\overset{③}{\leq} & F_k(\boldsymbol{x}_k) + \frac{\eta}{2c_1}\left\|\nabla F(\boldsymbol{x}_k) - \boldsymbol{m}_k\right\|_2^2 - \frac{\eta}{4c_2}\left\|\boldsymbol{u}_k\right\|_2^2 - \frac{\eta}{2c_2}\left\|\nabla F(\boldsymbol{x}_k) + \lambda_k \boldsymbol{x}_k \odot \boldsymbol{v}_k\right\|_2^2 \\
\overset{④}{\leq} & F_k(\boldsymbol{x}_k) + \frac{\eta}{2c_1}\left\|\nabla F(\boldsymbol{x}_k) - \boldsymbol{m}_k\right\|_2^2 - \frac{\eta}{4c_2}\left\|\boldsymbol{m}_k + \lambda_k \boldsymbol{x}_k \odot \boldsymbol{v}_k\right\|_2^2 - \frac{\eta}{2c_2}\left\|\nabla F(\boldsymbol{x}_k) + \lambda_k \boldsymbol{x}_k \odot \boldsymbol{v}_k\right\|_2^2,
\end{aligned} \tag{32}$$

where ① uses $-\langle \boldsymbol{a}, \boldsymbol{b}\rangle = \frac{1}{2}\|\boldsymbol{a}-\boldsymbol{b}\|_2^2 - \frac{1}{2}\|\boldsymbol{a}\|_2^2 - \frac{1}{2}\|\boldsymbol{b}\|_2^2$. ② holds because of $c_1 := \delta^p \leq \|\boldsymbol{v}_k\|_\infty \leq c_2 := (c_\infty^2 + \delta)^{0.5}$. ③ holds since we set $\eta \leq \frac{c_1^2}{2c_2(L+\lambda c_1)}$ such that $\frac{\eta}{4c_2} \geq \frac{L\eta^2}{2c_1^2} + \frac{\lambda_k \eta^2}{2c_1}$ in which we use $\lambda_k \leq \lambda$.

By comparison, by using the techniques in [8] on AdamW, we can only obtain

$$F_{k+1}(\boldsymbol{x}_{k+1}) \leq F_k(\boldsymbol{x}_k) + \frac{\eta}{2c_1}\|\nabla F(\boldsymbol{x}_k) - \boldsymbol{m}_k\|_2^2 - \frac{\eta}{4c_2}\|\boldsymbol{m}_k + \lambda_k \boldsymbol{x}_k \odot \boldsymbol{v}_k\|_2^2, \tag{33}$$

while using our own techniques on AdamW, we can obtain

$$F_{k+1}(\boldsymbol{x}_{k+1}) \leq F_k(\boldsymbol{x}_k) + \frac{\eta}{2c_1}\left\|\nabla F(\boldsymbol{x}_k) - \boldsymbol{m}_k\right\|_2^2 - \frac{\eta}{4c_2}\left\|\boldsymbol{m}_k + \lambda_k \boldsymbol{x}_k \odot \boldsymbol{v}_k\right\|_2^2 - \frac{\eta}{2c_2}\left\|\nabla F(\boldsymbol{x}_k) + \lambda_k \boldsymbol{x}_k \odot \boldsymbol{v}_k\right\|_2^2. \tag{34}$$

This means that the improvement of Eqn. (10) over Eqn. (9) in Appendix comes from their different techniques instead of the algorithmic algorithms. Note the extra term $-\frac{\eta}{2c_2}\left\|\nabla F(\boldsymbol{x}_k) + \lambda_k \boldsymbol{x}_k \odot \boldsymbol{v}_k\right\|_2^2$ in our bound can help cancel many terms related to $\left\|\nabla F(\boldsymbol{x}_k) + \lambda_k \boldsymbol{x}_k \odot \boldsymbol{v}_k\right\|_2^2$ and greatly simplify the proof as shown in our Appendix D. □

References

[1] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," in *Int'l Conf. Learning Representations*, 2018.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Int'l Conf. Learning Representations*, 2020.

[4] Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael W Mahoney, "Pyhessian: Neural networks through the lens of the hessian," in *Int'l conf. Big Data*, 2020, pp. 581–590.

[5] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang, "On large-batch training for deep learning: Generalization gap and sharp minima," *arXiv preprint arXiv:1609.04836*, 2016.

[6] U. Simsekli, L. Sagun, and M. Gurbuzbalaban, "A tail-index analysis of stochastic gradient noise in deep neural networks," in *Proc. Int'l Conf. Machine Learning*, 2019.

[7] Pan Zhou, Jiashi Feng, Chao Ma, Caiming Xiong, Steven Chu Hong Hoi, et al., "Towards theoretically understanding why SGD generalizes better than adam in deep learning," in *Proc. Conf. Neural Information Processing Systems*, 2020, vol. 33, pp. 21285–21296.

[8] Xingyu Xie, Pan Zhou, Huan Li, Zhouchen Lin, and Shuicheng Yan, "Adan: Adaptive nesterov momentum algorithm for faster optimizing deep models," *arXiv preprint arXiv:2208.06677*, 2022.

[9] Zhishuai Guo, Yi Xu, Wotao Yin, Rong Jin, and Tianbao Yang, "A novel convergence analysis for algorithms of the adam family," *arXiv preprint arXiv:2112.03459*, 2021.

[10] Xiangyi Chen, Sijia Liu, Ruoyu Sun, and Mingyi Hong, "On the convergence of a class of adam-type algorithms for non-convex optimization," in *Int'l Conf. Learning Representations*, 2018.

[11] 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 *Proc. Int'l Joint Conf. Artificial Intelligence*, 2021, pp. 3267–3275.

[12] Matthew Staib, Sashank Reddi, Satyen Kale, Sanjiv Kumar, and Suvrit Sra, "Escaping saddle points with adaptive gradient methods," in *Proc. Int'l Conf. Machine Learning*, 2019, pp. 5956–5965.

[13] Stanislaw Jastrzkebski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey, "Three factors influencing minima in SGD," in *Int'l Conf. Learning Representations*, 2017.

[14] James Martens, "New insights and perspectives on the natural gradient method," *J. of Machine Learning Research*, vol. 21, no. 1, pp. 5776–5851, 2020.

[15] Levent Sagun, Utku Evci, V Ugur Guney, Yann Dauphin, and Leon Bottou, "Empirical analysis of the Hessian of over-parametrized neural networks," *arXiv preprint arXiv:1706.04454*, 2017.

[16] Stephan Mandt, Matthew Hoffman, and David Blei, "A variational analysis of stochastic gradient algorithms," in *Proc. Int'l Conf. Machine Learning*, 2016, pp. 354–363.

[17] Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma, "The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects," in *Proc. Int'l Conf. Machine Learning*, 2019.

[18] Samuel L Smith and Quoc V Le, "A Bayesian perspective on generalization and stochastic gradient descent," in *Int'l Conf. Learning Representations*, 2018.

[19] Mandt Stephan, Matthew D Hoffman, David M Blei, et al., "Stochastic gradient descent as approximate bayesian inference," *J. of Machine Learning Research*, vol. 18, no. 134, pp. 1–35, 2017.

[20] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Int'l Conf. Learning Representations*, 2015.

[21] David A McAllester, "Some PCA-Bayesian theorems," *Machine Learning*, vol. 37, no. 3, pp. 355–363, 1999.

[22] Leandro Pardo, *Statistical inference based on divergence measures*, Chapman and Hall/CRC, 2018.