VII. APPENDIX

A Assumption 2

Regarding Assumption 2 in general, the distribution of the Bernoulli random variable representing the ReLU activation function is not required to be stationary for all t. Since all loss functions are considered separately, we only need to assume that for every z^t, there is a corresponding ρ^t such that $\mathrm{E}\left[\omega_{1,*}^T\omega_{2,*}z^t\right] = \rho^t\omega_{1,*}^T\omega_{2,*}z^t = y^t$, then, later in the proof, those ρ^t's are absorbed into $\mathrm{E}\left[l_t(\omega_{1,t},\omega_{2,t}|\mathrm{F}^t\right]$. Therefore, the algorithms can dynamically adapt to the new patterns in the dataset. In the proof, we simplify this process by using a constant ρ. Then, given σ,σ_1,σ_2 are i.i.d, $\mathrm{E}_\sigma\left[\left\|\omega_{1,t}^T\sigma\left(\omega_{2,t}z^t\right)-y^t\right\|^2/2\right] = \mathrm{E}_\sigma\left[\left\|\omega_{1,t}^T\sigma\left(\omega_{2,t}z^t\right)-\omega_{1,*}^T\sigma\left(\omega_{2,*}z^t\right)\right\|^2/2\right] = \frac{\rho}{2}(\omega_{1,t}^T\omega_{2,t}z^t - \omega_{1,*}^T\omega_{2,*}z^t)^2$. At the same time, the new loss function is $\mathrm{E}_{\sigma_1,\sigma_2}\left[\left(\omega_{1,t}^T\sigma_1\left(\omega_{2,t}z^t\right)-y^t\right)\left(\omega_{1,t}^T\sigma_2\left(\omega_{2,t}z^t\right)-y^t\right)/2\right] = \mathrm{E}_{\sigma_1}\left[\omega_{1,t}^T\sigma_1\left(\omega_{2,t}z^t\right)-y^t\right]\mathrm{E}_{\sigma_2}\left[\omega_{1,t}^T\sigma_2\left(\omega_{2,t}z^t\right)-y^t\right]/2 = \frac{\rho^2}{2}\left(\omega_{1,t}^T\omega_{2,t}z^t-\omega_{1,*}^T\omega_{2,*}z^t\right)^2$. Therefore, minimizing our new loss function is the same as minimizing the original loss given that ρ is a positive constant.

B Numerical Study

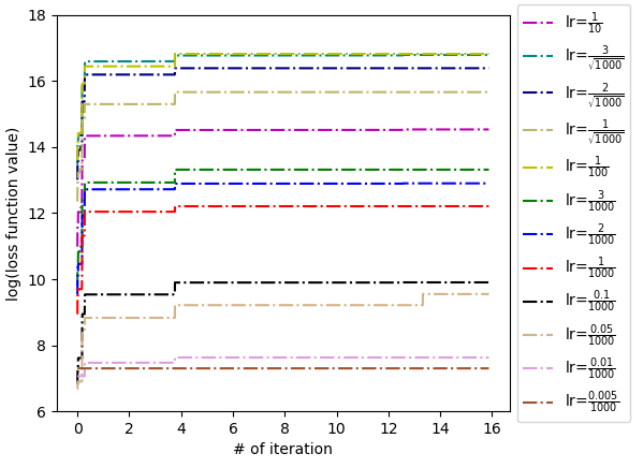
In this section, we compare the CONVGADAM method with OGD [1] for solving problem [st-problem] with a long sequence of data points (mimicking streaming). We conduct experiments on the MNIST8M dataset and two other different-size real datasets from the Yahoo! Research Alliance Webscope program. For all of these datasets, we train multi-class hinge loss support vector machines (SVM) [29] and we assume that the samples are streamed one by one based on a certain random order. The computing infrastructure was an on-premise Kubernetes cluster node on Linux operating system with 12 CPUs, 16GB RAM and three NVIDIA TITAN X GPUs, each with 12GB VRAM. For all the figures provided in this section, the horizontal axis is in 10^5 scale. Moreover, we set $\beta_1 = 0.8$ and $\beta_2 = 0.81$ in CONVGADAM. We mostly capture the log of the loss function value which is defined as $\max_{p\in\mathbb{N}}\min_{(\omega_t)_{t\in\mathbb{N}}}\sum_{t=p}^{T+p} f_t(\omega_t)$.

B.1 Multiclass SVM with Yahoo! Targeting User Modeling Dataset: We first compare CONVGADAM with OGD using the Yahoo! user targeting and interest prediction dataset consisting of Yahoo user profiles[3]. It contains 1,589,113 samples (i.e., user profiles), represented by a total of 13,346 features and 380 different classification problems (called labels in the supporting documentation) each one with 3 classes.

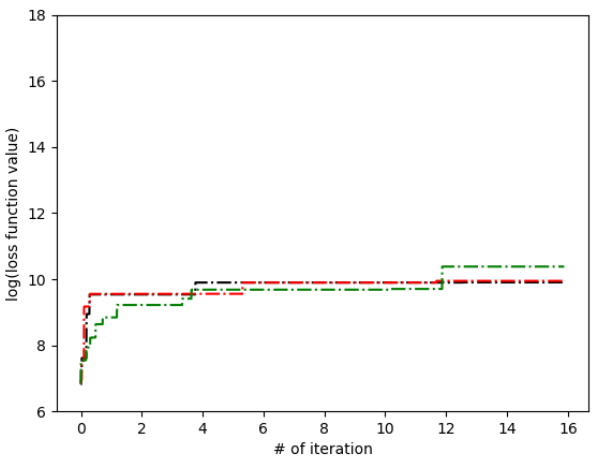
First, we pick the first label out and conduct a sequence of experiments with respect to this label. The most important results are presented in Figure 2 for OGD and Figure 3 for CONVGADAM. In Figures 2(a) and 3(a), we consider the cases when the learning rate or stepsize varies from 0.1 to $5\cdot10^{-6}$ while keeping the order and T fixed at 1,000. Figures 2(b) and 3(b) provide the influence of the order of the sequence. Figures 2(c) and 3(c) represent the case where T varies from 10 to 10^5 with a fixed learning rate or stepsize. Lastly, in Figure 3(d), we compare the performance of CONVGADAM and OGD with certain learning rates and stepsizes.

In these plots, we observe that CONVGADAM outperforms OGD for most of the learning rates and stepsizes, and definitely for promissing choices. More precisely, in Figure 2(a) and 3(a), we discover that 0.1/1000 and $3/\sqrt{1000}$ are two high-quality learning rate and stepsize values which have relatively low error and are learning for OGD and CONVGADAM, respectively. Therefore, we apply those two learning rates for the remaining experiments on this dataset. In Figures 2(b) and 3(b), we observe that the perturbation caused by the change of the order is negligible especially when compared to the loss value, which is a positive characteristic. Thus, in the remaining experiments, we no longer need to consider the impact of the order of the sequence. From Figure 2(c) and Figure 2(d), we discover that the loss and T have a significantly positive correlation as we expect. Notice that changing T but fixing the learning rate or stepsize essentially means containing more samples in the regret, in other words, the regret for $T = 100$ is roughly 10 times the regret for $T = 10$. Since the pattern in the figures is preserved for the different T values for OGD and CONVGADAM, in the remaining experiments we fix T. In Figure 3(c), we discover that too big T or too small T causes poor performance and therefore, for the remaining experiments, we set $T = 1,000$ whenever T is fixed. From Figure 3(d), we observe that CONVGADAM outperforms OGD.

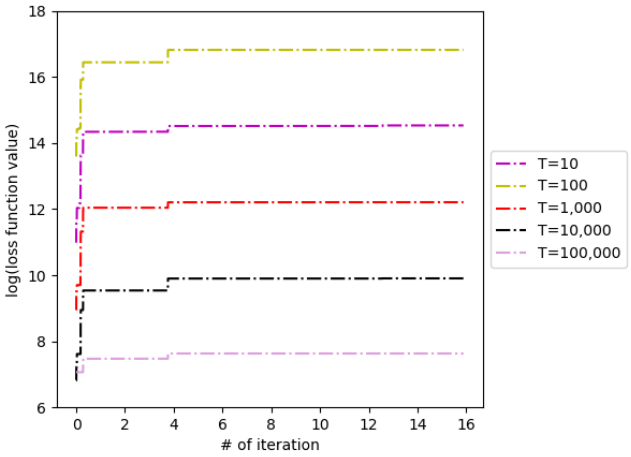
[3]https://webscope.sandbox.yahoo.com/catalog.php?datatype=a



(a)



(b)



(c)

Figure 2: Comparison of OGD for Different Orders, Learning Rates and T

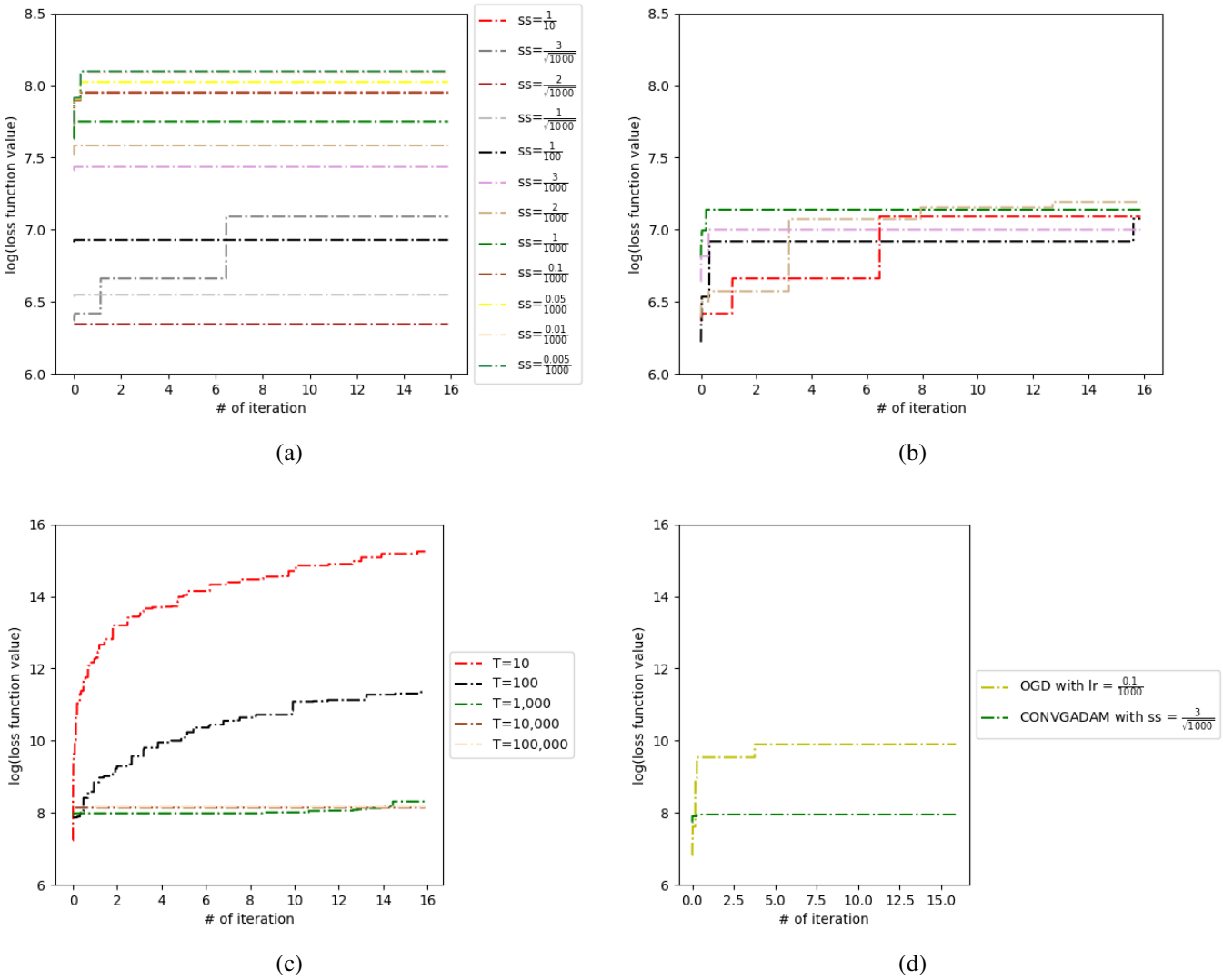


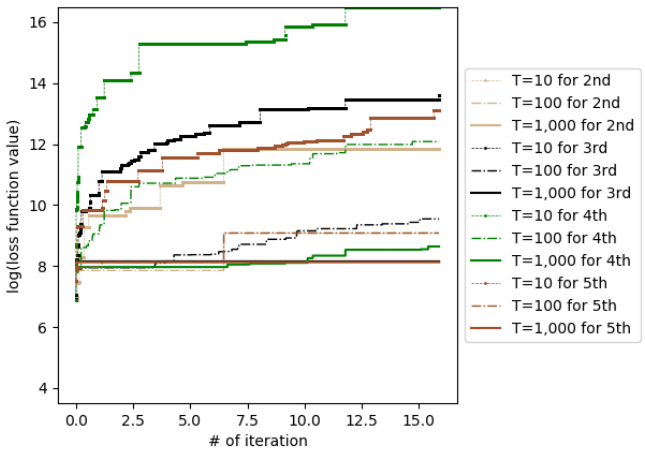
Figure 3: Comparison of CONVGADAM for Different Orders, Stepsizes and T

After studying the algorithms on the first label, we test them on the next four labels. In Figure 4, we compare the performance of CONVGADAM for different T and the difference with OGD on the four labels. In these plots, we observe that $T = 1000$ provides a more stable and better performance than the other two values. Moreover, CONVGADAM outperforms OGD for all considered learning rates and stepsizes.

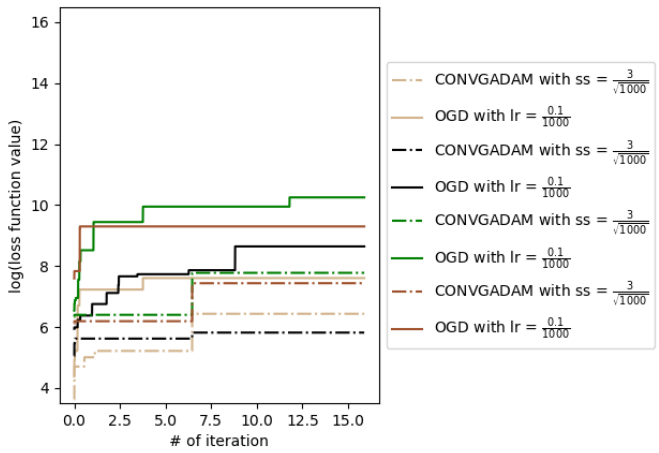
B.2 Multiclass SVM with MNIST8M Dataset: In this set of experiments, we study the performances of CONVGADAM and OGD on MNIST8M Dataset[4]. The dataset is generated on the fly by performing careful elastic deformation of the original MNIST training set. The dataset contains 8,100,000 samples, represented by a total of 784 features and 10 classes.

In Figures 5(a) and 5(b), we compare the performances of OGD and CONVGADAM for different learning rates and stepsizes. Figure 5(c) shows that performance of CONVGADAM for different T. Lastly, Figure 5(d) depicts the comparison of CONVGADAM and OGD. From Figures 5(a) and 5(b), we select the stepsize $2/\sqrt{1000}$ and the learning rate of 1/1000. As we observe, CONVGADAM always exhibits a better performance than OGD.

[4]https://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/

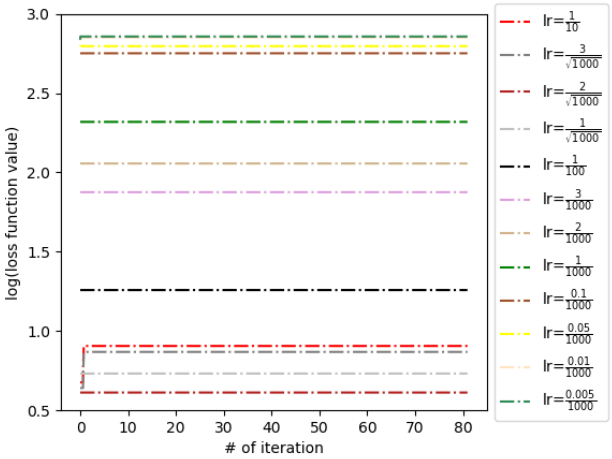


(a) Comparison of CONVGADAM for Different T

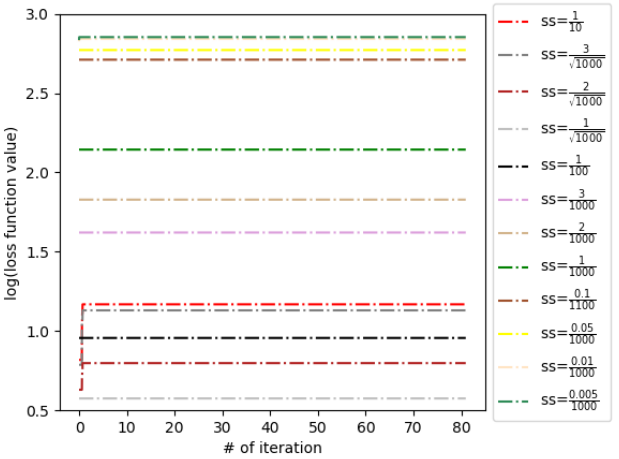


(b) Comparison of OGD and CONVGADAM

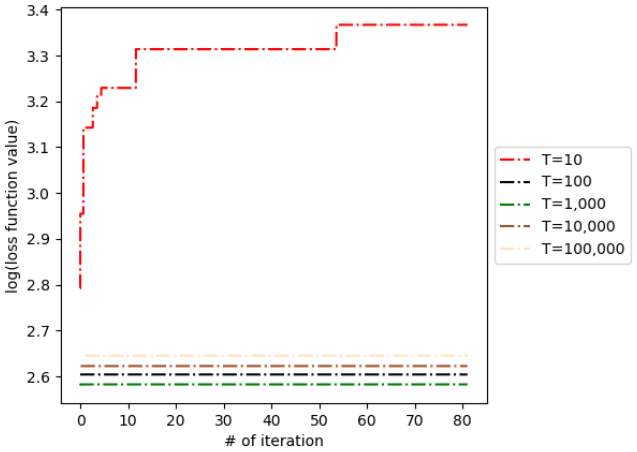
Figure 4: Performance of CONVGADAM and OGD on the Remaining Labels



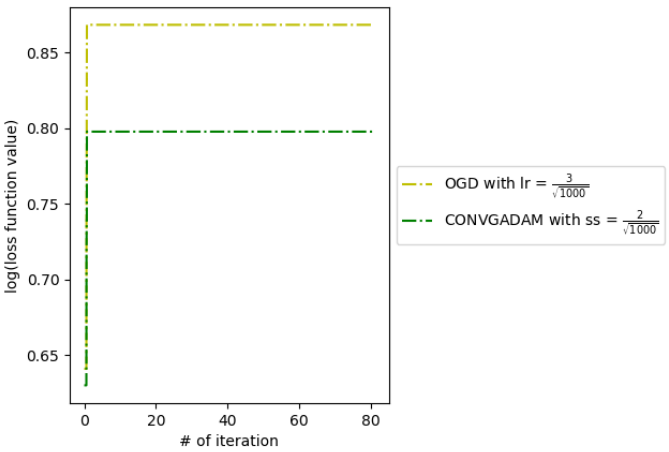
(a) Comparison of OGD for Different Learning Rates



(b) Comparison of CONVGADAM for Different Stepsizes



(c) Comparison of CONVGADAM for Different T



(d) Comparison of CONVGADAM and OGD

Figure 5: Performance of CONVGADAM on MINST8M Dataset

C Extensions

We first introduce techniques to guarantee boundedness of the weight ω, i.e. how to remove condition 1 in Assumption 1. We then point out problems in the proofs of AMSGRAD [4] and ADABOUND [3] and provide a different proof for AMSGRAD.

C.1 Unbounded Case: Projection is a popular technique to guarantee that a weight does not exceed a certain bound ([30], [18], [19], [3]). For unbounded weight $\hat{\omega}$, we introduce the following notation. Given convex sets $\mathcal{P}_1$, $\mathcal{P}_2$, vectors ω_1, ω_1', g_1 and matrix $\hat{v}$, we define projections

$$\begin{aligned}
&\Pi_{\mathcal{P}_1}(\hat{\omega}) = \operatorname*{argmin}_{\omega \in \mathcal{P}_1} \|\omega - \hat{\omega}\| \\
&\Pi^1_{\mathcal{P}_1,\mathcal{P}_2,\omega_1,g_1,\omega_1'}(\hat{\omega}_2) \\
&= \operatorname*{argmin}_{\omega_2':\omega_2' \cdot \left[\left\|\omega_1' - \eta g_1\right\|/\sqrt{\frac{1}{2}+\xi_1}\right] \in \mathcal{P}_2} \left\| \omega_2' - \operatorname*{argmin}_{\omega_2:\omega_1^T\omega_2 \in \mathcal{P}_1} \left\|\omega_1^T\omega_2 - \omega_1^T\hat{\omega}_2\right\| \right\| \\
&\Pi^2_{\mathcal{P}_1,\mathcal{P}_2,\omega_1,g_1,\omega_1',\hat{v}}(\hat{\omega}_2) \\
&= \operatorname*{argmin}_{\omega_2':\omega_2' \cdot \left[\left\|\omega_1' - \eta g_1\right\|/\sqrt{\frac{1}{2}+\xi_2}\right] \in \mathcal{P}_2} \left\| \omega_2' - \operatorname*{argmin}_{\omega_2:\omega_1^T\omega_2 \in \mathcal{P}_1} \left\| \left(\sqrt[4]{\hat{v}} \odot \omega_2\right)^T \omega_1 - \left(\sqrt[4]{\hat{v}} \odot \hat{\omega}_2\right)^T \omega_1 \right\| \right\|.
\end{aligned}$$

Projection Π is the standard projection which maps vector $\hat{\omega}$ into set $\mathcal{P}_1$. If an optimal weight ω_* is such that $\omega_* \in \mathcal{P}_1$, then we have

$$\left\|\Pi_{\mathcal{P}_1}(\hat{\omega}_{t+1}) - \omega_*\right\| \leq \left\|\hat{\omega}_{t+1} - \omega_*\right\|,$$

which could be directly applied in the proofs of Theorem 1 and 2.

For Π^1 and Π^2, we could regard them as a combination of two standard projections. Note that, for the outer projection, we require that it does not affect the product of $\omega_1^T\omega_2$, which could be done by projection methods for linear equality constraints. In this way, we have

$$\begin{aligned}
&\left\|\omega_{1,t+1}^T \Pi^1_{\mathcal{P}_1,\mathcal{P}_2,\omega_{1,t+1},g_{1,t},\omega_{1,t}}(\hat{\omega}_{2,t+1}) - \omega_{1,*}^T\omega_{2,*}\right\| \leq \left\|\omega_{1,t+1}^T\hat{\omega}_{2,t+1} - \omega_{1,*}^T\omega_{2,*}\right\| \\
&\left\|\left(\sqrt[4]{\hat{v}_{2,t}} \odot \Pi^2_{\mathcal{P}_1,\mathcal{P}_2,\omega_{1,t+1},g_{1,t},\omega_{1,t},\hat{v}_{2,t}}(\hat{\omega}_{2,t+1})\right)^T \omega_{1,t+1} - \left(\sqrt[4]{\hat{v}_{2,t}} \odot \omega_{2,*}\right)^T \omega_{1,*}\right\| \\
&\quad \leq \left\|\left(\sqrt[4]{\hat{v}_{2,t}} \odot (\hat{\omega}_{2,t+1})\right)^T \omega_{1,t+1} - \left(\sqrt[4]{\hat{v}_{2,t}} \odot \omega_{2,*}\right)^T \omega_{1,*}\right\|,
\end{aligned}$$

which could also be directly applied in the proofs of Theorem 3 and 4.

C.2 Standard setting of ADAM*:* First, let us point out the problem in AMSGRAD [4]. At the bottom of Page 18 in [4], the authors obtain an upper bound for the regret which has a term containing $\sum_{t=1}^T \frac{\beta_{1t}\hat{v}_{t,i}^{1/2}}{\alpha_t}$. Without assuming that β_{1t} is exponentially decaying, it is questionable to establish $\mathcal{O}(\sqrt{T})$ given $\alpha_t = \frac{1}{\sqrt{t}}$ since $\sum_{t=1}^T \sqrt{t} > \mathcal{O}(\sqrt{T})$. Although this questionable term can be bounded by assumptions on β_{1t}, the last term in Theorem 4 is $\mathcal{O}(\log(T)\sum_{i=1}^d \|g_{1:T,i}\|_2) = \mathcal{O}(\log(T)\sqrt{T})$ since $g_{1:T,i}$ is the concatenation of the gradients from 0 to current time T in the i^{th} coordinate. Moreover, the authors argue that decaying β_{1t} is crucial to guarantee the convergence, however, our proof shows $\mathcal{O}(\sqrt{T})$ regret for AMSGRAD with constant β and both constant and diminishing stepsizes, which is more practically relevant. For a diminishing stepsize, the slight change we need to make in the proof is that η_t needs to be considered together with $\sqrt{\hat{v}_{t,j}}$ in (11) and the rest of proof of Theorem 2. Applying the fact that $\frac{\sqrt{\hat{v}_{t,j}}}{\eta_t} \geq \frac{\sqrt{\hat{v}_{t-1,j}}}{\eta_t}$ and $\sum_{t=1}^T \frac{1}{\sqrt{t}} = 2\sqrt{T} - 1$ yields $\mathcal{O}(\sqrt{T})$ regret in standard online setting.

Table I summarizes the various regret bounds in different convex settings.

Table I: Summary of Known Regret Bounds for Online Learning and Streaming in Convex Setting

	gradient descent		Adam	
	constant	diminishing	constant	diminishing
standard online	$\mathcal{O}(\sqrt{T})$(us) $\mathcal{O}(T)$ [1]	$\mathcal{O}(\sqrt{T})$(us) $\mathcal{O}(\sqrt{T})$ [1]	$\mathcal{O}(\sqrt{T})$(us)	$\mathcal{O}(\sqrt{T})$(us) $\mathcal{O}(\sqrt{T})$ [4] (flawed) $\mathcal{O}(\log(T)\sqrt{T})$ [4] (true)
streaming	$\mathcal{O}(\sqrt{T})$(us)	$\mathcal{O}(T)$(us)	$\mathcal{O}(\sqrt{T})$(us)	$\mathcal{O}(T)$(us)

D Regret with Rolling Window Analysis of OGD

The highlevel overview of the proof is as follows. Under the assumptions in Assumption 1, by finding a relationship for the sequence of the weight error $\omega_t - \omega_p^*$ and employing the property of convexity from condition 3 from Assumption 1, we prove that OGD obtains the regret with rolling window which is proportional to the square root of the size of the rolling window when given the constant learning rate. This is consistent with the regret of OGD in the standard online setting.

Proof of Theorem 1:

Proof. For any $p \in \mathbb{N}$ and fixed T, from step 4 in Algorithm 1, for any ω^*, we obtain

$$\begin{aligned}\left\|\omega_{t+1} - \omega^*\right\|^2 &= \left\|\omega_t - \eta \bigtriangledown f_t(\omega_t) - \omega^*\right\|^2 \\ &= \left\|\omega_t - \omega^*\right\|^2 - 2\eta \left\langle \omega_t - \omega^*, \bigtriangledown f_t(\omega_t)\right\rangle + \eta^2 \left\|\bigtriangledown f_t(\omega_t)\right\|^2,\end{aligned}$$

which in turn yields

$$\left\langle \omega_t - \omega^*, \bigtriangledown f_t(\omega_t)\right\rangle = \frac{\left\|\omega_t - \omega^*\right\|^2 - \left\|\omega_{t+1} - \omega^*\right\|^2}{2\eta} + \frac{\eta}{2}\left\|\bigtriangledown f_t(\omega_t)\right\|^2. \tag{1}$$

Applying convexity of f_t yields

$$f_t(\omega_t) - f_t(\omega^*) \leq \left\langle \omega_t - \omega^*, \bigtriangledown f_t(\omega_t)\right\rangle. \tag{2}$$

Inserting (1) into (2) gives

$$f_t(\omega_t) - f_t(\omega^*) \leq \frac{\left\|\omega_t - \omega^*\right\|^2 - \left\|\omega_{t+1} - \omega^*\right\|^2}{2\eta} + \frac{\eta}{2}\left\|\bigtriangledown f_t(\omega_t)\right\|^2.$$

By summing up all differences, we obtain

$$\begin{aligned}\sum_{t=p}^{T+p}\left[f_t(\omega_t) - f_t(\omega^*)\right] &\leq \frac{1}{2}\sum_{t=p}^{T+p}\left[\frac{\left\|\omega_t - \omega^*\right\|^2 - \left\|\omega_{t+1} - \omega^*\right\|^2}{\eta} + \eta\left\|\bigtriangledown f_t(\omega_t)\right\|^2\right] \\ &\leq \frac{1}{2}\left(\frac{\left\|\omega_p - \omega^*\right\|^2}{\eta}\right) + dG_\infty \sum_{t=p}^{T+p}\eta \\ &\leq \frac{D_\infty^2\sqrt{T}}{2\eta_1} + dG_\infty\eta_1\sqrt{T} = \mathcal{O}(\sqrt{T}).\end{aligned} \tag{3}$$

The second inequality holds due to 2 in Assumption 1 and the last inequality uses 4 in Assumption 1 and the definition of η. Since (3) holds for any p and ω^*, setting $\omega^* = \omega_p^*$ for each p yields the statement in Theorem 1. □

E Regret with Rolling Window Analyses of CONVGADAM

The very technical proof follows the following steps. Based on the updating procedure in steps 4-8, we establish a relationship for the sequence of the weight error $\omega_t - \omega_p^*$. Meanwhile, considering condition 4 in Assumption 1, we obtain another relationship between the loss function error $f_t(\omega_t) - f_t(\omega_p^*)$ and $\langle \omega_t - \omega_*, \bigtriangledown f_t(\omega_t)\rangle$. Assembling these two relationships provide a relationship between the weight error $\omega_t - \omega_*$ and the loss function error $f_t(\omega_t) - f_t(\omega_*)$. By deriving upper bounds for all of the remaining terms based on conditions from Assumption 1, we are able to argue the same regret with rolling window of $\mathcal{O}(\sqrt{T})$.

Lemma 1. *Under the conditions assumed in Theorem 2, we have*

$$\sum_{t=p}^{T+p}\left\|\frac{1}{\sqrt[4]{\hat{v}_t}} \odot m_t\right\|^2 \leq \mathcal{O}(T).$$

Proof of Lemma 1. By the definition of $\hat{v}_t$, for any $t = p, p+1, \cdots, T+p$, we obtain

$$\begin{aligned}\left\| \frac{1}{\sqrt[4]{\hat{v}_t}} \odot m_t \right\|^2 &= \sum_{j=1}^{d} \frac{m_{t,j}^2}{\sqrt{\hat{v}_{t,j}}} \leq \sum_{j=1}^{d} \frac{m_{t,j}^2}{\sqrt{v_{t,j}}} = \sum_{j=1}^{d} \frac{\left((1-\beta_1)\sum_{i=1}^{t} \beta_1^{t-i} g_{i,j}\right)^2}{\sqrt{(1-\beta_2)\sum_{i=1}^{t} \beta_2^{t-i} g_{i,j}^2}} \\ &\leq \frac{(1-\beta_1)^2}{\sqrt{1-\beta_2}} \sum_{j=1}^{d} \frac{\left(\sum_{i=1}^{t} \beta_1^{t-i}\right)\left(\sum_{i=1}^{t} \beta_1^{t-i} g_{i,j}^2\right)}{\sqrt{\sum_{i=1}^{t} \beta_2^{t-i} g_{i,j}^2}} \\ &\leq \frac{1-\beta_1}{\sqrt{1-\beta_2}} \sum_{j=1}^{d} \frac{\sum_{i=1}^{t} \beta_1^{t-i} g_{i,j}^2}{\sqrt{\sum_{i=1}^{t} \beta_2^{t-i} g_{i,j}^2}} \leq \frac{1-\beta_1}{\sqrt{1-\beta_2}} \sum_{j=1}^{d} \sum_{i=1}^{t} \left(\frac{\beta_1}{\sqrt{\beta_2}}\right)^{t-i} \|g_{i,j}\|_2 \\ &= \frac{1-\beta_1}{\sqrt{1-\beta_2}} \sum_{j=1}^{d} \sum_{i=1}^{t} \lambda^{t-i} \|g_{i,j}\|_2 . \end{aligned} \tag{4}$$

The second equality follows from the updating rule of Algorithm 2. The second inequality follows from the Cauchy-Schwarz inequality, while the third inequality follows from the inequality $\sum_{i=1}^{t} \beta_1^{t-i} \leq \frac{1}{1-\beta_1}$. Using (4) for all time steps yields

$$\begin{aligned}&\sum_{t=p}^{T+p} \frac{1}{\sqrt{\hat{v}_t}} \odot (m_t \odot m_t) \\ \leq &\frac{1-\beta_1}{\sqrt{1-\beta_2}} \sum_{t=p}^{T+p} \sum_{j=1}^{d} \sum_{i=1}^{t} \lambda^{t-i} \|g_{i,j}\|_2 \\ = &\frac{1-\beta_1}{\sqrt{1-\beta_2}} \sum_{j=1}^{d} \sum_{t=p}^{T+p} \left(\sum_{i=p+1}^{t} \lambda^{t-i} \|g_{i,j}\|_2 + \sum_{i=1}^{p} \lambda^{t-i} \|g_{i,j}\|_2 \right) \\ = &\frac{1-\beta_1}{\sqrt{1-\beta_2}} \sum_{j=1}^{d} \left(\sum_{t=p+1}^{T+p} \sum_{i=p+1}^{t} \lambda^{t-i} \|g_{i,j}\|_2 + \sum_{t=p}^{T+p} \sum_{i=1}^{p} \lambda^{t-i} \|g_{i,j}\|_2 \right) \\ = &\frac{1-\beta_1}{\sqrt{1-\beta_2}} \sum_{j=1}^{d} \left(\sum_{t=p+1}^{T+p} \sum_{i=p+1}^{t} \lambda^{t-i} \|g_{i,j}\|_2 + \left(\sum_{i=1}^{p} \lambda^{p-i} \|g_{i,j}\|_2 \right) \left(\sum_{i=0}^{T} \lambda^{i} \right) \right). \end{aligned} \tag{5}$$

We first bound the first term in (5) for each j as follows,

$$\begin{aligned}&\sum_{t=p+1}^{T+p} \sum_{i=p+1}^{t} \lambda^{t-i} \|g_{i,j}\|_2 = \sum_{i=p+1}^{T+p} \|g_{i,j}\|_2 \sum_{t=i}^{T+p} \lambda^{T+p-t} \\ \leq &\frac{1}{1-\lambda} \sum_{t=p+1}^{T+p} \|g_{i,j}\|_2 \leq \frac{TG_\infty}{1-\lambda}. \end{aligned} \tag{6}$$

The first inequality follows from the fact that $\sum_{t=i}^{T+p} \lambda^{T+p-t} < \frac{1}{1-\lambda}$ and the last inequality is due to IV-B in Assumption 1. Using a similar argument, we further bound the second term in (5) as follows,

$$\begin{aligned}&\left(\sum_{i=1}^{p} \lambda^{p-i} \|g_{i,j}\|_2 \right) \left(\sum_{i=0}^{T} \lambda^{i} \right) \leq \frac{1}{1-\lambda} \left(\sum_{i=1}^{p} \lambda^{p-i} \|g_{i,j}\|_2 \right) \\ \leq &\frac{G_\infty}{1-\lambda} \left(\sum_{i=1}^{p} \lambda^{p-i} \right) \leq \frac{G_\infty}{(1-\lambda)^2}. \end{aligned} \tag{7}$$

Inserting (6) and (7) into (5) implies

$$\sum_{t=p}^{T+p} \left\| \frac{1}{\sqrt[4]{\hat{v}_t}} \odot m_t \right\|^2 \leq \frac{d\,(1-\beta_1)}{\sqrt{1-\beta_2}} \left(\frac{TG_\infty}{1-\lambda} + \frac{G_\infty}{(1-\lambda)^2} \right).$$

This completes the proof of the lemma. □

In order to establish the regret analysis of Algorithm 2, we further need the following intermediate result.

Lemma 2. *Under the conditions in Theorem 2, we have*

$$\sum_{t=p}^{T+p} \|m_{t-1}\|^2 \leq \mathcal{O}(T).$$

Proof of Lemma 2. By the definition of m_t, we obtain

$$\begin{aligned}
&\sum_{t=p}^{T+p} \|m_{t-1}\|^2 = \sum_{t=p}^{T+p}\sum_{j=1}^{d} m_{t-1,j}^2 \\
=&\sum_{t=p}^{T+p}\sum_{j=1}^{d}\left((1-\beta_1)\sum_{i=1}^{t}\beta_1^{t-i}g_{i,j}\right)^2 \\
\leq&(1-\beta_1)^2\sum_{t=p}^{T+p}\sum_{j=1}^{d}\left(\sum_{i=1}^{t}\beta_1^{t-i}\right)\left(\sum_{i=1}^{t}\beta_1^{t-i}g_{i,j}^2\right) \\
\leq&(1-\beta_1)\sum_{t=p}^{T+p}\sum_{j=1}^{d}\left(\sum_{i=1}^{t}\beta_1^{t-i}g_{i,j}^2\right) \leq (1-\beta_1)\sum_{t=p}^{T+p}\sum_{j=1}^{d}\left(G_\infty\sum_{i=1}^{t}\beta_1^{t-i}\right) \\
\leq&\sum_{t=p}^{T+p}\sum_{j=1}^{d} G_\infty = dTG_\infty.
\end{aligned}$$

The first inequality follows from the Cauchy-Schwarz inequality. The second and the last inequalities use the fact that $\sum_{i=1}^{t}\beta_1^{t-i} \leq \frac{1}{1-\beta_1}$. The third inequality is due to 2 in Assumption 1. This completes the proof of the lemma. □

Proof of Theorem 2:

Proof. Based on the update step 8 in Algorithm 2 and given any $\omega^* \in \mathbb{R}^d$, we obtain

$$\begin{aligned}
&\|\omega_{t+1}-\omega^*\|^2 = \left\|\omega_t - \frac{\eta}{\sqrt{\hat{v}_t}}\odot m_t - \omega^*\right\|^2 \\
=&\|\omega_t-\omega^*\|^2 - 2\left\langle \omega_t-\omega^*, \frac{\eta}{\sqrt{\hat{v}_t}}\odot m_t\right\rangle + \left\|\frac{\eta}{\sqrt{\hat{v}_t}}\odot m_t\right\|^2 \\
=&\|\omega_t-\omega^*\|^2 - 2\left\langle \omega_t-\omega^*, \frac{\eta(1-\beta_1)}{\sqrt{\hat{v}_t}}\odot g_t\right\rangle - 2\left\langle \omega_t-\omega^*, \frac{\eta\beta_1}{\sqrt{\hat{v}_t}}\odot m_{t-1}\right\rangle \\
&+\left\|\frac{\eta}{\sqrt{\hat{v}_t}\odot m_t}\right\|^2.
\end{aligned} \tag{8}$$

The first inequality uses the same argument as those used in Theorem 1. Rearranging (8) gives

$$\begin{aligned}
\langle \omega_t-\omega^*, g_t\rangle =& \frac{\left[\left\|\sqrt[4]{\hat{v}_t}\odot(\omega_t-\omega^*)\right\|^2 - \left\|\sqrt[4]{\hat{v}_t}\odot(\omega_{t+1}-\omega^*)\right\|^2\right]}{2\eta(1-\beta_1)} \\
&-\frac{\beta_1}{1-\beta_1}\left\langle\frac{\omega_t-\omega^*}{\sqrt{\eta}}, m_{t-1}\sqrt{\eta}\right\rangle + \frac{1}{2\eta(1-\beta_1)}\left\|\frac{\eta}{\sqrt[4]{\hat{v}_t}}\odot m_t\right\|^2 \\
\leq& \frac{\left[\left\|\sqrt[4]{\hat{v}_t}\odot(\omega_t-\omega^*)\right\|^2 - \left\|\sqrt[4]{\hat{v}_t}\odot(\omega_{t+1}-\omega^*)\right\|^2\right]}{2\eta(1-\beta_1)} \\
&+\frac{\beta_1}{1-\beta_1}\left[\frac{\|\omega_t-\omega^*\|^2}{2\eta} + \frac{m_{t-1}\odot m_{t-1}\eta}{2}\right] + \frac{\eta}{2(1-\beta_1)}\left\|\frac{1}{\sqrt[4]{\hat{v}_t}}\odot m_t\right\|^2.
\end{aligned} \tag{9}$$

From the strong convexity property of f_t in 4 in Assumption 1, we obtain

$$f_t(\omega_t) - f_t(\omega^*) \leq \langle \omega_t-\omega^*, \bigtriangledown f_t(\omega_t)\rangle - \frac{H}{2}\|\omega_t-\omega^*\|^2.$$

Using (9) in the above inequality and summing up over all time steps yields

$$\begin{aligned}
&\sum_{t=p}^{T+p}\left[f_t(\omega_t)-f_t(\omega^*)\right]\\
\leq&\sum_{t=p}^{T+p}\left\{\frac{\left[\left\|\sqrt[4]{\hat{v}_t}\odot(\omega_t-\omega^*)\right\|^2-\left\|\sqrt[4]{\hat{v}_t}\odot(\omega_{t+1}-\omega^*)\right\|^2\right]}{2\eta(1-\beta_1)}+\|\omega_t-\omega^*\|^2\left[\frac{\beta_1}{2\eta(1-\beta_1)}-\frac{H}{2}\right]\right.\\
&\left.+\frac{\eta}{2(1-\beta_1)}\left[\beta_1 m_{t-1}\odot m_{t-1}+\left\|\frac{1}{\sqrt[4]{\hat{v}_t}}\odot m_t\right\|^2\right]\right\}. \qquad (10)
\end{aligned}$$

We proceed by separating (10) into 3 parts and find upper bounds for each one of them. Considering the first part in (10), we have

$$\begin{aligned}
&\sum_{t=p}^{T+p}\frac{\left[\left\|\sqrt[4]{\hat{v}_t}\odot(\omega_t-\omega^*)\right\|^2-\left\|\sqrt[4]{\hat{v}_t}\odot(\omega_{t+1}-\omega^*)\right\|^2\right]}{2\eta(1-\beta_1)}\\
\leq&\frac{\left\|\sqrt[4]{\hat{v}_p}\odot(\omega_p-\omega^*)\right\|^2}{2\eta(1-\beta_1)}+\frac{1}{2\eta(1-\beta_1)}\sum_{t=p+1}^{T+p}\left(\left\|\sqrt[4]{\hat{v}_t}\odot(\omega_t-\omega^*)\right\|^2\right.\\
&\left.-\left\|\sqrt[4]{\hat{v}_{t-1}}\odot(\omega_t-\omega^*)\right\|^2\right)\\
=&\frac{1}{2\eta(1-\beta_1)}\left[\left\|\sqrt[4]{\hat{v}_p}\odot(\omega_p-\omega^*)\right\|^2+\sum_{t=p+1}^{T+p}\left(\sum_{j=1}^{d}\sqrt{\hat{v}_{t,j}}(\omega_{t,j}-\omega^{*,j})^2\right.\right.\\
&\left.\left.-\sum_{j=1}^{d}\sqrt{\hat{v}_{t-1,j}}(\omega_{t,j}-\omega^{*,j})^2\right)\right]\\
=&\frac{1}{2\eta(1-\beta_1)}\left[\left\|\sqrt[4]{\hat{v}_p}\odot(\omega_p-\omega^*)\right\|^2+\sum_{t=p+1}^{T+p}\left(\sum_{j=1}^{d}(\omega_{t,j}-\omega^{*,j})^2\left(\sqrt{\hat{v}_{t,j}}-\sqrt{\hat{v}_{t-1,j}}\right)\right)\right]. \qquad (11)
\end{aligned}$$

Since $\hat{v}_{t,j}$ is maximum of all $v_{t,j}$ for each j until the current time step, i.e. $\sqrt{\hat{v}_{t,j}}-\sqrt{\hat{v}_{t-1,j}}\geq 0$, by using 1 in Assumption 1, (11) can be further bounded as follows,

$$\begin{aligned}
&\sum_{t=p}^{T+p}\frac{\left[\left\|\sqrt[4]{\hat{v}_t}\odot(\omega_t-\omega^*)\right\|^2-\left\|\sqrt[4]{\hat{v}_t}\odot(\omega_{t+1}-\omega^*)\right\|^2\right]}{2\eta(1-\beta_1)}\\
\leq&\frac{1}{2\eta(1-\beta_1)}\left[\left\|\sqrt[4]{\hat{v}_p}\odot(\omega_p-\omega^*)\right\|^2+D_\infty^2\sum_{j=1}^{d}\sum_{t=p+1}^{T+p}\left(\sqrt{\hat{v}_{t,j}}-\sqrt{\hat{v}_{t-1,j}}\right)\right]\\
\leq&\frac{1}{2\eta(1-\beta_1)}\left[D_\infty^2\sum_{j=1}^{d}\sqrt{\hat{v}_{p,j}}+D_\infty^2\sum_{j=1}^{d}\sum_{t=p+1}^{T+p}\left(\sqrt{\hat{v}_{t,j}}-\sqrt{\hat{v}_{t-1,j}}\right)\right]\\
=&\frac{1}{2\eta(1-\beta_1)}D_\infty^2\sum_{j=1}^{d}\sqrt{\hat{v}_{p+T,j}}.
\end{aligned}$$

By the definition of $\hat{v}_t$ in step 6 in Algorithm 2, for any t and j, we have

$$v_{t,j}=(1-\beta_2)\sum_{i=1}^{t}\beta_2^{t-i}g_{i,j}^2\leq(1-\beta_2)G_\infty^2\sum_{i=1}^{t}\beta_2^{t-i}\leq G_\infty^2,$$

which in turn yields

$$\sum_{t=p}^{T+p}\frac{\left[\left\|\sqrt[4]{\hat{v}_t}\odot(\omega_t-\omega^*)\right\|^2-\left\|\sqrt[4]{\hat{v}_t}\odot(\omega_{t+1}-\omega^*)\right\|^2\right]}{2\eta(1-\beta_1)}\leq\frac{dD_\infty^2G_\infty}{2\eta(1-\beta_1)}=\mathcal{O}(\sqrt{T}). \qquad (12)$$

The last equality is due to the setting of the stepsize, i.e. $\eta = \frac{\eta_1}{\sqrt{T}}$. For the second term in (10), from the relationship between β_1 and H, we obtain

$$\frac{\beta_1}{1-\beta_1} \leq H\eta,$$

which in turn yields

$$\frac{\beta_1}{2\eta(1-\beta_1)} - \frac{H}{2} \leq 0. \tag{13}$$

Thus, (13) guarantees negativity of the second term in (10). For the third term in (10), by using Lemmas 1 and 2, we assert

$$\frac{\eta}{2(1-\beta_1)}\left[\beta_1 m_{t-1} \odot m_{t-1} + \left\| \frac{1}{\sqrt[4]{\hat{v}_t}} \odot m_t \right\|^2\right] \leq \mathcal{O}(\frac{1}{\sqrt{T}}) \cdot \mathcal{O}(T) = \mathcal{O}(\sqrt{T}). \tag{14}$$

The desired result follows directly from (10), (12), (13) and (14). □

F Regret with Rolling Window Analysis of nnOGD for Two-Layer ReLU Neural Network

The steps to study the regret with rolling window are as follows. Based on steps 4 - 6, we expand $\omega_{1,t+1}^T\omega_{2,t+1} - \omega_{1,*}^T\omega_{2,*}$ to establish a relationship for the sequence of the weight error $\omega_{1,t}^T\omega_{2,t} - \omega_{1,*}^T\omega_{2,*}$. In association with explicit formulas of gradients and condition 4 in Assumption 2, we obtain the loss function error $f_t(\omega_{1,t}, \omega_{2,t}) - f_t(\omega_{1,*}, \omega_{2,*})$. Meanwhile, all of the remaining terms are bounded due to condition 3 from Assumption 2. Combined with the fact that $\omega_{1,t}$ has a constant norm, the regret with rolling window bound of NNGD is achieved by applying the law of iterated expectation.

For a two-layer ReLU neural network, we first introduce $\mathcal{F}^t$ that records all previous iterates up until t.

Lemma 3. *If conditions 1 and 2 hold from Assumption 2, we have*

$$\mathbb{E}\left[l_t\left(\omega_{1,t}, \omega_{2,t}\right) \mid \mathcal{F}^t\right] = \frac{\rho^2}{2}(\omega_{1,t}^T\omega_{2,t}z^t - \omega_{1,*}^T\omega_{2,*}z^t)^2. \tag{15}$$

Proof of Lemma 3. Based on condition 2 in Assumption 2, we obtain

$$\begin{aligned}
&\mathbb{E}_{\sigma_1,\sigma_2}\left[f_t(\omega_{1,t}, \omega_{2,t}) \mid \mathcal{F}^t\right] \\
=&\frac{1}{2}\mathbb{E}_{\sigma_1}\left[\omega_{1,t}^T\sigma_1\left(\omega_{2,t}z^t\right) - y^t \mid \mathcal{F}^t\right] \cdot \mathbb{E}_{\sigma_2}\left[\omega_{1,t}^T\sigma_2\left(\omega_{2,t}z^t\right) - y^t \mid \mathcal{F}^t\right] \\
=&\frac{1}{2}\left(\rho\omega_{1,t}^T\omega_{2,t}z^t - y^t\right) \cdot \left(\rho\omega_{1,t}^T\omega_{2,t}z^t - y^t\right) = \frac{1}{2}\left(\rho\omega_{1,t}^T\omega_{2,t}z^t - y^t\right)^2 \\
=&\frac{\rho^2}{2}(\omega_{1,t}^T\omega_{2,t}z^t - \omega_{1,*}^T\omega_{2,*}z^t)^2.
\end{aligned}$$

On the other hand, we get

$$\begin{aligned}
&\mathbb{E}_{\sigma_1,\sigma_2}\left[f_t(\omega_{1,*}, \omega_{2,*}) \mid \mathcal{F}^t\right] \\
=&\frac{1}{2}\mathbb{E}_{\sigma_1}\left[\omega_{1,*}^T\sigma_1\left(\omega_{2,*}z^t\right) - y^t \mid \mathcal{F}^t\right] \cdot \mathbb{E}_{\sigma_2}\left[\omega_{1,*}^T\sigma_2\left(\omega_{2,*}z^t\right) - y^t \mid \mathcal{F}^t\right] \\
=&\frac{1}{2}\left(\mathbb{E}_{\sigma_1}\left[\omega_{1,*}^T\sigma_1\left(\omega_{2,*}z^t\right) \mid \mathcal{F}^t\right] - y^t\right) \cdot \left(\mathbb{E}_{\sigma_2}\left[\omega_{1,*}^T\sigma_2\left(\omega_{2,*}z^t\right) \mid \mathcal{F}^t\right] - y^t\right) \\
=&0,
\end{aligned}$$

which in turn yields

$$\begin{aligned}
&\mathbb{E}\left[l_t\left(\omega_{1,t}, \omega_{2,t}\right) \mid \mathcal{F}^t\right] \\
=&\mathbb{E}_{\sigma_1,\sigma_2}\left[f_t(\omega_{1,t}, \omega_{2,t}) \mid \mathcal{F}^t\right] - \mathbb{E}_{\sigma_1,\sigma_2}\left[f_t(\omega_{1,*}, \omega_{2,*}) \mid \mathcal{F}^t\right] \\
=&\frac{\rho^2}{2}(\omega_{1,t}^T\omega_{2,t}z^t - \omega_{1,*}^T\omega_{2,*}z^t)^2.
\end{aligned}$$

This completes the proof of the lemma. □

Lemma 4. *Under the conditions assumed in Theorem 3, we have*

$$\mathbb{E}_{\sigma_1,\sigma_2}\left[g_{1,t} \mid \mathcal{F}^t\right] = \rho^2\left(\omega_{1,t}^T\omega_{2,t}z^t - \omega_{1,*}^T\omega_{2,*}z^t\right)\omega_{2,t}z^t \tag{16}$$

$$\mathbb{E}_{\sigma_1,\sigma_2}\left[g_{2,t} \mid \mathcal{F}^t\right] = \rho^2\left(\omega_{1,t}^T\omega_{2,t}z^t - \omega_{1,*}^T\omega_{2,*}z^t\right)\omega_{1,t}\left(z^t\right)^T. \tag{17}$$

Proof of Lemma 4. From [nn-updates], we have

$$\begin{aligned}
&\mathbb{E}_{\sigma_1,\sigma_2}\left[g_{1,t} \mid \mathcal{F}^t\right] \\
=&\mathbb{E}_{\sigma_1,\sigma_2}\left[\bigtriangledown_{\omega_1}\left(\frac{1}{2}\left(\omega_{1,t}^T\sigma_1\left(\omega_{2,t}z^t\right)-y^t\right)\left(\omega_{1,t}^T\sigma_2\left(\omega_{2,t}z^t\right)-y^t\right)\right) \mid \mathcal{F}^t\right] \\
=&\mathbb{E}_{\sigma_1,\sigma_2}\left[\left(\omega_{1,t}^T\sigma_1\left(\omega_{2,t}z^t\right)-y^t\right)\sigma_2\left(\omega_{2,t}z^t\right) \mid \mathcal{F}^t\right] \\
=&\mathbb{E}_{\sigma_1}\left[\omega_{1,t}^T\sigma_1\left(\omega_{2,t}z^t\right)-\omega_{1,*}^T\sigma_1\left(\omega_{2,*}z^t\right) \mid \mathcal{F}^t\right]\mathbb{E}_{\sigma_2}\left[\sigma_2\left(\omega_{2,t}z^t\right) \mid \mathcal{F}^t\right] \\
=&\rho\left(\omega_{1,t}^T\omega_{2,t}z^t-\omega_{1,*}^T\omega_{2,*}z^t\right)\rho\omega_{2,t}z^t=\rho^2\left(\omega_{1,t}^T\omega_{2,t}z^t-\omega_{1,*}^T\omega_{2,*}z^t\right)\omega_{2,t}z^t.
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\mathbb{E}_{\sigma_1,\sigma_2}\left[g_{2t} \mid \mathcal{F}^t\right] \\
=&\mathbb{E}_{\sigma_1,\sigma_2}\left[\bigtriangledown_{\omega_2}\left(\frac{1}{2}\left(\omega_{1,t}^T\sigma_1\left(\omega_{2,t}z^t\right)-y^t\right)\left(\omega_{1,t}^T\sigma_2\left(\omega_{2,t}z^t\right)-y^t\right)\right) \mid \mathcal{F}^t\right] \\
=&\mathbb{E}_{\sigma_1,\sigma_2}\left[\left(\omega_{1,t}^T\sigma_1\left(\omega_{2,t}z^t\right)-y^t\right)\omega_{1,t}\left(\sigma_2(z^t)\right)^T \mid \mathcal{F}^t\right] \\
=&\mathbb{E}_{\sigma_1}\left[\omega_{1,t}^T\sigma_1\left(\omega_{2,t}z^t\right)-\omega_{1,*}^T\sigma_1\left(\omega_{2,*}z^t\right) \mid \mathcal{F}^t\right]\mathbb{E}_{\sigma_2}\left[\omega_{1,t}\left(\sigma_2(z^t)\right)^T \mid \mathcal{F}^t\right] \\
=&\rho\left(\omega_{1,t}^T\omega_{2,t}z^t-\omega_{1,*}^T\omega_{2,*}z^t\right)\rho\omega_{1,t}\left(z^t\right)^T \\
=&\rho^2\left(\omega_{1,t}^T\omega_{2,t}z^t-\omega_{1,*}^T\omega_{2,*}z^t\right)\omega_{1,t}\left(z^t\right)^T.
\end{aligned}$$

This completes the proof of the lemma. □

Proof of Theorem 3

Proof. First, based on the update step 5 and 6 in Algorithm 3, we obtain

$$\begin{aligned}
&\mathbb{E}_{\sigma_1,\sigma_2}\left[\left\|\omega_{1,t+1}^T\omega_{2,t+1}-\omega_{1,*}^T\omega_{2,*}\right\|^2 \mid \mathcal{F}^t\right]=\mathbb{E}_{\sigma_1,\sigma_2}\left[\left\|\omega_{2,t+1}^T\omega_{1,t+1}-\omega_{2,*}^T\omega_{1,*}\right\|^2 \mid \mathcal{F}^t\right] \\
=&\mathbb{E}_{\sigma_1,\sigma_2}\left[\left\|\left(\omega_{2,t}-\eta g_{2,t}\right)^T\left(\omega_{1,t}-\eta g_{1,t}\right)-\omega_{2,*}^T\omega_{1,*}\right\|^2 \mid \mathcal{F}^t\right] \\
=&\mathbb{E}_{\sigma_1,\sigma_2}\left[\left\|\omega_{2,t}^T\omega_{1,t}-\omega_{2,*}^T\omega_{1,*}-\eta\left(g_{2,t}^T\omega_{1,t}+\omega_{2,t}^Tg_{1,t}\right)+\eta^2g_{1,t}^Tg_{2,t}\right\|^2 \mid \mathcal{F}^t\right] \\
=&\mathbb{E}_{\sigma_1,\sigma_2}\left[\left\|\omega_{2,t}^T\omega_{1,t}-\omega_{2,*}^T\omega_{1,*}\right\|^2-2\eta\left\langle\omega_{2,t}^T\omega_{1,t}-\omega_{2,*}^T\omega_{1,*},g_{2,t}^T\omega_{1,t}+\omega_{2,t}^Tg_{1,t}\right\rangle\right. \\
&\quad\left.+\eta^2\left(2\left\langle\omega_{2,t}^T\omega_{1,t}-\omega_{2,*}^T\omega_{1,*},g_{2,t}^Tg_{1,t}\right\rangle+\left\|\eta g_{1,t}^Tg_{2,t}-\left(g_{2,t}^T\omega_{1,t}+\omega_{2,t}^Tg_{1,t}\right)\right\|^2\right) \mid \mathcal{F}^t\right] \\
=&\left\|\omega_{2,t}^T\omega_{1,t}-\omega_{2,*}^T\omega_{1,*}\right\|^2-2\eta\left\langle\omega_{2,t}^T\omega_{1,t}-\omega_{2,*}^T\omega_{1,*},\mathbb{E}_{\sigma_1,\sigma_2}\left[g_{2,t}^T \mid \mathcal{F}^t\right]\omega_{1,t}\right. \\
&\quad\left.+\omega_{2,t}^T\mathbb{E}_{\sigma_1,\sigma_2}\left[g_{1,t} \mid \mathcal{F}^t\right]\right\rangle+\eta^2\mathbb{E}_{\sigma_1,\sigma_2}\left[2\left\langle\omega_{2,t}^T\omega_{1,t}-\omega_{2,*}^T\omega_{1,*},g_{2,t}^Tg_{1,t}\right\rangle\right. \\
&\quad\left.+\left\|\eta g_{1,t}^Tg_{2,t}-\left(g_{2,t}^T\omega_{1,t}+\omega_{2,t}^Tg_{1,t}\right)\right\|^2 \mid \mathcal{F}^t\right]. \qquad (18)
\end{aligned}$$

By Lemma 4 we conclude that $\mathbb{E}\left[\|g_{1,t}\| \mid \mathcal{F}^t\right]$ and $\mathbb{E}\left[\|g_{2,t}\| \mid \mathcal{F}^t\right]$ are bounded due to 3 in Assumption 2, which in turn yields

$$\begin{aligned}
&\mathbb{E}_{\sigma_1,\sigma_2}\left[2\left\langle\omega_{2,t}^T\omega_{1,t}-\omega_{2,*}^T\omega_{1,*},g_{2,t}^Tg_{1,t}\right\rangle+\left\|\eta g_{1,t}^Tg_{2,t}-\left(g_{2,t}^T\omega_{1,t}+\omega_{2,t}^Tg_{1,t}\right)\right\|^2 \mid \mathcal{F}^t\right] \\
&\quad\le\left\|\omega_{2,t}^T\omega_{1,t}-\omega_{2,*}^T\omega_{1,*}\right\|^2\cdot\mathbb{E}_{\sigma_1,\sigma_2}\left[\left\|g_{2,t}^Tg_{1,t}\right\|^2 \mid \mathcal{F}^t\right] \\
&\qquad\qquad+\mathbb{E}_{\sigma_1,\sigma_2}\left[\left\|\eta g_{1,t}^Tg_{2,t}-\left(g_{2,t}^T\omega_{1,t}+\omega_{2,t}^Tg_{1,t}\right)\right\|^2 \mid \mathcal{F}^t\right] \\
&\quad\le M_1, \qquad (19)
\end{aligned}$$

where M_1 is a fixed positive number. The first inequality comes from the Cauchy-Schwarz inequality and the second inequality is due to the boundedness of $\omega_{1,t}$, $\omega_{2,t}$, $\omega_{1,*}$, $\omega_{2,*}$, $g_{1,t}$, $g_{2,t}$ and η. Inserting (19) into (18) gives

$$\begin{aligned}
&\left\langle\omega_{2,t}^T\omega_{1,t}-\omega_{2,*}^T\omega_{1,*},\mathbb{E}_{\sigma_1,\sigma_2}\left[g_{2,t}^T \mid \mathcal{F}^t\right]\omega_{1,t}\right\rangle+\left\langle\omega_{2,t}^T\omega_{1,t}-\omega_{2,*}^T\omega_{1,*},\omega_{2,t}^T\mathbb{E}_{\sigma_1,\sigma_2}\left[g_{1,t} \mid \mathcal{F}^t\right]\right\rangle \\
=&\left\langle\omega_{2,t}^T\omega_{1,t}-\omega_{2,*}^T\omega_{1,*},\mathbb{E}_{\sigma_1,\sigma_2}\left[g_{2,t}^T \mid \mathcal{F}^t\right]\omega_{1,t}+\omega_{2,t}^T\mathbb{E}_{\sigma_1,\sigma_2}\left[g_{1,t} \mid \mathcal{F}^t\right]\right\rangle \\
\le&\frac{\left\|\omega_{2,t}^T\omega_{1,t}-\omega_{2,*}^T\omega_{1,*}\right\|^2-\mathbb{E}_{\sigma_1,\sigma_2}\left[\left\|\omega_{1,t+1}^T\omega_{2,t+1}-\omega_{1,*}^T\omega_{2,*}\right\|^2 \mid \mathcal{F}^t\right]}{2\eta}+\frac{\eta M_1}{2}. \qquad (20)
\end{aligned}$$

Using (16) yields

$$\begin{aligned}
&\left\langle \omega_{2,t}^T\omega_{1,t} - \omega_{2,*}^T\omega_{1,*}, \mathbb{E}_{\sigma_1,\sigma_2}\left[g_{2,t}^T \mid \mathcal{F}^t\right]\omega_{1,t}\right\rangle \\
=&\left\langle \omega_{2,t}^T\omega_{1,t} - \omega_{2,*}^T\omega_{1,*}, \rho^2\left(\omega_{1,t}^T\omega_{2,t}z^t - \omega_{1,*}^T\omega_{2,*}z^t\right)z^t\omega_{1,t}^T\omega_{1,t}\right\rangle \\
=&\rho^2\left(\omega_{1,t}^T\omega_{2,t}z^t - \omega_{1,*}^T\omega_{2,*}z^t\right)\left(\omega_{1,t}^T\omega_{2,t} - \omega_{1,*}^T\omega_{2,*}\right)z^t\left\|\omega_{1,t}\right\|^2 \\
=&\rho^2\left(\omega_{1,t}^T\omega_{2,t}z^t - \omega_{1,*}^T\omega_{2,*}z^t\right)^2\left\|\omega_{1,t}\right\|^2 = \mathbb{E}\left[l_t\left(\omega_{1,t},\omega_{2,t}\right) \mid \mathcal{F}^t\right]\cdot 2\left\|\omega_{1,t}\right\|^2.
\end{aligned} \tag{21}$$

The last equality follows from (15) in Lemma 3. Then, we have

$$\begin{aligned}
&\left|\left\langle \omega_{2,t}^T\omega_{1,t} - \omega_{2,*}^T\omega_{1,*}, \omega_{2,t}^T\,\mathbb{E}_{\sigma_1,\sigma_2}\left[g_{1,t} \mid \mathcal{F}^t\right]\right\rangle\right| \\
=&\left|\left\langle \omega_{2,t}^T\omega_{1,t} - \omega_{2,*}^T\omega_{1,*}, \omega_{2,t}^T\rho^2\left(\omega_{1,t}^T\omega_{2,t}z^t - \omega_{1,*}^T\omega_{2,*}z^t\right)\omega_{2,t}z^t\right\rangle\right| \\
=&\left|\rho^2\left(\omega_{1,t}^T\omega_{2,t}z^t - \omega_{1,*}^T\omega_{2,*}z^t\right)\left(\omega_{1,t}^T\omega_{2,t} - \omega_{1,*}^T\omega_{2,*}\right)\omega_{2,t}^T\omega_{2,t}z^t\right| \\
\leq&\rho^2\left(\omega_{1,t}^T\omega_{2,t}z^t - \omega_{1,*}^T\omega_{2,*}z^t\right)^2\frac{\left\|\omega_{1,t}^T\omega_{2,t} - \omega_{1,*}^T\omega_{2,*}\right\|\left\|\omega_{2,t}^T\omega_{2,t}\right\|\left\|z^t\right\|}{\left|\left(\omega_{1,t}^T\omega_{2,t} - \omega_{1,*}^T\omega_{2,*}\right)z^t\right|} \\
\leq&\rho^2\left(\omega_{1,t}^T\omega_{2,t}z^t - \omega_{1,*}^T\omega_{2,*}z^t\right)^2\left\|\omega_{2,t}^T\omega_{2,t}\right\|\frac{\left\|\omega_{1,t}^T\omega_{2,t} - \omega_{1,*}^T\omega_{2,*}\right\|\left\|z^t\right\|}{\left|\left(\omega_{1,t}^T\omega_{2,t} - \omega_{1,*}^T\omega_{2,*}\right)z^t\right|} \\
\leq&\rho^2\left(\omega_{1,t}^T\omega_{2,t}z^t - \omega_{1,*}^T\omega_{2,*}z^t\right)^2\frac{\alpha}{\cos\left(\epsilon\right)} = \mathbb{E}\left[l_t\left(\omega_{1,t},\omega_{2,t}\right) \mid \mathcal{F}^t\right]\cdot\frac{2\alpha}{\cos\left(\epsilon\right)}.
\end{aligned} \tag{22}$$

Note that $\left\|\omega_{2,t}^T\omega_{2,t}\right\| = \sigma_{max}(\omega_{2,t}^T) \leq \left\|\omega_{2,t}^T\right\|_F \leq \alpha$ by 3 in Assumption 2. If $\left(\omega_{1,t}^T\omega_{2,t} - \omega_{1,*}^T\omega_{2,*}\right)z^t = 0$, then the inequality holds trivially. Using (20), (21) and (22) we obtain

$$\begin{aligned}
&\mathbb{E}_{\sigma_1,\sigma_2}\left[l_t\left(\omega_{1,t},\omega_{2,t}\right) \mid \mathcal{F}^t\right]\cdot 2\left(\left\|\omega_{1,t}\right\|^2 - \frac{\alpha}{\cos(\epsilon)}\right) \\
\leq&\frac{\left\|\omega_{2,t}^T\omega_{1,t} - \omega_{2,*}^T\omega_{1,*}\right\|^2 - \mathbb{E}_{\sigma_1,\sigma_2}\left[\left\|\omega_{1,t+1}^T\omega_{2,t+1} - \omega_{1,*}^T\omega_{2,*}\right\|^2 \mid \mathcal{F}^t\right]}{2\eta} + \frac{\eta M_1}{2}.
\end{aligned} \tag{23}$$

From update step 5 we notice that $\left\|\omega_{1,t}\right\|^2 = \frac{1}{2} + \xi_1 = \frac{1}{2} + \frac{\alpha}{\cos(\epsilon)}$, thus, (23) could be further simplified as

$$\begin{aligned}
\mathbb{E}_{\sigma_1,\sigma_2}\left[l_t\left(\omega_{1,t},\omega_{2,t}\right) \mid \mathcal{F}^t\right] \leq& \frac{\left\|\omega_{2,t}^T\omega_{1,t} - \omega_{2,*}^T\omega_{1,*}\right\|^2 - \mathbb{E}_{\sigma_1,\sigma_2}\left[\left\|\omega_{1,t+1}^T\omega_{2,t+1} - \omega_{1,*}^T\omega_{2,*}\right\|^2 \mid \mathcal{F}^t\right]}{2\eta} \\
&+ \frac{\eta M_1}{2}.
\end{aligned}$$

Applying the law of iterated expectation implies

$$\mathbb{E}\left[l_t\left(\omega_{1,t},\omega_{2,t}\right)\right] \leq \frac{\mathbb{E}\left[\left\|\omega_{2,t}^T\omega_{1,t} - \omega_{2,*}^T\omega_{1,*}\right\|^2\right] - \mathbb{E}\left[\left\|\omega_{1,t+1}^T\omega_{2,t+1} - \omega_{1,*}^T\omega_{2,*}\right\|^2\right]}{2\eta} + \frac{\eta M_1}{2}$$

By summing up all differences, we obtain

$$\begin{aligned}
\sum_{t=p}^{T+p}\mathbb{E}\left[l_t\left(\omega_{1,t},\omega_{2,t}\right)\right] \leq& \frac{1}{2}\sum_{t=p}^{T+p}\frac{\mathbb{E}\left[\left\|\omega_{2,t}^T\omega_{1,t} - \omega_{2,*}^T\omega_{1,*}\right\|^2\right] - \mathbb{E}\left[\left\|\omega_{1,t+1}^T\omega_{2,t+1} - \omega_{1,*}^T\omega_{2,*}\right\|^2\right]}{\eta} \\
&+ \frac{M_1}{2}\eta T \\
=& \frac{1}{2}\frac{\mathbb{E}\left[\left\|\omega_{2,p}^T\omega_{1,p} - \omega_{2,*}^T\omega_{1,*}\right\|^2\right] - \mathbb{E}\left[\left\|\omega_{1,p+T+1}^T\omega_{2,p+T+1} - \omega_{1,*}^T\omega_{2,*}\right\|^2\right]}{\eta} + \frac{M_1}{2}\eta T \\
=& \mathcal{O}(\sqrt{T}).
\end{aligned} \tag{24}$$

The last equality uses 3 from Assumption 2 and the definition of $\eta = \frac{\eta_1}{\sqrt{T}}$. The desired result in Theorem 3 follows directly from (24) since it holds for any p.

□

G Regret with Rolling Window Analyses of nnAdam for Two-Layer ReLU Neural Network

The structure of the technical proof is similar to that of Theorem 3. We first establish a relationship for the sequence of the weight error $\omega_{1,t}^T\omega_{2,t} - \omega_{1,*}^T\omega_{2,*}$ by multiplying $\sqrt[4]{\hat{v}_{2,t}}$. Then, using the definitions of β's, λ's and γ's, we bound all the terms without the stepsize by constants except those which potentially can contribute to the loss function. To this end, we obtain a relationship between the weight error $\omega_{1,t}^T\omega_{2,t} - \omega_{1,*}^T\omega_{2,*}$ and the loss function. Finally, combined with $\hat{v}_{2t}$ in step 5 and the law of iterated expectation, we are able to argue $\mathcal{O}(\sqrt{T})$ regret with rolling window for NNADAM.

Lemma 5. *In Algorithm 4, given $\omega_{2,t}, \omega_{1,t}, \omega_{2,*}, \omega_{1,*}$ and $\hat{v}_{2,t}$, there exists a bounded matrix $\tilde{v}_{2,t}$ such that*

$$\left(\sqrt{\hat{v}_{2,t}} \odot \omega_{2,t}\right)^T \omega_{1,t} - \left(\sqrt{\hat{v}_{2,t}} \odot \omega_{2,*}\right)^T \omega_{1,*} = \left(\sqrt{\tilde{v}_{2,t}}\right)^T \left(\omega_{2,t}^T\omega_{1,t} - \omega_{2,*}^T\omega_{1,*}\right), \tag{25}$$

where $\odot$ is an element-wise multiplication operation.

Proof of Lemma 5. From step 5 in Algorithm 4, v_{2t} is a matrix with same value in the same column, which in turn yields

$$\left(\sqrt{\hat{v}_{2,t}} \odot \omega_{2,t}\right)^T \omega_{1t} = \left(\sqrt{\tilde{v}_{2,t}}\right)^T \omega_{2,t}^T\omega_{1,t},$$

where $\tilde{v}_{2,t}$ is a diagonal matrix with $\tilde{v}_{2,t} = \operatorname{diag}\left(\left[\hat{v}_{2,t}\right]_{1,:}\right)$, and $\left[\hat{v}_{2,t}\right]_{1,:}$ is the 1^{st} row of matrix $\hat{v}_{2,t}$. Applying the same argument for $\left(\sqrt{\hat{v}_{2,t}} \odot \omega_{2,*}\right)^T \omega_{1,*}$ yields (25). Next, let us show that $\tilde{v}_{2,t}$ is bounded. It is sufficient to show that $\hat{v}_{2,t}$ is bounded. From $\hat{v}_{2t}$ in step 5 and $\dot{v}_{2,t}$ in [nn-adam-updates], we conclude that

$$\hat{v}_{2,t} \leq \max\left(v_{2,1}, v_{2,2}, \cdots, v_{2,t}\right).$$

Therefore, it is sufficient to show that $\dot{v}_{2,t}$ is bounded for all t. For each entry in the matrix, since $\left|\left[g_{2,t}\right]_{ij}\right| \leq G_{2,\infty}$, we obtain

$$\left|\left[\dot{v}_{2,t}\right]_{ik}\right| = \left|(1-\beta_{22})\sum_{j=1}^{t}\beta_{22}^{t-j}\left(\max_p \left[g_{2,j}\right]_{pk}^2\right)\right| \leq \left|(1-\beta_{22})\sum_{j=1}^{t}\beta_{22}^{t-j}G_{2,\infty}^2\right| \leq G_{2,\infty}^2. \tag{26}$$

By combining with the fact that g_2 is bounded due to [nn-adam-updates] and the boundedness of $\omega_{1,t}, \omega_{2,t}, z^t$ and y^t from condition 3 in Assumption 2, Lemma 5 follows. □

Lemma 6. *In Algorithm 4, given $m_{1,t-1}, m_{1,t}, \hat{v}_{1,t} \in \mathbb{R}^n$ and $m_{2,t}, \hat{v}_{2,t} \in \mathbb{R}^{n\times d}$ for any t, and $\beta_{111}, \beta_{121}, \beta_{21}$ and β_{22} are constants between 0 and 1 such that $\lambda_1 := \frac{\beta_{111}}{\beta_{21}} < 1$ and $\lambda_2 := \frac{\beta_{121}}{\beta_{22}} < 1$, then*

$$\left\|\frac{1}{\sqrt{\hat{v}_{1,t}}} \odot m_{1,t-1}\right\|^2 \leq \frac{n}{(1-\beta_{111})(1-\beta_{21})(1-\lambda_1)} \tag{27}$$

$$\left\|\frac{1}{\sqrt{\hat{v}_{1,t}}} \odot m_{1,t}\right\|^2 \leq \frac{n}{(1-\beta_{111})(1-\beta_{21})(1-\lambda_1)} \tag{28}$$

$$\left\|\frac{1}{\sqrt{\hat{v}_{2,t}}} \odot m_{2,t}\right\|^2 \leq \frac{nd}{(1-\beta_{121})(1-\beta_{21})(1-\lambda_2)} \tag{29}$$

$$\left\|\frac{1}{\sqrt[4]{\hat{v}_{2,t}}} \odot m_{2,t}\right\|^2 \leq \frac{ndG_{2,\infty}}{(1-\beta_{121})\sqrt{1-\beta_{21}}(1-\lambda_2)}. \tag{30}$$

Proof of Lemma 6. Based on steps in [nn-adam-updates] - step 5 in Algorithm 4, we obtain

$$m_{1,t} = \sum_{j=1}^{t}\left[(1-\beta_{11j})\prod_{k=1}^{t-j}\beta_{11(t-k+1)}g_{1,j}\right] \tag{31}$$

$$m_{2,t} = \sum_{j=1}^{t}\left[(1-\beta_{12j})\prod_{k=1}^{t-j}\beta_{12(t-k+1)}g_{2,j}\right] \tag{32}$$

$$\hat{v}_{1,t} \geq (1-\beta_{21})\sum_{j=1}^{t}\beta_{21}^{t-j}g_{1,j} \odot g_{1,j} \tag{33}$$

$$\hat{v}_{2,t} \geq (1-\beta_{22})\sum_{j=1}^{t}\beta_{22}^{t-j}g_{2,j} \odot g_{2,j}. \tag{34}$$

Then, combining (31) and (33) yields

$$\begin{aligned}
\left\| \frac{1}{\sqrt{\hat{v}_{1,t}}} \odot m_{1,t} \right\|^2 &\leq \sum_{i=1}^{n} \frac{\left(\sum_{j=1}^{t} \left[(1-\beta_{11j}) \prod_{k=1}^{t-j} \beta_{11(t-k+1)} \left[g_{1,j} \right]_i \right] \right)^2}{\left((1-\beta_{21}) \sum_{j=1}^{t} \beta_{21}^{t-j} \left[g_{1,j} \right]_i^2 \right)} \\
&\leq \sum_{i=1}^{n} \frac{\left(\sum_{j=1}^{t} \prod_{k=1}^{t-j} \beta_{11(t-k+1)} \left[g_{1,j} \right]_i \right)^2}{\left((1-\beta_{21}) \sum_{j=1}^{t} \beta_{21}^{t-j} \left[g_{1,j} \right]_i^2 \right)} \\
&\leq \sum_{i=1}^{n} \frac{\left(\sum_{j=1}^{t} \prod_{k=1}^{t-j} \beta_{11(t-k+1)} \right) \left(\sum_{j=1}^{t} \prod_{k=1}^{t-j} \beta_{11(t-k+1)} \left[g_{1,j} \right]_i^2 \right)}{\left((1-\beta_{21}) \sum_{j=1}^{t} \beta_{21}^{t-j} \left[g_{1,j} \right]_i^2 \right)} \\
&\leq \sum_{i=1}^{n} \frac{\left(\sum_{j=1}^{t} \beta_{111}^{t-j} \right) \left(\sum_{j=1}^{t} \beta_{111}^{t-j} \left[g_{1,j} \right]_i^2 \right)}{\left((1-\beta_{21}) \sum_{j=1}^{t} \beta_{21}^{t-j} \left[g_{1,j} \right]_i^2 \right)} \\
&\leq \frac{1}{(1-\beta_{111})(1-\beta_{21})} \sum_{i=1}^{n} \sum_{j=1}^{t} \frac{\beta_{111}^{t-j} \left[g_{1,j} \right]_i^2}{\beta_{21}^{t-j} \left[g_{1,j} \right]_i^2} \\
&\leq \frac{1}{(1-\beta_{111})(1-\beta_{21})} \sum_{i=1}^{n} \sum_{j=1}^{t} \lambda_1^{t-j} \\
&\leq \frac{n}{(1-\beta_{111})(1-\beta_{21})(1-\lambda_1)}.
\end{aligned}$$

The first inequality follows from the definition of $\hat{v}_{1,t}$, which is maximum of all $v_{1,t}$ until the current time step. The third inequality follows from the Cauchy-Schwarz inequality and the forth inequality uses the fact that $\beta_{11t} \leq \beta_{111}$ for any t. Applying the same argument to $\left\| \frac{1}{\sqrt{\hat{v}_{2,t}}} \odot m_{2,t} \right\|^2$ implies (29). Then, applying the fact that $\hat{v}_{1,t} \geq v_{1,\hat{t-1}}$ yields

$$\left\| \frac{1}{\sqrt{\hat{v}_{1,t}}} \odot m_{1,t-1} \right\|^2 \leq \left\| \frac{1}{\sqrt{\hat{v}_{1,t-1}}} \odot m_{1,t-1} \right\|^2 \leq \frac{n}{(1-\beta_{111})(1-\beta_{21})(1-\lambda_1)},$$

where the last inequality follows from (28). Lastly, $\lambda_2 = \frac{\beta_{121}}{\beta_{22}} < 1$ implies $\frac{\beta_{121}}{\sqrt{\beta_{22}}} < \lambda_2 < 1$. By combining (32) and (34), we get

$$\begin{aligned}
\left\| \frac{1}{\sqrt[4]{\hat{v}_{2,t}}} \odot m_{2,t} \right\|^2 &\leq \sum_{p=1}^{n} \sum_{q=1}^{d} \frac{\left(\sum_{j=1}^{t} \left[(1-\beta_{12j}) \prod_{k=1}^{t-j} \beta_{12(t-k+1)} \left[g_{2,j} \right]_{pq} \right] \right)^2}{\sqrt{\left((1-\beta_{21}) \sum_{j=1}^{t} \beta_{22}^{t-j} \left[g_{2,j} \right]_{pq}^2 \right)}} \\
&\leq \sum_{p=1}^{n} \sum_{q=1}^{d} \frac{\left(\sum_{j=1}^{t} \left[\prod_{k=1}^{t-j} \beta_{12(t-k+1)} \left[g_{2,j} \right]_{pq} \right] \right)^2}{\sqrt{\left((1-\beta_{21}) \sum_{j=1}^{t} \beta_{22}^{t-j} \left[g_{2,j} \right]_{pq}^2 \right)}} \\
&\leq \sum_{p=1}^{n} \sum_{q=1}^{d} \frac{\left(\sum_{j=1}^{t} \prod_{k=1}^{t-j} \beta_{12(t-k+1)} \right) \left(\sum_{j=1}^{t} \prod_{k=1}^{t-j} \beta_{12(t-k+1)} \left[g_{2,j} \right]_{pq}^2 \right)}{\sqrt{\left((1-\beta_{21}) \sum_{j=1}^{t} \beta_{22}^{t-j} \left[g_{2,j} \right]_{pq}^2 \right)}} \\
&\leq \sum_{p=1}^{n} \sum_{q=1}^{d} \frac{\left(\sum_{j=1}^{t} \beta_{121}^{t-j} \right) \left(\sum_{j=1}^{t} \beta_{121}^{t-j} \left[g_{2,j} \right]_{pq}^2 \right)}{\sqrt{\left((1-\beta_{21}) \sum_{j=1}^{t} \beta_{22}^{t-j} \left[g_{2,j} \right]_{pq}^2 \right)}} \\
&\leq \frac{1}{(1-\beta_{121})\sqrt{1-\beta_{21}}} \sum_{p=1}^{n} \sum_{q=1}^{d} \sum_{j=1}^{t} \frac{\beta_{121}^{t-j} \left[g_{2,j} \right]_{pq}^2}{\sqrt{\beta_{22}^{t-j} \left[g_{2,j} \right]_{pq}^2}} \\
&\leq \frac{1}{(1-\beta_{121})\sqrt{1-\beta_{21}}} \sum_{p=1}^{n} \sum_{q=1}^{d} \sum_{j=1}^{t} \lambda_2^{t-j} \left| \left[g_{2,j} \right]_{pq} \right| \\
&\leq \frac{ndG_{2,\infty}}{(1-\beta_{121})\sqrt{1-\beta_{21}}\,(1-\lambda_2)}.
\end{aligned}$$

□

Proof of Theorem 4

Proof. Now, let us multiply $\left\|\omega_{2,t+1}^T\omega_{1,t+1}-\omega_{2*}^T\omega_{1*}\right\|^2$ by $\sqrt{\hat{v}_{2,t}}$, then take expectation given all records until time t. Then, from steps 4 - 7, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\left\|\left(\sqrt[4]{\hat{v}_{2,t}}\odot\omega_{2,t+1}\right)^T\omega_{1,t+1}-\left(\sqrt[4]{\hat{v}_{2,t}}\odot\omega_{2,*}\right)^T\omega_{1,*}\right\|^2\mid\mathcal{F}^t\right] &(35)\\
=&\mathbb{E}\left[\left\|\left(\sqrt[4]{\hat{v}_{2,t}}\odot\left(\omega_{2,t}-\frac{\eta}{\sqrt{\hat{v}_{2,t}}}\odot m_{2,t}\right)\right)^T\left(\omega_{1,t}-\frac{\eta}{\sqrt{\hat{v}_{1,t}}}\odot m_{1,t}\right)\right.\right.\\
&\left.\left.-\left(\sqrt[4]{\hat{v}_{2,t}}\odot\omega_{2,*}\right)^T\omega_{1,*}\right\|^2\mid\mathcal{F}^t\right]\\
=&\mathbb{E}\left[\left\|\left(\sqrt[4]{\hat{v}_{2,t}}\odot\omega_{2,t}\right)^T\omega_{1,t}-\left(\sqrt[4]{\hat{v}_{2,t}}\odot\omega_{2,*}\right)^T\omega_{1,*}\right\|^2\mid\mathcal{F}^t\right]\\
&-2\,\mathbb{E}\left[\left\langle\left(\sqrt{\hat{v}_{2,t}}\odot\omega_{2,t}\right)^T\omega_{1,t}-\left(\sqrt{\hat{v}_{2,t}}\odot\omega_{2,*}\right)^T\omega_{1*},\omega_{2,t}^T\frac{\eta}{\sqrt{\hat{v}_{1,t}}}\odot m_{1,t}\right\rangle\mid\mathcal{F}^t\right] &(36)\\
&-2\,\mathbb{E}\left[\left\langle\left(\sqrt{\hat{v}_{2,t}}\odot\omega_{2,t}\right)^T\omega_{1,t}-\left(\sqrt{\hat{v}_{2,t}}\odot\omega_{2,*}\right)^T\omega_{1,*},\left(\frac{\eta}{\sqrt{\hat{v}_{2,t}}}\odot m_{2,t}\right)^T\omega_{1,t}\right\rangle\mid\mathcal{F}^t\right] &(37)\\
&+2\eta^2\,\mathbb{E}\left[\left\langle\left(\sqrt{\hat{v}_{2,t}}\odot\omega_{2,t}\right)^T\omega_{1,t}-\left(\sqrt{\hat{v}_{2,t}}\odot\omega_{2,*}\right)^T\omega_{1,*},\right.\right.\\
&\qquad\left.\left.\left(\frac{\eta}{\sqrt{\hat{v}_{2,t}}}\odot m_{2,t}\right)^T\left(\frac{\eta}{\sqrt{\hat{v}_{1,t}}}\odot m_{1,t}\right)\right\rangle\mid\mathcal{F}^t\right] &(38)\\
&+\eta^2\,\mathbb{E}\left[\left\|\left(\sqrt[4]{\hat{v}_{2,t}}\odot\omega_{2,t}\right)^T\left(\frac{\eta}{\sqrt{\hat{v}_{1,t}}}\odot m_{1,t}\right)+\left(\sqrt[4]{\hat{v}_{2,t}}\odot\left(\frac{\eta}{\sqrt{\hat{v}_{2,t}}}\odot m_{2,t}\right)\right)^T\omega_{1,t}+\right.\right.\\
&\qquad\left.\left.\left(\sqrt[4]{\hat{v}_{2,t}}\odot\left(\frac{\eta}{\sqrt{\hat{v}_{2,t}}}\odot m_{2,t}\right)\right)^T\left(\frac{\eta}{\sqrt{\hat{v}_{1,t}}}\odot m_{1,t}\right)\right\|^2\mid\mathcal{F}^t\right]. &(39)
\end{aligned}$$

Let us first consider the expectations in (38) and (39). From (26), we conclude that $\hat{v}_{2,t}$ is bounded. Similarly, given $\beta_{11t}=\beta_{111}\gamma_1^t$ and $\beta_{12t}=\beta_{121}\gamma_2^t$ with $0<\gamma_1,\gamma_2<1$, for each entry, we attain

$$\begin{aligned}
\left|[m_{1,t}]_i\right|&\le\left|(1-\beta_{111})\sum_{j=1}^t\beta_{111}^{t-j}\left[g_{1,j}\right]_i\right|\le\max_j\left|[g_{1,j}]_i\right|\\
\left|[m_{2,t}]_{ik}\right|&\le\left|(1-\beta_{121})\sum_{j=1}^t\beta_{121}^{t-j}\left[g_{2,j}\right]_{ik}\right|\le\max_j\left|[g_{2,j}]_{ik}\right|.
\end{aligned}$$

Since $\left\|\frac{1}{\sqrt{\hat{v}_{1,t}}}\odot m_{1,t}\right\|^2$, $\left\|\frac{1}{\sqrt{\hat{v}_{2,t}}}\odot m_{2,t}\right\|^2$ and $\left\|\frac{1}{\sqrt[4]{\hat{v}_{2,t}}}\odot m_{2,t}\right\|^2$ are bounded from Lemma 6 and $\omega_{1,t},\omega_{2,t},\omega_{1,*},\omega_{2,*},\hat{v}_{2,t}$ are also bounded from Assumption 2 and Lemma 5, applying Lemma 6 and Cauchy-Schwarz inequality yields

$$
\begin{aligned}
&2\eta^2\,\mathbb{E}\left[\left\langle\left(\sqrt{\hat{v}_{2,t}}\odot\omega_{2,t}\right)^T\omega_{1,t}-\left(\sqrt{\hat{v}_{2,t}}\odot\omega_{2,*}\right)^T\omega_{1,*},\left(\frac{\eta}{\sqrt{\hat{v}_{2,t}}}\odot m_{2,t}\right)^T\right.\right.\\
&\left.\left.\left(\frac{\eta}{\sqrt{\hat{v}_{1,t}}}\odot m_{1,t}\right)\right\rangle\mid\mathcal{F}^t\right]+\eta^2\,\mathbb{E}\left[\left\|\left(\sqrt[4]{\hat{v}_{2,t}}\odot\omega_{2,t}\right)^T\left(\frac{\eta}{\sqrt{\hat{v}_{1,t}}}\odot m_{1,t}\right)\right.\right.\\
&+\left(\sqrt[4]{\hat{v}_{2,t}}\odot\left(\frac{\eta}{\sqrt{\hat{v}_{2,t}}}\odot m_{2,t}\right)\right)^T\omega_{1,t}+\left(\sqrt[4]{\hat{v}_{2,t}}\odot\left(\frac{\eta}{\sqrt{\hat{v}_{2,t}}}\odot m_{2,t}\right)\right)^T\\
&\left.\left.\left(\frac{\eta}{\sqrt{\hat{v}_{1,t}}}\odot m_{1,t}\right)\right\|^2\mid\mathcal{F}^t\right]\\
=&2\eta^2\,\mathbb{E}\left[\left\langle\left(\sqrt{\hat{v}_{2,t}}\odot\omega_{2,t}\right)^T\omega_{1,t}-\left(\sqrt{\hat{v}_{2,t}}\odot\omega_{2,*}\right)^T\omega_{1,*},\left(\frac{\eta}{\sqrt{\hat{v}_{2,t}}}\odot m_{2,t}\right)^T\right.\right.\\
&\left.\left.\left(\frac{\eta}{\sqrt{\hat{v}_{1,t}}}\odot m_{1,t}\right)\right\rangle\mid\mathcal{F}^t\right]+\eta^2\,\mathbb{E}\left[\left\|\left(\sqrt[4]{\hat{v}_{2,t}}\odot\omega_{2,t}\right)^T\left(\frac{\eta}{\sqrt{\hat{v}_{1,t}}}\odot m_{1,t}\right)\right.\right.\\
&\left.\left.+\left(\frac{\eta}{\sqrt[4]{\hat{v}_{2,t}}}\odot m_{2,t}\right)^T\omega_{1,t}+\left(\frac{\eta}{\sqrt[4]{\hat{v}_{2,t}}}\odot m_{2,t}\right)^T\left(\frac{\eta}{\sqrt{\hat{v}_{1,t}}}\odot m_{1,t}\right)\right\|^2\mid\mathcal{F}^t\right]\\
\le&\eta^2\,\mathbb{E}\left[\left\|\left(\sqrt{\hat{v}_{2,t}}\odot\omega_{2,t}\right)^T\omega_{1,t}-\left(\sqrt{\hat{v}_{2,t}}\odot\omega_{2,*}\right)^T\omega_{1,*},\right\|^2+\left\|\frac{\eta}{\sqrt{\hat{v}_{2,t}}}\odot m_{2,t}\right\|^2\right.\\
&\left.\cdot\left\|\frac{\eta}{\sqrt{\hat{v}_{1,t}}}\odot m_{1,t}\right\|^2\mid\mathcal{F}^t\right]+2\eta^2\,\mathbb{E}\left[\left\|\sqrt[4]{\hat{v}_{2,t}}\odot\omega_{2,t}\right\|^2\left\|\frac{\eta}{\sqrt{\hat{v}_{1,t}}}\odot m_{1,t}\right\|^2\right.\\
&\left.+\left\|\frac{\eta}{\sqrt[4]{\hat{v}_{2,t}}}\odot m_{2,t}\right\|^2\|\omega_{1,t}\|^2+\left\|\frac{\eta}{\sqrt[4]{\hat{v}_{2,t}}}\odot m_{2,t}\right\|^2\left\|\frac{\eta}{\sqrt{\hat{v}_{1,t}}}\odot m_{1,t}\right\|^2\mid\mathcal{F}^t\right]\\
\le&\,\eta^2\cdot M_1,
\end{aligned}
\tag{40}
$$

where M_1 is a fixed constant. Now, let us proceed to show an upper bound for the term in (36). Applying Lemma 5 to (36) yields

$$
\begin{aligned}
&\mathbb{E}\left[\left\langle\left(\sqrt[2]{\hat{v}_{2,t}}\odot\omega_{2,t}\right)^T\omega_{1,t}-\left(\sqrt[2]{\hat{v}_{2,t}}\odot\omega_{2,*}\right)^T\omega_{1,*},\omega_{2,t}^T\frac{\eta}{\sqrt{\hat{v}_{1,t}}}\odot m_{1,t}\right\rangle\mid\mathcal{F}^t\right]\\
=&\,\mathbb{E}\left[\left\langle\left(\sqrt{\tilde{v}_{2,t}}\right)^T\left(\omega_{2,t}^T\omega_{1,t}-\omega_{2,*}^T\omega_{1,*}\right),\omega_{2,t}^T\frac{\eta}{\sqrt{\hat{v}_{1,t}}}\odot m_{1,t}\right\rangle\mid\mathcal{F}^t\right]\\
=&\,\mathbb{E}\left[\left\langle\left(\sqrt{\tilde{v}_{2,t}}\right)^T\left(\omega_{2,t}^T\omega_{1,t}-\omega_{2,*}^T\omega_{1,*}\right),\omega_{2,t}^T\frac{\eta}{\sqrt{\hat{v}_{1,t}}}\odot(\beta_{11t}m_{1,t-1}+(1-\beta_{11t})g_{1,t})\right\rangle\mid\mathcal{F}^t\right]\\
=&\,\mathbb{E}\left[\left\langle\left(\sqrt{\tilde{v}_{2,t}}\right)^T\left(\omega_{2,t}^T\omega_{1,t}-\omega_{2,*}^T\omega_{1,*}\right),\omega_{2,t}^T\frac{\eta}{\sqrt{\hat{v}_{1,t}}}\odot\beta_{11t}m_{1,t-1}\right\rangle\mid\mathcal{F}^t\right] \qquad (41)\\
&+\mathbb{E}\left[\left\langle\left(\sqrt{\tilde{v}_{2,t}}\right)^T\left(\omega_{2,t}^T\omega_{1,t}-\omega_{2,*}^T\omega_{1,*}\right),\omega_{2,t}^T\frac{\eta}{\sqrt{\hat{v}_{1,t}}}\odot(1-\beta_{11t})g_{1,t}\right\rangle\mid\mathcal{F}^t\right]. \qquad (42)
\end{aligned}
$$

Since $\omega_{2,t}, \omega_{1,t}, \omega_{2,*}, \omega_{1,*}, \frac{m_{1,t-1}}{\sqrt{\hat{v}_{1,t}}}, \tilde{v}_{2,t}$ and $m_{1,t-1}$ are all bounded, for the term in (41), there exists a constant M_2 such that

$$\begin{aligned}
&\mathbb{E}\left[\left\langle \left(\sqrt{\tilde{v}_{2,t}}\right)^T \left(\omega_{2,t}^T\omega_{1,t} - \omega_{2,*}^T\omega_{1,*}\right), \omega_{2,t}^T \frac{\eta}{\sqrt{\hat{v}_{1,t}}} \odot \beta_{11t} m_{1,t-1} \right\rangle \mid \mathcal{F}^t\right] \\
=&\eta\beta_{11t}\,\mathbb{E}\left[\left(\omega_{1,t}^T\omega_{2,t} - \omega_{1,*}^T\omega_{2,*}\right)\sqrt{\tilde{v}_{2,t}}\omega_{2,t}^T \frac{1}{\sqrt{\hat{v}_{1,t}}} \odot m_{1,t-1} \mid \mathcal{F}^t\right] \\
\leq&\frac{\eta\beta_{11t}}{2}\,\mathbb{E}\left[\left\|\left(\omega_{1,t}^T\omega_{2,t} - \omega_{1,*}^T\omega_{2,*}\right)\sqrt{\tilde{v}_{2,t}}\omega_{2,t}^T\right\|^2 + \left\|\frac{1}{\sqrt{\hat{v}_{1,t}}} m_{1,t-1}\right\|^2 \mid \mathcal{F}^t\right] \leq \eta\beta_{11t}M_2. \qquad (43)
\end{aligned}$$

Next, let us bound the term in (42). Based on Lemma 6, we have

$$\begin{aligned}
&\left|\mathbb{E}\left[\left\langle \left(\sqrt{\tilde{v}_{2,t}}\right)^T \left(\omega_{2,t}^T\omega_{1,t} - \omega_{2,*}^T\omega_{1,*}\right), \omega_{2,t}^T \frac{\eta}{\sqrt{\hat{v}_{1,t}}} \odot (1-\beta_{11t}) g_{1,t} \right\rangle \mid \mathcal{F}^t\right]\right| \\
=&\eta(1-\beta_{11t})\left|\mathbb{E}\left[\left(\omega_{1,t}^T\omega_{2,t} - \omega_{1,*}^T\omega_{2,*}\right)\sqrt{\tilde{v}_{2,t}}\omega_{2,t}^T \frac{1}{\sqrt{\hat{v}_{1,t}}} \odot g_{1,t} \mid \mathcal{F}^t\right]\right| \\
\leq&\eta(1-\beta_{11t})\left\|\omega_{1,t}^T\omega_{2,t} - \omega_{1,*}^T\omega_{2,*}\right\| \mathbb{E}\left[\left\|\sqrt{\tilde{v}_{2,t}}\omega_{2,t}^T \frac{1}{\sqrt{\hat{v}_{1,t}}} \odot g_{1,t}\right\| \mid \mathcal{F}^t\right].
\end{aligned}$$

Now, let us focus on the product in the expectation. Since $\sqrt{\tilde{v}_{2,t}} \in \mathbb{R}^{d\times d}$ is a diagonal matrix, let us denote the i_{th} element on diagonal as $[\tilde{v}_{2,t}]_i$. Then,

$$\sqrt{\tilde{v}_{2,t}}\omega_{2,t}^T \frac{1}{\sqrt{\hat{v}_{1,t}}} \odot g_{1,t} = \left(\mathcal{V}_t \odot \omega_{2,t}\right)^T g_{1,t},$$

where $\mathcal{V}_{12} \in \mathbb{R}^{n\times d}$ such that $[\mathcal{V}_{12}]_{ij} = \sqrt{\frac{[\tilde{v}_{2,t}]_j}{[\hat{v}_{1,t}]_i}}$. Then, we obtain

$$\begin{aligned}
&\left|\mathbb{E}\left[\left\langle \left(\sqrt{\tilde{v}_{2,t}}\right)^T \left(\omega_{2,t}^T\omega_{1,t} - \omega_{2,*}^T\omega_{1,*}\right), \omega_{2,t}^T \frac{\eta}{\sqrt{\hat{v}_{1,t}}} \odot (1-\beta_{11t}) g_{1,t} \right\rangle \mid \mathcal{F}^t\right]\right| \\
\leq&\eta(1-\beta_{11t})\left\|\omega_{1,t}^T\omega_{2,t} - \omega_{1,*}^T\omega_{2,*}\right\| \mathbb{E}\left[\left\|\left(\mathcal{V}_t \odot \omega_{2,t}\right)\right\| \left\|g_{1,t}\right\| \mid \mathcal{F}^t\right].
\end{aligned}$$

Based on (26) and condition 5 from Assumption 2, we discover

$$[\mathcal{V}_{12}]_{ij} = \sqrt{\frac{[\tilde{v}_{2,t}]_j}{[\hat{v}_{1,t}]_i}} \leq \frac{G_{2,\infty}}{\mu}, \qquad (44)$$

which in turn yields

$$\begin{aligned}
&\left|\mathbb{E}\left[\left\langle \left(\sqrt{\tilde{v}_{2,t}}\right)^T \left(\omega_{2,t}^T\omega_{1,t} - \omega_{2,*}^T\omega_{1,*}\right), \omega_{2,t}^T \frac{\eta}{\sqrt{\hat{v}_{1,t}}} \odot (1-\beta_{11t}) g_{1,t} \right\rangle \mid \mathcal{F}^t\right]\right| \\
\leq&\eta(1-\beta_{11t})\frac{\alpha G_{2,\infty}}{\mu}\left\|\omega_{1,t}^T\omega_{2,t} - \omega_{1,*}^T\omega_{2,*}\right\| \mathbb{E}\left[\left\|g_{1,t}\right\| \mid \mathcal{F}^t\right]. \qquad (45)
\end{aligned}$$

Note that in (44), we assume that $[\hat{v}_{1,t}]_i$ is nonzero on the i_{th} coordinate. On the other hand, if $[\hat{v}_{1,t}]_i$ is zero on the i_{th} coordinate, then it implies $[g_{1,j}]_i = 0$ for $j = 1, 2, \cdots, t$ on the i_{th} coordinate, which in turn yields $[g_{1,t}]_i = 0$. Thus, (45)

directly follows. Then, based on $g_{1,t}$ in [nn-adam-updates] in Algorithm 4, we obtain

$$
\begin{aligned}
&\left|\mathbb{E}\left[\left\langle\left(\sqrt{\tilde{v}_{2,t}}\right)^T\left(\omega_{2,t}^T\omega_{1,t}-\omega_{2,*}^T\omega_{1,*}\right),\omega_{2,t}^T\frac{\eta}{\sqrt{\hat{v}_{1,t}}}\odot(1-\beta_{11t})g_{1,t}\right\rangle\mid\mathcal{F}^t\right]\right| \\
=&\eta(1-\beta_{11t})\frac{\alpha G_{2,\infty}}{\mu}\left\|\omega_{1,t}^T\omega_{2,t}-\omega_{1,*}^T\omega_{2,*}\right\|\mathbb{E}\left[\left\|\left(\omega_{1,t}^T\sigma_1\left(\omega_{2,t}z^t\right)-y^t\right)\sigma_2\left(\omega_{2,t}z^t\right)\right\|\mid\mathcal{F}^t\right] \\
\leq&\eta(1-\beta_{11t})\frac{\alpha G_{2,\infty}}{\mu}\left\|\omega_{1,t}^T\omega_{2,t}-\omega_{1,*}^T\omega_{2,*}\right\|\mathbb{E}\left[\left|\omega_{1,t}^T\sigma_1\left(\omega_{2,t}z^t\right)-\omega_{1,*}^T\sigma_1\left(\omega_{2,*}z^t\right)\right|\mid\mathcal{F}^t\right] \\
&\qquad\cdot\mathbb{E}\left[\left\|\sigma_2\left(\omega_{2,t}z^t\right)\right\|\mid\mathcal{F}^t\right] \\
=&\eta(1-\beta_{11t})\frac{\alpha G_{2,\infty}}{\mu}\left\|\omega_{1,t}^T\omega_{2,t}-\omega_{1,*}^T\omega_{2,*}\right\|\rho\left|\left(\omega_{1,t}^T\omega_{2,t}-\omega_{1,*}^T\omega_{2,*}\right)z^t\right|\rho\left\|\omega_{2,t}z^t\right\| \\
=&\eta(1-\beta_{11t})\frac{\alpha G_{2,\infty}}{\mu}\left\|\omega_{1,t}^T\omega_{2,t}-\omega_{1,*}^T\omega_{2,*}\right\|\rho^2\left|\left(\omega_{1,t}^T\omega_{2,t}-\omega_{1,*}^T\omega_{2,*}\right)z^t\right|\left\|\omega_{2,t}z^t\right\| \\
\leq&\eta(1-\beta_{11t})\frac{\alpha G_{2,\infty}}{\mu}\rho^2\left\|\omega_{2,t}\right\|\left(\omega_{1,t}^T\omega_{2,t}z^t-\omega_{1,*}^T\omega_{2,*}z^t\right)^2\frac{\left\|\omega_{1,t}^T\omega_{2,t}-\omega_{1,*}^T\omega_{2,*}\right\|\left\|z^t\right\|}{\left|\omega_{1,t}^T\omega_{2,t}z^t-\omega_{1,*}^T\omega_{2,*}z^t\right|} \\
\leq&2\eta\frac{\alpha G_{2,\infty}(1-\beta_{11t})}{\mu\cos\epsilon}\mathbb{E}\left[l_t\mid\mathcal{F}^t\right].
\end{aligned}
\tag{46}
$$

The last inequality follows by applying conditions 3 and 4 in Assumption 2. Next, Let us deal with the term in (37). Based on $m_{2,t}$ in [nn-adam-updates] in Algorithm 4, we observe

$$
\begin{aligned}
&\mathbb{E}\left[\left\langle\left(\sqrt[2]{\hat{v}_{2,t}}\odot\omega_{2,t}\right)^T\omega_{1,t}-\left(\sqrt[2]{\hat{v}_{2,t}}\odot\omega_{2,*}\right)^T\omega_{1,*},\left(\frac{\eta}{\sqrt{\hat{v}_{2,t}}}\odot m_{2,t}\right)^T\omega_{1,t}\right\rangle\mid\mathcal{F}^t\right] \\
=&\eta\,\mathbb{E}\left[\left\langle\omega_{2,t}^T\omega_{1,t}-\omega_{2,*}^T\omega_{1,*},m_{2,t}^T\omega_{1,t}\right\rangle\mid\mathcal{F}^t\right] \\
=&\eta\,\mathbb{E}\left[\left\langle\omega_{2,t}^T\omega_{1,t}-\omega_{2,*}^T\omega_{1,*},\left(\beta_{12t}m_{2,t-1}+(1-\beta_{12t})\,g_{2,t}\right)^T\omega_{1,t}\right\rangle\mid\mathcal{F}^t\right] \\
=&\eta\left[\beta_{12t}\left\langle\omega_{2,t}^T\omega_{1,t}-\omega_{2,*}^T\omega_{1,*},m_{2,t-1}^T\omega_{1t}\right\rangle+(1-\beta_{12t})\left\langle\omega_{2,t}^T\omega_{1,t}-\omega_{2,*}^T\omega_{1,*},\mathbb{E}\left[g_{2,t}^T\mid\mathcal{F}^t\right]\omega_{1,t}\right\rangle\right] \\
=&\eta\beta_{12t}\left\langle\omega_{2,t}^T\omega_{1,t}-\omega_{2,*}^T\omega_{1,*},m_{2,t-1}^T\omega_{1t}\right\rangle+\eta\,(1-\beta_{12t}) \\
&\qquad\cdot\left\langle\omega_{2,t}^T\omega_{1,t}-\omega_{2,*}^T\omega_{1,*},\left(\mathbb{E}\left[\left(\omega_{1,t}^T\sigma_1\left(\omega_{2,t}z^t\right)-y^t\right)\omega_{1,t}\left(\sigma_2\left(z^t\right)\right)^T\mid\mathcal{F}^t\right]\right)^T\omega_{1,t}\right\rangle \\
=&\eta\beta_{12t}\left(\omega_{1,t}^T\omega_{2,t}-\omega_{1,*}^T\omega_{2,*}\right)m_{2,t-1}^T\omega_{1,t} \qquad (47)\\
&\qquad+\eta\rho^2\,(1-\beta_{12t})\left\langle\omega_{2,t}^T\omega_{1,t}-\omega_{2,*}^T\omega_{1,*},\left(\omega_{1,t}^T\omega_{2,t}z^t-\omega_{1,*}^T\omega_{2,*}z^t\right)z^t\omega_{1,t}^T\omega_{1,t}\right\rangle. \qquad (48)
\end{aligned}
$$

The last equality holds true due to (17) in Lemma 4. By using the fact that $\omega_{1,t},\omega_{2,t},\omega_{1,*},\omega_{2,*}$ and $m_{2,t-1}$ are all bounded, for the term in (47), there exists a constant M_3 such that

$$
\left|\eta\beta_{12t}\left(\omega_{1,t}^T\omega_{2,t}-\omega_{1,*}^T\omega_{2,*}\right)m_{2,t-1}^T\omega_{1,t}\right|\leq\eta\beta_{12t}M_3. \tag{49}
$$

At the same time, by inserting (15) from Lemma 3 into (48) we get

$$
\begin{aligned}
&\eta\rho^2\,(1-\beta_{12t})\left\langle\omega_{2,t}^T\omega_{1,t}-\omega_{2,*}^T\omega_{1,*},\left(\omega_{1,t}^T\omega_{2,t}z^t-\omega_{1,*}^T\omega_{2,*}z^t\right)z^t\omega_{1,t}^T\omega_{1,t}\right\rangle \\
=&\eta\rho^2(1-\beta_{12t})\left(\omega_{1,t}^T\omega_{2,t}z^t-\omega_{1,*}^T\omega_{2,*}z^t\right)\left(\omega_{1,t}^T\omega_{2,t}-\omega_{1,*}^T\omega_{2,*}\right)z^t\omega_{1,t}^T\omega_{1,t} \\
=&2(1-\beta_{12t})\eta\left\|\omega_{1,t}\right\|^2\mathbb{E}\left[l_t\mid\mathcal{F}^t\right] \\
\geq&2(1-\beta_{121})\eta\left\|\omega_{1,t}\right\|^2\mathbb{E}\left[l_t\mid\mathcal{F}^t\right].
\end{aligned}
\tag{50}
$$

By inserting (40),(43),(46), (49), and (50), into (35) we obtain

$$
\begin{aligned}
&2\,\mathbb{E}\left[l_t\mid\mathcal{F}^t\right]\left((1-\beta_{121})\left\|\omega_{1,t}\right\|^2-\frac{\alpha G_{2,\infty}(1-\beta_{11t})}{\mu\cos\epsilon}\right) \\
\leq&\frac{1}{\eta}\left\{\mathbb{E}\left[\left\|\left(\sqrt[4]{\hat{v}_{2,t}}\odot\omega_{2,t+1}\right)^T\omega_{1,t+1}-\left(\sqrt[4]{\hat{v}_{2,t}}\odot\omega_{2,*}\right)^T\omega_{1,*}\right\|^2\mid\mathcal{F}^t\right]\right. \\
&\qquad\left.-\mathbb{E}\left[\left\|\left(\sqrt[4]{\hat{v}_{2,t}}\odot\omega_{2,t}\right)^T\omega_{1,t}-\left(\sqrt[4]{\hat{v}_{2,t}}\odot\omega_{2,*}\right)^T\omega_{1,*}\right\|^2\mid\mathcal{F}^t\right]\right\} \\
&\qquad+2\left(\beta_{11t}M_2+\beta_{12t}M_3\right)+\eta M_1.
\end{aligned}
$$

Since $\|\omega_{1,t}\| = \sqrt{\left[\frac{1}{2}+\xi_2\right]/(1-\beta_{121})} = \sqrt{\left[\frac{1}{2}+\frac{\alpha G_{2,\infty}}{\mu\cos(\epsilon)}\right]/(1-\beta_{121})}$, which in turn yields

$$2\left((1-\beta_{121})\|\omega_{1,t}\|^2 - \frac{\alpha G_{2,\infty}(1-\beta_{11t})}{\mu\cos\epsilon}\right) \geq 1.$$

Therefore, by recalling the law of iterated expectations and summing up all loss functions for $t = p, p+1, \cdots, p+T$, we get

$$\begin{aligned}\sum_{t=p}^{T+p}\mathbb{E}\left[l_t\right] \leq \frac{1}{\eta}\sum_{t=p}^{p+T}\Bigg\{&\mathbb{E}\left[\left\|\left(\sqrt[4]{\hat{v}_{2,t}}\odot\omega_{2,t+1}\right)^T\omega_{1,t+1} - \left(\sqrt[4]{\hat{v}_{2,t}}\odot\omega_{2,*}\right)^T\omega_{1,*}\right\|^2\right]\\ &-\mathbb{E}\left[\left\|\left(\sqrt[4]{\hat{v}_{2,t}}\odot\omega_{2,t}\right)^T\omega_{1,t} - \left(\sqrt[4]{\hat{v}_{2,t}}\odot\omega_{2,*}\right)^T\omega_{1,*}\right\|^2\right]\Bigg\}\\ &+2\sum_{t=p}^{p+T}\left(\beta_{11t}M_2+\beta_{12t}M_3\right)+T\eta M_1. \end{aligned} \tag{51}$$

Applying the definition of β_{11t} and β_{12t} implies

$$\begin{aligned}&\sum_{t=p}^{p+T}\left(\beta_{11t}M_2+\beta_{12t}M_3\right) = \sum_{t=p}^{p+T}\left(\beta_{111}\gamma_1^t M_2+\beta_{121}\gamma_2^t M_3\right)\\ =&\beta_{111}M_2\sum_{t=p}^{p+T}\gamma_1^t+\beta_{121}M_3\sum_{t=p}^{p+T}\gamma_2^t \leq \frac{\beta_{111}M_2}{1-\gamma_1}+\frac{\beta_{121}M_3}{1-\gamma_2}.\end{aligned} \tag{52}$$

Since $z\in\mathbb{R}^d$, we notice that $\tilde{v}_{2,t}\in\mathbb{R}^{d\times d}$. Applying Lemma 5 yields

$$\begin{aligned}&\sum_{t=p}^{p+T}\Bigg\{\mathbb{E}\left[\left\|\left(\sqrt[4]{\hat{v}_{2,t}}\odot\omega_{2,t+1}\right)^T\omega_{1,t+1} - \left(\sqrt[4]{\hat{v}_{2,t}}\odot\omega_{2,*}\right)^T\omega_{1,*}\right\|^2\right]\\ &\quad-\mathbb{E}\left[\left\|\left(\sqrt[4]{\hat{v}_{2,t}}\odot\omega_{2,t}\right)^T\omega_{1,t} - \left(\sqrt[4]{\hat{v}_{2,t}}\odot\omega_{2,*}\right)^T\omega_{1,*}\right\|^2\right]\Bigg\}\\ =&\sum_{t=p}^{p+T}\Bigg\{\mathbb{E}\left[\left\|\left(\sqrt[4]{\tilde{v}_{2,t}}\right)^T\left(\omega_{2,t+1}^T\omega_{1,t+1}-\omega_{2,*}^T\omega_{1,*}\right)\right\|^2\right]\\ &\quad-\mathbb{E}\left[\left\|\left(\sqrt[4]{\tilde{v}_{2,t}}\right)^T\left(\omega_{2,t}^T\omega_{1,t}-\omega_{2,*}^T\omega_{1,*}\right)\right\|^2\right]\Bigg\}\\ =&\sum_{t=p}^{p+T}\Bigg\{\mathbb{E}\left[\sum_{i=1}^d\left[\sqrt{\tilde{v}_{2,t}}\right]_i\left[\omega_{2,t+1}^T\omega_{1,t+1}-\omega_{2,*}^T\omega_{1,*}\right]_i^2\right]\\ &\quad-\mathbb{E}\left[\sum_{i=1}^d\left[\sqrt{\tilde{v}_{2,t}}\right]_i\left[\omega_{2,t}^T\omega_{1,t}-\omega_{2,*}^T\omega_{1,*}\right]_i^2\right]\Bigg\}\\ =&\mathbb{E}\left[\sum_{i=1}^d\left[\sqrt{\tilde{v}_{2,p}}\right]_i\left[\omega_{2,p}^T\omega_{1,p}-\omega_{2,*}^T\omega_{1,*}\right]_i^2\right]+\sum_{t=p+1}^{T+p}\Bigg\{\mathbb{E}\left[\sum_{i=1}^d\left[\sqrt{\tilde{v}_{2,t}}\right]_i\left[\omega_{2,t}^T\omega_{1,t}-\omega_{2,*}^T\omega_{1,*}\right]_i^2\right]\\ &\quad-\mathbb{E}\left[\sum_{i=1}^d\left[\sqrt{\tilde{v}_{2,t-1}}\right]_i\left[\omega_{2,t}^T\omega_{1,t}-\omega_{2,*}^T\omega_{1,*}\right]_i^2\right]\Bigg\}\\ =&\mathbb{E}\left[\sum_{i=1}^d\left[\sqrt{\tilde{v}_{2,p}}\right]_i\left[\omega_{2,p}^T\omega_{1,p}-\omega_{2,*}^T\omega_{1,*}\right]_i^2\right]\\ &\quad+\sum_{t=p+1}^{T+p}\sum_{i=1}^d\left\{\mathbb{E}\left[\left[\sqrt{\tilde{v}_{2,t}}\right]_i\left[\omega_{2,t}^T\omega_{1,t}-\omega_{2,*}^T\omega_{1,*}\right]_i^2-\left[\sqrt{\tilde{v}_{2,t-1}}\right]_i\left[\omega_{2,t}^T\omega_{1,t}-\omega_{2,*}^T\omega_{1,*}\right]_i^2\right]\right\}\\ =&\mathbb{E}\left[\sum_{i=1}^d\left[\sqrt{\tilde{v}_{2,p}}\right]_i\left[\omega_{2,p}^T\omega_{1,p}-\omega_{2,*}^T\omega_{1,*}\right]_i^2\right]\\ &\quad+\sum_{t=p+1}^{T+p}\sum_{i=1}^d\mathbb{E}\left[\left[\left(\sqrt{\tilde{v}_{2,t}}-\sqrt{\tilde{v}_{2,t-1}}\right)\right]_i\left[\omega_{2,t}^T\omega_{1,t}-\omega_{2,*}^T\omega_{1,*}\right]_i^2\right],\end{aligned} \tag{53}$$

where $\left[\sqrt{\tilde{v}_{2,t}}\right]_i$ represents the i_{th} element on diagonal in matrix $\tilde{v}_{2,t}$ and $\left[\omega_{2,t}^T\omega_{1,t}-\omega_{2,*}^T\omega_{1,*}\right]_i$ represents the i_{th} coordinate in vector $\omega_{2,t}^T\omega_{1,t}-\omega_{2,*}^T\omega_{1,*}$. Since $\omega_{1,t},\omega_{2,t},\omega_{1,*}$ and $\omega_{2,*}$ are all bounded for any t, e.g. $\left|\left[\omega_{2,t}^T\omega_{1,t}-\omega_{2,*}^T\omega_{1,*}\right]_i^2\right|\leq W_\infty$ and $\tilde{v}_{2,t}\geq\tilde{v}_{2,t-1}$ due to the fact that $\hat{v}_{2,t}\geq\hat{v}_{2,t-1}$, (53) can be further simplified as

$$\begin{aligned}
&\sum_{t=p}^{p+T}\left\{\mathbb{E}\left[\left\|\left(\sqrt[4]{\hat{v}_{2,t}}\odot\omega_{2,t+1}\right)^T\omega_{1,t+1}-\left(\sqrt[4]{\hat{v}_{2,t}}\odot\omega_{2,*}\right)^T\omega_{1,*}\right\|^2\right]\right.\\
&\qquad\left.-\mathbb{E}\left[\left\|\left(\sqrt[4]{\hat{v}_{2,t}}\odot\omega_{2,t}\right)^T\omega_{1,t}-\left(\sqrt[4]{\hat{v}_{2,t}}\odot\omega_{2,*}\right)^T\omega_{1,*}\right\|^2\right]\right\}\\
\leq&W_\infty\sum_{i=1}^{d}\mathbb{E}\left[\left[\sqrt{\tilde{v}_{2,p}}\right]_i\right]+W_\infty\sum_{t=p+1}^{T+p}\sum_{i=1}^{d}\mathbb{E}\left[\left[\left(\sqrt{\tilde{v}_{2,t}}-\sqrt{\tilde{v}_{2,t-1}}\right)\right]_i\right]\\
=&W_\infty\sum_{i=1}^{d}\mathbb{E}\left[\left[\sqrt{\tilde{v}_{2,p+T}}\right]_i\right].
\end{aligned}\tag{54}$$

Substituting (52) and (54) in (51) gives

$$\sum_{t=p}^{T+p}\mathbb{E}\left[l_t\right]\leq\frac{1}{\eta}W_\infty\sum_{i=1}^{d}\mathbb{E}\left[\left[\sqrt{\tilde{v}_{2,p+T}}\right]_i\right]+2\left(\frac{\beta_{111}M_2}{1-\gamma_1}+\frac{\beta_{121}M_3}{1-\gamma_2}\right)+T\eta M_1=\mathcal{O}(\sqrt{T}).\tag{55}$$

The last equality uses the definition of $\eta=\frac{\eta_1}{\sqrt{T}}$. The desired result in Theorem 3 follows directly from (55) since it holds for any p. □