Appendix

A. Additional Plots

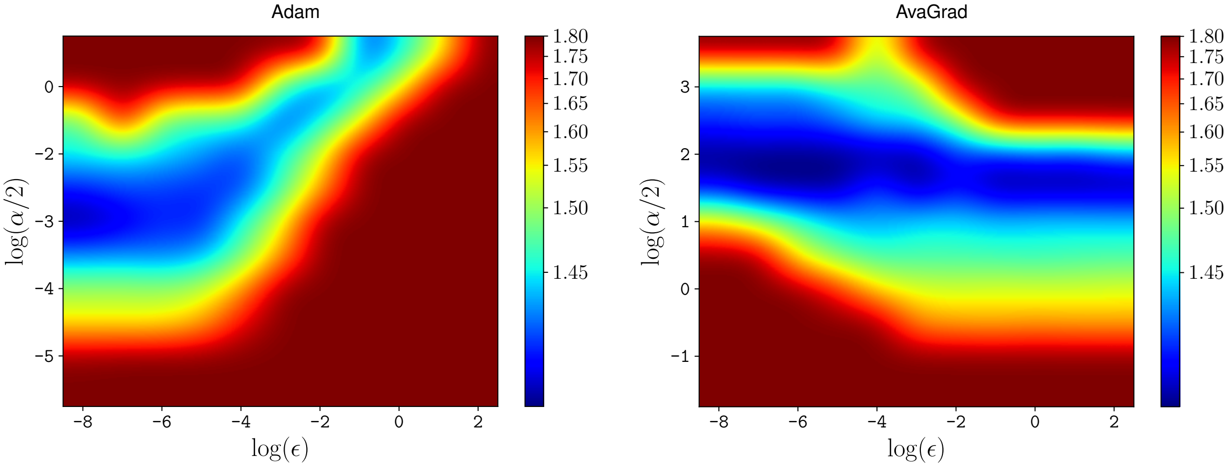

Figure 4: Performance of Adam and AvaGrad with different learning rate α and adaptability parameter ϵ, measured in terms of validation BPC (*lower is better*) on PTB of a 3-layer LSTM. Best performance is achieved with high adaptability/small ϵ.

B. Full Statement and Proof of Theorem 1

Theorem 4. *For any $\epsilon \geq 0$ and constant $\beta_{2,t} = \beta_2 \in [0,1)$, there is a stochastic optimization problem for which Adam does not converge to a stationary point.*

Proof. Consider the following stochastic optimization problem:

$$\min_{w\in[0,1]} f(w) := \mathbb{E}_{s\sim\mathcal{D}}\left[f_s(w)\right] \qquad f_s(w) = \begin{cases} C\frac{w^2}{2}, & \text{with probability} \quad p := \frac{1+\delta}{C+1} \\ -w, & \text{otherwise} \end{cases}, \tag{12}$$

where δ is a positive constant to be specified later, and $C > \frac{1-p}{p} > 1 + \frac{\epsilon}{w_1\sqrt{1-\beta_2}}$ is another constant that can depend on δ, β_2 and ϵ, and will also be determined later. Note that $\nabla f(w) = pCw - (1-p)$, and f is minimized at $w^\star = \frac{1-p}{Cp} = \frac{C-\delta}{C(1+\delta)}$.

The proof follows closely from [34]. We assume w.l.o.g. that $\beta_1 = 0$. We first consider the difference between two consecutive iterates computed by Adam with a constant learning rate α:

$$\Delta_t = w_{t+1} - w_t = -\alpha\frac{g_t}{\sqrt{v_t}+\epsilon} = -\alpha\frac{g_t}{\sqrt{\beta_2 v_{t-1} + (1-\beta_2)g_t^2}+\epsilon}, \tag{13}$$

and then we proceed to analyze the expected change in iterates divided by the learning rate. First, note that with probability p we have $g_t = \nabla(C\frac{w_t^2}{2}) = Cw_t$, and while $g_t = \nabla(-w) = -1$ with probability $1-p$. Therefore, we have

$$\begin{aligned} \frac{\mathbb{E}\left[\Delta_t\right]}{\alpha} &= \frac{\mathbb{E}\left[w_{t+1} - w_t\right]}{\alpha} = -\mathbb{E}\left[\frac{g_t}{\sqrt{\beta_2 v_{t-1} + (1-\beta_2)g_t^2}+\epsilon}\right] \\ &= p\mathbb{E}\left[\underbrace{\frac{-Cw_t}{\sqrt{\beta_2 v_{t-1} + (1-\beta_2)C^2 w_t^2}+\epsilon}}_{T_1}\right] + (1-p)\mathbb{E}\left[\underbrace{\frac{1}{\sqrt{\beta_2 v_{t-1} + (1-\beta_2)}+\epsilon}}_{T_2}\right], \end{aligned} \tag{14}$$

where the expectation is over all the randomness in the algorithm up to time t, as all expectations to follow in the proof. We will proceed by computing lower bounds for the terms T_1 and T_2 above. Note that $T_1 = 0$ for $w_t = 0$, while for $w_t > 0$ we can bound T_1 by

$$T_1 = \frac{-Cw_t}{\sqrt{\beta_2 v_{t-1} + (1-\beta_2)C^2 w_t^2} + \epsilon} \geq \frac{-Cw_t}{\sqrt{(1-\beta_2)C^2 w_t^2}} = \frac{-1}{\sqrt{1-\beta_2}} \,. \tag{15}$$

Combining the cases $w_t = 0$ and $w_t > 0$ (note that the feasible region is $w \in [0,1]$), we have that, generally, $T_1 \geq \min(0, \frac{-1}{\sqrt{1-\beta_2}}) = \frac{-1}{\sqrt{1-\beta_2}}$.

Next, we bound the expected value of T_2 using Jensen's inequality coupled with the convexity of $x^{-1/2}$ as

$$\mathbb{E}\left[T_2\right] = \mathbb{E}\left[\frac{1}{\sqrt{\beta_2 v_{t-1} + 1 - \beta_2} + \epsilon}\right] \geq \frac{1}{\sqrt{\beta_2 \mathbb{E}\left[v_{t-1}\right] + 1 - \beta_2} + \epsilon} \,. \tag{16}$$

Let us consider $\mathbb{E}\left[v_{t-1}\right]$ now. Note that

$$\begin{aligned}
v_{t-1} &= \beta_2 v_{t-2} + (1-\beta_2) g_{t-1}^2 \\
&= \beta_2 \left(\beta_2 v_{t-3} + (1-\beta_2) g_{t-2}^2\right) + (1-\beta_2) g_{t-1}^2 \\
&= \beta_2^2 v_{t-3} + \beta_2(1-\beta_2) g_{t-2}^2 + (1-\beta_2) g_{t-1}^2 \\
&= \beta_2^2 \left(\beta_2 v_{t-4} + (1-\beta_2) g_{t-3}^2\right) + \beta_2(1-\beta_2) g_{t-2}^2 + (1-\beta_2) g_{t-1}^2 \\
&= \beta_2^3 v_{t-4} + \beta_2^2(1-\beta_2) g_{t-3}^2 + \beta_2(1-\beta_2) g_{t-2}^2 + (1-\beta_2) g_{t-1}^2 \\
&\;\;\vdots \\
&= \beta_2^{t-1} v_0 + \beta_2^{t-2}(1-\beta_2) g_1^2 + \beta_2^{t-3}(1-\beta_2) g_2^2 + \cdots + (1-\beta_2) g_{t-1}^2 \\
&= (1-\beta_2) \sum_{i=1}^{t-1} \beta_2^{t-i-1} g_i^2 \,,
\end{aligned} \tag{17}$$

where we used the fact that $v_0 = 0$ (i.e. the second-moment estimate is initialized as zero).

Taking the expectation of the above expression for v_{t-1}, we get

$$\begin{aligned}
\mathbb{E}\left[v_{t-1}\right] &= (1-\beta_2) \sum_{i=1}^{t-1} \beta_2^{t-i-1} \mathbb{E}\left[g_i^2\right] \\
&= (1-\beta_2) \sum_{i=1}^{t-1} \beta_2^{t-i-1} \left(1 - p + pC^2 \mathbb{E}\left[w_t^2\right]\right) \,,
\end{aligned} \tag{18}$$

where we can use the fact that $w_t \in [0,1]$, so $w_t^2 \leq 1$ to get

$$\begin{aligned}
\mathbb{E}\left[v_{t-1}\right] &\leq (1-\beta_2) \sum_{i=1}^{t-1} \beta_2^{t-i-1} \left(1 - p + pC^2\right) \\
&= (1-\beta_2) \left(1 - p + pC^2\right) \sum_{i=1}^{t-1} \beta_2^{t-i-1} \\
&= (1-\beta_2) \left(1 - p + pC^2\right) \sum_{i=0}^{t-2} \beta_2^{i} \\
&= \left(1 - p + pC^2\right) \sum_{i=0}^{t-2} \left(\beta_2^i - \beta_2^{i+1}\right) \\
&= \left(1 - p + pC^2\right) \left(1 - \beta_2^{t-1}\right) \\
&\leq (1+\delta) C^2 \,,
\end{aligned} \tag{19}$$

where $\sum_{i=0}^{t-2} \left(\beta_2^i - \beta_2^{i+1}\right) = 1 - \beta_2^{t-1}$ follows from the fact that the sum telescopes.

Plugging the above bound in (16) yields

$$\mathbb{E}\left[T_2\right] \geq \frac{1}{\sqrt{\beta_2(1+\delta)C + 1 - \beta_2} + \epsilon} \tag{20}$$

Combining the bounds for T_1 and T_2 in (14) gets us that

$$\frac{\mathbb{E}\left[\Delta_t\right]}{\alpha} \geq \frac{1+\delta}{C+1}\frac{-1}{\sqrt{1-\beta_2}} + \left(1 - \frac{1+\delta}{C+1}\right)\frac{1}{\sqrt{\beta_2(1+\delta)C + 1 - \beta_2} + \epsilon} \tag{21}$$

Now, recall that $w^\star = \frac{C-\delta}{C(1+\delta)}$, so for C sufficiently large in comparison to δ we get $w^\star \approx \frac{1}{1+\delta}$. On the other hand, the above quantity can be made non-negative for large enough C, so $\mathbb{E}\left[w_t\right] \geq \mathbb{E}\left[w_{t-1}\right] \geq \dots \geq w_1$. In other words, Adam will, in expectation, update the iterates towards $w = 1$ even though the stationary point is $w^* \approx \frac{1}{1+\delta}$ and we have $\left\|\nabla f(1)\right\|^2 = \delta$ at $w = 1$. Setting $\delta = 1$, for example, implies that $\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[\left\|\nabla f(w_t)\right\|^2\right] = 1$, and hence Adam presents nonconvergence in terms of stationarity. To see that $w = 1$ is not a stationary point due to the feasibility constraints, check that $\nabla f(1) = 1 > 0$: that is, the negative gradient points *towards* the feasible region. □

C. Technical Lemmas

This section presents intermediate results that are used in the proofs given in the next sections.

For simplicity we adopt the following notation for all following results:

$$H_t = \max_i \eta_{t,i} \qquad L_t = \min_i \eta_{t,i}\,, \tag{22}$$

where $\eta_t \in \mathbb{R}^d$ denotes the parameter-wise learning rates computed at iteration t (the method being considered and consequently the exact expression for η_t will be specified in each result).

For the following Lemmas we rely extensively on the assumption that $\left\|\nabla f_s(w)\right\|_\infty \leq G_\infty$ for some constant G_∞, and also that this assumption implies that there exists G_2 such that $\left\|\nabla f_s(w)\right\| \leq G_2$ for all $s \in \mathcal{S}$ and $w \in \mathbb{R}^d$, which can be seen by noting that

$$\left\|\nabla f_s(w)\right\| = \left(\sum_{i=1}^d (\nabla f_s(w))_i^2\right)^{\frac{1}{2}} \leq \left(d\left\|\nabla f_s(w)\right\|_\infty^2\right)^{\frac{1}{2}} = \sqrt{d}\cdot\left\|\nabla f_s(w)\right\|_\infty \leq \sqrt{d}\cdot G_\infty\,, \tag{23}$$

hence such constant G_2 must exist as any $G_2 \geq \sqrt{d}\cdot G_\infty$ satisfies $\left\|\nabla f_s(w)\right\| \leq G_2$.

Lemma 1. *Assume that there exists a constant G_∞ such that $\left\|\nabla f_s(w)\right\|_\infty \leq G_\infty$ for all $s \in \mathcal{S}$ and $w \in \mathbb{R}^d$, and let G_2 be a constant such that $\left\|\nabla f_s(w)\right\| \leq G_2$ for all $s \in \mathcal{S}$ and $w \in \mathbb{R}^d$. Moreover, assume that $\beta_{1,t} \in [0,1)$ for all $t \in \mathbb{N}$.*

Let $m_t \in \mathbb{R}^d$ be given by

$$m_t = \beta_{1,t}m_{t-1} + (1-\beta_{1,t})g_t \quad \textit{and} \quad m_0 = 0\,,$$

where $\beta_{1,t} \in [0,1)$ for all $t \in \mathbb{N}$.

Then, we have

$$\left\|m_t\right\|_\infty \leq G_\infty \qquad \textit{and} \qquad \left\|m_t\right\| \leq G_2$$

for all $t \in \mathbb{N}$ and all possible sample sequences $(s_1, \dots, s_t) \in \mathcal{S}^t$.

Proof. Assume for the sake of contradiction that $\left\|m_t\right\|_\infty > G_\infty$ for some $t \in \mathbb{N}$ and some sequence of samples $(s_1, \dots, s_t)$. Moreover, assume w.l.o.g. that $\left\|m_{t'}\right\|_\infty \leq G_\infty$ for all $t' \in \{1, \dots, t-1\}$ and note that there is no loss of generality since $(m_{t'})_{t'=0}^t$ must indeed have a first element that satisfies $\left\|m_{t'}\right\|_\infty > G_\infty$, which cannot be m_0 since we have $m_0 = 0$ by definition.

Then, we have that $m_{t,i} > G_\infty$ for some $i \in [d]$, but

$$\begin{aligned} m_{t,i} &= \beta_{1,t}m_{t-1,i} + (1-\beta_{1,t})g_{t,i} \\ &\leq \beta_{1,t}\left\|m_{t-1}\right\|_\infty + (1-\beta_{1,t})\left\|g_t\right\|_\infty \\ &\leq \beta_{1,t}G_\infty + (1-\beta_{1,t})G_\infty \\ &= G_\infty\,, \end{aligned} \tag{24}$$

where we used $\beta_{1,t} \in [0,1)$ and the assumptions $\|m_{t-1}\|_\infty \leq G_\infty$ and $\|g_t\|_\infty \leq G_\infty$.

To show that $\|m_t\| \leq G_2$, note that if we assume $\|m_t\| > G_2$ and $\|m_{t'}\| \leq G_2$ for all $t' \in \{1, \dots, t-1\}$, we again get a contradiction since, by the triangle inequality,

$$\begin{aligned}
\|m_t\| &= \|\beta_{1,t} m_{t-1} + (1-\beta_{1,t}) g_t\| \\
&\leq \beta_{1,t} \|m_{t-1}\| + (1-\beta_{1,t}) \|g_t\| \\
&\leq \beta_{1,t} G_2 + (1-\beta_{1,t}) G_2 \\
&= G_2 \, ,
\end{aligned} \tag{25}$$

therefore it must indeed follow that $\|m_t\| \leq G_2$.

□

Lemma 2. *Assume that there exists a constant G_∞ such that $\|\nabla f_s(w)\|_\infty \leq G_\infty$ for all $s \in \mathcal{S}$ and $w \in \mathbb{R}^d$, and let G_2 be a constant such that $\|\nabla f_s(w)\| \leq G_2$ for all $s \in \mathcal{S}$ and $w \in \mathbb{R}^d$. Moreover, assume that $\beta_{2,t} \in [0,1)$ for all $t \in \mathbb{N}$.*

Let $v_t \in \mathbb{R}^d$ be given by

$$v_t = \beta_{2,t} v_{t-1} + (1-\beta_{2,t}) g_t^2 \quad \textit{and} \quad v_0 = 0 \, ,$$

where $\beta_{2,t} \in [0,1)$ for all $t \in \mathbb{N}$.

Then, we have

$$\|v_t\|_\infty \leq G_\infty^2 \qquad \textit{and} \qquad \|v_t\| \leq G_2^2$$

for all $t \in \mathbb{N}$ and all possible sample sequences $(s_1, \dots, s_t) \in \mathcal{S}^t$.

Proof. The proof follows closely from the one of Lemma 1. Assume for the sake of contradiction that there exists $t \in \mathbb{N}$ and some sequence of samples $(s_1, \dots, s_t)$ such that $\|v_t\|_\infty > G_\infty^2$ and $\|v_{t'}\|_\infty \leq G_\infty^2$ for all $t' \in \{1, \dots, t-1\}$.

Then $v_{t,i} > G_\infty^2$ for some $i \in [d]$ but

$$\begin{aligned}
v_{t,i} &= \beta_{2,t} v_{t-1,i} + (1-\beta_{2,t}) g_{t,i}^2 \\
&\leq \beta_{2,t} \|v_{t-1}\|_\infty + (1-\beta_{2,t}) \|g_t\|_\infty^2 \\
&\leq \beta_{2,t} G_\infty^2 + (1-\beta_{2,t}) G_\infty^2 \\
&= G_\infty^2 \, ,
\end{aligned} \tag{26}$$

where we used $\beta_{2,t} \in [0,1)$ and the assumptions $\|v_{t-1}\|_\infty \leq G_\infty^2$ and $\|g_t\|_\infty \leq G_\infty$, which raises a contradiction and shows that indeed $\|v_t\|_\infty \leq G_\infty^2$.

For the ℓ_2 case, assume that $\|v_t\| > G_2^2$ and $\|v_{t'}\| \leq G_2^2$ for all $t' \in \{1, \dots, t-1\}$, which yields

$$\begin{aligned}
\|v_t\| &= \left\|\beta_{2,t} v_{t-1} + (1-\beta_{2,t}) g_t^2\right\| \\
&\leq \beta_{2,t} \|v_{t-1}\| + (1-\beta_{2,t}) \left\|g_t^2\right\| \\
&\leq \beta_{2,t} G_2^2 + (1-\beta_{2,t}) G_2^2 \\
&= G_2^2 \, ,
\end{aligned} \tag{27}$$

where we used the assumption $\|g_t\| \leq G_2$ which also implies that

$$\begin{aligned}
\|g_t^2\| &= \left[\sum_{i=1}^{d} g_{t,i}^4\right]^{\frac{1}{2}} \\
&\leq \left[\sum_{i=1}^{d} g_{t,i}^4 + \sum_{i=1}^{d}\sum_{j=1}^{d} g_{t,i}^2 g_{t,j}^2\right]^{\frac{1}{2}} \\
&= \left[\left(\sum_{i=1}^{d} g_{t,i}^2\right)^2\right]^{\frac{1}{2}} \\
&= \left[\left(\sum_{i=1}^{d} g_{t,i}^2\right)^{\frac{1}{2}}\right]^2 \\
&\leq G_2^2 \, .
\end{aligned} \tag{28}$$

Checking that (27) yields a contradiction completes the argument.

□

Lemma 3. *Under the same assumptions of Lemma 1, we have*

$$\|m_{t'} \odot \eta_t\| \leq \min\left(G_\infty \|\eta_t\|, G_2 H_t\right) \, , \tag{29}$$

for all $t, t' \in \mathbb{N}$ and all possible sample sequences $(s_1, \ldots, s_{\max(t,t')})$.

Proof. By definition,

$$\begin{aligned}
\|m_{t'} \odot \eta_t\|^2 &= \sum_{i=1}^{d} m_{t,i}^2 \cdot \eta_{t,i}^2 \\
&\leq \sum_{i=1}^{d} (\max_{j \in [d]} m_{t',j}^2) \cdot \eta_{t,i}^2 \\
&\leq \|m_{t'}\|_\infty^2 \sum_{i=1}^{d} \eta_{t,i}^2 \\
&= \|m_{t'}\|_\infty^2 \cdot \|\eta_t\|^2 \, ,
\end{aligned} \tag{30}$$

hence invoking Lemma 1 and taking the square root yields $\|m_{t'} \odot \eta_t\| \leq G_\infty \|\eta_t\|$.

Additionally, we have

$$\begin{aligned}
\|m_{t'} \odot \eta_t\|^2 &\leq \sum_{i=1}^{d} m_{t',i}^2 (\max_{j \in [d]} \eta_{t,j}^2) \\
&\leq \|\eta_t\|_\infty^2 \sum_{i=1}^{d} m_{t',i}^2 \\
&= \|\eta_t\|_\infty^2 \cdot \|m_{t'}\|^2 \, ,
\end{aligned} \tag{31}$$

hence recalling that $\|\eta_t\|_\infty = H_t$ and by Lemma 1 we get $\|m_{t'} \odot \eta_t\| \leq G_2 H_t$.

Combining the two bounds completes the proof. □

Lemma 4. *Under the same assumptions of Lemma 1, we have*

$$\langle \nabla f(w_t), m_t \odot \eta_t \rangle \geq (1 - \beta_{1,t}) \langle \nabla f(w_t), g_t \odot \eta_t \rangle - \beta_{1,t} G_2 \|m_{t-1} \odot \eta_t\| \, , \tag{32}$$

for all $t \in \mathbb{N}$ and all possible sample sequences $(s_1, \ldots, s_t) \in \mathcal{S}^t$.

Proof. Using the definition of m_t, we have

$$\begin{aligned} \langle \nabla f(w_t), m_t \odot \eta_t \rangle &= \Big\langle \nabla f(w_t), \big(\beta_{1,t} m_{t-1} + (1-\beta_{1,t}) g_t\big) \odot \eta_t \Big\rangle \\ &= (1-\beta_{1,t}) \langle \nabla f(w_t), g_t \odot \eta_t \rangle + \beta_{1,t} \langle \nabla f(w_t), m_{t-1} \odot \eta_t \rangle \\ &\geq (1-\beta_{1,t}) \langle \nabla f(w_t), g_t \odot \eta_t \rangle - \beta_{1,t} \|\nabla f(w_t)\| \cdot \|m_{t-1} \odot \eta_t\| \ , \end{aligned} \tag{33}$$

where we used Cauchy-Schwarz in the last step.

Next, by Jensen's inequality and the fact that $\|\cdot\|$ is convex we have, for all $w \in \mathbb{R}^d$,

$$\|\nabla f(w)\| = \|\mathbb{E}_s [\nabla f_s(w)]\| \leq \mathbb{E}_s [\|\nabla f_s(w)\|] \leq \mathbb{E}_s [G_2] = G_2 \,. \tag{34}$$

Applying this bound in (33) yields the desired inequality.

□

Lemma 5. *Assume that there exists a constant G_∞ such that $\|\nabla f_s(w)\|_\infty \leq G_\infty$ for all $s \in \mathcal{S}$ and $w \in \mathbb{R}^d$, and let G_2 be a constant such that $\|\nabla f_s(w)\| \leq G_2$ for all $s \in \mathcal{S}$ and $w \in \mathbb{R}^d$. Moreover, assume that $\beta_{1,t} \in [0,1)$ and $\beta_{1,t} \leq \beta_{1,t-1}$ for all $t \in \mathbb{N}$.*

If η_t is independent of s_t for all $t \in \mathbb{N}$, i.e. $P(\eta_t = \eta, s_t = s) = P(\eta_t = \eta) P(s_t = s)$ for all $\eta \in \mathbb{R}^d, s \in \mathcal{S}$, then

$$\mathbb{E}_{s_t} [\langle \nabla f(w_t), m_t \odot \eta_t \rangle] \geq (1-\beta_1) L_t \|\nabla f(w_t)\|^2 - \beta_{1,t} G_2 \|m_{t-1} \odot \eta_t\| \ , \tag{35}$$

for all $t \in \mathbb{N}$ and all possible sample sequences $(s_1, \ldots, s_t) \in \mathcal{S}^t$.

Proof. From Lemma 4 we have that

$$\langle \nabla f(w_t), m_t \odot \eta_t \rangle \geq (1-\beta_{1,t}) \langle \nabla f(w_t), g_t \odot \eta_t \rangle - \beta_{1,t} G_2 \|m_{t-1} \odot \eta_t\| \ . \tag{36}$$

Then, taking the expectation over the draw of $s_t \in \mathcal{S}$ and recalling that w_t, and hence also $\nabla f(w_t)$, is computed from $(s_1, \ldots, s_{t-1})$,

$$\mathbb{E}_{s_t} [\langle \nabla f(w_t), m_t \odot \eta_t \rangle] \geq (1-\beta_{1,t}) \langle \nabla f(w_t), \mathbb{E}_{s_t} [g_t \odot \eta_t] \rangle - \beta_{1,t} G_2 \mathbb{E}_{s_t} [\|m_{t-1} \odot \eta_t\|] \ . \tag{37}$$

Now, note that since we assume that η_t is independent of s_t, we get

$$\mathbb{E}_{s_t} [g_t \odot \eta_t] = \eta_t \odot \mathbb{E}_{s_t} [g_t] = \eta_t \odot \mathbb{E}_{s_t} [\nabla f_{s_t}(w_t)] = \eta_t \odot \nabla f(w_t) \,, \tag{38}$$

and also

$$\mathbb{E}_{s_t} [\|m_{t-1} \odot \eta_t\|] = \|m_{t-1} \odot \eta_t\| \ . \tag{39}$$

Combining (39) and (38) into (37) yields

$$\mathbb{E}_{s_t} [\langle \nabla f(w_t), m_t \odot \eta_t \rangle] \geq (1-\beta_{1,t}) \langle \nabla f(w_t), \nabla f(w_t) \odot \eta_t \rangle - \beta_{1,t} G_2 \|m_{t-1} \odot \eta_t\| \ . \tag{40}$$

Moreover, we have

$$\begin{aligned} \langle \nabla f(w_t), \nabla f(w_t) \odot \eta_t \rangle &= \sum_{i=1}^d (\nabla f(w_t))_i (\nabla f(w_t))_i \eta_{t,i} \\ &= \sum_{i=1}^d (\nabla f(w_t))_i^2 \eta_{t,i} \\ &\geq \sum_{i=1}^d (\nabla f(w_t))_i^2 (\min_j \eta_{t,j}) \\ &= L_t \sum_{i=1}^d (\nabla f(w_t))_i^2 \\ &= L_t \|\nabla f(w_t)\|^2 \ , \end{aligned} \tag{41}$$

which, when applied to (40) yields

$$\mathbb{E}_{s_t}\left[\langle \nabla f(w_t), m_t \odot \eta_t \rangle\right] \geq (1-\beta_{1,t}) L_t \left\|\nabla f(w_t)\right\|^2 - \beta_{1,t} G_2 \left\|m_{t-1} \odot \eta_t\right\| \,, \tag{42}$$

where we also used that $\beta_{1,t} \in [0,1)$ and $L_t \geq 0$. Using the fact that $\beta_{1,t} \leq \beta_{1,t-1} \leq \beta_1$ for all $t \in \mathbb{N}$ and hence $1-\beta_{1,t} \geq 1-\beta_1$ yields the desired inequality.

□

D. Proof of Theorem 3

We organize the proof as follows: we first prove an intermediate result (Lemma 6) and split the proof of the bounds in (8) and (9) in two, where the latter can be seen as a refinement of (8) given the additional assumption that $Z := \sum_{t=1}^{T} \alpha_t \min_i \eta_{t,i}$ is independent of each s_t.

Throughout the proof we use the following notation for clarity:

$$H_t = \max_i \eta_{t,i} \qquad L_t = \min_i \eta_{t,i} \,. \tag{43}$$

Lemma 6. *Assume that f is M-smooth, lower-bounded by some f^* (i.e. $f^* \leq f(w)$ for all $w \in \mathbb{R}^d$), and that there exists a constant G_∞ such that $\left\|\nabla f_s(w)\right\|_\infty \leq G_\infty$ for all $s \in \mathcal{S}$ and $w \in \mathbb{R}^d$, and let G_2 be a constant such that $\left\|\nabla f_s(w)\right\| \leq G_2$ for all $s \in \mathcal{S}$ and $w \in \mathbb{R}^d$.*

Consider any optimization method that performs updates following

$$w_{t+1} = w_t - \alpha_t \cdot \eta_t \odot m_t \,, \tag{44}$$

where we further assume assume that for all $t \in \mathbb{N}$ we have $\alpha_t \geq 0$, $\beta_{1,t} = \frac{\beta_1}{\sqrt{t}}$ for some $\beta_1 \in [0,1)$, and $\eta_{t,i} \geq 0$ for all $i \in [d]$.

If η_t is independent of s_t for all $t \in \mathbb{N}$, i.e. $P(\eta_t = \eta, s_t = s) = P(\eta_t = \eta) P(s_t = s)$ for all $\eta \in \mathbb{R}^d, s \in \mathcal{S}$, then

$$\begin{aligned}\sum_{t=1}^{T} \alpha_t L_t \left\|\nabla f(w_t)\right\|^2 \leq \frac{1}{1-\beta_1} \Bigg(\sum_{t=1}^{T} \left(f(w_t) - \mathbb{E}_{s_t}\left[f(w_{t+1})\right]\right) + \sum_{t=1}^{T} \alpha_t \beta_{1,t} G_2 \left\|m_{t-1} \odot \eta_t\right\| \\ + \frac{M}{2} \sum_{t=1}^{T} \alpha_t^2 \mathbb{E}_{s_t}\left[\left\|m_t \odot \eta_t\right\|^2\right] \Bigg) \,,\end{aligned} \tag{45}$$

for all $T \in \mathbb{N}$ and all possible sample sequences $(s_1, \dots, s_T) \in \mathcal{S}^T$.

Proof. We start from the assumption that f is M-smooth, which yields

$$f(w_{t+1}) \leq f(w_t) + \langle \nabla f(w_t), w_{t+1} - w_t \rangle + \frac{M}{2} \left\|w_{t+1} - w_t\right\|^2 \,. \tag{46}$$

Plugging the update expression $w_{t+1} = w_t - \alpha_t \cdot \eta_t \odot m_t$,

$$f(w_{t+1}) \leq f(w_t) - \alpha_t \langle \nabla f(w_t), m_t \odot \eta_t \rangle + \frac{\alpha_t^2 M}{2} \left\|m_t \odot \eta_t\right\|^2 \,. \tag{47}$$

Now, taking the expectation over the random sample $s_t \in \mathcal{S}$, we get

$$\mathbb{E}_{s_t}\left[f(w_{t+1})\right] \leq f(w_t) - \alpha_t \mathbb{E}_{s_t}\left[\langle \nabla f(w_t), m_t \odot \eta_t \rangle\right] + \frac{\alpha_t^2 M}{2} \mathbb{E}_{s_t}\left[\left\|m_t \odot \eta_t\right\|^2\right] \,, \tag{48}$$

where we used the fact that w_t and α_t are not functions of of s_t – in particular, recall that w_t is deterministically computed from $(s_1, \dots, s_{t-1})$.

Using the assumption that η_t is independent of s_t and applying Lemma 5, we get

$$\begin{aligned}\mathbb{E}_{s_t}\left[f(w_{t+1})\right] \leq f(w_t) - \alpha_t (1-\beta_1) L_t \left\|\nabla f(w_t)\right\|^2 + \alpha_t \beta_{1,t} G_2 \left\|m_{t-1} \odot \eta_t\right\| \\ + \frac{\alpha_t^2 M}{2} \mathbb{E}_{s_t}\left[\left\|m_t \odot \eta_t\right\|^2\right] \,,\end{aligned} \tag{49}$$

which can be re-arranged into

$$\alpha_t L_t \|\nabla f(w_t)\|^2 \leq \frac{1}{1-\beta_1}\Bigg(f(w_t) - \mathbb{E}_{s_t}\left[f(w_{t+1})\right] + \alpha_t \beta_{1,t} G_2 \|m_{t-1} \odot \eta_t\| + \frac{\alpha_t^2 M}{2} \mathbb{E}_{s_t}\left[\|m_t \odot \eta_t\|^2\right]\Bigg), \tag{50}$$

where we used the assumption that $\beta_1 \in [0,1)$, hence $1-\beta_1 > 0$ which was used to divide both sides of the inequality.

Now, summing over $t = 1$ to T,

$$\sum_{t=1}^{T} \alpha_t L_t \|\nabla f(w_t)\|^2 \leq \frac{1}{1-\beta_1}\Bigg(\sum_{t=1}^{T} \left(f(w_t) - \mathbb{E}_{s_t}\left[f(w_{t+1})\right]\right) + \sum_{t=1}^{T} \alpha_t \beta_{1,t} G_2 \|m_{t-1} \odot \eta_t\| + \sum_{t=1}^{T} \frac{\alpha_t^2 M}{2} \mathbb{E}_{s_t}\left[\|m_t \odot \eta_t\|^2\right]\Bigg), \tag{51}$$

which yields the desired result.

□

D.1. Proof of the first guarantee (8)

Proof. We start from the bound given in Lemma 6:

$$\sum_{t=1}^{T} \alpha_t L_t \|\nabla f(w_t)\|^2 \leq \frac{1}{1-\beta_1}\Bigg(\sum_{t=1}^{T} \left(f(w_t) - \mathbb{E}_{s_t}\left[f(w_{t+1})\right]\right) + \sum_{t=1}^{T} \alpha_t \beta_{1,t} G_2 \|m_{t-1} \odot \eta_t\| + \frac{M}{2} \sum_{t=1}^{T} \alpha_t^2 \mathbb{E}_{s_t}\left[\|m_t \odot \eta_t\|^2\right]\Bigg). \tag{52}$$

Now, using Lemma 3 to upper bound both $\|m_{t-1} \odot \eta_t\|$ and $\|m_t \odot \eta_t\|$ by $G_2 H_t$,

$$\sum_{t=1}^{T} \alpha_t L_t \|\nabla f(w_t)\|^2 \leq \frac{1}{1-\beta_1}\Bigg(\sum_{t=1}^{T} \left(f(w_t) - \mathbb{E}_{s_t}\left[f(w_{t+1})\right]\right) + \sum_{t=1}^{T} \alpha_t \beta_{1,t} G_2^2 H_t + \frac{M}{2} \sum_{t=1}^{T} \alpha_t^2 G_2^2 H_t^2\Bigg), \tag{53}$$

where we used that $\mathbb{E}_{s_t}\left[H_t^2\right] = H_t^2$ since H_t is deterministically computed from η_t, which in turn is independent of s_t.

Next, from the assumption in Theorem 3 that there are positive constants L and H such that $L \leq \eta_{t,i} \leq H$ for all $t \in \mathbb{N}, i \in [d]$ and sample sequences $(s_1, \dots, s_t)$, it follows that

$$L \leq L_t = \min_{i \in [d]} \eta_{t,i} \quad \text{and} \quad H \geq H_t = \max_{i \in [d]} \eta_{t,i}$$

for all $t \in \mathbb{N}$, therefore

$$L \sum_{t=1}^{T} \alpha_t \|\nabla f(w_t)\|^2 \leq \frac{1}{1-\beta_1}\Bigg(\sum_{t=1}^{T} \left(f(w_t) - \mathbb{E}_{s_t}\left[f(w_{t+1})\right]\right) + G_2^2 H \sum_{t=1}^{T} \alpha_t \beta_{1,t} + \frac{M G_2^2 H^2}{2} \sum_{t=1}^{T} \alpha_t^2\Bigg). \tag{54}$$

Dividing both sides by $L \geq 0$ and letting $\alpha_t = \alpha'/\sqrt{T}$ yields

$$\begin{aligned}\sum_{t=1}^{T} \frac{\alpha'}{\sqrt{T}} \left\|\nabla f(w_t)\right\|^2 \leq \frac{1}{L(1-\beta_1)} \Bigg(\sum_{t=1}^{T} \left(f(w_t) - \mathbb{E}_{s_t}\left[f(w_{t+1})\right]\right) + G_2^2 H \sum_{t=1}^{T} \frac{\alpha'}{\sqrt{T}} \beta_{1,t} \\ + \frac{M G_2^2 H^2}{2} \sum_{t=1}^{T} \frac{\alpha'^2}{T} \Bigg) ,\end{aligned} \tag{55}$$

and, rearranging and using the fact that

$$\sum_{t=1}^{T} \beta_{1,t} = \beta_1 \sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq \beta_1 \int_0^T \frac{1}{\sqrt{t}} dt \leq 2\beta_1 \sqrt{T} ,$$

which implies that $\sum_{t=1}^{T} \frac{\alpha'}{\sqrt{T}} \beta_{1,t} \leq 2\alpha' \beta_1$, we get

$$\begin{aligned}\frac{\alpha'}{\sqrt{T}} \sum_{t=1}^{T} \left\|\nabla f(w_t)\right\|^2 \leq \frac{1}{L(1-\beta_1)} \Bigg(\sum_{t=1}^{T} \left(f(w_t) - \mathbb{E}_{s_t}\left[f(w_{t+1})\right]\right) + 2\alpha' \beta_1 G_2^2 H \\ + \frac{\alpha'^2 M G_2^2 H^2}{2} \Bigg) .\end{aligned} \tag{56}$$

Now, taking the expectation over the full sample sequence $(s_1, \ldots, s_T)$ yields

$$\begin{aligned}\frac{\alpha'}{\sqrt{T}} \sum_{t=1}^{T} \mathbb{E}\left[\left\|\nabla f(w_t)\right\|^2\right] \leq \frac{1}{L(1-\beta_1)} \Bigg(\sum_{t=1}^{T} \left(\mathbb{E}\left[f(w_t)\right] - \mathbb{E}\left[f(w_{t+1})\right]\right) + 2\alpha' \beta_1 G_2^2 H \\ + \frac{\alpha'^2 M G_2^2 H^2}{2} \Bigg) .\end{aligned} \tag{57}$$

Note that, by telescoping sum,

$$\sum_{t=1}^{T} \mathbb{E}\left[f(w_t)\right] - \mathbb{E}\left[f(w_{t+1})\right] = \mathbb{E}\left[f(w_1)\right] - \mathbb{E}\left[f(w_{T+1})\right] \leq f(w_1) - f^* , \tag{58}$$

where the last step follows since w_1 (the parameters at initialization) is independent of the drawn samples, and also from the assumption that f^* lower bounds f.

Combining the above with (57) and dividing both sides by $\alpha' \cdot \sqrt{T}$,

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[\left\|\nabla f(w_t)\right\|^2\right] \leq \frac{1}{L\sqrt{T}(1-\beta_1)} \left(\frac{f(w_1) - f^*}{\alpha'} + 2\beta_1 G_2^2 H + \frac{\alpha' M G_2^2 H^2}{2} \right) ,. \tag{59}$$

Finally, we will use Young's inequality with $p = 2$ and the conjugate exponent $q = 2$, which states that $xy \leq \frac{x^2}{2} + \frac{y^2}{2}$ for any nonnegative numbers x, y.

In that context, let

$$x = \frac{1}{\sqrt{\alpha'}} \quad \text{and} \quad y = \sqrt{\alpha'} H , \tag{60}$$

which yields

$$H = xy \leq \frac{x^2}{2} + \frac{y^2}{2} = \frac{1}{\alpha'} + \alpha' H^2 , \tag{61}$$

and hence

$$2\beta_1 G_2^2 \cdot H \leq \frac{2\beta_1 G_2^2}{\alpha'} + 2\beta_1 G_2^2 \cdot \alpha' H^2 . \tag{62}$$

Plugging the above in (59) yields

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla f(w_t)\|^2\right] \leq \frac{1}{L\sqrt{T}(1-\beta_1)}\left(\frac{2\beta_1 G_2^2 + f(w_1) - f^*}{\alpha'} + \alpha' H^2 \frac{G_2^2(M+2\beta_1)}{2}\right),. \tag{63}$$

Verifying that the above is $\mathcal{O}\left(\frac{1}{L\sqrt{T}}\left(\frac{1}{\alpha'} + \alpha' H^2\right)\right)$ in terms of T, α', L and H finishes the proof.

□

D.2. Proof of the second guarantee (9)

Proof. As before, we start from Lemma 6, which states that

$$\begin{aligned}\sum_{t=1}^{T}\alpha_t L_t \|\nabla f(w_t)\|^2 \leq \frac{1}{1-\beta_1}\Bigg(&\sum_{t=1}^{T}\left(f(w_t) - \mathbb{E}_{s_t}\left[f(w_{t+1})\right]\right) + \sum_{t=1}^{T}\alpha_t\beta_{1,t}G_2\|m_{t-1}\odot\eta_t\| \\ &+ \frac{M}{2}\sum_{t=1}^{T}\alpha_t^2\mathbb{E}_{s_t}\left[\|m_t\odot\eta_t\|^2\right]\Bigg).\end{aligned} \tag{64}$$

We then invoke Lemma 3 to upper bound $\|m_{t-1}\odot\eta_t\|$ by $G_2 H_t$ and $\|m_t\odot\eta_t\|$ by $G_\infty\|\eta_t\|^2$:

$$\begin{aligned}\sum_{t=1}^{T}\alpha_t L_t \|\nabla f(w_t)\|^2 \leq \frac{1}{1-\beta_1}\Bigg(&\sum_{t=1}^{T}\left(f(w_t) - \mathbb{E}_{s_t}\left[f(w_{t+1})\right]\right) + \sum_{t=1}^{T}\alpha_t\beta_{1,t}G_2^2 H_t \\ &+ \frac{M}{2}\sum_{t=1}^{T}\alpha_t^2 G_\infty^2\mathbb{E}_{s_t}\left[\|\eta_t\|^2\right]\Bigg).\end{aligned} \tag{65}$$

Next, define the unormalized probability distribution $\tilde{p}(t) = \alpha_t L_t$, so that $p(t) = \tilde{p}(t)/Z$ with $Z = \sum_{t=1}^{T}\tilde{p}(t) = \sum_{t=1}^{T}\alpha_t L_t$ is a valid distribution over $t \in \{1, \dots T\}$. Dividing both sides by Z yields

$$\sum_{t=1}^{T}p(t)\|\nabla f(w_t)\|^2 \leq \frac{1}{Z(1-\beta_1)}\sum_{t=1}^{T}\left(f(w_t) - \mathbb{E}_{s_t}\left[f(w_{t+1})\right] + \alpha_t\beta_{1,t}G_2^2 H_t + \frac{\alpha_t^2 M G_\infty^2\mathbb{E}_{s_t}\left[\|\eta_t\|^2\right]}{2}\right) \tag{66}$$

Now, taking the conditional expectation over all samples S given Z:

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}p(t)\|\nabla f(w_t)\|^2 \Big| Z\right] &\leq \frac{1}{Z(1-\beta_1)}\Big(\sum_{t=1}^{T}\left(\mathbb{E}\left[f(w_t)|Z\right] - \mathbb{E}\left[\mathbb{E}_{s_t}\left[f(w_{t+1})\right]|Z\right]\right) \\ &\quad + \sum_{t=1}^{T}\mathbb{E}\left[\alpha_t\beta_{1,t}G_2^2 H_t + \frac{\alpha_t^2 M G_\infty^2\mathbb{E}_{s_t}\left[\|\eta_t\|^2\right]}{2}\Big| Z\right]\Big) \\ &\leq \frac{1}{Z(1-\beta_1)}\Big(\sum_{t=1}^{T}\left(\mathbb{E}\left[f(w_t)|Z\right] - \mathbb{E}\left[f(w_{t+1})|Z\right]\right) \\ &\quad + \sum_{t=1}^{T}\mathbb{E}\left[\alpha_t\beta_{1,t}G_2^2 H_t + \frac{\alpha_t^2 M G_\infty^2\|\eta_t\|^2}{2}\Big| Z\right]\Big) \\ &= \frac{1}{Z(1-\beta_1)}\Big(f(w_1) - f^* \\ &\quad + \sum_{t=1}^{T}\mathbb{E}\left[\alpha_t\beta_{1,t}G_2^2 H_t + \frac{\alpha_t^2 M G_\infty^2\|\eta_t\|^2}{2}\Big| Z\right]\Big).\end{aligned} \tag{67}$$

where in the second step we used $\mathbb{E}\left[\mathbb{E}_{s_t}\left[\cdot\right]|Z\right] = \mathbb{E}\left[\cdot|Z\right]$ which follows from the assumption that $p(Z|s_t) = p(Z)$, and the third step follows from telescoping sum and the assumption that f^* lower bounds f.

Then, taking the expectation over Z and re-arranging:

$$\mathbb{E}\left[\sum_{t=1}^{T} p(t)\left\|\nabla f(w_t)\right\|^2\right] \leq \mathbb{E}\left[\frac{1}{Z(1-\beta_1)}\sum_{t=1}^{T}\left(\frac{f(w_1)-f^*}{T} + \alpha_t\beta_{1,t}G_2^2 H_t + \frac{\alpha_t^2 M G_\infty^2\left\|\eta_t\right\|^2}{2}\right)\right] . \tag{68}$$

Setting $\beta_1 = 0$ for simplicity yields

$$\mathbb{E}\left[\sum_{t=1}^{T} p(t)\left\|\nabla f(w_t)\right\|^2\right] \leq \mathbb{E}\left[\frac{1}{Z}\sum_{t=1}^{T}\left(\frac{f(w_1)-f^*}{T} + \frac{\alpha_t^2 M G_\infty^2\left\|\eta_t\right\|^2}{2}\right)\right] . \tag{69}$$

Now, let $\alpha_t = \alpha'_t/\sqrt{T}$

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T} p(t)\left\|\nabla f(w_t)\right\|^2\right] &\leq \mathbb{E}\left[\frac{1}{Z}\sum_{t=1}^{T}\left(\frac{f(w_1)-f^*}{T} + \frac{\alpha_t'^2 M G_\infty^2\left\|\eta_t\right\|^2}{2T}\right)\right] \\ &= \frac{1}{T}\cdot\mathbb{E}\left[\frac{1}{Z}\sum_{t=1}^{T}\left(f(w_1)-f^* + \frac{1}{2}\alpha_t'^2 M G_\infty^2\left\|\eta_t\right\|^2\right)\right] .\end{aligned} \tag{70}$$

Now, recall that $Z = \sum_{t=1}^{T}\alpha_t L_t = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\alpha'_t L_t$, hence

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T} p(t)\left\|\nabla f(w_t)\right\|^2\right] &\leq \frac{1}{\sqrt{T}}\cdot\mathbb{E}\left[\frac{\sum_{t=1}^{T} f(w_1)-f^* + \frac{1}{2}\alpha_t'^2 M G_\infty^2\left\|\eta_t\right\|^2}{\sum_{t=1}^{T}\alpha'_t L_t}\right] \\ &\leq \mathcal{O}\left(\frac{1}{\sqrt{T}}\cdot\mathbb{E}\left[\frac{\sum_{t=1}^{T} 1 + \alpha_t'^2\left\|\eta_t\right\|^2}{\sum_{t=1}^{T}\alpha'_t L_t}\right]\right) .\end{aligned} \tag{71}$$

Finally, checking that $\sum_{t=1}^{T} p(t)\left\|\nabla f(w_t)\right\|^2 = \mathbb{E}_{t\sim\mathcal{P}(t|S)}\left[\left\|\nabla f(w_t)\right\|^2\right]$:

$$\mathop{\mathbb{E}}_{\substack{S\sim\mathcal{D}^T \\ t\sim\mathcal{P}(t|S)}}\left[\left\|\nabla f(w_t)\right\|^2\right] \leq \mathcal{O}\left(\frac{1}{\sqrt{T}}\cdot\mathbb{E}\left[\frac{\sum_{t=1}^{T} 1 + \alpha_t'^2\left\|\eta_t\right\|^2}{\sum_{t=1}^{T}\alpha'_t L_t}\right]\right) . \tag{72}$$

Recalling that $L_t = \min_i \eta_{t,i}$ completes the proof. □

E. Full Statement and Proof of Theorem 2

We organize the formal statement and proof of Theorem 2 as follows: we first state a general convergence result for Adam which depends on the step-wise adaptivity parameter ϵ_t and the learning rates α_t in Theorem 5, and then present a Corollary that shows how a $\mathcal{O}(1/\sqrt{T})$ rate follows from such result (Corollary 6). This section proceeds the proof of Theorem 3 (Appendix D) as the proof presented here is more easily seen as a small variant (although overall simpler) of the analysis given in the previous section. Steps which also appear in the proof of Theorem 3 are not necessarily described in full detail, hence the following arguments can be better understood with the previous section in context.

Throughout the proof we use the following notation for clarity:

$$H_t = \max_i \eta_{t,i} \qquad L_t = \min_i \eta_{t,i} \, . \tag{73}$$

Theorem 5. *Assume that f is smooth and f_s has bounded gradients. If $\epsilon_t \geq \epsilon_{t-1} > 0$ for all $t \in [T]$, then for the iterates $\{w_1, \dots, w_T\}$ produced by Adam we have*

$$\mathbb{E}\left[\left\|\nabla f(w_t)\right\|^2\right] \leq \mathcal{O}\left(\frac{1 + \sum_{t=1}^{T}\frac{\alpha_t}{\epsilon_{t-1}^2}\left(1 + \alpha_t + \epsilon_t - \epsilon_{t-1}\right)}{\sum_{t=1}^{T}\frac{\alpha_t}{1+\epsilon_{t-1}}}\right) , \tag{74}$$

where w_t is sampled from $p(t) \propto \frac{\alpha_t}{G_\infty + \epsilon_{t-1}}$.

Corollary 6. *Setting $\epsilon_t = \Theta(T^{p_1} t^{p_2})$ for any $p_1, p_2 > 0$ such that $p_1 + p_2 \geq \frac{1}{2}$ (e.g. $\epsilon_t = \Theta(\sqrt{T}), \epsilon_t = \Theta(\sqrt[4]{Tt}), \epsilon_t = \Theta(\sqrt{t})$) and $\alpha_t = \Theta\left(\frac{\epsilon_t}{\sqrt{T}}\right)$ on Theorem 5 yields a bound of $\mathcal{O}(1/\sqrt{T})$ for Adam.*

Proof. Similarly to the proof of Theorem 3, we plug the update rule $w_{t+1} = w_t - \alpha_t \cdot \eta_t \odot m_t$ in

$$f(w_{t+1}) \leq f(w_t) + \langle \nabla f(w_t), w_{t+1} - w_t \rangle + \frac{M}{2} \left\| w_{t+1} - w_t \right\|^2 . \tag{75}$$

yielding

$$f(w_{t+1}) \leq f(w_t) - \alpha_t \langle \nabla f(w_t), m_t \odot \eta_t \rangle + \frac{\alpha_t^2 M}{2} \left\| m_t \odot \eta_t \right\|^2 . \tag{76}$$

By Lemmas 3 and 4, we have

$$f(w_{t+1}) \leq f(w_t) - \alpha_t (1 - \beta_{1,t}) \langle \nabla f(w_t), g_t \odot \eta_t \rangle + \alpha_t \beta_{1,t} G_2^2 H_t + \frac{\alpha_t^2 M G_\infty^2 \left\| \eta_t \right\|^2}{2} . \tag{77}$$

Now, note that we can write

$$\langle \nabla f(w_t), g_t \odot \eta_t \rangle = \langle \nabla f(w_t), g_t \odot \eta_{t-1} \rangle + \langle \nabla f(w_t), g_t \odot (\eta_t - \eta_{t-1}) \rangle ,$$

therefore we have that

$$\begin{aligned} f(w_{t+1}) &\leq f(w_t) - \alpha_t (1 - \beta_{1,t}) \langle \nabla f(w_t), g_t \odot \eta_{t-1} \rangle + \alpha_t \beta_{1,t} G_2^2 H_t + \frac{\alpha_t^2 M G_\infty^2 \left\| \eta_t \right\|^2}{2} \\ &\quad - \alpha_t (1 - \beta_{1,t}) \langle \nabla f(w_t), g_t \odot (\eta_t - \eta_{t-1}) \rangle \\ &\leq f(w_t) - \alpha_t (1 - \beta_{1,t}) \langle \nabla f(w_t), g_t \odot \eta_{t-1} \rangle + \alpha_t \beta_{1,t} G_2^2 H_t + \frac{\alpha_t^2 M G_\infty^2 \left\| \eta_t \right\|^2}{2} \\ &\quad + \alpha_t (1 - \beta_{1,t}) \left| \langle \nabla f(w_t), g_t \odot (\eta_t - \eta_{t-1}) \rangle \right| \end{aligned} \tag{78}$$

We will proceed to bound $|\langle \nabla f(w_t), g_t \odot (\eta_t - \eta_{t-1}) \rangle|$. By Cauchy-Schwarz we have

$$|\langle \nabla f(w_t), g_t \odot (\eta_t - \eta_{t-1}) \rangle| \leq \| \nabla f(w_t) \| \cdot \| g_t \odot (\eta_t - \eta_{t-1}) \| \leq G_2 \left\| g_t \odot (\eta_t - \eta_{t-1}) \right\| , \tag{79}$$

and moreover

$$\begin{aligned} \| g_t \odot (\eta_t - \eta_{t-1}) \| &= \left(\sum_{i=1}^d g_{t,i}^2 |\eta_{t,i} - \eta_{t-1,i}|^2 \right)^{1/2} \\ &\leq \left(\sum_{i=1}^d G_\infty^2 |\eta_{t,i} - \eta_{t-1,i}|^2 \right)^{1/2} \\ &= G_\infty \left\| \eta_t - \eta_{t-1} \right\| , \end{aligned} \tag{80}$$

therefore we get

$$\begin{aligned} f(w_{t+1}) &\leq f(w_t) - \alpha_t (1 - \beta_{1,t}) \langle \nabla f(w_t), g_t \odot \eta_{t-1} \rangle + \alpha_t \beta_{1,t} G_2^2 H_t + \frac{\alpha_t^2 M G_\infty^2 \left\| \eta_t \right\|^2}{2} \\ &\quad + \alpha_t (1 - \beta_{1,t}) G_\infty G_2 \left\| \eta_t - \eta_{t-1} \right\| \end{aligned} \tag{81}$$

Using the fact that η_{t-1} is independent of s_t and that $\mathbb{E}_{s_t}[g_t] = \nabla f(w_t)$, taking expectation over s_t yields

$$\begin{aligned} \mathbb{E}_{s_t}[f(w_{t+1})] &\leq f(w_t) - \alpha_t (1 - \beta_{1,t}) \langle \nabla f(w_t), \nabla f(w_t) \odot \eta_{t-1} \rangle + \alpha_t \beta_{1,t} G_2^2 \mathbb{E}_{s_t}[H_t] + \frac{\alpha_t^2 M G_\infty^2 \mathbb{E}_{s_t}\left[\left\| \eta_t \right\|^2 \right]}{2} \\ &\quad + \alpha_t (1 - \beta_{1,t}) G_\infty G_2 \mathbb{E}_{s_t}\left[\| \eta_t - \eta_{t-1} \| \right] \\ &\leq f(w_t) - \alpha_t (1 - \beta_1) \left\| \nabla f(w) \right\|^2 L_{t-1} + \alpha_t \beta_{1,t} G_2^2 \mathbb{E}_{s_t}[H_t] + \frac{\alpha_t^2 M G_\infty^2 \mathbb{E}_{s_t}\left[\left\| \eta_t \right\|^2 \right]}{2} \\ &\quad + \alpha_t (1 - \beta_1) G_\infty G_2 \mathbb{E}_{s_t}\left[\| \eta_t - \eta_{t-1} \| \right] , \end{aligned} \tag{82}$$

where in the second step we used $\beta_{1,t} \leq \beta_1$ and

$$\langle \nabla f(w_t), \nabla f(w_t) \odot \eta_{t-1} \rangle = \sum_{i=1}^{d} \nabla f(w)_i^2 \eta_{t-1,i} \geq \min_j \eta_{t-1,j} \sum_{i=1}^{d} \nabla f(w)_i^2 = L_{t-1} \left\| \nabla f(w) \right\|^2 .$$

Re-arranging,

$$\begin{aligned} \alpha_t L_{t-1}(1-\beta_1) \left\| \nabla f(w_t) \right\|^2 \leq f(w_t) - \mathbb{E}_{s_t}\left[f(w_{t+1}) \right] + \alpha_t \beta_{1,t} G_2^2 \mathbb{E}_{s_t}\left[H_t \right] + \frac{\alpha_t^2 M G_\infty^2 \mathbb{E}_{s_t}\left[\left\| \eta_t \right\|^2 \right]}{2} \\ + \alpha_t (1-\beta_1) G_\infty G_2 \mathbb{E}_{s_t}\left[\left\| \eta_t - \eta_{t-1} \right\| \right] , \end{aligned} \tag{83}$$

Next, we will bound $L_{t-1}, H_t, \|\eta_t\|$, and $\|\eta_t - \eta_{t-1}\|$. Recall that, for Adam, we have

$$\eta_t = \frac{1}{\sqrt{v_t} + \epsilon_t} ,$$

and since $v_{t,i} \leq G_\infty^2$, we also have that

$$\frac{1}{G_\infty + \epsilon_t} \leq \eta_{t,i} \leq \frac{1}{\epsilon_t} .$$

From the above it follows that

$$\frac{1}{G_\infty + \epsilon_{t-1}} \leq L_{t-1}$$

and

$$H_t \geq \frac{1}{\epsilon_t} ,$$

which also implies that $\|\eta_t\| \leq \frac{\sqrt{d}}{\epsilon_t}$.

As for $\|\eta_t - \eta_{t-1}\|$, check that

$$\left| \eta_{t,i} - \eta_{t-1,i} \right| \leq \frac{1}{\epsilon_{t-1}} - \frac{1}{G_\infty + \epsilon_t} = \frac{G_\infty + \epsilon_t - \epsilon_{t-1}}{G_\infty \epsilon_{t-1} + \epsilon_t \epsilon_{t-1}} \leq \frac{G_\infty + \epsilon_t - \epsilon_{t-1}}{\epsilon_{t-1}^2} , \tag{84}$$

where we used the assumption that $\epsilon_t \geq \epsilon_{t-1}$. The above implies that $\|\eta_t - \eta_{t-1}\| \leq \sqrt{d} \cdot \frac{G_\infty + \epsilon_t - \epsilon_{t-1}}{\epsilon_{t-1}^2}$.

Applying the bounds given above to (83) yields

$$\begin{aligned} \frac{\alpha_t}{G_\infty + \epsilon_{t-1}} (1-\beta_1) \left\| \nabla f(w_t) \right\|^2 \leq f(w_t) - \mathbb{E}_{s_t}\left[f(w_{t+1}) \right] + \beta_{1,t} G_2^2 \frac{\alpha_t}{\epsilon_t} + \frac{\alpha_t^2 M d G_\infty^2}{\epsilon_t^2} \\ + \alpha_t (1-\beta_1) \sqrt{d} G_\infty G_2 \cdot \frac{G_\infty + \epsilon_t - \epsilon_{t-1}}{\epsilon_{t-1}^2} , \end{aligned} \tag{85}$$

Next, define the unormalized probability distribution $\tilde{p}(t) = \frac{\alpha_t}{G_\infty + \epsilon_{t-1}}$, so that $p(t) = \tilde{p}(t)/Z$ with $Z = \sum_{t=1}^{T} \tilde{p}(t) = \sum_{t=1}^{T} \frac{\alpha_t}{G_\infty + \epsilon_{t-1}}$ is a valid distribution over $t \in \{1, \dots T\}$. Adopting this notation and dividing both sides by $Z(1-\beta_1)$:

$$\begin{aligned} p(t) \left\| \nabla f(w_t) \right\|^2 \leq \frac{1}{Z(1-\beta_1)} \Big(f(w_t) - \mathbb{E}_{s_t}\left[f(w_{t+1}) \right] + \beta_{1,t} G_2^2 \frac{\alpha_t}{\epsilon_t} + \frac{\alpha_t^2 M d G_\infty^2}{\epsilon_t^2} \\ + \alpha_t (1-\beta_1) \sqrt{d} G_\infty G_2 \cdot \frac{G_\infty + \epsilon_t - \epsilon_{t-1}}{\epsilon_{t-1}^2} \Big) . \end{aligned} \tag{86}$$

Taking the expectation over all samples and summing over t yields

$$\begin{aligned}\sum_{t=1}^{T} p(t)\mathbb{E}\left[\|\nabla f(w_t)\|^2\right] &\leq \frac{1}{Z(1-\beta_1)}\sum_{t=1}^{T}\Big(\mathbb{E}\left[f(w_t)\right] - \mathbb{E}\left[f(w_{t+1})\right] + \beta_{1,t}G_2^2\frac{\alpha_t}{\epsilon_t} + \frac{\alpha_t^2 MdG_\infty^2}{\epsilon_t^2} \\ &\qquad + \alpha_t(1-\beta_1)\sqrt{d}G_\infty G_2 \cdot \frac{G_\infty + \epsilon_t - \epsilon_{t-1}}{\epsilon_{t-1}^2}\Big) \\ &\leq \frac{1}{Z(1-\beta_1)}\Big[f(w_1) - f^* + \sum_{t=1}^{T}\Big(\beta_{1,t}G_2^2\frac{\alpha_t}{\epsilon_t} + \frac{\alpha_t^2 MdG_\infty^2}{\epsilon_t^2} \\ &\qquad + \alpha_t(1-\beta_1)\sqrt{d}G_\infty G_2 \cdot \frac{G_\infty + \epsilon_t - \epsilon_{t-1}}{\epsilon_{t-1}^2}\Big)\Big] .\end{aligned} \tag{87}$$

where we used the fact that $\sum_{t=1}^{T}\mathbb{E}\left[f(w_t)\right] - \mathbb{E}\left[f(w_{t+1})\right] = f(w_1) - \mathbb{E}\left[f(w_{T+1})\right] \leq f(w_1) - f^*$ by telescoping sum and where f^* lower bounds f.

For simplicity, assume that $\beta_{1,t} = 0$ (or, alternatively, let $\beta_{1,t} = \frac{\beta_1}{\sqrt{T}}$ and apply Young's inequality as in the proof of Theorem 3). In this case, we get

$$\sum_{t=1}^{T} p(t)\mathbb{E}\left[\|\nabla f(w_t)\|^2\right] \leq \frac{1}{Z(1-\beta_1)}\Big[f(w_1) - f^* + \sum_{t=1}^{T}\frac{\alpha_t}{\epsilon_{t-1}^2}\Big(\alpha_t MdG_\infty^2 + (1-\beta_1)\sqrt{d}G_\infty G_2 \cdot (G_\infty + \epsilon_t - \epsilon_{t-1})\Big)\Big] , \tag{88}$$

and recalling that $Z = \sum_{t=1}^{T}\frac{\alpha_t}{G_\infty + \epsilon_{t-1}}$ yields

$$\begin{aligned}\mathbb{E}_{t\sim P(t)}\left[\mathbb{E}\left[\|\nabla f(w_t)\|^2\right]\right] &\leq \frac{f(w_1) - f^* + \sum_{t=1}^{T}\frac{\alpha_t}{\epsilon_{t-1}^2}\Big(\alpha_t MdG_\infty^2 + (1-\beta_1)\sqrt{d}G_\infty G_2 \cdot (G_\infty + \epsilon_t - \epsilon_{t-1})\Big)}{(1-\beta_1)\sum_{t=1}^{T}\frac{\alpha_t}{G_\infty + \epsilon_{t-1}}} \\ &\leq \mathcal{O}\left(\frac{1 + \sum_{t=1}^{T}\frac{\alpha_t}{\epsilon_{t-1}^2}\left(1 + \alpha_t + \epsilon_t - \epsilon_{t-1}\right)}{\sum_{t=1}^{T}\frac{\alpha_t}{1+\epsilon_{t-1}}}\right) ,\end{aligned} \tag{89}$$

which completes the argument. □

F. Details on Hyperparameter Optimization

This section contains additional details on the experiments performed in Section 7. We use Gradientless Descent (GLD [10]), a recently-proposed zeroth-order optimization method to tune both α and ϵ for Adam, AMSGrad and AvaGrad. The search space consists of 21 values for α and 21 values for ϵ, yielding a discrete search space composed of 441 hyperparameter settings. We use a projected isotropic Gaussian constrained to $[0, 21] \times [0, 21]$ for sampling: we first sample (x, y) and then round both x and y to the nearest integer to associate the continuous samples to elements in the discrete 21×21 search space. We use a search radius of 4 and 3 samples per iteration.

For the coordinate-wise hyperparameter optimization, we run GLD separately on α and ϵ, in an alternating fashion. In practice, this amounts to using univariate Gaussians during sampling, where in one iteration the distribution is over α only, and in the other it is over ϵ. We denote GLD when run in this manner by "CGLD", standing for coordinate gradientless descent.