AdaNorm: Adaptive Gradient Norm Correction based Optimizer for CNNs

Shiv Ram Dubey*, Satish Kumar Singh*, Bidyut Baran Chaudhuri†
*Computer Vision and Biometrics Lab, Indian Institute of Information Technology, Allahabad
†Techno India University, Kolkata, India and Indian Statistical Institute, Kolkata, India
srdubey@iiita.ac.in, sk.singh@iiita.ac.in, bidyutbaranchaudhuri@gmail.com

Supplementary

A. Algorithms

This section provides the Algorithms for different optimization techniques, including diffGrad (Algorithm 1), diffGradInject (Algorithm 2), Radam (Algorithm 3), RadamInject (Algorithm 4), AdaBelief (Algorithm 5) and AdaBeliefInject (Algorithm 6).

Algorithm 1: diffGrad Optimizer

Initialize: $\boldsymbol{\theta}_0, \boldsymbol{m}_0 \leftarrow 0, \boldsymbol{v}_0 \leftarrow 0, t \leftarrow 0$
Hyperparameters: α, β_1, β_2
While $\boldsymbol{\theta}_t$ not converged
$t \leftarrow t+1$
$\boldsymbol{g}_t \leftarrow \nabla_\theta f_t(\boldsymbol{\theta}_{t-1})$
$\boldsymbol{\xi}_t \leftarrow 1/(1+e^{-|\boldsymbol{g}_t - \boldsymbol{g}_{t-1}|})$
$\boldsymbol{m}_t \leftarrow \beta_1 \boldsymbol{m}_{t-1} + (1-\beta_1)\boldsymbol{g}_t$
$\boldsymbol{v}_t \leftarrow \beta_2 \boldsymbol{v}_{t-1} + (1-\beta_2)\boldsymbol{g}_t^2$
Bias Correction
$\widehat{\boldsymbol{m}_t} \leftarrow \boldsymbol{m}_t/(1-\beta_1^t)$, $\widehat{\boldsymbol{v}_t} \leftarrow \boldsymbol{v}_t/(1-\beta_2^t)$
Update
$\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \alpha \boldsymbol{\xi}_t \widehat{\boldsymbol{m}_t}/(\sqrt{\widehat{\boldsymbol{v}_t}} + \epsilon)$

B. Convergence Proof

Lemma 1. *Let* $\eta \triangleq \frac{\beta_1^2}{\sqrt{\beta_2}}$. *For* $\beta_1, \beta_2 \in [0,1)$ *that satisfy* $\frac{\beta_1^2}{\sqrt{\beta_2}} < 1$ *and bounded* g_t, $||g_t||_2 \leq G$, $||g_t||_\infty \leq G_\infty$, $e_t \leq G_\infty$, $\frac{e_t}{||g_t||_2} \leq \frac{G_\infty}{G}$, *the following inequality holds,*

$$\sum_{t=1}^{T} \frac{\hat{m}_{t,i}^2}{\sqrt{t\hat{v}_{t,i}}} \leq \frac{2G_\infty^3}{G^2(1-\eta)^2\sqrt{1-\beta_2}} ||g_{1:T,i}||_2$$

Proof. Under the assumption, $\frac{\sqrt{1-\beta_2^t}}{(1-\beta_1^t)^2} \leq \frac{1}{(1-\beta_1)^2}$. We can use the update rules of AdamNorm and expand the last term

Algorithm 2: diffGradNorm (diffGrad + AdaNorm) Optimizer

Initialize: $\boldsymbol{\theta}_0, \boldsymbol{m}_0 \leftarrow 0, \boldsymbol{v}_0 \leftarrow 0, e_0 \leftarrow 0, t \leftarrow 0$
Hyperparameters: $\alpha, \beta_1, \beta_2, \gamma$
While $\boldsymbol{\theta}_t$ not converged
$t \leftarrow t+1$
$\boldsymbol{g}_t \leftarrow \nabla_\theta f_t(\boldsymbol{\theta}_{t-1})$
$\boldsymbol{\xi}_t \leftarrow 1/(1+e^{-|\boldsymbol{g}_t - \boldsymbol{g}_{t-1}|})$
$g_{norm} \leftarrow L_2Norm(\boldsymbol{g}_t)$
$e_t = \gamma e_{t-1} + (1-\gamma) g_{norm}$
$\boldsymbol{s}_t = \boldsymbol{g}_t$
If $e_t > g_{norm}$
$\boldsymbol{s}_t = (e_t/g_{norm})\boldsymbol{g}_t$
$\boldsymbol{m}_t \leftarrow \beta_1 \boldsymbol{m}_{t-1} + (1-\beta_1)\boldsymbol{s}_t$
$\boldsymbol{v}_t \leftarrow \beta_2 \boldsymbol{v}_{t-1} + (1-\beta_2)\boldsymbol{g}_t^2$
Bias Correction
$\widehat{\boldsymbol{m}_t} \leftarrow \boldsymbol{m}_t/(1-\beta_1^t)$, $\widehat{\boldsymbol{v}_t} \leftarrow \boldsymbol{v}_t/(1-\beta_2^t)$
Update
$\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \alpha \boldsymbol{\xi}_t \widehat{\boldsymbol{m}_t}/(\sqrt{\widehat{\boldsymbol{v}_t}} + \epsilon)$

in the summation,

$$\begin{aligned}
&\sum_{t=1}^{T} \frac{\hat{m}_{t,i}^2}{\sqrt{t\hat{v}_{t,i}}} \\
&= \sum_{t=1}^{T-1} \frac{\hat{m}_{t,i}^2}{\sqrt{t\hat{v}_{t,i}}} \\
&\quad + \frac{\sqrt{1-\beta_2^T}}{(1-\beta_1^T)^2} \frac{(\sum_{k=1}^{T}(1-\beta_1)\beta_1^{T-k}s_{k,i})^2}{\sqrt{T\sum_{j=1}^{T}(1-\beta_2)\beta_2^{T-j}g_{j,i}^2}} \\
&\leq \sum_{t=1}^{T-1} \frac{\hat{m}_{t,i}^2}{\sqrt{t\hat{v}_{t,i}}} \\
&\quad + \frac{\sqrt{1-\beta_2^T}}{(1-\beta_1^T)^2} \sum_{k=1}^{T} \frac{T((1-\beta_1)\beta_1^{T-k}s_{k,i})^2}{\sqrt{T\sum_{j=1}^{T}(1-\beta_2)\beta_2^{T-j}g_{j,i}^2}} \\
&\leq \sum_{t=1}^{T-1} \frac{\hat{m}_{t,i}^2}{\sqrt{t\hat{v}_{t,i}}} + \frac{\sqrt{1-\beta_2^T}}{(1-\beta_1^T)^2} \sum_{k=1}^{T} \frac{T((1-\beta_1)\beta_1^{T-k}s_{k,i})^2}{\sqrt{T(1-\beta_2)\beta_2^{T-k}g_{k,i}^2}}
\end{aligned}$$

Algorithm 3: Radam Optimizer

Initialize: $\boldsymbol{\theta}_0, \boldsymbol{m}_0 \leftarrow 0, \boldsymbol{v}_0 \leftarrow 0, t \leftarrow 0$
Hyperparameters: α, β_1, β_2
While $\boldsymbol{\theta}_t$ not converged
$t \leftarrow t+1$
$\boldsymbol{g}_t \leftarrow \nabla_\theta f_t(\boldsymbol{\theta}_{t-1})$
$\boldsymbol{m}_t \leftarrow \beta_1 \boldsymbol{m}_{t-1} + (1-\beta_1)\boldsymbol{g}_t$
$\boldsymbol{v}_t \leftarrow \beta_2 \boldsymbol{v}_{t-1} + (1-\beta_2)\boldsymbol{g}_t^2$
$\rho_\infty \leftarrow 2/(1-\beta_2) - 1$
$\rho_t = \rho_\infty - 2t\beta_2^t/(1-\beta_2^t)$
If $\rho_t \geq 5$
$\rho_u = (\rho_t - 4)(\rho_t - 2)\rho_\infty$
$\rho_d = (\rho_\infty - 4)(\rho_\infty - 2)\rho_t$
$\rho = \sqrt{(1-\beta_2)\rho_u/\rho_d}$
$\alpha_1 = \rho\alpha/(1-\beta_1^t)$
Update
$\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \alpha_1 \boldsymbol{m}_t/(\sqrt{\boldsymbol{v}_t} + \epsilon)$
Else
$\alpha_2 = \alpha/(1-\beta_1^t)$
Update
$\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \alpha_2 \boldsymbol{m}_t$

Algorithm 4: RadamNorm (i.e., Radam + AdaNorm) Optimizer

Initialize: $\boldsymbol{\theta}_0, \boldsymbol{m}_0 \leftarrow 0, \boldsymbol{v}_0 \leftarrow 0, e_0 \leftarrow 0, t \leftarrow 0$
Hyperparameters: $\alpha, \beta_1, \beta_2, \gamma$
While $\boldsymbol{\theta}_t$ not converged
$t \leftarrow t+1$
$\boldsymbol{g}_t \leftarrow \nabla_\theta f_t(\boldsymbol{\theta}_{t-1})$
$g_{norm} \leftarrow L_2Norm(\boldsymbol{g}_t)$
$e_t = \gamma e_{t-1} + (1-\gamma) g_{norm}$
$\boldsymbol{s}_t = \boldsymbol{g}_t$
If $e_t > g_{norm}$
$\boldsymbol{s}_t = (e_t/g_{norm})\boldsymbol{g}_t$
$\boldsymbol{m}_t \leftarrow \beta_1 \boldsymbol{m}_{t-1} + (1-\beta_1)\boldsymbol{s}_t$
$\boldsymbol{v}_t \leftarrow \beta_2 \boldsymbol{v}_{t-1} + (1-\beta_2)\boldsymbol{g}_t^2$
$\rho_\infty \leftarrow 2/(1-\beta_2) - 1$
$\rho_t = \rho_\infty - 2t\beta_2^t/(1-\beta_2^t)$
If $\rho_t \geq 5$
$\rho_u = (\rho_t - 4)(\rho_t - 2)\rho_\infty$
$\rho_d = (\rho_\infty - 4)(\rho_\infty - 2)\rho_t$
$\rho = \sqrt{(1-\beta_2)\rho_u/\rho_d}$
$\alpha_1 = \rho\alpha/(1-\beta_1^t)$
Update
$\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \alpha_1 \boldsymbol{m}_t/(\sqrt{\boldsymbol{v}_t} + \epsilon)$
Else
$\alpha_2 = \alpha/(1-\beta_1^t)$
Update
$\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \alpha_2 \boldsymbol{m}_t$

Algorithm 5: AdaBelief Optimizer

Initialize: $\boldsymbol{\theta}_0, \boldsymbol{m}_0 \leftarrow 0, \boldsymbol{v}_0 \leftarrow 0, t \leftarrow 0$
Hyperparameters: α, β_1, β_2
While $\boldsymbol{\theta}_t$ not converged
$t \leftarrow t+1$
$\boldsymbol{g}_t \leftarrow \nabla_\theta f_t(\boldsymbol{\theta}_{t-1})$
$\boldsymbol{m}_t \leftarrow \beta_1 \boldsymbol{m}_{t-1} + (1-\beta_1)\boldsymbol{g}_t$
$\boldsymbol{v}_t \leftarrow \beta_2 \boldsymbol{v}_{t-1} + (1-\beta_2)(\boldsymbol{g}_t - \boldsymbol{m}_t)^2$
Bias Correction
$\widehat{\boldsymbol{m}_t} \leftarrow \boldsymbol{m}_t/(1-\beta_1^t)$, $\widehat{\boldsymbol{v}_t} \leftarrow \boldsymbol{v}_t/(1-\beta_2^t)$
Update
$\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \alpha\widehat{\boldsymbol{m}_t}/(\sqrt{\widehat{\boldsymbol{v}_t}} + \epsilon)$

Algorithm 6: AdaBeliefNorm (AdaBelief + AdaNorm) Optimizer

Initialize: $\boldsymbol{\theta}_0, \boldsymbol{m}_0 \leftarrow 0, \boldsymbol{v}_0 \leftarrow 0, e_0 \leftarrow 0, t \leftarrow 0$
Hyperparameters: $\alpha, \beta_1, \beta_2, \gamma$
While $\boldsymbol{\theta}_t$ not converged
$t \leftarrow t+1$
$\boldsymbol{g}_t \leftarrow \nabla_\theta f_t(\boldsymbol{\theta}_{t-1})$
$g_{norm} \leftarrow L_2Norm(\boldsymbol{g}_t)$
$e_t = \gamma e_{t-1} + (1-\gamma) g_{norm}$
$\boldsymbol{s}_t = \boldsymbol{g}_t$
If $e_t > g_{norm}$
$\boldsymbol{s}_t = (e_t/g_{norm})\boldsymbol{g}_t$
$\boldsymbol{m}_t \leftarrow \beta_1 \boldsymbol{m}_{t-1} + (1-\beta_1)\boldsymbol{s}_t$
$\boldsymbol{v}_t \leftarrow \beta_2 \boldsymbol{v}_{t-1} + (1-\beta_2)(\boldsymbol{g}_t - \boldsymbol{m}_t)^2$
Bias Correction
$\widehat{\boldsymbol{m}_t} \leftarrow \boldsymbol{m}_t/(1-\beta_1^t)$, $\widehat{\boldsymbol{v}_t} \leftarrow \boldsymbol{v}_t/(1-\beta_2^t)$
Update
$\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} - \alpha\widehat{\boldsymbol{m}_t}/(\sqrt{\widehat{\boldsymbol{v}_t}} + \epsilon)$

Further, we can simplify as,

$$
\begin{aligned}
&\sum_{t=1}^{T} \frac{\hat{m}_{t,i}^2}{\sqrt{t\hat{v}_{t,i}}} \\
&\leq \sum_{t=1}^{T-1} \frac{\hat{m}_{t,i}^2}{\sqrt{t\hat{v}_{t,i}}} + \frac{1}{(1-\beta_1)^2} \sum_{k=1}^{T} \frac{T((1-\beta_1)\beta_1^{T-k} s_{k,i})^2}{\sqrt{T(1-\beta_2)\beta_2^{T-k} g_{k,i}^2}} \\
&= \sum_{t=1}^{T-1} \frac{\hat{m}_{t,i}^2}{\sqrt{t\hat{v}_{t,i}}} + \frac{T}{\sqrt{T(1-\beta_2)}} \sum_{k=1}^{T} \frac{(\beta_1^{T-k} s_{k,i})^2}{\sqrt{\beta_2^{T-k} g_{k,i}^2}} \\
&= \sum_{t=1}^{T-1} \frac{\hat{m}_{t,i}^2}{\sqrt{t\hat{v}_{t,i}}} + \frac{T}{\sqrt{T(1-\beta_2)}} \sum_{k=1}^{T} \left(\frac{\beta_1^2}{\sqrt{\beta_2}}\right)^{T-k} \frac{s_{k,i}^2}{g_{k,i}} \\
&= \sum_{t=1}^{T-1} \frac{\hat{m}_{t,i}^2}{\sqrt{t\hat{v}_{t,i}}} + \frac{T}{\sqrt{T(1-\beta_2)}} \sum_{k=1}^{T} \eta^{T-k} \left(\frac{s_{k,i}}{\sqrt{g_{k,i}}}\right)^2 \\
&\leq \sum_{t=1}^{T-1} \frac{\hat{m}_{t,i}^2}{\sqrt{t\hat{v}_{t,i}}} \\
&\quad + \frac{T}{\sqrt{T(1-\beta_2)}} \sum_{k=1}^{T} \eta^{T-k} \left(\frac{\max(1, \frac{e_k}{||g_k||_2}) g_{k,i}}{\sqrt{g_{k,i}}}\right)^2
\end{aligned}
$$

By considering the bound of e_k and $||g_k||_2$, we can rewrite the above relation as,

$$\sum_{t=1}^{T} \frac{\hat{m}_{t,i}^2}{\sqrt{t\hat{v}_{t,i}}} \leq \sum_{t=1}^{T-1} \frac{\hat{m}_{t,i}^2}{\sqrt{t\hat{v}_{t,i}}} + \frac{T}{\sqrt{T(1-\beta_2)}} \sum_{k=1}^{T} \eta^{T-k} \frac{G_\infty^2}{G^2} ||g_{k,i}||_2$$

Similarly, after considering the upper bound of the rest of the terms in the summation, we can get as follows,

$$\sum_{t=1}^{T} \frac{\hat{m}_{t,i}^2}{\sqrt{t\hat{v}_{t,i}}} \leq \frac{G_\infty^2}{G^2\sqrt{(1-\beta_2)}} \sum_{t=1}^{T} \frac{||g_{t,i}||_2}{\sqrt{t}} \sum_{j=0}^{T-t} t\eta^j \leq \frac{G_\infty^2}{G^2\sqrt{(1-\beta_2)}} \sum_{t=1}^{T} \frac{||g_{t,i}||_2}{\sqrt{t}} \sum_{j=0}^{T} t\eta^j$$

We can obtain $\sum_t t\eta^t < \frac{1}{(1-\eta)^2}$ for $\eta < 1$ using the upper bound on the arithmetic-geometric series. Hence,

$$\sum_{t=1}^{T} \frac{\hat{m}_{t,i}^2}{\sqrt{t\hat{v}_{t,i}}} \leq \frac{G_\infty^2}{G^2(1-\eta)^2\sqrt{1-\beta_2}} \sum_{t=1}^{T} \frac{||g_{k,i}||_2}{\sqrt{t}}$$

By applying Lemma 10.3 of [1], we can get,

$$\sum_{t=1}^{T} \frac{\hat{m}_{t,i}^2}{\sqrt{t\hat{v}_{t,i}}} \leq \frac{2G_\infty^3}{G^2(1-\eta)^2\sqrt{1-\beta_2}} ||g_{1:T,i}||_2$$

□

Theorem 1. *Let the bounded gradients for function f_t (i.e., $||g_{t,\theta}||_2 \leq G$ and $||g_{t,\theta}||_\infty \leq G_\infty$) for all $\theta \in R^d$. Also assume that AdamNorm produces the bounded distance between any θ_t (i.e., $||\theta_n - \theta_m||_2 \leq D$ and $||\theta_n - \theta_m||_\infty \leq D_\infty$ for any $m, n \in \{1, ..., T\}$). Let $\eta \triangleq \frac{\beta_1^2}{\sqrt{\beta_2}}$, $\beta_1, \beta_2 \in [0, 1)$ satisfy $\frac{\beta_1^2}{\sqrt{\beta_2}} < 1$, $\alpha_t = \frac{\alpha}{\sqrt{t}}$, and $\beta_{1,t} = \beta_1\lambda^{t-1}, \lambda \in (0, 1)$ with λ is typically close to 1, e.g., $1 - 10^{-8}$. For all $T \geq 1$, the proposed AdamNorm optimizer shows the following guarantee:*

$$R(T) \leq \frac{D^2}{2\alpha(1-\beta_1)} \sum_{i=1}^{d} \sqrt{T\hat{v}_{T,i}} + \frac{\alpha(1+\beta_1)G_\infty^3}{(1-\beta_1)\sqrt{1-\beta_2}(1-\eta)^2G^2} \sum_{i=1}^{d} ||g_{1:T,i}||_2 + \sum_{i=1}^{d} \frac{D_\infty^2 G_\infty \sqrt{1-\beta_2}}{2\alpha(1-\beta_1)(1-\lambda)^2}$$

Proof. Using Lemma 10.2 of Adam [1], we can write as

$$f_t(\theta_t) - f_t(\theta^*) \leq g_t^T(\theta_t - \theta^*) = \sum_{i=1}^{d} g_{t,i}(\theta_{t,i} - \theta_{,i}^*)$$

We can write following from the AdamNorm update rule, ignoring ϵ,

$$\theta_{t+1} = \theta_t - \frac{\alpha_t \hat{m}_t}{\sqrt{\hat{v}_t}} = \theta_t - \frac{\alpha_t}{(1-\beta_1^t)} \Big(\frac{\beta_{1,t}}{\sqrt{\hat{v}_t}} m_{t-1} + \frac{(1-\beta_{1,t})}{\sqrt{\hat{v}_t}} g_t \Big)$$

where $\beta_{1,t}$ is the 1^{st} order moment coefficient at t^{th} iteration and β_1^t is the t^{th} power of initial 1^{st} order moment coefficient.
For i^{th} dimension of parameter vector $\theta_t \in R^d$, we can write

$$(\theta_{t+1,i} - \theta_{,i}^*)^2 = (\theta_{t,i} - \theta_{,i}^*)^2 - \frac{2\alpha_t}{1-\beta_1^t} \Big(\frac{\beta_{1,t}}{\sqrt{\hat{v}_{t,i}}} m_{t-1,i} + \frac{(1-\beta_{1,t})}{\sqrt{\hat{v}_{t,i}}} g_{t,i} \Big) (\theta_{t,i} - \theta_{,i}^*) + \alpha_t^2 (\frac{\hat{m}_{t,i}}{\hat{v}_{t,i}})^2$$

The above equation can be reordered as

$$g_{t,i}(\theta_{t,i} - \theta_{,i}^*) = \frac{(1-\beta_1^t)\sqrt{\hat{v}_{t,i}}}{2\alpha_t(1-\beta_{1,t})} \Big((\theta_{t,i} - \theta_{,i}^*)^2 - (\theta_{t+1,i} - \theta_{,i}^*)^2 \Big) + \frac{\beta_{1,t}}{1-\beta_{1,t}} (\theta_{,i}^* - \theta_{t,i}) m_{t-1,i} + \frac{\alpha_t(1-\beta_1^t)}{2(1-\beta_{1,t})} \frac{(\hat{m}_{t,i})^2}{\sqrt{\hat{v}_{t,i}}}.$$

Further, it can be written as

$$\begin{aligned} & g_{t,i}(\theta_{t,i} - \theta_{,i}^*) \\ &= \frac{(1-\beta_1^t)\sqrt{\hat{v}_{t,i}}}{2\alpha_t(1-\beta_{1,t})} \Big((\theta_{t,i} - \theta_{,i}^*)^2 - (\theta_{t+1,i} - \theta_{,i}^*)^2 \Big) \\ &\quad + \sqrt{\frac{\beta_{1,t}}{\alpha_{t-1}(1-\beta_{1,t})} (\theta_{,i}^* - \theta_{t,i})^2 \sqrt{\hat{v}_{t-1,i}}} \sqrt{\frac{\beta_{1,t}\alpha_{t-1}(m_{t-1,i})^2}{(1-\beta_{1,t})\sqrt{\hat{v}_{t-1,i}}}} \\ &\quad + \frac{\alpha_t(1-\beta_1^t)}{2(1-\beta_{1,t})} \frac{(\hat{m}_{t,i})^2}{\sqrt{\hat{v}_{t,i}}} \end{aligned}$$

Based on Young's inequality, $ab \leq a^2/2 + b^2/2$ and fact that $\beta_{1,t} \leq \beta_1$, the above equation can be reordered as

$$\begin{aligned} g_{t,i}(\theta_{t,i} - \theta_{,i}^*) &\leq \frac{1}{2\alpha_t(1-\beta_1)} \Big((\theta_{t,i} - \theta_{,i}^*)^2 - (\theta_{t+1,i} - \theta_{,i}^*)^2 \Big) \sqrt{\hat{v}_{t,i}} \\ &\quad + \frac{\beta_{1,t}}{2\alpha_{t-1}(1-\beta_{1,t})} (\theta_{,i}^* - \theta_{t,i})^2 \sqrt{\hat{v}_{t-1,i}} \\ &\quad + \frac{\beta_1\alpha_{t-1}(m_{t-1,i})^2}{2(1-\beta_1)\sqrt{\hat{v}_{t-1,i}}} \\ &\quad + \frac{\alpha_t}{2(1-\beta_1)} \frac{(\hat{m}_{t,i})^2}{\sqrt{\hat{v}_{t,i}}} \end{aligned}$$

We use the Lemma 1 and derive the regret bound by aggregating it across all the dimensions for $i \in \{1, \ldots, d\}$ and all the sequence of convex functions for $t \in \{1, \ldots, T\}$ in the upper bound of $f_t(\theta_t) - f_t(\theta^*)$ as

$$
\begin{aligned}
&R(T) \\
&\leq \sum_{i=1}^{d} \frac{1}{2\alpha_1(1-\beta_1)}(\theta_{1,i} - \theta^*_{,i})^2 \sqrt{\hat{v}_{1,i}} \\
&\quad + \sum_{i=1}^{d} \sum_{t=2}^{T} \frac{1}{2(1-\beta_1)}(\theta_{t,i} - \theta^*_{,i})^2 (\frac{\sqrt{\hat{v}_{t,i}}}{\alpha_t} - \frac{\sqrt{\hat{v}_{t-1,i}}}{\alpha_{t-1}}) \\
&\quad + \frac{\beta_1 \alpha G_\infty^3}{(1-\beta_1)\sqrt{1-\beta_2}(1-\eta)^2 G^2} \sum_{i=1}^{d} ||g_{1:T,i}||_2 \\
&\quad + \frac{\alpha G_\infty^3}{(1-\beta_1)\sqrt{1-\beta_2}(1-\eta)^2 G^2} \sum_{i=1}^{d} ||g_{1:T,i}||_2 \\
&\quad + \sum_{i=1}^{d} \sum_{t=1}^{T} \frac{\beta_{1,t}}{2\alpha_t(1-\beta_{1,t})}(\theta^*_{,i} - \theta_{t,i})^2 \sqrt{\hat{v}_{t,i}}
\end{aligned}
$$

By utilizing the assumptions that $\alpha = \alpha_t\sqrt{t}$, $||\theta_t - \theta^*||_2 \leq D$ and $||\theta_m - \theta_n||_\infty \leq D_\infty$, we can write as

$$
\begin{aligned}
R(T) &\leq \frac{D^2}{2\alpha(1-\beta_1)} \sum_{i=1}^{d} \sqrt{T\hat{v}_{T,i}} \\
&\quad + \frac{\alpha(1+\beta_1)G_\infty^3}{(1-\beta_1)\sqrt{1-\beta_2}(1-\eta)^2 G^2} \sum_{i=1}^{d} ||g_{1:T,i}||_2 \\
&\quad + \frac{D_\infty^2}{2\alpha} \sum_{i=1}^{d} \sum_{t=1}^{t} \frac{\beta_{1,t}}{(1-\beta_{1,t})} \sqrt{t\hat{v}_{t,i}} \\
&\leq \frac{D^2}{2\alpha(1-\beta_1)} \sum_{i=1}^{d} \sqrt{T\hat{v}_{T,i}} \\
&\quad + \frac{\alpha(1+\beta_1)G_\infty^3}{(1-\beta_1)\sqrt{1-\beta_2}(1-\eta)^2 G^2} \sum_{i=1}^{d} ||g_{1:T,i}||_2 \\
&\quad + \frac{D_\infty^2 G_\infty \sqrt{1-\beta_2}}{2\alpha} \sum_{i=1}^{d} \sum_{t=1}^{t} \frac{\beta_{1,t}}{(1-\beta_{1,t})} \sqrt{t}
\end{aligned}
$$

It is shown in Adam [1] that $\sum_{t=1}^{t} \frac{\beta_{1,t}}{(1-\beta_{1,t})}\sqrt{t} \leq \frac{1}{(1-\beta_1)(1-\eta)^2}$. Thus, the regret bound can be written as

$$
\begin{aligned}
R(T) &\leq \frac{D^2}{2\alpha(1-\beta_1)} \sum_{i=1}^{d} \sqrt{T\hat{v}_{T,i}} \\
&\quad + \frac{\alpha(1+\beta_1)G_\infty^3}{(1-\beta_1)\sqrt{1-\beta_2}(1-\eta)^2 G^2} \sum_{i=1}^{d} ||g_{1:T,i}||_2 \\
&\quad + \sum_{i=1}^{d} \frac{D_\infty^2 G_\infty \sqrt{1-\beta_2}}{2\alpha(1-\beta_1)(1-\lambda)^2}
\end{aligned}
$$

□

References

[1] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, 2015.