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: \( \theta_0, m_0 \leftarrow 0, v_0 \leftarrow 0, t \leftarrow 0 \)

Hyperparameters: \( \alpha, \beta_1, \beta_2 \)

While \( \theta_t \) not converged

\[
\begin{align*}
g_t &\leftarrow \nabla_{\theta_t}f_t(\theta_{t-1}) \\
x_t &\leftarrow 1/(1 + e^{-|g_t - g_{t-1}|}) \\
m_t &\leftarrow \beta_1 m_{t-1} + (1 - \beta_1) g_t \\
v_t &\leftarrow \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \\
\text{Bias Correction: } & \quad \hat{m}_t \leftarrow m_t/(1 - \beta_1^t), \hat{v}_t \leftarrow v_t/(1 - \beta_2^t) \\
\text{Update: } & \quad \theta_t \leftarrow \theta_{t-1} - \alpha \hat{x}_t \hat{m}_t/\sqrt{\hat{v}_t + \epsilon}
\end{align*}
\]

B. Convergence Proof

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

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

Proof. Under the assumption, \( \frac{1 - \beta_2^t}{(1 - \beta_1^t)} \leq \frac{1}{(1 - \beta_1^t)} \). We can use the update rules of AdamNorm and expand the last term in the summation,
### Algorithm 3: Radam Optimizer

**Initialize:** $\theta_0, m_0 \leftarrow 0, v_0 \leftarrow 0, t \leftarrow 0$

**Hyperparameters:** $\alpha, \beta_1, \beta_2$

**While** $\theta_t$ not converged

$t \leftarrow t + 1$

$g_t \leftarrow \nabla f_t(\theta_{t-1})$

$m_t \leftarrow \beta_1 m_{t-1} + (1 - \beta_1) g_t$

$v_t \leftarrow \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$

$\rho_\infty \leftarrow 2/(1 - \beta_2) - 1$

$\rho_t = \rho_\infty - 2\beta_2^2/(1 - \beta_2^2)$

**If** $\rho_t \geq 5$

$\rho_u = (\rho_t - 4)(\rho_t - 2)\rho_\infty$

$\rho_d = (\rho_\infty - 4)(\rho_\infty - 2)\rho_t$

$\rho = \sqrt{(1 - \beta_2)\rho_u/\rho_d}$

$\alpha_1 = \rho\alpha/(1 - \beta_1^2)$

**Update**

$\theta_t \leftarrow \theta_{t-1} - \alpha_1 m_t/(\sqrt{v_t} + \epsilon)$

**Else**

$\alpha_2 = \alpha/(1 - \beta_1^2)$

**Update**

$\theta_t \leftarrow \theta_{t-1} - \alpha_2 m_t$

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

**Initialize:** $\theta_0, m_0 \leftarrow 0, v_0 \leftarrow 0, t \leftarrow 0$

**Hyperparameters:** $\alpha, \beta_1, \beta_2, \gamma$

**While** $\theta_t$ not converged

$t \leftarrow t + 1$

$g_t \leftarrow \nabla f_t(\theta_{t-1})$

$g_{\text{norm}} \leftarrow L_2\text{Norm}(g_t)$

$e_t = \gamma e_{t-1} + (1 - \gamma) g_{\text{norm}}$

$s_t = g_t$

**If** $e_t > g_{\text{norm}}$

$s_t = (e_t/g_{\text{norm}})g_t$

$m_t \leftarrow \beta_1 m_{t-1} + (1 - \beta_1) s_t$

$v_t \leftarrow \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$

$\rho_\infty \leftarrow 2/(1 - \beta_2) - 1$

$\rho_t = \rho_\infty - 2\beta_2^2/(1 - \beta_2^2)$

**If** $\rho_t \geq 5$

$\rho_u = (\rho_t - 4)(\rho_t - 2)\rho_\infty$

$\rho_d = (\rho_\infty - 4)(\rho_\infty - 2)\rho_t$

$\rho = \sqrt{(1 - \beta_2)\rho_u/\rho_d}$

$\alpha_1 = \rho\alpha/(1 - \beta_1^2)$

**Update**

$\theta_t \leftarrow \theta_{t-1} - \alpha_1 m_t/(\sqrt{v_t} + \epsilon)$

**Else**

$\alpha_2 = \alpha/(1 - \beta_1^2)$

**Update**

$\theta_t \leftarrow \theta_{t-1} - \alpha_2 m_t$

### Algorithm 5: AdaBelief Optimizer

**Initialize:** $\theta_0, m_0 \leftarrow 0, v_0 \leftarrow 0, t \leftarrow 0$

**Hyperparameters:** $\alpha, \beta_1, \beta_2, \gamma$

**While** $\theta_t$ not converged

$t \leftarrow t + 1$

$g_t \leftarrow \nabla f_t(\theta_{t-1})$

$g_{\text{norm}} \leftarrow L_2\text{Norm}(g_t)$

$e_t = \gamma e_{t-1} + (1 - \gamma) g_{\text{norm}}$

$s_t = g_t$

**If** $e_t > g_{\text{norm}}$

$s_t = (e_t/g_{\text{norm}})g_t$

$m_t \leftarrow \beta_1 m_{t-1} + (1 - \beta_1) s_t$

$v_t \leftarrow \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$

**Bias Correction**

$\hat{m}_t = m_t/(1 - \beta_1^t), \hat{v}_t = v_t/(1 - \beta_2^t)$

**Update**

$\theta_t \leftarrow \theta_{t-1} - \alpha\hat{m}_t/(\sqrt{\hat{v}_t} + \epsilon)$

### Algorithm 6: AdaBeliefNorm (AdaBelief + AdaNorm) Optimizer

**Initialize:** $\theta_0, m_0 \leftarrow 0, v_0 \leftarrow 0, e_0 \leftarrow 0, t \leftarrow 0$

**Hyperparameters:** $\alpha, \beta_1, \beta_2, \gamma$

**While** $\theta_t$ not converged

$t \leftarrow t + 1$

$g_t \leftarrow \nabla f_t(\theta_{t-1})$

$g_{\text{norm}} \leftarrow L_2\text{Norm}(g_t)$

$e_t = \gamma e_{t-1} + (1 - \gamma) g_{\text{norm}}$

$s_t = g_t$

**If** $e_t > g_{\text{norm}}$

$s_t = (\epsilon_t/g_{\text{norm}})g_t$

$m_t \leftarrow \beta_1 m_{t-1} + (1 - \beta_1) s_t$

$v_t \leftarrow \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$

**Bias Correction**

$\hat{m}_t = m_t/(1 - \beta_1^t), \hat{v}_t = v_t/(1 - \beta_2^t)$

**Update**

$\theta_t \leftarrow \theta_{t-1} - \alpha\hat{m}_t/(\sqrt{\hat{v}_t} + \epsilon)$

Further, we can simplify as,

$$
\sum_{t=1}^{T} \hat{m}^2_{t,i} \left/ \sqrt{\hat{v}_{t,i}} \right.
\leq \sum_{t=1}^{T-1} \frac{\hat{m}^2_{t,i}}{\sqrt{\hat{v}_{t,i}}} + \frac{1}{(1 - \beta_1)^2} \sum_{k=1}^{T} \frac{\left(T((1 - \beta_1)\beta^{T-k} s_{k,i})^2 \right)}{\sqrt{\hat{v}_{t,i}}^2}
$$

$$
\leq \sum_{t=1}^{T-1} \frac{\hat{m}^2_{t,i}}{\sqrt{\hat{v}_{t,i}}} + \frac{T}{\sqrt{T(1 - \beta_2)}} \sum_{k=1}^{T} \frac{\left(\beta^{T-k} s_{k,i})^2 \right)}{\sqrt{\hat{v}_{t,i}}^2}
$$

$$
= \sum_{t=1}^{T-1} \frac{\hat{m}^2_{t,i}}{\sqrt{\hat{v}_{t,i}}} + \frac{T}{\sqrt{T(1 - \beta_2)}} \sum_{k=1}^{T} \frac{\left(\beta^{T-k} s_{k,i})^2 \right)}{\sqrt{\hat{v}_{t,i}}^2}
$$

$$
\leq \sum_{t=1}^{T-1} \frac{\hat{m}^2_{t,i}}{\sqrt{\hat{v}_{t,i}}} + \frac{T}{\sqrt{T(1 - \beta_2)}} \sum_{k=1}^{T} \frac{(\beta^{T-k} s_{k,i})^2}{\sqrt{\hat{v}_{t,i}}^2}
$$

$$
\leq \sum_{t=1}^{T-1} \frac{\hat{m}^2_{t,i}}{\sqrt{\hat{v}_{t,i}}} + \frac{T}{\sqrt{T(1 - \beta_2)}} \sum_{k=1}^{T} \frac{(\beta^{T-k} s_{k,i})^2}{\sqrt{\hat{v}_{t,i}}^2}
$$

$$
= \sum_{t=1}^{T-1} \frac{\hat{m}^2_{t,i}}{\sqrt{\hat{v}_{t,i}}} + \frac{T}{\sqrt{T(1 - \beta_2)}} \sum_{k=1}^{T} \frac{(\beta^{T-k} s_{k,i})^2}{\sqrt{\hat{v}_{t,i}}^2}
$$

$$
\leq \sum_{t=1}^{T} \frac{\hat{m}^2_{t,i}}{\sqrt{\hat{v}_{t,i}}} + \frac{T}{\sqrt{T(1 - \beta_2)}} \sum_{k=1}^{T} \frac{(\beta^{T-k} s_{k,i})^2}{\sqrt{\hat{v}_{t,i}}^2}
$$

$$
= \sum_{t=1}^{T} \frac{\hat{m}^2_{t,i}}{\sqrt{\hat{v}_{t,i}}} + \frac{T}{\sqrt{T(1 - \beta_2)}} \sum_{k=1}^{T} \frac{(\beta^{T-k} s_{k,i})^2}{\sqrt{\hat{v}_{t,i}}^2}
$$

$$
+ \frac{T}{\sqrt{T(1 - \beta_2)}} \sum_{k=1}^{T} \frac{(\beta^{T-k} s_{k,i})^2}{\sqrt{\hat{v}_{t,i}}^2}
$$
By considering the bound of $e_k$ and $\|g_k\|_2$, we can rewrite the above relation as,

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

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

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

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

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

We can write following from the AdamNorm update rule, ignoring $\epsilon$,

$$\theta_{t+1} = \theta_t - \frac{\alpha_t \hat{m}_t}{\sqrt{v_t}}$$

$$= \theta_t - \frac{\alpha_t}{(1 - \beta_1)} \left( \frac{\beta_{1,t}}{\sqrt{v_{t,i}}} g_{t,i} + \alpha_t^2 \frac{\hat{m}_{t,i}^2}{v_{t,i}} \right)$$

where $\beta_{1,t}$ is the $t^{th}$ order moment coefficient at $t^{th}$ iteration and $\beta_1^2$ is the $t^{th}$ power of initial $1^{st}$ order moment coefficient.

For $t^{th}$ dimension of parameter vector $\theta_t \in R^d$, we can write

$$(\theta_{t+1,i} - \theta^*_i)^2 = (\theta_{t,i} - \theta^*_i)^2 - 2 \alpha_t \left( \frac{\beta_{1,t}}{\sqrt{v_{t,i}}} g_{t,i} + \alpha_t^2 \frac{\hat{m}_{t,i}^2}{v_{t,i}} \right)$$

$$+ \left( \frac{\beta_{1,t}}{1 - \beta_1} \right) \beta_1 m_{t-1,i} - \beta_1 \eta t m_{t-1,i}$$

The above equation can be reordered as

$$g_{t,i}(\theta_{t,i} - \theta^*_i) = \frac{(1 - \beta_1^2)}{2 \alpha_t (1 - \beta_{1,t})} (\theta_{t,i} - \theta^*_i)^2$$

$$- (\theta_{t+1,i} - \theta^*_i)^2$$

$$+ \beta_1 t g_{t,i} + \alpha_t^2 \frac{\hat{m}_{t,i}^2}{v_{t,i}}$$

Further, it can be written as

$$g_{t,i}(\theta_{t,i} - \theta^*_i) = \frac{(1 - \beta_1^2)}{2 \alpha_t (1 - \beta_{1,t})} (\theta_{t,i} - \theta^*_i)^2$$

$$- (\theta_{t+1,i} - \theta^*_i)^2$$

$$+ \beta_1 t g_{t,i} + \alpha_t^2 \frac{\hat{m}_{t,i}^2}{v_{t,i}}$$

Based on Young’s inequality, $ab \leq a^2/2 + b^2/2$ and fact that $\beta_1 \leq \beta_1$, the above equation can be reordered as

$$g_{t,i}(\theta_{t,i} - \theta^*_i) \leq \frac{1}{2 \alpha_t (1 - \beta_{1,t})} (\theta_{t,i} - \theta^*_i)^2$$

$$- (\theta_{t+1,i} - \theta^*_i)^2$$

$$+ \beta_1 t g_{t,i} + \alpha_t^2 \frac{\hat{m}_{t,i}^2}{v_{t,i}}$$

$$+ \frac{\beta_1}{2(1 - \beta_{1,t})} \beta_1 (\theta_{t,i} - \theta^*_i)^2$$

$$+ 2 \alpha_t^2 \frac{\hat{m}_{t,i}^2}{v_{t,i}}$$

$$+ \frac{\beta_1}{2(1 - \beta_{1,t})} \beta_1 \eta t m_{t-1,i}$$

$$+ \frac{\alpha_t}{2(1 - \beta_{1,t})} \frac{\hat{m}_{t,i}^2}{v_{t,i}}$$

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

$$R(T) \leq \frac{D^2}{(1 - \beta_1)} \sum_{i=1}^{d} \sqrt{T \hat{v}_{T,i}}$$

$$+ \frac{\alpha_t (1 + \beta_1) G_\infty^3}{(1 - \beta_1) \sqrt{1 - \beta_2 (1 - \eta)^2} G^2} \sum_{i=1}^{d} \|g_{1:T,i}\|_2$$

$$+ \frac{D_{\infty} G_\infty \sqrt{1 - \beta_2}}{2 \alpha_t (1 - \beta_1)(1 - \lambda)^2} \sum_{i=1}^{d} \|g_{1:T,i}\|_2$$

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

$$f_t(\theta_t) - f_t(\theta^*) \leq \frac{d}{\alpha_t} g_t(\theta_t - \theta^*) = \sum_{i=1}^{d} g_{t,i}(\theta_{t,i} - \theta^*_{i})$$
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 \( f_t(\theta_i) - f_t(\theta^*) \) as

\[
R(T) \leq \sum_{i=1}^{d} \frac{1}{2\alpha_t(1-\beta_1)}(\theta_{t,i} - \theta^*_{t,i})^2 \sqrt{\hat{v}_{t,i}}
+ \sum_{i=1}^{d} \sum_{t=2}^{T} \frac{1}{2(1-\beta_1)}(\theta_{t,i} - \theta^*_{t,i})^2 \left( \frac{\sqrt{\hat{v}_{t,i}}}{\alpha_t} - \frac{\sqrt{\hat{v}_{t-1,i}}}{\alpha_{t-1}} \right)
+ \frac{\beta_1 \alpha G^3}{(1-\beta_1)\sqrt{1-\beta_2(1-\eta)^2}G^2} \sum_{i=1}^{d} ||g_{1:T,i}||_2
+ \frac{\alpha G^3}{(1-\beta_1)\sqrt{1-\beta_2(1-\eta)^2}G^2} \sum_{i=1}^{d} ||g_{1:T,i}||_2
+ \frac{\beta_1 t}{2\alpha_t(1-\beta_1,t)}(\theta^*_{t,i} - \theta_{t,i})^2 \sqrt{\hat{v}_{t,i}}
\]

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

\[
R(T) \leq \frac{D^2}{2\alpha(1-\beta_1)} \sum_{i=1}^{d} \sqrt{T \hat{v}_{T,i}}
+ \frac{\alpha(1 + \beta_1)G^3}{(1-\beta_1)\sqrt{1-\beta_2(1-\eta)^2}G^2} \sum_{i=1}^{d} ||g_{1:T,i}||_2
+ \frac{D^2}{2\alpha} \sum_{i=1}^{d} \sum_{t=1}^{T} \frac{\beta_{1,t}}{(1-\beta_{1,t})} \sqrt{T \hat{v}_{t,i}}
\]

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

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

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

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.