Miles Pophal
Date: April 30, 2024

AdamW can Fail

Review of Adam Failure Proof

We start with the Adam Algorithm with bias correction $\hat{m}, \hat{v}$ although it may be removed for simplicity. Our candidate for the counterexample is $\mathcal{F} = [-1, 1]$ and for some $C > 2$,

Algorithm 1 The Adam Algorithm

Require: $x_1 \in \mathcal{F}$ initial point, $\{\alpha_t\}_{t=1}^T$ step sizes, $\beta_1, \beta_2 \geq 0$ and $\alpha < \sqrt{1-\beta_2}$.
  $m_0, v_0 \leftarrow 0$
  for $t = 1, \ldots, T$ **do**
    $g_t \leftarrow \nabla f_t(x_t)$
    $m_t \leftarrow \beta_1 m_{t-1} + (1-\beta_1) g_t$
    $v_t \leftarrow \beta_2 v_{t-1} + (1-\beta_2) g_t^2$
    $\hat{m}_t \leftarrow m_t/(1-\beta_1^t)$
    $\hat{v}_t \leftarrow v_t/(1-\beta_2^t)$
    $\hat{x}_{t+1} \leftarrow x_t - \alpha_t \hat{m}_t/\sqrt{\hat{v}_t}$
    $x_{t+1} \leftarrow \Pi_{\mathcal{F}}(\hat{x}_{t+1})$ ▷ Projection
  end for
  return x_T

$$f_t(x) = \begin{cases} Cx, & t \mod 3 = 1 \\ -x, & \text{else} \end{cases} \implies \nabla f_t(\cdot) = \begin{cases} C, & t \mod 3 = 1 \\ -1, & \text{else} \end{cases}$$

we can see $x = -1$ produces minimal regret because for fixed x the regret cycles as

$$Cx + (-x) + (-x) = (C-2)x.$$

This is monotonic increasing on $[-1, 1]$ and hence has the minimum at $x = -1$. WLOG let $x_1 = 1$ (translate the system otherwise), $\beta_1 = 0, \beta_2 = 1/(1+C^2)$ to satisfy $\beta_1^2 \leq \sqrt{\beta_2}$ from theorem 4.1 in [1]. We also set $\alpha_t = \alpha_{\sqrt{t}}$, although the experiments show that it holds even with the default setting for learning rate. As in the paper, we prove $x_t > 0$ with $x_{3t+1} = 1$ (this allows us to keep cyclic behavior) via induction (base case done by assumption). Our assumption is $x_{3t+1} = 1$ and $x_t > 0$ up until the fixed $3t+1$. From the definition of Adam 1, the $3t+2$ update is given by

$$\beta_1 = 0 \implies m_t = g_t = \hat{m}_t$$

so $m_{3t+1} = C$. Hence,

$$\hat{x}_{3t+2} \geq x_{3t+1} - \frac{\alpha C}{\sqrt{(3t+1)(\beta_2 v_{3t} + (1-\beta_2)C^2)}} = 1 - \frac{\alpha C}{\sqrt{(3t+1)(\beta_2 v_{3t} + (1-\beta_2 C^2))}}, \tag{1}$$

where we substitute in $\alpha_{3t+1} = \alpha/\sqrt{3t+1}$ and likewise the update for $\sqrt{\hat{v}_{3t+1}}$. The reason for the inequality is because $\hat{v} \geq v$ and since we take v here, it becomes a lower bound (note Reddi et. al. noted the analysis is similar, but they ignored this term). Since $\beta_2 v_{3t}$ is positive, removing it from the denominator increases the value, hence

$$\frac{\alpha C}{\sqrt{(3t+1)(\beta_2 v_{3t} + (1-\beta_2 C^2))}} \leq \frac{\alpha C}{\sqrt{(3t+1)(1-\beta_2 C^2)}} = \frac{\alpha}{\sqrt{(3t+1)(1-\beta_2)}}$$

because the C terms cancel. This can be bounded strictly by 1 because $\alpha < \sqrt{1-\beta_2}$, so

$$\frac{\alpha}{\sqrt{(3t+1)(1-\beta_2)}} < \frac{\sqrt{1-\beta_2}}{\sqrt{(3t+1)(1-\beta_2)}} = \frac{1}{\sqrt{3t+1}} < 1$$

and hence, $0 < \hat{x}_{3t+2} < 1$. When we project this to get $x_{3t+2} = \Pi_{\mathcal{F}}(\hat{x}_{3t+2}) = \hat{x}_{3t+2}$ because it already is inside $[-1, 1]$. Furthermore because it is strictly positive,

$$\hat{x}_{3t+3} = x_{3t+2} + \frac{\alpha}{\sqrt{(3t+2)(\beta_2 v_{3t+1} + (1-\beta_2))}} > 0 \tag{2}$$

since $\alpha > 0$ and the gradient is -1 which flips the sign. Hence, $\hat{x}_{3t+3} > 0$, but it may also be above 1 so all we can say for the x_{3t+4} iterate is

$$\hat{x}_{3t+4} = x_{3t+3} + \frac{\alpha}{\sqrt{(3t+3)(\beta_2 v_{3t+2} + (1-\beta_2))}} = \min\{\hat{x}_{3t+3}, 1\} + \frac{\alpha}{\sqrt{(3t+3)(\beta_2 v_{3t+2} + (1-\beta_2))}}$$

as the minimum is the projection operation when we know the inside is positive. If $\hat{x}_{3t+3} \geq 1$, we are done because adding to a positive term makes $\hat{x}_{3t+4} \geq 1 \implies x_{3t+4} = 1$ which is what we want to prove. Otherwise, we have

$$\hat{x}_{3t+4} = \underbrace{\hat{x}_{3t+3}}_{\text{equal to projection}} + \frac{\alpha}{\sqrt{(3t+3)(\beta_2 v_{3t+2} + (1-\beta_2))}},$$

and we can then unwind this all the way back to the known quantity $\hat{x}_{3t+2}$ to get

$$\hat{x}_{3t+4} = x_{3t+2} + \frac{\alpha}{\sqrt{(3t+2)(\beta_2 v_{3t+1} + (1-\beta_2))}} + \frac{\alpha}{\sqrt{(3t+3)(\beta_2 v_{3t+2} + (1-\beta_2))}}$$

by substituting in equation (2). As we showed $x_{3t+2} = \hat{x}_{3t+2}$, we can perform one more unwind to get

$$\begin{aligned}\hat{x}_{3t+4} \geq 1 - &\underbrace{\frac{\alpha C}{\sqrt{(3t+1)(\beta_2 v_{3t} + (1-\beta_2 C^2))}}}_{:=T_1} \qquad \{\text{unwound } x_{3t+2}\\ &+ \underbrace{\frac{\alpha}{\sqrt{(3t+2)(\beta_2 v_{3t+1} + (1-\beta_2))}} + \frac{\alpha}{\sqrt{(3t+3)(\beta_2 v_{3t+2} + (1-\beta_2))}}}_{:=T_2},\end{aligned}$$

using equation (1). Now we just have to show $T_1 \leq T_2$ like we did for the iterate x_{3t+2}. Following that train, using the fact v_{3t} is positive, we can say

$$T_1 \leq \frac{\alpha}{\sqrt{(3t+1)(1-\beta_2)}}. \tag{3}$$

Next, we can note because v_t is a convex combination of v_{t-1} and g_t^2, we can say $v_t \leq C^2$ for all t (since each gradient is bounded by C too) using induction. With this bound, we can use it as a lower bound because it's in the denominator to get

$$T_2 \geq \frac{\alpha}{\sqrt{\beta_2 C^2 + (1-\beta_2)}} \left(\frac{1}{\sqrt{3t+2}} + \frac{1}{\sqrt{3t+3}} \right).$$

We would like to get the denominators like $3t+1$ to compare with T_1 while bounding this from below, but immediately reducing them actually decreases the denominator and ruins the lower bound. We can also multiply everything inside by 2 because $2(3t+1) > 3t+3 > 3t+2$ for all $t \geq 1$. Hence,

$$T_2 \geq \frac{\alpha}{\sqrt{\beta_2 C^2 + (1-\beta_2)}} \left(\frac{1}{\sqrt{2(3t+1)}} + \frac{1}{\sqrt{2(3t+1)}} \right) = \frac{\alpha\sqrt{2}}{(3t+1)(\beta_2 C^2 + (1-\beta_2))}.$$

For this to be comparable we need to get rid of the $\sqrt{2}$ and fix the denominator, i.e., we want (compare with T_1 in equation (3))

$$\frac{\sqrt{2}}{\sqrt{\beta_2 C^2 + (1-\beta_2)}} = \frac{1}{\sqrt{(1-\beta_2)}} \iff 2(1-\beta_2) = \beta_2 C^2 + 1 - \beta_2 \tag{4}$$

$$\iff 2 - \beta_2 = \beta_2 C^2 + 1 \iff 1 = \beta_2 + \beta_2 C^2 \iff \beta_2 = \frac{1}{1+C^2}, \tag{5}$$

motivating our choice for this β_2. Hence,

$$T_2 \geq \frac{\alpha\sqrt{2}}{(3t+1)(\beta_2 C^2 + (1-\beta_2))} = \frac{\alpha}{(3t+1)(1-\beta_2)} = T_1 \implies \hat{x}_{3t+4} > 1$$

and therefore $x_{3t+4} = \Pi_{\mathcal{F}}(\hat{x}_{3t+4}) = 1$ as desired and hence, we are done.

This is a counterexample because if we look at the regret along a cycle, we have

$$f_{3t+1}(x_{3t+1}) + f_{3t+2}(x_{3t+2}) + f_{3t+3}(x_{3t+3}) = C - x_{3t+2} - x_{3t+3} \geq C - 2$$

as both $x_{3t+2}, x_{3t+3} \leq 1$. We compare this with the optimal $x = -1$ to see

$$f_{3t+1}(-1) + f_{3t+2}(-1) + f_{3t+3}(-1) = -C + 2 \implies R \geq (C-2) - (-C+2) = 2C - 4$$

and hence for each cycle the regret will increase as $C > 2$. Since this happens every 3 timesteps, the average regret

$$R_T/T \geq \frac{1}{3}[2C - 4] \not\to 0$$

giving us the counterexample.

***Ada*pted to AdamW**

For AdamW, we have the algorithm 2 below. for weight decay parameter γ with default initialization $\gamma = 1/100$. Instead

---

Algorithm 2 The AdamW Algorithm

Require: $x_1 \in \mathcal{F}$ initial point, $\{\alpha_t\}_{t=1}^T$ step sizes, $\beta_1, \beta_2 \geq 0$ and $\alpha < \sqrt{1-\beta_2}$.

  $m_0, v_0 \leftarrow 0$

  for $t = 1, \ldots, T$ **do**

    $g_t \leftarrow \nabla f_t(x_t)$

    $x_t \leftarrow (1 - \alpha_t \gamma) x_{t-1}$

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

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

    $\hat{m}_t \leftarrow m_t / (1 - \beta_1^t)$

    $\hat{v}_t \leftarrow v_t / (1 - \beta_2^t)$

    $\hat{x}_{t+1} \leftarrow x_t - \alpha_t \hat{m}_t / \sqrt{\hat{v}_t}$

    $x_{t+1} \leftarrow \Pi_{\mathcal{F}}(\hat{x}_{t+1})$ ▷ Projection

  end for

  return x_T

---

of showing all terms are positive, we can try and show each is bounded below by 1/2. If we repeat the same calculation for $\hat{x}_{3t+2}$, we have

$$\hat{x}_{3t+2} \geq (1 - \alpha_{3t+1}\gamma)\, x_{3t+1} - \frac{\alpha C}{\sqrt{(3t+1)(\beta_2 v_{3t} + (1-\beta_2)C^2)}} = \left(1 - \frac{\gamma\alpha}{\sqrt{3t+1}}\right) - \frac{\alpha C}{\sqrt{(3t+1)(\beta_2 v_{3t} + (1-\beta_2 C^2))}},$$

meaning we now want

$$\frac{\alpha}{\sqrt{(3t+1)(1-\beta_2)}} < \left(1 - \frac{\alpha\gamma}{\sqrt{3t+1}}\right).$$

We can use the same assumption on $\alpha < \sqrt{1-\beta_2}$ and bring the decay term over to say

$$\frac{\alpha}{\sqrt{(3t+1)(1-\beta_2)}} + \frac{\gamma\alpha}{\sqrt{3t+1}} \leq \frac{1+\gamma\alpha}{\sqrt{3t+1}} < 1/2 < 1$$

and this time we can use the bound $\alpha < 1$ to conclude because γ is small already. Hence, $1/2 < \hat{x}_{3t+2} < 1$, but if there isn't a $t : x_{3t+1} = 1$, take $t = 0$ and note the above still holds. Therefore,

$$\hat{x}_{3t+3} = \left(1 - \frac{\gamma\alpha}{\sqrt{3t+2}}\right) x_{3t+2} + \frac{\alpha}{\sqrt{(3t+2)(\beta_2 v_{3t+1} + (1-\beta_2))}} > 1/2$$

because

$$\begin{aligned} \hat{x}_{3t+3} \geq \left(1 - \frac{\gamma\alpha}{\sqrt{3t+2}}\right) \frac{1}{2} + \frac{\alpha}{\sqrt{(3t+2)(\beta_2 v_{3t+1} + (1-\beta_2))}} > 1/2 \\ \iff \frac{\gamma\alpha}{\sqrt{3t+2}} < \frac{\alpha}{\sqrt{(3t+2)(\beta_2 v_{3t+1} + (1-\beta_2))}}, \end{aligned}$$

which happens when

$$\gamma \leq \frac{1}{\sqrt{\beta_2 C^2 + (1-\beta_2)}} \leq \frac{1}{\sqrt{\beta_2 v_{3t+2} + (1-\beta_2)}},$$

and by our choice of $\beta_2 = 1/(1+C^2)$, we have

$$\gamma \leq \frac{1}{\sqrt{2(1-\beta_2)}} = \sqrt{\frac{C^2+1}{2C^2}} \tag{6}$$

following the calculations in equation (5). All that remains is finding conditions such that $\hat{x}_{3t+4} > 1/2$ to force $x_{3t+4} = 1$. If $(1 - \alpha_{3t+3}\gamma)x_{3t+3} \geq 1/2$, we are done and otherwise,

$$\begin{aligned} \hat{x}_{3t+4} &= (1 - \alpha_{3t+3}\gamma)x_{3t+3} + \frac{\alpha}{\sqrt{(3t+3)(\beta_2 v_{3t+2} + (1-\beta_2))}} \\ &\geq \left(1 - \frac{\alpha\gamma}{\sqrt{3t+3}}\right)\frac{1}{2} + \frac{\alpha}{\sqrt{(3t+3)(\beta_2 v_{3t+2} + (1-\beta_2))}}, \end{aligned}$$

then

$$1/2 - \frac{\alpha\gamma}{\sqrt{3t+3}} + \frac{\alpha}{\sqrt{(3t+3)(\beta_2 v_{3t+2} + (1-\beta_2))}} \overset{?}{>} 1/2$$

whenever we have

$$\begin{aligned} \frac{\alpha\gamma}{\sqrt{3t+3}} &\leq \frac{\alpha}{\sqrt{(3t+3)(\beta_2 v_{3t+2} + (1-\beta_2))}} \\ \gamma &\leq \frac{1}{\sqrt{\beta_2 C^2 + (1-\beta_2)}} \leq \frac{1}{\sqrt{\beta_2 v_{3t+2} + (1-\beta_2)}}, \end{aligned}$$

which we know to be true from the same remarks about equation (5) shown in equation (6). This is relatively tame because $(C^2+1)/C^2$ is monotonically decreasing and in the limit we require $\gamma \leq 1/\sqrt{2}$ which should always be satisfied. Now we have established $x_t > 1/2$ for all t, so our regret bound becomes (even though no longer cyclic)

$$f_{3t+1}(x_{3t+1}) + f_{3t+2}(x_{3t+2}) + f_{3t+3}(x_{3t+3}) = Cx_{3t+1} - x_{3t+2} - x_{3t+3} \geq C/2 - 2$$

with the same optimal $-C+2$ making our regret over this interval

$$R \geq C/2 - 2 - (-C+2) = \frac{3C}{2} - 4 > 0 \iff 3C > 8$$

is our necessary condition. Note this is satisfied in my experiments where I let $C = 4$. By summing over intervals of this type, we have

$$R_T \sim \sum_{t=1}^{T/3} [f_{3t+1}(x_{3t+1}) + f_{3t+2}(x_{3t+2}) + f_{3t+3}(x_{3t+3}) - f_{3t+1}(-1) + f_{3t+2}(-1) + f_{3t+3}(-1)] \geq \frac{T}{3}\left(\frac{3C}{2} - 4\right)$$

showing that $R_T/T \not\to 0$ as $T \to +\infty$.