CMPT 409/981: Optimization for Machine Learning

Lecture 20

Sharan Vaswani

November 19, 2024

Non-convergence of Adam

- **Adam**: $w_{k+1} = \Pi_{\mathcal{C}}^{k}[w_k - \eta_k\, A_k^{-1} m_k]$ where $A_k = G_k^{\frac{1}{2}}$, $G_0 = 0$ and $G_k = \beta_2 G_{k-1} + (1-\beta_2)\nabla f_k(w_k) \nabla f_k(w_k)^{\mathsf{T}}$, $m_k = \beta_1 m_{k-1} + (1-\beta_1)\nabla f_k(w_k)$, for $\beta_1, \beta_2 \in (0,1)$.
- **Scalar Adam**: $v_k = \Pi_{\mathcal{C}}\left[w_k - \frac{\eta_k\, m_k}{\sqrt{\beta_2 G_{k-1} + (1-\beta_2)\|\nabla f_k(w_k)\|^2}}\right]$, $w_{k+1} = \Pi_{\mathcal{C}}[v_k]$, where $G_0 = 0$ and $m_k = \beta_1 m_{k-1} + (1-\beta_1)\nabla f_k(w_k)$.
- For $C > 2$, run scalar Adam with $\beta_1 = 0$ (no momentum), $\beta_2 = \frac{1}{1+C^2}$ and $\eta_k = \frac{\eta}{\sqrt{k}}$ such that $\eta < \sqrt{1-\beta_2}$ on the following problem:
- Consider $\mathcal{C} = [-1, 1]$ and the following sequence of linear functions.

$$f_k(w) = \begin{cases} C\, w & \text{for } k \bmod 3 = 1 \\ -w & \text{otherwise} \end{cases}$$

We will prove that Adam results in linear regret for the above example.

Non-convergence of Adam

- **Update**: $w_1 = 1$ and for $k \geq 1$,

$$v_{k+1} := w_k - \frac{\eta_k}{\sqrt{\beta_2\, G_{k-1} + (1-\beta_2)\,\|\nabla f_k(w_k)\|^2}}\, \nabla f_k(w_k) \text{ and } w_{k+1} = \Pi_{[-1,1]}[v_{k+1}]$$

- We will compare Adam to the "best" fixed decision (w^*) that minimizes the regret. To compute w^*, consider the sequence of 3 functions from iteration $3k$ to $3k+2$ for $k \geq 0$. In this case,

$$w^* := \arg\min_{[-1,1]} [f_{3k}(w) + f_{3k+1}(w) + f_{3k+2}(w)] = \arg\min_{[-1,1]} [(C-2)w] = -1 \qquad \text{(Since } C > 2\text{)}$$

Claim: For Adam's iterates, for $k \geq 0$, for all $i \leq [3k+1]$, $w_i > 0$ and $w_{3k+1} = 1$.

Proof: Let us prove the statement by induction. **Base case**: For $k = 0$, $w_{3k+1} = w_1 = 1$.

Inductive hypothesis: Assume that for $i \leq [3k+1]$, $w_i > 0$ and $w_{3k+1} = 1$. We need to prove that (a) $w_{3k+2} > 0$, (b) $w_{3k+3} > 0$ and (c) $w_{3k+4} = 1$.

In order to show this, note that $\nabla f_i(w) = C$ for i mod $3 = 1$ and $\nabla f_i(w) = -1$ otherwise.

Non-convergence of Adam

Consider the update at iteration $(3k+1)$. By the induction hypothesis, we know that $w_{3k+1} = 1$.

$$
\begin{aligned}
v_{3k+2} &= w_{3k+1} - \left[\frac{\eta_{3k+1}}{\sqrt{\beta_2\, G_{3k} + (1-\beta_2)\left\|\nabla f_{3k+1}(w_{3k+1})\right\|^2}} \nabla f_{3k+1}(w_{3k+1})\right] \\
&= 1 - \left[\frac{C\eta}{\sqrt{(3k+1)\left(\beta_2\, G_{3k} + (1-\beta_2)C^2\right)}}\right] && \text{(Using the value of } \eta_{3k+1}\text{)} \\
&\geq 1 - \left[\frac{C\eta}{\sqrt{(3k+1)(1-\beta_2)C^2}}\right] = 1 - \left[\frac{\eta}{\sqrt{(3k+1)(1-\beta_2)}}\right] && \text{(Since } G_{3k} \geq 0\text{)} \\
\implies v_{3k+2} &> 1 - \frac{1}{\sqrt{3k+1}} > 0 && \text{(Since } \eta < \sqrt{1-\beta_2} \text{ and } k \geq 0\text{)}
\end{aligned}
$$

Since $\left[\frac{C\eta}{\sqrt{(3k+1)\left(\beta_2\, G_{3k}+(1-\beta_2)C^2\right)}}\right] > 0$, $v_{3k+2} < 1$. Since $v_{3k+2} \in (0,1)$, $w_{3k+2} = v_{3k+2} < 1$ which proves (a).

Non-convergence of Adam

- For the update at iteration $(3k+2)$, since $\nabla f_{3k+2}(w) = -1$ for all w,

$$v_{3k+3} = w_{3k+2} + \left[\frac{\eta}{\sqrt{(3k+2)\,(\beta_2\, G_{3k+1} + (1-\beta_2))}}\right]$$

Since $w_{3k+2} \in (0,1)$ and $\frac{\eta}{\sqrt{(3k+2)\,(\beta_2\, G_{3k+1} + (1-\beta_2))}} > 0$, $v_{3k+3} > 0$ and hence $w_{3k+3} > 0$ which proves (b).

- In order to prove (c), consider iteration $3k+3$. Since $\nabla f_{3k+3}(w) = -1$ for all w,

$$v_{3k+4} = w_{3k+3} + \left[\frac{\eta}{\sqrt{(3k+3)\,(\beta_2\, G_{3k+2} + (1-\beta_2))}}\right]$$

From the above update, we can conclude that $v_{3k+4} > w_{3k+3}$.

To prove (c), we will show that $v_{3k+4} \geq 1$ and hence $w_{3k+4} = \Pi_{[-1,1]} v_{3k+4} = 1$. For this, we consider two cases – when $v_{3k+3} \geq 1$ or when $v_{3k+3} < 1$.

Non-convergence of Adam

Case 1: When $v_{3k+3} \geq 1 \implies w_{3k+3} = 1 \implies v_{3k+4} \geq 1 \implies w_{3k+4} = 1$.

Case 2: When $v_{3k+3} < 1 \implies w_{3k+3} = v_{3k+3} < 1$. Combining iterations $(3k+4)$ and $(3k+3)$,

$$
\begin{aligned}
v_{3k+4} &= v_{3k+3} + \left[\frac{\eta}{\sqrt{(3k+3)\,(\beta_2\, G_{3k+2} + (1-\beta_2))}}\right] \\
&= w_{3k+2} + \left[\frac{\eta}{\sqrt{(3k+2)\,(\beta_2\, G_{3k+1} + (1-\beta_2))}}\right] + \left[\frac{\eta}{\sqrt{(3k+3)\,(\beta_2\, G_{3k+2} + (1-\beta_2))}}\right] \\
&= 1 - \underbrace{\left[\frac{C\eta}{\sqrt{(3k+1)\,(\beta_2\, G_{3k} + (1-\beta_2)\,C^2)}}\right]}_{:=T_1} \qquad \text{(Since } v_{3k+2} = w_{3k+2} \text{ and } w_{3k+1} = 1\text{)} \\
&\quad + \underbrace{\left[\frac{\eta}{\sqrt{(3k+2)\,(\beta_2\, G_{3k+1} + (1-\beta_2))}}\right] + \left[\frac{\eta}{\sqrt{(3k+3)\,(\beta_2\, G_{3k+2} + (1-\beta_2))}}\right]}_{:=T_2}
\end{aligned}
$$

In order to show that $v_{3k+4} \geq 1$, it is sufficient to show that $T_1 \leq T_2$.

Non-convergence of Adam

Recall from Slide 15, $T_1 \leq \left[\frac{\eta}{\sqrt{(3k+1)\,(1-\beta_2)}}\right]$. Let us lower-bound T_2.

$$T_2 := \left[\frac{\eta}{\sqrt{(3k+2)\,(\beta_2\, G_{3k+1} + (1-\beta_2))}}\right] + \left[\frac{\eta}{\sqrt{(3k+3)\,(\beta_2\, G_{3k+2} + (1-\beta_2))}}\right]$$

$$\geq \left[\frac{\eta}{\sqrt{(3k+2)\,(\beta_2\, C^2 + (1-\beta_2))}}\right] + \left[\frac{\eta}{\sqrt{(3k+3)\,(\beta_2\, C^2 + (1-\beta_2))}}\right]$$

(Since $G_k \leq C^2$ for all k)

$$= \frac{\eta}{\sqrt{(\beta_2\, C^2 + (1-\beta_2))}} \left[\sqrt{\frac{1}{3k+2}} + \sqrt{\frac{1}{3k+3}}\right]$$

$$\geq \frac{\eta}{\sqrt{(\beta_2\, C^2 + (1-\beta_2))}} \left[\sqrt{\frac{1}{2(3k+1)}} + \sqrt{\frac{1}{2(3k+1)}}\right] = \frac{\sqrt{2}\eta}{\sqrt{(\beta_2\, C^2 + (1-\beta_2))}} \left[\frac{1}{\sqrt{3k+1}}\right]$$

$$\implies T_2 \geq \left[\frac{\eta}{\sqrt{(3k+1)\,(1-\beta_2)}}\right] \geq T_1 \qquad \left(\text{Since } \beta_2 = \frac{1}{1+C^2} \implies \frac{\beta_2 C^2 + (1-\beta_2)}{2} = 1-\beta_2\right)$$

Non-convergence of Adam

Since we have proved that $T_2 \geq T_1$, $v_{3k+4} = 1 - T_1 + T_2 \geq 1 \implies w_{3k+4} = 1$. This completes the induction proof.

Hence, for the Adam iterates, for $k \geq 0$, for all $i \leq [3k+1]$, $w_i > 0$ and $w_{3k+1} = 1$. Now that we have bounds on the Adam iterates, let us compute its regret $R_{[3k \to 3k+2]}(w^*)$ w.r.t $w^* = -1$ for iterations $3k$ to $3k+2$.

$$\begin{aligned} R_{[3k \to 3k+2]}(w^*) &= [f_{3k}(w_{3k}) - f_{3k}(-1)] + [f_{3k+1}(w_{3k+1}) - f_{3k+1}(-1)] + [f_{3k+2}(w_{3k+2}) - f_{3k+2}(-1)] \\ &= [-w_{3k} - 1] + [C\, w_{3k+1} + C] + [-w_{3k+2} - 1] > 2C - 4 > 0 \\ & \qquad \text{(Since } w_{3k} \text{ and } w_{3k+2} \text{ are in } (0,1),\ w_{3k+1} = 1 \text{ and } C > 2) \end{aligned}$$

- Hence for every three functions, Adam has a regret $> 2C - 4$ and hence $R_T(w^*) = O(T)$.
- Both OGD and AdaGrad achieve sublinear regret when run on this example.

Non-convergence of Adam

- The example takes advantage of the non-monotonicity in the Adam step-sizes – resulting in smaller updates for $k = 1 \bmod 3$ (when the gradient is positive and will push the iterates towards -1) and larger updates for the other k (when the gradient is negative and will push the iterates towards 1).
- In the example, as $C > 2$ increases, the regret increases, $\beta_2 = \frac{1}{1+C^2} \to 0$. [**?**] show that using a "large" β_2 and ensuring that $\beta_1 \leq \sqrt{\beta_2}$ (often the choice in practice) can bypass the lower-bound resulting in convergence for Adam (without modifying the update).