CMPT 409/981: Optimization for Machine Learning

Lecture 17

Sharan Vaswani

November 14, 2022

Recap - AdaGrad

$$v_{k+1} = w_k - \eta\, A_k^{-1} \nabla f_k(w_k) \quad ; \quad w_{k+1} = \Pi_{\mathcal{C}}^k[v_{k+1}] := \arg\min_{w \in \mathcal{C}} \frac{1}{2} \|w - v_{k+1}\|_{A_k}^2 \ .$$

For $G_k \in \mathbb{R}^{d \times d} := \sum_{s=1}^k [\nabla f_s(w_s) \nabla f_s(w_s)^\intercal]$,

$$A_k = \begin{cases} \sqrt{\sum_{s=1}^k \|\nabla f_s(w_s)\|^2}\, I_d & \text{(Scalar AdaGrad)} \\ \mathrm{diag}(G_k^{\frac{1}{2}}) & \text{(Diagonal AdaGrad)} \\ G_k^{\frac{1}{2}} & \text{(Full-Matrix AdaGrad)} \end{cases}$$

For convex, G-Lipschitz losses, AdaGrad has regret $R_T(u) \leq \left(\frac{D^2}{2\eta} + \eta\right) G \sqrt{d} \sqrt{T}$.

For convex, L-smooth losses, AdaGrad has regret,
$R_T(u) \leq 2dL \left(\frac{D^2}{2\eta} + \eta\right)^2 + \sqrt{2dL} \left(\frac{D^2}{2\eta} + \eta\right) \zeta \sqrt{T}$, where $\zeta^2 := \max_k [f_k(u) - f_k^*]$.

Adaptive Gradient Methods

Update for a generic method: For $k \geq 1$ with $m_0 := 0$, $\beta \geq 0$,

$$w_{k+1} = \Pi_{\mathcal{C}}^k[w_k - \eta_k A_k^{-1} m_k]; \qquad m_k = \beta m_{k-1} + (1-\beta)\nabla f_k(w_k)$$

$$\text{where, } \Pi_{\mathcal{C}}^k[v] := \arg\min_{w \in \mathcal{C}} \frac{1}{2} \|w - v\|_{A_k}^2 .$$

Instantiating the generic method:

- **SGD**: $A_k = I_d$, $\beta = 0$. Resulting update: $w_{k+1} = w_k - \eta_k \nabla f_k(w_k)$.
- **Stochastic Heavy-Ball Momentum**: $A_k = I_d$. For $\alpha_k = \eta_k (1-\beta)$ and $\gamma_k = \frac{\beta \eta_k}{\eta_{k-1}}$, Resulting update: $w_{k+1} = w_k - \alpha_k \nabla f_k(w_k) + \gamma_k(w_k - w_{k-1})$ (Prove in Assignment 4!)
- **AdaGrad**: $A_k = G_k^{\frac{1}{2}}$ where $G_0 = 0$ and $G_k = G_{k-1} + \nabla f_k(w_k) \nabla f_k(w_k)^\intercal$, $\beta = 0$, $\eta_k = \eta$. Resulting update: $w_{k+1} = w_k - \eta A_k^{-1} \nabla f_k(w_k)$.
- **Adam**: $A_k = G_k^{\frac{1}{2}}$ where $G_0 = 0$ and $G_k = \beta_2 G_{k-1} + (1-\beta_2) \nabla f_k(w_k) \nabla f_k(w_k)^\intercal$, $\beta = \beta_1$ for $\beta_1, \beta_2 \in (0,1)$. Resulting update: $w_{k+1} = w_k - \eta_k A_k^{-1} m_k$ where $m_k = \beta_1 m_{k-1} + (1-\beta_1) \nabla f_k(w_k)$.

Adam

Recall the update: $w_{k+1} = \Pi_{\mathcal{C}}^{k}[w_k - \eta_k \, A_k^{-1} m_k]$; $m_k = \beta m_{k-1} + (1-\beta)\nabla f_k(w_k)$.

For Adam, $G_k = (1-\beta_2)\sum_{i=1}^{k} \beta_2^{k-i}[\nabla f_i(w_i)\nabla f_i(w_i)^\intercal]$ and $m_k = (1-\beta_1)\sum_{i=1}^{k} \beta_1^{k-i}[\nabla f_i(w_i)]$.

Hence, the influence of the past gradients is decayed exponentially which ensures that G_k and m_k are both primarily influenced by the most recent gradient $\nabla f_k(w_k)$.

Consider scalar Adam for which $G_k = (1-\beta_2)\sum_{i=1}^{k} \beta_2^{k-i} \left\|\nabla f_i(w_i)\right\|^2$. Unlike scalar AdaGrad (for which $G_k = \sum_{i=1}^{k} \left\|\nabla f_i(w_i)\right\|^2$), for scalar Adam, G_k is not guaranteed to increase monotonically (i.e. $G_{k+1} > G_k$). Hence $\tilde{\eta_k} := \frac{\eta}{\sqrt{G_k}}$ is not guaranteed to decrease.

Hence, to ensure convergence, Adam requires $\eta_k = \tilde{\eta_k}\alpha_k$ for some decreasing sequence α_k.

However, the non-monotonic behaviour of G_k can result in non-convergence of Adam even with an explicitly decreasing sequence of η_k.

Non-convergence of Adam

We will construct an example on which Adam can result in linear regret in the online setting (and is hence not guaranteed to converge to the minimizer in the stochastic setting) [RKK19].

Consider $\mathcal{C} = [-1, 1]$ and the following sequence of linear functions. For $C \geq 2$,

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

Run Adam with $\beta_1 = 0$ (no momentum), $\beta_2 = \frac{1}{1+C^2}$ and $\eta_k = \frac{\eta}{\sqrt{k}}$ such that $\eta < \sqrt{1-\beta_2}$. These parameters were chosen to prove the Adam regret bound in the original paper [KB14].

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

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

Non-convergence of Adam

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,

$$w^* := \underset{[-1,1]}{\arg\min} \left[f_{3k}(w) + f_{3k+1}(w) + f_{3k+2}(w)\right] = \underset{[-1,1]}{\arg\min} \left[(C-2)w\right] = -1 \qquad \text{(Since } C \geq 2\text{)}$$

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 $w_{3k+1} = 1$.

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

$$= 1 - \left[\frac{C\eta}{\sqrt{(3k+1)\left(\beta_2 \, G_{3k} + (1-\beta_2) C^2\right)}} \right] \qquad \text{(Using the value of } \eta_{3k+1}\text{)}$$

$$\geq 1 - \left[\frac{C\eta}{\sqrt{(3k+1)(1-\beta_2) C^2}} \right] = 1 - \left[\frac{\eta}{\sqrt{(3k+1)(1-\beta_2)}} \right] \qquad \text{(Since } G_{3k} \geq 0\text{)}$$

$$\implies v_{3k+2} \geq 1 - \frac{1}{\sqrt{3k+1}} > 0 \qquad \text{(Since } \eta < \sqrt{1-\beta_2} \text{ and } k \geq 1\text{)}$$

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 $v_{3k+3} \leq 1 \implies w_{3k+3} = v_{3k+3} \leq 1$. Combining iterations $(3k+4)$ and $(3k+3)$,

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

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

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

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

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

$$
\begin{aligned}
T_2 &:= \left[\frac{\eta}{\sqrt{(3k+2)\,(\beta_2\, G_{3k+1} + (1-\beta_2))}}\right] + \left[\frac{\eta}{\sqrt{(3k+3)\,(\beta_2\, G_{3k+2} + (1-\beta_2))}}\right] \\
&\geq \left[\frac{\eta}{\sqrt{(3k+2)\,(\beta_2\, C^2 + (1-\beta_2))}}\right] + \left[\frac{\eta}{\sqrt{(3k+3)\,(\beta_2\, C^2 + (1-\beta_2))}}\right]
\end{aligned}
$$

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

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

$$
\implies T_2 \geq \left[\frac{\eta}{\sqrt{(3k+1)\,(1-\beta_2)}}\right] \geq T_1
$$

(Since $\beta_2 = \frac{1}{1+C^2} \implies \frac{\beta_2 C^2 + (1-\beta_2)}{2} = 1 - \beta_2$)

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 $3k+2$.

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

Hence for every three functions, Adam has a regret $> 2C$ 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 $k = 1 \mod 3$ (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).

The example can be modified [RKK19] to consider:

- Updates of the form $w_{k+1} = w_k - \frac{\eta_k}{\sqrt{G_k + \epsilon}}$ for $\epsilon > 0$.
- Constant η_k (rather than $O(1/\sqrt{k})$).
- Stochastic setting (rather than the more general online convex optimization setup).
- Decreasing, non-zero β_1 (the momentum parameter).

- To bypass such examples where Adam fails to converge, AMSGrad [RKK19] modifies the update to ensure monotonically decreasing step-sizes and prove convergence.
- In the example, as $C \geq 2$ increases, the regret increases, $\beta_2 = \frac{1}{1+C^2} \to 0$. [ZCS$^+$22] 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).

Questions?

AMSGrad – fixing the convergence of Adam

Since the non-decreasing step-size for Adam is problematic, AMSGrad [RKK19] fixes this issue by making a small modification (in red) to Adam. It has the following update – for $\beta_1, \beta_2 \in (0, 1)$,

$$G_k = \beta_2 G_{k-1} + (1 - \beta_2)\,\text{diag}\,[\nabla f_k(w_k) \nabla f_k(w_k)^\intercal] \quad ; \quad A_k = \max\{G_k^{\frac{1}{2}}, A_{k-1}\}$$

$$w_{k+1} = \Pi_{\mathcal{C}}^k[w_k - \eta_k\, A_k^{-1} m_k]; \quad ; \quad m_k = \beta_1 m_{k-1} + (1 - \beta_1) \nabla f_k(w_k)$$

$$\Pi_{\mathcal{C}}^k[v_{k+1}] := \arg\min_{w \in \mathcal{C}} \frac{1}{2} \|w - v_{k+1}\|_{A_k}^2 \,,$$

where $C = \max\{A, B\}$ for diagonal matrices A and B implies that for all $i \in [d]$, $C_{i,i} = \max\{A_{i,i}, B_{i,i}\}$.

The AMSGrad update ensures that $A_k \succeq A_{k-1}$ and hence the step-sizes η_k are non-increasing, which guarantees convergence.

- Diederik P Kingma and Jimmy Ba, *Adam: A method for stochastic optimization*, arXiv preprint arXiv:1412.6980 (2014).
- Sashank J Reddi, Satyen Kale, and Sanjiv Kumar, *On the convergence of adam and beyond*, arXiv preprint arXiv:1904.09237 (2019).
- Yushun Zhang, Congliang Chen, Naichen Shi, Ruoyu Sun, and Zhi-Quan Luo, *Adam can converge without any modification on update rules*, arXiv preprint arXiv:2208.09632 (2022).