CMPT 409/981: Optimization for Machine Learning

Lecture 19

Sharan Vaswani

November 14, 2024

Recap

- **Scalar AdaGrad**:

$$w_{k+1} = \Pi_C[w_k - \eta_k \nabla f_k(w_k)] \quad ; \quad \eta_k = \frac{\eta}{\sqrt{\sum_{s=1}^{k} \|\nabla f_s(w_s)\|^2}}$$

- We proved that if the convex set $\mathcal{C}$ has diameter D i.e. for all $x, y \in \mathcal{C}$, $\|x - y\| \leq D$, for an arbitrary sequence of losses such that each f_k is convex, differentiable and G-Lipschitz, scalar AdaGrad with $\eta_k = \frac{\eta}{\sqrt{\sum_{s=1}^{k} \|\nabla f_s(w_s)\|^2}}$ and $w_1 \in \mathcal{C}$ has the following regret for all $u \in \mathcal{C}$,

$$R_T(u) \leq \left(\frac{D^2}{2\eta} + \eta \right) G \sqrt{T}$$

- Unlike OGD, scalar AdaGrad does not require the knowledge of G.
- Scalar AdaGrad uses one step-size for each coordinate. In practice, using one step-size per coordinate results in better empirical performance.

AdaGrad

- Let us consider the more practical variants of AdaGrad.
- The corresponding update is similar to preconditioned GD with the preconditioner A_k^{-1}:

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

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

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

- For the commonly-used diagonal variant, AdaGrad results in a per-coordinate update, i.e. $\forall i \in [d]$, if $g_{k,i} := [\nabla f_k(w_k)]_i$, then,

$$v_{k+1}[i] = w_k[i] - \eta \frac{g_{k,i}}{\sqrt{\sum_{s=1}^{k} g_{s,i}^2}} \quad ; \quad w_{k+1} = \arg\min_{w \in \mathcal{C}} \left[\sum_{i=1}^{d} \sqrt{\sum_{s=1}^{k} g_{s,i}^2}\; (w[i] - v_{k+1}[i])^2 \right]$$

AdaGrad

- We will assume that A_k is invertible (a small ϵI_d can be added to ensure invertibility).

Claim: If the convex set $\mathcal{C}$ has diameter D, for an arbitrary sequence of losses such that each f_k is convex and differentiable, AdaGrad with the general update $w_{k+1} = \Pi_{\mathcal{C}}^{k}[w_k - \eta A_k^{-1} \nabla f_k(w_k)]$ and $w_1 \in \mathcal{C}$ has the following regret for $u \in \mathcal{C}$,

$$R_T(u) \leq \left(\frac{D^2}{2\eta} + \eta\right) \text{Tr}[A_T]$$

Proof: Starting from the update, $v_{k+1} = w_k - \eta A_k^{-1} \nabla f_k(w_k)$,

$$v_{k+1} - u = w_k - \eta A_k^{-1} \nabla f_k(w_k) - u \implies A_k[v_{k+1} - u] = A_k[w_k - u] - \eta \nabla f_k(w_k)$$

Multiplying the above equations,

$$\begin{aligned}
[v_{k+1} - u]^{\mathsf{T}} A_k [v_{k+1} - u] &= [w_k - u - \eta A_k^{-1} \nabla f_k(w_k)]^{\mathsf{T}} \left[A_k[w_k - u] - \eta \nabla f_k(w_k)\right] \\
\|v_{k+1} - u\|_{A_k}^2 &= \|w_k - u\|_{A_k}^2 - 2\eta \langle \nabla f_k(w_k), w_k - u \rangle + \eta^2 [A_k^{-1} \nabla f_k(w_k)]^{\mathsf{T}} [\nabla f_k(w_k)] \\
\implies \|v_{k+1} - u\|_{A_k}^2 &= \|w_k - u\|_{A_k}^2 - 2\eta \langle \nabla f_k(w_k), w_k - u \rangle + \eta^2 \|\nabla f_k(w_k)\|_{A_k^{-1}}^2
\end{aligned}$$

AdaGrad

Recall that $\|v_{k+1} - u\|_{A_k}^2 = \|w_k - u\|_{A_k}^2 - 2\eta\langle \nabla f_k(w_k), w_k - u\rangle + \eta^2 \|\nabla f_k(w_k)\|_{A_k^{-1}}^2$. Using the update $w_{k+1} = \Pi_{\mathcal{C}}^k[v_{k+1}]$, $u \in \mathcal{C}$ with the non-expansiveness of projections,

$$\begin{aligned}
&\|w_{k+1} - u\|_{A_k}^2 = \|\Pi_{\mathcal{C}}[v_{k+1}] - \Pi_{\mathcal{C}}[u]\|_{A_k}^2 \le \|v_{k+1} - u\|_{A_k}^2 \\
\implies &\|w_{k+1} - u\|_{A_k}^2 \le \|w_k - u\|_{A_k}^2 - 2\eta\langle \nabla f_k(w_k), w_k - u\rangle + \eta^2 \|\nabla f_k(w_k)\|_{A_k^{-1}}^2 \\
&\qquad\qquad \le \|w_k - u\|_{A_k}^2 - 2\eta[f_k(w_k) - f_k(u)] + \eta^2 \|\nabla f_k(w_k)\|_{A_k^{-1}}^2 \qquad \text{(Convexity)} \\
\implies &f_k(w_k) - f_k(u) \le \frac{\|w_k - u\|_{A_k}^2 - \|w_{k+1} - u\|_{A_k}^2}{2\eta} + \frac{\eta}{2} \|\nabla f_k(w_k)\|_{A_k^{-1}}^2
\end{aligned}$$

Summing from $k = 1$ to T,

$$\implies R_T(u) \le \frac{1}{2\eta} \underbrace{\sum_{k=1}^{T} \left[\|w_k - u\|_{A_k}^2 - \|w_{k+1} - u\|_{A_k}^2\right]}_{\text{Term (i)}} + \frac{\eta}{2} \sum_{k=1}^{T} \|\nabla f_k(w_k)\|_{A_k^{-1}}^2$$

Let us now bound Term (i).

$$\text{Term (i)} = \sum_{k=1}^{T} \left[\|w_k - u\|_{A_k}^2 - \|w_{k+1} - u\|_{A_k}^2 \right]$$

$$= \sum_{k=2}^{T} [(w_k - u)^\intercal [A_k - A_{k-1}](w_k - u)] + \|w_1 - u\|_{A_1}^2 - \|w_{T+1} - u\|_{A_T}^2$$

$$\leq \sum_{k=2}^{T} \|w_k - u\|^2 \; \lambda_{\max}[A_k - A_{k-1}] + \|w_1 - u\|_{A_1}^2 \leq \sum_{k=2}^{T} D^2 \, \lambda_{\max}[A_k - A_{k-1}] + \|w_1 - u\|_{A_1}^2$$

(Since $A_{k-1} \preceq A_k$, $\lambda_{\max}[A_k - A_{k-1}] \geq 0$ and $\|w_k - u\|^2 \leq D$)

$$\implies \sum_{k=1}^{T} \left[\|w_k - u\|_{A_k}^2 - \|w_{k+1} - u\|_{A_k}^2 \right] \leq D^2 \sum_{k=2}^{T} \mathrm{Tr}[A_k - A_{k-1}] + \|w_1 - u\|_{A_1}^2$$

(For any PSD matrix B, $\lambda_{\max}[B] \leq \mathrm{Tr}[B]$)

AdaGrad

Continuing the proof from the previous slide,

$$
\begin{aligned}
\text{Term (i)} &= \sum_{k=1}^{T}\left[\|w_k - u\|_{A_k}^2 - \|w_{k+1} - u\|_{A_k}^2\right] \leq D^2 \sum_{k=2}^{T} \mathrm{Tr}[A_k - A_{k-1}] + \|w_1 - u\|_{A_1}^2 \\
&= D^2 \ \mathrm{Tr}\left[\sum_{k=2}^{T}[A_k - A_{k-1}]\right] + \|w_1 - u\|_{A_1}^2 \qquad \text{(Linearity of Trace)} \\
&= D^2 \ \mathrm{Tr}\left[A_T - A_1\right] + \|w_1 - u\|_{A_1}^2 \leq D^2 \ \mathrm{Tr}\left[A_T - A_1\right] + \lambda_{\max}[A_1] \, \|w_1 - u\|^2 \\
\implies \text{Term (i)} &\leq D^2 \ \mathrm{Tr}[A_T] - D^2 \, \mathrm{Tr}[A_1] + D^2 \, \mathrm{Tr}[A_1] = D^2 \ \mathrm{Tr}[A_T]
\end{aligned}
$$

Putting everything together,

$$
R_T(u) \leq \frac{D^2 \ \mathrm{Tr}[A_T]}{2\eta} + \frac{\eta}{2} \underbrace{\sum_{k=1}^{T} \|\nabla f_k(w_k)\|_{A_k^{-1}}^2}_{\text{Term (ii)}}
$$

Let us now bound Term (ii).

AdaGrad

Claim: Term (ii) $= \sum_{k=1}^{T} \|\nabla f_k(w_k)\|^2_{A_k^{-1}} \leq 2\,\mathrm{Tr}[A_T]$

Proof: Let us prove by induction. For convenience, define $g_k := \nabla f_k(w_k)$.

Base case: For $k = 1$, LHS $= \mathrm{Tr}[g_1^\mathsf{T} A_1^{-1} g_1] = \mathrm{Tr}[A_1^{-1} g_1 g_1^\mathsf{T}] = \mathrm{Tr}[A_1^{-1} A_1 A_1] \leq 2\,\mathrm{Tr}[A_1] =$ RHS.

Here, we used the cyclic property of trace i.e. $\mathrm{Tr}[ABC] = \mathrm{Tr}[BCA]$.

Inductive Hypothesis: If the statement is true for $T-1$, we need to prove it for T.

$$\sum_{k=1}^{T-1} \|g_k\|^2_{A_k^{-1}} + \|g_T\|^2_{A_T^{-1}} \leq 2\,\mathrm{Tr}[A_{T-1}] + \|g_T\|^2_{A_T^{-1}} = 2\,\mathrm{Tr}[\left(A_T^2 - g_T g_T^\mathsf{T}\right)^{1/2}] + \mathrm{Tr}[A_T^{-1} g_T g_T^\mathsf{T}]$$

For any $X \succeq Y \succeq 0$, we have [DHS11, Lemma 8], $2\,\mathrm{Tr}[(X-Y)^{1/2}] + \mathrm{Tr}[X^{-1/2} Y] \leq 2\,\mathrm{Tr}[X^{1/2}]$.

Using this for $X = A_T^2$, $Y = g_T g_T^\mathsf{T}$, $\sum_{k=1}^{T} \|g_k\|^2_{A_k^{-1}} \leq 2\,\mathrm{Tr}[A_T]$, which completes the proof.

Putting everything together,

$$R_T(u) \leq \left(\frac{D^2}{2\eta} + \eta\right) \mathrm{Tr}[A_T]\,.$$

Diagonal AdaGrad vs OGD

- We have proved that for both the diagonal and full-matrix variants of AdaGrad, $R_T(u) \leq \left(\frac{D^2}{2\eta} + \eta\right) \text{Tr}[A_T]$.
- By doing a tighter analysis for the diagonal variant, we can prove that the corresponding regret bound is: $R_T(u) \leq \left(\frac{D_\infty^2}{2\eta} + \eta\right) \text{Tr}[A_T]$ where $D_\infty = \max_{x,y \in \mathcal{C}} \|x - y\|_\infty$. Setting $\eta = \frac{D_\infty}{\sqrt{2}}$, $R_T(u) \leq \sqrt{2} D_\infty \sum_{i=1}^d \sqrt{\sum_{k=1}^T g_{k,i}^2}$.
- Compare the above bound to the regret for OGD (with $\eta = D/\sqrt{2}G$), $R_T(u) \leq \sqrt{2}\, D \sqrt{\sum_{i=1}^d \sum_{k=1}^T g_{k,i}^2}$ where $D = \max_{x,y \in \mathcal{C}} \|x - y\|_2$.
- If $\mathcal{C}$ is the unit hypercube, then, $D = \sqrt{d}$ and $D_\infty = 1$. If the gradients are sparse (e.g. corresponding to one-hot features for logistic regression), diagonal AdaGrad will result in a better regret bound than OGD.
- For other convex sets, such as the Euclidean ball, and when the gradients are dense, the regret of OGD can be better than that of diagonal AdaGrad.

AdaGrad

Recall that $R_T(u) \leq \left(\frac{D^2}{2\eta} + \eta\right) \mathrm{Tr}[A_T]$. In the worst-case, $\mathrm{Tr}[A_T] \leq \sqrt{d}\sqrt{\sum_{k=1}^{T} \|\nabla f_k(w_k)\|^2}$.

$$\mathrm{Tr}[A_T] = \mathrm{Tr}[G_T^{\frac{1}{2}}] = \sum_{j=1}^{d} \sqrt{\lambda_j[G_T]} = d\,\frac{\sum_{j=1}^{d} \sqrt{\lambda_j[G_T]}}{d} \leq d\sqrt{\frac{\sum_{j=1}^{d} \lambda_j[G_T]}{d}}$$

(Jensen's inequality for $\sqrt{x}$)

$$= \sqrt{d}\sqrt{\sum_{j=1}^{d} \lambda_j[G_T]} = \sqrt{d}\sqrt{\mathrm{Tr}[G_T]} = \sqrt{d}\sqrt{\mathrm{Tr}\left[\sum_{k=1}^{T} \nabla f_k(w_k)\nabla f_k(w_k)^\intercal\right]}$$

$$\mathrm{Tr}[A_T] \leq \sqrt{d}\sqrt{\left[\sum_{k=1}^{T} \mathrm{Tr}\,\nabla f_k(w_k)\nabla f_k(w_k)^\intercal\right]} = \sqrt{d}\sqrt{\sum_{k=1}^{T} \|\nabla f_k(w_k)\|^2} \quad \text{(Linearity of Trace)}$$

Putting everything together, in the worst-case, the regret can be bounded as:

$$R_T(u) \leq \left(\frac{D^2}{2\eta} + \eta\right)\sqrt{d}\sqrt{\sum_{k=1}^{T} \|\nabla f_k(w_k)\|^2}$$

AdaGrad - Convex, Lipschitz functions

Claim: If the convex set $\mathcal{C}$ has diameter D, for an arbitrary sequence of losses such that each f_k is convex, differentiable and G-Lipschitz, AdaGrad with the general update $w_{k+1} = \Pi_{\mathcal{C}}^{k}[w_k - \eta A_k^{-1} \nabla f_k(w_k)]$ with $\eta = \frac{D}{\sqrt{2}}$ and $w_1 \in \mathcal{C}$ has the following regret for $u \in \mathcal{C}$,

$$R_T(u) \leq \sqrt{2} DG \sqrt{d} \sqrt{T}$$

Proof: Using the general result for AdaGrad and that each f_k is G-Lipschitz,

$$R_T(u) \leq \left(\frac{D^2}{2\eta} + \eta\right) \sqrt{d} \sqrt{\sum_{k=1}^{T} \|\nabla f_k(w_k)\|^2} \leq \left(\frac{D^2}{2\eta} + \eta\right) \sqrt{d}\, G \sqrt{T}$$

$$R_T(u) \leq \sqrt{2} DG \sqrt{d} \sqrt{T} \qquad \text{(Setting } \eta = \tfrac{D}{\sqrt{2}}\text{)}$$

- Unlike scalar AdaGrad, when using the diagonal or full-matrix variant, the worst-case regret has a dimension dependence.
- Similar to scalar AdaGrad, we can derive regret bounds for the strongly-convex Lipschitz and smooth convex losses.

Questions?

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 $G_k = G_{k-1} + \nabla f_k(w_k) \nabla f_k(w_k)^\mathsf{T}$, $\beta = 0$, $\eta_k = \eta$. 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 $G_k = \beta_2 G_{k-1} + (1-\beta_2) \nabla f_k(w_k) \nabla f_k(w_k)^\mathsf{T}$, $\beta = \beta_1$ 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, $G_k = (1-\beta_2)\sum_{i=1}^{k} \beta_2^{k-i}[\nabla f_i(w_i) \nabla f_i(w_i)^\mathsf{T}]$ 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 $\nabla f_k(w_k)$. This results in better empirical performance.

- 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 $G_k = \sum_{i=1}^{k} \left\|\nabla f_i(w_i)\right\|^2$), G_k is not guaranteed to increase monotonically (i.e. $G_{k+1} > G_k$). Hence the "effective step-size" $\tilde{\eta}_k$ equal to $\frac{\eta}{\sqrt{G_k}}$ is not guaranteed to decrease.

Hence, to ensure convergence, Adam requires $\eta_k = \tilde{\eta}_k \alpha_k$ for some decreasing sequence α_k. The original paper [KB14] claimed convergence for $\eta_k = O(1/\sqrt{k})$, $\beta_2 \in [0, 1)$ and $\beta_1 \in [0, 1)$.

- However, the non-monotonic behaviour of G_k can result in non-convergence of Adam even with an explicitly decreasing sequence of η_k, constant $\beta_2 \in (0, 1)$ and $\beta_1 = 0$ (no momentum).

Non-convergence of Adam

- For $C > 2$, run Adam with $\beta_1 = 0$ (no momentum), $\beta_2 = \frac{1}{1+C^2}$ and $\eta_k = \frac{\eta}{\sqrt{k}}$ such that $\eta < \sqrt{1-\beta_2}$ on the following problem:
- Consider $\mathcal{C} = [-1, 1]$ and the following sequence of linear functions.

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

In the next class, we will prove that Adam results in linear regret for the above example [RKK19].

- 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).

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 $\beta_1, \beta_2 \in (0, 1)$,

$$G_k = \beta_2 G_{k-1} + (1 - \beta_2)\, \text{diag}\,[\nabla f_k(w_k) \nabla f_k(w_k)^\mathsf{T}] \quad ; \quad A_k = \max\{G_k^{\frac{1}{2}}, A_{k-1}\}$$
$$w_{k+1} = \Pi_{\mathcal{C}}^{k}[w_k - \eta_k\, A_k^{-1} m_k]; \quad ; \quad m_k = \beta_1 m_{k-1} + (1 - \beta_1) \nabla f_k(w_k)$$
$$\Pi_{\mathcal{C}}^{k}[v_{k+1}] := \arg\min_{w \in \mathcal{C}} \frac{1}{2} \|w - v_{k+1}\|_{A_k}^2 \,,$$

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.

References i

- John Duchi, Elad Hazan, and Yoram Singer, *Adaptive subgradient methods for online learning and stochastic optimization.*, Journal of machine learning research **12** (2011), no. 7.
- 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).