CMPT 409/981: Optimization for Machine Learning

Lecture 17

Sharan Vaswani

November 14, 2022

Recap - AdaGrad

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

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

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

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 $\zeta^2 := \arg\max_k [f_k(u) - f_k^*]$.

Adaptive Gradient Methods

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

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

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

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 $G_k = G_{k-1} + \nabla f_k(w_k) \nabla f_k(w_k)^\intercal$, $\beta = 0$. Resulting update: $w_{k+1} = w_k - \eta_k \, A_k^{-1} \nabla f_k(w_k)$.
- **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, $G_k = (1-\beta_2) \sum_{i=1}^{k} \beta_2^{k-i} [\nabla f_i(w_i) \nabla f_i(w_i)^{\mathsf{T}}]$ 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 $\nabla f_k(w_k)$.

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,

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

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

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

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

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

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

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

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

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

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

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 $R_{[3k \rightarrow 3k+2]}(w^*)$ w.r.t $w^* = -1$ for iterations $3k$ to $3k+2$.

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

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 $k = 1 \bmod 3$ (when the gradient point is positive and will push the iterates towards the optimal point -1) and larger updates for the other k (when the gradient point is negative and will push the iterates towards the sub-optimal point 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 is not guaranteed 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 $\beta_1, \beta_2 \in (0, 1)$,

$$G_k = \beta G_{k-1} + (1-\beta_2)\,\mathrm{diag}\left[\nabla f_k(w_k) \nabla f_k(w_k)^\mathsf{T}\right] \quad ; \quad A_k = \max\{G_k^{\frac{1}{2}}, A_{k-1}\}$$

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

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

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.

Convergence of AMSGrad

For a sequence of convex, G-Lipschitz functions,

- [RKK19] prove an $O(D^2 \, Gd \sqrt{T})$ regret bound for AMSGrad. The proof requires $\eta_k = O(1/\sqrt{k})$ and $\beta_1 = O(\exp(-t))$ (decreasing step-size and momentum).
- [AMMC20] prove the same regret guarantee with a decreasing step-size, but constant β_1.

Since AMSGrad is typically used with a constant step-size and momentum term, [VLK$^+$20] analyze the convergence of this variant of AMSGrad for smooth, convex functions.

For this analysis, we will assume that the eigenvalues of A_k are bounded for all iterations, i.e. for all k, there exists constant $a_{\min}, a_{\max} > 0$ such that $a_{\min} I_d \preceq A_k \preceq a_{\max} I_d$. This condition can be algorithmically ensured for AMSGrad.

Moreover, we will consider the setting where interpolation is approximately satisfied, i.e. there exists a $\zeta < \infty$ such that $\zeta^2 := \mathbb{E}_i[f_i(w^*) - f_i^*]$ is small.

Minimizing convex, smooth functions using AMSGrad

Let us prove the convergence of AMSGrad when minimizing a finite-sum of L-smooth, convex functions. As a warm-up, let us first analyze the case where $\beta_1 = 0$.

Claim: For minimizing a finite-sum of L-smooth functions lower-bounded by f^*, T iterations of the AMSGrad update such that $a_{\min} I_d \preceq A_k \preceq a_{\max} I_d$, with $\eta = \frac{a_{\min}}{2L}$, $\beta_1 = 0$ returns an iterate $\bar{w} = \sum_{k=1}^{T} w_k / T$ such that,

$$\mathbb{E}[f(\bar{w}_T) - f(w^*)] \leq \frac{D^2 \, 2dL \, a_{\max}}{a_{\min} \, T} + \zeta^2 \quad \text{where} \quad \zeta^2 := \mathbb{E}_i[f_i(w^*) - f_i^*].$$

Proof: Define $P_k := \frac{A_k}{\eta}$. Starting from the update, $v_{k+1} = w_k - P_k^{-1} \nabla f_{ik}(w_k)$ and using the same steps as the AdaGrad proof,

$$v_{k+1} - w^* = w_k - P_k^{-1} \nabla f_{ik}(w_k) - w^* \implies P_k[v_{k+1} - w^*] = P_k[w_k - w^*] - \nabla f_{ik}(w_k)$$

$$\implies [v_{k+1} - w^*]^\intercal P_k [v_{k+1} - w^*] = [w_k - w^* - P_k^{-1} \nabla f_{ik}(w_k)]^\intercal \left[P_k[w_k - w^*] - \nabla f_{ik}(w_k)\right]$$

$$\|v_{k+1} - w^*\|_{P_k}^2 = \|w_k - w^*\|_{P_k}^2 - 2\langle \nabla f_{ik}(w_k), w_k - w^* \rangle + [P_k^{-1} \nabla f_{ik}(w_k)]^\intercal [\nabla f_{ik}(w_k)]$$

$$\implies \|v_{k+1} - w^*\|_{P_k}^2 = \|w_k - w^*\|_{P_k}^2 - 2\langle \nabla f_{ik}(w_k), w_k - w^* \rangle + \|\nabla f_{ik}(w_k)\|_{P_k^{-1}}^2$$

Minimizing convex, smooth functions using AMSGrad

Recall that $\|v_{k+1} - w^*\|^2_{P_k} = \|w_k - w^*\|^2_{P_k} - 2\langle \nabla f_{ik}(w_k), w_k - w^* \rangle + \|\nabla f_{ik}(w_k)\|^2_{P_k^{-1}}$. Using the update $w_{k+1} = \Pi^k_{\mathcal{C}}[v_{k+1}]$, $w^* \in \mathcal{C}$ with the non-expansiveness of projections,

$$\|w_{k+1} - w^*\|^2_{P_k} = \frac{\|w_{k+1} - w^*\|^2_{A_k}}{\eta} = \frac{\|\Pi_{\mathcal{C}}[v_{k+1}] - \Pi_{\mathcal{C}}[w^*]\|^2_{A_k}}{\eta} \leq \frac{\|v_{k+1} - w^*\|^2_{A_k}}{\eta} = \|v_{k+1} - w^*\|^2_{P_k}$$

$$\implies \|w_{k+1} - w^*\|^2_{P_k} \leq \|w_k - w^*\|^2_{P_k} - 2\langle \nabla f_{ik}(w_k), w_k - w^* \rangle + \|\nabla f_{ik}(w_k)\|^2_{P_k^{-1}}$$

$$f_{ik}(w_k) - f_{ik}(w^*) \leq \frac{\|w_k - w^*\|^2_{P_k} - \|w_{k+1} - w^*\|^2_{P_k}}{2} + \frac{1}{2}\|\nabla f_{ik}(w_k)\|^2_{P_k^{-1}} \qquad \text{(Convexity of } f_{ik}\text{)}$$

$$\implies \mathbb{E}[f(w_k) - f(w^*)] \leq \mathbb{E}\left[\frac{\|w_k - w^*\|^2_{P_k} - \|w_{k+1} - w^*\|^2_{P_k}}{2}\right] + \frac{1}{2}\mathbb{E}\left[\|\nabla f_{ik}(w_k)\|^2_{P_k^{-1}}\right]$$

$$\mathbb{E}\|\nabla f_{ik}(w_k)\|^2_{A_k^{-1}} \leq \tfrac{\eta}{a_{\min}}\mathbb{E}\left[\|\nabla f_{ik}(w_k)\|^2\right] \leq \tfrac{2L\eta}{a_{\min}}\mathbb{E}\left[f_{ik}(w_k) - f^*_{ik}\right] \leq \tfrac{2L\eta}{a_{\min}}\mathbb{E}\left[f(w_k) - f(w^*)\right] + \tfrac{2L\eta\,\zeta^2}{a_{\min}}$$

$$\implies \mathbb{E}[f(w_k) - f(w^*)] \leq \mathbb{E}\left[\frac{\|w_k - w^*\|^2_{P_k} - \|w_{k+1} - w^*\|^2_{P_k}}{2}\right] + \frac{L\eta}{a_{\min}}\mathbb{E}\left[f(w_k) - f(w^*)\right] + \frac{L\eta\,\zeta^2}{a_{\min}}$$

Minimizing convex, smooth functions using AMSGrad

Recall that $\mathbb{E}[f(w_k) - f(w^*)] \leq \mathbb{E}\left[\frac{\|w_k - w^*\|_{P_k}^2 - \|w_{k+1} - w^*\|_{P_k}^2}{2}\right] + \frac{L\eta}{a_{\min}} \mathbb{E}\left[f(w_k) - f(w^*)\right] + \frac{L\eta\,\zeta^2}{a_{\min}}$.

Setting $\eta = \frac{a_{\min}}{2L}$ and rearranging,

$$\mathbb{E}[f(w_k) - f(w^*)] \leq \mathbb{E}\left[\|w_k - w^*\|_{P_k}^2 - \|w_{k+1} - w^*\|_{P_k}^2\right] + \zeta^2$$

Taking expectation w.r.t the randomness in iterations $k = 1$ to T and summing,

$$\sum_{k=1}^{T} \mathbb{E}[f(w_k) - f(w^*)] \leq \sum_{k=1}^{T} \mathbb{E}\left[\|w_k - w^*\|_{P_k}^2 - \|w_{k+1} - w^*\|_{P_k}^2\right] + \zeta^2\, T$$

Dividing by T, using Jensen's inequality on the LHS and the definition of $\bar{w}_T$

$$\mathbb{E}[f(\bar{w}_T) - f(w^*)] \leq \frac{\sum_{k=1}^{T} \mathbb{E}\left[\|w_k - w^*\|_{P_k}^2 - \|w_{k+1} - w^*\|_{P_k}^2\right]}{T} + \zeta^2$$

Minimizing convex, smooth functions using AMSGrad

Recall that $\mathbb{E}[f(\bar{w}_T) - f(w^*)] \leq \frac{\sum_{k=1}^{T} \mathbb{E}\left[\|w_k - w^*\|_{P_k}^2 - \|w_{k+1} - w^*\|_{P_k}^2\right]}{T} + \zeta^2$.

$$\sum_{k=1}^{T} \left[\|w_k - w^*\|_{P_k}^2 - \|w_{k+1} - w^*\|_{P_k}^2\right]$$

$$= \sum_{k=2}^{T} [(w_k - w^*)^{\mathsf{T}}[P_k - P_{k-1}](w_k - w^*)] + \|w_1 - u\|_{P_1}^2 - \|w_{T+1} - u\|_{P_T}^2$$

$$\leq \sum_{k=2}^{T} \|w_k - w^*\|^2 \; \lambda_{\max}[P_k - P_{k-1}] + \|w_1 - w^*\|_{P_1}^2 \leq \sum_{k=2}^{T} D^2 \, \lambda_{\max}[P_k - P_{k-1}] + \|w_1 - u\|_{P_1}^2$$

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

$$\sum_{k=1}^{T} \left[\|w_k - w^*\|_{P_k}^2 - \|w_{k+1} - w^*\|_{P_k}^2\right] \leq D^2 \sum_{k=2}^{T} \mathrm{Tr}[P_k - P_{k-1}] + \|w_1 - u\|_{P_1}^2 \leq D^2 \, \mathrm{Tr}[P_T]$$

(By linearity of trace, and bounding $\|w_1 - u\|_{P_1}^2 \leq D^2 \; \mathrm{Tr}[P_1]$)

Minimizing convex, smooth functions using AMSGrad

Recall that $\mathbb{E}[f(\bar{w}_T) - f(w^*)] \leq \frac{D^2 \, \text{Tr}[P_T]}{T} + \zeta^2$.

$$D^2 \, \text{Tr}[P_T] \leq \frac{D^2}{\eta} \, \text{Tr}[A_T] = \frac{D^2 \, 2L \, \text{Tr}[A_T]}{a_{\text{min}}} \leq \frac{D^2 \, 2L \, d \, \lambda_{\text{max}}[A_T]}{a_{\text{min}}} \leq \frac{D^2 \, 2L \, d \, a_{\text{max}}}{a_{\text{min}}}$$

$$\implies \mathbb{E}[f(\bar{w}_T) - f(w^*)] \leq \frac{D^2 \, 2dL \, a_{\text{max}}}{a_{\text{min}} \, T} + \zeta^2$$

When minimizing smooth, convex functions, AMSGrad with a constant step-size without momentum will converge to a neighbourhood of the solution. Similar to SGD, this neighbourhood depends on ζ, the extent to which interpolation is violated.

Next, we will consider the $\beta_1 \neq 0$ case and prove a similar convergence result for constant step-size AMSGrad.

Questions?

References i

- Ahmet Alacaoglu, Yura Malitsky, Panayotis Mertikopoulos, and Volkan Cevher, *A new regret analysis for adam-type algorithms*, International conference on machine learning, PMLR, 2020, pp. 202–210.
- 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).
- Sharan Vaswani, Issam H Laradji, Frederik Kunstner, Si Yi Meng, Mark Schmidt, and Simon Lacoste-Julien, *Adaptive gradient methods converge faster with over-parameterization (and you can do a line-search)*.
- 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).