CMPT 409/981: Optimization for Machine Learning

Lecture 18

Sharan Vaswani

November 17, 2022

Recap

Adam: $w_{k+1} = \Pi_{\mathcal{C}}^{k}[w_k - \eta_k \, A_k^{-1} m_k]$; $m_k = \beta m_{k-1} + (1-\beta)\nabla f_k(w_k)$.
$G_k = (1-\beta_2)\sum_{i=1}^{k} \beta_2^{k-i}[\nabla f_i(w_i)\nabla f_i(w_i)^\intercal]$ and $m_k = (1-\beta_1)\sum_{i=1}^{k}\beta_1^{k-i}[\nabla f_i(w_i)]$.

Adam does not guarantee that $A_k \succeq A_{k-1}$ for all k. There are simple counter-examples that exploit this and can result in the non-convergence of Adam.

AMSGrad – fixing the convergence of Adam

AMSGrad [RKK19] fixes the non-convergence of Adam by making a small modification (in red) to Adam. It has the following update – for $\beta_1, \beta_2 \in (0,1)$,

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

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

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

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 make the following simplifying assumptions,

- **Bounded eigenvalues**: The eigenvalues of A_k are bounded for all iterations, i.e. for all k, there exists constants $a_{\min}, a_{\max} > 0$ such that $a_{\min} I_d \preceq A_k \preceq a_{\max} I_d$. This condition can be algorithmically ensured for the diagonal preconditioner.
- **Near-interpolation**: There exists a $\zeta < \infty$ such that $\zeta^2 := \mathbb{E}_i[f_i(w^*) - f_i^*]$ is small.
- **Bounded iterates**: The domain is unconstrained i.e. $\mathcal{C} = \mathbb{R}^d$ but the iterates remain bounded in a set of diameter D, i.e. for all k, $\|w_k - w^*\|^2 \leq D^2$.

Minimizing convex, smooth functions using AMSGrad

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

Claim: For minimizing a finite-sum of convex, L-smooth functions, assuming that for all $k \in [T]$, $\|w_k - w^*\|^2 \leq D^2$, $a_{\min} I_d \preceq A_k \preceq a_{\max} I_d$, T iterations of the AMSGrad update 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,

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

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

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

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

Minimizing convex, smooth functions using AMSGrad

Recall that $\|v_{k+1} - w^*\|_{P_k}^2 = \|w_k - w^*\|_{P_k}^2 - 2\langle \nabla f_{ik}(w_k), w_k - w^* \rangle + \|\nabla f_{ik}(w_k)\|_{P_k^{-1}}^2$. Since $\mathcal{C} = \mathbb{R}^d$, $w_{k+1} = v_{k+1}$,

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

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

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

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

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

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

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

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

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

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

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

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

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

Minimizing convex, smooth functions using AMSGrad

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

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

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

When minimizing smooth, convex functions, AMSGrad with a constant step-size *without momentum* will converge to a neighbourhood of the solution at an $O(1/T)$ rate. 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?

Minimizing convex, smooth functions using AMSGrad

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

$$\mathbb{E}[f(\bar{w}_T) - f(w^*)] \leq \left(\frac{1+\beta}{1-\beta}\right)^2 \frac{D^2\, 2dL\, a_{\max}}{a_{\min}\; T} + \zeta^2 \quad \text{where} \quad \zeta^2 := \mathbb{E}_i[f_i(w^*) - f_i^*].$$

Proof: Proceeding similar to the case for $\beta_1 = 0$, define $P_k := \frac{A_k}{\eta}$ and $\beta := \beta_1$. Starting from the update, $v_{k+1} = w_k - P_k^{-1} m_k$ where $m_k = \beta m_{k-1} + (1-\beta) \nabla f_{ik}(w_k)$.

$$v_{k+1} - w^* = w_k - P_k^{-1} m_k - w^* \implies P_k[v_{k+1} - w^*] = P_k[w_k - w^*] - m_k$$

$$[v_{k+1} - w^*]^\mathsf{T} P_k [v_{k+1} - w^*] = [w_k - w^* - P_k^{-1} m_k]^\mathsf{T} \, [P_k[w_k - w^*] - m_k]$$

$$\|v_{k+1} - w^*\|_{P_k}^2 = \|w_k - w^*\|_{P_k}^2 - 2\langle m_k, w_k - w^*\rangle + [P_k^{-1} m_k]^\mathsf{T}[m_k]$$

$$\|w_{k+1} - w^*\|_{P_k}^2 = \|w_k - w^*\|_{P_k}^2 - 2(1-\beta)\,\langle w_k - w^*, \nabla f_{ik}(w_k)\rangle - 2\beta\,\langle w_k - w^*, m_{k-1}\rangle + \|m_k\|_{P_k^{-1}}^2 .$$

(Since $\mathcal{C} = \mathbb{R}^d$, $w_{k+1} = v_{k+1}$)

Minimizing convex, smooth functions using AMSGrad

$$\|w_{k+1} - w^*\|_{P_k}^2 = \|w_k - w^*\|_{P_k}^2 - 2(1-\beta)\langle w_k - w^*, \nabla f_{ik}(w_k)\rangle - 2\beta\langle w_k - w^*, m_{k-1}\rangle + \|m_k\|_{P_k^{-1}}^2.$$

To simplify the $\langle w_k - w^*, m_{k-1}\rangle$ term, we will prove the following lemma: for any set of vectors a, b, c, d, if $a = b + c$, then, $-2\langle c, a - d\rangle = \|b - d\|^2 + \|a - b\|^2 - \|a - d\|^2$.

$$\|a - d\|^2 = \|b + c - d\|^2 = \|b - d\|^2 + 2\langle a - b, b - d\rangle + \|a - b\|^2 \quad (a = b + c,\ c = b - a)$$

$$\|a - d\|^2 = \|b - d\|^2 + 2\langle a - b, b - a + a - d\rangle + \|a - b\|^2 = \|b - d\|^2 + 2\langle c, a - d\rangle - \|a - b\|^2$$

$$\implies -2\langle c, a - d\rangle = \|b - d\|^2 + \|a - b\|^2 - \|a - d\|^2$$

$$-2\langle w_k - w^*, m_{k-1}\rangle = -2\langle w_k - w^*, P_{k-1}(w_k - w_{k-1})\rangle = -2\langle P_{k-1}^{1/2}(w_k - w^*), P_{k-1}^{1/2}(w_k - w_{k-1})\rangle$$

$$= -2\langle \underbrace{P_{k-1}^{1/2}(w_k - w^*)}_{=c}, \underbrace{P_{k-1}^{1/2}(w_k - w^*)}_{=a} - \underbrace{P_{k-1}^{1/2}(w_{k-1} - w^*)}_{=d}\rangle$$

Using the above lemma with $a = c = P_{k-1}^{1/2}(w_k - w^*)$, $b = 0$, $d = P_{k-1}^{1/2}(w_{k-1} - w^*)$,

$$-2\langle w_k - w^*, m_{k-1}\rangle \leq \|m_{k-1}\|_{P_{k-1}^{-1}}^2 + \|w_k - w^*\|_{P_k}^2 - \|w_{k-1} - w^*\|_{P_{k-1}}^2$$

$$\text{(Since } P_{k-1}(w_k - w_{k-1}) = m_{k-1})$$

Minimizing convex, smooth functions using AMSGrad

Putting everything together,

$$
\begin{aligned}
\|w_{k+1} - w^*\|^2_{P_k} &= \|w_k - w^*\|^2_{P_k} - 2(1-\beta)\left\langle w_k - w^*, \nabla f_{ik}(w_k)\right\rangle - 2\beta\left\langle w_k - w^*, m_{k-1}\right\rangle + \|m_k\|^2_{P_k^{-1}} \\
&= \|w_k - w^*\|^2_{P_k} - 2(1-\beta)\left\langle w_k - w^*, \nabla f_{ik}(w_k)\right\rangle \\
&\quad - 2\beta\left[\|m_{k-1}\|^2_{P_{k-1}^{-1}} + \|w_k - w^*\|^2_{P_k} - \|w_{k-1} - w^*\|^2_{P_{k-1}}\right] + \|m_k\|^2_{P_k^{-1}} \\
&= \|w_k - w^*\|^2_{P_k} - 2(1-\beta)\left[f_{ik}(w_k) - f_{ik}(w^*)\right] && \text{(By convexity)} \\
&\quad - 2\beta\left[\|m_{k-1}\|^2_{P_{k-1}^{-1}} + \|w_k - w^*\|^2_{P_k} - \|w_{k-1} - w^*\|^2_{P_{k-1}}\right] + \|m_k\|^2_{P_k^{-1}}
\end{aligned}
$$

$$
\begin{aligned}
\implies 2(1-\beta)\left[f_{ik}(w_k) - f_{ik}(w^*)\right] \\
\leq \underbrace{\left[\|w_k - w^*\|^2_{P_k} - \|w_{k+1} - w^*\|^2_{P_k}\right]}_{\text{Will telescope}} + \beta \underbrace{\left[\|w_k - w^*\|^2_{P_k} - \|w_{k-1} - w^*\|^2_{P_{k-1}}\right]}_{\text{Will telescope}} \\
+ \underbrace{\left[\beta\|m_{k-1}\|^2_{P_{k-1}^{-1}} + \|m_k\|^2_{P_k^{-1}}\right]}_{\text{Will handle next}}
\end{aligned}
$$

Minimizing convex, smooth functions using AMSGrad

Let us focus on bounding the $\beta \left\|m_{k-1}\right\|^2_{P^{-1}_{k-1}} + \left\|m_k\right\|^2_{P^{-1}_k}$ term.

$$\beta \left\|m_{k-1}\right\|^2_{P^{-1}_{k-1}} + \left\|m_k\right\|^2_{P^{-1}_k}$$

$$= \beta \left\|m_{k-1}\right\|^2_{P^{-1}_{k-1}} + (1+\delta) \left\|m_k\right\|^2_{P^{-1}_k} - \delta \left\|m_k\right\|^2_{P^{-1}_k} \qquad \text{(For some } \delta > 0\text{)}$$

$$= \beta \left\|m_{k-1}\right\|^2_{P^{-1}_{k-1}} + (1+\delta) \left\|\beta m_{k-1} + (1-\beta)\nabla f_{ik}(w_k)\right\|^2_{P^{-1}_k} - \delta \left\|m_k\right\|^2_{P^{-1}_k}$$

$$\leq \beta \left\|m_{k-1}\right\|^2_{P^{-1}_{k-1}} + (1+\delta) \left[(1+\epsilon)\,\beta^2 \left\|m_{k-1}\right\|^2_{P^{-1}_k} + (1+1/\epsilon)\,(1-\beta)^2 \left\|\nabla f_{ik}(w_k)\right\|^2_{P^{-1}_k}\right] - \delta \left\|m_k\right\|^2_{P^{-1}_k}$$

(By Young's inequality: for some $\epsilon > 0$, $(a+b)^2 = a^2 + 2ab + b^2 \leq a^2(1+\epsilon) + b^2(1+1/\epsilon)$)

$$= \left[\left(\beta + (1+\delta)(1+\epsilon)\,\beta^2\right) \left\|m_{k-1}\right\|^2_{P^{-1}_{k-1}} - \delta \left\|m_k\right\|^2_{P^{-1}_k}\right] + (1+\delta)(1+1/\epsilon)\,(1-\beta)^2 \left\|\nabla f_{ik}(w_k)\right\|^2_{P^{-1}_k}$$

In order to obtain a telescoping sum, we want $\beta + (1+\delta)(1+\epsilon)\,\beta^2 = \delta$. Hence, $\delta = \frac{\beta + \beta^2(1+\epsilon)}{1-(1+\epsilon)\beta^2}$. Since $\delta > 0 \implies \beta < \frac{1}{\sqrt{1+\epsilon}}$. With these parameter settings,

$$\beta \left\|m_{k-1}\right\|^2_{P^{-1}_{k-1}} + \left\|m_k\right\|^2_{P^{-1}_k} \leq \delta \left[\left\|m_{k-1}\right\|^2_{P^{-1}_{k-1}} - \left\|m_k\right\|^2_{P^{-1}_k}\right] + (1+\delta)(1+1/\epsilon)\,(1-\beta)^2 \left\|\nabla f_{ik}(w_k)\right\|^2_{P^{-1}_k}$$

Minimizing convex, smooth functions using AMSGrad

Putting everything together and taking expectation w.r.t randomness at iteration k,

$$
\begin{aligned}
&2(1-\beta)\,\mathbb{E}[f(w_k)-f(w^*)] \\
&\le \mathbb{E}\left[\|w_k-w^*\|^2_{P_k}-\|w_{k+1}-w^*\|^2_{P_k}\right]+\beta\,\mathbb{E}\left[\|w_k-w^*\|^2_{P_k}-\|w_{k-1}-w^*\|^2_{P_{k-1}}\right] \\
&+\delta\,\mathbb{E}\left[\|m_{k-1}\|^2_{P^{-1}_{k-1}}-\|m_k\|^2_{P^{-1}_k}\right]+(1+\delta)(1+1/\epsilon)\,(1-\beta)^2\,\mathbb{E}\,\|\nabla f_{ik}(w_k)\|^2_{P^{-1}_k}
\end{aligned}
$$

Bounding $\mathbb{E}\,\|\nabla f_{ik}(w_k)\|^2_{P^{-1}_k}$ using smoothness of f_{ik},

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

$$
\begin{aligned}
&\left[\underbrace{2(1-\beta)-(1+\delta)(1+1/\epsilon)\,(1-\beta)^2\,{}^{2L\eta}/_{a_{\min}}}_{:=\alpha}\right]\mathbb{E}[f(w_k)-f(w^*)] \\
&\le \mathbb{E}\left[\|w_k-w^*\|^2_{P_k}-\|w_{k+1}-w^*\|^2_{P_k}\right]+\beta\,\mathbb{E}\left[\|w_k-w^*\|^2_{P_k}-\|w_{k-1}-w^*\|^2_{P_{k-1}}\right] \\
&+\delta\,\mathbb{E}\left[\|m_{k-1}\|^2_{P^{-1}_{k-1}}-\|m_k\|^2_{P^{-1}_k}\right]+(1+\delta)(1+1/\epsilon)\,(1-\beta)^2\,\frac{2L\eta\,\zeta^2}{a_{\min}}
\end{aligned}
$$

Minimizing convex, smooth functions using AMSGrad

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

$$\alpha \sum_{k=1}^{T} \mathbb{E}[f(w_k) - f(w^*)]$$

$$\leq \mathbb{E} \underbrace{\sum_{k=1}^{T} \left[\|w_k - w^*\|_{P_k}^2 - \|w_{k+1} - w^*\|_{P_k}^2 \right]}_{:= T_1} + \beta\, \mathbb{E} \underbrace{\sum_{k=1}^{T} \left[\|w_k - w^*\|_{P_k}^2 - \|w_{k-1} - w^*\|_{P_{k-1}}^2 \right]}_{:= T_2}$$

$$+ \delta\, \mathbb{E} \underbrace{\sum_{k=1}^{T} \left[\|m_{k-1}\|_{P_{k-1}^{-1}}^2 - \|m_k\|_{P_k^{-1}}^2 \right]}_{:= T_3} + (1+\delta)(1 + {}^1\!/\!_\epsilon)\,(1-\beta)^2\, \frac{2L\eta\, \zeta^2\, T}{a_{\min}}$$

As before, $T_1 \leq \frac{D^2}{\eta} \operatorname{Tr}[A_T] \leq \frac{D^2\, d\, a_{\max}}{\eta}$. $T_2 = \frac{1}{\eta} \|w_T - w^*\|_{A_T}^2 \leq \frac{D^2\, d\, a_{\max}}{\eta}$. $T_3 = \frac{1}{\eta} \|m_0\|_{A_0}^2 = 0$.

$$\implies \alpha \sum_{k=1}^{T} \mathbb{E}[f(w_k) - f(w^*)] \leq \frac{D^2\, d\, a_{\max}\,(1+\beta)}{\eta} + (1+\delta)(1 + {}^1\!/\!_\epsilon)\,(1-\beta)^2\, \frac{2L\eta\, \zeta^2\, T}{a_{\min}}$$

Minimizing convex, smooth functions using AMSGrad

Recall that $\alpha \sum_{k=1}^{T} \mathbb{E}[f(w_k) - f(w^*)] \leq \frac{D^2\, d\, a_{\max}\,(1+\beta)}{\eta} + (1+\delta)(1+1/\epsilon)\,(1-\beta)^2\, \frac{2L\eta\, \zeta^2\, T}{a_{\min}}$. Here, $\delta = \frac{\beta + \beta^2(1+\epsilon)}{1-(1+\epsilon)\beta^2}$, $\beta < \frac{1}{\sqrt{1+\epsilon}}$ and $\alpha = 2(1-\beta) - (1+\delta)(1+1/\epsilon)\,(1-\beta)^2\, {}^{2L\eta}/_{a_{\min}}$. For $\epsilon > 0$, setting

$$\beta = \frac{1}{1+\epsilon} < \frac{1}{\sqrt{1+\epsilon}} \implies \delta = \frac{\beta + \beta^2\,\frac{1}{\beta}}{1 - \frac{1}{\beta}\beta^2} = \frac{2\beta}{1-\beta}$$

$$\alpha = 2(1-\beta) - \left(1 + \frac{2\beta}{1-\beta}\right)(1+1/\epsilon)\,(1-\beta)^2\, {}^{2L\eta}/_{a_{\min}} = 2(1-\beta) - (1+\beta)\, {}^{2L\eta}/_{a_{\min}}$$

For $\alpha > 0$, we want that $\eta < \frac{1-\beta}{1+\beta}\,\frac{a_{\min}}{L}$. Setting $\eta = \frac{1-\beta}{1+\beta}\,\frac{a_{\min}}{2L}$, $\alpha = 1-\beta$. With these settings,

$$\implies \sum_{k=1}^{T} \mathbb{E}[f(w_k) - f(w^*)] \leq \frac{D^2\, d\, a_{\max}\,(1+\beta)}{\alpha\eta} + \frac{(1-\beta)\,\zeta^2\, T}{\alpha}$$

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

$$\mathbb{E}[f(\bar{w}) - f(w^*)] \leq \left(\frac{1+\beta}{1-\beta}\right)^2 \frac{D^2\, 2dL\, a_{\max}}{a_{\min}} + \zeta^2$$

Minimizing convex, smooth functions using AMSGrad

When minimizing smooth, convex functions, AMSGrad with a constant step-size will converge to a neighbourhood of the solution at an $O(1/T)$ rate. Similar to SGD, this neighbourhood depends on ζ, the extent to which interpolation is violated.

Since Stochastic Heavy Ball (SHB) is a special case of AMSGrad with $A_k = I_d$, we can prove a similar $O(1/T + \zeta^2)$ rate of convergence (Prove in Assignment 4!).

Questions?

- 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.
- 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)*.