CMPT 409/981: Optimization for Machine Learning

Lecture 16

Sharan Vaswani

November 10, 2022

Recap - Scalar AdaGrad

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

For any $\eta > 0$, Scalar AdaGrad achieves the following regret for a sequence of convex losses:

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

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

Recap - AdaGrad

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

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

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 any $\eta > 0$, AdaGrad achieves the following regret for a sequence of convex losses:

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

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 regret depends on the dimension d.

AdaGrad - Convex, Smooth functions

Recall that for convex functions, the regret for AdaGrad is bounded as:

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

In order to bound the regret for smooth functions, we define ζ^2 such that $f_k(u) - f_k^* \leq \zeta^2$. Hence, if the learner is competing against a fixed decision u that minimizes each f_k, then $\zeta^2 = 0$. ζ^2 characterizes the analog of interpolation in the online setting.

Using L-smoothness of f_k to bound the gradient norm term (for each k) in the regret expression,

$$\|\nabla f_k(w_k)\|^2 \leq 2L[f_k(w_k) - f_k^*] = 2L[f_k(w_k) - f_k(u)] + 2L[f_k(u) - f_k^*] \leq 2L[f_k(w_k) - f_k(u)] + 2L\,\zeta^2$$

$$\implies \sum_{k=1}^{T} \|\nabla f_k(w_k)\|^2 \leq 2L \sum_{k=1}^{T} [f_k(w_k) - f_k(u)] + 2L \sum_{k=1}^{T} \zeta^2 = 2L\,[R_T(u) + \zeta^2\,T]$$

$$R_T(u) \leq \left(\frac{D^2}{2\eta} + \eta\right) \sqrt{d} \sqrt{2L\,[R_T(u) + \zeta^2\,T]}$$

AdaGrad - Convex, Smooth functions

Recall that $R_T(u) \leq \left(\frac{D^2}{2\eta} + \eta\right)\sqrt{d}\sqrt{2L\left[R_T(u) + \zeta^2\, T\right]}$. Squaring this expression,

$$[R_T(u)]^2 \leq \underbrace{2dL\left(\frac{D^2}{2\eta} + \eta\right)^2}_{:=\alpha}[\underbrace{R_T(u)}_{:=x} + \underbrace{\zeta^2 T}_{:=\beta}]$$

$$\implies x^2 \leq \alpha(x+\beta) \implies x \leq \frac{\alpha + \sqrt{\alpha^2 + 4\alpha\beta}}{2} \leq \alpha + \sqrt{\alpha\beta}$$

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

Note that the above bound holds for all $\eta > 0$ and AdaGrad does not need to know ζ or L. The regret depends on ζ^2, the upper-bound on $\max_{k \in [T]}[f_k(u) - f_k^*]$. Such bounds that depend on the fixed decision that we are comparing against are called *first-order regret bounds*.

For example, when $u = w^* := \arg\min_w \sum_{k=1}^{T} f_k(w)$ and $\zeta = 0$, then AdaGrad only incurs a *constant regret* that is independent of T. This observation has been used to explain the good performance of IL algorithms when using over-parameterized (convex) models [YBC20, LVS22].

Questions?

Scalar AdaGrad - Minimizing smooth, non-convex functions

We have seen that AdaGrad can results in $O(\sqrt{T})$ regret (and hence $O(1/\sqrt{T})$ convergence using the online-to-batch conversion) for a sequence of convex, smooth losses. Two problems with the practical applicability of these results:

- The regret depends on the diameter of the constrained domain $\mathcal{C}$, but for typical ML applications the optimization is unconstrained. In order to use a similar analysis for unconstrained domains, we need to make a (strong) assumption that the iterates remain bounded i.e. $\|w_k - w^*\|^2 \leq D$ for all iterations k.
- Adaptive methods like AdaGrad are heavily used in the non-convex setting (e.g. for training neural networks) but the bounds we proved heavily rely on convexity.

Similar to the standard SGD analysis, let us analyze AdaGrad Norm (the scalar variant) for minimizing a finite-sum of smooth functions on $\mathbb{R}^d$: $\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(w)$. We will use the proof in [FTC+22] and make the following standard assumptions:

- **Unbiasedness**: $\mathbb{E}_i[\nabla f_i(w)] = \nabla f(w)$
- **Bounded Variance**: $\mathbb{E}_i \|\nabla f_i(w) - \nabla f(w)\|^2 \leq \sigma^2$

Scalar AdaGrad - Minimizing smooth, non-convex functions

Scalar AdaGrad: $w_{k+1} = w_k - \eta_k g_k$ where $g_k := \nabla f_{ik}(w_k)$, and $\eta_k = \frac{\eta}{b_k}$ where $b_k^2 = b_{k-1}^2 + \|g_k\|^2 = b_0^2 + \sum_{s=1}^{k} \|g_s\|^2$.

Claim: For minimizing a finite-sum of L-smooth functions lower-bounded by f^*, T iterations of the scalar AdaGrad update (for any $\eta > 0$) returns an iterate $\hat{w}$ such that,

$$\mathbb{E}\left[\|\nabla f(\hat{w})\|\right] \leq \frac{\sqrt{2}\left(\frac{C}{\eta}\right) + \sqrt{\frac{C}{\eta}}\sqrt{b_0}}{\sqrt{T}} + \frac{\sqrt{\frac{C}{\eta}\sigma}}{\sqrt[4]{T}},$$

where, $C := 2[f(w_1) - f^*] + \left[4\eta\sigma + L\eta^2\right]\mathbb{E}\left[1 + \log\left(1 + \frac{\sum_{k=1}^{T}\|g_s\|^2}{b_0^2}\right)\right] = O(\log(T))$

SGD with $\eta_k = \frac{1}{L}\frac{1}{\sqrt{k+1}}$ has the following guarantee in the same setting (Lecture 8, Slide 6):

$$\mathbb{E}[\|\nabla f(\hat{w})\|^2] \leq \frac{2L\left[f(w_0) - f^*\right]}{\sqrt{T}} + \frac{\sigma^2\left(1 + \log(T)\right)}{\sqrt{T}}$$

Scalar AdaGrad can attain the *noise-adaptive* rate without dependence on the diameter, knowledge of L or σ for smooth, non-convex functions! Moreover, this rate holds for all η.

Scalar AdaGrad - Minimizing smooth, non-convex functions

Proof: For the analysis, we define a proxy step-size $\tilde{\eta}_k := \frac{\eta}{\sqrt{b_{k-1}^2+\sigma^2+\|\nabla f(w_k)\|^2}}$. Since $\tilde{\eta}_k$ depends on $\nabla f(w_k)$ and b_{k-1}, it does not depend on i_k. By L-smoothness of f,

$$f(w_{k+1}) \leq f(w_k) - \eta_k \langle \nabla f(w_k), g_k \rangle + \frac{L\eta_k^2}{2} \|g_k\|^2$$

$$= f(w_k) - \tilde{\eta}_k \langle \nabla f(w_k), g_k \rangle + (\tilde{\eta}_k - \eta_k)\langle \nabla f(w_k), g_k \rangle + \frac{L\eta^2}{2} \frac{\|g_k\|^2}{b_{k-1}^2 + \|g_k\|^2}$$

$$\leq f(w_k) - \tilde{\eta}_k \langle \nabla f(w_k), g_k \rangle + |\tilde{\eta}_k - \eta_k| \, \|\nabla f(w_k)\| \, \|g_k\| + \frac{L\eta^2}{2} \frac{\|g_k\|^2}{b_{k-1}^2 + \|g_k\|^2}$$

$$\mathbb{E}[f(w_{k+1})] \leq f(w_k) - \tilde{\eta}_k \|\nabla f(w_k)\|^2 + \eta \underbrace{\mathbb{E}\left[\left|\frac{\tilde{\eta}_k - \eta_k}{\eta}\right| \|\nabla f(w_k)\| \, \|g_k\|\right]}_{(*)} + \frac{L\eta^2}{2}\mathbb{E}\left[\frac{\|g_k\|^2}{b_{k-1}^2 + \|g_k\|^2}\right]$$

$$\implies \tilde{\eta}_k \|\nabla f(w_k)\|^2 \leq f(w_k) - \mathbb{E}[f(w_{k+1})] + \eta \, (*) + \frac{L\eta^2}{2}\mathbb{E}\left[\frac{\|g_k\|^2}{b_{k-1}^2 + \|g_k\|^2}\right]$$

Scalar AdaGrad - Minimizing smooth, non-convex functions

Recall that $\tilde{\eta}_k \left\|\nabla f(w_k)\right\|^2 \leq f(w_k) - \mathbb{E}[f(w_{k+1})] + \eta\,(*) + \frac{L\eta^2}{2}\mathbb{E}\left[\frac{\|g_k\|^2}{b_{k-1}^2 + \|g_k\|^2}\right]$ where $(*) = \mathbb{E}\left[\left|\frac{\tilde{\eta}_k - \eta_k}{\eta}\right| \|\nabla f(w_k)\| \; \|g_k\|\right]$. In order to bound (*), we will first bound $\left|\frac{\tilde{\eta}_k - \eta_k}{\eta}\right|$. Let $a = b_{k-1}^2 + \|g_k\|^2$ and $b = b_{k-1}^2 + \sigma^2 + \|\nabla f(w_k)\|^2$, implying that $\frac{\eta_k}{\eta} = \frac{1}{\sqrt{a}}$ and $\frac{\tilde{\eta}_k}{\eta} = \frac{1}{\sqrt{b}}$.

$$\left|\frac{\tilde{\eta}_k - \eta_k}{\eta}\right| = \left|\frac{1}{\sqrt{a}} - \frac{1}{\sqrt{b}}\right| = \left|\frac{b-a}{(\sqrt{a}+\sqrt{b})\sqrt{ab}}\right| = \left|\frac{\sigma^2 + \|\nabla f(w_k)\|^2 - \|g_k\|^2}{(\sqrt{a}+\sqrt{b})\sqrt{ab}}\right|$$

$$= \left|\frac{(\|\nabla f(w_k)\| + \|g_k\|)(\|\nabla f(w_k)\| - \|g_k\|)}{(\sqrt{a}+\sqrt{b})\sqrt{ab}} + \frac{\sigma^2}{(\sqrt{a}+\sqrt{b})\sqrt{ab}}\right|$$

Note that $\sqrt{a} + \sqrt{b} > \|\nabla f(w_k)\| + \|g_k\|$ and $\sqrt{a} + \sqrt{b} > \sigma$. Using these coarse bounds,

$$\left|\frac{\tilde{\eta}_k - \eta_k}{\eta}\right| \leq \left|\frac{\|\nabla f(w_k)\| - \|g_k\|}{\sqrt{ab}} + \frac{\sigma}{\sqrt{ab}}\right| \leq \left|\frac{\|\nabla f(w_k)\| - \|g_k\|}{\sqrt{ab}}\right| + \frac{\sigma}{\sqrt{ab}} \leq \frac{\|\nabla f(w_k) - g_k\|}{\sqrt{ab}} + \frac{\sigma}{\sqrt{ab}}$$

$$(|a+b| \leq |a| + |b| \text{ and } \big|\,\|a\| - \|b\|\,\big| \leq \|a-b\|)$$

Using the previous inequality to bound (*), we obtain that

$(*) \leq (**) = \mathbb{E}\left[\left[\frac{\|\nabla f(w_k) - g_k\| + \sigma}{\sqrt{b_{k-1}^2 + \|g_k\|^2}\sqrt{b_{k-1}^2 + \sigma^2 + \|\nabla f(w_k)\|^2}}\right] \|\nabla f(w_k)\| \; \|g_k\|\right]$. Let us simplify the first term in (**).

With $X = \|\nabla f(w_k) - g_k\|^2$, $Y = \frac{\|g_k\|^2}{b_{k-1}^2 + \|g_k\|^2}$, using Holders inequality: $\mathbb{E}[\sqrt{XY}] \leq \sqrt{\mathbb{E}[X]\,\mathbb{E}[Y]}$,

$$\text{First term in } (**) = \frac{\|\nabla f(w_k)\|}{\sqrt{b_{k-1}^2 + \sigma^2 + \|\nabla f(w_k)\|^2}} \mathbb{E}\left[\frac{\|\nabla f(w_k) - g_k\| \;\; \|g_k\|}{\sqrt{b_{k-1}^2 + \|g_k\|^2}}\right]$$

$$\leq \frac{\|\nabla f(w_k)\|}{\sqrt{b_{k-1}^2 + \sigma^2 + \|\nabla f(w_k)\|^2}} \sqrt{\mathbb{E}[\|\nabla f(w_k) - g_k\|^2]} \sqrt{\mathbb{E}\left[\frac{\|g_k\|^2}{b_{k-1}^2 + \|g_k\|^2}\right]}$$

$$\text{First term in } (**) \leq \frac{\|\nabla f(w_k)\| \; \sigma}{\sqrt{b_{k-1}^2 + \sigma^2 + \|\nabla f(w_k)\|^2}} \sqrt{\mathbb{E}\left[\frac{\|g_k\|^2}{b_{k-1}^2 + \|g_k\|^2}\right]}$$

Scalar AdaGrad - Minimizing smooth, non-convex functions

Let us simplify the second term in (**).With $X = \frac{\|g_k\|^2}{b_{k-1}^2 + \|g_k\|^2}$ and $Y = 1$, using Holder's inequality that $\mathbb{E}[\sqrt{XY}] \leq \sqrt{\mathbb{E}[X]\,\mathbb{E}[Y]}$

$$\text{Second term in (**)} = \frac{\sigma\,\|\nabla f(w_k)\|}{\sqrt{b_{k-1}^2 + \sigma^2 + \|\nabla f(w_k)\|^2}}\mathbb{E}\left[\frac{\|g_k\|}{\sqrt{b_{k-1}^2 + \|g_k\|^2}}\right]$$

$$\text{Second term in (**)} \leq \frac{\sigma\,\|\nabla f(w_k)\|}{\sqrt{b_{k-1}^2 + \sigma^2 + \|\nabla f(w_k)\|^2}}\sqrt{\mathbb{E}\left[\frac{\|g_k\|^2}{b_{k-1}^2 + \|g_k\|^2}\right]}$$

Putting everything together,

$$(*) \leq (**) \leq \frac{2\,\sigma\,\|\nabla f(w_k)\|}{\sqrt{b_{k-1}^2 + \sigma^2 + \|\nabla f(w_k)\|^2}}\sqrt{\mathbb{E}\left[\frac{\|g_k\|^2}{b_{k-1}^2 + \|g_k\|^2}\right]}$$

Scalar AdaGrad - Minimizing smooth, non-convex functions

Recall that $(*) \leq \frac{2\,\sigma\,\|\nabla f(w_k)\|}{\sqrt{b_{k-1}^2+\sigma^2+\|\nabla f(w_k)\|^2}} \sqrt{\mathbb{E}\left[\frac{\|g_k\|^2}{b_{k-1}^2+\|g_k\|^2}\right]}$.

With $a = \frac{\|\nabla f(w_k)\|}{\left[b_{k-1}^2+\sigma^2+\|\nabla f(w_k)\|^2\right]^{1/4}}$ and $b = \frac{2\,\sigma}{\left[b_{k-1}^2+\sigma^2+\|\nabla f(w_k)\|^2\right]^{1/4}} \sqrt{\mathbb{E}\left[\frac{\|g_k\|^2}{b_{k-1}^2+\|g_k\|^2}\right]}$, using that $ab \leq \frac{a^2}{2} + \frac{b^2}{2}$,

$$(*) \leq \frac{\|\nabla f(w_k)\|^2}{2\sqrt{b_{k-1}^2 + \sigma^2 + \|\nabla f(w_k)\|^2}} + \frac{2\,\sigma^2}{\sqrt{b_{k-1}^2 + \sigma^2 + \|\nabla f(w_k)\|^2}}\,\mathbb{E}\left[\frac{\|g_k\|^2}{b_{k-1}^2 + \|g_k\|^2}\right]$$

$$(*) \leq \frac{\tilde{\eta}_k}{2\eta}\,\|\nabla f(w_k)\|^2 + 2\sigma\,\mathbb{E}\left[\frac{\|g_k\|^2}{b_{k-1}^2 + \|g_k\|^2}\right]$$

(By definition of $\tilde{\eta}_k$ and since $b_{k-1}^2 + \sigma^2 + \|\nabla f(w_k)\|^2 \geq \sigma^2$)

Scalar AdaGrad - Minimizing smooth, non-convex functions

Putting back the value of (*) in

$\tilde{\eta}_k \|\nabla f(w_k)\|^2 \le f(w_k) - \mathbb{E}[f(w_{k+1})] + \eta\,(*) + \frac{L\eta^2}{2}\mathbb{E}\left[\frac{\|g_k\|^2}{b_{k-1}^2 + \|g_k\|^2}\right]$,

$$\tilde{\eta}_k \|\nabla f(w_k)\|^2 \le 2\left[f(w_k) - \mathbb{E}[f(w_{k+1})]\right] + \left[4\eta\sigma + L\eta^2\right]\mathbb{E}\left[\frac{\|g_k\|^2}{b_{k-1}^2 + \|g_k\|^2}\right]$$

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

$$\mathbb{E}\left[\sum_{k=1}^{T} \tilde{\eta}_k \|\nabla f(w_k)\|^2\right] \le 2[f(w_1) - f^*] + \left[4\eta\sigma + L\eta^2\right]\mathbb{E}\left[\sum_{k=1}^{T}\left[\frac{\|g_k\|^2}{b_{k-1}^2 + \|g_k\|^2}\right]\right]$$

$$\le \underbrace{2[f(w_1) - f^*] + \left[4\eta\sigma + L\eta^2\right]\mathbb{E}\left[\sum_{k=1}^{T}\left[\frac{\|g_k\|^2}{b_0^2 + \sum_{s=1}^{k}\|g_s\|^2}\right]\right]}_{:=C}$$

Scalar AdaGrad - Minimizing smooth, non-convex functions

Recall that $\mathbb{E}\left[\sum_{k=1}^{T} \tilde{\eta}_k \left\|\nabla f(w_k)\right\|^2\right] \leq C$ where $C = 2[f(w_1) - f^*] + \left[4\eta\sigma + L\eta^2\right] \mathbb{E}\left[\sum_{k=1}^{T} \left[\frac{\|g_k\|^2}{b_0^2 + \sum_{s=1}^{k} \|g_s\|^2}\right]\right]$. In order to bound C, we will use the following relation: for $a_s \geq 0$, need to prove in Assignment 4 that

$$\sum_{k=1}^{T} \left[\frac{a_k}{1 + \sum_{s=1}^{k} a_s}\right] \leq 1 + \log\left(1 + \sum_{k=1}^{T} a_k\right)$$

$$\implies \sum_{k=1}^{T} \left[\frac{\|g_k\|^2}{b_0^2 + \sum_{s=1}^{k} \|g_s\|^2}\right] = \sum_{k=1}^{T} \left[\frac{\|g_k\|^2 / b_0^2}{1 + \sum_{s=1}^{k} \|g_s\|^2 / b_0^2}\right] \leq 1 + \log\left(1 + \frac{\sum_{k=1}^{T} \|g_k\|^2}{b_0^2}\right)$$

$$\implies \mathbb{E}\left[\sum_{k=1}^{T} \tilde{\eta}_k \left\|\nabla f(w_k)\right\|^2\right] \leq \underbrace{2[f(w_1) - f^*] + \left[4\eta\sigma + L\eta^2\right] \mathbb{E}\left[1 + \log\left(1 + \frac{\sum_{k=1}^{T} \|g_k\|^2}{b_0^2}\right)\right]}_{O(\log(T))}$$

Scalar AdaGrad - Minimizing smooth, non-convex functions

Recall that $\mathbb{E}\left[\sum_{k=1}^{T} \tilde{\eta}_k \|\nabla f(w_k)\|^2\right] \leq C$ where $C = O(\log(T))$. Now we need to simplify the LHS. For this, define $\tilde{\eta}_{\underline{T}} := \frac{\eta}{\sqrt{b_{T-1}^2 + \sigma^2 + \sum_{k=1}^{T} \|\nabla f(w_k)\|^2}}$. Note that $\tilde{\eta}_{\underline{T}} < \tilde{\eta}_k$ for all $k \in [T]$. Hence,

$$\mathbb{E}\left[\sum_{k=1}^{T} \tilde{\eta}_k \|\nabla f(w_k)\|^2\right] \geq \mathbb{E}\left[\tilde{\eta}_{\underline{T}} \sum_{k=1}^{T} \|\nabla f(w_k)\|^2\right]$$

Note that $\mathbb{E}\left[\sum_{k=1}^{T} \|\nabla f(w_k)\|^2\right] = \mathbb{E}\left[\left[\tilde{\eta}_{\underline{T}} \sum_{k=1}^{T} \|\nabla f(w_k)\|^2\right] \frac{1}{\tilde{\eta}_{\underline{T}}}\right]$. Using Holder's inequality with $X = \tilde{\eta}_{\underline{T}} \left[\sum_{k=1}^{T} \|\nabla f(w_k)\|^2\right]$ and $Y = \frac{1}{\tilde{\eta}_{\underline{T}}}$, $(\mathbb{E}[\sqrt{XY}])^2 \leq \mathbb{E}[X]\,\mathbb{E}[Y]$

$$\left(\mathbb{E}\left[\sqrt{\sum_{k=1}^{T} \|\nabla f(w_k)\|^2}\right]\right)^2 \leq \mathbb{E}\left[\tilde{\eta}_{\underline{T}} \sum_{k=1}^{T} \|\nabla f(w_k)\|^2\right] \mathbb{E}\left[\frac{1}{\tilde{\eta}_{\underline{T}}}\right]$$

$$\implies \mathbb{E}\left[\sum_{k=1}^{T} \tilde{\eta}_k \|\nabla f(w_k)\|^2\right] \geq \mathbb{E}\left[\tilde{\eta}_{\underline{T}} \sum_{k=1}^{T} \|\nabla f(w_k)\|^2\right] \geq \frac{\eta\left(\mathbb{E}\left[\sqrt{\sum_{k=1}^{T} \|\nabla f(w_k)\|^2}\right]\right)^2}{\mathbb{E}\left[\frac{\eta}{\tilde{\eta}_{\underline{T}}}\right]}$$

Scalar AdaGrad - Minimizing smooth, non-convex functions

Recall $\mathbb{E}\left[\sum_{k=1}^{T} \tilde{\eta}_k \|\nabla f(w_k)\|^2\right] \geq \frac{\eta\left(\mathbb{E}\left[\sqrt{\sum_{k=1}^{T} \|\nabla f(w_k)\|^2}\right]\right)^2}{\mathbb{E}\left[\eta/\tilde{\eta}_T\right]}$. Now let us upper-bound $\mathbb{E}\left[\eta/\tilde{\eta}_T\right]$.

$$\mathbb{E}\left[\eta/\tilde{\eta}_T\right] = \mathbb{E}\sqrt{b_{T-1}^2 + \sigma^2 + \sum_{k=1}^{T} \|\nabla f(w_k)\|^2} = \mathbb{E}\sqrt{b_0^2 + \sigma^2 + \underbrace{\left[\sum_{k=1}^{T-1} \|g_k\|^2\right]}_{:=A} + \left[\sum_{k=1}^{T} \|\nabla f(w_k)\|^2\right]}$$

$$A = \sum_{k=1}^{T-1} \|g_k\|^2 = \sum_{k=1}^{T-1} \left[\|g_k - \nabla f(w_k)\|^2 + \|\nabla f(w_k)\|^2 + 2\langle \nabla f(w_k), g_k - \nabla f(w_k)\rangle\right]$$

$$\mathbb{E}\left[\eta/\tilde{\eta}_T\right] \leq \mathbb{E}\sqrt{b_0^2 + \sigma^2 + \sum_{k=1}^{T-1} \left[\|g_k - \nabla f(w_k)\|^2 + 2\langle \nabla f(w_k), g_k - \nabla f(w_k)\rangle\right] + 2\sum_{k=1}^{T} \|\nabla f(w_k)\|^2}$$

$$\mathbb{E}\left[\eta/\tilde{\eta}_T\right] \leq \mathbb{E}\sqrt{b_0^2 + \sigma^2 + \sum_{k=1}^{T-1} \left[\|g_k - \nabla f(w_k)\|^2 + 2\langle \nabla f(w_k), g_k - \nabla f(w_k)\rangle\right]} + \mathbb{E}\sqrt{2\sum_{k=1}^{T} \|\nabla f(w_k)\|^2}$$

$(\sqrt{a+b} \leq \sqrt{a} + \sqrt{b})$

Scalar AdaGrad - Minimizing smooth, non-convex functions

Recall $\mathbb{E}\left[\sum_{k=1}^{T} \tilde{\eta}_k \left\|\nabla f(w_k)\right\|^2\right] \geq \frac{\eta\left(\mathbb{E}\left[\sqrt{\sum_{k=1}^{T} \|\nabla f(w_k)\|^2}\right]\right)^2}{\mathbb{E}\left[\eta/\tilde{\eta}_T\right]}$. Using Jensen's inequality for $\sqrt{x}$,

$$\mathbb{E}\left[\frac{\eta}{\tilde{\eta}_T}\right]$$

$$\leq \sqrt{b_0^2 + \sigma^2 + \sum_{k=1}^{T-1} \mathbb{E}\left[\left\|g_k - \nabla f(w_k)\right\|^2 + 2\mathbb{E}\langle \nabla f(w_k), g_k - \nabla f(w_k)\rangle\right]} + \mathbb{E}\sqrt{2\sum_{k=1}^{T} \left\|\nabla f(w_k)\right\|^2}$$

$$\implies \mathbb{E}\left[\frac{\eta}{\tilde{\eta}_T}\right] \leq \sqrt{b_0^2 + T\sigma^2} + \sqrt{2}\,\mathbb{E}\sqrt{\sum_{k=1}^{T} \left\|\nabla f(w_k)\right\|^2}$$

Putting everything together,

$$\mathbb{E}\left[\sum_{k=1}^{T} \tilde{\eta}_k \left\|\nabla f(w_k)\right\|^2\right] \geq \frac{\eta\left(\mathbb{E}\left[\sqrt{\sum_{k=1}^{T} \left\|\nabla f(w_k)\right\|^2}\right]\right)^2}{\sqrt{b_0^2 + T\sigma^2} + \sqrt{2}\,\mathbb{E}\sqrt{\sum_{k=1}^{T} \left\|\nabla f(w_k)\right\|^2}}$$

Scalar AdaGrad - Minimizing smooth, non-convex functions

Recall that $C \geq \mathbb{E}\left[\sum_{k=1}^{T} \tilde{\eta}_k \left\|\nabla f(w_k)\right\|^2\right] \geq \frac{\eta\left(\mathbb{E}\left[\sqrt{\sum_{k=1}^{T}\|\nabla f(w_k)\|^2}\right]\right)^2}{\sqrt{b_0^2+T\sigma^2}+\sqrt{2}\,\mathbb{E}\sqrt{\sum_{k=1}^{T}\|\nabla f(w_k)\|^2}}$. Putting everything together,

$$\implies \left(\mathbb{E}\sqrt{\left[\sum_{k=1}^{T}\left\|\nabla f(w_k)\right\|^2\right]}\right)^2 \leq \underbrace{\frac{C}{\eta}}_{:=\alpha}\left[\underbrace{\sqrt{b_0^2+T\,\sigma^2}}_{:=\beta}+\sqrt{2}\,\underbrace{\mathbb{E}\sqrt{\left[\sum_{k=1}^{T}\left\|\nabla f(w_k)\right\|^2\right]}}_{:=x}\right]$$

$$\implies x^2 \leq \alpha\,(\sqrt{2}x+\beta) \implies x \leq \frac{\sqrt{2}\alpha+\sqrt{2\alpha^2+4\alpha\beta}}{2} \leq \sqrt{2}\alpha+\sqrt{\alpha\beta}$$

$$\implies \mathbb{E}\left[\sqrt{\sum_{k=1}^{T}\left\|\nabla f(w_k)\right\|^2}\right] \leq \sqrt{2}\left(\frac{C}{\eta}\right)+\sqrt{\frac{C}{\eta}}\sqrt[4]{b_0^2+T\,\sigma^2}$$

Scalar AdaGrad - Minimizing smooth, non-convex functions

Recall that $\mathbb{E}\left[\sqrt{\sum_{k=1}^{T} \|\nabla f(w_k)\|^2}\right] \leq \sqrt{2}\left(\frac{C}{\eta}\right) + \sqrt{\frac{C}{\eta}}\sqrt[4]{b_0^2 + T\sigma^2}$

$$\sqrt{T}\,\mathbb{E}\left[\sqrt{\frac{\sum_{k=1}^{T} \|\nabla f(w_k)\|^2}{T}}\right] \leq \sqrt{2}\left(\frac{C}{\eta}\right) + \sqrt{\frac{C}{\eta}}\sqrt[4]{b_0^2 + T\sigma^2}$$

$$\implies \mathbb{E}\left[\|\nabla f(\hat{w})\|\right] \leq \frac{\sqrt{2}\left(\frac{C}{\eta}\right)}{\sqrt{T}} + \frac{\sqrt{\frac{C}{\eta}}\sqrt[4]{b_0^2 + T\sigma^2}}{\sqrt{T}} \qquad (\hat{w} := \min_{k \in [T] \|\nabla f(w_k)\|^2})$$

$$\implies \mathbb{E}\left[\|\nabla f(\hat{w})\|\right] \leq \frac{\sqrt{2}\left(\frac{C}{\eta}\right) + \sqrt{\frac{C}{\eta}}\sqrt{b_0}}{\sqrt{T}} + \frac{\sqrt{\frac{C}{\eta}}\,\sigma}{\sqrt[4]{T}} \qquad (\sqrt{a+b} \leq \sqrt{a} + \sqrt{b})$$

Can use the above result and prove the rate in high-probability (rather than just expectation) using Markov's Theorem.

Questions?

References i

- Matthew Faw, Isidoros Tziotis, Constantine Caramanis, Aryan Mokhtari, Sanjay Shakkottai, and Rachel Ward, *The power of adaptivity in sgd: Self-tuning step sizes with unbounded gradients and affine variance*, arXiv preprint arXiv:2202.05791 (2022).
- Jonathan Wilder Lavington, Sharan Vaswani, and Mark Schmidt, *Improved policy optimization for online imitation learning*, arXiv preprint arXiv:2208.00088 (2022).
- Xinyan Yan, Byron Boots, and Ching-An Cheng, *Explaining fast improvement in online policy optimization*, arXiv preprint arXiv:2007.02520 (2020).