# CMPT 409/981: Optimization for Machine Learning

Lecture 18

Sharan Vaswani

November 12, 2024

# Adaptive step-sizes

- Recall the claim we proved earlier: If the convex set $\mathcal{C}$ has diameter $D$, for an arbitrary sequence of losses such that each $f_k$ is convex and differentiable, OGD with the update $w_{k+1} = \Pi_{\mathcal{C}}[w_k - \eta_k \nabla f_k(w_k)]$ such that $\eta_k \leq \eta_{k-1}$ and $w_1 \in \mathcal{C}$ has the following regret for $u \in \mathcal{C}$,

$$R_T(u) \leq \frac{D^2}{2\eta_T} + \sum_{k=1}^{T} \frac{\eta_k}{2} \left\|\nabla f_k(w_k)\right\|^2 = \frac{D^2}{2\eta} + \frac{\eta}{2} \sum_{k=1}^{T} \left\|\nabla f_k(w_k)\right\|^2 \quad \text{(If } \eta_k = \eta \text{ for all } k\text{)}$$

In order to find the optimal $\eta$, differentiating the RHS w.r.t $\eta$ and setting it to zero,

$$-\frac{D^2}{2\eta^2} + \frac{1}{2} \sum_{k=1}^{T} \left\|\nabla f_k(w_k)\right\|^2 = 0 \implies \eta^* = \frac{D}{\sqrt{\sum_{k=1}^{T} \left\|\nabla f_k(w_k)\right\|^2}}$$

Since the second derivative equal to $\frac{2D^2}{\eta^3} > 0$, $\eta^*$ minimizes the RHS. Setting $\eta = \eta^*$,

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

# Adaptive step-sizes

- Choosing $\eta = \eta^* = \frac{D}{\sqrt{\sum_{k=1}^{T} \|\nabla f_k(w_k)\|^2}}$ minimizes the upper-bound on the regret. However, this is not practical since setting $\eta$ requires knowing $\nabla f_k(w_k)$ for all $k \in [T]$.
- To approximate $\eta^*$ to have a practical algorithm, we can set $\eta_k$ as follows:

$$\eta_k = \frac{D}{\sqrt{\sum_{s=1}^{k} \|\nabla f_s(w_s)\|^2}}$$

Hence, at iteration $k$, we only use the gradients upto that iteration.

- Algorithmically, we only need to maintain the running sum of the squared gradient norms.
- Moreover, this choice of step-size ensures that $\eta_k \leq \eta_{k-1}$ (since we are accumulating gradient norms in the denominator so the step-size cannot increase) and hence we can use our general result for bounding the regret.

# Scalar AdaGrad

Hence, we have the following update for any $\eta > 0$,

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

This is exactly the AdaGrad update without a per-coordinate scaling and is referred to as scalar AdaGrad or AdaGrad Norm [WWB20].

- For a sequence of convex, differentiable losses, using the general result,

$$R_T(u) \leq \frac{D^2}{2\eta_T} + \sum_{k=1}^{T} \frac{\eta_k}{2} \|\nabla f_k(w_k)\|^2 = \frac{D^2}{2\eta} \sqrt{\sum_{k=1}^{T} \|\nabla f_k(w_k)\|^2} + \frac{\eta}{2} \sum_{k=1}^{T} \frac{\|\nabla f_k(w_k)\|^2}{\sqrt{\sum_{s=1}^{k} \|\nabla f_s(w_s)\|^2}}$$

In order to bound the regret for AdaGrad, we need to bound the last term.

# Scalar AdaGrad

We prove the following general claim and will use it for $a_s = \|\nabla f_s(w_s)\|^2$.

**Claim**: For all $T$ and $a_s \geq 0$, $\sum_{k=1}^{T} \frac{a_k}{\sqrt{\sum_{s=1}^{k} a_s}} \leq 2\sqrt{\sum_{k=1}^{T} a_k}$.

**Proof**: Let us prove by induction. **Base case**: For $T = 1$, LHS $= \sqrt{a_1} < 2\sqrt{a_1} =$ RHS.

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

$$\sum_{k=1}^{T} \frac{a_k}{\sqrt{\sum_{s=1}^{k} a_s}} = \sum_{k=1}^{T-1} \frac{a_k}{\sqrt{\sum_{s=1}^{k} a_s}} + \frac{a_T}{\sqrt{\sum_{s=1}^{T} a_s}} \leq 2\sqrt{\sum_{s=1}^{T-1} a_s} + \frac{a_T}{\sqrt{\sum_{s=1}^{T} a_s}} = 2\sqrt{Z-x} + \frac{x}{\sqrt{Z}}$$

$$(x := a_T,\ Z := \textstyle\sum_{s=1}^{T} a_s)$$

The derivative of the RHS w.r.t to $x$ is $-\frac{1}{\sqrt{Z-x}} + \frac{1}{\sqrt{Z}} < 0$ for all $x \geq 0$ and hence the RHS is maximized at $x = 0$. Setting $x = 0$ completes the induction proof.

$$\implies \sum_{k=1}^{T} \frac{a_k}{\sqrt{\sum_{s=1}^{k} a_s}} \leq 2\sqrt{Z} = 2\sqrt{\sum_{s=1}^{T} a_s}$$

# Scalar AdaGrad

Recall that $R_T(u) \leq \frac{D^2}{2\eta} \sqrt{\sum_{k=1}^T \|\nabla f_k(w_k)\|^2} + \frac{\eta}{2} \sum_{k=1}^T \frac{\|\nabla f_k(w_k)\|^2}{\sqrt{\sum_{s=1}^k \|\nabla f_s(w_s)\|^2}}$.

Using the claim in the previous slide with $a_s := \|\nabla f_s(w_s)\|^2 \geq 0$,

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

The step-size that minimizes the above bound is equal to $\eta^* = \frac{D}{\sqrt{2}}$. With this choice,

$$R_T(u) \leq \sqrt{2} D \sqrt{\sum_{k=1}^T \|\nabla f_k(w_k)\|^2}$$

Comparing to the regret for the optimal (impractical) constant step-size on Slide 1,

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

Hence, AdaGrad is only sub-optimal by $\sqrt{2}$ when compared to the best constant step-size!

# Scalar AdaGrad - Convex, Lipschitz functions

**Claim**: 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 $u \in \mathcal{C}$,

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

**Proof**: Using the general result from the previous slide,

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

(Since each $f_k$ is $G$-Lipschitz)

With $\eta = \frac{D}{\sqrt{2}}$, $R_T(u) \leq \sqrt{2}\, D\, G\sqrt{T}$.

- Hence, for convex, Lipschitz functions, AdaGrad achieves the same regret as OGD but is adaptive to $G$.

# Scalar AdaGrad - Convex, Smooth 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 $L$-smooth and $\zeta^2 := \max_{k \in [T]} [f_k(u) - f_k^*]$ where $f_k^* = \min_{w \in \mathcal{C}} f_k(w)$, 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 $u \in \mathcal{C}$,

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

- The regret depends on $\zeta^2$ which depends on $u$. Such bounds that depend on the fixed decision that we are comparing against are called *first-order regret bounds*.
- If the learner is competing against a fixed decision $u$ that minimizes each $f_k$, i.e. $u \in \arg\min_w f_k(w)$ for all $k$, then $\zeta^2 = 0$. Hence, $\zeta^2$ characterizes the analog of interpolation in the online setting. In this setting, 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].
- Note that the above bound holds for all $\eta > 0$ and AdaGrad does not need to know $\zeta$ or $L$.

# Scalar AdaGrad - Convex, Smooth functions

**Proof**: Using the general result for scalar AdaGrad,

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

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

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

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

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

# Scalar AdaGrad - Convex, Smooth functions

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

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

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

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

# Scalar AdaGrad - Strongly-Convex, Lipschitz functions

**Claim**: 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 $\mu$ strongly-convex, differentiable and $G$-Lipschitz, scalar AdaGrad with $\eta_k = \frac{G^2/\mu}{1+\sum_{s=1}^{k}\|\nabla f_s(w_s)\|^2}$ and $w_1 \in \mathcal{C}$ has the following regret for all $u \in \mathcal{C}$,

$$R_T(u) \leq \frac{D^2\mu}{2\,G^2} + \frac{G^2}{2\mu}\,\left[1 + \log\left(1 + G^2\,T\right)\right]$$

**Proof**: Need to prove this in Assignment 4!

- Though AdaGrad can achieve logarithmic regret for strongly-convex, Lipschitz functions similar to OGD and FTL, it requires knowledge of both $G$ and $\mu$.

**Questions?**

# References i


- Jonathan Wilder Lavington, Sharan Vaswani, and Mark Schmidt, *Improved policy optimization for online imitation learning*, arXiv preprint arXiv:2208.00088 (2022).
- Rachel Ward, Xiaoxia Wu, and Leon Bottou, *Adagrad stepsizes: Sharp convergence over nonconvex landscapes*, The Journal of Machine Learning Research **21** (2020), no. 1, 9047–9076.
- Xinyan Yan, Byron Boots, and Ching-An Cheng, *Explaining fast improvement in online policy optimization*, arXiv preprint arXiv:2007.02520 (2020).