

# 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  $u \in \mathcal{C}$ ,

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

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

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

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

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

# 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  $T = 1$ ,  $\text{LHS} = \sqrt{a_1} < 2\sqrt{a_1} = \text{RHS}$ .

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

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

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

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.

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

# 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 16,

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

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 - 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  $u \in \mathcal{C}$ ,

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

**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  $G$  and  $\mu$ .



## 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  $\zeta^2 := \max_{k \in [T]} f_k(u) - f_k^*$ , 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.  $u = w^* := \arg \min_w \sum_{k=1}^T f_k(w)$ , 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,

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

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

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

## Scalar AdaGrad - Convex, Smooth functions

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

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

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

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

Questions?

# AdaGrad

- Let us consider the more practical variant of AdaGrad that uses a per-coordinate step-size.
- The corresponding update is similar to preconditioned GD with the preconditioner  $A_k^{-1}$ :

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

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

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

- 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,

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

- We will assume that  $A_k$  is invertible (a small  $\epsilon I_d$  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  $u \in \mathcal{C}$ ,

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

**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,

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

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  $w_{k+1} = \Pi_C^k[v_{k+1}]$ ,  $u \in \mathcal{C}$  with the non-expansiveness of projections,

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

Summing from  $k = 1$  to  $T$ ,

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

Let us now bound Term (i).

$$\begin{aligned}
\text{Term (i)} &= \sum_{k=1}^T \left[ \|w_k - u\|_{A_k}^2 - \|w_{k+1} - u\|_{A_k}^2 \right] \\
&= \sum_{k=2}^T [(w_k - u)^\top [A_k - A_{k-1}] (w_k - u)] + \|w_1 - u\|_{A_1}^2 - \|w_{T+1} - u\|_{A_T}^2 \\
&\leq \sum_{k=2}^T \|w_k - u\|^2 \lambda_{\max}[A_k - A_{k-1}] + \|w_1 - u\|_{A_1}^2 \leq \sum_{k=2}^T D^2 \lambda_{\max}[A_k - A_{k-1}] + \|w_1 - u\|_{A_1}^2 \\
&\quad \text{(Since } A_{k-1} \preceq A_k, \lambda_{\max}[A_k - A_{k-1}] \geq 0 \text{ and } \|w_k - u\|^2 \leq D) \\
&\implies \sum_{k=1}^T \left[ \|w_k - u\|_{A_k}^2 - \|w_{k+1} - u\|_{A_k}^2 \right] \leq D^2 \sum_{k=2}^T \text{Tr}[A_k - A_{k-1}] + \|w_1 - u\|_{A_1}^2 \\
&\quad \text{(For any PSD matrix } B, \lambda_{\max}[B] \leq \text{Tr}[B])
\end{aligned}$$



Continuing the proof from the previous slide,

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

Putting everything together,

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

Let us now bound Term (ii).

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

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

**Base case:** For  $k = 1$ ,  $\text{LHS} = \text{Tr}[\nabla_1^T A_1^{-1} \nabla_1] = \text{Tr}[A_1^{-1} \nabla_1 \nabla_1^T] = \text{Tr}[A_1^{-1} A_1 A_1] \leq 2 \text{Tr}[A_1] = \text{RHS}$ . Here, we used the cyclic property of trace i.e.  $\text{Tr}[ABC] = \text{Tr}[BCA]$ .

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

$$\sum_{k=1}^{T-1} \|\nabla_k\|_{A_k^{-1}}^2 + \|\nabla_T\|_{A_T^{-1}}^2 \leq 2 \text{Tr}[A_{T-1}] + \|\nabla_T\|_{A_T^{-1}}^2 = 2 \text{Tr}[(A_T^2 - \nabla_T \nabla_T^T)^{1/2}] + \text{Tr}[A_T^{-1} \nabla_T \nabla_T^T]$$

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

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

Putting everything together,

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

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

$$\begin{aligned}\text{Tr}[A_T] &= \text{Tr}[G_T^{\frac{1}{2}}] = \sum_{j=1}^d \sqrt{\lambda_j[G_T]} = d \frac{\sum_{j=1}^d \sqrt{\lambda_j[G_T]}}{d} \leq d \sqrt{\frac{\sum_{j=1}^d \lambda_j[G_T]}{d}} \\ &\hspace{20em} (\text{Jensen's inequality for } \sqrt{x}) \\ &= \sqrt{d} \sqrt{\sum_{j=1}^d \lambda_j[G_T]} = \sqrt{d} \sqrt{\text{Tr}[G_T]} = \sqrt{d} \sqrt{\text{Tr} \left[ \sum_{k=1}^T \nabla f_k(w_k) \nabla f_k(w_k)^\top \right]} \\ \text{Tr}[A_T] &\leq \sqrt{d} \sqrt{\left[ \sum_{k=1}^T \text{Tr} \nabla f_k(w_k) \nabla f_k(w_k)^\top \right]} = \sqrt{d} \sqrt{\sum_{k=1}^T \|\nabla f_k(w_k)\|^2} \quad (\text{Linearity of Trace})\end{aligned}$$

Putting everything together,

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

## 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,

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

$$R_T(u) \leq \sqrt{2}DG \sqrt{d} \sqrt{T} \quad \left( \text{Setting } \eta = \frac{D}{\sqrt{2}} \right)$$

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

Questions?

-  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.
-  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).