CMPT 409/981: Optimization for Machine Learning

Lecture 15

Sharan Vaswani

November 7, 2022

Recap

Online Optimization

```
Online Optimization (w_0, Algorithm A, Convex set C)
for k = 1, ..., T do
    Algorithm A chooses point (decision) w_k ∈ C
    Environment chooses and reveals the (potentially adversarial) loss function f_k : C → ℝ
    Algorithm suffers a cost f_k(w_k)
end for
```

Regret: For any fixed decision $u \in \mathcal{C}$, $R_T(u) := \sum_{k=1}^{T} [f_k(w_k) - f_k(u)]$.

Recap

Online Gradient Descent (OGD): At iteration k, OGD chooses w_k. After the loss function f_k is revealed, OGD uses the function to compute

$$w_{k+1} = \Pi_C[w_k - \eta_k \nabla f_k(w_k)] \text{ where } \Pi_C[x] = \arg\min_{y \in \mathcal{C}} \frac{1}{2} \|y - x\|^2 .$$

If the convex set $\mathcal{C}$ has a diameter D i.e. for all $x, y \in \mathcal{C}$, $\|x - y\| \leq D$, for an arbitrary sequence losses such that each f_k is convex, differentiable and G-Lipschitz, OGD with $\eta_k = \frac{D}{\sqrt{2}\, G \sqrt{k}}$ and $w_1 \in \mathcal{C}$, has regret $R_T(u) \leq \sqrt{2} DG \sqrt{T}$.

Additionally, if each f_k is μ_k strongly-convex, OGD with $\eta_k = \frac{1}{\sum_{i=1}^{k} \mu_i}$ has regret $R_T(u) \leq \frac{G^2}{2\mu} (1 + \log(T))$.

Recap

Follow the Leader (FTL): At iteration k, FTL chooses the point w_k. After the loss function f_k is revealed, FTL uses it to compute

$$w_{k+1} = \arg\min_{w \in \mathcal{C}} \sum_{i=1}^{k} f_i(w)\,.$$

Running FTL on a quadratic lower-bound for the loss recovers OGD in the strongly-convex case.

For strongly-convex, G-Lipschitz losses, FTL has regret $R_T(u) \leq \frac{G^2}{2\mu}\,(1 + \log(T))$ that matches OGD, but does not require knowledge of μ (Proof today).

If the losses are not necessarily strongly-convex, then FTL can result in $O(T)$ regret.

Recap

Idea: Add an explicit regularization to fix FTL for a convex sequence of losses.

Follow the Regularized Leader (FTRL): At iteration $k \geq 0$, FTRL chooses the point w_k. After the loss function f_k is revealed, FTRL uses it to compute

$$w_{k+1} = \arg\min_{w \in \mathcal{C}} \sum_{i=1}^{k} \left[f_i(w) + \frac{\sigma_i}{2} \|w - w_i\|^2 \right] + \frac{\sigma_0}{2} \|w\|^2 ,$$

where $\sigma_i \geq 0$ is the regularization strength. If we set $\sigma_i = 0$ for all i, FTRL reduces to FTL.

Running FTRL on a linear lower-bound for the loss recovers OGD in the convex case.

FTRL has the following regret for a general sequence of convex losses,

$$R_T(u) \leq \sum_{k=1}^{T} \left[\frac{1}{2\lambda_{k+1}} \|\nabla f_k(w_k)\|^2 \right] + \sum_{k=1}^{T} \frac{\sigma_k}{2} \|u - w_k\|^2 + \frac{\sigma_0}{2} \|u\|^2 \ \text{ where } \lambda_k = \sum_{i=1}^{k-1} [\mu_i] + \sum_{i=0}^{k} [\sigma_i] \,.$$

For convex, G-Lipschitz losses, FTRL has regret $R_T(u) \leq \sqrt{2}\sqrt{D^2 + \|u\|^2}\, G \sqrt{T}$.

Follow the Leader - Strongly-Convex, Lipschitz functions

Claim: If the convex set $\mathcal{C}$ has diameter D, for an arbitrary sequence losses such that each f_k is μ_k strongly-convex (s.t. $\mu := \min_{k=1}^{T} \mu_k > 0$), G-Lipschitz and differentiable, then FTL with $w_1 \in \mathcal{C}$ satisfies the following regret bound for all $u \in \mathcal{C}$,

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

Proof: Using the general result for FTRL, for $\lambda_{k+1} = \sum_{i=1}^{k} \mu_i + \sum_{i=0}^{k} \sigma_i$. Since f_k is μ_k strongly-convex, we will set $\sigma_i = 0$ for all i. Hence, $\lambda_{k+1} = \sum_{i=1}^{k} \mu_i \geq \mu\, k$.

$$R_T(u) \leq \sum_{k=1}^{T} \left[\frac{1}{2\lambda_{k+1}} \left\|\nabla f_k(w_k)\right\|^2\right] + \sum_{i=1}^{T} \frac{\sigma_i}{2} \left\|u - w_i\right\|^2 + \frac{\sigma_0}{2} \left\|u\right\|^2 \leq \frac{G^2}{2\mu} \sum_{k=1}^{T} \left[\frac{1}{k}\right]$$

(Since f_k is G-Lipschitz)

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

Hence, FTL matches the regret for OGD for strongly-convex, Lipschitz functions, but does not require knowledge of μ.

Questions?

Adaptive step-sizes

Recall the claim we proved in Lecture 14 (Slide 6): 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 η, differentiating the RHS w.r.t η 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 $\frac{2D^2}{\eta^3} > 0$, η^* 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 η requires knowing $\nabla f_k(w_k)$ for all $k \in [T]$.

To approximate η^* to have a practical algorithm, we can set η_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 3,

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

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 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 losses such that each f_k is μ 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 $u \in \mathcal{C}$,

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

Though AdaGrad can achieve logarithmic regret for strongly-convex, Lipschitz functions similar to OGD and FTL, it requires knowledge of G and μ and is not adaptive to these quantities.
Proof: Need to prove this in Assignment 4!

Questions?

AdaGrad

Let us consider a more general and practical variant of AdaGrad that uses a per-coordinate step-size. The corresponding update is:

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

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

where $G_k \in \mathbb{R}^{d \times d} := \sum_{s=1}^{k} [\nabla f_s(w_s) \nabla f_s(w_s)^\mathsf{T}]$. For the subsequent analysis, we will assume that A_k is invertible (a small ϵI_d can be added to ensure invertibility)

AdaGrad

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,

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

AdaGrad

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

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

Summing from $k = 1$ to T,

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

Let us now bound the first term in the above expression.

AdaGrad

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

AdaGrad

Continuing the proof from the previous slide,

$$\sum_{k=1}^{T}\left[\|w_k - u\|_{A_k}^2 - \|w_{k+1} - u\|_{A_k}^2\right] \leq D^2 \sum_{k=2}^{T} \mathrm{Tr}[A_k - A_{k-1}] + \|w_1 - u\|_{A_1}^2$$

$$= D^2 \ \mathrm{Tr}\left[\sum_{k=2}^{T}[A_k - A_{k-1}]\right] + \|w_1 - u\|_{A_1}^2 \qquad \text{(Linearity of Trace)}$$

$$= D^2 \ \mathrm{Tr}\left[A_T - A_1\right] + \|w_1 - u\|_{A_1}^2 \leq D^2 \ \mathrm{Tr}\left[A_T - A_1\right] + \lambda_{\max}[A_1]\,\|w_1 - u\|^2$$

$$\sum_{k=1}^{T}\left[\|w_k - u\|_{A_k}^2 - \|w_{k+1} - u\|_{A_k}^2\right] \leq D^2 \ \mathrm{Tr}[A_T] - D^2\,\mathrm{Tr}[A_1] + D^2\,\mathrm{Tr}[A_1] = D^2 \ \mathrm{Tr}[A_T]$$

Putting everything together,

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

Let us now bound the second term in the above expression.

Claim: $\sum_{k=1}^{T} \|\nabla f_k(w_k)\|_{A_k^{-1}}^2 \leq 2\,\mathrm{Tr}[A_T]$
Proof: Let us prove by induction. For convenience, define $\nabla_k := \nabla f_k(w_k)$.
Base case: For $k = 1$, LHS $= \mathrm{Tr}[\nabla_1^\intercal A_1^{-1} \nabla_1] = \mathrm{Tr}[A_1^{-1} \nabla_1 \nabla_1^\intercal] = \mathrm{Tr}[A_1^{-1} A_1 A_1] \leq 2\,\mathrm{Tr}[A_1] =$ RHS. Here, we used the cyclic property of trace i.e. $\mathrm{Tr}[ABC] = \mathrm{Tr}[BCA]$.

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

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

For any $X \succeq Y \succeq 0$, we have [DHS11, Lemma 8], $2\,\mathrm{Tr}[(X - Y)^{1/2}] + \mathrm{Tr}[X^{-1/2} Y] \leq 2\,\mathrm{Tr}[X^{1/2}]$.
Using this for $X = A_T^2$, $Y = \nabla_T \nabla_T^\intercal$, $\sum_{k=1}^{T} \|\nabla_k\|_{A_k^{-1}}^2 \leq 2\,\mathrm{Tr}[A_T]$, which completes the proof.

Putting everything together,

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

AdaGrad

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

$$\mathrm{Tr}[A_T] = \mathrm{Tr}[G_T{}^{\frac{1}{2}}] = \sum_{j=1}^{d} \sqrt{\lambda_j[G_T]} = d\,\frac{\sum_{j=1}^{d} \sqrt{\lambda_j[G_T]}}{d} \leq d\sqrt{\frac{\sum_{j=1}^{d} \lambda_j[G_T]}{d}}$$

(Jensen's inequality for $\sqrt{x}$)

$$= \sqrt{d}\sqrt{\sum_{j=1}^{d} \lambda_j[G_T]} = \sqrt{d}\sqrt{\mathrm{Tr}[G_T]} = \sqrt{d}\sqrt{\mathrm{Tr}\left[\sum_{k=1}^{T} \nabla f_k(w_k)\nabla f_k(w_k)^\intercal\right]}$$

$$\mathrm{Tr}[A_T] \leq \sqrt{d}\sqrt{\mathrm{Tr}\left[\sum_{k=1}^{T} \nabla f_k(w_k)\nabla f_k(w_k)^\intercal\right]} = \sqrt{d}\sqrt{\sum_{k=1}^{T} \|\nabla f_k(w_k)\|^2} \quad \text{(Linearity of Trace)}$$

Putting everything together,

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

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.

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.