CMPT 409/981: Optimization for Machine Learning

Lecture 17

Sharan Vaswani

November 7, 2024

Recap

Generic Online Optimization (w_0, Algorithm $\mathcal{A}$, Convex set $\mathcal{C} \subseteq \mathbb{R}^d$)

for $k = 1, \ldots, T$ **do**
    Algorithm $\mathcal{A}$ chooses point (decision) $w_k \in \mathcal{C}$
    Environment chooses and reveals the (potentially adversarial) loss function $f_k : \mathcal{C} \to \mathbb{R}$
    Algorithm suffers a cost $f_k(w_k)$
end for

Examples: In imitation learning, $f_k(\pi) = \mathbb{E}_{s \sim d^{\pi_k}}[\mathrm{KL}(\pi(\cdot|s) \,||\, \pi_{\text{expert}}(\cdot|s)]$ where d^{π_k} is a distribution over the states induced by running policy π_k. In online control such as LQR (linear quadratic regulator) with unknown costs/perturbations, f_k is quadratic.

- **Regret**: For any fixed decision $u \in \mathcal{C}$, $R_T(u) := \sum_{k=1}^{T}[f_k(w_k) - f_k(u)]$.
- **Online Gradient Descent** (OGD): $w_{k+1} = \Pi_C[w_k - \eta_k \nabla f_k(w_k)]$.
- **Claim**: 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 of losses such that each f_k is convex, differentiable and G-Lipschitz, OGD with $\eta_k = \frac{\eta}{\sqrt{k}}$ and $w_1 \in \mathcal{C}$ has the following regret for all $u \in \mathcal{C}$, $R_T(u) \leq \frac{D^2 \sqrt{T}}{2\eta} + G^2 \sqrt{T}\, \eta$.

Online Gradient Descent - Strongly-convex, Lipschitz functions

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

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

Proof: Similar to the convex proof, use the update $w_{k+1} = \Pi_{\mathcal{C}}[w_k - \eta_k \nabla f_k(w_k)]$. Since $u \in \mathcal{C}$,

$$\begin{aligned}
\|w_{k+1} - u\|^2 &= \|\Pi_{\mathcal{C}}[w_k - \eta_k \nabla f_k(w_k)] - u\|^2 = \|\Pi_{\mathcal{C}}[w_k - \eta_k \nabla f_k(w_k)] - \Pi_{\mathcal{C}}[u]\|^2 \\
&\leq \|w_k - u\|^2 - 2\eta_k \langle \nabla f_k(w_k), w_k - u \rangle + \eta_k^2 \|\nabla f_k(w_k)\|^2 \\
&\leq \|w_k - u\|^2 (1 - \mu_k \eta_k) - 2\eta_k [f_k(w_k) - f_k(u)] + \eta_k^2 \|\nabla f_k(w_k)\|^2
\end{aligned}$$

(Since f_k is μ_k strongly-convex)

$$\implies R_T(u) \leq \sum_{k=1}^{T} \left[\frac{\|w_k - u\|^2 (1 - \mu_k \eta_k) - \|w_{k+1} - u\|^2}{2\eta_k} \right] + \frac{G^2}{2} \sum_{k=1}^{T} \eta_k$$

(Since f_k is G-Lipschitz)

Online Gradient Descent - Strongly-convex, Lipschitz functions

Recall that $R_T(u) \leq \sum_{k=1}^{T} \left[\frac{\|w_k - u\|^2 (1 - \mu_k \eta_k) - \|w_{k+1} - u\|^2}{2\eta_k} \right] + \frac{G^2}{2} \sum_{k=1}^{T} \eta_k$.

$$\sum_{k=1}^{T} \left[\frac{\left\|w_k - u\right\|^2 (1 - \mu_k \eta_k) - \left\|w_{k+1} - u\right\|^2}{2\eta_k} \right]$$

$$= \sum_{k=2}^{T} \left[\left\|w_k - u\right\|^2 \underbrace{\left(\frac{1}{2\eta_k} - \frac{1}{2\eta_{k-1}} - \frac{\mu_k}{2} \right)}_{=0} \right] + \left\|w_1 - u\right\|^2 \underbrace{\left[\frac{1}{2\eta_1} - \frac{\mu_1}{2} \right]}_{=0} - \frac{\left\|w_{T+1} - u\right\|^2}{2\eta_T} \leq 0$$

(Since $\eta_k = \frac{1}{\sum_{i=1}^{k} \mu_i}$)

Putting everything together,

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

(Since $\mu := \min_{k \in [T]} \mu_k$ and $\sum_{k=1}^{T} 1/k \leq 1 + \log(T)$)

Lower Bound: There is an $\Omega(\log(T))$ lower-bound on the regret for strongly-convex, Lipschitz functions and hence OGD is optimal (in terms of T) for this setting!

Questions?

Follow the Leader

Common algorithm that achieves logarithmic regret for strongly-convex losses.

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

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

- × Needs to solve a deterministic optimization sub-problem which can be expensive.
- × Needs to store all the previous loss functions and requires $O(T)$ memory.
- ✓ Does not require any step-size and is hyper-parameter free.
- In applications such Imitation Learning (IL), interacting with the environment and getting access to f_k is expensive. FTL allows multiple policy updates (when solving the sub-problem) and helps better reuse the collected data. FTL is a standard method to solve online IL problems and the resulting algorithm is known as DAGGER [RGB11].
- Compared to FTL, OGD requires an environment interaction for each policy update.

Follow the Leader and OGD

To connect FTL and OGD, consider the case when $\mathcal{C} = \mathbb{R}^d$.

$$w_{k+1} = \arg\min_{w \in \mathbb{R}} \sum_{i=1}^{k} [f_i(w)] \implies \sum_{i=1}^{k} \nabla f_i(w_{k+1}) = 0$$

- If we define $\tilde{f}_i(w)$ to be a lower-bound on the original μ_i strongly-convex function as $\tilde{f}_i(w) := f_i(w_i) + \langle \nabla f_i(w_i), w - w_i \rangle + \frac{\mu_i}{2} \|w - w_i\|^2$, then $\nabla \tilde{f}_i(w) = \nabla f_i(w_i) + \mu_i [w - w_i]$.
- Using FTL on $\tilde{f}_k$ instead and using that $\sum_{i=1}^{k} \nabla \tilde{f}_i(w_{k+1}) = 0$ and $\sum_{i=1}^{k-1} \nabla \tilde{f}_i(w_k) = 0$,

$$\sum_{i=1}^{k} \nabla f_i(w_i) + w_{k+1} \left[\sum_{i=1}^{k} \mu_i\right] = \sum_{i=1}^{k} \mu_i w_i \quad ; \quad \sum_{i=1}^{k-1} \nabla f_i(w_i) + w_k \left[\sum_{i=1}^{k-1} \mu_i\right] = \sum_{i=1}^{k-1} \mu_i w_i$$

$$\nabla f_k(w_k) + (w_{k+1} - w_k) \left[\sum_{i=1}^{k} \mu_i\right] = 0 \implies w_{k+1} = w_k - \eta_k \nabla f_k(w_k). \text{ (where } \eta_k := 1/\textstyle\sum_{i=1}^{k} \mu_i)$$

(Adding $\mu_k w_k$ to the second equation, and subtracting the two equations)

Hence, in the strongly-convex setting, running FTL on $\tilde{f}_k$ (a quadratic lower-bound on f_k) recovers OGD on f_k.

Follow the Leader

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

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

Hence, FTL achieves the same regret as OGD when the sequence of losses is strongly-convex and Lipschitz (we will prove this later today).

- What about when the losses are convex but not strongly-convex?

Consider running FTL on the following problem. $\mathcal{C} = [-1, 1]$ and $f_k(w) = \langle z_k, w \rangle$ where

$$z_1 = -0.5; \quad z_k = 1 \quad \text{for } k = 2, 4, \ldots; \quad z_k = -1 \quad \text{for } k = 3, 5, \ldots$$

In round 1, FTL suffers $-0.5w_1$ cost and will compute $w_2 = 1$. It will suffer cost of 1 in round 2 and compute $w_3 = -1$. In round 3, it will thus suffer a cost of 1 and so on. Hence, FTL will suffer $O(T)$ regret if the losses are not strongly-convex.

Follow the Regularized Leader

A way to fix the performance of FTL for a convex sequence of losses is to add an explicit regularization resulting in *Follow the Regularized Leader*.

Follow the Regularized Leader (FTRL): At iteration $k \geq 0$, the algorithm chooses w_{k+1} as:

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

where $\sigma_i > 0$ is the regularization strength.

- Intuitively, since FTRL is equivalent to running FTL on a sequence of strongly-convex (because of the additional regularization) losses, it can obtain sublinear regret even for convex f_k.
- If we set $\sigma_i = 0$ for all i, FTRL reduces to FTL.

Follow the Regularized Leader and OGD

To connect FTRL and OGD, consider the case when $\mathcal{C} = \mathbb{R}^d$ and set $\sigma_0 = 0$.

$$w_{k+1} = \operatorname*{arg\,min}_{w \in \mathbb{R}} \sum_{i=1}^{k} \left[f_i(w) + \frac{\sigma_i}{2} \left\| w - w_i \right\|^2 \right] \implies \sum_{i=1}^{k} \nabla f_i(w_{k+1}) + w_{k+1} \left[\sum_{i=1}^{k} \sigma_i \right] = \sum_{i=1}^{k} \sigma_i w_i$$

- If we define $\tilde{f}_i(w)$ to be a lower-bound on the original convex function as $\tilde{f}_i(w) := f_i(w_i) + \langle \nabla f_i(w_i), w - w_i \rangle$, then, $\forall w$, $\nabla \tilde{f}_i(w) = \nabla f_i(w_i)$.
- Using FTRL on $\tilde{f}_k$ instead and computing the gradients at w_{k+1} and w_k,

$$\sum_{i=1}^{k} \nabla f_i(w_i) + w_{k+1} \left[\sum_{i=1}^{k} \sigma_i \right] = \sum_{i=1}^{k} \sigma_i w_i \quad ; \quad \sum_{i=1}^{k-1} \nabla f_i(w_i) + w_k \left[\sum_{i=1}^{k-1} \sigma_i \right] = \sum_{i=1}^{k-1} \sigma_i w_i$$

$$\nabla f_k(w_k) + (w_{k+1} - w_k) \left(\sum_{i=1}^{k} \sigma_i \right) = 0 \implies w_{k+1} = w_k - \eta_k \nabla f_k(w_k),$$

(Adding $\sigma_k w_k$ to the second equation, and subtracting the two equations)

where $\eta_k := 1/(\sum_{i=1}^{k} \sigma_i)$. Hence, in the general convex setting, running FTRL on $\tilde{f}_k$ (a linear lower-bound on f_k) recovers OGD on f_k.

Questions?

Follow the Regularized Leader

- To analyze FTRL, define $\psi_k(w) := \sum_{i=1}^{k-1} \frac{\sigma_i}{2} \|w - w_i\|^2 + \frac{\sigma_0}{2} \|w\|^2$. At iteration $k-1$, FTRL uses the knowledge of the losses upto $k-1$ and computes the decision for iteration k as:

$$w_k = \arg\min_{w \in \mathcal{C}} F_k(w) \quad \text{where} \quad F_k(w) := \sum_{i=1}^{k-1} f_i(w) + \psi_k(w) \,.$$

- Hence F_k is $\lambda_k := \sum_{i=1}^{k-1} \mu_i + \sum_{i=0}^{k-1} \sigma_i$ strongly-convex. The regularizer ψ_k is known as a *proximal regularizer* and satisfies the condition that,

$$w_k = \arg\min \left[\psi_{k+1}(w) - \psi_k(w)\right] \implies \nabla \psi_{k+1}(w_k) - \nabla \psi_k(w_k) = 0$$

- In order to simplify the analysis, we will assume that w_k lies in the interior of $\mathcal{C}$. This assumption is not necessary and can be handled by augmenting the loss with an indicator function $\mathcal{I}_{\mathcal{C}}$ (see [Ora19, Sec 7.2]).

- We will also assume that the minimization for the w_k update is done exactly. Hence $\nabla F_k(w_k) = 0$ for all k.

Follow the Regularized Leader

Claim: For an arbitrary sequence losses such that each f_k is convex and differentiable, FTRL with the update $w_k = \arg\min_{w \in \mathcal{C}} F_k(w)$ satisfies the following regret for all $u \in \mathcal{C}$,

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

Proof: For $k \geq 1$,

$$F_{k+1}(w_k) - F_{k+1}(w_{k+1}) \leq \langle \nabla F_{k+1}(w_{k+1}), w_k - w_{k+1} \rangle + \frac{1}{2\lambda_{k+1}} \left\| \nabla F_{k+1}(w_k) - \nabla F_{k+1}(w_{k+1}) \right\|^2$$

(By λ_{k+1} strong-convexity of F_{k+1})

$$\leq \frac{1}{2\lambda_{k+1}} \left\| \nabla F_{k+1}(w_k) \right\|^2 \qquad \text{(Since } \nabla F_{k+1}(w_{k+1}) = 0\text{)}$$

$$\implies F_{k+1}(w_k) - F_{k+1}(w_{k+1}) \leq \frac{1}{2\lambda_{k+1}} \left\| \sum_{i=1}^{k} \nabla f_i(w_k) + \nabla \psi_{k+1}(w_k) \right\|^2 \qquad \text{(By def. of } F_{k+1}\text{)}$$

Follow the Regularized Leader

Recall that $F_{k+1}(w_k) - F_{k+1}(w_{k+1}) \leq \frac{1}{2\lambda_{k+1}} \left\| \sum_{i=1}^{k} \nabla f_i(w_k) + \nabla \psi_{k+1}(w_k) \right\|^2$

$$
\begin{aligned}
&F_{k+1}(w_k) - F_{k+1}(w_{k+1}) \\
&\leq \frac{1}{2\lambda_{k+1}} \left\| \left[\sum_{i=1}^{k-1} \nabla f_i(w_k) + \nabla \psi_k(w_k) \right] + \nabla f_k(w_k) + [\nabla \psi_{k+1}(w_k) - \nabla \psi_k(w_k)] \right\|^2 \\
&= \frac{1}{2\lambda_{k+1}} \left\| \nabla f_k(w_k) + [\nabla \psi_{k+1}(w_k) - \nabla \psi_k(w_k)] \right\|^2 \qquad \text{(Since } \nabla F_k(w_k) = 0\text{)}
\end{aligned}
$$

$$
\implies F_{k+1}(w_k) - F_{k+1}(w_{k+1}) \leq \frac{1}{2\lambda_{k+1}} \left\| \nabla f_k(w_k) \right\|^2 \qquad \text{(Since } \nabla \psi_{k+1}(w_k) - \nabla \psi_k(w_k) = 0\text{)}
$$

$$
\begin{aligned}
&F_{k+1}(w_k) - F_{k+1}(w_{k+1}) = [F_{k+1}(w_k) - F_k(w_k)] + [F_k(w_k) - F_{k+1}(w_{k+1})] \\
&= [f_k(w_k) + \psi_{k+1}(w_k) - \psi_k(w_k)] + [F_k(w_k) - F_{k+1}(w_{k+1})]
\end{aligned}
$$

Putting everything together,

$$
\implies [f_k(w_k) + \psi_{k+1}(w_k) - \psi_k(w_k)] + [F_k(w_k) - F_{k+1}(w_{k+1})] \leq \frac{1}{2\lambda_{k+1}} \left\| \nabla f_k(w_k) \right\|^2
$$

Follow the Regularized Leader

Recall that $[f_k(w_k) + \psi_{k+1}(w_k) - \psi_k(w_k)] + [F_k(w_k) - F_{k+1}(w_{k+1})] \leq \frac{1}{2\lambda_{k+1}} \|\nabla f_k(w_k)\|^2$.

$$[f_k(w_k) - f_k(u)] + [F_k(w_k) - F_{k+1}(w_{k+1})] \leq \frac{1}{2\lambda_{k+1}} \|\nabla f_k(w_k)\|^2 + [\psi_k(w_k) - \psi_{k+1}(w_k)] - f_k(u)$$

$$R_T(u) + F_1\underbrace{(w_1)}_{=\frac{\sigma_0}{2}\|w_1\|^2 \geq 0} - F_{T+1}(w_{T+1}) \leq \sum_{k=1}^{T}\left[\frac{1}{2\lambda_{k+1}}\|\nabla f_k(w_k)\|^2\right] + \underbrace{\sum_{k=1}^{T}[\psi_k(w_k) - \psi_{k+1}(w_k)]}_{=-\frac{\sigma_k}{2}\|w_k - w_k\|^2 = 0} - \sum_{k=1}^{T} f_k(u)$$

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

$$\leq \sum_{k=1}^{T}\left[\frac{1}{2\lambda_{k+1}}\|\nabla f_k(w_k)\|^2\right] + \underbrace{[F_{T+1}(w_{T+1}) - F_{T+1}(u)]}_{\text{Non-Positive since } w_{T+1} := \arg\min F_{T+1}(w)} + \psi_{T+1}(u)$$

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

Follow the Regularized Leader - Convex, Lipschitz functions

Claim: If the convex set $\mathcal{C}$ has a diameter D and for an arbitrary sequence of losses such that each f_k is convex, G-Lipschitz and differentiable, then FTRL with $\eta_k := \frac{1}{\sum_{i=0}^{k} \sigma_i} = \frac{\sqrt{D^2 + \|u\|^2}}{\sqrt{2}\, G \sqrt{k}}$ satisfies the following regret bound for all $u \in \mathcal{C}$,

$$R_T(u) \leq \sqrt{2} \sqrt{D^2 + \|u\|^2}\, G \sqrt{T}$$

Proof: Using the general result from the previous slide, for $\lambda_{k+1} = \sum_{i=1}^{k} \mu_i + \sum_{i=0}^{k} \sigma_i$. Since f_k is not necessarily strongly-convex, $\lambda_{k+1} = \sum_{i=0}^{k} \sigma_i$

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

$$\leq \sum_{k=1}^{T} \left[\frac{1}{2 \sum_{i=0}^{k} \sigma_i} \|\nabla f_k(w_k)\|^2 \right] + \frac{D^2 + \|u\|^2}{2} \sum_{i=0}^{T} \sigma_i \qquad \text{(Since } \|u - w_i\|^2 \leq D\text{)}$$

$$R_T(u) \leq \frac{G^2}{2} \sum_{k=1}^{T} \left[\frac{1}{\sum_{i=0}^{k} \sigma_i} \right] + \frac{D^2 + \|u\|^2}{2} \sum_{i=0}^{T} \sigma_i \qquad \text{(Since } f_k \text{ is } G\text{-Lipschitz)}$$

Follow the Regularized Leader - Convex, Lipschitz functions

Recall that $R_T(u) \leq \frac{G^2}{2} \sum_{k=1}^{T} \left[\frac{1}{\sum_{i=0}^{k} \sigma_i} \right] + \frac{D^2 + \|u\|^2}{2} \sum_{i=0}^{T} \sigma_i$. Denoting $\eta_k := \frac{1}{\sum_{i=0}^{k} \sigma_i}$,

$$R_T(u) \leq \frac{G^2}{2} \sum_{k=1}^{T} \eta_k + \frac{(D^2 + \|u\|^2)}{2\eta_T} = G^2\, \eta \sqrt{T} + \frac{(D^2 + \|u\|^2) \sqrt{T}}{2\eta} \qquad \text{(Since } \eta_k = \tfrac{\eta}{\sqrt{k}}\text{)}$$

Using $\eta = \frac{\sqrt{D^2 + \|u\|^2}}{\sqrt{2} G}$,

$$R_T(u) \leq \sqrt{2} \sqrt{D^2 + \|u\|^2}\, G \sqrt{T}$$

- If $0 \in \mathcal{C}$, then $\|u\|^2 \leq D^2$, and this is the regret bound we derived for OGD (upto a $\sqrt{2}$ factor)!
- Hence, though FTL incurs linear regret for convex, Lipschitz losses, FTRL can attain the optimal $\Theta(\sqrt{T})$ regret.

Follow the Leader - Strongly-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 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 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 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 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 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 $u \in \mathcal{C}$,

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

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 G and μ.

Questions?

- Francesco Orabona, *A modern introduction to online learning*, arXiv preprint arXiv:1912.13213 (2019).
- Stéphane Ross, Geoffrey Gordon, and Drew Bagnell, *A reduction of imitation learning and structured prediction to no-regret online learning*, Proceedings of the fourteenth international conference on artificial intelligence and statistics, JMLR Workshop and Conference Proceedings, 2011, pp. 627–635.
- 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.