CMPT 409/981: Optimization for Machine Learning

Lecture 14

Sharan Vaswani

November 3, 2022

Recap

Function class	L-smooth + convex	L-smooth + μ-strongly convex	G-Lipschitz + convex	G-Lipschitz + μ-strongly convex
GD	$O\left(1/\epsilon\right)$	$O\left(\kappa \log\left(1/\epsilon\right)\right)$	$\Theta\left(1/\epsilon^2\right)$	$\Theta\left(1/\epsilon\right)$
SGD	$\Theta\left(1/\epsilon^2\right)$	$\Theta\left(1/\epsilon\right)$	$\Theta\left(1/\epsilon^2\right)$	$\Theta\left(1/\epsilon\right)$

Table 1: Number of iterations required for obtaining an ϵ-sub-optimality.

Today, we will consider online convex optimization for Lipschitz functions.

Online Optimization

Online Optimization

Online Optimization (w_0, Algorithm $\mathcal{A}$, Convex set $\mathcal{C}$)
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

Application: Prediction from Expert Advice – Given n experts, $\mathcal{C} = \Delta_n = \{w_i | w_i \geq 0 \; ; \; \sum_{i=1}^{n} w_i = 1\}$ and $f_k(w_k) = \langle c_k, w_k \rangle$ where $c_k \in \mathbb{R}^n$ is the loss vector.

Application: Imitation Learning – Given access to an expert that knows what action $a \in [A]$ to take in each state $s \in [S]$, learn a policy $\pi : [S] \to [A]$ that imitates the expert, i.e. we want that $\pi(a|s) \approx \pi_{\text{expert}}(a|s)$. Here, $w = \pi$ and $\mathcal{C} = \Delta_A \times \Delta_A \ldots \Delta_A$ (simplex for each state) and f_k is a measure of discrepancy between π_k and π_{expert}.

Online Optimization

Recall that the sequence of losses $\{f_k\}_{k=1}^{T}$ is potentially adversarial and can also depend on w_k.

Objective: Do well against the *best fixed decision in hindsight*, i.e. if we knew the entire sequence of losses beforehand, we would choose $w^* := \arg\min_{w \in \mathcal{C}} \sum_{k=1}^{T} f_k(w)$.

Regret: For any fixed decision $u \in \mathcal{C}$,

$$R_T(u) := \sum_{k=1}^{T} [f_k(w_k) - f_k(u)]$$

When comparing against the best decision in hindsight,

$$R_T := \sum_{k=1}^{T} [f_k(w_k)] - \min_{w \in \mathcal{C}} \sum_{k=1}^{T} f_k(w).$$

We want to design algorithms that achieve a *sublinear regret* (that grows as $o(T)$). A sublinear regret implies that the performance of our sequence of decisions is approaching that of w^*.

Online Convex Optimization

Online Convex Optimization (OCO): When the losses f_k are (strongly) convex loss functions.

Example 1: In prediction with expert advice, $f_k(w) = \langle c_k, w \rangle$ is a linear function.

Example 2: 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.

Example 3: In online control such as LQR (linear quadratic regulator) with unknown costs/perturbations, f_k is quadratic.

In Examples 2-3, the loss at iteration $k+1$ depends on the *learner*'s decision at iteration k.

Online Convex Optimization

Online-to-Batch conversion: If the sequence of loss functions is i.i.d from some fixed distribution, we can convert the regret guarantees into the traditional convergence guarantees for the resulting algorithm.

Formally, if f_k are convex and $R(T) = O(\sqrt{T})$, then taking the expectation w.r.t the distribution generating the losses,

$$\mathbb{E}\left[\frac{R_T}{T}\right] = \mathbb{E}\left[\frac{\sum_{k=1}^{T}[f_k(w_k)] - \sum_{k=1}^{T} f_k(w^*)}{T}\right] \geq \sum_{k=1}^{T}[f(\bar{w}_T) - f(w^*)] = O\left(\frac{1}{\sqrt{T}}\right)$$

where $f(w) := \mathbb{E}[f_k(w)]$ (since the losses are i.i.d) and $\bar{w}_T := \frac{\sum_{k=1}^{T} w_k}{T}$ (since the losses are convex, we used Jensen's inequality).

If the distribution generating the losses is a uniform discrete distribution on n fixed data-points, then $f(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w)$ and we are back in the finite-sum minimization setting.

Hence, algorithms that attain $R(T) = O(\sqrt{T})$ can result in an $O\left(\frac{1}{\sqrt{T}}\right)$ convergence (in terms of the function values) for convex losses.

Questions?

Online Gradient Descent

The simplest algorithm that results in sublinear regret for OCO is *Online Gradient Descent*.

Online Gradient Descent (OGD): At iteration k, the algorithm chooses the point w_k. After the loss function f_k is revealed, OGD suffers a cost $f_k(w_k)$ and uses the function to compute

$$w_{k+1} = \Pi_C[w_k - \eta_k \nabla f_k(w_k)]$$

where $\Pi_C[x] = \arg\min_{y \in \mathcal{C}} \frac{1}{2} \|y - x\|^2$.

Claim: If the convex set $\mathcal{C}$ has a diameter D i.e. for all $x, y \in \mathcal{C}$, $\|x - y\|^2 \leq D$, for an arbitrary sequence losses such that each f_k is convex and differentiable, OGD with a non-increasing sequence of step-sizes i.e. $\eta_k \leq \eta_{k-1}$ and $w_1 \in \mathcal{C}$ has the following regret for all $u \in \mathcal{C}$,

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

Online Gradient Descent - Convex functions

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

$$\|w_{k+1} - u\|^2 = \|\Pi_{\mathcal{C}}[w_k - \eta_k \nabla f_k(w_k)] - u\|^2 = \|\Pi_{\mathcal{C}}[w_k - \eta_k \nabla f_k(w_k)] - \Pi_{\mathcal{C}}[u]\|^2$$

Since projections are non-expansive i.e. for all x, y, $\|\Pi_{\mathcal{C}}[y] - \Pi_{\mathcal{C}}[x]\| \leq \|y - x\|$,

$$\begin{aligned}
&\leq \|w_k - \eta_k \nabla f_k(w_k) - u\|^2 \\
&= \|w_k - u\|^2 - 2\eta_k \langle \nabla f_k(w_k), w_k - u \rangle + \eta_k^2 \|\nabla f_k(w_k)\|^2 \\
&\leq \|w_k - u\|^2 - 2\eta_k [f_k(w_k) - f_k(u)] + \eta_k^2 \|\nabla f_k(w_k)\|^2
\end{aligned}$$

(Since f_k is convex)

$$\implies 2\eta_k [f_k(w_k) - f_k(u)] \leq [\|w_k - u\|^2 - \|w_{k+1} - u\|^2] + \eta_k^2 \|\nabla f_k(w_k)\|^2$$

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

Online Gradient Descent - Convex functions

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

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

$$= \sum_{k=2}^{T} \left[\|w_k - u\|^2 \underbrace{\left(\frac{1}{2\eta_k} - \frac{1}{2\eta_{k-1}}\right)}_{\text{Non-negative since } \eta_k \le \eta_{k-1}} \right] + \frac{\|w_1 - u\|^2}{2\eta_1} - \frac{\|w_{T+1} - u\|^2}{2\eta_T}$$

$$\le D^2 \sum_{k=2}^{T} \left[\frac{1}{2\eta_k} - \frac{1}{2\eta_{k-1}}\right] + \frac{D^2}{2\eta_1} = D^2 \left[\frac{1}{2\eta_T} - \frac{1}{2\eta_1}\right] + \frac{D^2}{2\eta_1} = \frac{D^2}{2\eta_T}$$

(Since $\|x - y\| \le D$ for all $x, y \in \mathcal{C}$)

Putting everything together,

$$R_T(u) \le \frac{D^2}{2\eta_T} + \sum_{k=1}^{T} \frac{\eta_k}{2} \|\nabla f_k(w_k)\|^2$$

Online Gradient Descent - Convex, Lipschitz functions

Claim: If the convex set $\mathcal{C}$ has a diameter D i.e. for all $x, y \in \mathcal{C}$, $\|x-y\|^2 \leq D$, for an arbitrary sequence 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 $u \in \mathcal{C}$,

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

Proof: Since the step-size is decreasing, we can use the general result from the previous slide,

$$R_T(u) \leq \frac{D^2}{2\eta_T} + \sum_{k=1}^{T} \frac{\eta_k}{2}\|\nabla f_k(w_k)\|^2 \leq \frac{D^2}{2\eta_T} + \frac{G^2}{2}\sum_{k=1}^{T}\eta_k \qquad \text{(Since } f_k \text{ is } G\text{-Lipschitz)}$$

$$\implies R_T(u) \leq \frac{D^2\sqrt{T}}{2\eta} + \frac{G^2\eta}{2}\sum_{k=1}^{T}\frac{1}{\sqrt{k}} \leq \frac{D^2\sqrt{T}}{2\eta} + G^2\sqrt{T}\,\eta \qquad \left(\text{Since } \textstyle\sum_{k=1}^{T}\frac{1}{\sqrt{k}} \leq 2\sqrt{T}\right)$$

In order to find the "best" η, set it such that $D^2/2\eta = G^2\eta$, implying that $\eta = D/\sqrt{2}G$ and $R_T(u) \leq \sqrt{2}\,DG\sqrt{T}$. Hence, OGD with a decreasing step-size attains sublinear $\Theta(\sqrt{T})$ regret for convex, Lipschitz functions.

Online Gradient Descent - Strongly-convex, Lipschitz functions

Claim: If the convex set $\mathcal{C}$ has a diameter D, for an arbitrary sequence 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 $u \in \mathcal{C}$,

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

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 $u \in \mathcal{C}$,

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

(Since f_k is μ_k strongly-convex)

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

(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 $\sum_{k=1}^{T} 1/k \leq 1 + \log(T)$)

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

Questions?

Follow the Leader

Another algorithm that achieves logarithmic regret for strongly-convex losses is *Follow the Leader*.

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 the 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 $\mathcal{C} = \mathbb{R}$.

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

If we redefine $f_i(w)$ to be a lower-bound on the original μ_i strongly-convex function as $f_i(w) := f_i(w_i) + \langle \nabla f_i(w_i), w - w_i \rangle + \frac{\mu_i}{2} \|w - w_i\|^2$, then $\nabla f_i(w) = \nabla f_i(w_i) + \mu_i [w - w_i]$.
Computing the gradients at w_{k+1} and w_k,

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

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

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

where $\eta_k := \frac{1}{\sum_{i=1}^{k} \mu_i}$. Hence, running FTL on the lower-bound for the loss (instead of the loss itself) recovers OGD in the strongly-convex case!

Follow the Leader

Claim: If the convex set $\mathcal{C}$ has a diameter D, for an arbitrary sequence 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 are strongly-convex and Lipschitz (we will prove this later)

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 cost $-0.5 w_1$ 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:

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

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

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 $\mathcal{C} = \mathbb{R}$ and set $\sigma_0 = 0$.

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

If we redefine $f_i(w)$ to be a lower-bound on the original convex function as $f_i(w) := f_i(w_i) + \langle \nabla f_i(w_i), w - w_i \rangle$, then, $\nabla f_i(w) = \nabla f_i(w_i)$.

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, running FTRL on a lower-bound for the loss (instead of the loss itself) recovers OGD in the convex case!

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:

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

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 $\mathcal{C}$. Hence $\nabla F_k(w_k) = 0$ for all k. 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]).

Follow the Regularized Leader

Claim: For an arbitrary sequence losses such that each f_k is convex and differentiable, FTRL with the update $w_k = \arg\min_{w\in\mathcal{C}} F_k(w) = \sum_{i=1}^{k-1} f_i(w) + \psi_k(w)$ such that $\psi_k(w) = \sum_{i=1}^{k-1} \frac{\sigma_i}{2}\left\|w - w_i\right\|^2 + \frac{\sigma_0}{2}\left\|w\right\|^2$ and $\lambda_k = \sum_{i=1}^{k-1}[\mu_i] + \sum_{i=0}^{k}[\sigma_i]$ 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 $F_{k+1}(w_k) - F_{k+1}(w_{k+1}) \leq \frac{1}{2\lambda_{k+1}} \left\| \sum_{i=1}^{k} \nabla f_i(w_k) + \nabla \psi_{k+1}(w_k) \right\|^2$

$$\implies F_{k+1}(w_k) - F_{k+1}(w_{k+1})$$

$$= \frac{1}{2\lambda_{k+1}} \left\| \left[\sum_{i=1}^{k-1} \nabla f_i(w_k) + \nabla \psi_k(w_k) \right] + \nabla f_k(w_k) + [\nabla \psi_{k+1}(w_k) - \nabla \psi_k(w_k)] \right\|^2$$

$$= \frac{1}{2\lambda_{k+1}} \left\| \nabla f_k(w_k) + [\nabla \psi_{k+1}(w_k) - \nabla \psi_k(w_k)] \right\|^2 \qquad (\text{Since } \nabla F_k(w_k) = 0)$$

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

$$F_{k+1}(w_k) - F_{k+1}(w_{k+1}) = [F_{k+1}(w_k) - F_k(w_k)] + [F_k(w_k) - F_{k+1}(w_{k+1})]$$

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

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 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, $\lambda_{k+1} = \sum_{i=0}^{k} \sigma_i$

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

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

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

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 $\eta_k := \frac{1}{\sum_{i=0}^{k} \sigma_i}$,

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

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 exactly 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.

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.