Lecture 19: Online Convex Optimization

Instructor: Vatsal Sharan *Scribe: Sophie Hsu*

Recap:

Follow the regularized leader (FTRL)

FTRL algorithm: At every time t, choose

$$w_t \in \text{argmin}_{w \in S}(\psi(w) + \sum_{i=1}^{t-1} f_i(w)).$$

Note: FTL is FTRL with $\psi = 0$.

Linear f_t, quadratic ψ

Theorem 1. *For any $\eta > 0$, FTRL with $S \subseteq \mathbb{R}^d$ a convex set and $\psi(w) = \dfrac{\|w\|_2^2}{2\eta}$, $f_t(w) =< w, v_t >$ satisfies,*

$$\text{Regret}(u, T) \leq \frac{1}{2\eta}\|u\|_2^2 + \eta \sum_{t=1}^{T} \|v_t\|_2^2 .$$

If $\|u\|_2 \leq B$ and $\|v_t\|_2 \leq L$, then choosing $\eta = \dfrac{B}{L\sqrt{T}}$ gives $\text{Regret}(T) = O(BL\sqrt{T})$.

Beyond linear functions: online convex optimization

Computing the FTL or FTRL solution is computationally intractable without further assumptions on the functions. Therefore, we consider convex functions to avoid intractability, and for convex functions, a linear approximation to the function suffices.

Algorithm 1 Online gradient descent (OGD)

Let $w_1 = 0$, $\theta_1 = 0$
for for $t = 1, \ldots,$ T **do**
  predict w_t, receive f_t.
  find gradient $v_t = \partial f_t\left(w_t\right)$.
  if $S = \mathbb{R}^d$ **then**
    $w_{t+1} = w_t - \eta v_t$
  else
    $w_{t+1} = \Pi_S\left(\eta \theta_{t+1}\right), \quad \theta_{t+1} = \theta_t - v_t$

Note: If f is not differentiable, we can use subgradients.

Theorem 2. *OGD enjoys the following regret bound for every $w^* \in S$,*

$$\text{Regret}(w^*, T) \leq \frac{\|w^*\|^2}{2\eta} + \eta \sum_{t=1}^{T} \|v_t\|_2^2$$

If $\|v_t\| \leq \rho \; \forall \; t$ (which is true if f_t is ρ-Lipschitz for all t) and $\|w^*\|_2 \leq B$, than setting $\eta = \dfrac{B}{\rho\sqrt{T}}$ yields,

$$\text{Regret}(T) \leq B\rho\sqrt{T}.$$

Proof. From our earlier analysis of FTRL, we have a regret bound for linearized losses:

$$\begin{aligned} &\sum_{t=1}^{T} [< w_t, v_t > - < w^*, v_t >] \\ &\leq \frac{\|w^*\|^2}{2\eta} + \eta \sum_{t=1}^{T} \|v_t\|_2^2 \,. \end{aligned}$$

The actual regret is $\sum_{t=1}^{T} [f_t(w_t) - f_t(w^*)]$

Using convexity, since v_t is gradient at w_t :

$$\begin{aligned} f_t(w^*) &\geq f_t(w_t) + < v_t, w^* - w_t > \\ \therefore \quad f_t(w_t) - f_t(w^*) &\leq < w_t, v_t > - < w^*, v_t > . \end{aligned}$$

and the result follows. ☐

Example 1: Leaning with expert advice
$S =$ Simplex in $\mathbb{R}_d \left(\Delta_d = \left\{ w : w \in \mathbb{R}^d, w_i \geq 0 \in [d], \Sigma w_i = 1 \right\} \right)$
$f_t(w) = < w, v_t >$
$v_t = (\ell(h_1(x_t), y_t), \ell(h_2(x_t), y_t), \cdots, \ell(h_d(x_t), y_t))$, where $\ell(h_i(x_t), y_t) \in [0, 1]$, $\forall \; i \in [d]$.

Corollary 3. *FTRL with quadratic regularizer (or OGD) for the experts setting gets Regret $(T) \leq \sqrt{dT}$.*

Proof. We first derive a bound on the ℓ_2 norm B of the weigh vectors corresponding to the set of experts.

Since the experts are in Δd
$\|u\|_1 = 1, \forall u \in \Delta d$
Since $\|u\|_2 \leq \|u\|_1 \Rightarrow \|u\|_2 \leq 1, \forall u \in \Delta_d$
$\therefore B \leq 1$

Gradient: v_t
since $(v_t)_i \in [0, 1] \therefore \|v_t\|_2 \leq \sqrt{d}$
$\therefore \quad L \leq \sqrt{d}$
$\therefore \quad \text{Regret}(T) \leq \sqrt{dT}$. ☐

Note: This depends on $\sqrt{d}$ instead of $\sqrt{\log(d)}$.
Using the entropic regularizer

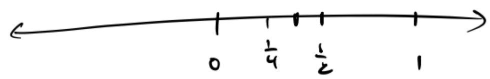
$$\psi(\omega) = \sum_{i=1}^{d} w_i \log(w_i) \quad \text{(negative entropy of } w \text{)}$$

gives the $O(\sqrt{T\log(d)})$ regret bound (and recovers the Weighted-Majority algorithm).

Example 2: Online Perception

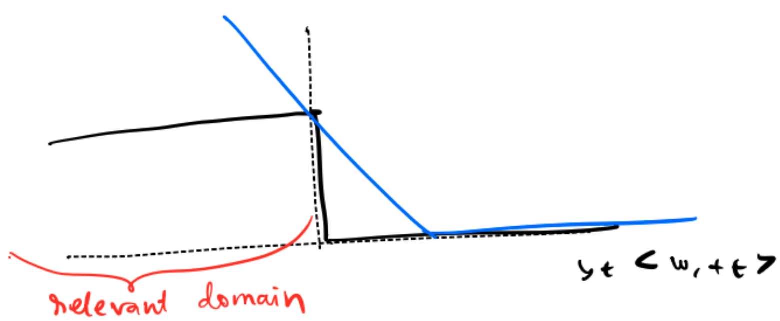
$$x = \mathbb{R}^d$$
$$y = \{\pm 1\}$$

At every time t, learner receives $x_t \in \mathbb{R}^d$.
Maintain $w_t \in \mathbb{R}^d$ and predict $p_t = \text{sign}(< w_t, x_t >)$. We showed earlier that thresholds have $Ldim = \infty$, therefore, no hope of getting small mistake bound without further assumptions.



Convex surrogate loss to avoid this

$$\ell_{0,1}(w,(x,y)) = \mathbb{1}(y < w, x > \leq 0)$$



$*$ Whenever the algorithm makes a mistake, we use the hinge loss:

$$f_t(w) = \max\{0, 1 - y_t < w, x_t >\} = [1 - y_t < w, x_t >]_+$$

* On rounds on which the algorithm is correct, define

$$f_t(w) = 0$$

Note:

- $f_t(w)$ is convex
- for all $\ell_{0,1}(w,(x_t,y_t)) \leq f_t(w)$

Use OGD to learn, $\nabla f_t(w_t) = \begin{cases} 0, & \text{if } y_t\langle w_t, x_t\rangle \ > \ 0 \\ & \text{(since } f_t(w) = 0) \\ -y_t x_t, & \text{if } y_t\langle w_t, x_t\rangle \ < 0 \\ & \text{since } f_t(w) = [1 - y_t\langle w, x_t\rangle]_+ = 1 - y_t < w, x_t > \end{cases}$

$\therefore$ OGD updates:

- $w_1 = 0$
- $w_{t+1} = \begin{cases} w_t, & \text{if } y_t\langle w_t, x_t\rangle \ > 0 \\ w_t + \eta y_t x_t, & \text{otherwise} \end{cases}$

Algorithm 2 Perceptron

Initialize $w_1 = 0$
for for $t = 1, \ldots, T$ **do**
  receive x_t
  predict $p_t = \text{sign}\langle w_t, x_t\rangle$
  if $y_t\langle w_t, x_t\rangle \leq 0$ **then**
    $w_{t+1} = w_t + y_t x_t$ (we can drop η since we just use the sign)
  else
    $w_{t+1} = w_t$

Theorem 4. *Suppose that the Perceptron algorithm runs on a sequence $(x_1, y_1), \ldots, (x_T, y_T)$ and let $R = \max_t \|x_t\|_2$. Let $\mathcal{M}$ be the rounds on which the Perceptron makes a mistake and let $f_t(w) = \mathbb{1}(t \in \mathcal{M})\,[1 - y_t < w, x_t >]_+$. Then, for every w^**

$$|\mathcal{M}| \leq \sum_t f_t(w^*) + R\|w^*\| \sqrt{\sum_t f_t(w^*)} + R^2\|w^*\|^2 \tag{1}$$

If there exists w^* such that $\|w^*\| = 1$ and $y_t < w^*, x_t > \geq \gamma$ for all t, then

$$|\mathcal{M}| \leq R^2/\gamma^2 \tag{2}$$

Proof. By OGD guarantee,

$$\sum_{t=1}^{T} f_t(w_t) - \sum_{t=1}^{T} f_t(w^*) \leq \frac{1}{2\eta}\|w^*\|_2^2 + \frac{\eta}{2}\sum_{t=1}^{T}\|v_t\|_2^2$$

where v_t is the gradient

$$\|v_t\| = \begin{cases} 0 & \text{if } t \notin \mathcal{M} \\ \|x_t\| & \text{if } t \in \mathcal{M} \end{cases}$$

$$\therefore \sum_{t=1}^{T} f_t(\omega) - \sum_{t=1}^{T} f_t(w^*) \leq \frac{\|w^*\|_2^2}{2\eta} + \frac{\eta}{2}|\mathcal{M}|R^2$$

Since $\ell_{0,1}\left(w,\left(x_t, y_t\right)\right) \leq f_t(w)$

$$\sum_{t=1}^{T} f_t(w_t) \geq |\mathcal{M}|$$

$$|\mathcal{M}| - \sum_{t=1}^{T} f_t\left(\omega^*\right) \leq \frac{1}{2\eta}\left\|w^*\right\|_2^2 + \frac{\eta}{2}|\mathcal{M}|R^2$$

This is true $\forall\, \eta \geq 0$. Setting $\eta = \dfrac{\|w^*\|}{R\sqrt{|\mathcal{M}|}}$, we get

$$|\mathcal{M}| \leq \sum_{t=1}^{T} f_t\left(w^*\right) + R\left\|w^*\right\|\sqrt{|\mathcal{M}|}$$

By solving the quadratic, we get Eq (1).

$$|\mathcal{M}| \leq \sum_{t=1}^{T} f_t\left(w^*\right) + \frac{1}{2\eta} + \frac{\eta}{2}|\mathcal{M}|R^2$$

□

Claim 5. $f_t\left(w^*\right) \leq |\mathcal{M}|(1-\gamma)$

Proof. For all $t \notin \mathcal{M}$, $f_t(w^*) = 0$ and for all $t \in \mathcal{M}$, $f_t(w^*) \leq 1-\gamma$,

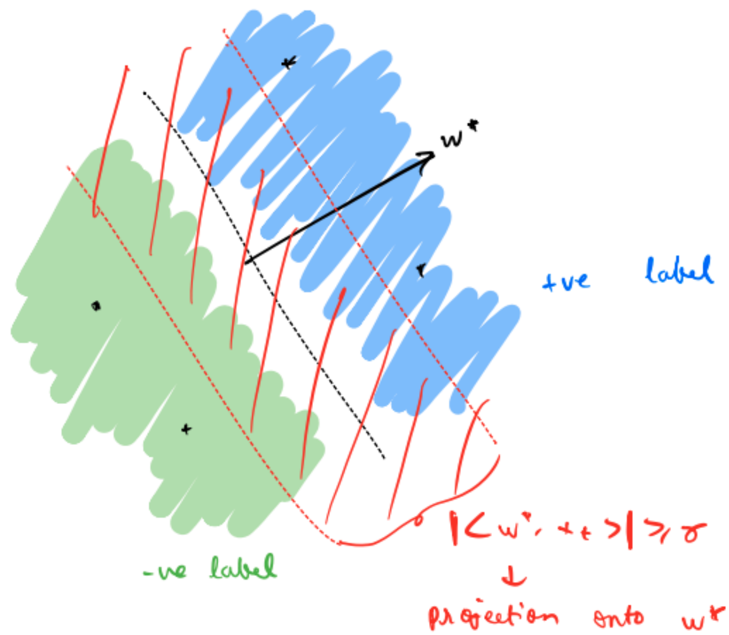
$$|\mathcal{M}| \leq |\mathcal{M}|(1-\gamma) + \frac{1}{2\eta} + \frac{\eta}{2}|\mathcal{M}|R^2$$

Setting $\eta = \dfrac{1}{R\sqrt{|\mathcal{M}|}}$,

$$\gamma|\mathcal{M}| \leq R\sqrt{|\mathcal{M}|} \Rightarrow |\mathcal{M}| \leq R^2/\gamma^2$$

□

The assumption that $y_t\langle w, x_t\rangle \geq \gamma$ is called separability with a margin. If we set $\|w^*\| = 1$, $y_t\langle w^*, x_t\rangle \geq \gamma$ implies that the projection of x_t onto the direction w^* cannot be smaller than γ, which means points cannot lie too close to the separating hyperplane:



w*
+ve label
-ve label
|<w*, x_t>| ≥ γ
projection onto w*