Lecture 18

* Please mark pages on Gradescope when you submit.
* Presentation check-ins start today.

RECAP:

Def (Littlestone dimension) Littlestone dimension of hypothesis class $\mathcal{H}$ ($\text{Ldim}(\mathcal{H})$) is the maximum T s.t. $\exists$ exists a tree of depth T shattered by $\mathcal{H}$.

Online learning in the unrealizable case.

Def (Regret) The regret of an algo A relative to hypothesis h when run on a sequence of T examples is:

$$\text{Regret}_A(h, T) = \sup_{(x_1,y_1),\ldots,(x_T,y_T)} \left[\sum_{t=1}^{T} |p_t - y_t| - \sum_{t=1}^{T} |h(x_t) - y_t| \right]$$

The regret of A relative to a hypothesis class $\mathcal{H}$ is:

$$\text{Regret}_A(\mathcal{H}, T) = \sup_{h \in \mathcal{H}} \text{Regret}_A(h, t).$$

Alg : Weighted Majority (Multiplicative Weights / Hedge)

1. Initialize $w^{(1)} = (1, \dots, 1)$ (d-dimensional)
2. for $t = 1 \dots T$
3. Set $\tilde{w}^{(t)} = w^{(t)} / z_t$ where $z_t = \sum_i w_i^{(t)}$
4. Choose expert i at random according to $P[i] = \tilde{w}_i^{(t)}$.
5. Receive costs of all experts $v_t \in [0,1]^d$
6. Pay expected cost : $\langle \tilde{w}^{(t)}, v_t \rangle$
7. Update : $\forall i : w_i^{(t+1)} = w_i^{(t)} e^{-\eta v_{t,i}}$

Thm Assuming $T \geq 2\log(d)$, the Weighted-Majority algorithm enjoys the bound

$$\sum_{t=1}^{T} \langle w^{(t)}, v_t \rangle - \min_{i \in [d]} \sum_{t=1}^{T} v_{t,i} \leq \sqrt{2\log(d)\, T}$$

Thm: Let $\mathcal{H} = \{h_1, \dots, h_d\}$ be finite hypothesis class. Then Weighted-Majority achieves

$$\sum_{t=1}^{T} |p_t - y_t| - \min_{h \in \mathcal{H}} \sum_{t=1}^{T} |h(x_t) - y_t| \leq \sqrt{2\log(|\mathcal{H}|)\, T}.$$

TODAY:

What if $|\mathcal{H}| = \infty$?

Thm For every hypothesis class $\mathcal{H}$, there exists an algorithm for online learning with a regret bound:

$$\sum_{t=1}^{T} |p_t - y_t| - \min_{h \in \mathcal{H}} \sum_{t=1}^{T} |h(x_t) - y_t| \leq \sqrt{2 L \dim(\mathcal{H}) \log(eT) T}.$$

Online Convex Optimization

At each timestep, learner chooses $w_t \in S$ (for some convex domain S)

Convex loss $f_t : S \to \mathbb{R}$ at every time t.

Learner suffers $f_t(w_t)$ at time t.

$$\text{Regret}(T) = \sum_{t=1}^{T} f_t(w_t) - \min_{w \in S} \sum_{t=1}^{T} f_t(w)$$

Example 1 (linear regression)

$S = \mathbb{R}^d$

$f_t(w) = (\langle w, x_t \rangle - y_t)^2$ (convex as a function of w)

(x_t, y_t) are baked into f_t.

Example 2 (experts)

Set of experts are discrete.

However, we can randomize over experts.

S = Simplex over $\mathbb{R}^d$ $\left[\Delta_d = \{ w : w \in \mathbb{R}^d, w_i \geq 0 \ \forall i \in [d], \sum w_i = 1 \}\right)$

$f_t(w) = \langle w, v_t \rangle$

$v_t = \left(l(h_1(x_t), y_t), l(h_2(x_t), y_t), \ldots, l(h_d(x_t), y_t) \right)$

$\left(l(h(x), y) \in [0,1] \right)$

f_t is merging loss function & data.

Algorithmic frameworks (even beyond convex functions).

Follow-the-leader (FTL)

- At every timestep t, we play

$$w_t \in \operatorname*{argmin}_{w \in S} \sum_{i=1}^{t-1} f_i(w) \quad - (1)$$

Lemma (compare FTL with lookahead oracle).

Let $w_1, \dots, w_T$ be produced by the FTL algorithm according to ①. For any $u \in S$, define

$$\text{Regret}(u, T) = \sum_{t=1}^{T} \left[f_t(w_t) - f_t(u) \right]$$

how stable FTL is

Then $$\text{Regret}(u, T) \le \sum_{t=1}^{T} \left[f_t(w_t) - f_t(w_{t+1}) \right]$$

$$\therefore \text{Regret}(T) = \max_{u \in S} \text{Regret}(u, T) \le \sum_{t=1}^{T} \left[f_t(w_t) - f_t(w_{t+1}) \right]$$

Proof Suffices to show that $\forall u \in S$

$$\sum_{t=1}^{T} f_t(w_{t+1}) \le \sum_{t=1}^{T} f_t(u)$$

Proof by induction:

* Assume inductive hypothesis on $T-1$.

$$\sum_{t=1}^{T-1} f_t(w_{t+1}) \le \sum_{t=1}^{T-1} f_t(u) \qquad \forall u \in S.$$

$$\Rightarrow \sum_{t=1}^{T} f_t(w_{t+1}) \le \sum_{t=1}^{T-1} f_t(u) + f_T(w_{T+1}) \qquad \forall u \in S.$$

$\therefore$ this holds for $u = w_{T+1}$.

$$\sum_{t=1}^{T} f_t(w_{t+1}) \le \sum_{t=1}^{T} f_t(w_{T+1})$$

Since $w_{T+1} \in \operatorname{argmin}_{w \in S} \sum_{t=1}^{T} f_t(w)$

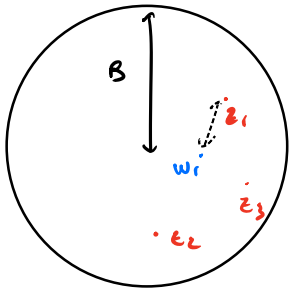
$$\sum_{t=1}^{T} f_t(w_{t+1}) \leq \sum_{t=1}^{T} f_t(u) \quad \forall u \in S,$$

which completes induction step. ∎

Quadratic optimization (FTL works).

Lemma Assume $f_t(w) = \frac{1}{2}\|w - z_t\|^2$ where $\|z_t\| \leq B$ $\forall t \in [T]$. Then FTL has regret $O(B^2 \log(T))$.

Proof



FTL has closed form solution:

$$w_t = \frac{1}{t-1} \sum_{i=1}^{t-1} z_i$$

$$\Rightarrow w_{t+1} = \frac{(t-1)w_t + z_t}{t}$$

$$= \left(1 - \frac{1}{t}\right) w_t + \frac{z_t}{t}$$

Using Lemma (FTL vs. lookahead oracle),

$$\text{Regret}(T) \le \sum_{t=1}^{T} \left[f_t(w_t) - f_t(w_{t+1}) \right]$$

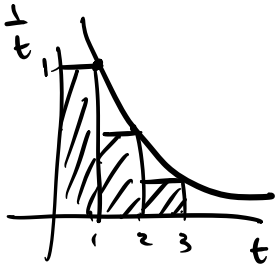
$$f_t(w_t) - f_t(w_{t+1}) = \frac{1}{2}\|w_t - z_t\|^2 - \frac{1}{2}\left\|\left(1-\frac{1}{t}\right)w_t + \frac{z_t}{t} - z_t\right\|^2$$

$$= \frac{1}{2}\left(1 - \left(1-\frac{1}{t}\right)^2\right)\|w_t - z_t\|^2$$

$$\le \left(\frac{1}{t}\right)\|w_t - z_t\|^2$$

$$\le \frac{4B^2}{t}.$$

$$\text{Regret}(T) \le 4B^2 \sum_{t=1}^{T} \left(\frac{1}{t}\right)$$



$$\sum_{t=1}^{T} \frac{1}{t} \le 1 + \int_{1}^{T} \frac{1}{t}\,dt$$

$$= 1 + \log(T)$$

$$\therefore \quad \text{Regret}(T) = O\left(B^2 \log(T)\right)$$

Linear optimization (FTL fails!)

Let $S = [-1, 1]$ be FTL's possible predictions.

Consider linear functions $f_t(w) = w v_t$ (in $d=1$ dimension)

where $(v_1, v_2, \ldots) = (-0.5, 1, -1, 1, -1, \ldots)$

What does FTL do here?

Initialize $w_1 = 0 \rightarrow$ loss $\ell(w_1, v_1) = w_1 v_1 = 0$

Since $f_1(w) = -0.5w \Rightarrow w_2 = 1 \rightarrow$ loss $(w_2, v_2) = w_2 v_2 = 1$

$f_1(w) + f_2(w) = 0.5w \Rightarrow w_3 = -1 \rightarrow$ loss $(w_3, v_3) = 1$

$\vdots$

$\therefore$ FTL gets loss 1 on every example except 1st.

Expert $w = 0$ gets 0 loss.

$\therefore$ Regret $(w, T) = T - 1$

$\therefore$ Regret $(T) \geq T - 1$:(

→ for quadratic functions, w_t & w_{t+1} get closer & closer (we get low regret).

→ for linear functions, w_t & w_{t+1} do not get closer (we get high regret).

Follow the regularized leader (FTRL)

Let $\psi : S \rightarrow \mathbb{R}$ be a function called a regularizer.

Let $f_1, \ldots, f_T$ be the sequence of loss functions played by the environment.

• FTRL algorithm: At every time t, choose

$$w_t \in \operatorname*{argmin}_{w \in S} \left(\psi(w) + \sum_{i=1}^{t-1} f_i(w) \right)$$

(FTL is FTRL with $\psi = 0$).

linear f_t, quadratic ψ

Thm For any $\eta > 0$, FTRL with $S \subseteq \mathbb{R}^d$ a convex set & $\psi(w) = \frac{\|w\|_2^2}{2\eta}$, $f_t(w) = \langle w, v_t \rangle$ satisfies,

$$\text{Regret}(u, T) \leq \underbrace{\frac{1}{2\eta}\|u\|^2}_{\text{large when } \eta \text{ is small, not learning}} + \underbrace{\eta \sum_{t=1}^{T} \|v_t\|_2^2}_{\text{large when } \eta \text{ is large, not stable}}.$$

If $\|u\|_2 \leq B$ & $\|v_t\| \leq L$, then choosing $\eta = \frac{B}{L\sqrt{T}}$

gives $\text{Regret}(T) = O(BL\sqrt{T})$.

Proof

$$\text{FTRL}: w_t = \underset{w \in S}{\operatorname{argmin}} \left(\frac{1}{2\eta}\|w\|^2 - \sum_{i=1}^{t-1} \langle w, \theta_t \rangle \right)$$

where $\theta_t = -\sum_{i=1}^{t-1} v_i$.

By adding $\frac{\eta}{2}\|\theta_t\|^2$ (independent of w) to complete squares,

$$w_t = \underset{w \in S}{\operatorname{argmin}} \left(\frac{1}{2\eta}\|w - \eta\theta_t\|_2^2 \right)$$

Therefore

① for $S = \mathbb{R}^d$, $w_{t+1} = \eta \theta_t$

Rewriting, $w_t = -\eta (v_1 + v_2 + \dots v_{t-1})$

$$w_{t+1} = w_t - \eta v_t$$

② for $S \neq \mathbb{R}^d$, $w_t = \Pi_S (\eta \theta_t)$

where $\Pi_S (x) = \underset{w \in S}{\operatorname{argmin}} \|w - x\|_2^2$

Called a <u>lazy projection step</u>: accumulate gradients θ_t, only project for making predictions (also called Nesterov's dual averaging method).

By using Lemma (FTL vs one-step lookahead) on the sequence $\Psi, f_1, \dots, f_T$, for any $u \in S$

$$\left[\Psi(w_0) - \Psi(u)\right] + \sum_{t=1}^{T} \left[f_t(w_t) - f_t(u)\right]$$
$$\leq \left[\Psi(w_0) - \Psi(w_1)\right] + \sum_{t=1}^{T} \left[f_t(w_t) - f_t(w_{t+1})\right].$$

Since $\Psi(w_1) \geq 0$,

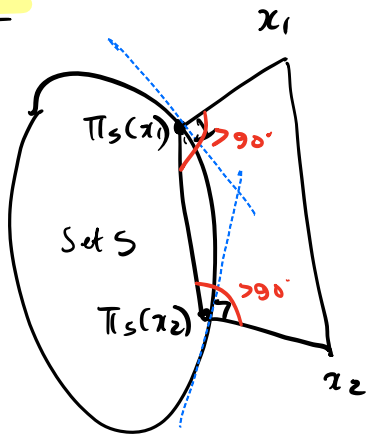
$$\text{Regret}(u, T) = \sum_{t=1}^{T} \left[f_t(w_t) - f_t(u)\right]$$
$$\leq \frac{1}{2\eta} \|u\|_2^2 + \sum_{t=1}^{T} \left[f_t(w_t) - f_t(w_{t+1})\right]$$

$$f_t(w_t) - f_t(w_{t+1}) \le \langle v_t, w_t - w_{t+1} \rangle$$
$$\le \|v_t\|_2 \, \|w_t - w_{t+1}\|_2$$
$$= \|v_t\|_2 \, \|\Pi_S(\eta\theta_t) - \Pi_S(\eta\theta_{t+1})\|_2$$

Claim (Projections onto convex sets are contractive)
If S is convex, then for any $x_1, x_2 \in \mathbb{R}^d$,

$$\|\Pi_S(x_1) - \Pi_S(x_2)\|_2 \le \|x_1 - x_2\|_2.$$

Proof by picture



$$f_t(w_t) - f_t(w_{t+1}) \le \|v_t\|_2 \, \|\eta\theta_t - \eta\theta_{t+1}\|_2$$
$$= \eta \|v_t\|_2^2$$

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

Beyond linear : online convex optimization.

Consider convex functions to avoid intractability.

for convex functions, a linear approximation suffices.

Algorithm : online gradient descent (OGD).

1. Let $w_1 = 0$, $\theta_1 = 0$
2. for $t = 1, \dots, T$
3. Predict w_t, receive f_t
4. Find gradient $v_t = \nabla f_t(w_t)$
5. If $S = \mathbb{R}^d$
6. $w_{t+1} = w_t - \eta v_t$
7. Else
8. $w_{t+1} = \Pi_S(\eta \theta_{t+1})$, $\theta_{t+1} = \theta_t - v_t$

Note : If f is not differentiable, use subgradients.

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

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

If $\|v_t\| \leq \rho \ \forall t$ (f_t is ρ-Lipschitz $\forall t$) & $\|w^*\|_2 \leq B$, then setting $\eta = \frac{B}{\rho\sqrt{T}}$ yields,

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