5.6 Online subgradient descent

Handle general losses efficiently — run FTRL on a linear approximation of f_t.

[Algorithm 4] Online sublinear descent (OGD)

Let $w_1 = 0$.

For iteration $t = 1, \cdots, T$:

(i) Predict w_t and recieve f_t.

(ii) Take any subgradient $z_t \in \partial f_t(w_t)$

(iii) If $S = \mathbb{R}^d$: $w_{t+1} = w_t - \eta z_t$.

If $S \subseteq \mathbb{R}^d$: $w_{t+1} = \Pi_S(\eta \theta_{t+1})$, $\theta_{t+1} = \theta_t - z_t$

Analyzing Regret:

(i) Theorem 30 gives the bound on:

$$\sum_{t=1}^{T} [w_t \cdot z_t - u \cdot z_t]$$

(ii) We want to control regret w.r.t. f_t (convex):

$$\sum_{t=1}^{T} [f_t(w_t) - f_t(u)].$$

Since $z_t \in \partial f_t(w_t)$ is a subgradient, we have by def:

$$f_t(u) \geq f_t(w) + z_t \cdot (u - w_t)$$

$$\Rightarrow \quad f_t(w) - f_t(u) \leq w_t \cdot z_t - u \cdot z_t$$

We get the same bound for general losses as it for linear losses.

[Example 30] Online SVM

Apply OGD on the hinge loss for classification: $x_t \in \mathbb{R}^d$, $y_t \in \{+1, -1\}$.

$$f_t(w) = \max\{0, 1 - y_t(w \cdot x_t)\}$$

The algorithm ($S = \mathbb{R}^d$):

(i) If $y_t(w_t \cdot x_t) \geq 1$ (classify correctly with margin 1): do nothing.

(ii) Else: $w_{t+1} = w_t + \eta y_t x_t$

(ii) Else $w_{t+1} = w_t + \eta y_t x_t$

Analysis:

- Assume $\|x_t\|_2 \le L$, then $\|z_t\|_2 \le L$ $\quad (z_t \in \partial f_t(w_t))$
- Assume $\|u\|_2 \le B$

The regret bound from Th.30 is:

$$\text{Regret} \le BL\sqrt{T}$$

[Example 31] Learning with expert advice

* maintain $w_t \in \Delta_d$ over d expert and predict by sampling.
* Assume $\ell(y_t, p_t) = \mathbb{1}\{y_t \ne p_t\}$
* loss is linear: $f_t(w_t) = w_t \cdot z_t$, where

$$z_t = [\ell(y_t, h_1(x_t)), \cdots, \ell(y_t, h_d(x_t))] \in \{0,1\}^d$$

- Bound on set of experts (B):

  experts $\in$ simplex $S = \Delta_d$, 2-norm bounded by $B=1$
- Bound on loss gradient (L):

  Lipschitz constant is bounded by $\|z_t\|_2 \le \sqrt{d}$

Therefore, by Th 30:

$$\text{Regret} \le BL\sqrt{T} = \sqrt{dT}.$$