Online Optimization/Learning

- **Learner** vs. **Environment**

  $t = 1, 2, \ldots$

  L selects $f_t \in \mathcal{F}$

  E selects $\ell_t : \mathcal{F} \to \mathbb{R}$, reveals* to L

  L incurs loss $\ell_t(f_t)$

  *: full info feedback

Regret of the Learner after T rounds:

$$R_T := \sum_{t=1}^{T} \ell_t(f_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{T} \ell_t(f)$$

want **sublinear** regret: $R_T/T \to 0$ as $T \to \infty$

- **Online Cvx Opt.** (Zinkevich; Hazan, Abernethy, Rakhlin)

  $\mathcal{F}$: closed bdd cvx set in $\mathbb{R}^d$

  $D := \max\{\|f - f'\| : f, f' \in \mathcal{F}\} < \infty$

  $(\ell_t)_{t \geq 0}$: cvx, L-Lip, m-strongly convex

  Projected GD:

  $$f_{t+1} = \Pi(f_t - \alpha_t \nabla \ell_t(f_t))$$

  $$R_T = \begin{cases} O(\sqrt{T}) & \text{for cvx, Lip } (\alpha_t = c/\sqrt{t}) \\ O(\log t) & \text{for m-sc., Lip. } (\alpha_t = \frac{c'}{t}) \end{cases}$$

— can be shown that this is "optimal"...

Perceptron Mistake Bound

(F. Rosenblatt, 1958)

data: $\tilde{z}_i = (\tilde{x}_i, \tilde{y}_i)$ $\quad i \in [n]$

$\tilde{x}_i \in \mathbb{R}^d$, $\quad \tilde{y}_i \in \{\pm 1\}$

classifiers: $x \mapsto \operatorname{sgn}\langle f, x\rangle$ $\quad (f \in \mathbb{R}^d)$

Assume the data are separable: $\exists f^* \in \mathbb{R}^d$ s.t.

$$\tilde{y}_i \langle f^*, \tilde{x}_i\rangle \geq 1 \qquad \forall i \in [n]$$

Goal: find $\hat{f} \in \mathbb{R}^d$ s.t. all examples are classified correctly

Perceptron: $\quad (\alpha > 0$: parameter)

- $f_1 = 0$ $\quad$ (init)
- for $t = 1, 2, \ldots$:

$$I_t \in [n] \leftarrow \text{if } \exists i \in [n] \text{ s.t. } \tilde{y}_i \langle f_t, \tilde{x}_i\rangle \leq 0, \text{ then } I_t = i; \text{ arbitrary o/w}$$

$$z_t = \tilde{z}_{I_t} = (\tilde{x}_{I_t}, \tilde{y}_{I_t})$$

$$f_{t+1} = \begin{cases} f_t & \text{if } y_t \langle f_t, x_t\rangle \geq 0 \\ f_t + \alpha y_t x_t & \text{if } y_t \langle f_t, x_t\rangle < 0 \end{cases}$$

- if f_{t+1} makes no mistakes on data, STOP

Suppose L, B are chosen s.t.:

$$L \geq \max_{i \in [n]} \|\tilde{x}_i\|$$

$$B \geq \min\{\|f^*\| : \min_{i \in [n]} \tilde{y}_i \langle f^*, x_i^* \rangle \geq 1\}$$

[Then Perceptron will find a separating classifier in at most $L^2 B^2$ steps.

Proof

1) map to Online Convex Opt

time t:

- Learner generates f_t (acc. to algo)
- Environment responds with

$$\ell_t(f) := \begin{cases} 0 & \text{if } f_t \text{ is correct on } z_t \\ (1 - y_t \langle f, x_t \rangle)_+ & \text{o/w} \end{cases}$$

$\downarrow$
$\in \mathbb{R}^d$

Note $\ell_t(\cdot)$ are cvx, L-Lip.

$$\nabla \ell_t(f) = \begin{cases} 0, & \text{if } f_t \text{ is correct on } z_t \\ -y_t x_t & \text{o/w} \end{cases}$$

$f_{t+1} = f_t - \alpha \nabla \ell_t(f_t)$: Perceptron $f_1 = 0$

2) Let f^* be s.t. $\|f^*\| \leq B$

and $\tilde{y}_i \langle f^*, \tilde{x}_i \rangle \geq 1 \quad \forall i \in [n]$

Regret (in T rounds) w.r.t. f^*

$$V_t := \|f_t - f^*\|^2 \qquad V_1 = \|f^*\|^2 \le B^2$$

$$V_{t+1} = \|f_{t+1} - f^*\|^2$$

$$= \|f_t - \alpha g_t - f^*\|^2 \qquad g_t := \nabla \ell_t(f_t)$$

$$= V_t - 2\alpha \langle g_t, f_t - f^* \rangle + \alpha^2 \|g_t\|^2$$

$$\le V_t - 2\alpha \langle g_t, f_t - f^* \rangle + \alpha^2 L^2$$

$$\Rightarrow 2\langle g_t, f_t - f^* \rangle \le \frac{V_t - V_{t+1}}{\alpha} + \alpha L^2$$

$\ell_t(\cdot)$ cvx $\Rightarrow$

$$\ell_t(f^*) - \ell_t(f_t) \ge \langle g_t, f^* - f_t \rangle$$

$$\boxed{2\left(\ell_t(f_t) - \ell_t(f^*)\right) \le \frac{V_t - V_{t+1}}{\alpha} + \alpha L^2}$$

3) $\ell_t(f^*) = 0 \quad \forall t$

suppose that $f_1, \dots, f_T$ do not separate the data

$\Rightarrow \ell_t(f_t) \ge 1 \quad \forall t$

$$2 \sum_{t=1}^{T} [\ell_t(f_t) - \ell_t(f^*)] \ge 2T$$

$$2 \sum_{t=1}^{T} [\ell_t(f_t) - \ell_t(f^*)] \le \frac{V_1}{\alpha} + \alpha L^2 T \le \frac{B^2}{\alpha} + L^2 T$$

$\therefore$ if none of $f_1, \dots, f_T$ separate the data, then

$$2T \le \frac{B^2}{\alpha} + \alpha L^2 T \qquad \forall \alpha > 0$$

$$2T \le \inf_{\alpha > 0} \left(\frac{B^2}{\alpha} + \alpha L^2 T \right) = 2LB\sqrt{T}$$

$$\Rightarrow \sqrt{T} \le LB \qquad \blacksquare$$

Online to Batch, Generalization

$z_1, z_2, \dots, z_n \overset{iid}{\sim} P$

- $T = n$ rounds
- $f_1 \in \mathcal{F}$ (init)
- $t = 1, 2, \dots, n$:
  - L generates f_t
  - E "reveals" z_t
  - L incurs loss $\ell(f_t, z_t)$
  - $f_{t+1} = A(f_1, \dots, f_t, z_1, \dots, z_t)$

$$L(f) := \mathbb{E}_P[\ell(f, z)]$$

$$f_t = A(f_1, \dots, f_{t-1}, z_1, \dots, z_{t-1}) \perp\!\!\!\perp z_t$$

$$\Rightarrow \mathbb{E}[\ell(f_t, z_t) \mid z_1, \dots, z_{t-1}] = L(f_t) \quad \forall t$$

Now we can analyze $\frac{1}{T}\sum_{t=1}^{T} L(f_t)$

vs. $\frac{1}{T}\sum_{t=1}^{T} \ell(f_t, z_t)$

- Assume $\ell(f, z) \in [0, 1]$

- $(Y_t)_{t=0}^{T}$: $Y_0 := 0$

$$Y_t := \sum_{s=1}^{t} [L(f_s) - \ell(f_s, z_s)]$$

$$\mathbb{E}[L(f_s) - \ell(f_s, z_s) \mid z_1, \dots, z_{s-1}] = 0$$

$\Rightarrow$ $(Y_t)_{t=0}^{T}$ is a martingale

$\mathbb{E}[Y_t \mid z_1, \dots, z_s] = Y_s$ if $0 \le s < t$

By Azuma-Hoeffding inequality,

$$\mathbb{P}\{Y_T \ge \varepsilon T\} \le \exp\left(-\frac{T\varepsilon^2}{2}\right) \quad \forall \varepsilon > 0$$

Set $\varepsilon = \sqrt{\frac{2\log(1/\delta)}{T}}$, so

$$\frac{1}{T} Y_T \le \sqrt{\frac{2\log(1/\delta)}{T}} \quad \text{w.p. } \ge 1 - \delta$$

or

$$\frac{1}{T}\sum_{t=1}^{T} L(f_t) \le \frac{1}{T}\sum_{t=1}^{T} \ell(f_t, z_t) + \sqrt{\frac{2\log(1/\delta)}{T}}.$$

Regret: $$\frac{1}{T}\sum_{t=1}^{T} \ell(f_t, z_t) \le \underbrace{\inf_{f \in \mathcal{F}} \frac{1}{T}\sum_{t=1}^{T} \ell(f, z_t)}_{\text{min-empirical risk}} + \frac{R_T}{T}$$

- w.p. $\ge 1 - \delta$,

$$\frac{1}{T}\sum_{t=1}^{T} L(f_t) \le \frac{1}{T}\inf_{f \in \mathcal{F}} L_T(f) + \frac{R_T}{T} + \sqrt{\frac{2\log(1/\delta)}{T}}$$

- take f^* s.t. $L(f^*) = \min_{f \in \mathcal{F}} L(f)$

$$\frac{1}{T}\sum_{t=1}^{T} \ell(f_t, z_t) \le \frac{1}{T}\sum_{t=1}^{T} \ell(f^*, z_t) + \frac{R_T}{T}$$

$$= L_T(f^*) + \frac{R_T}{T}$$

$$\le L(f^*) + \sqrt{\frac{2\log(1/\delta)}{T}} + \frac{R_T}{T}$$

w.p. $\ge 1-\delta$ by Hoeffding $(z_1, \dots, z_T$ iid$)$

$\Rightarrow$ w.p. $\ge 1-2\delta$,

$$\frac{1}{T}\sum_{t=1}^{T} L(f_t) \le L(f^*) + \frac{R_T}{T} + \sqrt{\frac{8\log(1/\delta)}{T}}$$

- if $f \mapsto \ell(f, z)$ is convex, then

$\bar{f}_T := \frac{1}{T}\sum_{t=1}^{T} f_t$ satisfies

$$L(\bar{f}_T) \le \frac{1}{T}\sum_{t=1}^{T} L(f_t)$$

$$\le L(f^*) + \frac{R_T}{T} + \sqrt{\frac{8\log(1/\delta)}{T}} \quad \text{w.p.} \ge 1-2\delta.$$