(See note from previous lecture first)

$$2\langle f_t - f^*, \nabla l_t(f_t)\rangle \le \frac{\|f_t - f^*\|^2 - \|f_{t+1} - f^*\|^2}{\alpha_t} + \alpha_t L^2$$

By (strong) convexity

$$l_t(f^*) - l_t(f_t) \ge \langle f^* - f_t, \nabla l_t(f_t)\rangle + \frac{m}{2}\|f^* - f_t\|^2$$

$$2(l_t(f_t) - l_t(f^*)) \le 2\langle f_t - f^*, \nabla l_t(f_t)\rangle - m\|f^* - f_t\|^2$$

so

$$\boxed{2(l_t(f_t) - l_t(f^*)) \le \frac{\|f_t - f^*\|^2 - \|f_{t+1} - f^*\|^2}{\alpha_t} + \alpha_t L^2 - m\|f_t - f^*\|^2}$$

we use

$$\frac{\|f_t - f^*\|^2 - \|f_{t+1} - f^*\|^2}{\alpha_t} = \frac{\|f_t - f^*\|^2}{\alpha_t} - \frac{\|f_{t+1} - f^*\|^2}{\alpha_{t+1}} + \left(\frac{1}{\alpha_{t+1}} - \frac{1}{\alpha_t}\right)\|f_{t+1} - f^*\|^2$$

In HW we have f_t^*

sum $t = 1$ to T

$$2(J_T(l(f_t)) - J_T(f^*)) \le \left(\frac{1}{\alpha_1} - m\right)\overbrace{\|f_1 - f^*\|^2}^{\le D^2} - \frac{1}{\alpha_T}\|f_{T+1} - f^*\|$$

$$+ \sum_{t=1}^{T-1}\left(\frac{1}{\alpha_{t+1}} - \frac{1}{\alpha_t} - m\right)\underbrace{\|f_{t+1} - f^*\|^2}_{D^2} + L^2\sum_{t=1}^{T}\alpha_t$$

$$\le D^2\left(\frac{1}{\alpha_1} - m + \sum_{t=1}^{T-1}\left(\frac{1}{\alpha_{t+1}} - \frac{1}{\alpha_t} - m\right)\right) + L^2\sum_{t=1}^{T}\alpha_t$$

$$\le D^2\left(\frac{1}{\alpha_T} - mT\right) + L^2\sum_{t=1}^{T}\alpha_t$$

(a) $\left(\alpha_t = \frac{c}{\sqrt{t}},\ m = 0\right)$ (b) $\left(\alpha_t = \frac{1}{mt},\ m > 0\right)$

$\le DL\sqrt{2T}$ $\le \frac{L^2 \log T}{m}$

Online perceptron algorithm

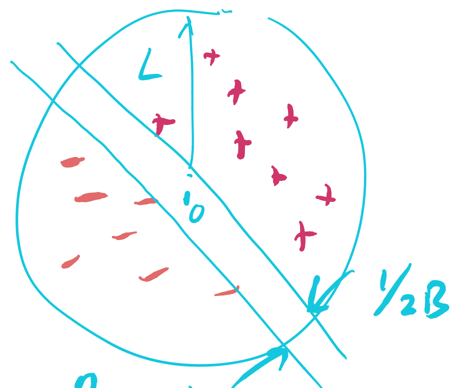
SV

$x_t \in \mathbb{R}^d$ for each t

$y_t \in \{-1, 1\}$

Classifier:

$y = sgn(\langle f, x \rangle)$



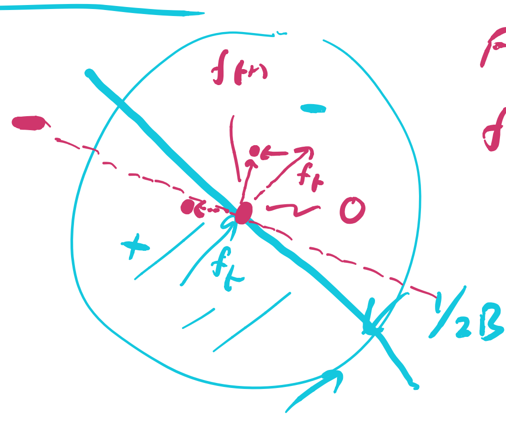
$(x_t, y_t)_{t=1}^n = Z^n$ $Z_t = (x_t, y_t)$

Perceptron Algorithm

Set $f_0 = 0 \in \mathbb{R}^d$.

l_t

$$f_{t+1} = \begin{cases} f_t & \text{if } y_t = sgn(\langle f_t, x_t \rangle) \\ f_t + \alpha_t \underline{y_t x_t} & \text{if } y_t \neq sgn(\langle f_t, x_t \rangle) \end{cases}$$



For this example

$f_{t+1} = f_t - x_t$

similar to gradient descent for hinge loss function

$$\ell(f_t, z_t) = (1 - \langle f_t, x_t\rangle y_t)_+ \qquad \varphi(u) = (1+u)_+$$

$z_t = (x_t, y_t)$

$\varphi(-\langle f_t, x_t\rangle y_t)$

$$\nabla \ell(f_t, z_t) = \begin{cases} -x_t y_t & \text{if } \langle f_t, x_t\rangle y_t < 1 \\ 0 & \text{if } \langle f_t, x_t\rangle y_t > 1 \end{cases}$$

gradient alg.

0

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

Zinkevich gives a guarantee

Use a different adversarial sequence:

- If $y_t\langle f_t, x_t\rangle \ge 0$ then let $\ell_t \equiv 0$

  If $y_t\langle f_t, x_t\rangle < 0$ then $\ell_t \equiv \tilde{\ell}_t(f) = (1 - \langle f, x_t\rangle y_t)_+$

  $(\ell_t(f_t) \ge 1)$

Proposition

Let $L \ge \max_{1 \le i \le n} \|x_i\|$ $\quad B \ge \min\{\|f^*\| : y\langle f^*, x_i\rangle \ge 1$ for all $i\}$

Then at most $B^2 L^2$ updates are needed to get error $= 0$ on data.

Proof Use

$$\sum_{t=1}^{T} 2\left(\ell_t(f_t) - \ell_t(f^*)\right) \le \frac{\|f_t - f^*\|^2 - \|f_{t+1} - f^*\|^2}{\alpha} + \alpha L^2$$

$$2\left(\begin{array}{c}\#\text{ of}\\ \text{updates}\end{array}\right) \le 2\left(J_T(f_t) - J_T(f^*)\right) \le \frac{\|f_1 - f^*\|^2}{\alpha} + \alpha L^2 T, \quad \|f_1 - f^*\|^2 \le B^2$$

$$\alpha = \frac{B}{L\sqrt{T}} \qquad \cancel{2}T \le \cancel{2}BL\sqrt{T}$$

or $T \le (BL)^2$