Jake Schinto

# Homework 4

## Convex optimization

1. R-Lipschitz: $|f(w_1) - f(w_2)| \leq R\|w_1 - w_2\|$

$|\max(0, 1 - y w_1^T x) - \max(0, 1 - y w_2^T x)| \leq R\|w_1 - w_2\|$

there are 4 cases of max:

$(0,0)$: $|0 - 0| \leq R\|w_1 - w_2\| \rightarrow 0 \leq R\|w_1 - w_2\|$

this will be true for any $R \geq 0$ bec $\|w_1 - w_2\|$ is positive

$(0, 1 - y w_2^T x)$:

$|0 - 1 + y w_2^T x| \leq R\|w_1 - w_2\| \rightarrow R \geq \frac{|y w_2^T x - 1|}{\|w_0 - w_2\|}$

$(1 - y w_1^T x, 0)$:

$|1 - y w_1^T x| \leq R\|w_1 - w_2\| \rightarrow R \geq \frac{|1 - y w_1^T x|}{\|w_1 - w_2\|}$

$(1 - y w_1^T x, 1 - y w_2^T x)$

$|1 - y w_1^T x - 1 + y w_2^T x| \leq R\|w_1 - w_2\|$

$|y w_2^T x - y w_1^T x| \leq R\|w_1 - w_2\| \rightarrow R \geq \frac{|y(w_2^T - w_1^T)x|}{\|w_1 - w_2\|}$

In all cases there exists some $R > 0$ that can satisfy all 4 cases, assuming $(x, y)$ is constant

## Learning Theory

1.a. with some distribution over $X$ and a training set of $m$ samples $(x_m, f(x_m))$ an algorithm can be defined as:

- if $y_i = 1$ for some $(x_i, f(x_i))$ (there can only be 1 because of realizability), then $h_{x_i}$ is output
- else output $h^-$

This fits with ERM because 0 is the lowest possible loss because $L_S(h_S) = 0$ if $h_S$ is outputed

1.b. $S_X = \{x_1, x_2, \ldots x_m\}$ and $\epsilon, \delta$ is in $(0,1)$

$$P(\{S_X \mid L_{(D,f)}(h_S) > \epsilon\}) < \delta$$

where true error: $L_{(D,f)}(h) = P(\{h(x) \neq f(x)\})$

The algorithm has 2 cases:

1. $x_i \in S_x$ which will return $h_{x_i}$.
   - This case will lead to 0 error
2. $x_i \notin S_x$ which will return $h^-$
   - if $h^-$ is the real hypothesis then true error will be zero and the probability will satisfy $< \delta$
   - else:

$$P(\{S_x \mid L_{(D,f)}(h_s) > \epsilon\}) \leq P(\{x_i \notin S_x\})$$

$$P(\{x_i \notin S_x\}) = (1 - P(\{x_i\}))^m$$ ← probability that $x_i$ is not sampled in $m$ trys

$f$ and $h^-$ are different at $x_i$ only

$$\epsilon < L_{(D,f)}(h^-) = P(\{h^-(x) \neq f(x)\}) = P(\{x_i\})$$

Therefore: $(1 - P(\{x_i\}))^m < (1 - \epsilon)^m$

$$(1-\epsilon)^m \leq \delta \rightarrow \quad m \geq \left\lceil \frac{\ln(\frac{1}{\delta})}{\ln(1-\epsilon)} \right\rceil$$

which is the upper bound on the sample complexity of $H$
This shows that $H$ is PAC learnable given a large but finite $m$.

2. Create a polynomial that has value of $\geq 0$ if $f(x) = 1$
   - Should have shape of negative quadratic

$$P_S(x) = -\prod_{i=1}^{m} \left( h_S(x_i)(x - x_i)^2 + (1 - h_S(x_i)) \right)$$ ← will max at 0 when $h_S(x) = 1$