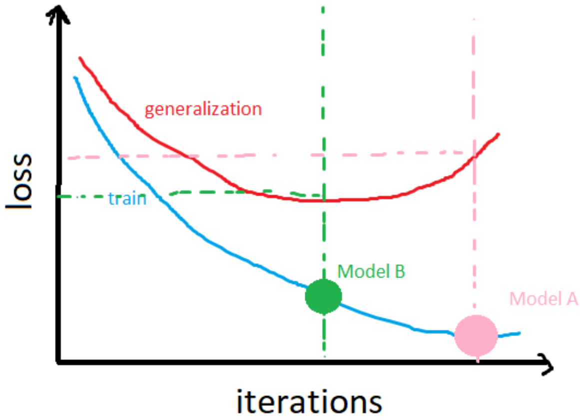
Michael Potter
UID: UID: 705644667
mlpotter124@g.uclda.edu

Homework Assignment 2
CS260: Machine Learning Algorithms

2021-02-16

Problem 1

(a) **False**: As shown by the learning curve, it is possible for model A to have smaller training error compared to model B but larger generalization error compared to B.



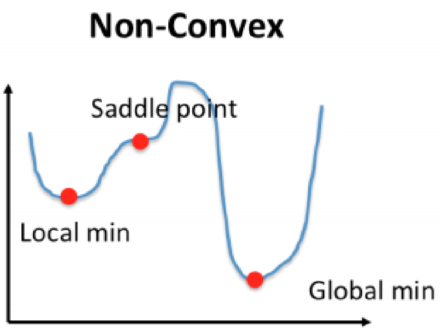
(b) **False**: The VC-dimension is just an *effective* measure of the number of parameters, and therefore is not always equal to the exact number of parameters in the model.

$$\text{Hypothesis space } H = f_t : t \in \mathbb{R}$$

$$f_t(x) = \begin{cases} +1 & \text{if } sin(tx) \geq 0 \\ -1 & \text{if } sin(tx) < 0 \end{cases}$$

This example has an infinite VC-dimension, but the functions are only parameterized by 1 parameter.

(c) **False**: A non-convex function may have local minima or saddle points (stationary points) that are not global minimum and therefore the model may not converge to a global minimum for a non-convex function.



Problem 2

The following is not a possible growth function $m_H(N)$ for a hypothesis set:

$$(2) \quad \boldsymbol{m_H(N) = 2^{\lfloor \sqrt{N} \rfloor}}$$

$m_H(N) = 2^{\lfloor \sqrt{N} \rfloor}$ cannot be a growth function of any parameterizable function as **the growth function must be a polynomial function or exponential function, but cannot be something inbetween (Lecture 6 Slide 29 and https://en.wikipedia.org/wiki/Growth_function).** On the other hand, option (1) is exponential, option (3) is a 0 order polynomial, and option (4) is a second order polynomial in N.

Problem 3

(a) The objective function is non-differentiable because of the introduction of the ℓ_1-norm, which produces sharp-non-differntiable points in the objective function as

$$\ell_1 - norm = \|\boldsymbol{w}\|_1 = \sum_i |w_i| \tag{1}$$

(b)

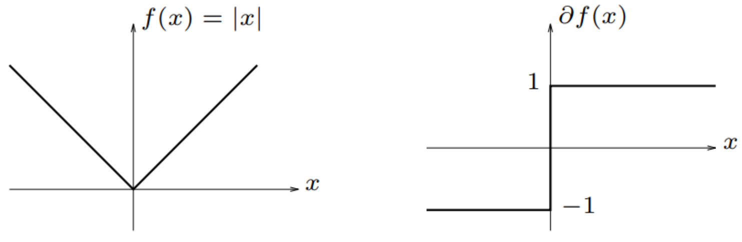
$$\boldsymbol{w}_{t+1} = \underset{\boldsymbol{w}}{\operatorname{argmin}}\, \hat{g}(\boldsymbol{w}) + \lambda \|\boldsymbol{w}\|_1$$
$$\boldsymbol{w}_{t+1} = \underset{\boldsymbol{w}}{\operatorname{argmin}}\, \frac{\eta}{2}\left\|\boldsymbol{w} - (\boldsymbol{w}_t - \frac{1}{\eta}\nabla g(\boldsymbol{w_t}))\right\|_2^2 + \lambda \|\boldsymbol{w}\|_1$$
$$\boldsymbol{z} = \boldsymbol{w}_t - \frac{1}{\eta}\nabla g(\boldsymbol{w_t})$$
$$\boldsymbol{w}_{t+1} = \underset{\boldsymbol{w}}{\operatorname{argmin}}\, \frac{\eta}{2}\|\boldsymbol{w} - \boldsymbol{z}\|_2^2 + \lambda \|\boldsymbol{w}\|_1$$

May express the argmin of $\boldsymbol{w}$ by each element w_i

$$\|\boldsymbol{w}\|_2 = \sum_i w_i^2$$
$$\|\boldsymbol{w}\|_1 = \sum_i |w_i|$$
$$\boldsymbol{w}_{t+1} = \underset{\boldsymbol{w}}{\operatorname{argmin}}\, \frac{\eta}{2}\|\boldsymbol{w} - \boldsymbol{z}\|_2^2 + \lambda \|\boldsymbol{w}\|_1 = \underset{\boldsymbol{w}}{\operatorname{argmin}} \sum_i \frac{\eta}{2}(w_i - z_i)^2 + \lambda|w_i|$$
$$w_i = \underset{w}{\operatorname{argmin}}\, \frac{\eta}{2}(w - z_i)^2 + \lambda|w|$$

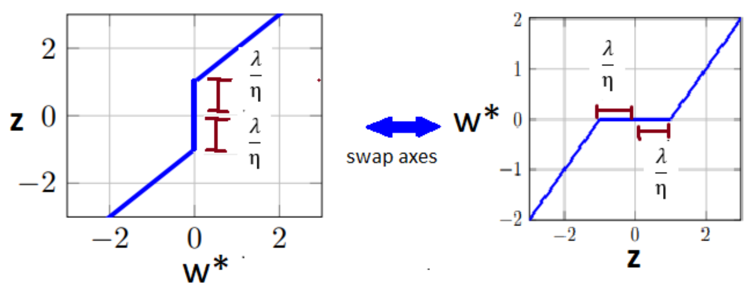
Although $|w|$ is not differentiable, it is convex and $(w-z)^2$ is convex and differentiable. Therefore, I use the fact that the minimum of a convex non-differentiable function is x^* if and only if f is subdifferentiable at x^* and $0 \in \partial f(x^*)$.

The subgradient of $|w|$ is $sgn(w)$



$$0 \in \partial f(w^*)$$
$$0 \in \eta(w^* - z) + \lambda sgn(w^*) \iff z = w^* + \frac{1}{\eta}\lambda sgn(w^*)$$

The graphical solution may be shown below, stemming from the current set of equations.



$$\boldsymbol{w}^* = sgn(\boldsymbol{z}) \odot max(|\boldsymbol{z}| - \frac{1}{\eta}\lambda, 0)$$

To analyze time complexity of one proximal gradient descent iteration, the gradient of g wrt w should be derived, and all matrix and vector time complexity calculations should be analyzed.

$$\begin{aligned}
\nabla_w g &= \boldsymbol{X^T(Xw - y)} \\
\boldsymbol{X^T X} &\rightarrow O(d^2 n) \\
\boldsymbol{(X^T X)w} &\rightarrow O(d^2) \\
\boldsymbol{X^T y} &\rightarrow O(nd) \\
sgn(vector) &\rightarrow O(d) \\
max(vector) &\rightarrow O(d) \\
O(d^2 n + nd + d) &\rightarrow O(d^2 n)
\end{aligned}$$

solution

$\boldsymbol{w_{t+1}} = sgn(\boldsymbol{w_t} - \frac{1}{\eta}\nabla_g(\boldsymbol{w_t})) \odot max(|\boldsymbol{w_t} - \frac{1}{\eta}\nabla_g(\boldsymbol{w_t})| - \frac{1}{\eta}\lambda, 0)$

Time complexity for one iteration $\approx O(nd^2)$