Lecture Notes 01/27/2021

Admin: HW1 due 01/29/2021 @ 11:30PM

HW2 will be posted by Monday 02/01/2021

01/25/2021 Summary

- We want to find the linear model $\hat{y} = w^T x + b$ that is parameterized by w
- Define a cost $l(w) = \sum_{i=1}^{n} (y_i - w^T x_i)^2 = (\bar{y} - Xw)^T (\bar{y} - Xw)$

  ground truth, prediction

- Solution: $w = (X^T X)^{-1} X^T y$

  $X^T X \rightarrow$ dimension? d x d

  d by d matrix

  $\rightarrow (X^T X)$ must be invertible

  $\rightarrow$ What if n and/or p are very large?

Batch

- process all of the data at once

Online

- process one data sample at a time

The error of each sample in the data set $i = 1, \ldots, n$

$$E_i = (y_i - w^T x_i)^2$$

[error of the ith sample]

$$E_T = \sum_{i=1}^{n} E_i$$

We are taking the derivative with respect to a vector w.

$$\nabla E_i = -2(y_i - w^T x_i)x_i \qquad w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_j \\ \vdots \\ w_d \end{bmatrix}$$

$$\nabla E_T = \sum_{i=1}^{n} \nabla E_i$$

We can use an algorithm called stochastic gradient descent to find w.

Online Gradient Descent

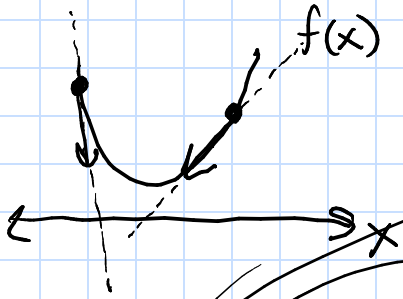
$\eta > 0$

$w^0 = N(0, \sigma)$

$$w^{t+1} = w^t - \eta \tilde{\nabla} E_n$$

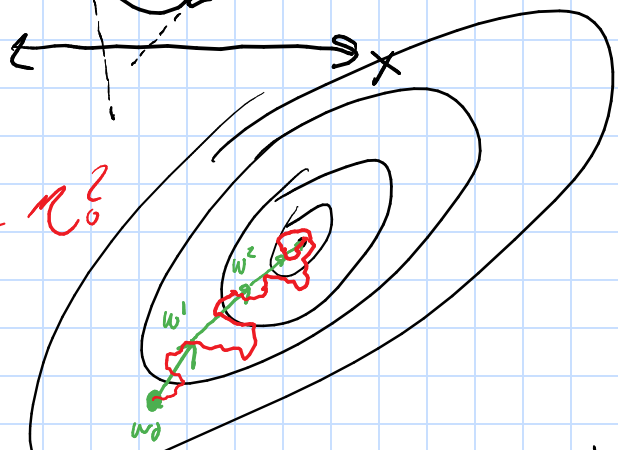
$$= w^t + \eta(y_i - w^T x_i)x_i$$

How do we choose η?

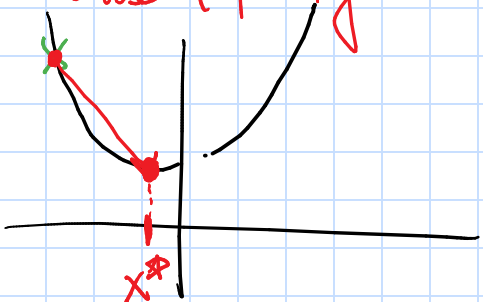


LMS
Widrow-Hoff

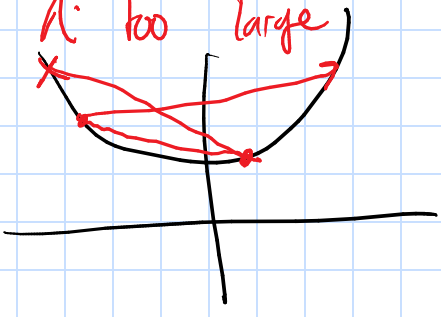
- GD



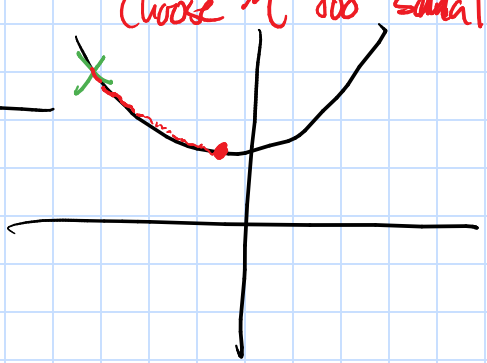
Choose η perfectly

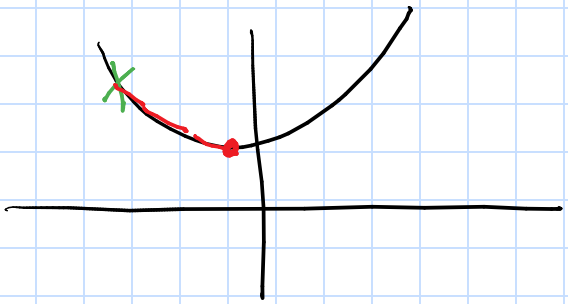


η too large



Choose η too small!





Overfitting

regularization are data independent.

We can add in regularization to our cost function that we want to minimize. We are going to add in terms to the cost function that are not data dependent

Data dependent error ★

$$w^* = \arg\min_{w \in \mathbb{R}^d} E_D(w, x, y) + \lambda E_R(w)$$

$\lambda > 0$

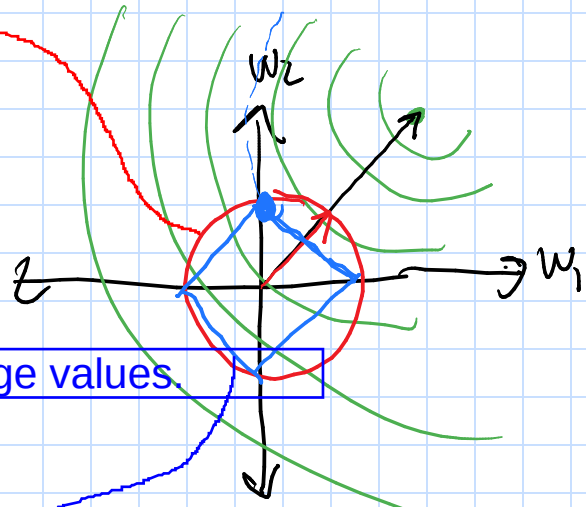
- Not data dependent
- Only uses w
- $E_R(w) \geq 0$

$$\sum_{i=1}^{n} (y_i - w^T x_i)^2 + \frac{\lambda}{2} \|w\|_2^2$$

L_2 Regularization

$$\|w\|_2^2 = w^T w, \quad E_R(w) = \frac{1}{2} w^T w$$

What is going on? $cx, w \in \mathbb{R}^2$



We basically panalizing w for taking large values.

L_1 Reg

$$\|w\|_1 = \sum_{j=1}^{d} |w_j|$$

With L1 Reg we tend to get sparse solution.