Unit 1 - Lecture 4: Linear Classification and Generalization

Objectives:

- Understanding optimization view of learning
- Apply optimization algorithms such as gradient descent, stochastic gradient descent quadratic program.

Lambda parameter:

Machine learning problems are often cast as optimization problems

Objective function = average loss + regularization

Large margin linear classification as optimization : Support vector Machine

$$J(\theta,\theta_0) = \underbrace{\frac{1}{m}\sum_{i=1}^{m} \text{loss}_i\,(\overbrace{y^i\,(\theta \cdot x^i + \theta_0)}^{\text{agreement}})}_{\text{Hinge loss}} + \frac{\lambda}{2}\|\theta\|^2$$

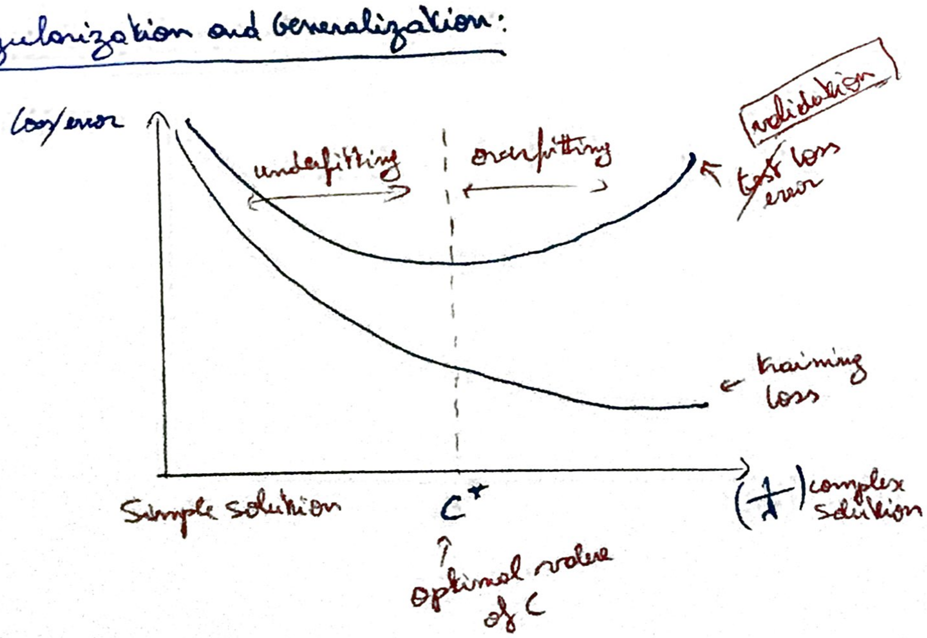
Regularization parameter

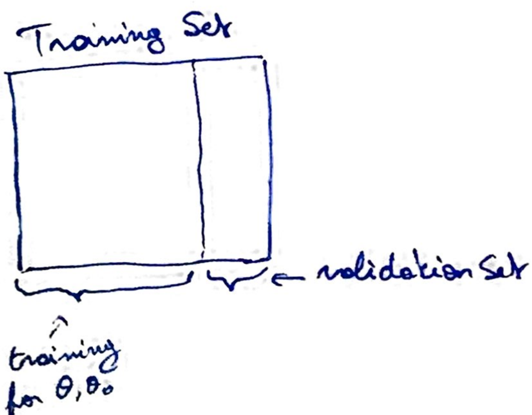
When $\lambda \searrow \implies d \searrow$, So when we change the Regula. parameter lambda
$\lambda \nearrow \implies d \nearrow$ we change the balance between these two terms

Given $d: ax + by + c = 0$ and $P = (x_0, y_0)$

$$d = \frac{a.x_0 + b.y_0 + c}{\sqrt{a^2+b^2}}$$

Regularization and Generalization:



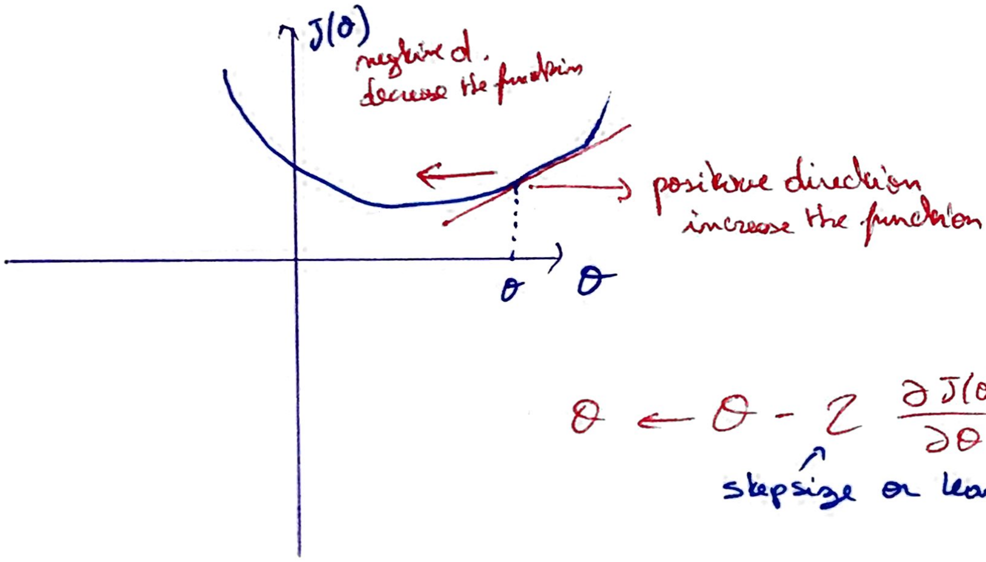


We evaluate C and find it by evaluating θ, θ_0 in the validation set

Optimization Algorithms to minimize the objective function:

1. gradient descent
2. Stochastic gradient descent
3. quadratic program.

Gradient Descent:



$$\theta \leftarrow \theta - \eta \frac{\partial J(\theta)}{\partial \theta}$$

stepsize or learning Rate

My gradient descent update rule here, is getting a new parameter value θ in term of the old one and moving in the negative direction of the derivative of the gradient.

$$\theta \leftarrow \theta - \eta \nabla J(\theta) \quad \text{with} \quad \nabla J(\theta) = \begin{bmatrix} \frac{\partial J(\theta)}{\partial \theta_1} \\ \vdots \\ \frac{\partial J(\theta)}{\partial \theta_d} \end{bmatrix}$$

Stochastic gradient descent: SGD

$$J(\theta, \theta_0) = \underbrace{\frac{1}{m} \sum_{i=1}^{m} J_i(\theta)}$$

We sample a training example i at random and we perform a gradient descent update with respect to the selected sampled term.

$$\theta \leftarrow \theta - \eta \nabla J_i(\theta) \qquad \eta_t \longrightarrow 0$$

$$\theta \leftarrow \theta - \eta \nabla_\theta \left[\text{Loss}(y^i \theta . x^i) + \frac{\lambda}{2} \|\theta\|^2 \right]$$

$$\theta \leftarrow \theta - \eta \left[\begin{cases} 0, & \text{Loss} = 0 \\ -y^i x^i, & \text{Loss} \neq 0 \end{cases} + \lambda \theta \right]$$