## Regularization — Part 1(of 2)

By the end of this lecture, you will understand notion of regularization as sensitivity to input and how to change a loss function to its regularized version. by the end of this recture, you will understand how to change a loss function to its regularized version and how to change a loss function to its regularized version.

- Consider squared error loss function
- $L(w) = (y w x)^2$
- How do you know how sensitive is loss function with respect to input?

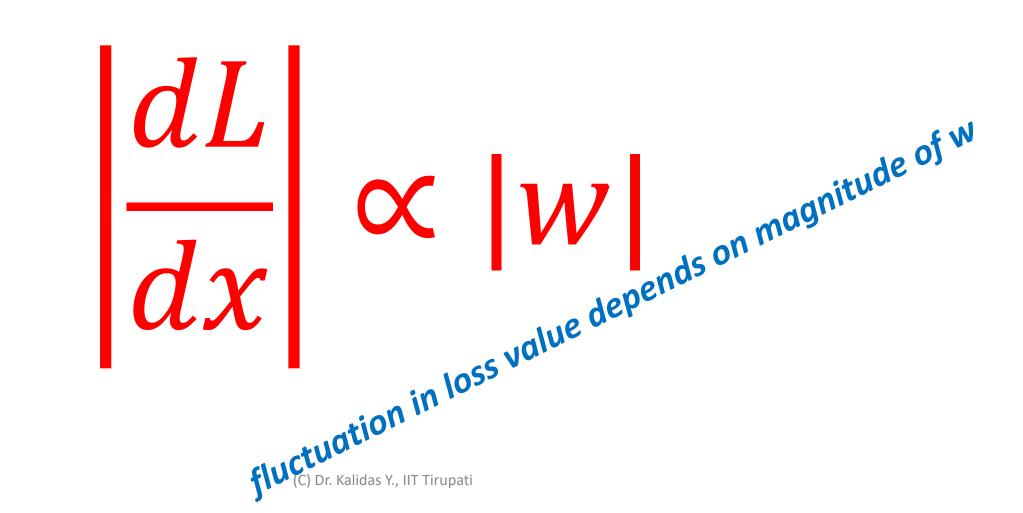
- Consider squared error loss function
- $L(w) = (y w x)^2$
- How do you know how sensitive is loss function with respect to input?
- Consider x as variable and fix w for a while

- Consider squared error loss function
- $L(w) = (y w x)^2$
- How do you know how sensitive is loss function with respect to input?
- Consider x as variable and fix w for a while

$$L(x) = (y - w x)^2$$

$$\bullet \frac{dL}{dx} = 2 * (y - w x) * -w$$

- Consider squared error loss function
- $L(w) = (y w x)^2$
- How do you know how sensitive is loss function with respect to input?
- Consider x as variable and fix w for a while
- $L(x) = (y w x)^2$
- $\bullet \frac{dL}{dx} = 2 * (y w x) * -w$
- Magnitude of change is  $\left|\frac{dL}{dx}\right| \propto |w|*(w|x|-y)| = |w|*\epsilon_{max}$  (if we assume difference between prediction and actual is a maximum of  $\epsilon_{max}$ )



## 40) key phrase "Regularized Loss function" or "Regularization"

Loss function + Weight magnitude component

#### Consider a multi variate loss function

•Lasso regularization  $-|\nabla L(w)|_1$ 

•Ridge regularization –  $|\nabla L(w)|_2$ 

#### 41) key phrase "Lasso loss function"

• 
$$L(w) = \sum_{i=1}^{i=N} \left( \left( \sum_{j=1}^{j=k} w_{i,j} * x_{i,j} \right) - y_i \right)^2 + \sum_{j=1}^{j=k} |w_j|$$

- Correspondingly it is called,
  - "L1 regularization"
  - "lasso regularization"
  - "lasso regression"

#### 42) key phrase Ridge loss function

• 
$$L(w) = \sum_{i=1}^{i=N} \left( \left( \sum_{j=1}^{j=k} w_{i,j} * x_{i,j} \right) - y_i \right)^2 + \sum_{j=1}^{j=k} w_j^2$$

- Correspondingly it is called,
  - "L2 regularization"
  - "ridge regression"

#### Lasso for feature elimination or feature selection

Property of interest	Without regularization	Lasso regression	Ridge regression
After Thousands of iterations of the solver  Final magnitude of the w <sub>j</sub> values	Can't say – each weight parameter's magnitude may be high or low	Magnitude of some of the weights reduces and gets close to zero	Magnitudes only reduce, but not as much as in lasso
		Let a weight be 0.1  Because direct magnitude is used, its contribution is 0.1.  So, further iterations reduce it beyond 0.1, to become 0.01 or so.	Let a weight be 0.1.  Because square of magnitude is used (w <sub>j</sub> <sup>2</sup> ), its contribution is only 0.01.  So, further iterations may not reduce it beyond 0.1
Feature elimination	NO (C) Dr. Kalid	YES as Y., IIT Tirupati	NO