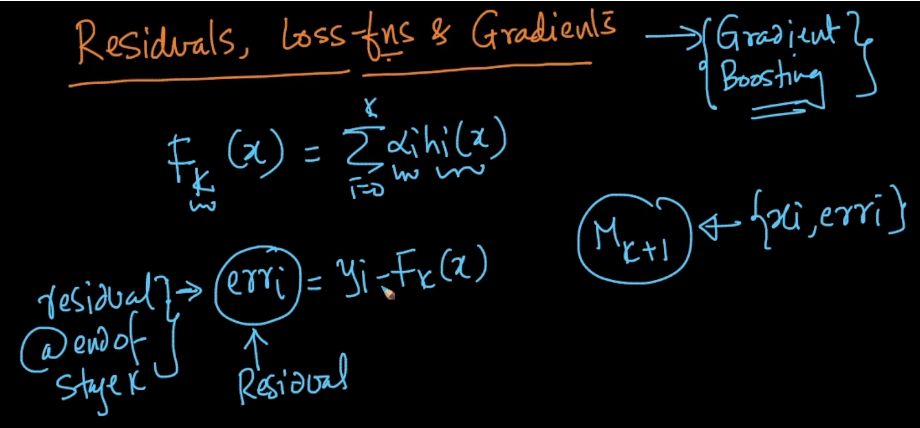
**Residuals, Loss functions and gradients**

Below image shows Fk(x) , residual at end of style k (erri)

And model Mk+1 is trained by {xi , erri}



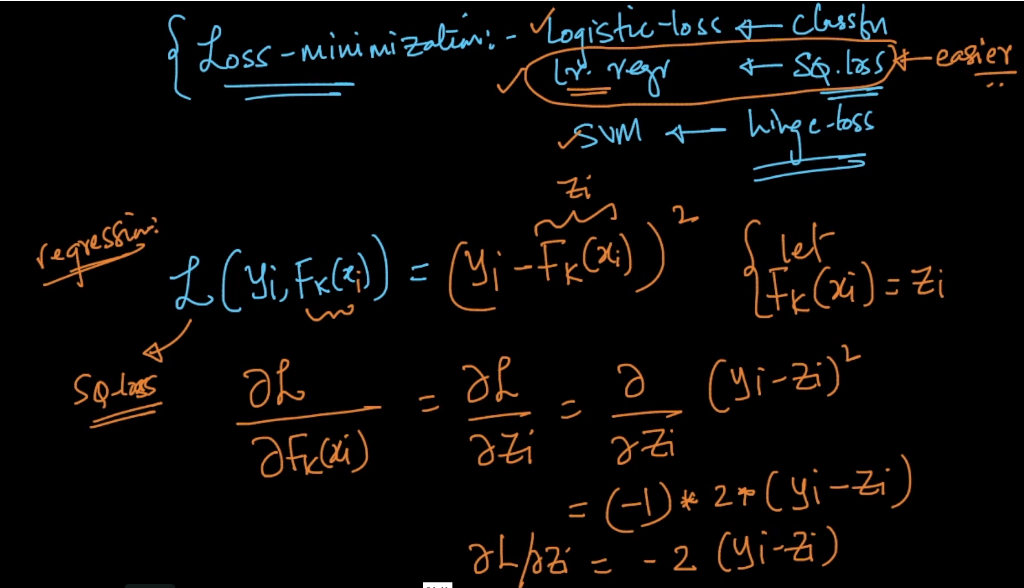
Loss-minimization :

For classification we have logistic loss,

For regression we have squared loss

For svm we have hinge loss

In below image it shows regression loss L(yi , Fk(xi)) and then took derivative of loss function L with respect to Model at stage k Fk(xi)



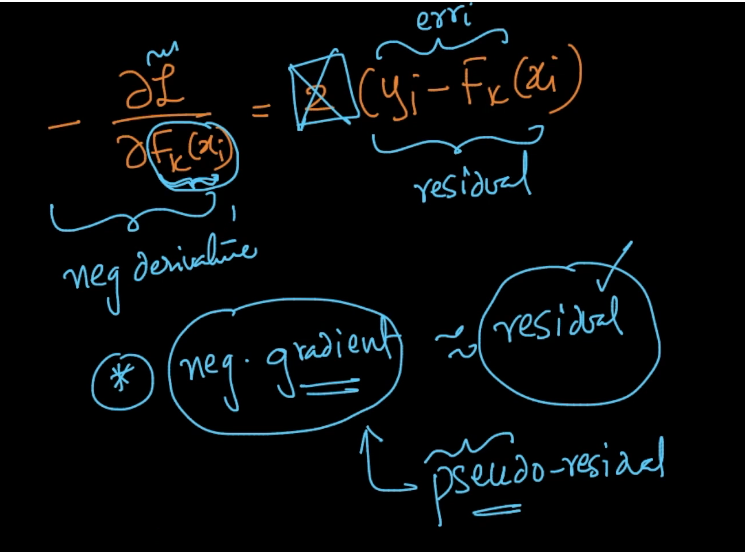
So after derivation we got residual error as yi – Fk(xi)

And the derivative is neg derivative because we multiply both sides by –

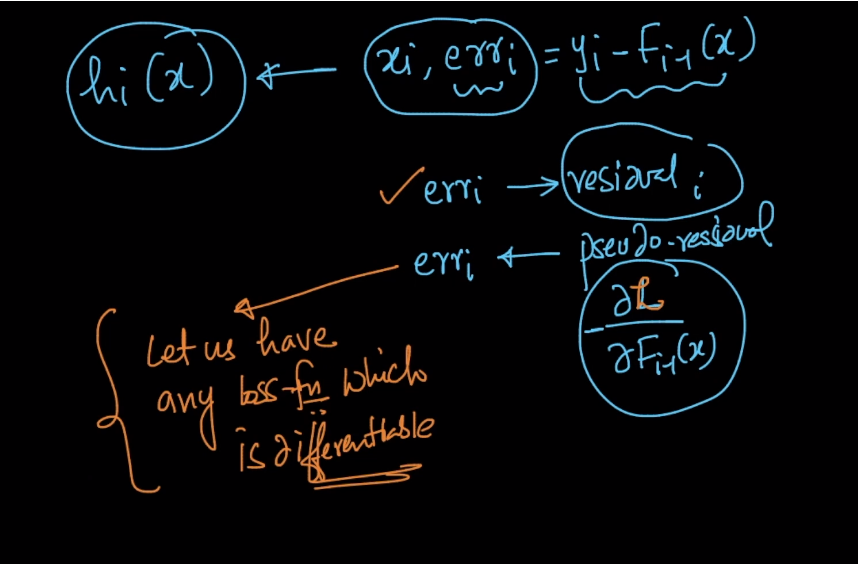
Derivative is also called gradient, so what we got after derivative is

Neg. gradient/derivative is approximately equal to residual

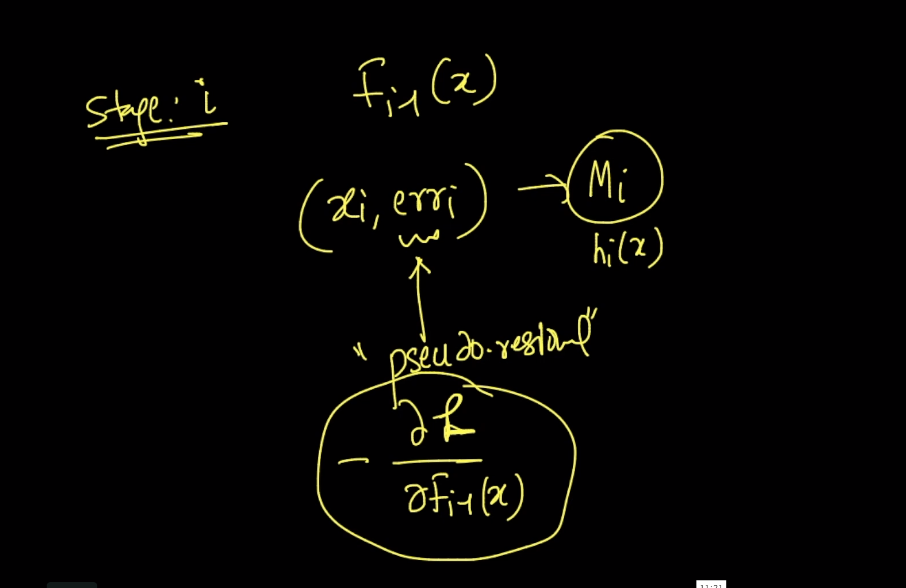
And this neg. gradient/derivative is pseudo residual



So why we use pseudo residual when we have residual because pseudo residual minimize any loss function which is differentiable.

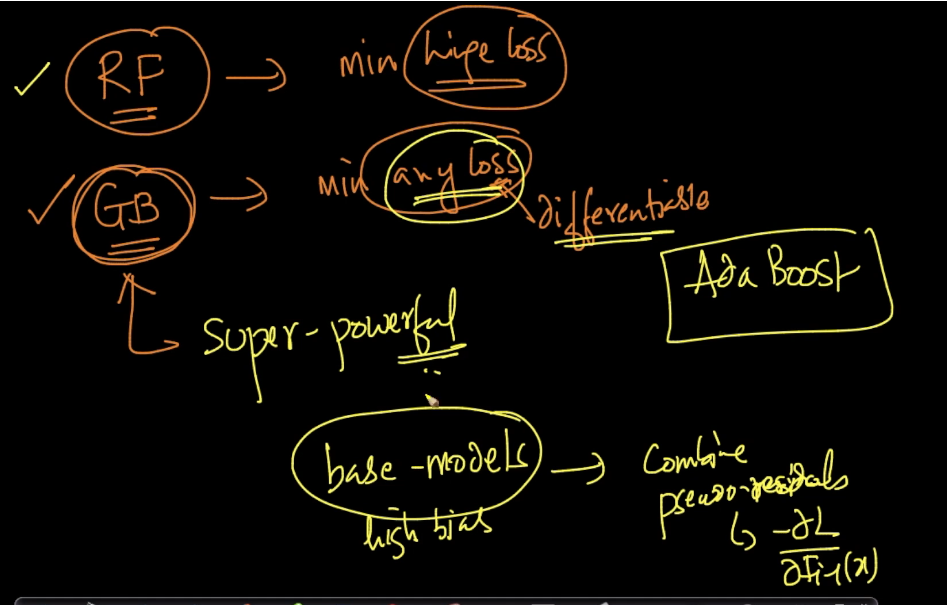


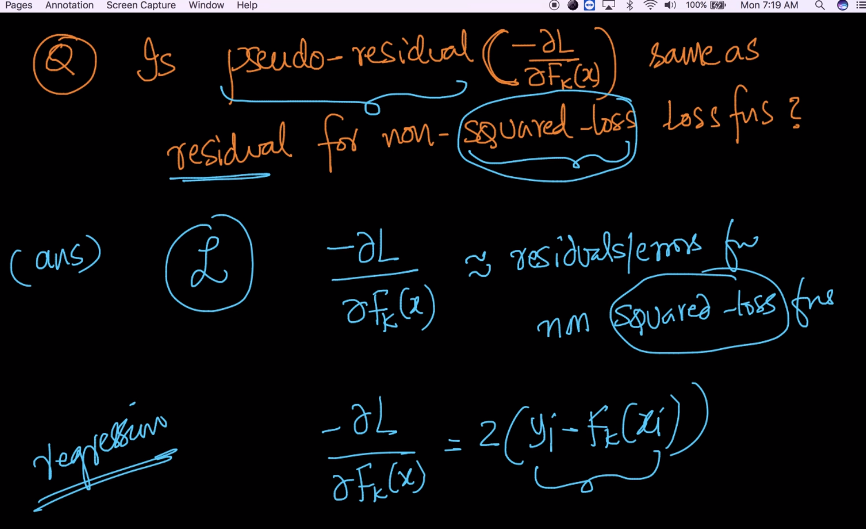
So the algorithm looks like at any stage i we train model Mi(hi(x)) by (xi , errori) where errori is pseudo residual del\_l / del\_Fi-1(x)



So what is the use of gradient boost is that it minimize any loss which is differentiable whereas RF only minimize only entropy but it cannot minimize any other loss like hinge loss therefore we use gradient boost

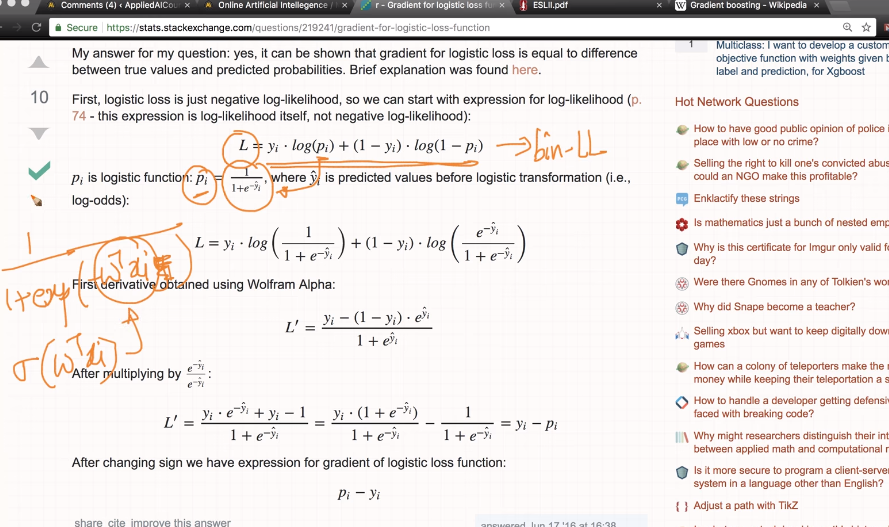
There is also one more type of boosting ADA boost in this it also minimize only 1 type of loss unlike Gradient boost therefore gradient boost is super powerful technique

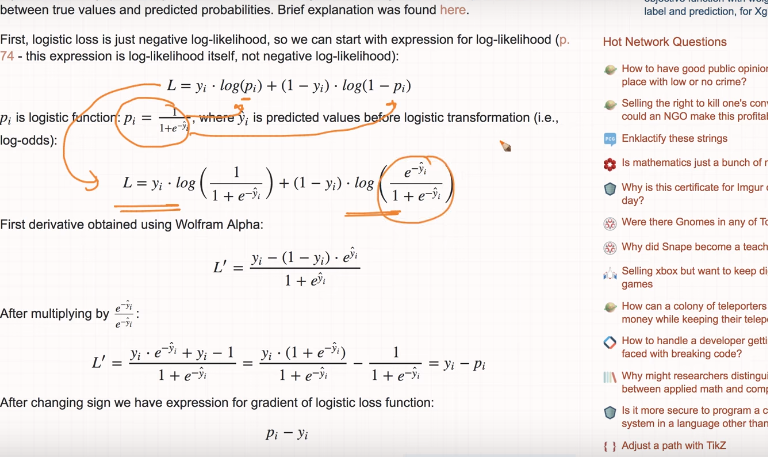


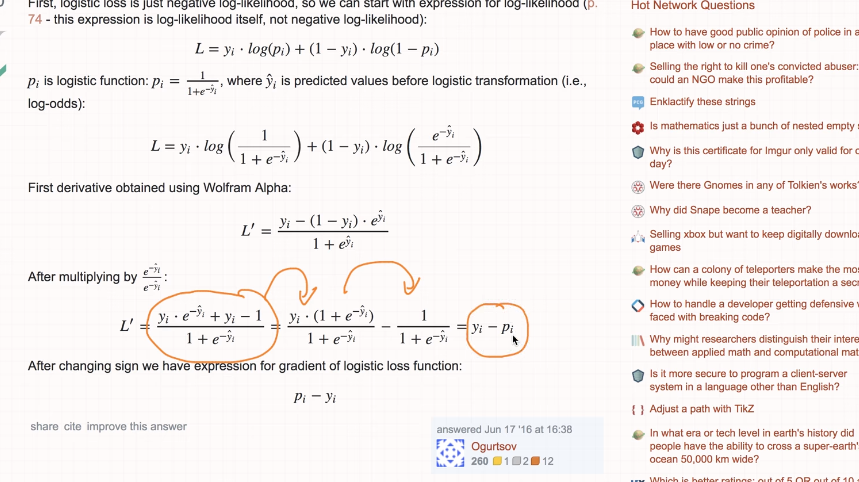


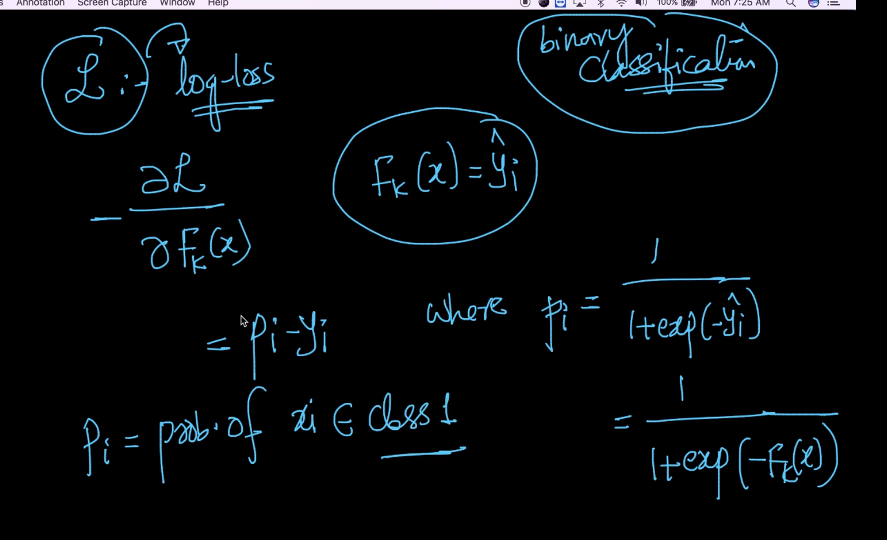
Above image show pseudo residual for regression (squared loss function) now we find pseudo residual for classification (non-squared loss function)

In below image we use log loss error and pi is logistic function i.e probability of xi belongs to class 1

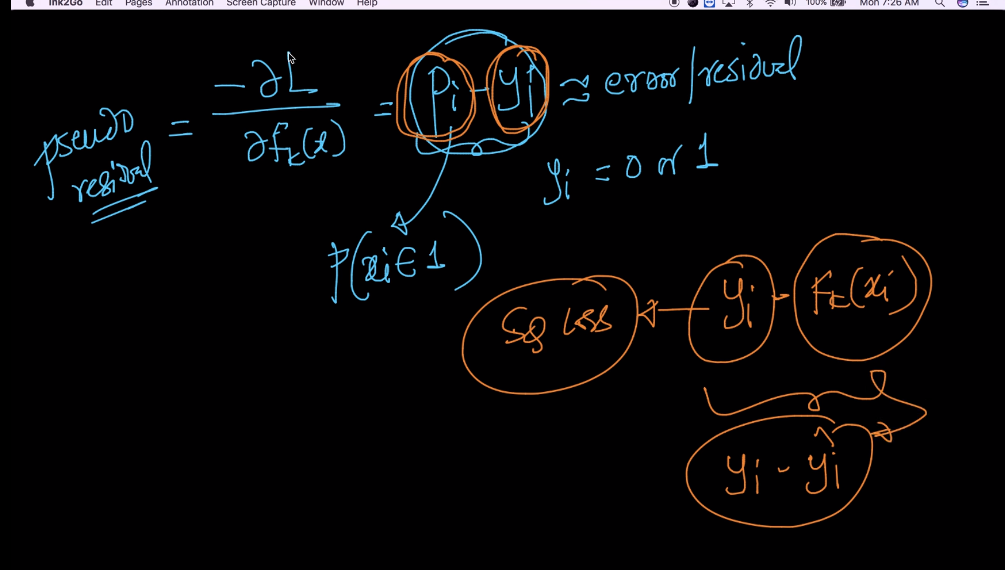


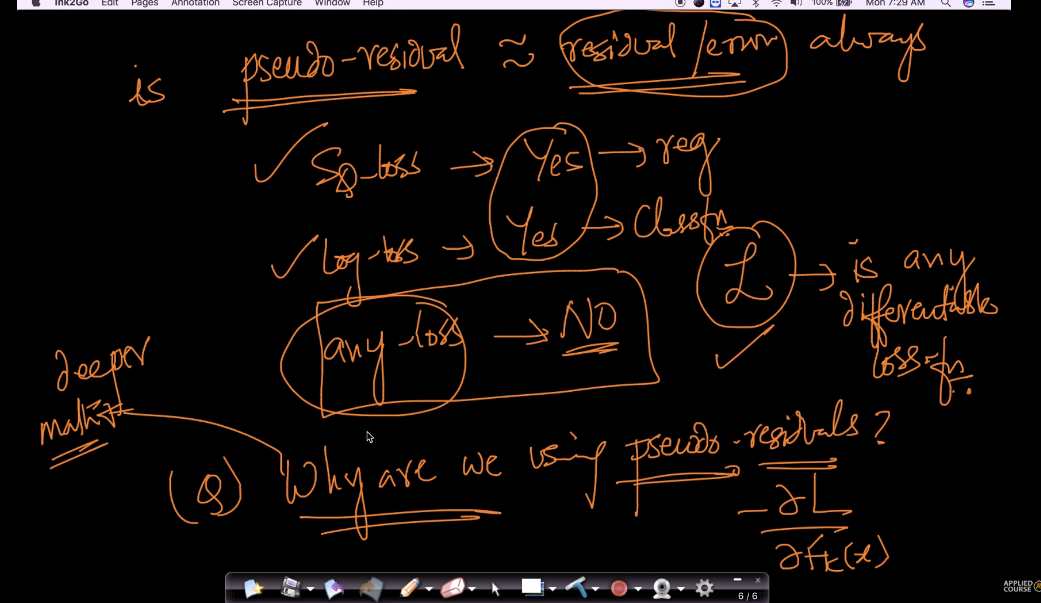






Therefore pseudo residual of non squared function is probability of xi belongs to class 1 which is approximately same as error or residual which is some what same as of squared function





Link :

### <https://youtu.be/qEZvOS2caCg>

<https://stats.stackexchange.com/questions/219241/gradient-for-logistic-loss-function>

### The mathematical/rigorous argument for using pseudo-residuals:

### <https://youtu.be/1o0yd6eMmA0>

<https://en.wikipedia.org/wiki/Gradient_boosting#Algorithm>

comments :

