Jerry Chen

CS541

Artificial Intelligence

Prof. Shen

HW3

Gradient Calculation

Sigmoid: $F(w) = \frac{1}{1+e^{-x*w}}$, Gradient: $\frac{dF(w)}{dw} = \frac{xe^{-xw}}{(1+e^{-xw})^2}$

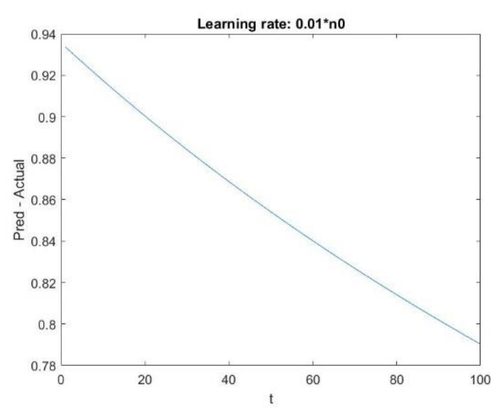
Logistic Loss: $F(w) = \log(1 + e^{-yx*w})$, Gradient: $-\frac{xye^{-xy*w}}{1+e^{-xy*w}}, y - \frac{1}{1+e^{-xw}}$

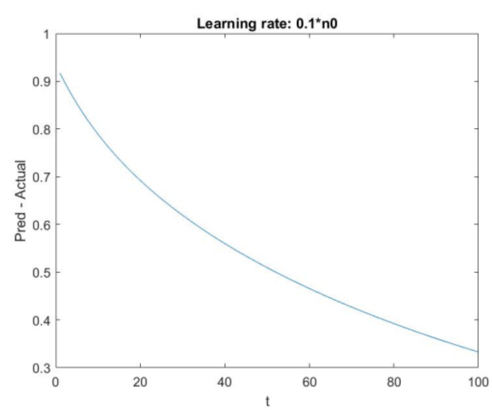
Linear Regression

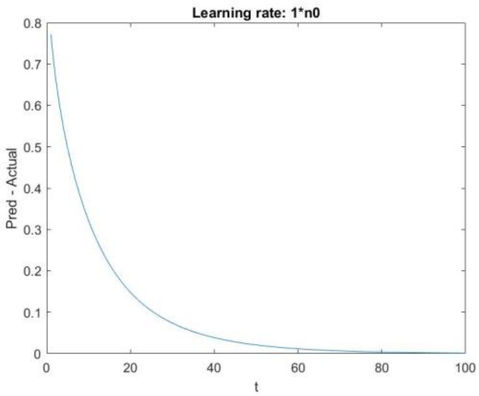
1. $\min w \in R^d, F(w) = \frac{1}{2}\left||y - Xw|\right|_2^2$, Gradient: $\frac{dF(w)}{dw} = X^T Xw - X^T y$, Hessian: $\frac{d^2F(w)}{dw^2} = X^T X$

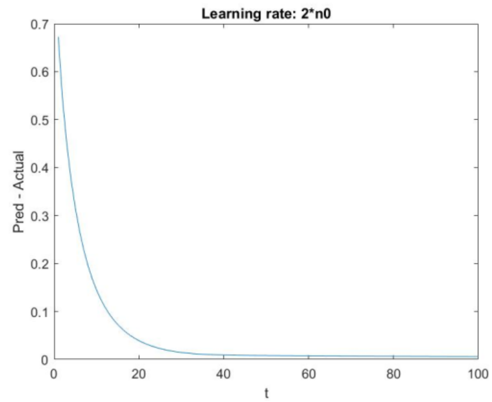
   1 is a convex program because the sum of convex functions is convex and the function x^2 is convex. The composition with an affine function is also convex, therefore $\frac{1}{2}\sum_{i=1}^{n}|y_i - Xw|^2$ is a convex program.

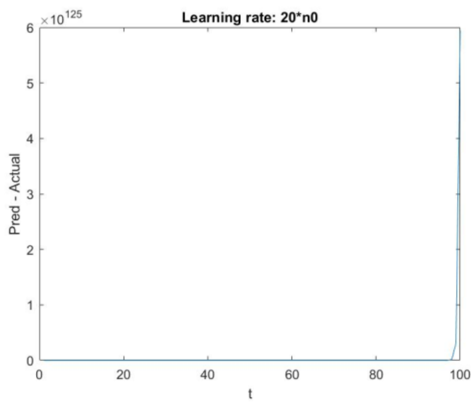
2. We stick with the least squares formulation because it is more efficient to compute and gives us the essential information to compute error (squaring out the negatives).
3. F(w) is strongly convex when the eigenvalues of the hessian of F(w) can be lower bounded by a value α. If the minimum eigenvalue becomes infinitely small (goes towards negative infinity), then the function is not strongly convex. F(w) is not strongly convex when d>n.
4. N=100, D=40

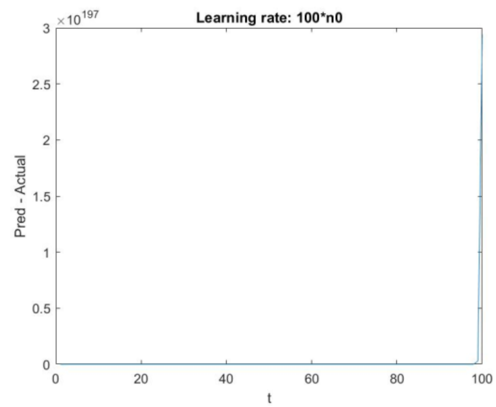






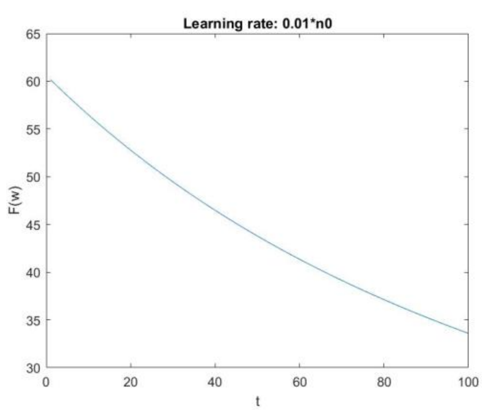


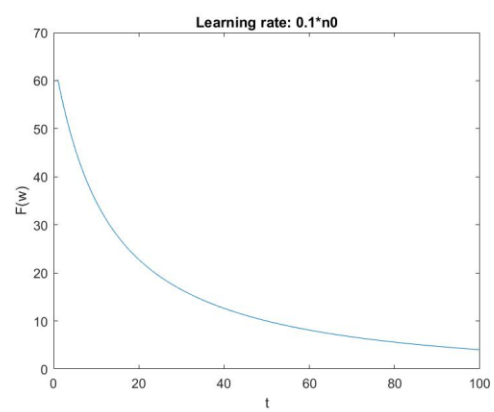


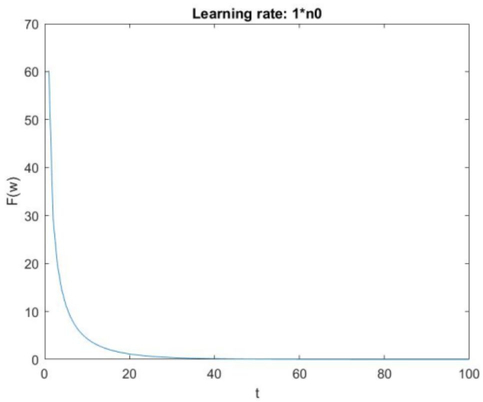


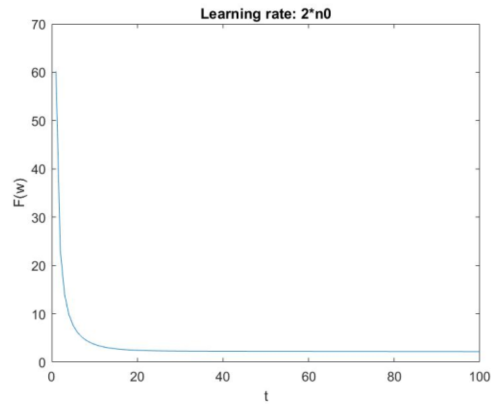
The best learning rates seem to be at 1*n0 and 2*n0, the smaller learning rates are way too slow while the larger learning rates just oversteps and increases the error rather than minimize.

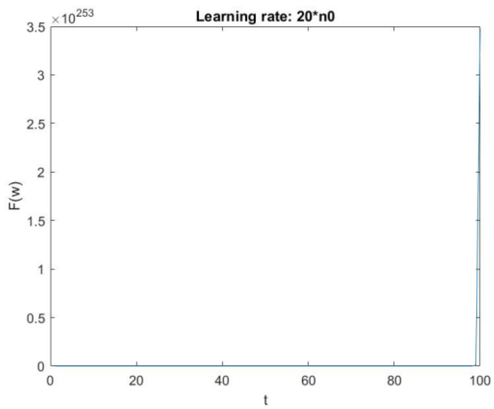
5. When trying to calculate the solution w^*, it turns out that the eigenvalues of the hessian matrix are no longer lower bounded by a value α as the minimum eigenvalue has gone to negative infinity. This means that the function is no longer strongly convex, but rather just convex. We can still apply GD, but w_{opt} will not be as easy to identify.

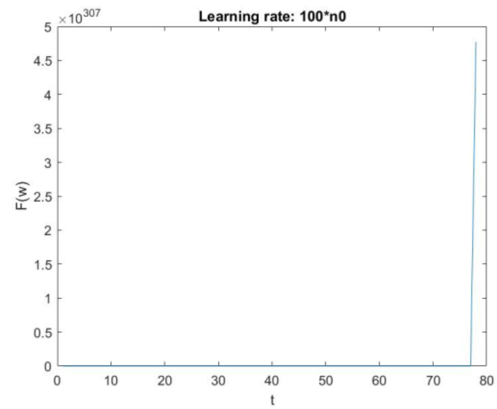












Similar to question 4, we get the best results with the learning rate of 1*n0 with the smaller values taking too long to converge to 0 and the larger values going the opposite direction instead.