CS 541 Artifical Intelligence: Homework 3

Tarquin Bennett

November 8, 2020

1 Gradient Calculations

$$F(w) = \frac{1}{(1 + e^{-x \cdot w})}$$

$$\nabla F(w) = \frac{\frac{d}{dw}(e^{-x \cdot w} + 1)}{(e^{-x \cdot w} + 1)^2} = -\frac{e^{-x \cdot w} * \frac{d}{dw}(-x \cdot w)}{(e^{-x \cdot w} + 1)^2} = \frac{e^{x \cdot w} \cdot x}{(e^{x \cdot w} + 1)^2}$$

$$F(w) = log(1 + e^{-yx \cdot w})$$

$$\nabla F(w) = \frac{1}{1 + e^{-yx \cdot w}} * \frac{d}{dw}(e^{-yx \cdot w} + 1)$$

$$= \frac{1}{1 + e^{-yx \cdot w}} * e^{-yx \cdot w} * \frac{d}{dw}(-yx \cdot w)$$

$$= \frac{1}{1 + e^{-yx \cdot w}} * e^{-yx \cdot w} * -yx$$

$$\nabla F(w) = \frac{-yxe^{yx \cdot w}}{1 + e^{yx \cdot w}}$$

2 Linear Regression

1.)

$$F(w) = \frac{1}{2}||y - Xw||_2^2$$

$$||y - Xw||_2^2 = (y - Xw)^T(y - Xw) = y^Ty - (Xw)^Ty - y^TXw + w^TX^TXw$$

$$= y^Ty - 2y^TXw + w^TX^TXw = y^Ty - 2(X^Ty)^Tw + w^TX^TXw$$

$$\nabla F(w) = \frac{d}{dw}(\frac{1}{2}||y - Xw||_2^2) = \frac{1}{2} * \frac{d}{dw}(||y - Xw||_2^2)$$

$$= \frac{1}{2} * \frac{d}{dw}(y^Ty - 2(X^Ty)^Tw + w^TX^TXw)$$

$$= \frac{1}{2}(2(X^TXw - X^Ty)) = X^TXw - X^Ty$$

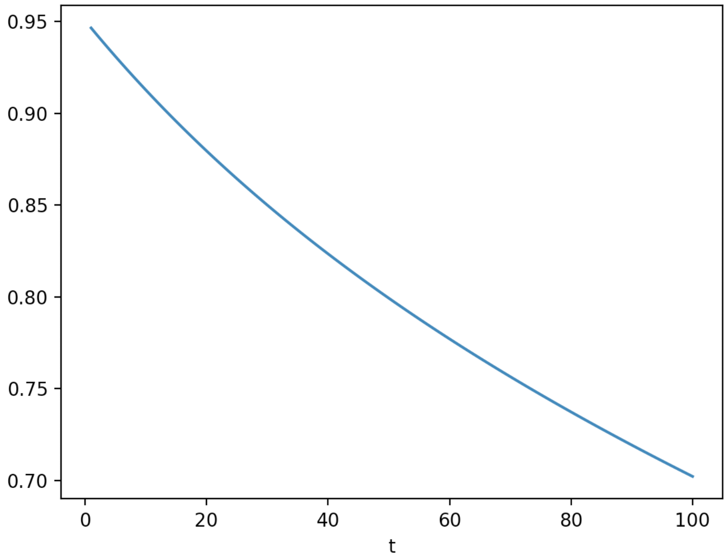
$$H(F(w)) = \nabla^2 F(w) = X^T X$$

Since the second gradient is positive definite, and is greater than 0, the program is convex.

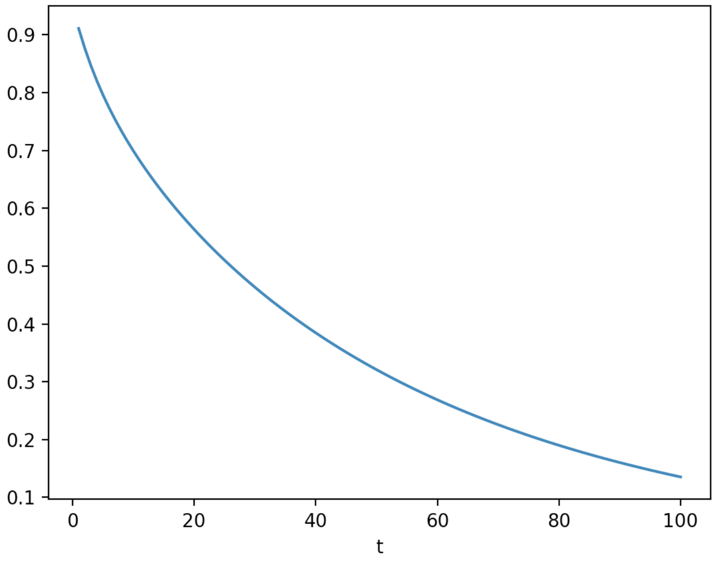
2.) We stick with least-squares because compared to the formula to the 100 power it allows us to get closer to the correct estimate faster since the new formula will generate big numbers.

3.) A condition of X such that F(w) is strongly-convex is the eigenvalues of its Hessian are positive. To make it not strongly-convex the eigenvalues would have to be singular cause then the second derivative test would be inconclusive.

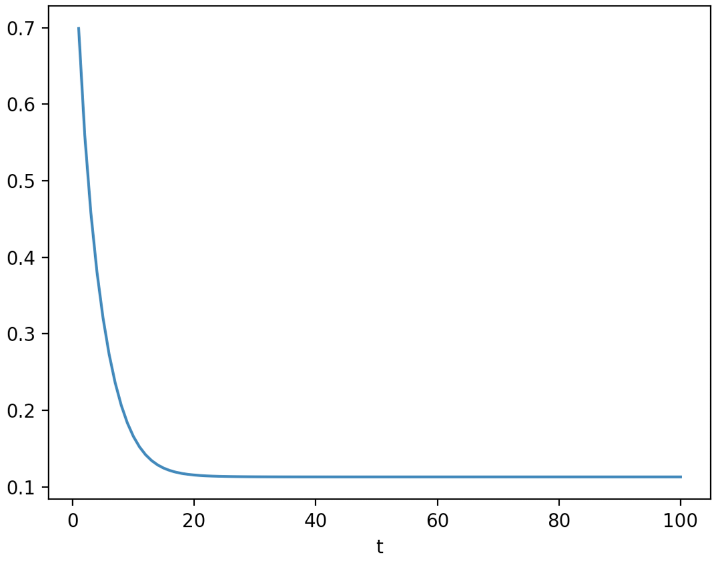
4.) Plot for $0.01\eta_0$



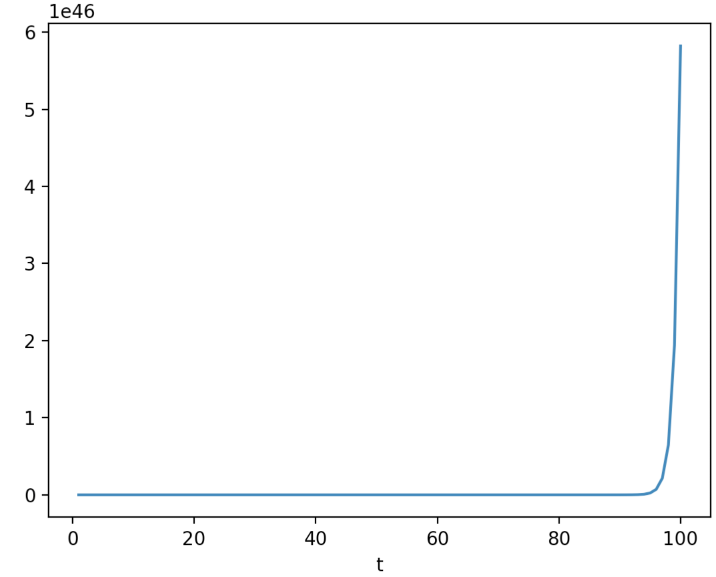
Plot for $0.1\eta_0$



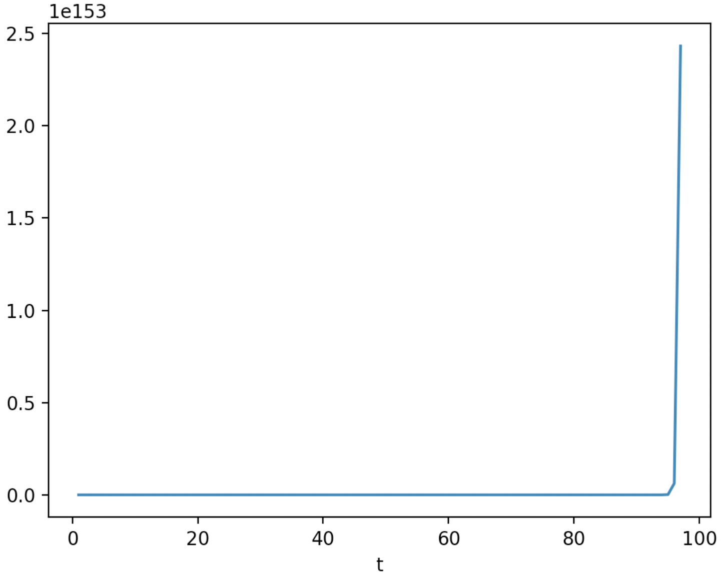
Plot for η_0



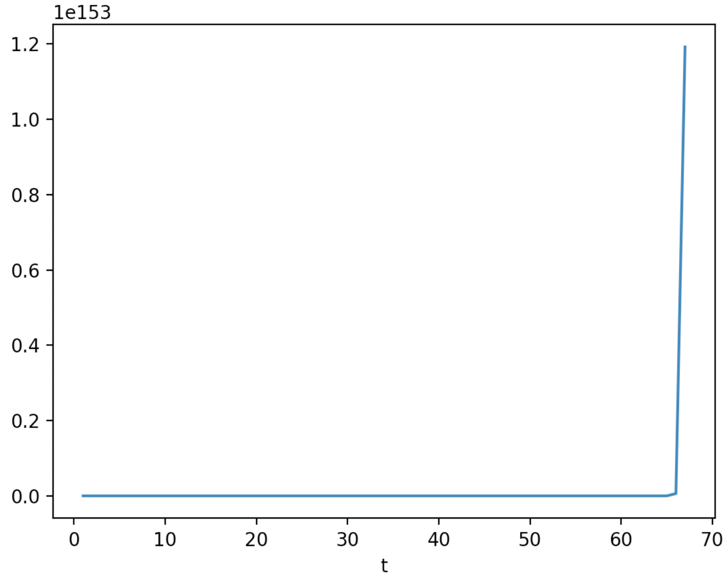
Plot for $2\eta_0$



Plot for $20\eta_0$

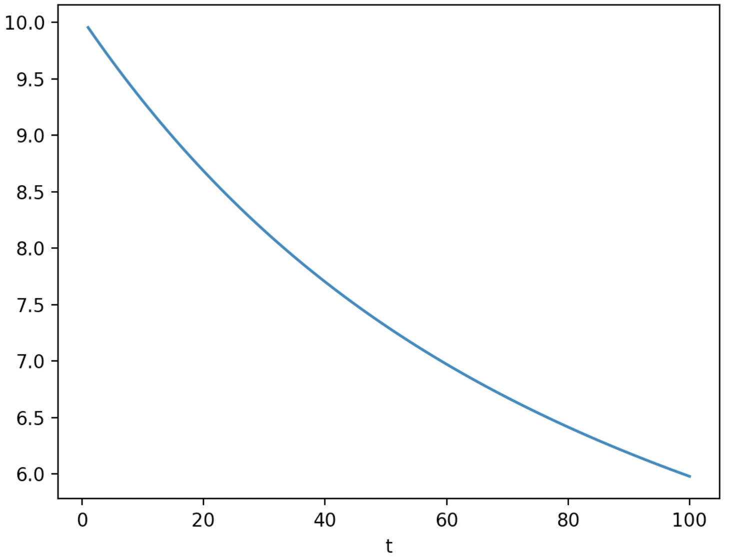


Plot for $100\eta_0$

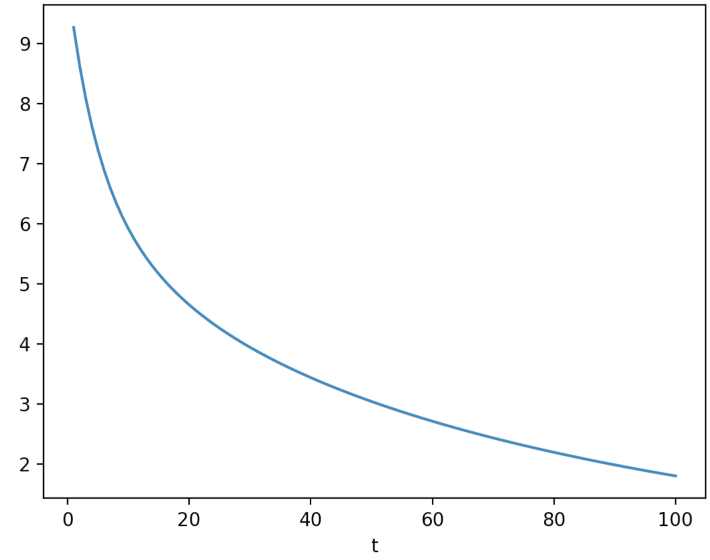


As we increase the learning rate to L and then surpass it, the graphs flip because the the orignal inequailty does not hold anymore. The best learning rate was L which makes sense fromt the inequailty.

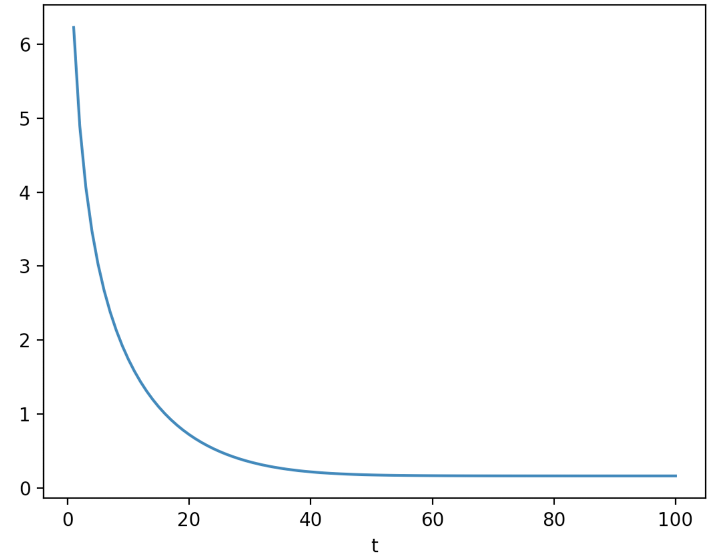
5.) plot for $0.01\eta_0$



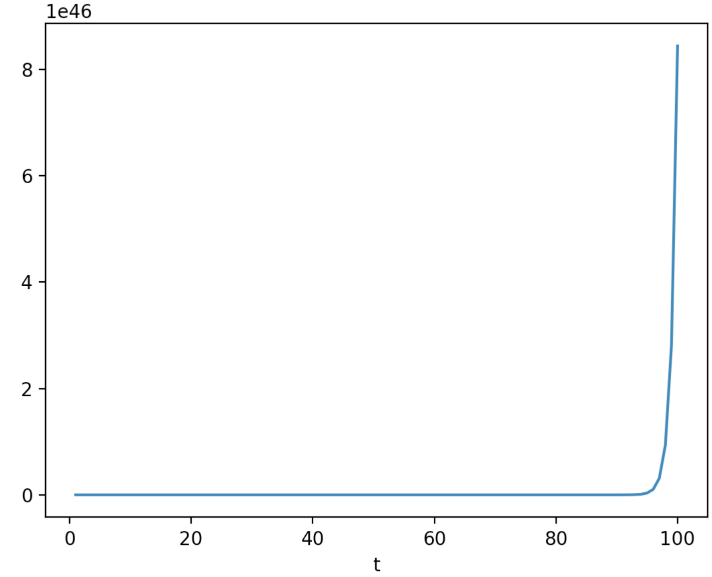
plot for $0.1\eta_0$



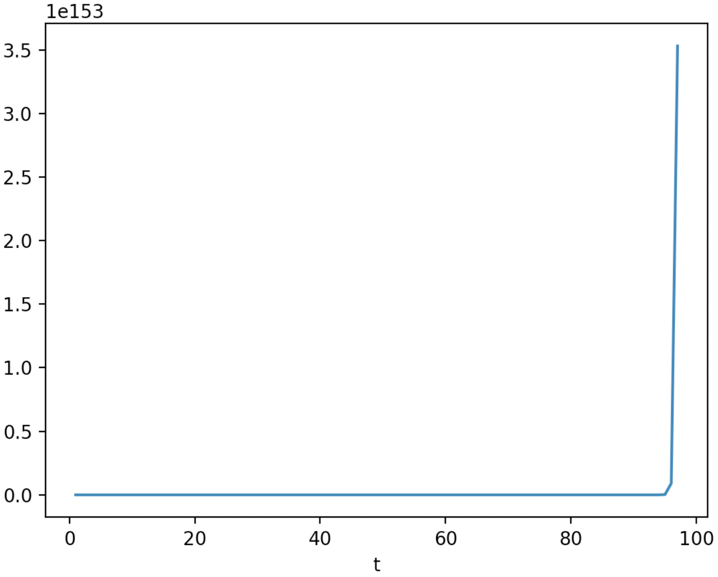
plot for η_0



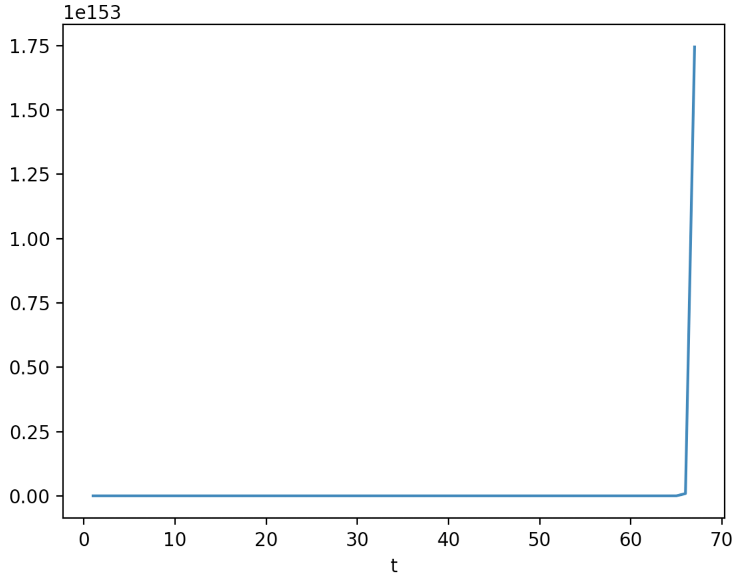
plot for $2\eta_0$



plot for $20\eta_0$



plot for $100\eta_0$



When you calculate the closed-form solution to w* the values are larger and no longer between -1 and 1. However we are still able to GD. What ends up happening is the learning rate is not as good as before. This could be due to the eigenvalues being higher therefore making the the learning rate lower. This causes the GD to take longer to reach the leveling out stage.