constant and the other parameters varied. From using the best hyper-parameters, where 10 epochs for each data set was chosen. The test set `astro/original/test` scored a classification accuracy of 0.90625, while the scaled data did much worse and `astro/scaled/test` had a classification accuracy of 0.50000. These results agree pretty well with the last homework assignment which is reassuring on the implemented algorithm working correctly. Figure 4 shows the negative log likelihood function of each of the test files at the end of each epoch.

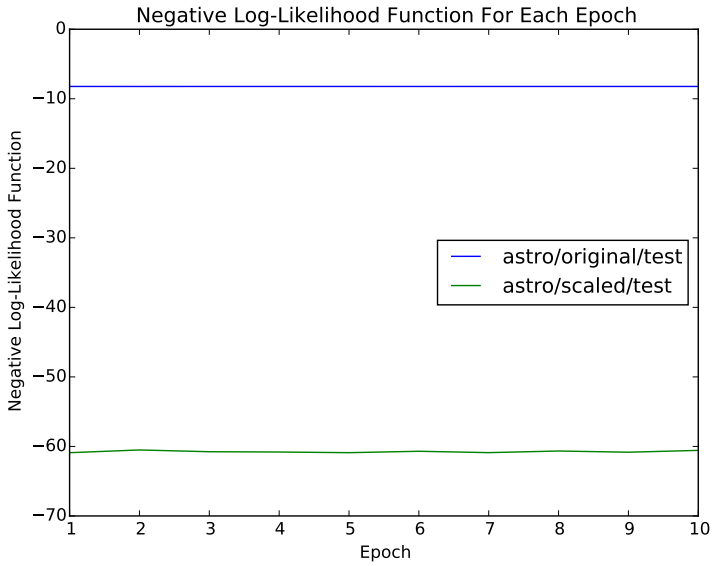

Figure 1: Negative Log Likelihood Function of $J(\mathbf{w})$

As mentioned in previous homeworks, you may use any programming language for your implementation. Upload your code along with a script so the TAs can run your solution in the CADE environment.

4 HAPPY HOLIDAYS Extra Credit

25pts You've seen stochastic gradient decent applied to logistic regression. Now we ask why this is a viable strategy for optimizing this objective. Prove that SGD will find the optimal value for this function. This can be done by demonstrating that the objective is convex. There are many ways to prove this. One of the most straightforward ways to show this is demonstrate that the Hessian is positive-semidefninite.

1. [5 points] Find the Gradient of:

$$\sum_{i=1}^{m} \log\left(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i)\right) + \frac{1}{\sigma^2}\mathbf{w}^T\mathbf{w}$$

- The gradient of any function is defined as

$$\vec{\nabla} f(\mathbf{x}) = \sum_{i=1}^{\mathcal{D}} \frac{\partial f(\mathbf{x})}{\partial x_i} \hat{e}_i \tag{47}$$

where $\mathcal{D}$ is the dimensionality/number of dependent variables in $f(\mathbf{x})$ and $\hat{e}_i$ is the normal vector for that coordinate – a value of 1 for all dimensions in cartesian space. This requires a derivative over $\mathbf{x}$ and $\mathbf{w}$ which is

$$\vec{\nabla} f(\mathbf{x}_i) = \sum_{i=1}^{\mathcal{D}} \frac{\partial f(\mathbf{x}_i)}{\partial \mathbf{w}} + \frac{\partial f(\mathbf{x}_i)}{\partial \mathbf{x}_i} \tag{48}$$

We can rewrite our equation $f(\mathbf{x}_i)$ to make it easier to differentiate

$$f(\mathbf{x}_i) = \sum_{i=1}^{n} \log\left(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i)\right) + \frac{||\mathbf{w}||^2}{\sigma^2} \tag{49}$$

where the derivative of a sum of a function is the sum of the derivative of that function, so the differentials can be brought inside the sum and the respected differentials can be taken and the partial derivative outside the braces denotes that segment representing that differential of $f(\mathbf{x}_i)$

$$\vec{\nabla} f(\mathbf{x}_i) = \left\{ \sum_{i=1}^{n} \frac{-y_i \mathbf{x}_i}{\exp(y_i \mathbf{w}^T \mathbf{x} + 1)} + \frac{2\,||\mathbf{w}||}{\sigma^2} \right\}_{\frac{\partial f}{\partial \mathbf{w}}} + \left\{ \sum_{i=1}^{n} \frac{-y_i \mathbf{w}^T}{\exp(y_i \mathbf{w}^T \mathbf{x}) + 1} \right\}_{\frac{\partial f}{\partial \mathbf{x}}} \tag{50}$$

2. [5 points] Find the Hessian of (a).

- The Hessian Matrix is defined as

$$\boldsymbol{\mathcal{H}}\left(f(\mathbf{x}_i)\right) = \begin{bmatrix} \frac{\partial^2 f}{\partial \mathbf{x}_i^2} & \frac{\partial^2 f}{\partial \mathbf{x}_i \partial \mathbf{w}} \\ \frac{\partial^2 f}{\partial \mathbf{w} \partial \mathbf{x}_i} & \frac{\partial^2 f}{\partial \mathbf{w}^2} \end{bmatrix} \tag{51}$$

$$\frac{\partial^2 f(\mathbf{x}_i)}{\partial \mathbf{x}_i^2} = \sum_{i=1}^{n} \frac{||\mathbf{w}||^2 y_i^2 e^{\varsigma_i}}{\left(e^{\varsigma_i} + 1\right)^2} \tag{52}$$

$$\frac{\partial^2 f(\mathbf{x}_i)}{\partial \mathbf{x}_i \partial \mathbf{w}} = \sum_{i=1}^{n} \frac{y_i \left[e^{\varsigma_i}\left(\varsigma_i - 1\right) - 1\right]}{\left(e^{\varsigma_i} + 1\right)^2} \tag{53}$$

$$\frac{\partial^2 f(\mathbf{x}_i)}{\partial \mathbf{w} \partial \mathbf{x}_i} = \frac{\partial^2 f(\mathbf{x}_i)}{\partial \mathbf{x}_i \partial \mathbf{w}} = \sum_{i=1}^{n} \frac{y_i \left[e^{\varsigma_i}\left(\varsigma_i - 1\right) - 1\right]}{\left(e^{\varsigma_i} + 1\right)^2} \tag{54}$$

$$\frac{\partial^2 f(\mathbf{x}_i)}{\partial \mathbf{w}^2} = \sum_{i=1}^{n} \frac{||\mathbf{x}_i||^2 y_i^2 e^{\varsigma_i}}{\left(e^{\varsigma_i} + 1\right)^2} + \frac{2}{\sigma^2} \tag{55}$$

where $\varsigma_i = \mathbf{w}^T \mathbf{x}_i y_i$

3. [10 points] prove that the Hessian is positive semidefinite.

- A matrix is *positive semidefinite* if and only if $\mathbf{b}^T\boldsymbol{\mathcal{H}}\mathbf{b} > \vec{0}\ \forall\ \mathbf{b} \neq \vec{0}$, where $\mathbf{b}$ is a vector. $\boldsymbol{\mathcal{H}}(f(\mathbf{x}_i))$ can be re-written as, by using Einstein Summation Notation

$$\boldsymbol{\mathcal{H}} = \begin{bmatrix} \frac{||\mathbf{w}||^2 y_i^2 e^{\varsigma_i}}{(e^{\varsigma_i}+1)^2} & \frac{y_i[e^{\varsigma_i}(\varsigma_i-1)-1]}{(e^{\varsigma_i}+1)^2} \\ \frac{y_i[e^{\varsigma_i}(\varsigma_i-1)-1]}{(e^{\varsigma_i}+1)^2} & \frac{||\mathbf{x}_i||^2 y_i^2 e^{\varsigma_i}}{(e^{\varsigma_i}+1)^2} + \frac{2}{\sigma^2} \end{bmatrix} \tag{56}$$

where $\varsigma_i = \mathbf{w}^T\mathbf{x}_i y_i$

We only care about if it is going to be negative, so without loss of generality we can remove all the norms, squares, and denominators since they are *always* positive. We can also generalize our vector $\mathbf{b}$ as $\mathbf{b} = (\alpha_1, \alpha_2)^T$, where α_1 and α_2 can be anything

$$\boldsymbol{\mathcal{H}}' = \begin{bmatrix} e^{\varsigma_i} & e^{\varsigma_i}(\varsigma_i - 1) - 1 \\ e^{\varsigma_i}(\varsigma_i - 1) - 1 & e^{\varsigma_i} + \frac{2}{\sigma^2} \end{bmatrix} \tag{57}$$

where $\mathbf{b}^T\boldsymbol{\mathcal{H}}'\mathbf{b}$ can be generalized with the following result

$$\mathbf{b}^T \begin{bmatrix} x_1 & x_2 \\ x_3 & x_4 \end{bmatrix} \mathbf{b} = x_1\alpha_1^2 + x_4\alpha_2^2 + \alpha_1\alpha_2(x_2 + x_3) \tag{58}$$

when applied to $\boldsymbol{\mathcal{H}}'$ results in

$$\mathbf{b}^T\boldsymbol{\mathcal{H}}'\mathbf{b} = \alpha_1^2 e^{\varsigma_i} + \alpha_2^2\left(e^{\varsigma_i} + \frac{2}{\sigma^2}\right) + 2\alpha_1\alpha_2\left(e^{\varsigma_i}(\varsigma_i - 1) - 1\right) \tag{59}$$

Where there's two instances that we need to prove. (1) If $\alpha_1, \alpha_2 > 0$ then $\alpha_1\alpha_2(x_2+x_3) > 0$ and (2) if $\alpha_1 > 0$ and $\alpha_2 < 0$, then $\alpha_1^2 x_1 + \alpha_2^2 x_2 > \alpha_1\alpha_2(x_2+x_3)$.

$$\alpha_1, \alpha_2 > 0 \Rightarrow \alpha_1\alpha_2(x_2 + x_3) > 0 \tag{60}$$

$$\alpha_1\alpha_2(x_2 + x_3) = 2\alpha_1\alpha_2\varsigma_i e^{\varsigma_i} - 2e^{\varsigma_i} - 2 > 0 \tag{61}$$

$$\underbrace{\varsigma_i e^{\varsigma_i} - e^{\varsigma_i} - 1}_{>\ 0 \text{ if } \varsigma_i e^{\varsigma_i} > e^{\varsigma_i} - 1}; \quad \varsigma_i \neq 0 \tag{62}$$

$$\Rightarrow \varsigma_i e^{\varsigma_i} > e^{\varsigma_i} - 1 \tag{63}$$

$$\Rightarrow \varsigma_i > 1 - e^{-\varsigma_i} \quad \square \tag{64}$$

where $\varsigma_i \geq 1\ \forall\ i$. Now for the alternative case where if $\alpha_1 > 0$ and $\alpha_2 < 0$, that $\alpha_1^2 x_1 + \alpha_2^2 x_2 > \alpha_1\alpha_2(x_2 + x_3)$

$$\alpha_1^2 e^{\varsigma_i} + \alpha_2^2\left(e^{\varsigma_i} + \frac{2}{\sigma^2}\right) > \alpha_1\alpha_2\left[e^{\varsigma_i}(\varsigma_i - 1) - 1\right] \tag{65}$$

$$\left[\alpha_1^2 + \alpha_2^2 + \frac{2}{\sigma^2 e^{\varsigma_i}}\right] > \left[\alpha_1\alpha_2(\varsigma_i - 1) - e^{-\varsigma_i}\right] \tag{66}$$

if $\varsigma_i \gg 1$

$$\underbrace{\alpha_1^2 + \alpha_2^2}_{>0} > \underbrace{\alpha_1 \alpha_2}_{<0} \underbrace{(\varsigma_i - 1)}_{>0} \quad \square \tag{67}$$

for $\varsigma_i \geq 0$ (always)

$$\underbrace{\alpha_1^2 + \alpha_2^2}_{>0} + \underbrace{e^{-\varsigma_i}}_{\geq 1} \underbrace{\left(\frac{2}{\sigma^2} + 1\right)}_{>0} > \underbrace{\alpha_1 \alpha_2}_{<0} \underbrace{(\varsigma_i - 1)}_{>0} \quad \square \tag{68}$$

4. [5 points] Why does this prove that we are GUARANTEED to have within $\epsilon > 0$ of the right answer?

   - Because the Hessian Matrix of the function is positive-semidefinite, the given function is convex. On top of which, since we are maximizing the function, we are required to step with the gradient and stop when the slope is less than some tolerance (ϵ). The gradient gives the slope/direction of the greatest change and we "walk" in teh direction which makes the gradient zero. We are guarenteed that one exists since the Hessian is positive-semidefinite.