3.1

Solving the Least Squares problem for the student debt data gives the trend line as $y = b + wx$, where x is year, y is the amount of debt in trillions of dollars, and optimal parameters b and w are found respectively as $b = -160.7$ and $w = .08$ (see our code repository). Therefore in year 2050 ($x_{new} = 2050$), the student debt will reach $y_{new} = -160.7 + .08(2050) \approx 3.9$ trillion dollars if this trend continues.

3.3

(a)

$$\mathbf{Q} = 2\sum_{p=1}^{P} \tilde{\mathbf{x}}_p \tilde{\mathbf{x}}_p^T, \; \mathbf{r} = -2\sum_{p=1}^{P} y_p \tilde{\mathbf{x}}_p \text{ and } d = \sum_{p=1}^{P} y_p^2.$$

(b)

$$\mathbf{z}^T\mathbf{Q}\mathbf{z} = \mathbf{z}^T \left(2\sum_{p=1}^{P} \tilde{\mathbf{x}}_p \tilde{\mathbf{x}}_p^T\right) \mathbf{z} = \sum_{p=1}^{P} 2\left(\tilde{\mathbf{x}}_p^T \mathbf{z}\right)^2 \geq 0. \text{ So } \mathbf{Q} \text{ is positive semidefinite}$$

with all nonnegative eigenvalues

$$\nabla^2 g\left(\tilde{\mathbf{w}}\right) = \tfrac{1}{2}\left(\mathbf{Q} + \bar{\mathbf{Q}}\right) = \mathbf{Q}.$$ (c)

Q has all nonnegative eigenvalues as shown in (b), the function g is convex according to the second order definition of convexity.

(d)

The 1st Newton step is the solution to the following linear system of equation:

$$\left[\nabla^2 g\left(\mathbf{v}\right)\right] \tilde{\mathbf{w}} = \left[\nabla^2 g\left(\mathbf{v}\right)\right] \mathbf{v} - \nabla g\left(\mathbf{v}\right)$$

Plugging $\nabla^2 g\left(\mathbf{v}\right) = \mathbf{Q} = 2\sum^{P} \tilde{\mathbf{x}}_p \tilde{\mathbf{x}}_p^T$ and $\nabla g\left(\mathbf{v}\right) = \mathbf{Q}\mathbf{v} + \mathbf{r} =$ into the equation above have

$$\mathbf{Q}\tilde{\mathbf{w}} = \mathbf{Q}\mathbf{v} - \left(\mathbf{Q}\mathbf{v} + \mathbf{r}\right) = -\mathbf{r}.$$

Plugging again $\mathbf{Q} = 2\sum_{p=1}^{P} \tilde{\mathbf{x}}_p \tilde{\mathbf{x}}_p^T$ and $\mathbf{r} = -2\sum_{p=1}^{P} y_p \tilde{\mathbf{x}}_p$ from part a) into this equation gives

$$\left(\sum_{p=1}^{P} \tilde{\mathbf{x}}_p \tilde{\mathbf{x}}_p^T\right) \tilde{\mathbf{w}} = \sum_{p=1}^{P} y_p \tilde{\mathbf{x}}_p.$$

3.10

a) $\sigma^{-1}\left(\sigma\left(t\right)\right)=\log\left(\frac{\sigma(t)}{1-\sigma(t)}\right)=\log\left(\frac{\frac{1}{1+e^{-t}}}{1-\frac{1}{1+e^{-t}}}\right)=\log\left(\frac{\frac{1}{1+e^{-t}}}{\frac{e^{-t}}{1+e^{-t}}}\right)=\log\left(\frac{1}{e^{-t}}\right)=t.$

b) From (3.23) we have $\sigma\left(b+x_pw\right)\approx y_p$ for all p. Taking the *sigmoid-inverse* from both sides then gives $\sigma^{-1}\left(\sigma\left(b+x_pw\right)\right)\approx\sigma^{-1}\left(y_p\right)$, which simplifies to $b+x_pw\approx\log\left(\frac{y_p}{1-y_p}\right)$.

c) Similar to part b) of Exercise 3.8 we can form the Least Squares cost function

$$g\left(b,w\right)=\sum_{p=1}^{P}\left(b+x_pw-\log\left(\frac{y_p}{1-y_p}\right)\right)^2,$$

which can be written more compactly as

$$g\left(\tilde{\mathbf{w}}\right)=\sum_{p=1}^{P}\left(\tilde{\mathbf{x}}_p^T\tilde{\mathbf{w}}-\log\left(\frac{y_p}{1-y_p}\right)\right)^2,$$

introducing the familiar notation

$$\tilde{\mathbf{x}}_p=\begin{bmatrix}1\\x_p\end{bmatrix}\quad\text{and}\quad\tilde{\mathbf{w}}=\begin{bmatrix}b\\w\end{bmatrix}.$$

Setting the gradient of g to zero, we have

$$\nabla g\left(\tilde{\mathbf{w}}\right)=2\sum_{p=1}^{P}\left(\tilde{\mathbf{x}}_p^T\tilde{\mathbf{w}}-\log\left(\frac{y_p}{1-y_p}\right)\right)\tilde{\mathbf{x}}_p=\mathbf{0}_{2\times1}.$$

This system can be solved for *optimal* parameters using a linear solver or algebraically as $\tilde{\mathbf{w}}=\left(\sum_{p=1}^{P}\tilde{\mathbf{x}}_p\tilde{\mathbf{x}}_p^T\right)^{-1}\left(\sum_{p=1}^{P}\log\left(\frac{y_p}{1-y_p}\right)\tilde{\mathbf{x}}_p\right)$

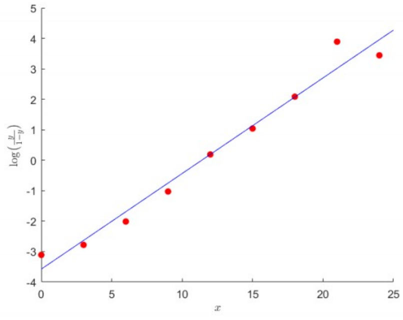


Figure 1: A plot of linearized data and the best fit line.

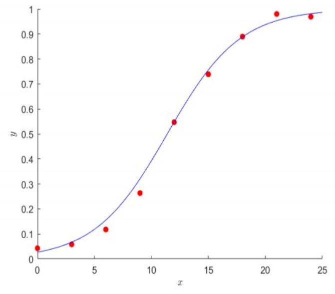


Figure 2: A plot of original data and the logistic sigmoid fit to the data.

3.11

a) Using the compact notation the cost function can be written as

$$g\left(\tilde{\mathbf{w}}\right)=\sum_{p=1}^{P}\left(\sigma\left(\tilde{\mathbf{x}}_{p}^{T}\tilde{\mathbf{w}}\right)-y_{p}\right)^{2}.$$

Now, the partial derivative with respect to the i th entry in $\tilde{\mathbf{w}}$ can be computed as

$$\frac{\partial}{\partial\tilde{w}_{i}}g\left(\tilde{\mathbf{w}}\right)=2\sum_{p=1}^{P}\left(\sigma\left(\tilde{\mathbf{x}}_{p}^{T}\tilde{\mathbf{w}}\right)-y_{p}\right)\frac{\partial}{\partial\tilde{w}_{i}}\left(\sigma\left(\tilde{\mathbf{x}}_{p}^{T}\tilde{\mathbf{w}}\right)-y_{p}\right)$$

$$=2\sum_{p=1}^{P}\left(\sigma\left(\tilde{\mathbf{x}}_{p}^{T}\tilde{\mathbf{w}}\right)-y_{p}\right)\sigma'\left(\tilde{\mathbf{x}}_{p}^{T}\tilde{\mathbf{w}}\right)\frac{\partial}{\partial\tilde{w}_{i}}\left(\tilde{\mathbf{x}}_{p}^{T}\tilde{\mathbf{w}}\right)$$

$$=2\sum_{p=1}^{P}\left(\sigma\left(\tilde{\mathbf{x}}_{p}^{T}\tilde{\mathbf{w}}\right)-y_{p}\right)\sigma\left(\tilde{\mathbf{x}}_{p}^{T}\tilde{\mathbf{w}}\right)\left(1-\sigma\left(\tilde{\mathbf{x}}_{p}^{T}\tilde{\mathbf{w}}\right)\right)\tilde{x}_{p,i}.$$

The full gradient of g is then given by

$$\nabla g\left(\tilde{\mathbf{w}}\right)=2\sum_{n=1}^{P}\left(\sigma\left(\tilde{\mathbf{x}}_{p}^{T}\tilde{\mathbf{w}}\right)-y_{p}\right)\sigma\left(\tilde{\mathbf{x}}_{p}^{T}\tilde{\mathbf{w}}\right)\left(1-\sigma\left(\tilde{\mathbf{x}}_{p}^{T}\tilde{\mathbf{w}}\right)\right)\tilde{\mathbf{x}}_{p}.$$

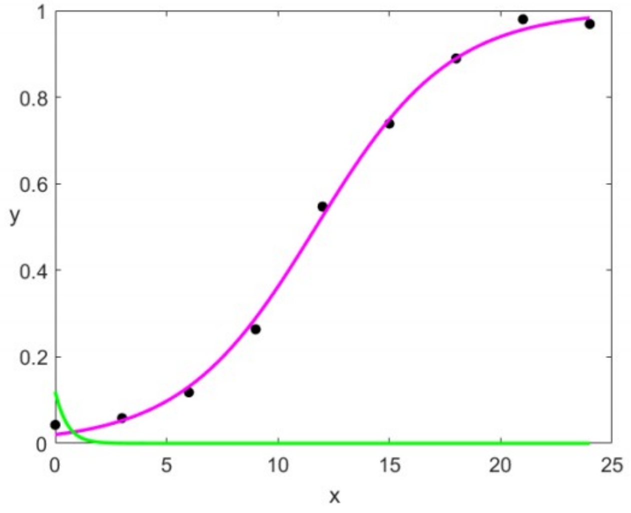


Figure 3: A dataset along with two sigmoidal fits (shown in magenta and green), each found via minimizing the Least Squares cost in (3.26) using gradient descent with a different initialization.

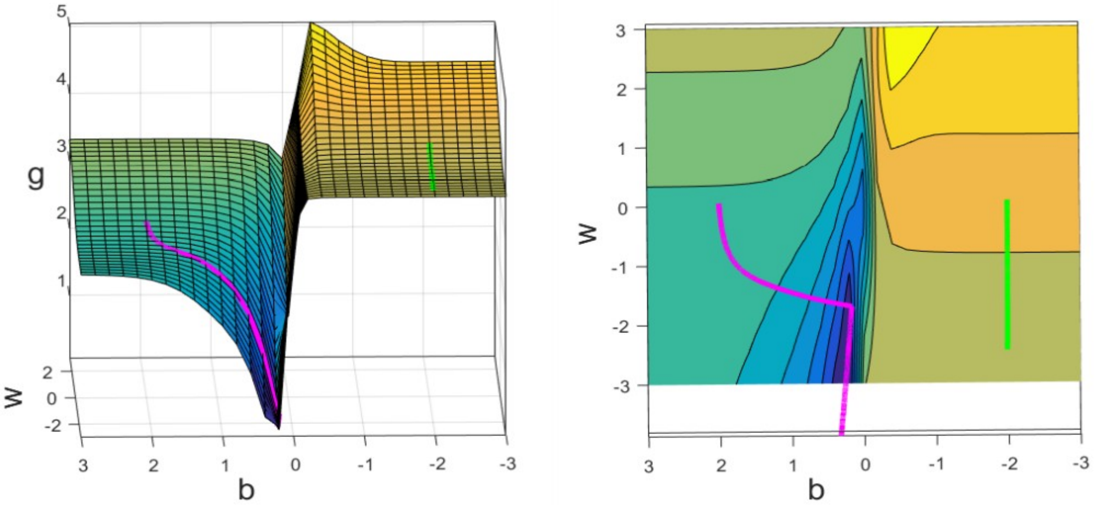


Figure 4: A surface (left) and contour (right) plot of this cost function, along with the paths taken by the two runs of gradient descent.

3.13
(a)

the regularized cost function can be written as $g\left(\tilde{\mathbf{w}}\right)=\sum\limits_{p=1}^{P}\left(\sigma\left(\tilde{\mathbf{x}}_p^T\tilde{\mathbf{w}}\right)-y_p\right)^2+\lambda\mathbf{w}^T\mathbf{w}$. The gradient of g with respect to $\tilde{\mathbf{w}}$ can then be written as

$$\nabla g\left(\tilde{\mathbf{w}}\right)=\nabla_{\tilde{\mathbf{w}}}\left(\sum_{p=1}^{P}\left(\sigma\left(\tilde{\mathbf{x}}_p^T\tilde{\mathbf{w}}\right)-y_p\right)^2\right)+\nabla_{\tilde{\mathbf{w}}}\left(\lambda\mathbf{w}^T\mathbf{w}\right).$$

The first part, as shown in Exercise 3.11, can be written as

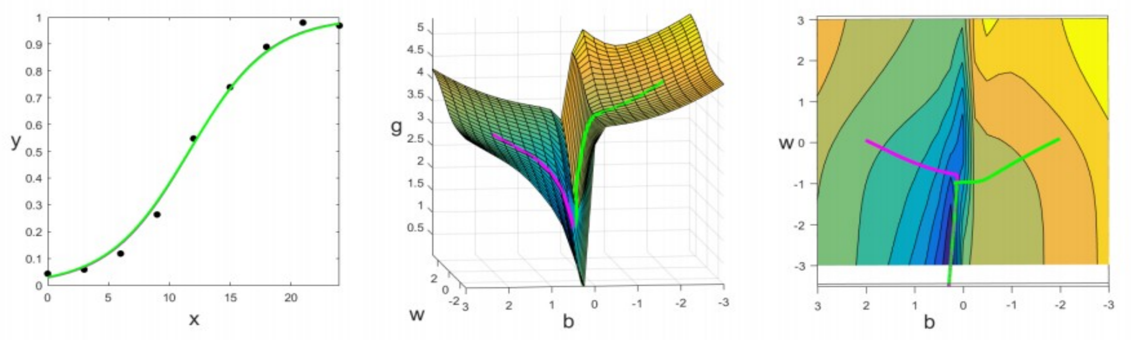
$$\nabla_{\tilde{\mathbf{w}}}\left(\sum_{p=1}^{P}\left(\sigma\left(\tilde{\mathbf{x}}_p^T\tilde{\mathbf{w}}\right)-y_p\right)^2\right)$$

$$=2\sum_{p=1}^{P}\left(\sigma\left(\tilde{\mathbf{x}}_p^T\tilde{\mathbf{w}}\right)-y_p\right)\sigma\left(\tilde{\mathbf{x}}_p^T\tilde{\mathbf{w}}\right)\left(1-\sigma\left(\tilde{\mathbf{x}}_p^T\tilde{\mathbf{w}}\right)\right)\tilde{\mathbf{x}}_p,$$

and the second part as

$$\nabla_{\tilde{\mathbf{w}}}\left(\lambda\mathbf{w}^T\mathbf{w}\right)=\left[\begin{array}{c}\frac{\partial}{\partial b}\left(\lambda\mathbf{w}^T\mathbf{w}\right)\\ \nabla_{\mathbf{w}}\left(\lambda\mathbf{w}^T\mathbf{w}\right)\end{array}\right]=\left[\begin{array}{c}0\\ \nabla_{\mathbf{w}}\left(\lambda\mathbf{w}^T\mathbf{w}\right)\end{array}\right]$$

$$=\left[\begin{array}{c}0\\ 2\lambda\mathbf{w}\end{array}\right]=2\lambda\left[\begin{array}{c}0\\ \mathbf{w}\end{array}\right].$$



y
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
5
10
15
20
x
g
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
w
2
0
-2
3
2
1
0
-1
-2
-3
b
w
3
2
1
0
-1
-2
-3
3
2
1
0
-1
-2
-3
b