CS/ECE/ME532 Period 16 Activity

Estimated Time: 40 minutes for Q1 and 30 minutes for Q2.

1. The squared-error cost function $f(\boldsymbol{w}) = ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||_2^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})$ may be rewritten as a perfect square in the form

$$f(\boldsymbol{w}) = (\boldsymbol{w} - \boldsymbol{w}_{LS})^T \boldsymbol{X}^T \boldsymbol{X} (\boldsymbol{w} - \boldsymbol{w}_{LS}) + c$$

where $\boldsymbol{w}_{LS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ and $c = \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$. This assumes the n-by-p ($p < n$) matrix $\boldsymbol{X}$ is full rank. $f(\boldsymbol{w})$ is called a "quadratic form" in $\boldsymbol{w}$ since it is a quadratic function of $\boldsymbol{w}$.

a) Prove that the minimum value of $f(\boldsymbol{w}) = c$ when $\boldsymbol{w} = \boldsymbol{w}_{LS}$ and all other $\boldsymbol{w}$ result in higher values of $f(\boldsymbol{w})$.

b) Suppose $\boldsymbol{y} = \begin{bmatrix} 1 \\ 1/2 \\ 1 \\ 0 \end{bmatrix}$ and the 4-by-2 $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ has singular value decomposition $\boldsymbol{U} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$, $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1/2 \end{bmatrix}$, and $\boldsymbol{V} = \boldsymbol{I}$. Sketch a contour plot of $f(\boldsymbol{w})$ in the w_1-w_2 plane.

c) Suppose $\boldsymbol{y} = \begin{bmatrix} 1 \\ 1/5 \\ 1 \\ 0 \end{bmatrix}$, $\boldsymbol{U} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$, $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1/5 \end{bmatrix}$, and $\boldsymbol{V} = \boldsymbol{I}$. Sketch a contour plot of $f(\boldsymbol{w})$ in the w_1-w_2 plane. How do the singular values of $\boldsymbol{X}$ affect the shape of the contours?

d) In this case assume $\boldsymbol{y} = \begin{bmatrix} \sqrt{2} \\ 0 \\ 1 \\ 0 \end{bmatrix}$, $\boldsymbol{U} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$, $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1/2 \end{bmatrix}$, and $\boldsymbol{V} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$. Sketch a contour plot of $f(\boldsymbol{w})$ in the w_1-w_2 plane. How do the right singular vectors of $\boldsymbol{X}$ affect the contours?

e) Sketch the gradient of $f(\boldsymbol{w})$ at $\boldsymbol{w} = \boldsymbol{w}_o$ on the contour plot for the previous case when

i. $\boldsymbol{w}_o = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$

ii. $\boldsymbol{w}_o = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$

iii. $\boldsymbol{w}_o = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$

2. The provided script will compute a specified number of iterations of the gradient descent algorithm and help you display the path taken by the weights in the gradient descent iteration superimposed on a contour plot. Assume $\boldsymbol{y} = \begin{bmatrix} \sqrt{2} \\ 0 \\ 1 \\ 0 \end{bmatrix}$, the 4-by-2 $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ has singular value decomposition $\boldsymbol{U} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$, $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1/2 \end{bmatrix}$, and $\boldsymbol{V} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$. Complete 20 iterations of gradient descent in each case specified below.

Include the plots you generate below with your submission.

a) What is the maximum value for the step size τ that will guarantee convergence?

b) Start gradient descent from the point $\boldsymbol{w} = \begin{bmatrix} 1.5 \\ -0.5 \end{bmatrix}$ and use a step size of $\tau = 0.5$. How do you explain the trajectory the weights take toward the optimum, e.g., why is it shaped this way? What would the trajectory be if you started from the point $\boldsymbol{w} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$? From $\boldsymbol{w} = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$?

c) Start gradient descent from the point $\boldsymbol{w} = \begin{bmatrix} 1.5 \\ -0.5 \end{bmatrix}$ and use a step size of $\tau = 2.5$. Complete 20 iterations. What happens and why?

d) Now change $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1/4 \end{bmatrix}$, start from $\boldsymbol{w} = \begin{bmatrix} 1.5 \\ -0.5 \end{bmatrix}$ and use $\tau = 0.5$. What happens to the cost function? How does the change in the cost function lead to a change in the trajectory of the gradient descent weights? What would happen if you further decreased the smaller singular value?

e) Discuss how changing the ratio of the singular values changes the shape of the cost function and how that might affect the number of iterations it takes for gradient descent to get close to the optimum.