

CS/ECE/ME532 Activity 16

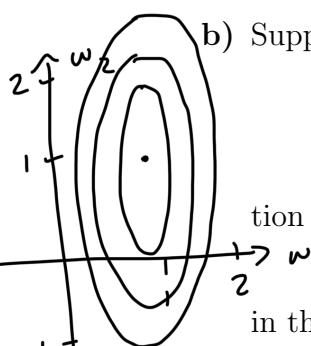
Estimated Time: 40 minutes for Q1 and 30 minutes for Q2.

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

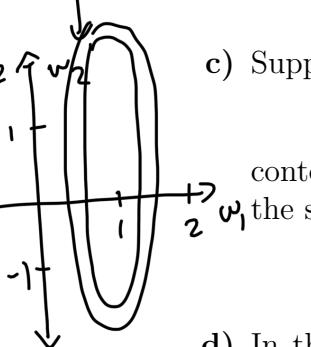
$$f(\mathbf{w}) = (\mathbf{w} - \mathbf{w}_{LS})^T \mathbf{X}^T \mathbf{X} (\mathbf{w} - \mathbf{w}_{LS}) + c$$

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

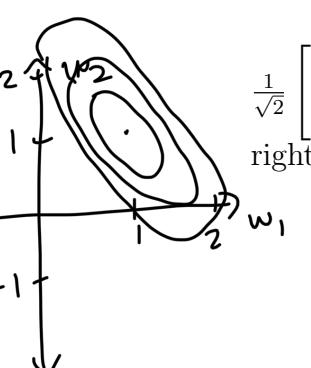
- Prove that the minimum value of $f(\mathbf{w}) = c$ when $\mathbf{w} = \mathbf{w}_{LS}$ and all other \mathbf{w} result in higher values of $f(\mathbf{w})$. When $\mathbf{w} = \mathbf{w}_{LS}$, $(\mathbf{w} - \mathbf{w}_{LS})$ becomes 0, so the first term disappears, leaving $f(\mathbf{w}) = c$. \mathbf{w}_{LS} is the lowest value possible so all other \mathbf{w} will make $(\mathbf{w} - \mathbf{w}_{LS})$ positive.



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



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



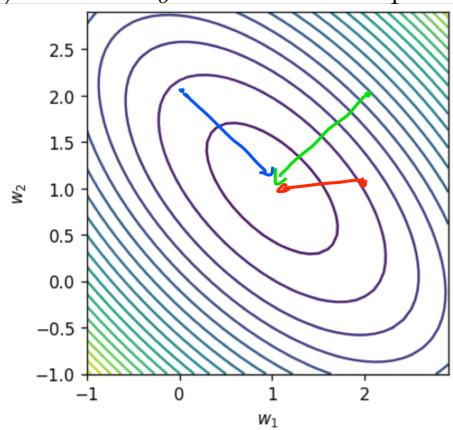
d) In this case assume $\mathbf{y} = \begin{bmatrix} \sqrt{2} \\ 0 \\ 1 \\ 0 \end{bmatrix}$, $\mathbf{U} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$, $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1/2 \end{bmatrix}$, and $\mathbf{V} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$. Sketch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

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

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

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

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



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 $\mathbf{y} = \begin{bmatrix} \sqrt{2} \\ 0 \\ 1 \\ 0 \end{bmatrix}$, the 4-by-2

$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T$ has singular value decomposition $\mathbf{U} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$, $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1/2 \end{bmatrix}$, and

$\mathbf{V} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$. 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?

$$0 < \tau < \frac{2}{\|\mathbf{A}\|_2 \|\mathbf{b}\|^2}$$

$$\text{max step size is } \frac{2}{1^2} = 2$$

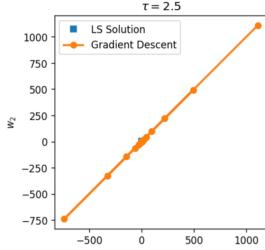
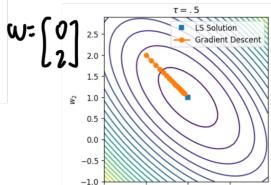
- b) Start gradient descent from the point $\mathbf{w} = \begin{bmatrix} 1.5 \\ -0.5 \end{bmatrix}$ and use a step size of $\tau = 0.5$.

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

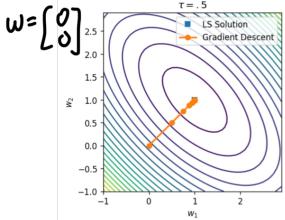
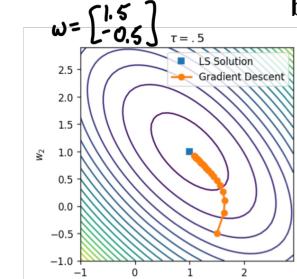
$\mathbf{w} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$? From $\mathbf{w} = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$? The trajectory curves to first reach the area/line with fastest descent and then follows that line to the minimum at $(1, 1)$. For these \mathbf{w} , they already start at a fastest descent line, which are the axes of the ellipse.

- c) Start gradient descent from the point $\mathbf{w} = \begin{bmatrix} 1.5 \\ -0.5 \end{bmatrix}$ and use a step size of $\tau = 2.5$.

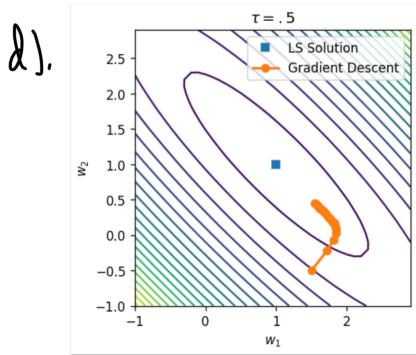
Complete 20 iterations. What happens and why?



Because step size is greater than max for convergence, 2, each step overshoots the minimum more and more so the descent never converges to the minimum



- d) Now change $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1/4 \end{bmatrix}$, start from $w = \begin{bmatrix} 1.5 \\ -0.5 \end{bmatrix}$ 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.



The contours became longer, basically meaning that the cost function is less steep, so 20 iterations of the descent is no longer enough to reach the minimum. We would need a larger stepsize or more iterations.

Further decreasing the smaller singular value would make an even flatter cost function, so the descent would take even longer.

- e). A larger ratio between singular values, meaning one is much smaller than the other, creates a flatter cost function, especially along the major axis. This reflects in the contours because they appear longer and manifests in gradient descent because it takes more iterations to get close to the optimum, since each step now covers way less ground.