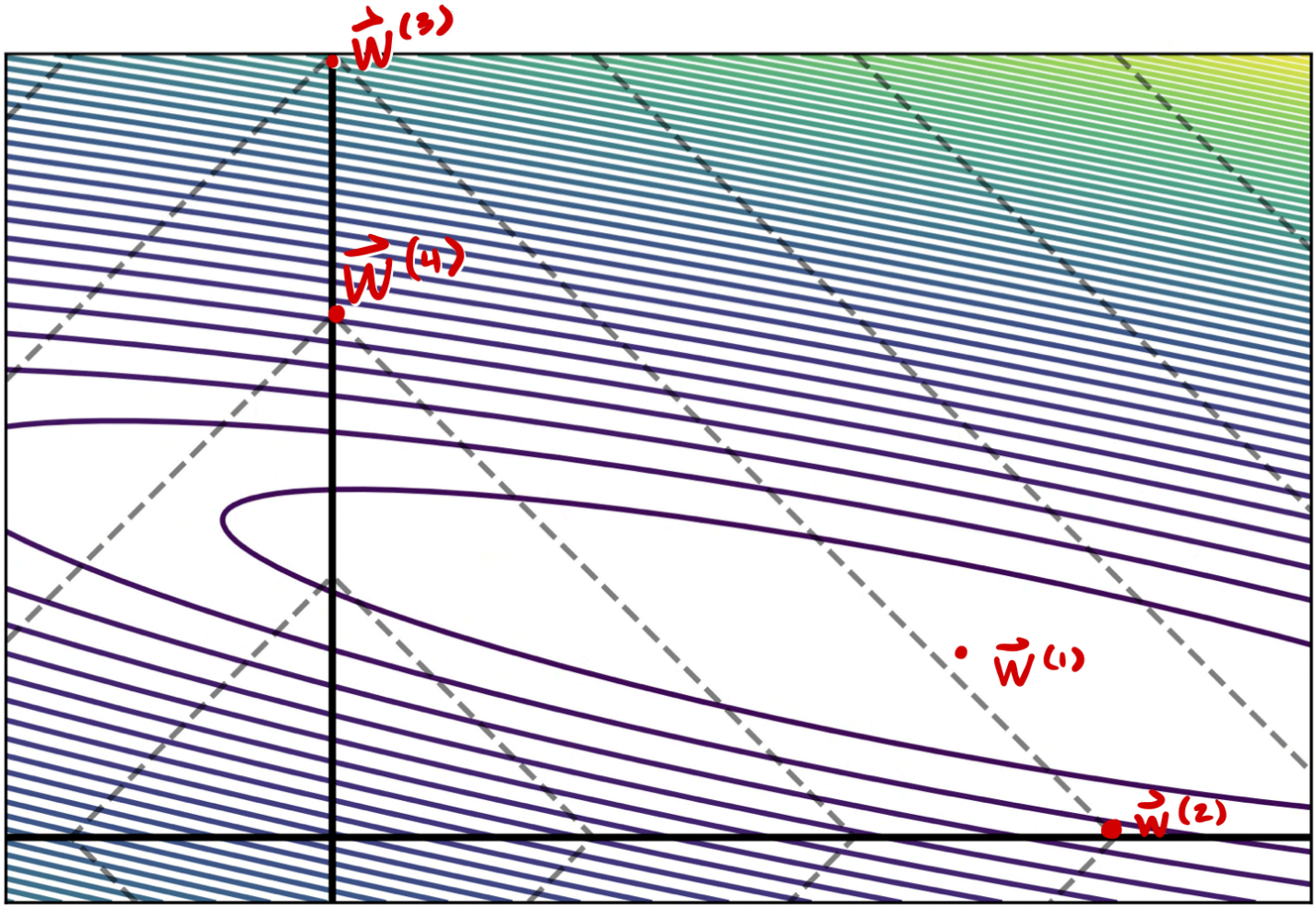
The image below shows the contours of the expected square loss:

$$R(\vec{w}) = \frac{1}{n}\sum_{i=1}^{n}(\text{Aug}(\vec{x}^{(i)}) \cdot \vec{w} - y_i)^2.$$

The dashed lines show the contours of $\rho(\vec{w}) = \|\vec{w}\|_1$; that is, all points along the same dashed line have the same 1-norm.



Which of the points labeled could possibly be the minimizer of the unregularized risk, $R(\vec{w})$?

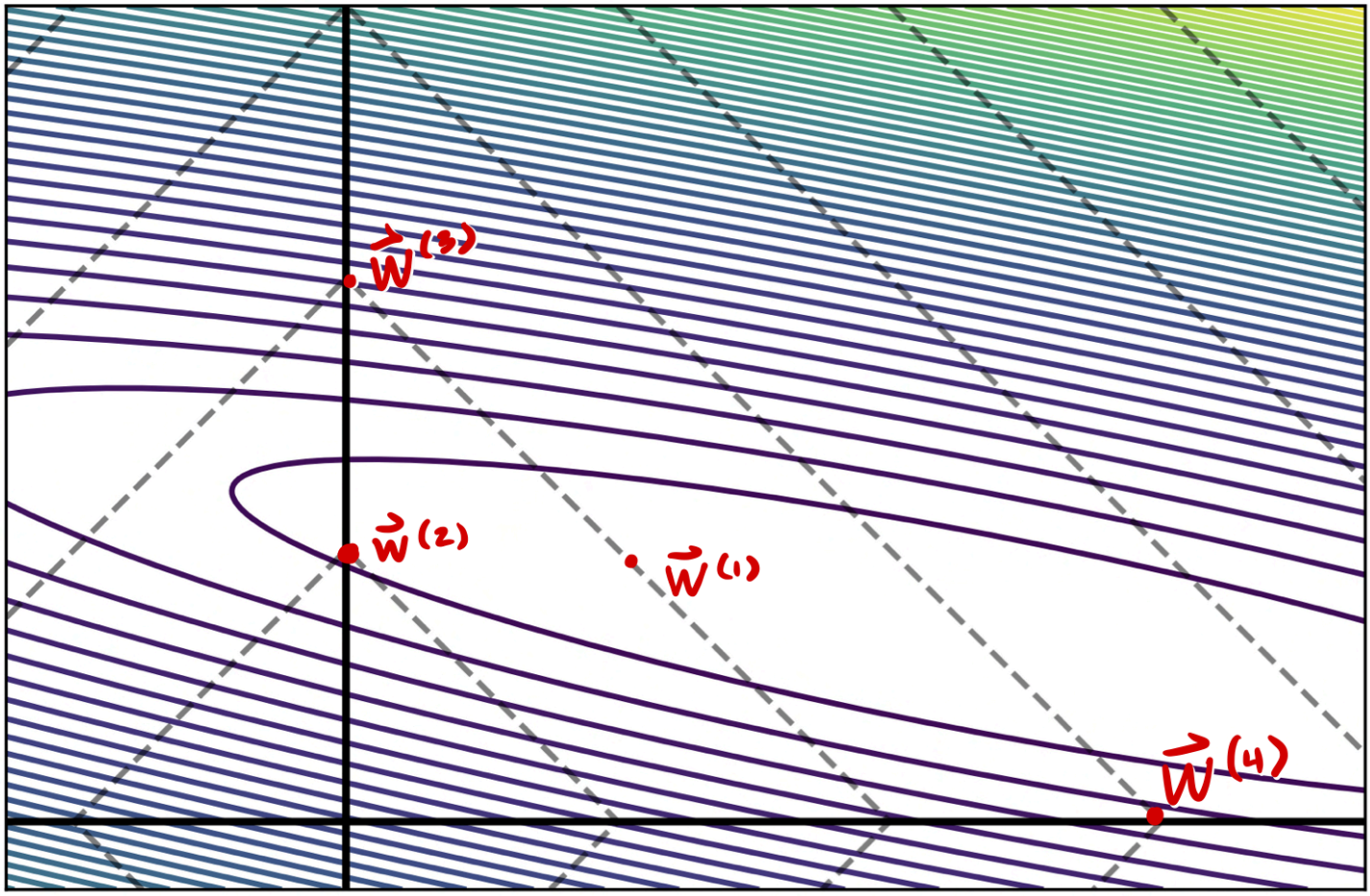
(X) $\vec{w}^{(1)}$
( ) $\vec{w}^{(2)}$
( ) $\vec{w}^{(3)}$
( ) $\vec{w}^{(4)}$

Since we are asking for the minimizer of unregularized risk, we can ignore the contours of $\rho(\vec{w})$ which are the dashed lines. If we look at just the contours of the risk (the solid lines), we can see the lowest point would be at $\vec{w}^{(1)}$ as risk is a convex function and can be geometrically interpreted as a bowl.

The image below shows the contours of the expected square loss:

$$R(\vec{w}) = \frac{1}{n}\sum_{i=1}^{n}(\text{Aug}(\vec{x}^{(i)}) \cdot \vec{w} - y_i)^2.$$

The dashed lines show the contours of $\rho(\vec{w}) = \|\vec{w}\|_1$; that is, all points along the same dashed line have the same 1-norm.



Suppose that when the LASSO is performed with regularization parameter λ_1, the optimal choice of parameter vector is found to be $\vec{w}^{(1)}$.

Suppose the LASSO is repeated, this time with the regularization parameter $\lambda_2 > \lambda_1$. Assuming that one of the labeled points is the optimal choice of parameter vector with this new regularization parameter, which point could it be?

You may assume that $\lambda_1 > 0$.

( ) $\vec{w}^{(1)}$
(X) $\vec{w}^{(2)}$

( ) $\vec{w}^{(3)}$
( ) $\vec{w}^{(4)}$

Now we repeat LASSO but with a stronger emphasis on regularization: $\lambda_2 > \lambda_1$. This means that in comparison with LASSO with λ_1, we want to have a smaller 1-norm for our weight vector $\vec{w}$ when doing LASSO with λ_2. Since $\vec{w}^{(1)}$ is the optimal choice for LASSO with λ_1, we can see that the only other choice of $\vec{w}$ with a smaller 1-norm (the contours corresponding to the dashed lines) is $\vec{w}^{(2)}$.

Recall that the ridge regression objective function is

$$\tilde{R}(\vec{w}) = \frac{1}{n}\sum_{i=1}^{n}(\text{Aug}(\vec{x}^{(i)}) \cdot \vec{w} - y_i)^2 + \|\vec{w}\|^2.$$

True or False: $\tilde{R}(\vec{w})$ is convex as a function of $\vec{w}$.

(X) True
( ) False

We have learned a few different ways to prove the convexity of a function. One of these is to use properties of convex functions (as done in Homework 03 Problem 2). We know that squared loss is convex as proven in Lecture 05, and since the sum of convex functions is convex, $\sum_{i=1}^{n}(y_i - \vec{w} \cdot \text{Aug}(\vec{x}^{(i)}))^2$ is convex. Finally, multiplying by a positive scalar, $\frac{1}{n}$ does not change its convexity.

From Homework 03 Problem 2, we also know that the squared L-2 norm is convex. One way to show this is by showing that its Hessian is PSD (or all eigenvalues of the Hessian are non-negative). Since $\lambda > 0$, multiplying $\|\vec{w}\|^2$ by a positive constant does not change its convexity (the Hessian is now $2\lambda I \succeq 0$ which is a diagonal matrix with positive diagonal entries, namely 2λ.)

Therefore, $\tilde{R}(\vec{w})$ which is the sum of two convex functions is convex.

Recall that in ridge regression, we solve the following optimization problem:

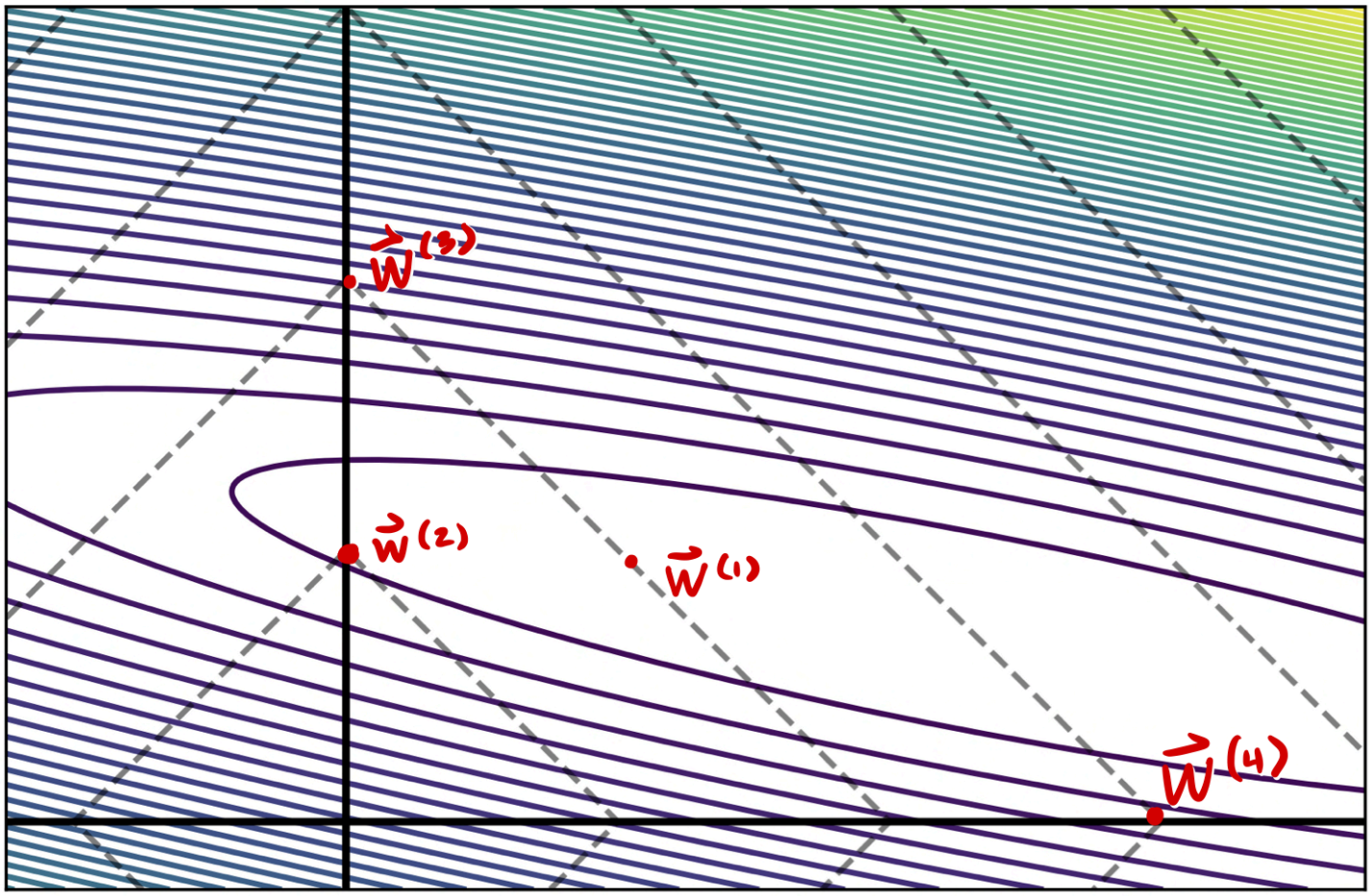
$$\arg\min_{\vec{w}} \sum_{i=1}^{n}(y_i - \vec{w} \cdot \text{Aug}(\vec{x}^{(i)}))^2 + \lambda\|\vec{w}\|^2.$$

where $\lambda > 0$ is a hyperparameter controlling the strength of regularization.

Suppose you solve the ridge regression problem with $\lambda = 2$, and the resulting solution has a mean squared error of 10.

Now suppose you increase the regularization strength to $\lambda = 4$ and solve the ridge regression problem again. True or False: it is possible that the mean squared error of the new solution is less than 10.

By "mean squared error," we mean $\frac{1}{n}\sum_{i=1}^{n}(y_i - \vec{w} \cdot \text{Aug}(\vec{x}^{(i)})^2$ for the training data.



( ) True
(X) False

Recall from Lecture 09 that the main idea behind regularization is that it “trades an increase in risk for a decrease in complexity.” In comparison to the ridge regression model fitted with $\lambda = 2$, when we fit a ridge regression model with a regularization parameter of $\lambda = 4$, we are essentially trading an increase in risk (i.e. a higher MSE) for a less complex model (i.e. a higher λ).

Recall that in ridge regression, we solve the following optimization problem:

$\arg\min_{\vec{w}} \sum_{i=1}^{n} (y_i - \vec{w} \cdot \text{Aug}(\vec{x}^{(i)}))^2 + \lambda \|\vec{w}\|^2$.

where $\lambda > 0$ is a hyperparameter controlling the strength of regularization.

Suppose you solve the ridge regression problem with $\lambda = 2$, and the resulting solution is the weight vector $\vec{w}_{\text{old}}$. You then solve the ridge regression problem with $\lambda = 4$ and find a weight vector $\vec{w}_{\text{new}}$.

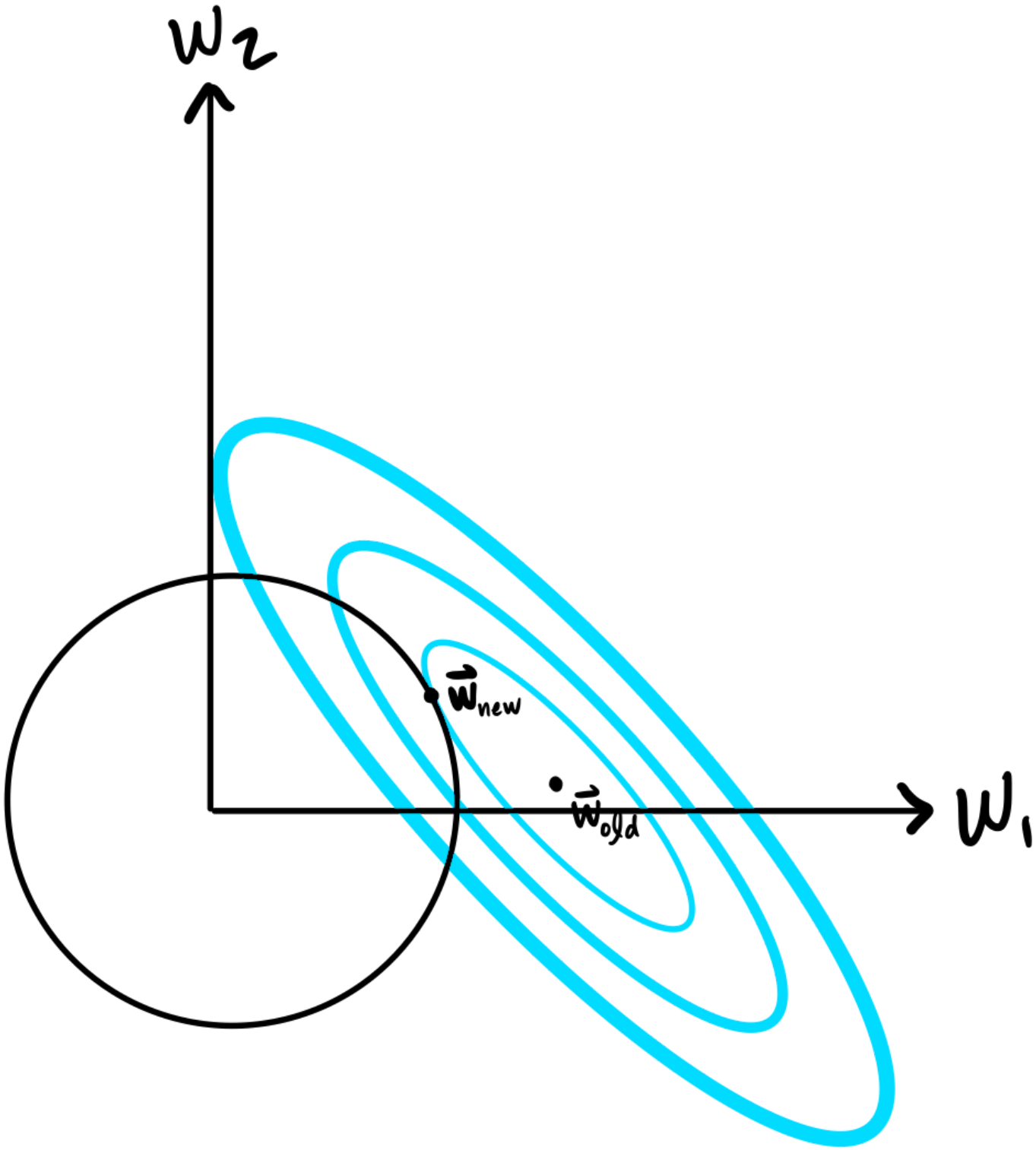
True or False: each component of the new solution, $\vec{w}_{\text{new}}$, must be less than or equal to the corresponding component of the old solution, $\vec{w}_{\text{old}}$.

( ) True
(X) False

While it is true that $\|\vec{w}_{\text{new}}\| \leq \|\vec{w}_{\text{old}}\|$, this does not imply that each component of $\vec{w}_{\text{new}}$ is less than or equal to the corresponding component of $\vec{w}_{\text{old}}$.

The picture to have in mind is that of the contour lines of the mean squared error (which are ovals), along with the circles representing where $\|\vec{w}\| = c$ for some constant c. The question asked us to consider going from $\lambda = 2$ to $\lambda = 4$, but to gain an intuition we can think of going from no regularization ($\lambda = 0$) to some regularization ($\lambda > 0$); this won't affect the outcome, but will make the story easier to tell.

Consider the situation shown below:



When we had no regularization, the solution was $\vec{w}_{\text{old}}$, as marked. Suppose we add regularization, and we're told that the regularization is such that when we solve the ridge regression problem, the norm of $\vec{w}_{\text{new}}$ will be equal to c, and that the radius of the circle we've drawn is c. Then the solution $\vec{w}_{\text{new}}$ will be the point marked, since that is the point on the circle that is on the lowest contour.

Notice that the point $\vec{w}_{\text{new}}$ is closer to the origin, and it's first component is much smaller than the first component of $\vec{w}_{\text{old}}$. However, the second component of $\vec{w}_{\text{new}}$ is actually *larger* than the second component of $\vec{w}_{\text{old}}$.