**Learning rate and features scaling**

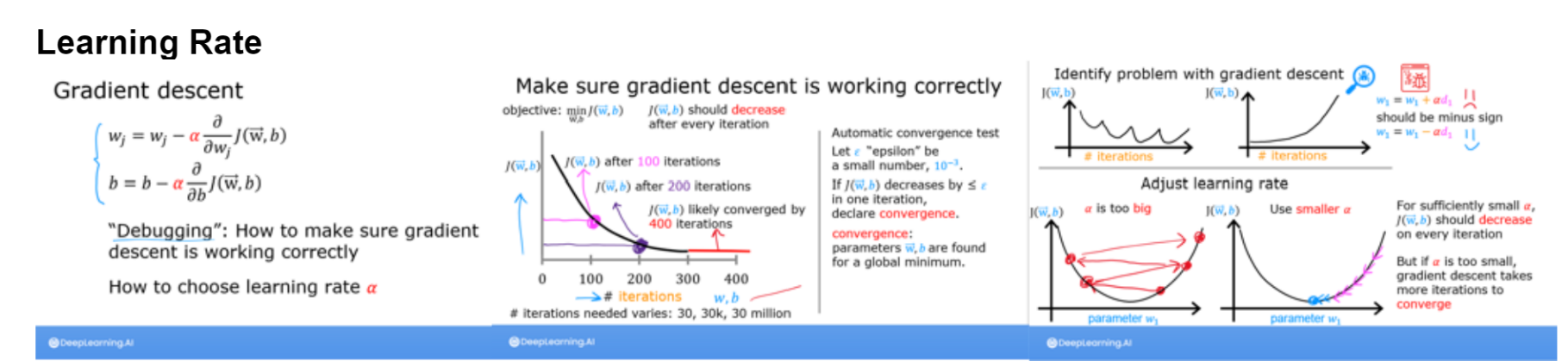


Figure 1 Learning rates basics

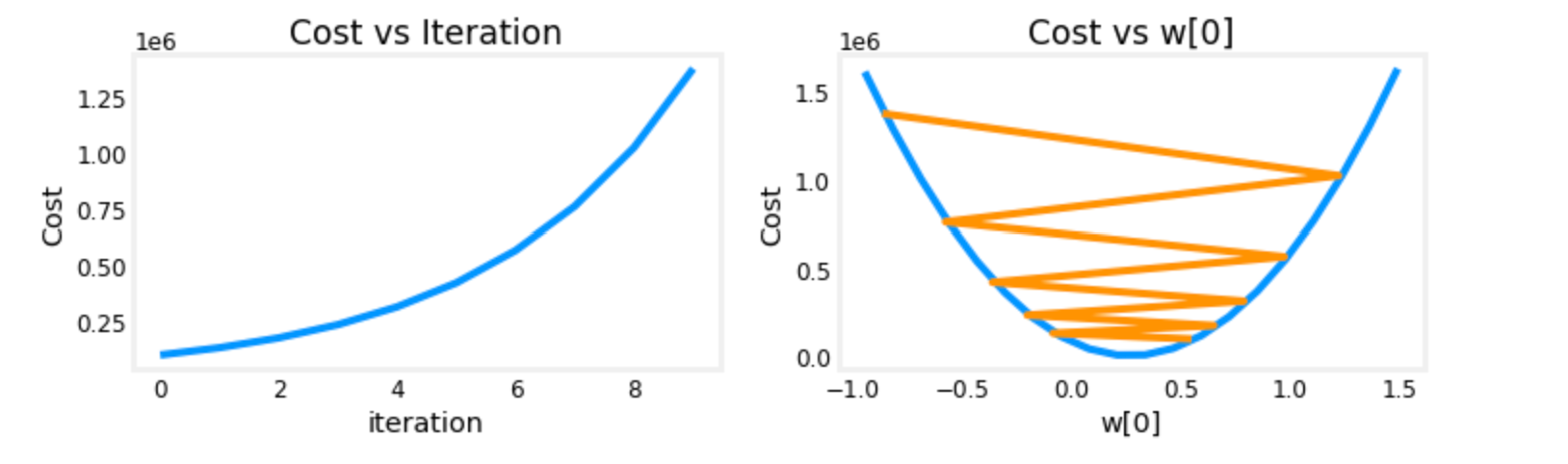


Figure 2 Learning rate way too high

The fig2 show a learning rate way too high. Easy to identify, the cost function increases.

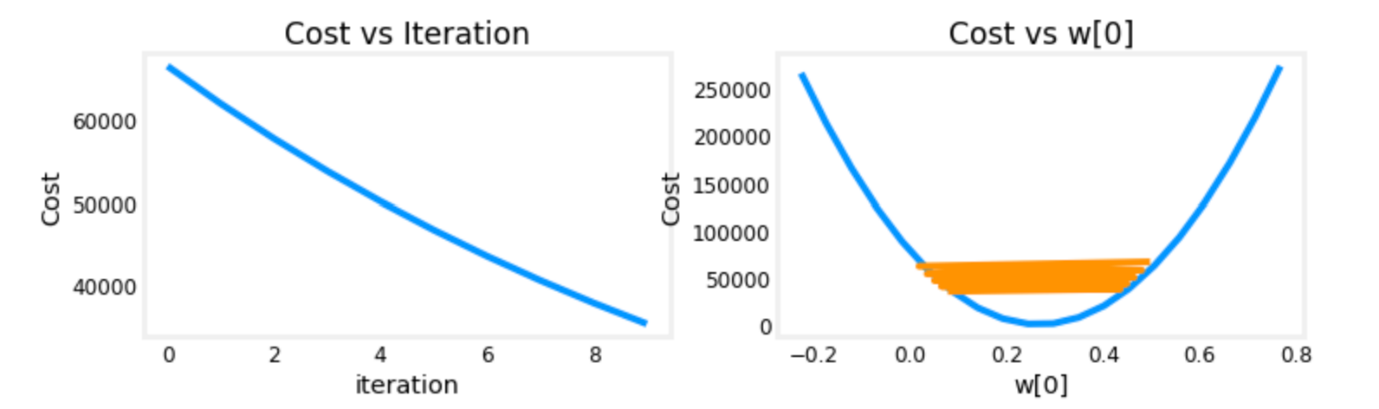


Figure 3 Learning rate too high

Indications that the learning rate is too high :

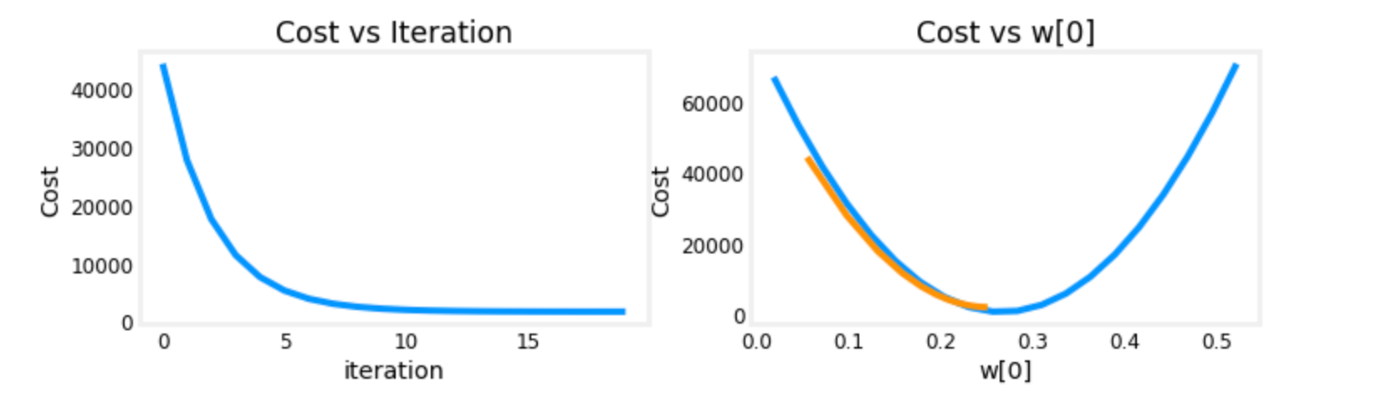
* The cost function value increase after parameters update
* The solution doesn’t converge
* Overshoot the optimal values of the parameters (fig2)
* Oscillation around the right values of the parameters (fig3)

Figure 4 Good learning rate selection

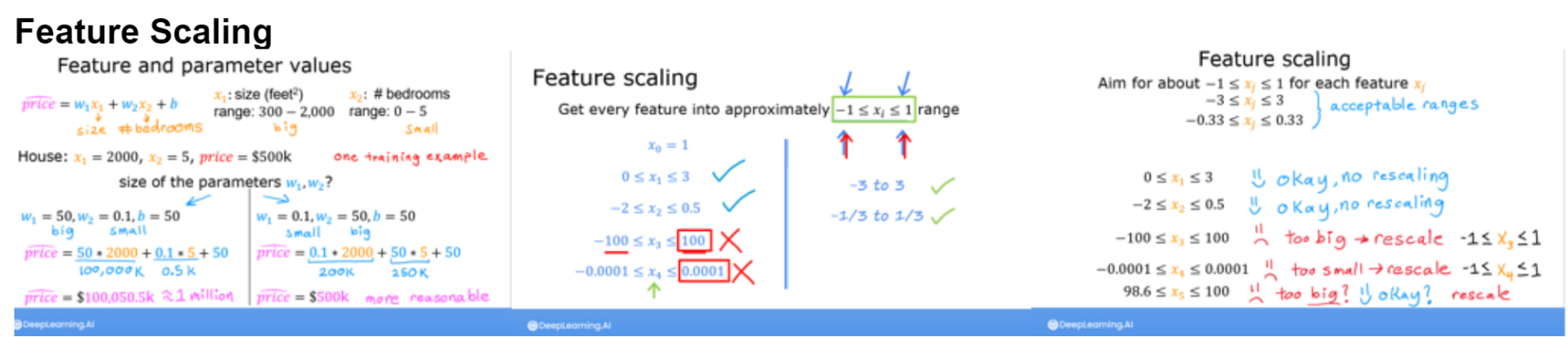


Figure 5 Feature scaling basics

Feature scaling issue indices :

* Some parameters converge way faster than other
* The gradient at each iteration of gradient descent is way larger for some parameters

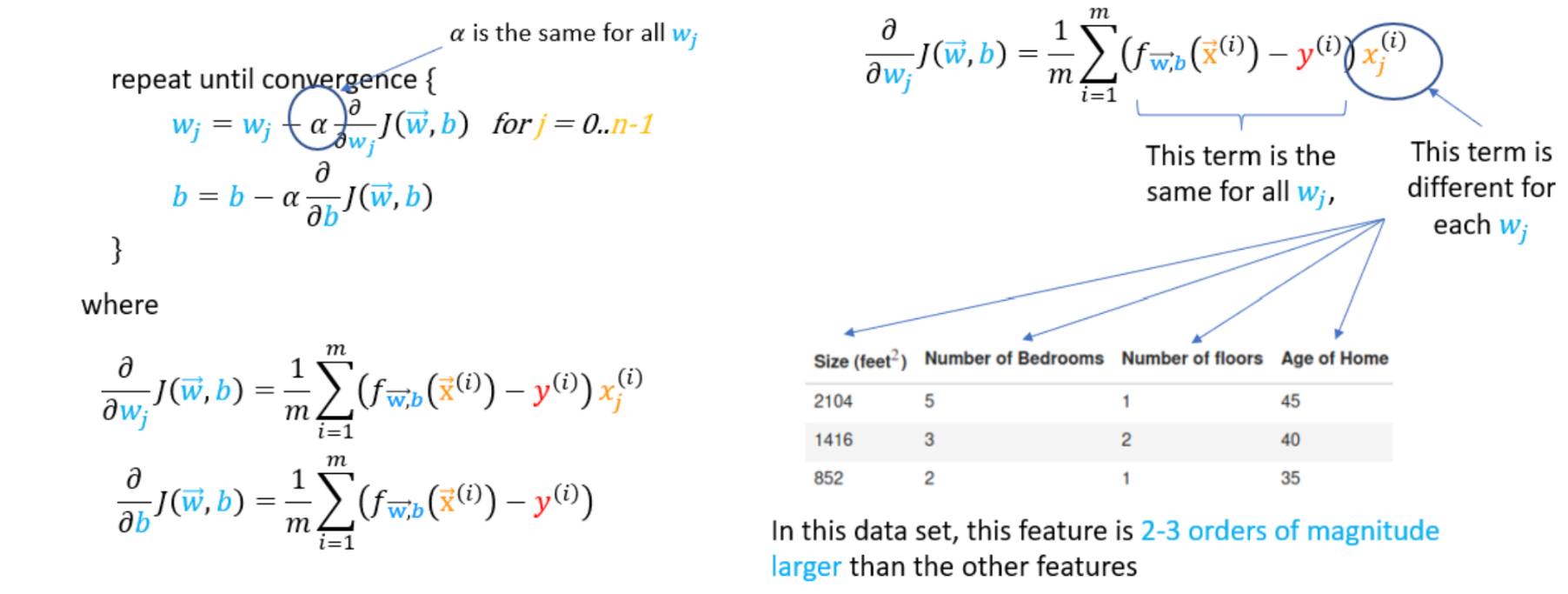
All parameters share the same learning rate. The error term is multiplied by the feature “j”. Each feature vary significantly in magnitude making some feature update faster than other. The solution is feature scalling.

Figure 6 Explanation of the difference between parameters scales

Three different techniques:

* Feature scaling, essentially dividing each positive feature by its maximum value, or more generally, rescale each feature by both its minimum and maximum values using (x-min)/(max-min). Both ways normalizes features to the range of -1 and 1, where the former method works for positive features which is simple and serves well for the lecture's example, and the latter method works for any features.
* Mean normalization: 𝑥𝑖:=𝑥𝑖−𝜇𝑖𝑚𝑎𝑥−𝑚𝑖𝑛xi:=xi−μimax−min
* Z-score normalization

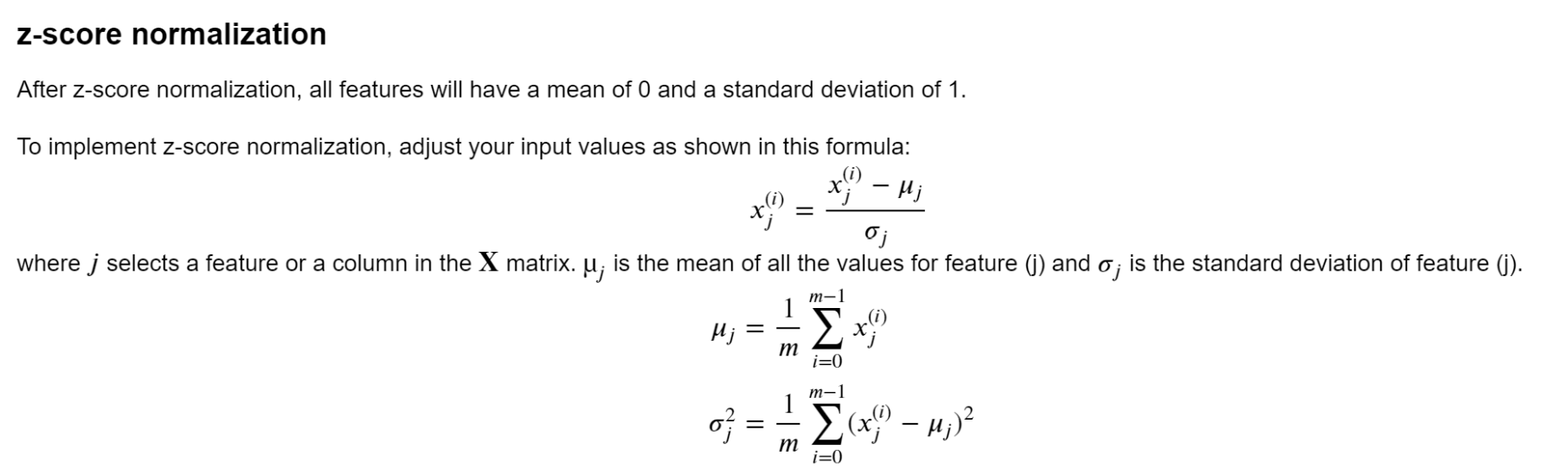


Figure 7 z-score normalization

**Implementation Note:** When normalizing the features, it is important to store the values used for normalization - the mean value and the standard deviation used for the computations. After learning the parameters from the model, we often want to predict the prices of houses we have not seen before. Given a new x value (living room area and number of bed- rooms), we must first normalize x using the mean and standard deviation that we had previously computed from the training set.

After feature scaling, you can use larger learning rate values, this will speed up the gradient descent. The scaled feature will get accurate results much faster

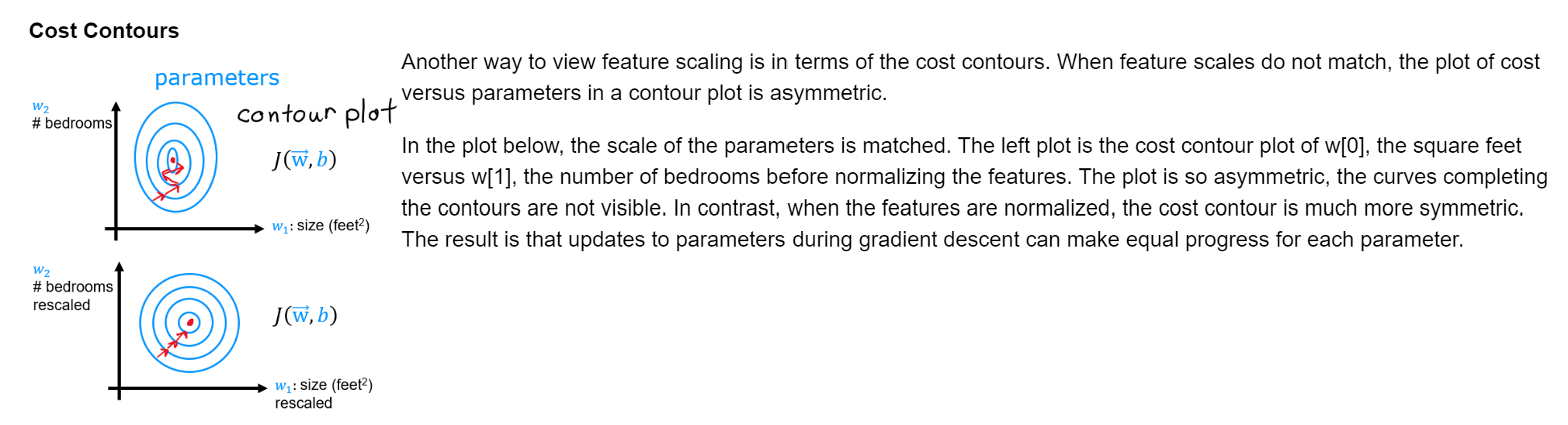


Figure 8 Cost contours

On the first plot, we can see that the parameters update is oscillating (for w1) even if the cost function value is decreasing. The update on the parameters w1 and w2 is not of the same magnitude at each iteration.

On the second plot, the is no oscillation, the updates are in the same directions and the parameters converge at the same rate smoothly.