Gradient Descent in Practice II: Learning Rate

Multivariate Linear Regression

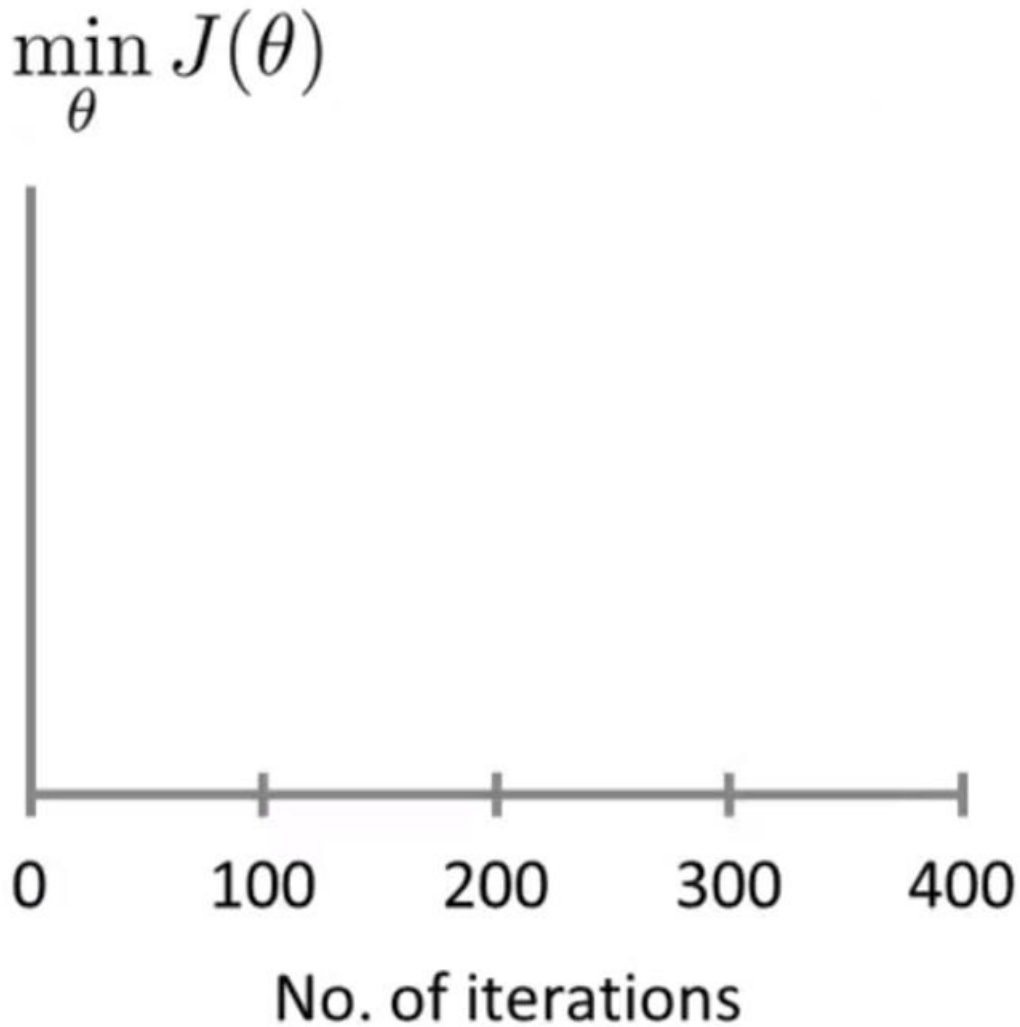
Linear Regression with Multiple Variables

Gradient descent

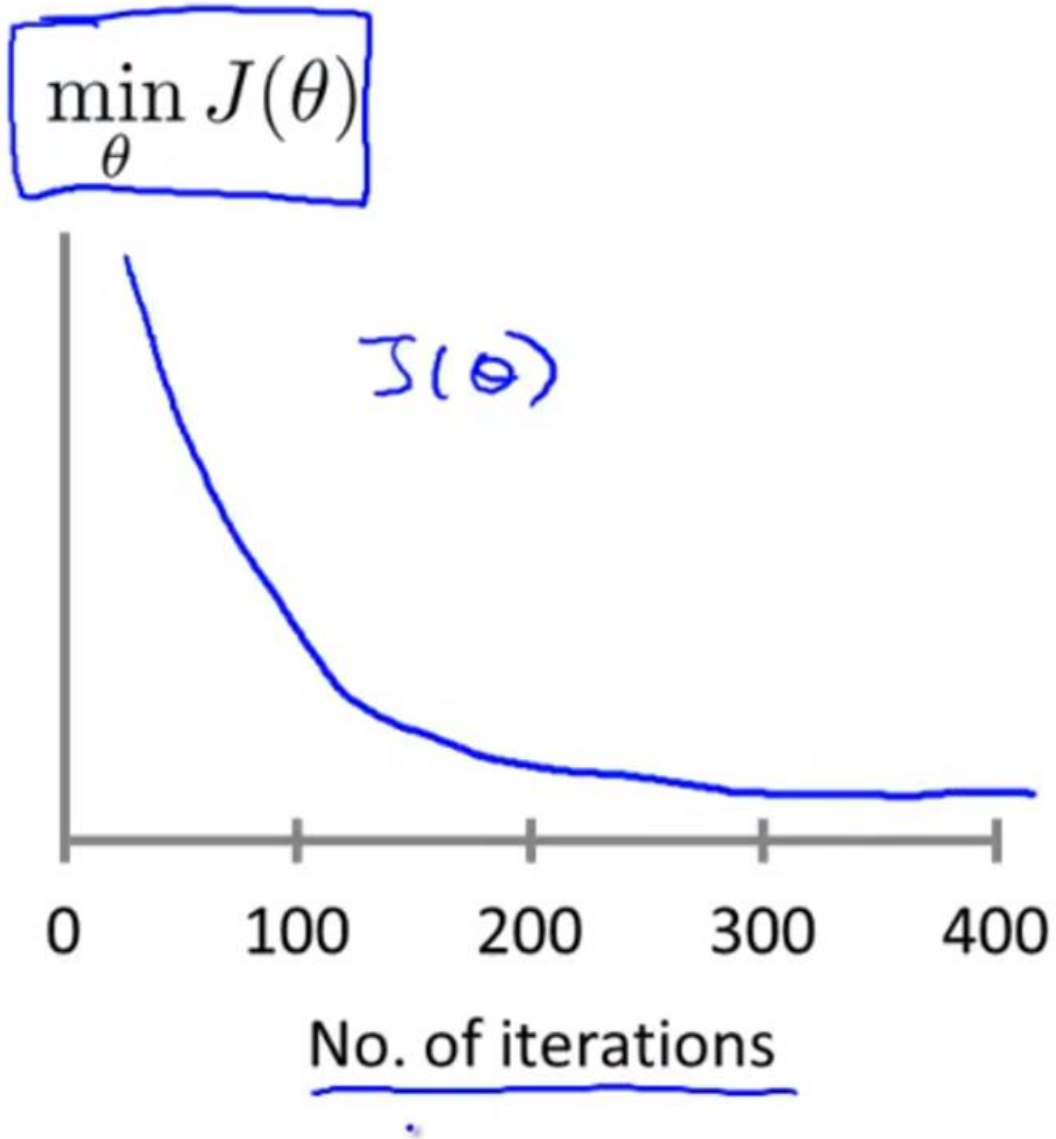
$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

- "Debugging": How to make sure gradient descent is working correctly.
- How to choose learning rate α.

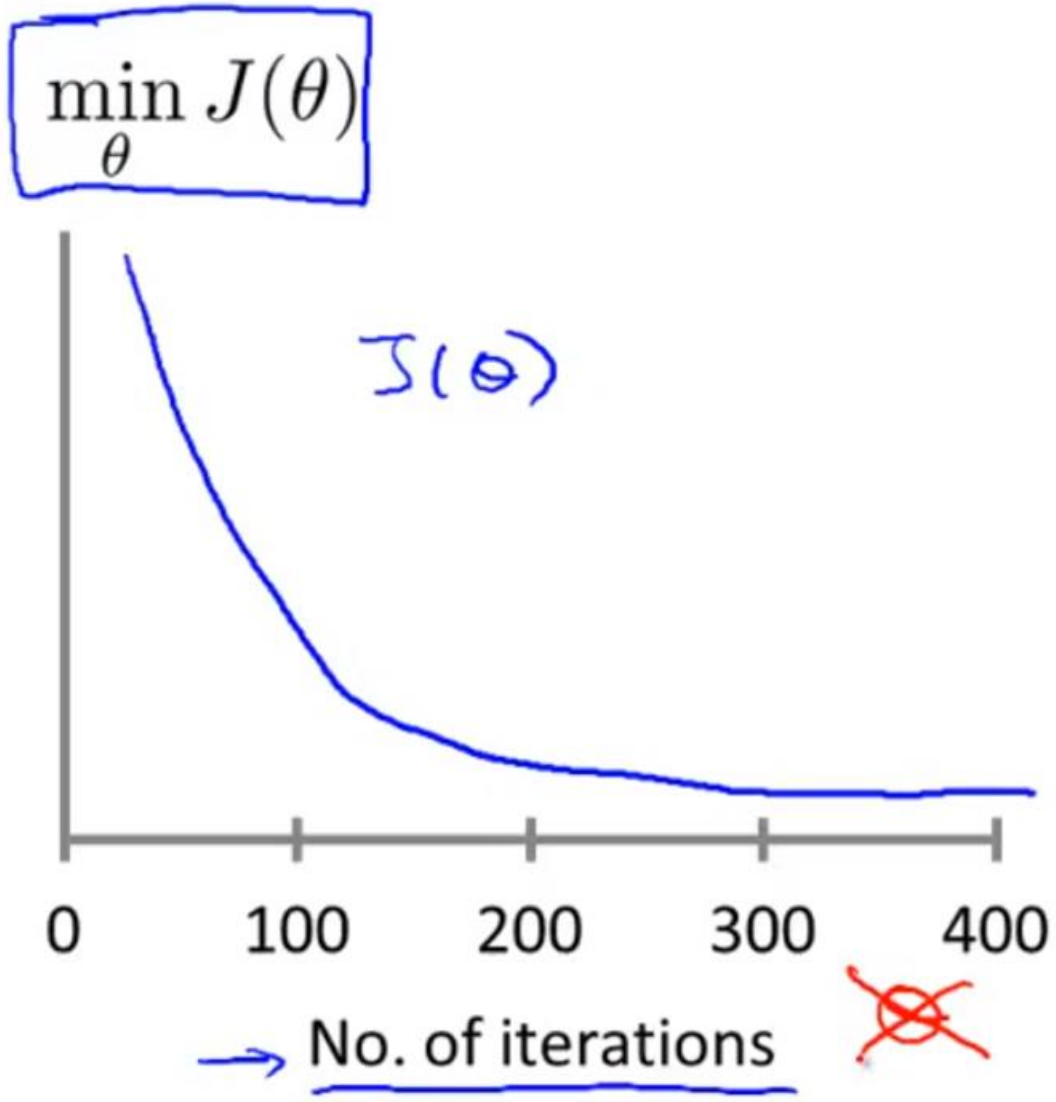
Making sure gradient descent is working correctly.



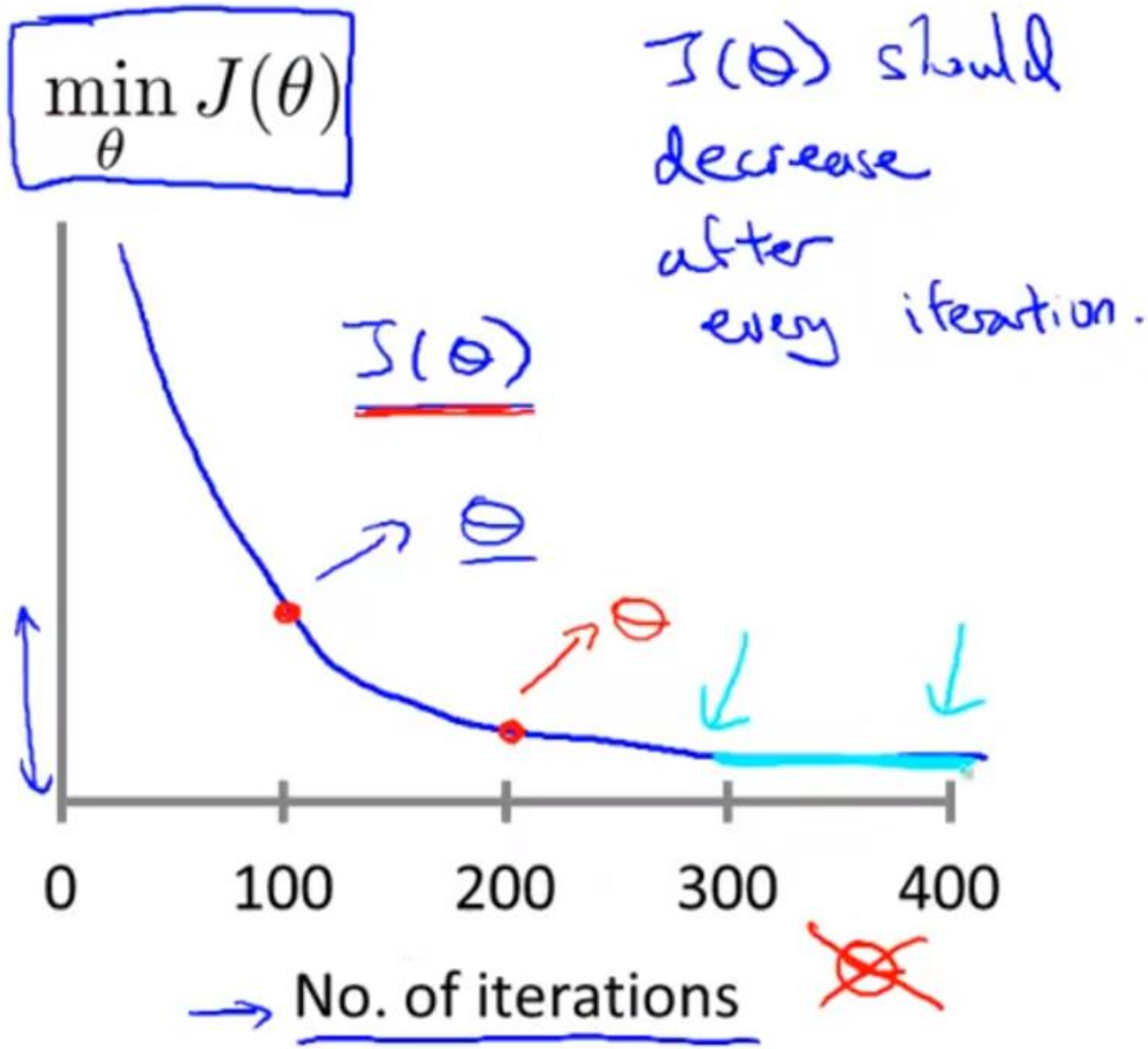
Making sure gradient descent is working correctly.



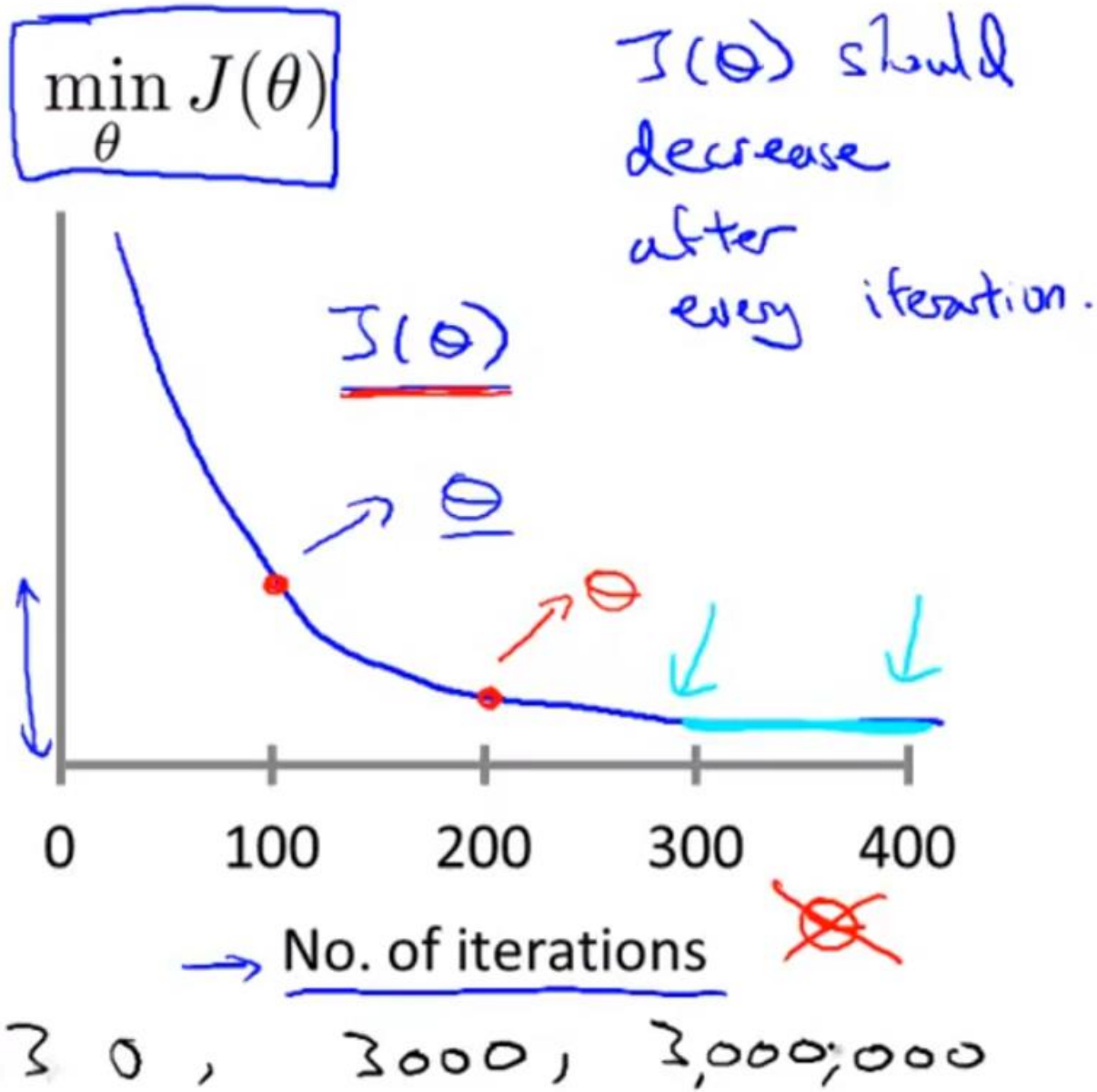
Making sure gradient descent is working correctly.



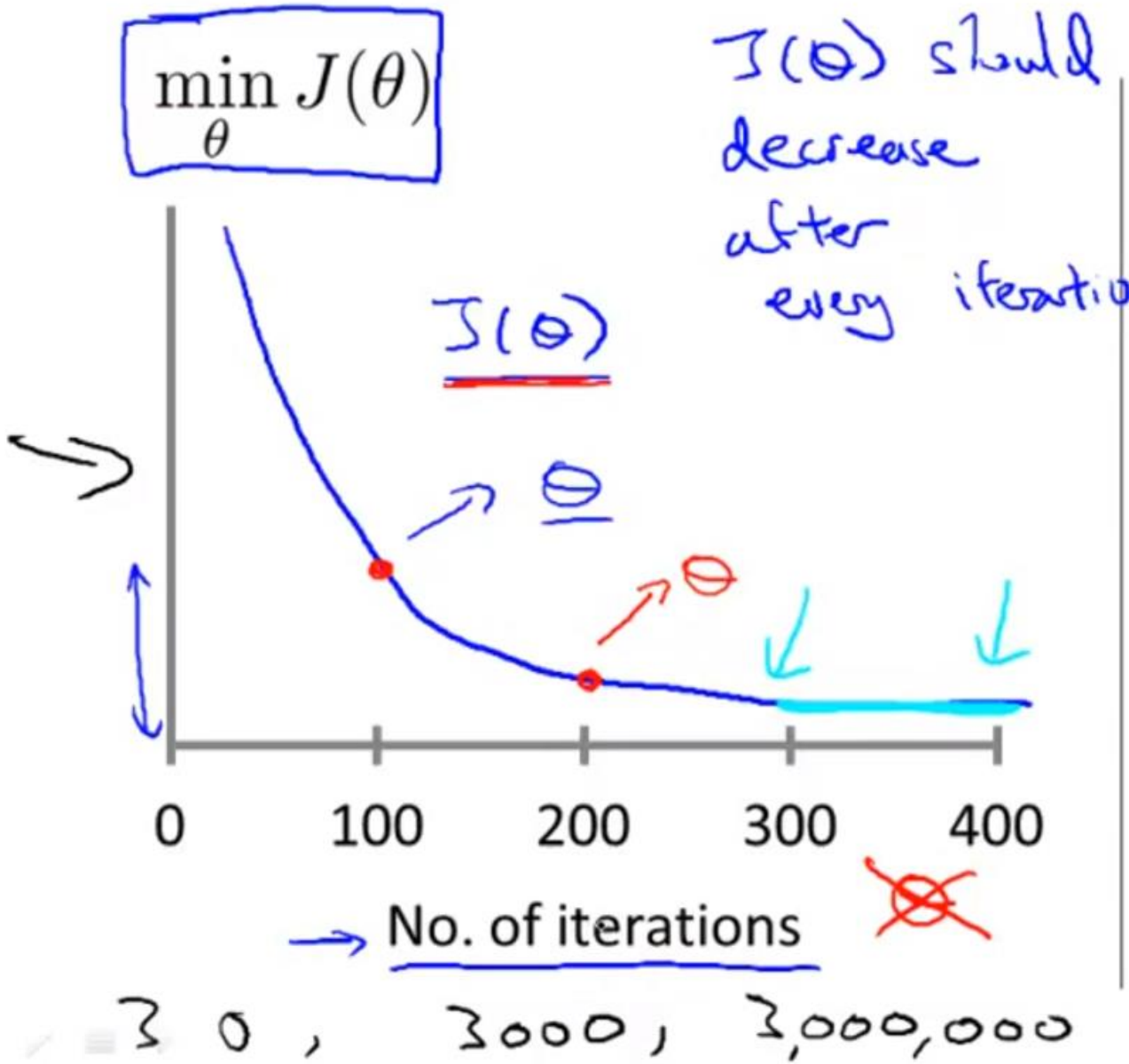
Making sure gradient descent is working correctly.



Making sure gradient descent is working correctly.



Making sure gradient descent is working correctly.

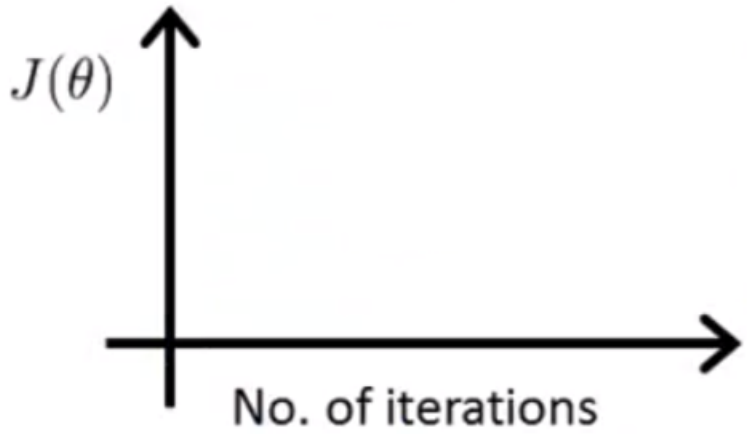


Example automatic convergence test:

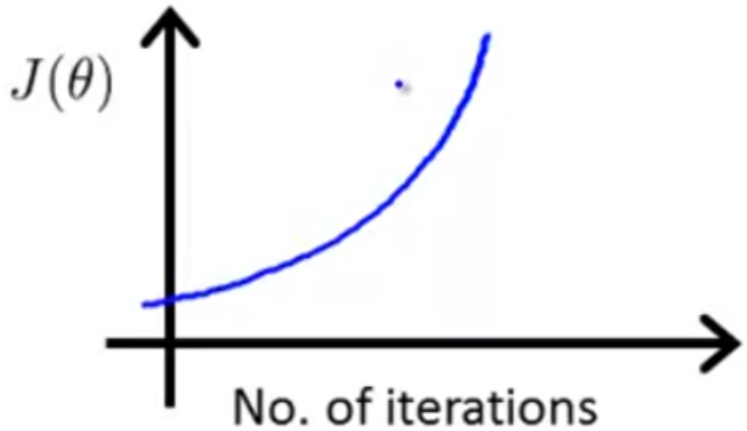
Declare convergence if $J(\theta)$ decreases by less than 10^{-3} in one iteration.

Epsilon

Making sure gradient descent is working correctly.



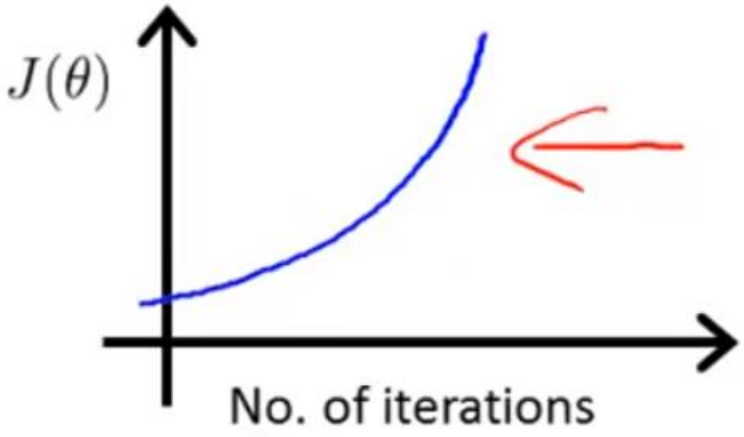
Making sure gradient descent is working correctly.



Gradient descent not working.

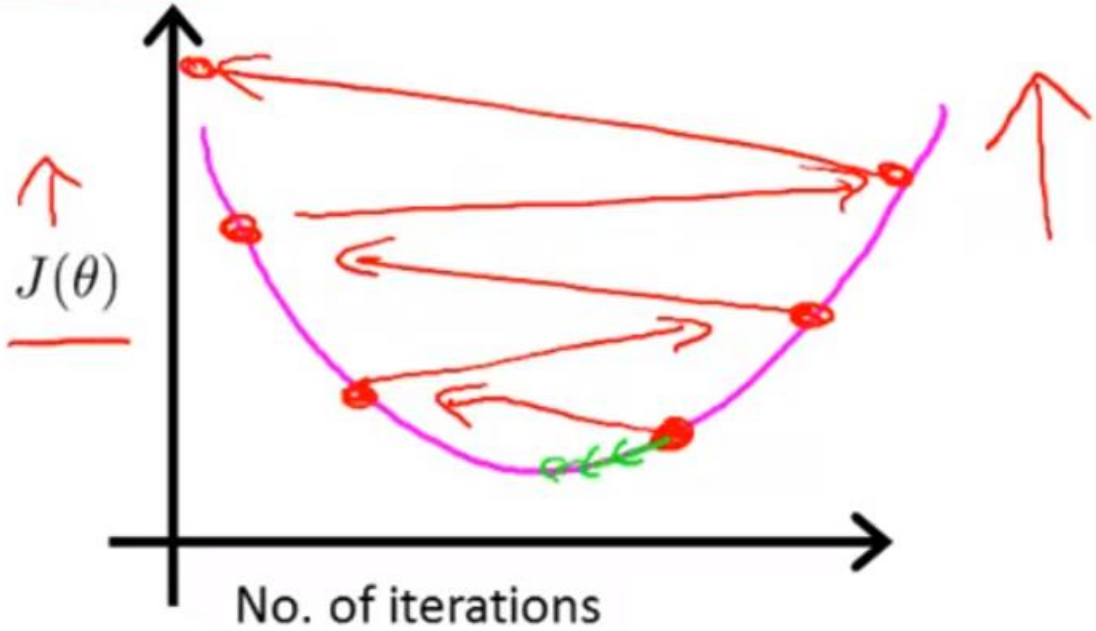
Use smaller α.

Making sure gradient descent is working correctly.

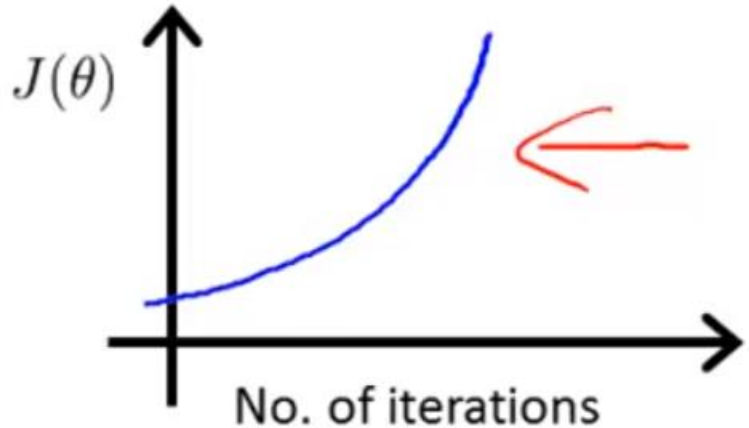


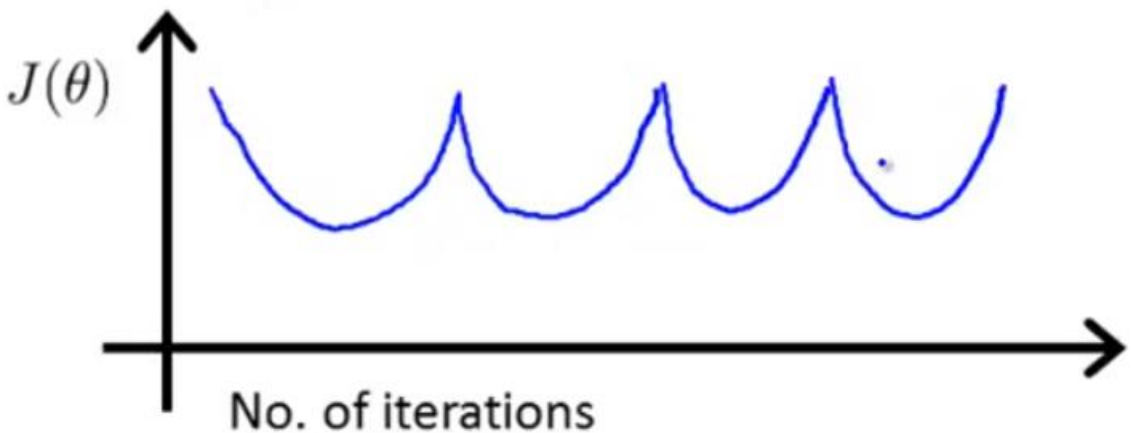
Gradient descent not working.

Use smaller α.



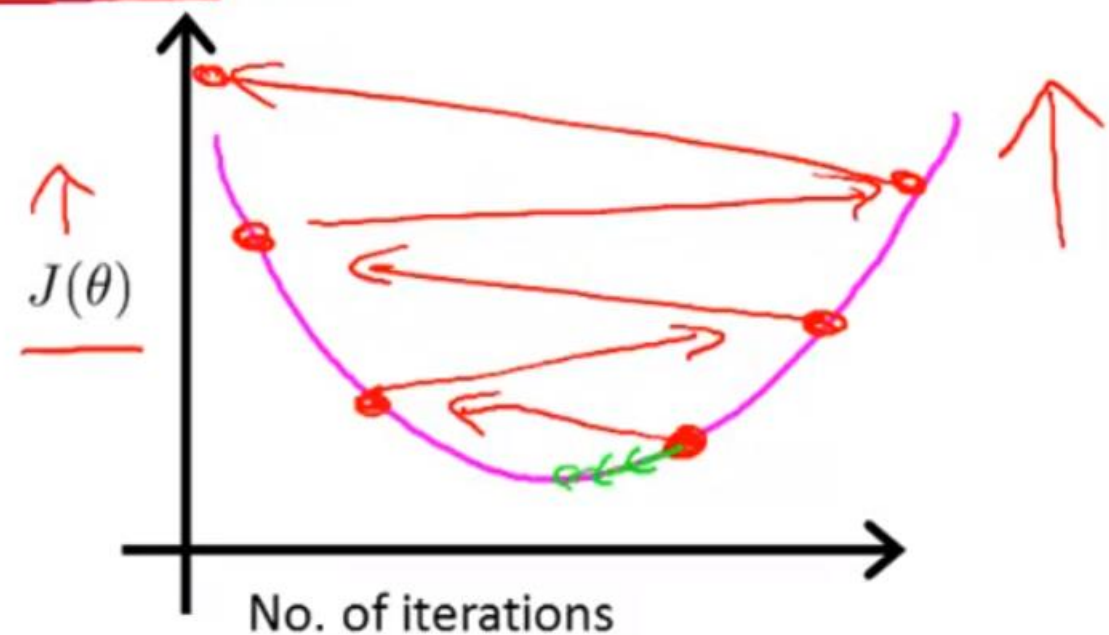
Making sure gradient descent is working correctly.



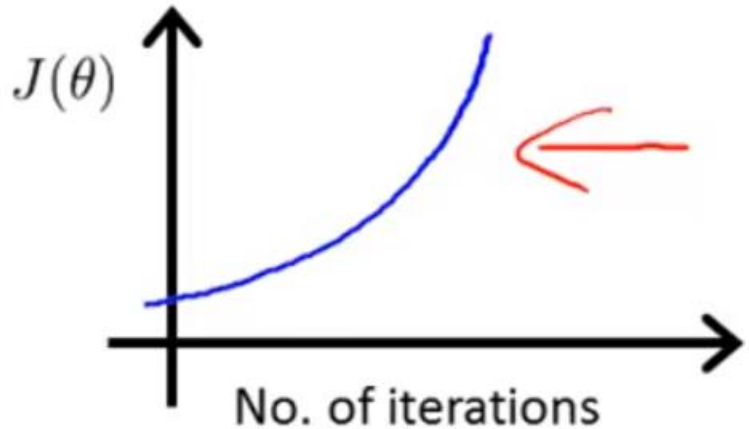


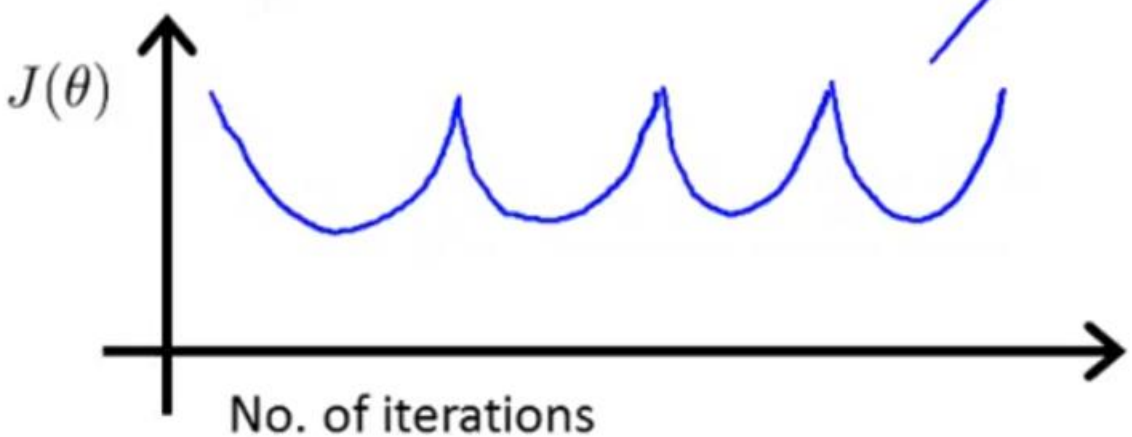
Gradient descent not working.

Use smaller α.



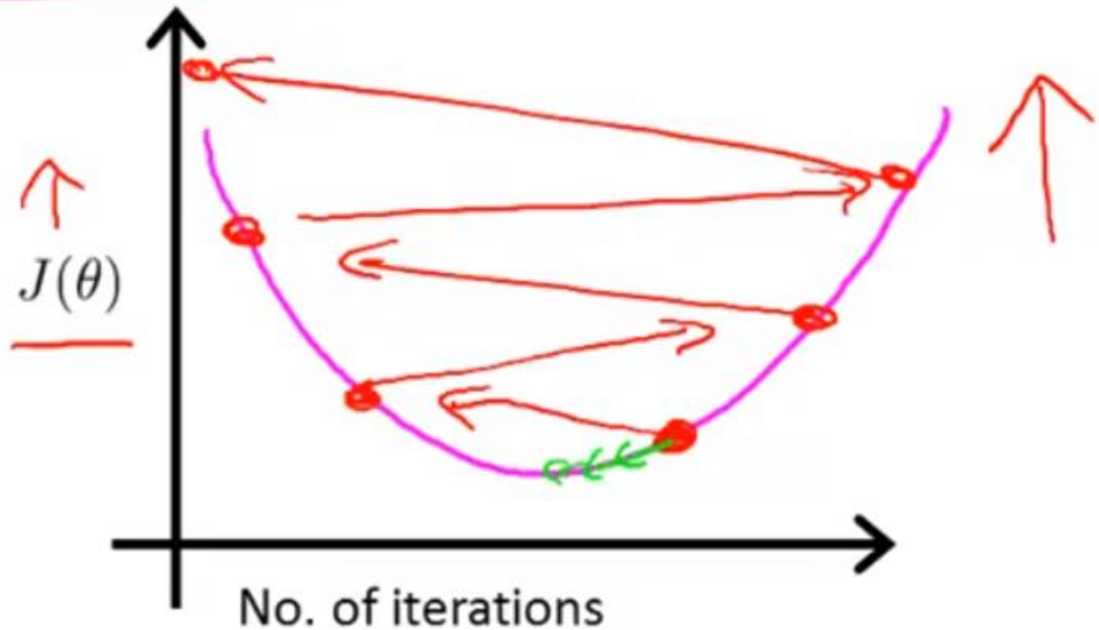
Making sure gradient descent is working correctly.



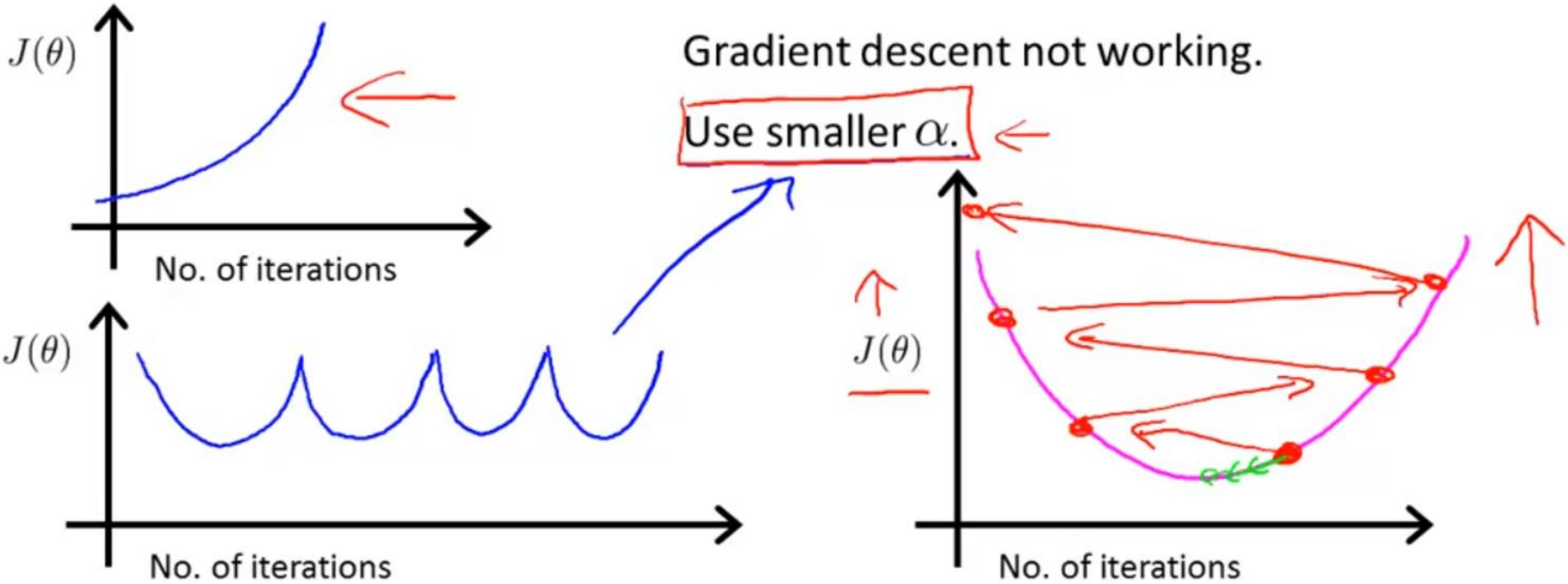


Gradient descent not working.

Use smaller α.



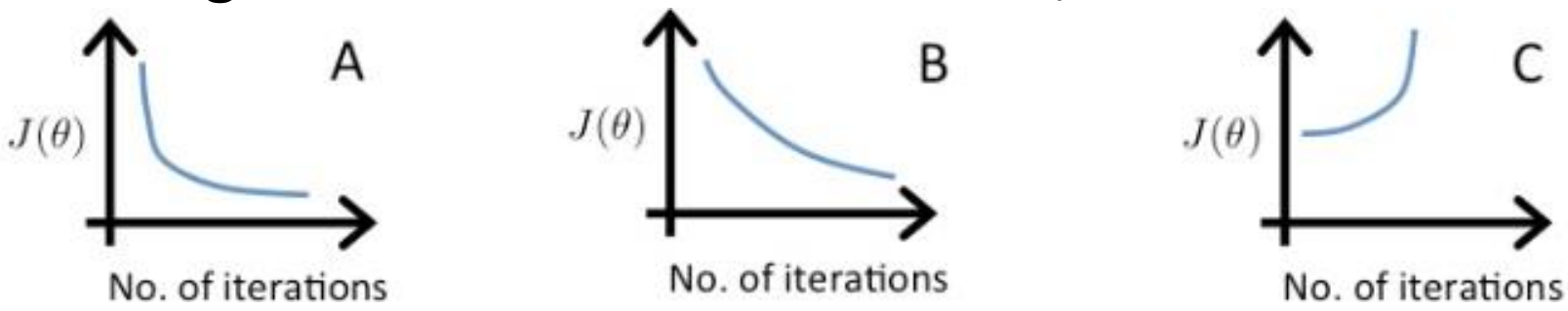
Making sure gradient descent is working correctly.



- For sufficiently small α, $J(\theta)$ should decrease on every iteration.
- But if α is too small, gradient descent can be slow to converge.

Exercise

- Suppose a friend ran gradient descent three times, with
- α=0.01,
- α=0.1,
- α=1,



- and got the following three plots (labeled A, B, and C):
- Which plots corresponds to which values of α?

	A	B	C
A	0.01	0.1	1
B	0.1	0.01	1
C	1	0.1	0.01
D	1	0.01	0.1

Summary:

- If α is too small: slow convergence.
- If α is too large: $J(\theta)$ may not decrease on every iteration; may not converge.

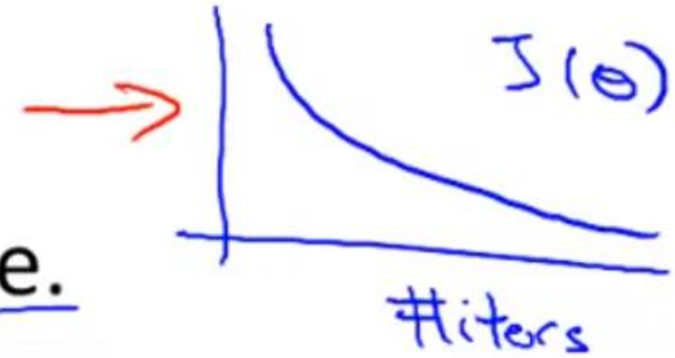
Summary:

- If α is too small: slow convergence.
- If α is too large: $J(\theta)$ may not decrease on every iteration; may not converge.

To choose α, try

$$\ldots, 0.001, \quad , 0.01, \quad , 0.1, \quad , 1, \ldots$$

Summary:



- If α is too small: slow convergence.
- If α is too large: $J(\theta)$ may not decrease on every iteration; may not converge. (Slow converge also possible.)

To choose α, try

$$\ldots, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1, \ldots$$