COMP2261 ARTIFICIAL INTELLIGENCE / MACHINE LEARNING

Gradient Descent

-- Intuition

Dr SHI Lei

Cost Function

$$J(\theta_0, \theta_1) = \frac{1}{2m}\sum_{i=1}^{m}(predicted_i - actual_i)^2 = \frac{1}{2m}\sum_{i=1}^{m}(\theta_0 + \theta_1 x^{(i)} - y^{(i)})^2$$

- Cost Function $J(\theta_0, \theta_1)$ is a function of θ_0 and θ_1, i.e.
  - θ_0 and θ_1 are the cost function's independent variables.
  - $x^{(i)}$ and $y^{(i)}$ are constants (training data).
- We want to find the θ_0, θ_1 pair, so that the cost function is minimised.
  - This pair of θ_0 and θ_1 will then be used as the parameters of our linear regression model.
- We want this process to be automatic, so we need to implement it in our learning algorithm.

Gradient Descent

Learning Objectives

- Understand what is gradient descent.
- Understand how gradient descent works.
- Understand what is learning rate and overshooting.
- Understand the pitfalls of the gradient descent.

Gradient Descent

General Idea

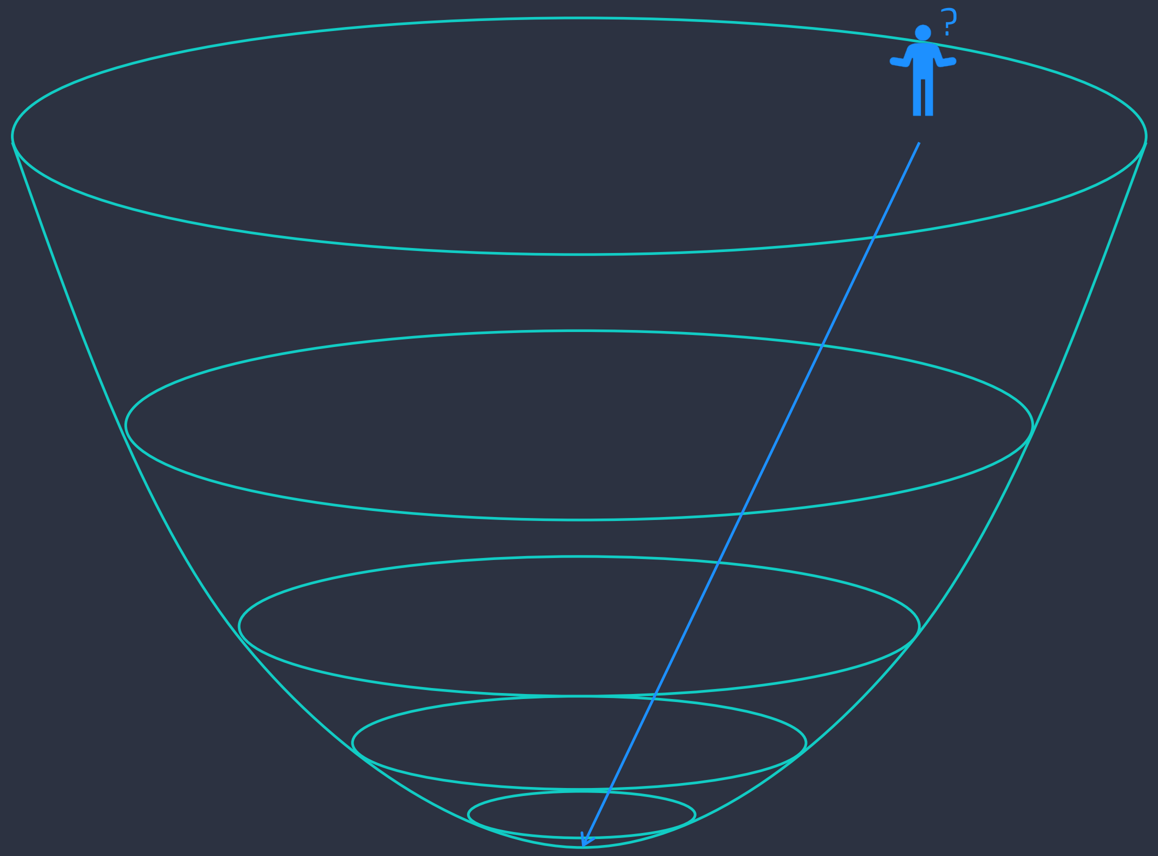
Try different pairs of θ_0 and θ_1 $J(\theta_0, \theta_1)$ to reduce the cost.

in iterations

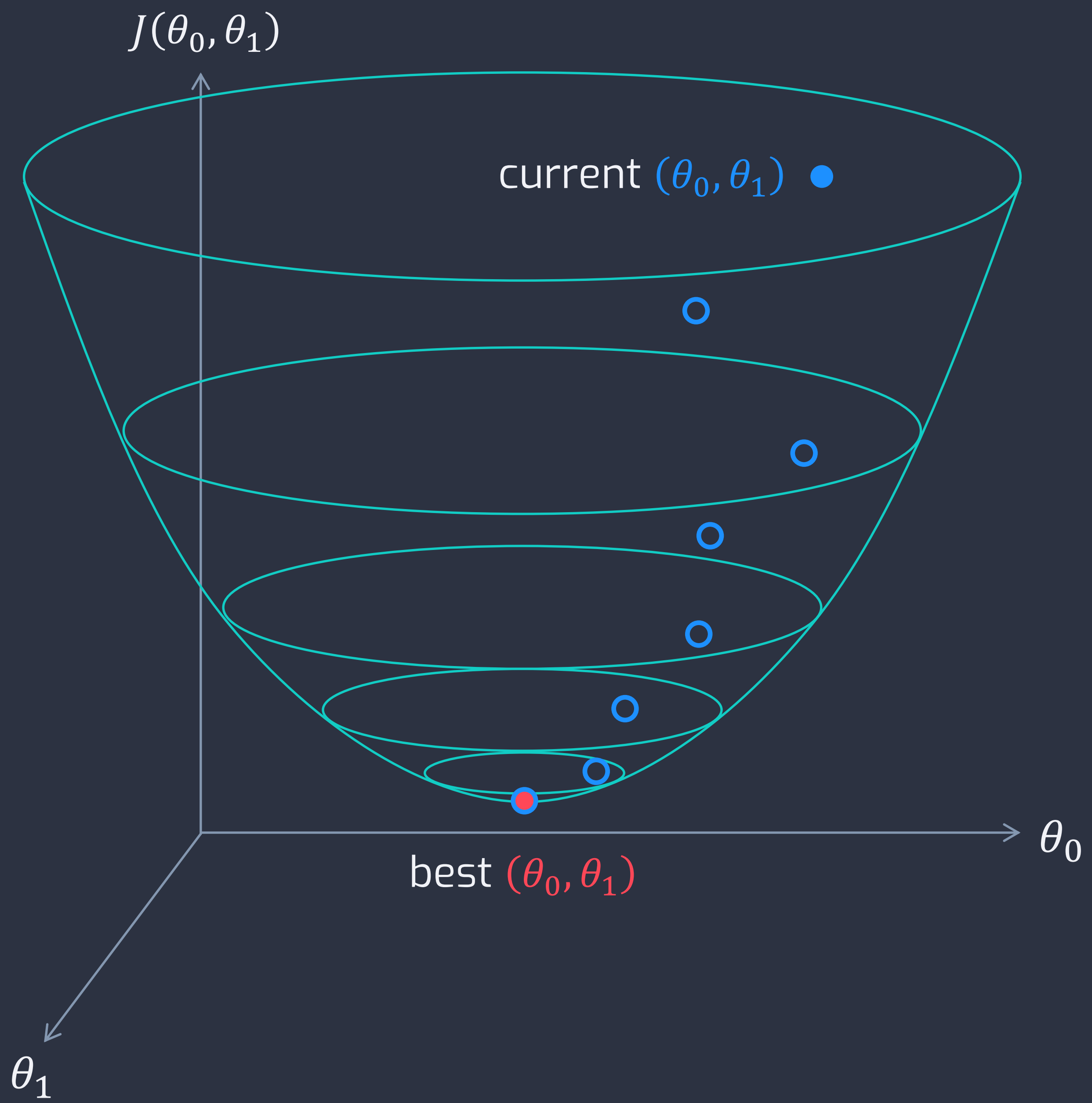
✓ Gradient Descent help us on how to select θ_0 and θ_1 to try and when to stop.

Intuition

Go downhill in the direction of the steepest slop.

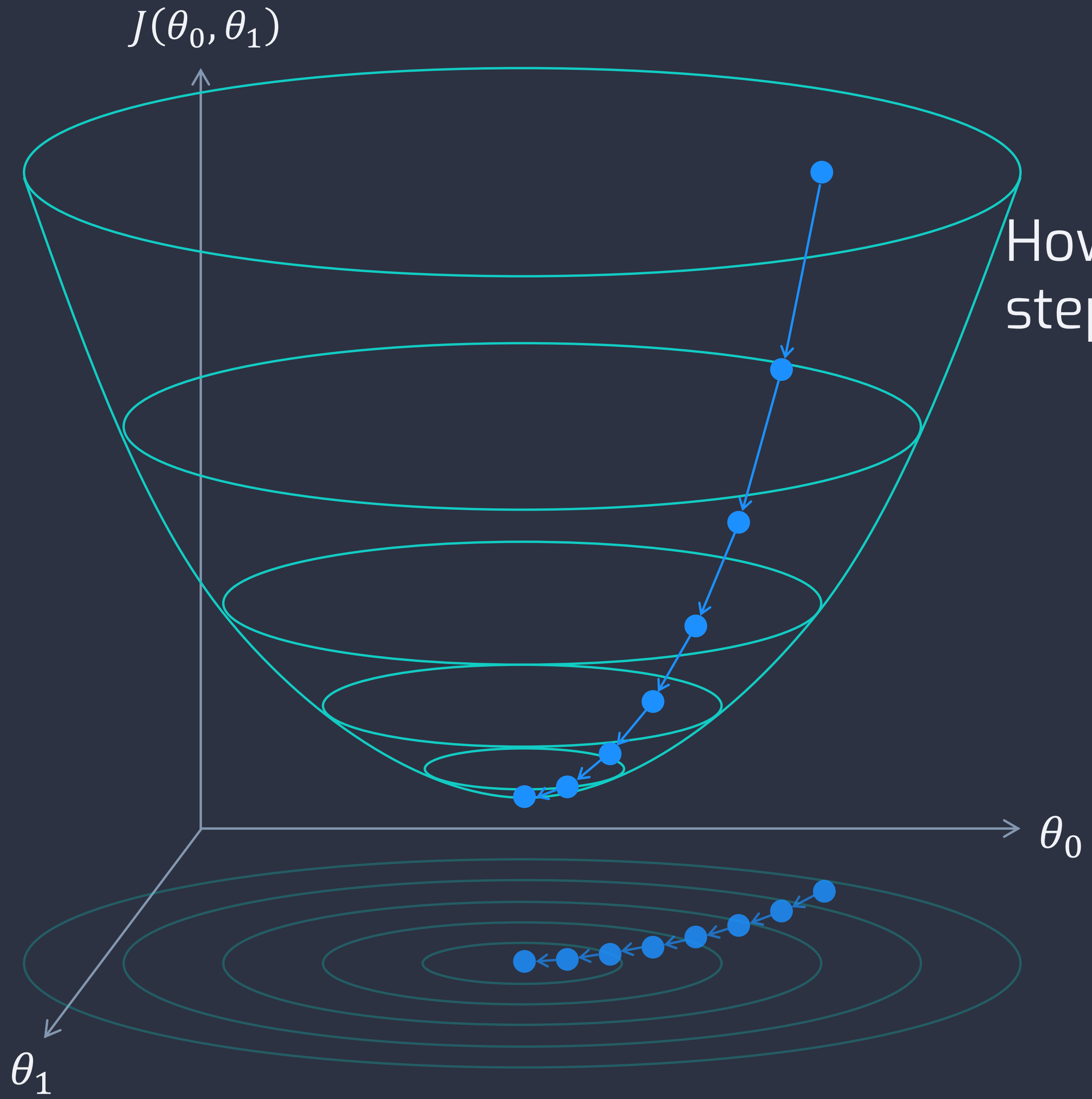


Intuition



$J(\theta_0, \theta_1)$
current (θ_0, θ_1)
best (θ_0, θ_1)
θ_0
θ_1

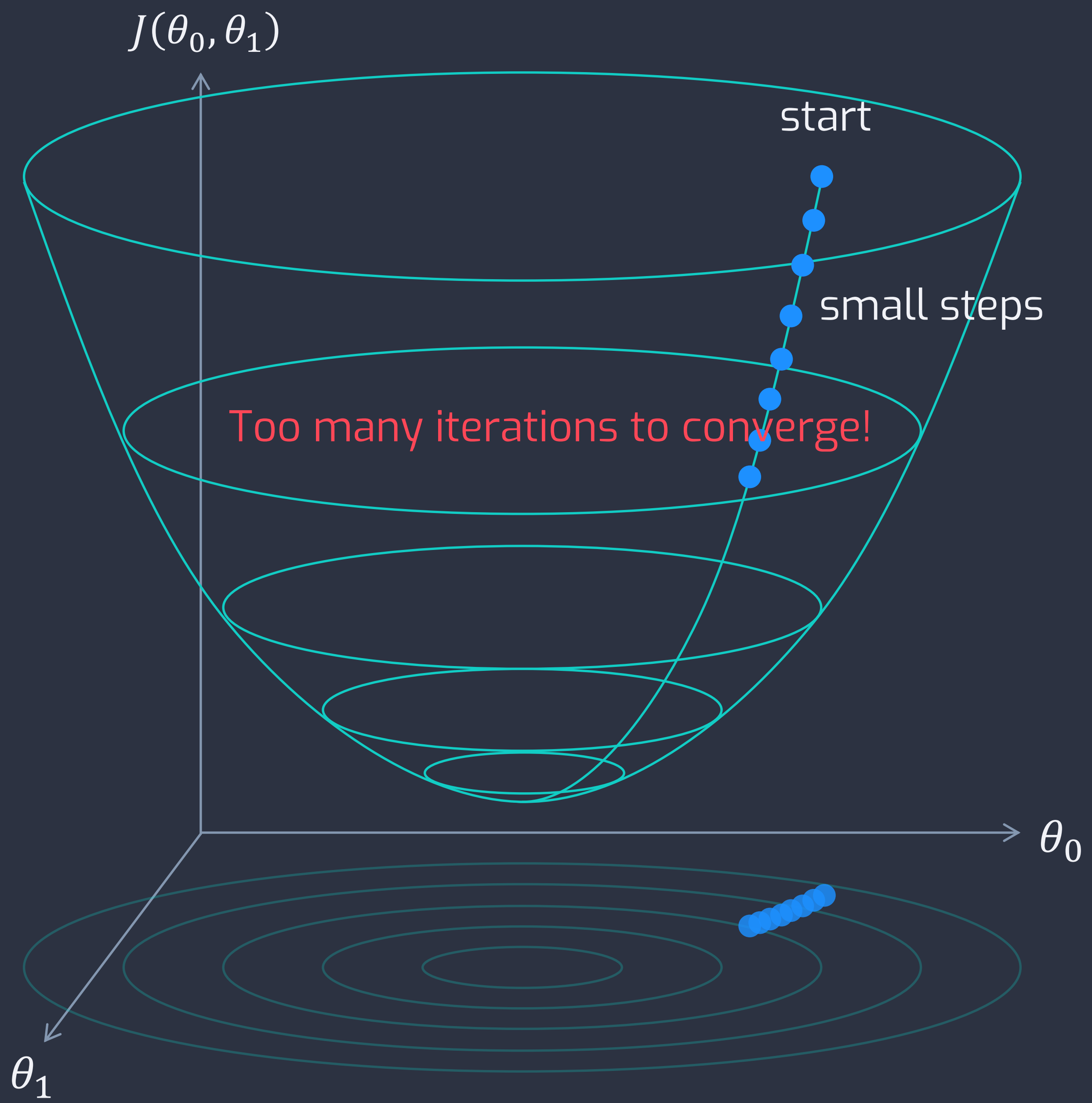
Intuition



How large /small a step should we take?

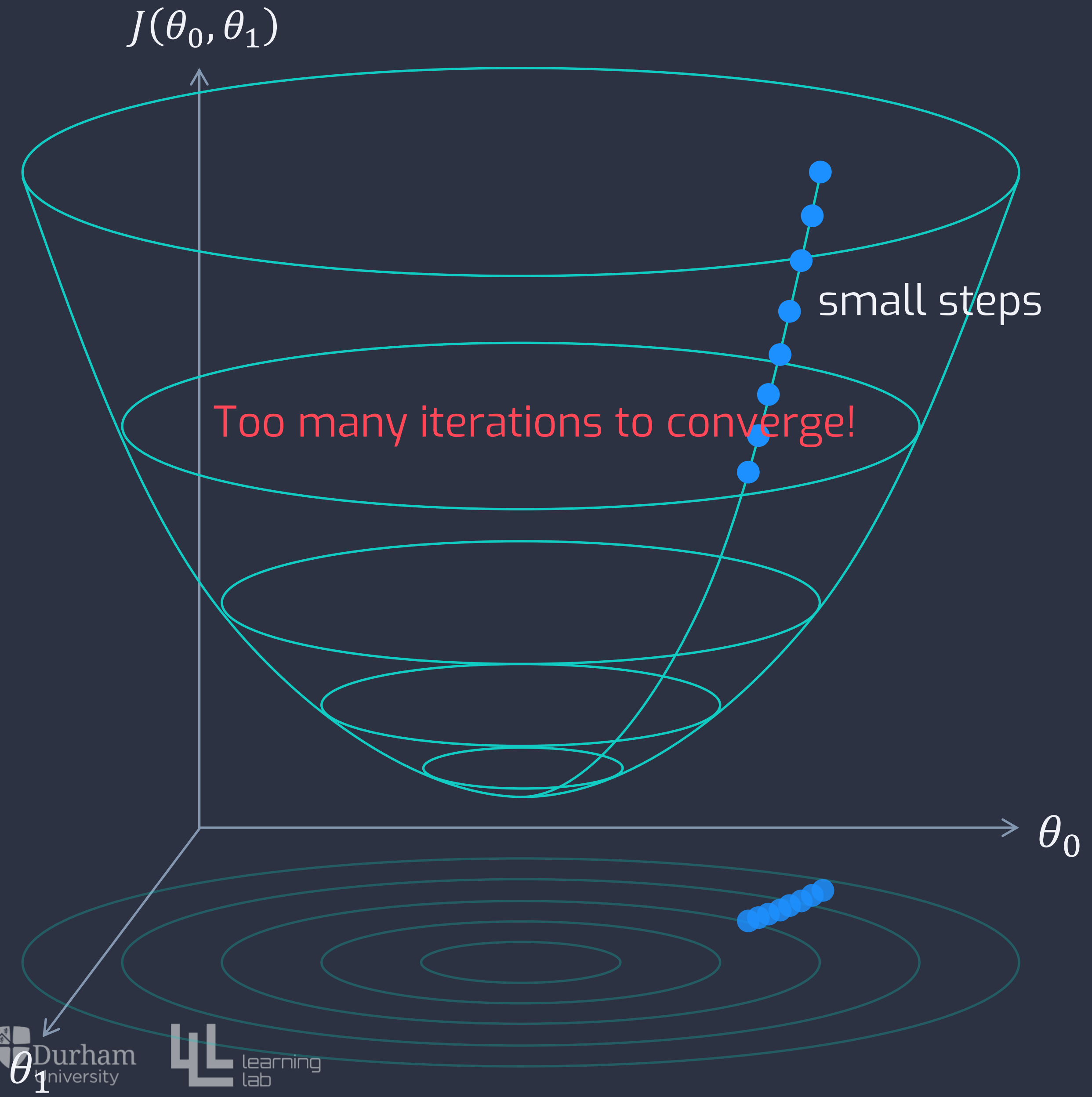
Intuition

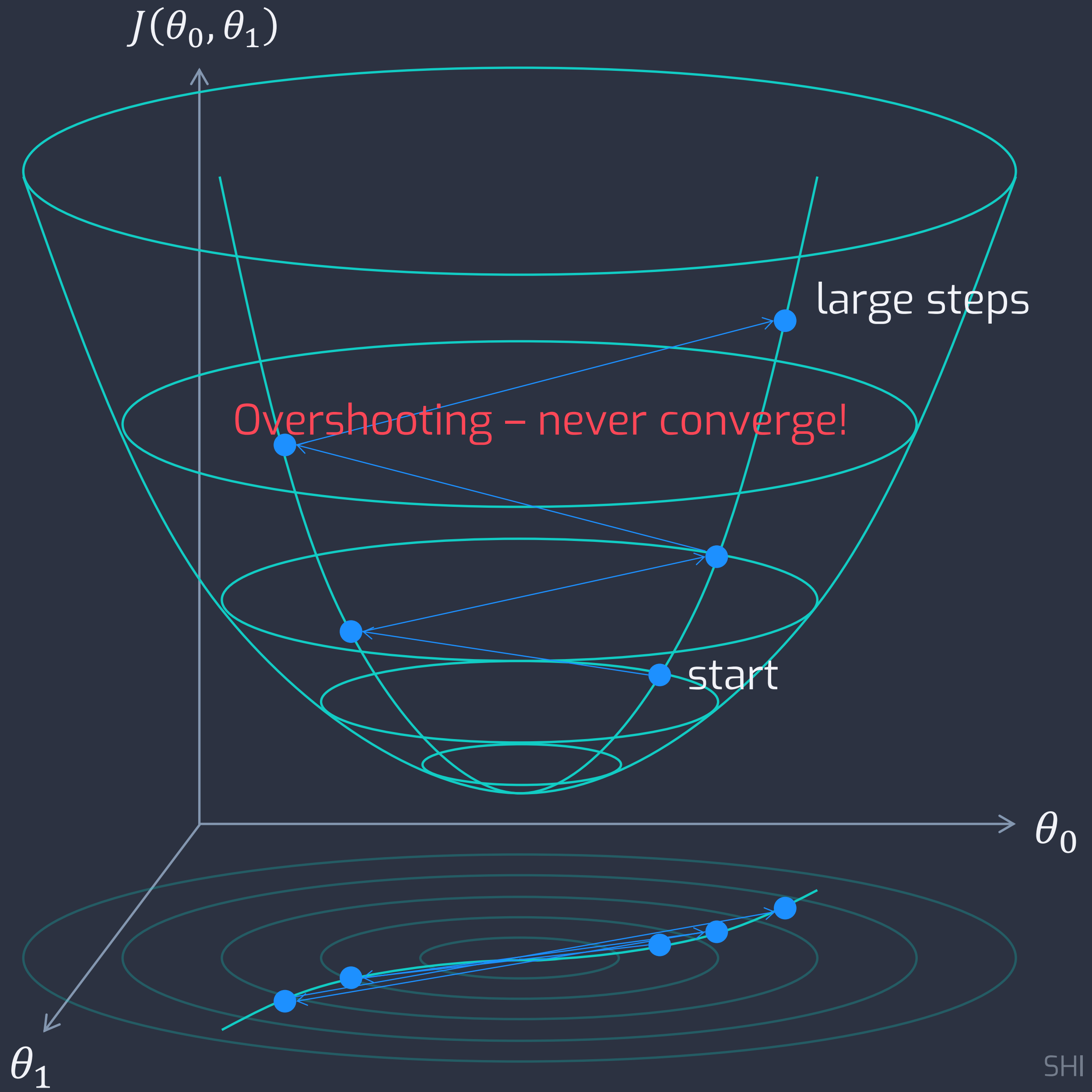
Hyperparameter: learning rate (size of step)



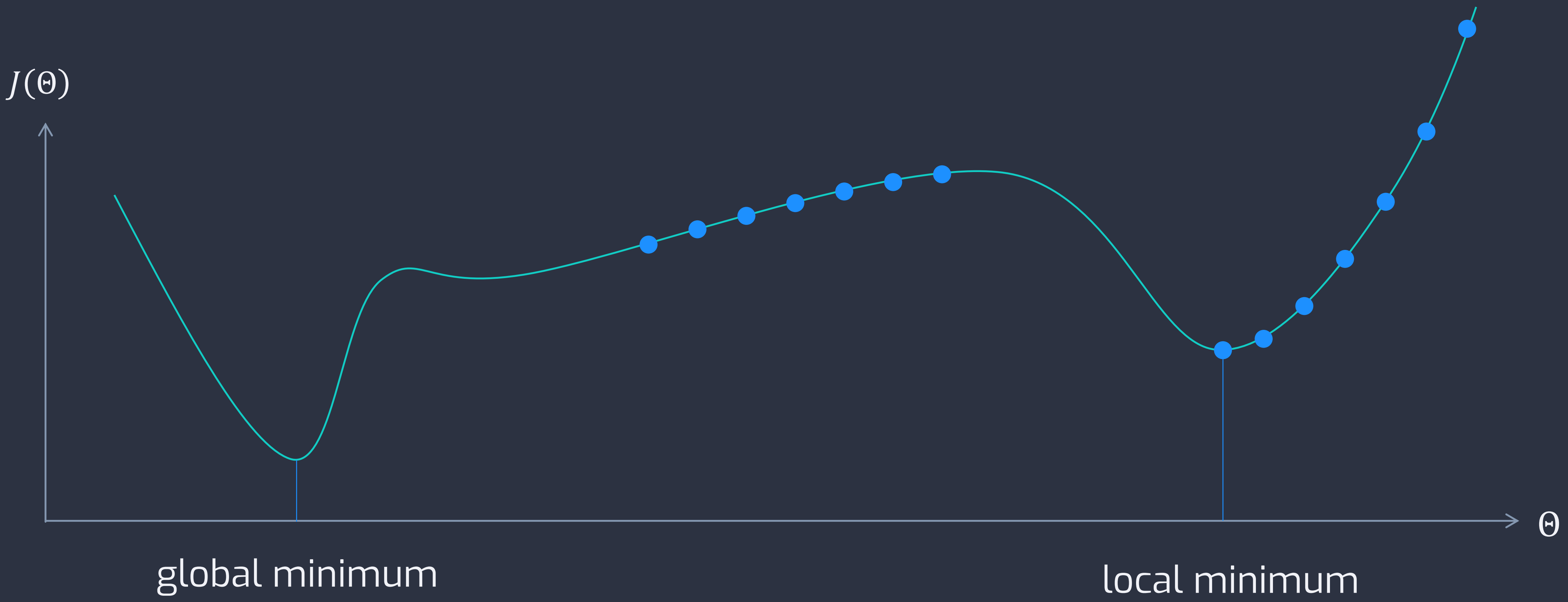
Intuition

Hyperparameter: learning rate (size of step)



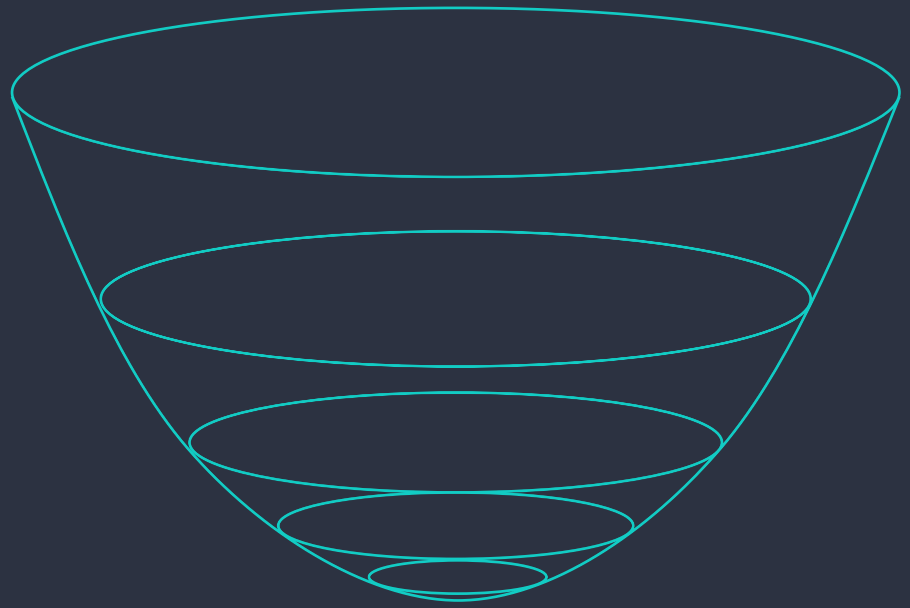


Pitfall



- Learning algorithm may end up with the local minimum not the global minimum.
- Learning algorithm may stops before reaching the minimum.

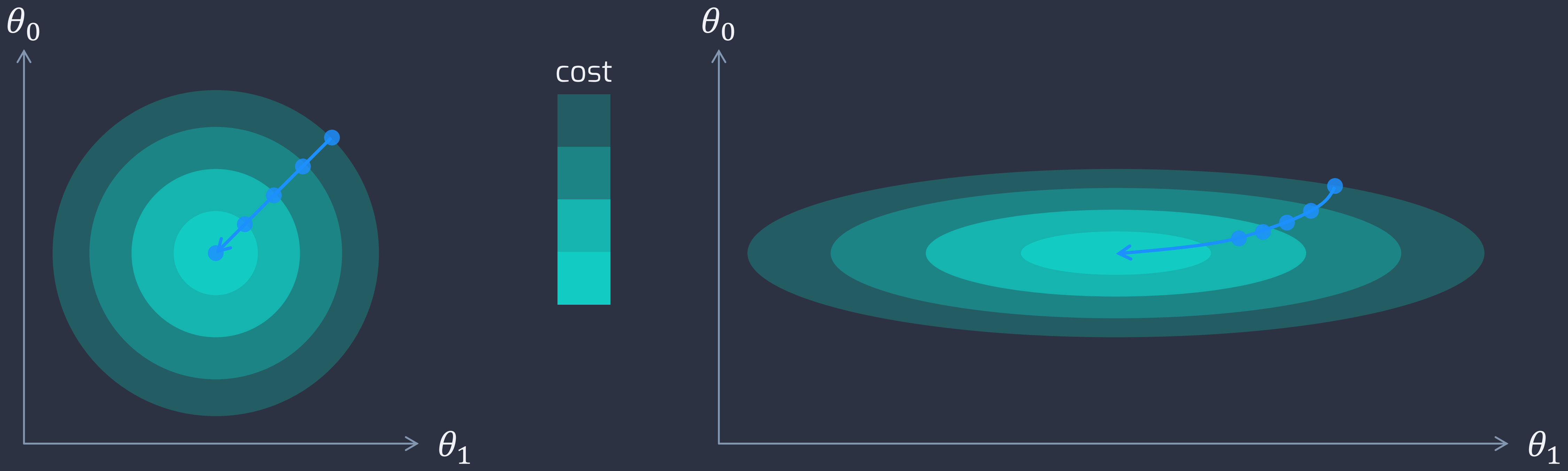
Pitfall



- MSE cost function for a linear regression model is always convex, meaning they are always shaped like a giant bowl with only one global minimum.
- As long as we start somewhere on the giant bowl and we take steps in reasonable sizes and we follow the gradient, it is guaranteed that we will eventually approach to the global minimum.

Pitfall

- If features in training set have very different scales, and this means the parameters of our model or the variables of our cost function have very different scales, the cost function will be in an elongated bowl shape, and so the learning algorithm will take much longer time to converge.



Gradient Descent with Feature Scaling

Gradient Descent without Feature Scaling

Takeaway Points

- Gradient Descent helps choose parameters to try.
- Appropriate learning rate (hyperparameter) to avoid taking to long for the learning algorithm to converge, or overshooting.
- Gradient Descent may end up with local minimum.
- MSE cost function for a linear regression model is always convex.