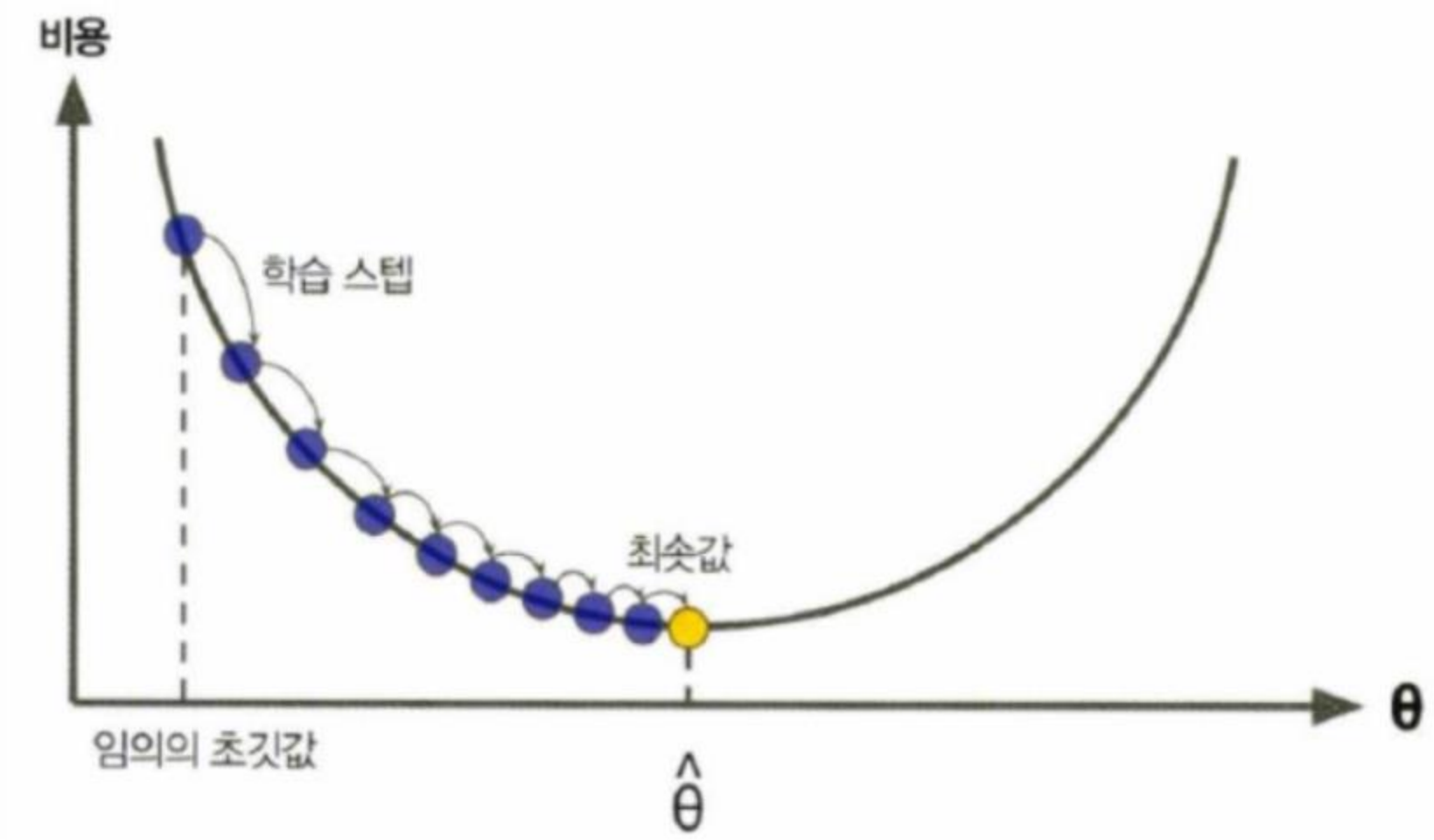
Gradient Descent

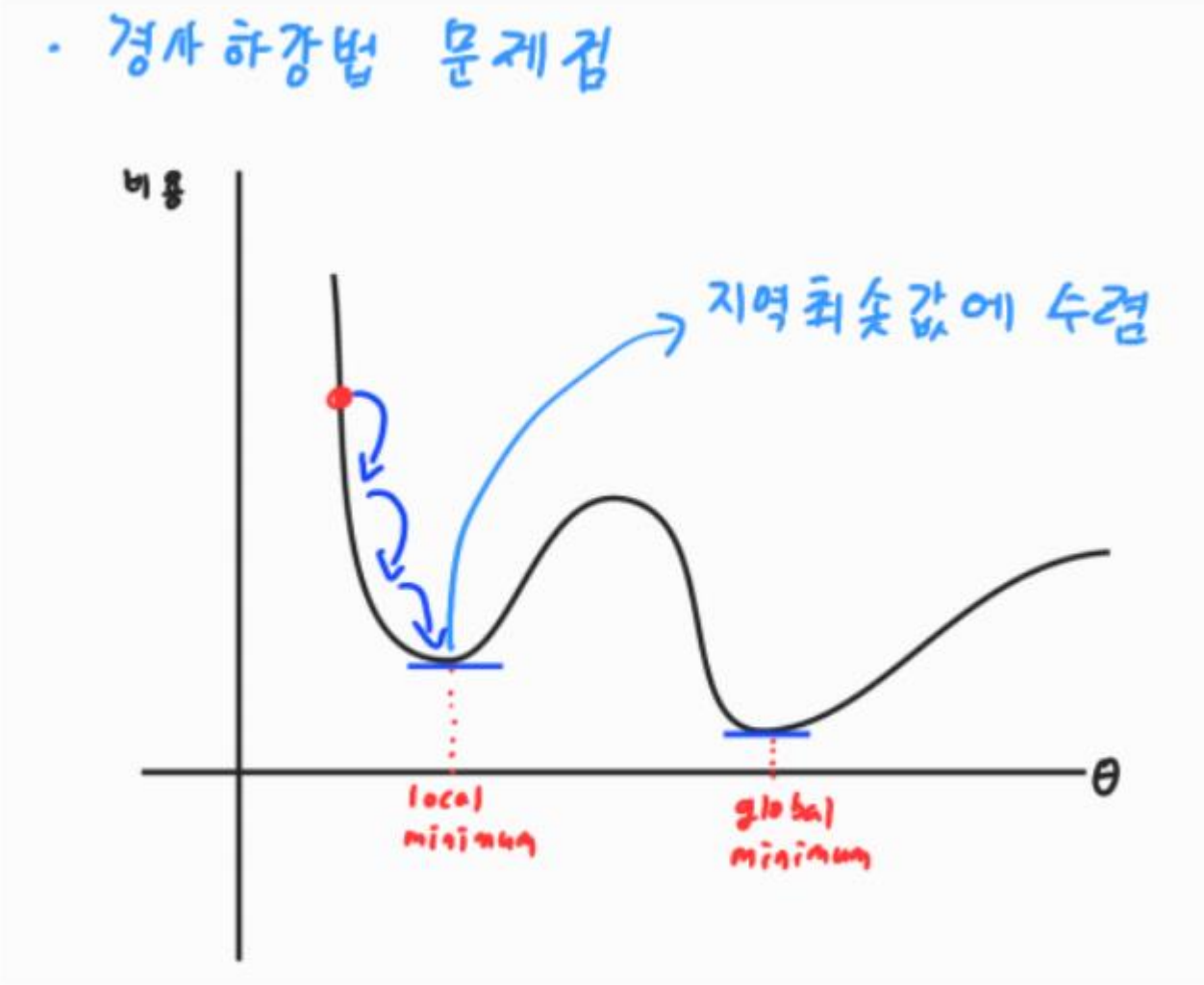
Gradient Descent



GD

최솟값 theta hat을 찾기 위해 step만큼을 내려간다

Gradient Descent의 문제점

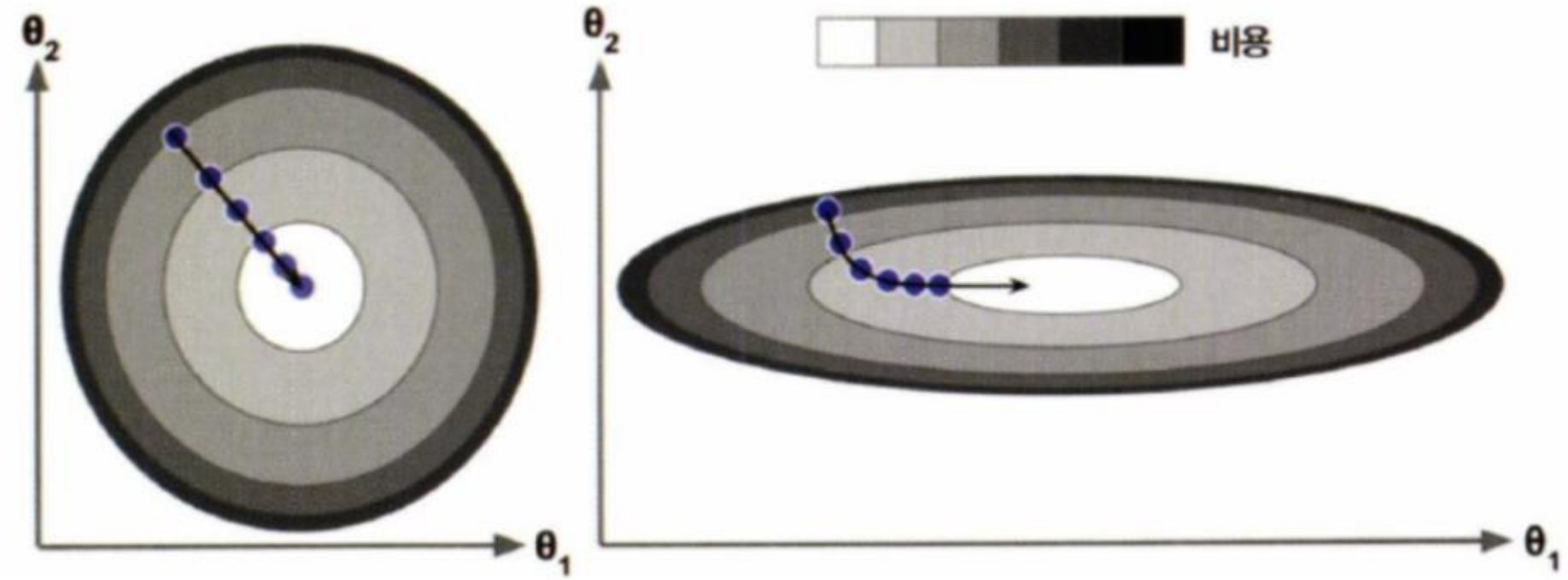


GD의 문제점

$$\frac{1}{m}\sum_{i=1}^{m}\left(\theta^T \mathbf{x}^{(i)} - y^{(i)}\right)^2$$

MSE는 볼록함수 이므로 지역 최솟값이 존재 하지 않는다.

Gradient Descent 주의점



스케일링 적용

스케일링 미적용

Batch Gradient Descent

전체 훈련세트 대한 그라디언트 계산을 함으로써 계산이 느림

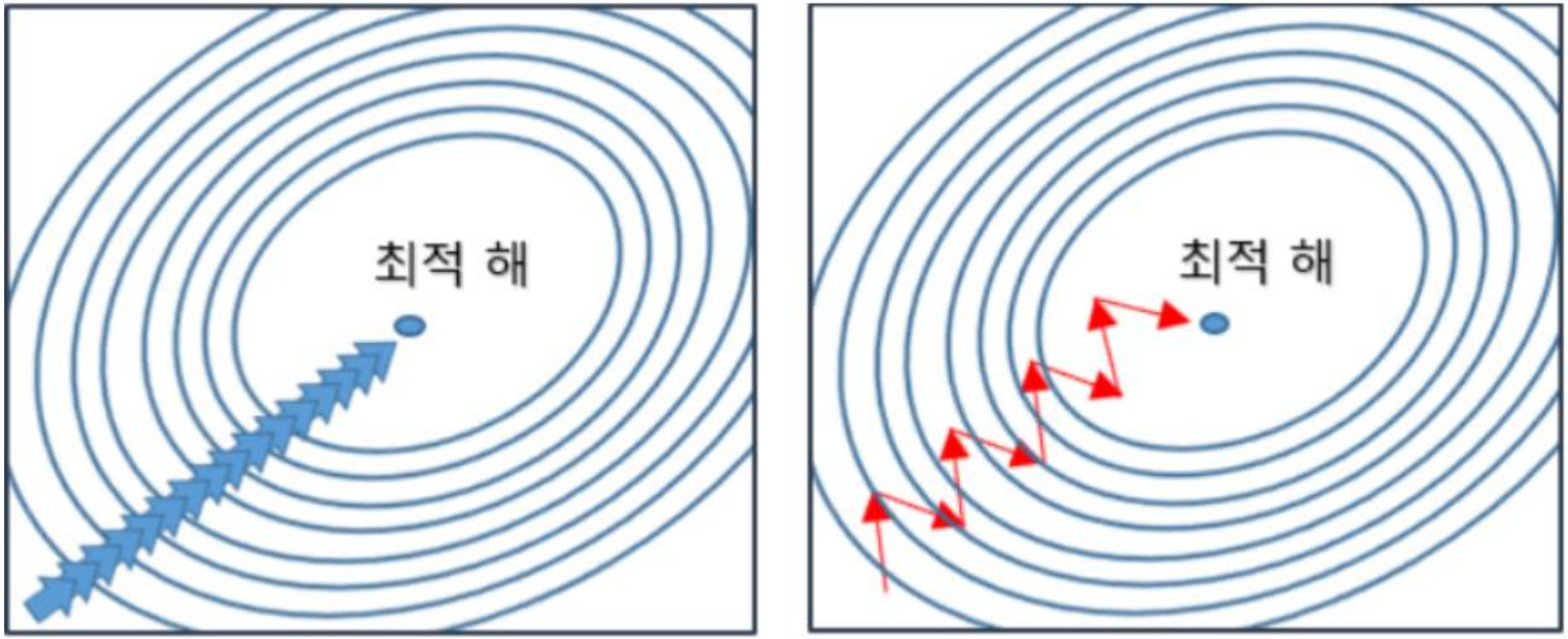
$$\nabla_\theta \mathrm{MSE}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial}{\partial \theta_0} \mathrm{MSE}(\boldsymbol{\theta}) \\ \frac{\partial}{\partial \theta_1} \mathrm{MSE}(\boldsymbol{\theta}) \\ \vdots \\ \frac{\partial}{\partial \theta_n} \mathrm{MSE}(\boldsymbol{\theta}) \end{pmatrix} = \frac{2}{m} \mathbf{X}^T (\mathbf{X}\boldsymbol{\theta} - \mathbf{y})$$

비용 함수의 그라디언트 벡터

Stochastic Gradient Descent

데이터 중 무작위로 하나 선택하고 그 샘플에 대한 그라디언트를 계산

- 확률적(=무작위)이기 때문에 불안정하다. 따라서 반복 횟수를 높여야한다.
- 지역 최솟값을 뛰어넘을 수 있으므로 전역 최솟값을 찾을 확률이 높다(than batch)



경사 하강법 확률적 경사 하강법

Mini Batch Gradient Descent

데이터 중 하나가 아닌 작은 배치(=묶음)으로 그라디언트 계산

- SGD보다 안정적.
- 지역 최솟값에서 못 벗어날 가능성 존재

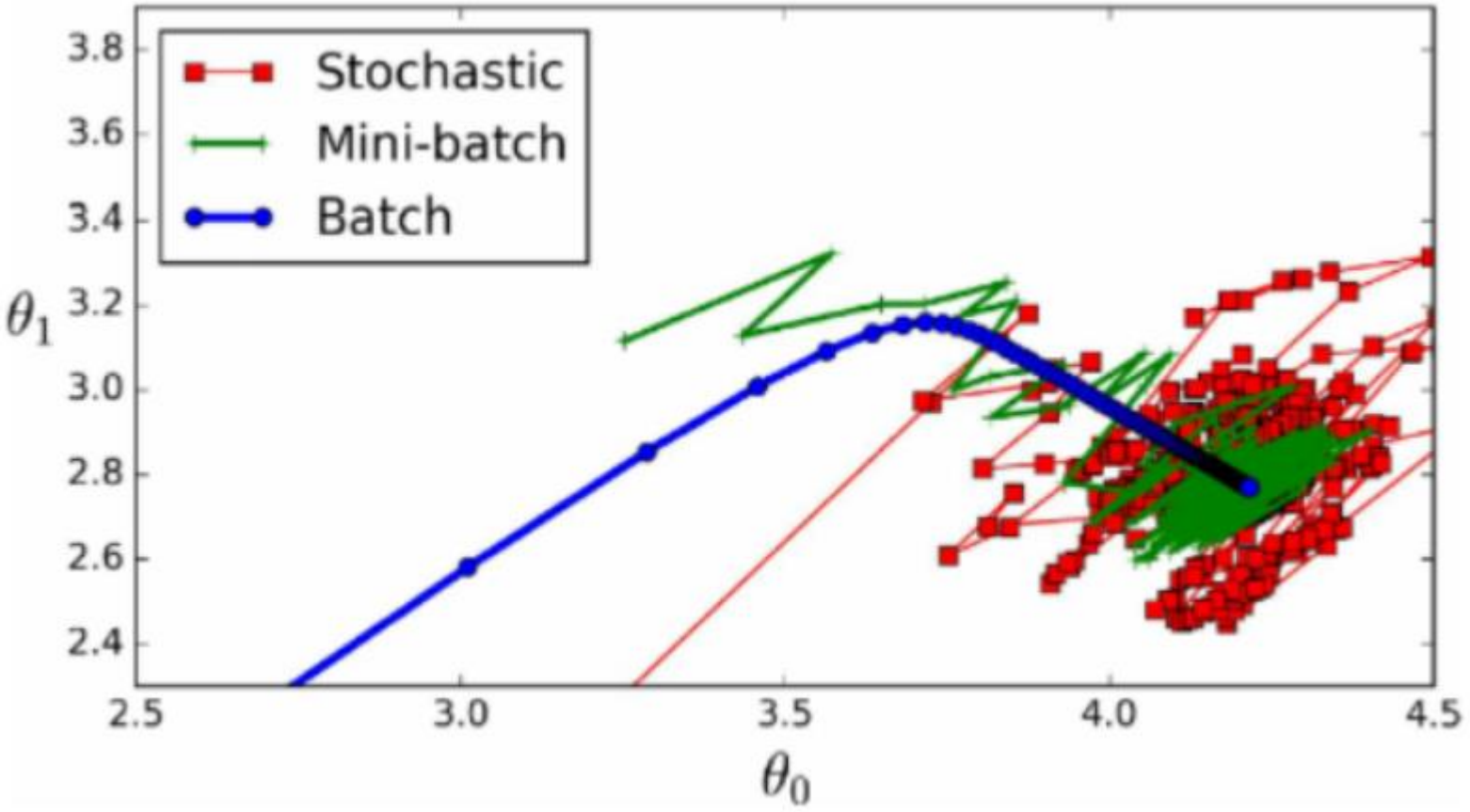


Figure 4-11. Gradient Descent paths in parameter space