# Gradient Descent

$\theta_0 \leftarrow$ Random

$$\theta_k \leftarrow \theta_{k-1} - \alpha \nabla_\theta J(\theta_{k-1}) \quad (*)$$

GD

$$J(\theta) = \frac{1}{n} \sum_{i=1}^{n} \text{cost}(y_i, \hat{y}_i)$$

$$\nabla_\theta J(\theta) = \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta \text{cost}(y_i, \hat{y}_i)$$

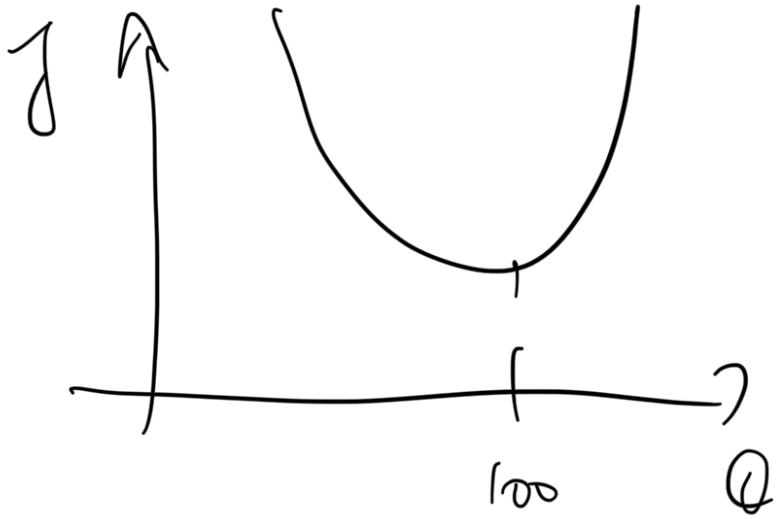
$$\hat{\nabla}_\theta J_\theta(\theta) = \frac{1}{s} \sum_{i=1}^{s} \nabla_\theta \text{cost}(y_i, \hat{y}_i)$$

$s \leq n$

SGD

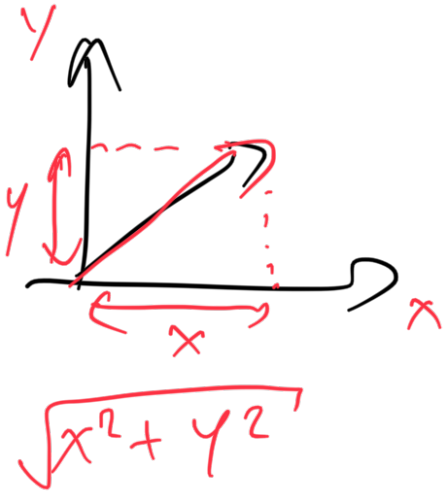
Approximation in a faster time (seconds)

steps are faster
but may need more steps.



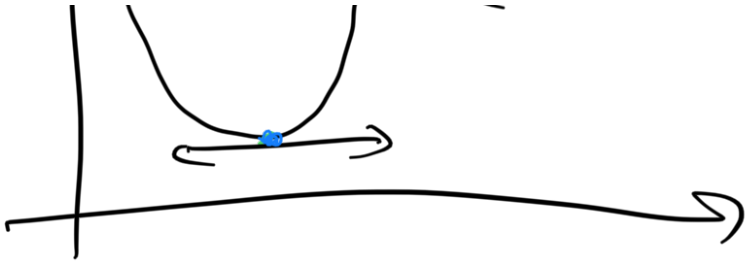

stopping criteria

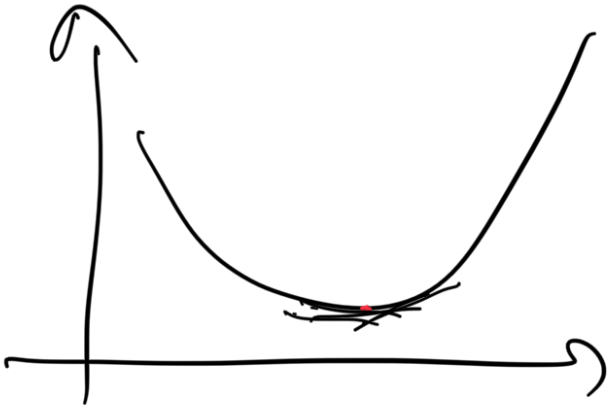
$$\|\nabla_\theta J\|_2 < \varepsilon$$




$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$$

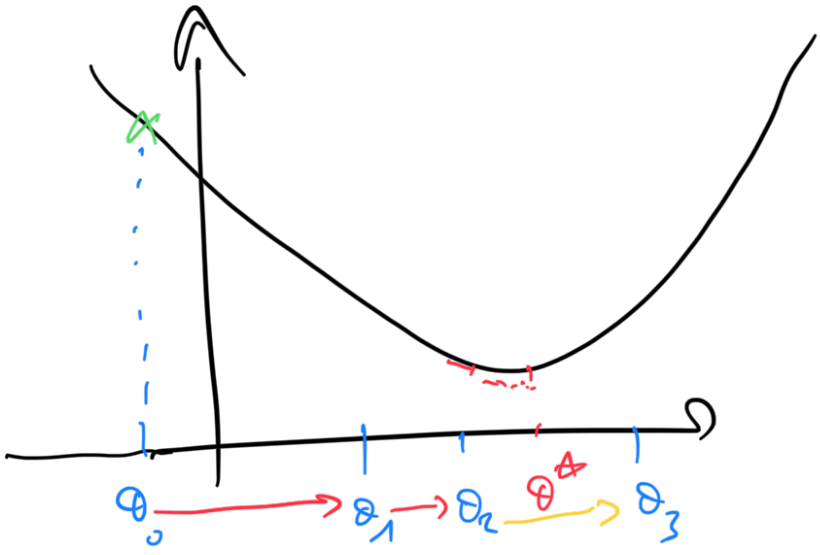
$$\|x\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_d^2}$$





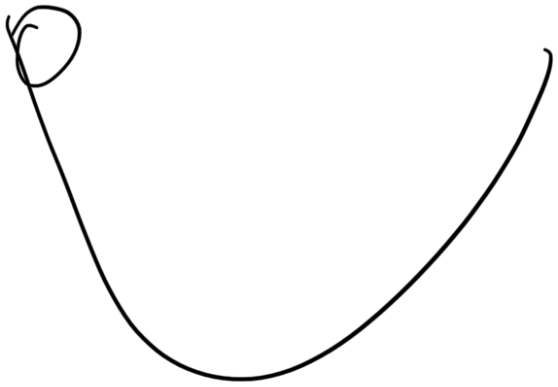
$$\|\nabla J\|_2 \leq \varepsilon$$

---




Decrease learning rate over the steps

eg. multiply by 0.9 every step



$$\begin{cases} V = \mu V - \alpha \nabla_\theta J(\theta_{k-1}) \\ \theta_k \leftarrow \theta_{k-1} + V \end{cases}$$

with momentum