

GD Gradient Descent

$$\frac{\partial}{\partial \underline{w}} L(D; \underline{w}) = \sum_{i \in D} \frac{\partial}{\partial \underline{w}} l(x_i, y_i; \underline{w})$$

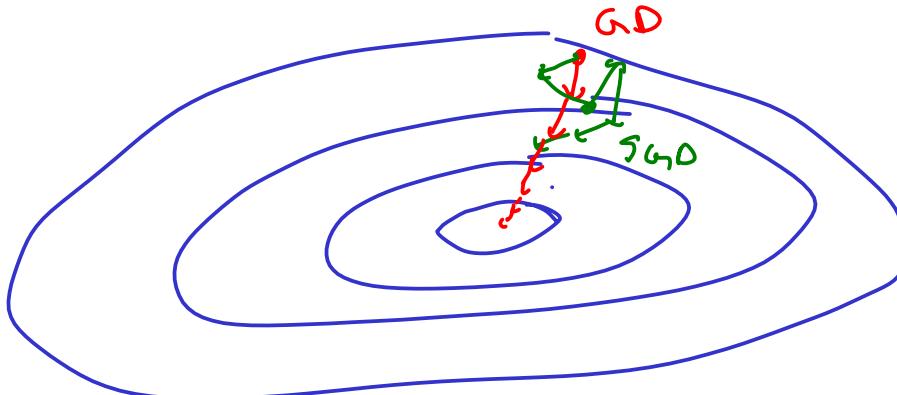
GD is problematic
if $|D|$ is big

while $\left| \frac{\partial L}{\partial \underline{w}} \right| > 0.0001$:

$$\underline{w}_{t+1} = \underline{w}_t - \alpha \begin{bmatrix} \frac{\partial L(D; \underline{w})}{\partial \underline{w}} \end{bmatrix}^T$$

Advantages of SGD

- ① Lower memory requirement
- ② Making progress with weights without having to wait full epoch



SGD Stochastic Gradient Descent

while

{ for B in D :

$$\underline{w}_{t+1} = \underline{w}_t - \alpha \sum_{i \in B} \frac{\partial}{\partial \underline{w}} l(x_i, y_i; \underline{w})$$

↓ EPOCH

[One iteration
over entire
Dataset]

↑ Randomly chosen
BATCH of data
from the dataset
 D

$|B|=1$ - SGD

$|B| \approx 64, 32, 128, 256, 512$

Batch SGD