GD
Gradient Descent $\xrightarrow{?}$ SGD Stochastic Gradient Descent

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

GD is problematic if $|D|$ is big

while $\left|\frac{\partial}{\partial w^L}\right| > 0.00001$:

$$\underline{w}_{t+1} = \underline{w}_t - \alpha \left[\frac{\partial}{\partial \underline{w}} L(D;\underline{w})\right]^T$$

while
- for B in D:
  - $\underline{w}_{t+1} = \underline{w}_t - \alpha \sum_{i \in B} \frac{\partial}{\partial \underline{w}} l(\underline{x}_i, y_i; \underline{w})$
- EPOCH

[One iteration over entire Dataset]

Randomly chosen BATCH of data from the dataset D

$|B| = 1$ - SGD

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

Batch SGD

Advantages of SGD

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