Stochastic Gradient Descent

$$w_{new} = w_{old} - \eta \frac{\partial L}{\partial w_{old}}$$

- n data points → gradient descent $L = \sum_{i=1}^{n} (y - \hat{y})^2$
- 1 data point - Stochastic GD $L = (y - \hat{y})^2$
- $K < n$ data points - Mini Batch SGD $L = \sum_{i=1}^{K} (y - \hat{y})^2$

Problem with SGD is the noise in the convergence. We need to remove this noise. We use SGD with momentum for this very purpose.

$$w_{new} = w_{old} - \eta \frac{\partial L}{\partial w_{old}} - \gamma V_{t-1}$$

$$V_{t-1} = 1 \times \left[\frac{\partial L}{\partial w_{old}}\right]_t + \gamma \left[\frac{\partial L}{\partial w_{old}}\right]_{t-1} + \gamma^2 \left[\frac{\partial L}{\partial w_{old}}\right]_{t-2} + \text{---}$$