# Обучение на больших выборках

Основы Deep Learning

# Gradient Descent

$$Loss(\hat{y}, y) = \frac{1}{2n}\sum_{i=1}^{n}(\hat{y_i} - y_i)^2 = \frac{1}{2n}\sum_{i=1}^{n}(\sigma(w \cdot X_i) - y_i)^2$$

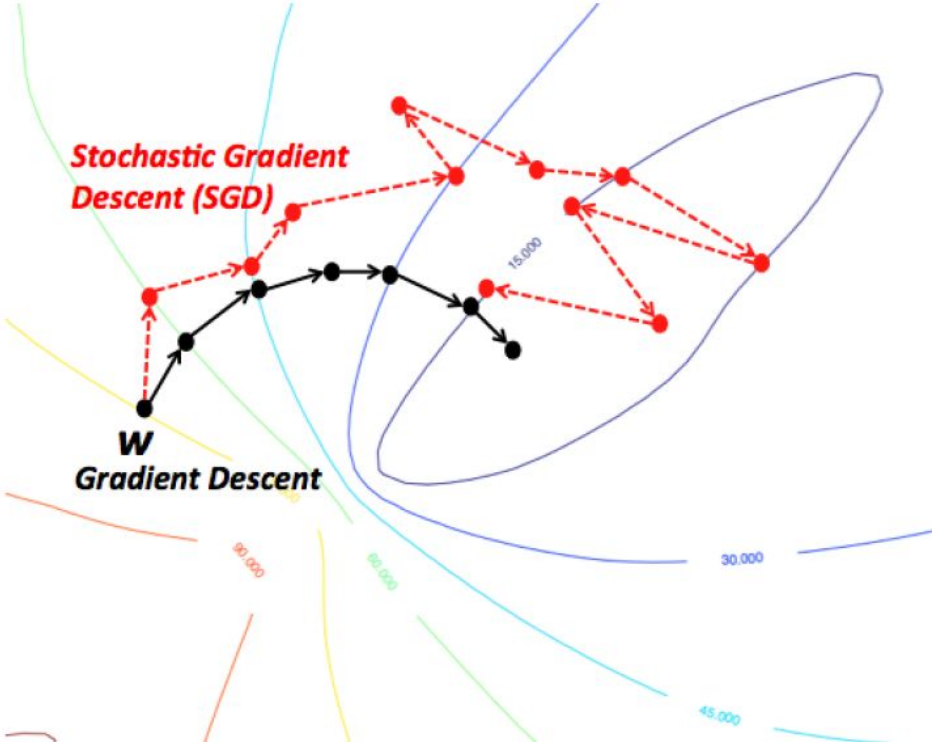
$$\frac{\partial Loss}{\partial w} = \frac{1}{n}X^T(\sigma(w \cdot X) - y)\sigma(w \cdot X)(1 - \sigma(w \cdot X))$$

# Stochastic Gradient Descent

$$Loss(\hat{y}, y) = (\hat{y}_i - y_i)^2 = (\sigma(w \cdot X_i) - y_i)^2$$

$$\frac{\partial Loss}{\partial w} = X_i^T(\sigma(w \cdot X_i) - y)\sigma(w \cdot X_i)(1 - \sigma(w \cdot X_i))$$

# Stochastic Gradient Descent

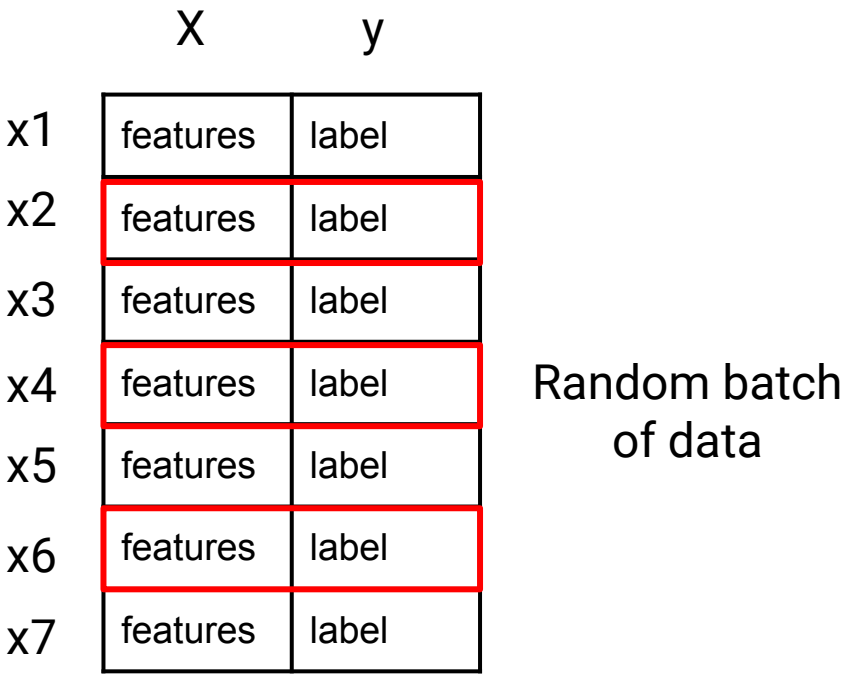


# Batch

| | X | y |
|---|---|---|
| x1 | features | label |
| x2 | features | label |
| x3 | features | label |
| x4 | features | label |
| x5 | features | label |
| x6 | features | label |
| x7 | features | label |

Batch of data

# Batch

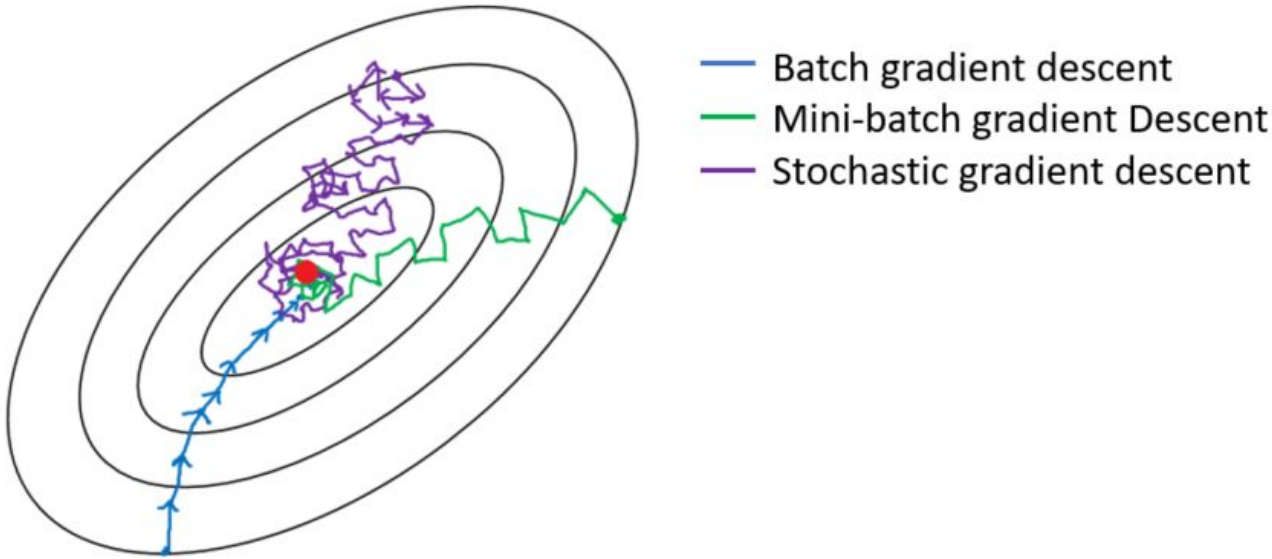

X
y
x1
features
label
x2
features
label
x3
features
label
x4
features
label
x5
features
label
x6
features
label
x7
features
label
Random batch of data

# Mini-batch Gradient Descent

$$Loss(\hat{y}, y) = \frac{1}{2 \cdot \text{batch\_size}} \sum_{i=1}^{\text{batch\_size}} (\hat{y_i} - y_i)^2 = \frac{1}{2 \cdot \text{batch\_size}} \sum_{i=1}^{\text{batch\_size}} (\sigma(w \cdot X_i) - y_i)^2$$

$$\frac{\partial Loss}{\partial w} = \frac{1}{\text{batch\_size}} X_{batch}^T (\sigma(w \cdot X_{batch}) - y) \sigma(w \cdot X_{batch})(1 - \sigma(w \cdot X_{batch}))$$

# Визуализация на линиях уровня



# Итерация и эпоха

- **Итерация**: оптимизация по одному батчу данных
- **Эпоха**: оптимизация по всей выборке, т.е. когда мы прошлись по всем батчам
- Например, если в выборке **N** элементов, а в батче **B** элементов, то **в одной эпохе будет N // B итераций** (но можно сделать и больше)