

Производная

$$f'(x) = \lim_{\Delta x \rightarrow 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \rightarrow 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}.$$

Частная производная

$$\frac{\partial z}{\partial x} = f'_x(x, y) \text{ (где } y = \text{const),}$$

$$\frac{\partial z}{\partial y} = f'_y(x, y) \text{ (где } x = \text{const).}$$

Градиент



вектор, который показывает направление наибольшего роста функции

Градиентный спуск

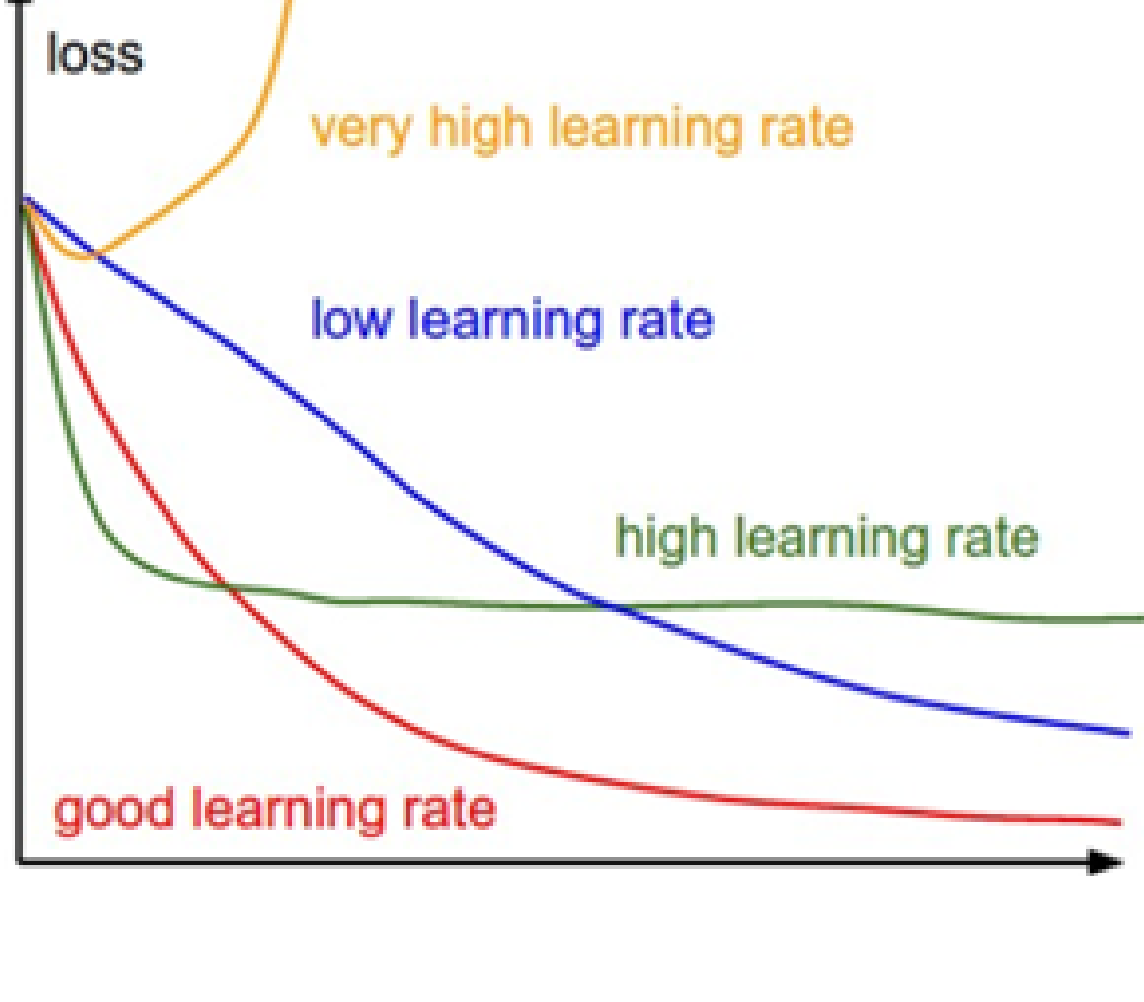
Задача: найти минимум функции

Алгоритм:

- 1. Берем произвольный x и считаем в этой точке градиент.
- 2. Делаем цикл пока grad(x) < eps
- 3. Если он больше, то высчитываем новый x по формуле:

$$\theta = \theta - \eta \cdot \nabla_{\theta} J(\theta).$$

Learning rate



Стохастический градиентный спуск

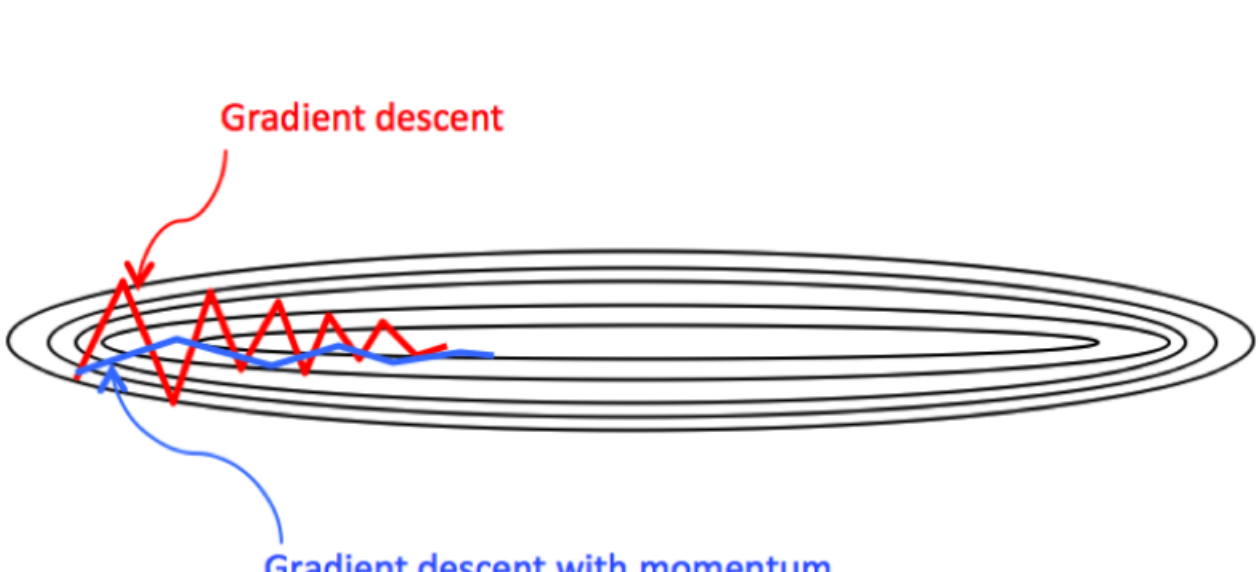
Градиент оптимизируемой функции считается как градиент от случайно выбранного подмножества данных.

$$\theta = \theta - \eta \cdot \nabla_{\theta} J(\theta; x^{(i)}; y^{(i)}).$$

Момент

$$v_t = \gamma v_{t-1} + \eta \nabla_{\theta} J(\theta)$$

$$\theta = \theta - v_t$$



Ускоренный градиент Нестерова

Дает нам приблизительное представление о следующем положении параметров.

$$v_t = \gamma v_{t-1} + \eta \nabla_{\theta} J(\theta - \gamma v_{t-1})$$

$$\theta = \theta - v_t$$

Метод отжига

Применяя его к градиентному спуску, мы каждый раз уменьшаем понемногу ленинг рэйт по мере приближения к минимуму функции.

Adagrad

$$\theta_{t+1,i} = \theta_{t,i} - \frac{\eta}{\sqrt{G_{t,ii} + \epsilon}} \cdot g_{t,i}.$$

Adadelata и RMSprop

RMSprop и Adadelata были разработаны независимо друг от друга примерно в одно и то же время из-за необходимости решить проблему радикального снижения learning rate в Adagrad.

$$E[g^2]_t = \gamma E[g^2]_{t-1} + (1 - \gamma) g_t^2.$$

$$\Delta \theta_t = - \frac{\eta}{\sqrt{E[g^2]_t + \epsilon}} g_t.$$

Adam

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$$

1. Вычисляем параметры

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t}$$

2. Считаем оценку параметров

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t + \epsilon}} \hat{m}_t.$$

3. Подставляем в формулу