Название	Формула	Недостатки
Gradient Descent	$\theta = \theta - \alpha \nabla_\theta J(\theta)$	Градиент вычисл. проходом по всему под. данных => Треб. много памяти. Также из-за слишком большого градиента алгоритм может не сойтись
Stohastic Gradient Descent	$\theta = \theta - \alpha \nabla_\theta J(\theta, S)$, где S - параметры для обучения	На каждом парам. обуч меняются веса => • аномалии сильно влияют • долгое обучение
Mini-Batch Gradient Descent	$\theta = \theta - \alpha \cdot \nabla_\theta J(\theta, Ns)$, Ns - группы обуч. параметров	Параметр N оказывает основное влияние на обучение => необходимо потратить много времени на его подбор
SGD + Momentum	$v = \gamma \cdot v + \eta \nabla_\theta J(\theta)$, $\theta = \theta - \alpha v$	Также необходимо потратить много времени на ручной подбор параметра. К тому же импульс

		может стать слишком большим ⇒ начнёт пропускать лок. min.
SGD + Momentum + Acceleration	$v = \gamma \cdot v + \eta \nabla_\theta J(\theta - \gamma \cdot v)$ $\theta = \theta - \alpha v$	Аналогично предыдущ. сложно подобрать параметр
Adagard	$g_t = \nabla_\theta \cdot J(\theta_t)$, $\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{G_t + \varepsilon}} \cdot g_t$	Знаменатель всё время увелич. ⇒ скорость обуч. уменьшается и может останов.
Adadelta	$\theta_{t+1,i} = \theta_{t,i} - \frac{\eta}{\sqrt{E[G_{t,ii}] + \varepsilon}} \cdot \nabla_{\theta_{t,i}} J(\theta_{t,i})$	Вычислительно дорого
Adam	$\theta_{t+1,i} = \theta_{t,i} - \frac{\eta}{\sqrt{E[G_{t,ii}] + \varepsilon}} \cdot E[g_{t,i}]$	Вычислительно дорого Хуже обобщ. данные
Nadam	$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{E[G_t] + \varepsilon}} \cdot \left(\beta_1 E[g_t] + \frac{(1-\beta_1)\nabla_{\theta_t} J(\theta_t)}{1 - \beta_1^t}\right)$	Низкая скорость обучения

Accelerate - создан для ускорения импульса, даёт возможность вычислять его относительно предположит. будущего.

Adagard:

Меняет скорость обуч. на η для каждого парам. на каждом шаге. Работает на ∀ ф-ции ошибок.

$$g_t = \nabla_\theta J(\theta_t)$$

$$G_t = G_t + g_t^2$$

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{G_{t+\varepsilon}}} \cdot g_t$$

η – скорость обуч., которая изменяется для заданного θ_i.

Adadelta:

Расширение Adagard. Огранич окно накопленных градиентов до фиксир. размера.

Используется exp скользящая средняя

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

$$RMS[g]_t = \sqrt{E[g^2]_t + \varepsilon}$$

$$\theta_{t+1} = \theta_t - \frac{RMS[\Delta\theta]_{t-1}}{RMS[g]_t} \cdot g_t$$

Adam

Работает с имп. 1-го и 2-го порядка.

Уменьшается скорость во избежание проскак. min.

Сохраняются скользящие средние прошлых ∇.

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t}, \quad \hat{v} = \frac{v_t}{1-\beta_2^t}$$

$$\Theta_{t+1} = \Theta_t - \frac{\eta}{\sqrt{\hat{v}_t} + \varepsilon} \cdot \hat{m}_t$$