Accelerate - создан для управления импульсом, дает нам возможность его вычислять не относительно прошлого состояния, а относительно предположительного будущего. Это не дает импульсу сильно разогнаться и начать пропускать локальные минимумы.

Название	Формула	Недостатки
Gradient Descent	$\theta = \theta - \alpha \cdot \nabla_\theta J(\theta)$	Из-за прохода по всему датасету накапливается слишком большой градиент, из-за чего алгоритм может вообще не сойтись
Stochastic Gradient Descent	$\theta = \theta - \alpha \cdot \nabla_\theta J(\theta; sample)$	Из-за обновления весов на каждом сэмпле - аномалии сильно влияют на градиент. Увеличение времени обучения.
Mini-Batch Gradient Descent	$\theta = \theta - \alpha \cdot \nabla_\theta J(\theta; Nsamples)$	Все начинает зависеть от параметра N, алогритм может долго или плохо обучаться. Необходимо потратить время на его подбор.
SGD + Momentum	$v = \gamma \cdot v + \eta \cdot \nabla_\theta J(\theta)$ $\theta = \theta - \alpha v$	Необходимо потратить время и настроить параметр, так же импульс может стать слишком большим и алгоритм начнет пропускать локальные минимумы.
SGD + Momentum + Acceleration	$v = \gamma \cdot v + \eta \cdot \nabla_\theta J(\theta - \gamma \cdot v)$ $\theta = \theta - \alpha v$	Сложно подобрать параметр, повышается время работы
Adagrad	$\theta_{t+1,i} = \theta_{t,i} - \frac{\eta}{\sqrt{G_{t,ii}+\epsilon}} \nabla_{\theta_{t,i}} J(\theta_{t,i})$	Из-за постоянного увеличения знаменателя скорость обучения начинает уменьшаться и в конечном итоге модель может вообще перестать обучаться.
Adadelta	$\theta_{t+1,i} = \theta_{t,i} - \frac{\eta}{\sqrt{E[G_{t,ii}]+\epsilon}} \nabla_{\theta_{t,i}} J(\theta_{t,i})$	На некоторых данных будет справляться чуть хуже (конкретных примеров не нашел).
Adam	$\theta_{t+1,i} = \theta_{t,i} - \frac{\eta}{\sqrt{E[G_{t,ii}]+\epsilon}} \times E[g_{t,i}]$	Возможно чуть хуже обобщает данные
Nadam	$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{E[G_t]+\epsilon}} \left(\beta_1 \cdot E[g_t] + \frac{(1-\beta_1)\nabla_{\theta_t} J(\theta_t)}{1-\beta_1^t} \right)$	На некоторых данных будет справляться чуть хуже (конкретных примеров не нашел).