Adam Optimizer

Устинов Н.А., г. 5030102/20101

13 октября 2025 г.

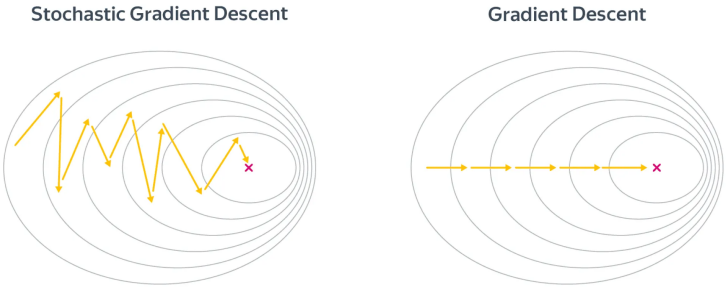
Оптимизация параметров в ML

Задача поиска минимума:

$$\nabla f_t(w^*) = 0$$

Классический градиентный спуск:

$$w_{t+1} = w_t - \alpha \nabla f(w_t)$$



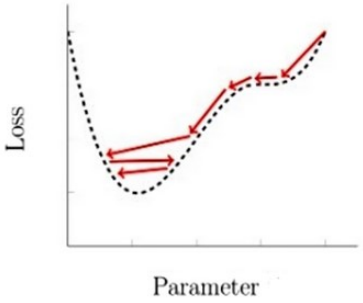
AdaGrad

Накопление момента:

$$v_t = v_{t-1} + (\nabla f(w_t))^2$$

Переход к следующей точке:

$$w_{t+1} = w_t - \frac{\alpha}{\sqrt{v_t} + \varepsilon} \nabla f(w_t)$$



RMSprop

Накопление момента:

$$v_t = \beta v_{t-1} + (1 - \beta)\nabla f(w_t)^2$$

Переход к следующей точке:

$$w_{t+1} = w_t - \frac{\alpha}{\sqrt{v_t} + \epsilon}\nabla f(w_t)$$

Adam

Первый и второй моменты:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1)\nabla f(w_t)$$

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

Смещение моментов:

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

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

Переход к следующей точке:

$$w_{t+1} = w_t - \frac{\alpha}{\sqrt{\hat{v}_t} + \epsilon}\hat{m}_t$$

Параметры: $\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$.

Алгоритм оптимизатора Adam

$m_0 \leftarrow 0$ ▷ Первый момент
$v_0 \leftarrow 0$ ▷ Второй момент
$t \leftarrow 0$
while θ_t not converged **do**
    $t \leftarrow t + 1$
    $g_t \leftarrow \nabla_\theta f_t(\theta_{t-1})$ ▷ Градиент функции в момент времени t
    $m_t \leftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$ ▷ Обновление первого момента
    $v_t \leftarrow \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$ ▷ Обновление второго момента
    $\hat{m}_t \leftarrow m_t / (1 - \beta_1^t)$ ▷ Смещение первого момента
    $\hat{v}_t \leftarrow v_t / (1 - \beta_2^t)$ ▷ Смещение второго момента
    $\theta_t \leftarrow \theta_{t-1} - \alpha \cdot \hat{m}_t / (\sqrt{\hat{v}_t} + \epsilon)$
end while
return θ_t

Модификации оптимизатора Adam

- AdamW

$$w_{t+1} = w_t - \left(\frac{\alpha}{\sqrt{\hat{v}_{t+1} + \epsilon}} \hat{m}_{t+1} + \lambda w_t \right)$$

- AdaMax

$$u_t = \max(\beta_2 u_{t-1}, |g_t|)$$

- Nesterov Momentum

Пример

$$f(x) = x^2, x_0 = 1, \alpha = 0.4$$

$$g_1 = \nabla f(x_0) = 2x_0 = 2$$

$$m_1 = \beta_1 m_0 + (1 - \beta_1) g_1 = 0.9 \times 0 + 0.1 \times 2 = 0.2$$

$$v_1 = \beta_2 v_0 + (1 - \beta_2) g_1^2 = 0.999 \times 0 + 0.001 \times 4 = 0.004$$

$$\hat{m}_1 = \frac{m_1}{1 - \beta_1^1} = \frac{0.2}{1 - 0.9} = 2$$

$$\hat{v}_1 = \frac{v_1}{1 - \beta_2^1} = \frac{0.004}{1 - 0.999} = 4$$

$$x_1 = x_0 - \alpha \frac{\hat{m}_1}{\sqrt{\hat{v}_1} + \varepsilon} = 1 - 0.4 \times \frac{2}{\sqrt{4} + 10^{-8}} = 0.6$$

Пример

График функции $f(x) = x^2$ и точек алгоритма

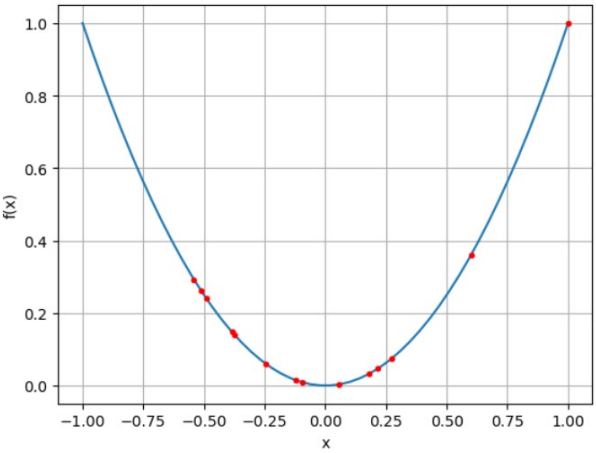


Таблица: Пример работы оптимизатора Adam для $f(x) = x^2$

t	x_t	$f(x_t)$	$\nabla f(x_t)$	m_t	v_t	$\hat{m}_t$	$\hat{v}_t$
0	1	1	0	0	0	0	0
1	0.600000	0.360000	2.000000	0.200000	0.004000	2.000000	4.000000
2	0.217004	0.047091	1.200000	0.300000	0.005436	1.578947	2.719360
3	-0.120833	0.014601	0.434008	0.313401	0.005619	1.156460	1.874850
4	-0.372553	0.138795	-0.241667	0.257894	0.005672	0.749910	1.420056
5	-0.510416	0.260524	-0.745105	0.157594	0.006221	0.384836	1.246735
6	-0.541234	0.292934	-1.020831	0.039752	0.007257	0.084838	1.212543
7	-0.490651	0.240739	-1.082467	-0.072470	0.008422	-0.138911	1.206696
8	-0.384862	0.148118	-0.981303	-0.163354	0.009376	-0.286820	1.176122
9	-0.246100	0.060565	-0.769723	-0.223990	0.009959	-0.365651	1.111012
10	-0.093865	0.008811	-0.492201	-0.250812	0.010192	-0.385081	1.023745
11	0.053658	0.002879	-0.187730	-0.244503	0.010217	-0.356321	0.933431
12	0.179761	0.032314	0.107316	-0.209321	0.010218	-0.291709	0.856182
13	0.271171	0.073534	0.359523	-0.152437	0.010337	-0.204390	0.799928

Список источников

Diederik P. Kingma, Jimmy Lei Ba Adam: A method for stochastic optimization, 2015

Yandex: Учебник по машинному обучению.
https://education.yandex.ru/handbook/ml

Хабр: Методы оптимизации в машинном и глубоком обучении
https://habr.com/ru/articles/813221/