RMAD without momentum

Stochastic gradient descent without momentum

1. input: initial $\mathbf{w}_1$, learning rates α, loss function $L(\mathbf{w}, \theta, t)$
2. initialize $\mathbf{v}_1 = 0$
3. for $t = 1$ to T do
4. $g_t \sim \nabla_w L(\mathbf{w}_t, \theta, t)$
5. $v_t = -g_t$
6. $w_{t+1} = w_t + \alpha_t v_t$
7. end for
8. output: trained parameters w_T

Мы не накапливаем информацию о градиентах в переменной v_t, а просто вычитаем стох. оценку градиента. Далее используем данный алгоритм и перепишем в режиме обратного дифференцирования:

1. input: $\mathbf{w}_T$, $\mathbf{v}_T$, α, train loss $L(\mathbf{w}, \theta, t)$, loss $f(\mathbf{w})$
2. initialize $d\mathbf{v} = 0$, $d\theta = 0$, $d\alpha = 0$
3. initialize $d\mathbf{w} = \nabla_{\mathbf{w}} f(\mathbf{w}_T)$
4. for $t = T$ counting down to 1 do
5. $d\alpha_t = d\mathbf{w}^\top \mathbf{v}_t$
6. $\mathbf{w}_{t-1} = \mathbf{w}_t - \alpha_t \mathbf{v}_t$
7. $\mathbf{g}_{t-1} \sim \nabla_{\mathbf{w}} L\left(\mathbf{w}_{t-1}, \boldsymbol{\theta}, t-1\right)$
8. $\mathbf{v}_{t-1} = -\mathbf{g}_{t-1}$

9. $d\mathbf{v} = d\mathbf{v} + \alpha_t d\mathbf{w}$

10. $d\mathbf{w} = d\mathbf{w} - d\mathbf{v}\nabla_{\mathbf{w}}\nabla_{\mathbf{w}} L\left(\mathbf{w}_t, \boldsymbol{\theta}, t\right)$

11. $d\boldsymbol{\theta} = d\boldsymbol{\theta} - d\mathbf{v}\nabla_{\boldsymbol{\theta}}\nabla_{\mathbf{w}} L\left(\mathbf{w}_t, \boldsymbol{\theta}, t\right)$

12. end for

13. output gradient of $f\left(\mathbf{w}_T\right)$ w.r.t $\mathbf{w}_1, \mathbf{v}_1, \gamma, \boldsymbol{\alpha}$ and $\boldsymbol{\theta}$

Здесь мы не используем параметр затухания гамма, по сути мы его кладем равным нулю, и используем то, что мы прибавляем оценку антиградиента на каждом шаге. Такой алгоритм даже проще, чем алгоритм с инерцией.