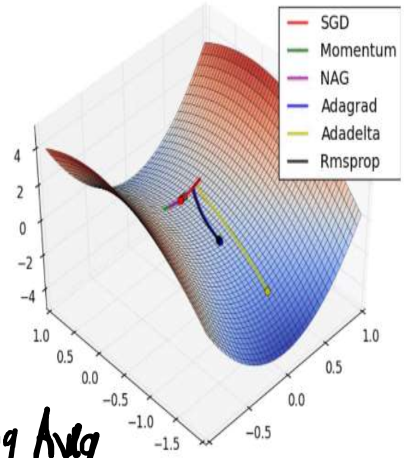
Optimization algorithms (part 3)

RMSProp

Edit



RMSProp is an unpublished adaptive learning rate optimizer proposed by Geoff Hinton. The motivation is that the magnitude of gradients can differ for different weights, and can change during learning, making it hard to choose a single global learning rate. RMSProp tackles this by keeping a moving average of the squared gradient and adjusting the weight updates by this magnitude. The gradient updates are performed as:

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

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{E[g^2]_t + \epsilon}} g_t$$

$E(g^2)$ → Running Avg.

Hinton suggests $\gamma = 0.9$, with a good default for η as 0.001.

Image: Alec Radford

$$E_t = E(g^2)_{t-1}$$

$$g_t^2 = (g_t)^2 = (\nabla \ell)^2$$

$$\boxed{\begin{aligned} &\text{Updation rule (RMSProp)} \\ &E_t = \gamma E_{t-1} + (1-\gamma) g_t^2 \\ &w_{t+1} = w_t - \frac{\eta}{\sqrt{E_t + \varepsilon}} g_t \end{aligned}}$$

Consider $\ell(\omega) = 4\omega^2 - 12\omega + 9$

$\ell'(\omega) = g = 8\omega - 12 \Rightarrow \boxed{g_t = 8\omega_t - 12}$

We start with $\omega_0 = 5$, $\eta = 0.2$, $\varepsilon = 0$, $\gamma = 0.9$

$g_0 = 8(5) - 12 = 28$

$g_0^2 = 784$

$E_0 = E(g_0^2) = \frac{784}{1} = 784$

Now, $\omega_1 = \omega_0 - \frac{\eta}{\sqrt{E_0 + \varepsilon}} g_0$

$\Rightarrow \omega_1 = 5 - \frac{0.2}{\sqrt{784}} (28)$

$\Rightarrow \boxed{\omega_1 = 4.8}$

$g_t = 8 w_t - 12$

$\Rightarrow \boxed{w_1 = 4.8}$ $g_t = 8w_t - 12$

We get $g_1 = 8(4.8) - 12$ $\quad g_1 = 8w_1 - 12$

$\boxed{g_1 = 26.4}$

We finish for $t = 0$ over here & proceed with $t = 1$

Now, $E_1 = \gamma E_0 + (1-\gamma) g_1^2$ $\quad \left[E_t = \gamma E_{t-1} + (1-\gamma) g_t^2\right]$

$= 0.9(784) + 0.1(26.4)^2$

$= 775.296$

Now, $w_2 = w_1 - \frac{0.2}{\sqrt{E_1}} g_1$ $\quad \left[w_{t+1} = w_t - \frac{\eta}{\sqrt{E_t}} g_t\right]$

$= 4.8 - \frac{0.2}{\sqrt{775.296}} (26.4)$

$\boxed{w_2 = 4.6104}$

Now, $g_2 = 8w_2 - 12$

$\boxed{g_2 = 24.883}$

We finish for $t = 1$ & proceed for $t = 2$,

Now, $E_2 = \gamma E_1 + (1-\gamma) g_2^2$

$= 0.9(775.296) + 0.1(24.883)^2$

$= 759.683$

Now, $w_3 = w_2 - \frac{\eta}{\sqrt{E_2}} g_2$

$= 4.6104 - \frac{0.2}{\sqrt{759.683}} (24.883)$

$\boxed{w_3 = 4.4299}$

Also $g_3 = 8w_3 - 12 = 47.4387$

Adam optimizer

Algorithm 1: *Adam*, our proposed algorithm for stochastic optimization. See section 2 for details, and for a slightly more efficient (but less clear) order of computation. g_t^2 indicates the elementwise square $g_t \odot g_t$. Good default settings for the tested machine learning problems are $\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. All operations on vectors are element-wise. With β_1^t and β_2^t we denote β_1 and β_2 to the power t.

Require: α: Stepsize
Require: $\beta_1, \beta_2 \in [0, 1)$: Exponential decay rates for the moment estimates
Require: $f(\theta)$: Stochastic objective function with parameters θ
Require: θ_0: Initial parameter vector
  $m_0 \leftarrow 0$ (Initialize 1st moment vector)
  $v_0 \leftarrow 0$ (Initialize 2nd moment vector)
  $t \leftarrow 0$ (Initialize timestep)
  while θ_t not converged **do**
    $t \leftarrow t + 1$
    $g_t \leftarrow \nabla_\theta f_t(\theta_{t-1})$ (Get gradients w.r.t. stochastic objective at timestep t)
    $m_t \leftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$ (Update biased first moment estimate)
    $v_t \leftarrow \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$ (Update biased second raw moment estimate)
    $\widehat{m}_t \leftarrow m_t / (1 - \beta_1^t)$ (Compute bias-corrected first moment estimate)
    $\widehat{v}_t \leftarrow v_t / (1 - \beta_2^t)$ (Compute bias-corrected second raw moment estimate)
    $\theta_t \leftarrow \theta_{t-1} - \alpha \cdot \widehat{m}_t / (\sqrt{\widehat{v}_t} + \epsilon)$ (Update parameters)
  end while
  return θ_t (Resulting parameters)

$\delta, \gamma, \eta, \epsilon$

Adam updation rule :-

$$m_t = \delta m_{t-1} + (1-\delta) g_t$$
$$v_t = \gamma v_{t-1} + (1-\gamma) g_t^2$$
$$\hat{m}_t = \frac{m_t}{1-\delta^t} \quad \& \quad \hat{v}_t = \frac{v_t}{1-\gamma^t}$$
$$\Rightarrow w_{t+1} = w_t - \frac{\eta}{\sqrt{\hat{v}_t + \varepsilon}} \cdot \hat{m}_t$$

$g_t^2 = g_t * g_t$

Consider $l(w) = 4w^2 - 12w + 9$

$g_t = 8w_t - 12$

We assume :- $w_0 = 5$

$\delta = 0.9$

$\gamma = 0.999$, $\eta = 0.2$

$v_0 = 0 = \hat{v}_0$

$m_0 = 0 = \hat{m}_0$, $\varepsilon = 0.00005$

Here $g_0 = 8(5) - 12 = 28$

$g_0^2 = 28^2 = 784$

$w_1 = w_0 - \frac{0.2}{\sqrt{v_0 + \varepsilon}} \cdot m_0 = 5 - \frac{0.2(0)}{\sqrt{\varepsilon}} = 5 \rightarrow \boxed{g_1 = g_0}$

$w_1 = 5$

$g_1 = 8(5) - 12$

$= g_0$

$= 28$

Now, $m_1 = \delta m_0 + (1-\delta) g_1$

$\Rightarrow m_1 = \underline{\underline{0.9}}(\underline{0}) + \underline{\underline{0.1}}(\underline{\underline{28}})$

$\boxed{m_1 = 2.8}$ ✓

Now, $v_1 = 0.999(0) + 0.001(784) = 0.784$ ✓

$\hat{m}_1 = \dfrac{m_1}{1-0.9} = \underline{\underline{28}}$ ✓ $\dfrac{m_1}{1-\delta'}$

$\hat{V}_1 = \dfrac{0.784}{1-0.999} = \underline{\underline{784}}$ $\dfrac{V_1}{1-\gamma'}$

Now $w_2 = w_1 - \dfrac{0.2}{\sqrt{784 + 0.00005}}(28)$

$\boxed{w_2 = 4.80}$

$\Rightarrow g_2 = 8(w_2) - 12 = 8(4.8) - 12$

$\boxed{g_2 = 26.4}$

We will proceed to $t = 2$.

$m_2 = \delta m_1 + (1-\delta) g_2$

$\Rightarrow m_2 = 0.9(2.8) + (1-0.9)(26.4)$

$\Rightarrow \boxed{m_2 = 5.16}$

Now, $V_2 = \gamma V_1 + (1-\gamma)\, g_2^2$

$$= 0.999(0.784) + (1-0.999)(26.4)^2$$

$$\boxed{V_2 = 1.480176}$$

Now $\hat{m}_2 = \dfrac{m_2}{1-\delta^2} = \dfrac{5.16}{1-0.9^2} = 27.158$

$$\hat{V}_2 = \frac{V_2}{1-\gamma^2} = \frac{1.48}{1-0.999^2} = 740.37$$

Now, $w_3 = w_2 - \dfrac{0.2}{\sqrt{\cancel{27.158}\ 740.37} + 0.00005}\,(27.158)$

$$\boxed{w_3 = 4.60}$$

We get $g_3 = 8(4.60) - 12$

$$= 24.80$$