General Framework

1. $g_t = \nabla f(w)$
2. Calculate $m_t = \phi(g_1, g_2 \cdots g_t)$, $v_t = \psi(g_1, g_2 \cdots g_t)$
3. $\eta_t = \alpha \cdot m_t / \sqrt{v_t}$
   Calculate the decent gradient of current moment
4. $w_{t+1} = w_t - \eta_t$ update weights

SGD

$m_t = g_t$; $v_t = I^2$

$\eta_t = \alpha \cdot g_t$ the biggest disadvantage of SDG is its decent is very slow

SGD with Momentum

$m_t = \beta_1 m_{t-1} + (1 - \beta_1) \cdot g_t$ $g_t = 0.9$ as experience.

SGD with Nesterov Acceleration

$g_t = \nabla f(w_t - \alpha \cdot m_{t-1} / \sqrt{v_{t-1}})$ Take a step forward

AdaGrad

$v_t = \sum_{r=1}^{t} g_r^2$ the sum of history gradient

$\eta_t = \alpha \cdot m_t / \sqrt{v_t}$

Learning rate turns to $\frac{\alpha}{\sqrt{v_t}}$ ← add a bias to prevent

denominator to be 0

AdaDelta / RMSProp

$V_t = \beta_2 \cdot V_{t-1} + (1-\beta_2) g_t^2$

Adam

$m_t = \beta_1 \cdot m_{t-1} + (1-\beta_1) \cdot g_t$

$V_t = \beta_2 \cdot V_{t-1} + (1-\beta_2) g_t^2$

Nadam

$g_t = \nabla f(w_t - \alpha \cdot m_{t-1} / \sqrt{V_t})$

$m_0 = 0$ $V_0 = 0$ will cause denominator to be zero.

initialize $\hat{m}_t = m_t / (1-\beta_1^t)$

$\tilde{V}_t = V_t / (1-\beta_2^t)$

Adam 不收敛，学习率不单调

$V_t = \max(\beta_2 \cdot V_{t-1} + (1-\beta_2) g_t^2, V_{t-1})$

保证 $\|V_t\| \geq \|V_{t-1}\|$

先用 adam 再用 SGD