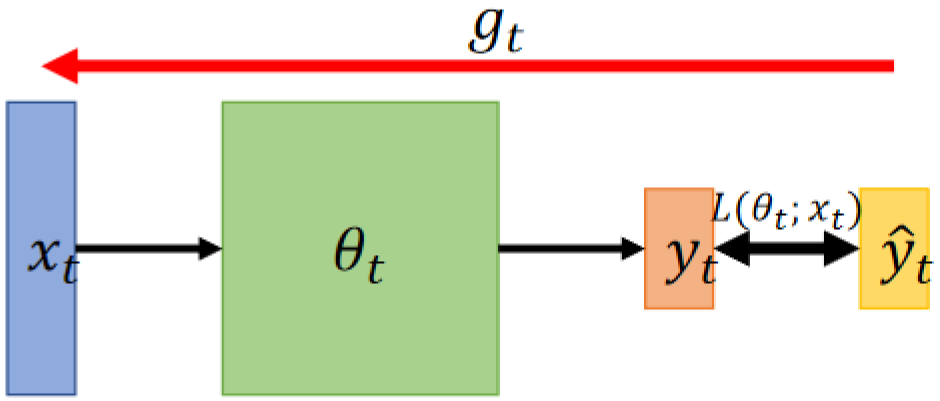
Some Notations

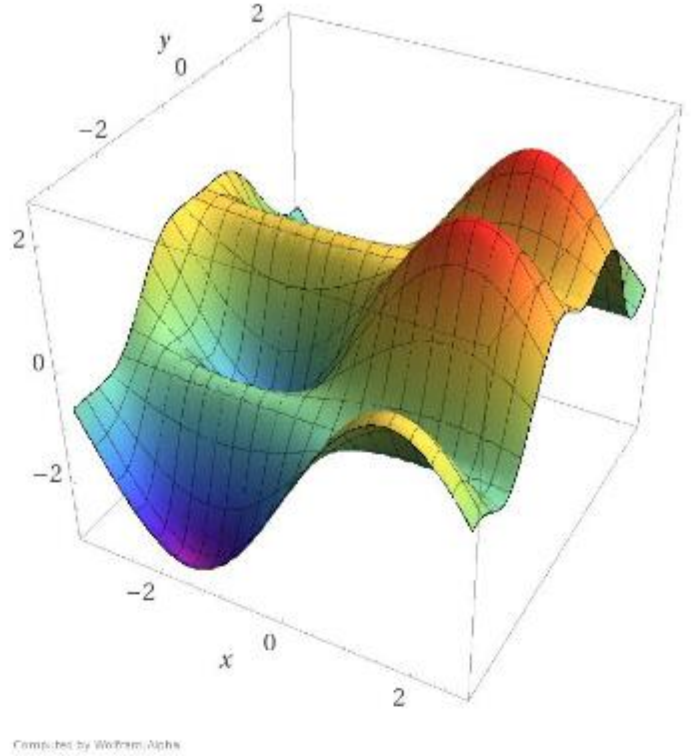
- θ_t : model parameters at time step t
- $\nabla L(\theta_t)$ or g_t: gradient at θ_t, used to compute θ_{t+1}
- m_{t+1}: momentum accumulated from time step 0 to time step t, which is used to compute θ_{t+1}



Credit to 簡仲明助教

What is Optimization about?

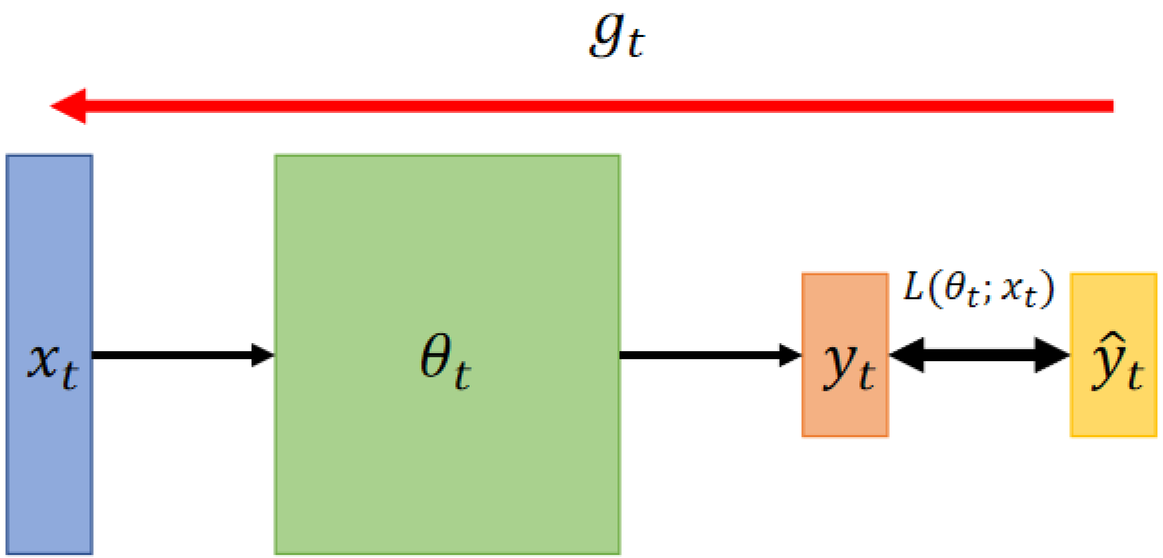
- Find a θ to get the lowest $\sum_x L(\theta; x)$!!
- Or, Find a θ to get the lowest $L(\theta)$!!



Credit to 簡仲明助教

On-line vs Off-line

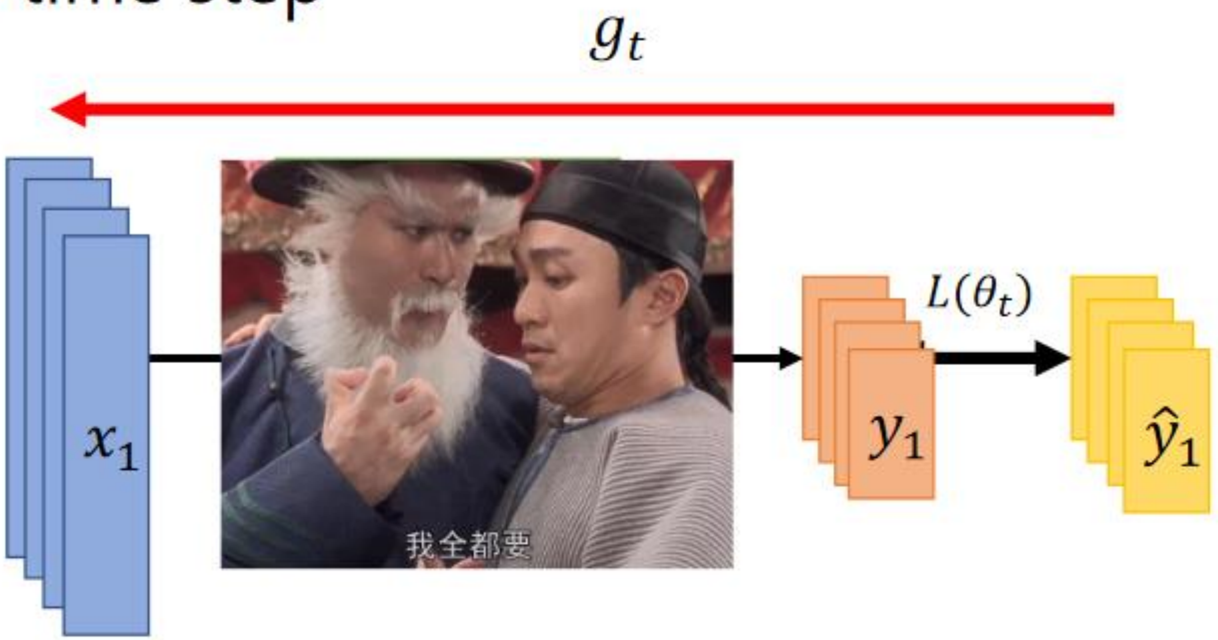
- On-line：one pair of $(x_t, \hat{y}_t)$ at a time step



Credit to 簡仲明助教

On-line vs Off-line

- Off-line：pour all $(x_t, \hat{y}_t)$ into the model at every time step

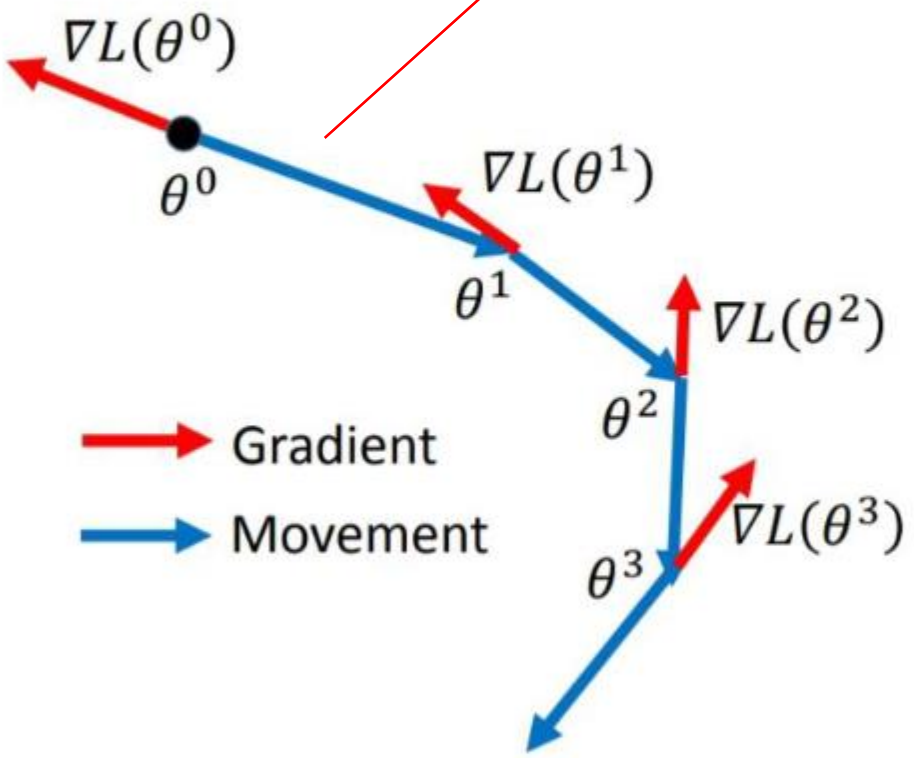


- The rest of this lecture will focus on the off-line cases

Credit to 簡仲明助教

SGD

因為gradient 的方向為L 增加的方向,
所以往反方向移動, 才可找到有最小L 的θ



Start at position θ^0

Compute gradient at θ^0

Move to $\theta^1 = \theta^0 - \eta\nabla L(\theta^0)$

Compute gradient at θ^1

Move to $\theta^2 = \theta^1 - \eta\nabla L(\theta^1)$

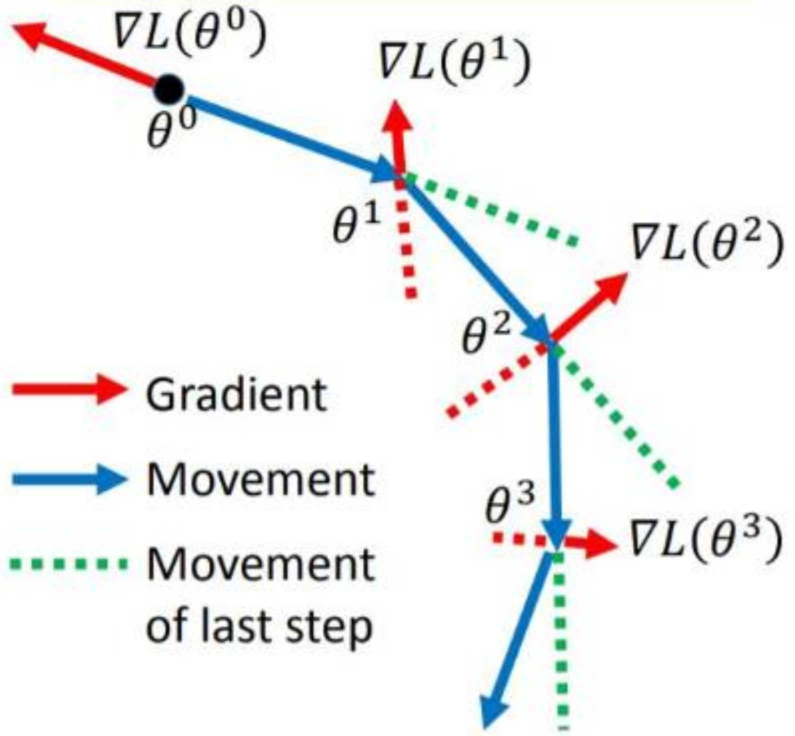
⋮

Stop until $\nabla L(\theta^t) \approx 0$

Credit to 李宏毅老師上課投影片

SGD with Momentum(SGDM)

Movement: movement of last step minus gradient at present



Start at point θ^0

Movement $v^0=0$

Compute gradient at θ^0

Movement $v^1 = \lambda v^0 - \eta \nabla L(\theta^0)$

Move to $\theta^1 = \theta^0 + v^1$

Compute gradient at θ^1

Movement $v^2 = \lambda v^1 - \eta \nabla L(\theta^1)$

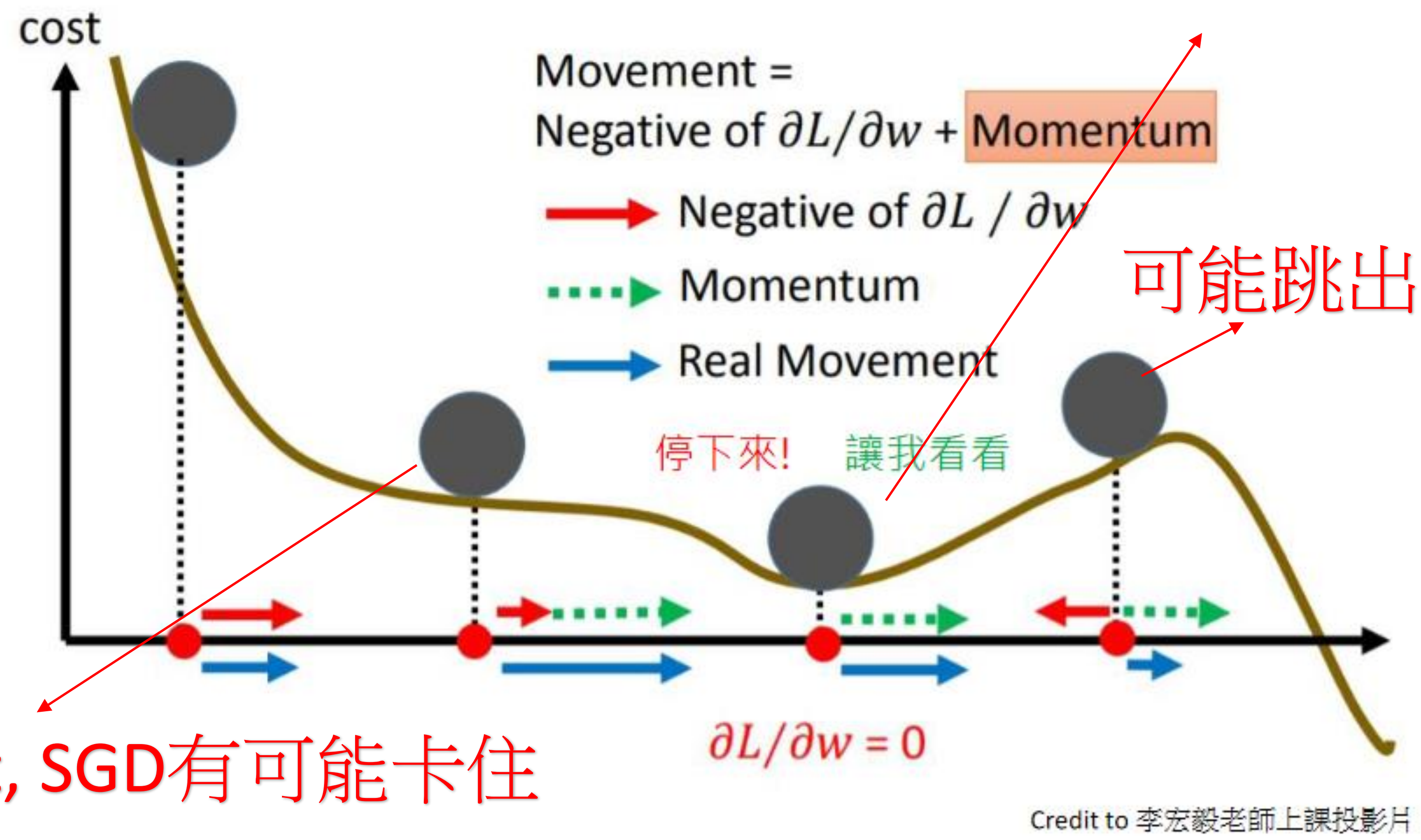
Move to $\theta^2 = \theta^1 + v^2$

Movement not just based on gradient, but previous movement.

Credit to 李宏毅老師上課投影片

Why momentum?

Gradient = 0, 若使用SGD 會停止



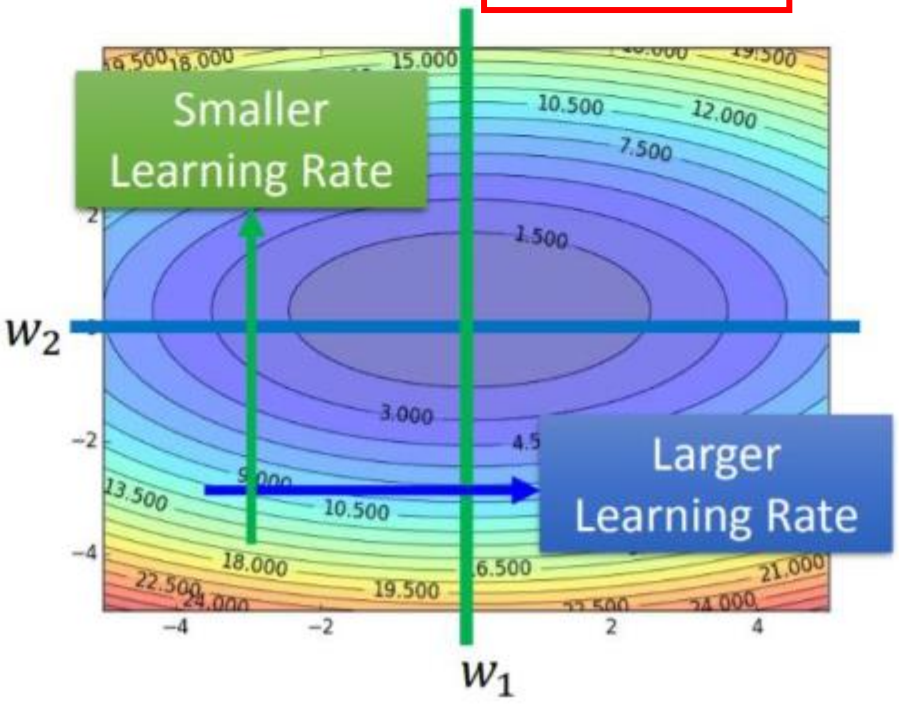
可能跳出local minima

Saddle point, SGD有可能卡住

Adagrad

過去所有time step gradient 的和

$$\theta_t = \theta_{t-1} - \frac{\eta}{\boxed{\sqrt{\sum_{i=0}^{t-1}(g_i)^2}}} g_{t-1}$$



What if the gradients at the first few time steps are extremely large...

Credit to 李宏毅老師上課投影片

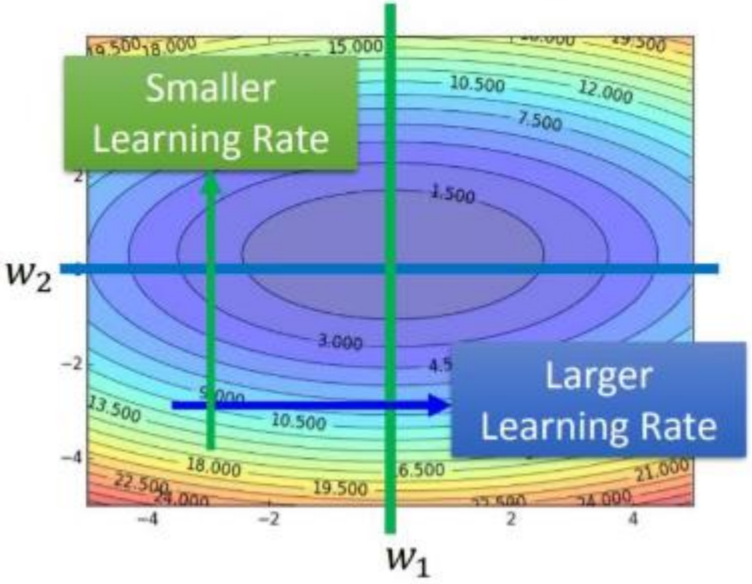
RMSProp

避免gradient 持續變大, 導致Optimizer 卡住

$$\theta_t = \theta_{t-1} - \frac{\eta}{\sqrt{v_t}} g_{t-1}$$

$$v_1 = {g_0}^2$$

$$\boxed{v_t = \alpha v_{t-1} + (1-\alpha)(g_{t-1})^2}$$



Credit to 李宏毅老師上課投影片

Exponential moving average (EMA) of squared gradients is not monotonically increasing

Adam

- SGDM

$$\theta_t = \theta_{t-1} - \eta m_t$$
$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_{t-1}$$

- RMSProp

$$\theta_t = \theta_{t-1} - \frac{\eta}{\sqrt{v_t}} g_{t-1}$$
$$v_1 = {g_0}^2$$
$$v_t = \beta_2 v_{t-1} + (1-\beta_2)(g_{t-1})^2$$

$$\boxed{\theta_t = \theta_{t-1} - \frac{\eta}{\sqrt{\hat{v}_t + \varepsilon}} \hat{m}_t}$$

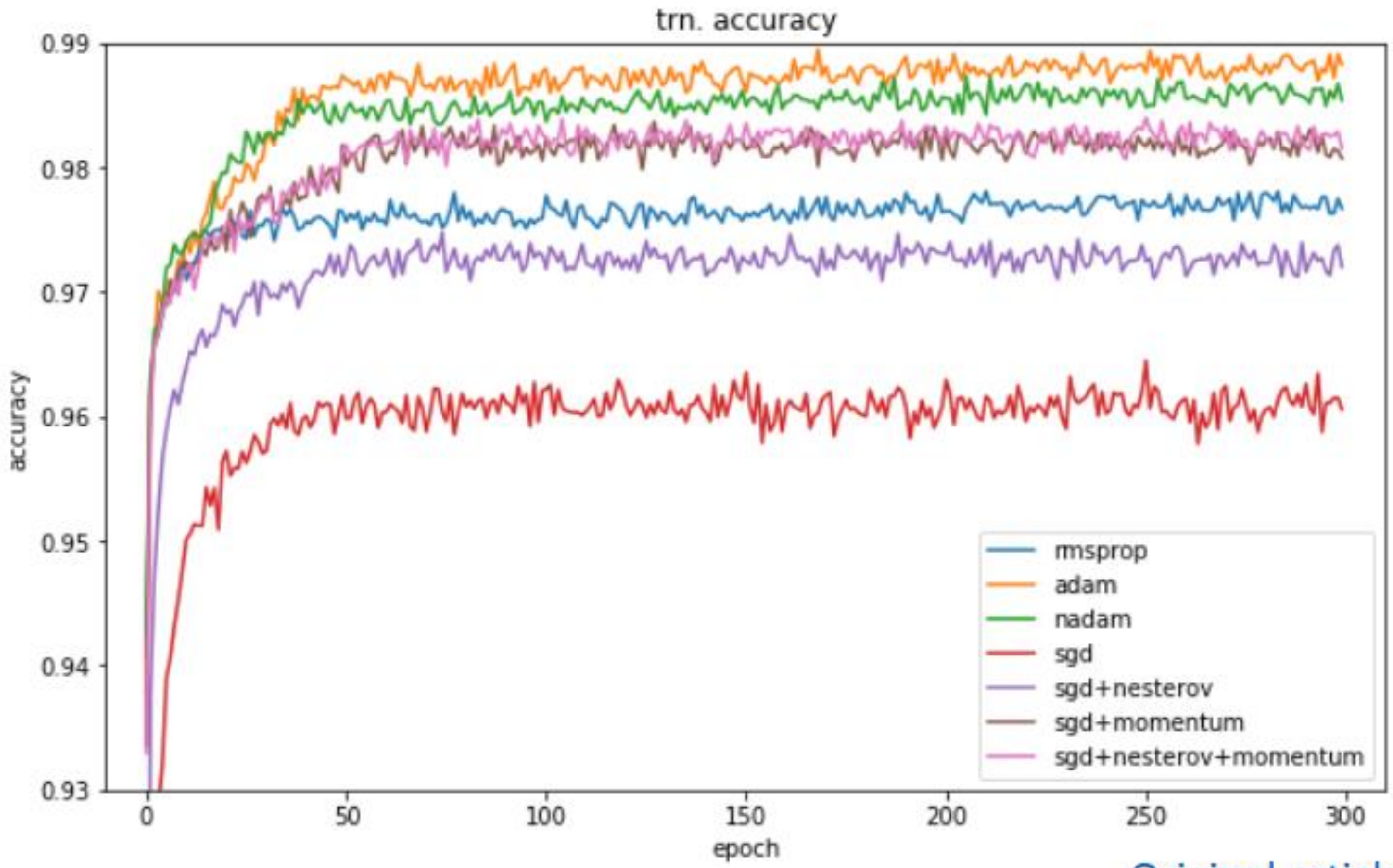
$$\hat{m}_t = \frac{m_t}{1-{\beta_1}^t}$$
$$\hat{v}_t = \frac{v_t}{1-{\beta_2}^t}$$

de-biasing

$$\beta_1 = 0.9$$
$$\beta_2 = 0.999$$
$$\varepsilon = 10^{-8}$$

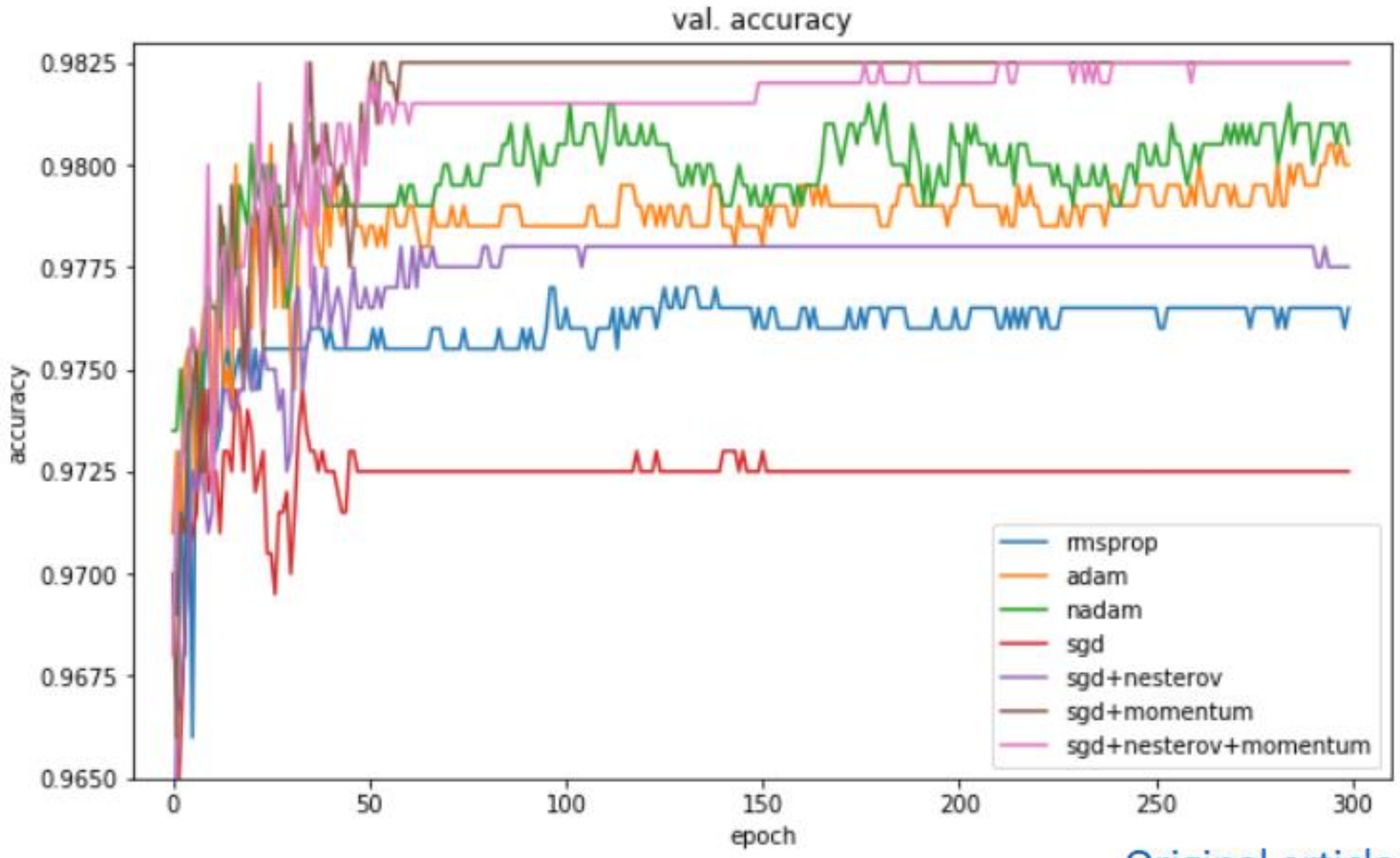
Credit to 簡仲明助教

Adam vs SGDM



Original article

Adam vs SGDM



Original article