

★ AdamW GD. Optimizer

• For each param. keeps track of how small gradients, gives step sizes accordingly

$$g_t \rightarrow \nabla f_\theta(X_t)$$

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

$$v_t \rightarrow \beta_2 \cdot v_{t-1} + (1 - \beta_2) g_t^2$$

$$X_t \rightarrow X_{t-1} - \eta \left(\frac{m_t}{\sqrt{v_t} + \epsilon} \right) - \lambda X_{t-1}$$

Divide by $\sqrt{v_t}$ to avoid ∞ for $X/m_t = 0$

weight decay \rightarrow Smaller bias \rightarrow simpler model

• Store for each param., use w/ a learning rate scheduler.