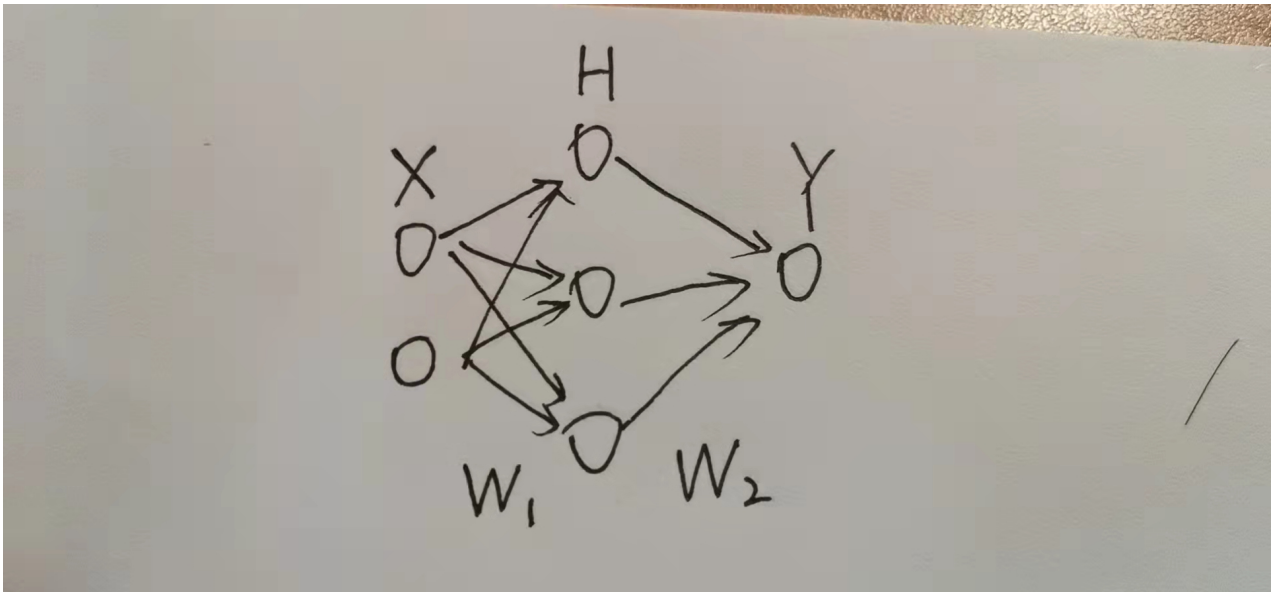
神经网络学习笔记



$$X \overset{W_1}{\to} Z_1 \overset{sigmod}{\to} H \overset{W_2}{\to} Z_2 \overset{sigmod}{\to} \hat{Y} \overset{f}{\to} Cost$$

$$sigmod(XW_1) = H$$
$$sigmod(HW_2) = \hat{Y}$$
$$Cost = (\hat{Y} - Y)^2$$

矩阵乘法求偏导：

$$M \cdot \frac{\partial AX}{\partial X} = A^T \cdot M$$
$$M \cdot \frac{\partial XA}{\partial X} = M \cdot A^T$$

$$\frac{\partial Cost}{\partial W_2} = \frac{\partial Cost}{\partial \hat{Y}} \frac{\partial \hat{Y}}{\partial W_2} = H^T \cdot [2(\hat{Y} - Y) \cdot sigmod'(HW_2)]$$

$$\frac{\partial Cost}{\partial W_1} = \frac{\partial Cost}{\partial \hat{Y}} \frac{\partial \hat{Y}}{\partial H} \frac{\partial H}{\partial W_1} = X^T \cdot \left(\left([2(\hat{Y} - Y) \cdot sigmod'(HW_2)] \cdot W_2^T \right) sigmod'(XW_1) \right)$$

adam优化：

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)