$(y_1, x_1, \cdots, x_T)$ $\quad x_t \in \mathbb{R}^d$

sequence classification / regression

$h_0 = \phi$

$$f(x) = R\sigma(Wx_T + U\sigma(Wx_{T-1} + U\sigma(\cdots \sigma(Wx_1 + Uh_0 + b)\cdots) + b) + b) + c$$

$f(x) = Rh_T + c$, where

$$h_t = \sigma(Wx_t + Uh_{t-1} + b)$$

$$l(f(x), y) = \|f(x) - y\|^2$$

$$\frac{\partial l}{\partial R} = \frac{\partial l}{\partial f} \cdot \frac{\partial f}{\partial R}, \quad \frac{\partial l}{\partial c} = \frac{\partial l}{\partial f} \cdot \frac{\partial f}{\partial c}$$

$$f(x) = R\sigma([V]^T x_T + [U]^T\sigma([W]^{T-1}x_{T-1} + [U]^{T-1}\sigma(\cdots \sigma([W]^1 x_1 + [U]^1 h_0 + [b]^1)\cdots) + [b]^{T-1}) + [b]^T) + c$$

$$\frac{\partial l}{\partial U} = \frac{\partial l}{\partial f}\left(\sum_{t=1}^{T} \frac{\partial f}{\partial h_t} \cdot \frac{\partial h_t}{\partial [U]^t}\right)$$

$$\frac{\partial l}{\partial W} = \frac{\partial l}{\partial f} \cdot \left(\sum_{t=1}^{T} \frac{\partial f}{\partial h_t} \cdot \frac{\partial h_t}{\partial [W]^t}\right)$$

$$\frac{\partial l}{\partial b} = \frac{\partial l}{\partial f} \cdot \left(\sum_{t=1}^{T} \frac{\partial f}{\partial h_t} \cdot \frac{\partial h_t}{\partial [b]^t}\right)$$

$$\frac{\partial f}{\partial h_t} = \frac{\partial f}{\partial h_T} \cdot \frac{\partial h_T}{\partial h_{T-1}} \cdots \frac{\partial h_{t+1}}{\partial h_t}$$

$\frac{\partial h_T}{\partial h_{T-1}}(h_{T-1})$

temporal derivative

$$\frac{\partial h_t}{\partial h_{t-1}} = \mathrm{diag}(\sigma'(a_t))\, U$$

$$h_t = \sigma(\underbrace{Wx_t + Uh_{t-1} + b}_{= a_t})$$

$$\sigma'(a) = \begin{cases} 0 & \text{if } a \le 0 \\ 1 & \text{o.w} \end{cases}, \text{ when } \sigma(a) = \max(0, a)$$

$\sigma : \mathbb{R}^{d'} \to (-1, 1)^{d'}$ $\quad \sigma(a) = \tanh(a)$ ✓

$0 < \tanh'(a) \leq 1$.

$$\left\| \frac{\partial h_t}{\partial h_{t-1}} \right\| = \left\| \operatorname{diag}(\sigma'(a_t)) U \right\| \leq \left\| \operatorname{diag}(\sigma'(a_{t-1})) \right\| \|U\|$$

$$\| \operatorname{diag}(\sigma'(a_t)) \| \|U\| = \max \sigma'(a_t) \|U\| \leq \|U\|$$

$$\left\| \frac{\partial L}{\partial h_t} \right\| \leq \left\| \frac{\partial L}{\partial h_T} \right\| \left\| \frac{\partial h_T}{\partial h_{T-1}} \right\| \cdots \left\| \frac{\partial h_{t+1}}{\partial h_t} \right\| \leq \left\| \frac{\partial L}{\partial h_T} \right\| \cdot \|U\|^{T-t}$$

- exploding gradient may happen when $\|U\| >> 1$.
- vanishing gradient when $\|U\| < 1$.

$$\left\| \frac{\partial l}{\partial [u]^t} \right\| \to 0 \quad \cdots \quad \left\| \frac{\partial l}{\partial u} \right\| \to 0$$

Linear shortcuts.

$$h_t = \sigma(U h_{t-1} + W x_t + b) + \sum_{t'=1}^{t-1} g_{t'}(h_{t'})$$

$$h_t = \sigma(U h_{t-1} + W x_t + b) + \sum_{t'=1}^{t-1} g_{t'}([h_{t'}]^t)$$

$$\frac{\partial h_t}{\partial h_{t'}} = \prod_{t''=1}^{t-t'} \frac{\partial h_{t-t''+1}}{\partial h_{t-t''}} + \frac{\partial h_t}{\partial g_{t'}([h_{t'}]^t)} \cdot \frac{\partial g_{t'}([h_{t'}]^t)}{\partial [h_{t'}]^t}$$

$t' << t$

① Residual connection → $g_{t+1}(a) = Ia$

$g_{t'}(a) = 0$ for $t' < t-1$

$$h_t = \sigma(Uh^{t-1} + Wx^t + b) + h_{t-1}$$

$$= \sigma(Uh_{t-2} + Wx_{t-1} + b) + h_{t-2}$$

$$\vdots$$

$$h_t = \sigma(Uh^{t-1} + Wx^t + b) + \sigma(Uh_{t-2} + Wx_{t-1} + b) + h_{t-2}$$

$$= \sigma(Uh_{t-1} + Wx_t + b) + \sigma(Uh_{t-2} + Wx_{t-1} + b) + \sigma(Uh_{t-3} + Wx_{t-2} + b) + h_{t-3}$$

$$= \cdots$$

$$= \underbrace{\sum_{t' \le t} \sigma(Uh_{t'-1} + Wx_{t'} + b)}_{r}$$

$$\|h_t\| \le \sum_{t' \le t} \|\sigma(Uh_{t'-1} + Wx_{t'} + b)\|$$ formal explosion

② GRU

$$h_t = u_t \odot \sigma_h(Uh^{t-1} + Wx^t + b) + (1-u_t) \odot h_{t-1}$$

$$u_t = \sigma_u(U_u h_{t-1} + W_u x_t + b_u) \in [0,1]^d$$

update gate

$$h_t = u_t \odot \sigma_h(U(r_t \odot h^{t-1}) + Wx^t + b) + (1-u_t) \odot h_{t-1}$$

$$r_t = \sigma_r(U_r h_{t-1} + W_r x_t + b_r) \in [0,1]^d$$

element-wise multiplication