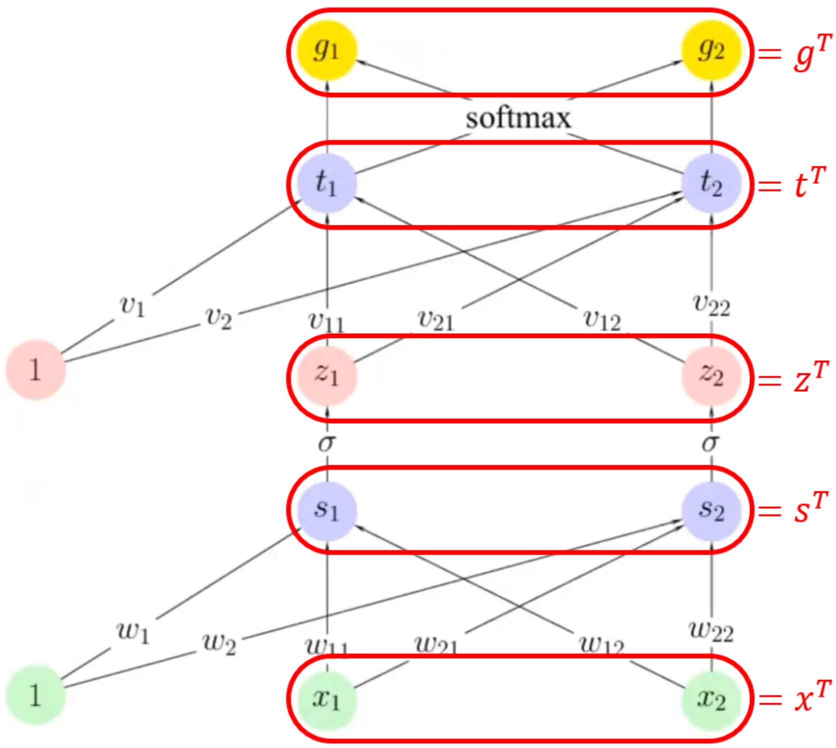
6. Алгоритм обратного распространения ошибки на матричном языке.

Прямой ход

$$
\begin{aligned}
&s_1 = w_1 + w_{11}x_1 + w_{12}x_2 \\
&s_2 = w_2 + w_{21}x_1 + w_{22}x_2 \\
&\begin{pmatrix} s_1 \\ s_2 \end{pmatrix} = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} + \begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\
&s = w + Wx \\
&z = \sigma(s) \\
&\begin{pmatrix} t_1 \\ t_2 \end{pmatrix} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} + \begin{pmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \\
&t = v + Vz \\
&R^{(i)} = logloss(g) \\
&g = softmax(t)
\end{aligned}
$$

σ применяется покомпонентно.



Обратный ход

$$\delta_t = \begin{pmatrix} \delta_{t1} \\ \delta_{t2} \end{pmatrix} = g - \begin{pmatrix} y_1^{(i)} \\ y_2^{(i)} \end{pmatrix}$$

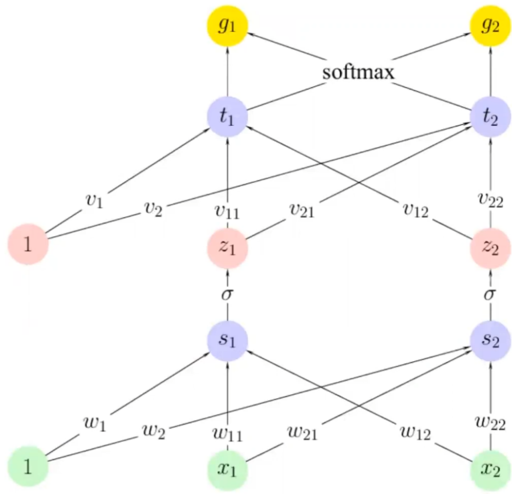
$$\delta_{z1} = \delta_{t1} v_{11} + \delta_{t2} v_{21}$$

$$\delta_{z2} = \delta_{t1} v_{12} + \delta_{t2} v_{22}$$

$$(\delta_{z1}, \delta_{z2}) = (\delta_{t1}, \delta_{t2}) V$$

$$\delta_s = \delta_z \circ \sigma'(s)$$

$$\delta_x = \delta_s W = x \delta_{x_0}$$



$\circ$ обозначает покомпонентное применение.

$g = softmax(V(\sigma(Wx)))$

$R^{(i)} = logloss(g) = \text{logloss}(\text{softmax}(V(\sigma(Wx))))$

Обозначим функцию logloss как L, функцию softmax как g:

$R^{(i)} = L(g(V \cdot \sigma(Wx)))$

$t(x) = V \cdot \sigma(Wx)$

$z(x) = \sigma(Wx)$

$s(x) = Wx$

$R^{(i)} = L(g(t(x)))$

Тогда с помощью матрично-векторного дифференцирования можно получить:

$$\frac{\partial R^{(i)}}{\partial x} = \frac{\partial L}{\partial g}\frac{\partial g}{\partial x} = \frac{\partial L}{\partial g}\frac{\partial g}{\partial t}\frac{\partial t}{\partial x} = \frac{\partial L}{\partial t}\frac{\partial t}{\partial x} = \frac{\partial L}{\partial t}\frac{\partial (V \cdot \sigma(s))}{\partial x} = \frac{\partial L}{\partial t}\frac{\partial (V \cdot \sigma(s))}{\partial \sigma}\frac{\partial \sigma}{\partial x} =$$

$$= \frac{\partial R^{(i)}}{\partial t}\frac{\partial (V \cdot \sigma(Wx))}{\partial \sigma(Wx)}\frac{\partial \sigma(Wx)}{\partial Wx}\frac{\partial Wx}{\partial x}$$

$$\frac{\partial (V \cdot \sigma(Ax))}{\partial \sigma(Wx)} = V, \frac{\partial \sigma(Wx)}{\partial Wx} = \text{diag}(\sigma'), \frac{\partial Wx}{\partial x} = W$$

$$\frac{\partial R^{(i)}}{\partial t} = g - y = \delta_t$$

$$\frac{\partial R^{(i)}}{\partial x} = (g - y) \cdot V \cdot \text{diag}(\sigma') \cdot W = \delta_x$$

$$\frac{\partial R^{(i)}}{\partial W} = (g - y) \cdot V \cdot \text{diag}(\sigma') \cdot x = \delta_s \cdot x$$

$$\frac{\partial R^{(i)}}{\partial V} = (g - y) \cdot \sigma(Wx) = \delta_t \cdot z$$

Нас интересует вектор частных производных $R^{(i)}$ по каждой компоненте вектора x. Для этого находим частные производные и делаем шаг стохастического градиентного спуска. Шаг делается в пространстве весов (коэффициент $\gamma > 0$ - learning rate).

$$
\begin{aligned}
&W \leftarrow W - \gamma \frac{\partial R^{(i)}}{\partial W}, w \leftarrow w - \gamma \frac{\partial R^{(i)}}{\partial w}, \\
&V \leftarrow V - \gamma \frac{\partial R^{(i)}}{\partial V}, v \leftarrow v - \gamma \frac{\partial R^{(i)}}{\partial v}
\end{aligned}
$$