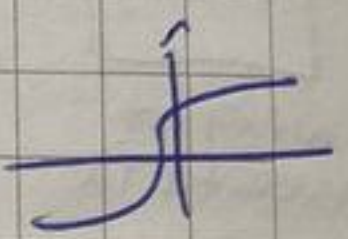
Логистическая функция $f(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{1+e^x}$, $x \in \mathbb{R}$



Обобщение на высокие размерности:

Softmax $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ $S(x) = \begin{pmatrix} \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \end{pmatrix} \in \mathbb{R}^n$

Свойства

1) $\|S(x)\|_1 = \sum_i \frac{e^{x_i}}{\sum e^{x_j}} = 1$

$\Rightarrow$ вероятностная интерпретация

2) вычислительная стабильность (numerical stability) Softmax'а

– это $\frac{e^{x_j+D}}{\sum e^{x_k+D}} = \frac{e^D \cdot e^{x_j}}{e^D \cdot \sum_k e^{x_k}} = \frac{e^{x_j}}{\sum_k e^{x_k}}$

Свойство исп. при "нормализации" данных, потому что в Python'

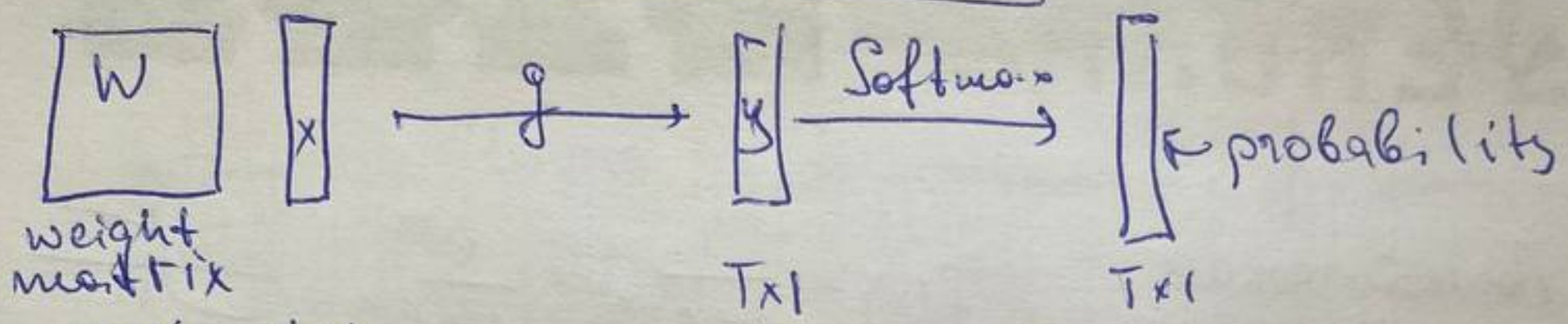
$\max(\text{float64}) = 10^{38}$

$\Rightarrow$ softmax $(1000) \sim e^{1000} = NaN$

Нормализуем: $e^{x_j} \longrightarrow e^{x_j - \max_i(x_i)}$ $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$

(в предположении, что данные сгруппированы)

Вычисление градиента (сложной ф-ции)



$$\nabla S(g(W)) = \nabla(Sg)\,(\nabla g W)$$

① для Softmax'а

$$\nabla S = \begin{bmatrix} \nabla_1 S_1 & \dots & \nabla_N S_1 \\ \vdots & & \\ \nabla_1 S_N & & \nabla_N S_N \end{bmatrix}$$ якобиан

$$\nabla_i S_j = \frac{\partial S_j}{\partial x_i} = \frac{\partial}{\partial x_i}\left(\frac{e^{x_j}}{\sum_{k=1}^{N} e^{x_k}}\right)$$

$$= \begin{cases} i \neq j: & \frac{-e^{x_i} e^{x_j}}{(\Sigma x_k)^2} = -S_i \cdot S_j \\ i = j & \frac{e^{x_i}(\Sigma x_k) - e^{2x_i}}{(\Sigma x_k)^2} = \frac{e^{x_i}}{\Sigma} \cdot \frac{\Sigma - e^{x_i}}{\Sigma} = S_i(1 - S_i) \end{cases}$$

$$= \begin{cases} -S_i S_j, & i \neq j \\ S_i(1 - S_i), & i = j \end{cases} = S_j(\delta_{ij} - S_i)$$

$$\delta_{ij} = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}$$

② $$\nabla g = \begin{bmatrix} \nabla_1 g_1 & \dots & \nabla_{NT} g_1 \\ \nabla_1 g_T & \dots & \nabla_{NT} g_T \end{bmatrix}$$

$g_i = W_{i1}x_1 + \dots + W_{iN}x_N$ (матричное умножение)

$$\nabla_{ij} g_t = \frac{\partial(W_{i1}x_1 + \dots)}{\partial W_{ij}} = \begin{cases} x_i, & i = t \\ 0, & i \neq t \end{cases}$$

Итого для сложной ф-ции:

$$\nabla_{ij}\left(S_t(g(w))\right) = \sum_k \nabla_k S_t \underbrace{\nabla_{ij} g_k}_{=0,\ i\neq k} =$$

$$= \nabla S_t \underbrace{\nabla_{ij} g_i}_{\textcircled{2}=x_j} = \underbrace{\nabla_i S_t}_{\textcircled{1}} x_j = S_t(\delta_{ti} - S_i) x_j$$

Итого: • градиент от Softmax'а - это сам Softmax (почти)

• Градиент линейной функции от Softmax'а - это сам Softmax (почти)

⇒ НЕТ МАТРИЧНОГО УМНОЖЕНИЯ! ♥