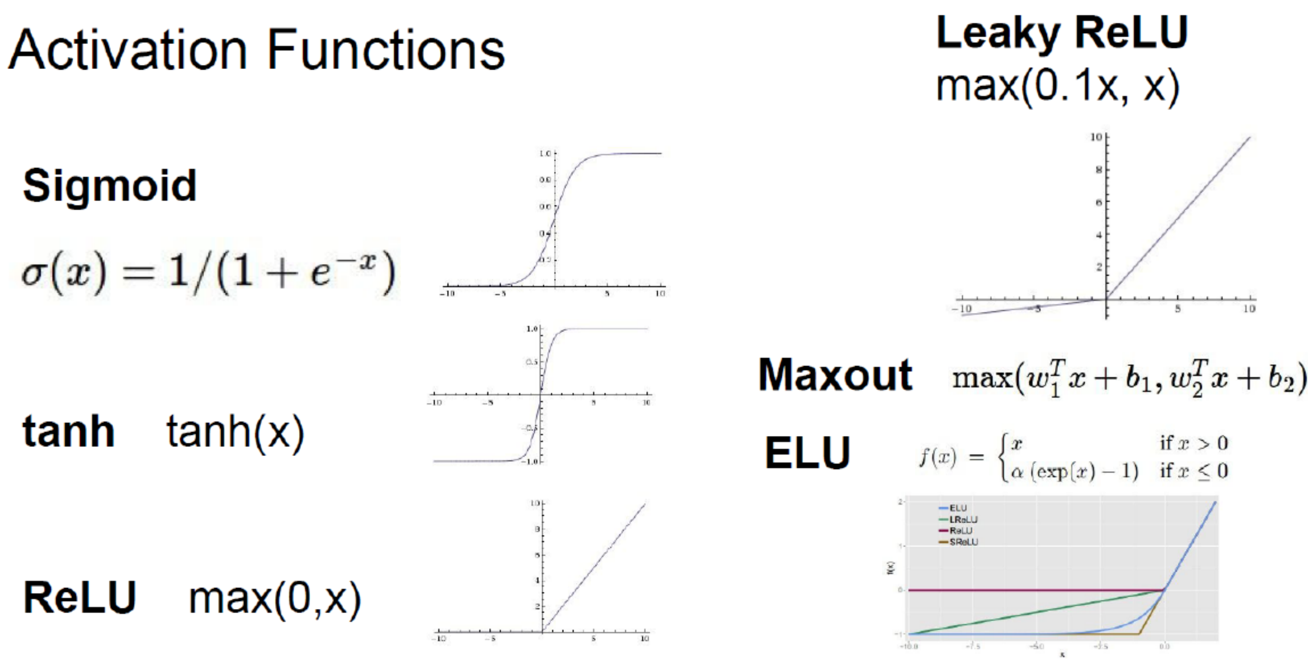
Neural Network and Back Propagation

Activation functions



Gradient Descent

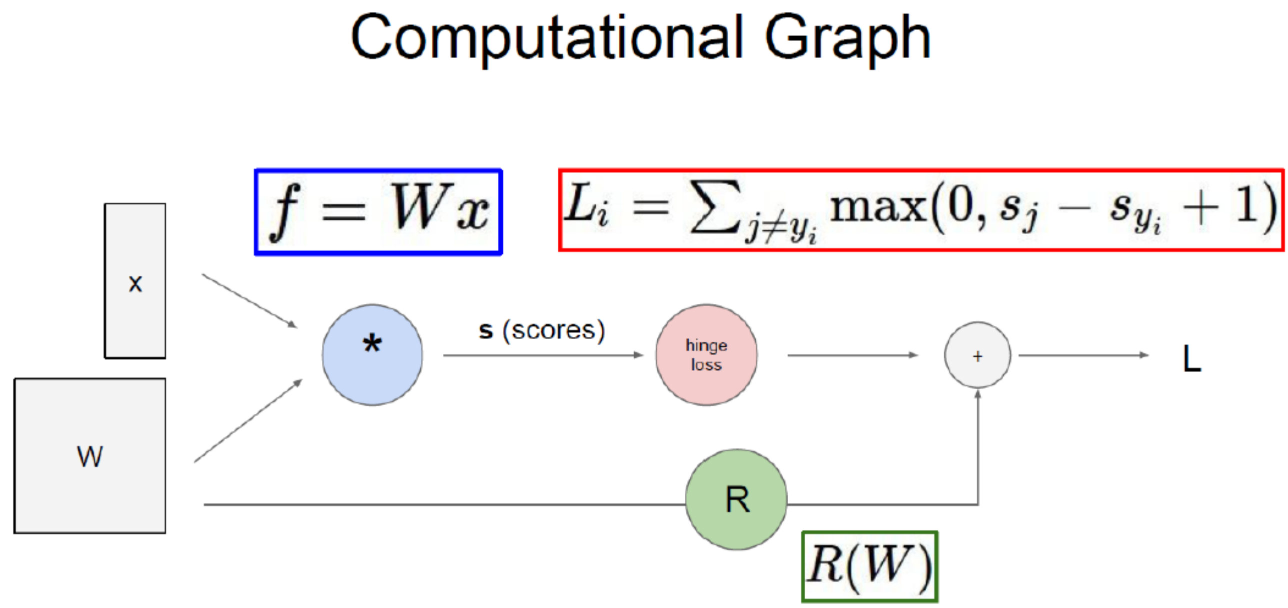
$$s = f(x; W) = Wx \qquad \text{scores function}$$

$$L_i = \sum_{j \neq y_i} \max(0, s_j - s_{y_i} + 1) \qquad \text{SVM loss}$$

$$L = \frac{1}{N} \sum_{i=1}^{N} L_i + \sum_k W_k^2 \qquad \text{data loss + regularization}$$

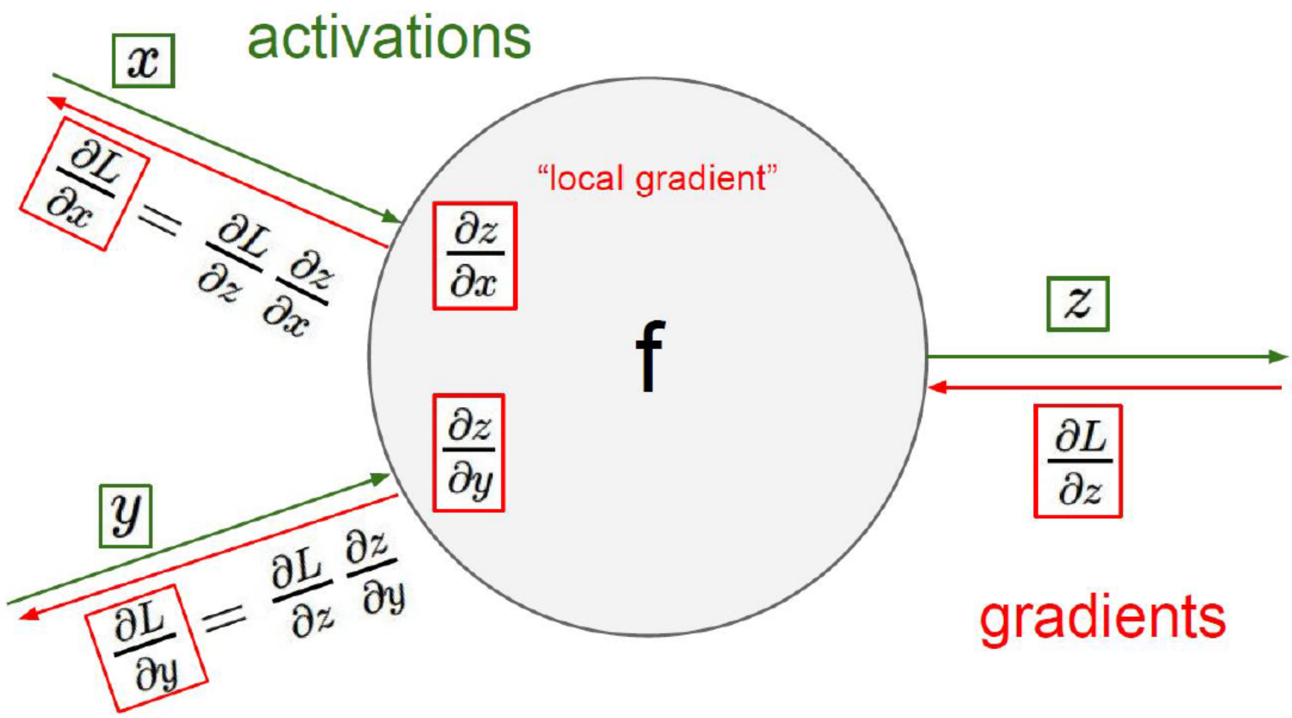
want $\boxed{\nabla_W L}$

注：对于第三个损失函数，L_i 代表单个测试数据的损失；$\sum_k W_k^2$ 代表对权重的累加，这是一种正则化的方法，因为我们有很多种 W 可以选择，但太过复杂的 W 可能导致过拟合，因此使用这种方式可以使模型选择简单的 W。最后我们希望求出梯度。



上面是 SVM 的计算图，R(W) 就是上面的 $\sum_k W_k^2$。

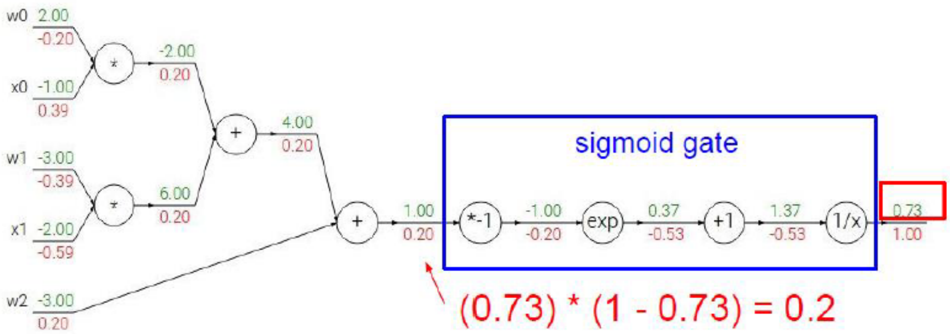
BP



上面是 BP 算法中 chain rule 的一个示例。

$$f(w,x) = \frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}} \qquad \boxed{\sigma(x) = \frac{1}{1+e^{-x}}} \quad \text{sigmoid function}$$

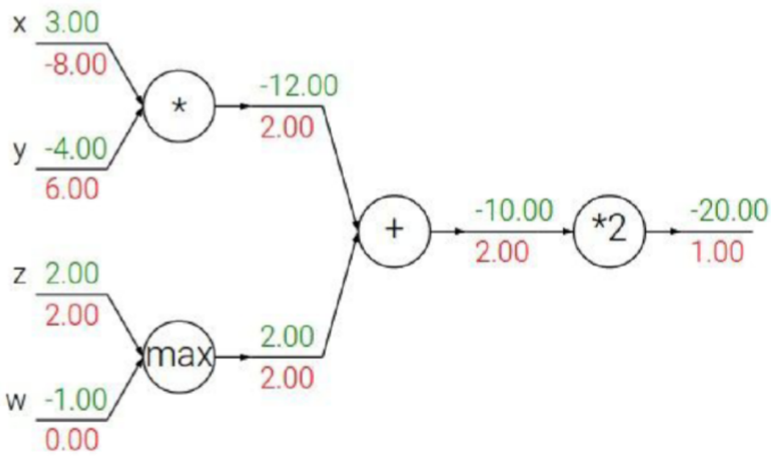
$$\frac{d\sigma(x)}{dx} = \frac{e^{-x}}{(1+e^{-x})^2} = \left(\frac{1+e^{-x}-1}{1+e^{-x}}\right)\left(\frac{1}{1+e^{-x}}\right) = (1-\sigma(x))\,\sigma(x)$$



这里可以把 sigmoid 部分看作一个整体，求出梯度。

Patterns in backward flow

add gate: gradient distributor
max gate: gradient router
mul gate: gradient… “switcher”?



加法门：梯度分配器，分配相同的梯度值。

最大值门：梯度路由，反向传播时，它会把梯度值分配给输入值最大的线路。

乘法门：梯度转换器，把梯度转换成不一样的值。

Exercise

Pooling units take n values x_i, $\mathrm{i} \in [1, \mathrm{n}]$ and compute a scalar output whose value is invariant to permutations of the inputs.

1. The Lp-pooling module takes positive inputs and computes $\mathrm{y} = (\sum_i x_i^p)^{\frac{1}{p}}$, assuming we know that $y' = \frac{\partial L}{\partial y}$, what is $x_i' = \frac{\partial L}{\partial x_i}$?
2. The log-average module computes $\mathrm{y} = \frac{1}{\beta} \ln(\frac{1}{n} \sum_i \exp(\beta x_i))$, assuming we know that $y' = \frac{\partial L}{\partial y}$, what is $x_i' = \frac{\partial L}{\partial x_i}$?

第一问:

$$x_i' = \frac{\partial L}{\partial x_i} = \frac{\partial L}{\partial y} \cdot \frac{\partial y}{\partial x_i}$$

$$= \frac{\partial L}{\partial y} \cdot \left[\frac{\partial (\sum_i x_i^p)^{\frac{1}{p}}}{\partial (\sum_i x_i^p)} \cdot \frac{\partial (\sum_i x_i^p)}{\partial x_i^p} \cdot \frac{\partial x_i^p}{\partial x_i} \right]$$

$$= \frac{\partial L}{\partial y} \cdot \left[\frac{1}{p} (\sum_i x_i^p)^{1-\frac{1}{p}} \cdot 1 \cdot p x_i^{p-1} \right]$$

$$= \frac{\partial L}{\partial y} \cdot (\sum_i x_i^p)^{\frac{1-p}{p}} \cdot x_i^{p-1}$$

第二问:

$$x_i' = \frac{\partial L}{\partial x_i} = \frac{\partial L}{\partial y} \cdot \frac{\partial y}{\partial x_i}$$

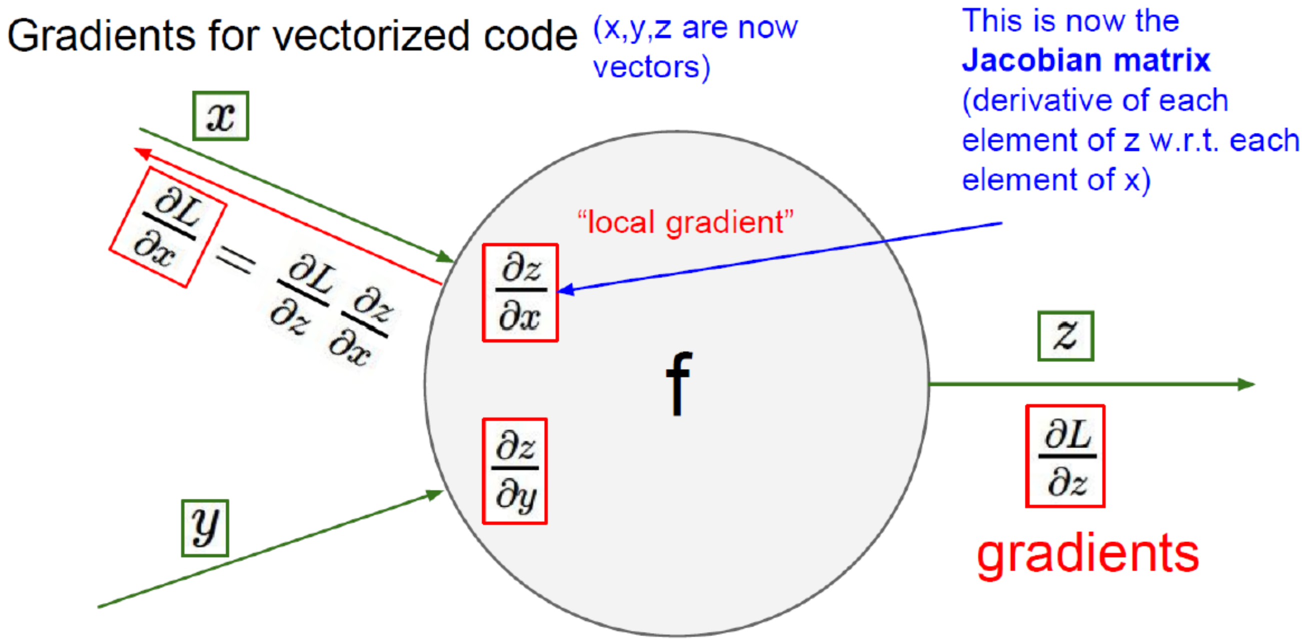
$$= \frac{\partial L}{\partial y} \cdot \left\{ \underset{\substack{\Downarrow \\ \frac{1}{\beta}}}{\frac{\partial\{\frac{1}{\beta}\ln[\frac{1}{n}\Sigma_i \exp(\beta x_i)]\}}{\partial\{\ln[\frac{1}{n}\Sigma_i \exp(\beta x_i)]\}}} \cdot \underset{\substack{\Downarrow \\ \frac{1}{\frac{1}{n}\Sigma_i \exp(\beta x_i)}}}{\frac{\partial\{\ln[\frac{1}{n}\Sigma_i \exp(\beta x_i)]\}}{\partial[\frac{1}{n}\Sigma_i \exp(\beta x_i)]}} \right.$$

$$\cdot \underset{\substack{\Downarrow \\ \frac{1}{n}}}{\frac{\partial[\frac{1}{n}\Sigma_i \exp(\beta x_i)]}{\partial[\Sigma_i \exp(\beta x_i)]}} \cdot \underset{\substack{\Downarrow \\ 1}}{\frac{\partial[\Sigma_i \exp(\beta x_i)]}{\partial[\exp(\beta x_i)]}} \cdot \underset{\substack{\Downarrow \\ \exp(\beta x_i)}}{\frac{\partial[\exp(\beta x_i)]}{\partial(\beta x_i)}}$$

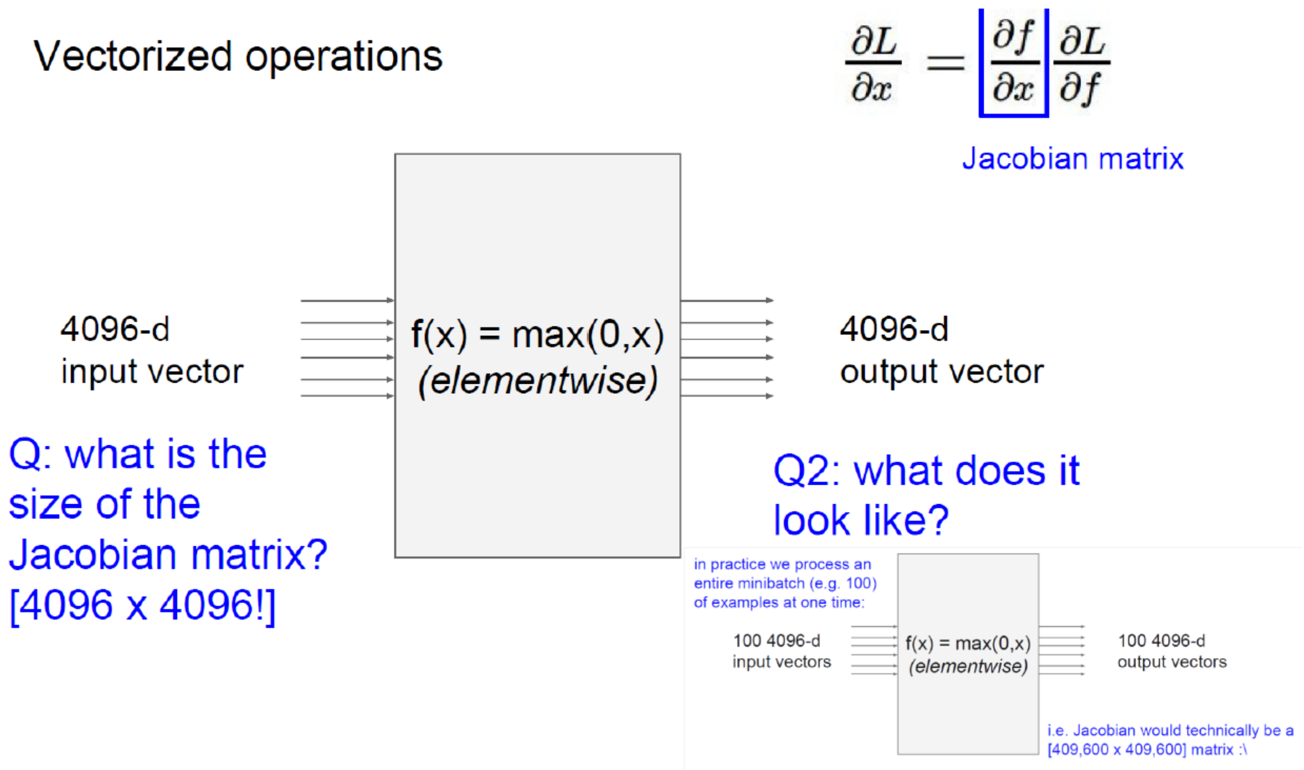
$$\left. \cdot \underset{\substack{\Downarrow \\ \beta}}{\frac{\partial(\beta x_i)}{\partial x_i}} \right\}$$

$$\Rightarrow \frac{\partial L}{\partial y} \cdot \frac{\exp(\beta x_i)}{\Sigma_i \exp(\beta x_i)}$$

Gradients for vector

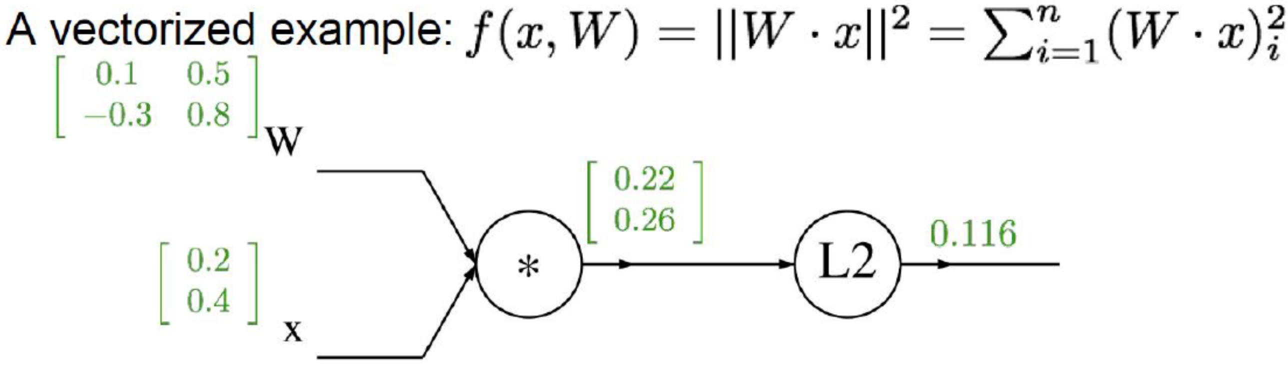


之前输入单个值的时候，梯度是单个值；现在输入的是向量，梯度也变成同等大小的向量，即 jacobian matrix (其中是 local gradient)。



注：jacobian matrix 的 shape 和输入的 shape 一致。

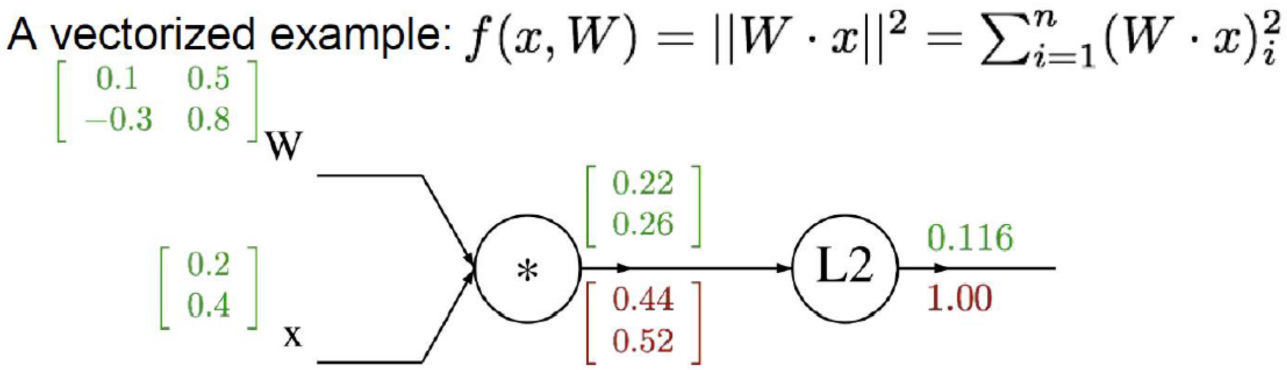
如下面的例子所示，我们对其进行 BP 算法：



$$q = W \cdot x = \begin{pmatrix} W_{1,1}x_1 + \cdots + W_{1,n}x_n \\ \vdots \\ W_{n,1}x_1 + \cdots + W_{n,n}x_n \end{pmatrix}$$

$$f(q) = ||q||^2 = q_1^2 + \cdots + q_n^2$$

前向传播之后，进行反向传播:



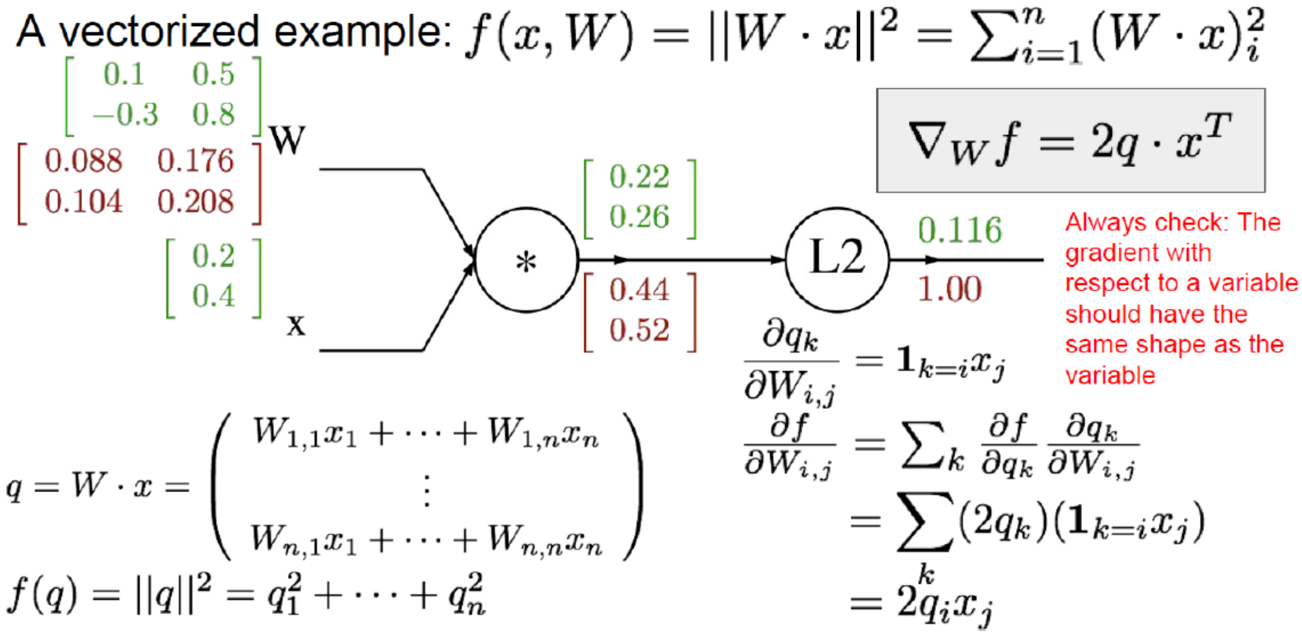
$$q = W \cdot x = \begin{pmatrix} W_{1,1}x_1 + \cdots + W_{1,n}x_n \\ \vdots \\ W_{n,1}x_1 + \cdots + W_{n,n}x_n \end{pmatrix}$$

$$f(q) = ||q||^2 = q_1^2 + \cdots + q_n^2$$

$$\frac{\partial f}{\partial q_i} = 2q_i$$

$$\boxed{\nabla_q f = 2q}$$

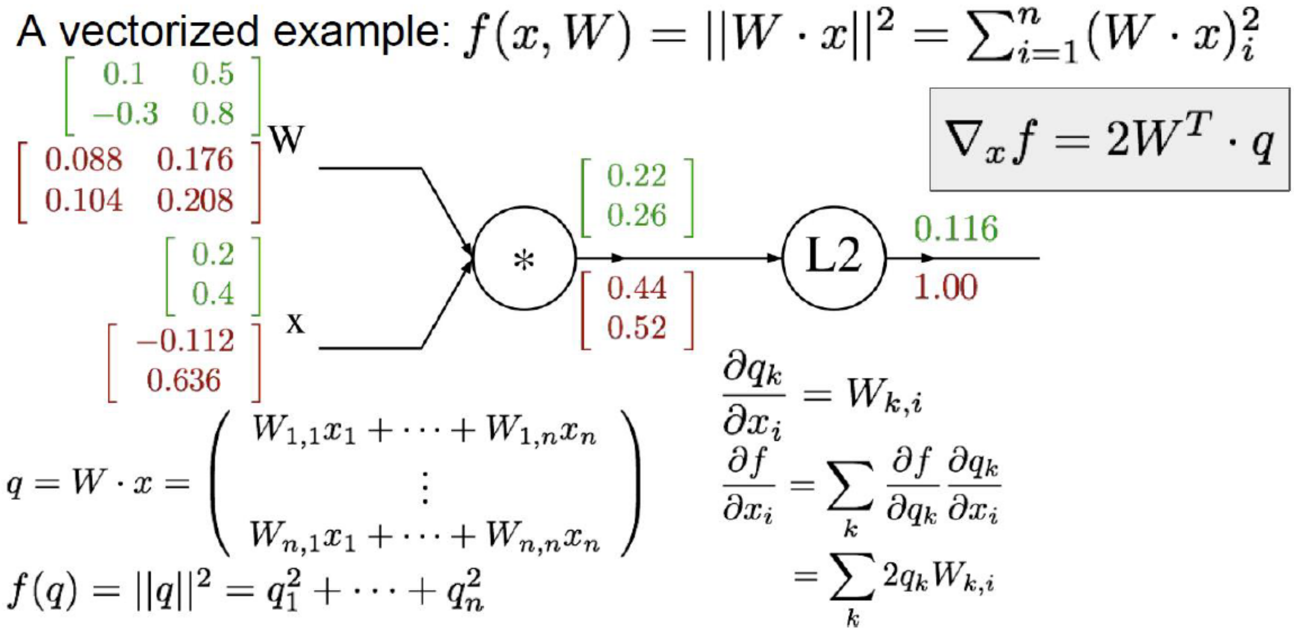
之后求 W 的梯度:



注：比如要求 $W_{1,1}$ 的梯度，包含 $W_{1,1}$ 的式子只有 q_1，因为 $q_1 = W_{1,1}x_1 + \ldots W_{1,n}x_n$。

因此，可以得到：$\frac{\partial f}{\partial w_{i,j}} = \frac{\partial f}{\partial q_k}\cdot\frac{\partial q_k}{\partial w_{i,j}} = 2q_k x_j$。

最后是 x 的梯度：



注：比如要求 x_1 的梯度，所有 q 中都包含 x_1。因此我们要考虑所有的 q。

包含 x_1 的式子：$f = (w_{1,1}x_1)^2 + \ldots + (w_{n,1}x_1)^2$，我们要求这个式子对 x_1 的偏导，就要求出每一个的偏导，再相加。先求 q_1 中 x_1 的偏导：$\frac{\partial q_1}{\partial x_1} = 2(w_{1,1}x_1)w_{1,1} = 2q_1 w_{1,1}$。

对每一个 q 都这样操作，f 对 x_1 的偏导最后就是：
$\frac{\partial f}{\partial x_1} = 2q_1 w_{1,1} + \ldots + 2q_n w_{n,1} = \sum_k 2q_k W_{k,1}$。

因此，对于 i 个 x，最后就得到了 $\sum_k 2q_k W_{k,i}$。