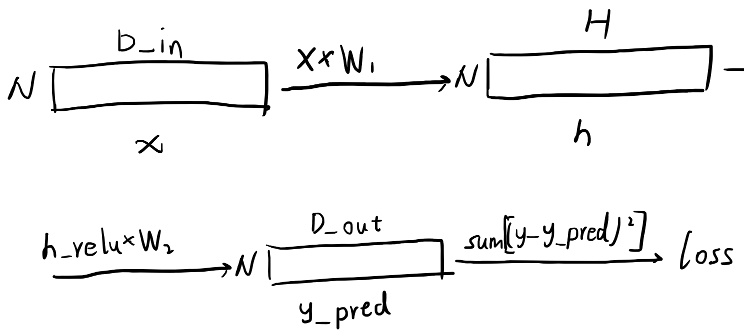
正向传播



反向传播（不妨假设N=1）约定"·"表示对位相乘，"×"表示矩阵乘法，"d"表示偏微分

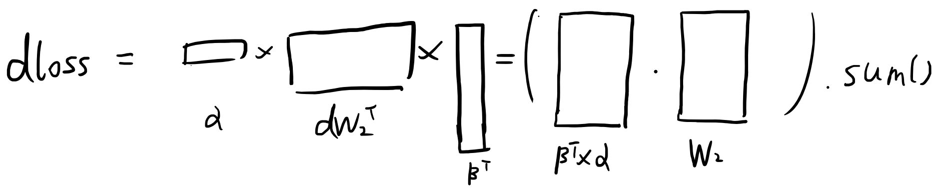
$$dloss = d(sum(y-y\_pred)^2) = sum[2(y\_pred-y)\cdot d(y\_pred)] = 2(y\_pred-y)\times d(y\_pred)^T$$

$$= 2(y\_pred-y)\times d(W_2^T)\times h\_relu^T$$

记 $2(y\_pred-y)=\alpha$，$h\_relu=\beta$. 则α,β均为行向量.

$$dloss = \alpha\times d(W_2^T)\times\beta^T = sum[(\alpha^T\times\beta)\cdot d(W_2^T)] = sum[(\beta^T\times\alpha)\cdot d(W_2)]$$

故W_2的梯度即为 $\beta^T\times\alpha = h\_relu^T\times 2(y\_pred-y)$



同理，有

$$dloss = 2(y\_pred-y)\times d(y\_pred^T) = 2(y\_pred-y)\times W_2^T\times d(h\_relu^T)$$

$$= 2(y\_pred-y)\times W_2^T\times diag(h>0)\times d(h^T) = 2(y\_pred-y)\times W_2^T\times diag(h>0)\times d(W_1^T)\times X^T$$

同理可得W_1的梯度为 $x^T\times 2(y\_pred-y)\times W_2^T\times diag(h>0) = x^T\times[2(y\_pred-y)\times W_2^T)\cdot(h>0)]$

故令 $W_1 -= grad\_W_1\cdot learning\_rate$　即完成一次梯度下降过程.

$W_2 -= grad\_W_2\cdot learning\_rate$