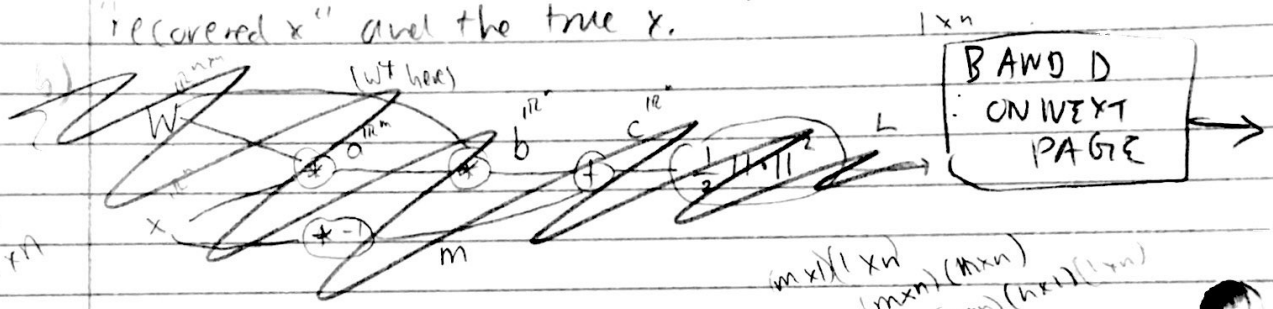
Scalar chain: $\frac{dz}{dx} = \frac{dz}{dy}\frac{dy}{dx}$ $(W^TWx - x)(Wx + W^Tx)$ $n \times 1$

$\frac{dz}{dx} = \frac{dy}{dx}\frac{dz}{dy}$ $a = m \times 1$ $c = n \times 1$ $W = m \times n$

#1

a) We know that the transformation Wx would preserve information perfectly if we "reverse" this transpose via $W^T(Wx)$ and get an approximation a that is "close" to x, for some closeness metric. Essentially if we approximate x by W^TWx where Wx reduces the dimensionality of x, then W would have to be selected such that it learns important features about x, so that it can be recovered. In order to do this, we'd want to minimize a cost such as the L2 squared distance between our "recovered x" and the true x.



c) The two paths to w should be summed. This is because we can think of the 2 paths out of w as separate functions that w affects. For example, we could assign each f and g as functions coming out of the w's node. Then, f and g are both functions of w, and contribute some incoming gradient $\nabla_w f$ and $\nabla_w g$ to the node w, so these gradients should be summed: $\boxed{\nabla_w f + \nabla_w g}$

d) Let $\frac{dL}{dl} = 1$.

$c \in \mathbb{R}^n$.

$\frac{dL}{dc}$ (scalar wrt to vector) $= \frac{1}{2}\|c\|_2^2 = c$

$\frac{dL}{db} = \frac{dc}{db}\frac{dL}{dc} = \frac{dc}{db}c = \frac{d}{db}(b+m)(c) = c$

dL/dW^T (from b) $= \left(\frac{db}{dW}\frac{dL}{db}\right) = \frac{d}{dW}W^Ta(c) = \boxed{ac^T}$

$\frac{dL}{da} = \frac{db}{da}\frac{dL}{db} = \frac{d}{da}W^Ta(c) = Wc$

$\frac{dL}{dW}$ (from a) $= \frac{da}{dW}\frac{dL}{da} = \frac{d}{dW}(Wx)(Wc) = \boxed{(Wcx^T)}$

$\rightarrow \frac{dL}{dW} = ac^T + Wcx^T$

$= Wx(W^TWx - x)^T + W(W^TWx - x)x^T$

$n \times m$ $n \times m$ $a \in \mathbb{R}^{m \times 1}$ $n \times 1$ $1 \times n$ $1 \times m$ $n \times 1$ n $1 \times m$

$(ac^T)^T = c^T a^T$ $(a^T c)^T = c^T a$ $1_n a^T$

a) $\frac{dL}{dL} = 1$

$\frac{dL}{dc} = \frac{dL}{dc} \frac{d}{dc}\left(\frac{1}{2}\|c\|_2^2\right) = c$ $n \times m$

$\frac{dL}{db} = \frac{dL}{dc}\frac{dc}{db} = \frac{d}{db}(b+m)x = c$ dx

$\frac{dL}{dW^T} = \frac{dL}{dc}\frac{dc}{db}\frac{db}{dW^T} = \frac{d}{dW^T}(W^T a)(c)$ $W^T a = (n \times m)(m \times 1)$ $n \times 1$ $\mathbb{R}^m$

W $n \times m$ transpose $m \times n$ $a \in \mathbb{R}^{m \times 1}$ $b \in \mathbb{R}^{n \times 1}$ $c \in \mathbb{R}^{n \times 1}$ $\frac{1}{2}\|\cdot\|_2^2$ L x $n \times 1$ $*-1$ m $1_n x^T$

d) $dL/dL = 1$

$\frac{dL}{dc} = \frac{d}{dc}\left(\frac{1}{2}\|c\|_2^2\right) = c$ (as shown in class)

$\frac{dL}{db} = c$ as well, since the $+$ node acts like a gradient gate that lets the upstream through

$\frac{dL}{da} = \frac{db}{da}\frac{dL}{db} = \frac{d}{da}(W^T a)\frac{dL}{db} = \boxed{Wc}$

$\frac{dL}{dW}$ (from a) $= \frac{dL}{da}\frac{da}{dW} = Wc\frac{d}{dW}Wx$

Since $\frac{dL}{dW} \in \mathbb{R}^{m \times n}$, we can infer that $\frac{dL}{dW} = Wc\frac{d}{dW}Wx = \boxed{Wcx^T}$

$\frac{dL}{dW^T}$ (from b) $= \frac{dL}{db}\frac{db}{dW^T} = c\frac{d}{dW^T}W^T a$. Since $\frac{dL}{dW^T} \in \mathbb{R}^{n \times m}$,

we can infer that $dL/dW^T = \boxed{ca^T}$

Using $\frac{dL}{dW^T} = \left(\frac{dL}{dW}\right)^T$, $\frac{dL}{dW} = ac^T$. Sum both terms: $\frac{dL}{dW} = ac^T + Wcx^T$

$\rightarrow \boxed{Wx(W^TWx - x)^T + W(W^TWx - x)x^T}$