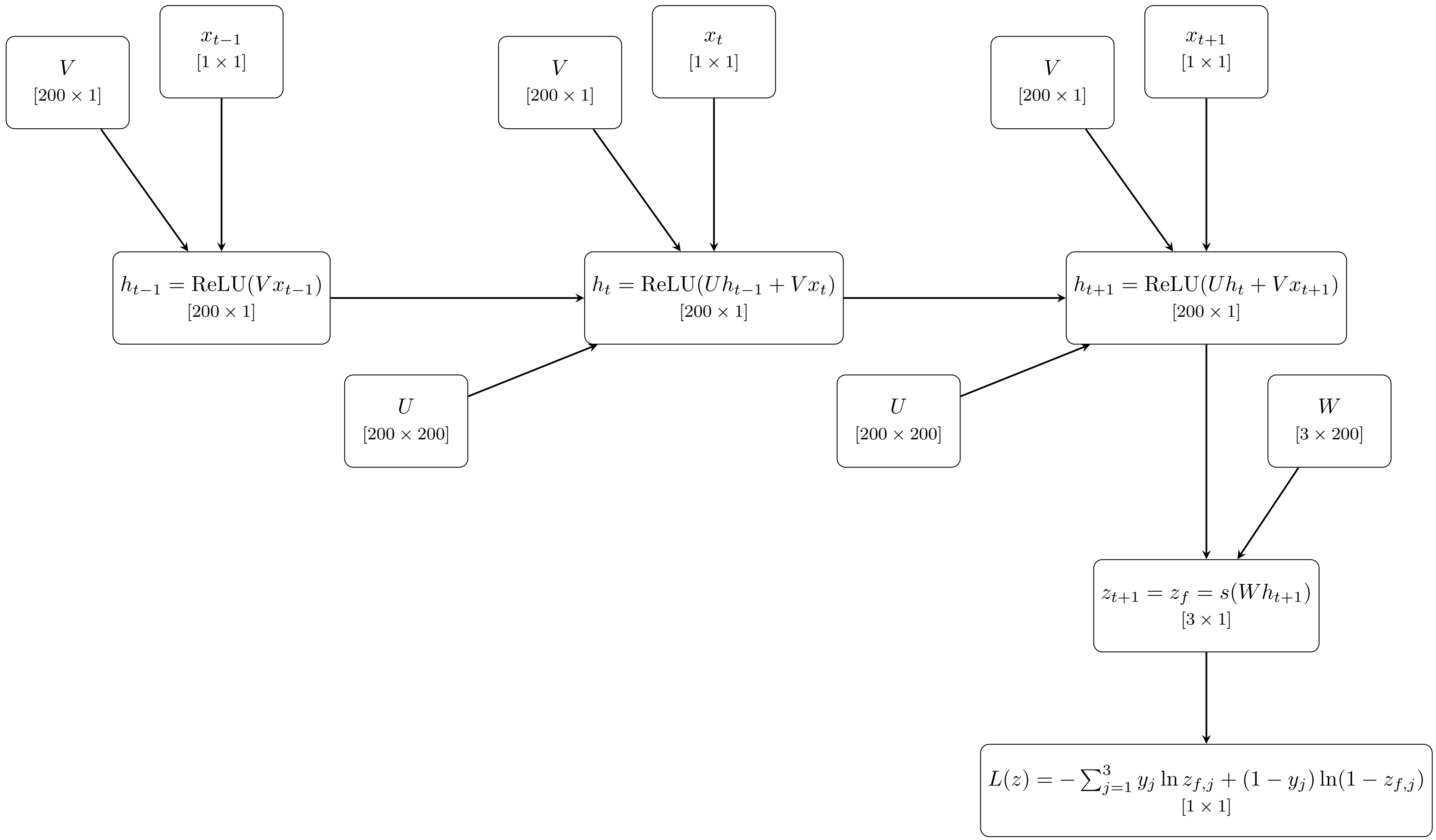
RNN Diagram - ReLU Activation on Hidden Layer

The constructed neural net follows the diagram below. $x \in \mathbb{R}^{24}$, a feature vector with (4 hours $\times$ 6 intervals) = 24 features corresponding to occupancies every 10 minutes for 4 hours before the current time. For clarity, shapes of the solutions are given in brackets. Subscript j denotes a row index. A_j^T indicates the transpose of A_j.



$A_k = Uh_{k-1} + Vx_k$

$\nabla_V L =$
$\left(\nabla_{z_f} L\right)\left(\nabla_V z_{t+1}\right) =$
$\left(\nabla_{z_f} L\right)\left(\nabla_{h_{t+1}} z_{t+1}\right)\left(\nabla_V h_{t+1}\right) =$
$\left(\frac{1-y_j}{1-z_j} - \frac{y_j}{z_j}\right)\left(z_j(1-z_j)W_j^T\right)\left(\nabla_V h_{t+1}\right) =$
$(z_j - y_j)\, W_j^T \left(\nabla_V h_{t+1}\right) =$
$W_j^T\,(z_j - y_j)\left(\nabla_V h_{t+1}\right) =$
$W^T(z-y)\left(\nabla_V h_{t+1}\right) =$
$W^T(z-y)\left(\nabla_{A_{t+1}} h_{t+1}\right)\left(\nabla_V A_{t+1}\right) =$
$W^T(z-y)\left(\nabla_{A_{t+1}} h_{t+1}\right)\left(\nabla_V Uh_t + \nabla_V Vx_{t+1}\right) =$
$W^T(z-y)\left(\nabla_{A_{t+1}} h_{t+1}\right)\left(\nabla_V Uh_t\right) + W^T(z-y)\left(\nabla_{A_{t+1}} h_{t+1}\right)\left(\nabla_V Vx_{t+1}\right) =$
$W^T(z-y)\left(\nabla_{A_{t+1}} h_{t+1}\right)\left(U\nabla_V h_t\right) + W^T(z-y)\left(\nabla_{A_{t+1}} h_{t+1}\right)\left(I^{[200\times 200]}x_{t+1}\right) =$
$W^T(z-y)\left(\nabla_{A_{t+1}} h_{t+1}\right)U\left(\nabla_{A_t} h_t\right)\left(\nabla_V A_t\right) + W^T(z-y)\left(\nabla_{A_{t+1}} h_{t+1}\right)\left(I^{[200\times 200]}x_{t+1}\right) =$
$W^T(z-y)\left(\nabla_{A_{t+1}} h_{t+1}\right)U\left(\nabla_{A_t} h_t\right)\left(\nabla_V Uh_{t-1} + \nabla_V Vx_t\right) + W^T(z-y)\left(\nabla_{A_{t+1}} h_{t+1}\right)\left(I^{[200\times 200]}x_{t+1}\right) =$
$W^T(z-y)\left(\nabla_{A_{t+1}} h_{t+1}\right)U\left(\nabla_{A_t} h_t\right)\left(\nabla_V Uh_{t-1}\right) + W^T(z-y)\left(\nabla_{A_{t+1}} h_{t+1}\right)U\left(\nabla_{A_t} h_t\right)\left(\nabla_V Vx_t\right) + W^T(z-y)\left(\nabla_{A_{t+1}} h_{t+1}\right)\left(I^{[200\times 200]}x_{t+1}\right) =$
$W^T(z-y)\left(\nabla_{A_{t+1}} h_{t+1}\right)U\left(\nabla_{A_t} h_t\right)U\left(\nabla_V h_{t-1}\right) + W^T(z-y)\left(\nabla_{A_{t+1}} h_{t+1}\right)U\left(I^{[200\times 200]}x_t\right)\left(\nabla_{A_t} h_t\right) + W^T(z-y)\left(I^{[200\times 200]}x_{t+1}\right)\left(\nabla_{A_{t+1}} h_{t+1}\right) =$

$\nabla_U L =$
$\left(\nabla_{z_{t+1}} L\right)\left(\nabla_U z_{t+1}\right) =$
$\left(\nabla_{z_{t+1}} L\right)\left(\nabla_{h_{t+1}} z_{t+1}\right)\left(\nabla_U h_{t+1}\right) =$
$\left(\frac{1-y_j}{1-z_j} - \frac{y_j}{z_j}\right)\left(z_j(1-z_j)W_j^T\right)\left(\nabla_U h_{t+1}\right) =$
$(z_j - y_j)\, W_j^T \left(\nabla_U h_{t+1}\right) =$
$W_j^T\,(z_j - y_j)\left(\nabla_U h_{t+1}\right) =$
$W^T(z-y)\left(\nabla_U h_{t+1}\right) =$
$W^T(z-y)\left(\nabla_{A_{t+1}} h_{t+1}\right)\left(\nabla_U A_{t+1}\right) =$
$W^T(z-y)\left(\nabla_{A_{t+1}} h_{t+1}\right)\left(\nabla_U Uh_t\right) =$
$W^T(z-y)\left(\nabla_{A_{t+1}} h_{t+1}\right)\left[U\nabla_U h_t + h_t\right] =$
$W^T(z-y)\left(\nabla_{A_{t+1}} h_{t+1}\right)U\left(\nabla_U h_t\right) + W^T(z-y)\left(\nabla_{A_{t+1}} h_{t+1}\right)h_t =$
$W^T(z-y)\left(\nabla_{A_{t+1}} h_{t+1}\right)U\left(\nabla_{A_t} h_t\right)\left(\nabla_U A_t\right) + W^T(z-y)\left(\nabla_{A_{t+1}} h_{t+1}\right)h_t =$
$W^T(z-y)\left(\nabla_{A_{t+1}} h_{t+1}\right)U\left(\nabla_{A_t} h_t\right)\left(\nabla_U Uh_{t-1}\right) + W^T(z-y)\left(\nabla_{A_{t+1}} h_{t+1}\right)h_t =$
$W^T(z-y)\left(\nabla_{A_{t+1}} h_{t+1}\right)U\left(\nabla_{A_t} h_t\right)\left[U\nabla_U h_{t-1} + h_{t-1}\right] + W^T(z-y)\left(\nabla_{A_{t+1}} h_{t+1}\right)h_t =$
$W^T(z-y)\left(\nabla_{A_{t+1}} h_{t+1}\right)U\left(\nabla_{A_t} h_t\right)U\left(\nabla_U h_{t-1}\right) + W^T(z-y)\left(\nabla_{A_{t+1}} h_{t+1}\right)U\left(\nabla_{A_t} h_t\right)h_{t-1} + W^T(z-y)\left(\nabla_{A_{t+1}} h_{t+1}\right)h_t =$

$\nabla_W L =$
$\left(\nabla_{z_f} L\right)\left(\nabla_W z_f\right)$

$(\nabla_{z_j} L)(\nabla_{W_j} z_j)$
$\left(\frac{1-y_j}{1-z_j} - \frac{y_j}{z_j}\right) z_j(1-z_j)h_f^T$
$(z_j - y_j)\, h_f^T$
$(z-y)h_f^T \qquad [3 \times 200]$

$\nabla_{z_j} L = \frac{1-y_j}{1-z_j} - \frac{y_j}{z_j}$
$\nabla_{W_j} z_j = z_j(1-z_j)h^T \qquad [1 \times 200]$
$\nabla_h z_j = z_j(1-z_j)W_j^T \qquad [201 \times 1]$
$\nabla_A h = (\text{elementwise}) \begin{cases} 1, & h_j > 0 \\ 0, & h_j < 0 \end{cases}$
$\nabla_U h = \text{recursive expansion}$
$\nabla_U h = \text{recursive expansion}$

We can substitute the following expressions:

$\nabla_h z_j = z_j(1-z_j)W_j^T \qquad [201 \times 1]$
$\nabla_{V_j} h_j = \text{sech}^2(V_j x)x^T \qquad [1 \times 785]$

We can also use backpropagation to show:

$\nabla_{W_j} L = \left(\frac{1-y_j}{1-z_j} - \frac{y_j}{z_j}\right)\left(z_j(1-z_j)h^T\right) \qquad [1 \times 201]$
$\nabla_h L = \sum_{j=1}^{26}\left(\frac{1-y_j}{1-z_j} - \frac{y_j}{z_j}\right)\left(z_j(1-z_j)W_j^T\right) \qquad [201 \times 1]$
$\nabla_{V_j} L = (\nabla_h L)_j\, \text{sech}^2(V_j x)x^T \qquad [1 \times 785]$

To enhance calculation efficiency, we can reduce some of these equations into matrices. First, let $\mathcal{Q}$ be the matrix

$$\mathcal{Q} = \begin{bmatrix} z_1(1-z_1)\left(\frac{1-y_1}{1-z_1} - \frac{y_1}{z_1}\right) \\ z_2(1-z_2)\left(\frac{1-y_2}{1-z_2} - \frac{y_2}{z_2}\right) \\ \vdots \\ z_{26}(1-z_{26})\left(\frac{1-y_{26}}{1-z_{26}} - \frac{y_{26}}{z_{26}}\right) \end{bmatrix} = \begin{bmatrix} z_1 - y_1 \\ z_2 - y_2 \\ \vdots \\ z_{26} - y_{26} \end{bmatrix}$$

With this, we can reexpress the above equations.
(1)

$\nabla_{W_j} L = \mathcal{Q}_j h^T \quad [1 \times 201]$
or

$$\nabla_W L = \begin{bmatrix} \mathcal{Q}_1 h^T \\ \mathcal{Q}_2 h^T \\ \vdots \\ \mathcal{Q}_{26} h^T \end{bmatrix} = \mathcal{Q} h^T = \mathcal{Q} \otimes h \qquad [26 \times 201]$$

(2)
$\nabla_h L = \sum_{j=1}^{26} \mathcal{Q}_j W_j^T \quad [201 \times 1]$
or

$$\nabla_h L = \left[\mathcal{Q}_1 W_1^T + \mathcal{Q}_2 W_2^T + ... + \mathcal{Q}_{26} W_{26}^T\right]$$
$$\nabla_h L = \left[W_1^T \mathcal{Q}_1 + W_2^T \mathcal{Q}_2 + ... + W_{26}^T \mathcal{Q}_{26}\right]$$
$$\nabla_h L = \begin{bmatrix} W_1^T & W_2^T & \dots & W_{26}^T \end{bmatrix} \begin{bmatrix} \mathcal{Q}_1 \\ \mathcal{Q}_2 \\ \vdots \\ \mathcal{Q}_{26} \end{bmatrix} = W^T \mathcal{Q} \qquad [201 \times 1]$$

(3)
Additionally, let

$$\mathcal{S} = \begin{bmatrix} \operatorname{sech}^2(V_1 x) \\ \operatorname{sech}^2(V_2 x) \\ \vdots \\ \operatorname{sech}^2(V_{200} x) \end{bmatrix} = \operatorname{sech}^2(V x)$$

Then, $\nabla_{V_j} L = (W^T \mathcal{Q})_j \operatorname{sech}^2(V_j x) x^T \quad [1 \times 785]$
or

$$\nabla_V L = \begin{bmatrix} (W^T \mathcal{Q})_1 \operatorname{sech}^2(V_1 x) x^T \\ (W^T \mathcal{Q})_2 \operatorname{sech}^2(V_2 x) x^T \\ \vdots \\ (W^T \mathcal{Q})_{200} \operatorname{sech}^2(V_{200} x) x^T \end{bmatrix} = \begin{bmatrix} (W^T \mathcal{Q})_1 \operatorname{sech}^2(V_1 x) \\ (W^T \mathcal{Q})_2 \operatorname{sech}^2(V_2 x) \\ \vdots \\ (W^T \mathcal{Q})_{200} \operatorname{sech}^2(V_{200} x) \end{bmatrix} x^T$$

$$\nabla_V L = \left(\begin{bmatrix} (W^T \mathcal{Q})_1 \\ (W^T \mathcal{Q})_2 \\ \vdots \\ (W^T \mathcal{Q})_{200} \end{bmatrix} \circ \begin{bmatrix} \operatorname{sech}^2(V_1 x) \\ \operatorname{sech}^2(V_2 x) \\ \vdots \\ \operatorname{sech}^2(V_{200} x) \end{bmatrix} \right) x^T = (W^T \mathcal{Q}) \circ \mathcal{S} x^T$$

Since we are updating our matrices V and W using stochastic gradient descent, we repeat the following process:

$V, W \leftarrow$ weight matrices initialized randomly from normal distribution with mean $\mu = 0$ and $\sigma^2 = (...)$
while (continue = True or $L(z) > 0$)
  Forward calculation [$h = \tanh(Vx) \quad \rightarrow \quad z = s(Wh) \quad \rightarrow \quad L(z)$]
  Backward calculation to return $\nabla_V L$ and $\nabla_W L$
  $V \leftarrow V - \epsilon \nabla_V L)$
  $W \leftarrow W - \epsilon \nabla_W L$
return V, W

where, using our derived equations from above, the update rules are more specifically

$$V \leftarrow V - \epsilon(W^T \mathcal{Q}) \circ \mathcal{S} x^T$$
$$W \leftarrow W - \epsilon(\mathcal{Q} \otimes h).$$

$h = \mathrm{ReLU}(Uh_{t-1} + Vx_t)$
$\partial h / \partial h_{t-1} = \mathrm{RLD}(Uh_{t-1} + Vx_t) U^T$
$\partial h / \partial V = \mathrm{RLD}(Uh_{t-1} + Vx_t)(\partial(Uh_{t-1})/\partial V + \partial(Vx_t)/\partial V) = \mathrm{RLD}(Uh_{t-1} + Vx_t)(U^T \frac{\partial h_{t-1}}{\partial h_{t-2}}(...)\frac{\partial h_0}{\partial V} + Ix_t)$
$\partial h_0 / \partial V =$
$\partial h / \partial U = \mathrm{RLD}(Uh_{t-1} + Vx_t)(\partial(Uh_{t-1})/\partial U + \partial(Vx_t)/\partial U) = \mathrm{RLD}(Uh_{t-1} + Vx_t)(U^T \frac{\partial h_{t-1}}{\partial h_{t-2}}(...)\frac{\partial h_0}{\partial V} + Ix_t)$

$y = \sin(Ax)$
$\partial y / \partial x = \frac{\partial(Ax)}{\partial x} \cos(Ax) = A^T \cos(Ax)$
$[200 \times 200] = [200 \times 1](...)$