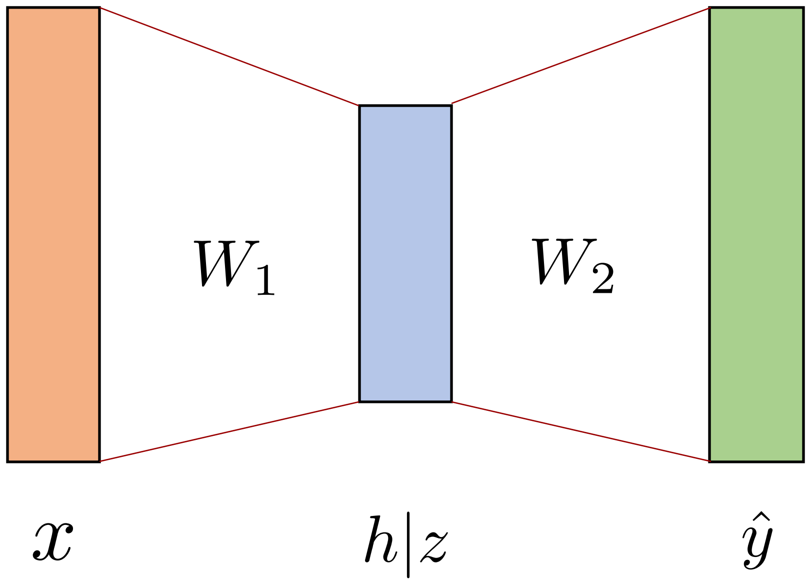
Linear neural network: Single-batch

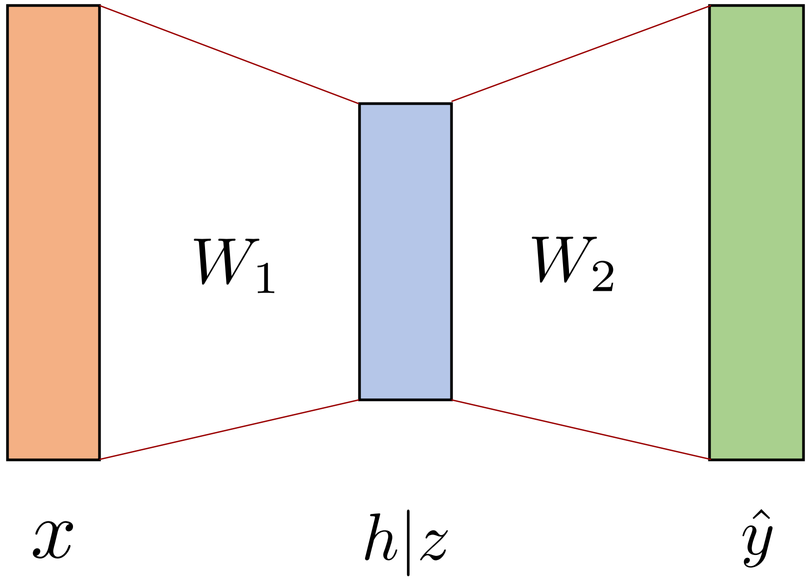


dims: d (p,d) p (q,p) q

Forward	Backward
$h = W_1 x$	$\frac{\partial f}{\partial W_1} = \frac{\partial f}{\partial h} x^T$
$z = \sigma(h)$	$\frac{\partial f}{\partial h} = \frac{\partial f}{\partial z} \odot \nabla \sigma(h)$
$\hat{y} = W_2 z$	$\frac{\partial f}{\partial W_2} = \frac{\partial L}{\partial \hat{y}} z^T, \quad \frac{\partial f}{\partial z} = W_2 \frac{\partial L}{\partial \hat{y}}$
$f = L(\hat{y})$	$\frac{\partial f}{\partial \hat{y}} = \nabla L(\hat{y})$

Store h, z and $\hat{y}$ **Store $\nabla_{W_1} f(W_1)$ and $\nabla_{W_2} f(W_2)$**

Linear neural network: Mini-batch



dims: d (p,d) p (q,p) q

$$h_b = W_1 x_b$$

$$z_b = \sigma(h_b)$$

$$\hat{y}_b = W_2 z_b$$

$$f = \frac{1}{B}\sum_{b=1}^{B} L(\hat{y}_b)$$

Forward

$$\frac{\partial f}{\partial W_1} = \frac{1}{B}\sum_{b=1}^{B} \frac{\partial f}{\partial h_b} x_b^T,$$

$$\frac{\partial f}{\partial h_b} = \frac{\partial f}{\partial z_b} \odot \nabla\sigma(h_b)$$

$$\frac{\partial f}{\partial W_2} = \frac{1}{B}\sum_{b=1}^{B} \frac{\partial L}{\partial \hat{y}_b} z_b^T, \quad \frac{\partial f}{\partial z_b} = W_2 \frac{\partial f}{\partial \hat{y}_b}$$

$$\frac{\partial f}{\partial \hat{y}_b} = \frac{\partial L}{\partial \hat{y}_b}$$

Backward

Store $\{\mathbf{h_b}, \mathbf{z_b}, \hat{y}_b\}_{b=1}^{B}$

Store $\nabla_{W_1} f(W_1)$ and $\nabla_{W_2} f(W_2)$