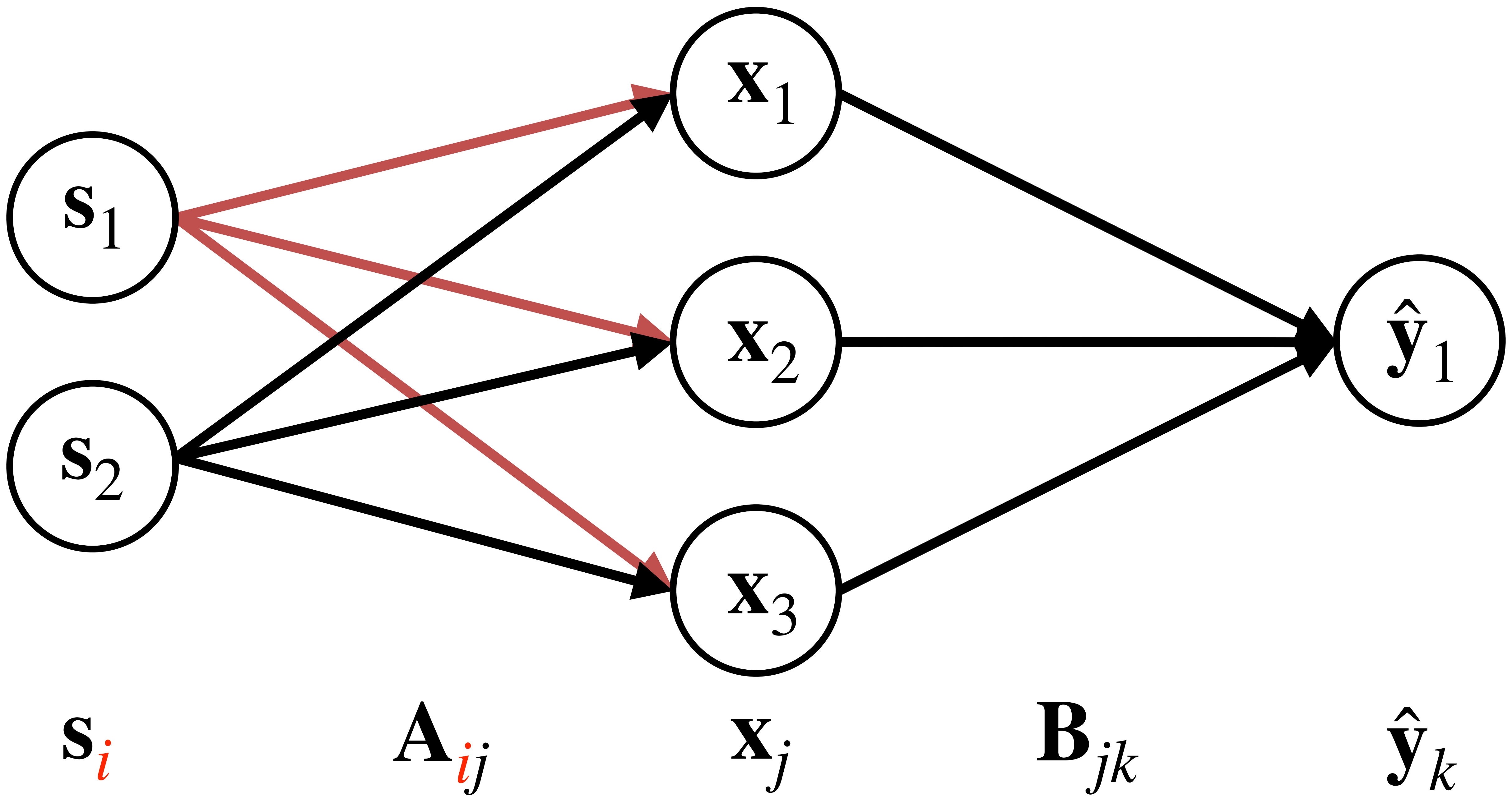
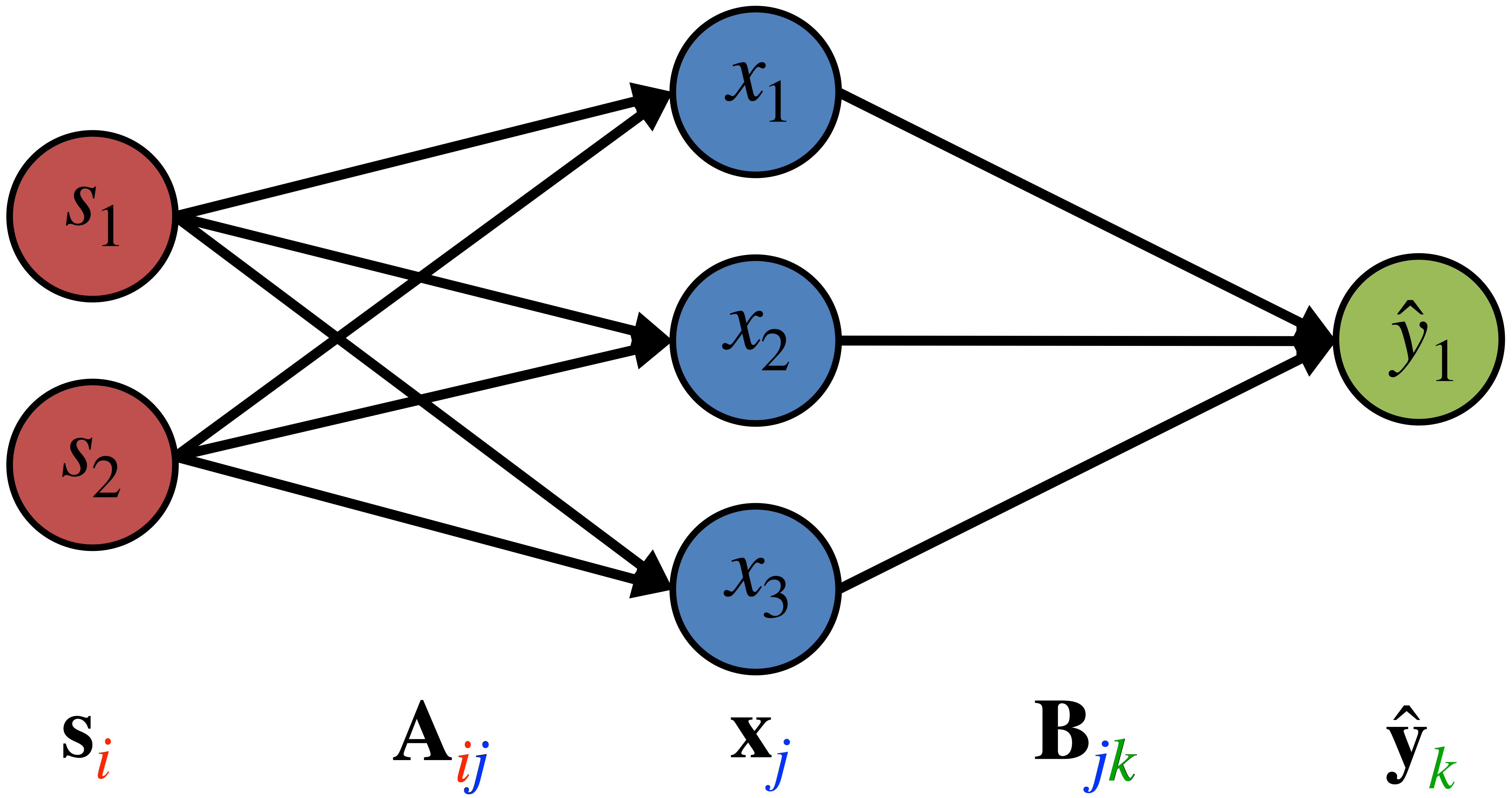


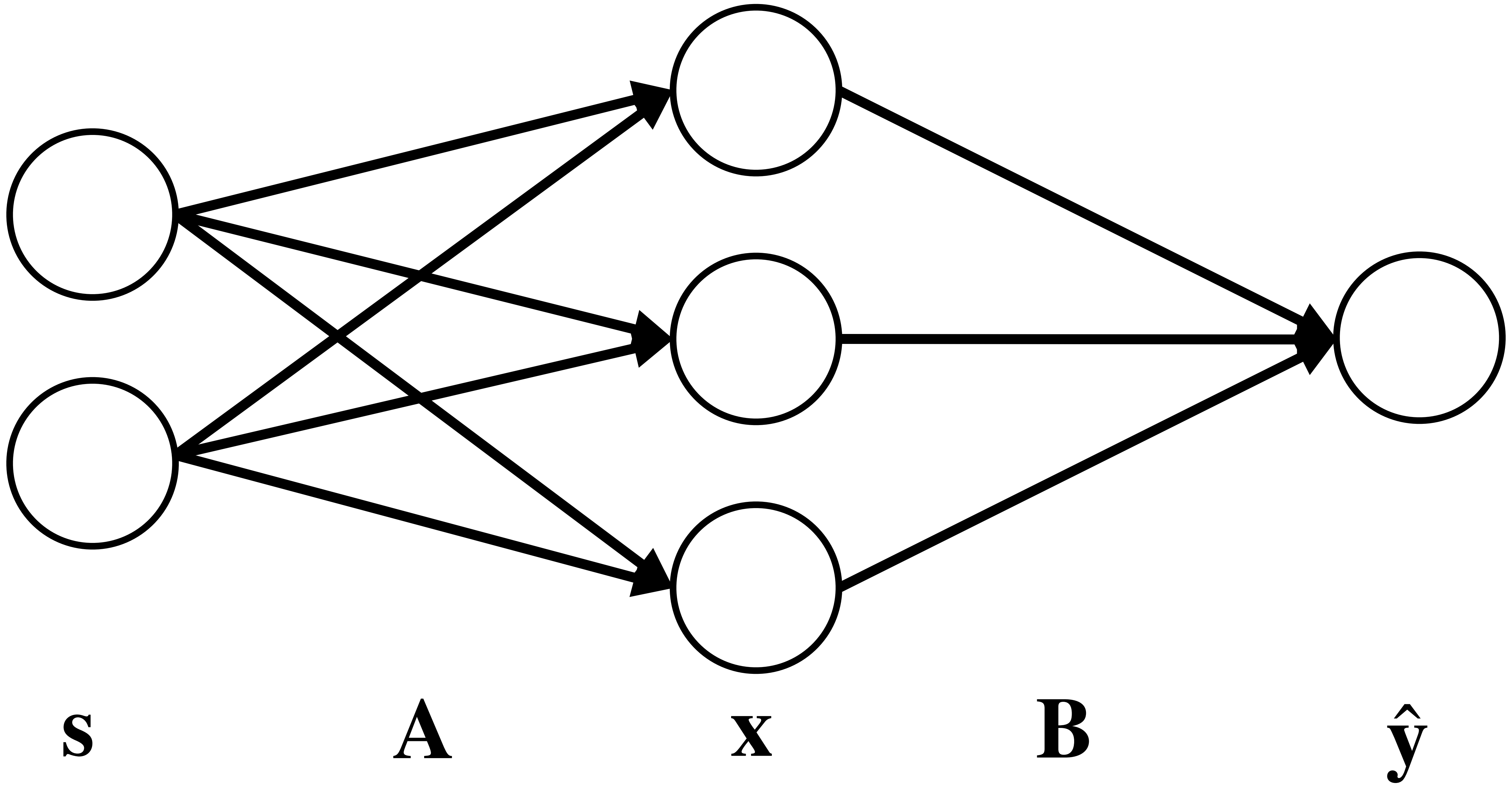

Prediction and Control with Approximation

# Gradient Descent for Training Neural Networks

# Notation

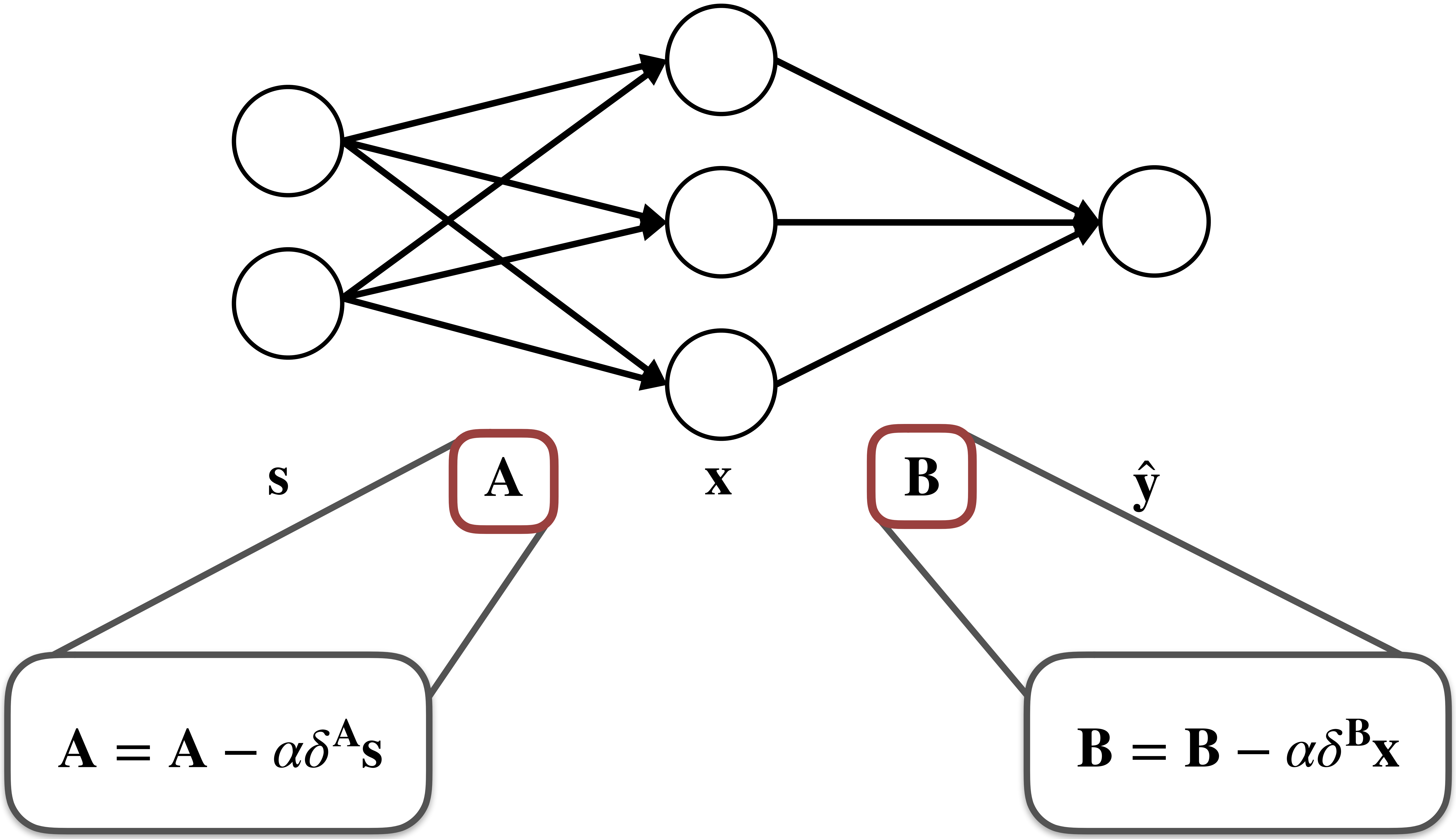

$s_1$
$s_2$
$x_1$
$x_2$
$x_3$
$\hat{y}_1$
$\mathbf{s}_i$
$\mathbf{A}_{ij}$
$\mathbf{x}_j$
$\mathbf{B}_{jk}$
$\hat{\mathbf{y}}_k$

# Notation




$$L(\hat{y}_k, y_k) = (\hat{y}_k - y_k)^2$$

# Goal



$\mathbf{s}$
$\mathbf{A}$
$\mathbf{x}$
$\mathbf{B}$
$\hat{\mathbf{y}}$
$\mathbf{A} = \mathbf{A} - \alpha\delta^{\mathbf{A}}\mathbf{s}$
$\mathbf{B} = \mathbf{B} - \alpha\delta^{\mathbf{B}}\mathbf{x}$

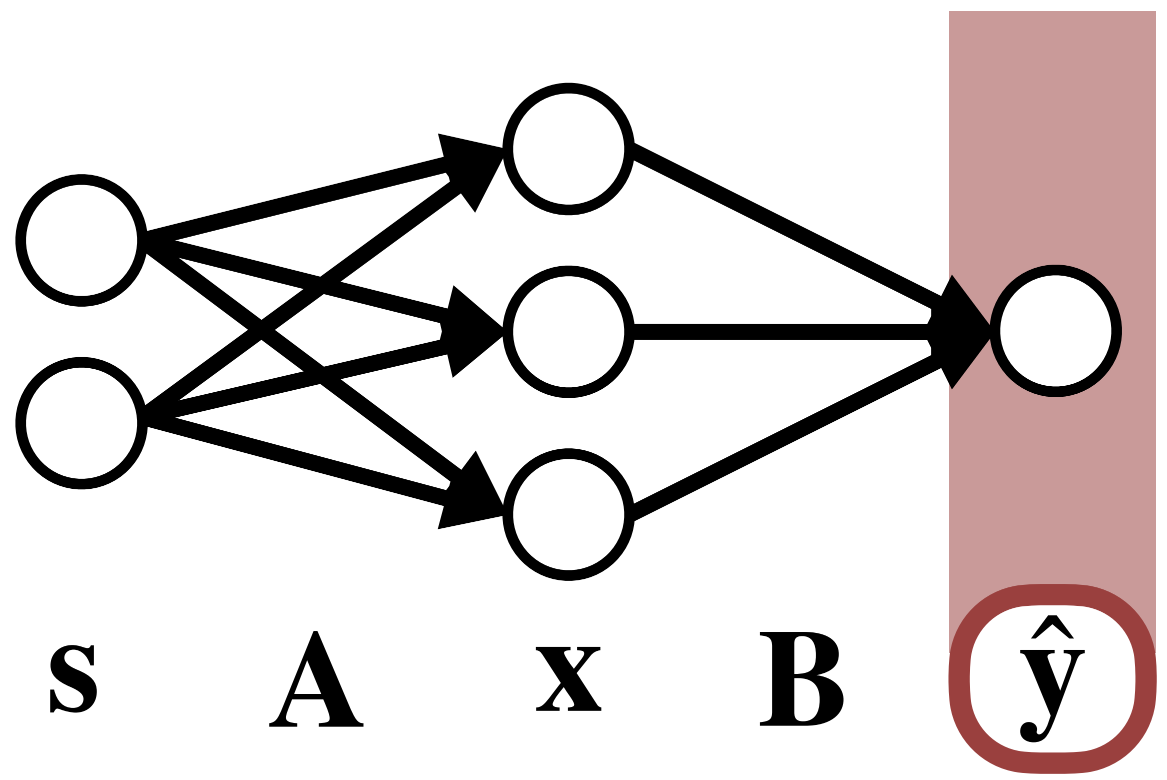
# Deriving the gradient

$$\mathbf{x} \doteq f_{\mathbf{A}}(\mathbf{s}\mathbf{A})$$

$$\hat{\mathbf{y}} \doteq f_{\mathbf{B}}(\mathbf{x}\mathbf{B})$$

$$\frac{\partial L(\hat{\mathbf{y}}_k, \mathbf{y}_k)}{\partial \mathbf{B}_{jk}} = \frac{\partial L(\hat{\mathbf{y}}_k, \mathbf{y}_k)}{\partial \hat{\mathbf{y}}_k} \boxed{\frac{\partial \hat{\mathbf{y}}_k}{\partial \mathbf{B}_{jk}}}$$

$$\theta \doteq \mathbf{x}\mathbf{B}$$



# Deriving the gradient

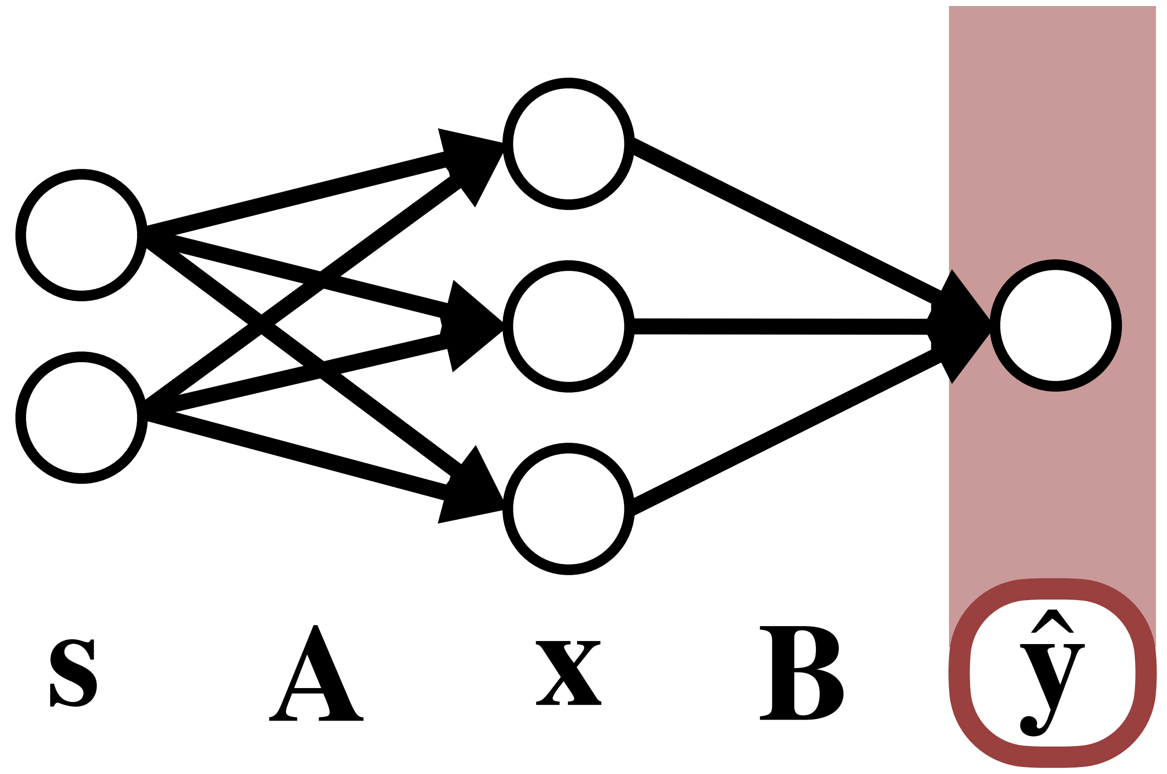
$$\mathbf{x} \doteq f_{\mathbf{A}}(\mathbf{sA})$$

$$\theta \doteq \mathbf{xB}$$

$$\hat{\mathbf{y}} \doteq f_{\mathbf{B}}(\theta)$$

$$\frac{\partial L(\hat{\mathbf{y}}_k, \mathbf{y}_k)}{\partial \mathbf{B}_{jk}} = \frac{\partial L(\hat{\mathbf{y}}_k, \mathbf{y}_k)}{\partial \hat{\mathbf{y}}_k} \frac{\partial \hat{\mathbf{y}}_k}{\partial \mathbf{B}_{jk}}$$

$$= \frac{\partial L(\hat{\mathbf{y}}_k, \mathbf{y}_k)}{\partial \hat{\mathbf{y}}_k} \frac{\partial f_{\mathbf{B}}(\theta_k)}{\partial \theta_k} \frac{\partial \theta_k}{\partial \mathbf{B}_{jk}}$$



# Deriving the gradient

$$\mathbf{x} \doteq f_{\mathbf{A}}(\mathbf{sA})$$

$$\theta \doteq \mathbf{xB}$$

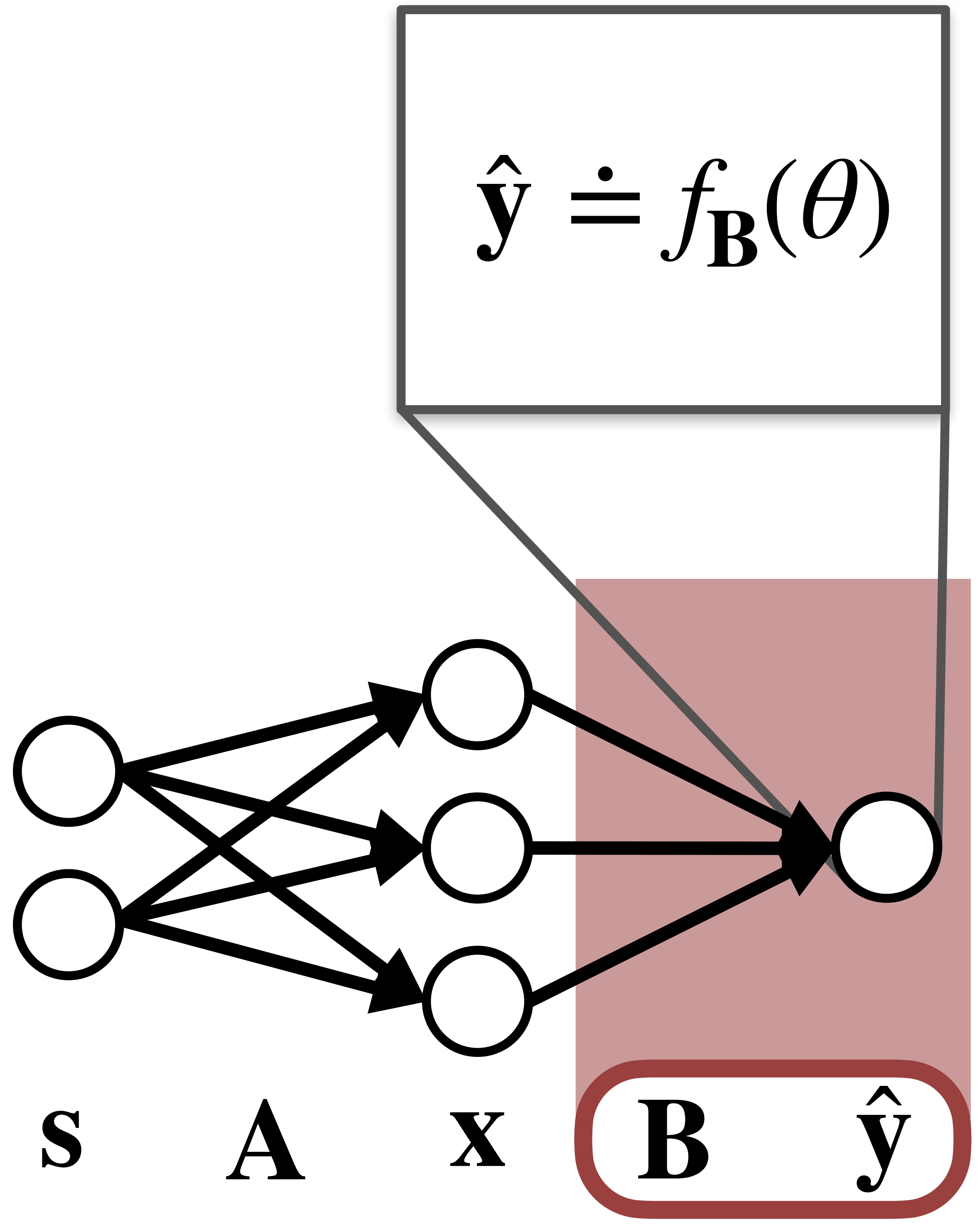
$$\hat{\mathbf{y}} \doteq f_{\mathbf{B}}(\theta)$$

$$\frac{\partial L(\hat{\mathbf{y}}_k, \mathbf{y}_k)}{\partial \mathbf{B}_{jk}} = \frac{\partial L(\hat{\mathbf{y}}_k, \mathbf{y}_k)}{\partial \hat{\mathbf{y}}_k} \frac{\partial \hat{\mathbf{y}}_k}{\partial \mathbf{B}_{jk}}$$

$$= \frac{\partial L(\hat{\mathbf{y}}_k, \mathbf{y}_k)}{\partial \hat{\mathbf{y}}_k} \frac{\partial f_{\mathbf{B}}(\theta_k)}{\partial \theta_k} \frac{\partial \theta_k}{\partial \mathbf{B}_{jk}}$$

$$\frac{\partial \theta_k}{\partial \mathbf{B}_{jk}} = \mathbf{x}_j$$

$$\boxed{= \frac{\partial L(\hat{\mathbf{y}}_k, \mathbf{y}_k)}{\partial \hat{\mathbf{y}}_k} \frac{\partial f_{\mathbf{B}}(\theta_k)}{\partial \theta_k} \mathbf{x}_j}$$

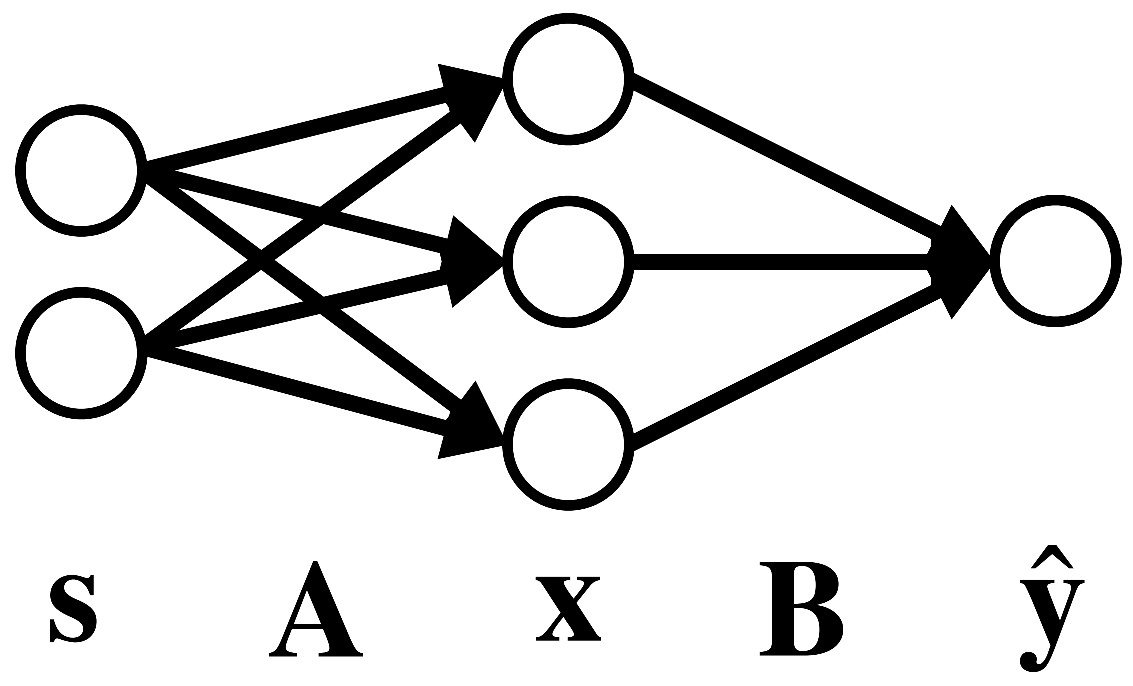


# An example of the gradient

$$\mathbf{x} \doteq f_{\mathbf{A}}(\mathbf{s}\mathbf{A})$$
$$\theta \doteq \mathbf{x}\mathbf{B}$$
$$\hat{\mathbf{y}} \doteq f_{\mathbf{B}}(\theta)$$

| | | |
|---|---|---|
| *Loss:* | $L = \frac{1}{2}(\hat{\mathbf{y}}_k - \mathbf{y}_k)^2$ | $\frac{\partial L(\hat{\mathbf{y}}_k, \mathbf{y}_k)}{\partial \hat{\mathbf{y}}_k} = (\hat{\mathbf{y}}_k - \mathbf{y}_k)$ |
| *Activation:* | $f_{\mathbf{B}}(\theta_k) = \theta_k$ | $\frac{\partial f_{\mathbf{B}}(\theta_k)}{\partial \theta_k} = \frac{\partial \theta_k}{\partial \theta_k} = 1$ |

$$\frac{\partial L(\hat{\mathbf{y}}_k, \mathbf{y}_k)}{\partial \mathbf{B}_{jk}} = \frac{\partial L(\hat{\mathbf{y}}_k, \mathbf{y}_k)}{\partial \hat{\mathbf{y}}_k} \frac{\partial f_{\mathbf{B}}(\theta_k)}{\partial \theta_k} \mathbf{x}_j = (\hat{\mathbf{y}}_k - \mathbf{y}_k)\mathbf{x}_j$$



# Deriving the gradient

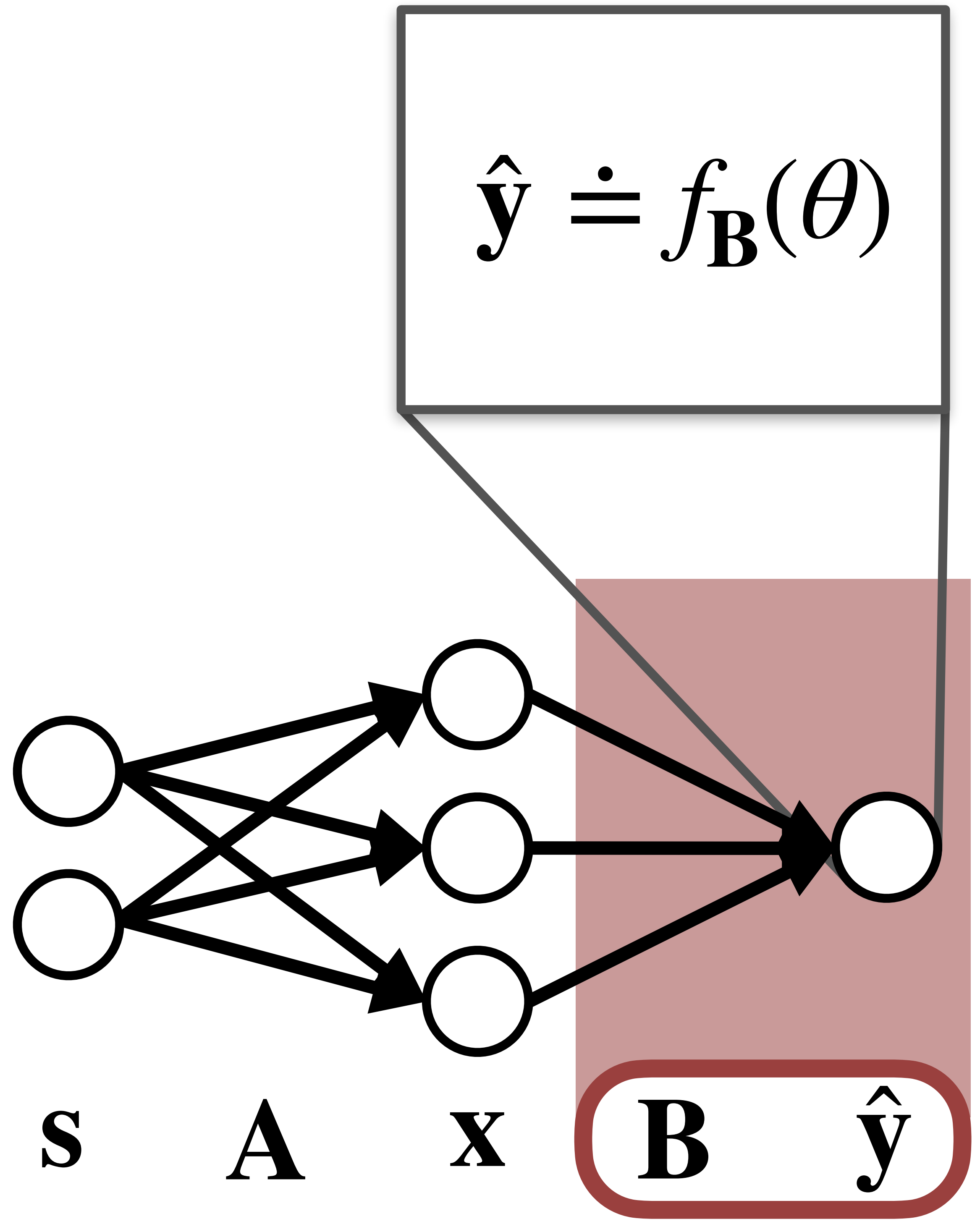
$$\mathbf{x} \doteq f_{\mathbf{A}}(\mathbf{s}\mathbf{A})$$
$$\theta \doteq \mathbf{x}\mathbf{B}$$
$$\hat{\mathbf{y}} \doteq f_{\mathbf{B}}(\theta)$$

$$\frac{\partial L(\hat{\mathbf{y}}_k, \mathbf{y}_k)}{\partial \mathbf{B}_{jk}} = \frac{\partial L(\hat{\mathbf{y}}_k, \mathbf{y}_k)}{\partial \hat{\mathbf{y}}_k} \frac{\partial \hat{\mathbf{y}}_k}{\partial \mathbf{B}_{jk}}$$

$$= \frac{\partial L(\hat{\mathbf{y}}_k, \mathbf{y}_k)}{\partial \hat{\mathbf{y}}_k} \frac{\partial f_{\mathbf{B}}(\theta_k)}{\partial \theta_k} \frac{\partial \theta_k}{\partial \mathbf{B}_{jk}}$$

$$= \boxed{\frac{\partial L(\hat{\mathbf{y}}_k, \mathbf{y}_k)}{\partial \hat{\mathbf{y}}_k} \frac{\partial f_{\mathbf{B}}(\theta_k)}{\partial \theta_k}} \mathbf{x}_j \qquad \delta_k^{\mathbf{B}} = \frac{\partial L(\hat{\mathbf{y}}_k, \mathbf{y}_k)}{\partial \hat{\mathbf{y}}_k} \frac{\partial f_{\mathbf{B}}(\theta_k)}{\partial \theta_k}$$

$$= \delta_k^{\mathbf{B}} \mathbf{x}_j$$



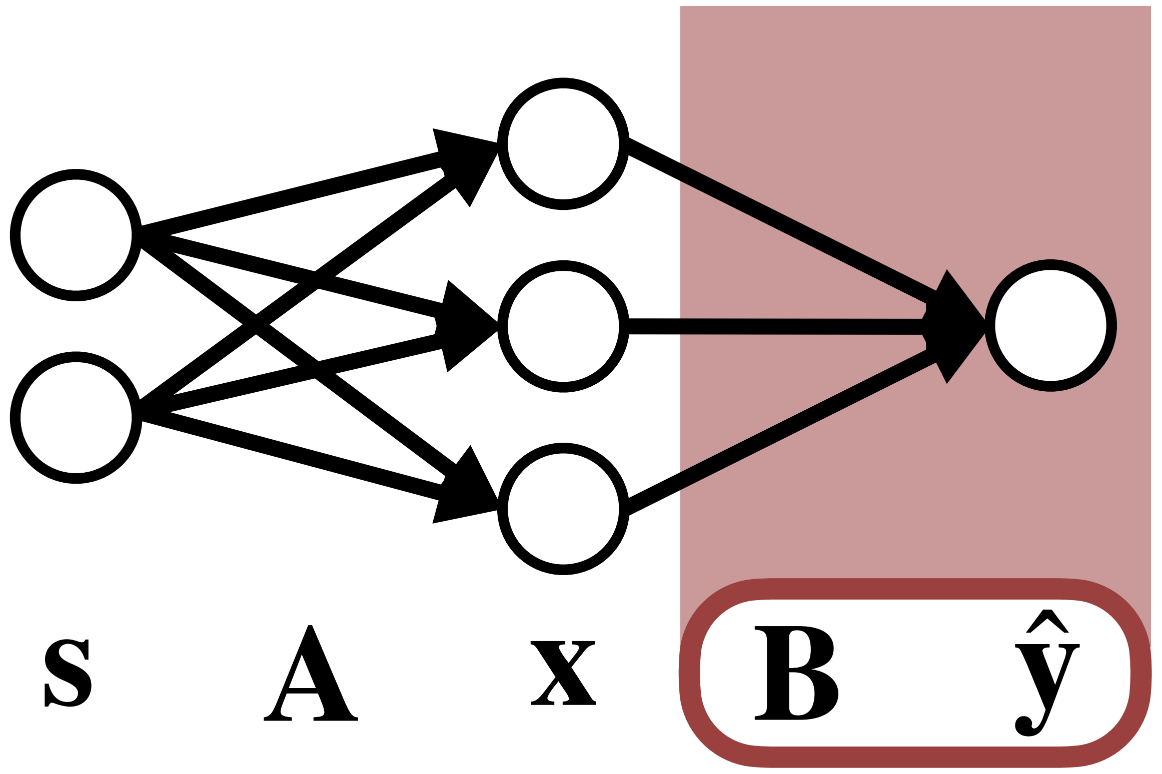
# Deriving the gradient

$$\frac{\partial L(\hat{\mathbf{y}}_k, \mathbf{y}_k)}{\partial \mathbf{B}_{jk}} = \delta_k^{\mathbf{B}} \frac{\partial \boldsymbol{\theta}_k}{\partial \mathbf{B}_{jk}}$$

$$\mathbf{x} \doteq f_{\mathbf{A}}(\mathbf{s}\mathbf{A})$$

$$\theta \doteq \mathbf{x}\mathbf{B}$$

$$\hat{\mathbf{y}} \doteq f_{\mathbf{B}}(\theta)$$



# Deriving the gradient

$$\frac{\partial L(\hat{\mathbf{y}}_k, \mathbf{y}_k)}{\partial \mathbf{A}_{ij}} = \delta_k^{\mathbf{B}} \boxed{\frac{\partial \theta_k}{\partial \mathbf{A}_{ij}}}$$

$$= \delta_k^{\mathbf{B}} \mathbf{B}_{jk} \frac{\partial \mathbf{x}_j}{\partial \mathbf{A}_{ij}}$$

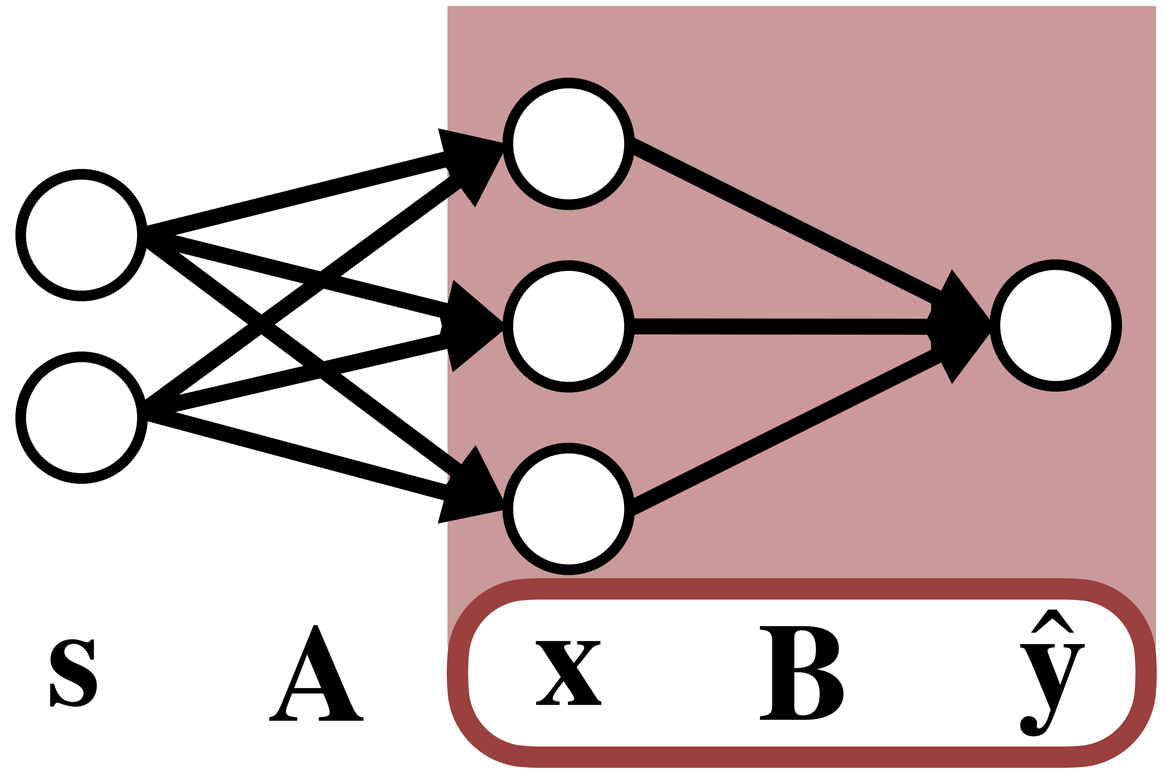
$$\frac{\partial \theta_k}{\partial \mathbf{A}_{ij}} = \mathbf{B}_{jk} \frac{\partial \mathbf{x}_j}{\partial \mathbf{A}_{ij}}$$

$$\psi \doteq \mathbf{s}\mathbf{A}$$

$$\mathbf{x} \doteq f_{\mathbf{A}}(\mathbf{s}\mathbf{A})$$

$$\theta \doteq \mathbf{x}\mathbf{B}$$

$$\hat{\mathbf{y}} \doteq f_{\mathbf{B}}(\theta)$$



# Deriving the gradient

$$\frac{\partial L(\hat{\mathbf{y}}_k, \mathbf{y}_k)}{\partial \mathbf{A}_{ij}} = \delta_k^{\mathbf{B}} \frac{\partial \theta_k}{\partial \mathbf{A}_{ij}}$$

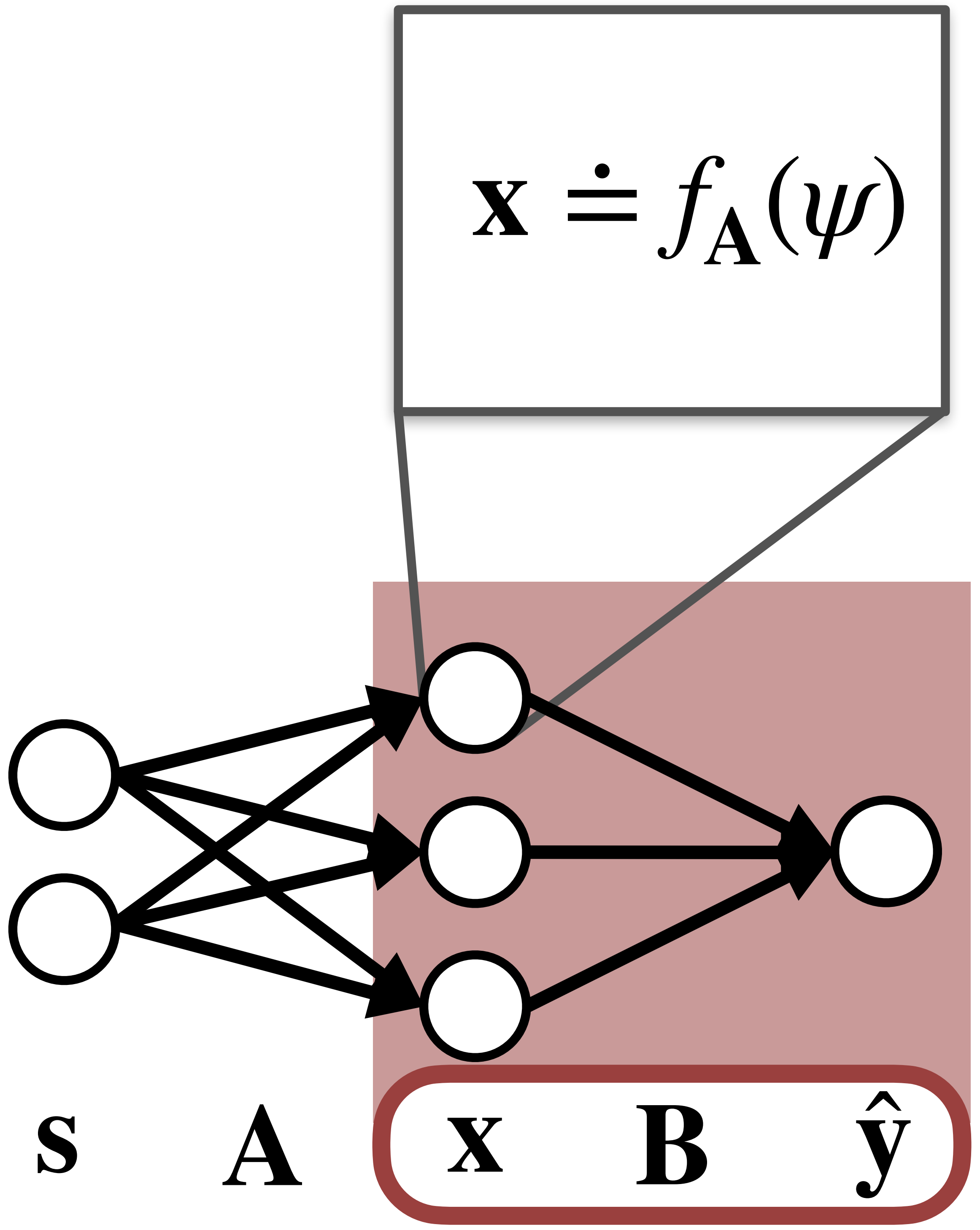
$$= \delta_k^{\mathbf{B}} \mathbf{B}_{jk} \boxed{\frac{\partial \mathbf{x}_j}{\partial \mathbf{A}_{ij}}} \qquad \frac{\partial \mathbf{x}_j}{\partial \mathbf{A}_{ij}} = \frac{\partial f_{\mathbf{A}}(\psi_j)}{\partial \psi_j} \frac{\partial \psi_j}{\partial \mathbf{A}_{ij}}$$

$\psi \doteq \mathbf{s}\mathbf{A}$

$\mathbf{x} \doteq f_{\mathbf{A}}(\psi)$

$\theta \doteq \mathbf{x}\mathbf{B}$

$\hat{\mathbf{y}} \doteq f_{\mathbf{B}}(\theta)$



# Deriving the gradient

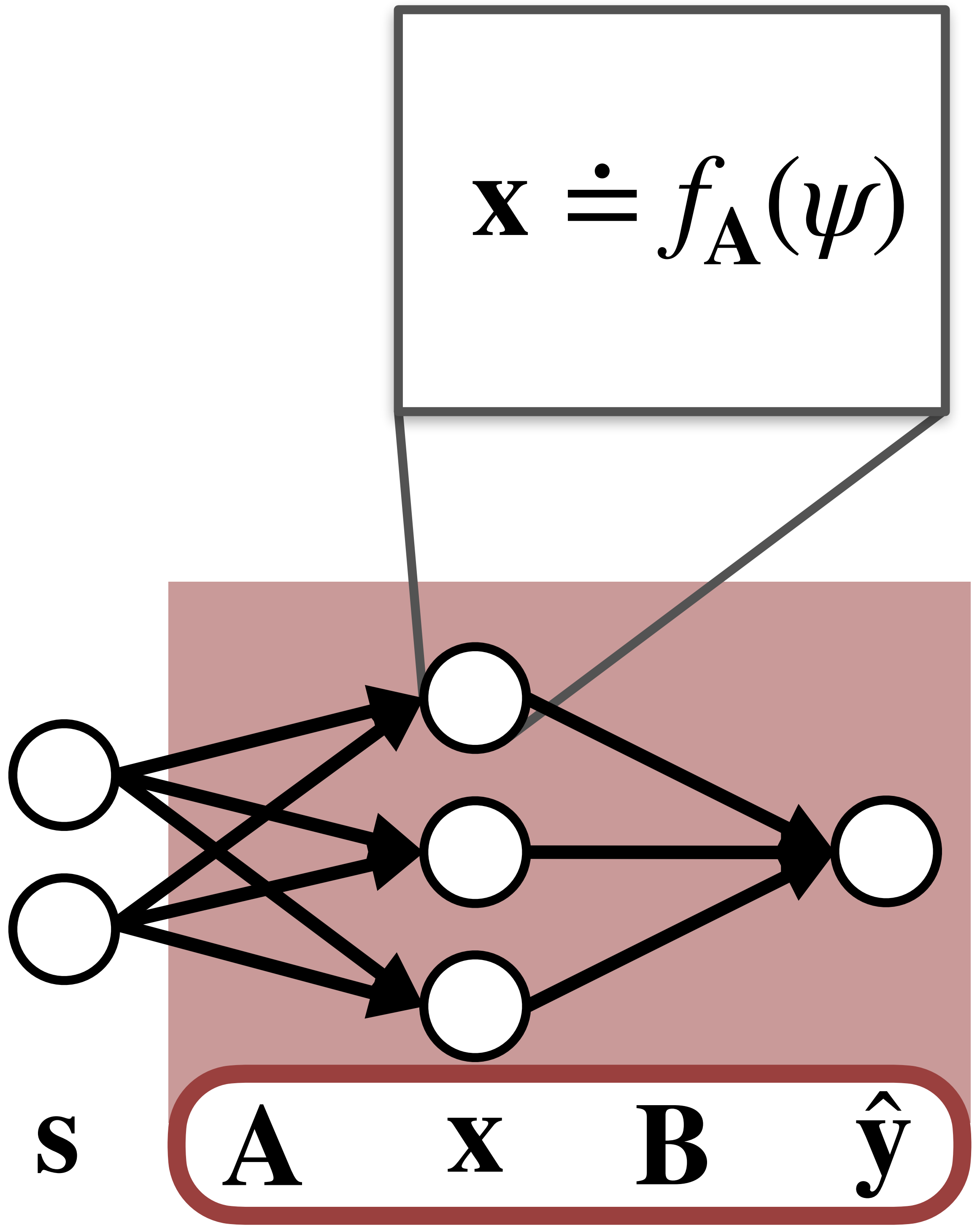
$$\psi \doteq \mathbf{s}\mathbf{A}$$
$$\mathbf{x} \doteq f_{\mathbf{A}}(\psi)$$
$$\theta \doteq \mathbf{x}\mathbf{B}$$
$$\hat{\mathbf{y}} \doteq f_{\mathbf{B}}(\theta)$$

$$\frac{\partial L(\hat{\mathbf{y}}_k, \mathbf{y}_k)}{\partial \mathbf{A}_{ij}} = \delta_k^{\mathbf{B}} \frac{\partial \theta_k}{\partial \mathbf{A}_{ij}}$$

$$= \delta_k^{\mathbf{B}} \mathbf{B}_{jk} \frac{\partial \mathbf{x}_j}{\partial \mathbf{A}_{ij}}$$

$$= \delta_k^{\mathbf{B}} \mathbf{B}_{jk} \frac{\partial f_{\mathbf{A}}(\psi_j)}{\partial \psi_j} \frac{\partial \psi_j}{\partial \mathbf{A}_{ij}} \qquad \frac{\partial \psi_j}{\partial \mathbf{A}_{ij}} = \mathbf{s}_i$$

$$= \delta_k^{\mathbf{B}} \mathbf{B}_{jk} \frac{\partial f_{\mathbf{A}}(\psi_j)}{\partial \psi_j} \mathbf{s}_i$$

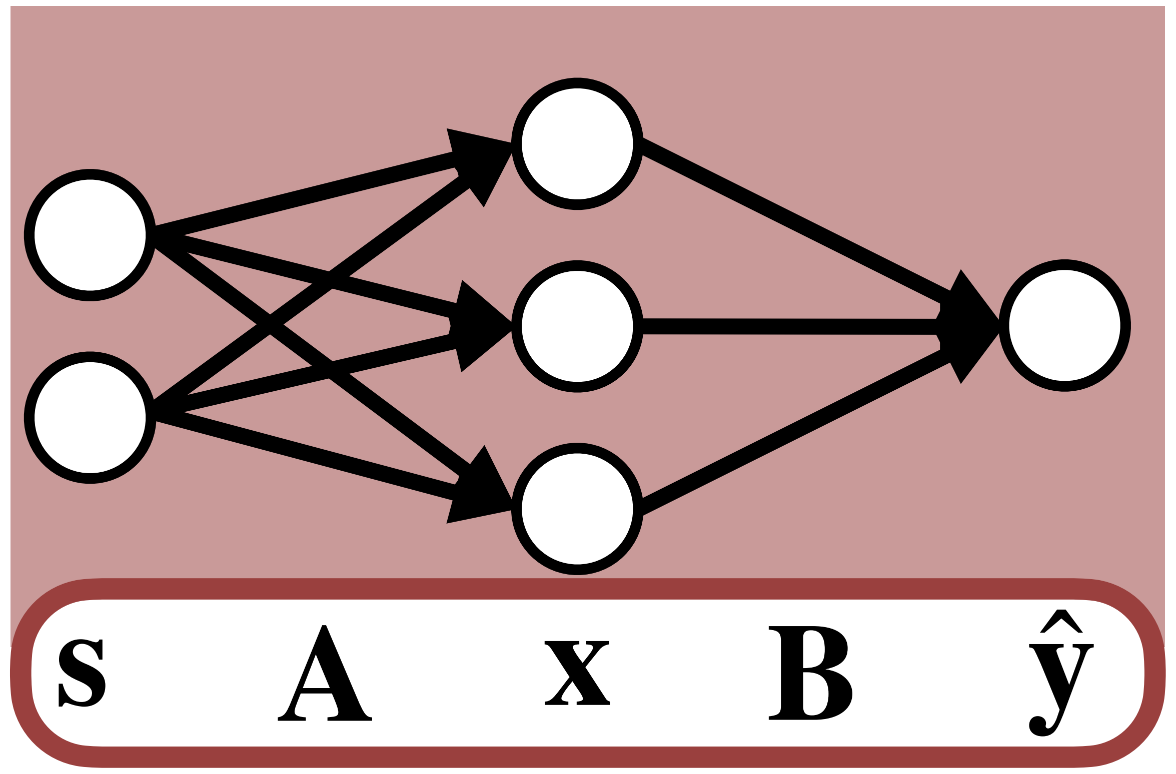


# Deriving the gradient

$$\psi \doteq \mathbf{s}\mathbf{A}$$
$$\mathbf{x} \doteq f_{\mathbf{A}}(\psi)$$
$$\theta \doteq \mathbf{x}\mathbf{B}$$
$$\hat{\mathbf{y}} \doteq f_{\mathbf{B}}(\theta)$$

$$\frac{\partial L(\hat{\mathbf{y}}_k, \mathbf{y}_k)}{\partial \mathbf{A}_{ij}} = \delta_k^{\mathbf{B}} \mathbf{B}_{jk} \frac{\partial f_{\mathbf{A}}(\psi_j)}{\partial \psi_j} \mathbf{s}_i \qquad \delta_j^{\mathbf{A}} = \left(\mathbf{B}_{jk} \delta_k^{\mathbf{B}}\right) \frac{\partial f_{\mathbf{A}}(\psi_j)}{\partial \psi_j}$$

$$= \delta_j^{\mathbf{A}} \mathbf{s}_i$$

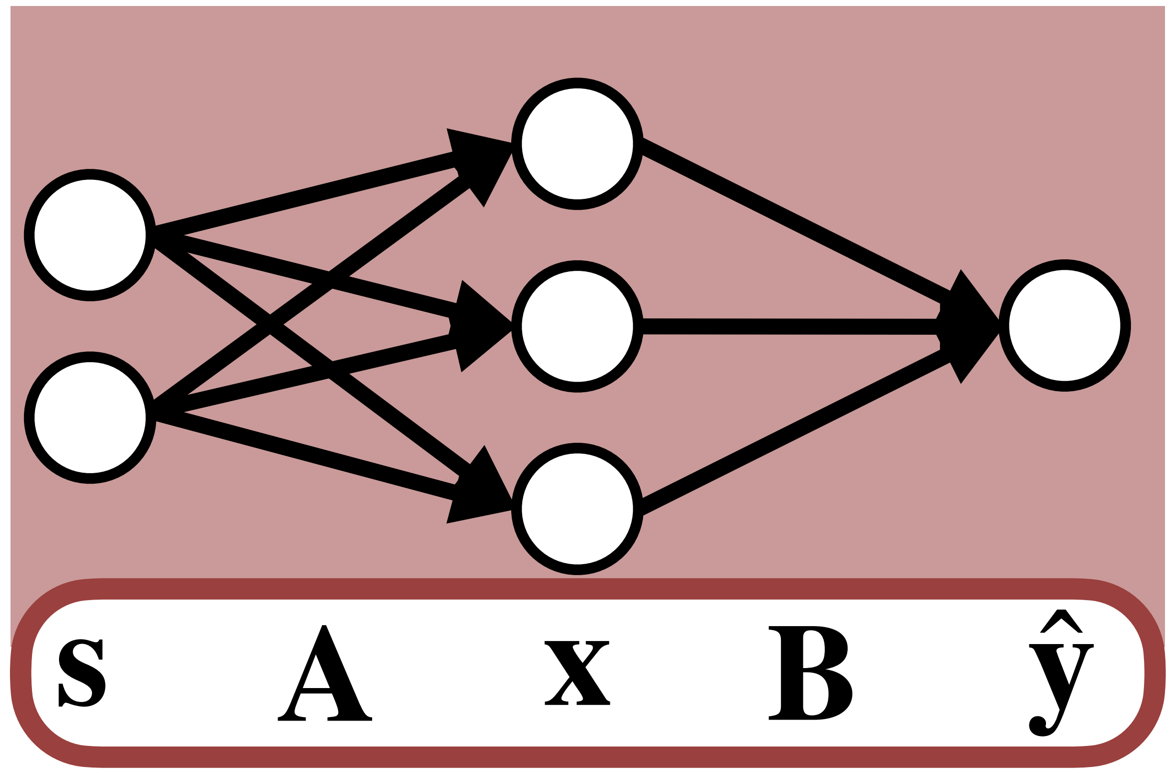


# Deriving the gradient

$$\frac{\partial L(\hat{\mathbf{y}}_k, \mathbf{y}_k)}{\partial \mathbf{A}_{ij}} = \delta_j^{\mathbf{A}} \mathbf{s}_i$$

$$\frac{\partial L(\hat{\mathbf{y}}_k, \mathbf{y}_k)}{\partial \mathbf{B}_{jk}} = \delta_k^{\mathbf{B}} \mathbf{x}_j$$

$$\psi \doteq \mathbf{s}\mathbf{A}$$
$$\mathbf{x} \doteq f_{\mathbf{A}}(\psi)$$
$$\theta \doteq \mathbf{x}\mathbf{B}$$
$$\hat{\mathbf{y}} \doteq f_{\mathbf{B}}(\theta)$$



# The backprop algorithm

**for each** $(s, y)$ **in** $D$**:**

$$\delta_k^{\mathbf{B}} = \frac{\partial L(\hat{\mathbf{y}}_k, \mathbf{y}_k)}{\partial \hat{\mathbf{y}}_k} \frac{\partial f_{\mathbf{B}}(\theta_k)}{\partial \theta_k}$$

$$\nabla_{\mathbf{B}}^{jk} = \delta_k^{\mathbf{B}} \mathbf{x}_j$$

$$\mathbf{B} = \mathbf{B} - \alpha_{\mathbf{B}} \nabla_{\mathbf{B}}$$

$$\delta_j^{\mathbf{A}} = \left(\mathbf{B}_{jk} \delta_k^{\mathbf{B}}\right) \frac{\partial f_{\mathbf{A}}(\psi_j)}{\partial \psi_j}$$

$$\nabla_{\mathbf{A}}^{ij} = \delta_j^{\mathbf{A}} \mathbf{s}_i$$

$$\mathbf{A} = \mathbf{A} - \alpha_{\mathbf{A}} \nabla_{\mathbf{A}}$$

# The backprop algorithm

**for each** $(s, y)$ **in** $D$**:**

$$\delta_k^{\mathbf{B}} = (\hat{\mathbf{y}}_k - \mathbf{y}_k)\mathbf{x}_j \qquad \delta_k^{\mathbf{B}} = \frac{\partial L(\hat{\mathbf{y}}_k, \mathbf{y}_k)}{\partial \hat{\mathbf{y}}_k}\frac{\partial f_{\mathbf{B}}(\theta_k)}{\partial \theta_k}$$

$$\nabla_{\mathbf{B}}^{jk} = \delta_k^{\mathbf{B}}\mathbf{x}_j$$

$$\mathbf{B} = \mathbf{B} - \alpha_{\mathbf{B}} \nabla_{\mathbf{B}}$$

$$u = \begin{cases} \psi & \psi > 0 \\ 0 & otherwise \end{cases} \qquad u = \frac{\partial f_{\mathbf{A}}(\psi_j)}{\partial \psi_j}$$

$$\delta_j^{\mathbf{A}} = \left(\mathbf{B}_{jk}\delta_k^{\mathbf{B}}\right)u \qquad \delta_j^{\mathbf{A}} = \left(\mathbf{B}_{jk}\delta_k^{\mathbf{B}}\right)\frac{\partial f_{\mathbf{A}}(\psi_j)}{\partial \psi_j}$$

$$\nabla_{\mathbf{A}}^{ij} = \delta_j^{\mathbf{A}}\mathbf{s}_i$$

$$\mathbf{A} = \mathbf{A} - \alpha_{\mathbf{A}} \nabla_{\mathbf{A}}$$

# Summary

- The gradient can be used to update the parameters of a neural network with stochastic gradient descent
- Backprop can save computation by computing gradients starting at the output of the network.