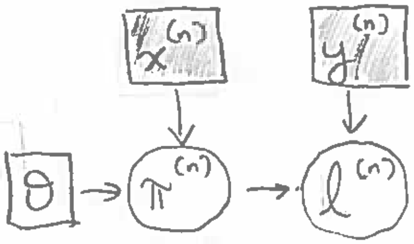
Training Feature Discovery Networks

① So, given the causal model



$$\left.\begin{aligned} \ell^{(n)} &= L^{(n)}(\pi^{(n)}, y^{(n)}) \\ \pi^{(n)} &= f(x^{(n)}, \theta) \end{aligned}\right\}$$

we want to compute $\frac{\partial L^{(n)}}{\partial \theta}$.

By the Chain Rule (since $\pi^{(n)}$ separates θ from $L^{(n)}$):

$$\frac{\partial L^{(n)}}{\partial \theta} = \underbrace{\frac{\partial}{\partial \pi^{(n)}} \ell^{(n)}} \cdot \frac{\partial}{\partial \theta} \pi^{(n)}$$

---

② This part should be straightforward to compute (assuming we chose some reasonable loss function).

e.g.
$$\frac{\partial}{\partial \pi^{(n)}} \ell^{(n)} = \frac{\partial}{\partial \pi^{(n)}} (y^{(n)} - \pi^{(n)})^2$$

$$= -2(y^{(n)} - \pi^{(n)}) \qquad \left(\text{for } \ell^{(n)} = L_{lin}^{(n)}(\pi^{(n)}, y^{(n)})\right)$$

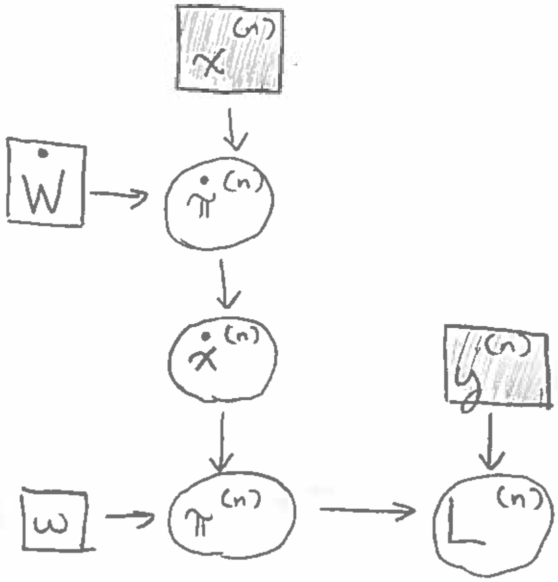
$$\frac{\partial}{\partial \pi^{(n)}} \ell^{(n)} = \frac{\partial}{\partial \pi^{(n)}} (1 - y^{(n)})\pi^{(n)} + \log(1 + e^{-\pi^{(n)}})$$

$$= (1 - y^{(n)}) - \frac{e^{\pi^{(n)}}}{1 + e^{-\pi^{(n)}}} \qquad \left(\text{for } \ell^{(n)} = L_{logistic}^{(n)}(\pi^{(n)}, y^{(n)})\right)$$

Training Feature Discovery Networks

3) The challenge lies in computing $\frac{\partial}{\partial \theta} \pi^{(n)} = \frac{\partial}{\partial \theta} f(x^{(n)}, \theta)$

Let's suppose it's our "feature discovery" network:

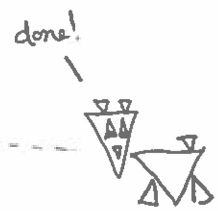


4) Recall that $\theta = \{w_i \in w\} \cup \{\dot{w}_{ij} \in \dot{W}\}$,

so to compute $\frac{\partial}{\partial \theta} \pi^{(n)}$, we need to compute:

(a) for each $w_i \in w$:

$$\frac{\partial}{\partial w_i} \pi^{(n)} = \frac{\partial}{\partial w_i} w^T \dot{x}^{(n)} = \dot{x}_i^{(n)}$$

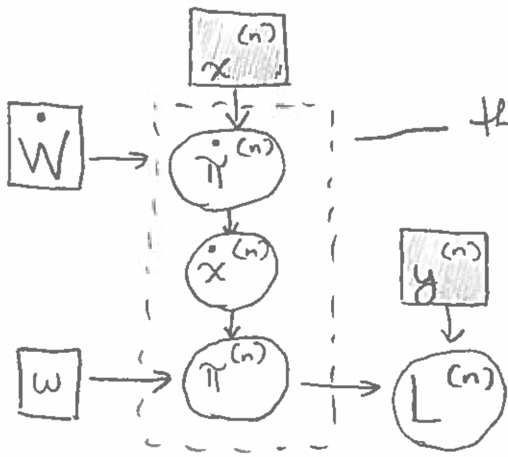


(b) for each $\dot{w}_{ij} \in \dot{W}$:

$$\frac{\partial}{\partial \dot{w}_{ij}} \pi^{(n)}$$

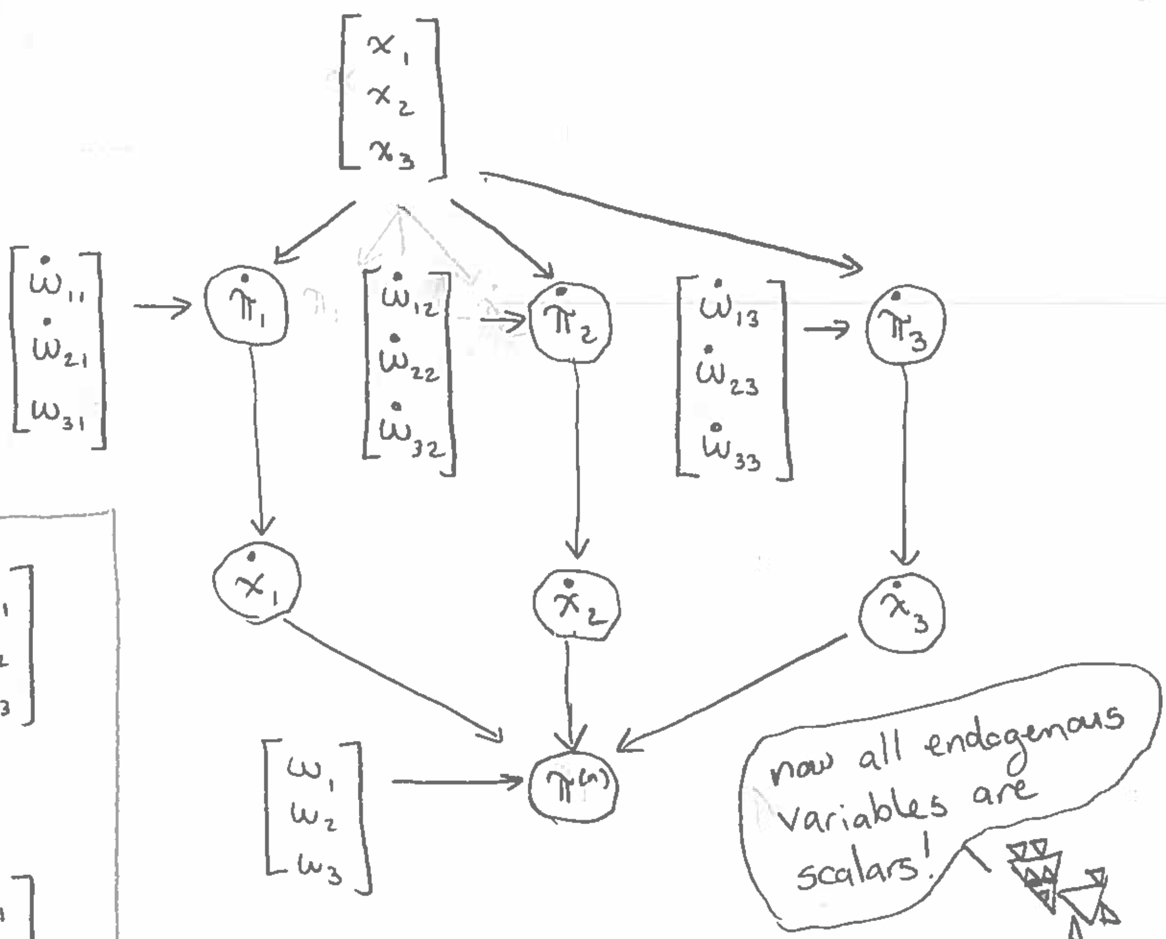
Training Feature Discovery Networks

⑤ Sounds like a job for the Chain Rule of Partial Derivatives, but there's one issue:



these are vectors, and we've only shown how to use the Chain Rule on a causal diagram when these are scalars.

⑥ No problem though, let's just explicitly represent the components of these vectors and matrices:



$$\dot{\pi}_j \leftarrow \begin{bmatrix} \dot{w}_{1j} \\ \dot{w}_{2j} \\ \dot{w}_{3j} \end{bmatrix}^T \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

$$\dot{x}_j \leftarrow a(\dot{\pi}_j)$$

$$\pi^{(n)} \leftarrow \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix}^T \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix}$$

TRAINING FEATURE DISCOVERY NETWORKS

7) Now we can compute $\frac{\partial \pi^{(n)}}{\partial \dot{\omega}_{ij}}$ by repeated application of the Chain Rule:

$$\frac{\partial \pi^{(n)}}{\partial \dot{\omega}_{ij}} = \frac{\partial \pi^{(n)}}{\partial \dot{x}_j} \cdot \frac{\partial \dot{x}_j}{\partial \dot{\omega}_{ij}}$$

[$\dot{x}_j$ separates $\pi^{(n)}$ from $\dot{\omega}_{ij}$, so Chain Rule applies]

$$= \frac{\partial \pi^{(n)}}{\partial \dot{x}_j} \cdot \frac{\partial \dot{x}_j}{\partial \dot{\pi}_j} \cdot \frac{\partial \dot{\pi}_j}{\partial \dot{\omega}_{ij}}$$

[$\dot{\pi}_j$ separates $\dot{x}_j$ from $\dot{\omega}_{ij}$, so Chain Rule applies]

$$= \left(\frac{\partial}{\partial \dot{x}_j} \omega^T \dot{x}^{(n)}\right)\left(\frac{\partial}{\partial \dot{\pi}_j} a(\dot{\pi}_j)\right)\left(\frac{\partial}{\partial \dot{\omega}_{ij}} \begin{bmatrix} \dot{\omega}_{1j} \\ \dot{\omega}_{2j} \\ \dot{\omega}_{3j} \end{bmatrix}^T \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}\right)$$

$$= \omega_j \cdot \underbrace{a'(\dot{\pi}_j)}_{\text{this is just the standard derivative of the ReLU function.}} \cdot x_i$$

(8) Putting this together with the result from (1):

$$\frac{\partial L^{(n)}}{\partial \dot{w}_{ij}} = \left(\frac{\partial \ell^{(n)}}{\partial \pi^{(n)}}\right) w_j\, a'(\dot{\pi}_j)\, x_i$$

and
$$\frac{\partial L^{(n)}}{\partial w_i} = \left(\frac{\partial \ell^{(n)}}{\partial \pi^{(n)}}\right) \dot{x}_i^{(n)}$$

So we can compute the partial derivative of the loss with respect to any of our weights. Thus we can use gradient descent to compute the weights that minimize our loss.