Bellman Equation as Dot Products
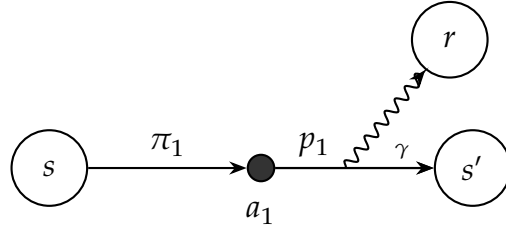

$$g = \vec{\gamma} \cdot \vec{r}$$

$$\vec{o} = \vec{r} + \gamma \vec{v}'$$

$$q = \vec{p} \cdot \vec{o} \qquad = E[\,g \mid s,a\,]$$

$$v = \vec{\pi} \cdot \vec{q} \qquad = E[\,g \mid s\,]$$

$$\vec{\gamma} = \begin{bmatrix} \gamma^0 \\ \gamma^1 \\ \vdots \end{bmatrix}$$

$$\vec{r} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \end{bmatrix} = \begin{bmatrix} r(s,a,s_1') \\ r(s,a,s_2') \\ \vdots \end{bmatrix}$$

$$\vec{v} = \begin{bmatrix} v_0 \\ v_1 \\ \vdots \end{bmatrix} = \begin{bmatrix} v(s_0) \\ v(s_1) \\ \vdots \end{bmatrix} = \begin{bmatrix} \vec{\pi_1} \cdot \vec{q_1} \\ \vec{\pi_2} \cdot \vec{q_2} \\ \vdots \end{bmatrix}$$

$$\vec{v}' = \begin{bmatrix} v_0' \\ v_1' \\ \vdots \end{bmatrix} = \begin{bmatrix} v(s_0') \\ v(s_1') \\ \vdots \end{bmatrix}$$
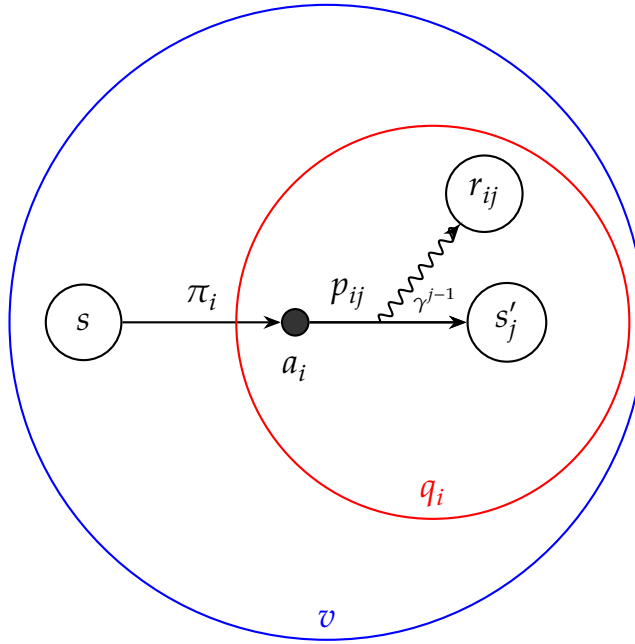
$$\vec{o} = \begin{bmatrix} o_1 \\ o_2 \\ \vdots \end{bmatrix} = \begin{bmatrix} r_1 + \gamma v_1' \\ r_2 + \gamma v_2' \\ \vdots \end{bmatrix}$$

$$\vec{p} = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \end{bmatrix} = \begin{bmatrix} p(s_1' \mid a,s) \\ p(s_2' \mid a,s) \\ \vdots \end{bmatrix}$$

$$\vec{q} = \begin{bmatrix} q_1 \\ q_2 \\ \vdots \end{bmatrix} = \begin{bmatrix} \vec{p_1} \cdot \vec{o_1} \\ \vec{p_2} \cdot \vec{o_2} \\ \vdots \end{bmatrix}$$

$$\vec{\pi} = \begin{bmatrix} \pi_1 \\ \pi_2 \\ \vdots \end{bmatrix} = \begin{bmatrix} p(a_1 \mid s) \\ p(a_2 \mid s) \\ \vdots \end{bmatrix}$$

Einstein Notation

(implicit summation over repeated indices)

$i = $ **actions**

$j = $ **next states**

$$g = \gamma^{j-1} r_j \qquad \text{(discounted sum over trajectory steps)}$$

$$o_{ij} = r_{ij} + \gamma v'_j \qquad \text{(elementwise — no summation)}$$

$$q_i = p_{ij}\, o_{ij} \qquad \text{(sum over next states } j\text{)}$$

$$v = \pi_i\, q_i \qquad \text{(sum over actions } i\text{)}$$

$$v = \pi_i\, p_{ij}\left(r_{ij} + \gamma v'_j\right) \qquad \text{(full Bellman equation — sum over } i \text{ and } j\text{)}$$

GLOSSARY:

r = reward *(rewarded **after** action a is taken)*

γ = discounting factor

$g = g_0 = r_1 + \gamma r_2 + \gamma^2 r_3 + \cdots =$ return *(reward now plus discounted future reward)*

$o = r + \gamma v' =$ outcome *($v = E[g|s] = E[o|s] \leftarrow$ Recursion: Bellman's insight!)*

$s = s_0 =$ current state

$s' = s_1 =$ next state

$v = v(s) = v(s_0)$

$v' = v(s') = v(s_1)$

$a =$ action

$p =$ transition probability, dynamics

$v =$ state value function

$q =$ state-action value function

$\pi =$ policy