① $Q(s,a) = Q(s,a) + \alpha\left(r + \gamma \max_a Q(s',a) - Q(s,a)\right)$

TD error

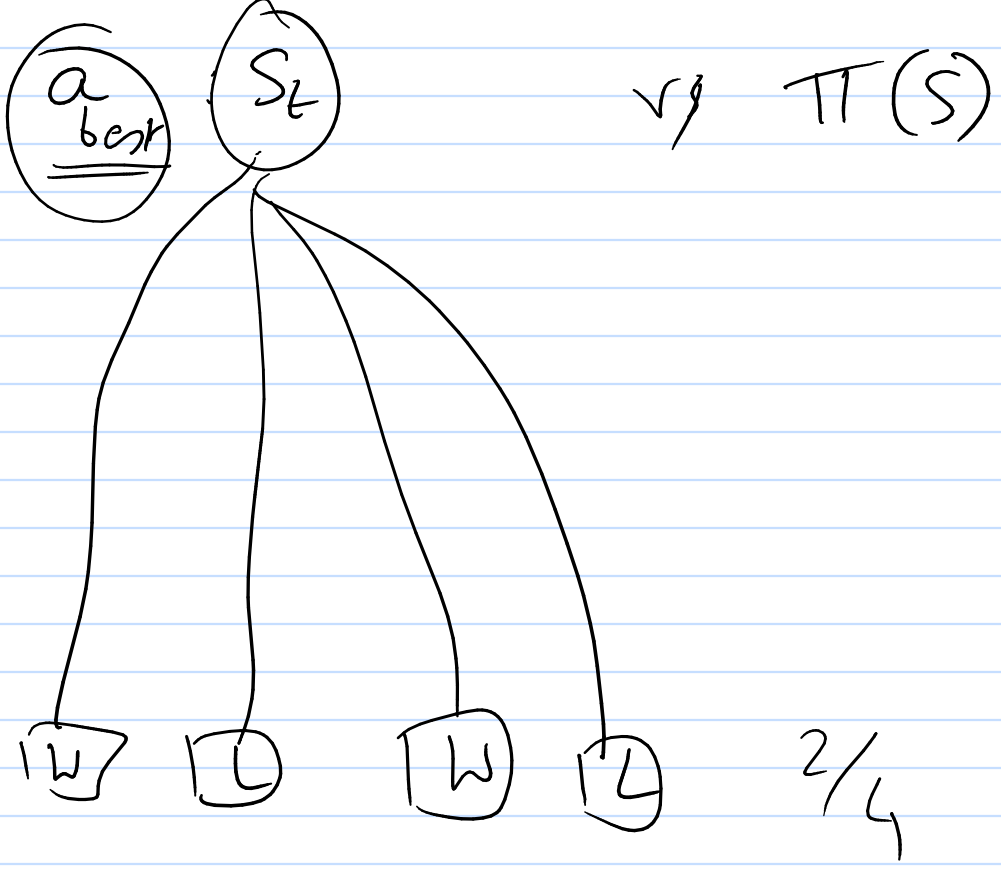
② batch

$\underline{w}^{t+1} = \underline{w}^{t} - \alpha\left(\text{TD error}\right) \nabla_w Q(s,a;w)$

$\frac{d}{dw}\left(r + \gamma \max_a Q(s',a,w^-) - Q(s,a;w)\right)^2$

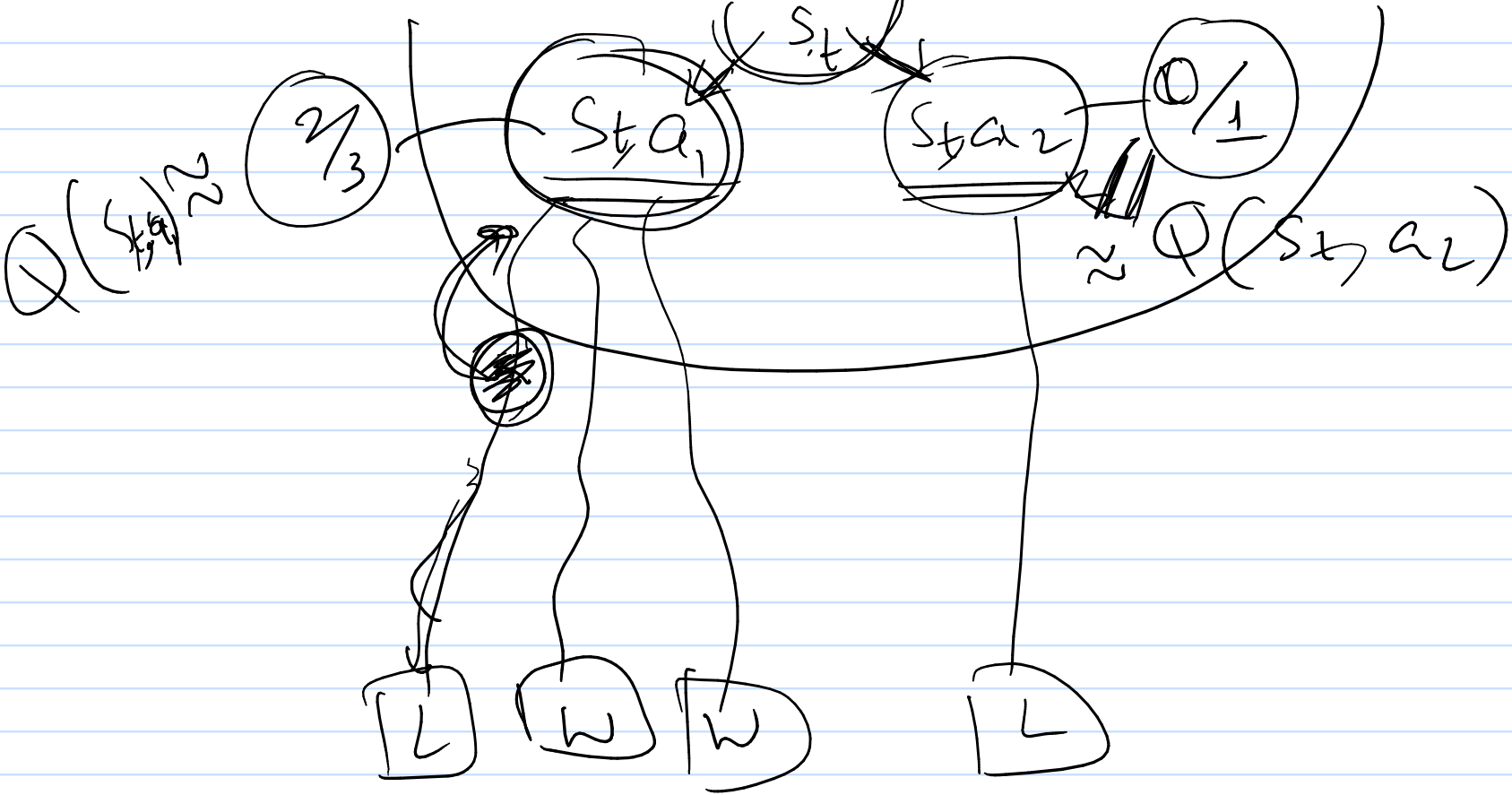
w^-

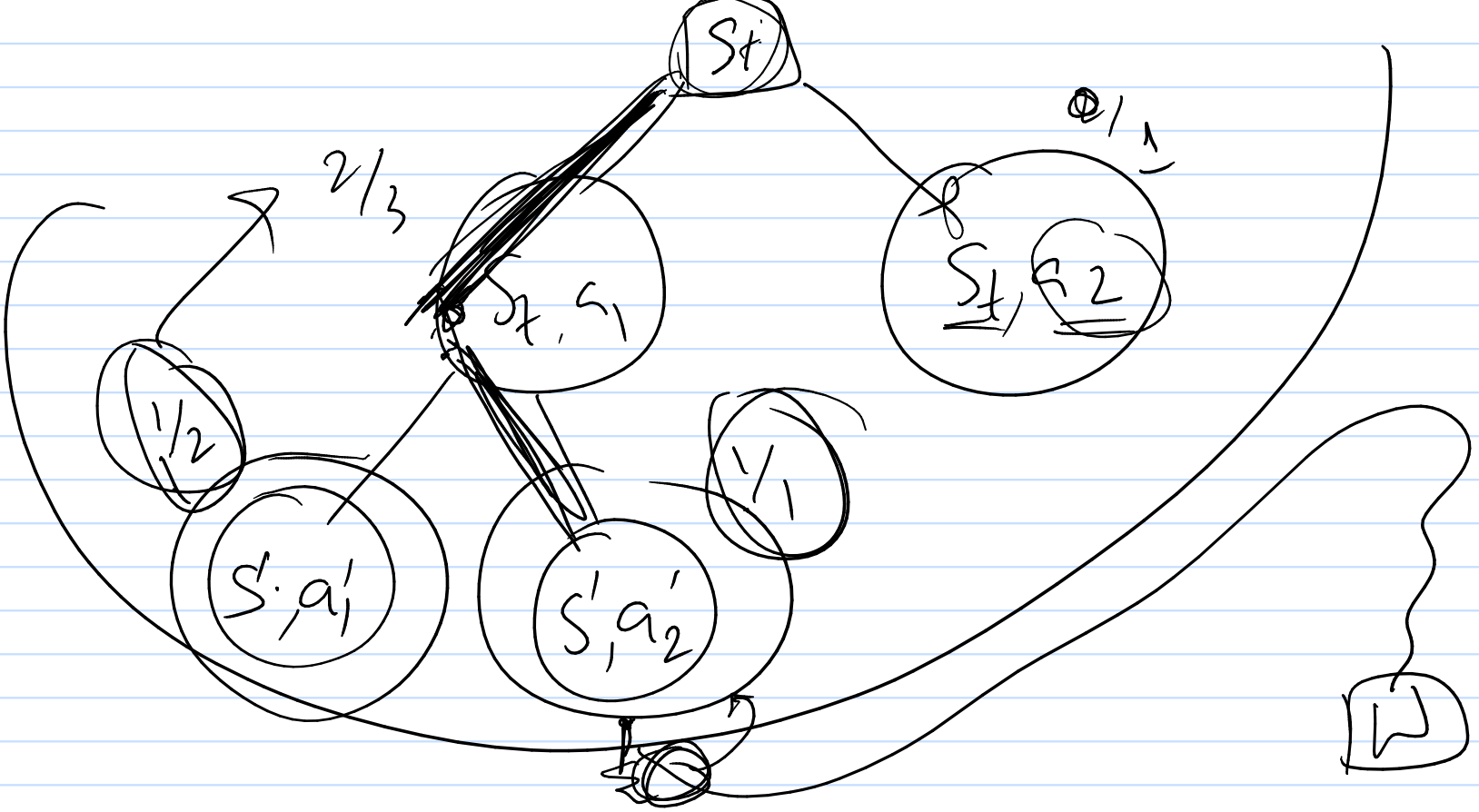
Goal:



$\approx \pi(s)$

MCS


s_t
s_t, a_1
$s_t a_2$
2/3
0/1
$Q(s_t, a_1) \approx$
$\approx Q(s_t, a_2)$
L
W
W
L


s_t
$2/3$
$1/1$
s_t, a_1
s_t, a_2
$1/2$
$1/1$
s', a_1'
s', a_2'
1