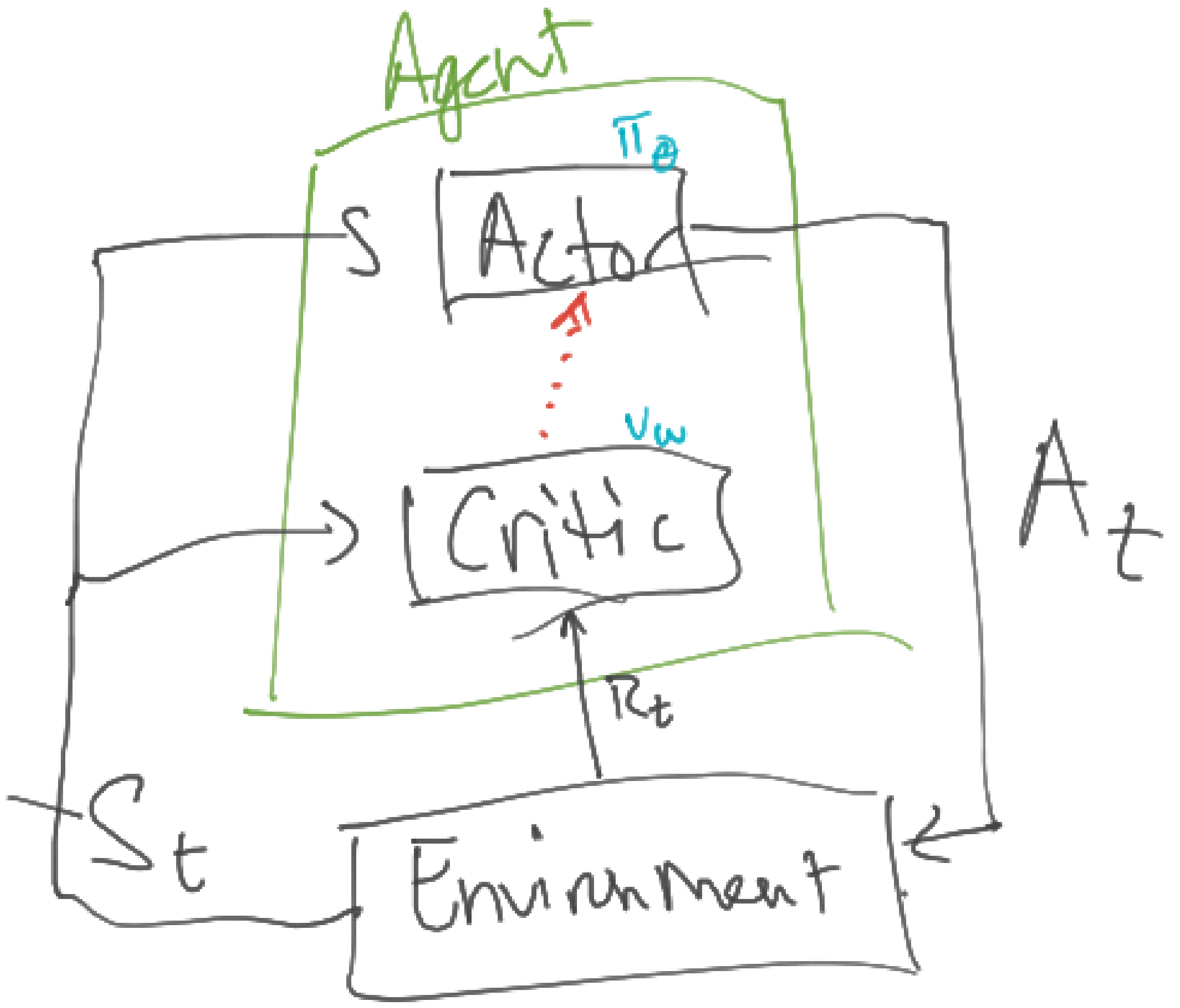
Actor - Critic

$$v^{\pi}(s) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} \,\middle|\, S_t = s ; \pi\right]$$



Hyperparams:
- Initial policy params θ
- Policy representation (ANN? Linear?)
- Actor step size α
- Critic step size β.
- Value function representation, v_w.
- Initial value function weights w.

For each episode:

For each time t

Agent observes S_t

Agent selects action A_t using π_θ

Env responds with S_{t+1} and R_t

$\delta_t = R_t + \gamma v_w(S_{t+1}) - v_w(S_t)$ // TD-error

$\forall i, \ \theta_i \leftarrow \theta_i + \alpha \, \delta_t \dfrac{\partial \ln(\pi_\theta(S_t, A_t))}{\partial \theta_i}$ // Actor update

$\forall j, \ w_j \leftarrow w_j + \beta \, \delta_t \dfrac{\partial v_w(S_t)}{\partial w_j}$ // critic update.

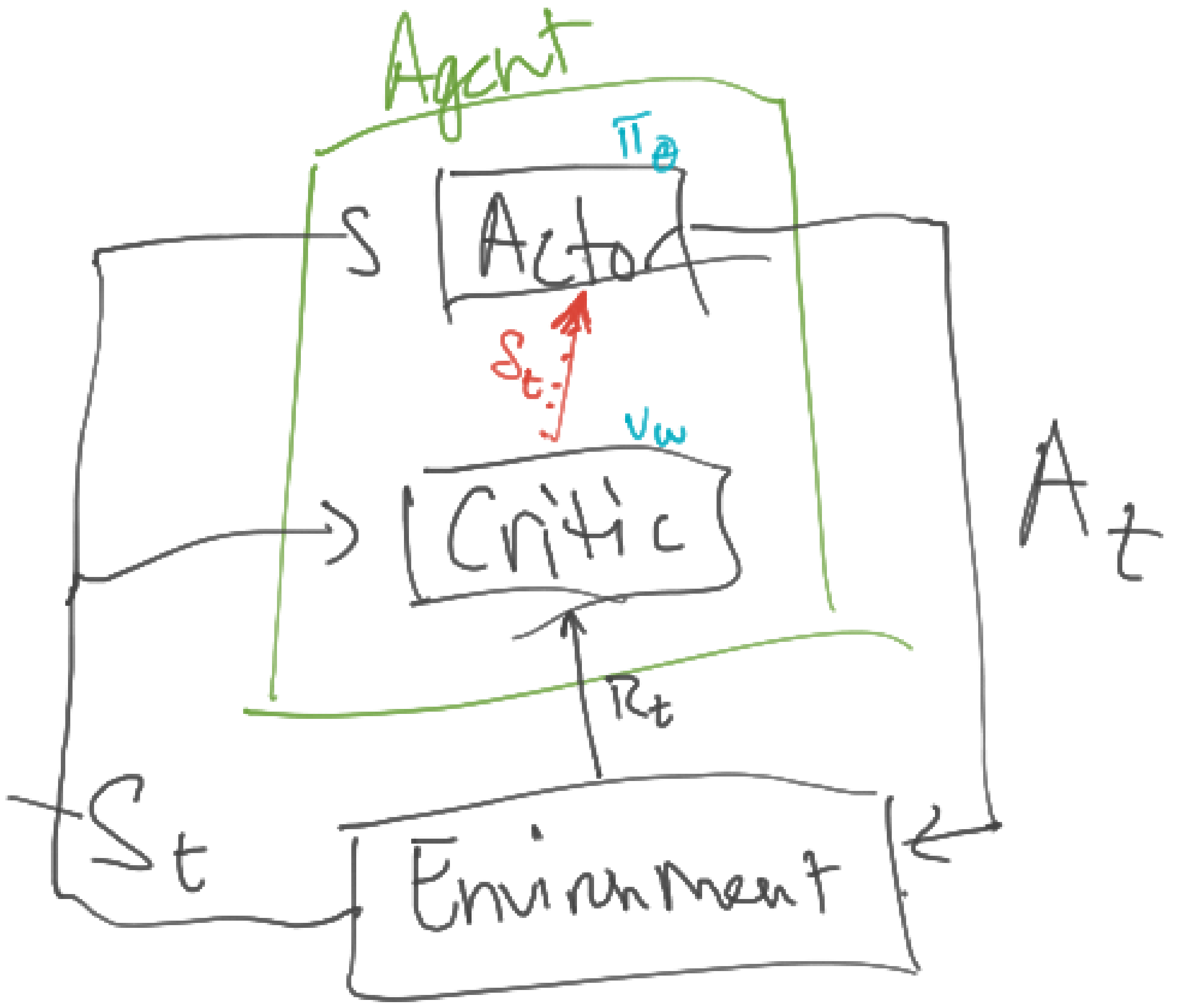
~~For each time t~~

~~$\forall j, \ w_j \leftarrow w_j + \beta \left(\cancel{\sum_{k=0}^{\infty} \gamma^k R_{t+k} - v_w(S_t)} \right) \dfrac{\partial v_w(S_t)}{\partial w_j}$ // critic update~~

δ_t

Actor - Critic

$$v^{\pi}(s) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} \,\middle|\, S_t = s ; \pi\right]$$



Hyperparams:

- Initial policy params θ
- Policy representation (ANN? Linear?)
- Actor step size α
- Critic step size β.
- Value function representation, v_w.
- Initial value function weights w.

For each episode:

  For each time t

    Agent observes S_t

    Agent selects action A_t using π_θ

    Env responds with S_{t+1} and R_t

    $\delta_t = R_t + \gamma v_w(S_{t+1}) - v_w(S_t)$ // TD-error

    $\forall i,\ \theta_i \leftarrow \theta_i + \alpha \boxed{\gamma^t} \delta_t \dfrac{\partial \ln(\pi_\theta(S_t, A_t))}{\partial \theta_i}$ // Actor update

    $\forall j,\ w_j \leftarrow w_j + \beta \delta_t \dfrac{\partial v_w(S_t)}{\partial w_j}$ // critic update.

Theory says to include
In practice it is <u>bad</u>. Almost nobody includes this term.

Learning v^π

- Supervised learning problem!
  → regression

input	output
S_0	$\sum_{k=0}^{\infty} \gamma^k R_{0+k}$
S_1	$\sum_{k=0}^{\infty} \gamma^k R_{1+k}$
$\vdots$	

Parametric model (f_w), v_w. $v_w(s)$ is an estimate of $v^\pi(s)$

Recall: (grad descent on least squares loss) (pointwise like backprop)

$w_{k,j}$ in old notation is w_j here.

$$\forall j,\ w_j \leftarrow w_j + \alpha\Big(\overset{\sum_{k=0}^{\infty}\gamma^k R_{t+k}}{\cancel{y}} - \underset{v_w(S_t)}{\cancel{f_w(x)}}\Big) \frac{\partial \overset{v_w(S_t)}{\cancel{f_w(x)}}}{\partial w_j}$$

$$w_j \leftarrow w_j + \alpha\Big(\sum_{k=0}^{\infty} \gamma^k R_{t+k} - v_w(S_t)\Big) \frac{\partial v_w(S_t)}{\partial w_j}$$

linear func approx

$$v_w(S_t) = \sum_j w_j \phi_j(S_t)$$

$$w_j \leftarrow w_j + \alpha\Big(\sum_{k=0}^{\infty} \gamma^k R_{t+k} - v_w(S_t)\Big) \phi_j(S_t)$$

Moving critic update into the episode:

$$\forall_j \quad w_j \leftarrow w_j + \alpha \left(\sum_{k=0}^{\infty} \gamma^k R_{t+k} - v_w(S_t) \right) \frac{\partial v_w(S_t)}{\partial w_j}$$

This is a (stochastic) gradient update on least squares loss.

$$R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} \cdots$$

$$\boxed{\forall_j \quad w_j \leftarrow w_j + \alpha \delta_t \frac{\partial v_w(S_t)}{\partial w_j}}$$

$$\ldots, S_t, A_t, R_t, S_{t+1}$$

NOT gradient update. (Not grad descent on least squares) Can diverge!

$$R_t + \gamma \underbrace{\left(R_{t+1} + \gamma R_{t+2} + \cdots \right)}_{v_w \cancel{v^\pi} (S_{t+1})}$$

$$\forall_j \quad w_j \leftarrow w_j + \alpha \Big(\overbrace{\underbrace{R_t + \gamma v_w(S_{t+1})}_{\text{label}} - \overbrace{v_w(S_t)}^{\text{prediction}}}^{\delta_t} \Big) \frac{\partial v_w(S_t)}{\partial w_j}$$

Policy Gradient

Run gradient ascent on $J(\theta) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t ; \pi_\theta\right]$

$$\forall j, \quad \theta_j \leftarrow \theta_j + \alpha \frac{\partial J(\theta)}{\partial \theta_j}$$

Algs presented are approximately doing this. Hence, called "policy gradient".

Not on exams or HW.

- REINFORCE → wait until episode end to update, use $\sum_{k=0}^{\infty} \gamma^k R_{t+k}$ not δ_t
  - → Missing γ^t term. Include $\ln$
  - → Do updates simultaneously
- Standard Actor-Critic.
  - → Many AC algorithms.
  - → This is just one.

Advanced Topics in RL

"Hierarchical RL"

Skills or options.

A Skill is a policy that the agent can choose to run.

- Go to UMass
  - Go to car
  - leave room
    - get out of chair
    - walk to door
    - open door
      - reach to handle
      - turn handle
      - pull open door.
  - look up directions
  - light BBQ
  - Drive
  - walk to location

skills

low level motor control.

Skill discovery.

- How can the agent learn what skills it should learn.

Off-policy evaluation

→ Have data from current policy

→ Goal: predict (with confidence bound) how good a new policy would be if we used it instead, but without actually using it.