Actor - Critic

Hyperparams:
- Initial policy params $\theta$
- Policy representation (ANN? Linear?)
- Actor step size $\alpha$
- Critic step size $\beta$
- 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 $\pi_{\theta}$
    Env responds with $s_{t+1}$ and $r_t$
    $v_t = r_t + \gamma v_w(s_{t+1}) - v_w(s_t)$ // TD-error
    $\theta_i \leftarrow \theta_i + \alpha \frac{\partial}{\partial \theta_i} J(\theta)$ // Actor update
    $w_j \leftarrow w_j + \beta \frac{\partial}{\partial w_j} J_w(w)$ // Critic update.

$$v(\sigma) = \mathbb{E}[\sum_{k=0}^{\infty} \gamma^k r_{t+k} | s_t = \sigma; \pi]$$
Psychology

Operant conditioning: learning process through which the strength of a behavior is modified by reward or punishment.

Classical conditioning: learning procedure in which a biologically potent stimulus (e.g., food) is paired with a previously neutral stimulus (e.g., bell).

- Thorndike's Puzzle Boxes
  1898
Neuroscience

- Dopamine
  - Contraction of 3,4-di hydroxyphenethylamine
  - Chemical neurotransmitters.

Dopamine neuron
(Dopaminergic neuron)
is a neuron that emits the neurotransmitter dopamine.

Two clusters in mammals:
- SNpc and VTA
  - Substantia nigra pars compacta
  - Ventral tegmental area.

SNpc → Striatum
- Coordinates motor & action planning, decision making, motivation.

VTA → numerous areas including prefrontal cortex
- Planning, personality, decision making.
Reward Prediction Error Hypothesis for dopamine.

TD error:
Olds & Milner 1954: Dopamine X reward.

Brains do not implement backpropagation.

Duplicate network in reverse.

Each "heurin" can update using only its input, output, and context.

Training RBN networks without backprop.
Reward devaluation Studies.

- early learning is model-based planning
- transitions to a model-free policy over time.

Addictive

\[ U(\text{cake in front}) \]

\[ U(\text{heroin (drug)}) = \]