Lecture 11. Reinforcement Learning

*Reading Assignments:**

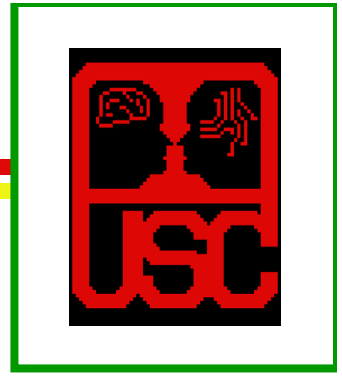
HBTNN:

Reinforcement Learning (Barto)

Reinforcement Learning in Motor Control (Barto)

* This week the HBTNN material is the required reading

Learning Feedback



In supervised learning, training information is in the form of desired, or 'target', responses.

The aspect of real training that corresponds most closely to the supervised learning paradigm is the trainer's role in telling or showing the learner what to do, or explicitly guiding his or her movements.

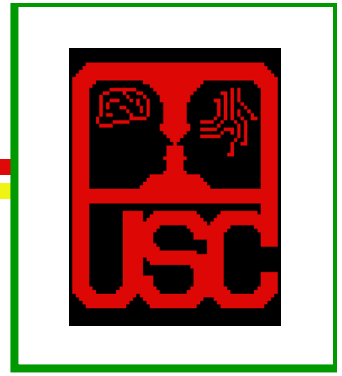
When motor skills are acquired without the help of an explicit teacher or trainer, learning feedback must consist of intrinsic feedback automatically generated by the movement and its

consequences on the environment.

E.g., the "feel" of a successfully completed movement and the sight of a basketball going through the hoop

A teacher or trainer can augment intrinsic feedback by providing extrinsic feedback

Reinforcement learning



Reinforcement: the occurrence of an event, in the proper relation to a response,that tends to increase the probability that the response will occur again in the same situation.

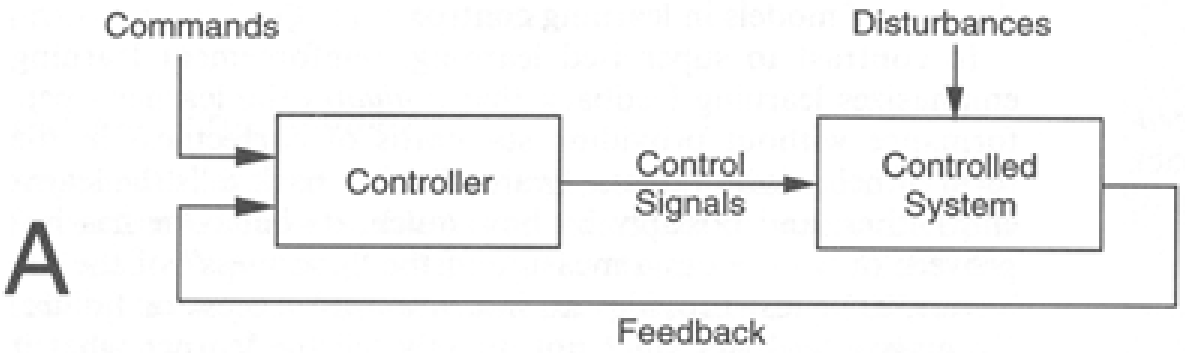
Reinforcement learning emphasizes learning feedback that evaluates the learner's performance without providing standards of correctness in the form of behavioral targets.

Evaluative feedback

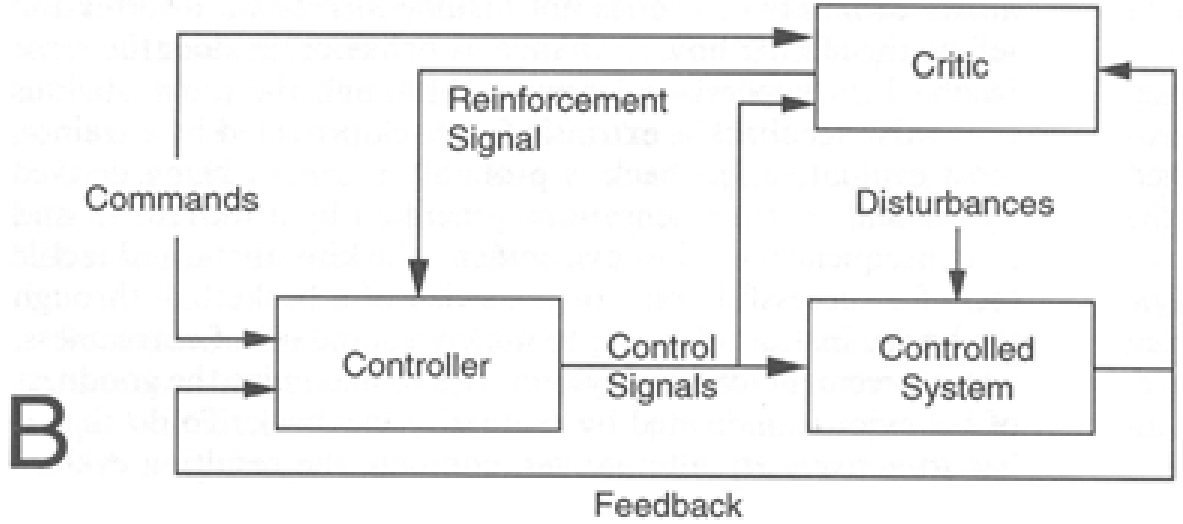
- tells the learner whether or not, and possibly by how much, its behavior has improved; or
- provides a measure of the 'goodness' of the behavior; or
- just provides an indication of success or failure.

Evaluative feedback does not directly tell the learner what it should have done, as does the feedback of supervised learning.

Learning From Consequences 1



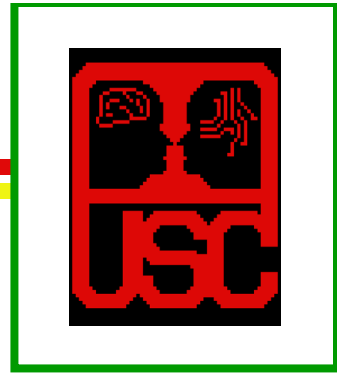
classical control system



What is the ***Critic?***

The control loop is augmented with another feedback loop that provides learning feedback to the controller.

Learning From Consequence 2



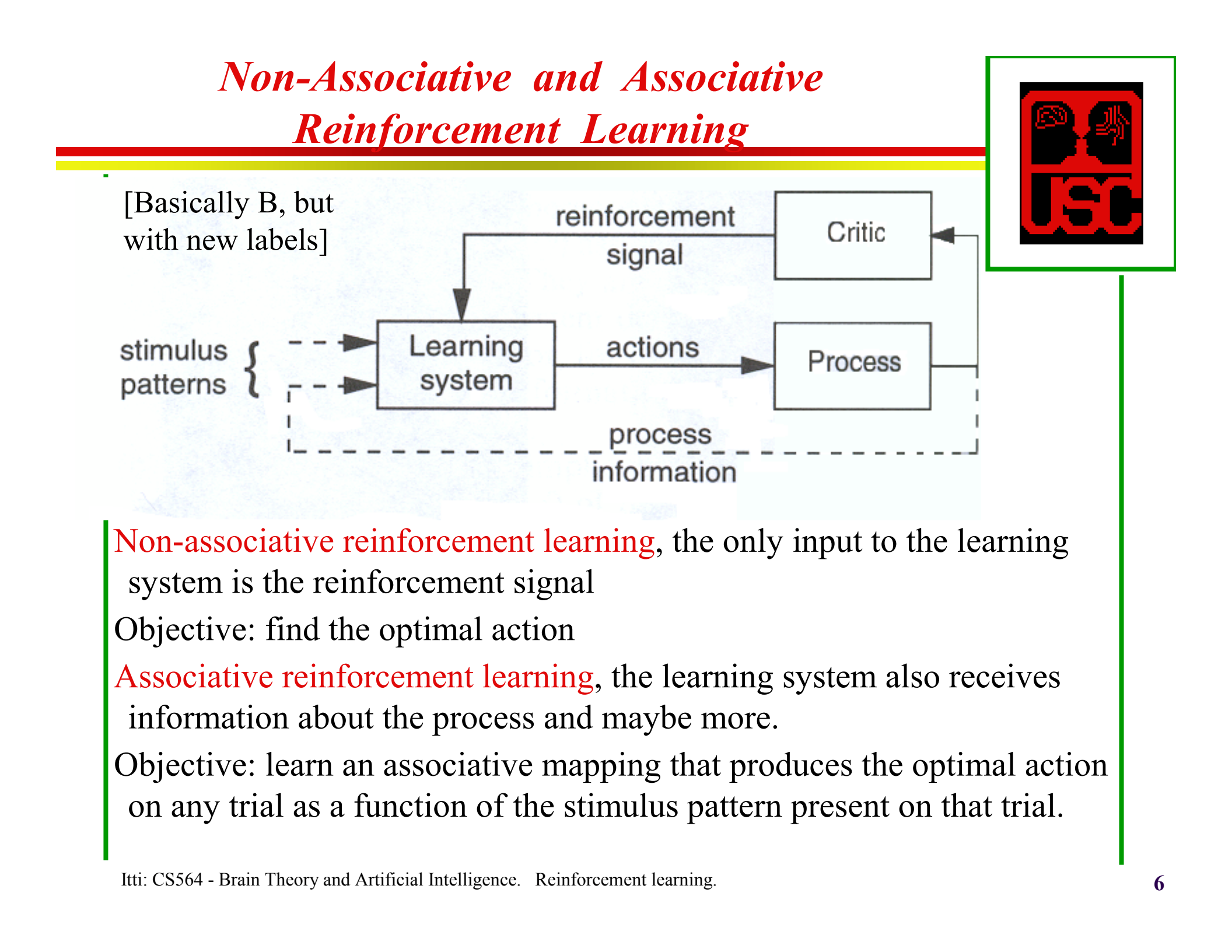
From Teacher to **Critic**:

The critic generates evaluative learning feedback on the basis of observing the control signals and their consequences on the behavior of the controlled system.

The critic also needs to know the command to the controller because its evaluations must be different depending on what the controller should be trying to do.

The critic is an abstraction of whatever process supplies evaluative learning feedback, both intrinsic and extrinsic, to the learning system.

Non-Associative and Associative Reinforcement Learning



[Basically B, but with new labels]

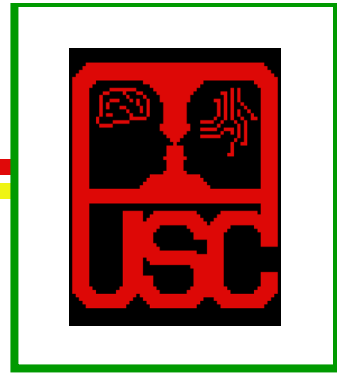
Non-associative reinforcement learning, the only input to the learning system is the reinforcement signal

Objective: find the optimal action

Associative reinforcement learning, the learning system also receives information about the process and maybe more.

Objective: learn an associative mapping that produces the optimal action on any trial as a function of the stimulus pattern present on that trial.

An example of non-associative reinforcement learning 1



The learning system has m actions a1, a2, ..., am.

The reinforcement signal simply indicates

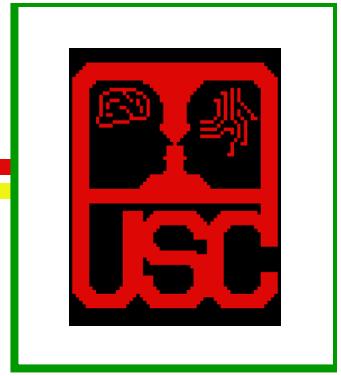
'success' or 'failure'.

The influence of the learning system's actions on the reinforcement signal can be modeled as a collection of success probabilities d1,d2, ..., dm

The learning system's objective is to eventually maximize the probability of receiving 'success'. This occurs if it always performs the action aj such that

$dj = \max \{di | i = 1, ..., m\}$.

An example of non-associative reinforcement learning 2



Desired outcome:

dj = max {di|i = 1, ..., m}.

Stochastic learning automaton:

On each trial, the system selects an action a(t) from its set of m actions according to a probability vector

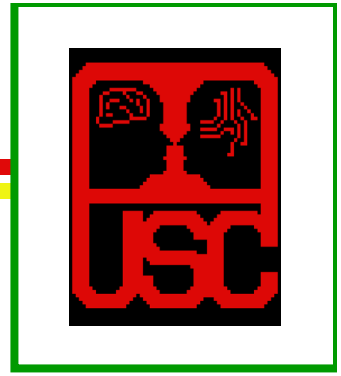
(p1(t),...,pm (t)), where pi(t) = Pr{a(t) = ai}.

Learning rule:

If action ai is chosen on trial t and the critic's feedback is 'success', then pi(t) is increased and the other probabilities are decreased;

If the critic indicates 'failure', then pi(t) is decreased and the probabilities of the other actions are increased.

A related associative reinforcement learning problem



Suppose that on trial t the learning system senses

stimulus pattern x(t) and selects an action a(t) = ai through a process that can depend on x(t).

After this action is executed, the critic signals success with probability di(x(t)) and failure with probability 1 - di(x(t)).

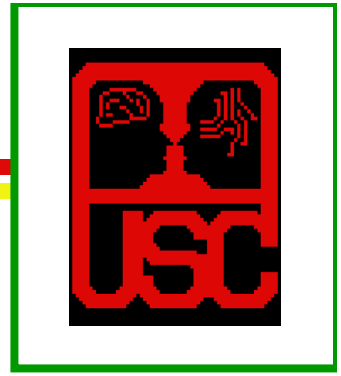
The objective of learning is to maximize success probability, i.e., to obtain a(x) = the action aj such that

dj(x(t)) = max {di(x(t))|i = 1,...,m}.

Unlike supervised learning:

Examples of optimal actions are not provided during training; they have to be discovered through "exploration".

Key Observations About Reinforcement Learning



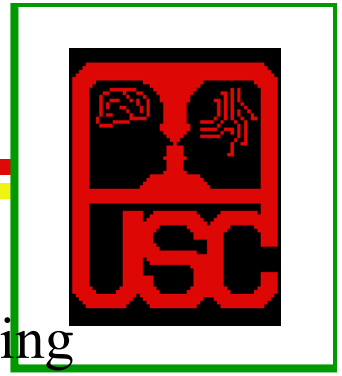
The reinforcement signal can be any signal evaluating the learning system's actions, not just a success/failure signal

Often it takes on real values, and the objective of learning is to maximize its expected value.

The critic does not directly tell the learning system how to change its actions.

Reinforcement learning algorithms are selectional processes. There must be variety in the action-generation process so that the consequences of alternative actions can be compared to select the best.

Exploitation and exploration



Behavioral variety = exploration

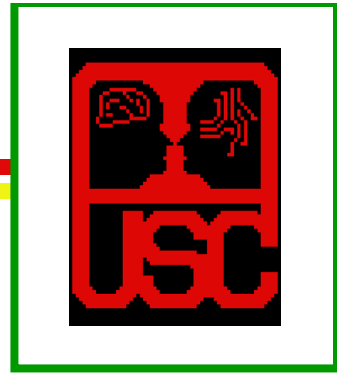
It is often generated through randomness (as in stochastic learning automata), but need not be.

Reinforcement learning involves a conflict between exploitation and exploration.

- exploiting what it has already learned to obtain high evaluations, vs
- exploring to learn more.

Reinforcement learning systems have to balance these strategies. cf. the conflict between control and identification.

Associative Reinforcement Learning Rules



Consider a neuron-like unit receiving a stimulus pattern as input in addition to the critic's reinforcement signal.

x(t), stimulus vector; w(t), weight vector; a(t), action; r(t).

$$s(t) = \sum_{i=1}^{n} w_i(t)x_i(t)$$

Associative Search Unit: The associative search rule, based on Klopf's (1982) self-interested neuron - the unit's output is a random variable depending on the activation level:

$$a(t) = \begin{cases} 1 \text{ with probability } p(t) \\ 0 \text{with probability } 1 - p(t) \end{cases}$$

where p(t), between 0 and 1, is an increasing function of s(t).

If the critic takes time τ to evaluate an action, the weights are updated according to: $\Delta w(t) = \eta\, r(t)\, a(t - \tau)\, x(t - \tau)$

where r(t) is +1 (success) or -1 (failure), and $\eta > 0$ is the learning rate parameter.

This is basically the Hebbian rule with the reinforcement signal acting as an additional modulatory factor.

The structural credit assignment problem

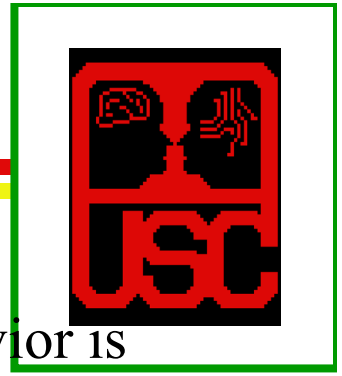
How is credit assigned to the internal workings of a complex structure?

- The backpropagation algorithm addresses structural credit assignment for artificial neural networks]
- Reinforcement learning principles lead to a number of alternatives: In these methods , a single reinforcement signal is uniformly broadcast to all the sites of learning, either neurons or individual synapses.
- Any task that can be learned via error backpropagation can also be learned using this approach, although possibly more slowly.

These network learning methods are consistent with the role of diffusely projecting neural pathways by which neuromodulators (cf. TMB2 §6.1) can be widely and nonspecifically distributed.

Hypothesis: Dopamine mediates synaptic enhancement in the corticostriatal pathway in the manner of a broadcast reinforcement signal (Wickens, 1990).

The Temporal Credit Assignment Problem

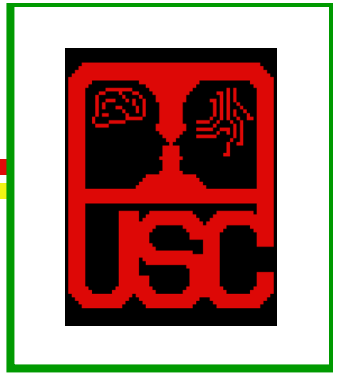


How can reinforcement learning work when the learner's behavior is temporally extended and evaluations occur at varying and unpredictable times?

It is especially relevant in motor control because movements extend over time and evaluative feedback may become available, for example, only after the end of a movement.

To address this, reinforcement learning is not only the process of improving behavior according to given evaluative feedback; it also includes learning how to improve the evaluative feedback itself: adaptive critic methods.

Dynamic Programming



Sequential reinforcement learning problems are examples of stochastic optimal control problems.

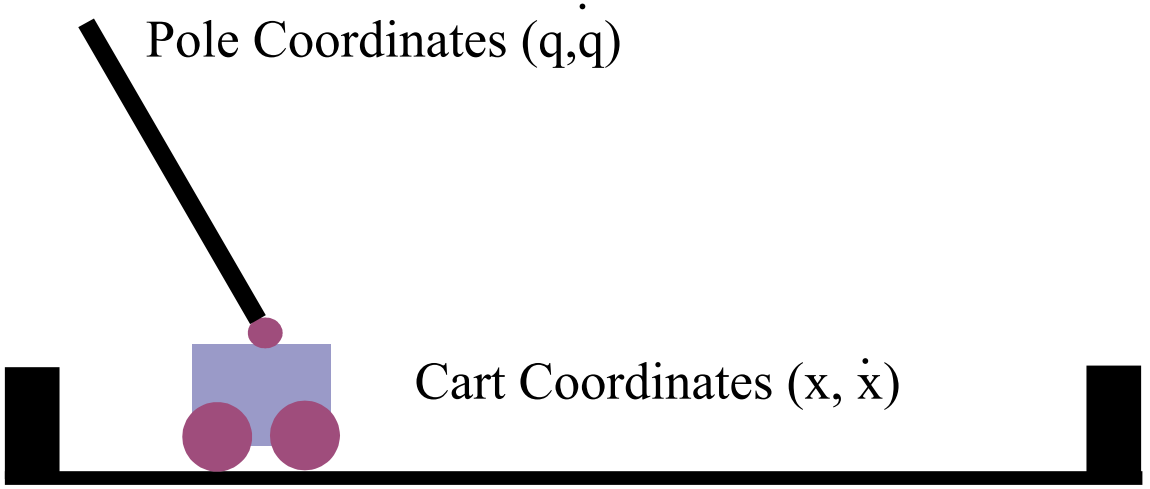
Among the traditional methods for solving these problems are dynamic programming (DP - Richard Bellman of USC!) algorithms.

A basic operation in all DP algorithms is "backing up" evaluations in a manner similar to the operation used in Samuel's method and in the adaptive critic.

But because conventional DP algorithms require multiple exhaustive "sweeps" of the process state set, they are not practical for problems with very large state sets or high-dimensional continuous state spaces.

Sequential reinforcement learning provides a collection of heuristic methods providing computationally feasible approximations of DP solutions to stochastic optimal control problems.

A Classic Example: Pole Balancing



If we used 5 values to discretize all 4 coordinates, we would have a state space of 5^4 values.

Problem: We know failure when we see it - when the cart hits the buffers or the pole falls over?

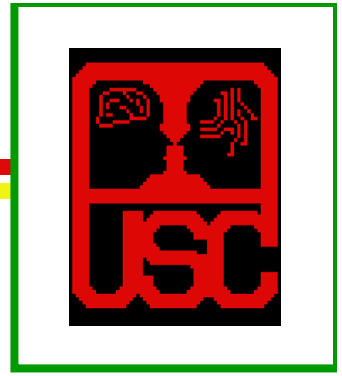
But how do we evaluate the other states?

The Adaptive Critic Solution: Climb the hill.

But there is no hill!

Build the hill - and then climb it!!

Samuel's Checkers Player 1



Samuel's (1959) checkers playing program (cf. TMB2, §3.4)

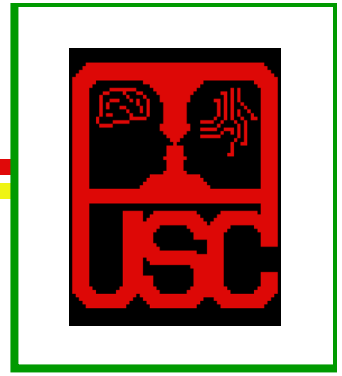
has been a major influence on adaptive critic methods.

The checkers player uses an evaluation function to assign a score to each board configuration, and

makes the move expected to lead to the configuration with the highest score.

Samuel used a method to improve the evaluation function through a process that compared the score of the current board position with the score of a board position likely to arise later in the game. As a result of this process of "backing up" board evaluations, the evaluation function improved in its ability to evaluate the long-term consequences of moves.

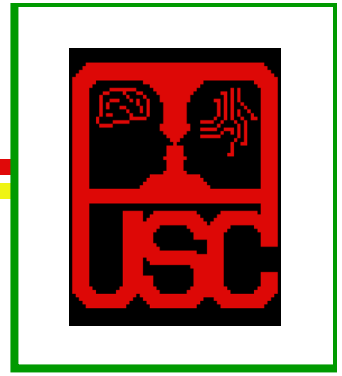
Samuel's Checkers Player



If the evaluation function can be made to score each board configuration according to its true promise of eventually leading to a win, then the best strategy for playing is to myopically select each move so that the next board configuration is the most highly scored.

If the evaluation function is optimal in this sense, then it already takes into account all the possible future courses of play.

Building the Hill You Climb



When there is no immediate reinforcement until a goal state is reached we have a delayed reward problem in which the learning system has to learn how to make the process enter a goal state.

The temporal credit-assignment problem:

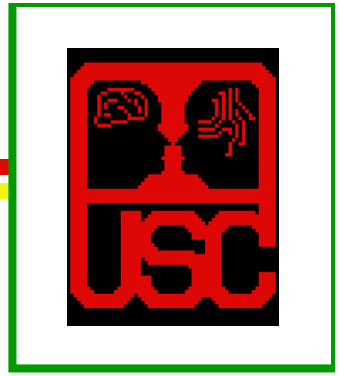
When a goal state is finally reached, which of the decisions made earlier deserve credit for the resulting reinforcement?

An approach:

Learn an internal evaluation function that is more informative than the evaluation function implemented by the external critic. "Build the Hill!!"

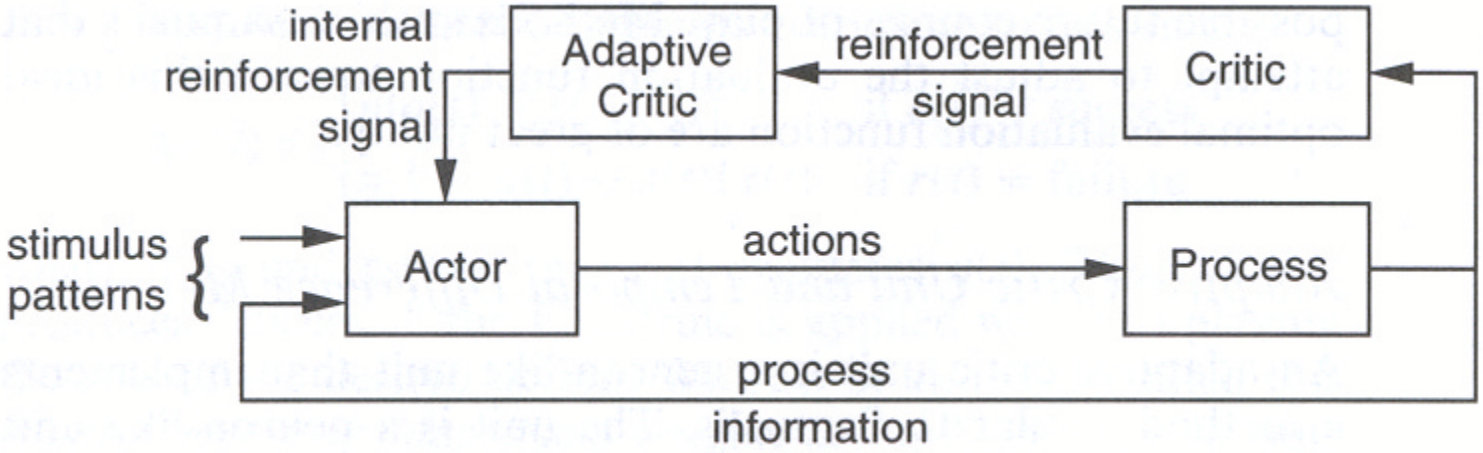
An adaptive critic is a system that learns such an internal evaluation function.

Sequential Reinforcement Learning

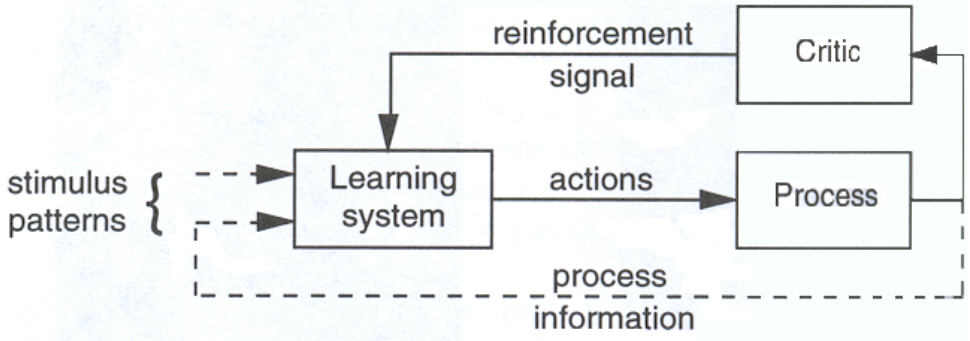


Sequential reinforcement requires improving the long-term consequences of a strategy:

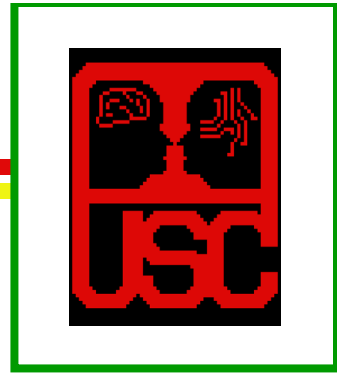
Actor-Critic Architecture:



Recall and Compare:



Actor-Critic Architectures



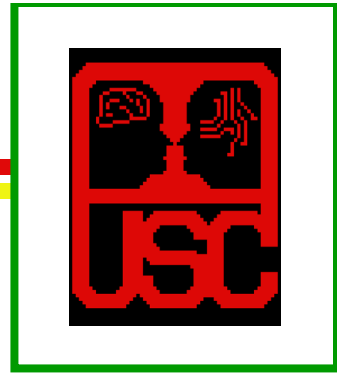
To distinguish the adaptive critic's signal from the reinforcement signal supplied by the original, non-adaptive critic, we call it the internal reinforcement signal.

The actor tries to maximize the immediate internal reinforcement signal

The adaptive critic tries to predict total future reinforcement.

To the extent that the adaptive critic's predictions of total future reinforcement are correct given the actor's current policy, the actor actually learns to increase the total amount of future reinforcement.

Sequential Reinforcement Learning 2



A sequential reinforcement learning system tries to influence the behavior of the process to

maximize the total amount of reinforcement received over time

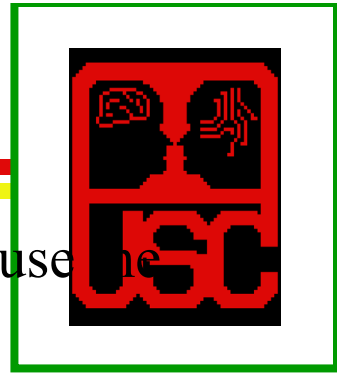
In the simplest case, this measure is the sum of the future reinforcement values, and the objective is to learn an associative mapping that at each time step t selects, as a function of the stimulus pattern x(t), an action a(t) that maximizes

$$(6) \qquad \sum_{k=0}^{\infty} r(t+k)$$

where r(t + k) is the reinforcement signal at step t + k.

Such an associative mapping is called a policy.

Discounting Future Rewards



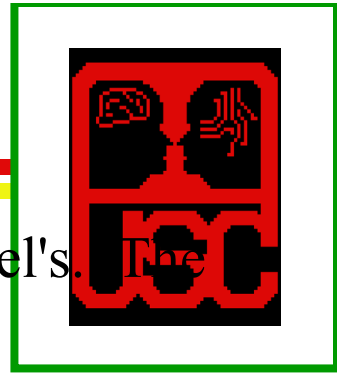
Because this sum might be infinite in some problems, and because the learning system usually has control only over its expected value, researchers often consider the following expected discounted sum instead:

(7) $E\{r(t) + \gamma\, r(t+1) + \gamma^2\, r(t+2) + \ldots\}$

where E is the expectation over all possible future behavior patterns of the process.

The discount factor determines the present value of future reinforcement: a reinforcement value received k time steps in the future is worth γ^k times what it would be worth if it were received now. If $0 < \gamma < 1$, this infinite discounted sum is finite as long as the reinforcement values are bounded.

An adaptive critic unit 1



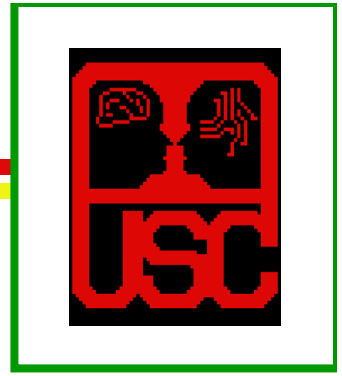
is a neuron-like unit that implements a method similar to Samuel's. The unit's output at time step t is

(8) $$P(t) = \sum_{i=1}^{n} w_i(t)x_i(t)$$

It is denoted by P because it is a prediction of the discounted sum of future reinforcement.

The adaptive critic learning rule rests on noting that correct predictions must satisfy a consistency condition relating predictions at adjacent time steps.

An adaptive critic unit 2



If the predictions at any two successive time steps, say steps t and t + 1, are correct, then

(9) $P(t) = E\{r(t) + \gamma r(t+1) + \gamma^2 r(t+2) + \ldots\}$

(10) $P(t+1) = E\{r(t+1) + \gamma r(t+2\} + \gamma^2 r(t+3) + \ldots\}$

But (11) $P(t) = E\{r(t) + \gamma\,[r(t+1\} + \gamma r(t+2) + \ldots]\}$

so that $P(t) = E\{r(t)\} + \gamma P(t+1)$.

An estimate of the error by which any two adjacent predictions fail to satisfy this consistency condition is called the

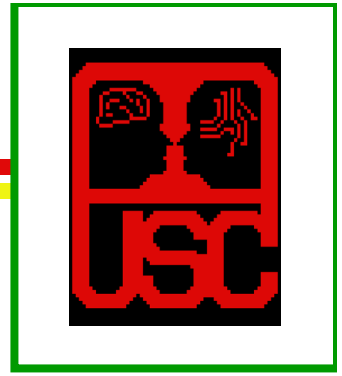
temporal difference (TD) error (Sutton,1988):

$$r(t) + \gamma P(t+1) - P(t)$$

where r(t) is used as an unbiased estimate of $E\{r(t)\}$.

The error essentially depends on the difference between the critic's predictions at successive time steps.

An Adaptive Critic Unit 3



$$P(t) = \sum_{i=1}^{n} w_i(t)x_i(t)$$

yields error estimated by $r(t) + \gamma P(t+1) - P(t)$

The adaptive critic unit adjusts its weights according to the following learning rule:

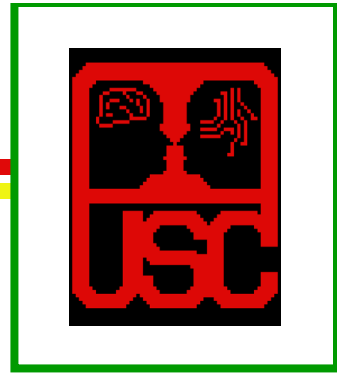
(12) $$\Delta w(t) = \eta[r(t) + \gamma P(t+1) - P(t)]\, x(t)$$

This rule changes the weights to decrease the magnitude of the TD error.

If $\gamma = 0$, it is equal to the LMS learning rule (Equation 3).

Think of $r(t) + \gamma P(t+1)$ as the prediction target: it is the quantity that each P(t) should match.

An Adaptive Critic Unit 4



$$P(t) = \sum_{i=1}^{n} w_i(t)x_i(t)$$

yields error estimated by $r(t) + \gamma P(t+1) - P(t)$

The adaptive critic unit adjusts its weights according to the following learning rule:

(12) $\Delta w(t) = \eta[r(t) + \gamma P(t+1) - P(t)]\, x(t)$

This rule changes the weights to decrease the magnitude of the TD error.

If $\gamma = 0$, it is equal to the LMS learning rule (Equation 3).

Think of $r(t) + \gamma P(t+1)$ as the prediction target: it is the quantity that each P(t) should match.