Reinforcement Learning

With help from

FIAS Frankfurt Institute
for Advanced Studies

A Taxonomoy of Learning

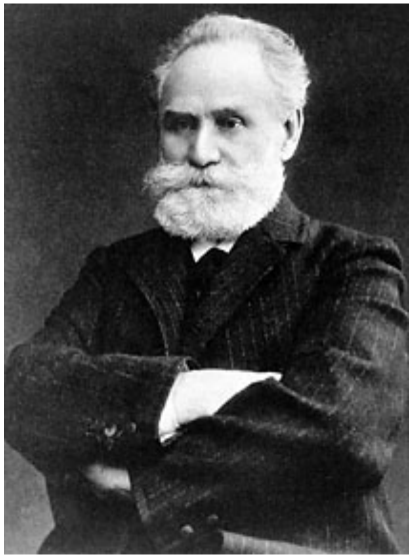
- L. of representations, models, behaviors, facts, …
  - Unsupervised L.
  - Self-supervised L.
  - Reinforcement L.
  - Imitation L.
  - Instruction-based L.
  - Supervised L.

increasing amount of “help” from the environment

- In addition, there’s evolutionary “learning”

Classical Conditioning

Pavlov's finding (1890-1900s):
Initially, sight of food leads to dog salivating. If sound of bell consistently accompanies or precedes presentation of food, then after a while the sound of the bell leads to salivating.



Terminology:

food →
salivating
unconditioned stimulus, US unconditioned response, UR
(reward)

Sound of bell consistently precedes food. Afterwards, bell leads to salivating: →

bell salivating
conditioned stimulus, CS conditioned response, CR
(expectation of reward)

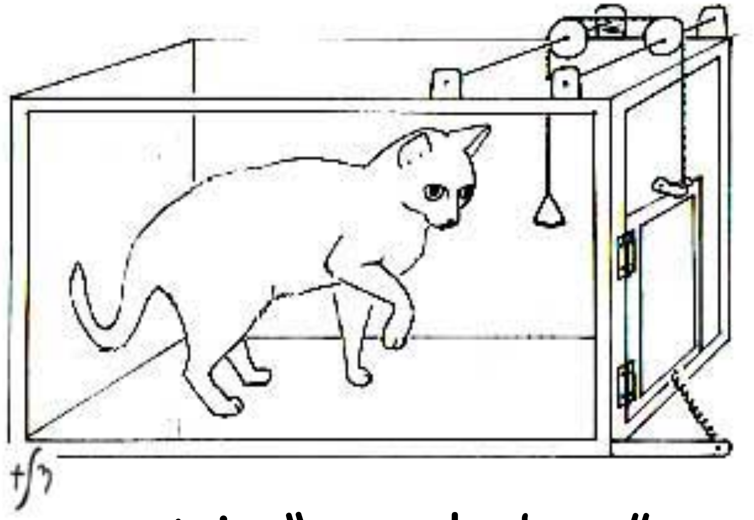
Instrumental Conditioning

- Classical Conditioning: only concerned with prediction of reward; doesn't consider agent's actions. Reward usually depends on what you've done!

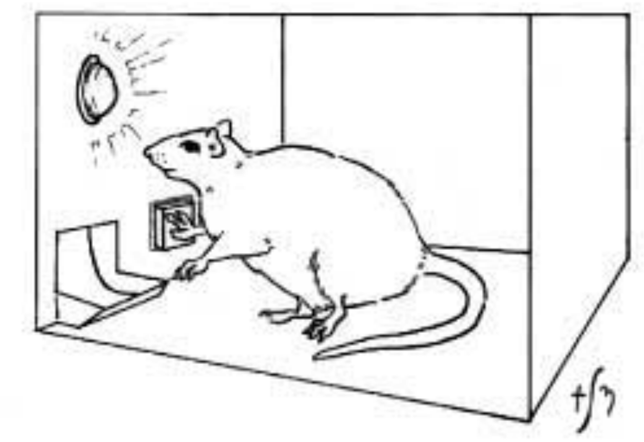
Edward L. Thorndike's „Law of effect" (1911):

- Responses to a situation that are followed by satisfaction are strengthened
- Responses that are followed by discomfort are weakened.

> "Of several responses made to the same situation, those which are accompanied or closely followed by satisfaction to the animal will, other things being equal, be more firmly connected with the situation, so that, when it recurs, they will be more likely to recur; those which are accompanied or closely followed by discomfort to the animal will, other things being equal, have their connections with that situation weakened, so that, when it recurs, they will be less likely to occur. The greater the satisfaction or discomfort, the greater the strengthening or weakening of the bond"
>
> (Thorndike, 1911, p. 244)



cat in "puzzle box"



rat in "Skinner box"

What is Reinforcement Learning?

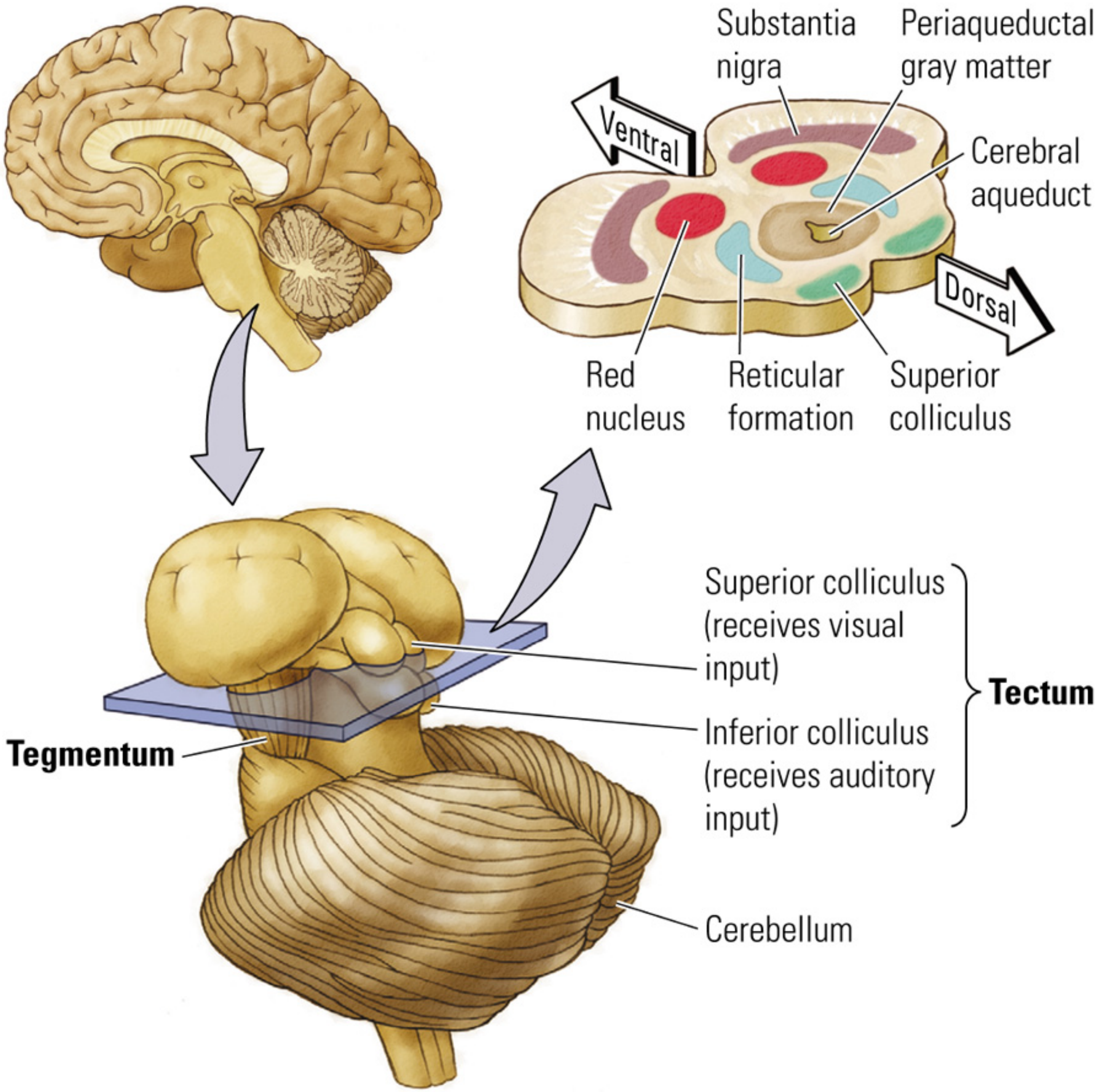
Central aspects:

- learning what to do: mapping situations to actions
- focus on learning through interaction with environment (there is no explicit teacher but agent engages in different behaviors and observes outcomes)
- evaluative feedback from environment (pleasure&pain) that must be predicted

Two key problems:

- trial and error learning leads to the *exploration versus exploitation dilemma*
- delayed rewards lead to the *temporal credit assignment problem*

Reward Signals in the Brain



Midbrain Dopamine Neurons

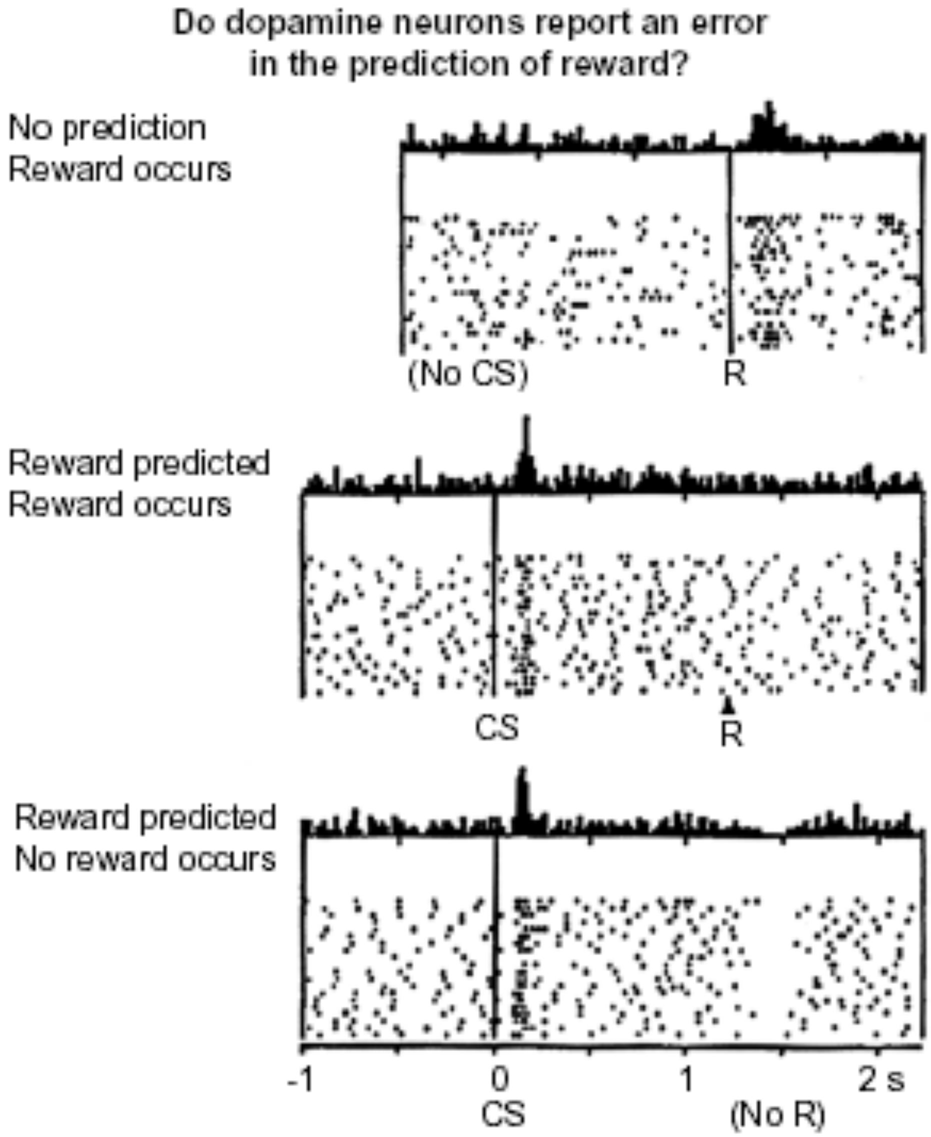
Source: (http://www.scholarpedia.org/article/Reward_Signals)

- phasic activation following primary food and liquid rewards, visual, auditory and somatosensory reward-predicting stimuli and physically intense visual and auditory stimuli
- briefly depressed by reward omission and by stimuli predicting the absence of reward
- respond very little to aversive stimuli and not at all to inedible objects and known neutral stimuli unless they are very intense or large
- seemingly signals like difference between actual and predicted reward

SCIENCE • VOL. 275 • 14 MARCH 1997

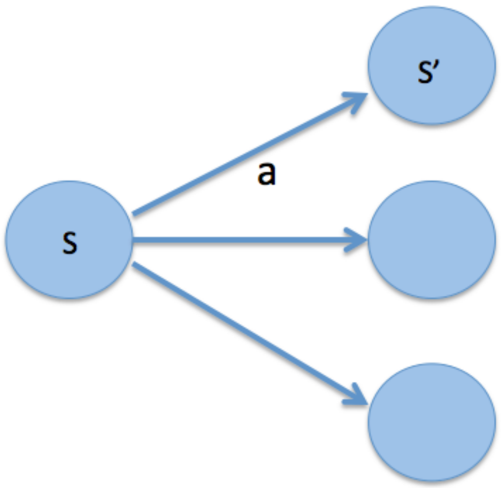
A Neural Substrate of Prediction and Reward

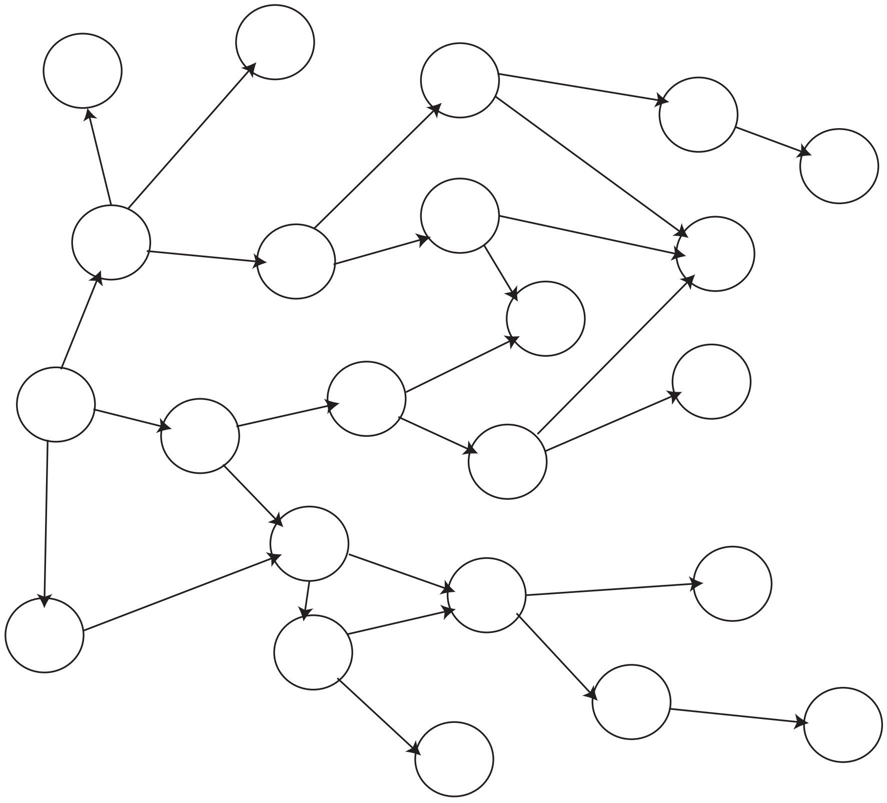
Wolfram Schultz, Peter Dayan, P. Read Montague*

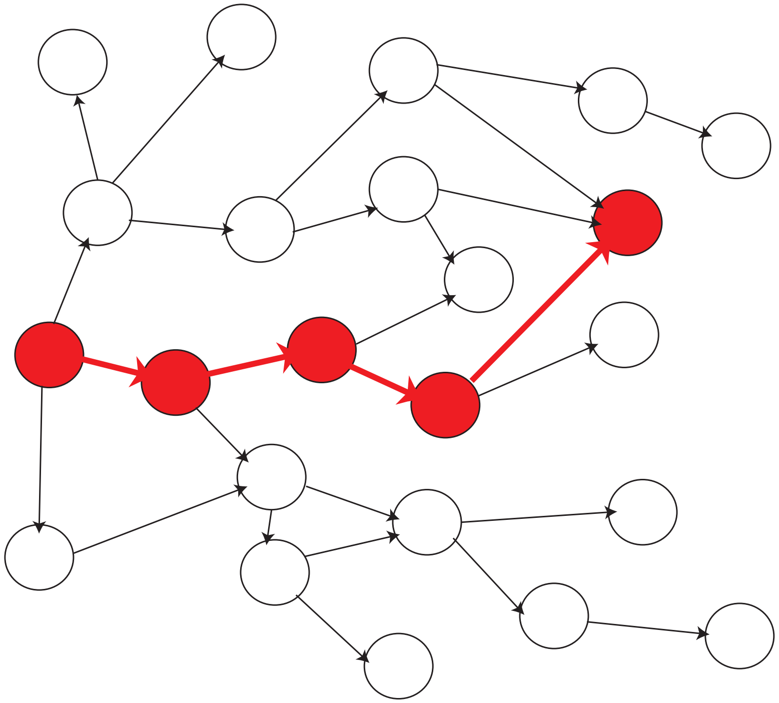


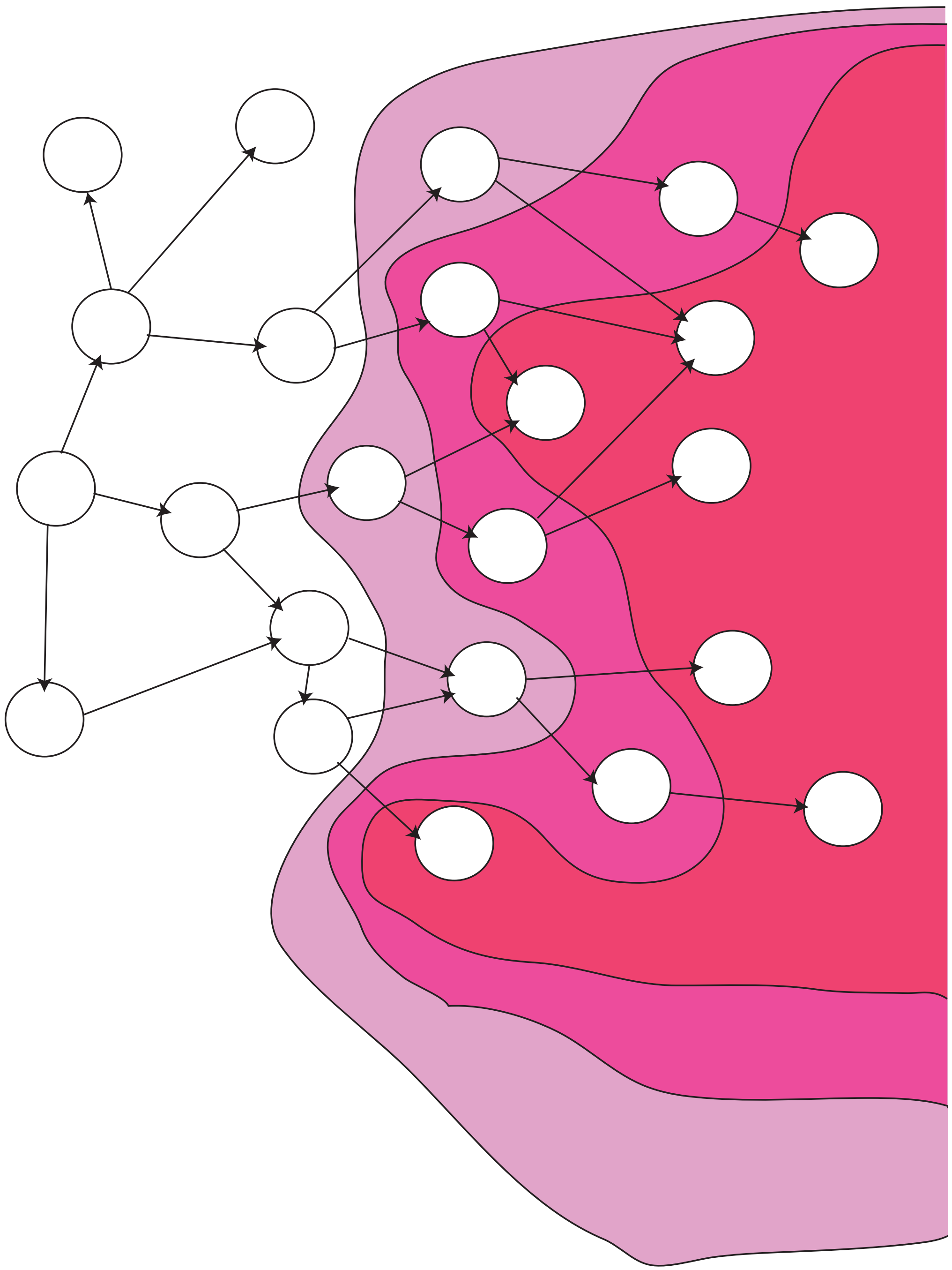
dopaminergic neurons can signal unexpected rewards in a way that qualitatively resembles a so-called *temporal difference error*

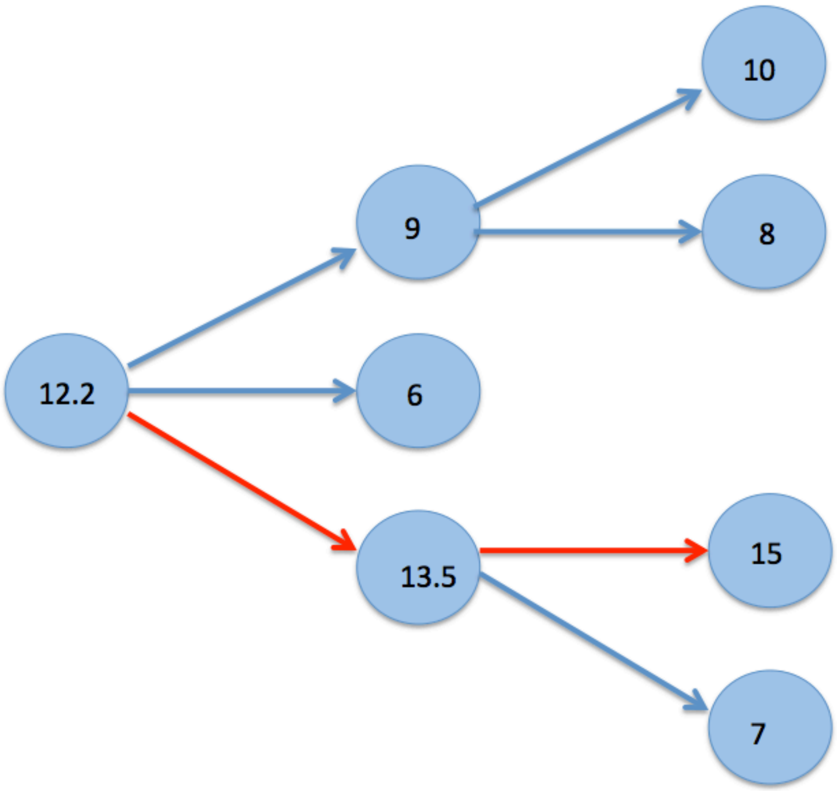
Basic idea: search










12.2
9
6
13.5
10
8
15
7

n-armed Bandit Problem

Non-associative case: no different states to distinguish

Aim: focus on exploration vs. exploitation dilemma

each round pull one of *n* arms and receive random reward from unknown distribution specific to that slot machine. How to maximize the reward over time?

Action Values

Def.: *action value* $Q^*(a)$ = true value of action
= average reward when playing *a.*

This needs to be estimated since it's unknown. Define the estimate $Q_t(a)$ after choosing action a k_a times as the sample mean:

$$Q_t(a) = \frac{r_1 + r_2 + \ldots + r_{k_a}}{k_a}$$

If action *a* tried often enough: $Q_t(a)$ converges to $Q^*(a)$
We also need an initial estimate before *a* has been tried.
Let's choose a default value $Q_0(a)$

Non-stationary problems

- Problem: what if rewards are changing over time?
- Idea: want to forget about old *r* values. Can do so by using an incremental update rule:

$$\begin{aligned} Q_{k+1} &= \mathrm{oldEstimate} + \mathrm{stepsize(target - oldEstimate)} \\ &= Q_k + \alpha\left(r_{k+1} - Q_k\right) \\ &= \alpha r_{k+1} + (1-\alpha)Q_k \longleftarrow \end{aligned}$$

weighted average of new target and old estimate

- can also be seen as weighted average of all previous rewards, with exponentially smaller weights for rewards far back in time:

$$Q_k = (1-\alpha)^k Q_0 + \sum_{i=1}^{k} \alpha(1-\alpha)^{k-i} r_i$$

greedy and ε-greedy action selection

greedy policy:

- always chose action *a* whose $Q_t(a)$ is maximal
- but: no effort spent on exploring seemingly inferior actions to see if they aren't better than previously thought.

ε-greedy policy:

- with small probability ε choose action *at random* in order to explore, otherwise choose greedy action

Example test bed:

- 10-armed bandit task, 1000 plays total (=1 experiment)
- repeat experiment 2000 times with different normally distributed rewards for each arm

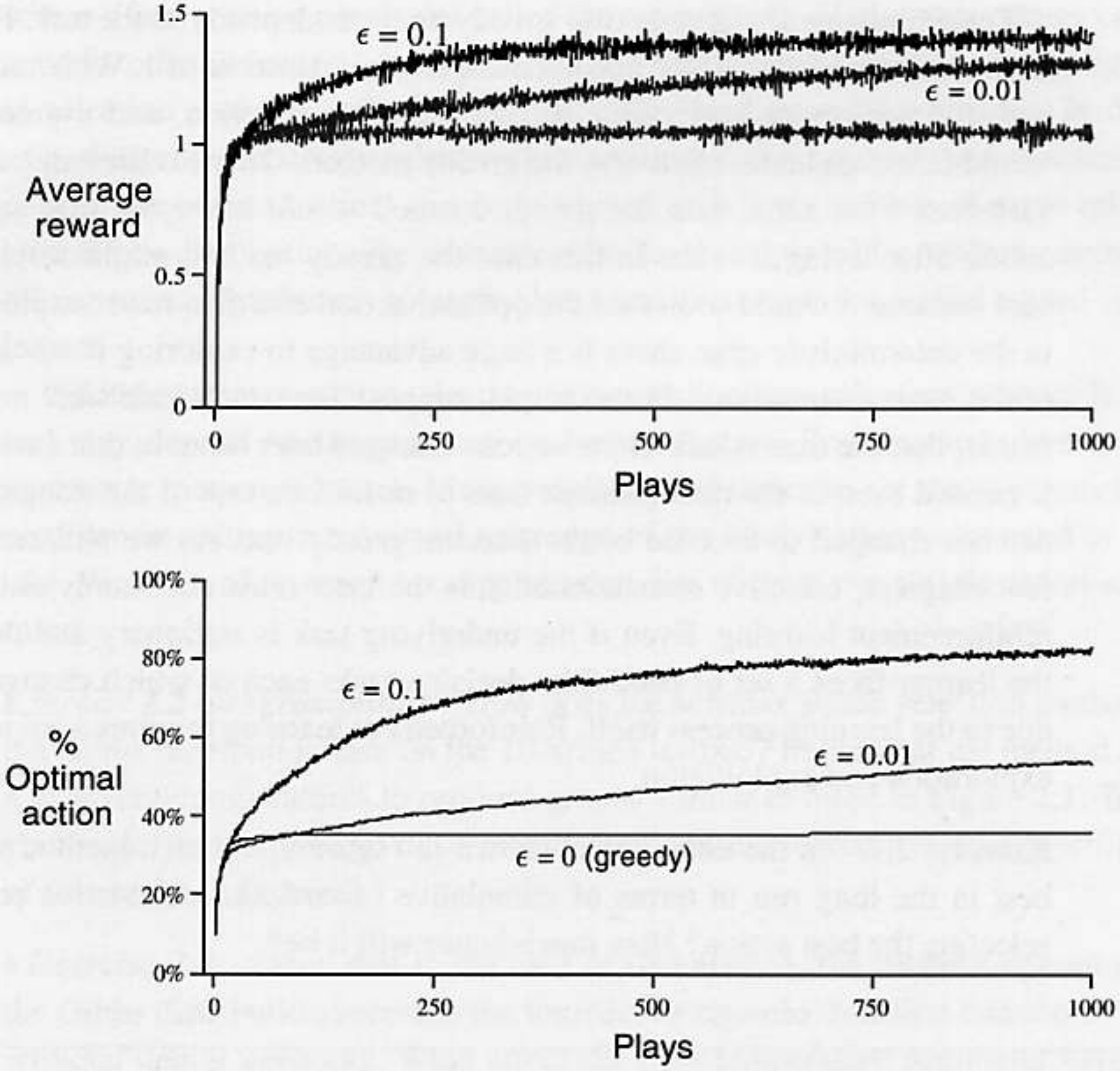


Figure 2.1 Average performance of ϵ-greedy action-value methods on the 10-armed testbed. These data are averages over 2000 tasks. All methods used sample averages as their action-value estimates.

figure taken from Sutton&Barto

Results of Experiments on 10-armed bandit test bed:

- greedy method tends to perform poorly in the long run
- higher ε leads to faster learning
- higher ε also leads to less exploitation since optimal action only chosen with probability ($1-\varepsilon$) and inferior actions are chosen otherwise

- this is an instance of the famous:

"Exploration Exploitation Dilemma"

Softmax Action Selection

Drawback of ε-greedy method: very bad actions are explored as frequently as very good ones

Idea: "better" actions should be explored more often

$$p(a) = \frac{\exp\left(Q_t(a)/\tau\right)}{\sum_{b=1}^{n} \exp\left(Q_t(b)/\tau\right)}$$

Boltzmann or *Gibbs* distribution,
τ: "temperature"

Note: $\Sigma p(a)=1$

Limiting cases:

- τ → 0: choose action with highest $Q_t(a)$ (greedy policy)
- τ → ∞: choose all actions with same probability
- sometimes the inverse temperature $\beta=1/\tau$ is used

Returns for Continuing Tasks

Continuing tasks: interaction does not have natural episodes.

Discounted return:

$$R_t = r_{t+1} + \gamma\, r_{t+2} + \gamma^2 r_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1},$$

where $\gamma, 0 \le \gamma \le 1$, is the **discount rate**.

shortsighted $0 \leftarrow \gamma \rightarrow 1$ farsighted

Temporal Difference RL

Want $$V(\boldsymbol{x}_t) = \gamma V(s_{t+1})$$

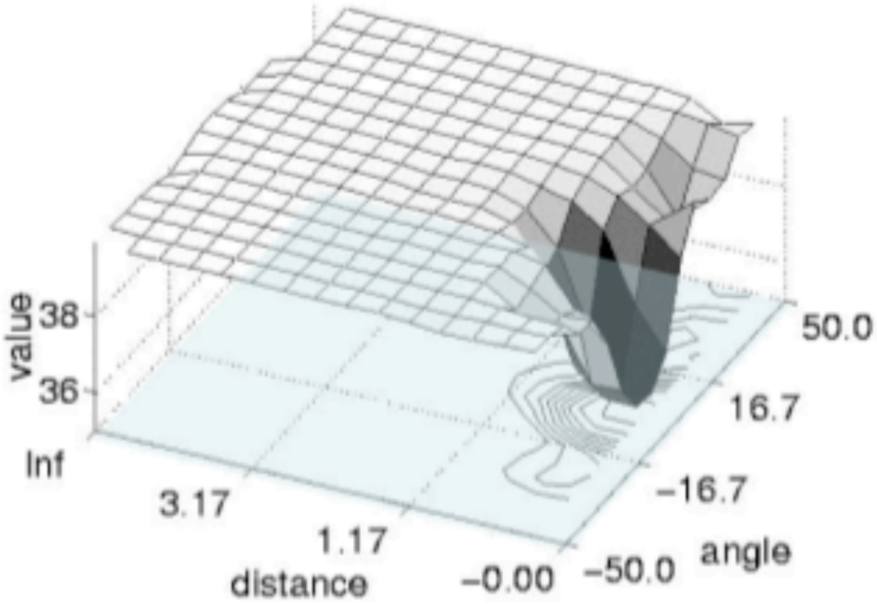
So use ... $$\Delta V(s_t) = \alpha[V(s_t) - \gamma V(s_{t+1})]$$

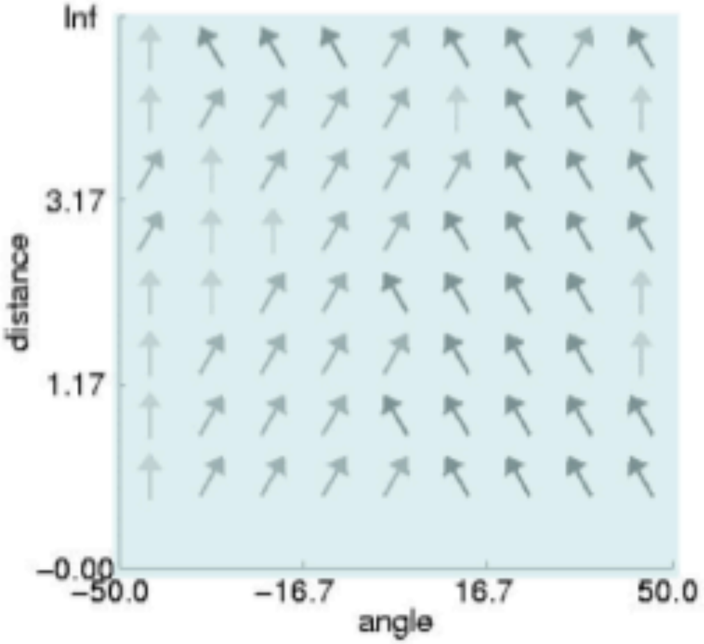
Q Learning variation

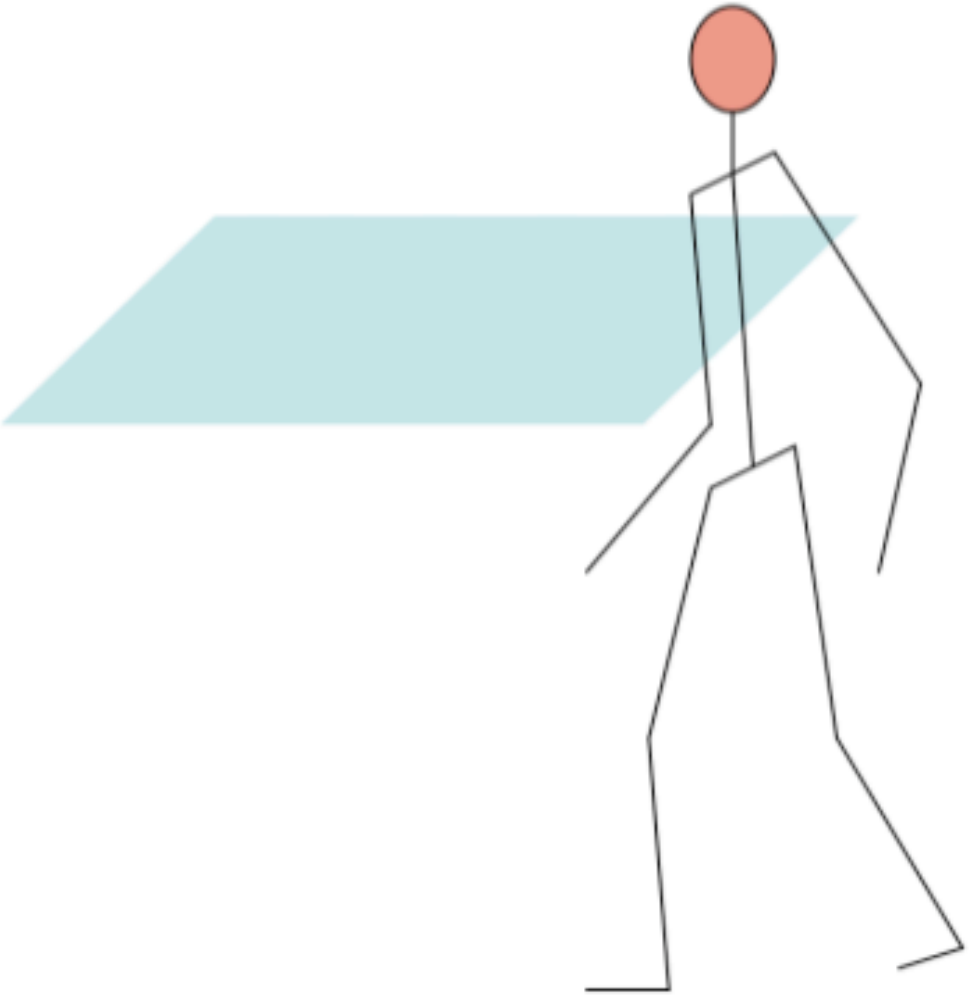
$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\delta_Q$$

$$\delta_Q = r_t + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)$$

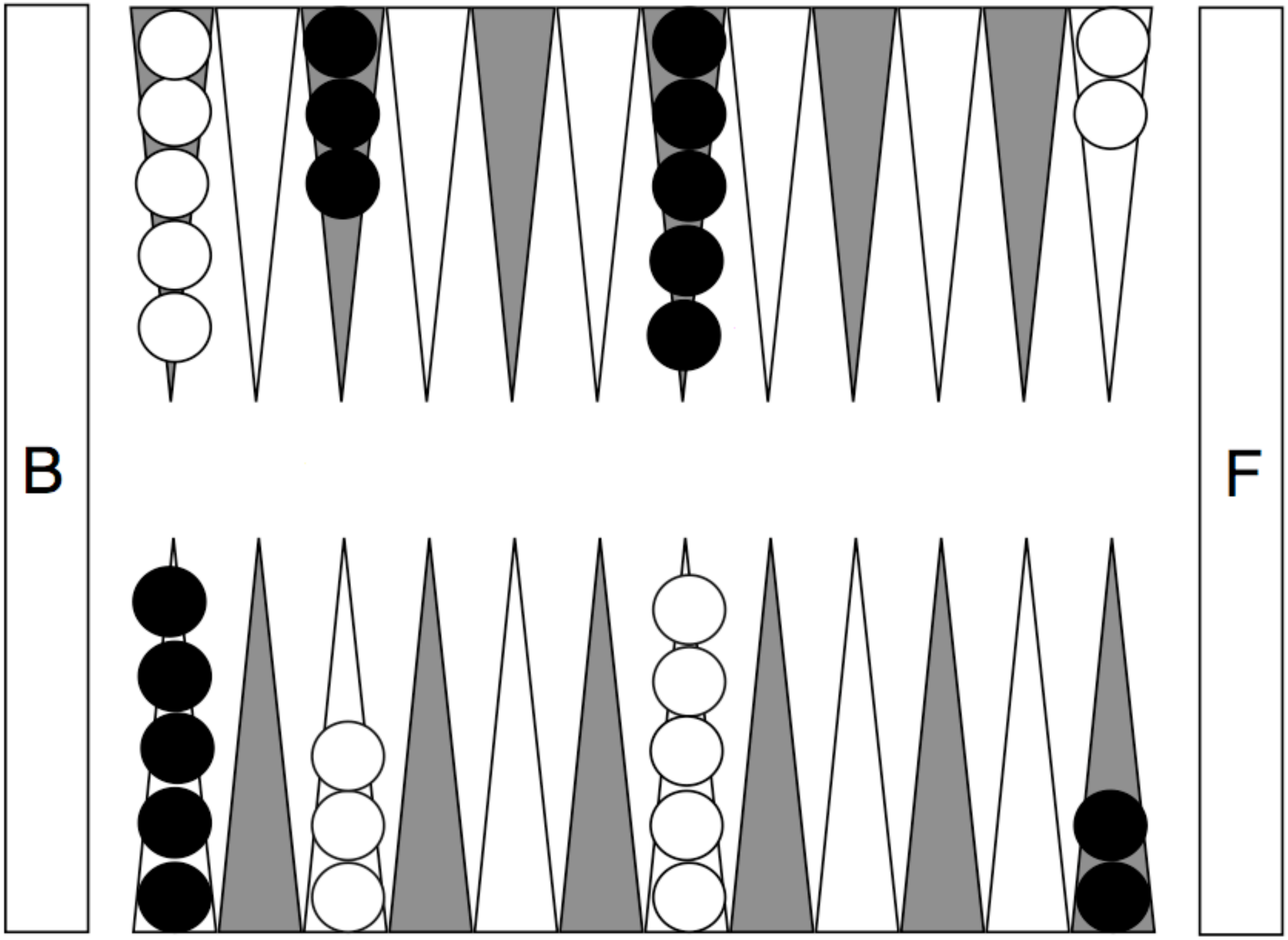
State Space


value
38
36
Inf
3.17
1.17
distance
-0.00
-50.0
-16.7
16.7
50.0
angle

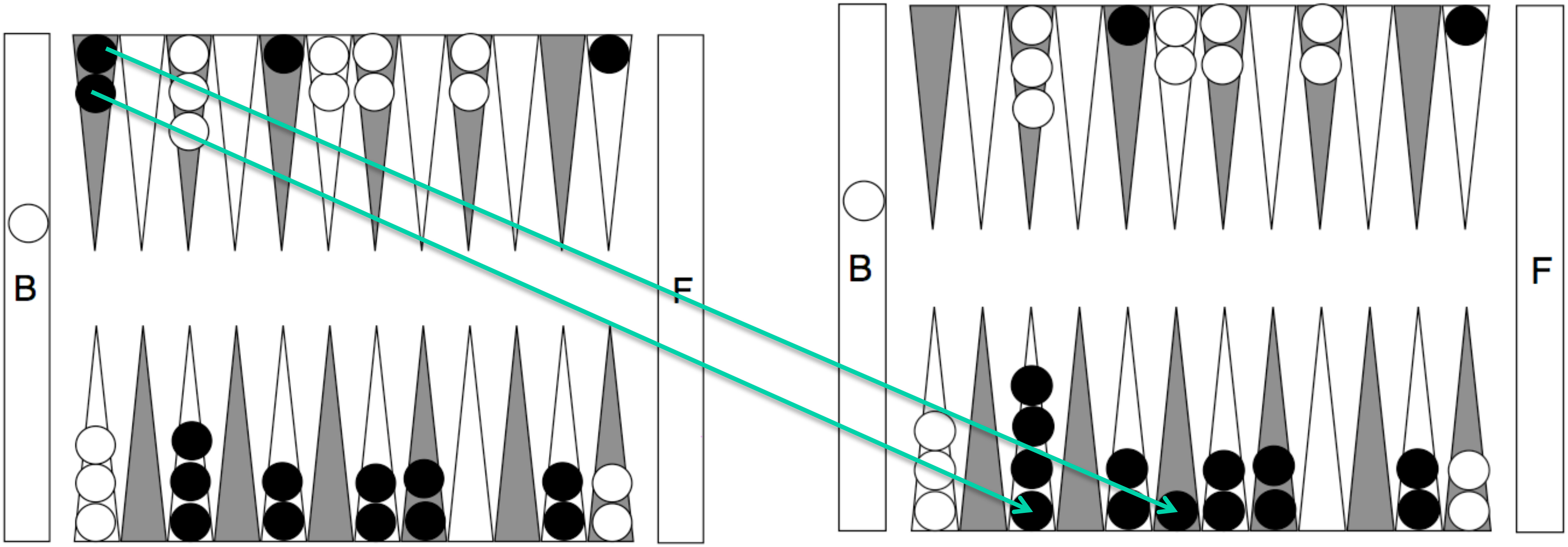

Inf
3.17
1.17
-0.00
distance
-50.0
-16.7
16.7
50.0
angle



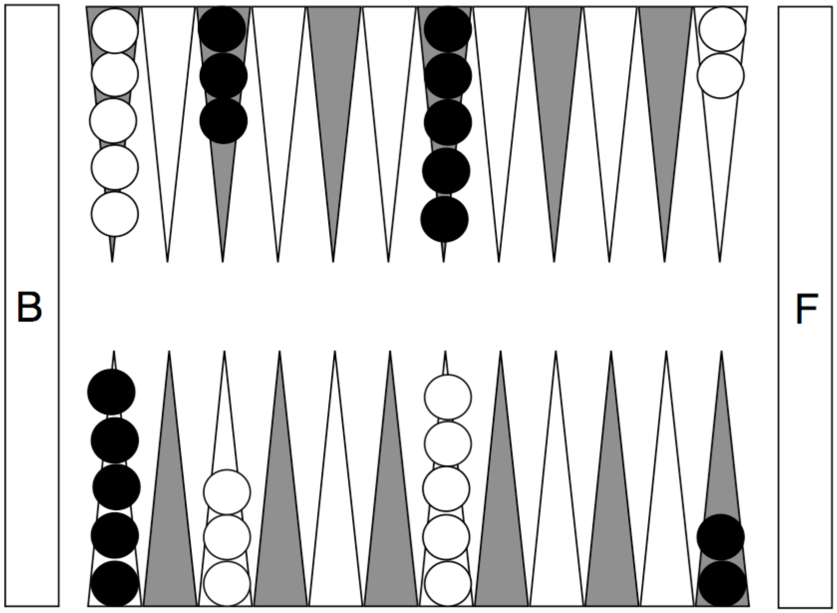
Backgammon



A move in Backgammon

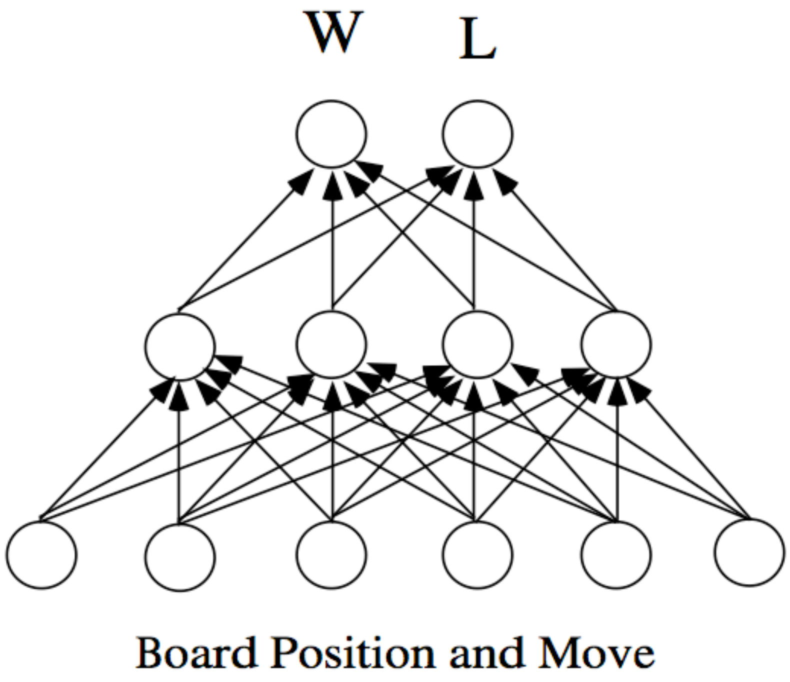


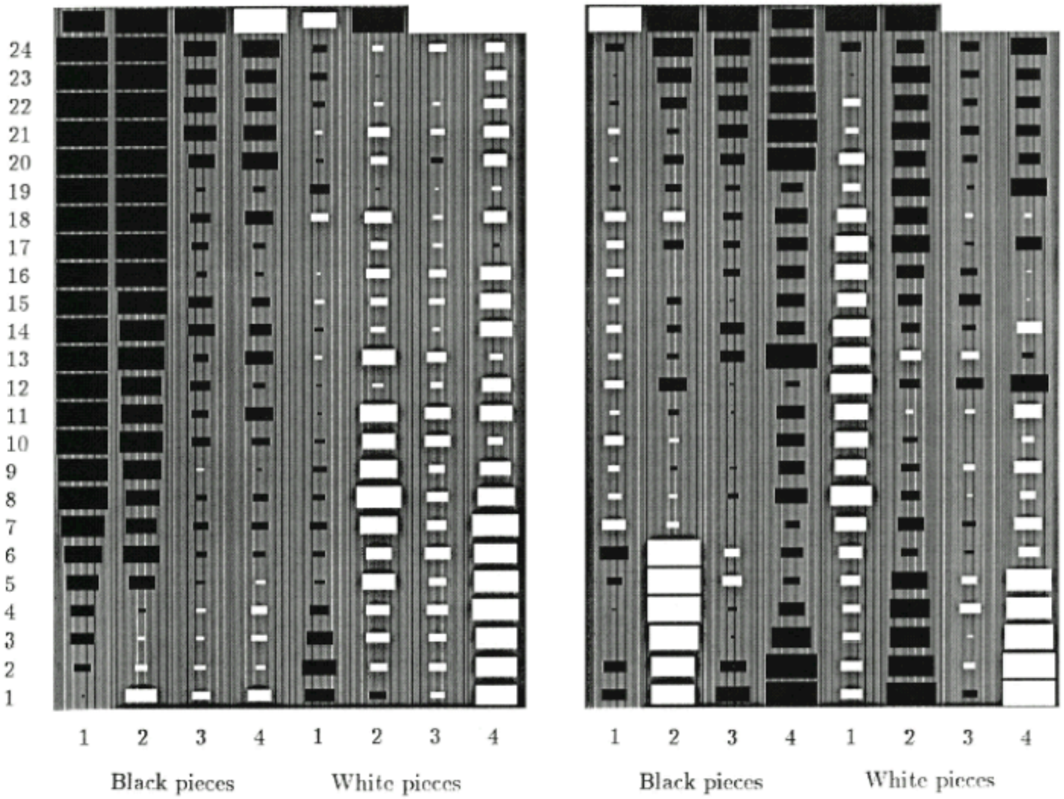
Coding the input



W	1	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
	2	●	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	●
	3	○	○	○	○	○	○	○	○	○	○	○	○	○	○	●	○	○	○	○	○	○	○	○	○
	4	○	○	○	○	○	○	○	○	○	○	○	●	○	○	○	○	○	○	●	○	○	○	○	○
B	1	○	○	○	○	○	○	○	○	○	○	○	○	●	○	○	○	○	○	○	○	○	○	○	○
	2	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
	3	○	○	○	○	○	○	○	○	○	●	○	○	○	○	○	○	○	○	○	○	○	○	○	○
	4	○	○	○	○	○	●	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
		1																							24

Backgammon played with RL and Backpropagation


W
L
Board Position and Move


Black pieces
White pieces
Black pieces
White pieces