# Model-Based Reinforcement Learning

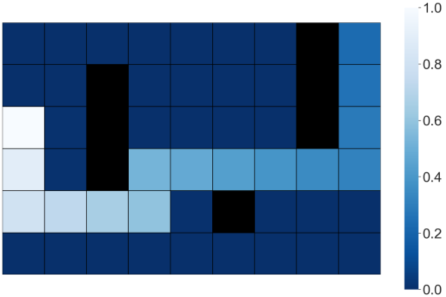
## Eligibility traces and Dyna architectures

Olivier Sigaud

Sorbonne Université
http://people.isir.upmc.fr/sigaud

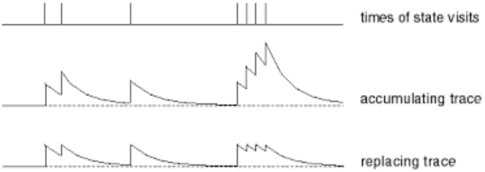
# DYNA approaches

## Eligibility traces: the replay idea




- Goal: improve over Q-LEARNING
- Naive approach: store all $(s, a)$ pair and back-propagate values
- Limited to finite horizon trajectories
- Converges faster, but needs more memory
- Speed of convergence versus memory trade-off

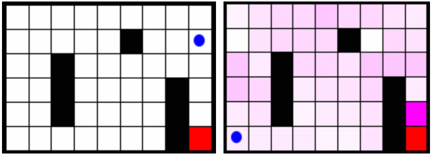
## Eligibility traces: efficient implementation




- TD($\lambda$), SARSA ($\lambda$) and Q($\lambda$): more sophisticated approach to deal with infinite horizon trajectories
- A variable $e(s)$ is reinitialized to 1 each time $s$ is visited again and decayed with a factor $\lambda$ after each time step
- $e_{t+1}(s) = \lambda e_t(s)$
- TD($\lambda$): $V(s) \leftarrow V(s) + \alpha e(s)\delta$, (similar for SARSA ($\lambda$) and Q($\lambda$)),
- The most recently visited states are updated more
- If $\lambda = 0$, $e(s)$ goes to 0 immediately, thus we get TD=TD(0), SARSA or Q-LEARNING
- If $\lambda = 1$, $e(s)$ does not decay, we get Monte Carlo updates...

Singh, S. P. and Sutton, R. S. (1996) Reinforcement learning with replacing eligibility traces. *Machine learning*, 22:123–158

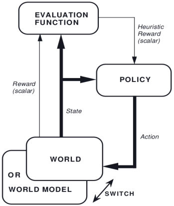
## Model-based Reinforcement Learning: general approach

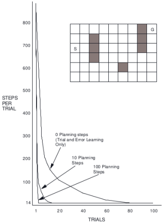


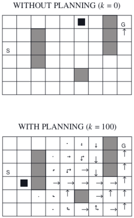
- Make profit of a learnt model of $T$ and $r$
- Learning $T$ and $r$ is an incremental self-supervised learning problem
- Central phenomenon in MBRL: model acquisition is driven by the current policy
- If the policy is stuck, the model will not improve

Sutton, R. S. (1991) DYNA, an integrated architecture for learning, planning and reacting. *SIGART Bulletin*, 2:160–163

# The Dyna-AHC family



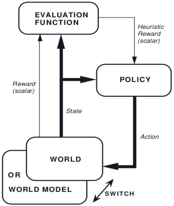



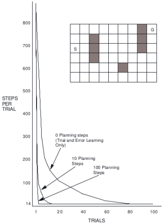

- AHC stands for Adaptive Heuristic Critic
- Interleaves back-ups in the model and in the environment
- Thanks to the model of transitions, DYNA can propagate values more often
- Fun fact: two nearly identical versions: ICML and NIPS

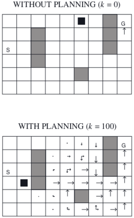
Sutton, R. S. (1990) Integrating architectures for learning, planning, and reacting based on approximating dynamic programming. In *Proceedings of the Seventh International Conference on Machine Learning*, pages 216–224, San Mateo, CA. Morgan Kaufmann

Sutton, R. S. (1990) Integrated modeling and control based on reinforcement learning and dynamic programming. *Advances in neural information processing systems*, 3

## Dyna-PI, Dyna-Q, Dyna-AC



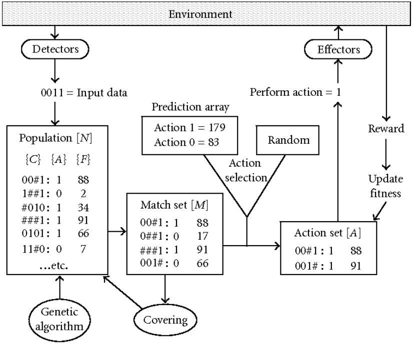




- Several approaches:
  - Draw random transitions in the model ("in the agent's head") and apply TD back-ups: DYNA-Q, DYNA-AC
  - Also called learning in imagination
  - Do the same along Monte Carlo trajectories: DYNA-PI
  - Better propagation: Prioritized Sweeping (see later)
  - Other point: model generalization

Sutton, R. S. (1990) Integrating architectures for learning, planning, and reacting based on approximating dynamic programming. In *Proceedings of the Seventh International Conference on Machine Learning*, pages 216–224, San Mateo, CA. Morgan Kaufmann

Sutton, R. S. (1990) Integrated modeling and control based on reinforcement learning and dynamic programming. *Advances in neural information processing systems*, 3

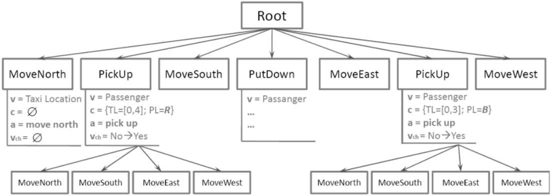
## Generalization 1: Anticipatory Learning Classifier Systems




- Problem: in the stochastic case, the model of transitions is in $card(S) \times card(S) \times card(A)$
- Usefulness of compact models
- MACS: DYNA with generalisation
- Lots of heuristic processes

Gérard, P., Meyer, J.-A., & Sigaud, O. (2005) Combining latent learning with dynamic programming in MACS. *European Journal of Operational Research*, 160:614–637

# Generalization 2: Factored MDPs




- SPITI: DYNA with generalisation (Factored MDPs)
- Much better founded formalization
- Could address very large MDPs
- A bot at CounterStrike with boolean features

Degris, T., Sigaud, O., & Wuillemin, P.-H. (2006) Learning the Structure of Factored Markov Decision Processes in Reinforcement Learning Problems. *Proceedings of the 23rd International Conference on Machine Learning (ICML'2006)*, pages 257–264

Degris, T., Sigaud, O., and Wuillemin, P.-H. (2006).
Learning the Structure of Factored Markov Decision Processes in Reinforcement Learning Problems.
In *Proceedings of the 23rd International Conference on Machine Learning*, pages 257–264, CMU, Pennsylvania.

Gérard, P., Meyer, J.-A., and Sigaud, O. (2005).
Combining latent learning with dynamic programming in MACS.
*European Journal of Operational Research*, 160:614–637.

Singh, S. P. and Sutton, R. S. (1996).
Reinforcement learning with replacing eligibility traces.
*Machine learning*, 22:123–158.

Sutton, R. S. (1990a).
Integrated modeling and control based on reinforcement learning and dynamic programming.
*Advances in neural information processing systems*, 3.

Sutton, R. S. (1990b).
Integrating architectures for learning, planning, and reacting based on approximating dynamic programming.
In *Proceedings of the Seventh International Conference on Machine Learning*, pages 216–224, San Mateo, CA. Morgan Kaufmann.

Sutton, R. S. (1991).
DYNA, an integrated architecture for learning, planning and reacting.
*SIGART Bulletin*, 2:160–163.