# Learning with options: Just deliberate and relax

Pierre-Luc Bacon and Doina Precup

Reasoning and Learning Lab (RLLAB), McGill University

## Summary

- From the point of view of absolute optimality, temporal abstractions in reinforcement learning are not necessary.
- We propose bounded rationality as a lens through which we can describe the desiderata for constructing temporal abstractions.
- We formalize the idea that *good* options are those which result in fast planning (or inference).

## Options framework

Options (Richard S. Sutton, Precup, and Singh, 1999) formalize the idea of temporally extended actions (also sometimes called *skills* or *macro-actions*).

A Markovian option $\omega \in \Omega$ is a triple $\langle \mathcal{I}_\omega, \pi_\omega, \beta_\omega \rangle$:

- Initiation set $\mathcal{I}_\omega \subseteq \mathcal{S}$
- Policy $\pi_\omega$ (stochastic or deterministic)
- Termination function $\beta_\omega : \mathcal{S} \to [0, 1]$.

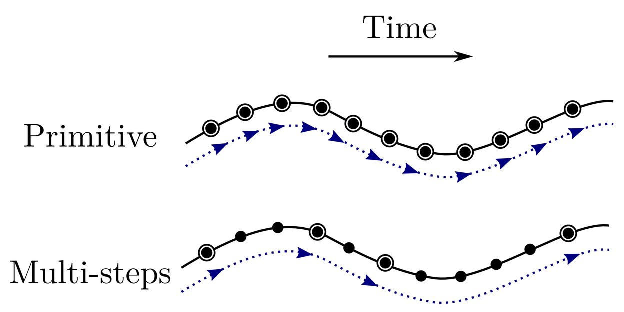
## Deliberation cost

We define the cost of a one-step backup for $Q^\star_\Omega(s, \omega)$:

$$c(s, \omega) = \sum_{s'} \mathbf{1}_{P(s' \mid s, \omega) > \epsilon} |\Omega(s')|$$

where $\epsilon \in [0, 1]$ is a constant that can be used to allow next states to be ignored (or, can be set to 0 if we want to take into account all successor states).

## Cost of a trajectory




In our model, there is no deliberation cost incurred within an option once initiated. During the option's execution, its policy will be in effect and choices do not require any deliberation.

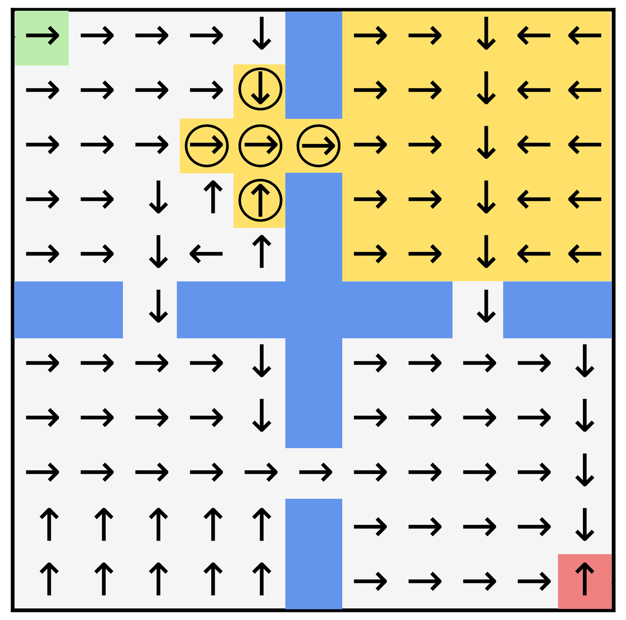
## Expected value of control

We define a joint objective which expresses the desire to seek reward under a reasonable deliberation (or *cognitive*) effort:

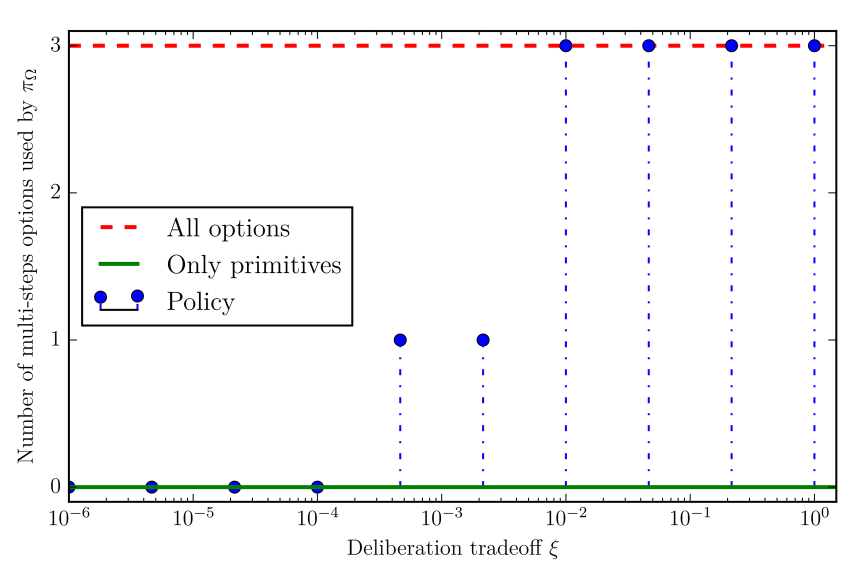
$$Q_{VC}(s, \omega) = Q_\Omega(s, \omega) + \xi Q_c(s, \omega)$$

where $\xi$ controls the trade-off between *value* and *computation cost*.

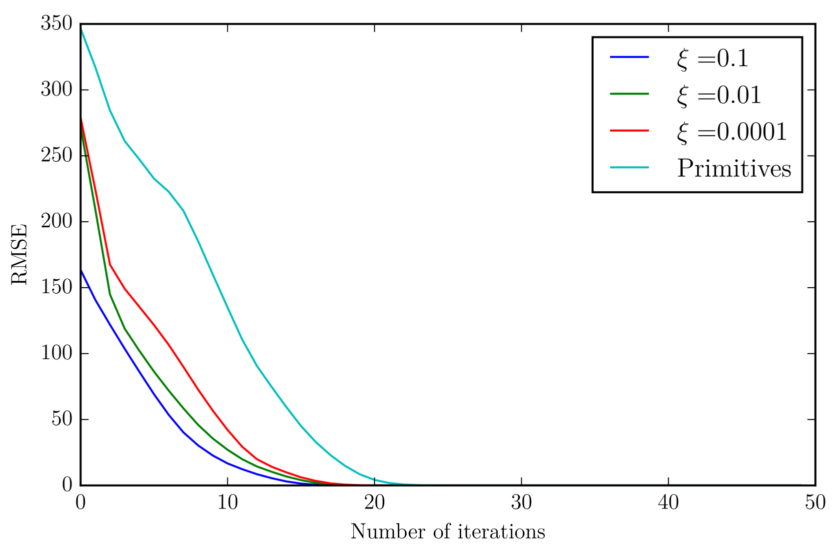
## Experiments



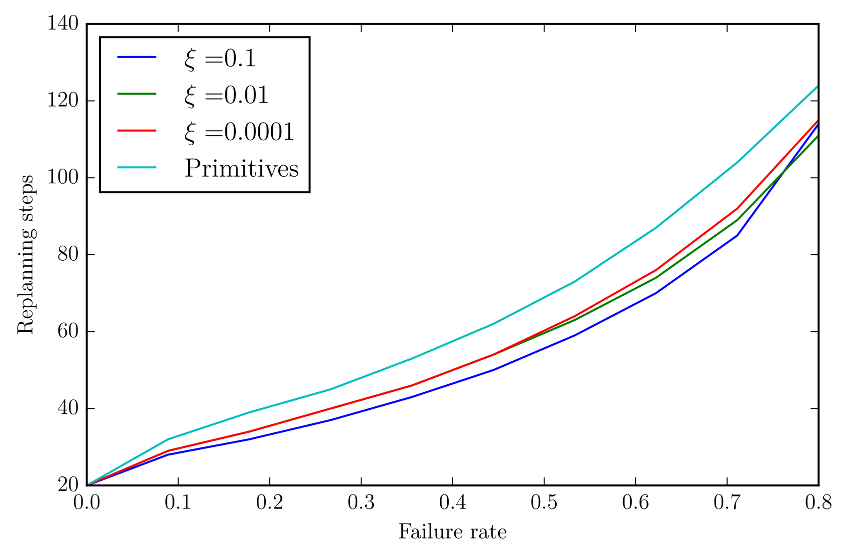
(a) Four-rooms domain (Richard S. Sutton, Precup, and Singh, 1999)




(b) Number of options recruited by $\pi_\Omega$ as a function of $\xi$




(c) The error in planning with regularized options decreases faster




(d) Replanning cost for different perturbation levels

We optimized a set of four options under the option-critic architecture (Bacon and Precup, 2015). For larger values of $\xi$ in the expected value of control, the policy over options uses more temporally extended actions (fig. 1b). The root mean square error in $V_\Omega$ decreases faster with a set of options optimized with a larger $\xi$ and leads to faster planning (fig. 1c). Stronger regularization also protects against perturbations in the MDP. Figure 1d) shows the number of replanning steps for different noise levels in the MDP.

## Unified view

- The deliberation concept subsumes dedicated cost functions for *switching* and *commitment*.
- Reiterates the idea that *simple* options (Maisto, Donnarumma, and Pezzulo, 2015) are preferable
- Low deliberation corresponds to *sparse* option models
- Smaller set of terminating states imples less variance in the sample backups (R. S. Sutton and Barto, 1998) (cheaper by definition)
- Sparse models (especially in linear form) are computationally cheaper than dense ones
- In a partially observable setting, sparse models would skip over regions of the state space with high uncertainty.
- Robustness: $Q_c(s, \omega)$ can be interpreted as the average replanning cost. $\xi$ controls the degree of robustness against perturbations.

## Future work

- Study the relationship with the successor state representation of Dayan, 1993
- Learn initiation sets for options: we need to be able to "chain" options together
- Model-free setting: a new definition for the deliberation cost might be necessary
- Interplay of our deliberation cost with value function approximation

## References

- R. S. Sutton and A. G. Barto (1998). *Introduction to Reinforcement Learning*. MIT Press. ISBN: 0262193981
- Richard S. Sutton, Doina Precup, and Satinder P. Singh (1999). "Between MDPs and Semi-MDPs: A Framework for Temporal Abstraction in Reinforcement Learning". In: *Artif. Intell.* 112.1-2, pp. 181–211
- Pierre-Luc Bacon and Doina Precup (2015). "The option-critic architecture". In: *NIPS Deep Reinforcement Learning Workshop*
- D. Maisto, F. Donnarumma, and G. Pezzulo (2015). "Divide et impera: subgoaling reduces the complexity of probabilistic inference and problem solving". In: *Journal of The Royal Society Interface* 12.104
- P. Dayan (1993). "Improving generalisation for temporal difference learning: The successor representation". In: *Neural Computation* 5, pp. 613–624