Learning and Planning with Timing Information in Markov Decision Processes

Pierre-Luc Bacon, Borja Balle, Doina Precup
Reasoning and Learning Lab, McGill University

Find out more in our UAI 2015 paper or at cs.mcgill.ca/~pbacon/timing

McGill

Highlights

We consider the problems of learning and planning in Markov decision processes with temporally extended actions represented in the options framework (Sutton et al., 1999).

- We propose to use predictions about the duration of extended actions to represent the state.
- We develop a consistent and efficient spectral learning algorithm.
- We show how such timing features can be used for planning.

Motivation

Timing information is very cheap to measure and process, even with simple hardware.

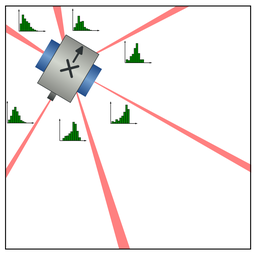

Figure : Imagine that a robot has models for options that move radially out from the current position, this would allow localizing with respect to all neighbouring walls.

Timing models can be advantegeous for resource-constrained devices:

- Roomba-like robots (as in our experiments)
- Cellphones (on which using a lot of sensors or computation drains the battery)

Building full models might also be too data and computation-hungry:

- Investment problems, depending on long histories or news
- Robotics, with sensors producing too much data for real-time processing

Being able to exploit a simpler model is important.

Options Duration Model (ODM)

Instead of predicting a full model at the end of an option, we only consider when an **option will terminate** given the history. We have a dynamical system with observations from $\Omega \times \{\sharp, \perp\}$, where $\sharp$ (*sharp*) denotes **continuation** and $\perp$ (*bottom*) is for **termination**. Trajectories are of the form:

$$(\omega_1, \sharp, \ldots, \omega_1, \sharp, \omega_1, \perp, \omega_2, \sharp, \ldots, \omega_2, \sharp, \omega_2, \perp, \ldots) = (\omega_1, \sharp)^{d_1 - 1}(\omega_1, \perp)(\omega_2, \sharp)^{d_2 - 1}(\omega_2, \perp)\ldots$$

Representing ODM with Predictive State Representation (PSR)

Let $\delta(\alpha, \omega)$ be a random variable representing the duration of option ω when started from $s \sim \alpha$. The probability of a sequence of options $\bar{\omega} = \omega_1 \cdots \omega_t$ and their durations $\bar{d} = d_1 \cdots d_t$, $d_i > 0$ is given by:

$$\mathbb{P}[\bar{d}|\alpha, \bar{\omega}] = \boldsymbol{\alpha}^\top \mathbf{A}_{\omega_1,\sharp}^{d_1 - 1} \mathbf{A}_{\omega_1,\perp} \mathbf{A}_{\omega_2,\sharp}^{d_2 - 1} \mathbf{A}_{\omega_2,\perp} \cdots \mathbf{A}_{\omega_t,\sharp}^{d_t - 1} \mathbf{A}_{\omega_t,\perp} \mathbf{1} \ .$$

$$\mathbf{A}_{\omega,\sharp}(s, s') = \sum_{a \in A} \pi(s, a) P(s, a, s')(1 - \beta(s')), \quad \mathbf{A}_{\omega,\perp}(s, s') = \sum_{a \in A} \pi(s, a) P(s, a, s')\beta(s'), \quad \boldsymbol{\alpha} = \mathbb{I}[s = s_0]$$

Theorem 1: Existence of an ODM for an MDP

Let M be an MDP with n states, Ω a set of options, and $\Sigma = \Omega \times \{\sharp, \perp\}$. For every distribution α over the states of M, there exists a PSR $\mathcal{A} = \langle \boldsymbol{\alpha}, \mathbf{1}, \{\mathbf{A}_\sigma\} \rangle$ with at most n states that computes the distributions over durations of options executed from a state sampled according to α.

Learning ODM

The probabilities over sequences in our embedded system can be summarized in a **Hankel matrix**, a bi-infinite (conceptually) matrix $\mathbf{H}_f \in \mathbb{R}^{\Sigma^\star \times \Sigma^\star}$ with rows and columns indexed by strings in $\Sigma^\star$. We estimate the Hankel matrix for a fixed finite subsets of rows and columns. The SVD decomposition of $\mathbf{H}_f$ provides a way to recover a PSR (see Boots et al., 2011).

Planning with ODM

We consider PSR states obtained by the state-update procedure. Given a valid trajectory $u \in V$, the updated state is $\boldsymbol{\theta}_{\alpha,u}^\top = \frac{\boldsymbol{\theta}_\alpha^\top \mathbf{A}_u}{\boldsymbol{\theta}_\alpha^\top \mathbf{A}_u \boldsymbol{\alpha}_\infty}$.

Theorem 3: Linearity of the state-option value function

Let $\pi_\Omega : S \times \Omega \to [0, 1]$ be a stochastic stationary policy over options on the MDP M. For every $\omega \in \Omega$ there exists a vector $\boldsymbol{\rho}_\omega \in \mathbb{R}^{n'}$ so that for every distribution α over states in M and every history $u \in V$, we have $\mathbb{E}_s[Q^{\pi_\Omega}(s, \omega)] = \boldsymbol{\theta}_{\alpha,u}^\top \boldsymbol{\rho}_\omega$, where the expectation is over states s sampled from the distribution induced by observing u after starting in a state drawn from α.

Experiments

We considered both discrete and continuous nagivation tasks where the options allow the agent to move radially until it hits a wall. We used Fitted-Q iteration (Ersnt et al, 2005) for planning over ODM states.

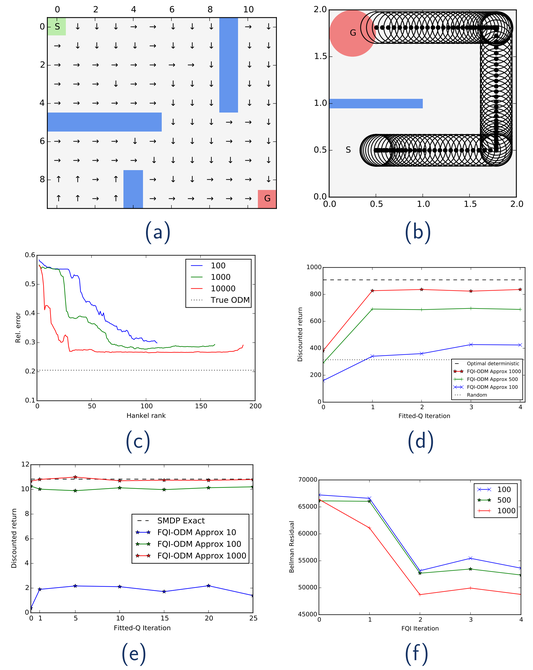

Figure : **(a)** grid layout and optimal policy over options **(b)** continuous navigation environment **(c)** relative error in prediction vs rank **(d)** average discounted cumulative return (continuous task) **(e)** average discounted cumulative return (discrete task) **(f)** mean square Bellman residual (continuous task)

Conclusion

We presented an approach to learn a predictive model for option durations that is useful for planning. Timing models get around the problems of:

- large action spaces (by using a finite set of options)
- large observation spaces (by focusing only on continuation and termination).

A theoretical analysis that fully characterizes the error of planning with timing models instead of true transition models is left for future work.