A Matrix Splitting Perspective on Planning with Options

Pierre-Luc Bacon and Doina Precup

Reasoning and Learning Lab (RLLAB), McGill University

Summary

- We show that when planning with options, the corresponding Bellman operator involves a matrix splitting (Varga 1962).
- Equivalently, a set of options and a policy over them is shown to specify a matrix preconditioner.
- A choice of options is therefore a choice of a preconditioned fixed point iteration algorithm.

Options Framework

A Markovian option (Sutton, Precup, and Singh 1999) $w \in \mathcal{W}$ is a triple $(\mathcal{I}_w, \pi_w, \beta_w)$ where:

- $\mathcal{I}_w \subseteq \mathcal{S}$ is the initiation set
- π_w is a policy
- $\beta_w : \mathcal{S} \to [0, 1]$ is a termination function.

The policy over options is $\mu : \mathcal{S} \to (\mathcal{W} \to [0, 1])$.

Generalized Bellman Operator

Value iteration typically propagates values for one time step only. However, multi-steps backups are also possible (Sutton 1995). We consider the following generalized Bellman operator in which the number of backups K is a random variable:

$$(Lv)(s) \doteq \mathbb{E}\left[\sum_{k=0}^{K-1} \gamma^k r(S_k, A_k) + \gamma^K v(S_K) \,\middle|\, S_0 = s\right]$$

Linear Representation of L and Options Models

We now assume that the number of backups K in L is controlled by the termination functions of a set of Markovian options. By linearity, we can decompose the generalized Bellman operator in a *reward model* b and a *transition model* F:

$$b \doteq (I - \gamma H)^{-1} r_\sigma, \qquad \text{and} \qquad F \doteq \gamma (I - \gamma H)^{-1}(P_\sigma - H) \ ,$$

where

$$\sigma(a \mid s) \doteq \sum_w \mu(w \mid s) \pi_w(a \mid s) \ ,$$

and

$$H(s, s') \doteq \sum_w \mu(w \mid s) \sum_a \pi_w(a \mid s) P(s' \mid s, a)(1 - \beta_w(s')) \ .$$

The generalized Bellman operator L then becomes:

$$Lv = b + Fv = (I - \gamma H)^{-1} r_\sigma + \gamma (I - \gamma H)^{-1}(P_\sigma - H)v \ .$$

The Preconditioning Effect of Options

The generalized Bellman equations can also be written as:

$$v = v + (I - \gamma H)^{-1}(r_\sigma - (I - \gamma P_\sigma)v)$$

Options therefore yield the following preconditioned linear system:

$$(I - \gamma H)^{-1}(I - \gamma P_\sigma)v = (I - \gamma H)^{-1} r_\sigma \ .$$

A *good* set of options is therefore one for which M *close* to $I - \gamma P_\sigma$ but whose inverse M^{-1} is easier to compute.

A Family of Successive Approximation Methods

Options specify a family of successive approximation methods for solving Markov Decision Processes containing the following two extreme members:

1. $L^{(\infty)} v \doteq (I - \gamma P_\sigma)^{-1} r_\sigma$, when options always continue.
2. $L^{(0)} v \doteq r_\sigma + \gamma P_\sigma v$, when options always stop.

Theorem 1 : Options Induce a Regular Splitting

Let $A \doteq I - \gamma P_\sigma$, $M \doteq I - \gamma H$ and $N \doteq \gamma(P_\sigma - H)$, then $A = M - N$ is a regular splitting.

Corollary 1 : Convergence

For the regular splitting of theorem 1,

1. *The spectral radius of the iteration matrix is $\rho(\gamma(I - \gamma H)^{-1}(P_\sigma - H)) < 1$*
2. *The successive approximation method based on the generalized Bellman operator L converges for any initial vector v_0.*

Theorem 2 : Consistency

The iterative method

$$v_{k+1} = (I - \gamma H)^{-1} r_\sigma + \gamma (I - \gamma H)^{-1}(P_\sigma - H)v_k, \ \ k \geq 0$$

associated with the matrix splitting is a consistent policy evaluation method if the set of options and policy over them is such that

$$\sigma(a \mid s) = \sum_w \mu(w \mid s) \pi_w(a \mid s) \, \forall a \in \mathcal{A}, s \in \mathcal{S}$$

where σ is the target policy to be evaluated.

Theorem 3 : Predict Further, Plan Faster

If a set of options $\widetilde{\mathcal{W}}$ has the same intra-option policies and policy over options with some other set $\mathcal{W}$ but whose termination functions are such that $\beta_{\tilde{w}}(s) \leq \beta_w(s) \, \forall w \in \mathcal{W}, s \in \mathcal{S}$, then:
$0 \leq \rho(\widetilde{M}^{-1}\widetilde{N}) \leq \rho(M^{-1}N) < 1$.

Implications

- We now have formal framework to define what *good* options are.
- We can compare options through the splitting that they induce.
- It opens up new opportunities for options discovery.
- The idea of transfer learning with options is natural:
  - The preconditioner M can be reused for different RHS.
- Preconditioning can also regularize ill-conditioned linear systems:
  - Options for off-policy learning
  - Options to deal with partial observability, feature aliasing

References

- Richard S. Varga. *Matrix iterative analysis.* Prentice-Hall, 1962
- Richard S. Sutton. "TD Models: Modeling the World at a Mixture of Time Scales". In: *ICML*. 1995
- Richard S. Sutton, Doina Precup, and Satinder P. Singh. "Between MDPs and Semi-MDPs: A Framework for Temporal Abstraction in Reinforcement Learning". In: *Artif. Intell.* 112.1-2 (1999), pp. 181–211