Approximate Value Iteration with Temporally Extended Actions (Extended Abstract)*

Timothy A. Mann
DeepMind,
London, UK
timothymann@google.com

Shie Mannor
The Technion,
Haifa, Israel
shie@ee.technion.ac.il

Doina Precup
McGill University,
Montreal, Canada
dprecup@cs.mcgill.ca

Abstract

The options framework provides a concrete way to implement and reason about temporally extended actions. Existing literature has demonstrated the value of planning with options empirically, but there is a lack of theoretical analysis formalizing when planning with options is more efficient than planning with primitive actions. We provide a general analysis of the convergence rate of a popular Approximate Value Iteration (AVI) algorithm called Fitted Value Iteration (FVI) with options. Our analysis reveals that longer duration options and a pessimistic estimate of the value function both lead to faster convergence. Furthermore, options can improve convergence even when they are suboptimal and sparsely distributed throughout the state space. Next we consider generating useful options for planning based on a subset of landmark states. This suggests a new algorithm, Landmark-based AVI (LAVI), that represents the value function only at landmark states. We analyze OFVI and LAVI using the proposed landmark-based options and compare the two algorithms. Our theoretical and experimental results demonstrate that options can play an important role in AVI by decreasing approximation error and inducing fast convergence.

1 Introduction

We consider planning in Markov Decision Processes (MDPs; [Puterman, 1994], see Section 2) with large or infinite state-spaces. Traditional planning algorithms, such as Value Iteration (VI) and Policy Iteration (PI), are intractable in this setting because the computational and memory complexities at each iteration scale (polynomially and linearly, respectively; [Littman *et al.*, 1995]) with the number of states in the target MDP. Approximate Value Iteration (AVI) algorithms are more scalable than VI, because they compactly represent the value function [Bertsekas and Tsitsiklis, 1996]. This allows AVI algorithms to achieve per iteration computational and memory complexities that are independent of the size of the state-space. However, there are many challenges to using AVI algorithms in practice. AVI and VI often need many iterations to solve the MDP [Munos and Szepesvári, 2008]. It turns out that temporally extended actions can play an important role in reducing the number of iterations.

Options are a unified abstraction for representing both temporally extended actions and primitive actions [Sutton *et al.*, 1999]. Options provide a valuable tool for efficient planning [Sutton *et al.*, 1999; Silver and Ciosek, 2012]. Under most analyses of AVI, one iteration corresponds to planning one timestep into the future. On the other hand, performing a single iteration of AVI with temporally extended actions, one iteration could instead correspond to planning several timesteps into the future. We derive bounds that help us reason about when AVI with temporally extended actions converges faster than AVI with only primitive actions.

Motivation: We focus on analyzing the convergence rate of AVI with options, because a faster convergence rate implies a solution with fewer iterations. Using the convergence rate we can determine the total computational cost of planning by bounding the computational cost at each iteration. If the total computational cost with options is smaller than with primitive actions, planning with options is faster than planning with primitive actions.

Contributions: The main contributions of this paper are the following: (1) We analyze Options Fitted Value Iteration (OFVI) in Theorem 1, characterizing the asymptotic loss and the convergence behavior of planning with a given set of options. (2) We analyze the asymptotic loss and convergence behavior of Landmark-based Approximate Value Iteration (LAVI) in Theorem 2. Comparing the bounds of LAVI and OFVI suggests that LAVI may converge faster than OFVI. However, their asymptotic losses are not directly comparable. (3) Our experimental comparison in a complex inventory management problem demonstrates that LAVI achieves a favorable performance versus time trade-off.

2 Background

An MDP is defined by $\langle X, A, P, R, \gamma \rangle$ [Puterman, 1994] where X is a set of states, A is a finite set of primitive actions, P maps from state-action pairs to probability distributions over states, R is a mapping from state-action pairs to reward distributions bound to the interval $[-R_{\max}, R_{\max}]$, and $\gamma \in [0, 1)$ is a discount factor. Let $B(X; V_{\max})$ denote

*This paper is an extended abstract of an article in the Journal of Artificial Intelligence Research [Mann *et al.,* 2015].

the set of functions with domain X and range bounded by $[-V_{\text{MAX}}, V_{\text{MAX}}]$ where $V_{\text{MAX}} \leq \frac{R_{\text{MAX}}}{1-\gamma}$.

A policy $\pi : X \to A$ is a mapping from states to actions. We denote the value function for π by V^π, the Bellman operator by $\mathcal{T}^\pi$, and the Bellman optimality operator by $\mathcal{T}$. Value Iteration (VI) is defined by repeatedly applying $\mathcal{T}$. The algorithm produces a series of value function estimates $V_0, V_1, V_2, \ldots, V_K$ and the greedy policy π_K is constructed based on V_K.

Approximate Value Iteration (AVI): iteratively produces a sequence of $K \geq 1$ approximations $\{V_k\}_{k=1}^K$ of the optimal value function and returns a greedy policy π_K with respect to the final iterate V_K.

Semi-Markov Decision Processes (SMDPs): are a generalization of the Markov Decision Process (MDP) model that incorporates temporally extended actions. For each state $x \in X$, we denote the set of options [Sutton *et al.*, 1999] that can be initialized from x by $\mathcal{O}_x = \{o \in \mathcal{O} \mid x \in \mathcal{I}_o\}$. Options encompass, not only primitive actions and temporally extended actions, but also stationary policies and other control structures. Here we take actions to be options that always terminate after a finite number of timesteps.

For an option $o = \langle \mathcal{I}_o, \pi_o, \beta_o \rangle$, we denote the probability that o is initialized from a state x and terminates in a subset of states $Y \subseteq X$ in exactly t timesteps by $P_t^o(Y|x)$. We denote the SMDP Bellman operator by $\mathbb{T}$.

3 AVI Algorithms

Options Fitted Value Iteration (OFVI): is a generalization of the multisample FVI algorithm to the case where samples are generated by options. We refer to the special case of OFVI with primitive actions as PFVI. The algorith takes as arguments positive integers n, m, K, μ a sampling distribution, an initial value function estimate V_0, and a simulator $\mathbb{S}$. At each iteration $k = 1, 2, \ldots, K$, states $x_i \sim \mu$ for $i = 1, 2, \ldots, n$ are sampled, and for each option $o \in \mathcal{O}_{x_i}$, m next states, rewards, and option execution times $\langle y_{i,j}^o, r_{i,j}^o, \tau_{i,j}^o \rangle \sim \mathbb{S}(x_i, o)$ are sampled for $j = 1, 2, \ldots, m$. Then the update resulting from applying the Bellman operator to the previous iterate V_{k-1} is estimated by $\hat{V}_k(x_i) \leftarrow \max_{o \in \mathcal{O}_{x_i}} \frac{1}{m} \sum_{j=1}^m \left[r_{i,j}^o + \gamma^{\tau_{i,j}^o} V_{k-1}(y_{i,j}^o) \right]$, and we apply a supervised learning algorithm to obtain the best fit. The given simulator $\mathbb{S}$ differs from the simulator for PFVI. It returns the state where the option returned control to the agent, the total cumulative, discounted reward received during execution, and the number of timesteps that the option executed.

Landmark-based AVI: One limitation of planning with options is that options typically need to be designed by an expert. In this section, we consider an approach similar in spirit to the successful FF-Replan algorithm [Yoon *et al.*, 2007], which plans on a deterministic projection of the target MDP. The algorithm replans whenever the agent enters a state not in the current plan. Unlike FF-Replan, our approach is more scalable as it does not plan over the entire problem. We argue for using landmark-based options where the designer specifies a collection of landmark regions and creates options by planning a path from one landmark region to another. For this algorithm, we only maintain value estimates for a finite set of landmark points.

Landmark-based OFVI: It is also possible to consider using landmark-based options with OFVI. We refer to the resulting case of the algorithm as Landmark-based Options Fitted Value Iteration (LOFVI).

OFVI Analysis

Our approach is based on a contraction mapping argument. By applying the MDP Bellman operator $\mathcal{T}$ to V, we obtain a contraction mapping where γ (the discount factor) serves as the contraction coefficient. Smaller values of γ imply a faster convergence rate, but γ is part of the problem description and cannot be changed. However, if we apply $\mathcal{T}$, $\tau > 1$ times, we obtain a contraction mapping where the contraction coefficient is $\gamma^\tau < \gamma$. Temporally extended options have a similar effect. Options can speed up the convergence rate by inducing a smaller contraction coefficient.

Options with a long duration are desirable for planning because options that execute for many timesteps enable OFVI to look far into the future during a single iteration. However, the duration depends on both an option and the state where the option is initialized. We denote by $D_{x,Y}^o$ the random variable representing the duration of initializing option o from state x and terminating in $Y \subseteq X$. For a set of options $\mathcal{O}$, we define the **minimum duration** to be $d_{\min} = \min_{x \in X, o \in \mathcal{O}_x} \inf_{Y \subseteq X} \mathbb{E}\left[D_{x,Y}^o\right]$.

The duration of an option is a random variable that depends on the state where the option was initialized. This complicates the analysis compared to assuming that all temporally extended actions terminate after a fixed number of timesteps, but it allows for much greater flexibility when selecting options to use for planning.

Similar to the analysis of PFVI, the analysis of OFVI depends on the concentrability of future state distributions. We denote this coefficient by $\mathbb{C}_{\nu,\mu}$ and assume it is finite. This assumption is analogous to the concentrability coefficient assumption from [Munos and Szepesvári, 2008]. Despite the fact that options are a more general framework than the set of primitive actions, the concentrability coefficient for temporally extended actions is smaller than the coefficient for primitive actions.

The important properties of temporally extended actions that cause faster convergence are (1) the quality of the policy they follow, and (2) how long the action executes for (or its duration). The following definition describes the set of states where there exists an option that follows a near-optimal policy and has sufficient duration.

We define the set $\omega_{\alpha,d}$ to be states with particularly long duration and follow an α-optimal policy. However, the states outside of $\omega_{\alpha,d}$ do not. At these other states, either the available options are not sufficiently temporally extended or they follow a suboptimal policy. To obtain faster convergence, we need a way of connecting the convergence rates of the states outside of $\omega_{\alpha,d}$ with the convergence rates of the states in $\omega_{\alpha,d}$.

Assumption 1. $[A1(\alpha, d, \psi, \nu, j)]$ *Let* $\alpha, \psi, j \geq 0$, $d \geq d_{\min}$, *and* $\nu \in M(X)$. *For any* $m \geq 0$ *option*

policies $\varphi_1, \varphi_2, \ldots, \varphi_m$*, let* $\rho = \nu P^{\varphi_1} P^{\varphi_2} \ldots P^{\varphi_m}$*. There exists an* α*-optimal option policy* $\hat{\varphi}$ *such that either (1)* $\Pr_{x\sim\rho}[x \in \omega_{\alpha,d}] \geq 1 - \psi$ *or (2)* $\exists_{i\in\{1,2,\ldots,j\}} \Pr_{y\sim\eta_i}[y \in \omega_{\alpha,d}] \geq 1 - \psi$ *where* $\eta_i = \nu P^{\varphi_1} P^{\varphi_2} \ldots P^{\varphi_m} \left(P^{\hat{\varphi}}\right)^i$ *for* $i = 1, 2, \ldots, j$.

Assumption 1 points out three key features that impact planning performance with options: (1) Quality of the option set controlled by α, (2) Duration of options specified by d, and (3) Sparsity of $\omega_{\alpha,d}$ characterized by j and ψ.

We call $\hat{\varphi}$ the "bridge" policy, because it bridges states in $\omega_{\alpha,d}$ and other states. Notice that we do not assume that the planner has any knowledge of $\hat{\varphi}$. It is enough that such a policy exists. Assumption 1 says that no matter what policies are followed, either (1) the agent will end up in $\omega_{\alpha,d}$ with high probability or (2) there exists a near-optimal option policy that will transport the agent to $\omega_{\alpha,d}$ in at most j timesteps with high probability. This enables us to account for problems where only a few states have temporally extended actions, but these states can be reached quickly without following a policy that is too suboptimal.

The following theorem provides a comprehensive description of the convergence behavior of OFVI (with PFVI as a special case where $\mathcal{O} = A$).

Theorem 1. *Let* $\varepsilon_S, \delta > 0$*,* $\alpha, \psi, j \geq 0$*,* $K, p \geq 1$*,* $d \geq d_{\min}$*,* $0 \leq Z \leq K$*, and* $\nu, \mu \in M(X)$*. Suppose that* $A1(\alpha, d, \psi, \nu, j)$ *(Assumption 1) holds and* $\mathbb{C}_{\nu,\mu} < \infty$*. Given* $V_0 \in B(X, V_{\max})$*, if the first* Z *iterates* $\{V_k\}_{k=0}^{Z}$ *produced by the algorithm are pessimistic (i.e.,* $V_k(x) \leq V^{\Phi^*}(x)$ *for all* $x \in X$*), then there exists positive integers* n *and* m *such that when OFVI is executed,*

$$\begin{aligned}\|V^* - V^{\varphi_K}\|_{p,\nu} &\leq \|V^* - V^{\Phi^*}\|_{p,\nu} \\ &+ \frac{2\gamma^{d_{\min}}}{(1-\gamma)^2} \mathbb{C}_{\nu,\mu}^{1/p} \left(b_{p,\mu}(\mathbb{T}\mathcal{F}, \mathcal{F}) + \alpha\right) + \varepsilon_S \\ &+ \left(\gamma^{d_{\min}(K+1) + (1-\psi)(d-d_{\min})\lfloor Z/\hat{j}\rfloor}\right)^{1/p} \left(\frac{2\left\|V^{\Phi^*} - V_0\right\|_\infty}{(1-\gamma)^2}\right) \end{aligned} \quad (1)$$

holds with probability at least $1 - \delta$ *where* Φ^* *is the optimal option policy with respect to the given options* $\mathcal{O}$ *and* $\hat{j} = j + 1$.

Theorem 1 bounds the loss of the option policy φ_K returned after performing $K \geq 1$ iterations of value iteration with respect to a (p, ν)-norm. The distribution ν can be thought of as an initial state distribution. It places more probability mass on the regions of the state space where we want the policy φ_K to have the best performance. The value of $p \geq 1$ is generally determined by the function approximation procedure. For $p = 1$, the function approximation procedure minimizes the L_1-norm and for $p = 2$, the function approximation procedure minimizes the L_2-norm.

The right hand side of (1) contains four terms.

1. The first term bounds the abstraction loss, which is the loss between the optimal policy over primitive actions and the optimal option policy.
2. The second term bounds the approximation error, which is the error caused by the inability of the function approximation architecture to exactly fit $\hat{V}_k(x_i)$ during each iteration and α which shows up in this term is due to bootstrapping off options that follow α-optimal policies to gain faster convergence. Notice that $\frac{\gamma^{d_{\min}}}{(1-\gamma)^2}$ shrinks as $d_{\min}$ grows. Thus option sets with longer minimum duration shrink the approximation error.
3. The third term ε_S is the sample error, which is controlled by the number of samples taken at each iteration.
4. The last term controls the convergence error. Notice that γ, the discount factor, is in $[0, 1)$ and therefore the last term shrinks rapidly as its exponent grows. While OFVI does not actually converge in the sense that the loss may never go to zero, this last term goes to zero as $K \to \infty$. In the worst case, the convergence rate is controlled by $\gamma^{d_{\min}(K+1)}$, but the convergence rate can be significantly faster if Z and d are large and j is small.

An iterate $\hat{V} : X \to \mathbb{R}$ is pessimistic if $\forall_{y\in X}\ \hat{V}(y) \leq V^{\Phi^*}(y)$, where Φ^* is the optimal option policy. Whether iterates are pessimistic (or not) has a critical impact on the convergence rate of OFVI. To understand why, suppose that $q \in \mathcal{O}_x$ is an option that can be initialized from a state $x \in X$ where q is α-optimal with respect to Φ^* (i.e., $Q^{\Phi^*}(x, q) \geq V^{\Phi^*}(x) - \alpha$) and has a long duration (at least d timesteps). Since $\mathbb{T}$ is known to be a monotone operator, $V^{\Phi^*}(x) \geq (\mathbb{T}\hat{V})(x)$, even if an option other than q was selected for the update, $(\mathbb{T}\hat{V})(x)$ is at least as close to $V^{\Phi^*}(x)$ as if q was selected. This allows us to prove that when the iterates are pessimistic, the convergence rate of OFVI is rapid (depending on d). Unfortunately, when the iterates are not pessimistic, this reasoning no longer holds and convergence may depend on options with duration $d_{\min}$ instead.

Using Theorem 1 it is possible to prove convergence rates of OFVI on a wide range of planning problems. See the corollaries to Theorem 1 in [Mann *et al.*, 2015] for more details.

LAVI Analysis

We provide a theoretical analysis of LAVI along two dimensions. (1) We bound the loss associated with policies returned by LAVI compared to the optimal policy over primitive actions, and (2) we analyze the convergence rate of LAVI.

Theorem 2. *(LAVI Convergence) Let* $\varepsilon_S > 0, \delta \in (0, 1]$*. There exists* $m = O\left(\frac{1}{(\varepsilon_S(1-\gamma)^2(1-\gamma^{d_{\min}}))^2} \ln\left(\frac{LK}{\delta}\right)\right)$ *such that with probability greater than* $1 - \delta$*, if LAVI is executed for* $K \geq 1$ *iterations, the greedy policy* φ_K *derived from* V_K *satisfies*

$$\begin{aligned}\|V^* - V^{\varphi_K}\|_{1,\nu} \leq &\left(\frac{2(\varepsilon_{\mathbb{L}} + \varepsilon_{\mathcal{P}})}{1 - \gamma^{\widehat{d}_{\min}}} + \varepsilon_R\right) + \tilde{\varepsilon} + \varepsilon_S \\ &+ \gamma^{(K+1)d_{\min}} \left(\frac{\left\|V_M^{\Phi^*} - V_0\right\|_{\mathbb{L}}}{1 - \gamma^{d_{\min}}}\right) , \end{aligned} \quad (2)$$

where $\tilde{\varepsilon} = \left(\frac{\gamma^{d_{\min}}}{1-\gamma^{d_{\min}}}\right)\left(1 + \frac{(1-\psi)\gamma^{d_{\min}}}{1-\gamma^{d_{\min}}}\right)(\psi V_{\max} + (1-\psi)\varepsilon_H)$ *and* $\widehat{d}_{\min}$ *and* $d_{\min}$ *are the minimum duration of any*

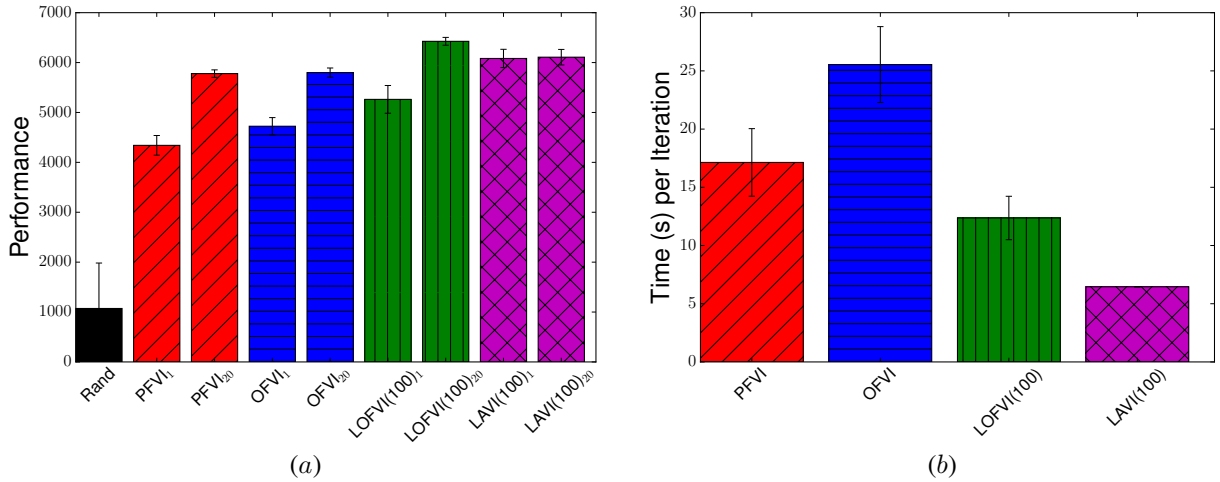

Figure 1: (a) Comparison of performance of the first and last policies derived by PFVI, OFVI, and LAVI. (b) Comparison of time per iteration in seconds. Results were averaged over 20 trials. Error bars represent ± 1 standard deviation.

landmark-option pair in $\widehat{M}$ and M, respectively.

Here m controls the amount of sampling LAVI does at each landmark state. The first four terms on the right hand side of (2) describe the worst case loss of the policy derived by LAVI as $K \to \infty$. $\varepsilon_{\mathbb{L}}$ is the landmark error and $\varepsilon_{\mathcal{P}}$ is the planning error. The first term corresponds to the error associated with the choice of landmarks and using a suboptimal local planner. If LAVI uses an optimal local planner, such as A^*, then $\varepsilon_{\mathcal{P}} = 0$. The second term ε_R is the relaxation error. $\tilde{\varepsilon}$ is controlled by the stochastic plan failure ψ and local Lipschitz error ε_H. If both, ψ and ε_H are small then $\tilde{\varepsilon}$ will be small. In addition, longer duration options (i.e., larger $d_{\min}$) decreases $\tilde{\varepsilon}$. The sample error ε_S is decreased by increasing m. See [Mann *et al.*, 2015] for formal definitions of each term.

The last term is the convergence rate. $\gamma^{d_{\min}}$ is smaller than γ, indicating a faster convergence rate than PFVI [Munos and Szepesvári, 2008; Mann and Mannor, 2014]. $d_{\min}$ is controlled by the minimum time between landmark regions. So convergence is faster when the landmarks provide greater mobility throughout the state-space.

4 Experiments and Results

We compared PFVI, OFVI, LAVI, and LOFVI in an eight commodity inventory management problem [Mann and Mannor, 2014]. We simulated options by simulating individual primitive actions, until the selected option terminates or a maximum number of timesteps (100 in our experiments) occurs. This potentially places options-based planning methods at a disadvantage. Nevertheless, our experiments provide strong evidence that options can speed up the convergence rate of planning, which leads to a smaller time-to-solution.

All experiments were implemented in Java and executed using OpenJDK 1.7 on a desktop computer running Ubuntu 12.04 64-bit with an 8 core Intel Core i7-3370 CPU 3.40GHz and 8 gigabytes of memory.

In a basic inventory management task, the objective is to maintain stock of one or more commodities to meet customer demand while at the same time minimizing ordering costs and storage costs [Scarf, 1959; Sethi and Cheng, 1997]. At each time period, the agent is given the opportunity to order shipments of commodities to resupply its warehouse. See [Mann, 2014] for implementation details of the cyclic inventory management problem used in these experiments. See [Mann *et al.*, 2015] for details on function approximation and how we created hand-crafted options.

Figure 1a compares the performance of a policy that selects primitive actions uniformly at random and policies derived from the first and last iterates of PFVI, OFVI, LOFVI, and LAVI. In this task, LAVI outperforms PFVI and OFVI after on its first iteration, while LOFVI ultimately has higher performance. Figure 1b compares the time per iteration in seconds of PFVI, OFVI, LOFVI, and LAVI. In this task, LAVI is significantly faster than PFVI, OFVI, and LOFVI.

5 Discussion

Option discovery has been investigated extensively, and many approaches explore heuristics related to finding useful subgoals [McGovern and Barto, 2001; Simsek and Barto, 2004; Stolle and Precup, 2002; Wolfe and Barto, 2005], which is similar in spirit to finding landmarks. In all of these approaches, however, the emphasis is on finding only useful subgoals. Our analysis provides instead a way to use any arbitrary set of landmarks, and quantify the quality of the obtained policy. Because of this less careful approach in selecting landmarks, and because of the use of local planning on a deterministic problem, the scalability of LAVI is significantly better, especially in high-dimensional problems.

Acknowledgments

This work was funded in part by the NSERC Discovery grant program and the European Research Council under the European Unions Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n.306638.

References

[Bertsekas and Tsitsiklis, 1996] Dimitri P. Bertsekas and John Tsitsiklis. *Neuro-dynamic programming*. Athena Scientific, 1996.

[Littman *et al.*, 1995] Michael L. Littman, Thomas L. Dean, and Leslie Pack Kaelbling. On the complexity of solving Markov decision problems. In *Proceedings of the* 11^{th} *conference on Uncertainty in artificial intelligence*, pages 394–402, 1995.

[Mann and Mannor, 2014] Timothy A Mann and Shie Mannor. Scaling up approximate value iteration with options: Better policies with fewer iterations. In *Proceedings of the* 31^{st} *International Conference on Machine Learning*, 2014.

[Mann *et al.*, 2015] Timothy A. Mann, Shie Mannor, and Doina Precup. Approximate Value Iteration with Temporally Extended Actions. *Journal of Artificial Intelligence Research*, 53:375–438, 2015.

[Mann, 2014] Timothy A. Mann. Cyclic Inventory Management (CIM). https://code.google.com/p/rddlsim/source/browse/trunk/files/rddl2/examples/cim.rddl2, 2014. Accessed: 2015-06-29.

[McGovern and Barto, 2001] Amy McGovern and Andrew G Barto. Automatic discovery of subgoals in reinforcement learning using diverse density. In *Proceedings of the* 18^{th} *International Conference on Machine Learning*, pages 361 – 368, San Fransisco, USA, 2001.

[Munos and Szepesvári, 2008] Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9:815–857, 2008.

[Puterman, 1994] Martin L Puterman. *Markov Decision Processes - Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., 1994.

[Scarf, 1959] Herbert Scarf. The optimality of (s,S) policies in the dynamic inventory problem. Technical Report NR-047-019, Office of Naval Research, April 1959.

[Sethi and Cheng, 1997] Suresh P. Sethi and Feng Cheng. Optimality of (s,S) policies in inventory models with markovian demand. *Operations Research*, 45(6):931–939, 1997.

[Silver and Ciosek, 2012] David Silver and Kamil Ciosek. Compositional planning using optimal option models. In *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, 2012.

[Simsek and Barto, 2004] Özgür Simsek and Andrew G. Barto. Using relative novelty to identify useful temporal abstractions in reinforcement learning. In *Proceedings of the* 21^{st} *International Conference on Machine Learning*, pages 95–102, New York, NY, USA, 2004. ACM.

[Stolle and Precup, 2002] Martin Stolle and Doina Precup. Learning options in reinforcement learning. In *Abstraction, Reformulation, and Approximation*, pages 212–223. Springer, 2002.

[Sutton *et al.*, 1999] Richard S Sutton, Doina Precup, and Satinder Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1):181–211, August 1999.

[Wolfe and Barto, 2005] Alicia P. Wolfe and Andrew G. Barto. Identifying useful subgoals in reinforcement learning by local graph partitioning. In *Proceedings of the* 22^{nd} *International Conference on Machine Learning*, pages 816–823, 2005.

[Yoon *et al.*, 2007] Sung Wook Yoon, Alan Fern, and Robert Givan. FF-Replan: A Baseline for Probabilistic Planning. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 7, pages 352–359, 2007.