Appendix

John Asmuth and **Michael Littman**
Department of Computer Science
Rutgers University
Piscataway, NJ 08854

This document presents the detailed argument that **BFS3** is near Bayes-optimal.

1 Terms

- N is the maximum number of transitions that will be observed from any particular state-action pair. All subsequent transitions will not be recorded.
- The history $h \in H$ is the collection of all observed transitions $(s, a) \rightarrow s'$.
- ϕ is the MDP prior. $\phi|h$ is the MDP posterior.
- m_0 is the unknown true MDP.
- $P_m(s'|s, a)$ is the probability of going from s to s' when performing action a in some MDP m.
- $P_{\phi|h}(s'|s, a) = \int_m \phi(m|h) P_m(s'|s, a) dm$ is the posterior transition likelihood, integrating out all possible MDPs.
- $V_m(s)$ is the true value of state s according to MDP m.
- $\textbf{FSSS}_m(s, B)$ is the value of state s estimated by **FSSS** when using s as a root, m as an oracle (that samples next-states given a state-action pair), and with computational resource bounds represented by B. In this context, B is the number of trajectories and the search depth that **FSSS** uses to perform roll-outs.

2 Prerequisites

1. The **planning assumption**: For any s, w.p. $1 - \delta_p$, $|\textbf{FSSS}_{m_0}(s, B) - V_{m_0}(s)| \leq \epsilon_p$.
   In other words, **FSSS** will estimate values accurately if it is given the true model and at least B computation time.

2. The **accuracy assumption**: For some state-action pair (s, a), if h includes N samples from $P_{m_0}(s'|s, a)$, then w.p. $1 - \delta_c/(SA)$,

$$\forall_{s'} \; |P_{\phi|h}(s'|s, a) - P_{m_0}(s'|s, a)| \leq \alpha_c.$$

As a result, if N examples of each of the $S \cdot A$ state-action pairs are observed, w.p. $1 - \delta_c$, all transition probability deviations are bounded by α_c.

Note: this condition removes the need to say $m_0 \sim \phi$. All of our closeness assumptions are already met.

3. The **MCTS simulation conjecture**, described in Section 3.

3 MCTS simulation conjecture

Let an MCTS algorithm $\mathcal{A}_m(s, B)$ be an estimator of the value of state s in MDP m, bounded by computational resources B, whose sole source of stochasticity comes from using m as a generator of $s' \sim s, a|m$. **FSSS** and **UCT** (Kocsis & Szepesvári, 2006) both fall into this category; their resources are the number of roll-outs and maximum search depth.

Intuitively, we conjecture that if $\mathcal{A}$ can reliably find accurate values for states in one MDP, then it can also reliably find accurate values for the MDP when using an approximate second MDP as its next-state generator.

If,

1. for all s, a, s', $|P_{m_a}(s'|s, a) - P_{m_b}(s'|s, a)| \leq \alpha_c$,
2. for all s, w.p. $1 - \delta_p/(CA)$, $|\mathcal{A}_{m_a}(s, B) - V_{m_a}(s)| \leq \epsilon_p$,

then

- for all s, w.p. $1 - \delta_{p'}/(CA)$, $|\mathcal{A}_{m_b}(s, B) - V_{m_a}(s)| \leq \epsilon_{p'}$,

where $1/\delta_{p'} = \text{poly}(1/(1-\alpha_c), 1/\delta_p, 1/\epsilon_p, m_a, B)$ and $\epsilon_{p'} = \text{poly}(1/(1-\alpha_c), 1/\delta_p, 1/\epsilon_p, m_a, B)$.

3.1 Note on $C \cdot A$

The probability $1 - \delta_p/(CA)$ is used because, for each step, **BFS3** will invoke **FSSS** $C \cdot A$ times, or C times for each action. This choice was made to ensure that all next states get enough resources to ensure accuracy when their dynamics are known. We use the union bound to say, w.p. $1 - \delta_p$, *all* such next states have accurate estimates.

4 PAC-MDP behavior in the belief-MDP

Let

- the belief-MDP be $m_{\phi|h}$,
- the set of all possible histories be H,

- the state space of the belief-MDP be the set of belief-states $\mathbb{S} = S \times H$, and
- the action space of the belief-MDP be the same as the regular MDP.

We shall limit H to contain only histories with at most N examples of transitions from any state-action pair. That is, when the agent observes the N_{th} transition from state s and action a, it shall never again update its history when making a transition from that state with that action. This forgetfulness causes the set H to be finite.

In a discrete state and action MDP, the history $h \in H$ may be summarized by a set of histograms, one for each state-action pair. Since we ignore all transitions from any state-action pair after the N_{th} transition from that state-action pair, we know that the sum of all entries in all histograms cannot exceed $S \cdot A \cdot N$.

Since, whenever the history *is* updated, a single entry for a single histogram will be incremented by 1, we can infer that at most $S \cdot A \cdot N$ unique histories can possibly be experienced by an agent over the course of a single experiment.

The agent can also visit at most S true states (all of them). Since there are S true states and $S \cdot A \cdot N$ possible histories, the number of belief-states that an agent can visit, in a single experiment, is $M = S^2 \cdot A \cdot N$.

Our strategy will be to say that, for each of these at most M possible belief-states visited by the agent, the agent will either choose an action that is approximately optimal in the belief-MDP (and therefore approximately Bayes-optimal in the true MDP), or exploratory in the belief-MDP (with no guarantees about how good this action is in the true MDP).

We say that an agent is near Bayes-optimal if we can limit the number of exploratory actions taken to a polynomial of the domain parameters.

To show PAC-MDP behavior in the belief-MDP (and, therefore, near Bayes-optimal behavior), we need to show that three conditions hold (Li, 2009):

1. bounded discoveries,
2. accuracy, and
3. optimism.

To understand these criteria and how they connect PAC-MDP behavior in the belief-MDP to near Bayes-optimal behavior in the true MDP, we need to define the concept of *known* and *unknown states* in m_0 and *known* and *unknown belief-states* in m_ϕ.

A *state* s is considered known if the history h contains N examples of transitions from s for each action $a \in A$.

A *belief-state* $\langle s, h \rangle$ is considered known if **FSSS**'s estimate of its value is accurate.

Figure 1 illustrates how unknown states in a belief-state's **FSSS** subtree may cause a belief-state's estimate to be inaccurate. Because of Prerequisite 1, the planning assumption, we know that if there are no unknown states in a belief-state's subtree, then the belief-state's estimate will be accurate, and we can consider the belief-state to be known.

It is possible for a belief-state to be known if there are unknown states in its subtree; in the limit of computation, **FSSS** will be accurate for all belief-states. But, it is not

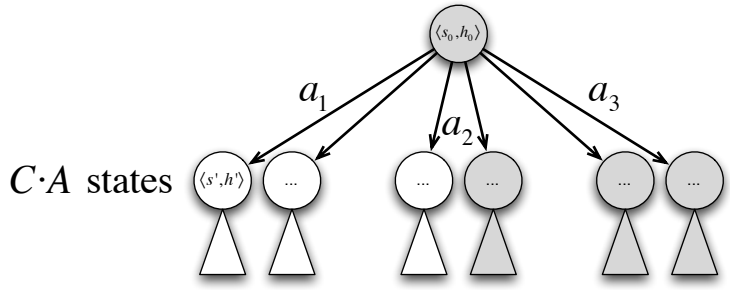

Figure 1: **BFS3** will run **FSSS** on each of the $C \cdot A$ next belief-states, $\langle s', h' \rangle$ and its peers. Triangles below belief-states represent search trees explored by **FSSS**. A greyed search tree indicates that some of the true states visited in that search tree are unknown, and the value estimate may be inaccurate. Because we are using **FSSS**, an inaccurate estimate is guaranteed to be optimistic, and can bubble up to the root.

possible for all states in the belief-state's subtree to be known if the belief-state is not known.

As a result, an unknown belief-state indicates an unknown state that is at most D steps away, where D is the search depth given to **FSSS**. Since unknown belief-states always have optimistic estimates when they are evaluated with **FSSS**, the agent is pulled towards unknown states.

4.1 Satisfying the PAC-MDP criteria

BFS3 is PAC-MDP for the belief-state MDP because it satisfies the 3 criteria.

Condition 1 holds because if the agent reaches an unknown belief state (one with an inaccurate, optimistic value), it must be because there is an unknown state-action pair beneath it in the tree *and* this node will be reached by the current policy with high probability. (Quantifying this claim involves an argument that closely parallels the "explore or exploit" lemma from PAC-MDP theory.) Note that when an unknown state-action pair is reached, it moves closer to being known. In fact, an agent can only do this update $S \cdot A \cdot N$ times, so, the number of discoveries is bounded.

Condition 2 holds because of Prerequisite 2 (the accuracy assumption). Once we have N examples of a state-action pair, we have an accurate estimate of the next-state distribution for that pair. Once we have N examples of all state-action pairs, we have an accurate estimate of the entire MDP.

Condition 3 holds because of our definition of known and unknown belief-states and the behavior of **FSSS**. Known belief-states have accurate values, and unknown belief-states have optimistic values. No belief-state ever has a pessimistic value.

5 How accurate, and how likely?

We have shown that **BFS3** is near Bayes-optimal, meaning with high probability, it will be accurate for all but a small number of steps. This section details how accurate **BFS3** is and with what probability it achieves this accuracy.

5.1 Bayes-optimal behavior for a converged posterior

We declare the posterior to be converged if, for each s, a, our history h includes N examples of a transition from s, a.

Let m_0 and $\phi|h$ take the place of m_a and m_b in the MCTS simulation conjecture. By Prerequisite 2, we have Condition 1 for the MCTS simulation conjecture. By Prerequisite 1, we have Condition 2 for the MCTS simulation conjecture. For each step in the environment, **BFS3** runs **FSSS** C times for each action, resulting in $C \cdot A$ executions of **FSSS**. We can put a lower bound on the likelihood that they are all $\epsilon_{p'}$-accurate of $1 - \delta_{p'}$.

The estimate for the value of taking action a from the root state s_0 is calculated as

$$\hat{Q}(s_0, a) = R(s_0, a) + \gamma \frac{1}{C} \sum_{i=1}^{C} \textbf{FSSS}_{\phi|h}(s_i, B). \tag{1}$$

Let the Monte Carlo estimate of $Q(s_0, a)$, using the true value for all next states, be

$$\tilde{Q}(s_0, a) = R(s_0, a) + \gamma \frac{1}{C} \sum_{i=1}^{C} V^*(s_i). \tag{2}$$

From the Bellman equation, we know that

$$Q^*(s_0, a) = R(s_0, a) + \gamma \sum_{s_i} P_{m_0}(s_i|s, a) V^*(s_i). \tag{3}$$

First, let's bound the difference between $\hat{Q}(s_0, a)$ and $\tilde{Q}(s_0, a)$.

$$\begin{aligned}
|\hat{Q}(s_0, a) - \tilde{Q}(s_0, a)| &= \left| \gamma \frac{1}{C} \sum_{i=1}^{C} \text{FSSS}_{\phi|h}(s_i, B) - \gamma \frac{1}{C} \sum_{i=1}^{C} V^*(s_i) \right| \\
&= \left| \gamma \frac{1}{C} \sum_{i=1}^{C} \left(\text{FSSS}_{\phi|h}(s_i, B) - V^*(s_i) \right) \right| \\
&\leq \gamma \frac{1}{C} \sum_{i=1}^{C} \left| \text{FSSS}_{\phi|h}(s_i, B) - V^*(s_i) \right| \\
&\leq \gamma \frac{1}{C} \sum_{i=1}^{C} \epsilon_{p'} \\
|\hat{Q}(s_0, a) - \tilde{Q}(s_0, a)| &\leq \epsilon_{p'}.
\end{aligned} \tag{4}$$

Second, let's bound the difference between $\tilde{Q}(s_0, a)$ and $Q^*(s_0, a)$.

$$\begin{aligned} |Q^*(s_0,a) - \tilde{Q}(s_0,a)| &= \left| \gamma \sum_{s_i} P_{m_0}(s_i|s,a) V^*(s_i) - \gamma \frac{1}{C} \sum_{i=1}^{C} V^*(s_i) \right| \\ &= \gamma \left| \sum_{s_i} P_{m_0}(s_i|s,a) V^*(s_i) - \frac{1}{C} \sum_{i=1}^{C} V^*(s_i) \right| \\ |Q^*(s_0,a) - \tilde{Q}(s_0,a)| &\leq \lambda, \end{aligned} \tag{5}$$

w.p. $1 - e^{-\lambda^2 C / V_{\max}^2}$ (Kearns et al., 1999).

This bounds the error in the estimated Q-values:

$$\begin{aligned} |\hat{Q}(s_0,a) - Q^*(s_0,a)| &= |\hat{Q}(s_0,a) - \tilde{Q}(s_0,a) + \tilde{Q}(s_0,a) - Q^*(s_0,a)| \\ &\leq |\hat{Q}(s_0,a) - \tilde{Q}(s_0,a)| + |\tilde{Q}(s_0,a) - Q^*(s_0,a)| \\ |\hat{Q}(s_0,a) - Q^*(s_0,a)| &\leq \epsilon_{p'} + \lambda, \end{aligned} \tag{6}$$

w.p. $1 - \delta_{p'} - e^{-\lambda^2 C / V_{\max}^2}$.

Let $\epsilon = \epsilon_{p'} + \lambda$, and $\delta = \delta_c + S\left(\delta_{p'} + e^{-\lambda^2 C / V_{\max}^2}\right)$.

Once the model has converged, every time the agent takes a step from some state s_0 it remembers what action it took, and will take that action in the future. The likelihood of it choosing an ϵ-optimal action for each of the S possible states is no less than $1 - \delta$.

5.2 Bayes-optimal behavior for an unconverged posterior

Since the agent will not update the history h on a transition from s, a when transitions s, a have been observed N times in the past, we know there are only $S \cdot A \cdot N$ different histories possible over the course of a single experiment.

Given a particular history h and state s that has been experienced before, **BFS3** will choose the same action as last time, limiting the total amount of computation possible, and also the total number of times **BFS3** has to succeed is limited at $M = S^2 \cdot A \cdot N$, or the number of states times the number of possible histories.

In Sections 5.2.1 and 5.2.2, we set conditions for **BFS3** to choose either an exploratory action or an optimal action for each of these M possible events.

On a given step, **BFS3** will run **FSSS** on each of A possible actions and, for each action, C possible next-states.

5.2.1 In the limit of infinite computation

Each of the next-state possibilities will be fully explored and FSSS will agree with Sparse Sampling, giving an error no more than

$$\epsilon_\infty = \frac{\lambda(1 - \gamma^{D+1})}{1 - \gamma} + \gamma^D V_{\max},$$

w.p. at least $1 - (C \cdot A)^D e^{-\lambda^2 C / V_{\max}^2}$, where D is the maximum search depth. Let $\delta_\infty / M = (C \cdot A)^D e^{-\lambda^2 C / V_{\max}^2}$; the agent will then be able to pick an ϵ_∞-optimal action for each of the M events w.p. at least $1 - \delta_\infty$.

5.2.2 Limited computation

In cases where there is limited computation and all states in the subtree have converged according to Prerequisite 2, we are guaranteed to act ϵ_f-optimal w.p. at least $1-\delta_f/M$, according to Section 5.1, where

$$\epsilon_f = \epsilon_{p'} + \lambda$$

and

$$\delta_f/M = \delta_{p'} + e^{-\lambda^2 C/V_{\max}^2}.$$

In cases with limited computation and when there are states in the subtree that are not converged, we are guaranteed to have an ϵ_f*-optimistic* estimate w.p. at least $1 - \delta_f/M$. This is true because of the guarantees provided by **FSSS**.

An action a's estimate is ϵ-optimistic if

$$\hat{Q}(s,a) \geq Q^*(s,a) - \epsilon.$$

Note that an ϵ-optimal estimate is also ϵ-optimistic.

Since for each of the M events we have an ϵ_f-optimistic estimate w.p. $1 - \delta_f/M$, we will have an ϵ_f-optimistic estimate for each event simultaneously w.p. $1 - \delta_f$.

References

Kearns, M., Mansour, Y., & Ng, A. Y. (1999). A sparse sampling algorithm for near-optimal planning in large Markov decision processes. *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI-99)* (pp. 1324–1331).

Kocsis, L., & Szepesvári, C. (2006). Bandit based Monte-Carlo planning. *Proceedings of the 17th European Conference on Machine Learning (ECML-06)* (pp. 282–293). Springer Berlin / Heidelberg.

Li, L. (2009). *A unifying framework for computational reinforcement learning theory (pp 78-79)*. Doctoral dissertation, Rutgers University.