
Feature Markov Decision Processes

Marcus Hutter

RSISE @ ANU and SML @ NICTA
Canberra, ACT, 0200, Australia
marcus@hutter1.net www.hutter1.net

24 December 2008

Abstract

General purpose intelligent learning agents cycle through (complex, non-MDP) sequences of observations, actions, and rewards. On the other hand, reinforcement learning is well-developed for small finite state Markov Decision Processes (MDPs). So far it is an art performed by human designers to extract the right state representation out of the bare observations, i.e. to reduce the agent setup to the MDP framework. Before we can think of mechanizing this search for suitable MDPs, we need a formal objective criterion. The main contribution of this article is to develop such a criterion. I also integrate the various parts into one learning algorithm. Extensions to more realistic dynamic Bayesian networks are developed in the companion article [Hut09].

Keywords: evolutionary algorithms, ranking selection, tournament selection, equivalence, efficiency.

1 Introduction

Background & motivation. Artificial General Intelligence (AGI) is concerned with designing agents that perform well in a wide range of environments [GP07, LH07]. Among the well-established “narrow” AI approaches, arguably Reinforcement Learning (RL) pursues most directly the same goal. RL considers the general agent-environment setup in which an agent interacts with an environment (acts and observes in cycles) and receives (occasional) rewards. The agent’s objective is to collect as much reward as possible. Most if not all AI problems can be formulated in this framework.

The simplest interesting environmental class consists of finite state fully observable Markov Decision Processes (MDPs) [Put94, SB98], which is reasonably well understood. Extensions to continuous states with (non)linear function approximation [SB98, Gor99], partial observability (POMDP) [KLC98, RPPCd08], structured MDPs (DBNs) [SDL07], and others have been considered, but the algorithms are much more brittle.

In any case, a lot of work is still left to the designer,

namely to extract the right state representation (“features”) out of the bare observations. Even if *potentially* useful representations have been found, it is usually not clear which one will turn out to be better, except in situations where we already know a perfect model. Think of a mobile robot equipped with a camera plunged into an unknown environment. While we can imagine which image features are potentially useful, we cannot know which ones will actually be useful.

Main contribution. Before we can think of mechanically searching for the “best” MDP representation, we need a formal objective criterion. Obviously, at any point in time, if we want the criterion to be effective it can only depend on the agents past experience. The main contribution of this article is to develop such a criterion. Reality is a non-ergodic partially observable uncertain unknown environment in which acquiring experience can be expensive. So we want/need to exploit the data (past experience) at hand optimally, cannot generate virtual samples since the model is not given (need to be learned itself), and there is no reset-option. In regression and classification, penalized maximum likelihood criteria [HTF01, Chp.7] have successfully been used for semi-parametric model selection. It is far from obvious how to apply them in RL. Ultimately we do not care about the observations but the rewards. The rewards depend on the states, but the states are arbitrary in the sense that they are model-dependent functions of the data. Indeed, our derived Cost function cannot be interpreted as a usual model+data code length.

Relation to other work. As partly detailed later, the suggested Φ MDP model could be regarded as a scaled-down practical instantiation of AIXI [Hut05, Hut07], as a way to side-step the open problem of learning POMDPs, as extending the idea of state-aggregation from planning (based on bi-simulation metrics [GDG03]) to RL (based on code length), as generalizing U-Tree [McC96] to arbitrary features, or as an alternative to PSRs [SLJ⁺03] for which proper learning algorithms have yet to be developed.

Notation. Throughout this article, \log denotes the binary logarithm, ϵ the empty string, and $\delta_{x,y} = \delta_{xy} = 1$ if $x=y$ and 0 else is the Kronecker symbol. I generally omit separating commas if no confusion arises, in particular in

indices. For any x of suitable type (string,vector,set), I define string $\mathbf{x} = x_{1:l} = x_1 \dots x_l$, sum $x_+ = \sum_j x_j$, union $x_* = \bigcup_j x_j$, and vector $\mathbf{x}_* = (x_1, \dots, x_l)$, where j ranges over the full range $\{1, \dots, l\}$ and $l = |x|$ is the length or dimension or size of x . \hat{x} denotes an estimate of x . $P(\cdot)$ denotes a probability over states and rewards or parts thereof. I do not distinguish between random variables X and realizations x , and abbreviation $P(x) := P[X = x]$ never leads to confusion. More specifically, $m \in \mathbb{N}$ denotes the number of states, $i \in \{1, \dots, m\}$ any state index, $n \in \mathbb{N}$ the current time, and $t \in \{1, \dots, n\}$ any time. Further, in order not to get distracted at several places I gloss over initial conditions or special cases where inessential. Also $0 * \text{undefined} = 0 * \text{infinity} = 0$.

2 Feature Markov Decision Process (Φ MDP)

This section describes our formal setup. It consists of the agent-environment framework and maps Φ from observation-action-reward histories to MDP states. I call this arrangement “Feature MDP” or short Φ MDP.

Agent-environment setup. I consider the standard agent-environment setup [RN03] in which an *Agent* interacts with an *Environment*. The agent can choose from actions $a \in \mathcal{A}$ (e.g. limb movements) and the environment provides (regular) observations $o \in \mathcal{O}$ (e.g. camera images) and real-valued rewards $r \in \mathcal{R} \subseteq \mathbb{R}$ to the agent. The reward may be very scarce, e.g. just +1 (-1) for winning (losing) a chess game, and 0 at all other times [Hut05, Sec.6.3]. This happens in cycles $t = 1, 2, 3, \dots$: At time t , after observing o_t , the agent takes action a_t based on history $h_t := o_1 a_1 r_1 \dots o_{t-1} a_{t-1} r_{t-1} o_t$. Thereafter, the agent receives reward r_t . Then the next cycle $t+1$ starts. The agent’s objective is to maximize his long-term reward. Without much loss of generality, I assume that \mathcal{A} , \mathcal{O} , and \mathcal{R} are finite. Implicitly I assume \mathcal{A} to be small, while \mathcal{O} may be huge.

The agent and environment may be viewed as a pair or triple of interlocking functions of the history $\mathcal{H} := (\mathcal{O} \times \mathcal{A} \times \mathcal{R})^* \times \mathcal{O}$:

$$\begin{aligned} \text{Env} : \mathcal{H} \times \mathcal{A} \times \mathcal{R} &\rightsquigarrow \mathcal{O}, & o_n &= \text{Env}(h_{n-1} a_{n-1} r_{n-1}), \\ \text{Agent} : \mathcal{H} &\rightsquigarrow \mathcal{A}, & a_n &= \text{Agent}(h_n), \\ \text{Env} : \mathcal{H} \times \mathcal{A} &\rightsquigarrow \mathcal{R}, & r_n &= \text{Env}(h_n a_n). \end{aligned}$$

where \rightsquigarrow indicates that mappings \rightarrow might be stochastic.

The goal of AI is to design agents that achieve high (expected) reward over the agent’s lifetime.

(Un)known environments. For known $\text{Env}()$, finding the reward maximizing agent is a well-defined and formally solvable problem [Hut05, Chp.4], with computational efficiency being the “only” matter of concern. For most real-world AI problems $\text{Env}()$ is at best partially known.

Narrow AI considers the case where function $\text{Env}()$ is either known (like in blocks world), or essentially known

(like in chess, where one can safely model the opponent as a perfect minimax player), or $\text{Env}()$ belongs to a relatively small class of environments (e.g. traffic control).

The goal of AGI is to design agents that perform well in a large range of environments [LH07], i.e. achieve high reward over their lifetime with as little as possible assumptions about $\text{Env}()$. A minimal necessary assumption is that the environment possesses *some* structure or pattern.

From real-life experience (and from the examples below) we know that usually we do not need to know the complete history of events in order to determine (sufficiently well) what will happen next and to be able to perform well. Let $\Phi(h)$ be such a “useful” summary of history h .

Examples. In full-information *games* (like chess) with static opponent, it is sufficient to know the current state of the game (board configuration) to play well (the history plays no role), hence $\Phi(h_t) = o_t$ is a sufficient summary (Markov condition). Classical *physics* is essentially predictable from position and velocity of objects at a single time, or equivalently from the locations at two consecutive times, hence $\Phi(h_t) = o_t o_{t-1}$ is a sufficient summary (2nd order Markov). For *i.i.d. processes* of unknown probability (e.g. clinical trials \simeq Bandits), the frequency of observations $\Phi(h_n) = (\sum_{t=1}^n \delta_{o_t o})_{o \in \mathcal{O}}$ is a sufficient statistic. In a *POMDP planning* problem, the so-called belief vector at time t can be written down explicitly as some function of the complete history h_t (by integrating out the hidden states). $\Phi(h_t)$ could be chosen as (a discretized version of) this belief vector, showing that Φ MDP generalizes POMDPs. Obviously, the *identity* $\Phi(h) = h$ is always sufficient but not very useful, since $\text{Env}()$ as a function of \mathcal{H} is hard to impossible to “learn”.

This suggests to look for Φ with small codomain, which allow to learn/estimate/approximate Env by $\widehat{\text{Env}}$ such that $o_t \approx \widehat{\text{Env}}(\Phi(h_{t-1}))$ for $t = 1 \dots n$.

Example. Consider a robot equipped with a camera, i.e. o is a pixel image. Computer vision algorithms usually extract a set of features from o_{t-1} (or h_{t-1}), from low-level patterns to high-level objects with their spatial relation. Neither is it possible nor necessary to make a precise prediction of o_t from summary $\Phi(h_{t-1})$. An approximate prediction must and will do. The difficulty is that the similarity measure “ \approx ” needs to be context dependent. Minor image nuances are irrelevant when driving a car, but when buying a painting it makes a huge difference in price whether it’s an original or a copy. Essentially only a bijection Φ would be able to extract *all potentially* interesting features, but such a Φ defeats its original purpose.

From histories to states. It is of utmost importance to properly formalize the meaning of “ \approx ” in a general, domain-independent way. Let $s_t := \Phi(h_t)$ summarize all relevant information in history h_t . I call s a state or feature (vector) of h . “Relevant” means that the future is predictable from s_t (and a_t) alone, and that the relevant future is coded in $s_{t+1} s_{t+2} \dots$. So we pass from the

complete (and known) history $o_1 a_1 r_1 \dots o_n a_n r_n$ to a “compressed” history $sar_{1:n} \equiv s_1 a_1 r_1 \dots s_n a_n r_n$ and seek Φ such that s_{t+1} is (approximately a stochastic) function of s_t (and a_t). Since the goal of the agent is to maximize his rewards, the rewards r_t are always relevant, so they (have to) stay untouched (this will become clearer below).

The Φ MDP. The structure derived above is a classical Markov Decision Process (MDP), but the primary question I ask is not the usual one of finding the value function or best action or comparing different models of a given state sequence. I ask how well can the state-action-reward sequence generated by Φ be modeled as an MDP compared to other sequences resulting from different Φ .

3 Φ MDP Coding and Evaluation

I first review optimal codes and model selection methods for i.i.d. sequences, subsequently adapt them to our situation, and show that they are suitable in our context. I state my Cost function for Φ and the Φ selection principle.

I.i.d. processes. Consider i.i.d. $x_1 \dots x_n \in \mathcal{X}^n$ for finite $\mathcal{X} = \{1, \dots, m\}$. For known $\theta_i = P[x_t = i]$ we have $P(x_{1:n}|\theta) = \theta_{x_1} \dots \theta_{x_n}$. It is well-known that there exists a code (e.g. arithmetic or Shannon-Fano) for $x_{1:n}$ of length $-\log P(x_{1:n}|\theta)$, which is asymptotically optimal with probability one.

For unknown θ we may use a frequency estimate $\hat{\theta}_i = n_i/n$, where $n_i = |\{t : x_t = i\}|$. Then $-\log P(x_{1:n}|\hat{\theta}) = n H(\hat{\theta})$, where $H(\hat{\theta}) := -\sum_{i=1}^m \hat{\theta}_i \log \hat{\theta}_i$ is the Entropy of $\hat{\theta}$ ($0 \log 0 := 0 = 0 \log \frac{0}{0}$). We also need to code (n_i) , which naively needs $\log n$ bits for each i . One can show that it is sufficient to code each $\hat{\theta}_i$ to accuracy $O(1/\sqrt{n})$, which requires only $\frac{1}{2} \log n + O(1)$ bits each. Hence the overall code length of $x_{1:n}$ for unknown frequencies is

$$\text{CL}(x_{1:n}) = \text{CL}(\mathbf{n}) := n H(\mathbf{n}/n) + \frac{m'-1}{2} \log n \quad (1)$$

for $n > 0$ and 0 else, where $\mathbf{n} = (n_1, \dots, n_m)$ and $n = n_+ = n_1 + \dots + n_m$ and $m' = |\{i : n_i > 0\}| \leq m$ is the number of non-empty categories. The code is optimal (within $+O(1)$) for all i.i.d. sources. It can be rigorously derived from many principles: MDL, MML, combinatorial, incremental, and Bayesian [Grü07].

In the following I will ignore the $O(1)$ terms and refer to (1) simply as *the* code length. Note that $x_{1:n}$ is coded exactly (lossless). Similarly (see MDP below) sampling models more complex than i.i.d. may be considered, and the one that leads to the shortest code is selected as the best model [Grü07].

MDP definitions. Recall that a sequence $sar_{1:n}$ is said to be sampled from an MDP $(\mathcal{S}, \mathcal{A}, T, R)$ iff the probability of s_t only depends on s_{t-1} and a_{t-1} ; and r_t only on s_{t-1} , a_{t-1} , and s_t . That is,

$$\begin{aligned} P(s_t | h_{t-1} a_{t-1}) &= P(s_t | s_{t-1}, a_{t-1}) &=: T_{s_{t-1} s_t}^{a_{t-1}} \\ P(r_t | h_t) &= P(r_t | s_{t-1}, a_{t-1}, s_t) &=: R_{s_{t-1} s_t}^{a_{t-1} r_t} \end{aligned}$$

For simplicity of exposition I assume a deterministic dependence of r_t on s_t only, i.e. $r_t = R_{s_t}$. In our case, we can identify the state-space \mathcal{S} with the states s_1, \dots, s_n “observed” so far. Hence $\mathcal{S} = \{s^1, \dots, s^m\}$ is finite and typically $m \ll n$, i.e. states repeat. Let $s \xrightarrow{a} s'(r')$ be shorthand for “action a in state s resulted in state s' (reward r')”. Let $\mathcal{T}_{ss'}^{ar'} := \{t \leq n : s_{t-1} = s, a_{t-1} = a, s_t = s', r_t = r'\}$ be the set of times $t-1$ at which $s \xrightarrow{a} s'r'$, and $n_{ss'}^{ar'} := |\mathcal{T}_{ss'}^{ar'}|$ their number ($n_{++}^+ = n$).

Coding MDP sequences. For some fixed s and a , consider the subsequence $s_{t_1} \dots s_{t_{n'}}$ of states reached from s via a ($s \xrightarrow{a} s_{t_i}$), i.e. $\{t_1, \dots, t_{n'}\} = \mathcal{T}_{ss'}^{a*}$, where $n' = n_{ss'}^{a+}$. By definition of an MDP, this sequence is i.i.d. with s' occurring $n'_{s'} := n_{ss'}^{a+}$ times. By (1) we can code this sequence in $\text{CL}(n')$ bits. The whole sequence $s_{1:n}$ consists of $|\mathcal{S} \times \mathcal{A}|$ i.i.d. sequences, one for each $(s, a) \in \mathcal{S} \times \mathcal{A}$. We can join their codes and get a total code length

$$\text{CL}(s_{1:n} | a_{1:n}) = \sum_{s,a} \text{CL}(n_{s,a}^{a+}) \quad (2)$$

Similarly to the states we code the rewards. There are different “standard” reward models. I consider only the simplest case of a small discrete reward set \mathcal{R} like $\{0, 1\}$ or $\{-1, 0, +1\}$ here and defer generalizations to \mathbb{R} and a discussion of variants to the Φ DBN model [Hut09]. By the MDP assumption, for each state s' , the rewards at times $\mathcal{T}_{+s'}^{+*}$ are i.i.d. Hence they can be coded in

$$\text{CL}(r_{1:n} | s_{1:n}, a_{1:n}) = \sum_{s'} \text{CL}(n_{+s'}^{+*}) \quad (3)$$

bits. I have been careful to assign zero code length to non-occurring transitions $s \xrightarrow{a} s'r'$ so that large but sparse MDPs don’t get penalized too much.

Reward \leftrightarrow state trade-off. Note that the code for \mathbf{r} depends on \mathbf{s} . Indeed we may interpret the construction as follows: Ultimately we/the agent cares about the reward, so we want to measure how well we can predict the rewards, which we do with (3). But this code depends on \mathbf{s} , so we need a code for \mathbf{s} too, which is (2). To see that we need both parts consider two extremes.

A simplistic state transition model (small $|\mathcal{S}|$) results in a short code for \mathbf{s} . For instance, for $|\mathcal{S}| = 1$, nothing needs to be coded and (2) is identically zero. But this obscures potential structure in the reward sequence, leading to a long code for \mathbf{r} .

On the other hand, the more detailed the state transition model (large $|\mathcal{S}|$) the easier it is to predict and hence compress \mathbf{r} . But a large model is hard to learn, i.e. the code for \mathbf{s} will be large. For instance for $\Phi(h) = h$, no state repeats and the frequency-based coding breaks down.

Φ selection principle. Let us define the *Cost* of $\Phi : \mathcal{H} \rightarrow \mathcal{S}$ on h_n as the length of the Φ MDP code for \mathbf{sr} given \mathbf{a} :

$$\begin{aligned} \text{Cost}(\Phi | h_n) &:= \text{CL}(s_{1:n} | a_{1:n}) + \text{CL}(r_{1:n} | s_{1:n}, a_{1:n}), \quad (4) \\ &\text{where } s_t = \Phi(h_t) \text{ and } h_t = oar_{1:t-1} o_t \end{aligned}$$

The discussion above suggests that the minimum of the joint code length, i.e. the Cost, is attained for a Φ that keeps all and only relevant information for predicting rewards. Such a Φ may be regarded as best explaining the rewards. So we are looking for a Φ of minimal cost:

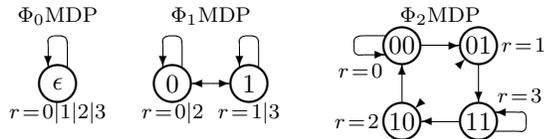
$$\Phi^{best} := \arg \min_{\Phi} \{\text{Cost}(\Phi|h_n)\} \quad (5)$$

The state sequence generated by Φ^{best} (or approximations thereof) will usually only be approximately MDP. While $\text{Cost}(\Phi|h)$ is an optimal code only for MDP sequences, it still yields good codes for approximate MDP sequences. Indeed, Φ^{best} balances closeness to MDP with simplicity. The primary purpose of the simplicity bias is *not* computational tractability, but generalization ability [LH07, Hut05].

4 A Tiny Example

The purpose of the tiny example in this section is to provide enough insight into how and why Φ MDP works to convince the reader that our Φ selection principle is reasonable. Consider binary observation space $\mathcal{O} = \{0,1\}$, quaternary reward space $\mathcal{R} = \{0,1,2,3\}$, and a single action $\mathcal{A} = \{0\}$. Observations o_t are independent fair coin flips, i.e. Bernoulli($\frac{1}{2}$), and reward $r_t = 2o_{t-1} + o_t$ a deterministic function of the two most recent observations.

Considered features. As features Φ I consider $\Phi_k: \mathcal{H} \rightarrow \mathcal{O}^k$ with $\Phi_k(h_t) = o_{t-k+1} \dots o_t$ for various $k=0,1,2,\dots$ which regard the last k observations as “relevant”. Intuitively Φ_2 is the best observation summary, which I confirm below. The state space $\mathcal{S} = \{0,1\}^k$ (for sufficiently large n). The Φ MDPs for $k=0,1,2$ are as follows.



Φ_2 MDP with all non-zero transition probabilities being 50% is an exact representation of our data source. The missing arrow (directions) are due to the fact that $s = o_{t-1}o_t$ can only lead to $s' = o'_t o'_{t+1}$ for which $o'_t = o_t$. Note that Φ MDP does not “know” this and has to learn the (non)zero transition probabilities. Each state has two successor states with equal probability, hence generates (see previous paragraph) a Bernoulli($\frac{1}{2}$) state subsequence and a constant reward sequence, since the reward can be computed from the state = last two observations. Asymptotically, all four states occur equally often, hence the sequences have approximately the same length $n/4$.

In general, if s (and similarly r) consists of $x \in \mathcal{N}$ i.i.d. subsequences of equal length n/x over $y \in \mathcal{N}$ symbols, the code length (2) (and similarly (3)) is

$$\begin{aligned} \text{CL}(s|\mathbf{a}; x_y) &= n \log y + x \frac{|\mathcal{S}|-1}{2} \log \frac{n}{x}, \\ \text{CL}(r|\mathbf{s}, \mathbf{a}; x_y) &= n \log y + x \frac{|\mathcal{R}|-1}{2} \log \frac{n}{x} \end{aligned}$$

where the extra argument x_y just indicates the sequence property. So for Φ_2 MDP we get

$$\text{CL}(s|\mathbf{a}; 4_2) = n + 6 \log \frac{n}{4} \quad \text{and} \quad \text{CL}(r|\mathbf{s}, \mathbf{a}; 4_1) = 6 \log \frac{n}{4}$$

The log-terms reflect the required memory to code (or the time to learn) the MDP structure and probabilities. Since each state has only 2 realized/possible successors, we need n bits to code the state sequence. The reward is a deterministic function of the state, hence needs no memory to code given s .

The Φ_0 MDP throws away all observations (left figure above), hence $\text{CL}(s|\mathbf{a}; 1_1) = 0$. While the reward sequence is *not* i.i.d. (e.g. $r_{t+1} = 3$ cannot follow $r_t = 0$), Φ_0 MDP has no choice regarding them as i.i.d., resulting in $\text{CL}(s|\mathbf{a}; 1_4) = 2n + \frac{3}{2} \log n$.

The Φ_1 MDP model is an interesting compromise (middle figure above). The state allows a partial prediction of the reward: State 0 allows rewards 0 and 2; state 1 allows rewards 1 and 3. Each of the two states creates a Bernoulli($\frac{1}{2}$) state successor subsequence and a binary reward sequence, wrongly presumed to be Bernoulli($\frac{1}{2}$). Hence $\text{CL}(s|\mathbf{a}; 2_2) = n + \log \frac{n}{2}$ and $\text{CL}(r|\mathbf{s}, \mathbf{a}; 2_2) = n + 3 \log \frac{n}{2}$.

Summary. The following table summarizes the results for general $k=0,1,2$ and beyond:

$\text{Cost}(\Phi_0 h)$	$\text{Cost}(\Phi_1 h)$	$\text{Cost}(\Phi_2 h)$	$\text{Cost}(\Phi_{k \geq 2} h)$
$2n + \frac{3}{2} \log n$	$2n + 4 \log \frac{n}{2}$	$n + 12 \log \frac{n}{4}$	$n + \frac{2^k + 2}{2^{1-k}} \log \frac{n}{2^k}$

For large n , Φ_2 results in the shortest code, as anticipated. The “approximate” model Φ_1 is just not good enough to beat the vacuous model Φ_0 , but in more realistic examples some approximate model usually has the shortest code. In [Hut09] I show on a more complex example how Φ^{best} will store long-term information in a POMDP environment.

5 Cost(Φ) Minimization

I have reduced the reinforcement learning problem to a formal Φ -optimization problem. I briefly explain what we have gained by this reduction, and provide some general information about problem representations, stochastic search, and Φ neighborhoods. Finally I present a simplistic but concrete algorithm for searching context tree MDPs.

Φ search. I now discuss how to find good summaries Φ . The introduced generic cost function $\text{Cost}(\Phi|h_n)$, based on only the known history h_n , makes this a well-defined task that is completely decoupled from the complex (ill-defined) reinforcement learning objective. This reduction should not be under-estimated. We can employ a wide range of optimizers and do not even have to worry about overfitting. The most challenging task is to come up with creative algorithms proposing Φ 's.

There are many optimization methods: Most of them are search-based: random, blind, informed, adaptive, local, global, population based, exhaustive, heuristic, and

other search methods [AL97]. Most are or can be adapted to the structure of the objective function, here $\text{Cost}(\cdot|h_n)$. Some exploit the structure more directly (e.g. gradient methods for convex functions). Only in very simple cases can the minimum be found analytically (without search).

General maps Φ can be represented by/as programs for which variants of Levin search [Sch04, Hut05] and genetic programming are the major search algorithms. Decision trees/lists/grids are also quite powerful, especially rule-based ones in which logical expressions recursively divide domain \mathcal{H} into “true/false” regions [San08] that can be identified with different states.

Φ neighborhood relation. Most search algorithms require the specification of a neighborhood relation or distance between candidate Φ . A natural “minimal” change of Φ is splitting and merging states (state refinement and coarsening). Let Φ' split some state $s^a \in \mathcal{S}$ of Φ into $s^b, s^c \notin \mathcal{S}$

$$\Phi'(h) := \begin{cases} \Phi(h) & \text{if } \Phi(h) \neq s^a \\ s^b \text{ or } s^c & \text{if } \Phi(h) = s^a \end{cases}$$

where the histories in state s^a are distributed among s^b and s^c according to some splitting rule (e.g. randomly). The new state space is $\mathcal{S}' = \mathcal{S} \setminus \{s^a\} \cup \{s^b, s^c\}$. Similarly Φ' merges states $s^b, s^c \in \mathcal{S}$ into $s^a \notin \mathcal{S}$ if

$$\Phi'(h) := \begin{cases} \phi(h) & \text{if } \Phi(h) \neq s^a \\ s^a & \text{if } \Phi(h) = s^b \text{ or } s^c \end{cases}$$

where $\mathcal{S}' = \mathcal{S} \setminus \{s^b, s^c\} \cup \{s^a\}$. We can regard Φ' as being a neighbor of or similar to Φ .

Stochastic Φ search. Stochastic search is the method of choice for high-dimensional unstructured problems. Monte Carlo methods can actually be highly effective, despite their simplicity [Liu02]. The general idea is to randomly choose a neighbor Φ' of Φ and replace Φ by Φ' if it is better, i.e. has smaller Cost. Even if $\text{Cost}(\Phi'|h) > \text{Cost}(\Phi|h)$ we may keep Φ' , but only with some (in the cost difference exponentially) small probability. Simulated annealing is a version which minimizes $\text{Cost}(\Phi|h)$. Apparently, Φ of small cost are (much) more likely to occur than high cost Φ .

Context tree example. The Φ_k in Section 4 depended on the last k observations. Let us generalize this to a context dependent variable length: Consider a finite complete suffix free set of strings (= prefix tree of reversed strings) $\mathcal{S} \subset \mathcal{O}^*$ as our state space (e.g. $\mathcal{S} = \{0, 01, 011, 111\}$ for binary \mathcal{O}), and define $\Phi_{\mathcal{S}}(h_n) := s$ iff $o_{n-|s|+1:n} = s \in \mathcal{S}$, i.e. s is the part of the history regarded as relevant. State splitting and merging works as follows: For binary \mathcal{O} , if history part $s \in \mathcal{S}$ of h_n is deemed too short, we replace s by $0s$ and $1s$ in \mathcal{S} , i.e. $\mathcal{S}' = \mathcal{S} \setminus \{s\} \cup \{0s, 1s\}$. If histories $1s, 0s \in \mathcal{S}$ are deemed too long, we replace them by s , i.e. $\mathcal{S}' = \mathcal{S} \setminus \{0s, 1s\} \cup \{s\}$. Large \mathcal{O} might be coded binary and then treated similarly. The idea of using suffix trees as state space is from [McC96]. For small \mathcal{O} we have the following simple Φ -optimizer:

Φ Improve($\Phi_{\mathcal{S}}, h_n$)

```
[ Randomly choose a state  $s \in \mathcal{S}$ ;
  Let  $p$  and  $q$  be uniform random numbers in  $[0, 1]$ ;
  if  $(p > 1/2)$  then split  $s$  i.e.  $\mathcal{S}' = \mathcal{S} \setminus \{s\} \cup \{os : o \in \mathcal{O}\}$ 
  else if  $\{os : o \in \mathcal{O}\} \subseteq \mathcal{S}$ 
  then merge them, i.e.  $\mathcal{S}' = \mathcal{S} \setminus \{os : o \in \mathcal{O}\} \cup \{s\}$ ;
  if  $(\text{Cost}(\Phi_{\mathcal{S}'}|h_n) - \text{Cost}(\Phi_{\mathcal{S}}|h_n) > \log(q))$  then  $\mathcal{S} := \mathcal{S}'$ ;
] return ( $\Phi_{\mathcal{S}}$ );
```

6 Exploration & Exploitation

Having obtained a good estimate $\hat{\Phi}$ of Φ^{best} in the previous section, we can/must now determine a good action for our agent. For a finite MDP with known transition probabilities, finding the optimal action is routine. For estimated probabilities we run into the infamous exploration-exploitation problem, for which promising approximate solutions have recently been suggested [SL08]. At the end of this section I present the overall algorithm for our Φ MDP agent.

Optimal actions for known MDPs. For a known finite MDP $(\mathcal{S}, \mathcal{A}, T, R, \gamma)$, the maximal achievable (“optimal”) expected future discounted reward sum, called (Q) Value (of action a) in state s , satisfies the following (Bellman) equations [SB98]

$$Q_s^{*a} = \sum_{s'} T_{ss'}^a [R_{ss'}^a + \gamma V_{s'}^*] \quad \text{and} \quad V_s^* = \max_a Q_s^{*a} \quad (6)$$

where $0 < \gamma < 1$ is a discount parameter, typically close to 1. See [Hut05, Sec.5.7] for proper choices. The equations can be solved in polynomial time by a simple iteration process or various other methods [Put94]. After observing o_{n+1} , the optimal next action is

$$a_{n+1} := \arg \max_a Q_{s_{n+1}}^{*a}, \quad \text{where } s_{n+1} = \Phi(h_{n+1}) \quad (7)$$

Estimating the MDP. We can estimate the transition probability T by

$$\hat{T}_{ss'}^a := \frac{n_{ss'}^{a+}}{n_{s+}^{a+}} \quad \text{if } n_{s+}^{a+} > 0 \quad \text{and } 0 \quad \text{else.} \quad (8)$$

It is easy to see that the Shannon-Fano code of $s_{1:n}$ based on $P_{\hat{T}}(s_{1:n}|a_{1:n}) = \prod_{t=1}^n \hat{T}_{s_{t-1}s_t}^{a_{t-1}a_t}$ plus the code of the non-zero transition probabilities $T_{ss'}^a > 0$ to relevant accuracy $O(1/\sqrt{n_{s+}^{a+}})$ has length (2), i.e. the frequency estimate (8) is consistent with the attributed code length. The expected reward can be estimated as

$$\hat{R}_{ss'}^a := \sum_{r' \in \mathcal{R}} \hat{R}_{ss'}^{ar'} r', \quad \hat{R}_{ss'}^{ar'} := \frac{n_{ss'}^{ar'}}{n_{ss'}^{a+}} \quad (9)$$

Exploration. Simply replacing T and R in (6) and (7) by their estimates (8) and (9) can lead to very poor behavior, since parts of the state space may never be explored, causing the estimates to stay poor.

Estimate \hat{T} improves with increasing $n_{s^+}^a$, which can (only) be ensured by trying all actions a in all states s sufficiently often. But the greedy policy above has no incentive to explore, which may cause the agent to perform very poorly: The agent stays with what he *believes* to be optimal without trying to solidify his belief. Trading off exploration versus exploitation optimally is computationally intractable [Hut05, PVHR06, RP08] in all but extremely simple cases (e.g. Bandits). Recently, polynomially optimal algorithms (Rmax,E3,OIM) have been invented [KS98, SL08]: An agent is more explorative if he expects a high reward in the unexplored regions. We can “deceive” the agent to believe this by adding another “absorbing” high-reward state s^e to \mathcal{S} , not in the range of $\Phi(h)$, i.e. never observed. Henceforth, \mathcal{S} denotes the extended state space. For instance + in (8) now includes s^e . We set

$$n_{ss^e}^a = 1, \quad n_{s^e s}^a = \delta_{s^e s}, \quad R_{ss^e}^a = R_{max}^e \quad (10)$$

for all s, a , where exploration bonus R_{max}^e is polynomially (in $(1-\gamma)^{-1}$ and $|\mathcal{S} \times \mathcal{A}|$) larger than $\max \mathcal{R}$ [SL08].

Now compute the agent’s action by (6)-(9) but for the extended \mathcal{S} . The optimal policy p^* tries to find a chain of actions and states that likely leads to the high reward absorbing state s^e . Transition $\hat{T}_{ss^e}^a = 1/n_{s^+}^a$ is only “large” for small $n_{s^+}^a$, hence p^* has a bias towards unexplored (state,action) regions. It can be shown that this algorithm makes only a polynomial number of sub-optimal actions.

The overall algorithm for our Φ MDP agent is as follows.

Φ MDP-Agent(\mathcal{A}, \mathcal{R})

```
[ Initialize  $\Phi \equiv \epsilon$ ;  $\mathcal{S} = \{\epsilon\}$ ;  $h_0 = a_0 = r_0 = \epsilon$ ;
for  $n = 0, 1, 2, 3, \dots$ 
[ Choose e.g.  $\gamma = 1 - 1/(n+1)$ ;
Set  $R_{max}^e = \text{Polynomial}((1-\gamma)^{-1}, |\mathcal{S} \times \mathcal{A}|) \cdot \max \mathcal{R}$ ;
While waiting for  $o_{n+1}$   $\{\Phi := \Phi \text{Improve}(\Phi, h_n)\}$ ;
Observe  $o_{n+1}$ ;  $h_{n+1} = h_n a_n r_n o_{n+1}$ ;
 $s_{n+1} := \Phi(h_{n+1})$ ;  $\mathcal{S} := \mathcal{S} \cup \{s_{n+1}\}$ ;
Compute action  $a_{n+1}$  from Equations (6)-(10);
Output action  $a_{n+1}$ ;
] [ Observe reward  $r_{n+1}$ ;
```

7 Improved Cost Function

As discussed, we ultimately only care about (modeling) the rewards, but this endeavor required introducing and coding states. The resulted $\text{Cost}(\Phi|h)$ function is a code length of not only the rewards but also the “spurious” states. This likely leads to a too strong penalty of models Φ with large state spaces \mathcal{S} . The proper Bayesian formulation developed in this section allows to “integrate” out the states. This leads to a code for the rewards only, which better trades off accuracy of the reward model and state space size.

For an MDP with transition and reward probabilities $T_{ss'}^a$ and $R_{ss'}^{ar'}$, the probabilities of the state and reward

sequences are

$$P(s_{1:n}|a_{1:n}) = \prod_{t=1}^n T_{s_{t-1}s_t}^{a_{t-1}}, \quad P(r_{1:n}|s_{1:n}a_{1:n}) = \prod_{t=1}^n R_{s_{t-1}s_t}^{a_{t-1}r_t}$$

The probability of $\mathbf{r}|\mathbf{a}$ can be obtained by taking the product and marginalizing \mathbf{s} :

$$P_U(r_{1:n}|a_{1:n}) = \sum_{s_{1:n}} \prod_{t=1}^n U_{s_{t-1}s_t}^{a_{t-1}r_t} = \sum_{s_n} [U^{a_0 r_1} \dots U^{a_{n-1} r_n}]_{s_0 s_n}$$

where for each $a \in \mathcal{A}$ and $r' \in \mathcal{R}$, matrix $U^{ar'} \in \mathbb{R}^{m \times m}$ is defined as $[U^{ar'}]_{ss'} \equiv U_{ss'}^{ar'} := T_{ss'}^a R_{ss'}^{ar'}$. The right n -fold matrix product can be evaluated in time $O(m^2 n)$. This shows that \mathbf{r} given \mathbf{a} and U can be coded in $-\log P_U$ bits. The unknown U needs to be estimated, e.g. by the relative frequency $\hat{U}_{ss'}^{ar'} := n_{ss'}^{ar'} / n_{s^+}^a$. The $M := m(m-1)|\mathcal{A}|(|\mathcal{R}|-1)$ (independent) elements of \hat{U} can be coded to sufficient accuracy in $\frac{1}{2} M \log n$ bits. Together this leads to a code for $\mathbf{r}|\mathbf{a}$ of length

$$\text{ICost}(\Phi|h_n) := -\log P_{\hat{U}}(r_{1:n}|a_{1:n}) + \frac{1}{2} M \log n \quad (11)$$

In practice, M can and should be chosen smaller like done in the original Cost function, where we have used a restrictive model for R and considered only non-zero transitions in T .

8 Conclusion

I have developed a formal criterion for evaluating and selecting good “feature” maps Φ from histories to states and presented the feature reinforcement learning algorithm Φ MDP-Agent(). The computational flow is $h \rightsquigarrow \hat{\Phi} \rightsquigarrow (\hat{T}, \hat{R}) \rightsquigarrow (\hat{V}, \hat{Q}) \rightsquigarrow a$. The algorithm can easily and significantly be accelerated: Local search algorithms produce sequences of “similar” Φ , which naturally suggests to compute/update $\text{Cost}(\Phi|h)$ and the value function V incrementally. The primary purpose of this work was to introduce and explore Φ -selection on the conveniently simple (but impractical) unstructured finite MDPs. The results of this work set the stage for the more powerful Φ DBN model developed in the companion article [Hut09] based on Dynamic Bayesian Networks. The major open problems are to develop smart Φ generation and smart stochastic search algorithms for Φ^{best} , and to determine whether minimizing (11) is the right criterion.

References

- [AL97] E. H. L. Aarts and J. K. Lenstra, editors. *Local Search in Combinatorial Optimization*. Discrete Mathematics and Optimization. Wiley-Interscience, Chichester, England, 1997.
- [GDG03] R. Givan, T. Dean, and M. Greig. Equivalence notions and model minimization in Markov decision processes. *Artificial Intelligence*, 147(1–2):163–223, 2003.
- [Gor99] G. Gordon. *Approximate Solutions to Markov Decision Processes*. PhD thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 1999.

- [GP07] B. Goertzel and C. Pennachin, editors. *Artificial General Intelligence*. Springer, 2007.
- [Grü07] P. D. Grünwald. *The Minimum Description Length Principle*. The MIT Press, Cambridge, 2007.
- [HTF01] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. Springer, 2001.
- [Hut05] M. Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin, 2005. 300 pages, <http://www.hutter1.net/ai/uaibook.htm>.
- [Hut07] M. Hutter. Universal algorithmic intelligence: A mathematical top→down approach. In *Artificial General Intelligence*, pages 227–290. Springer, Berlin, 2007.
- [Hut09] M. Hutter. Feature dynamic Bayesian networks. In *Artificial General Intelligence (AGI'09)*. Atlantis Press, 2009.
- [KLC98] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101:99–134, 1998.
- [KS98] M. J. Kearns and S. Singh. Near-optimal reinforcement learning in polynomial time. In *Proc. 15th International Conf. on Machine Learning*, pages 260–268. Morgan Kaufmann, San Francisco, CA, 1998.
- [LH07] S. Legg and M. Hutter. Universal intelligence: A definition of machine intelligence. *Minds & Machines*, 17(4):391–444, 2007.
- [Liu02] J. S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer, 2002.
- [McC96] A. K. McCallum. *Reinforcement Learning with Selective Perception and Hidden State*. PhD thesis, Department of Computer Science, University of Rochester, 1996.
- [Put94] M. L. Puterman. *Markov Decision Processes — Discrete Stochastic Dynamic Programming*. Wiley, New York, NY, 1994.
- [PVHR06] P. Poupart, N. A. Vlassis, J. Hoey, and K. Regan. An analytic solution to discrete Bayesian reinforcement learning. In *Proc. 23rd International Conf. on Machine Learning (ICML'06)*, volume 148, pages 697–704, Pittsburgh, PA, 2006. ACM.
- [RN03] S. J. Russell and P. Norvig. *Artificial Intelligence. A Modern Approach*. Prentice-Hall, Englewood Cliffs, NJ, 2nd edition, 2003.
- [RP08] S. Ross and J. Pineau. Model-based Bayesian reinforcement learning in large structured domains. In *Proc. 24th Conference in Uncertainty in Artificial Intelligence (UAI'08)*, pages 476–483, Helsinki, 2008. AUAI Press.
- [RPPCd08] S. Ross, J. Pineau, S. Paquet, and B. Chaib-draa. Online planning algorithms for POMDPs. *Journal of Artificial Intelligence Research*, 2008(32):663–704, 2008.
- [San08] S. Sanner. *First-Order Decision-Theoretic Planning in Structured Relational Environments*. PhD thesis, Department of Computer Science, University of Toronto, 2008.
- [SB98] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.
- [Sch04] J. Schmidhuber. Optimal ordered problem solver. *Machine Learning*, 54(3):211–254, 2004.
- [SDL07] A. L. Strehl, C. Diuk, and M. L. Littman. Efficient structure learning in factored-state MDPs. In *Proc. 27th AAAI Conference on Artificial Intelligence*, pages 645–650, Vancouver, BC, 2007. AAAI Press.
- [SL08] I. Szita and A. Lörincz. The many faces of optimism: a unifying approach. In *Proc. 12th International Conference (ICML 2008)*, volume 307, Helsinki, Finland, June 2008.
- [SLJ⁺03] S. Singh, M. Littman, N. Jong, D. Pardoe, and P. Stone. Learning predictive state representations. In *Proc. 20th International Conference on Machine Learning (ICML'03)*, pages 712–719, 2003.