A Markovian Model for Carcinoid Tumors

Marcel A. J. van Gerven
Institute for Computing and Information Sciences
Radboud University Nijmegen
Heyendaalseweg 135
6525 AJ Nijmegen, The Netherlands

Francisco J. Díez
Department of Artificial Intelligence
UNED
Juan del Rosal, 16
28040 Madrid, Spain

Abstract

DLIMIDs, which can represent decision problems with partial observability and large horizons, constitute an alternative to POMDPs, or rather, they can be viewed as almost the same type of model with two different types of policies and, consequently, two paradigms of evaluation. In this paper, we describe a Markovian model for carcinoid tumors and discuss the effect of evaluating it as a DLIMID or as a POMDP.

1. Introduction

Limited-memory influence diagrams (LIMIDs), a formalism for decision-making under uncertainty (Lauritzen and Nilsson 2001), were later generalized to *dynamic LIMIDs* (DLIMIDs), a type of model that can represent both finite- and infinite-horizon Markov processes (van Gerven and Díez 2006; van Gerven et al. 2007). From the point of view of the representation of decision problems, DLIMIDs can be seen as an extension of factored POMDPs (Boutilier and Poole 1996), as they allow to have several decisions per time slice. Therefore, modeling issues are the same for POMDPs and DLIMIDs, with the exception that in DLIMIDs it is necessary in general to introduce *memory variables* (cf. Sec. 2.4) to circumvent the limited-memory restriction that characterizes this type of models.

This paper is structured as follows: In Section 2. we offer a review of the main properties of DLIMIDs, with special emphasis on stationary DLIMIDs, which are the basis of infinite-horizon DLIMIDs. In Section 3. we present three algorithms that can be used for evaluating infinite-horizon DLIMIDs. The second and third of these algorithms, namely, SRU and SA, have been used to evaluate the medical DLIMID presented in Section 4., which constitutes the core of this paper: in Section 4.1 we describe the construction of the model and in Section 4.2 the process and the results of evaluating it. In Section 5. we analyze the similarities and differences between DLIMIDs and POMDPs and discuss the lines open for future research, and we conclude in Section 6.

ICAPS 2010 POMDP Practitioners Workshop, May 12, 2010, Toronto, Canada.

2. Background: DLIMIDs

2.1 Definition of DLIMID

A *dynamic limited memory influence diagram* (DLIMID) is a tuple $\mathcal{L} = (h, \mathbf{C}, \mathbf{D}, \mathbf{U}, G, P, U, \gamma)$. In this tuple, $h \in \mathbb{N}^+ \cup \{\infty\}$, is the *horizon*. $\mathbf{C}$ is a set of chance variables. $\mathbf{D}$ is a set of decision variables (actions). We assume in this paper that all the variables in $\mathbf{C}$ and $\mathbf{D}$ are discrete. $\mathbf{U}$ is the set of utility variables (rewards). Variables are indexed by time: $t \in \{1, \ldots, h\}$.

G is an acyclic directed graph containing a node for each variable X^t; for this reason we will use the terms variable and node indifferently. Links to the past, i.e., of the form $X^t \rightarrow Y^{t'}$ with $t > t'$, are not allowed.

P is a set of conditional distributions, such that for each variable X^t in $\mathbf{C}$ and each configuration $pa(X^t)$ of the parents of X^t in the graph—that we call $Pa(X^t)$—there is a conditional probability distribution of the form $P(x^t|pa(X^t))$. The join probability distribution for the variables in $\mathbf{C}$, conditioned on the decisions, is

$$P(\mathbf{c}|\mathbf{d}) = \prod_{C \in \mathbf{C}} P(c|pa(C)) \tag{1}$$

In this equation, c is the result of projecting configuration $(\mathbf{c}, \mathbf{d})$ onto C and $pa(C)$ is the result of projecting $(\mathbf{c}, \mathbf{d})$ onto $Pa(C)$. The next equations in this paper must be understood in the same way.

Similarly, the set U contains, for each node U^t, a function $u(pa(U^t))$, where $pa(U^t)$ is a configuration of the parents of U^t in the graph.

Finally, $\gamma \in [0, 1]$ is the discount factor. Usually $\gamma < 1$, which implies that delayed rewards are less valuable to the decision maker.

The subgraph that contains all the nodes X^t having the same temporal index t and the links between them is known as the t-th slice. We use $\mathbf{C}^t$, $\mathbf{D}^t$, and $\mathbf{U}^t$ to denote the chance, decision, and utility nodes at the t-th slice, respectively. If there is a $k \in \mathbb{N}^+$ such that every link $X^t \rightarrow Y^{t'}$ satisfies that $t' - t \leq k$, we say that the DLIMID is *of k-th order*. In practice, it is usual to work with first-order DLIMIDs ($k = 1$).

2.2 Stationary DLIMIDs

A DLIMID of order k is stationary if the structure of the graph, the probabilities, and the utilities repeat themselves

after the k-th slice, i.e., for each $t > k+1$,

- there is a link $X^t \rightarrow Y^t$ if and only if there is a link $X^{k+1} \rightarrow Y^{k+1}$,
- $P(x^t|pa(X^t)) = P(x^{k+1}|pa(X^{k+1}))$, and
- $U^t(pa(U^t)) = U^{k+1}(pa(U^{k+1}))$.

Given that the slices repeat themselves after the k-th, i.e., that all subsequent slices are the same as the $(k+1)$-th, we can have a compact representation in which we only represent the first $k+1$ slices. If the horizon h of the DLIMID is finite, we can unroll it completely by cloning the last slice $h-(k+1)$ times. In the case of a first-order DLIMID, which is the most common case, stationarity means that only the first slice differs from the others and, hence, we can represent the DLIMID by only the first two time slices; this compact form of DLIMIDs is very similar to the standard representation of (factored) MDPs and POMDPs. Alternatively, a DLIMID can be represented by a *temporal LIMID* (TLIMID), which consists of a prior model $\mathcal{L}_0$, representing the initial state of the system, and a transition model $\mathcal{L}_t$, representing the evolution over time; this representation is very similar to the representation of dynamic Bayesian networks using 2TBNs (Murphy 2002, Sec. 2.2).[1]

2.3 Policies and strategies

A link $X \rightarrow D$, where X is a chance variable or a decision and D is a decision, means that the value of X is known when making decision D. For this reason, the parents of D are also known as its *informational predecessors*.[2] A *stochastic policy* for a decision $D \in \mathbf{D}$ is defined as a distribution $P_D(d|pa(D))$ that maps each configurations of $Pa(D)$ to a distribution over alternatives (i.e., actions) for D; thus, $P_D(d|pa(D))$ is the probability of selecting alternative (action) d when the information available is $pa(D)$. If P_D is degenerate (consisting of ones and zeros only) then we say that the policy is *deterministic*.

A *strategy* is a set of policies $\Delta = \{P_D : D \in \mathbf{D}\}$, which induces the following joint distribution over the variables in $\mathbf{C} \cup \mathbf{D}$:

$$P_\Delta(\mathbf{c}, \mathbf{d}) = P(\mathbf{c} : \mathbf{d}) \prod_{D \in \mathbf{D}} P_D(d|pa(D)) \; . \tag{2}$$

We define

$$\mathcal{U}_t(\mathbf{c}^t, \mathbf{d}^t) = \sum_{U^t \in \mathbf{U}^t} u(pa(U^t)) \; , \tag{3}$$

which is the utility for the t-th time slice for a configuration $(\mathbf{c}^t, \mathbf{d}^t)$ of $\mathbf{C}^t \cup \mathbf{D}^t$. The *utility* for a configuration $(\mathbf{c}, \mathbf{d})$ of $\mathbf{C} \cup \mathbf{D}$ is defined as

$$\mathcal{U}(\mathbf{c}, \mathbf{d}) = \sum_{t=0}^{h} \gamma^t \mathcal{U}_t(\mathbf{c}^t, \mathbf{d}^t)$$

and the *expected utility* of a strategy Δ as

$$\mathrm{EU}(\Delta) = \sum_{\mathbf{c}} \sum_{\mathbf{d}} P_\Delta(\mathbf{c}, \mathbf{d}) \cdot \mathcal{U}(\mathbf{c}, \mathbf{d}) \; . \tag{4}$$

The *optimal strategy* is the one that maximizes the expected utility:

$$\Delta^* = \arg\max_{\Delta} \; \mathrm{E}_\Delta(\mathcal{U}) \; . \tag{5}$$

2.4 Memory variables

An important shortcoming of the limited-memory assumption is that it may lead to neglecting important findings obtained in the past. For instance, let us consider the case of a patient who showed an allergic reaction to a drug in the past. If we forget this fact and administer the same drug again, we may cause more harm than benefit. Apparently, a situation like this cannot be addressed with stationary DLIMIDs, because a stationary DLIMID is, by definition, of k-th order, and therefore it cannot have an informational link $F^t \rightarrow D^{t'}$ with $t' - t > k$; i.e., when making decision D at time t' we cannot take into account the value of finding F^t in a distant past. In particular, in a first-order DLIMID, the informational predecessors of a decision variable D^t can only occur in time-slices t or $t-1$. Thus, the limited-memory assumption implies that observations made earlier in time will not be taken into account and, as a result, states that are qualitatively different can appear the same to the decision maker, which leads to suboptimal decisions.

However, there is a way of circumventing this problem: to introduce in the DLIMID *memory variables* that represent a summary of the observed history. In the above example we may have a binary variable M indicating whether this patient has ever had an allergic reaction to that drug or not. Memory variables are not necessarily associated to chance variables: we can also "remember" whether we have made a decision (chosen a particular option) at any moment in the past. This way, in a k-th order stationary LIMID, we can add a memory variable M, of type chance ($M \in \mathbf{C}$), whose parents are in the previous k slices. The domain of M and the conditional probability tables for each M^t will depend on the meaning that the knowledge engineer assigns to this variable.

For example, when variable M is intended to "remember" the last k' values of finding F, with $k' \leq k$, the parents of M^t will be $\{F^{t-k'-1}, \ldots, F^t\}$. If variable F represents the result of a test, it may take on three values, $\{u, n, p\}$, where u stands for *unobserved* (test not performed), n for *negative*, and p for *positive*, and we wish to remember the last three values of F ($k' = 3$) then the domain of M will be $\Omega_M = \{u, n, p\} \times \{u, n, p\} \times \{u, n, p\}$, i.e., M will have 27 possible values. The value of M^t represented by upn means that the test was not performed at time $t-2$, gave a positive result at $t-1$, and a negative result at t. Alternatively, if F were binary and the purpose of M were to "remember" whether F has ever taken a positive value, the parents of M^t would be M^{t-1} and F^t and the conditional probability table for M^t would be

$$P(+m^t|m^{t-1}, f^t) = \begin{cases} 1 & \text{if } m^{t-1} = +m^{t-1} \vee f^t = +f^t \\ 0 & \text{otherwise} \end{cases}$$

[1] 2TBNs were introduced in (Boutilier, Dean, and Hanks 1996). The reason for representing DLIMIDs as TLIMIDs is the possibility of converting it into 2TBNs, as explained in Section 3.1.

[2] In a LIMID and a DLIMID, the informational predecessors or a decision are only its parents, while in an influence diagram the informational predecessors of a decision are its parents and the parents of any decision made before D—see (Lauritzen and Nilsson 2001; van Gerven et al. 2007).

The advantage of memory variables is that they allow us to keep information from the distant past and, at the same time, to fulfill *the Markov condition*, which states that the future is independent of the past given the present, i.e., the current state of the system. This is just the function of memory variables, to augment the state of the system by including information from the past.

It is not always easy to decide the number of memory variables that should be introduced in a DLIMID and the meanings of each one. We cannot think of any algorithmic procedure to make this decision. It is the knowledge engineers who, using the available domain knowledge, their common sense, and their experience, should make the best modeling decisions.

3. Evaluation of infinite-horizon DLIMIDs

As finite-horizon DLIMIDs are a particular case of LIMIDs, they can be evaluated with the methods proposed in (Lauritzen and Nilsson 2001). On the contrary, infinite-horizon DLIMIDs, which contain an infinite number of nodes, cannot be solved with those methods. It is possible, however, to use any of the three algorithms proposed in (van Gerven and Díez 2006), that we describe briefly below. Given that all the three need to compute the expected utility of strategies, we first present a method for this task.

3.1 Computing the expected utility of a strategy

In order to compute the expected utility for a TLIMID, we resort to an indirect approach, which combines the value iteration algorithm for MDPs (Bellman 1957) with the transformation of influence diagrams (IDs) into Bayesian networks (BNs) (Cooper 1988).[3]

Given a strategy Δ, Cooper's method converts an ID with chance variables $\mathbf{C}$ and decisions $\mathbf{D}$ into a BN. Each decision variable D in the ID is converted into a chance variable in the BN whose conditional probability is given by the policy for D in Δ, namely $P_D(d|pa(D))$. Additionally, each utility node U in the ID is converted into a binary variable in the BN by means of a linear transformation such that the minimum utility is mapped onto 0 and the maximum onto 1, and this way each utility is converted into a probability. This way it is possible to compute the expected utility $\mathrm{EU}(\Delta)$ given a strategy Δ by computing the probability of $+u$ in the BN and transforming back this probability into a utility (Cooper 1988).

The same idea can be applied to LIMIDs and DLIMIDs. We denote by $B(\mathcal{L}, \Delta)$ the conversion of a LIMID $\mathcal{L}$, given a strategy Δ, into a Bayesian network $\mathcal{B}$. Given Δ, we may convert a TLIMID $(\mathcal{L}_0, \mathcal{L}_t)$ into the pair $(\mathcal{B}_0, \mathcal{B}_t)$, which is a dynamic Bayesian network, where $\mathcal{B}_0 = B(\mathcal{L}_0, \Delta_0)$ and $\mathcal{B}_t = B(\mathcal{L}_t, \Delta_t)$; $\mathcal{B}_t$ is a two time-slice Bayesian network (2TBN). This dynamic Bayesian network can be evaluated with the methods proposed in (Murphy 2002); in fact, in our experiments we used the BNT package (Murphy 2001).

In order to compute an approximation to the expected utility for a (first-order) DLIMID given Δ, we assume that the DLIMID can be represented by a TLIMID $(\mathcal{L}_0, \mathcal{L}_t)$ and Δ can be expressed as a pair (Δ_0, Δ_t), where Δ_0 is the initial strategy and Δ_t is a *stationary* strategy that does not depend on t. Recall that, for infinite-horizon Markov decision processes (Ross 1983), the optimal strategy is deterministic and stationary. However, in the partially observable case, we can only expect to find approximations to the optimal strategy by using memory variables that represent part of the observational history. The approximation $\mathrm{EU}^{\kappa}(\Delta)$ to the expected utility is made by computing the discounted expected utility ($\gamma < 1$) using $(B(\mathcal{L}_0, \Delta_0), B(\mathcal{L}_t, \Delta_t))$ for a finite number of time-slices κ. Here, κ may be chosen based on the problem characteristics, or based on some error criterion ϵ. For instance, by choosing

$$\kappa = \log_{\gamma}(\epsilon(1-\gamma)/2u_{\max}),$$

where $u_{\max}$ stands for the maximum utility obtainable during one time-slice, we ensure that at most $\epsilon/2$ error is introduced into the approximation (Ng and Jordan 2000).

3.2 Single policy updating

The single policy updating (SPU) method proposed in (Lauritzen and Nilsson 2001) for LIMIDs can be adapted to infinite-horizon DLIMIDs by computing the expected utility of the strategies with the method presented above. This way we can obtain a strategy that is locally optimal, i.e., it cannot be improved by changing just one policy.

In the case of LIMIDs, which include finite-horizon DLIMIDs as a particular case, locally optimal policies can be found by optimizing each single rule independently of the others, such that we need to evaluate km^r different policies at each decision variable D, where k denotes the cardinality of Ω_D, and r is the number of informational predecessors of D, assuming that the cardinality of $\Omega_{D'}$ equals m for all $D' \in pa(D)$. However, in case of infinite-horizon DLIMIDs, the optimal rule for a certain scenario at time t depends on the policies applied at future times, which leads to a coupling of the rules. The number of policies that need to be evaluated at each decision variable D therefore grows as $k^{(m^r)}$, which makes SPU unfeasible even for many small DLIMIDs.

3.3 Single rule updating

Given the impossibility of iterating over the set of policies in DLIMIDs, we proposed in (van Gerven and Díez 2006) a hill-climbing method, called *single rule updating* (SRU). Instead of exhaustively searching over all possible policies for each decision variable, we try to increase the expected utility by local changes to the decision rules within the policy; i.e., at each step we change one decision-rule within the policy, accepting the change when the expected utility increases. Similarly to SPU, we keep iterating until there is no further increase in the expected utility. Using SRU, we decrease the number of policies that need to be evaluated in each *local* cycle for a decision node to only km^r (as when applying SPU for LIMIDs), albeit at the expense of replacing the exhaustive search by a hill-climbing strategy, which further increases the risk of ending up in a local maximum, and having to run local cycles until convergence.

[3] A similar transformation was use by (2006) for solving MDPs.

3.4 Simulated annealing

In order to improve upon the strategies found by SRU, we resort to *simulated annealing* (SA) (Kirkpatrick, Gelatt Jr, and Vecchi 1983), which is a heuristic search method that tries to avoid getting trapped into local maximum solutions found by hill-climbing techniques. SA chooses candidate solutions by looking at neighbors of the current solution as defined by a *neighborhood function*. Local maxima are avoided by sometimes accepting worse solutions according to an *acceptance function*. In (van Gerven and Díez 2006), we used the acceptance function

$$\mathrm{P}(a(\Delta') = \text{yes} \mid eu, eu', t) = \begin{cases} 1 \text{ if } eu' > eu \\ \mathrm{e}^{\frac{eu'-eu}{T(t)}} \text{ otherwise} , \end{cases}$$

where $a(\Delta')$ stands for the acceptance of the proposed strategy Δ', $eu' = \mathrm{EU}^{\kappa}(\Delta')$, $eu = \mathrm{EU}^{\kappa}(\Delta)$ for the current strategy Δ, and T represents the temperature in an *annealing schedule* defined as

$$T(t+1) = \alpha \cdot T(t)$$

where $T(0) = \beta$ with $\alpha < 1$ and $\beta > 0$. The annealing schedule ensures that initially a random search through the space of strategies is performed, which gradually changes into a hill-climbing search. We refer to (Eglese 1990) for a discussion about choices that can be made for SA parameters α and β. With respect to strategy finding in dynamic LIMIDs, we propose an initial simulated annealing scheme and a subsequent application of SRU in order to greedily find a local maximum solution.

An experiment with a small LIMID for a hypothetical problem, using 20 different initial strategies Δ^0, showed that in most of the cases SA yields better strategies than SRU (van Gerven and Díez 2006). It also showed that in general it is possible to reach an additional increase in the expected utility by applying SRU after SA. We will see in Section 4.2 that SA gives better results than SRU also for the medical model presented in this paper.

4. A DLIMID for treating carcinoid tumors

We have applied DLIMIDs to the problem of treatment selection for high-grade carcinoid tumor patients.[4] A carcinoid tumor is a type of neuroendocrine tumor predominantly found in the midgut and is normally characterized by the production of excessive amounts of biochemically active substances, such as serotonin.[5] The dynamic decision problem then is whether or not to administer chemotherapy at each decision moment.

In order to solve this problem, we have constructed a DLIMID as a model of high-grade carcinoid tumor pathophysiology in collaboration with an expert physician. Our aim is to validate if the treatment strategy that is used in practice will also be found by a TLIMID as a formal domain model, thereby confirming the quality of the employed strategy. If the policy returned by the DLIMID differs from that applied by human doctors, we should perform a sensitivity analysis, first parametric and then structural, to find out the cause of the disagreement. This would lead to a refinement of the model, for instance, by adding other relevant variables, or to the conclusion that the strategy currently used in practice is not optimal and should be changed.

4.1 Description of the model

Figure 1 depicts the structure of the model, where shaded variables are observable. Since patients return to the clinic for follow-up every three months, we assume that each time slice represents patient status at three-month intervals, at which time treatment can be adjusted.

In the model, the patient's *general health status* (ghs) is of central importance. In oncology, one way to estimate the general health status is by means of the *performance status*, which is distinguished into *normal* (0), *mild complaints* (1), *ambulatory* (2), *nursing care* (3), *intensive care* (4), and *death* (5). Modeling the evolution of ghs is a non-trivial task; it depends on the current general health status, and on patient properties such as age and gender, since these are risk factors that may lead to patient death due to causes other than the disease. Furthermore, ghs is influenced by the tumor mass (mass) and the treatment strategy. Tumor mass has a negative influence on the general health status and is the first cause of death for patients with high-grade carcinoid tumors. Hepatic metastases normally account for the majority of the tumor mass and the primary tumor does not normally contribute significantly to the tumor mass.

Most patients with high-grade tumors have extensive metastatic disease when admitted to the hospital. If there is no tumor response due to treatment then the physician estimates an exponential growth in tumor mass: $x(t) = x_0 \cdot \mathrm{e}^{1.41t}$. If there is a tumor response due to treatment then we will see a reduction in tumor mass according to Table 1. If no chemotherapy is given, then we use nt (no treatment) to denote the absence of tumor response. Finally, if $\mathsf{ghs} = dead$ then there is no change in tumor mass.

Chemotherapy (chemo), with states $\{none, reduced, standard\}$, is the only available treatment to reduce tumor growth. We use treathist with states $\{0, 1, 2, 3\}$ as a memory variable to represent the patient's relevant treatment history, such that $\mathsf{treathist} = i$ represents continued chemotherapy over the past i trimesters. Reductions in tumor mass due to chemotherapy are often described by means of the WHO criteria for tumor response (bmdhist), as defined in Table 1. If a patient has been treated previously, then the effectiveness of treatment changes. In case $\mathsf{resp}(t-1)$ is either pr or cr, then it is assumed that continued chemotherapy will lead to stable disease (sd). If, on the other hand, $\mathsf{resp}(t-1) = sd$ then continued chemotherapy will become less effective. Even when chemotherapy is discontinued, we expect some residual effect of chemotherapy due to the knock-out effect on tumor-cells. It is estimated that after three months, the effect of chemotherapy is at 70% of its normal effectiveness.

[4] Although a patient's life-span is bounded, it is useful to describe a treatment selection problem as an infinite-horizon POMDP, where the process has an exponentially decreasing but non-zero probability of continuing at each time slice.

[5] Further details and medical references used to build this model can be found in (van Gerven et al. 2007).

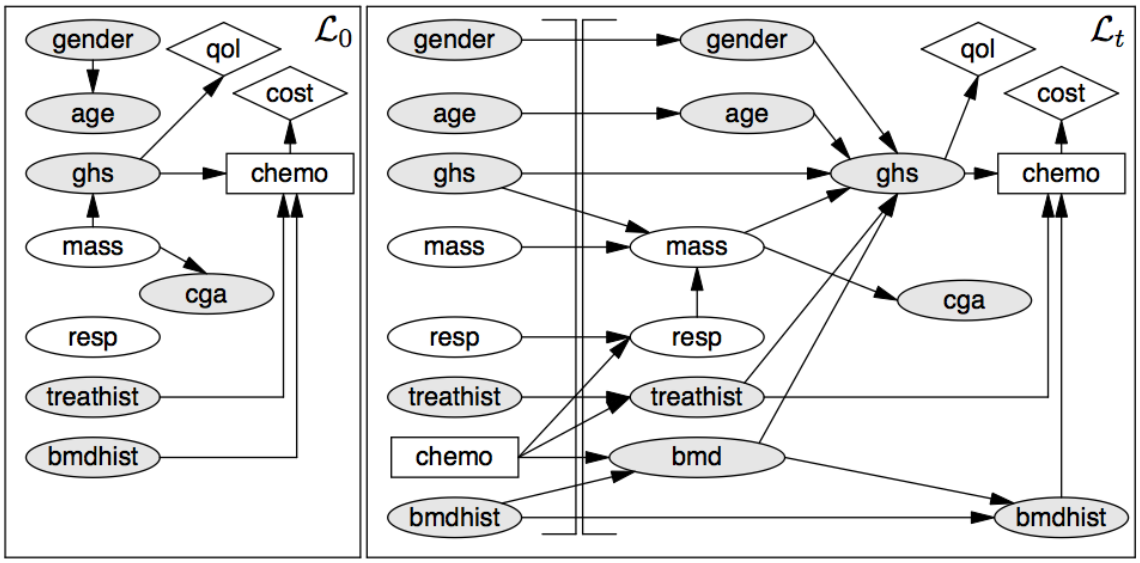

Figure 1: TLIMID for high-grade carcinoid tumor pathophysiology. Chemotherapy has not yet been given at the initial time, which renders the tumor response to chemotherapy (RESP) independent of all other variables at $\mathcal{L}_0$.

Table 1: The WHO criteria for tumor response.

Tumor Response	**Criteria**
Complete remission (cr)	Disappearance of all lesions.
Partial remission (pr)	More than 50% decrease in tumor mass.
Progressive disease (pd)	More than 25% increase in lesions, or a new lesion.
Stable disease (sd)	Neither pr nor pd.

Note that chemotherapy may have both positive and negative effects on general health status. Positive due to reductions in tumor mass, and negative due to severe bone-marrow depression (bmd) and damage associated with prolonged chemotherapy. Severe bone-marrow depression may cause patient death due to associated neutropenic sepsis and/or internal bleeding. We use bmdhist with states $no\text{-}bmd$ and bmd as a memory variable to represent whether or not the patient has experienced bmd in the past. No decision variable has been defined that determines whether or not to assess bmd status, since this status is assumed to be given by routine laboratory tests.

The global utility is defined as a discounted additive combination of the *quality of life* (qol) and the cost of chemotherapy (cost):

$$\mathcal{U} = \sum_{t=0}^{n} \gamma^t \left(\mathsf{qol}^t(\mathsf{ghs}^t) - \mathsf{cost}^t(\mathsf{chemo}^t) \right) .$$

Our measure of quality of life is based on *quality-adjusted life-years*, or QALYs (Weinstein and Stason 1977), which simultaneously captures gains in quantity and quality of life. QALYs are computed by multiplying a *quality-adjustment weight* for each health state by the discounted time spent in this state. We associate quality-adjustment weights with the states of ghs based on the *quality of well-being* scale, taking into account that each time slice stands for a three-month period (Table 2). We have associated a small economical cost with chemotherapy, that is regarded insignificant compared with the benefit gained in terms of quality of life.

Table 2: Quality-adjustment weights for ghs.

ghs	0	1	2	3	4	5
weight	0.214	0.184	0.168	0.121	0.109	0.000

In our model, we used a discounting factor of 0.95 as suggested in the literature, such that the three-month discount factor is $\gamma \approx 0.987$. The expected utility then becomes:

$$\mathrm{EU}(\Delta) = \mathrm{E}_\Delta \left(\sum_{t=0}^{n} \gamma^t \mathsf{qol}^t(\mathsf{ghs}^t)) \right) - \mathrm{E}_\Delta \left(\sum_{t=0}^{n} \gamma^t \mathsf{cost}^t(\mathsf{chemo}^t)) \right) . \quad (6)$$

The first term in Eq. (6) is the discounted quality-adjusted life expectancy (QALE) and the second term is the discounted expected cost of treatment. The goal of our model then is to find a policy for chemotherapy that maximizes this expression.

The physician has indicated that the informational predecessors of chemo are given by ghs, treathist and bmdhist, where both treathist and bmdhist are used as memory variables within the model. Changes in treatment history are specified as follows. Given that chemo equals *standard* or *reduced*, treathist increases from x to $x + 1$ until the maximum of 3 is reached, and given that chemo $= none$, treathist decreases from x to $x - 1$ until the minimum of 0 is reached. In order to represent whether or not a patient has ever experienced bone-marrow depression, we assume that bmdhist$' = no\text{-}bmd$ if bmd$' = no$ and bmdhist $= no\text{-}bmd$. Otherwise, it is assumed that bmdhist$' = bmd$. Note that in this case, we represent memory of infinite length by restricting ourselves to the event whether or not severe bone-marrow depression has occurred. Contrary to what may be expected, cga, as a correlate of tumor mass, is not regarded to be an informational predecessor by the physician since a patient who is known to have a high-grade carcinoid tumor is treated as often as possible, irrespective of the current state of the tumor.

4.2 Evaluation of the model

Our aim was to find a treatment strategy for high-grade carcinoid tumors using the developed model and the described algorithms. We applied the simulated annealing scheme, followed by SRU, as suggested in Section 3.4. Since the informational predecessors are equal for chemo in $\mathcal{L}_0$ and $\mathcal{L}_t$, we assumed that $\Delta_0 = \Delta_t$. We use Δ to denote this strategy, containing a stationary policy for chemo. The number of possible policies for chemo is then given by:

$$\Omega_{\text{CHEMO}}^{\Omega_{\text{ghs}} \cdot \Omega_{\text{treathist}} \cdot \Omega_{\text{bmdhist}}} = 3^{5 \cdot 4 \cdot 2} \approx 1.22 \cdot 10^{19} .$$

Note that single policy updating would require an exhaustive search through this space of possible policies, which is clearly computationally intractable. For our model, we have used $\kappa = 40$ as the stopping criterion for the approximation to the expected utility, based on the observation that ten-year survival is rarely attained for this aggressive form of cancer. After some initial experiments, we have chosen $\alpha = 0.995$, $\beta = 0.5$ and $T_{\min} = 1.225 \cdot 10^{-3}$ for the simulated annealing parameters.

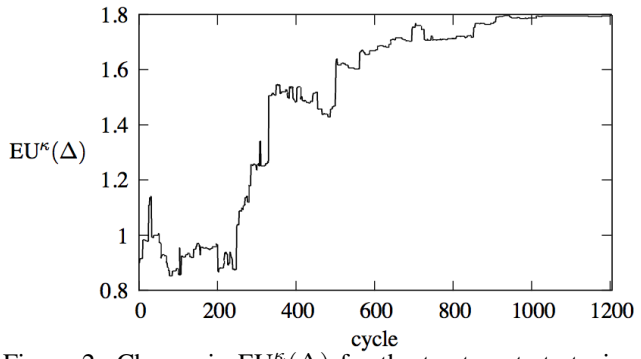

Figure 2: Change in EU$^{\kappa}(\Delta)$ for the treatment strategies, selected during simulated annealing, followed by single rule updating at the end.

The SA algorithm was repeated twenty times, starting from random initial strategies. It consistently found the same treatment strategy Δ, with an expected utility of 1.795. Figure 2 shows the subsequent values of EU$^{\kappa}(\Delta)$ of the strategies found during one of these experiments. The figure depicts how the initial explorative behavior of the simulated annealing scheme gradually changes into a hill-climbing strategy. The application of single rule updating after the simulated annealing phase caused a small increase in expected utility from 1.795 to 1.798. For this particular example, the solution found by simulated annealing (followed by SRU) was the same as the solution found by SRU alone, although this does not hold in general.

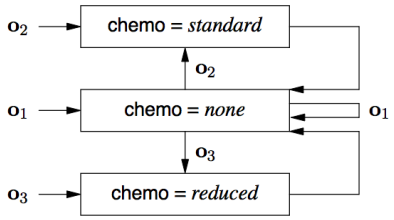

Figure 3: Policy graph for the best strategy that was found by simulated annealing, where $\mathbf{o}_1$ = ghs > 3, $\mathbf{o}_2$ = ghs $\leq 3 \wedge$ bmdhist $=$ *no-bmd* and $\mathbf{o}_3$ = ghs $\leq 3 \wedge$ bmdhist $=$ *bmd*.

A strategy for a DLIMID can be represented by a *policy graph* (Meuleau et al. 1999); a finite state controller that depicts state transitions, where states represent actions and transitions are induced by observations. The policy graph for this medical problem is shown in Figure 3. We can see that the policy only depends on the patient's general health status (ghs) and on the antecedents of severe bone-marrow depression (bmd), summarized in the memory variable bmdhist, while it does not depend on the treatments applied previously, as summarized in variable treathist, which implies that this memory variable was not necessary.

The experts collaborating with us confirmed that this is the policy that they apply in clinical practice.

5. Discussion

5.1 Knowledge engineering of DLIMIDs and POMDPs

DLIMIDs and factored POMDPs are very similar in their structure: both contain chance variables (some of which are observable and some are not), decisions, and utility nodes. An advantage of DLIMIDs is that they may include several decisions per time slice; for instance, whether to perform a test or not, and whether to apply a therapy or not, possibly depending on the result of the test—see a toy example in (van Gerven and Díez 2006; van Gerven et al. 2007). POMDPs cannot easily represent this kind of problems. In many domains, it may be an important reason for building a DLIMID instead of a POMDP.

Given that the DLIMID for carcinoid tumors presented above only contains one decision per time slice, it can be

viewed and evaluated as a POMDP, which has two unobservable variables (the mass of the tumor, mass, and the response to the treatment, resp), seven observable variables (the seven shaded ovals on the right part of Figure 1), one decision (chemo), and two rewards (quality of life, qol, and the cost of the chemotherapy).

The main difference would be in the returned policies. In a DLIMID, the policy for a decision depends on its informational predecessors, which are a subset of the **observable** variables.[6] Therefore, if all the informational predecessors are discrete, then the domain of the policy is discrete, as in the case of fully observable MDPs.[7] In our medical model, the informational predecessor for the only decision, chemo, are the global health state (ghs) and two memory variables, bmdhist and treathist, that represent the antecedents of severe bone-marrow depression and the previous treatments; however, after the evaluation it turned out that the policy does not depend on treathist. Reciprocally, we might have included more variables in the domain of the policy, as we discuss below.

On the contrary, in the most usual way of solving POMDPs, the policy depends on the belief states of the **unobservable** variable and the solution is expressed as a partition of the belief space. In a factored POMDP, the policy may depend on both the belief state of unobservable variables and on the configuration of observable variables, which makes the solution much more difficult to communicate. This is an advantage of evaluating this model as a DLIMID, because the policy shown in Figure 3 is very clear and intuitive.

However, an alternative way to solve a POMDP is to build a finite-state controller (Hansen 1998; Meuleau et al. 1999). This way, the algoritms presented in this paper can be understood as a way of building finite-state controllers for factored POMDPs, thus reducing the difference between DLIMIDs and POMDPs.

We would like to conclude this section with two remarks. First, we said at the beginning of the paper that POMDPs do not need memory variables, because they store the history of observations implicitly in the belief state of non-observable variables. Then, one might think that when evaluating the model for carcinoid tumors as a POMDP we can get rid of the two memory variables bmdhist and treathist without affecting the performance of the model. However, this is not true, because in this case the model would not take into account the influence of chemotherapy into the global health state, which is mediated by variables bmd and treathist—see Figure 1. If we removed bmdhist and treathist from the POMDP, we would need to add unobservable variables to represent two features specific of each patient: the response to chemotherapy and the probability of suffering from severe bone-marrow depression. Therefore, building POMDPs avoids the difficulty to make modeling decisions about memory variables, but may require to include additional unobservable variables in the model; choosing the domain and obtaining the conditional probabilities for such variables may be as difficult as introducing memory variables.

Second, when building a DLIMID we have to decide which variables are allowed to make part of the domain of each decision, i.e., to be "observed"—or, better, "remembered"—by the decision maker at each moment. In our DLIMID the decision about chemotherapy was based on three observations: ghs, bmdhist, and treathist. We can see that in Figure 3 there is an information link from each of them to the decision node therapy. We did not draw links from age or gender to therapy because we assumed, using domain knowledge, that the treatment policy does not depend on the patient's gender and age. But perhaps it was a wrong assumption. This difficulty of selecting the informational predecessors of each decision in a DLIMID is avoided when building a POMDP, but, on the other hand, DLIMIDs give us more flexibility, because the fewer variables we include for each decision, the faster the algorithms will be and the easier the policies will be to understand for human users. This way, we can reach a trade-off between computational complexity, intuitiveness of the solutions, and higher expected utilities, while in POMDPs we do not have this degree of freedom.

5.2 Future work

One of the lines for future research is the development of other algorithms for DLIMIDs that avoid as much as possible getting trapped in local maxima without incurring excessive computational costs. A possibility would be the application of well-known optimization methods such as genetic algorithms or tabu search.

Another line of research is related with the use of memory variables. As mentioned above, the evaluation of the DLIMID showed that the policy does not depend on treathist, but perhaps if we had modeled this variable such that it summarized previous treatments in a different way, it would be relevant for the policy and might have led to an increased expected utility. In same way, the inclusion of more variables in the policy for decision chemo, such as age and gender, might also increase the expected utility. It is a possibility that should be tested empirically.

We might also compare the expected utility obtained when evaluating the carcinoid model as a DLIMID with that obtained when evaluating it as a POMDP. There are several algorithms for DLIMIDs and many for POMDPs, based on different approximations and heuristic techniques. Thus, we have a chance to trade-off reduction in expected utility in order to save time and space.

It would also be interesting to analyze how intuitive the policies returned by each model are for human users. For our carcinoid model, the policy obtained when evaluating it as a DLIMID (see Fig. 3) is very easy to understand, but other policies containing more variables in their domains might at-

[6] Please note that the algorithms for the evaluation of DLIMIDs allow that the policies depend on unobservable variables. However, such policies would be useless, because it would impossible to make a decision based on information that is not available.

[7] In this sense, DLIMIDs are similar to factored fully observable MDPs (FOMDPs), but the latter assume that all the variables are observable and that policies depend on all the variables. Therefore, DLIMIDs are a generalization of FOMDPs and permit to solve a much wider range of problems.

tain higher expected utilities, as mentioned above; this is another possible trade-off: intuitiveness vs. quality of the policy. On the other hand, we believe that these DLIMID policies, having discrete domains, would be much more intuitive for human users than the POMDP policies based on partitions of the belief state, but it is something that we should prove empirically. We should also compare our algorithms with those proposed for building finite-state controllers that represent near-optimal policies for POMDPs.

There is, finally, another research line also related with the methods that build controllers for POMDPs: how to adapt them to DLIMIDs, which may contain several decisions per time slice.

6. Conclusion

Both POMDPs and DLIMIDs are designed for Markovian infinite-horizon problems in which only some of the variables can be observed. DLIMIDs can be seen as a generalization of factored POMPDs, as they allow to have several decisions per time slice. In fact, a DLIMID having only one decision per time slice, such as the carcinoid model presented in this paper, can be interpreted and solved as a POMDP, and vice versa. Therefore, we can say that they are not two different types of models, but two ways of evaluating a model.

There are, however, subtle differences from the point of view of knowledge engineering. DLIMIDs in general need memory variables to summarize past observations, while POMDPs summarize the whole observed history implicitly in the belief state the unobservable variables. However, it does not imply that when interpreting and evaluating a DLIMID as a POMDP we can always remove the memory variables; it may be necessary to replace them with hidden variables.

If a POMDP is evaluated as a DLIMID without introducing memory variables, in general we will obtain a much lower expected utility. But if is an open issue to determine if a model including memory variables gives higher expected utilities when evaluated as a DLIMID than when evaluated as a POMDP, or vice versa. It would also be interesting to study which policies are easier to understand by human users.

References

Bellman, R. E. 1957. *Dynamic Programming*. Princeton, NJ: Princeton University Press.

Boutilier, C., and Poole, D. 1996. Computing optimal policies for partially observable decision processes using compact representations. In Clancey, W. J., and Weld, D., eds., *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, 1168–1175. Portland, OR: AAAI Press.

Boutilier, C.; Dean, T.; and Hanks, S. 1996. Planning under uncertainty: Structural assumptions and computational leverage. In Ghallab, M., and Milani, A., eds., *New Directions in AI Planning*. Amsterdam, The Netherlands: IOS Press. 157–171.

Cooper, G. F. 1988. A method for using belief networks as influence diagrams. In Shachter, R.; Levitt, T.; Kanal, L. N.; and Lemmer, J. F., eds., *Proceedings of the Fourth Workshop on Uncertainty in Artificial Intelligence*, 55–63. New York, NY: Elsevier Science.

Eglese, R. W. 1990. Simulated annealing: A tool for operational research. *Eur J Oper Res* 46:271–281.

Hansen, E. A. 1998. Solving POMDPs by searching in policy space. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI'98)*, 211–219. Madison, WI: Morgan Kaufmann, San Francisco, CA.

Kirkpatrick, S.; Gelatt Jr, C. D.; and Vecchi, M. P. 1983. Optimization by simulated annealing. *Science* 220:671–680.

Lauritzen, S., and Nilsson, D. 2001. Representing and solving decision problems with limited information. *Management Science* 47:1238–1251.

Meuleau, N.; Kim, K. E.; Kaelbling, L. P.; and Cassandra, A. R. 1999. Solving POMDPs by searching the space of finite policies. In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (UAI'99)*, 417–426. Stockholm, Sweden: Morgan Kaufmann, San Francisco, CA.

Murphy, K. 2001. The Bayes Net Toolbox for Matlab. *Computing Science and Statistics* 33:1–20.

Murphy, K. 2002. *Dynamic Bayesian Networks: Representation, Inference and Learning*. Ph.D. Dissertation, Computer Science Division, University of California, Berkeley.

Ng, A. Y., and Jordan, M. I. 2000. PEGASUS: A policy search method for large MDPs and POMDPs. In Boutilier, C., and Goldszmidt, M., eds., *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, 406–415. San Francisco, CA: Morgan Kaufmann.

Ross, S. 1983. *Introduction to Stochastic Dynamic Programming*. New York, NY: Academic Press.

Toussaint, M., and Storkey, A. J. 2006. Probabilistic inference for solving discrete and continuous state Markov Decision Processes. In *Proceedings of the 23nd International Conference on Machine Learning (ICML 2006)*, 945–952.

van Gerven, M. A. J., and Díez, F. J. 2006. Selecting strategies for infinite-horizon dynamic LIMIDs. In Studený, M., and Vomlel, J., eds., *Proceedings of the Third European Workshop on Probabilistic Graphical Models (PGM'06)*, 131–138.

van Gerven, M. A. J.; Díez, F. J.; Taal, B. G.; and Lucas, P. J. F. 2007. Selecting treatment strategies with dynamic limited-memory influence diagrams. *Artificial Intelligence in Medicine* 40:171–186.

Weinstein, M., and Stason, W. 1977. Foundations of cost-effectiveness analysis for health and medical practices. *New Engl J Med* 296(13):716–721.