University of Innsbruck

Faculty of Mathematics, Computer Science and Physics

Institute for Theoretical Physics

Bachelor Thesis

submitted for the degree of

Bachelor of Science

An information-theoretical approach to internal models in a Partially Observable Markov Decision Process

by

Lukas Prader

Matriculation Nr.: 12115058

SE Seminar with Bachelor Thesis

Submission Date: 24.07.2024
Supervisors: Alexander Vining, Hans J Briegel

Abstract

Intelligent biological agents are able to make sense of their environment and learn to act in it. In order to accurately replicate these behaviours and other complex phenomena in artificial agents, we need better models of the information processing happening in biological organisms. When only provided with partial information about a system, current artificial agents rely on sufficiently complex internal structures (i.e. a sufficient number of optimizable parameters) to enable them to learn complex tasks, yet the question remains how these internal structures come to be. In this work, an information theoretical approach to internal model generation was explored with a simple delayed action task. The results show great potential for the framework to enable dynamic internal model generation, with further analysis necessary concerning the scenario of an agent estimating information theoretical quantities from finite samples.

Keywords: reinforcement learning, information theory, Partially Observable Markov Decision Process, internal model

Contents

1 Introduction

The rise of artificial intelligence in recent years has produced new research trying to create complex models able to perform intelligent tasks. In many disciplines it has been shown that there are so-called emergent phenomena, which are not easily explained by lower-level fundamental laws, requiring additional concepts to accurately explain them (Anderson 1972). Exactly how these complex phenomena can emerge from the set of currently known fundamental laws is an ongoing field of research, with interdisciplinary approaches taking from the fields of physics, chemistry, biology, psychology, philosophy computer science and others.

Research into so-called complex adaptive systems, for example connected to biology, aims to understand the emergence of intelligent and adaptive behaviour in biological organisms. Gaining insight into the mechanisms which enable biological systems to exhibit complex behaviour can in turn be used to improve the ways in which we attempt to create systems capable of these behaviours ourselves.

Complex system research can thus provide valuable insights into the emergence of intelligent behaviour, which also includes processes connected to adaptive behaviour and learning in animals. Works by the likes of Pavlov and Skinner have probed into the mechanisms that influence behavioural patterns in animals (Pavlov 1906; Skinner 1957). Exactly how conditionable behaviour like this can emerge just from interaction with the environment is still not fully understood.

Information theory has been fundamental in furthering our understanding of such complex processes. With the work of Shannon (Shannon 1948) as a basis, modern information theory has enabled researchers to quantify correlations and changes in complexity of dynamic systems. It has successfully been applied to many examples in behavioural biology, such as looking into the collective behaviour of ants, information exchange in slime moulds and group behaviour in bat populations (Kim et al. 2021).

Information theory can also be used to examine how biological agents, such as animals, acquire an understanding of their surroundings in order to then act in response, a so-called internal model of their environment. Current approaches to create artificial agents able to learn certain behaviours rely on trial and error to find the number of parameters necessary to explain the complexity of a given environment, especially if only limited information about an environment is available to the agent.

Real life organisms do not seem to be limited as much, still being able to infer information about their environment even with limited information at their disposal. Logical conclusions based on indirectly connected information, called transitive

inference, has been shown in multiple animals such as pigeons, honey bees, rats and others (Vasconcelos 2008). Here, animals are able to infer the relations between stimuli only with information about other stimuli and their relations to each other. Also in the field of Artificial General Intelligence, tests like the ARC test set (Chollet 2019) show that current state-of-the-art machine learning models are not able to replicate reasoning tasks routinely performed by humans (Johnson et al. 2021).

Crutchfield and Feldman have proposed information theoretical quantities able to explain processes of inference for complex systems (Crutchfield et al. 2003). They can quantify the amount of "synchronization" a system has achieved in comparison to a different system (like an environment), which it can interact with. This framework can provide tools which may enable an agent to autonomously modify the current internal model in order to more accurately reflect the processes of the observed environment.

This thesis aims to apply some of the quantities proposed by Crutchfield and Feldman to a simple example of an agent acting with partial information of its environment. The goal is to show the capabilities of this framework to enable evaluation of the current internal model and subsequent improvements to accurately reflect the environment, even with only partial information available to the agent. We will specifically examine how to determine if an internal model has reached sufficient complexity to explain the processes it observes and how this can be used to create an adequate internal model.

2 Methods

We will look at a small system related to behavioural biology, which can be modelled as a Markov Decision Process. A Markov Decision Process (MDP) (Bellman 1957) is defined by a set of states S, which are connected by transition probabilities, and a set of actions that an agent can perform in a given state. These actions are determined by the agent's policy, commonly denoted with π. One usually imposes the Markovian property, which implies that the transition probability from one state to the next only depends on the state itself and not the previous states of the system (Cover et al. 1999).

2.1 The delayed action task

The system we will analyse is a delayed action task, which we want our agent to learn. One can imagine a rat in a box, very similar to the aforementioned experiments by Skinner. In this box, the rat observes a light switching on and off and has access to a button it can press. The rat can obtain rewards, such as food, based on its actions. The light's state (on or off) changes in discrete time steps depending on the rat's actions. In our setup, the rat receives a reward only if it presses the button at the correct time after the light turns on, specifically in the second time step. This process can be visualized as the graph of a MDP, using three states and the possible transition probabilities as edges between them (Fig. 1).
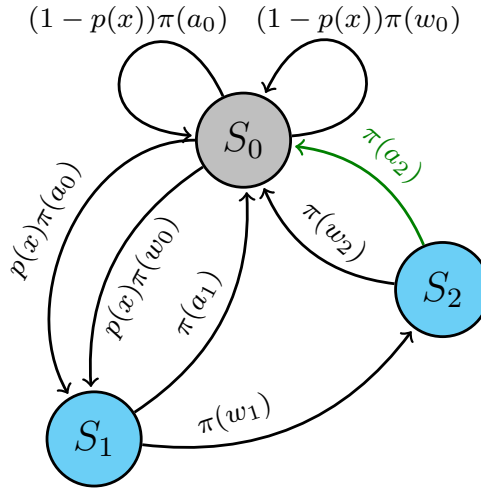

Figure 1: Visualization of the delayed action MDP. Colours of the states describe the state of the light (grey = dark, blue = light) Probabilities coming from the agent policy, acting a_i or waiting w_i for a state with index i, are denoted with π. The probability of the light turning on in the dark state is given as $p(x)$.

In the most general case, the policy of the agent can be different in every state

S_i. We will define the policy to be a probability, with probabilities of choosing to either wait $\pi(w_i)$ or act and press the button $\pi(a_i)$. One can specify the policy only by defining the probability to act, consequently choosing the probability to wait as the inverse probability $1 - \pi(a_i)$. We will assume that the state of the light is fully determined by the agent's actions when being on, but if it is off, there is a probability of $p(x)$ for the light to turn on, independent of the agents action taken in the dark state.

The given MDP has two main properties, stationarity and irreducibility. Being irreducible means that every state is reachable from every other state with positive probability in a finite number of steps (Cover et al. 1999), while stationarity implies that the transition probabilities do not change over time. This is the case for our MDP, if we assume the policy to be fixed.

If we are specifically interested in the transitions for the states of the light, we can write the state transitions of this MDP into a state transition matrix P (assuming some fixed policy π):

$$P_{MDP} = \begin{bmatrix} 1 - p(x) & p(x) & 0 \\ \pi(a_1) & 0 & 1 - \pi(a_1) \\ 1 & 0 & 0 \end{bmatrix}, \tag{1}$$

with each row summing up to one.

One can see that the system reflected by this transition matrix does not capture the whole information about our initial system any more, ignoring the reward and the actions that do not directly influence the transitions of the light.

With the state transition matrix we can calculate the probability of ending up in a state in the next time step, given the probability of being in any of the states in the step before. If we let the system transition for many time steps, the state distribution will converge to the so-called stationary distribution (Cover et al. 1999). This stationary distribution μ, a row vector by convention, satisfies the following equation:

$$\mu = P\mu. \tag{2}$$

This means that μ can be calculated by finding the matrix eigenvector with eigenvalue 1. The eigenvector should be re-scaled if necessary, such that the stationary distribution also has a sum of 1.

2.2 Entropy and entropy rate

Since we are interested in modelling an agent learning the task, we will treat it as a process generating observations. In this way, we are able to analyse how an internal model can be formed, especially if there is no direct interaction to the task. One can picture an observer just receiving the observations, without having any additional context to aid in interpretation. The agent can observe parameters of the system, in particular the current state of the light and the action that was just performed. Over multiple time steps, these observations will form sequences made up of symbols, which correspond to the particular state that was observed. The sequences will have a symbol distribution dependent on the nature of the generating process, motivating the use of information theory to analyse the properties of these sequences.

The standard Shannon entropy,

$$H = -\sum_{x \epsilon \mathcal{X}} p(x) \log p(x), \tag{3}$$

is defined with the sum over all symbols x in a given alphabet $\mathcal{X}$, $\log$ meaning the base two logarithm here, as well as in the rest of this thesis, implying "bits" as the unit of information. Given a sequence, we can also calculate the entropy of tuples, which we will call blocks, of symbols. We can define the block entropy $H(s^L)$ as the entropy of blocks of symbols with length L:

$$H(s^L) = -\sum_{s_i^L \epsilon S^L} p(s_i^L) \log p(s_i^L), \tag{4}$$

looking at all possible blocks s_i^L of length L, given a set of symbols S. One simple intuition to understand this is to think about the entropy of words. Instead of calculating the entropy of the individual letters, we calculate the entropy of length L words in a given sentence.

For a given block length, one can define the change of entropy at this length using the discrete derivative:

$$\Delta H(s^L) = H(s^L) - H(s^{L-1}). \tag{5}$$

This change in block entropy is equivalent to an estimate of the entropy rate $\hat{h}_\mu(L)$ at length L, which measures how random the system appears if only blocks up to length L are observed. For an infinitely long sequence, the change in entropy

converges to the final entropy rate (Crutchfield et al. 2003):

$$h_\mu = \lim_{L\to\infty} \frac{H(s^L)}{L}.$$ (6)

This entropy rate can be interpreted as the inherent randomness of sequences obtained from the generating process, stemming from the full complexity of the system. It can also be shown that $\Delta H(s^L)$ decreases monotonically for increasing L, implying that a finite estimate of h_μ will tend to overestimate the randomness of an incoming sequence, and thus of the generating process (Crutchfield et al. 2003).

For a stationary Markov Process, the final entropy rate can be calculated if the stationary distribution and the transition matrix are known (Cover et al. 1999):

$$h_\mu = -\sum_i \mu_i \sum_j P_{ij} \log P_{ij}.$$ (7)

2.3 Synchronization and predictability gain

Knowing how exactly $\Delta H(s^L)$ of a system converges to the final entropy rate h_μ provides information about the complexity and structure of the system. Depending on the length of blocks used, the estimated complexity will be different from the true complexity of the system. This opens the question of when an agent observing a sequence from a generating process can be said to have obtained all the information about the system. At this point the agent would have all the information necessary to understand the nature of the generating process. So-called "synchronization" is achieved once the discrete change in entropy is equal to the true final entropy rate of the system (Crutchfield et al. 2003):

$$h_\mu - \Delta H(s^L) = 0.$$ (8)

It means that there is no new information in blocks larger than L, for which synchronization was achieved.

This criterion depends on knowing the true entropy rate of the generating process, or at least having a very good estimate of it. In general, this might not be feasible to obtain for an agent, especially since the agent will have no prior knowledge about the generating process, which it could use to estimate the entropy rate.

A different, although weaker condition to characterize synchronization is that the entropy rate has to be constant from then on. This means that the second derivative

$\Delta^2 H(s^L)$ will be 0. We define

$$\Delta^2 H(s^L) = \Delta H(s^L) - \Delta H(s^{L-1}), \tag{9}$$

also called predictability gain, which quantifies how much randomness is lost when using information of length L blocks (Crutchfield et al. 2003). In order to sensibly define the predictability gain for $L = 1$, one defines $\Delta H(s^0) = \log |S|$, using the number of symbols in S defined by the cardinality $|S|$. This is motivated by the idea that the randomness of the system is assumed to be maximal if no sequences have yet been observed (Crutchfield et al. 2003). It is important to note that for synchronization to be achieved, the predictability gain has to be zero for all following block lengths. In some cases like periodic processes, it can happen that $\Delta^2 H(s^L)$ will be zero for some blocks, but different from zero for following blocks due to the periodic nature of the process (Crutchfield et al. 2003).

2.4 Observing the delayed action task

In the case of the delayed action task (section 2.1), the ability for an agent to learn this task heavily depends on the type of information it has about the system. If the agent is able to observe the true labels of each state, namely S_0, S_1 and S_2 (Fig. 2), it can easily find the optimal policy maximizing the reward.
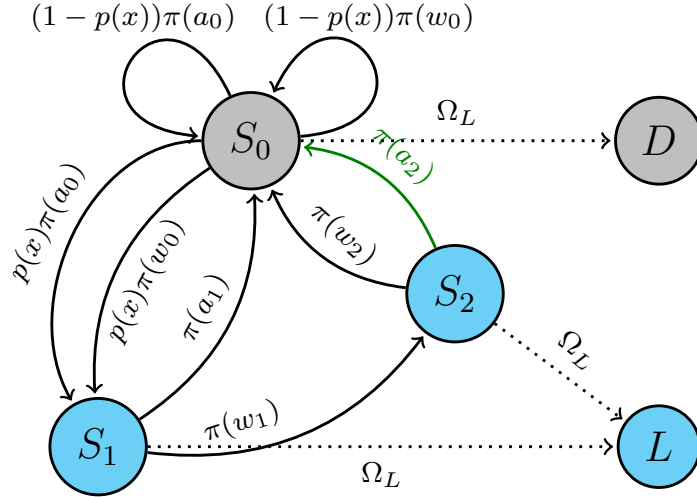

Figure 2: The process of reducing the delayed action task (Fig. 1) to the observable light states using the mapping Ω_L. One can see that S_1 and S_2 produce the same observation.

If the agent can only distinguish between light and dark states based on observations of the light, it can not distinguish S_1 and S_2, making it a Partially Observable Markov Decision Process (POMDP). In this case, the agent will not be able to reach

the optimal policy, since it can not distinguish the rewarded light state S_2 from the unrewarded S_1. Yet, due to the temporal structure of the MDP, one can make a difference between S_1 and S_2 as soon as the prior state of the light is also known. If the agent were able to use this information, it would again be in a position to distinguish S_1 and S_2, making it possible to learn the delayed action task once more.

This motivates the idea to use the measures shown in the previous sections, since they should be able to find this temporal correlation. In particular, we should be able to show this temporal structure in the change of block entropy, so a different $\Delta H(s^L)$ for $L \leq 2$ and $L > 2$. We will thus analyse sequences of the original MDP, reduced to partial observations, and try to quantify the information encoded in them with higher block lengths.

In order to obtain these sequences, we have to first generate a sequence from the true MDP and then reduce the symbols to the partial observation we want to look at (i.e. Fig. 2 for observing the light). This reduction is performed by an observation mapping Ω_K, with K being the chosen observation. Two main sources of information in this process are the state of the light ($L, D \rightarrow$ light, dark) and the actions taken by the agent ($A, W \rightarrow$ act, wait), which is why we will analyse the behaviour of both of them in the following sections.

2.5 Quantity estimation

Since information entropy and the quantities derived thereof use probabilities, their true values are results of using infinite samples. Since this is not achievable for a real agent, we will have to use the estimated probabilities. The agent can estimate the likelihood of a block by calculating the frequency of it in a given observation sequence, i.e. for the probability of a specific block s_i^L we get

$$\hat{p}(s_i^L) = \frac{n(s_i^L)}{N}, \tag{10}$$

with N being the total number of blocks of length L in the given sequence.

As an illustrative example, we take the sequence D L L D L D. For blocks of length 2, we get DL, LL, LD, DL and LD. From this, one could now calculate the frequencies of all possible blocks of length 2. With the defined probability estimator, we can define the so-called plug-in-estimator for the block entropy:

$$\hat{H}(s^L) = - \sum_{s_i^L \epsilon S^L} \hat{p}(s_i^L) \log \hat{p}(s_i^L), \tag{11}$$

which we will use later to compare estimated values to the analytical results.

3 Results

3.1 Stationary distribution and partial models

To be able to calculate the block entropy curve for the delayed action task, one first calculates the stationary distribution of the system. Using equation (2) with the transition matrix P (1), we get the probabilities of encountering each state when observing the system at some random point in time:

$$\mu_{MDP} = [p(S_0),\ p(S_1),\ p(S_2)] =$$
$$= [\frac{1}{1 + p(x)(2 - \pi(a_1))},\ \frac{p(x)}{1 + p(x)(2 - \pi(a_1))},\ \frac{p(x)(1 - \pi(a_1))}{1 + p(x)(2 - \pi(a_1))}] \tag{12}$$

Using this result, we can actually reduce the MDP of the delayed action task to Markov processes using either the light or the possible actions as states. This results in two partial models of the MDP, which we will call MP1 for the light states and MP2 for the action states (see Fig. 3). For easier notation, we will define the total probability to act and wait as follows:

$$p(a) = \sum_i p(S_i)\pi(a_i),\ p(w) = 1 - p(a). \tag{13}$$

Sequences generated by these partial models can be compared to observations generated by the MDP, showing if these partial models are accurate representations of the true generating process. If they are, it would imply that they can be used as accurate internal representations of the system.
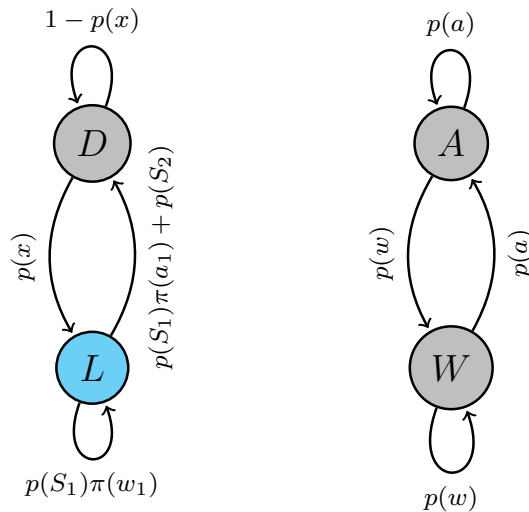

Figure 3: Partial models for observations of light (left, MP1) and action (right, MP2). The transitions are derived from the original MDP (Fig. 1) and weighted using its stationary distribution (Eq. 12).

9

3.2 Block entropy curves

We will now look at the block entropy behaviour of the previously obtained partial models, comparing it to the block entropy curves obtained by observations of the true MDP. This will provide us with information about the adequacy of MP1 and MP2, showing if they are sufficiently complex to describe the MDP. If they are, they could potentially be used by an agent to successfully learn the delayed action task. In order to visualize the curves, we will choose the following numerical values for the free parameters:

$$p(x) = 0.6, \ \pi(a_0) = 0.5, \ \pi(a_1) = 0.3, \ \pi(a_2) = 0.7. \tag{14}$$

With this, we can calculate the stationary distributions of our processes, as well as the block entropy values for a given length, just by using the stationary distribution and the transition matrices to calculate the block probabilities (Fig. 4). We can also calculate the final entropy rates for each system by using equation (7), which makes it possible to visualize the behaviour of convergence.

There is one elementary difference between the results obtained for MP1 or MP2 and the results for the MDP. The results for the internal models can be seen as predictions the agent makes about its environment, while the results for the MDP are those that the agent obtains using observations. This means that figure (4) illustrates how predictions of the internal models compare to the real observations an agent would receive.
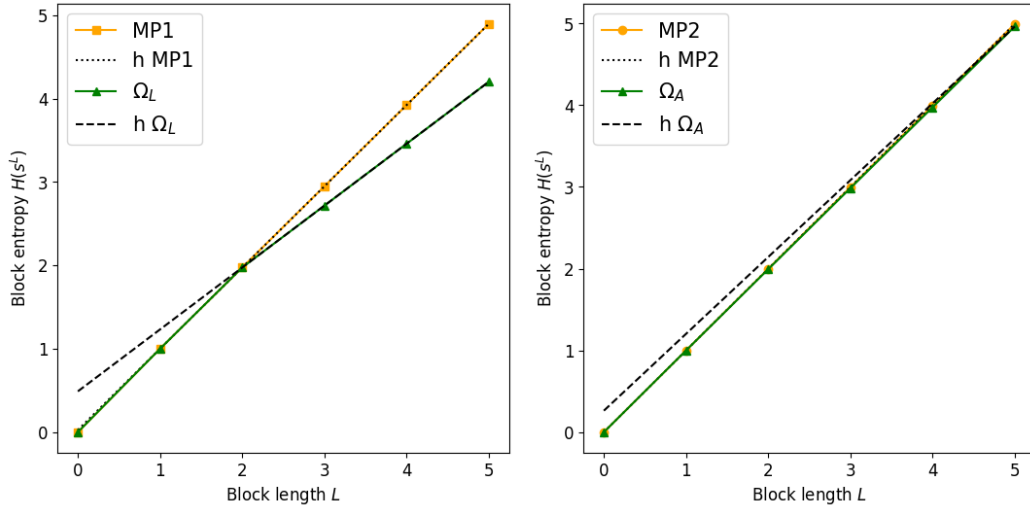

Figure 4: Plots showing the block entropies of the partial models in orange (MP1 left, MP2 right) compared to the block entropies generated by observing the true MDP in green. The dashed and dotted lines show the final entropy rate to which the systems converge for the observations and the partial models respectively.

The first thing we can notice is that in both systems, there is a mismatch between the block entropies of sequences generated by the MDP (green triangles) and block entropies of sequences generated by the partial observation models (orange squares). In case of the light (Fig. 4 left), the difference is very distinct, showing a divergence exactly after $L = 2$. For the action (Fig. 4 right) it is more subtle, but one can see that the two curves definitely start to separate at $L = 5$. There also are differences in how Ω_L and Ω_A converge to their respective final entropy rates $h\,\Omega_L$ and $h\,\Omega_A$.

The behaviours observed for Ω_L and Ω_A can both be explained with results from Crutchfield and Feldman (Crutchfield et al. 2003).

Observing the light shows the block entropy curve of an order-R Markov Process, with R being the amount of states relevant for the transition probabilities of the next state. The order of the process reflects in the point of synchronization, where the entropy rate reaches its final value. This point is illustrated with the green points of Ω_L meeting the dashed line indicated by the final entropy rate of the observation $h\Omega_L$ (Fig. 4, left). In our case, this implies R=2, where the entropy rate changes to match the black dashed line, meaning that the current state and the state before are needed to know the transition probabilities to the next state.

Looking at the structure of the delayed action task again (Fig. 1), this result highlights how the light states are related to each other. If an agent can only observe the current state of the light, S_1 and S_2 can not be distinguished since in both of them the light is on. Yet, if the agent is able to look at the current state of the light and the state in the time step just before, it is possible to distinguish S_1 from S_2. Since S_2 is only reachable from S_1 and S_1 only from S_0, there must have been a light state before being in S_2 and a dark state before being in S_1. This means that with DL for S_1 and LL for S_2 as distinguishable sequences, an agent capable of remembering the current state and the last one just before, so a temporal memory of length 2, would be able to distinguish all states of the MDP.

Comparing this behaviour to MP1, we can see that MP1 does not have any change in entropy rate $\Delta H(s^L)$, it matches the final block entropy (dotted line) already from the beginning. Since Ω_L has exactly one change in entropy rate exactly at $L = 2$, it signifies that there is some structure underlying the generating process that influences the block entropy for blocks of length 2, but no higher order complexity. This means that MP1 does not have the order 2 complexity needed to accurately represent the light states of the MDP.

Yet it is interesting to note that MP1 and Ω_L do match in block entropy values for $L \leq 2$, suggesting that MP1 correctly models the behaviour of the MDP up to sequences of length 2. This arises from how MP1 is computed, using the stationary

distribution and transitions to the next state from the MDP, which implies accurate representation of blocks with $L \leq 2$. It also means, that an agent with only the ability to look at blocks up to $L = 2$ would not be able to see the shortcomings of MP1, seeing it as the correct model for this process. This issue already shows an additional limitation for an agent trying to find the correct internal model from estimates with obtained observations, which will be discussed in section (4). The choice of using stationary results of the MDP additionally implies inherent stationarity for MP1, explaining why it is already synchronized from the beginning.

When observing the actions of the agent, the behaviour can be characterized as that of a Hidden Markov Process. This stems from the fact that in every state of the MDP, the agent will either act or wait, making the underlying states "hidden". The convergence to the final entropy rate in this case is said to be exponential, implying complete synchronization only in the limit of infinity. A visual result from this is that the block entropy of Ω_A will never match the dashed line of $h\Omega_A$, only asymptotically converge to it with every increase of L. This means that there will always be some small uncertainty about the exact state of the system, even with many observed transitions. In comparison to that, MP2 is again a model with no higher order structure, resulting in a straight line. While this might not be easily visible in figure (4), one can make the same argument as made for MP1, arguing that the way of creating MP2 results in a completely stationary process due to the stationary distributions and transitions used from the MDP.

Regarding the accuracy of MP1 and MP2 as models of the true MDP, we can now conclude that both are not exact representations of the underlying MDP. This can also be understood as the internal model predicting a certain behaviour, the orange values, but failing at matching the observations in green, implying that the current internal model is not sufficiently complex.

3.3 A sufficiently complex internal model

In the case of observing the light state in the delayed action task, synchronization has been shown to be achieved at finite block lengths, specifically $L = 2$. Using this information, we can create a model using length 2 blocks as states (Fig. 5).
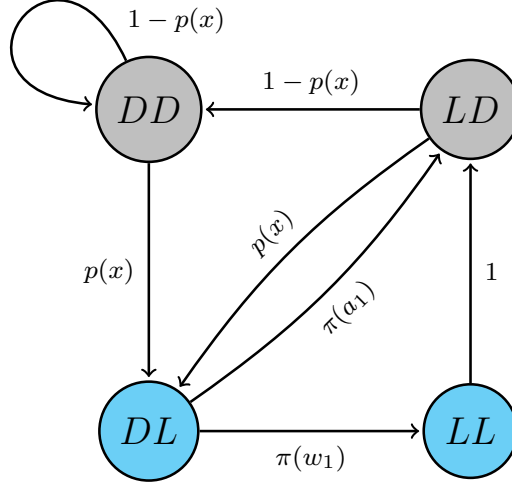
Figure 5: Internal model with sufficient complexity to explain the light state of the delayed action task, created using length 2 blocks as states.

An internal model of this form would be sufficient for the agent to synchronize to the MDP, potentially making it possible to learn the optimal policy for the delayed action task.

3.4 Estimating block entropy and predictability gain

Now knowing the analytical result for the necessary complexity of the internal model, we want to further investigate how an agent would be able to come to this conclusion by itself. We will only analyse the case of observing the light now, since the desired complexity is reachable with finite block length. The agent would have to estimate the predictability gain for the incoming observation in order to find the block length needed to accurately represent the observation.

To simulate the estimation process for the delayed action task, 1000 replications using sequences of length 50 were computed using the transition matrix of the MDP (Eq. 1), using the MDP stationary distribution (Eq. 12) to initialize each sequence. The sequences were reduced to the light observations (see section 2.4) and block entropy values were then estimated using equation (11). The resulting values can then be compared with the analytical results from before (Fig. 6).
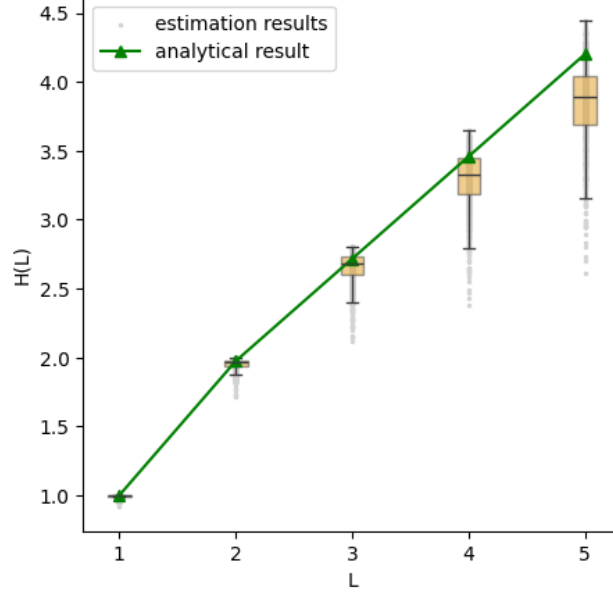
Figure 6: Plot showing the results when estimating block entropy from sequences of length 50. The histograms show the combined results from 1000 replicates.

One can see that the estimate of the block entropy becomes worse for larger block lengths, not only having higher variance, but also resulting in values further below the analytical results. This is in part due to the nature of the used plug-in-estimator, which is known to underestimate the true result for finite sample size, which is also called negative bias (Basharin 1959). A more dominant reason here though is the temporal correlation between samples, since samples come from the same trajectory of the MDP for each sequence.

When estimating the predictability gain, this issue with bias and variance has significant implications for the results the agent obtains. Since the needed complexity is determined by the last block length with non-zero predictability gain, negative bias will wrongly shift the complexity estimate to higher block lengths, as seen in figure (7). This means that the agent would estimate the system to be more complex than it actually is.
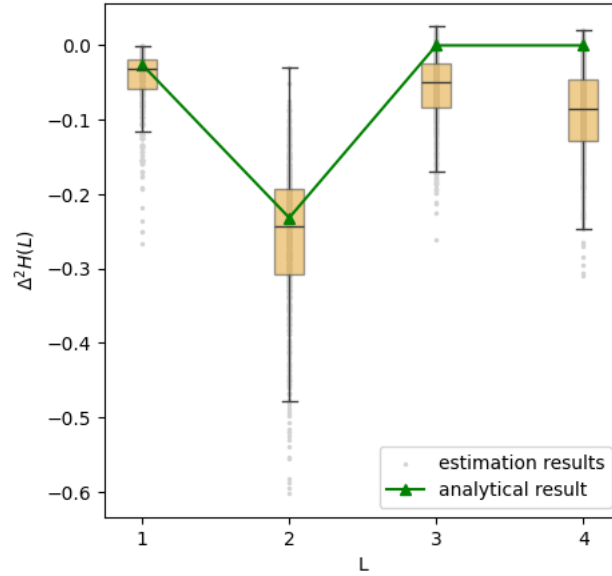
Figure 7: Plot showing the estimated predictability gain for sequences of length 50. The histograms show the combined results from 1000 replicates.

Nicely enough, the plug-in-estimator is a consistent estimator, implying that with sequence length approaching infinity, the bias and variance will vanish, making the estimate converge to the analytical result (Basharin 1959). While infinite sample size is of course not feasible for any agent, it would also be desirable in general to have a minimal sample size, since this would make the agent more efficient. The simulations here are just a demonstration of the difficulty encountered when trying to estimate the complexity of an observation. More analysis would need to be done to see how exactly this estimation could be implemented efficiently for use in an agent.

4 Discussion

The results presented show that information theoretical measures can be applied to a Partially Observable Markov Process in order to find sufficiently complex internal models, though mostly analytically. From observing the light state of the system, one can obtain an internal model which is sufficiently complex to explain the observed block entropy of the light state. The found representation does not correctly reflect the true MDP though, i.e. with four states instead of three.

Other work by Crutchfield shows that for unifilar processes, in which each state only emits distinct symbols, once a sufficiently complex model is found, the true system can be recovered by correctly reducing the model if necessary. This model is then called the ϵ-machine of the system (Crutchfield 1994). A proposed way to reconstruct the ϵ-machine of a system is to iteratively increase model complexity based on the incoming observations. This procedure could also be implemented for the delayed action task.

It has also been shown that this procedure can be used for Hidden Markov Processes (Jurgens et al. 2021). This means that the action of the agent could also be modelled with an ϵ-machine, potentially enabling the agent to infer the impact of its actions in relation to the light state by using states which combine light state and action. With this, an agent might be able to infer the optimal policy just from its internal model without having to learn the policy directly, as in other approaches of reinforcement learning like q-learning. It may also enable the agent to simulate the environment by itself, which could be used to improve the policy similarly to methods in Projective Simulation (Briegel et al. 2012).

For this, further work would need to be done to clarify the difficulties in estimation shown in section (3.4). There are papers which propose certain thresholds for sequence lengths in order to mitigate bias and get sufficiently accurate results. (Lesne et al. 2009) show that the needed sequence length N to accurately estimate block entropy of length n with k symbols depends on the final entropy rate h_μ:

$$n \leq \frac{N h_\mu}{\ln(k)}.$$ (15)

They propose first estimating the entropy rate using Lempel-Ziv complexity, which is not biased in the same way as the plug-in-estimator, then selecting the sequence length accordingly. Other works use slightly different sequence lengths, but all of them incorporate the influence of h_μ into their results (Larson et al. 2011; Dębowski 2016).

It should also be possible to calculate confidence intervals for the estimates, provid-

ing additional information for the agent. Estimation is a major point of concern in this framework and should be addressed more thoroughly in future work.

It is important to note that the result of this estimation now strongly depends on the capacity of the agent to store long sequences. Depending on the length of memory, high block lengths can not be estimated confidently. This also means that if the memory is too short, an agent is not able to find the highest order complexity for a given sequence. If, for example, only blocks up to $L = 2$ could be estimated by an agent observing the light state of our delayed action task, the agent would not find an accurate model. It would actually find MP1 do be an accurate representation, since it matches blocks entropies up to $L = 2$. This means that an agent will converge to an insufficient internal model, if the length of observed sequences is too short compared to the complexity of the generating process.

One way for the agent to be aware of this limitation would be to estimate a final entropy rate $\hat{h}_\mu$ for the internal model and compare it to the one used to determine sequence length. The difference between those values would tell the agent how far the internal model is away from an accurate representation, even if the current model is the optimal one for the obtainable sequence length.

5 Conclusion

Creating more realistic and capable artificial agents promises a more complete understanding of agency and the emergence of complex behaviour in complex adaptive systems. It has been shown that information theory is a very powerful tool, which can be used to infer a lot about a system even in circumstances with limited information.

The model-based approach shown here has the potential to provide strong methods for policy optimization, with additional adaptability due to a dynamic, iterative approach to internal model generation. Current methods for solving POMDPs usually rely on a correct internal model to begin with, which would make the proposed framework a huge improvement in terms of adaptability and usefulness. With internal models representing the observed environment, this approach offers an excellent way to study the learning process of agents, clearly illustrating the structures an agent is able to identify from its own experiences.

In order to prove the viability of this framework, more work has to be done to show how exactly it can be realized and deployed in an agent. The potential benefits give strong motivation to pursue more research focused on internal model generation.

6 Acknowledgements

Declaration of Authorship

I hereby solemnly declare, by my own signature, that I have independently authored the presented work and have not used any sources or aids other than those indicated. All passages taken verbatim or in content from the specified sources are identified as such.

I consent to the archiving of this Bachelor thesis.

Innsbruck, 24.07.2024   Lukas Prader

References

Anderson, Philip W (1972). 'More Is Different: Broken symmetry and the nature of the hierarchical structure of science.' In: *Science* 177.4047, pp. 393–396. URL: `https://www.jstor.org/stable/pdf/1734697.pdf`.

Basharin, Georgij P (1959). 'On a statistical estimate for the entropy of a sequence of independent random variables'. In: *Theory of Probability & Its Applications* 4.3, pp. 333–336. DOI: `10.1137/1104033`.

Bellman, Richard (1957). 'A Markovian decision process'. In: *Journal of mathematics and mechanics*, pp. 679–684. URL: `http://www.jstor.com/stable/24900506`.

Briegel, Hans J and Gemma De las Cuevas (2012). 'Projective simulation for artificial intelligence'. In: *Scientific reports* 2.1, p. 400. DOI: `10.1038/srep00400`.

Chollet, François (2019). 'On the measure of intelligence'. In: *arXiv preprint arXiv:1911.01547*. URL: `https://arxiv.org/pdf/1911.01547`.

Cover, Thomas M and Joy A Thomas (1999). *Elements of information theory*. John Wiley & Sons.

Crutchfield, James P (1994). 'The calculi of emergence: computation, dynamics and induction'. In: *Physica D: Nonlinear Phenomena* 75.1-3, pp. 11–54. URL: `https://csc.ucdavis.edu/~cmg/papers/CalcEmerg.pdf`.

Crutchfield, James P and David P Feldman (2003). 'Regularities unseen, randomness observed: Levels of entropy convergence'. In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 13.1, pp. 25–54. DOI: `10.1063/1.1530990`.

Dębowski, Łukasz (2016). 'Consistency of the plug-in estimator of the entropy rate for ergodic processes'. In: *2016 IEEE International Symposium on Information Theory (ISIT)*. IEEE, pp. 1651–1655. DOI: `10.1109/ISIT.2016.7541579`.

Johnson, Aysja et al. (2021). 'Fast and flexible: Human program induction in abstract reasoning tasks'. In: *arXiv preprint arXiv:2103.05823*. URL: `https://arxiv.org/pdf/2103.05823`.

Jurgens, Alexandra M and James P Crutchfield (2021). 'Shannon entropy rate of hidden Markov processes'. In: *Journal of Statistical Physics* 183.2, p. 32. DOI: `10.1007/s10955-021-02769-3`.

Kim, Hyunju et al. (2021). 'Informational architecture across non-living and living collectives'. In: *Theory in Biosciences*, pp. 1–17. DOI: `10.1007/s12064-020-00331-5`.

Larson, J Walter, Peter R Briggs and Michael Tobis (2011). 'Block-entropy analysis of climate data'. In: *Procedia Computer Science* 4, pp. 1592–1601. DOI: `10.1016/j.procs.2011.04.172`.

Lesne, Annick, Jean-Luc Blanc and Laurent Pezard (2009). 'Entropy estimation of very short symbolic sequences'. In: *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics* 79.4, p. 046208. DOI: `10.1103/PhysRevE.79.046208`.

Pavlov, Ivan P (1906). 'The scientific investigation of the psychical faculties or processes in the higher animals'. In: *Science* 24.620, pp. 613–619. DOI: `10.1126/science.24.620.613`.

Shannon, Claude Elwood (1948). 'A mathematical theory of communication'. In: *The Bell system technical journal* 27.3, pp. 379–423. URL: `https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6773024`.

Skinner, Burrhus Frederic (1957). 'The experimental analysis of behavior'. In: *American scientist* 45.4, pp. 343–371. URL: `https://www.jstor.org/stable/pdf/27826953.pdf`.

Vasconcelos, Marco (2008). 'Transitive inference in non-human animals: An empirical and theoretical analysis'. In: *Behavioural Processes* 78.3, pp. 313–334. DOI: `10.1016/j.beproc.2008.02.017`.