Model Selection in Markovian Processes

Assaf Hallak IFOGPH@GMAIL.COM
Dotan Di-Castro DOTAN.DICASTRO@GMAIL.COM
Shie Mannor SHIE@EE.TECHNION.AC.IL
Technion institue of technology, Haifa, Israel

Abstract

In this work we address the problem of how to use time series data to choose from a finite set of candidate discrete state spaces, where these spaces are constructed by a domain expert. We formalize the notion of model selection consistency in the proposed setup. We then discuss the difference between our proposed framework and the classical Maximum Likelihood (ML) framework, and give an example where ML fails. Afterwards, we suggest alternative selection criteria and show them to be weakly consistent. Finally, we test the performance of the suggested criteria on both simulated and real world data.

1. Introduction

Markov decision processes (MDPs) can describe dynamical problems found in artificial intelligence, control, operations research and many other fields. Algorithms that use MDPs for optimizing and evaluating policies in different decision problems typically start with the assumption that the state space is known. In practice, this is generally not the case. In many situations the practitioner must choose from a candidate set of state spaces, usually constructed by a domain expert, before applying an optimization algorithm.

Our work is motivated by the following scenario: a stream of data describing some goal oriented dynamics is given and a domain expert analyzes the observations and suggests different models that might generate the suggested data. We focus on selecting the most suitable model among these suggested. Our findings offer conceptual and practical contributions. The conceptual contribution include a new framework for model selection of stochastic processes, which deviates from the classical *maximum likelihood* (ML) framework.

We present alternative criteria for model selection in MDPs. Our methods are then tested on a real world marketing problem where each client is modeled using a Markov chain and the goal is to optimize the company's mail requests to maximize future utility.

2. Setup

The setup is defined in the Markov decision process framework (Puterman, 1994); We begin with a formal definition:

Definition 2.1 *An observable* Markov Decision Process *(MDP) is a tuple* $(\mathcal{S},\mathcal{U},P,R,O)$*, where* $\mathcal{S}$ *is the state space set,* $\mathcal{U}$ *is the actions space,* $P:\mathcal{S}\times\mathcal{S}\times\mathcal{U}\mapsto[0,1]$ *is the transition probability function, the reward* $R\in\mathbb{R}$ *is a random variable dependant on the state and the action, and the observation* $O\in\mathcal{O}$*, where* $\mathcal{O}$ *is the observation space, is a random variable dependant on the state.*

The system dynamics are the following: in each time step $t=0,1,...$, the system is at some state $s_t\in\mathcal{S}$. An observation o_t is generated according to the current state and viewed as an output to the user. The user then chooses an action $u_t\in\mathcal{U}$. A reward r_t is generated according to the last state and action, and the state in the succeeding time step $t+1$ is chosen according to the transition matrix, s_t and u_t such that $s_{t+1}\sim P(\cdot|s_t,u_t)$. The time t is incremented by 1 and the process repeats itself.

Throughout this work, we assume some regularity conditions regarding the MDP since other cases are of less interest in our context. These conditions are summarized in the following assumptions.

Assumption 2.2 *For increasingly more data samples from the MDP, each state-action pair appears infinitely often.*

Appearing in *Proceedings of the* 30^{th} *International Conference on Machine Learning*, Atlanta, Georgia, USA, 2013.

Assumption 2.3 *The data were generated by applying a constant policy.*

Assumption 2.4 *For every $s \in \mathcal{S}, o \in \mathcal{O}$, if $P(o|s) > 0$ then $\forall s' \in \mathcal{S} \setminus \{s\} : P(o|s') = 0$, i.e., for each observation $o \in \mathcal{O}$ there is a unique possible state $s \in \mathcal{S}$ that it could have originated from, denoted by $s(o)$.*

Assumption 2.2 guarantees estimates of the MDP's parameters $P, \mathbb{E}[R]$ based on increasingly more samples will converge to their correct values. Assumption 2.3 guarantees estimates of the incorrect MDP's parameters will converge to some policy dependent value as well. Thus, these are crucial to the notion of weak consistency which will be presented later. Assumption 2.4 may seem too harsh and it is in fact used to simplify some technicalities. Moreover, in the framework we have in mind the observations hold excessive information on the state, which means Assumption 2.4 will hold at least with high probability on such cases.

Our basic setup is known as the offline batch setup: We observe a sequence of T observations, actions and rewards that occur in some space $\mathcal{O} \times \mathcal{U} \times \mathbb{R}$. The observation space $\mathcal{O}$ is possibly high dimensional, continuous, or processed in an unknown way that does not allow us to compute its probability density function. Denote the trajectory by

$$D_T = (o_1, u_1, r_1, o_2, u_2, r_2, \ldots, o_T, u_T, r_T). \quad (1)$$

These observations and rewards come from an underlying finite state MDP, denoted by M^*.

Definition 2.5 *A candidate MDP $M = (F^M, \mathcal{S}^M)$ is the empirically induced MDP by the mapping $F^M : \mathcal{O} \to \mathcal{S}^M$.*

In our problem formulation we are given K candidate MDPs $\{M^i\}_{i=1}^K$ where $M^i = (F^i, S^i)$. Each candidate is in fact a mapping that describes some underlying MDP. Following Assumption 2.4 we can define a true candidate model as one which perfectly represents the underlying state.

Definition 2.6 *Given data generated by an MDP M, a candidate MDP $M = (F^*, S^*)$ is defined to be the correct model if $\forall o_1, o_2 \in \mathcal{O} : s(o_1) = s(o_2)$ iff $F^*(o_1) = F^*(o_2)$.*

Note that we do not describe how the mappings $\{F^i\}_{i=1}^K$ are formed. Usually, these mappings are constructed by a domain expert who applies the appropriate methods for doing feature extraction. We can now define our setup of identification.

Definition 2.7 *A model selection criterion takes as input D_T and the candidate models $M^1, \ldots, M^K$, and chooses one of the K models as the proposed best model. We denote a generic model selector by $\hat{M}(D_T)$.*

We begin with a nesting assumption on the MDPs, which we relax in Section 6.

Assumption 2.8 *For all $i = 1, \ldots, K, 1 \leq j < i$ and $\forall o_1, o_2 \in \mathcal{O}$ if $F^i(o_1) = F^i(o_2)$ then $F^j(o_1) = F^j(o_2)$.*

In other words, Assumption 2.8 states that the candidate model M_i is a refinement of all candidate models $M_j, 1 \leq j < i$. When the nesting assumption holds, it is much easier to ascertain one candidate is preferable to another since the model selection problem becomes whether or not a group of states should be aggregated. In addition, although Assumption 2.8 seems harsh, hierarchical clustering algorithms naturally create a family of nested candidate models.

Finally, we give a formal definition of criterion's weak consistency which implies that for enough samples it will select the correct model.

Definition 2.9 *Consider a model M, a model selection criterion $\hat{M}(D_T)$ and a set of candidate models $\{M^i\}_{i=1}^K$. Define $\hat{M}(D_T)$ to be a* weakly consistent *criterion with respect to the given correct model and set of models, if for $1 \leq i \leq K$, $i \neq j$:*

$$\mathbb{P}^j\left(\hat{M}(D_T) = i\right) \to 0 \quad \text{as } T \to \infty,$$

where $\mathbb{P}^j$ is the induced probability when model j is the correct model.

We conclude this section with an example which will demonstrate the setup.

Example 2.10 *Consider an MDP $M = (\mathcal{S}, \mathcal{U}, P, R, O)$ with $\mathcal{S} = \{1, 2, 3\}, O = s + n_1, \mathcal{U} = \{u\}, R = s + n_2$, where $n_1 \sim U([-0.2, 0.2]), n_2 \sim \mathcal{N}(0, 1)$ and the transitions are uniform for the only action u. An observation realization may be:*

$$\begin{aligned} o &= (0.99, 1.98, 1.99, 3.0, 3.0, 3.0, 2.0, 2.0, 1.08), \\ r &= (0.9, 1.97, 2.06, 3.1, 2.9, 3.13, 2.07, 2.0, 1.0). \end{aligned} \quad (2)$$

Suppose we have 4 candidate models, $M^1, \ldots, M^4$, where the function F^i is the induced clustering from applying the k-means clustering algorithm (Duda et al., 2001) on the observations to i clusters, and the transition matrix and the reward for each such model are

found empirically from the induced trajectory. Expressing the states abstractly using the finest state space $\mathcal{S}^4 = \{a, b, c, d\}$ *yields*

$$D_T = \begin{pmatrix} a & a & a & a & a & a & a & a & a \\ a & b & b & b & b & b & b & b & a \\ a & b & b & c & c & c & b & b & a \\ a & b & b & c & c & c & b & b & d \end{pmatrix},$$

where line i *depicts the* i*'th model's induced trajectory.*

3. Previous Work

Previous works investigating model selection in Markov processes have largely focused on a single state space ((Fard & Pineau, 2010) and (Farahmand & Szepesvári, 2011)), selecting state representations in RL focusing on the regret (Maillard et al., 2011), or minimizing the errors of the Bellman operator (Farahmand & Szepesvári, 2011). Unlike our model based approach, these works focused largely on the Q-function.

There has also been substantial work on state aggregation in the RL literature, proposing different aliased states definitions (Li et al., 2006). Givan et al. (2003) suggested the bisimulation definition for aliased states which we adopt in this paper, but other aliasing definitions have been proposed as well (for example according to the Q-function in McCallum (1996) or policy invariance in Jong & Stone (2005)). Li et al. (2006) reviewed the different definitions and found relations between them. We see our work as another layer in unifying model selection theory as we focus on the offline problem where historical data are available.

Another aspect in which much work has been done is finding the aggregated states. For instance one can use the spectral properties of the transition matrix (see Mahadevan 2009 and references therein), while Ravindran (2003) suggested defining and finding aliased states using homomorphisms. In this aspect our work is most closely related to the works of Jong & Stone (2005) who proposed statistical testing on the Q-function, while we use them on the models' transition probabilities and rewards.

Finally, there are substantial amount of works on finding a good policy in a dynamic marketing environment. In their paper on catalog mailing policies, Simester et al. (2006) suggested a discretizing heuristic for a continuous state space with a geometric structure. Although our method of designing a state space is similar, we were able to provide some theoretical reasoning to it. (Pednault et al., 2002) conducted experiments showing that a dynamic policy on data from the KDD cup in 1998 (Hettich & Bay, 1999) outperforms a myopic policy which ignores the underlying dynamics. In contrast to this work and other works in this area, we focus on a rigorous method to build the state space which is based on the underlying dynamics.

4. Penalized Likelihood Criteria

Penalized Likelihood Criteria are criteria that measure the fitness of a model based on available data. Suppose we are given a parameterized set of candidate statistical models of degree i, $\{M^i(\theta)\}_{\theta\in\Theta}$, that describe the generation of data. A conventional way to choose between the models is to use *Maximum Likelihood Estimation* (MLE; (Duda et al., 2001)), which assumes that the best value for missing parameters is the one that maximizes the observations' probability. But in many cases, when comparing between models with a varying number of parameters, the MLE is prone to choose the model with the highest number of parameters.

The *Minimum Description Length* (MDL; (Rissanen, 1978)) principle is a formalization of the celebrated Occam's Razor principle that copes with the over-fitting problem. According to this principle, the best hypothesis for a given data set is the one that leads to the best compression of the data. Define the maximum likelihood (ML) of the model to be

$$L^i(T) = \max_\theta \{P(y_1, \ldots, y_T | M^i(\theta))\}.$$

We denote the dimension of θ by $|M^i|$. Then, an MDL-style model estimator has the following structure

$$\mathrm{MDL}(i) \triangleq |M^i| f(T) - \log L^i(T), \tag{3}$$

where $f(T)$ is some sub-linear function. In this model, the goal is to find i such that the MDL(i) is minimized. The rationale behind this criterion is simple: we look for for a model that best fits the data but is still "simple" in terms of missing parameters.

The most popular ones MDL-style criteria are Akaike Information Criterion ($\mathrm{AIC}(i) = 2|M^i| - 2\log L^i(T)$; (Akaike, 1974)) and Bayesian Information Criteria ($\mathrm{BIC}(i) = |M^i|\log(T) - \log L^i(T)$; (Schwarz, 1978)). We will show that in our setting, where the observations probabilities cannot be used due to their high dimensionality, continuous and processed nature, these criteria can fail to find the right model.

Theorem 4.1 *There exists an MDP for which an MDL criterion in the form of* (3) *is not consistent.*

Proof We construct an example for the general criterion (3). Suppose the correct model, M^*, is an MDP with a single action $\mathcal{U}^* = \{u\}$ and three states,

$\mathcal{S}^* = \{a, b, c\}$ where $\Pr(s_{t+1}|s_t, u) = 1/2$ if $s_{t+1} \neq s_t$. An illustration of the process is given in Figure 1. The reward function is $r(a) = 0$ and $r(b) = r(c) = 1$. Consider a candidate model, denoted by M^1, that is a single-state MDP. For the correct model M^* the likelihood will be for any trajectory affected only by the transitions. For the second model the likelihood will be for any trajectory affected only by the distribution of the rewards. A straightforward calculation yields:

$$\mathrm{L}^*(T) = 0.5^T, \quad \mathrm{L}^1(T) = (\frac{1}{3})^{\frac{T}{3}}(\frac{2}{3})^{\frac{2T}{3}} \approx 0.53^T.$$

Now, the likelihood ratio of the two models is:

$$\lim_{T\to\infty} \frac{\mathrm{L}^*(T)}{\mathrm{L}^1(T)} = 0.$$

Recalling the MDL-like criteria (3), we see that the penalizing term can be neglected asymptotically since it scales sub-linearly with T, while the logarithm of the likelihood ratio scales linearly. Therefore, the wrong model M^1 is chosen. The model M^1 is in fact a bad model to describe the data since the reward sequence of $r_t = 0, r_{t+1} = 0$ cannot appear in the actual data, yet the model M_1 allows it.

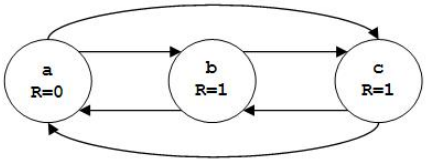

Figure 1. The example given in Theorem 4.1's proof.

We remark that this counter example follows the framework discussed above where the models' features can be thought of being constructed by a domain expert and therefore do not convey a particular probabilistic behavior. Although the true model M^* is one of the candidate models, the candidate model M^1 was chosen. In other words, the feature selection procedure done before applying the ML criterion leads to the ML approach failure to identify the right model. In the next section we propose an alternative criterion for choosing the right model and show that this criterion is consistent.

5. Aggregation Based Criterion

We begin with defining aliased states, followed by more intuitive explanation of this technical and lengthy definition. This definition is directly related to the containment relation in Assumption 2.8.

Definition 5.1 *Consider models M and $\tilde{M}$, where $\tilde{M}$ is a refinement of M, and with state spaces $S = \{s_1, \ldots, s_i\}$ and $\tilde{S} = \{\tilde{s}_1, \ldots, \tilde{s}_{i+k-1}\}$, respectively. Let P and $\tilde{P}$, be the transition matrices of M and $\tilde{M}$, respectively. Let $R(\cdot)$ and $\tilde{R}(\cdot)$ be the reward functions of M and $\tilde{M}$, respectively. Define C to be the set of states common to both S and $\tilde{S}$ (i.e., the mappings from observations to states have the same inverse image for any one of these states), and let $s^* \in S$ be aggregation of k states in $\tilde{S}$, denoted by A, such that $C \bigcup \{s^*\} = S$ and $C \bigcup A = \tilde{S}$. Suppose that*

1. $\tilde{P}(c_2|c_1, u) = P(c_2|c_1, u), \quad \forall c_1, c_2 \in C, u \in \mathcal{U}$;
2. $\sum_{a\in A} \tilde{P}(a|c, u) = P(s^*|c, u), \quad \forall c \in C, u \in \mathcal{U}$;
3. $\tilde{P}(c|a_1, u) = \tilde{P}(c|a_2, u), \quad \forall c \in C, a_1, a_2 \in A, u \in \mathcal{U}$;
4. $\sum_{a\in A} \tilde{P}(a|a_1, u) = \sum_{a\in A} \tilde{P}(a|a_2, u) \quad \forall a_1, a_2 \in A, u \in \mathcal{U}$;
5. $\tilde{R}(a_1, u) \sim \tilde{R}(a_2, u), \forall a_1, a_2 \in A, u \in \mathcal{U}$.

Then, we say that the states A in model $\tilde{M}$ are aliased with respect to model M *(or simply* aliased*).*

The notion of aliased states can be interpeted as follows. In model M, there is a state, s^*, that is split into k states in model $\tilde{M}$ (denoted by A). I.e, if we take the states belonging A, and we cannot provide a statistical test that differentiate between them (conditions 2-5) based on the MDPs parameters, then for all practical purposes we can aggregate these states. For example, the value function for two MDPs that differ by having aliased states is the same (Givan et al., 2003).

Testing whether a two states are aliased or not can be done using hypothesis testing (Cover & Thomas, 2006) on the empirical probabilities and average rewards. Let A^i be the set of possibly aliased states in model M^i, $\hat{p}^{(i)}_{kj,u}$ be the empirical probability for the transition from state k to state j in model i after choosing action u, and $\hat{r}^{(i)}_{k,u}$ be the empirical reward of choosing action u in state k. An examination of conditions $1-5$ is now needed, where conditions 1 and 2 are trivially satisfied from the nesting assumption. Define:

$$
\begin{aligned}
h_1^{(i)} &\triangleq \bigcap_{j\in C,l,m\in A^i,u\in\mathcal{U}} \left\{\left|\hat{p}_{lj,u}^{(i)} - \hat{p}_{mj,u}^{(i)}\right| < \epsilon^{i,lm,u}\right\}, \\
h_2^{(i)} &\triangleq \bigcap_{l,m\in A^i,u\in\mathcal{U}} \left\{\left|\sum_{j\in A}\hat{p}_{lj,u}^{(i)} - \sum_{j\in A}\hat{p}_{mj,u}^{(i)}\right| < \epsilon^{i,lm,u}\right\}, \\
h_3^{(i)} &\triangleq \bigcap_{l,m\in A^i,u\in\mathcal{U}} \left\{\left|\hat{r}_{l,u}^{(i)} - \hat{r}_{m,u}^{(i)}\right| < \epsilon^{i,lm,u}\right\},
\end{aligned} \tag{4}
$$

where $\{\epsilon^{i,lm,u}\}_{l,m\in A^i,u\in\mathcal{U},i=2..K}$ are tolerance parameters that are to be determined according to the desired level of error balancing different sources of error. The value of ϵ represents a tradeoff: if it is too large we may choose a model that is too refined while if it is too small we may choose a model that is too fine.

We note that $h_1^{(i)}, h_2^{(i)}$, and $h_3^{(i)}$ are the empirical analogies to conditions 3-5 above. Define $H_{i-1,i} \triangleq h_1^{(i)} \bigcap h_2^{(i)} \bigcap h_3^{(i)}$ to be the event that models M_{i-1} and M_i are statistically aliased. Based on this, we define a comparison test:

$$C_i = \mathbf{1}\{\text{Outcome contained in } H_{i-1,i}\},$$

and the model selector in this case is

$$\hat{M}_C = \max_i \{i : C_i = 0\}. \tag{5}$$

I.e., it is the first index for which aliased states are identified. For clarity, we summarize how to use our proposed model selection criterion (5). We set the tolerance parameters $\{\epsilon^{i,lm,u}\}_{l,m\in A^i,u\in\mathcal{U},i=2..K}$ for each test to a value depending on the type of significance test (proportions / mean) and the desirable significance level. Specifically, we set the tolerance in the following manner:

$$\lim_{T\to\infty} \epsilon_T^{i,lm,u}\sqrt{T} = \infty, \quad \lim_{T\to\infty} \epsilon_T^{i,lm,u} = 0, \tag{6}$$

in order to guarantee consistency as shown. Next we compute $h_1^{(i)}$, $h_2^{(i)}$ and $h_3^{(i)}$ for each pair of consecutive candidate models $(i-1, i)$. Based on their value we compute the event $H_{i-1,i}$. Then, we try to identify the greatest index i such that $C_i = 0$, i.e., identifying the finest model that does not contain aliased states.

We conclude this section with a theorem that states that the criterion in (5) is weakly consistent.

Theorem 5.2 *Suppose Assumptions 2.2 and 2.8 hold and that the correct model contains no aliased states. In additon, assume $\{\epsilon^{i,lm,u}\}_{l,m\in A^i,u\in\mathcal{U},i=2..K}$ are chosen as specified in Eq. (6). Then, for any set of candidate models the model selector $\hat{M}_C$ is weakly consistent.*

Due to space limitation, this proof and the rest of the proofs in the paper are omitted and will be available in a more extensive version of this work.

6. Extension to arbitrary candidates set

In Section 5 we used Assumption 2.8 that requires a containment relation between the models. Yet, strict containment between models is a harsh assumption that will not always hold. In this section we show that we can still establish consistency when the set of candidate models $\mathcal{M}$ has no structure. We emphasize that we still assume that one of the candidate models is the true model.

We begin by formalizing the nested approach in partial order formulation (similarly to Li et al. 2006).

Definition 6.1 *For two candidate models M^1 and M^2 define the* aggregation order*: $M^1 <_{Agg} M^2$ if aliased states in M^1 can be aggregated to obtain M^2.*

It is easy to see the $<_{Agg}$ order is partial, and that the aggregation criterion $\hat{M}_C$ is equivalent to choosing the candidate model with the least number of states among all the maxima candidates in the given set of nested models. We can fix the aggregation order such that the aggregation criterion will simply choose the only maximum as the correct model in any given set.

Definition 6.2 *For two candidate models $M^1 = (F^1, \mathcal{S}^1)$ and $M^2 = (F^2, \mathcal{S}^2)$ define the* fixed aggregation order *as following: let $M^{1\times 2} = ((F^1, F^2), \mathcal{S}^1 \times \mathcal{S}^2)$, then $M^1 <_{fAgg} M^2$ if $M^{1\times 2} <_{Agg} M^2$ and not $M^2 <_{Agg} M^1$.*

The motivation behind Definition 6.2 is the following: Assume that we compare the correct model M^1 and some other model M^2. Since the correct model contains all the information on the system's dynamics, it is unnecessary to use the other model as an additional information source by looking at $M^{1\times 2}$. Therefore $M^{1\times 2}$ can be aggregated to the correct model M^1. In other words, the fixed aggregation order asserts whether one model contains all the information on the dynamics that is contained by the other model.

Like the original aggregation order, we can expand the fixed aggregation order to a model selection criterion and show it is weakly consistent.

Definition 6.3 *Given a set of models* $\{M^i\}_{i=1}^K$ *define the fixed aggregation criterion:*

$$\hat{M}_{fAgg} = \arg\max_{<_{fAgg}} \{M^i\}. \quad (7)$$

Theorem 6.4 *Suppose that Assumption 2.2 holds and that the correct model contains no aliased states. If the tolerance parameters are chosen as specified in Eq. (6), then for any set of candidate models the model selector* $\hat{M}_{fAgg}$ *is weakly consistent.*

7. Reward based criteria

In the previous sections we introduced two aggregation based orders. However, in the improper case when the correct model is not in the given set of candidate models aggregation based criteria hold no ground. In this section we suggest another reward-based criterion that has a meaning in the predictive sense on the MDP.

Definition 7.1 *For a given model* M*, a trajectory* $D_T = (o_t, a_t, r_t)_{t=1}^T$ *and a constant* $d \in \mathbb{N}_0$ *define the d-delayed Reward Error (*RE_d*) value as*

$$RE_d(M) = \frac{1}{T}\sum_{t=1}^{T-d}(r_{t+d} - \hat{\mathbb{E}}[R_{t+d}|s_t, a_t])^2 + |\mathcal{S}|\frac{f(T)}{T}, \quad (8)$$

where $\hat{\mathbb{E}}[R_{t+d}|s_t, a_t]$ *is the empirical expectation of rewards obtained from the state-action pair* (s_t, a_t) ***after*** *d steps, and* $f(T)$ *is a sublinear function that satisfies* $\lim_{T\to\infty}\frac{f(T)}{\sqrt{T}} = \infty$.

The RE_d score for a given model is the reward prediction error, with an additional penalty function which prevents empirical fluctuations from tilting the score to more refined models.

Definition 7.2 *The* RE_d *order is the induced order by the* RE_d *score, and the* RE_d *model criterion as selecting the minimal model with respect to the* RE_d *order.*

Observe for instance the example given in Theorem 4.1's proof. The rewards for the correct model M^* are deterministic, while the rewards for the one-state model M^1 are distributed Bernoulli(1/3). Therefore, $\mathrm{RE}_0(M^*) = 0 + 3\frac{f(T)}{T}$ and $\mathrm{RE}_0(M^1) = \frac{2}{9} + \frac{f(T)}{T}$, and the chosen model asymptotically will be M^*.

Theorem 7.3 *If* $\forall s^1, s^2 \in \mathcal{S} : \mathbb{E}[R_{t+d}|s_t = s^1] \neq \mathbb{E}[R_{t+d}|s_t = s^2]$*, then the* RE_d *criterion is weakly consistent.*

The RE_0 criterion was suitable for the example in Theorem 4.1's proof, but it will often fail in real world problems where the rewards are sparse, which means many candidate models will have the same RE_0 value. For example, in (Simester et al., 2006) the reward is zero in most of the states. In this case higher values of d can be used, since these include the dynamics of the system as well as the immediate rewards. While on one hand the d-step reward is spread over more states and therefore might be less distinctive, it originates from the transition probabilities and therefore considers model information not available in the RE_0 criterion. An example where the RE_0 criterion fails but the RE_1 criterion works is illustrated in Figure 2.

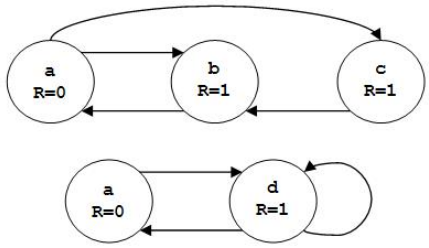

Figure 2. An example where RE_0 fail and RE_1 succeeds.

In Figure 2, the upper drawing is the correct single-action MDP. Assume that the data are generated from the given MDP, and two candidate models: The correct model M^1, and another model M^2 given in the lower drawing with 2 states - a and another state d - the aggregation of the states b and c. According to the RE_0 criterion, both models will produce the same score and thus the wrong model M^2 will be chosen since it contains less states. However, applying the RE_1 criterion we obtain asymptotically that $\mathrm{RE}_1(M^1) = 0$ while $\mathrm{RE}_1(M^2) > 0$, i.e., the RE_1 criterion will select the correct model relying on enough data.

8. Experiments

8.1. Simulated data

We simulated an MDP with 20 non aliased states with noisy rewards and observations consisting of 7 independent features. Next, we generated two data trajectories using the simulator. Using Matlab's k-means clustering algorithm (Duda et al., 2001) on the observations from the first trajectory we constructed candidate models of increasing state space size from 2 to 40, where the candidate model of size 20 was set to be the correct model. The first trajectory was only used to create data independent candidate models.

The second trajectory was used for evaluation of different MDL criteria, $\mathrm{RE}_0/\mathrm{RE}_1$ criteria and the optimal average value function based on the estimated model.

This simulation process was averaged over 100 simulations, and we used trajectories of different sizes - 100, 1K and 10K. The results are shown in Figure 3.

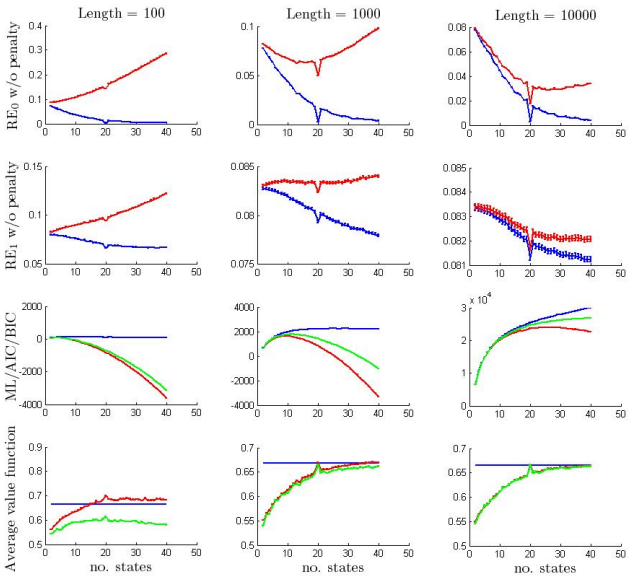

Figure 3. Performance of the different criteria on simulated data. [1st, 2nd row] The RE_d score with / without (red/blue) the regularizing summand. [3rd row] The ML, AIC and BIC (blue/red/green) scores. [4th row] The correct value for the optimal policy, the value for the estimated optimal policy and its real value (blue/red/green).

It seems that the RE_d works best among the inspected criteria on our simulations. The penalized MDL scores favors overly refined models for increasingly more data. The value function exhibits an interesting property - when there's not enough data the estimated value is higher than the correct one. This phenomenon is more severe for more refined state spaces, i.e. sometimes choosing a smaller incorrect model can lead to better performance. With that in mind, the value function itself can be used as a model selection criterion, perhaps with some additional regularization summand.

8.2. KDD Cup 1998 Data

As a test bench, we used the donation data set from the KDD Cup 1998 competition (Hettich & Bay, 1999) in which the goal was to estimate the return for a direct mailing task. As observations we used the first 8 features given by (Pednault et al., 2002) with some rescaling. We then tested the different criteria similarly to before: we used a small portion of the data (1K trajectories of length 22) to construct candidate models using k-means. In order to compensate for unknown penalty for frequent donation requests, we have decreased the reward for sending a donation request by 2. Over 100 simulations, we randomly chose the data from which candidate models are formed, and used the most of what's left of the data (8K trajectories) to evaluate the different criteria on the proposed models. The remaining 1K trajectories were used to estimate the optimal/myopic policy for the infinite horizon value function with a discount factor 0.9 (normalized to $[0, 1]$). The results are shown in Figure 4.

It is important to emphasize that in our scheme of cross validation, instead of using the same data to construct the state space and to estimate the induced MDP, we used disjoint parts from the data. When the state space is constructed only according to the observations, this partition is not necessary. However, building the state space according to the dynamics of the problem and then estimating the same dynamics yields a statistical dependence which undermines the generality of the proposed solution.

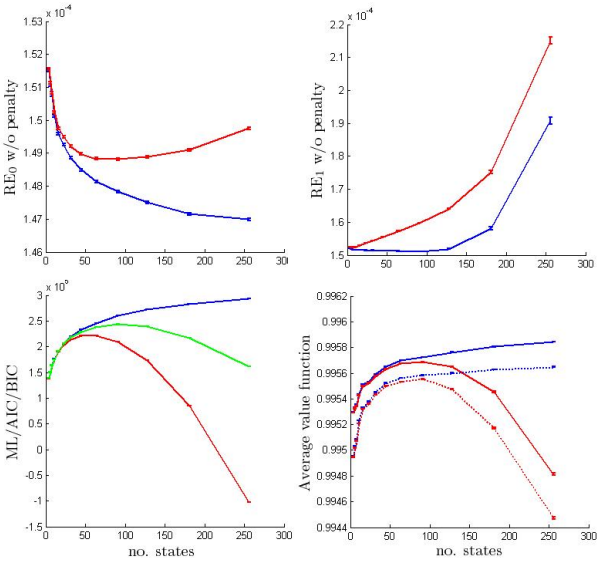

Figure 4. Performance of the different criteria on data acquired from the KDD cup 1998, where the state space was constructed by k-means clustering. [1st row] The RE_d score with/without the regularizing summand (red/blue). [2nd row, left] The ML, AIC and BIC (blue/red/green) scores. [2nd row, right] The estimated value for the optimal policy, estimated value for the greedy policy (blue/dashed blue), and their sampled value on the general population (red/dashed red).

In our results, it seems that all criteria point towards state spaces with roughly 60 states, supporting each of the suggested criteria. In addition, as was shown before (Pednault et al., 2002), dynamic policy distinctively outperform a myopic policy. Another interesting

property is the saturation behavior of the estimated value function, while its cross evaluation receives its maximum and decreases significantly afterwards. This phenomenon can be described as overfitting - the found optimal policy is less accurate since the number of samples per state decreases.

9. Conclusions

Estimating or optimizing a Markov decision process requires three steps: identifying the correct model, estimating the parameters, and applying an optimization algorithm. While considerable research has been conducted on estimation procedures and optimization algorithms (Singh et al., 2009), much less work has been done on identifying the right model. In this paper we propose a framework for statistical identification of Markovian models from data.

Our work concentrated mainly on asymptotic notions and definitions. Yet, providing finite sample analysis for the proposed criteria is not hard as we employ standard tools of statistical hypothesis testing. As a result, the tolerance parameters can be chosen in a simple fashion and exponential bounds on the error probabilities can be derived.

In our experiments, we had examined different model selection criteria. The aggregation criteria has good theoretical guarantees and the reward based criteria showed results as good as ML based methods. We extended our hypothesis testing method to build a consistent model construction algorithm that works in manageable complexity. Finally, our methods were used on real world donation data from the KDD cup 1998, yielding interesting results.

References

Akaike, H. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974.

Cover, TM and Thomas, J. Elements of information theory, 2nd Ed. 2006.

Duda, R.O., Hart, P.E., and Stork, D.G. *Pattern classification*, volume 2. wiley New York:, 2001.

Farahmand, A. and Szepesvári, C. Model selection in reinforcement learning. *Machine Learning*, pp. 1–34, 2011.

Fard, M.M. and Pineau, J. PAC-Bayesian model selection for reinforcement learning. *Advances in Neural Information Processing Systems (NIPS)*, 2010.

Givan, R., Dean, T., and Greig, M. Equivalence notions and model minimization in Markov decision processes. *Artificial Intelligence*, 147(1-2):163–223, 2003.

Hettich, S. and Bay, S. D. The UCI KDD Archive, 1999. URL http://kdd.ics.uci.edu.

Jong, N.K. and Stone, P. State abstraction discovery from irrelevant state variables. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence*, pp. 752–757, 2005.

Li, L., Walsh, T.J., and Littman, M.L. Towards a unified theory of state abstraction for MDPs. In *Proceedings of the Ninth International Symposium on Artificial Intelligence and Mathematics*, pp. 531–539, 2006.

Mahadevan, S. *Learning Representation and Control in Markov Decision Processes*. Now Pub, 2009.

Maillard, O.A., Munos, R., and Ryabko, D. Selecting the State-Representation in Reinforcement Learning. In *Advances in Neural Information Processing Systems*, 2011.

McCallum, A.K. *Reinforcement learning with selective perception and hidden state*. PhD thesis, University of Rochester, 1996.

Pednault, E., Abe, N., and Zadrozny, B. Sequential cost-sensitive decision making with reinforcement learning. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 259–268. ACM, 2002.

Puterman, M.L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, Inc., 1994.

Ravindran, B. SMDP homomorphisms: An algebraic approach to abstraction in semi markov decision processes. 2003.

Rissanen, J. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

Schwarz, G. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

Simester, D.I., Sun, P., and Tsitsiklis, J.N. Dynamic catalog mailing policies. *Management science*, 52 (5):683, 2006.

Singh, S., Lewis, R.L., and Barto, A.G. Where do rewards come from. In *Proceedings of the Annual Conference of the Cognitive Science Society*, pp. 2601–2606. Citeseer, 2009.