Model Selection in Markovian Processes

Assaf Hallak
Technion
Haifa, Israel
ifogph@gmail.com

Dotan Di-Castro
Technion
Haifa, Israel
dotan.dicastro@gmail.com

Shie Mannor
Technion
Haifa, Israel
shie@ee.technion.ac.il

ABSTRACT

When analyzing data that originated from a dynamical system, a common practice is to encompass the problem in the well known frameworks of Markov Decision Processes (MDPs) and Reinforcement Learning (RL). The state space in these solutions is usually chosen in some heuristic fashion and the formed MDP can then be used to simulate and predict data, as well as indicate the best possible action in each state. The model chosen to characterize the data affects the complexity and accuracy of any further action we may wish to apply, yet few methods that rely on the dynamic structure to select such a model were suggested.

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. We then define weak consistency for a model construction algorithm and show a simple algorithm that is weakly consistent. Finally, we test the performance of the suggested criteria and algorithm on both simulated and real world data.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Markov Processes; J.1 [**Computer Applications**]: Administrative Data Processing—*Marketing*

Keywords

Model Selection, Reinforcement Learning, Markov Decision Processes, Dynamic Mailing Policies

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 almost always 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. Our proposed framework is more suitable for the path that engineers usually undertake: (1) receiving the raw data; (2) applying different preprocessing, discretization and feature selection procedures; and (3) choosing the model that represents their data most faithfully. In contrast, in the ML framework all these stages are performed at once *based on the observations directly*. Thus, integrating domain expert knowledge regarding the feature selection is more difficult.

Next, we discuss the practical contribution. A natural question that arises in this context is the following: *Does an ML based approach still yields a reasonable result?* The first result in this paper establishes that this standard approach, which works well in some settings, may fail to choose the correct model for MDPs. We then present alternative criteria for model selection in MDPs, one that is based on transitions and one that is based on rewards; We show that these criteria are consistent under appropriate assumptions. In addition, the computation of these criteria scales linearly with the size of the data set and they contain a natural way of regularizing the number of states according to the amount of data available. At last, these criteria can be extended to build a simple model construction algorithm which converges to a refinement of the correct state space.

Finally, we make use of our methods in a marketing problem in which a firm decides whether to send each client a mail, and the reward depends on the client's response. Specifically, we examine the data from the KDD cup in 1998 [9], where donation requests were sent to many individuals

KDD'13, August 11–14, 2013, Chicago, Illinois, USA.

and the reward was based on the donation received. We construct candidate models of different sizes using a simple clustering algorithm to inspect the behavior of the different criteria, and examine the performance of our own model construction in comparison.

The paper is organized as follows. In Section 2 we describe the setup and define the notations. In Section 3 we review previous research. In Section 4 we discuss penalized maximum likelihood based criteria and show that they are not necessarily consistent in MDPs. In Section 5 we describe a criterion for choosing models among a nested set, where in Section 6 we expand the results to any general case. We propose different reward based criteria in Section 7. Section 8 presents the notion of weak consistency for algorithms that build a specific model, as well as a simple algorithm that is weakly consistent. In Section 9 we illustrate the findings on simulated and real-world data. We conclude in Section 10.

2. SETUP

The setup is defined in the Markov decision process framework [16]; We begin with a formal definition:

Definition 1. A *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 1. For increasingly more data samples from the MDP, each state-action pair appears infinitely often.

Assumption 2. The data were generated by applying a constant policy.

Assumption 3. 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 1 guarantees estimates of the MDP's parameters $P, \mathbb{E}[R]$ based on increasingly more samples will converge to their correct values. Assumption 2 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 3 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 3 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

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

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

Definition 2. 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 3 we can define a true candidate model as one which perfectly represents the underlying state.

Definition 3. 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 4. 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 4. 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 4 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 4 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 5. 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 1. 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:

$$o = (1.0, 1.01, 0.99, 1.98, 1.99, 3.0, 3.0, 3.0, 2.0, 2.0, 1.08),$$
$$r = (0.86, 1.05, 0.9, 1.97, 2.06, 3.1, 2.9, 3.13, 2.07, 2.0, 1.0).$$

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 [4] on the observations to i clusters, and the transition matrix and the reward for each such model are found empirically from the induced trajectory. In our case, for M^4 the centers vector is $(1, 1.9925, 3, 1.07)$, for M^3 it is $(1.028, 1.9925, 3)$, for M^2 the centers vector is $(1.028, 2.4243)$, and for M^1 the center is (1.842). Therefore, expressing the states abstractly using the finest state space $\mathcal{S}^4 = \{a, b, c, d\}$ yields

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

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

3. PREVIOUS WORK

Previous research on model selection for dynamic random processes includes works on Hidden Markov Models (HMMs; Elliott et al. 5) and Dynamic Bayesian Networks (DBNs; Dean and Kanazawa 3). In HMMs, one obtains (corrupted) observations of a Markov process, where the observations may be a stochastic function of the underlying process. The goal is to find the best model that describes the underlying process. The problem of identification in DBNs is to find a graphical model that compactly describes the relation between the components of a multivariate random process. Our setting is somewhat different as we consider observations which may have gone through preprocessing and from which domain experts suggest different mappings to candidate state spaces.

More recent works investigating model selection in Markov processes have largely focused on a single state space (see, for example, Farahmand and Szepesvári 6, Fard and Pineau 7), selecting state representations in RL focusing on the regret; see [13], or minimizing the errors of the Bellman operator [6]. These works focused largely on the Q-function rather than on the model selection thus following a different approach from ours.

There has also been substantial work on state aggregation in the RL literature, proposing different aliased states definitions [11]. Givan et al. [8] 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 [14] or policy invariance in Jong and Stone [10]). Li et al. [11] 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 12 and references therein), while Ravindran [17] suggested defining and finding aliased states using homomorphisms. In this aspect our work is most closely related to the works of Jong and Stone [10] 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. [20] 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. [15] conducted experiments showing that a dynamic policy on data from the KDD cup in 1998 [9] 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 have a statistical model M that produces data samples $y_1, y_2, \ldots, y_T$. We are given a parameterized set of candidate statistical models of degree i that describe the generation of data denoted by $\{M^i(\theta)\}_{\theta \in \Theta}$. A conventional way to choose between the models is to use *Maximum Likelihood Estimation* (MLE; Duda et al. 4), 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 over-fit the data, i.e., it chooses the model with the highest number of parameters.

The *Minimum Description Length* (MDL; [18]) 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

$$\mathrm{MDL}(i) \triangleq |M^i| f(T) - \log L^i(T), \tag{2}$$

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 a model that best fits the data but is still "simple" in terms of missing parameters.

Many MDL-style criteria exist and some of them were developed from an information theory perspective, we mention the two most popular ones as we later compare them to our algorithm. The first is the Akaike Information Criterion (AIC; Akaike 1). This criterion has the form AIC$(i) = 2|M^i| - 2 \log L^i(T)$ and it tries to minimize the Kullback-Leibler divergence between the statistics of the true model and the candidate model. The second criterion is the Bayesian Information Criteria (BIC; Schwarz 19) that has the form

$\mathrm{BIC}(i) = |M^i| \log(T) - \log L^i(T)$ and is similar in its nature to the AIC but was developed in a Bayesian framework. We will next 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. We do so by presenting an example that shows the counter-intuitive behavior of standard MDL criteria.

THEOREM 1. *For MDPs, there does not exist a consistent MDL-style criterion in the form of* (2).

PROOF. We construct a counter example for the general criterion (2). 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 (2), 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. □

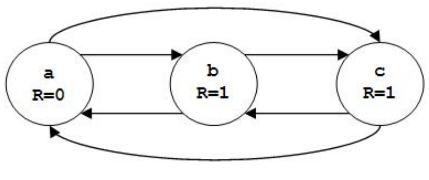

Figure 1: The counterexample given in Theorem 1's proof.

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 4.

Definition 6. 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. $\tilde{P}\left(c_2 | c_1, u\right) = P\left(c_2 | c_1, u\right), \quad \forall c_1, c_2 \in C, u \in \mathcal{U};$
2. $\sum_{a \in A} \tilde{P}\left(a | c, u\right) = P\left(s^* | c, u\right), \quad \forall c \in C, u \in \mathcal{U};$
3. $\tilde{P}\left(c | a_1, u\right) = \tilde{P}\left(c | a_2, u\right), \quad \forall c \in C, a_1, a_2 \in A, u \in \mathcal{U};$
4. $\sum_{a \in A} \tilde{P}\left(a | a_1, u\right) = \sum_{a \in A} \tilde{P}\left(a | a_2, u\right) \quad \forall a_1, a_2 \in A, u \in \mathcal{U};$
5. $\tilde{R}\left(a_1, u\right) \sim \tilde{R}\left(a_2, u\right), \forall a_1, a_2 \in A, u \in \mathcal{U}.$

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

Discussion: Intuitively, the meaning of aliased states is the following. In model M, there is a state, s^*, that is split into k states in model $\tilde{M}$ (denoted by A). Condition 1 suggests that transitions between states that are not in A are the same in both models. If the probabilities related to S^* in model M and the states A in $\tilde{M}$, satisfy conditions 2-5 we have aliased states. In other words, if we take the states that belong to 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 and get the same result on M and $\tilde{M}$. For example, computing the value function for two MDPs that differ by having aliased states yields the same result [8]. Based on this, a natural criterion for identifying the right model is the following. We look for a model that *best fits the data*, but does not contain any aliased states which unnecessarily complicate it.

How to test whether two states are aliased? A criterion for that may be the following. The observer examines the empirical probabilities, analogously to those of Definition 6, of the candidate aliased states. Then, using a *significance test* (or *hypothesis testing*; see Cover and Thomas 2) it is decided whether these states are aliased. I.e., the comparison between models is not carried out by applying a scalar score on the models (an MDL-like score), but by comparing two models directly and doing some statistical test.

The statistical test examines if a finer model adds information comparing to the coarser model. If so, and if the finer model does not have aliased states, then the observer may choose the highest order model that does not contain aliased states. We formally summarize this test for two models.

The idea in the base of statistical testing is the following. 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: Conditions 1 and 2 are trivially satisfied from the nesting assumption, but the rest of the conditions have to be tested.

Define

$$\begin{aligned}
h_1^{(i)} &\triangleq \left\{ \left| \hat{p}^{(i)}_{lj,u} - \hat{p}^{(i)}_{mj,u} \right| < \epsilon^{i,lm,u}, \forall j \in C, \forall l, m \in A^i, \forall u \in \mathcal{U} \right\}, \\
h_2^{(i)} &\triangleq \left\{ \left| \sum_{j \in A} \hat{p}^{(i)}_{lj,u} - \sum_{j \in A} \hat{p}^{(i)}_{mj,u} \right| < \epsilon^{i,lm,u}, \forall l, m \in A^i, \forall u \in \mathcal{U} \right\}, \\
h_3^{(i)} &\triangleq \left\{ \left| \hat{r}^{(i)}_{l,u} - \hat{r}^{(i)}_{m,u} \right| < \epsilon^{i,lm,u}, \forall l, m \in A^i, \forall u \in \mathcal{U} \right\},
\end{aligned} \tag{3}$$

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

$$\hat{M}_C = \max_i \{i : C_i = 0\}. \tag{4}$$

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 (4). 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. In effect, if the tolerance is set too high then the test will falsely mark non-aliased states as aliased, however low tolerance can cause failure to identify aliased states when data are limited. Specifically, we set the tolerance in the following manner:

$$\lim_{T \to \infty} \epsilon_T^{i,lm,u} \sqrt{T} = \infty, \quad \lim_{T \to \infty} \epsilon_T^{i,lm,u} = 0, \tag{5}$$

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.

As a final note, we point out that hypothesis testing could have been done using other methods. For example, we could have used χ^2 score to compare the transition probabilities of different states, this choice was done arbitrarily. We could have also used some characteristic of the reward distribution other than the expectation, but since in MDPs this is the deciding factor our choice here is probably the most suited.

We conclude this section with a theorem that states that the criterion in (4) is weakly consistent. The proof is a technical use of Hoeffding's inequality and is therefore omitted.

Theorem 2. *Suppose Assumptions 1 and 4 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. (5). Then, for any set of candidate models the model selector $\hat{M}_C$ is weakly consistent.*

6. EXTENSION TO ARBITRARY CANDIDATES SET

In Section 5 we used Assumption 4 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. 11).

Definition 7. 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 8. 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 $M^{1\times2} = ((F^1, F^2), \mathcal{S}^1 \times \mathcal{S}^2)$, then $M^1 <_{fAgg} M^2$ if $M^{1\times2} <_{Agg} M^2$ and not $M^2 <_{Agg} M^1$.

The motivation behind Definition 8 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 $M^{1\times2}$. Therefore $M^{1\times2}$ 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 9. Given a set of models $\{M^i\}_{i=1}^K$ define the fixed aggregation criterion:

$$\hat{M}_{fAgg} = \underset{<_{fAgg}}{\arg\max} \left\{ M^i \right\}. \tag{6}$$

Theorem 3. *Suppose that Assumption 1 holds and that the correct model contains no aliased states. In addition, assume that the tolerance parameters are chosen as specified in Eq. (5). Then, for any set of candidate models the model selector $\hat{M}_{fAgg}$ is weakly consistent.*

A sketch of the proof is as follows: We prove that if the correct model has no aliased states it is strictly $<_{fAgg}$ bigger than any other candidate model by going over the possible nesting relations between the two models. The main difficulty is to show that no model can be $<_{fAgg}$ bigger than the correct model; we solve this by proving that it contradicts the assumption there are no aliased states.

To conclude this section we would like to discuss the computational aspect of our solution. In order to find the correct model among a set of given models we need to find the maximum in this set with respect to the suggested order. When two models cannot be compared using the $<_{fAgg}$ order, evidently none of them can be correct therefore the maximum can be found in a single sweep on the candidates. However, the computation of the order between any two models can be expensive since naively it requires finding aliased states from $|S_1| \cdot |S_2|$ states. Even so, there are cases when only few states from $|S_1| \cdot |S_2|$ exists; For example, if the models are nested then there are only $\max(|S_1|, |S_2|)$ different states.

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 10. 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

$$\mathrm{RE}_d(M) = \frac{1}{T}\sum_{t=1}^{T-d}(r_{t+d} - \hat{\mathbb{E}}[R_{t+d}|s_t, a_t])^2 + |\mathcal{S}|\,\frac{f(T)}{T}, \quad (7)$$

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. Another important property of the RE$_d$ score is that if two sets of data were generated from different policies, asymptotically their RE$_d$ score would be different even for the correct model. We can formalize a RE$_d$ based criterion by trying to minimize it:

Definition 11. Define the RE$_d$ order as 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. If there are multiple candidate models achieving the minimal value, then the RE$_d$ criterion chooses arbitrarily among these with the least number of states.

Observe for instance the example given in the proof of Theorem 1. 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, we obtain that $\mathrm{RE}_0(M^*) = 0 + 3\frac{f(T)}{T}$ and $\mathrm{RE}_0(M^1) = \frac{2}{9} + \frac{f(T)}{T}$. So that the chosen model asymptotically will be M^*.

THEOREM 4. *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.*

A proof sketch is as follows: we would like to show that the minimal RE$_d$ is achieved by the correct model. Refinements can be shown to have asymptotically lower RE$_d$ score due to the penalty function added. Models that are neither the correct model nor its refinements necessarily contain an abstract state that originated from two original states. According to the assumption in the theorem, these two states ought to have different expected rewards. Therefore, the estimated mean reward for the abstract space is composed of two different means and its prediction will yield higher error than estimating these means separately.

The RE$_0$ criterion was suitable for the example in Theorem 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 [20] 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.

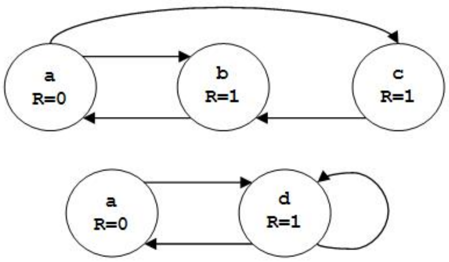

Figure 2: An example where RE$_0$ fail and RE$_1$ succeeds.

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 which is 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. MODEL CONSTRUCTION

In this section we expand the notion of consistency to algorithms that construct one specific candidate model. We begin with a formal definition of a model construction algorithm:

Definition 12. A model construction algorithm $\mathcal{A}$ is given an input data trajectory D_T, and returns a candidate model $M = (F, S)$, i.e., a mapping $F : \mathcal{O} \to \mathcal{S}$.

Following Assumption 3, we can define a model construction algorithm to be weakly consistent by demanding that for increasingly more data the constructed mapping will converge to the true mapping. Different partitions on the observations' space might use different state spaces, so a logical way of comparing two such partitions is by checking their agreement on pairs of observations. Since we allow the constructed model to be a refined version of the correct one, we define weak consistency as follows:

Definition 13. Assume Assumptions 1, 2 and 3, and denote $F_T = \mathcal{A}(D_T)$. We define a model construction algorithm to be weakly consistent if:

$$\mathbb{P}\left(F_T(o_1) = F_T(o_2), F^*(o_1) \neq F^*(o_2)\right) \to 0 \quad \text{as } T \to \infty,$$

where the probability density over the observation o is the induced probability from the stationarity of the process.

A trivial example of a weakly consistent model construction is one that assigns each new observation to a new state. However, this property is immidate: general observations based clustering methods are not weakly consistent. We present a non trivial algorithm that is weakly consistent.

Algorithm 1 Naive model construction algorithm

```
1: Assume all observations belong to the same state.
2: Choose a state and a feature $d$ that were not chosen
   before.
3: Find the median of the observations in the current states
   according to this feature.
4: Partition the current state to two states, to one of them
   add the observations in the state holding $o_t(d)$ <median
   and all the rest add the other.
5: If the states are aliased according to our hypothesis test-
   ing, reunite them.
6: If there are more states and features not visited, or if
   you reached a predefined desirable number of states, go
   back to step 2.
```

THEOREM 5. *Assume the observations have a continuous distribution and that all state space partitions to non aliased states results in two states with different mixtures of original states, then Algorithm 1 is weakly consistent.*

An intuitive explanation as to why this simple decision tree [4] based algorithm is weakly consistent is as follows: Once a state generated by Algorithm 1 contains only observations coming from the same state, it will not be divided anymore. However, if a state contains observations from several distinct states given enough samples it will be divided since each half of the samples contain a different mixture of original states. Each such division diminishes the probability given in Definition 13 for the current state by an order of 2, implying convergence. The rigorous proof is left out due to space constraints.

Algorithm 1 has some additional advantages: Since it is based on the median, it is naturally robust to outliers. In addition, the complexity is linear in the size of the data set and the tree depth. Finally, due to its hierarchical nature, it is possible to extend it when more data is available without rebuilding the entire tree.

9. EXPERIMENTS

Our experiments were done both on simulated data and on real data taken from the KDD cup 1998 [9]. Initially, we evaluate our suggested criteria and the classic MDL based criteria on a simple randomized simulation. Our goal is to examine which criteria find the correct model most distinctively and exhibit correlation between the different criteria.

Next, we test our methods on data from the KDD cup 1998. These data describes donation requests over a time period of 22 months from a given set of individuals. For each person in the mailing list there is some meta-data available such as his age and income level. Over the course of time, the number of mail requests and donations received by each person is documented. The meta-data and personal history for each person can be transformed into a feature vector we use as observations. The action in this case is whether or not to send a donation request and the reward is given by the donation accepted.

9.1 Simulated data

We simulated an MDP with 20 non aliased states with noisy rewards and observations consisting of 7 independent features. The MDPs were generated in the following manner: each transition probability, in the transition matrix for $N = 20$ states, was sampled uniformly over the simplex. The rewards' expectations were generated from the uniform distribution in the interval $[0, 1]$. The observations expectation in each of the 7 dimensions were generated from the uniform distribution in the interval $[0, 20]$. Whenever sampled, states' rewards and observations were added a normally distributed noise with variance 0.05^2 and 1 respectively.

Next, we generated two data trajectories using the simulator. Using Matlab's k-means clustering algorithm [4] 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, RE_0/RE_1 criteria, our aggregation method 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 Figures 3 and 4.

According to these results, we can see that the RE_d works best among the inspected criteria on our simulations. The penalized MDL scores show decent results; However for increasingly more data we can see the weight is tilted towards more refined models. Looking at the value function we can see 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, which means sometimes choosing a smaller, yet incorrect model can lead to better performance. With that in mind, we can see the value function itself can be used as a model selection criterion, perhaps with some additional regularization summand. As for the aggregation criterion, although it was slower, it seems to produce similar results to the RE_d criterion, identifying the correct model starting from trajectories of length 1K.

9.2 KDD Cup 1998 Data

As a test bench, we used the donation data set from the KDD Cup 1998 competition [9] in which the goal was to estimate the return for a direct mailing task. As observations we used the first 8 features given by [15]: the amount in dollars of the last gift, total amount of gifts to date, number of gifts to date, number of promotions to date, their divi-

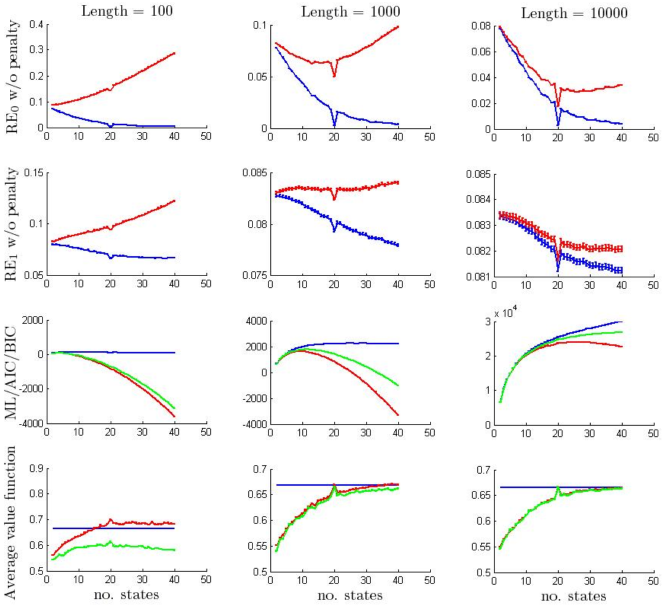

Figure 3: Performance of the different criteria on simulated data. [First two rows] The blue plot is the RE$_d$ score without the regularizing summand. [Third row] The ML (blue), AIC (red) and BIC (green) scores. [Last row] The correct value for the optimal policy (blue), the value for the estimated optimal policy (red) and its real value (green).

sion, number of months since the last gift, age and income bracket. We rescaled the first 3 mentioned features using $f(x) = log(x + 1)$. We then tested the different criteria in a similar manner as 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 requesting donation too often, 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 5.

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.

An analysis of the results is in order. First, it seems that all criteria point towards state spaces with roughly 60 states, which implies a correlation between the different criteria. In addition, as was shown before [15], we see again that dynamic policy distinctively outperform a myopic policy. Another interesting property is the saturation behavior of the estimated value function, while its evaluation on a different portion of the data receives its maximum and decreases significantly afterwards. This phenomenon can be describes as overfitting - the suspected optimal policy is less accurate since the number of samples per state decreases.

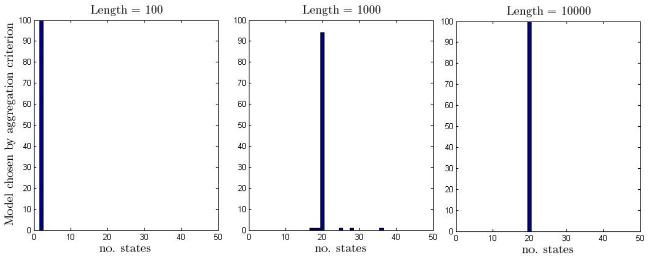

Figure 4: Histogram of the chosen state space size for the aggregation criterion.

In Figure 6 we can see the different scores applied to models constructed using Algorithm 1 with varying number of states. Instead of choosing arbitrarily among the unvisited states and features on each iteration, we visited the states according to the number of observations they encompass, and partitioned according to the feature for which the splitted states are least similar according to our hypothesis testing. We applied the algorithm on observations made from the 17 features given by [15].

We can observe the same saturation phenomena as previously seen using k-means, though now the value itself is higher (at the cost of a longer run-time). This means models constructed by Algorithm 1 are likely to perform better on this data set than models constructed by k-means in terms of accumulated reward. The model order has not changed and it is still around 50 states by all checked criteria, implying this is the true order of the model.

10. 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 [21], 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. The methods themselves are easy to implement and their computational complexity is low.

In our experiments, we had examined different model selection criteria. The RE$_d$ criterion showed results as good as ML based methods. Our aggregation criterion required more computation power, but its theoretical guarantees are better. We extended these ideas to build a weakly consistent decision tree based model construction algorithm that works in manageable complexity. Finally, our methods were

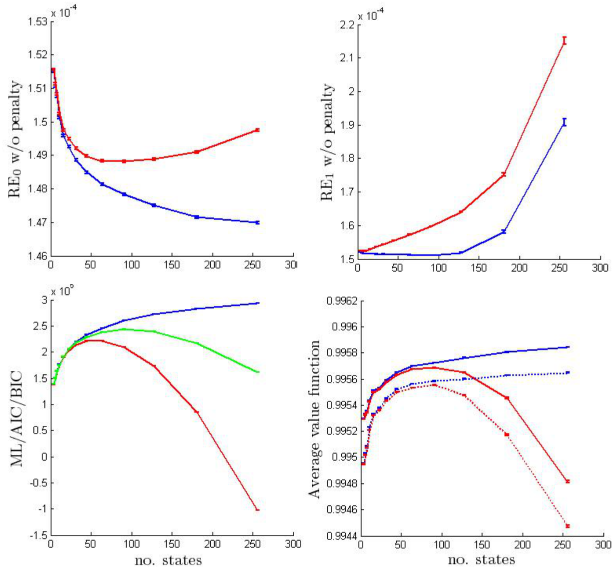

Figure 5: Performance of the different criteria on real data acquired from the KDD cup 1998, where the state space was constructed by k-means clustering. [1st row] The RE$_d$ score with/without the regularizing summand (red/blue). [2nd row, left] The ML (blue), AIC (red) and BIC (green) scores. [2nd row, right] The estimated value optimal policy (blue), estimated value for the greedy policy (dashed blue), and their sampled value on the general population (red and dashed red correspondingly).

used on real world donation data from the KDD cup 1998, yielding promising results.

11. ACKNOWLEDGMENTS

We thank Georgios Theocharous and Duncan Simester for helpful discussions. The research leading to these results has received funding from the European Research Counsel under the European Union's Seventh Framework Program (FP7/2007-2013) / ERC Grant Agreement No 306638 and from the Israel Science Foundation under grant no. 920/12.

References

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

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

[3] Dean, T., K. Kanazawa. 1989. A model for reasoning about persistence and causation. *Computational intelligence* **5**(2) 142–150.

[4] Duda, R.O., P.E. Hart, D.G. Stork. 2001. *Pattern classification*, vol. 2. wiley New York:.

[5] Elliott, R.J., L. Aggoun, J.B. Moore. 1995. *Hidden Markov models: estimation and control*, vol. 29. Springer.

[6] Farahmand, A., C. Szepesvári. 2011. Model selection in reinforcement learning. *Machine Learning* 1–34.

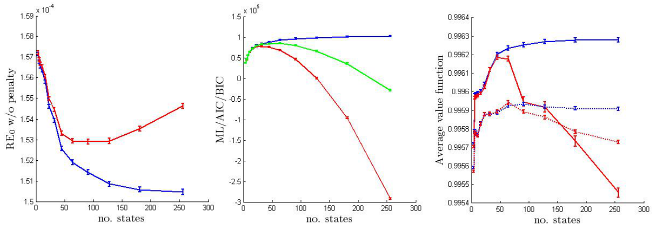

Figure 6: Performance of the different criteria on real data acquired from the KDD cup 1998, state space formed by Algorithm 1. [Left] The RE$_d$ score with/without the regularizing summand (red/blue). [Middle] The ML (blue), AIC (red) and BIC (green) scores. [Right] The estimated value optimal policy (blue), estimated value for the greedy policy (dashed blue), and their sampled value on the general population (red and dashed red correspondingly).

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

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

[9] Hettich, S., S. D. Bay. 1999. The UCI KDD Archive. URL `http://kdd.ics.uci.edu`.

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

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

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

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

[14] McCallum, A.K. 1996. Reinforcement learning with selective perception and hidden state. Ph.D. thesis, University of Rochester.

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

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

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

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

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

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

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