Proceedings of International Joint Conference on Neural Networks, Orlando, Florida, USA, August 12-17, 2007

Optimality of LSTD and its Relation to MC

Steffen Grünewälder, Sepp Hochreiter and Klaus Obermayer

Abstract— **In this analytical study we compare the risk of the Monte Carlo (MC) and the least-squares TD (LSTD) estimator. We prove that for the case of acyclic Markov Reward Processes (MRPs) LSTD has minimal risk for any convex loss function in the class of unbiased estimators. When comparing the Monte Carlo estimator, which does not assume a Markov structure, and LSTD, we find that the Monte Carlo estimator is equivalent to LSTD if both estimators have the same amount of information. Theoretical results are supported by an empirical evaluation of the estimators.**

I. Introduction

One of the important theoretical issues in reinforcement learning are rigorous statements on convergence properties of so called *value estimators* (e.g. [12], [14], [4], [3]) which provide an empirical estimate of the expected future reward for every given state. Most of these convergence results so far were restricted to the asymptotic case rather than providing statements about the deviation of the estimate from the true value for the case of a finite number of observations. In practice, however, one wants to choose the estimator which yields the best result for a given number of examples or in the shortest time.

Current approaches to the finite example case are mostly empirical and few non-empirical approaches exist. [6] present upper bounds on the generalization error for *Temporal Difference estimators (TD)*. They use these bounds to formally verify the intuition that TD methods are subject to a "bias-variance" trade-off and to derive schedules for estimator parameters. Comparisons of different estimators with respect to the bounds were not performed. The issue of *bias and variance* in reinforcement learning is also addressed in other works ([9], [8]). [9] provide analytical expressions of the *mean squared error (MSE)* for various *Monte Carlo (MC)* and TD value estimators. They further provide a software that yields the exact mean squared error curves given a complete description of a *Markov Reward Process (MRP)*. The method can be used to compare different estimators for concrete MRPs and concrete parameter values. But it is not possible to prove general statements with this method. In [8] a MC like estimator was analyzed and a second order approximation of the expectation and the covariance was given, but a comparison between estimators was not performed.

In this paper we follow a new approach to the finite example case using tools from statistical estimation theory (e.g. [11]). Rather than relying on bounds, approximations or on results to be recalculated for every specific MRP this allows us to derive rigorous and more general statements - applicable to at least the class of acyclic MRPs. The most important result of our work is that for acyclic MRPs the *least-squares temporal difference (LSTD)* estimator from [3] has the lowest risk for any convex loss function in the class of unbiased estimators (Section III-A). This is intuitive because LSTD is the estimator which makes optimal use of the underlying Markov structure. We further show that Monte Carlo estimation, which does not use the Markov structure, is equivalent to LSTD, if the Markov structure does not provide additional information (Section III-B).

Symbols are explained at their first occurrence. Proofs are presented in Appendix VI.

Steffen Grünewälder and Klaus Obermayer are with the Department of Electrical Engineering and Computer Science, University of Technology Berlin, GER, Sepp Hochreiter is with the Department of Information Systems, JKU Linz, AUS (email: gruenew@cs.tu-berlin.de).

II. Estimation in Reinforcement Learning

Reinforcement learning methods typically consist of a value estimation and a policy update step (value/policy iteration, [13]). A common assumption underlying the value estimation is that the environment can be described by a Markov Decision Process (MDP). This assumption allows us to improve estimation performance beyond these of general estimators like the sample mean (Monte Carlo) estimator.

In our work, we focus on systems which are modeled as Markov Reward Processes (MRP). The difference of a MRP to a MDP is that only one action exists. Therefore, there is only one policy and "learning" is restricted to the estimation of the value function. Results obtained for the different estimators, however, apply to general MDPs, as long as the policy remains the same (e.g. no online update). The reason for this is that it is possible to account for the policy through the transition distribution. For example, if we got a deterministic policy then the transition probabilities of the MRP at a state are given by the transition probabilities of the MDP corresponding to that state and to the action choosen by the policy at that state.

A. Markov Reward Processes and Value Estimators

A Markov Reward Process (MRP) consists of a state space $\mathbb{S}$ (in the following we will consider a finite state space), starting probabilities p_i for the initial states, transition probabilities $p_{i,j}$ between states s_i and s_j, and a reward function $r : \mathbb{S} \times \mathbb{S} \to \mathbb{R}^{\Omega}$, which maps a state transition to a real valued random variable. The random variables $r(s, s')$ can have an arbitrary distribution, for example Gaussian, binomial or simply be deterministic. We assume that $r(s, s')$ has finite expectation and variance.

Our goal is to estimate the value V_s of each state s, i.e. the expected future reward received after visiting the state.

This value function is given by

$$V_s = \sum_{s' \in \mathbb{S}} p_{ss'} \left(\gamma V_{s'} + \mathbb{E}[r(s,s')] \right) = \sum_{i=1}^{\infty} \gamma^i \mathbf{P}^i \mathbf{r},$$

where γ is a discount factor, $\mathbf{r}$ is the vector of the expected reward ($\mathbf{r}_i = \sum_{s' \in \mathbb{S}} \mathbb{E}[r(s_i, s')]$), and $\mathbf{P}$ the transition matrix of the Markov process.

We compare different value estimators with respect to their risk (not the empirical risk)

$$\mathbb{E}[l(\bar{V}_s, V_s)],$$

where $\bar{V}_s$ is the estimator of the value of state s and $l(\bar{V}_s, V_s)$ the loss function, which penalizes the deviation of the estimator from the true value V_s. We will mainly use the mean squared error

$$\mathbb{E}[(\bar{V}_s - V_s)^2], \tag{1}$$

which can be split into a *bias* and a *variance* term

$$mse(\bar{V}_s) = \underbrace{\mathbb{V}[\bar{V}_s]}_{\text{Variance}} + \underbrace{(\mathbb{E}[\bar{V}_s - V_s])^2}_{\text{Bias}}.$$

The MSE used here corresponds to the value obtained by averaging the empirical mean-squared-error values from an infinite number of learning tasks on a given problem. An estimator is called *unbiased* if the bias term is zero.

B. Monte Carlo Estimation

The Monte Carlo estimator is the sample mean estimator of the future reward. In [13] it is defined as $1/n \sum_{i=1}^{n} \text{Returns}(i)$, where n is the number of trajectories for a given MRP and Returns(i) the cumulated future reward for a given trajectory i. The Monte-Carlo estimator can be interpreted as a special case of $TD(\lambda)$ with $\lambda = 1$ and $\alpha_i = 1/i$. The estimator is unbiased [10], converges almost sure and in the average to the correct value.

III. Comparison of Estimators: Theory

The structure of MRPs, given by the transition matrix $\mathbf{P}$, introduces dependencies between the values of different states. These dependencies can be used to improve estimation performance. The linear least squares temporal difference estimator (LSTD) [3] optimally utilizes "structure" at the expanse of a high numerical cost ($O(|\mathbb{S}|^3)$, compared to MC with a cost of $O(|\mathbb{S}|)$). We prove that firstly, LSTD is the optimal unbiased estimator for acyclic MRPs for any convex loss (Section III-A) and secondly, that Monte Carlo estimation is equivalent to LSTD, if the Markov structure provides no further information (Section III-B).

A. Linear Least-Squares Temporal Difference Learning

The LSTD estimator was first introduced by [3] and extensively analyzed in [1] and [2]. Empirical studies showed that LSTD often outperforms massively the TD and the Monte Carlo estimator with respect to convergence speed per sample size. An analytical statement for the higher convergence speed of LSTD, however, is missing. Here, we prove that - for acyclic MRPs and for any convex loss function $l(\bar{V}_s, V_s)$ - the LSTD estimator has the minimal risk of all unbiased estimators. We derive the optimality not directly for LSTD, but for a maximum likelihood estimator which is equivalent to LSTD (for equivalence see Section III-A.5).

1) Maximum Likelihood: Let $\theta_{i,j}$ be the transition probability of s_i to s_j,ϑ_i the probability to start in s_i and x a sample consisting of n iid state sequences $x_1, \ldots, x_n$. The log-likelihood of the sample is given by

$$\log \mathbb{P}[x|\theta, \vartheta] \overset{\text{ind}}{=} \log \prod_{k=1}^{n} \mathbb{P}[x_k|\theta, \vartheta].$$

The corresponding maximization problem is given by

$$\max_{\theta, \vartheta} \log \prod_{i=1}^{n} \mathbb{P}[x_i|\theta, \vartheta], \qquad \text{s.t.:} \sum_{j=1}^{\sharp\mathbb{S}} \theta_{ij} = \sum_{j=1}^{\sharp\mathbb{S}} \vartheta_j = 1.$$

The unique solution for θ and ϑ (Lagrange multipliers) is given by

$$\theta_{ij} = \frac{\mu_{s_i,s_j}}{K_{s_i}} =: \bar{p}_{s_i s_j} \quad \text{and} \quad \vartheta_i = K_{s_i} - \sum_{s' \in \mathbb{S}} \mu_{s_i,s'} =: \bar{p}_{s_i}, \tag{2}$$

where K_s denotes the number of visits of state s, $\mu_{s,s'}$ the number of direct transitions from s to s', $\bar{p}_{ss'}$ the estimate of the true transition probability $p_{ss'}$ and $\bar{p}_s$ the estimate of the true starting probability p_s. Using the Markov structure it is possible to calculate state values in a manner similar to dynamic programming, where the true probabilities are replaced by maximum likelihood parameter estimates. We start by defining an estimator for $\bar{P}_\pi$ for path probabilities,

$$\bar{P}_\pi := \prod_i \bar{p}_{\pi_{i-1}\pi_i}, \tag{3}$$

with π being a path and π_i the ith state in the path. An estimator for the probability of reaching state s' from state s (through different state sequences) is given by

$$\bar{P}_{ss'} := \sum_{\pi \in \Pi_{ss'}} \bar{P}_\pi, \tag{4}$$

where $\Pi_{ss'}$ is the set of paths from s to s'. In general, we have no probability model for the reward of a state transition, hence maximum likelihood is not applicable. As a natural alternative we use the sample mean estimator,

$$\bar{R}_s := \frac{H_s}{K_s}, \tag{5}$$

where H_s denotes the summed reward of state transitions from state s. The maximum likelihood value estimator is then given by

$$\bar{V}_s := \bar{R}_s + \sum_{s' \in \mathbb{S}} \bar{P}_{ss'} \bar{R}_{s'}. \tag{6}$$

In this section we define an estimator which is based on a maximum likelihood estimate of the MRP parameters. The approach results in the same estimator as a maximum likelihood estimator that is briefly sketched in [12].Our definition is especially useful to proof the theorems 3.1, 3.4 and to evaluate the relation of LSTD.

2) Sufficient Statistics for the MRP Parameters: Information about a sample is typically available through a *statistic* $\mathcal{S}$ of the data (for example $\mathcal{S} = \sum_i x_i$, where x is a sample). A statistic which contains all information about a sample is called *sufficient*. Important properties of sufficient statistics are *minimality* and *completeness*. The minimal sufficient statistics is the sufficient statistic with the smallest dimension (typically the same dimension as the parameter space). Formally, suppose that a statistic $\mathcal{S}$ is sufficient for a parameter θ. Then $\mathcal{S}$ is minimally sufficient if $\mathcal{S}$ is a function of any other statistic $\mathcal{T}$ that is sufficient for θ. Formally, a statistic $\mathcal{S}$ is complete if $\mathbb{E}_\theta[h(\mathcal{S})] = 0$ for all θ implies $h = 0$ almost sure. The theorem from *Rao and Blackwell* [11] states that for a complete and minimal sufficient statistics $\mathcal{S}$ and any unbiased estimator A of a parameter θ the estimator $\mathbb{E}[A|\mathcal{S}]$ is the optimal unbiased estimator with respect to any convex loss function and hence the unbiased estimator with minimal MSE.

The maximum likelihood solution is a *sufficient statistics* for the MRP parameters. We demonstrate this with the help of the *Fisher-Neyman factorization theorem* [11]. It states that a statistic is sufficient if and only if the density $f(\mathbf{x}|\theta)$ can be factored into a product $g(\mathcal{T}, \theta)h(\mathbf{x})$. For an MRP we can factor the density as needed by the Fisher-Neyman theorem ($h(\mathbf{x}) = 1$ in our case),

$$\begin{aligned}\mathbb{P}(\mathbf{x}|\theta, \vartheta) &= \prod_{i=1}^{n}\Big(\vartheta_{\phi_\mathbf{x}(i,1)} \prod_{j=2}^{L_i} \theta_{\phi_\mathbf{x}(i,j-1)\phi_\mathbf{x}(i,j)}\Big) \\ &= \prod_{i=1}^{n}\Big(\Big(\sum_{j=1}^{|\mathbb{S}|} \vartheta_{s_j}^{\delta(\mathbf{x}_{1,i},j)}\Big) \prod_{j=2}^{N}\Big(\sum_{k,l}^{|\mathbb{S}|} \theta_{s_k s_l}^{\delta(\mathbf{x}_{i,j-2},k)\delta(\mathbf{x}_{i,j},l)}\Big)\Big) \\ &= \prod_{s\in\mathbb{S}} \vartheta_s^{(K_s - \sum_{s'} \mu_{ss'})} \prod_{s,s'\in\mathbb{S}} \theta_{ss'}^{K_s \mu_{ss'}},\end{aligned}$$

where ϕ is an enumeration of the visited states in the trajectories ($\phi(i, j)$ is the jth state in the ith trajectory), δ is the Dirac delta function, n the number of trajectories and L_i the length of the ith trajectory. $K_s\mu_{ss'}$ is sufficient for $\theta_{ss'}$ and because sufficiency is sustained by one-to-one mappings [11] this holds true also for $\mu_{ss'}$. The sufficient statistics is *minimal* because the maximum likelihood solution is unique [11].The sufficient statistic is also *complete* because the sample distribution induced by an acyclic MRP forms an exponential family of distributions (Lemma 6.1, Appendix VI). Due to [7] any *exponential family* of distributions is complete.

3) Optimality: The *Rao-Blackwell theorem* [11] states that for any unbiased estimator A the estimator $\mathbb{E}[A|\mathcal{S}]$ is the optimal unbiased estimator, given $\mathcal{S}$ is a minimal and complete sufficient statistic. For the case of value estimation this means that we can use any unbiased value estimator (e.g. the Monte Carlo estimator) and condition it with the statistic induced by the maximum likelihood parameter estimate to get the optimal unbiased value estimator.

If an estimator is a function of the sufficient statistic (e.g. $A = f(\mathcal{S})$) then the conditional estimator is equal to the original estimator, $A = \mathbb{E}[A|\mathcal{S}]$. If the estimator A is further unbiased then it is due to the Rao-Blackwell theorem the optimal unbiased estimator. The defined maximum likelihood estimator is a function of a minimal and complete sufficient statistic. It is further unbiased and therefore the optimal unbiased estimator.

Theorem 3.1 (Unbiased): Given an acyclic MRP with finite state space and n iid sequences, the maximum likelihood estimator is unbiased, i.e. for $n > 0$

$$\mathbb{E}[\bar{V}_s|K_s = n] = \mathbb{E}[r_s] + \sum_{s'\in\mathbb{S}} P_{ss'} \cdot \mathbb{E}[r_{s'}].$$

Corollary 3.1 (Optimality): The maximum likelihood estimator is optimal with respect to any convex loss function in the class of unbiased estimators, especially for any unbiased estimator $\bar{\bar{V}}_s$ of the state value it holds that

$$MSE[\bar{V}_s] \leq MSE[\bar{\bar{V}}_s]$$

4) The LSTD Estimator: The LSTD algorithm computes analytically the parameters which minimize the empirical quadratic error for the case of a linear system. [3] show that the resulting algorithm converges almost sure to the true solution. In [1] a further characterization of the least-square solution is given. This turns out to be very useful to establish the relation to the maximum likelihood estimator. According to this characterization, LSTD finds the solution for which the equation

$$\bar{V}_s = \frac{H_s}{K_s} + \sum_{s'\in\mathbb{S}} \frac{\mu_{s,s'}}{K_s} \bar{V}_{s'} \tag{7}$$

holds.

5) Equivalence between the Maximum Likelihood Estimator and LSTD: The maximum likelihood estimator defined through eq. (2) to (6) is equivalent to LSTD.

Theorem 3.2: Given a MRP with a finite state space $\mathbb{S}$, $s, s' \in \mathbb{S}$ and iid sequences, the following equality holds for the maximum likelihood estimator:

$$\bar{V}_s = \bar{R}_s + \sum_{t\in\mathbb{S}} \bar{p}_{st} \bar{V}_t. \tag{8}$$

This theorem follows from Lemma 6.2 (Appendix VI-B). Notice the similarity to the classical consistency condition of value functions [13]: $V_s = R_s + \sum_{s'} P_{ss'} V_{s'}$ (Here, V_s denotes the true value). The equivalence to LSTD (eq. 7) becomes apparent by substituting the definitions of $\bar{p}_{st}$ and $\bar{R}_s$:

Corollary 3.2: The maximum likelihood estimator defined through equations (2) to (6) is equivalent to LSTD.

B. Monte Carlo Estimation

The estimation approaches of LSTD and Monte Carlo are at the first glance quite different. LSTD makes massively use of the underlying MRP structure to propagate information from successor states whereas the Monte Carlo estimator uses only the sequences which visit the state of interest. The increased amount of information is actually the only, and also the major advantage of structure using estimators like LSTD. Both estimators are equivalent if the following criterion is fulfilled for the corresponding state.

Criterion 3.1 (Full Information): We say that a state s has full information if every path to the successors of s includes s itself and if the starting probability for the successors is zero.

We call the criterion the *full information* criterion because all information-containing trajectories must hit state s. To proof the equivalence we first transform the Monte Carlo estimator into a form suitable for comparison with LSTD.

Theorem 3.3 (MC Reformulation): Given an acyclic MRP with a finite state space $\mathbb{S}$ and a sample x of n iid sequences, then the Monte Carlo estimator is equal to $\bar{\bar{V}}_s$, where $\bar{\bar{V}}_s$ is defined as

$$\bar{\bar{V}}_s := \sum_{s' \in \mathbb{S}} \bar{\bar{P}}_{ss'} \bar{\bar{R}}_{s'|s},$$

$$\bar{\bar{P}}_{ss'} := \frac{K_{s'|s}(x)}{K_s}, \qquad \bar{\bar{R}}_{s'|s} := \frac{H_{s'|s}}{K_{s'|s}},$$

where $K_{s'|s}$ denotes the number of visits to state s', which followed visits of state s (conditional). Similarly, $H_{s'|s}$ denotes the sum of the direct reward for which only examples are used which visited s. Hence $\bar{\bar{P}}_{ss'}$ is the sample mean estimator of the transition probability from s to s' and $\bar{\bar{R}}_{s'|s}$ the sample mean estimator of the reward in s', where for estimation only samples are used which contain state s.

In this notation the Monte Carlo estimator looks already very similar to the maximum likelihood estimator (eq. 6). The difference lies in the transfer and reward estimators ($\bar{\bar{P}}_{ss'}$, $\bar{\bar{R}}_{s'|s}$). For the special case that the Monte Carlo estimator considers all trajectories, these estimators are equivalent to the ones of LSTD and Monte Carlo estimation is optimal.

Theorem 3.4 (Equality Theorem): Given an acyclic MRP with finite state space $\mathbb{S}$, if the full information criterion (3.1) is fulfilled for state s it holds that

$$\bar{P}_{ss'}(x) = \bar{\bar{P}}_{ss'}(x).$$

Finally, we get the optimality of Monte Carlo estimation for this case because the estimator $\bar{\bar{R}}_{s'|s}(x)$ is equal to $\bar{R}_{s'}(x)$. The estimators are equal because every path to s' contains s.

Corollary 3.3: With s being like in theorem 3.4 and with the assumptions of that theorem

$$\bar{V}_s(x) = \bar{\bar{V}}_s(x),$$

and the Monte Carlo estimator in state s is the estimator with minimal risk for any convex loss function in the class of unbiased estimators.

IV. Simulations

We performed three experiments for analyzing the estimators. In the first experiment we measured the MSE in dependence on the number of trajectories. In the second experiment we analyzed how the MRP structure effects the estimation performance. As we can see from the equality theorem (Theorem 3.4) the difference of the performance between LSTD and MC depends strongly on the ratio between the number of sequences hitting state s itself and the number of sequences entering the subgraph of successor states without hitting s. We varied this ratio in the second experiment and measured the MSE.

a) Basic Experimental Setup: We generated randomly acyclic MRPs for the experiments. The generation process was the following: First, we defined a state s for which we want to estimate the value. Then we generated randomly a graph of successor states. We used different layers with a random number of states in each layer. Connections were only possible between adjacent layers. Given these constraints, the transition matrix was generated randomly (uniform distribution). For the different experiments, a specific number of starts in state s was defined. Beside that, a number of starts in other states were defined. Starting states are all states in the first layers (typically the first 4). Other layers which are further apart from s were omitted as trajectories starting in these contribute few to the estimate, but consume computation time. The distribution over the starting states was chosen to be uniform. Finally, we defined randomly a reward for the different transitions (between 0 and 1), while a small percentage (1 to 5 percent) got a high reward (reward 1000). Beside the reward definition, this class of MRPs contains a wide range of acyclic MRPs. We tested the performance (empirical MSE) of the LSTD and MC estimator. The simulations were repeated 10000 times.

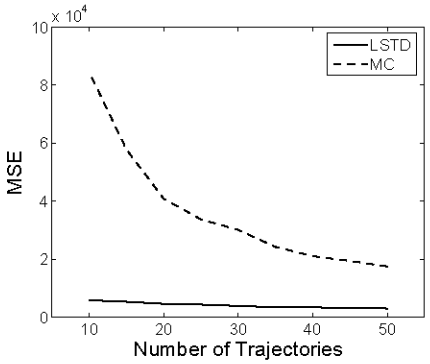

Fig. 1. MSE of LSTD and MC in relation to the number of trajectories. The state space consisted of 10 layers with 20 states per layer.

1) Experiment 1: MSE in Relation to the Number of Trajectories: In the first experiment, we analyzed the effect of the number of trajectories given a fixed rate of 0.2 for starts in state s. The starting probability for state s is high and beneficial to MC (The effect of $P(s)$ is analyzed in the second experiment). LSTD is even for few trajectories strongly superior and already produces a good estimate for 10 trajectories. Due to the scale the improvement of LSTD is hard to observe.

2) Experiment 2: MSE in Relation to the Starting Probability: In the second experiment we tested how strong the different estimators use the Markov structure. To do so, we varied the ratio of starts in state s (the estimator state) to starts in the subgraph. The trajectories which start in the subgraph can only improve the estimation quality of state s if the Markov structure is used. Figure 2 shows the results of

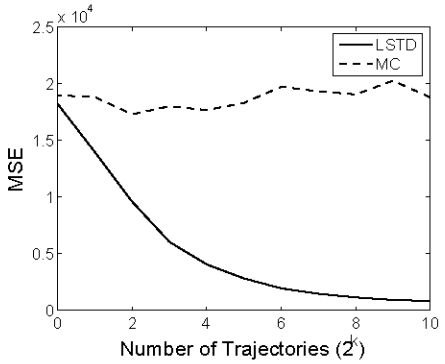

Fig. 2. MSE of LSTD and MC in relation to the starting probability of the estimated state. State space: 10 layers with 20 states per layer.

the simulations. The x-axis gives the number of starts in the subgraph while the number of starts in state s was set to 10. We increased the number exponentially, while on the x-axis the exponential factor is printed. x=0 is equivalent to always start in s. One can see that the MC and LSTD estimator are equivalent if in each run the trajectory starts in s ($k = 0$).

V. Summary

In this work, we explored the relation of estimation based on the MRP structure (model based estimation) to model free estimation with the help of statistical estimation theory. We proved that the optimal unbiased estimator, with respect to any convex loss and for acyclic MRPs, is a model based estimator (the LSTD estimator). The core ingredients of the proof are that LSTD is unbiased in the acyclic case and that it is a function of a complete and minimal sufficient statistics of the MRP parameters. Monte Carlo estimation is unbiased and can therefore have at best the same risk as LSTD[1]. Interestingly, in the special case where the Monte Carlo estimator "observes" all relevant trajectories it is equal to LSTD for acyclic MRPs. This can be interpreted in the sense that it is not the way the estimate is computed but the amount of information that makes model based estimators superior.

Appendix

VI. Proofs

A. Markov Reward Process

Lemma 6.1: An acyclic MRP with finite state space and iid sequences forms an s-dimensional exponential family, where s is the number of free MRP parameters.

Proof: A family $\{P_\theta\}$ of distributions is said to form an s-dimensional exponential family if the distributions P_θ have densities of the form

$$p_\theta(x) = \exp\left(\sum_{i=1}^{s} \eta_i(\theta) T_i(x) - A(\theta)\right) h(x) \qquad (9)$$

with respect to some common measure μ [7]. Here, the η_i and A are real-valued functions of the parameters and the T_i are real-valued statistics, and x is a point in the sample space. The η's are called *natural parameters*. It is important that the natural parameters are not functionally related, for example, that no f exists with $\eta_2 = f(\eta_1)$. Otherwise, the family forms only a *curved exponential family* [7]. Firstly, we demonstrate that the transition distribution forms an exponential family. The density can be written as

$$\begin{aligned}\mathbb{P}(X_1 = \pi_{i_1}, \ldots, X_n = \pi_{i_n}) &= m(i_1, \ldots, i_n) P_{\pi_0}^{c_0} \ldots P_{\pi_L}^{c_L} \\ &= m(i_1, \ldots, i_n) \exp\Big(c_0 \log P_{\pi_0} + \ldots + c_L \log P_{\pi_L}\Big),\end{aligned}$$

with π being the observed paths, m an input dependent function (for example multinomial), L the number of paths in the MRP, c_i the number of times path i has occurred and P the probability of the path. The parameters P_π are redundant with this representation. We explore now the MRP structure to find natural parameters that are not functionally dependent. The size of this set of parameters is the number of necessary MRP parameters, that is

$$\sharp\text{Starting States} - 1 + \sum_{s \in \mathbb{S}} \big(\sharp\text{Direct Successors of } s - 1\big).$$

We have per state as many η's as outgoing connections (-1 if not a starting state, respectively for the first starting state). We reformulate the exponential expression to reduce the number of parameters. We first define the expression for one specific starting state that has no predecessors. For this state the following expression is used:

$$n \log(\vartheta_1 p_{10_{(1)}}) + \sum_{i_{(1)} > 0_{(1)}} \mu_{1 i_{(1)}} \log \frac{(\vartheta_1 p_{1 i_{(1)}})}{(\vartheta_1 p_{10_{(1)}})}, \qquad (10)$$

where n is the total number of runs, ϑ the starting probability, p the transition probability and $i_{(1)}$ an enumeration of the direct successors of state 1 ($0_{(1)}$ is the first successor state with respect to the enumeration). The term $A(\theta)$ of the exponential family is $-n \log(\vartheta_1 p_{10_{(1)}})$ and the first η's are the $\log \frac{(\vartheta_1 p_{1 i_{(1)}})}{(\vartheta_1 p_{10_{(1)}})}$ terms. Notice that the parameter $p_{10_{(1)}}$ has a coefficient of $n - \sum \mu_{1 i_{(1)}}$ and ϑ_1 a coefficient of n. In the end the coefficient for $p_{10_{(1)}}$ needs to be $n_1 - \sum \mu_{1 i_{(1)}}$ and for ϑ_1 it must be n_1, where n_1 is the number of starts in state 1. Further, $n_1 = n - \sum_{k>1} n_k$, where k enumerates all starting states. This leads us directly to the following term which must be included for every other starting state

$$n_i \log \frac{(\vartheta_i p_{i 0_{(i)}})}{(\vartheta_1 p_{10_{(1)}})}. \qquad (11)$$

We now have the problem that the number of visits of a state depends on the taken paths (the data). This is at first viewing problematic as the straight forward approach

$$n_j \log(\vartheta_j p_{j 0_{(j)}}) + \sum_{i_{(j)} > 0_{(j)}} \mu_{j i_{(j)}} \log \frac{(\vartheta_j p_{j i_{(j)}})}{(\vartheta_j p_{j 0_{(j)}})}, \qquad (12)$$

introduces one η too much, as n_j is data dependent. The solution to the problem is that n_j equals the μ's of the incoming connections, $n_j = \sum_l \mu_{lj}$ where l enumerates the

[1]Demonstrated only for acyclic MRPs.

direct predecessors. Hence, we can remove the n_j term by modifying the incoming terms, as follows

$$\mu_{lj} \log \frac{(\vartheta_l p_{lj})}{(\vartheta_l p_{l0_{(l)}})} \rightarrow \mu_{lj} \log \frac{(\vartheta_l p_{lj})(\vartheta_j p_{j0_{(j)}})}{(\vartheta_l p_{l0_{(l)}})}. \quad (13)$$

The exponential term is defined by the terms (10), (11), (12) and by the modification (13). ■

B. Maximum Likelihood Approach

Theorem 3.1 *Proof:* The sample mean estimators $\bar{p}$ and $\bar{R}$ are unbiased [5]. The main problem is to show that $\bar{P}_{ss'}$ is unbiased (eq. (4) and (6)). For this we start with

$$\mathbb{E}[\bar{P}_{ss'}|K_s = n] = \sum_{\pi \in \Pi_{ss'}} \mathbb{E}\Big[\prod_i \bar{p}_{\pi_{i-1}\pi_i} \Big| K_s = n\Big]$$

The last of these estimators (denote it with $p_{\hat{s}\hat{s}}$) is conditionally independent of the others given the number of visits of state $\hat{s}$ ($K_{\hat{s}}$). This is also the main point where acyclicity is needed. Using this together with the law of total probability and the fact that $\bar{p}$ is unbiased, leads to the following statement (with L being the length of the path π):

$$\mathbb{E}\Big[\prod_i^L \bar{p}_{\pi_{i-1}\pi_i} \Big| K_s = n\Big] =$$

$$\sum_{l=1}^n \mathbb{E}\Big[\prod_i^L \bar{p}_{\pi_{i-1}\pi_i} \Big| K_s = n, K_{\hat{s}} = l\Big] \mathbb{P}[K_{\hat{s}} = l | K_s = n] \overset{ind}{=}$$

$$\sum_{l=1}^n \mathbb{E}\Big[\prod_i^{L-1} \bar{p}_{\pi_{i-1}\pi_i} \Big| K_s = n, K_{\hat{s}} = l\Big] p_{\hat{s}\hat{s}} \mathbb{P}[K_{\hat{s}} = l | K_s = n] =$$

$$p_{\hat{s}\hat{s}} \mathbb{E}\Big[\prod_i^{L-1} \bar{p}_{\pi_{i-1}\pi_i} \Big| K_s = n\Big] \quad (14)$$

We used that for $l = 0$ the last estimator $\bar{p}$ in the product is zero. The procedure has to be repeated for every $\bar{p}$. As a result the expectation of this estimator is equal to the path probability. $\bar{R}$ can be handled similarly ■

Lemma 6.2: Given a MRP with a finite state space $\mathbb{S}$, $s, s' \in \mathbb{S}$ and iid sequences.

$$1.\ \bar{P}_{ss'} = \sum_{t \in \mathbb{S}} \bar{p}_{st} \cdot \bar{P}_{ts'} + \bar{p}_{ss'} \qquad 2.\ \bar{P}_{ss'} = \sum_{t \in \mathbb{S}} \bar{P}_{st} \cdot \bar{p}_{ts'} + \bar{p}_{ss'}$$

Proof: We present only the derivation of equation 1. $\bar{P}_{ss'} = \sum_{\pi \in \Pi_{ss'}} \bar{P}_\pi$. Set $A_t := \{\pi | \pi = (s, t, ..., s'), (t, ..., s') \in \Pi_{ts'}\}$. Thus $\sum_{\pi \in A_t} \bar{P}_\pi = \bar{p}_{st} \bar{P}_{ts'}$.

$$\Pi_{ss'} = \Big(\dot{\bigcup_{t \in \mathbb{S}}} A_t\Big) \dot{\cup} \{(s, s') | \text{if } s' \text{ is a direct successor of } s\},$$

with $\dot{\cup}$ denoting disjoint union. Therefore $\bar{P}_{ss'} = \bar{p}_{ss'} + \sum_{t \in \mathbb{S}} \sum_{\pi \in A_t} \bar{P}_\pi = \bar{p}_{ss'} + \sum_{t \in \mathbb{S}} \bar{p}_{st} \bar{P}_t$. ■

C. Monte-Carlo Estimation

Theorem 3.3 *Proof:*

$$MC_s = \sum_{s' \in \mathbb{S}} \frac{1}{K_s} \sum_{i=1}^{K_s} \psi^i_{s'|s} = \frac{1}{K_s} \sum_{i=1}^{K_s} \overbrace{\Big(\sum_{s' \in SUCC(s)} \psi^i_{s'|s}\Big)}^{returns(s,i)},$$

where $\psi^i_{s'|s}$ equals the reward received through the transition from state s' in run i if the path includes state s or is 0. ■

Theorem 3.4 *Proof:* We make an inductive proof. To be able to do so we first enumerate the states of $SUCC(s) \cup \{s\}$ through $\phi : \mathbb{N} \rightarrow \mathbb{S}$ with $i < j \Rightarrow \phi(i) \notin SUCC(\phi(j))$. Here, $SUCC(s)$ denotes the set of successor states of s. This way $\phi(1)$ is state s itself.

Induction Step (n-1 → n): Denote the direct predecessors of $\phi(n)$ with $a(1), ..., a(k)$ and let $z_{\tilde{s}} := \mu_{\tilde{s}\phi(n)}$. For s being no direct predecessor of $\phi(n)$ ($\Rightarrow \bar{p}_{s\phi(n)} = 0$, other case is similar):

$$\begin{aligned}
\bar{\bar{P}}_{s\phi(n)} &= \frac{K_{\phi(n)}}{K_{\phi(1)}} = \frac{z_{a(1)}}{K_{\phi(1)}} + ... + \frac{z_{a(k)}}{K_{\phi(1)}} \\
&= \frac{z_{a(1)}}{K_{a(1)}} \frac{K_{a(1)}}{K_{\phi(1)}} + ... + \frac{z_{a(k)}}{K_{a(k)}} \frac{K_{a(k)}}{K_{\phi(1)}} \\
&= \frac{z_{a(1)}}{K_{a(1)}} \bar{\bar{P}}_{sa(1)} + ... + \frac{z_{a(k)}}{K_{a(k)}} \bar{\bar{P}}_{sa(k)} \\
&\overset{\text{I.H.}}{=} \frac{z_{a(1)}}{K_{a(1)}} \bar{P}_{sa(1)} + ... + \frac{z_{a(k)}}{K_{a(k)}} \bar{P}_{sa(k)} \\
&= \bar{P}_{sa(1)} \cdot \bar{p}_{a(1)\phi(n)} + ... + \bar{P}_{sa(k)} \cdot \bar{p}_{a(k)\phi(n)} \\
&\overset{\text{Lem 6.2}}{=} \bar{P}_{s\phi(n)},
\end{aligned}$$

while $\frac{K_{\tilde{s}}}{K_s} = \bar{\bar{P}}_{s\tilde{s}}$ holds, because the full information criterion applies to state s. ■

References

[1] J. Boyan. *Learning Evaluation Functions for Global Optimization*. PhD thesis, School of Computer Science Carnegie Mellon University, 1998.

[2] J. Boyan. Least-squares temporal difference learning. In *International Conference Machine Learning*, 1999.

[3] S. J. Bradtke and A. G. Barto. Linear least-squares algorithms for temporal difference learning. *Machine Learning*, 22(1/2/3):33–57, 1996.

[4] Tommi Jaakkola, Michael I. Jordan, and Satinder P. Singh. On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation*, 1994.

[5] Steven M. Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall International, Inc., 1993.

[6] M. Kearns and S. Singh. Bias-variance error bounds for temporal difference updates. In *Conference on Computational Learning Theory*, 2000.

[7] Erich Leo Lehmann and George Casella. *Theory of Point Estimation*. Springer Texts in Statistics, 1998.

[8] Shie Mannor, Duncan Simester, Peng Sun, and John N. Tsitsiklis. Bias and variance in value function estimation. In *International Conference Machine Learning*, 2004.

[9] Satinder Singh and Peter Dayan. Analytical mean squared error curves for temporal difference learning. *Machine Learning*, 32, 1998.

[10] Satinder Singh and Richard Sutton. Reinforcement learning with replacing eligibility traces. *Machine Learning*, 22(1), 1996.

[11] Stuart and Ord. *Kendall's Advanced Theory of Statistics*. Edward Arnold, fifth edition, 1991.

[12] Richard S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3:9–44, 1988.

[13] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

[14] Chris Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8, 1992.