Mach Learn (2011) 83: 289–330
DOI 10.1007/s10994-010-5220-9

The optimal unbiased value estimator and its relation to LSTD, TD and MC

Steffen Grünewälder · Klaus Obermayer

Received: 13 July 2009 / Revised: 20 June 2010 / Accepted: 29 September 2010 /
Published online: 29 October 2010

Abstract In this analytical study we derive the optimal unbiased value estimator (MVU) and compare its statistical risk to three well known value estimators: Temporal Difference learning (TD), Monte Carlo estimation (MC) and Least-Squares Temporal Difference Learning (LSTD). We demonstrate that LSTD is equivalent to the MVU if the Markov Reward Process (MRP) is acyclic and show that both differ for most cyclic MRPs as LSTD is then typically biased. More generally, we show that estimators that fulfill the Bellman equation can only be unbiased for special cyclic MRPs. The reason for this is that at each state the bias is calculated with a different probability measure and due to the strong coupling by the Bellman equation it is typically not possible for a set of value estimators to be unbiased with respect to each of these measures. Furthermore, we derive relations of the MVU to MC and TD. The most important of these relations is the equivalence of MC to the MVU and to LSTD for undiscounted MRPs in which MC has the *same amount of information*. In the discounted case this equivalence does not hold anymore. For TD we show that it is essentially unbiased for acyclic MRPs and biased for cyclic MRPs. We also order estimators according to their risk and present counter-examples to show that no general ordering exists between the MVU and LSTD, between MC and LSTD and between TD and MC. Theoretical results are supported by examples and an empirical evaluation.

Keywords Optimal unbiased value estimator · Maximum likelihood value estimator · Sufficient statistics · Lehmann-Scheffe theorem

Editor: P. Tadepalli.

S. Grünewälder · K. Obermayer
Department of Computer Science, Berlin University of Technology, Berlin 10587, Germany

S. Grünewälder (✉)
Centre for Computational Statistics and Machine Learning, University College London, Gower Street, London WC1E 6BT, UK
e-mail: steffen@cs.ucl.ac.uk

1 Introduction

One of the important theoretical issues in reinforcement learning are rigorous statements on convergence properties of so called *value estimators* (e.g. Sutton 1988; Watkins and Dayan 1992; Jaakkola et al. 1994; Bradtke and Barto 1996) which provide an empirical estimate of the expected future reward for every given state. So far most of these convergence results were restricted to the asymptotic case and did not provide statements 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. Kearns and Singh (2000) 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 Singh and Dayan (1998), Mannor et al. (2007). Singh and Dayan (1998) provide analytical expressions of the *mean squared error* (*MSE*) for various *Monte Carlo* (*MC*) and TD estimators. Furthermore, they provide 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. But it is not possible to prove general statements with their method. The most relevant works for our analysis are provided by Mannor et al. (2007) and by Singh and Sutton (1996).

In Mannor et al. (2007) the bias and the variance in value function estimates is studied and closed-form approximations are provided for these terms. The approximations are used in a large sample approach to derive asymptotic confidence intervals. The underlying assumption of normally distributed estimates is tested empirically on a dataset of a mail-order catalogue. In particular, a Kolmogorov-Smirnov test was unable to reject the hypothesis of normal distribution with a confidence of 0.05. The value function estimates are based on sample mean estimates of the MRP parameters. The parameter estimates are used in combination with the value equation to produce the value estimate. Different assumptions are made in the paper to simplify the analysis. A particularly important assumption is that they assume that the number of visits of a state is not random but fixed and that each state is visited at least once (Mannor et al. 2007, p. 310, top of right column). Under this assumption the sample mean parameter estimates are unbiased. We show that the application of the value equation onto these unbiased parameter estimates results in biased value estimates. Without this assumption the sample mean estimates underestimate the parameters in the average and the value estimates can be unbiased. We state special cases in which this happens. We address this point in detail in Sect. 3.4.

In Singh and Sutton (1996) different kinds of eligibility traces are introduced and analyzed. It is shown that TD(1) is unbiased if the *replace-trace* is used and that it is biased if the usual eligibility trace is used. What is particularly important for our work is one of their side findings: The Maximum Likelihood and the MC estimates are equivalent in a special case. We characterize this special case with Criterion 3 (p. 303) and we make frequent use of this property. We call the criterion the *Full Information Criterion* because all paths that are relevant for value estimation in a state s must hit this state. The idea behind this criterion is to make a "fair" comparison of value estimators that use only paths that hit state s with value estimators that propagate information back to state s (for details see p. 303). This same property is already described in Singh and Sutton (1996, Theorem 5, p. 11).

In this paper we follow a new approach to the finite example case using tools from statistical estimation theory (e.g. Stuart and Ord 1991). Rather than relying on bounds, on

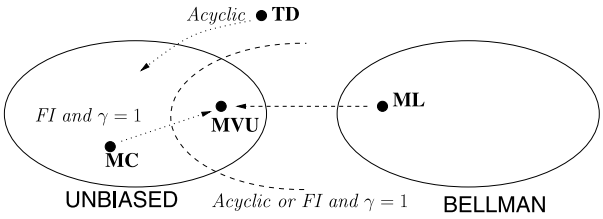

Fig. 1 The figure shows two value estimator classes and four value estimators. On the *left* the class of unbiased value estimators is shown and on the *right* the class of Bellman estimators. The graph visualises to which classes the estimators belong and how the two classes are related. The cursive texts state conditions under which different estimators are equivalent, respectively, under which the two classes overlap: (1) The modified TD estimator (Appendix B) is unbiased if the MRP is acyclic, (2) MC is equivalent to the MVU if the Full Information Criterion (*FI*) is fulfilled and $\gamma = 1$, (3) ML is equivalent to the MVU if the MRP is acyclic or if FI is fulfilled and $\gamma = 1$, (4) in general, the class of unbiased value estimators does not overlap with the class of Bellman estimators. However, both overlap if the MRP is acyclic or if FI is fulfilled and $\gamma = 1$ (indicated by the *dashed semi-ellipse*). The *dashed semi-ellipse* suggests that the classes are not equivalent. In Sect. D.3 (p. 329) we present two simple examples that show that this is indeed the case

approximations, or on results to be recalculated for every specific MRP this approach allows us to derive general statements. Our main results are sketched in Fig. 1. The major contribution is the derivation of the optimal unbiased value estimator (*Minimum Variance Unbiased estimator* (*MVU*), Sect. 3.3). We show that the *Least-Squares Temporal Difference estimator* (*LSTD*) from Bradtke and Barto (1996) is equivalent to the *Maximum Likelihood value estimator* (*ML*) (Sect. 3.4.6) and that both are equivalent to the MVU if the discount $\gamma = 1$ (undiscounted) and the Full Information Criterion is fulfilled or if an acyclic MRP is given (Sect. 3.4.3). In general the ML estimator differs from the MVU because ML fulfills the Bellman equation and because estimators that fulfill the Bellman equation can in general not be unbiased (we refer to estimators that fulfill the Bellman equation in the future as *Bellman estimators*). The main reason for this effect being the probability measures with which the expectations are taken (Sect. 3.1). The bias of the Bellman estimators vanishes exponentially in the number of observed paths. As both estimators differ in general it is natural to ask which of them is better? We show that in general neither the ML nor the MVU estimator are superior to each other, i.e. examples exist where the MVU is superior and examples exist where ML is superior (Sect. D.2).

The *first-visit MC* estimator is unbiased (Singh and Sutton 1996) and therefore inferior to the MVU as the MVU is the optimal unbiased estimator. However, we show that for $\gamma = 1$ the estimator becomes equivalent to the MVU if the Full Information Criterion applies (Sect. 3.5). Furthermore, we show that this equivalence is restricted to the undiscounted case. Finally, we compare the estimators to TD(λ). We show that TD(λ) is essentially unbiased for acyclic MRPs (Appendix B) and is thus inferior to the MVU and to the ML estimator for this case. In the cyclic case TD is biased (Sect. 3.6).

An early version of this work was presented in Grünewälder et al. (2007). The analysis was restricted to acyclic MRPs and to the MC and LSTD estimator. The main findings were that LSTD is unbiased and optimal for acyclic MRPs and that MC equals LSTD in the acyclic case if the Full Information Criterion applies and $\gamma = 1$. It turned out that the second finding was already shown in more generality by Singh and Sutton (1996, Theorem 5). The restriction to acyclic MRPs simplified the analysis considerably compared to the general case which we approach here.

Theoretical findings are summarized in two tables in Sect. 3.7 (p. 310). Symbols are explained at their first occurrence and a table of notations is included in Appendix A. For the sake of readability proofs are presented in Appendix C.

2 Estimation in reinforcement learning

A common approach to the optimization of a control policy is to iterate between estimating the current performance (value estimation) and updating the policy based on this estimate (policy improvement). Such an approach to optimization is called *policy iteration* (Sutton and Barto 1998; Bertsekas and Tsitsiklis 1996). The value estimation part is of central importance as it determines the direction of the policy improvement step.

In this work we focus on this value estimation problem and we study it for Markov Reward Processes. In Reinforcement Learning Markov Decision Processes are typically used. An MRP is the same with the only difference being that the policy does not change over time.

2.1 Markov reward processes

A Markov Reward Process consists of a state space $\mathbb{S}$ (in our case a finite state space), probabilities p_i to start in state i, transition probabilities p_{ij} and a random reward R_{ij} between states i and j. The MRP is *acyclic* if no state i and no path $\pi = (s_1, s_2, s_3, \ldots)$ exists such that $P(\pi) := p_{s_1 s_2} p_{s_2 s_3} \ldots > 0$ and state i is included at least twice in π.

Our goal is to estimate the values V_i of the states in $\mathbb{S}$, i.e. the expected future reward received after visiting state i. The value is defined as

$$V_i = \sum_{j \in \mathbb{S}} p_{ij} \big(\mathbb{E}[R_{ij}] + \gamma V_j \big) \text{ and in vector notation by } \mathbf{V} = \sum_{t=0}^{\infty} \gamma^t \mathbf{P}^t \mathbf{r} = (\mathbf{I} - \gamma \mathbf{P})^{-1} \mathbf{r},$$

where $\mathbf{P} = (p_{ij})$ is the transition matrix of the Markov process, $\mathbf{I}$ the identity matrix, $\gamma \in (0, 1]$ a discount factor and $\mathbf{r}$ is the vector of the expected one step reward ($\mathbf{r}_i = \sum_{j \in \mathbb{S}} p_{ij} \mathbb{E}[R_{ij}]$). In the undiscounted case ($\gamma = 1$) we assume that with probability one a path reaches a terminal state after a finite number of steps.

A large part of this work is concerned with the relation between the maximum likelihood value estimator and the optimal unbiased value estimator. In particular, we are interested in equivalence statements for these two estimators. Equivalence between these estimators can only hold if the estimates for the reward are equivalent, meaning that the maximum likelihood estimator for the reward distribution matches with the optimal unbiased estimator. We therefore restrict our analysis to reward distributions with this property, i.e. we assume throughout that the following assumption holds:

Assumption 1 *The maximum likelihood estimate of the mean reward is unbiased and equivalent to the optimal unbiased estimate.*

The assumption is certainly fulfilled for *deterministic rewards*. Other important cases are *normal distributed, binomial and multinomial distributed rewards*.

2.2 Value estimators and statistical risk

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

$$\mathbb{E}[\mathcal{L}(\bar{V}_i, V_i)], \tag{1}$$

where $\bar{V}_i$ is a value estimator of state i and $\mathcal{L}$ is a loss function, which penalizes the deviation from the true value V_i. We will mainly use the mean squared error

$$\mathrm{MSE}[\bar{V}_i] := \mathbb{E}[(\bar{V}_i - V_i)^2], \tag{2}$$

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

$$\mathrm{MSE}[\bar{V}_i] = \underbrace{\mathbb{V}[\bar{V}_i]}_{\text{Variance}} + \underbrace{(\mathbb{E}[\bar{V}_i - V_i])^2}_{\text{Bias}}. \tag{3}$$

An estimator is called *unbiased* if the bias term is zero. The unbiasedness of an estimator depends on the underlying probability distribution with which the mean is calculated.

Typically, there is a chance that a state is not visited at all by an agent and it makes no sense to estimate the value if this event occurs. We encode the probability event that state i has not been visited with $\{N_i = 0\}$ and that it has been visited at least once with $\{N_i \geq 1\}$, where N_i denotes the number of visits of state i. Unbiased estimators are estimators that are correct in the mean. However, if we take the (unconditional) mean for an MRP then we include the term $\mathbb{E}[\bar{V}_i|\{N_i = 0\}]$ into the calculation, i.e. the value estimate for the case that the estimator has not seen a single example. This is certainly not what we want. We therefore measure the bias of an estimator using the conditional expectation $\mathbb{E}[\cdot\,|\{N_i \geq 1\}]$.

Equal weighting of examples We conclude this section by citing a simple criterion with which it is possible to verify unbiasedness and minimal MSE in special cases. This criterion provides an intuitive interpretation of a weakness of the TD(λ) estimator (see Sect. 3.6). Let $x_i, i = 1, \ldots, n$ be a sample consisting of $n \geq 1$ *independent and identically distributed (iid)* elements of an arbitrary distribution. The estimator

$$\sum_{i=1}^{n} \alpha_i x_i, \quad \text{with} \quad 0 \leq \alpha_i \leq 1, \quad \text{and} \quad \sum_{i=1}^{n} \alpha_i = 1, \tag{4}$$

is unbiased. It has the lowest variance for $\alpha_i = 1/n$ (Stuart and Ord 1991). The x_i could, for example, be the summed rewards for n different paths starting in the same state s, i.e. $x_i := \sum_{t=0}^{\infty} \gamma^t R_t^{(i)}$, where $R_t^{(i)}$ denotes here the reward at time t in path i. The criterion states that for estimators which are linear combinations of iid examples all examples should have an equal influence and none should be preferred over another. However, it is important to notice that not all unbiased estimators must be linear combinations of such sequences and that better unbiased estimators might exist. In fact this is the case for MRPs. The structure of an MRP allows better value estimates.

2.3 Temporal difference learning

A commonly used value estimator for MRPs is the TD(λ) estimator (Sutton 1988). It converges on average (L^1-convergence, Sutton 1988) and it converges almost surely to the correct value (Watkins and Dayan 1992; Jaakkola et al. 1994). In practical tasks it seems to

outperform the MC estimator with respect to convergence speed and its computational costs are low. Analyses for the TD(0) estimator are often less technical. We therefore restrict some statements to this estimator. TD(0) can be defined by means of an update equation:

$$\bar{V}_s^{(i+1)} = \bar{V}_s^{(i)} + \alpha_{i+1}(R_{ss'}^{(i+1)} + \gamma \bar{V}_{s'}^{(j_{i+1})} - \bar{V}_s^{(i)}), \tag{5}$$

where $V_s^{(i)}$ denotes the ith value estimate of state s, i.e. the value estimate after state s has been visited i times. s' denotes the successor state of state s in the corresponding path. $\bar{V}_{s'}^{(j_{i+1})}$ denotes the value estimate of state s' when state s is visited for the $(i+1)$th time. The use of (i) differs slightly from the last section as not all paths need to hit state s anymore, respectively can hit state s multiple times. Finally, α_{i+1} is the learning rate and $R_{ss'}^{(i+1)}$ is the reward which occurred during the transition from s to s'. The general TD(λ) update equation is heavy in notation if we want to keep track of all the variables. We therefore drop some indices that are not of major importance for our analysis. The update equation is given by

$$\bar{V}_s^{(i+1)} = \bar{V}_s^{(i)} + \Delta \bar{V}_s^{(i+1)} \quad \text{and} \quad \Delta \bar{V}_s^{(i+1)} = \alpha_{\phi(i+1)}(R_{tt'} + \gamma \bar{V}_{t'} - \bar{V}_t) e_s^{(i+1)}.$$

We suppressed the indices of $R_{tt'}$, $\bar{V}_{t'}$ and $\bar{V}_t$. $\bar{V}_s^{(i+1)}$ is like before the ith value estimate of state s. t denotes the state at the $i+1$th update and t' the successor state. $R_{tt'}$ is the reward gained at that transition and $\bar{V}_{t'}$ and $\bar{V}_t$ are the value estimates for state t and t'. We want the learning rate to change with the number of visits of state s and not with each update of $\bar{V}_s$, therefore we use the function $\phi(i+1)$ to map $i+1$ to the corresponding number of visits of state s. Finally, $e_s^{(i+1)}$ is an *eligibility trace*. The update equation can be applied after each transition (*online*), when a terminal state is reached (*offline*) or after an entire set of paths has been observed (*batch update*). The eligibility trace can be defined in various ways. Two important definitions are the *accumulating trace* and the *replacing trace* (Singh and Sutton 1996). In Singh and Sutton (1996) it is shown that for $\lambda = 1$ the TD(λ) estimator corresponding to the accumulating trace is biased while the one corresponding to the replacing trace is unbiased. The replacing trace is defined by

$$e_s^{(i+1)} = \begin{cases} 1 & \text{if } s = t, \\ \gamma\lambda e_s^{(i)} & \text{else.} \end{cases} \tag{6}$$

The idea is here that the value estimate for state s gets a TD(0) like update in step $i+1$ if state s is visited in that step ($s = t$) and otherwise it gets a scaled down update. For acyclic MRPs both definitions are equivalent. For $\lambda < 1$ the estimators are biased towards their initialization value. However, a minor modification is sufficient to delete the bias for acyclic MRPs (Appendix B on p. 319). We will mostly use this modified version.

2.4 Monte Carlo estimation

The Monte Carlo estimator is the sample mean estimator of the summed future reward (Sutton and Barto 1998). For acyclic MRPs the MC estimator is given by

$$\frac{1}{n} \sum_{i=1}^{n} \left(\sum_{t=0}^{\infty} \gamma^t R_t^{(i)} \right), \tag{7}$$

where n is the number of paths that have been observed.

In the cyclic case there are two alternative MC estimators: *First-visit MC* and *every-visit MC*. First-visit MC makes exactly one update for each visited state. It uses the part of the

path which follows upon the first visit of the relevant state. The first-visit MC estimator $\bar{V}_i$ is unbiased for every state i, i.e. $\mathbb{E}[\bar{V}_i | N_i \geq 1] = V_i$. Every-visit MC makes an update for each visit of the state. The advantage of the every-visit MC estimator is that it has more samples available for estimation, however, the paths overlap and the estimator is therefore biased (Singh and Sutton 1996). Both estimators converge almost surely and on average to the correct value.

The MC estimators are special cases of TD(λ). The every-visit MC estimator is equivalent to TD(λ) for the *accumulate trace* and the first-visit MC estimator for the *replace trace* if $\lambda = 1$ and $\alpha_i = 1/i$.

3 Comparison of estimators: theory

The central theme of this paper is the relation between two important classes of value estimators and between four concrete value estimators. One can argue that the two most important estimator classes are the estimators that *fulfill the Bellman equation* and estimators that are *unbiased*. The former class is certainly of great importance as the Bellman equation is the central equation in Reinforcement Learning. The latter class proved its importance in statistical estimation theory, where it is the central class of estimators that is studied. We analyse the relation between these two classes.

We concentrate on popular Reinforcement Learning estimators (the Monte-Carlo and the Temporal Difference estimator) and on estimators that are optimal in the two classes. These are: (1) The optimal unbiased value estimator which we derive in Sect. 3.1. (2) The Maximum Likelihood (ML) estimator for which one can argue (yet not prove!) that it is the best estimator in the class of Bellman estimators.

Parts of this section are very technical. We therefore conclude this motivation with a high level overview of the main results.

Estimator classes: unbiased vs. Bellman estimators The key finding for these two estimator classes is that cycles in an MRP essentially separate them. That means if we have an MRP with cycles then the estimators either do not fulfill the Bellman equation or some of them must be biased. The main factor that is responsible for this effect is the "normalization" $\{N_i \geq 1\}$. The Bellman equation couples the estimators, yet the estimators must be "flexible" to be unbiased with respect to different probability measures, i.e. the conditional probabilities $\mathbb{P}[\cdot \,|\{N_i \geq 1\}]$.

Furthermore, we show that the discount has an effect on the bias of Bellman estimators. Estimators that use the Bellman equation are based on parameter estimates $\bar{p}_{ij}$. We show that these parameter estimates must be discount dependent. Otherwise, a "further bias" is introduced.

We show that these factors are the main factors for the separation of the classes: (1) If the MRP is acyclic or (2) if the problem with the normalization and the discount is not present then Bellman estimators can be unbiased.

Estimator comparison and ordering: MVU, ML, TD and MC The key contribution in this part is the derivation of the optimal unbiased value estimator. We derive this estimator by conditioning the first-visit Monte Carlo estimator with "all the information" that is available through the observed paths and we show that the resulting estimator is optimal with respect to any convex loss function. The conditioning has two effects: (1) The new estimator uses the Markov structure to make use of (nearly) all paths. (2) It uses "consistent" alternative cycles

beside the observed ones. For example, if a cyclic connection from state $1 \to 1$ is observed once in the first run and three times in the second run, then the optimal estimator will use paths with the cyclic connection being taken 0 to 4 times. Consistent with this finding, we show that if the first-visit MC estimator observes all paths and the modification of cycles has no effect, then the first-visit MC estimator is already optimal.

Furthermore, the methods from statistical estimation theory allow us to establish a strong relation between the MVU and the Maximum Likelihood estimator. The ML estimator also uses all the information, but it is typically biased as it fulfills the Bellman equation. However, in the cases where the ML estimator is unbiased it is equivalent to the MVU. In particular, the ML estimator is unbiased and equivalent to the MVU for acyclic MRPs or for MRPs without discount where the Full Information Criterion applies.

In the final theory part we are addressing the Temporal Difference estimator. In contrast to MC and ML the theoretical results for TD are not as strong. The reason being that the tools from statistical estimation theory that we are applying can be used to compare estimators inside one of the two estimator classes. However, TD is typically neither contained in the class of unbiased estimators nor in the class of Bellman estimators. We are therefore falling back to a more direct comparison of TD to ML. The analysis makes concrete the relation of the optimal value estimator to TD and demonstrates the power of the Lehmann-Scheffe theorem.

Beside the mentioned equivalence statements between different estimators we are also establishing orderings like "the MVU is at least as good as the first-visit MC estimator" or we are giving counter-examples if no ordering exists.

3.1 Unbiased estimators and the Bellman equation

In this section we analyze the relation between unbiased estimators and Bellman estimators. Intuitively, we mean by "a value estimator $\bar{\mathbf{V}}$ fulfills the Bellman equation" that $\bar{\mathbf{V}} = \bar{\mathbf{r}} + \gamma \bar{\mathbf{P}} \bar{\mathbf{V}}$, where $\bar{\mathbf{r}}, \bar{\mathbf{P}}$ are the rewards, respectively the transition matrix, of a well defined MRP. We make this precise with the following definition:

Definition 1 (Bellman Equation for Value Estimators) An estimator $\bar{\mathbf{V}}$ fulfills the Bellman equation if an MRP $\bar{\mathbf{M}}$ exists with the same state space as the original MRP, with a transition matrix $\bar{\mathbf{P}}$, deterministic rewards $\bar{\mathbf{r}}$ and with value $\bar{\mathbf{V}}$, i.e. $\bar{\mathbf{V}} = \bar{\mathbf{r}} + \gamma \bar{\mathbf{P}} \bar{\mathbf{V}}$. Furthermore, $\bar{\mathbf{M}}$ is not allowed to have additional connections, i.e. $\bar{\mathbf{P}}_{ij} = 0$ if in the original MRP $\mathbf{P}_{ij} = 0$ holds.

Two remarks: Firstly, we restrict the MRP $\bar{\mathbf{M}}$ to have deterministic rewards for simplicity. Secondly, the last condition is used to enforce that the MRP $\bar{\mathbf{M}}$ has a "similar structure" as the original MRP. However, it is possible for $\bar{\mathbf{M}}$ to have fewer connections. For example, this will be the case if not every transition $i \to j$ has been observed.

Constraining the estimator to fulfill the Bellman equation restricts the class of estimators considerably. Essentially, the only degree of freedom is the parameter estimate $\bar{\mathbf{P}}$. If $\mathbf{I} - \gamma \bar{\mathbf{P}}$ is invertible then

$$V(\bar{\mathbf{P}}, \bar{\mathbf{r}}) := \bar{\mathbf{V}} = (\mathbf{I} - \gamma \bar{\mathbf{P}})^{-1} \bar{\mathbf{r}}, \tag{8}$$

i.e. $\bar{\mathbf{V}}$ is completely specified by $\bar{\mathbf{P}}$ and $\bar{\mathbf{r}}$. Here, $V(\bar{\mathbf{P}}, \bar{\mathbf{r}})$ denotes the value function for an MRP with parameters $\bar{\mathbf{P}}$ and rewards $\bar{\mathbf{r}}$. In particular, the Bellman equation couples the value estimates of different states. This coupling of the value estimates introduces a bias. The intuitive explanation of the bias is the following: Assume we have two value estimators $\bar{V}_i, \bar{V}_j$ and both are connected with a connection $i \to j$ and $\bar{p}_{ij} = 1$ holds. Fixing, $\mathbb{E}[\bar{V}_j | \{N_j \geq 1\}] = V_j$ defines then essentially the value for $\bar{V}_i$ as $\bar{V}_i = r_{ij} + \gamma \bar{V}_j$. Yet,

the value for $\bar{V}_i$ must be flexible to allow $\bar{V}_i$ to depend on the probability of $\{N_i \geq 1\}$, as $\mathbb{E}[\bar{V}_i | \{N_i \geq 1\}] = V_i$ must hold. It is in general not possible to fulfill both constraints simultaneously in the cyclic case, i.e. constraining $\bar{V}_i$ for all states i and enforcing the Bellman equation. However, value estimators for single states can be unbiased, even if the Bellman equation is fulfilled.

Another factor that influences the bias is the discount γ. If the Bellman equation is fulfilled by $\bar{\mathbf{V}}$ then the value estimate can be written as $\sum_{t=0}^{\infty} \gamma^t \bar{\mathbf{P}}^t \bar{\mathbf{r}}$, i.e. γ^t weights the estimate $\bar{\mathbf{P}}^t$ of $\mathbf{P}^t$. If $\mathbb{E}[\bar{\mathbf{P}}^t] \neq \mathbf{P}^t$ and the parameter estimate $\bar{\mathbf{P}}$ is independent of γ then with varying γ the deviations of $\bar{\mathbf{P}}^t$ from $\mathbf{P}^t$ are weighted differently and it is intuitive that we can find a γ for which the weighted deviations do not cancel out and the estimator is not unbiased. This effect can be circumvented by making the parameter estimator $\bar{\mathbf{P}}$ discount dependent.

3.1.1 Normalization $\mathbb{P}[\{N_i \geq 1\}]$ and value estimates on $\{N_i = 0\}$

Consider the MRP shown in Fig. 2(B) and let the number of observed paths be one ($n = 1$). The agent starts in state 2 and has a chance of p to move on to state 1. The value of state 1 and 2 is

$$V_1 = V_2 = (1-p) \sum_{i=0}^{\infty} i p^i = \frac{p}{1-p}. \tag{9}$$

Using the sample mean parameter estimate $\bar{p} = i/(i+1)$, we get the following value estimate for state 2:

$$\bar{V}_2(i) = \frac{\bar{p}}{1-\bar{p}} = i \quad \Rightarrow \quad \mathbb{E}[\bar{V}_2(i)] = (1-p) \sum_{i=0}^{\infty} \bar{V}_2(i) p^i = V_2, \tag{10}$$

where $\bar{V}_2(i)$ denotes the value estimate, given the cyclic transition has been taken i times. The estimator fulfills the Bellman equation. Therefore, $\bar{V}_1(i) = \bar{V}_2(i) = i$, given at least one visit of state 1, i.e. conditional on the event $\{N_1 \geq 1\}$. The expected value estimate for state 1 is therefore

$$\mathbb{E}[\bar{V}_1(i) | \{N_1 \geq 1\}] = \frac{(1-p) \sum_{i=1}^{\infty} i p^i}{(1-p) \sum_{i=1}^{\infty} p^i} = \frac{p/(1-p)}{1-(1-p)} = \frac{V_1}{p}, \tag{11}$$

where $(1-p) \sum_{i=1}^{\infty} p^i = p$ is the normalization. Hence, the estimator is biased.

Intuitively the reasons for the bias are: Firstly, $\bar{V}_1$ equals $\bar{V}_2$ on $\{N_1 \geq 1\}$ but the estimators differ (in general) on $\{N_1 = 0\}$. In the example, we did not exploit this issue. We could make use of it by introducing a reward for the transition $2 \to 3$. Secondly, the normalization differs, i.e. $\mathbb{E}[\cdot]$ versus $\mathbb{E}[\cdot \,| \{N_1 \geq 1\}]$. In our example we exploited this issue. Both estimators are 0 on $\{N_1 = 0\}$ and are therefore always equivalent. However, the expectation is calculated differently and introduces the bias.

The following Lemma shows that this problem does not depend on the parameter estimate we used:

Lemma 1 (p. 323) *For the MRP from Fig. 2(B) there exists no parameter estimator $\bar{p}$ such that $V_i(\bar{p})$ is unbiased for all states i.*

How do these effects behave with respect to the number n of observed paths? Let p_i denote the probability to visit state i in one sampled path. Then the probability of the

event $\{N_i = 0\}$ drops exponentially fast, i.e. $\mathbb{P}[\{N_i = 0\}] \leq (1 - p_i)^n$ and the normalization $1/\mathbb{P}[\{N_i \geq 1\}]$ approaches one exponentially fast. Therefore, if the estimates are upper bounded on $\{N_i = 0\}$ then the bias drops exponentially fast in n.

3.1.2 Discount

Consider the MRP from Fig. 2(A) with modified reward $R_{11} = 0, R_{12} = 1$ for one run ($n = 1$) and for $\gamma < 1$. We use again the sample mean parameter estimate, i.e. $\bar{p} = i/(i+1)$ if the cyclic transition has been taken i times. The value of state 1 is

$$V_1 = (1-p)\sum_{i=0}^{\infty} \gamma^i p^i = \frac{1-p}{1-\gamma p} \quad \text{and the value estimate is} \quad \bar{V}_1 = \frac{1 - i/(i+1)}{1 - \gamma i/(i+1)}.$$

The estimator is unbiased if and only if

$$(1-p)\sum_{i=0}^{\infty} \gamma^i p^i \stackrel{?}{=} \mathbb{E}[\bar{V}_1] = (1-p)\sum_{i=0}^{\infty} \frac{1 - i/(i+1)}{1 - \gamma i/(i+1)} p^i. \tag{12}$$

The equality marked with $\stackrel{?}{=}$ holds if and only if

$$\sum_{i=0}^{\infty} \left(\gamma^i - \frac{1 - i/(i+1)}{1 - \gamma i/(i+1)} \right) p^i = 0. \tag{13}$$

With induction one sees that $\gamma^i \leq \frac{1-i/(i+1)}{1-\gamma i/(i+1)}$. *Induction basis* ($i = 1$):

$$\gamma \leq \frac{1 - \frac{1}{2}}{1 - \frac{\gamma}{2}} \quad \Leftrightarrow \quad \gamma - \frac{\gamma^2}{2} \leq \frac{1}{2} \quad \Leftrightarrow \quad 0 \leq (1-\gamma)^2 = 1 - 2\gamma + \gamma^2. \tag{14}$$

Induction step:

$$\begin{aligned} \gamma^{i+1} &\stackrel{\text{I.H.}}{\leq} \gamma \frac{1 - \frac{i}{i+1}}{1 - \gamma \frac{i}{i+1}} \stackrel{?}{\leq} \frac{1 - \frac{i+1}{i+2}}{1 - \gamma \frac{i+1}{i+2}} \quad \Leftrightarrow \\ \gamma \left(1 - \frac{i}{i+1} - \frac{\gamma}{i+2} \right) &\leq 1 - \frac{i+1}{i+2} - \frac{\gamma i}{(i+1)(i+2)} \quad \Leftrightarrow \\ (i - \gamma i)(\gamma - 1) &\leq (1-\gamma)^2 \quad \Leftrightarrow \quad -(i - \gamma i) \leq 1 - \gamma, \end{aligned} \tag{15}$$

where the last inequality holds, because $-(i - \gamma i) \leq 0$ and $(1 - \gamma) \geq 0$. I.H. denotes *Induction Hypothesis*.

Hence, the estimator is unbiased if and only if the left and the right side of the inequality are equivalent. This is the case only for $\gamma = 1$: (14) shows that only $\gamma = 1$ is a candidate and it is easy to verify that for $\gamma = 1$ both sides of the inequality are equivalent. Thus, the estimator is biased for all $\gamma < 1$ and is only unbiased for $\gamma = 1$.

In general, value estimators that fulfill the Bellman equation must at least be discount dependent to be able to be unbiased for general MRPs, as the following Lemma shows:

Lemma 2 (p. 323) *For the MRP from Fig. 2(A) with modified reward $R_{11} = 0, R_{12} = 1$ and for $n = 1$ there exists no parameter estimator $\bar{p}$ that is independent of γ such that $V_1(\bar{p})$ is unbiased for all parameters p and all discounts γ.*

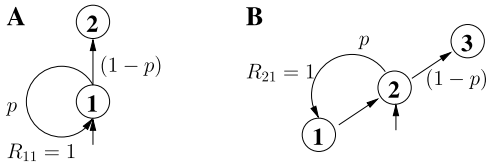

Fig. 2 **A** A cyclic MRP with starting state 1 and with probability p for the cyclic transition. The reward is 0 for the cyclic transition and 1 otherwise. **B** A cyclic MRP with starting state 2 and with probability p for the cyclic transition. The reward is 1 for the cyclic transition from state 2 to state 1 and 0 otherwise

3.2 Maximum likelihood parameter estimates and sufficient statistics

We start this section with a derivation of the maximum likelihood parameter estimates. After that we introduce a minimal sufficient statistics for MRPs and we show that this statistic equals the maximum likelihood estimates.

3.2.1 Maximum likelihood parameter estimates

Let p_{ij} be the transition probability of state i to j, p_i the probability to start in i and x a sample consisting of n iid state sequences $x_1, \ldots, x_n$. The log-likelihood of the sample is

$$\log \mathbb{P}[x|p] = \sum_{k=1}^{n} \log \mathbb{P}[x_k|p]. \tag{16}$$

The corresponding maximization problem is given by

$$\max_{p_{ij}, p_i} \sum_{k=1}^{n} \log \mathbb{P}[x_k|p_{ij}, p_i], \quad \text{s.t.: } \sum_{j\in\mathbb{S}} p_{ij} = \sum_{j\in\mathbb{S}} p_j = 1, \quad \text{for all } i. \tag{17}$$

The unique solution for p_{ij} and p_i (Lagrange multipliers) is given by

$$p_{ij} = \frac{\mu_{ij}}{K_i} =: \bar{p}_{ij} \quad \text{and} \quad p_i = \frac{1}{n}\left(K_i - \sum_{j\in\mathbb{S}} \mu_{ji}\right) =: \bar{p}_i, \tag{18}$$

where K_i denotes the number of visits of state i, μ_{ij} the number of direct transitions from i to j, $\bar{p}_{ij}$ the estimate of the true transition probability p_{ij} and $\bar{p}_i$ the estimate of the true starting probability p_i.

3.2.2 Sufficient statistics

We need some results and concepts from point estimation theory (e.g. Lehmann and Casella 1998). The first important concept that we need is that of a *sufficient statistic* $\mathcal{S}$. A statistic $\mathcal{S}$ is a function of data, e.g. $\mathcal{S}(x) = \sum_{i=1}^{n} x_i$, with $x_1, \ldots, x_n$ being a sample of an arbitrary distribution, is a statistic. The intuition is here that the statistic contains information from the sample in a compressed form. For example, to predict the mean of a distribution it might be enough to know $\sum_{i=1}^{n} x_i$ and the concrete values x_i do not provide any useful information beside that.

The question is now, when does a statistic $\mathcal{S}(x)$ contain all the useful information from a sample? This question leads to the concept of sufficiency. Intuitively, a statistic which

contains all information about a sample is called *sufficient*. In detail, a statistic is sufficient if the probability of a sample is completely determined by the statistic $\mathcal{S}$, i.e. if it is independent of the underlying data generating distribution:

$$\mathbb{P}(X = x|\mathcal{S}(x), \theta) = \mathbb{P}(X = x|\mathcal{S}(x)), \tag{19}$$

where θ parameterises the data generating distribution. If a concrete function of a sample is a sufficient statistic or not depends on the underlying family of distributions. E.g. for some families the sample mean is a sufficient statistic for others it is not.

Why is this now relevant for estimation? The idea is here that an estimator is a function of a sample, e.g. a function that takes $x_1, \ldots, x_n$ as an input and predicts the mean by averaging over the sample. The estimator itself is again a stochastic object as it varies with the drawn sample. Thus, it has a distribution on its own which depends on θ as θ parameterises the data generating process. Now, given a sufficient statistic $\mathcal{S}(x)$ the sample probability is completely described by $\mathcal{S}(x)$ and so is the distribution of the estimator. For example, if for two estimators A and B that are functions of the sufficient statistic $\mathcal{S}(x)$ it holds that $A(\mathcal{S}(x)) = B(\mathcal{S}(x))$ then both have the same distribution over estimates and are essentially equivalent.

Completeness is another important property of a statistic $\mathcal{S}$ which has also to do with this concept of equivalence of estimators. A statistic $\mathcal{S}$ is complete if $\mathbb{E}_\theta[h(\mathcal{S})] = 0$ for all θ implies $h = 0$ almost surely. Let us again take two estimators $A(\mathcal{S}(x))$ and $B(\mathcal{S}(x))$, let us assume that the statistic $\mathcal{S}$ is complete and let us define the function $h(\mathcal{S}) := A(\mathcal{S}) - B(\mathcal{S})$. If $\mathbb{E}_\theta[h(\mathcal{S})] = \mathbb{E}_\theta[A(\mathcal{S}) - B(\mathcal{S})] = \mathbb{E}_\theta[A(\mathcal{S})] - \mathbb{E}_\theta[B(\mathcal{S})] = 0$ for all θ then due to completeness $A(\mathcal{S}(x)) = B(\mathcal{S}(x))$ almost surely. Hence, if the two estimators have the same mean for each θ then they are equivalent almost surely.

A final property of a sufficient statistic that we need is *minimality*. The minimal sufficient statistic is the sufficient statistic with the smallest dimension. For example, the sample itself is a sufficient statistic: $\mathcal{S}(x) = x = (x_1, \ldots, x_n)$. This concrete statistic has dimension n. For some probability distributions the sample mean is a sufficient statistic, i.e. $\mathcal{S}(x) = \sum_{i=1}^n x_i$. This statistic has dimension 1. The minimal statistic has typically the same dimension as the parameter space. To define this formally, suppose that a statistic $\mathcal{S}$ is sufficient for a family of parameters θ. Then $\mathcal{S}$ is minimally sufficient if $\mathcal{S}$ is a function of any other statistic $\mathcal{T}$ that is sufficient for that family.

We need all these concepts to state one important theorem, the so called *Lehmann-Scheffe theorem*. The theorem says that for a complete and minimal sufficient statistics $\mathcal{S}$ and any unbiased estimator A the estimator $\mathbb{E}[A|\mathcal{S}] = \mathbb{E}[A|\mathcal{S}(x)]$ is the optimal unbiased estimator with respect to any convex loss function and hence the unbiased estimator with minimal MSE. If the estimator A is already a function of the sufficient statistic $\mathcal{S}(x)$ then $\mathbb{E}[A|\mathcal{S}] = A$. Otherwise, $\mathbb{E}[A|\mathcal{S}(x)]$ is the average estimate of A, whereas the average is taken over all samples x' that are consistent with $\mathcal{S}(x)$. These are the samples x' for which $\mathcal{S}(x') = \mathcal{S}(x)$ holds.

The proof of the theorem is actually not too difficult and it provides insight into why the different properties of $\mathcal{S}$ are needed. We go through the main steps to help the reader build some intuition:

We already said that completeness is important for the equivalence of estimators. Especially, if two estimators A, B are functions of a sufficient statistic and if both have the same mean then they are equivalent almost surely. This is useful for the theorem as it is assumed that the estimator A is unbiased. Conditioning with $\mathcal{S}$ preserves the unbiasedness and hence $\mathbb{E}[A|\mathcal{S}]$ is unbiased and a function of the sufficient statistic $\mathcal{S}$. Assume now that B is another

unbiased estimator, then $\mathbb{E}[B|\mathcal{S}]$ is also an unbiased estimator and a function of the sufficient statistic and hence equivalent to $\mathbb{E}[A|\mathcal{S}]$. So completeness guarantees a unique point in the class of unbiased estimators to which $\mathbb{E}[\cdot|\mathcal{S}]$ projects.

Sufficiency of the statistic $\mathcal{S}$ is needed as it makes $\mathbb{E}[A|\mathcal{S}]$ independent of the true θ. If $\mathcal{S}$ would not be sufficient then $\mathbb{E}[A|\mathcal{S}]$ could only be calculated with the help of the true θ and in this sense would not be a valid estimator.

To understand the optimality of the estimator $\mathbb{E}[A|\mathcal{S}]$ one needs to understand two things: First, that conditioning with $\mathcal{S}$ reduces the estimation error. This is due to a straight forward application of the Jensen inequality for conditional expectations. Let us take the MSE as an error measure for convenience, then

$$\begin{aligned}\mathrm{MSE}(\mathbb{E}[A|\mathcal{S}]) &= \mathbb{E}[(\mathbb{E}[A|\mathcal{S}]-V)^2] = \mathbb{E}[\mathbb{E}[A-V|\mathcal{S}]^2] \\ &\leq \mathbb{E}[\mathbb{E}[(A-V)^2|\mathcal{S}]] = \mathbb{E}[(A-V)^2] = \mathrm{MSE}(A),\end{aligned} \tag{20}$$

where we denote the true value that we want to estimate with V. Second, that the "maximal" conditioning is achieved by conditioning with a minimal sufficient statistic. Let us assume that $\mathcal{S}$ is a minimal sufficient statistic and $\mathcal{T}$ another (non-minimal) sufficient statistic then we can improve the estimator $\mathbb{E}[A|\mathcal{T}]$ by conditioning it with $\mathcal{S}$. On the contrary, conditioning the estimator $\mathbb{E}[A|\mathcal{S}]$ with $\mathcal{T}$ does not change anything as $\mathcal{S}$ is a function of $\mathcal{T}$ due to the minimality of $\mathcal{S}$.

We want to apply this theorem now to value estimation. The first thing that we need is the minimal sufficient statistic for an MRP.

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

$$\mathbb{P}(\mathbf{x}|p) = \prod_{i=1}^{n}\left(p_{\mathbf{x}_i(1)}\prod_{j=2}^{L_i}p_{\mathbf{x}_i(j-1)\mathbf{x}_i(j)}\right) = \prod_{s\in\mathbb{S}}p_s^{(K_s-\sum_{s'}\mu_{s's})}\prod_{s,s'\in\mathbb{S}}p_{ss'}^{\mu_{ss'}}, \tag{21}$$

where $\mathbf{x}_i(j)$ is the jth state in the ith path, n the number of observed paths and L_i the length of the ith path. $\mu_{ss'}$ is sufficient for $p_{ss'}$ and because sufficiency is sustained by one-to-one mappings (Stuart and Ord 1991) this holds true also for the maximum likelihood estimates $\bar{p}_{ss'}=\mu_{ss'}/K_s$. The sufficient statistic is *minimal* because the maximum likelihood solution is unique (Stuart and Ord 1991).[1] The sufficient statistic is also *complete* because the sample distribution induced by an MRP forms an *exponential family* of distributions (Lemma 4, p. 324). 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

$$p_\theta(x) = \exp\left(\sum_{i=1}^{s}\eta_i(\theta)T_i(x)-A(\theta)\right)h(x) \tag{22}$$

with respect to some common measure μ (Lehmann and Casella 1998). Here, the η_i and A are real-valued functions of the parameters, the T_i are real-valued statistics and x is a point in

[1] It is needed to use the minimal parameter set of the MRP to be formally correct. The minimal sufficient statistics excludes also one value $\mu_{ss'}$, however the missing value is defined by the other μ's.

the sample space. The η's are called *natural parameters*. It is important that the natural parameters are not functionally related. In other words no f should exist with $\eta_2 = f(\eta_1)$. If the natural parameters are not functionally related, then the distribution is complete (Lehmann and Casella 1998). Otherwise, the family forms only a *curved exponential family* and a curved exponential family is not complete.

3.3 Optimal unbiased value estimator

The Lehmann-Scheffe theorem states that for any unbiased estimator A the estimator $\mathbb{E}[A|\mathcal{S}]$ is the optimal unbiased estimator *with probability one* (w.p.1), 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. From now on, we refer to this estimator as the *Minimum Variance Unbiased estimator* (*MVU*).

Theorem 2 (MVU) *Let $\bar{V}$ be the first-visit Monte-Carlo estimator and $\mathcal{S}$ the sufficient and complete statistics for a given MRP. The estimator $\mathbb{E}[\bar{V}|\mathcal{S}]$ is unbiased and the optimal unbiased estimator with respect to any convex loss function w.p.*1. *Especially, it has minimal MSE w.p.*1. *For deterministic reward and $n := |\mathbb{S}|$ the sufficient statistic is*

$$\mathcal{S} = (\bar{p}_1, \ldots, \bar{p}_n, \bar{p}_{11}, \bar{p}_{12}, \ldots, \bar{p}_{nn}),$$

where $\bar{p}_i$ and $\bar{p}_{ij}$ are the maximum likelihood estimates of the MRP parameters and the estimator $\mathbb{E}[\bar{V}|\mathcal{S}]$ is given by

$$\mathbb{E}[\bar{V}|\mathcal{S}] = \frac{1}{|\mathbf{\Pi}(\mathcal{S})|} \sum_{\boldsymbol{\pi} \in \mathbf{\Pi}(\mathcal{S})} \bar{V}(\boldsymbol{\pi}), \tag{23}$$

where $\boldsymbol{\pi} := (\pi_1, \ldots, \pi_i)$ denotes a vector of paths, $\mathbf{\Pi}(\mathcal{S})$ denotes the set of vectors of paths which are consistent with the observation $\mathcal{S}$, $|\cdot|$ is the size of a set and $\bar{V}(\boldsymbol{\pi})$ is the first-visit MC estimate for the vector of paths $\boldsymbol{\pi}$.

Essentially, $\boldsymbol{\pi}$ is an ordered set of paths and it is an element of $\mathbf{\Pi}(\mathcal{S})$ (i.e. consistent with $\mathcal{S}$) if it produces the observed transitions, starts and rewards. In detail, assume again deterministic reward and let us denote the observed sample with $\boldsymbol{\pi}$. $\boldsymbol{\pi}'$ is consistent with $\mathcal{S}(\boldsymbol{\pi}) = (\bar{p}_1, \ldots, \bar{p}_n, \bar{p}_{11}, \bar{p}_{12}, \ldots, \bar{p}_{nn})$ if $\mathcal{S}(\boldsymbol{\pi}') = \mathcal{S}(\boldsymbol{\pi})$. The MC estimate in (23) is simply the average value for the paths in $\boldsymbol{\pi}$. The estimator $\mathbb{E}[\bar{V}|\mathcal{S}]$ is thus the average over all paths which could explain the (compressed) data $\mathcal{S}$. As an example, take the two state MRP from Fig. 2(A). Assume that an agent starts twice in state 1, takes two times the cycle in the first run and zero times in the second. The paths which are consistent with this observation are:

$$\Pi(\mathcal{S}) = \{((1,1,1,2),(1,2)), ((1,1,2),(1,1,2)), ((1,2),(1,1,1,2))\}. \tag{24}$$

The example also shows that the ordering of paths matters. Here, each path $((1,2),(1,1,2)$ and $(1,1,1,2))$ occurs two times while without ordering $(1,1,2)$ would be counted twice as often as the other two paths. The MC estimator for the value of a state s does not consider paths which do not hit s. On the contrary the conditioned estimator uses these paths. To see this take a look at the MRP from Fig. 8(A) at p. 327. Assume, that two paths were sampled:

$(1, 2, 4)$ and $(2, 3)$. The MC value estimate for state one uses only the first path. Taking a look at

$$\mathbf{\Pi}(\mathcal{S}) = \{((1, 2, 4), (2, 3)), ((\mathbf{1}, \mathbf{2}, \mathbf{3}), (2, 4)), ((2, 3), (1, 2, 4)), ((2, 4), (\mathbf{1}, \mathbf{2}, \mathbf{3}))\}, \quad (25)$$

we can observe that the conditioned estimator uses the information. The bold paths are paths that do not occur in the sample and start in state 1, i.e. paths that are used by the MC value estimator for state 1.

3.3.1 Costs of unbiasedness

The intuition that the MVU uses all paths is, however, not totally correct. Let us take a look at the optimal unbiased value estimator of state 1 of the MRP in Fig. 2(B) for $\gamma = 1$. Furthermore, assume that one run is made and that the path $(2, 1, 2, 3)$ is observed. No permutations of this path are possible and the estimate of state 1 is therefore the MC estimate of path $(1, 2, 3)$, which is 0. In general, if we make one run and we observe i transitions from state 2 to state 1, then the estimate is $(i - 1)$. That is we ignore the first transition. As a consequence, we have on average the following estimate:

$$(1 - p) \sum_{i=1}^{\infty} (i - 1) p^i = p \frac{p}{1 - p} = p V_1. \quad (26)$$

The term p is exactly the probability of the event $\{N_1 \geq 1\}$ and the estimator is conditionally unbiased on this event. The intuition is, that the estimator needs to ignore the first transition to achieve (conditional) unbiasedness.

Hence, unbiasedness has its price. Another cost beside this loss in information is that the Bellman equation cannot be fulfilled. In Sect. 3.1 we started with Bellman estimators and we showed that the estimators are biased. Here, we have a concrete example of an unbiased estimator that does not fulfill the Bellman equation, as $\bar{V}_1 = (i - 1) \neq i = \bar{V}_2$. For this example this is counterintuitive as $p_{12} = 1$ and essentially no difference between the states exists in the undiscounted case.

3.3.2 Undiscounted MRPs

In the undiscounted case permutations of paths do not change the cumulative reward. For example, $\sum_{i=1}^{n} R_{\pi_i \pi_{i+1}} = \sum_{i=1}^{n} R_{\pi_{\sigma(i)} \pi_{\sigma(i)+1}}$, if σ is a permutation of $(1, \ldots, n)$, because the time at which a reward is observed is irrelevant. This invariance to permutations implies already a simple fact. We need the following criterion to state this:

Criterion 3 (Full Information) *A state s has full information if, for every successor state $s'(\neq s)$ of s (not only immediate successor states) and any path π, it holds that*

$$\pi_i = s' \quad \Rightarrow \quad \exists j \quad \textit{with } j < i \textit{ and } \pi_j = s.$$

π_i denotes the ith state in the path. In Fig. 2(A) state 1 and 2 have this property and in (B) state 2 and 3. In general, if it is only possible to reach successor states of s by visiting s then the Full Information Criterion applies to state s. For example, if there is a single start state then the criterion applies to that state. A more visual example is a room in a building with a single one-way door located at state s. In this case the Full Information criterion applies to

state s. If there is a second door to the room through which one can enter then the criterion does not apply to s.

Let $\boldsymbol{\pi}$ be a vector of paths following the first visit of state s that are consistent with the observations. $\bar{V}(\boldsymbol{\pi})$ is then given by $(1/|\boldsymbol{\pi}|)\sum_i \sum_j R^{(i)}_{jj+1}$, where $|\boldsymbol{\pi}|$ is the number of paths contained in $\boldsymbol{\pi}$ and $R^{(i)}_{jj+1}$ is the observed reward in path i at position j. Rearranging the path does not change the sum and the normalizing term. Therefore each consistent path results in the same first-visit MC estimate and the MVU equals the first-visit MC estimator.

Corollary 1 *Let $\bar{V}$ be the first-visit MC estimator and let the value function be undiscounted. If the Full Information Criterion applies to a state s, then*

$$\mathbb{E}[\bar{V}_s|\mathcal{S}] = \bar{V}_s. \tag{27}$$

The undiscounted setting allows alternative representations of the optimal estimator. As an example, suppose we observed one path $\pi := (1, 1, 1, 2)$ with reward $R(\pi) = 2R_{11} + 1R_{12}$. The optimal estimator is given by $R(\pi)$. Alternatively, we can set the reward for a path π with j-cycles to $R(\pi) := jR_{11} + R_{12}$ and define a new probability measure $\hat{\mathbb{P}}[\{j \text{ cycles}\}]$ such that $\sum_{j=0}^{\infty} j\hat{\mathbb{P}}[\{j \text{ cycles}\}] = i$, i.e. we average over the set of paths with 0 to "∞" many cycles using the probability measure $\hat{\mathbb{P}}[\{j \text{ cycles}\}]$. If this measure is constrained to satisfy $\sum_{j=0}^{\infty} j\hat{\mathbb{P}}[\{j \text{ cycles}\}] = i$, then

$$\sum_{j=0}^{\infty} \hat{\mathbb{P}}[\{j \text{ cycles}\}](jR_{11} + R_{12}) = iR_{11} + R_{12} = \text{MVU}. \tag{28}$$

We pronounce this point here, because the ML value estimator, which we discuss in the next section, can be interpreted in this way.

3.3.3 Convergence

Intuitively, the estimator should converge because MC converges in L^1 and almost surely. Furthermore, conditioning reduces norm-induced distances to the true value. This is already enough to follow L^1 convergence but the almost sure convergence is not induced by a norm. We therefore refer to an integral convergence theorem which allows us to follow a.s. under the assumption that the MC estimate is upper bounded by a random variable $Y \in L^1$. Details are given in Sect. C.3.

Theorem 4 (p. 325) $\mathbb{E}[\bar{V}|\mathcal{S}]$ *converges on average to the true value. Furthermore, it converges almost surely if the MC value estimate is upper bounded by a random variable $Y \in L^1$.*

Such a Y exists for example, if the reward is upper bounded by R_{max} and if $\gamma < 1$ as in this case each MC estimate is smaller than $R_{max}\sum_{i=0}^{\infty}\gamma^i = R_{max}/(1-\gamma)$.

An MVU algorithm can be constructed using (23). However, the algorithm needs to iterate through all possible paths and therefore has an exponential computation time.

3.4 Least-squares temporal difference learning

In this section we discuss the relation of the MVU to the LSTD estimator. The LSTD estimator was introduced by Bradtke and Barto (1996) and extensively analyzed in Boyan (1998)

and Boyan (1999). Empirical studies showed that LSTD often outperforms TD and MC with respect to convergence speed based on sample size.

In this section we support these empirical findings by showing that the LSTD estimator is equivalent to the MVU for acyclic MRPs and closely related to the MVU for undiscounted MRPs. We derive our statements not directly for LSTD, but for the maximum likelihood value estimator (ML) which is equivalent to LSTD (Sect. 3.4.6). The estimator is briefly sketched in Sutton (1988), where it is also shown that batch TD(0) is in the limit equivalent to the ML estimator. The estimator is also implicitly used in the *certainty-equivalence* approach, where a maximum likelihood estimate of an MDP is typically used for optimization.

3.4.1 Maximum likelihood estimator

The ML value estimator is given by $V(\bar{\mathbf{P}}, \bar{\mathbf{r}})$, where $\bar{\mathbf{P}} := (\bar{p}_{ij})$ is the maximum likelihood estimate of the transition matrix and $\bar{\mathbf{r}}$ is the vector of the maximum likelihood estimates of the expected one step reward. Hence, the ML value estimator is given by:

$$\bar{\mathbf{V}} = \sum_{i=0}^{\infty} \gamma^i \bar{\mathbf{P}}^i \bar{\mathbf{r}} = (\mathbf{I} - \gamma \bar{\mathbf{P}})^{-1} \bar{\mathbf{r}}, \tag{29}$$

whereas the Moore-Penrose pseudoinverse is used if $\bar{\mathbf{P}}$ is singular (e.g. too few samples).

3.4.2 Unbiasedness and the MVU

If an estimator is a function of the sufficient statistic (e.g. $\bar{V} = f(\mathcal{S})$) then the conditional estimator is equal to the original estimator, $\bar{V} = \mathbb{E}[\bar{V}|\mathcal{S}]$. If the estimator $\bar{V}$ is also unbiased then it is the optimal unbiased estimator w.p.1 due to the Lehmann-Scheffe theorem. The defined maximum likelihood estimator is a function of a minimal and complete sufficient statistic. This statistic consists of the counts for the different transitions. These counts are used to compute the estimate of the transition matrix and the maximum likelihood value estimator is a function of this estimated transition matrix. Therefore, the following relation holds between the ML estimator and the MVU:

Corollary 2 *The ML estimator is equivalent to the MVU w.p.*1, *if and only if it is unbiased.*

The following two subsections address two cases where ML is unbiased.

3.4.3 Acyclic MRPs

The ML estimator is unbiased in the acyclic case and therefore equivalent to the MVU.

Theorem 5 (p. 325) *The ML estimator is unbiased if the MRP is acyclic.*

Corollary 3 *The ML estimator is equivalent to the MVU w.p.*1 *if the MRP is acyclic.*

3.4.4 Undiscounted MRPs

It is also possible that ML value estimates for specific states are unbiased even if the MRP is cyclic. One important case in which ML value estimates are unbiased is characterized by the Full Information Criterion. If it applies to a state i then the problem of Sect. 3.1.1 does

not affect state i. For example, in Sect. 3.1.1 we showed that the value estimator for state 1 is biased. However, (10) shows that in the same example the value estimator for state 2 is unbiased and for state 2 the Full Information Criterion applies.

This can be shown by using Theorem 5 from Singh and Sutton (1996), which states that the ML estimator equals the first-visit MC estimator if the Full Information Criterion holds and $\gamma = 1$. Furthermore, in this case the first-visit MC estimator is equivalent to the MVU w.p.1 (Corollary 1). Hence, ML is unbiased and optimal w.p. 1. We state this as a corollary:

Corollary 4 *The ML estimator of a state i is unbiased and equivalent to the MVU w.p.*1 *if the Full Information Criterion applies to state i and if $\gamma = 1$.*

We analyze this effect using a simple MRP and we give two interpretations.

Example: Cyclic MRP—unbiased We start with calculating the bias of ML explicitly for a simple MRP and thus "verifying" the Corollary. The value of state 1 for the MRP of Fig. 2(A) with $\gamma = 1$ is $V_1 = (1-p)\sum_{i=0}^{\infty} i p^i$. The Full Information Criterion applies to state 1. The ML estimate for a sample of n paths is

$$\bar{V}_1 = \left(1 - \frac{k}{k+n}\right) \sum_{i=0}^{\infty} i \left(\frac{k}{k+n}\right)^i = \left(1 - \frac{k}{k+n}\right) \frac{k/(k+n)}{(1-k/(k+n))^2} = \frac{k}{n}, \tag{30}$$

where k is the number of cycles taken (summed over all observed paths). Therefore

$$\mathbb{E}[\bar{V}_1] = \mathbb{E}\left[\frac{k}{n}\right] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[k_i]. \tag{31}$$

Furthermore,

$$\mathbb{E}[k_i] = (1-p) \sum_{k_i=0}^{\infty} k_i p^{k_i} = V_1 \tag{32}$$

and the ML estimator is unbiased. Now, Corollary 2 tells us that the ML estimator is equivalent to the MVU w.p.1.

It is also possible to show this equivalence using simple combinatorial arguments. The MVU and the MC estimate for this MRP is $\frac{k}{n}$: Let u be the number of ways of splitting k into n paths. For each split the summed reward is k and the MC estimate is therefore $\frac{k}{n}$. Hence, the MVU is $\frac{uk/n}{u} = \frac{k}{n}$.

Interpretation I: Non-linearity vs. underestimated parameters It is interesting that ML is unbiased in this example. In general nonlinear transformations of unbiased parameter estimates produce biased estimators, as

$$\mathbb{E}[f(\bar{\theta})] = f(\theta) = f(\mathbb{E}[\bar{\theta}]) \tag{33}$$

essentially means that f is a linear transformation as f and $\mathbb{E}$ commute. Furthermore, the value function is a nonlinear function. Yet, in our example the parameter estimator $\bar{\theta}$ is actually not unbiased. For $n = 1$:

$$\mathbb{E}\left[\frac{k}{k+1}\right] = (1-p) \sum_{k=0}^{\infty} \frac{k}{k+1} p^k < (1-p) \sum_{k=1}^{\infty} p^k = (1-p) \sum_{k=0}^{\infty} p^{k+1} = p. \tag{34}$$

The parameter is underestimated on average. The reason for this lies in the dependency between the visits of state 1. For a fixed number of visits with *iid* observations, the parameter estimate would be unbiased. The relation between these two estimation settings is very similar to the first-visit and every-visit MC setting. The first-visit MC estimator is unbiased because it uses only one observation per path while the every-visit MC estimator is biased. In our case, the effect is particularly paradoxical as for the *iid* case the value estimator is biased.

Interpretation II: Consistency of the set of paths The ML estimator differs in general from the MVU because it uses paths that are inconsistent with the observation $\mathcal{S}$. For example, given the MRP from Fig. 2(A) with the observation $(1, 1, 1, 2)$. The set of paths consistent with this observation is again $\{(1, 1, 1, 2)\}$. The ML estimator, however, uses the following set of paths

$$\{(1, 2), (1, 1, 2), (1, 1, 1, 2), (1, 1, 1, 1, 2) \ldots\}, \tag{35}$$

with a specific weighting $\hat{\mathbb{P}}[\{j \text{ cycles}\}]$ for a path that contains j cycles. In general, this representation will result in an estimate that is different from the MVU estimate. However, if Corollary 4 applies then both representations are equivalent. State 1 fulfills the properties of the Full Information Criterion. The ML estimator can therefore be represented as a sum over the cycle times with each summand being a product between the estimated path probability and the reward of the path if $\gamma = 1$. One can see this easily for the example (one run with $i = 2$ cycles being taken): The path probability is in this case simply $\hat{\mathbb{P}}[\{j \text{ cycles}\}] = \bar{p}^j(1 - \bar{p})$ and because $\sum_{j=0}^{\infty} j\,\bar{p}^j(1 - \bar{p}) = i = 2$ ((30) with $n = 1$) the estimate is equal to $2R_{11} + R_{12}$ which is exactly the MVU estimate (compare to (28) on p. 304).

3.4.5 Which estimator is better? The MVU or ML?

The MVU is optimal in the class of unbiased estimators. However, this does not mean that the ML estimator is worse than the MVU. The ML estimator is also a function of the sufficient statistics, it is just not unbiased. To demonstrate this, we present two examples based on the MRP from Fig. 2(A) in Sect. D.2 (p. 328). One for which the MVU is superior and one where the ML estimator is superior. We summarize this in a corollary:

Corollary 5 *MRPs exist in which the MVU has a smaller MSE than the ML estimator and MRPs exist in which the ML estimator has a smaller MSE than the MVU.*

3.4.6 The LSTD estimator

The LSTD algorithm computes analytically the parameters which minimize the empirical quadratic error for the case of a linear system. Bradtke and Barto (1996) showed that the resulting algorithm converges almost surely to the true value. In Boyan (1998) a further characterization of the least-squares solution is given. This turns out to be useful to establish the relation to the ML value estimator. According to this characterization, the LSTD estimate $\bar{V}$ is the unique solution of the Bellman equation, i.e.

$$\bar{\mathbf{V}} = \bar{\mathbf{r}} + \gamma \bar{\mathbf{P}} \bar{\mathbf{V}}, \tag{36}$$

where $\bar{\mathbf{r}}$ is the sample mean estimate of the reward and $\bar{\mathbf{P}}$ is the maximum likelihood estimate of the transition matrix.

Comparing (36) with (29) of the ML estimator it becomes obvious that both are equivalent if the sample mean estimate of the reward equals the maximum likelihood estimate.

Corollary 6 *The ML value estimator is equivalent to LSTD if the sample mean and the maximum likelihood estimator of the expected reward are equivalent.*

3.5 Monte Carlo estimation

We first summarize Theorem 5 from Singh and Sutton (1996) and Corollary 1 from p. 304:

Corollary 7 *The (first-visit) MC estimator of a state i is equivalent to the MVU and to the ML estimator w.p.*1 *if the Full Information Criterion applies to state i and an undiscounted MRP is given.*

Essentially, the corollary tells us that in the undiscounted case it is only the "amount" of information that makes the difference between the MC and MVU estimators, and the MC and ML estimators. Amount of information refers here to the observed paths. If MC observes every path then the estimators are equivalent.

From a different point of view this tells us that in the undiscounted case the MRP structure is only useful for passing information between states, but yields no advantage beyond that.

3.5.1 Discounted MRPs

In the discounted cyclic case the MC estimator differs from the ML and the MVU estimator. It differs from ML because ML is biased. The MC estimator is equivalent to the MVU in the undiscounted case because the order in which the reward is presented is irrelevant. That means the time at which a cycle occurs is irrelevant. In the discounted case this is not true anymore. Consider again the MRP from Fig. 2(A) with the following two paths $\boldsymbol{\pi} = ((1,1,1,2),(1,2))$. The MC estimate is $1/2((1+\gamma)+0)$. The set of paths consistent with this observation is $\boldsymbol{\Pi}(\mathcal{S}) = \{((1,1,1,2),(1,2)),((1,1,2),(1,1,2)),((1,2),(1,1,1,2))\}$. Hence, the MVU uses the cycle $(1,1,2)$ besides the observed ones. The MVU estimate is $1/3((1+\gamma)/2+2/2+(1+\gamma)/2) = 1/3(2+\gamma)$. Both terms are equivalent if and only if $\gamma = 1$. For this example the Full Information Criterion applies.

Similarly, for acyclic MRPs the MC estimator is different from the ML/MVU estimator if $\gamma < 1$. Consider a 5 state MRP with the following observed paths: $((1,3,4),(1,2,3,5))$, a reward of $+1$ for $3 \to 4$ and -1 for $3 \to 5$. The ML estimate for state 1 is $(1/4\gamma^2 + 1/4\gamma)(1-1) = 0$, while the MC estimate is $1/2(-\gamma^2+\gamma)$ which is 0 if and only if $\gamma = 1$. Again the Full Information Criterion applies.

3.5.2 Ordering with respect to other value estimators

Beside the stated equivalence the MVU is for every MRP at least as good as the first-visit MC estimator, because the first-visit MC estimator is unbiased. The relation to ML is not that clear cut. In general MRPs exist where the first visit MC estimator is superior and MRPs exist where the ML estimator is superior (see Sect. D.2, p. 328 for examples). How about TD(λ)? Again the relation is not clear cut. In the case that the MRP is acyclic and that Corollary 7 applies the first-visit MC estimator is at least as good as TD(λ). In general, however, no ordering exists (see Sect. D.1, p. 326 for examples).

3.6 Temporal difference learning

One would like to establish inequalities between the estimation error of TD and the error of other estimators like the MVU or the ML estimator. For the acyclic case TD(λ) is essentially unbiased and the MVU and the ML estimator are superior to TD. However, for the cyclic case the analysis is not straightforward, as TD(λ) is biased for $\lambda < 1$ and does not fulfill the Bellman equation. So TD is in a sense neither in the estimator class of the MVU nor of the ML estimator and conditioning with a sufficient statistics does not project TD to either of these estimators.

The bias of TD can be verified with the MRP from Fig. 2(A) with a discount of $\gamma = 1$ and with $n = 1$. If we take the TD(0) estimator with a learning rate of $\alpha_j = 1/j$ then the value estimate for state 0 is $i/(i+1)\sum_{j=1}^{i} 1/j$ if i cyclic transitions have been observed. The estimate should on average equal i to be unbiased. Yet, for $i > 0$ it is strictly smaller than i.

While our tools are not usable to establish inferiority of TD, we can still interpret the weaknesses of TD with it. In the following we focus on the TD(0) update rule.

3.6.1 Weighting of examples and conditioning

In the examples comparing TD(λ) and MC (Sect. D.1.1, p. 326) one observes that a weakness of TD(0) is that not all of the examples are weighted equally. In particular, (4) on p. 293 suggests that no observation should be preferred over another. Intuitively, conditioning suggests so too: For an acyclic MRP TD(0) can be written as $\bar{V}_i = \hat{p}_{ij}(R_{ij} + \gamma \bar{V}_j)$, whereas $\hat{p}_{ij}$ differs from the maximum likelihood parameter estimates $\bar{p}_{ij}$ due to the weighting. Generally, conditioning with a sufficient statistics permutes the order of the observations and resolves the weighting problem. Therefore, one would assume that conditioning with the element $\bar{p}_{ij}$ of the sufficient statistics changes $\bar{V}_i$ to $\bar{p}_{ij}(R_{ij} + \gamma \bar{V}_j)$. As conditioning improves the estimate, the new estimator would be superior to TD(0). However, conditioning with just a single element $\bar{p}_{ij}$ must not modify the estimator at all, as the original path might be reconstructed from the other observations. E.g. if one observes a transition $1 \to 2$ and $2 \to 3$, with $2 \to 3$ being the only path from state 2 to state 3, then it is enough to know that transition $1 \to 2$ occurred and state 3 was visited.

Despite these technical problems, the superiority of $\bar{p}_{ij}$ over $\hat{p}_{ij}$ and the weighting problem are reflected in the contraction properties of TD(0). Due to Sutton (1988) TD(0) contracts towards the ML solution. Yet, the contraction is slow compared to the case where each example is weighted equally.

3.6.2 Weighting of examples and contraction factor

We continue with another look at the familiar ML equation: $\bar{\mathbf{V}} = \bar{\mathbf{r}} + \gamma \bar{\mathbf{P}} \bar{\mathbf{V}} =: \bar{\mathbf{T}} \bar{\mathbf{V}}$. If the matrix $\bar{\mathbf{P}}$ is of full rank then the ML estimate is the sole fixed point of the Bellman operator $\bar{\mathbf{T}}$. The ML estimate can be gained by solving the equation, i.e. $\bar{\mathbf{V}} = (\mathbf{I} - \gamma \bar{\mathbf{P}})^{-1} \bar{\mathbf{r}}$. Alternatively, it is possible to make a fixed point iteration. I.e. starting with an initial guess $\bar{\mathbf{V}}^{(0)}$ and iterating the equation, i.e $\bar{\mathbf{V}}^{(n)} = \bar{\mathbf{T}} \bar{\mathbf{V}}^{(n-1)}$. Convergence to the ML solution is guaranteed by the *Banach Fixed Point Theorem*, because $\bar{\mathbf{T}}$ is a contraction. The contraction factor is upper bounded by $\gamma \|\bar{\mathbf{P}}\| \leq \gamma$, where $\|\cdot\|$ denotes in the following the *operator norm*. The bound can be improved by using better suited norms (e.g. Bertsekas and Tsitsiklis 1996). Hence, for n updates the distance to the ML solution is reduced by a factor of at least γ^n.

Applying the TD(0) update (5) to the complete value estimate $\bar{\mathbf{V}}$ using $\bar{\mathbf{P}}$ and a learning rate of $1/n$ results in

$$\bar{\mathbf{V}}^{(n)} = \bar{\mathbf{V}}^{(n-1)} + \frac{1}{n}\left(\mathbf{r} + \gamma\bar{\mathbf{P}}\bar{\mathbf{V}}^{(n-1)} - \bar{\mathbf{V}}^{(n-1)}\right) = \left(\frac{n-1}{n} + \frac{1}{n}\bar{\mathbf{T}}\right)\bar{\mathbf{V}}^{(n-1)}. \tag{37}$$

In this equation the weighting problem becomes apparent: The contraction $\bar{\mathbf{T}}$ affects only a part of the estimate. Yet, the operators $\bar{\mathbf{S}}^{(n)} := (\frac{n-1}{n} + \frac{1}{n}\bar{\mathbf{T}})$ are still contractions. For $\bar{\mathbf{V}}$ and $\bar{\mathbf{W}}$:

$$\|\bar{\mathbf{S}}^{(n)}\bar{\mathbf{V}} - \bar{\mathbf{S}}^{(n)}\bar{\mathbf{W}}\| \leq \frac{n-1}{n}\|\bar{\mathbf{V}} - \bar{\mathbf{W}}\| + \frac{1}{n}\|\bar{\mathbf{T}}\|\|\bar{\mathbf{V}} - \bar{\mathbf{W}}\| \leq \frac{n-1+\gamma}{n}\|\bar{\mathbf{V}} - \bar{\mathbf{W}}\|. \tag{38}$$

The contraction coefficient is therefore at least $\frac{n-1+\gamma}{n}$. The ML solution (in the following $\bar{\mathbf{V}}$) is a fixed point for the $\bar{\mathbf{S}}^{(i)}$ and for n iterations the distance is bounded by

$$\|\bar{\mathbf{S}}^{(n)}\ldots\bar{\mathbf{S}}^{(1)}\bar{\mathbf{V}}^{(0)} - \bar{\mathbf{V}}\| \leq \frac{\prod_{i=0}^{n-1}(i+\gamma)}{n!}\|\bar{\mathbf{V}}^{(0)} - \bar{\mathbf{V}}\|. \tag{39}$$

The smaller γ the faster the contraction. Yet, even in the limit the contraction is much slower than the contraction with the ML fixed point iteration, i.e. for γ tending to 0 the norm decreases at least with $1/n$ while for the ML fixed point iteration it decreases with γ^n. For $\gamma = 0.1$ and two applications of the Bellman operator the contraction is at least $\gamma^2 = 1/100$ and it needs 100 iterations with the TD(0) equation to reach the same distance.

TD(0) is applied only to the current state and not to the full value vector. The same can be done with the ML fixed point iteration, i.e. $\bar{V}_i = \bar{p}_{ij}(\bar{R}_{ij} + \gamma\bar{V}_j)$. We analyze the contraction properties of this estimator in the empirical part and we refer to the estimator as the iterative Maximum Likelihood (iML) estimator. The costs of the algorithm are slightly higher than the TD(0) costs: $O(|\mathbb{S}|)$ (time) and $O(|\mathbb{S}|^2)$ (space).

The restriction to the current path does not affect the convergence, i.e. the restricted iteration converges to the ML solution. Intuitively, the convergence is still guaranteed, as a contraction of γ is achieved by visiting each state once and because each state is visited infinitely often. Using that idea the following theorem can be proved:

Theorem 6 *iML is unbiased for acyclic MRPs, converges on average and almost surely to the true value.*

We use this algorithm only for the analysis and we therefore omit the proof.

3.7 Summary of theory results

We conclude the theory section with two tables that summarize central properties of estimators and established orderings. Footnotes are used to reference the corresponding theorems, corollaries or sections. We start with a table that summarizes the properties of the different estimators (Table 1). The row **Optimal** refers to the class of unbiased estimators and to convex loss functions. The statement that ML is unbiased if the Full Information Criterion is fulfilled and $\gamma = 1$ applies state-wise. I.e. for a cyclic MRP there will exist a state for which the ML estimator is biased. However, if the Full Information Criterion applies to a state, then the ML estimator for this particular state is unbiased. Finally, F-visit MC denotes the first-visit Monte-Carlo estimator.

Table 1 The table summarizes properties of different value estimators. The references for the statements are: (1) Theorem 4, p. 304. (2) (23), p. 302. (3) Theorem 2, p. 302. (4) Corollary 3, p. 305. (5) Corollary 4, p. 306. (6) Minorly modified TD estimator. Theorem 8, p. 322. (7) Corollary 7, p. 308. Counterexamples for $\gamma < 1$: Sect. 3.5.1, p. 308

Estimator	MVU	ML/LSTD	TD(λ)	(F-visit) MC
Convergence	L^1, a.s.$^{(1)}$	L^1, a.s.	L^1, a.s.	L^1, a.s.
Cost (Time)	exp?$^{(2)}$	$O(\|\mathbb{S}\|^2)$	$O(\|\mathbb{S}\|)$	$O(\|\mathbb{S}\|)$
Cost (Space)		$O(\|\mathbb{S}\|^2)$	$O(\|\mathbb{S}\|)$	$O(\|\mathbb{S}\|)$
Unbiased	$\surd^{(3)}$	Acyclic$^{(4)}$ or Cr. 3 and $\gamma = 1^{(5)}$	Acyclic$^{(6)}$	$\surd$
Bellman	Acyclic$^{(4)}$ or Cr. 3 and $\gamma = 1^{(5)}$	$\surd$		
Optimal	$\surd^{(3)}$	Acyclic$^{(4)}$ or Cr. 3 and $\gamma = 1^{(5)}$		Cr. 3 and $\gamma = 1^{(7)}$

Table 2 The table summarizes orderings between value estimators with respect to estimation quality. The references for the statements are: (1) Corollary 2, p. 305. (2) Counterexamples: Sect. D.2, p. 328. (3) Minorly modified TD estimator. Theorem 8, p. 322. (4) Theorem 2, p. 302. (5) Corollary 7, p. 308. (6) Theorem 5 in Singh and Sutton (1996). (7) Corollary 3, p. 305. (8) Counterexamples: Sect. D.1, p. 326

	ML/LSTD	TD(λ)	(F-visit) MC
MVU	ML unbiased: $=^{(1)}$ In general: $\neq^{(2)}$	Acyclic: $\leq^{(3,4)}$	Cr. 3 and $\gamma = 1$: $=^{(5)}$ In general: $\leq^{(4)}$
ML/LSTD		Acyclic: $\leq^{(3,4,7)}$	Cr. 3 and $\gamma = 1$: $=^{(6)}$ In general: $\neq^{(2)}$
TD(λ)			In general: $\neq^{(8)}$

Table 2 summarizes established orderings between value estimators. The legend is the following: $=$ means equivalent, $\neq$ means not comparable, $\leq$ means that the estimator in the corresponding row has a smaller risk (estimation error) than the estimator in the corresponding column. With **In general** we mean for $\neq$ that there exist MRPs where the row estimator is superior and MRPs where the column estimator is superior. However, for a subclass of MRPs, like acyclic MRPs, one of the estimators might be superior or they might be equivalent. For $\leq$ **in general** means that the row estimator is always as good as the column estimator, however, both might be equivalent on a subclass of MRPs.

4 Comparison of estimators: experiments

In this section we make an empirical comparison of the estimators. We start with a comparison using acyclic MRPs. For this case the ML estimator equals the MVU and the MVU solution can be computed efficiently. This allows us to make a reasonable comparison of the MVU/ML estimator with other estimators. In a second set of experiments we compare the MVU with the ML estimator using a very simple cyclic MRP. In a final set of experiments we compare the contraction properties of iML and TD(0).

4.1 Acyclic MRPs

We performed three experiments with acyclic MRPs for analyzing the estimators. In the first experiment we measured the MSE w.r.t. the number of observed paths. In the second experiment we analyzed how the MRP structure affects the estimation performance. As we can see from Corollary 1 the performance difference between "MDP" based estimators such as TD or ML and model-free estimators like MC depends on the ratio between the number of sequences hitting a 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. The third experiment was constructed to analyze the practical usefulness of the different estimators. We measured the MSE in relation to the calculation time.

Basic experimental setup We generated randomly acyclic MRPs for the experiments. The generation process was the following: We started by defining 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 fixed number of states in each layer. Connections were only allowed 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 were all states in the first layers (typically the first 4). Other layers which were further apart from s were omitted as paths starting in these contribute insignificantly to the estimate, but consume computation time. The distribution over the starting states was chosen to be uniform. Finally, we randomly defined rewards 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 ML, iML, MC and TD estimators. For the first two experiments the simulations were repeated 300000 times for each parameter setting. We split these runs into 30 blocks with 10000 examples each and calculated the mean and standard deviation for these. In the third experiment we only calculated the mean using 10000 examples. We used the modified TD(0) version which is unbiased with a learning rate of $1/i$ for each state, where i is the iteration index of the corresponding state. The ML solution was computed at the end and not at each run. This means no intermediate estimates were available, which can be a drawback. We also calculated the standard TD(0) estimates. The difference to the modified TD(0) version is marginal and therefore we did not include the results in the plots.

4.1.1 Experiment 1: MSE in relation to the number of observed paths

In the first experiment, we analyzed the effect of the number of observed paths given a fixed rate of $p_s = 0.2$ 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). Apart from ML, all three estimators perform quite similarly with a small advantage for iML and MC (Fig. 3). ML is even for few paths strongly superior and the estimate is already good for 10 paths. Note that, due to the scale the improvement of ML is hard to observe.

4.1.2 Experiment 2: MSE in relation to the starting probability

In the second experiment we tested how strongly 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

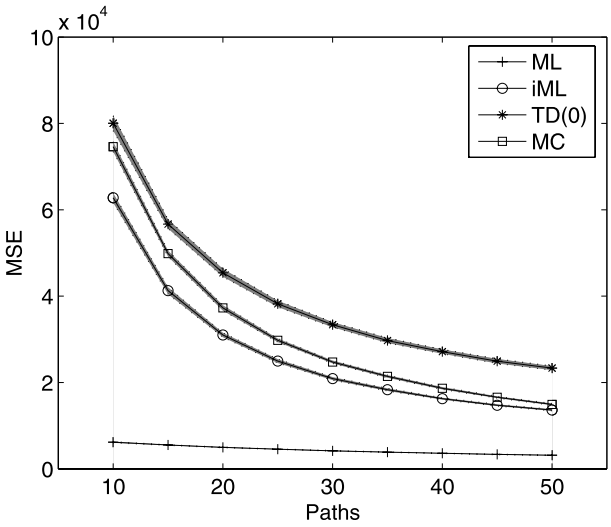

Fig. 3 MSE of ML, iML, TD(0) and MC in relation to the number of observed paths. The state space consisted of 10 layers with 20 states per layer

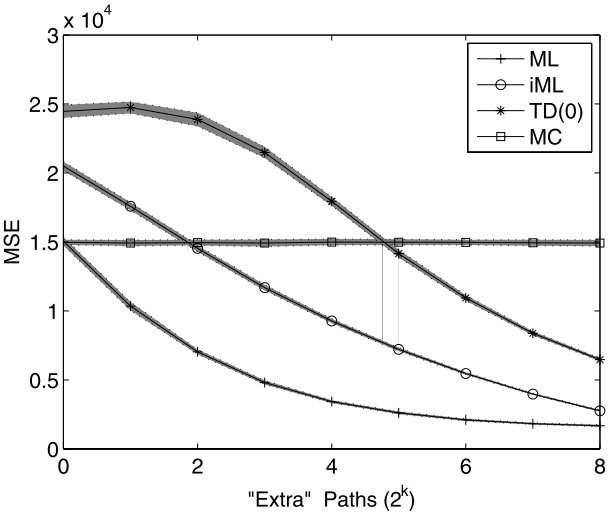

Fig. 4 MSE of ML, iML, TD(0) and MC in relation to the starting probability of the estimated state. The state space consisted of 10 layers with 20 states per layer

subgraph. The paths which start in the subgraph can only improve the estimation quality of state s if the Markov structure is used. Figure 4 shows the results of 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. The exponential factor is printed on the x-axis. $x = 0$ is equivalent to always start in s. One can see that the MC and ML estimator are equivalent if in each run the path starts in s. Furthermore, for this case MC outperforms TD due to the weighting problem of TD (Sect. 3.6.2). Finally, TD, iML and ML make a strong use of paths which does not visit state s itself. Therefore, TD becomes superior to MC for a higher number of paths. The initial plateau for the TD estimator appeared in the modified and the standard version. We assume that it is an effect of the one step error propagation of TD(0). For the one step error propagation a path starting in a state s' in the

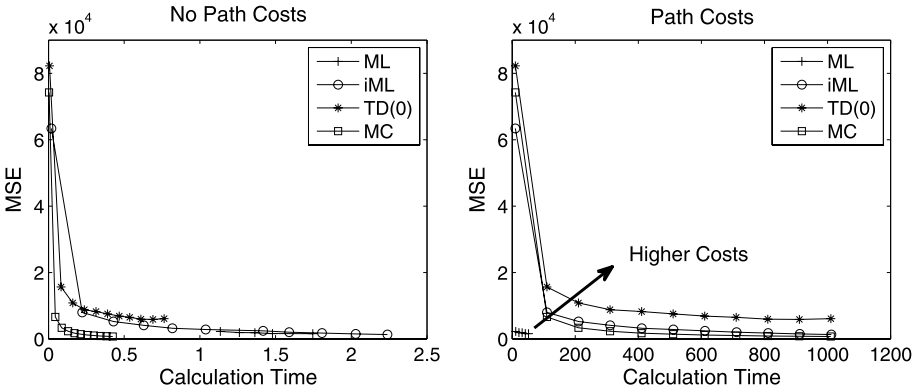

Fig. 5 MSE in relation to the computation time of the ML, iML, TD(0) and MC estimator. The *left* plot shows pure computation time (we excluded computation time for MRP calculations like state changes). In the *right* plot, an extra factor for each observed path is included (one second per path). The state space consisted of 10 layers with 20 states per layer. We tracked for a given number of paths (ML: 10–50, iML, TD(0), MC: 10–1000) the MSE and the computation time. The plot was constructed with the mean values for every number of paths

ith layer can only improve the estimate if i paths are observed that span the gap between s and s'. The probability of such an event is initially very small but increases with more paths.

4.1.3 Experiment 3: MSE in relation to calculation time

In many practical cases the convergence speed per sample is not the important measure. It is the time which is needed to produce a good estimate. This time depends on how many samples an estimator needs, the computation time to calculate the estimate from the sample and the time needed to generate the sample. We constructed an experiment to evaluate this relation (Fig. 5). We first tested which estimator is superior if only the pure estimator computation time is regarded (left part). For this specific MRP the MC estimator converges fastest with respect to time. The rate for starts in state s was 0.2, which is an advantage for MC. The ratio will typically be much lower. The other three estimators seem to be more or less equivalent. In the second plot a constant cost of 1 was introduced for generating a path. Through this the pure computation time becomes less important while the needed number of paths for reaching a specific MSE level becomes relevant. As ML needs only very few paths, it becomes superior to the other estimators. Further, iML catches up on MC. For higher costs the estimators will be drawn further apart from ML (indicated by the arrow). The simulations suggest that MC or TD (dependent on the MRP) are a good choice if the path costs are low. For higher costs ML and iML are alternatives.

4.2 Cyclic MRPs: MVU—ML comparison

Calculating the MVU is infeasible without some algebraic rearrangements. Yet, the algebraic rearrangements get tricky, even for simple MRPs. We therefore restrict the comparison of the MVU and the ML estimator to the simplest possible cyclic MRP, i.e. the MRP from Fig. 2(A). The MC and ML value estimates are

$$\frac{1}{n}\sum_{u=1}^{n}\sum_{j=0}^{i_u-1}\gamma^j = \frac{1}{1-\gamma} - \frac{1}{n}\sum_{u=1}^{n}\frac{\gamma^{i_u}}{1-\gamma} \tag{40}$$

and

$$(1-\bar{p})\sum_{i=0}^{\infty}\frac{1-\gamma^i}{1-\gamma}\bar{p}^i=\frac{1}{1-\gamma}\left(1-\frac{1-\bar{p}}{1-\gamma\,\bar{p}}\right)=\frac{\bar{p}}{1-\gamma\,\bar{p}}, \tag{41}$$

where i_u denotes the number of times the cycle has been taken in run u. The MVU sums the MC estimates over all consistent sets, i.e. over all vectors $(k_1,\ldots,k_n)$ which fulfill $\sum_{u=1}^{n}k_u=s:=\sum_{u=1}^{n}i_u$. Let the normalization $\mathcal{N}$ being the size of this set and $\mathrm{MC}(k_i)$ being the MC estimate for k_i cycles. The MVU is given by

$$\begin{aligned}&\frac{1}{n\mathcal{N}}\sum_{|(k_1,\ldots,k_n)|=s}\mathrm{MC}(k_1)+\cdots+\mathrm{MC}(k_n)\\&\quad=\frac{1}{n\mathcal{N}}\sum_{k_1=0}^{s}\cdots\sum_{k_{n-1}=0}^{s-k_1\ldots-k_{n-2}}\mathrm{MC}(k_1)+\cdots+\mathrm{MC}(k_n),\end{aligned} \tag{42}$$

where in the second line $k_n=s-k_1\ldots-k_{n-1}$. The number of times k_u takes a value j is independent of u, i.e. $\mathrm{MC}(k_1)$ appears equally often as $\mathrm{MC}(k_i)$ if $k_1=k_i$. Hence, it is enough to consider $\mathrm{MC}(k_1)$ and the MVU is

$$\frac{1}{\mathcal{N}}\sum_{k_1=0}^{s}\mathrm{MC}(k_1)\sum_{k_2=0}^{s-k_1}\cdots\sum_{k_{n-1}=0}^{s-k_1\ldots-k_{n-2}}1=:\frac{1}{\mathcal{N}}\sum_{k_1=0}^{s}MC(k_1)\mathcal{C}(k_1). \tag{43}$$

Finally, the coefficient is $\mathcal{C}(k_1)=\binom{s+n-2-k_1}{n-2}$ and the normalization is $\mathcal{N}=\binom{s+n-1}{n-1}$. The derivation can be done in the following way. First, observe that $1=\binom{k_n}{0}$. Then, that $\sum_{k_{n-1}=0}^{s-k_1\ldots-k_{n-2}}\binom{k_n}{0}=\binom{1+(s-k_1\ldots-k_{n-2})}{1}$ (e.g. rule 9 in Aigner 2006, p. 13). And finally that $\sum_{k_{n-2}=0}^{s-k_1\ldots-k_{n-3}}\binom{1+(s-k_1\ldots-k_{n-2})}{1}=\sum_{k_{n-2}=0}^{s-k_1\ldots-k_{n-3}}\binom{1+k_{n-2}}{1}$. Iterating the steps leads to the normalization and the coefficients. Alternatively, these coefficients can be derived by assuming that you have s indistinguishable objects (the cycles) that you want to place into n boxes (the runs). In summary the MVU is

$$\frac{1}{1-\gamma}-\frac{1}{(1-\gamma)\binom{s+n-1}{n-1}}\sum_{i=0}^{s}\binom{s+n-2-i}{n-2}\gamma^i. \tag{44}$$

We compared the MVU to the ML estimator in two experiments. The results are shown in Fig. 6. One can observe in Fig. 6(A) that high probabilities for cycles are beneficial for ML and that the discount which is most beneficial to ML depends on the probability for the cycle. We have seen in Sect. 3.4.4 that the Bellman equation enforces the estimator to use all cycle times from 0 to "∞" and thus in a sense "overestimates" the effect of the cycle. Furthermore, the probability for the cycle is underestimated by ML, i.e. $\mathbb{E}[\bar{p}]<p$ (Sect. 3.4.4), which can be seen as a correction for the "overestimate". The parameter estimate is independent of the true probability and the discount. Therefore, a parameter must exist which is most beneficial for ML, i.e. ML is biased towards this parameter. The experiment suggests that the most beneficial parameter p is close to 1, meaning that ML is biased towards systems with high probabilities for cycles.

In Fig. 6(B) the results of the second experiment are shown. In this experiment $\gamma=p=0.9$ and the number of paths is varied. One can observe that the difference between the ML and the MVU estimator is marginal in comparison to the difference to the MC estimator.

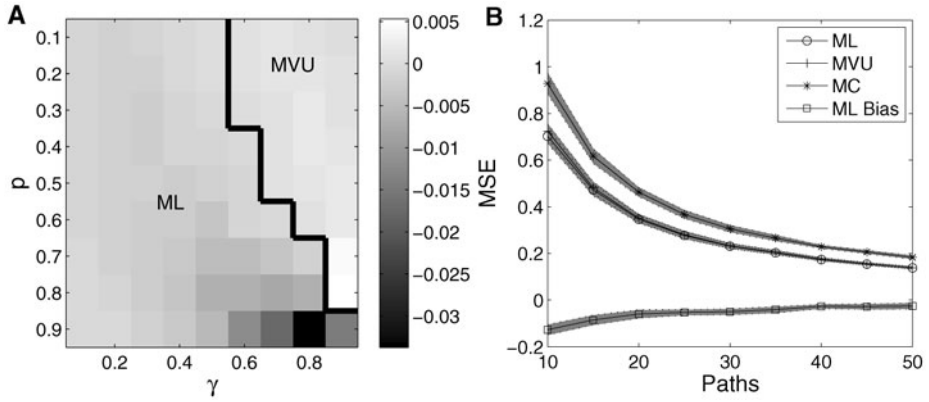

Fig. 6 **A** The plot shows the difference in MSE between the ML estimator and the MVU (MSE(ML)-MSE(MVU)) for 10 paths and different values of γ and p. In the *top right* part the MVU is superior and in the remaining part the ML estimator. **B** The plot shows the MSE of the ML, the MVU and the MC estimator and the bias of ML with respect to the number of paths for $p = \gamma = 0.9$. 30000 samples were used for the mean and the standard deviation (30 blocks with 1000 examples)

Furthermore, the bias of ML approaches quickly to 0 and the MVU and the ML estimator become even more similar.

4.3 Contraction: ML, iML and TD(0)

In a final set of experiments we compared the contraction factor of different operators. We generated randomly transition matrices for a state space size of 100 and applied the different operators. The results are shown in Fig. 7. The left plot shows the results for the usual Bellman operator and the bound for different discount values. In the middle the TD(0) update equation is used and in the right plot the Bellman operator is applied state wise, whereas the state is chosen randomly from different priors. The prior probabilities for states $1, \ldots, n := |\mathbb{S}|$ are given by: $p_1 = (1 - c)m$, $p_i = (1 - c + 2c(i - 1)/(n - 1))m, \ldots, p_n = (1 + c)m$, where $m = 1/n$ (mean) and c denotes the deviation from the uniform prior. If $c = 0$ then we have a uniform distribution. If $c = 0.1$ then $p_1 = 0.9m$, $p_2 = (0.9 + 0.2/(n - 1))m, \ldots, p_n = 1.1m$.

While we were not able to prove that TD is in general inferior to ML or to iML the plots suggest this to be the case for typical MRPs. Especially, the contraction of TD (middle plot) to the ML solution is magnitudes slower than the contraction using the Bellman operator. The state-wise update reduces the contraction speed further. The right plot shows the difference between the fixed point iteration and the state-wise update with the Bellman operator (corresponding to iML). The contraction factor of the state-wise update depends crucially on the distribution of visits of the different states. At best (i.e. uniform distribution over the states) the contraction is about $|\mathbb{S}|$-times slower than the contraction with the Bellman operator applied to the full state vector, i.e. $|\mathbb{S}|$ many update steps of single states have roughly the same effect as one full update.

5 Summary

In this work we derived the MVU and compared it to different value estimators. In particular, we analyzed the relation between the MVU and the ML estimator. It turned out that the

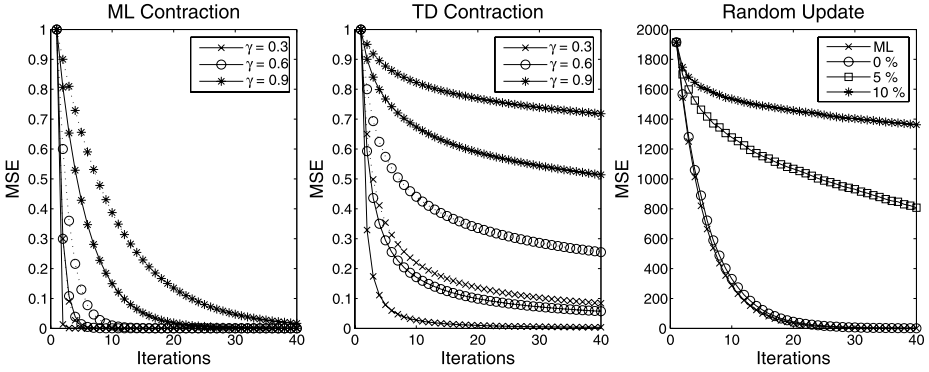

Fig. 7 The three plots show the contraction rate of different operators to the ML solution. The x-axis denotes the number of applications of the operators and the y-axis shows the distance to the ML solution. The *left* and the *center* plot are normed with the initial distance (before the first application). *Left*: The Bellman operator is used. The discount γ varies from 0.3 to 0.9. For each discount value the empirical distance and the bound (*dotted line*) is plotted. *Center*: Same setting as in the left plot but with the "TD(0)" operator. *Right*: In this plot $\gamma = 0.9$. The ML curve corresponds again to the Bellman operator. For the other three curves only single states are updated with the Bellman operator, whereas the states which are updated are chosen randomly due to the prior probabilities from Sect. 4.3. The percent values denote the deviation from the uniform prior for the states (0% means uniform). $|\mathbb{S}|$ updates were performed per iteration instead of just one, i.e. $|\mathbb{S}|$ many states were drawn from the prior probabilities and the values of these states were updated sequentially in the order in which they were drawn

relation between these estimators is directly linked to the relation between the class of unbiased estimators and Bellman estimators. If the ML estimator is unbiased then it is equivalent to the MVU and more generally the difference between the estimators depends on the bias of ML. This relation is interesting, in particular as the estimators are based on two very different algorithms and proving equivalence using combinatorial arguments is a challenging task. Furthermore, we demonstrated in this paper that the MC estimator is equivalent to the MVU in the undiscounted case if both estimators have the same amount of information. The relation to TD is harder to characterize. TD is essentially unbiased in the acyclic case and therefore inferior to the MVU and the ML estimator in this case. In the cyclic case TD is biased and our tools are not applicable.

We want to conclude the section with open problems. Possibly, the most interesting problem is the derivation of an efficient MVU algorithm. The combinatorial problems that must be solved appear to be formidable. Therefore, it is astonishing that in the undiscounted case the calculation essentially boils down to calculating the ML estimate. In particular, the exponential runtime of a brute force MVU algorithm which is intractable even for simple MRPs decreases in this case to an $\mathrm{O}(n^3)$ factor. This efficiency is mainly due to the irrelevance of the time at which a reward is observed. In the discounted case the time of an observation matters and the algorithmical difficulties increase considerably. Instead of the full geometric series of ML with arbitrary long paths it seems to be needed to make a cut-off at a maximum number of cycles, i.e. replacing $(\mathbf{I}-\gamma\bar{\mathbf{P}})^{-1}$ with something like $(\mathbf{I}-(\gamma\bar{\mathbf{P}})^{s})(\mathbf{I}-\gamma\bar{\mathbf{P}})^{-1}$, where s is the maximum number of cycles. Yet, (44) shows that a weighting factor is associated with each time step and the MVU equation is not that simple.

Another interesting question concerns the bias of the ML estimator. We showed that the normalizations $\{N_i \geq 1\}$ are the reason for the bias. Furthermore, if the Full Information Criterion applies then the normalization problem is not present and we used a theorem from Singh and Sutton (1996) to deduce unbiasedness of ML for this case. Yet, there seems to be

a deeper reason for the unbiasedness of the ML estimator and the theorem from Singh and Sutton (1996) appears to be an implication from this and from Corollary 1 (MVU = MC).

5.1 Discussion

In the discussion section we address two questions: (1) What is the convergence speed of the MVU? (2) Which estimator is to be preferred in which setting? In this section the emphasis is put onto gaining intuition and not on mathematical rigor.

Convergence Speed We are interested in the MSE and in the small deviation probability of the MVU. First, let us state the variance and the Bernstein inequality (e.g. Lugosi 2006, Theorem 3, p. 219) for the first-visit MC estimator with n paths available for estimation:

$$\mathbb{V}[\bar{V}^{(n)}] = \mathrm{MSE}[\bar{V}^{(n)}] = \frac{1}{n}\mathbb{V}[R] \tag{45}$$

and

$$\mathbb{P}(|\bar{V}^{(n)} - V| \geq \epsilon) \leq 2\exp\left(-\frac{\epsilon^2 n}{2\mathbb{V}[R] + 2d\epsilon/3}\right), \tag{46}$$

where $\mathbb{V}[R]$ is the variance in the cumulative reward (see Sobel 1982 for the variance of an MRP) and d is an upper bound for the difference between the cumulative reward of any path and the mean V, i.e. $|\sum_{t=0}^{\infty} \gamma^t R_t - V| < d$.

How about the MVU? In the undiscounted case the MVU has the same variance and small deviation probability if the Full Information Criterion applies. The quality increases with further paths into the graph of successor states. Intuitively, the improvement in quality depends on the "distance" of the entry point s' in the successor state graph to the state s of which we want to estimate the value. For example, if the shortest path from s to s' visits 10 intermediate states then the information gained at s' will in general not improve the estimate at state s too much. Opposite to that, if s' is a direct successor of state s then paths that enter the successor states graph at state s' will improve the estimate considerably. A natural distance measure for this setting is the probability to move from state s to s'. Furthermore, the improvement will depend on the variation in the cumulative reward of paths starting in s'. Paths, that run through regions in which the reward has high variance will yield a better performance increase than paths which run through near deterministic regions. The performance will, however, be lower bounded by the case that all of the paths start directly in s. Therefore, for undiscounted MRPs the rough lower bound $(1/N)\mathbb{V}[R]$ will hold:

$$\frac{1}{N}\mathbb{V}[R] \leq \mathrm{MSE}[\mathbb{E}[\bar{V}^{(n)}|\mathcal{S}]] \leq \frac{1}{n}\mathbb{V}[R], \tag{47}$$

where n denotes here the number of paths that visit state s and N denotes the number of paths that visits the graph of successor states of s (including the paths that visit state s). If starts in the successor graph are c times more often than starts in s, i.e. $N = (c+1)n$ then

$$\frac{1}{c+1}\mathrm{MSE}[\bar{V}^{(n)}] \approx \mathrm{MSE}[\mathbb{E}[\bar{V}^{(n)}|\mathcal{S}]]. \tag{48}$$

Similarly, a "reasonable" Bernstein bound of the small deviation probability will lie between

$$2\exp\left(-\frac{(c+1)\epsilon^2 n}{2\mathbb{V}[R] + 2d\epsilon/3}\right) \quad \text{and} \quad 2\exp\left(-\frac{\epsilon^2 n}{2\mathbb{V}[R] + 2d\epsilon/3}\right). \tag{49}$$

Choosing an estimator Our study shows that we have essentially a tradeoff between computation time and convergence speed per sample. As one would expect, the methods which converge faster have a higher computation time. It seems that the fast methods with bad convergence speed are superior if we consider pure computation time (Experiment 3, Sect. 4.1.3). However, if there are costs involved for producing examples, then the computationally expensive methods become competitive. In a high cost scenario it currently seems best to choose the ML/LSTD estimator. The MVU might become an alternative, but an efficient algorithm is currently missing. Furthermore, the algorithmic problems restricted the numerical comparison to ML and it is unclear in which setting which estimator is superior.

Acknowledgements We would like to thank Sepp Hochreiter for invaluable help in the beginning of this study. Furthermore, we would like to thank the anonymous reviewers for thorough reading and for helpful suggestions. In particular, the alternative derivation of the MVU equation of Sect. 4.2 and the link between the ML and LSTD value estimator were found by reviewers. This work was funded by BMBF (grant no. 01GQ1001B) and through the ARAGORN project from the European Union.

Appendix A: Notation

$\mu_{ss'}$	The number of direct transitions from state s to s'
π	Path
π_i	ith state in the path
$\Pi_{ss'}$	Set of all paths from state s to s'
$\mathbf{\Pi}(\mathcal{S})$	Set of paths that are consistent with $\mathcal{S}$
$\mathbb{P}, \mathbb{E}, \mathbb{V}$	Probability measure, expectation and variance
H_s	Sum of the reward received through transitions from state s
K_s	Number of visits of state s
$\bar{p}_{ss'}$	Estimate of the probability for a direct transition from state s to s'
$\bar{P}_{ss'}$	Estimate of the transition probability from state s to s'
$\bar{R}_s$	Estimator of the reward received through transitions from state s
$\mathbb{S}$	State space
$\mathcal{S}, \mathcal{T}$	Sufficient Statistics
V_s	True value of state s
$\bar{V}_s$	Estimated value of state s. Value estimator varies with the section
$\bar{V}_s^{(i)}, \bar{P}_{ss'}^{(i)}, \ldots$	Superscripts denote values after the ith run
$\mathbf{V}, \mathbf{P}, \ldots$	Vectors and Matrices

Appendix B: Unbiased TD(λ)

In this section we introduce a (minorly) modified TD(λ) estimator. The estimates are, in contrast to the standard TD(λ) estimator, independent of the initialization. In the acyclic case this is already enough to guarantee unbiasedness of TD(λ). We first discuss the TD(0) case. This case contains the major arguments in an accessible form.

B.1 TD(0)

We first restate the TD(0) equation through unfolding the recursive definition ((5), p. 294).

Lemma 3 *If the TD*(0) *estimator is initialized with* 0 *then for an acyclic MRP it equals*

$$\bar{V}_s^{(n)} = \sum_{i=1}^{n} \beta_i R^{(i)} + \sum_{s' \in \mathbb{S}} \left(\sum_{i=1}^{n} T_{i,s,s'} \beta_i \gamma \bar{V}_{s'}^{(j_i)} \right), \tag{50}$$

where $\beta_i := (\alpha_i \prod_{j=i+1}^{n} (1-\alpha_j))$, $R^{(i)}$ *is the received reward in path* i *and* $T_{i,s,s'}$ *a random variable which is one if in run* i *the state* s' *followed immediately upon state* s *and is zero otherwise.*

Proof The recursive TD(0) definition (5) can be written as: $\bar{V}_s^{(n)} = \bar{V}_s^{(n-1)}(1-\alpha_n) + \alpha_n(R^{(n)} + \gamma \bar{V}_{s'}^{(j_n)})$. Substituting $\bar{V}_s^{(n-1)}$ and denoting the i-th level successor of state s with $s^{(i-1)}$:

$$\begin{aligned}
\bar{V}_s^{(n)} &= \left(\gamma \bar{V}_s^{(n-2)}(1-\alpha_{n-1}) + \alpha_{n-1}(R^{(n-1)} + \gamma \bar{V}_{s''}^{(j_{n-1})})\right)(1-\alpha_n) + \alpha_n(R^{(n)} + \gamma \bar{V}_{s'}^{(j_n)}) \\
&= \gamma \bar{V}_s^{(n-2)}(1-\alpha_{n-1})(1-\alpha_n) + \alpha_{n-1}(1-\alpha_n)(R^{(n-1)} + \gamma \bar{V}_{s''}^{(j_{n-1})}) + \alpha_n(R^{(n)} + \gamma \bar{V}_{s'}^{(j_n)}) \\
&= \cdots = \sum_{i=1}^{n} \left(\alpha_i \cdot \prod_{j=i+1}^{n} (1-\alpha_j) \right) \left(R^{(i)} + \gamma \bar{V}_{s^{(i)}}^{(j_i)} \right) =: \sum_{i=1}^{n} \beta_i \left(R^{(i)} + \gamma \bar{V}_{s^{(i)}}^{(j_i)} \right).
\end{aligned} \tag{51}$$

□

The estimate contains values $\bar{V}_{s'}^{(0)}$ if state s' has not been visited before. This biases the estimator towards the initialization. The estimator can be made unbiased for acyclic MRPs by excluding these values and by guarantying that the β_i terms sum to one. Modification 1 does exactly this.

Modification 1 Modified TD(0) for acyclic MRPs

Let s be the current state, s' the successor state of s in the i-th path and $\bar{V}_s^{(i)}$ the value estimate for state s in path i.
if s is a terminal state **then**
  Set $\bar{V}_s^{(i)} = 0$.
else
  if $\bar{V}_s^{(i)}$ has never been updated **then**
    set the learning rate for this step to 1.
  end if
  if $\bar{V}_{s'}^{(i)}$ has never been updated **then**
    first update the estimate $\bar{V}_{s'}^{(i)}$
  end if
  Use the TD(0) update rule (5).
end if

Setting the learning rate for the first example to 1 eliminates the initialization of $\bar{V}_s$. The second rule assures that the initialization of the estimators of the successor states is eliminated. Setting the learning rate α_1 to 1 has also the effect that the weighting factors β_i sum to one, independent of the learning rate. For example for $n = 3$, we have $\sum_{i=1}^{3} \beta_i = 1(1-\alpha_2)(1-\alpha_3) + \alpha_2(1-\alpha_3) + \alpha_3 = 1$.

Theorem 7 *The modified TD*(0) *estimator is unbiased if the MRP is acyclic*.

Proof We prove this by induction. Order the states such that for each state s the successor states s' of s come before s. Go from the beginning through this ordered set.
Induction Basis (first state in the ordered set): The first state is a terminal state. For a terminal state s the value estimator is unbiased as $\bar{V}_s = 0 = V_s$ holds.
Induction Step (next state in the ordered set): Due to the induction we know that the modified TD(0) estimator is unbiased for any successor state of the current state s as these come earlier in the ordering. The expected estimate has the form (Lemma 3):

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^{n}\beta_i R^{(i)} + \sum_{s'\in\mathbb{S}}\Big(\sum_{i=1}^{n} T_{i,s,s'}\beta_i\gamma\,\bar{V}_{s'}^{(j_i)}\Big)\Big|K_s = n\right] \\
&\quad= \mathbb{E}\left[\sum_{i=1}^{n}\beta_i R^{(i)}\Big|K_s = n\right] + \sum_{s'\in\mathbb{S}}\sum_{i=1}^{n}\beta_i\gamma\,\mathbb{E}\big[T_{i,s,s'}\bar{V}_{s'}^{(j_i)}\big|K_s = n\big] \\
&\quad= \mathbb{E}\left[\sum_{i=1}^{n}\beta_i R^{(i)}\Big|K_s = n\right] + \sum_{s'\in\mathbb{S}} p_{ss'}\sum_{i=1}^{n}\beta_i\gamma\,\mathbb{E}\big[\bar{V}_{s'}^{(j_i)}\big|K_s = n\big]. \qquad (52)
\end{aligned}$$

It remains to show that $\mathbb{E}[\bar{V}_{s'}^{(j_i)}|K_s = n]$ is unbiased. For $j_i \geq 2$ this follows from the induction hypothesis. For the case $j_i = 1$ Modification 1 guarantees that the estimator has at least one example for estimation and is unbiased due to the induction hypothesis. Furthermore, the β_i's sum to one due to the modification and $\sum_{s'\in\mathbb{S}} p_{ss'}\gamma V_{s'}\sum_{i=1}^{n}\beta_i = \sum_{s'\in\mathbb{S}} p_{ss'}\gamma V_{s'}$. □

B.2 TD(λ)

The TD(λ) case is essentially the same. The main difference is that the estimates of all states of a path are used. Therefore, it is not enough that the estimators of the direct successor states are set to "reasonable" values, but all states of the path must be:

Modification 2 Modified TD(λ) for acyclic MRPs

Let s be the current state, s' the successor state of s in the i-th path and $\bar{V}_s^{(i)}$ the value estimate for state s in path i.
if s is a terminal state **then**
  Set $\bar{V}_s^{(i)} = 0$.
else
  if $\bar{V}_s^{(i)}$ has never been updated **then**
    set the learning rate for this step to 1.
  end if
  if for any successor (not necessarily a direct successor) s' of s in the i-th path $\bar{V}_{s'}^{(i)}$ has never been updated **then**
    first update the estimate $\bar{V}_{s'}^{(i)}$
  end if
  Use the TD(λ) update rule.
end if

Theorem 8 *The modified TD(λ) estimator is unbiased if the MRP is acyclic.*

Proof Proof by induction. **Induction Hypothesis:** $\mathbb{E}[\bar{V}_s|K_s=n]=V_s$.
Induction Basis: For terminal states the Hypothesis trivially holds.
Induction Step: Let $\pi(i+1,n)$ be state i-th successor state of state s in path n, $\pi(1,n)=s$ and let $R_{\pi(i+1,n)}$ be the reward received at the transition from the i-th successor state of state s in path n. It is more convenient for the proof to enumerate the value estimates of state i with the number of paths that have been observed. Thus we use $\bar{V}_s^n$ to denote the value estimate of state s after observing n path and after applying the TD(λ) update rule at each transition. In the acyclic case TD(λ) can be written as

$$\begin{aligned}\bar{V}_s^n &= (1-\alpha_n)\bar{V}_s^{n-1}+\alpha_n\left(\sum_{i=1}(\gamma\lambda)^{i-1}R_{\pi(i,n)}+\gamma(1-\lambda)\sum_{i=2}(\gamma\lambda)^{i-2}\bar{V}_{\pi(i,n)}\right)\\ &= (1-\alpha_1)\bar{V}_s^0+\sum_{j=1}^{n}\beta_j\left(\sum_{i=1}(\gamma\lambda)^{i-1}R_{\pi(i,j)}+\gamma(1-\lambda)\sum_{i=2}(\gamma\lambda)^{i-2}\bar{V}_{\pi(i,j)}\right). \quad (53)\end{aligned}$$

We suppressed the "iteration" index of $\bar{V}_{\pi(i,j)}$ for readability. Like in the TD(0) case $\beta_j := (\alpha_j\prod_{k=j+1}^{n}(1-\alpha_k))$. Applying the expectation operator and using $\alpha_1=1$, we get

$$\begin{aligned}\mathbb{E}[\bar{V}_s^n|K_s=n] &= \mathbb{E}\left[\sum_{j=1}^{n}\beta_j\left(\sum_{i=1}(\gamma\lambda)^{i-1}R_{\pi(i,j)}+\gamma(1-\lambda)\sum_{i=2}(\gamma\lambda)^{i-2}\bar{V}_{\pi(i,j)}\right)\middle| K_s=n\right]\\ &= \sum_{j=1}^{n}\beta_j\left(\sum_{i=1}(\gamma\lambda)^{i-1}\mathbb{E}[R_{\pi(i,j)}|K_s=n]\right.\\ &\quad \left.+\gamma(1-\lambda)\sum_{i=2}(\gamma\lambda)^{i-2}\mathbb{E}[\bar{V}_{\pi(i,j)}|K_s=n]\right). \quad (54)\end{aligned}$$

Instead of $\mathbb{E}[R_{\pi(i,j)}]$ and $\mathbb{E}[\bar{V}_{\pi(i,j)}]$ we use $\mathbb{E}[R_i]$ and $\mathbb{E}[\bar{V}_i]$ in the following to denote the expected reward in step i, respectively the expected value estimate in step i (expected state times expected value estimate for that state). Due to the induction hypothesis

$$\mathbb{E}[\bar{V}_i|K_s=n]=V_i=\sum_{j=i}^{\infty}\gamma^{j-i}\mathbb{E}[R_j]. \quad (55)$$

Substituting this term into (54):

$$\sum_{j=1}^{n}\beta_j\left(\sum_{i=1}(\gamma\lambda)^{i-1}\mathbb{E}[R_i]+\gamma(1-\lambda)\sum_{i=2}(\gamma\lambda)^{i-2}\sum_{u=i}\gamma^{u-i}\mathbb{E}[R_u]\right). \quad (56)$$

Taking a specific $\mathbb{E}[R_i]$, we see that for the coefficient

$$\begin{aligned}&(\gamma\lambda)^{i-1}+\gamma(1-\lambda)(\gamma^{i-2}\lambda^{i-2}+\ldots+\gamma^{i-2}\lambda+\gamma^{i-2})\\ &= \gamma^{i-1}\left(\lambda^{i-1}+(1-\lambda)\sum_{j=0}^{i-2}\lambda^j\right)=\gamma^{i-1}\left(\lambda^{i-1}+(1-\lambda)\frac{1-\lambda^{i-1}}{1-\lambda}\right)=\gamma^{i-1} \quad (57)\end{aligned}$$

holds. We know already that the β_j sum to one. Hence, the modified TD(λ) is unbiased. □

Appendix C: Proofs

C.1 Unbiased estimators—Bellman equation

Lemma 1 *For the MRP from Fig. 2(B) there exists no parameter estimator $\bar{p}$ such that $V_i(\bar{p})$ is unbiased for all states i.*

Proof Assume that $\bar{V}_1, \bar{V}_2$ are unbiased, i.e. $\mathbb{E}[\bar{V}_1|\{N_1 \geq 1\}] = \mathbb{E}[\bar{V}_2] = V_1 = V_2$ and the estimator fulfills the Bellman equation, i.e. $\bar{V}_1 = \bar{V}_2$ on $N_1 := \{N_1 \geq 1\}$. Then

$$\mathbb{E}[\bar{V}_2|N_1] \overset{\text{Bellm.}}{=} \mathbb{E}[\bar{V}_1|N_1] \overset{\text{unb.}}{=} V_1 \overset{\text{Bellm.}}{=} V_2 \overset{\text{unb.}}{=} \mathbb{E}[\bar{V}_2]. \tag{58}$$

We used for the first equality that $\bar{p}_{12}$ must equal 1 as there is only one connection leading away from state 1. The derived equality shows that the average value estimate $\bar{V}_2$ must be the same as the average estimate for the case that the connection $2 \to 1$ has been taken at least once. This implies that the value estimate for the case that only the connection $2 \to 3$ has been taken must be the same as the average value estimate:

$$\mathbb{E}[\bar{V}_2] = \mathbb{E}[\bar{V}_2|N_1]\mathbb{P}[N_1] + \mathbb{E}[\bar{V}_2|N_1^c]\mathbb{P}[N_1^c] \overset{\text{unb.}}{=} \mathbb{E}[\bar{V}_2]\mathbb{P}[N_1] + \mathbb{E}[\bar{V}_2|N_1^c]\mathbb{P}[N_1^c], \tag{59}$$

where N_1^c denotes the event $N_1 = 0$. This implies that $\mathbb{E}[\bar{V}_2|N_1] = \mathbb{E}[\bar{V}_2] = \mathbb{E}[\bar{V}_2|N_1^c]$.

There are two possibilities to achieve this equality: (1) All three terms are 0. In particular, $\mathbb{E}[\bar{V}_2] = 0 \neq V_2$. This contradicts unbiasedness. (2) $\mathbb{E}[\bar{V}_2|N_1^c] \neq 0$: As $R_{23} = 0$ that means that $\bar{p}_{21} \neq 0$, despite the fact that this transition has not been observed. Furthermore, this implies that $\bar{p}_{12} = 1$ as otherwise no valid MRP is defined. As a consequence $\bar{V}_1 = \bar{V}_2$ on both the events N_1 and N_1^c, in particular $\mathbb{E}[\bar{V}_1|N_1^c] = \mathbb{E}[\bar{V}_2|N_1^c] \neq 0$. Now, we get a contradiction with the following argument:

$$V_1 + \mathbb{E}[\bar{V}_1|N_1^c] \overset{\text{unb.}}{=} \mathbb{E}[V_1|N_1] + \mathbb{E}[\bar{V}_1|N_1^c] = \mathbb{E}[V_2|N_1] + \mathbb{E}[\bar{V}_2|N_1^c] = \mathbb{E}[\bar{V}_2] \overset{\text{unb.}}{=} V_2 \overset{\text{Bellm.}}{=} V_1, \tag{60}$$

as the equality can only hold if $\mathbb{E}[\bar{V}_1|N_1^c] = 0$. □

Lemma 2 *For the MRP from Fig. 2(A) with modified reward $R_{11} = 0$, $R_{12} = 1$ and for $n = 1$ there exists no parameter estimator $\bar{p}$ that is independent of γ such that $V_1(\bar{p})$ is unbiased for all parameters p and all discounts γ.*

Proof For $V(\bar{p})$ to be unbiased, it must hold that

$$\mathbb{E}\left[(1-\bar{p})\sum_{i=0}^{\infty}\gamma^i\bar{p}^i\right] = (1-p)\sum_{i=0}^{\infty}\gamma^i p^i \quad \Rightarrow$$

$$\sum_{i=0}^{\infty}\gamma^i(\mathbb{E}[(1-\bar{p})\bar{p}^i] - (1-p)p^i) = 0. \tag{61}$$

If the equality holds for all $\gamma \in (0,1)$, then $\mathbb{E}[(1-\bar{p})\bar{p}^i] = (1-p)p^i$ for $i \geq 0$. Otherwise, with $x_i := \mathbb{E}[(1-\bar{p})\bar{p}^i] - (1-p)p^i$ and x_n being the first term different from 0 ($|x_n| > 0$): $|\gamma^n x_n| = |\sum_{i=n+1}^{\infty}\gamma^i x_i|$ and therefore $|x_n| = |\gamma\sum_{i=n+1}^{\infty}\gamma^{i-(n+1)}x_i|$. We can now adjust the discount to downscale the right hand side arbitrary low, while the left side stays unaffected. The sequence $|x_i|$ is bounded, i.e. $|x_i| \leq \max\{\mathbb{E}[(1-\bar{p})\bar{p}^i], (1-p)p^i\}$ for all i.

Furthermore, both terms are bounded by $\max_{a\in[0,1]}(1-a)a^i$. The maximum is reached for $a = i/(i+1)$ and the maximal value over all i is reached for $i = 0$. The value for $i = 0$ is 1. Therefore,

$$\left|\gamma \sum_{i=n+1}^{\infty} \gamma^{i-(n+1)} x_i\right| < \gamma \sum_{i=n+1}^{\infty} \gamma^{i-(n+1)} 1 = \frac{\gamma}{(1-\gamma)}. \tag{62}$$

For $\gamma < |x_n|/(1-|x_n|)$ the term and the remaining part of the sum becomes smaller than $|x_n|$. $|x_n|/(1-|x_n|)$ is always larger than 0 as $|x_n| > 0$ and for $|x_n| \to 1$ the discount γ can be chosen arbitrary large. In summary this contradicts the assumption and $\mathbb{E}[(1-\bar{p})\bar{p}^i]$ must equal $(1-p)p^i$ for all $i \geq 0$.

Therefore, $\mathbb{E}[1-\bar{p}] = 1-p \Rightarrow \mathbb{E}[\bar{p}] = p$ and $\mathbb{E}[(1-\bar{p})\bar{p}] = (1-p)p \Rightarrow \mathbb{E}[\bar{p}^2] = p^2$. Consequently, $\bar{p}$ must be a constant. Otherwise, we get a contradiction with the following argument: The possible values of $\bar{p}$ are countable (countable many outcomes). We denote the values with a_i and with q_i the probabilities for the values a_i. From $\mathbb{E}[\bar{p}^2] = \mathbb{E}[\bar{p}]^2$ it follows that $\sum_{i=0}^{\infty} q_i a_i^2 = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} q_i q_j a_i a_j \Rightarrow \sum_{i=0}^{\infty}\sum_{j\neq i} q_i q_j a_i a_j = 0$. Furthermore, $q_i, a_i \geq 0$ for all i and therefore $q_i q_j a_i a_j = 0$ for all $i \neq j$. As a consequence, there can be only one $a_i > 0$ with $q_i > 0$. Because $\mathbb{E}[\bar{p}] = p$ it holds that $a_i = p/q_i$ and because $\mathbb{E}[\bar{p}^2] = p^2$ it holds that $a_i = p/\sqrt{q_i}$. Hence, $q_i = 1$ and the parameter estimate is almost surely a constant p. Hence, for an MRP with parameter $p/2$ the estimator will not be unbiased. □

C.2 Markov reward process

Lemma 4 *Assume that each state has a reward distribution which comes from a d-dimensional exponential family. An MRP with finite state space and iid sequences forms then an $(s+|\mathbb{S}|d)$-dimensional exponential family, where s is the number of free MRP parameters.*

Proof Firstly, we demonstrate that the transition distribution forms an exponential family. The density can be written as

$$\mathbb{P}(X_1 = \pi^{(1)}, \ldots, X_l = \pi^{(l)}) = \prod_{i=1}^{\infty} P_{\pi_i}^{c_i} = \exp\left(\sum_{i=1}^{\infty} c_i \log P_{\pi_i}\right), \tag{63}$$

with $\pi^{(j)}$ being the observed paths, $\{\pi_i\}_{i\in\mathbb{N}}$ the set of all paths that can be generated by the MRP (which are countable infinite many paths), c_i is the number of observed paths equivalent to π_i and P_π the probability of path π. The parameters P_π are redundant. 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

$$\sharp\text{Starting States} - 1 + \sum_{i\in\mathbb{S}} \big(\sharp\text{Direct Successors of } i - 1\big). \tag{64}$$

We reformulate the exponential expression to reduce the number of parameters. First, one can observe that $\prod_{i=1}^{\infty} P_{\pi_i}^{c_i}$ is equivalent to $\prod_{i\in\mathbb{S}}(p_i^{n_i}\prod_{j\in\mathbb{S}} p_{ij}^{\mu_{ij}})$, where n_i is the number of starts in state i. The parameters are still redundant: Let state 1 be a starting state and S the remaining set of starting states, then $p_1 = 1 - \sum_{j\in S} p_j$. Furthermore, we have one redundant parameter p_{ij} for every state i. The first problem can be overcome by using $A(\theta)$ in the following way: $n\log(1-\sum_{i\in S} p_i) + \sum_{i\in S} n_i \log\frac{p_i}{(1-\sum_{j\in S} p_j)}$. Here, $A(\theta)$ equals the

n term and n_i is the number of starts in state i. Using the same approach for the transition parameters results in $K_i \log(1-\sum_{j\in S(i)} p_{ij}) + \sum_{j\in S(i)} \mu_{ij} \log \frac{p_{ij}}{(1-\sum_{u\in S(i)} p_{iu})}$, with $S(i)$ being the set of successor states of i without the first successor. This time the K_i term cannot be moved into $A(\theta)$, as K_i is data dependent. This problem can be overcome by observing that $K_i = n_i + \sum_{j\in\mathbb{S}} \mu_{ji}$ and by splitting the K_i terms. As a result we get

$$\exp\Bigg(n \log\left(1-\sum_{i\in S} p_i\right)\left(1-\sum_{u\in S(1)} p_{1u}\right) + \sum_{i\in S} n_i \log \frac{p_i(1-\sum_{u\in S(i)} p_{iu})}{(1-\sum_{j\in S} p_j)} + \sum_{i\in S}\sum_{j\in S(i)} \mu_{ij} \log \frac{p_{ij}(1-\sum_{u\in S(j)} p_{ju})}{(1-\sum_{u\in S(i)} p_{iu})}\Bigg). \tag{65}$$

This case is enough if the reward is deterministic. In the general case where each reward distribution comes from a d-dimensional exponential family the natural parameters and the statistics of the reward distributions must be included in the sum of (22). Similarly, $A(\theta)$ and $h(x)$ need to be adapted. This is possible as there is no functional relation between the MRP parameters and the parameters of the reward distributions. □

C.3 MVU

Theorem 4 $\mathbb{E}[\bar{V}|\mathbb{S}]$ *converges on average to the true value. Furthermore, it converges almost surely if the MC value estimate is upper bounded by a random variable* $Y \in L^1$.

Proof The estimator converges on average, because $\mathbb{E}[|\mathbb{E}[\bar{V}|\mathbb{S}] - V|] \leq \mathbb{E}[|\bar{V} - V|] \overset{n\to\infty}{\longrightarrow} 0$, where n denotes the number of observed paths and convergence follows from the MC convergence. The inequality follows from the Lehmann-Scheffe Theorem, respectively from the Jensen inequality because $|\cdot|$ is convex.

We need to show that

$$\lim_{n\to\infty} \mathbb{E}[\bar{V}|\mathbb{S}] = V \quad \text{a.s.}, \tag{66}$$

where n denotes again the number of observed paths for almost sure convergence.

Due to Bauer and Burckel (1995, Sect. 15 Conditional Expectation, (15.14)) $\lim_{n\to\infty} \mathbb{E}[\bar{V}|\mathbb{S}] = V$ a.s. if $\bar{V}$ converges almost surely to V and if it is upper bounded by a random variable $Y \in L^1$. $\bar{V}$ converges almost surely as $\bar{V}$ is the MC estimator. The second condition is fulfilled due to the assumptions of our Theorem. □

C.4 ML estimator

Theorem 5 *The ML estimator is unbiased if the MRP is acyclic.*

Proof The value function can be written as

$$V_s = \mathbb{E}[R_s] + \sum_{s'\in\mathbb{S}} \sum_{\pi\in\Pi_{ss'}} P_\pi \gamma^{|\pi|} \mathbb{E}[R_{s'}], \tag{67}$$

where $\Pi_{ss'}$ is the set of paths from s to s', P_π the probability of path π, $|\pi|$ the length of the path and $\mathbb{E}[R_s] = \sum_{s'\in\mathbb{S}} p_{ss'} \mathbb{E}[R_{ss'}]$. The ML estimator can be written in the same form,

whereas P_π is replaced with $\bar{P}_\pi := \prod_i \bar{p}_{\pi_i\pi_{i+1}}$ and the expected reward with the reward estimator. The sample mean estimator $\bar{p}$ is unbiased and the reward estimator is unbiased because of our initial assumption. The main problem is to show that $\bar{P}_\pi$ is unbiased, i.e. that

$$\mathbb{E}\left[\prod_i \bar{p}_{\pi_i\pi_{i+1}} \middle| K_s = n\right] \stackrel{?}{=} \prod_i p_{\pi_i\pi_{i+1}}. \tag{68}$$

The estimator for the last transition in a path (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):

$$\begin{aligned}
\mathbb{E}\left[\prod_{i=1}^{L-1} \bar{p}_{\pi_i\pi_{i+1}} \middle| K_s = n\right] &= \sum_{l=1}^{n} \mathbb{E}\left[\prod_{i=1}^{L-1} \bar{p}_{\pi_i\pi_{i+1}} \middle| K_s = n, K_{\hat{s}} = l\right] \mathbb{P}[K_{\hat{s}} = l | K_s = n] \\
&\stackrel{ind}{=} \sum_{l=1}^{n} \mathbb{E}\left[\prod_{i=1}^{L-2} \bar{p}_{\pi_i\pi_{i+1}} \middle| K_s = n, K_{\hat{s}} = l\right] p_{\hat{s}\hat{s}} \mathbb{P}[K_{\hat{s}} = l | K_s = n] \\
&= p_{\hat{s}\hat{s}} \sum_{l=1}^{n} \mathbb{E}\left[\prod_{i=1}^{L-2} \bar{p}_{\pi_i\pi_{i+1}} \middle| K_s = n, K_{\hat{s}} = l\right] \mathbb{P}[K_{\hat{s}} = l | K_s = n] \\
&= p_{\hat{s}\hat{s}} \mathbb{E}\left[\prod_{i=1}^{L-2} \bar{p}_{\pi_i\pi_{i+1}} \middle| K_s = n\right],
\end{aligned} \tag{69}$$

where "ind" abbreviates "independence". 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. One can handle the reward estimator with the same procedure. In summary we find that the value estimator is unbiased. □

Appendix D: Counterexamples

D.1 MC—TD

We present two examples in this section. In the first example MC has a lower MSE than TD(0) and is at least as good as TD(λ) for every λ. In the second example TD(0) is superior to MC.

D.1.1 MC superior to TD

Figure 8(A) shows an example for which the MC estimator is superior. We assume that the learning rate α_i of TD(0) is between 0 and 1, that the learning rate in the first step is 1 ($\alpha_1 = 1$) and that the estimator is initialized to 0 (we use this assumption for readability, it is also possible to use the unbiased TD(0) estimator (Modification 1)). Let state 1 be the starting state, n be the number of observed paths and let $\gamma = 1$ for simplicity.

The MC estimator for state 2 is $1/n \sum_{i=1}^{n} Y_i$, where $Y_i = R_{23}$ or $Y_i = R_{24}$ are the rewards received after a transition from state 2 to state 3 or 4. For state 1 we obtain $1/n \sum_{i=1}^{n} (Y_i + R_{12}^{(i)})$, where the Y_i are the same as before and $R_{12}^{(i)}$ is the received reward after a transition

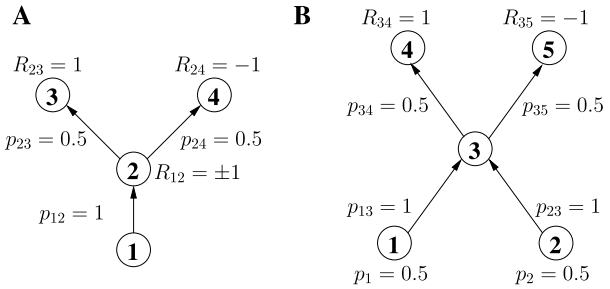

Fig. 8 **A** An MRP for which TD is inferior to MC. The transition from state 1 to state 2 is followed by a reward $R_{12} = +1$ and $R_{12} = -1$ with probability $p = 0.5$ each. **B** An MRP for which MC is inferior to TD. No reward is received for transitions $1 \to 2$ and $1 \to 3$. p_1 and p_2 are the probabilities to start in state 1 and 2

to state 2. The MC estimator is a weighted average of the examples and it is the optimal unbiased linear estimator (4) as $\alpha_i = 1/n$ for all i.

We now analyze the TD(0) estimator. Consider two different sequences α_i and $\tilde{\alpha}_i$, $i = 1, \ldots, n$, of learning rates for the TD(0) estimators $\bar{V}_1$ and $\bar{V}_2$. The TD(0) estimator $\bar{V}_2$ can be written as (Lemma 3, Appendix B)

$$\bar{V}_2^{(n)} = \sum_{i=1}^{n} \left(\tilde{\alpha}_i \prod_{j=i+1}^{n} (1 - \tilde{\alpha}_j) \right) Y_i =: \sum_{i=1}^{n} \tilde{\beta}_i Y_i. \tag{70}$$

The estimator is unbiased and has minimal variance if and only if $\tilde{\beta}_i = 1/n$. This can be enforced by choosing $\tilde{\alpha}_i = 1/i$. For state 1 we obtain

$$\begin{aligned} \bar{V}_1^{(n)} &= \sum_{i=1}^{n} \beta_i (\bar{V}_2^{(i-1)} + R_{12}^{(i)}) = \sum_{i=1}^{n} \beta_i \left(\sum_{j=1}^{i-1} \tilde{\beta}_j Y_j + R_{12}^{(i)} \right) \\ &= \left(\sum_{i=1}^{n} \beta_i R_{12}^{(i)} \right) + \left(\sum_{i=1}^{n-1} \left(\tilde{\beta}_i \sum_{j=i+1}^{n} \beta_j \right) Y_i \right) =: \left(\sum_{i=1}^{n} \beta_i R_{12}^{(i)} \right) + \left(\sum_{i=1}^{n-1} \xi_i Y_i \right), \end{aligned} \tag{71}$$

where $\beta_i = \alpha_i \prod_{j=i+1}^{n} (1 - \alpha_j)$. Using the Bienaymé equality (e.g. Bauer and Burckel 1995) the variance of the estimator takes the following form

$$\mathbb{V}(\bar{V}_1^{(n)}) \overset{ind}{=} \mathbb{V}\left(\sum_{i=1}^{n} \beta_i R_{12}^{(i)} \right) + \mathbb{V}\left(\sum_{i=1}^{n-1} \xi_i Y_i \right) \overset{iid}{=} \mathbb{V}(R_{12}^{(1)}) \sum_{i=1}^{n} \beta_i^2 + \mathbb{V}(Y_1) \sum_{i=1}^{n-1} \xi_i^2. \tag{72}$$

Y_1 and $R_{12}^{(1)}$ have the same variance. With $\xi_n = 0$

$$\mathbb{V}(\bar{V}_1^{(n)}) = \mathbb{V}(Y_1) \left(\sum_{i=1}^{n} \beta_i^2 + \sum_{i=1}^{n} \xi_i^2 \right). \tag{73}$$

Because $0 \le \beta_i, \xi_i \le 1$ and $\sum_{i=1}^{n} \beta_i = \sum_{i=1}^{n} \xi_i = 1$ (see Appendix B) this term would be minimal if and only if $\beta_i = \xi_i = 1/n$. From $\beta_i = \tilde{\beta}_i = 1/n$, however, it follows that $\xi_i = 1/n \sum_{i+1}^{n-1} 1/n = (n - i - 2)/n^2 \neq 1/n$. Hence optimality cannot be achieved. Since both MC and TD are unbiased, we obtain MSE[MC] < MSE[TD].

This example demonstrates a major weakness of TD, namely that it is impossible for TD to weight the observed paths equally, even for simple MRPs. Furthermore, MC is for this example the optimal unbiased value estimator and TD(λ) is unbiased. The optimality of MC

is a direct implication of Corollary 7 from Sect. 3.5. Therefore MSE[MC] ≤ MSE[TD(λ)] for each λ.

D.1.2 TD superior to MC

Figure 8(B) shows an example where TD(0) is superior. Let the number of observed paths be $n = 2$ and $\gamma = 1$. The value of all states is zero. TD(0) and MC are unbiased for this example. The variance of the MC estimator for states 1, 2 and 3 is therefore given by

$$\begin{aligned}\mathbb{E}[\bar{V}_3^2] &= \mathbb{P}[R^{(1)} = 1, R^{(2)} = 1] \cdot 1^2 + \mathbb{P}[R^{(1)} = -1, R^{(2)} = -1] \cdot (-1)^2 \\ &\quad + \mathbb{P}[R^{(1)} = 1, R^{(2)} = -1] \cdot 0 + \mathbb{P}[R^{(1)} = -1, R^{(2)} = 1] \cdot 0 \\ &= (1/4)1^2 + (1/4)(-1)^2 + (2/4) \cdot 0 = 1/2, \\ \mathbb{E}[\bar{V}_1^2] &= \mathbb{E}[\bar{V}_2^2] = (1/4)1/2 + (1/2)1 + 0 = 5/8,\end{aligned} \tag{74}$$

where $R^{(i)}$ denotes the received reward in run i. The first term in the second line results from starting two times in state 1 or 2 and the second term in the second line from a single start in state 1. Setting the learning rate α_i to $\alpha_i = 1/i$ for TD, the estimator for state 3 is equivalent to the corresponding MC estimator and therefore the variance is 1/2. In the first run the standard TD(0) update rule uses the initialization value of state 3 to calculate the estimate in state 1 or 2. This is advantageous and results in a variance of 1/2. Without exploiting this advantage the variance is 17/32. This is still lower than the variance of the MC estimator. Since both estimators are unbiased we obtain MSE[TD(0)] < MSE[MC].

In this example we used the propagation of value estimates which TD performs to show superiority. Without propagation of value estimates TD cannot outperform MC if we use an unbiased TD version and a discount of $\gamma = 1$. This is because MC observes all relevant paths and is hence equivalent to the optimal unbiased value estimator. If the discount is smaller than one then MC is not optimal anymore and also differs from the ML estimator. The second example in Sect. 3.5.1 shows that MC weights transitions from a state s' according to the length of the path taken from state s (the state for which MC estimates the value) to s'. The length should, however, play no role. TD has here a slight advantage as the TD value estimate at state s' treats all transitions independent of the length of the path taken from state s to s'.

D.2 MVU/MC—ML

We show by means of counterexamples that neither the MVU is superior to the ML estimator nor is the ML estimator superior to the MVU or to the MC estimator. We use again the MRP from Fig. 2(A) on p. 299 with modified reward $R_{11} = 0, R_{12} = 1$ and with $n = 1$. As we showed before, the value for state 1 is $(1 - p)/(1 - \gamma p)$ and the ML estimate is $(1 - \bar{p})/(1 - \gamma \bar{p})$, where $\bar{p} = i/(i + 1)$ and i denotes the number of times the cyclic connection has been taken. The MC estimate and therefore the MVU estimate is given by γ^i. Because of the unbiasedness of the MVU/MC estimator the MSE is given by:

$$\mathrm{MSE}[\bar{V}_1] = \mathbb{E}[\bar{V}_1^2] - V_1^2 = (1 - p) \sum_{i=0}^{\infty} \gamma^{2i} p^i - \frac{(1 - p)^2}{(1 - \gamma p)^2} = \frac{p(1 - p)}{(1 - \gamma p)^2 (1 - \gamma^2 p)} (1 - \gamma)^2, \tag{75}$$

where $\bar{V}_1$ denotes the MVU/MC estimator. For the MSE of the ML estimator $\bar{\bar{V}}_1$ we need to calculate the first and the second moment. The first moment:

$$\mathbb{E}[\bar{\bar{V}}_1] = (1-p)\sum_{i=0}^{\infty}\frac{1-\bar{p}}{1-\gamma\bar{p}}p^i = (1-p)\sum_{i=0}^{\infty}\frac{1}{1+(1-\gamma)i}p^i. \tag{76}$$

In the following, we chose γ such that $(1-\gamma)^{-1} = m \in \mathbb{N}$. The sum can then be written as

$$\begin{aligned}\frac{m(1-p)}{p^m}\sum_{i=0}^{\infty}\frac{1}{m+i}p^{m+i} &= \frac{m(1-p)}{p^m}\left(\sum_{i=1}^{\infty}\frac{p^i}{i} - \sum_{i=1}^{m-1}\frac{p^i}{i}\right)\\ &= \frac{m(1-p)}{p^m}\left(\ln\frac{1}{1-p} - \sum_{i=1}^{m-1}\frac{p^i}{i}\right).\end{aligned} \tag{77}$$

The second moment:

$$\begin{aligned}\mathbb{E}[\bar{\bar{V}}_1^2] &= (1-p)\sum_{i=0}^{\infty}\frac{(1-\bar{p})^2}{(1-\gamma\bar{p})^2}p^i = (1-p)\sum_{i=0}^{\infty}\frac{1}{(1+(1-\gamma)i)^2}p^i\\ &= \frac{(1-p)m^2}{p^m}\sum_{i=0}^{\infty}\frac{1}{(m+i)^2}p^{m+i} = \frac{(1-p)m^2}{p^m}\left(\sum_{i=1}^{\infty}\frac{p^i}{i^2} - \sum_{i=1}^{m-1}\frac{p^i}{i^2}\right).\end{aligned} \tag{78}$$

The infinite sum is called *Spence function* or *dilogarithm* and is denoted with $Li_2(p)$. Using these terms one can derive the MSE:

$$\begin{aligned}&\frac{(1-p)^2m^2}{p^m(m(1-p)+p)}\left(\frac{m(1-p)+p}{(1-p)}\left(Li_2(p) - \sum_{i=1}^{m-1}\frac{p^i}{i^2}\right)\right.\\ &\quad - \frac{2}{m(1-p)}\left(\ln\frac{1}{1-p} - \sum_{i=1}^{m-1}\frac{p^i}{i}\right)\\ &\quad \left. + \frac{p^m}{(m(1-p)+p)}\right).\end{aligned} \tag{79}$$

For $\gamma = p = 1/2$ the MSE of the MVU/MC estimator is 0.127 and 0.072 for the ML estimator. Contrary, for $p = 0.99$ the MSE of the MVU/MC estimator is 0.0129 and 0.0219 for the ML estimator.

D.3 Bellman—unbiased

Figure 1 in the introduction gives an overview of the relation between two classes of estimators: Bellman estimators and unbiased estimators. In particular, the figure shows that both classes overlap if the MRP is acyclic or if the Full Information criterion holds and $\gamma = 1$. We present in this section two examples that show that neither of the classes is fully contained in the other, even if these assumptions hold.

Example 1: Bellman $\not\subset$ unbiased This relation is easy to see: Use an estimator of the transition probabilities that does not use data, i.e. $\bar{p}_{ij} = const$. If the constant differs from the true p_{ij} then the corresponding value estimator is biased while it fulfills the Bellman equation. This approach can be applied on any MRP to get a biased Bellman estimator. In particular, it can be applied to any acyclic MRP or to the case where the Full Information criterion holds and $\gamma = 1$.

Example 2: Unbiased $\not\subset$ Bellman The other direction is also not too difficult. We can use Fig. 8(A) with $R_{12} = 0$, $\gamma = 1$ and state 1 being the start state. Then all the assumptions from above are fulfilled. Using the unbiased TD(0) version from Appendix B it is easy to see that this estimator does not fulfill the Bellman equation: Assume that after observing the first path the estimates fulfill the Bellman equation. For the second path we then have two choices. Either it can go through states 1, 2, 3 or 1, 2, 4. Now assume that the Bellman equation will be fulfilled if path 1, 2, 3 occurs (otherwise we are already done). This means that an estimator $\bar{p}_{12}$ exists such that $\bar{V}_1 = \bar{p}_{12}\bar{V}_2$ holds. If instead of path 1, 2, 3 the path 1, 2, 4 occurs then $\bar{V}_1$ and $\bar{p}_{12}$ will be the same as for the path 1, 2, 3 as the TD(0) estimator uses only the old value $\bar{V}_2$ to calculate the new estimates. $\bar{V}_2$ will, however, be different as a different reward is observed.

References

Aigner, M. (2006). *A course in enumeration*. Berlin: Springer.

Bauer, H., & Burckel, R. B. (1995). *Probability theory*. Berlin: de Gruyter.

Bertsekas, D. P., & Tsitsiklis, J. N. (1996). *Neuro-dynamic programming*. Athena Scientific: Nashua.

Boyan, J. (1998). *Learning evaluation functions for global optimization*. PhD thesis, School of Computer Science Carnegie Mellon University.

Boyan, J. (1999). Least-squares temporal difference learning. In *International conference machine learning*.

Bradtke, S. J., & Barto, A. G. (1996). Linear least-squares algorithms for temporal difference learning. *Machine Learning*, *22*.

Grünewälder, S., Hochreiter, S., & Obermayer, K. (2007). Optimality of lstd and its relation to mc. In *Proceedings of the international joint conference of neural networks*.

Jaakkola, T., Jordan, M. I., & Singh, S. P. (1994). On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation*.

Kearns, M., & Singh, S. (2000). Bias-variance error bounds for temporal difference updates. In *Conference on computational learning theory*.

Lehmann, E. L., & Casella, G. (1998). *Theory of point estimation*. *Springer texts in stat*. Berlin: Springer.

Lugosi, G. (2006). *Concentration-of-measure inequalities*. Lecture notes.

Mannor, S., Simester, D., Sun, P., & Tsitsiklis, J. N. (2007). Bias and variance approximation in value function estimates. *Management Science*, *53*.

Singh, S., & Dayan, P. (1998). Analytical mean squared error curves for temporal difference learning. *Machine Learning*, *32*.

Singh, S., & Sutton, R. (1996). Reinforcement learning with replacing eligibility traces. *Machine Learning*, *22*, 123–158.

Sobel, M. J. (1982). The variance of discounted Markov decision processes. *Journal of Applied Probability*, *19*.

Stuart, A., & Ord, K. (1991). *Kendall's advanced theory of statistics* (5th ed.). Sevenoaks: Edward Arnold.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, *3*.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: an introduction*. Cambridge: MIT Press.

Watkins, C., & Dayan, P. (1992). Q-learning. *Machine Learning*, *8*.