Policy Evaluation with Temporal Differences: A Survey and Comparison (Extended Abstract)

Christoph Dann
cdann@cdann.de
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh PA, 15213, USA

Gerhard Neumann
geri@robot-learning.de
Technische Universität Darmstadt
Karolinenplatz 5
64289 Darmstadt, Germany

Jan Peters[2]
mail@jan-peters.net
Max Planck Institute for Intelligent Systems
Spemannstraße 38
72076 Tübingen, Germany

Extended abstract of the article: Christoph Dann, Gerhard Neumann, Jan Peters (2014) *Policy Evaluation with Temporal Differences: A Survey and Comparison* Journal of Machine Learning Research, 15, 809–883.

Introduction

Value functions are an essential tool for solving sequential decision making problems such as Markov decision processes (MDPs). Computing the value function for a given policy (policy evaluation) is not only important for determining the quality of the policy but also a key step in prominent policy-iteration-type algorithms. In common settings where a model of the Markov decision process is not available or too complex to handle directly, an approximation of the value function is usually estimated from samples of the process. Linearly parameterized estimates are often preferred due to their simplicity and strong stability guarantees. Since the late 1980s, research on policy evaluation in these scenarios has been dominated by temporal-difference (TD) methods because of their data-efficiency. However, several core issues have only been tackled recently, including stability guarantees for off-policy estimation where the samples are not generated by the policy to evaluate. Together with improving sample efficiency and probabilistic treatment of uncertainty in the value estimates, these efforts have lead to numerous new temporal-difference algorithms. These methods are scattered over the literature and usually only compared to most similar approaches. The article therefore aims at presenting the state of the art of policy evaluation with temporal differences and linearly parameterized value functions in discounted MDPs as well as a more comprehensive comparison of these approaches.

We put the algorithms in a unified framework of function optimization, with focus on surrogate cost functions and optimization strategies, to identify similarities and differences between the methods. In addition, important extensions of the base methods such as off-policy estimation and eligibility traces for better bias-variance trade-off, as well as regularization in high dimensional feature spaces, are discussed.

We further present the results of the first extensive empirical evaluation, comparing temporal-difference algorithms for value function estimation. These results shed light on the strengths and weaknesses of the methods which will hopefully not only help practitioners in applying these methods but also lead to improvements of the algorithms. As an example, novel versions of the LSTD and LSPE algorithms with drastically improved off-policy performance are presented. For details of these new algorithms, see the full journal article. In the remainder of this extended abstract, we first list temporal-difference algorithms categorized according to their surrogate loss functions and optimization strategies and subsequently summarize the most important results of the empirical analysis. [1]

Temporal Difference Policy Evaluation

Most temporal difference methods for value function estimation can be understood as optimization procedures, that is, combinations of objective functions and optimization strategies. In the following, we first highlight different choices of loss functions and subsequently categorize algorithms according to their optimization procedures.

The general goal of value function estimation is to minimize the *mean squared error (MSE)* which measures the average squared distance between the true value function $V^\pi(s)$ and the linear estimate $\hat{V}(s) = \hat{\theta}^\top \phi(s)$ for each state s. As Monte-Carlo estimation of V^π often has high variance, surrogate loss functions (MSBE, MSPBE, MSTDE, NEU, OPE/FPE) are minimized instead which do not depend on the true value function V^π. They measure in different ways how well the estimate $\hat{V}$ satisfies the Bellman equation $\hat{V}(s) \stackrel{!}{=} \mathbb{E}[r(s_t, a_t) + \gamma \hat{V}(s_{t+1}) | s_t = s]$. Surrogate loss functions are usually easier to estimate from samples due to smaller variance but come at the price of additional bias. The following list categorizes temporal-difference methods according to their objective function:

MSE: Least-squares Monte-Carlo estimation, Kalman TD learning (KTD), Gaussian process TD learning (GPTD)

MSTDE: Bellman residual minimization (BRM), residual gradient (RG);

[1] For the sake of brevity, we omit references in this extended abstract. All references are available in the full journal article.

[2] Also at *Technische Universität Darmstadt, Karolinenplatz 5, Darmstadt, Germany*.

MSBE: BRM with double sampling (BRM DS), residual gradient with double sampling (RG DS);

MSPBE: GTD2, TDC, Least-squares TD-learning (LSTD);

OPE/FPE: Least-squared policy evaluation (LSPE), fixed-point Kalman filtering (FPKF), TD learning;

NEU: GTD.

While the loss function mostly determines the quality of the value function after convergence, the optimization strategy defines the convergence speed and the computational costs of an algorithm. There are roughly three classes of strategies.

Gradient-based methods: (TD learning, GTD, GTD2, TDC, residual gradient) Steps along the stochastic gradient are taken. Algorithms can be employed online and have per-step-runtime linear in the number of features but heavily depend on good choices of step lengths.

Least-squares methods: (LSTD, LSPE, FPKF, BRM) These methods compute the minimum of their quadratic and convex loss function directly and converge much faster than gradient-based methods. However, they have quadratic runtime complexity.

Bayesian approaches: (KTD, GPTD) Bayesian approaches minimize the MSE but mitigate the issue of high variance by incorporating prior beliefs which introduce a bias. Their optimization strategy directly originates from belief updates and usually has quadratic runtime complexity.

Empirical Comparison

While there have been several impressive theoretical analyses of temporal-difference methods, which lead to guarantees and better understanding of them, many question remained open. For example, it has been shown that the MSPBE minimum can be arbitrarily far away from the desired MSE minimum, while the distance between MSBE and MSE minimum (its bias) can be bounded. However, how relevant is this bound in practice? Which surrogate objective has usually lower bias? The article therefore presents a systematic experimental study that aims at empirically answering these questions.

To this end, twelve benchmark problems have been selected, including classic benchmarks, tasks with continuous and discrete state spaces, on- and off-policy scenarios, as well as problems of different size. The performance of all algorithms (including eligibility traces, if available) has been evaluated on these benchmarks in terms of the different objective functions. In addition, the effect of algorithm parameters across all benchmarks has been investigated in extensive parameter studies. The following list summarizes the main results:

- Empirically, the magnitudes of objective biases, that is, the distance of surrogate objective minima to the MSE optimum, are: bias(MSTDE) $\geq$ bias(MSBE) $\geq$ bias(MSPBE) = bias(NEU) = bias(OPE/FPE).
- Optimizing for the MSBE objective instead of the MSTDE objective by using double samples introduces high variance in the estimate. Particularly, Bellman residual minimization requires stronger regularization which results in slower convergence than relying on one sample per transition.
- Interpolating between the MSPBE/MSTDE surrogate objectives and the MSE cost function with eligibility traces can improve the performance of policy evaluation.
- Normalization of features for the linear parameterization of the value-function estimate is crucial for the prediction quality of gradient-based temporal-difference methods.
- The GTD algorithm consistently performs worse than its successors GTD2 and TDC. The TDC method finds the surrogate objective minimum faster than the other gradient-based algorithms GTD, GTD2 and TD learning.
- By optimizing algorithm parameters, the TDC algorithm performs always at least as good as TD-learning, but comes at the price of optimizing an additional hyper-parameter. Often, hyper-parameter optimization yields very small values for the second learning rate, in which case TDC reduces to TD-learning.
- In general, the LSTD and LSPE algorithms produce the predictions with lowest errors for sufficiently many observations.
- In practice, LSTD and LSPE algorithms perform well with off-policy samples only if transition reweighting (proposed in the journal article) is used. The variance of LSTD with standard reweighting makes the algorithm unusable in practice.
- For a modest number of features, least-squares methods are superior to gradient-based approaches both in terms of data-efficiency and even CPU-time to reach the same error level. For a very large number of features (e.g., $\geq$ 20,000), gradient-based methods should be preferred as least-squares approaches become prohibitively time- and memory-consuming.

Conclusion

The article presents a concise survey of temporal difference methods for linear value function estimation including recent trends such as regularization and feature generation to deal with high-dimensional feature spaces. It also presents the results of the first comprehensive empirical comparison that provides evidence for several important questions regarding algorithm stability, parameter choices and surrogate loss functions. The article could therefore be a helpful guide for practitioners to choose the most suitable algorithm for the problem at hand. It further aids researchers to identify possible opportunities for improvements, as shown by the new sample reweighting strategy for LSTD and LSPE proposed and empirically validated in the journal article.