Efficient Reinforcement Learning in Parameterized Models: The Parameter Elimination Approach (Extended Abstract)

Kirill Dyagilev
kirilld@tx.technion.ac.il
Department of EE
Technion
Haifa 32000, Israel

Shie Mannor*
shie@ee.technion.ac.il
Department of EE
Technion
Haifa 32000, Israel

Nahum Shimkin
shimkin@ee.technion.ac.il
Department of EE
Technion
Haifa 32000, Israel

1 Problem Formulation

We consider on-line Reinforcement Learning (RL)[12] in parameterized control models. Namely, we assume that the actual model belongs to a family of MDPs parameterized with a parameter vector which belongs to a compact set. Our goal is to design a learning algorithm that minimizes the total number of suboptimal decisions made throughout the learning period. We shall refer to the latter as the *total mistake count*. In sections below we explain motivation behind this work, proposed algorithm and its performance guarantees. The complete version of this work may be found in [4].

2 Why Parameterized Models?

Several on-line RL algorithms were recently proposed and shown to provide polynomial bounds (in a PAC sense) on the total mistake count. Such algorithms, that depend on efficient resolution of the exploration-exploitation tradeoff, include the R-max algorithm [3],MBIE [11], and Delayed Q-learning [10]. As these algorithms make no structural assumptions on the model involved, they essentially rely on the empirical estimation of the model parameters for each state and action independently; hence, their convergence-rate bounds are at least proportional to the cardinality of the state and action spaces. This may be unacceptable for large problems.

Possible approaches to reduce the complexity of learning in large problems include various approximation schemes, such as parametric representations of the value function, [2], and state aggregation methods (e.g., [1]). In this paper we consider the situation where a parameterized model of the system in question is available. The potential problem simplification offered by such a model in an RL setting can be best demonstrated through a simple example.

Example 0.1 Consider a discrete time queue, with an input buffer of size K and a single server. The control decision may be whether to admit an arrival customer to the queue, or perhaps idle the server; the specifics are not important here. Without prior knowledge, estimating the system transition structure would require independent sampling at each of the possible B states. However, if we know that the arrival and service processes are geometric with rates that do not depend on the buffer occupancy, then two parameters are sufficient to describe the state dynamics, and these parameters can be estimated by monitoring the arrivals and departures at any system states.

The above example becomes even more distinctive when we consider N queues in parallel (say, with a joint routing controller). Here the state space increases exponentially in N, while the number of parameters increases linearly in N. Obviously, simple-minded learning of the transition probabilities at each state separately makes no sense here.

The concept of parameterized control models has been introduced and extensively studied in the stochastic adaptive control literature [9]. However, that research is focused on asymptotic convergence results, rather than on PAC-like convergence bounds which are our concern here.

3 Proposed Solution

We consider the problem of efficient learning in parameterized control models with a *continuous* (and compact) parameter space. We assume a finite state and action MDP model with the discounted reward criterion. We extend our previous work in [8] which considered the case of a finite parameter set.

The algorithm proposed here, called the Approximate Parameter Elimination (APEL) algorithm, starts with a discrete grid that is sufficiently dense in the parameter space. Thus, the compact set of MDPs is approximated (to a specified precision) by a finite set of representative models. At the heart of the algorithm is the hypothesis testing procedure which is used to eliminate unlikely parameter values based on the obtained measurements. The necessary exploration component is handled in the APEL algorithm by choosing a policy that is optimistic with respect to the set of *remaining* parameters, namely those representatives that were not eliminated thus far.

The parameter elimination part in the proposed algorithm is clearly meant to make efficient use of available statistical information, by permanently discarding parameter values that are distinctly not commensurate with the observed data, and thereby reducing over time the set of parameters that needs to be considered. Obviously, it is required to ensure that with high probability the true parameter (or rather, one of its close representatives) is not eliminated in the process.

We further note that the problem considered here is substantially different from the finite parameter case of [8], since

*The author is also with the Department of Electrical and Computer Engineering at McGill University.

the true parameter will generally not belong to the finite grid of representative parameters. Namely, there may be a "mismatch" between the model corresponding to the true parameter and the model induced by the closest grid point. This mismatch requires several major adjustments in the design and analysis of the algorithm. In particular, Wald's Sequential Probability Ratio Test (SPRT) [13] used in the discrete parameter case [8] ceases to provide acceptable guarantees on error probability and convergence rate. The main novelty of this work is in proposition of the Mismatched Probability Ratio Test, a modified version of the SPRT test. This test preserves the error probability and the convergence rate of the original SPRT in the presence of mismatch.

4 APEL's Performance

We provide a PAC bound on APEL's total mistake bound that grows linearly (up to a logarithmic term) in the number of grid points. The proof of this bound relies on the "worst-case" assumptions on the informational structure of models.

We argue that without further assumptions on the model, the linear dependence of the error bound on the number of the grid points can not be improved upon. Moreover, it can be shown that this dependence is the best that can be obtained for this model by *any* learning algorithm.

We further consider a classical Multi-Armed Bandit model [5], a Markov bandit model [7] and a single-server multiclass queueing system [6]. We argue that these models have a favorable information structure and provide sub-linear (logarithmic) model-specific bounds on the APEL's total mistake bound. Moreover, as in these examples the parametrization is decomposable, computational and memory complexities and the total mistake count of the APEL algorithm may be significantly reduced by tuning the algorithm to a specific information structure.

One can also argue that linear dependence on the number of grid points of the mistake bound can be obtained by the following simple–minded algorithm: (a) Try out all optimal policies (one for grid point) for a long enough period of time; (b) choose the policy that proved most profitable, and use it ad infinitum. It can be seen that the error bound on this algorithm is linear in the number of the optimal policies, which is bounded by the number of representative values (some representatives may induce identical optimal policies). However, the APEL algorithm makes a more efficient use of the parametrization of the model, therefore in a large family of models it outperforms the above simple-minded algorithm. In particular, it can be shown that APEL outperforms this simple–minded algorithm in the Markov Bandit model and the queueing example mention above.

5 Conclusion

In this work we considered the case of parameterized models with with a compact parameter set. We proposed a learning algorithm that incorporates efficient exploration to achieve polynomial mistake bounds in the PAC sense. As may be expected, these bounds are independent of the cardinality of the state and action spaces, and in fact may well apply to continuous spaces under reasonable regularity conditions.

The APEL's performance depends on the informational structure of the model. For instance, when the model has a "generous" information structure, the APEL's total mistake bound is small. Generally speaking, the APEL algorithm is expected to perform well when an efficient and hopefully decomposable parametrization of the model is available.

We note that despite the fact that the main focus of this work is on theoretical performance bounds, the resulting algorithm is very intuitive and is simple to implement. We refer the reader to [4] for the complete details on the algorithm and its analysis.

References

[1] Andrey Bernstein and Nahum Shimkin. Adaptive aggregation for reinforcement learning with efficient exploration: Deterministic domains. In *COLT*, pages 323–334, 2008.

[2] D. P. Bertsekas and Tsitsiklis J.N. *Neuro-dynamic Programming*. Athena Scientific.

[3] R.I. Brafman and M. Tennenholtz. R-max - a general polynomial time algorithm for near-optimal reinforcement learning. *JMLR*, 3:213–231, 2002.

[4] Kirill Dyagilev, Shie Mannor, and Nahum Shimkin. Efficient reinforcement learning in parameterized models: Continuous case. Technical report, Technion, 2008. `http://www.ee.technion.ac.il/people/shimkin/PREPRINTS/APELFull.pdf`.

[5] Robbins H. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, pages 527–535, 1952.

[6] Haviv M. Hassin, R. *To Queue or Not to Queue: Equilibrium. Behavior in Queueing Systems.* Kluwer Academic, 2003.

[7] Gittins J.C. *Multi–Armed Bandit Allocation Indices*. Wiley, New York, 1989.

[8] Dyagilev K., Mannor S., and Shimkin N. Efficient reinforcement learning in parameterized models: Discrete parameter case. In *Recent Advances in Reinforcement Learning: 8th European Workshop, EWRL 2008, Revised and Selected Papers*. Springer, 2009.

[9] P.R. Kumar and P. Varaiya. *Stochastic Systems: Estimation, Identification and Adaptive Control*. The MIT Press, 1998.

[10] A.L. Strehl, L. Li, E. Wiewiora, J. Langford, and M.L. Littman. PAC model-free reinforcement learning. *Proceedings of the ICML-06*.

[11] A.L. Strehl and M.L. Littman. A theoretical analysis of model-based interval estimation. *Proceedings of ICML–05*, pages 857–864.

[12] R.S. Sutton and A.G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, 1998.

[13] A. Wald. *Sequential Analysis*. Wiley, 1952.