Dynamic State Space Partitioning for Adaptive Simulation Algorithms

Tobias Helms
Institute of Computer Science
University of Rostock
18059 Rostock, Germany
tobias.helms@uni-rostock.de

Steffen Mentel
Institute of Computer Science
University of Rostock
18059 Rostock, Germany
steffen.mentel@uni-rostock.de

Adelinde M. Uhrmacher
Institute of Computer Science
University of Rostock
18059 Rostock, Germany
adelinde.uhrmacher@uni-rostock.de

ABSTRACT
Adaptive simulation algorithms can automatically change their configuration during runtime to adapt to changing computational demands of a simulation, e.g., triggered by a changing number of model entities or the execution of a rare event. These algorithms can improve the performance of simulations. They can also reduce the configuration effort of the user. By using such algorithms with machine learning techniques, the advantages come with a cost, i.e., the algorithm needs time to learn good adaptation policies and it must be equipped with the ability to observe its environment. An important challenge is to partition the observations to suitable macro states to improve the effectiveness and efficiency of the learning algorithm. Typically, aggregation algorithms, e.g., the adaptive vector quantization algorithm (AVQ), that dynamically partition the state space during runtime are preferred here. In this paper, we integrate the AVQ into an adaptive simulation algorithm.

Categories and Subject Descriptors
I.6.8 [Simulation and Modeling]: Types of Simulation—Adaptive Simulation; I.2.6 [Artificial Intelligence]: Learning—Reinforcement Learning

Keywords
Adaptive Algorithms, Reinforcement Learning, Component-based Simulation Software, Dynamic State Space Representations

1. INTRODUCTION
Adaptive simulation algorithms change their configuration during runtime automatically to improve the overall performance of a simulation [3]. Adaptations can be necessary due to changing computational demands during a simulation run, e.g., caused by a changing number of entities or the execution of a rare event. In general, it is challenging for a user to select a suitable algorithm and configuration for a specific experiment, let alone changing the algorithms during runtime. Even developer of complex algorithms themselves are typically not able to evaluate the performance of their algorithms for a specific problem properly [5]. Adaptive algorithms are an approach to deal with these challenges.

Machine learning techniques can be used to learn when to use which configuration automatically. Adaptive algorithms that use such techniques can be reused for different models, modeling languages and simulation experiments [3]. A suitable machine learning technique is reinforcement learning [11]. In reinforcement learning, the following procedure is repeated until a termination criterion is fulfilled, e.g., a simulation run terminates. An agent observes the environment and performs an action. Then, it receives a reward from the environment and it updates its knowledge base. Typically, the agent starts with little knowledge about the environment and it must find a suitable trade-off between the exploration and the exploitation of knowledge. Based on reinforcement learning, we developed a generic adaptive simulation algorithm [3]. However, the advantages of the learning algorithms come with a cost. For example, they must automatically determine how the features of the model and the environment influence the performance. In reinforcement learning, aggregation algorithms are used to deal with this problem, e.g., [9, 10, 6]. These algorithms partition the state space into disjunct macro states dynamically. A macro state represents a region of the state space with a homogeneous performance behavior. All states within the same macro state are treated equally, i.e., knowledge learned for one state is reused for all other states of the same macro state. The performance of these aggregation algorithms strongly depends on the given problem and the used configuration — a proper usage of these algorithms is a challenge in itself. In this paper, we discuss several aggregation algorithms referring to their applicability for adaptive simulation algorithms and we adjust the adaptive vector quantization algorithm [6] so that it can be used for the adaptive simulation algorithm developed in [3].

2. ADAPTIVE SIMULATOR
The adaptive simulation algorithm called “adaptive simulator” developed in [3] uses reinforcement learning to learn when to use which simulation algorithm and configuration. Referring to reinforcement learning, the adaptive simulator represents the agent, all available simulation algorithms and configurations represent the available actions, and all observable information like the model state represent the environment. The adaptive simulator is organized as a wrapper for other simulation algorithms, i.e., it uses other algorithms to compute the actual simulation and it exchanges the current simulation algorithm during an adaptation. The available simulation algorithms and configurations are listed in a finite
Algorithm splits two adjacent macro states \(s_i\) and \(s_j\) at the mid-point of their longest dimension if the following three conditions based on the current knowledge hold. First, the best action in both macro states differ. Second, the difference of the utilities of the best action of \(ms_i\) and the best action of \(ms_j\) is in \(ms_i\) or \(ms_j\) higher than \(\Delta_{min} \in \mathbb{R}\). Third, all actions of both macro states have been visited at least \(v_{min}\) times. The conditions guarantee that only reasonable splits of macro states are executed. Figure 1 (left) illustrates an exemplary state space partitioning that could have been built with the decision boundary partitioning algorithm. This algorithm can be used for the adaptive simulator, but there are various challenges. For example, if many observed states occur inside the same macro state but not inside its neighbors, this algorithm will not split this macro state although it could be useful. In the worst case, no splits are executed at all because the initial macro states have been set poorly.

Another group of aggregation algorithms uses the idea of the nearest neighbor vector quantization to create macro states, e.g., [6, 13]. These algorithms maintain a codebook \(CB \subseteq S\) containing specific states that are called codewords. A nearest vector quantizer is used to map a state \(s \in S\) onto the nearest codeword \(c \in CB\) available in the current codebook. Basically, this mapping creates a partitioning of the state space into disjoint regions, i.e., the macro states (see Figure 1). The idea of these algorithms is to frequently change the codebook, so that all states mapped to a macro state represent a similar q-value behavior. A promising algorithm of this group to be applied to the adaptive simulator is the adaptive vector quantization algorithm (AVQ). Because it does not depend on the definition of a goal state, b) it allows complex shapes of macro states, and c) no initial partitioning for each dimension must be defined. Further, compared to other aggregation algorithms that use the nearest neighbor vector quantization, the configuration effort of this algorithm is low. To decide whether new codewords shall be added to the codebook, this algorithm uses a concept based on the accumulated reward \([acc\text{Reward}]\), that “with respect to a particular action is the sum of the total rewards received by continuously taking the same action within a particular cell” [6]. Consequently, this algorithm directly influences the action selection. This approach has to be replaced when applying this algorithm to the adaptive simulator, because it is possible that a poor performing action is reused frequently.

3. AGGREGATED STATE SPACES

Due to high-dimensional and real-valued state spaces, it is usually not feasible to learn a suitable selection policy for each state individually. Aggregation algorithms, e.g., [9, 10, 6, 13, 1], dynamically partition the state space of a reinforcement learning problem into disjoint macro states do deal with this problem. Typically, these algorithms start with a coarse-grained partitioning of the state space and refine the state space based on various conditions.

For example, the parti-game algorithm [9] assumes that a) the goal is to reach a specific region of the state space, b) that there exists a continuous path from the start position to the goal region, c) that state transitions are deterministic, and d) that the agent can move deliberately through the state space. Due to these requirements, this algorithm cannot be applied to the adaptive simulator. First, the goal of the adaptive simulator is not to reach a specific region of the state space, but to maximize the received rewards. Second, the adaptive simulator cannot move deliberately through the state space — it cannot influence deliberately model variables that are possibly used as dimensions for the state space.

An extended and less restricted version of the parti-game algorithm is the decision boundary partitioning algorithm [10]. This algorithm splits two adjacent macro states \(ms_i\) and \(ms_j\) at the middle of their longest dimension if the following three conditions based on the current knowledge hold. First, the best action in both macro states differ. Second, the difference of the utilities of the best action of \(ms_i\) and the best action of \(ms_j\) is in \(ms_i\) or \(ms_j\) higher than \(\Delta_{min} \in \mathbb{R}\). Third, all actions of both macro states have been visited at least \(v_{min}\) times. The conditions guarantee that only reasonable splits of macro states are executed. Figure 1 (left) illustrates an exemplary state space partitioning that could have been built with the decision boundary partitioning algorithm. This algorithm can be used for the adaptive simulator, but there are various challenges. For example, if many observed states occur inside the same macro state but not inside its neighbors, this algorithm will not split this macro state although it could be useful. In the worst case, no splits are executed at all because the initial macro states have been set poorly.

Another group of aggregation algorithms uses the idea of the nearest neighbor vector quantization to create macro states, e.g., [6, 13]. These algorithms maintain a codebook \(CB \subseteq S\) containing specific states that are called codewords. A nearest vector quantizer is used to map a state \(s \in S\) onto the nearest codeword \(c \in CB\) available in the current codebook. Basically, this mapping creates a partitioning of the state space into disjoint regions, i.e., the macro states (see Figure 1). The idea of these algorithms is to frequently change the codebook, so that all states mapped to a macro state represent a similar q-value behavior. A promising algorithm of this group to be applied to the adaptive simulator is the adaptive vector quantization algorithm (AVQ). Because it does not depend on the definition of a goal state, b) it allows complex shapes of macro states, and c) no initial partitioning for each dimension must be defined. Further, compared to other aggregation algorithms that use the nearest neighbor vector quantization, the configuration effort of this algorithm is low. To decide whether new codewords shall be added to the codebook, this algorithm uses a concept based on the accumulated reward \([acc\text{Reward}]\), that “with respect to a particular action is the sum of the total rewards received by continuously taking the same action within a particular cell” [6]. Consequently, this algorithm directly influences the action selection. This approach has to be replaced when applying this algorithm to the adaptive simulator, because it is possible that a poor performing action is reused frequently.

4. AVQ AND ADAPTIVE SIMULATOR

Algorithm 1 illustrates the integration of the AVQ algorithm in the adaptive simulator. As motivated above, we replaced the concept of the accumulated reward. Instead, we use a condition inspired by the decision boundary partitioning algorithm. A state \(s\), that is mapped to a codeword \(c\), is added to the codebook if the absolute difference of the current reward and the last reward achieved by the same action for any other state \(s'\) mapped to \(c\) is higher than a threshold \(\alpha \in \mathbb{R}\) (see l. 16). Generally speaking, a state is added to the codebook if the rewards of an action differ significantly within its macro state. Further, we add a limit \(m \in \mathbb{N}\) for the number of macro states. The reward of the adaptive simulator is calculated by computing the logarithm of base 2 of the event throughput. We set \(\alpha = log_2(1.5) \approx 0.585\) by default, i.e., a throughput difference of at least 50% must occur so that a codeword is added to the codebook. We reused the merging routine of the AVQ (l. 35-41). Thus, at the end of each simulation run, i.e., after finishing the adaptation loop, the merging routine is executed. We set \(\rho = \sigma^2 \approx 0.342\) by default, so that this parameter is linked to the condition to add codewords to the codebook.

We demonstrate the effectiveness and efficiency of our approach with a benchmark using a two-dimensional state space \((x, y) \in \mathbb{R}^2\)
Algorithm 1: Pseudo-code for the adaptive simulator [3] extended by the changed AVQ.

| Line | Description |
|------|-------------|
| 1    | \( c := \text{nearest}\_\text{neighbor}(s_0, CB) \) |
| 2    | \( s := s_0 \) |
| 3    | \( a := a_0 \) |
| 4    | \( \text{initialize}(a) \) |
| 5    | \( \text{repeat} \{ / /	ext{Adaptation loop} \) |
| 6    | \( N[c,a] := N[c,a] + 1 \) |
| 7    | \( \tau := [] \) |
| 8    | \( \text{repeat} \{ / /	ext{Simulation loop} \) |
| 9    | \( \text{simulate}\_\text{next}\_\text{event}(a) \) |
| 10   | \( \sigma := \text{observe}(a) \) |
| 11   | \( \tau := \tau + \sigma \) |
| 12   | \( \text{until adaptation}\_\text{condition}(\tau, f, R) \) |
| 13   | \( r := R(\tau) / /	ext{Calculate reward} \) |
| 14   | \( \text{//Check whether to add s to the codebook} \) |
| 15   | \( \text{if} \ (|CB| < m \&\& |L[c,a] − r| > \alpha) \) |
| 16   | \( c := s \) |
| 17   | \( L[c,a] := r \) |
| 18   | \( / /	ext{Simplified Q-learning} : \) |
| 19   | \( Q[c,a] := Q[c,a] + \frac{1}{|\text{AVQ}|} \cdot (r − Q[c,a]) \) |
| 20   | \( / /	ext{Aggregate base states to a state} \) |
| 21   | \( s := p(\tau) \) |
| 22   | \( / /	ext{Get codeword for computed state} \) |
| 23   | \( c := \text{nearest}\_\text{neighbor}(s, CB) \) |
| 24   | \( / /	ext{Select next action based on c, Q, N} \) |
| 25   | \( a_{\text{next}} := f(c, Q, N) \) |
| 26   | \( / /	ext{Adapt internal SA:} \) |
| 27   | \( a := \text{adapt}(a, a_{\text{next}} + \sigma) \) |
| 28   | \( / /	ext{Merging process} \) |
| 29   | \( \text{for} \ (c \in CB) \) |
| 30   | \( c' := \text{nearest}\_\text{neighbor}(c, CB) \) |
| 31   | \( \text{if} \ (\sum_{a \in A} (Q[c,a] − Q[c', a])^2 / |A| < \rho) \) |
| 32   | \( CB := (CB \cup \{c, c'\}) \cup \{(c + c') / 2\} \) |
| 33   | \( \text{update}_B(Q, N, c, c') \) |
| 34   | \} |

Figure 2: State space partitionings built by using the changed AVQ with \( m = 10 \) (left) and \( m = 100 \) (right).

Figure 3: Success rates for the benchmark using the changed AVQ (left) and a fixed grid (right), once for the size \( x \in [0.2 \cdot \pi, y \in [-1, 1] \) (circles), once for the scaled down size \( x \in [0.52, 0.54], y \in [0.39, 0.41] \) (crosses). The dotted lines show the success rate of a random choice, i.e., \( \frac{1}{2} \).

In other actions. If the state observed by the agent is between the sine and cosine curve, it receives reward 1 for choosing action \( a_1 \), and reward 0 for the other actions. If the state observed by the agent is below the sine and cosine curve, it receives reward 1 for choosing action \( a_2 \), and reward 0 for the other actions. In the last case, i.e., if the state observed by the agent is below the sine and cosine curve, it receives reward 1 for choosing action \( a_2 \), and reward 0 for the other actions. The actions chosen by the agent do not influence the next state of the environment — the next state is chosen randomly. Finally, 100 trials are executed with 1000 steps per trial, i.e., the merging process is executed every 1000 steps. Although simple, this setting reflects important aspects of the adaptive simulator scenario, e.g., the actions of the agent do not influence the observed environment states. Figure 2 illustrates the final partitioning of the state space using \( m \in \{10, 100\} \). The success rates (the rate of correct decisions) are shown in Figure 3 (left). As expected, the more macro states are allowed, i.e., the higher \( m \) is chosen, the more accurate gets the state space partitioning and the higher is the success rate. However, the success rate converges to a value \( \approx 0.87 \), i.e., at some degree it is not worth to increase the number of macro states, because most of the occurring states are already covered. Figure 3 (right) also shows the success rate of a static rectangular grid with different macro state sizes. Here, the smaller the length and width of a macro state, the worse the success rate, because the more redundant knowledge is learned.

To demonstrate that the results do not depend on the granularity of state space dimensions, we repeat the same benchmark experiment with a scaled down state space with scaled down sine and cosine curves so that \( x \in [0.52, 0.54] \) and \( y \in [0.39, 0.41] \). The results of this scaled down benchmark are also shown in Figure 3 (crosses). Whereas the results using the changed AVQ are almost equal, the results using a static grid state space differ significantly, i.e., a granularity that is effective for one problem is usually unsuitable for another.
For the changed A VQ, we used the default configuration of the adaptive simulator with its configuration effort.

7. ACKNOWLEDGMENTS
This research was supported by the German research foundation (DFG), via the research grant ESCeMMo (UH-66/14).

8. REFERENCES
[1] A. Bonarini, A. Lazaric, and M. Restelli. Reinforcement Learning in Complex Environments Through Multiple Adaptive Partitions. In AI*IA 2007: Artificial Intelligence and Human-Oriented Computing, volume 4733, pages 531–542. 2007.
[2] R. Ewald, J. Himmelspach, and A. M. Uhrmacher. An Algorithm Selection Approach for Simulation Systems. In Proceedings of the 22nd Workshop on Principles of Advanced and Distributed Simulation (PADS’08), pages 91–98, 2008.
[3] T. Helms, R. Ewald, S. Rybacki, and A. M. Uhrmacher. Automatic Runtime Adaptation for Component-based Simulation Algorithms. ACM Transactions on Modeling and Computer Simulation (TOMACS), 2015. to appear.
[4] T. Helms, O. Reinhardt, and A. M. Uhrmacher. Bayesian Changepoint Detection for Generic Adaptive Simulation Algorithms. In Proceedings of the 48th Annual Simulation Symposium (ANSS’15), pages 62–69, 2015.
[5] H. H. Hoos. Programming by Optimization. Communications of the ACM, 55(2):70–80, 2012.
[6] I. S. Lee and H. Y. Lau. Adaptive state space partitioning for reinforcement learning. Engineering Applications of Artificial Intelligence, 17(6):577–588, 2004.
[7] C. Maus, S. Rybacki, and A. M. Uhrmacher. Rule-based multi-level modeling of cell biological systems. BMC Systems Biology, 5(166), 2011.
[8] O. Mazemondet, R. Hubner, J. Frahm, D. Koczan, B. Bader, D. Weiss, A. M. Uhrmacher, J. F. Rech, A. Rolfs, and J. Luo. Quantitative and kinetic profile of Wnt/β-catenin signaling components during human neural progenitor cell differentiation. Cellular & Molecular Biology Letters, 16(4):515–538, 2011.
[9] A. W. Moore and C. G. Atkeson. The parti-game algorithm for variable resolution reinforcement learning in multidimensional state-spaces. Machine Learning, 21(3):199–233, 1995.
[10] S. I. Reynolds. Decision boundary partitioning: Variable resolution model- free reinforcement learning. In Proceedings of the 17th International Conference on Machine Learning (ICML’00), pages 783–790, 2000.
[11] R. S. Sutton and A. G. Barto. Reinforcement Learning: An Introduction. MIT Press, 1998.
[12] J. Veromel and M. Mohri. Multi-armed Bandit Algorithms and Empirical Evaluation. In Proceedings of the 16th European conference on Machine Learning (ECML’05), pages 437–448, 2005.
[13] D. Vieira, P. Adeodato, and P. Goncalves. A Temporal Difference GNG-Based Algorithm That Can Learn to Control in Reinforcement Learning Environments. In Proceedings of the 12th International Conference on Machine Learning and Applications (ICMLA), pages 329–332, 2013.
[14] C. J. C. H. Watkins. Learning from delayed rewards. PhD thesis, University of Cambridge, England, 1989.