A Parallel Point-Based POMDP Algorithm Leveraging GPUs

Kyle Hollins Wray and Shlomo Zilberstein

College of Information and Computer Sciences
University of Massachusetts, Amherst, MA 01003
{wray, shlomo}@cs.umass.edu

Abstract

We parallelize the Point-Based Value Iteration (PBVI) algorithm, which approximates the solution to Partially Observable Markov Decision Processes (POMDPs), using a Graphics Processing Unit (GPU). We detail additional optimizations, such as leveraging the bounded size of non-zero values over all belief point vectors, usable by serial and parallel algorithms. We compare serial (CPU) and parallel (GPU) implementations on 10 distinct problem domains, and demonstrate that our approach provides an order of magnitude improvement.

Introduction

Partially Observable Markov Decision Processes (POMDPs) (Kaelbling et al. 1998) are a powerful mathematical model which describe the interactions of an agent in a stochastic environment, with numerous real-world applications such as semi-autonomous driving (Wray and Zilberstein 2015). Since POMDPs are PSPACE-hard, approximation algorithms have been developed, such as Point-Based Value Iteration (PBVI) (Pineau et al. 2003) and Perseus (Sapna and Vlassis 2005). This algorithm performs a single update equation (single instruction) over multiple parts of the POMDP (multiple data), which is amenable to parallelization. We present a General-Purpose Graphics Processing Unit (GPGPU) approach to PBVI, as well as discuss key optimizations which vastly improve performance all parallel-based algorithms. To the best of our knowledge, our approach provides an order of magnitude improvement, and index of our CPU and GPU implementations.

Optimizing Parallel PBVI

A POMDP is a tuple \( \langle S, A, \Omega, T, O, R \rangle \). \( S \) is a set of \( n \) states. \( A \) is a set of \( m \) actions. \( \Omega \) is a set of \( z \) observations. \( T : S \times A \times S \rightarrow [0, 1] \) is a state transition such that \( T(s,a,s') = Pr(s'|s,a) \). \( O : A \times S \times \Omega \rightarrow [0, 1] \) is an observation transition such that \( O(a,s',\omega) = Pr(\omega|a,s') \). \( R : S \times A \rightarrow \mathbb{R} \) is a reward (denoted \( R(s,a) \)). POMDPs have a horizon \( h \in \mathbb{N} \) and a discount factor \( \gamma \in (0,1) \). The agent maintains a set of \( r \) beliefs \( B \subseteq \Delta^n \) over the true state (standard \( n \)-simplex \( \Delta^n \)). A policy \( \pi : B \rightarrow A \) determines which action to take based on a belief. The goal is to find a policy that maximizes the expected reward over time, or value function \( V : B \rightarrow \mathbb{R} \). Since it is piecewise linear and convex, we use a set of \( \alpha \)-vectors \( \Gamma = \{ \alpha_1, \ldots, \alpha_r \} \). At time \( t \), the value at \( b \) is:

\[
V^t(b) = \max_{a \in A} \sum_{s \in S} b(s)R(s,a) + \sum_{\omega \in \Omega} \max_{a \in A} \sum_{s \in S} b(s)V^t_{s,a,\omega}.
\]

\[
V^t_{s,a,\omega} = \gamma \sum_{s' \in S} O(a,s',\omega)T(s,a,s')\alpha(s')
\]

PBVI updates the \( \alpha \)-vectors \( \Gamma^t \) only at a fixed set of beliefs:

\[
\Gamma^t_{a,\omega} = \left\{ [V^t_{s_1,a,\omega}, \ldots, V^t_{s_n,a,\omega}]^T, \forall \alpha_i \in \Gamma^{t-1} \right\}
\]

\[
\Gamma^t = \left\{ \arg\max_{a \in A} \alpha \cdot \beta, \forall \beta \in B \right\}
\]

There are two main optimizations required for GPU algorithms: (1) coalesced memory access (each thread accesses contiguous lines in memory; i.e., the same cache line) and (2) avoid branching (e.g., avoid if-statements and keep threads synchronized to the same instructions). Thus, all variables (\( T, B, \text{etc.} \)) are structured arrays which ensure that thread “execution groups” called warps are aligned memory accesses to the same region (coalesced access), and we use a reduction for \( \arg\max \) (avoid branching).

Our approach has two steps which fully parallelize computing \( \Gamma^t_b \) and \( \Gamma^t \). First, we launch threads that compute the optimal \( \alpha \in \Gamma^t_{a,\omega} \) by evaluating each \( \alpha \cdot \beta \) in parallel (\( rmz \) blocks of \( r \) threads). Importantly, we only store the value and index of \( \alpha \cdot \beta \), use a reduction to compute the maximum in \( \log(r) \), and recompute it compute \( \alpha \cdot \beta \in \Gamma^t_\beta \). Second, we perform a similar technique for \( \Gamma^t \) for each \( B \).

We limit array sizes to improve performance and reduce RAM usage so that we can solve much larger POMDPs. Specifically, we only use state-action pairs for \( R \) and limit the size of \( T \) and \( B \) (optionally, \( O \) and \( A \)). We let \( n_s \) denote the maximum number of successors such that \( n_s = \)
max_{s,a} |\{ s' \in S | T(s,a,s') > 0 \}|, define S : S \times A \rightarrow \mathcal{P}(S) as the mapping to available successor states, and redefine T : S \times A \times \{1, \ldots, n_s\} \rightarrow \{0, 1\}. State sizes tend to have a large impact on performance and memory. This common optimization enables larger problems to be quickly solved.

To the best of our knowledge, no one has discussed a similar optimization with respect to belief points: remove zero values from all b \in B. All uses of b are dot-products in our PBVI equations, for which computation can be greatly reduced. Similarly, we let r_z denote the maximum number of non-zero belief values such that r_z = \max_{b \in B} |\{ s \in S | b(s) > 0 \}|, define Z : B \rightarrow \mathcal{P}(S) as the mapping to non-zero belief points, and redefine all b \in B to be parallel arrays to the state-arrays in Z(b) of only the non-zero values over beliefs.

### Experimentation

We implemented our parallel approach using CUDA\(^1\) on an Intel(R) Core(TM) i7-4702HQ CPU at 2.20GHz, 8GB of RAM, and an Nvidia(R) GeForce GTX 870M graphics card using C++, CUDA(C) 7.0, and Python 3.4.3. We compared the computation time of a single update step of GPU and CPU implementations. These two algorithms were run on 10 problem domains, each with a varying number of states, actions, observations, and structure regarding the state and observation transitions, linearly varying the number of belief points within each domain. We include 4 cities from the semi-autonomous driving domain (Wray and Zilberstein 2015), denoted Drive, using the road distance cost function. Table 1 shows the domains and final run time results. They represent averages over 10 trials in order to reduce any noise resulting from our belief point expansion.

The complexity of our CPU implementation’s update step is \( O(n \times \sum_{z} (n_z + r_z)) \). In comparison, our GPU implementation’s complexity is \( O(r_z \times (n_n + n_z) + \log(r_z) + n_z + n) \). This assumes an ideal scenario in which enough threads are created to perform instructions in parallel. The exponential increase of hardware performance and capability will likely make this increasingly more realizable with time.

### Conclusion

Our results demonstrate approximately an order of magnitude improvement is realizable with the current generation of GPUs. The vast asymptotic difference between our implementations, however, suggests that this can be greatly improved by increasing the number of threads. Additionally, the comparison between Drive (\( r_z \ll n \)) and domains such as Tiger Grid (\( r_z \approx n \)) provides strong evidence that the maximum number of non-zero belief points \( r_z \) significantly affects performance. In general, proactively incorporating how many non-zero values are in all probability vectors (e.g., our \( B \) optimization “trick”) can yield large performance gains.

Our work presents a new perspective on designing parallel algorithms for POMDPs by leveraging the power of GPUs. Experimentation shows that our GPU-based implementation successfully outperforms our CPU-based implementation, utilizing the aforementioned optimizations. We compiled these implementations into a single open-source library entitled nova, which also includes parallel solvers for other MDP models, so that others may build upon this work to produce high-performance parallel algorithms.

### Acknowledgments

This work was supported in part by NSF grant IIS-1405550.

### References


---

1https://developer.nvidia.com/about-cuda

<table>
<thead>
<tr>
<th>Domain</th>
<th>n</th>
<th>m</th>
<th>z</th>
<th>( r_1 )</th>
<th>( n_s )</th>
<th>( r_z )</th>
<th>( 1 \times r_1 )</th>
<th>( 2 \times r_1 )</th>
<th>( 3 \times r_1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tiger</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>16</td>
<td>2</td>
<td>2</td>
<td>0.002</td>
<td>0.003</td>
<td>0.004</td>
</tr>
<tr>
<td>Grid (4x3)</td>
<td>11</td>
<td>4</td>
<td>6</td>
<td>32</td>
<td>9</td>
<td>4</td>
<td>0.014</td>
<td>0.013</td>
<td>0.027</td>
</tr>
<tr>
<td>Tiger Grid</td>
<td>36</td>
<td>5</td>
<td>17</td>
<td>128</td>
<td>5</td>
<td>32</td>
<td>1.475</td>
<td>5.507</td>
<td>3.154</td>
</tr>
<tr>
<td>Hallway2</td>
<td>92</td>
<td>5</td>
<td>17</td>
<td>256</td>
<td>88</td>
<td>88</td>
<td>7.896</td>
<td>55.886</td>
<td>23.271</td>
</tr>
<tr>
<td>Tag</td>
<td>870</td>
<td>5</td>
<td>30</td>
<td>256</td>
<td>5</td>
<td>29</td>
<td>6.607</td>
<td>41.081</td>
<td>23.952</td>
</tr>
<tr>
<td>Rock Sample (7x8)</td>
<td>12545</td>
<td>13</td>
<td>2</td>
<td>256</td>
<td>1</td>
<td>256</td>
<td>8.140</td>
<td>45.897</td>
<td>17.735</td>
</tr>
<tr>
<td>Drive: San Franc.</td>
<td>172</td>
<td>8</td>
<td>2</td>
<td>256</td>
<td>2</td>
<td>2</td>
<td>0.073</td>
<td>0.216</td>
<td>0.144</td>
</tr>
<tr>
<td>Drive: Seattle</td>
<td>432</td>
<td>10</td>
<td>2</td>
<td>512</td>
<td>2</td>
<td>2</td>
<td>0.233</td>
<td>1.180</td>
<td>0.597</td>
</tr>
<tr>
<td>Drive: N.Y.C.</td>
<td>1064</td>
<td>12</td>
<td>2</td>
<td>1024</td>
<td>2</td>
<td>2</td>
<td>1.083</td>
<td>7.256</td>
<td>2.997</td>
</tr>
<tr>
<td>Drive: Boston</td>
<td>2228</td>
<td>12</td>
<td>2</td>
<td>2048</td>
<td>2</td>
<td>2</td>
<td>5.288</td>
<td>27.700</td>
<td>19.155</td>
</tr>
</tbody>
</table>

Table 1: Update step computation time (in seconds) for 6 well-known domains and 4 Drive cities averaged over 10 trials.