

# COMP767 - Reinforcement Learning

## Labeled RTDP: Improving the Convergence of Real-Time Dynamic Programming

Claudio Sole

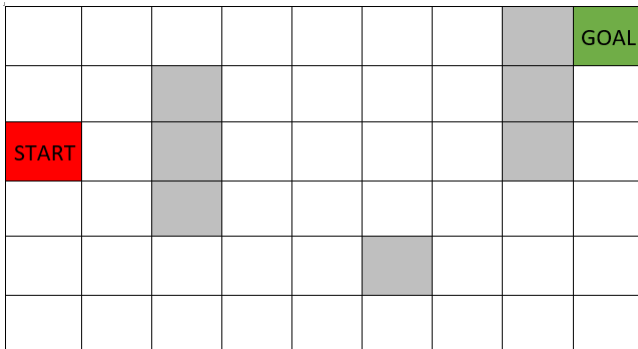
École polytechnique de Montréal

March 9, 2017

- 1 The problem
- 2 DP algorithms seen so far
- 3 Labeled RTDP
- 4 References

## (Deterministic) Shortest Path Problem on a grid

- For each state the possible action are  $\{up, down, right, left\}$
- If an action takes the agent out of the grid or against a wall, then the position on the grid remains the same (self-transitions)



- 1 The problem
- 2 DP algorithms seen so far
  - Synchronous DP
  - Gauss-Seidel
  - Real-Time Dynamic Programming (RTDP)
- 3 Labeled RTDP
- 4 References

---

## Algorithm 1 Synchronous DP

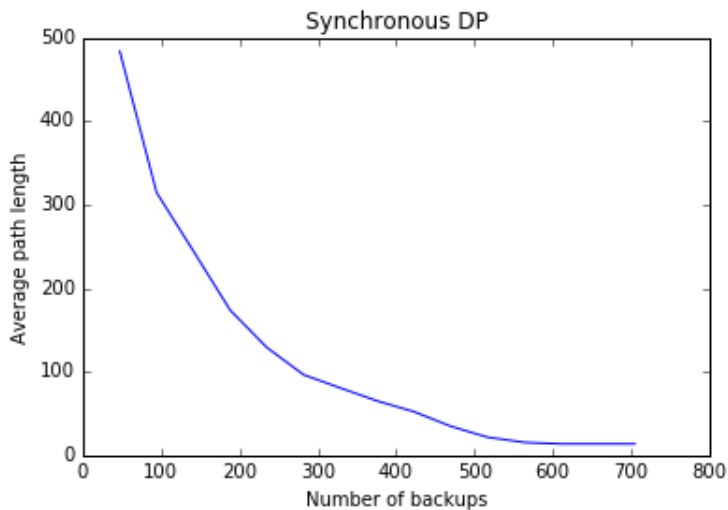
---

```

1: Initialize  $V$  arbitrarily (e.g.  $V(s) = 0 \forall s \in S$ )
2:  $\epsilon \leftarrow 10^{-4}$ 
3:  $\Delta \leftarrow \epsilon + 1$ 
4: while  $\Delta > \epsilon$  do
5:      $\Delta \leftarrow 0$ 
6:      $V_{old} \leftarrow V$ 
7:     for  $s \in S$  do
8:          $V(s) \leftarrow \max_a \sum_{s',r} p(s', r | s, a) [r + \gamma V_{old}(s')]$ 
9:          $\Delta \leftarrow \max\{\Delta, |V_{old}(s) - V(s)|\}$ 
    
```

---

Remark : states ordering has no influence



---

## Algorithm 2 Gauss-Seidel DP

---

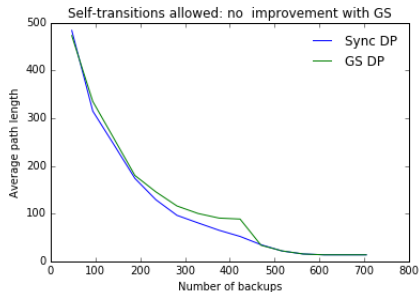
```

1: Initialize  $V$  arbitrarily (e.g.  $V(s) = 0 \forall s \in S$ )
2:  $\epsilon \leftarrow 10^{-4}$ 
3:  $\Delta \leftarrow \epsilon + 1$ 
4: while  $\Delta > \epsilon$  do
5:    $\Delta \leftarrow 0$ 
6:    $V_{old} \leftarrow V$ 
7:   for  $s \in S$  do
8:      $v \leftarrow V(s)$ 
9:      $V(s) \leftarrow \max_a \sum_{s',r} p(s', r | s, a) [r + \gamma V_{old}(s')]$ 
10:     $\Delta \leftarrow \max\{\Delta, |v - V(s)|\}$ 

```

---

Remark: states ordering influences convergence rate



self-transitions	Yes		No	
	Sync	GS	Sync	GS
# sweeps	15	15	15	11
# backups	705	705	705	517



---

### Algorithm 3 TRIAL-BASED RTDP

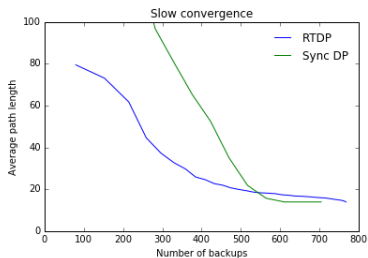
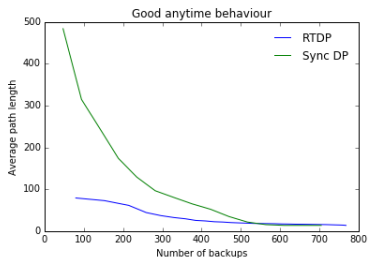
---

```

1: function RTDP( ):
2:   while not converged do
3:     RTDP_TRIAL(start)

4: function RTDP_TRIAL(s):
5:   while s is not goal do
6:      $greedy\_a \leftarrow \operatorname{argmax}_a \sum_{s',r} p(s', r | s, a) [r + \gamma V(s')]$ 
7:      $V(s) \leftarrow \sum_{s',r} p(s', r | s, greedy\_a) [r + \gamma V(s')]$ 
8:      $s \leftarrow s'$  with probability  $P(s' | s, greedy\_a)$ 
    
```

---



- 1 The problem
- 2 DP algorithms seen so far
- 3 Labeled RTDP**
- 4 References

## Definitions:

- **Greedy graph( $s$ )**: graph made up of the states reachable from  $s$  following the greedy policy
- **Greedy envelope( $s$ )**: set of states in the greedy graph of  $s$
- **Residual( $s$ )** =  $|V(s) - Q(s, greedy\_a)|$

## Idea

*We keep track of the states over which the value function has converged and avoid visiting those states again*

---

### Algorithm 4 CheckSolved( $s$ )

---

```
1: if  $\nexists s' \in greedyEnvelope(s) : Residual(s') > \epsilon$  then
2:   mark as SOLVED  $s$ 
3:   mark as SOLVED each states in the greedy envelope of  $s$ 
4:   return TRUE
5: else
6:   backup states in the greedy envelope of  $s$ 
7:   return FALSE
```

---

## Algorithm 5 TRIAL-BASED LRTDP

---

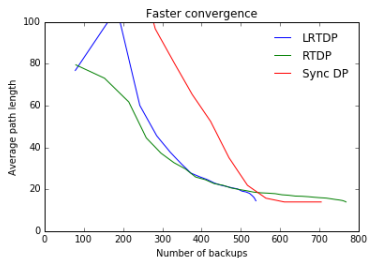
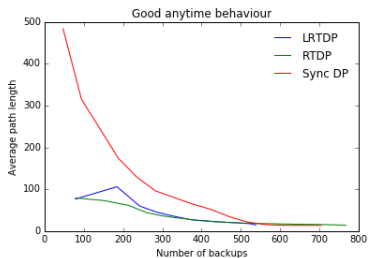
```

1: function LRTDP( ):
2:   while start not SOLVED do
3:     LRTDP_TRIAL(start)

4: function LRTDP_TRIAL(s):
5:   visited = empty Stack
6:   while s is not SOLVED do
7:     visited.push(s)
8:     if s is goal then
9:       break
10:     $greedy\_a \leftarrow \operatorname{argmax}_a \sum_{s',r} p(s', r | s, a) [r + \gamma V(s')]$ 
11:     $V(s) \leftarrow \sum_{s',r} p(s', r | s, greedy\_a) [r + \gamma V(s')]$ 
12:    s  $\leftarrow s'$  with probability  $P(s' | s, greedy\_a)$ 

13:  while visited  $\neq$  EMPTY do
14:    s = visited.pop()
15:    if not checkSolved(s) then
16:      break
  
```

---



- 1 The problem
- 2 DP algorithms seen so far
- 3 Labeled RTDP
- 4 References**





Barto, Andrew G. and Bradtke, Steven J. and Singh, Satinder P.

*Learning to act using real-time dynamic programming*  
Artificial Intelligence  
1995



Bonet, Blai and Geffner

*Labeled RTDP : Improving the Convergence of Real-Time Dynamic Programming*  
Proceedings of Thirteenth International Conference on Automated Planning and Scheduling  
2003