BRTDP for Stochastic Games

Maxi Weininger

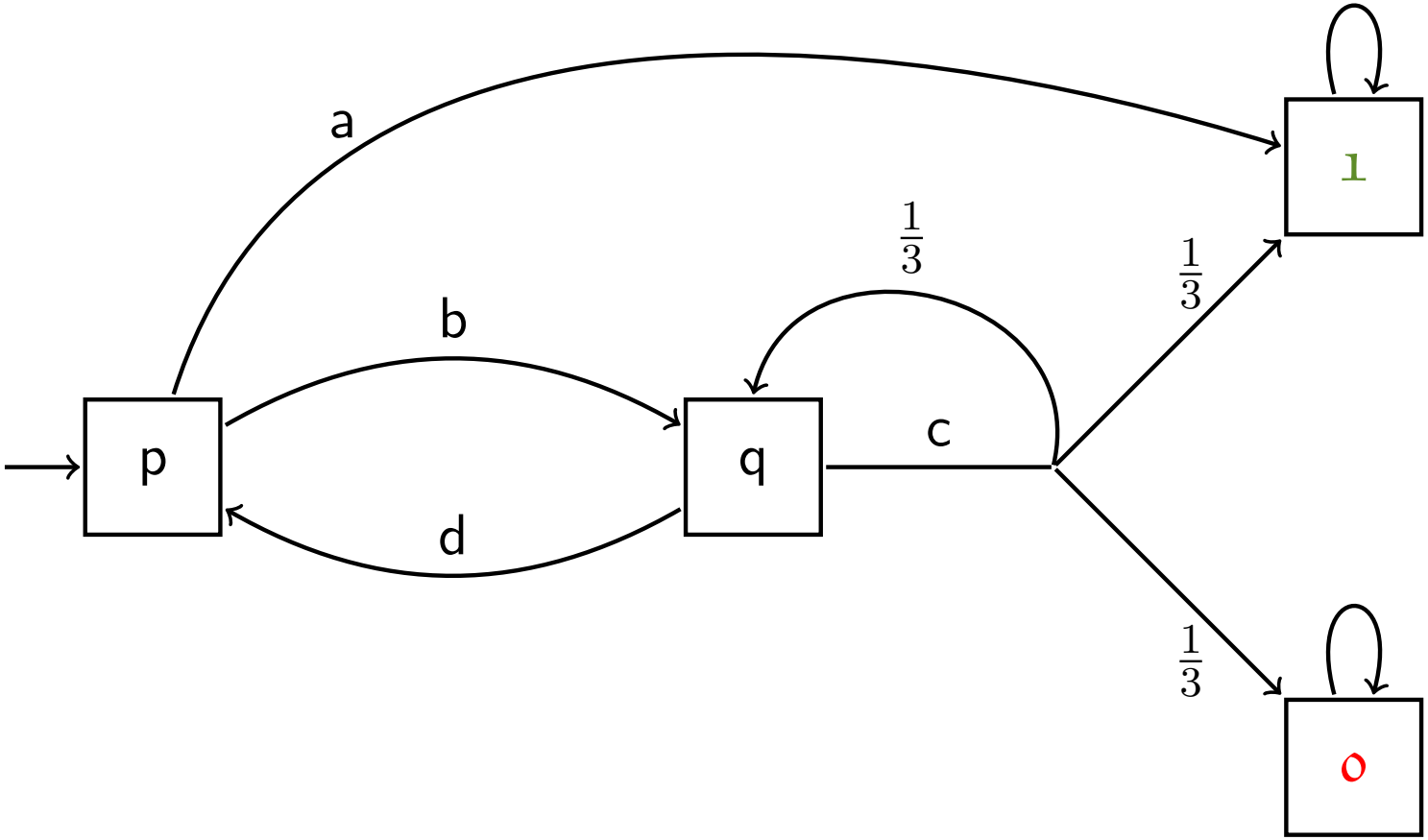
Technische Universität München

06.04.2019

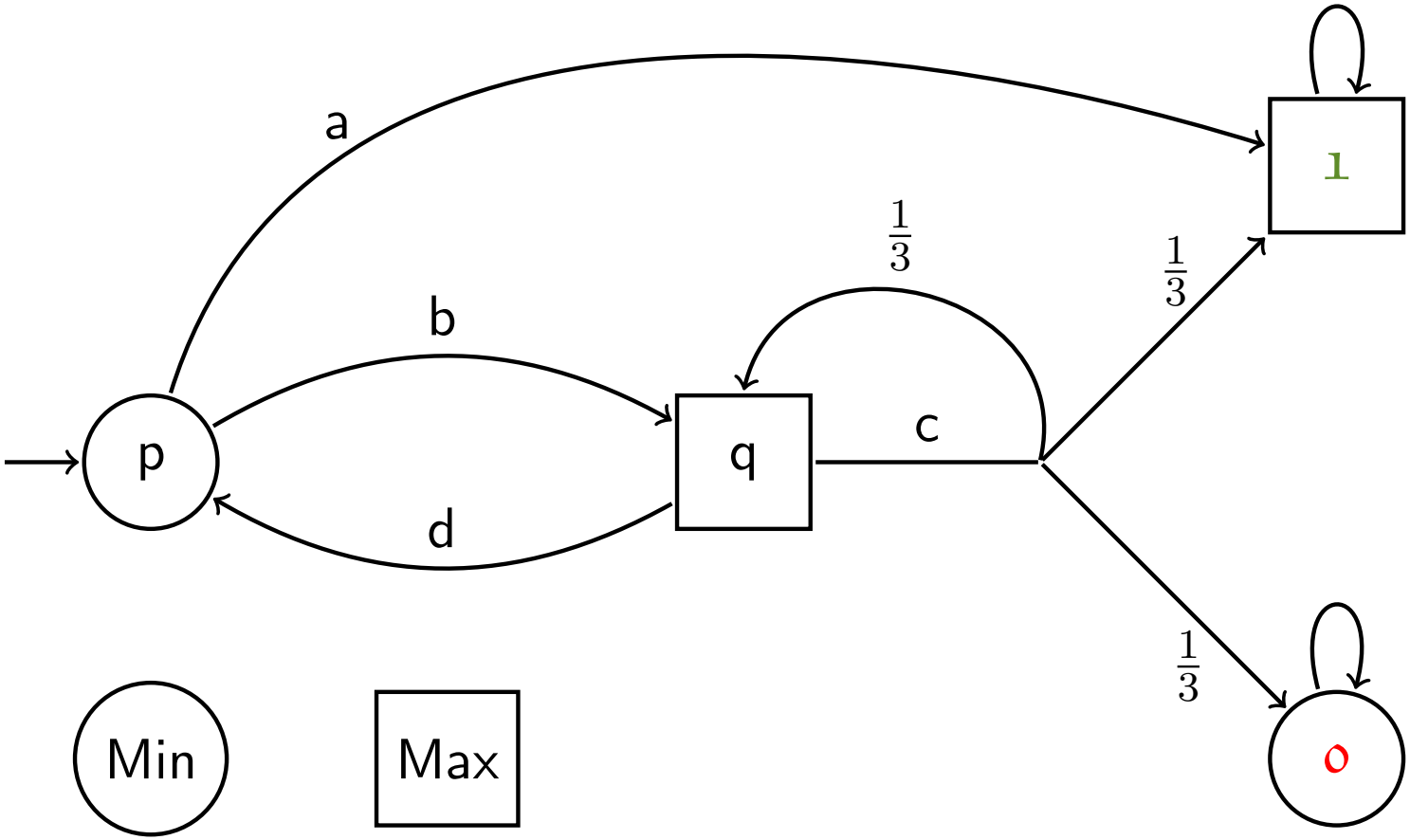
The talk in one slide

- Reachability in stochastic games
- Bounded value iteration
- BRTDP (bounded real-time dynamic programming)
  - Sometimes a lot faster
  - Considers subset of states
  - Guided by bounds

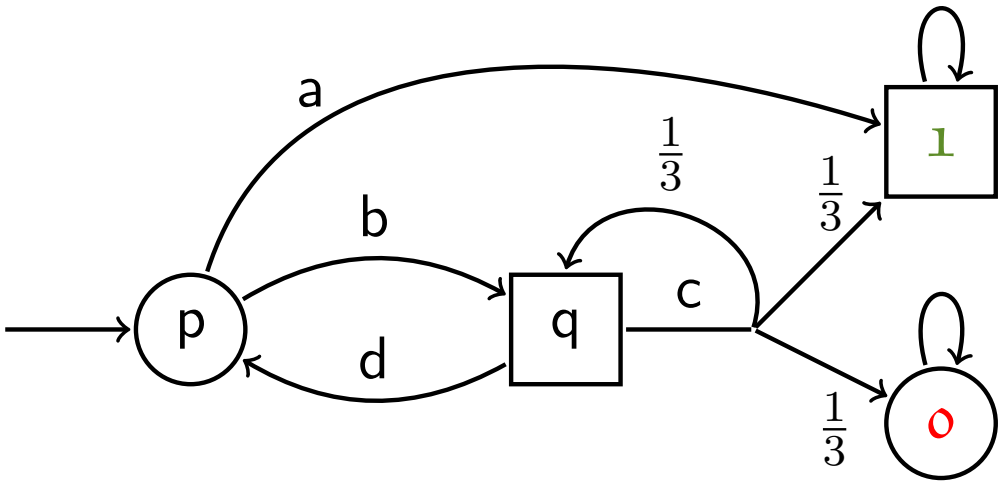
Reachability in stochastic games



Reachability in stochastic games

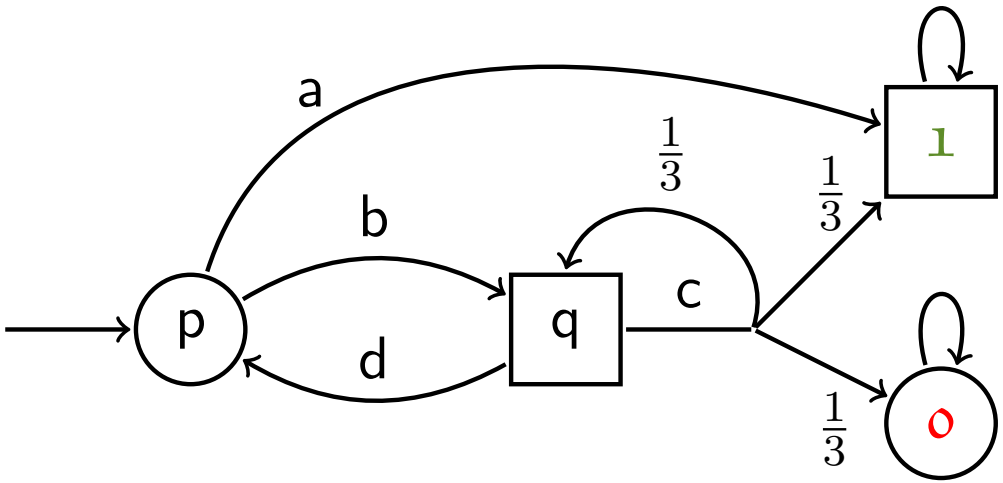


Bounded value iteration



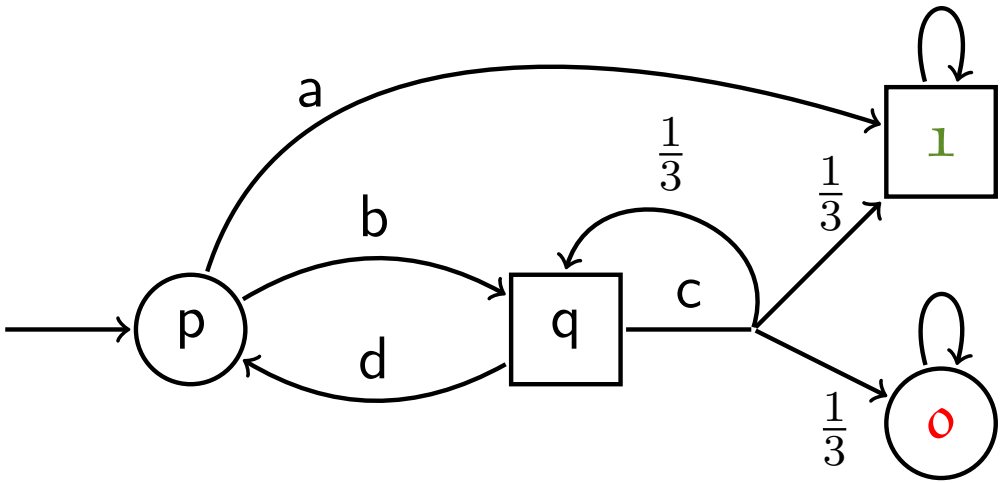
Iteration	L(p)	L(q)	L(1)	L(0)
0	0	0	1	0
1				
2				
3				

Bounded value iteration



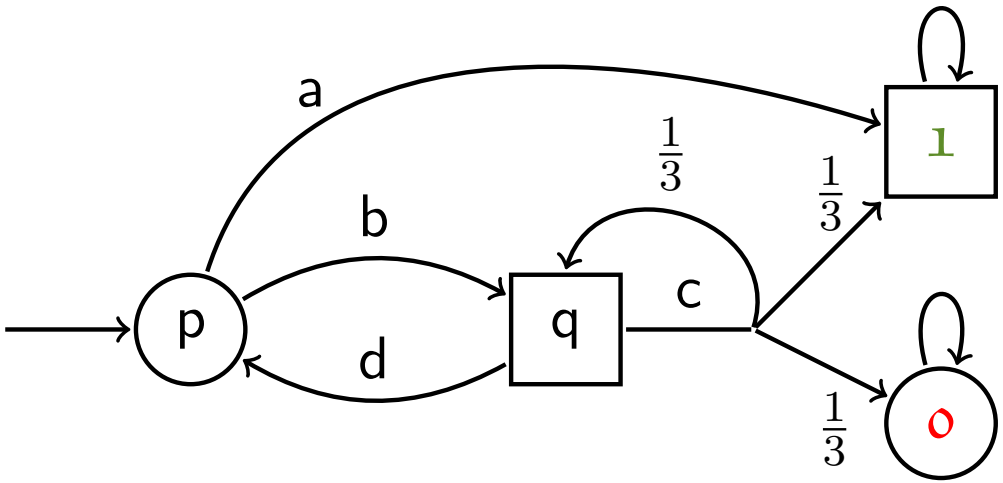
Iteration	L(p)	L(q)	L(1)	L(0)
0	0	0	1	0
1			1	
2			1	
3			1	

Bounded value iteration



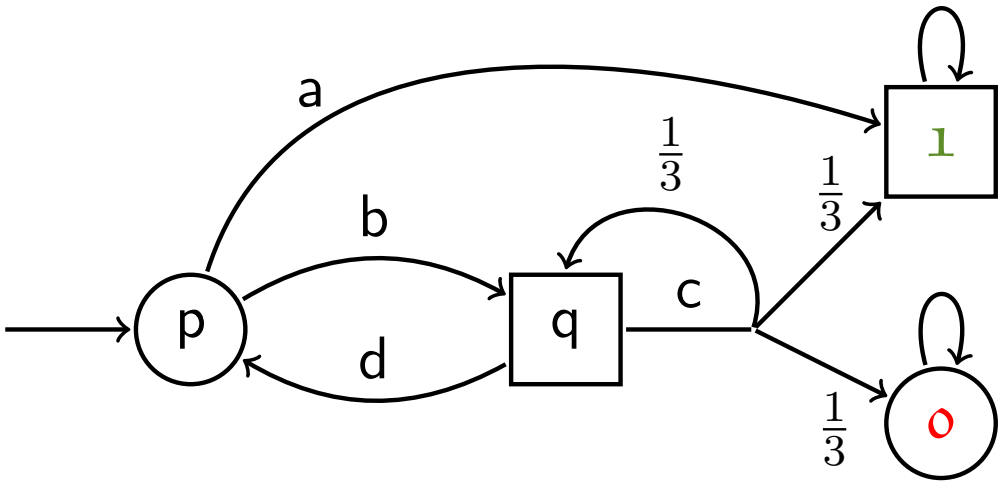
Iteration	L(p)	L(q)	L(1)	L(0)
0	0	0	1	0
1			1	0
2			1	0
3			1	0

Bounded value iteration



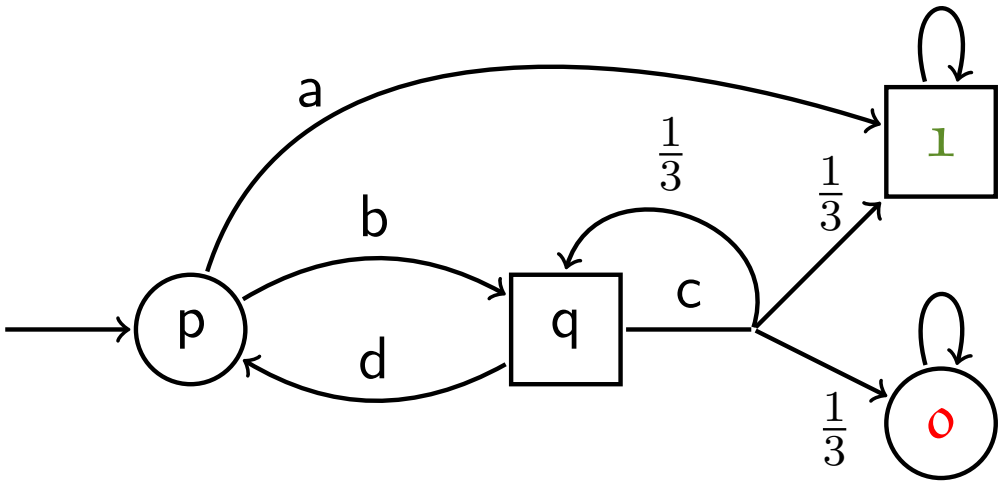
Iteration	L(p)	L(q)	L(1)	L(0)
0	0	0	1	0
1	0		1	0
2			1	0
3			1	0

Bounded value iteration



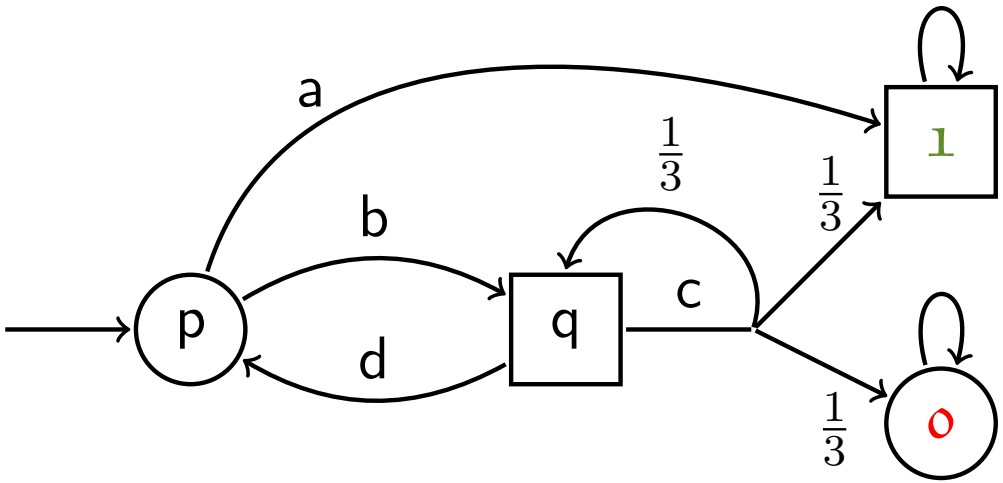
Iteration	L(p)	L(q)	L(1)	L(0)
0	0	0	1	0
1	0	$1/3$	1	0
2			1	0
3			1	0

Bounded value iteration



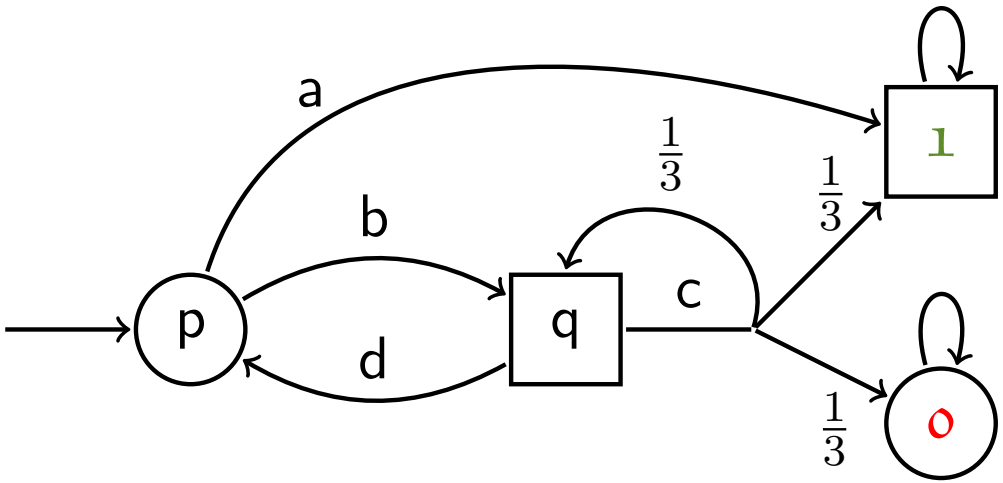
Iteration	L(p)	L(q)	L(1)	L(0)
0	0	0	1	0
1	0	$1/3$	1	0
2	$1/3$		1	0
3			1	0

Bounded value iteration



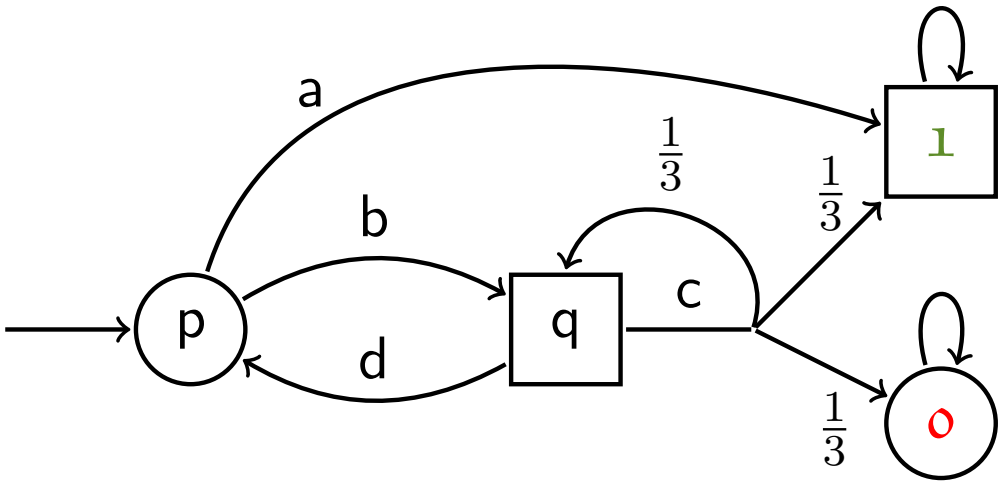
Iteration	L(p)	L(q)	L(1)	L(0)
0	0	0	1	0
1	0	$1/3$	1	0
2	$1/3$	$4/9$	1	0
3			1	0

Bounded value iteration



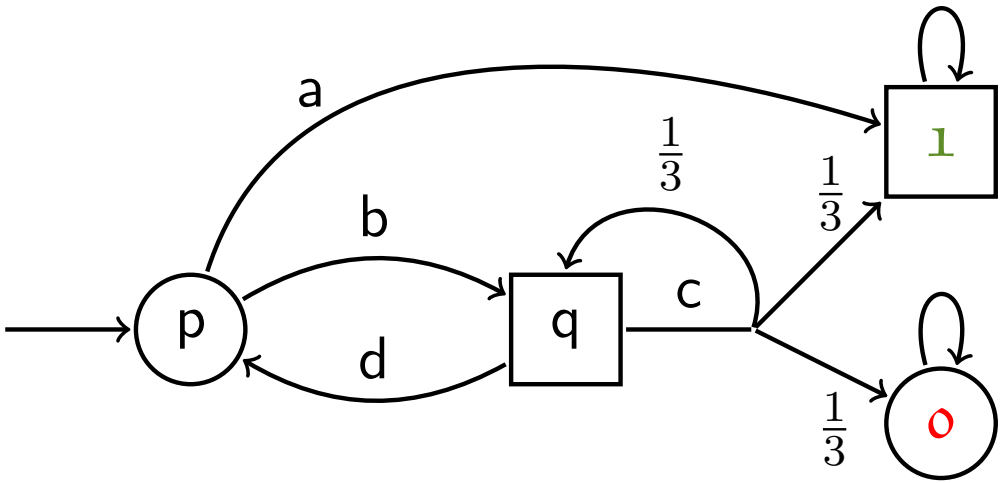
Iteration	L(p)	L(q)	L(1)	L(0)
0	0	0	1	0
1	0	$1/3$	1	0
2	$1/3$	$4/9$	1	0
3	$4/9$		1	0

Bounded value iteration



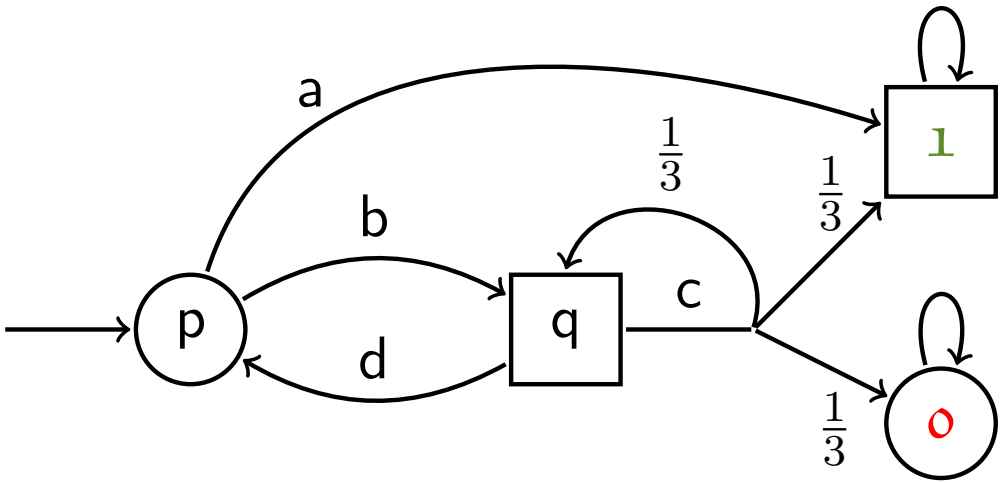
Iteration	L(p)	L(q)	L(1)	L(0)
0	0	0	1	0
1	0	$1/3$	1	0
2	$1/3$	$4/9$	1	0
3	$4/9$	$13/27$	1	0

Bounded value iteration



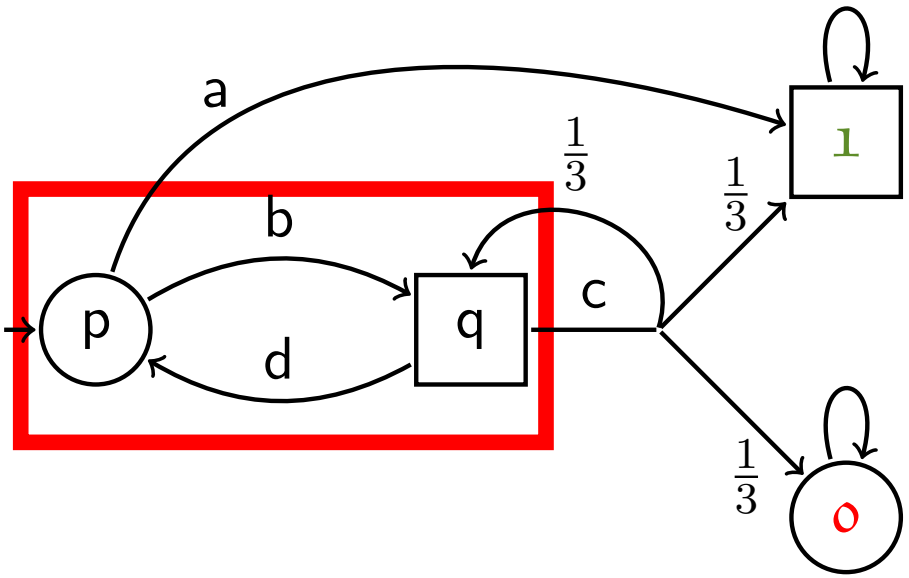
Iteration	L(p)	L(q)	L(1)	L(0)	U(p)	U(q)
0	0	0	1	0	1	1
1	0	$1/3$	1	0		
2	$1/3$	$4/9$	1	0		
3	$4/9$	$13/27$	1	0		

Bounded value iteration

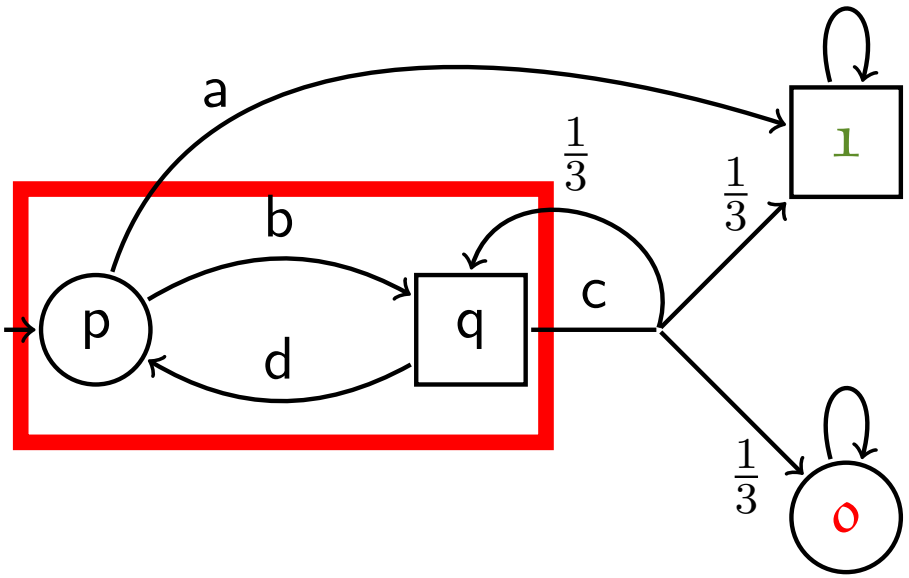


Iteration	L(p)	L(q)	L(1)	L(0)	U(p)	U(q)
0	0	0	1	0	1	1
1	0	$1/3$	1	0	1	1
2	$1/3$	$4/9$	1	0	1	1
3	$4/9$	$13/27$	1	0	1	1

The problem of **end components**



The problem of **end components**

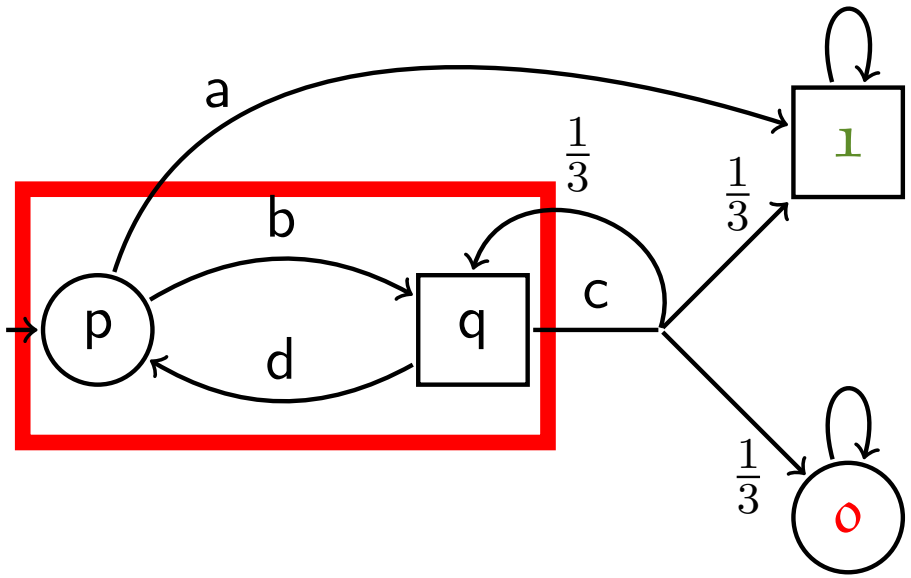


Deflating

For all states in EC:

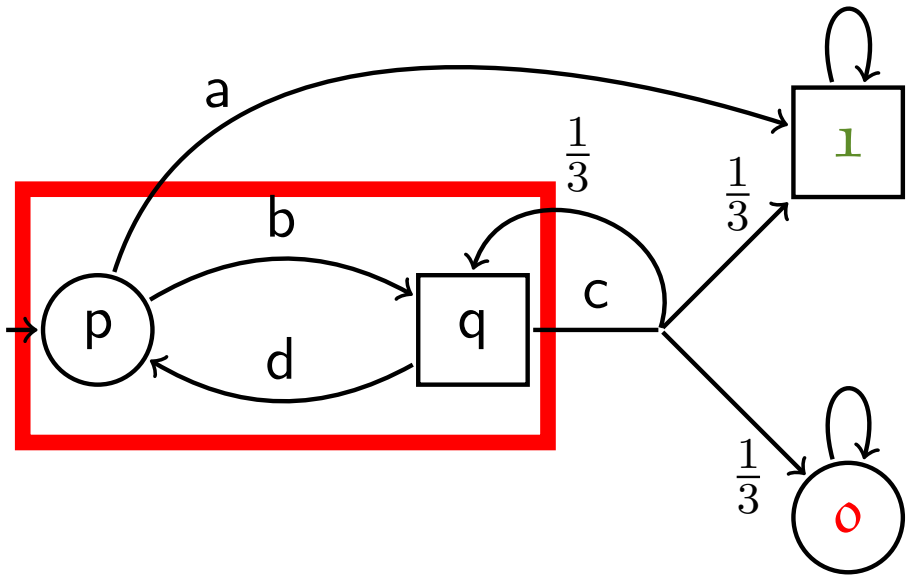
Decrease U to **U(best-□-exit)**

Deflating end components



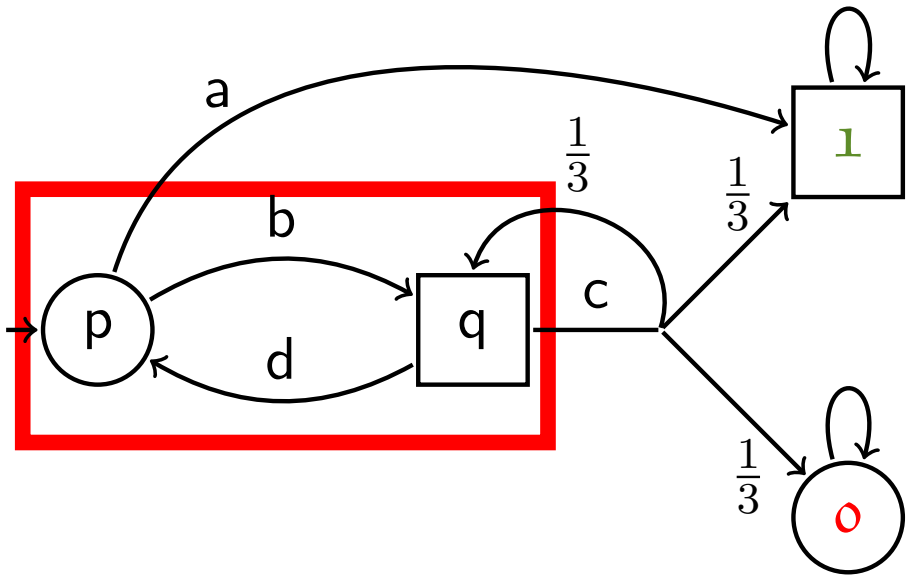
Iteration	L(p)	L(q)	L(1)	L(0)	U(p)	U(q)
0	0	0	1	0	1	1
1	0	$1/3$	1	0		
2	$1/3$	$4/9$	1	0		
3	$4/9$	$13/27$	1	0		

Deflating end components



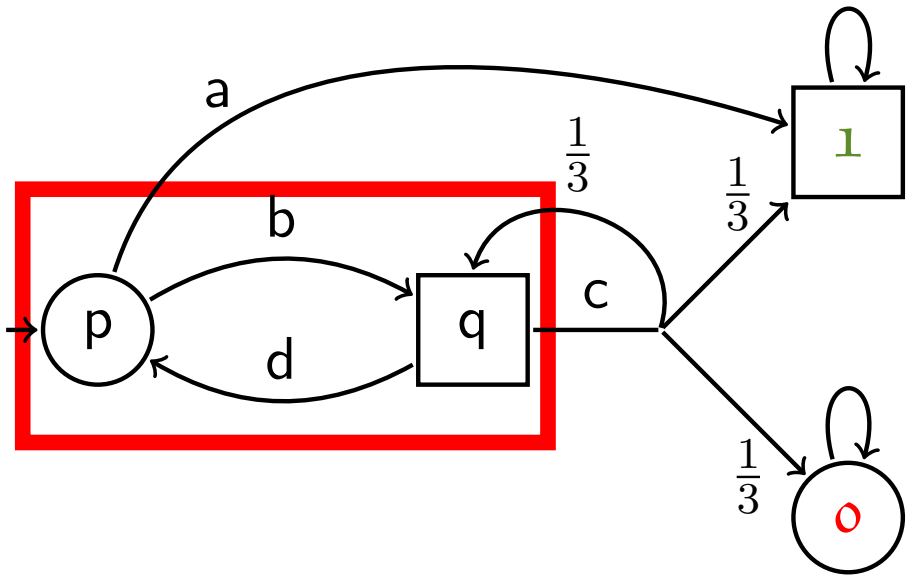
Iteration	L(p)	L(q)	L(1)	L(0)	U(p)	U(q)
0	0	0	1	0	1	1
1	0	$1/3$	1	0	$2/3$	$2/3$
2	$1/3$	$4/9$	1	0		
3	$4/9$	$13/27$	1	0		

Deflating end components



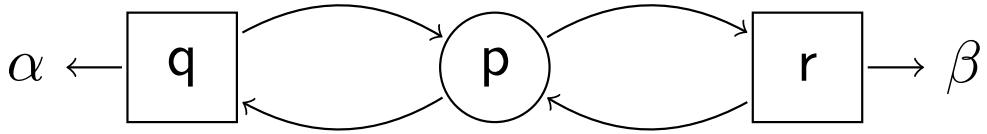
Iteration	L(p)	L(q)	L(1)	L(0)	U(p)	U(q)
0	0	0	1	0	1	1
1	0	$1/3$	1	0	$2/3$	$2/3$
2	$1/3$	$4/9$	1	0	$5/9$	$5/9$
3	$4/9$	$13/27$	1	0		

Deflating end components

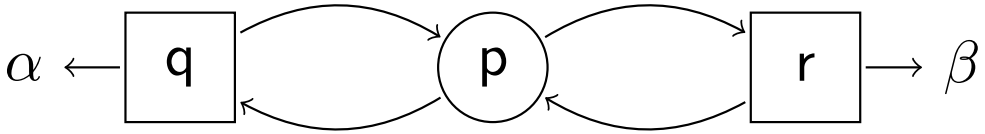


Iteration	L(p)	L(q)	L(1)	L(0)	U(p)	U(q)
0	0	0	1	0	1	1
1	0	$1/3$	1	0	$2/3$	$2/3$
2	$1/3$	$4/9$	1	0	$5/9$	$5/9$
3	$4/9$	$13/27$	1	0	$14/27$	$14/27$

The final problem: Simple end components

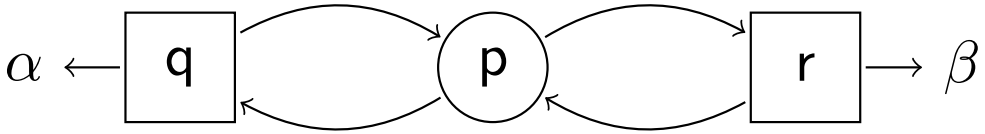


The final problem: Simple end components



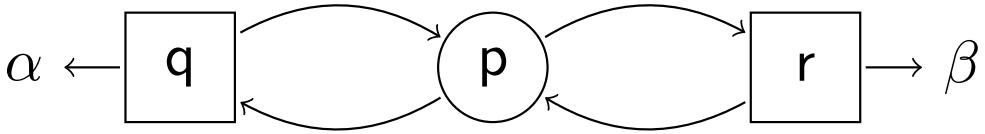
$\min(\alpha, \beta)$	α	β	α, β
SEC			

The final problem: Simple end components



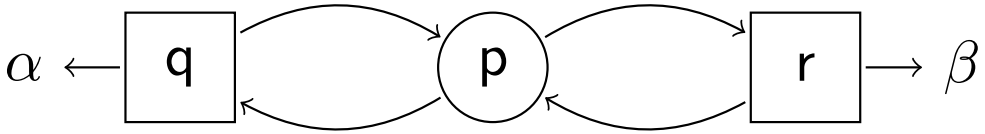
$\min(\alpha, \beta)$	α	β	α, β
SEC	{p,q}		

The final problem: Simple end components



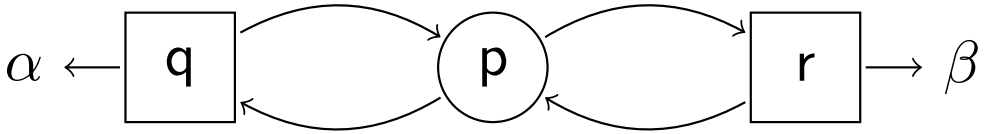
$\min(\alpha, \beta)$	α	β	α, β
SEC	{p,q}	{p,r}	

The final problem: Simple end components



$\min(\alpha, \beta)$	α	β	α, β
SEC	{p,q}	{p,r}	{p,q,r}

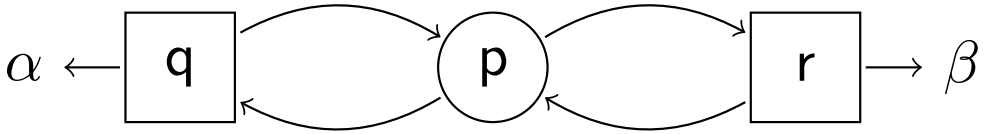
The final problem: Simple end components



$\min(\alpha, \beta)$	α	β	α, β
SEC	{p,q}	{p,r}	{p,q,r}

How do we know the minimal value?

The final problem: Simple end components



$\min(\alpha, \beta)$	α	β	α, β
SEC	{p,q}	{p,r}	{p,q,r}

How do we know the minimal value? Guess it according to L.

Bounded value iteration - full picture

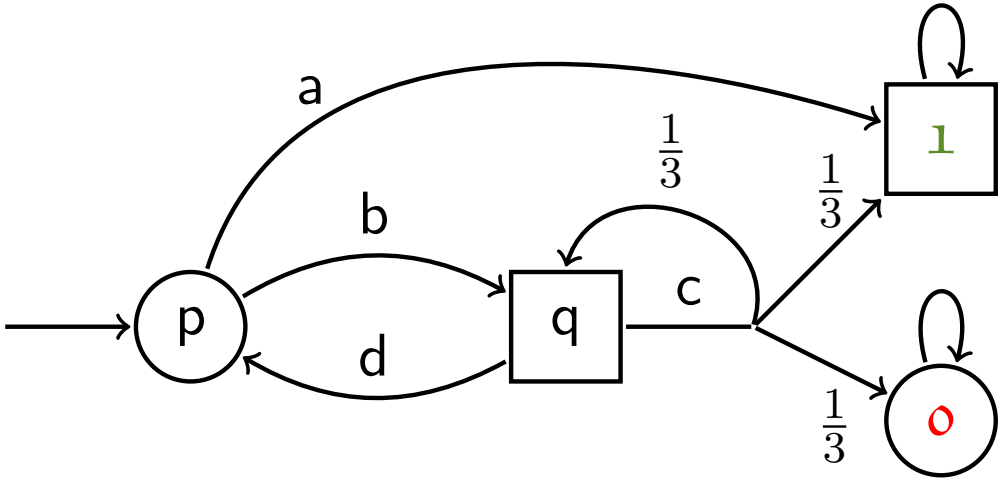
- Compute probability to reach target in SG
- How:
  - Lower and upper bound
  - Start at 0/1
  - Iterative backpropagation
  - Deflate SECs on the fly

Bounded value iteration - full picture

- Compute probability to reach target in SG
- How:
  - Lower and upper bound
  - Start at 0/1
  - Iterative backpropagation
  - Deflate SECs on the fly

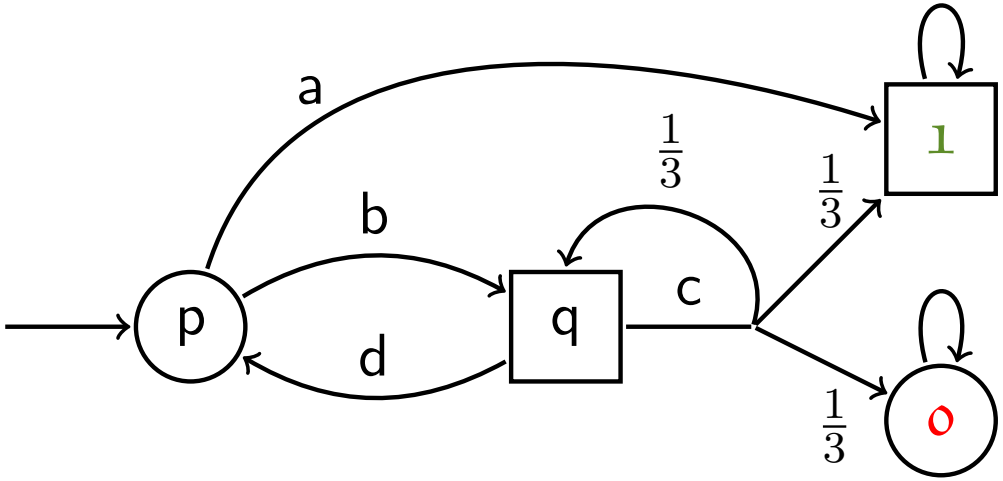
Convergent anytime algorithm with guarantees

Anytime algorithm with guarantees?



Iteration	p	q	1	0
0	[0,1]	[0,1]	[1,1]	[0,0]
1	$[0,\frac{2}{3}]$	$[\frac{1}{3},\frac{2}{3}]$	[1,1]	[0,0]
2	$[\frac{1}{3},\frac{5}{9}]$	$[\frac{4}{9},\frac{5}{9}]$	[1,1]	[0,0]

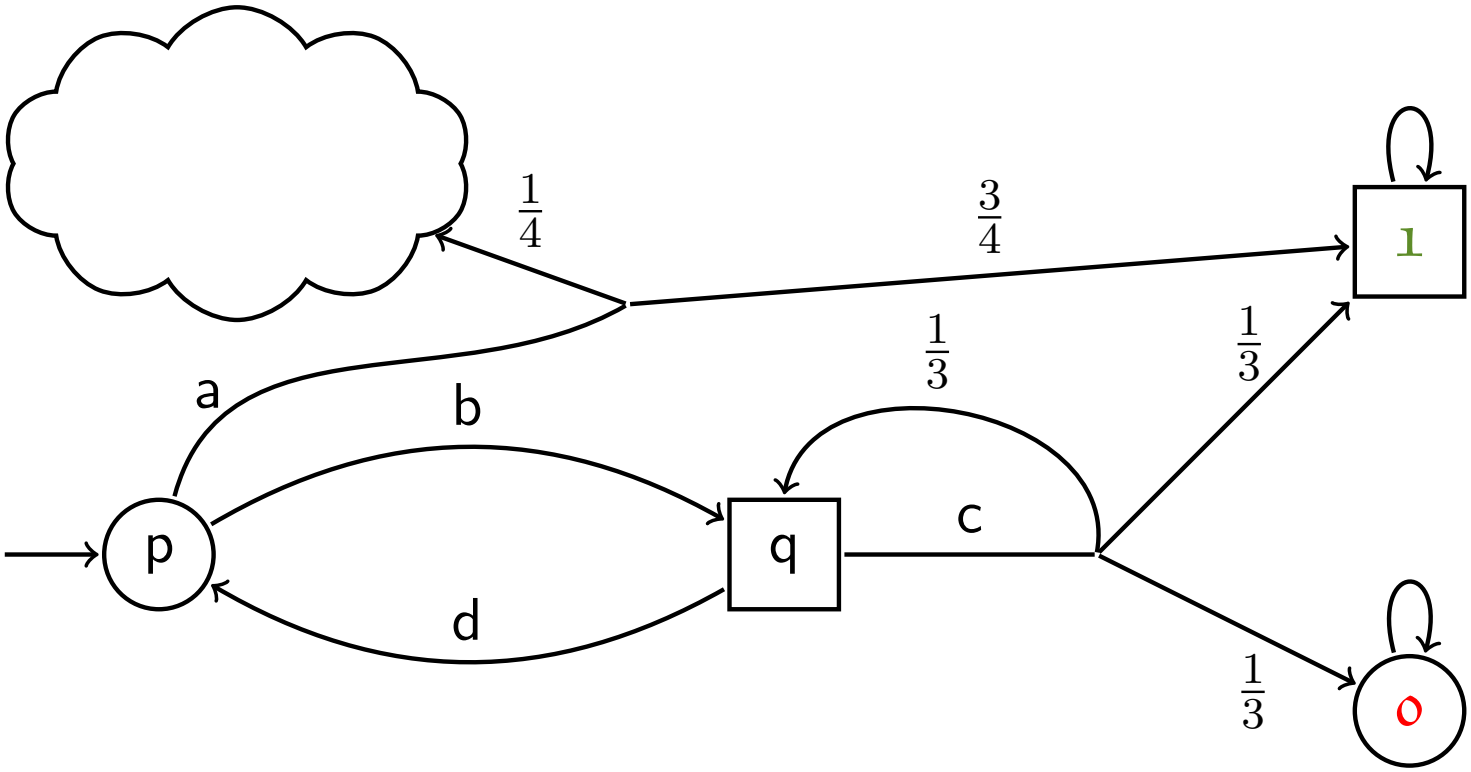
Anytime algorithm with guarantees?



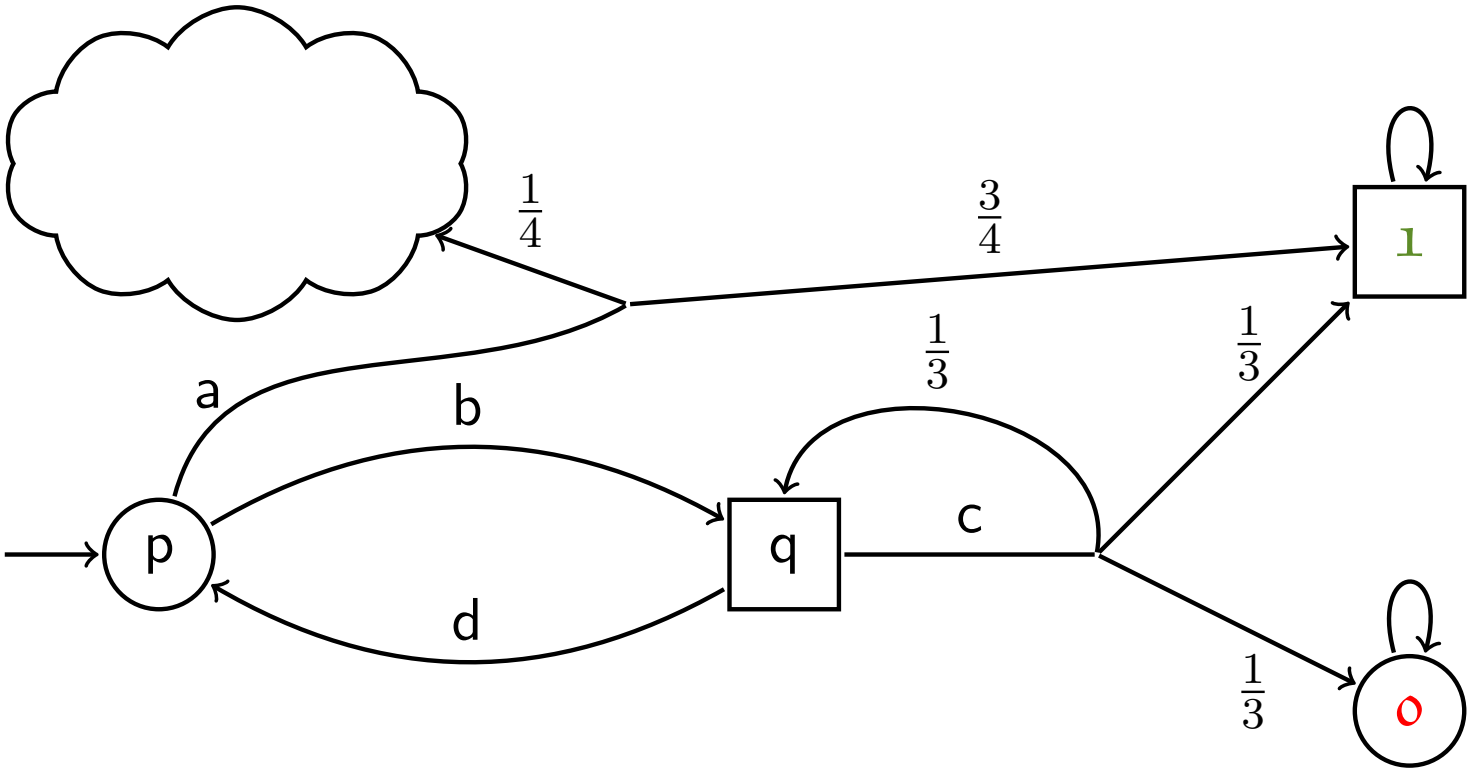
Iteration	p	q	1	0
0	[0,1]	[0,1]	[1,1]	[0,0]
1	$[0,\frac{2}{3}]$	$[\frac{1}{3},\frac{2}{3}]$	[1,1]	[0,0]
2	$[\frac{1}{3},\frac{5}{9}]$	$[\frac{4}{9},\frac{5}{9}]$	[1,1]	[0,0]

Selectively update bounds, guided by estimates and precision

BRTDP by example

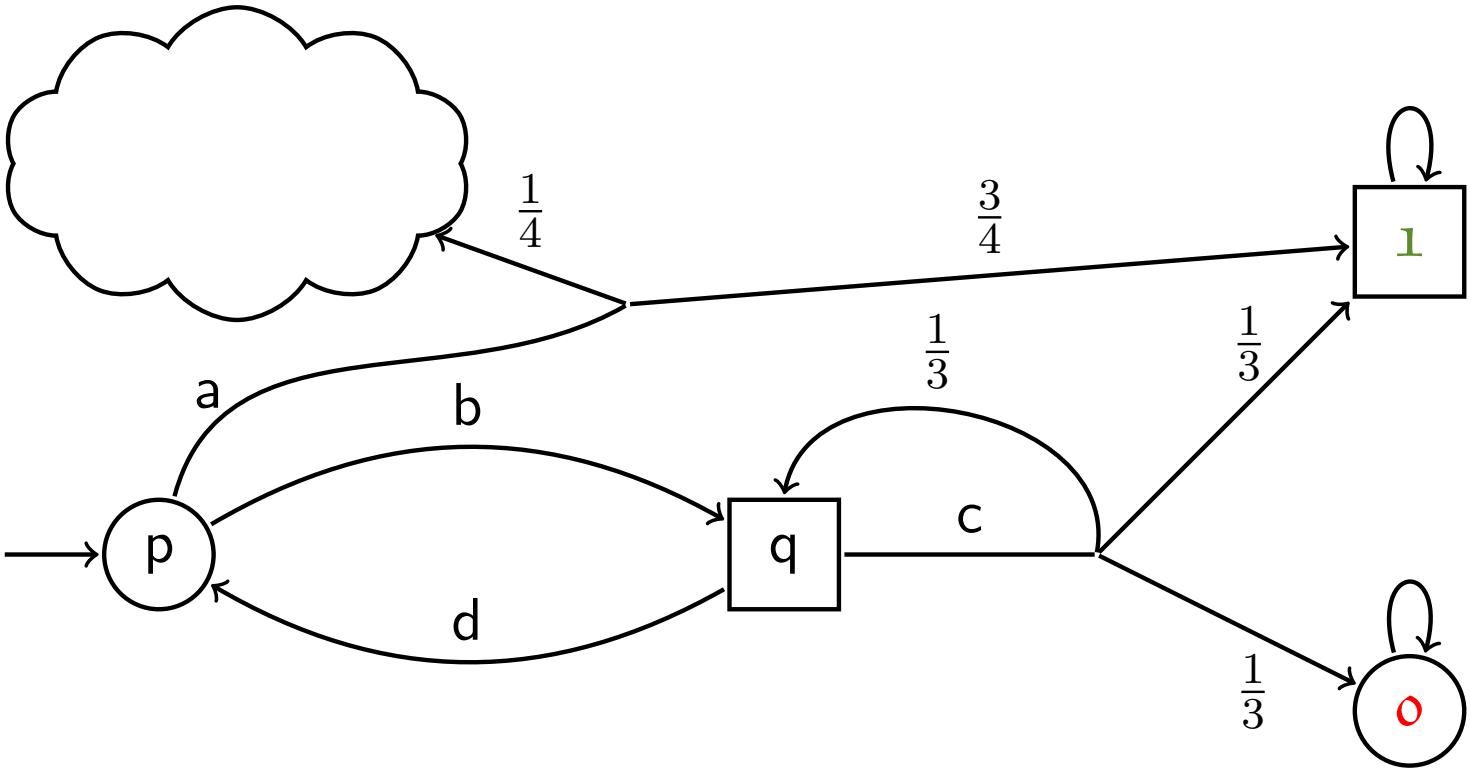


BRTDP by example



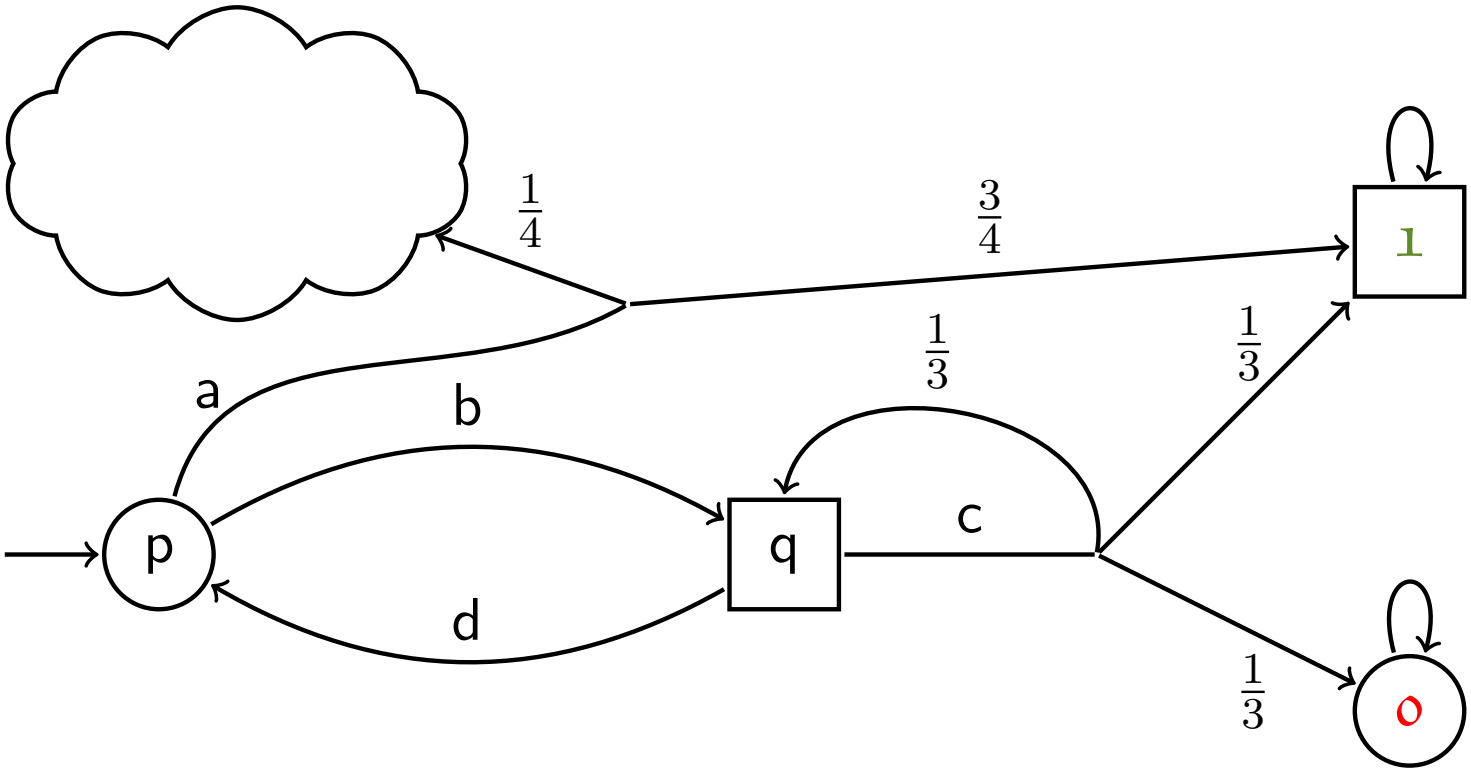
$$L(p,a) = \frac{3}{4}$$

BRTDP by example



$$L(p,a) = \frac{3}{4} \qquad U(p,b) = \frac{2}{3}$$

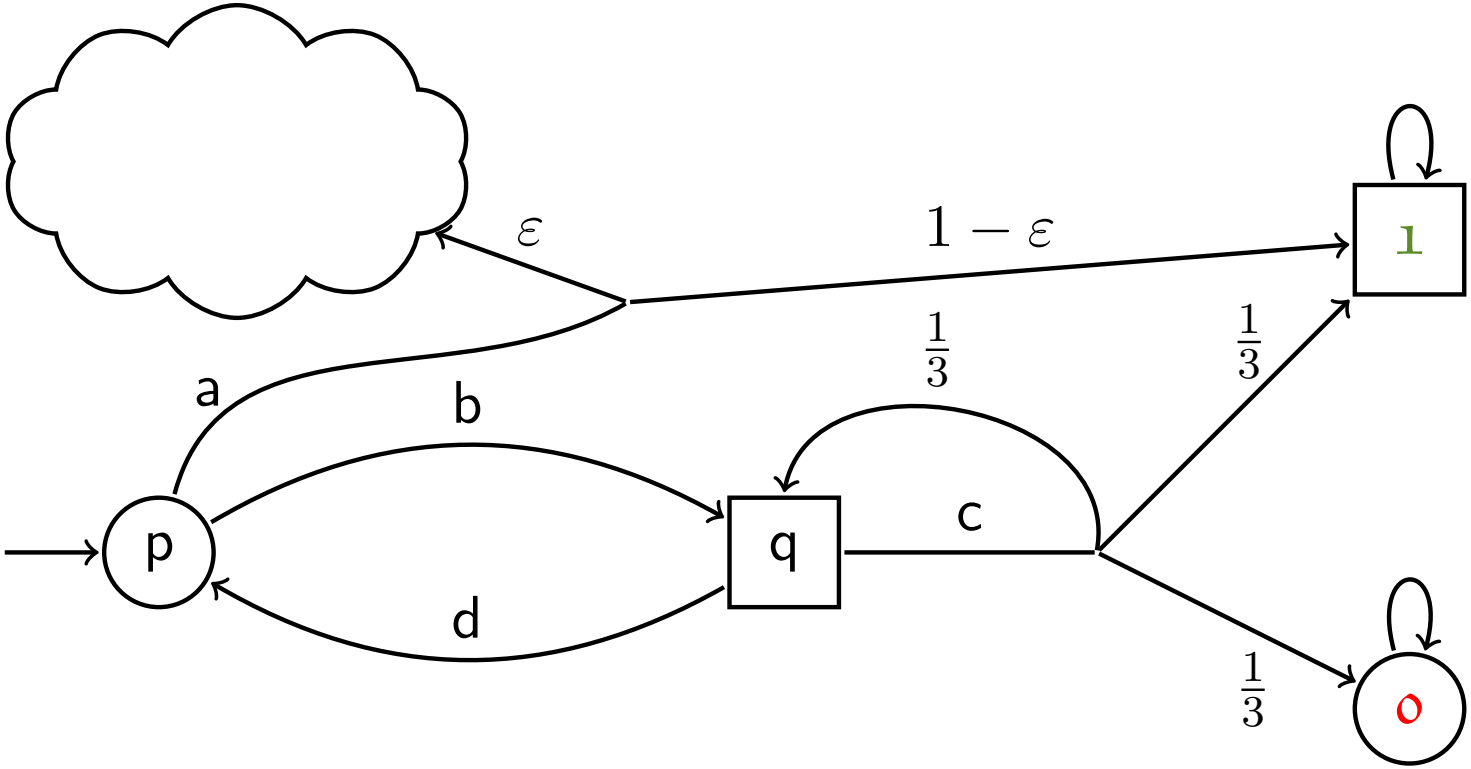
BRTDP by example



$$L(p,a) = \frac{3}{4} \qquad U(p,b) = \frac{2}{3}$$

$\longrightarrow$ **No need to explore cloud**

BRTDP by example



No need to explore cloud for ϵ-optimality

BRTDP algorithm

Simulate path:

- Pick "best" actions
- Pick "interesting" successors

Update bounds:

- When reaching target/sink, backpropagate
- **When stuck in EC, deflate**

BRTDP strengths and weaknesses

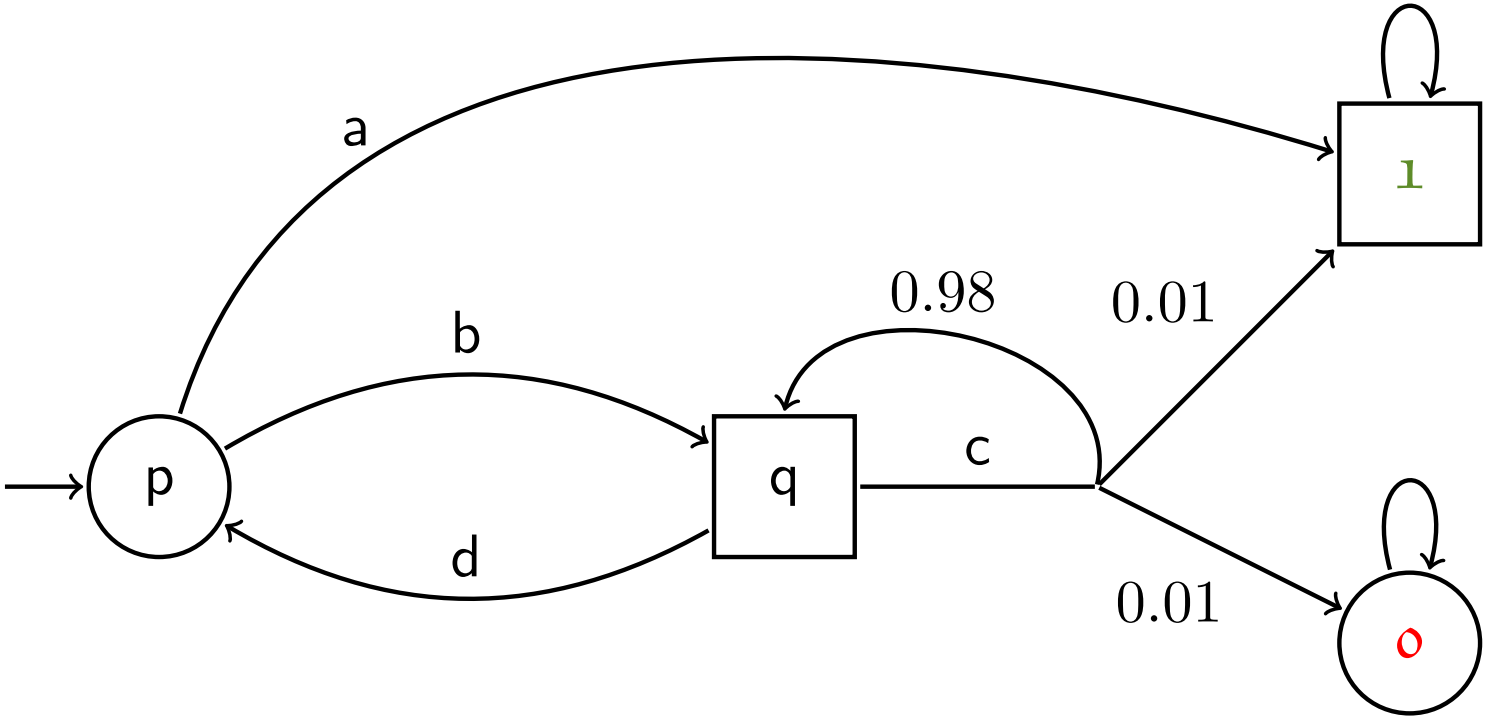
Works well, if

- actions can be pruned.
- reachability probabilities become smaller than ϵ.

BRTDP strengths and weaknesses

Works well, if

- actions can be pruned.
- reachability probabilities become smaller than ϵ.
- stuck-detection does **not** have many false positives.



Experimental results

Model	PRISM	BVI	BRTDP
cloud	6s	59s	4s
mdsm	8s	8s	17s
zeroconf	7s	24s	3s
csma	2s	4s	86s

Experimental results

Model	PRISM	BVI	BRTDP
cloud	**6s**	59s	**4s**
mdsm	**8s**	**8s**	17s
zeroconf	**7s**	24s	**3s**
csma	**2s**	**4s**	86s

Experimental results

Model	PRISM	BVI	BRTDP
cloud	6s	59s	4s
mdsm	8s	8s	17s
zeroconf	7s	24s	3s
csma	2s	4s	86s

Conclusion

- Convergent guaranteed anytime algorithm for stochastic games
- Exploiting bounds for BRTDP
- Great speedup on some models

Conclusion

- Convergent guaranteed anytime algorithm for stochastic games
- Exploiting bounds for BRTDP
- Great speedup on some models

Future work:

- Other objectives
- Limited information setting
- Other learning algorithms