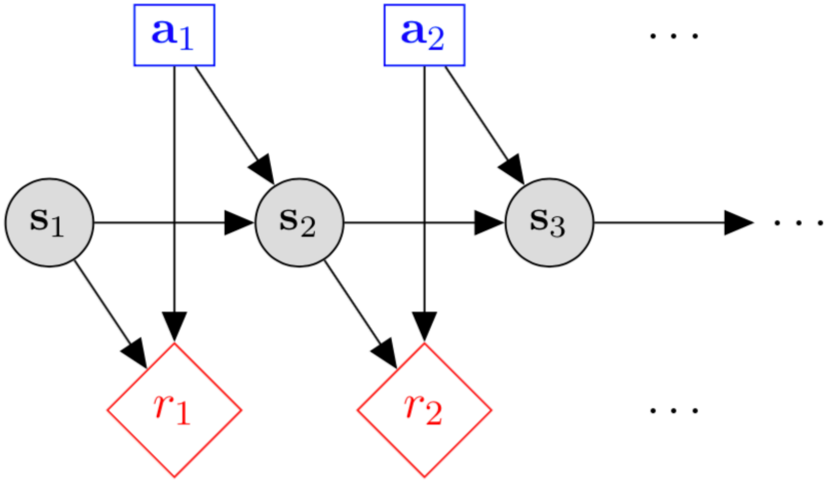
CS181: Introduction to Machine Learning

Lecture 21 (MDPs and RL)

Spring 2021

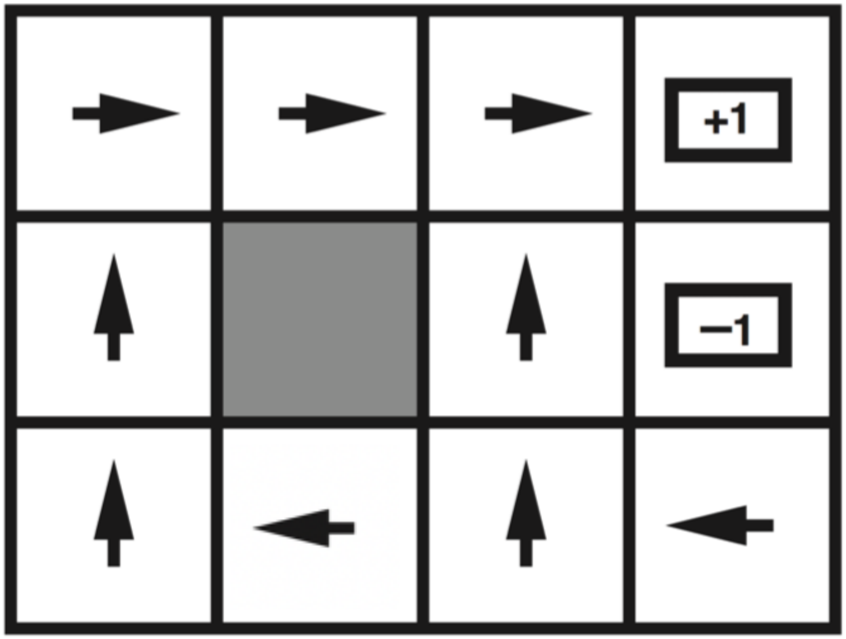
Finale Doshi-Velez and David C. Parkes
Harvard Computer Science

Example: House cleaning robot



- States: physical location, objects in environment
- Actions: move, pick-up, drop, ...
- Reward: +1 if pick up dirty clothes, -1 if break dish, ...
- Transition model: describe actuators and uncertain environment

GridWorld

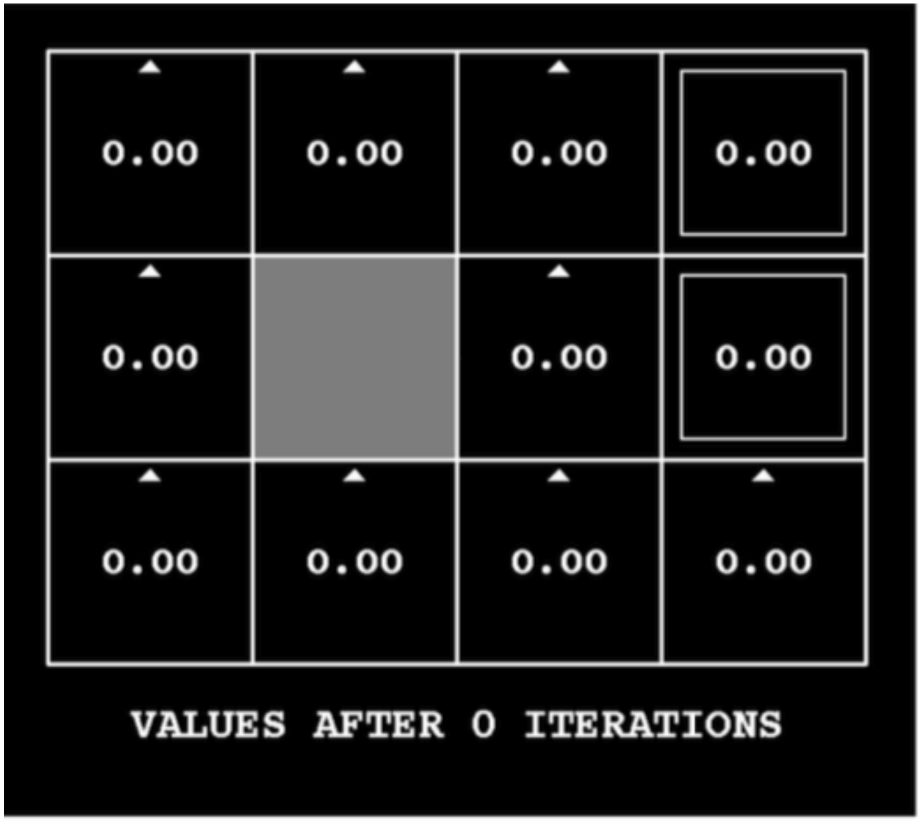


(D. Klein, P. Abbeel)

- r(s,a) = 0, except states (1,4), (2,4). In these states get +1 or -1 when take ANY action. Then no more actions
- Bounce off obstacles. Actuator has 20% noise; e.g., w/ prob 0.1 goes L, prob 0.1 goes R when moving U
- Discounting 0.9 (r + 0.9 r + 0.9^2 r + …)

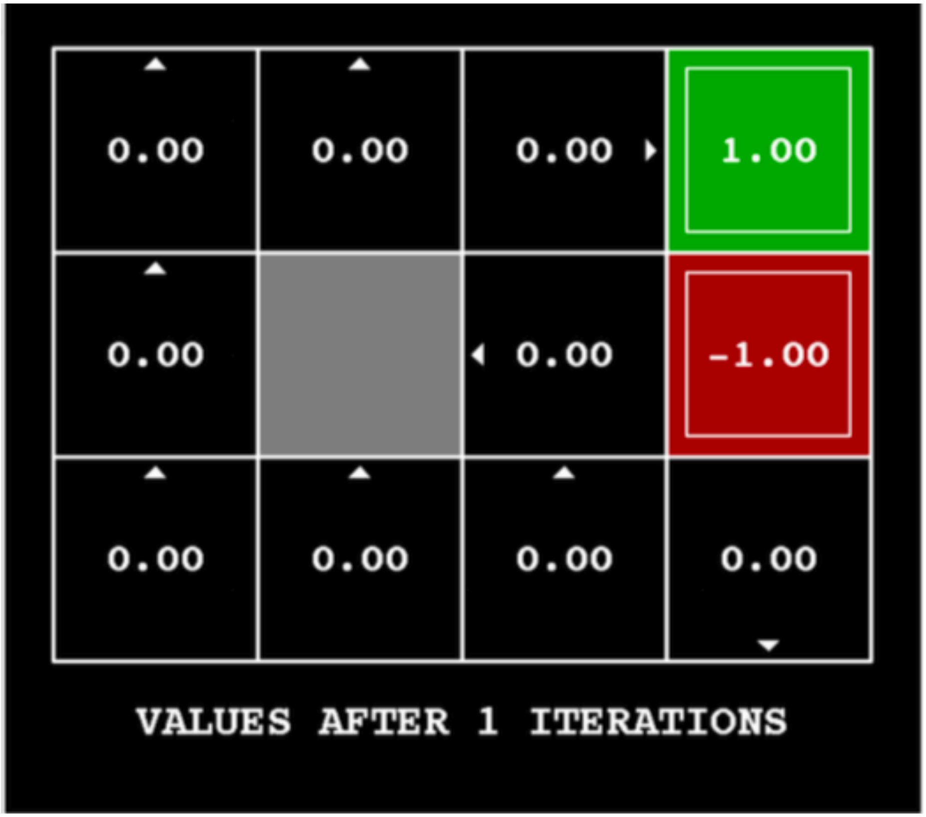
VI, GridWorld

$$V'(s) \leftarrow \max_{a \in A} \left[r(s,a) + \gamma \sum_{s' \in S} p(s' \mid s,a) V(s') \right], \quad \forall s$$



▲ 0.00	▲ 0.00	▲ 0.00	0.00
▲ 0.00	▨	▲ 0.00	0.00
▲ 0.00	▲ 0.00	▲ 0.00	▲ 0.00

VALUES AFTER 0 ITERATIONS



▲ 0.00	▲ 0.00	0.00 ▶	1.00
▲ 0.00	▨	◀ 0.00	-1.00
▲ 0.00	▲ 0.00	▲ 0.00	0.00 ▼

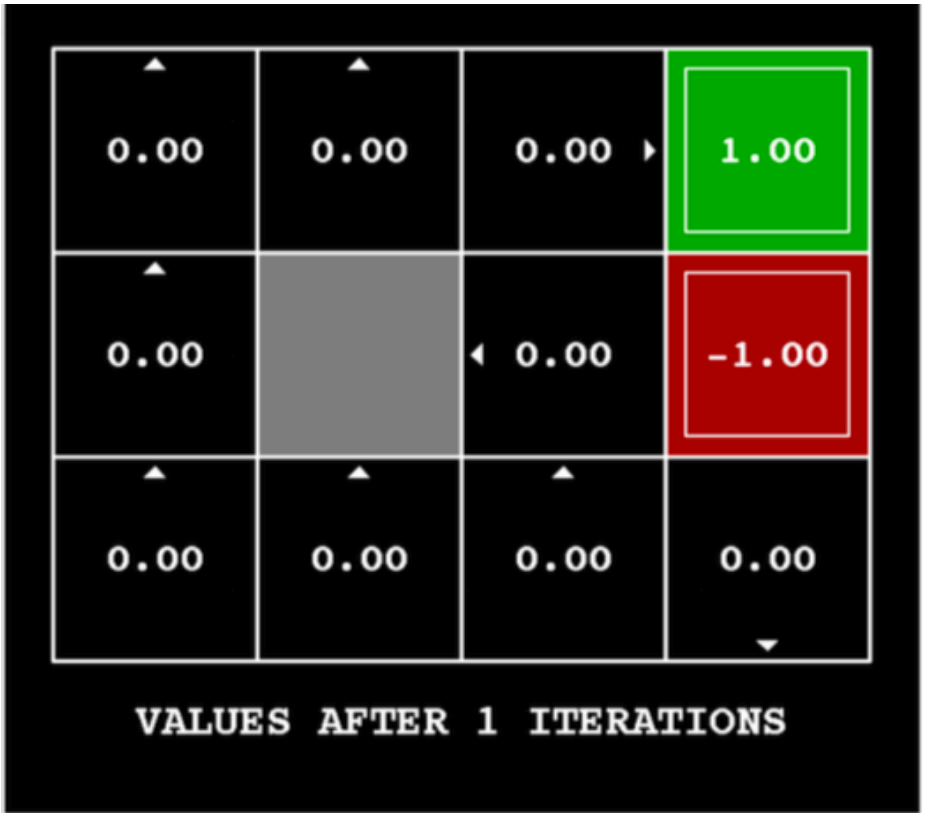
VALUES AFTER 1 ITERATIONS

(D. Klein, P. Abbeel)

Noise 0.2, Discount 0.9

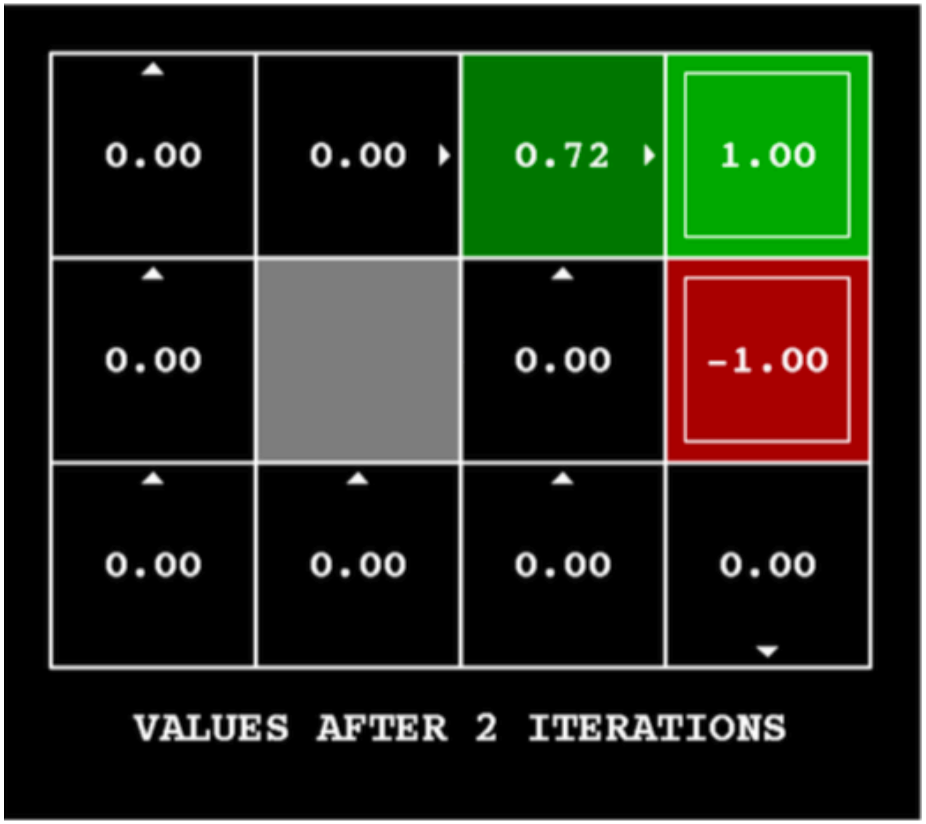
VI, GridWorld

$$V'(s) \leftarrow \max_{a\in A}\left[r(s,a) + \gamma \sum_{s'\in S} p(s' \mid s,a)V(s')\right], \quad \forall s$$



▲ 0.00	▲ 0.00	0.00 ▶	1.00
▲ 0.00		◀ 0.00	-1.00
▲ 0.00	▲ 0.00	▲ 0.00	0.00 ▼

VALUES AFTER 1 ITERATIONS



▲ 0.00	0.00 ▶	0.72 ▶	1.00
▲ 0.00		▲ 0.00	-1.00
▲ 0.00	▲ 0.00	▲ 0.00	0.00 ▼

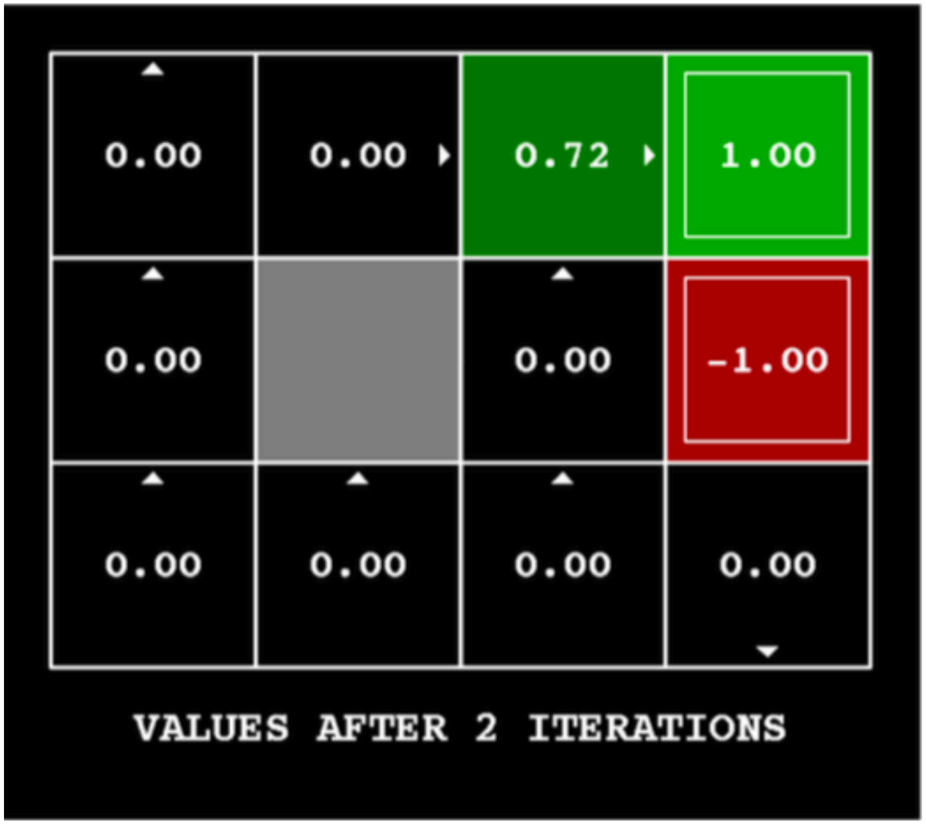
VALUES AFTER 2 ITERATIONS

(D. Klein, P. Abbeel)

Noise 0.2, Discount 0.9

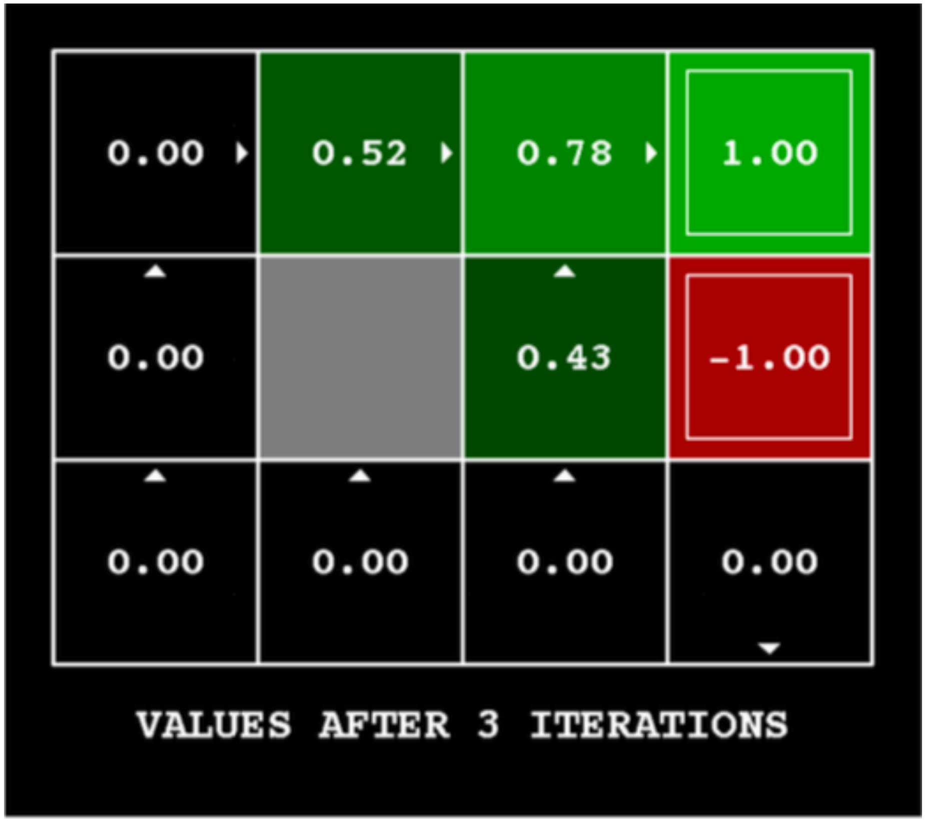
VI, GridWorld

$$V'(s) \leftarrow \max_{a \in A} \left[r(s,a) + \gamma \sum_{s' \in S} p(s' \mid s,a) V(s') \right], \quad \forall s$$



▲ 0.00	0.00 ▶	0.72 ▶	1.00
▲ 0.00		▲ 0.00	-1.00
▲ 0.00	▲ 0.00	▲ 0.00	0.00 ▼

VALUES AFTER 2 ITERATIONS



0.00 ▶	0.52 ▶	0.78 ▶	1.00
▲ 0.00		▲ 0.43	-1.00
▲ 0.00	▲ 0.00	▲ 0.00	0.00 ▼

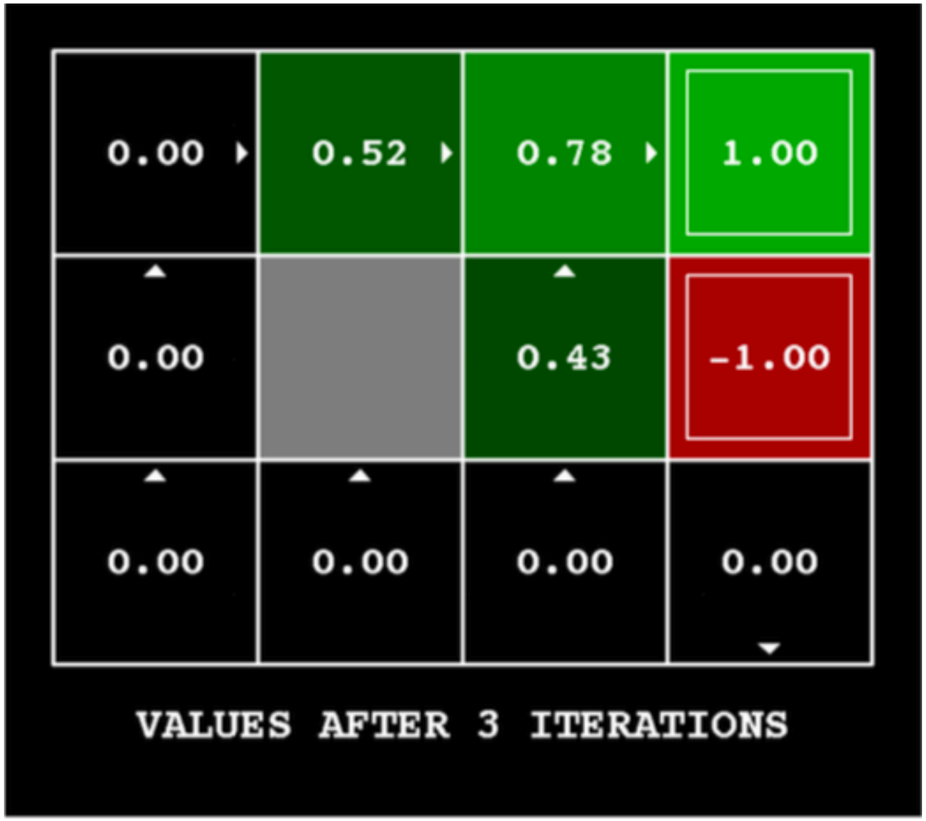
VALUES AFTER 3 ITERATIONS

(D. Klein, P. Abbeel)

Noise 0.2, Discount 0.9

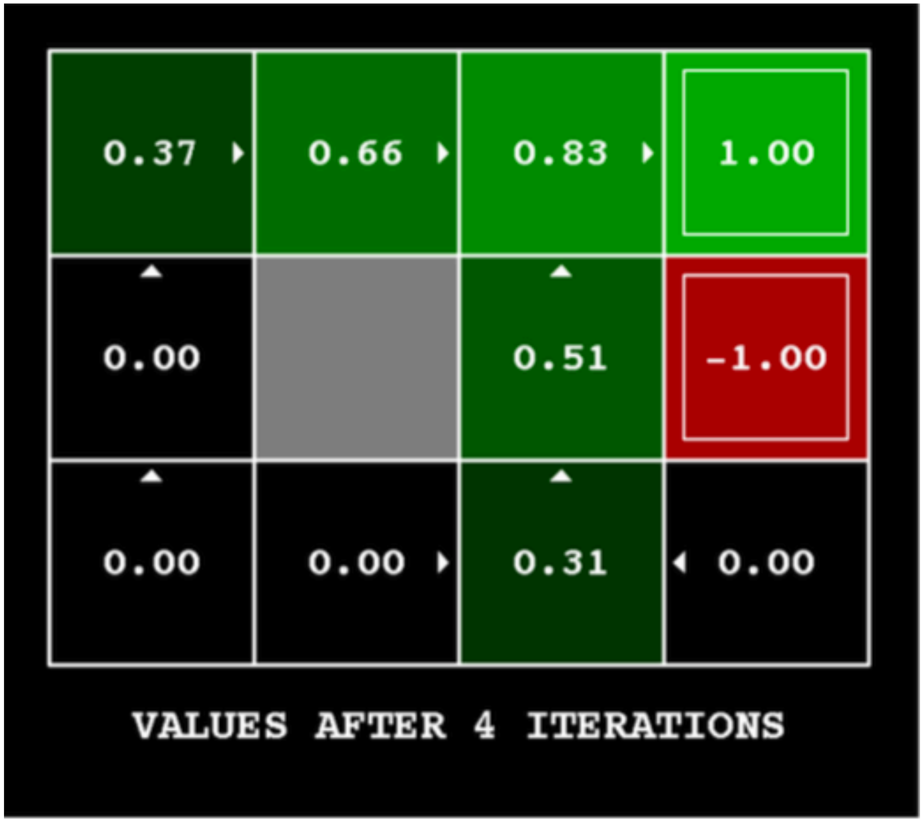
VI, GridWorld

$$V'(s) \leftarrow \max_{a \in A} \left[r(s,a) + \gamma \sum_{s' \in S} p(s' \mid s,a) V(s') \right], \quad \forall s$$



0.00 ▸	0.52 ▸	0.78 ▸	1.00
▴ 0.00		▴ 0.43	-1.00
▴ 0.00	▴ 0.00	▴ 0.00	0.00 ▾

VALUES AFTER 3 ITERATIONS



0.37 ▸	0.66 ▸	0.83 ▸	1.00
▴ 0.00		▴ 0.51	-1.00
▴ 0.00	0.00 ▸	▴ 0.31	◂ 0.00

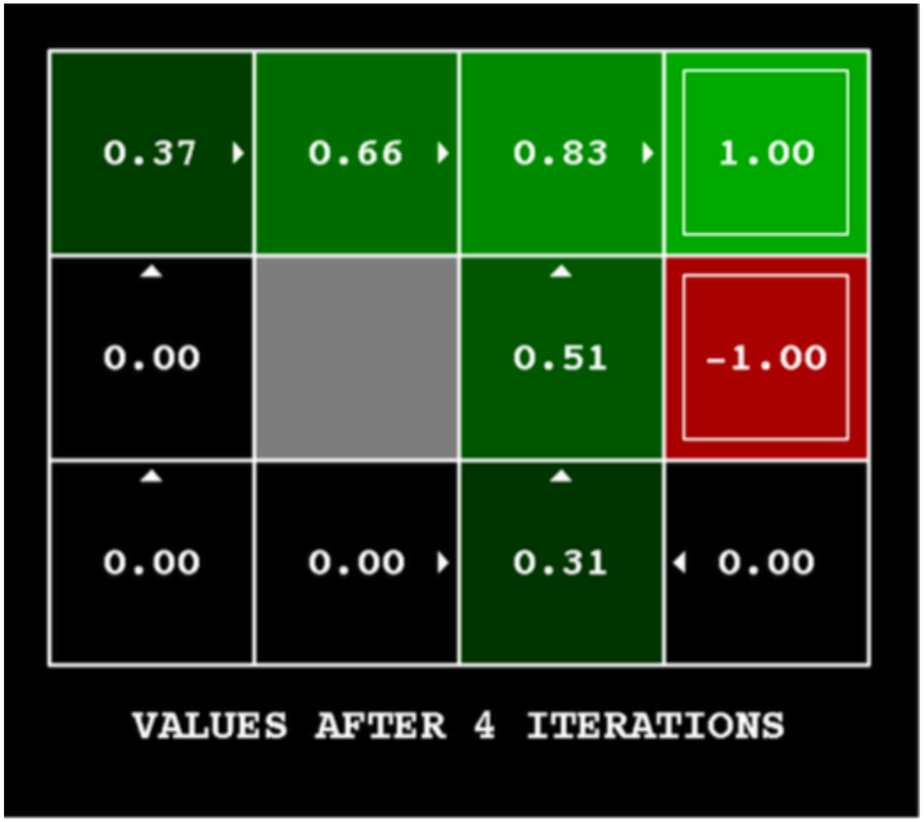
VALUES AFTER 4 ITERATIONS

(D. Klein, P. Abbeel)

Noise 0.2, Discount 0.9

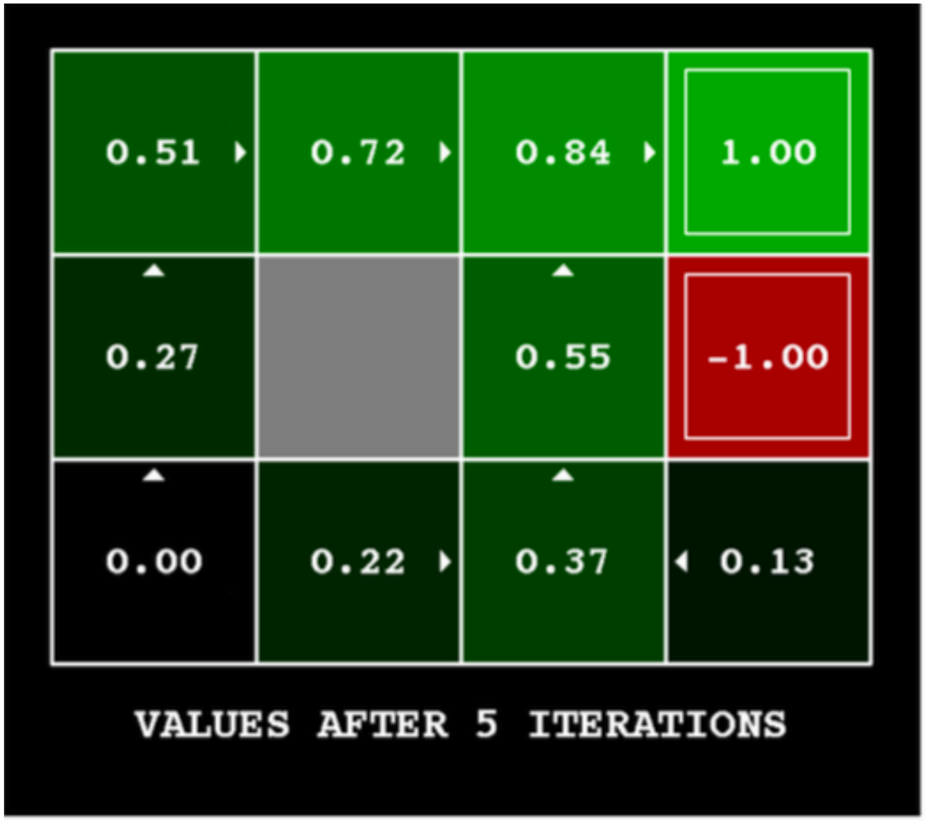
VI, GridWorld

$$V'(s) \leftarrow \max_{a \in A} \left[r(s,a) + \gamma \sum_{s' \in S} p(s' \mid s,a) V(s') \right], \quad \forall s$$



0.37 ▸	0.66 ▸	0.83 ▸	1.00
▴ 0.00		▴ 0.51	-1.00
▴ 0.00	0.00 ▸	▴ 0.31	◂ 0.00

VALUES AFTER 4 ITERATIONS



0.51 ▸	0.72 ▸	0.84 ▸	1.00
▴ 0.27		▴ 0.55	-1.00
▴ 0.00	0.22 ▸	▴ 0.37	◂ 0.13

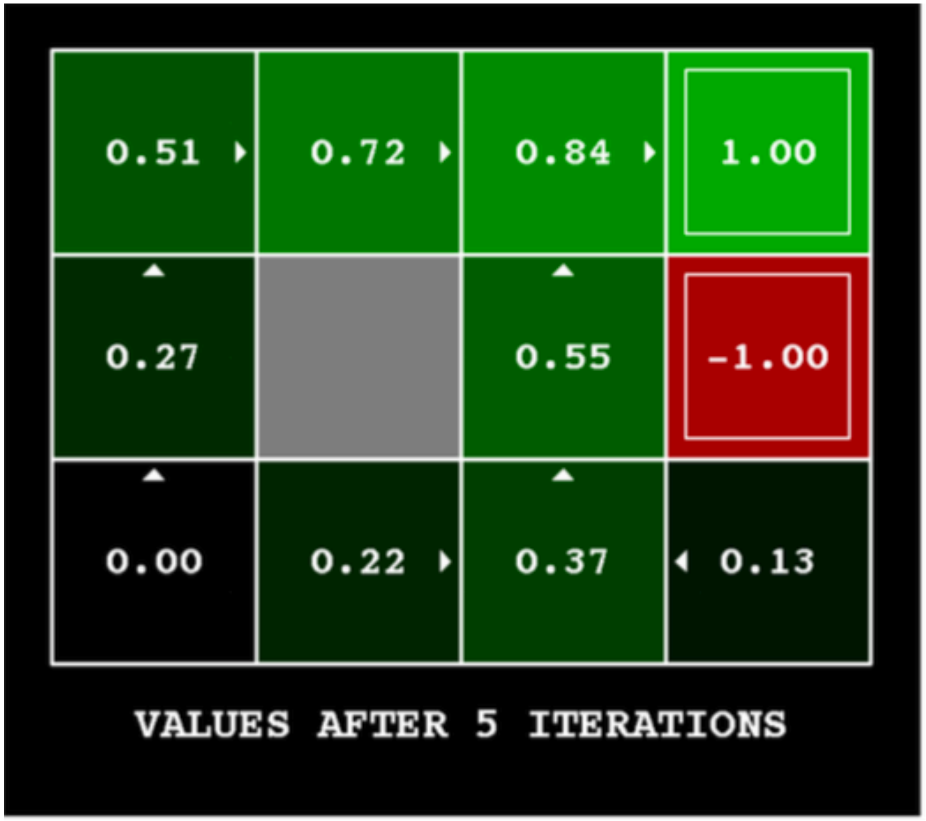
VALUES AFTER 5 ITERATIONS

(D. Klein, P. Abbeel)

Noise 0.2, Discount 0.9

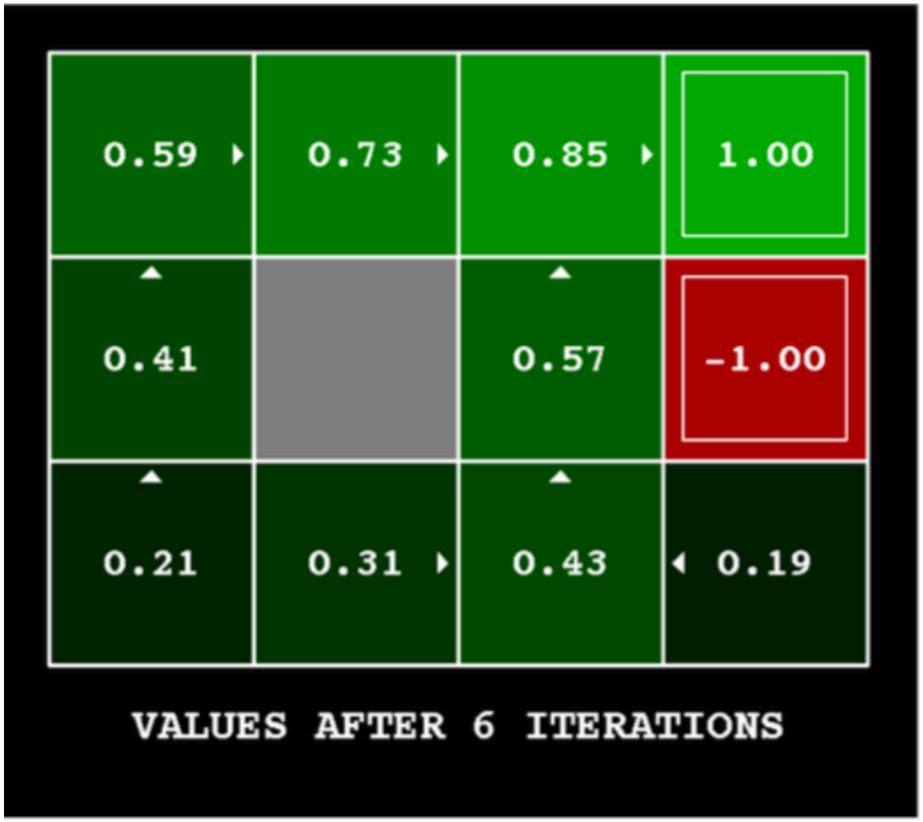
VI, GridWorld

$$V'(s) \leftarrow \max_{a \in A} \left[r(s,a) + \gamma \sum_{s' \in S} p(s' \mid s,a) V(s') \right], \quad \forall s$$



0.51 ▸	0.72 ▸	0.84 ▸	1.00
▴ 0.27		▴ 0.55	-1.00
▴ 0.00	0.22 ▸	▴ 0.37	◂ 0.13

VALUES AFTER 5 ITERATIONS



0.59 ▸	0.73 ▸	0.85 ▸	1.00
▴ 0.41		▴ 0.57	-1.00
▴ 0.21	0.31 ▸	▴ 0.43	◂ 0.19

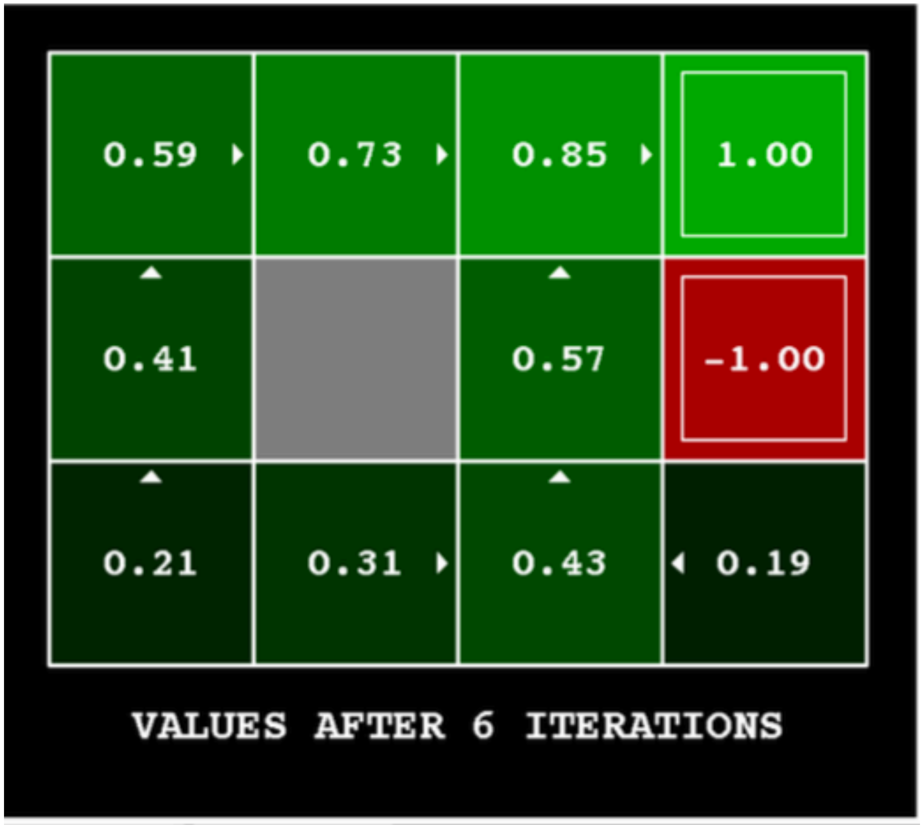
VALUES AFTER 6 ITERATIONS

(D. Klein, P. Abbeel)

Noise 0.2, Discount 0.9

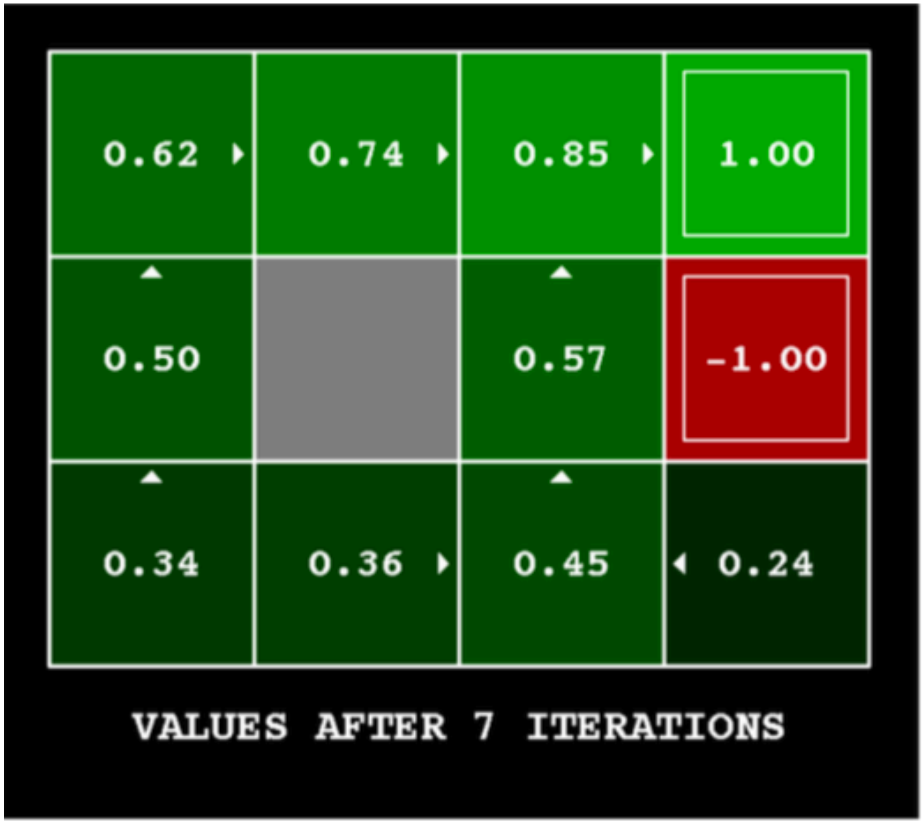
VI, GridWorld

$$V'(s) \leftarrow \max_{a \in A} \left[r(s,a) + \gamma \sum_{s' \in S} p(s' \mid s, a) V(s') \right], \quad \forall s$$



0.59 ▸	0.73 ▸	0.85 ▸	1.00
▴ 0.41		▴ 0.57	-1.00
▴ 0.21	0.31 ▸	▴ 0.43	◂ 0.19

VALUES AFTER 6 ITERATIONS



0.62 ▸	0.74 ▸	0.85 ▸	1.00
▴ 0.50		▴ 0.57	-1.00
▴ 0.34	0.36 ▸	▴ 0.45	◂ 0.24

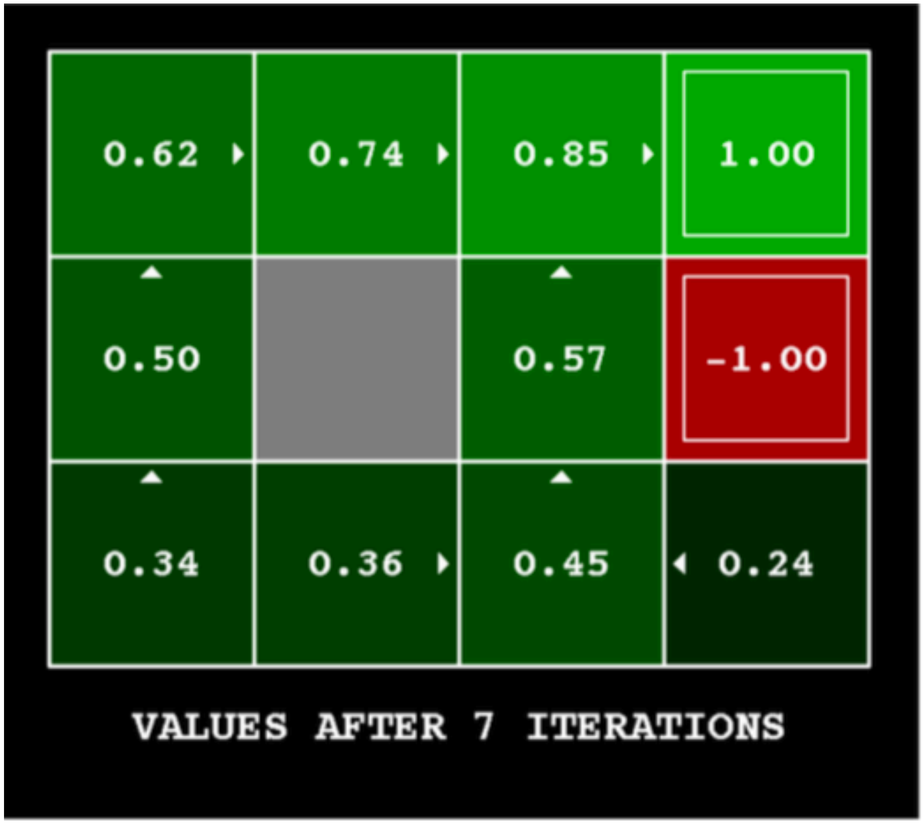
VALUES AFTER 7 ITERATIONS

(D. Klein, P. Abbeel)

Noise 0.2, Discount 0.9

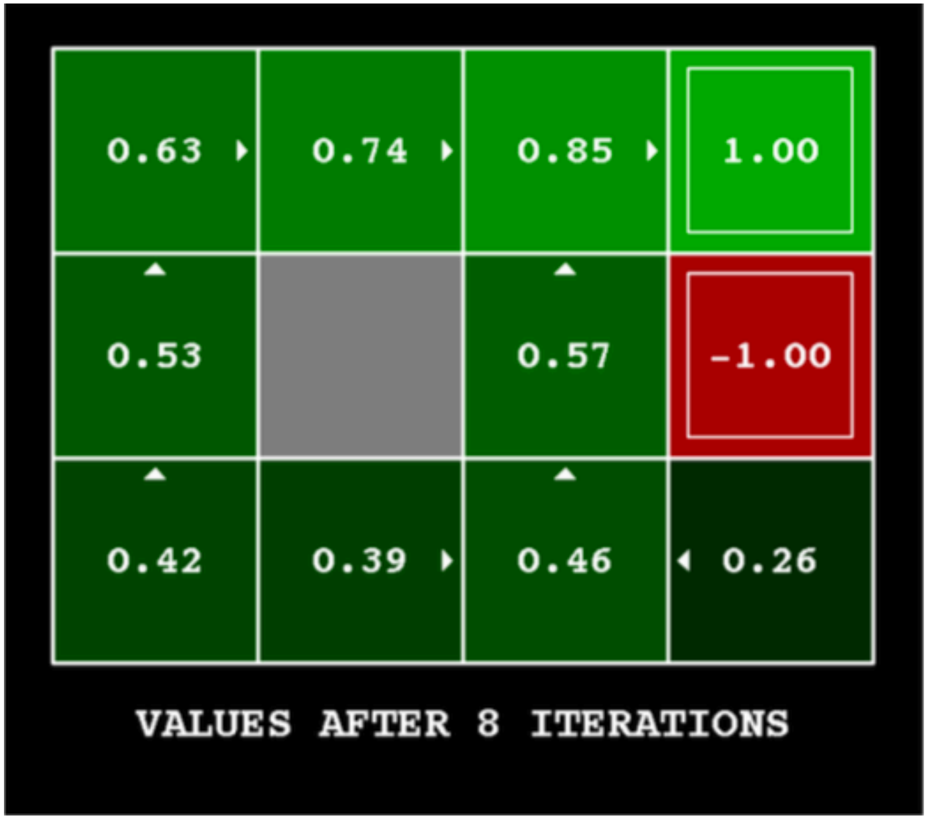
VI, GridWorld

$$V'(s) \leftarrow \max_{a \in A} \left[r(s,a) + \gamma \sum_{s' \in S} p(s' \mid s,a) V(s') \right], \quad \forall s$$



0.62 ▸	0.74 ▸	0.85 ▸	1.00
▴ 0.50		▴ 0.57	-1.00
▴ 0.34	0.36 ▸	▴ 0.45	◂ 0.24

VALUES AFTER 7 ITERATIONS



0.63 ▸	0.74 ▸	0.85 ▸	1.00
▴ 0.53		▴ 0.57	-1.00
▴ 0.42	0.39 ▸	▴ 0.46	◂ 0.26

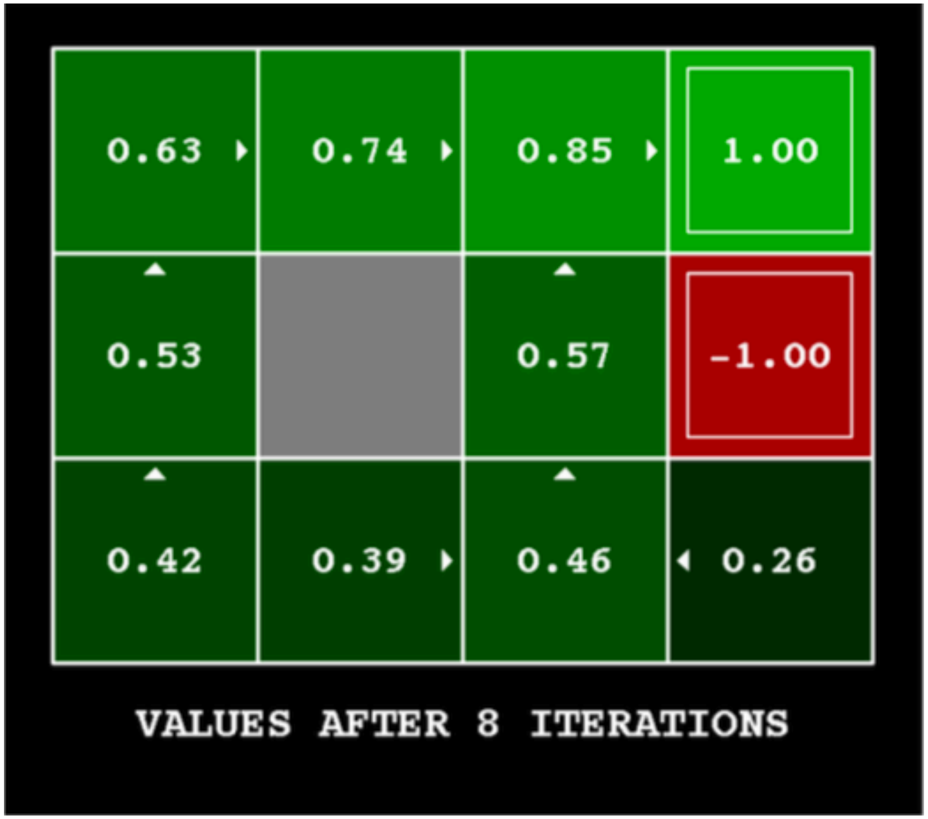
VALUES AFTER 8 ITERATIONS

(D. Klein, P. Abbeel)

Noise 0.2, Discount 0.9

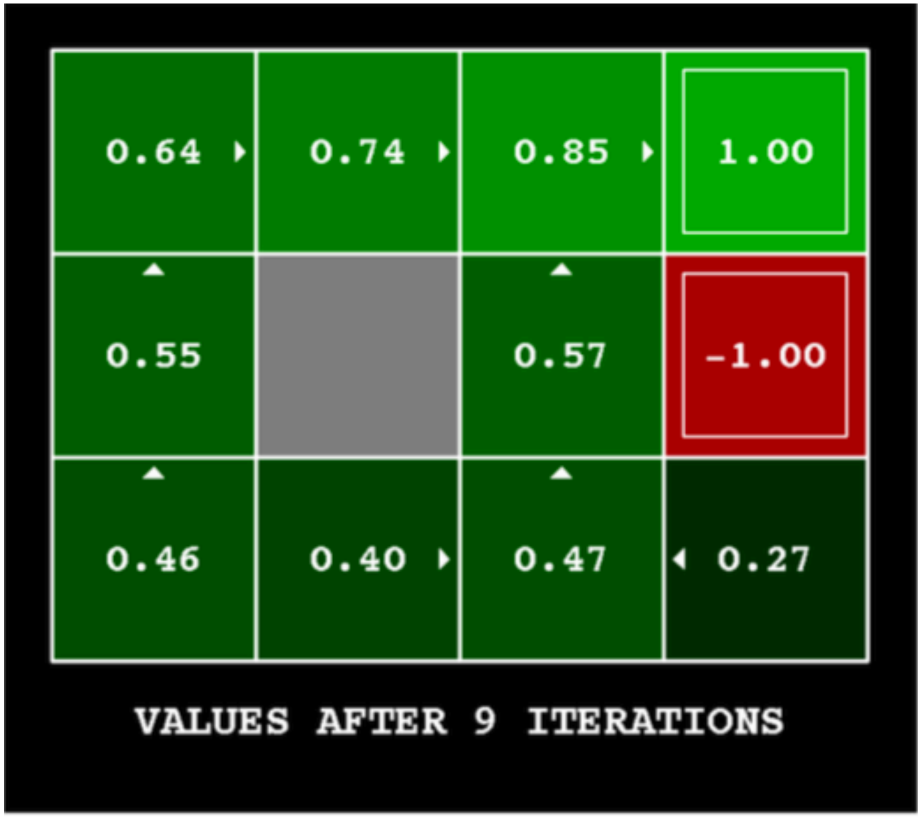
VI, GridWorld

$$V'(s) \leftarrow \max_{a \in A} \left[r(s,a) + \gamma \sum_{s' \in S} p(s' \mid s, a) V(s') \right], \quad \forall s$$



0.63 ▸	0.74 ▸	0.85 ▸	1.00
▴ 0.53		▴ 0.57	-1.00
▴ 0.42	0.39 ▸	▴ 0.46	◂ 0.26

VALUES AFTER 8 ITERATIONS



0.64 ▸	0.74 ▸	0.85 ▸	1.00
▴ 0.55		▴ 0.57	-1.00
▴ 0.46	0.40 ▸	▴ 0.47	◂ 0.27

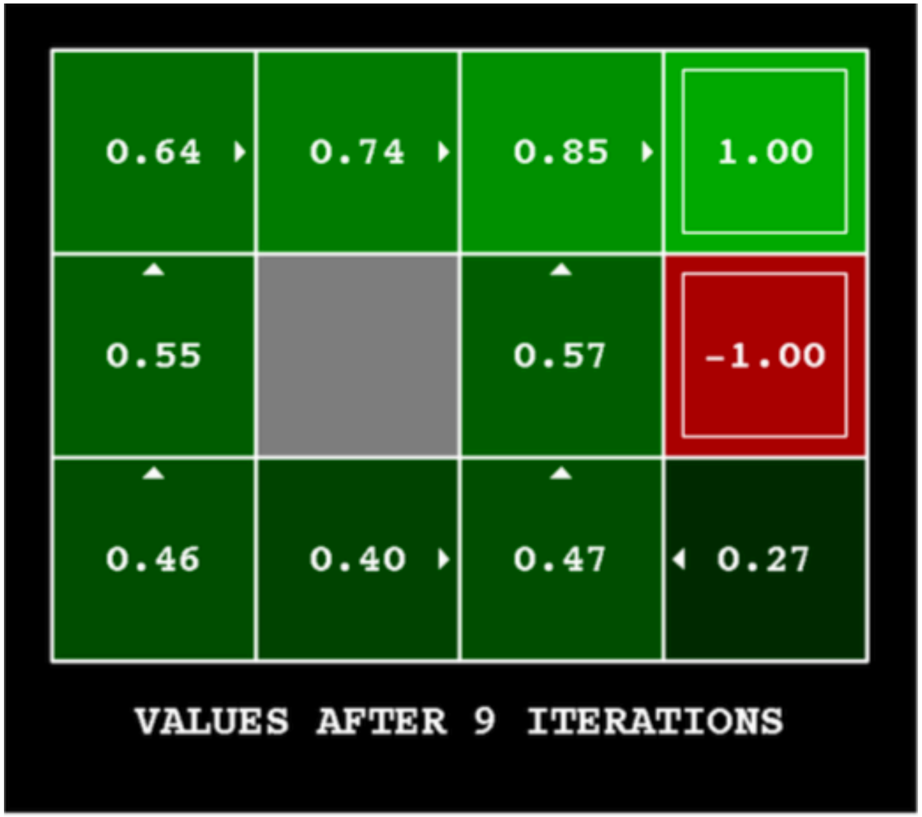
VALUES AFTER 9 ITERATIONS

(D. Klein, P. Abbeel)

Noise 0.2, Discount 0.9

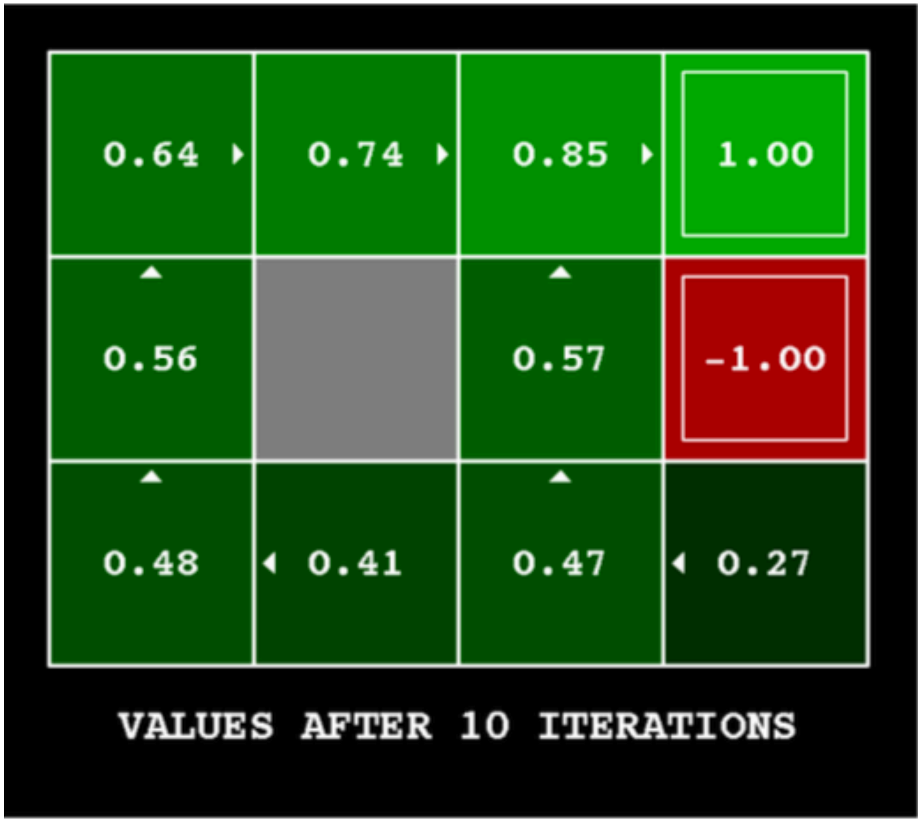
VI, GridWorld

$$V'(s) \leftarrow \max_{a\in A}\left[r(s,a)+\gamma\sum_{s'\in S}p(s'\mid s,a)V(s')\right], \quad \forall s$$



0.64 ▸	0.74 ▸	0.85 ▸	1.00
▴ 0.55		▴ 0.57	-1.00
▴ 0.46	0.40 ▸	▴ 0.47	◂ 0.27

VALUES AFTER 9 ITERATIONS



0.64 ▸	0.74 ▸	0.85 ▸	1.00
▴ 0.56		▴ 0.57	-1.00
▴ 0.48	◂ 0.41	▴ 0.47	◂ 0.27

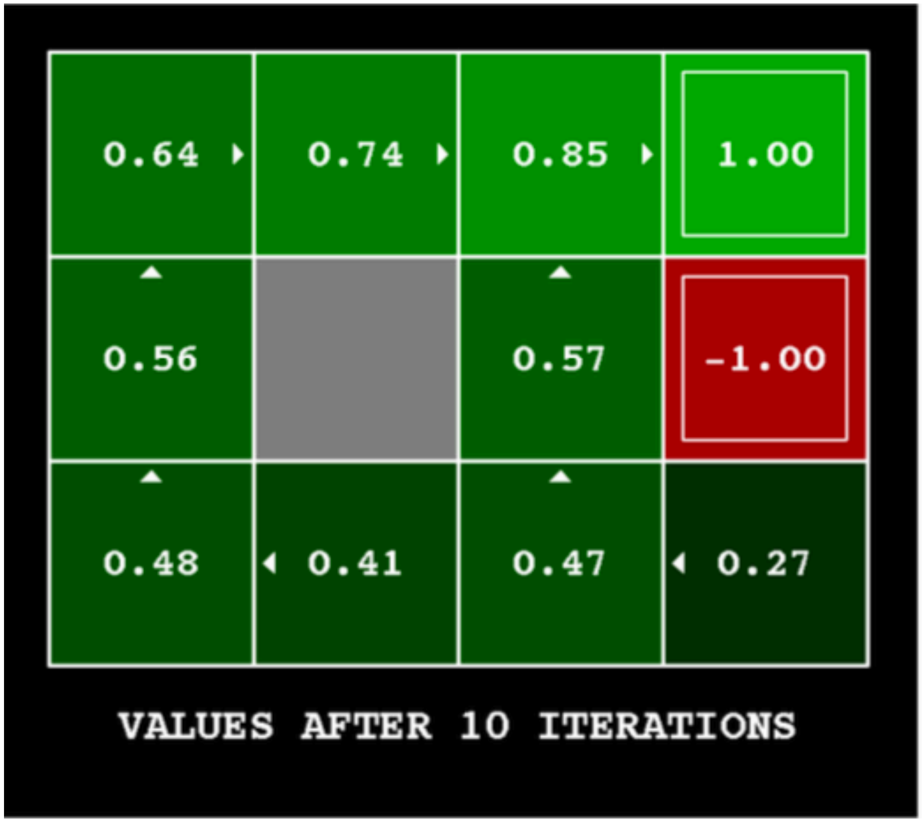
VALUES AFTER 10 ITERATIONS

(D. Klein, P. Abbeel)

Noise 0.2, Discount 0.9

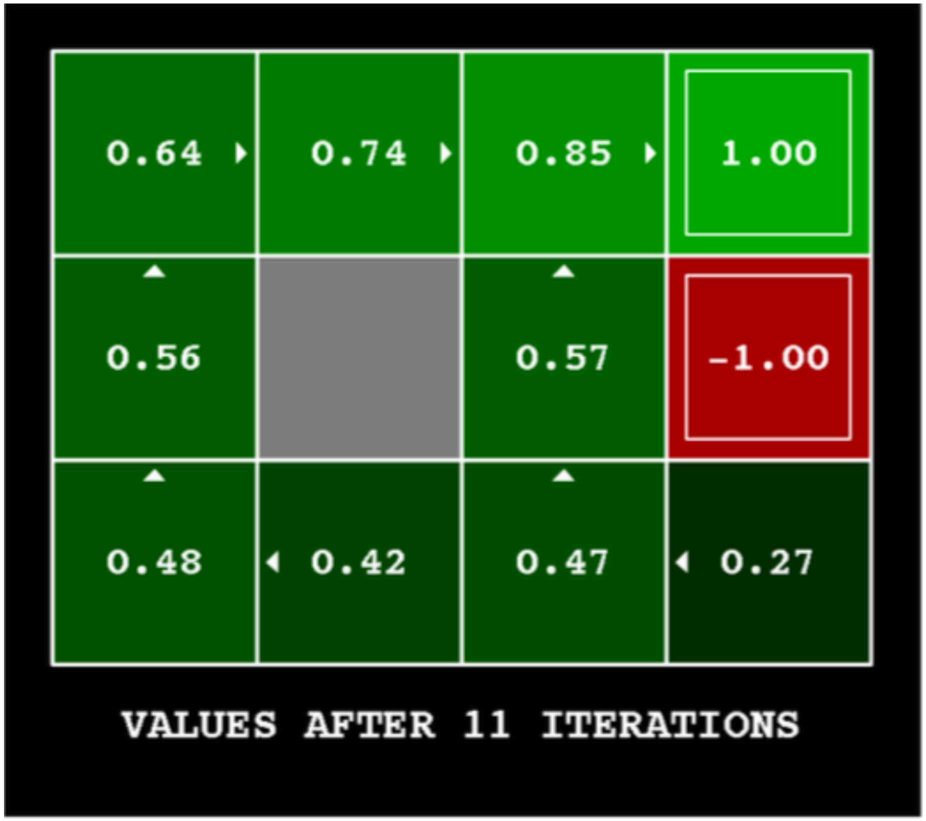
VI, GridWorld

$$V'(s) \leftarrow \max_{a \in A} \left[r(s,a) + \gamma \sum_{s' \in S} p(s' \mid s,a) V(s') \right], \quad \forall s$$



0.64 ▸	0.74 ▸	0.85 ▸	1.00
▴ 0.56		▴ 0.57	-1.00
▴ 0.48	◂ 0.41	▴ 0.47	◂ 0.27

VALUES AFTER 10 ITERATIONS



0.64 ▸	0.74 ▸	0.85 ▸	1.00
▴ 0.56		▴ 0.57	-1.00
▴ 0.48	◂ 0.42	▴ 0.47	◂ 0.27

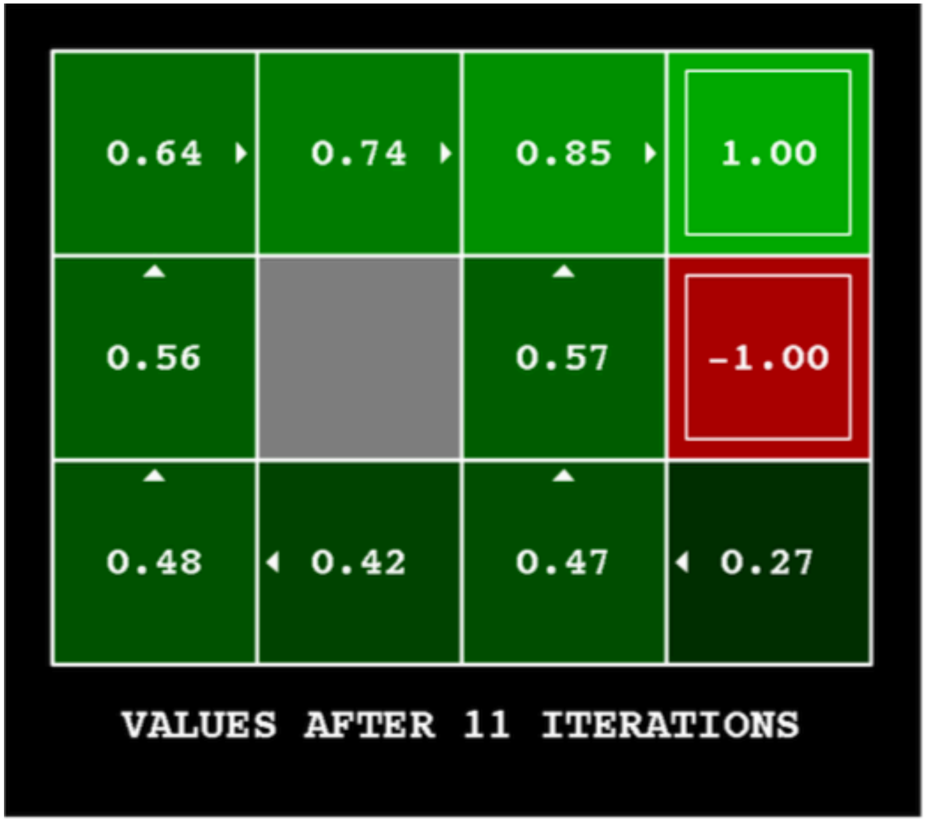
VALUES AFTER 11 ITERATIONS

(D. Klein, P. Abbeel)

Noise 0.2, Discount 0.9

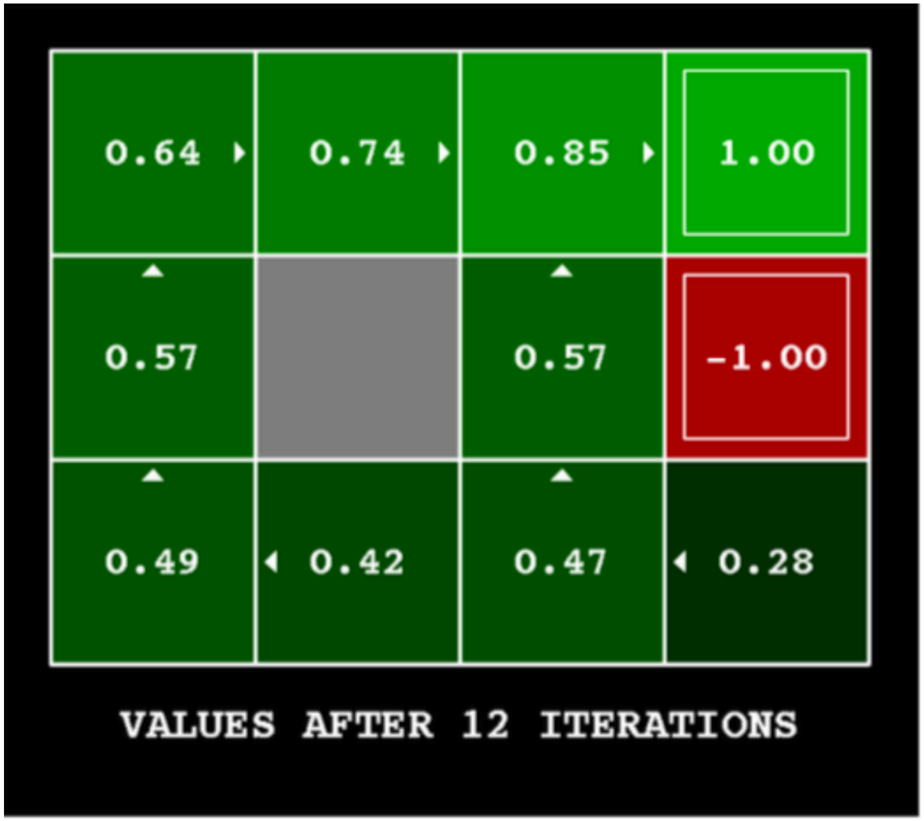
VI, GridWorld

$$V'(s) \leftarrow \max_{a \in A} \left[r(s,a) + \gamma \sum_{s' \in S} p(s' \mid s, a) V(s') \right], \quad \forall s$$



0.64 ▸	0.74 ▸	0.85 ▸	1.00
▴ 0.56		▴ 0.57	-1.00
▴ 0.48	◂ 0.42	▴ 0.47	◂ 0.27

VALUES AFTER 11 ITERATIONS



0.64 ▸	0.74 ▸	0.85 ▸	1.00
▴ 0.57		▴ 0.57	-1.00
▴ 0.49	◂ 0.42	▴ 0.47	◂ 0.28

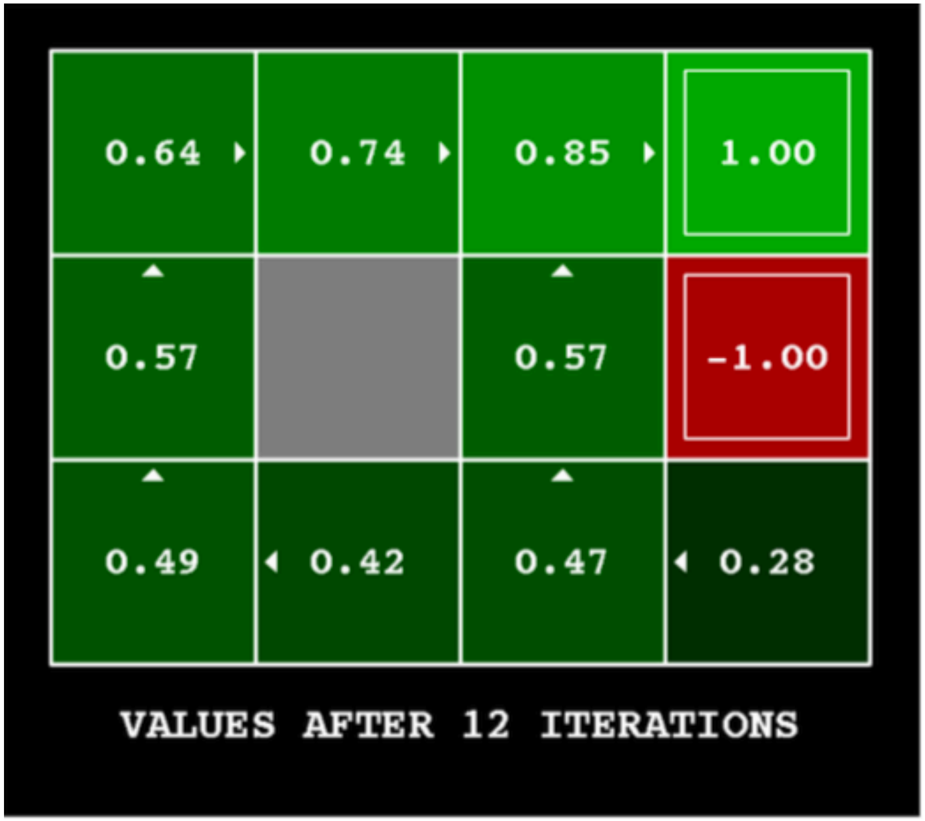
VALUES AFTER 12 ITERATIONS

(D. Klein, P. Abbeel)

Noise 0.2, Discount 0.9

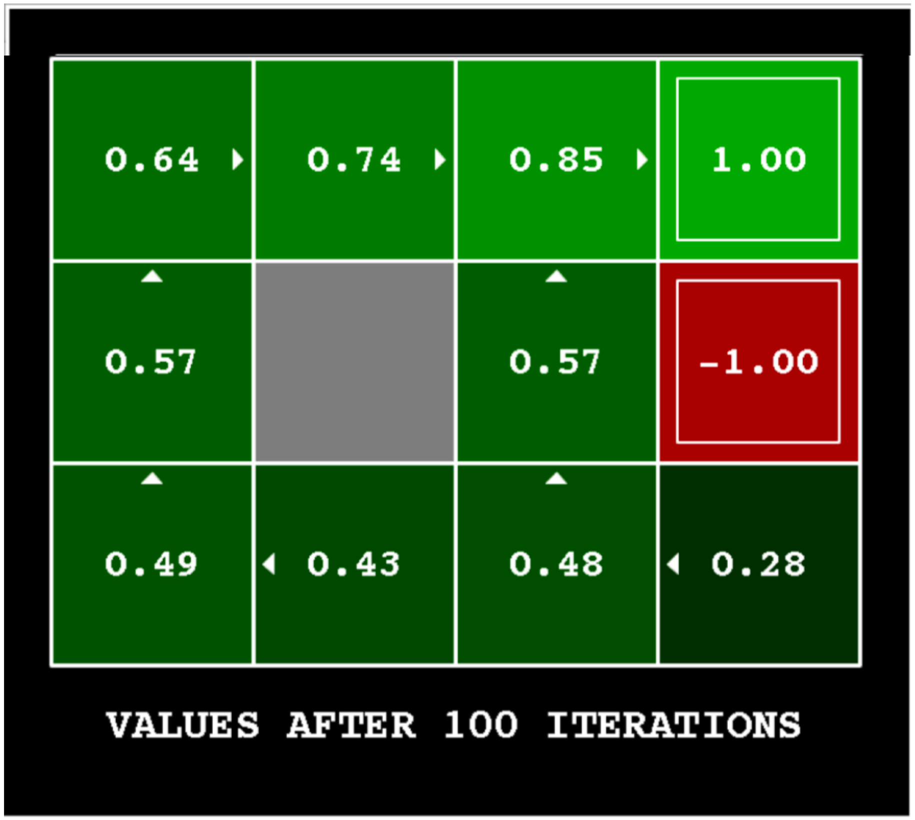
VI, GridWorld

$$V'(s) \leftarrow \max_{a \in A} \left[r(s,a) + \gamma \sum_{s' \in S} p(s' \mid s,a) V(s') \right], \quad \forall s$$



0.64 ▸	0.74 ▸	0.85 ▸	1.00
▴ 0.57		▴ 0.57	-1.00
▴ 0.49	◂ 0.42	▴ 0.47	◂ 0.28

VALUES AFTER 12 ITERATIONS



0.64 ▸	0.74 ▸	0.85 ▸	1.00
▴ 0.57		▴ 0.57	-1.00
▴ 0.49	◂ 0.43	▴ 0.48	◂ 0.28

VALUES AFTER 100 ITERATIONS

(D. Klein, P. Abbeel)

Noise 0.2, Discount 0.9

Policy Iteration

Don’t stop in goal states in this Grid World

0	0	0	1
0	■	0	-100
0	0	0	0

Rewards

(Z. Kolter)

Initialize “up” everywhere

0.418	0.884	2.331	6.367
0.367	■	-8.610	-105.7
-0.168	-4.641	-14.27	-85.05

After 1 improvement step

pi0-> V0 -> pi1 -> V1

Policy Iteration

Don’t stop in goal states in this Grid World

0	0	0	1
0	■	0	-100
0	0	0	0

Rewards

(Z. Kolter)

5.414	6.248	7.116	8.634
4.753	■	2.881	-102.7
2.251	1.977	1.849	-8.701

After 2 improvement steps

pi0-> V0 -> pi1 -> V1 -> pi2 -> V2

Policy Iteration

Don’t stop in goal states in this Grid World

0	0	0	1
0	■	0	-100
0	0	0	0

Rewards

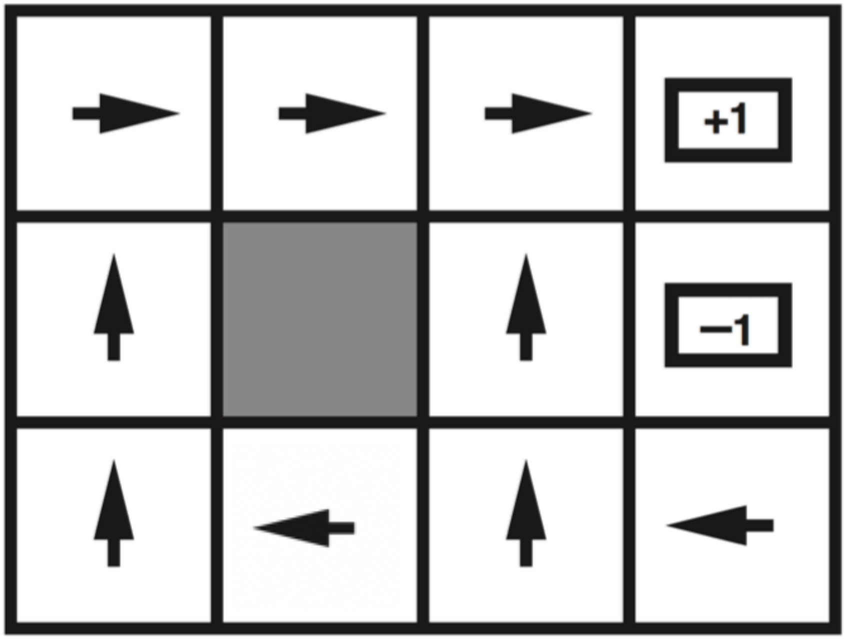
(Z. Kolter)

5.470	6.313	7.190	8.669
4.803	■	3.347	-96.67
4.161	3.654	3.222	1.526

After 3 improvement steps (converged!)

pi0-> V0 -> pi1 -> V1 -> pi2 -> V2 -> pi3 -> V3

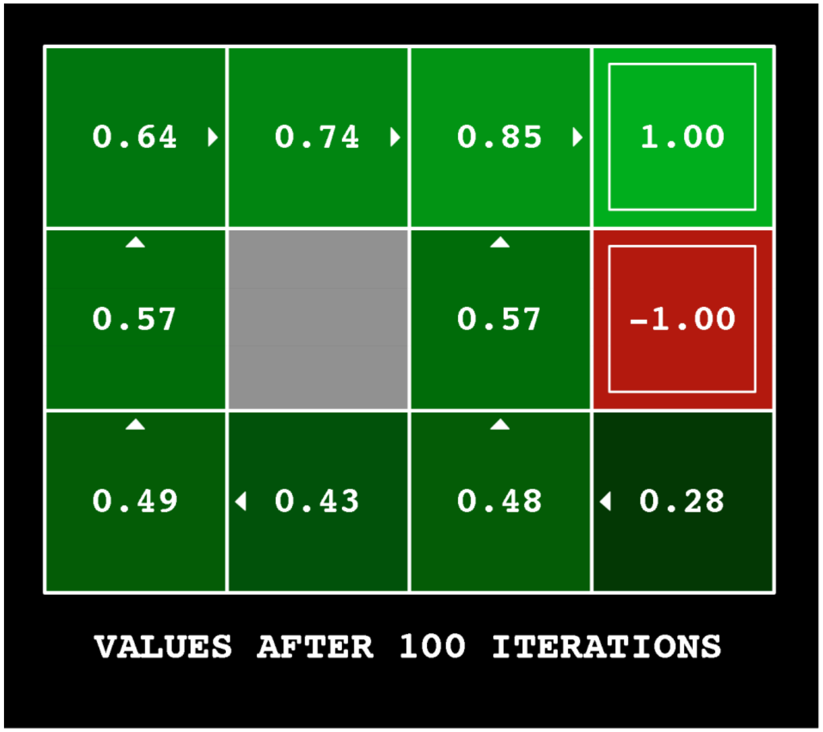
GridWorld



(D. Klein, P. Abbeel)

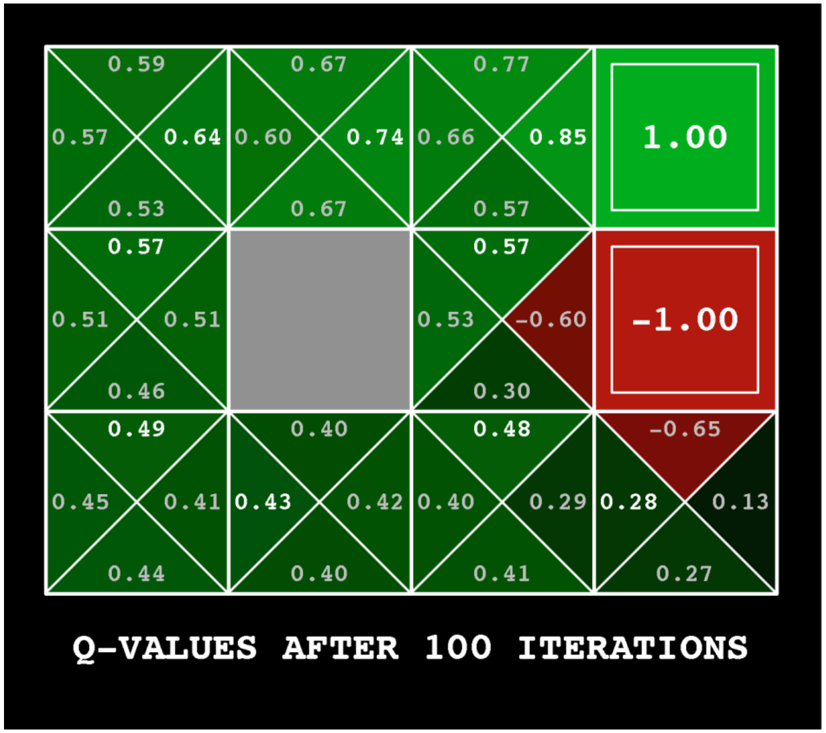
- r(s,a) = 0, except states (1,4), (2,4). In these states get +1 or -1 when take ANY action. Then no more actions
- Bounce off obstacles. Actuator has 20% noise; e.g., w/ prob 0.1 goes L, prob 0.1 goes R when moving U
- Discounting 0.9 (r + 0.9 r + 0.9^2 r + …)

Can also look at Q-Values



0.64 ▸	0.74 ▸	0.85 ▸	1.00
▴ 0.57		▴ 0.57	-1.00
▴ 0.49	◂ 0.43	▴ 0.48	◂ 0.28

VALUES AFTER 100 ITERATIONS



0.59 0.57 0.64 0.53	0.67 0.60 0.74 0.67	0.77 0.66 0.85 0.57	1.00
0.57 0.51 0.51 0.46		0.57 0.53 -0.60 0.30	-1.00
0.49 0.45 0.41 0.44	0.40 0.43 0.42 0.40	0.48 0.40 0.29 0.41	-0.65 0.28 0.13 0.27

Q-VALUES AFTER 100 ITERATIONS

(D. Klein, P. Abbeel)

Noise 0.2, Discount 0.9