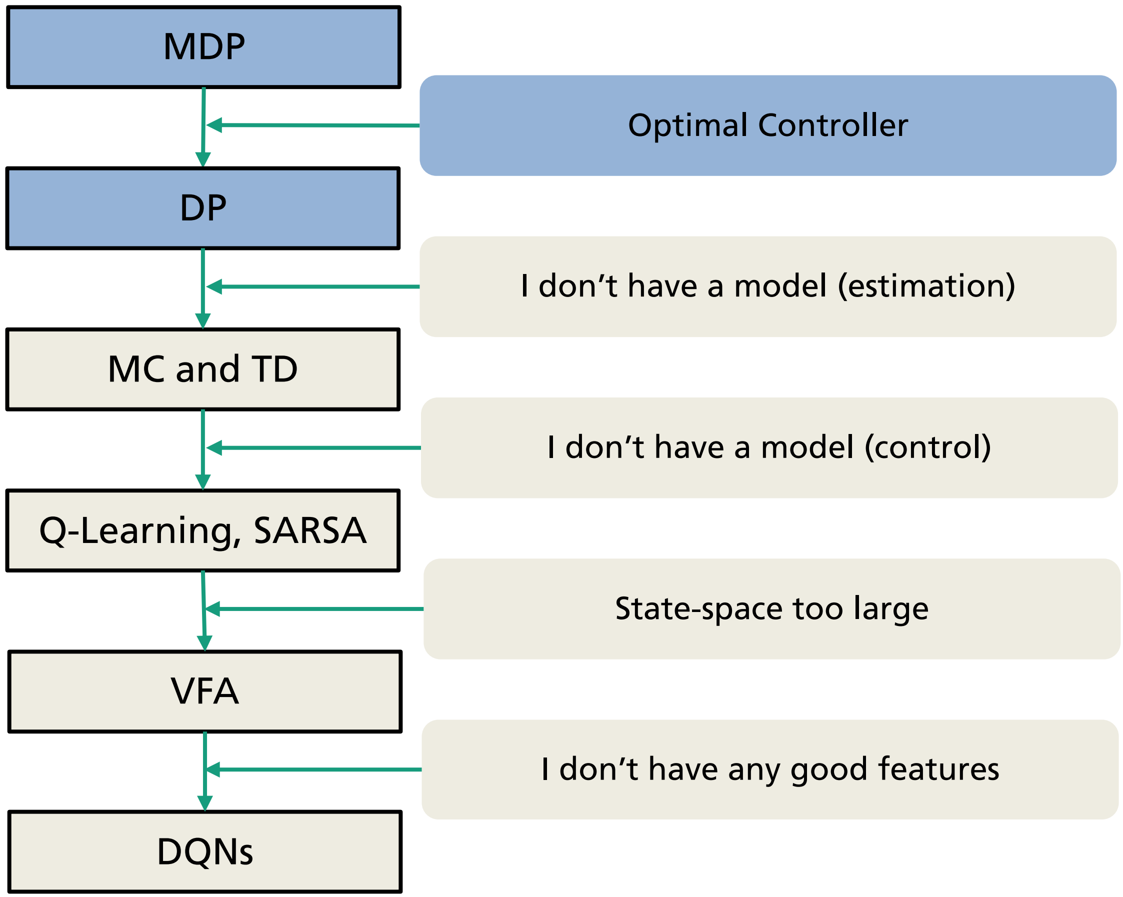
Dynamic Programming: Value Iteration

Christopher Mutschler

Overview



MDP
Optimal Controller
DP
I don’t have a model (estimation)
MC and TD
I don’t have a model (control)
Q-Learning, SARSA
State-space too large
VFA
I don’t have any good features
DQNs

Dynamic Programming: Value Iteration

- How do we find optimal controllers for given (known) MDPs?
- Optimal Solver #1: Value Iteration (convergence guaranteed)

Value Iteration, for estimating $\pi \approx \pi_*$

Algorithm parameter: a small threshold $\theta > 0$ determining accuracy of estimation
Initialize $V(s)$, for all $s \in \mathcal{S}^+$, arbitrarily except that $V(terminal) = 0$

Loop:
| $\Delta \leftarrow 0$
| Loop for each $s \in \mathcal{S}$:
| $v \leftarrow V(s)$
| $V(s) \leftarrow \max_a \sum_{s',r} p(s', r \mid s, a)\big[r + \gamma V(s')\big]$
| $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
until $\Delta < \theta$

Output a deterministic policy, $\pi \approx \pi_*$, such that
  $\pi(s) = \operatorname{argmax}_a \sum_{s',r} p(s', r \mid s, a)\big[r + \gamma V(s')\big]$

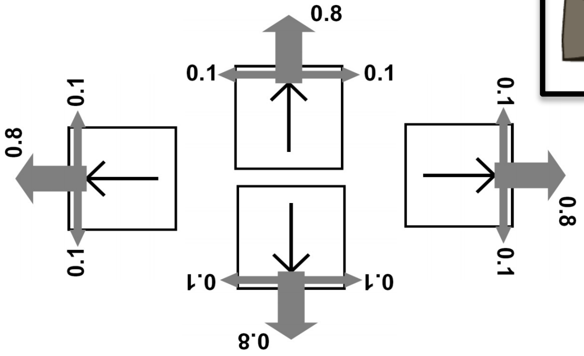
Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction. MIT press.

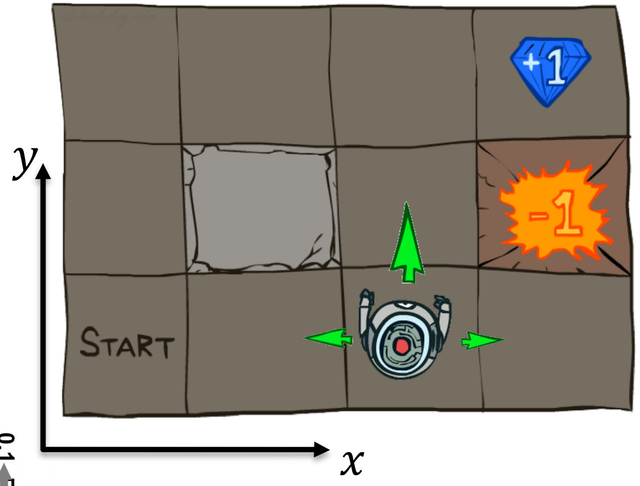
Dynamic Programming: Value Iteration

- How do we find optimal controllers for given (known) MDPs?
- Optimal Solver #1: Value Iteration (convergence guaranteed)

$$V_i(S^{x,y}) = \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}, r} \mathcal{P}(s', r \mid s, a)\,[r + \gamma V_{i-1}(s')]$$

- Value Iteration Example:
  - Noise = 0.2 (it is windy)
  - γ = 0.9
  - Living reward = 0.0
    (Transitioning from state to state)





http://ai.berkeley.edu/reinforcement.html

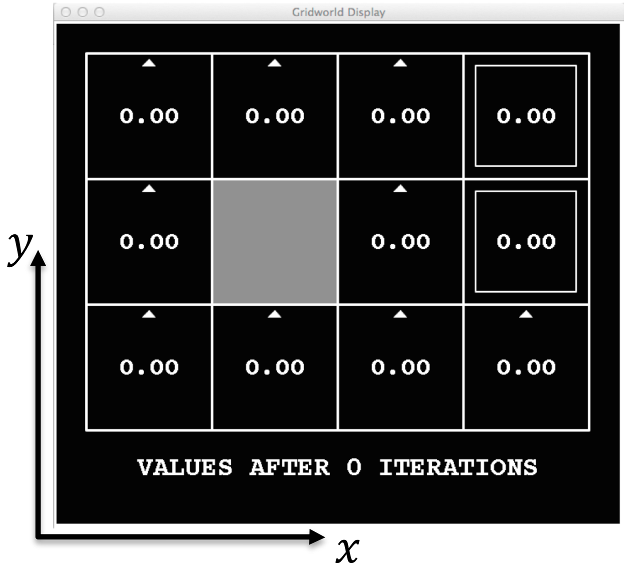
Dynamic Programming: Value Iteration

- How do we find optimal controllers for given (known) MDPs?
- Optimal Solver #1: Value Iteration (convergence guaranteed)

$$V_i(S^{x,y}) = \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}, r} \mathcal{P}(s', r \mid s, a)\,[r + \gamma V_{i-1}(s')]$$

- Value Iteration Example: noise $= 0.2$, $\gamma = 0.9$, $r = 0.0$

$V_1(S^{3,2}) = 1.00$ (terminal state with reward 1.0)
$V_1(S^{2,2}) = 0.00$
$V_1(S^{1,2}) = 0.00$
…
$V_1(S^{3,1}) = -1.00$ (terminal state with reward -1.0)



http://ai.berkeley.edu/reinforcement.html

Dynamic Programming: Value Iteration

- How do we find optimal controllers for given (known) MDPs?
- Optimal Solver #1: Value Iteration (convergence guaranteed)

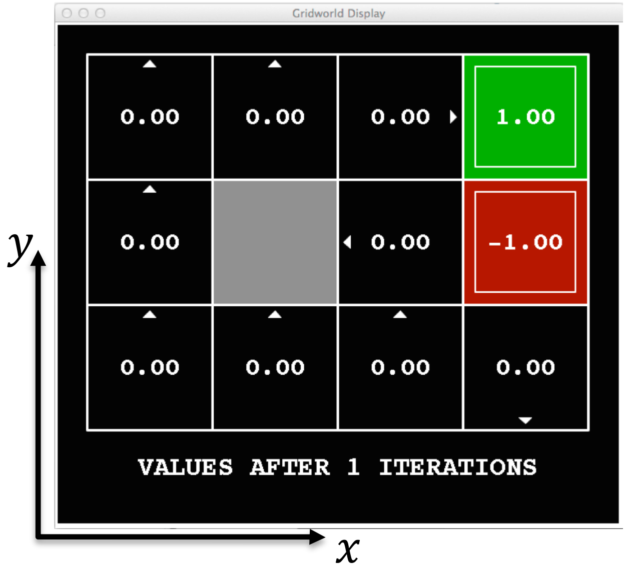
$$V_i(S^{x,y}) = \max_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S},r} \mathcal{P}(s',r|\,s,a)\,[r+\gamma V_{i-1}(s')]$$

- Value Iteration Example: noise $= 0.2$, $\gamma = 0.9$, $r = 0.0$

$$V_2(S^{2,2}) = \max_{a\in\mathcal{A}} \begin{Bmatrix} a=R: 0.8*[0.0+0.9*1.0]+0.1*[0.0+0.9*0.0]+0.1*[0.0+0.9*0.0] \\ a=L: 0.8*[0.0+0.9*0.0]+0.1*[0.0+0.9*0.0]+0.1*[0.0+0.9*0.0] \\ a=U: 0.8*[0.0+0.9*0.0]+0.1*[0.0+0.9*0.0]+0.1*[0.0+0.9*1.0] \\ a=D: 0.8*[0.0+0.9*0.0]+0.1*[0.0+0.9*1.0]+0.1*[0.0+0.9*0.0] \end{Bmatrix}$$

$$= \max_{a\in\mathcal{A}} \begin{Bmatrix} 0.8*0.9*1.0+0.1=0.72 \\ 0 \\ 0.1*0.9*1.0=0.09 \\ 0.1*0.9*1.0=0.09 \end{Bmatrix} = 0.72$$

$$V_2(S^{2,1}) = \cdots$$



http://ai.berkeley.edu/reinforcement.html

Dynamic Programming: Value Iteration

- How do we find optimal controllers for given (known) MDPs?
- Optimal Solver #1: Value Iteration (convergence guaranteed)

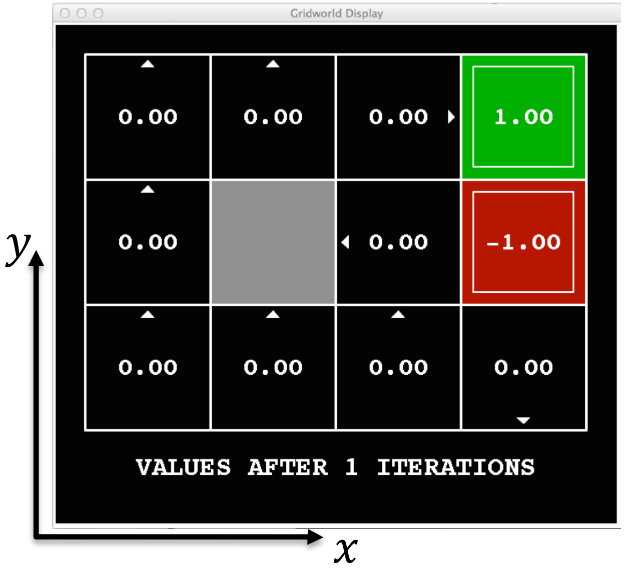
$$V_i(S^{x,y}) = \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}, r} \mathcal{P}(s', r | \, s, a) \, [r + \gamma V_{i-1}(s')]$$

- Value Iteration Example: noise $= 0.2$, $\gamma = 0.9$, $r = 0.0$

$$V_2(S^{2,2}) = 0.72$$

$$V_2(S^{2,1}) = \max_{a \in \mathcal{A}} \begin{Bmatrix} a = R: 0.8 * [0.0 + 0.9 * -1.0] + 0.1 \; * [0.0 + 0.9 \; * 0.0] + 0.1 \; * [0.0 + 0.9 \; * 0.0] \\ a = L: 0.8 * [0.0 + 0.9 * 0.0] + 0.1 \; * [0.0 + 0.9 \; * 0.0] + 0.1 \; * [0.0 + 0.9 \; * 0.0] \\ a = U: 0.8 * [0.0 + 0.9 * 0.0] + 0.1 \; * [0.0 + 0.9 \; * 0.0] + 0.1 \; * [0.0 + 0.9 \; * -1.0] \\ a = D: 0.8 * [0.0 + 0.9 * 0.0] + 0.1 \; * [0.0 + 0.9 \; * -1.0] + 0.1 \; * [0.0 + 0.9 \; * 0.0] \end{Bmatrix}$$

$$= \max_{a \in \mathcal{A}} \begin{Bmatrix} 0.8 * 0.9 * -1.0 + 0.1 = -0.72 \\ 0 \\ 0.1 * 0.9 * -1.0 = -0.09 \\ 0.1 * 0.9 * -1.0 = -0.09 \end{Bmatrix} = 0.00$$



http://ai.berkeley.edu/reinforcement.html

Dynamic Programming: Value Iteration

- How do we find optimal controllers for given (known) MDPs?
- Optimal Solver #1: Value Iteration (convergence guaranteed)

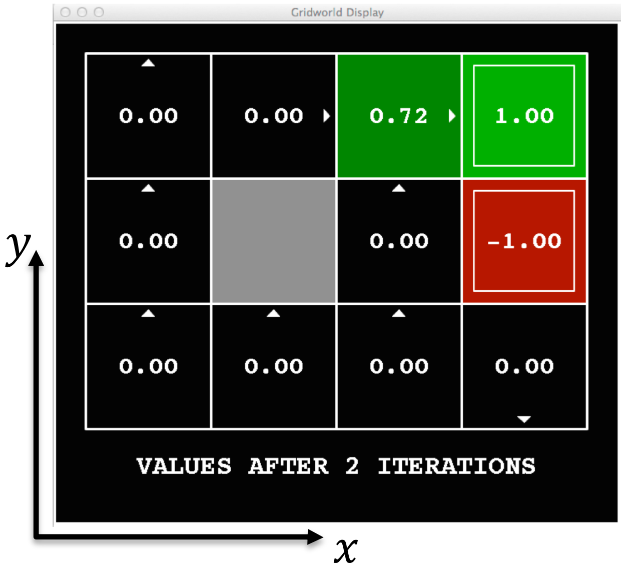
$$V_i(S^{x,y}) = \max_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S},r} \mathcal{P}(s',r|\,s,a)\,[r+\gamma V_{i-1}(s')]$$

- Value Iteration Example: noise $= 0.2$, $\gamma = 0.9$, $r = 0.0$

$$V_3(S^{2,2}) = \max_{a\in\mathcal{A}} \begin{cases} a = R: 0.8 * [0.0 + 0.9 * 1.0] + 0.1 * [0.0 + 0.9 * 0.72] + 0.1 * [0.0 + 0.9 * 0.0] \\ a = L: 0.8 * [0.0 + 0.9 * 0.0] + 0.1 * [0.0 + 0.9 * 0.0] + 0.1 * [0.0 + 0.9 * 0.72] \\ a = U: 0.8 * [0.0 + 0.9 * 0.72] + 0.1 * [0.0 + 0.9 * 0.0] + 0.1 * [0.0 + 0.9 * 1.0] \\ a = D: 0.8 * [0.0 + 0.9 * 0.0] + 0.1 * [0.0 + 0.9 * 1.0] + 0.1 * [0.0 + 0.9 * 0.0] \end{cases}$$

$$= \max_{a\in\mathcal{A}} \left\{ \begin{matrix} 0.8 * 0.9 * 1.0 + 0.1 * 0.9 * 0.72 = 0.72 + 0.0648 \approx 0.78 \\ 0.1 * 0.9 * 0.72 = 0.0648 \approx 0.06 \\ 0.8 * 0.9 * 0.72 + 0.1 * 0.9 * 1.0 = 0.5184 + 0.09 \approx 0.61 \\ 0.1 * 0.9 * 1.0 = 0.09 \end{matrix} \right\} = 0.78$$

$$V_3(S^{2,1}) = \cdots$$



http://ai.berkeley.edu/reinforcement.html

Dynamic Programming: Value Iteration

- How do we find optimal controllers for given (known) MDPs?
- Optimal Solver #1: Value Iteration (convergence guaranteed)

$$V_i(S^{x,y}) = \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}, r} \mathcal{P}(s', r \mid s, a)\, [r + \gamma V_{i-1}(s')]$$

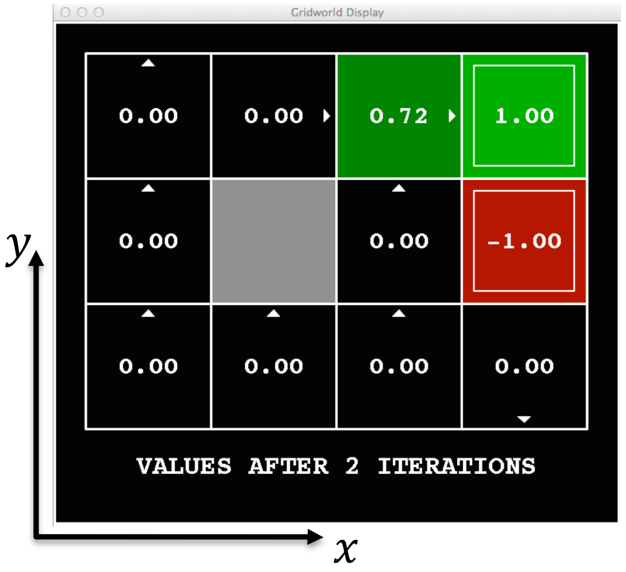
- Value Iteration Example: noise $= 0.2$, $\gamma = 0.9$, $r = 0.0$

$$V_3(S^{2,2}) = 0.78$$

$$V_3(S^{2,1})$$

$$= \max_{a \in \mathcal{A}} \begin{cases} a = R: 0.8 * [0.0 + 0.9 * -1.0] + 0.1 * [0.0 + 0.9 * 0.72] + 0.1 * [0.0 + 0.9 * 0.0] \\ a = L: 0.8 * [0.0 + 0.9 * 0.0] + 0.1 * [0.0 + 0.9 * 0.0] + 0.1 * [0.0 + 0.9 * 0.72] \\ a = U: 0.8 * [0.0 + 0.9 * 0.72] + 0.1 * [0.0 + 0.9 * 0.0] + 0.1 * [0.0 + 0.9 * -1.0] \\ a = D: 0.8 * [0.0 + 0.9 * 0.0] + 0.1 * [0.0 + 0.9 * -1.0] + 0.1 * [0.0 + 0.9 * 0.0] \end{cases}$$

$$= \max_{a \in \mathcal{A}} \left\{ \begin{matrix} 0.8 * 0.9 * -1.0 + 0.1 * 0.9 * 0.72 = -0.72 + 0.0648 \approx -0.65 \\ 0.1 * 0.9 * 0.72 = 0.0648 \approx 0.06 \\ 0.8 * 0.9 * 0.72 - 0.1 * 0.9 * 1.0 = 0.5184 - 0.09 \approx 0.43 \\ 0.1 * 0.9 * -1.0 = -0.09 \end{matrix} \right\} = 0.43$$



http://ai.berkeley.edu/reinforcement.html

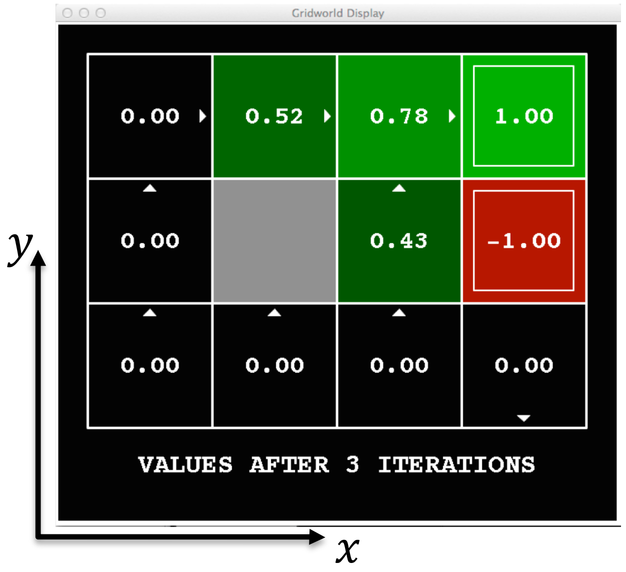
Dynamic Programming: Value Iteration

- How do we find optimal controllers for given (known) MDPs?
- Optimal Solver #1: Value Iteration (convergence guaranteed)

$$V_i(S^{x,y}) = \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}, r} \mathcal{P}(s', r | s, a)\,[r + \gamma V_{i-1}(s')]$$

- Value Iteration Example: noise $= 0.2$, $\gamma = 0.9$, $r = 0.0$

$V_4(S^{2,2}) = \cdots$
$V_4(S^{2,1}) = \cdots$



http://ai.berkeley.edu/reinforcement.html

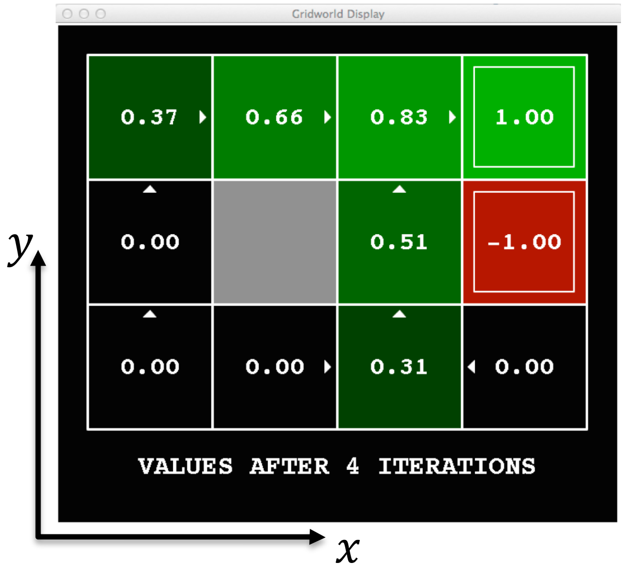
Dynamic Programming: Value Iteration

- How do we find optimal controllers for given (known) MDPs?
- Optimal Solver #1: Value Iteration (convergence guaranteed)

$$V_i(S^{x,y}) = \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}, r} \mathcal{P}(s', r \mid s, a)\,[r + \gamma V_{i-1}(s')]$$

- Value Iteration Example: noise $= 0.2$, $\gamma = 0.9$, $r = 0.0$

$V_5(S^{2,2}) = \cdots$
$V_5(S^{2,1}) = \cdots$



http://ai.berkeley.edu/reinforcement.html

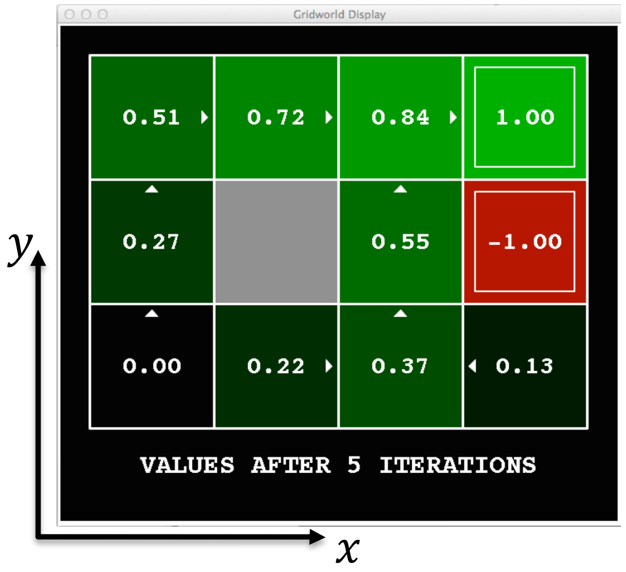
Dynamic Programming: Value Iteration

- How do we find optimal controllers for given (known) MDPs?
- Optimal Solver #1: Value Iteration (convergence guaranteed)

$$V_i(S^{x,y}) = \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}, r} \mathcal{P}(s', r \mid s, a)\,[r + \gamma V_{i-1}(s')]$$

- Value Iteration Example: noise $= 0.2$, $\gamma = 0.9$, $r = 0.0$

$V_6(S^{2,2}) = \cdots$
$V_6(S^{2,1}) = \cdots$



http://ai.berkeley.edu/reinforcement.html

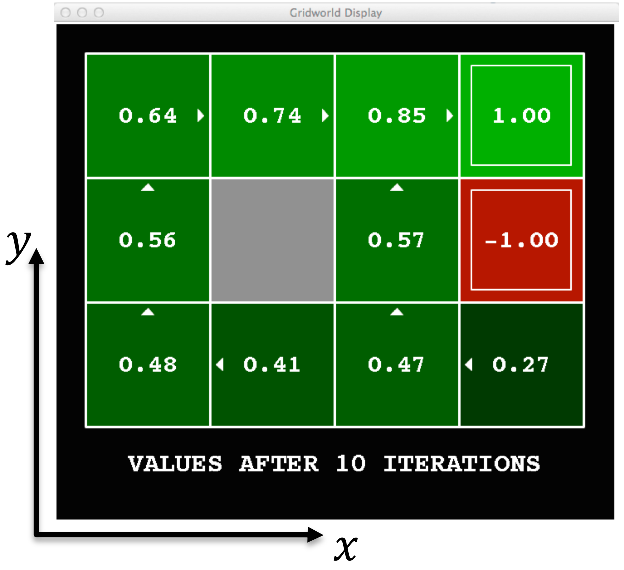
Dynamic Programming: Value Iteration

- How do we find optimal controllers for given (known) MDPs?
- Optimal Solver #1: Value Iteration (convergence guaranteed)

$$V_i(S^{x,y}) = \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S},\, r} \mathcal{P}(s', r \mid s, a)\,[r + \gamma V_{i-1}(s')]$$

- Value Iteration Example: noise $= 0.2$, $\gamma = 0.9$, $r = 0.0$

$V_{11}(S^{2,2}) = \cdots$
$V_{11}(S^{2,1}) = \cdots$



http://ai.berkeley.edu/reinforcement.html

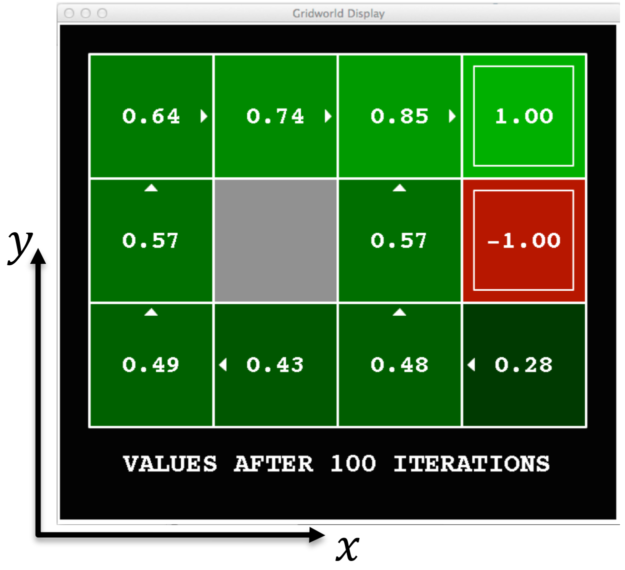
Dynamic Programming: Value Iteration

- How do we find optimal controllers for given (known) MDPs?
- Optimal Solver #1: Value Iteration (convergence guaranteed)

$$V_i(S^{x,y}) = \max_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S},r} \mathcal{P}(s',r|\,s,a)\,[r+\gamma V_{i-1}(s')]$$

- Value Iteration Example: noise $= 0.2$, $\gamma = 0.9$, $r = 0.0$

$V_{101}(S^{2,2}) = \cdots$
$V_{101}(S^{2,1}) = \cdots$



http://ai.berkeley.edu/reinforcement.html

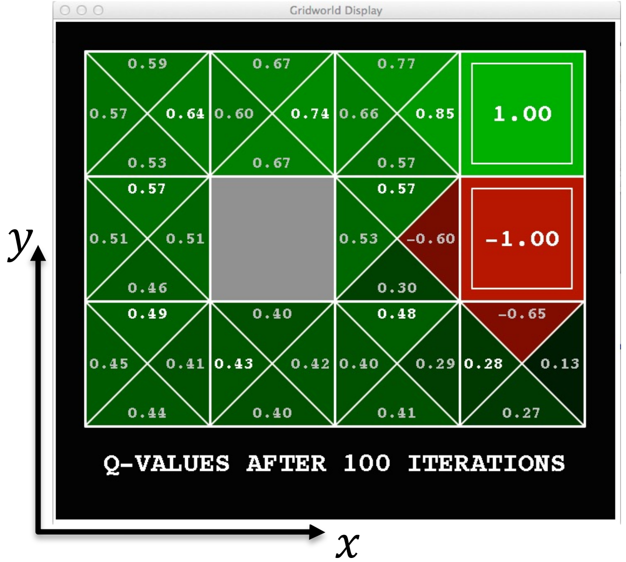
Dynamic Programming: Value Iteration

- How do we find optimal controllers for given (known) MDPs?
- Optimal Solver #1: Value Iteration (convergence guaranteed)

$$V_i(S^{x,y}) = \max_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S},r} \mathcal{P}(s',r|\,s,a)\,[r+\gamma V_{i-1}(s')]$$

- Value Iteration Example: noise $= 0.2$, $\gamma = 0.9$, $r = 0.0$

$V_{101}(S^{2,2}) = \cdots$
$V_{101}(S^{2,1}) = \cdots$



http://ai.berkeley.edu/reinforcement.html

Dynamic Programming: Value Iteration

- How do we find optimal controllers for given (known) MDPs?
- Optimal Solver #1: Value Iteration (convergence guaranteed)

$$V_i(S^{x,y}) = \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}, r} \mathcal{P}(s', r | \, s, a) \, [r + \gamma V_{i-1}(s')]$$

- Value Iteration Example: noise $= 0.2$, $\boldsymbol{\gamma} = \mathbf{1}$, $r = 0.0$
- How would it look like?

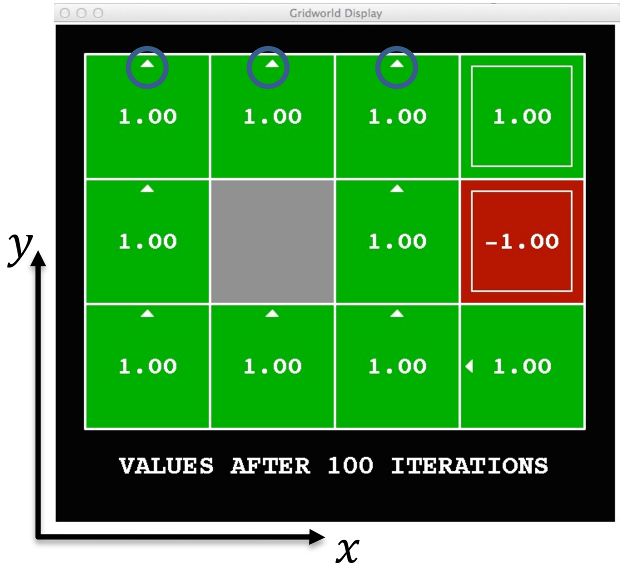
http://ai.berkeley.edu/reinforcement.html

Dynamic Programming: Value Iteration

- How do we find optimal controllers for given (known) MDPs?
- Optimal Solver #1: Value Iteration (convergence guaranteed)

$$V_i(S^{x,y}) = \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}, r} \mathcal{P}(s', r | s, a) \left[r + \gamma V_{i-1}(s') \right]$$

- Value Iteration Example: noise $= 0.0$, $\boldsymbol{\gamma} = \mathbf{1}$, $r = 0.0$
- How would it look like?



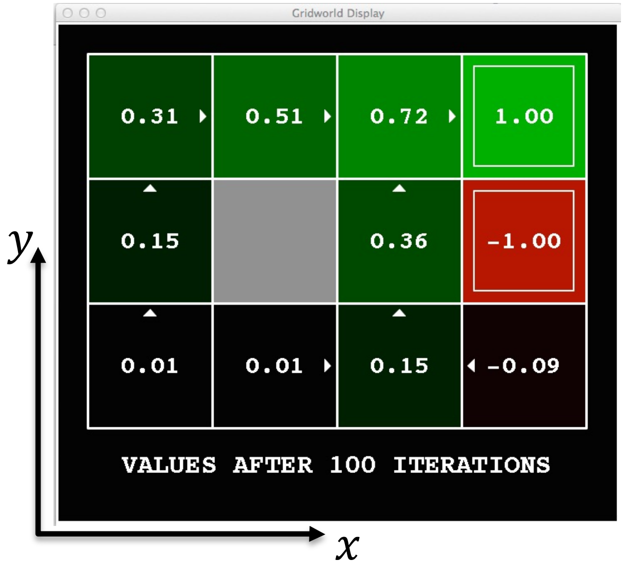
http://ai.berkeley.edu/reinforcement.html

Dynamic Programming: Value Iteration

- How do we find optimal controllers for given (known) MDPs?
- Optimal Solver #1: Value Iteration (convergence guaranteed)

$$V_i(S^{x,y}) = \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}, r} \mathcal{P}(s', r | \, s, a) \, [r + \gamma V_{i-1}(s')]$$

- Value Iteration Example: noise $= 0.2$, $\gamma = 0.9$, $\boldsymbol{r} = \mathbf{-0.1}$
- How would it look like?



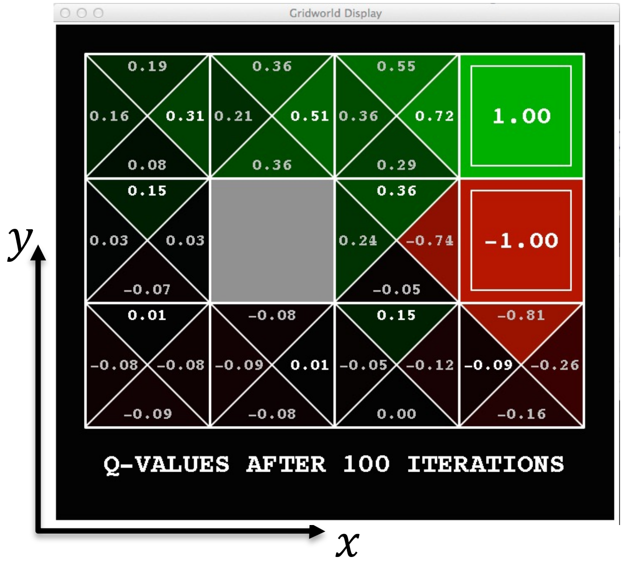
http://ai.berkeley.edu/reinforcement.html

Dynamic Programming: Value Iteration

- How do we find optimal controllers for given (known) MDPs?
- Optimal Solver #1: Value Iteration (convergence guaranteed)

$$V_i(S^{x,y}) = \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}, r} \mathcal{P}(s', r \mid s, a)\,[r + \gamma V_{i-1}(s')]$$

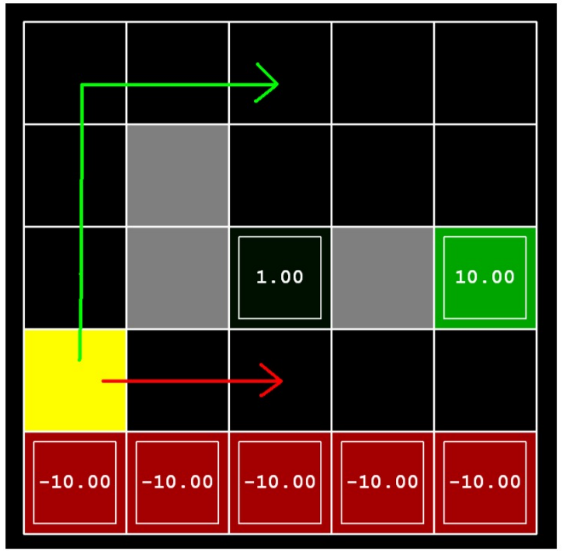
- Value Iteration Example: noise $= 0.2$, $\gamma = 0.9$, $r = -0.1$
- How would it look like?



http://ai.berkeley.edu/reinforcement.html

Dynamic Programming: Value Iteration

- How do we find optimal controllers for given (known) MDPs?
- Important: effect of environment noise and γ



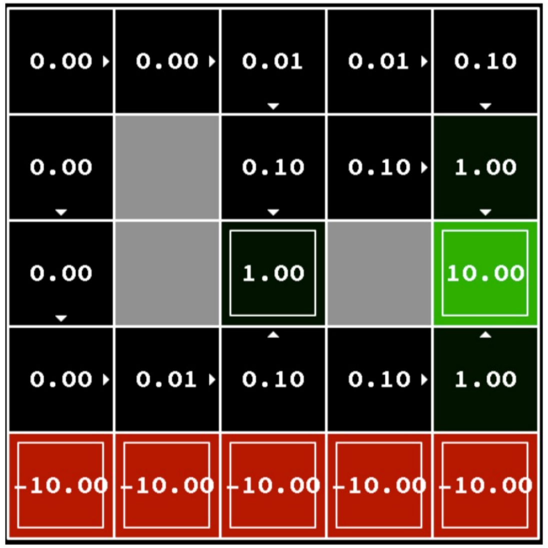
http://ai.berkeley.edu/reinforcement.html

Goals:
- Close exit (Reward +1.0)
- Distant exit (Reward +10.0)

Avoid:
- Cliff on bottom (Reward -10.0)

Dynamic Programming: Value Iteration

- How do we find optimal controllers for given (known) MDPs?
- Important: effect of environment noise and γ



http://ai.berkeley.edu/reinforcement.html

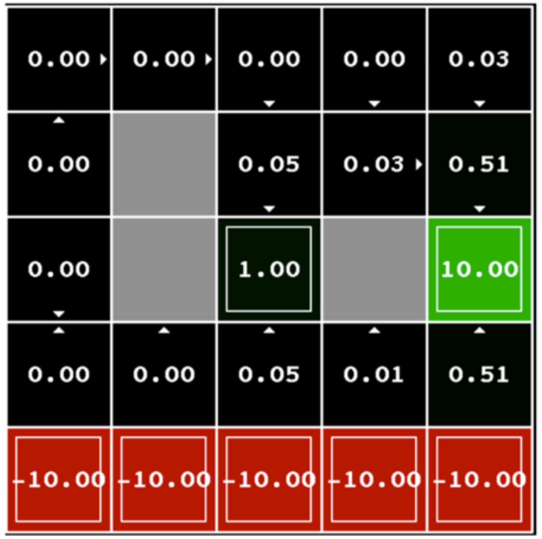
Solution for:
- $\gamma = 0.1$
- Noise = 0.0

Behavior:
- Prefers close exit
- Avoids cliff: No

Why?
- Since noise = 0.0 there is no risk
- $\gamma = 0.1$ forces early termination

Dynamic Programming: Value Iteration

- How do we find optimal controllers for given (known) MDPs?
- Important: effect of environment noise and γ



http://ai.berkeley.edu/reinforcement.html

Solution for:
- $\gamma = 0.1$
- Noise = 0.5

Behavior:
- Prefers close exit
- Avoids cliff: Yes

Why?
- Since noise = 0.5 there is high risk
- $\gamma = 0.1$ forces early termination

Dynamic Programming: Value Iteration

- How do we find optimal controllers for given (known) MDPs?
- Important: effect of environment noise and γ

http://ai.berkeley.edu/reinforcement.html

Solution for:
- $\gamma = 0.99$
- Noise = 0.0

Behavior:
- Prefers distant exit
- Avoids cliff: No

Why?
- Since noise = 0.0 there is no risk
- $\gamma = 0.99$ allows for distant exit

Dynamic Programming: Value Iteration

- How do we find optimal controllers for given (known) MDPs?
- Important: effect of environment noise and γ

http://ai.berkeley.edu/reinforcement.html

Solution for:
- $\gamma = 0.99$
- Noise = 0.5

Behavior:
- Prefers distant exit
- Avoids cliff: Yes

Why?
- Since noise = 0.5 there is high risk
- $\gamma = 0.99$ allows for distant exit

Dynamic Programming: Value Iteration

Hands-On:

https://cs.stanford.edu/people/karpathy/reinforcejs/gridworld_dp.html