CS 4700: Foundations of Artificial Intelligence

Fall 2019

Prof. Haym Hirsh

Lecture 9

October 16, 2019

Homework 2:
Due Friday, Oct 16, 2:40pm

This is the late submission
deadline with penalty waived
=> 0% credit if submitted after this

Grading being done on Saturday
(in time for studying for prelim)

Prelim Reminder:

Tuesday, October 22, 7:30-9:00 PM Statler Aud
(If you have a conflict you should have already informed us)

Covers material through today's lecture
Review questions are up on website

Review sessions:

Saturday, 3:00-5:00 PM Phillips 101
Prepared material, with TAs

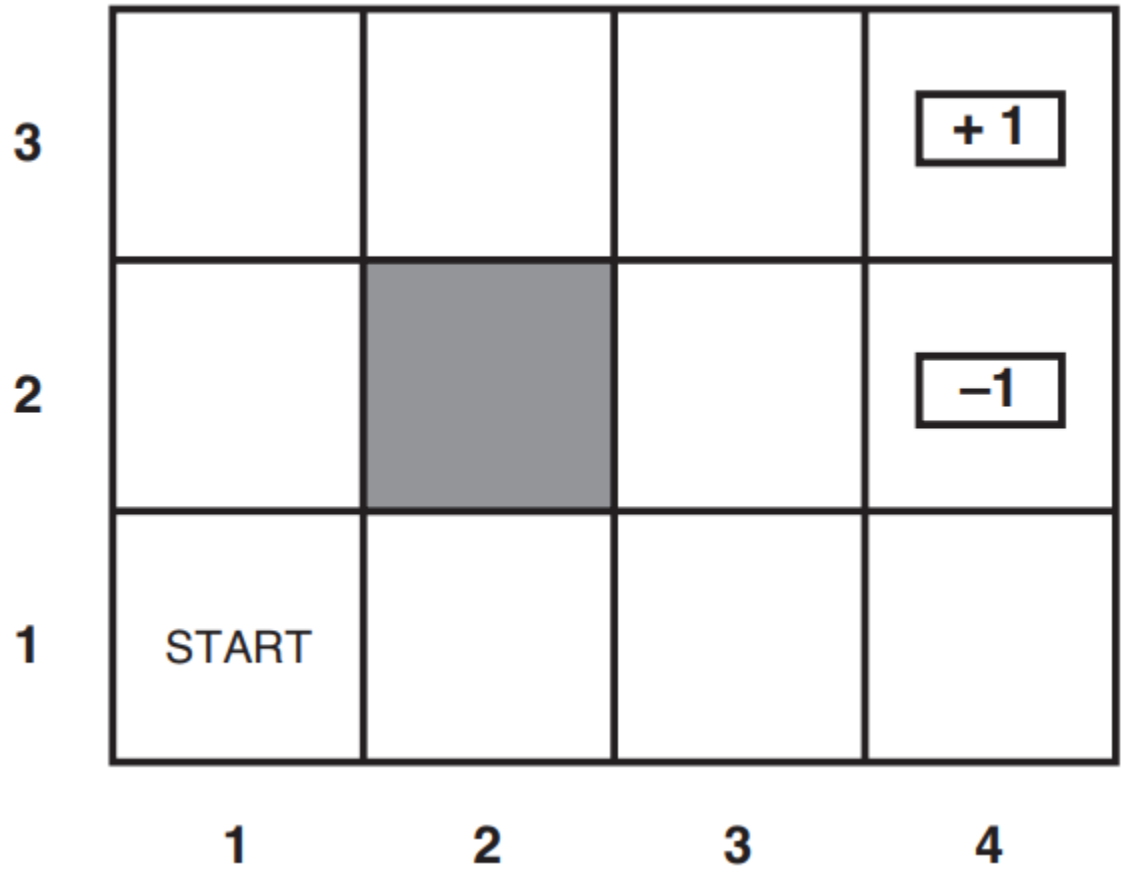
Sunday, 8:00-9:30pm, Gates G01
Answering questions from students, with Prof. Hirsh

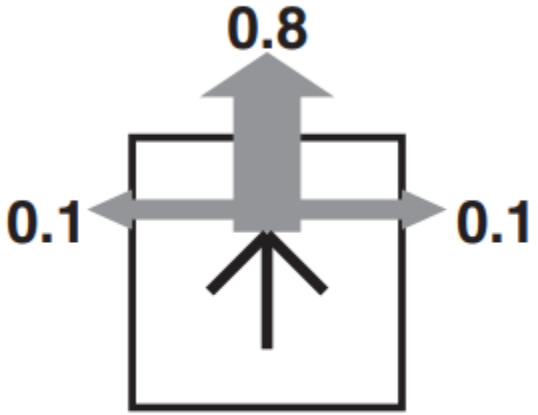
Karma Lectures

Th 10/17 11:40am, Gates G01	Martín Abadi, Google	"On the Theory and Practice of Software that Learns"
Th 10/24 4:30pm, Statler Aud	David Lazer, Northeastern	"Democracy, Today: Fake news, Social Networks, and Algorithms"
Th 11/14 11:40am, Gates G01	Christopher Potts, Stanford	TBD (Computational Linguistics/NLP)
Tu 11/14 4:30pm, Physical Sciences Building, Room 120	Marion Fourcade, UC Berkeley	"A Maussian Bargain: Accumulation by Gift in Digital Capitalism" (Algorithms and Fairness seminar)
Th 11/21 11:40am, Gates G01	Peter Stone, TU Austin	TBA (Machine Learning, Robotics)
Th 12/5 11:40am, Gates G01	Jim Baker	TBD (AI and Policy)

Th 10/31 11:40am, Gates G01	Tim Roughgarden (Cornell PhD '02)	TBD?

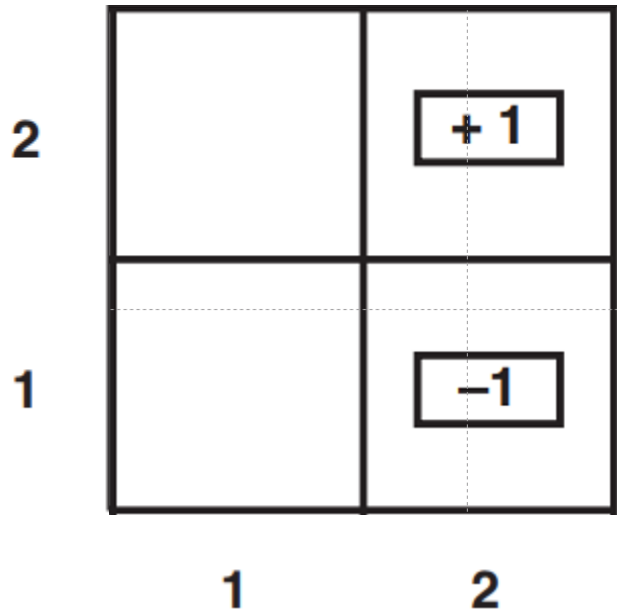
Example (Figure 16.1)

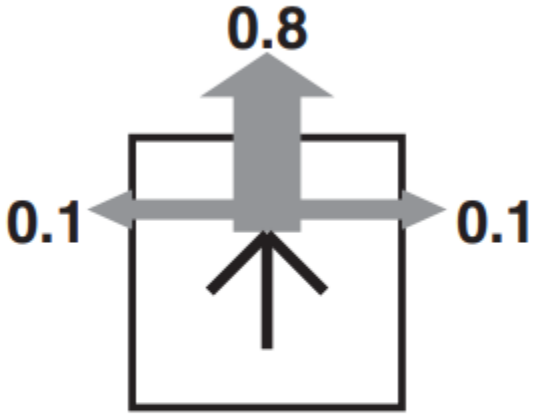




R(s,a,s')= - 0.04
(except in <4,2> and <4,3>)

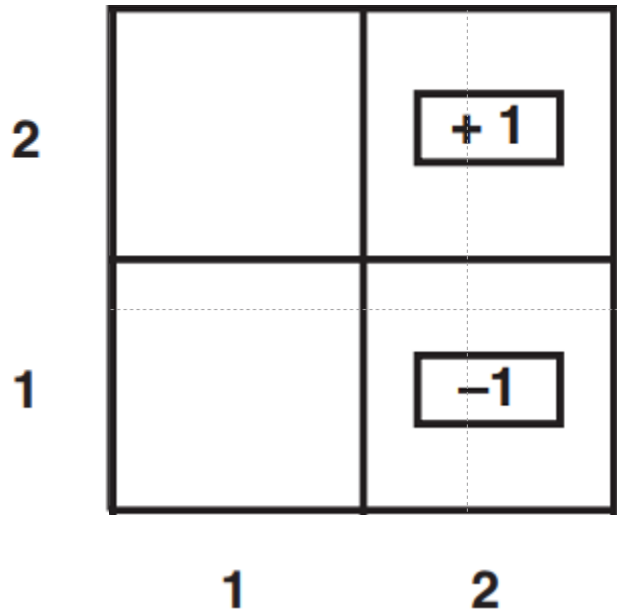
Simplified Example

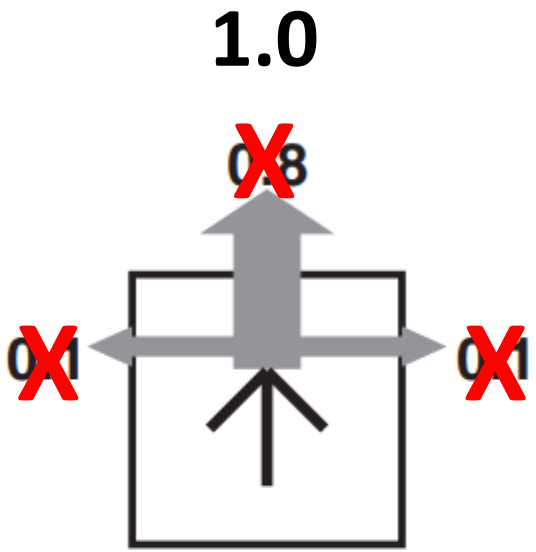




R(s,a,s')= - 0.04
(except in <2,1> and <2,2>)

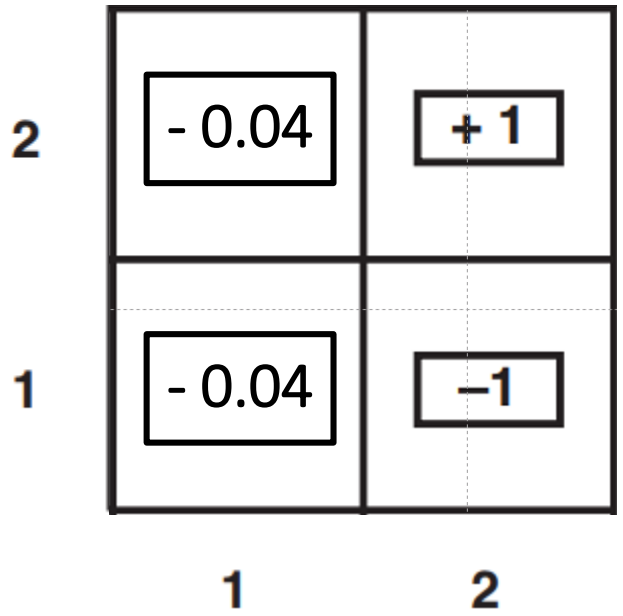
New Example

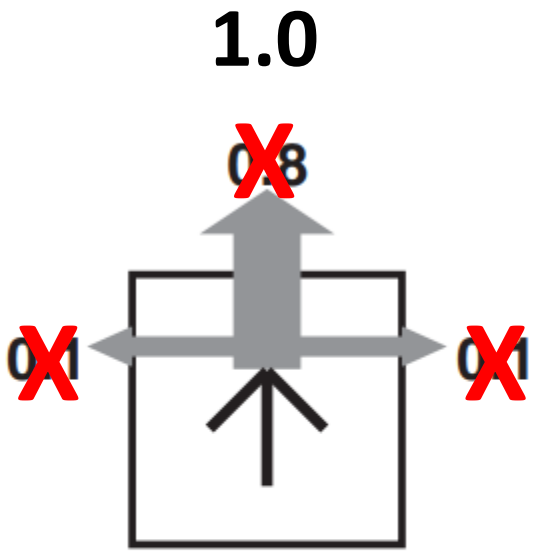




R(s,a,s')= - 0.04
(except in <2,1> and <2,2>)

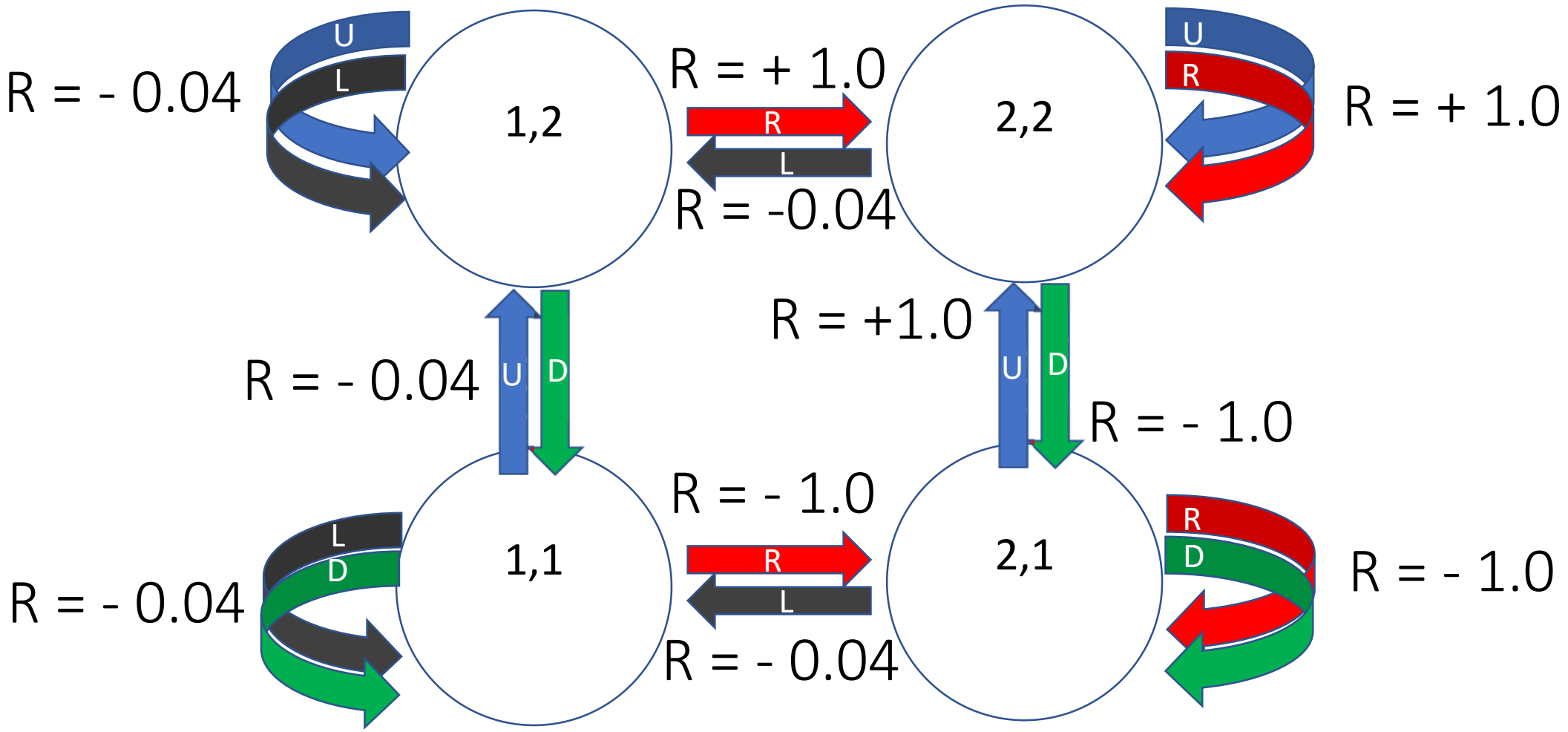
New Example



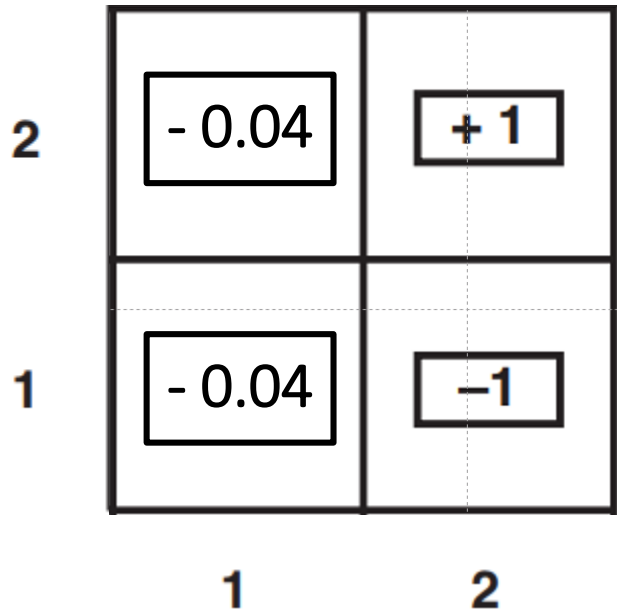


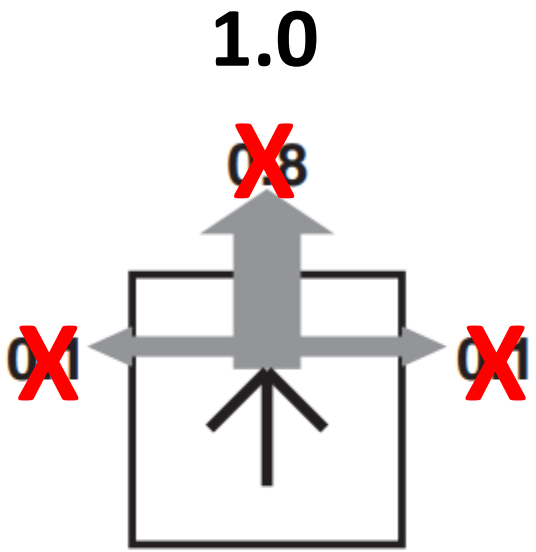
R(s,a,s')= - 0.04
(except in <2,1> and <2,2>)

Shorthand for This



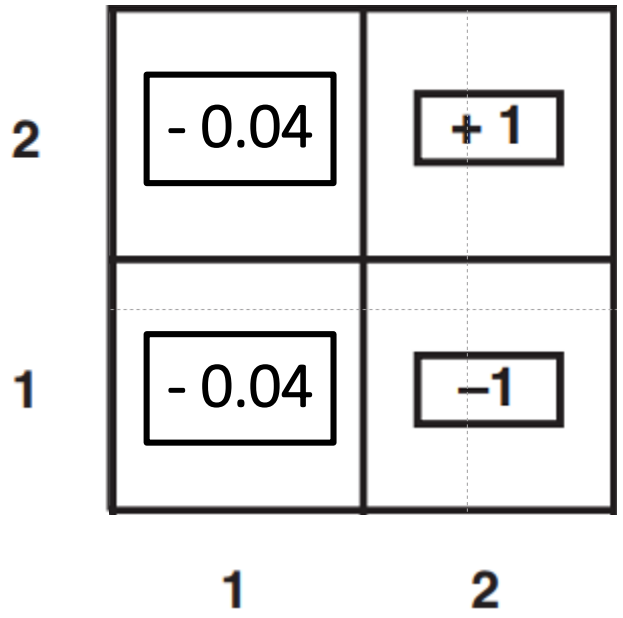
New Example

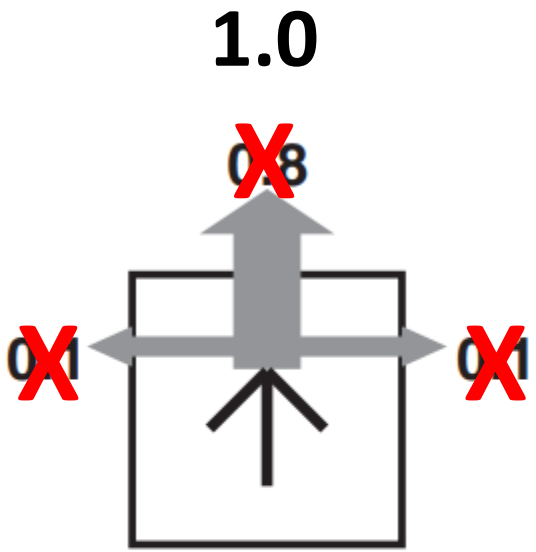




New Example

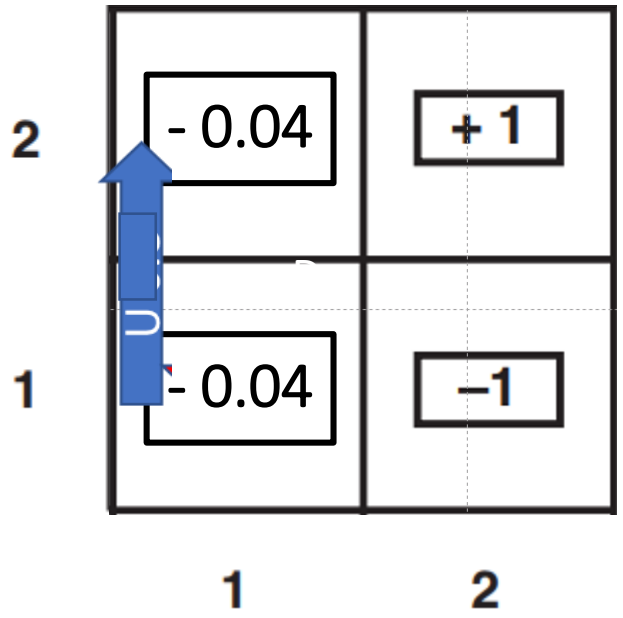
Π* = What is the best thing to do in each state?

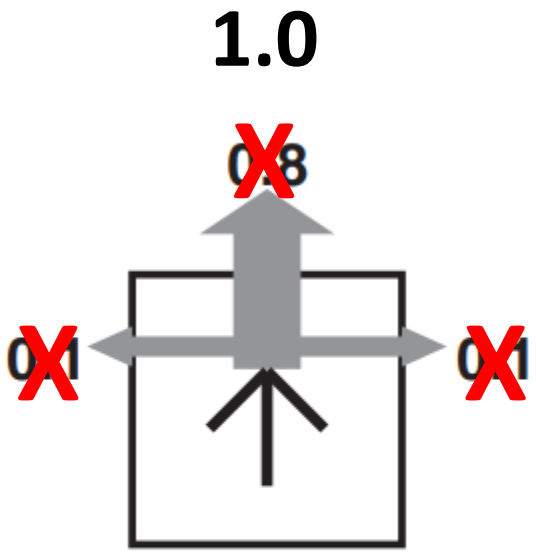




New Example

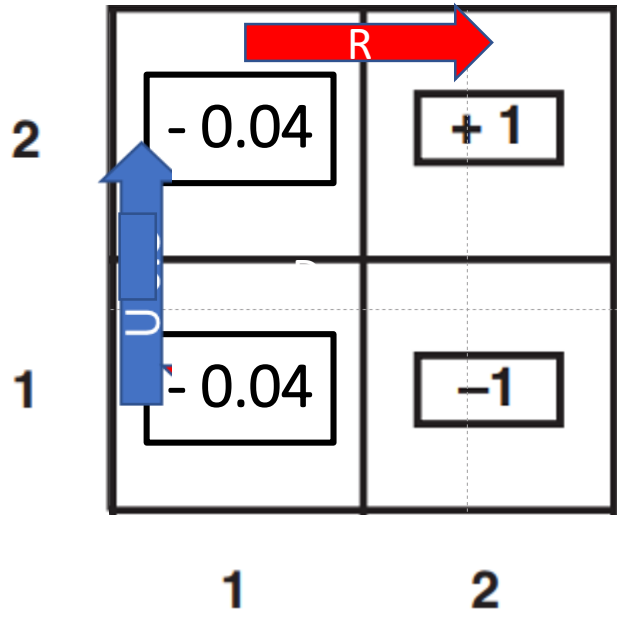
Π* = What is the best thing to do in each state?

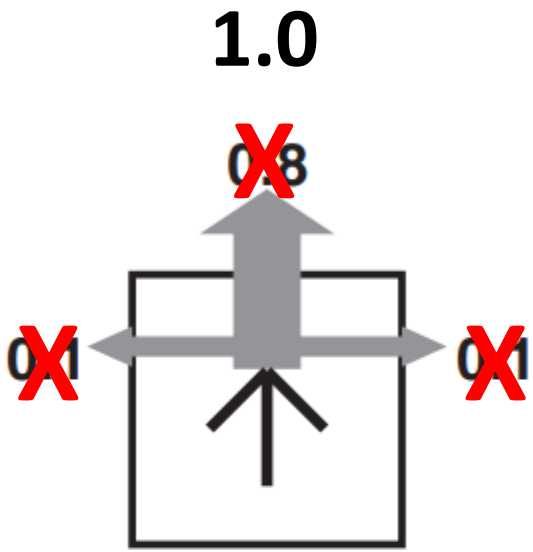




New Example

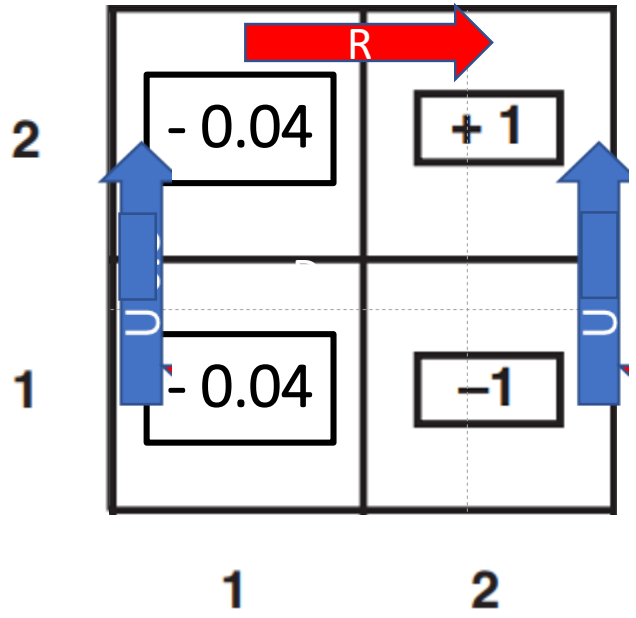
Π* = What is the best thing to do in each state?

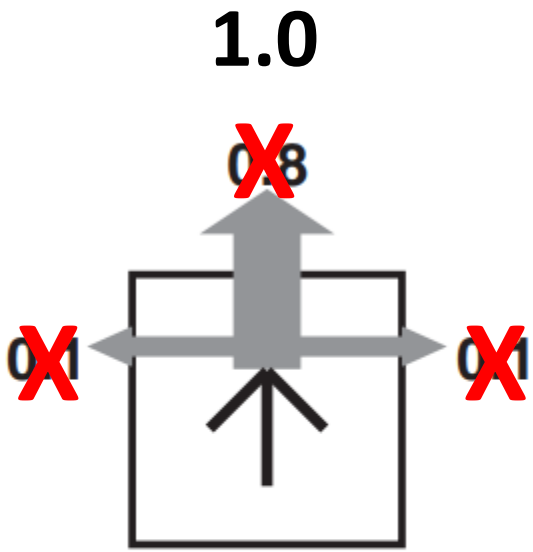




New Example

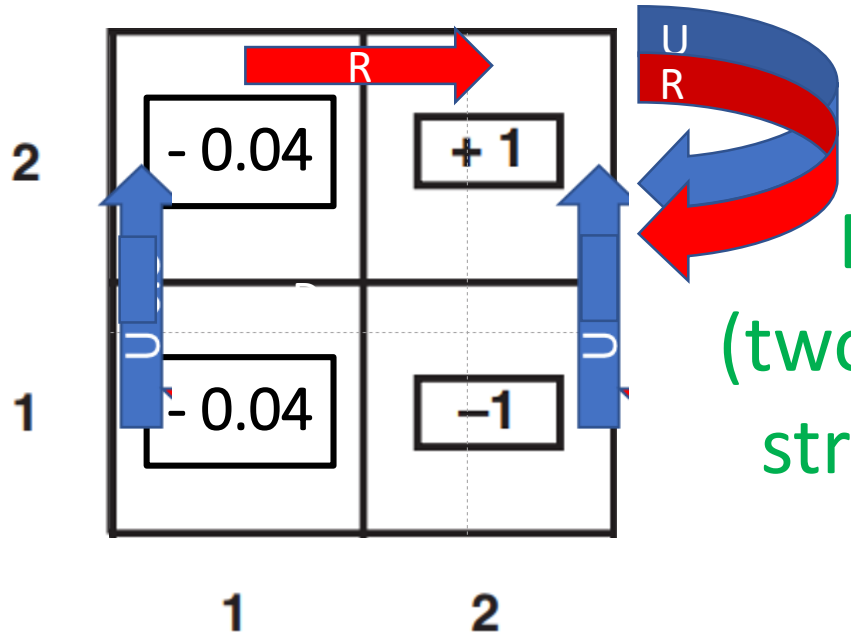
Π* = What is the best thing to do in each state?



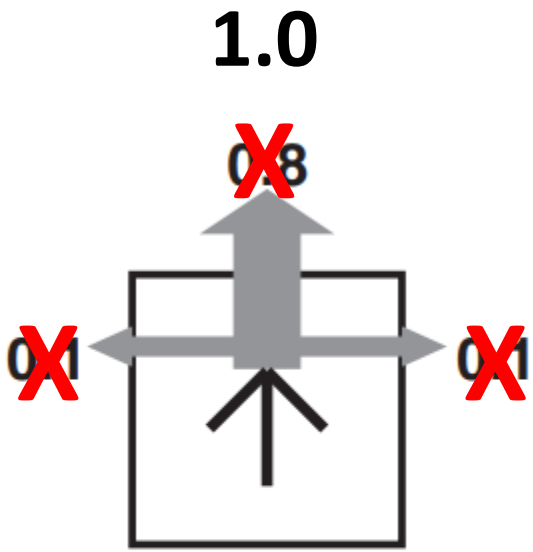


New Example

Π* = What is the best thing to do in each state?

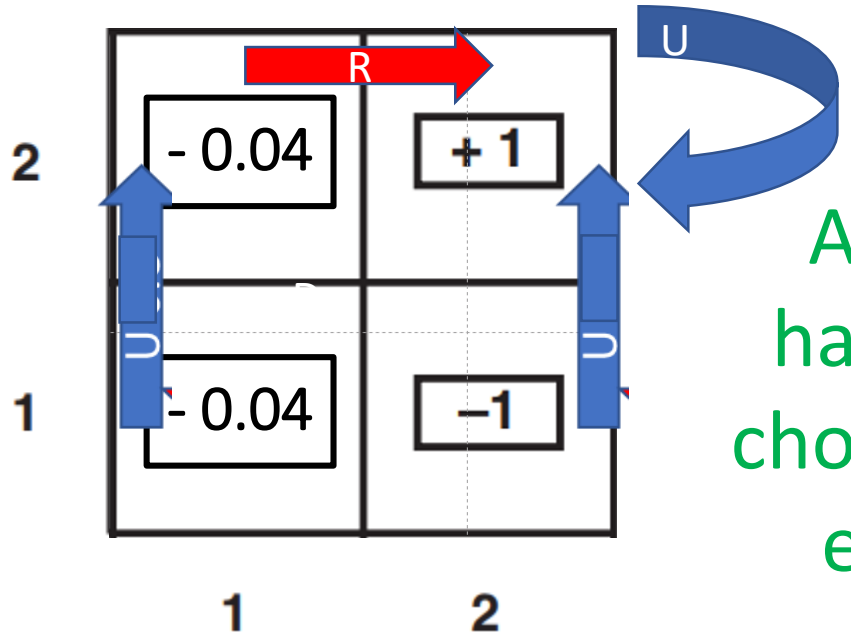


Either (two optimal strategies)

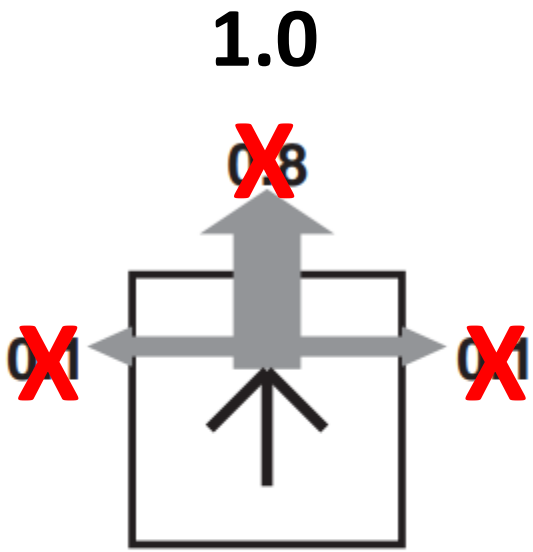


New Example

Π* = What is the best thing to do in each state?

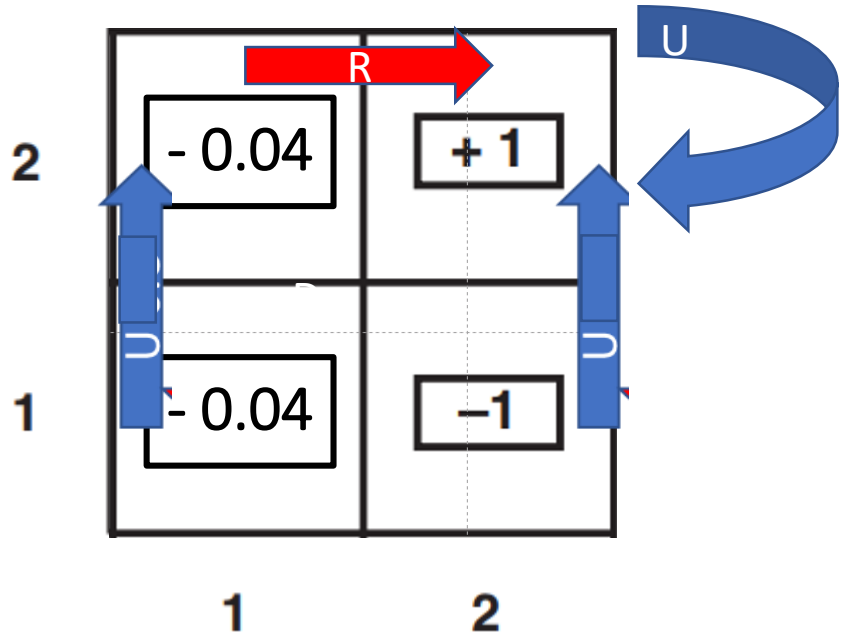


A policy will have only one choice (could be either one)



New Example

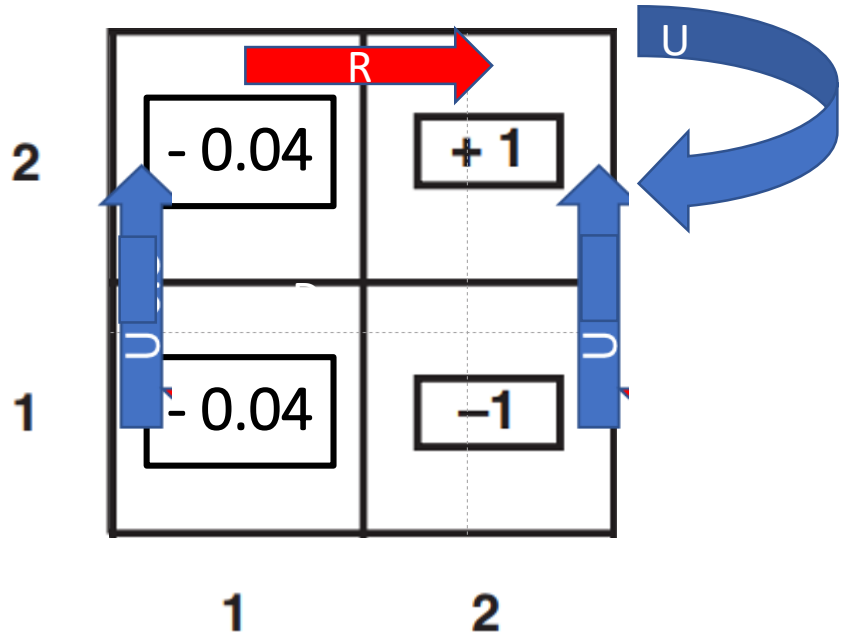
Π* = What is the best thing to do in each state?



U*(<2,2>) =
U*(<1,2>) =
U*(<2,1>) =
U*(<1,1>) =

New Example

Π* = What is the best thing to do in each state?



Value of optimal policy (parametrized by γ)

U*(<2,2>) =

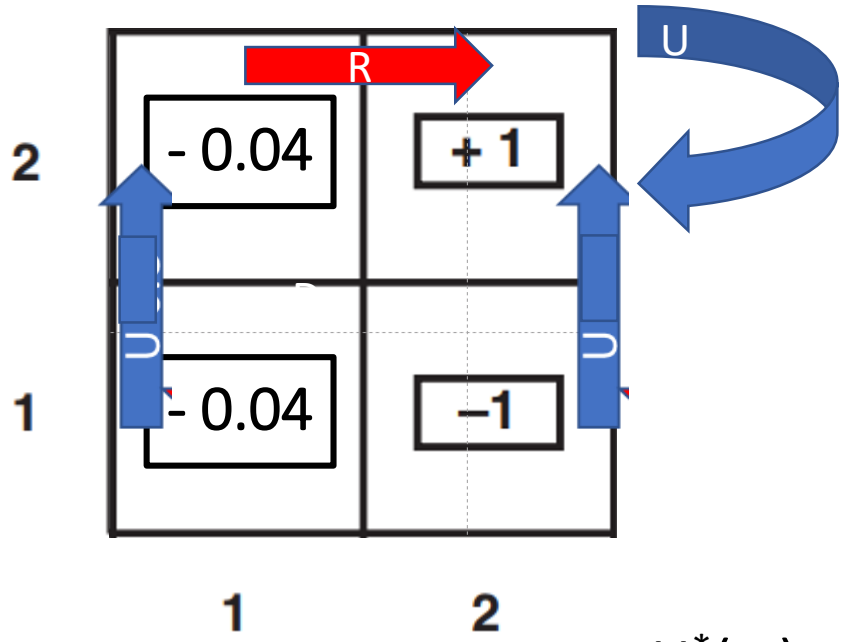
U*(<1,2>) =

U*(<2,1>) =

U*(<1,1>) =

New Example

Π* = What is the best thing to do in each state?



Value of optimal policy (parametrized by γ)

U*(<2,2>) =

U*(<1,2>) =

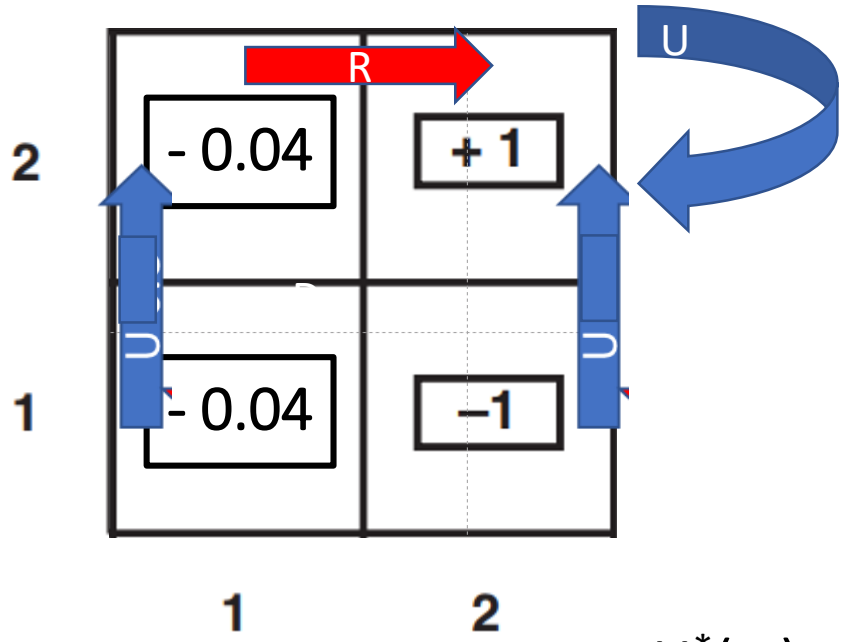
U*(<2,1>) =

U*(<1,1>) =

$U^*(s_0) = U^{\Pi^*}(s_0) = E \sum_{t=0}^{\infty} \gamma^t R(s_t, \Pi^*(s_t), \text{apply}(s_t, \Pi^*(s_t)))$

New Example

Π* = What is the best thing to do in each state?



Value of optimal policy (parametrized by γ)

U*(<2,2>) =

U*(<1,2>) =

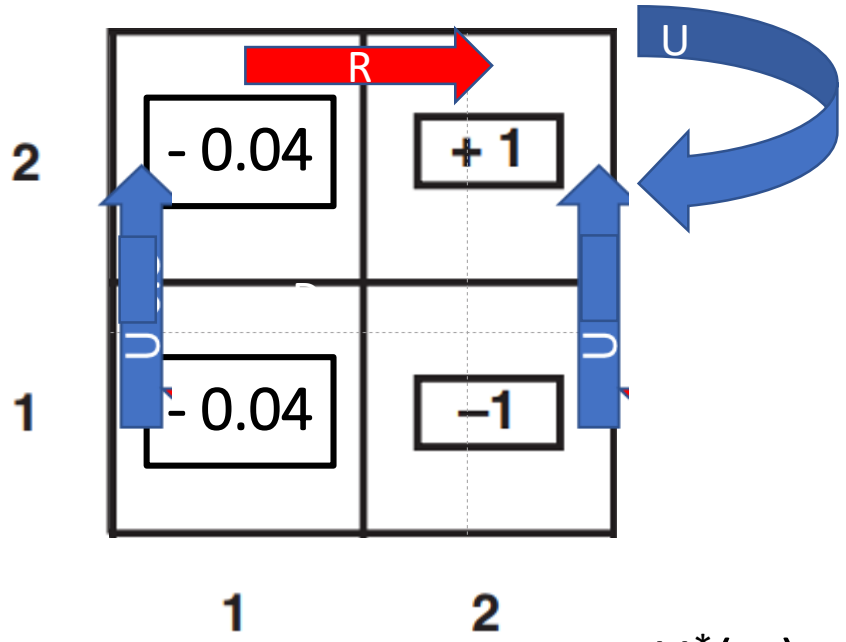
U*(<2,1>) =

U*(<1,1>) =

$U^*(s_0) = U^{\Pi^*}(s_0) = \sum_{t=0}^{\infty} \gamma^t R(s_t, \Pi^*(s_t), \text{apply}(s_t, \Pi^*(s_t)))$

New Example

Π* = What is the best thing to do in each state?



Value of optimal policy (parametrized by γ)

$U^*(\langle 2,2\rangle) = 1/(1 - \gamma)$

$U^*(\langle 1,2\rangle) = 1/(1 - \gamma)$

$U^*(\langle 2,1\rangle) = 1/(1 - \gamma)$

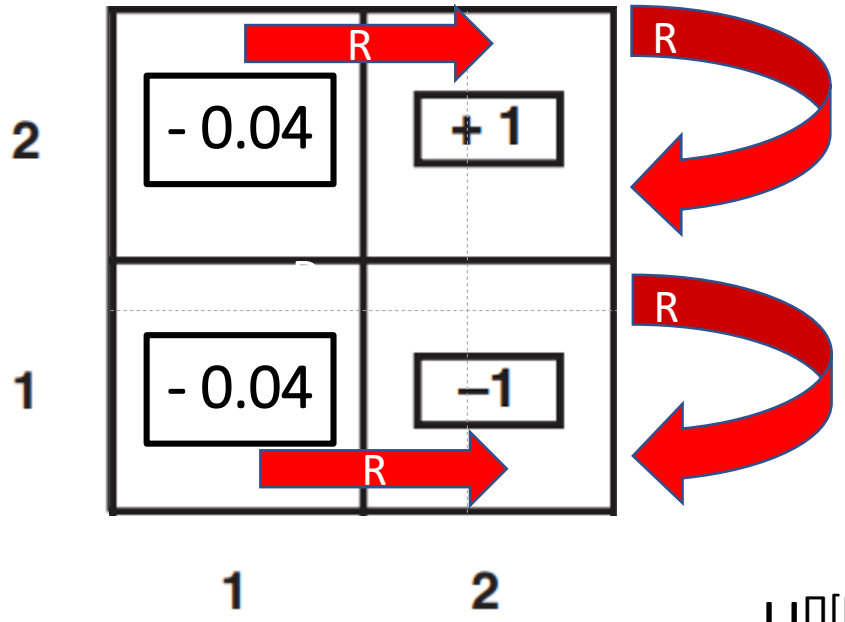
$U^*(\langle 1,1\rangle) = -0.04 + \gamma/(1 - \gamma)$

$U^*(s_0) = U^{\Pi^*}(s_0) = \sum_{t=0}^{\infty} \gamma^t R(s_t, \Pi^*(s_t), \text{apply}(s_t, \Pi^*(s_t)))$

New Example

$\Pi^{[R]}$ = Always go right

$U^{[R]}$ = …



Value of a particular policy

$U^{[R]}(<2,2>) = 1/(1 - \gamma)$

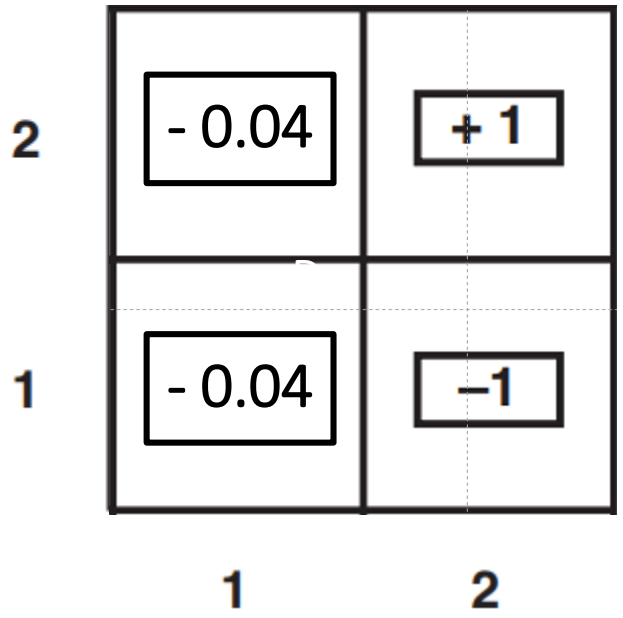
$U^{[R]}(<1,2>) = 1/(1 - \gamma)$

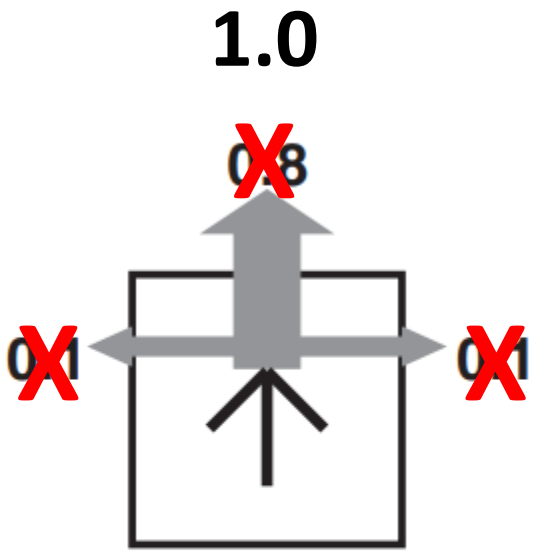
$U^{[R]}(<2,1>) = -1/(1 - \gamma)$

$U^{[R]}(<1,1>) = -1/(1 - \gamma)$

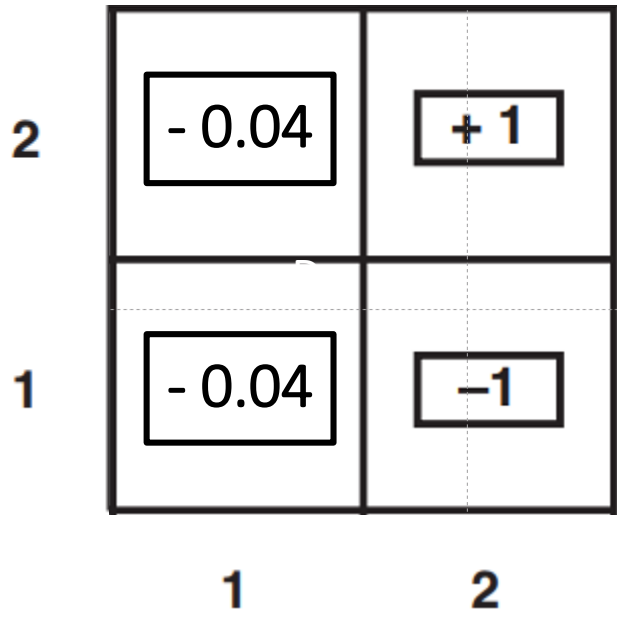
$$U^{\Pi[R]}(s_0) = \sum_{t=0}^{\infty} \gamma^t R(s_t, \Pi^{[R]}(s_t), \text{apply}(s_t, \Pi^{[R]}(s_t)))$$

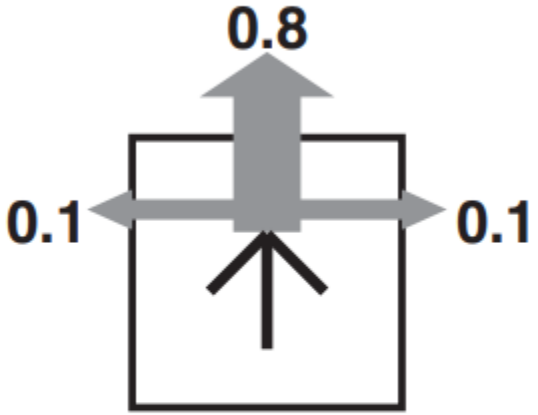
New Example





Original Simplified Example

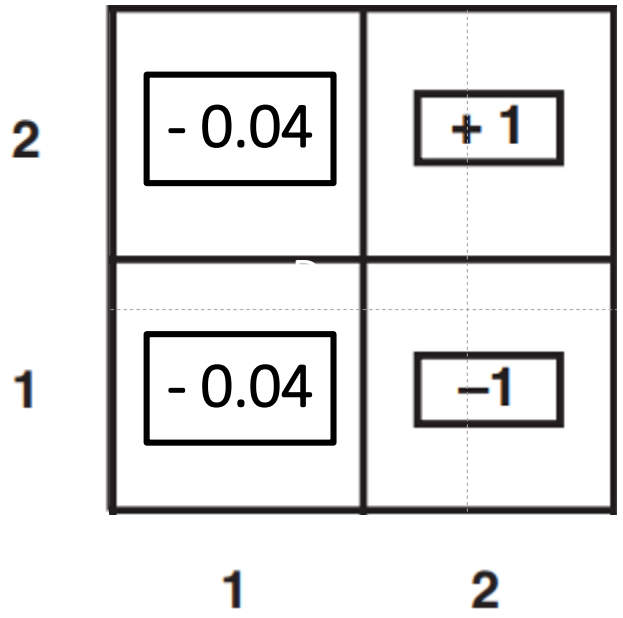


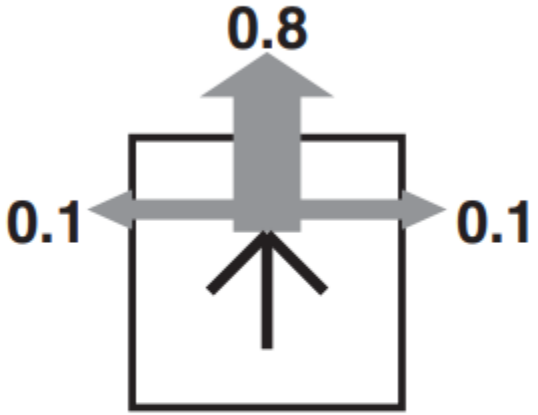


Original Simplified Example

Π* = What is the best thing to do in each state?

U* = What is its value in each state?

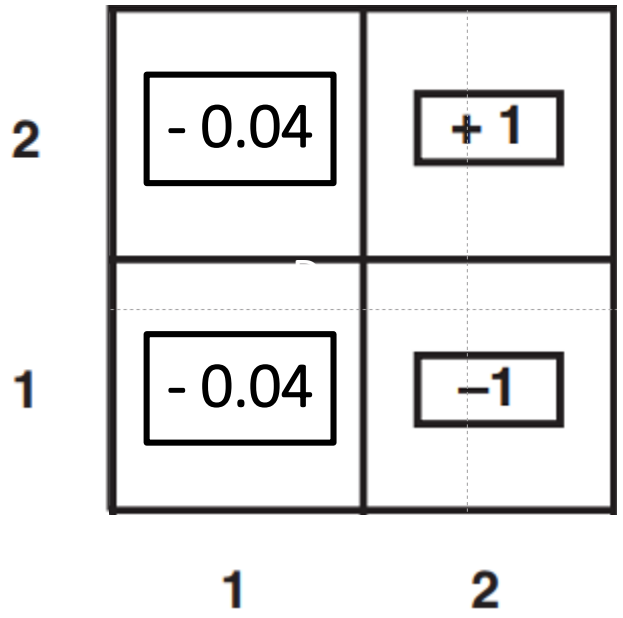


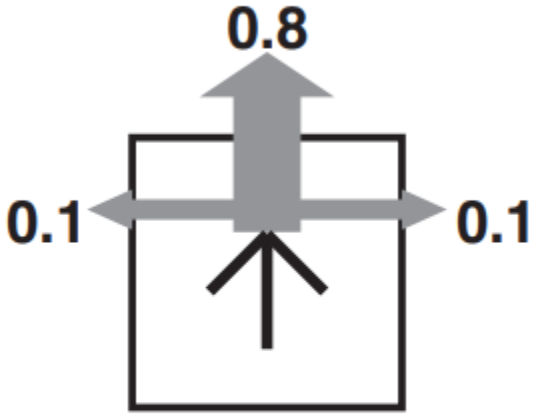


Original Simplified Example

Π^{UR} = Policy to go up in row 1, right in row 2

U^{UR} = What is its value in each state?





Example

Policy Π_{UR} = Go up if you're in row 1, otherwise go right

$$U^{UR}(<1,1>) = 0.8 \times [R(<1,1>,U,<1,2>) + \gamma \times U^{UR}(<1,2>)]$$
$$+ 0.1 \times [R(<1,1>,U,<1,1>) + \gamma \times U^{UR}(<1,1>)]$$
$$+ 0.1 \times [R(<1,1>,U,<2,1>) + \gamma \times U^{UR}(<2,1>)]$$

Example

Policy Π_{UR} = Go up if you're in row 1, otherwise go right

$$U^{UR}(<1,1>) = 0.8 \times [R(<1,1>,U,<1,2>) + \gamma \times U^{UR}(<1,2>)]$$
$$+ 0.1 \times [R(<1,1>,U,<1,1>) + \gamma \times U^{UR}(<1,1>)]$$
$$+ 0.1 \times [R(<1,1>,U,<2,1>) + \gamma \times U^{UR}(<2,1>)]$$
$$= 0.8 \times [-0.04 + \gamma \times U^{UR}(<1,2>)]$$
$$+ 0.1 \times [-0.04 + \gamma \times U^{UR}(<1,1>)]$$
$$+ 0.1 \times [-1.0 + \gamma \times U^{UR}(<2,1>)]$$
$$= -0.136 + \gamma \times [0.8 \times U^{UR}(<1,2>) + 0.1 \times U^{UR}(<1,1>) + 0.1 \times U^{UR}(<2,1>)]$$

Example

Policy Π_{UR} = Go up if you're in row 1, otherwise go right

$$U^{UR}(<1,1>) = 0.8 \times [R(<1,1>,U,<1,2>) + \gamma \times U^{UR}(<1,2>)]$$
$$+ 0.1 \times [R(<1,1>,U,<1,1>) + \gamma \times U^{UR}(<1,1>)]$$
$$+ 0.1 \times [R(<1,1>,U,<2,1>) + \gamma \times U^{UR}(<2,1>)]$$
$$= 0.8 \times [-0.04 + \gamma \times U^{UR}(<1,2>)]$$
$$+ 0.1 \times [-0.04 + \gamma \times U^{UR}(<1,1>)]$$
$$+ 0.1 \times [-1.0 + \gamma \times U^{UR}(<2,1>)]$$
$$= -0.136 + \gamma \times [0.8 \times U^{UR}(<1,2>) + 0.1 \times U^{UR}(<1,1>) + 0.1 \times U^{UR}(<2,1>)]$$

And similarly for <1,2>, <2,1>, and <2,2>

Example

Policy Π_{UR} = Go up if you're in row 1, otherwise go right

$$U^{UR}(<1,2>) = 0.8 \times [R(<1,2>,U,<2,2>) + \gamma \times U^{UR}(<2,2>)]$$
$$+ 0.1 \times [R(<1,2>,U,<1,1>) + \gamma \times U^{UR}(<1,1>)]$$
$$+ 0.1 \times [R(<1,2>,U,<1,2>) + \gamma \times U^{UR}(<1,2>)]$$
$$= 0.8 \times [+1.0 + \gamma \times U^{UR}(<2,2>)]$$
$$+ 0.1 \times [-0.04 + \gamma \times U^{UR}(<1,1>)]$$
$$+ 0.1 \times [-0.04 + \gamma \times U^{UR}(<1,2>)]$$
$$= +0.792 + \gamma \times [0.8 \times U^{UR}(<2,2>) + 0.1 \times U^{UR}(<1,1>) + 0.1 \times U^{UR}(<1,2>)]$$

Example

Policy Π_{UR} = Go up if you're in row 1, otherwise go right

$$U^{UR}(<2,1>) = 0.8 \times [R(<2,1>,U,<2,2>) + \gamma \times U^{UR}(<2,2>)]$$
$$+ 0.1 \times [R(<2,1>,U,<2,1>) + \gamma \times U^{UR}(<2,1>)]$$
$$+ 0.1 \times [R(<2,1>,U,<1,1>) + \gamma \times U^{UR}(<1,1>)]$$
$$= 0.8 \times [+1.0 + \gamma \times U^{UR}(<2,2>)]$$
$$+ 0.1 \times [-1.0 + \gamma \times U^{UR}(<2,1>)]$$
$$+ 0.1 \times [-0.04 + \gamma \times U^{UR}(<1,1>)]$$
$$= +0.696 + \gamma \times [0.8 \times U^{UR}(<2,2>) + 0.1 \times U^{UR}(<2,1>) + 0.1 \times U^{UR}(<1,1>)]$$

Example

Policy Π_{UR} = Go up if you're in row 1, otherwise go right

$$U^{UR}(<2,2>) = 0.8 \times [R(<2,2>,U,<2,2>) + \gamma \times U^{UR}(<2,2>)]$$
$$+ 0.1 \times [R(<2,2>,U,<2,2>) + \gamma \times U^{UR}(<2,2>)]$$
$$+ 0.1 \times [R(<2,2>,U,<2,1>) + \gamma \times U^{UR}(<2,1>)]$$
$$= 0.8 \times [+1.0 + \gamma \times U^{UR}(<2,2>)]$$
$$+ 0.1 \times [+1.0 + \gamma \times U^{UR}(<2,2>)]$$
$$+ 0.1 \times [-1.0 + \gamma \times U^{UR}(<2,1>)]$$
$$= +0.80 + \gamma \times [0.9 \times U^{UR}(<2,2>) + 0.1 \times U^{UR}(<2,1>)]$$

Example

Policy Π_{UR} = Go up if you're in row 1, otherwise go right

$U^{UR}(<1,1>) = -0.136 + \gamma \times [0.8 \times U^{UR}(<1,2>) + 0.1 \times U^{UR}(<1,1>) + 0.1 \times U^{UR}(<2,1>)]$

$U^{UR}(<1,2>) = +0.792 + \gamma \times [0.8 \times U^{UR}(<2,2>) + 0.1 \times U^{UR}(<1,1>) + 0.1 \times U^{UR}(<1,2>)]$

$U^{UR}(<2,1>) = +0.696 + \gamma \times [0.8 \times U^{UR}(<2,2>) + 0.1 \times U^{UR}(<2,1>) + 0.1 \times U^{UR}(<1,1>)]$

$U^{UR}(<2,2>) = +0.800 + \gamma \times [0.9 \times U^{UR}(<2,2>) + 0.1 \times U^{UR}(<2,1>)]$

Example

Policy Π_{UR} = Go up if you're in row 1, otherwise go right

If we use A for $U^{UR}(<1,1>)$, B for $U^{UR}(<2,1>)$, C for $U^{UR}(<1,2>)$ and D for $U^{UR}(<2,2>)$
(to make it easier to read)

$$A = -0.136 + \gamma \times [0.8 \times C + 0.1 \times A + 0.1 \times B]$$
$$C = +0.792 + \gamma \times [0.8 \times D + 0.1 \times A + 0.1 \times C]$$
$$B = +0.696 + \gamma \times [0.8 \times D + 0.1 \times B + 0.1 \times A]$$
$$D = +0.800 + \gamma \times [0.9 \times D + 0.1 \times B]$$

Example

Policy Π_{UR} = Go up if you're in row 1, otherwise go right

$$
\begin{aligned}
0.136 &= (0.1\gamma-1)A + 0.1\gamma B + 0.8\gamma C \\
-0.792 &= 0.1\gamma A + (0.1\gamma-1)C + 0.8\gamma D \\
-0.696 &= 0.1\gamma A + (0.1\gamma-1)B + 0.8\gamma D \\
-0.800 &= 0.1\gamma B + (0.9\gamma-1)D
\end{aligned}
$$

Four equations, four unknowns, for a given γ can now solve
(for example, with Gaussian Elimination)

Example: γ =0.5

Policy Π_{UR} = Go up if you're in row 1, otherwise go right

$$0.136 = -0.95A + 0.05B + 0.4C$$
$$-0.792 = 0.05A + 0.95C + 0.4D$$
$$-0.696 = 0.05A + -0.95B + 0.4D$$
$$-0.800 = 0.05B + -0.55D$$

Solve for A, B, C, and D to get the $U^{\Pi_{UR}}$ values for each state

MDPs and Reinforcement Learning: Game Plan

- How to figure out value of a given Π if you know S, R, and P
  - One approach: Gaussian Elimination
- How to figure out Π* if you know S and P but not R

(|S| finite and ~small)

Properties of Π*

- Optimal Policy Π*

  $\Pi^*(s) = E \sum_{t=0}^{\infty} \gamma^t R(s_t, \Pi^*(s_t), \text{apply}(s_t, \Pi^*(s_t)))$

  $= \text{argmax}_{\Pi} U^{\Pi}(s)$

  $= \text{argmax}_{a \in A} \sum_{s' \in S} P(s'|s,a)[R(s,a,s') + \gamma U^{\Pi^*}(s')]$

  [This says that some action gives the best outcome. Find it. To do so consider each action a and figure out its expected immediate reward R(s,a,s') together with what you would get if you acted optimally from that point onwards $U^{\Pi^*}(s')$. However that second term is what you get one step in the future, so you have to discount it by γ.]

- Bellman Equation

  $U^*(s) = U^{\Pi^*}(s) = \max_{a \in A} \sum_{s' \in S} P(s'|s,a)[R(s,a,s') + \gamma U^{\Pi^*}(s')]$

  [This is the same as for Π*(s) above, just max rather than argmax]

  [This will be important – it is the foundation of various algorithms]

$$U^*(<1,2>) = \max_{a \in A} \{0.8 \times [-0.4 + \gamma U^*(<1,2>)] + 0.1 \times [-0.4 + \gamma U^*(<1,2>)] + 0.1 \times [-0.4 + \gamma U^*(<2,2>)] \quad \text{UP}$$

$$0.8 \times [-0.4 + \gamma U^*(<1,1>)] + 0.1 \times [-0.4 + \gamma U^*(<1,2>)] + 0.1 \times [-0.4 + \gamma U^*(<2,2>)] \quad \text{DOWN}$$

$$0.8 \times [-0.4 + \gamma U^*(<1,2>)] + 0.1 \times [-0.4 + \gamma U^*(<1,2>)] + 0.1 \times [-0.4 + \gamma U^*(<1,1>)] \quad \text{LEFT}$$

$$0.8 \times [-0.4 + \gamma U^*(<2,2>)] + 0.1 \times [-0.4 + \gamma U^*(<1,2>)] + 0.1 \times [-0.4 + \gamma U^*(<1,1>)]\} \quad \text{RIGHT}$$

$$\Pi^*(<1,2>) = \text{argmax}_{a \in A} \{0.8 \times [-0.4 + \gamma U^*(<1,2>)] + 0.1 \times [-0.4 + \gamma U^*(<1,2>)] + 0.1 \times [-0.4 + \gamma U^*(<2,2>)] \quad \text{UP}$$

$$0.8 \times [-0.4 + \gamma U^*(<1,1>)] + 0.1 \times [-0.4 + \gamma U^*(<1,2>)] + 0.1 \times [-0.4 + \gamma U^*(<2,2>)] \quad \text{DOWN}$$

$$0.8 \times [-0.4 + \gamma U^*(<1,2>)] + 0.1 \times [-0.4 + \gamma U^*(<1,2>)] + 0.1 \times [-0.4 + \gamma U^*(<1,1>)] \quad \text{LEFT}$$

$$0.8 \times [-0.4 + \gamma U^*(<2,2>)] + 0.1 \times [-0.4 + \gamma U^*(<1,2>)] + 0.1 \times [-0.4 + \gamma U^*(<1,1>)]\} \quad \text{RIGHT}$$

If you write down U* for every state it no longer gives a set of linear equations to which you can apply Gaussian elimination

If you write down U* for every state it no longer gives a set of linear equations that you can apply Gaussian elimination to

What to do?

Policy Iteration algorithm
(among various algorithms for doing this)

Uses $U^*(s) = U^{\Pi^*}(s) = \max_{a \in A} \sum_{s' \in S} P(s'|s,a)[R(s,a,s') + \gamma U^{\Pi^*}(s')]$