DMR Worked Examples

Yordan Hristov

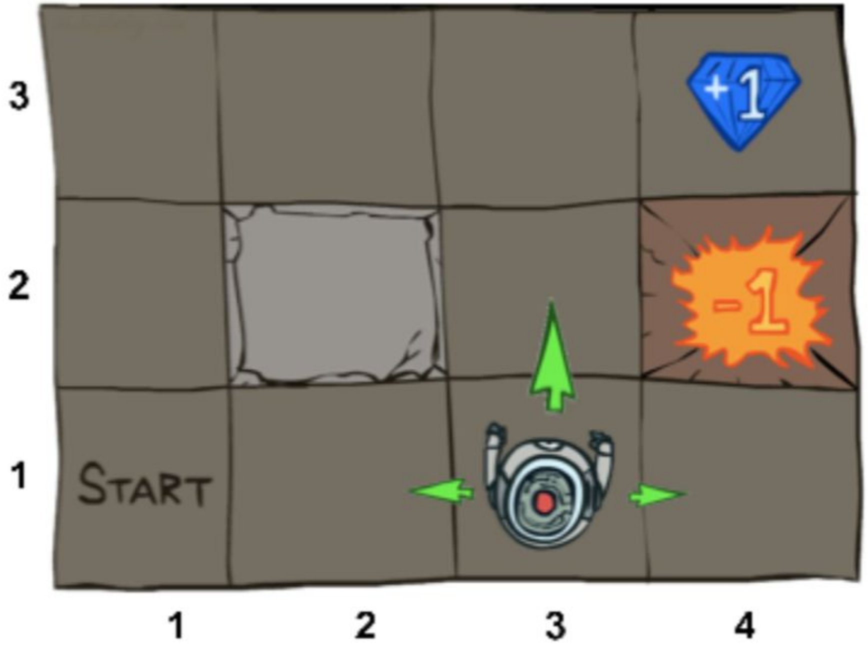
Announcement:

CW2 delayed due to Admin reasons
Released tomorrow

Agenda

1. Value Iteration & Policy Iteration
2. Causality
3. Game Theory (optional)

Value Iteration & Policy Iteration



[1]

An MDP is defined by:

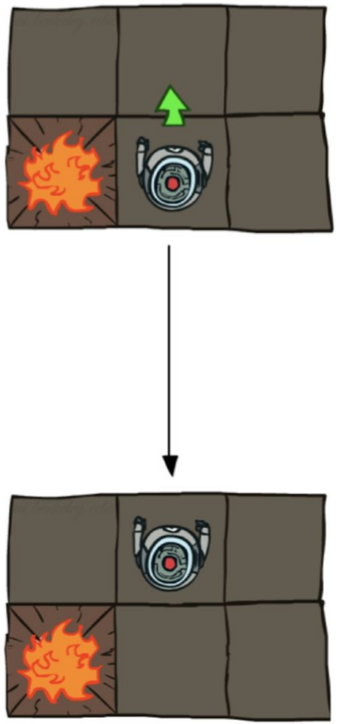
- Set of states S
- Set of actions A
- Transition function $P(s' \mid s, a)$
- Reward function $R(s, a, s')$
- Start state s_0
- Discount factor γ
- Horizon H

[1]

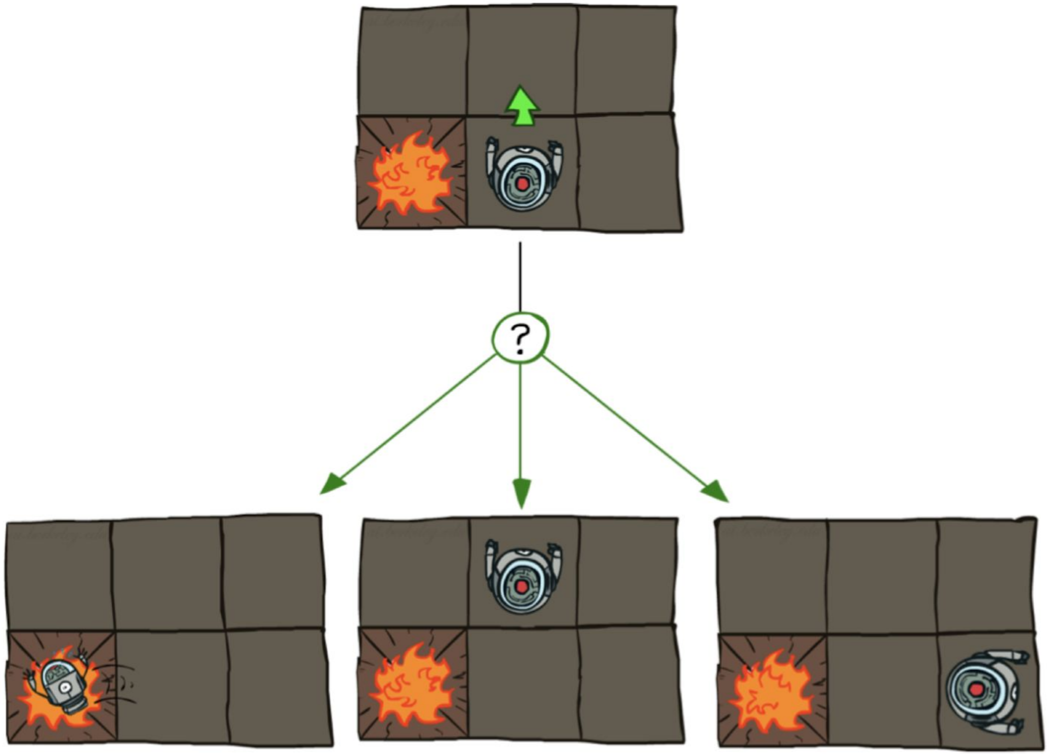
$$V^*(s) = \max_a \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V^*(s')]$$

[1]

Deterministic Grid World



Stochastic Grid World



[1]

Value Iteration & Policy Iteration

Algorithm:

Start with $V_0^*(s) = 0$ for all s.

For k = 1, ... , H:

For all states s in S:

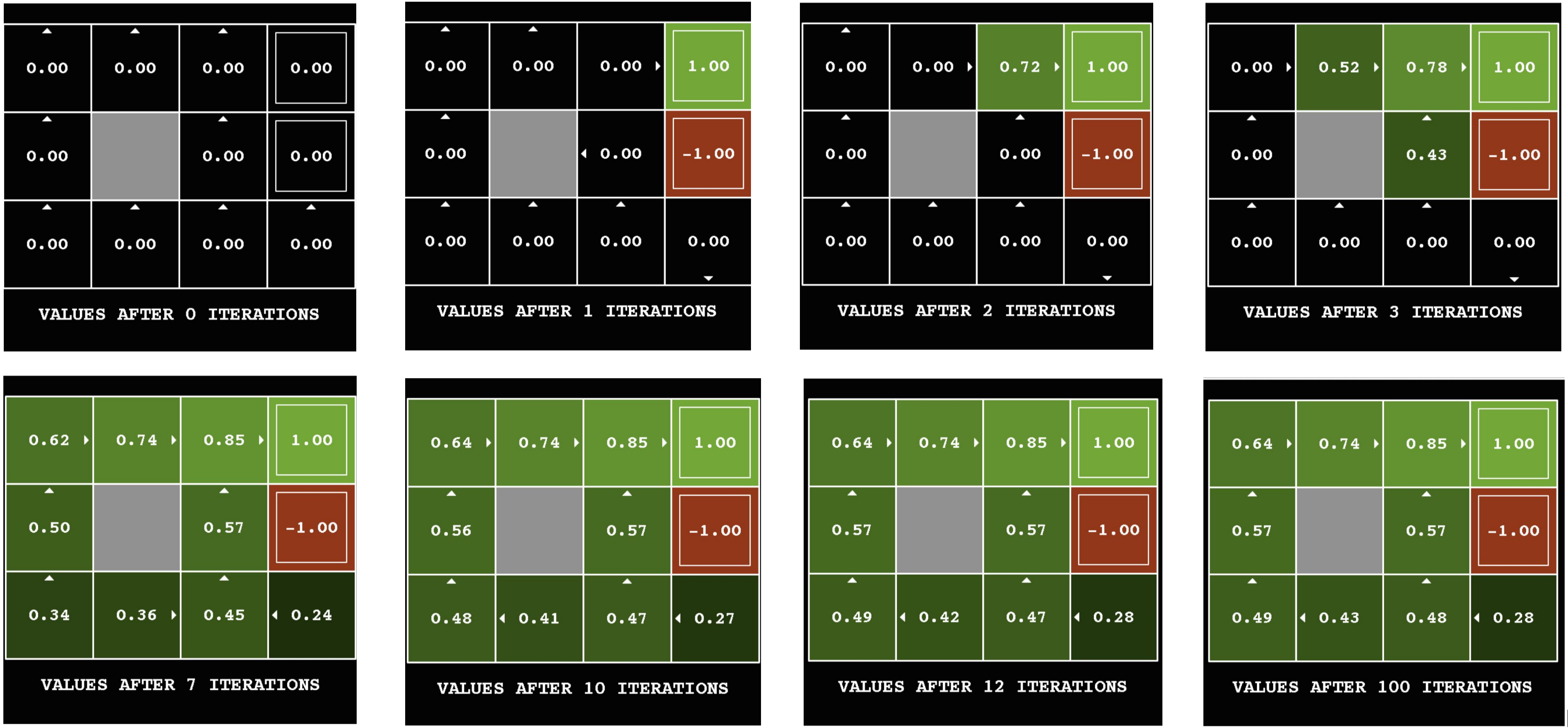
$$V_k^*(s) \leftarrow \max_a \sum_{s'} P(s'|s,a)\left(R(s,a,s') + \gamma V_{k-1}^*(s')\right)$$

$$\pi_k^*(s) \leftarrow \arg\max_a \sum_{s'} P(s'|s,a)\left(R(s,a,s') + \gamma V_{k-1}^*(s')\right)$$

This is called a value update or Bellman update/back-up

[1]

Value Iteration & Policy Iteration

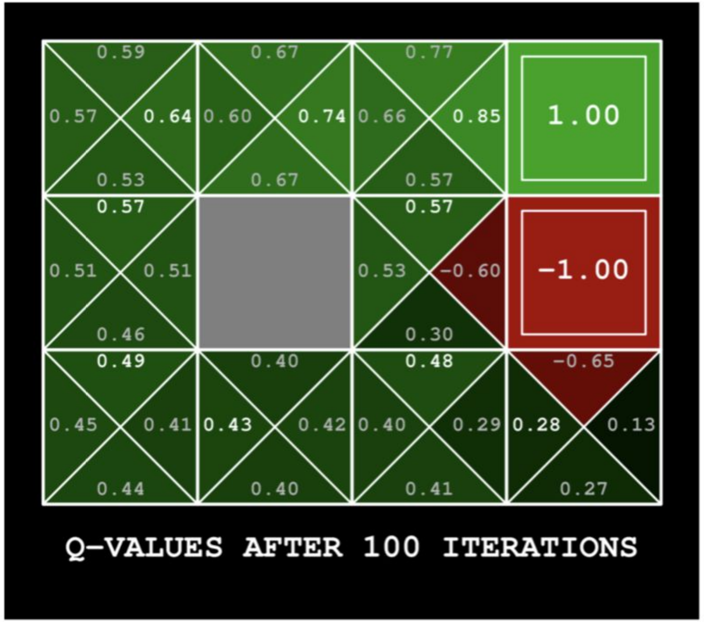


[1]

Value Iteration & Policy Iteration

$$Q^*_{k+1}(s,a) \leftarrow \sum_{s'} P(s'|s,a)(R(s,a,s') + \gamma \max_{a'} Q^*_k(s',a'))$$ [1]

k = 100



[1]

- Same like value iteration but instead of only keeping the max utility function - max Q(s,a), keep track of the utility values for all actions in a given state - Q(s,a).
- Policy is still greedily derived by taking the action with max utility

Value Iteration & Policy Iteration

- Policy evaluation for current policy π_k :
  - Iterate until convergence

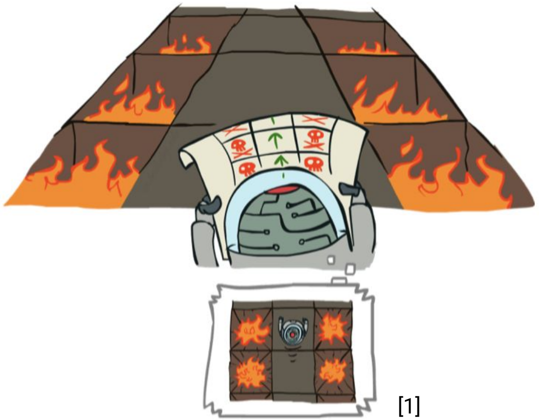
$$V_{i+1}^{\pi_k}(s) \leftarrow \sum_{s'} P(s'|s,\pi_k(s))\left[R(s,\pi(s),s') + \gamma V_i^{\pi_k}(s')\right]$$

- Policy improvement: find the best action according to one-step look-ahead

$$\pi_{k+1}(s) \leftarrow \arg\max_a \sum_{s'} P(s'|s,a)\left[R(s,a,s') + \gamma V^{\pi_k}(s')\right]$$

[1]

Value Iteration & Policy Iteration



[1]

Always Go Right

[1]

Always Go Forward

[1]

Value Iteration & Policy Iteration

- Both value iteration and policy Iteration compute optimal values and policies

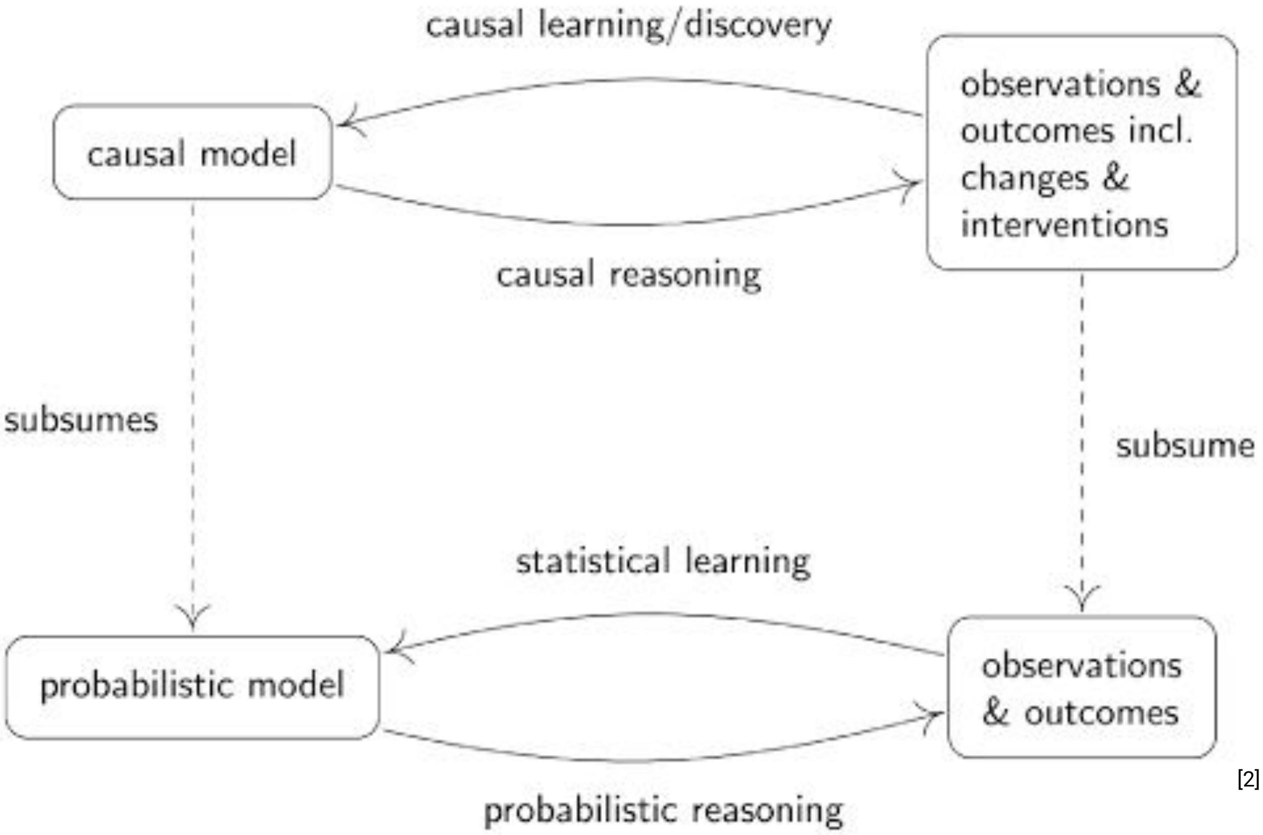
- In value iteration:
    - Every iteration updates both the value and (implicitly) the policy
    - The policy is not tracked but is easily accessible through the max over actions

- In policy iteration:
    - We do several passes that update the value function of a fixed policy. Each pass is fast since we consider only one action, not all of them
    - After the policy is evaluated - value function converges/is calculated, a new policy is extracted
    - The new policy will be better or the same => done

- Both are dynamic programs for solving MDPs

Causality

Causality



causal learning/discovery
causal model
observations & outcomes incl. changes & interventions
causal reasoning
subsumes
subsume
statistical learning
probabilistic model
observations & outcomes
probabilistic reasoning

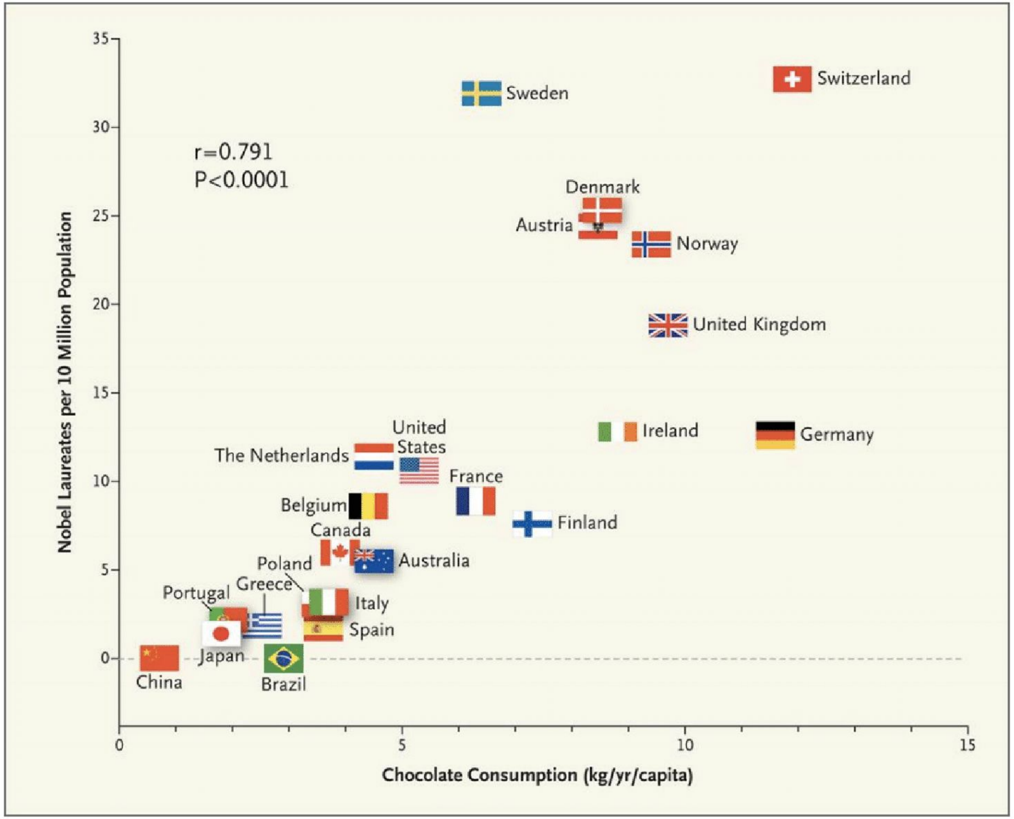
[2]

Causality

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y\|x)$	Seeing	What is? How would seeing X change my belief in Y?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y\|do(x), z)$	Doing Intervening	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x\|x', y')$	Imagining, Retrospection	Why? Was it X that caused Y? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past 2 years?

[2]

Causality



F. H. Messerli: *Chocolate Consumption, Cognitive Function, and Nobel Laureates*, N Engl J Med 2012 [2]

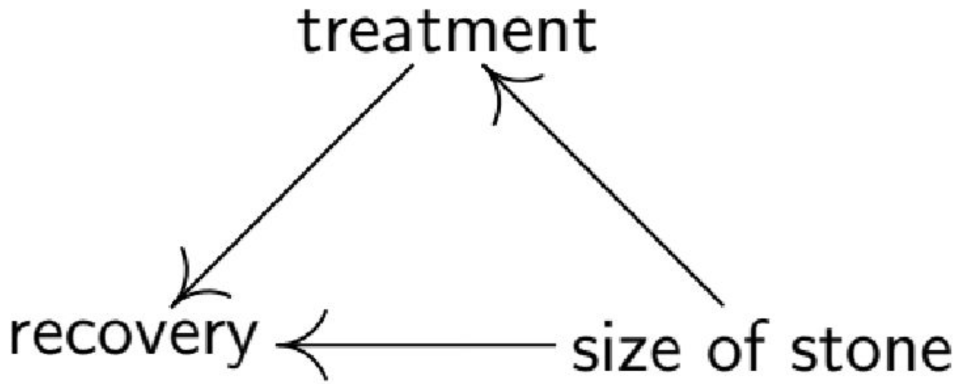
Causality - Kidney Stones Example

	Treatment A	Treatment B
Small Stones ($\frac{357}{700} = 0.51$)	$\frac{81}{87} = 0.93$	$\frac{234}{270} = 0.87$
Large Stones ($\frac{343}{700} = 0.49$)	$\frac{192}{263} = 0.73$	$\frac{55}{80} = 0.69$
	$\frac{273}{350} = 0.78$	$\frac{289}{350} = 0.83$
	$\frac{562}{700} = 0.80$	

Charig et al.: *Comparison of treatment of renal calculi by open surgery, (...)* , British Medical Journal, 1986 [2]

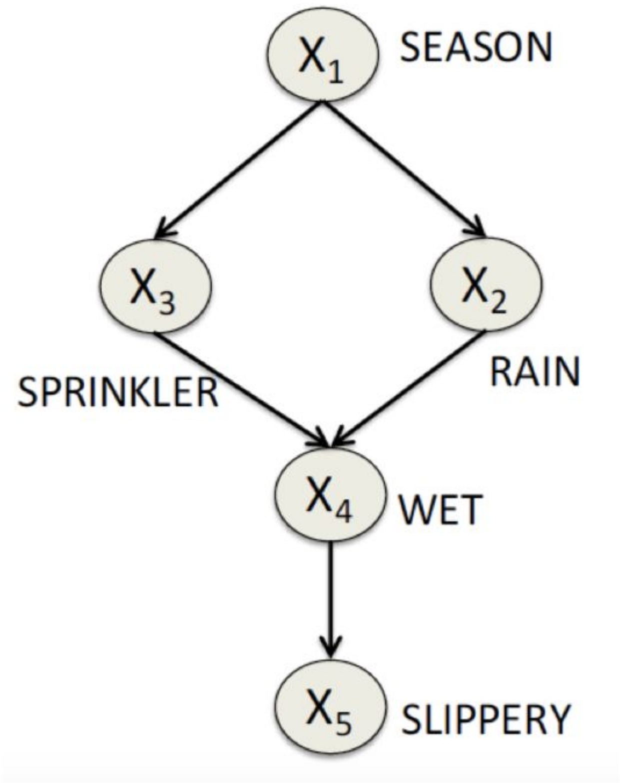
Causality - Kidney Stones Example

underlying ground truth:



[2]

Causality - Slipperiness Counterfactual Example

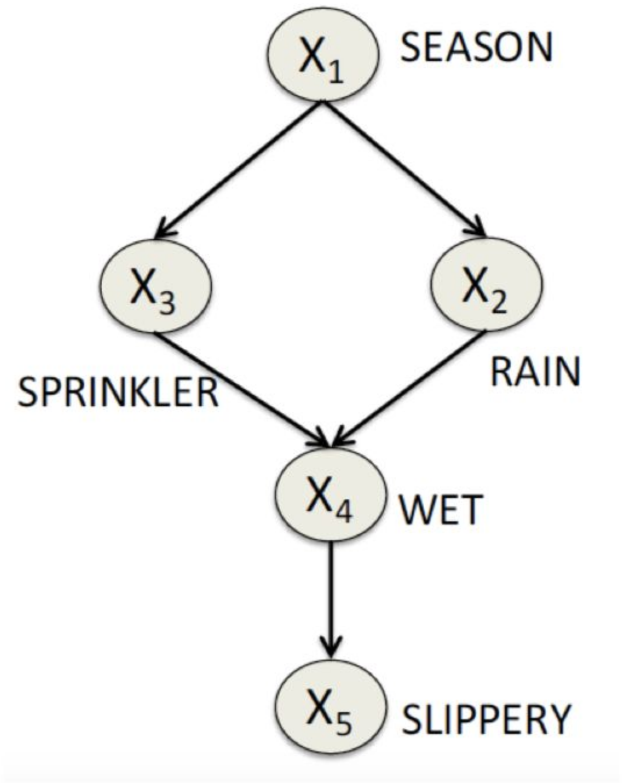


[2]

Observe that it is slippery (SL=True) and the sprinkler is on (S=ON).

Wish to access the probability that the ground would be slippery, had the sprinkler been OFF.

Causality - Slipperiness Counterfactual Example



[2]

Sprinkler=OFF should still be treated as interventional surgery, but only after we fully account for the evidence given: Slippery=True and Sprinkler=ON.

1. *Abduction*: Interpret the past in light of the evidence
2. *Action*: Bend the course of history (minimally) to account for the hypothetical Sprinkler=OFF.
3. *Prediction*: Project the consequences to the future.

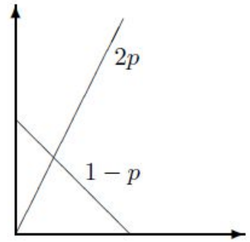
For more details check pages 1-10 from http://ftp.cs.ucla.edu/pub/stat_ser/r260-reprint.pdf

Pick-a-Hand Example

- Hider has 2 coins
  - Puts 1 in Left hand OR
  - Puts 2 in Right hand

- Chooser guesses

		hider	
		$L1$	$R2$
chooser	L	1	0
	R	0	2



Chooser:

$P(L) = 1-p$
$P(R) = p$

$E[L] = 1-p$
$E[R] = 2p$

max min $\{2p, 1-p\}$
$p = ⅓$

Hider:

$P(L) = 1 - q$
$P(R) = q$

$E[L] = 1-q$
$E[R] = 2q$

max min $\{2q, 1-q\}$
$q = ⅓$

Thus, by choosing R with probability ⅓ and L with probability ⅔ , chooser assures expected payoff of ⅔ , regardless of whether hider knows their strategy

Choose can assure expected gain of at least ⅔, hider can assure an expected loss of no more than ⅔, regardless of what either knows of the other's strategy.

Acknowledgements

Examples and images were taken from the following resources:

1. Value Iteration + Policy Iteration Resources
    - Introduction to Artificial Intelligence, CS188 course, Berkeley
    - CS 188: Artificial Intelligence Markov Decision Processes
    - CS 188: Artificial Intelligence Markov Decision Processes II
    - Deep RL Bootcamp 2017, Lecture 1, Peter Abbeel

2. Causality Resources
    - Jonas Peters Causality 4-part series
    - Probabilities Of Causation: Three Counterfactual Interpretations And Their Identification, Judea Pearl
    - Causality, Second Edition, Judea Pearl

Thanks. If you have questions:

yordan.hristov@ed.ac.uk

Bellman Equations:

$$V^*(s) = \max_a \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V^*(s')]$$

$$Q^*(s,a) = \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V^*(s')]$$

$$V^*(s) = \max_a Q^*(s,a)$$

Bellman Update:

$$V_k^*(s) = \max_a \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V_{k-1}^*(s')]$$

↓

The best we could do in state s given k timesteps to go - the utility of the action whose expectation over discounted rewards is maximised, assuming we keep acting optimally in the remaining k-1 timesteps.

$V_0^*(s)$ - the best we could get if in state s and no more timesteps to go.

$V_1^*(s)$ - if in state s and 1 timestep to go

$$\underline{\underline{V_1^*(s)}} = \max_a \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma \underline{\underline{V_0^*(s')}}]$$

We ~~calculate~~ approximate $V^*(s)$ by starting with $V_0^*(s)$ at $k=0$ and then gradually increment k until $k=H$ - a predefined horizon.

In the limit $k \to +\infty$ it can be proven that $V_k^*(s) = V_{k+1}^*(s)$

$\Rightarrow$ the utility values for all the states converge

Slide 7

$H = 100$

$\gamma = 0.9$

noise $= 0.2 \rightarrow$ Agent: $\rightarrow 0.8$, $\uparrow 0.1$, $\downarrow 0.1$ | Agent: $\downarrow 0.8$, $\leftarrow 0.1$, $\rightarrow 0.1$ | Agent: $\leftarrow 0.8$, $\uparrow 0.1$, $\downarrow 0.1$ | Agent: $\uparrow 0.8$, $\leftarrow 0.1$, $\rightarrow 0.1$

If the agent tries to go in a certain direction, there is a ~~0.8~~ 80% chance it would end up where intended and 20% chance it would go in the neighbouring ~~-of-the-intended~~ directions.

The stochasticity in the environment is modelled in the transition function $T(s,a,s')$

$\boxed{k=0} \Rightarrow$ 0 timesteps to go

$V_0^*(s) = 0$ for $\forall s$

$\boxed{k=1} \Rightarrow$ 1 time step to go $\rightarrow$ we can only exit if in the terminal states

$V_1^*(4,3) = 1$

$V_1^*(4,2) = -1$

$V_1^*(s) = 0$ for $\forall s \notin \{(4,3),(4,2)\}$

$\boxed{k=2}$

$V_2^*(4,3) = 1$

$V_2^*(4,2) = -1$

$$V_2^*(3,3) = \max_a \begin{cases} a = \text{left}: 0.8 \times 0.9 \times V_1^*(2,3) + 0.1 \times 0.9 \times V_1^*(3,3) + 0.1 \times 0.9 \times V_1^*(3,2) = 0 \\ a = \text{top}: 0.1 \times 0.9 \times V_1^*(4,3) + 0.8 \times 0.9 \times V_1^*(3,3) + 0.1 \times 0.9 \times V_1^*(2,3) = 0.09 \\ a = \text{right}: 0.1 \times 0.9 \times V_1^*(3,3) + 0.8 \times 0.9 \times V_1^*(4,3) + 0.1 \times 0.9 \times V_1^*(3,2) = 0.72 \\ a = \text{down}: 0.1 \times 0.9 \times V_1^*(4,3) + 0.8 \times 0.9 \times V_1^*(3,2) + 0.1 \times 0.9 \times V_1^*(2,3) = 0.09 \end{cases}$$

$= 0.72$ for $a = \text{right}$

$$V_2^*(3,2) = \max_a \begin{Bmatrix} a = \text{left}: 0 \\ a = \text{top}: -0.09 \\ a = \text{right}: -0.72 \\ a = \text{down}: -0.09 \end{Bmatrix} = 0 \text{ for } a = \text{left}$$

$$V_2^*(4,1) = \max_a \begin{Bmatrix} a = \text{left}: -0.09 \\ a = \text{top}: -0.72 \\ a = \text{right}: -0.09 \\ a = \text{down}: 0 \end{Bmatrix} = 0 \text{ for } a = \text{down}$$

$V_2^*(s) = 0$ for all other states

If we have 2 timesteps to go, the agent can't reach the +1 terminal state $\Rightarrow$ it would try to avoid any chance of ending up in the −1 terminal state [for states (3,2) and (4,1)]. Therefore it chooses to bump into the wall.

$\boxed{K = 3}$

$V_3^*(4,3) = 1$

$V_3^*(4,2) = -1$

$$V_3^*(3,3) = \max_a \begin{Bmatrix} a = \text{left}: 0.072 \\ a = \text{top}: 0.51 \\ a = \text{right}: 0.78 \\ a = \text{down}: 0.072 \end{Bmatrix} = 0.78 \text{ for } a = \text{right}$$

$$V_3^*(3,2) = \max_a \begin{cases} a = left: & 0.07 \\ a = top: & 0.43 \\ a = right: & -0.65 \\ a = down & -0.09 \end{cases} = 0.43 \text{ for } a = top$$

$$V_3^*(2,3) = \max_a \begin{cases} a = left: & 0 \\ a = top: & 0.07 \\ a = right: & 0.52 \\ a = down: & 0.07 \end{cases} = 0.52 \text{ for } a = right$$

$$V_3^*(4,1) = \max_a \begin{cases} a = left: & -0.08 \\ a = top: & -0.72 \\ a = right: & -0.08 \\ a = down: & 0 \end{cases} = 0 \text{ for } a = down$$

$V_3^*(s) = 0$ for all other states

Given 3 timesteps it is worth to risk going to (4,3) from (3,2) $\Rightarrow$ change in policy. However, it is not worth it from (4,1)

Continue untill $k = H$ ...

DMR Worked Examples

Slide 16
Slide 17

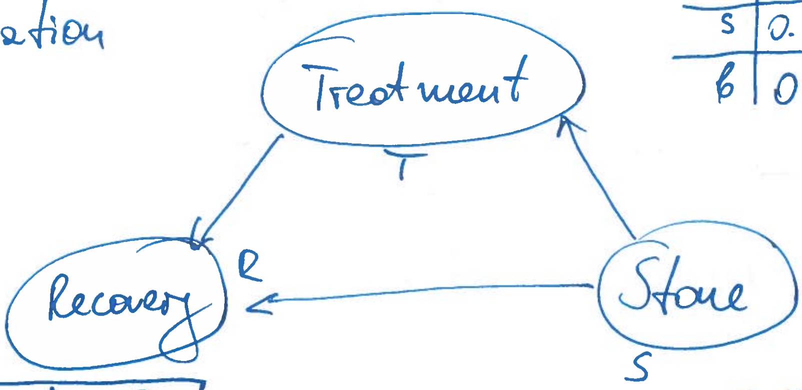
II. Causality - Kidney Stones Example

	T = A	T = B
Small Stones (0.51)	81/87 = 0.93	234/270 = 0.87
Big Stones (0.49)	192/263 = 0.73	55/80 = 0.69
	273/350 = 0.78	289/350 = 0.83

Structural Equations:

1) $S \sim P(s)$
2) $T = F_1(S)$
3) $R = F_2(T, S)$

1) Observation



S	T = A	T = B
s	0.24	0.76
b	0.76	0.24

S = s	S = b
0.51	0.49

T	S	R = 1	R = 0
A	s	0.93	0.07
A	b	0.73	0.27
B	s	0.87	0.13
B	b	0.69	0.31

Joint Probability Factorisation

$$P(S,T,R) = P(S)P(T|S)P(R|T,S)$$

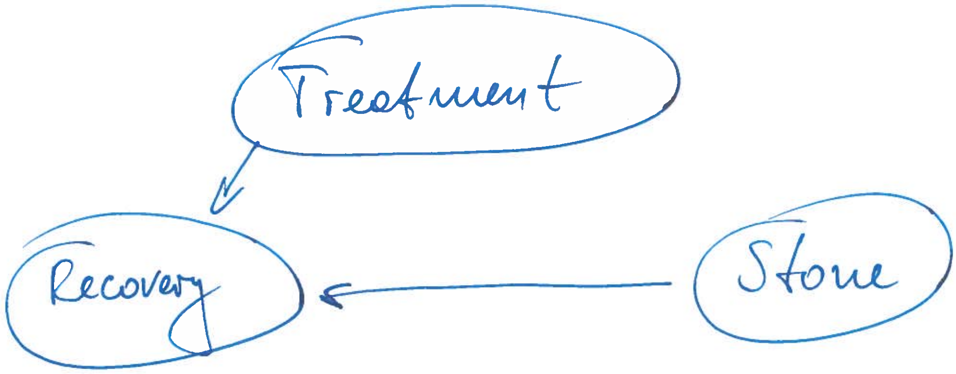
$$P(R=1|T=A) = \frac{P(R=1,T=A)}{P(T=A)} = \alpha \sum_s P(R=1,T=A,s) = \alpha \sum_s P(s)P(T=A|s)P(R=1|T=A,s)$$

$$= \alpha[0.93 \times 0.24 \times 0.51 + 0.73 \times 0.76 \times 0.49] = \alpha[0.38]$$

$$P(R=0|T=A) = \alpha[.10] \Longrightarrow P(R=1|T=A) = \underline{\underline{0.78}}$$

$$\left.\begin{aligned} P(R=1|T=B) &= \beta[0.41] \\ P(R=0|T=B) &= \beta[0.08] \end{aligned}\right\} \Longrightarrow P(R=1|T=B) = \underline{\underline{0.83}}$$

2) Intervention



Structural Equations:

1) $S \sim P(s)$
2) $T = A$ ← Intervene only on 2)
3) $R = F_2(T, S)$

<u>Joint Probability Factorisation</u>

$$P_{do(T=A)}(S,R,T) = P_{do(T=A)}(S)\,P_{do(T=A)}(R \mid S, T=A)\,\underbrace{P(T=A)}_{\text{deterministic / set} \Rightarrow 1}$$

$$P_{do(T=A)}(R=1) = \sum_s P_{do(T=A)}(R=1, S=s, T=A) =$$

$$= \sum_s P_{do(T=A)}(R=1 \mid T=A, S=s)\, P_{do(T=A)}(S=s)$$

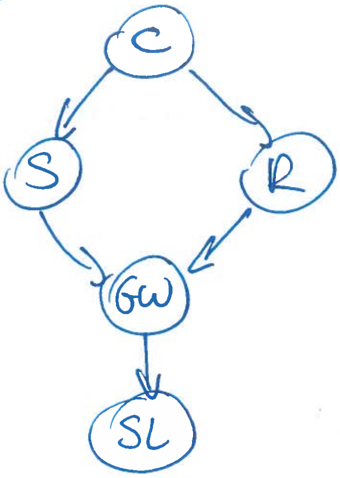
$$= \sum_s P(R=1 \mid T=A, S=s)\, P(S=s) =$$

$$= 0.51 \times 0.93 + 0.49 \times 0.73 = \underline{\underline{0.83}}$$

$$P_{do(T=B)}(R=1) = \sum_s P(R=1 \mid T=B, S=s)\,P(S=s) =$$

$$= 0.51 \times 0.87 + 0.49 \times 0.69 = \underline{\underline{0.78}}$$

)Counterfactual



$$P(c)\begin{cases} P(C=0)=0.5 \\ P(C=1)=0.5 \end{cases}$$

Observe $SL = \text{True}$ and $S = ON$
Wish to access the probability that the ground would be slippery had the sprinkler been OFF! $P_{do(S=OFF)}(SL=1 \mid SL=1, S=ON)$

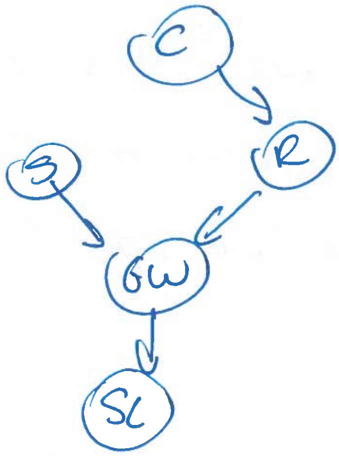
tructural Equations

1) $C \sim P(c)$
2) $S = \bar{C}$
3) $R = C$
4) $GW = R \vee S$
5) $SL = GW$

* abduction - update $P(c)$ to $P(c \mid SL=1, S=ON)$

$$P(c \mid e) = \begin{cases} 1 & \text{if } c=0 \\ 0 & \text{if } c=1 \end{cases}$$

* action $do(S=OFF)$



Revised
Structura Equations

1) $C \sim P(c \mid SL=1, S=ON)$
2) $S = OFF$
3) $R = C$
4) $GW = R \vee S$
5) $SL = GW$

$$P(c \mid SL=\text{True}, S=ON) = \begin{cases} P(c=1 \mid SL=1, S=ON) = 0 \\ P(c=0 \mid SL=1, S=ON) = 1 \end{cases}$$

if we see it is slippery and the sprinkler was ON it must have been not cloudy

Sum over all states of the exogeneous variables - c - that are compatible with the information at hand, the evidence.

$$P_{do(S=OFF)}(SL=True \mid SL=True, S=ON) =$$

$$= \sum_c P_{do(S=OFF)}(SL=True \mid c)\, \overbrace{P_{do(S=OFF)}(c \mid SL=True, S=ON)}^{\text{we have that on previous page}}$$

$$= 1 \times 0 \quad + \quad 0 \times 1 \quad = \quad 0$$

the chance it would be slippery if cloudy and sprinkler is OFF

the chance it would have been cloudy if sprinkler ~~is~~ was observed ON

the chance it would be slippery if not cloudy and ~~was~~ sprinkler OFF

chance it would have not been cloudy if sprinkler ~~is~~ was observed ON

If we observe that the sprinkler is ON & the floor is slippery, then the chance it would have been slippery, if the sprinkler was OFF, is $\underline{0}$