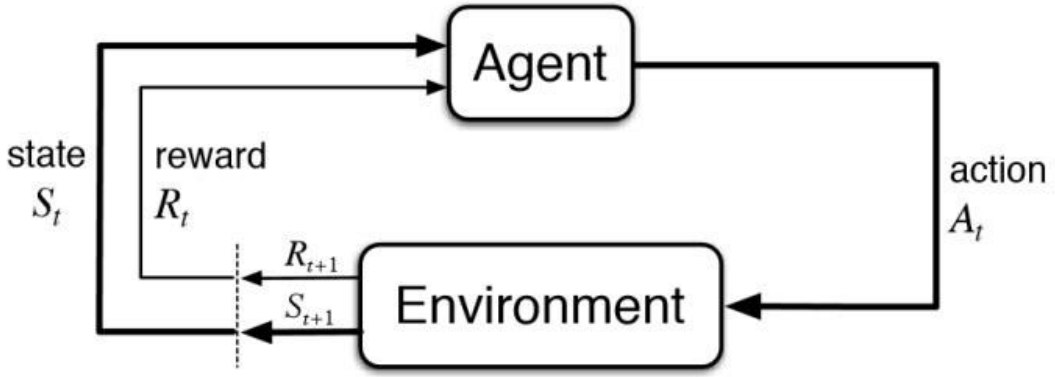
Clinical Reinforcement Learning

Aniruddh Raghu

Outline

Part 1: Reinforcement Learning

- Goals
- Definitions
- Methods



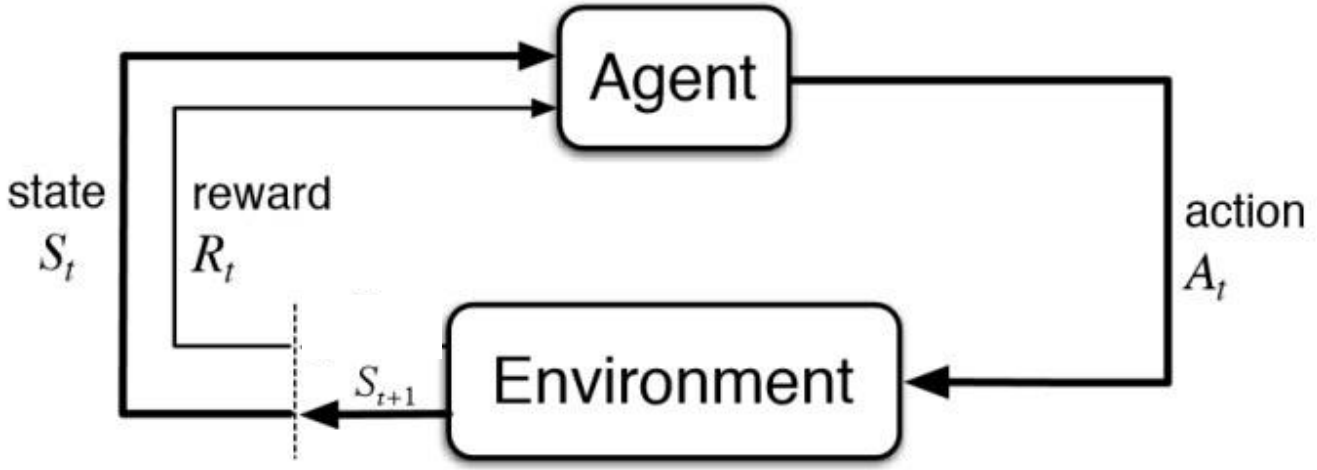
Part 2: Clinical applications

- Challenges
- Case study: sepsis treatment

Part 1: Reinforcement Learning (RL)

RL: Fundamentals

Goal: learn to make good decisions in an uncertain environment to maximise accumulated reward

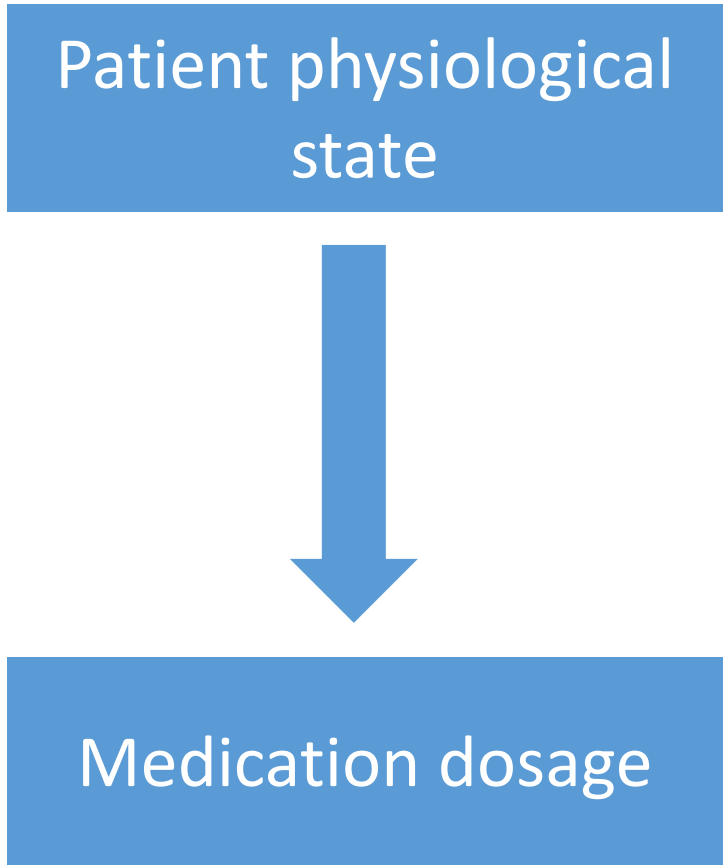


RL: Examples

Game playing:

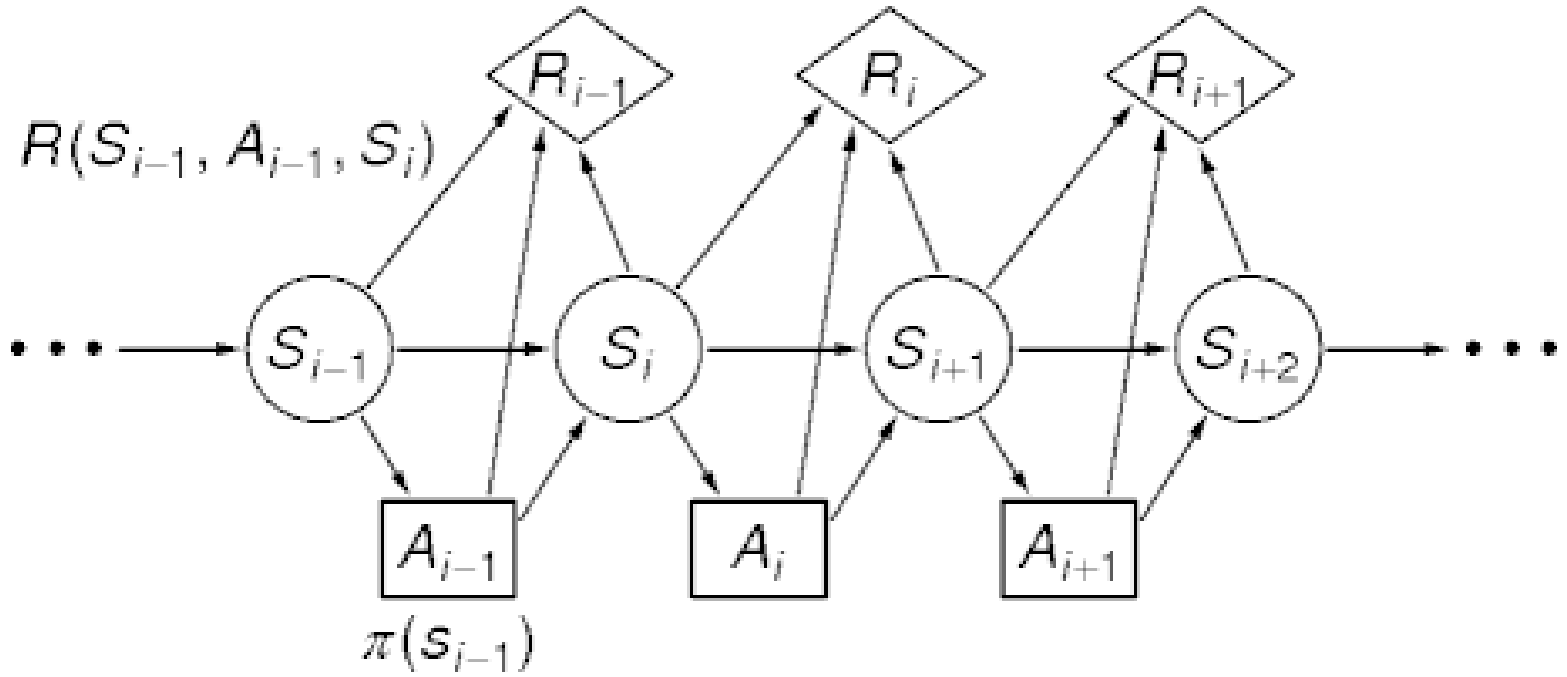
https://www.cs.toronto.edu/~vmnih/docs/dqn.pdf

Medical decision support systems:



RL Problem Formulation: Markov Decision Processes (MDPs)

$$\text{MDP} : (\mathcal{S}, \mathcal{A}, P_{s_0}(\cdot), P(\cdot|s,a), R(s,a,s'), \gamma)$$



Some definitions

Policy: $\pi(a|s)$

Return: $R_{T-1} = \sum_{t=0}^{T-1} \gamma^t r_t$

Goal of agent: Find an optimal policy:

$$\pi^* = \arg\max_\pi \mathbb{E}_\pi \left[R_{T-1} \right]$$

Action-Value function (or Q-function)

$$Q^{\pi}(s,a) = \mathbb{E}_{\mathrm{MDP},\pi}\left[R_{T-1}|s_t = s, a_t = a, \pi\right]$$

Recursive definition:

$$Q^{\pi}(s,a) = \mathbb{E}_{\mathrm{MDP}}\left[r(s,a,s') + \gamma \mathbb{E}_{\pi}\left[Q^{\pi}(s',a')\right]\right]$$

$$Q^{\pi}(s,a) = \sum_{s'} P(s'|s,a)\left[r(s,a,s') + \gamma \sum_{a'} \pi(a'|s') Q^{\pi}(s',a')\right]$$

Optimal action value function

$$\pi^* = \arg\max_\pi \mathbb{E}_\pi \left[R_{T-1}\right]$$

$$Q^*(s,a) = \mathbb{E}_{\text{MDP}} \left[r(s,a,s') + \gamma \max_{a'} Q^*(s',a')\right]$$

$$Q^*(s,a) = \sum_{s'} P(s'|s,a) \left[r(s,a,s') + \gamma \max_{a'} Q^*(s',a')\right]$$

$$\pi^*(a|s) = \arg\max_a Q^*(s,a)$$

Methods

Key question: How to learn optimal policy?

- Q-value iteration
- Q-learning
- Fitted Q-Iteration (FQI)
- Deep Q-learning

Q-value iteration: Algorithm

Discrete state space and discrete action space

Keep a table of Q(s,a), for every (s,a) pair; initialise to zero

Requires P(s’| s, a), R(s, a, s’)

Iterate:

$$Q^{(k+1)}(s,a) \leftarrow \sum_{s'} P(s'|s,a) \left[r(s,a,s') + \gamma \max_{a'} Q^{(k)}(s',a') \right]$$

Q-value iteration: Challenges

Requires discrete state and action space

Requires known $P(s' \mid s, a)$

Requires reward function to be specified

Q-learning: Algorithm

Discrete state and action spaces

Keep a table of Q(s,a) for every (s,a) pair; initialise to zero

Don’t require P(s’| s, a): use data, tuples of (s, a, r, s’) instead

Iterate:

Observe transition (s, a, r, s’)

$$Q(s,a) \leftarrow Q(s,a) + \alpha\left[r + \gamma \max_{a'} Q(s',a') - Q(s,a)\right]$$

NOTE: $\text{TD-error} = r + \gamma \max_{a'} Q(s',a') - Q(s,a)$

Q-learning: Challenges

Discrete state and action spaces are strong assumptions

Data inefficient

Q-learning/Q-value iteration: Visually

Initialisation

Action space

State space

	Up	Down	Left	Right
0	0	0	0	0
1	0	0	0	0
2	0	0	0	0
3	0	0	0	0
4	0	0	0	0
5	0	0	0	0
6	0	0	0	0

After a few updates

	Up	Down	Left	Right
0	0	0.25	0	0
1	0	0	0.5	0
2	0	0.1	0	0
3	0.5	0	0	0.5
4	0	1	0	0
5	0	0	1	0
6	0	1	0	0

At convergence

	Up	Down	Left	Right
0	0	0.5	0	0
1	0	0	1	0
2	0	0.8	0	0
3	1	0	0	0.75
4	0	1.5	0	0
5	0	0	2	1
6	0	3	0	0

Fitted Q-Iteration (FQI)

Continuous state space

Parameterise $Q^{(k)}(s, a; \theta^{(k)})$ with parameter vector $\theta^{(k)}$; $Q^{(0)}(s, a; \theta^{(0)}) = 0$

Uses dataset of N (s, a, r, s') tuples

Iterate:

TD ERROR

$$\theta^{(k)} = \arg\min_\theta \sum_{i=1}^{N} \left(r^{(i)} + \gamma \max_{a'} Q^{(k-1)}(s'^{(i)}, a'; \theta^{(k-1)}) - Q^{(k)}(s^{(i)}, a^{(i)}; \theta) \right)^2$$

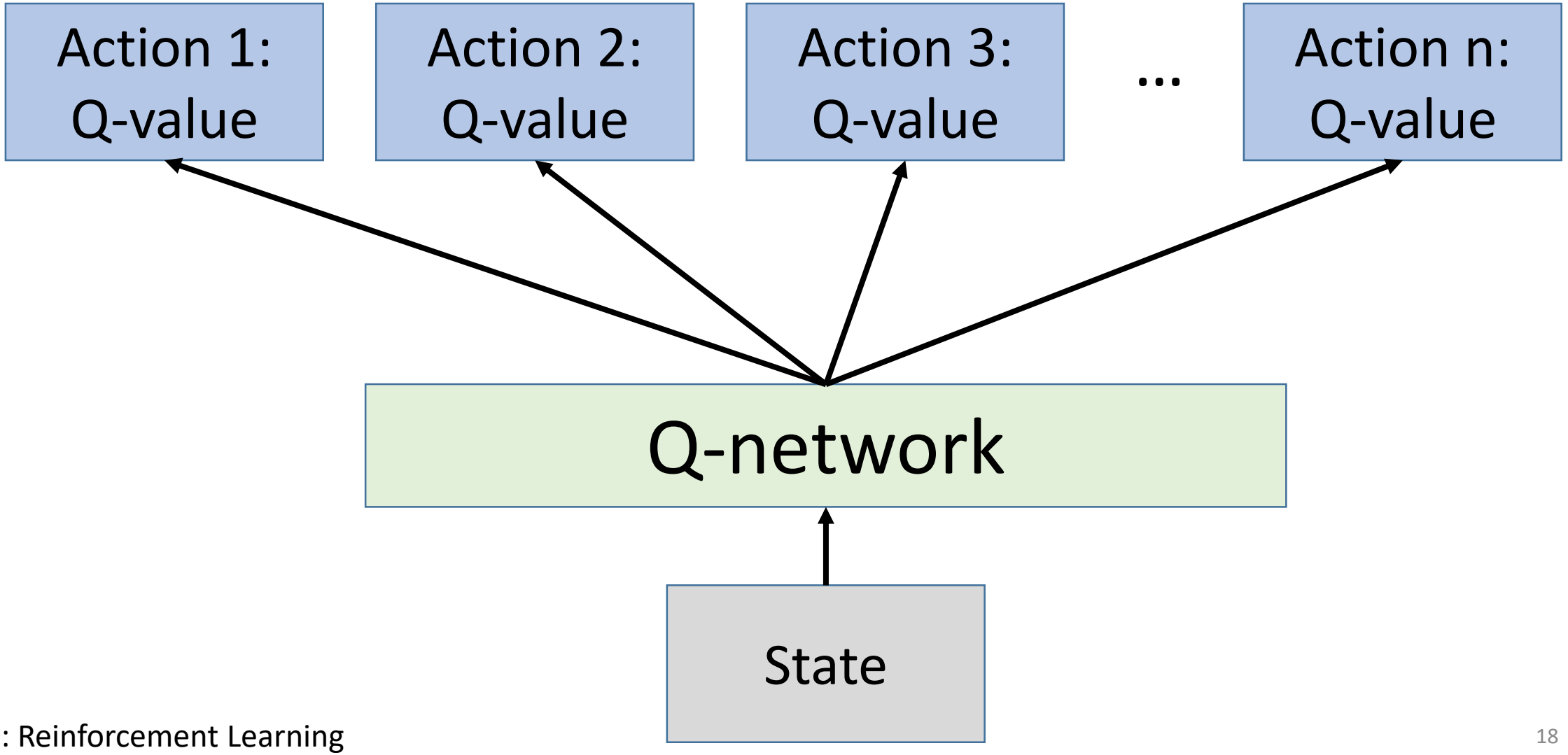
Deep Q learning

In healthcare, similar to FQI, except:

- Q-function represented with neural network – a Deep Q Network (DQN)
- Main and target networks, with different parameters, to define loss function

Various tricks to improve learning speed and learning stability

Deep Q learning visualised (discrete action space)



Summary of algorithms

Algorithm	State space	Action space	Transition?	Reward?	Data?
Q –value iteration	Discrete	Discrete	Yes	Yes	No
Q-learning	Discrete	Discrete	No	No	Yes
Fitted Q-Iteration	Continuous	Either	No	No	Yes
Deep Q learning	Continuous	Either	No	No	Yes

Part 2: Clinical applications

RL in medicine: Themes

Lack of simulator

Small datasets

Unobserved, confounding variables

RL in medicine: Challenges

1. Defining state space, action space, reward function

2. Computing **safe** policy in data-efficient manner

3. Evaluating policy

Case study: Treatment of Sepsis in ICU

Clinical motivation

Sepsis: severe infection, typically involving organ dysfunction

Leading cause of mortality

Expensive to treat

Suboptimal medical treatment

Dataset

Cohort: sepsis patients from ICU

Patient ID	Timestep	Demographics	Vitals	Lab values	Treatments	Outcome
00001	1	Age, Gender, …	HR, BP, …	Albumin, …	IV, vasopressor, …	Survived
00001	2	…	…	…	…	…

State space

Patient ID	Timestep	Demographics	Vitals	Lab values	Treatments	Outcome
00001	1	Age, Gender, …	HR, BP, …	Albumin, …	IV, vasopressor, …	Survived
00001	2	…	…	…	…	…

Discrete [1], [2]

~ 1000 Cluster centroids from k-means

Continuous vector from raw ICU data [3], [4]

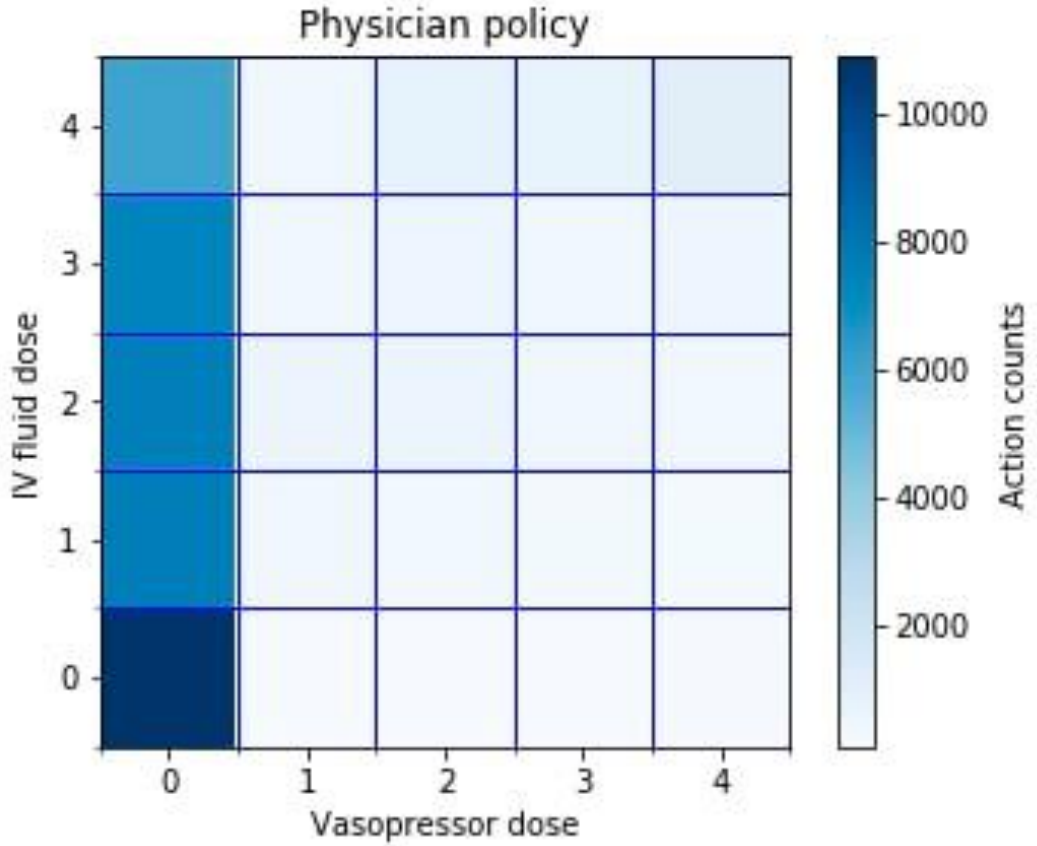
Continuous state using recurrent autoencoder embedding [5]

Capture historical information in latent representation of autoencoder

Action space

Patient ID	Timestep	Demographics	Vitals	Lab values	Treatments	Outcome
00001	1	Age, Gender, …	HR, BP, …	Albumin, …	IV, vasopressor, …	Survived
00001	2	…	…	…	…	…

Discretised over dosage amounts of IV fluids and vasopressors



Reward function

Patient ID	Timestep	Demographics	Vitals	Lab values	Treatments	Outcome
00001	1	Age, Gender, …	HR, BP, …	Albumin, …	IV, vasopressor …	Survived
00001	2	…	…	…	…	…

Mortality event [1], [2], [3]

- Can add intermediate signal [4]
- *Issue: sparse, hard to learn from*

Log odds of mortality [5]

- Regressor: predict mortality probability from (s, a, s')
- Reward: change in mortality probability
- *Issue: dependent on quality of reward predictor*

Methods

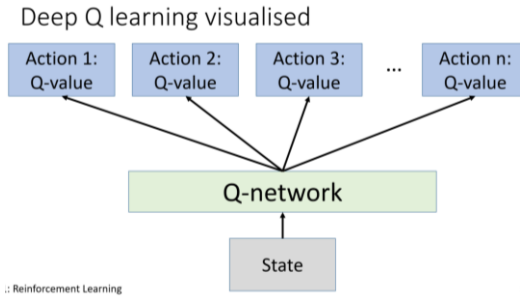
Q-value iteration [1], [2] :

Discrete state space,
estimate P(s’|s, a) and R(s, a, s’) from data

	Up	Down	Left	Right
0	*0*	*0.5*	*0*	*0*
1	*0*	*0*	*1*	*0*
2	*0*	*0.8*	*0*	*0*
3	*1*	*0*	*0*	*0.75*

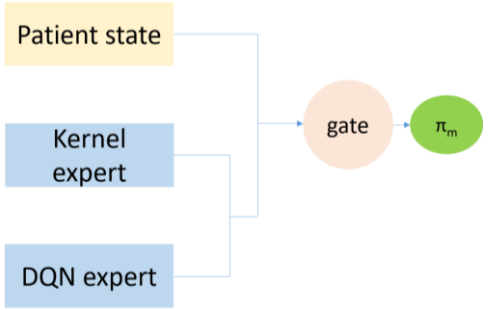
Deep Q learning [3], [4] :

Standard Q network, with some additional tricks



Mixture of Experts: Deep Q Learning & Kernel-based expert [5] :

Safe policy discovery



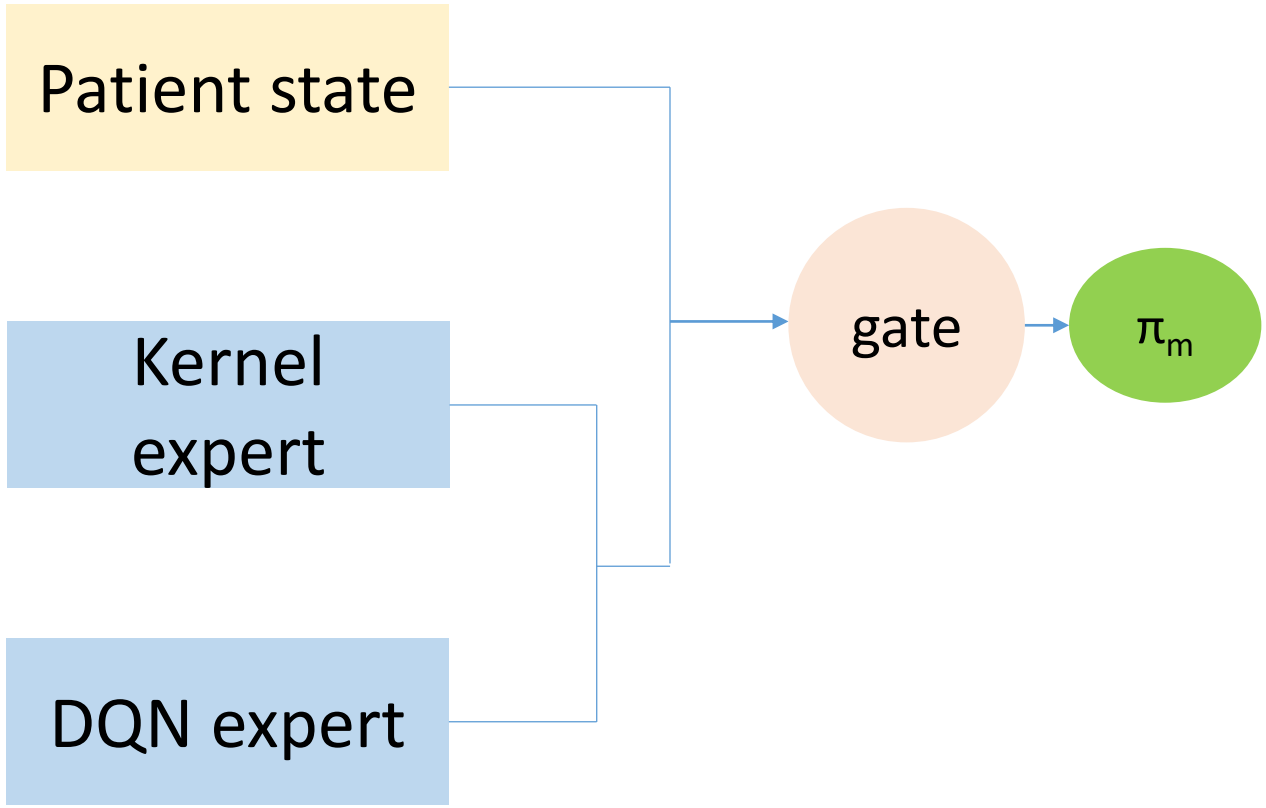
Mixture of Experts

DQN expert: Deep Q learning, restricted to actions seen in neighbour states

Kernel expert: based on clinical actions in neighbour states

Learn gating function to choose between them

Safe policy – will not recommend unseen actions

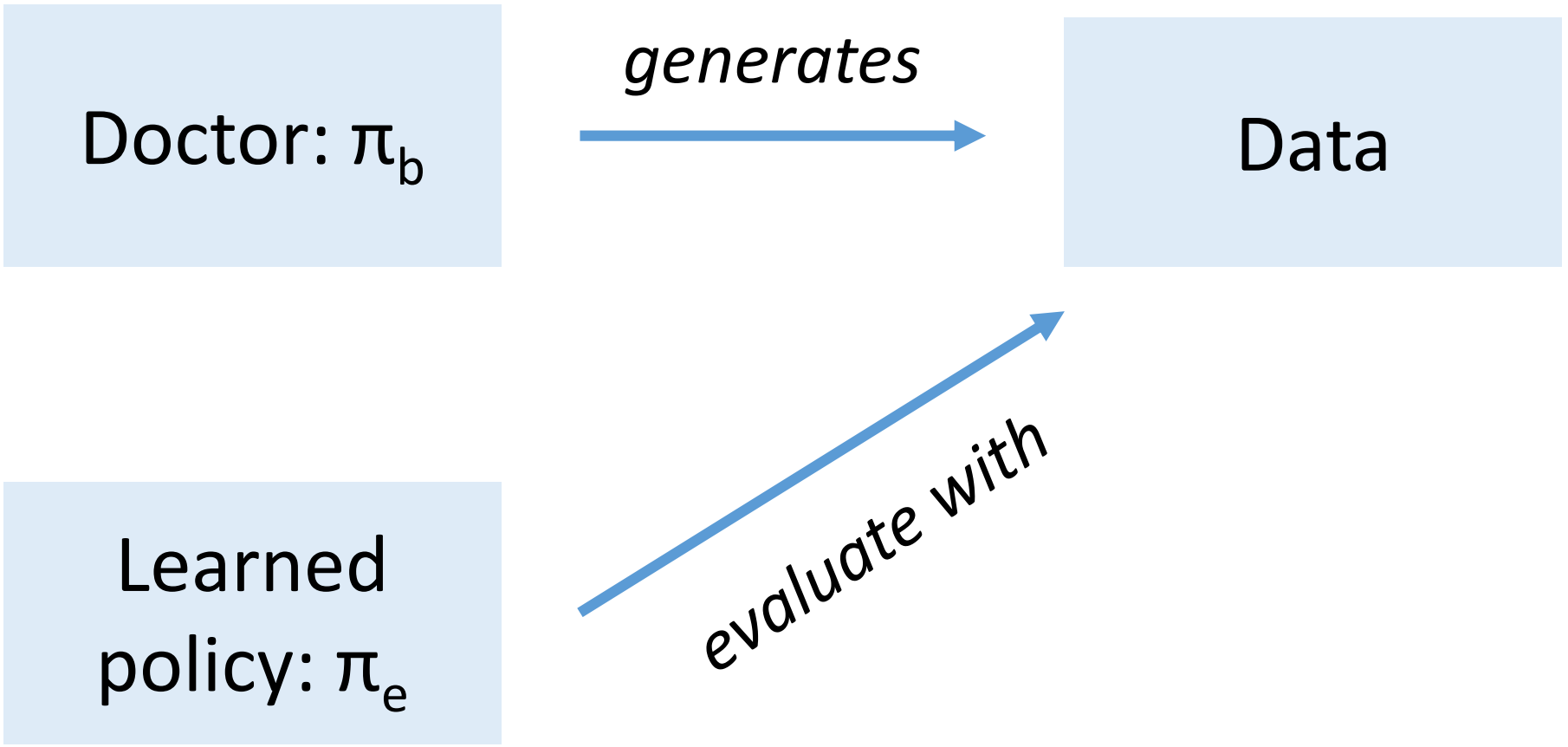


Off-Policy Evaluation

Objective: find quality of learned policy π_e:

$$V^{\pi_e} = \mathbb{E}_{P(\cdot|s,a),P_{s_0},\pi_e}\left[R_{T-1}\right]$$

Estimate given data from π_b – clinical policy

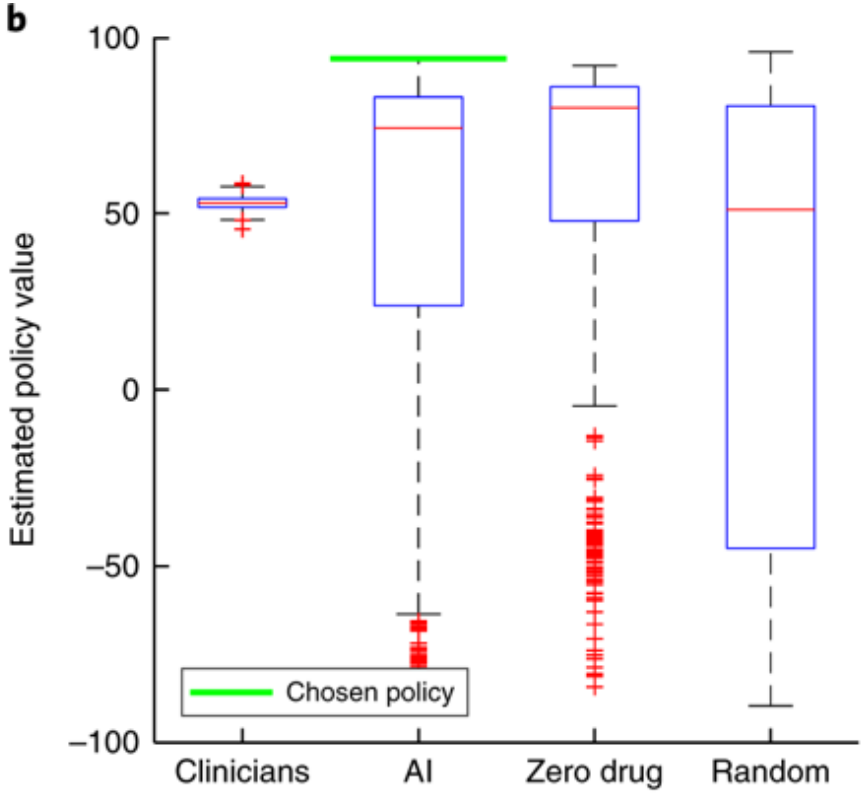

Doctor: π_b
generates
Data
Learned policy: π_e
evaluate with

Off-Policy Evaluation – challenges (from [6])

Importance sampling: high variance, crucial to estimate π_b accurately

Model-based estimation: unknown bias

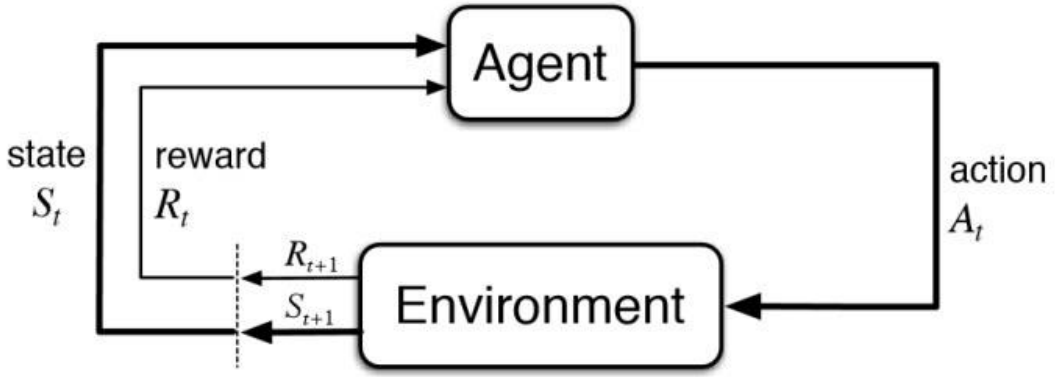
From [5]:



Summary

Part 1: Reinforcement Learning

- Goals
- Definitions
- Methods



Part 2: Clinical applications

- Challenges
- Case study: sepsis treatment

Areas to consider

Best state representation?

Can we formulate better rewards?

How do we better utilise clinical policy in our treatment strategies?

Different methods for evaluation with lower variance?

What do doctors want to see with RL in medicine?

References

[1] Komorowski, Matthieu, et al. "A Markov Decision Process to suggest optimal treatment of severe infections in intensive care." *Neural Information Processing Systems Workshop on Machine Learning for Health*. 2016.

[2] Komorowski, Matthieu, et al. "The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care." *Nature Medicine* 24.11 (2018): 1716.

[3] Raghu, Aniruddh, et al. "Continuous state-space models for optimal sepsis treatment-a deep reinforcement learning approach." *arXiv preprint arXiv:1705.08422* (2017).

[4] Raghu, Aniruddh, et al. "Deep reinforcement learning for sepsis treatment." *arXiv preprint arXiv:1711.09602* (2017).

[5] Peng, Xuefeng, et al. "Improving Sepsis Treatment Strategies by Combining Deep and Kernel-Based Reinforcement Learning." *arXiv preprint arXiv:1901.04670* (2019).

[6] Gottesman, Omer, et al. "Evaluating reinforcement learning algorithms in observational health settings." *arXiv preprint arXiv:1805.12298* (2018).