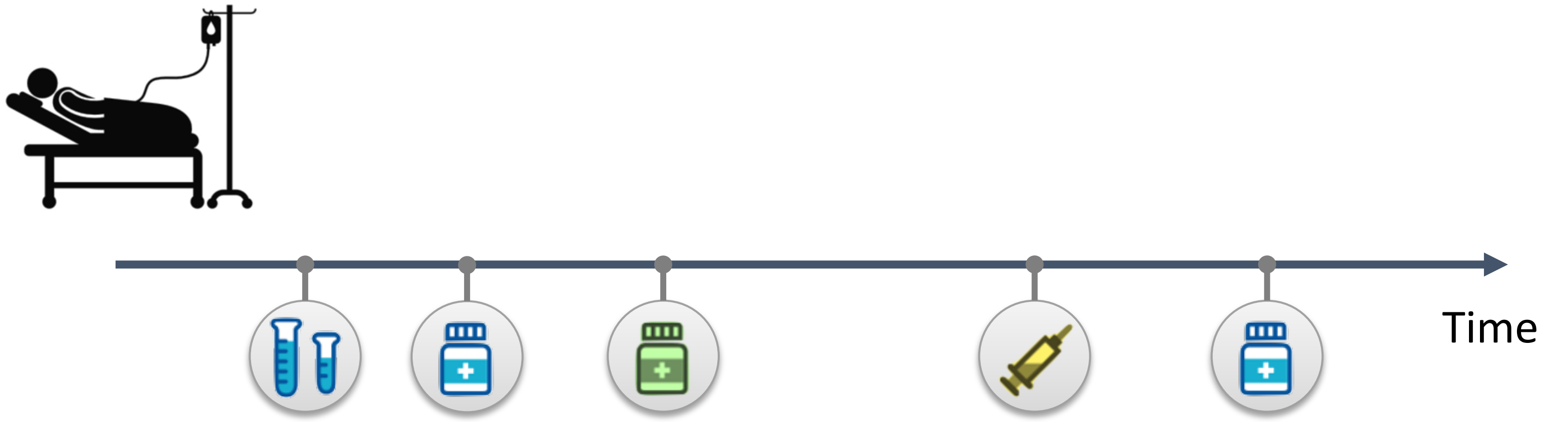
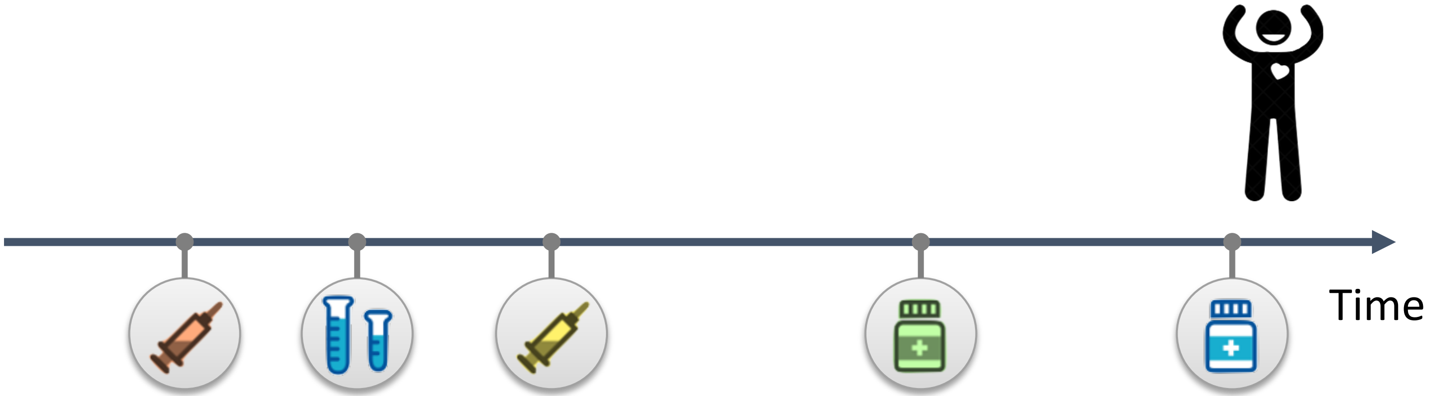
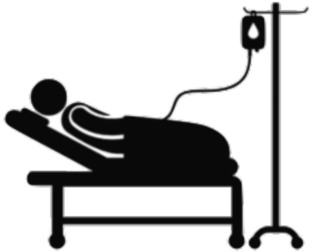


Clinician-in-the-Loop Decision-Making:
Reinforcement Learning with Near-Optimal Set-Valued Policies

Shengpu Tang et al., ICML 2020.

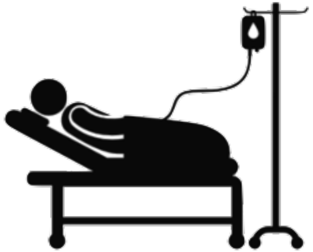


Reinforcement Learning (RL)



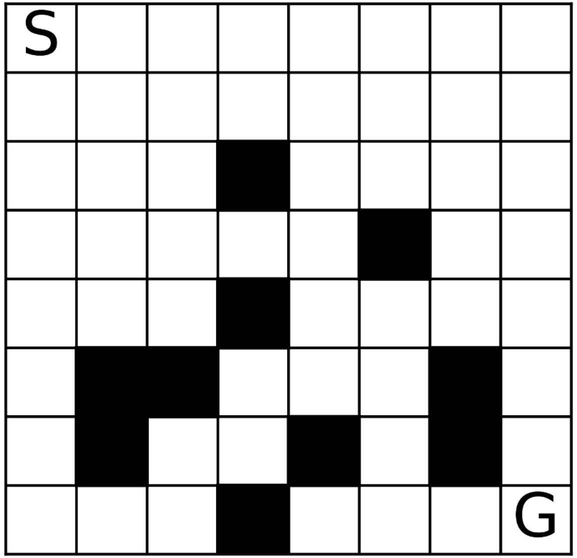
RL agent

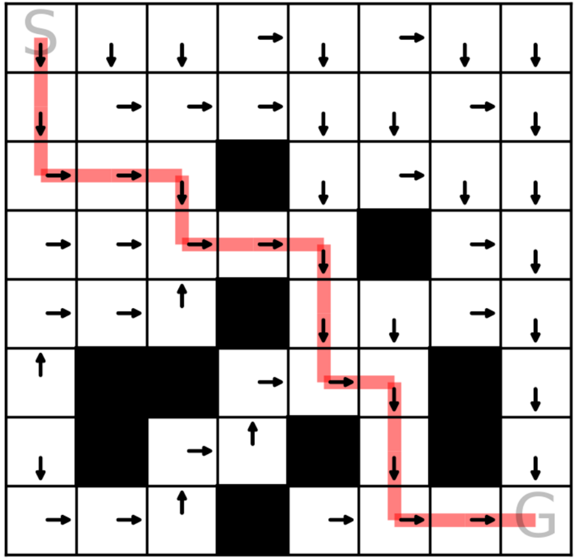
Policy

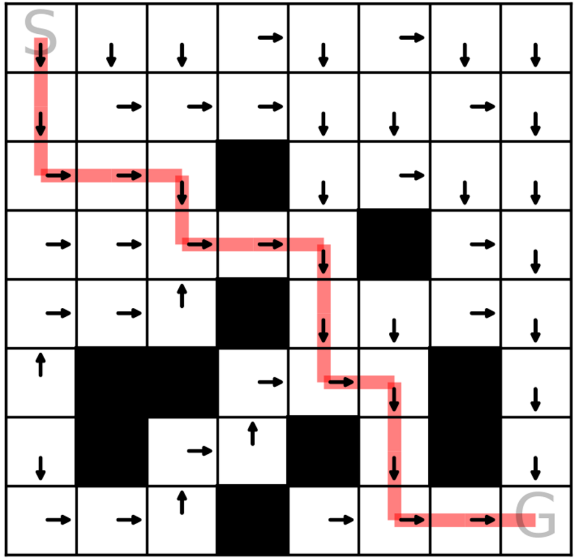


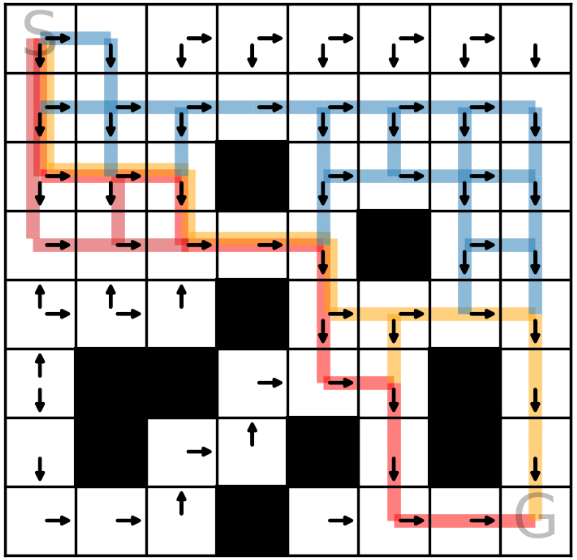
near-equivalent
actions

Set-Valued Policy


S
G


S
G


S
G


S
G

Temporal Difference learning

$$Q(s,a) \leftarrow Q(s,a) + \alpha[Q_{\text{target}}(s,a) - Q(s,a)]$$

Near-greedy action selection

$$\pi(s) = \{a : Q^{\pi}(s,a) \geq (1-\zeta)V^{*}(s)\}$$

ζ
sub-optimality margin

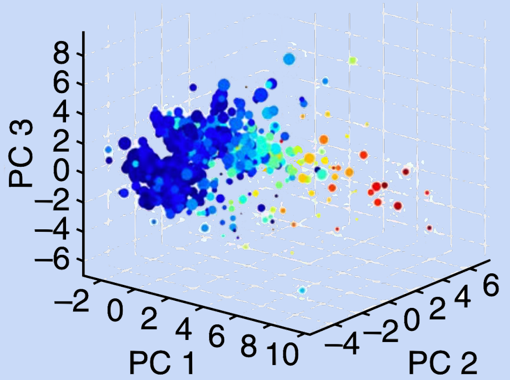
Proposed algorithm
near-greedy TD learning

π
near-optimal set-valued policy

Clinical Task

Sepsis Treatment in ICUs

Patient health state

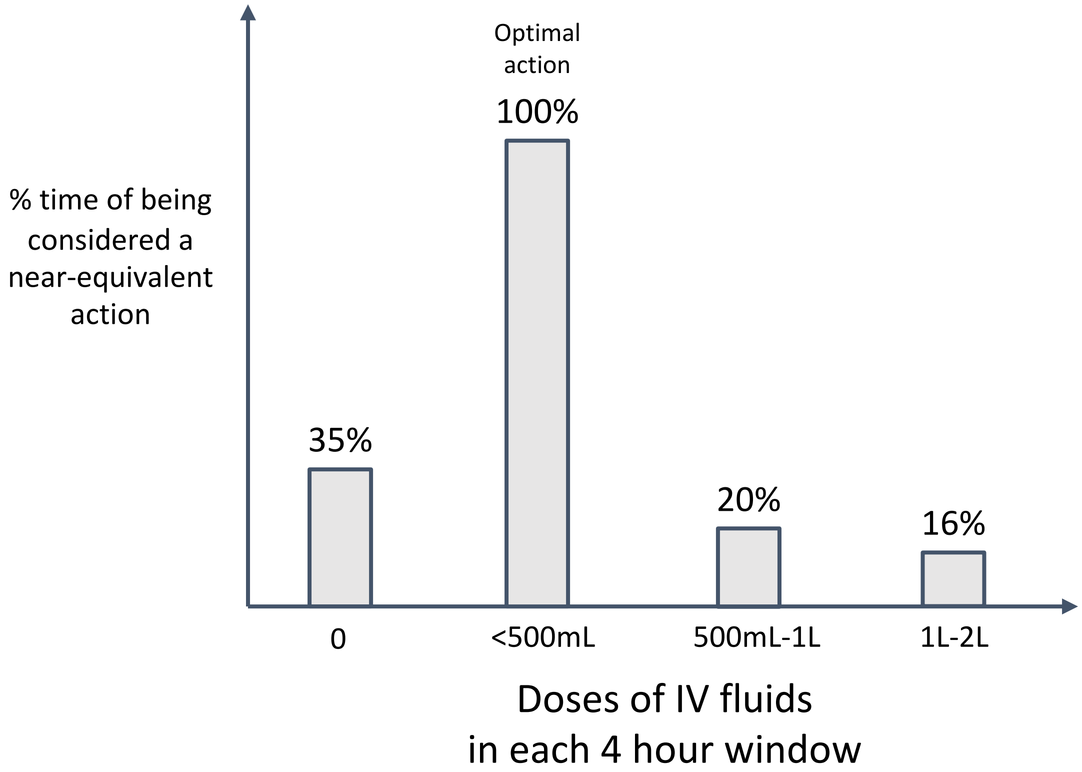


Treatment actions

		Dose of vasopressor				
		1	2	3	4	5
Dose of i.v. fluid	1	1	2	3	4	5
	2	6	7	8	9	10
	3	11	12	13	14	15
	4	16	17	18	19	20
	5	21	22	23	24	25

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

Similar IV fluid doses are near-equivalent when **no vasopressors** are used.



We propose a **new algorithm** for learning **near-optimal set-valued policies**, which can provide action choices while maintaining near-optimality

- An important step for clinician/human-in-the-loop decision support
- Humans incorporate additional knowledge to select among near-equivalent actions
- Potential broader impact to other applications beyond healthcare

MICHIGAN MEDICINE
UNIVERSITY OF MICHIGAN

COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF MICHIGAN


S. Tang


A. Modi


M.W. Sjoding


J. Wiens

https://gitlab.eecs.umich.edu/MLD3/RL-Set-Valued-Policy

This work was supported by the National Library of Medicine (NLM grant no. R01LM013325).

More Details

Clinician-in-the-Loop Decision Making:
RL with Near-Optimal Set-Valued Policies

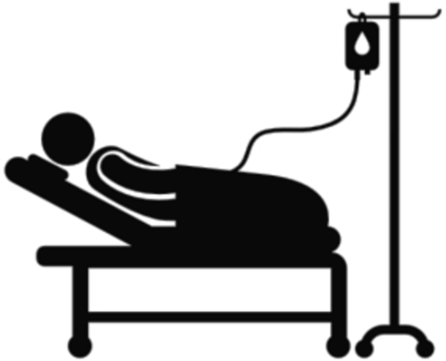
Presented by: Shengpu Tang

Co-authors: Aditya Modi; Michael W. Sjoding, MD; Jenna Wiens, PhD

ICML, July 2020

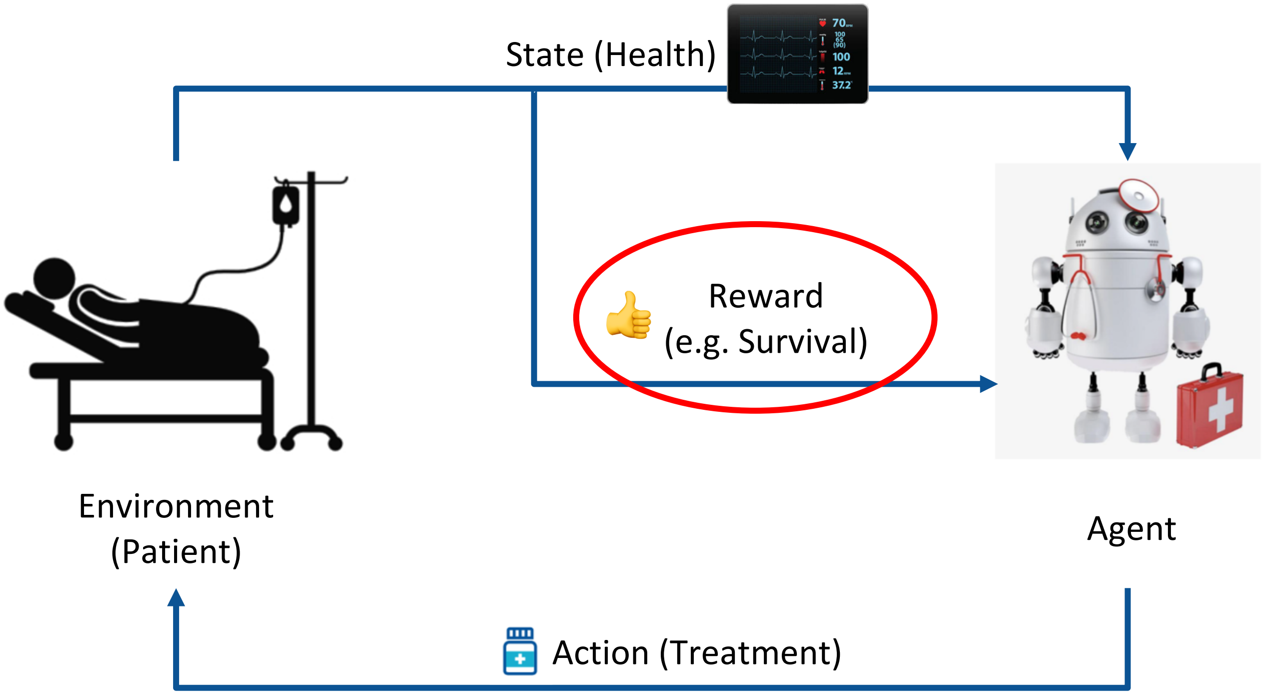
Decision-Making in Healthcare

Matching patients to *the right* ***treatment*** at *the right* ***time***.



Data-driven methods

Reinforcement Learning for Healthcare



Motivation: Near-equivalent actions

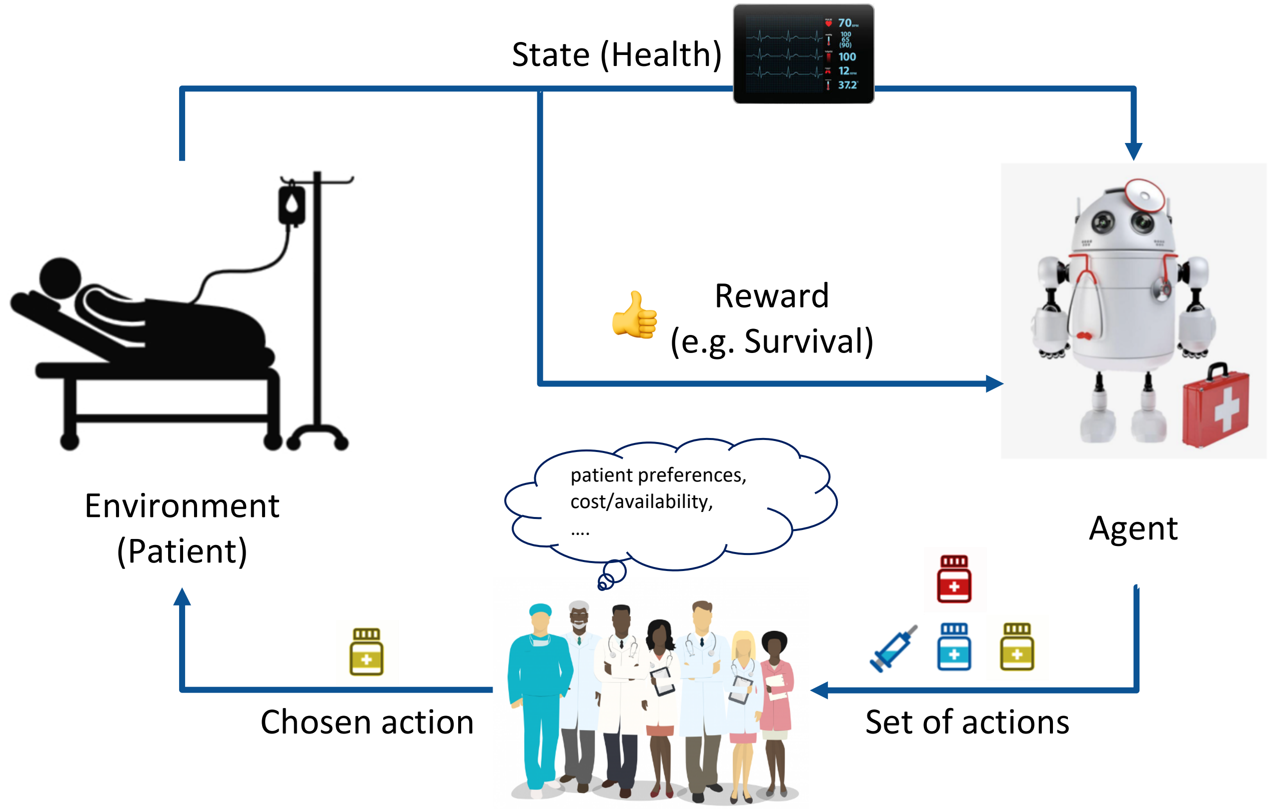
Action 1 → 90.0% survival **+ expensive**

Action 2 → 90.1% survival **+ side effects**

Action 3 → 89.1% survival **+ invasive**

- Many actions could be ***near-equivalent*** with respect to survival **but differ otherwise**
- Challenging to quantify a **single reward** that **captures different goals** for different individuals
- Impractical to incorporate all aspects **at training time**

Our Goal: Learn a mapping from each state to a **set** of **near-equivalent** actions.



Why is this challenging?

The **sequential** nature of decisions makes learning such policies non-trivial

Learning agent should consider actions as ***near-equivalent*** only if these actions are

- both **similar in the short term** (instantaneous reward)
- and **similar for any possible future** trajectory (expected cumulative returns)

Previous work on learning set-valued policies

Fard & Pineau (2011) proposed a **model-based solution** for finite-horizon planning, formulated as a **mixed-integer program**

- requires knowledge of the **MDP model**
- **exhaustive search** over all (s,a) pairs

Existing approach does not apply to more complex settings (e.g., clinical applications)

We aim to develop an approach that:

- applies in **model-free** settings
- can be **solved efficiently**

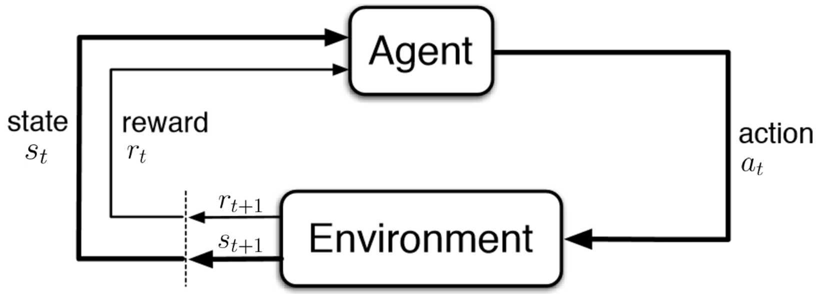
Fard, M. M., & Pineau, J. (2009). MDPs with non-deterministic policies. In Advances in Neural Information Processing Systems (pp. 1065-1072).
Fard, M. M., & Pineau, J. (2011). Non-deterministic policies in Markovian decision processes. Journal of Artificial Intelligence Research, 40, 1-24.

Our Contributions

We propose a **new algorithm** for learning **near-optimal set-valued policies** that can support clinician/human-in-the-loop decision-making

- Provide **theoretical analyses** that prove convergence in directed acyclic graphs (DAG)
- Demonstrate **empirical behavior** across synthetic environments including non-DAGs
- Show that the algorithm discovers meaningful action near-equivalencies on a **clinical task of sepsis treatment**

Problem Setting & Notation



(from Sutton & Barto's RL book pg 48)

Markov Decision Process $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$

- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$ transition model
- $\mathcal{R} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ reward function
- $\gamma \in [0, 1]$ discount factor

Trajectory $s_0, a_0, r_1, s_1, a_1, r_2, s_2, a_2, r_3, \ldots$

Return $G_0 = r_1 + \gamma r_2 + \gamma^2 r_3 + \cdots = \sum_{t=0}^{\infty} \gamma^t r_{t+1}$

Optimal value function V^*

Set-Valued Policy & Worst-Case Value Functions

SVP

$$\pi : \underset{\text{state}}{\mathcal{S}} \to \underbrace{2^{\mathcal{A}} \setminus \{\varnothing\}}_{\text{non-empty set of actions}}$$

$$V^{\pi}(s) = \boxed{\min_{a \in \pi(s)}} \{Q^{\pi}(s,a)\}$$

$$Q^{\pi}(s,a) = r(s,a) + \gamma \mathbb{E} \boxed{\min_{a' \in \pi(s')}} \{Q^{\pi}(s',a')\}$$

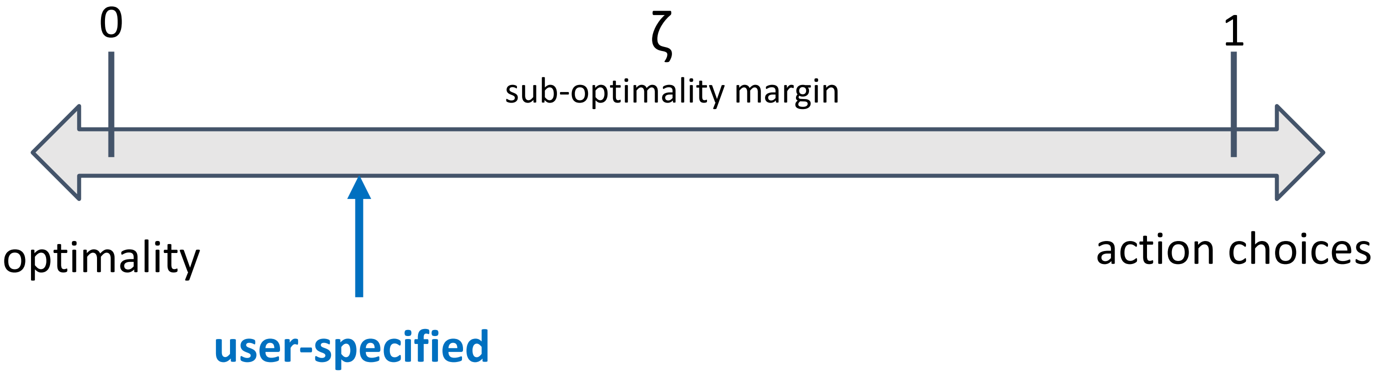
Considers a **worst-case analysis**

Fard, M. M., & Pineau, J. (2009). MDPs with non-deterministic policies. In Advances in Neural Information Processing Systems (pp. 1065-1072).
Fard, M. M., & Pineau, J. (2011). Non-deterministic policies in Markovian decision processes. Journal of Artificial Intelligence Research, 40, 1-24.

Near-optimality: multiplicative constraint

Requires $V^*(s) \geq 0 \ \forall s$

$$V^\pi(s) \geq (1-\zeta)V^*(s), \ \forall s \in \mathcal{S}$$



Fard, M. M., & Pineau, J. (2009). MDPs with non-deterministic policies. In Advances in Neural Information Processing Systems (pp. 1065-1072).
Fard, M. M., & Pineau, J. (2011). Non-deterministic policies in Markovian decision processes. Journal of Artificial Intelligence Research, 40, 1-24.

Objective

Given ζ, learn an SVP π that satisfies near-optimality.

$$V^{\pi}(s) \geq (1-\zeta)V^{*}(s), \ \forall s \in \mathcal{S}$$

Trivial Solution: equivalent to the greedy optimal policy,

$$\pi(s) = \{\pi^{*}(s)\}$$

→ We want to learn π to recommend **more actions** when possible

Key Idea

Standard RL setup: An optimal policy is a **fixed-point solution** to the following equation

$$\forall s \in \mathcal{S}, \quad \pi^*(s) = \underbrace{\arg\max_a}_{\text{Greedy action selection}} \overbrace{Q^{\pi^*}(s,a)}^{\text{Optimal value function}}$$

"an ***optimal*** policy is ***greedy*** with respect to its own Q-function"

Key Idea: Near-greedy heuristic

For set-valued policies: We formulate a similar equation and seek the **fixed-point solution**

$$\forall s \in \mathcal{S}, \quad \pi(s) = \{a : \overbrace{Q^\pi(s,a)}^{\text{Worst-case value of SVP } \pi} \underbrace{\geq (1-\zeta)V^*(s)}_{\text{Near-greedy action selection}}\}$$

"a ***near-optimal*** SVP should be ***near-greedy*** with respect to its Q-function"

Learning near-greedy SVPs

$$\forall s \in \mathcal{S}, \quad \pi(s) = \{a : Q^{\pi}(s,a) \geq (1-\zeta)V^{*}(s)\}$$

Using this equation to modify the Bellman backup, we can derive a family of **value-based algorithms** for learning SVPs :

policy iteration → near-greedy policy iteration

value iteration → near-greedy value iteration

Q-learning → near-greedy TD-learning

model-free
function approximator

etc.

*see paper for details

Theoretical analyses

$$\forall s \in \mathcal{S}, \quad \pi(s) = \{a : Q^{\pi}(s, a) \geq (1 - \zeta)V^{*}(s)\}$$

- The modified Bellman update operator is ***generally not a contraction***

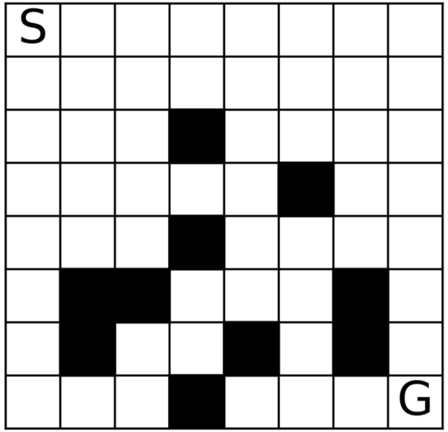
- **Thm 1:** If MDP is a *DAG with non-negative rewards*, then the near-greedy ζ-optimal SVP ***exists*** and is ***unique***.
- **Thm 2:** If MDP is a *DAG with non-negative rewards*, then "near-greedy TD-learning" ***converges*** to the **unique** solution, under the same convergence conditions for Q-learning.

Experiments

* Please refer to our paper for other experiments & results

1. Empirical behavior on non-DAG environments (FrozenLake)
   - can **converge** to **non-trivial solutions**
2. Application to a real clinical problem (MIMIC-sepsis)
   - discovers **meaningful near-equivalencies** among actions

1. Empirical behavior on non-DAG

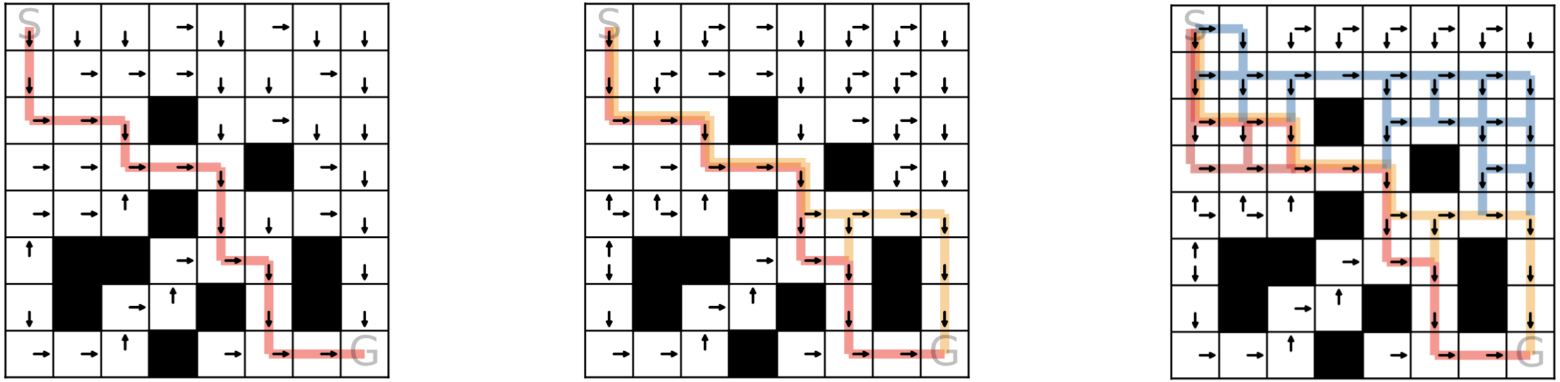


FrozenLake-8x8

Brockman et al. Openai Gym. arXiv:1606.01540, 2016.
https://gym.openai.com/envs/FrozenLake-v0/

- Gridworld
  - Task: get from **S** to **G** without falling into
  - Actions: ↑ ↓ ← →
- Base reward
  - +1 for transition to **G**
  - 0 otherwise
- Reward modifiers
  to induce near-equivalent actions
  - Randomly sampled from
    {0.001, 0.002, 0.003, 0.004}

1. Empirical behavior on non-DAG



$\zeta = 0.00$ Avg. policy size: 1.00

$\zeta = 0.01$ Avg. policy size: 1.25

$\zeta = 0.03$ Avg. policy size: 1.42

Despite a lack of theoretical guarantees for non-DAGs,
the proposed algorithm can converge to **useful solutions**.

2. Clinical task

*see paper for details

MIMIC-sepsis * (Komorowski 2018)

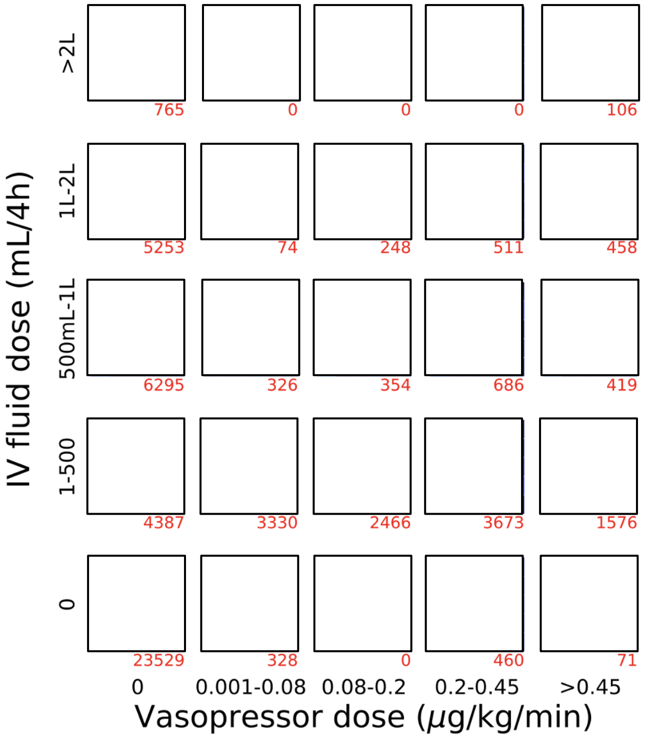
Goal: learn optimal treatment strategies for patients with sepsis in the ICU

- State space: derived from 48 physiological signals at 4h timesteps
- Action space: 25 treatment options, (5 vasopressor doses) x (5 intravenous fluids)
- Reward: survival (+100) vs death (-100)
- $\gamma = 0.99$

For illustration, we visualize action near-equivalencies at $\zeta = 0.05$

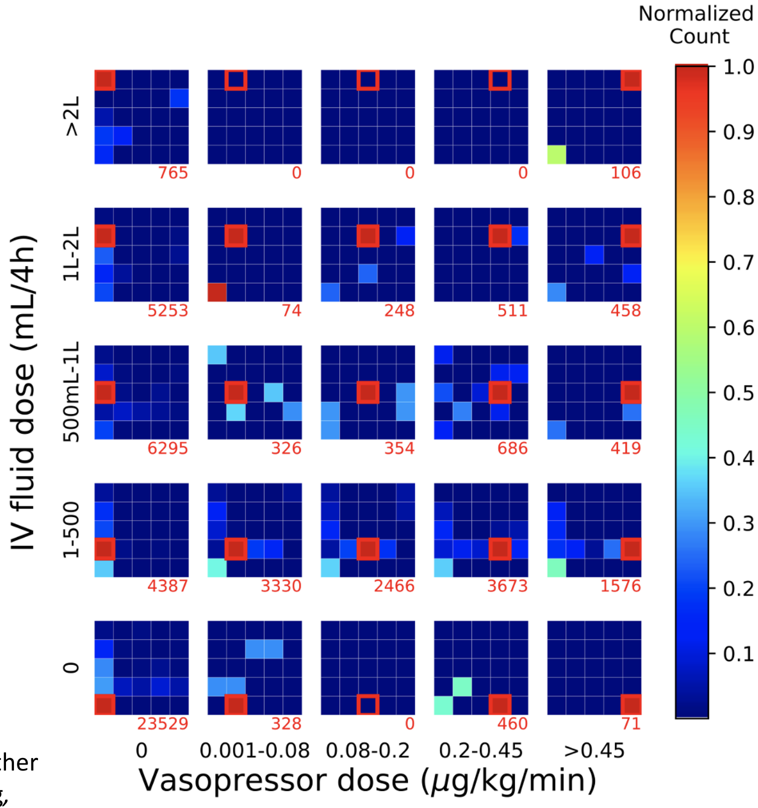
Komorowski, Matthieu, et al. "The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care." *Nature Medicine* 24.11 (2018): 1716.

2. Clinical task - Visualizing action near-equivalencies



Note: The red numbers indicate how often that action is considered **optimal** by the learned policy over all states in the **test set**.

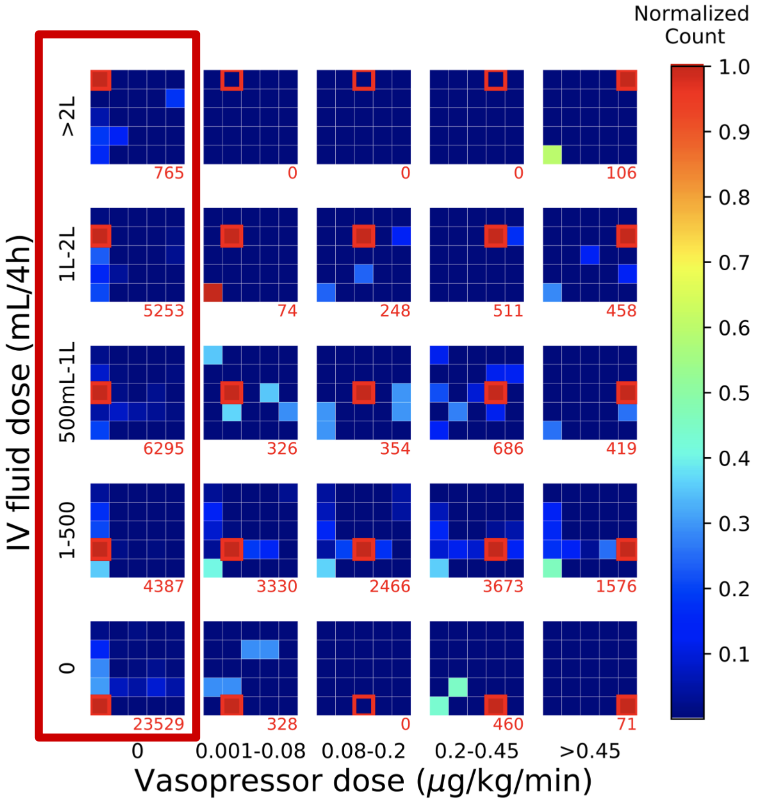
2. Clinical task - Visualizing action near-equivalencies



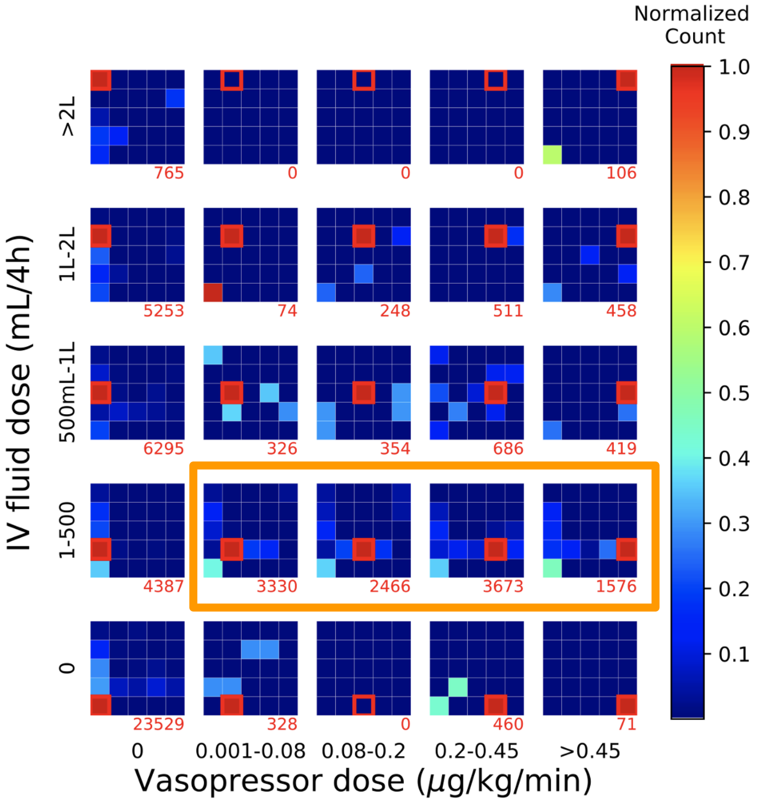
Note: Interpretation of results was conducted together with a *critical care physician,* Dr. Michael W. Sjoding, who treats patients with sepsis

2. Clinical task - Visualizing action near-equivalencies

- Most frequent actions have no vasopressors
- Similar doses of IV fluids considered near-equivalent

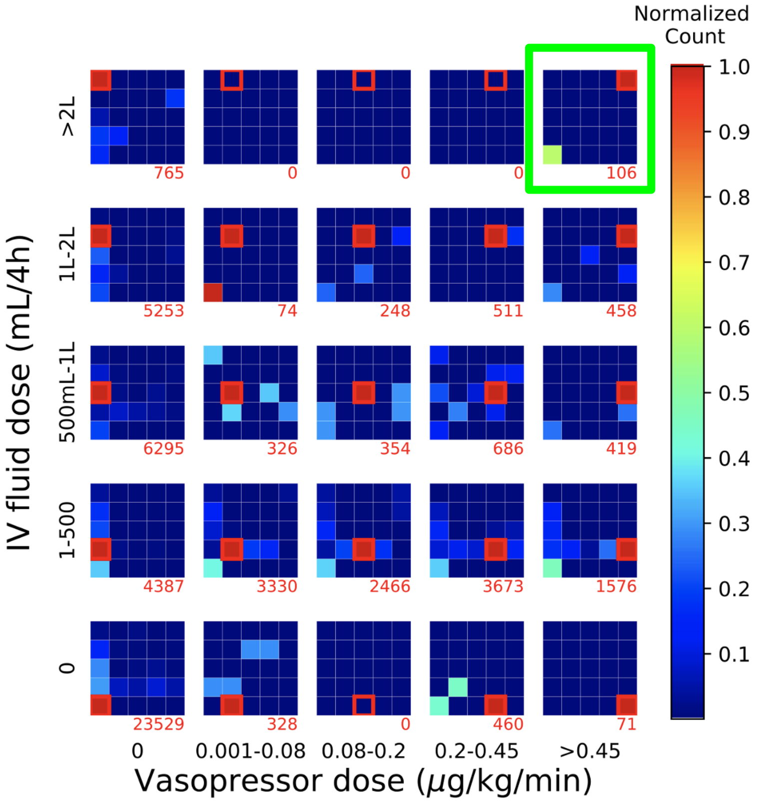


2. Clinical task - Visualizing action near-equivalencies



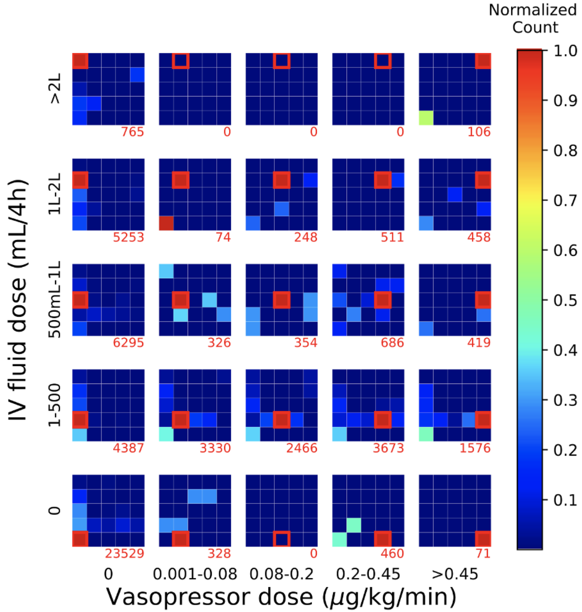
When a low dose of IV fluids is used, similar doses of vasopressors considered near-equivalent

2. Clinical task - Visualizing action near-equivalencies



Possibly very sick states, where "do-nothing" and "do-everything" could lead to similarly bad outcomes

2. Clinical task - Visualizing action near-equivalencies



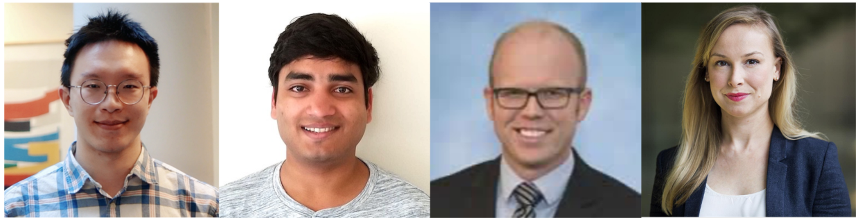
The proposed algorithm uncovers clinically **meaningful near-equivalencies** in terms of treating patients with sepsis.

We propose a **new algorithm** for learning **near-optimal set-valued policies**, which can provide action choices while maintaining near-optimality

- An important step for clinician/human-in-the-loop decision support
- Humans incorporate additional knowledge to select among near-equivalent actions
- Potential broader impact to other applications beyond healthcare

MICHIGAN MEDICINE
UNIVERSITY OF MICHIGAN

COMPUTER SCIENCE
& ENGINEERING
UNIVERSITY OF MICHIGAN



S. Tang A. Modi M.W. Sjoding J. Wiens

https://gitlab.eecs.umich.edu/MLD3/RL-Set-Valued-Policy

This work was supported by the National Library of Medicine (NLM grant no. R01LM013325).