Sepsis Treatment via Personalized Reinforcement Learning: A Multi-Head Dueling Double Deep Q-Network Approach

Selma Tabakovic, Chalmers University of Technology
Department of Applied Mathematics and Statistics
Supervisor: Marina Axelson-Fisk (Chalmers and University of Gothenburg)

CHALMERS UNIVERSITY OF TECHNOLOGY

1. Motivation and Research Goals

Sepsis is a life-threatening organ dysfunction caused by a dysregulated host response to infection, and remains a leading cause of death in intensive care units worldwide. An optimal treatment strategy is still unknown, leading to a significant variability in sepsis treatment. Recently, deep reinforcement learning have shown promise as a decision-aiding tool for the administration of intravenous fluids and vasopressors to septic patients. However, these models are limited in their ability to accommodate different patient profiles, and thus fail to provide personalized treatment recommendations.We propose a Multi-Head Dueling Double Deep Q-Network (MH-DQN) model that incorporates patient characteristics to enable personalized treatment recommendations.

2. Markov Decision Process Formulation

A Markov Decision Process is the foundation of reinforcement learning, consisting of the following four parts:

1. **State Space**: Patient health states S_t at time t.
2. **Action Space**: Physician actions A_t affecting S_{t+1}.
3. **Transition Probability**: Probability of S_{t+1} given S_t and A_t.
4. **Reward Function**: Reward for taking A_t in S_t resulting in S_{t+1}.

Physician's actions: Vasopressors and/or IV fluids, represented as a 5×5 grid of maximum vasopressor dose and total IV fluid volume over 4 hours.

3. Data and Patient Profiles

- **Data Source**: MIMIC-III database [1].
- **Clustering**: Fuzzy C-Means (FCM) on temporal SOFA score features.
- **Patient Profiles**: Clusters represent different sepsis severity levels.

Profile	Patients (n)	Mortality (%)
Mild	$5\,954\ (5\,942, 5\,966)$	$14.41\%\ (14.33\%, 14.49\%)$
Moderate	$5\,258\ (5\,249, 5\,267)$	$18.67\%\ (18.59\%, 18.75\%)$
Severe	$2\,042\ (2\,032, 2\,052)$	$36.02\%\ (35.85\%, 36.19\%)$

4. Model Architecture

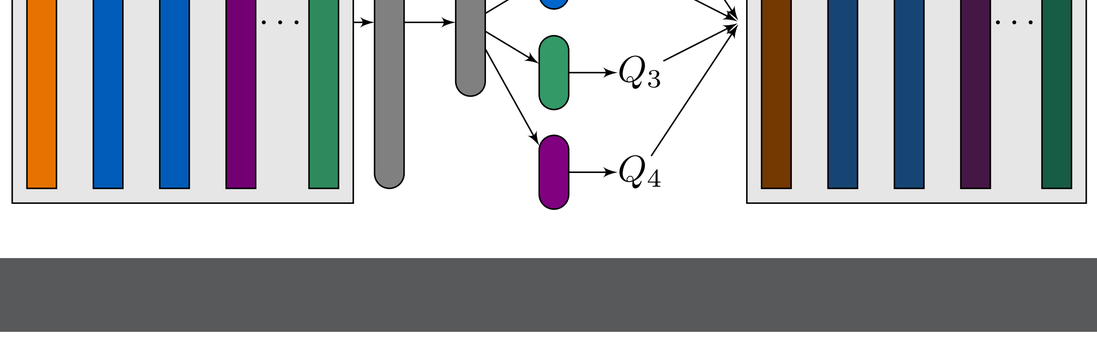
We compare our MH-DQN model, with its personalized multi-head architecture, to the existing Dueling DDQN to assess performance differences.

Dueling Double Deep Q-Learning (Dueling DDQN)

- **Loss Function**: $L(\theta) = \mathbb{E}\left[(Q_{\text{target}} - Q(S_t, A_t; \theta))^2\right]$ [2].
- **Q-Target Update**: $Q_{\text{target}} = R_{t+1} + \gamma \max_{a'} Q(S_{t+1}, a'; \theta')$.
- **Network Structure**: Separate networks estimate Q_{target} , with output streams for value and advantage.

Multi-Head Dueling DDQN (MH-DQN)

- **Personalization**: Enhances Dueling DDQN with a multi-head architecture tailored to patient profiles.
- **Output Weighing**: Output is a weighted sum of head outputs.
- **Weighting Strategies**: Hard weighing uses binary weights; soft weighing uses FCM membership scores.

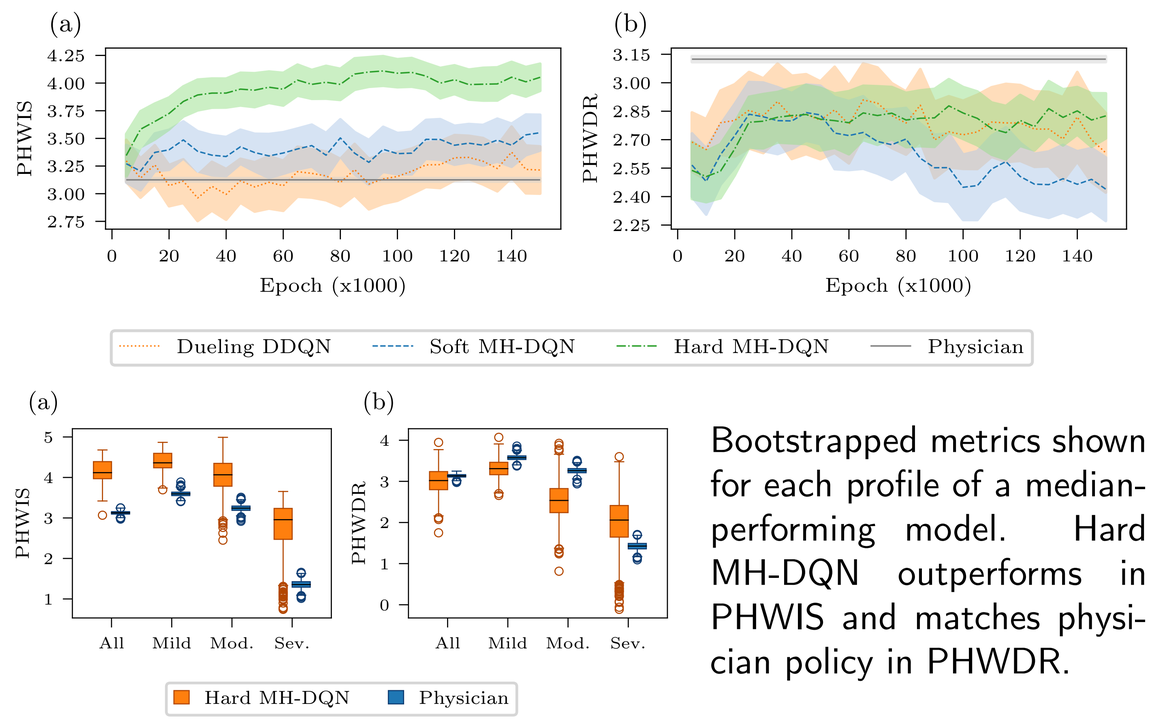


5. Off-Policy Evaluation

Off-policy evaluation assesses model performance using hospital data collected under a physician's treatment policy. Two metrics are used:

- Per-Horizon Weighted Importance Sampling (PHWIS) [3].
- Per-Horizon Weighted Doubly Robust (PHWDR) [4].

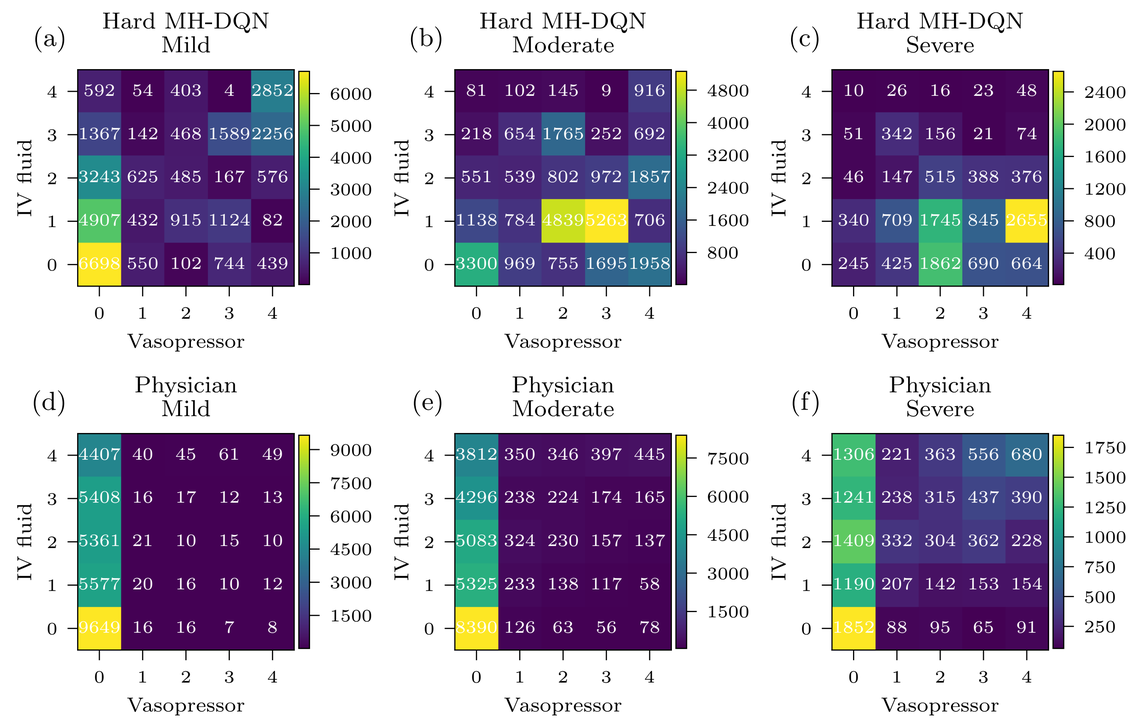
Evaluation on test sets using 50 models trained on different train-test splits. Hard MH-DQN excels, outperforming all models in PHWIS.



Bootstrapped metrics shown for each profile of a median-performing model. Hard MH-DQN outperforms in PHWIS and matches physician policy in PHWDR.

6. Action matrices

Policies are evaluated by comparing its action frequencies to the physician's, where action 0 means no drugs and higher actions indicate larger dosages.



The hard MH-DQN policy adapts more to the patient profile, resembling the physician's policy for mild profile and differing more for severe profile.

7. References

[1] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng *et al.*, "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, p. 160035, 2016.
[2] A. Raghu, M. Komorowski, L. A. Celi, P. Szolovits, and M. Ghassemi, "Continuous State-Space Models for Optimal Sepsis Treatment - a Deep Reinforcement Learning Approach," 2017.
[3] S. Doroudi, P. S. Thomas, and E. Brunskill, "Importance Sampling for Fair Policy Selection," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 2018, pp. 5239–5243.
[4] A. Raghu, O. Gottesman, Y. Liu, M. Komorowski, A. Faisal, F. Doshi-Velez, and E. Brunskill, "Behaviour Policy Estimation in Off-Policy Policy Evaluation: Calibration Matters," 2018.