Improving Sepsis Treatment Strategies using Deep Reinforcement Learning and Mixture-of-Experts

Xuefeng Peng, MSc.[1], Yi Ding, MSc.[2], David Wihl, A.L.B.[1], Omer Gottesman, PhD[1], Matthieu Komorowski, MD[3], Li-wei H. Lehman, PhD[4], Andrew Ross, MSE[1], Aldo Faisal, PhD[3], Finale Doshi-Velez, PhD[1]
[1]Harvard University, Paulson School of Engineering and Applied Sciences, Cambridge, MA
[2]Harvard University, T.H. Chan School of Public Health, Cambridge, MA
[3] Imperial College London, London, UK
[4]MIT, Institute for Medical Engineering & Science, Cambridge, MA

Abstract

Sepsis is the leading cause of mortality in the ICU. It is challenging to manage because different patients respond differently to treatment. Thus, tailoring treatment to the individual patient is essential for the best outcomes. In this paper, we take steps toward this goal by applying a mixture-of-experts framework to individualize sepsis treatment. The mixture model switches between neighbor-based (kernel) and deep reinforcement learning (DRL) experts depending on patient's current history. On a large retrospective cohort, this mixture-based approach outperforms physician, kernel only, and DRL-only experts.

Introduction

Sepsis is a medical emergency that requires rapid treatment.[1] In 2011 alone, the US spent $20.3 billion dollars on hospital care for patients with sepsis.[2] Sepsis is the cause of 6.0% of hospital admissions but 15.0% of hospital mortality.[3] Despite its significance, treating sepsis remains challenging, in part because there exists large variation in patient response to sepsis management techniques.[4]

In this work, we focus on two treatments in the context of sepsis management: intravenous (IV) fluid (adjusted for fluid tonicity) and vasopressor (VP). These two drugs are used to correct the hypovolemia and counteract sepsis-induced vasodilation, respectively. While hypovolemia and vasodilation are common among patients with sepsis, there exists little clinical consensus about when and how these should be treated;[5] however, these choices can have large effects on patient mortality.[4] Thus, it is essential to identify ways to personalize treatment. We chose to personalize treatment by combining two very different methods: a kernel-based approach and a deep reinforcement learning (DRL) approach. We then employ a switch function to discern which method is appropriate in a given patient context leading to a personalized treatment policy.

The availability of large observational critical care data sets[6] has made it possible to hypothesize improved sepsis management strategies, and prior work[7,8] has used this resource to suggest optimal treatment strategies for patients with sepsis. Our work extends these prior efforts in three important ways:

1. *Encoding patient ICU history recurrently.* It is often the case that evaluating the patient's current set of measurements using a Markov assumption are insufficient to represent all the aspects of the ICU history that are relevant for treatment. To retain more of this decision-relevant information, we use a recurrent autoencoder instead of a standard autoencoder to compress the patient's entire history rather than simply the most recent set of measurements.

2. *Safe-guarding DRL-based approach.* The deep Q-networks (DQN) used by DRL are prone to instability. We introduce safe-guards in our policy learning to prevent our models from suggesting actions rarely taken by clinicians, moving us toward policies that are more credible for clinical application.

3. *Mixture-of-experts.* Finally, we create a MoE model to switch between the restricted DRL approach and a neighbor-based approach resulting in significant improvements over earlier work. With a richer set of inputs, the MoE is able to dynamically switch policies during the clinical course and if necessary propose an action different from either expert.

Related Work

RL has been applied to a number of applications in healthcare, ranging from emergency decision support,[9] treating malaria,[10] and managing HIV.[11] In the realm of critical care, Prasad et al.[12] use RL to identify when to the wean patients from mechanical ventilation in ICUs.

Within the area of sepsis management, Komorowski et al.[8] use a discrete Markov Decision Process approach to identify when to administer fluids and vasopressors. Raghu et al.[7] extend this work by considering a much more expressive continuous representation of patient state. They use a traditional, non-recurrent autoencoder to first compress measurements from each time step into a continuous state representation, and then they learn a mapping from the state representation to an appropriate treatment via a Dueling Double-Deep Q Network (Dueling DDQN). Our work uses an even more expressive state representation that compresses the patient's entire history, and we also add safe-guards against insensible actions and develop richer policies through our mixture of experts.

Our mixture of experts approach builds from ideas presented by Parbhoo et al.[13] for HIV management. In their case, they switch between a kernel-based policy and a discrete Bayesian Partially Observable Markov Decision Process (POMDP). We follow the idea of combining experts, but use the DDQN as an expert rather than a discrete POMDP, as Raghu et al.[7] have already demonstrated that an expressive state representation is valuable for the sepsis management task.

Background

An RL agent interacts with the environment over time. At each time step t, the RL agent observes a state s from the state space S, selects an action a from the action space A following a policy π, a mapping from state to action. The agent receives a reward r, and transitions to a new state s'. This process continues until a terminal state is reached. The return is the discounted accumulated rewards $\sum_t \gamma^t r_t$. At each time step, the agent selects actions to maximize the expectation of the return. The optimal value function is defined as $V^*(s) = max_\pi \mathbf{E}[\sum_t \gamma^t r_t | s_0 = s, \pi]$. The optimal state-action value function $Q^*(s,a) = max_\pi \mathbf{E}[\sum_t \gamma^t r_t | s_0 = s, a_0 = a, \pi]$ is the maximum return achievable by any policy π for s and a. In Q-learning, the optimal state-action value function $Q^*(s,a)$ is approximated by Bellman Equation, $Q^*(s,a) = r + \gamma max_{a'} Q^*(s', a')$, where γ is the discount factor determines the trade-off between immediate and future rewards. $Q^*(s,a)$ is estimated based on the subsequent estimates; thus, temporal difference (TD) error, defined as $r + \gamma Q^*(s', a') - Q^*(s,a)$, are used as the criterion for either value iteration or function approximator learning.

Cohort and Data Processing

Cohort. We used the same patient set as in Raghu et al.[7] which applied the Sepsis-3 criteria to the Multi-parameter Intelligent Monitoring in Intensive Care (MIMIC-III v1.4) database.[6] Our cohort consisted of 15,415 adults (age range of 18 to 91), summarized in Table 1.

Table 1: Comparison of cohort statistics for subjects that fulfilled the Sepsis-3 criteria

	% Female	Mean Age	Total Population
Survivors	44.1%	63.9	13,535
Non-survivors	44.3%	67.4	1,880

Cleaning and Preprocessing. As in Raghu et al.[7], patient histories were partitioned into 4-hour windows each containing fifty attributes, ranging from vitals (e.g. heart rate, mean blood pressure) to clinician assessments of the patient's conditions (e.g. sequential organ failure assessment (SOFA) score). The full list of attributes is included in the Appendices. We removed the patients from their processed cohort that had any missing values.

The observations range from demographics, lab values and vital signs all of which have different scales. A table in the Appendices summarizes the preprocessing methods. Following Raghu et al.,[7] we performed log transformations of observations with large values and standardized the remaining values. After the standardization and log transformation, all values were rescaled into $[0-1]$. The data set was split into a fixed 75% training and validation set and a 25% test set.

Treatment Discretization. In this work, we focus on administrating two drugs: intravenous (IV) fluid and vasopressor (VP). In the cohort, the usage of IV and VP for each patient are recorded at each 4-hour window; and follow Raghu's work[7], the dosages for each drug are discretized into 5 bins, resulting in a 5×5 action space indexed from 0 to 24. Note that the first action ($a = 0$) means "no action" – neither IV nor VP are prescribed.

Method Overview

Applying RL to the sepsis management required several parts. While we have defined our patient histories and treatment space above, we must still define our objective, how we evaluate it, and how we transform patient histories into treatment policies. Below, we first present how we compress a patient's history into a state via a recurrent autoencoder. Next, we describe how we attribute rewards to each state, and also how we determine the quality of some policy given observational data. We finally describe how to derive treatment policies via a kernel-based expert, a DRL expert, and our mixture-of-experts (MoE) approach

Compressing Patient Histories

In treatment, having the patient trend information is key to deciding the appropriate course of action. Prior efforts[7,8] made a Markov assumption and therefore do not capture any trend or temporal information about the patient's ICU-stay histories. To capture more of this temporal information, we encoded patient states recurrently using an LSTM autoencoder representing the cumulative history for each patient. The LSTM had a single layer of 128 hidden units for both encoder and decoder—resulting in a state s that consisted of 128 real-valued vector. The autoencoder was trained to minimize the reconstruction loss (MSE) between the original measurements and the decoded measurements. We trained the model with mini-batches of 128 and the Adam optimizer for 50 epochs until its convergence.

Reward Formulation

Broadly, we are interested in reducing mortality among patients with sepsis. However, mortality is a challenging objective to optimize for because it is only observed after a long sequence of decisions; it can be hard to ascertain which action was responsible for a good or bad outcome. Thus, for the purposes of our training, we introduce an *intermediate* reward that can give a preliminary signal for whether our sequence of treatment decisions is likely to reduce mortality. Specifically, we consider the change in the negative mortality log-odds value between states as the reward. Using the change in mortality probability directly as the reward would constrain the reward value to a very small range; thus, we scaled the probabilities by log-odds, and applied negation over the log-odds values so that state with low mortality probability has higher log-odds value. Let $R(s, a, s')$ be the reward for taking action a at state s and transit to state s', $f(s)$ be the probability of mortality in state s. The reward function is described in Equation 1. The rewards from both training and testing sets are in the interval $[-3, 3]$.

$$R(s, a, s') = -\log \frac{f(s')}{1 - f(s')} f(s') + \log \frac{f(s)}{1 - f(s)} \tag{1}$$

To predict the probability of mortality, we trained a two-layer neural network with L1-regularized gradients (see Ross et al.[14] for details). The L1 regularization encourages sparse local linear approximations, and thus makes its behaviors more interpretable. Additionally, to better interpret the predicted behaviors, we trained the network using the unencoded patient observations described in the preprocessing section of the Appendices. We trained the network on a sampled, balanced training set which achieved a 73.1% 10-fold crossed validated accuracy. We present the mortality log-odds distribution for each class in the Appendices

Off-Policy Evaluation via the WDR Estimator

The natural question is how to evaluate the quality of a policy π_e given only retrospective data that was collected according to some other clinician behavioral policy π_b. The weighted doubly robust (WDR) estimator[15] is widely used for off-policy evaluation in RL. It requires a value function, V and a state-action value function Q. Accurate V and

Q functions reduce variance of the WDR.[16] The DQN's Q-function is learned during training by $\arg\max_a Q(s, a; \theta)$ (see Van Hasselt et al.[17] for additional details). Since the kernel and physician expert policies are not determined via RL, they do not have associated V and Q functions, and must therefore be estimated in order to calculate the WDR. We used the DQN's Q-function, using the $mean$ rather than $\arg\max$ in order to derive the equivalent physician policy. $\hat{V}$ was obtained by taking the max of the approximated $\hat{Q}$.

$$\text{WDR}(D) := \sum_{i=1}^{I}\sum_{t=0}^{T}\gamma^t w_i^t R_t^{H_i} - \sum_{i=1}^{I}\sum_{t=0}^{T}\gamma^t (w_t^i \hat{Q}^{\pi_e}(S_t^{H_i}, A_t^{H_i}) - w_{t-1}^i \hat{V}^{\pi_e}(S_t^{H_i})) \tag{2}$$

In Equation 2, n is the number of patients, t is the time step, and H_i refers to the i_{th} patient's ICU-stay state trajectory. The importance weight of the state of the patient i at time step t is defined as $\rho_t^i = \prod_{i=0}^{t} \frac{\pi_e(A_i^H|S_i^H)}{\pi_b(A_i^H|S_i^H)}$, and $w_t^i = \frac{\rho_t^i}{\sum_{j=1}^{J}\rho_t^j}$. R is the reward described above; $\hat{V}$ is estimate of the state value function; $\hat{Q}$ is the estimate of the state-action value function, and γ is a discount factor. To evaluate a learned policy, the behavior policy π_b is needed, which is the observed physician policy based on interventions performed on the patients. However, it is deterministic over each patient state, this makes the evaluation of other stochastic expert policies using WDR to suffer from high variances. Thus, we derive a stochastic physician policy representing the π_b. As physicians usually prescribe treatments to patients not only based on their expertise but also their prior experience of treating similar patients,[18] we define the behavior policy for some state s as being the empirical distribution over actions of the 300 neighbors in the training set with states closest to s.

Deriving Policies

With a means of representing patient history, and a metric for evaluating policies, we now discuss the mechanism for each expert, independently and in combination, to learn their respective policies.

Kernel Policy. One simple way to derive a treatment decision rule is to look at the nearest neighbors to the current state s, and choose the actions in the distribution that corresponding to only the survivors, instead of all the patients. Figure 1 summarizes the kernel policy derivation process. Specifically,

1. Recurrently encode a given patient's history of ICU observations from time 0 to t to identify the 128-dimensional vector describing their current state.
2. Search k nearest neighbors in the training set in this encoded representation space using Euclidean distance.
3. The policy is the distribution of actions taken over the surviving nearest neighbors.

We cross-validated the k values ranging from 200 to 500 using WDR, and proceeded with $k = 300$.

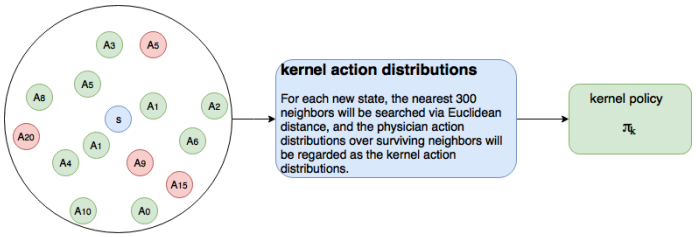


Figure 1: The circle in the left shows an example of the neighborhoods of a new state s, red and green marks the mortality and surviving states respectively, and each of these states is associated with a physician action A_i.

DQN Policy. Double DQN (DDQN) with dueling structure[19], which is a variant of DQN, has been successfully applied to derive a policy that outperforms the physician policy[7]. The structure of dueling DDQN is particularly suitable for

sepsis treatment strategy learning, as it differentiates the value function V into the value of the patient's underlying physiological condition, called the *Value* stream, and the value of the treatment given, called the *Advantage* stream.

We train the dueling DDQN for 200,000 steps with $batch\ size = 32$ to minimize the TD-error. At each given state, the agent is trained to take an action with the highest Q-value, in order to achieve the ultimate goal of improving the overall survival rate. To stabilize the training process and improve the performance, we applied two techniques: 1) a regularization term λ is added to the Q-network loss shown in Equation 3 to penalize output Q-values which exceed the allowed threshold $r_{max} = 3$; to learn the optimal policy more efficiently, 2) prioritized experience replay[20].

$$Ł(\theta) = E[(Q_{double-target} - Q(s, a; \theta))^2] + \lambda \max(|Q(s, a; \theta) - r_{max}|, 0) \tag{3}$$

where

$$Q_{double-target} = r + \gamma Q(s, \arg\max_{a'} Q(s, a'; \theta')) \tag{4}$$

Finally, given a set of action-values $Q(s, a)$, we must still define a policy π. Typically, these actions are chosen by the $\max Q$-value which makes the resulting policy deterministic. Combining a deterministic policy with the stochastic kernel policy to construct the mixture-of-experts is sub-optimal. To avoid this, we took advantage of the dueling double DQN structure to make the DQN policy stochastic. The dueling double DQN estimates the Q-value by approximating the underlying state value (*Value* stream) and the action value (*Advantage* stream) separately.[19] In this way, the *Advantage* stream reflects the qualities of the actions without being affected by the underlying value of the state. Thus, we created a policy action distribution for the DQN expert at each state by applying a *softmax* layer on its *Advantage* stream.

Mixture-of-Experts (MoE)

Due to the heterogeneity of the patient response, it is difficult for one single model to perform well on all types of patients.[13] For patient states which are atypical, i.e. farther Euclidean distance away from any neighbors, the kernel expert is unable to generalize resulting in a policy that may be too rigid and conservative. In contrast, DQN experts can over-generalize for these unusual patients. The mixture-of-experts (MoE) combines the kernel and the DQN policies in order to choose the appropriate policy for the appropriate context. The MoE generates mixed policy for any given state based on several characteristics of that state. The Figure 2 shows the architecture of MoE.

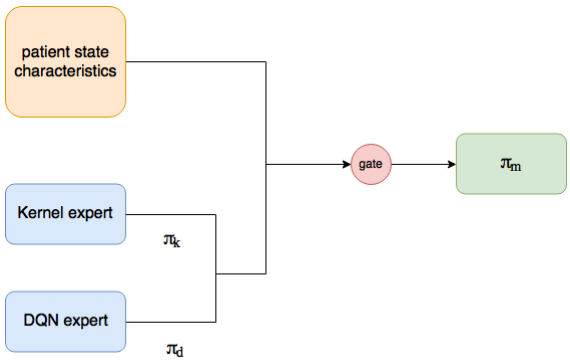


Figure 2: The architecture of MoE, it produces a mixed policy via combining kernel.

Action Restriction In situations where the DDQN has little data, the DDQN can arbitrary policies, including those that suggest actions that were rarely or never performed by clinicians. To safe guard against actions which may be potentially dangerous, we restricted the actions from the DDQN by the distribution of observed physician actions. Specifically, we discarded any DDQN suggested action taken less than 1% of the time by the physicians among its 300 nearest neighbors. Let $\pi_d(A|S)$ be the DDQN policy, and $\pi_b(A|S)$ be the physician policy. The DDQN suggests $\pi_d(a|S), a \in A$. If $\pi_b(a|S) < 0.01$, then, $\pi_d(a|S) \leftarrow 0$. Finally, $\pi_d(A|S)$ was normalized to be a valid probability distribution.

We did not restrict kernel policy as it is effectively the physician policy over survivors, and thus cannot deviate too far from clinician actions. As the MoE policy is a mixture of the kernel and DDQN policy, no additional restriction was needed to safe-guard it.

Logistic-based Switching The primary idea of logistic-based switching is to create a mixed policy per patient per time step that is overall superior to the distinct kernel and DQN policies. A probability as a function of the current state is assigned to each expert. The model used to predict the mortality probability described in Reward Formulation demonstrates the associations between the mortality and various patient demographic and lab values can be found in the Appendices. We examined several medical sources[21–23] to determine which features are most useful in the expert selection and chose: age, Elixhauser, SOFA, FiO_2, BUN, GCS, Albumin, trajectory length, and max distance from neighbors.

The MoE model's switching function learns the weights and bias for these input features in order to assign probabilities for each expert. The function is described in Equation 5, where X denotes the input features, W_k and b_k denote the weights for the input features and bias in terms of the kernel expert. p_k and p_d denote the assigned probability for kernel and DQN expert respectively.

$$\text{sigmoid}(W_k X + b_k) = p_k \tag{5}$$

$$p_d = 1 - p_k \tag{6}$$

MoE Policy Derivation With the expert probabilities assigned for the kernel and DQN experts, the MoE policy is defined as the sum of the products of the expert probabilities and the expert policy action probability distributions. Let $\pi_k(A|S)$ and $\pi_d(A|S)$ respectively denote the probability distribution of actions of the kernel and DQN policies over states. Then, the mixed policy is defined as $\pi_m(A|S) = p_k\pi_k(A|S) + p_d\pi_d(A|S)$.

The weights and bias for the input features in the switching function was trained to produce $\pi_m(A|S)$ that maximizes the WDR[15] estimate of the discounted expected return. The non-concavity of the WDR function makes the optimization challenging, so we conducted a random search of parameters to find local maxima.

Results

The estimates of the discounted expected return for each policy are presented in Table 2. Both kernel and DQN experts demonstrate an improvement over physician policy. We provide two columns for the mixture of experts policy because it is challenging to derive accurate value estimates V and Q for the WDR estimator in this case (as hard as the original policy evaluation task). Thus, we consider two sensible options: using the value estimates V and Q from the clinican policy, and using the estimates from the DDQN policy. Regardless of choice of evaluation covariate, the MoE projects a further improvement.

Table 2: Estimate of the discounted expected return for policies over test set, $\gamma = 0.99$)

	Physician	Kernel	DQN	MoE_{V_b,Q_b}	MoE_{V_d,Q_d}
Estimate	3.76	4.46	4.23	5.03	5.72

Analysis of discovered policies. We present the expert actions in Figure 3a. As shown, $a = 0$, i.e. no treatment action, is dominant for all policies. Actions which are in favored by physicians, such as IV but no vasopressor are also prescribed often by the kernel expert. But, our kernel expert tends to be more conservative than the physicians as it suggests no action approximately twice the physicians' frequency–perhaps reflecting a bias toward the fact that those patients who were not treated were somehow healthier and thus survived. In addition, the kernel expert prescribes more fluids when compared with that of physicians. While kernel expert and physician give action high fluids and no vasopressor almost in the same frequency, the kernel expert very rarely suggests actions with vasopressor. Our kernel expert action distribution shares the general behavior of a conservative physician.

The DQN expert, like physicians, favors giving actions a range of fluid values. But, over the test set, DQN expert

prescribes more extreme values when compared with physicians. Physicians frequently prescribe high fluids and low vasopressor; however, the DQN expert also tends to give more high vasopressor dosage actions in addition to fluids.

Over the test set, the behavior of the MoE is much akin to that of the kernel expert, as expected, as most of the states from patients can be mapped into a group of similar states. But driven by the DQN expert, MoE prescribes more high dosage actions. That said, for many given patient states, the actions suggested by experts overlap (see Table 3). However in 4.4% of circumstances, the MoE policy follows neither that of kernel nor that of DQN.

Table 3: Percent similarity of different policies over test set patient states

	kernel	DQN	MoE
physician	0.305	0.151	0.296
kernel	-	0.182	0.871
DQN	-	-	0.258

Finally, to better understand the behaviors of the MoE on forming $\pi_m(S|A)$, we show its switching function's parameters W_k in Table 4.

Table 4: MoE model parameters

	age	elixhauser	SOFA	GCS	FiO2	BUN	albumin	traj. len.	max dist.
Wk	0.0100	-0.3328	0.0556	-0.0331	-0.2714	-0.3713	-0.4085	0.1429	-0.4134

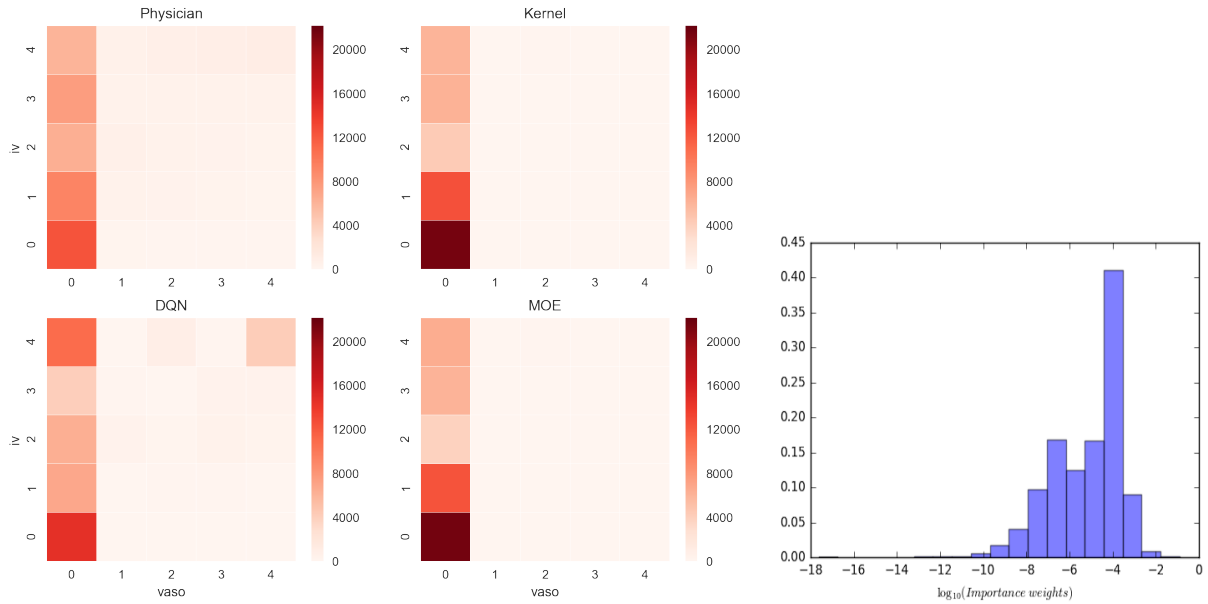
Evaluation Quality Assessment The WDR estimator of policy quality relies on having a large enough collection of patient histories in the evaluation set having non-zero weight w_t^i. For the MoE policy, 90% of the importance sampling weights are non-zero and 86% final weights in the sequences are non-zero. These high numbers of non-zero importance weights indicates that nearly all of our data was used in the evaluation of the policy. We plot the full distribution of weights in figure 3b. A significant number of weights lie in the range of $[10^{-4}, 10^{-3}]$ and only very few observations have weights significantly larger than that range (the samples with significantly smaller weights are unlikely to have a significant influence on the estimate). However, the few observations with weights on the order of 10^{-1} could potentially have large influence.

Discussion

A key part of managing sepsis patients is to restore their blood pressure through the administration of intravenous fluids and/or vasopressors. Prescribing an optimal balance between vasopressors and fluids remains challenging. Vasopressors are known to have harmful effects on patients; recent studies have also demonstrated the association between fluid-overload and negative outcomes in the ICUs.[24] Overall, the DQN recommended a treatment strategy with more aggressive use of both vasopressors and fluids. In comparison to the physician policy, DQN recommended 70% more actions involving medium-to-high fluid volume and vasopressor dosage (actions 18,19,23, and 24). Most notably, frequency for the DQN action corresponding to maximum levels of both fluid and vasopressor (action 24) increased by 3.8 fold from the physician policy. These results suggest that while recent advances in deep reinforcement learning aim to learn optimal strategies to guide interventions, further investigations are required, and careful clinical judgment should be exercised to guard against potentially high-risk actions introduced due to instability of non-linear function approximation.

The proposed kernel expert, in it's current implementation, recommended far fewer actions involving vasopressors in comparison to both the physician policy and DQN. While we have chosen to favor a more conservative strategy, e.g. regularizing rare actions as a safety mechanism to guard against unduly aggressive actions, we recognize that both over and under-treatment can potentially lead to adverse outcomes. Further investigation is required to examine clinical effects of a conservative kernel expert policy, and whether a more personalized approach may be possible to automatically detect when rare actions are beneficial on an individual patient basis.

Limitations. While it appears that our MoE policy significantly outperforms the clinician policy (as well as each individual expert), and we have ensured that the actions it suggests are at least sensible, that is, often taken by clinicians,



(a) Action Distributions for physician and experts over test set

(b) WDR importance weights distribution

Figure 3

there still exist a number of limitations. When encoding the patient clinical course in an LSTM representation, in spite of our high prediction accuracy, we cannot be certain that there are no hidden confounding factors. Using only survivors in the kernel policy may have biased us toward considering healthier patients, and all optimizations were prone to local optima.

On the evaluation side, the WDR estimator requires either the behavior policy estimate π_b to be accurate or the covariate estimates V and Q to be accurate to be unbiased; in our case, we estimated these values to the best of our ability, and showed that our results were at least insenstive to different choices of V and Q, but we do not have the true values of either.

More broadly, our work focused on a very specific subproblem with sepsis management, with a very specific reward structure. To apply off-policy evaluation (WDR) in a statistically credible way, we considered the accumulation of low mortality risk as the objective, rather than mortality itself (as the latter could not be evaluated reliably). There also exist many other interventions, such as antibiotics use and mechanical ventilation, that also affect the patient's outcomes. Finally, our data were limited to the aforementioned fifty variables and only during patients' ICU stays (aside from pre-ICU fluid balance). Future work remains to investigate policies that incorporated a broader scope of patient history as well as a larger variety of interventions.

Conclusion

We presented a MoE framework to learn improved fluid and vasopressor administration strategies for sepsis patients in ICUs using observational data. We demonstrated that the proposed mixture model approach can automatically adapt to patient states at each time step, and dynamically switch between a conservative kernel expert and a more aggressive DQN-generated policy to achieve better expected outcomes than physician, kernel only, and DRL-only experts. While further investigation—such as shadow prospective evaluation—is required to truly validate the efficacy of our approach, the proposed MoE framework represents a novel approach in dynamically switching between two treatment policies, and could potentially provide a safer mechanism to progressively deploy new policies, safe-guarding against actions potentially harmful to the patients.

Acknowledgments

We would like to thank the other students in Harvard CS282R - Reinforcement Learning for Healthcare, Fall 2017 for their insights, encouragement and feedback. Omer Gottesman was supported by the Harvard Data Science Initiative.

References

1. Christopher W Seymour, Foster Gesten, Hallie C Prescott, Marcus E Friedrich, Theodore J Iwashyna, Gary S Phillips, Stanley Lemeshow, Tiffany Osborn, Kathleen M Terry, and Mitchell M Levy. Time to treatment and mortality during mandated emergency care for sepsis. *New England Journal of Medicine*, 376(23):2235–2244, 2017.

2. Anne Pfuntner, Lauren M Wier, and Claudia Steiner. Costs for hospital stays in the united states, 2011: Statistical brief# 168. 2006.

3. Chanu Rhee, Raymund Dantes, Lauren Epstein, David J Murphy, Christopher W Seymour, Theodore J Iwashyna, Sameer S Kadri, Derek C Angus, Robert L Danner, Anthony E Fiore, et al. Incidence and trends of sepsis in us hospitals using clinical vs claims data, 2009-2014. *Jama*, 318(13):1241–1249, 2017.

4. Jason Waechter, Anand Kumar, Stephen E Lapinsky, John Marshall, Peter Dodek, Yaseen Arabi, Joseph E Parrillo, R Phillip Dellinger, Allan Garland, Cooperative Antimicrobial Therapy of Septic Shock Database Research Group, et al. Interaction between fluids and vasoactive agents on mortality in septic shock: a multicenter, observational study. *Critical care medicine*, 42(10):2158–2168, 2014.

5. PE Marik. The demise of early goal-directed therapy for severe sepsis and septic shock. *Acta Anaesthesiologica Scandinavica*, 59(5):561–567, 2015.

6. Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3, 2016.

7. Aniruddh Raghu, Matthieu Komorowski, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Continuous state-space models for optimal sepsis treatment - a deep reinforcement learning approach. *CoRR*, abs/1705.08422, 2017.

8. M Komorowski, A Gordon, LA Celi, and A Faisal. A markov decision process to suggest optimal treatment of severe infections in intensive care. In *Neural Information Processing Systems Workshop on Machine Learning for Health*, 2016.

9. Devinder Thapa, In-Sung Jung, and Gi-Nam Wang. Agent based decision support system using reinforcement learning under emergency circumstances. In *International Conference on Natural Computation*, pages 888–892. Springer, 2005.

10. Pranav Rajpurkar, Vinaya Polamreddi, and Anusha Balakrishnan. Malaria likelihood prediction by effectively surveying households using deep reinforcement learning. *arXiv preprint arXiv:1711.09223*, 2017.

11. Sonali Parbhoo. *A reinforcement learning design for HIV clinical trials*. PhD thesis, 2014.

12. Niranjani Prasad, Li-Fang Cheng, Corey Chivers, Michael Draugelis, and Barbara E Engelhardt. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. *arXiv preprint arXiv:1704.06300*, 2017.

13. Sonali Parbhoo, Jasmina Bogojeska, Maurizio Zazzi, Volker Roth, and Finale Doshi-Velez. Combining kernel and model based learning for hiv therapy selection. *AMIA Summits on Translational Science Proceedings*, 2017:239, 2017.

14. Andrew Ross. Transparent and interpretable machine learning in safety critical environments. NIPS, 2017.

15. Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148, 2016.

16. Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. *arXiv preprint arXiv:1511.03722*, 2015.

17. Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *AAAI*, volume 16, pages 2094–2100, 2016.

18. Geoffrey Norman. Research in clinical reasoning: past history and current trends. *Medical education*, 39(4):418–427, 2005.

19. Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Van Hasselt, Marc Lanctot, and Nando De Freitas. Dueling network architectures for deep reinforcement learning. *arXiv preprint arXiv:1511.06581*, 2015.

20. Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015.

21. Alan E Jones, Stephen Trzeciak, and Jeffrey A Kline. The sequential organ failure assessment score for predicting outcome in patients with severe sepsis and evidence of hypoperfusion at the time of emergency department presentation. *Critical care medicine*, 37(5):1649, 2009.

22. Kevin Beier, Sabitha Eppanapally, Heidi S Bazick, Domingo Chang, Karthik Mahadevappa, Fiona K Gibbons, and Kenneth B Christopher. Elevation of bun is predictive of long-term mortality in critically ill patients independent of'normal'creatinine. *Critical care medicine*, 39(2):305, 2011.

23. F Tamion. Albumin in sepsis. In *Annales francaises d'anesthesie et de reanimation*, volume 29, pages 629–634, 2010.

24. Diana Kelm, Jared T. Perrin, Rodrigo Cartin-Ceba, Ognjen Gajic, Louis Schenck, and Cassie Kenned. Fluid overload in patients with severe sepsis and septic shock treated with early-goal directed therapy is associated with increased acute need for fluid-related medical interventions and hospital death. *Shock*, 43(1):68–73, 2015.

Appendices

Due to space constraints, all appendices are available at
https://dtak.github.io/cs282-f17-xuefeng-yi-david/peng_sepsis_moe_appendix.pdf