Interpretable Off-Policy Evaluation in Reinforcement Learning by Highlighting Influential Transitions – Appendix

1. Derivations for Kernel-Based FQE

1.1. Proof of Proposition 1

Proposition 1. *For all transitions in the dataset, the values for corresponding state-action pairs are given by*

$$\hat{\mathbf{q}}'_t = \left(\sum_{t'=1}^{t} \gamma^{t'-1}\mathbf{M}'^{t'}\right)\mathbf{r} \equiv \mathbf{\Phi}'_t\mathbf{r} \tag{1}$$

$$\hat{\mathbf{q}}_t = \mathbf{M}\left(\sum_{t'=1}^{t} \left(\gamma\mathbf{M}'\right)^{t'-1}\right)\mathbf{r} \equiv \mathbf{\Phi}_t\mathbf{r}. \tag{2}$$

where $\hat{q}'_{t,i}$ and $\hat{q}_{t,i}$ are the estimated policy values at $(x'^{(i)}, \pi_e(x'^{(i)}))$ and $(x^{(i)}, a^{(i)})$, respectively, for the observed transition i

Proof. We first prove 1 by induction. We start by noting that for a given observed transition, i, averaging over all observations j such that Δ_{ij} holds can be written as $\frac{1}{N_i}\sum_{j:\Delta_{ij}} = \sum_j \frac{\mathbb{I}(\Delta_{ij})}{N_i} = \sum_j M_{ij}$. Similarly, averaging over all j such that $\Delta_{i'j}$ holds can be written as $\sum_j M'_{ij}$. Therefore, if $u(x,a)$ is some function over the sttate-action space and $\mathbf{u}$ is a vector containing the quantity $u_i = u(x^{(i)}, a^{(i)})$ for every $(x^{(i)}, a^{(i)})$, then the nearest-neighbors estimation of $u(x'^{(i)}, \pi(x'^{(i)}))$ is given by $[\mathbf{M}'\mathbf{u}]_i$.

Given the formulation above, for $t = 1$, $\hat{q}'_{t,i}$ estimates the reward at $(x'^{(i)}, \pi(x'^{(i)}))$, and can be written as:

$$\hat{\mathbf{q}}'_1 = \mathbf{M}'\mathbf{r}. \tag{3}$$

For $t > 1$, assume $\hat{\mathbf{q}}'_{t-1} = \left(\sum_{t'=1}^{t-1} \gamma^{t'-1}\mathbf{M}'^{t'}\right)\mathbf{r}$. Then

$$\begin{aligned}
\hat{\mathbf{q}}'_t &= \mathbf{M}'\left(\mathbf{r} + \gamma\hat{\mathbf{q}}'_{t-1}\right) \\
&= \mathbf{M}'\left[\mathbf{r} + \gamma\left(\sum_{t'=1}^{t-1} \gamma^{t'-1}\mathbf{M}'^{t'}\right)\mathbf{r}\right] \\
&= \left(\mathbf{M}' + \sum_{t'=1}^{t-1} \gamma^{t'}\mathbf{M}'^{t'+1}\right)\mathbf{r} \\
&= \left(\sum_{t'=0}^{t-1} \gamma^{t'}\mathbf{M}'^{t'+1}\right)\mathbf{r} \\
&= \left(\sum_{t'=1}^{t} \gamma^{t'-1}\mathbf{M}'^{t'}\right)\mathbf{r} \equiv \mathbf{\Phi}'_t\mathbf{r},
\end{aligned} \tag{4}$$

completing the proof of 1. To estimate $\hat{\mathbf{q}}_t$, we write $\hat{q}_{t,i} = \frac{1}{N_i}\sum_{j:\Delta_{ij}}\left(r^{(j)} + \gamma\hat{q}'_{t-1,j}\right)$ or in matrix notation.

$$\begin{aligned}
\hat{\mathbf{q}}_t &= \mathbf{M}\left(\mathbf{r} + \gamma\hat{\mathbf{q}}'_{t-1}\right) \\
&= \mathbf{M}\left(\mathbf{I} + \gamma\left(\sum_{t'=1}^{t-1} \gamma^{t'-1}\mathbf{M}'^{t'}\right)\right)\mathbf{r} \\
&= \mathbf{M}\left(\sum_{t'=1}^{t} \left(\gamma\mathbf{M}'\right)^{t'-1}\right)\mathbf{r} \equiv \mathbf{\Phi}_t\mathbf{r}.
\end{aligned} \tag{5}$$

□

1.2. Computational complexity.

Computation of a single influence value, $\tilde{I}_{i,j}$ requires summation over all transitions k that satisfy $\Delta_{k'j}$. Denote the number of such neighbors by $N^*_{j'}$ [1]. We expect $N^*_{j'}$ to be small and not scale with the size of the dataset, and also $\tilde{I}_{i,j}$ is inversely proportional to $N^*_{j'}$. Thus, if we only compute the influence of transitions such that $N^*_{j'} < \frac{v_{\max}}{\tilde{v}\tilde{I}_c} \equiv N^*_{j',c}$, where $v_{\max}$ is the maximum possible value, we are guaranteed not to miss any transitions with influence larger than our threshold $\tilde{I}_c$. Since $N^*_{j',c}$ does not scale with the size of the data, computation of a single individual influence can effectively be done in constant time. Performing influence analysis on a full dataset requires computing the influences of all transitions on all initial transitions, and therefore takes $\mathcal{O}(N|\mathcal{D}^*_0|)$ time.

In our matrix formulation, the FQE evaluation itself is bottlenecked by computing the matrix Φ, which includes computation of powers of M'. Because M' is a sparse matrix (each row i only has N_i nonzero elements), the matrix multiplication itself can be done in $\mathcal{O}(N^2)$ rather than $\mathcal{O}(N^3)$ time, and the entire evaluation is done in $\mathcal{O}(N^2T)$ time. Importantly, the influence analysis analyzing all transitions has lower complexity than the OPE, and should not significantly increase the computational cost of the evaluation pipeline.

[1] Note that $N^*_{j'}$, which counts all k that satisfy $\Delta_{k'j}$, is subtly different from the quantity $N_{j'}$ introduced in section 5.1, which counts all k that satisfy $\Delta_{j'k}$.

2. Derivations Linear Least-Squares FQE

Proposition 2. *The the linear least square solution of fitted Q evaluation is* $(\mathbf{\Psi}^\top\mathbf{\Psi} - \gamma\mathbf{\Psi}^\top\mathbf{\Psi}_p)^{-1}\mathbf{\Psi}^\top\mathbf{r}$

Proof. The least-square solution of parameter vector $\mathbf{w}$ can be found by minimizing the following square error of the Bellman equation for all (x, a) in the dataset:

$$\left(\hat{q}(x,a) - r(x,a) - \gamma\hat{q}(x', \pi_e(x'))\right)^2 \tag{6}$$

Plugging in $\hat{q}(x,a) = \boldsymbol{\psi}(x,a)^\top\mathbf{w}$, the square error is

$$\left(\boldsymbol{\psi}(x,a)^\top\mathbf{w} - r(x,a) - \gamma\boldsymbol{\psi}_\pi(x')^\top\mathbf{w}\right)^2 \tag{7}$$

By definition of $\mathbf{\Psi}$ and $\mathbf{\Psi}_p$, the mean square error over the N samples is:

$$\|\mathbf{\Psi}\mathbf{w} - \mathbf{r} - \gamma\mathbf{\Psi}_p\mathbf{w}\|_2^2 \tag{8}$$

The least square solution is:

$$\mathbf{w} = (\mathbf{\Psi}^\top\mathbf{\Psi})^{-1}\mathbf{\Psi}^\top(\mathbf{r} + \gamma\mathbf{\Psi}_p\mathbf{w}) \tag{9}$$

$$(\mathbf{\Psi}^\top\mathbf{\Psi})\mathbf{w} = \mathbf{\Psi}^\top(\mathbf{r} + \gamma\mathbf{\Psi}_p\mathbf{w}) \tag{10}$$

$$(\mathbf{\Psi}^\top\mathbf{\Psi} - \gamma\mathbf{\Psi}^\top\mathbf{\Psi}_p)\mathbf{w} = \mathbf{\Psi}^\top\mathbf{r} \tag{11}$$

$$\mathbf{w} = (\mathbf{\Psi}^\top\mathbf{\Psi} - \gamma\mathbf{\Psi}^\top\mathbf{\Psi}_p)^{-1}\mathbf{\Psi}^\top\mathbf{r} \tag{12}$$

□

Proposition 3. *Let* $\mathbf{C}_{-j} = (\mathbf{\Psi}_{-j}^\top\mathbf{\Psi}_{-j} - \gamma\mathbf{\Psi}_{-j}^\top\mathbf{\Psi}_{p,-j})$ *and* $\mathbf{C} = (\mathbf{\Psi}^\top\mathbf{\Psi} - \gamma\mathbf{\Psi}^\top\mathbf{\Psi}_p)$.

$$\mathbf{w}_{-j} = (\mathbf{C}_{-j})^{-1}\left(\mathbf{\Psi}^\top\mathbf{r} - r^{(j)}\boldsymbol{\psi}_j\right)$$

$$(\mathbf{C}_{-j})^{-1} = \mathbf{B}_j - \frac{\gamma\mathbf{B}_j\boldsymbol{\psi}_j\boldsymbol{\psi}_{\pi,j}^\top\mathbf{B}_j}{1 + \gamma\boldsymbol{\psi}_{\pi,j}^\top\mathbf{B}_j\boldsymbol{\psi}_j}$$

where

$$\mathbf{B}_j = \mathbf{C}^{-1} + \frac{\mathbf{C}^{-1}\boldsymbol{\psi}_j\boldsymbol{\psi}_j^\top\mathbf{C}^{-1}}{1 - \boldsymbol{\psi}_j^\top\mathbf{C}^{-1}\boldsymbol{\psi}_j}$$

Proof. By the list squares solution of FQE, $\mathbf{w}_{-j}$ equals $(\mathbf{\Psi}_{-j}^\top\mathbf{\Psi}_{-j} - \gamma\mathbf{\Psi}_{-j}^\top\mathbf{\Psi}_{p,-j})^{-1}\mathbf{\Psi}_{-j}^\top\mathbf{r}_{-i}$. Since $\mathbf{\Psi}_{-j}^\top\mathbf{r}_{-j} = \mathbf{\Psi}^\top\mathbf{r} - r^{(i)}\boldsymbol{\psi}_j$, we have that $\mathbf{w}_{-j} = (\mathbf{C}_{-j})^{-1}\left(\mathbf{\Psi}^\top\mathbf{r} - r^{(j)}\boldsymbol{\psi}_j\right)$. Then

$$\mathbf{C}_{-j} = \mathbf{C} - \boldsymbol{\psi}_j\boldsymbol{\psi}_j^\top + \gamma\boldsymbol{\psi}_j\boldsymbol{\psi}_{\pi,j}^\top, \tag{13}$$

because

$$\mathbf{\Psi}_{-j}^\top\mathbf{\Psi}_{-j} = \mathbf{\Psi}^\top\mathbf{\Psi} - \boldsymbol{\psi}_j\boldsymbol{\psi}_j^\top \tag{14}$$

$$\mathbf{\Psi}_{-j}^\top\mathbf{\Psi}_{p,-j} = \mathbf{\Psi}^\top\mathbf{\Psi}_p - \boldsymbol{\psi}_j\boldsymbol{\psi}_{\pi,j}^\top \tag{15}$$

This indicate $\mathbf{C}_{-j}$ equals $\mathbf{C}$ plus two rank-1 matrices. Fortunately, we can store $\mathbf{C}^{-1}$ when we compute $\mathbf{w}$ and $\hat{q}$. The following result named Sherman–Morrison formula allow us to compute $\mathbf{C}_{-j}^{-1}$ from $\mathbf{C}^{-1}$ in an efficient way. For any invertible matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ and vector $u, v \in \mathbb{R}^d$:

$$(\mathbf{A} + uv^\top)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}uv^\top\mathbf{A}^{-1}}{1 + v^\top\mathbf{A}^{-1}u} \tag{16}$$

Then if we define $\mathbf{B}_j \equiv \left(\mathbf{C} - \boldsymbol{\psi}_j\boldsymbol{\psi}_j^\top\right)^{-1}$, we have that

$$\mathbf{B}_j = \mathbf{C}^{-1} + \frac{\mathbf{C}^{-1}\boldsymbol{\psi}_j\boldsymbol{\psi}_j^\top\mathbf{C}^{-1}}{1 - \boldsymbol{\psi}_j^\top\mathbf{C}^{-1}\boldsymbol{\psi}_j} \tag{17}$$

$$(\mathbf{C}_{-j})^{-1} = \mathbf{B}_j - \frac{\gamma\mathbf{B}_j\boldsymbol{\psi}_j\boldsymbol{\psi}_{\pi,j}^\top\mathbf{B}_j}{1 + \gamma\boldsymbol{\psi}_{\pi,j}^\top\mathbf{B}_j\boldsymbol{\psi}_j} \tag{18}$$

□

3. Influence computation of importance sampling methods

The standard importance sampling (IS) estimator is given by

$$\hat{v}_{IS}^{\pi_e} = \frac{1}{N}\sum_{n=1}^{N} w_{0:T}^{(n)} g_T^{(n)}, \tag{19}$$

where the summation is over all N trajectories in the dataset, and the importance sampling weight $w_{0:t}$ is given by

$$w_{0:t}^{(n)} = \prod_{t'=0}^{t} \frac{\pi_e(a_{t'}^{(n)}|s_{t'}^{(n)})}{\pi_b(a_{t'}^{(n)}|s_{t'}^{(n)})}. \tag{20}$$

The total influence of trajectory j is then

$$\begin{aligned} I_j &= \hat{v}_{-j} - \hat{v} \\ &= \frac{1}{N-1}\sum_{n\neq j} w_{0:T}^{(n)} g_T^{(n)} - \frac{1}{N}\sum_{n=1}^{N} w_{0:T}^{(n)} g_T^{(n)} \\ &= \sum_{n=1}^{N} w_{0:T}^{(n)} g_T^{(n)} \left(\frac{1}{N-1} - \frac{1}{N}\right) - \frac{1}{N-1} w_{0:T}^{(j)} g_T^{(j)} \\ &= \frac{1}{N(N-1)}\sum_{n=1}^{N} w_{0:T}^{(n)} g_T^{(n)} - \frac{1}{N-1} w_{0:T}^{(j)} g_T^{(j)} \\ &= \frac{1}{N-1}\left(\hat{v}_{IS} - w_{0:T}^{(j)} g_T^{(j)}\right). \end{aligned} \tag{21}$$

This relation is nothing more then the fact that removing the j^{th} sample from an average over N samples, $\bar{x} = \frac{1}{N}\sum x^{(n)}$, changes the average by $\frac{1}{N-1}(\bar{x} - x^{(j)})$.

Using the same derivation we can compute the influence of the per-decision importance sampling estimator (PDIS) and doubly-robust importance sampling estimator (DR):

PDIS For the PDIS estimator, given by

$$\hat{v}^{\pi_e}_{PDIS} = \frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T-1} w^{(n)}_{0:t}\gamma^t r^{(n)}_t, \tag{22}$$

the total influence of trajectory j is given by

$$I_j = \frac{1}{N-1}\left(\hat{v}_{PDIS} - \sum_{t=0}^{T-1} w^{(j)}_{0:t}\gamma^t r^{(j)}_t\right). \tag{23}$$

DR For the DR estimator (Jiang & Li, 2015), given by

$$\hat{v}^{\pi_e}_{DR} = \frac{1}{N}\sum_{n=1}^{N}\left(\sum_{t=0}^{T-1}\gamma^t\left(w^{(n)}_{0:t} r^{(n)}_t - w^{(n)}_{0:t}\tilde{q}(x^{(n)}_t, a^{(n)}_t) + w^{(n)}_{0:t-1}\tilde{v}(x^{(n)}_t)\right)\right), \tag{24}$$

where $\tilde{v}$ and $\tilde{q}$ are independent estimates of the value function, the total influence of trajectory j is given by

$$I_j = \frac{1}{N-1}\left(\hat{v}_{DR} - \sum_{t=0}^{T-1}\gamma^t\left(w^{(j)}_{0:t} r^{(n)}_t - w^{(j)}_{0:t}\tilde{q}(x^{(j)}_t, a^{(j)}_t) + w^{(j)}_{0:t-1}\tilde{v}(x^{(j)}_t)\right)\right). \tag{25}$$

3.1. Influence of weighted IS estimators

For weighted estimators such as weighted importance sampling (WIS) given by

$$\hat{v}^{\pi_e}_{WIS} = \frac{1}{\sum_{n=1}^{N} w^{(n)}_{0:T}}\sum_{n=1}^{N} w^{(n)}_{0:T} g^{(n)}_T, \tag{26}$$

the influence calculation is slightly different, and requires caching the sum of weights of all trajectories in the data.

$$\begin{aligned} I_j &= \hat{v}_{-j} - \hat{v} \\ &= \frac{1}{\sum_{n\neq j} w^{(n)}_{0:T}}\sum_{n\neq j} w^{(n)}_{0:T} g^{(n)}_T \\ &\quad - \frac{1}{\sum_{n=1}^{N} w^{(n)}_{0:T}}\sum_{n=1}^{N} w^{(n)}_{0:T} g^{(n)}_T \\ &= \sum_{n=1}^{N} w^{(n)}_{0:T} g^{(n)}_T \left(\frac{1}{\sum_{n\neq j} w^{(n)}_{0:T}} - \frac{1}{\sum_{n=1}^{N} w^{(n)}_{0:T}}\right) \\ &\quad - \frac{1}{\sum_{n\neq j} w^{(n)}_{0:T}} w^{(j)}_{0:T} g^{(j)}_T \\ &= \frac{w^{(j)}_{0:T}}{(\sum_{n\neq j} w^{(n)}_{0:T})(\sum_{n=1}^{N} w^{(n)}_{0:T})}\sum_{n=1}^{N} w^{(n)}_{0:T} g^{(n)}_T \\ &\quad - \frac{1}{\sum_{n\neq j} w^{(n)}_{0:T}} w^{(j)}_{0:T} g^{(j)}_T \\ &= \frac{w^{(j)}_{0:T}}{W - w^{(n)}_{0:T}}\left(\hat{v}_{WIS} - g^{(j)}_T\right). \end{aligned} \tag{27}$$

In the last expression, $W = \sum_{n=1}^{N} w^{(n)}_{0:T}$ is the cached sum of weights.

WDR For the weighted doubly robust estimator (WDR) (Thomas & Brunskill, 2016) the influence calculation is conceptually similar, but the fact that the sum of weights which normalizes the estimator is time dependant makes it a little more tedious and requires caching a number of values which scales with the horizon, T. The estimator is given by

$$\hat{v}^{\pi_e}_{WDR} = \sum_{n=1}^{N}\left(\sum_{t=0}^{T-1}\gamma^t\left(\frac{w^{(n)}_{0:t}}{W_t} r^{(n)}_t - \frac{w^{(n)}_{0:t}}{W_t}\tilde{q}(x^{(n)}_t, a^{(n)}_t) + \frac{w^{(n)}_{0:t-1}}{W_{t-1}}\tilde{v}(x^{(n)}_t)\right)\right), \tag{28}$$

where we define $W_t = \sum_{n=1}^{N} w^{(n)}_{0:t}$.

If we switch the order of summation and treat the three terms in the sum independently, we can think of the estimator as being composed of $3T$ terms:

$$\hat{v}^{\pi_e}_{WDR} = \sum_{t=0}^{T-1}\gamma^t\left(\sum_{n=1}^{N}\frac{w^{(n)}_{0:t}}{W_t} r^{(n)}_t - \sum_{n=1}^{N}\frac{w^{(n)}_{0:t}}{W_t}\tilde{q}(x^{(n)}_t, a^{(n)}_t) + \sum_{n=1}^{N}\frac{w^{(n)}_{0:t-1}}{W_{t-1}}\tilde{v}(x^{(n)}_t)\right). \tag{29}$$

For a given t, let's look at the resulting difference in the first term if trajectory j is removed from the data:

$$\begin{aligned}
&\sum_{n\neq j} \frac{w_{0:t}^{(n)}}{(W_t)_{-j}} r_t^{(n)} - \sum_{n=1}^{N} \frac{w_{0:t}^{(n)}}{W_t} r_t^{(n)} \qquad (30)\\
&= \sum_{n=1}^{N} w_{0:t}^{(n)} r_t^{(n)} \left(\frac{1}{(W_t)_{-j}} - \frac{1}{W_t} \right)\\
&- \frac{w_{0:t}^{(j)}}{(W_t)_{-j}} r_t^{(j)}\\
&= \frac{w_{0:t}^{(j)}}{(W_t)_{-j}} \left(\sum_{n=1}^{N} \frac{w_{0:t}^{(n)} r_t^{(n)}}{W_t} - r_t^{(j)} \right)\\
&= \frac{w_{0:t}^{(j)}}{W_t - w_{0:t}^{j}} \left(A_t - r_t^{(j)} \right),
\end{aligned}$$

where we defined $A_t = \sum_{n=1}^{N} \frac{w_{0:t}^{(n)} r_t^{(n)}}{W_t}$.

If we similarly define $B_t = \sum_{n=1}^{N} \frac{w_{0:t}^{(n)} \tilde{q}(x_t^{(n)}, a_t^{(n)})}{W_t}$ and $C_t = \sum_{n=1}^{N} \frac{w_{0:t-1}^{(n)} \tilde{v}(x_t^{(n)})}{W_{t-1}}$ and repeat the calculation in Equation 30 for the second and third terms in Equation 29 (note the time offset in the definition of C_t) we see that the influence of the WDR is given by

$$\begin{aligned}
I_j = \sum_{t=0}^{T-1} \gamma^t \Bigg(& \frac{w_{0:t}^{(j)}}{W_t - w_{0:t}^{j}} \left(A_t - r_t^{(j)} \right) \qquad (31)\\
& - \frac{w_{0:t}^{(j)}}{W_t - w_{0:t}^{j}} \left(B_t - \tilde{q}(x_t^{(j)}, a_t^{(j)}) \right)\\
& + \frac{w_{0:t-1}^{(j)}}{W_{t-1} - w_{0:t-1}^{j}} \left(C_t - \tilde{v}(x_t^{(j)}) \right) \Bigg).
\end{aligned}$$

.

4. Preprocessing and experimental details for MIMIC-III acute hypotension dataset

In this section, we describe the preprocessing we performed on the raw MIMIC-III database to convert it into a dataset amenable to modeling with RL. This preprocessing procedure was done in close consultation with the intensivist collaborator on our team.

4.1. Cohort Selection

We use MIMIC-III v1.4 (Johnson et al., 2016), which contains information from about 60,000 intensive care unit (ICU) admissions to Beth Israel Deaconess Medical Center. We filter the initial database on the following features: admissions where data was collected using the Metavision clinical information system; admissions to a medical ICU (MICU); adults (age $\geq$ 18 years); initial ICU stays for hospital admissions with multiple ICU stays; ICU stays with a total length of stay of at least 24 hours; and ICU stays where there are 7 or more mean arterial pressure (MAP) values of 65mmHg or less, indicating probable acute hypotension. For long ICU stays, we limit to only using information captured during the inital 48 hours after admission, as our intensivist advised that care for hypotension during later periods of an ICU stay often look very different. After this filtering, we have a final cohort consisting of 1733 distinct ICU admissions. For computational convenience, we further down-sample this cohort, and use 20% (346) ICU stays to use to learn a policy, and another 20% (346) ICU stays to evaluate the policy via FQE and our proposed influence analysis.

4.2. Clinical Variables Considered

Given our final cohort of patients admitted to the ICU, we next discuss the different clinical variables that we extract that are relevant to our task of acute hypotension management.

The two first-line treatments are intravenous (IV) fluid bolus therapy, and vasopressor therapy. We construct fluid bolus variables in the following way:

1. We filter all fluid administration events to only include NaCl 0.9%, lactated ringers, or blood transfusions (packed red blood cells, fresh frozen plasma, or platelets).
2. Since a fluid bolus should be a nontrivial amount of fluid administered over a brief period of time, we further filter to only fluid administrations with a volume of at least 250mL and over a period of 60 minutes or shorter.

Each fluid bolus has an associated volume, and a starting time (since a bolus is given quickly / near-instantaneously, we ignore the end-time of the administration). To construct vasopressors, we first normalize vasopressor infusion rates across different drug types as follows, using the same normalization as in Komorowski et al. (2018):

1. Norepinephrine: this is our "base" drug, as it's the most commonly administered. We will normalize all other drugs in terms of this drug. Units for vasopressor rates are in mcg per kg body weight per minute for all drugs except vasopressin.
2. Vasopressin: the original units are in units/min. We first clip any values above 0.2 units/min, and then multiply the final rates by 5.
3. Phenylephrine: we multiply the original rate by 0.45.

4. Dopamine: we multiply the original rate by 0.01.

5. Epinephrine: this drug is on the same scale as norepinephrine and is not rescaled.

As vasopressors are given as a continuous infusion, they consist of both a treatment start time and stop time, as well as potentially many times in the middle where the rates are changed. More than a single vasopressor may be administered at once, as well.

We also use 11 other clinical variables as part of the state space in our application: serum creatinine, FiO_2, lactate, urine output, ALT, AST, diastolic/systolic blood pressure, mean arterial pressure (MAP; the main blood pressure variable of interest), PO_2, and the Glasgow Coma Score (GCS).

4.3. Selecting Action Times

Given a final cohort, clinical variables, and treatment variables, we still must determine how to discretize time and choose at which specific time points actions should be chosen. To arrive at a final set of "action" times for a specific ICU stay, we use the following heuristic-based algorithm:

1. We start by including all times a treatment is started, stopped, or modified.

2. Next, we remove consecutive treatment times if there are no MAP measurements between treatments. We do this because without at least one MAP measurement in between treatments, we would not be able to assess what effect the treatment had on blood pressure. This leaves us with a set of time points when treatments were started or modified.

3. At many time points, the clinician consciously chooses not to take an action. Unfortunately, this information is not generally recorded (although, on occasion, may exist in clinical notes). As a proxy, we consecutively add to our existing set of "action times" any time point at which an abnormally low MAP is observed ($<$ 60mmHg) and there are no other "action times" within a 1 hour window either before or after. This captures the relatively fine-granularity with which a physician may choose not to treat despite some degree of hypotension.

4. Last, we add additional time points to fill in any large gaps where no "action times" exist. We do this by adding time points between existing "action times" until there are no longer any gaps greater than 4 hours between actions. This makes some clinical sense, as patients in the ICU are being monitored relatively closely, but if they are more stable, their treatment decisions will be made on a coarser time scale.

Now that we have a set of action times for each trajectory, we can count up the total number of transitions in our training and evaluation datasets (both of which consist of 346 trajectories). The training trajectories contain a total of 6777 transitions, while there are 6863 total transitions in the evaluation data. Trajectories vary in length from a minimum of 7 transitions to a maximum of 49, with 16, 18, and 23 transitions comprising the 25%, 50%, and 75% quantiles, respectively.

4.4. Action Space Construction

Given treatment timings, doses, and manually identified "action times" at which we want to assess what type of clinical decision was made, we can now construct our action space. We choose to operate in a discrete action space, which means we need to decide how to bin each of the continuous-valued treatment amounts.

Binning of IV fluids is more natural and easier, as fluid boluses are generally given in discrete amounts. The most common bolus sizes are 500mL and 1000mL, so we bin fluid bolus volumes into the following 4 bins, which correspond to "none"/"low"/"medium"/"high" (in mL): $\{0, [250, 500), [500, 1000), [1000, \infty]\}$, although in practice very few boluses of more than 2L are ever given. Given this binning scheme, we can simply add up the total amount of fluids administered during any adjacent action times to determine which discrete fluid amount we should code the action as.

Binning of vasopressors is slightly more complex. These drugs are dosed at a specific rate, and there may be many rate changes made between action times, or sometimes there are several vasopressors being given at once. We chose to first add up the cumulative amount of (normalized) vasopressor drug administered between action times, and then normalize this amount by the size of the time window between action times to account for the irregular spacing. Finally, we also bin vasopressors into 4 discrete bins corresponding to "none"/"low"/"medium"/"high" amounts: $\{0, (0, 8.1), [8.1, 21.58), [21.58, \infty]\}$. The relevant units here are total mcg of drug given each hour, per kg body weight. Since the distribution of values for vasopressors is not as naturally discrete, we chose our bin sizes using the 33.3% and 66.7% quantiles of dose amounts.

In the end, we have an action space with 16 possible discrete actions, considering all combinations of each of the 4 vasopressor amounts and fluid bolus amounts.

4.5. State Construction

Given a patient cohort, decision/action times, and discrete actions, we are now ready to construct a state space. For simplicity in this initial work, we first start with the 11 clin-

ical time series variables previously listed. If a variable is never measured, we use the population median as a placeholder. If a variable has been measured before, we use the most recent measurement. The sole exception to this is the 3 blood pressure variables. For the blood pressures, we instead use the minimum (or worst) value observed since the last action.

We add to these a number of indicator variables that denote whether a particular variable was recently measured or not. Due to the strongly missing-not-at-random nature of clinical time series, there is often considerable signal in knowing that certain types of measurements were recently taken, irrespective of the measurement values (Agniel et al., 2018). We choose to construct indicator variables denoting whether or not a urine output was taken since the last action time, and whether a GCS was recorded since the last action. We also include state features denoting whether the following labs/vitals were *ever* ordered: creatinine, FiO_2, lactate, ALT, AST, PO_2. We do not include these indicators for all 11 clinical variables, as most of the vitals are recorded at least once an hour, and sometimes even more frequently. In total, 8 indicators comprise part of our state space.

Last, we include 10 additional variables that summarize past treatments administered, if any. We first include 6 indicator variables (3 for each treatment type) denoting which dose of fluid and vasopressor, if any, was chosen at the last action time. Last, for each treatment type we include two final features that summarize past actual amounts of treatments administered (the total amount of this treatment administered up until the current time, and the total amount of this treatment administered within the last 8 actions.

In total, our final state space has 29 dimensions. In future work we plan to explore richer state representations.

4.6. Reward Function Construction

In this preliminary work, we use a simple reward that is a piecewise linear function of the MAP in the next state. In particular, the reward takes on a value of -1 at 40mmHg, the lowest attainable MAP in the data. It increases linearly to -0.15 at 55mmHg, linearly from there to -0.05 at 60mmHg, and achieves a maximum value of 0 at 65mmHg, a commonly used target for blood pressure in the ICU (Asfar et al., 2014). However, if a patient has a urine output of 30mL/hour or higher, then any MAP values of 55mmHg or higher are reset to 0. This attempts to mimic the fact that a clinician will not be too concerned if a patient is slightly hypotensive but otherwise stable, since a modest urine output indicates that the modest hypotension is not a real problem.

4.7. Choice of Kernel Function

In order to use kernel-based FQE, we need to define a kernel that defines similarity between states. In consultation with our intensivist collaborator, we chose a simple weighted Euclidean distance, where each state variable receives a different weight based on its estimated importance to the clinical problem. We show all weights in Table 1.

Since technically we need a kernel over both all possible states and actions for FQE and influence analysis, we augment our kernel with extremely large weights so that effectively the kernel only compares pairs (s, a) and (s', a') for $a = a'$. Other choices should be made for continuous action spaces.

4.8. Hyperparameters

We use the training set of 6777 trajectories to learn a policy to then evaluate using FQE and influence analysis. In particular, we learn a deterministic policy by taking the most common action within the 50 nearest neighbors of a given state, with respect to the kernel in Table 1. We use a discount of $\gamma = 1$ so that all time steps are treated equally, and use a neighborhood radius of 7 for finding nearest neighbors in FQE. Lastly, for the influence analysis, we use a threshold of 0.05, or 5%, so that transitions which will affect the FQE value estimate by more than 5% are flagged for expert review.

5. Additional Results from MIMIC-III acute hypotension dataset

In the main body of the paper, we showed two qualitative results figures showing 2 of the 6 highly influential transitions flagged by influence analysis. In this section, we show the remaining 4 influential transitions.

Table 1. Weights for each state variable in our kernel function.

State Variable	Kernel Weight
Creatinine	3
FiO2	15
Lactate	10
Urine Output	15
Urine Output since last action?	5
ALT	5
AST	5
Diastolic BP	5
MAP	15
PO2	3
Systolic BP	5
GCS	15
GCS since last action?	5
Creatinine ever taken?	3
FiO2 ever taken?	15
Lactate ever taken?	10
ALT ever taken?	5
AST ever taken?	5
PO2 ever taken?	3
Low vasopressor done last time?	15
Medium vasopressor done last time?	15
High vasopressor done last time?	15
Low fluid done last time?	15
Medium fluid done last time?	15
High fluid done last time?	15
Total vasopressors so far	15
Total fluids so far	15
Total vasopressors last 8 actions	15
Total fluids last 8 actions	15

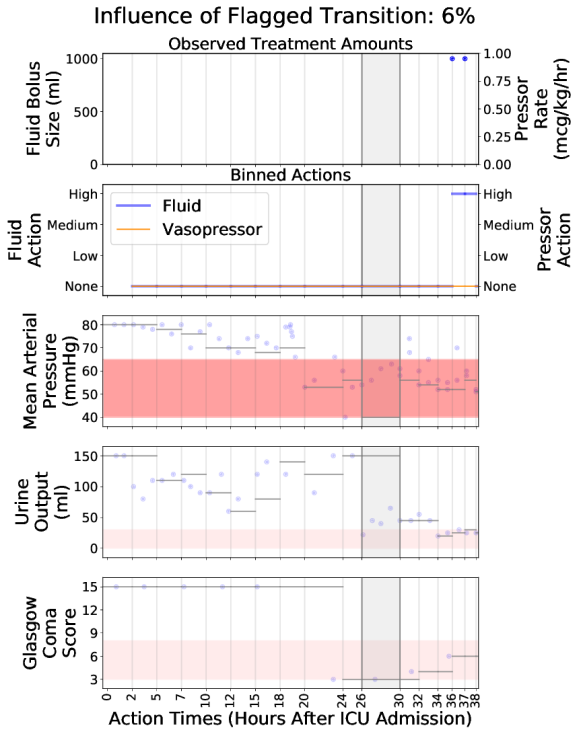

Figure 1. An additional example identified by our influence analysis as having an especially high effect on the OPE value estimate.

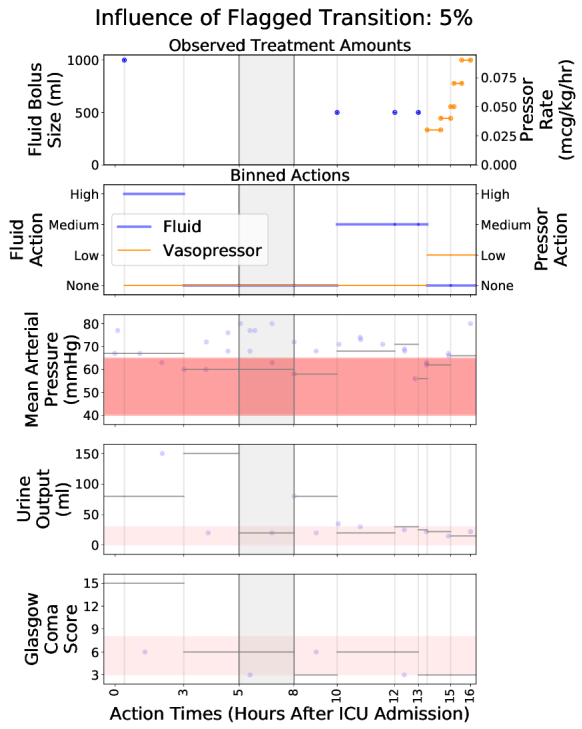

Figure 2. An additional example identified by our influence analysis as having an especially high effect on the OPE value estimate. Note that this transition is from the same trajectory as the influential transition highlighted in Figure 3

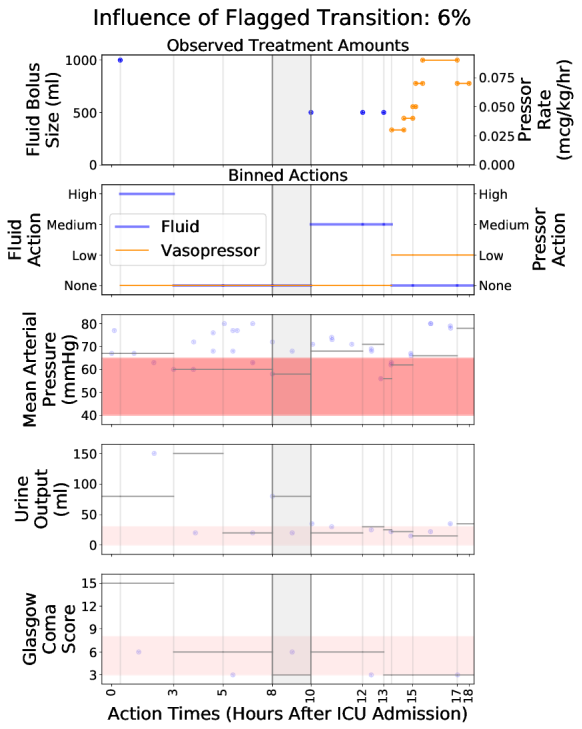

Figure 3. An additional example identified by our influence analysis as having an especially high effect on the OPE value estimate. Note that this transition is from the same trajectory as the influential transition highlighted in Figure 2

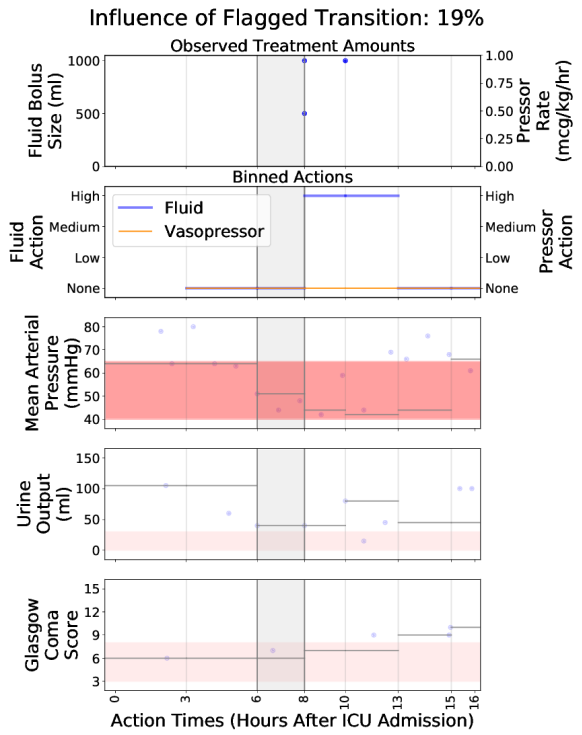

Figure 4. An additional example identified by our influence analysis as having an especially high effect on the OPE value estimate.

References

Agniel, D., Kohane, I. S., and Weber, G. M. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *Bmj*, 361:k1479, 2018.

Asfar, P., Meziani, F., Hamel, J.-F., Grelon, F., Megarbane, B., Anguel, N., Mira, J.-P., Dequin, P.-F., Gergaud, S., Weiss, N., et al. High versus low blood-pressure target in patients with septic shock. *N Engl J Med*, 370:1583–1593, 2014.

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

Johnson, A. E., Pollard, T. J., Shen, L., Li-wei, H. L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., and Mark, R. G. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C., and Faisal, A. A. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716–1720, 2018.

Thomas, P. and Brunskill, E. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 2139–2148, 2016.