Biometrika (2015), **99**, 1, *pp*. 1–15
© 2015 Biometrika Trust
Printed in Great Britain

Advance Access publication on 31 July 2015

Supplemental Material for Constructing Dynamic Treatment Regimes in Infinite-Horizon Settings

BY ASHKAN ERTEFAIE AND ROBERT L. STRAWDERMAN

Department of Biostatistics and Computational Biology, University of Rochester
265 Crittenden Boulevard, Rochester, New York, 14642, U.S.A.

ashkan_ertefaie@urmc.rochester.edu robert_strawderman@urmc.rochester.edu

S1. THE STOCHASTIC MINIMIZATION ALGORITHM

We base our minimization procedure on the approximate gradient descent approach, in which sub-gradients are defined as Frechet sub-gradients of the objective function $M(\theta)$. Following Maei et al. (2010) and under assumptions A.1-8, the algorithm converges to the minimizer of our objective function.

The sub-gradient $\partial M(\theta)$ of $M(\theta)$ with respect to θ is

$$\partial M(\theta) = -E\left\{\sum_t \delta_{t+1}(\theta)\varphi(s_t,a_t)^\top\right\} + \gamma E\left\{\sum_t \varphi(s_{t+1},\pi^*(s_{t+1}))\varphi(s_t,a_t)^\top\right\}\varpi,$$

where $\varpi = E\{\sum_t \varphi(s_t,a_t)\varphi(s_t,a_t)^\top\}^{-1}E\{\sum_t \delta_{t+1}(\theta)\varphi(s_t,a_t)^\top\}$. Using the weight-doubling trick introduced by Sutton et al. (2009), we summarize the steps toward minimizing the objective function $M(\theta)$ as follows:

1. Set initial values for the p dimensional vectors of θ and ϖ. Using grid search, the initial value θ_1 can be set as the one that minimizes the objective function and the initial value $w_1 = \mathbb{P}_n\{\sum_t \varphi(s_t,a_t)\varphi(s_t,a_t)^\top\}^{-1}\mathbb{P}_n\{\sum_t \delta_{t+1}(\theta_0)\varphi(s_t,a_t)^\top\}$.
2. Start from the first individual's trajectory and obtain θ_{k+1} from the following iterative equations:

$$\theta_{k+1} = \theta_k + \alpha_k\nu\sum_t\left[\delta_{t+1}(\theta_k)\varphi(s_t,a_t) - \gamma\{\varpi_k^\top\varphi(s_t,a_t)\}\varphi\{s_{t+1},\pi^*_{\theta_k}(s_{t+1})\}^\top\right] \quad (1)$$

$$\varpi_{k+1} = \varpi_k + \beta_k\nu\sum_t\left[\delta_{t+1}(\theta_k) - \{\varphi(s_t,a_t)^\top\varpi_t\}^\top\right]\varphi(s_t,a_t)^\top, \quad (2)$$

   where α_k, β_k and ν are tuning parameters (step sizes) and $\pi^*_{\theta_k}(.)$ is the optimal policy estimated as a function of θ_k.
3. Use step 2 to continue updating the parameters to the last individual.
4. Continue steps 2 and 3 until $\|\theta_{k+1} - \theta_k\|_2 < c$ where c is a constant.

The tuning parameters, i.e., step sizes, α_k and β_k need to satisfy assumptions P.1-P.4 in Section S7. The parameter ν tunes the step sizes and lies in the interval (0,1). Our simulation studies show that the best choice of ν would be close to $\frac{1}{T}$, where T is the maximal length of the trajectories in our data. See our discussion in Section S5.

S2. LEMMA

LEMMA S1. *Let $\{a_i\}_1^K$ and $\{b_i\}_1^K$ be two sets of elements, then*

I. $\lim_{\|b\|\to 0} \frac{\max_{1\le i\le K}(a_i+b_i)-\max_{i\in\pi^*}(a_i+b_i)}{\|b\|} = 0$*;*

II. $\max_{1\le i\le K}(a_i+b_i) - \max_{i\in\pi^*}(a_i+b_i)$ *is non-negative and bounded above by*

$$\max_{i\notin\pi^*} b_i - \max_{i\in\pi^*} b_i \le \max_{1\le i\le K} b_i - \max_{i\in\pi^*} b_i$$

where $\pi^* = \arg\max_{1\le i\le K} a_i$.

Proof. Part I. Since the set π^* is a subset of $1 \le i \le K$, we have

$$\begin{aligned}
0 &\le \frac{\max_{1\le i\le K}(a_i+b_i) - \max_{i\in\pi^*}(a_i+b_i)}{\|b\|} \\
&= \max_{1\le i\le K}\left(a_i - a_{i^*} + b_i - \max_{i\in\pi^*} b_i\right)\frac{1}{\|b\|}, \qquad \forall i^*\in\pi^* \\
&= \max\left\{\max_{i\in\pi^*}(b_i - \max_{i\in\pi^*} b_i), \max_{i\notin\pi^*}(a_i - a_{i^*} + b_i - \max_{i\in\pi^*} b_i)\right\}\frac{1}{\|b\|} \\
&= \max\left\{0, \max_{i\notin\pi^*}\left(\frac{a_i - a_{i^*}}{\|b\|} + \frac{b_i - \max_{i\in\pi^*} b_i}{\|b\|}\right)\right\} \\
&\le \max\left\{0, \max_{i\notin\pi^*}\left(\frac{a_i - a_{i^*}}{\|b\|} + \frac{2|\max_{1\le i\le K} b_i|}{\|b\|}\right)\right\}
\end{aligned}$$

Since $\frac{a_i - a_{i^*}}{\|b\|} \to -\infty$ as $\|b\| \to 0$ and $\frac{|\max_{i\in\pi} b_i|}{\|b\|} \le 1$, part I is proved. Part II can be proved similarly.

S3. PROOF OF THEOREM 1

We first show that the objective function $M(\theta)$ is continuous. Then using the results of Lemma S1, we show that $M(\theta)$ satisfies the two required conditions of Theorem 3.2.1 in Van Der Vaart and Wellner (1996), which completes the proof of consistency. To prove the asymptotic normality, under the additional assumption $A.8$, we define a function $V(b)$ such that $\sqrt{n}(\hat{\theta} - \theta_0) \to_d \arg\min_b V(b)$, where $\arg\min_b V(b)$ is normally distributed.

First, we show that the function $M(\theta)$ is continuous by proving the continuity of $D(\theta)$ around $\theta = \theta_0$. Since

$$\begin{aligned}
\|D(\theta) - D(\theta_0)\| = \|\sum_{t=0}^{T-1} E\Big[\{&\max_{a\in\mathcal{A}_{S_{t+1}}}\{\theta^\top\varphi(S_{t+1},a)\} - \max_{a\in\mathcal{A}_{S_{t+1}}}\{\theta_0^\top\varphi(S_{t+1},a)\} \\
&-\theta^\top\varphi(S_t,A_t) + \theta_0^\top\varphi(S_t,A_t)\}\varphi(S_t,A_t)^\top\Big]\|,
\end{aligned}$$

by the Cauchy-Schwartz inequality and the fact that $|\max_a f(a) - \max_a g(a)| \le \max_a |f(a) - g(a)|$, we have

$$E\left[\max_{a\in\mathcal{A}_{S_{t+1}}}\{\theta^\top\varphi(S_{t+1},a)\} - \max_{a\in\mathcal{A}_{S_{t+1}}}\{\theta_0^\top\varphi(S_{t+1},a)\}\varphi(S_t,A_t)^\top\right] \le ||\theta-\theta_0||E\left[\sum_a \|\varphi(S_{t+1},a)\|\|\varphi(S_t,A_t)\|\right],$$

which under assumption $A.4$ implies the continuity of $D(\theta)$ around $\theta = \theta_0$.

Part I (Consistency). We show that $M(\theta)$ satisfies the two conditions listed in Theorem 3.2.1 Van Der Vaart and Wellner (1996). For the first condition, we need to show that for some $\epsilon > 0$ and $c > 0$ with $||\theta - \theta_0|| < \epsilon$,

$$D(\theta_0)W^{-1}D(\theta_0)^\top - D(\theta)W^{-1}D(\theta)^\top \leq -c||\theta - \theta_0||^2,$$

and since $D(\theta_0) = 0$,

$$D(\theta)W^{-1}D(\theta)^\top \geq c||\theta - \theta_0||^2. \tag{3}$$

The left hand side of the above inequality can be written as

$$\begin{aligned} D(\theta)W^{-1}D(\theta)^\top &= \{D(\theta) - D(\theta_0) - \dot{D}_{\theta_0}(\theta - \theta_0)\}W^{-1}\{D(\theta) - D(\theta_0) - \dot{D}_{\theta_0}(\theta - \theta_0)\}t^\top \\ &\quad + \dot{D}_{\theta_0}(\theta - \theta_0)W^{-1}\dot{D}_{\theta_0}(\theta - \theta_0)^\top + 2\dot{D}_{\theta_0}(\theta - \theta_0)W^{-1}\{D(\theta) - D(\theta_0) - \dot{D}_{\theta_0}(\theta - \theta_0)\}^\top, \end{aligned}$$

where

$$\dot{D}_{\theta_0}(b) = E\left[\sum_{t=0}^{T-1}\{\gamma \max_{a\in\pi^*(S_{t+1})} b^\top\varphi(S_{t+1}, a) - b^\top\varphi(S_t, A_t)\}\varphi(S_t, A_t)^\top\right],$$

and $\pi^*(S_{t+1}) = \arg\max_{a\in\mathcal{A}_{S_{t+1}}} \theta_0^\top\varphi(S_{t+1}, a)$. Note that $\pi^*(S_{t+1})$ may be a set of actions. Now, we show that $\|D(\theta) - D(\theta_0) - \dot{D}_{\theta_0}(\theta - \theta_0)\| = o(\|\theta - \theta_0\|)$.

$$\begin{aligned} &\frac{\|D(\theta) - D(\theta_0) - \dot{D}_{\theta_0}(\theta - \theta_0)\|}{\|\theta - \theta_0\|} = \\ &\frac{\gamma}{||\theta - \theta_0||}\left\|E\left[\sum_{t=0}^{T-1}\left\{\max_{a\in\mathcal{A}_{S_{t+1}}}(\theta_0 + b||\theta - \theta_0||)^\top\varphi(S_{t+1}, a) - \max_{a\in\pi^*(S_{t+1})}\theta_0^\top\varphi(S_{t+1}, a)\right.\right.\right. \\ &\qquad\qquad \left.\left.\left. - \max_{a\in\pi^*(S_{t+1})} b^\top\varphi(S_{t+1}, a)\|\theta - \theta_0\|\right\}\varphi(S_t, A_t)^\top\right]\right\|, \end{aligned}$$

where $b = (\theta - \theta_0)/\|\theta - \theta_0\|$. Then, since, for any $a, a' \in \pi^*(S_{t+1})$, we have $\theta_0^\top\varphi(S_{t+1}, a) = \theta_0^\top\varphi(S_{t+1}, a')$. Thus the following equality holds:

$$\max_{a\in\pi^*(S_{t+1})}\theta_0^\top\varphi(S_{t+1}, a) + \max_{a\in\pi^*(S_{t+1})}(\theta - \theta_0)^\top\varphi(S_{t+1}, a) = \max_{a\in\pi^*(S_{t+1})}\theta^\top\varphi(S_{t+1}, a).$$

Thus,

$$\frac{||D(\theta)-D(\theta_0)-\dot{D}_{\theta_0}(\theta-\theta_0)||}{||\theta-\theta_0||}=$$

$$\frac{\gamma}{||\theta-\theta_0||}\left\|E\left[\sum_{t=0}^{T-1}\left\{\max_{a\in\mathcal{A}_{S_{t+1}}}(\theta_0+b||\theta-\theta_0||)^\top\varphi(S_{t+1},a)-\max_{a\in\pi^*(S_{t+1})}(\theta_0+b\|\theta-\theta_0\|)^\top\varphi(S_{t+1},a)\right\}\varphi(S_t,A_t)^\top\right]\right\|$$

$$\leq\gamma E\left[\sum_{t=0}^{T-1}\|\varphi(S_{t+1},a)\|\left\{\max_{a\in\mathcal{A}_{S_{t+1}}}\left(\frac{\theta_0}{||\theta-\theta_0||}+b\right)^\top\varphi(S_{t+1},a)-\max_{a\in\pi^*(S_{t+1})}\left(\frac{\theta_0}{\|\theta-\theta_0\|}+b\right)^\top\varphi(S_{t+1},a)\right\}\right]$$

$$\leq\gamma E\left[\sum_{t=0}^{T-1}\|\varphi(S_t,A_t)\|\left\{\max_{a\in\mathcal{A}_{S_{t+1}}}b^\top\varphi(S_{t+1},a)-\max_{a\in\pi^*(S_{t+1})}b^\top\varphi(S_{t+1},a)\right\}\right]$$

$$\leq\gamma E\left[\sum_{t=0}^{T-1}\|\varphi(S_t,A_t)\|\sum_{a\in\mathcal{A}_{S_{t+1}}}\|b\|\|\varphi(S_{t+1},a)\|\right]<\infty.$$

The second inequality follows from Lemma S1 part (II) and the last inequality follows from $\|b\|=1$ and assumption A.5. Also, using Lemma S1 part (I), we have

$$\lim_{\|\theta-\theta_0\|\to 0}\left\{\max_{a\in\mathcal{A}_{S_{t+1}}}\left(\frac{\theta_0}{||\theta-\theta_0||}+b\right)^\top\varphi(S_{t+1},a))-\max_{a\in\pi^*(S_{t+1})}\left(\frac{\theta_0}{\|\theta-\theta_0\|}+b\right)^\top\varphi(S_{t+1},a)\right\}=0.$$

We just showed that $\|D(\theta)-D(\theta_0)-\dot{D}_{\theta_0}(\theta-\theta_0)\|=o(\|\theta-\theta_0\|)$. Since W^{-1} is of full rank matrix,

$$\{D(\theta)-D(\theta_0)-\dot{D}_{\theta_0}(\theta-\theta_0)\}W^{-1}\{D(\theta)-D(\theta_0)-\dot{D}_{\theta_0}(\theta-\theta_0)\}^\top=o(\|\theta-\theta_0\|^2).$$

Now, we need to show that $\dot{D}_{\theta_0}(\theta-\theta_0)\dot{D}_{\theta_0}(\theta-\theta_0)^\top\geq c''\|\theta-\theta_0\|^2$. By definition,

$$\begin{aligned}\dot{D}_{\theta_0}(\theta-\theta_0)&=E\left[\sum_{t=0}^{T-1}\left\{\gamma\max_{a\in\pi^*(S_{t+1})}(\theta-\theta_0)^\top\varphi(S_{t+1},a)-(\theta-\theta_0)^\top\varphi(S_t,A_t)\right\}\varphi(S_t,A_t)^\top\right]\\&=(\theta-\theta_0)^\top E\left[\sum_{t=0}^{T-1}\left\{\gamma I_{|\pi^*(S_{t+1})|=1}\varphi(S_{t+1},\pi^*(S_{t+1}))-\varphi(S_t,A_t)\right\}\varphi(S_t,A_t)^\top\right]\\&\quad+\gamma E\left[\sum_{t=0}^{T-1}\left\{\gamma I_{|\pi^*(S_{t+1})|>1}\max_{a\in\pi^*(S_{t+1})}\frac{(\theta-\theta_0)^\top}{\|\theta-\theta_0\|}\varphi(S_{t+1},a)\right\}\varphi(S_t,A_t)^\top\right]\|\theta-\theta_0\|.\end{aligned}$$

Let

$$M_1=E\left[\sum_{t=0}^{T-1}\left\{\gamma I_{|\pi^*(S_{t+1})|=1}\varphi(S_{t+1},\pi^*(S_{t+1}))-\varphi(S_t,A_t)\right\}\varphi(S_t,A_t)^\top\right],$$

$$M_2=E\left[\sum_{t=0}^{T-1}\left\{I_{|\pi^*(S_{t+1})|>1}\max_{a\in\pi^*(S_{t+1})}\frac{(\theta-\theta_0)^\top}{\|\theta-\theta_0\|}\varphi(S_{t+1},a)\right\}\varphi(S_t,A_t)^\top\right].$$

Then

$$\dot{D}_{\theta_0}(\theta-\theta_0)\dot{D}_{\theta_0}(\theta-\theta_0)^\top=(\theta-\theta_0)^\top M_1M_1^\top(\theta-\theta_0)+2\gamma(\theta-\theta_0)^\top M_1M_2^\top\|\theta-\theta_0\|+\gamma^2\|\theta-\theta_0\|^2M_2M_2^\top.$$

Assuming that M_1 is of full rank, i.e., Assumption A.7, we have

$$(\theta-\theta_0)^\top M_1 M_1^\top(\theta-\theta_0) \geq \|\theta-\theta_0\|^2\lambda_{min},$$

where λ_{min} is the smallest eigenvalue of $M_1^\top M_1$. Also, using singular value decomposition we have

$$(\theta-\theta_0)^\top M_1 M_2^\top\|\theta-\theta_0\| \leq \|\theta-\theta_0\|^2\sqrt{\lambda_{max}}\|M_2\|,$$

where λ_{max} is a maximum eigenvalue of $M_1 M_1^\top$. Thus

$$\dot{D}_{\theta_0}(\theta-\theta_0)\dot{D}_{\theta_0}(\theta-\theta_0)^\top \geq \|\theta-\theta_0\|^2\left[\lambda_{min}-2\gamma\sqrt{\lambda_{max}}\|M_2\|-\gamma^2\|M_2\|^2\right].$$

Therefore, function $M(.)$ satisfies the first condition of Theorem 3.2.5 in Van Der Vaart and Wellner (1996) for any small enough γ such that $\left(\lambda_{min}-2\gamma\sqrt{\lambda_{max}}\|M_2\|-\gamma^2\|M_2\|^2\right)>0$. Note that under assumption $A.8$ when $|\pi^*(S_{t+1})|=1$, the latter condition is satisfied automatically. Since W is of full rank (Assumption $A.6$), $\dot{D}_{\theta_0}(\theta-\theta_0)W^{-1}\dot{D}_{\theta_0}(\theta-\theta_0)^\top \geq c\|\theta-\theta_0\|^2$ for $c>0$.

For the second condition, we need to show that for every large enough n, sufficiently small δ_n and $c>0$

$$E\sup_{\|\theta-\theta_0\|\leq\delta_n}\left|\{\hat{M}(\theta)-M(\theta)\}-\{\hat{M}(\theta_0)-M(\theta_0)\}\right| \leq c\delta_n^2.$$

Since by definition $\hat{D}(\hat{\theta})=D(\theta_0)=0$, we have

$$\begin{aligned}\left|\{\hat{M}(\theta)-M(\theta)\}-\{\hat{M}(\theta_0)-M(\theta_0)\}\right| = \Big|&\{\hat{D}(\theta)-\hat{D}(\hat{\theta})\}\hat{W}^{-1}\{\hat{D}(\theta)-\hat{D}(\hat{\theta})\}^\top\\ &-\{\hat{D}(\theta_0)-\hat{D}(\hat{\theta})\}\hat{W}^{-1}\{\hat{D}(\theta_0)-\hat{D}(\hat{\theta})\}^\top\\ &-\{D(\theta)-D(\theta_0)\}\hat{W}^{-1}\{D(\theta)-D(\theta_0)\}^\top\Big|.\end{aligned}$$

We show that for every large n such that $\|\theta-\hat{\theta}\|\leq\delta_n$ and $\|\theta_0-\hat{\theta}\|\leq\delta_n$

$$\begin{aligned}&E\sup_{\|\theta-\hat{\theta}\|\leq\delta_n}[\{\hat{D}(\theta)-\hat{D}(\hat{\theta})\}\hat{W}^{-1}\{\hat{D}(\theta)-\hat{D}(\hat{\theta})\}^\top] \leq c_1\delta_n^2\\ &E\sup_{\|\hat{\theta}-\theta_0\|\leq\delta_n}[\{\hat{D}(\theta_0)-\hat{D}(\hat{\theta})\}\hat{W}^{-1}\{\hat{D}(\theta_0)-\hat{D}(\hat{\theta})\}^\top] \leq c_2\delta_n^2\\ &E\sup_{\|\theta-\theta_0\|\leq\delta_n}[\{D(\theta)-D(\theta_0)\}^\top\hat{W}^{-1}\{D(\theta)-D(\theta_0)\}^\top] \leq c_3\delta_n^2,\end{aligned}$$

where c_1, c_2 and c_3 are positive constants. Here we show the first inequality and the rest can be shown similarly. By the Cauchy-Schwartz inequality and the fact that $|\max_a f(a)-\max_a g(a)| \leq \max_a|f(a)-g(a)|$, we have

$$\begin{aligned}|\hat{D}(\theta)-\hat{D}(\hat{\theta})| &= \left|\mathbb{P}_n\left[\sum_{t=0}^{T-1}\left\{\max_{a\in\mathcal{A}_{S_{t+1}}}\theta^\top\varphi(S_{t+1},a)-\max_{a\in\mathcal{A}_{S_{t+1}}}\hat{\theta}^\top\varphi(S_{t+1},a)-(\theta-\hat{\theta})^\top\varphi(S_t,A_t)\right\}\varphi(S_t,A_t)^\top\right]\right|\\ &\leq \mathbb{P}_n\left[\sum_{t=0}^{T-1}\|\varphi(S_t,A_t)\|\left\{\sum_{a\in\mathcal{A}_{S_{t+1}}}\|\varphi(S_{t+1},a)\|\|\theta-\hat{\theta}\|+\|\varphi(S_t,A_t)\|\|\theta-\hat{\theta}\|\right\}\right].\end{aligned}$$

Thus,

$$|\hat{D}(\theta) - \hat{D}(\hat{\theta})| \leq m(S, A)\|\theta - \hat{\theta}\|,$$

where

$$m(S, A) = \mathbb{P}_n \left[\sum_{t=0}^{T-1} \|\varphi(S_t, A_t)\| \left\{ \sum_{a \in \mathcal{A}_{S_{t+1}}} \|\varphi(S_{t+1}, a)\| + \|\varphi(S_t, A_t)\| \right\} \right].$$

Therefore

$$E \sup_{\|\theta - \hat{\theta}\| \leq \delta_n} [\{\hat{D}(\theta) - \hat{D}(\hat{\theta})\}^\top \hat{W}^{-1} \{\hat{D}(\theta) - \hat{D}(\hat{\theta})\}] \leq c_1 \delta_n^2,$$

where $c_1 = E\{m(S, A)^2 \|\hat{W}^{-1}\|\}$. Define $c = c_1 + c_2 + c_3$. This shows that our objective function satisfies the second condition of Theorem 3.2.5 in Van Der Vaart and Wellner (1996) as well. This completes the proof of consistency.

Part II (Asymptotic Normality). Let $\theta = \theta_0 + \frac{b}{\sqrt{n}}$ and

$$\hat{V}(b) = n\hat{D}(\theta_0 + b/\sqrt{n})\hat{W}^{-1}\hat{D}(\theta_0 + b/\sqrt{n})^\top - n\hat{D}(\theta_0)\hat{W}^{-1}\hat{D}(\theta_0)^\top.$$

Then,

$$\begin{aligned}
\hat{D}(\theta_0 + b/\sqrt{n}) &= \mathbb{P}_n \Bigg[\sum_t \Bigg\{ R_{t+1} + \gamma \max_{a \in \mathcal{A}_{S_{t+1}}} \theta_0^\top \varphi(S_{t+1}, a) - \theta_0^\top \varphi(S_t, A_t) - b^\top/\sqrt{n}\varphi(S_t, A_t) \\
&\qquad - \gamma \max_{a \in \mathcal{A}_{S_{t+1}}} \theta_0^\top \varphi(S_{t+1}, a) + \gamma \max_{a \in \mathcal{A}_{S_{t+1}}} (\theta_0 + b/\sqrt{n})^\top \varphi(S_{t+1}, a) \Bigg\} \varphi(S_t, A_t)^\top \Bigg] \\
&= \mathbb{P}_n \Bigg[\sum_t \Bigg\{ \delta_{t+1} - b^\top/\sqrt{n}\varphi(S_t, A_t) - \gamma \max_{a \in \mathcal{A}_{S_{t+1}}} \theta_0^\top \varphi(S_{t+1}, a) \\
&\qquad + \gamma \max_{a \in \mathcal{A}_{S_{t+1}}} (\theta_0 + b/\sqrt{n})^\top \varphi(S_{t+1}, a) \Bigg\} \varphi(S_t, A_t)^\top \Bigg].
\end{aligned}$$

Using the above equation and the defined δ_{t+1}, the function $\hat{V}(b)$ can be written as

$$\begin{aligned}
&n\mathbb{P}_n \Bigg[\sum_t \Bigg\{ 2\delta_{t+1} - b^\top/\sqrt{n}\varphi(S_t, A_t) - \gamma \max_{a \in \mathcal{A}_{S_{t+1}}} \theta_0^\top \varphi(S_{t+1}, a) + \gamma \max_{a \in \mathcal{A}_{S_{t+1}}} (\theta_0 + b/\sqrt{n})^\top \varphi(S_{t+1}, a) \Bigg\} \varphi(S_t, A_t)^\top \Bigg] \hat{W}^{-1} \\
&\quad \times \mathbb{P}_n \Bigg[\sum_t \Bigg\{ -b^\top/\sqrt{n}\varphi(S_t, A_t) - \gamma \max_{a \in \mathcal{A}_{S_{t+1}}} \theta_0^\top \varphi(S_{t+1}, a) + \gamma \max_{a \in \mathcal{A}_{S_{t+1}}} (\theta_0 + b/\sqrt{n})^\top \varphi(S_{t+1}, a) \Bigg\} \varphi(S_t, A_t)^\top \Bigg]^\top.
\end{aligned}$$

The $\hat{V}(b)$ can be decomposed to the following parts:

I. $n\mathbb{P}_n \left\{\sum_t b^\top/\sqrt{n}\varphi(S_t, A_t)\varphi(S_t, A_t)^\top\right\} \hat{W}^{-1} \mathbb{P}_n \left\{\sum_t b^\top/\sqrt{n}\varphi(S_t, A_t)\varphi(S_t, A_t)^\top\right\}^\top \to_p b^\top W b$

II. $n\mathbb{P}_n \left\{\sum_t \zeta_{t+1}(\theta)\varphi(S_t, A_t)^\top\right\} \hat{W}^{-1} \mathbb{P}_n \left\{\sum_t \zeta_{t+1}(\theta)\varphi(S_t, A_t)^\top\right\}^\top \to_p E\left\{\sum_t \psi_{t+1}\varphi(S_t, A_t)^\top\right\} W^{-1} \times E\left\{\sum_t \psi_{t+1}\varphi(S_t, A_t)^\top\right\}^\top$

III. $-2n\mathbb{P}_n \left\{\sum_t \delta_{t+1}\varphi(S_t, A_t)^\top\right\} \hat{W}^{-1} \mathbb{P}_n \left\{\sum_t b^\top/\sqrt{n}\varphi(S_t, A_t)\varphi(S_t, A_t)^\top\right\}^\top \to_d -2Z_\infty b$

IV. $2n\mathbb{P}_n \left\{\sum_t \delta_{t+1}\varphi(S_t, A_t)^\top\right\} \hat{W}^{-1} \mathbb{P}_n \left\{\sum_t \zeta_{t+1}(\theta)\varphi(S_t, A_t)^\top\right\}^\top \to_d 2Z_\infty W^{-1} E\left\{\sum_t \psi_{t+1}\varphi(S_t, A_t)^\top\right\}^\top$

V. $-2n\mathbb{P}_n\left\{\sum_t b^\top/\sqrt{n}\varphi(S_t,A_t)\varphi(S_t,A_t)^\top\right\}\hat{W}^{-1}\mathbb{P}_n\left\{\sum_t \zeta_{t+1}(\theta)\varphi(S_t,A_t)^\top\right\}^\top \to_p$
$-2b^\top E\left\{\sum_t \psi_{t+1}\varphi(S_t,A_t)^\top\right\}^\top$

where $\zeta_{t+1}(\theta) = -\gamma \max_{a\in\mathcal{A}_{S_{t+1}}} \theta_0^\top\varphi(S_{t+1},a) + \gamma\max_{a\in\mathcal{A}_{S_{t+1}}}(\theta_0 + b/\sqrt{n})^\top\varphi(S_{t+1},a)$ and $\psi_{t+1} = \gamma\max_{a\in\pi^*} b^\top\varphi(S_{t+1},a)$. The first part follows from a law of large numbers. Here, we prove Part II and the rest follow similarly.

By adding and subtracting $\gamma\max_{a\in\pi^*} b^\top/\sqrt{n}\varphi(S_{t+1},a)$ to $\zeta_{t+1}(\theta)$, we have

$$\zeta_{t+1}(\theta) = -\gamma\max_{a\in\pi^*}(\theta_0 + b/\sqrt{n})^\top\varphi(S_{t+1},a) + \gamma\max_{a\in\mathcal{A}_{S_{t+1}}}(\theta_0 + b/\sqrt{n})^\top\varphi(S_{t+1},a) + \gamma\max_{a\in\pi^*} b^\top/\sqrt{n}\varphi(S_{t+1},a).$$

Thus, by Lemma S1, when $\|b/\sqrt{n}\| \to 0$ as $n\to\infty$, we have

$$\sqrt{n}[\zeta_{t+1}(\theta)] \to \psi_{t+1},$$

and by law of large numbers

$$\sqrt{n}\mathbb{P}_n\left\{\sum_t \zeta_{t+1}(\theta)\varphi(S_t,A_t)^\top\right\} \to_p E\left\{\sum_t \psi_{t+1}\varphi(S_t,A_t)^\top\right\}^\top.$$

We just showed that $\hat{V}(b)$ converges in distribution to

$$V(b) = -2Z_\infty b + 2Z_\infty W^{-1}E\left\{\sum_t \psi_{t+1}\varphi(S_t,A_t)^\top\right\}^\top + b^\top W b - 2b^\top E\left\{\sum_t \psi_{t+1}\varphi(S_t,A_t)^\top\right\}^\top$$
$$+ E\left\{\sum_t \psi_{t+1}\varphi(S_t,A_t)^\top\right\}W^{-1}E\left\{\sum_t \psi_{t+1}\varphi(S_t,A_t)^\top\right\}^\top.$$

By assuming that $V(b)$ is uniquely minimized in b and by continuity and local convexity of $V(b)$,

$$\sqrt{n}(\hat{\theta} - \theta_0) = \arg\min_b \hat{V}(b) \to_d \arg\min_b V(b),$$

which is a consequence of the epi-convergence results of Geyer (1994). Note that when $\gamma = 0$, $\arg\min_b V(b) = Z_\infty W^{-1}$. Also, under assumption $A.8$, that is, when $|\pi^*| = 1$, $\arg\min_b V(b) = Z_\infty\Gamma$, where

$$\Gamma = \left[I - \gamma\, W^{-1}E\left\{\sum_t \varphi(S_{t+1},\pi^*)\varphi(S_t,A_t)^\top\right\}\right]^\top$$
$$\left[W + \gamma^2/2E\left\{\sum_t \varphi(S_{t+1},\pi^*)\varphi(S_t,A_t)^\top\right\}W^{-1}E\left\{\sum_t \varphi(S_{t+1},\pi^*)\varphi(S_t,A_t)^\top\right\}^\top\right.$$
$$+ \gamma^2/2E\left\{\sum_t \varphi(S_{t+1},\pi^*)\varphi(S_t,A_t)^\top\right\}^\top W^{-1}E\left\{\sum_t \varphi(S_{t+1},\pi^*)\varphi(S_t,A_t)^\top\right\}$$
$$\left.-\gamma E\left\{\sum_t \varphi(S_{t+1},\pi^*)\varphi(S_t,A_t)^\top\right\} - \gamma E\left\{\sum_t \varphi(S_{t+1},\pi^*)\varphi(S_t,A_t)^\top\right\}^\top\right]^{-1},$$

where I is an identity matrix. Hence,

$$\sqrt{n}(\hat{\theta} - \theta_0) = \arg\min_b \hat{V}(b) \to_d N(0, \Gamma^\top \Sigma \Gamma),$$

with $\Sigma = E\left[\{\sum_t \delta_{t+1}\varphi(S_t, A_t)^\top\}^\top \{\sum_t \delta_{t+1}\varphi(S_t, A_t)^\top\}\right]$.

S4. THE EFFECT OF THE CHOICE OF γ ON THE ESTIMATED OPTIMAL REGIME.

In this section, we have estimated the optimal treatment regime under the simulation scenario discussed in the manuscript for different values of γ. To better reflect the effect of γ, we assume that there is no death (i.e., C_t=0 for all $t = 0, ..., 15$), and the reward function is defined as

- $R_t = 1$ if $A1c_t < 7$, -5 if $7 < A1c_t \& D_t = 1$ and zero otherwise.

For smaller values of γ ($\gamma = 0.1$), the estimated optimal policy will be more myopic and does not suggest augmenting any medication. This happens because the side effect outweighs the treatment effect. However, as γ gets larger, the optimal policy suggests to augment more treatments simply because the long-term effect of treatments outweighs the side effects. This shows that γ balances the immediate and long-term effect of treatments. Results are presented in Figure S1.

S5. THE EFFECT OF THE CHOICE OF TUNING PARAMETERS ON THE ESTIMATED OPTIMAL REGIME.

In this section, we discuss the effect of the tuning parameters (α, β) on the estimated optimal treatment regime under the simulation scenario discussed in the manuscript. We generated 500 datasets of size 2000 and applied the proposed methods using various choices of tuning parameters:

1. $\alpha = \frac{\nu}{k\log(k)}$ and $\beta = \frac{\nu}{k}$ for ν =0.050, 0.025, and 0.010
2. $\alpha = \frac{\nu}{k}$ and $\beta = \frac{\nu}{k^{3/4}}$ for ν =0.050, 0.025, and 0.010
3. $\alpha = \frac{\nu}{k}$ and $\beta = \frac{\nu}{k^{1/3}}$ for ν =0.050, and 0.010.

The parameter ν specifies the step size (increment size) for each choice of the tuning parameters. Figure S2 shows that as long as the step sizes are not very small, the stochastic minimization algorithm has a good performance. However, when the step sizes are small ($\nu = 0.010$), the algorithm fails to converge to the true values because it cannot reach the true minimizers of the objective function. Table S1 presents the value of the objective function at the estimated $\hat{\theta}$ and the number of required iterations to converge (K) using different tuning parameters. The value of the objective function $M(\theta)$ for small ν is more than twice the value of $M(\theta)$ for larger values of ν, which indicates the lack of convergence to the true minimizers. Based on this result, the first choice of tuning parameter, $\nu = 0.05$, outperforms the other choices.

Figure S3 displays the effect of tuning parameters on standard errors. The vertical axis is the ratios of the standard errors obtained by different simulation scenarios over the standard error obtained by $\alpha = \frac{\nu}{k\log(k)}$ and $\beta = \frac{\nu}{k}$ and $\nu = 0.050$. For example, the vertical axis in the first plot is

$$\frac{\text{S.D. of } \hat{\theta} \text{ when } \alpha = \frac{\nu}{k\log(k)}, \beta = \frac{\nu}{k} \text{ and } \nu = 0.025}{\text{S.D. of } \hat{\theta} \text{ when } \alpha = \frac{\nu}{k\log(k)}, \beta = \frac{\nu}{k} \text{ and } \nu = 0.050}.$$

Fig. S1: Simulation: The effect of the choice of γ on the estimated optimal regime.

Table S1: Simulation: The effect of the choice of tuning parameters on the estimated optimal regime.

	$\alpha = \frac{1}{k\log(k)}, \beta = \frac{1}{k}$			$\alpha = \frac{1}{k}, \beta = \frac{1}{k^{3/4}}$			$\alpha = \frac{1}{k}, \beta = \frac{1}{k^{1/3}}$	
	$\nu = 0.05$	$\nu = 0.025$	$\nu = 0.01$	$\nu = 0.05$	$\nu = 0.025$	$\nu = 0.01$	$\nu = 0.05$	$\nu = 0.01$
$M(\theta)$	0.007	0.011	0.025	0.007	0.008	0.017	0.007	0.018
K	14.03	13.22	12.86	21.35	20.66	19.44	21.77	19.45

Note that the reference S.D.s in the denominator includes the tuning parameter values used in the main simulation study in Section 4, which is shown to converge to the true values. In the first two rows, the ratios deviate more from one as the step size (ν) gets smaller. This can be due to stoping the updates before converging to the minimizer of the objective function.

Fig. S2: Simulation: The effect of the choice of tuning parameters on the estimated optimal regime.

S6. GEOMETRIC INTERPRETATION OF $M(\theta)$

The objective function $M(\theta)$ has a nice geometric interpretation which also explains why $\hat{W}$ improves the performance of the minimization algorithm. Define the Bellman operator acting on the optimal action-value function as

$$T^{\pi^*}Q^*(s,a) = \int [r(s,a,s') + \gamma Q^*\{s', \pi^*(s')\}] f(s'|s,a) ds'.$$

Also , let $\Pi T^{\pi^*}Q^*(s,a)$ denote the projection of the Bellman operator to the space of features $\varphi(.,.)$. Then using the Bellman equation we have $Q^*(s,a) = \Pi T^{\pi^*}Q^*(s,a) = \theta_0\varphi(s,a)$. Note

Fig. S3: Simulation: The effect of the choice of tuning parameters on the standard errors.

that

$$
\begin{aligned}
& E\left[\sum_{t=1}^{T-1}\left\{Q^*(S_t,A_t)-\Pi T^{\pi^*}Q^*(S_t,A_t)\right\}^\top\left\{Q^*(S_t,A_t)-\Pi T^{\pi^*}Q^*(S_t,A_t)\right\}\right] \\
&= E\left[\sum_{t=1}^{T-1}\left\{\Pi(Q^*(S_t,A_t)-T^{\pi^*}Q^*(S_t,A_t))\right\}^\top\left\{\Pi(Q^*(S_t,A_t)-T^{\pi^*}Q^*(S_t,A_t))\right\}\right] \\
&= E\left\{\sum_{t=1}^{T-1}\delta_{t+1}(\theta)\varphi(S_t,A_t)^\top\right\}E\left\{\sum_{t=1}^{T-1}\varphi(S_t,A_t)\varphi(S_t,A_t)^\top\right\}^{-1}E\left\{\sum_{t=1}^{T-1}\varphi(S_t,A_t)\varphi(S_t,A_t)^\top\right\} \\
&\qquad\qquad E\left\{\sum_{t=1}^{T-1}\varphi(S_t,A_t)\varphi(S_t,A_t)^\top\right\}^{-1}E\left\{\sum_{t=1}^{T-1}\delta_{t+1}(\theta)\varphi(S_t,A_t)^\top\right\} \\
&= D(\theta)W^{-1}D(\theta)^\top = M(\theta).
\end{aligned}
$$

The third equality follows from the definition of the projection operator

$$\Pi\{Q^*(S_t, A_t) - T^{\pi^*}Q^*(S_t, A_t)\} = \varphi(S_t, A_t)E\left\{\sum_{t=1}^{T-1}\varphi(S_t, A_t)\varphi(S_t, A_t)^\top\right\} \times E\left[\sum_{t=1}^{T-1}\{Q^*(S_t, A_t) - T^{\pi^*}Q^*(S_t, A_t)\}\varphi(S_t, A_t)^\top\right].$$

Thus a minimizer of $M(\theta)$, in fact minimizes the distance between the parametrized optimal state-value function $Q^*(s, a)$ with the projected Bellman error. Note that W^{-1} appears as a result of the projection operator.

When the postulated parametrized model for the optimal action-value function does not exactly match the true optimal action value function, Bellman equation may fail to hold, i.e., $Q^*(s, a) \neq T^{\pi^*}Q^*(s, a)$. However, using the projected Bellman operator ensures that there exist a θ for which $Q^*(s, a) = \Pi T^{\pi^*}Q^*(s, a)$ holds. In other words, minimizing the objective function $M(\theta)$ leads to an estimate of the state-value function that is optimal given the postulated parametric model.

S7. TUNING PARAMETERS

The tuning parameters α_k and β_k in the GGQ algorithm need to satisfy the following assumptions (Maei et al., 2010):

P.1 $\alpha_k, \beta_k\ \forall k$ and are deterministic.
P.2 $\sum_{k=0}^{\infty}\alpha_k = \sum_{k=0}^{\infty}\beta_k = \infty$.
P.3 $\sum_{k=0}^{\infty}(\alpha_k^2 + \beta_k^2) < \infty$.
P.4 $\alpha_k/\beta_k \to 0$.

S8. FEATURE FUNCTIONS

The feature functions are constructed using the radial basis functions and

$$\varphi(s, a) = I(s \neq \emptyset)\{\varphi_1(s, a), \varphi_2(s, a), \varphi_3(s, a), \varphi_4(s, a), \varphi_5(s, a), \varphi_6(s, a), \varphi_7(s, a), \varphi_8(s, a), \varphi_9(s, a)\},$$

where

$$\varphi_1(s,a) = I(A=0, NAT=0)[1, \exp\{-h(A1c-q_{11})^2\}, \exp\{-h(A1c-q_{12})^2\}, \phi(BP), \phi(Weight)]$$
$$\varphi_2(s,a) = I(A=0, NAT=1)[1, \exp\{-h(A1c-q_{21})^2\}, \exp\{-h(A1c-q_{22})^2\}, \exp\{-h(A1c-8.0)^2\}, d, \phi(BP), \phi(Weight)]$$
$$\varphi_3(s,a) = I(A=0, NAT=2)[1, \exp\{-h(A1c-q_{31})^2\}, \exp\{-h(A1c-q_{33})^2\}, d, \phi(BP), \phi(Weight)]$$
$$\varphi_4(s,a) = I(A=0, NAT=3)[1, \exp\{-h(A1c-q_{41})^2\}, \exp\{-h(A1c-q_{42})^2\}, \exp\{-h(A1c-8.5)^2\}, d, \phi(BP), \phi(Weight)]$$
$$\varphi_5(s,a) = I(A=0, NAT=4)[1, \exp\{-h(A1c-q_{51})^2\}, \exp\{-h(A1c-q_{52})^2, \exp\{-h(A1c-8.0)^2\}, d, \phi(BP), \phi(Weight)]$$
$$\varphi_6(s,a) = I(A=1, NAT=0)[1, \exp\{-h(A1c-6.5)^2\}, \exp\{-h(A1c-7.5)^2\}, \phi(BP), \phi(Weight)]$$
$$\varphi_7(s,a) = I(A=2, NAT=1)[1, \exp\{-h(A1c-6.5)^2\}, \exp\{-h(A1c-q_{71})^2\}, \exp\{-h(A1c-q_{73})^2\} d, \phi(BP), \phi(Weight)]$$
$$\varphi_8(s,a) = I(A=3, NAT=2)[1, \exp\{-h(A1c-q_{82})^2\}, \exp\{-h(A1c-8.5)^2\}, d, \phi(BP), \phi(Weight)]$$
$$\varphi_9(s,a) = I(A=4, NAT=3)[1, \exp\{-h(A1c-6.5)^2\}, \exp\{-h(A1c-q_{92})^2, \exp\{-h(A1c-8.5)^2\}, d, \phi(BP), \phi(Weight)],$$

where $\phi(BP) = [\exp\{-h(\text{BP}-q_{b1})^2\}, \exp\{-h(\text{BP}-q_{b3})^2\}]$ and $\phi(\text{Weight}) = [\exp\{-h(\text{Weight}-q_{w1})^2\}, \exp\{-h(\text{Weight}-q_{w3})^2\}]$. h is a positive constant and $q_{.j}$ is the observed jth quantile of the corresponding variable. For example, q_{11} and q_{12} are the first and second quantiles of $A1c$ given $A=0$ and NAT$=0$. Similarly, q_{b1} and q_{b3} are the first and third quantiles of BP. Note that, in our generative model, BP and Weight are independent of A and NAT.

Remark 1. Number of quantiles used in each φ_k and $\phi(.)$ is a bias-variance trade-off such that increasing the number of quantiles decreases the bias but increases the variance of the estimated parameters. Similarly, decreasing the value of h may decrease the bias but increase the variance of the estimators. In our simulation, we set $h = 0.5$.

S9. Simulation generative model

Here are the steps we take to generate the dataset:

- Baseline variables: Standardized variables $(BP_0, Weight_0)$ are generated independently standard normal distribution, where BP is the systolic blood pressure. Also, $NAT_0 = D_0 = C_0 = 0$ and $A1c_0$ is generated a normal distribution with mean 9.0 and variance 1.
- Assigned treatment at time t: Given the state variable NAT_t, the sets of available treatments are $\mathcal{A}_{NAT_t=0} = \{0, Metformin\}$, $\mathcal{A}_{NAT_t=1} = \{0, Sulfonylurea\}$, $\mathcal{A}_{NAT_t=2} = \{0, Glitazone\}$, $\mathcal{A}_{NAT_t=3} = \{0, Insulin\}$, and $\mathcal{A}_{NAT_t=4} = \{0\}$, where 0 means continue with the treatment received at the previous decision point. Although the ideal $A1c$ level is below 7%, Timbie et al. (2010) raises concern about the feasibility and polypharmacy burden needed for treating patients whose $7 < A1c < 8$. Our simulation study investigates the optimal treatment regime for these patients. More specifically,
  - if $A1c_t < 7$, the treatment is not augmented because $A1c$ is under control and $NAT_t = NAT_{t-1}$.
  - if $A1c_t > 8$ and $NAT_{t-1} < 4$, the treatment is augmented with the next available treatment. Hence, $NAT_t = NAT_{t-1} + 1$. Note that these are patients whose $A1c$ is too high. Thus, the only option is augmenting their treatment.

- If $7 < A1c_t < 8$ and $NAT_{t-1} < 4$, then a binary variable Z_t is generated from $Z_t \sim Ber\left(\frac{\exp[-0.2A1c_{t-1}+0.5NAT_{t-1}+0.5D_{t-1}]}{1+\exp[-0.2A1c_{t-1}+0.5NAT_{t-1}+0.5D_{t-1}]}\right)$, where D_t denotes the discontinuation indicator. $Z_t = 1$ implies that the patient continues with the same treatment as time $t-1$ and we set $A_t = 0$ ($NAT_t = NAT_{t-1}$), while $Z_t = 0$ implies that the patient takes the next available treatment (treatment is augmented) and we set $NAT_t = NAT_{t-1} + 1$. For example, if $7 < A1c_t < 8$ and $NAT_{t-1} = 3$, a patient can be assigned to either augmenting the treatment taken at time $t-1$ with $A_t = insulin$ or continuing with the same treatment as time $t-1$ ($A_t = 0$), depending on the generated variable Z_t.
  Note: When $Z_t = 1$, no new treatment is added. Hence, $E[A1c_t|Z_t = 1, A1c_{t-1}] = E[A1c_t|A_t = 0, A1c_{t-1}] = A1c_{t-1}$.

- Treatment discontinuation indicator at time t: A binary variable D_t is generated from a Bernoulli distribution given the last augmented treatment. For all t, the treatment discontinuation rates are $p(D_t|A_{t-1} = metformin) = p(D_t|A_{t-1} = sulfonylurea) = p(D_t|A_{t-1} = glitazone) = 0.20$, and $p(D_t|A_{t-1} = insulin) = 0.35$.
  Note: We assume no treatment discontinuation for patients who are taking the same treatment at time t as at time $t-1$ (i.e., $p[D_t = 1|A_{t-1} = 0] = 0$).
- Intermediate outcome $A1c$ at time t: To avoid variance inflation through time, we use the following generative model for $A1c$ at time t, $A1c_t = \frac{A1c_{t-1}-\mu_{t-1}+\epsilon}{\sqrt{(1+\sigma_\epsilon^2)}} + \mu_t$, where $\epsilon \sim N(0, \sigma_\epsilon = 0.5)$ and

$$\mu_t = E[A1c_t|A1c_{t-1}, NAT_{t-1}, A_{t-1}, D_t] = \begin{cases} \mu_{t-1}(1-\tau_{A_t}) & \text{if } A1c_{t-1} > 7, NAT_{t-1} < 4, A_{t-1} \neq 0, D_t \neq 1, \\ \mu_{t-1} & \text{o.w.} \end{cases}$$

  with τ_{A_t} being the treatment effect of the augmented treatment A_t. The treatment effects of metformin, sulfonylurea, glitazone* and insulin are 0.14, 0.20, 0.02, and 0.14, respectively. Note that the treatment effects are reported as a percentage reduction in $A1c$. The treatment effect of glitazone is listed as 0.12 in Timbie et al. (2010), which is similar to metformin. However, in order to study the effect of the treatment discontinuation on the optimal regime, we set its treatment effect to 0.02 and, from now on, denote it by glitazone*.
- Time-varying variables at time t: $BP_t = (BP_{t-1} + \epsilon)/(\sqrt{1+\sigma_\epsilon^2})$ and $Weight_t = (Weight_{t-1} + \epsilon)/(\sqrt{1+\sigma_\epsilon^2})$.
- Death indicator at time t: A binary variable C_t is generated from a Bernoulli distribution with probability $\frac{\exp\{-10+0.08I(A1c_{t-1}>7)A1c_{t-1}^2+0.5NAT_{t-1}\}}{1+\exp\{-10+0.08I(A1c_{t-1}>7)A1c_{t-1}^2+0.5NAT_{t-1}\}}$. $C_t = 1$ is the indicator of death.
- Reward function at time t: In order to be able to find an optimal treatment regime, we need an operational definition of controlled $A1c$. Hence, we define the following reward function at time t as a function of $A1c_t$, D_t and C_t,
  - $R_t = 1$ if $A1c_t < 7$, -2 if $7 < A1c_t \& D_t = 1$, -10 if $C_t = 1$ and zero otherwise.
  This reward structure helps us to identify treatments whose discontinuation rate outweighs their efficacy while reducing the chance of death.

Note that the state space at time t includes $S_t = (NAT_t, D_t, A1c_t, BP_t, Weight_t)$. However, the Markov property holds with $(NAT_t, D_t, A1c_t)$, and variables BP and $Weight$ are noise variables.

REFERENCES

Geyer, C. J. (1994). On the asymptotics of constrained m-estimation. *The Annals of Statistics*, 2(4):1993–2010.

Maei, H., Szepesvári, C., Bhatnagar, S., and Sutton, R. (2010). Toward off-policy learning control with function approximation. *Proc. ICML 2010*, pages 719–726.

Sutton, R., Szepesvári, C., and Maei, H. (2009). A convergent o (n) algorithm for off-policy temporal-difference learning with linear function approximation. *Advances in Neural Information Processing Systems 21*.

Timbie, J., Hayward, R., and Vijan, S. (2010). Diminishing efficacy of combination therapy, response-heterogeneity, and treatment intolerance limit the attainability of tight risk factor control in patients with diabetes. *Health Services Research*, 45(2):437–456.

Van Der Vaart, A. and Wellner, J. (1996). *Weak convergence and empirical processes*. Springer New York.