Electronic Journal of Statistics
Vol. 8 (2014) 1309–1311
ISSN: 1935-7524
DOI: 10.1214/14-EJS907

Discussion of "Dynamic treatment regimes: Technical challenges and applications"*

Moulinath Banerjee

Department of Statistics
University of Michigan
Ann Arbor, Michigan 48109
e-mail: moulib@umich.edu

Abstract: The contributions of the review paper on the technical challenges and applications of dynamic treatment regimes are briefly discussed and potential extensions to continuous action spaces and high dimensional problems are indicated.

MSC 2010 subject classifications: Primary 62C05.
Keywords and phrases: Asymptotic bias, dynamic treatment regime, non-regularity, Q-function.

Received May 2014.

I would like to start out by congratulating the authors of the article on both a lucid and precise exposition of some of the key inferential challenges that arise in the prescription of optimal dynamic treatment regimes (DTR from now on) which have found growing interest across the clinical sciences but also have intimate connections to other forms of adaptive decision making (in engineering/robotics contexts). This particular discussant has had the good fortune to hear his colleague, Professor Murphy, and also members of her group, speak on several occasions on this area and this review paper is the perfect icing on that cake, providing a great degree of clarity about the subtle mathematical issues that arise naturally in very canonical versions of this problem: non-regularity of least squares estimates of the parameters associated with the optimal DTR and issues related to the associated asymptotic bias.

The paper is formulated within the context of Q-learning, an indirect estimation method for optimal DTR that is attractive when model building can be aided by expert opinion. For ease of exposition, the authors concentrate on two treatment times which is sufficient to illustrate the non-regularity issue. The Q-functions are taken to have linear forms and the action space (decision space) taken to be binary – generalization to a finite action space is straightforward though notationally (and algebraically) tedious. A version of Q-learning that uses a least squares based algorithm is described in Section 2.1. It appears that for this particular algorithm to be consistent with the definitions of the Q_2 and

*Main article 10.1214/14-EJS920.

Q_1 functions that precede it, Y_1 should be part of H_2, the history at time 2. This does not seem to be explicated in the paper.

The source of non-regularity lies in the estimation of $\beta_1^\star$, the optimal population regression coefficient at time 1, as its estimation involves regressing the predicted outcome $\tilde{Y}$ on (H_1, A_1) and the computation of $\tilde{Y}$ depends on a maximizing operation over the action space at time 2. This leads to an expression for $\hat{\beta}_1$, the least squares estimate of $\beta_1^\star$, whose normalized version contains the expression

$$\mathbb{U}_n = \sqrt{n}\,\left([H_{2,1}^T\,\hat{\beta}_{2,1}]_+ - [H_{2,1}^T\,\beta_{2,1}^\star]_+\right)\,.$$

Owing to the appearance of positive parts the distribution of $\mathbb{U}_n$ depends in an abrupt way on whether $\beta_{2,1}^\star$ is orthogonal to $H_{2,1}$, the part of the history at time 2 that is relevant to decision making, leading to a 'boundary phenomenon' and subsequent non-regularity (with a rigorous description in Theorem 4.1). The asymptotic bias of $\hat{\beta}_1$ is also shown to be strictly positive (Theorem 3.1). While, strictly speaking, non-regularity is not possible with continuous distributions for $H_{2,1}$, relatively small magnitudes for the second stage treatment effect may push the scenario close towards the non-regular situation for modest sample sizes. This motivates the formulation of moving parameter asymptotics expounded in Section 3 that provide a more appealing way of studying the behavior of the non-smooth estimator $\hat{\beta}_1$. This is formalized via a standard Hellinger–differentiability type formulation (A3) along a sequence of local alternatives. A particular insight gleaned from the local alternatives exposition is the fact that ample caution needs to be exercised while thresholding the predicted outcome $\tilde{Y}$ with an eye towards reducing the asymptotic bias of $\hat{\beta}_1$: aggressive thresholding can blow up the supremum of the c-directional asymptotic bias over the direction of approach of the local parameter to ∞. The toy example using a randomized two-arm study is particularly illustrative in this context.

Section 4.2 describes the construction of adaptive confidence intervals for linear combinations of the Stage 1 regression coefficients through upper and lower bounding of the term involving the problematic quantity $\mathbb{U}_n$ by regular uniformly convergent stochastic quantities, a strategy motivated by earlier work of some of the authors. The asymptotic properties of the bounds are investigated in Theorem 4.1: a pleasing property of these bounds is that they adapt to the situation where all patients experience a treatment effect. An important issue here would seem to be how these bounds are reliably computed in practice as they involve taking the suprema or infima over a quantity varying continuously over a Euclidean space of a certain dimension, which could be be large in applications. There seem to be no pointers in this direction in the paper.

The authors end by discusing some interesting challenges in this area. I would like to focus on two more aspects of dynamic treatment regimes which seem natural. The first has to do with continuous action spaces as opposed to the discrete action space of this paper and the second with high-dimensional versions of these problems.

Continuous action space It would be desirable to accommodate a continuous action space into this framework. To motivate this, consider a situation of an un-manned vehicle and think of the action at time t as deflecting the vehicle by a certain angle from its course so as to avoid a collision or some other mishap. In this case a_t would be the angle by which the vehicle is deviated at time t and is therefore naturally modeled as being in a continuous space. In the examples discussed in the paper the non-regularity appears, in a sense, to be tied to the discrete nature of the action space. In what situations might the issue of non-regularity arise in such a context? Of course, a continuous action space may not be natural in a clinical context but should be more relevant in engineering/technology.

High-dimensional action space Here I am thinking about controlling a complex system over time where the action vector a_t could be high-dimensional. To put this into context, suppose that at time t action involves altering a large number of settings, say p of these, on a control panel. For concreteness, suppose $a_{t,j}$, the j'th component of a_t assumes values $-1, 0, 1$: 0 corresponds to maintaing the current setting and invokes no cost, -1 and 1 are changes of setting of two different kinds that involve a cost. For starters, consider a scenario without histories in the spirit of the toy example that the authors discuss in their paper. Let $Q_t(a_t, \beta_t) = \Psi(\langle a_t, \beta_t \rangle)$ where Ψ, say, is a monotone function; even the identity function for the purpose of this discussion. Naturally one would like to make a sparsity assumption on β_t, the idea being that many of the settings may not require any tweaking at time t and one wants to optimize over a parsimonious number of actions. In this case, the Stage 2 regression step could involve solving a least squares problem penalized by the l_1 norm of β_2. Having obtained a parsimonious solution, $\hat{\beta}_{2,\lambda}$, one can now maximize $Q_2(a_2, \hat{\beta}_{2,\lambda})$ over all vectors a_2 but because of the sparsity of $\hat{\beta}_{2,\lambda}$ this becomes a lower dimensional optimization problem.

Of course in a practical setting, histories will need to be accommodated. In a situation where the relevant part of the history, say $h_{t,1}$ (using notation from the paper), is a vector capturing information on the p settings $Q_t(h_t, a_t, \beta_t)$ might be modeled as $\Psi(\langle a_t \cdot h_t \ , \ \beta_t \rangle)$ where $\cdot$ denotes the co-ordinate wise product of two vectors and a similar optimization procedure as above could be used. The key question here, of course, is how tractable such models would be at an analytical level.

Last but not least, given Professor Murphy's expertise in semiparametrics, the possibility of using semiparametric approaches to Q-learning would seem a natural mode of approach, though as an outsider to this field it is difficult for me to judge the marginal gains that might accrue from them in practical contexts.