Clairon, Q., Henderson, R., Young, N., Wilson, E. and Taylor, C.J. (2020) Adaptive treatment and robust control, *Biometrics*. DOI:10.1111/biom.13268

Biometrics 00, 1–26 DOI: 1111111111
XXXXX 0000

Adaptive Treatment and Robust Control

Q. CLAIRON[1,*], R. HENDERSON[2],N.J. YOUNG[2],E.D. WILSON[3] andC.J. TAYLOR[4]

[1]University of Bordeaux, Inria Bordeaux Sud-Ouest,

Inserm, Bordeaux Population Health Research Center, SISTM Team, UMR 1219, 33000 Bordeaux France.

[2]School of Mathematics, Statistics and Physics, Newcastle University, NE1 7RU, U.K.

[3]School of Computing and Communications, Lancaster University, LA1 4WA

[4]Department of Engineering, Lancaster University, LA1 4YF, U.K.

email: quentin.clairon@u-bordeaux.fr

Summary: A control theory perspective on determination of optimal dynamic treatment regimes is considered. The aim is to adapt statistical methodology that has been developed for medical or other biostatistical applications so as to incorporate powerful control techniques that have been designed for engineering or other technological problems. Data tend to be sparse and noisy in the biostatistical area and interest has tended to be in statistical inference for treatment effects. In engineering fields, experimental data can be more easily obtained and reproduced and interest is more often in performance and stability of proposed controllers rather than modelling and inference per se. We propose that modelling and estimation be based on standard statistical techniques but subsequent treatment policy be obtained from robust control. To bring focus, we concentrate on A-learning methodology as developed in the biostatistical literature and H^∞-synthesis from control theory. Simulations and two applications demonstrate robustness of the H^∞ strategy compared to standard A-learning in the presence of model misspecification or measurement error.
Key words: A-learning; Anticoagulation; Control; H^∞-synthesis; Misspecification; Personalized medicine; Robustness.

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

Murphy (2003) introduced to a wide statistical audience the concepts of optimal dynamic treatment allocation. In brief, decision rules are sought to allow treatment or other actions to adapt to accruing information in an optimal way. With increasing interest in personalized medicine, Murphy's ideas have been taken up widely in the biostatistical literature and approaches such as A-learning, Q-learning and outcome-weighted learning have become popular. These methods are closely related to reinforcement learning in the machine learning literature, to dynamic programming in general and to other sequential methods such as the g-computation approach of Robins (1986). Chakraborty and Moodie (2013) provide a good overview.

Murphy's original approach is a form of A-learning, with the A standing for *advantage*, and under which contrasts between expected outcomes under different treatment regimes are modelled. Examples include Murphy (2003); Robins (2004); Henderson et al. (2010), and Henderson et al. (2011). In Q-learning, where Q is taken from *quality*, the response itself is modelled at each decision time as a function of history to date, and optimal actions are determined sequentially (Chakraborty et al., 2013; Laber et al., 2014; Moodie et al., 2014; Wallace and Moodie, 2015; Song et al., 2015; Linn et al., 2017). A- and Q- learning are reviewed by Chakraborty and Moodie (2013) and Schulte et al. (2014). Outcome weighted learning (Zhao et al., 2012; Zhang et al., 2012) is a form of direct search based on direct estimation of the decision rule itself (Zhao et al., 2015; Zhou et al., 2017). Most of the methods that have been developed to date are applicable to the case of a binary treatment. In this paper we are interested in situations in which the treatment can take a large number of values, such as the dose of a drug, and can be considered as effectively continuous. In principle both A- and Q-learning and other methods can still apply, and indeed have been applied by a small number of authors (Rosthøj et al., 2012; Barrett et al., 2014; Rich et al.,

2014). Outcome learning is more problematic, though we note promising work for single timepoints by Chen et al. (2016) and also the tree-based approach of Laber and Zhao (2015). Nonetheless it is clear that this area is much less well developed than the binary treatment situation.

When treatments are considered as continuous and there are a reasonable number of decision times, there are close similarities between the dynamic treatment problems considered in statistics and some of the problems considered by control analysts (Sontag, 1998; Taylor et al., 2013). Control theory has been useful for a number of statistical problems, such as optimal experimental design (Pronzato, 2008; Hooker et al., 2015), theoretical analysis of treatment allocation (Orellana, 2010; Zhang and Xu, 2016), or for control of biomarkers (Deshpande et al., 2014; Chakrabarty et al., 2017). Still, to our knowledge, rather few of the vast array of control techniques have been transported into the statistical methodology literature. To illustrate, R is probably the computing environment that is most used amongst researchers in statistics. Yet a search of the over 14,000 contributed packages that are publicly available through the Comprehensive R Archive Network CRAN, showed none at all related to control theory. In contrast there are multiple toolboxes in MATLAB, which is popular with engineers, for almost all aspects of control theory.

More specifically, few of the many robust control methodologies have been transported for optimal dynamic treatment problems, one exception being Bekiroglu et al. (2017), who proposed a model predictive control approach assuming a sequence of binary treatments mimicking behavioural change experiments. This is despite the robust control framework complying well with biostatistical optimal dynamic treatment problems, where there are multiple sources of uncertainty, including measurement noise, model misspecification, inter-patient variability and so on. This paper takes a step in this direction by developing so-called H^{∞}-synthesis for A-learning using Murphy's regret functions. We selected A-learning

simply to bring focus: the methods can also be applied to other optimal dynamic treatment methodologies.

We do not consider modelling, estimation or inference, relying for these on existing A-learning techniques. Instead, our aim is show how H^∞ methods can be applied after the modelling stage so as to develop treatment rules that are robust to a raft of departures from the assumed model (Glover and Doyle, 1988; Doyle et al., 1989). These aims are partially shared with some attempts by statistical researchers to develop either robust or transportable decision rules (Orellana et al., 2010; Qian and Murphy, 2008; Zhang et al., 2012, 2013; Wallace et al., 2016; Xu et al., 2016). As will be seen, the methods are very different.

In the next section, we present our general framework, summarise the A-learning methodology and show how regret and state-space models can be linked. In Section 3 we concentrate on linear state-space models and discuss a selection of treatment strategies from the statistical and control literatures, including a summary of the H^∞ approach, with fuller information provided in online supplementary material. In Section 4 we present simulations in which the approach developed here gives a more robust strategy than a treatment policy derived directly by application of A-learning. This advantage is confirmed in two applications presented in Section 5. In the first, our method and standard A-learning procedures are compared using an experimental ventilation chamber which allows evaluation and comparison of proposed strategies. In the second, we conduct a retrospective comparison between model-based and robust anticoagulation decision rules and those selected in practice by clinicians.

2. Preliminaries

2.1 *Informal background*

We assume that treatment allocations or other decisions are to be made for individual subjects longitudinally, with the expectation that the choice will depend upon accruing

information. For a generic subject, assessments are to be made online at times $1, \ldots, T$. At time t the available information on the current *state* of the individual is S_t. This might be, for example, a vector of biomarker values, results of a battery of psychological tests, or some other summary of current condition or change in condition.

Immediately after S_t becomes known, an *action* A_t is selected, such as treatment to be administered or discontinued, dosage to be applied, or perhaps timing of the next assessment. For $t > 1$ let $\overline{S}_t = (S_1, \ldots S_t)$ and $\overline{A}_{t-1} = (A_1, \ldots A_{t-1})$. In addition set $\overline{S}_1 = S_1$ and adopt the convention that $\overline{A}_0$ is null. Hence the information available to the decision maker at time t is $H_t = (\overline{S}_t, \overline{A}_{t-1})$ and the objective is to develop a decision function $d_t = d_t(H_t)$ that proposes an action given the information to hand. A treatment strategy then means a sequence of decision functions $d = \{d_t\}_{1 \leqslant t \leqslant T-1}$ chosen to meet some overall aim generally given under the form of an outcome of interest $Y(d)$ to maximise. In our case we restrict ourselves to the objective of stabilizing a patient state $\{S_t\}$ as closely as possible to a target s^*. For this, we take as outcome:

$$Y(d) = -\sum_{t=1}^{T} \{S_t(d) - s^*\}^{\mathsf{T}} \{S_t(d) - s^*\} = -\sum_{t=1}^{T} \|S_t(d) - s^*\|_2^2, \tag{1}$$

and the optimal strategy d^{opt} is defined as the maximizer over strategies d of $\mathbb{E}[Y(d)]$. For simplicity we will assume the target to be time-fixed, but the methods can be extended to time-varying targets if required. This is a typical objective for long-term maintenance therapies, such as control of insulin, white blood cell count, CD4 levels or blood clotting times, as in the example of Section 5.2.

Here $Y(d)$ and $S_t(d)$ are counterfactuals that would arise if treatment strategy d were selected. As our focus is not on estimation and inference, we will not introduce any special notation for counterfactuals, nor discuss conditions required for causal inference. See Chakraborty and Moodie (2013) for fuller information and discussion if required.

2.2 *A-learning and regret modelling*

A-learning relies on modelling contrasts between expected outcomes under different decision rules (Schulte et al., 2014; Moodie et al., 2007; Robins, 2004; Rosthøj et al., 2006). The *regret* version introduced by Murphy (2003) is based on functions of the form

$$\mu_t\left(a_t \mid H_t\right) = \mathbb{E}\{Y(\underline{d}_t^{opt}) \mid H_t\} - \mathbb{E}\{Y(a_t, \underline{d}_{t+1}^{opt}) \mid H_t\} \tag{2}$$

with $\underline{d}_t^{opt} = (d_t^{opt}, \ldots, d_{T-1}^{opt})$. Thus $\mathbb{E}\{Y(\underline{d}_t^{opt}) \mid H_t\}$ denotes the expected value of Y if the optimal policy is followed starting from t and if the patient has prior history H_t. In contrast $\mathbb{E}\{Y(a_t, \underline{d}_{t+1}^{opt}) \mid H_t\}$ is the expected value if, given the same prior history, action a_t is selected at t but the optimal policy followed thereafter.

The function (2) quantifies the loss made by choosing a_t as the action at time t instead of $d_t^{opt}(H_t)$, and is consequently the regret caused by taking a_t in comparison with the best possible decision. Since we have assumed that the objective is to maximize the expected value of Y, by definition the regret is non-negative and $\mu_t(a_t \mid H_t) = 0$ if and only if a_t corresponds to the optimal decision $d_t^{opt}(H_t)$.

The optimal decision rule is of course unknown. To estimate it, Murphy proposed two steps:

(1) A parametric form $\mu_t\left(a_t \mid H_t\right) \simeq \mu_t\left(a_t \mid H_t; \psi\right)$ is assumed, and an estimator $\widehat{\psi}$ is constructed from the data. Usually only a subset of H_t is assumed to be important and the dimension of ψ is kept modest.

(2) An estimator of the optimal treatment strategy value $\widehat{d}_t^{opt}\left(H_t\right)$ is derived by solving the equation $\mu_t(a_t \mid H_t; \widehat{\psi}) = 0$.

The first step is delicate and has been the focus of most research in the biostatistical area. But once it is achieved, the second step is usually taken as straightforward provided the regret function is parametrized so that zero is achievable. We simply choose the action that leads to zero regret.

2.3 *Link with state-space models*

To adopt a control theory point of view we need to make a link between $\mu_t(a_t \mid H_t)$ and a state-space model of S_t. First we introduce some notation. We use $\mathscr{S}_t$ and $\mathscr{A}_t$ to denote the supports of the state and action variables S_t and A_t respectively. Similarly, we define $\overline{\mathscr{S}}_t$ and $\overline{\mathscr{A}}_t$ as the supports of the histories $\overline{S}_t$ and $\overline{A}_t$. Recall that the information available immediately prior to choice of A_t is $H_t = \left(\overline{S}_t, \overline{A}_{t-1}\right)$ and that $d_t^{opt}(H_t)$ is the optimal but unknown decision at time t. For $r > t$ now let $\underline{d}_{r\setminus t}^{opt} = (d_{t+1}^{opt}, \ldots, d_r^{opt})$ be the next $r-t$ optimal decisions. Obviously these will depend on past values but this is suppressed in the notation.

We will make three assumptions.

A1. The target for maximisation is $Y(d)$ given by (1), that is

$$Y(d) = -\sum_{t=1}^{T} \|S_t(d) - s^*\|_2^2.$$

A2. The system evolution linking the next patient state S_{t+1} with its history H_t and the chosen action A_t can be described by a state-space model

$$S_{t+1} = f_t(H_t, A_t) + \varepsilon_t^{dyn} \tag{3}$$

for some function f_t, and with ε_t^{dyn} a random variable representing stochastic innovations acting on the system dynamic and independent of (H_t, A_t). Without loss of generality we assume $E[\varepsilon_t^{dyn}] = 0$.

A3. At each time t there is an action which will allow the expected state to reach its target value. Thus for all $\overline{S}_t \in \overline{\mathscr{S}}_t$ and $\overline{A}_{t-1} \in \overline{\mathscr{A}}_{t-1}$, there is an $a_t \in \mathscr{A}_t$ such that

$$f_t(\overline{S}_t, \overline{A}_{t-1}, a_t) = s^*.$$

Ensuring s^* is reachable given a state-space model is known as the reachability/controllability problem in control theory. When the model is linear and time-invariant, a rank condition on model components has to be satisfied (see Sontag (1998) Chapter 3). For more general models there is no standardised way to check for controllability, and bespoke methods are needed for the model and target to hand.

Each of these assumptions can be relaxed to at least some extent. For example we might adopt a non-myopic strategy by allowing the target s^* to be time-varying and always within the range that is achievable at any time. Then the expected state can be walked in the desired direction rather than forcing a dramatic short-term change. The assumptions are kept to allow us to focus on the main ideas rather than considering too much detail and our results follow from the assumptions as given. Proofs of the propositions are in supplementary material. The first result relies on the form of Y and the independence of ε_t^{dyn} from, specifically, the immediately preceding action.

PROPOSITION 1: Under A1 and A2, the regret function at time $t \in \{1, \ldots, T-1\}$ has the general expression

$$\mu_t\left(a_t \mid H_t\right) = \left\|f_t(H_t, a_t) - s^*\right\|_2^2 - \left\|f_t(H_t, d_t^{opt}(H_t)) - s^*\right\|_2^2$$

$$+ \sum_{r=t+1}^{T-1} \mathbb{E}\left\{\left\|f_r(\overline{S}_r, \{\overline{A}_{t-1}, a_t, d_{r\setminus t}^{opt}\}) + \varepsilon_r^{dyn} - s^*\right\|_2^2 \mid H_t\right\}$$

$$- \sum_{r=t+1}^{T-1} \mathbb{E}\left\{\left\|f_r(\overline{S}_r, \{\overline{A}_{t-1}, d_{r\setminus(t-1)}^{opt}\}) + \varepsilon_r^{dyn} - s^*\right\|_2^2 \mid H_t\right\}.$$

The addition of assumption A3 leads to a simpler result.

PROPOSITION 2: Under A1, A2 and A3 the regret function at time $t \in \{1, \ldots, T-1\}$ has the general expression

$$\mu_t\left(a_t \mid H_t\right) = \left\|f_t(H_t, a_t) - s^*\right\|_2^2.$$

After fitting a regret or other model, a decision strategy can be selected. Performance in future practice may then be affected by at least three sources of uncertainty.

(1) Parametric uncertainty. Error due to estimation of the assumed parameter values, i.e. $\widehat{\psi} = \psi + \triangle$, with $\triangle$ an unknown error term.

(2) Measurement uncertainty. We may not have access to S_t but only to noisy observations $\mathbb{S}_t = S_t + \varepsilon_t^m$ with ε_t^m measurement noise.

(3) Model misspecification: dynamic and model uncertainty. The regret function parametrization may not be appropriate for the training data, which we think of as model uncertainty. Or it may be correctly specified for the training data but not fully suitable for all new patients, which we refer to as dynamic uncertainty.

Parametric uncertainty is of course acknowledged in standard statistical approaches, at least during the inference stage. A consequence of measurement uncertainty is that estimates and treatment strategies can be based only on the observed history $\mathbb{H}_t = (\overline{\mathbb{S}}_t, \overline{A}_{t-1})$, where $\overline{\mathbb{S}}_t = (\mathbb{S}_1, \ldots, \mathbb{S}_t)$, rather than the true history H_t. In principle measurement error can be allowed in standard optimal dynamic treatment approaches, though in practice this possibility seems to be overlooked. Model and, particulary, dynamic uncertainty are more problematic from a statistical viewpoint but these are precisely the forms of disturbances that robust control policies are designed to give protection against.

3. Treatment Strategies for Linear State-Space Models

3.1 *Linear State-Space Model*

The definition of a discrete optimal control problem generally requires a state-space model describing the evolution of S_t and a cost function to minimize which takes as argument acceptable treatment strategies d.

A state-space model follows from Proposition 2. As soon as we assume a parametric formulation for the regret function $\mu_t(a_t \mid H_t) = \mu_t(a_t \mid H_t; \psi)$, we automatically have a parametric form for $f_t(H_t, a_t) = f_t(H_t, a_t; \psi)$. For this work, we restrict ourselves to parametric regret functions of the form:

$$\mu_t(a_t \mid H_t; \psi) = \|s^* - F_S(\psi)\overline{S}_t^r - F_A(\psi)\overline{A}_{t-1}^q - F_a(\psi)a_t\|_2^2, \tag{4}$$

where $\overline{S}_t^r = (S_t, S_{t-1}, \ldots, S_{t-r+1})^\mathsf{T}$, $\overline{A}_{t-1}^q = (A_{t-1}, A_{t-2}, \ldots A_{t-q})^\mathsf{T}$ and $F_S(\psi)$, $F_A(\psi)$ and $F_a(\psi)$ are time-constant matrices of appropriate dimension. This means that

$$f_t(H_t, a_t; \psi) = F_S(\psi)\overline{S}_t^r + F_A(\psi)\overline{A}_{t-1}^q + F_a(\psi)a_t.$$

Our regret estimator is $\mu_t\left(a_t \mid H_t; \widehat{\psi}\right)$, with $\widehat{\psi}$ an estimator of ψ obtained using any of the existing methods. Note that even for a correctly specified model $\widehat{\psi}$ may be inconsistent in the presence of ignored measurement error in the training data. Proposition 2 leads from the regret to an uncertain linear time-invariant model for the true state S_t and its observed version $\mathbb{S}_t$:

$$\begin{aligned} S_{t+1} &= F_S(\widehat{\psi})\overline{S}_t^r + F_A(\widehat{\psi})\overline{A}_{t-1}^q + F_a(\widehat{\psi})a_t + \varepsilon_t^{glob} \\ \mathbb{S}_t &= S_t + \varepsilon_t^m. \end{aligned} \tag{5}$$

Here ε_t^{glob} is the error caused by using $f_t(H_t, a_t; \widehat{\psi})$ instead of the true model $f_t(H_t, a_t) + \varepsilon_t^{dyn}$ as the state space model. It can be decomposed as

$$\begin{aligned} \varepsilon_t^{glob} &= f_t(H_t, a_t) - f_t(H_t, a_t; \psi) \\ &+ f_t(H_t, a_t; \psi) - f_t(H_t, a_t; \widehat{\psi}) \\ &+ \varepsilon_t^{dyn} \end{aligned}$$

The first term is the error when we assume a wrong parametric form for the model (model uncertainty), the second is the error due to estimation from the possibly noisy training data (parametric uncertainty), and the third is stochastic disturbance acting on the system which cannot be taken into account (dynamic uncertainty).

In (5) the term ε_t^m represents measurement error at time t, which we collect with ε_t^{glob} in the vector

$$w_t = \begin{pmatrix} \varepsilon_t^{glob} \\ \\ \varepsilon_t^m \end{pmatrix}. \tag{6}$$

having twice the dimension of S_t.

3.2 *Treatment strategies*

A decision strategy that aims to produce zero regret would respect $f_t(\overline{S}_t, \overline{A}_{t-1}, d_t^{opt}(\mathbb{H}_t)) = s^*$, which under (5) would allow the decision rule to be expressed as a linear combination of the elements in $\mathbb{H}_t$. In control terminology this would be described as a *deadbeat* strategy. These are little used in practice, because there is often high variability from one time step to the next and because of a lack of robustness to any form of uncertainty, including the variants summarised in Section 2.3. Uncertainty is of course the rule rather than the exception in real applications. Consequently it is common to accept some underperformance under ideal conditions in return for robustness in non-ideal circumstances.

One quick and simple alternative would be to change $Y(d)$. For example, we might replace the outcome $Y(d)$ in (1) with a penalised version

$$Y(d, \lambda) = -\sum_{t=1}^{T} \left\| S_t(d, \overline{w}_{t-1}^*) - s^* \right\|_2^2 - \lambda \| d(\mathbb{H}_{t-1}) \|_2^2, \tag{7}$$

with λ a positive constant. Here $S_t(d, \overline{w}_{t-1}^*)$ is the solution of equation (5) for the treatment strategy d, and the sequence $\overline{w}_{t-1}^* := (w_1^*, ..., w_{t-1}^*)$ is a realisation of w_t up to time $t-1$. Using (7) allows a penalty to be applied to overly-aggressive treatments, assuming that treatments are parametrized so that small absolute values are preferred, and easily adapted otherwise. Moreover, the use of a linear state-space representation (5) and the addition of a Tikhonov regularization term in equation (7) leads our approach to be robust to some level of misspecification in the way treatment action is modeled (El Ghaoui and Lebret (1997) and Sra et al. (2012) Chapter 14). At a more theoretical perspective, for a given disturbance realization $\overline{w}_T^*$, adding a quadratic penalty term to the input ensures the existence and uniqueness of the treatment strategy minimizing $Y(d, \lambda)$ as well as its continuous dependence with respect to observations. This is a classic result from Linear-Quadratic theory (see Sontag (1998)).

An alternative approach is to concentrate on directly parametrizing the decision rule itself,

say $d_t(\mathbb{H}_t, \beta)$, and then seek the β that gives the optimal policy within this reduced class of rules, with as few additional assumptions as possible. An important paper in this class is the robust procedure of Zhang et al. (2013), which builds on Zhang et al. (2012). In brief, for each candidate value of β we look for the subset of patients in a training data set whose actual treatments are consistent with the rules $d_t(\mathbb{H}_t, \beta)$ at all t. Then we find the β for which such subsets give good responses. In a little more detail, let $C_j = 1$ if all decisions for patient j match those given by $d_t(\mathbb{H}_t, \beta)$, with $C_j = 0$ otherwise. Then in its simplest form this approach would choose the value of β that maximizes

$$\frac{1}{n}\sum_{j=1}^{n}\frac{C_j Y_j}{\pi_j(\beta)},$$

where Y_j is the observed value of (1) for patient j in a training set of size n, and $1/\pi_j(\beta)$ is a weight function selected to provide consistent estimation. In finite samples of course rather few patients will have $C_j = 1$, especially if there are multiple timepoints and multiple treatment possibilities. More efficient versions are available which make use of more patient information (Zhang et al., 2013) but nonetheless the method is not realistic when there are many treatment options, and is not possible for continuous treatments.

There are a vast number of control theory approaches for bringing robustness. Many, as above, involve writing down a parametric expression for the decision rule and then either theoretically or empirically seeking the parameter values that lead to desirable performance. For example, a simple proportional integral controller is conventionally expressed in the form

$$d_t(\mathbb{H}_t, K_1, K_2) = -K_1\mathbb{S}_t - K_2\sum_{r=1}^{t-1}(\mathbb{S}_r - s^*),$$

where matrices K_1 and K_2 determine how the rule responds to short and long term responses respectively. A proportional integral derivative controller extends this by including an additional term corresponding to the current rate of change, and there are many further variants.

The method that we will concentrate on for the remainder of this paper is taken from

H^∞-synthesis in control theory. Recall that in (5) and (6) we collected all uncertainties in a vector w_t, including measurement error, estimation and modelling errors. We have not so far made any assumptions about w_t other than the additive effect in (5). This is sufficient for the H^∞ approach, which assumes there is an acceptable set $\mathscr{D}$ of decision rules and looks within this set for the rule d^{inf} that minimizes the maximum possible output-to-noise ratio over all possible non-zero realisations w_t^* of w_t. Thus

$$d^{inf} = \arg\min_{d\in\mathscr{D}} \sup_{\overline{w}^*_{T-1}\neq 0} \left\{ \frac{\sum_{t=1}^{T} \left\| S_t(d, \overline{w}^*_{t-1}) - s^* \right\|_2^2 + \lambda \| d(\mathbb{H}_{t-1}) \|_2^2}{\sum_{t=1}^{T-1} \| w_t^* \|_2^2} \right\}. \tag{8}$$

The goal of this strategy is to reach the target value while uniformly minimizing the impact of exogenous perturbation on the system. There is no claim of course that d^{inf} will always lead to good performance when the disturbances are large, but it will provide a least-bad strategy.

To progress, we obviously need to be able to solve the optimization problem in (8). This is feasible if we are willing to restrict the set of acceptable strategies $\mathscr{D}$ to linear feedback form

$$\begin{aligned} d_t(\mathbb{H}_t) &= K_1^1 \overline{\mathbb{S}}^r_{t-1} + K_1^2 \overline{A}^q_{t-1} + K_1^3 O^f_t \\ O^f_{t+1} &= K_2^1 \overline{\mathbb{S}}^r_{t-1} + K_2^2 \overline{A}^q_{t-1} + K_2^3 O^f_t, \end{aligned} \tag{9}$$

where O^f is an inner state-variable, appearing only in (9) and driven by it. Such restriction of $\mathscr{D}$ ensures we have necessary and sufficient testable criteria for the existence of a solution for the problem (8), as described in supplementary material. Further, the coefficient matrices $\{K_i^j\}$ can be derived and hence d^{inf} can be obtained, as also summarised in supplementary material. In practice, given the cost function in (8) and the state-space equation (5), easy-to-use software is available, such as the `hinfsyn` function in MATLAB. This produces a frequency-domain transfer function that can be used to derive the coefficient matrices in (9) and hence provide a dynamic treatment rule.

4. Simulations

4.1 *Experimental design*

We have defined d^{inf} to be the H^∞ strategy. Let d^{nom} be the regret-based strategy based on the nominal model assumed in (4). To compare d^{nom} and d^{inf} through simulation we generated 100 training data sets with scalar states and actions. Each set consisted of 100 longitudinal data sequences of length $T = 15$. True states S_t were generated using (3) with mean functions

$$f_t(\overline{s}_t, \overline{a}_t; \psi) = f_t^*(\overline{s}_t, \overline{a}_t; \psi) + g_t(\overline{s}_t, \overline{a}_t),$$

where g_t is included in order to simulate model misspecification. Treatments A_t were drawn uniformly between -2 and 2, and all targets s^* were set to zero. After each training data set was generated, a regret model $\mu_i\left(a_t \mid H_t; \psi\right)$ was specified and the parameters estimated from the training data using the regret-regression approach developed by Henderson et al. (2010). The two treatment policies d^{nom} and d^{inf} were then obtained and applied so as to generate a further 100 longitudinal data sequences of the same length $T = 15$. Actions were determined by policy but states generated using the same state-space model as for the training data. We modelled the different kinds of uncertainty as follows.

(1) Model misspecification. To add stochastic disturbance at time t, we took an N(0, σ^2_{dyn}) distribution for ε_t^{dyn}. In order to imitate model uncertainty, we took as assumed regret functions

$$\mu_t\left(a_t \mid H_t, \psi\right) = \left\{f_t^*(H_t, a_t; \psi) - s^*\right\}^2,$$

so that the functions g_t are missing.

(2) Measurement uncertainty. At each time t, we took $\mathbb{S}_t = S_t + \varepsilon_t^m$ where $\varepsilon_t^m \sim N(0, \sigma_m^2)$.

(3) Parametric uncertainty. Instead of ψ, we used in generating d^{nom} and d^{inf} the estimator $\widehat{\psi}$ obtained from the training data set.

We took the assumed state evolution model to be

$$f_t^*(\overline{s}_t, \overline{a}_t; \psi) = \psi_1 s_t + \psi_2 a_t + \psi_3 s_{t-1} + \psi_4 a_{t-1}, \tag{10}$$

with true parameter values (ψ_1, ψ_2, ψ_3, ψ_4) = (0.6, 0.2, 0.25, 0.15). For model uncertainty we assumed an interaction was ignored so $g_t(\overline{s}_t, \overline{a}_t) = g_1 \times s_{t-1}a_{t-1}$. We took four different values for the coefficient g_1, namely 0, 0.005, 0.02 or 0.04, We used three levels each for the variances of the stochastic perturbations and measurement noise, $\sigma^2_{dyn} = 0$, 0.1 or 0.3 and σ^2_m = 0, 0.1 or 0.3 respectively. Figure 1 gives examples of generated data under these scenarios.

[Figure 1 about here.]

For each scenario and treatment rule we estimated the quantities

$$\mathrm{ERR}\left\{d\right\} = -\mathbb{E}_{g_t, \sigma^2_{dyn}, \sigma^2_m}\left\{Y(d)\right\}, \quad \mathrm{VERR}\left\{d\right\} = \mathrm{var}_{g_t, \sigma^2_{dyn}, \sigma^2_m}\left\{Y(d)\right\}. \tag{11}$$

The first quantity, ERR, is the negative of the original criterion, and so ideally minimized by d^{opt}. The second, VERR, is the variance of the outcome computed on the patient set.

4.2 *Sensitivity analysis and adaptive selection method for* λ

We now consider the role of the penalty parameter λ in the H^∞ strategy (8), and denote the decision rule as d^{inf}_λ so as explicitly to acknowledge the dependence on λ. In the upper section of Figure 2 we show $\mathrm{ERR}\left\{d^{inf}_\lambda\right\}$ for different values of $\left(g_1, \sigma^2_{dyn}, \sigma^2_m\right)$ as λ is varied. The pattern is similar in all three panels, with the quality of d^{inf}_λ quickly increasing with λ before starting to slowly decreases later. This calls for an adaptive method to select λ.

Given training data $\mathbb{H}^j_T = \left(\overline{\mathbb{S}}^j_T, \overline{A}^j_{T-1}\right)$ for $j = 1, 2, \ldots n$, we propose an algorithm based on scoring the number of occasions at which the H^∞ strategy is retrospectively assumed to provide a better decision than that chosen in practice. We total, over all time points t and all training data individuals j, the number of occasions on which either

$$\mathbb{S}^j_{t+1} > s^* \text{ and } d^{inf}_\lambda(\mathbb{H}^j_t) < A^j_t$$

or

$$\mathbb{S}_{t+1}^j < s^* \text{ and } d_\lambda^{inf}(\mathbb{H}_t^j) > A_t^j.$$

We retain the value of λ that corresponds to the highest score. To illustrate, the lower section of Figure 2 shows how the scores change with λ. The optimal values of λ in the lower plots would all lead to acceptable decision rules, in that the corresponding ERR $\left\{d_\lambda^{inf}\right\}$ in the upper plots are in the low, flat, regions.

[Figure 2 about here.]

4.3 *Results*

Results are presented in Table 1. Those for d^{inf} are based on use of the algorithm of the previous subsection for choice of penalty λ. The quantity denoted "Ratio" is the mean value of

$$\sum_{t=1}^{T-1} \frac{g_t(\overline{S}_t, \overline{A}_t)^2}{f_t^*(\overline{S}_t, \overline{A}_t)^2},$$

which is the ratio of the unknown over the known part of the model computed on the training set. It is used to quantify the level of misspecification.

When there is little or no model misspecification or uncertainty, d^{nom} slightly outperforms d^{inf}, as expected, though the performance of d^{inf} is still good. However, as noise and misspecification levels increase, d^{inf} is much better than d^{nom} in maintaining the true S_t close to target s^*. Further, the low values of VERR for d^{inf} suggest uniform good performance despite high inter-subject variability.

[Table 1 about here.]

5. Applications

5.1 *Ventilation chamber experiment*

We tested the d^{nom} and d^{inf} strategies using an experimental ventilation chamber (Taylor, 2004), which allows us to generate training data for the modelling and estimation phase, and then test data following any of the recommended strategies. Unlike a simulation experiment there is no known true model, and unlike a standard application we have the opportunity to test different strategies.

The aim is to control the internal temperature of the chamber by adjusting the voltage applied to a heating element. Internal temperature is additionally affected by external temperature, which is not under our control, and by air flow, which is determined in part by external conditions and in part by two fans inside the chamber, one outlet and one inlet. In the experiments we considered air flow to represent environmental conditions, with the outlet fan used to give run-to-run variability and the inlet fan used to add time-varying noise.

In the following our true state S_t is the difference between ventilation chamber temperature and outside temperature, A_t is the chosen input voltage to the heater, and ε_t^{dyn} represents exogenous disturbance due to air flow and potential other unmeasured variables.

To demonstrate use of the heater, the target was set to be 6 C warmer than the outside temperature i.e $s^* = 6$. The first step was to derive from training data an appropriate regret model within the class (4) and estimate its parameters. As training data we generated 30 trajectories, each of 15 sampling times, and with the input heater voltage A_t set at either 2V or 4V. The initial choice was random with equal probabilities, and then at samples 5 and 10 we randomly either changed to the alternative voltage or left it at the current value. The outlet fan input was set at a different level for each trial, selected randomly in the interval 1–3V and held constant throughout the trial. For time-varying disturbances we changed

the voltage to the inlet fan each second, with the changes drawn from a centered Gaussian distribution with standard deviation 0.2V. Five examples of state and action training data are given in Figure 3.

[Figure 3 about here.]

We found that a very simple regret model

$$\mu_t\left(a_t \mid H_t; \psi\right) = \left(s^* - \psi_1 S_t - \psi_2 a_t\right)^2$$

was adequate for the data, with $\widehat{\psi}_1 = 0.7$ and $\widehat{\psi}_2 = 0.5$ estimated by using the regret regression approach proposed by Henderson et al. (2010). From this, we obtain the regret-based decision rule

$$d_t^{nom}\left(\mathbb{H}_t \mid \widehat{\psi}\right) = \frac{s^* - \widehat{\psi}_1 \mathbb{S}_t}{\widehat{\psi}_2}$$

and the uncertain state-space model

$$S_{t+1} = \widehat{\psi}_1 S_t + \widehat{\psi}_2 a_t + \varepsilon_t^{glob},$$

which is used to define d^{inf} using the methodology of Section 3.

Next we performed additional trials in which the input voltage, now on a continuous scale, was selected using either the regret or H^∞ strategies, d^{nom} and d^{inf} respectively. We took two versions of the H^∞ strategy, one with $\lambda = 0.01$ and one with $\lambda = 0.001$. In each of the three cases we performed ten new trials, this time of 30 sampling points each. In all cases the target remained at $s^* = 6$ and we set the fan inputs in the same way as for the training data.

Results are summarised in Table 2 for both the total response over the whole test period and over the 15 final sampling times. The latter was chosen to represent steady state after transition from the initial conditions. In terms of average mean square error (MSE) and mean maximum absolute error (MAE), the H^∞ controllers gave better results for both choices of λ

than the regret rule d^{nom}. The difference in performance was more pronounced at the steady state, that is, once the system is stabilized near the target value.

[Table 2 about here.]

5.2 *Warfarin data analysis*

We now consider a more typical application of A-learning methodology, which is the choice of drug dose for patients prescribed Warfarin as long-term anticoagulation prophylaxis. At each observation time t we take the measured state $\mathbb{S}_t$ to be the log-transform of the blood clotting speed, measured through the international normalized ratio (Baglin et al., 2006). We take the input A_t to be the prescribed Warfarin dose in mg. If clotting time is too low there is a risk of thrombosis and an increase in dose is suggested. If clotting time is too high there is a risk of hemmorrhage and a decrease of dose might be warranted. A typical example of patient history is given in Figure 4.

[Figure 4 about here.]

Our purpose is to make a retrospective comparison between dose levels that could have been suggested using regret-based or H^{∞} strategies with those actually chosen by the healthcare providers. The available data consists of the records of 152 patients receiving Warfarin anticoagulation in Newcastle-upon-Tyne during 2013. Our analyses of the records indicated no effect of time intervals between clinic visits and so we consider the data as being in discrete time, indexed by clinic visit number. For our approach we need a sufficient number of observations for each patient, so we narrowed the analysis to the 120 patients with at least 20 successive measurements. To have data sequences of the same length for comparison, we took only the first $T = 20$ entries for each of these patients. We randomly divided the patients into a training set of 20 patients for calibrating the required models and a testing set of 100 patients for comparing treatment strategies. This balance was selected because of

our focus on treatment comparison rather than modelling and inference. We repeated the random selection ten times.

The international normalized ratio is usually considered as acceptable if it lies between two and three, so we choose as target value $s^* = \log(2.5)$. For each training set we fit the very simple regret model

$$\mu_t(a_t \mid H_t, \psi) = (s^* - \psi_1 S_t - \psi_2 a_t)^2 .$$

More complex functions taking into account past outputs and inputs showed little consistency between the training sets, which is unsurprising given their small size. Once ψ has been estimated, we define d^{nom} and d^{inf} as presented in Section 3 for each patient and for each decision time, using $\lambda = 0.001$ for d^{inf}.

For the comparison, we divided the decisions in the test data into three groups, corresponding to when the healthcare provider chose a dose a_t^{hp} at time t that seemed to be good, that seemed to be too low, and that seemed to be too high. We assumed a decision to be good if measured blood clotting speed at the next observation time was within the target range of log(2) to log(3). There were 52% of decisions in this category and in these cases we quantified the difference between the chosen dose and each of $d = d^{nom}$ and $d = d^{inf}$ by

$$GD_{j,t}(d) = \frac{|d(\mathbb{H}_{j,t}) - a_{j,t}^{hp}|}{2\{d(\mathbb{H}_{j,t}) + a_{j,t}^{hp}\}}$$

for patient j at time t.

For 18% of decision times a_t^{hp} was followed by a clotting time above the upper limit of the target range. Here we assume that a lower dose would have been preferred and use

$$HD_{j,t}(d) = 1\{d(\mathbb{H}_{j,t}) < a_{j,t}^{hp}\}$$

as performance measure. For the remaining 30% of decisions a_t^{hp} was followed by a clotting time below the lower limit of the target range and for these we assume that a higher dose would have been preferred and measure performance through

$$LD_{j,t}(d) = 1\{d(\mathbb{H}_{j,t}) > a_{j,t}^{hp}\}.$$

Table 3 presents the mean values of these statistics, averaged over patients and decision times, for each of our ten random splits of the data. When the actual decision is acceptable d^{nom} gives results that are on average closer to those selected by the healthcare provider than are decisions d^{inf}. In both cases the mean differences are low however. Otherwise, if the actual decision is poor then d^{inf} outperformed d^{nom} in all ten trials. The actual d^{inf} decisions were invariably more cautious than d^{nom} when the latter was to recommend a relatively large change in dose.

In the absence of model or dynamic uncertainty the regret-based policy would be optimal. In this case the estimated proportion of visits for which the INR would be in the target range is 67%, which is to be compared with the previously-mentioned observed value of 52%.

[Table 3 about here.]

6. Discussion

We have tried to show how H^∞ methods can be of use in dynamic treatment allocation, after the modelling and estimation stages. We envisage that these and other control methods will be of most use in applications involving quantitative choice of drug dosage for patients with chronic conditions, in which the aim is to maintain a biomarker at, or close to, some target level. The warfarin application of Section 5.2 provides a typical example. Although not considered in the paper, covariates can easily be taken into account because the H^∞ method simply takes as input an assumed model. Thus we can build a covariate-dependent model on training data, and simply input the model with individual-specific covariates at the decision stage. We do not propose control methods when treatment options are binary or categorical or when there are few decision times.

For simplicity we have concentrated on the simple cost function (1). More involved, perhaps asymmetric, costs might be considered in future work, though sometimes a simple

transformation of the response data might be sufficient, such as the use of the logarithm of blood clotting speed in Section 5.2. Function (1) as given leads naturally to the little-used deadbeat control and would probably not be selected in engineering applications because of lack of robustness, as seen in our simulations. We will investigate alternative cost functions, including integral-of-error components, in future work. We also concentrated in this work on linear regret models with constant coefficients but note that the methods can be extended readily to linear Q-learning models. In principle extension to non-linear models and to include time-varying coefficients is possible, though H^∞ control is less well developed for these cases. There are, however, other robust control methods that could be transportable, including the non-minimal state-space family (Taylor et al., 2013). Model predictive control is another area that might very fruitfully be exploited by statisticians. In this approach, at each decision time a sequence of future decisions is planned rather than just the next, and the sequence is allowed to change dynamically as new information is provided. It is in some ways close to the history-adjusted marginal structural modelling methods of van der Laan et al. (2005) and Petersen et al. (2007), though presented in quite a different manner. In the other direction, we suspect that there is not widespread familiarity within the control community of the latest statistical methods for dealing with noisy or missing observations, or for the careful consideration of causal effects (Wilson et al., 2018). As control methods are now being used far beyond traditional engineering applications, and in particular in biomedical areas where data may be sparse and repeatability is problematic, full attention to modelling and estimation is becoming ever more important and requires properly grounded and efficient statistical methodology.

Acknowledgements

We are grateful for the helpful suggestions of two referees and an associate editor. This work was supported by EPSRC grant EP/M015637/1.

Data Availability Statement

The data (Matlab code and real data) that supports the findings of this study are available in the supplementary material of this article

References

Baglin, T. P., Keeling, D. M., and Watson, H. G. (2006). Guidelines on oral anticoagulation (warfarin): third edition - 2005 update. *British Journal of Haematology* **132,** 277–285.

Barrett, J., Henderson, R., and Rosthøj, S. (2014). Doubly robust estimation of optimal dynamic treatment regimes. *Statistics in Biosciences* **6,** 244–260.

Bekiroglu, K., Lagoa, C., Murphy, S. A., and Lanza, S. T. (2017). Control engineering methods for the design of robust behavioral treatments. *IEEE Transactions on Control Systems Technology* **3,** 979–990.

Chakrabarty, A., Zavitsanou, S., F., D. I., and Dassau, E. (2017). Event triggered model predictive control for embedded artificial pancreas systems. *IEEE Transactions on Biomedical Engineering.* .

Chakraborty, B., Laber, E., and Zhao, Y. (2013). Inference for optimal dynamic treatment regimes using an adaptive m-out-of-n bootstrap scheme. *Biometrics* **69,** 714–723.

Chakraborty, B. and Moodie, E. (2013). *Statistical Methods for Dynamic Treatment Regimes.* Springer, New York.

Chen, G., Zeng, D., and Kosorok, M. (2016). Personalized dose finding using outcome weighted learning. *Journal of the American Statistical Association* **111,** 1509–547.

Deshpande, S., Nandola, N. N., Rivera, D. E., and Younger, J. W. (2014). Optimized treatment of fibbromyalgia using system identfication and hybrid model predictive control. *Control Engineering Practice* **33,** 161–173.

Doyle, J., Glover, K., Khargonekar, P., and Francis, B. (1989). State-space solutions to

standard h2 and h-infinity control problems. *IEEE Transactions on Automatic Control* **34,** 831–847.

El Ghaoui, L. and Lebret, H. (1997). Robust solutions to least-squares problems with uncertain data. *SIAM Journal on matrix analysis and applications* **18,** 1035–1064.

Glover, K. and Doyle, J. (1988). State-space formulae for all stabilizing controllers that satisfy an h-infinity-norm bound and relations to risk sensitivity. *Systems and Control Letters* **11,** 167–172.

Henderson, R., Ansell, P., and Alsibani, D. (2010). Regret-regression for optimal dynamic treatment regimes. *Biometrics* **66,** 1192–1201.

Henderson, R., Ansell, P., and Alsibani, D. (2011). Optimal dynamic treatment methods. *Revstat Statistical Journal* **9,** 19–36.

Hooker, G., Lin, K., and Rogers, B. (2015). Control theory and experimental design in diffusion processes. *SIAM* **3,** 234–164.

Laber, E. and Zhao, Y. (2015). Tree-based methods for individualized treatment regimes. *Biometrika* **102,** 501–514.

Laber, E. B., Linn, K. A., and Stefanski, L. A. (2014). Interactive model building for Q-learning. *Biometrika* **101,** 831—-847.

Linn, K., Laber, E., and Stefanski, L. (2017). Interactive Q-learning for quantiles. *Journal of the American Statistical Association* **112,** 638–649.

Moodie, E., Dean, N., and Sun, Y. (2014). Q-learning: flexible learning about useful utilities. *Statistics in Biosciences* **6,** 223–243.

Moodie, E., T.S., R., and Stephens, D. (2007). Demystifying optimal dynamic treatment regimes. *Biometrics* **63,**.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society Series B* **65,** 331–355.

Orellana, J. (2010). Optimal drug scheduling for hiv therapy efficiency improvement. *Biomedical signal Processing and Control* **6,** 376–386.

Orellana, L., Rotnitzky, A., and Robins, J. (2010). Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes. part i: main content. *International Journal of Biosciences* **6,** article 8.

Petersen, M., Deeks, S., Martin, J., and van der Laan, M. (2007). History-adjusted marginal structural models for estimating time-varying effect modification. *American Journal of Epidemiology* **166,** 985–993.

Pronzato, L. (2008). Optimal experimental design and some related control problems. *Automatica* **44,** 303–325.

Qian, M. and Murphy, S. A. (2008). Performance guarantees for individualized treatment rules. *Annals of Statistics* **27,** 1180–1210.

Rich, B., Moodie, E., and Stephens, D. (2014). Simulating sequential multiple assignment randomized trials to generate optimal personalized warfarin dosing strategies. *Clinical trials* **11,** 435–444.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling* **7,** 1393–1512.

Robins, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In Lin, D. Y. and Heagerty, P., editors, *Proceedings of the Second Symposium on Biostatistics*, pages 189–326. New York: Springer.

Rosthøj, S., Fullwood, C., Henderson, R., and Stewart, S. (2006). Estimation of optimal dynamic anticoagulation regimes from observational data: a regret-based approach. *Statistics in Medicine* **4197-4215,**.

Rosthøj, S., Keiding, N., and Schmiegelow, K. (2012). Estimation of dynamic treatment

strategies for maintenance therapy of children with acute lymphoblastic leukemia: an application of history-adjusted marginal structural models. *Statistics in Medicine* **31,** 470–488.

Schulte, P., Tsiatis, A., Laber, E., and Davidian, M. (2014). Q- and A-learning methods for estimating optimal dynamic treatment regimes. *Statistical Science* **69,** 640–661.

Song, R., Wang, W., Zeng, D., and Kosorok, M. (2015). Penalized Q-learning for dynamic treatment regimens. *Statistica Sinica* **25,** 901–920.

Sontag, E. (1998). *Mathematical Control Theory: Deterministic finite-dimensional systems.* Springer-Verlag: New-York.

Sra, S., Nowozin, S., and Wright, S. J. (2012). *Optimization for machine learning.* Mit Press.

Taylor, J. (2004). Environmental test chamber for the support of learning and teaching in intelligent control. *International Journal of Electrical Engineering Education* **41,** 375–387.

Taylor, J., Young, P., and Chotai, A. (2013). *True Digital Control: Statistical Modelling and Non-Minimal State Space Design.* John Wiley and Sons: Chichester.

van der Laan, M., Petersen, M., and Joffe, M. (2005). History-adjusted marginal structural models and statically-optimal dynamic treatment regimens. *International Journal of Biostatistics* **1,**.

Wallace, M. and Moodie, E. (2015). Doubly-robust dynamic treatment regimen estimation via weighted least squares. *Biometrics* **71,** 636–644.

Wallace, M., Moodie, E., and Stephens, D. (2016). Model assessment in dynamic treatment regimen estimation via double robustness. *Biometrics* **72,** 855–864.

Wilson, E., Clairon, Q., Henderson, R., and Taylor, J. (2018). Dealing with observational data in control. *Annual Reviews in Control* **46,** 94–106.

Xu, Y., Müller, P., Wahed, A., and Thall, P. (2016). Bayesian nonparametric estimation for

dynamic treatment regimes with sequential transition times. *Journal of the American Statistical Association* **111,** 921–950.

Zhang, B., Tsiatis, A., Laber, E., and Davidian, M. (2012). A robust method for estimating optimal treatment regimes. *Biometrics* **68,** 1010–1018.

Zhang, B., Tsiatis, A., Laber, E., and Davidian, M. (2013). Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika* **100,** 681–694.

Zhang, S. and Xu, X. (2016). Dynamic analysis and optimal control for a model of hepatitis c with treatment. *Commun Nonlinear Sci Numer Simulat* **46,** 14–25.

Zhao, Y., Zeng, D., Laber, E., Song, R., Yuan, M., and Kosorok, M. (2015). Doubly robust learning for estimating individualized treatment with censored data. *Biometrika* **102,** 151–168.

Zhao, Y., Zeng, D., Rush, A., and Kosorok, M. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association* **107,** 1106=1118.

Zhou, X., Mayer-Hamblett, N., Khan, U., and Kosorok, M. (2017). Residual weighted learning for estimating individualized treatment rules. *Journal of the American Statistical Association* **112,** 169–187.

Supporting Information

Web Appendices referenced in Sections 2.3 and 3.2 are available with this paper at the Biometrics website on Wiley Online Library.

Received November 2018. *Revised November* 2019. *Accepted March* 2020.

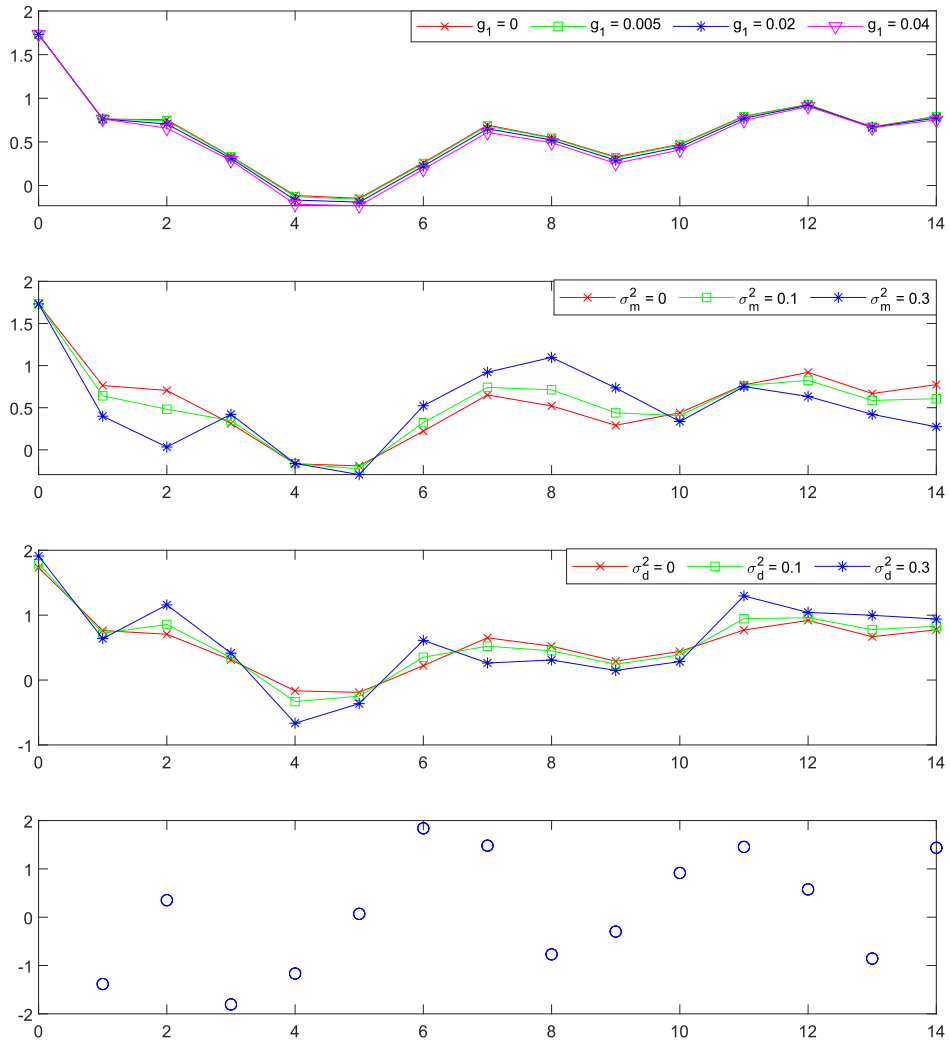

Figure 1. Example of generated observations for a patient when one value among $\left(g_1, \sigma^2_{dyn}, \sigma^2_m\right)$ is varied while the others are set to 0. The upper plots show the observed responses $\mathbb{S}_t$ and lower show the actions A_t, which are the same in all examples. This figure appears in color in the electronic version of this article, and any mention of color refers to that version.

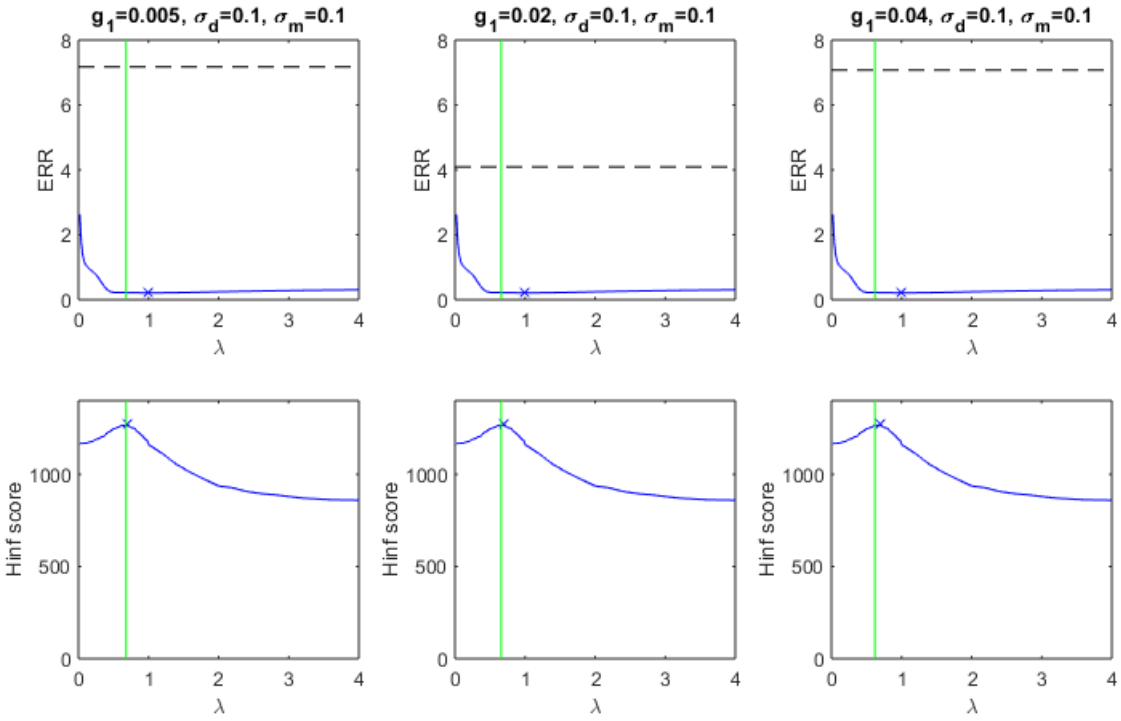

Figure 2. The achieved ERR $\left\{d_\lambda^{inf}\right\}$ (upper panels) and decision scores (lower panels) as the penalty λ is varied. The dashed horizontal lines are the values obtained with d^{nom} and the vertical lines correspond to the λ values giving the highest score.

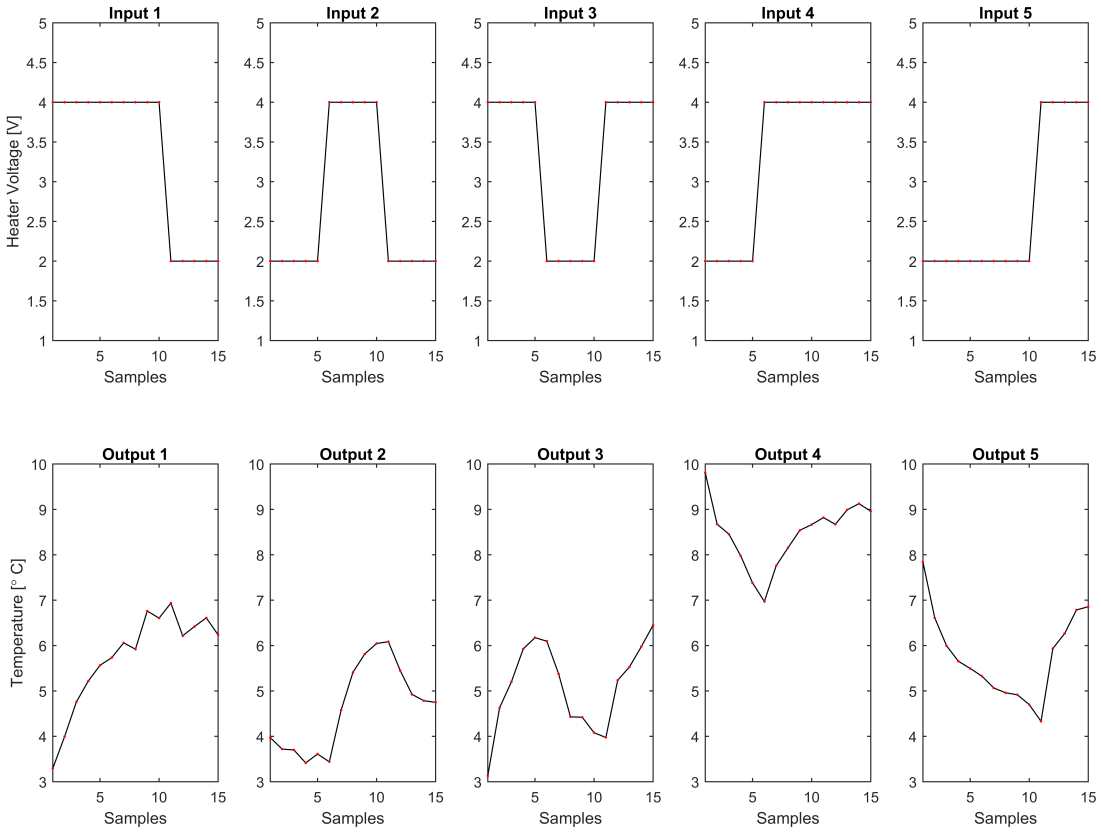

Figure 3. Example training trials for ventilation chamber experiment.

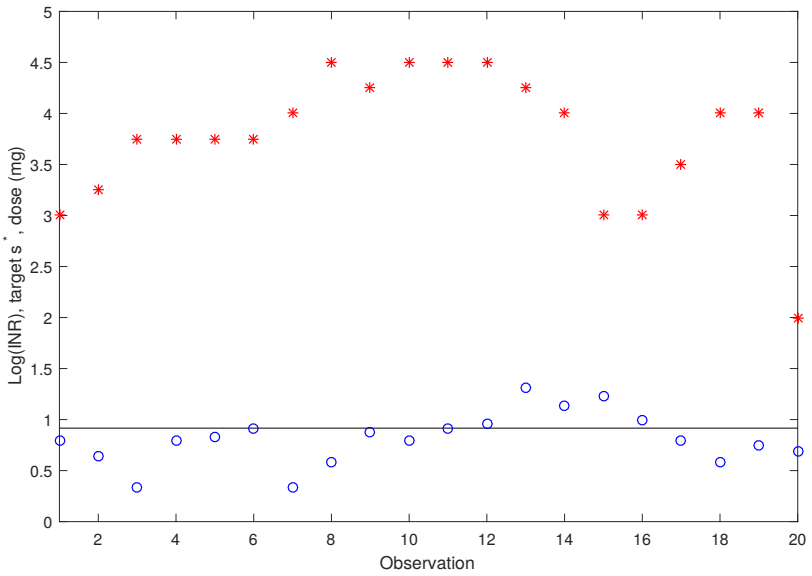

Figure 4. An example of patient history: log(INR) (circle), prescribed dose (star) and target value s^* (solid line).

Table 1
Simulation results for both treatment strategies for model (10). ERR and VERR are defined in Eq. (11).

	$(\sigma^2_{dyn}, \sigma^2_m)$								
	(0, 0)	(0, 0.1)	(0, 0.3)	(0.1, 0)	(0.3, 0)	(0.1, 0.1)	(0.1, 0.3)	(0.3, 0.1)	(0.3, 0.3)
(a) $g_1 = 0$									
Ratio	0	0	0	0	0	0	0	0	0
ERR $\{d^{inf}\}$	0.03	0.03	0.08	0.22	1.74	0.23	0.28	1.75	1.84
ERR $\{d^{nom}\}$	0.00	0.97	10.36	4.22	7.67	6.80	11.36	9.08	9.35
VERR $\{d^{inf}\}$	0.00	0.00	0.00	0.02	0.840	0.02	0.03	0.84	0.99
VERR $\{d^{nom}\}$	0.00	9.90	98.75	48.08	48.95	64.08	99.84	66.30	122.89
(b) $g_1 = 0.005$									
Ratio	0.24	0.24	0.24	0.24	0.24	0.24	0.24	0.26	0.26
ERR $\{d^{inf}\}$	0.03	0.03	0.08	0.22	1.75	0.23	0.28	1.76	1.85
ERR $\{d^{nom}\}$	0.00	1.64	11.40	4.30	6.80	7.17	7.66	6.53	11.20
VERR $\{d^{inf}\}$	0.00	0.00	0.00	0.02	0.85	0.02	0.03	0.86	1.03
VERR $\{d^{nom}\}$	0.00	19.81	139.15	72.41	47.24	73.94	90.88	43.04	87.23
(c) $g_1 = 0.02$									
Ratio	0.96	0.97	0.94	0.95	0.94	0.96	0.99	0.93	0.96
ERR $\{d^{inf}\}$	0.01	0.03	0.08	0.21	1.74	0.22	0.28	1.74	1.82
ERR $\{d^{nom}\}$	0.00	2.20	11.10	4.10	6.12	4.09	10.03	4.34	7.08
VERR $\{d^{inf}\}$	0.00	0.00	0.00	0.01	0.85	0.02	0.03	0.82	0.96
VERR $\{d^{nom}\}$	0.000	25.51	89.48	55.07	49.84	44.56	120.07	37.35	24.26
(d) $g_1 = 0.04$									
Ratio	1.89	1.89	1.88	1.99	2.15	1.92	1.87	1.79	1.93
ERR $\{d^{inf}\}$	0.01	0.02	0.09	0.21	1.72	0.22	0.28	1.69	1.80
ERR $\{d^{nom}\}$	0.33	3.44	10.56	4.40	6.00	3.90	9.87	5.89	8.52
VERR $\{d^{inf}\}$	0.00	0.00	0.00	0.01	0.79	0.01	0.03	0.75	0.89
VERR $\{d^{nom}\}$	3.45	43.15	104.40	42.01	55.46	34.62	87.99	59.23	80.17

Table 2

Summary of test data, using complete sequences (Total) or the final 15 sampling points only (SS), assumed to represent steady state. MSE is mean squared error between state and target, and MAE is the maximum absolute error.

Method	Test no.	Total		SS	
		MSE	MAE	MSE	MAE
Nominal	1	1.61	4.69	0.18	0.62
	2	2.16	5.34	0.21	0.91
	3	1.43	4.34	0.32	1.03
	4	1.52	4.75	0.22	0.73
	5	2.73	5.51	0.14	0.67
	6	1.37	4.61	0.26	0.90
	7	1.56	4.57	0.17	0.73
	8	1.68	4.78	0.25	0.73
	9	1.64	4.69	0.20	0.61
	10	2.18	5.45	0.14	0.77
	Mean	1.79	4.79	0.21	0.77
H^∞, $\lambda = 0.01$	1	1.49	4.41	0.07	0.50
	2	1.27	4.43	0.14	0.75
	3	1.23	4.22	0.17	0.78
	4	1.11	4.28	0.15	0.72
	5	1.22	4.03	0.15	0.63
	6	1.19	4.67	0.13	0.52
	7	1.67	4.83	0.13	0.73
	8	1.38	4.57	0.11	0.50
	9	1.20	4.09	0.10	0.50
	10	1.41	4.82	0.11	0.56
	Mean	1.31	4.43	0.13	0.62
H^∞, $\lambda = 0.001$	1	0.71	3.63	0.10	0.58
	2	1.17	4.16	0.21	0.85
	3	1.27	4.25	0.15	0.72
	4	1.39	4.64	0.12	0.60
	5	1.81	4.77	0.11	0.61
	6	1.09	4.25	0.15	0.60
	7	1.67	4.69	0.09	0.57
	8	1.49	4.43	0.09	0.44
	9	2.55	5.39	0.09	0.48
	10	1.22	4.69	0.13	0.69
	Mean	1.44	4.49	0.12	0.61

Table 3

Comparison measures between actual and model-based dose levels for Warfarin data. Columns $GD(.)$ measure average relative distance between good true dose decisions and those proposed by $d = d^{nom}$ and $d = d^{inf}$. Columns $HD(.)$ and $LD(.)$ relate to true decisions which led to a dose which is assumed to be too high or too low respectively. The tabulated values in these columns are the proportions of occasions in which the model-based dose decision is assumed to be better. Low values of GD are preferred, high values of LD and HD.

Trial	$GD(d^{nom})$	$GD(d^{inf})$	$HD(d^{nom})$	$HD(d^{inf})$	$LD(d^{nom})$	$LD(d^{inf})$
1	0.009	0.012	0.42	0.53	0.70	0.77
2	0.009	0.013	0.47	0.59	0.66	0.74
3	0.010	0.014	0.70	0.73	0.44	0.54
4	0.009	0.013	0.45	0.56	0.61	0.70
5	0.009	0.014	0.43	0.54	0.65	0.71
6	0.008	0.012	0.46	0.55	0.66	0.73
7	0.009	0.013	0.48	0.58	0.62	0.70
8	0.010	0.014	0.57	0.68	0.57	0.63
9	0.010	0.014	0.47	0.59	0.66	0.75
10	0.009	0.012	0.45	0.54	0.64	0.75

Adaptive Treatment and Robust Control

Q. Clairon, R. Henderson, N.J. Young, E.D. Wilson, and C.J. Taylor

Supplementary Material

Proof of Propositions 1 and 2

To compute the regret function we need

$$\begin{aligned} Y(\underline{d}_t^{opt}) &= -\sum_{r=1}^{t} \|S_r - s^*\|_2^2 - \sum_{r=t+1}^{T} \|S_r - s^*\|_2^2 \\ &= -\sum_{r=1}^{t} \|S_r - s^*\|_2^2 - \sum_{r=t}^{T-1} \left\| f_r(\overline{S}_r, \left\{ \overline{A}_{t-1}, d_{r\setminus(t-1)}^{opt} \right\}) + \varepsilon_r^{dyn} - s^* \right\|_2^2 \end{aligned}$$

and

$$\begin{aligned} Y(a_t, \underline{d}_{t+1}^{opt}) &= -\sum_{r=1}^{t} \|S_r - s^*\|_2^2 - \|S_{t+1} - s^*\|_2^2 - \sum_{r=t+2}^{T} \|S_r - s^*\|_2^2 \\ &= -\sum_{r=1}^{t} \|S_r - s^*\|_2^2 - \left\| f_t(H_t, a_t) + \varepsilon_t^{dyn} - s^* \right\|_2^2 \\ &\quad - \sum_{r=t+1}^{T-1} \left\| f_r(\overline{S}_r, \left\{ \overline{A}_{t-1}, a_t, d_{r\setminus t}^{opt} \right\}) + \varepsilon_r^{dyn} - s^* \right\|_2^2 . \end{aligned}$$

On conditioning upon H_t and taking expectations we obtain

$$\mathbb{E}\left[Y(\underline{d}_t^{opt}) \mid H_t \right] = -\sum_{r=1}^{t} \|S_r - s^*\|_2^2 - \sum_{r=t}^{T-1} \mathbb{E}\left[\left\| f_r(\overline{S}_r, \left\{ \overline{A}_{t-1}, d_{r\setminus(t-1)}^{opt} \right\}) + \varepsilon_r^{dyn} - s^* \right\|_2^2 \mid H_t \right]$$

and

$$\begin{aligned} \mathbb{E}\left[Y(a_t, \underline{d}_{t+1}^{opt}) \mid H_t \right] &= -\sum_{r=1}^{t} \|S_r - s^*\|_2^2 - \mathbb{E}\left[\left\| f_t(H_t, a_t) + \varepsilon_t^{dyn} - s^* \right\|_2^2 \mid H_t \right] \\ &\quad - \sum_{r=t+1}^{T-1} \mathbb{E}\left[\left\| f_r(\overline{S}_r, \left\{ \overline{A}_{t-1}, a_t, d_{r\setminus t}^{opt} \right\}) + \varepsilon_r^{dyn} - s^* \right\|_2^2 \mid H_t \right] . \end{aligned}$$

Hence

$$\begin{aligned}
\mu_t\left(a_t \mid H_t\right) &= -\sum_{r=t}^{T-1}\mathbb{E}\left[\left\|f_r(\overline{S}_r,\left\{\overline{A}_{t-1},d^{opt}_{r\backslash(t-1)}\right\})+\varepsilon^{dyn}_r-s^*\right\|_2^2 \mid H_t\right] \\
&\quad +\mathbb{E}\left[\left\|f_t(H_t,a_t)+\varepsilon^{dyn}_t-s^*\right\|_2^2 \mid H_t\right] \\
&\quad +\sum_{r=t+1}^{T-1}\mathbb{E}\left[\left\|f_r(\overline{S}_r,\left\{\overline{A}_{t-1},a_t,d^{opt}_{r\backslash t}\right\})+\varepsilon^{dyn}_r-s^*\right\|_2^2 \mid H_t\right] \\
&= \sum_{r=t+1}^{T-1}\mathbb{E}\left[\left\|f_r(\overline{S}_r,\left\{\overline{A}_{t-1},a_t,d^{opt}_{r\backslash t}\right\})+\varepsilon^{dyn}_r-s^*\right\|_2^2 \mid H_t\right] \\
&\quad -\sum_{r=t+1}^{T-1}\mathbb{E}\left[\left\|f_r(\overline{S}_r,\left\{\overline{A}_{t-1},d^{opt}_{r\backslash(t-1)}\right\})+\varepsilon^{dyn}_r-s^*\right\|_2^2 \mid H_t\right] \\
&\quad +\mathbb{E}\left[\left\|f_t(H_t,a_t)+\varepsilon^{dyn}_t-s^*\right\|_2^2 \mid H_t\right]-\mathbb{E}\left[\left\|f_t(H_t,d^{opt}_t(H_t))+\varepsilon^{dyn}_t-s^*\right\|_2^2 \mid H_t\right].
\end{aligned}$$

By Assumption A2, ε^{dyn}_t is independent of H_t and so

$$\begin{aligned}
\mathbb{E}\left[\left\|f_t(H_t,a_t)+\varepsilon^{dyn}_t-s^*\right\|_2^2 \mid H_t\right] &= \mathbb{E}\left[\|f_t(H_t,a_t)-s^*\|_2^2 \mid H_t\right]+\mathbb{E}\left[\left\|\varepsilon^{dyn}_t\right\|_2^2 \mid H_t\right] \\
&= \|f_t(H_t,a_t)-s^*\|_2^2+\mathbb{E}\left[\left\|\varepsilon^{dyn}_t\right\|_2^2\right].
\end{aligned}$$

Similarly

$$\mathbb{E}\left[\left\|f_t(H_t,d^{opt}_t(H_t))+\varepsilon^{dyn}_t-s^*\right\|_2^2 \mid H_t\right]=\left\|f_t(H_t,d^{opt}_t(H_t))-s^*\right\|_2^2+\mathbb{E}\left[\left\|\varepsilon^{dyn}_t\right\|_2^2\right].$$

Taking the difference of these last two quantities, we have indeed as Proposition 1 claims

$$\begin{aligned}
\mu_t\left(a_t \mid H_t\right) &= \|f_t(H_t,a_t)]-s^*\|_2^2-\left\|f_t(H_t,d^{opt}_t(H_t))-s^*\right\|_2^2 \\
&\quad +\sum_{r=t+1}^{T-1}\mathbb{E}\left[\left\|f_r(\overline{S}_r,\left\{\overline{A}_{t-1},a_t,d^{opt}_{r\backslash t}\right\})+\varepsilon^{dyn}_r-s^*\right\|_2^2 \mid H_t\right] \\
&\quad -\sum_{r=t+1}^{T-1}\mathbb{E}\left[\left\|f_r(\overline{S}_r,\left\{\overline{A}_{t-1},d^{opt}_{r\backslash(t-1)}\right\})+\varepsilon^{dyn}_r-s^*\right\|_2^2 \mid H_t\right].
\end{aligned}$$

Turning to Proposition 2, let us define a blip function for any strategy $d^\dagger$ as

$$\begin{aligned}\nu_t^{d^\dagger}(a_t \mid H_t) &= \|f_t(H_t,a_t)-s^*\|_2^2 - \left\|f_t(H_t,d_t^\dagger(H_t))]-s^*\right\|_2^2 \\ &+ \sum_{r=t+1}^{T-1}\mathbb{E}\left[\left\|f_r(\overline{S}_r,\left\{\overline{A}_{t-1},a_t,d^\dagger_{r\setminus t}\right\})+\varepsilon_r^{dyn}-s^*\right\|_2^2 \mid H_t\right] \\ &- \sum_{r=t+1}^{T-1}\mathbb{E}\left[\left\|f_r(\overline{S}_r,\left\{\overline{A}_{t-1},d^\dagger_{r\setminus(t-1)}\right\})+\varepsilon_r^{dyn}-s^*\right\|_2^2 \mid H_t\right].\end{aligned}$$

By construction we have $\nu_t^{d^{opt}}(a_t \mid H_t) = \mu_t(a_t \mid H_t)$, and for all strategies $d^\dagger$ not necessarily optimal $\nu_t^{d^\dagger}(a_t \mid H_t) \geqslant \mu_t(a_t \mid H_t)$. Thus necessarily $\nu_t^{d^\dagger}\left(d^\dagger(H_t) \mid H_t\right) = 0$ implies that $d^\dagger(H_t) = d^{opt}(H_t)$. Now, we introduce the treatment strategy d^* defined by

$$f_r(\overline{s}_r,\overline{a}_{r-1},d_r^*(\overline{s}_r,\overline{a}_{r-1})) = s^*$$

for $t \leqslant r < T-1$ and for all $\overline{s}_r \in \overline{\mathscr{S}}_r$ and $\overline{a}_{r-1} \in \overline{\mathscr{A}}_{r-1}$. Existence of such a strategy is guaranteed by Assumption A3. Hence, for $r > t$,

$$\mathbb{E}\left[\left\|f_r(\overline{S}_r,\left\{\overline{A}_{t-1},a_t,d^*_{r\setminus t}\right\})+\varepsilon_r^{dyn}-s^*\right\|_2^2 \mid H_t\right] = \mathbb{E}\left[\left\|\varepsilon_r^{dyn}\right\|_2^2 \mid H_t\right] = Var(\varepsilon_r^{dyn})$$

and

$$\mathbb{E}\left[\left\|f_r(\overline{S}_r,\left\{\overline{A}_{t-1},d^*_{r\setminus(t-1)}\right\})+\varepsilon_r^{dyn}-s^*\right\|_2^2 \mid H_t\right] = \mathbb{E}\left[\left\|\varepsilon_r^{dyn}\right\|_2^2 \mid H_t\right] = Var(\varepsilon_r^{dyn}).$$

Thus

$$\begin{aligned}\nu_t^{d^*}(a_t \mid H_t) &= \|f_t(H_t,a_t)-s^*\|_2^2 - \|f_t(H_t,d_t^*(H_t))-s^*\|_2^2 \\ &+ \sum_{r=t+1}^{T-1}\mathbb{E}\left[\left\|f_r(\overline{S}_r,\left\{\overline{A}_{t-1},a_t,d^*_{r\setminus t}\right\})+\varepsilon_r^{dyn}-s^*\right\|_2^2 \mid H_t\right] \\ &- \sum_{r=t+1}^{T-1}\mathbb{E}\left[\left\|f_r(\overline{S}_r,\left\{\overline{A}_{t-1},d^*_{r\setminus(t-1)}\right\})+\varepsilon_r^{dyn}-s^*\right\|_2^2 \mid H_t\right] \\ &= \|f_t(H_t,a_t)-s^*\|_2^2 - \|f_t(H_t,d_t^*(H_t))-s^*\|_2^2 \\ &= \|f_t(H_t,a_t)-s^*\|_2^2.\end{aligned}$$

From this, it is easy to see that $\nu_t^{d^*}(d_t^*(H_t) \mid H_t) = 0$ which implies $d^* = d^{opt}$ and so

$$\mu_t(a_t \mid H_t) = \|f_t(H_t,a_t)-s^*\|_2^2$$

as required.

Construction of d^{inf} via the use of H^∞ theory

In this section, we explain how to use H^∞ theory to implement our robust control strategy. In the first subsection, we reformulate the cost function defining d^{inf} as an H^∞ control problem. Then, in the second subsection we briefly summarise the theory for H^∞-synthesis in the context of a linear, discrete and time-invariant state-space model to construct d^{inf} in practice. See Francis (1987), Glover and Doyle (1988) or Doyle et al. (1989) for more detail.

Equations in the main paper are referenced here as (M1), (M2) and so on.

Reformulation of the treatment strategy design problem as an H^∞ control synthesis problem

First of all, to comply with classic control notation we need to reformulate the S_t evolution process, the cost function and the observation process as a linear time-invariant state-space model with no delay terms. This is

$$\left.\begin{array}{rcl} X_{t+1} &=& AX_t + B_1 w_t + B_2 a_t \\ V_t &=& C_1 X_t + D_{12} a_t \\ \mathbb{S}_t &=& C_2 X_t + D_{21} w_t, \end{array}\right\} \tag{1}$$

where:

(1) $X_t = (S_t, \ldots, S_{t-r}, A_{t-1}, \ldots, A_{t-q}, s^*)^T$ is the realised extended state vector at time t, containing the current and past state and past input values that determinine S_{t+1} together with s^*; and

(2) $V_t = \begin{pmatrix} S_t - s^* \\ \\ \lambda a_t \end{pmatrix}$ is the instantaneous cost we want to minimize at each timestep, assuming a_t close to zero is preferred.

To comply with the state-space model (M5) and cost $Y(d, \lambda)$ of equation (M7), the coefficient matrices in (1) need to be

$$A = \begin{pmatrix} F_S(\psi) & F_A(\psi) & 0 \\ (I_{r-1}\, 0_{r-1,1}) & 0_{r-1,q} & 0_{r-1,1} \\ 0_{1,r} & 0_{1,q} & 0 \\ 0_{q-1,r} & (I_{q-1}\, 0_{q-1,1}) & 0_{q-1,1} \\ 0_{1,r} & 0_{1,q} & 1 \end{pmatrix}, \quad B_1 = \begin{pmatrix} 1 & 0 \\ 0_{r+q,1} & 0_{r+q,1} \end{pmatrix}, \quad B_2 = \begin{pmatrix} F_a(\psi) \\ 0_{r-1,1} \\ 1 \\ 0_{q,1} \end{pmatrix}$$

for the extended state-space representation and:

$$C_1 = \begin{pmatrix} 1 & 0_{1,r+q-1} & -1 \\ 0 & 0_{1,r+q-1} & 0 \end{pmatrix}, \quad D_{12} = \begin{pmatrix} 0 \\ \lambda \end{pmatrix}, \quad C_2 = \begin{pmatrix} 1 & 0_{1,q+r} \end{pmatrix}, \quad D_{21} = (0 \quad 1)$$

for the cost function and observation process respectively. From this, d^{inf} is defined as the minimizer of

$$\sup_{\overline{w}^*_{T-1}} \left\{ \frac{\sum_{t=1}^{T} \left\| V_t(d, \overline{w}^*_{t-1}) \right\|_2^2}{\sum_{t=1}^{T} \|w_t^*\|_2^2} \right\}^{\frac{1}{2}}, \tag{2}$$

where $V_t(d, \overline{w}^*_{t-1})$ is the objective value obtained at time t when the strategy d and the sequence of disturbances $\overline{w}^*_{t-1}$ are applied to (1), as in (M8) of the main paper.

In order to present the theoretical results and numerical methods required for H^∞ synthesis, it is convenient to work in the frequency domain. As we are working in discrete time we make use of the z-transform version of (2). The z-transform of the sequence $\{X_t\}$ is

$$\tilde{X}(z) = X_0 + X_1 z + X_2 z^2 + \dots. \tag{3}$$

Formally, z is an indeterminate, but in practice it is almost always possible to take equation (3) to define $\tilde{X}$ as a function which is analytic in the whole complex plane except for poles. Define the z-transforms $\tilde{w}(z), \tilde{a}(z), \tilde{V}(z)$ and $\tilde{\mathbb{S}}(z)$ similarly. Multiplying the equations in (1) by z^t and summing gives, in the presence of zero initial conditions, the equivalent complex

frequency domain model as

$$\begin{aligned} z\tilde{X}(z) &= A\tilde{X}(z) + B_1\tilde{w}(z) + B_2\tilde{a}_t(z) \\ \tilde{V}(z) &= C_1\tilde{X}(z) + D_{12}\tilde{a}(z) \\ \tilde{\mathbb{S}}(z) &= C_2\tilde{X}(z) + D_{21}\tilde{w}(z). \end{aligned}$$

Let $P(z)$ be a matrix-valued transfer function linking the transformed output to the transformed input after substitution for $\tilde{X}(z)$, ie

$$\begin{pmatrix} \tilde{V}(z) \\ \tilde{\mathbb{S}}(z) \end{pmatrix} = P(z) \begin{pmatrix} \tilde{w}(z) \\ \tilde{a}(z) \end{pmatrix}.$$

From (1), $P(z)$ is given by

$$P(z) = \begin{pmatrix} 0 & D_{12} \\ D_{21} & 0 \end{pmatrix} + \begin{pmatrix} C_1 \\ C_2 \end{pmatrix} (zI - A)^{-1} \begin{pmatrix} B_1 & B_2 \end{pmatrix} = \begin{pmatrix} P_{11}(z) & P_{12}(z) \\ P_{21}(z) & P_{22}(z) \end{pmatrix}.$$

Now let the decision rule d be represented by a feedback transfer function $K = K(z)$ that links the transformed input $\tilde{a}(z)$ to the transformed observables $\tilde{\mathbb{S}}(z)$, i.e. $\tilde{a}(z) = K(z)\tilde{\mathbb{S}}(z)$. In turn let $G(K, z)$ be the transfer function that links the transformed objective $\tilde{V}(z)$ to the uncontrolled disturbances assuming the inputs are controlled through $K(z)$, namely $\tilde{V}(z) = G(K, z)\tilde{w}(z)$ where

$$G(K, z) = P_{11}(z) + P_{12}(z)K(I - P_{22}(z)K)^{-1}P_{21}(z).$$

With H^∞ the set of bounded analytic functions on the unit disc $\mathbb{D}$, the H^∞-norm of a function $f(z)$ is given by

$$\|f\|_{H^\infty} = \sup_{z \in \mathbb{D}} |f(z)| \quad \text{for } f \in H^\infty.$$

Since $\tilde{V}(z) = G(K, z)\tilde{w}(z)$, the sub-multiplicative property of subordinate norms leads to $\|\tilde{V}\|_2 / \|\tilde{w}\|_2 \leqslant \|G(K, .)\|_{H^\infty}$. Plancherel's theorem gives the same inequality in the temporal domain and since there are disturbances w which make the difference between the left and

right side arbitrarily small, it follows that for a given K, we have:

$$\|G(K,.)\|_{H^\infty} = \sup_{\overline{w}^*_{T-1}\neq 0} \left\{ \frac{\sum_{t=1}^{T} \left\|V_t(d,\overline{w}^*_{t-1})\right\|_2^2}{\sum_{t=1}^{T} \|w_t^*\|_2^2} \right\}^{\frac{1}{2}}.$$

This is the ratio we want to minimize. Hence we have turned our variational problem in the temporal domain into an optimization problem in the frequency domain. We are looking for the transfer function K, such that the norm $\|G(K,.)\|_{H^\infty}$, known as the $\mathcal{L}_2$-gain, is minimal, which is the problem H^∞ theory aims to address.

The decision rule, or equivalently transfer function K, which minimizes the $\mathcal{L}_2$-gain is generally not unique and usually there is no explicit expression. Instead, in practice the aim is to seek policies which do not necessarily achieve the minimal $\mathcal{L}_2$-gain but for which the $\mathcal{L}_2$-gain is bounded above by a small constant. For linear systems, as given and under assumptions, the global minimum exists and can be estimated numerically with an arbitrary level of precision. Nonetheless, in control theory it is often considered sufficient to find a policy K^γ which gives

$$\|G(K^\gamma,.)\|_{H^\infty} \leqslant \gamma \tag{4}$$

for a given $\gamma > 0$. Necessary and sufficient criteria for the existence of such a K^γ together with an outline of how to obtain it are the subjects of the next subsection.

Finding policy K

We need some technical conditions and definitions in order to determine whether at least one control law satisfying (4) exists, and if so to find K^γ. In this subsection we give an outline of the ideas and method. It is based on Francis (1987), Glover and Doyle (1988) and Doyle et al. (1989), which together provide full information.

First some terminology. A linear time-invariant system is *controllable* if it is possible to reach any location in the state-space in finite time by suitable choice of input. The weaker *stabilizable* assumption requires that any non-controllable states are *stable*, which means

that if the system reaches or begins in that state then it stays close thereafter. A system is *observable* if it is always possible to determine the current state x_t given the system outputs $\{y_i : i = 0, 1, \ldots, t\}$. A system is *detectable* if any non-observable state is stable.

Now we give some requirements for the matrices involved in (1):

(1) $(A,\ B_1)$ is stabilizable and $(C_1,\ A)$ is detectable.

(2) $(A,\ B_2)$ is stabilizable and $(C_2,\ A)$ is detectable.

(3) $D_{12}^T \begin{pmatrix} C_1 & D_{12} \end{pmatrix} = \begin{pmatrix} 0 & I_{j_1} \end{pmatrix}$.

(4) $\begin{pmatrix} B_1 \\ D_{21} \end{pmatrix} D_{21}^T = \begin{pmatrix} 0 \\ I_{j_2} \end{pmatrix}$.

In the third and fourth conditions, which can be achieved by an appropriate scaling, I_{j_1} and I_{j_2} are identity matrices of appropriate dimension.

Next we introduce the matrix space RIC. This is the space of $2p$ square Hamiltonian matrices H of the form

$$H = \begin{pmatrix} F & R \\ Q & -F^T \end{pmatrix},$$

with Q and R symmetric. Moreover, H should have no eigenvalues on the imaginary axis and, further, the vector space corresponding to the eigenvalues of H with a negative real part should be complementary to the image

$$Im \begin{pmatrix} 0 \\ I_p \end{pmatrix},$$

where I_p is the $p \times p$ identity matrix. The space RIC is introduced because it allows the derivation of conditions ensuring the existence and uniqueness of solutions to so-called algebraic Riccati matrix equations. These are expressed in the following paraphrase of Lemma 1 of Doyle et al. (1989). Suppose

$$H = \begin{pmatrix} F & R \\ Q & -F^T \end{pmatrix} \in RIC.$$

Then there exists a unique solution to the algebraic Riccati equation

$$F^T X + XF + XRX - Q = 0,$$

denoted $X_\infty = Ric(H)$.

Solutions to algebraic Riccati equations of the above form are in turn needed to ensure the existence of K^γ and the subsequent derivation of an expression for it. Following Doyle et al. (1989) it can be shown that there exists a K^γ such that

$$\|G\left(K^\gamma, .\right)\|_{H^\infty} < \gamma$$

if and only if the matrices

$$H_\infty := \begin{pmatrix} A & \gamma^{-2} B_1 B_1^T - B_2 B_2^T \\ -C_1^T C_1 & -A^T \end{pmatrix}, \quad J_\infty := \begin{pmatrix} A & \gamma^{-2} C_1 C_1^T - C_2 C_2^T \\ -B_1^T C_1 & -A^T \end{pmatrix}$$

belong to RIC space, and the solutions X_∞ and Y_∞ of the algebraic Riccati equations $Ric(H_\infty)$ and $Ric(J_\infty)$ are such that $\rho\left(X_\infty Y_\infty\right) < \gamma^2$, where ρ is the spectral radius, i.e. the largest absolute eigenvalue.

The first condition only requires the computation of the eigenvalues of H_∞ and J_∞ in order to be verified. It then ensures existence of the solutions introduced in the second condition, which can be obtained through reasonably straightforward matrix manipulations.

Once X_∞ and Y_∞ have been obtained, there is a simple expression for a valid control law:

$$K^\gamma = \begin{pmatrix} A_\infty & -Z_\infty L_\infty \\ F_\infty & 0 \end{pmatrix},$$

where

$$\begin{aligned} A_\infty &= A + \gamma^{-2} B_1 B_1^T X_\infty + B_2 F_\infty + Z_\infty L_\infty C_2, \\ F_\infty &= -B_2^T X_\infty, \\ L_\infty &= -Y_\infty C_2^T, \\ Z_\infty &= (I - \gamma^{-2} X_\infty Y_\infty)^{-1}. \end{aligned}$$

Glover and Doyle (1988) and Doyle et al. (1989) suggest the following iterative method for computing in practice the optimal policy K: (a) select a positive number γ; (b) test if there is

a K^γ such that $\|G(K^\gamma, .)\|_{H^\infty} < \gamma$ by calculating the eigenvalues of H_∞ and J_∞; (c) increase or decrease γ accordingly. The limiting value of γ is reached when either $\rho\left(X_\infty Y_\infty\right) = \gamma^2$ or there are no solutions X_∞ or Y_∞. Any scaling should of course be reversed before application of the selected K.

References

Doyle, J., Glover, K., Khargonekar, P., and Francis, B. (1989). State-space solutions to standard h2 and h-infinity control problems. *IEEE Transactions on Automatic Control* **34,** 831–847.

Francis, B. (1987). *Lecture Notes in Control and Information Sciences.* Springer-Verlag: New York.

Glover, K. and Doyle, J. (1988). State-space formulae for all stabilizing controllers that satisfy an h-infinity-norm bound and relations to risk sensitivity. *Systems and Control Letters* **11,** 167–172.