Bayesian Experimentation and Learning Treatment Sets

Martin W. Cripps*

June 30, 2016

Abstract

This paper addresses the problem of a Bayesian policy maker learning which group of subjects should be treated, when there is sequential sampling of the subjects. This is modelled as a continuum-armed bandit problem with imperfect control of the arm that is may be pulled. The arms are correlated by the structure of the potential treatment sets. We show that the for every discount factor the policy maker correctly learns the true treatment set. *Journal of Economic Literature* Classification Numbers: C11, C44, D83

Keywords: Experimentation, Bandit Problems, Treatment Choice, Treatment Response.

PRELIMINARY AND INCOMPLETE : DO NOT CITE

1. INTRODUCTION

In this paper I introduce a model of a policy maker who learns about how to allocate policy treatments to the members of a heterogeneous population. This is a continuum-armed bandit model that has correlated arms and imperfect control. The policy maker faces the usual trade-off between exploration and exploitation that is found in experimentation problems. However, unusually for bandit models, we will show that for any discount factor the policy maker correctly learns the set of subjects (arms) that benefit from the policy. It is the continuum of subjects and the imperfect control that gives this result on learning, and the result fails to hold when the set of arms is discrete (Banks and Sundaram (1992)) or if the control is perfect (Callander (2011)). One useful implication of the result is that even if the participation in the policy is voluntary and the subjects themselves determine their treatment, the policy maker will eventually acquire enough subjects willing to be treated by a policy to learn the true "treatment" set.

The choice of a policy intervention is often thought of as a static optimisation: There is a trial where a representative sample is taken from a heterogenous population. The subjects in the sample are allocated to treatments and the outcome of the treatment is observed. (This allocation usually has an element of randomisation, although there may be constraints on experimenters' abilities to do this.[1]) Once the trial is completed and the policy is chosen, the characteristics of the subjects in the general population who will receive the treatment is completely determined. This set of subjects is called here the treatment set. The learning about the optimal treatment set is essentially a one-off event and the policy maker does

*Department of Economics, University College London; m.cripps@ucl.ac.uk . My thanks are due to Amanda Friedenberg, Toru Kitagawa, Peter Sørensen and Aleksey Tetenov for their comments and suggestions.

[1]This is particularly true if the subjects are human and the experimenters are economists; see for example Glennnerster and Takavarasha (2013).

not consider what might be later learned from implementing the policy in the population as a whole. It is likely that the policy chosen in this way is suboptimal and it is later learned that subjects are being treated to the policy who should not be. It is also likely that there are subjects who are excluded from the treatment set who ought to be included, but it is less likely that it is learned who these subjects are. A consequence of this failure to learn is that much of the focus of the theoretical work on this issue and the related topic of pattern recognition (for example, Manski (2004), Kitagawa and Tetenov (2016), Devroye, Györfi, and Lugosi (1996)) has been on reducing the need for ex-post adjustment by imposing the criterion of regret minimisation in the initial policy choice.[2]

Nevertheless, it appears desirable to have a theory of how the initial treatment set should be adjusted in the light of the new information that comes from implementing the policy. And a theory of how this possibility of later adjustment should affect the initial sampling.[3] These are inherently dynamic issues. Here we address these issues by introducing a model of a Bayesian policy maker who sequentially samples subjects from the population and updates their beliefs about who should be treated to the policy in the light of all available information on the effects of the policy.

The important distinction between the static and dynamic model of policy choice is that the allocation of heterogeneous subjects to treatments is not random (as it was in the static model). The Bayesian policy maker uses her current beliefs about the optimal treatment set and the characteristics of the currently sampled subject to determine whether it is worth applying the policy intervention. Over time these beliefs evolve and so will the policy maker's willingness to experiment with the treatment on some subjects. In models with discrete arms, it is possible that this willingness to experiment gets permanently curtailed if a sequence of bad outcomes occurs for an arm. As a result, the Bayesian experimenter gets stuck at a sub-optimal policy with positive probability. We will show that in the model with a continuum of arms and imperfect control the policy maker will always conduct enough sampling in the limit to find the true treatment set.

The intuition for why learning occurs is ultimately quite simple. The subject who gets sampled each period is not determined by the policy maker (there is imperfect control of the arm that gets pulled). Hence, there is a chance that the sampled subject is not in the class of subjects that the policy maker is confident ought to be treated, but is nevertheless sufficiently close to this set for the policy maker to think it is likely that this subject would benefit from treatment. As a result, the policy maker continues to learn about the subjects who are on the margins of the subjects who are certain to be treated and the exploration phase of the learning does not end. If the set of marginal subjects does not shrink too rapidly as time passes, then the policy maker will continue to explore the subject space, by a non-trivial amount, for suitable candidates. Ultimately, finding all suitable candidates for treatment. (The intuition given here is incomplete, because there is no explanation of why the margins for exploration do not collapse to zero; this will follow below.)

[2] One exception is Chamberlain (2011) who describes a Bayesian approach to this static problem.

[3] A recent approach to this is Perchet, Rigollet, Chassang, and Snowberg (2016) where a finite number of rounds of sampling are considered.

It is clear from the above story that the imperfect control plays an important role in this result, because it obliges the policy maker to try marginal subjects and not just continue to use subjects known to benefit from treatment. Furthermore, the existence of these marginal subjects depends on the continuity of the subject-space and the correlation of the arms. Without a continuous set of subjects (arms), it is quite possible that a positive-probability sequence of bad news leads the set of marginal subjects to vanish entirely. Indeed we will give an example with discrete arms (but imperfect control) where the result on learning fails. Thus, it is the combination of the two properties, imperfect control and a continuous set of arms, that generates our result.

We now address the missing step in the above intuition. For the set of marginal subjects to become negligible, it is necessary that the policy maker's posterior probability of a successful treatment declines very rapidly as the sampled subject moves outside the treatment set. This rapid decline in the posteriors can only occur if the policy maker has acquired lots of negative information about the outcome of the policy in this region. In models with discrete arms, only a finite decline is necessary and a finite amount of negative information is sufficient for this. However, with a continuum of arms it is necessary that there is an unbounded amount of negative information, for the margin set to shrink to zero. This unbounded amount of negative information is a probability zero event in the case where treatment is optimal. Thus margins do not shrink rapidly when there is still suitable candidates for treatment. And, in contrast to the discrete-arms case, the learning at the margin does not end.

The literature on bandit models with a continuum of arms can be divided into those models where the decision taker gets to choose the the arm that gets pulled each period and the those where there is imperfect control of the arm being pulled. In the first category there are many papers in economics (notably McClennan (1984), Aghion, Bolton, Harris, and Jullien (1991) and Callander (2011)) as well as a multitude of papers in other fields.[4] These exhibit the phases of experimentation and exploitation common to bandit problems that are affected by the large number of potential arms. In such models the desire to explore new arms from the multitude available eventually weakens and the Bayesian learning stops before the state is fully known. (Except in Aghion, Bolton, Harris, and Jullien (1991) where arms are analytic functions of the state and arbitrarily small amounts of exploration can be very informative.)

The literature on bandit problems with a continuum of arms and incomplete control, was begun by Woodroofe (1979) and developed by, among others, Goldenshluger and Zeevi (2009), Perchet and Rigollet (2013). Their terminology is different. That which we have called a bandit with imperfect control is termed a "bandit with concomitant variables" or a "contextual bandit". In all cases, the idea is that in each period the policy maker observes a random variable, that is correlated with the outcome of the arms and can decide, after seeing the random variable, whether to pull the arm to learn about the relationship between the random variable and the outcome. Much of this literature focusses on the regret minimisation objective and the rate at which this regret vanishes as the size of the sample increases. This literature has also tended to focus on treatment effects that are continuous, so a small change in the observable results in a small change in the outcome in every state. This makes the arms that are being pulled by

[4]For example Agrawal (1995).

the experimenter very highly correlated. In such a setting, learning the treatment set is often possible by observing the outcomes for a very limited set of arms/covariates. For example, in Woodroofe (1979), if the policy maker knew the treatment outcome of one particular subject with certainty, then she would also know exactly which subjects lay in the treatment set. As the arms that are being pulled are highly correlated, the need to explore the space of subjects for possible candidates for treatment is quite weak. In this paper we look at the opposite case where the arms are only minimally correlated and the need for exploration is stronger. Here knowing that a particular subject is in the treatment set places a bound on the possible values of the treatment set, but is otherwise uninformative. To be precise, the treatment sets we consider will be arbitrary, finite, convex, polytopes and the response to treatment is discontinuous on the boundary of the polytope. All subjects in the treatment set have the same (random) outcome from treatment and all subjects outside the treatment set also have the same (inferior) outcome from treatment. Thus, the policy maker's objective is to learn a polytope of subjects, rather than a continuous functional relationship between subject and treatment outcome. The learning in the bandit problem studied here depends a lot on the structure of the set of all polytopes and hence also relates to the ideas in the literature on Vapnik-Chervonenkis dimension Devroye, Györfi, and Lugosi (1996), Al-Najjar (2009).

There is also a literature on the strategic issues that arise when policy trials are undertaken, for example Di Tillio, Ottaviani, and Sørensen (2015). This looks at the problems the policy maker faces when she tries to get accurate information from a trial conducted by an organisation or individual with distinct objectives. We have nothing to say about this, as the policy maker does not delegate the trial process. We do have something to say about the game played between the policy maker and the subjects who participate in the trial, however. It is clear that there is a conflict in objectives between the policy maker and the subjects even if the policy maker is entirely benign. For example, a policy maker might want to treat a subject that does not benefit from a treatment to gain information about other subjects' response to treatment. As a result subjects may not agree to be treated even if this policy maker is would like to impose it, Chan and Hamilton (2006) give a dynamic example of this. The results in this paper apply to policy makers who have a zero discount factor. Such a policy maker does not value the information generated from experimentation and thus has identical preferences to the current sampled subject. Thus we are able to show that even if the policy maker shared all information with the current subject and also delegated the treatment decision to the subject, then the subjects themselves would learn the true treatment set.

This paper will be organised in the following way. In Section 2 we describe the policy maker's problem and also sketch an example that will re-emerge at various stages. Then, in Section 3 we show how the policy maker's optimisation can be described by an HJB recursion. This generates some insight into the nature of the policy maker's decision taking. In Section 4 the main result is proven, this begins by presenting some results on learning and then divides the result into two parts, first showing that no subject outside the treatment set will be treated in the limit and then showing that all subjects inside the treatment set will be treated. In Section 5 we prove a general property of these models, which we will call limiting myopia.

1.1. An Example with Discrete Space of Subjects

Before embarking on the model we provide a simple example of how the result obtained here will fail if the subject space is discrete. The example will have the imperfect control of the arms that get pulled. There is a physician who must decide how to treat her patients. The patients have two types, $x = 1$ or $x = 2$, the simplest discrete space one could imagine. In every period one patient arrives at the physician to be treated; they are sampled from the set $\mathcal{X} = \{1, 2\}$ independently and uniformly. The physician believes that either all patients will benefit from treatment, or only patients with $x = 1$ benefit from the treatment, that is, the treatment set is either $\{1, 2\}$ or $\{1\}$. She attaches the prior $1 > \pi > 0$ to the hypothesis that all patients benefit from treatment. The discussion below shows that the physician will fail to learn that the treatment set is $\{1, 2\}$ and stop offering treatment to the $x = 2$ patients with positive probability.

The patients who are in the treatment set have a good outcome ($u = +1$) with probability $1 - \gamma$ and a bad outcome ($u = -1$) with probability $\gamma < 1/2$. The patients that are outside the treatment set always have a bad outcome ($u = -1$), if they are treated. If the physician decides not to treat any patient, then the outcome is a default $u = 0$. Both the patient and the physician can observe these outcomes and the physician's utility is simply the sum of the patients' utilities in each period.

This is a simple model of good-news learning, where observing a good outcome for a patient with $x = 2$ reveals that it is optimal for the physician to treat all patients. But if the treatment set is truly $\{1\}$, then only half the patients benefit from treatment so there is a potential cost to treating the $x = 2$'s. The problem the physician faces is to decide how many bad outcomes from $x = 2$ patients to tolerate before giving up and only treating the $x = 1$ patients. Notice that if the physician could decide which arm of the bandit problem to pull (which patient to sample each period), then she would always choose $x = 1$ patients. She would choose never to learn anything about the $x = 2$ patients, because she knows a maximal payoff can be received from treating the $x = 1$'s. When she has imperfect control she is forced to consider treating the $x = 2$'s. If she is sufficiently pessimistic about success with an $x = 2$, she will still prefer no treatment and waiting for the next $x = 1$ to arrive to the expected negative outcome of treating the $x = 2$ patient today.

It is a positive probability event that the physician chooses to stop treating patients with $x = 2$ even though the treatment set is genuinely $\{1, 2\}$. First, we calculate some utility benchmarks. If the physician sees a good outcome for an $x = 2$ patient, she will get the maximum expected utility of $1 - 2\gamma = (1 - \gamma)(+1) + \gamma(-1)$ per period. If she decides to only treat those she is certain are in the treatment set, that is $x = 1$ patients, then half the patients that arrive at her office with have the default outcome and her expected utility will be $\frac{1}{2} - \gamma$ per period. Now consider a physician with the prior $\pi < 1$ and facing a patient with $x = 2$. One possible policy for the physician to adopt is to treat this patient and, if the treatment is unsuccessful, to never treat an $x = 2$ patient again. But, if the outcome is successful then all future patients will be treated. Thus this policy has the expected payoff

$$\pi(1-\gamma)\left(1(1-\delta)+\delta(1-2\gamma)\right)+\left(1-\pi(1-\gamma)\right)\left(-1(1-\delta)+\delta((1/2)-\gamma)\right),$$

where $\delta < 1$ is the physician's discount factor. Simply refusing to treat the $x = 2$ patient and henceforth only treating patients with $x = 1$ gives the physician the discounted expected utility $\delta(\frac{1}{2} - \gamma)$. Comparing these two payoffs we can see that refusing to treat the $x = 2$ patient is optimal if

$$\pi < \frac{1-\delta}{(1-\gamma)(2-\delta(1-\gamma)-\delta/2)}.$$

This condition on the physician's beliefs ensure that no further experimentation with $x = 2$ patients is optimal. If the prior of the physician is above this level, it will be revised down each time an $x = 2$ patient is unsuccessfully treated. For each prior (below unity) there is a finite number of times this can happen before this threshold is breached. Thus, there is strictly positive probability of ceasing to treat the $x = 2$ patients even though it is optimal to do so.

2. MODEL

The model we describe below has a heterogeneous population of subjects that is sampled sequentially by a policy maker. In each period the policy maker observes a characteristic of the sampled subject and then chooses whether or not to treat them. The outcome of the treatment (if it has been applied) is observed and the period ends. However, only a subset of the population will get a good outcome from the treatment and this "treatment set" is unknown to the policy maker. The policy maker's long-term aim is treat those in the treatment set but not to treat those outside it. Thus, the decision to treat this period's subject has a direct utility benefit/cost as well as longer term informational benefit for the policy maker.

Before play begins a treatment set $\tilde{\theta} \subset \mathcal{X}$ will be determined. Time is discrete and denoted $t = 0, 1, \ldots$. In each period t one subject with characteristics $x_t \in \mathcal{X}$ is independently sampled from the set of possible characteristics, $\mathcal{X}$, where

$$\mathcal{X} := \{x \in \mathbb{R}^n : \|x\| \leq M\}, \qquad M > 1.$$

The subjects are sampled according to the measure υ, which is the uniform (Borel) measure on $\mathcal{X}$.[5]

On observing this period's sampled subject, x_t, the policy maker has two possible actions $D_t = 0$ or $D_t = 1$. The action $D_t = 0$ is a default non-treatment action and generates the utility $y_t = 0$. The action $D_t = 1$ is a decision to treat the subject, which generates a random outcome. The possible utilities after treatment are $y_t \in \{-1, +1\}$ and these occur with probabilities that depend on whether x_t is in the treatment set $\tilde{\theta}$ or not:

$$\Pr(y_t = 1 | x_t \in \tilde{\theta}) = \frac{1+\beta}{2}, \qquad \Pr(y_t = 1 | x_t \notin \tilde{\theta}) = \frac{1-\beta}{2},$$
$$\Pr(y_t = -1 | x_t \in \tilde{\theta}) = \frac{1-\beta}{2}, \qquad \Pr(y_t = -1 | x_t \notin \tilde{\theta}) = \frac{1+\beta}{2};$$

where $0 < \beta < 1$. Thus choosing to treat a subject ($D_t = 1$) when they are in the treatment set ($x_t \in \tilde{\theta}$) generates a positive expected utility, β, but treating a subject when they are not in the treatment set ($x_t \notin \tilde{\theta}$) generates an expected utility, $-\beta$, that is negative. Recall that untreated subjects always have

[5]The Euclidean norm, $\|\cdot\|$, is used throughout this paper.

utility zero, so it is optimal to treat subjects in $\tilde{\theta}$ but not treat those outside it. The parameter β determines the informativeness of the policy outcomes and as β approaches zero treating a single individual reveals more about the location of $\tilde{\theta}$.[6]

At the end of each period, the policy maker observes the decision that has been taken and the outcome of the policy. Hence, at the start of period t, a history for the policy maker consists of a sequence of subjects who have arrived for treatment, $(x_0, x_1, \ldots, x_{t-1}) \in \mathcal{X}^t$, a sequence of past treatment decisions, and a sequence of past utility outcomes $(y_0, y_1, \ldots, y_{-1}) \in \{-1, 0, 1\}^t$.[7] We write this history as $h_t \in \mathcal{X}^t \times \{-1, 0, 1\}^t := \mathcal{H}_t$. A treatment policy σ for the policy maker, then, consists of a sequence of functions $\{\sigma_t\}_{t=0}^{\infty}$ that map the history $h_t \in \mathcal{H}_t$ and the current sampled subject $x_t \in \mathcal{X}$ to a treatment decision:

$$\sigma_t : \mathcal{H}_t \times \mathcal{X} \to \Delta(\{0, 1\}), \qquad t = 0, 1, \ldots;$$

where $\mathcal{H}_0$ is taken as an arbitrary singleton set. This definition of a policy allows for randomisation in the treatment decision, but it will be made clear below that there is always an optimal deterministic policy. We treat the policy maker's objective as those of a benevolent social planner, so her objective is to maximise the sum of utilities of the agents that arrive each period. Given a sequence of utility outcomes $\{y_t\}_{t=0}^{\infty}$ for the population, she receives a utility that is a normalised discounted sum of these values: $(1 - \delta) \sum_{t=0}^{\infty} \delta^t y_t$, where $0 < \delta < 1$.

The treatment sets, denoted by $\tilde{\theta} \subset \mathcal{X}$, are restricted to lie in a class $\tilde{\Theta}$ of possible treatment sets and this class will be determined by a finite-dimensional parameterization. First we will use the parameters $(\lambda, \gamma) \in \mathcal{S}^{n-1} \times [0, 1]$ to determine a half-space $H_{\lambda\gamma} \subset \mathbb{R}^n$, where[8]

$$H_{\lambda\gamma} := \{x \in \mathbb{R}^n : \lambda^T x \leq \gamma\}.$$

So, λ is the normal to a hyperplane and $\gamma \geq 0$ is its distance from the origin and $H_{\lambda\gamma}$ is the halfspace containing the origin bounded by the hyperplane $\lambda^T x = \gamma$. The treatment set $\tilde{\theta}$ will be determined by the intersection of K such halfspaces. The parametric representation of the treatment set is a list of parameters that describe the normals and distances of K planes: $\theta := \{(\gamma_k, \lambda_k)\}_{k=1}^{K}$. The treatment set itself is the resulting intersection of the K halfspaces

$$\tilde{\theta} := \bigcap_{k=1}^{K} H_{\lambda_k \gamma_k} \subset \mathbb{R}^n, \qquad \text{for some} \qquad \theta = \{(\gamma_k, \lambda_k)\}_{k=1}^{K} \in ([0, 1] \times S^{n-1})^K.$$

One interpretation of this description of the set $\tilde{\theta}$ is that a subject who will benefit from being treated must simultaneously satisfy a number of different *eligibility* criteria. These eligibility criteria, $\lambda^T x \leq \gamma$, are linear (but unknown) functions of the subject's characteristics.[9] The class $\tilde{\Theta}$ is a collection of possible

[6] More general outcome distributions are possible provided the supports are finite and the distributions are distinct this simple parametric pair of distributions is not necessary for the result.

[7] Here it is not necessary to record directly the treatment decisions as these are encoded in the utility outcomes, although in more general models this is not the case.

[8] S^{n-1} is the $n-1$-sphere $\{x \in \mathbb{R}^n : \|x\| = 1\}$.

[9] This model of a random convex set is also called the intersection of linear eligibility scores. The use of support functions to describe random sets is well established practice, see for example Molchanov (2005), and can be used to embed the convex sets into a Banach space.

treatment sets and the set Θ is a collection of possible parameter values where:

$$\tilde{\Theta} := \left\{ \tilde{\theta} \subset \mathbb{R}^n : \tilde{\theta} = \bigcap_{k=1}^{K} H_{\lambda_k \gamma_k}, \{(\gamma_k, \lambda_k)\}_{k=1}^{K} \in \Theta \right\},$$
$$\Theta := ([0,1] \times S^{n-1})^K.$$

This construction defines a one-to-one map, f, from the finite-dimensional parameter space Θ to the convex polytopes in $\mathbb{R}^n$, that we will denote by f where

$$f(\theta) \equiv f\left(\{(\gamma_k, \lambda_k)\}_{k=1}^{K}\right) := \bigcap_{k=1}^{K} H_{\lambda_k \gamma_k}.$$

All beliefs that the policy maker has about the treatment set will be described here by beliefs on the parameter set Θ rather than directly on $\tilde{\Theta}$. The prior beliefs of the policy maker on the parameters $\theta \in \Theta$, will be denoted by a Borel probability measure $\mu \in \Delta(\Theta)$.[10] This measure defines a probability measure on $\tilde{\Theta}$ indirectly by defining a probability measure on its parametric representation $\tilde{\Theta}$. However, this indirect representation μ and the function f does generate a random set in $\mathcal{X}$ in the sense of Molchanov (2005).[11]

The uniform sampling density v, the treatment set $f(\theta)$ and the sequence of of policies $\sigma := \{\sigma_t\}_{t=0}^{\infty}$ determine a probability measure on the set of histories $\mathcal{H}_\infty$. When combined with the policy maker's prior $\mu \in \Delta(\Theta)$, this determines a probability measure $\mathbb{P}$ on the states of the world $\Theta \times \mathcal{H}_\infty$. We take an expectation with respect to this measure to calculate the policy maker's expected payoff from the policy:

$$E_{\mu\sigma}\left((1-\delta)\sum_{t=0}^{\infty} y_t\right), \qquad \mu \in \Delta(\Theta).$$

Her objective is to find a policy, σ, that maximises this value. Hence we define the value function, $U : \Delta(\Theta) \to [-1, 1]$, as

$$U(\mu) := \sup_{\sigma} E_{\mu\sigma}\left((1-\delta)\sum_{t=0}^{\infty} \delta^t y_t\right).$$

This completes a bald description of the model. We end with an example of this structure.

2.1. Example

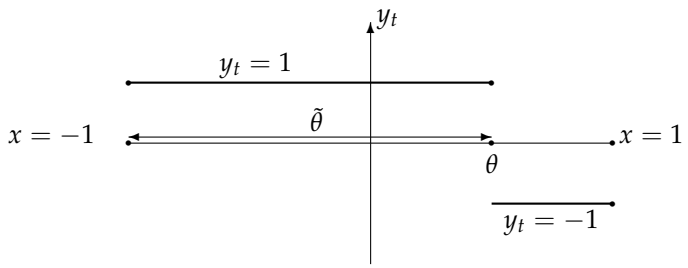
Our lead example of this model will be one-dimensional. This is very similar to an example studied by Aghion, Bolton, Harris, and Jullien (1991) in the context of perfect control (where the policy maker chooses which subject to sample. And an example of Ghosh and Ramamoorthi (2003, p. 22), where it is used to study the properties Bayesian learning without optimization. Here the population characteristic is uniformly distributed on the set $\mathcal{X} = [-1, +1]$. Thus $\{x_t\}_{t=0}^{\infty}$ is a sequence of *i.i.d.* uniformly

[10] We use $\Delta(Z)$ to denote the set of Borel probability measures on the set Z.

[11] To verify this it is necessary to check that for every closed set $F \subset \mathcal{X}$ the set $\{\theta : f(\theta) \cap F \neq \emptyset\}$ is Borel measurable (see Molchanov (2005)).

distributed random variables. The parameter θ that determines the treatment set is also uniformly distributed on the interval $[0,1]$, so μ is the uniform distribution on $[0,1]$. The treatment set itself will be a one-sided interval $\tilde{\theta} = [-1, \theta]$. Thus we have the function $f : \theta \mapsto [-1, \theta]$.

The payoffs will also be particularly simple in the example as we assume there is perfect information about the outcome of treatment. That is we will take $\beta = 1$, so treating a subject in the treatment set is always beneficial and treating a subject outside that set is always harmful. Further, outcomes of treatment are particularly informative and observing $(x_t, y_t) = (x_t, +1)$ implies that $\tilde{\theta} \geq x_t$ with probability one. Whilst observing $(x_t, y_t) = (x_t, -1)$ implies that $\tilde{\theta} < x_t$.



We will re-visit this example to illustrate the concepts developed below.

3. The Policy Maker's Optimisation Problem

This section deals with the necessary preliminaries. It begins by describing some simple properties of the policy maker's value function $U(\mu)$. Then, we write down the HJB equation that characterises $U(\mu)$, which will allow us to describe the optimal policy and provide some intuition for the nature of this optimization. At the end of this section the example is revisited and we provide a concrete demonstration of these features.

The result in this section also explain why it is not possible to have a more general model of the treatment sets. Here we have treatment set that is determined by θ, a parameter sampled from a compact and finite-dimensional set of parameters Θ. The priors on this set will be in $\Delta(\Theta)$ and are the state-space for the policy maker's optimization problem. A more general treatment set, one that is not determined by a finite parameter space, will result in an enlarged state-space. This new state space does not necessarily satisfy the conditions required (by the results used below) for the existence of an HJB equation. Thus the approach taken here is the most general model of a treatment set that is consistent with the existence of well-defined stationary solutions to the policy maker's optimization.

We begin by restating some well-know facts about the value functions for experimentation problems like this. First, there is a boundary condition. If there is no uncertainty about θ, then the optimal policy is to treat the subject x_t if and only if $x_t \in f(\theta)$. This occurs in each period with probability $v(f(\theta))$ and gives the expected utility β. Thus the value function must satisfy the boundary condition $U(\mathbb{1}_\theta) = \beta v(f(\theta))$,

where $\mathbb{1}_\theta$ is the Dirac measure on θ.

The second fact is that the function U is convex and non-negative on $\Delta(\Theta)$. Non-negativity is trivial as the default treatment gives a zero utility. The convexity follows from the fact that U is the upper envelope of linear functions of σ, see for example DeGroot (1970, p. 125). For example, if before implementing the optimal policy for μ, the policy maker observed a signal that told her the true state was either μ' or μ'', (where $\mu = \omega\mu' + (1-\omega)\mu''$; $\mu', \mu'' \in \Delta(\Theta)$; and $0 < \omega < 1$). Then, her maximal utility could not decrease if she observed the information, that is,

$$U(\mu) = U(\omega\mu' + (1-\omega)\mu'') \leq \omega U(\mu') + (1-\omega)U(\mu'').$$

In Lemma 1 we show that the policy maker's value function, $U(\mu)$, is a continuous solution to an HJB equation. And, hence, that there is a stationary and deterministic optimal treatment policy. The HJB equation also provides a simple interpretation of the nature of the optimal policy. The state-space for this problem is $\Delta(\Theta)$, so this is a non-standard HJB equation and we employ results of Easley and Kiefer (1988) and Maitra (1968) to establish these claims. The statement of this result contains some notation that will now be formally defined. First there is $\pi_\mu(x)$ which denotes the probability the subject x is believed to be in the treatment set given prior μ:

$$\pi_\mu(x) := \mu(I_x) \qquad \text{where} \qquad I_x := \{\theta : x \in f(\theta)\}. \tag{1}$$

(The condition $x \in f(\theta)$ can obviously be written more explicitly as a set of linear inequalities involving x and (λ_k, γ_k).) This is also known as the containment functional in the literature on random sets. The second piece of terminology is

$$\rho_\mu(x) := \frac{1}{2}(1+\beta)\pi_\mu(x) + \frac{1}{2}(1-\beta)(1-\pi_\mu(x)),$$

which denotes the probability that the outcome of treatment is $+1$ when subject x is treated. This, obviously, depends on the probability that x is in the treatment set and the probability of the outcome $+1$. Finally, $\mu^+, \mu^- \in \Delta(\Theta)$ will denote the posteriors when treatment outcomes $+1$ and -1 (respectively) are observed for subject x. (These are conditional on the current x and denote the next period's state.) Let G be any Borel measurable subset of Θ, then from Bayes' rule we have:

$$\mu^+(G) := \Pr(\theta \in G | x, y = 1) = \frac{(1+\beta)\mu(G \cap I_x) + (1-\beta)\mu(G \setminus I_x)}{2\rho_\mu(x)}; \tag{2}$$

$$\mu^-(G) := \Pr(\theta \in G | x, y = -1) = \frac{(1-\beta)\mu(G \cap I_x) + (1+\beta)\mu(G \setminus I_x)}{2(1-\rho_\mu(x))}. \tag{3}$$

Now we are able to state the HJB equation. The proof of this relation borrows heavily on Easley and Kiefer (1988) and Maitra (1968). It is, however, slightly different. The final HJB is an analogous condition to those found by Woodroofe (1979), Goldenshluger and Zeevi (2009) and other papers on bandit problems with exogenous variables.

LEMMA 1: *The function $U : \Delta(\Theta) \to [0,1]$ is the unique continuous solution to*

$$rU(\mu) = \int_{\mathcal{X}} \left[r\beta(2\pi_\mu(x) - 1) + \rho_\mu(x)U\left(\mu^+\right) + (1-\rho_\mu(x))U\left(\mu^-\right) - U(\mu) \right]^+ dv \tag{4}$$

where $r = (1-\delta)/\delta$ *and where* $[z]^+ := \max\{0, z\}$. *The stationary optimal policy in state* μ *is to treat all subjects* x *satisfying*

$$r\beta(2\pi_\mu(x) - 1) + \rho_\mu(x)U\left(\mu^+\right) + (1 - \rho_\mu(x))U\left(\mu^-\right) - U(\mu) \geq 0,$$

we will denote this policy as σ^*.

The proof of this lemma is relegated to the appendix, but now we will provide some intuition for the relationship (4). Observe that (4) is written in terms of the growth of the value function rather than its level. The expression still reflects the usual trade off in experimentation between the value of extra information and the short-run costs of acquiring that information weighted by the discount factor. Here, however, they are captured by the size of the increase in the value function rather than its level and so some explanation is required. The size of the increase in value is the expectation of a non-negative part of

$$H_\mu(x) := r\beta(2\pi_\mu(x) - 1) + \rho_\mu(x)U(\mu^+) + (1 - \rho_\mu(x))U(\mu^-) - U(\mu). \tag{5}$$

The expression $H_\mu(x)$ can be divided into two effects:

The first, $r\beta(2\pi_\mu(x) - 1)$, describes the increase in expected utility obtained from treating subject x at state μ. This can be negative if the subject is more likely to be outside the treatment set ($\pi_\mu(x) < \frac{1}{2}$) given the current beliefs. Or positive, if the subject is likely to be in the treatment set ($\pi_\mu(x) \geq \frac{1}{2}$). Thus, as the sampled x varies over $\mathcal{X}$ the expectation of this first term is the expected increase in utility from the policy of treating the current sampled subject.

The second part of the above, $\rho_\mu(x)U(\mu^+) + (1 - \rho_\mu(x))U(\mu^-) - U(\mu)$, is the expected increase in value as a result of the information gained by treating the subject x in state μ. Or, equivalently, the value of the information obtained from the experiment of treating x in state μ. Recall that $\rho_\mu(x)$ is the probability that treating subject x generates the outcome $y = 1$ and μ^+ is the posterior that results from $y = 1$. So $\rho_\mu(x)U(\mu^+) + (1 - \rho_\mu(x))U(\mu^-)$ is the expectation of next period's value when subject x is treated. The difference between this and $U(\mu)$ is the change in value as a result of the information gained from the treatment decision. This second part of the expected growth in value is non-negative. We have (from the martingale property of posteriors) that $\rho_\mu(x)\mu^+ + (1 - \rho_\mu(x))\mu^- = \mu$ and also recall that U is a convex function. Thus,

$$\rho_\mu(x)U\left(\mu^+\right) + (1 - \rho_\mu(x))U\left(\mu^-\right) \geq U\left(\rho_\mu(x)\mu^+ + (1 - \rho_\mu(x))\mu^-\right) = U(\mu).$$

The convexity of the value function is equivalent to there being a non-negative value to additional information.

The integrand $H_\mu(x)$, describing the increase in the value experienced when playing out the optimal policy, is the sum of two parts one non-negative (the value of additional information from experimentation) and the other ambiguous (representing the short-run costs and benefits of acquiring this information by treating the sampled subject).

$$H_\mu(x) := r\beta(2\pi_\mu(x) - 1) + \overbrace{\rho_\mu(x)U(\mu^+) + (1 - \rho_\mu(x))U(\mu^-) - U(\mu)}^{\geq 0}$$

Whenever the sum of these two terms is positive, or $H_\mu(x) \geq 0$. Then, treating subject x has a combined short-term cost/benefit and experimentation value that is greater than the current value function. It is, therefore, optimal to treat the subjects x that satisfy $H_\mu(x) \geq 0$ in state μ, because such a decision results in the value function growing. Whenever the sum of these two terms is negative, or $H_\mu(x) < 0$. Then, treating subject x has a combined short-term cost and experimentation value that is less than the current value function. It is, therefore, optimal *not* to treat these subjects and the value function does not grow in these states. These optimal choices (of treating x's if and only if $H_\mu(x)$ is positive) are described by the $[H_\mu(x)]^+$ function that appears inside the integral (4).

Because of the positive value of information, a sufficient condition for it to be optimal to treat subject x at state μ is that there are no expected costs to this action. Thus if $r\beta(2\pi_\mu(x) - 1) \geq 0$, or $\pi_\mu(x) \geq \frac{1}{2}$, it is optimal to treat x independently of this decision's informational effects. Of course, when $r\beta(2\pi_\mu(x) - 1) < 0$ it still may be optimal to treat x because of the weight the agent places on the future informational gains. Nevertheless, a sufficient condition for treating a subject is $\pi_\mu(x) \geq \frac{1}{2}$ and we will call this the myopic policy.[12]

DEFINITION 1: *The policy, σ^o, of treating subject x in state μ if and only if $\pi_\mu(x) \geq \frac{1}{2}$ will be called the* myopic policy.

From the discussion above it is trivial that the optimal policy treats more subjects than the myopic policy.

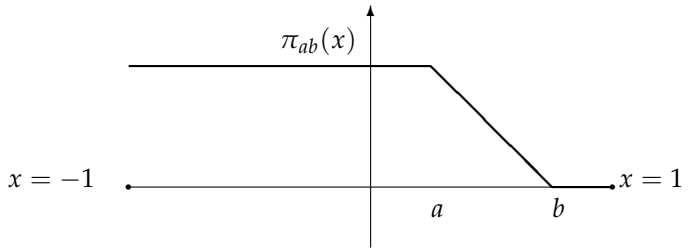
3.1. Example 2.1 Continued

Recall that the population is uniformly distributed on the set $\mathcal{X} = [-1, +1]$ and the treatment set is an interval $\tilde{\theta} = [-1, \theta]$ where θ is uniformly distributed on $[0, 1]$. As there is perfect monitoring, the state of the policy maker's beliefs about θ are described by an interval $[a, b]$. Where a is the largest non-negative x value that has been seen to generate $y = 1$ and b is the smallest x value that has been see to generate $y = -1$. (In the case where no $y = -1$ has been observed $b = 1$ and in the case where no $y = 1$ has been observed non-negative x then $a = 0$.) Given the data (a, b) the posterior on θ is uniformly distributed on the interval $[a, b]$. Hence we can write the probability that a given observation x is in the treatment set, $\pi_\mu(x)$ as depending only on the pair (a, b) in the following way.

$$\pi_{ab}(x) := \Pr(\theta > x | \theta \in (a, b)) = \begin{cases} 1 & x \leq a, \\ \frac{b-x}{b-a} & x \in (a, b), \\ 0 & x \geq b. \end{cases}$$

(This definition only applies when $b > a$, when $b = a$ we define π to be the a left-continuous step function.)

[12] Although in other fields this is called a greedy policy.



We can treat the pair (a,b) as a state variable for this problem and can write the HJB equation (4) in the following form

$$rU(a,b) = \frac{1}{2}\int_{-1}^{1} [r(2\pi_{ab}-1) + U(x,b)\pi_{ab} + U(a,x)(1-\pi_{ab}) - U(a,b)]^{+}\, dx,$$

(where we define $U(a,b) = U(b,b)$ if $a > b$). The first term is the expectation of the payoff when a subject x is sampled. The terms $U(x,b)$ and $U(a,x)$ are the new value functions when the outcomes $+1$ and -1 from treating subject x occur. Thus good news at x moves the interval containing θ upwards from $[a,b]$ to $[x,b]$ and bad news downwards to $[a,x]$. The subjects $x < a$ are known to be in the treatment set and will always be treated and subjects with $x > b$ will not be treated. Thus the above can be simplified to

$$\begin{aligned} rU(a,b) =& \frac{1}{2}\int_{a}^{b} [r(2\pi_{ab}-1) + U(x,b)\pi_{ab}(x) + U(a,x)(1-\pi_{ab}) - U(a,b)]^{+}\, dx \\ &+ \frac{1}{2}(a+1)r. \end{aligned} \tag{6}$$

It is possible to make some further simplifications to this function (including a reduction in the dimension of the problem), which are stated in the following Lemma

LEMMA 2: *The solution to (6) satisfies*

$$U(a,b) = \frac{1}{2}(1+b) + \frac{r^2}{b-a} C\left(\frac{b-a}{r}\right),$$

where C is the unique solution to

$$C(v) + \frac{v^2}{2} = \frac{1}{2}\int_{0}^{v} [v - 2u + C(v-u) + C(u) - C(v)]^{+}\, du.$$

The proof of this lemma is given in the Appendix. It is possible to use numerical methods to describe U, but an analytic solution to this is beyond this author.

4. LEARNING THE TREATMENT SET

In this section we describe the properties of the priors and the learning that we will repeatedly use in the proof of the main result. Then we will state and prove our main result, that for all discount factors the policy maker correctly learns the treatment set. This will be divided into two parts that addresses two possible ways the policy maker might err. The first, Theorem 1, shows that for all discount factors the

policy maker does not treat subjects outside the true treatment set indefinitely. The second, Theorem 2, shows that the policy maker does eventually treat all subjects that are inside the true treatment set (again for all discount factors). So, there is neither over-treatment nor under-treatment in the limit.

To be a little more precise, recall $T^0(\mu^t) := \{x : \pi_{\mu^t}(x) \geq \frac{1}{2}\}$ was defined as the set of subjects treated by the myopic policy when the prior is μ^t. As the myopic policy is played out, this results in random sequence of sets $\{T^0(\mu^t)\}_{t=0}^{\infty}$ of subjects who (potentially) will be treated at each date. For such a sequence

$$\mathcal{L} := \limsup_{t\to\infty} T^0(\mu^t)$$

is the (random) set of subjects who are expected to be treated infinitely often by the myopic policy.[13] The results in this section shows that $\mathcal{L}$ is arbitrarily close to the true treatment set, $\tilde{\theta}$, with probability one. If this is true for the myopic policy it must also be true for policy makers with positive discount factors, because, by the results in Section 5, all policies converge to the myopic policy.

4.1. Properties of Bayesian Learning

In this section some well-known properties of Bayesian learning are applied to the policy maker's beliefs that a particular subject is in the treatment set. These properties are essential for the results that follow. In (1), $\pi_\mu(x)$ was used to denote the prior belief that the subject x is in the treatment set. We want to consider how this prior gets updated, so let $\pi^t(x)$ denote this updated probability at time t, that is $\pi^t(x) := \Pr(x \in \tilde{\theta}|h_t)$.

Bayes' rule determines how, $\pi^t(x)$, the beliefs about subject x being in the treatment set evolve as observations accrue. Suppose that in period t the subject x_t was sampled and there was the outcome y_t, then

$$\begin{aligned}
\pi^{t+1}(x) =& \Pr(x \in \tilde{\theta} \mid x_t, y_t, h_t) \\
=& \frac{\pi^t(x)\Pr(y_t \mid x \in \tilde{\theta}, x_t)}{\Pr(y_t \mid x_t, h_t)} \\
=& \pi^t(x)\frac{\Pr(y_t \mid x_t \in \tilde{\theta})\Pr(x_t \in \tilde{\theta}|x \in \tilde{\theta}, h_t) + \Pr(y_t \mid x_t \notin \tilde{\theta})\Pr(x_t \notin \tilde{\theta}|x \in \tilde{\theta}, h_t)}{\Pr(y_t \mid x_t, h_t)}.
\end{aligned}$$

When no treatment occurred this reduces to $\pi^{t+1}(x) = \pi^t(x)$ and beliefs are not revised as no new information has arrived. However when subject x_t is treated, how beliefs about subject x are revised depends upon the nature of the class of treatment sets $\tilde{\Theta}$. To be precise, what x being in the treatment set implies about the probability that subject x_t is in the treatment set. For example, if x_t is an extreme observation it may be that x_t is in the treatment set *only if* x is in the treatment set. Thus getting positive news about x_t will also be positive news about x and so $\pi^t(x)$ will be revised a lot. Conversely, if x is extreme in one dimension and x_t more central in the same dimension, then it could be highly likely that x_t is in the treatment set independently of subject x being in the set. In which case, good news about

[13] If $\{F_t\}_{t=0}^{\infty}$ is a sequence of measurable sets, then $\limsup_{t\to\infty} F_t := \cap_{s=0}^{\infty} \cup_{t\geq s} F_t$ and includes all those points that appear in infinitely many F_t's.

subject x_t may not be very informative about the possibility of x being in the treatment set and there is little revision of $\pi^t(x)$ as a result. What matters very much in the belief revision is how the sets in the class of potential treatment sets $\tilde{\Theta}$ overlap or nest each other. Here we have adopted a particularly simple class, (convex polytopes), which makes the resultant belief revision easy to handle. For more complex classes of sets (of higher VC dimension see for example Devroye, Györfi, and Lugosi (1996), Al-Najjar (2009)) the above relation still applies, but its interpretation is less simple. It is, furthermore, not clear how the other results here, such as the HJB relation, generalise to more complex classes $\tilde{\Theta}$.

Now a lemma is stated that describes two key properties of Bayesian updating that are repeatedly used in our results, these are a well-known and are stated here just because these properties are fundamental in the proof below. (The proof of this lemma is given in the appendix for completeness.)

LEMMA 3: *If $\pi^t(x) := \Pr(x \in \tilde{\theta} \mid h^t)$, then for all $x \in \mathcal{X}$ $\pi^t(x)$ converges $\mathbb{P}$ almost surely. Furthermore,*

$$E\left(\frac{1-\pi^{t+1}(x)}{\pi^{t+1}(x)} \;\middle|\; x \in \tilde{\theta}, h_t\right) = \frac{1-\pi^t(x)}{\pi^t(x)}, \qquad E\left(\frac{\pi^{t+1}(x)}{1-\pi^{t+1}(x)} \;\middle|\; x \notin \tilde{\theta}, h_t\right) = \frac{\pi^t(x)}{1-\pi^t(x)}. \tag{7}$$

If x_t was sampled and treated at time t, then

$$\left|\pi^{t+1}(x) - \pi^t(x)\right| \geq 2\beta\pi^t(x)\left|\pi^t(x_t|x) - \pi^t(x_t)\right|. \tag{8}$$

Here the notation $\pi^t(x_t|x) := \Pr(x_t \in \tilde{\theta} \mid x \in \tilde{\theta}, h_t)$ is used.

The conditions (7) imply (from Jensen's Inequality) that posteriors are sub- or super-martingales: $E(\pi^{t+1}(x)|x \in \tilde{\theta}, h_t) \geq \pi^t(x)$ and $E(\pi^{t+1}(x)|x \notin \tilde{\theta}, h_t) \leq \pi^t(x)$. Thus on average the belief that x lies in the treatment set increase when it is true that $x \in \tilde{\theta}$ and decrease when it is true that $x \notin \tilde{\theta}$. So, the probability attached to the true state of the world is revised upwards under Bayesian learning. We, however, will use another property of the equalities (7). This says that the Bayesian revision of beliefs will not allow an individual to be certain that $x \in \tilde{\theta}$ when in fact $x \notin \tilde{\theta}$, except on a zero probability set of histories. To see why (7) implies this, suppose that the random variable $\pi^{t+1}(x) \to 1$ on a set of positive measure in the RHS equality. This would say that the individual is becoming convinced that the x is *in* the treatment set when in fact it is not. As $\pi^{t+1}(x) \to 1$ on a set of positive measure, the unconditional expectation of $\pi^{t+1}(x)/(1-\pi^{t+1}(x))$ would become unbounded. This would violate the fact that the unconditional expectation of this martingale sequence has to equal the (finite) $\pi^0(x)/(1-\pi^0(x))$. Thus it is impossible for $\pi^{t+1}(x) \to 1$ on a set of positive probability when $x \notin \tilde{\theta}$. We will show that events like this will happen with positive probability if the learning about the treatment set goes permanently wrong.

The second part of the lemma, (8), is often called the merging property of beliefs. As the random variable $\pi^t(x)$ converges almost surely, (8) says that

$$2\beta\pi^t(x)\left|\pi^t(x_t|x) - \pi^t(x_t)\right| \to 0.$$

Thus, if there continues to be x_t's that are treated, at least one of the two terms in this expression must converge to zero. That is, the policy maker ultimately believes subject x is not in the treatment set

($\pi^t(x) \to 0$). Or, the data, x_t, that the policy maker continues to observe has the property that knowing $x \in \tilde{\theta}$ is uninformative about the likelihood that $x_t \in \tilde{\theta}$; $\left|\pi^t(x_t|x) - \pi^t(x_t)\right| \to 0$. Possibly, both of these outcomes can occur. To see how this works in practice suppose that x is an extreme observation relative to x_t, and that knowing $x \in \tilde{\theta}$ implies $x_t \in \tilde{\theta}$. In this case $\pi^t(x_t|x) = 1$ so the above says that all the observations made satisfy $\pi^t(x_t) \to 1$, or $\pi^t(x) \to 0$. Hence, the policy maker only continues to sample observations that are believed to be in the treatment set with certainty or extreme observations are believed to be outside the treatment set. If, for example, the policy maker continued to experiment with x_t's which are not certain to be in the treatment set, then it would have to be that $\pi^t(x) \to 0$.

4.2. Subjects Outside the Treatment Set

We will now prove the first result: that subjects outside the true treatment set will not receive treatment in the limit as $t \to \infty$, whatever the discount factor. The proof of the first result will be by contradiction. Suppose that $\mathcal{L} := \limsup_{t\to\infty} T^0(\mu^t)$ contains subjects outside a neighbourhood of $\tilde{\theta}$ with positive probability. Pick such a subject, say x', then as this subject continues to be in the treatment set the posterior $\pi^t(x')$ converges to something above $\frac{1}{2}$. This is not enough—we want to find a a subject outside the treatment set with $\pi^t(x)$ converging to unity for a contradiction. As $\pi^t(x')$ does not go to zero, from (8), it then must be that $\left|\pi^t(x_t|x') - \pi^t(x_t)\right| \to 0$, for all x_t's that continue to be sampled. (That is, the subjects that continue to be treated are believed to be in the treatment set with a probability that is unaffected by knowing x' is in the treatment set.) Now consider a subject x'' that is on the line joining the origin to x'. This subject is *more* likely to be in the treatment set than x', thus the subject x'' will definitely continue to be sampled if x' is sampled. Furthermore, if $x' \in \tilde{\theta}$ the construction of the treatment set implies that $x'' \in \tilde{\theta}$; that is $\pi^t(x''|x') = 1$. Given this, the only way $\left|\pi^t(x_t|x') - \pi^t(x_t)\right| \to 0$ can hold is if $\pi^t(x'') \to 1$. Hence, if subject x' is in the treatment set then all points joining x' to the origin must have $\pi^t(x'') \to 1$. Thus we have achieved our aim of find a subject who is actually outside the treatment set (x'' sufficiently close to x') that is believed to be in the treatment set with certainty. As (7) shows a Bayesian cannot believe a falsity with certainty except on a set of measure zero. It is this that will give us a contradiction.

When the dimension of the subject set is bigger than unity ($n > 1$), the set of subjects on the line to the origin is a zero-measure set. Thus the argument above will fail, because subjects, like x'', on this ray may never be sampled. To avoid this problem (and some other zero-probability event issues) we make an assumption that there is a positive measure set of subjects that are always in the treatment set. This enlarges the line joining x' to the origin into an n-dimensional set.

ASSUMPTION 1: *There exists $\phi > 0$, such that $\mu(\{\tilde{\theta} : \Phi \subseteq \tilde{\theta}\}) = 1$, where $\Phi := \{x \in \mathcal{X} : \|x\| \leq \phi\}$*

Also, to avoid conditioning on zero probability events it is necessary to condition on a positive-measure neighbourhood of a given treatment set $\tilde{\theta}$. Let $\eta > 0$ sufficiently small be given and let $B_\eta(x)$ denote the closed ball centred at x radius η.[14] We will define $\tilde{\theta}_\eta$ as an enlargement or upper bound on

[14] $B_\eta(x) := \{\tilde{x} \in \mathcal{X} : \|\tilde{x} - x\| \leq \eta\}$.

$\tilde{\theta}$ as the closed set that results from enlarging $\tilde{\theta}$ by η: $\tilde{\theta}_\eta := \bigcup_{x\in\tilde{\theta}} B_\eta(x)$. And, we will define $\tilde{\theta}_{-\eta}$ as a shrinking or a lower bound on $\tilde{\theta}$: $\tilde{\theta}_{-\eta} := \{x \in \mathcal{X} : B_\eta(x) \subseteq \tilde{\theta}\}$. (Clearly, if $\tilde{\theta}$ is not of dimension n this set is empty and we will exclude this zero probability event in our results below.) Finally, we define a subset of the parameter space $B(\theta^*, \eta) \subset \Theta$ to be the neighbourhood of the parameter θ that ensures that for all θ' the set $f(\theta')$ is bounded below by $\tilde{\theta}_{-\eta}$ and above by $\tilde{\theta}_\eta$.

$$B(\theta, \eta) := \left\{ \theta' \in \Theta : \tilde{\theta}_{-\eta} \subseteq f(\theta') \subseteq \tilde{\theta}_\eta \right\}.$$

We are now able to state and prove the first result.

THEOREM 1: *Let $\eta > 0$ and $\theta^* \in \Theta$ be given. Assume that the prior μ is absolutely continuous with respect to Lebesgue measure and satisfies Assumption 1, then*

$$\mathbb{P}^0 \left(\mathcal{L} \subseteq \tilde{\theta}^*_\eta \;\middle|\; \theta \in B(\theta^*, \eta) \right) = 1, \tag{9}$$

where $\mathbb{P}^0$ is the probability measure on states induced by the myopic policy σ^0.

Proof. Let $\kappa > 0$ and small be given. We begin by defining the set of subjects that are strictly inside the treatment set at the state μ: $T^0_\kappa(\mu) := \{x : \pi_\mu(x) \geq \frac{1}{2} + \kappa\}$. And showing that

$$\mathbb{P}^0 \left(\mathcal{L}_\kappa \subseteq \tilde{\theta}^*_\eta \;\middle|\; \theta \in B(\theta^*, \eta) \right) = 1. \tag{10}$$

where $\mathcal{L}_\kappa := \limsup_{t\to\infty} T^0_\kappa(\mu^t)$. Suppose the claim (9) is false and there exists θ^*, η and $\varepsilon, \nu > 0$ such that

$$\mathbb{P}^0 \left(\mathcal{L}_\kappa \subseteq \tilde{\theta}^*_{\eta+\nu} \;\middle|\; \theta \in B(\theta^*, \eta) \right) \leq 1 - \varepsilon. \tag{11}$$

It will be useful to have a positive-measure set of subjects who are more likely to be in the treatment set than any given subject x. To that end for any $x \in \mathcal{X}$, let Φ_x denote the convex hull of the point x and the neighbourhood of the origin that is known to be in the treatment set Φ:

$$\Phi_x := \left\{ x' \in \mathbb{R}^n : x' = \omega x + (1-\omega)\tilde{x}, \text{for some } \tilde{x} \in \Phi, \omega \in [0,1] \right\}.$$

If $x \in \tilde{\theta}$, then $\Phi_x \subset \tilde{\theta}$ by the convexity of $\tilde{\theta}$. The probability that x is in the treatment set is, therefore, a lower bound on the probability that an element of Φ_x is in the treatment set: $\pi^t(x) \leq \pi^t(x^t)$ for all $x^t \in \Phi_x$. Thus if $x \in T^0_\kappa(\mu^t)$, then $\Phi_x \subseteq T^0_\kappa(\mu^t)$.

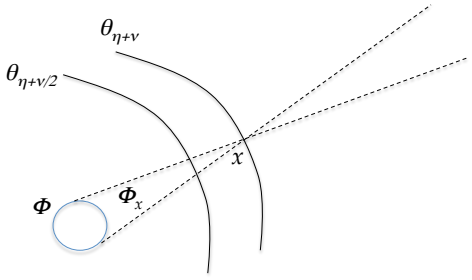


Figure 1 : The Cone Φ_x

Step 1: As a first step in the proof, we show that (11) implies there exists $\varepsilon' > 0$ and an $\bar{x} \in \mathcal{X} \setminus \tilde{\theta}^*_\eta$ such that

$$\mathbb{P}^0 \left(\bar{x} \in \mathcal{L}_\kappa \mid \theta \in B(\theta^*, \eta) \right) > \varepsilon'.$$

Observe that if $x \notin \tilde{\theta}^*_{\eta+\nu}$, then it is at least distance $\nu/2$ from $\tilde{\theta}^*_{\eta+\nu/2}$ and the truncated cone Φ_x must intersect the boundary of $\tilde{\theta}^*_{\eta+\nu/2}$ on a set of at least length $\phi\nu M^{-1}$ (recall ϕ is the radius of the ball Φ). This implies that we can pick a finite set of points, $\{x^1, x^2, \ldots, x^w\}$, on the boundary of $\tilde{\theta}_{\eta+\nu/2}$, such that every member of the class $\{\Phi_x : x \notin \tilde{\theta}_{\eta+\nu}\}$ has a non-empty intersection with this set. Thus, (by (11)) one of these points on the boundary of $\tilde{\theta}_{\eta+\nu/2}$ (call it $\bar{x}$) must have probability of at least ε/w of being in $T^0_\kappa(\mu^t)$. Because, whenever $\limsup_{t\to\infty} T^0_\kappa(\mu^t) \not\subset \tilde{\theta}_{\nu+\eta}$ it must contain a point $x \notin \tilde{\theta}_{\nu+\eta}$ and also contain Φ_x which has a non-zero intersection one of the points $\{x^1, x^2, \ldots, x^w\}$. Hence we have found $\bar{x} \notin \mathcal{X} \setminus \tilde{\theta}$ and ε' such that

$$\mathbb{P}^0 \left(\bar{x} \in \mathcal{L}_\kappa \mid \theta \in B(\theta^*, \eta) \right) > \varepsilon'.$$

Step 2: We will now show that if $\bar{x} \in \mathcal{L}_\kappa$, then $\pi^t(x) \to 1$ for every x in the interior of $\Phi_{\bar{x}}$. Hence, there exists a point $\hat{x}$ arbitrarily close to $\bar{x}$ (and outside $\tilde{\theta}_\eta$) such that $\pi^t(\hat{x}) \to 1$ with probability at least ε'. Take (8) and choose $x = \bar{x}$, then as $\pi^t(\bar{x})$ converges almost surely we have that

$$\mathbb{1}_{x_t \in T^0(\mu^t)} \pi^t(\bar{x}) \left| \pi^t(x_t|\bar{x}) - \pi^t(x_t) \right| \to 0, \qquad \mathbb{P}^0 \text{ almost surely.} \tag{12}$$

As $\bar{x} \in \mathcal{L}_\kappa$ with probability ε' when $\theta \in B(\theta^*, \eta)$, it is true that $\limsup_t \pi^t(\bar{x}) \geq \frac{1}{2} + \kappa$ with at least probability ε' when $\theta \in B(\theta^*, \eta)$. And, as $\pi^t(\bar{x})$ converges almost surely, we can also say with that $\lim_t \pi^t(\bar{x}) \geq \frac{1}{2} + \kappa$ with at least probability ε' when $\theta \in B(\theta^*, \eta)$.

From (12) it follows that

$$(\frac{1}{2} + \kappa) \mathbb{1}_{x_t \in T^0(\mu^t)} \left| \pi^t(x_t|\bar{x}) - \pi^t(x_t) \right| \to 0, \qquad \mathbb{P}^0 \text{ almost surely.}$$

on a set of states with at least probability ε' conditional on $\theta \in B(\theta^*, \eta)$. Applying Egoroff's Theorem (see Royden (1988) p.72) we can find a subset of these states with probability $0 < \varepsilon'' < \varepsilon'$ and a τ such that $\pi^t(\bar{x}) \geq \frac{1}{2}$ for all $t > \tau$. The construction of $\Phi_{\bar{x}}$ ensures that $\pi^t(x) \geq \pi^t(\bar{x}) \geq \frac{1}{2}$ for all $x \in \Phi_{\bar{x}}$. So on this positive-probability set of states for all $t > \tau$ *every* $x_t \in \Phi_{\bar{x}}$ will be treated if it is sampled. Hence, there exists a set of states with at least probability ε'' conditional on $\theta \in B(\theta^*, \eta)$ and a τ, such that

$$\mathbb{1}_{\{x_t \in T^0(\mu^t)\}} \left| \pi^t(x_t|\bar{x}) - \pi^t(x_t) \right| \to 0, \qquad \mathbb{P}^0 \text{ a. s.;}$$

and $\Phi_{\bar{x}} \subseteq T^0(\mu^t)$ for all $t > \tau$.

As the set $\Phi_{\bar{x}}$ has strictly positive measure, it is without loss of generality to assume that on this set of states there are infinitely many x_t's sampled from the set $\Phi_{\bar{x}}$. However, $\pi^t(x_t|\bar{x}) = 1$ for all $x_t \in \Phi_{\bar{x}}$. Hence the above implies there exists a set of states with at least probability ε'' conditional on $\theta \in B(\theta^*, \eta)$ and a τ, such that

$$\mathbb{1}_{\{x_t \in \Phi_{\bar{x}} \subseteq T^0(\mu^t)\}} \left| 1 - \pi^t(x_t) \right| \to 0, \qquad \mathbb{P}^0 \text{ a. s.} \qquad \forall t > T.$$

This implies that $\pi^t(x) \to 1$ for all x in the interior of $\Phi_{\bar{x}}$ on this positive probability set of states, because for any x^o in the interior of $\Phi_{\bar{x}}$ there will exist a sequence of x_t's sampled from $\Phi_{\bar{x}}$ satisfying $x^o \in \Phi_{x^t}$ and

so $\pi^t(x^o) \geq \pi^t(x^t) \to 1$. To complete this step, choose a point $\hat{x}$ in the interior of $\Phi_{\bar{x}}$ but not in $\tilde{\theta}^*_\eta$. Such a point exists as $\tilde{\theta}^*_\eta$ is closed. This point will satisfy $\pi^t(\hat{x}) \to 1$ with probability at least ε'' conditional on $\theta \in B(\theta^*, \eta)$.

Step 3: The final step in the proof is to recall (7) for $x = \hat{x}$

$$E\left(\frac{\pi^{t+1}(\hat{x})}{1-\pi^{t+1}(\hat{x})} \;\middle|\; \hat{x} \notin f(\theta), h_t\right) = \frac{\pi^t(\hat{x})}{1-\pi^t(\hat{x})}.$$

Taking an unconditional expectation this gives

$$E\left(\frac{\pi^{t+1}(\hat{x})}{1-\pi^{t+1}(\hat{x})} \;\middle|\; \hat{x} \notin f(\theta)\right) = \frac{\pi^0(\hat{x})}{1-\pi^0(\hat{x})}.$$

Note that $\hat{x} \notin f(\theta)$ for all $\theta \in B(\theta^*, \eta)$ and so we have show that on this positive measure set there is strictly positive probability that the LHS expectation is of a random variable that becomes arbitrarily large; $\pi^{t+1}(\hat{x}) \to 1$; on a positive probability set. This is a contradiction, as the RHS is finite.

Thus our initial assertion (11) is false and $\mathcal{L}_\kappa \subseteq \tilde{\theta}^*_\eta$ with probability one for all $\theta \in B(\theta^*, \eta)$ for all $\kappa > 0$. As $T^0_\kappa(\mu^t)$ and $\tilde{\theta}^*_\eta$ are closed sets and $\pi^t(x)$ is continuous in x this result also holds for $\kappa = 0$. □

4.3. Subjects Inside the Treatment Set

Now the second part of the result is proved, that is, all subjects in the treatment set eventually get treated by the policy maker. This is considerably more difficult to do than the previous case, but the overall strategy of proof is the same. We will ultimately derive a contradiction by finding a subject that is actually inside the treatment set but is believed to be outside the set with probability one. The intuition will be reminiscent of reputation arguments.

Suppose that there is a subject, say x', in the treatment set who is not in $\mathcal{L}$ with positive probability. Then we would like to argue that $\pi^t(x') \to 0$, but all we know is that it converges to something less than $\frac{1}{2}$. Again we will consider the subjects, x'', that are sampled on the line joining x' to the origin. The policy maker will be certain that some of these subjects are in the treatment set, so gathering further data on these subjects is not particularly informative about x'. However, as $\pi^t(x') < 1/2$ and π^t is continuous, some of the subjects sampled on the line segment will be on the margins of the set of subjects to be sampled and satisfy $1 - \xi > \pi^t(x'') > 1/2$. These subjects are not known to be in the treatment set and so their outcomes do generate a lot of information about x'. In particular, for these subjects $\pi^t(x''|x') - \pi^t(x'') > \xi$. Hence, it is impossible for x' to be in the treatment set and to be uninformative about x''. Thus observing these marginal x'''s continues to be informative about x' being in the treatment set and ultimately (by (8)) drives $\pi^t(x')$ to zero. Thus the policy maker must continue to observe significant negative information about x', there is not point at which this stops. The continuity here ensures that there continues to be informative signals about x' and as is x' not be treated in the limit it has to converge to zero.

THEOREM 2: *Let $\eta > 0$ and $\theta^* \in \Theta$ be given. Assume that the prior μ is absolutely continuous with respect to Lebesgue measure and satisfies Assumption 1, then*

$$\mathbb{P}^0\left(\mathcal{L} \supseteq \tilde{\theta}^*_{-\eta} \;\middle|\; \theta \in B(\theta^*, \eta)\right) = 1, \tag{13}$$

where $\mathbb{P}^0$ is the probability measure on states of the world induced the prior μ and the myopic policy.

Proof. Suppose the claim (13) is false and there exists θ^*, η and $\varepsilon, \nu > 0$ such that

$$\mathbb{P}^0\left(\mathcal{L} \supseteq \tilde{\theta}^*_{-(\eta+\nu)} \;\middle|\; \theta \in B(\theta^*, \eta)\right) \leq 1 - \varepsilon. \tag{14}$$

It will be useful in this proof to have a positive measure set of subjects that are *less* likely to be in the treatment set than any given subject x. To that end consider the cone of points

$$C_x := \{x' \in \mathbb{R}^n : x' = (1+\omega)x - \omega\check{x}, \text{for some } \check{x} \in \Phi, \omega \in \mathbb{R}_+\}.$$

This is a cone with the apex x that extends away from the set Φ. It is the continuation of the cone Φ_x considered in the proof of Theorem 1.

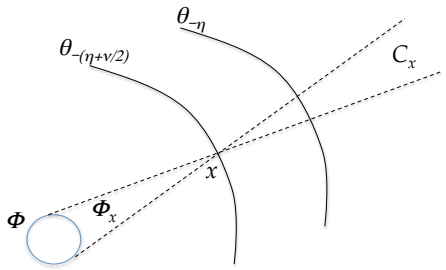


Figure 2 : The Cones C_x and Φ_x

The set has the property that if $x \notin \tilde{\theta}$, then no element of C_x is in $\tilde{\theta}$. To see this observe that if $x \notin \tilde{\theta}$, then there exists a hyperplane defining $\tilde{\theta}$ that x is on the wrong side of. That is, there exists (λ_k, γ_k) such that $\lambda_k^T x > \gamma_k$. Now pick $x' \in C_x$. This means there exists $\tilde{x} \in \Phi$ and $\omega > 0$ so that $x' = (1+\omega)x - \omega\tilde{x}$, (by the definition of C_x). If we perform the calculation to test whether the point x' in the treatment set, we find it is not, that is, $\lambda_k^T x' > \gamma_k$.

$$\lambda_k^T x' = (1+\omega)\lambda_k^T x - \omega\lambda_k^T \tilde{x} > \gamma_k + \omega(\gamma_k - \lambda_k^T \tilde{x}) \geq \gamma_k$$

(The equality in this calculation follows from a substitution for $\tilde{x}$. The first inequality holds as $\lambda_k^T x > \gamma_k$ and $\omega > 0$. The second inequality holds as $\tilde{x}$ is in the treatment set (because it is in Φ) and so satisfies $\lambda_k^T \tilde{x} \leq \gamma_k$.)

The fact that $x \notin \tilde{\theta}$ implies $C_x \cap \tilde{\theta} = \emptyset$. Hence, if the prior μ satisfies Assumption 1, the probability that x is believed to be in the treatment set is greater than the probability that any point in C_x is in the treatment set

$$\pi_\mu(x) \geq \pi_\mu(x'), \qquad \forall x' \in C_x.$$

Step 1: We now show that the assertion (14) implies there exists a neighbourhood of a subject in $\tilde{\theta}^*_{-\eta}$, that is not in $\mathcal{L}$ with positive probability. That is, there exists $\varepsilon' > 0$, $\zeta > 0$, an $\bar{x} \in \tilde{\theta}^*_{-\eta}$ and a ball $B_\zeta(\bar{x}) \subset \mathcal{X}$ such that[15]

$$\mathbb{P}^0 \left(\; B_\zeta(\bar{x}) \cap \mathcal{L} = \emptyset \;\middle|\; \theta \in B(\theta^*, \eta) \; \right) > \varepsilon'. \tag{15}$$

Consider two points $\hat{x}$ and $\tilde{x}$ both on the boundary of the set $\tilde{\theta}^*_{-(\eta+\nu/2)}$, satisfying $\|\tilde{x} - \hat{x}\| < \kappa$, where $\kappa < \phi\nu/(2M)$. There exists $\omega > 0$ such that the point $(1+\omega)\hat{x}$ is on the boundary of the larger set $\tilde{\theta}^*_{-\eta}$. Trivially this point is in the cone of $C_{\hat{x}}$, that is, $(1+\omega)\hat{x} \in C_{\hat{x}}$. We will show that $(1+\omega)\hat{x}$ is also in the cone $C_{\tilde{x}}$ with apex $\tilde{x}$. For $(1+\omega)\hat{x} \in C_{\tilde{x}}$ there must exist ω' and $\check{x} \in \Phi$ so that

$$(1+\omega)\hat{x} = (1+\omega')\tilde{x} - \omega'\check{x}.$$

Choose $\omega' = \omega$. Making this choice, then adding and subtracting $\hat{x}$ into the above gives

$$(1+\omega)\hat{x} = (1+\omega)\hat{x} + (1+\omega)\left[(\tilde{x} - \hat{x}) - \frac{\omega}{1+\omega}\check{x}\right].$$

As any $\check{x}$ satisfying $\|\check{x}\| < \phi$ can be chosen, the equality is true (and the final parentheses equal zero) for some $\check{x}$ provided

$$\|\tilde{x} - \hat{x}\| \leq \frac{\omega}{1+\omega}\|\check{x}\| \leq \frac{\omega}{1+\omega}\phi \leq \omega\phi.$$

As $\hat{x}$ is on the boundary of $\tilde{\theta}^*_{-(\eta+\nu/2)}$ and $(1+\omega)\hat{x}$ is on the boundary of $\tilde{\theta}^*_{-(\eta+\nu)}$ the triangle inequality implies $\|\hat{x} - (1+\omega)\hat{x}\| \geq \nu/2$, or $\omega > \nu/(2M)$. Hence we have that $\omega\phi > \omega\nu/(2M) > \kappa$ which is what we needed to establish.

Now cover the boundary of $\tilde{\theta}^*_{-(\eta+\nu/2)}$ with a finite number of $\kappa/2$ closed neighbourhoods numbered $w = 1, 2, \ldots, W$. For each of these neighbourhoods, denoted K_w, there exists one point x_w on the boundary of $\tilde{\theta}^*_{-\eta}$ that is strictly in the cone C_x for all $x \in K_w$ (by the proof above). Hence, if any point $x \in K_w$ satisfies $\pi_\mu(x) < 1/2$ then $\pi_\mu(x_w) < 1/2$. By (14) we know with probability at least ε at least one point in K_w (for some w) is not in $\mathcal{L}$. Thus with probability at least $\varepsilon/W := \varepsilon'$ one of the points $x_w := \bar{x}$ is not in $\mathcal{L}$. As $x_w \in \tilde{\theta}^*_{-(\eta+\nu/2)} \subset \tilde{\theta}^*_{-\eta}$ this establishes

$$\mathbb{P}^0 \left(\; \bar{x} \notin \mathcal{L} \;\middle|\; \theta \in B(\theta^*, \eta) \; \right) > \varepsilon'.$$

Now choose $\zeta > 0$ sufficiently small so that the ball $B_\zeta(\bar{x})$ is in the cone C_x for all $x \in K_w$, which is possible by the strict inclusion and the closed neighbourhoods used. This choice and the above property then satisfies (15).

Step 2: We now show that not only can we draw a ball around the subject $\bar{x}$ that lies outside $\mathcal{L}$ with positive probability, but there is a hyperplane through $\bar{x}$ that also lies outside $\mathcal{L}$ with positive probability. That is, there exists a $\varepsilon'' > 0$, $\bar{\lambda} \in \mathcal{S}^{n-1}$ and $\bar{\gamma}$ such that: $\bar{\lambda}^T\bar{x} = \bar{\gamma}$ and

$$\mathbb{P}^0 \left(\; \bar{\lambda}^T x \leq \bar{\gamma}, \; \forall x \in \mathcal{L} \;\middle|\; \theta \in B(\theta^*, \eta) \; \right) > \varepsilon''.$$

[15] Recall $B_\zeta(x) := \{\tilde{x} \in \mathcal{X} : \|\tilde{x} - x\| \leq \zeta\}$.

First, we will now show that the interior of the set $\mathcal{L}$ is convex. Suppose, x', x'' are in the interior of $\mathcal{L}$. Consider $x^\omega = \omega x' + (1-\omega)x''$ for $\omega \in (0,1)$. By linearity if $\lambda^T x' \leq \gamma$ and $\lambda^T x'' \leq \gamma$ then $\lambda^T x^\omega \leq \gamma$. Hence,

$$\Pr(x^\omega \in \tilde{\theta}) \geq 1 - \Pr(x' \notin \tilde{\theta}) - \Pr(x'' \notin \tilde{\theta}). \tag{16}$$

Or in our usual terminology $\pi^t(x^\omega) \geq \pi^t(x') + \pi^t(x'') - 1$. Step 2 in the proof of Theorem 1 shows that if $\bar{x} \in \mathcal{L}$, then $\pi^t(x) \to 1$ on the interior of $\Phi_{\bar{x}}$, thus $\pi^t(x'), \pi^t(x'') \to 1$ and from this bound $\pi^t(x^\omega) \to 1$ (and $x^\omega \in \mathcal{L}$), which is what we set out to prove.

Choose a set $\Lambda := \{\lambda_1, \ldots, \lambda_\ell\} \subset \mathcal{S}^{n-1}$, so that for any $\tilde{\lambda} \in \mathcal{S}^{n-1}$ there exists $\lambda_k \in \Lambda$ such that $\|\tilde{\lambda} - \lambda_k\| < \zeta/4M$. Suppose that $\tilde{\lambda} \in \mathcal{S}^{n-1}$ defines a hyperplane separating a convex set $\tilde{X} \subset \mathcal{X}$ from the ball $B_\zeta(\bar{x}) \subset \mathcal{X}$, that is there exists γ so that, $\tilde{\lambda}^T \tilde{x} \leq \gamma \leq \tilde{\lambda}^T y$ for all $\tilde{x} \in \tilde{X}$ and all $y \in B_\zeta(\bar{x})$. We will show that there exists $\lambda_k \in \Lambda$ such that $\lambda_k^T \tilde{x} \leq \lambda_k^T \bar{x}$ for all $\tilde{x} \in \tilde{X}$. To see this first observe that as $\bar{x}$ is the centre of the ball radius ζ and $\tilde{\lambda}$ is a unit vector, it follows $\tilde{\lambda}^T \bar{x} - \zeta \geq \tilde{\lambda}^T \tilde{x}$ for all $\tilde{x} \in \tilde{X}$. But this is equivalent to

$$\lambda_k^T \bar{x} \geq \lambda_k^T \tilde{x} + \zeta - (\tilde{\lambda}^T - \lambda_k^T)(\bar{x} - \tilde{x}) \qquad \forall \tilde{x} \in \tilde{X}. \tag{17}$$

Choose λ_k so that $\|\tilde{\lambda} - \lambda_k\| < \zeta/4M$. Then,

$$(\tilde{\lambda}^T - \lambda_k^T)(\bar{x} - \tilde{x}) \leq \frac{\zeta}{4M}\|\bar{x} - \tilde{x}\| \leq \frac{\zeta}{2}.$$

Hence for this choice of λ_k the inequality (17) implies $\lambda_k^T \bar{x} \geq \lambda_k^T \tilde{x}$ for all $\tilde{x} \in \tilde{X}$, which is what we wanted to show.

As the interior of $\mathcal{L}$ is convex, when it is disjoint from the convex set $B_\zeta(\bar{x})$ there exists a separating hyperplane between these two sets. That is, there exists $\lambda \in \mathcal{S}^{n-1}$ and γ so that

$$\lambda^T y \geq \gamma \geq \lambda^T x, \qquad \forall y \in B_\zeta(\bar{x}),\ \forall x \in \mathcal{L}.$$

($\mathcal{L}$ is connected, so this may hold with equality on the boundary of $\mathcal{L}$.) But this implies $\lambda_k^T \bar{x} \geq \lambda_k^T x$ (for all $x \in \mathcal{L}$) for some $\lambda_k \in \Lambda$, by the argument in the previous paragraph. As the set Λ is finite there must exist $\lambda_k \in \Lambda$ such that

$$\mathbb{P}^0\left(\lambda_k^T x \leq \lambda_k^T \bar{x}\ \forall x \in \mathcal{L} \;\middle|\; \theta \in B(\theta^*, \eta) \right) > \frac{\varepsilon'}{\ell} := \varepsilon''.$$

By choosing $\bar{\lambda} = \lambda_k$ and $\bar{\gamma} = \lambda_k^T \bar{x}$ we have proved our claim.

Step 3: We will now show that the set of subjects with $\pi^t(x) \in [1/2, 1-\xi]$ does not shrink to zero, when only subjects satisfying $\pi^t(x_t) \geq 1 - \xi$ are treated. To do this we will derive an explicit expression for $\pi^t(x)$ and then show that the ratio of the posteriors $\pi^t(x)/\pi^t(x')$ is bounded by a term that is independent of t, when $\pi^t(x), \pi^t(x') \in [1/2, 1-\xi]$.

We begin by deriving an explicit expression (18) for $\pi^t(x)$. At time t the history h_t is a list of pairs (x_s, y_s); $s = 0, 1, \ldots, t-1$. Let $\tilde{a}_t$ be the empirical measure that describes the subjects x_s ($s < t$) for whom $y_s = 1$, and $\tilde{b}_t$ is the empirical measure for subjects with $y_s = -1$. That is, for any measurable $A \subset \mathcal{X}$:

$$\tilde{a}_t(A) := \sharp\{\ (x_s, y_s) : s < t, x_s \in A, y_s = 1\ \},$$

$$\tilde{b}_t(A) := \sharp\{\ (x_s, y_s) : s < t, x_s \in A, y_s = -1\ \}.$$

Then the probability that the data at time t was generated by the treatment set $\tilde{\theta}$ is

$$\begin{aligned}\left(\frac{1+\beta}{2}\right)^{\tilde{a}_t(\tilde{\theta})} \left(\frac{1-\beta}{2}\right)^{\tilde{b}_t(\tilde{\theta})} \left(\frac{1+\beta}{2}\right)^{\tilde{b}_t(\mathcal{X})-\tilde{b}_t(\tilde{\theta})} \left(\frac{1-\beta}{2}\right)^{\tilde{a}_t(\mathcal{X})-\tilde{a}_t(\tilde{\theta})} \\ = \frac{(1+\beta)^{\tilde{b}_t(\mathcal{X})}(1-\beta)^{\tilde{a}_t(\mathcal{X})}}{2^{\tilde{a}_t(\mathcal{X})+\tilde{b}_t(\mathcal{X})}} \left(\frac{1+\beta}{1-\beta}\right)^{\tilde{a}_t(\tilde{\theta})-\tilde{b}_t(\tilde{\theta})} \\ := K_t \rho^{\tilde{c}_t(\tilde{\theta})},\end{aligned}$$

where $\rho := (1+\beta)/(1-\beta) > 1$, $\tilde{c}_t(\tilde{\theta}) = \tilde{a}_t(\tilde{\theta}) - \tilde{b}_t(\tilde{\theta})$ and K_t is a term that is independent of $\tilde{\theta}$. Recall that μ is the prior on the parameter space Θ and that when the parameters are θ the resultant treatment set is $\tilde{\theta} = f(\theta)$, hence we can write the unconditional probability of the outcomes h_t as

$$\Pr(h_t) = K_t \int_\Theta \rho^{\tilde{c}_t(f(\theta))} d\mu \equiv K_t \int_\Theta \rho^{c_t(\theta)} d\mu,$$

where $c_t(\theta) := \tilde{c}_t(f(\theta))$. As $\pi^t(x)$ is the probability $x \in f(\theta)$ conditional on the data we can, therefore write

$$\pi^t(x) = \frac{\int_{\Theta_x} \rho^{c_t(\theta)} d\mu}{\int_\Theta \rho^{c_t(\theta)} d\mu} \tag{18}$$

where $\Theta_x := \{\theta : x \in f(\theta)\}$.

Now we will pick the two subjects x and x' to show that the ratio $\pi^t(x)/\pi^t(x')$ is bounded. First, consider the convex hull of the set of subjects with $\pi^t(x) \geq 1 - \xi$, that is, $D := \text{co}\{\text{x} : \text{ß}^\text{t}(\text{x}) \geq 1 - \text{¸}\}$. Let $(\hat{\lambda}, \hat{\gamma}) \in \mathcal{S}^{n-1} \times [0,1]$ be a supporting hyperplane for D, that is, $\hat{\lambda}^T x \leq \hat{\gamma}$ for all $x \in D$ and there exists $\hat{x} \in D$ for which $\hat{\lambda}^T \hat{x} = \hat{\gamma}$. The half-line $\hat{x} + v\hat{\lambda}$ (where $v \geq 0$) is in the normal cone of the set D at $\bar{x}$. On this line we will choose three points. The first, x^0, will be chosen sufficiently close to $\bar{x}$ for $\pi^t(x^0) > 1 - 3\xi$ (which is possible by continuity) and the fact that $\pi^t(x) \geq 1 - 2\xi$ for all $x \in D$, (by (16) and $\pi^t(x^\omega) \geq \pi^t(x') + \pi^t(x'') - 1$). The second point x^m will be chosen so that the hyperplane $(\hat{\lambda}, \gamma^m) \in \mathcal{S}^{n-1} \times [0,1]$ through x^m ($\hat{\lambda}^T x^m = \gamma^m$) lies above every hyperplane through x^0 that does not intersect in $\mathcal{X}$ the supporting hyperplane $(\hat{\lambda}, \hat{\gamma})$. That is, if $\lambda^T x^m = \gamma$ and $H_{\lambda\gamma} \cap \mathcal{X} \subset H_{\hat{\lambda}\gamma^m} \cap \mathcal{X}$, then $H_{\hat{\lambda}\hat{\gamma}} \cap \mathcal{X} \subset H_{\lambda\gamma} \cap \mathcal{X}$. The final point on the half-line, x^1, will be chosen outside the halfspace $H_{\hat{\lambda}\gamma^m}$.

We use E to denote the event that x^0 is in the treatment set, but the points on the plane $\hat{\lambda}^T x = \gamma^m$ are not: formally $E = \{\tilde{\theta} : x^0 \in \tilde{\theta}, \hat{\lambda}^T x = \gamma^m\ \forall x \in \tilde{\theta}\}$. And we use $\pi^t(E)$ to denote the prior probability of this event. Now we consider the ratio

$$\frac{\pi^t(x^1)}{\pi^t(E)} = \frac{\int_{\Theta_{x^1}} \rho^{c_t(\theta)} d\mu}{\int_{\Theta_E} \rho^{c_t(\theta)} d\mu},$$

where $\Theta_E := \{\theta : x^m \in f(\theta) \subset H_{\hat{\lambda}\gamma^m}\}$. Suppose that after some time T only subjects, x_t, satisfying $\hat{\lambda}^T x_t \leq \hat{\gamma}$ are treated. The sets $\tilde{\theta} \in E$ all contain the half space $H_{\hat{\lambda}\hat{\gamma}}$ and so for $\theta \in \Theta_E$ we can write $c_t(\theta) = c_T(\theta) + f_t$ where $f_t = \tilde{c}_t(H_{\hat{\lambda}\hat{\gamma}}) - \tilde{c}_T(H_{\hat{\lambda}\hat{\gamma}})$ is the data collected in the halfspace $H_{\hat{\lambda}\hat{\gamma}}$ and is independent

of θ. Further define $\Theta'_{x^1} \subset \Theta_{x^1}$ to be those treatment sets that contain x^1 and the halfspace $H_{\hat{\lambda}\hat{\gamma}}$: that is, $\Theta'_{x^1} := \{\theta : x^1 \in f(\theta), H_{\hat{\lambda}\hat{\gamma}} \subset \mathcal{X} \subset f(\theta)\}$. For $\theta \in \Theta'_{x^1}$ we can also write $c_t(\theta) = c_T(\theta) + f_t$, hence

$$\frac{\pi^t(x^1)}{\pi^t(E)} = \frac{\int_{\Theta_{x^1}} \rho^{c_t(\theta)} d\mu}{\int_{\Theta_E} \rho^{c_t(\theta)} d\mu} \geq \frac{\int_{\Theta'_{x^1}} \rho^{c_T(\theta)+f_t} d\mu}{\int_{\Theta_E} \rho^{c_T(\theta)+f_t} d\mu} = \frac{\rho^{f_t}}{\rho^{f_t}} \frac{\int_{\Theta'_{x^1}} \rho^{c_T(\theta)} d\mu}{\int_{\Theta_E} \rho^{c_T(\theta)} d\mu}.$$

Thus this ratio is bounded below by a value that is independent of t and only depends on T. Thus it cannot shrink to zero as t increases. Hence there will always be points in the interval $\pi^t(x) \in [1/2, 1-\xi]$ does not shrink to zero, when only subjects satisfying $\pi^t(x_t) \geq 1-\xi$ are treated.

Step 4: By Step 2 there is a hyperplane $\bar{\lambda}^T x = \bar{\gamma}$ that is in the treatment set but not in $\mathcal{L}$, with positive probability (conditional on $\theta \in B(\theta^*, \eta)$). We will now show that there exists a point $x^\dagger$ on this hyperplane, such that $\pi^t(x^\dagger) \to 0$ on these positive probability set of states of the world.

Every time a subject x_t satisfying $\pi^t(x_t) \in (1/2, 1-\xi)$ is sampled, all $\check{x} \in C_{x_t}$ have

$$\left|\pi^t(x_t|\check{x}) - \pi^t(x_t)\right| = 1 - \pi^t(x_t) \geq \xi.$$

The hyperplane $\bar{\lambda}^T x = \bar{\gamma}$ has a strictly positive intersection with the cone C_{x_t} when x_t is sampled on the appropriate side on $T^0(\mu^t)$. By Step 3 there is an infinite number of x_t's sampled in this way on every positive probability event. Thus there exists a point $x^\dagger$ on the hyperplane $\bar{\lambda}^T x = \bar{\gamma}$ that is has infinitely many x_t's sampled with

$$\left|\pi^t(x_t|x^\dagger) - \pi^t(x_t)\right| = 1 - \pi^t(x_t) \geq \xi.$$

For such a point

$$\mathbb{1}_{\{x_t \in \Phi_{\check{x}} \cap T^0(\mu^t) \cap S(\mu^t)\}} \pi^t(x^\dagger)\xi \to 0, \qquad \mathbb{P}^0 \text{ almost surely.}$$

On a set of positive measure.

Step 5: We will now derive a contradiction to the claim (14). By Step 4 we have found $x^\dagger \in \tilde{\theta}^*_{-\nu}$, so that $\pi^t(x^\dagger) \to 0$ with probability at least $\varepsilon'' > 0$ conditional on $\theta \in B(\theta^*, \eta)$. But by (7)

$$E\left(\frac{1-\pi^{t+1}(x^\dagger)}{\pi^{t+1}(x^\dagger)} \;\middle|\; x^\dagger \in f(\theta), h_t\right) = \frac{1-\pi^t(x^\dagger)}{\pi^t(x^\dagger)}.$$

Taking an unconditional expectation this gives

$$E\left(\frac{1-\pi^{t+1}(x^\dagger)}{\pi^{t+1}(x^\dagger)} \;\middle|\; x^\dagger \in f(\theta)\right) = \frac{1-\pi^0(x^\dagger)}{\pi^0(x^\dagger)}.$$

Note that $x^\dagger \in f(\theta)$ for all $\theta \in B(\theta^*, \eta)$ and so we have show that on this positive measure set there is strictly positive probability that the LHS expectation is of a random variable that becomes arbitrarily large; as $\pi^{t+1}(x^\dagger) \to 0$; on a positive probability set. This is a contradiction, as the RHS is finite.

Thus our initial assertion (14) is false and $\tilde{\theta}^*_\eta \subseteq \mathcal{L}$ with probability one for all $\theta \in B(\theta^*, \eta)$. □

5. Asymptotic Myopia

In this section we describe a property of the optimal policy that we term asymptotic myopia. Informally, we mean by this that the behaviour of a policy maker who implements the optimal policy (for

some discount factor) converges to the behaviour of a policy maker who to treats a subject if and only if the the short run benefits of treatment outweigh the short run costs. Below we will give a result that says the optimal policy approaches the myopic policy as time passes. Results similar to this have been derived in other contexts (Woodroofe (1979) and Easley and Kiefer (1988)) and this appears to be a very robust property of Bayesian learning models. At the end of this section the example is used to show this property in action. That is, we show that the myopic policy is a very good approximation to the optimal policy when the discount factor tends to 1.

We will begin by giving some intuition for the result. Given a treatment set $\tilde{\theta}$ and a choice of policy, σ, the posteriors in periods $t = 0, 1, \ldots$ are a $\Delta(\Theta)$-valued stochastic process. We will use $\{\tilde{\mu}_t\}_{t=0}^{\infty}$ to denote this stochastic process (where each $\tilde{\mu}^t$ is a random variable with a realisation μ^t). The martingale property of the priors implies that this sequence of random variables $\{\tilde{\mu}_t\}$ converges almost surely to a limit (by Easley and Kiefer (1988)) . As this convergence occurs, the state in the next period (either μ_t^+, μ_t^-, or μ_t if no treatment occurred) is very unlikely to be much different from the state today. When this is combined with the continuity of the value function, this means that the extra value that is acquired from experimentation in the limit, converges to zero. So, the terms, $\rho_\mu(x)U(\mu^+) + (1-\rho_\mu(x))U(\mu^-) - U(\mu)$, in the integrand (5) tend almost surely to zero along every stationary policy including the optimal one. Thus, the set of subjects x for which

$$H_\mu(x) \equiv r\beta(2\pi_\mu(x) - 1) + \rho_\mu(x)U(\mu^+) + (1-\rho_\mu(x))U(\mu^-) - U(\mu) \geq 0$$

becomes closer and closer to the set of subjects, x, for which

$$2\pi_\mu(x) - 1 \geq 0.$$

This implies that over time the optimal policy approaches the myopic policy.

In fact we will show a slightly stronger result by showing that the optimal policy approaches the the myopic policy with a uniformity property. The statement in Lemma 4 is required to be probabilistic, however, because even after very many periods it is possible that the history of observations has been sufficiently perverse that a substantial amount of learning has yet to occur.

To state and prove our result we will define two potential sets of subjects that will be treated for each state μ. The first is the set who would be treated under the optimal policy which we denote $T^*(\mu)$, the second is the set who would be treated by a policy maker implementing a policy that resembles the myopic one but which has a threshold $\frac{1}{2} - \eta$ for treatment which we denote $T^\eta(\mu)$.

$$T^*(\mu) = \{\, x : H_\mu(x) \geq 0 \,\} \qquad T^\eta(\mu) = \left\{ x : \pi_\mu(x) \geq \frac{1}{2} - \eta \right\} \tag{19}$$

To state the result some additional definitions are necessary. For a given initial value of θ, the policy σ^*, prior μ, and the uniform sampling measure υ generate a probability measure over the space of infinite histories of play $\mathcal{H}_\infty$. We will denote this measure $\mathbb{P}_\theta$. We will also use $\{\mathcal{F}_t\}_{t=0}^{\infty}$ to denote the filtration induced by the observations $h_t \in \mathcal{H}_t$ at the start of each period (before the current subject is sampled).

Now we can state our result, it says that the probability that $T^*(\mu^t) \subset T^\eta(\mu^t)$ for all periods $t > \tau$ can be made arbitrarily large if τ is chosen sufficiently big.

LEMMA 4: *For $\theta \in \Theta$, $\delta < 1$ and any $\eta > 0$ there exists τ (dependent on these values) such that*

$$\mathbb{P}_\theta \left(T^*(\mu^t) \subset T^\eta(\mu^t) \; \forall t > \tau \;\middle|\; \mathcal{F}_t \right) > 1 - \eta.$$

Proof. Fix an initial value of θ, a prior μ, and the policy σ^*. This generates a sequence of $\Delta(\Theta)$-valued random variables $\{\tilde{\mu}_t\}_{t=0}^\infty$, where $\tilde{\mu}_0 = \mu$, that are the random posteriors at the start of period t. Furthermore, if U is the function defined by (4) the $[0,1]$-valued sequence of random variables $\{U(\mu_t)\}_{t=0}^\infty$ is a submartingale, relative to the filtration $\{\mathcal{F}_t\}_{t=0}^\infty$. This follows as for each x_t sampled we will have one of:

$$\begin{aligned} U(\mu_t) &\leq \rho_\mu(x_t) U\left(\mu_t^+\right) + (1 - \rho_\mu(x_t)) U\left(\mu_t^-\right), \\ U(\mu_t) &= U(\mu_{t+1}). \end{aligned}$$

As $\{U(\mu_t)\}_{t=0}^\infty$ is a bounded submartingale it converges $\mathbb{P}_\theta$ almost surely. This convergence implies that

$$E_\theta \left(\sup_{s \geq t} |U(\mu_s) - U(\mu_t)| \;\middle|\; \mathcal{F}_t \right) \to 0, \qquad \mathbb{P}_\theta \text{ almost surely.} \tag{20}$$

To see this, define the random variable U^∞ to be the almost sure limit of the sequence of random variables $\{U(\mu_t)\}$ and let $Z_t := \sup_{s \geq t} |U(\mu_s) - U^\infty|$. As Z_t is a non-increasing sequence of random variables converging almost surely to zero, it follows that $E(Z_t \mid \mathcal{F}_t)$ is a bounded supermartingale, that must also, then, converge to some limit (say Z^∞). But taking unconditional expectations we have $E(E(Z_t \mid \mathcal{H}_t)) = E(Z_t)$ and $E(Z_t) \to 0$. It follows that $E(Z_t \mid \mathcal{H}_t) \to Z^\infty = 0$ and as $2Z_t \geq \sup_{s \geq t} |U(\mu_s) - U(\mu_t)|$ the claim follows.

Applying Egoroff's Theorem to the almost sure convergence in (20), (see Royden (1988) p.72) implies that for any $\eta > 0$ and $K > 0$ there exists a τ such that for all $t > \tau$

$$\mathbb{P}_\theta \left(\sup_{s \geq t} |U(\mu_s) - U(\mu_t)| < K\frac{\eta}{2} \;\middle|\; \mathcal{F}_t \right) \geq 1 - \eta.$$

If $\sup_{s \geq t} |U(\mu_s) - U(\mu_t)| < \frac{\eta}{2}$ then $|U(\mu_{s+1}) - U(\mu_s)| < \eta$ for all $s \geq t$. Hence we have for all $t > \tau$

$$\mathbb{P}_\theta \left(\sup_{s \geq t} |U(\mu_{s+1}) - U(\mu_s)| < K\eta \;\middle|\; \mathcal{F}_t \right) \geq 1 - \eta,$$

Now let us consider the condition $H_{\mu_t}(x) \geq 0$. This can be written as

$$\begin{aligned} 0 &\leq r\beta(2\pi_{\mu_t}(x) - 1) + \rho_{\mu_t}(x)[U(\mu_t^+) - U(\mu_t)] + (1 - \rho_{\mu_t}(x))[U(\mu_t^-) - U(\mu_t)] \\ &\leq r\beta(2\pi_{\mu_t}(x) - 1) + \max\left\{|U(\mu_t^+) - U(\mu_t)|, |U(\mu_t^-) - U(\mu_t)|\right\} \end{aligned}$$

Thus if $|U(\mu_{t+1}) - U(\mu_t)| < \eta$ then $H_{\mu_t}(x) \geq 0$ implies

$$0 \leq r\beta(2\pi_{\mu_t}(x) - 1) + K\eta.$$

Now choose $K = r\beta$ then we have $H_{\mu_t}(x) \geq 0$ implies $\pi_{\mu_t}(x) \geq \frac{1}{2} - \eta$.

Hence if $\sup_{s \geq t} |U(\mu_{s+1}) - U(\mu_s)| < K\eta$, then $T^*(\mu^s) \subset T^\eta(\mu^s)$ for all $s > t$. The result now follows.

□

In the context of the usual two-armed bandit model of Rothschild (1974) for example, this fact is not so interesting. It just says that when the experimenter is unlikely to learn any more they are happy with the arm they are currently playing. However, in the context of this model where there are a continuum of arms it is useful. This is because, it is always optimal to engage in a little experimentation of nearby untried arms it gives a useful lower bounds on the amount of exploration that can occur. It is this that was used in the theorems of the previous section. However, when agents are particularly patient their utility is mainly influenced the limit

5.1. Example 2.1 Continued

We have established that the optimal policy in our example satisfies the functional equation (6) which was then simplified in Lemma 2. I have not be able to explicitly solve the recursion in Lemma 2 for the value function for the optimal policy, although it may be possible to do this numerically.

In this section the payoff from two suboptimal stationary policies is calculated. These give lower bounds on $U(a,b)$. The two policies that can be evaluated are the myopic policy, σ^o, where a subject is treated iff it is SR optimal to do so, and the permanent experimentation policy where all subjects are treated. (We will denote this latter policy σ^∞.) This gives the two stationary policies with the treatment sets:

$$T^o(\mu) = \left\{ x : \pi_\mu(x) \geq \frac{1}{2} \right\}, \qquad T^\infty(\mu) = \left\{ x : \pi_\mu(x) > 0 \right\}. \tag{21}$$

The following Lemma gives the payoffs from these policies

LEMMA 5: *The policy σ^o has the value function*

$$U^0(a,b) = \frac{1}{2}(1+b) + 3r\left(1 - \left(1 + \frac{4r}{b-a}\right)\log\left(1 + \frac{b-a}{4r}\right)\right),$$

and the policy σ^∞ has the value function

$$U^\infty(a,b) = \frac{1}{2}(1+b) + 2r\left(1 - \left(1 + \frac{2r}{b-a}\right)\log\left(1 + \frac{b-a}{2r}\right)\right).$$

where $r = (1-\delta)/\delta$.

The proof of this lemma is given in the Appendix. What is surprising here is which of these two lower bounds is tighter. In fact (when δ is close to unity) the better lower bound is given by the myopic policy σ^0. So we behaving myopically is good for the policy maker—particularly if they are quite patient! To make good on this claim we compare $U^0(0,1)$ and $U^\infty(0,1)$. If $a = 0$ and $b = 1$ are substituted into the expressions in Lemma 5 we get:

$$U^\infty(0,1) = 1 + 2r[1 - (1+2r)\log(1+1/2r)],$$
$$U^0(0,1) = 1 + 3r\left[1 - (1+4r)\log(1+1/4r)\right].$$

Taking a series expansion of these expressions in the neighbourhood of the origin, one gets:

$$U^\infty(0,1) = 1 + 2r + 2r\log r + r\log 4 + O(r^2),$$

$$U^0(0,1) = 1 + 3r + 3r\log(4r) + O(r^2).$$

This shows that for δ sufficiently close to unity (r sufficiently small) the value of following the myopic policy is greater than the value from the policy σ^∞.

REFERENCES

AGHION, P., P. BOLTON, C. HARRIS, AND B. JULLIEN (1991): "Optimal Learning by Experimentation," *Review of Economic Studies*, 58, 621–654.

AGRAWAL, R. (1995): "The Continuum-Armed Bandit Problem," *SIAM Journal of Control and Optimization*, 33, 1926–1951.

AL-NAJJAR, N. I. (2009): "Decision Makers as Statisticians: Diversity, Ambiguity, and Learning," *Econometrica*, 77(5), 1371–1401.

BANKS, J. S., AND R. K. SUNDARAM (1992): "Denumerable-Armed Bandits," *Econometrica*, 60, 1071–1096.

CALLANDER, S. (2011): "Searching and Learning by Trial and Error," *American Economic Review*, 101(6), 2277–2308.

CHAMBERLAIN, G. (2011): "Bayesian Aspects of Treatment Choice," in *The Oxford Handbook of Bayesian Econometrics*, ed. by J. Geweke, G. Koop, and H. van Dijk, pp. 11–39. Oxford University Press, Oxford, UK.

CHAN, T. Y., AND B. H. HAMILTON (2006): "Learning, Private Information and the Economic Evaluation of Randomized Experiments," *Journal of Political Economy*, 114(6), 997–1040.

DEGROOT, M. H. (1970): *Optimal Statistical Decisions*. McGraw-Hill, New York.

DEVROYE, L., L. GYÖRFI, AND G. LUGOSI (1996): *A Probabilistic Theory of Pattern Recognition*. Springer, Berlin.

DI TILLIO, A., M. OTTAVIANI, AND P. N. SØRENSEN (2015): "Persuasion Bias in Science: Can Economics Help?," Working paper, University of Bocconi, Innocenzo Gasparini Institute for Economic Research.

EASLEY, D., AND N. M. KIEFER (1988): "Controlling a Stochastic Process with Unknown Parameters," *Econometrica*, 56(5), 1045–1064.

GHOSH, J., AND R. RAMAMOORTHI (2003): *Bayesian Nonparmetrics*. Springer-Verlag, New York.

GLENNNERSTER, R., AND K. TAKAVARASHA (2013): *Running Randomized Evaluations: A Practical Guide*. Princeton University Press, Princeton, NJ.

GOLDENSHLUGER, A., AND A. ZEEVI (2009): "Woodroofe's One-Armed Bandit Problem Revisited," *Annals of Applied Probability*, 19(4), 1603–1633.

KITAGAWA, T., AND A. TETENOV (2016): "Who Should be Treated? Empirical Welfare Maximization Methods for Treatment Choice," *CEMMAP Working Paper*.

MAITRA, A. (1968): "Discounted Dynamic Programming on Compact Metric Spaces," *Sankhya, Series A*, 30, 211–216.

MANSKI, C. F. (2004): "Statistical Treatment Rules for Heterogenous Populations," *Econometrica*, 72, 1221–1246.

MCCLENNAN, A. (1984): "Price Dispersion and Incomplete Learning in the Long Run," *Journal of Economic Dynamics and Control*, 7, 331–347.

MOLCHANOV, I. (2005): *Theory of Random Sets*. Springer-Verlag, London, first edn.

PERCHET, V., AND P. RIGOLLET (2013): "The Multi-Armed Bandit Problem with Covariates," *Annals of Statistics*, 41(2), 693—721.

PERCHET, V., P. RIGOLLET, S. CHASSANG, AND E. SNOWBERG (2016): "Batched Bandit Problems," *Annals of Statistics*, 44(2), 660—681.

ROTHSCHILD, M. (1974): "A Two-Armed Bandit Theory of Market Pricing," *Journal of Economic Theory*, 9, 185–202.

ROYDEN, H. L. (1988): *Real Analysis*. Prentice Hall, Englewood Cliffs, NJ, 3rd edn.

WOODROOFE, M. (1979): "A One-armed Bandit Problem with a Concomitant Variable," *Journal of the American Statistical Association*, 74(368), 799–806.

A. APPENDIX

A.1. Proof of Lemma 1

Proof. Consider a policy maker who has observed the subject x and is deciding whether to treat them or not. If θ were known, then the expected payoff from treating subject x is $\beta(2\mathbf{1}_{x\in f(\theta)}-1)$. Thus, the flow utility of the policy maker is the function

$$u(x,\theta,d)=\begin{cases}0 & \text{if } d=0,\\ \beta(2\mathbf{1}_{x\in f(\theta)}-1) & \text{if } d=1.\end{cases}$$

(Here d denotes the decision to treat ($d=1$) or not ($d=0$).) As $f(\theta)$ is a closed set, the function $u(.)$ is upper semi continuous on $\mathcal{X}\times\Theta\times\{0,1\}$. When the policy maker has the beliefs μ, the expected payoff from not treating the subject is zero and the expected payoff from treating them is $\beta(2\pi_\mu(x)-1)$.

The prior μ on the parameters θ is subject to Bayesian updating when a treatment decision is taken. The up-date is random (it depends on the outcome of treatment), so we define the map $q_1:\Delta(\Theta)\times\mathcal{X}\to\Delta(\Delta(\Theta))$ that describes the distribution over posteriors induced when treatment is chosen:

$$q_1(\mu,x)=\begin{cases}\mu^+ & \text{with probability } \rho_\mu(x);\\ \mu^- & \text{with probability } 1-\rho_\mu(x).\end{cases}$$

When no treatment is applied there is no updating, so define $q_0(\mu,x)$ to be the probability measure in $\Delta(\Delta(\Theta))$ that puts probability one on μ. Together q_0 and q_1 define a map $q:\Delta(\Theta)\times\mathcal{X}\times\{0,1\}\to\Delta(\Delta(\Theta))$ that the prior μ, observation x, and treatment decision and map them to a distribution over posteriors. By Easley and Kiefer (1988) Theorem 2 p.1050 the map q is continuous (given the weak topology on $\Delta(\Delta(\Theta))$).

Initially, we treat $\Delta(\Theta)\times\mathcal{X}$ as the state space in our optimisation, with a value function

$$U(\mu,x):=\max_\sigma E_{\mu\sigma}\left(\sum_{t=0}^{\infty}\delta^t y_t \mid x_0=x\right).$$

In this problem the utility function is bounded and upper semi continuous, the state space $\Delta(\Theta)\times\mathcal{X}$ is Polish, the state updating relation q is continuous, and the action space is discrete; the result of Maitra (1968) p.216 applies. From this we can conclude: (1) there exists a stationary optimal policy, (2) the value function $U(\mu,x)$ is upper semi continuous on $\Delta(\Theta)\times\mathcal{X}$, and (3) it is the unique solution to the HJB equation.

$$U(\mu,x)=\max\left\{\delta E_v U(\mu,x),(1-\delta)\beta(2\pi_\mu-1)+\delta\rho_\mu E_v U(\mu^+,x)+\delta(1-\rho_\mu)E_v U(\mu^-,x)\right\}$$

Where E_v represents an expectation taken over the measure v on x and we have suppressed the arguments of π_μ and ρ_μ in our notation. We define $U(\mu)=E_v U(\mu,x)$, the above implies that $U(\mu)$ is the unique solution to

$$U(\mu)=E_v\left[\max\left\{\delta U(\mu),(1-\delta)\beta(2\pi_\mu-1)+\delta\rho_\mu U(\mu^+)+\delta(1-\rho_\mu)U(\mu^-)\right\}\right]. \tag{A.1}$$

The function $U(\mu)$ is the fixed point of an operator that maps continuous functions to continuous functions, so it is continuous.

Subtracting $\delta U(\mu)$ from both sides and writing $r=(1-\delta)/\delta$ we get

$$rU(\mu)=\int_{\mathcal{X}}\left[r\beta(2\pi_\mu(x)-1)+\rho_\mu(x)U(\mu^+)+(1-\rho_\mu(x))U(\mu^-)-U(\mu)\right]^+dv,$$

where $[x]^+:=\max\{0,x\}$. This is the result claimed above. □

A.2. Proof of Lemma 2

First make a substitution for π_{ab} into (6). Then, define $W(a,b) = (b-a)[U(a,b) - \frac{1}{2}(a+1)]$ and substitute this for $U(a,b)$. As a result of these transformations (6) becomes

$$rW(a,b) = \frac{1}{2}\int_a^b \left[r(b+a-2x) + \frac{1}{2}(b-x)(x-a) + W(x,b) + W(a,x) - W(a,b)\right]^+ dx.$$

Now we can turn this into a 1-dimensional problem. Letting $y = x - a$ and changing the variable in the integrand, gives

$$rW(a,b) = \frac{1}{2}\int_0^{b-a} \left[r(b-a-2y) + \frac{y(b-a-y)}{2} + W(a+y,b) + W(a,a+y) - W(a,b)\right]^+ dy.$$

Observe that this implies $W(a,b)$ only depends on $b-a$. Hence, it is possible to transform the recursion using the substitution $W(a,b) \equiv r^2 D(b-a) + \frac{1}{2}(b-a)^2$. Getting a new recursion:

$$D(b-a) + \frac{(b-a)^2}{2r^2} = \frac{1}{2r}\int_0^{b-a} \left[\frac{1}{r}(b-a-2y) + D(b-a-y) + D(y) - D(b-a)\right]^+ dy.$$

Now perform a final change of variable so that $(b-a)/r = v$, $y/r = u$ and $D(b-a) \equiv C(v)$, then this equation reduces to

$$C(v) + \frac{v^2}{2} = \frac{1}{2}\int_0^v [v - 2u + C(v-u) + C(u) - C(v)]^+ du. \tag{A.2}$$

Reversing all the transformations used here we get the relationship between $C(.)$ and $W(.)$ claimed in the Lemma.

A.3. Proof of Lemma 3

Proof. When no treatment occurs the equalities (7) are trivially true. When treatment occurs at x_t, then a substitution from the monitoring technology gives:

$$\pi^{t+1}(x) = \pi^t(x)\frac{(1+\beta)\pi^t(x_t|x) + (1-\beta)(1-\pi^t(x_t|x))}{(1+\beta)\pi^t(x_t) + (1-\beta)(1-\pi^t(x_t))} \qquad \text{if } y_t = 1,$$
$$\pi^{t+1}(x) = \pi^t(x)\frac{(1-\beta)\pi^t(x_t|x) + (1+\beta)(1-\pi^t(x_t|x))}{(1-\beta)\pi^t(x_t) + (1+\beta)(1-\pi^t(x_t))} \qquad \text{if } y_t = -1.$$

Hence the updated likelihood ratios are:

$$\frac{1-\pi^{t+1}(x)}{\pi^{t+1}(x)} = \frac{2\beta\pi^t(x)(1-\pi^t(x_t|x)) + (1-\pi^t(x))(1-\beta)}{\pi^t(x)(1+\beta)\pi^t(x_t|x) + \pi^t(x)(1-\beta)(1-\pi^t(x_t|x))} \qquad \text{if } y_t = 1,$$
$$\frac{1-\pi^{t+1}(x)}{\pi^{t+1}(x)} = \frac{-2\beta\pi^t(x)(1-\pi^t(x_t|x)) + (1-\pi^t(x))(1+\beta)}{\pi^t(x)(1-\beta)\pi^t(x_t|x) + \pi^t(x)(1+\beta)(1-\pi^t(x_t|x))} \qquad \text{if } y_t = -1.$$

A simple weighted sum of these expectations (or their reciprocals) then gives

$$E\left(\frac{1-\pi^{t+1}(x)}{\pi^{t+1}(x)} \,\middle|\, x \in \tilde{\theta}, x_t, h_t\right) = \frac{1-\pi^t(x)}{\pi^t(x)}, \qquad E\left(\frac{\pi^{t+1}(x)}{1-\pi^{t+1}(x)} \,\middle|\, x \notin \tilde{\theta}, x_t, h_t\right) = \frac{\pi^t(x)}{1-\pi^t(x)}.$$

As these hold for all x_t we can take an expectation over x_t which gives the result that conditional on $x \in f(\theta)$ the odds ratio $\frac{1-\pi^t(x)}{\pi^t(x)}$ is a martingale and conditional on $x \notin f(\theta)$ the odds ratio $\frac{\pi^t(x)}{1-\pi^t(x)}$ is a martingale.

Now we suppose that subject x_t is treated. Taking a difference We, also, have

$$\pi^{t+1}(x) - \pi^t(x) = \frac{2\beta\pi^t(x)(\pi^t(x_t|x) - \pi^t(x_t))}{(1+\beta)\pi^t(x_t) + (1-\beta)(1-\pi^t(x_t))} \qquad \text{if } y_t = 1,$$

$$\pi^{t+1}(x) - \pi^t(x) = \frac{-2\beta\pi^t(x)(\pi^t(x_t|x) - \pi^t(x_t))}{(1-\beta)\pi^t(x_t) + (1+\beta)(1-\pi^t(x_t))} \qquad \text{if } y_t = -1.$$

As the denominator in the fractions above is less than unity this implies

$$\left|\pi^{t+1}(x) - \pi^t(x)\right| \geq 2\beta\pi^t(x)\left|\pi^t(x_t|x) - \pi^t(x_t)\right|.$$

This proves (8). □

A.4. Proof of Lemma 5

Proof. Under the myopic policy σ^0 the recursion for the value function, $U^0(a,b)$, would satisfy

$$rU^0(a,b) = \frac{1}{2}\int_a^{\frac{1}{2}(b+a)} r(2\pi_{ab} - 1) + U^0(x,b)\pi_{ab}(x) + U^0(a,x)(1-\pi_{ab}) - U^0(a,b)dx + \frac{1}{2}(a+1)r.$$

If this is transformed using $v = (b-a)/r$ and $U^\infty(a,b) = \frac{1}{2}(1+b) + r^2C(v)/(b-a)$, into a single variable problem in the manner described by Lemma 2 this becomes

$$C(v) + \frac{1}{2}v^2 = \frac{1}{2}\int_0^{v/2} v - 2u + C(v-u) + C(u) - C(v)du.$$

Differentiating with respect to v (and the boundary condition $C(0) = 0$ is applied) we can derive and solve a differential equation for C.

$$\begin{aligned} & (4+v)C'(v) - C(v) = -3v \\ \Rightarrow \quad & C(v) = 3v - 3(4+v)\log(1 + (v/4)) \end{aligned}$$

If is returned to the original form by using the transformation $v = (b-a)/r$ and $U^0(a,b) = \frac{1}{2}(1+b) + r^2C(v)/(b-a)$, we get

$$U^0(a,b) = \frac{1}{2}(1+b) + 3r\left(1 - \left(1 + \frac{4r}{b-a}\right)\log\left(1 + \frac{b-a}{4r}\right)\right).$$

This is what was given in the Lemma.

Under the policy σ^∞ the recursion for the value function, $U^\infty(a,b)$, would satisfy

$$rU^\infty(a,b) = \frac{1}{2}\int_a^{b} r(2\pi_{ab} - 1) + U^\infty(x,b)\pi_{ab}(x) + U^\infty(a,x)(1-\pi_{ab}) - U^\infty(a,b)dx + \frac{1}{2}(a+1)r.$$

If this is transformed using $v = (b-a)/r$ and $U^\infty(a,b) = \frac{1}{2}(1+b) + r^2C(v)/(b-a)$, into a single variable problem in the manner described by Lemma 2 this becomes

$$C(v) + \frac{1}{2}v^2 = \frac{1}{2}\int_0^{v} v - 2u + C(v-u) + C(u) - C(v)du.$$

Differentiation with respect to v and the boundary condition $C(0) = 0$ allows us to derive and solve a differential equation for C.

$$\begin{aligned} & (2+v)C'(v) - C(v) = -2v, \qquad C(0) = 0, \\ \Rightarrow \quad & C(v) = 2v - 2(v+2)\log(1 + (v/2)). \end{aligned}$$

If is returned to the original form by substituting $v = (b-a)/r$ and $U^{\infty}(a,b) = \frac{1}{2}(1+b) + r^2 C(v)/(b-a)$, we get

$$U^{\infty}(a,b) = \frac{1}{2}(1+b) + 2r\left(1 - \left(1 + \frac{2r}{b-a}\right)\log\left(1 + \frac{b-a}{2r}\right)\right).$$

Which is what was claimed in the Lemma. □