Biometrics 000, 000–000 DOI: 000
000 0000

Thompson Sampling for mHealth and Precision Health Applications

John Sperger, Eric B. Laber, Michael R. Kosorok

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

1. Introduction

Technological advancements in mobile-health (mHealth) have made it possible to deliver personalized healthcare at scale. Mobile devices, such as smart phones, allow patients to receive care if, where, and when it is needed, while wearables, such as continuous glucose monitors and accelerometers, allow for efficient collection of rich patient-level data which can be used to tailor and refine treatment decisions. Recent mHealth studies cover a wide range of diseases and disorders including addiction (Carpenter et al., 2020), diet and exercise planning for persons with type I diabetes (Luckett et al., 2019), supportive care for cancer pain (Fisher et al., 2021), and HIV/STI prevention (Mustanski et al., 2022).

However, while it is increasingly recognized that mHealth holds immense potential for scaling and democratizing healthcare (Hernández-Neuta et al., 2019), clinical trials targeting the evaluation and optimization of personalized mHealth-based interventions largely remains in the purview of a small number of specialists. A primary goal of this chapter is to provide an accessible introduction to Thompson Sampling (TS Thompson, 1933; Russo et al., 2017) as a framework for adaptive randomization in sequential decision problems with a long or indefinite horizon when the goal is to maximize cumulative utility (e.g., patient benefit in mHealth). TS is general and extensible. It applies with continuous, categorical, or time-to-event data with censoring, under a Bayesian or frequentist paradigm, and with parametric, semi-parametric, or non-parametric models. Thus, we believe TS makes an excellent starting point for researchers designing an adaptive trial in mHealth and other settings with many decision points.

We present TS from the perspective of precision medicine and clinical trial design though, as anticipated by its generality, it is applicable much more broadly (see Russo et al. (2017), Lattimore and Szepesvári (2020), and Slivkins et al. (2019)). TS was first proposed by Thompson (1933) nearly a century ago for adaptive treatment allocation with binary treat-

ments. This seminal paper was the antecedent to long and fruitful lines of work on multi-arm bandits (Robbins, 1956), sequential designs (Wald, 1947; Robbins, 1952; Chernoff, 1959), and adaptive trials (Armitage, 1960). Furthermore, TS has been a central idea of Bayesian adaptive design for decades (Berry and Fristedt, 1985) and has been applied in multiple cancer clinical trials (see Trippa et al., 2012; Thall and Wathen, 2005, and references therein). Nevertheless, rigorous theoretical and empirical study of TS in realistic environments occurred only in the past 15 years or so. (Chapelle and Li, 2011; Agrawal and Goyal, 2012, 2013). Distributional approximations and inferential procedures have been developed even more recently (Zhang et al., 2020; Hadad et al., 2021; Bibaut et al., 2021; Wager and Xu, 2021; Zhang et al., 2022). These recent innovations have been instrumental in driving more widespread adoption of TS in sequential decision problems.

The remainder of this chapter is organized as follows. In Section 2, as means of building intuition, we introduce TS using a simple adaptive trial with two treatments. In Section 3, we present a general version of TS. In Section 4, we discuss some of the statistical properties of TS including regret, power, inference and prior sensitivity. In Section 5, we discuss some practical considerations with modern applications of TS. We close with a summary and brief discussion of future research directions in Section 6.

2. Thompson Sampling in the simplest case

To introduce Thompson Sampling (TS), we begin with the original scenario considered by Thompson (1933) in which a set of patients (all with the same ailment) arrive at the clinic one-by-one and, upon arriving, are assigned one of two treatments the result of which is either a success or a failure. In this simple setting, the outcome of the present patient is observed before the next one arrives. Thus, the clinician has available the treatment and outcome history of all prior patients to inform each treatment decision. The goal is to allocate treatments in such a way that the expected number of successes (ENS) is maximized.

It can be seen that traditional one-to-one randomization only maximizes ENS if there is no difference in the success rate of the two treatments (in which case, any allocation of the treatments is equally good). Conversely, greedy selection, i.e., always selecting the treatment with the highest estimated probability of success (say, after a burn-in period of equal allocation), need not optimize ENS as it can become stuck allocating the inferior treatment by chance. An ENS-maximizing allocation strategy must balance optimization based on current information with experimentation (choosing the estimated suboptimal treatment). The need to strike this balance is known as the 'exploration-exploitation' dilemma, and is at the heart of sequential decision making (Sutton and Barto, 2018). Intuitively, as evidence accumulates that a given treatment is optimal, an optimal adaptive treatment allocation strategy should become more likely to recommend that treatment. TS operationalizes this intuition by setting the probability of treatment assignment at each time step equal to the posterior probability that the treatment is optimal; for this reason, TS is sometimes called posterior probability matching. Under TS, as the posterior becomes increasingly concentrated on the true parameter values, the probability of assigning an optimal treatment will increase to one.

To make these ideas concrete, consider a trial which will enroll a total of n subjects. Each subject, $t = 1, \ldots, n$, will be assigned a treatment $A_t \in \mathcal{A} = \{0, 1\}$ and their outcome, success or failure, subsequently observed. The 'complete' set of outcomes are $\{(Y_t^0, Y_t^1)\}_{t=1}^n$ which comprise n i.i.d. copies of (Y^0, Y^1), where Y^a is the potential outcome under treatment $a \in \mathcal{A}$. The observed outcome for subject t is $Y_t = A_t Y_t^1 + (1 - A_t) Y_t^0$, i.e., the observed outcome is the potential outcome under the treatment actually given.

Let $\mu_a = P(Y^a = 1)$ be the success probability under treatment a and let N_a denote the (random) number of subjects assigned to treatment a at the completion of the trial so that $n = N_0 + N_1$. The ENS is thus $\mathcal{E}_n = \mathbb{E}(N_0)\mu_0 + \mathbb{E}(N_1)\mu_1$. Application of classic,

aka Bayesian, TS requires a prior $p_0(\mu_0, \mu_1)$ over the success probabilities. A natural choice is to specify independent beta distributions for the success probabilities, i.e., $p(\mu_0, \mu_1) = \rho(\mu_0; \alpha_{0,0}, \beta_{0,0})\rho(\mu_1; \alpha_{1,0}, \beta_{1,0})$ with

$$\rho(\mu_a; \alpha_{a,0}, \beta_{a,0}) = \frac{1}{B(\alpha_{a,0}, \beta_{a,0})} \mu_a^{\alpha_{a,0}-1}(1-\mu_a)^{\beta_{a,0}-1},$$

for $\mu_a \in [0,1]$, where $B(\alpha_{a,0}, \beta_{a,0}) = \Gamma(\alpha_{a,0})\Gamma(\beta_{a,0})/\Gamma(\alpha_{a,0}, \beta_{a,0})$ and $\alpha_{a,0}, \beta_{a,0} \geqslant 0$ are hyper-parameters.

Under this model, the posterior for μ_a after processing the outcome for the tth subject follows a beta distribution with parameters

$$\begin{aligned} \alpha_{a,t} &= \alpha_{a,t-1} + Y_t 1_{A_t=a} \\ \beta_{a,t} &= \beta_{a,t-1} + (1-Y_t) 1_{A_t=a}, \end{aligned}$$

where 1_u is an indicator that the clause u is true. Use an overline to denote history so that $\overline{\mathbf{A}}_t = (A_1, \ldots, A_t)$ and $\overline{\mathbf{Y}}_t = (Y_1, \ldots, Y_t)$. Thus, when the tth subject enters the trial, the information available to the clinician is $\mathcal{H}^b_{t-1} = (\overline{\mathbf{A}}_{t-1}, \overline{\mathbf{Y}}_{t-1})$, where $\mathcal{H}^b_0 = \emptyset$ (the superscript 'b' is a mnemonic for bandit). Under TS, when the tth subject enters the trial, they are assigned treatment a with probability $P(\mu_a \geqslant \mu_{1-a} | \mathcal{H}^b_{t-1})$ (we assume the posterior probability that $\mu_a = \mu_{1-a}$ is zero, if not, one can use random tie-breaking). Thus, one could compute $\theta_{0,t} = P(\mu_0 \geqslant \mu_1 | \mathcal{H}^b_{t-1})$ and then draw $A_t \sim \text{Bernoulli}(1-\theta_{0,t})$. An implementation that is simpler, especially in more complex settings, is to draw a sample from the posterior, say $\widetilde{\mu}_{a,t} \sim \rho(\mu_a; \alpha_{a,t-1}, \beta_{a,t-1})$ for each $a \in \mathcal{A}$, and then to select treatment so as to optimize the mean outcome if the sampled parameters were correct, e.g., $A_t = \arg\max_a \widetilde{\mu}_{a,t}$. Algorithm 1 provides a schematic for using TS in an adaptive trial with n subjects.

2.1 *Simple TS with clipping*

The preceding version of TS does not place guardrails on the treatment assignment probabilities, i.e., the probability of assigning one treatment may converge to zero or one. In clinical

Algorithm 1: Beta-Bernoulli TS adaptive trial with two treatments

Input: Hyperparameters $(\alpha_{0,0}, \beta_{0,0}, \alpha_{1,0}, \beta_{1,0})$, trial size n

for *Subjects* $t = 1, \ldots, n$ **do**

  for *Treatments* $a = 0, 1$ **do**

    Sample $\tilde{\mu}_{a,t} \sim \rho(\mu_a; \alpha_{a,t-1}, \beta_{a,t-1})$

  Assign treatment $A_t = \arg\max_a \tilde{\mu}_{a,t}$

  Observe outcome Y_t

  Update posterior parameters

$$\alpha_{a,t} = \alpha_{a,t-1} + Y_t 1_{A_t=a}$$

$$\beta_{a,t} = \beta_{a,t-1} + (1 - Y_t) 1_{A_t=a}$$

settings in which there are a multiple secondary analyses of interest, this behavior may not be desirable as we might not have sufficient power for these analyses. A common remedy is to truncate (i.e., clip) the probabilities to the interval $[c_0, c_1]$ where $0 < c_0 < c_1 < 1$ (Zhang et al., 2020). Clipped-TS will select action a with probability

$$P(A_t = a|\mathcal{H}^b_{t-1}) = \max\left[c_0, \min\left\{c_1, P\left(a = \arg\max_{a'} \mu_{a'}|\mathcal{H}^b_{t-1}\right)\right\}\right].$$

Thus, under clipped-TS, the expected number of subjects assigned to treatment a is bounded below by $c_0 n$ and above by $c_1 n$.

Because Clipped-TS, as described, requires explicitly computing the probabilities of each action, rather than simply computing a draw from the posterior, it can be burdensome to execute in more complex settings. One sampling-based approach that ensures each treatment is selected with some minimal probability is ϵ-TS (Li et al., 2022). In ϵ-TS, for a pre-specified value $\epsilon \in (0, 1)$, when subject t enters the trial they are assigned treatment according to TS with probability $(1-\epsilon)$ and they are assigned treatment uniformly at random with probality ϵ. Thus, under ϵ-TS with two treatments, the probability of assigning treatment a at time t is always bounded below by $\epsilon/2$. This strategy of mixing uniform random treatment assignment

with TS can be applied much more broadly, e.g., with continuous treatments or treatment sets which depend on the decision context (see Section 3).

2.2 *Simple TS in basket trials*

To illustrate how TS can be easily extended to more complex trial designs, we consider a hypothetical basket trial of cancer therapeutic agents. This hypothetical trial begins with two agents, say $\mathcal{A}_0 = \{0, 1\}$, but after (say) the 76th subject is processed, a new agent becomes available so that our set of allowable agents becomes $\mathcal{A}_1 = \{0, 1, 2\}$. Further suppose that at the time the new agent is introduced, there have been 39 successes and 9 failures under agent 0, and 17 successes and 11 failures under agent 1. Assuming uniform priors for all three agents, the posterior distributions for the success probabilities, μ_0, μ_1, μ_2, are Beta$(40, 10)$, Beta$(18, 10)$, and Beta$(1, 1)$ respectively. When the 77th subject enters the trial, the treatment assignment probabilities for each arm are approximately 78%, 2%, and 20% for treatments 0,1, and 2 respectively. The estimated mean for treatment 1 is 0.6 while the estimated mean for treatment 2 is 0.5, yet treatment 3 is ten times more likely to be selected. This illustrates the dependence of Thompson Sampling on both the estimated mean *and* uncertainty around this estimated mean.

[Figure 1 about here.]

3. Beyond the simplest case: contextual bandits

In the preceding section, we considered a one-size-fits-all approach to treatment selection, i.e., the goal was to identify a single treatment which was best (on average) for the entire population. We now consider the setting in which treatment is tailored to individual patient characteristics. As in the preceding section, we consider a trial in which a total of n subjects will be enrolled. However, we now assume that the data generated by the trial will be of the form $\{(\mathbf{X}_t, A_t, Y_t)\}_{t=1}^n$, where $\mathbf{X}_t \in \mathcal{X} \subseteq \mathbb{R}^p$ are characteristics for the tth subject,

$A_t \in \mathcal{A} = \{0, 1\}$ is their assigned intervention, and $Y_t \in \mathbb{R}$ is their outcome, coded so that higher values are preferred.

Let Y_t^a denote the potential outcome for the tth subject under treatment $a \in \mathcal{A}$. Under the contextual bandit model, the set of contexts and all potential outcomes $\{(\mathbf{X}_t, Y_t^0, Y_t^1)\}_{t=1}^n$ are assumed to be independent copies of $(\mathbf{X}, Y^0, Y^1)$. We assume the following standard causal conditions hold: (i) no unmeasured confounders, $(Y^0, Y^1) \perp A \big| \mathbf{X}$; (ii) consistency, $Y = Y^A$; and (iii) positivity, there exists $\epsilon > 0$ such that $P(A = a|\mathbf{X} = \mathbf{x}) \geqslant \epsilon$ for all $a \in \mathcal{A}$ and $\mathbf{x} \in \mathcal{X}$. In addition, we assume there is no interference nor are there multiple versions of treatment (Hernan and Robins, 2020).

For simplicity, we assume that subjects enroll one-by-one so that the outcome for one subject is observed before the next one is assigned their treatment. As previously, use an overline to denote history so that $\overline{\mathbf{X}}_t = (\mathbf{X}_1, \ldots, \mathbf{X}_t)$, $\overline{\mathbf{A}}_t = (A_1, \ldots, A_t)$, and $\overline{\mathbf{Y}}_t = (Y_1, \ldots, Y_t)$. Define $\mathcal{H}_{t-1}^c = (\overline{\mathbf{X}}_{t-1}, \overline{\mathbf{A}}_{t-1}, \overline{\mathbf{Y}}_{t-1})$ to be the information collected through the first $(t-1)$ subjects with $\mathcal{H}_0^c = \emptyset$ (the superscript 'c' is a mnemonic for contextual bandit). The information available to a clinician in selecting treatment for the tth subject is thus $\mathcal{H}_t^{c-} = (\mathcal{H}_{t-1}^c, \mathbf{X}_t)$.

To illustrate TS in this setting, we assume a linear model for the outcome of the form

$$Y_t = \psi(\mathbf{X}_t, A_t)^\intercal \boldsymbol{\gamma} + \epsilon_t,$$

where $\psi(\mathbf{X}_t, A_t) \in \mathbb{R}^q$ is a feature vector constructed from $\mathbf{X}_t$ and A_t, $\boldsymbol{\gamma} \in \boldsymbol{\Gamma} \subseteq \mathbb{R}^q$ is a vector of unknown coefficients, and ϵ_t is an independent error term with mean zero and finite variance. We assume that $\epsilon_1, \ldots, \epsilon_n$ are drawn *i.i.d.* from a distribution with density $f(\epsilon; \boldsymbol{\eta})$ which is indexed by unknown parameters $\boldsymbol{\eta} \in \mathcal{N} \subseteq \mathbb{R}^d$. Set $\boldsymbol{\theta} = (\boldsymbol{\gamma}^\intercal, \boldsymbol{\eta}^\intercal)^\intercal \in \boldsymbol{\Theta} = \boldsymbol{\Gamma} \times \mathcal{N}$. To apply TS, we specify a prior $\rho(\boldsymbol{\theta})$ over $\boldsymbol{\Theta}$. When the tth subject arrives, presenting with context $\mathbf{X}_t$, treatment a is selected with probability

$$P(A_t = a|\mathcal{H}_t^{c-}) = P\left\{a = \arg\max_{a'} \psi(\mathbf{X}_t, a')^\intercal \boldsymbol{\theta} \Big| \mathcal{H}_t^{c-}\right\}.$$

For example, we might assume normal errors so that $\epsilon_t \sim \text{Normal}(0, \tau^{-1})$ and assume an

improper prior of the form $\rho(\boldsymbol{\theta}) \propto \tau^{-1}$. In this case, the posterior distribution for $\boldsymbol{\gamma}$ given $\mathcal{H}_t^{c-}$ follows a multivarite t-distribution with $t-1-q$ degrees of freedom, centered at the OLS estimator

$$\widehat{\boldsymbol{\gamma}}_{t-1} = \left\{ \sum_{v=1}^{t-1} \psi(\mathbf{X}_v, A_v)\psi(\mathbf{X}_v, A_v)^\intercal \right\}^{-1} \sum_{v=1}^{t-1} \psi(\mathbf{X}_v, A_v) Y_v, \tag{3.1}$$

and with variance equal to

$$\widehat{\Sigma}_{t-1} = \widehat{\sigma}_{t-1}^2 \left\{ \sum_{v=1}^{t-1} \psi(\mathbf{X}_v, A_v)\psi(\mathbf{X}_v, A_v)^\intercal \right\}^{-1},$$

where $\widehat{\sigma}_{t-1}^2 = (t-1-q)^{-1} \sum_{v=1}^{t-1} \left\{ Y_v - \psi(\mathbf{X}_v, A_v)^\intercal \widehat{\boldsymbol{\theta}}_{t-1} \right\}^2$, is the usual estimator of the residual variance.

In practice, under the above prior, $\boldsymbol{\gamma}$ does not have a proper posterior distribution until sufficient data have been collected. Thus, in early stages of the trial, one may follow simple 1:1 randomization or some other (possibly stratified) randomization scheme. If at time t the posterior distribution of $\boldsymbol{\gamma}$ is proper, one can compute the TS treatment assignment by first drawing $\widetilde{\boldsymbol{\gamma}}_t \sim$ Multivariate–t $\left(\widehat{\boldsymbol{\gamma}}_{t-1}, \widehat{\boldsymbol{\Sigma}}_{t-1}, t-1-q\right)$ and then setting $A_t = \arg\max_{a \in \mathcal{A}} \psi(\mathbf{X}_t, a)^\intercal \widetilde{\boldsymbol{\gamma}}_t$.

3.1 *Frequentist TS for contextual bandits*

The fully Bayesian approach to TS for contextual bandits requires significant modeling and can become computationally burdensome if one does not use conjugate priors. As an alternative, one can use the (estimated) sampling distribution in place of the posterior to obtain a frequentist version of TS that requires fewer modeling assumptions and is often much more computationally tractable.

Consider again the linear model $Y_t = \psi(\mathbf{X}_t, A_t)^\intercal \boldsymbol{\gamma} + \epsilon_t$, where $\boldsymbol{\gamma} \in \boldsymbol{\Gamma}$ is an unknown parameter vector and the error ϵ_t satisfies $\mathbb{E}(\epsilon_t | \mathcal{H}_t^{c-}) = 0$ and $\text{Var}(\epsilon_t | \mathcal{H}_t^{c-}) = \sigma_t^2$ where $0 < \sigma_t^2 < C$ for some constant C and all t. The ordinary least squares estimator $\widehat{\boldsymbol{\gamma}}_{t-1}$ of $\boldsymbol{\gamma}$

based on $\mathcal{H}_t^{c-}$ is given in (3.1). Of course, $\widehat{\boldsymbol{\gamma}}_{t-1}$ also solves the normal equations

$$\sum_{v=1}^{t-1} \{Y_v - \psi(\mathbf{X}_v, A_v)^\intercal \boldsymbol{\gamma}\} \psi(\mathbf{X}_v, A_v) = 0.$$

We can approximate the sampling distribution of $\widehat{\boldsymbol{\gamma}}_{t-1}$ using a normal approximation (e.g., Heyde, 1997, see), however, we prefer to use a generalized bootstrap instead as it is trivial to implement and avoids cumbersome derivations (Chatterjee and Bose, 2005). For each t, let $\lambda_{t,1}, \ldots, \lambda_{t,t} \sim_{\text{i.i.d.}} \text{Exp}(1)$ and let $\widehat{\boldsymbol{\gamma}}_{t-1}^{(b)}$ denote the solution to the bootstrap normal equations

$$\sum_{v=1}^{t-1} \lambda_{t,v} \{Y_v - \psi(\mathbf{X}_v, A_v)^\intercal \boldsymbol{\gamma}\} \psi(\mathbf{X}_v, A_v) = 0, \tag{3.2}$$

then $\widehat{\boldsymbol{\gamma}}_{t-1}^{(b)}$ is a draw from the generalized bootstrap estimator of the sampling distribution of $\widehat{\boldsymbol{\gamma}}_{t-1}$. Under bootstrap TS, the action at time t is $A_t = \arg\max_a \psi(\mathbf{X}_t, a)^\intercal \widehat{\boldsymbol{\gamma}}_{t-1}^{(b)}$.

Frequentist TS with the bootstrap is semi-parametric as it only requires specification of the mean structure and regularity conditions needed for bootstrap consistency (Chatterjee and Bose, 2005). In addition, it avoids specification of a prior and potentially expensive computation (e.g., MCMC). If there is historical data or scientific evidence that can be used to construct an informative prior, this information can be incorporated into frequentist TS via data augmentation (DA). We illustrate using a simple version of DA; other more sophisticated approaches are possible. Suppose we wish to use an informative prior ρ on $\boldsymbol{\Gamma}$. We posit a working prior model for the error $\epsilon_1, \epsilon_2, \ldots \sim_{\text{i.i.d.}} F_\epsilon$. Let $M \in \mathbb{Z}_+$ be a positive integer which reflects the number of 'prior samples' we wish to generate. Let $\mathcal{D}_0 = \emptyset$, at time $t = 1$, upon observing $\mathbf{X}_1$ we then draw $\widetilde{\boldsymbol{\gamma}}_1^1 \sim \rho(\boldsymbol{\gamma})$ and set $A_1 = \arg\max_a \psi(\mathbf{X}_1, a)^\intercal \widetilde{\boldsymbol{\gamma}}_1^1$ and subsequently observe Y_1. In addition, draw a second sample as follows. Set $\widetilde{\mathbf{X}}_1 = \mathbf{X}_1$, draw $\widetilde{\boldsymbol{\gamma}}_1^2 \sim \rho(\boldsymbol{\gamma})$, $\widetilde{\epsilon}_1 \sim F_\epsilon$, and $\widetilde{A}_1 \sim \text{Uniform}(\mathcal{A})$, and set $\widetilde{Y}_1 = \psi(\widetilde{\mathbf{X}}_1, \widetilde{A}_1)^\intercal \widetilde{\boldsymbol{\gamma}}_1^2 + \widetilde{\epsilon}_1$. We continue to generate data in this way so that after m steps we have data $\mathcal{D}_m = \left\{(\mathbf{X}_i, A_i, Y_i), (\widetilde{\mathbf{X}}_i, \widetilde{A}_i, \widetilde{Y}_i)\right\}_{i=1}^m$. Once m is sufficiently large so that $\widehat{\boldsymbol{\gamma}}_m$ is well-defined from $\mathcal{D}_m$, at iterations $t = m+1, m+2, \ldots$ we proceed as follows. We observe $\mathbf{X}_t$, compute the bootstrap estimator $\widehat{\boldsymbol{\gamma}}_{t-1}^{(b)}$, set $A_t =$

$\arg\max_a \psi(\mathbf{X}_t, a)^\intercal \widehat{\boldsymbol{\gamma}}_{t-1}^{(b)}$ and observe Y_t. If $t \leqslant M$, we construct a second sample in which we set $\widetilde{\mathbf{X}}_t = \mathbf{X}_t$, draw $\widetilde{\epsilon}_t \sim F_\epsilon$, $\widetilde{\gamma}_t \sim \rho(\boldsymbol{\gamma})$, and $A_t \sim \text{Uniform}(\mathcal{A})$, and set $\widetilde{Y}_t = \psi(\widetilde{\mathbf{X}}_t, A_t)^\intercal \widetilde{\gamma}_t + \widetilde{\epsilon}_t$.

The preceding algorithm generates a set of M artificial samples from the 'prior.' Each sample was generated at an observed context value and, in this way, avoided having to posit a model for the contexts. However, if one were willing to posit a context model, it would have been possible to simply generate these M samples before collecting any data. This idea of simulating artificial data from a prior is general and applies to more general decision problems, e.g., Markov decision processes.

4. Thompson Sampling in more general settings

We have thus far discussed TS for a simple two-arm clinical trial and for a linear contextual bandit. We now illustrate how TS can be applied in multi-stage (possibly non-Markov) decision problems (Tsiatis et al., 2019) and infinite horizon Markov decision processes (MDPs; Puterman, 2014).

4.1 *TS for multi-stage decision problems*

We consider an adaptive sequential multiple assignment randomized trial (SMART, Lavori and Dawson, 2004; Murphy, 2005a) with T treatment stages and a planned enrollment size of n subjects. We assume that subjects arrive in cohorts of size k and that $n = km$ so that there are a total of m cohorts. To simplify notation, we assume that one cohort finishes the trial before the next one begins (for a treatment of the more general setting with overlapping and random cohort sizes, see Manschot et al., 2022).

The observed data after the ℓth cohort completes the trial is

$$\mathcal{D}_\ell = \{(\mathbf{X}_{1,i}, A_{1,i}, Y_{1,i}, \mathbf{X}_{2,i}, A_{2,i}, Y_{2,i}, \ldots, \mathbf{X}_{T,i}, A_{T,i}, Y_{T,i})\}_{i=1}^{\ell k},$$

which comprises ℓk trajectories, one per subject, where $\mathbf{X}_{1,i} \in \mathbb{R}^p$ is baseline information for subject i, $A_{t,i} \in \mathcal{A} = \{1, \ldots, K\}$ is the intervention assigned to subject i at time t,

$\mathbf{X}_{t,i} \in \mathbb{R}^p$ contains interim information collected on subject i during the course of treatment $A_{t-1,i}$ for $t = 2, \ldots, T$, and $Y_{t,i} \in \mathcal{Y} \subseteq \mathbb{R}$ is an immediate (momentary) outcome measured on subject i after treatment $A_{t,i}$. The trajectories need not be independent across subjects as accumulated data on past subjects is used in treatment selection. We omit a subscript i when discussing a generic subject, i.e., $(\mathbf{X}_1, A_1, Y_1, \mathbf{X}_2, A_2, Y_2, \ldots, \mathbf{X}_T, A_T, Y_T)$.

Let $\mathbf{H}_1 = \mathbf{X}_1$ and $\mathbf{H}_t = (\mathbf{H}_{t-1}, A_{t-1}, Y_{t-1}, \mathbf{X}_t)$ for $t \geqslant 2$. Thus, $\mathbf{H}_t$ represents the available information to inform treatment selection for a subject in the trial at time t. In many contexts, the set of allowable treatments depends on a subject's health status (van der Laan and Petersen, 2007), e.g., in the context of schizophrenia, one cannot prescribe a type I antipsychotic to a subject with tardive dyskinesia (Lieberman et al., 2005). We operationalize such constraints as a sequence of functions $\boldsymbol{\zeta} = \{\zeta_t\}_{t=1}^T$ with $\zeta_t : \operatorname{dom} \mathbf{H}_t \to 2^{\mathcal{A}}$ so that $\zeta_t(\mathbf{h}_t) \subseteq \mathcal{A}$ is the set of allowable treatments for a subject with history $\mathbf{H}_t = \mathbf{h}_t$.

A treatment regime, in this context is a sequence of decision rules $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_T)$ such that $\pi_t : \mathcal{H}_t \to \mathcal{A}$ and $\pi_t(\mathbf{h}_t) \in \zeta_t(\mathbf{h}_t)$ for all $\mathbf{h}_t \in \operatorname{dom} \mathbf{H}_t$, $t = 1, 2, \ldots, T$. An optimal treatment regime maximizes the expectation of the cumulative outcome, $\sum_{t=1}^T Y_t$, if applied to select treatments in the target population. The optimal regime is formalized using potential outcomes. As in previous sections, write $\overline{\mathbf{a}}_t = (a_1, \ldots, a_t)$. Let $\mathbf{H}_t^{\overline{\mathbf{a}}_{t-1}}$ be the potential history at time t under treatment sequence $\mathbf{a}_{t-1}$ and $Y_t^{\overline{\mathbf{a}}_t}$ the potential outcome at time t under treatment sequence $\overline{\mathbf{a}}_t$; for convenience, we define $\mathbf{H}_1^{\overline{\mathbf{a}}_0} \equiv \mathbf{H}_1$. The potential outcome at time t under sequence of decision rules $\overline{\boldsymbol{\pi}}_t = (\pi_1, \ldots, \pi_t)$ is

$$Y_t^{\overline{\boldsymbol{\pi}}_t} = \sum_{\overline{a}_t} Y_t^{\overline{\mathbf{a}}_t} \prod_{v=1}^{t} 1\left\{\pi_v\left(\mathbf{H}_v^{\overline{\mathbf{a}}_{v-1}}\right) = a_v\right\}.$$

The value of a regime $\boldsymbol{\pi}$ is $V(\boldsymbol{\pi}) = \mathbb{E}\left(\sum_{v=1}^T Y_v^{\overline{\boldsymbol{\pi}}_v}\right)$, and the optimal regime, $\boldsymbol{\pi}^{\text{opt}}$, satisfies $V(\boldsymbol{\pi}^{\text{opt}}) \geqslant V(\boldsymbol{\pi})$ for all feasible regimes $\boldsymbol{\pi}$.

To identify $\boldsymbol{\pi}^{\text{opt}}$ from the data-generating model, we make the following standard assump-

tions. Define

$$\mathcal{W} = \left\{ \left(\mathbf{H}_t^{\overline{\mathbf{a}}_{t-1}}, Y_t^{\overline{\mathbf{a}}_t} \right) : \overline{\mathbf{a}}_t \in \mathcal{A}^t, a_v \in \zeta_v(\mathbf{H}_v^{\overline{\mathbf{a}}_{v-1}}) \, \forall \, 1 \leqslant v \leqslant t \right\}_{t=1}^{T}$$

to be the set of realizable (feasible) potential outcomes. We assume the following conditions hold: (C1) strong ignorability, $\mathcal{W} \perp A_t | \mathbf{H}_t$ for all $t = 1, \ldots, T$; (C2) positivity, there exists $\epsilon > 0$ such that $P(A_t = a | \mathbf{H}^t = \mathbf{h}_t) \geqslant \epsilon$ for all $\mathbf{h}_t \in \operatorname{dom} \mathbf{H}_t$, $a \in \zeta_t(\mathbf{h}_t)$, and $t = 1, \ldots, T$; and (C3) consistency, $\mathbf{H}_t = \mathbf{H}_t^{\overline{\mathbf{A}}_{t-1}}$ and $Y_t = Y_t^{\overline{\mathbf{A}}_t}$ for all $t = 1, \ldots, T$, i.e., the observed history and outcomes are the potential history and outcomes under treatment actually assigned. We also assume that there are not multiple versions of treatment or interference among subjects (Tsiatis et al., 2019). In the context of a SMART, (C1) and (C2) can be guaranteed by design. In practice, the (approximate) validity of the other conditions must be argued on the basis of the underlying science and implementation of the trial (see Tsiatis et al., 2019; Hernan and Robins, 2020, and references therein). Hereafter, we implicitly assume these conditions hold.

We characterize the optimal regime, $\boldsymbol{\pi}^{\text{opt}}$, using dynamic programming (Bellman, 1952). Define the Q-function at stage-T as

$$Q_T(\mathbf{h}_T, a_T) = \mathbb{E}\left(Y_T \middle| \mathbf{H}_T = \mathbf{h}_T, A_T = a_T \right),$$

and recursively for $t = T-1, T-2, \ldots, 1$ define

$$Q_t(\mathbf{h}_t, a_t) = \mathbb{E}\left\{ Y_t + \max_{a_{t+1} \in \zeta_{t+1}(\mathbf{H}_{t+1})} Q_{t+1}(\mathbf{H}_{t+1}, a_{t+1}) \middle| \mathbf{H}_t = \mathbf{h}_t, A_t = a_t \right\}.$$

It follows that an optimal regime is given by $\pi_t^{\text{opt}}(\mathbf{h}_t) = \arg\max_{a_t \in \zeta_t(\mathbf{h}_t)} Q_t(\mathbf{h}_t, a_t)$ (see Murphy, 2005b; Schulte et al., 2014). Thus, given data after ℓ cohorts, one can construct an estimator of $\boldsymbol{\pi}^{\text{opt}}$ by constructing estimators $\widehat{Q}_{t,\ell}$ of Q_t for $t = 1, \ldots, T$, and subsequently $\widehat{\pi}_{t,\ell}(\mathbf{h}_t) = \arg\max_{a_t \in \zeta_t(\mathbf{h}_t)} \widehat{Q}_{t,\ell}(\mathbf{h}_t, a_t)$. We illustrate this approach using parametric models for the Q-functions, though more flexible non-parametric models are possible (Ernst et al., 2005).

For each t we posit a working model $Q_t(\mathbf{h}_t, a_t; \boldsymbol{\theta}_t)$ indexed by $\boldsymbol{\theta}_t \in \boldsymbol{\Theta}_t \subseteq \mathbb{R}^{q_t}$. We assume that

$Q_t(\mathbf{h}_t, a_t; \boldsymbol{\theta}_t)$ is defined and continuously differentiable for all $\boldsymbol{\theta}_t$ in an open set that contains $\boldsymbol{\Theta}_t$. Define $\widehat{\boldsymbol{\theta}}_{T,\ell}$ as the solution to the so-called conditional least-squares score equation

$$\sum_{i=1}^{\ell k} \{Y_{T,i} - Q_T(\mathbf{H}_{T,i}, A_{T,i}; \boldsymbol{\theta}_T)\} \nabla_{\boldsymbol{\theta}_T} Q_T(\mathbf{H}_{T,i}, A_{T,i}; \boldsymbol{\theta}_T) = 0, \tag{4.1}$$

and similarly, for $t = T-1, T-2, \ldots, 1$ define $\widehat{\boldsymbol{\theta}}_{t,\ell}$ as the solution to

$$\sum_{i=1}^{\ell k} \left\{ Y_{t,i} + \max_{a_{t+1}} Q_{t+1}(\mathbf{H}_{t+1,i}, a_{t+1}; \widehat{\boldsymbol{\theta}}_{t+1,\ell}) - Q_t(\mathbf{H}_{t,i}, A_{t,i}; \boldsymbol{\theta}_t) \right\} \nabla_{\boldsymbol{\theta}_t} Q_t(\mathbf{H}_{t,i}, A_{t,i}; \boldsymbol{\theta}_t) = 0. \tag{4.2}$$

The estimated optimal regime using data from the first ℓ cohorts is thus $\widehat{\pi}_{t,\ell}(\mathbf{h}_t) = \arg\max_{a_t \in \zeta_t(\mathbf{h}_t)} Q_t(\mathbf{h}_t, a_t; \widehat{\boldsymbol{\theta}}_{t,\ell})$.

We use the preceding estimating equations with the multiplier bootstrap to implement a frequentist version of TS. However, in early cohorts, the estimating equations will not have unique solutions so one needs a base strategy to begin. One natural approach is to use uniform randomization at each stage for fixed number of cohorts, say L, so that for any subject i in cohort $\ell \leqslant L$ with history $\mathbf{H}_{t,i}$ will be assigned treatment $A_{t,i} \sim \text{Uniform}\{\zeta_t(\mathbf{H}_{t,i})\}$. For a subject i in cohort $\ell = L+1, \ldots, m$ draw $\lambda_{i,1}, \ldots, \lambda_{i,(\ell-1)k} \sim_{i.i.d.} \text{Exp}(1)$, compute $\widehat{\boldsymbol{\theta}}^{(b)}_{T,\ell-1,i}$ as the solution to

$$\sum_{j=1}^{(\ell-1)k} \lambda_{i,j} \{Y_{T,j} - Q_T(\mathbf{H}_{T,j}, A_{T,j}; \boldsymbol{\theta}_T)\} \nabla_{\boldsymbol{\theta}_T} Q_T(\mathbf{H}_{T,j}, A_{T,j}; \boldsymbol{\theta}_T) = 0,$$

and, recursively, compute $\widehat{\boldsymbol{\theta}}^{(b)}_{t,\ell-1,i}$ for $t = T-1, \ldots, 1$ as the solution to

$$\sum_{j=1}^{(\ell-1)k} \lambda_{i,j} \left\{ Y_{t,j} + \max_{a_{t+1}} Q_{t+1}(\mathbf{H}_{t+1,j}, a_{t+1}; \widehat{\boldsymbol{\theta}}^{(b)}_{t+1,\ell-1,i}) - Q_t(\mathbf{H}_{t,j}, A_{t,j}; \boldsymbol{\theta}_t) \right\} \nabla_{\boldsymbol{\theta}_t} Q_t(\mathbf{H}_{t,j}, A_{t,j}; \boldsymbol{\theta}_t) = 0.$$

Thus, a subject i in cohort $\ell = L+1, \ldots, m$ with history $\mathbf{H}_{t,i}$ at time t, is assigned treatment $A_{t,i} = \arg\max_{a_t \in \zeta_t(\mathbf{H}_{t,i})} Q_t(\mathbf{H}_{t,i}, a_t; \widehat{\boldsymbol{\theta}}^{(b)}_{t,\ell-1,i})$.

The version of TS for multi-stage decision problems we describe uses the estimated sampling distribution of parameters indexing the Q-functions in place of a proper posterior distribution. To see this, let $\widehat{P}_{t,\ell}$ denote the estimated sampling distribution of $\widehat{\boldsymbol{\theta}}_{t,\ell-1}$ based on the multiplier bootstrap applied to data from the first $\ell - 1$ cohorts. For a subject with history $\mathbf{H}_t = \mathbf{h}_t$ in cohort ℓ and any $a \in \zeta_t(\mathbf{h}_t)$ the event $A_t = a$ is equivalent to the event

$a = \arg\max_{a_t \in \zeta_t(\mathbf{h}_t)} Q_t(\mathbf{h}_t, a_t; \widehat{\boldsymbol{\theta}}^{(b)}_{t,\ell-1})$ which occurs with probability

$$\int 1\left\{a = \arg\max_{a_t \in \zeta_t(\mathbf{h}_t)} Q_t(\mathbf{h}_t, a_t; \boldsymbol{\theta})\right\} d\widehat{P}_{t,\ell-1}(\boldsymbol{\theta}).$$

In a fully-Bayesian approach, $\widehat{P}_{t,\ell-1}$ would be replaced by the posterior distribution over $\boldsymbol{\theta}_t$ given data on the first $\ell - 1$ cohorts.

In some settings, one may wish to incorporate prior information into frequentist TS for multi-stage decision problems. One way to do this is to posit a prior for the joint trajectory distribution, $(\mathbf{X}_1, A_1, Y_1, \mathbf{X}_2, A_2, Y_2, \ldots, \mathbf{X}_T, A_T, Y_T)$ and then to simulate data from this prior to augment the observed data; i.e., one can simulate J i.i.d. trajectories under the prior as a kind of zeroth cohort. This can be an effective strategy for folding in domain knowledge or historical data when it is available at the trajectory level.

Remark 1: We did not use partial information of subjects within a cohort; e.g., if the enrollment and/or completion times of the stages vary across subjects within a cohort, it is possible to use data from subjects who have advanced further in the trial to inform the treatment decisions for subjects at earlier stages. This can increase efficiency though at the expense of more complex bookkeeping. See Norwood et al. (2022) for details.

4.2 *TS for MDPs*

In decision problems with a long or indefinite time horizon one needs to impose more structure on the data-generating model than in the preceding section to facilitate extrapolation in time. Most commonly, one assumes that the data are from a stationary and homogeneous MDP (Sutton, 1997). We present a frequentist version of TS in this setting. However, we first discuss the construction of a homogeneous and stationary MDP from raw (possibly non-Markov) data. This is a critically important issue in application but has received little attention in the precision medicine literature (see Wang et al., 2017; Ma et al., 2023, for references).

4.2.1 *Pre-processing and the Markov assumption.* As in the preceding section, we consider longitudinal data on n subjects in a sequential randomized trial.[1] However, we now assume that subjects enroll in a single cohort and that the treatment decisions are aligned in time for all subjects in the cohort. At any time t, the raw observed data are of the form

$$\{(\mathbf{X}_{1,i}, A_{1,i}, Y_{1,i}, \mathbf{X}_{2,i}, A_{2,i}, Y_{2,i}, \ldots, \mathbf{X}_{t,i}, A_{t,i}, Y_{t,i})\}_{i=1}^{n}, \tag{4.3}$$

which comprises n trajectories of the form $(\mathbf{X}_1, A_1, Y_1, \mathbf{X}_2, A_2, Y_2, \ldots, \mathbf{X}_t, A_t, Y_t)$, where $\mathbf{X}_1 \in \mathbb{R}^p$ are baseline measurements, $\mathbf{A}_t \in \mathcal{A} = \{1, 2, \ldots, K\}$ is the assigned intervention at time t, $\mathbf{X}_t \in \mathcal{X}$ are interim measurements taken during the course of A_t, and $Y_t \in \mathcal{Y} \subseteq [0, 1]$ are outcomes coded so that higher values are better. Let the history $\mathbf{H}_t$ be defined as in the preceding section and let $\boldsymbol{\Pi}$ denote the class of feasible regimes. We write $Y_t^{\boldsymbol{\pi}}$ to denote the potential outcome at time t under $\boldsymbol{\pi} \in \boldsymbol{\Pi}$.

For any $\boldsymbol{\pi} \in \boldsymbol{\Pi}$ an $t \geqslant 1$ write $\underline{\boldsymbol{\pi}}_t = (\pi_t, \pi_{t+1}, \ldots)$. Given history $\mathbf{H}_t = \mathbf{h}_t$ and $\boldsymbol{\pi} \in \boldsymbol{\Pi}$, define the state-value function at time t as

$$V_t(\boldsymbol{\pi}, \mathbf{h}_t) = V_t(\underline{\boldsymbol{\pi}}_t, \mathbf{h}_t) = \mathbb{E}\left(\sum_{v \geqslant 0} \gamma^v Y_{t+v}^{\boldsymbol{\pi}} \big| \mathbf{H}_t = \mathbf{h}_t\right),$$

where $\gamma \in (0, 1)$ is a discount factor. The optimal feasible regime, $\boldsymbol{\pi}^{\text{opt}} \in \boldsymbol{\Pi}$, satisfies $V_t(\boldsymbol{\pi}^{\text{opt}}, \mathbf{h}_t) \geqslant V_t(\boldsymbol{\pi}, \mathbf{h}_t)$ for all $\boldsymbol{\pi} \in \boldsymbol{\Pi}$ and $\mathbf{h}_t \in \operatorname{dom} \mathbf{H}_t$. It is clear that without additional structure, one cannot recover $\boldsymbol{\pi}^{\text{opt}}$ from n trajectories of length t as in (4.3) even as $n \to \infty$ (as one will have no information about $\underline{\boldsymbol{\pi}}_{t+1}^{\text{opt}}$). The most common approach to estimating $\boldsymbol{\pi}^{\text{opt}}$ in practice is to assume that, after some suitable transformation, the observed data can be represented as a homogeneous MDP; we now describe how such a transformation might be constructed.

Assume that there exists a sequence of summary functions $\{\psi_t\}_{t \geqslant 1}$ with $\psi_t : \operatorname{dom} \mathbf{H}_t \to \mathcal{S} \subseteq \mathbb{R}^q$ and we call $\mathbf{S}_t = \psi_t(\mathbf{H}_t)$ the state of the system at time t. For example, the state

[1] We omit a discussion of the necessary causal assumptions in this section delaying a formal statement of these assumptions to the the next section.

might be constructed by concatenating interim measurements, treatments, and outcomes over fixed look-back period (Ma et al., 2023), taking a weighted average over past measurements (Laber and Staicu, 2018), or using data-driven feature selection, e.g., using recurrent neural networks (Wang et al., 2018). We assume that the summary function induces a homogeneous MDP so that

$$\mathbf{S}^{t+1} \perp (\mathbf{H}^{t-1}, A_{t-1}, Y_{t-1}) \big| (\mathbf{S}^t, A^t),$$

and the conditional distribution of $\mathbf{S}_{t+1}$ given $(\mathbf{S}_t, A_t)$ does not depend on time t. We also assume that the summary is such that $Y_t = u(\mathbf{S}_t, A_t, \mathbf{S}_{t+1})$ for some fixed and known function $u : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathcal{Y}$, and that there exists function $\upsilon : \mathcal{S} \to 2^{\mathcal{A}}$ such that $\upsilon\{\psi_t(\mathbf{h}_t)\} = \zeta_t(\mathbf{h}_t)$ for all $\mathbf{h}_t \in \operatorname{dom} \mathbf{H}_t$ and all t. Let $\mathbf{\Pi}_{\mathcal{M}}$ denote the set of maps, $\varpi : \mathcal{S} \to \mathcal{A}$, such that $\varpi(\mathbf{s}) \in \upsilon(\mathbf{s})$ for all $\mathbf{s} \in \mathcal{S}$. Let Y_t^{ϖ} denote the potential outcome under $\varpi \in \mathbf{\Pi}_{\mathcal{M}}$. For each t, $\mathbf{h}_t \in \operatorname{dom} \mathbf{H}_t$, and $a_t \in \zeta_t(\mathbf{h}_t)$ define

$$Q_t^{\text{opt}}(\mathbf{h}_t, a_t) = \sup_{\boldsymbol{\pi} \in \mathbf{\Pi}} \mathbb{E}\left\{ \sum_{v=0}^{\infty} \gamma^v Y_{t+v}^{\boldsymbol{\pi}} \big| \mathbf{H}_t = \mathbf{h}_t, A_t = a_t \right\},$$

then it follows (e.g., see Puterman, 2014; Bertsekas, 2012) that

$$Q_t^{\text{opt}}(\mathbf{h}_t, a_t) = \sup_{\boldsymbol{\pi} \in \mathbf{\Pi}} \mathbb{E}\left\{ Y_t^{\boldsymbol{\pi}} + \gamma \max_{a_{t+1} \in \zeta_t(\mathbf{H}_{t+1})} Q_{t+1}^{\text{opt}}(\mathbf{H}_{t+1}, a_{t+1}) \big| \mathbf{H}_t = \mathbf{h}_t, A_t = a_t \right\},$$

and an optimal decision strategy based on the raw data, say $\boldsymbol{\pi}^{\text{opt}}$, is given by $\pi_t^{\text{opt}}(\mathbf{h}_t) = \arg\max_{\mathbf{a}_t \in \zeta_t(\mathbf{h}_t)} Q_t^{\text{opt}}(\mathbf{h}_t, a_t)$. If $\left\{ (Y_{t+1}, \max_{a_{t+1} \in \zeta_{t+1}(\mathbf{H}_{t+1})} Q_{t+1}^{\text{opt}}(\mathbf{H}_{t+1}, a_{t+1}) \right\} \perp \mathbf{H}_t | (\mathbf{S}_t, A_t)$, then it follows that $Q_t^{\text{opt}}(\mathbf{h}_t, a_t)$ depends on $\mathbf{h}_t$ only through $\mathbf{s}_t = \psi_t(\mathbf{h}_t)$ and therefore $\pi_t^{\text{opt}}(\mathbf{h}_t)$ depends on $\mathbf{h}_t$ only through $\psi_t(\mathbf{h}_t)$ (Wang et al., 2017). Furthermore, the optimal value

starting from $\mathbf{H}_t = \mathbf{h}_t$ satisfies

$$\begin{aligned} V_t^{\text{opt}}(\mathbf{h}_t) &= Q_t^{\text{opt}}\left\{\mathbf{h}_t, \pi_t^{\text{opt}}(\mathbf{h}_t)\right\} \\ &= \mathbb{E}\left\{\sum_{v=0}^{\infty} \gamma^v Y_{t+v}^{\boldsymbol{\pi}^{\text{opt}}} \big| \mathbf{H}_t = \mathbf{h}_t\right\} \\ &= \mathbb{E}\left\{\sum_{v=0}^{\infty} \gamma^v Y_{t+v}^{\boldsymbol{\pi}^{\text{opt}}} \big| \mathbf{S}^t = \psi_t(\mathbf{h}_t)\right\} \\ &= \sup_{\boldsymbol{\varpi} \in \boldsymbol{\Pi}_{\mathcal{M}}^{\infty}} \mathbb{E}\left\{\sum_{v=0}^{\infty} \gamma^v Y_{t+v}^{\boldsymbol{\varpi}} \big| \mathbf{S}^t = \psi_t(\mathbf{h}_t)\right\} \\ &= \sup_{\varpi \in \boldsymbol{\Pi}_{\mathcal{M}}} \mathbb{E}\left\{\sum_{v=0}^{\infty} \gamma^v Y_{t+v}^{\varpi} \big| \mathbf{S}^t = \psi_t(\mathbf{h}_t)\right\} \end{aligned} \tag{4.4}$$

where $\boldsymbol{\Pi}_{\mathcal{M}}^{\infty}$ is the space of sequences in $\boldsymbol{\Pi}_{\mathcal{M}}$ and the last equality follows from the fact that the best treatment in a state $\mathbf{S}^t = \mathbf{s}$ does not depend on t (see Puterman, 2014, for additional details). Let ϖ^{opt} attain the sup in (4.4). It follows that the reduced process

$$\{(\mathbf{S}_{1,i}, A_{1,i}, Y_{1,i}, \mathbf{S}_{2,i}, A_{2,i}, Y_{2,i}, \ldots, \mathbf{S}_{t,i}, A_{t,i}, Y_{t,i})\}_{i=1}^{n} \tag{4.5}$$

comprises trajectories from a homogeneous MDP and that $\pi_t^{\text{opt}}(\mathbf{h}_t) = \varpi^{\text{opt}}\{\psi_t(\mathbf{h}_t)\}$ is optimal; i.e., $V_t(\boldsymbol{\pi}^{\text{opt}}, \mathbf{h}_t) \geqslant V_t(\boldsymbol{\pi}, \mathbf{h}_t)$ for all $\boldsymbol{\pi} \in \boldsymbol{\Pi}$ and $\mathbf{h}_t \in \operatorname{dom} \mathbf{H}_t$. Furthermore, $\boldsymbol{\pi}^{\text{opt}}$ can be estimated using only the data from the reduced process (4.5) as we describe in the next section. Constructing a suitable reduced process that is parsimonious, homogeneous, Markov, and has the same optimal regime as the original process is not trivial. While data-driven methods for constructing the maps ψ_t exist (Wang et al., 2017; Ma et al., 2023), this is more often done using *ad hoc* transformations and justified using clinical theory.

4.2.2 *Q-learning in MDPs.* We assume that the observed data (possibly after transformation) are of the form

$$\{(\mathbf{S}_{1,i}, A_{1,i}, \mathbf{S}_{2,i,}, A_{2,i}, \ldots, \mathbf{S}_{t,i}, A_{t,i}, \mathbf{S}_{t+1,i})\}_{i=1}^{n},$$

which comprise n trajectories, one for each subject, of the form $(\mathbf{S}_1, A_1, \mathbf{S}_2, A_2, \ldots, \mathbf{S}_t, A_t, \mathbf{S}_{t+1})$, where: $\mathbf{S}_t \in \mathcal{S} \subseteq \mathbb{R}^q$ is a summary of the subject's health status at time t and $A_t \in \mathcal{A} =$

$\{1, 2, \ldots, K\}$ is the treatment assigned at time t. In the context of MDPs, the term action is often used in place of treatment; we shall use the terms interchangeably. We assume that there exists a fixed function $u : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$, so that the outcome $Y_t = u(\mathbf{S}_t, A_t, \mathbf{S}_{t+1})$ captures the utility associated with the state-treatment-next state triple $(\mathbf{S}_t, A_t, \mathbf{S}_{t+1})$.

We assume that the data-generating model is a homogeneous MDP so that for any measurable set $\mathcal{B} \subseteq \mathcal{S}$ and time t

$$P\left(\mathbf{S}_{t+1} \in \mathcal{B} \middle| \mathbf{S}_1, \ldots, \mathbf{S}_t, A_1, \ldots, A_t\right) = P\left(\mathbf{S}_{t+1} \middle| \mathbf{S}_t, A_t\right)$$

with probability one, and the probability does not depend on t.

We assume that there exists a set-valued function $\nu : \mathcal{S} \to 2^{\mathcal{A}}$ so that $\nu(\mathbf{s}) \subseteq \mathcal{A}$ is the set of allowable treatments for a subject in state $\mathbf{s}$; we assume $\nu(\mathbf{s})$ is non-empty for all $\mathbf{s} \in \mathcal{S}$. A treatment regime in this context is a map $\pi : \mathcal{S} \to \mathcal{A}$ that satisfies $\pi(\mathbf{s}) \in \nu(\mathbf{s})$ for all $\mathbf{s} \in \mathcal{A}$. Let $\mathbf{\Pi}$ denote the set of all treatment regimes. Under a regime $\pi \in \mathbf{\Pi}$, a subject with $\mathbf{S}_t = \mathbf{s}$ at time t will be recommended treatment $\pi(\mathbf{s})$. An optimal treatment regime maximizes expected discounted cumulative utility if used to select treatments for patients in the target population. As in previous sections, we formalize this definition using potential outcomes. Let $\mathbf{S}_t^{\overline{\mathbf{a}}_{t-1}}$ denote the potential state under treatment sequence $\overline{\mathbf{a}}_{t-1} = (a_1, \ldots, a_{t-1})$; for convenience, we follow the notational convention that $\mathbf{S}_1^{\overline{\mathbf{a}}_0} \equiv \mathbf{S}_1$. The potential outcome under treatment sequence $\overline{\mathbf{a}}_t$ is thus

$$Y_t^{\mathbf{a}_t} = u\left(\mathbf{S}_t^{\overline{\mathbf{a}}_{t-1}}, a_t, \mathbf{S}_{t+1}^{\overline{\mathbf{a}}_t}\right),$$

and the potential outcome at time t under a regime π is

$$Y_t^{\pi} = \sum_{\overline{\mathbf{a}}_t} Y_t^{\overline{\mathbf{a}}_t} \prod_{v=1}^{t-1} 1\left\{\pi(\mathbf{S}_v^{\overline{\mathbf{a}}_{v-1}}) = a_v\right\}.$$

For any $\mathbf{s} \in \mathcal{S}$ and regime π define the state-value function

$$V(\pi, \mathbf{s}) = \mathbb{E}\left(\sum_{v \geqslant 0} \gamma^v Y_{t+v}^{\pi} \middle| \mathbf{S}_t = \mathbf{s}\right),$$

where $\gamma \in (0, 1)$ is a discount factor. The optimal regime, π^{opt}, satisfies $V(\pi^{\text{opt}}, \mathbf{s}) \geqslant V(\pi, \mathbf{s})$ for all $\mathbf{s} \in \mathcal{S}$ and $\pi \in \mathbf{\Pi}$. To identify π^{opt} in terms of the data-generating model we make use

of the following causal assumptions which mirror those made in previous sections. Let

$$\mathcal{W} = \left\{ \left(\mathbf{S}_t^{\overline{\mathbf{a}}_{t-1}}, Y_t^{\overline{\mathbf{a}}_t} \right) \,:\, \overline{\mathbf{a}}_t \in \mathcal{A}^t,\, a_v \in \nu(\mathbf{S}_v^{\overline{\mathbf{a}}_{v-1}}) \,\forall\, 1 \leqslant v \leqslant t \right\}_{t \geqslant 1};$$

we assume: (C1) strong ignorability, $\mathcal{W} \perp A_t | (\overline{\mathbf{S}}_t, \overline{\mathbf{A}}_{t-1})$, for all $t \geqslant 1$; (C2) positivity, there exists $\epsilon > 0$ such that $P(A_t = a | \overline{\mathbf{S}}_t, \overline{\mathbf{A}}_{t-1}) \geqslant \epsilon$ for all $a \in \nu(\mathbf{S}_t)$ with probability one; and (C3) consistency, $\mathbf{S}_t = \mathbf{S}_t^{\overline{\mathbf{A}}_{t-1}}$ for all t. In addition, we assume that there is no interference nor are there multiple versions of treatment. We note that because $Y_t = u(\mathbf{S}_t, A_t, \mathbf{S}_{t+1})$ it follows from (C3) that $Y_t = Y_t^{\overline{\mathbf{a}}_t}$. For any $\mathbf{s} \in \mathcal{S}$ and $a \in \nu(\mathbf{s})$, define the optimal Q-function as

$$Q(\mathbf{s}, a) = \sup_{\pi \in \mathbf{\Pi}} \mathbb{E} \left(\sum_{v \geqslant 0} \gamma^v Y_{t+v}^{\pi} \big| \mathbf{S}_t = \mathbf{s}, A_t = a \right),$$

then it follows (see Ertefaie and Strawderman, 2018) under (C1)-(C3) that

$$Q(\mathbf{s}, a) = \mathbb{E} \left\{ Y_t + \gamma \max_{a_{t+1} \in \nu(\mathbf{S}_{t+1})} Q(\mathbf{S}_{t+1}, a_{t+1}) \big| \mathbf{S}_t = \mathbf{s}, A_t = a \right\}, \tag{4.6}$$

where, critically, the expectation is taken with respect to the data-generating model rather than a counterfactual distribution. Let $\psi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ be an arbitrary function of state. It follows that

$$\begin{aligned} Q(\mathbf{S}_t, A_t) &= \mathbb{E} \left\{ Y_t + \gamma \max\nolimits_{a_{t+1} \in \nu(\mathbf{S}_{t+1})} Q(\mathbf{S}_{t+1}, a_{t+1}) \Big| \mathbf{S}_t, A_t \right\} \\ \implies \quad 0 &= \mathbb{E} \left\{ Y_t + \gamma \max\nolimits_{a_{t+1} \in \nu(\mathbf{S}_{t+1})} Q(\mathbf{S}_{t+1}, a_{t+1}) - Q(\mathbf{S}_t, A_t) \Big| \mathbf{S}_t, A_t \right\} \\ \implies \quad 0 &= \mathbb{E} \left[\left\{ Y_t + \gamma \max\nolimits_{a_{t+1} \in \nu(\mathbf{S}_{t+1})} Q(\mathbf{S}_{t+1}, a_{t+1}) - Q(\mathbf{S}_t, A_t) \right\} \psi(\mathbf{S}_t, A_t) \right], \end{aligned}$$

where the last equality follows from multiplying the second equality through by $\psi(\mathbf{S}_t, A_t)$ and taking an expectation. Q-learning uses this last equality to construct an estimating function for the Q-function. We illustrate this idea using a linear model for the Q-function of the form $Q(\mathbf{s}, a) = \phi(\mathbf{s}, a)^\intercal \boldsymbol{\theta}$ where $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ is a feature vector and $\boldsymbol{\theta} \in \mathbf{\Theta} \subseteq \mathbb{R}^d$ is a vector of unknown coefficients. We take $\psi(\mathbf{s}, a) = \nabla_{\boldsymbol{\theta}} Q(\mathbf{s}, a; \boldsymbol{\theta}) = \phi(\mathbf{s}, a)$, and construct $\widehat{\boldsymbol{\theta}}_{t,n}$ as the solution to

$$0 = \sum_{i=1}^{n} \sum_{v=1}^{t} \left\{ Y_{t,i} + \gamma \max_{a_{t+1} \in \nu(\mathbf{S}_{t+1,i})} \phi(\mathbf{S}_{t+1,i}, a_{t+1})^\intercal \boldsymbol{\theta} - \phi(\mathbf{S}_{t,i}, A_{t,i})^\intercal \boldsymbol{\theta} \right\} \phi(\mathbf{S}_{t,i}, A_{t,i}), \tag{4.7}$$

so that the estimated optimal regime is $\widehat{\pi}_{t,n}(\mathbf{s}) = \arg\max_{a\in\nu(\mathbf{s})} Q(\mathbf{s}, a; \widehat{\boldsymbol{\theta}}_{t,n})$.

To implement TS in this context we again use the estimated sampling distribution of $\widehat{\boldsymbol{\theta}}_{t,n}$, (based on an asymtptotic approximation in which n grows large). To select a treatment at time $t+1$, we draw $\lambda_{1,n}, \ldots, \lambda_{n,n} \sim_{i.i.d.} \text{Exp}(1)$, compute $\widetilde{\boldsymbol{\theta}}_{t,n}$ as the solution to

$$0 = \sum_{i=1}^{n} \lambda_{i,n} \sum_{v=1}^{t} \left\{ Y_{t,i} + \gamma \max_{a_{t+1}\in\nu(\mathbf{S}_{t+1,i})} \phi(\mathbf{S}_{t+1,i}, a_{t+1})^{\intercal}\boldsymbol{\theta} - \phi(\mathbf{S}_{t,i}, A_{t,i})^{\intercal}\boldsymbol{\theta} \right\} \phi(\mathbf{S}_{t,i}, A_{t,i}),$$

and assign treatments at time point $t+1$ according to the regime $\widetilde{\pi}_t(\mathbf{s}) = \arg\max_{a\in\nu(\mathbf{s})} Q(\mathbf{s}, a; \widetilde{\boldsymbol{\theta}}_{t,n})$; i.e., $A_{t+1,i} = \arg\max_{a\in\nu(\mathbf{S}_{t+1,i})} \psi(\mathbf{S}_{t+1,i}, A_{t+1,i})^{\intercal}\widetilde{\boldsymbol{\theta}}_{t,n}$, for $i = 1, \ldots, n$.

The preceding version of TS uses a single bootstrap resample at each time point. An alternative is to compute a separate resample for each subject, i.e., for subject j, we draw $\lambda_{1,n,j}, \ldots, \lambda_{n,n,j} \sim_{i.i.d.} \text{Exp}(1)$, compute $\widetilde{\boldsymbol{\theta}}_{t,n,j}$ as the solution to

$$0 = \sum_{i=1}^{n} \lambda_{i,n,j} \sum_{v=1}^{t} \left\{ Y_{t,i} + \gamma \max_{a_{t+1}\in\nu(\mathbf{S}_{t+1,i})} \phi(\mathbf{S}_{t+1,i}, a_{t+1})^{\intercal}\boldsymbol{\theta} - \phi(\mathbf{S}_{t,i}, A_{t,i})^{\intercal}\boldsymbol{\theta} \right\} \phi(\mathbf{S}_{t,i}, A_{t,i}),$$

and set $A_{t+1,j} = \arg\max_{a\in\nu(\mathbf{S}_{t+1,i})} \phi(\mathbf{S}_{t+1,j}, a)^{\intercal}\widetilde{\boldsymbol{\theta}}_{t,n,j}$. This approach, while computationally more expensive, often provides better balance in terms of treatment allocation across subject states.

4.3 *Inference for TS in MDPs*

Statistical inference under adaptive sampling, i.e., when accumulated are used to select interventions, is markedly more complex than non-adaptive sampling (Lai and Wei, 1982; Zhan et al., 2021; Zhang et al., 2020, 2022). Intuitively, a key challenge is ensuring sufficient information generation across the entire state-action (state-treatment) space $\mathcal{S} \times \mathcal{A}$. An adaptive algorithm attempting to maximize cumulative reward may quickly become concentrated around an optimal regime so that little data is available for estimation and inference about the performance of other regimes of interest (say business-as-usual, or a less intensive regime, etc.). In this section, we introduce some basic technical tools that are often useful for

analyzing TS in MDPs (as well as in other settings such as bandits or partially observable MDPs).

We treat the number of subjects, n, as fixed and consider asymptotic approximations as the number of time points, t, grows large. As in the preceding section, to simplify notation, we assume that subjects are aligned in time. Let $\mathcal{F}_t$ denote the σ-algebra generated by $\{(\mathbf{S}_{1,i}, A_{1,i}, \mathbf{S}_{2,i},, A_{2,i}, \ldots, \mathbf{S}_{t-1,i}, A_{t-1,i}, \mathbf{S}_{t,i})\}_{i=1}^n$, and for any $\boldsymbol{\theta}$ define

$$u_t(\theta) = \sum_{i=1}^{n} \left\{ Y_{t,i} + \gamma \max_{a_{t+1}} \phi(\mathbf{S}_{t+1,i}, a_{t+1})^\intercal \boldsymbol{\theta} - \phi(\mathbf{S}_{t,i}, A_{t,i})^\intercal \boldsymbol{\theta} \right\} \phi(\mathbf{S}_{t,i}, A_{t,i}).$$

The Q-learning estimating equations (4.7) can thus be written as $\mathcal{U}_t(\boldsymbol{\theta}) = 0$, where

$$\mathcal{U}_t(\boldsymbol{\theta}) = \sum_{v=1}^{t} u_v(\boldsymbol{\theta}).$$

Suppose that the model is correctly specified so that $Q(\mathbf{s}, a) = \phi(\mathbf{s}, a)^\intercal \boldsymbol{\theta}^*$ for some $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$ and all $(\mathbf{s}, a) \in \mathcal{S} \times \mathcal{A}$. Then it follows that $\mathcal{U}_t(\boldsymbol{\theta}^*)$ is a Martingale with respect to the filtration $\{\mathcal{F}_t\}_{t \geqslant 1}$ as

$$\begin{aligned}
\mathbb{E}\left\{\mathcal{U}_t(\boldsymbol{\theta}^*) \middle| \mathcal{F}_t\right\} &= \mathbb{E}\left\{u_t(\boldsymbol{\theta}^*) \middle| \mathcal{F}_t\right\} + \mathcal{U}_{t-1}(\boldsymbol{\theta}^*) \\
&= \mathbb{E}\left[\sum_{i=1}^{n} \left\{ Y_{t,i} + \gamma \max_{a_{t+1}} \phi(\mathbf{S}_{t+1,i}, a_{t+1})^\intercal \boldsymbol{\theta}^* - \phi(\mathbf{S}_{t,i}, A_{t,i})^\intercal \boldsymbol{\theta}^* \right\} \phi(\mathbf{S}_{t,i}, A_{t,i}) \middle| \mathcal{F}_t \right] \\
&\quad + \mathcal{U}_{t-1}(\boldsymbol{\theta}^*) \\
&= \mathbb{E}\left[\sum_{i=1}^{n} \left\{ Y_{t,i} + \gamma \max_{a_{t+1}} Q(\mathbf{S}_{t+1,i}, a_{t+1}) - Q(\mathbf{S}_{t,i}, A_{t,i}) \right\} \phi(\mathbf{S}_{t,i}, A_{t,i}) \middle| \mathcal{F}_t \right] \\
&\quad + \mathcal{U}_{t-1}(\boldsymbol{\theta}^*) \\
&= \mathbb{E}\left[\sum_{i=1}^{n} \left\{ Y_{t,i} + \gamma \max_{a_{t+1}} Q(\mathbf{S}_{t+1,i}, a_{t+1}) - Q(\mathbf{S}_{t,i}, A_{t,i}) \right\} \phi(\mathbf{S}_{t,i}, A_{t,i}) \middle| \mathbf{S}_t, A_t \right] \\
&\quad + \mathcal{U}_{t-1}(\boldsymbol{\theta}^*) \\
&= \mathcal{U}_{t-1}(\boldsymbol{\theta}^*).
\end{aligned}$$

Thus, $\mathcal{U}_t(\boldsymbol{\theta})$ is a Martingale estimating function (MEF; Godambe, 1991; Heyde, 1997; Hwang and Basawa, 2014), and the operating characteristics of $\widehat{\boldsymbol{\theta}}_{t,n}$ can be derived through properties of the functions $\boldsymbol{\theta} \mapsto \mathcal{U}_t(\boldsymbol{\theta})$. Our focus will be on conditions under which $\Sigma_{t,n}^{-1/2}(\boldsymbol{\theta}^*) \left\{ \widehat{\boldsymbol{\theta}}_{t,n} - \boldsymbol{\theta}^* \right\} \rightsquigarrow$

$\mathrm{N}(0, I_d)$ as $t \to \infty$, where $\Sigma_{t,n}(\boldsymbol{\theta}^*)$ is a (possibly random) scaling matrix. The conditions we provide are standard in MEF-theory. While these conditions are seemingly mild, they can be difficult to verify in practice.

Let $||J||_F = \sqrt{\text{trace}(J^\intercal J)}$ denote the Frobenius norm. Define $\xi_t(\boldsymbol{\theta}^*) \triangleq \text{Var}\left\{u_t(\boldsymbol{\theta}^*)|\mathcal{F}_t\right\} = \mathbb{E}\left\{u_t(\boldsymbol{\theta})u_t(\boldsymbol{\theta})^\intercal \middle| \mathcal{F}_t\right\}$. We assume (C1) that $||\xi_t(\boldsymbol{\theta}^*)|| \to \infty$ almost surely, as $t \to \infty$. Condition (C1) is a regularity condition which ensures sufficient information is generated across the state-action space. To see this, write

$$\text{Var}\left\{u_t(\boldsymbol{\theta}^*)|\mathcal{F}_t\right\} = \sum_{i=1}^{n} \mathbb{E}\left\{\delta_{t,i}^2(\boldsymbol{\theta}^*)|\mathbf{S}_{t,i}, A_{t,i}\right\} \phi(\mathbf{S}_{t,i}, A_{t,i})\phi(\mathbf{S}_{t,i}, A_{t,i})^\intercal,$$

where $\delta_{t,i}(\boldsymbol{\theta}) = Y_{t,i} + \gamma \max_{a_{t+1}} Q(\mathbf{S}_{t+1,i}, a_{t+1}) - Q(\mathbf{S}_{t,i}, A_{t,i})$ is the temporal difference error. If we assume that $\mathbb{E}\left\{\delta_{t,i}^2(\boldsymbol{\theta}^*)|\mathbf{S}_{t,i}, A_{t,i}\right\}$ is bounded below by some constant $c > 0$ with probability one, then a sufficient condition for (C1) is that the mininum eigenvalue of $\sum_{v=1}^{t}\sum_{i=1}^{n} \phi(\mathbf{S}_{v,i}, A_{v,i})\phi(\mathbf{S}_{v,i}, A_{v,i})^\intercal$ diverges to ∞, a condition that appears commonly in asymnptotics for time-series and other stochastic regression settings (Lai and Wei, 1982).

The second condition we require is (C2) that the MEF is regular, i.e., $\boldsymbol{\theta}^*$ is an interior point of $\boldsymbol{\Theta}$, $\mathcal{U}_t(\boldsymbol{\theta})$ is continuously differentiable almost everywhere in a neighborhood $\boldsymbol{\theta}^*$, and for any sequence $\overline{\boldsymbol{\theta}}_t$ converging in probability to $\boldsymbol{\theta}^*$ as $t \to \infty$, we have

$$\left|\left|\xi_t^{-1/2}(\boldsymbol{\theta}^*)\left\{\nabla_{\boldsymbol{\theta}}\mathcal{U}_t(\overline{\boldsymbol{\theta}}_t) - \nabla_{\boldsymbol{\theta}}\mathcal{U}_t(\boldsymbol{\theta}^*)\right\}\xi_t^{-1/2}(\boldsymbol{\theta}^*)\right|\right| \to_p 0,$$

as $t \to \infty$. Condition (C2) is a smoothness condition that rules out the possibility of multiple optimal treatments in any state (at such points, the max operator is not differentiable). It is possible to weaken this condition but at the expense of more complex asymptotic arguments (see Laber et al., 2014).

The third condition we require is (C3) that there exists a constant (non-stochastic) matrix $\boldsymbol{\Omega} \in \mathbb{R}^{d\times d}$ such that

$$\xi_t^{-1/2}(\boldsymbol{\theta}^*)\nabla_{\boldsymbol{\theta}}\mathcal{U}_t(\boldsymbol{\theta}^*)\xi_t^{-1/2}(\boldsymbol{\theta}^*) \to_p \boldsymbol{\Omega},$$

as $t \to \infty$. Condition (C3) is a regularity condition that can typically be verified using strong laws for dependent data (Prakasa Rao, 1987).

Finally, we require (C4) that $-\{\xi_t(\boldsymbol{\theta}^*)\}^{-1/2}\mathcal{U}_t(\boldsymbol{\theta}^*) \rightsquigarrow \mathrm{N}(0, I_d)$. This condition can be established using a Martingale central limit theorem (Hall and Heyde, 2014).

Under (C1)-(C4) and mild moment conditions, it can be shown (Hwang, 2015) that $\boldsymbol{\Omega}\xi_t^{1/2}(\boldsymbol{\theta}^*)\left(\widehat{\boldsymbol{\theta}}_t - \boldsymbol{\theta}^*\right) \rightsquigarrow N(0, I_d)$, which is the desired result with $\Sigma_{t,n}^{-1/2}(\boldsymbol{\theta}^*) = \boldsymbol{\Omega}\xi_t^{1/2}(\boldsymbol{\theta}^*)$. This shows that $\widehat{\boldsymbol{\theta}}_{t,n} = O_p(||\xi_t^{-1/2}(\boldsymbol{\theta}^*)||)$, which in turn can be used to derive the (asymptotic) rate of the cumulative regret. To use this result to construct a confidence set for $\boldsymbol{\theta}^*$ we can use a projection interval as follows. Suppose that if $\boldsymbol{\theta}^*$ were known, one could construct a consistent estimator $\widehat{\boldsymbol{\Sigma}}_{t,n}^{-1/2}(\boldsymbol{\theta}^*)$ of $\boldsymbol{\Sigma}_{t,n}^{-1/2}(\boldsymbol{\theta}^*)$. Let $\chi^2_{d,1-\alpha}$ be the upper $(1-\alpha)\times 100\%$ percentile of a chi-squared distribution with d-degrees of freedom and define

$$\boldsymbol{\Gamma}_{t,n,1-\alpha} = \left\{\boldsymbol{\Theta} \in \boldsymbol{\Theta} \,:\, \widehat{\Sigma}_{t,n}^{-1/2}(\boldsymbol{\theta})\left(\widehat{\boldsymbol{\theta}}_{t,n} - \boldsymbol{\theta}\right) \leqslant \chi^2_{d,1-\alpha}\right\}.$$

It follows that $P\left\{\boldsymbol{\theta}^* \in \Gamma_{t,n,1-\alpha}\right\} \geqslant 1-\alpha+o_P(1)$. The set $\Gamma_{t,n,1-\alpha}$ is thus a valid (asymptotic) confidence region for $\boldsymbol{\theta}^*$ which in turn can be used to construct projection sets for other functions of $\boldsymbol{\theta}^*$, e.g., the value of the optimal regime (see also Zhang et al., 2022).

5. Open problems and ongoing work

Our goal in this chapter was to introduce TS as a flexible and extensible methodology for adaptive clinical trials especially in the context of mobile- and tele-health. However, despite a long history of empirical and theoretical study, there are a number of pressing open problems associated with TS. One such problem is statistical efficiency. The estimating equations we described are used widely in practice but they need not lead to the smallest asymptotic variance among the class of regular MEFs. Furthermore, if the posited class of models for the Q-function is misspecified, the solution to the MEF need not recover the projection of the true Q-function on the model class (see Baird, 1995; Leete and Laber, 2022). An

important open question is how to construct the estimating equations to obtain efficiency and the projection property. In principle, the efficient weights for the estimating equations can be obtained using the theory of optimal MEFs (Hwang and Basawa, 2011). However, the efficient weights depend on the unknown system dynamics and the cost of estimating the optimal weights risks further misspecifiction and/or inflated variance (Leete and Laber, 2022).

Another important open problem is interim analysis and optimal stopping for adaptive experiments under TS. In theory, one could obtain (approximate) joint asymptotic normality for the estimated parameters at multiple pre-specified analysis points and subsequently derive stopping boundaries (Jennison and Turnbull, 1999). However, the derivations of these boundaries are likely to be intricate.

Lastly, we note that TS may fail to perform well if the underlying system is non-stationary. One *ad hoc* approach is to limit the look-back period and only use estimating equations constructed from recent data. An important problem is how to adaptively choose the look-back period to optimally balance bias and variance.

References

Agrawal, S. and Goyal, N. (2012). Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1. JMLR Workshop and Conference Proceedings.

Agrawal, S. and Goyal, N. (2013). Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135. PMLR.

Armitage, P. (1960). *Sequential medical trials.* Blackwell Scientific Publications.

Baird, L. (1995). Residual algorithms: Reinforcement learning with function approximation. In *Machine Learning Proceedings 1995*, pages 30–37. Elsevier.

Bellman, R. (1952). On the theory of dynamic programming. *Proceedings of the National Academy of Sciences* **38,** 716–719.

Berry, D. A. and Fristedt, B. (1985). Bandit problems: sequential allocation of experiments (monographs on statistics and applied probability). *London: Chapman and Hall* **5,** 7–7.

Bertsekas, D. (2012). *Dynamic programming and optimal control: Volume I*, volume 1. Athena scientific.

Bibaut, A., Dimakopoulou, M., Kallus, N., Chambaz, A., and van der Laan, M. (2021). Post-contextual-bandit inference. *Advances in Neural Information Processing Systems* **34,**.

Carpenter, S. M., Menictas, M., Nahum-Shani, I., Wetter, D. W., and Murphy, S. A. (2020). Developments in mobile health just-in-time adaptive interventions for addiction science. *Current Addiction Reports* **7,** 280–290.

Chapelle, O. and Li, L. (2011). An empirical evaluation of thompson sampling. *Advances in Neural Information Processing Systems* **24,**.

Chatterjee, S. and Bose, A. (2005). Generalized bootstrap for estimating equations. *Annals of Statistics* **33,** 414–436.

Chernoff, H. (1959). Sequential design of experiments. *The Annals of Mathematical Statistics* **30,** 755–770.

Ernst, D., Geurts, P., and Wehenkel, L. (2005). Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research* **6,**.

Ertefaie, A. and Strawderman, R. L. (2018). Constructing dynamic treatment regimes over indefinite time horizons. *Biometrika* **105,** 963–977.

Fisher, H. M., Winger, J. G., Miller, S. N., Wright, A. N., Plumb Vilardaga, J. C., Majestic, C., Kelleher, S. A., and Somers, T. J. (2021). Relationship between social support, physical symptoms, and depression in women with breast cancer and pain. *Supportive*

Care in Cancer **29,** 5513–5521.

Godambe, V. P. (1991). *Estimating functions*. Oxford University Press.

Hadad, V., Hirshberg, D. A., Zhan, R., Wager, S., and Athey, S. (2021). Confidence intervals for policy evaluation in adaptive experiments. *Proceedings of the National Academy of Sciences* **118,** e2014602118.

Hall, P. and Heyde, C. C. (2014). *Martingale limit theory and its application*. Academic press.

Hernan, M. A. and Robins, J. M. (2020). *Causal Inference: What If*. CRC Press.

Hernández-Neuta, I., Neumann, F., Brightmeyer, J., Ba Tis, T., Madaboosi, N., Wei, Q., Ozcan, A., and Nilsson, M. (2019). Smartphone-based clinical diagnostics: towards democratization of evidence-based health care. *Journal of Internal Medicine* **285,** 19–39.

Heyde, C. C. (1997). *Quasi-likelihood and its application: a general approach to optimal parameter estimation*. Springer.

Hwang, S. (2015). Some characterizations of non-ergodic estimating functions for stochastic processes. *Journal of the Korean Statistical Society* **44,** 661–667.

Hwang, S. and Basawa, I. (2011). Godambe estimating functions and asymptotic optimal inference. *Statistics & Probability Letters* **81,** 1121–1127.

Hwang, S. and Basawa, I. (2014). Martingale estimating functions for stochastic processes: A review toward a unifying tool. *Contemporary Developments in Statistical Theory: A Festschrift for Hira Lal Koul* pages 9–28.

Jennison, C. and Turnbull, B. W. (1999). *Group sequential methods with applications to clinical trials*. CRC Press.

Laber, E. B., Lizotte, D. J., Qian, M., Pelham, W. E., and Murphy, S. A. (2014). Dynamic treatment regimes: Technical challenges and applications. *Electronic Journal of Statistics* **8,** 1225.

Laber, E. B. and Staicu, A.-M. (2018). Functional feature construction for individualized treatment regimes. *Journal of the American Statistical Association* **113,** 1219–1227.

Lai, T. L. and Wei, C. Z. (1982). Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics* **10,** 154–166.

Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

Lavori, P. W. and Dawson, R. (2004). Dynamic treatment regimes: practical design considerations. *Clinical trials* **1,** 9–20.

Leete, O. and Laber, E. (2022). Model-assisted v-learning. Technical report, Duke University.

Li, Z., Laber, E., and Meyer, N. (2022). Thompson sampling for pursuit-evasion problems. In *2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, pages 1–8. IEEE.

Lieberman, J. A., Stroup, T. S., McEvoy, J. P., Swartz, M. S., Rosenheck, R. A., Perkins, D. O., Keefe, R. S., Davis, S. M., Davis, C. E., Lebowitz, B. D., et al. (2005). Effectiveness of antipsychotic drugs in patients with chronic schizophrenia. *New England Journal of Medicine* **353,** 1209–1223.

Luckett, D. J., Laber, E. B., Kahkoska, A. R., Maahs, D. M., Mayer-Davis, E., and Kosorok, M. R. (2019). Estimating dynamic treatment regimes in mobile health using v-learning. *Journal of the American Statistical Association* **115,**.

Ma, T., Cai, H., Qi, Z., Shi, C., and Laber, E. B. (2023). Sequential knockoffs for variable selection in reinforcement learning. *arXiv preprint arXiv:2303.14281* .

Manschot, C., Laber, E., and Davidian, M. (2022). Interim monitoring of sequential multiple assignment randomized trials using partial information. *arXiv preprint* .

Murphy, S. A. (2005a). An experimental design for the development of adaptive treatment strategies. *Statistics in medicine* **24,** 1455–1481.

Murphy, S. A. (2005b). A generalization error for q-learning. *Journal of Machine Learning Research* **6,** 1073–1097.

Mustanski, B., Saber, R., Macapagal, K., Matson, M., Laber, E., Rodrgiuez-Diaz, C., Moran, K. O., Carrion, A., Moskowitz, D. A., and Newcomb, M. E. (2022). Effectiveness of the smart sex ed program among 13–18 year old English and Spanish speaking adolescent men who have sex with men. *AIDS and Behavior* pages 1–12.

Norwood, P., Davidian, M., and Laber, E. (2022). Adapative randomization methods for sequential multiple assignment randomized trials via thompson sampling. *Technical report: North Carolina State University, Department of Statistics* .

Prakasa Rao, B. L. S. (1987). *Asymptotic theory of statistical inference*. John Wiley & Sons, Inc.

Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society* **58,** 527–535.

Robbins, H. (1956). A sequential decision problem with a finite memory. *Proceedings of the National Academy of Sciences* **42,** 920–923.

Russo, D., Van Roy, B., Kazerouni, A., Osband, I., and Wen, Z. (2017). A Tutorial on Thompson Sampling. *arXiv:1707.02038 [cs]* arXiv: 1707.02038.

Schulte, P. J., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2014). Q- and a-learning methods for estimating optimal dynamic treatment regimes. *Statistical Science: a review journal of the Institute of Mathematical Statistics* **29,** 640.

Slivkins, A. et al. (2019). Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning* **12,** 1–286.

Sutton, R. S. (1997). On the significance of Markov decision processes. In *Artificial Neural*

Networks—ICANN'97: 7th International Conference Lausanne, Switzerland, October 8–10, 1997 Proceeedings 7, pages 273–282. Springer.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Thall, P. F. and Wathen, J. K. (2005). Covariate-adjusted adaptive randomization in a sarcoma trial with multi-stage treatments. *Statistics in Medicine* **24,** 1947–1964.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* **25,** 285–294.

Trippa, L., Lee, E. Q., Wen, P. Y., Batchelor, T. T., Cloughesy, T., Parmigiani, G., and Alexander, B. M. (2012). Bayesian adaptive randomized trial design for patients with recurrent glioblastoma. *Journal of Clinical Oncology* **30,** 3258.

Tsiatis, A. A., Davidian, M., Holloway, S. T., and Laber, E. B. (2019). *Dynamic treatment regimes: Statistical methods for precision medicine*. Chapman and Hall/CRC.

van der Laan, M. J. and Petersen, M. L. (2007). Causal effect models for realistic individualized treatment and intention to treat rules. *The international journal of biostatistics* **3,**.

Wager, S. and Xu, K. (2021). Diffusion asymptotics for sequential experiments. *arXiv preprint arXiv:2101.09855* .

Wald, A. (1947). *Sequential analysis*. John Wiley.

Wang, L., Laber, E. B., and Witkiewitz, K. (2017). Sufficient Markov decision processes with alternating deep neural networks. *arXiv preprint arXiv:1704.07531* .

Wang, L., Zhang, W., He, X., and Zha, H. (2018). Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2447–2456.

Zhan, R., Hadad, V., Hirshberg, D. A., and Athey, S. (2021). Off-policy evaluation via

adaptive weighting with data from contextual bandits. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2125–2135.

Zhang, K., Janson, L., and Murphy, S. (2020). Inference for batched bandits. *Advances in Neural Information Processing Systems* **33,** 9818–9829.

Zhang, K. W., Janson, L., and Murphy, S. A. (2022). Statistical inference after adaptive sampling in non-Markovian environments. *arXiv preprint arXiv:2202.07098* .

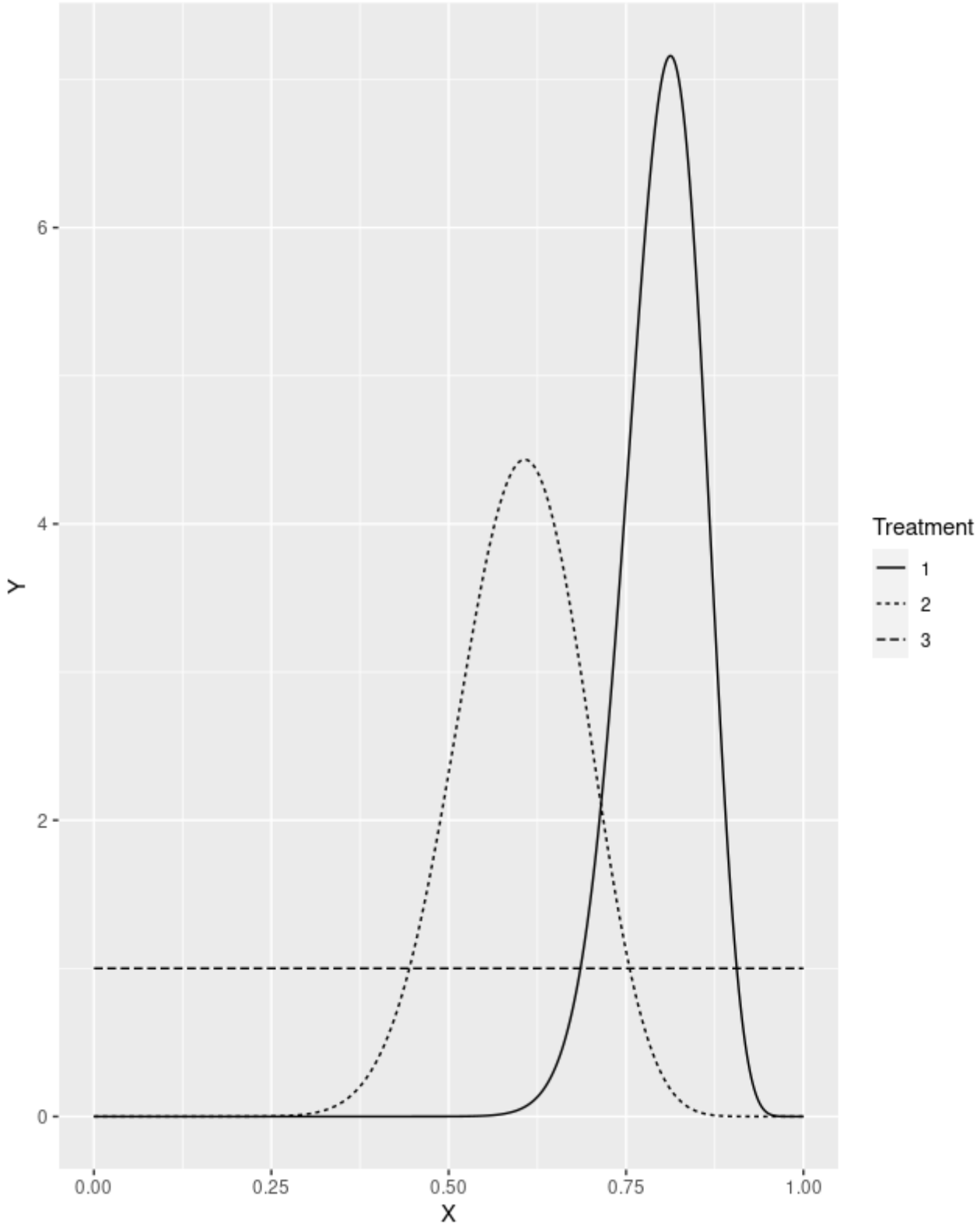

Figure 1. Beta distribution densities for hypothetical basket trial.