Reinforcing RCTs with Multiple Priors while Learning about External Validity *

Frederico Finan †
UC Berkeley

Demian Pouzo‡
UC Berkeley

April 2022

Abstract

This paper presents a framework for how to incorporate prior sources of information into the design of a sequential experiment. These sources can include previous experiments, expert opinions, or the experimenter's own introspection. We formalize this problem using a multi-prior Bayesian approach that maps each source to a Bayesian model. These models are aggregated according to their associated posterior probabilities. We evaluate a broad class of policy rules according to three criteria: whether the experimenter learns the parameters of the payoff distributions, the probability that the experimenter chooses the wrong treatment when deciding to stop the experiment, and the average rewards. We show that our framework exhibits several nice finite sample properties, including robustness to any source that is not externally valid.

Keywords: Reinforcement Learning, External Validity, RCTs, Multiple Priors, Bayesian Learning.

JEL: C11, C50, C90, O12.

Table of Contents

1 Introduction

Governments around the world are importing policies or programs that have been shown to be successful in other settings. Take for example, Mexico's conditional cash transfer program, *Oportunidades*. Since its inception in 1997, it has been replicated in over 52 countries around the world.[1] Other examples of policy interventions that have been exported to various settings include chartiable giving (Karlan and List, 2007), pay-for-performance schemes for teachers, charter schools (Chabrier et al., 2016), access to microcredit (Banerjee et al., 2015b), and BRAC's ultra-poor graduation program (Banerjee et al., 2015a).

When a policymaker decides to adopt a policy based on evidence from previous evaluations, she must assess whether those results will extrapolate to her setting. And depending on her degree of uncertainty, the policymaker may want to experiment. On the one hand, if the policymaker is certain that the benefits would extrapolate then the learning gains from experimentation may not

---

[1]See https://www.worldbank.org/en/news/feature/2014/11/19/un-modelo-de-mexico-para-el-mundo.

justify the costs of withholding the program's benefits from her beneficiaries. On the other hand, if her uncertainty is high, she may want to experiment first before expanding the program to scale.

Two issues lie at the heart of this decision. One is how much experimentation (versus exploitation) should our policymaker do? And two, how do we incorporate knowledge from experts or previous experiments into our decision process? The first question is relatively well understood and a few recent studies have shown how we can use algorithms, such as Thompson Sampling or ϵ-greedy, to solve this problem and achieve efficiency gains over a standard randomized control trial. But within this framework, the second question remains relatively unexplored. One of the key contributions of this paper is to provide a simple, but novel approach for doing so.

Setup   We consider a policymaker who has to decide how to assign a set of treatments sequentially to an eligible population and when to stop the experiment. Subjects arrive in stages and at the beginning of each stage, the policymaker must first decide whether to stop the experiment. If she stops the experiment, she then assigns what she thinks is the best treatment to all subsequent subjects. But if the policymaker decides to continue the experiment, she assigns treatment just to the new arrivals and then moves onto a new stage. At each stage, the policymaker knows the history of previous treatment assignments and the corresponding realized outcomes, but does not know the probability distributions of potential outcomes, which she tries to learn about using the observed data. The policymaker does, however, have prior information about these distributions, which can arise from many sources, including her own introspection and knowledge, previous experiments, or expert opinions.

The policymaker is unwilling or unable — due to lack of enough a-priori information — to *ex-ante* aggregate these sources into a single prior. Thus, we depart from the (single-prior) Bayesian paradigm and enhance a multi-armed bandit setup with a multi-prior Bayesian learning model (e.g. Epstein and Schneider (2003) and references therein), wherein each source of information is treated as a different prior. Multi-prior Bayesian learning models, while not commonly used in applied economics, provide a natural framework for incorporating multiple sources of information whose informativeness cannot be determined a priori.

As the policymaker gathers more data, she updates each of these priors using Bayes' rule and then takes a weighted average of each source's posterior where the weights depend on how well the sources fit the observed data.[2] On the basis of these beliefs, the policymaker then decides whether to stop the experiment and which treatment to assign. By incorporating potentially useful information,

---

[2]In contrast, a single prior Bayesian agent would have aggregated the different priors using fixed weights that do not depend on the observed data.

our policymaker may be able to stop the experiment sooner, thereby generating efficiency gains without increasing the risk of adopting the incorrect treatment.

In settings in which the policymaker must learn the truth, it is common not to use the optimal assignment rule. This rule (i.e. the one that maximizes her *subjective* payoff) can have undesirable properties, such as failing to learn the correct treatment effects or being hard to compute and implement.[3] As a result, the literature on multi-armed bandits have studied different heuristic rules such as ϵ-greedy (Watkins, 1989) and Thompson Sampling (Thompson, 1933) and its refinements (e.g. Upper Confidence Bounds (Lai and Robbins, 1985), or exploration sampling (Kasy and Sautmann, 2021; Russo, 2016)). We take a different approach and study a large class of policy rules that encompass, among others, the aforementioned examples. Importantly, we find that the only feature of the policy rule that matters for performance is the exploration structure – a sequence quantifying the amount of experimentation that occurs under a given policy rule at each stage of the experiment.

Performance Criteria   Given that optimality from the perspective of the policymaker may not be desirable, we evaluate our class of assignment rules on the basis of three regularly-used outcomes that are considered to be important from the point of view of an outside observer. Specifically, we explore whether the policymaker learns the true average treatment effects and at what rate. We also consider the likelihood that the policymaker does not choose the most beneficial treatment arm when deciding to stop the experiment. The third outcome measures the average payoff of the policymaker. Unlike the other two criteria, which are statistical in nature (i.e. they describe statistical properties of the experiment and its assignment rule), this outcome captures how much subjects benefit in net from the experiment both during and afterwards. When evaluated along these criteria, we can show, both theoretically and via Monte Carlo simulations, that our setup exhibits several nice finite sample properties, including robustness to incorrect priors.

Main Findings   We show that our policymaker will learn the average treatment effects, in the sense that her posterior mean of the potential outcome distribution concentrates around the true mean, and it does so at a rate of $1/(\sqrt{t}h_t^2)$, where t is the number of stages and h_t is the amount of experimentation. That this concentration result holds was not, ex ante, obvious: in contrast to

---

[3]To illustrate this point, consider a simple model with two treatments, A and B. For simplicity, suppose the policymaker knows that the average effect of treatment A is zero. The policymaker, however, does not know the true average effect of treatment B and incorrectly believes that it is negative. In this simple example, an optimal policy is to never assign treatment B; and without feedback, the policymaker will never update her (incorrect) prior that treatment B is bad. While this assignment rule is optimal from the perspective of the policymaker, it is undesirable from an objective point of view. This example also illustrates the need for experimentation because such a situation would not occur if the policy rule involved some degree of experimentation.

a standard randomized control trial setting, the policy functions in our setup are quite general and can depend on the entire history of play, thus creating time-dependence in the data. Nevertheless, by exploiting the concept of the exploration structure and Azuma-Hoeffding type concentration inequalities for Martingales, we not only obtain the rate of $1/(\sqrt{t}h_t^2)$, but we can also characterize and quantify how this rate depends on the initial parameters of the setup.

Importantly, we are able to show that our aggregation method exhibits an attractive robustness property. In other words, our model discards sources that do not extrapolate well to the current experiment, thereby exhibiting robustness to sources of information that are not externally valid. To aggregate her multiple priors, our policymaker uses a Bayesian approach that weights each prior according to the posterior probability that a particular model best fits the observed data within the class of sources being considered. Thus, if relative to the other priors, one of the policymaker's priors (about the average effects of the treatments) puts "low probability" on the true mean, then our approach will place close to zero weight on this source when aggregating across sources. Consequently, this prior will have little to no effect on the policymaker's decisions or the learning rate. Similarly, sources whose priors put high probability on the truth receive higher weights that can approach one in finite samples. This feature gives rise to an oracle type property wherein our concentration rates are close to those associated to the best source (the one with priors more concentrated around the truth) provided the other sources are sufficiently separated from this one.

Besides assigning treatments, our policymaker also has to consider when to stop the experiment and subsequently, what treatment to adopt. Both the duration of the experiment and adopting the correct treatment can have important welfare consequences. In our setup, the policymaker works with a class of stopping rules that stops the experiment when the average effect of a treatment is sufficiently above the others. This class of rules resembles the standard test of two means, but takes into account the fact that the data are not IID. Of course, whenever we stop an experiment, we worry about the possibility of making a mistake (i.e. not choosing the most beneficial treatment). We characterize the bounds on the probability of making a mistake for our setup. We show that these bounds decay exponentially fast with the length of the experiment, and that they are non-increasing in the degree of experimentation and in the size of the treatment effects. Moreover, we propose stopping rules that for any given tolerance level will yield a lower probability of making a mistake.

Finally, we also compute bounds for the rate at which the average observed outcomes converges to the maximum expected outcome. We show that the rate of convergence for these bounds are governed by an "exploitation versus exploration" trade-off. If we increase the degree of experimen-

tation (less exploitation, more exploration) our data become more independent and the underlying uncertainty decreases. However, by exploring more, we are also increasing the bias associated with not choosing the optimal treatment. Unfortunately, these bounds are sufficiently complicated that we cannot characterize analytically the "optimal" degree of experimentation. Nevertheless, the results do suggest that pure experimentation (as in the case of an RCT) is unlikely to be optimal, and we verify this numerically in a series of simulations.

Charitable Giving   To further illustrate our procedure, we also present a proof-of-concept using data from a recent charitable fundraising experiment (Karlan and List, 2020). These types of experiments provide a nice test case because charitable giving is an outcome that responds relatively quickly to treatment. It is also an experiment that has been replicated in various settings, thus allowing for multiple priors (e.g. Karlan and List (2007)). Using these data for our potential outcome distributions, we show that by incorporating multiple priors, our policymaker can stop the experiment in a third of the time, without a significant increase to the probability of making a mistake, thereby resulting in large performance gains relative to a standard RCT.

Contributions to the Literature   Our paper relates to three strands of the literature. First, we speak to an extensive multi-disciplinary literature on adaptive experimental design. Much of the focus of this literature has been on the multi-arm bandit problem, which considers how best to assign experimental units sequentially across treatment arms. Depending on the objective function, numerous studies have proposed a variety of alternative algorithms that, on average, outperform the static assignment mechanisms of traditional RCTs.[4] In this paper, we focus less about constructing an alternative policy function than about on how to introduce information from different sources for a given class of policy functions. By doing so, the fundamental 'earn vs learn' tradeoff that characterizes the multi-arm bandit problem is not only a function of sampling variability in target data, but also uncertainty over the data generating process of the source data. To our knowledge, this is the first paper to introduce multiple priors into the design of an adaptive experiment.

Much of the literature on multi-armed bandits has focused on deriving bounds on expected regret for specific solution heuristics.[5] Instead, we focus on alternative performance criteria, such as average outcomes, the probability of making a mistake, and concentration rates for posterior means,

---

[4]See Athey and Imbens (2019) for a survey of machine learning techniques as it applies to experimental design and problems in economics.

[5]For example, related to bounds on regret, see Agrawal and Goyal (2017) and Russo and Van Roy (2016) for regret bounds for Thompson sampling; or Cesa-Bianchi and Lugosi (2006) for a broad discussion about multi-armed bandit problems and bounds on regret.

which to the best of our knowledge have not been formalized in a multi-prior multi-arm Bayesian bandit framework.[6] Moreover, the results we derive are for a general class of solution heuristics, not a specific one. For these reasons, even though we do not view the technical results as the primary contribution of the paper, we do believe that they might be of independent interest even in standard multi-arm bandit problems. Furthermore, we view our paper as complementary to this existing literature, as techniques tailored for particular solution heuristics can be combined with our multi-prior Bayesian setting to obtain sharper theoretical guarantees.

By introducing issues of externality validity into the multi-arm bandit problem, our study also connects to the literature on measuring the generalizability of experiments. In general, scholars have taken three approaches for assessing external validity. One common approach is to measure how well treatment effect heterogeneity extrapolates to 'left out' study sites. Under the assumption that study site characteristics are independent of potential outcomes, a number of studies applying alternative estimators have interpreted the out-of-sample prediction errors as a measure or test of external validity.[7] A related approach uses local average treatment effects across different complier populations to test for evidence of external validity (e.g. Angrist and Fernández-Val (2013); Kowalski (2016); Bisbee et al. (2017)). The general idea being that if differences in observable characteristics across subgroups explain differences in treatment effect heterogeneity then we can make some claim for external validity. A third common approach adopted in the meta-analysis literature is the use of hierarchical models to aggregate treatment effects across different study sites. A byproduct of this framework is a "pooling factor" across study sites that has a natural interpretation of generalizability. The factor compares the sampling variation of a particular study site to the underlying variation in treatment heterogeneity: the higher the measure, the larger the sampling error and the less informative the study site is about the overall treatment effect (e.g. Vivalt (2020), Gelman and Carlin (2014), Gelman and Pardoe (2006), Meager (2020)).[8]

Our paper contributes to these approaches in two ways. First, we provide a formal definition for a subjective Bayesian model to be externally invalid using a Kullback-Leibler (KL) divergence criteria. Importantly, our definition offers a way to quantify or rank external invalidity among

---

[6]Average outcomes is related to regret. However, we do not provide bounds for the expected value, but instead provide exponential inequalities for the tail probability. There are classical results related to the probability of making a mistake stemming from the foundational work by Chernoff (1959) and Wald (1945).

[7]See for example Dehejia et al. (2021), Stuart et al. (2011), Buchanan et al. (2018), Imai and Ratkovic (2013), Joseph Hotz et al. (2005) and the references cited therein.

[8]The first and third approaches — and hence our paper as well — relates to a burgeoning sub-branch of machine learning called transfer learning (see Pan and Yang (2010) for a survey) wherein a model developed for a task is re-used as the starting point for a model on a second task. Even though elements of our problem are conceptually similar, to the best of our knowledge both our setup and approach are different to those considered in transfer learning.

models. Second, we provide a link between this ranking of external invalidity and our aggregation method. We show that, as t diverges, the weights are only positive for the least externally invalid models, allowing us to interpret these weights as measures of external validity.

While it is natural to interpret our measure of external validity in the context of other experiments, our setup is agnostic as to the source of the information and its level of uncertainty. Whether the policymaker's priors come from previous experiments, observational studies, or expert opinions is immaterial for our setup. In this respect, our study also relates to a nascent, but growing literature measuring the extent to which experts can forecast experimental results (e.g. DellaVigna and Pope (2018); DellaVigna et al. (2020)). Our paper provides a method for incorporating these forecasts in the design of policy evaluations in a manner that is robust to misspecified priors or behavioral biases (Vivalt and Coville, 2021).

Organization of the Paper   The structure of the paper proceeds as follows. In Section 2, we set up the problem. We present two versions of the setup, one for the general model and the other for a Gaussian model. In Section 3, we provide analytical results for the Gaussian model. We then illustrate the main analytical results by simulation in Section 4. In Section 5, we illustrate our procedure using data from a charitable giving experiment. Section 6 concludes.

2   Setup

In this section, we describe the problem our policymaker (PM) aims to solve. We first present the general model, followed by a more specialized problem that is the main focus of the paper.

2.1   General Model

Our PM's problem consists of three parts: the experiment, the learning framework, and the policy functions.

The Experiment

The PM has to decide how to assign a treatment to a given unit (e.g. individuals or firms) and when to stop the experiment. We define an experiment by a number of instances $T \in \mathbb{N}$; a discrete set of observed characteristics of the unit, $\mathbb{X}$; a set of treatments $\mathbb{D} := \{0, ..., M\}$; and the set of potential outcomes. For now, we do not include a payoff function.

At this point, it is useful to introduce some notation. For each $(d,x) \in \mathbb{D} \times \mathbb{X}$, let $Y_t(d,x) \in \mathbb{R}$ denote the potential outcome associated with treatment d and characteristic x in instance t; also, let $Y_t(d) := (Y_t(d,x))_{x \in \mathbb{X}}$. Let $D_t(x) \in \mathbb{D}$ be the treatment assigned to the unit with characteristic x in instance t. We denote the *observed* outcome of the unit with characteristic x in instance t as $Y_t(D_t(x),x)$.

The experiment has the following timing. At each instance, $t \in \{1,...,T\}$, the PM is confronted with $|\mathbb{X}| < \infty$ units, one for each value of the observed characteristic. At the beginning of the period, the PM decides whether to stop the experiment.

- If the PM decides to stop the experiment,

    - she chooses a treatment assignment at instance t that will be applied to all subsequent units.

- If the PM does not stop the experiment,

    - she chooses a treatment assignment for each unit x at time t.

    - Nature draws potential outcomes, $Y_t(d,x)$, for each unit.

    - The PM only observes the outcome corresponding to the assigned treatment, i.e. $Y_t(D_t(x),x)$.

We impose the following restriction on the data generating process for the vector of potential outcomes.

Assumption 1. *For each $t \in \{1,...,T\}$ and each $x \in \mathbb{X}$, $(Y_t(d,x))_{d \in \mathbb{D}}$ is drawn IID and $Y(d,x) \sim P(\cdot|d,x) \in \Delta(\mathbb{R})$.*

Assumption 1 implies that units do not self-select across instances, i.e., the types of unit treated in instance t are the same as the types treated in instance t'. Implicit in this assumption and framework is also the absence of any selection into treatment or attrition, which is reasonable to assume for most experimental settings.

Finally, the assumption that the PM is confronted with $|\mathbb{X}| < \infty$ units, one for each value of the observed characteristic, is made out of convenience: it is straightforward to extended our theory to situations where the PM receives a random number of units, including zero, for each characteristic, provided this random number is exogenous. However, to extend the assumption of discrete covariates — $|\mathbb{X}| < \infty$ — to continuous ones is non-trivial. For learning in multi-arm bandits with continuous covariates, we refer the reader to Dimakopoulou et al. (2017) and references therein, as well as to Qin and Russo (2022) where the authors adapt the Thompson Sampling algorithm to

handle a potentially non-stationary sequence of covariates influencing the arms' performance.

The Learning Model

The PM does not know the probability distribution of potential outcomes P, but does have prior beliefs about it. This prior knowledge can come from many sources: the PM's own prior knowledge, expert opinions, or past experiments. Importantly, we allow for multiple sources, in case the PM is unwilling or unable to discard one in favor of the others. If her prior sources of information extrapolate to the current experiment, then she should use them because they contain relevant information. But if some sources are not externally valid, then incorporating them in her assignment of treatment may lead to incorrect decisions, at least in finite samples. Thus, our PM not only faces the question of whether to incorporate the different sources, but how to aggregate them as well. We formalize this "external validity dilemma" by using a multiple prior Bayesian model.

Formally, for each $(d,x) \in \mathbb{D} \times \mathbb{X}$, the PM has a family of PDFs indexed by a finite dimensional parameter $\theta \in \Theta$, $\mathcal{P}_{d,x} := \{p_\theta : \theta \in \Theta\}$, that describes what she believes are plausible descriptions of the true probability of the potential outcome $Y(d,x)$. The PM also has $L+1$ prior beliefs, $(\mu_0^o(d,x))_{o=0}^{L}$, regarding which elements of $\mathcal{P}_{d,x}$ are more likely; these prior beliefs summarize the prior knowledge obtained from the $L+1$ different sources.

For each $(d,x) \in \mathbb{D} \times \mathbb{X}$, the family $\mathcal{P}_{d,x}$ and the collection of prior beliefs give rise to $L+1$ subjective Bayesian models for $P(\cdot|d,x)$. Given the observed data of past treatments and outcomes, at instance $t \geq 1$, the PM will observe the realized outcome $Y_t(D_t(x),x)$ and the treatment assignment $D_t(x)$. Using Bayesian updating, she will then form posterior beliefs for each model given by

$$\mu_t^o(d,x)(A) = \frac{\int_A p_\theta(Y_t(D_t(x),x))^{1\{D_t(x)=d\}} \mu_{t-1}^o(d,x)(d\theta)}{\int_\Theta p_\theta(Y_t(D_t(x),x))^{1\{D_t(x)=d\}} \mu_{t-1}^o(d,x)(d\theta)}, \ \forall A \subseteq \Theta \ Borel.$$

Observe that the belief is updated using observed data, $(Y_t(D_t(x),x), D_t(x))$ and using p_θ as the PDF $Y_t(D_t,x)$ given $D_t(x) = d$.[9] That the belief for (d,x) is only updated if $D_t(x) = d$ is analogous to the missing data problem featured in experiments under the frequentist approach.

It is worth noting that we specify subjective models for $P(\cdot|d,x)$ for each (d,x), as opposed to the joint distribution over $(Y(d,x))_{d\in\mathbb{D}}$. This decision is innocuous when the PM's objective is to learn about the distributions or moments of each $Y(d,x)$, such as the average or quantiles. If, however, the objective is to learn features of the joint distribution — e.g., the correlation between potential

---

[9]Because the PM already knows the probability of $D_t(x)$, she does not need to include it as part of the Bayesian updating problem.

10

outcomes — then we would have to modify the learning model. The subjective model would now be a family of joint probability distributions over $(Y(d,x))_{d \in \mathbb{D}}$. We present such a model in the Supplemental Material I, in which we also show how to obtain the learning model presented here as a particular case.

Model Aggregation & External Validity. Faced with $L+1$ distinct subjective Bayesian models, $\{\langle \mathcal{P}_{d,x}, \mu_0^o(d,x) \rangle\}_{o=0}^L$, our PM has to aggregate this information. There are different ways to do this; we choose one that, at each instance t, averages the posterior beliefs of each model using as weights the posterior probability that model o best fits the observed data within the class of models being considered, i.e.,

$$\mu_t^\alpha(d,x)(A) := \sum_{o=0}^L \alpha_t^o(d,x)\mu_t^o(d,x)(A), \ \forall A \subseteq \Theta \ Borel, \tag{2.1}$$

where

$$\alpha_t^o(d,x) := \frac{\int \prod_{s=1}^t p_\theta(Y_s(d,x))^{1\{D_s(x)=d\}} \mu_0^o(d,x)(d\theta)}{\sum_{o=0}^L \int \prod_{s=1}^t p_\theta(Y_s(d,x))^{1\{D_s(x)=d\}} \mu_0^o(d,x)(d\theta)}.$$

We can interpret $\alpha_t^o(d,x)$ as a measure of the PM's subjective probability that the prior belief associated with source o for (d,x) is externally valid for her current experiment. To expound on this last point, we introduce a definition of "external validity" that we can relate to the behavior of $(\alpha_t^o(d,x))_{o=0}^L$.

For each $(d,x) \in \mathbb{D} \times \mathbb{X}$ and $\mathcal{P}_{d,x}$, let

$$\theta \mapsto KL_{d,x}(\theta) := E_{p(\cdot|d,x)} \left[\log \frac{p(Y(d,x) \mid d,x)}{p_\theta(Y(d,x))} \right]$$

be the Kullblack-Leibler (KL) divergence, which acts as a notion of distance between the true PDF of $Y(d,x)$ — given by $p(\cdot|d,x)$ — and a "subjective" one $p_\theta \in \mathcal{P}_{d,x}$.[10] By combining the KL with the prior, $\mu_0(d,x)$ — that determines the likelihood of each $\theta \in \Theta$ —, we construct a notion of distance between the true PDF and the subjective Bayesian model, $\langle \mathcal{P}_{d,x}, \mu_0(d,x) \rangle$, and in turn, propose a definition of external (in)validity.

---

[10]Indeed, $KL_{d,x} \geq 0$ and $KL_{d,x}(\theta) = 0$ iff $p_\theta = p(\cdot|d,x)$. It is not, however, a distance in the formal sense as it does not satisfy the triangle inequality. Finally, the KL divergence does not depend on o as all models are assumed to have the same family $\mathcal{P}_{d,x}$.
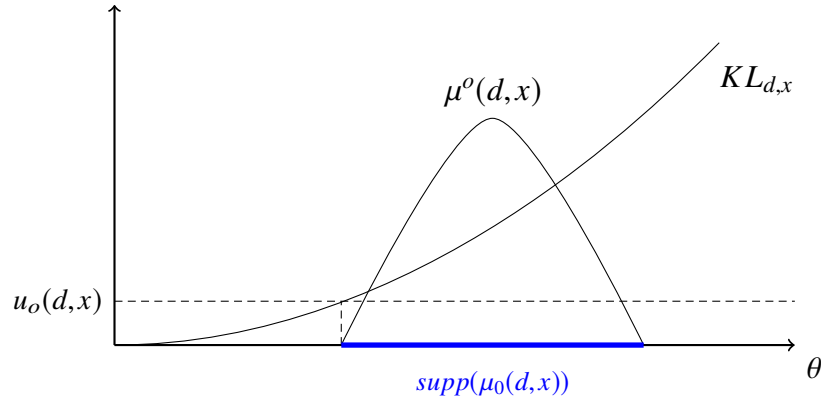
11

Definition 1 (Externally Invalid Subjective Bayesian Model). *We say a subjective Bayesian model* $\langle \mathcal{P}_{d,x}, \mu_0^o(d,x) \rangle$ *is externally invalid for* (d,x) *if*

$$u_o(d,x) > 0$$

where $u_o(d,x) := \inf_{\theta \in supp(\mu_0^o(d,x))} KL_{d,x}(\theta)$.[11]

According to this definition, a model is externally invalid if the associated source (i.e., the prior) puts zero probability to any PDF that is equivalent — as measured by the KL divergence — to the true PDF. If no $u_o(d,x) > 0$ exists, we say the subjective Bayesian model is externally valid for (d,x).

To illustrate this definition, consider the graph below. The horizontal axis indicates different values of θ, where the θ of the true PDF is set at the origin. The vertical axis corresponds to the resulting KL distance between the true PDF of $Y(d,x)$ and the "subjective" one p_θ which is depicted as the curve $KL_{d,x}$. We also plot a subjective belief $\mu^o(d,x)$ over the set of θs, for which the model places positive probability. According to our definition, the model $\mu^o(d,x)$ is externally invalid because as the graph depicts there exists a $u_o(d,x) > 0$.

A couple of remarks about this definition are in order. First, within the "frequentist" setup, where priors are degenerate, we believe this definition offers a new formalization of what is commonly understood as external validity (or rather, lack thereof): a model that puts probability one to, say, $\bar{\theta}$ — is externally valid if $p_{\bar{\theta}}(\cdot) = p(\cdot|d,x)$ almost surely under $P(\cdot|d,x)$. Second, because this definition quantifies how far the true PDF is from the closest PDF within the subjective Bayesian model, it offers a way to quantify or rank external invalidity among models: **model o' is less externally invalid than model o for** (d,x)**, if** $u_{o'}(d,x) < u_o(d,x)$; i.e., as illustrated in the graph

---

[11] $supp(\mu_0^o(d,x))$ is the support of probability measure $\mu_0^o(d,x)$.

below, model o' is "closer" to the truth than model o.
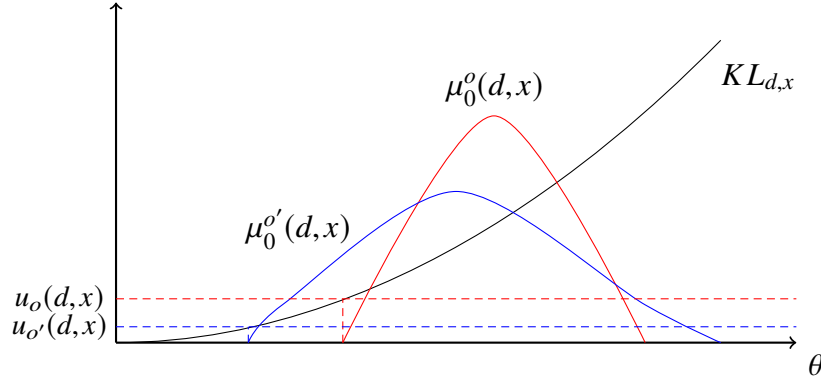
The next proposition provides a link between this ranking of external invalidity and our weights $(\alpha_t^o(d,x))_{o=0}^L$. It shows that, under some technical regularity assumptions, as t diverges, the weights are only positive for the least externally invalid models, provided (d,x) is played sufficiently often.[12]

Proposition 2.1. *Suppose* $\Theta \subseteq \mathbb{R}^{|\Theta|}$ *is compact and* $\theta \mapsto \log p_\theta$ *is continuous with* $\sup_{\theta \in \Theta} \log p_\theta$ *having a finite second moment. For any* $(d,x) \in \mathbb{D} \times \mathbb{X}$ *such that* $\inf_t t^{-1} \sum_{s=1}^t \Pr_{s-1}(D_s(x) = d) > 0$, *if model o is less externally valid than model o', then*

$$\frac{\alpha_t^o(d,x)}{\alpha_t^{o'}(d,x)} = o_{\mathbf{P}}(1)$$

Proof. See Appendix B.[13]   □

The proposition implies that if there exists an externally valid model among externally invalid models, the weight of the externally valid model will approach one as t diverges. This is why we interpret $\alpha_t^o(d,x)$ as a measure of external validity of our sources. It assigns higher probability — approaching 1 in the limit — to the source that is externally valid, and lower probability to sources that are externally invalid. Similarly, the proposition also suggests that our proposed method for aggregating the $L+1$ distinct subjective Bayesian models enjoys a "robustness property" in the sense that externally invalid models carry little weight and therefore have little influence on the PM's decisions.

---

[12]Lemma B.2 in the Appendix B provides a non-asymptotic version of this proposition.
[13]The probability measure $\mathbf{P}$ is formally defined in Appendix A.

The Policy Rule

The policy rule associated with this experiment defines the behavior of the PM. We define it as a sequence of two policy functions that, at each instance t, determine the probability of stopping the experiment and the probability of treatment for each $x \in \mathbb{X}$.

The first policy function, $(y^{t-1}, d^{t-1}) \mapsto \sigma_t(y^{t-1}, d^{t-1})(x) \in [0,1]$, specifies the probability of stopping the experiment for unit $x \in \mathbb{X}$ given the observed history y^{t-1}, d^{t-1}. The second policy function, $(y^{t-1}, d^{t-1}) \mapsto \delta_t(y^{t-1}, d^{t-1})(\cdot|x) \in \Delta(\mathbb{D})$ for each $x \in \mathbb{X}$, specifies the probability distribution over treatments for each $x \in \mathbb{X}$; i.e., $\delta_t(y^{t-1}, d^{t-1})(d|x)$ is the probability that $x \in \mathbb{X}$ receives treatment d given the past history. When there is no risk of confusion, we will omit the dependence on the history.

The policy rule defines two consecutive stages: a first stage of exploitation *and* exploration and a second stage of pure exploitation, in which the PM has stopped the experiment and has selected what she believes to be the best treatment. How the PM regulates the trade-off between exploitation and exploration in the first stage will be key for the results presented in Section 3. With this in mind, we now define a **structure of exploration for the policy rule** $(\delta_t)_t$ as two positive-valued sequences $(h_t, \omega_t)_t$ such that for any $(d,x) \in \mathbb{D} \times \mathbb{X}$ and any $t \geq 0$, $\omega_t(d,x) \in [0,1]$, $h_t(\cdot|x) \in \Delta(\mathbb{D})$, and

$$\mathbf{P}\left(t^{-1} \sum_{s=1}^{t} \delta_s(d|x) \geq h_t(d|x) \right) \geq 1 - \omega_t(d,x). \tag{2.2}$$

We call $(h_t)_t$ the **degree of exploration** of the policy rule and $(1 - \omega_t)_t$ the **likelihood of exploration** of the policy rule.

By providing a lower bound on the (average) propensity score, the structure of exploration quantifies the extent to which experimentation occurs under the policy rule $(\delta_t)_t$. It turns out that this structure is the *only* feature of the policy rule that matters for our performance criteria. We present these results formally in Section 3.[14]

2.2 The Gaussian Bayesian Learning Model

The general setup provides a useful conceptual framework to study experiments and policy recommendations. But at this level of generality, it becomes difficult to understand the dynamics of the

---

[14]The structure of exploration is not unique (e.g. $\omega_t = 0$ and $h_t = 0$ or $\omega_t = 1$ and $h_t = 0$), however, the results in Section 3 provide a criteria for ranking the different structures.

problem. It requires characterizing, for each t, the subjective PDF over the (d,x)-outcomes

$$\int p_\theta(\cdot)\mu_t^\alpha(d,x)(d\theta),$$

that the PM uses to form recommendations and decisions. This object, in turn, requires characterizing the process of the posterior beliefs $(\mu_t^\alpha(d,x))_{t=0}^T$, which is an infinite dimensional object.

Therefore, we focus on a setup wherein the PM is interested in the average effect of treatments

$$\theta(d,x) := E_{P(\cdot|d,x)}[Y(d,x)], \ \forall(d,x) \in \mathbb{D}\times\mathbb{X},$$

and takes subjective models within the Gaussian family. This, and the corresponding conjugate priors, imply that the posterior belief is fully characterized by a finite dimensional object, which is more tractable. We now present some definitions and discuss the scope of this Gaussian Bayesian learning model.

Even though this new setup is more restrictive than the original one, it is sufficiently general to encompass the canonical RCT setup for estimation of average treatment effects, even with the Gaussianity assumption. To see this, note that even if the PM's subjective model for potential outcomes is misspecified (i.e. she incorrectly assumes that $Y(d,x)$ is Gaussian) the PM can still accurately learn the true average effect because, for each (d,x), there always exists a θ such that $\theta = E_{P(\cdot|d,x)}[Y(d,x)]$. We show this is the case in Section 3.1.

Formally, the Gaussian learning model is constructed assuming that, for each $(d,x) \in \mathbb{D}\times\mathbb{X}$, $\mathcal{P}_{d,x}$ is a family of Gaussian PDFs given by $\{\phi(\cdot;\theta,1) : \theta \in \mathbb{R}\}$ and the prior for every source is also assumed to be Gaussian with mean $\zeta_0^o(d,x)$ and variance $1/v_0^o(d,x)$.[15] The quantity $v_0^o(d,x)$ can be interpreted as the number of units with characteristics x that were assigned treatment d in a past experiment. The higher the $v_0^o(d,x)$, the more certain source o is about $\phi(\cdot;\zeta_0^o(d,x),1)$ being the correct model. Throughout, we will assume $(\zeta_0^o,v_0^o)_{o=0}^L$ are non-random.

Given the observed data of past treatments and observed outcomes, at instance t the posterior belief is also Gaussian with mean and inverse of the variance given by the following recursion: For any

---

[15]Throughout, $\phi(\cdot;\theta,\sigma^2)$ is the Gaussian PDF with mean θ and variance σ^2.

$t \geq 1$,

$$\zeta_t^o(d,x) = \frac{1\{D_t(x) = d\}}{v_{t-1}^o(d,x) + 1\{D_t(x) = d\}} Y_t(d,x) + \frac{v_{t-1}^o(d,x)}{v_{t-1}^o(d,x) + 1\{D_t(x) = d\}} \zeta_{t-1}^o(d,x)$$

$$= \frac{J_t(d,x)}{f_t(d,x) + v_0^o(d,x)/t} + \frac{v_0^o(d,x)/t}{f_t(d,x) + v_0^o(d,x)/t} \zeta_0^o(d,x) \tag{2.3}$$

where

$$v_t^o(d,x) = N_t(d,x) + v_0^o(d,x), \quad f_t(d,x) := N_t(d,x)/t \tag{2.4}$$

$$\text{and } J_t(d,x) := t^{-1} \sum_{s=1}^{t} Y_s(d,x) 1\{D_s(x) = d\}. \tag{2.5}$$

From these expressions, we can see how Gaussianity simplifies the dynamics of the problem as we only need to analyze $(\zeta_t^o(d,x), v_t^o(d,x))_{t=0}^T$, a finite dimensional object, as opposed to $(\mu_t^o(d,x))_{t=0}^T$, an infinite dimensional object that is quite intractable.

At each instance t, the (subjective) average effect of treatment d for unit x is given by

$$\int y \left(\int_\Theta \phi(y; \theta, 1) \mu_t^\alpha(d,x)(d\theta) \right) dy;$$

i.e., the expected $Y(d,x)$ where the expectation is taking with respect to $\phi(.; \theta, 1)$ — the PDF describing the subjective model of the PM — where each parameter θ is weighted according to the posterior belief defined in Expression 2.1. By re-arranging the order of the integrals, it follows that

$$\int y \left(\int_\Theta \phi(y; \theta, 1) \mu_t^\alpha(d,x)(d\theta) \right) dy = \sum_{o=0}^{L} \alpha_t^o(d,x) \int y \phi(y; \zeta_t^o(d,x), 1/v_t^o(d,x)) dy$$

$$= \sum_{o=0}^{L} \alpha_t^o(d,x) \zeta_t^o(d,x) =: \zeta_t^\alpha(d,x).$$

Hence, the (subjective) average effect of treatment d on unit x at instance t is given by $\zeta_t^\alpha(d,x)$. Thus, the PM uses this quantity to assign treatment. In Section 3, we establish some finite sample properties of this quantity, such as the rate at which it concentrates around the true average effect.

Remark 2.1. *Our results and methodology extend to any subjective model whose posterior beliefs can be fully described by low finite-dimensional objects. For instance, in cases where $Y(d,x) \in \{0,1\}$, they extend to the Bernoulli-Beta model wherein the t instance posterior is given by a Beta density with parameters given by $(\sum_{s=1}^{t} 1\{D_s(x) = d\} Y_s(d,x) + v_0^o(d,x) \zeta_0^o(d,x), \sum_{s=1}^{t} 1\{D_s(x) = $*

$d\}(1 - Y_s(d,x)) + v_0^o(d,x)(1 - \zeta_0^o(d,x)))$. *More generally, our methodology can be extended to the entire exponential family — which includes the models considered here and more (see Schlaifer and Raiffa (1961) for examples), however, the interpretation of $\zeta_t(d,x)$ may change.* △

As discussed above, our definition of external invalidity allows us to distinguish between models that are more or less externally invalid. But it is silent about comparisons within equally externally (in)valid models. This is not an issue within a "frequentist setup", where the priors are degenerate, as any two externally valid models are identical.[16] In a Bayesian setup, however, there could be different degrees of external validity which are not captured by our general definition; for instance, if two models are both externally valid, but the prior of one is more concentrated around the true PDF than the other one. The next lemma makes progress on this issue for Gaussian subjective Bayesian models wherein the weights $(\alpha_t^o(d,x))_{o=0}^L$ get simplified.

Lemma 2.1. *For any $o \in \{0,...,L\}$, any $t \geq 1$, and any $(d,x) \in \mathbb{D} \times \mathbb{X}$,*

1.

$$\alpha_t^o(d,x) = \frac{\phi(J_t(d,x)/f_t(d,x) - \zeta_0^o(d,x); 0, (N_t(d,x) + v_0^o(d,x))/(N_t(d,x)v_0^o(d,x)))}{\sum_{o=0}^L \phi(J_t(d,x)/f_t(d,x) - \zeta_0^o(d,x); 0, (N_t(d,x) + v_0^o(d,x))/(N_t(d,x)v_0^o(d,x)))}.$$

2. $\lim_{|\zeta_0^o(d,x)-\theta(d,x)|\to\infty} \alpha_t^o(d,x) = 0$.

3. *If $\inf_t t^{-1} \sum_{s=1}^t \delta_s(d|x) > 0$, then* $\alpha_t^o(d,x) = \frac{\phi(\sqrt{v_0^o(d,x)}(\theta(d,x)-\zeta_0^o(d,x)); 0,1)\sqrt{v_0^o(d,x)}}{\sum_{o'=0}^L \phi(\sqrt{v_0^{o'}(d,x)}(\theta(d,x)-\zeta_0^{o'}(d,x)); 0,1)\sqrt{v_0^{o'}(d,x)}} + o_{\mathbf{P}}(1)$.

Proof. See Appendix C. □

Part (1) of the lemma characterizes $\alpha_t^o(d,x)$ as the odds ratio of Gaussian PDFs, which with probability approaching one are evaluated at $\theta(d,x) - \zeta_0^o(d,x)$ with mean 0 and variance $1/v_0^o(d,x)$. Part (2) shows that α_t^o offers certain robustness properties against external invalidity, in the sense that, if $\zeta_0^o(d,x)$ is "far away" from $\theta(d,x)$ then the associated weight of that model is approximately 0. Similarly, the weight will be higher for models with priors more concentrated around the true parameters. Finally, Part (3) of the lemma complements Proposition 2.1 for the Markov Gaussian learning model. It offers an asymptotic characterization of the weights when all the models are, according to our definition, equally externally invalid. It shows that not only will $\alpha_t^o(d,x)$ not equal 0 or 1 even with infinite data, but it also suggests a partial ordering among the sources. To see this, it is useful to introduce some nomenclature: We call $|\zeta_0^o(d,x) - \theta(d,x)|$ the **bias** of model o,

---

[16]By "identical" we mean that each model has a component, p^o and $p^{o'}$ that nullifies the KL divergence.

and $v_0^o(d,x)$, the **degree of conviction**. We call $|\zeta_0^o(d,x) - \theta(d,x)|\sqrt{v_0^o(d,x)}$, **the degree of stubbornness** of model o and a model with zero stubbornness and high conviction, **confident**. Part (3) indicates that, asymptotically, $\alpha_t^o(d,x)$ will put more weight on models that are less stubborn and more confident. In particular, if $v_0^o(d,x)$ diverges — intuitively, if all the prior sources have large sample sizes —, $\alpha_t^o(d,x)$ will concentrate around the least stubborn model.

Examples of Policy Rules. We conclude this section by presenting a series of examples of policy rules — and their associated exploration structure — in the context of the Gaussian learning framework.

Example 1 (Generalized ϵ-Greedy Policy Rule). *A commonly-used policy function that is admissible in our framework is the so-called Epsilon-Greedy policy rule, given by*

$$\delta_t(y^{t-1}, d^{t-1})(d|x) = (M+1)\,\epsilon\,\frac{1}{M+1} + (1 - (M+1)\,\epsilon)\mathbf{1}\{d = \arg\max_a \zeta_{t-1}^\alpha(a,x)\}, \ \forall t. \qquad (2.6)$$

That is, with probability $(M+1)\,\epsilon$, the treatment is assigned randomly, and with one minus this probability, the treatment assigned is the one with highest posterior mean.

A generalization of this policy rule is one where δ is Markov and yields "uniform exploration". Formally, for any past history (y^{t-1}, d^{t-1}),

$$\delta_t(y^{t-1}, d^{t-1})(\cdot|x) = \delta(\zeta_{t-1}, v_{t-1}, \alpha_{t-1})(\cdot|x), \ \forall x \in \mathbb{X},$$

where $\zeta_t := (\zeta_t^o)_{o=0}^L$ (the other variables are similarly defined), and

Assumption 2. *There exists an $\epsilon \in (0, 1/(M+1))$ such that $\delta(\cdot)(\cdot|x) \geq \epsilon$ for all $x \in \mathbb{X}$.*

Under this assumption, each treatment arm is chosen with positive probability, thus ensuring some experimentation.

It is straightforward to show that a structure of exploration for this class of policy rules is given by $h_t(d|x) = \epsilon$ and $\omega_t(d,x) = 0$. However, exploiting the Markov assumption, in the Supplemental Material L.2 we present an alternative — and perhaps more useful — structure of exploration, which is arbitrary close to $t^{-1}\sum_{s=1}^t E[\delta(\zeta_s, v_s, \alpha_s)(d|x)]$, and $\omega_t(d,x)$ converges faster than polynomial to zero. △

Example 2 (Optimal policy function). *The optimal policy function of this problem solves the Bellman equation problem with a per-period payoff given by the $\sum_{x\in\mathbb{X}}\zeta^\alpha(d,x)$ (or some other aggregator for x). Our framework does allow for such policy function but there are no guarantees that*

18

it will have a non-trivial exploration structure. Instead, one can consider a "perturbed" version of the form:[17]

$$\delta_t(d|x) = \frac{\exp\{h\Pi_t(\zeta_t,\nu_t,\alpha_t)(d,x)\}}{\sum_{d'=0}^{M}\exp\{h\Pi_t(\zeta_t,\nu_t,\alpha_t)(d',x)\}}, \ \forall (d,x) \in \mathbb{D} \times \mathbb{X}$$

where $\Pi_t(\zeta_t,\nu_t,\alpha_t)(d,x)$ is the instance t payoff of choosing treatment d for unit x given beliefs μ_t and weights α_t; $h > 0$ is a tuning parameter that governs the size of the perturbation. △

Example 3 (Thompson Sampling & refinements)**.** *Sampling schemes like Thompson's (*Thompson *(1933)) and others can be viewed as $\delta_t(d|x) = \pi_t(d|x)$ where $\pi_t(d|x)$ is a probability that treatment d yields the highest expected outcome and it is associated with the beliefs of the PM at time t, $(\zeta_{t-1},\nu_{t-1},\alpha_{t-1})$. For instance, in Thompson sampling $\pi_t(d|x)$ is constructed using the posterior beliefs $\mu_{t-1}^{\alpha}(d,x)$.*

For Thompson sampling, it is easy to show that Assumption 2 holds within the Markov Gaussian model and with $Y(d,x)$ having bounded support. Thus the exploration structure is such that $\epsilon :=\inf_t h_t(d|x) > 0$ and $\omega_t(d,x) = 0$. In the other cases, Assumption 2 may not hold if $\pi_t(d|x)$ fails to be uniformly bounded from below, but a non-trivial exploration structure can still be obtained exploiting the fact that the subjective probability has full support and that $Y(d,x)$ is bounded with high probability. We establish the formal result in the Supplemental Material L. △

Finally, we provide an example of the policy rule for stopping the experiment, σ. While most of our results do not require any restriction on this policy rule, this rule does govern the probability of making mistakes when stopping the experiment. As such, a desirable property for this rule is that, for a given tolerance level $\beta \in (0,1)$ chosen by the PM, the probability of making a mistake when stopping the experiment is no larger than β. Below, we propose a family of stopping rules for which it will be shown in Section 3.2 that, by appropriately choosing certain parameters, the rule has such desirable property.

Example 4 (Threshold Stopping Rule)**.** *For any $x \in \mathbb{X}$ and any positive-valued non-increasing sequence $(\gamma_t)_t$ and $B \in \mathbb{N}$, the stopping rule parameterized by $((\gamma_t)_t,B)$ is such that, for any $t \geq B$,*

$$\sigma_t(Y^{t-1},D^{t-1})(x) = 1, \ iff \ \max_d\{\min_{m \neq d} \zeta_{t-1}^{\alpha}(d,x) - \zeta_{t-1}^{\alpha}(m,x) - c_{t-1}(\gamma_{t-1},d,m,x)\} > 0,$$

---

[17]This idea of perturbing the optimal policy is by no means new; it is commonly used in economics and can be traced back to Harsanyi's trembling hand idea.

and if $t < B$, $\sigma_t(Y^{t-1}, D^{t-1})(x) = 0$, where, for any $d, m \in \{0, ..., M\}$ and any $o \in \{0, ..., L\}$,

$$c_t(\gamma_t, d, m, x) : c_t(\gamma_t, d, x) + c_t(\gamma_t, m, x) := \sum_{o=0}^{L} \gamma_t \frac{\alpha_t^o(d, x)}{f_t(d, x) + v_0^o(d, x)/t} + \sum_{o=0}^{L} \gamma_t \frac{\alpha_t^o(m, x)}{f_t(m, x) + v_0^o(m, x)/t}$$

where $f_t(d, x) := t^{-1} \sum_{s=1}^{t} 1\{D_s(x) = d\}$. While the expression for the cutoff is a bit involved, the constant γ_t is the key element – the other terms are convenient scaling factors. Loosely speaking, the proposition proposes to stop the experiment after B instances and as soon as the distance between the highest average posterior and the rest — measured by $\max_d \min_{m \neq d} (\zeta_t^\alpha(d, x) - \zeta_t^\alpha(m, x))$ — is far enough from zero, where "far enough" is essentially measured by $Constant \times \gamma_t$. This rule is akin to a test of two means wherein the hypothesis is rejected when the difference in means is above a multiple of the standard error. This intuition suggests that γ_t should be of order $1/\sqrt{t}$, however in this problem, as there are unknown quantities, the appropriate order is $\sqrt{\log t / t}$; see Section 3.2 for a more through discussion. $\triangle$

3 Analytical Results for the Gaussian Bayesian Learning Model

In this section, we derive analytical results for the Gaussian model presented in Section 2.2. But before we do so, a bit of housekeeping is required. Moving forward, we will omit x from the notation and derive our results for $|\mathbb{X}| = 1$. Given our assumptions, we can learn the fundamentals for each $x \in \mathbb{X}$ by treating them as separate and independent problems. Thus, we can extend all our results to the case of $|\mathbb{X}| > 1$ by treating the relevant quantities (e.g. $\theta(d)$, $Y(d)$, etc.) as vectors of dimension $|\mathbb{X}|$. Furthermore, to derive the results below we will need some assumptions on the (true) distribution of the potential outcomes,

Assumption 3. *There exists a $\upsilon < \infty$ such that for any $\lambda > 0$ and any $d \in \mathbb{D}$, $E[e^{\lambda(Y(d) - \theta(d))}] \leq e^{\upsilon \sigma(d)^2 \lambda^2}$ where $\sigma(d)^2 := Var(Y(d))$.*

This assumption imposes that $Y(d)$ is sub-gaussian, which, in effect, ensures that the probability $Y(d)$ takes large values decays at the same rate as the Normal does. Sub-gaussianity plays two roles in our results. First, it ensures that some higher moments, like the variance, exist. Second, and more importantly, it is used to derive how fast the average outcome concentrates around certain population quantities (see Lemma E.1 in the Appendix E). We could relax this assumption, but at the cost of getting slower concentration rates; see Remark E.1 in that appendix for more details.

Before presenting these results formally, it is useful to present our general approach for how we derived them. As we discussed in the previous section, a key of object of interest is $\zeta_t^\alpha = \sum_{o=0}^{L} \alpha_t^o \zeta_t^o$,

the subjective average effect of treatment at instance t. Most of our results hinge on understanding how this object concentrates around the true expected value θ.

For each treatment d, the randomness of $\zeta_t^\alpha(d)$ comes from two quantities: the frequency of play, $f_t(d) = t^{-1} \sum_{s=1}^t 1\{D_s = d\}$ and the treatment-outcome average, defined as $J_t(d) := t^{-1} \sum_{s=1}^t 1\{D_s = d\}Y_s(d)$. Hence, to derive the concentration rate of $\zeta_t^\alpha(d)$, we first need to understand how $f_t(d)$ and $J_t(d)$ concentrate. For $J_t(d)$, we can employ exponential inequalities for martingale differences (see Lemma E.1 in the Appendix E) to determine how fast the treatment-outcome average concentrates around $f_t(d)\theta(d)$. The case of $f_t(d)$ is a bit more nuanced because we care not only about how fast it concentrates around the average propensity score, $t^{-1} \sum_{s=1}^t \delta_s(d)$, but also about how far the average propensity score is from zero (i.e. the degree of exploration). Our structure of exploration enables us to separate the problem into two parts: we use exponential inequalities for martingale differences to determine the concentration rate and the structure of exploration to assess how far the average propensity score is from zero.

The next important step is to understand how the concentration rates of $f_t(d)$ and $J_t(d)$ translate into the concentration rate of $\zeta_t^\alpha(d)$ and how the parameters of the model and the exploration structure affect this rate. To do this, take any $\gamma, \eta > 0$ where γ and η quantify the concentration rate of $J_t(d)$ around $f_t(d)\theta(d)$ and $f_t(d)$ around $t^{-1} \sum_{s=1}^t \delta_s(d)$ resp. Given this, Lemma D.6 in the Appendix D.3 shows that, given an exploration structure $(h_t, \omega_t)_t$,

$$|\zeta_t^\alpha(d) - \theta(d)| \leq \Gamma(\gamma, h_t(d) - \eta, |\zeta_0(d) - \theta(d)|, v_0(d)) \tag{3.1}$$

where $h_t(d)$ is the degree of exploration and $\Gamma : \mathbb{R}_+ \times [0,1] \times \mathbb{R}^{L+1} \times \mathbb{N}^{L+1} \to \mathbb{R}$ is a function defined in Appendix D.3. Thus, Γ maps the concentration rate of $J_t(d)$ and $f_t(d)$ (given by γ and η respectively) to the concentration rate of the posterior mean around the true parameter. In fact, we show in Lemma D.5 in the Appendix D.3 that Γ is increasing in the first argument and decreasing in the second one, thereby implying that a faster concentration rate of $J_t(d)$ and $f_t(d)$ translate into a faster concentration rate of the posterior mean. Moreover, Γ also quantifies how a higher degree of exploration translates into a faster concentration rate, as well as how the source's parameters, $(\zeta_0(d), v_0(d))$, affect this rate.

3.1 Concentration bounds on the Posterior Mean

The next proposition establishes the rate at which the posterior mean concentrates around the true expected outcome.

Proposition 3.1. *For any $d \in \{0,...,M\}$, any $t \in \mathbb{N}$ and any $\varepsilon \geq 0$ such that $th_t(d)^2 \geq \varepsilon$,*

$$\mathbf{P}\left(|\zeta_t^\alpha(d) - \theta(d)| > \Gamma\left(\sqrt{\frac{2\nu\varepsilon}{h_t(d)^2 t}}\sigma(d), 0.5h_t(d), |\zeta_0(d) - \theta(d)|, v_0(d)\right)\right) \leq 4(e^{-\varepsilon} + \omega_t(d)).$$

Proof. See Appendix F. □

The intuition behind the proof relies on the arguments discussed above that explain how the concentration rate of the posterior mean depends on two factors: the concentration rates of the random quantities $J_t(d)$ and $f_t(d)$ and how these get distorted by the function Γ. More precisely, we show that by employing concentration inequalities for Martingale difference sequences (see Lemma E.1 in Appendix E), $J_t(d)$ and $f_t(d)$ are (up to constants) within $\gamma = \sqrt{\delta/t}$ and $\eta = h_t(d)\sqrt{\delta/t}$ of their respective population values with probability higher than $1 - 4e^{-\delta/h_t^2(d)}$ for any $\delta > 0$. To obtain the result, we simply plug these quantities into expression 3.1, while noting that $h_t(d) - \eta \geq 0.5h_t(d)$ for large enough t and that $\varepsilon = \delta/h_t^2(d)$.

Through the term Γ and the probability bound, the proposition illustrates the effect of the structure of exploration, $(h_t, \omega_t)_t$, on the concentration rates. In particular, Γ is of order $O\left(\frac{(th_t^2(d))^{-1/2}}{h_t(d) + t^{-1}}\right)$ (see Lemma D.5(3) in the Appendix D.3). For policy functions with $h_t(d) \geq \epsilon > 0$ (e.g. ϵ-greedy) the concentration rate is of order $t^{-1/2}$, but for those with $h_t(d) = o(1)$ then the concentration rate is slower than this rate and consistency of the posterior mean to the truth is only ensured if $\sqrt{t}h_t^2(d) \to \infty$.

Our method for aggregating multiple priors offers an attractive feature regarding our concentration rates. Sufficiently stubborn models, i.e. $|\zeta_0^o(d) - \theta(d)|\sqrt{v_0^o(d)}$ is sufficiently large, will have close to zero effect on the concentration rate of $\zeta_t^\alpha(d)$, as they are essentially dropped from the weighted average. This implies an *oracle* property in the sense that the concentration rate becomes arbitrary close to the least stubborn model, provided there is enough separation between the stubbornness of this model and the others. We formalize this property in the next corollary.

Corollary 3.1. *Take any (t,d,ε) as in Proposition 3.1. Furthermore, let model $o = 0$ denote the least stubborn model and suppose that for any given $\delta > 0$, there exists a C such that $\min_{o \neq 0} |\zeta_0^o(d) - \theta(d)|\sqrt{v_0^o(d)} \geq C$. Then,*

$$\mathbf{P}\left(|\zeta_t^\alpha(d) - \theta(d)| > \Omega\left(\sqrt{\frac{2\nu\varepsilon}{h_t(d)^2 t}}\sigma(d), 0.5h_t(d), |\zeta_0^o(d) - \theta(d)|, v_0^o(d)\right) + \delta\right) \leq 4(e^{-\varepsilon} + \omega_t(d))$$

Proof. See Appendix F. □

The function Ω, which is formally defined in Appendix D.2, acts as Γ but for one model; i.e., for any $o \in \{0,\ldots L\}$ and any $\gamma \geq 0$, assuming $J_t(d)$ and $f_t(d)$ are within γ of their population analogues,

$$|\zeta_t^o(d) - \theta(d)| \leq \Omega(\gamma, h_t(d) - \gamma, |\zeta_0^o(d) - \theta(d)|, v_0^o(d)).$$

Thus, the expression inside the probability in the corollary is in fact the concentration rate of the least stubborn model.

We summarize the implications of the previous proposition in the following remark and illustrate them numerically in Section 4.

Remark 3.1 (Properties of the Concentration Rate)**.** *The following properties are based on Lemma D.3 in Appendix D.2.*

1. *All else equal, the concentration rate decreases as the bias increases; it also decreases with the degree of stubbornness, i.e. $|\zeta_0^o(d) - \theta(d)|\sqrt{v_0^o(d)}$. The concentration rate is fastest when the bias is zero.*

2. *For confident models, the concentration rate increases with the degree of conviction, i.e. $v_0^o(d)$ increases. The intuition behind this result is as follows: If $v_0^o(d)$ increases but $|\zeta_0^o(d) - \theta(d)|\sqrt{v_0^o(d)}$ remains constant — equal to 0, in particular —, then necessarily, the model is becoming more convinced about a prior that is unbiased, thereby implying a faster convergence rate.*

3. *The effects of the degree of stubbornness and conviction on the concentration rate decrease as t increases.*

4. *An increase in the degree of the exploration, $h_t(d)$, improves the concentration rate. This comes from the fact that $h_t(d) \mapsto \Omega\left(\sqrt{\frac{2v\varepsilon}{h_t^2(d)t}}\sigma(d), 0.5h_t(d), |\zeta_0^o(d) - \theta(d)|, v_0^o(d)\right)$ is decreasing (see Lemma D.3 in the Appendix D.2). Intuitively, increasing $h_t(d)$ implies having more observations to estimate $\theta(d)$ — "more information" about treatment d implies a faster concentration rate.*

△

3.2 Probability of making a mistake

In this section, we provide bounds on the probability of making a mistake when following the stopping rule proposed in Example 4. Suppose treatment M has the largest expected effect, i.e., $\Delta := \theta(M) - \max_{d \neq M} \theta(d) > 0$. We define a mistake as recommending a treatment different than M at the instance t in which the experiment was stopped. Because recommendations are based on the PM's posteriors, we can express a mistake as

$$\max_{d \neq M} \zeta_\tau^\alpha(d) - \zeta_\tau^\alpha(M) > 0,$$

where τ indicates when the experiment is stopped, i.e., is the first instance after B such that $\max_d \min_{m \neq d} \{\zeta_t^\alpha(d) - \zeta_t^\alpha(m) - c_t(\gamma_t, d, m)\} > 0$ where the cutoffs c_t are defined in Example 4.

The following proposition provides an upper bound for the probability of making a mistake associated with this stopping rule, when all sources are unbiased or are biased "in the right direction".[18]

Proposition 3.2. *Suppose for each $d \in \{0, ..., M-1\}$, $\zeta_0(d) \leq \theta(d)$ and $\zeta_0(M) \geq \theta(M)$. Consider the stopping rule defined in Example 4 with parameters $((\gamma_t)_t, B)$ then for any $t \geq 1$,*

$$\mathbf{P}\left(\max_{d \neq M}\{\zeta_\tau^\alpha(d) - \zeta_\tau^\alpha(M)\} > 0\right) \leq 2 \sum_{d=0}^{M} \sum_{t=B}^{T} e^{-0.5t \frac{\gamma_t^2}{\nu\sigma(d)^2}}. \tag{3.2}$$

Proof. See Appendix G. □

The intuition behind this proposition is as follows. Mistakes occur when, at some instance t greater than B, the posterior mean of some treatment d — different from M — is "much larger" than the others. This implies that the posterior has to be "much larger" than its population mean, $\theta(d)$, where "much larger" depends on the pre-specified cutoff γ_t. Hence, given our assumption on the priors, the probability of a mistake is essentially given by the probability that, for some instance after B, the outcome-treatment average exceeds its population value by an amount given by γ_t. Lemma E.1 in Appendix E provides the bound of $e^{-0.5t \frac{\gamma_t^2}{\nu\sigma(d)^2}}$. Finally, since the stopping time is random and can ocurr at any instance after B, we sum over all possible such values of t.

In Proposition 3.2, we assumed unbiased sources or that the priors ranked treatment M as the highest. In the next collorary, we can prove that when some sources are biased, there still exists an oracle property akin to the one demonstrated for the concentration rates. In particular, we show that

---

[18]By "in the right direction" we mean the priors rank treatment M as the highest one. For the general case where sources can be biased (in any direction), see Lemma G.1 in the Appendix G.

upper bound is arbitrary close to the unbiased source, provided the other sources are sufficiently biased.[19]

Corollary 3.2. *Let $o = 0$ denoted the unbiased source. There exists a C such that, if $\min_{o \neq 0} |\zeta_0^o(.) - \theta(.)| \geq C$ and $\zeta_0^0(.) = \theta(.)$, then*

$$\mathbf{P}\left(\max_{d \neq M}\{\zeta_\tau^\alpha(d) - \zeta_\tau^\alpha(M)\} > 0\right) \leq \sum_{d=0}^{M}\sum_{t=B}^{T} 2e^{-0.5t\frac{(\gamma_t)^2}{v\sigma(d)^2}}.$$

Proof. See Appendix G. □

Proposition 3.2 also reveals how by properly choosing $((\gamma_t)_t, B)$, the probability of a mistake associated with the stopping rule is bounded by β, where $\beta \in (0,1)$ is any tolerance level. The next corollary presents such result.[20]

Corollary 3.3. *Suppose all the conditions of Proposition 3.2 hold, and, for any t, $\gamma_t \geq \sqrt{\frac{\log t}{t}}\sqrt{A\log t}$ with (A, B) such that $\log B \geq \max_d 2v\sigma(d)^2$, and*

$$\frac{3(M+1)}{A-1}(B^{-(A-1)} - T^{-(A-1)}) \leq \beta. \tag{3.3}$$

Then

$$\mathbf{P}\left(\max_{d \neq M}\{\zeta_\tau^\alpha(d) - \zeta_\tau^\alpha(M)\} > 0\right) \leq \beta.$$

Proof. See Appendix G. □

The choice of $(\gamma_t)_t$ and the extra restrictions in B are so that both terms in the upper bound in Proposition 3.2 are less than t^{-A}. By simple arguments, one can show that $\sum_{t=B}^{T} t^{-A} \leq \frac{1}{A-1}(B^{-(A-1)} - T^{-(A-1)})$ and so expression G.4 ensures the desired result.

The sequence $(\gamma_t)_t$ has to decay, at most, at $\log t/\sqrt{t}$ rate. Compared to the $1/\sqrt{t}$ rate that arises in the canonical difference of means test in statistics, we lose a factor of $\log t$. This factor acts as an upper bound for the unknown population quantities. If one knew or could estimate these quantities — the same way one estimates the standard deviations in the difference in means test — one could lose this extra $\log t$ factor.

---

[19]A more general statement that relaxes the unbiased assumption of source $o = 0$ is proven in Lemma G.2 in Appendix G.

[20]For the general result allowing for biased sources, please see Lemma G.3 in Appendix G.

Finally, we note that the sequence $(\gamma_t)_t$ can decay much slower than $\log t / \sqrt{t}$ — in fact, it may not decay at all. However, large values of γ are undesirable because, the larger the γ, the less likely it is to stop the experiment at any instance thereby implying longer — and more costly — experiments. We therefore recommend to set $\gamma_t = O\left(\frac{\log t}{\sqrt{t}}\right)$.

3.3 Average Observed Outcomes

In this section, we characterize the behavior of the average outcome $t^{-1} \sum_{s=1}^{t} Y_s$. By Lemma E.1 in Appendix E, $t^{-1} \sum_{s=1}^{t} Y_s$ will concentrate around a weighted average of $\theta(\cdot)$, with the time average of the propensity score as weights, i.e.,

$$t^{-1} \sum_{s=1}^{t} \sum_{d=0}^{M} \theta(d)\delta_s(d).$$

Without further knowledge of $(\delta_t)_t$, it is nearly impossible to characterize this average any further. However, for generalized ϵ-*greedy* policy functions, indexed by a non-random sequence $\Xi := (\Xi_t)_t$:

$$\delta_t(d) = \Xi_t(M+1)^{-1} + (1 - \Xi_t)\mathbb{1}\{d = \arg\max_a \zeta_{t-1}^{\alpha}(a)\}, \ \forall t \in \{1,...,T\},$$

we can establish the following proposition for unbiased sources (the general result for biased sources can be found in Lemma H.2 in the Appendix H).

Proposition 3.3. *Suppose all sources are unbiased. For any $\gamma > 0$ and any $t \in \{1,...,T\}$*

$$\mathbf{P}\left(\max_d \theta(d) - t^{-1} \sum_{s=1}^{t} Y_s > -\mathcal{S}(t) - \mathcal{E}(t,\gamma,\bar{\Xi}_t) - \mathcal{B}(\bar{\Xi}_t)\right) \leq 5e^{-\gamma}.$$

where

$$\mathcal{S}(t,\gamma) := \sqrt{\frac{\gamma}{t}}\left(\sqrt{2\upsilon}\sigma(d) + \frac{||\theta||_1}{2}\right) \ and \ \mathcal{E}(t,\gamma,\bar{\Xi}_t) := 2||\theta||_1\sqrt{1 - \bar{\Xi}_t}\sqrt{e^{\gamma}\sum_{d=0}^{M} t^{-1}\sum_{s=1}^{t} \omega_{s-1}(d)}$$

$$with \ \mathcal{B}(\bar{\Xi}_t) := ||\theta||_1 \frac{\bar{\Xi}_t}{M+1}, \ with \ \bar{\Xi}_t := t^{-1} \sum_{s=1}^{t} \Xi_s.$$

Proof. See Appendix H. □

Despite the length of the proposition, its parts are quite intuitive. The term $\mathcal{S}(t,\gamma)$ controls the

26

stochastic error that arises from the difference between $t^{-1} \sum_{s=1}^{t} Y_s = \sum_{d=0}^{M} t^{-1} \sum_{s=1}^{t} Y_s(d) 1\{D_s = d\}$ and its conditional expectation $\sum_{d=0}^{M} t^{-1} \sum_{s=1}^{t} \theta(d) \delta_s(d)$. This term is essentially of order $O(\sqrt{\gamma/t})$. The term $\mathcal{E}(t, \gamma, \bar{\Xi}_t)$ arises from choosing the wrong treatment in the "exploitation" part because the policy function depends on ζ^{α} and not θ. It is decreasing on the quantity $\bar{\Xi}_t$, which regulates the trade-off between exploitation and exploration and can be viewed as the degree of exploration. A higher degree of exploration will result on more information about the treatment and in turn a smaller likelihood of choosing the wrong treatment. Finally, the term $\mathcal{B}(\bar{\Xi}_t)$ is a non-random bias that stems from the "exploration" part of the policy function: With probability $\bar{\Xi}_t (M + 1)$ the treatment is chosen at random, producing $\sum_{d=0}^{M} \theta(d)/(M + 1)$.

If $\bar{\Xi}_t = t^{-1} \sum_{s=1}^{t} \sum_{d=0}^{M} \omega_{s-1}(d) = o(1)$, i.e., if the exploration part of the policy function vanishes, then γ can be chosen to diverge with t, however slowly, and so, with probability approaching one, $t^{-1} \sum_{s=1}^{t} Y_s$ converges to $\max_d \theta(d)$.

The term $\mathcal{E}(t, \gamma, \bar{\Xi}_t) + \mathcal{B}(\bar{\Xi}_t)$ illustrates the so-called "exploration vs. exploitation" tradeoff and how it is regulated by $\bar{\Xi}_t$. This tradeoff suggests a choice for $\bar{\Xi}_t$ that balances $\mathcal{B}(\bar{\Xi}_t)$ and $\mathcal{E}(t, \gamma, \bar{\Xi}_t)$. Unfortunately, such a choice is infeasible as both terms depend on unknown quantities. Nevertheless, we can conclude that $\Xi = 1$ — the choice used in RCTs — will typically not be optimal. In fact, as t increases, the "optimal" Ξ_t will decrease to 0, favoring "exploitation" to "exploration". We explore the choice of Ξ further, when we simulate our model in the next section.

4 Model Simulations

In this section, we present Monte Carlo simulations of our model using the generalized ϵ-Greedy policy rule presented in Section 2.1. The purpose of these simulations is to highlight different aspects of our analytical results and to provide a sense of the tightness of our analytic bounds. We consider the case with only two treatment arms, $D \in \{0,1\}$, and assume that $Y(0) \sim N(1,1)$ and $Y(1) \sim N(1.3, 1)$. We assess the performance of our model according to the three outcomes outlined in Section 3: concentrations bounds, probability of making a mistake, and average earnings. We simulate each experiment 1000 times, with each experiment lasting at most 1000 instances.

Multiple Priors, External Validity, Robustness

We begin by illustrating how our setup weights the different models over the course of the experiment. Recall that to aggregate across several distinct subjective Bayesian models, our setup will average the posterior beliefs of each model using as weights, $\alpha_t^o(d)$ – the posterior probability that

model o best fits the observed data within the class of models being considered. We demonstrated in Proposition 2.1 for the general case, and Lemma 2.1 for Gaussianity, that if there exists an externally valid model among externally *invalid* models, then $\alpha_t^o(d)$ will approach one for the externally valid model. Conversely, $\alpha_t^o(d)$ will approach zero if models are far from the true $\theta(d)$.

To illustrate this property, we simulate our model under different sets of priors. For each simulation, we assume that our policymaker has two sets of priors about the potential outcomes distributions. One is her initial set of priors, which we will assume are correct (i.e. $\zeta_o^o = \theta$) but diffuse (i.e. v=1). For the other set of priors, we consider four alternative scenarios varying in their degree of stubbornness.

In Figure 1, we plot $\alpha_t^o(d)$ corresponding to the second set of priors over the course of the experiment. The graph on the left is for the $d = 0$ arm and the one on the right is for the $d = 1$ arm. Each line corresponds to a different set of priors, and the lighter the line, the more stubborn the prior. Starting with the top and darkest line, we see that $\alpha_t^o(d)$ increases over time putting more and more weight on an externally valid model. By the end of the experiment, $\alpha_t^o(d)$ is close to 95% for both arms. As we consider more stubborn models, we can see that the corresponding $\alpha_t^o(d)$ becomes smaller. So much so that for extremely stubborn models (i.e. the lowest line) $\alpha_t^o(d)$ becomes essentially zero by the 600th instance. This is why we interpret the parameter $\alpha_t^o(d)$ as a measure of external validity: the more externally valid the model, the higher the corresponding $\alpha_t^o(d)$.

An important feature of how we aggregate across models is that it generates a robustness property. Because $\alpha_t^o(d)$ will place less weight on models that are not externally valid, over time they will have limited influence on the PM's beliefs and consequent decisions. We illustrate this Figure 2. In the top graphs, we plot the policymaker's posterior beliefs about the mean of the potential outcome distributions over time. The plot distinguishes between three posterior means. The bottom (dashed) line corresponds to one set of priors, which we assume to be unbiased (i.e. $\zeta_o^o = \theta$), but diffuse. The top (dash-dotted) line refers to an alternative set of priors, which contains some degree of stubbornness (i.e. $\zeta_o^o = \theta + .5, v = 250$). The middle (solid) line comes from the combined model, which is a weighted average of the two sets of priors using $\alpha_t^o(d)$ as weights. We see that even though our policymaker starts with a stubborn prior, the combined model converges relatively quickly to the non-stubborn model. This is the result of both the oracle property – concentrating on the least stubborn model – and robustness property – putting less and less weight on sufficiently stubborn models.

In the bottom graphs, we consider the case in which the alternative model is confident. Thus, both sets of priors are unbiased; the alternative prior simply comes with a higher degree of conviction.

28

Because both priors are correct, the combined model does not immediate converge to one of the models as we saw in case with stubborn priors. As we started in Lemma 2.1, our parameter α is more responsive to bias than conviction.

Concentration Bounds

Effects of ϵ. We now simulate our model's concentration bounds and some its key properties. Recall from Remark 3.1 in Section 3.1, the concentration rate increases with the parameter ϵ. We demonstrate this property in the top panel of Figure 3, in which we plot concentration bounds for three different values of $\epsilon \in \{0.1, 0.5, 0.9\}$. That is, for a given ϵ, we compute the difference over time between the policymaker's posterior belief of the true mean, $\zeta_t^o(d)$, and the true mean, $\theta(d)$. We then plot the probability that these differences are greater than 0.1. For these simulations, we assume that our policymaker has correct, but diffuse priors (i.e. $\zeta_0^o = \theta$ and $v_0^o = [1,1]$).

In the top panel, we see that except for early on, our concentration bounds decrease over time and in the case of $\zeta_t^o(0)$ decrease faster, the higher the ϵ. For instance, after 1000 instances, $Pr(\zeta_t^o(0) - \theta(0) > 0.1)$ is almost zero for the case of $\epsilon = 0.9$, but is still close to 0.5 for $\epsilon = 0.1$. For the other treatment arm, the patterns are reversed. All three lines decrease relatively quickly, with the lower ϵ lines decreasing faster.

The intuition for these patterns is straightforward and speaks to the point about frequency of play in Remark 3.1. When the PM selects a treatment arm, she will only learn about the distribution of potential outcomes for that arm. As she become more confident in which arm is better, she will play the other arm only when forced to by the ϵ-greedy algorithm. In this case, the higher the ϵ the more the PM will be forced to play treatment $d = 0$ and the more she learns about $\theta(0)$. We can see this clearly in the bottom panel, which depicts the cumulative number of times the treatment has been played over time by different values of ϵ's. As we compare the two panels, the more we play a particular arm, the more we learn about it, and the sooner our beliefs converge to the truth.

Effects of Priors. In Figure 4, we investigate the effects of different priors on the concentration bounds. In particular, we plot different concentration bounds for priors with different degrees of stubbornness and confidence. For example, in the bottom two lines, we consider two unbiased priors, but with different levels of confidence. According to Remark 3.1, concentration rates increase as the degree of conviction increases and this precisely what we see. It is also the case, that the concentration rate decreases faster with less stubborn models. We can see this pattern clearly by comparing the top two lines. By comparing the two middle lines, we can also see that conditional

on the degree of stubbornness, the higher the bias, the slower the concentration rate. Lastly, as before, the concentration rates for $\theta(1)$ tend to be faster than those for $\theta(0)$ because of the frequency of play.

Probability of Making a Mistake

In Section 3.2, we defined a mistake as recommending a treatment arm different from the one that yields the largest expected effect at the instance in which the experiment was stopped. In Figure 5, we plot the average stopping period (left axis) and the probability of making a mistake at that stopping period (right axis) by ϵ. It is clear from the graph that the more we experiment across treatment arms (i.e., higher ϵ), the faster we stop the experiment. This makes sense. As we experiment more, the data become more IID and we are able to better learn the true means of the potential outcome distributions. According to these simulations, the degree of experimentation does not have to be particularly high. Even though at low levels of ϵ the experiment lasts for almost its entire duration, the drop off is fairly quick. Once ϵ is greater than 0.5, the difference gained in stopping periods from additional experimentation is minimal.

Shorter stopping periods do not come at the cost of making more mistakes. This result is to some extent an artifact of our stopping rule, whose parameters control the probability of type I errors. As the graph depicts, the probability of making a mistake varies little with ϵ and is always below 1%.

In Figure 6, we explore how the initial priors affect the probability of making a mistake. We again consider two sets of priors, both with $v_0 = [250,250]$. One, however, is confident with $\zeta_0^o = \theta$, whereas the other is stubborn, with $\zeta_0^o = [\theta(0)+\delta, \theta(1)-\delta]$, where δ is indicated by a point on the x-axis. For $\delta \in (0,0.15)$, the priors are biased, but have a proper ranking of the treatment arms. For $\delta > 0.15$, the priors are not only biased, but reverse the ranking of the arms. On the y-axis, we plot the probability of making a mistake associated with each set of priors, as well as for the combined model.

We can see that for $\delta \in (0,0.15)$, the probability of making mistake is small, less than 1%, for all three models. But once $\delta > 0.15$, and the ranking of treatment arms are reversed, the probability of making a mistake for the stubborn model increases significantly and approaches 1 by $\delta \geq 0.3$. Importantly, the probability of making a mistake for the combined model mirrors the one for the confident model, which again illustrates the robustness property of $\alpha_t^o(d)$.

Expected Earnings

The final outcome we evaluate is expected earnings. According to Proposition 3.3, the distance between the average outcomes and maximum expected outcome is decreasing in ϵ. In Figure 7, we plot by ϵ, the difference between the policymaker's average impact and the maximum expected outcome, $\max_d \theta(d)$, for an experiment that lasts 1000 instances. The figure also distinguishes between our two familiar sets of priors, a confident one and a stubborn one.

Two important observations emerge from this figure. First, there is a steep negative monotonic relationship between expected earning and ϵ. In fact, the 10% quantile of the average earnings distribution for $\epsilon = 0.10$ lies above the 90% quantile of the average earnings distribution for $\epsilon = 0.90$. Second, if we compare across the two plots, we can see that starting off with a stubborn prior affects average earnings, but only minimally. Again, this result is a product of the robustness property that our model aggregation approach provides.

The fact that average earnings declines with experimentation does not imply that our policymaker should set ϵ close to zero. Because as we saw in Figure 5, lower ϵ's result in longer experiments, which can come with costs. Moreover, as we show in Proposition 3.2, the upper bound the probability of making a mistake is weakly smaller for higher levels of ϵ. Thus, to properly capture the experimentation versus exploitation tradeoff inherent in multi-armed bandit problems, we need to specify a payoff function.

We consider the following payoff function:

$$\Pi_{\beta,c}^I = \sum_{d=0}^{M} \sum_{t=0}^{T^*} \beta^t 1\{D_t = d\}(Y_t(d) - c_1) + \sum_{t=T^*+1}^{\infty} \beta^t 1\{D_{T^*} = d\}(\theta(d) - c_2) \tag{4.1}$$

$$= \sum_{d=0}^{M} \sum_{t=0}^{T^*} \beta^t 1\{D_t = d\}(Y_t(d) - c_1) + \frac{\beta^{T^*+1}}{1-\beta} 1\{D_{T^*} = d\}(\theta(d) - c_2) \tag{4.2}$$

where c_1 indicates the costs of running the experiment, c_2 cost of administering the treatment, β^t represents a discount factor, and T^* denotes the stopping period. This payoff function comprises of two parts. The first part is the earnings during the experiment net of cost. The second part captures the expected future benefits under the chosen treatment, net of cost.

In Figure 8, we compute the payoff function for our model simulations by different values of ϵ. In contrast with the previous figure, we see that the average payoffs are increasing with ϵ until approximately $\epsilon = 0.38$, at which point the payoffs start to decline. While this "optimal" value of ϵ is clearly a function of an arbitrary set of parameter choices, our conjecture is that the inverted u-shape

31

relationship is likely to hold more generally, suggesting that some combination of experimentation and exploitation is optimal.

5   Charitable Fundraising Experiment

In this section, we present a real-world numerical example to show that by incorporating multiple priors, our policymaker can stop the experiment sooner without significantly increasing the probability of making a mistake. This results in large performance gains relative to a standard RCT.

Our numerical example uses data from a direct mail fundraising experiment reported in Karlan and List (2020). The experiment, which we will refer to as the BMGF experiment, consisted of sending 51,971 solicitation letters to previous donors of a charity focused on international development and poverty reduction. Donors were randomly assigned to receive letters with or without information about a \$2:\$1 limited-time matching grant offered by the Bill and Melinda Gates Foundation. Letters were mailed in December 2009 and responses were tracked until March 2011.

The authors find that a matching grant from the Bill and Melinda Gates Foundation was effective at increasing donations. Over the course of the experiment, the treatment increased the total unconditional amount given by \$0.36 relative to a control mean of \$0.26, an increase of 38%.

We use the data from this experiment to conduct a series of Monte Carlo simulations. We chose this experiment for two reasons. First, charitable giving is an outcome that responds relatively quickly to treatment: conditional on donating, letter recipients will typically respond within a month. Second, and more importantly for our setup, similar experiments have been conducted in various settings, even by the same authors. Thus, we can use the results from these other charitable fundraising experiments as priors when simulating the BMGF experiment.

To run our simulations, we sample from the empirical distributions of the outcomes for the treatment and control groups. We focus on the log amount given during the experiment conditional on donating.[21] In the treatment group, 225 individuals gave a donation, at an average log amount of 3.56. In the control group, 121 individuals donated, resulting in an average log amount of 3.34. Given these two empirical distributions, we simulate our model 1000 times for 600 instances, which is the minimum number of observations needed for the average difference between treatment and control to become statistically significant.

---

[21]We focus on the amount given conditional on donating because only a small fraction of people donate and for computational reasons, we wanted to avoid running the Monte Carlos for hundreds of thousands of instances.

Our simulations incorporate five sets of priors. The first two sets of priors come from an experiment that Karlan and List (2020) ran in conjunction with the BMGF experiment, but for a different population of donors. The experimental arms were also different in that both treatment and control were offered a $2:$1 limited-time matching grant. The treatment group, however, was told that the Bill and Melinda Gates Foundation was the matching donor, whereas this information was kept anonymous for the control group. This treatment led to 16.6% (robust standard error = 0.099) increase in donations conditional on giving. We generate two different priors from this experiment that differ in their level of confidence. We set the first prior to $\zeta_0 = [3.11, 3.27], v_0 = [284, 223]$, where the v_0 reflects the number of treated in each arm. We set the second prior to $\zeta_0 = [3.11, 3.27], v_0 = [0.01, 0.01]$, which makes the prior diffuse. By introducing a diffuse version of this prior, we allow the policymaker to reject priors that are overly stubborn sooner.

The second set of priors comes from another fundraising experiment that the authors conducted in 2005 (Karlan and List, 2007). Similar to the BMGF experiment, this experiment also offered, as one of its treatment, a $2:$1 limited-time matching grant relative to a no-matching grant control. However, both the organization requesting the donation and the pool of donors were likely quite different. In this case, the treatment effect only led to 1.5% increase in donations conditional on giving. As before, we generate two other sets of priors from this experiment that again only differ in their level of confidence. We set the first prior to $\zeta_0 = [3.42, 3.44], v_0 = [782, 252]$, where the v_0 reflects the number of treated in each arm. We set the second prior to $\zeta_0 = [3.42, 3.44], v_0 = [0.01, 0.01]$, which makes the prior diffuse. The last set of priors is completely diffuse. We set $\zeta_0 = [0, 0], v_0 = [0.01, 0.01]$.

We present the results of the simulation in Table 1. In column 1, we present the results from simulating the RCT for 600 periods. In columns 2-5, we display the results from simulating our model with multiple priors. We present our model for several values of ϵ to assess the robustness of the findings to different degrees of exploration. In columns 6-10, we again report simulation results for different values of ϵ, but for a model that does not incorporate the use of priors.

The average stopping period of our multi-prior model is less than 200 periods, across each ϵ. This is much lower than the stopping period of the standard RCT (by construction), as well as of the non multi-prior models, which average around 525 instances across the various ϵ. Importantly, the shorter stopping periods of our multi-prior models do not come with a substantial increase in the probability of making a mistake, which is less than 3 percent across the different ϵ.

One of the main advantages of our model, relative to the RCT, can be seen when comparing the

average cumulative payoff distributions. We define the cumulative payoff as the sum of donations received at each instance over the 600 periods. At each instance, the policymaker receives a donation amount depending on which treatment was selected. In cases in which the experiment was stopped, the policymaker received the average payoff of the potential outcomes distribution corresponding to her selected treatment.

When comparing cumulative payoffs, our model outperforms the RCT by a wide margin. The 90th percentile of the payoff distribution of the RCT is lower than the even the 10th percentile of our multi-prior model. In fact, the maximum the policymaker could achieve (should she always choose treatment) is 2136 on average. Our framework achieves 99% of this maximum.

Our multi-prior models also outperform the models without priors, but only minimally. This is not too surprising since these models are also adaptive and thus, engages in a fair amount of exploitation. Nevertheless, it is worth highlighting that we are only comparing payoffs. If there were important costs associated with running longer experiments (as is usually the case) then in terms of net payoffs, our multi-prior model would outperform the models without priors to a much larger extent.

In Table 2, we report our measure of external validity associated with each of the initial priors for the model with $\epsilon = 0.3$.[22] On the basis of average outcomes, the least biased model is the experiment from Karlan and List (2007). Our measure assigns a weight of 0.646 for θ_0 and 0.869 for θ_1. The α's are more concentrated for θ_1 because that treatment arms has been played more often.

In sum, these simulations provide a proof-of-concept for our model. By incorporating multiple priors, our policymaker not only stopped the experiment sooner, but could do so without risk of making a mistake. When compared to the RCT, this translated into larger performance gains in terms of payoffs.

6   Conclusions

This paper presents a conceptual framework for how to incorporate prior sources of information into the design of a sequential experiment. An obvious issue is how to handle the potential lack of external validity of each of these sources. We address this issue by first presenting a formal definition of external validity that can be used to differentiate sources with different degrees of external invalidity and second, by showing that our framework is robust to including externally-invalid sources. This last property relaxes the burden on the policymaker of having to correctly choose

---

[22]Our choice of $\epsilon = 0.3$ was for the sake of parsimony. The results are qualitatively similar for other values of ϵs.

relevant sources of information based on limited ex-ante information. As "stubborn" sources are harder to discard, we believe it is useful to incorporate many priors, including versions that are diffuse.

For the common problem of learning about average treatment effects, we show that our framework offers several nice properties. As we illustrated for the case of charitable giving, these properties translate into substantial gains in performance — such as reducing the duration of experiment and increasing the average payoffs while keeping an acceptable probability of making a mistake — over both standard RCTs and adaptive experiments.

References

Agrawal, S. and Goyal, N. (2017). Near-optimal regret bounds for thompson sampling. *J. ACM*, 64(5).

Angrist, J. D. and Fernández-Val, I. (2013). *ExtrapoLATE-ing: External Validity and Overidentification in the LATE Framework*, volume 3 of *Econometric Society Monographs*, pages 401–434. Cambridge University Press.

Athey, S. and Imbens, G. (2019). Machine learning methods economists should know about.

Banerjee, A., Duflo, E., Goldberg, N., Karlan, D., Osei, R., Parienté, W., Shapiro, J., Thuysbaert, B., and Udry, C. (2015a). A multifaceted program causes lasting progress for the very poor: Evidence from six countries. *Science*, 348(6236). Publisher Copyright: © 2015 by the American Association for the Advancement of Science; all rights reserved.

Banerjee, A., Karlan, D., and Zinman, J. (2015b). Six randomized evaluations of microcredit: Introduction and further steps. *American Economic Journal: Applied Economics*, 7(1):1–21.

Bisbee, J., Dehejia, R., Pop-Eleches, C., and Samii, C. (2017). Local instruments, global extrapolation: External validity of the labor supply–fertility local average treatment effect. *Journal of Labor Economics*, 35(S1):S99–S147.

Buchanan, A. L., Hudgens, M. G., Cole, S. R., Mollan, K. R., Sax, P. E., Daar, E. S., Adimora, A. A., Eron, J. J., and Mugavero, M. J. (2018). Generalizing evidence from randomized trials using inverse probability of sampling weights. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(4):1193–1209.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge university press.

Chabrier, J., Cohodes, S., and Oreopoulos, P. (2016). What can we learn from charter school lotteries? *Journal of Economic Perspectives*, 30(3):57–84.

Chernoff, H. (1959). Sequential design of experiments. *The Annals of Mathematical Statistics*, 30(3):755–770.

Dehejia, R., Pop-Eleches, C., and Samii, C. (2021). From local to global: External validity in a fertility natural experiment. *Journal of Business & Economic Statistics*, 39(1):217–243.

DellaVigna, S., Otis, N., and Vivalt, E. (2020). Forecasting the results of experiments: Piloting an elicitation strategy. *AEA Papers and Proceedings*, 110:75–79.

DellaVigna, S. and Pope, D. (2018). Predicting experimental results: Who knows what? *Journal of Political Economy*, 126(6):2410–2456.

Dimakopoulou, M., Zhou, Z., Athey, S., and Imbens, G. (2017). Estimation considerations in contextual bandits. *arXiv preprint arXiv:1711.07077*.

Epstein, L. G. and Schneider, M. (2003). Recursive multiple-priors. *Journal of Economic Theory*, 113(1):1–31.

Gelman, A. and Carlin, J. (2014). Beyond power calculations: Assessing type s (sign) and type m (magnitude) errors. *Perspectives on Psychological Science*, 9(6):641–651. PMID: 26186114.

Gelman, A. and Pardoe, I. (2006). Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. *Technometrics*, 48(2):241–251.

Imai, K. and Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443 – 470.

Joseph Hotz, V., Imbens, G. W., and Mortimer, J. H. (2005). Predicting the efficacy of future training programs using past experiences at other locations. *Journal of Econometrics*, 125(1-2):241–270.

Karlan, D. and List, J. A. (2007). Does price matter in charitable giving? evidence from a large-scale natural field experiment. *American Economic Review*, 97(5):1774–1793.

Karlan, D. and List, J. A. (2020). How can bill and melinda gates increase other people's donations to fund public goods? *Journal of Public Economics*, 191:104296.

Kasy, M. and Sautmann, A. (2021). Adaptive treatment assignment in experiments for policy choice. *Econometrica*, 89(1):113–132.

Kowalski, A. E. (2016). Doing More When You're Running LATE: Applying Marginal Treatment Effect Methods to Examine Treatment Effect Heterogeneity in Experiments. NBER Working Papers 22363, National Bureau of Economic Research, Inc.

Lai, T. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22.

Liebscher, E. (2005). Towards a unified approach for proving geometric ergodicity and mixing properties of nonlinear autoregressive processes. *Journal of Time Series Analysis*, 26(5):669–689.

Meager, R. (2020). Aggregating distributional treatment effects: Abayesian hierarchical analysis of the microcredit literature.

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Qin, C. and Russo, D. (2022). Adaptivity and confounding in multi-armed bandit experiments. *arXiv*.

Roberts, G. O. and Rosenthal, J. S. (2004). General state space markov chains and mcmc algorithms. *Probability surveys*, 1:20–71.

Russo, D. (2016). Simple bayesian algorithms for best arm identification.

Russo, D. and Van Roy, B. (2016). An information-theoretic analysis of thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471.

Schlaifer, R. and Raiffa, H. (1961). *Applied statistical decision theory*.

Stuart, E. A., Cole, S. R., Bradshaw, C. P., and Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(2):369–386.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294.

Vivalt, E. (2020). How Much Can We Generalize From Impact Evaluations? *Journal of the European Economic Association*, 18(6):3045–3089.

Vivalt, E. and Coville, A. (2021). How do policy-makers update their beliefs?

Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186.

Watkins, C. J. C. H. (1989). *Learning from Delayed Rewards*. PhD thesis, King's College, Cambridge, UK.

Yu, B. (1994). Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, pages 94–116.
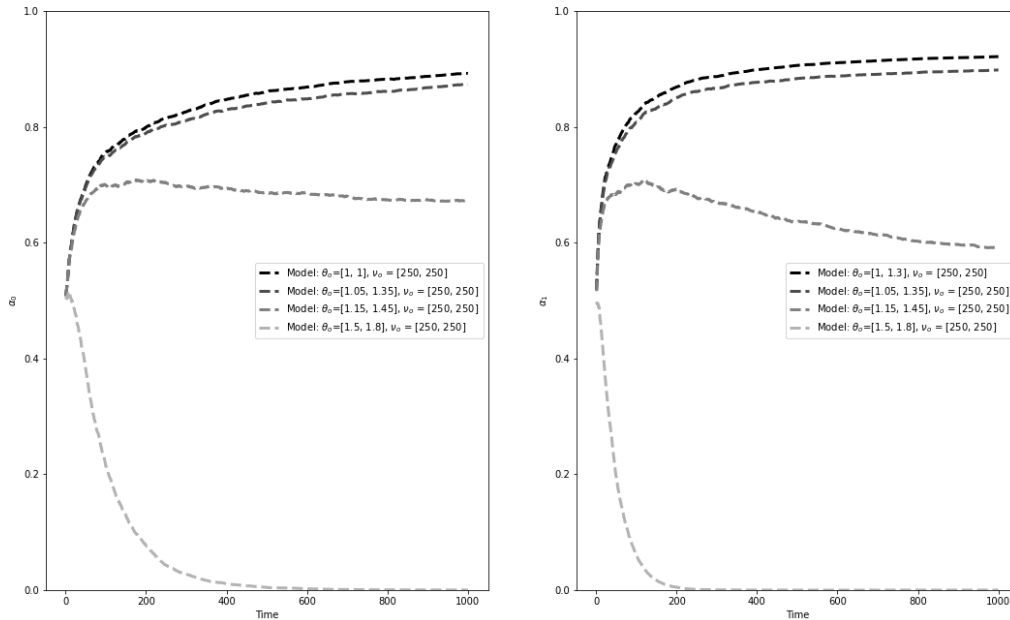
Appendix: Figures & Tables

Figure 1: External Validity - α_t^o

Notes: This figure plots $\alpha^0(d = 0, x)$ (left plot) and $\alpha^0(d = 1, x)$ (right plot) under two alternative sets of priors. For the confident model, the initial priors are: $\zeta_0^0 = \zeta_0^1 = \theta; \nu_0^0 = [1,1]; \nu_0^1 = [250,250]$. For the stubborn model, the initial priors are: $\zeta_0^0 = \theta; \zeta_0^1 = \theta + 0.3; \nu_0^0 = [1,1]; \nu_0^1 = [250,250]$. These figures are based on 1,000 simulations using the following parameters: $\theta = [1, 1.3], \epsilon = 0.5$.
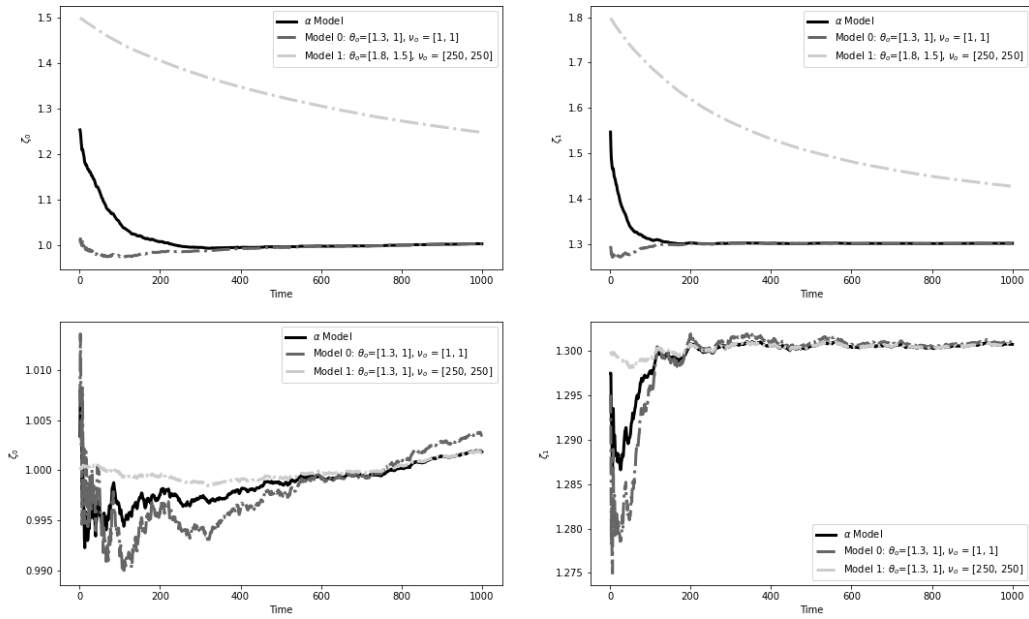
Figure 2: Posterior Beliefs Over Time, Holding Behavior Constant

Notes: This figure plots the policymakers posterior beliefs (i.e. $[\zeta_t^o(0,x), \zeta_t^o(1,x)]$) over time, distinguishing between two alternative sets of initial priors. In the top panel, one of the initial priors is stubborn; and in the bottom panel, one of the initial priors is confident. For the stubborn model, the initial priors are: $\zeta_0^0 = \theta; \zeta_0^1 = \theta + 0.3; \nu_0^0 = [1,1]; \nu_0^1 = [250, 250]$. For the confident model, the initial priors are: $\zeta_0^0 = \zeta_0^1 = \theta; \nu_0^0 = [1,1]; \nu_0^1 = [250, 250]$. These figures are based on 1,000 simulations using the following parameters: $\theta = [1, 1.3]$, $\epsilon = 0.5$.
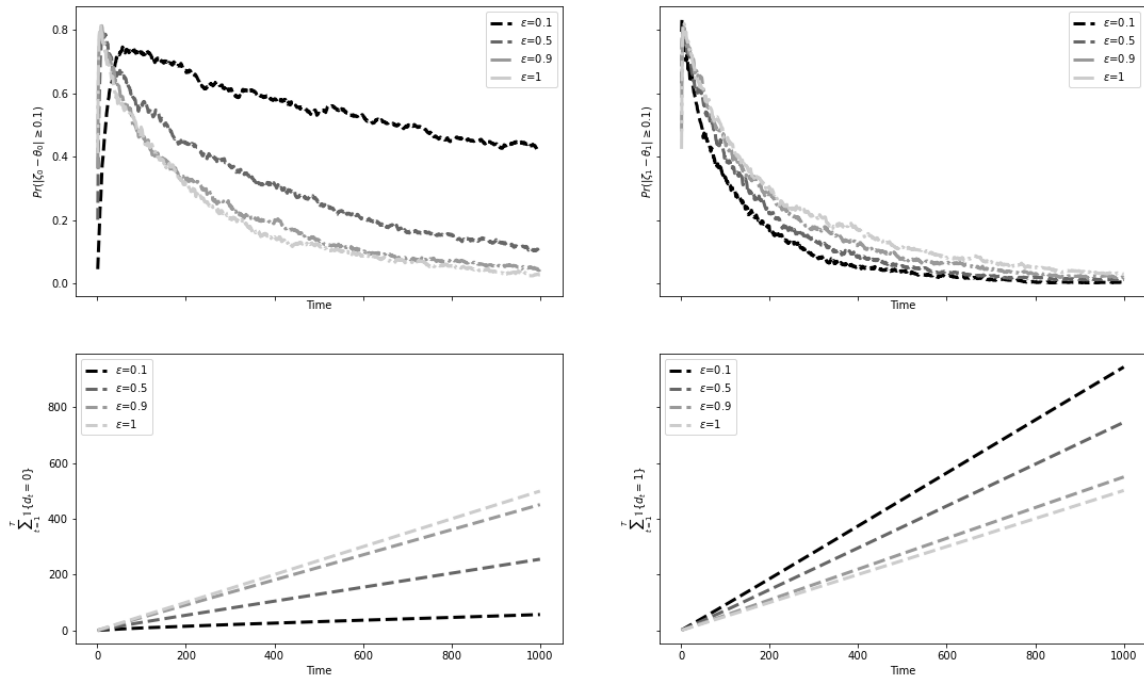
Figure 3: Concentration Bounds and Frequency of Play

Notes: The top panel plots concentration bounds over time for different values of ϵ. The bottom panel plots the number of times the experimental arm was played at time t for different values of ϵ. The graphs on the left correspond treatment arm $d = 0$; the graphs on the right correspond to treatment arm $d = 1$. These figures are based on 1,000 simulations using the following parameters: $\theta = [1, 1.3]$; $\zeta_0^o = \theta$; $\zeta_1^o = \theta$; $\nu_0^o = [1, 1]$; $\nu_1^o = [1, 1]$.
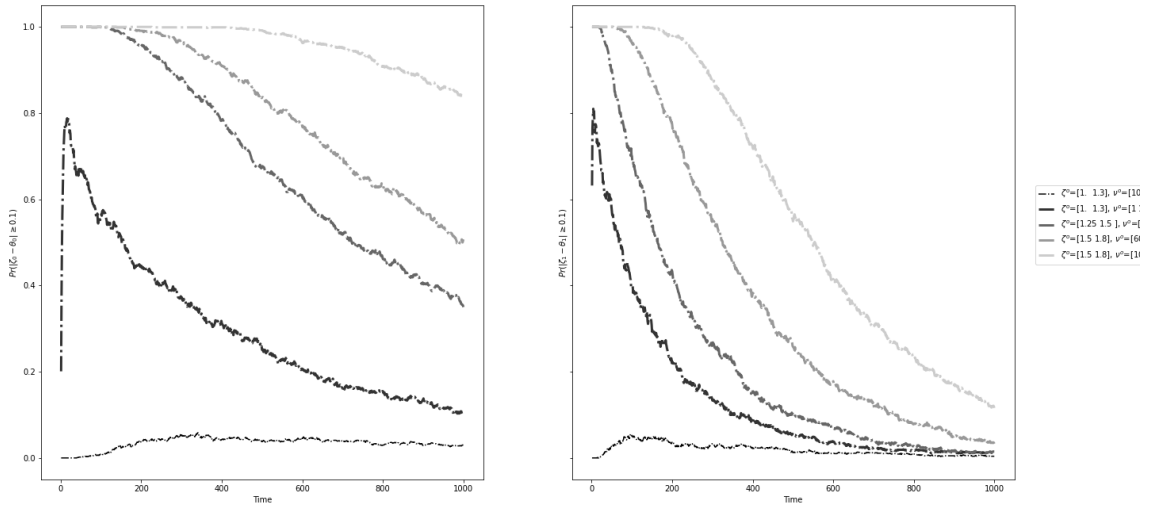
Figure 4: Concentration Bounds by Model Stubbornness

Notes: The figure plots concentration bounds over time for different degrees of model stubbornness. The lines in these plots appear in descending order of stubbornness, with the top line being most stubborn and the bottom line being the most confident. The graphs on the left correspond treatment arm $d = 0$; the graphs on the right correspond to treatment arm $d = 1$. These figures are based on 1,000 simulations using the following parameters: $\theta = [1, 1.3]$, $\epsilon = 0.5$. The initial priors are specified in the legend.
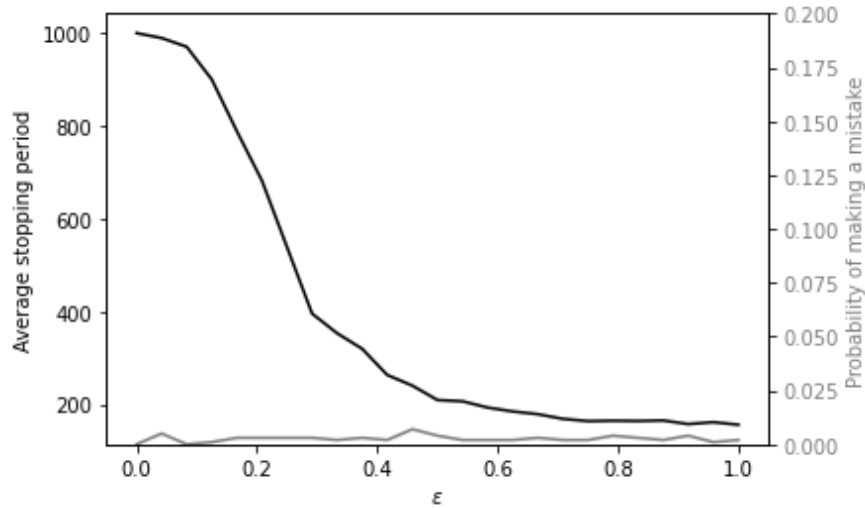

Figure 5: Stopping Period and Probability of Making a Mistake

Notes: This figure plots the average stopping period (left axis) and the probability of making a mistake at the stopping period (right axis) by ϵ. These figures are based on 1,000 simulations using the following parameters: $\theta = [1, 1.3]$; $\zeta_0^o = \theta$; $\zeta_1^o = \theta$; $v_0^o = [1,1]$; $v_1^o = [1,1]$; $B = 100$.
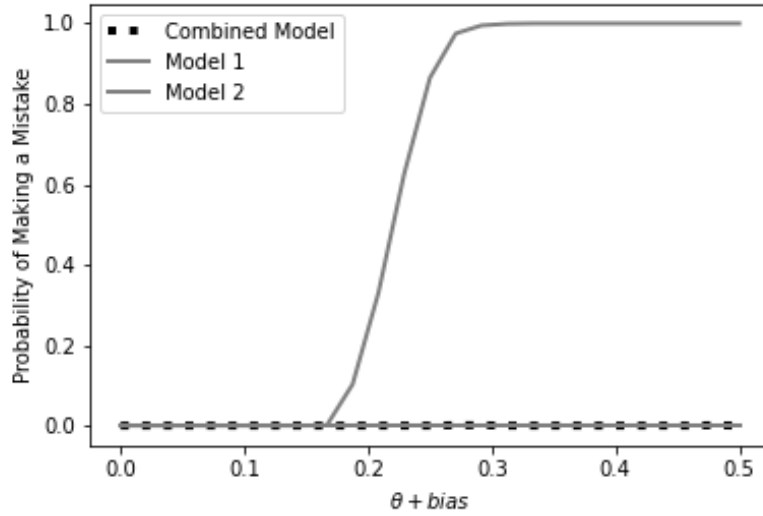
41

Figure 6: Probability of Making a Mistake by Model Bias

Notes: The figure plots the probability of making a mistake at the stopping period by the degree of bias in model 1's initial priors. These figures are based on 1,000 simulations using the following parameters: $\theta = [1,1.3]$; $v_0^0 = v_0^1 = [250,250]$; $\zeta_0^0 = [\theta(0) + bias, \theta(1) - bias]$; $\zeta_0^1 = \theta$, $\epsilon = 0.5$.
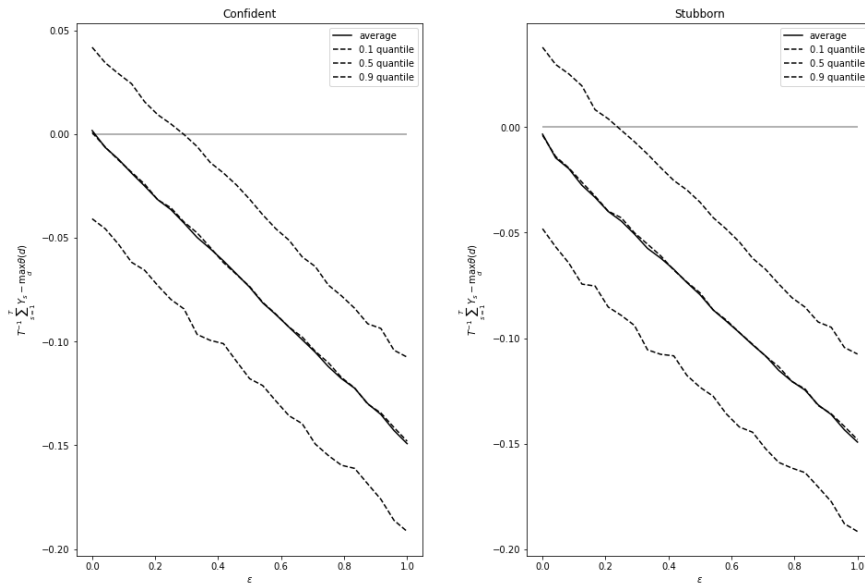

Figure 7: Relative Average Earnings During the Experiment

Notes: This figure plots by ϵ, the average earnings net of maximal earnings. These figures are based on 1,000 simulations using the following parameters: $\theta = [1,1.3]$; $\zeta_0^o = \theta$; $\zeta_1^o = \theta$; $v_0^o = [1,1]$; $v_1^o = [1,1]$.

Figure 8: Experimentation versus Exploitation – Expected Payoffs

Notes: This figure plots by ϵ, the expected payoffs as defined by Equation 4.1. These figures are based on 1,000 simulations using the following parameters: $\theta = [1, 1.3]$; $\zeta_0^o = \thet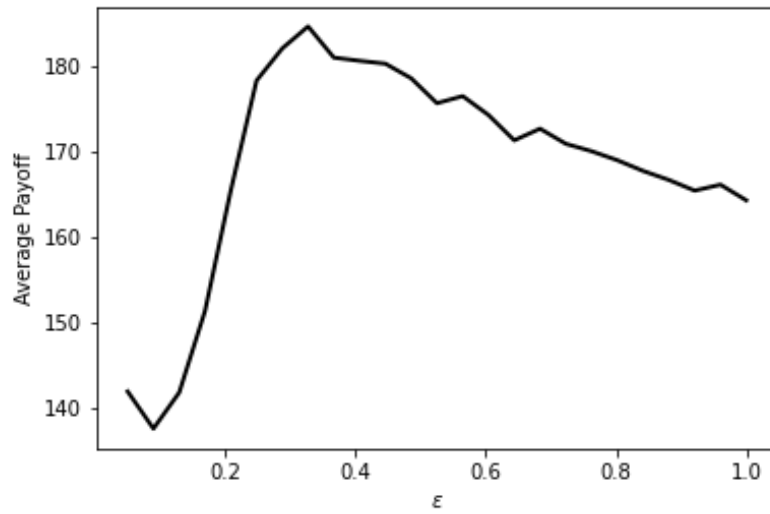a$; $\zeta_1^o = \theta$; $v_0^o = [1, 1]$; $v_1^o = [1, 1]$; $B = 100$; $\beta^t = 0.994$; $c = 1.15$; $\lambda = 1,100$.

Table 1: Fundraising Matching Grant Experiment

	RCT	α-Model				Non-α Model			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
ϵ-Greedy	1	0.2	0.25	0.3	0.35	0.2	0.25	0.3	0.35
Average Stopping Period	600	198.3	171.5	180.6	182.4	548.8	527.5	511.6	509.2
Probability of Making a Mistake at Stopping	0.006	0.024	0.027	0.028	0.029	0.033	0.022	0.028	0.022
Average Payoffs	2071.57	2127.94	2127.87	2126.12	2125.46	2117.05	2116.97	2112.32	2111.30
10th Quantile of Payoff Distribution	2032.11	2106.23	2108.70	2104.645	2102.71	2079.99	2080.34	2073.30	2074.08
50th Quantile of Payoff Distribution	2072.69	2131.21	2131.15	2129.36	2129.60	2118.51	2118.63	2114.34	2113.83
90th Quantile of Payoff Distribution	2109.09	2149.75	2149.75	2149.75	2149.75	2153.62	2153.62	2153.62	2153.62

Notes: This table reports the results of re-simulating the Bill Melinda Gates Foundation Experiment (Karlan and List, 2020). Each simulation lasted for at most 600 periods, and the results are averaged over 1,000 simulations. In column 1, we report the results for a standard randomized control trial. In columns 2-5, we report the results for several multi-prior models with different degrees of experimentation, as indicated by ϵ. For the multi-prior models, we consider 5 sets of priors, corresponding to: $\zeta_0^0 = [0,0], \nu = [0.01, 0.01], \zeta_0^1 = [3.424909, 3.4407649], \nu = [782, 252], \zeta_0^2 = [3.424909, 3.4407649], \nu = [0.01, 0.01], \zeta_0^3 = [3.106395, 3.2722909], \nu = [284, 223], \zeta_0^4 = [3.106395, 3.2722909], \nu = [0.01, 0.01]$. For the stopping rule, we set $B = 100$, In columns 6-10, we report the results for models with different degrees of experimentation, as indicated by ϵ, but no priors.

Table 2: Fundraising Matching Grant Experiment - External Validity

$\theta_0 = 3.34$			$\theta_1 = 3.56$		
ζ_0	ν_0	α_0	ζ_0	ν_0	α_1
0	0.01	0.013	0	0.01	0.038
3.42	782	0.646	3.44	252	0.869
3.42	0.01	0.014	3.44	0.01	0.041
3.11	284	0.313	3.27	223	0.012
3.11	0.01	0.014	3.27	0.01	0.014

Notes: Notes: This table reports the α corresponding to the initial priors when re-simulating the Bill Melinda Gates Foundation Experiment (Karlan and List, 2020), with $\epsilon = 0.30$. Each simulation lasted for at most 600 periods, and the results are averaged over 1,000 simulations. For the multiprior models, we consider 5 sets of priors, corresponding to: $\zeta_0^0 = [0,0], \nu = [0.01, 0.01]$, $\zeta_0^1 = [3.424909, 3.4407649], \nu = [782, 252]$, $\zeta_0^2 = [3.424909, 3.4407649], \nu = [0.01, 0.01]$, $\zeta_0^3 = [3.106395, 3.2722909], \nu = [284, 223]$, $\zeta_0^4 = [3.106395, 3.2722909], \nu = [0.01, 0.01]$. For the stopping rule, we set $B = 100$.

Online Appendix

A   Notation and some definitions

For any set S, let $\Delta(S)$ be the set of Borel probability measures over S.

For each $t \in \mathbb{N}$ and each $(d,x) \in \mathbb{D} \times \mathbb{X}$, let $D^t(x) := (D_1(x),...,D_t(x))$, $Y^t(x) := (Y_1(D_1(x),x),...,Y_t(D_t(x),x))$ and $Y^t(d,x) := (Y_1(d,x),...,Y_t(d,x))$. Let $(Y^{t-1}(x),D^{t-1}(x)) \mapsto \delta_t(Y^{t-1},D^{t-1})(.|x) \in \Delta(\mathbb{D})$ be the treatment assignment policy rule and $(Y^{t-1}(x),D^{t-1}(x)) \mapsto \sigma_t(Y^{t-1},D^{t-1})(x) \in [0,1]$ be the stopping policy rule. When there is no risk of confusion we will simply use $\delta_t(.|x)$ and $\sigma_t(x)$ to denote these rules.

We define the probability measure $\mathbf{P}$ that is used in the probability statements in our proofs. Formally, let $\mathbf{P}$ be a probability measure over histories $((Y^T(d,x))_{(d,x) \in \mathbb{D} \times \mathbb{X}}, (D^T(x))_{x \in \mathbb{X}})$ (and easily extended to infinite histories) constructed as follows: By assumption, for all $(d,x) \in \mathbb{D} \times \mathbb{X}$, $Y_1(d,x)$ is IID drawn from $P(.|d,x)$ and $D_1(x) \sim \delta_1(.|x)$. For any $t > 1$, given the past history $((Y^{t-1}(d,x))_{(d,x) \in \mathbb{D} \times \mathbb{X}}, (D^{t-1}(x))_{x \in \mathbb{X}})$, with probability $\sigma_t(x)$ the experiment is stopped and $D_t(x)$ is the same for all subsequent instances; with probability $1 - \sigma_t(x)$ the experiment is not stopped and $D_t(x) \sim \delta_t(.|x)$; $Y_t(d,x)$ is is IID drawn from $P(.|d,x)$.

For each $(d,x) \in \mathbb{D} \times \mathbb{X}$ and each $t \in \mathbb{N}$, let

$$N_t(d,x) := \sum_{s=1}^{t} 1\{D_s(x) = d\} \ and \ f_t(d,x) := N_t(d,x)/t \tag{A.1}$$

$$\iota_t(d,x) := t^{-1} \sum_{s=1}^{t} \delta_s(d|x) \tag{A.2}$$

$$J_t(d,x) := t^{-1} \sum_{s=1}^{t} 1\{D_s(x) = d\}Y_s(d,x). \tag{A.3}$$

B   Proof of Proposition 2.1

To show Proposition 2.1 we use the following lemmas whose proofs are relegated to Supplemental Material J. The first lemma studies the process $g(Y_s, D_s, \theta) := (1\{D_s(x) = d\}(\ell(Y_s,\theta) - E[\ell(Y_s,\theta) \mid D_s(x)])$ for all $s \geq 1$, where the expectation is taken with respect to $P(.|d,x)$ and $\ell(.,\theta) := \log p_\theta(.)/p(. \mid d,x)$ and $p(\cdot|d,x)$ is the true PDF of $Y(d,x)$.

In what follows, let $B(\theta,\delta)$ as the ball with radius $\delta > 0$, centered at $\theta \in \Theta$.

Lemma B.1. *Suppose $\Theta \subseteq \mathbb{R}^{|\Theta|}$ is compact and there exists a $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ increasing, such that $\varphi(0) = 0$*

and

$$\max_{\theta' \in \Theta} E[\sup_{\theta \in B(\theta',\delta)} |\ell(Y(d,x),\theta) - \ell(Y(d,x),\theta')|^2] \le \varphi(\delta)^2, \ \forall \delta > 0.$$

Then there exists a $C > 0$ such that for any $t \in \mathbb{N}$ and $\gamma > 0$,

$$\mathbf{P}\left(\sup_{\theta \in \Theta} |t^{-1} \sum_{s=1}^{t} g(Y_s, D_s, \theta)| \ge \sqrt{1/(2C\gamma)}\sqrt{\Lambda(t,|\Theta|)}\right) \le \gamma$$

where $\Lambda(t,|\Theta|) := \min_{\delta \ge 0}(t^{-1}\delta^{-|\Theta|} + \varphi(\delta))$ and is decreasing in t and increasing in $|\Theta|$ and $\lim_{t \to \infty} \Lambda(t,|\Theta|) = 0$.

The following lemma provides a non-asymptotic bound for the ratio of the weights for any two models. In particular, it relates the weights, α_t^o, with the Laplace transform of the CDF

$$u \mapsto G_{d,x}^o(u) := \mu_0^o(d,x)(KL_{d,x}(\theta) \le u).$$

To ou knowledge, this result is new and might be of independent interest.

Lemma B.2. *Take any $o,o' \in \{0,...,L\}$ and $(d,x) \in \mathbb{D} \times \mathbb{X}$. Suppose $\Theta \subseteq \mathbb{R}^{|\Theta|}$ is compact and there exists a $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ increasing, such that $\varphi(0) = 0$ and*

$$\max_{\theta' \in \Theta} E[\sup_{\theta \in B(\theta',\delta)} |\ell(Y(d,x),\theta) - \ell(Y(d,x),\theta')|^2] \le \varphi(\delta)^2, \ \forall \delta > 0.$$

Then, for any $\gamma > 0$,

$$\frac{\alpha_t^o(d,x)}{\alpha_t^{o'}(d,x)} \le \frac{\int G_{d,x}^o(u)e^{-N_t(d,x)u}\,du}{\int G_{d,x}^{o'}(u)e^{-N_t(d,x)u}\,du}e^{tr_t} \ \forall t,$$

with probability larger than $1 - \gamma$, where $r_t := \sqrt{1/(2C\gamma)}\sqrt{\Lambda(t,|\Theta|)}$ and C is as in lemma B.1.

Proof of Proposition 2.1. Suppose the conditions in Lemma B.2 hold (we show they do towards the end of the proof).

Throughout, fix a $(d,x) \in \mathbb{D} \times \mathbb{X}$ and $o,o' \in \{0,...,L\}$.

For any $\gamma > 0$, let $O_t(\gamma)$ be the set of histories such that

$$\frac{\alpha_t^o(d,x)}{\alpha_t^{o'}(d,x)} \le \frac{\int G_{d,x}^o(u)e^{-N_t(d,x)u}\,du}{\int G_{d,x}^{o'}(u)e^{-N_t(d,x)u}\,du}e^{tr_t},$$

where $(r_t)_t$ is the sequence defined in Lemma B.2; observe that r_t depends on γ, and thus, so does the set $O_t(\gamma)$. By Lemma B.2, $\mathbf{P}(O_t(\gamma)^C) \le \gamma$ for any $t \in \mathbb{N}$.

2

For any $\varepsilon > 0$, let $S_t(\varepsilon)$ be the set of histories such that $N_t(d,x)/t \geq \iota_t(d,x) - \varepsilon$. By Lemma E.1 in Appendix E, there exists a t_ε such that $\mathbf{P}(S_t(\varepsilon)^C) \leq \varepsilon$ for all $t \geq t_\varepsilon$.

In this case, it follows that for any $\varepsilon > 0$ any $\gamma > 0$,

$$\mathbf{P}\left(\frac{\alpha_t^o(d,x)}{\alpha_t^{o'}(d,x)} \geq \varepsilon\right) \leq \mathbf{P}\left(\frac{\alpha_t^o(d,x)}{\alpha_t^{o'}(d,x)} \geq \varepsilon \cap S_t(\varepsilon) \cap O_t(\gamma)\right) + \mathbf{P}\left(O_t(\gamma)^C\right) + \mathbf{P}\left(S_t(\varepsilon)^C\right)$$

for all $t \in \mathbb{N}$.

In what follows, take an arbitrary sequence in $O_t(\gamma)$. Let $u_o(d,x)$ as in the definition. If $u_o(d,x) - u_{o'}(d,x) =: A > 0$, then,

$$
\begin{aligned}
\frac{\alpha_t^o(d,x)}{\alpha_t^{o'}(d,x)} &\leq e^{2tr_t} \frac{N_t(d,x) \int_0^\infty \mu_0^o(d,x)\left(KL_{d,x}(\theta) \leq v\right) e^{-N_t(d,x)v} dv}{N_t(d,x) \int_0^\infty \mu_0^{o'}(d,x)\left(KL_{d,x}(\theta) \leq v\right) e^{-N_t(d,x)v} dv} \\
&\leq e^{2tr_t} \frac{N_t(d,x) \int_{u_o(d,x)}^\infty \mu_0^o(d,x)\left(KL_{d,x}(\theta) \leq v\right) e^{-N_t(d,x)v} dv}{N_t(d,x) \int_{u_{o'}(d,x)}^{u_{o'}(d,x)+0.5A} \mu_0^{o'}(d,x)\left(KL_{d,x}(\theta) \leq v\right) e^{-N_t(d,x)v} dv} \\
&\leq e^{2tr_t} \frac{e^{-N_t(d,x)u_o(d,x)} \int_{u_o(d,x)}^\infty \mu_0^o(d,x)\left(KL_{d,x}(\theta) \leq v\right) dv}{e^{-N_t(d,x)(m_{o'}(d,x)+0.5A)} \int_{m_{o'}(d,x)}^{u_{o'}(d,x)+0.5A} \mu_0^{o'}(d,x)\left(KL_{d,x}(\theta) \leq v\right) dv} \\
&= e^{-N_t(d,x)0.5A+2tr_t} \frac{\int_{u_o(d,x)}^\infty \mu_0^o(d,x)\left(KL_{d,x}(\theta) \leq v\right) dv}{\int_{u_{o'}(d,x)}^{u_{o'}(d,x)+0.5A} \mu_0^{o'}(d,x)\left(KL_{d,x}(\theta) \leq v\right) dv}
\end{aligned}
$$

where the last line follows from definition of A.

The fraction in the previous display is a fixed number. Under the set $S_t(\varepsilon)$, $\frac{N_t(d,x)}{t}0.5A + 2r_t \geq (\iota_t(d,x) - \varepsilon)0.5A - 2\varepsilon$. Under the assumption that $\inf_t \iota_t(d,x) > 0$, there exists a $\bar{\varepsilon} > 0$ (not dependant on t) such that for any $\varepsilon \leq \bar{\varepsilon}$, $\frac{N_t(d,x)}{t}0.5A + 2r_t \geq c > 0$. Thus, we obtain $\frac{\alpha_t^o(d,x)}{\alpha_t^{o'}(d,x)} = O(e^{-tc})$, which in turn implies that for any $\varepsilon \in (0,\bar{\varepsilon})$ any $\gamma > 0$,

$$\mathbf{P}\left(\frac{\alpha_t^o(d,x)}{\alpha_t^{o'}(d,x)} \geq \varepsilon\right) \leq \mathbf{P}\left(e^{-tc} \geq \varepsilon\right) + \mathbf{P}\left(O_t(\gamma)^C\right) + \mathbf{P}\left(S_t(\varepsilon)^C\right)$$

for all $t \in \mathbb{N}$. By choosing $\gamma = \varepsilon$, for any $t \geq \max\{t_\varepsilon, c^{-1}\log 1/\varepsilon\}$ the RHS is less than 2ε and the desired result holds.

We now verify that the conditions in Lemma B.2 (and thus, Lemma B.1) hold. Θ is compact by assumption, so we "just" need to verify the continuity condition.

By Assumption, $\theta \mapsto \log p_\theta(\cdot)$ is continuous a.s.-$\mathbf{P}$. This implies that $\theta \mapsto \ell(\cdot,\theta)$ is continuous a.s.-$\mathbf{P}$. Continuous functions over compact sets are uniformly continuous, thus $f(Y(d,x),\delta) := \sup_{||\theta-\theta'||\leq\delta} |\ell(Y(d,x),\theta) -$

3

$\ell(Y(d,x),\theta')|$ converges to 0 as δ vanishes a.s.-**P**. By Assumption,

$$E[\sup_{\theta\in\Theta}|\log p_\theta(Y(d,x))|^2] < \infty,$$

and thus by the Dominated convergence theorem, $\lim_{\delta\to 0} E[f(Y(d,x),\delta)^2] = 0$ as desired.

$\square$

C Proof of Lemma 2.1

Proof of Lemma 2.1 . Let p_θ denote a Gaussian PDF with mean θ and variance 1.

(1) It follows that

$$\int \prod_{s=1}^{t} (p_\theta(Y_s))^{1\{D_s(x)=d\}} \mu_0^o(d,x)(d\theta) = \int (2\pi)^{-0.5\sum_{s=1}^{t}1\{D_s(x)=d\}} \exp\left\{-\frac{1}{2}\sum_{s=1}^{t}1\{D_s(x)=d\}(Y_s(d,x)-m_t(d,x))^2\right\}$$

$$\times \exp\left\{-\frac{1}{2}\sum_{s=1}^{t}1\{D_s(x)=d\}(m_t(d,x)-\theta)^2\right\}$$

$$\times \exp\left\{-\sum_{s=1}^{t}1\{D_s(x)=d\}(Y_s(d,x)-m_t(d,x))(m_t(d,x)-\theta)\right\}\phi(\theta;\zeta_0^o(d,x),1/v_0^o(d,x))d\theta$$

where $m_t(d,x) := \sum_{s=1}^{t}1\{D_s(x)=d\}Y_s(d,x)/\sum_{s=1}^{t}1\{D_s(x)=d\}$. Observe that

$$\sum_{s=1}^{t}1\{D_s(x)=d\}(Y_s(d,x)-m_t(d,x)) = 0,$$

so, letting $N_t(d,x) := \sum_{s=1}^{t}1\{D_s(x)=d\}$ it follows that

$$\int \prod_{s=1}^{t}(p_\theta(Y_s))^{1\{D_s(x)=d\}}\mu_0^o(d,x)(d\theta) = \frac{\exp\left\{-\frac{1}{2}\sum_{s=1}^{t}1\{D_s(x)=d\}(Y_s(d,x)-m_t(d,x))^2\right\}}{(2\pi)^{0.5\sum_{s=1}^{t}1\{D_s(x)=d\}+0.5}N_t(d,x)^{0.5}}$$

$$\times \int (2\pi/N_t(d,x))^{-1/2}\exp\left\{-\frac{1}{2}(m_t(d,x)-\theta)^2 N_t(d,x)\right\}$$

$$\phi(\theta;\zeta_0^o(d,x),1/v_0^o(d,x))d\theta.$$

The expression of the integral can be viewed as a convolution between to Gaussian PDFs one indexed by $(0,1/N_t(d,x))$ and $(\zeta_0^o(d,x),1/v_0^o(d,x))$ resp, which in turn is equivalent to PDF of the sum of the correspond-

4

ing random variables evaluated at $m_t(d,x)$. Therefore,

$$\int \prod_{s=1}^{t}(p_\theta(Y_s))^{1\{D_s(x)=d\}}\mu_0^o(d,x)(d\theta) = C\phi(m_t(d,x);\zeta_0^o(d,x),(N_t(d,x)+v_0^o(d,x))/(N_t(d,x)v_0^o(d,x)))$$

where $C := (2\pi)^{-0.5\sum_{s=1}^{t}1\{D_s(x)=d\}+0.5}N_t(d,x)^{-1/2}\exp\left\{-\frac{1}{2}\sum_{s=1}^{t}1\{D_s(x)=d\}(Y_s(d,x)-m_t(d,x))^2\right\}$ which, importantly, doesn't depend on the model o.

Hence

$$\alpha_t^o(d,x) = \frac{\phi(m_t(d,x);\zeta_0^o(d,x),(N_t(d,x)+v_0^o(d,x))/(N_t(d,x)v_0^o(d,x)))}{\sum_{o=0}^{L}\phi(m_t(d,x);\zeta_0^o(d,x),(N_t(d,x)+v_0^o(d,x))/(N_t(d,x)v_0^o(d,x)))}$$

and since $m_t = J_t/f_t$, the desired result follows.

(2) Follows from the definition of ϕ.

(3) By assumption and by Lemma E.1, it can be shown that $m_t(d,x) = \theta(d,x)+o_{\mathbf{P}}(1)$ and that $1/N_t(d,x) = o_{\mathbf{P}}(1)$. By continuity of ϕ, the result follows. □

D   Non-stochastic Bounds

D.1   Non-stochastic Bounds for α_t

For each $t \in \{1,...,T\}$ and each $o \in \{1,...,L\}$, let $\overline{\alpha}_t^o : \mathbb{R}_+ \times [0,1] \times \mathbb{R}^{1+L} \times \mathbb{N}^{1+L} \to \mathbb{R}_+$ and $\underline{\alpha}_t^o : \mathbb{R}_+ \times [0,1] \times \mathbb{R}^{1+L} \times \mathbb{N}^{1+L} \to \mathbb{R}_+$ be defined as follows

$$\overline{\alpha}_t^o(\delta,g,a,b) := \min\left\{1, \frac{e^{\overline{\ell}_t(\delta,g,a^o,b^o)}}{\sum_{o'=0}^{L}e^{\underline{\ell}_t(\delta,g,a^{o'},b^{o'})}}\right\}, \text{ and } \underline{\alpha}_t^o(\delta,g,a,b) := \frac{e^{\underline{\ell}_t(\delta,g,a^o,b^o)}}{\sum_{o'=0}^{L}e^{\overline{\ell}_t(\delta,g,a^{o'},b^{o'})}}$$

for any $(\delta,g,a,b) \in \mathbb{R}_+ \times [0,1] \times \mathbb{R}^{1+L} \times \mathbb{N}^{1+L}$, where

$$\overline{\ell}_t(\delta,g,a^o,b^o) := -\log\underline{\sigma}_t - 0.5\frac{\max\{g^2(a^o)^2 - 2\delta a^o, 0\}}{\overline{\sigma}_t^2}$$

$$\underline{\ell}_t(\delta,g,a^o,b^o) := -\log\overline{\sigma}_t - 0.5\frac{(\delta+a^o)^2}{g^2\underline{\sigma}_t^2}$$

with $(1+b^o/T)/b^o =: \underline{\sigma}_t^2$ and $(g+b^o/t)/(gb) =: \overline{\sigma}_t^2$.

Lemma D.1. *For any $o \in \{0,...,L\}$, any $t \in \{1,...,T\}$, any $(d,x) \in \mathbb{D} \times \mathbb{X}$ and any $\eta \in (0,h_t(d,x))$ and $\delta > 0$*

5

such that $f_t(d,x) - h_t(d,x) \geq -\eta$ and $|J_t(d,x) - f_t(d,x)\theta(d,x)| \leq \delta$, it follows that

$$\underline{\alpha}_t^o(\delta, h_t(d,x) - \eta, |\bar{\zeta}_0(d,x)|, v_0(d,x)) \leq \alpha_t^o(d,x) \leq \overline{\alpha}_t^o(\delta, h_t(d,x) - \eta, |\bar{\zeta}_0(d,x)|, v_0(d,x)).$$

Proof of Lemma D.1 . Note that

$$\phi(J_t(d,x)/f_t(d,x); \zeta_0^o(d,x), (N_t(d,x) + v_0^o(d,x))/(N_t(d,x)v_0^o(d,x)))$$

$$= \frac{\sqrt{(N_t(d,x)v_0^o(d,x))}}{2\pi\sqrt{(N_t(d,x) + v_0^o(d,x))}} \exp\left\{-0.5 \frac{\left(J_t(d,x) - f_t(d,x)\zeta_0^o(d,x)\right)^2}{f_t(d,x)} \frac{N_t(d,x)v_0^o(d,x)}{f_t(d,x)(N_t(d,x) + v_0^o(d,x))}\right\}$$

$$= \frac{\sqrt{f_t(d,x)v_0^o(d,x)}}{2\pi\sqrt{(f_t(d,x) + v_0^o(d,x)/t)}} \exp\left\{-0.5 \frac{\left(\bar{J}_t(d,x) - f_t(d,x)\bar{\zeta}_0^o(d,x)\right)^2}{f_t(d,x)} \frac{v_0^o(d,x)}{(f_t(d,x) + v_0^o(d,x)/t)}\right\}$$

where $\bar{}$ indicates centered at $\theta(d,x)$. Henceforth, let $\sigma_t^2 := (f_t(d,x) + v_0^o(d,x)/t)/(f_t(d,x)v_0^o(d,x))$.

Under $|\bar{J}_t(d,x)| \leq \delta$, if follows that

$$\begin{aligned}(\bar{J}_t(d,x) - f_t(d,x)\bar{\zeta}_0^o(d,x))^2 &= (\bar{J}_t(d,x))^2 + (f_t(d,x)\bar{\zeta}_0^o(d,x))^2 - 2\bar{J}_t(d,x)f_t(d,x)\bar{\zeta}_0^o(d,x)\\ &\leq \delta^2 + (f_t(d,x)\bar{\zeta}_0^o(d,x))^2 + 2\delta f_t(d,x)|\bar{\zeta}_0^o(d,x)|\\ &\leq (\delta + |\bar{\zeta}_0^o(d,x)|)^2\end{aligned}$$

and, in addition, if $f_t(d,x) - h_t(d,x) \geq -\eta$ with $\eta \leq h_t(d,x)$, then

$$\begin{aligned}(\bar{J}_t(d,x) - f_t(d,x)\bar{\zeta}_0^o(d,x))^2 &= (\bar{J}_t(d,x))^2 + (f_t(d,x)\bar{\zeta}_0^o(d,x))^2 - 2\bar{J}_t(d,x)f_t(d,x)\bar{\zeta}_0^o(d,x)\\ &\geq (f_t(d,x)\bar{\zeta}_0^o(d,x))^2 - 2f_t(d,x)\delta|\bar{\zeta}_0^o(d,x)|\\ &\geq (h_t(d,x) - \eta))^2(\bar{\zeta}_0^o(d,x))^2 - 2\delta|\bar{\zeta}_0^o(d,x)|.\end{aligned}$$

Also, under these conditions,

$$\begin{aligned}\sigma_t^2 &\geq (1 + v_0^o(d,x)/t)/(v_0^o(d,x)) \geq (1 + v_0^o(d,x)/T)/(v_0^o(d,x)) =: \underline{\sigma}_t^2\\ &\leq (h_t(d,x) - \eta + v_0^o(d,x)/t)/((h_t(d,x) - \eta)v_0^o(d,x)) =: \overline{\sigma}_t^2.\end{aligned}$$

Therefore,

$$\log\phi(\bar{m}_t(d,x) - \bar{\zeta}_0^o(d,x); 0, \sigma_t^2(d,x)) \geq -\log\sigma_t - 0.5\frac{(\delta + |\bar{\zeta}_0^o(d,x)|)^2}{f_t(d,x)^2\sigma_t^2} + Cte \geq -\log\overline{\sigma}_t - 0.5\frac{(\delta + |\bar{\zeta}_0^o(d,x)|)^2}{(h_t(d,x) - \eta)^2\underline{\sigma}_t^2} + Cte$$

$$=: \underline{\ell}_t(\delta, h_t(d,x) - \eta, |\bar{\zeta}_0^o(d,x)|, v_0^o(d,x)) + Cte$$

6

and

$$\log \phi(\bar{m}_t(d,x) - \bar{\zeta}_0^o(d,x); 0, \sigma_t^2(d,x)) \leq -\log \sigma_t - 0.5 \frac{\max\{(h_t(d,x) - \eta))^2 (\bar{\zeta}_0^o(d,x))^2 - 2\delta|\bar{\zeta}_0^o(d,x)|, 0\}}{f_t(d,x)^2 \sigma_t^2} + Cte$$

$$\leq -\log \underline{\sigma}_t - 0.5 \frac{\max\{(h_t(d,x) - \eta))^2 (\bar{\zeta}_0^o(d,x))^2 - 2\delta|\bar{\zeta}_0^o(d,x)|, 0\}}{\overline{\sigma}_t^2} + Cte$$

$$=: \overline{\ell}_t(\delta, h_t(d,x) - \eta, |\bar{\zeta}_0^o(d,x)|, v_0^o(d,x)).$$

$\square$

Lemma D.2. *The following properties are true:*

1. *$\delta \mapsto \underline{\ell}_t(\delta, g, |\bar{\zeta}_0^o(d,x)|, v_0^o(d,x))$ is decreasing and $\delta \mapsto \overline{\ell}_t(\delta, g, |\bar{\zeta}_0^o(d,x)|, v_0^o(d,x))$ is non-decreasing.*

2. *$g \mapsto \underline{\ell}_t(\delta, g, |\bar{\zeta}_0^o(d,x)|, v_0^o(d,x))$ is increasing and $g \mapsto \overline{\ell}_t(\delta, g, |\bar{\zeta}_0^o(d,x)|, v_0^o(d,x))$ is decreasing.*

3. *$\delta \mapsto \overline{\alpha}_t^o(\delta, g, |\zeta_0(d,x)|, v_0(d,x))$ is increasing and $\delta \mapsto \underline{\alpha}_t^o(\delta, g, |\zeta_0(d,x)|, v_0(d,x))$ is decreasing.*

4. *$g \mapsto \overline{\alpha}_t^o(\delta, g, |\zeta_0(d,x)|, v_0(d,x))$ is decreasing and $g \mapsto \underline{\alpha}_t^o(\delta, g, |\zeta_0(d,x)|, v_0(d,x))$ is increasing.*

5. *$t \mapsto \overline{\alpha}_t^o(\delta, g, |\zeta_0(d,x)|, v_0(d,x))$ is increasing and $t \mapsto \underline{\alpha}_t^o(\delta, g, |\zeta_0(d,x)|, v_0(d,x))$ is decreasing.*

Proof of Lemma D.2. (1) It is easy to see that $\underline{\ell}_t(\delta, g, |\bar{\zeta}_0^o(d,x)|, v_0^o(d,x), h_t(d,x))$ is decreasing in δ and $\overline{\ell}_t(\delta, g, |\bar{\zeta}_0^o(d,x)|, v_0^o(d,x)$ is non-decreasing in δ.

(2) We first observe that $\underline{\sigma}_t^2$ is constant as a function of $g = h_t(d,x) - \eta$ and $\overline{\sigma}_t^2$ is an decreasing function of $g := h_t(d,x) - \eta$. Also, note that $\frac{d\underline{\ell}_t(\delta,g,|\bar{\zeta}_0^o(d,x)|,v_0^o(d,x))}{dg} = -\frac{1}{\underline{\sigma}} \frac{d\underline{\sigma}_t}{dg} + \frac{(\delta + |\bar{\zeta}_0^o(d,x)|)^2}{(g)^3 \underline{\sigma}_t^2}$, thus, since $g \geq 0$, $\underline{\ell}_t(\delta, g, |\bar{\zeta}_0^o(d,x)|, v_0^o(d,x))$ is increasing as a function of g. Similarly, $\overline{\ell}_t(\delta, g, |\bar{\zeta}_0^o(d,x)|, v_0^o(d,x))$ is increasing as a function of $\overline{\sigma}_t^2$ and decreasing as a direct function of g, thus by computing the derivative it can be shown that it is decreasing in g.

(3-4) By parts (1), it readily follows that $\overline{\alpha}_t^o(\delta, g, |\zeta_0(d,x)|, v_0(d,x))$ is increasing in δ and $\underline{\alpha}_t^o(\delta, g, |\zeta_0(d,x)|, v_0(d,x))$ is decreasing in δ. And by part (2) $\overline{\alpha}_t^o(\delta, g, |\zeta_0(d,x)|, v_0(d,x))$ is decreasing in g and $\underline{\alpha}_t^o(\delta, g, |\zeta_0(d,x)|, v_0(d,x))$ is increasing in g.

(5) It follows that $t \mapsto \overline{\sigma}_t^2$ is decreasing and $t \mapsto \underline{\sigma}_t^2$ is constant. Since $\overline{\ell}_t$ is non-increasing in $\overline{\sigma}_t^2$ it follows that it is non-decreasing in t. Similarly, $\underline{\ell}_t$ is increasing in $\overline{\sigma}_t^2$ and thus increasing in t.

These results imply that $t \mapsto \overline{\alpha}_t^o(\delta, g, |\zeta_0(d,x)|, v_0(d,x))$ is increasing and $t \mapsto \underline{\alpha}_t^o(\delta, g, |\zeta_0(d,x)|, v_0(d,x))$ is decreasing. $\square$

D.2 Non-stochastic Bounds for ζ_t^o

For any $t \in \mathbb{N}$, let $\Omega_0 : (D(\Omega_0) := [0,1] \times \mathbb{R} \times \mathbb{N}) \to \mathbb{R}$ and $\Omega : D(\Omega) := \mathbb{R}_+ \times D(\Omega_0) \to \mathbb{R}$ be such that

$$\Omega(\gamma, g, \zeta_0^o(d,x), v_0^o(d,x)) := \frac{\gamma}{g + v_0^o(d,x)/t} + \Omega_0(g, \zeta_0^o(d,x), v_0^o(d,x))$$

and

$$\Omega_0(g, \zeta_0^o(d,x), v_0^o(d,x)) := v_0^o(d,x) \left(\frac{(\zeta_0^o(d,x))_+/t}{g + v_0^o(d,x)/t} + \frac{(\zeta_0^o(d,x))_-/T}{1 + v_0^o(d,x)/t} \right)$$

for any $(\gamma, g, \zeta_0^o(d,x), v_0^o(d,x)) \in D(\Omega)$, where for any real number a, $a_+ := \max\{a, 0\}$ and $a_- := \min\{a, 0\}$.

Lemma D.3. *For any $o \in \{0, ..., L\}$, any $(d,x) \in \mathbb{D} \times \mathbb{X}$ and any $t \in \mathbb{N}$, the following are true:*

1. *$\gamma \mapsto \Omega(\gamma, g, \zeta_0^o(d,x), v_0^o(d,x))$ is increasing.*

2. *$g \mapsto \Omega(\gamma, g, \zeta_0^o(d,x), v_0^o(d,x))$ is decreasing and $g \mapsto \Omega_0(g, \zeta_0^o(d,x), v_0^o(d,x))$ is non-increasing.*

3. *If $\zeta_0(d,x) \leq 0$, $v_0(d,x) \mapsto \Omega(\gamma, g, \zeta_0^o(d,x), v_0^o(d,x))$ is decreasing; and if $\zeta_0(d,x) \leq (\geq) 0$ $v_0(d,x) \mapsto \Omega_0(g, \zeta_0^o(d,x), v_0^o(d,x))$ is non-increasing (non-decreasing).*

Proof of Lemma D.3. (1) Trivial.

(2) By inspection, $g \mapsto \Omega_0(g, \zeta_0^o(d,x), v_0^o(d,x))$ is non-increasing. In addition $g \mapsto \gamma/(g + v_0^o(d,x)/t)$ is decreasing.

(3) Trivial. $\qquad\square$

Lemma D.4. *For any $o \in \{0, ..., L\}$, any $(d,x) \in \mathbb{D} \times \mathbb{X}$ and any $t \in \mathbb{N}$, suppose $f_t(d,x) - h_t(d,x) \geq -\eta$ for some $\gamma \geq 0$ and $0 \leq \eta \leq h_t(d,x) \leq \iota_t(d,x)$. Then:*

1. *$|\zeta_t^o(d,x) - \theta(d,x)| \leq \Omega(\gamma, h_t(d,x) - \eta, |\zeta_0^o(d,x) - \theta(d,x)|, v_0^o(d,x))$.*

2. *$\zeta_t^o(d,x) - \theta(d,x) \leq \Omega(\gamma, h_t(d,x) - \eta, \zeta_0^o(d,x) - \theta(d,x), v_0^o(d,x))$.*

3. *$-(\zeta_t^o(d,x) - \theta(d,x)) \leq \Omega(\gamma, h_t(d,x) - \eta, -(\zeta_0^o(d,x) - \theta(d,x)), v_0^o(d,x))$.*

Proof of Lemma D.4 . (1) Under the conditions, it easy to see that

$$|\zeta_t^o(d,x) - \theta(d,x)| \leq \frac{\gamma + |\zeta_0^o(d,x) - \theta(d,x)| v_0^o(d,x)/t}{\iota_t(d,x) - \eta + v_0^o(d,x)/t} \leq \frac{\gamma + |\zeta_0^o(d,x) - \theta(d,x)| v_0^o(d,x)/t}{h_t(d,x) - \eta + v_0^o(d,x)/t}$$

$$= \Omega(\gamma, h_t(d,x) - \eta, |\zeta_0^o(d,x) - \theta(d,x)|, v_0^o(d,x)).$$

8

(2-3) The proof is analogous and thus omitted. □

D.3 Non-stochastic Bounds for ζ_t^α

Let $\Gamma_0 : D(\Gamma) := \mathbb{R}_+ \times [0,1] \times \mathbb{R}^{1+L} \times \mathbb{N}^{1+L} \to \mathbb{R}$ be such that

$$\Gamma_0(\gamma, g, \zeta_0(d,x), v_0(d,x)) := \sum_{o=0}^{L} \overline{\alpha}_t^o(\gamma, g, \zeta_0(d,x), v_0(d,x)) \Omega_0^+(g, \zeta_0^o(d,x), v_0^o(d,x))$$

$$+ \sum_{o=0}^{L} \underline{\alpha}_t^o(\gamma, g, \zeta_0(d,x), v_0(d,x)) \Omega_0^-(g, \zeta_0^o(d,x), v_0^o(d,x))$$

where $\Omega^+ := \max\{\Omega, 0\}$ and $\Omega^- := \min\{\Omega, 0\}$, and $\Gamma : D(\Gamma) \to \mathbb{R}$

$$\Gamma(\gamma, g, \zeta_0(d,x), v_0(d,x)) := \gamma \sum_{o=0}^{L} \frac{\overline{\alpha}_t^o(\gamma, g, \zeta_0(d,x), v_0(d,x))}{g + v_0^o(d,x)/t} + \Gamma_0(\gamma, g, \zeta_0(d,x), v_0(d,x))$$

for any $(\gamma, g, \zeta_0(d,x), v_0(d,x)) \in D(\Gamma)$.

Remark D.1. *Another possible formulation for $\Gamma(\gamma, g, \zeta_0(d,x), v_0(d,x))$ is $\max_{o \in \{0,...,L\}} \Omega(\gamma, g, \zeta_0^o(d,x), v_0^o(d,x))$, so one can define Γ as the minimum of this expression and the one above, and depending on the context one can use one bound or the other. The same applies to Γ_0 and Ω_0. For the sake of the exposition, however, we do not make this bound explicit.* △

Lemma D.5. *The following properties are true*

1. *$\delta \mapsto \Gamma(\delta, g, \zeta_0(d,x), v_0(d,x))$ is increasing and $\delta \mapsto \Gamma(\delta, g, \zeta_0(d,x), v_0(d,x))$ is non-decreasing.*

2. *$g \mapsto \Gamma(\delta, g, \zeta_0(d,x), v_0(d,x))$ and $g \mapsto \Gamma_0(\delta, g, \zeta_0(d,x), v_0(d,x))$ are decreasing.*

3. *For any positive sequences $(\delta_t, g_t)_t$, $\Gamma(\delta_t, g_t, \zeta_0(d,x), v_0(d,x)) = O\left(\frac{\delta_t + t^{-1}}{g_t + t^{-1}}\right)$ and $\Gamma_0(\delta_t, g_t, \zeta_0(d,x), v_0(d,x)) = O\left(\frac{t^{-1}}{g_t + t^{-1}}\right)$.*

Proof of Lemma D.5. (1) (we only establish the results for Γ as for Γ_0 is analogous) $\Gamma(\delta, g, \zeta_0(d,x), v_0(d,x))$ is the sum of products

$$\overline{\alpha}_t^o(\delta, g, |\zeta_0(d,x)|, v_0(d,x)) \Omega^+(\delta, g, \zeta_0^o(d,x), v_0^o(d,x)) + \underline{\alpha}^o(\delta, g, |\zeta_0(d,x)|, v_0(d,x)) \Omega^-(\delta, g, \zeta_0^o(d,x), v_0^o(d,x)).$$

Observe that, by Lemma D.2(3), $\overline{\alpha}_t^o(\delta, g, |\zeta_0(d,x)|, v_0(d,x))$ is increasing as a function of δ and $\Omega^+(\delta, g, \zeta_0^o(d,x), v_0^o(d,x))$ is non-decreasing as a function of δ by Lemma D.3(1). Since both quantities are positive, it follows that $\overline{\alpha}_t^o(\delta, g, |\zeta_0(d,x)|, v_0(d,x)) \Omega^+(\delta, g, \zeta_0^o(d,x), v_0^o(d,x))$ is non-decreasing (increasing if $\Omega^+(\delta, g, \zeta_0^o(d,x), v_0^o(d,x)) > 0$). Similarly, by Lemmas D.2(3) and D.3(1), $\underline{\alpha}_t^o$ is decreasing and Ω^- is non-positive and non-decreasing

9

as a function of δ, so the product is is non-decreasing (increasing if $\Omega^- < 0$). Thus, $\Gamma(\delta, g, \zeta_0(d,x), v_0(d,x))$ is increasing as a function of δ.

(2) (we only establish the results for Γ as for Γ_0 is analogous) By a similar argument, Lemma D.3(2) and Lemma D.2(4) it follows that $\Gamma(\delta, g, \zeta_0(d,x), v_0(d,x))$ is decreasing as a function of g.

(3) Clearly, $\max_{o \in \{0,...,L\}} \Omega(\delta_t, g_t, \zeta_0(d,x), v_0(d,x)) = O\left(\frac{\delta_t + t^{-1}}{g_t + t^{-1}}\right)$. Thus $\Gamma(\delta_t, g_t, \zeta_0(d,x), v_0(d,x))$ inherits the same rate. Similarly, $\max_{o \in \{0,...,L\}} \Omega_0(g_t, \zeta_0(d,x), v_0(d,x)) = O\left(\frac{t^{-1}}{g_t + t^{-1}}\right)$ and $\Gamma_0(\delta_t, g_t, \zeta_0(d,x), v_0(d,x))$ inherits the same rate. $\qquad \square$

Lemma D.6. *For any $(d,x) \in \mathbb{D} \times \mathbb{X}$ and any $t \in \mathbb{N}$, suppose $|J_t(d,x) - f_t(d,x)\theta(d,x)| \leq \gamma$ and $f_t(d,x) - h_t(d,x) \geq -\eta$ for some $\gamma \geq 0$ and $0 \leq \eta \leq h_t(d,x) \leq \iota_t(d,x)$. Then:*

1. $|\zeta_t^\alpha(d,x) - \theta(d,x)| \leq \Gamma(\gamma, h_t(d,x) - \eta, \zeta_0(d,x) - \theta(d,x), v_0(d,x))$.

Proof of Lemma D.6 . Follows directly from the definition of ζ_t^α and Lemmas D.4 and D.1. $\qquad \square$

D.3.1 Relationship between Γ and Ω

For each $o \in \{0,...,L\}$ and $d \in \mathbb{D}$, let $\zeta_0^{-o}(d,x)$ be the $L \times 1$ vector of all coordinates of $\zeta_0(d,x)$ except for $\zeta_0^o(d,x)$.

Lemma D.7. *For each $o \in \{0,...,L\}$, $(d,x) \in \mathbb{D} \times \mathbb{X}$, $\gamma, g \geq 0$ and $v_0(d,x)$,*

$$\lim_{\bar{\zeta}_0^{-o}(d,x) \to \infty} \left| \Gamma(\gamma, g, |\zeta_0(d,x)|, v_0(d,x)) - \Omega(\gamma, g, |\zeta_0^o(d,x)|, v_0^o(d,x)) \right| = 0$$

Proof of Lemma D.7. By construction of $\bar{\ell}_t$ and $\underline{\ell}_t$ it is easy to see that for any $o' \in \{1,...,L\}$,

$$\lim_{\bar{\zeta}_0^{o'}(d,x) \to \infty} \underline{\ell}_t(\gamma, g, |\zeta_0^{o'}(d,x)|, v_0^{o'}(d,x)) = \lim_{\bar{\zeta}_0^{o'}(d,x) \to \infty} \bar{\ell}_t(\gamma, g, |\zeta_0^{o'}(d,x)|, v_0^{o'}(d,x)) = -\infty$$

Moreover, in both cases the rate is $O(-|\zeta_0^{o'}(d,x)|^2)$ (observe that the Oh depends on $(\gamma, g, v_0^{o'}(d,x))$). Hence, for any $o' \neq o$,

$$\underline{\alpha}_t^{o'}(\gamma, g, |\zeta_0(d,x)|, v_0(d,x)) = O(e^{-|\zeta_0^{o'}(d,x)|^2}) \text{ and } \overline{\alpha}_t^{o'}(\gamma, g, |\zeta_0(d,x)|, v_0(d,x)) = O(e^{-|\zeta_0^{o'}(d,x)|^2}).$$

That is, they converge to 0 at exponential rate.

On the other hand, $\Omega(\gamma,g,|\zeta_0^{o'}(d,x)|,v_0^{o'}(d,x)) = O(|\zeta_0^{o'}(d,x)|)$, hence

$$\overline{\alpha}_t^{o'}(\gamma,g,|\zeta_0(d,x)|,v_0(d,x))\Omega^+(\gamma,g,|\zeta_0^{o'}(d,x)|,v_0^{o'}(d,x))$$

$$+\underline{\alpha}_t^{o'}(\gamma,g,|\zeta_0(d,x)|,v_0(d,x))\Omega^-(\gamma,g,|\zeta_0^{o'}(d,x)|,v_0^{o'}(d,x)) = O(e^{-|\zeta_0^{o'}(d,x)|^2}|\zeta_0^{o'}(d,x)|)$$

which clearly converges to 0 as $|\zeta_0^{o'}(d,x)|$ diverges.

On the other hand, these results imply that

$$\lim_{\zeta_0^{-o}(d,x)\to\infty} \overline{\alpha}_t^o(d,x) = \min\left\{1, \frac{e^{\overline{\ell}_t\gamma,g,|\zeta_0^o(d,x)|,v_0^o(d,x))}}{e^{\underline{\ell}_t\gamma,g,|\zeta_0^o(d,x)|,v_0^o(d,x))}}\right\} = 1$$

where the last equality follows because $\overline{\ell}_t(\gamma,g,|\zeta_0^o(d,x)|,v_0^o(d,x)) \geq \underline{\ell}_t(\gamma,g,|\zeta_0^o(d,x)|,v_0^o(d,x))$.

Therefore $\lim_{\zeta_0^{-o}(d,x)\to\infty} \Gamma(\gamma,g,|\zeta_0(d,x)|,v_0(d,x)) = \Omega(\gamma,g,|\zeta_0^o(d,x)|,v_0^o(d,x))$, as desired. $\qquad\square$

E Concentration Inequalities

Recall that for any $d \in \{0,...,M\}$ and any $t \geq 0$, let $(h_t(d),\omega_t(d)) \in [0,1]^2$ be such that

$$\mathbf{P}(\iota_t(d) \geq h_t(d)) \geq 1 - \omega_t(d),$$

and $\sum_{d=0}^M h_t(d) = 1$, where $\iota_t(d) := t^{-1}\sum_{s=1}^t \delta_s(d)$.

The next lemma presents a Azuma-Hoeffding-type concentration inequality for $(J_t)_t$ and $(f_t)_t$ which are the basis of our theoretical results.

Lemma E.1. *For any $d \in \{0,...,M\}$, any $a \geq 0$ and any $t \geq 0$,*

$$\mathbf{P}\left(\left|t^{-1}\sum_{s=1}^t (Y_s(d) - \theta(d))1\{D_s = d\}\right| \geq a\right) \leq 2e^{-0.5t\frac{a^2}{v\sigma(d)^2}},$$

and

$$\mathbf{P}\left(\left|t^{-1}\sum_{s=1}^t 1\{D_s = d\} - \iota_t(d)\right| \geq a\right) \leq 2e^{-4ta^2}.$$

It readily follows that a common bound is given by $2e^{-0.5t\frac{a^2}{\max\{1/8,v\sigma(d)^2\}}}$.

Remark E.1 (Remarks on Lemma E.1)**.** *We use Assumption 3(i) in the first part of the lemma. in particular, it is used in order to get an upper bound with exponential decay. The assumption, however, could be replaced*

11

by sub-exponential or any other type of control on the MGF of $Y(d)$, e.g., $E[e^{\lambda(Y(d)-\theta(d))}] \leq e^{\kappa(\lambda)}$ for some decreasing function $\lambda \mapsto \kappa(\lambda)$. This change, however, will affect the upper bound obtained in the lemma; it will decay slower than the current one. In fact, up to constant, the result in the lemma will change to

$$\mathbf{P}\left(\left|t^{-1}\sum_{s=1}^{t}(Y_s(d)-\theta(d))1\{D_s=d\}\right| \geq a\right) \leq 2e^{-t\max_{\lambda\geq0}\{a\lambda-\kappa(\lambda)\}}.$$

$\triangle$

Proof of Lemma E.1. Let $W_s(d) := (Y_s(d) - \theta(d))1\{D_s = d\}$. By the Markov inequality, it follows that, for any $\lambda > 0$,

$$\mathbf{P}\left(t^{-1}\sum_{s=1}^{t}W_s(d) \geq a\right) \leq E\left[\prod_{s=1}^{t}e^{\lambda W_s(d)}\right]e^{-a\lambda t}.$$

Observe that

$$E\left[\prod_{s=1}^{t}\exp\{\lambda W_s(d)\}\right] = E\left[\prod_{s=1}^{t-1}\exp\{\lambda W_s(d)\}E_t\left[\exp\{\lambda W_t(d)\}\right]\right]$$

where $E_t[.]$ denotes the conditional expectation under $\mathbf{P}$ given $(Y_s)_{s=1}^{t-1}$ and $(D_s)_{s=1}^{t}$ (but not $Y_t(d)$). Observe that $Y_t(d)$ is independent of past Y's, given D_t. This observation and the fact that $Y_t(d)$ is sub-gaussian (Assumption 3) imply

$$E_t\left[\exp\{\lambda W_t(d)\}\right] = E_t\left[\exp\{\lambda 1\{D_t = d\}(Y_t(d)-\theta(d))\}\right] \leq \exp\{0.5\upsilon\sigma(d)^2 1\{D_t = d\}\lambda^2\} \leq \exp\{0.5\upsilon\sigma(d)^2\lambda^2\}.$$

Iterating in this fashion,

$$E\left[\prod_{s=1}^{t}\exp\{\lambda W_s(d)\}\right] \leq e^{0.5\upsilon\sigma^2(d)\lambda^2}E\left[\prod_{s=1}^{t-1}\exp\{\lambda W_s(d)\}\right] \leq e^{0.5\upsilon\sigma^2(d)t\lambda^2}$$

Therefore, for any $\lambda > 0$

$$\mathbf{P}\left(t^{-1}\sum_{s=1}^{t}W_s(d) \geq a\right) \leq \exp\{0.5\upsilon\sigma^2(d)\lambda^2 t - a\lambda t\}.$$

Choosing $\lambda = a/(\upsilon\sigma^2(d))$, it follows that

$$\mathbf{P}\left(t^{-1}\sum_{s=1}^{t}W_s(d) \geq a\right) \leq \exp\{-0.5t(a^2/(\upsilon\sigma(d)^2))\}.$$

By analogous calculations, it is easy to show that

$$\mathbf{P}\left(|t^{-1}\sum_{s=1}^{t}W_s(d)|\ge a\right)\le 2\exp\{-0.5t(a^2/(\upsilon\sigma(d)^2)\}.$$

Now let $W_t(d) := 1\{D_t = d\} - \delta_t(d)$ and observe that $\left|t^{-1}\sum_{s=1}^{t}W_s(d)\right|\ge a$ implies that either $t^{-1}\sum_{s=1}^{t}1\{D_s = d\} - \iota_t(d)\ge a$ or $(t^{-1}\sum_{s=1}^{t}1\{D_s = d\} - \iota_t(d))\le -a$. We only do the proof for the first case since the second one is analogous.

By the Markov inequality, it follows that, for any $\lambda > 0$,

$$\mathbf{P}\left(t^{-1}\sum_{s=1}^{t}W_s(d)\ge a\right)=E\left[1\{t^{-1}\sum_{s=1}^{t}W_s(d)\ge a\}\right]\le e^{-\lambda at}E\left[\prod_{s=1}^{t}e^{\lambda W_s(d)}\right]=e^{-\lambda at}E\left[\prod_{s=1}^{t-1}e^{\lambda W_s(d)}E_{t-1}[e^{\lambda W_t(d)}]\right]$$

where the last line follows by LIE, where E_{t-1} is the expectation conditional on (Y^{t-1},D^{t-1}).

Given (Y^{t-1},D^{t-1}), $\delta_t(.)$ is non-random as it is measurable with respect to these variables. Thus,

$$E_{t-1}\left[e^{\lambda W_t(d)}\right]=e^{-\lambda\delta_t(d)}\left(\delta_t(d)e^{\lambda}+(1-\delta_t(d))\right)=e^{L(\lambda)}$$

where $L(\lambda)=-\lambda\delta_t(d)+\log\left(\delta_t(d)e^{\lambda}+(1-\delta_t(d))\right)$. Observe that $L(0)=0$, $L'(\lambda)=-\delta_t(d)+\delta_t(d)\frac{e^{\lambda}}{\delta_t(d)e^{\lambda}+(1-\delta_t(d))}$ so that $L'(0)=0$ and $L''(\lambda)=\delta_t(d)\left(\frac{e^{\lambda}(1-\delta_t(d))}{(\delta_t(d)e^{\lambda}+(1-\delta_t(d)))^2}\right)$. The global maximum of L'' is at $\lambda=\log((1-\delta_t(d))/\delta_t(d))$ and thus $L''(\lambda)\le L''(\log((1-\delta_t(d))/\delta_t(d)))=\frac{(1-\delta_t(d))^2}{4(1-\delta_t(d))^2}=0.25$. Therefore, by the Mean Value Theorem,

$$E_{t-1}\left[e^{\lambda W_t(d)}\right]\le e^{L(\lambda)}\le e^{\frac{1}{8}\lambda^2}.$$

Iterating over this, $E\left[\prod_{s=1}^{t}e^{\lambda W_s(d)}\right]\le\prod_{s=1}^{t}e^{\frac{\lambda^2}{8}}=e^{t\frac{\lambda^2}{8}}$. Therefore, for any $\lambda > 0$

$$\mathbf{P}\left(t^{-1}\sum_{s=1}^{t}W_s(d)\ge a\right)\le e^{t\frac{\lambda^2}{8}-a\lambda t}.$$

Choosing $\lambda = 4a$, it follows that $\mathbf{P}\left(t^{-1}\sum_{s=1}^{t}W_s(d)\ge a\right)\le e^{-2ta^2}$. $\qquad\square$

13

F   Appendix for Section 3.1

Recall that for any $d \in \{0, \ldots, M\}$ and $t \geq 0$,

$$\iota_{t+1}(d) := \sum_{s=1}^{t+1} \delta_s(d), \; J_{t+1}(d) := \sum_{s=1}^{t+1} 1\{D_s = d\} Y_s(d)/(t+1), \; and \; f_{t+1}(d) := N_{t+1}(d)/(t+1) = \sum_{s=1}^{t+1} 1\{D_s = d\}/(t+1).$$

We now prove Proposition 3.1.

Proof of Proposition 3.1. Recall that $\bar{\zeta}_t(d) := \zeta_t(d) - \theta(d)$, $\bar{Y}_s(d) := (Y_s(d) - \theta(d))$ and $\bar{J}_t(d) := \sum_{s=1}^{t} 1\{D_s = d\} \bar{Y}_s(d)/t$.

For any t, any $\gamma \geq 0$ and any $\eta \in [0, h_t(d)]$, let $S(t, \gamma) := \{|\bar{J}_t(d)| \leq \gamma\}$, and $R(t, \eta) := \{|f_t(d) - \iota_t(d)| \leq \eta\}$, and $U(t) := \{\iota_t(d) \geq h_t(d)\}$.

Conditional on these sets, by Lemma D.6,

$$|\zeta_t^\alpha(d) - \theta(d)| \leq \Gamma(\gamma, h_t(d) - \eta, \bar{\zeta}_0(d), v_0(d))$$

Therefore, for any $a > 0$,

$$\mathbf{P}\left(|\zeta_t^\alpha(d) - \theta(d)| > a\right) \leq 1\{\Gamma(\gamma, h_t(d) - \eta, \bar{\zeta}_0(d), v_0(d)) > a\} + \mathbf{P}\left(S(t, \gamma)^C\right) + \mathbf{P}\left(R(t, \eta)^C\right) + \mathbf{P}\left(U(t)^C\right).$$

By Lemma E.1 and the definition of exploration structure, it follows that

$$\mathbf{P}\left(S(t, \gamma)^C\right) + \mathbf{P}\left(R(t, \eta)^C\right) + \mathbf{P}\left(U(t, \eta)^C\right) \leq 2\left(e^{-4t\eta^2} + e^{-0.5t \frac{\gamma^2}{v\sigma(d)^2}} + \omega_t(d)\right).$$

We now choose γ, η and a for any $\varepsilon > 0$. Let $\eta = \eta_t = h_t(d)\sqrt{0.25t^{-1}\varepsilon}$, $\gamma = \gamma_t = \sqrt{2t^{-1}v\varepsilon}\sigma(d)$ (which satisfies $\eta \leq h_t(d)$). By Lemma D.5(1), $g \mapsto \Gamma(\gamma_t, g, \bar{\zeta}_0(d), v_0(d))$ is decreasing and since $h_t(d)(1 - \sqrt{0.25t^{-1}\varepsilon}) \geq 0.5h_t(d) \iff t0.5^2 \geq 0.25\varepsilon \iff t \geq \varepsilon$, then $\Gamma(\gamma_t, h_t(d)(1 - \sqrt{0.25t^{-1}\varepsilon}), \bar{\zeta}_0(d), v_0(d)) \leq \Gamma(\gamma_t, 0.5h_t(d), \bar{\zeta}_0(d), v_0(d)) =: a_t = a$. With these choices,

$$\mathbf{P}\left(|\zeta_t^\alpha(d) - \theta(d)| > a_t\right) \leq 2\left(e^{-\varepsilon} + e^{-\varepsilon h_t(d)^2} + \omega_t(d)\right).$$

Since $0 \leq h_t(d) \leq 1$, it follows that

$$\mathbf{P}\left(|\zeta_t^\alpha(d) - \theta(d)| > a_t\right) \leq 4\left(e^{-\varepsilon h_t(d)^2} + \omega_t(d)\right).$$

Re-normalizing ε to $\varepsilon/h_t(d)^2$, the desired result follows. □

14

We now prove Corollary 3.1.

Proof of Corollary 3.1. We can prove the result using limits. For any given $o \neq 0$, let $|\bar{s}_0^o(d)| := \sqrt{v_0^o(d)}|\bar{\zeta}_0^o(d)|$ and let $|\bar{s}_0^{-0}(d)|$ be the $L \times 1$ vector, excluding $|\bar{s}_0^0(d)|$. We consider the limit of this quantity going to ∞.

By Lemma D.7 applied to $o = 0$, for each $\gamma \geq 0$ and $\eta \leq h_t(d)$,

$$\lim_{|\bar{s}_0^{-0}(d)| \to \infty} |\Gamma(\gamma, h_t(d) - \eta, |\bar{\zeta}_0(d)|, v_0(d)) - \Omega(\gamma, h_t(d) - \eta, |\bar{\zeta}_0^o(d)|, v_0^o(d))| = 0.$$

Thus, this result implies that for any given $\delta > 0$, there exists a C such that

$$\Omega(\gamma, h_t(d) - \eta, |\bar{\zeta}_0^o(d)|, v_0^o(d)) \geq \Gamma(\gamma, h_t(d) - \eta, |\bar{\zeta}_0(d)|, v_0(d)) - \delta$$

for any $|\bar{s}_0^{-0}(d)| \geq C$.

The result follows by setting $\eta = 0.5h_t(d)$ and $\gamma = \sqrt{2\upsilon\varepsilon/(h_t(d)^2 t)}\sigma(d)$. $\qquad \square$

G Appendix for Section 3.2

Proposition 3.2 follows from this more general lemma that allows for biased sources. To state this lemma we define, for each $d \in \mathbb{D}$, $\eta_d^* : \mathbb{N} \times [0,1] \times \mathbb{R}_+ \to \mathbb{R}_+ \cup \{+\infty\}$ as follows: For any $(t, h_t(d), \Delta) \in \mathbb{N} \times [0,1] \times \mathbb{R}_+$, if $\Gamma_0(\gamma_t, h_t(d) - \eta, (-1)^{1\{d=M\}}\zeta_0(d), v_0(d)) < 0.5\Delta$ for all η, then we choose $\eta_d^*(t, h_t(d), \Delta) = +\infty$; otherwise,

$$\eta_d^*(t, h_t(d), \Delta) := \max\left\{\eta : \Gamma_0(\gamma_t, h_t(d) - \eta, (-1)^{1\{d=M\}}\zeta_0(d), v_0(d)) \leq 0.5\Delta \text{ and } \eta \leq h_t(d)\right\}$$

and if the set is empty, set $\eta_d^*(t, h_t(d), \Delta) = 0$.

Lemma G.1. *Consider the stopping rule defined in Example 4 with parameters $((\gamma_t)_t, B)$ then for any $t \geq 1$,*

$$\mathbf{P}\left(\max_{d \neq M}\{\zeta_\tau^\alpha(d) - \zeta_\tau^\alpha(M)\} > 0\right) \leq 2\sum_{t=B}^{T}\sum_{d=0}^{M}\left(e^{-0.5t\frac{\gamma_t^2}{\upsilon\sigma(d)^2}} + e^{-4t\eta_d^*(t, h_t(d), \Delta)^2}\right) \tag{G.1}$$

$$+ 1\{\forall d : (-1)^{1\{d=M\}}\bar{\zeta}_0(d) > 0\}\sum_{d=0}^{M}\omega_T(d), \tag{G.2}$$

where $\eta_d^*(t, h_t(d), \Delta) \in \mathbb{R}_+ \cup \{+\infty\}$ is defined in Appendix G and is non-decreasing in t, $h_t(d)$, and Δ; and if $(-1)^{1\{d=M\}}\bar{\zeta}_0(d) \leq 0$, then $\eta_d^*(t, h_t(d), \Delta) = +\infty$.

This lemma shows that the quantity $\eta_d^*(t, h_t(d), \Delta)$, which defines the concentration rate of $f_t(d)$, is key for understanding how the primitives of our setup — i.e. the exploration structure and Δ — affect the upper bound for the probability of a mistake. The upper bound for the probability of a mistake decays

exponentially with t and is non-increasing in $h_t(d)$ and Δ. Intuitively, as the degree of exploration increases, the data become less dependent on the past and thus more informative, resulting in a tighter bound. Also, as Δ becomes more positive, so does the difference between the PM's posteriors, which also decreases the probability of making a mistake.

Proof of Lemma G.1. We divide the proof into several steps.

STEP 1 We first provide a bound

$$\max_d \left\{ \min_{m \neq d} \zeta_t^\alpha(d) - \zeta_t^\alpha(m) - c_t(\gamma_t, d, m) \cap \tau = t \right\} > 0$$

for any t, where, recall that,

$$\tau := \min \left\{ t \geq B : \max_d \left\{ \min_{m \neq d} \zeta_t^\alpha(d) - \zeta_t^\alpha(m) - c_t(\gamma_t, d, m) \right\} > 0 \right\}.$$

The first display implies that

$$\max_d \left\{ \zeta_t^\alpha(d) - \zeta_t^\alpha(M) - c_t(\gamma_t, d, M) \right\} > 0.$$

Thus, the event $\{\max_{d \neq M} \{\zeta_t^\alpha(d) - \zeta_t^\alpha(M)\} > 0 \cap \tau = t\}$ implies

$$\{\max_{d \neq M} \left\{ \zeta_t^\alpha(d) - \zeta_t^\alpha(M) - c_t(\gamma_t, d, M) \right\} > 0\}.$$

Suppose the max is achieved by $d(t) \neq M$, then the above expression is equivalent to $\bar{\zeta}_t^\alpha(d(t)) - \bar{\zeta}_t^\alpha(M) - c_t(\gamma_t, d, M) > \theta(M) - \theta(d(t))$. Since $\theta(M) - \theta(d(t)) \geq \Delta$ — recall, $\Delta := \min_d \theta(M) - \theta(d)$ —, it follows that

$$\{\max_{d \neq M} \left\{ \bar{\zeta}_t^\alpha(d) - \bar{\zeta}_t^\alpha(M) - c_t(\gamma_t, d, M) \right\} > \Delta\}.$$

Observe that

$$c_t(\gamma_t, d, M) =: c_t(\gamma_t, d) + c_t(\gamma_t, M)$$

where $(\gamma, d) \mapsto c_t(\gamma, d) := \gamma \sum_{o=0}^{L} \frac{\alpha_t^o(d)}{f_t(d) + \nu_0^o(d)/t}$.

Thus, the event $\{\max_{d \neq M} \left\{ \bar{\zeta}_t^\alpha(d) - \bar{\zeta}_t^\alpha(M) - c_t(\gamma_t, d, M) \right\} > \Delta\}$, is included in the event

$$\cup_{d \neq M} \{\{\bar{\zeta}_t^\alpha(d) - \bar{\zeta}_t^\alpha(M) - c_t(\gamma_t, d, M)\} > \Delta\} \cap \{\bar{\zeta}_t^\alpha(M) + c_t(\gamma_t, M)) \geq -0.5\Delta\} \cup \{\bar{\zeta}_t^\alpha(M) + c_t(\gamma_t, M)) < -0.5\Delta\}$$

$$= \cup_{d \neq M} \{\bar{\zeta}_t^\alpha(d) > c_t(\gamma_t, d) + 0.5\Delta\} \cup \{\bar{\zeta}_t^\alpha(M) < -(c_t(\gamma_t, M) + 0.5\Delta)\}.$$

16

By the definition of c_t, it follows that for any $d \in \{0, \ldots, M-1\}$,

$$\{\bar{\zeta}_t^\alpha(d) > c_t(\gamma_t, d) + 0.5\Delta\} \subseteq \{\bar{\zeta}_t^\alpha(d) > c_t(\gamma_t, d) + 0.5\Delta\} \cap \mathcal{J}_t(\gamma_t, d) \cup \mathcal{J}_t(\gamma_t, d)^C$$

$$\subseteq \left\{ \sum_{o=0}^{L} \alpha_t^o(d) \frac{\bar{\zeta}_0^o(d) v_0^o(d)/t}{f_t(d) + v_0^o(d)/t} > 0.5\Delta \right\} \cap \mathcal{J}_t(\gamma_t, d) \cup \mathcal{J}_t(\gamma_t, d)^C,$$

where $\mathcal{J}_t(\gamma, d) := \{|\bar{J}_t(d)| \le \gamma\}$ for any $\gamma > 0$. Similarly,

$$\{\bar{\zeta}_t^\alpha(M) > -(c_t(\gamma_t, M) + 0.5\Delta)\} \subseteq \left\{ \sum_{o=0}^{L} \alpha_t^o(M) \frac{(-\bar{\zeta}_0^o(M)) v_0^o(M)/t}{f_t(M) + v_0^o(M)/t} > 0.5\Delta \right\} \cap \mathcal{J}_t(\gamma_t, M) \cup \mathcal{J}_t(\gamma_t, M)^C.$$

STEP 2. We now bound

$$\mathbf{P}\left(\max_{d \neq M} \{\zeta_\tau^\alpha(d) - \zeta_\tau^\alpha(M)\} > 0 \right)$$

when $\forall d \colon (-1)^{1\{d=M\}} \bar{\zeta}_0(d) \le 0$. By the union bound and Step 1,

$$\mathbf{P}\left(\max_{d \neq M} \{\zeta_\tau^\alpha(d) - \zeta_\tau^\alpha(M)\} > 0 \right) \le \sum_{t=B}^{T} \mathbf{P}\left(\max_{d \neq M} \{\bar{\zeta}_t^\alpha(d) - \bar{\zeta}_t^\alpha(M)\} > \Delta \right)$$

$$\le \sum_{t=B}^{T} \mathbf{P}\left(\cup_d \left\{ \sum_{o=0}^{L} \alpha_t^o(d) \frac{(-1)^{1\{d=M\}} \bar{\zeta}_0^o(d) v_0^o(d)/t}{f_t(d) + v_0^o(d)/t} > 0.5\Delta \right\} \cap \mathcal{J}_t(\gamma_t, d) \right)$$

$$+ \sum_{t=B}^{T} \mathbf{P}(\mathcal{J}_t(\gamma_t, d)^C).$$

By the assumption that $\forall d \colon (-1)^{1\{d=M\}} \bar{\zeta}_0(d) \le 0$, the first term in the RHS is 0. So the result follows from Lemma E.1.

STEP 3. We now bound $\mathbf{P}\left(\max_{d \neq M} \{\zeta_\tau^\alpha(d) - \zeta_\tau^\alpha(M)\} > 0 \right)$ when $\forall d \colon (-1)^{1\{d=M\}} \bar{\zeta}_0(d) \le 0$ *does not* hold.

Let $\mathcal{V}(T, d) := \{\forall t \le T \colon \iota_t(d) \ge h_t(d)\}$ and observe that

$$\mathbf{P}\left(\max_{d \neq M} \{\zeta_\tau^\alpha(d) - \zeta_\tau^\alpha(M)\} > 0 \right) \le \mathbf{P}\left(\max_{d \neq M} \{\zeta_\tau^\alpha(d) - \zeta_\tau^\alpha(M)\} > 0 \cap \mathcal{V}(T, d) \right) + \mathbf{P}\left(\mathcal{V}(T, d)^C \right)$$

$$\le \mathbf{P}\left(\max_{d \neq M} \{\zeta_\tau^\alpha(d) - \zeta_\tau^\alpha(M)\} > 0 \cap \mathcal{V}(T, d) \right) + \omega_T(d)$$

where, recall that,

$$\tau := \min\left\{ t \ge B \colon \max_d \left\{ \min_{m \neq d} \zeta_t^\alpha(d) - \zeta_t^\alpha(m) - c_t(\gamma_t, d, m) \right\} > 0 \right\}.$$

17

We now bound $\mathbf{P}\left(\max_{d\neq M}\{\zeta_\tau^\alpha(d)-\zeta_\tau^\alpha(M)\}>0\cap\mathcal{V}(T,d)\right)$. This probability can be bounded by

$$\sum_{t=B}^{T}\mathbf{P}\left(\max_{d\neq M}\sum_{o=0}^{L}\alpha_t^o(d)\zeta_t^o(d)-\sum_{o=0}^{L}\alpha_t^o(M)\zeta_t^o(M)>0\cap\tau=t\cap\mathcal{V}(T,d)\right),$$

where $\{\max_{d\neq M}\sum_{o=0}^{L}\alpha_t^o(d)\zeta_t^o(d)-\sum_{o=0}^{L}\alpha_t^o(M)\zeta_t^o(M)>0\cap\tau=t\}$ is the event wherein the experiment is stopped at time t but one choose a treatment that is not M (recall that by construction, M is the treatment with highest expected outcome).

The fact that $\tau=t$ implies that $\max_d\left\{\min_{m\neq d}\zeta_t^\alpha(d)-\zeta_t^\alpha(m)-c_t(\gamma_t,d,m)\right\}>0$. By Step 1, it follows that for any $d\in\{0,...,M-1\}$,

$$\{\bar\zeta_t^\alpha(d)>c_t(\gamma_t,d)+0.5\Delta\}\subseteq\{\bar\zeta_t^\alpha(d)>c_t(\gamma_t,d)+0.5\Delta\}\cap\mathcal{J}_t(\gamma_t,d)\cup\mathcal{J}_t(\gamma_t,d)^C$$

$$\subseteq\left\{\sum_{o=0}^{L}\alpha_t^o(d)\frac{\bar\zeta_0^o(d)v_0^o(d)/t}{f_t(d)+v_0^o(d)/t}>0.5\Delta\right\}\cap\mathcal{J}_t(\gamma_t,d)\cap\mathcal{E}_t(\eta,d)\cup\mathcal{J}_t(\gamma_t,d)^C\cup\mathcal{E}_t(\eta,d)^C$$

where $\mathcal{J}_t(\gamma,d):=\{|\bar{J}_t(d)|\leq\gamma\}$ and $\mathcal{E}_t(\eta,d):=\{|f_t(d)-\iota_t(d)|\leq\eta\}$ for any $t\in\mathbb{N}$ and any $\eta,\gamma>0$. Similarly,

$$\{\bar\zeta_t^\alpha(M)>-(c_t(\gamma_t,M)+0.5\Delta)\}\subseteq\{\sum_{o=0}^{L}\alpha_t^o(M)\frac{(-\bar\zeta_0^o(M))v_0^o(M)/t}{f_t(M)+v_0^o(M)/t}>0.5\Delta\}\cap\mathcal{J}_t(\gamma_t,M)\cap\mathcal{E}_t(\eta,M)\cup\mathcal{J}_t(\gamma_t,M)^C\cup\mathcal{E}_t(\eta,M)^C.$$

Observe that $\mathcal{E}_t(\eta,d)$ and $\mathcal{V}(T,d)\subseteq\mathcal{V}(t,d)$, imply $\{f_t(d)\geq h_t(d)-\eta\}$. This fact and Lemmas D.4(2) and D.1, imply that under $\mathcal{J}_t(\gamma_t,d)\cap\{f_t(d)\geq h_t(d)-\eta\}$, it follows that for any $d\in\mathbb{D}$,

$$\alpha_t^o(d)\frac{\bar\zeta_0^o(d)v_0^o(d)/t}{f_t(d)+v_0^o(d)/t}\leq\alpha_t^o(d)\Omega_0(h_t(d)-\eta,(-1)^{d=M}\bar\zeta_0^o(d),v_0^o(d))$$

$$\leq\bar\alpha_t^o(\gamma_t,h_t(d)-\eta,|\zeta_0(d)|,v_0(d))\Omega_0^+(h_t(d)-\eta,(-1)^{d=M}\bar\zeta_0^o(d),v_0^o(d))$$
$$+\underline\alpha_t^o(\gamma_t,h_t(d)-\eta,|\zeta_0(d)|,v_0(d))\Omega_0^-(h_t(d)-\eta,(-1)^{d=M}\bar\zeta_0^o(d),v_0^o(d)),$$

where the RHS coincides with Γ_0 defined above. Thus, for any $d\in\mathbb{D}$,

$$\left\{\sum_{o=0}^{L}\alpha_t^o(d)\frac{(-1)^{d=M}\bar\zeta_0^o(d)v_0^o(d)/t}{f_t(d)+v_0^o(d)/t}>0.5\Delta\right\}\cap\mathcal{J}_t(\gamma_t,d)\cap\mathcal{E}_t(\eta,d)$$
$$\subseteq\left\{\Gamma_0(\gamma_t,h_t(d)-\eta,(-1)^{1\{d=M\}}\bar\zeta_0(d),v_0(d))>0.5\Delta\right\}\cap\mathcal{J}_t(\gamma_t,M)\cap\mathcal{E}_t(\eta,M).$$

Let $\mathcal{U}_d(\gamma_t, h_t(d) - \eta, \Delta) := \{\Gamma_0(\gamma_t, h_t(d) - \eta, (-1)^{1\{d=M\}}\bar{\zeta}_0(d), v_0(d)) > 0.5\Delta\}$. It thus follows that

$$\sum_{t=B}^{T} \mathbf{P}\left(\max_{d \neq M}\{\bar{\zeta}_t^\alpha(d) - \bar{\zeta}_t^\alpha(M) - c_t(\gamma_t, d, M)\} > \Delta \cap \mathcal{V}(T, d)\right)$$

$$\leq \sum_{t=B}^{T} \mathbf{P}\left(\cup_d \mathcal{U}_d(\gamma_t, h_t(d) - \eta, \Delta) \cup \cup_d \mathcal{J}_t(\gamma_t, d)^C \cup \cup_d \mathcal{E}_t(\eta, d)^C \cap \mathcal{V}(T, d)\right)$$

$$\leq \sum_{t=B}^{T} 1\{\cup_d \mathcal{U}_d(\gamma_t, h_t(d) - \eta, \Delta)\} + \mathbf{P}\left(\cup_d \mathcal{J}_t(\gamma_t, d)^C\right) + \mathbf{P}\left(\cup_d \mathcal{E}_t(\eta, d)^C\right) \tag{G.3}$$

where the second inequality follows from the union bound.

We now choose η as follows. If $\Gamma_0(\gamma_t, h_t(d) - \eta, (-1)^{1\{d=M\}}\bar{\zeta}_0(d), v_0(d)) < 0.5\Delta$ for all η, then we choose $\eta_d^*(t, h_t(d), \Delta) = +\infty$; otherwise,

$$\eta_d^*(t, h_t(d), \Delta) := \max\left\{\eta : \Gamma_0(\gamma_t, h_t(d) - \eta, (-1)^{1\{d=M\}}\bar{\zeta}_0(d), v_0(d)) \leq 0.5\Delta \text{ and } \eta \leq h_t(d)\right\}$$

and if the set is empty, set $\eta_d^*(t, h_t(d), \Delta) = 0$.

If $\eta_d^*(t, h_t(d), \Delta) = 0$, the expression G.3 yields the trivial bound of 1. The expression in the proposition also implies an upper bound greater than 1 (since $\eta_d^*(t, h_t(d), \Delta) = 0$). Thus the proposition is proven. We now study the case if $\eta_d^*(t, h_t(d), \Delta) > 0$ which is more involved. Under this choice of η, it follows that

$$\sum_{t=B}^{T} \mathbf{P}\left(\max_{d \neq M}\{\bar{\zeta}_t^\alpha(d) - \bar{\zeta}_t^\alpha(M) - c_t(\gamma_t, d, M)\} > \Delta\right) \leq \sum_{t=B}^{T} \mathbf{P}\left(\cup_d \mathcal{J}_t(\gamma_t, d)^C\right) + \sum_{t=B}^{T} \mathbf{P}\left(\cup_d \mathcal{E}_t(\eta_d^*(t, h_t(d), \Delta), d)^C\right).$$

By Lemma E.1, it follows that

$$\sum_{t=B}^{T} \mathbf{P}\left(\max_{d \neq M}\{\bar{\zeta}_t^\alpha(d) - \bar{\zeta}_t^\alpha(M) - c_t(\gamma_t, d, M)\} > \Delta\right) \leq 2\sum_{t=B}^{T}\sum_{d=0}^{M}\left(e^{-0.5t\frac{\gamma_t^2}{v\sigma(d)^2}} + e^{-4t\eta_d^*(t, h_t(d), \Delta)^2}\right).$$

We conclude the proof by showing some properties of η_d^*. First, $t \mapsto \eta_d^*(t, h_t(d), \Delta)$ is non-decreasing. To show this, first note that $\Gamma_0(\gamma_t, h_t(d) - \eta, (-1)^{1\{d=M\}}\bar{\zeta}_0(d), v_0(d))$ is (implicitly) a function of t and thus it suffices to show it is non-increasing (for a fixed $h_t(d)$) and $\eta \mapsto \Gamma_0(\gamma_t, h_t(d) - \eta, (-1)^{1\{d=M\}}\bar{\zeta}_0(d), v_0(d))$ is non-decreasing. This follows from Lemma D.5(2) and the fact that g (in that lemma) equals $h_t(d) - \eta$. ; we now show the former.

By construction of Γ_0 it suffices to show that $t \mapsto K_t(\gamma_t, h_t(d) - \eta) := \overline{\alpha}_t^o(\gamma_t, h_t(d) - \eta, |\bar{\zeta}_0(d)|, v_0(d))\Omega_0^+(h_t(d) - \eta, (-1)^{d=M}\bar{\zeta}_0^o(d), v_0^o(d)) + \underline{\alpha}_t^o(\gamma_t, h_t(d) - \eta, |\bar{\zeta}_0(d)|, v_0(d))\Omega_0^-(h_t(d) - \eta, (-1)^{d=M}\bar{\zeta}_0^o(d), v_0^o(d))$ is non-increasing for each o. If $(-1)^{d=M}\bar{\zeta}_0^o(d) \geq 0$ then Ω_0 is positive and decreasing as a function of t (see its definition). In

addition, $\underline{\sigma}_t^o$ and $\overline{\sigma}_t^o$ are non-increasing and decreasing in t resp. Hence $\overline{\alpha}_t^o$ is decreasing in t and positive. Thus, $K_t(\gamma, h_t(d) - \eta) = \overline{\alpha}_t^o(\gamma, h_t(d) - \eta, |\overline{\zeta}_0(d)|, v_0(d))\Omega_0(h_t(d) - \eta, (-1)^{d=M}\overline{\zeta}_0^o(d), v_0^o(d))$ is decreasing. In addition, by the proof of Lemma D.5(1) it follows that $\gamma \mapsto K_t(\gamma, h_t(d) - \eta)$ is increasing and since $t \mapsto \gamma_t$ is non-increasing it follows that $t \mapsto K_t(\gamma_t, h_t(d) - \eta)$ is non-increasing for this case. If $(-1)^{d=M}\overline{\zeta}_0^o(d) \leq 0$ then Ω_0 is negative and decreasing as a function of t (see its definition). Also, $\underline{\alpha}_t^o$ is increasing in t (see Lemma D.2(5)) and positive. Thus, $t \mapsto K_t(\gamma, h_t(d) - \eta) = \underline{\alpha}_t^o(\gamma_t, h_t(d) - \eta, |\overline{\zeta}_0(d)|, v_0(d))\Omega_0^-(h_t(d) - \eta, (-1)^{d=M}\overline{\zeta}_0^o(d), v_0^o(d))$ is decreasing in this case. We thus showed that $t \mapsto \Gamma_0(\gamma_t, h_t(d) - \eta, (-1)^{1\{d=M\}}\overline{\zeta}_0(d), v_0(d))$ is non-increasing.

Second, $\Delta \mapsto \eta_d^*(t, h_t(d), \Delta)$ is non-decreasing. To show this is sufficient to show that $\eta \mapsto \Gamma_0(\gamma_t, h_t(d) - \eta, (-1)^{1\{d=M\}}\overline{\zeta}_0(d), v_0(d))$ is non-decreasing. This follows from Lemma D.5(2) and the fact that g (in that lemma) equals $h_t(d) - \eta$.

Third, $h_t(d) \mapsto \eta_d^*(t, h_t(d), \Delta)$ is non-decreasing. As before, this follows from Lemma D.5(2). $\qquad\square$

The proof of Corollary 3.2 follows from this more general lemma that allows for biased sources.

Lemma G.2. *Suppose all the conditions of Lemma G.1 hold and* $\frac{|\zeta_0^0(d) - \theta(d)|v_0^0(d)/t}{h_t(d) + v_0^0(d)/t} \leq 0.5\Delta.$[23] *Then, for any* $\varepsilon > 0$, *there exists a* C *such that for all* $\min_{o \neq 0}|\zeta_0^o(d) - \theta(d)| \geq C$, *it follows that*

$$\mathbf{P}\left(\max_{\substack{d \neq M}}\{\zeta_\tau^\alpha(d) - \zeta_\tau^\alpha(M)\} > 0\right) \leq \sum_{d=0}^{M}\sum_{t=B}^{T}(2e^{-\frac{0.5t(\gamma_t)^2}{v\sigma(d)^2}} + e^{-4t(\eta_d^{oracle}(t,h_t(d),(\Delta-\varepsilon)/(1+\varepsilon)))^2}) + 1\{\forall d: (-1)^{1\{d=M\}}\overline{\zeta}_0(d) \leq 0\}\omega_T(d).$$

where $\eta_d^{oracle}(t, h_t(d), (\Delta - \varepsilon)/(1 + \varepsilon))$ *is defined as*

$$\max\left\{\eta: \frac{|\zeta_0^0(d) - \theta(d)|v_0^0(d)/t}{h_t(d) - \eta + v_0^0(d)/t} \leq 0.50\frac{\Delta - \varepsilon}{1 + \varepsilon} \text{ and } \eta \leq \epsilon\right\}.$$

Moreover, if $\overline{\zeta}_0^0(d) = 0$, *then* $\eta_d^{oracle}(t, h_t(d), \Delta) = \infty$.

The behavior of η_d^* determines whether the upper bound embodies an oracle property similar to the one we demonstrated for the concentration rates. Given the properties of the weights illustrated in Proposition 2.1 and Lemma 2.1, it is easy to show that if sources other than $o = 0$ are sufficiently stubborn, then η_d^* becomes arbitrarily close to η_d^{oracle}, where η_d^{oracle} is defined as the largest η such that $\frac{|\zeta_0^0(d) - \theta(d)|v_0^0(d)/t}{h_t(d) - \eta + v_0^0(d)/t} \leq 0.5\Delta$, which is the relevant quantity determining the probability of mistake for the *least stubborn* source. It then follows that the bound obtained in Proposition 3.2 would be arbitrary close to the oracle one; the corollary below formalizes this discussion.

Proof of Lemma G.2. We only prove the result for the case where $(-1)^{1\{d=M\}}\overline{\zeta}_0(d) \leq 0$ does not hold (the proof for the other case is analogous).

---

[23]This last condition always holds for sufficiently small biases or for large values of t.

By the same calculations as those in Step 3 of the proof of Proposition 3.2, for all $d \in \mathbb{D}$,

$$\{\bar{\zeta}_t^\alpha(d) > (-1)^{1\{d=M\}}(c_t(\gamma_t,d)+0.5\Delta)\} \cap \mathcal{V}(T,d)$$

$$\subseteq \left\{\sum_{o=0}^L \alpha_t^o(d)\frac{(-1)^{1\{d=M\}}\bar{\zeta}_0^o(d)v_0^o(d)/t}{f_t(d)+v_0^o(d)/t} > 0.5\Delta\right\} \cap \mathcal{J}_t(\gamma_t,d) \cap \mathcal{E}_t(\eta,d) \cap \mathcal{V}(T,d)$$

$$\cup \mathcal{J}_t(\gamma_t,d)^C \cup \mathcal{E}_t(\eta,d)^C.$$

Observe that $\mathcal{E}_t(\eta,d)$ and $\mathcal{V}(T,d) \subseteq \mathcal{V}(t,d)$, imply $\{f_t(d) \geq h_t(d)-\eta\}$. This fact, Lemma D.1 and the proof of Lemma D.7 imply that under $\mathcal{J}_t(\gamma_t,d) \cap \{f_t(d) \geq h_t(d)-\eta\}$, $\lim_{|\bar{\zeta}_0^{-0}(d)|\to\infty} \alpha_t^o(d)\max\{|\bar{\zeta}_0^o(d)|,1\} = 0$ a.s., for all $o \neq 0$. Since the $(\alpha_t^o(d))_{o=0}^L$ sum to one, this implies that $\lim_{|\bar{\zeta}_0^{-0}(d)|\to\infty} \alpha_t^0(d) = 1$ a.s.

Hence, for any $\varepsilon > 0$, there exists a C such that, if $|\bar{\zeta}_0^{-0}(d)| \geq C$, then

$$\{\bar{\zeta}_t^\alpha(d) > (-1)^{1\{d=M\}}(c_t(\gamma_t,d)+0.5\Delta)\}$$

$$\subseteq \left\{(1+\varepsilon)\frac{|\bar{\zeta}_0^0(d)|v_0^o(d)/t}{f_t(d)+v_0^o(d)/t} + \varepsilon > 0.5\Delta\right\} \cap \mathcal{J}_t(\gamma_t,d) \cap \mathcal{E}_t(\eta,d) \cup \mathcal{J}_t(\gamma_t,d)^C \cup \mathcal{E}_t(\eta,d)^C.$$

The rest of the proof follows the same steps as the proof of Proposition 3.2, but instead of using η_d^*, we use

$$\eta_d^{oracle}(t,h_t(d),(\Delta-\varepsilon)/(1+\varepsilon)) := \max\left\{\eta: \frac{|\zeta_0^0(d)-\theta(d)|v_0^0(d)/t}{h_t(d)-\eta+v_0^0(d)/t} \leq 0.50\frac{\Delta-\varepsilon}{1+\varepsilon} \text{ and } \eta \leq \epsilon\right\}.$$

Finally, if $|\zeta_0^0(d)-\theta(d)| = 0$, then for any $\varepsilon < 0.5\Delta$, $\{\bar{\zeta}_t^\alpha(d) > (-1)^{1\{d=M\}}(c_t(\gamma_t,d)+0.5\Delta)\} \subseteq \mathcal{J}_t(\gamma_t,d)^C \cup \mathcal{E}_t(\eta,d)^C$, so one can set $\eta_d^{oracle} = \infty$ and obtain that $\{\bar{\zeta}_t^\alpha(d) > (-1)^{1\{d=M\}}(c_t(\gamma_t,d)+0.5\Delta)\} \subseteq \mathcal{J}_t(\gamma_t,d)^C$. This result implies that there exists a C (the one constant corresponding to any $\varepsilon \leq 0.5\Delta$) such that, if $|\bar{\zeta}_0^{-0}(d)| \geq C$, then

$$\mathbf{P}\left(\max_{d \neq M}\{\zeta_\tau^\alpha(d) - \zeta_\tau^\alpha(M)\} > 0\right) \leq 2\sum_{d=0}^M \sum_{t=B}^T e^{-0.5t\frac{(\gamma_t)^2}{v\sigma(d)^2}}$$

$\square$

The proof of Corollary 3.3 follows from this more general Lemma that allows for biased sources.

Lemma G.3. *Suppose all the conditions of Proposition 3.2 hold, and, for any t, $\gamma_t \geq \sqrt{\frac{\log t}{t}}\sqrt{A\log t}$ with*

(A, B) *such that* $\log B \geq \max_d 2\upsilon\sigma(d)^2,$ $\frac{4\min_{d\in\mathbb{D}}(\eta_d^*(t,h_t(d),\Delta))^2}{A} \geq \frac{\log t}{t}$ *for all* $t \geq B$, *and*

$$\frac{3(M+1)}{A-1}(B^{-(A-1)} - T^{-(A-1)}) \leq \beta. \tag{G.4}$$

Then

$$\mathbf{P}\left(\max_{d\neq M}\{\zeta_\tau^\alpha(d) - \zeta_\tau^\alpha(M)\} > 0\right) \leq \beta.$$

Proof of Lemma G.3. We do the proof for where the expression for γ holds with equality. We do this because if the desired bound holds for this case, it will hold for any γ that is greater. By Lemma G.1

$$\mathbf{P}\left(\max_{d\neq M}\{\zeta_\tau^\alpha(d) - \zeta_\tau^\alpha(M)\} > 0\right) \leq \sum_{d=0}^{M}\sum_{t=B}^{T}(2e^{-0.5t\frac{(\gamma)^2}{\upsilon\sigma(d)^2}} + e^{-4t(\eta_d^*(t,h_t(d),\Delta))^2}). \tag{G.5}$$

By our choice of γ, the first term in the RHS is less or equal than $2t^{-A}$. We now need to check that for all $t \geq B(\beta)$, $e^{-4t(\eta_d^*(t,h_t(d),\Delta))^2} \leq t^{-A} \iff \frac{4(\eta_d^*(t,h_t(d),\Delta))^2}{A} \geq \frac{\log t}{t}$. Since $t \mapsto \frac{\log t}{t}$ is decreasing (as long as $\log B \geq 2$) and $\eta_d^*(t,h_t(d),\Delta) \geq \eta^*(B,h_t(d),\Delta)$, it thus suffices to check that $\frac{(\eta_d^*(B,h_t(d),\Delta))^2}{A} \geq \frac{\log B}{4B}$ which holds by assumption. $\qquad\square$

The next proposition provides bounds on the probability of making a mistake in any instance t, and how it depends on the prior of the model and the exploration structure. These results are used in Appendix H to establish the concentration rate of the average outcomes. The proof is relegated to the Supplemental Material K.

Proposition G.1. *For any* $t \in \mathbb{N}$ *and any* $d \in \mathbb{D}$, *suppose there exists a* $\gamma \leq h_t(d)$ *such that* $\max_{o\in\{0,...,L\}}\Omega_0(h_t(d) - \gamma, (-1)^{1\{d=M\}}\zeta_0^o(d), v_0^o(d)) \leq 0.5\Delta$, *then*

$$\mathbf{P}\left(\exists d \neq M : \zeta_t^\alpha(d) - \zeta_t^\alpha(M) > \varepsilon\right) \leq 4\sum_{d=0}^{M}\left(e^{-0.5t\frac{(H^{-1}(0.5(\Delta+\varepsilon),h_t(d),(-1)^{1\{d=M\}}\bar\zeta_0(d),v_0(d)))^2}{\max\{1/8,\upsilon\sigma(d)^2\}}} + \omega_t(d)\right),$$

where

$$\gamma \mapsto H(\gamma, h_t(d), (-1)^{1\{d=M\}}\bar\zeta_0(d), v_0(d)) := \max_{o\in\{0,...,L\}}\Omega_0(h_t(d) - \gamma, (-1)^{1\{d=M\}}\zeta_0^o(d), v_0^o(d)).$$

H   Appendix for Section 3.3

To show Proposition 3.3 we use the following lemmas.

Lemma H.1. *For any $t \in \{1,...,T\}$ and any $\gamma > 0$,*

$$\mathbf{P}\left(\max_d \theta(d) - t^{-1}\sum_{s=1}^{t} Y_s > -\sqrt{\frac{\gamma}{t}}\left(\sqrt{2\upsilon}\sigma(d) + \frac{||\theta||_1}{2}\right) + \max_d \theta(d) - \sum_{d=0}^{M}\theta(d)\iota_t(d)\right) \leq 4e^{-\gamma}.$$

Let $(1 - \omega_t)_t$ be any likelihood of exploration associated to the policy rule, $\delta_t(d) = \Xi_t(M+1)^{-1} + (1 - \Xi_t)1\{d = \arg\max_a \zeta_{t-1}^\alpha(a)\}$ for any $t \in \{1,...,T\}$.

Lemma H.2. *Suppose $\delta_t(d) = \Xi_t(M+1)^{-1} + (1 - \Xi_t)1\{d = \arg\max_a \zeta_{t-1}^\alpha(a)\}$ for any $t \in \{1,...,T\}$. Then, for any $t \in \{1,...,T\}$ and any $\gamma > 0$,*

$$\mathbf{P}\left(\max_d \theta(d) - t^{-1}\sum_{s=1}^{t} Y_s > \sqrt{\frac{\gamma}{t}}\left(\sqrt{2\upsilon}\sigma(d) + \frac{||\theta||_1}{2}\right) - ||\theta||_1\left(\sqrt{1 - \bar{\Xi}_t}\sqrt{e^\gamma \Lambda_t(\Delta)} + \frac{\bar{\Xi}_t}{M+1}\right)\right) \leq 5e^{-\gamma}.$$

where $H(\gamma, h_t(d), \bar{\zeta}_0(d), v_0(d)) := \max_{o \in \{0,...,L\}} \frac{|\bar{\zeta}_0^o(d)|v_0^o(d)/t}{h_t(d) - \gamma + v_0^o(d)/t}$ and

$$\Lambda_t(\Delta) := 4\sum_{d=0}^{M} t^{-1}\sum_{s=1}^{t}\left(e^{-0.5(s-1)\frac{H^{-1}(0.5\Delta, h_t(d), \bar{\zeta}_0(d), v_0(d))}{\max\{1/8, \upsilon\sigma(d)^2\}}} + \omega_{s-1}(d)\right).$$

Moreover, a feasible structure of exploration is given by $h_t(.) = \bar{\Xi}_t/(M+1)$ and $\omega_t(.) = 0$, so the previous bound can be specialized to this case.

Remark H.1. *The function H is an upper bound for the function Γ_0 and so, it can be replace by the latter function without altering the proof. This change will obviously provide tighter bounds but also yield a more cumbersome expression, thus, for the sake of presentation, we choose to work with H.* △

Proof of Proposition 3.3. The result follows by Lemma H.2. □

H.1 Proofs of Lemmas

Proof of Lemma H.1. Observe that $t^{-1}\sum_{s=1}^{t} Y_s = \sum_{d=0}^{M} t^{-1}\sum_{s=1}^{t} Y_s(D_s)1\{D_s = d\}$, and thus

$$t^{-1}\sum_{s=1}^{t} Y_s - \max_d \theta(d) = \left(\sum_{d=0}^{M} t^{-1}\sum_{s=1}^{t} 1\{D_s = d\}(Y_s(d) - \theta(d))\right) + \left(\sum_{d=0}^{M} \theta(d)(f_t(d) - \iota_t(d))\right) + \sum_{d=0}^{M}\theta(d)\iota_t(d) - \max_d \theta(d)$$

$$=:Term1 + Term2 + Term3.$$

Therefore, to obtain the desired result we just need to bound

$$\mathbf{P}(|Term_1| > \Sigma_1(\gamma,t)) + \mathbf{P}(|Term_2| > ||\theta||_1\Sigma_2(\gamma,t)).$$

23

By Lemma E.1, $\mathbf{P}(|Term_1| > \Sigma_1(\gamma,t)) \le 2e^{-\gamma}$ with $\Sigma_1(\gamma,t) = \sqrt{\frac{2\upsilon\gamma}{t}}\sigma(d)$ and $\mathbf{P}(|Term_2| > ||\theta||_1\Sigma_2(\gamma,t)) \le 2e^{-\gamma}$ with $\Sigma_2(\gamma,t) = \sqrt{\frac{\gamma}{4t}}$.

$\qquad\square$

Proof of Lemma H.2. By Lemma H.1,

$$\mathbf{P}\left(\max_d \theta(d) - t^{-1}\sum_{s=1}^t Y_s > -\sqrt{\frac{\gamma}{t}}\left(\sqrt{2\upsilon}\sigma(d) + \frac{||\theta||_1}{2}\right) + \max_d \theta(d) - \sum_{d=0}^M \theta(d)\iota_t(d)\right) \le 4e^{-\gamma}.$$

Thus,

$$\mathbf{P}\left(\max_d \theta(d) - t^{-1}\sum_{s=1}^t Y_s > -\sqrt{\frac{\gamma}{t}}\left(\sqrt{2\upsilon}\sigma(d) + \frac{||\theta||_1}{2}\right) - ||\theta||_1\left(\sqrt{t^{-1}\sum_{s=1}^t(1-\Xi_s)^2}\sqrt{e^\gamma\Lambda_t(\Delta)} + \frac{\bar{\Xi}_t}{M+1}\right)\right)$$

$$\le \mathbf{P}\left(\max_d \theta(d) - t^{-1}\sum_{s=1}^t Y_s > -\sqrt{\frac{\gamma}{t}}\left(\sqrt{2\upsilon}\sigma(d) + \frac{||\theta||_1}{2}\right) + \max_d \theta(d) - \sum_{d=0}^M \theta(d)\iota_t(d)\right)$$

$$+ \mathbf{P}\left(\max_d \theta(d) - \sum_{d=0}^M \theta(d)\iota_t(d) \le -||\theta||_1\left(\sqrt{t^{-1}\sum_{s=1}^t(1-\Xi_s)^2}\sqrt{e^\gamma\Lambda_t(\Delta)} + \frac{\bar{\Xi}_t}{M+1}\right)\right)$$

$$\le 4e^{-\gamma} + \mathbf{P}\left(\max_d \theta(d) - \sum_{d=0}^M \theta(d)\iota_t(d) \le -||\theta||_1\left(\sqrt{t^{-1}\sum_{s=1}^t(1-\Xi_s)^2}\sqrt{e^\gamma\Lambda_t(\Delta)} + \frac{\bar{\Xi}_t}{M+1}\right)\right)$$

So it suffices to bound $\mathbf{P}\left(\max_d \theta(d) - \sum_{d=0}^M \theta(d)\iota_t(d) \le -||\theta||_1\left(\sqrt{t^{-1}\sum_{s=1}^t(1-\Xi_s)^2}\sqrt{e^\gamma\Lambda_t(\Delta)} + \frac{\bar{\Xi}_t}{M+1}\right)\right)$. For this, note that

$$\sum_{d=0}^M \theta(d)t^{-1}\sum_{s=1}^t \mathbb{1}\{d = \arg\max_a \theta(a)\} - \sum_{d=0}^M \theta(d)\iota_t(d)$$

$$= t^{-1}\sum_{s=1}^t(1-\Xi_s)\sum_{d=0}^M \theta(d)(\mathbb{1}\{d = \arg\max_a \theta(a)\} - \mathbb{1}\{d = \arg\max_a \zeta_{s-1}^\alpha(d)\}) + \bar{\Xi}_t\left(\max_d \theta(d) - \sum_{d=0}^M \frac{\theta(d)}{M+1}\right).$$

For the first term in the RHS it follows that for each $s \ge 1$,

$$(1-\Xi_s)\sum_{d=0}^M \theta(d)(\mathbb{1}\{d = \arg\max_a \theta(a)\} - \mathbb{1}\{d = \arg\max_a \zeta_{s-1}^\alpha(d)\}) \le (1-\bar{\Xi}_t)\mathbb{1}\{\max_{a\ne M}\{\zeta_{s-1}^\alpha(a) - \zeta_{s-1}^\alpha(M)\} > 0\}||\theta||_1.$$

24

Thus,

$$\sum_{d=0}^{M} \theta(d) t^{-1} \sum_{s=1}^{t} 1\{d = \arg\max_{a} \theta(a)\} - \sum_{d=0}^{M} \theta(d) \iota_t(d) \geq -||\theta||_1 \left(t^{-1} \sum_{s=1}^{t} (1 - \Xi_s) \left(1\{\max_{a \neq M}\{\zeta_{s-1}^{\alpha}(a) - \zeta_{s-1}^{\alpha}(M)\} > 0\} \right) + \frac{\bar{\Xi}_t}{M+1} \right)$$

Hence, by the Cauchy-Swarchz inequality,

$$\mathbf{P}\left(\max_{d} \theta(d) - \sum_{d=0}^{M} \theta(d) \iota_t(d) \leq -||\theta||_1 \left(\sqrt{t^{-1} \sum_{s=1}^{t} (1 - \Xi_s)^2} \sqrt{e^{\gamma} \Lambda_t(\Delta)} + \frac{\bar{\Xi}_t}{M+1} \right) \right)$$

$$\leq \mathbf{P}\left(t^{-1} \sum_{s=1}^{t} 1\{\max_{d \neq M}\{\zeta_{s-1}^{\alpha}(d) - \zeta_{s-1}^{\alpha}(M)\} > 0\} \geq e^{\gamma} \Lambda_t(\Delta) \right),$$

Thus, by the Markov inequality,

$$\mathbf{P}\left(t^{-1} \sum_{s=1}^{t} 1\{\max_{d \neq M}\{\zeta_{s-1}^{\alpha}(d) - \zeta_{s-1}^{\alpha}(M)\} > 0\} \geq e^{\gamma} \Lambda_t(\Delta) \right) \leq e^{-\gamma} (\Lambda_t(\Delta))^{-1} t^{-1} \sum_{s=1}^{t} \mathbf{P}\left(\max_{d \neq M}\{\zeta_{s-1}^{\alpha}(d) - \zeta_{s-1}^{\alpha}(M)\} > 0 \right)$$

and by Proposition G.1 and the fact that $\Omega_0(g, \bar{\zeta}_0(d), \nu_0(d)) \leq \max_{o \in \{0,\dots,L\}} \frac{|\bar{\zeta}_0^o(d)| \nu_0^o(d)/t}{g + \nu_0^o(d)/t}$,

$$\mathbf{P}\left(t^{-1} \sum_{s=1}^{t} 1\{\max_{d \neq M}\{\zeta_{s-1}^{\alpha}(d) - \zeta_{s-1}^{\alpha}(M)\} > 0\} \geq e^{\gamma} \Lambda_t(\Delta) \right) \leq e^{-\gamma} (\Lambda_t(\Delta))^{-1} 4 \sum_{d=0}^{M} t^{-1} \sum_{s=1}^{t} (e^{-0.5(s-1)\frac{H^{-1}(0.5\Delta, \bar{\zeta}_0(d), \nu_0(d))}{\max\{1/8, \nu\sigma(d)^2\}}}$$

$$+ \omega_{s-1}(d)).$$

By our choice of $\Lambda_t(\Delta)$, the RHS is less than $e^{-\gamma}$ and the desired result follows. $\qquad \square$

Online Supplemental Material

I   General Learning Model

Next we present a learning model for the joint distribution of potential outcomes, and we also show that the learning model presented in the text is a particular case of this more general learning model.

Formally, for each $x \in \mathbb{X}$, the PM has a family of PDFs indexed by a finite dimensional parameter $\theta \in \Theta$, $\mathcal{P}_x := \{p_\theta \colon \theta \in \Theta\} \subseteq \Delta(\mathbb{R}^{M+1})$, that models what she believes are plausible descriptions of the true joint probability of the potential outcome $(Y(d,x))_{d \in \mathbb{D}}$. For each $p_\theta \in \mathcal{P}_x$, we use $p_{\theta,d}$ to denote the marginal PDF of p_θ for $Y(d,x)$. Observe that each $p_\theta \in \mathcal{P}_x$ induces a conditional PDF over the realized outcome $Y_t(x) = Y_t(D_t(x),x)$ given the treatment assignment $D_t(x)$:

$$p_\theta(Y_t(x) \mid D_t(x)) = p_{\theta,D_t(x)}(Y_t(x)).$$

Suppose the PM has $L+1$ prior beliefs regarding which elements of $\mathcal{P}_x$ are more likely; each of these prior beliefs summarize the prior knowledge obtained from the $L+1$ different sources; we use $(\mu_0^o(x))_{o=0}^L$ to denote such prior beliefs.

For each $x \in \mathbb{X}$, the family $\mathcal{P}_x$ and the collection of prior beliefs gives rise to $L+1$ subjective Bayesian models for $P(.|x)$. Given the realized outcome $Y_t(x) = Y_t(D_t(x),x)$ and the treatment assignment $D_t(x) = d$, each of these models will produce, with Bayesian updating, a posterior belief given by

$$\mu_t^o(x)(A) = \frac{\int_A p_{\theta,d}(Y_t(x))\mu_{t-1}^o(x)(d\theta)}{\int_\Theta p_{\theta,d}(Y_t(x))\mu_{t-1}^o(x)(d\theta)}$$

for any Borel set $A \subseteq \Theta$. Observe that it is possible that the policymaker's subjective model imposes "cross outcomes restrictions", meaning that the distribution of the different potential outcomes may have common components. Hence, in principle, the policymaker uses observations of $Y(d,x)$ to learn something about the distribution of $Y(d',x)$ with $d' \neq d$; we discuss this feature (or rather the lack of it) in the sub-section below.

Faced with $L+1$ distinct subjective Bayesian models, $\{\langle \mathcal{P}_x, \mu_0^o(x)\rangle\}_{o=0}^L$, our PM has to somehow aggregate this information. There are many ways of doing this; we choose a particular one whereby, at each instance t, the PM averages the posterior beliefs of each model using as weights the posterior probability that model o best fits the observed data within the class of models being considered, i.e.,

$$\bar{\mu}_t(x)(A) := \sum_{o=0}^L \alpha_t^o(x)\mu_t^o(x)(A)$$

for any Borel set $A \subseteq \Theta$, where

$$\alpha_t^o(x) := \frac{\int \prod_{s=1}^t p_{\boldsymbol{\theta}, D_s(x)}(Y_s(x)) \mu_0^o(x)(d\boldsymbol{\theta})}{\sum_{o=0}^L \int \prod_{s=1}^t p_{\boldsymbol{\theta}, D_s(x)}(Y_s(x)) \mu_0^o(x)(d\boldsymbol{\theta})}.$$

I.1 A special Case: The model in the text

One example of $\mathcal{P}_x$ that is of particular interest is one where $\boldsymbol{\Theta} = \prod_{d \in \mathbb{D}} \Theta$ and, for each $d \in \mathbb{D}$, $p_{\boldsymbol{\theta}, d} = p_{\theta_d, d}$ (i.e., it only depends on the d-th coordinate of $\boldsymbol{\theta}$; henceforth, we omit "d" from the θ_d); and also, for each $o \in \{0, ..., L\}$, $\mu_0^o(x) = \prod_{d \in \mathbb{D}} \mu_0^o(d, x)$. That is, each potential outcome has its own parameter and thus learning of each takes place individually and independently. Thus, there is no extrapolation, in the sense that having observed $Y_t(d, x)$ does not affect the beliefs about $Y_t(d', x)$ for any $d' \neq d$. To see this, the posterior for model o at instance $t = 1$ is given by

$$\int f(\boldsymbol{\theta}) \mu_1^o(x)(d\boldsymbol{\theta}) = \int f(\theta_0, ..., \theta_M) \frac{p_{\theta, d}(Y_1(x)) \mu_0^o(d, x)(d\theta) \prod_{d' \neq d} \mu_0^o(d', x)(d\theta)}{\int_\Theta p_{\theta, d}(Y_1(x)) \mu_0^o(d, x)(d\theta)}$$

for any $f : \Theta \to \mathbb{R}$. Now suppose we are interested in the posterior for $d' \neq d$; to do this we set $f(\boldsymbol{\theta}) = 1\{\theta_{d'} \in A\}$ for any $A \subseteq \Theta$ Borel. It is easy to see that

$$\mu_1^o(d', x)(A) = \mu_0^o(d', x)(A),$$

so the posterior is not updated. On the other hand, the posterior for θ_d is given by

$$\mu_1^o(d, x)(A) = \int_A \frac{p_{\theta, d}(Y_1(x)) \mu_0^o(d, x)(d\theta)}{\int_\Theta p_{\theta, d}(Y_1(x)) \mu_0^o(d, x)(d\theta)}.$$

That is, the posterior is only updated if $D_t(x) = d$, which is analogous to the missing data problem featured in experiments under the frequentist approach. Moreover, the above expressions imply that $\mu_1^o(x) = \prod_{d \in \mathbb{D}} \mu_1^o(d, x)$.

A more succinct notation that captures these nuances is given by

$$\mu_1^o(d, x)(A) = \int_A \frac{p_{\theta, D_1(x)}(Y_1(x))^{1\{D_1(x)=d\}} \mu_0^o(d, x)(d\theta)}{\int_\Theta p_{\theta, D_1(x)}(Y_1(x))^{1\{D_1(x)=d\}} \mu_0^o(d, x)(d\theta)}$$

for any $d \in \mathbb{D}$ and any $A \subseteq \Theta$ Borel. Applying this recursively, it follows that

$$\mu_t^o(d, x)(A) = \int_A \frac{p_{\theta, D_t(x)}(Y_t(x))^{1\{D_t(x)=d\}} \mu_{t-1}^o(d, x)(d\theta)}{\int_\Theta p_{\theta, D_t(x)}(Y_t(x))^{1\{D_t(x)=d\}} \mu_{t-1}^o(d, x)(d\theta)}$$

2

for any $t \geq 1$.

Setting $\mathcal{P}_{d,x} = \{p_{\theta,d} : \theta \in \Theta\}$ — and changing the notation from $p_{\theta,d}$ to p_θ — it is easy to see that the previous recursion describes the Bayesian updated presented in the paper.

J   Proof of Supplemental Lemmas for Appendix B

Proof of Lemma B.1 . Henceforth, we omit the notation "x" from the quantities as there is no risk of confusion. Let $\mathcal{F}^s$ denote the σ-algebra generated by $(D_1,...,D_s,Y_1,...,Y_{s-1})$. We now show that for each $\theta \in \Theta$, $(g(Y_s,D_s,\theta))_s$ is a MDS with respect to aforementioned σ-algebra. To do this, note that

$$E[g(Y_s,D_s,\theta) \mid \mathcal{F}^s] = 1\{D_s = d\}E[\ell(Y_s,\theta) - E[\ell(Y_s,\theta) \mid D_s] \mid \mathcal{F}^s].$$

Observe that $E[\ell(Y_s,\theta) \mid \mathcal{F}^s] = E[\ell(Y_s,\theta) \mid D_s]$ because Y_s is independent of the whole past once we condition on D_s. Since $E[[\ell(Y_s,\theta) \mid D_s] \mid \mathcal{F}^s] = E[\ell(Y_s,\theta) \mid D_s]$, it follows that $E[g(Y_s,D_s,\theta) \mid \mathcal{F}^s] = 0$.

Since Θ is assumed to be compact, for any $\delta > 0$, there exists a L_δ such that $\Theta \subseteq \cup_{l=1}^{L} B(\theta_l,\delta)$ where $B(\theta,\delta)$ is a δ-radius ball with center θ. Indeed, let L_δ be the smallest number of balls of radius δ needed to cover the set, and as $\Theta \subseteq \mathbb{R}^{|\Theta|}$, it follows that $L_\delta \leq C\delta^{-|\Theta|}$ where C is an universal constant. Thus, for any t and any $\delta > 0$,

$$\sup_{\theta \in \Theta} |t^{-1} \sum_{s=1}^{t} g(Y_s,D_s,\theta)| \leq \max_{l \in \{1,...,L_\delta\}} |t^{-1} \sum_{s=1}^{t} g(Y_s,D_s,\theta_l)|$$

$$+ \max_{l \in \{1,...,L_\delta\}} \sup_{\theta \in B(\theta_l,\delta)} |t^{-1} \sum_{s=1}^{t} \{g(Y_s,D_s,\theta) - g(Y_s,D_s,\theta_l)\}|.$$

By the triangle inequality and simple algebra, for any t, any $\delta > 0$ and any $l \in \{1,...,L_\delta\}$,

$$\sqrt{E[\sup_{\theta \in B(\theta_l,\delta)} |t^{-1} \sum_{s=1}^{t} \{g(Y_s,D_s,\theta) - g(Y_s,D_s,\theta_l)\}|^2]} \leq t^{-1} \sum_{s=1}^{t} \sqrt{E[\sup_{\theta \in B(\theta_l,\delta)} |g(Y_s,D_s,\theta) - g(Y_s,D_s,\theta_l)|^2]}$$

$$\leq 2\sqrt{E[\sup_{\theta \in B(\theta_l,\delta)} |\ell(Y(d,x),\theta) - \ell(Y(d,x),\theta_l))|^2]}$$

where the last line follows from IID-ness of $Y(d,x)$. By Assumption, the RHS is less than $\varphi(\delta)$.

By the Martingale difference property, for any t and any $\delta > 0$,

$$E[(\max_{l \in \{1,...,L_\delta\}} |t^{-1} \sum_{s=1}^{t} g(Y_s,D_s,\theta_l)|)^2] \leq L_\delta t^{-1} \max_{1 \leq s \leq t} E[(g(Y_s,D_s,\theta_l))^2] \leq C L_\delta t^{-1} \max_{1 \leq s \leq t} E[(\ell(Y(d,x),\theta_l))^2]$$

3

for some universal constant C. Compactness of Θ and the continuity assumption imply that $\max_l E[(\ell(Y(d,x),\theta_l))^2] \leq C$ for some finite constant.

Then, by the Markov inequality, for any $\varepsilon > 0$, any t and any $\delta > 0$,

$$\mathbf{P}\left(\sup_{\theta \in \Theta} |t^{-1}\sum_{s=1}^{t} g(Y_s, D_s, \theta)| \geq \varepsilon\right) \leq C(\varepsilon^{-2}t^{-1}L_\delta + \varepsilon^{-1}\varphi(\delta)) \leq C\varepsilon^{-2}(t^{-1}\delta^{-|\Theta|} + \varphi(\delta))$$

For any t, choose δ as the argmin of $\Lambda(t,|\Theta|) := \min_{\delta \geq 0}(t^{-1}\delta^{-|\Theta|} + \varphi(\delta))$. Observe that $\Lambda(.,.)$ is decreasing in t and increasing in $|\Theta|$ and $\lim_{t\to\infty}\Lambda(t,|\Theta|) = 0$; the first property is straightforward and the second one follows because $t^{-1}\delta^{-|\Theta|} + \varphi(\delta)$ converges to $\varphi(\delta)$ (pointwise) and $\varphi(0) = 0$.

Hence, by choosing $\varepsilon = \sqrt{(0.5/C)M}\sqrt{\Lambda(t,|\Theta|)}$, the previous display implies that for any $M > 0$

$$\mathbf{P}\left(\sup_{\theta \in \Theta} |t^{-1}\sum_{s=1}^{t} g(Y_s, D_s, \theta)| \geq \sqrt{(0.5/C)M}\sqrt{\Lambda(t,|\Theta|)}\right) \leq M^{-1}$$

for any t. Thus, by setting $r_t = \sqrt{(0.5/C)M}\sqrt{\Lambda(t,|\Theta|)}$ and re-defining M as $1/\gamma$ the desired result follows. $\square$

Proof of Lemma B.2. For any $(d,x) \in \mathbb{D} \times \mathbb{X}$ and any $o,o' \in \{0,...,L\}$ observe that

$$\frac{\alpha_t^o(d,x)}{\alpha_t^{o'}(d,x)} = \frac{\int \exp\left\{\sum_{s=1}^{t} 1\{D_s(x) = d\}\ell(Y_s,\theta)\right\} \mu_0^o(d,x)(d\theta)}{\int \exp\left\{\sum_{s=1}^{t} 1\{D_s(x) = d\}\ell(Y_s,\theta)\right\} \mu_0^{o'}(d,x)(d\theta)}$$

where $\ell(.,\theta) := \log p_\theta(.)/p(. \mid d,x)$ and $p(\cdot|d,x)$ is the true PDF of $Y(d,x)$. Moreover,

$$\int \exp\left\{\sum_{s=1}^{t} 1\{D_s(x) = d\}\ell(Y_s,\theta)\right\} \mu_0^o(d,x)(d\theta) = \int_1^\infty \mu_0^o(d,x)\left(\exp\left\{\sum_{s=1}^{t} 1\{D_s(x) = d\}\ell(Y_s,\theta)\right\} \geq u\right) du$$

$$= N_t(d,x) \int_0^\infty \mu_0^o(d,x)\left(-\frac{N_t(d,x)}{\sum_{s=1}^{t} 1\{D_s(x) = d\}\ell(Y_s,\theta)} \leq v\right) e^{-N_t(d,x)v} dv$$

where the second equality is obtained by a change of variables $v = -N_t^{-1}(d,x)\log u$.

In addition, note that

$$N_t^{-1}(d,x)\sum_{s=1}^{t} 1\{D_s(x) = d\}\ell(Y_s,\theta) = \frac{t}{N_t(d,x)}t^{-1}\sum_{s=1}^{t} g(Y_s, D_s, \theta) + N_t^{-1}(d,x)\sum_{s=1}^{t} 1\{D_s(x) = d\}E[\ell(Y_s,\theta) \mid D_s]$$

$$= \frac{t}{N_t(d,x)}t^{-1}\sum_{s=1}^{t} g(Y_s, D_s, \theta) + E[\ell(Y(d,x),\theta)]$$

where $g(Y_s, D_s, \theta) := (1\{D_s(x) = d\}(\ell(Y_s,\theta) - E[\ell(Y_s,\theta) \mid D_s(x)])$ and the last line follows because $1\{D_s(x) = $

4

$d\}E[\ell(Y_s,\theta) \mid D_s(x)] = 1\{D_s(x) = d\}E[\ell(Y_s,\theta) \mid D_s(x) = d] = 1\{D_s(x) = d\}E[\ell(Y(d,x),\theta)]$ as $Y(d,x)$ is IID and in particular independent of $D_s(x)$.

By Lemma B.1, for any $\gamma > 0$,

$$\mathbf{P}\left(\sup_{\theta \in \Theta}\left|t^{-1}\sum_{s=1}^{t} g(Y_s, D_s, \theta)\right| \geq r_t\right) \leq \gamma,$$

where $r_t = \sqrt{1/(2C\gamma)}\sqrt{\Lambda(t, |\Theta|)}$ (C is an universal constant defined inside the proof of the lemma).

Henceforth, fix γ. The previous result implies that, with probability greater than $1 - \gamma$,

$$\int \exp\left\{\sum_{s=1}^{t} 1\{D_s(x) = d\}\ell(Y_s,\theta)\right\}\mu_0^o(d,x)(d\theta) \leq N_t(d,x)\int_0^\infty \mu_0^o(d,x)\left(E[-\ell(Y(d,x),\theta)] - \frac{tr_t}{N_t(d,x)} \leq v\right)e^{-N_t(d,x)v}\,dv$$

and

$$\int \exp\left\{\sum_{s=1}^{t} 1\{D_s(x) = d\}\ell(Y_s,\theta)\right\}\mu_0^o(d,x)(d\theta) \geq N_t(d,x)\int_0^\infty \mu_0^o(d,x)\left(E[-\ell(Y(d,x),\theta)] + \frac{tr_t}{N_t(d,x)} \leq v\right)e^{-N_t(d,x)v}\,dv.$$

Hence, since $KL_{d,x}(\theta) := -E[\ell(Y(d,x),\theta)]$ it follows that, with probability greater or equal to $1 - \gamma$,

$$\frac{\alpha_t^o(d,x)}{\alpha_t^{o'}(d,x)} \leq \frac{N_t(d,x)\int_0^\infty \mu_0^o(d,x)\left(KL_{d,x}(\theta) - \frac{tr_t}{N_t(d,x)} \leq v\right)e^{-N_t(d,x)v}\,dv}{N_t(d,x)\int_0^\infty \mu_0^{o'}(d,x)\left(KL_{d,x}(\theta) + \frac{tr_t}{N_t(d,x)} \leq v\right)e^{-N_t(d,x)v}\,dv} = \frac{\int_0^\infty \mu_0^o(d,x)\left(KL_{d,x}(\theta) \leq v\right)e^{-N_t(d,x)v + tr_t}\,dv}{\int_0^\infty \mu_0^{o'}(d,x)\left(KL_{d,x}(\theta) \leq v\right)e^{-N_t(d,x)v - tr_t}\,dv}.$$

□

K Proof of Proposition G.1

Proof of Proposition G.1. By analogous calculations to those in the proof of Proposition 3.2, for any $\gamma \geq 0$ and $\eta \in [0, h_t(d))$,

$$\mathbf{P}\left(\exists d \neq M: \zeta_t^\alpha(d) - \zeta_t^\alpha(M) > \varepsilon\right) \leq 1\{\cup_d \mathcal{U}_d(\gamma, h_t(d) - \eta, \Delta + \varepsilon)\} + \mathbf{P}\left(\cup_d \mathcal{J}_t(\gamma, d)^C\right) + \mathbf{P}\left(\cup_d \mathcal{E}_t(\eta, d)^C\right) + \omega_t(d)$$

where the sets $\mathcal{U}_d$, $\mathcal{J}_t$ and $\mathcal{E}_t$ are defined as in the proof of proposition 3.2.

By Lemma E.1,

$$\mathbf{P}\left(\cup_d \mathcal{J}_t(\gamma, d)^C\right) + \mathbf{P}\left(\cup_d \mathcal{E}_t(\eta, d)^C\right) \leq 2e^{-0.5t\frac{\gamma^2}{v\sigma(d)^2}} + 2e^{-4t\eta^2}.$$

5

By remark D.1 and the definition of Ω_0, it follows that $\Gamma_0(\gamma, h_t(d) - \gamma, (-1)^{1\{d=M\}}\bar{\zeta}_0(d), v_0(d)) \leq \max_{o \in \{0,\ldots,L\}} \Omega_0(h_t(d) - \gamma, (-1)^{1\{d=M\}}\bar{\zeta}_0^o(d), v_0^o(d))$. Thus,

$$1\{\cup_d \mathcal{U}_d(\gamma, h_t(d) - \eta, \Delta + \varepsilon)\} \leq 1\{\cup_d\{\max_{o \in \{0,\ldots,L\}} \Omega_0(h_t(d) - \gamma, (-1)^{1\{d=M\}}\bar{\zeta}_0^o(d), v_0^o(d)) > 0.5(\Delta + \varepsilon)\}\}.$$

Since $g \mapsto \max_{o \in \{0,\ldots,L\}} \Omega_0(g, \zeta_0^o(d), v_0^o(d))$ is decreasing, then the generalized inverse exists, which we denote by H^{-1}. Let $\gamma_d := H^{-1}(0.5(\Delta + \varepsilon), h_t(d), (-1)^{1\{d=M\}}\bar{\zeta}_0(d), v_0(d))$, which is less than $h_t(d)$. $\qquad\square$

L Examples of policy rules and their corresponding exploration structure

In this appendix we further discuss examples of policy rules and their corresponding exploration structure.

L.1 Thompson Sampling and its refinements

Sampling schemes like Thompson's (Thompson (1933)) and others imply $\delta(\zeta_t, v_t, \alpha_t)(d|x) = \pi_t(d|x)$ where $\pi_t(d|x)$ is the subjective probability that treatment d yields the highest expected outcome and it is associated with the beliefs of the PM at time t, (ζ_t, v_t, α_t). For instance, in Thompson sampling $\pi_t(d|x)$ is given by $\mu_t^\alpha(d,x)$. The next proposition present a structure of exploration for this policy rule.[24]

Proposition L.1. *For any $t \in \{1,\ldots,T\}$, and any $(a_s', b_s)_{s=1}^t$ such that $a_s' \geq \max_o |\zeta_0^o(d,x)|$, and any $b_s \geq a_s'$ for all $s \leq t$, it follows that*

$$h_t(d|x) = t^{-1} \sum_{s=1}^t (1 - \max_{o \in \{0,\ldots,L\}} \Phi(a_s' + b_s; 0, 1/(s + v_0^o(d,x)))) \prod_{l \neq d} \prod_{o=0}^L \int_{-b_s+a_s'}^{b_s-a_s'} \phi(y; 0, 1/v_0^o(l,x)) dy,$$

$$\omega_t(d,x) = 1 - e^{\sum_{s=1}^{t-1} \log(\min_{d,x} P(-a_s' \leq Y(d,x) \leq a_s' | d,x))},$$

is an exploration structure for the Thompson sampling policy rule.

Proof. It suffices to show that, for any $t \in \{1,\ldots,T\}$,

$$\mathbf{P}(\forall s \leq t: \pi_s(d|x) \geq e_s(d|x)) \geq 1 - \omega_t(d,x)$$

with $e_s(d|x) := (1 - \max_{o \in \{0,\ldots,L\}} \Phi(a_s' + b_s; 0, 1/(s + v_0^o(d,x)))) \prod_{l \neq d} \prod_{o=0}^L \int_{-b_s+a_s'}^{b_s-a_s'} \phi(y; 0, 1/v_0^o(l,x)) dy$ as this implies that $h_t(d|x) = t^{-1} \sum_{s=1}^t e_s(d|x)$ is an exploration index.

---

[24]In the proof we use that the probability generating π_t has full support; so the proof can be extended to similar sampling schemes that satisfy this condition.

To do this, let, for each $t \in \{1,...,T\}$ and $\mathbf{a}' := (a'_s)_{s \leq t} > 0$, $S_t(\mathbf{a}') := \{(Y_s(.,.))_{s=1}^{t-1} : \forall s \leq t-1, |Y_s(.,.)| \leq a'_s\}$. Observe that under this set $\max_{d,x,o} |\zeta^o_s(d,x)| \leq \max\{a'_s, |\zeta^o_0(d,x)|\} =: a_s$. Thus, it suffices to show that

$$\mathbf{P}(\forall s \leq t : \pi_s(d|x) \geq e_s(d|x) \mid S_t(\mathbf{a}'))\mathbf{P}(S_t(\mathbf{a}')) \geq 1 - \omega_t(d,x).$$

We first show that $\mathbf{P}(\forall s \leq t : \pi_s(d|x) \geq e_s(d|x) \mid S_t(\mathbf{a}')) = 1$. To do this, fix a history of potential outcomes in $S_t(\mathbf{a}')$ and note that

$$\pi_s(d|x) = \Pr(\hat{\zeta}^\alpha_s(d,x) \geq \max_{a \neq d} \hat{\zeta}^\alpha_s(a,x)) = \int \Pr(\hat{\zeta}^\alpha_s(d,x) \geq z \mid z = \max_{u \neq d} \hat{\zeta}^\alpha_s(u,x)) \Pr(dz)$$

where Pr is the product measure induced by the posterior for each arm, which is a mixture of Gaussians with weights $\alpha^o_{s-1}(d,x)$ and each Gaussian PDF has mean $\zeta^o_{s-1}(d,x)$ and variance $1/v^o_{s-1}(d,x)$. Observe these quantities as non-random as we are fixing a history.

Thus,

$$\pi_s(d|x) = \Pr(\hat{\zeta}^\alpha_s(d,x) \geq \max_{u \neq d} \hat{\zeta}^\alpha_t(u,x)) = \int (1 - \sum_{o=0}^{L} \alpha^o_{t-1}(d,x)\Phi(z; \zeta^o_{s-1}(d,x), 1/v^o_{s-1}(d,x))) \Pr(dz)$$

$$\geq \int_K (1 - \sum_{o=0}^{L} \alpha^o_{s-1}(d,x)\Phi(z; \zeta^o_{s-1}(d,x), 1/v^o_{s-1}(d,x))) \Pr(dz)$$

$$\geq (1 - \max_o \max_{z \in K} \Phi(z; \zeta^o_{s-1}(d,x), 1/v^o_{s-1}(d,x))) \Pr(K)$$

for any $K \subseteq \mathbb{R}$ of the form $K = [-b,b]$.

As we are fixing a history of potential outcomes in $S_t(a')$, the previous display implies that

$$\pi_s(d|x) \geq (1 - \max_o \Phi(b; \zeta^o_{s-1}(d,x), 1/v^o_{s-1}(d,x))) \Pr(K)$$

$$\geq (1 - \max_o \Phi(b; -a, 1/v^o_{s-1}(d,x))) \Pr(K)$$

$$= (1 - \max_o \Phi(a+b; 0, 1/v^o_{s-1}(d,x))) \Pr(K)$$

$$\geq (1 - \max_o \Phi(a+b; 0, 1/(s+v^o_0(d,x)))) \Pr(K)$$

where the last line follows because $x \mapsto \Phi(c; 0, x)$ is decreasing for $c > 0$.

We now bound $\Pr(K)$. To do this, note that given past data, the $\hat{\zeta}^o_t(.,x)$ are independent and thus

$$\Pr(|\max_{l \neq d} \hat{\zeta}^\alpha_s(l,x)| \leq b) \geq \Pr(\max_{l \neq d} \max_o |\hat{\zeta}^o_s(l,x)| \leq b) = \prod_{l \neq d} \prod_{o=0}^{L} \Pr(|\hat{\zeta}^o_s(l,x)| \leq b).$$

7

Moreover, since $|\zeta_s^o(d,x)| \le a$,

$$
\begin{aligned}
\Pr(|\hat{\zeta}_s^o(l,x)| \le b) &= (\Phi(b; \zeta_{s-1}^o(l,x), 1/v_{s-1}^o(l,x)) - \Phi(-b; \zeta_{s-1}^o(l,x), 1/v_{s-1}^o(l,x))) \\
&= (\Phi(b - \zeta_{s-1}^o(l,x); 0, 1/v_{s-1}^o(l,x)) - \Phi(-b - \zeta_{s-1}^o(l,x); 0, 1/v_{s-1}^o(l,x))) \\
&\ge (\Phi(b - a; 0, 1/v_{s-1}^o(l,x)) - \Phi(-b + a; 0, 1/v_{s-1}^o(l,x))) \\
&\ge (\Phi(b - a; 0, 1/v_0^o(l,x)) - \Phi(-b + a; 0, 1/v_0^o(l,x)))
\end{aligned}
$$

where the third line follows because Φ is increasing in its first argument and the fourth line follows because $v_{s-1}^o \ge v_0^o$ and $b - a > 0$.

Therefore, we showed that $\mathbf{P}(\forall s \le t : \pi_t(d|x) \ge e_t(d|x) \mid S_t(\mathbf{a}')) = 1$.

We now show that $\mathbf{P}(S_t(a')) \ge 1 - \omega_t(d,x)$. To do this, observe that the potential outcomes are assumed to be IID, so $\mathbf{P}(S(\mathbf{a}')) \ge \prod_{s=1}^{t-1}(\min_{d,x}(P(-a_s' \le Y(d,x) \le a_s \mid d,x))$. $\qquad\square$

L.2 Markov Policy Rules

For any past history (y^{t-1}, d^{t-1}), let

$$
\delta_t(y^{t-1}, d^{t-1})(\cdot|x) = \delta(\zeta_{t-1}, v_{t-1}, \alpha_{t-1})(\cdot|x), \ \forall x \in \mathbb{X},
$$

where $\zeta_t := (\zeta_t^o)_{o=0}^L$ (the other variables are similarly defined), and

Assumption 4. *There exists an $\epsilon \in (0, 1/(M+1))$ such that $\delta(\cdot)(\cdot|x) \ge \epsilon$ for all $x \in \mathbb{X}$.*

Under this assumption, each treatment arm is chosen with positive probability, thus ensuring some experimentation.

It is straightforward to show that a structure of exploration for this class of policy rules is given by $e_t(d|x) = \epsilon$ and $\omega_t(d,x) = 0$. However, exploiting the Markov assumption, we can obtain a sharper exploration index. This result relies on the fact that $(D_t)_t$ is β-mixing, formally:

Lemma L.1. $(Z_t, D_t, Y_t)_t$ *is a Markov Chain and is β-mixing with mixing coefficients, $(\beta_k)_k$ such $\beta_k \le Ce^{-0.5k\varrho(\epsilon)}$ for all $k \ge 1$.*

where $\varrho(.)$ is a non-decreasing function defined in the proof that is relegated to the end of the section.

Given this result, we employ coupling results for β-mixing process to derive exponential inequalities for $f_t(d|x) - t^{-1}\sum_{s=1}^t E[1\{D_s(x) = d\}]$.

8

Proposition L.2. *Suppose the true PDF of $Y(d,x)$ has full support. Then, for any $a > 0$, $(h_t,\omega_t)_t$ given by*

$$h_t(d|x) = t^{-1} \sum_{s=1}^{t} E[\delta(\zeta_{s-1}, \nu_{s-1}, \alpha_{s-1})(\cdot|x)] - a \geq \epsilon - a, \ \forall t \geq 1$$

$$\text{and } \omega_t(d,x) = 4 \min_{q \in \{1,\ldots,t\}} (2e^{-\frac{0.5ta^2}{4q}} + Ctq^{-1}e^{-0.5q\varrho(\epsilon)}), \ \forall t \geq 1$$

is an exploration structure.

Proof of Proposition L.2. Fix an $(d,x) \in \mathbb{D} \times \mathbb{X}$. It is enough to show that for any positive sequence $(a_t)_t$ and any t,

$$\mathbf{P}\left(f_t(d|x) - t^{-1} \sum_{s=1}^{t} E[1\{D_s(x) = d\}] \geq a_t\right) \leq 2e^{-\frac{0.5t(a_t)^2}{4q}} + Cq^{-1}e^{-0.5q\varrho(\epsilon)}t,$$

for any $q \in \{1,\ldots,t\}$. Henceforth, we will omit x from the notation.

By Lemma L.1, $(D_t)_t$ is β-mixing with mixing coefficients, $(\beta_k)_k$ such $\beta_k \leq Ce^{-0.5k\varrho(\epsilon)}$ for all $k \geq 1$.

We use well-known coupling results for β-mixing sequences (see Yu (1994)). For any $q \in \mathbb{N}$, let $(\tilde{D}_t)_{t=-\infty}^{\infty}$ be such that: (a) for any $i \in \mathbb{N}_0$, $\bar{U}_i \equiv (\tilde{D}_{iq+1},\ldots,\tilde{D}_{iq+q})$ has the same distribution as $U_i \equiv (D_{iq+1},\ldots,D_{iq+q})$; (b) the sequence $(\bar{U}_i)_{i \ even}$ is i.i.d. and so is $(\bar{U}_i)_{i \ odd}$; (c) for any $i \in \mathbb{N}_0$, $\mathbb{P}(\bar{U}_i \neq U_i) \leq \beta_q$, where $\mathbb{P}$ is the product measure of the random processes $(D_t)_{t=-\infty}^{\infty}$ and $(\tilde{D}_t)_{t=-\infty}^{\infty}$.

Consider $t = qk$ for some positive integers q and k; if this does not hold, t is simply replaced with $qk(t)$, where $k(t)$ is the largest integer such that $t \geq qk$. Letting $M_j \equiv \{jq+1,\ldots,(j+1)q\}$, it follows that $\{1,\ldots,t\} = \cup_{j=0}^{k-1} M_j$ and thus,

$$f_t(d|x) - t^{-1} \sum_{s=1}^{t} E[1\{D_s(x) = d\}] = t^{-1}\left(\sum_{j \in E_k} \sum_{i \in M_j} K(D_i) + \sum_{j \in O_k} \sum_{i \in M_j} K(D_i)\right)$$

where $E_k := \{0 \leq j \leq k-1 : j \text{ is even}\}$ and $O_k := \{0 \leq j \leq k-1 : j \text{ is odd}\}$ and $K(D_i) := 1\{D_i = d\} - E[1\{D_i = d\}]$. Hence, it suffices to bound

$$\mathbf{P}\left(\sum_{j \in E_k} \sum_{i \in M_j} K(D_i) \geq 0.5ta_t\right) \text{ and } \mathbf{P}\left(\sum_{j \in O_k} \sum_{i \in M_j} K(D_i) \geq 0.5ta_t\right).$$

We bound the former as the latter is completely analogous.

9

To do this, note that by condition (c),

$$
\mathbf{P}\left(\sum_{j\in E_k}\sum_{i\in M_j}K(D_i)\geq 0.5ta_t\right)\leq\mathbb{P}\left(\left|\sum_{j\in E_k}\sum_{i\in M_j}K(\tilde{D}_i)\right|\geq 0.5ta_t\right)
$$

$$
+\mathbb{P}\left(\left|\sum_{j\in E_k}\sum_{i\in M_j}K(D_i)-\sum_{j\in E_k}\sum_{i\in M_j}K(\tilde{D}_i)\right|\geq 0.5ta_t\right)
$$

$$
\leq\mathbb{P}\left(\left|\sum_{j\in E_k}\sum_{i\in M_j}K(\tilde{D}_i)\right|\geq 0.5ta_t\right)+\mathbb{P}(\exists j\in E_k:\bar{U}_j\neq U_j)
$$

$$
\leq\mathbb{P}\left(\left|\sum_{j\in E_k}\sum_{i\in M_j}K(\tilde{D}_i)\right|\geq 0.5ta_t\right)+\beta_q k.
$$

Observe that $(\sum_{i\in M_j}K(\tilde{D}_i))_{j\in E_k}$ are mean 0, independent by condition (b) and bounded by $2q$. Hence, by Hoeffding inequality

$$
\mathbb{P}\left(\left|\sum_{j\in E_k}\sum_{i\in M_j}K(\tilde{D}_i)\right|\geq 0.5ta_t\right)\leq 2e^{-\frac{0.5t^2(a_t)^2}{4q^2k}}=2e^{-\frac{0.5t(a_t)^2}{4q}}.
$$

Therefore,

$$
\mathbf{P}\left(\sum_{j\in E_k}\sum_{i\in M_j}K(D_i)\geq 0.5ta_t\right)\leq 2e^{-\frac{0.5t(a_t)^2}{4q}}+\beta_q t/q\leq 2e^{-\frac{0.5t(a_t)^2}{4q}}+Cq^{-1}e^{-0.5q\varrho(\epsilon)t}
$$

Since the $q\in\{1,...,t\}$ was arbitrary we can optimize the RHS by minimizing over q. $\qquad\square$

L.2.1 Supplementary Material for Markov Decision Rules: Mixing Results

The following lemmas are used in the proof of Lemma L.1. Their proofs are relegate to the end of the section.

Lemma L.2. *Let π be any probability over $\mathbb{Z}$ such that $\int V(z)\pi(dz)<\infty$ (in particular, it could be different from the invariant distribution λ). Then, the process $(Z_t)_{t=1}^{\infty}$ with initial probabilty π and transition Q is β-mixing with mixing coefficients, $(\beta(k))_k$ such $\beta_k\leq C_M e^{-k0.5\varrho(\epsilon)}$ for all $k\geq 1$, where*

$$
C_M:=3\int V(z)\pi(dz)+\int V(z)\lambda(dz)
$$

The next lemma shows that for any t and any $o\in\{0,...,L\}$, ζ_t^o can be written as a function of (ζ_t^0,v_t^0) and their priors, and the same holds for v_t^o. This result implies that it suffices to study the evolution of $(\zeta_t^0,v_t^0)_t$ and not of all the models.

Lemma L.3. *For any $t \geq 0$, any $o \in \{0,...,L\}$ and any $d \in \{0,...,M\}$,*

$$v_{t+1}^o(d) = v_{t+1}^0(d) - v_0^0(d) + v_0^o(d)$$

$$\zeta_{t+1}^o(d) = \frac{\zeta_{t+1}^0(d)(v_{t+1}^0(d))}{v_{t+1}^0(d) + v_0^o(d) - v_0^0(d)} + \frac{v_0^o(d)\zeta_0^o(d) - v_0^0(d)\zeta_0^0(d)}{v_{t+1}^0(d) + v_0^o(d) - v_0^0(d)}$$

The next lemma shows that α_t^o can also be written as a function of ζ_t^0 and v_t^0, and the priors.

Lemma L.4. *For any $t \geq 0$, any $o \in \{0,...,L\}$ and any $d \in \{0,...,M\}$,*

$$\alpha_t^o(d) = \frac{\exp \ell_t^o(d)}{\sum_{o=0}^L \exp \ell_t^o(d)}$$

where

$$\ell_t^o(d)$$
$$= \log \phi \left(\frac{v_t^0(d)}{v_t^0(d) - v_0^0(d)} \zeta_t^0(d) - \frac{v_0^0(d)}{v_t^0(d) - v_0^0(d)} \zeta_0^0(d); \zeta_0^o(d), (v_t^0(d) - v_0^0(d) + v_0^o(d))/((v_t^0(d) - v_0^0(d))v_0^o(d))\right).$$

Proof of Lemma L.1. First, recall that $z = (\zeta^0, v^0)$ and Lemma L.2 establishes that this process is β-mixing. Since, for all $o \in \{0,...,L\}$, (ζ^o, v^o, α^o) can be written as a deterministic function of z (and the priors; see Lemmas L.3 and L.4), conditional on the priors, the process $(\zeta_t, v_t, \alpha_t)_t$ is also β-mixing with the same convergence rate.

By assumption, δ depends on $(\zeta_t^o, v_t^o, \alpha_t^o)_{o=0}^L$ and thus it depends on $z_t = (\zeta_t^0, v_t^0)$ (and the priors, but these are taken to be non-random). Henceforth, and abusing notation, we use $z \mapsto \delta(z)$ to denote this composition of functions.

Note that for any $s \in \{0, 1,...\}$, $1\{D_s = d\} = 1\{\delta(Z_s)(d) \geq U_s\}$ where $U_s \in U(0,1)$. It is easy to show that the "expanded" process $(Z_t, U_t)_t$ is a Markov Chain with transition $Q \times U(0,1)$ and it inherits all the properties of the original one. In particular, by Lemma L.2 , $(Z_t, U_t)_t$ is β-mixing with the same coefficients. Hence, $(Z_t, D_t)_t$ is also a Markov Chain with transition Q_{ext} and β-mixing with the same coefficients

Since $Y_t = Y_t(D_t)$ is a deterministic function of IID random variables, $(Y_t(d))_d$ and D_t. So to show that $(Z_t, D_t, Y_t)_t$ is β-mixing with the same coefficients, it suffices to show that $(Z_t, D_t, Y_t(0),...,Y_t(M))_t$ is β-mixing with the same coefficients. It is easy to show that the "expanded" process $(Z_t, D_t, Y_t(0),...,Y_t(M))_t$ is a Markov Chain with transition $Q_{ext} \times F(0) \times ... \times F(M)$ and it inherits all the properties of the original one. In particular, by Lemma L.2 , $(Z_t, D_t, Y_t(0),...,Y_t(M))_t$ is β-mixing with the same coefficients. Therefore, $(Z_t, D_t, Y_t)_t$ is a Markov Chain and is β-mixing with mixing coefficients, $(\beta(k))_k$ such $\beta_k \leq Ce^{-0.5k\varrho(\epsilon)}$ for all $k \geq 1$. $\qquad \square$

11

The proof of Lemma L.2 follows from Propostion 4 in Liebscher (2005). These result in turn relies on properties of Markov chains, thus we need to fit our problem in this framework. To do this, observe that the PM is given by $(\zeta_{t-1}, v_{t-1}, \alpha_{t-1})$; we want to understand this stochastic process. To do this, Lemmas L.3 and L.4 show that for any t and any $o \in \{0, ..., L\}$, ζ_t^o can be written as a function of (ζ_t^0, v_t^0) and their priors, and the same holds for v_t^o. This result implies that it suffices to study the evolution of $(\zeta_t^0, v_t^0)_t$ and not of all the models. Throughout we omit the super script "0" from (ζ_t^0, v_t^0) and, for each t, let

$$Z_t := (\zeta_t, v_t) \in \mathbb{Z};$$

we now define the state space, $\mathbb{Z}$. To do this, first let, for any $t \in \{0, ..., T\}$,

$$\mathcal{V}_t(v_0) := \{a \in \{v_0, 1 + v_0, ..., t + v_0\}^{M+1} : \sum_{d=0}^{M} a(d) - v_0 < t\}$$

$$\partial \mathcal{V}_t(v_0) := \{a \in \{v_0, 1 + v_0, ..., t + v_0\}^{M+1} : \sum_{d=0}^{M} a(d) - v_0 = t\}$$

$$\bar{\mathcal{V}}_t(v_0) := \mathcal{V}_t(v_0) \cup \partial \mathcal{V}_t(v_0).$$

Note that $v_t \in \bar{\mathcal{V}}_T(v_0)$ and if $v_t \in \partial \mathcal{V}_T(v_0)$, then $\sum_d 1\{D_t = d\} = T$ and thus the experiment stops. Hence, $\mathbb{Z} := \mathbb{R}^{M+1} \times \bar{\mathcal{V}}_T(v_0)$. Henceforth, we will omit v_0 from the set $\mathcal{V}_T$.

Henceforth, we will omit v_0 from the set $\mathcal{V}_T$.

The policy functions $\phi := (\sigma, \delta)$ — that are time homogeneous — and the DGP for Y induce a Markov chain over $(Z_t)_t$ with transition probability function Q. This transition probability function is characterize by the following recursion: For any $d \in \{0, ..., M\}$ and given $z_t = (\zeta_t, v_t)$,

$$v_{t+1}(d) = v_t(d) + 1\{D_{t+1} = d\}1\{v_t \in \mathcal{V}_T(v_0)\}$$
$$\zeta_{t+1}(d) = \frac{1\{D_{t+1} = d\}Y_{t+1}(d)}{v_t(d) + 1\{D_{t+1} = d\}} + \frac{v_t(d)\zeta_t(d)}{v_t(d) + 1\{D_{t+1} = d\}},$$

where $\Pr(D_{t+1} = d \mid z_t) = \delta(z_t)(d)$ and $Y_{t+1}(d) \sim F_d$ where F_d has mean $\theta(d)$, variance $\sigma^2(d)$, has full-support and, by Assumption 3, it is assumed to be sub-gaussian, i.e., $E[\exp \lambda(Y(d) - \theta(d))] \le C \exp \upsilon \sigma^2(d)\lambda^2$ for some constants $C = 1$, $\upsilon > 0$ and any $\lambda > 0$.

Under these assumptions, we show that the Markov Chain has two important properties: Existence of smalls sets and a Lyapounov drift. We now discuss these two concepts and the associated results.

Small sets. In this section we characterize the type of sets that are "small", i.e., a set C such that there exists a $\delta > 0$, a $n \in \mathbb{N}$, and a measure $\psi \in \Delta(\mathbb{Z})$ such that

$$\inf_{z \in C} Q^n(\cdot \mid z) \geq \delta \psi(.).$$

Lemma L.5. *Any set $C = \prod_{m=0}^{M} C(m)$ where $C(m) := \{(\zeta(m), v(m)) : \zeta(m) \in S(m), v(m) = a\}$ with $S(m)$ bounded and with non-empty interior and $a(.) \in \bar{\mathcal{V}}_T(v_0)$ is small, i.e.,*

$$Q^{M+1}(A \mid z) \geq \epsilon^{M+1} \delta_C \psi(A), \forall A \subseteq \mathbb{Z},$$

for any $z \in C$, where ψ is a probability measure such that

$$\psi(A) = \prod_{m=0}^{M} \mathbf{1}_{a(m)+1\{m < m^*\}}(A_v(m)) Leb(A_\zeta(m) \mid C(m)),$$

for any set $A := \prod_{m=0}^{M} A_\zeta(m) \times A_v(m)$, where m^* be the first $m \in \{0, ..., M\}$ such that $a(.) \notin \mathcal{V}_{T-m}$ (if this never happens, we simply set $m^* = T$).

Remark L.1 (A remark about Assumption 2). *By inspection of the proof of Lemma L.5, and the other lemmas used as building blocks, it follows that Assumption 2 could be relaxed to allow for ϵ to depend on the state, i.e., $z \mapsto \epsilon(z)$ provided that $\inf_{z \in C} \epsilon(z) > 0$ for any C compact.* $\triangle$

Drift Condition. We show that Q satisfies a drift condition.

Lemma L.6. *For any $a \geq 0$ and $A \geq 0$, the function $z \mapsto V(z) := 1 + a\|\bar{\zeta}\|_1 + A\|v - \partial\mathcal{V}_T\|_1$ where $d \mapsto \bar{\zeta}(d) := (\zeta(d) - \theta(d))/\sigma(d)$, satisfies*

$$Q[V](z) \leq \gamma(z)V(z) + b,$$

with $\gamma(z) := \epsilon \max_d \frac{v(d)}{v(d)+1} + (1 - \epsilon)$ and $b := 1 - \gamma + a \sum_d \frac{\delta(z)(d)}{1+v_0(d)}$. Observe that $\max_z \gamma(z) \leq \gamma := \epsilon \max_d \frac{v_0(d)+T}{v_0(d)+T+1} + (1 - \epsilon) < 1$.

Remark L.2. *Let*

$$C := \prod_{m=0}^{M} C(m) \tag{L.1}$$

where $C(m) := \{(\zeta, v) : |\bar{\zeta}(m)| \leq (R/a)/(M + 1) \text{ and } v(m) = \mathbf{v}(m)\}$ for some $\mathbf{v} \in \partial\mathcal{V}_T$ and $R > 0$. By Lemma L.5, this set is small.

13

Moreover, $\inf_{z \in C^C} V(z) \geq R + 1$. *Note that since* $\gamma < 1$, *there exists* R *and a*

$$\gamma + b/(R+1) = \gamma + (1 - \gamma + 0.5a)/(R+1) < 1 \iff \frac{0.5a}{1-\gamma} < R.$$

△

Given these results, we are able to show that the Markov Chain is Geometrically Ergodic, formally:

Lemma L.7. *There exists a constant* L *and an invariant distribution of* Q, λ, *such that, for any* $n \geq 1$ *and any* $z_0 \in \mathbb{Z}$,

$$||Q^n(.|z_0) - \lambda|| \leq LV(z_0)e^{-n\varrho(\epsilon)}$$

where $\epsilon \mapsto \varrho(\epsilon)$ *is positive valued,* $\varrho(0) = 0$ *and increasing on* ϵ; *formally defined in the proof.*

These results imply the conditions of Propostion 4 in Liebscher (2005) and thus we can show Lemma L.2.

Proof of Lemma L.2 . See Propostion 4 in Liebscher (2005). □

L.2.2 Proofs for Supplementary Lemmas

Proof of Lemma L.3 . Throughout, we omit the super script "0" from the quantities.

Observe that for any $t \geq 1$,

$$v_t(d) = \sum_{s=1}^{t} 1\{D_s = d\} + v_0(d) = v_t^o(d) + v_0(d) - v_0^o(d).$$

and

$$
\begin{aligned}
\zeta_{t+1}(d) &= \frac{1\{D_{t+1} = d\}Y_{t+1}(d)}{v_t(d) + 1\{D_{t+1} = d\}} + \frac{v_t(d)\zeta_t(d)}{v_t(d) + 1\{D_{t+1} = d\}} \\
&= \frac{1\{D_{t+1} = d\}Y_{t+1}(d)}{v_t(d) + 1\{D_{t+1} = d\}} + \frac{1\{D_t = d\}Y_t(d)}{v_t(d) + 1\{D_{t+1} = d\}} \\
&\quad + \frac{v_{t-1}(d)\zeta_{t-1}(d)}{v_t(d) + 1\{D_{t+1} = d\}}
\end{aligned}
$$

since $v_t(d) = v_{t-1}(d) + 1\{D_t = d\}$, and iterating in this fashion, it follows that

$$\zeta_{t+1}(d) = \sum_{s=1}^{t+1} \frac{1\{D_s = d\}Y_s(d)}{v_t(d) + 1\{D_{t+1} = d\}} + \frac{v_0(d)\zeta_0(d)}{v_{t+1}(d) + 1\{D_{t+1} = d\}}.$$

Hence, $\sum_{s=1}^{t+1} 1\{D_s = d\}Y_s(d) = \zeta_{t+1}(d)(v_t(d) + 1\{D_{t+1} = d\}) - v_0(d)\zeta_0(d)$. Since the same equation holds for

$\zeta_{t+1}^o(d)$, it follows that

$$\zeta_{t+1}(d)(v_t(d) + 1\{D_{t+1} = d\}) - v_0(d)\zeta_0(d) = \zeta_{t+1}^o(d)(v_t^o(d) + 1\{D_{t+1} = d\}) - v_0^o(d)\zeta_0^o(d),$$

which implies

$$
\begin{aligned}
\zeta_{t+1}^o(d) &= \frac{\zeta_{t+1}(d)(v_{t+1}(d))}{v_{t+1}^o(d)} + \frac{v_0^o(d)\zeta_0^o(d) - v_0(d)\zeta_0(d)}{v_{t+1}^o(d)} \\
&= \frac{\zeta_{t+1}(d)(v_{t+1}(d))}{v_{t+1}(d) + v_0^o(d) - v_0(d)} + \frac{v_0^o(d)\zeta_0^o(d) - v_0(d)\zeta_0(d)}{v_{t+1}(d) + v_0^o(d) - v_0(d)}
\end{aligned}
$$

$\square$

Proof of Lemma L.4 . Throughout we omit the super script "0" from the relevant quantities.

It readily follows from the fact that $\zeta_t(d) = \frac{N_t(d)}{v_t(d)} m_t(d) + \frac{v_0(d)}{v_t(d)} \zeta_0(d)$ iff $\frac{v_t(d)}{N_t(d)} \zeta_t(d) - \frac{v_0(d)}{N_t(d)} \zeta_0(d) = m_t(d)$ and $N_t(d) = v_t(d) - v_0(d)$. $\square$

To prove Lemma L.5 we use the following results. The next lemma provides a simple lower bound for Q using the fact that $\delta(.)(.) \geq \epsilon$.

Lemma L.8. *For any set $A := \prod_{m=1}^M A(m)$ and any $z_0 \in \mathbb{Z}$, it follows that*

$$Q^{L+1}(A \mid z_0) \geq \epsilon \int Q^L(A \mid z_1) Q(dz_1 \mid D_0 = d_0, z_0), \ \forall L \geq 0 \text{ and } d_0 \in \{0, ..., M\},$$

and

$$Q(A \mid z_0) \geq \epsilon \sum_{l=0}^M \prod_{d \neq l} \mathbf{1}_{z_0(d)}(A(d)) Q(A(l) \mid D = l, z_0).$$

Proof of Lemma L.8. First we observe that for any $L \in \{0, ..., M\}$, any z_0 and any set A,

$$
\begin{aligned}
Q^{L+1}(A \mid z_0) &= \int Q^L(A \mid z_1) Q(dz_1 \mid z_0) \\
&= \int \sum_{d_0} \delta(z_0)(d_0) Q^L(A \mid z_1) Q(dz_1 \mid D_0 = d_0, z_0) \\
&\geq \epsilon \int Q^L(A \mid z_1) \sum_{d_0} Q(dz_1 \mid D_0 = d_0, z_0),
\end{aligned}
$$

where $Q^0(. \mid z) := \mathbf{1}_z(.)$ and the third line follows because $\delta(.)(.) \geq \epsilon$. Observe further that for any $d \in$

15

$\{0,...,M\}$,

$$Q(A \mid D_0 = d, z_0) = \prod_{d' \neq d} \mathbf{1}_{z_0(d')}(A(d'))Q(A(d) \mid D_0 = d, z_0).$$

Thus, for $L = 0$, this implies that

$$Q(A \mid z_0) \geq \epsilon \sum_{d=0}^{M} \prod_{d' \neq d} \mathbf{1}_{z_0(d')}(A(d'))Q(A(d) \mid D_0 = d, z_0),$$

$\square$

The next lemma show that we can lower bound the transition Q^L for different values of $L \in \{0,...\}$ in terms of a product measure given by the family $(Q_m)_m$.

Lemma L.9. *For any set $A := \prod_{m=1}^{M} A(m)$ such that $A(m) := A_\zeta(m) \times A_\nu(m)$, it follows that*

1. *For any $z_0 \in \mathbb{Z}$ and any $L \in \{1,...,M\}$,[25]*

$$Q^{L+1}(A \mid z_0) \geq \epsilon^L \int Q(A \mid z_1(0), z_2(1), z_3(2), ..., z_L(L-1), z_0(L), ..., z_0(M)) \prod_{t=0}^{L-1} Q_t(dz_{t+1}(t) \mid z_0).$$

2. *For any $z_0 \in \mathbb{Z}$,*

$$Q^{M+1}(A \mid z_0) \geq \epsilon^{M+1} \prod_{m=0}^{M} Q_m(A(m) \mid z_0),$$

Proof of Lemma L.9. (1) By the proof of Lemma L.8, for $L \geq 1$ it follows that

$$Q^{L+1}(A \mid z) \geq \epsilon \int Q^L(A \mid z_1(0), z_0(1), ..., z_0(M))Q(dz_1(0) \mid D_0 = 0, z_0). \tag{L.2}$$

Analogously, for any z_1,

$$Q^L(A \mid z_1) \geq \epsilon \int Q^{L-1}(A \mid z_2)Q(dz_2 \mid D_1 = 1, z_1)$$

$$\geq \epsilon \int Q^{L-1}(A \mid z_1(0), z_2(1), z_1(2), ..., z_1(M))Q(dz_2(1) \mid D_1 = 1, z_1).$$

---

[25] D_t is the random variable corresponding to the treatment assignment in period t; z_t is the state at time t.

Plugging this expression in L.2 with $z_1 = z_1(0), z_0(1), ..., z_0(M)$, it follows that

$$Q^{L+1}(A \mid z_0) \geq \epsilon^2 \int Q^{L-1}(A \mid z_1(0), z_2(1), z_0(2), ..., z_0(M)) Q(dz_2(1) \mid D_1 = 1, z_1(0), z_0(1), ..., z_0(M))$$
$$\times Q(dz_1(0) \mid D_0 = 0, z_0).$$

Since

$$Q(A(d) \mid D_0 = d, z_0) = \mathbf{1}_{\{v_0(d)+1\{v_0(.)\in \mathcal{V}\}\}}(A_v(d)) Q(A_\zeta(d) \mid D_0 = d, z_0(d)), \tag{L.3}$$

it follows that

$$Q^{L+1}(A \mid z_0) \geq \epsilon^2 \int Q^{L-1}(A \mid z_1(0), z_2(1), z_0(2), ..., z_0(M)) \mathbf{1}_{\{v_0(1)+1\{v_1(.)\in \mathcal{V}_T\}\}}(dv_2(1)) Q(d\zeta_2(1) \mid D_1 = 1, z_0(1))$$
$$\times Q(dz_1(0) \mid D_0 = 0, z_0),$$

where $v_1(.) = (v_1(0), v_0(1), ..., v_0(M)) = (v_0(0) + 1\{v_0(.) \in \mathcal{V}_T\}, v_0(1), ..., v_0(M))$ because at time 0, $D_0 = 0$. Hence, $v_1(.) \in \mathcal{V}_T$ iff $v_0 \in \mathcal{V}_{T-1}$. Thus,

$$Q^{L+1}(A \mid z_0)$$
$$\geq \epsilon^2 \int Q^{L-1}(A \mid z_1(0), z_2(1), z_0(2), ..., z_0(M)) \mathbf{1}_{\{v_0(1)+1\{v_0(.)\in \mathcal{V}_{T-1}\}\}}(dv_2(1)) \mathbf{1}_{\{v_0(0)+1\{v_0(.)\in \mathcal{V}_T\}\}}(dv_1(0))$$
$$\times Q(d\zeta_1(0) \mid D_0 = 0, z_0(0)) Q(d\zeta_2(1) \mid D_1 = 1, z_0(1))$$
$$= \int Q^{L-1}(A \mid z_1(0), z_2(1), z_0(2), ..., z_0(M)) Q_1(dz_2(1) \mid z_0) Q_1(dz_1(0) \mid z_0)$$

Iterating in this fashion, it follows that

$$Q^{L+1}(A \mid z_0) \geq \epsilon^L \int Q(A \mid z_1(0), z_2(1), z_3(2), ..., z_L(L-1), z_0(L), ..., z_0(M))$$
$$\times \prod_{t=0}^{L-1} \mathbf{1}_{\{v_0(t)+1\{v_0(.)\in \mathcal{V}_{T-t}\}\}}(dv_{t+1}(t)) \prod_{t=0}^{L-1} Q(d\zeta_{t+1}(t) \mid D_t = t, z_0(t)).$$

(3) In particular, for $L = M$, these results imply that

$$Q^{M+1}(A \mid z_0) \geq \epsilon^M \int Q(A \mid z_1(0), z_2(1), z_3(2), ..., z_L(M-1), z_0(M))$$

$$\times \prod_{t=0}^{M-1} \mathbf{1}_{\{v_0(t)+1\{v_0(.)\in \mathcal{V}_{T-t}\}\}}(dv_{t+1}(t)) \prod_{t=0}^{M-1} Q(d\zeta_{t+1}(t) \mid D_t = t, z_0(t))$$

$$\geq \epsilon^{M+1} \int \prod_{t=0}^{M-1} \mathbf{1}_{z_{t+1}(t)}(A(t)) Q(A(M) \mid D_M = M, z_1(0), z_2(1), z_3(2), ..., z_L(M-1), z_0(M))$$

$$\times \prod_{t=0}^{M-1} \mathbf{1}_{\{v_0(t)+1\{v_0(.)\in \mathcal{V}_{T-t}\}\}}(dv_{t+1}(t)) \prod_{t=0}^{M-1} Q(d\zeta_{t+1}(t) \mid D_t = t, z_0(t))$$

$$\geq \epsilon^{M+1} \int \prod_{t=0}^{M-1} \mathbf{1}_{z_{t+1}(t)}(A(t)) 1_{\{v_0(M)+1\{v_0(.)\in \mathcal{V}_{T-M}\}\}}(A_v(M)) Q(A_\zeta(M) \mid D_M = M, z_0(M))$$

$$\times \prod_{t=0}^{M-1} \mathbf{1}_{\{v_0(t)+1\{v_0(.)\in \mathcal{V}_{T-t}\}\}}(dv_{t+1}(t)) \prod_{t=0}^{M-1} Q(d\zeta_{t+1}(t) \mid D_t = t, z_0(t)).$$

And using our definitions, this implies that

$$Q^{M+1}(A \mid z_0) \geq \epsilon^{M+1} \int \prod_{t=0}^{M} \mathbf{1}_{z(t)}(A(t)) Q(dz(0), ..., dz(M-1), dz(M) \mid z_0)$$

$$= \epsilon^{M+1} \prod_{t=0}^{M} Q_t(A(t) \mid z_0).$$

$\square$

The next lemma provides a lower bound for the probability $Q(\cdot \mid D = d, z_0)$ over $\mathbb{R}$, which in turn helps to construct a lower bound for Q_m

Lemma L.10. *For any $d \in \{0, ..., M\}$ and any $C := \prod_{m=0}^{M} C(m) \subseteq \mathbb{Z}$ where $C(d) := C_\zeta(d) \times C_v(d)$, with $C_\zeta(d)$ bounded with non-empty interior, it follows that for any set $E \subseteq \mathbb{R}$*

$$Q(\zeta_1(d) \in E \mid D = d, z) \geq \delta_C Leb(E \mid C_\zeta(d)), \ \forall z \in C$$

where

$$\delta_C := \inf_{y \in C_\zeta(d)} f_d(y)(1 + v_0(d))$$

Proof of Lemma L.10. Given z and $D = d$, let $y \mapsto e(y, z(d)) := \frac{y}{1+v(d)} + \frac{v(d)}{1+v(d)}\zeta(d)$; also, recall that $y(d) \sim F_d$

18

with PDF f_d. Thus

$$Q(\zeta_1(d) \in E \mid D = d, z) = \int 1\{e(y, z) \in E\} F_d(dy)$$

$$\geq \int 1\{e(y, z) \in E \cap C_\zeta(d)\} F_d(dy)$$

$$\geq \inf_{y \in C_\zeta(d)} f_d(y) \int 1\{e(y, z) \in E \cap C_\zeta(d)\} dy.$$

Observe that $de(y, z) = dy/(1 + v(d))$. Thus, with a change of variables,

$$Q(\zeta_1(d) \in E \mid D = d, z) \geq (1 + v(d)) \inf_{y \in C_\zeta(d)} f_d(y) \int 1\{\zeta'(d) \in E \cap C_\zeta(d)\} d\zeta'(d)$$

$$\geq (1 + v(d)) \inf_{y \in C_\zeta(d)} f_d(y) Leb(E \mid C_\zeta(d))$$

where the last line follows from the fact that $Leb(C_\zeta(d)) \in (0, 1]$. The result thus follows from the fact that $v(d) \geq v_0(d)$. $\qquad\square$

Proof of Lemma L.5. Abusing notation, let $Q(\cdot \mid D = d, z)$ be the probability over $\mathbb{Z}$ given $D = d$ and $Z = z$.

Let m^* be the first $m \in \{0, ..., M\}$ such that $a(.) \notin \mathcal{V}_{T-m}$ (if this never happens, we simply set $m^* = T$).

It is enough to show the results for "squares", $A := \prod_{m=0}^{M} A(m)$ and $A(m) := A_\zeta(m) \times A_v(m)$ with $A_v(m) \subseteq \{1, ..., T\}$ and $A_\zeta(m) \subseteq \mathbb{R}$.

First assume $a(.) \in \mathcal{V}_T$, then by Lemma L.9(2)

$$Q^{M+1}(A \mid z_0) \geq \epsilon^{M+1} \prod_{m=0}^{M} Q_m(A(m) \mid z_0)$$

$$= \epsilon^{M+1} \prod_{m=0}^{m^*-1} Q_m(A(m) \mid z_0) \prod_{m=m^*}^{M} Q_m(A(m) \mid z_0)$$

(if $m^* = 0$, then the first product is taken to be 1 and if $m^* > M$ the second product is taken to be 1).

Recall that for z_0 in C, $v_0(.) = a(.)$. Hence, for each $m \in \{0, ..., m^* - 1\}$,

$$Q_m(A(m) \mid z_0) = \mathbf{1}_{a(m)+1}(A_v(m)) Q(\zeta(m) \in A_\zeta(m) \mid D = m, z_0(m))$$

and for each $m \in \{m^*, ..., M\}$,

$$Q_m(A(m) \mid z_0) = \mathbf{1}_{a(m)}(A_v(m)) Q(\zeta(m) \in A_\zeta(m) \mid D = m, z_0(m)).$$

19

By Lemma L.10, it follows that, for each $m \in \{0, \ldots, m^* - 1\}$,

$$Q(\zeta(m) \in A_\zeta(m) \mid D = m, z_0(m)) \geq \mathbf{1}_{a(m)+1}(A_\nu(m))\delta_{C(m)}Leb(A_\zeta(m) \mid C(m)),$$

and for each $m \in \{m^*, \ldots, M\}$,

$$Q_m(A(m) \mid z_0) \geq \mathbf{1}_{a(m)}(A_\nu(m))\delta_{C(m)}Leb(A_\zeta(m) \mid C(m)).$$

Hence,

$$Q^{M+1}(A \mid z_0) \geq \epsilon^M \prod_{m=0}^{M} \delta_{C(m)}\mathbf{1}_{a(m)+1\{m<m^*\}}(A_\nu(m))Leb(A_\zeta(m) \mid C(m))$$

letting $\delta_C := \min_m \delta_{C(m)} > 0$, the result follows. $\qquad \square$

Proof of Lemma L.6. Suppose z is such that $||\nu||_1 = T - l$ for some $l \in \{1, \ldots, T\}$.

It follows that

$$Q[V](z) = 1 + a \int ||\bar{\zeta}'||_1 Q(d\zeta' \mid z) + A \int ||\nu' - \partial V_T||_1 Q(d\nu' \mid z)$$

$$= 1 + a \sum_d \delta(z)(d) \int \left(|\bar{\zeta}'(d)| + \sum_{m \neq d} |\bar{\zeta}(m)|\right) Q(d\zeta' \mid D = d, z)$$

$$+ A \sum_d \delta(z)(d) \int ||(\nu'(d), \nu(-d)) - \partial V_T||_1 Q(d\nu' \mid D = d, z).$$

where $(\nu'(d), \nu(-d))$ denotes the vector where the d-th coordinate is $\nu'(d)$ and the rest of the coordinates are given by $\nu(-d)$. Because $||\nu||_1 = T - l$, it follows that $||(\nu(d) + 1, \nu(-d))||_1 = T - l - 1$ and thus $||(\nu(d) + 1, \nu(-d)) - V_T||_1 = l - 1 \leq \frac{T-1}{T}||\nu - V_T||_1$. Moreover, if $l = 0$, then $\nu'(d) = \nu(d)$ and $||(\nu'(d), \nu(-d)) - V_T||_1 = 0$.

Also, observe that $|\bar{\zeta}'(d)| \leq |\bar{y}|/(1 + \nu(d)) + |\bar{\zeta}(d)|\nu(d)/(1 + \nu(d))|$ and $\nu(d) \geq \nu_0$. Thus,

$$\sum_d \delta(z)(d) \int \left(|\bar{\zeta}'(d)| + \sum_{m \neq d} |\bar{\zeta}(d)|\right) Q(d\zeta' \mid D = d, z)$$

$$\leq |\bar{\zeta}(0)| \left(\sum_d \delta(z)(d)\omega(d, 0)\right) + \ldots + |\bar{\zeta}(M)| \left(\sum_d \delta(z)(d)\omega(d, M)\right)$$

$$+ a \sum_d \frac{\delta(z)(d)}{1 + \nu(d)} \int |\bar{y}| f_d(y) dy$$

where $\omega(d', d) = 1$ if $d' \neq d$ and $= (\nu(d))/(1 + \nu(d))$ if $d' = d$. Hence, for any m, $\sum_d \delta(z)(d)\omega(d, m) =$

$\delta(z)(m)\frac{v(m)}{1+v(m)} + (1-\delta(z)(m)) \le \epsilon \frac{v(m)}{1+v(m)} + (1-\epsilon) := \gamma(v(m))$ because $\delta(z)(d) \ge \epsilon$ by Assumption 2. Let $\gamma(v) := \max_m \gamma(v(m))$. Thus

$$
\begin{aligned}
Q[V](z) &= 1 + a\int ||\bar{\zeta}'||_1 Q(d\zeta' \mid z) + A\int ||v' - \partial V_T||_1 Q(dv' \mid z)\\
&\le 1 + a\gamma(v)||\bar{\zeta}||_1 + A\frac{T-1}{T}||v - \partial V_T||_1 + \sum_d \frac{\delta(z)(d)}{1+v(d)}\int |\bar{y}|\phi(y;\theta(d),1)dy\\
&\le \gamma(v)V(z) + (1-\gamma(v)) + a\sum_d \frac{\delta(z)(d)}{1+v(d)}\int |\bar{y}|f_d(y)dy.
\end{aligned}
$$

Since $\int |\bar{y}|f_d(y)dy \le 1$ and $v(d) > v_0(d)$ the desired result follows. $\qquad\square$

Lemma L.11. *Let $C \subseteq \mathbb{Z}$ and a, A be as in Lemma L.6 and let $\bar{\gamma}$ and $R := \inf_{z \in C^C} V(z)$ be such that $\bar{\gamma} > \gamma$ and*

$$\gamma + (1 - \gamma + 0.5a)/R \le \bar{\gamma}$$

Then, for all $z \in \mathbb{Z}$,

$$Q[V](z) \le \bar{\gamma}V(z) + b1_C(z)$$

Proof. From Lemma L.6, for all $z \in \mathbb{Z}$,

$$Q[V](z) \le \gamma V(z) + b,$$

Thus

$$
\begin{aligned}
Q[V](z) &\le \bar{\gamma}V(z) - (\bar{\gamma} - \gamma)V(z) + b\\
&\le \bar{\gamma}V(z) - (\bar{\gamma} - \gamma)R + b.
\end{aligned}
$$

We now show that $-(\bar{\gamma} - \gamma)R + b < 0$. To do this, note that

$$
\begin{aligned}
-(\bar{\gamma} - \gamma)R + b &= (1 - \gamma) + 0.5a - (\bar{\gamma} - \gamma)R\\
&= 1 + \gamma(R-1) + 0.5a - \bar{\gamma}R.
\end{aligned}
$$

$\qquad\square$

Proof of Lemma L.7. This proof follows the proof of Theorems 9 and 12 in Roberts and Rosenthal (2004) (RR).

21

Lemma L.5 imply their condition 8 and Lemma L.6 with C chosen as in Remark L.2 implies their condition 10. Moreover, proposition 11 in RR holds with $\alpha := \alpha(a, R, \gamma)^{-1} := \gamma + (1 - \gamma + 0.5a)/(R + 1)$, which by Remark L.2 is less than 1; moreover, it is increasing in a, decreasing in R and decreasing on γ. Thus as pointed out in p. 47, their Theorem 12 holds. Thus,

$$||Q^n(.|z_0) - \pi|| \leq (1 - \epsilon^{M+1} \delta_C)^j + \alpha^{-n} B_M^{j-1} 0.5(V(z_0) + \int V(z)\pi(dz))$$

$$=: LV(z_0)\left((1 - \epsilon^{M+1} \delta_C)^j + \alpha^{-n} B_M^j\right)$$

for any $0 \leq j \leq n$, where $B_M := \max\{1, \alpha^{M+1}(1 - \epsilon^{M+1} \delta_C)\sup_{z \in C} Q^{M+1}[V](z)\}$.

We now bound the term $\alpha^{-n} B_M^j$. Observe that $\sup_{z \in C} Q^{M+1}[V](z) \leq \gamma^{M+1} \sup_{z \in C} V(z) + \frac{1 - \gamma^{M+1}}{1 - \gamma} b \leq \gamma^{M+1}((1 + R) - b) + \frac{b}{1 - \gamma}$ (recall that $b = 1 - \gamma + 0.5a$ and so

$$(1 - \epsilon^{M+1} \delta_C) \sup_{z \in C} Q^{M+1}[V](z) \leq (1 - \epsilon^{M+1} \delta_C)\left(\gamma^{M+1}((1 + R) - b) + \frac{b}{1 - \gamma}\right).$$

Abusing notation, we redefine γ as $\gamma := \max\{(1 - \epsilon^{M+1} \delta_C), \gamma\}$ and let $R(a, \eta, \gamma) = 0.5a/(1 - \gamma) + \eta$ for some $\eta > 0$. It is easy to see that the previous display is bounded by

$$G(a, \eta, \gamma) := \gamma\left(\gamma^{M+1}(0.5a/(1 - \gamma) + \eta + \gamma - 0.5a) + 1 + 0.5a/(1 - \gamma)\right)$$

$$= \gamma\left(\gamma^{M+1}(0.5a\gamma/(1 - \gamma) + \eta + \gamma) + 1 + 0.5a/(1 - \gamma)\right),$$

which is increasing in all of its arguments. Therefore,

$$\alpha^{-n} B_M^j \leq \alpha(a, R(a, \eta, \gamma), \gamma)^{-n+j(M+1+m)} (G(a, \eta, \gamma)/\alpha(a, R(a, \eta, \gamma), \gamma)^m)^j, \; \forall m \in \mathbb{N}.$$

We note that

$$G(0, \eta, \gamma) = \gamma\left(\gamma^{M+1}(\eta + \gamma) + 1\right) \; and \; \alpha(0, R(0, \eta, \gamma), \gamma)^{-1} := \gamma + (1 - \gamma)/(\eta + 1) < 1.$$

Therefore, there exists a $m := m(\gamma, \eta)$ such that $G(0, \eta, \gamma)/\alpha(0, R(0, \eta, \gamma), \gamma)^{m(\gamma, \eta)} < 1$. It is straightforward to show that $m(., .)$ is non-decreasing on γ and non-decreasing on η (at least for large values of η). By continuity, there exists a $a(\gamma, \eta)$ such that for all $a \leq a(\gamma, \eta)$, $G(a, \eta, \gamma)/\alpha(a, R(a, \eta, \gamma), \gamma)^{m(\gamma, \eta)} \leq 1$. Moreover, $a(., .)$ is non-increasing in γ.

Therefore, for any $\eta > 0$ and $a \leq a(\gamma, \eta)$, it follows that

$$\alpha^{-n} B_M^{j-1} \leq \alpha(a, R(a, \eta, \gamma), \gamma)^{-n+j(M+1+m(\gamma, \eta))}.$$

By choosing $j = 0.5n/(M+1+m(\gamma,\eta))$ (if it is not an integer, simply take the floor), it follows that

$$||Q^n(.|z_0) - \pi|| \leq LV(z_0))\alpha(a,R(a,\eta,\gamma),\gamma)^{0.5\frac{n}{M+1+m(\gamma,\eta)}} = LV(z_0))\exp\{-0.5n\frac{\log\alpha^{-1}(a,R(a,\eta,\gamma),\gamma)}{M+1+m(\gamma,\eta)}\}$$

Observe that by the computing the total derivative of $\gamma \mapsto \log\alpha^{-1}(a,R(a,\eta,\gamma),\gamma)$ with respect to γ it can be shown that this function is decreasing on γ. Since $\gamma \mapsto m(\gamma,\eta)$ is non-decreasing, it follows that $\gamma \mapsto \frac{\log\alpha^{-1}(a,R(a,\eta,\gamma),\gamma)}{M+1+m(\gamma,\eta)}$ is decreasing on γ. Since γ is decreasing on ϵ and increasing on T, it follows that $\epsilon \mapsto \varrho(\epsilon) := \frac{\log\alpha^{-1}(a,R(a,\eta,\gamma),\gamma)}{M+1+m(\gamma,\eta)}$ is increasing on ϵ. $\qquad\square$