Reinforcing RCTs with Multiple Priors while Learning about External Validity *

Frederico Finan †  Demian Pouzo‡
UC Berkeley  UC Berkeley

December 2021

Abstract

This paper presents a framework for how to incorporate prior sources of information into the design of a sequential experiment. This information can come from many sources, including 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 our framework 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.

*We thank Marina Dias for useful comments. We also thank Mengsi Gao for excellent research assistance. Usual disclaimer applies.
†Department of Economics, 508-1 Evans Hall, Berkeley, California 94720-3880. Email: ffinan@berkeley.edu; and BREAD, IZA, NBER
‡Department of Economics, 508-1 Evans Hall, Berkeley, California 94720-3880. Email: dpouzo@econ.berkeley.edu
# Table of Contents

1 **Introduction**  

2 **Setup**  
   2.1 General Model  
   2.2 The Markov Gaussian Model for Average Outcomes  

3 **Analytical Results for the Markov Gaussian Model**  
   3.1 Properties of the Stochastic Process for Outcomes and Treatment Assignments  
   3.2 Finite Sample Results  
      3.2.1 Concentration bounds on the Posterior Mean  
      3.2.2 Probability of making a mistake  
      3.2.3 Average Observed Outcomes  

4 **Model Simulations**  

5 **Conclusions**  

6 **Figures**  

A **General Learning Model**  
   A.1 A special Case: The model in the text  

B **Properties of \( \alpha \)**  
   B.1 Proof of Proposition 1  
   B.2 Proof of Lemma 1  
   B.3 Bounds for \( \alpha_t \)  
   B.4 Proof of Supplemental Lemmas  

C **Stochastic Properties of the Bayesian Posteriors and weights**  
   C.1 The Markov Chain for \( (Z_t)_t \)  
   C.2 Properties of the transition probability function \( Q \)
C.2.1 Small sets ................................................................. 61
C.2.2 Drift Condition ............................................................ 67
C.2.3 Geometric Ergodicity ..................................................... 69
C.3 Mixing Results ................................................................. 71

D Appendix for Section 3.2.1 72
D.1 Bounds for Posterior Means ................................................. 72
D.2 Concentration Bounds for \( (f_t), (J_t), \) and \( (m_t) \) ............ 74
D.3 Proof Proposition 3 and Corollary 1 ..................................... 76
D.4 Proofs of Supplemental Lemmas ........................................... 78

E Appendix for Section 3.2.2 83

F Appendix for Section 3.2.3 93
F.1 Proofs of Lemmas ............................................................. 95
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 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 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.

At the heart of this decision lies two issues. 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 $\epsilon$-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.

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 priori information about these distributions, which can arise from many sources, including her own introspection and knowledge, previous experiments, or expert opinions.

---

To model this problem, we take a multi-armed bandit setup and enhance it 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. As in most randomized control trials, our policymaker aims to learn the average treatment effects. As the policymaker gathers more data, she updates each of these priors using Bayes’ rule and then aggregates each source’s posterior according to their posterior model probability. On the basis of these beliefs, the policymaker then decides whether to stop the experiment and which treatment to assign.

In these types of sequential experiments, it is common for the policymaker to not use the optimal assignment rule. It is well known that in settings in which the policymaker must learn the truth, the optimal assignment 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.\(^2\) As a result, the literature on multi-armed bandits have studied certain properties of different heuristic strategies such as $\epsilon$-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)). We take a different approach and study a class of assignment rules, in which the policy functions are Markov – it does not depend on the stage – and whose probability of choosing any treatment is bounded away from zero. While these restrictions are not innocuous, especially the second one, they are sufficiently general to encompass, under certain conditions, many of the commonly-used solutions in multi-armed bandit problems, including those aforementioned.

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

\(^2\)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 policy maker, however, does not know the true average effect of treatment B and incorrectly believes that it is negative. In this simple example, the optimal policy function never assigns 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.
theoretically and via Monte Carlo simulations, that our setup exhibits several nice finite sample properties, including robustness to incorrect priors.

More precisely, we show that our policymaker will learn the average treatment effects, in the sense that the posterior mean concentrates around the true mean, and it does so at a rate of $\sqrt{\log t/t}$, where $t$ is the number of stages. That this result holds was not, ex ante, at all obvious: in contrast to 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 assumption that any treatment is played with positive probability, we are able to bring to bear results from Markov Chain theory to show that the data dependence vanishes “fast” (formally, that is $\beta$-mixing with exponential decay). In addition, we employ concentration inequalities for dependent processes to not only obtain the rate of $\sqrt{\log t/t}$, but also to characterize and quantify how this rate depends on the initial parameters of the setup, such as the amount of experimentation. Indeed, we show that the more the policymaker experiments, the faster she learns the average effects of the treatments.

Our technique of proof also allows us to quantify how the concentration rate is affected by the different priors. Importantly, we are able to show that our aggregation method exhibits an attractive robustness property. 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. 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. 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, which 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 a 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 probability of mistake below it.

Finally, we also characterize the behavior of the average observed outcomes by computing 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, in effect, governed by an “exploitation versus exploration” tradeoff. If we increase the degree of experimentation (less exploitation, more exploration) our data become more independent and we converge more quickly. 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 (the case of an RCT) is unlikely to be optimal, and we verify this numerically in a series of simulations.

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. 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.

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

---

3See Athey and Imbens (2019) for a survey of machine learning techniques as it applies to experimental design and problems in economics.
alternative estimators have interpreted the out-of-sample prediction errors as a measure or test of external validity.⁴ A related approach uses local average treatment effects across different complier populations to test for evidence of external validity (e.g. \textit{Angrist and Fernández-Val (2013); Kowalski (2016); Bisbee et al. (2017)}). The general idea being that if differences in observable characters 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. \textit{Vivalt (2020), Gelman and Carlin (2014), Gelman and Pardoe (2006), Meager (2020)}).⁵

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 Kullblack-Leibler (KL) divergence criteria. Importantly, our definition offers a way to quantify or rank external invalidity among models. Second, we provide a link between this ranking of external invalidity and our aggregation method. We shows 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 largely 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. \textit{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 (\textit{Vivalt and Coville, 2021}).

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 Markov Gaussian model. In Section 3, we provide analytical results for the Markov Gaussian model. We then illustrate the

---

⁴See for example Bo and Galiani (2021), 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.

⁵The 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.
main analytical results by simulation in Section 4. Section 5 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, $\mathcal{X}$; a set of treatments $\mathcal{D} := \{0, \ldots, M\}$; 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 $(d, x) \in \mathcal{D} \times \mathcal{X}$, let $Y_t(d, x) \in \mathbb{R}$ denote the potential outcome associated with treatment $d$ and characteristic $x$ in instance $t$; also, let $Y_t(d) := (Y_t(d, x))_{x \in \mathcal{X}}$. Let $D_t(x) \in \mathcal{D}$ 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, $t \in \{1, \ldots, T\}$, the PM is confronted with $|\mathcal{X}| < \infty$ units, one for each value of the observed characteristic. We assume this 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. 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 $t$. 

6
– 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 \( t \in \{1, \ldots, T\} \) and each \( x \in \mathcal{X} \), \( (Y_t(d,x))_{d \in \mathcal{D}} \) is drawn IID where \( Y(d,x) \sim P(\cdot|d,x) \in \Delta(\mathbb{R}) \).

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.

**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 \((d,x) \in \mathcal{D} \times \mathcal{X}\), 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 \( Y(d,x) \). Specifically, suppose the PM has \( L + 1 \) prior beliefs, \( (\mu^o_0(d,x))_{o=0}^L \), 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. By convention, we use \( o = 0 \) to denote the PM’s own prior and leave \( o > 0 \) to denote the other sources.

For each \((d,x) \in \mathcal{D} \times \mathcal{X}\), 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

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

for any Borel set \( A \subseteq \Theta \). Observe that the belief is updated using observed data, \((Y_t(D_t(x),x), D_t(x))\). Given \( D_t(x) = d \), the observed outcome corresponds to \( Y(d,x) \) and thus the conditional PDF of \( Y_t(x) \) given \( D_t(x) = d \) is described by \( p_\theta \). 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 was to learn features of the joint distribution — e.g., the correlation between potential outcomes — then we would have to modify the learning model. The subjective model would now be a family of probability distributions over \( (Y(d,x))_{d \in \mathbb{D}} \). We present such a model in Appendix A, 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, \( \{\langle \mathcal{P}_{d,x}, \mu_0^\circ_t (d,x) \rangle \}_{\circ = 0}^L \), 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 \( \circ \) best fits the observed data within the class of models being considered, i.e.,

\[
\mu_t^\circ_t (d,x)(A) := \sum_{\circ = 0}^L \alpha_t^\circ (d,x) \mu_t^\circ_t (d,x)(A)
\]

for any Borel set \( A \subseteq \mathbb{R} \), where

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

We can interpret \( \alpha_t^\circ (d,x) \) as a measure of the PM’s subjective probability that the prior belief associated with source \( \circ \) 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

\[\text{Because the PM already knows the probability of } D_t(x), \text{ she does not need to include it as part of the Bayesian updating problem.}\]
For each \((x, y) \in D \times X\) and \(P_d, x\), let

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

be the Kullback-Leibler (KL) divergence, which acts as a notion of distance between the true PDF of \(Y(x)\) — given by \(p(\cdot | d, x)\) — and a “subjective” one \(p_\theta \in P_d, x\). 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, \(\langle P_d, x, \mu_0(d, x) \rangle\), and in turn, to propose a definition of external (in)validity.

**Definition 1** (Externally Invalid Subjective Bayesian Model). We say a subjective Bayesian model \(\langle P_d, x, \mu_0(d, x) \rangle\) is externally invalid for \((x, y)\) if

\[
u_0(d, x) > 0
\]

where \(\nu_0(d, x)\) is the smallest \(u \geq 0\) such that \(\mu_0^u(d, x) (\theta : KL_d, x(\theta) \leq u) > 0\).

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 \(\nu_0(d, x) > 0\) exists, we say the subjective Bayesian model is externally valid for \((d, x)\). Moreover, \(\nu_0(d, x)\) quantifies how far the true PDF is from the closest PDF within the subjective Bayesian model.

To illustrate this definition, consider the following graph below. The horizontal axis indicates different values of \(\theta\), where the \(\theta\) of the true PDF is set at the origin. The vertical axis corresponds to the resulting KL distance between the true PDF of \(Y(x)\) and the “subjective” one \(p_\theta\). As the line labeled \(KL_d, x\) depicts, \(\theta\)s further from the origin are associated with larger distances. In addition, we also plot a subjective belief \(\mu^0(d, x)\) over the set of \(\theta\)s, for which the model places positive probability. According to our definition, the model \(\mu^0(d, x)\) is externally invalid because as the graph depicts there exists a \(\nu_0(d, x) > 0\), such that \(\mu_0^u(d, x) (\theta : KL_d, x(\theta) \leq u) > 0\).

---

\(^7\)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 triangle inequality. Finally, the KL divergence does not depend on \(o\) as all models are assumed to have the same family \(P_d, x\).
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 $P(\cdot|d,x)$.\(^{8}\) Second, this definition 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 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 $(a_t^o(d,x))_{t=0}^\infty$. It shows that, as $t$ diverges, the weights are only positive for the least externally invalid models.\(^{9}\)

**Proposition 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. Then, for any $(d,x) \in \mathcal{D} \times \mathcal{X}$, if model $o'$ is less externally valid

---

\(^{8}\)For an alternative formalization of external validity, see for example Bo and Galiani (2021) and references therein.

\(^{9}\)Lemma 3 in the Appendix B provides a non-asymptotic version of this proposition.
than model 0, then

\[
\frac{\alpha_o^\theta(d, x)}{\alpha_i^\theta(d, x)} = o_P(1)
\]

Proof. See Appendix B.

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 \( \alpha_o^\theta(d, x) \) 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 way 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.

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 \( G \in \mathcal{X} \).

The first policy function, \( \sigma_t(y^{t-1}, d^{t-1}) \in [0, 1] \), specifies the probability of stopping the experiment given the observed history \( y^{t-1}, d^{t-1} \). The second policy function, \( \delta_t(y^{t-1}, d^{t-1})(\cdot|x) \in \Delta(D) \) for each \( x \in \mathcal{X} \), specifies the probability distribution over treatments for each \( x \in \mathcal{X} \); i.e., \( \delta_t(y^{t-1}, d^{t-1})(d|x) \) is the probability that \( x \in \mathcal{X} \) receives treatment \( d \) given the past history.

Hence, the policy rule defines two consecutive stages in the experiment: A last stage of exploitation where the experiment has been stopped and what is thought to be the best treatment is chosen, and a first stage of exploitation and exploration. How the PM regulates this tradeoff will be discussed in more detail in the following section.

2.2 The Markov Gaussian Model for Average Outcomes

The general setup provides (in our opinion) 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 problem. It requires characterizing the process of the posterior beliefs \( (\mu_t^o(d, x))_{t=0}^T \) for each model \( o \), which is an infinite dimensional object. Moreover, the policy functions depend
on time, which introduces another level of time non-homogeneity. Therefore, we focus instead on a setup involving Gaussian subjective models, policy functions that are (mostly) Markov, and a PM who is interested in the average effects of treatments.

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 $\theta$ such that $\theta = E_{P(.|d,x)}[Y(d,x)]$. We show this is the case in Section 3.2.1.

Henceforth, for each $(d,x) \in D \times X$, let

$$\theta(d,x) := E_{P(.|d,x)}[Y(d,x)],$$

which will be the object of interest for the PM.

Formally, the Gaussian learning model is constructed assuming that, for each $(d,x) \in D \times X$, $P_{d,x}$ is a family of Gaussian PDFs given by $\{\phi(.; \theta, 1) : \theta \in \mathbb{R}\}$ and the prior for every source is also assumed to be Gaussian with mean $\zeta_0^o(d,x)$ and variance $1/\nu_0^o(d,x)$. The quantity $\nu_0^o(d,x)$ can be interpreted as the number of units with characteristics $G$ that were assigned treatment $3$ in a past experiment. The higher the $\nu_0^o(d,x)$, the more certain source $o$ is about $\phi(.; \zeta_0^o(d,x), 1)$ being the correct model. Throughout, we will assume $\zeta_0^o$ and $\nu_0^o$ 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 $t \geq 1$,

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

$$= \frac{1}{N_t(d,x) + \nu_0^o(d,x)} \sum_{s=1}^{t} Y_s(d,x) 1\{D_s(x) = d\} + \frac{\nu_0^o(d,x)}{N_t(d,x) + \nu_0^o(d,x)} \zeta_0^o(d,x)$$

$$= \frac{N_t(d,x)}{N_t(d,x) + \nu_0^o(d,x)} m_t(d,x) + \frac{\nu_0^o(d,x)}{N_t(d,x) + \nu_0^o(d,x)} \zeta_0^o(d,x)$$

(2) and

$$\nu_t^o(d,x) = \nu_{t-1}^o(d,x) + 1\{D_t(x) = d\} = N_t(d,x) + \nu_0^o(d,x),$$

(3)

(4)

---

10Throughout, $\phi(.; \theta, \sigma^2)$ is the Gaussian PDF with mean $\theta$ and variance $\sigma^2$.11
where \( N_t(d,x) := \sum_{s=1}^t 1\{D_s(x) = d\} \) and \( m_t(d,x) := \frac{\sum_{s=1}^t 1\{D_s(x) = d\} Y_s(d,x)}{N_t(d,x)} \).

From these expressions – and the assumption of Markov policy function (see below for details) – we can see how Gaussianity simplifies the dynamics of the problem; we only need to analyze \((z_t^0(d,x), v_t^0(d,x))_{t=0}^T\), a finite dimensional object with a Markov structure (see Section 3 for details), as opposed to \((\mu_t^0(d,x))_{t=0}^T\), an infinite dimensional object that is quite intractable.

**Remark 1.** Our results and methodology clearly extend to any subjective model whose posterior beliefs can be fully described by low finite-dimensional objects. For instance, in cases where the instance posterior is given by a Dirichlet density with parameters given by \( \alpha_t(j) = \alpha_0(j) + \sum_{s=1}^t 1\{D_s(x) = d\} 1\{Y_s(d,x) = j\} \) for any \( j \in \{1, \ldots, J\} \).

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. \( \Delta \)

As we 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 externally valid models. This is not an issue within a “frequentist setup”, where the priors are degenerate, as any two externally valid models are identical.\(^{11}\) 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 1.** For any \( o \in \{0, \ldots, L\} \), any \( t \geq 1 \), and any \((d,x) \in \mathcal{D} \times \mathcal{X}\),

1. 
\[
\alpha_t^o(d,x) := \frac{\phi(m_t(d,x) - z_0^o(d,x); 0, (N_t(d,x) + v_0^o(d,x))/(N_t(d,x)v_0^o(d,x)))}{\sum_{o=0}^L \phi(m_t(d,x) - z_0^o(d,x); 0, (N_t(d,x) + v_0^o(d,x))/(N_t(d,x)v_0^o(d,x)))}.
\]

2. \( \lim_{|z_0^o(d,x) - m_t(d,x)| \to \infty} \alpha_t^o(d,x) = 0 \).

3. \( \alpha_t^o(d,x) = \frac{\phi(\sqrt{v_0^o(d,x)}(\theta(d,x) - z_0^o(d,x)); 0, 1) \sqrt{v_0^o(d,x)}}{\sum_{o=0}^L \phi(\sqrt{v_0^o(d,x)}(\theta(d,x) - z_0^o(d,x)); 0, 1) \sqrt{v_0^o(d,x)}} + o_p(1) \).

**Proof.** See Appendix B. \( \square \)

\(^{11}\)By “identical” we mean that each model has a component, \( p^o \) and \( p'^o \) that nullifies the KL divergence.
The lemma characterizes $\alpha^o_t(d,x)$ as the odds ratio of Gaussian PDFs, which with probability approaching one are evaluated at $\theta(d,x) - \zeta^o_0(d,x)$ with mean 0 and variance $1/v^o_0(d,x)$. Moreover, it also illustrates how $\alpha^o_t$ offers certain robustness properties against lack of external validity. That is, if the prior mean, $\zeta^o_0(d,x)$, is “far away” from $m_t(d,x)$ — which, with enough observations, approximates the true $\theta(d,x)$ with high probability — 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.

Part (3) of the lemma complements Proposition 1 for the Markov Gaussian learning model. It offers an asymptotic characterization of the weights when all the models are, according to our definition, externally valid. It shows that not only will $\alpha^o_t(d,x)$ not equal 0 or 1 even with infinite data, but it also suggests a partial ordering among externally valid sources. To see this, it is useful to introduce some nomenclature: We call $|\zeta^o_0(d,x) - \theta(d,x)|$ the bias of model $o$, and $v^o_0(d,x)$, the degree of conviction. We call $|\zeta^o_0(d,x) - \theta(d,x)|/\sqrt{v^o_0(d,x)}$, the degree of stubbornness of model $o$ and a model with zero stubbornness and high conviction, confident. Part (3) indicates that, asymptotically, $\alpha^o_t(d,x)$ will put more weight on models that are less stubborn and more confident. In particular, if $v^o_0(d,x)$ diverges — intuitively, if all the prior sources have large sample sizes —, $\alpha^o_t(d,x)$ will concentrate around the least stubborn model.

**Policy Rule.** We now turn our attention to the policy rule. Our policy rule consists of two policy functions: the one that assigns treatments and the one that stops the experiment. The policy function that assigns treatments, $\delta$, is assumed to be Markov. For any past history $(y^{t-1}, d^{t-1})$,

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

where $\zeta_t := (\zeta^o_t)_{o=0}^L$ (the other variables are similarly defined). We require the policy function $\delta$ to be time homogeneous, only depending on the state – $(\mu_t, \alpha_t)$. When we derive our analytical results, this assumption ensures a Markov structure for our state variables and other variables of interest. We also impose the following additional assumption:

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

Under this assumption, each treatment arm is chosen with positive probability, thus ensuring some experimentation. This assumption will be key for deriving the results in Section 3.

Given these assumptions, we now present some examples of policy rules that are admissible in
our framework; we also discuss what type of behavior our framework does not accommodate.\footnote{We refer the reader to Wager and Xu (2021) for more examples of policy functions in a similar setup.}

Recall that our PM wants to discover the average effect of each treatment. At each instance $t$, the (subjective) average effect of treatment $d$ for unit $x$ is given by

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

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 $\theta$ is weighted according to the posterior belief defined in Expression 1. By re-arranging the order of the integrals, it follows that

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

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

Hence, the (subjective) average effect of treatment $d$ on unit $x$ at instance $t$ is given by $\xi_t^a (d,x)$. Thus, the PM uses this quantity to assign treatment. In Section 3.2, we establish some finite sample properties of this quantity, such as the rate at which it concentrates around the true average effect.

**Example 1** (Epsilon-Greedy Policy Function). A commonly-used policy function that is admissible in our framework is the so-called Epsilon-Greedy policy function:

$$\delta (\xi_t, v_t, \alpha_t) (d|x) = (M+1) \epsilon \frac{1}{M+1} + (1 - (M+1) \epsilon) 1\{d = \arg \max_a \xi_t^a (a,x)\}. \quad (5)$$

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. \triangle

**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 $\sum_{x \in \mathcal{X}} \xi^a (d,x)$ (or some other aggregator for $x$). Our framework does not allow for such policy function because it is not Markov. One could also consider the infinite time-horizon version of this problem (i.e., $T = \infty$), which is Markov, but is unlikely to satisfy Assumption 2. Instead, our framework allows for a “perturbed” version of
Finally, we provide an example of the policy rule for stopping the experiment, \( \sigma \). While most of our results do not require any restriction on this policy rule, 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 \( \beta \). Below, we propose a family of stopping rules for which it will be shown in Section 3.2.2 that, by appropriately choosing certain parameters, the rule has such desirable property.

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

\[
\sigma_t(Y^{t-1}, D^{t-1}) = 1, \text{ if } \max_d \min_{m \neq d} \left\{ \zeta_t^d - \zeta_{t-1}^m \right\} \geq c_{t-1}(\gamma_{t-1}, d, m) > 0,
\]

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

\[
c_t(\gamma_t, d, m) := \sum_{\alpha=0}^L \gamma_t \frac{\alpha_t^\alpha(d)}{f_t(d) + \nu_0^\alpha(d) / t} + \sum_{\alpha=0}^L \gamma_t \frac{\alpha_t^\alpha(m)}{f_t(m) + \nu_0^\alpha(m) / t}
\]

---

13This 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.
While the expression for the cutoff is a bit involved, the constant $\gamma_t$ 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 $\max_d \min_{m \neq d} (\xi_t^a (d) - \xi_t^a (m))$ — is far enough from zero, where “far enough” is essentially measured by $\text{Constant} \times \gamma_t$. 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 $\gamma_t$ should be of order $1/\sqrt{t}$, however in this problem, because the data are not IID, the correct order is $\sqrt{\log t/t}$; see Section 3.1 for a more through discussion. △

3 Analytical Results for the Markov Gaussian 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 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 taking 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. Thus, we impose the following assumption:

**Assumption 3.** (i) There exists a $\nu < \infty$ such that for any $\lambda > 0$ and any $d \in \mathbb{D}$, $E[e^{\lambda (Y(d) - \theta(d))}] \leq e^{\nu \sigma(d)^2 \lambda^2}$ where $\sigma(d)^2 := \text{Var}(Y(d))$; (ii) $Y(d)$ admits a PDF, denoted as $f_d$ that has full-support.

Part (i) of this assumption imposes that $Y(d)$ is sub-gaussian, which loosely speaking, 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 19 in the Appendix D.3). We could relax this assumption, but at the cost of getting slower concentration rates; see Remark 7 in the Appendix D.3 for more details. Part (ii) of the assumption is necessary for characterizing the properties of the stochastic process for the experiment’s data.

We now begin by presenting a key proposition that characterizes the dependence structure of the experiment’s data.
3.1 Properties of the Stochastic Process for Outcomes and Treatment Assignments

To derive finite sample results for our model, we need to first understand how the treatment assignments, \((D_t)_{t=1}^T\), and the realized outcomes, \((Y_t(D_t))_{t=1}^T\), evolve over time. In a standard randomized control trial setting, it is straightforward to characterize these processes: \((D_t)_{t=1}^T\) are IID random variables by construction, which implies the same for the realized outcomes \((Y_t(D_t))_{t=1}^T\). This is not the case in our setup. Because our policy functions depend on the PM’s beliefs, and hence on all past observed history, \((D_t)_{t=1}^T\) are no longer IID random variables. This complicates matters because we cannot establish our results by simply applying the law of large numbers, the central limit theorem, or other exponential inequalities for IID random variables. Thus to overcome this challenge, we first determine the properties of the stochastic process \((\zeta_t, \nu_t, \alpha_t)_{t=1}^T\), where, recall, \(\zeta_t := (\zeta_t^0)_{t=0}^L\) (the other variables are defined analogously). We then extrapolate these properties to \((D_t)_{t=1}^T\) and \((Y_t(D_t))_{t=1}^T\) to derive the necessary exponential inequalities.

The stochastic process \((\zeta_t, \nu_t, \alpha_t)_{t=1}^T\) has two important properties. First, as we show in Lemmas 6 and 7 in Appendix C.3, we only need to study the process for \(o = 0\), i.e., \((\zeta_t^0, \nu_t^0)\), because for any \(o \in \{0, ..., L\}\) and any \(t\), \((\zeta_t^o, \nu_t^o, \alpha_t^o)\) can be written as a function of \((\zeta_t^0, \nu_t^0)\) and the priors \((\zeta_0, \nu_0)\), which are taken to be non-random. Second, under the policy rule \(\delta\) described above, we show that \((\zeta_t^0, \nu_t^0)\) is a (non-stationary) Markov chain with transition probability function implied by expressions 3-4, and \((\zeta_0, \nu_0) \sim \pi\) where \(\pi\) is assumed to be degenerated; i.e., the priors are taken to be non-random.\(^{14}\) Based on this insight, we use known results from Markov Chains (see Douc et al. (2018) for a review) to understand the dependence structure of \((\zeta_t^0, \nu_t^0)\) and consequently that of \((D_t, Y_t(D_t))\). We summarize this in the following proposition:

**Proposition 2.** The process \((\zeta_t, \nu_t, D_t, Y_t(D_t))\) is \(\beta\)-mixing with mixing coefficients, \((\beta(k))_k\) such \(\beta_k = O\left(e^{-0.5k\varrho(\epsilon)}\right)\) for all \(k \geq 1\) where \(\epsilon \mapsto \varrho(\epsilon)\) is positive valued, \(\varrho(0) = 0\) and increasing on \(\epsilon\).\(^{15}\)

**Proof.** See Appendix C.3.

This proposition establishes that even though the process \((D_t, Y_t(D_t))\) is not IID, its dependence “dies off” at an exponentially fast rate governed by the parameter \(\epsilon\). The intuition behind the role of \(\epsilon\) is that, under Assumption 2, the policy function \(\delta\) can be thought of as follows: With probability \((M + 1)\;\epsilon\), the treatment is chosen at random, and with probability \(1 - (M + 1)\;\epsilon\), the treatment is chosen according to a function that depends on \((\zeta_t, \nu_t)\) and \(\alpha_t\). In effect, \(\epsilon\) captures the probability

\(^{14}\)More generally, \(\pi\) can be non-degenerate as long as it satisfies \(\int ||\zeta_0||^2 \pi(d\zeta_0) < \infty\).

\(^{15}\)The formal definition for \(\varrho\) is relegated to Lemma 14 in Appendix C.2.3.
that the Markov Chain for \((\xi_t, \nu_t)\) “forgets the past”. The larger the \(\epsilon\), the more likely the Markov Chain will “forget the past”, and the faster the dependence vanishes.

The proof of this proposition is (to our knowledge) novel, requiring different known results from Markov Chain theory and Mixing properties for non-stationary Markov Chains. In particular, we first show that the Markov Chain is Geometric Ergodic — i.e., for any initial condition, the \(t\)-step transition probability gets close (at a geometric rate) to the invariant probability (see Appendix C.2 for a formal definition). Since \((\xi_t, \nu_t)\) are not discrete random variables and the state space is unbounded, establishing Geometric Ergodicity requires showing that there exists subsets of the state space that once the Markov chain visits them it “forgets the past”; these are called “small” sets and their existence is established in Lemma 11 in the Appendix C.2; Assumption 3(ii) is used to achieve this result. In addition, we show that the Markov Chain tends to drift towards such sets — this is proved in Lemma 12 in the Appendix C.2. Geometric Ergodicity then follows by invoking the results by Roberts and Rosenthal (2004) and it is proven in Lemma 14 in the Appendix C.2. Given this result, we invoke the results by Liebscher (2005) for non-stationary Markov Chains to show that \((\xi_t, \nu_t, D_t, Y_t(D_t))\) becomes independent of the past as \(t\) diverges; formally, we show that the process \((\xi_t, \nu_t, D_t, Y_t(D_t))\) is \(\beta\)-mixing with exponential rate of decay.\(^{16}\)

### 3.2 Finite Sample Results

Given Proposition 2, we can now derive concentration bounds for the posterior mean, characterize the probability of making a mistake, and place bounds on the average outcomes. We also show how these quantities are affected by the initial priors of the model, \((\xi_0, \nu_0)\), and the \(\epsilon\) parameter of the policy function.

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 \(\zeta_t^o = \sum_{o=0}^M \alpha_t^o \xi_t^o\), 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 \(\theta\). For each treatment \(d\), the randomness of \(\zeta_t^o(d)\) comes from two quantities: the frequency of play, \(f_t(d) = t^{-1} \sum_{s=1}^t \mathbb{1}\{D_s = d\}\) and the treatment-outcome average, defined as

\[
J_t(d) := t^{-1} \sum_{s=1}^t \mathbb{1}\{D_s = d\} Y_s(d).
\]

---

\(^{16}\) An stochastic process \((W_t)_{t=1}^\infty\) jointly distributed according to \(P\) is \(\beta\)-mixing if \(\lim_{k \to \infty} \beta_k := \lim_{k \to \infty} \sup ||P_{t+k} - P_{t,k+1:t,\infty}||_\mathcal{F}_\mathcal{V} = 0\), where \(P_{a:b,c:d}\) denotes the probability of \((Z_d, ..., Z_c, Z_b, ..., Z_a)\) and \(\cdot\) denotes the product of two probabilities.
Thus, to derive the concentration rate of \( \zeta_t^a(d) \), we first need to understand how \( f_t(d) \) and \( J_t(d) \) concentrate. This is where Proposition 2 comes in. From this proposition and using exponential inequalities for dependent processes (e.g. Merlevède et al. (2009)), we can determine how fast the frequency of play, \( f_t(d) \), concentrates around the average propensity score at time \( t \), \( e_t(d) := t^{-1} \sum_{s=1}^{t} P_{\pi_s}(D_s = d) \) and how fast the average treatment-outcome for each treatment, \( J_t(d) \), concentrates around \( f_t(d)\theta(d) \). These concentration bounds are presented in Lemmas 18 and 19 in the Appendix D.2.

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 \( \zeta_t^a(d) \) and how the parameters of the model affect this rate. In particular, take any \( \gamma > 0 \) and suppose \( J_t(d) \) and \( f_t(d) \) are within \( \gamma \) of \( f_t(d)\theta(d) \) and \( e_t(d) \) respectively. We would like to know how the concentration rate of these quantities — given by \( \gamma \) — translates into the concentration rate of \( \zeta_t^a(d) - \theta(d) \). To answer this question, we rely on Lemma 17 in the Appendix D to show that

\[
|\zeta_t^a(d) - \theta(d)| \leq \Gamma(\gamma, |\zeta_0(d) - \theta(d)|, \nu_0(d), e_t(d))
\]

where \( \Gamma : \mathbb{R} \times \mathbb{R}^{L+1} \times \mathbb{N}^{L+1} \times [0,1] \rightarrow \mathbb{R} \), defined in Appendix D. Moreover, Lemma 17 in the Appendix D shows that \( \Gamma \) is non-decreasing as a function of \( \gamma \).

Finally, the results below make use of the functions \( \epsilon \mapsto C(\epsilon) \) and \( \epsilon \mapsto B(\epsilon) \) that are defined in Appendix D.3. Here, we just point out that \( C \) is non-decreasing.

### 3.2.1 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.** For any \( d \in \{0, \ldots, M\} \), any \( t \in \mathbb{N} \) and any \( \epsilon \geq 0 \) such that \( t \geq e^{\max\{4\sqrt{\frac{\epsilon}{t C(\epsilon)}}, 2\nu_0(d)\}} \) and \( \frac{\epsilon}{e_t(d)c_C(\epsilon)} \leq \frac{t}{\log t} \),

\[
P_{\pi_t}\left(|\zeta_t^a(d) - \theta(d)| > \Gamma\left(\sqrt{\log t / \epsilon}, \sqrt{\frac{\epsilon}{C(\epsilon)}}, |\zeta_0(d) - \theta(d)|, \nu_0(d), e_t(d)\right)\right) \leq 3e^{-\epsilon}.
\]

Moreover, the concentration rate is non-increasing on \( \epsilon \).

---

\(^{17}\)The on \( t \) restrictions stem from Lemmas 18 and 19; a detailed explanation can be found in their proofs. However, it is clear that there always exists a \( t \) large enough that all the restrictions are satisfied.
The intuition of the proof is as follows. As discussed above, the randomness of $\zeta_{i}^\sigma(d)$ comes from $J_i(d)$ and $f_i(d)$. By using concentration bound for mixing processes (e.g. Merlevède et al. (2009)), Lemmas 18 and 19 in Appendix D show that for any $\varepsilon \geq 0$, $J_i(d)$ and $f_i(d)$ are within $\sqrt{\frac{\log t}{t}} \sqrt{\frac{\varepsilon}{C(\varepsilon)}}$ of their population analogues with probability higher than $1 - 3e^{-\varepsilon}$. These concentration rates, however, get distorted by $\Gamma$ because the posterior mean is a non-linear transformation of $C(3)$ and $5C(3)$. By inspection, it is easy to see that $\Gamma = \sqrt{\frac{\log t}{t}} + \frac{1}{t}$, so the concentration rate is of order $\sqrt{\log t/t}$ where the additional factor of “$\log t$” arises from the lack of IID-ness in the data. Moreover, it is non-increasing on $\varepsilon$, thereby illustrating the fact that higher levels of experimentation — represented by a higher $\varepsilon$ — yield a faster order for the concentration rate.

Our method for aggregating multiple priors offers an attractive feature with regards to our concentration rates. Sufficiently stubborn models, i.e. $|\zeta_0^\sigma(d) - \theta(d)|\sqrt{v_0^\sigma(d)}$ is sufficiently large, will have close to zero effect on the concentration rate of $\zeta_t^\sigma(d)$, 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 1.** Take any $(t, d, \varepsilon)$ as in Proposition 3 and suppose all its assumptions hold. Furthermore, let model $o = 0$ denote the least stubborn model and suppose that for any given $\delta > 0$, there exists a $C$ such that $v_0^\sigma(d)|\zeta_0^\sigma(d) - \theta(d)| \geq C$ for all $o \neq 0$. Then,

$$P_\pi \left( |\zeta_t^\sigma(d) - \theta(d)| > \Omega \left( \sqrt{\frac{\log t}{t}} \sqrt{\frac{\varepsilon}{C(\varepsilon)}}, |\zeta_0^\sigma(d) - \theta(d)|, \frac{v_0^\sigma(d)}{t}, e_t(d) \right) + \delta \right) \leq 3e^{-\varepsilon}$$

**Proof.** See Appendix D.3. □

The function $\Omega$, which is formally defined in Appendix D.3, acts as $\Gamma$ but for one model; i.e., for any $o \in \{0, \ldots, L\}$ and any $\gamma \geq 0$, assuming $J_i(d)$ and $f_i(d)$ are within $\gamma$ of their population analogues,

$$|\zeta_0^\sigma(d) - \theta(d)| \leq \Omega(\gamma, |\zeta_0^\sigma(d) - \theta(d)|, \frac{v_0^\sigma(d)}{t}, e_t(d)).$$

Thus, $\Omega$, quantifies the effects on concentration rates of the model’s the priors and the expected frequency of play. We summarize its implications for the rate in the following remark and illustrate them numerically in Section 4.
**Remark 2** (Properties of the Concentration Rate).

1. All else equal, the concentration rate decreases as the bias increases; it also decreases with the degree of stubbornness, i.e. \( |\xi_0^\alpha(d) - \theta(d)| \sqrt{v_0^\alpha(d)} \). The concentrate rate is fastest when the bias is zero.

2. For confident models, the concentration rate increases with the degree of conviction, i.e. \( v_0^\alpha(d) \) increases. The intuition behind this result is as follows: If \( v_0^\alpha(d) \) increases but \( |\xi_0^\alpha(d) - \theta(d)| \sqrt{v_0^\alpha(d)} \) 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 decreases as \( C \) increases.

4. An increase of the frequency of play, \( e_t(d) \), improves the concentration rate. This comes from the fact that \( e_t(d) \leftrightarrow \Omega \left( \sqrt{\frac{\log t}{t}} \sqrt{\frac{\epsilon}{C(c)}}, |\xi_0^\alpha(d) - \theta(d)| / t, e_t(d) \right) \) is decreasing. Intuitively, increasing \( e_t(d) \) implies having more observations to estimate \( \theta(d) \) — “more information” about treatment \( d \) implies a faster concentration rate.

5. As \( C \) is non-decreasing and \( \Omega \) is increasing in the first argument, the concentration rate becomes faster with \( \epsilon \). Loosely speaking, from Proposition 2 it follows that the dependence between current and past realizations of \( (Y_t(D_t), D_t) \) decreases as \( \epsilon \) increases. Thus, the higher the \( \epsilon \), the more informative each realization becomes, thereby implying a higher concentration rate. It is, however, important to highlight that a change in the \( \epsilon \) will also affect \( e_t(d) \). Thus, in practice, the total effect on the concentration rate can be ambiguous.

\[ \triangle \]

### 3.2.2 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 arm 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

\[
\max_{d \neq M} \xi^\alpha_T(d) - \xi^\alpha_T(M) > 0,
\]
where \( \tau \) indicates when the experiment is stopped, i.e., is the first instance after \( B \) such that 
\[
\max_d \min_{m \neq d} \{ \zeta_t^a(d) - \zeta_t^a(m) - c_t(\gamma_t, d, m) \} > 0
\]
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.

**Proposition 4.** Consider the stopping rule defined in Example 4 with parameters \( ((\gamma_t), B) \) such that \( \log B \geq \max\{2, 4 \epsilon B(\epsilon)\} \). Then,
\[
P_\pi \left( \max_{d \neq M} \{ \zeta_\tau^a(d) - \zeta_\tau^a(M) \} > 0 \right) \leq \sum_{d=0}^{M-1} \sum_{t=0}^{T} \left( 2e^{-0.5t \gamma_t^2 \nu \sigma(d)^2} + e^{-\frac{t}{\log^2 \eta_d(t, \epsilon, \Delta)} C(\epsilon)} \right)
\]

where \( \eta_d^*(t, \epsilon, \Delta) \in \mathbb{R}_+ \cup \{+\infty\} \) is defined in Appendix E and is non-decreasing in \( t, \epsilon, \) and \( \Delta \). If \( \zeta_0(d) \leq \theta(d) \) and \( \zeta_0(M) \geq \theta(M) \), then \( \eta_d^*(t, \epsilon, \Delta) = +\infty \).

*Proof.* See Appendix E.

This proposition shows that the quantity \( \eta_d^*(t, \epsilon, \Delta) \) is key for understanding how the primitives of our setup – i.e. \( \epsilon \) and \( \Delta \), different priors, etc. – affect the upper bound for the probability of a mistake. As we prove in Appendix E, the upper bound for the probability of a mistake decays exponentially with \( t \) and is non-increasing in \( \epsilon \) and \( \Delta \). Intuitively, as \( \epsilon \) increases, the data becomes less dependent on the past and thus more informative, resulting in a tighter bound. As \( \Delta \) becomes more positive, so does the difference between the PM’s posteriors, which also decreases the probability of making a mistake.

This proposition also allows us to examine whether the upper bound embodies an oracle property similar to the one we demonstrated for the concentration rates. The key to assessing this again lies in understanding the behavior of \( \eta^* \). Given the properties of the weights illustrated in Proposition 1 and Lemma 5, it is easy to show that if the other sources are sufficiently different to the oracle source, then \( \eta_d^* \) becomes arbitrary close to \( \eta_d^{oracle} \), where \( \eta_d^{oracle} \) is defined as the largest \( \eta \) such that 
\[
\frac{|\zeta_0(d) - \theta(d)| \nu \sigma(d)/t}{\epsilon - \eta \nu \sigma(d)/t} \leq 0.5 \Delta \quad \text{for each} \quad d \in \mathcal{D}.
\]
It then follows that the bound obtained in Proposition 4 would be arbitrary close to the oracle one; the corollary below formalizes this discussion.

**Corollary 2.** Suppose all the conditions of Proposition 4 hold and \( B \) is such that 
\[
\frac{(1)^{-dM} (\zeta_0^0(d) - \theta(d)) \nu_0^0(d)/B}{\epsilon + \nu_0^0(d)/B} \leq
\]

\footnote{It should be understood that if the RHS is greater than one, then bound will be taken to be one.}
Then, for any \( \varepsilon > 0 \), there exists a \( M \) such that for all \( |\zeta_0^0(d) - \theta(d)| \geq M \), it follows that

\[
P_\pi \left( \max_{d \neq M} \left\{ \frac{\zeta^\alpha_\tau(d) - \zeta^\alpha_\tau(M)}{\zeta^\alpha_\tau(d) - \zeta^\alpha_\tau(M)} > 0 \right\} \right) \leq (1 + \varepsilon) \sum_{d=3}^{M} \sum_{t=0}^{T} \left( 2e^{-0.5t - (\gamma_0^0/d)^2} + e^{-\frac{t}{\sqrt{t}} (\eta^\alpha(t, \varepsilon, \Delta))^2 C(\varepsilon)} \right),
\]

where \( \eta^\alpha(t, \varepsilon, \Delta) \) is defined as

\[
\max\{\eta: \frac{|\zeta_0^0(d) - \theta(d)|/t}{\varepsilon - \eta + \nu_0^0(d)/t} \leq 0.50\Delta \text{ and } \eta \leq \varepsilon\}
\]

Proof. See Appendix E. \( \square \)

The previous proposition also shows how by choosing \((\gamma_t)_t, B)\) with some care, the probability of mistake associated with the stopping rule is bounded by \( \beta \), where \( \beta \in (0, 1) \) is any tolerance level. The next corollary presents such result.

**Corollary 3.** Suppose all the conditions of Proposition 4 hold, and, for any \( t, \gamma_t \geq \sqrt{\log t} \sqrt{A \log t} \) with \((A, B)\) such that \( \log B \geq \max_d 2\nu(d)^2 \), \( \min_d (\eta^\alpha(1, \varepsilon, \Delta))^2 C(\varepsilon) \geq (\log B)^2 \) and

\[
\frac{3(M + 1)}{A - 1}(B^{-(A-1)} - T^{-(A-1)}) \leq \beta.
\]

Then

\[
P_\pi \left( \max_{d \neq M} \left\{ \frac{\zeta^\alpha_\tau(d) - \zeta^\alpha_\tau(M)}{\zeta^\alpha_\tau(d) - \zeta^\alpha_\tau(M)} > 0 \right\} \right) \leq \beta.
\]

Proof. See Appendix E. \( \square \)

The choice of \((\gamma_t)_t\) and the extra restrictions in \( B \) are to ensure that both terms in the upper bound in Proposition 4 are less than \( t^{-A} \). By simple arguments it can be shown that \( \sum_{t=0}^{T} t^{-A} \leq \frac{1}{A-1}(B^{-(A-1)} - T^{-(A-1)}) \) and so expression 7 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 stems from two sources: First, our data are not IID and thus the concentration rate of the relevant quantities — the frequency \((f_t)_t\) in particular — are of order \( \sqrt{\log t} \) as opposed to \( \sqrt{1/t} \). Second, the extra \( \sqrt{\log t} \) factor acts as an upper bound for population quantities we do not know. 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 $\sqrt{\log t}$ factor.

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

3.2.3 Average Observed Outcomes

In this section, we characterize the behavior of the average outcome $t^{-1} \sum_{s=1}^{t} Y_s$. It is easy to show that $t^{-1} \sum_{s=1}^{t} Y_s$ will concentrate around a weighted average of $\theta(\cdot)$, with the expected frequency of play as weights.\(^{19}\)

\[
r^{-1} \sum_{s=1}^{t} \sum_{d=0}^{M} \theta(d) E_{\pi} [\delta(Z_s)(d)] = \sum_{d=0}^{M} \theta(d) e_s(d).
\]

But what is difficult to show is how close this quantity is to the maximum expected outcome, $\max_d \theta(d)$, when working with arbitrary policy functions. However, for the $\epsilon$-greedy policy function defined by Equation 5, we can establish the following proposition:

**Proposition 5.** Let $\max_d \theta(d)$ to be equal to $\theta(M)$ and suppose $\delta(\cdot)$ is the $\epsilon$-greedy policy function defined in Equation 5. Then, for any $\gamma > 0$ and any $t$ sufficiently large,\(^{20}\)

\[
\mathbb{P}_{\pi}\left[\left| t^{-1} \sum_{s=1}^{t} Y_s - \max \theta(d) \right| > \Sigma_1(\gamma, t, \epsilon) + ||\theta||_{t^{-1}} \sum_{s=1}^{t} \Lambda_s(\Delta, \zeta_0 - \theta, v_0, \epsilon) + \text{Bias}(\epsilon) \right] \leq \gamma
\]

where

\[
\text{Bias}(\epsilon) := \epsilon \left( M + 1 \right) \left( \max_d \theta(d) - \frac{\sum_{d=0}^{M} \theta(d)}{M + 1} \right),
\]

\[
\Sigma_1(\gamma, t, \epsilon) := \epsilon \left( M + 1 \right) \max_d \left( \min \left\{ \frac{\log t}{8\nu(\sigma(d))}, \left( \frac{0.5}{\theta(d)} \right)^2 \mathbb{C}(\epsilon) \right\} \right)^{-1/2} \sqrt{\frac{\log t}{t}} \sqrt{\frac{\log 3(M + 1)/\gamma}{\gamma}},
\]

\(^{19}\)The formal argument relies on noting that $t^{-1} \sum_{s=1}^{t} Y_s - t^{-1} \sum_{s=1}^{t} \sum_{d=0}^{M} \theta(d) 1\{D_s = d\}$ is a MDS and $t^{-1} \sum_{s=1}^{t} \sum_{d=0}^{M} \theta(d) 1\{D_s = d\} - \sum_{d=0}^{M} \theta(d)e_s(d)$ converges to zero at the concentration rate given by Lemma 19.

\(^{20}\)"Sufficiently large" is specified in the proof of the proposition.
and $\Sigma_1$ is decreasing in $\gamma$, decreasing in $t$ and non-increasing in $\epsilon$. Also,

$$
\Lambda_s(\Delta, \zeta_0 - \theta, \nu_0/s, \epsilon) := 3 \sum_{d=0}^{M-1} \exp \left\{ -\frac{s}{\log s} \left( \max \{ \Gamma^{-1}(0.5\Delta, \zeta_0(d) - \theta(d), \nu_0(d), \epsilon_s(d)) , 0 \} \right)^2 C(\epsilon) \right\}.
$$

and

$$
\Sigma_1(\gamma, t, \epsilon) := (M + 1) \max_d \left( \min \left\{ \frac{\log t}{8\nu(\sigma(d))^2} , \left( \frac{0.5}{\theta(d)} \right)^2 C(\epsilon) \right\} \right)^{-1/2} \sqrt{\frac{\log t}{t}} \sqrt{\log 3(M + 1)/\gamma},
$$

and $\Sigma_1$ is decreasing in $\gamma$, decreasing in $t$ and non-increasing in $\epsilon$.

Proof. See Appendix F. \qed

Despite the length of the proposition, its parts are quite intuitive. The term $\Sigma_1(\gamma, t, \epsilon)$ controls the stochastic error that arises from the difference between $t^{-1} \sum_{s=1}^{t} Y_s = \sum_{d=0}^{M-1} t^{-1} \sum_{s=1}^{t} Y_s(d) 1\{D_s = d\}$ and its expectation $\sum_{d=0}^{M-1} t^{-1} \sum_{s=1}^{t} Y_s(d) E_\pi[\delta(Z_s)(d)]$. This term is essentially of order $O(\sqrt{\log t/t})$, where the “Oh” depends on $\epsilon$ in a non-increasing way. If $\epsilon$ increases, all else things equal, the data becomes “more independent” and we converge faster. The term $||\theta|| t^{-1} \sum_{s=1}^{t} \Lambda_s(\Delta, \zeta_0, \nu_0, \epsilon)$ arises from choosing the wrong treatment, in expectation, because the policy function depends on $\zeta$ and not $\theta$; this expression is similar to the one we obtained in Proposition 6 in Appendix E. Finally, the term $Bias(\epsilon)$ is a non-random bias that stems from the “exploration” part of the $\epsilon$-greedy policy function. With probability $\epsilon(M + 1)$ the treatment is chosen at random, producing $\sum_{d=0}^{M-1} \theta(d)/(M + 1)$.

The term $||\theta|| t^{-1} \sum_{s=1}^{t} \Lambda_s(\Delta, \zeta_0, \nu_0, \epsilon)$ has a faster rate of convergence than $\Sigma_1(\gamma, t, \epsilon)$, thus the leading terms are $\Sigma_1(\gamma, t, \epsilon) + Bias(\epsilon)$. These leading terms nicely illustrate the so-called “exploration vs. exploitation” tradeoff and how it is regulated by $\epsilon$. By reducing $\epsilon$, we reduce the effect of “exploration” by lowering the $Bias(\epsilon)$, but at the cost of (weakly) increasing $\Sigma_1(\gamma, t, \epsilon)$.

This tradeoff suggests a choice for $\epsilon$ that balances $Bias(\epsilon)$ and $\Sigma_1(\gamma, t, \epsilon)$. Unfortunately, such a choice is infeasible as both terms depend on unknown quantities and $C(.)$ does not have an explicit expression. Nevertheless, we can conclude that $\epsilon = 1$ — the choice used in RCTs — will typically not be optimal. In fact, as $t$ increases, the “optimal” $\epsilon$ will decrease to 0, favoring “exploitation” to “exploration”. We explore the choice of $\epsilon$ further, when we simulate our model.
4 Model Simulations

In this section, we present Monte Carlo simulations of our model. 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 periods.

Multiple Priors, External Validity, Robustness

We begin by illustrating how our setup weights the different models over 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, $\alpha_i^o(d)$ – the posterior probability that model $o$ best fits the observed data within the class of models being considered. We demonstrated in Proposition 1 for the general case, and Lemma 1 for Gaussianity, that if there exists an externally valid model among externally invalid models, then $\alpha_i^o(d)$ will approach one for the externally valid model. Conversely, $\alpha_i^o(d)$ 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. $\xi_0 = \theta$) but diffuse (i.e. $\nu=1$). For the other set of priors, we consider four alternative scenarios varying in their degree of stubbornness.

In Figure 1, we plot $\alpha_i^o(d)$ 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 $\alpha_i^o(d)$ increases over time putting more and more weight on an externally valid model. By the end of the experiment, $\alpha_i^o(d)$ is close to 95% for both arms. As we consider more stubborn models, we can see that the corresponding $\alpha_i^o(d)$ becomes smaller. So much so that for extremely stubborn models (i.e. the lowest line) $\alpha_i^o(d)$ becomes essentially zero by the 600th instance. This is why we interpret the parameter $\alpha_i^o(d)$ as a measure of external validity: the more externally valid the model, the higher the corresponding $\alpha_i^o(d)$.

An important feature of how we aggregate across models is that it generates a robustness property.
Because $\alpha^o(d)$ 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 unbiased (i.e. $\zeta^o_0 = \theta$), but diffuse. The top (dash-dotted) line refers to an alternative set of priors, which contains some degree of stubbornness (i.e. $\zeta^o_0 = \theta + 0.5, \nu = 250$). The middle (solid) line comes from the combined model, which is a weighted average of the two sets of priors using $\alpha^o(d)$ 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. 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 Remark 2, our parameter $\alpha$ is more responsive to bias than conviction.

**Concentration Bounds**

**Effects of $\epsilon$.** We now simulate our model’s concentration bounds and some its key properties. Recall from Remark 2 in Section 3.2.1, the concentration rate increases with the parameter $\epsilon$. 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 $\epsilon$, we compute the difference over time between the policymaker’s posterior belief of the true mean, $\zeta^o_0(d)$, 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. $\zeta^o_0 = \theta$ and $\nu^o_0 = [1, 1]$).

In the top panel, we see that with the exception of early on, our concentration bounds decrease over time and in the case of $\zeta^o_0(0)$ decrease faster, the higher the $\epsilon$. For instance, after 1000 instances, $Pr(\zeta^o_0(0) - \theta(0) > 0.1)$ 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 $\epsilon$ lines decreasing faster.

The intuition for these patterns is straightforward and speaks to the point about frequency of play in Remark 2. 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 $\epsilon$-greedy algorithm. In this case, the higher the $\epsilon$ 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 $\epsilon$’s. As we compare the two panels, the more we play a particular arm, the more we learn about it, and 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 2, 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 a less stubborn the model. 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 3.2.2, 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 $\epsilon$. It is clear from the graph that the more we experiment across treatment arms (i.e., higher $\epsilon$), 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. While at low levels of $\epsilon$, the experiment lasts for almost its entire duration, the drop off appears fairly quick. Once $\epsilon$ 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 $\epsilon$ 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 \( \zeta^0_0 = \theta \), whereas the other is stubborn, with \( \zeta^\alpha_0 = [\theta(0) + \delta, \theta(1) - \delta] \), where \( \delta \) 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 significant 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 \( \alpha_i^\alpha(d) \).

**Expected Earnings**

The final outcome we evaluate is expected earnings. According to Proposition 5, the distance between the average outcomes and maximum expected outcome is decreasing in \( \epsilon \). In Figure 7, we plot by \( \epsilon \), 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 \( \epsilon \). 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, by 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 \( \epsilon \) close to zero. Because as we saw in Figure 5, lower \( \epsilon \)’s result in longer experiments, which can come with costs. Moreover, as we show in Proposition 4, the upper bound the probability of making a mistake is weakly smaller for higher levels of \( \epsilon \). 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:

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

(8)

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

(9)

where $c_1$ indicates the costs of running the experiment, $c_2$ cost of administering the treatment, $\beta^t$ 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 $\epsilon$. In contrast with the previous figure, we see that the average payoffs are increasing with $\epsilon$ until approximately $\epsilon=0.38$, at which point the payoffs start to decline. While this “optimal” value of $\epsilon$ is clearly a function of an arbitrary set of parameter choices, our conjecture is that the inverted u-shape relationship is likely to hold more generally, suggesting that some combination of experimentation and exploitation is optimal.

5 Conclusions

This paper presents a framework for how to incorporate prior sources of information into the design of a sequential experiment. We show that our setup offers several nice properties, including a robustness to “incorrect” priors. We also propose a formal definition of external validity that in the context of our setup allows us to differentiate across models in terms of their degree of external invalidity.

Even though we motivated our framework as a way for policymakers or researchers to incorporate prior evidence into their design of an adaptive experiment, we believe our framework is quite general and thus applicable to other types of diverse problems, ranging from online marketing campaigns to the targeting government programs.
References


6 Figures

Notes: This figure plots $\alpha_i^0(d = 0, x)$ (left plot) and $\alpha_i^0(d = 1, x)$ (right plot) under two alternative sets of priors. For the confident model, the initial priors are: $\zeta_0^0 = \zeta_1^0 = \theta; \nu_0^1 = [1, 1]; \nu_0^0 = [250, 250]$. For the stubborn model, the initial priors are: $\zeta_0^0 = \theta; \zeta_1^0 = \theta + 0.3; \nu_0^0 = [1, 1]; \nu_0^1 = [250, 250]$. These figures are based on 1,000 simulations using the following parameters: $\theta = [1, 1.3], \varepsilon = 0.5$. 
Figure 2: Posterior Beliefs Over Time, Holding Behavior Constant

Notes: This figure plots the policymakers posterior beliefs (i.e., $[\zeta^0_0(0,x), \zeta^1_0(1,x)]$) 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: $\zeta^0_0 = \theta; \zeta^1_0 = \theta + 0.3; \pi^0_0 = [1,1]; v^1_0 = [250,250]$. For the confident model, the initial priors are: $\zeta^0_0 = \zeta^1_0 = \theta; \pi^0_0 = [1,1]; v^1_0 = [250,250]$. 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 $\epsilon$. The bottom panel plots the number of times the experimental arm was played at time $t$ for different values of $\epsilon$. The graphs on the left correspond to 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: $\theta = [1, 1.3]; \ell^0 = \theta; \ell^1 = \theta; v^0 = [1, 1]; v^1 = [1, 1].$
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 $\epsilon$. These figures are based on 1,000 simulations using the following parameters: $\theta = [1, 1.3]; \zeta_0 = 0; \zeta_1 = 0; \nu_0 = [1, 1]; \nu_1 = [1, 1]; B = 100.$
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: $\theta = [1,1.3]$; $v_0^0 = v_1^0 = [250,250]$; $\epsilon_0^0 = [\theta(0) + bias, \theta(1) - bias]$ ; $\xi_0^1 = \theta$, $\epsilon = 0.5$. 
Figure 7: Relative Average Earnings During the Experiment

Notes: This figure plots by $\epsilon$, the average earnings net of maximal earnings. These figures are based on 1,000 simulations using the following parameters: $\theta = [1, 1.3]; \xi_0^\prime = \theta; \xi_1^\prime = \theta; \nu_0^\prime = [1, 1]; \nu_1^\prime = [1, 1].$
Figure 8: Experimentation versus Exploitation – Expected Payoffs

Notes: This figure plots by $\epsilon$, the expected payoffs as defined by Equation 8. These figures are based on 1,000 simulations using the following parameters: $\theta = [1, 1.3]; \zeta_0^\nu = \theta; \zeta_1^\nu = \theta; \nu_0^\nu = [1, 1]; \nu_1^\nu = [1, 1]; B = 100; \beta = 0.994; c = 1.15; \lambda = 1, 100.
A 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 $G \in X$, the PM has a family of PDFs indexed by a finite dimensional parameter $e^{\theta} \in \Theta$, $\mathcal{P}_G := \{p_\theta \in \Theta \} \subseteq \Delta(\mathbb{R}^{M+1})$, that describes what she believes are plausible descriptions of the true joint probability of the potential outcome $(Y(d,x))_{d \in D}$. For each $p_\theta \in \mathcal{P}_x$, we use $p_{\theta,d}$ to denote the marginal PDF of $p_\theta$ 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 $(\mu_0^\alpha(x))_{\alpha=0}^L$ to denote such prior beliefs. By convention, we use $\alpha=0$ to denote the PM’s own prior and leave $\alpha > 0$ to denote the other sources.

For each $x \in X$, 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

$$\mu_\alpha^\alpha(x)(A) = \frac{\int_T p_{\theta,d}(Y_t(x))\mu_\alpha^{\alpha-1}(x)(d\theta)}{\int_\Theta p_{\theta,d}(Y_t(x))\mu_\alpha^{\alpha-1}(x)(d\theta)}$$

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, $\{\mathcal{P}_x,\mu_\alpha^\alpha(x)\}_{\alpha=0}^L$, 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 $\alpha$ best fits the observed data within the class of models being considered,
i.e.,

\[ \tilde{\mu}_t(x)(A) := \sum_{o=0}^{L} \alpha_t^o(x) \mu_t^o(x)(A) \]

for any Borel set \( A \subseteq \Theta \), where

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

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

One example of \( \mathcal{P}_t \) that is of particular interest is one where \( \Theta = \prod_{d \in \mathbb{D}} \Theta \) and, for each \( d \in \mathbb{D} \), \( p_{\theta,d} = p_{\theta,d,d} \) (i.e., it only depends on the \( d \)-th coordinate of \( \theta \); henceforth, we omit "d" from the \( \theta_d \)); and also, for each \( o \in \{0, \ldots, L\} \), \( \mu_t^o(x) = \prod_{d \in \mathbb{D}} \mu_t^o(d,x) \). 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

\[
\int f(\theta) \mu_t^o(x)(d\theta) = \int f(\theta_0, \ldots, \theta_M) p_{\theta,d}(Y_1(x)) \mu_t^o(d,x)(d\theta) \frac{\prod_{d' \neq d} \mu_t^o(d',x)(d\theta)}{\int \theta_0 p_{\theta,d}(Y_1(x)) \mu_t^o(d,x)(d\theta)}
\]

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

\[ \mu_t^o(d',x)(A) = \mu_t^o(d',x)(A), \]

so the posterior is not updated. On the other hand, the posterior for \( \theta \) is given by

\[
\mu_t^\theta(d,x)(A) = \int_A \frac{p_{\theta,d}(Y_1(x)) \mu_t^o(d,x)(d\theta)}{\int \theta_0 p_{\theta,d}(Y_1(x)) \mu_t^o(d,x)(d\theta)}. 
\]

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 \( \mu_t^\theta(x) = \prod_{d \in \mathbb{D}} \mu_t^\theta(d,x) \).
A more succinct notation that captures these nuances is given by

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

for any \(d \in D\) and any \(A \subseteq \Theta\) Borel. Applying this recursively, it follows that

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

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 \(p_{\theta}\) — it is easy to see that the previous recursion describes the Bayesian updated presented in the paper.

## B Properties of \(\alpha\)

### B.1 Proof of Proposition 1

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

**Lemma 2.** Suppose \(\Theta \subseteq \mathbb{R}^{|\Theta|}\) is compact and there exists a \(\varphi : \mathbb{R}_+ \to \mathbb{R}_+\) such that \(\varphi(0) = 0\) and increasing and

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

Then for any \(t\) and \(\gamma > 0\),

\[
P\left(\sup_{\theta} |\sum_{s=1}^t g(Y_s, D_s, \theta)| \geq \sqrt{1/(2C\gamma)} \sqrt{\Lambda(t, |\Theta|)}\right) \leq \gamma
\]

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

The second lemma provides a non-asymptotic bound for the ratio of the weights for any two models.
In particular, it relates the weights, $\alpha_t^\circ$, with the Laplace transformation of

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

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

**Lemma 3.** Take any $o, o' \in \{0, \ldots, L\}$ and $(d, x) \in \mathbb{D} \times \mathbb{K}$. Suppose $\Theta \subseteq \mathbb{R}^{[\Theta]}$ is compact and there exists a $\varphi : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ such that $\varphi(0) = 0$ and increasing and

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

Then, for any $\gamma > 0$,

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

with probability larger than $1 - \gamma$, where $r_t := O(\sqrt{1/\gamma \Lambda(t, |\Theta|)})$.

**Proof of Proposition 1.** Suppose the conditions in Lemmas 2 and 3 hold (we show they do towards the end of the proof). In this case, it follows that for any $\varepsilon > 0$,

$$P_\pi \left( \frac{\alpha_t^o(d,x)}{\alpha_t^{o'}(d,x)} \geq \varepsilon \right) \leq P_\pi \left( \frac{\alpha_t^o(d,x)}{\alpha_t^{o'}(d,x)} \geq \varepsilon \cap S_t(\varepsilon) \right) + P_\pi \left( S_t(\varepsilon)^C \right)$$

where $S_t(\varepsilon)$ is the set such that $N_t(d,x)/t \geq e_t(d,x) - \varepsilon$ and $r_t^{-1} \sup_{\theta \in \Theta} \left| t^{-1} \sum_{s=1}^t g(Y_s, D_s, \theta) \right| \leq \varepsilon$. By Lemma 2 and Lemma 18, $\lim_{t \to \infty} P_\pi (S_t(\varepsilon)^C) = 0$.

In what follows, take an arbitrary sequence in $S_t(\varepsilon)$. Let $u_\circ(d,x)$ as in the definition. If $u_\circ(d,x)$ --
\[ u_{\alpha'}(d,x) =: A > 0, \text{ then by Lemma 3,} \]
\[
\frac{\alpha_{\ell'}^0(d,x)}{\alpha_{\ell'}(d,x)} \leq e^{2r_t} \frac{N_t(d,x) \int_0^\infty \mu_0^\ell(d,x) (KL_{d,x}(\theta) \leq \nu) e^{-N_t(d,x)\nu} d\nu}{N_t(d,x) \int_0^\infty \mu_0^\ell(d,x) (KL_{d,x}(\theta) \leq \nu) e^{-N_t(d,x)\nu} d\nu} \\
\leq e^{2r_t} \frac{N_t(d,x) \int_0^\infty \mu_0^\ell(d,x) (KL_{d,x}(\theta) \leq \nu) e^{-N_t(d,x)\nu} d\nu}{e^{-N_t(d,x)(m_\ell' + 0.5A)} \int_{m_\ell'}^\infty \mu_0^\ell(d,x) (KL_{d,x}(\theta) \leq \nu) d\nu}
\]
\[
= e^{-N_t(d,x)(0.5A + 2r_t)} \frac{\int_{m_\ell'}^\infty \mu_0^\ell(d,x) (KL_{d,x}(\theta) \leq \nu) d\nu}{\int_{m_\ell'}^\infty \mu_0^\ell(d,x) (KL_{d,x}(\theta) \leq \nu) d\nu}
\]
where the last line follows from definition of \( A \).

The fraction in the previous display is a fixed number. Under the set \( S_t(\varepsilon), \frac{N_t(d,x)}{t} 0.5A + 2r_t \geq (\varepsilon (d,x) - \varepsilon)0.5A - 2\varepsilon \). Under Assumption 2, 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 \( \frac{\alpha_{\ell'}^0(d,x)}{\alpha_{\ell'}(d,x)} = O(e^{-tc}), \) which in turn implies that
\[
P_\pi \left( \frac{\alpha_{\ell'}^0(d,x)}{\alpha_{\ell'}(d,x)} \geq \varepsilon \right) \leq P_\pi \left( e^{-tc} \geq \varepsilon \right) + P_\pi \left( S_t(\varepsilon) \right) = o(1),
\]
for any \( \varepsilon \leq \bar{\varepsilon} \). Thus the desired result follows.

We now verify that the conditions in Lemma 3 (and thus, Lemma 2) hold. \( \Theta \) 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.-\( P_\pi \). This implies that \( \theta \mapsto \ell(\cdot, \theta) \) is continuous a.s.-\( P_\pi \). Continuous functions over compact sets are uniformly continuous, thus \( f(Y(d,x),\delta) := \sup_{||\theta-\theta'|| \leq \delta} |\ell(Y(d,x),\theta) - \ell(Y(d,x),\theta')| \) converges to 0 as \( \delta \) vanishes a.s.-\( P_\pi \). 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.

\[ \square \]

B.2 \hspace{1em} Proof of Lemma 1

Proof of Lemma 1. Let \( p_\theta \) denote a Gaussian PDF with mean \( \theta \) and variance 1.
(1) It follows that
\[
\int \prod_{s=1}^{t} (p_{\theta}(Y_s))^{1\{D_s(x)=d\}} \mu_0^\alpha(d,x)(d\theta)
\]
\[
= \int (2\pi)^{-0.5} \Sigma_{s=1}^{t} 1\{D_s(x)=d\} \exp \left\{ - \frac{1}{2} \sum_{s=1}^{t} 1\{D_s(x)=d\} (Y_s - \theta)^2 \right\} \phi(\theta; \xi_0^\alpha(d,x), 1/\nu_0^\alpha(d,x)) d\theta
\]
\[
= \int (2\pi)^{-0.5} \Sigma_{s=1}^{t} 1\{D_s(x)=d\} \exp \left\{ - \frac{1}{2} \sum_{s=1}^{t} 1\{D_s(x)=d\} (Y_s(d,x) - m_\tau(d,x))^2 \right\}
\]
\[
\times \exp \left\{ - \frac{1}{2} \sum_{s=1}^{t} 1\{D_s(x)=d\} (m_\tau(d,x) - \theta)^2 \right\}
\]
\[
\times \exp \left\{ - \sum_{s=1}^{t} 1\{D_s(x)=d\} (Y_s(d,x) - m_\tau(d,x)) (m_\tau(d,x) - \theta) \right\}
\]
\[
\phi(\theta; \xi_0^\alpha(d,x), 1/\nu_0^\alpha(d,x)) d\theta,
\]
where \(m_\tau(d,x) := \sum_{s=1}^{t} 1\{D_s(x)=d\} Y_s(d,x) / \sum_{s=1}^{t} 1\{D_s(x)=d\} \). Observe that
\[
\sum_{s=1}^{t} 1\{D_s(x)=d\} (Y_s(d,x) - m_\tau(d,x)) = 0,
\]
so, letting \(N_\tau(d,x) := \sum_{s=1}^{t} 1\{D_s(x)=d\} \) it follows that
\[
\int \prod_{s=1}^{t} (p_{\theta}(Y_s))^{1\{D_s(x)=d\}} \mu_0^\alpha(d,x)(d\theta) = (2\pi)^{-0.5} \Sigma_{s=1}^{t} 1\{D_s(x)=d\} + 0.5 N_\tau(d,x)^{-1/2}
\]
\[
\times \exp \left\{ - \frac{1}{2} \sum_{s=1}^{t} 1\{D_s(x)=d\} (Y_s(d,x) - m_\tau(d,x))^2 \right\}
\]
\[
\times \int (2\pi/N_\tau(d,x))^{-1/2} \exp \left\{ - \frac{1}{2} (m_\tau(d,x) - \theta)^2 N_\tau(d,x) \right\}
\]
\[
\phi(\theta; \xi_0^\alpha(d,x), 1/\nu_0^\alpha(d,x)) d\theta.
\]
The expression of the integral can be viewed as a convolution between to Gaussian PDFs one indexed by \((0, 1/N_\tau(d,x))\) and \((\xi_0^\alpha(d,x), 1/\nu_0^\alpha(d,x))\) resp, which in turn is equivalent to PDF of
the sum of the corresponding random variables evaluated at \( m_t(d,x) \). Therefore,

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

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

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

(2) Follows from the definition of \( \phi \).

(3) By Lemmas 18 and 20 and Assumption 2, \( N_t(d,x) \) diverges with probability approaching 1 and \( m_t(d,x) = \theta(d,x) + o_P(1) \). By continuity of \( \phi \), the result follows.

\[ \square \]

**B.3 Bounds for \( \alpha_t \)**

It follows that, for any \( (d,x) \in \mathbb{D} \times \mathbb{X} \) and \( t \), \( \alpha_t^o(d,x) = \frac{\exp \ell_t^o(d,x)}{\sum_{o=0}^{L} \exp \ell_t^o(d,x)} \) where

\[
\ell_t^o(d,x) := \log \phi(m_t(d,x); \zeta_0^o(d,x), (N_t(d,x) + v_0^o(d,x))/(N_t(d,x)v_0^o(d,x)))
\]

We now provide non-random bounds for \( \ell_t^o \) and study some properties; all proofs are relegated to the end of the section. Finally, recall that \( f_t(d,x) := N_t(d,x)/t \).

**Lemma 4.** For any \( o \in \{0, ..., L\} \), any \( (d,x) \in \mathbb{D} \times \mathbb{X} \), any \( \pi_t(d,x) \in (0,1) \) and \( \eta \in (0,\epsilon) \), \( \delta > 0 \) such that \( |f_t(d,x) - \pi_t(d,x)| \leq \eta \) and \( |m_t(d,x) - \theta(d,x)| \leq \delta \), it follows that

\[
\ell(\eta, \delta, |\zeta_0^o(d,x)|, v_0^o(d,x), \pi_t(d,x)) \leq \ell_t^o(d,x) \leq \bar{\ell}(\eta, \delta, |\zeta_0^o(d,x)|, v_0^o(d,x), \pi_t(d,x))
\]

and thus

\[
\alpha^o(\eta, \delta, |\zeta_0(d,x)|, v_0(d,x), \pi_t(d,x)) \leq \alpha_t^o(d,x) \leq \bar{\alpha}^o(\eta, \delta, |\zeta_0(d,x)|, v_0(d,x), \pi_t(d,x))
\]
where
\[
\alpha^o(\eta, \delta, |\tilde{z}_0(d,x)|, \nu_0(d,x), \pi_t(d,x)) := \frac{e^{\ell(\eta, \delta, |\tilde{z}_0(d,x)|, \nu_0(d,x), \pi_t(d,x))}}{\sum_{\alpha = 0}^{L} e^{\ell(\eta, \delta, |\tilde{z}_0(d,x)|, \nu_0(d,x), \pi_t(d,x))}}
\]
and
\[
\bar{\alpha}^o(\delta, \eta, |\tilde{z}_0(d,x)|, \nu_0(d,x), \pi_t(d,x)) := \frac{e^{\bar{\ell}(\eta, \delta, |\tilde{z}_0(d,x)|, \nu_0(d,x), \pi_t(d,x))}}{\sum_{\alpha = 0}^{L} e^{\ell(\eta, \delta, |\tilde{z}_0(d,x)|, \nu_0(d,x), \pi_t(d,x))}},
\]
and
\[
\bar{\ell}(\eta, \delta, |\tilde{z}_0(d,x)|, \nu_0(d,x), \pi_t(d,x)) := -\log \sigma_t(d,x) - 0.5 \frac{(\delta - |\tilde{z}_0(d,x)|)^2}{\sigma_t^2(d,x)},
\]
\[
\ell(\eta, \delta, |\tilde{z}_0(d,x)|, \nu_0(d,x), \pi_t(d,x)) := \begin{cases}
1 \{ (\delta + |\tilde{z}_0(d,x)|)^2 < \sigma_t^2(d,x) \} \log \phi(\delta; -|\tilde{z}_0(d,x)|, \sigma_t^2(d,x)) \\
+ 1 \{ (\delta + |\tilde{z}_0(d,x)|)^2 > \sigma_t^2(d,x) \} \log \phi(\delta; -|\tilde{z}_0(d,x)|, \sigma_t^2(d,x))
\end{cases}.
\]
and
\[
\sigma_t^2(d,x) \geq \sigma^2_t(d,x) := 1/\nu_0^0(d,x) + r^{-1}/(\pi_t(d,x) + \eta)
\]
and \(\sigma_t^2(d,x) \leq \sigma^2_t(d,x) := 1/\nu_0^0(d,x) + r^{-1}/(\pi_t(d,x) - \eta)\).

**Remark 3.** For any \(\eta \in (0, \epsilon)\), we define
\[
\bar{\ell}(\eta, |\tilde{z}_0^o(d,x)|, \nu_0^0(d,x), \pi_t(d,x)) := \bar{\ell}(\eta, \eta, |\tilde{z}_0(d,x)|, \nu_0^0(d,x), \pi_t(d,x))
\]
and
\[
\ell(\eta, |\tilde{z}_0^o(d,x)|, \nu_0^0(d,x), \pi_t(d,x)) := \ell(\eta, \eta, |\tilde{z}_0^o(d,x)|, \nu_0^0(d,x), \pi_t(d,x)).
\]
\(\triangle\)

The next lemma summarizes some useful properties of \(\ell\) and \(\bar{\ell}\).

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

1. As \(|\tilde{z}_0^o(d,x)| \to \infty\), \(\bar{\ell}(\eta, \delta, |\tilde{z}_0^o(d,x)|, \nu_0^0(d,x), \pi_t(d,x))\) and \(\ell(\eta, \delta, |\tilde{z}_0^o(d,x)|, \nu_0^0(d,x), \pi_t(d,x))\) converge to \(-\infty\).
2. \( \eta \mapsto \ell(\eta, \delta, |\xi^0_0(d,x)|, \nu^0_0(d,x), \pi_t(d,x)) \) is increasing and \( \eta \mapsto \tilde{\ell}(\eta, \delta, |\xi^0_0(d,x)|, \nu^0_0(d,x), \pi_t(d,x)) \) is decreasing.

3. \( \delta \mapsto \tilde{\ell}(\eta, \delta, |\xi^0_0(d,x)|, \nu^0_0(d,x), \pi_t(d,x)) \) is non-decreasing, and \( \delta \mapsto \ell(\eta, \delta, |\xi^0_0(d,x)|, \nu^0_0(d,x), \pi_t(d,x)) \) is decreasing.

4. Suppose Assumption 2 holds and \( \pi_t = \epsilon_t \). For any \( \eta < \epsilon \), \( \overline{\sigma}_2^2(d,x) \geq 1/\nu^0_0(d,x) + t^{-1}/(1+\eta) =: \overline{\sigma}_t^2(d,x) \) and \( \overline{\sigma}_2^2(d,x) \leq 1/\nu^0_0(d,x) + t^{-1}/(\epsilon - \eta) =: \overline{\sigma}_t^2(d,x) \). Thus,

\[
\tilde{\ell}(\eta, \delta, |\xi^0_0(d,x)|, \nu^0_0(d,x), \pi_t(d,x)) \leq \tilde{\ell}(\eta, \delta, |\xi^0_0(d,x)|, \nu^0_0(d,x))
\]

and

\[
\ell(\eta, \delta, |\xi^0_0(d,x)|, \nu^0_0(d,x), \pi_t(d,x)) \geq \ell(\eta, \delta, |\xi^0_0(d,x)|, \nu^0_0(d,x))
\]

where the RHS bounds are defined as \( \tilde{\ell} \) and \( \tilde{\ell} \) but using \( \overline{\sigma} \) and \( \overline{\sigma} \) instead of \( \sigma \) and \( \sigma \). The usefulness of these bounds is that they do not depend on \( \epsilon_t \) but they still inherit properties 1-3.

Moreover,

\[
\alpha^o(\eta, \delta, |\xi_0(d,x)|, \nu_0(d,x), \epsilon_t(d,x)) \geq \alpha^o(\eta, \delta, |\xi_0(d,x)|, \nu_0(d,x)) := \frac{e^{\tilde{\ell}(\eta, \delta, |\xi^0_0(d,x)|, \nu^0_0(d,x))}}{\sum_{o^*=0}^L e^{\tilde{\ell}(\eta, \delta, |\xi^0_0(d,x)|, \nu^0_0(d,x))}}
\]

and

\[
\overline{\alpha}^o(\eta, \delta, |\xi_0(d,x)|, \nu_0(d,x), \epsilon_t(d,x)) \geq \overline{\alpha}^o(\eta, \delta, |\xi_0(d,x)|, \nu_0(d,x)) := \frac{e^{\ell(\eta, \delta, |\xi^0_0(d,x)|, \nu^0_0(d,x))}}{\sum_{o^*=0}^L e^{\ell(\eta, \delta, |\xi^0_0(d,x)|, \nu^0_0(d,x))}}.
\]

5. The functions \( \overline{\alpha}^o(\eta, \delta, |\xi_0(d,x)|, \nu_0(d,x), \epsilon_t(d,x)) \) and \( \overline{\alpha}^o(\eta, \delta, |\xi_0(d,x)|, \nu_0(d,x)) \) are increasing in \( \eta \) and \( \delta \).

Property 1 is key for our analysis. It shows that the upper and lower bound maintain a key property of the weights: If the bias of a model is “very large”, then the corresponding weight will be “small”.

### B.4 Proof of Supplemental Lemmas

Proof of Lemma 2. Henceforth, we omit the notation "x" from the quantities as there is no risk of confusion. Let \( \mathcal{F}^s \) denote the \( \sigma \)-algebra generated by \( (D_1, \ldots, D_s, Y_1, \ldots, Y_{s-1}) \). We now show that
for each \( \theta \in \Theta \), \((g(Y_s, D_s, \theta))_s\) is a MDS with respect to aforementioned \( \sigma \)-algebras. To do this, note that

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

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

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

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

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

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

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

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

\[
\leq 2t^{-1} \sum_{s=1}^{t} \sqrt{E\left[ \sup_{\theta \in B(\theta_l, \delta)} |\ell(Y_s(d, x), \theta) - \ell(Y_s(d, x), \theta_l)|^2 \right]}
\]

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

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 \(\delta > 0\),

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

52
for some universal constant $C$. Since $\Theta$ is bounded, the continuity assumption implies that \( \max \mathbb{E} \left[ (\ell(Y,d,x), \theta) \right] \leq C \) for some finite constant.

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

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

For any $t$, choose $\delta$ as the argmin of $\Lambda(t, |\Theta|) := \min_{\delta \geq 0} (t^{-1} \delta^{-\vert \Theta \vert} + \varphi(\delta))$. Observe that $\Lambda(\ldots)$ 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 $t^{-1} \delta^{-\vert \Theta \vert} + \varphi(\delta)$ converges to $\varphi(\delta)$ (pointwise) and $\varphi(0) = 0$.

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

\[
P \left( \sup_{\theta \in \Theta} |t^{-1} \sum_{s=1}^{t} g(Y_s, D_s, \theta)| \geq \sqrt{(0.5/C)M \Lambda(t, |\Theta|)} \right) \leq C (\varepsilon^{-1} t^{-1} L_\delta + \varepsilon^{-1} \delta^*) \leq M^{-1}
\]

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

\[\square\]

\textbf{Proof of Lemma 3.} For any $(d,x) \in \mathcal{D} \times \mathcal{X}$ and any $o, o' \in \{0, \ldots, L\}$ observe that

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

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

We observe that

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

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

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

53
In addition, note that

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

where \( g(Y_s, D_s, \theta) := (1\{D_s(x) = d\} (\ell(Y_s, \theta) - E[\ell(Y_s, \theta) | D_s(x)]) \) and the last line follows because \( 1\{D_s(x) = d\} E[\ell(Y_s, \theta) | D_s(x)] = E[\ell(Y_s, \theta) | D_s(x) = d] = E[\ell(Y_s(d,x), \theta)] \) as \( Y(d,x) \) is IID and in particular independent of \( D_s(x) \).

By Lemma 2, for any \( \gamma > 0 \),

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

where \( r_t = \sqrt{1/(2C\gamma)} t^{-0.5} \) (\( C \) is an universal constant defined inside the proof of the lemma).

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

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

and

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

Hence, since \( KL_{d,x}(\theta) := -E[\ell(Y(d,x), \theta)] \) it follows that, with probability greater or equal to
\[
1 - \gamma,
\]
\[
\frac{\alpha_t^\circ(d, x)}{\alpha_t^\circ(d, x)} \leq \frac{N_t(d, x) \int_0^\infty \mu_0^\circ(d, x) \left( KL_{d, x}(\theta) - \frac{\text{tr}_t}{N_t(d, x)} \leq v \right) e^{-N_t(d, x)v} d \nu}{N_t(d, x) \int_0^\infty \mu_0^\circ(d, x) \left( KL_{d, x}(\theta) + \frac{\text{tr}_t}{N_t(d, x)} \leq v \right) e^{-N_t(d, x)v} d \nu} = \frac{N_t(d, x) \int_0^\infty \mu_0^\circ(d, x) \left( KL_{d, x}(\theta) \leq v \right) e^{-N_t(d, x)v + \text{tr}_t} d \nu}{N_t(d, x) \int_0^\infty \mu_0^\circ(d, x) \left( KL_{d, x}(\theta) \leq v \right) e^{-N_t(d, x)v - \text{tr}_t} d \nu}
\]

and
\[
\frac{\alpha_t^\circ(d, x)}{\alpha_t^\circ(d, x)} \geq \frac{N_t(d, x) \int_0^\infty \mu_0^\circ(d, x) \left( KL_{d, x}(\theta) + \frac{\text{tr}_t}{N_t(d, x)} \leq v \right) e^{-N_t(d, x)v} d \nu}{N_t(d, x) \int_0^\infty \mu_0^\circ(d, x) \left( KL_{d, x}(\theta) - \frac{\text{tr}_t}{N_t(d, x)} \leq v \right) e^{-N_t(d, x)v} d \nu} = \frac{N_t(d, x) \int_0^\infty \mu_0^\circ(d, x) \left( KL_{d, x}(\theta) \leq v \right) e^{-N_t(d, x)v - \text{tr}_t} d \nu}{N_t(d, x) \int_0^\infty \mu_0^\circ(d, x) \left( KL_{d, x}(\theta) \leq v \right) e^{-N_t(d, x)v + \text{tr}_t} d \nu}
\]

where the second line(s) follow from a change of variable. Finally, it is easy to see that \( u \mapsto \mu_0^\circ(d, x) \left( KL_{d, x}(\theta) \leq u \right) =: G_{d, x}(u) \) is a CDF. \( \square \)

**Proof of Lemma 4.** Note that \( \phi(m_t(d, x); \xi_0^\circ(d, x), (N_t(d, x) + \nu_0^\circ(d, x)) / (N_t(d, x) \nu_0^\circ(d, x))) = \phi(\bar{m}_t(d, x) - \bar{\xi}_0^\circ(d, x); 0, \sigma_t^2(d, x)) \), where \( \sigma_t^2(d, x) := (N_t(d, x) + \nu_0^\circ(d, x)) / (N_t(d, x) \nu_0^\circ(d, x)) \) and \( \bar{\cdot} \) indicates centered at \( \theta(d, x) \).

Under \( |\bar{m}_t(d, x)| \leq \delta \), if follows that
\[
(\bar{m}_t(d, x) - \bar{\xi}_0^\circ(d, x))^2 = (\bar{m}_t(d, x))^2 + (\bar{\xi}_0^\circ(d, x))^2 - 2\bar{m}_t(d, x)\bar{\xi}_0^\circ(d, x)
\]
\[
\leq \delta^2 + (\bar{\xi}_0^\circ(d, x))^2 - 2\bar{m}_t(d, x)|\bar{\xi}_0^\circ(d, x)|
\]
\[
\leq \delta^2 + (\bar{\xi}_0^\circ(d, x))^2 - 2|\bar{m}_t(d, x)||\bar{\xi}_0^\circ(d, x)|
\]
\[
\leq \delta^2 + (\bar{\xi}_0^\circ(d, x))^2 + 2\delta|\bar{\xi}_0^\circ(d, x)|
\]
\[
= (\delta + |\bar{\xi}_0^\circ(d, x)|)^2
\]

Therefore,
\[
\log \phi(\bar{m}_t(d, x) - \bar{\xi}_0^\circ(d, x); 0, \sigma_t^2(d, x)) \geq -\log \sigma_t(d, x) - 0.5 \frac{(\delta + |\bar{\xi}_0^\circ(d, x)|)^2}{\sigma_t^2(d, x)} + Cte
\]
\[
= \log \phi(\delta; -|\bar{\xi}_0^\circ(d, x)|, \sigma_t^2(d, x)) + Cte,
\]

55
where $Cte$ is an irrelevant constant.

Observe that $\sigma \mapsto \log \phi(y;0,\sigma^2)$ is such that $\frac{d \log \phi(y;0,\sigma^2)}{d\sigma} = -\sigma^{-1} + \sigma^{-3} y^2 = \sigma^{-1} (\sigma^{-2} y^2 - 1)$, so it is decreasing if $y^2 < \sigma^2$ and increasing if $y^2 > \sigma^2$. Also, under $|f_t(d,x) - \pi_t(d,x)| \leq \eta$,

$$\sigma_t^2(d,x) \geq \tilde{\sigma}_t^2(d,x) := 1/\nu_0(d,x) + t^{-1}/(\pi_t(d,x) + \eta)$$

and $\sigma_t^2(d,x) \leq \tilde{\sigma}_t^2(d,x) := 1/\nu_0(d,x) + t^{-1}/(\pi_t(d,x) - \eta)$.

Hence, if $(\delta + |\tilde{\xi}_0^o(d,x)|)^2 < \sigma_t^2(d,x)$ then

$$\log \phi(\delta; -|\tilde{\xi}_0^o(d,x)|, \sigma_t^2(d,x)) \geq \log \phi(\delta; -|\tilde{\xi}_0^o(d,x)|, \tilde{\sigma}_t^2(d,x))$$

if $(\delta + |\tilde{\xi}_0^o(d,x)|)^2 > \sigma_t^2(d,x)$ then

$$\log \phi(\delta; -|\tilde{\xi}_0^o(d,x)|, \sigma_t^2(d,x)) \geq \log \phi(\delta; -|\tilde{\xi}_0^o(d,x)|, \tilde{\sigma}_t^2(d,x)).$$

Hence a possible lower bound is given by

$$\ell(\eta, \delta, |\tilde{\xi}_0^o(d,x)|, \nu_0(d,x), \pi_t(d,x)) := 1 \{ (\delta + |\tilde{\xi}_0^o(d,x)|)^2 < \sigma_t^2(d,x) \} \log \phi(\delta; -|\tilde{\xi}_0^o(d,x)|, \tilde{\sigma}_t^2(d,x)) + 1 \{ (\delta + |\tilde{\xi}_0^o(d,x)|)^2 > \sigma_t^2(d,x) \} \log \phi(\delta; -|\tilde{\xi}_0^o(d,x)|, \tilde{\sigma}_t^2(d,x)).$$

To verify this is a valid lower bound, note that $(\delta + |\tilde{\xi}_0^o(d,x)|)^2 < \sigma_t^2(d,x)$ implies that $(\delta + |\tilde{\xi}_0^o(d,x)|)^2 < \tilde{\sigma}_t^2(d,x)$ and thus

$$\ell(\eta, \delta, |\tilde{\xi}_0^o(d,x)|, \nu_0(d,x), \pi_t(d,x)) = \log \phi(\delta; -|\tilde{\xi}_0^o(d,x)|, \tilde{\sigma}_t^2(d,x))$$

which by our previous displays is indeed a valid lower bound when $(\delta + |\tilde{\xi}_0^o(d,x)|)^2 < \sigma_t^2(d,x)$.

Similarly, if $(\delta + |\tilde{\xi}_0^o(d,x)|)^2 \geq \sigma_t^2(d,x)$, then $(\delta + |\tilde{\xi}_0^o(d,x)|)^2 \geq \tilde{\sigma}_t^2(d,x)$ and

$$\ell(\delta, \eta, |\tilde{\xi}_0^o(d,x)|, \nu_0(d,x), \pi_t(d,x)) = \log \phi(\delta; -|\tilde{\xi}_0^o(d,x)|, \tilde{\sigma}_t^2(d,x))$$

which, again, by our previous calculations is a valid lower bound when $(\delta + |\tilde{\xi}_0^o(d,x)|)^2 > \sigma_t^2(d,x)$.

On the other hand,

$$(\tilde{m}_t(d,x) - \tilde{\xi}_0^o(d,x))^2 \geq \max\{ (\tilde{\xi}_0^o(d,x))^2 - 2\delta|\tilde{\xi}_0^o(d,x)|, 0\}.$$
Therefore,

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

$$\leq - \log \sigma_t(d,x) - 0.5 \frac{\max\{(\tilde{\zeta}_0^o(d,x))^2 - 2\delta|\tilde{\zeta}_0^o(d,x)|,0\}}{\sigma_t^2(d,x)} + Cte$$

$$= \ell(\eta,\delta,|\tilde{\zeta}_0^o(d,x)|,\nu_0^o(d,x),\pi_t(d,x)).$$

\[\square\]

**Proof of Lemma 5.**  (1) The proof is omitted as these property readily follows from Lemma 4.

(2) We first observe that $\sigma_t^2$ is decreasing as a function of $\eta$ and $\tilde{\sigma}_t^2$ is increasing as a function of $\eta$.

Second, observe that $\ell$ is decreasing as a function of $\sigma_t^2$ and increasing as a function of $\tilde{\sigma}_t^2$. Thus, $\bar{\ell}$ is increasing as a function of $\eta$.

Third, suppose $(\delta + |\tilde{\zeta}_0^o(d,x)|)^2 < \sigma_t^2(d,x)$, then

$$\ell(\eta,\delta,|\tilde{\zeta}_0^o(d,x)|,\nu_0^o(d,x),\pi_t(d,x)) = \log \phi(\delta;|\tilde{\zeta}_0^o(d,x)|,\tilde{\sigma}_t^2(d,x)).$$

Since $x \mapsto \log \phi(\delta + |\tilde{\zeta}_0^o(d,x)|;0,0)$ is decreasing when $(\delta + |\tilde{\zeta}_0^o(d,x)|)^2 < x$, it follows that the RHS of the display is decreasing as a function of $\eta$. Now suppose $(\delta + |\tilde{\zeta}_0^o(d,x)|)^2 > \sigma_t^2(d,x)$, then

$$\ell(\eta,\delta,|\tilde{\zeta}_0^o(d,x)|,\nu_0^o(d,x),\pi_t(d,x)) = \log \phi(\delta;|\tilde{\zeta}_0^o(d,x)|,\sigma_t^2(d,x))$$

and since $x \mapsto \log \phi(\delta + |\tilde{\zeta}_0^o(d,x)|;0,0)$ is increasing when $(\delta + |\tilde{\zeta}_0^o(d,x)|)^2 > x$ and $\sigma_t^2(d,x)$ is decreasing as a function of $\eta$, it follows that the RHS of the display is decreasing as a function of $\eta$.

Hence, we showed that $\ell(\eta,\delta,|\tilde{\zeta}_0^o(d,x)|,\nu_0^o(d,x),\pi_t(d,x))$ is decreasing as a function of $\eta$.

(3) Follows by inspection of $\underline{\ell}$ and $\bar{\ell}$.

(4) The proof is omitted as these property readily follows from Lemma 4.

(5) We do the proof for $\bar{\sigma}^o(\eta,\delta,|\tilde{\zeta}_0^o(d,x)|,\nu_0^o(d,x),\pi_t(d,x))$ as the proof of $\bar{\sigma}^o(\eta,\delta,|\tilde{\zeta}_0^o(d,x)|,\nu_0^o(d,x))$ is analogous. By property 2, $\exp \bar{\ell}$ is increasing and $\exp \underline{\ell}$ is decreasing. As both these quantities are positive and $\bar{\sigma}^o(\eta,\delta,|\tilde{\zeta}_0^o(d,x)|,\nu_0^o(d,x),\pi_t(d,x))$ is essentially the ratio of the first over the second one, it follows that $\bar{\sigma}^o(\eta,\delta,|\tilde{\zeta}_0^o(d,x)|,\nu_0^o(d,x),\pi_t(d,x))$ is increasing in $\eta$. A similar argument
delivers increasing in $\delta$. \hfill \Box

C  Stochastic Properties of the Bayesian Posteriors and weights

Timing. We first recall the timing. At period $t \in \{1, \ldots, T, \ldots\}$, the planner starts with beliefs $(\xi_{t-1}^o, v_{t-1}^o)^L_{o=0}$ and weights $(\alpha_{t-1}^o)^L_{o=0}$, it takes an action $D_t$, the corresponding outcome is realized $Y_t(D_t)$ and beliefs are updated given this new information.

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

Lemma 6. For any $t \geq 0$, any $o \in \{0, \ldots, L\}$ and any $d \in \{0, \ldots, M\}$,

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

$$\xi_{t+1}^o(d) = \frac{v_0^0(d) \xi_{t+1}^0(d)}{v_{t+1}^0(d) + v_0^o(d) - v_0^0(d)} + \frac{v_0^0(d) \xi_0^o(d) - v_0^o(d) \xi_0^0(d)}{v_{t+1}^0(d) + v_0^o(d) - v_0^0(d)}$$

Proof of Lemma 6. 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

$$\xi_{t+1}(d) = \frac{1\{D_{t+1} = d\} Y_{t+1}(d)}{v_t(d) + 1\{D_{t+1} = d\}} + \frac{v_t(d) \xi_t(d)}{v_t(d) + 1\{D_t = d\} Y_t(d)}$$

$$= \frac{1\{D_{t+1} = d\} Y_{t+1}(d)}{v_t(d) + 1\{D_{t+1} = d\}} + \frac{v_t(d) \xi_t(d)}{v_t(d) + 1\{D_t = d\} Y_t(d)}$$

$$+ \frac{v_t(d) \xi_{t-1}(d)}{v_t(d) + 1\{D_{t+1} = d\}}$$

since $v_t(d) = v_{t-1}(d) + 1\{D_t = d\}$, and iterating in this fashion, it follows that

$$\xi_{t+1}(d) = \sum_{s=1}^{t+1} \frac{1\{D_s = d\} Y_s(d)}{v_t(d) + 1\{D_{t+1} = d\}} + \frac{v_0(d) \xi_0(d)}{v_{t+1}(d) + 1\{D_{t+1} = d\}}.$$
Hence, \( \sum_{t=1}^{t+1} 1 \{ D_t = d \} Y_s(d) = \zeta_{t+1}(d) (v_t(d) + 1 \{ D_{t+1} = d \}) - \nu_0(d) \xi_0(d). \) Since the same equation holds for \( \zeta_{t+1}^o(d) \), it follows that

\[
\zeta_{t+1}(d) (v_t(d) + 1 \{ D_{t+1} = d \}) - \nu_0(d) \xi_0(d) = \zeta_{t+1}^o(d) (v_t^o(d) + 1 \{ D_{t+1} = d \}) - \nu_0^o(d) \xi_0^o(d),
\]

which implies

\[
\zeta_{t+1}^o(d) = \frac{\zeta_{t+1}(d) (v_{t+1}(d)) + \nu_0^o(d) \xi_0^o(d) - \nu_0(d) \xi_0(d)}{v_{t+1}^o(d)} + \frac{v_{t+1}^o(d)}{v_{t+1}^o(d) + \nu_0^o(d) - \nu_0(d)}
\]

\[
= \frac{\zeta_{t+1}(d) (v_{t+1}(d))}{v_{t+1}(d) + \nu_0^o(d) - \nu_0(d)} + \frac{\nu_0^o(d) \xi_0^o(d) - \nu_0(d) \xi_0(d)}{v_{t+1}(d) + \nu_0^o(d) - \nu_0(d)}
\]

\( \Box \)

The next lemma shows that \( \alpha_t^o \) can also be written as a function of \( \zeta_t^o \) and \( v_t^o \), and the priors.

**Lemma 7.** 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

\[
\ell_t^o(d) = \log \phi \left( \frac{v_t^o(d)}{v_t^o(d) - \nu_0^o(d)} \right) \xi_t^o(d) - \frac{\nu_0^o(d)}{v_t^o(d) - \nu_0^o(d)} \xi_t^o(d); \xi_t^o(d), (v_t^o(d) - \nu_0^o(d) + \nu_0^o(d))/(v_t^o(d) - \nu_0^o(d) + \nu_0^o(d)) \right).
\]

**Proof of Lemma 7.** Throughout we omit the super script "0" from the relevant quantities.

It readily follows from the fact that \( \zeta_t(d) = \frac{N_t(d)}{v_t(d)} m_t(d) + \frac{\nu_0(d)}{v_t(d)} \xi_0(d) \) iff \( \frac{v_t(d)}{N_t(d)} \zeta_t(d) - \frac{\nu_0(d)}{N_t(d)} \xi_0(d) \) = \( m_t(d) \) and \( N_t(d) = v_t(d) - \nu_0(d) \).

\( \Box \)

### C.1 The Markov Chain for \( (Z_t)_t \)

Throughout we omit the super script "0" from \( (\zeta_t^o, v_t^o) \).

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 \( t \in \{0, \ldots, T\} \),

\[
V_t(v_0) := \{ a \in \{ v_0, 1 + v_0, \ldots, t + v_0 \}^{M+1}; \sum_{d=0}^{M} a(d) - v_0 < t \} \\
\partial V_t(v_0) := \{ a \in \{ v_0, 1 + v_0, \ldots, t + v_0 \}^{M+1}; \sum_{d=0}^{M} a(d) - v_0 = t \} \\
\mathcal{V}_t(v_0) := V_t(v_0) \cup \partial V_t(v_0).
\]

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

Henceforth, we

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 characterized by the following recursion: For any \( d \in \{0, \ldots, M\} \) and given \( z_t = (\zeta_t, v_t) \),

\[
v_{t+1}(d) = v_t(d) + 1\{D_{t+1} = d\} 1\{v_t \in V_t(v_0)\} \\
\zeta_{t+1}(d) = \frac{1\{D_{t+1} = d\} Y_{t+1}(d)}{v_t(d) + 1\{D_{t+1} = d\}} + \frac{v_t(d) \zeta_t(d)}{v_t(d) + 1\{D_{t+1} = d\}},
\]

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 \( \sigma^2(d) \) and, by Assumption 3, it is assumed to have full support PDF and to be sub-gaussian, i.e., \( E[\exp \lambda(Y(d) - \theta(d))] \leq C \exp \nu \sigma^2(d) \lambda^2 \) for some constants \( C = 1, \nu > 0 \) and any \( \lambda > 0 \).

### C.2 Properties of the transition probability function \( Q \)

We now prove certain useful properties of \( Q \). For this, let \( 1_z(.) \) be the Dirac probability measure at \( z \). Also, we now define a family of conditional probability measures over \( \mathbb{Z} \). For any \( m \in \{0, \ldots\} \) and \( z_0 = (\zeta_0, v_0) \in \mathbb{Z} \), let \( Q_m(. \mid z_0) \) be such that

\[
Q_m(A_\zeta \times A_\nu \mid z_0) := 1_{\{v_0(m) + 1\{v_0(.) \in V_{T-m}\}\}}(A_\nu) Q(\zeta(m) \in A_\zeta \mid D = m, z_0(m)),
\]

for all \( A_\zeta \times A_\nu \) Borel in \( \mathbb{Z} \).
C.2.1 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

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

The next lemma provides a simple lower bound for $Q$ using the fact that $\delta(\cdot)(\cdot) \geq \epsilon$.

**Lemma 8.** For any set $A := \prod_{m=1}^M A(m)$ and any $z_0 \in \mathbb{Z}$, it follows that

$$Q^{L+1}(A \mid z_0) \geq \epsilon \int Q^L(A \mid z_1) Q(dz_1 \mid D_0 = d_0, z_0), \forall L \geq 0 \text{ and } d_0 \in \{0, \ldots, M\},$$

and

$$Q(A \mid z_0) \geq \epsilon \sum_{l=0}^M \prod_{d \neq l} 1_{z_0(d)}(A(d)) Q(A(l) \mid D = l, z_0).$$

**Proof of Lemma 8.** First we observe that for any $L \in \{0, \ldots, M\}$, any $z_0$ and any set $A$,

$$Q^{L+1}(A \mid z_0) = \int Q^L(A \mid z_1) Q(dz_1 \mid z_0)
= \int \sum_{d_0} \delta(z_0)(d_0) Q^L(A \mid z_1) Q(dz_1 \mid D_0 = d_0, z_0)
\geq \epsilon \int Q^L(A \mid z_1) \sum_{d_0} Q(dz_1 \mid D_0 = d_0, z_0),$$

where $Q^0(\cdot \mid z) := 1_z(\cdot)$ and the third line follows because $\delta(\cdot)(\cdot) \geq \epsilon$. Observe further that for any $d \in \{0, \ldots, M\}$,

$$Q(A \mid D_0 = d, z_0) = \prod_{d' \neq d} 1_{z_0(d')}(A(d')) Q(A(d) \mid D_0 = d, z_0).$$

Thus, for $L = 0$, this implies that

$$Q(A \mid z_0) \geq \epsilon \sum_{d=0}^M \prod_{d' \neq d} 1_{z_0(d')}(A(d')) Q(A(d) \mid D_0 = d, z_0).$$

$\square$
The next lemma shows that we can lower bound the transition $Q^L$ for different values of $L \in \{0,\ldots\}$ in terms of a product measure given by the family $(Q_m)_m$.

**Lemma 9.** For any set $A := \prod_{m=1}^M A(m)$ such that $A(m) := A_x(m) \times A_r(m)$, it follows that

1. For any $z_0 \in \mathbb{Z}$ and any $L \in \{1,\ldots,M\},^{21}$

   $$Q^{L+1}(A \mid z_0) \geq \epsilon^L \int Q(A \mid z_1(0), z_2(1), z_3(2), \ldots, z_L(L-1), z_0(L), \ldots, z_0(M)) \prod_{t=0}^{L-1} Q_r(dz_{t+1}(t) \mid z_0).$$

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 9.** (1) By the proof of Lemma 8, for $L \geq 1$ it follows that

$$Q^{L+1}(A \mid z) \geq \epsilon \int Q^L(A \mid z_1(0), z_0(1), \ldots, z_0(M))Q(dz_1(0) \mid D_0 = 0, z_0). \quad (10)$$

Analogously, for any $z_1,$

$$Q^L(A \mid z_1) \geq \epsilon \int Q^{L-1}(A \mid z_2)Q(dz_2 \mid D_1 = 1, z_1)$$

$$\geq \epsilon \int Q^{L-1}(A \mid z_1(0), z_2(1), z_1(2), \ldots, z_1(M))Q(dz_2(1) \mid D_1 = 1, z_1).$$

Plugging this expression in (10) with $z_1 = z_1(0), z_0(1), \ldots, z_0(M)$, it follows that

$$Q^{L+1}(A \mid z_0) \geq \epsilon^2 \int Q^{L-1}(A \mid z_1(0), z_2(1), z_0(2), \ldots, z_0(M))Q(dz_2(1) \mid D_1 = 1, z_1(0), z_0(1), \ldots, z_0(M))$$

$$\times Q(dz_1(0) \mid D_0 = 0, z_0).$$

Since

$$Q(A(d) \mid D_0 = d, z_0) = 1_{\{v_0(d)+1\{v_0(\cdot) \in \mathcal{V}\}\}}(A_r(d))Q(A_x(d) \mid D_0 = d, z_0(d)), \quad (11)$$

---

$^{21}D_r$ is the random variable corresponding to the treatment assignment in period $t$; $z_t$ is the state at time $t$. 

62
it follows that

$$Q^{L+1}(A \mid z_0) \geq \epsilon^2 \int Q^{L-1}(A \mid z_1(0), z_2(1), z_0(2), ..., z_0(M)) 1_{[v_0(1)+1\{v_0(.) \in V_T\}]}(dv_2(1)) Q(d\zeta_2(1) \mid D_1 = 1, z_0(1))$$

$$\times Q(dz_1(0) \mid D_0 = 0, z_0),$$

where $\nu_1(.) = (\nu_1(0), \nu_0(1), ..., \nu_0(M)) = (\nu_0(0) + 1\{\nu_0(.) \in V_T\}, \nu_0(1), ..., \nu_0(M))$ because at time $0, D_0 = 0$. Hence, $\nu_1(.) \in V_T$ iff $\nu_0 \in V_{T-1}$. Thus,

$$Q^{L+1}(A \mid z_0)$$

$$\geq \epsilon^2 \int Q^{L-1}(A \mid z_1(0), z_2(1), z_0(2), ..., z_0(M)) 1_{[v_0(1)+1\{v_0(.) \in V_{T-1}\}]}(dv_2(1)) 1_{[v_0(0)+1\{v_0(.) \in V_T\}]}(dv_1(0))$$

$$\times Q(d\zeta_1(0) \mid D_0 = 0, z_0(0)) Q(d\zeta_2(1) \mid D_1 = 1, z_0(1))$$

$$= \int Q^{L-1}(A \mid z_1(0), z_2(1), z_0(2), ..., z_0(M)) Q_1(dz_2(1) \mid z_0) Q_1(dz_1(0) \mid z_0)$$

Iterating in this fashion, it follows that

$$Q^{L+1}(A \mid z_0) \geq \epsilon^L \int Q(A \mid z_1(0), z_2(1), z_3(2), ..., z_L(L-1), z_0(L), ..., z_0(M))$$

$$\times \prod_{t=0}^{L-1} 1_{[v_0(t)+1\{v_0(.) \in V_{T-t}\}]}(dv_{t+1}(t))\prod_{t=0}^{L-1} Q(d\zeta_{t+1}(t) \mid D_t = t, z_0(t)).$$
(3) In particular, for $L = M$, these results imply that

$$Q^{M+1}(A \mid z_0) \geq e^{M} \int Q(A \mid z_1(0), z_2(1), z_3(2), \ldots, z_L(M-1), z_0(M))$$

$$\times \prod_{t=0}^{M-1} 1_{(y_0(t)+1 \{y_0(0) \in V_{r-t} \})} (d\nu_{t+1}(t)) \prod_{t=0}^{M-1} Q(d\zeta_{t+1}(t) \mid D_t = t, z_0(t))$$

$$\geq e^{M+1} \int \prod_{t=0}^{M-1} 1_{z_{r+1}(t)} (A(t)) Q(A(M) \mid D_M = M, z_1(0), z_2(1), z_3(2), \ldots, z_L(M-1), z_0(M))$$

$$\times \prod_{t=0}^{M-1} 1_{(y_0(t)+1 \{y_0(0) \in V_{r-M} \})} (d\nu_{t+1}(t)) \prod_{t=0}^{M-1} Q(d\zeta_{t+1}(t) \mid D_t = t, z_0(t))$$

$$\geq e^{M+1} \int \prod_{t=0}^{M-1} 1_{z_{r+1}(t)} (A(t)) 1_{(y_0(M)+1 \{y_0(0) \in V_{r-M} \})} (A_v(M)) Q(A_\zeta(M) \mid D_M = M, z_0(M))$$

$$\times \prod_{t=0}^{M-1} 1_{(y_0(t)+1 \{y_0(0) \in V_{r-M} \})} (d\nu_{t+1}(t)) \prod_{t=0}^{M-1} Q(d\zeta_{t+1}(t) \mid D_t = t, z_0(t)).$$

And using our definitions, this implies that

$$Q^{M+1}(A \mid z_0) \geq e^{M+1} \int \prod_{t=0}^{M} 1_{z(t)} (A(t)) Q(dz(0), \ldots, dz(M-1), dz(M) \mid z_0)$$

$$= e^{M+1} \prod_{t=0}^{M} Q_t(A(t) \mid z_0).$$

\[ \square \]

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 10.** For any $d \in \{0, \ldots, M\}$ 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 $E \subseteq \mathbb{R}$

$$Q(\zeta_1(d) \in E \mid D = d, z) \geq \delta_C \text{Leb}(E \mid C_\zeta(d)), \forall z \in C$$

where

$$\delta_C := \inf_{y \in C_\zeta(d)} f_d(y)(1 + v_0(d))$$

**Proof of Lemma 10.** Given $z$ and $D = d$, let $y \mapsto e(y, z(d)) := \frac{v}{1 + v(d)} + \frac{v(d)}{1 + v(d)} \zeta(d)$; also, recall that
$y(d) \sim F_d$ 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 $de(y, z) = dy/(1 + \nu(d))$. Thus, with a change of variables,

$$Q(\zeta_1(d) \in E \mid D = d, z) \geq (1 + \nu(d)) \inf_{y \in C_\zeta(d)} f_d(y) \int 1\{\zeta'(d) \in E \cap C_\zeta(d)\} d\zeta'(d)$$

$$\geq (1 + \nu(d)) \inf_{y \in C_\zeta(d)} f_d(y) \text{Leb}(E \mid C_\zeta(d))$$

where the last line follows from the fact that $\text{Leb}(C_\zeta(d)) \in (0, 1]$. The result thus follows from the fact that $\nu(d) \geq \nu_0(d)$.

□

Lemma 11. Any set $C = \prod_{m=0}^MC(m)$ where $C(m) := \{z(m), \nu(m)) : \zeta(m) \in S(m), \nu(m) = a\}$ with $S(m)$ bounded and with non-empty interior and $a(.) \in \mathcal{V}(\nu_0)$ is small, i.e.,

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

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

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

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 $m^* = T$).

Proof of Lemma 11. 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
\( A_v(m) \subseteq \{1, \ldots, T\} \) and \( A_\zeta(m) \subseteq \mathbb{R} \).

First assume \( a(.) \in \mathcal{V}_T \), then by Lemma 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 \( C \), \( \nu_0(.) = a(.) \). Hence, for each \( m \in \{0, \ldots, m^* - 1\} \),

\[
Q_m(A(m) \mid z_0) = 1_{a(m)+1}(A_v(m))Q(\zeta(m) \in A_\zeta(m) \mid D = m, z_0(m))
\]

and for each \( m \in \{m^*, \ldots, M\} \),

\[
Q_m(A(m) \mid z_0) = 1_{a(m)}(A_v(m))Q(\zeta(m) \in A_\zeta(m) \mid D = m, z_0(m)).
\]

By Lemma 10, it follows that, for each \( m \in \{0, \ldots, m^* - 1\} \),

\[
Q(\zeta(m) \in A_\zeta(m) \mid D = m, z_0(m)) \geq 1_{a(m)+1}(A_v(m))\delta_{C(m)} \text{Leb}(A_\zeta(m) \mid C(m)),
\]

and for each \( m \in \{m^*, \ldots, M\} \),

\[
Q_m(A(m) \mid z_0) \geq 1_{a(m)}(A_v(m))\delta_{C(m)} \text{Leb}(A_\zeta(m) \mid C(m)).
\]

Hence,

\[
Q^{M+1}(A \mid z_0) \geq \epsilon^M \prod_{m=0}^{M} \delta_{C(m)} 1_{a(m)+1(\{m < m^*\})} (A_v(m)) \text{Leb}(A_\zeta(m) \mid C(m))
\]

letting \( \delta_C := \min_m \delta_{C(m)} > 0 \), the result follows. \( \square \)

**Remark 4** (A remark about Assumption 2). By inspection of the proof of Lemma 11, and the other lemmas used as building blocks, it follows that Assumption 2 could be relaxed to allow for \( \epsilon \) to depend on the state, i.e., \( z \mapsto \epsilon(z) \) provided that \( \inf_{z \in C} \epsilon(z) > 0 \) for any \( C \) compact. \( \triangle \)
C.2.2 Drift Condition

In this section we show that $Q$ satisfies a drift condition.

**Lemma 12.** For any $a \geq 0$ and $A \geq 0$, the function $z \mapsto V(z) := 1 + a \| \vec{z} \|_1 + A \| v - \delta' V \|_1$ where $d \mapsto \vec{z}(d) := (\vec{z}(d) - \theta(d))/\sigma(d)$, satisfies

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

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

**Proof.** Suppose $z$ is such that $\|v\|_1 = T - l$ for some $l \in \{1, \ldots, T\}$.

It follows that

$$Q[V](z) = 1 + a \int \| \vec{z}' \|_1 Q(d \vec{z}' \mid z) + A \int \| v' - \delta' V \|_1 Q(d v' \mid z)$$

$$= 1 + a \sum_d \delta(z)(d) \int \left( \| \vec{z}'(d) \| + \sum_{m \neq d} |\vec{z}(m)| \right) Q(d \vec{z}' \mid D = d, z)$$

$$+ A \sum_d \delta(z)(d) \int \| (v'(d), v(-d)) - \delta' V \|_1 Q(d v' \mid D = d, z).$$

where $(v'(d), v(-d))$ denotes the vector where the $d$-th coordinate is $v'(d)$ and the rest of the coordinates are given by $v(-d)$. Because $\|v\|_1 = T - l$, it follows that $\| (v(d) + 1, v(-d)) \|_1 = T - l - 1$ and thus $\| (v(d) + 1, v(-d)) - V \|_1 = l - 1 \leq \frac{T-1}{T}\| v - V \|_1$. Moreover, if $l = 0$, then $v'(d) = v(d)$ and $\| (v'(d), v(-d)) - V \|_1 = 0$.

Also, observe that $|\vec{z}'(d)| \leq |\vec{y}|/(1 + v(d)) + |\vec{z}(d)|v(d)/(1 + v(d))$ and $v(d) \geq \nu_0$. Thus,

$$\sum_d \delta(z)(d) \int \left( |\vec{z}'(d)| + \sum_{m \neq d} |\vec{z}(d)| \right) Q(d \vec{z}' \mid D = d, z)$$

$$\leq |\vec{z}(0)| \left( \sum_d \delta(z)(d) \omega(d, 0) \right) + \ldots + |\vec{z}(M)| \left( \sum_d \delta(z)(d) \omega(d, M) \right)$$

$$+ a \sum_d \frac{\delta(z)(d)}{1 + v(d)} \int |\vec{y}| f_d(y) dy$$

where $\omega(d', d) = 1$ if $d' \neq d$ and $=(v(d))/(1 + v(d))$ if $d' = d$. Hence, for any $m$, $\sum_d \delta(z)(d) \omega(d, m) = \delta(z)(m) \frac{\nu(m)}{1+\nu(m)} + (1 - \delta(z)(m)) \leq \epsilon \frac{\nu(m)}{1+\nu(m)} + (1 - \epsilon) := \gamma(v(m))$ because $\delta(z)(d) \geq \epsilon$ by Assumption.
2. Let $\gamma(v) := \max_m \gamma(v(m))$. Thus

$$Q[V](z) = 1 + a \int \|\xi'\|_1 Q(d\xi' | z) + A \int \|\nu' - \partial'V_T\|_1 Q(d\nu' | z)$$

$$\leq 1 + a\gamma(v)\|\tilde{\xi}\|_1 + A \frac{T-1}{T} \|\nu - \partial'V_T\|_1 + \sum_d \frac{\delta(z)(d)}{1 + \nu(d)} \int |\tilde{y}| \phi(y; \theta(d), 1) dy$$

$$\leq \gamma(v) V(z) + (1 - \gamma(v)) + a \sum_d \frac{\delta(z)(d)}{1 + \nu(d)} \int |\tilde{y}| f_d(y) dy.$$ 

Since $\int |\tilde{y}| f_d(y) dy \leq 1$ and $\nu(d) > \nu_0(d)$ the desired result follows. \hfill $\square$

**Lemma 13.** Let $C \subseteq \mathbb{Z}$ and $a, A$ be as in Lemma 12 and let $\tilde{\gamma}$ and $R := \inf_{z \in C} V(z)$ be such that $\tilde{\gamma} > \gamma$ and

$$\gamma + (1 - \gamma + 0.5a)/R \leq \tilde{\gamma}$$

Then, for all $z \in \mathbb{Z}$,

$$Q[V](z) \leq \tilde{\gamma} V(z) + b 1_C(z)$$

**Proof.** From Lemma 12, for all $z \in \mathbb{Z}$,

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

Thus

$$Q[V](z) \leq \tilde{\gamma} V(z) - (\tilde{\gamma} - \gamma) V(z) + b$$

$$\leq \tilde{\gamma} V(z) - (\tilde{\gamma} - \gamma) R + b.$$

We now show that $-(\tilde{\gamma} - \gamma) R + b < 0$. To do this, note that

$-(\tilde{\gamma} - \gamma) R + b = (1 - \gamma) + 0.5a - (\tilde{\gamma} - \gamma) R$

$$= 1 + \gamma (R-1) + 0.5a - \tilde{\gamma} R.$$
Remark 5. Let

\[ C := \prod_{m=0}^{M} C(m) \]  

where \( C(m) := \{ (\zeta, \nu) : |\zeta(m)| \leq (R/a)/(M+1) \text{ and } \nu(m) = \nu(m) \} \) for some \( \nu \in \partial V_T \) and \( R > 0 \). By Lemma 11, this set is small.

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

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

\( \triangle \)

C.2.3 Geometric Ergodicity

Lemma 14. There exists a constant \( L \) and an invariant distribution of \( Q, \lambda \), such that, for any \( n \geq 1 \) and any \( z_0 \in \mathbb{Z} \),

\[ ||Q^n(.|z_0) - \lambda|| \leq LV(z_0)e^{-ng(\epsilon)} \]

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


Lemma 11 imply their condition 8 and Lemma 12 with \( C \) chosen as in Remark 5 implies their condition 10. Moreover, proposition 11 in RR holds with \( \alpha := \alpha(a, R, \gamma) = \gamma + (1 - \gamma + 0.5a)/(R+1) \), which by Remark 5 is less than 1; moreover, it is increasing in \( a \), decreasing in \( R \) and decreasing on \( \gamma \). Thus as pointed out in p. 47, their Theorem 12 holds. Thus,

\[ ||Q^n(.|z_0) - \pi|| \leq (1 - e^{M+1} \delta_C)^j + \alpha^{-n}B_M^{j-1}0.5(V(z_0) + \int V(z)\pi(dz)) \]

\[ = LV(z_0) \left( (1 - e^{M+1} \delta_C)^j + \alpha^{-n}B_M^j \right) \]

for any \( 0 \leq j \leq n \), where \( B_M := \max\{1, \alpha^{M+1}(1 - e^{M+1} \delta_C) \sup_{z \in C} Q^{M+1}[V](z) \} \).

We now bound the term \( \alpha^{-n}B_M^j \). Observe that \( \sup_{z \in C} Q^{M+1}[V](z) \leq \gamma^{M+1} \sup_{z \in C} V(z) + \frac{1 - \gamma^{M+1}}{1 - \gamma} b \leq \gamma^{M+1} \sup_{z \in C} V(z) \),
\( \gamma^{M+1}((1+R) - b) + \frac{b}{1-\gamma} \) (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 \( \gamma \) 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

\[
G(a, \eta, \gamma) := \gamma \left( \gamma^{M+1}(0.5a/(1-\gamma) + \eta + \gamma - 0.5a) + 1 + 0.5a/(1-\gamma) \right) \\
= \gamma \left( \gamma^{M+1}(0.5a\gamma/(1-\gamma) + \gamma) + 1 + 0.5a/(1-\gamma) \right),
\]

which is increasing in all of its arguments. Therefore,

\[
\alpha^{-n} B^j_M \leq \alpha(a, R(a, \eta, \gamma), \gamma)^{-n+j(M+1+m)} (G(a, \eta, \gamma)/\alpha(a, R(a, \eta, \gamma), \gamma)^m)^j, \quad \forall m \in \mathbb{N}.
\]

We note that

\[
G(0, \eta, \gamma) = \gamma (\gamma^{M+1}(\eta + \gamma) + 1) \text{ and } \alpha(0, R(0, \eta, \gamma), \gamma)^{-1} := \gamma + (1-\gamma)/(\eta + 1) < 1.
\]

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 \( m(\gamma, \eta) \) is non-decreasing on \( \gamma \) and non-decreasing on \( \eta \) (at least for large values of \( \eta \)). By continuity, there exists a \( a(\gamma, \eta) \) such that for all \( \alpha \leq a(\gamma, \eta), G(a, \eta, \gamma)/\alpha(a, R(a, \eta, \gamma), \gamma)^m(\gamma, \eta) \leq 1 \). Moreover, \( a(\gamma, \eta) \) is non-decreasing in \( \gamma \).

Therefore, for any \( \eta > 0 \) and \( \alpha \leq a(\gamma, \eta) \), it follows that

\[
\alpha^{-n} B^{j-1}_M \leq \alpha(a, R(a, \eta, \gamma), \gamma)^{-n+j(M+1+m(\gamma, \eta))}.
\]

By choosing \( j = 0.5n/(M+1+m(\gamma, \eta)) \) (if it is not an integer, simply take the floor), it follows that

\[
||Q^n(.,z_0) - \pi|| \leq LV(z_0)\alpha(a, R(a, \eta, \gamma), \gamma)^{0.5n/M+1+m(\gamma, \eta)} = LV(z_0) \exp\left\{-0.5n\frac{\log\alpha^{-1}(a, R(a, \eta, \gamma), \gamma)}{M+1+m(\gamma, \eta)}\right\}
\]

Observe that by the computing the total derivative of \( \gamma \mapsto \log\alpha^{-1}(a, R(a, \eta, \gamma), \gamma) \) with respect to \( \gamma \) it can be shown that this function is decreasing on \( \gamma \). 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 \( \gamma \). Since \( \gamma \) is decreasing on \( \epsilon \) 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 \( \epsilon \).
C.3 Mixing Results

Lemma 15. Let \( \pi \) be any probability over \( \mathbb{Z} \) such that \( \int V(z) \pi(dz) < \infty \) (in particular, it could be different from the invariant distribution \( \lambda \)). Then, the process \( (Z_t)_{t=1}^{\infty} \) with initial probability \( \pi \) and transition \( Q \) is \( \beta \)-mixing with mixing coefficients, \( (\beta(k))_k \) such \( \beta_k \leq C_M e^{-k0.5\varepsilon(\epsilon)} \) for all \( k \geq 1 \), where

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


Proof of Proposition 2. First, recall that \( z = (\zeta^0, \nu^0) \) and Lemma 15 establishes that this process is \( \beta \)-mixing. Since, for all \( o \in \{0,...,L\} \), \( (\zeta^o, \nu^o, \alpha^o) \) can be written as a deterministic function of \( z \) (and the priors; see Lemmas 6 and 7), conditional on the priors, the process \( (\zeta_t, \nu_t, \alpha_t)_t \) is also \( \beta \)-mixing with the same convergence rate.

By assumption, \( \delta \) depends on \( (\zeta^0_t, \nu^0_t, \alpha^0_t)_{o=0}^L \) and thus it depends on \( z_t = (\zeta^0_t, \nu^0_t) \) (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 15, \( (Z_t, U_t)_t \) is \( \beta \)-mixing with the same coefficients. Hence, \( (Z_t, D_t)_t \) is also 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 15, \( (Z_t, D_t, Y_t(0),...,Y_t(M))_t \) is \( \beta \)-mixing with the same coefficients. Therefore, \( (Z_t, D_t, Y_t)_t \) is a Markov Chain and is \( \beta \)-mixing with mixing coefficients, \( (\beta(k))_k \) such \( \beta_k \leq C e^{-0.5k\varepsilon(\epsilon)} \) for all \( k \geq 1 \). \( \square \)
Appendix for Section 3.2.1

Recall that for any $d \in \{0, \ldots, M\}$ and $t \geq 0$,\(^{22}\)

$$N_{t+1}(d) := \sum_{s=1}^{t+1} 1\{D_s = d\},$$

$$J_{t+1}(d) := \sum_{s=1}^{t+1} 1\{D_s = d\} Y_s(d)/(t+1),$$

$$f_{t+1}(d) := \sum_{s=1}^{t+1} 1\{D_s = d\} / (t+1),$$

$$m_{t+1}(d) := J_t(d)/f_t(d) = \sum_{s=1}^{t+1} 1\{D_s = d\} Y_s(d)/N_{t+1}(d),$$

and

$$e_{t+1}(d) := \sum_{s=1}^{t+1} E_\pi[1\{D_s = d\}] / (t+1) = \sum_{s=1}^{t+1} E_\pi[\delta(Z_s)(d)] / (t+1)$$

where the last equality follows from the LIE.

The quantity $\tilde{Y}_t$ is defined as $Y_t - \theta$; $\tilde{J}_t$ and $\tilde{m}_t$ are defined as the original quantities but with $\tilde{Y}_t$ instead of $Y_t$.

### D.1 Bounds for Posterior Means

Observe that for any $o \in \{0, \ldots, L\}$ and any $t \geq 0$,

$$\zeta_t^o(d) = \frac{1}{f_t(d) + \nu_0^o(d) / t} \left( J_t(d) + \zeta_0^o(d) \nu_0^o(d) / t \right).$$

The next lemma provides a non-stochastic bound for the RHS when $J_t(d)$ is within $\gamma \geq 0$ of 0 and $f_t(d)$ is within $\gamma \geq 0$ of $e_t(d)$; the proof is relegated to the end of the section. For this, let $\Omega : \mathbb{R} \times \mathbb{R} \times \mathbb{Q}_+ \times [0, 1] \rightarrow \mathbb{R}$ with

$$(a, b, c, d) \mapsto \Omega(a, b, c, d) := \frac{a}{d - a + c} + \frac{bc}{d - a + c} \left( \frac{1\{b \geq 0\}}{d - a + c} + \frac{1\{b < 0\}}{d + a + c} \right).$$

\(^{22}\)If $f_{t+1}(d) = 0$, then we set $m_{t+1} = 0.$
Lemma 16. For any $d \in \{0, \ldots, M\}$, any $o \in \{0, \ldots, L\}$, any $t \in \mathbb{N}$, if $|\bar{f}_t(d)| \leq \gamma$ and $|f_t(d) - e_t(d)| \leq \gamma$ for some $\gamma \geq 0$, then: (1)

$$|\zeta^o_t(d) - \theta(d)| \leq \Omega_t(\gamma) := \Omega(\gamma, |\bar{\zeta}^o_0(d)|, v_0(d)/t, e_t(d)).$$

and

1. It is increasing in the first argument (a).
2. It is increasing in the second argument (b).
3. If $b = 0$, then it is decreasing in the third argument (c).
4. If $b \geq 0$, then it is decreasing in its fourth argument (d).

and

$$\Omega(0, b, c, d) = \frac{bc}{d+c}.$$

(2) And,

$$\zeta^o_t(d) - \theta(d) \leq \Omega_{+,t}(\gamma) := \Omega(\gamma, \bar{\zeta}^o_0(d), v_0(d)/t, e_t(d)),$$

(3) And,

$$-(\zeta^o_t(d) - \theta(d)) \leq \Omega_{-,t}(\gamma) := \Omega(\gamma, -\bar{\zeta}^o_0(d), v_0(d)/t, e_t(d)),$$

The next lemma provides similar bounds for $(\zeta^o_t)$. To do this, let $\Gamma: \mathbb{R} \times \mathbb{R}^{L+1} \times \mathbb{N}^{L+1} \times [0, 1] \to \mathbb{R}$ be such that

$$\Gamma(\gamma, \zeta_0(o)(d) - \theta(d), v_0(d), e_t(d))$$

$$:= \sum_{o=0}^L \alpha^o(\gamma, \gamma/(e_t(d) - \gamma), |\zeta_0(o)(d) - \theta(d)|, v_0(d), e_t(d)) \Omega^+(\gamma, \zeta_0(o)(d) - \theta(d), v_0(d)/t, e_t(d))$$

$$+ \sum_{o=0}^L \alpha^o(\gamma, \gamma/(e_t(d) - \gamma), |\zeta_0(o)(d) - \theta(d)|, v_0(d), e_t(d)) \Omega^-(\gamma, \zeta_0(o)(d) - \theta(d), v_0(d)/t, e_t(d))$$

where $\Omega^+ := 1\{\Omega \geq 0\} \Omega$ and $\Omega^- := 1\{\Omega \leq 0\} \Omega$ and $\bar{\alpha}$ and $\underline{\alpha}$ are upper and lower bounds for weights and are defined in Lemma 4.

73
The following lemma is analogous to Lemma 16, its proof is also relegated to the end of this appendix.

**Lemma 17.** For any $d \in \{0, \ldots, M\}$, any $t \in \mathbb{N}$, if $|\bar{I}_t(d)| \leq \gamma$ and $|f_t(d) - e_t(d)| \leq \gamma$ for some $\gamma \geq 0$, then:

1. $|\zeta_t^a(d) - \theta(d)| \leq \Gamma(\gamma) := \Gamma(\gamma, |\bar{\zeta}_0(d)|, \nu_0(d), e_t(d))$,
2. and,
   
   $\xi_t^a(d) - \theta(d) \leq \Gamma_{+}(\gamma) := \Gamma(\gamma, \bar{\zeta}_0(d), \nu_0(d), e_t(d))$,
3. and,
   
   $-(\zeta_t^a(d) - \theta(d)) \leq \Gamma_{-}(\gamma) := \Gamma(\gamma, -\bar{\zeta}_0(d), \nu_0(d), e_t(d))$.
4. And $\gamma \mapsto \Gamma(\gamma, \bar{\zeta}_0(d), \nu_0(d), e_t(d))$ is non-decreasing.

**D.2 Concentration Bounds for $(f_t)_t$, $(J_t)_t$ and $(m_t)_t$**

The next lemmas provides concentration bounds for $(f_t)_t$, $(J_t)_t$ as well as for $(m_t)_t$. For this, we introduce the following constants: For any $\epsilon > 0$,

$$B(\epsilon) := \min\{0.5q(\epsilon), 1\}/C_M \quad (13)$$

$$C(\epsilon) := \frac{\min\{(0.5q(\epsilon))^2, 1\}}{181C_M^2} \quad (14)$$

where $C_M$ is as in Lemma 15. Note that $C$ and $B$ are non-decreasing as a function of $\epsilon$.

The proofs of the following lemmas are relegated to the end of this appendix.

**Lemma 18.** For any $d \in \{0, \ldots, M\}$, any initial probability $\pi$ such that $\int V(z) \pi(dz) < \infty$, any $a > 0$ and any $t \geq 2 \min\{0.5q(\epsilon), 2\}$:

$$P_\pi \left( \left| f_t(d) - t^{-1} \sum_{s=1}^t E_\pi[\delta(Z_s)(d)] \right| \geq a \right) \leq e^{-\frac{\min\{(0.5q(\epsilon))^2, 1\}a^2}{40\min\{|V|, \gamma|Q|, 1\}t\log t} \cdot 100C_M \log t}.$$
And if \( t \geq e^{4aB(\epsilon)} \), then

\[
P_\pi \left( \left| f_t(d) - t^{-1} \sum_{s=1}^{t} E_\pi [\delta(Z_s)(d)] \right| \geq a \right) \leq e^{-\log t a^2 C(\epsilon)}.
\]

**Remark 6.** The second expression shows the “cost” of not working with IID data. If \((D_t)_t\) were IID — as it is the case with \( \epsilon = 1 \) — then the bound will be \( 2e^{-ta^2} \). However, for general \( \epsilon < 1 \), \((D_t)_t\) are \( \beta \)-mixing with exponential decay (see Proposition 2). Because of this, we lose a factor \( \log t \) and a constant \( C(\epsilon) \). △

**Lemma 19.** For any \( d \in \{0, ..., M\} \), any \( a \geq 0 \) and any \( t \geq 0 \),

\[
P_\pi \left( \left| t^{-1} \sum_{s=1}^{t} (Y_s(d) - \theta(d)) 1\{D_s = d\} \right| \geq a \right) \leq 2e^{-t \lambda^*(a, \epsilon)}
\]

where

\[
\lambda^*(a, \epsilon) := \max_{\lambda \geq 0} \{a \lambda - F(\lambda, \epsilon)\},
\]

where \( F(\lambda, \epsilon) := \log(\epsilon M + (1 - \epsilon M)e^{0.5\nu\sigma(d^2)\lambda^2}) \), and \( \lambda^* \) is increasing in both arguments and \( \lambda^*(a, \epsilon) \geq \lambda^*(a, 0) = \frac{a^2}{2\nu\sigma(d^2)} \) and \( \lambda^*(0, \epsilon) = 0 \) and \( \lim_{a \to \infty} \frac{\lambda^*(a, \epsilon)}{a} = \infty \).

**Remark 7.** We use Assumption 3(i) in this lemma. In particular, it is used in order to get an upper bound with exponential decay. Indeed, the assumption could be replaced by sub-exponential or any other type of control on the MGF of \( Y(d) \), e.g., \( E[e^{\kappa Y(d) - \theta(d)}] \leq e^{\kappa(\lambda)} \) 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

\[
P_\pi \left( \left| t^{-1} \sum_{s=1}^{t} (Y_s(d) - \theta(d)) 1\{D_s = d\} \right| \geq a \right) \leq 2e^{-t \max_{\lambda \geq 0} \{a \lambda - F(\lambda, \epsilon)\}}.
\]

with where \( F(\lambda, \epsilon) := \log(\epsilon M + (1 - \epsilon M)e^{\kappa(\lambda)}) \). △

For completeness, we provide a concentration bound for \( m_t \) around the true average effect \( \theta \).

**Lemma 20.** For any \( d \in \{0, ..., M\} \), any initial probability \( \pi \) such that \( \int V(z)\pi(dz) < \infty \), any \( a > 0 \)
and any \( t \geq e^{4dB(c)} \),

\[
P_\pi (|\tilde{m}_t(d) - \theta(d)| > a) \leq 3 \max \left\{ e^{- \frac{t}{\log(\sqrt{\frac{\epsilon_t(d)d}{T}})^2} \min \left\{ C(\epsilon) \cdot 0.5 \frac{\log t}{\nu(\sigma(d))^2}, e \left( \log t - 0.5 \frac{a^2}{\nu(\sigma(d))^2} \right) \right\}}, 3\right\}
\]

\[
\leq 3e^{- \frac{t}{\log(\sqrt{\frac{\epsilon_t(d)d}{T}})^2} \min \left\{ C(\epsilon) \cdot 0.5 \frac{\log t}{\nu(\sigma(d))^2}, e \left( \log t - 0.5 \frac{a^2}{\nu(\sigma(d))^2} \right) \right\}}.
\]

**D.3 Proof Proposition 3 and Corollary 1**

We now prove Proposition 3.

Proof of Proposition 3. Recall that \( \tilde{\zeta}_t(d) := \zeta_t(d) - \theta(d) \), \( \tilde{Y}_s(d) := (Y_s(d) - \theta(d)) \), \( \tilde{J}_t(d) := \sum_{s=1}^t 1 \{ D_s = d \} \tilde{Y}_s(d) / t \) and \( f_t(d) := \sum_{s=1}^t 1 \{ D_s = d \} / t \).

For any \( \gamma \geq 0 \), let

\[
S(t, \gamma) := \{|\tilde{J}_t(d)| \leq \gamma\},
\]

and

\[
R(t, \gamma) := \{|f_t(d) - e_t(d)| \leq \gamma\},
\]

where, recall, \( e_t(d) = t^{-1} \sum_{s=1}^t E[\delta(Z_s)(d)] \). Given these sets and with \( \gamma \leq e_t(d) \), it also follows that

\[
|m_t(d)| = |\tilde{J}_t(d) / f_t(d)| \leq \gamma / (e_t(d) - \gamma).
\]

Therefore, by Lemma 4, it follows that, for any \( o \in \{0, \ldots, L\} \)

\[
\alpha_t^o(d) \leq \tilde{\alpha}(\gamma, \gamma / (e_t(d) - \gamma), |\tilde{\zeta}_t^o(d)|, \nu_t^o(d), e_t(d)).
\]
This, and the fact that for any $d$, any $o$ and any $t \geq 1$,
\[
\zeta_t^o(d) = \sum_{i=1}^t 1\{D_s = d\}Y_s(d) + \sum_{i=1}^t 1\{D_t = d\}Y_t(d) + \sum_{i=1}^t 1\{D_{i+1} = d\}Y_{i+1}(d) + \sum_{i=1}^t 1\{D_{i+1} = d\}Y_{i+1}(d) + \sum_{i=1}^t 1\{D_{i+1} = d\}Y_{i+1}(d)
\]

\[
= \frac{1}{f_t(d) + \nu_t^o(d)/t} \tilde{f}_t(d) + \frac{\nu_t^o(d)}{f_t(d) + \nu_t^0(d)/t} \tilde{\nu}_t^o(d),
\]

imply that conditional on $S(t, \gamma) \cap R(t, \gamma)$,
\[
\left| \sum_{o=0}^L \alpha_t^o(d) \zeta_t^o(d) - \theta(d) \right| \leq \sum_{o=0}^L \tilde{a}(\gamma, \gamma/(e_t(d) - \gamma), [\tilde{\zeta}_t^o(d)], v_0^o(d), e_t(d)) \Omega(\gamma, \tilde{\zeta}_t^o(d), v_0^o(d), e_t(d))
\]
\[
= \Gamma(\gamma, \tilde{\zeta}_t^o(d), v_0^o(d), e_t(d)).
\]

Therefore, for any $\delta \in \mathbb{R}$,
\[
P_\pi \left( \left| \sum_{o=0}^L \alpha_t^o(d) \zeta_t^o(d) - \theta(d) \right| > \delta \right) \leq \int \frac{1}{\Gamma(\gamma, \tilde{\zeta}_t^o(d), v_0^o(d), e_t(d)) > \delta} + P_\pi \left( S(t, \gamma)^C \right) + P_\pi \left( R(t, \gamma)^C \right).
\]

Now, set $\delta = \Gamma(\gamma_t^*(\varepsilon), \tilde{\zeta}_t^o(d), v_0^o(d), e_t(d))$, $\gamma = \gamma_t^*(\varepsilon) = \sqrt{\log t / \varepsilon} \sqrt{C(e)}$. Observe that $\gamma_t^*(\varepsilon) \leq e_t(d)$, or equivalently, $\log t / \varepsilon \leq e_t(d)^2 C(e)$ holds because it is assumed that $\frac{\varepsilon}{e_t(d)^2 C(e)} \leq t / \log t$.

We invoke Lemmas 19 and 18, for this we need to check that $t \geq e^{\max\{4\gamma_t^*(\varepsilon)B(e), 2\nu_t^o(d)^2\}}$. The function $t \mapsto \sqrt{\log t / \varepsilon}$ has its maximum in the interior at $t = e$ and equal to $\sqrt{\log e / \varepsilon} = 1 / \sqrt{\varepsilon} < 1$. Hence, $t \geq e^{\max\{4\gamma_t^*(\varepsilon)B(e), 2\nu_t^o(d)^2\}}$ is implied by $t \geq e^{\max\{4\sqrt{\log e / \varepsilon}B(e), 2\nu_t^o(d)^2\}}$, which is assumed. Therefore, by Lemmas 19 and 18 and our choice of $\delta$, the RHS in the previous display is less than $3e^{-t \gamma_t^*(\varepsilon)^2 C(e)}$. By our choice of $\gamma_t^*(\varepsilon)$, it follows that
\[
P_\pi \left( \left| \tilde{\zeta}_t^o(d) \right| > \delta \right) \leq 3e^{-\varepsilon},
\]
as desired.

The monotonicity claim follows from Lemma 17(4) and the fact that $C$ is non-decreasing.
We now proof Corollary 1.

**Proof of Corollary 1.** We can prove the result using limits. For any given \( \alpha \neq 0 \), let \( |\tilde{z}_0^\alpha(d)| := \nu_0^\alpha(d)\|\tilde{z}_0^\alpha(d)\| \) and consider the limit of this quantity going to \( \infty \).

By Lemmas 4 and 16, \( \overline{\sigma}^\alpha(\gamma, |\tilde{z}_0^\alpha(d)|, \nu_0^\alpha(d), \epsilon_t(d)) \to 0 \) and \( \Omega(\gamma, |\tilde{z}_0^\alpha(d)|, \nu_0^\alpha(d), \epsilon_t(d)) \to \infty \). However, the product of these quantities converges to 0 because the first quantity converges faster to 0 (at an exponential rate) than the second quantity, which diverges but linearly. Therefore,

\[
\Gamma(\gamma, |\tilde{z}_0^\alpha(d)|, \nu_0^\alpha(d), \epsilon_t(d)) \to \Omega(\gamma, |\tilde{z}_0^\alpha(d)|, \nu_0^\alpha(d)/t, \epsilon_t(d)).
\]

Since \( \Gamma \) is jointly continuous as a function of \( \gamma \) and \( |\tilde{z}_0^\alpha(d)| \) and \( \nu_0(d) \), this convergence is uniform over \( \gamma \) in compact sets and over \( t \in \{1, \ldots, T\} \).

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

\[
\Omega\left(\sqrt{\frac{\log t}{t}} \sqrt{\frac{\varepsilon}{C(\varepsilon)}}, |\tilde{z}_0^\alpha(d)|, \nu_0^\alpha(d)/t, \epsilon_t(d)\right) \geq \Gamma\left(\sqrt{\frac{\log t}{t}} \sqrt{\frac{\varepsilon}{C(\varepsilon)}}, |\tilde{z}_0^\alpha(d)|, \nu_0(d), \epsilon_t(d)\right) - \delta
\]

for any \( |\tilde{z}_0^\alpha(d)| \geq C \). \( \square \)

**D.4 Proofs of Supplemental Lemmas**

**Proof of Lemma 16.** (1) We omit \( \alpha \) from the notation. Observe that

\[
\tilde{z}_t(d) = \sum_{s=1}^t 1\{D_s = d\} \tilde{Y}_s(d) = \frac{\nu_0(d)}{f_t(d) + \nu_0(d)/t} \tilde{Y}_s(d) + \frac{\nu_0(d)/t}{f_t(d) + \nu_0(d)/t} \tilde{z}_0(d)
\]

where \( \tilde{z}_t(d) := \tilde{z}_t(\gamma) - \theta(d), \tilde{Y}_s(d) := (Y_s(d) - \theta(d)), \tilde{J}_t(d) := \sum_{s=1}^t 1\{D_s = d\} \tilde{Y}_s(d)/t \) and \( f_t(d) := \sum_{s=1}^t 1\{D_s = d\}/t \).

Under our assumptions, \( |\tilde{J}_t(d)| \leq \gamma \) and thus

\[
|\tilde{z}_t(d)| \leq \gamma \frac{\nu_0(d)/t|\tilde{z}_0(d)|}{f_t(d) + \nu_0(d)/t} + \frac{\nu_0(d)/t|\tilde{z}_0(d)|}{f_t(d) + \nu_0(d)/t}.
\]
Moreover, as \(|f_i - \pi_t(d)| \leq \gamma\) it also follows that

\[
|\tilde{\xi}_t(d)| \leq \frac{\gamma}{\pi_t(d) - \gamma + v_0(d)/t} + \frac{v_0(d)/t}{\pi_t(d) - \gamma + v_0(d)/t}.
\]

We now show the properties of \(\Omega\). Properties 2-4 are trivial, we thus only prove 1. To do this, note that

\[
\frac{d\Omega(a, b, c, d)}{da} = \frac{d + c}{(d - a + c)^2} + \frac{bc 1\{b \geq 0\}}{(d - a + c)^2} - \frac{bc 1\{b < 0\}}{(d + a + c)^2}
\]

which is positive as \(d, c\) are non-negative.

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

\(\square\)

**Proof of Lemma 17.** By Lemma 4,

\[
\tilde{\xi}_t^\alpha(d) = \sum_{o=0}^{L} \alpha_t^o(d) \tilde{\xi}_t^o(d)
\]

\[
\leq \sum_{o=0}^{L} \frac{\exp \tilde{\ell}(\gamma, \gamma/(e_t(d) - \gamma), |\xi_0^o(d) - \theta(d)|, v_0^o(d), e_t(d))}{\sum_{o'=0}^{L} \exp \tilde{\ell}(\gamma, \gamma/(e_t(d) - \gamma), |\xi_0^{o'}(d) - \theta(d)|, v_0^{o'}(d), e_t(d))} 1\{\tilde{\xi}_t^o(d) \geq 0\} \tilde{\xi}_t^o(d)
\]

\[
+ \sum_{o=0}^{L} \frac{\exp \tilde{\ell}(\gamma, \gamma/(e_t(d) - \gamma), |\xi_0^o(d) - \theta(d)|, v_0^o(d), e_t(d))}{\sum_{o'=0}^{L} \exp \tilde{\ell}(\gamma, \gamma/(e_t(d) - \gamma), |\xi_0^{o'}(d) - \theta(d)|, v_0^{o'}(d), e_t(d))} 1\{\tilde{\xi}_t^o(d) < 0\} \tilde{\xi}_t^o(d),
\]

and

\[
-\tilde{\xi}_t^\alpha(d) = \sum_{o=0}^{L} \alpha_t^o(d) (-\tilde{\xi}_t^o(d))
\]

\[
\leq \sum_{o=0}^{L} \frac{\exp \tilde{\ell}(\gamma, \gamma/(e_t(d) - \gamma), |\xi_0^o(d) - \theta(d)|, v_0^o(d), e_t(d))}{\sum_{o'=0}^{L} \exp \tilde{\ell}(\gamma, \gamma/(e_t(d) - \gamma), |\xi_0^{o'}(d) - \theta(d)|, v_0^{o'}(d), e_t(d))} 1\{-\tilde{\xi}_t^o(d) \geq 0\} (-\tilde{\xi}_t^o(d))
\]

\[
+ \sum_{o=0}^{L} \frac{\exp \tilde{\ell}(\gamma, \gamma/(e_t(d) - \gamma), |\xi_0^o(d) - \theta(d)|, v_0^o(d), e_t(d))}{\sum_{o'=0}^{L} \exp \tilde{\ell}(\gamma, \gamma/(e_t(d) - \gamma), |\xi_0^{o'}(d) - \theta(d)|, v_0^{o'}(d), e_t(d))} 1\{-\tilde{\xi}_t^o(d) < 0\} (-\tilde{\xi}_t^o(d)).
\]

By Lemma 16, we know that \(|\tilde{\xi}_t^o(d)| \leq \Omega_t(\gamma); \tilde{\xi}_t^o(d) \leq \Omega_{+,t}(\gamma)\) and \(-\tilde{\xi}_t^o(d) \leq \Omega_{-,t}(\gamma)\). These and the above inequalities imply 1-3.
Throughout the proof, let $\gamma > 0$. In this case, $\Gamma$ is the sum of the product of a positive function $-\frac{\exp(\gamma(1/(e_{i}(d)) - \gamma), L_{i}(d), \nu_{i}(d), e_{i}(d))}{\sum_{i=0}^{L_{i}(d)}}$ and a negative one given by $\Omega$. By Lemma 5 and the fact that $\gamma \mapsto \gamma/(e_{i}(d) - \gamma)$ is increasing in the relevant domain, the first function is decreasing in $\gamma$. By Lemma 16, $\Omega$ is non-decreasing as a function of $\gamma$. Thus, the product is non-decreasing and so is the sum. Analogous arguments prove the same result for the case $\Omega \geq 0$. □

Proof of Lemma 18. By proposition 2, $(Z_{i}, \nu_{i}, \Omega_{i})$ is $\beta$-mixing with the same coefficients, and this implies that the $\alpha$-mixing coefficients are also of order $O(e^{-0.5\ell_{i}(\epsilon)})$.

Throughout the proof, let $c := 0.5\ell_{i}(\epsilon)$.

By Corollary 12 in Merlevède et al. (2009), it follows that for any $t \geq 2(\max\{c, 2\})$

$$P_{\pi}\left(\left|\sum_{s=1}^{t} (1\{D_{s} = d\} - E_{\pi}[\delta(Z_{s})(d)])\right| \geq ta\right) \leq e^{-\frac{\min\{c, 1\}t^{2}a^{2}}{4\alpha t + \log t}} \leq e^{-\frac{\min\{c, 1\}t^{2}a^{2}}{180C_{M} + \log t}} \leq e^{-\frac{\min\{c, 1\}t^{2}a^{2}}{4\alpha t + \log t}} = e^{-\frac{\min\{c, 1\}t^{2}a^{2}}{180C_{M} \log t}}$$

If $C_{M} \log t \geq 4a \min\{c, 1\}$, then $4a \min\{c, 1\} + 180C_{M} \log t \leq 181C_{M} \log t$ and thus

$$P_{\pi}\left(\left|\sum_{s=1}^{t} (1\{D_{s} = d\} - E_{\pi}[\delta(Z_{s})(d)])\right| \geq ta\right) \leq e^{-\frac{\min\{c, 1\}t^{2}a^{2}}{181C_{M} \log t}}.$$ □

Proof of Lemma 19. Let $W_{d}(d) := (Y_{d}(d) - \theta(d))1\{D_{s} = d\}$. By the Markov inequality, it follows that, for any $\lambda > 0$,

$$P_{\pi}\left(t^{-1} \sum_{s=1}^{t} W_{d}(d) \geq a\right) \leq E_{\pi}\left[\prod_{s=1}^{t} e^{\lambda W_{d}(d)}\right] e^{-a\lambda t}$$
Observe that

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

where \( E_t[.\] denotes the conditional expectation given \((Z_s)_{s=1}^t\) (but not \(Y_t(d)\), also, observe that \(Y_t(d)\) is independent of past \(Y\), given \(D_t\)). By Assumption 3, \(Y_t(d)\) is sub-gaussian, hence

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

Hence,

\[
E_\pi \left[ \prod_{s=1}^{t} \exp\{\lambda W_s(d)\} \right] \leq E_\pi \left[ E_{t-1} \left[ \exp\{0.5\nu\sigma(d)^2 1\{D_t = d\} \lambda^2 + \lambda W_{t-1}(d)\} \right] \prod_{s=1}^{t-2} \exp\{\lambda W_s(d)\} \right].
\]

Given Assumption 2, we can set \(\bar{\delta}(\cdot)(d) = (M + 1) \epsilon \frac{1}{M+1} + (1 - \epsilon (M + 1)) \tilde{\delta}(\cdot)(d)\) for any \(d \in \mathbb{D}\) where \(\tilde{\delta} := \frac{\delta - \epsilon}{1 - \epsilon (M + 1)}\). That is, with probability \(\epsilon\), \(D_t\) is IID and given this event, \(D_t\) is not equal to \(d\) with probability \(1 - 1/(M + 1)\). Hence, the event that \(D_t\) is IID and not equal to \(d\) has probability \(\epsilon M\). Thus, conditional on information until \(t - 1\),

\[
E_{t-1} \left[ \exp\{0.5\nu\sigma(d)^2 1\{D_t = d\} \lambda^2 + \lambda W_{t-1}(d)\} \right] \leq \left( \epsilon M + (1 - \epsilon M) \exp\{0.5\nu\sigma(d)^2 \lambda^2\} \right) E_{t-1} [\exp\{\lambda W_{t-1}(d)\}].
\]

Iterating, it follows that

\[
E_\pi \left[ \prod_{s=1}^{t} \exp\{\lambda W_s(d)\} \right] \leq \left( \epsilon + (1 - \bar{\epsilon}) \exp\{0.5\nu\sigma(d)^2 \lambda^2\} \right)^t = \exp\left( t \log \left( \epsilon M + (1 - \epsilon M) \exp\{0.5\nu\sigma(d)^2 \lambda^2\} \right) \right) =: \exp t F(\lambda, \epsilon).
\]

Let

\[
\lambda^*(a, \epsilon) := \max_{\lambda \geq 0} \{a \lambda - F(\lambda, \epsilon)\}
\]

exists and is positive. Therefore,

\[
P_\pi \left( t^{-1} \sum_{s=1}^{t} W_s(d) \geq a \right) \leq \exp\{-t \lambda^*(a, \epsilon)\}
\]

81
By analogous calculations, it is easy to show that

\[ P_\pi \left( \left| t^{-1} \sum_{s=1}^{t} W_s(d) \right| \geq a \right) \leq 2 \exp \{-t \lambda^* (a, \epsilon) \}. \]

Observe that \( \epsilon \mapsto F(\lambda, \epsilon) \) is decreasing and so \( \epsilon \mapsto \lambda^* (a, \epsilon) \) is increasing. It is also easy to see that \( a \mapsto F(\lambda, \epsilon) \) is also increasing. Moreover, \( \lambda^* (a, 0) = \max_{\lambda \geq 0} \{ a \lambda - 0.5 \nu \sigma (d) \lambda^2 \} = \frac{a^2}{2 \nu \sigma (d)^2} \).

Finally, note that \( \lambda^* (a, \epsilon) / a \geq \lambda - F(\lambda, \epsilon) / a \) for any \( \lambda \geq 0 \). In particular, \( \lambda := F^{-1} (a, \epsilon) \). As \( a \) diverges, this quantity also diverges and thus \( \lambda^* (a, \epsilon) / a \) also diverges.

**Proof for Lemma 20.** Observe that

\[ c (|C(d) - C| \geq 0) \leq 2 \exp \{ -C_\epsilon (0, \epsilon) \}. \]

We provide two upper bounds and then combine them. Regarding the first bound, note that

\[ P_\pi (m_t (d) - \theta (d) > a) \leq P_\pi \left( \max_{0 \leq s \leq t} (Y_s (d) - \theta (d)) > a \right) \leq \sum_{s=1}^{t} P_\pi ((Y_s (d) - \theta (d)) > a) \]

\[ \leq t \exp \{-a \lambda \} E [\exp \lambda |Y(d) - \theta(d)|], \ \forall \lambda > 0. \]

where the last line follows from the Markov inequality and the fact that \( Y_s (d) \) is IID. Since \( Y(d) \) is also sub-gaussian (Assumption 3), \( E [\exp \lambda (Y(d) - \theta(d))] \leq \exp \{ 0.5 \nu \sigma (d)^2 \lambda^2 \} \) and thus

\[ P_\pi ((m_t (d) - \theta (d)) > a) \leq \exp \left\{ \log t - 0.5 \frac{a^2}{\nu \sigma (d)^2} \right\}. \]

An analogous bound holds for \( P_\pi (- (m_t (d) - \theta (d)) > a) \), and thus

\[ P_\pi (|m_t (d) - \theta (d)| > a) \leq 2 \exp \left\{ \log t - 0.5 \frac{a^2}{\nu \sigma (d)^2} \right\}. \]

We now derive the other bound. For this, let, for any \( \eta > 0 \) and \( \delta > 0 \),

\[ F(t, \delta) := \left\{ |f_t (d) - t^{-1} \sum_{s=1}^{t} E_\pi [\delta (Z_s) (d)]| \leq \delta \right\} \]
and

\[ A(t, \eta) := \left\{ |t^{-1} \sum_{s=1}^{t} 1 \{ D_s = d \} (Y_s(d) - \theta(d))| \leq \eta \right\}. \]

Thus, for any \( \eta > 0 \) and \( \delta > 0 \),

\[
P_\pi (|m_t(d) - \theta(d)| > a) = P_\pi \left( |t^{-1} \sum_{s=1}^{t} 1 \{ D_s = d \} (Y_s(d) - f_t(d) - \theta(d))| > \alpha \cap A(t, \eta) \right) \\
\leq P_\pi \left( |t^{-1} \sum_{s=1}^{t} 1 \{ D_s = d \} (Y_s(d) - \theta(d))| \geq \alpha \right) + P_\pi \left( A(t, \eta)^C \right) \\
\leq P_\pi \left( \eta > a \left( t^{-1} \sum_{s=1}^{t} E_\pi [\delta(Z_s)(d)] - \delta \right) \right) + P_\pi \left( F(t, \delta)^C \right) + P_\pi \left( A(t, \eta)^C \right).
\]

Letting \( \eta = \delta = \frac{t^{-1} \sum_{s=1}^{t} E_\pi [\delta(Z_s)(d)]}{1 + \alpha} \) the first term in the RHS is naught. By Lemmas 18 and 19, it follows that

\[
P_\pi (|m_t(d) - \theta(d)| > a) \leq 2 \exp \left\{ -0.5 t \nu^{-1} \left( \frac{t^{-1} \sum_{s=1}^{t} E_\pi [\delta(Z_s)(d)]}{1 + \alpha} \right)^2 \right\} \\
+ \exp \left\{ \min \{ (0.5 \sigma(\epsilon))^2, 1 \} t \left( \frac{t^{-1} \sum_{s=1}^{t} E_\pi [\delta(Z_s)(d)]}{1 + \alpha} \right)^2 \right\}.
\]

\[ \Box \]

E Appendix for Section 3.2.2

We now show Proposition 4. To do this, for each \( d \in \mathbb{D} \), we define \( \eta^* : \mathbb{N} \times [0, 1] \times \mathbb{R}_+ \to \mathbb{R}_+ \cup \{ +\infty \} \) as follows: For any \( (t, \epsilon, \Delta) \in \mathbb{N} \times \mathbb{R}_+ \times [0, 1] \times \mathbb{R}_+ \) and \( \eta \geq 0 \), if \( F^\alpha_{\eta^*}(t, \nu, \eta, \epsilon) \leq 0.5 \Delta \) for all \( \eta \), then
we choose $\eta^*_d(t, \epsilon, \Delta) = +\infty$, otherwise\footnote{If the set in the "max" is empty, then we set $\eta^*(t, \epsilon, \Delta) = 0.$}

$$
\eta^*_d(t, \epsilon, \Delta) := \max \left\{ \eta: \sum_{o=0}^L F^o_d(t, \gamma_t, \eta, \epsilon) \leq 0.50\Delta \text{ and } \eta \leq 0.99 \epsilon \right\},
$$

where for each $o \in \{0, \ldots, L\}$ and $d \in \mathbb{D}$, $F^o_d : \mathbb{N} \times \mathbb{R}_+ \times \mathbb{R}_+ \times [0, 1] \to \mathbb{R}_+$ is defined as

$$
\max_{x \in [\epsilon, 1]} \left\{ 1\{(1)^{d=M} \bar{\xi}^o(d) \leq 0\} \left( \frac{(1)^{d=M} \bar{\xi}^o(d) v^o_0(d) / t}{x + \eta + \nu^o_0(d) / t} \bar{\alpha}^o(\eta, \gamma_t / (x - \eta), |\bar{\xi}^o(d)|, v_0(d), x) \right) + 1\{(1)^{d=M} \bar{\xi}^o(d) > 0\} \left( \frac{(1)^{d=M} \bar{\xi}^o(d) v^o_0(d) / t}{x - \eta + \nu^o_0(d) / t} \bar{\alpha}^o(\eta, \gamma_t / (x - \eta), |\bar{\xi}^o(d)|, v_0(d), x) \right) \right\}
$$

where $\bar{\alpha}^o$ and $\bar{\alpha}^o$ are defined in Lemma 4, and $(\gamma_t)_t$ is as in the statement of Proposition 4.

\textbf{Proof of Proposition 4.} Recall that

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

Also, the probability of making a mistake associated to this stopping rule can be bounded by

$$
\sum_{t=B}^T P_x \left( \max_{d \neq M} \sum_{o=0}^L \alpha^o_t(d) \zeta^o_t(d) - \sum_{o=0}^L \alpha^o_t(M) \zeta^o_t(M) > 0 \cap \tau = t \right),
$$

where $\{\max_{d \neq M} \sum_{o=0}^L \alpha^o_t(d) \zeta^o_t(d) - \sum_{o=0}^L \alpha^o_t(M) \zeta^o_t(M) > 0 \cap \tau = t\}$ 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

$$
\max_d \left\{ \min_{m \neq d} \left( \zeta^o_t(d) - \zeta^o_t(m) - c_t(\gamma_t, d, m) \right) \right\} > 0
$$

which in turn implies that

$$
\max_d \left\{ \zeta^o_t(d) - \zeta^o_t(M) - c_t(\gamma_t, d, M) \right\} > 0.
$$
Thus, the event \( \{ \max_{d \neq M} \{ \xi_t^a(d) - \zeta_t^a(M) \} > 0 \cap \tau = t \} \) implies
\[
\{ \max_{d \neq M} \{ \xi_t^a(d) - \zeta_t^a(M) - c_i(\gamma_t, d, M) \} > 0 \}.
\]

Suppose the max is achieved by \( d(t) \neq M \), then the above expression is equivalent to \( \tilde{\xi}_t^a(d(t)) - \tilde{\zeta}_t^a(M) - c_i(\gamma_t, d, M) > \theta(M) - \theta(d(t)) \). Since \( \theta(M) - \theta(d(t)) \geq \Delta \) — recall, \( \Delta := \min_d \theta(M) - \theta(d) \) —, it follows that
\[
\{ \max_{d \neq M} \{ \tilde{\xi}_t^a(d) - \tilde{\zeta}_t^a(M) - c_i(\gamma_t, d, M) \} > \Delta \}.
\]

Observe that
\[
c_i(\gamma_t, d, M) =: c_i(\gamma_t, d) + c_i(\gamma_t, M)
\]

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

Thus, the event \( \{ \max_{d \neq M} \{ \tilde{\xi}_t^a(d) - \tilde{\zeta}_t^a(M) - c_i(\gamma_t, d, M) \} > \Delta \} \), is included in the event
\[
\cup_{d \neq M} \{ \{ \tilde{\xi}_t^a(d) - \tilde{\zeta}_t^a(M) - c_i(\gamma_t, d, M) \} > \Delta \} \cap \{ \tilde{\xi}_t^a(M) + c_i(\gamma_t, M) \) \geq -0.5\Delta \} \cup \{ \tilde{\zeta}_t^a(M) + c_i(\gamma_t, M) < -0.5\Delta \}
\]
\[
= \cup_{d \neq M} \{ \tilde{\xi}_t^a(d) > c_i(\gamma_t, d) + 0.5\Delta \} \cup \{ \tilde{\zeta}_t^a(M) < -(c_i(\gamma_t, M) + 0.5\Delta) \}
\]

By the definition of \( c_i \), it follows that for any \( d \in \{0, ..., M\} \),
\[
\{ \tilde{\xi}_t^a(d) > c_i(\gamma_t, d) + 0.5\Delta \} \subseteq \{ \tilde{\xi}_t^a(d) > c_i(\gamma_t, d) + 0.5\Delta \} \cap \mathcal{F}(\gamma_t, d) \cup \mathcal{F}(\gamma_t, d) \cup \mathcal{F}(\gamma_t, d) \cup \mathcal{F}(\gamma_t, d) \cup \mathcal{E}(\eta, d) \cup \mathcal{E}(\eta, d).
\]

where \( \mathcal{F}(\gamma, d) := \{ |f_t(d)| \leq \gamma \} \) and \( \mathcal{E}(\eta, d) := \{ |f_t(d) - e_t(d)| \leq \eta \} \) for any \( t \in \mathbb{N} \) and any \( \eta, \gamma > 0 \).

Similarly,
\[
\{ \tilde{\zeta}_t^a(M) > -(c_i(\gamma_t, M) + 0.5\Delta) \} \subseteq \left\{ \sum_{o=0}^{L} \frac{\alpha_t^o(M)(-\tilde{\xi}_t^a(M))}{f_t(M) + \nu_t^o(M) / t} > 0.5\Delta \right\} \cap \mathcal{F}(\gamma_t, M) \cup \mathcal{E}(\eta, M) \cup \mathcal{E}(\eta, M) \cup \mathcal{F}(\gamma_t, M) \cup \mathcal{F}(\gamma_t, M) \cup \mathcal{E}(\eta, M) \cup \mathcal{E}(\eta, M).
\]
By Lemma 4, under $\mathcal{F}(\gamma_t, d) \cap \mathcal{E}(\eta, d)$, it follows that for any $d \in \mathbb{D}$,
\[
\left( -1 \right)^{1(d=M)} \frac{\bar{\xi}_0^\circ (d) \nu_0^\circ (d) / t}{f_t^\circ (d) + \nu_0^\circ (d) / t} \alpha_t^\circ (d)
\leq 1 \left\{ \left( -1 \right)^{1(d=M)} \bar{\xi}_0^\circ (d) \leq 0 \right\} \left\{ \left( -1 \right)^{1(d=M)} \frac{\bar{\xi}_0^\circ (d) \nu_0^\circ (d) / t}{e_t(d) + \eta + \nu_0^\circ (d) / t} \alpha^\circ (\eta, \gamma_t / (e_t(d) - \eta), |\bar{\xi}_0(d)|), \nu_0(d), e_t(d)) \right\}
+ 1 \left\{ \left( -1 \right)^{1(d=M)} \bar{\xi}_0^\circ (d) > 0 \right\} \left\{ \left( -1 \right)^{1(d=M)} \frac{\bar{\xi}_0^\circ (d) \nu_0^\circ (d) / t}{e_t(d) - \eta + \nu_0^\circ (d) / t} \alpha^\circ (\eta, \gamma_t / (x - \eta), |\bar{\xi}_0(d)|), \nu_0(d), e_t(d)) \right\}
\leq \max_{x \in [0, 1]} \left\{ 1 \left\{ \left( -1 \right)^{1(d=M)} \bar{\xi}_0^\circ (d) \leq 0 \right\} \left\{ \left( -1 \right)^{1(d=M)} \frac{\bar{\xi}_0^\circ (d) \nu_0^\circ (d) / t}{x + \eta + \nu_0^\circ (d) / t} \alpha^\circ (\eta, \gamma_t / (x - \eta), |\bar{\xi}_0(d)|), \nu_0(d), x) \right\}
+ 1 \left\{ \left( -1 \right)^{1(d=M)} \bar{\xi}_0^\circ (d) > 0 \right\} \left\{ \left( -1 \right)^{1(d=M)} \frac{\bar{\xi}_0^\circ (d) \nu_0^\circ (d) / t}{x - \eta + \nu_0^\circ (d) / t} \alpha^\circ (\eta, \gamma_t / (x - \eta), |\bar{\xi}_0(d)|), \nu_0(d), x) \right\} \right\}
=: F_d^\circ (t, \gamma_t, \eta, \epsilon).
\]

Observe that $F_d^\circ$ is non-random. Also, it follows that for any $d \in \mathbb{D}$,
\[
\{ \bar{\xi}_t^\circ (d) > -(c_t(\gamma_t, d) + 0.5\Delta) \} \subseteq \mathcal{U}_d(t, \gamma_t, \eta, \epsilon, \Delta) \cup \mathcal{F}_d(t, \gamma_t, \eta, \epsilon, \Delta) \cup \mathcal{E}_t(\eta, d) \cap \mathcal{F}_d(t, \gamma_t, \eta, \epsilon, \Delta)
\]
where $\mathcal{U}_d(t, \gamma_t, \eta, \epsilon, \Delta) := \{ \sum_{\alpha=0}^L F_d^\circ (t, \gamma_t, \eta, \epsilon) > 0.5\Delta \}$. It thus follows that
\[
\sum_{t=B}^T P_\pi \left( \max_{d \neq M} \{ \bar{\xi}_t^\circ (d) - \bar{\xi}_t^\circ (M) - c_t(\gamma_t, d, M) \} > \Delta \right) \leq \sum_{t=B}^T P_\pi \left( \cup_{d=0}^L \mathcal{U}_d(t, \gamma_t, \eta, \epsilon, \Delta) \cup \cup_{d=0}^L \mathcal{F}_d(t, \gamma_t, \eta, \epsilon, \Delta) \cup \cup_{d=0}^L \mathcal{E}_t(\eta, d) \cap \mathcal{F}_d(t, \gamma_t, \eta, \epsilon, \Delta) \right)
\leq \sum_{t=B}^T \left\{ 1 \left\{ \cup_{d=0}^L \mathcal{U}_d(t, \gamma_t, \eta, \epsilon, \Delta) \right\} + P_\pi \left( \cup_{d=0}^L \mathcal{F}_d(t, \gamma_t, \eta, \epsilon, \Delta) \right) \right\}
\]
where the second inequality follows from the union bound.

We now choose $\eta$ as follows. If $F_d^\circ (t, \gamma_t, \eta, \epsilon) \leq 0.5\Delta$ for all $\eta$, then we choose $\eta^\circ (t, \epsilon, \Delta) = +\infty$, otherwise
\[
\eta^\circ (t, \epsilon, \Delta) := \max \left\{ \eta : \sum_{\alpha=0}^L F_d^\circ (t, \gamma_t, \eta, \epsilon) \leq 0.50\Delta \text{ and } \eta \leq 0.99 \epsilon \right\}
\]
If the set is empty, then $\eta^\circ (t, \epsilon, \Delta) = 0$. 

86
If \( \eta_d^*(t, \epsilon, \Delta) = 0 \), the expression 15 yields the trivial bound of 1. The expression in the proposition also implies an upper bound greater than 1 (since \( \eta_d^*(t, \epsilon, \Delta) = 0 \)). Thus the proposition is proven. We now study the case if \( \eta_d^*(t, \epsilon, \Delta) > 0 \) which is more involved. Under this choice of \( \eta \), it follows that

\[
\sum_{t=B}^{T} P_{\pi} \left( \max_{d \neq M} \left\{ \xi_{t}^d(d) - \bar{\xi}_{t}^d(M) - c_t(\gamma_t, d, M) \right\} > \Delta \right) \leq \sum_{t=B}^{T} P_{\pi} \left( \cup_{\gamma_t} (\gamma_t, d) \right) \\
+ \sum_{t=B}^{T} P_{\pi} \left( \cup_{\gamma_t} (\eta_d^*(t, \epsilon, \Delta), d) \right).
\]

By Lemmas 19 and 18, it follows that

\[
\sum_{t=B}^{T} P_{\pi} \left( \max_{d \neq M} \left\{ \xi_{t}^d(d) - \bar{\xi}_{t}^d(M) - c_t(\gamma_t, d, M) \right\} > \Delta \right) \leq \sum_{t=B}^{M} \sum_{d=0}^{T} \left( 2e^{-0.5I_t^{\eta_0}(d) + e^{-\frac{1}{\log_t(\gamma'(t, \epsilon, \Delta))}(C(\epsilon))}} \right).
\]

We conclude the proof by showing some properties of \( \eta_d^* \). First, \( t \mapsto \eta_d^*(t, \epsilon, \Delta) \) is non-decreasing. To show this, it suffices to show \( t \mapsto \sum_{o=0}^{L} F_d^o(t, \gamma_t, \eta, \epsilon) \) is non-increasing and \( \eta \mapsto \sum_{o=0}^{L} F_d^o(t, \gamma_t, \eta, \epsilon) \) is non-decreasing. We show the latter below, here we show the former. To do this it suffices to show that \( t \mapsto F_d^o(t, \gamma_t, \eta, \epsilon) \) is non-decreasing for each \( \eta \). To show this take any \( x \in [\epsilon, 1] \) and first consider the case \( (-1)^{1(d=M)} \xi_{t}^0(d) > 0 \); for this case, \( \frac{(-1)^{1(d=M)} \xi_{t}^0(d) v_0(d)/t}{x-\eta+\nu_0(d)/t} \) is decreasing in \( t \) and positive. By Lemma 5 and the definition of \( \alpha^o \) and \( \bar{\alpha}^o \) is increasing in \( \gamma_t \), moreover, \( t \mapsto \gamma_t \) is increasing in the relevant domain. Since \( \bar{\alpha}^o \) is positive as well, these results imply that \( t \mapsto F_d^o(t, \gamma_t, \eta, \epsilon) \) is non-decreasing for each \( \eta \) in this case. Now consider the case \( (-1)^{1(d=M)} \xi_{t}^0(d) \leq 0 \); in this case \( \frac{(-1)^{1(d=M)} \xi_{t}^0(d) v_0(d)/t}{x-\eta+\nu_0(d)/t} \) is increasing in \( t \) and negative. By Lemma 5 and the definition of \( \alpha^o \), \( \bar{\alpha}^o \) is non-increasing in \( \gamma_t \). Since \( \bar{\alpha}^o \) is positive as well, these results imply that \( t \mapsto F_d^o(t, \gamma_t, \eta, \epsilon) \) is also non-decreasing for each \( \eta \) in this case and thus the desired result holds.

Second, \( \Delta \mapsto \eta_d^*(t, \epsilon, \Delta) \) is non-decreasing. To show this is sufficient to show that \( \eta \mapsto \sum_{o=0}^{L} F_d^o(t, \gamma_t, \eta, \epsilon) \) is non-decreasing; which, in turn, it suffices to show that \( \eta \mapsto F_d^o(t, \gamma_t, \eta, \epsilon) \) is non-decreasing for each \( o \). To show this take any \( x \in [\epsilon, 1] \) and first consider the case \( (-1)^{1(d=M)} \xi_{t}^0(d) > 0 \); for this case, \( \frac{(-1)^{1(d=M)} \xi_{t}^0(d) v_0(d)/t}{x-\eta+\nu_0(d)/t} \) is increasing in \( \eta \) and positive and by Lemma 5 and the fact that \( \eta \mapsto \gamma_t \) is increasing in \( \eta \) (for all \( \eta \leq \gamma_t \leq x \)), \( \bar{\alpha}^o(\eta, \gamma_t, (\gamma_t - \eta), |\xi_{t}^0(d)|, \nu_0(d), x) \) increasing in \( \eta \) and positive. Thus, the product is also increasing and this shows that if \( (-1)^{1(d=M)} \xi_{t}^0(d) > 0 \), \( \eta \mapsto F^o(t, \gamma_t, \eta, \epsilon) \) is increasing. If \( (-1)^{1(d=M)} \xi_{t}^0(d) \leq 0 \), \( \frac{(-1)^{1(d=M)} \xi_{t}^0(d) v_0(d)/t}{x+\eta+\nu_0(d)/t} \) is increasing in \( \eta \) and negative, and \( \bar{\alpha}^o(\eta, \gamma_t, (\gamma_t - \eta), |\xi_{t}^0(d)|, \nu_0(d), x) \) is decreasing in \( \eta \) and positive. Thus, the product
is increasing in \(\eta\) and thus the desired result follows.

Third, \(\epsilon \mapsto \eta^*_{d}(t,\epsilon,\Delta)\) is non-decreasing. To show this, we show that for any \(\epsilon \leq \epsilon'\), the set of feasible \(\eta\)'s indexed by \(\epsilon\) is contained in that of \(\epsilon'\). From the definition of \(F^0\) is easy to see that \(\epsilon \mapsto F^0_{d}(t,\gamma,t,\epsilon,\Delta)\) is non-increasing; this fact and the fact that \(\eta \mapsto \sum_{o=0}^{L} F^o_{d}(t,\gamma_t,\eta,\epsilon)\) is non-decreasing imply that \(\{\eta: \sum_{o=0}^{L} F^o_{d}(t,\gamma_t,\eta,\epsilon) \leq 0.5\Delta \text{ and } \eta \leq 0.99 \epsilon\}\) is included in \(\{\eta: \sum_{o=0}^{L} F^o_{d}(t,\gamma_t,\eta',\epsilon) \leq 0.5\Delta \text{ and } \eta \leq 0.99 \epsilon'\}\).

\[\square\]

**Proof of Corollary 2.** Let \(\eta^\text{oracle}_d(t,\epsilon,\Delta)\) defined as in Proposition 4 and let

\[\eta^\text{oracle}_d(t,\epsilon,\Delta) := \max \left\{ \eta: \frac{|\tilde{\xi}^0_0(d)\xi^0_0(d)/t \leq 0.5\Delta \text{ and } \eta \leq 0.99 \epsilon} {\epsilon - \eta + \gamma^0_0(d)/t} \right\}.\]

Essentially, this quantity analogous to \(\eta^*_{d}(t,\gamma_t,\epsilon,\Delta)\) but putting all the weight to model \(o = 0\).

Suppose that (we show this at the end of the proof) given \(\Delta\) and \(\epsilon\), for any \(\delta > 0\), there exists a \(M_{\delta}\) such that for all \(|\tilde{\xi}^0_0(d)| \geq M_{\delta}\),

\[|\eta^\text{oracle}_d(t,\epsilon,\Delta) - \eta^*_{d}(t,\gamma_t,\epsilon,\Delta)| \leq \delta\]

and \(\eta^\text{oracle}_d(t,\epsilon,\Delta) > 0\) (assuming \(\epsilon,\Delta > 0\)) for all \(t \in \{B,\ldots,T\}\). Then, for any \(\delta < \eta^\text{oracle}_d(t,\epsilon,\Delta)\) and for all \(|\tilde{\xi}^0_0(d)| \geq M_{\delta}\),

\[P_{\pi} \left( \max_{d \neq M} \{\xi^{o}_T(d) - \xi^{a}_T(M)\} > 0 \right) \leq \sum_{d=0}^{M} \sum_{t=B}^{T} (2e^{-0.5t \frac{y(t)^2}{v^{2}(d)/\epsilon}} + e^{-\frac{t^{2}}{4\epsilon}} (\eta^\text{oracle}_d(t,\epsilon,\Delta) - \delta \epsilon)^2 C(\epsilon) \)

\[\leq \sum_{d=0}^{M} \sum_{t=B}^{T} (2e^{-0.5t \frac{y(t)^2}{v^{2}(d)/\epsilon}} + e^{-\frac{t^{2}}{4\epsilon}} (\eta^\text{oracle}_d(t,\epsilon,\Delta) + \delta \epsilon)^2 + \delta(\delta - 2\epsilon)) C(\epsilon) \]

where the second inequality follows because for any \(0 \leq a \leq b \leq \epsilon\), \((a - b)^2 \geq a^2 + b^2 - 2ab \geq b^2 + a(a - 2 \epsilon)\).

Now, for any \(\epsilon > 0\), choose \(\delta > 0\) such that \(e^{-\frac{t^{2}}{4\epsilon}} \delta(\delta - 2\epsilon) C(\epsilon) \leq 1 + \epsilon\). Then, for all \(|\tilde{\xi}^0_0(d)| \geq M_{\delta}\),

\[P_{\pi} \left( \max_{d \neq M} \{\xi^{a}_T(d) - \xi^{a}_T(M)\} > 0 \right) \leq (1 + \epsilon) \sum_{d=0}^{M} \sum_{t=B}^{T} (2e^{-0.5t \frac{y(t)^2}{v^{2}(d)/\epsilon}} + e^{-\frac{t^{2}}{4\epsilon}} (\eta^\text{oracle}_d(t,\epsilon,\Delta) + \delta \epsilon)^2 C(\epsilon) \)

and thus the desired result is proven.
We now show that for any $\delta > 0$, there exists a $M_0$ such that for all $|\tilde{\xi}_0^0(d)| \geq M$,

$$|\eta_d^{\text{oracle}}(t, \epsilon, \Delta) - \eta_d^*(t, \gamma_t, \epsilon, \Delta)| \leq \delta$$

$\eta_d^{\text{oracle}}(t, \epsilon, \Delta) > 0$ (provided $\epsilon, \Delta > 0$) for all $t \in \{B, \ldots, T\}$. To do this we invoke the theorem of the Maximum. In particular, we need to show that the correspondence

$$(|\tilde{\xi}_0(d)|) \mapsto C(|\tilde{\xi}_0(d)|) := \left\{ \eta \leq 0.99 \epsilon : \max_d \max_{x \in [\epsilon, 1]} F(t, \gamma, \eta, |\tilde{\xi}_0(d)|, \nu_0(d), x, \epsilon) \leq 0.50 \Delta \right\}$$

where $F$ defined in the proof of Proposition 4, is continuous and compact-valued and that $C(|\tilde{\xi}_0(d)|)$ converges to $\left\{ \eta \leq 0.99 \epsilon : \max_d \frac{|\tilde{\xi}_0^0(d)|}{\epsilon - \eta + \nu_0^0(d)/t} \leq 0.5 \Delta \right\}$ as $|\tilde{\xi}_0^0(d)|$ diverges.

By definition of $F$, we can see that is continuous in its arguments over $\{\eta : \eta \leq 0.99 \epsilon\}$. Applying the Theorem of the Maximum, this implies that $\max_d \max_{x \in [\epsilon, 1]} F(t, \gamma, \eta, |\tilde{\xi}_0(d)|, \nu_0(d), x, \epsilon)$ is also continuous and thus the correspondence $C$ is continuous too. In addition, $C$ is compact-valued. We now characterize the limit of $C$ as $|\tilde{\xi}_0^0(d)|$ diverges. In particular we show that for any $\delta > 0$, there exists a $M$ and a $K$ such that for all $|\tilde{\xi}_0^0(d)| \geq M$,

$$\sup_{(\eta, \gamma) \in [0.099 \epsilon] \times (0, K]} \left| \max_{x \in [\epsilon, 1]} F(t, \gamma, \eta, |\tilde{\xi}_0(d)|, \nu_0(d), x, \epsilon) - \frac{|\tilde{\xi}_0^0(d)|}{\epsilon - \eta + \nu_0^0(d)/t} \right| < \delta.$$  

By Lemma 5, $\lim_{|\tilde{\xi}_0^0(d)| \to \infty} \overline{\alpha}^0(\eta, \gamma/(\epsilon - \eta), |\tilde{\xi}_0(d)|, \nu_0(d), e_t(d)) = 0$ for any $\delta > 0$. As $e_t(d) \in [\epsilon, 1]$, the quantities $\overline{\alpha}^0_t$ and $\gamma^2_t$ that define $\overline{\alpha}$ are uniformly bounded and uniformly bounded away from 0 as functions of $e_t(d)$. Thus, by construction of $\overline{\alpha}^0$, the convergence above also holds uniformly with $e_t(d) \in [\epsilon, 1]$, i.e., $\lim_{|\tilde{\xi}_0^0(d)| \to \infty} \max_{x \in [\epsilon, 1]} \overline{\alpha}^0(\eta, \gamma/(\epsilon - \eta), |\tilde{\xi}_0(d)|, \nu_0(d), x) = 0$. By construction of $\overline{\alpha}^0$, it is easy to see that the convergence is exponentially fast on $|\tilde{\xi}_0^0(d)|$ and thus

$$\lim_{|\tilde{\xi}_0^0(d)| \to \infty} \max_{x \in [\epsilon, 1]} \overline{\alpha}^0(\eta, \gamma/(\epsilon - \eta), |\tilde{\xi}_0(d)|, \nu_0(d), x) \frac{|\tilde{\xi}_0^0(d)|}{\epsilon - \eta + \nu_0^0(d)/t} = 0$$

for every $\delta > 0$, and

$$\lim_{|\tilde{\xi}_0^0(d)| \to \infty} \max_{x \in [\epsilon, 1]} \overline{\alpha}^0(\eta, \gamma/(\epsilon - \eta), |\tilde{\xi}_0(d)|, \nu_0(d), x) = 1.$$  

It is easy to see that by definition of $\overline{\alpha}^0$ convergence holds uniformly over $\gamma \in [0, K]$ for $K < \infty$. In
addition, as \( \eta \leq 0.99 \epsilon \) and \( e_\ell(d) \geq \epsilon \), it follows that \( \bar{\sigma} \) and \( \bar{\sigma} \) in the definition of \( \bar{\sigma} \) are uniformly bounded as functions of \( \eta \) and thus the convergence also holds uniformly over \( \eta \in [0,0.99 \epsilon] \). Therefore,

\[
\sup_{(\eta,\gamma) \in [0,0.99\epsilon] \times [0,\bar{K}]} \max_{x \in [\epsilon,1]} F(t,\gamma,\eta,|\tilde{\xi}_0(d)|,\nu_0(d),x,\epsilon) - \frac{|\tilde{\xi}_0^0(d)|\nu_0^0(d)/t}{\epsilon - \eta + \nu_0^0(d)/t} < \delta.
\]

Therefore, \( C(|\tilde{\xi}_0(d)|) \) converges to \( \{ \eta \leq 0.99 \epsilon : \max_d \frac{|\tilde{\xi}_0^0(d)|\nu_0^0(d)/t}{\epsilon - \eta + \nu_0^0(d)/t} \leq 0.45\Delta \} \) as \( |\tilde{\xi}_0-0(d)| \) diverges. Hence, by the Theorem of the maximum,

\[
\lim_{|\tilde{\xi}_0^0(d)| \to \infty} |\eta_d^{oracle}(t,\epsilon,\Delta) - \eta_d^*(t,\gamma_t,\epsilon,\Delta)| = 0.
\]

Since \( t \in \{1,...,T\} \) and \( T < \infty \), the convergence is uniform in \( t \). Finally, \( \frac{|\tilde{\xi}_0^0(d)|\nu_0^0(d)/t}{\epsilon + \nu_0^0(d)/t} < 0.5\Delta \) by assumption; this implies that \( \eta_d^{oracle}(t,\epsilon,\Delta) > 0 \) for any \( t \in \{B,...,T\} \) because \( \eta^0 \) is chosen to be maximal.

\[\square\]

**Proof of Corollary 3.** We do the proof for where the expression for \( \gamma \) holds with equality. We do this because if the desired bound holds for this case, it will hold for any \( \gamma \) that is greater. By Proposition 4

\[
P_\pi \left( \max_{d \neq M} \{ \xi^\beta_r(d) - \zeta^\beta_r(M) \} > 0 \right) \leq \sum_{d=0}^M \sum_{t=B}^T (2e^{-0.5t \frac{(\gamma)^2}{\nu\sigma(d)^2}} + e^{-\frac{t}{\log(t)}}(rt_d(t,\epsilon,\Delta))^2C(\epsilon)). \tag{16}
\]

By our choice of \( \gamma \) and the fact that \( \log B \geq 2\nu\sigma(d)^2 \), the first term in the RHS is less or equal than \( 2t^{-A} \). We now need to check that for all \( t \geq B(\beta) \), \( e^{-\frac{t}{\log(t)}}(rt_d(t,\epsilon,\Delta))^2C(\epsilon) \leq t^{-A} \rightarrow A \leq (\log t)^2 \). Since \( t \mapsto (\log t)^2 \) is decreasing (as \( \log B \geq 2 \)) and \( \eta^*(t,\epsilon,\Delta) \geq \eta^*(1,\epsilon,\Delta) \), it thus suffices to check that \( (\eta^*(t,\epsilon,\Delta))^2C(\epsilon) \leq (\log B)^2 \) which holds by assumption. \[\square\]

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 parameter \( \epsilon \). These results are of particular interest for assessing the accuracy of the PM’s recommendation once the experiment is over.
Proposition 6. For any $\epsilon \geq 0$ and any $t \geq e^{\max\{\epsilon \max(\Gamma_+^{-1}(0.5(\epsilon+\Delta)),\Gamma_-^{-1}(0.5(\epsilon+\Delta)),0)\mathbf{B}(\epsilon),2\nu \max \sigma(d)^2\}}$, 24

$$P_\pi(\exists d \neq M : \zeta_t^a(d) - \zeta_t^a(M) > \epsilon) \leq 3 \sum_{d=0}^{M-1} \left( e^{-\frac{3}{\log(\max\{\Gamma^{-1}(0.5(\epsilon+\Delta),\bar{\nu}_d(d),\nu_t(d),0\})^2\})} + 3e^{-\frac{3}{\log(\max\{\Gamma^{-1}(0.5(\epsilon+\Delta),-\bar{\nu}_0(M),\nu_t^m(M),\nu_t(M),0\})^2\})} \right).$$

Proof of Proposition 6. For any $\gamma \geq 0$ and any $a \in \{0,\ldots,M\}$, let

$$S(t,a,\gamma) := \{|\bar{J}_t(a)| \leq \gamma\},$$

and

$$R(t,a,\gamma) := \{|f_t(a) - E[f_t(a)]| \leq \gamma\},$$

where $E[f_t(a)] = t^{-1} \sum_{s=1}^{t} E[\delta(Z_s)(a)] = e_t(a)$ and the expectation is constructed using the initial probability $\pi$ and the transition probability $Q$.

Observe that

$$P_\pi(\exists d \neq M : \zeta_t^a(d) - \zeta_t^a(M) > \epsilon) = P_\pi(\exists d \neq M : \bar{\zeta}_t^a(d) - \bar{\zeta}_t^a(M) > \epsilon + \Delta) \leq P_\pi(\exists d \neq M : \bar{\zeta}_t^a(d) - \bar{\zeta}_t^a(M) > \epsilon + \Delta \cap \bar{\zeta}_t^a(M) \geq -0.5(\epsilon + \Delta)) + P_\pi(\bar{\zeta}_t^a(M) < -0.5(\epsilon + \Delta))$$

$$\leq \sum_{d=0}^{M-1} P_\pi(\bar{\zeta}_t^a(d) > 0.5(\epsilon + \Delta)) + P_\pi(\bar{\zeta}_t^a(M) < -0.5(\epsilon + \Delta))$$

$$\leq \sum_{d=0}^{M-1} P_\pi(\bar{\zeta}_t^a(d) > 0.5(\epsilon + \Delta)) + P_\pi(\bar{\zeta}_t^a(M) < -0.5(\epsilon + \Delta)) + \sum_{d=0}^{M-1} \{P_\pi(R(t,d,\gamma)C) + P_\pi(S(t,d,\gamma)C)\}$$

$$+ P_\pi(-\bar{\zeta}_t^a(M) > 0.5(\epsilon + \Delta) \mid R(t,M,\eta) \cap S(t,M,\eta))$$

$$+ \{P_\pi(R(t,M,\eta)C) + P_\pi(S(t,M,\eta)C)\}$$

where $\bar{\zeta}_t^a(d) := \zeta_t^a(d) - \theta(d)$.

---

$^{24}$\(\Gamma_+ \text{ and } \Gamma_- \text{ are defined in Lemma 17.}\)
By Lemmas 19 and 18, and the fact that $t \geq \exp 2\nu_0(d)^2$ and $t \geq \exp 4\gamma B(\epsilon),$ 

$$P_\pi\left(S(t, d, \gamma)^C\right) + P_\pi\left(R(t, d, \gamma)^C\right) \leq 3 \exp\left\{-\frac{t}{\log t} \gamma^2 C(\epsilon)\right\}. $$

By Lemma 17(2), under the sets $R(t, d, \gamma) \cap S(t, d, \gamma),$ 

$$\tilde{\zeta}_t^\alpha(d) \leq \Gamma_+(\gamma) := \Gamma(\gamma, \xi_0(d), \nu_0(d), e_t(d)).$$

Hence, choosing $\gamma = \max\{\Gamma_+^{-1}(0.5(\epsilon + \Delta)), 0\},$ it follows that, if $\gamma > 0,$ then 

$$P_\pi\left(\tilde{\zeta}_t^\alpha(d) > 0.5(\epsilon + \Delta) \mid R(t, d, \gamma) \cap S(t, d, \gamma)\right) = 0$$

(if $\gamma = 0$ then the bound is trivial as probabilities are bounded by 1). Thus,

$$P_\pi\left(\tilde{\zeta}_t^\alpha(d) > 0.5(\epsilon + \Delta)\right) \leq 3 \exp\left\{-\frac{t}{\log t} \left(\max\{\Gamma_+^{-1}(0.5(\epsilon + \Delta)), 0\}\right)^2 C(\epsilon)\right\},$$

and it follows that

$$P_\pi\left(\exists d \neq M : \tilde{\zeta}_t^\alpha(d) - \zeta_t^\alpha(M) > \epsilon\right) \leq 3 \sum_{d=0}^{M-1} \exp\left\{-\frac{t}{\log t} \left(\max\{\Gamma_+^{-1}(0.5(\epsilon + \Delta)), \xi_0(d), \nu_0(d), e_t(d)\}, 0\right)^2 C(\epsilon)\right\}$$

$$+ P_\pi\left(-\tilde{\zeta}_t^\alpha(M) > 0.5(\epsilon + \Delta) \mid R(t, M, \eta) \cap S(t, M, \eta)\right)$$

$$+ \left\{P_\pi\left(R(t, M, \eta)^C\right) + P_\pi\left(S(t, M, \eta)^C\right)\right\}.$$ 

By Lemma 17(3), $-\tilde{\zeta}_t(M) \leq \Gamma_-(\eta) := \Gamma(\eta, -\xi_0(d), \nu_0(d), e_t(d)).$ Hence, choosing $\eta = \max\{\Gamma_+^{-1}(0.5(\epsilon + \Delta)), 0\}$ and by analogous arguments to those above, it follows that

$$P_\pi\left(\exists d \neq M : \zeta_t(d) - \zeta_t(M) > \epsilon\right) \leq 3 \sum_{d=0}^{M-1} \exp\left\{-\frac{t}{\log t} \left(\max\{\Gamma_-^{-1}(0.5(\epsilon + \Delta)), -\xi_0(d), \nu_0(d), e_t(d)\}, 0\right)^2 C(\epsilon)\right\}$$

$$+ 3 \exp\left\{-\frac{t}{\log t} \left(\max\{\Gamma_-^{-1}(0.5(\epsilon + \Delta)), -\xi_0(M), \nu_0(M), e_t(M)\}, 0\right)^2 C(\epsilon)\right\}. $$

□

As $C(.)$ is non-decreasing, the probability of making a mistake decreases with the $\epsilon$ (other things equal), illustrating the fact that as $\epsilon$ increases, the data becomes “less correlated” and thus more
informative thereby reducing the probability of making a mistake.

F Appendix for Section 3.2.3

The proof of Proposition 5 used the following lemma that establishes the rate at which \( t^{-1} \sum_{s=1}^{t} Y_s \) concentrates around its average \( t^{-1} \sum_{s=1}^{t} \sum_{d=0}^{M} \theta(d) E_\pi[\delta(Z_s)(d)] = \sum_{d=0}^{M} e_t(d) \theta(d) \). Its proof is relegated towards the end of the section.

**Lemma 21.** For any \( \gamma > 0 \), any \( d \in \{0, \ldots, M\} \) and for any \( t \geq e^{4eB(\epsilon)} \) where \( \epsilon = \max_d (M + 1) \sqrt{\left( \frac{0.5}{\theta(d)} \right)^2 C(\epsilon) } \),

\[
P_\pi \left( \left| t^{-1} \sum_{s=1}^{t} Y_s - \sum_{d=0}^{M} \theta(d) t^{-1} \sum_{s=1}^{t} E_\pi[\delta(Z_s)(d)] \right| > \Sigma_1(\gamma, t, \epsilon) \right) \leq \gamma,
\]

where

\[
\Sigma_1(\gamma, t, \epsilon) := (M + 1) \max_d \left( \min \left\{ \frac{\log t}{8\nu(\sigma(d))^2}, \left( \frac{0.5}{|\theta(d)|} \right)^2 C(\epsilon) \right\} \right)^{-1/2} \sqrt{\frac{\log t}{t} \log 3(M + 1)/\gamma},
\]

and \( \Sigma_1 \) is decreasing in \( \gamma \), decreasing in \( t \) and non-increasing in \( \epsilon \).

We now prove Proposition 5.

**Proof of Proposition 5.** Observe that

\[
t^{-1} \sum_{s=1}^{t} Y_s - \max_d \theta(d) = t^{-1} \sum_{s=1}^{t} Y_s - \sum_{d=0}^{M} \theta(d) e_t(d) + \sum_{d=0}^{M} \theta(d) e_t(d) - \sum_{d=0}^{M} \theta(d) \delta(\theta)(d)
\]

\[
+ \sum_{d=0}^{M} \theta(d) \delta(\theta)(d) - \max_d \theta(d)
\]

\[
= \left( t^{-1} \sum_{s=1}^{t} Y_s - \sum_{d=0}^{M} \theta(d) e_t(d) \right) + \left( \sum_{d=0}^{M} \theta(d) e_t(d) - \sum_{d=0}^{M} \theta(d) \delta(\theta)(d) \right)
\]

\[
+ \epsilon (M + 1) \left( \frac{\sum_{d=0}^{M} \theta(d)}{M + 1} - \max_d \theta(d) \right)
\]

=: Term1 + Term2 + Term3.
Therefore,

\[ P_\pi \left( |t^{-1} \sum_{s=1}^{t} Y_s - \max_d \theta(d)| > \Sigma_1(\gamma, t, \epsilon) + ||\theta||_1 t^{-1} \sum_{s=1}^{t} \Lambda_s(\Delta, \bar{\zeta}_0, \nu_0, \epsilon) + \text{Bias}(\epsilon) \right) \]

\[ \leq P_\pi \left( |\text{Term}_1| > \Sigma_1(\gamma, t, \epsilon) \right) + 1 \left\{ |\text{Term}_2| > ||\theta||_1 t^{-1} \sum_{s=1}^{t} \Lambda_s(\Delta, \bar{\zeta}_0, \nu_0, \epsilon) \right\} + 1 \left\{ |\text{Term}_3| > \text{Bias}(\epsilon) \right\}. \]

We now bound each of these terms.

Term 1 is less than \( \gamma \) by Lemma 21. Regarding Term 2, observe that

\[ \text{Term}_2 = \sum_{d=0}^{M} \theta(d) e_t(d) - \sum_{d=0}^{M} \theta(d) \delta(\theta)(d) \]

\[ = (1 - \epsilon (M + 1)) t^{-1} \sum_{s=1}^{t} \sum_{d=0}^{M} \theta(d) E_\pi \left[ 1 \{ d = \arg \max_{a} \xi^a_s(a) \} - 1 \{ d = \arg \max_{a} \theta(a) \} \right]. \]

By convention, \( \arg \max_{a} \theta(a) = M \). Hence,

\[ \text{Term}_2 = (1 - \epsilon (M + 1)) t^{-1} \sum_{s=1}^{t} \sum_{d=0}^{M} \theta(d) E_\pi \left[ 1 \{ d = \arg \max_{a} \xi^a_s(a) \} - 1 \{ d = \arg \max_{a} \theta(a) \} \right] \]

\[ = (1 - \epsilon (M + 1)) t^{-1} \sum_{s=1}^{t} \left( \sum_{d=0}^{M-1} \theta(d) P_\pi(\mathcal{M}_s(d)) - \theta(M)(1 - P_\pi(\mathcal{M}_s(M))) \right), \]

where for any \( d < M, \mathcal{M}_s(d) := \{ \xi^a_s(d) > \max_{a \neq d} \xi^a_s(a) \} \) and \( \mathcal{M}_s(M) := \{ \xi^a_s(M) \geq \max_{a \neq M} \xi^a_s(a) \} \), because, by convention, if there are ties, the last arm is played. Since, \( P_\pi(\mathcal{M}_s(d)) \leq 1 - P_\pi(\mathcal{M}_s(M)) \), it follows that,

\[ |\text{Term}_2| \leq (1 - \epsilon (M + 1)) \left( \sum_{d=0}^{M} |\theta(d)| t^{-1} \sum_{s=1}^{t} (1 - P_\pi(\mathcal{M}_s(M))) \right) \]

\[ = (1 - \epsilon (M + 1)) \left( ||\theta||_1 t^{-1} \sum_{s=1}^{t} P_\pi(\exists d: \xi^a_s(d) > \xi^a_s(M)) \right). \]

By Proposition 6 with \( \epsilon = 0 \), for any \( s \geq \exp\{ 4 \max\{ \Gamma_{-1}(0.5(\epsilon + \Delta)), \Gamma_{-1}(0.5(\epsilon + \Delta)) \}, 0 \} \mathcal{B}(\epsilon \]
Thus,\[1 \left\{ \left| \text{Term2} \right| > \left| \theta \right| t^{-1} \sum_{s=1}^{t} \Lambda_s (\Delta, \tilde{\xi}_0, \nu_0, \epsilon) \right\} = 0.\]

Finally, it is clear that \[1 \{ \left| \text{Term3} \right| > \epsilon (M + 1) \left( \max_d \theta (d) - \frac{\sum_{d=0}^{M-1} \theta (d)}{M+1} \right) \} = 0.\]

\[\square\]

### F.1 Proofs of Lemmas

**Proof of Lemma 21.** Observe that

\[t^{-1} \sum_{s=1}^{t} Y_s - \sum_{d=0}^{L} \theta (d) e_t (d) = \sum_{d=0}^{L} t^{-1} \sum_{s=1}^{t} 1 \{ D_s = d \} Y_s (d) - \sum_{d=0}^{L} \theta (d) e_t (d)\]

\[= \sum_{d=0}^{L} t^{-1} \sum_{s=1}^{t} 1 \{ D_s = d \} \tilde{Y}_s (d) + \sum_{d=0}^{M} \theta (d) \{ t^{-1} \sum_{s=1}^{t} 1 \{ D_s = d \} - e_t (d) \} \]

\[= \sum_{d=0}^{L} t^{-1} \sum_{s=1}^{t} 1 \{ D_s = d \} \tilde{Y}_s (d) + \sum_{d=0}^{M} \theta (d) \{ f_t (d) - e_t (d) \}.\]

Hence, for any \( \epsilon \geq 0, \)

\[P_\pi \left( \left| \left| t^{-1} \sum_{s=1}^{t} Y_s - \sum_{d=0}^{M} \theta (d) e_t (d) \right| \right| > \epsilon \right) \leq \sum_{d=0}^{M} \left\{ P_\pi \left( \left| t^{-1} \sum_{s=1}^{t} 1 \{ D_s = d \} \tilde{Y}_s (d) \right| > 0.5 \epsilon / (M + 1) \right) \right.\]

\[\left. + \ P_\pi \left( \left| \theta (d) \{ f_t (d) - e_t (d) \} \right| > 0.5 \epsilon / (M + 1) \right) \right\}.\]

\(^{25}\Gamma_+ \text{ and } \Gamma_- \text{ are defined in Lemma 17.}\)
Let \( \delta = \log 3(M + 1)/\gamma \) and \( \epsilon \) be equal to \( \max_d (M + 1) \sqrt{\frac{\log t}{t}} \left( \min \left\{ \frac{\log t}{8\nu(\sigma(d))^2}, (0.5\theta_d)^2 C(\epsilon) \right\} \right)^{-1} \delta \).

We now invoke Lemmas 18 and 19, and for this we need to check that for our choice of \( (t, \epsilon) \) it follows that \( t \geq e^{4\epsilon B(\epsilon)} \). To do this, note that \( t \mapsto \log t/t \) has its maximum at \( t = e \) and its value is \( 1/e < 1 \). Hence, \( \epsilon B(\epsilon) \leq \max_d (M + 1) \sqrt{\left( \left( \frac{0.5}{\theta_d} \right)^2 C(\epsilon) \right)^{-1} \delta B(\epsilon)} \). By assumption \( t \) is higher than this last quantity and thus the desired property holds.

Hence, it follows that
\[
P_\pi \left( \left| t^{-1} \sum_{s=1}^{t} 1\{D_s = d\} \hat{Y}_s(d) \right| > 0.5\epsilon/(M + 1) \right) \leq 2e^{-0.5t(0.5\epsilon/(M+1))^2/(\nu(\sigma(d))^2)}
\]
\[
= 2e^{-0.5t \log t (0.5\epsilon/(M+1))^2 \log t \nu(\sigma(d))^2}
\]
and
\[
P_\pi \left( \left| \theta(d) \{ f_i(d) - e_t(d) \} \right| > 0.5\epsilon/(M + 1) \right) \leq e^{-t \log t (0.5\epsilon/(M+1)\nu(d))^2 C(\epsilon)}.
\]

Therefore,
\[
P_\pi \left( \left| t^{-1} \sum_{s=1}^{t} Y_s - \sum_{d=0}^{M} \theta(d) e_t(d) \right| > \epsilon \right) \leq 3 \sum_{d=0}^{M} e^{-t \log t (0.5\epsilon/(M+1)\nu(d))^2 C(\epsilon)}.
\]

And by our choice of \( \epsilon \), it follows that
\[
P_\pi \left( \left| t^{-1} \sum_{s=1}^{t} Y_s - \sum_{d=0}^{M} \theta(d) e_t(d) \right| > (M + 1) \max_d \left\{ \log t, \frac{\log t}{8\nu(\sigma(d))^2}, (0.5\theta_d)^2 C(\epsilon) \right\}^{-1} \delta \right)
\]
\[
\leq 3(M + 1)e^{-\delta}.
\]

Since \( \delta = \log 3(M + 1)/\gamma \) the desired result follows. \( \square \)