Reinforcing RCTs with Multiple Priors while Learning about External Validity *

Frederico Finan † Demian Pouzo‡
UC Berkeley UC Berkeley

September 2024

Abstract

This paper introduces a framework for incorporating prior information into the design of sequential experiments. These sources may include past experiments, expert opinions, or the experimenter's intuition. We model the problem using a multi-prior Bayesian approach, mapping each source to a Bayesian model and aggregating them based on posterior probabilities. Policies are evaluated on three criteria: learning the parameters of payoff distributions, the probability of choosing the wrong treatment, and average rewards. Our framework demonstrates several desirable properties, including robustness to sources lacking external validity, while maintaining strong finite sample performance.

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

JEL: C11, C50, C90, O12.

*We thank Susan Athey, Marina Dias, Pat Kline, Chiara Motta, Chao Qin and Vira Semenova 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

1 Introduction

Adaptive experiments offer several potential advantages over standard randomized controlled trials (RCTs), including shorter experiment duration and more optimized treatments. This can lead to more effective outcomes and lower experimental costs. Such experimentation methods are increasingly used across various fields, including clinical research and marketing. Although their application in economics is still limited, adaptive experiments hold significant promise for improving RCTs aimed at identifying optimal policy variants.

When deciding to conduct adaptive experiments, policymakers or researchers often have some prior knowledge about the effectiveness of the treatments. This information might come from past experiments, pilot studies, or expert opinions. Depending on its external validity, this prior knowledge could allow for less experimentation. This raises two relatively unexplored questions: how does one incorporate this previous knowledge into an adaptive experiment? And in doing so, will this allow the experiment to stop sooner without increasing the risk of choosing the wrong treatment? This paper provides a tractable, but novel framework for how to incorporate prior sources of information into the design of a sequential experiment. Our framework is sufficiently general so that priors can include previous experiments, expert opinions, or the experimenter's own introspection.

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

As the policymaker gathers more data from own experiment, she uses Bayes' rule to update each of her prior sources and then takes a weighted average of each source's posterior where the weights depend on how well the sources fit the observed data. On the basis of these beliefs, the policymaker then decides whether to stop the experiment and which treatment to assign. By incorporating potentially useful information, our policymaker may be able to stop the experiment sooner, thereby generating efficiency gains without increasing the risk of adopting the incorrect treatment.

In settings, such as this one, in which the policymaker must learn the truth, it is common not to use the optimal assignment rule. This rule (i.e. the one that maximizes her *subjective* payoff) can have undesirable properties, such as failing to learn the correct treatment effects or being hard to compute and implement.[1] As a result, the literature on multi-armed bandits have studied different heuristic rules such as ϵ-greedy (Watkins, 1989) and Thompson Sampling (Thompson, 1933) and its refinements (e.g. Upper Confidence Bounds (Lai and Robbins, 1985), or exploration sampling (Kasy and Sautmann, 2021; Russo, 2016)). We take a different approach and study a large class of policy rules that encompass, among others, the aforementioned examples. Importantly, we find that the only feature of the policy rule that matters for performance is the exploration structure – a sequence quantifying the amount of experimentation that occurs under a given policy rule at each stage of the experiment.

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

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

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

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

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

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

Finally, we also compute bounds for the rate at which the average observed outcomes converges to the maximum expected outcome. We show that the rate of convergence for these bounds are governed by an "exploitation versus exploration" trade-off. If we increase the degree of experimentation (less exploitation, more exploration) our data become more independent and the underlying uncertainty decreases. However, by exploring more, we are also increasing the bias associated with not choosing the optimal treatment. Unfortunately, these bounds are sufficiently complicated that we cannot characterize analytically the "optimal" degree of experimentation. Nevertheless, the results do suggest that pure experimentation (as in the case of an RCT) is unlikely to be optimal, and we verify this numerically in a series of simulations.

Application to Debt Refinancing To further illustrate our procedure, we ran an experiment with the largest private bank in Argentina to enhance the marketing of their debt-refinancing program. This initiative came at a critical time, as Argentina was experiencing a surge in household debt, exacerbated by high inflation and interest rates. The bank aimed to engage over 300,000 delinquent clients, primarily from lower-income households, through targeted email campaigns. In February 2024, the bank conducted an RCT with a sample of around 15,491 clients to determine if displaying monthly payments instead of the high-interest rates in emails would increase engagement. Contrary to expectations, clients were more likely to click on the link when the interest rate was displayed, with response rates varying significantly across different regions of Argentina.

In June 2024, as interest rates started to decline, the bank launched a second, adaptive experiment using our algorithm. The experiment was conducted in three regions and used the results of the previous experiment as initial priors. The results showed that our multi-prior approach proved effect as fewer emails were needed to achieve significant results compared to traditional randomized controlled trials. In the capital city of Buenos Aires, for instance, the experiment was stopped after 294 emails, far fewer than would be needed in traditional methods.

Contributions to the Literature Our paper relates to three strands of the literature. First, we speak to an extensive multi-disciplinary literature on adaptive experimental design. Much of the focus of this literature has been on the multi-arm bandit problem, which considers how best to assign experimental units sequentially across treatment arms. Depending on the objective function, numerous studies have proposed a variety of alternative algorithms that, on average, outperform the static assignment mechanisms of traditional RCTs.[2] 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

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

data, but also uncertainty over the data generating process of the source data. To our knowledge, this is the first paper to introduce multiple priors into the design of an adaptive experiment.

Much of the literature on multi-armed bandits has focused on deriving bounds on expected regret for specific solution heuristics.[3] Instead, we focus on alternative performance criteria, such as average outcomes, the probability of making a mistake, and concentration rates for posterior means, which to the best of our knowledge have not been formalized in a multi-prior multi-arm Bayesian bandit framework.[4] Moreover, the results we derive are for a general class of solution heuristics, not a specific one. For these reasons, even though we do not view the technical results as the primary contribution of the paper, we do believe that they might be of independent interest even in standard multi-arm bandit problems. Furthermore, we view our paper as complementary to this existing literature, as techniques tailored for particular solution heuristics can be combined with our multi-prior Bayesian setting to obtain sharper theoretical guarantees.

By introducing issues of externality validity into the multi-arm bandit problem, our study also connects to the literature on measuring the generalizability of experiments. In general, scholars have taken three approaches for assessing external validity. One common approach is to measure how well treatment effect heterogeneity extrapolates to 'left out' study sites. Under the assumption that study site characteristics are independent of potential outcomes, a number of studies applying alternative estimators have interpreted the out-of-sample prediction errors as a measure or test of external validity.[5] A related approach uses local average treatment effects across different complier populations to test for evidence of external validity (e.g. Angrist and Fernández-Val (2013); Kowalski (2016); Bisbee et al. (2017)). The general idea being that if differences in observable characteristics across subgroups explain differences in treatment effect heterogeneity then we can make some claim for external validity. A third common approach adopted in the meta-analysis literature is the use of hierarchical models to aggregate treatment effects across different study sites. A byproduct of this framework is a "pooling factor" across study sites that has a natural interpretation of generalizability. The factor compares the sampling variation of a particular study site to the underlying variation in treatment heterogeneity: the higher the measure, the larger the sampling error and the less informative the study site is about the overall treatment effect (e.g. Vivalt (2020),

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

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

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

Gelman and Carlin (2014), Gelman and Pardoe (2006), Meager (2020)).[6]

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 show that, as t diverges, the weights are only positive for the least externally invalid models, allowing us to interpret these weights as measures of external validity.

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

Finally, our paper is related to the Hierarchical Bayes methodology and, albeit tangentially, to the ambiguity aversion literature. We employ the multi-prior Bayesian updating formalism from the ambiguity aversion literature in dynamic settings (e.g., Al-Najjar and Weinstein (2009); Epstein and Schneider (2007)) to describe the problem of a policymaker (henceforth, PM) with access to different sources of information. However, the literature has emphasized dynamic consistency within optimal, non-myopic agents. Our paper is unrelated to these issues, as we are not concerned with optimality, and therefore not with consistency either. As explained above, it is not obvious to us that optimality (from the point of view of the PM) is a desirable property in this context. Our paper simply aims to provide a method for incorporating prior information under a class of heuristic procedures, which encompasses commonly used methods and offers certain appealing theoretical guarantees, such as learning the true treatment effects and controlling the probability of making a mistake when stopping the experiment.

Our methodology is also related to the hierarchical Bayesian methodology. Our PM acknowledges (model) uncertainty but is not averse to it, as she *averages* across different models. In Appendix H, we show and discuss a type of *certainty equivalence result*, which states that the updating problem

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

we propose is mathematically equivalent to a certain empirical hierarchical Bayesian model (EHB). This result relies on an insight akin to the classical certainty equivalence result wherein, in a framework with risk, a risk-neutral agent makes decisions as if there is no risk. However, an important feature of our updating problem that differentiates it from classical hierarchical Bayesian models is that the weights are updated iteratively as more information arises, shifting towards models that best fit the data.[7] We prefer the "model uncertainty" interpretation to the EHB because we believe it better describes the motivation behind the PM's problem.[8]

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

2 Setup

In this section, we describe the problem our policymaker (PM) aims to solve. Our PM's problem consists of three parts: the experiment, the policy functions, and the learning framework.

2.1 The Experiment

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

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

[7]This last observation presents a dichotomy between learning—using relevant information from the data—and uncertainty aversion—being robust to potential model misspecification. In Appendix H, we discuss this further and present a possible extension that merges these ideas.

[8]Another, perhaps less important, reason to prefer the "model uncertainty" interpretation is that the EHB interpretation is only valid if the PM is uncertainty neutral; it breaks down if the PM is ambiguity averse. See Appendix H for more details.

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

- If the PM decides to stop the experiment,
  - she chooses a treatment assignment at instance t that will be applied to all subsequent units.
- If the PM does not stop the experiment,
  - she chooses a treatment assignment for each unit x at time t.
  - Nature draws potential outcomes, $Y_t(d,x)$, for each unit.
  - The PM only observes the outcome corresponding to the assigned treatment, i.e. $Y_t(D_t(x),x)$.

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

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

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

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

The parameter of interest. For each instance t, the learning setup gives raise to a subjective PDF over each of the (d,x)-outcomes given by $\int p_\theta(\cdot)\mu_t^\alpha(d,x)(d\theta)$ that the PM uses to form recommendations and decisions. While this setup is sufficiently general to allow for any parameter of interest, we focus on the average treatment effect setup wherein the PM wants to learn

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

2.2 The Policy Rule

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

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

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

$$\mathbf{P}\left(t^{-1}\sum_{s=1}^{t}\delta_s(d|x) \geq h_t(d|x)\right) \geq 1-\omega_t(d,x). \tag{2.1}$$

We call $(h_t)_t$ the **degree of exploration** of the policy rule and $(1-\omega_t)_t$ the **likelihood of exploration** of the policy rule. By providing a lower bound on the (average) propensity score, the structure of exploration quantifies the extent to which experimentation occurs under the policy rule $(\delta_t)_t$. This structure is the *only* feature of the policy rule that matters for our performance criteria. We present these results formally in Section 3.[9]

Supplemental material J presents several commonly-used policy rules — and their associated exploration structure — in the context of the Gaussian learning framework, which we describe next.

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

2.3 The Gaussian Learning Model

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

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

For each $(d,x) \in \mathbb{D} \times \mathbb{X}$, the family $\mathcal{P}_{d,x}$ and the collection of prior beliefs give rise to $L+1$ subjective Bayesian models for $P(\cdot|d,x)$. Given the observed data of past treatments and outcomes, at instance $t \geq 1$, the PM observes the realized outcome $Y_t(D_t(x),x)$ and the treatment assignment $D_t(x)$. Using Bayesian updating, she then forms posterior beliefs for each model, which we denote by $\mu_t^o(d,x)$. We observe that the belief is updated using observed data, $(Y_t(D_t(x),x), D_t(x))$ and using p_θ as the PDF of $Y_t(D_t,x)$ given $D_t(x) = d$, this feature is analogous to the missing data problem featured in experiments under the frequentist approach.[10]

In what follows, we will assume that the PM takes subjective models within the Gaussian family (see Section I in the Supplemental Material for the general setup). This assumption, and the corresponding conjugate priors, imply that the posterior belief is fully characterized by a finite dimensional object, which is more tractable. Formally, for each $(d,x) \in \mathbb{D} \times \mathbb{X}$, $\mathcal{P}_{d,x}$ is a family of Gaussian PDFs given by $\{\phi(\cdot;\theta,1) \colon \theta \in \mathbb{R}\}$ and the prior for every source is also assumed to be Gaussian with mean $\zeta_0^o(d,x)$ and variance $1/\nu_0^o(d,x)$.[11] The quantity $\nu_0^o(d,x)$ can be interpreted as the number of units with characteristics x that were assigned treatment d in a past experiment. The higher the $\nu_0^o(d,x)$, the more certain source o is about $\phi(\cdot;\zeta_0^o(d,x),1)$ being the correct model.

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

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

Throughout, we will assume $(\zeta_0^o, \nu_0^o)_{o=0}^L$ are non-random.

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

$$
\begin{aligned}
\zeta_t^o(d,x) =& \frac{1\{D_t(x)=d\}}{\nu_{t-1}^o(d,x)+1\{D_t(x)=d\}} Y_t(d,x) + \frac{\nu_{t-1}^o(d,x)}{\nu_{t-1}^o(d,x)+1\{D_t(x)=d\}} \zeta_{t-1}^o(d,x) \\
=& \frac{J_t(d,x)}{f_t(d,x)+\nu_0^o(d,x)/t} + \frac{\nu_0^o(d,x)/t}{f_t(d,x)+\nu_0^o(d,x)/t} \zeta_0^o(d,x) \qquad (2.2) \\
& where \\
\nu_t^o(d,x) =& N_t(d,x)+\nu_0^o(d,x), \;\; f_t(d,x) := N_t(d,x)/t \qquad (2.3) \\
& and\; J_t(d,x) := t^{-1} \sum_{s=1}^{t} Y_s(d,x) 1\{D_s(x)=d\}. \qquad (2.4)
\end{aligned}
$$

From these expressions, we can see how Gaussianity simplifies the dynamics of the problem. We only need to analyze $(\zeta_t^o(d,x), \nu_t^o(d,x))_{t=0}^T$, a finite dimensional object, as opposed to $(\mu_t^o(d,x))_{t=0}^T$, an infinite dimensional object that is quite intractable. Note however that even with the Gaussianity assumption, our setup remains quite general in practice as we describe in the following remark.

Remark 2.1. *(1) Since the PM cares about learning the ATE, this model is sufficiently general to encompass the canonical RCT setup for estimation of average treatment effects, even when the potential outcomes are not necessarily Gaussian. To see this, note that even if the PM's subjective model for potential outcomes is misspecified (i.e. she incorrectly assumes that $Y(d,x)$ is Gaussian) the PM can still accurately learn the true average effect because, for each (d,x), there always exists a θ such that $\theta = E_{P(.|d,x)}[Y(d,x)]$. We show this is the case in Section 3.1.*

(2) Our results and methodology extend to any subjective model whose posterior beliefs can be fully described by low finite-dimensional objects. For instance, in cases where $Y(d,x) \in \{0,1\}$, they extend to the Bernoulli-Beta model wherein the t instance posterior is given by a Beta density with parameters given by $(\sum_{s=1}^t 1\{D_s(x)=d\}Y_s(d,x) + \nu_0^o(d,x)\zeta_0^o(d,x), \sum_{s=1}^t 1\{D_s(x)=d\}(1 - Y_s(d,x)) + \nu_0^o(d,x)(1-\zeta_0^o(d,x)))$. More generally, our methodology can be extended to the entire exponential family — which includes the models considered here and more (see Schlaifer and Raiffa (1961) for examples), however, the interpretation of $\zeta_t(d,x)$ may change. $\triangle$

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

$$\mu_t^{\alpha}(d,x) := \sum_{o=0}^{L} \alpha_t^o(d,x)\mu_t^o(d,x) \tag{2.5}$$

where

$$\alpha_t^o(d,x) := \frac{\int \prod_{s=1}^{t} \phi(Y_s(d,x);\theta,1)^{1\{D_s(x)=d\}} \phi(\theta;\zeta_0^o(d,x),1/\nu_0^o(d,x))(\theta)d\theta}{\sum_{o=0}^{L}\int \prod_{s=1}^{t} \phi(Y_s(d,x);\theta,1)^{1\{D_s(x)=d\}} \phi(\theta;\zeta_0^o(d,x),1/\nu_0^o(d,x))(\theta)d\theta}.$$

We can interpret $\alpha_t^o(d,x)$ as the PM's subjective probability that source o for (d,x) is more externally valid for her current experiment. To expound on this last point, we introduce the concept of degree of external validity of a given source withing the Gaussian learning model, and then we relate this concept to the behavior of $(\alpha_t^o(d,x))_{o=0}^{L}$. But first, it useful to introduce some nomenclature. We call $|\zeta_0^o(d,x)-\theta(d,x)|$ the **bias of source** o and $\nu_0^o(d,x)$ the **degree of conviction of source** o — since a higher $\nu_0^o(d,x)$ implies a lower prior variance. As such, we can interpret $|\zeta_0^o(d,x)-\theta(d,x)|\sqrt{\nu_0^o(d,x)}$ as the **degree of stubbornness of source** o.

Definition 1 (Degree of external validity)**.** *For each $(d,x) \in \mathbb{D}\times\mathbb{X}$ the degree of external validity (DEV) of source o is given by*

$$\mathbb{EV}_{(d,x)}(o) := -\nu_0^o(d,x)(\theta(d,x)-\zeta_0^o(d,x))^2 + \log \nu_0^o(d,x).$$

The value of this quantity is that it allows us to define a partial ordering over sources given by the next definition

Definition 2 (Externally valid sources)**.** *For (d,x), a source o is more externally valid than source a o' if has a higher degree of external validity, i.e.,*

$$\mathbb{EV}_{(d,x)}(o) > \mathbb{EV}_{(d,x)}(o').$$

We denote this as $o \succ_{(d,x)} o'$. A source o is externally valid, if $\nu_0^o(d,x)(\theta(d,x)-\zeta_0^o(d,x))^2 = 0$ and $\nu_0^o(d,x) = +\infty$.

According to this definition the degree of external validity of a given source depends on its degree of stubbornness in a decreasing manner, i.e., the more stubborn the source is, the lower its DEV is. The DEV also depends on the degree of conviction, but its dependence is more nuanced. To see this, observe that $\frac{d\mathbb{EV}_{(d,x)}(o)}{d\nu_0^o(d,x)} = -(\theta(d,x) - \zeta_0^o(d,x))^2 + 1/\nu_0^o(d,x)$. Thus, the effect of the degree of conviction on the DEV depends on the level of the bias, and it does so in an intuitive way. To see this, first note that for unbiased sources, a higher degree of conviction can only increase the DEV. However, for biased sources, the effect is not uniform: For low initial values of conviction an increase on this quantity will increase the DEV, but for high enough initial levels of conviction an increase of it will lower the DEV; i.e., the source is becoming more stubborn, re-affirming the bias. Finally, we note that the "cutoff" conviction level is inversely proportional to the level of the bias: A higher bias imply a larger range of initial conviction levels for which $\frac{d\mathbb{EV}_{(d,x)}(o)}{d\nu_0^o(d,x)} < 0$. This behavior of the DEV with respect to conviction can be achieved with functional forms other than the current one. In particular, one can replace the $\log(.)$ in the definition by other increasing and concave function and still be able to obtain the aforementioned behavior. The particular choice of $\log(.)$ in this case stems from the fact that the prior and subjective model are both Gaussians.

Figure 1 illustrates this discussion. The plot on the left presents the case where $o \succ_{(d,x)} o'$ because even though both sources are unbiased, source o has a higher level of conviction. In the middle plot, $o \succ_{(d,x)} o'$, but now the the level of conviction is the same whereas the bias is lower for source o. Finally, the right plot illustrates the nuanced role of conviction: Source o'' (dashed black line) has a small bias but this gets amplified by a very high degree of conviction (i.e., it is highly stubborn), rendering it less externally valid than source o' (blue solid line) which is unbiased and with low degree of conviction; however source o (solid red line) has the same small bias as source o'' but a degree of conviction that is lower than o'' but higher than o', rendering *more* externally valid that source o' and o''.

In sum, definitions 1 and 2 extend the concept of external validity to a Bayesian framework. In the classical — non-Bayesian — setup, bias defines whether a sources is externally validity (unbiased) or not (bias), as stated in Definition 2. However, this view of external validity is perhaps too narrow within a (Gaussian) Bayesian framework with a conviction level that may not be infinite. In this framework, external validity ceases to be a *qualitative* notion and becomes a *quantitative* notion which depends both on the bias and the level of conviction. This is precisely what our DEV measure captures.

The next proposition provides a link between this ranking of external invalidity and our weights $(\alpha_t^o(d,x))_{o=0}^L$. It shows that under some technical regularity assumptions, the weights are only positive for the least externally invalid models as t diverges, provided (d,x) is played sufficiently often.

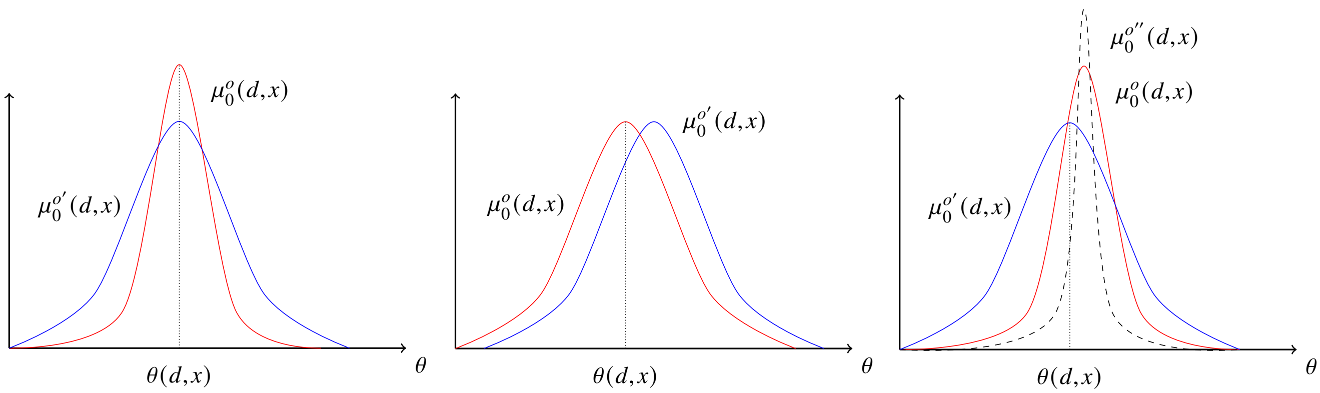


Figure 1: Different Degrees of External Validity

Proposition 2.1. *For any* $(d,x) \in \mathbb{D} \times \mathbb{X}$,

1. $\lim_{\mathbb{EV}_{(d,x)}(o) \to -\infty} \alpha_t^o(d,x) = 0$.
2. *If* $\inf_t t^{-1} \sum_{s=1}^t \delta_s(d|x) > 0$, *then*

$$\alpha_t^o(d,x) = \frac{e^{0.5\mathbb{EV}_{(d,x)}(o)}}{\sum_{o'=0}^L e^{0.5\mathbb{EV}_{(d,x)}(o')}} + o_{\mathbf{P}}(1).$$

In particular, if $o >_{(d,x)} o'$, *then* $\alpha_t^{o'}(d,x) < \alpha_t^o(d,x) + o_{\mathbf{P}}(1)$, *and if* o *is the externally valid source, then* $\alpha_t^o(d,x) = 1 - o_{\mathbf{P}}(1)$.[12]

Proof. See Appendix B. □

Part (1) of the proposition shows that α_t^o offers certain robustness properties against externally-invalid models. If a source has a high degree of external invalidity — i.e., a very negative $\mathbb{EV}_{(d,x)}(o)$ —, then the associated weight of that source is approximately 0. Part (2) offers a sharper characterization of this robustness property, but asymptotically. It links the degree of external validity and the weight each source will have in the limit. In particular, sources that are less externally valid will receive less weight, and externally valid sources will get weight approaching one.[13]

3 Analytical Results

As mentioned above, the object of interest is the average effect of each treatment, and, at each instance t and for each $(d,x) \in \mathbb{D} \times \mathbb{X}$, the PM will estimate it using a subjective average treatment

[12]The definition of externally valid source forces us to work with $+\infty$. We follow the standard convention that $a/(+\infty) = 0$ for any real number a.

[13]In Section H we discuss alternative interpretations of and potential extensions to our model.

effect given by

$$\zeta_t^{\alpha}(d,x) := \int y \int_{\Theta} p_{\theta}(y)\mu_t^{\alpha}(d,x)(d\theta)dy =: \sum_{o=0}^{L} \alpha_t^{o}(d,x)\zeta_t^{o}(d,x).$$

The middle expression is simply the mean of outcome computed with respect to the PM's subjective PDF of y constructed using the aggregate beliefs at time t, $\int_{\Theta} p_{\theta}(\cdot)\mu_t^{\alpha}(d,x)(d\theta)$. The right-most expression shows that such mean is the weighted average of the posterior mean for each source.

In this section, we establish some finite sample properties of this quantity, such as the rate at which it concentrates around the true average effect. Before we do so, a bit of housekeeping is required. Moving forward, we will omit x from the notation and derive our results for $|\mathbb{X}| = 1$. Given our assumptions, we can learn the fundamentals for each $x \in \mathbb{X}$ by treating them as separate and independent problems. Thus, we can extend all our results to the case of $|\mathbb{X}| > 1$ by treating the relevant quantities (e.g. $\theta(d)$, $Y(d)$, etc.) as vectors of dimension $|\mathbb{X}|$. Furthermore, to derive the results below we will need some assumptions on the (true) distribution of the potential outcomes,

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

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

Before presenting these results formally, it is useful to present our general approach for how we derived them. As we discussed above, a key of object of interest is $\zeta_t^{\alpha} = \sum_{o=0}^{L} \alpha_t^{o}\zeta_t^{o}$, the subjective average effect of treatment at instance t. Most of our results hinge on understanding how this object concentrates around the true expected value θ.

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

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

$$|\zeta_t^{\alpha}(d)-\theta(d)| \leq \Gamma(\gamma, h_t(d)-\eta, |\zeta_0(d)-\theta(d)|, \nu_0(d)) \tag{3.1}$$

where $h_t(d)$ is the degree of exploration and $\Gamma:\mathbb{R}_+\times[0,1]\times\mathbb{R}^{L+1}\times\mathbb{N}^{L+1}\to\mathbb{R}$ is a function defined in Appendix C.3. Thus, Γ maps the concentration rate of $J_t(d)$ and $f_t(d)$ (given by γ and η respectively) to the concentration rate of the posterior mean around the true parameter. In fact, we show in Lemma C.5 in the Appendix C.3 that Γ is increasing in the first argument and decreasing in the second one, thereby implying that a faster concentration rate of $J_t(d)$ and $f_t(d)$ translate into a faster concentration rate of the posterior mean. Moreover, Γ also quantifies how a higher degree of exploration translates into a faster concentration rate, as well as how the source's parameters, $(\zeta_0(d),\nu_0(d))$, affect this rate.

3.1 Concentration bounds on the Posterior Mean

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

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

$$\mathbf{P}\left(|\zeta_t^{\alpha}(d)-\theta(d)| > \Gamma\left(\sqrt{\frac{2\upsilon\varepsilon}{h_t(d)^2 t}}\sigma(d), 0.5h_t(d), |\zeta_0(d)-\theta(d)|, \nu_0(d)\right)\right) \leq 4(e^{-\varepsilon}+\omega_t(d)).$$

Proof. See Appendix E. □

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

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

Our method for aggregating multiple priors offers an attractive feature regarding our concentration rates. Sufficiently stubborn models, i.e. $|\zeta_0^o(d) - \theta(d)|\sqrt{\nu_0^o(d)}$ is sufficiently large, will have close to zero effect on the concentration rate of $\zeta_t^\alpha(d)$, as they are essentially dropped from the weighted average. This implies an *oracle* property in the sense that the concentration rate becomes arbitrary close to the least stubborn model, provided there is enough separation between the stubbornness of this model and the others. We formalize this property in the next corollary.

Corollary 3.1. *Take any* (t, d, ε) *as in Proposition 3.1. Furthermore, let model* $o = 0$ *denote the least stubborn model and suppose that for any given* $\delta > 0$*, there exists a* C *such that* $\min_{o \neq 0} |\zeta_0^o(d) - \theta(d)|\sqrt{\nu_0^o(d)} \geq C$*. Then,*

$$\mathbf{P}\left(|\zeta_t^\alpha(d) - \theta(d)| > \Omega\left(\sqrt{\frac{2\upsilon\varepsilon}{h_t(d)^2 t}}\sigma(d), 0.5h_t(d), |\zeta_0^o(d) - \theta(d)|, \nu_0^o(d)\right) + \delta\right) \leq 4(e^{-\varepsilon} + \omega_t(d))$$

Proof. See Appendix E. ◻

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

$$|\zeta_t^o(d) - \theta(d)| \leq \Omega(\gamma, h_t(d) - \gamma, |\zeta_0^o(d) - \theta(d)|, \nu_0^o(d)).$$

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

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

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

1. *All else equal, the concentration rate decreases as the bias increases; it also decreases with the degree of stubbornness, i.e.* $|\zeta_0^o(d) - \theta(d)|\sqrt{\nu_0^o(d)}$*. The concentration rate is fastest when the bias is zero.*

2. *For confident models, the concentration rate increases with the degree of conviction, i.e.* $\nu_0^o(d)$ *increases. The intuition behind this result is as follows: If* $\nu_0^o(d)$ *increases but* $|\zeta_0^o(d) - \theta(d)|\sqrt{\nu_0^o(d)}$ *remains constant — equal to 0, in particular —, then necessarily, the model is becoming more convinced about a prior that is unbiased, thereby implying a faster convergence rate.*

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

4. *An increase in the degree of the exploration,* $h_t(d)$*, improves the concentration rate. This comes from the fact that* $h_t(d) \mapsto \Omega\left(\sqrt{\frac{2\upsilon\varepsilon}{h_t^2(d)t}}\sigma(d), 0.5h_t(d), |\zeta_0^o(d) - \theta(d)|, \nu_0^o(d)\right)$ *is decreasing (see Lemma C.3 in the Appendix C.2). Intuitively, increasing* $h_t(d)$ *implies having more observations to estimate* $\theta(d)$ *— "more information" about treatment* d *implies a faster concentration rate.* $\triangle$

3.2 Probability of making a mistake

In this section, we provide bounds on the probability of making a mistake. To do so, we need to first define the policy rule for stopping the experiment, σ, since this rule does govern the probability of making mistakes when stopping the experiment. A desirable property for this rule is that, for a given tolerance level $\beta \in (0,1)$ chosen by the PM, the probability of making a mistake when stopping the experiment is no larger than β. We propose the following rule:

Example 1 (Threshold Stopping Rule)**.** *For any positive-valued non-increasing sequence* $(\gamma_t)_t$ *and* $B \in \mathbb{N}$*, the stopping rule parameterized by* $((\gamma_t)_t, B)$ *is such that, for any* $t \geq B$*,*

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

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

$$c_t(\gamma_t, d, m) := c_t(\gamma_t, d) + c_t(\gamma_t, m) := \left(\sum_{o=0}^{L} \frac{\alpha_t^o(d)\sqrt{t}\gamma_t(d)}{N_t(d) + \nu_0^o(d)} + \sum_{o=0}^{L} \frac{\alpha_t^o(m)\sqrt{t}\gamma_t(m)}{N_t(m) + \nu_0^o(m)} \right)$$

where $N_t(d) := \sum_{s=1}^{t} 1\{D_s = d\}$.

Loosely speaking, the rule proposes to stop the experiment after B instances and as soon as the distance between the highest average posterior and the rest — measured by $\max_d \min_{m \neq d}(\zeta_t^\alpha(d) - \zeta_t^\alpha(m))$ — is far enough from zero, where "far enough" is essentially measured by the cutoff $c_t(\gamma_t, d, m)$. While the expression for this cutoff is a bit involved, we highlight one key aspect of it. The appropriate scaling is given by $\sqrt{t}/(N_t(d) + \nu_0^o(d))$ as opposed to the standard scaling of $1/\sqrt{t}$ one obtains in hypothesis testing. The difference lies on the fact that our methodology uses prior information and thus leverages the "prior observations" represented by $\nu_0^o(d)$. Models with high levels of confidence (a high $\nu_0^o(.)$) will have lower cutoffs, prompting the rule to stop sooner. This effect naturally decreases as the sample size t increases, but for finite sample sizes it could still be sizable. $\triangle$

Suppose treatment M has the largest expected effect, i.e., $\Delta := \theta(M) - \max_{d \neq M} \theta(d) > 0$. We define a mistake as recommending a treatment different than M at the instance t in which the experiment was stopped. Because recommendations are based on the PM's posteriors, a mistake is given by

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

where τ indicates when the experiment is stopped, i.e., is the first instance after B such that $\max_d \min_{m \neq d}\{\zeta_t^\alpha(d) - \zeta_t^\alpha(m) - c_t(\gamma_t, d, m)\} > 0$ where the cutoffs c_t are defined in Example 1.[14]

The following proposition provides an upper bound for the probability of making a mistake associated with this stopping rule, when all sources are unbiased or are biased "in the right direction".[15]

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

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

[14] Once again, we will omit x from the notation in what follows.

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

Proof. See Appendix F. □

The intuition behind this proposition is as follows. Mistakes occur when, at some instance t greater than B, the posterior mean of some treatment d — different from M — is "much larger" than the others. This implies that the posterior has to be "much larger" than its population mean, $\theta(d)$, where "much larger" depends on the pre-specified cutoff. Hence, given our assumption on the priors, the probability of a mistake is essentially given by the probability that the outcome-treatment average exceeds its population value by an amount given by $\gamma_t/\sqrt{t}$. Lemma D.1 in Appendix D provides the bound of $e^{-0.5\frac{\gamma_t^2}{\upsilon\sigma(d)^2}}$.

In Proposition 3.2, we assumed unbiased sources or that the priors ranked treatment M as the highest. The next collorary proves that when some sources are biased, there still exists an oracle property akin to the one demonstrated for the concentration rates. In particular, we show that upper bound is arbitrary close to the unbiased source, provided the other sources are sufficiently biased.[16]

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

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

Proof. See Appendix F. □

Proposition 3.2 also reveals how by properly choosing $((\gamma_t)_t, B)$, the probability of a mistake associated with the stopping rule is bounded by β, where $\beta \in (0,1)$ is any tolerance level. The next corollary presents such result.[17]

Corollary 3.3. *Suppose all the conditions of Proposition 3.2 hold, and, for any t, $\gamma_t(d) \geq 2\sqrt{\upsilon}\sigma(d)A$ for all $d \in \mathbb{D}$ with A such that*

$$A \geq -\log\frac{\beta}{M+1}. \tag{3.3}$$

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

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

Then

$$\max_{t\in\{B,\ldots,T\}} \mathbf{P}\left(\max_{d\neq M}\{\zeta_\tau^\alpha(d)-\zeta_\tau^\alpha(M)\}>0\cap\{\tau=t\}\right)\leq\beta.$$

Proof. See Appendix F. □

The choice of $(\gamma_t)_t$ is so that the terms in the upper bound in Proposition 3.2 are less than β. The sequence $(\gamma_t)_t$ has to bounded below by $2\sqrt{\upsilon}\sigma(.)A$, the term $2\sqrt{\upsilon}\sigma(.)$ arises from Assumption 2. In cases where $\upsilon\sigma(.)$ is unknown, one can replace this value by a sample analog — the same way one estimates the standard deviations in the difference in means test — or simply by an upper bound such as $\log\log t$, both will be valid for large t. The term A is simply to ensure that $2\sum_{d\in\mathbb{D}}e^{-A}$ is less than the tolerance level. Finally, we note that the sequence $(\gamma_t)_t$ can be larger than $2\sqrt{\upsilon}\sigma(.)A$. However, we do not recommend this, large values of γ_t are undesirable because the larger the γ_t, the less likely it is to stop the experiment at any instance thereby implying longer — and more costly — experiments.

3.3 Average Observed Outcomes

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

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

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

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

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

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

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

where

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

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

Proof. See Appendix G. □

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

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

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

4 Model Simulations

In this section, we present Monte Carlo simulations of our model using the generalized ϵ-Greedy policy rule presented in Section 2.2. The purpose of these simulations is to highlight different aspects of our analytical results and to provide a sense of the tightness of our analytic bounds. We consider the case with only two treatment arms, $D \in \{0,1\}$, and assume that $Y(0) \sim N(1,1)$ and $Y(1) \sim N(1.3,1)$. We assess the performance of our model according to the three outcomes outlined in Section 3: concentrations bounds, probability of making a mistake, and average earnings. We simulate each experiment 1000 times, with each experiment lasting at most 1000 instances.

Multiple Priors, External Validity, Robustness

We begin by illustrating how our setup weights the different models over the course of the experiment. Recall that to aggregate across several distinct subjective Bayesian models, our setup will average the posterior beliefs of each model using as weights, $\alpha_t^o(d)$ – the posterior probability that model o best fits the observed data within the class of models being considered. We demonstrated in Proposition 2.1 for the general case, and Lemma B.1 for Gaussianity, that if there exists an externally valid model among externally *invalid* models, then $\alpha_t^o(d)$ will approach one for the externally valid model. Conversely, $\alpha_t^o(d)$ will approach zero if models are far from the true $\theta(d)$.

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

In Figure 2, we plot $\alpha_t^o(d)$ corresponding to the second set of priors over the course of the experiment. The graph on the left is for the $d = 0$ arm and the one on the right is for the $d = 1$ arm. Each line corresponds to a different set of priors, and the lighter the line, the more stubborn the prior. Starting with the top and darkest line, we see that $\alpha_t^o(d)$ increases over time putting more and more weight on an externally valid model. By the end of the experiment, $\alpha_t^o(d)$ is close to 95% for both arms. As we consider more stubborn models, we can see that the corresponding $\alpha_t^o(d)$ becomes smaller. So much so that for extremely stubborn models (i.e. the lowest line) $\alpha_t^o(d)$ becomes essentially zero by the 600^{th} instance. This is why we interpret the parameter $\alpha_t^o(d)$ as a measure of external validity: the more externally valid the model, the higher the corresponding $\alpha_t^o(d)$.

An important feature of how we aggregate across models is that it generates a robustness property.

Because $\alpha_t^o(d)$ will place less weight on models that are not externally valid, over time they will have limited influence on the PM's beliefs and consequent decisions. We illustrate this Figure 3. In the top graphs, we plot the policymaker's posterior beliefs about the mean of the potential outcome distributions over time. The plot distinguishes between three posterior means. The bottom (dashed) line corresponds to one set of priors, which we assume to be unbiased (i.e. $\zeta_o^o = \theta$), but diffuse. The top (dash-dotted) line refers to an alternative set of priors, which contains some degree of stubbornness (i.e. $\zeta_o^o = \theta + .5, \nu = 250$). The middle (solid) line comes from the combined model, which is a weighted average of the two sets of priors using $\alpha_t^o(d)$ as weights. We see that even though our policymaker starts with a stubborn prior, the combined model converges relatively quickly to the non-stubborn model. This is the result of both the oracle property – concentrating on the least stubborn model – and robustness property – putting less and less weight on sufficiently stubborn models.

In the bottom graphs, we consider the case in which the alternative model is confident. Thus, both sets of priors are unbiased; the alternative prior simply comes with a higher degree of conviction. Because both priors are correct, the combined model does not immediate converge to one of the models as we saw in case with stubborn priors. As we started in Lemma B.1, our parameter α is more responsive to bias than conviction.

Concentration Bounds

Effects of ϵ. We now simulate our model's concentration bounds and some its key properties. Recall from Remark 3.1 in Section 3.1, the concentration rate increases with the parameter ϵ. We demonstrate this property in the top panel of Figure 4, in which we plot concentration bounds for three different values of $\epsilon \in \{0.1, 0.5, 0.9\}$. That is, for a given ϵ, we compute the difference over time between the policymaker's posterior belief of the true mean, $\zeta_t^o(d)$, and the true mean, $\theta(d)$. We then plot the probability that these differences are greater than 0.1. For these simulations, we assume that our policymaker has correct, but diffuse priors (i.e. $\zeta_0^o = \theta$ and $\nu_0^o = [1, 1]$).

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

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

Effects of Priors. In Figure 5, we investigate the effects of different priors on the concentration bounds. We plot different concentration bounds for priors with different degrees of stubbornness and confidence. For example, in the bottom two lines, we consider two unbiased priors, but with different levels of confidence. According to Remark 3.1, concentration rates increase as the degree of conviction increases and this is precisely what we see. It is also the case, that the concentration rate decreases faster with less stubborn models. We can see this pattern clearly by comparing the top two lines. By comparing the two middle lines, we can also see that conditional on the degree of stubbornness, the higher the bias, the slower the concentration rate. Lastly, as before, the concentration rates for $\theta(1)$ tend to be faster than those for $\theta(0)$ because of the frequency of play.

Probability of Making a Mistake

In Section 3.2, we defined a mistake as recommending a treatment arm different from the one that yields the largest expected effect at the instance in which the experiment was stopped. In Figure 6, we plot the average stopping period (left axis) and the probability of making a mistake at that stopping period (right axis) by ϵ. It is clear from the graph that the more we experiment across treatment arms (i.e., higher ϵ), the faster we stop the experiment. This makes sense. As we experiment more, the data become more IID and we are able to better learn the true means of the potential outcome distributions. According to these simulations, the degree of experimentation does not have to be particularly high. Even though at low levels of ϵ the experiment lasts for almost its entire duration, the drop off is fairly quick. Once ϵ is greater than 0.5, the difference gained in stopping periods from additional experimentation is minimal.

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

In Figure 7, we explore how the initial priors affect the probability of making a mistake. We again consider two sets of priors, both with $\nu_0 = [250, 250]$. One, however, is confident with $\zeta_0^o = \theta$, whereas the other is stubborn, with $\zeta_0^o = [\theta(0)+\delta, \theta(1)-\delta]$, where δ is indicated by a point on the x-axis. For $\delta \in (0, 0.15)$, the priors are biased, but have a proper ranking of the treatment arms. For $\delta > 0.15$, the priors are not only biased, but reverse the ranking of the arms. On the y-axis, we plot the probability of making a mistake associated with each set of priors and for the combined model.

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

Expected Earnings

The final outcome we evaluate is expected earnings. According to Proposition 3.3, the distance between the average outcomes and maximum expected outcome is decreasing in ϵ. In Figure 8, we plot by ϵ, the difference between the policymaker's average impact and the maximum expected outcome, $\max_d \theta(d)$, for an experiment that lasts 1000 instances. The figure also distinguishes between our two familiar sets of priors, a confident one and a stubborn one.

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

The fact that average earnings declines with experimentation does not imply that our policymaker should set ϵ close to zero. Because as we saw in Figure 6, lower ϵ's result in longer experiments, which can come with costs. Moreover, as we show in Proposition 3.2, the upper bound the probability of making a mistake is weakly smaller for higher levels of ϵ. Thus, to properly capture the experimentation versus exploitation tradeoff inherent in multi-armed bandit problems, we need to specify a payoff function.

We consider the following payoff function:

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

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

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

In Figure 9, we compute the payoff function for our model simulations by different values of ϵ. In contrast with the previous figure, we see that the average payoffs are increasing with ϵ until approximately $\epsilon = 0.38$, at which point the payoffs start to decline. While this "optimal" value of ϵ is clearly a function of an arbitrary set of parameter choices, our conjecture is that the inverted u-shape relationship is likely to hold more generally, suggesting that some combination of experimentation and exploitation is optimal.

5 Debt Refinancing Experiment

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

In January 2024, we partnered with a major private bank in Argentina to implement our algorithm for marketing their debt-refinancing program. At that time, Argentina was experiencing a surge in household debt due to high inflation and interest rates. Consequently, the bank, one of the country's largest financial service providers, had over 300,000 clients who were delinquent on their loans and targeted for debt refinancing. These clients are located across the country and are predominantly from lower-income households.

The bank markets these refinancing loans via email, informing clients about the basic terms of the contract and providing a hyperlink to connect them with a bank representative. Getting clients to click on this link is a crucial first step in the refinancing process. As a result, the bank continuously experiments with the content and presentation of these emails to maximize engagement.

In February 2024, the bank ran a randomized control trial on a random sample of around 15,491 clients. The bank was interested in learning whether clients were more likely to click on the refinancing link and refinance their loans if the email displayed the monthly payments associated with their loan instead of the interest rate. With interest rates hovering above 100%, the bank thought that an alternative framing might help avoid with any potential sticker shock (see Appendix Figure 10 for the treatment and control messaging).

We present the results of this experiment in Table 1. Each column is a separate regression, corresponding to a particular region of Argentina. In columns 1-6, the dependent variable is whether the client clicked on the link; whereas, in columns 7-12, the dependent variable is whether the individual refinanced their loan. For ease of interpretation, the dependent variables have been scaled by 100. The independent variable ($1InterestRate$) is an indicator for whether the email provided the interest rate. Otherwise, the email provided just the implied monthly payments (and no interest rate).

From the table, we see that in contrast to the bank's priors, clients are much more likely to click on the link when the interest rate is displayed as opposed to the monthly payments. For example, in the west region of Buenos Aires (BA Zona Oeste) clients are 4.6 percentage points more likely to click on the refinancing link under the interest rate treatment. This effect is more than a doubling of the percentage of clients who clicked on the hyperlink in the control group (i.e., 3.049 percent). The effects are also quite heterogeneous across the different regions, ranging from the 4.6 percentage points BA Zone Oeste to a statistically insignificant 0.31 percentage points BA Zona Norte. It is also the case that clicks are necessary, but not sufficient, for getting the client to refinance. Only when the treatment effects on clicks are sufficiently large, as is the case in BA Zona Oeste or NE, do we see a corresponding treatment effect on refinance rates.

Adaptive Experiment. In June 2024, as interest rates began to fall below 50%, the bank was interested in rerunning their experiment, but adaptively and using our algorithm. This second experiment provided an ideal setting in which to pilot our algorithm. Even though the previous experiment had provided support for an interest rate messaging, it was no longer clear whether the previous treatment effects would extrapolate to the new period given how much the macroeconomy had changed. Moreover, with all the treatment effect heterogeneity across regions, our algorithm offered a robust approach to run the experiment by region, without sacrificing information from the other regions.

For this adaptive experiment, our experimental design consisted of the same two treatment arms as before (interest rate vs monthly payments) but sent out sequentially in batches of approximately forty emails. In the context of our setup, we assumed the policymaker (i.e., the bank) wanted to learn about the average click rate of each treatment arm. For this experiment, we decided to focus on click rate instead of refinancing rates, because the feedback loop for click rates is significantly shorter (1 to 2 days versus 1 to 2 weeks) and the relationship between click rates and refinance rates is both quite clear and stable. For our policy functions, we decided to use the epsilon greedy algorithm with $\epsilon = 0.20$. We chose this value of epsilon based on simulations using data from the previous experiment. The stopping rule is the one specified in our setup, above. We set parameters of the stopping rule such that the probability of making a mistake was no larger than 1 percent. We required that each experiment last for at least 4 batches, but no longer than 14 batches (per the bank's requirement), unless our algorithm stopped the experiment sooner.

The experiment was conducted three regions of Argentina, the capital city of Buenos Aires (CABA), the southern region of Buenos Aires (BA Zona Sur), and the northwest part of the country. Together these regions account for 20 percent of the population of Argentina. For priors, we used the average click rates estimated by treatment using data from the first experiment, but as a stress test of our algorithm excluded the own region (e.g. for the experiment in CABA we did not include the priors from CABA). We did include a "diffuse prior", which sets the average click rate of each treatment arm equal to zero and the variance close to zero. As we described in our setup, each of these other regions is a model that the policymaker will form posterior beliefs over as she accumulates more information from clients living in BA Zona Sur or CABA. She then aggregates these beliefs to determine how to allocate her batch of emails.

In Table 2, we report the results of the experiment. The table reports for each prior source by treatment arm: the posterior belief, the posterior probability the source fits the observed data, and the inverse of the variance. At the bottom of the table, we also report by treatment arm: the policymaker's subjective, aggregated, posterior beliefs at the end of the experiment, the average click rates among clients, and the number of emails sent.

In CABA, the experiment was stopped after sending out 294 emails. Of these, 84.4 percent (248 emails) included information about the interest rate, and 5.2 percent of recipients clicked on the message. In contrast, among the 46 emails that provided information on monthly payments, only 2.1 percent of clients clicked. This difference in click rates (3.07 percent) is slightly larger than the difference in beliefs (2.73 percent) because the initial priors for the control arms were higher.

The results highlight two key aspects of our approach. First, with a 1 percent probability of making a mistake, the policymaker could stop the experiment even though the difference in average click rates between the treatment arms was just 3.07 percent. In a standard randomized controlled trial

(RCT), at least 600 emails would be needed to detect this difference with 95 percent confidence. If we had used the same epsilon-greedy algorithm without incorporating prior information, more than 2,000 emails would have been required to stop the experiment.[18] By incorporating prior information, we effectively increased our sample size, allowing us to stop the experiment earlier.

Second, the table shows how our method combines different data sources. Given the similarities in initial priors and the speed with which we stopped the experiment, the weights assigned to the different regions are similar. The highest weight for the interest rate treatment appears in the Northwest Region, which, according to Definition 1, is the region most externally valid to CABA.

The results in the southern region of Buenos Aires (BA Zona Sur) were qualitatively similar to those in CABA. The experiment stopped after 366 emails. In this region, a standard RCT would have required at least 750 emails, and an epsilon-greedy algorithm without priors would have needed over 2,800 emails. However, the experiment in the Northwest region of Argentina was only stopped because the bank decided to end it. Unlike the other regions, clients in this region responded more to the monthly payment messaging. The average click rate for the monthly payment emails was 5.1 percent, compared to just 3.9 percent for the interest rate emails. Since our priors had suggested the opposite, the experiment did not stop within 14 batches, which would have also been the case in a simple RCT. In hindsight, it would have been interesting to see the experiment's results if we had included priors that had the opposite predictions.

6 Conclusions

This paper presents a conceptual framework for how to incorporate prior sources of information into the design of a sequential experiment. An obvious issue is how to handle the potential lack of external validity of each of these sources. We address this issue by first presenting a formal definition of external validity that can be used to differentiate sources with different degrees of external invalidity and second, by showing that our framework is robust to including externally-invalid sources. This last property relaxes the burden on the policymaker of having to correctly choose relevant sources of information based on limited ex-ante information. As "stubborn" sources are harder to discard, we believe it is useful to incorporate many priors, including versions that are diffuse.

For the common problem of learning about average treatment effects, we show that our framework offers several nice properties. As we illustrated for the case of a debt refinancing program, these

[18] This calculation assumes that the sample means remain constant with additional batches.

properties translate into substantial gains in performance — such as reducing the duration of experiment and increasing the average payoffs while keeping an acceptable probability of making a mistake — over both standard RCTs and adaptive experiments.

References

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

Al-Najjar, N. I. and Weinstein, J. (2009). The ambiguity aversion literature: a critical assessment. *Economics & Philosophy*, 25(3):249–284.

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

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

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

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

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

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

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

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

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

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

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

Epstein, L. G. and Schneider, M. (2007). Learning under ambiguity. *The Review of Economic Studies*, 74(4):1275–1303.

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

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

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

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

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

Klibanoff, P., Marinacci, M., and Mukerji, S. (2009). Recursive smooth ambiguity preferences. *Journal of Economic Theory*, 144(3):930–976.

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

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

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

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

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

Robbins, H. E. (1992). *An Empirical Bayes Approach to Statistics*, pages 388–394. Springer New York.

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

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

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

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

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

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

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

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

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

Appendix: Figures & Tables

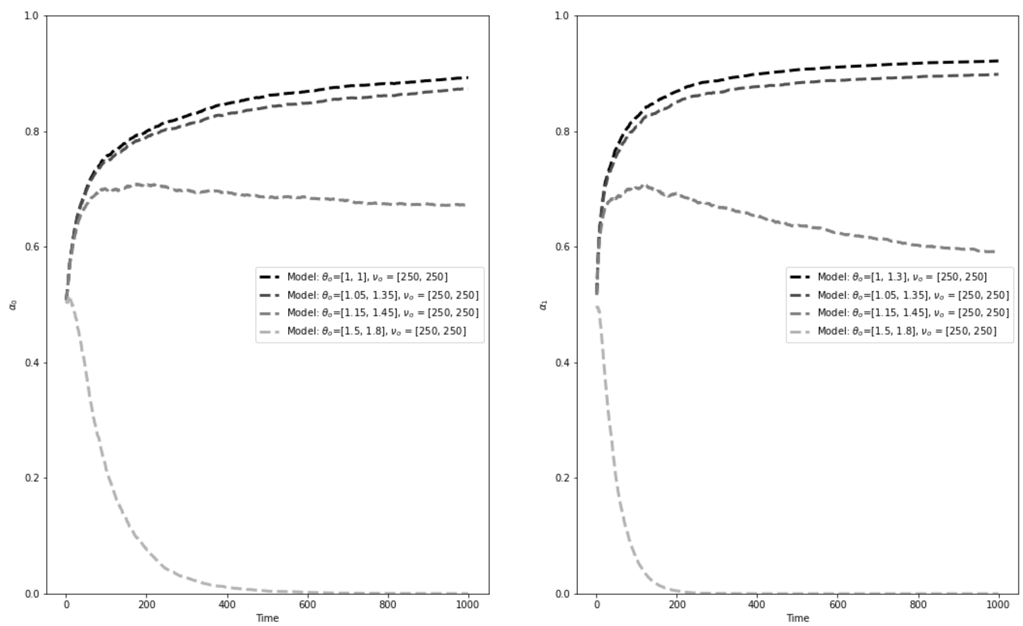


Figure 2: External Validity - α_t^o

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

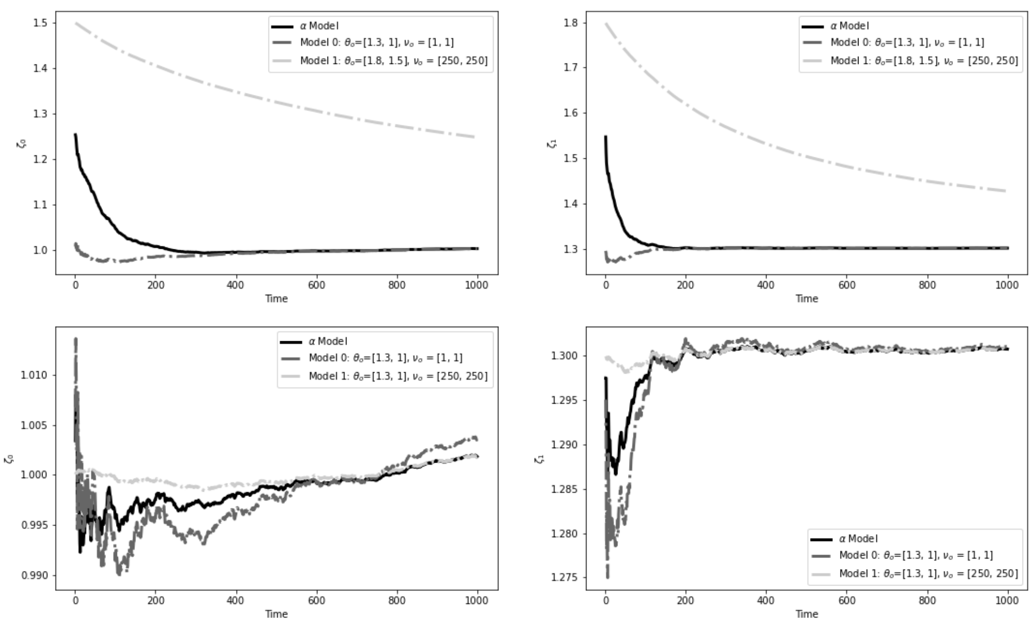


Figure 3: Posterior Beliefs Over Time, Holding Behavior Constant

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

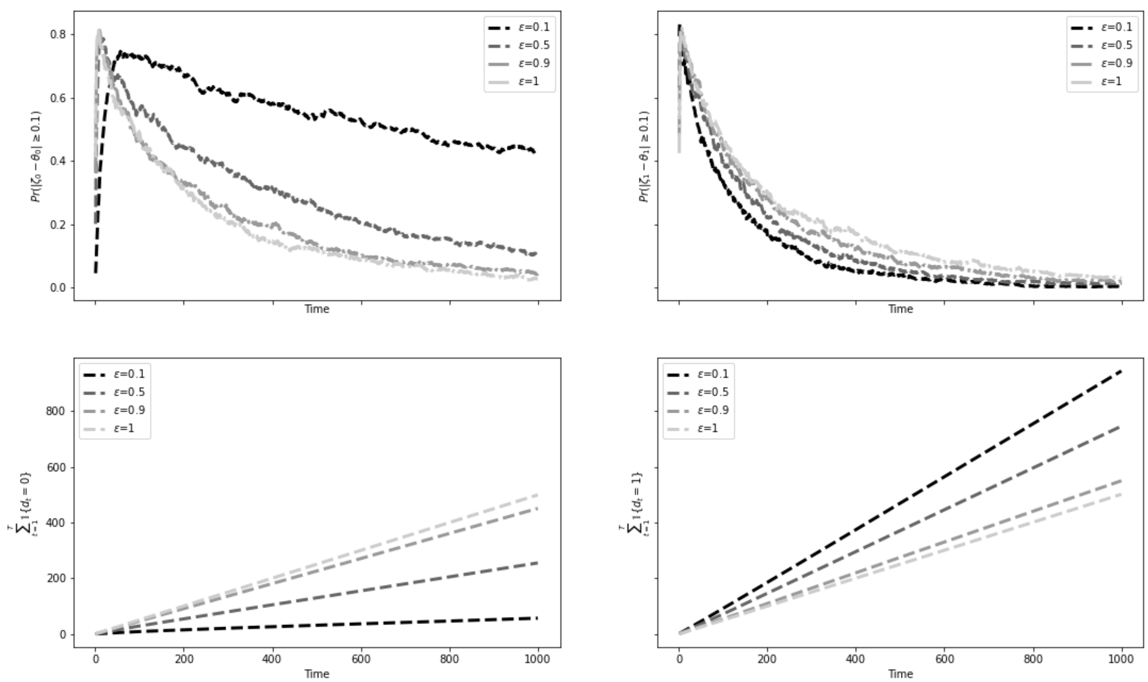


Figure 4: Concentration Bounds and Frequency of Play

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

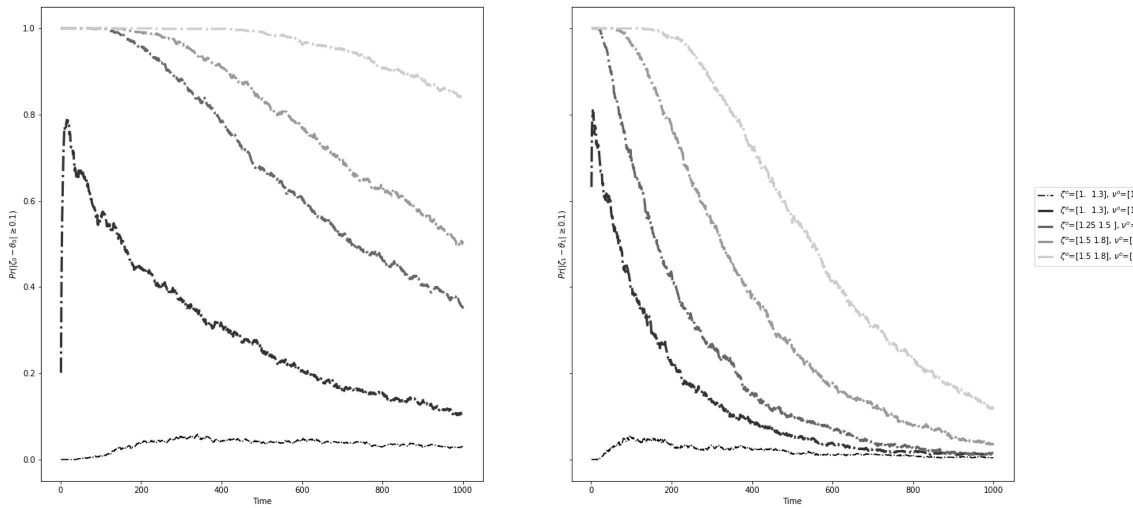


Figure 5: Concentration Bounds by Model Stubbornness

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: $\theta = [1, 1.3]$, ϵ= 0.5. The initial priors are specified in the legend.

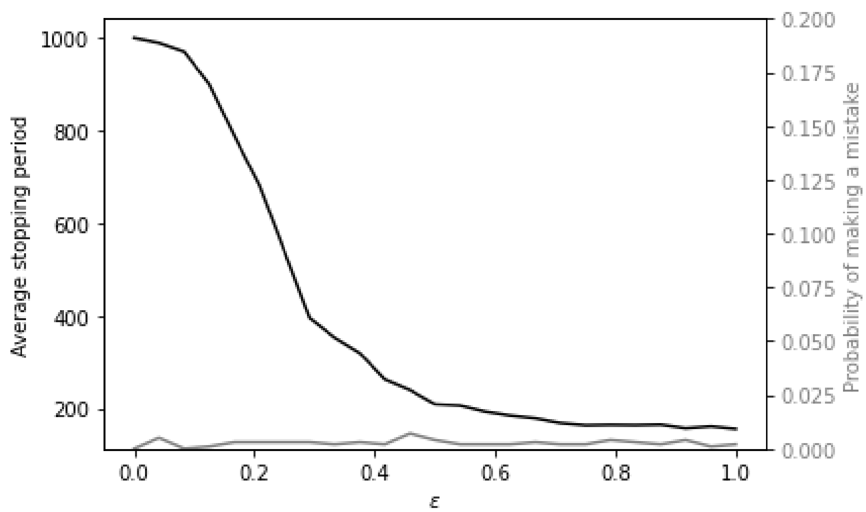


Figure 6: Stopping Period and Probability of Making a Mistake

Notes: This figure plots the average stopping period (left axis) and the probability of making a mistake at the stopping period (right axis) by ϵ. These figures are based on 1,000 simulations using the following parameters: $\theta = [1, 1.3]$; $\zeta_0^o = \theta$; $\zeta_1^o = \theta$; $\nu_0^o = [1, 1]$; $\nu_1^o = [1, 1]$; $B = 100$.

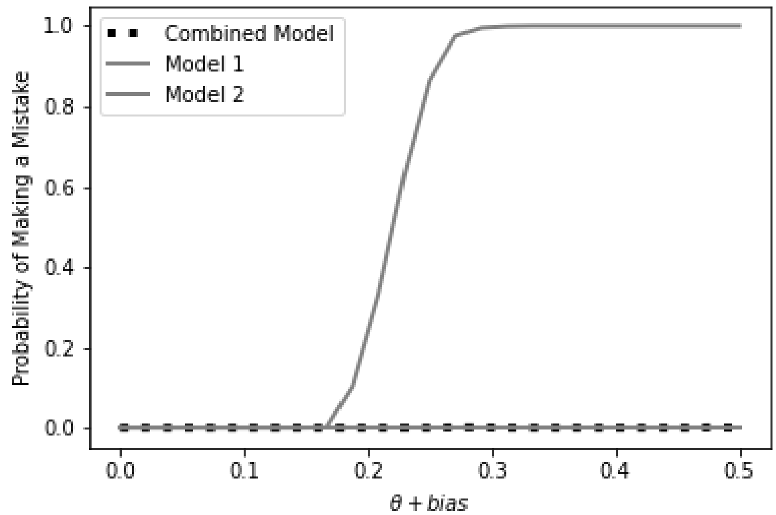


Figure 7: Probability of Making a Mistake by Model Bias

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: $\theta = [1, 1.3]$; $\nu_0^0 = \nu_0^1 = [250, 250]$; $\zeta_0^0 = [\theta(0) + bias, \theta(1) - bias]$; $\zeta_0^1 = \theta$, ϵ= 0.5.

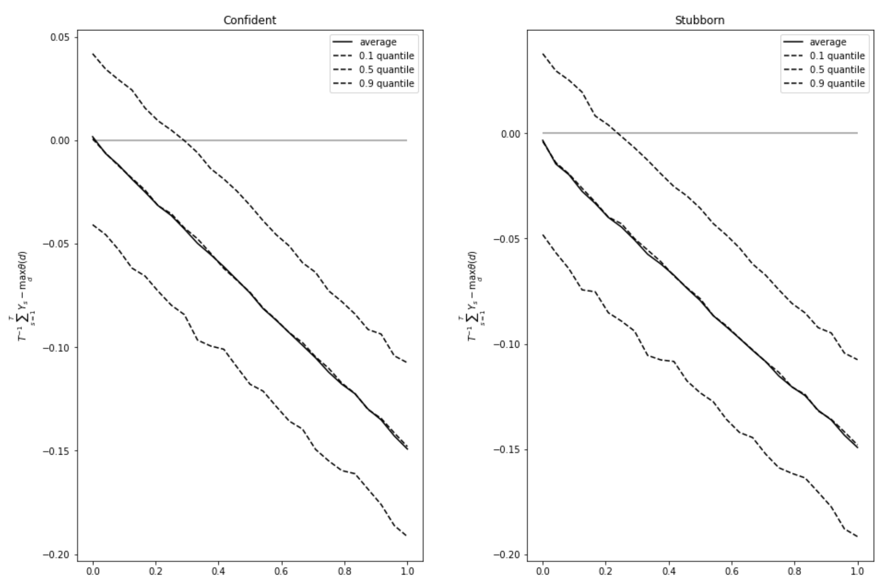


Figure 8: Relative Average Earnings During the Experiment

Notes: This figure plots by ϵ, the average earnings net of maximal earnings. These figures are based on 1,000 simulations using the following parameters: $\theta = [1, 1.3]$; $\zeta_0^o = \theta$; $\zeta_1^o = \theta$; $\nu_0^o = [1, 1]$; $\nu_1^o = [1, 1]$.

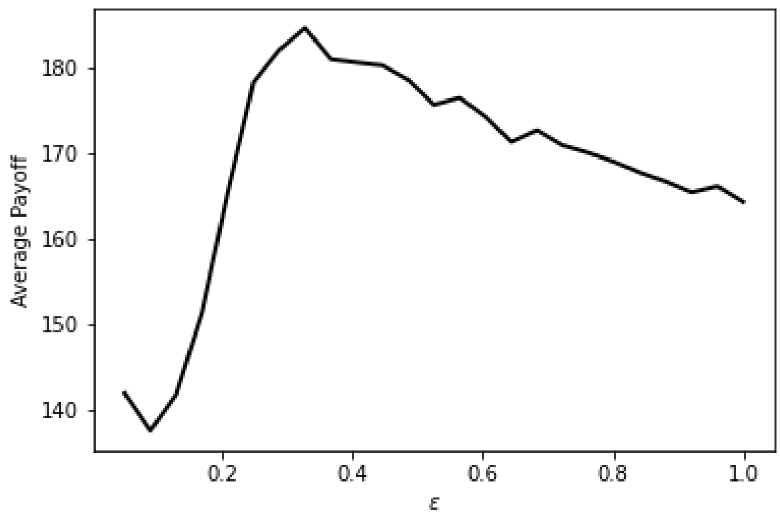
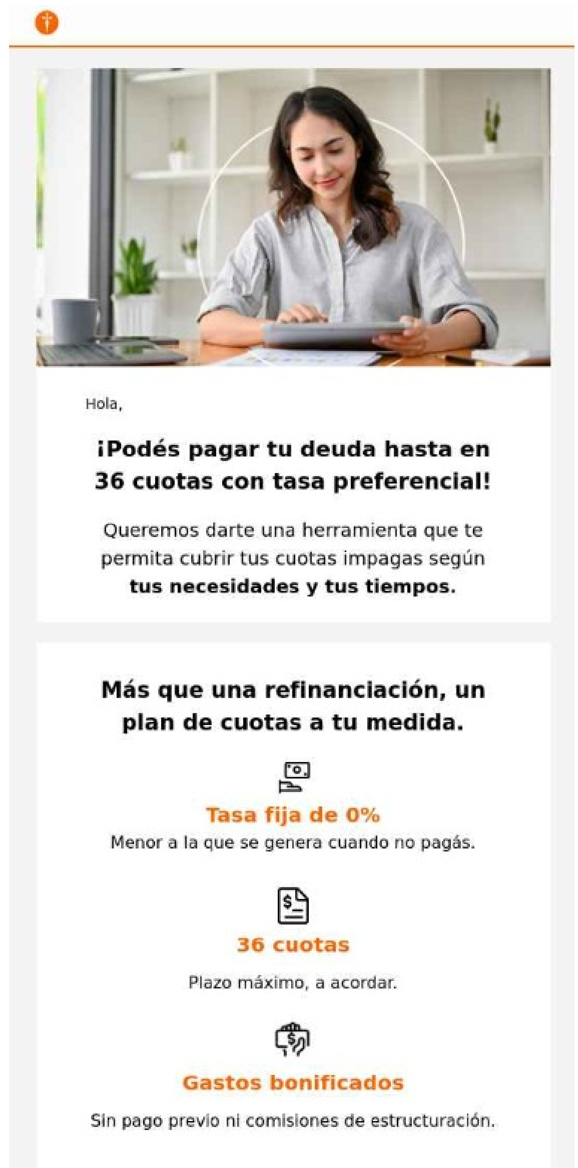


Figure 9: Experimentation versus Exploitation – Expected Payoffs

Notes: This figure plots by ϵ, the expected payoffs as defined by Equation 4.1. These figures are based on 1,000 simulations using the following parameters: $\theta = [1, 1.3]$; $\zeta_0^o = \theta$; $\zeta_1^o = \theta$; $\nu_0^o = [1, 1]$; $\nu_1^o = [1, 1]$; $B = 100$; $\beta^t = 0.994$; $c = 1.15$; $\lambda = 1,100$.

(a) Monthly Payments



(b) Interest Rate

Figure 10: Monthly Payment Messaging vs Interest Rate Messaging

Table 1: Debt Refinancing Experiment - RCT

	BA Zona Norte	BA Zona Oeste	BA Zona Sur	CABA	NE	NW
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A: Clicked on email hyper link					
$1\{InterestRate\}$	0.311	4.619	3.409	1.311	3.886	2.934
	(1.307)	(1.056)*	(0.940)*	(0.872)	(0.826)*	(0.751)*
Constant term	4.251	3.049	3.295	4.17	2.483	3.442
	(0.955)*	(0.601)*	(0.556)*	(0.596)*	(0.434)*	(0.444)*
Sample Size	995	1759	2121	2404	2498	3316
	Panel B: Refinanced Loan					
$1\{InterestRate\}$	-0.919	2.113	0.378	0.047	1.302	0.478
	(0.800)	(0.661)*	(0.567)	(0.516)	(0.502)*	(0.453)
Constant term	2.013	0.976	1.55	1.597	0.931	1.484
	(0.665)*	(0.343)*	(0.385)*	(0.374)*	(0.268)*	(0.295)*
Sample Size	995	1759	2121	2404	2498	3316

Notes: This table reports the results of the debt refinancing randomized control trial. In Panel A, we report the results on click rates and in Panel, we report the results on refinancing rates. Each column reports the results of separate regressions, corresponding to a particular region of Argentina. For ease of interpretation, the dependent variables have been scaled by 100. The independent variable ($1InterestRate$) is an indicator for whether the email provided the interest rate. Robust standard errors are reported in parentheses. * indicates that the estimated coefficient is statistically different from zero at 95 percent confidence.

Table 2: Debt Refinancing Experiment - Multi-prior Adaptive Experiment

REGION: CABA

	Monthly Payments			Interest Rates		
Sources	ζ_t^l	α_t^l	ν_t^l	ζ_t^l	α_t^l	ν_t^l
Buenos Aires- Otras regiones	0.037	0.167	1284.0	0.055	0.172	1408.0
Buenos Aires- Zona Norte	0.041	0.162	493.0	0.048	0.157	796.0
Buenos Aires- Zona Oeste	0.030	0.167	866.0	0.072	0.159	1187.0
Buenos Aires- Zona Sur	0.032	0.167	1078.0	0.064	0.167	1337.0
NE	0.025	0.168	1335.0	0.062	0.170	1457.0
NW	0.034	0.169	1731.0	0.062	0.174	1879.0
Diffuse	0.022	0.000	46.0	0.052	0.000	248.0
ζ_t^α	0.033			0.061		
Outcome Averages	0.022			0.052		
Sample Size	46			248		

REGION: BUENOS AIRES - ZONA SUR

	Monthly Payments			Interest Rates		
Sources	ζ_t^l	α_t^l	ν_t^l	ζ_t^l	α_t^l	ν_t^l
Buenos Aires- Otras regiones	0.037	0.168	1291.0	0.053	0.174	1473.0
Buenos Aires- Zona Norte	0.040	0.161	500.0	0.044	0.160	861.0
Buenos Aires- Zona Oeste	0.030	0.167	873.0	0.068	0.150	1252.0
CABA	0.041	0.167	1180.0	0.052	0.176	1590.0
NE	0.025	0.169	1342.0	0.059	0.168	1522.0
NW	0.034	0.169	1738.0	0.060	0.172	1944.0
Diffuse	0.019	0.000	53.0	0.042	0.000	313.0
ζ_t^α	0.034			0.056		
Outcome Averages	0.019			0.042		
Sample Size	53			313		

REGION: Northwest

	Monthly Payments			Interest Rates		
Sources	ζ_t^l	α_t^l	ν_t^l	ζ_t^l	α_t^l	ν_t^l
Buenos Aires- Otras regiones	0.039	0.170	1336.0	0.049	0.183	1971.0
Buenos Aires- Zona Norte	0.044	0.161	545.0	0.042	0.160	1359.0
Buenos Aires- Zona Oeste	0.033	0.165	918.0	0.059	0.138	1750.0
Buenos Aires- Zona Sur	0.035	0.168	1130.0	0.055	0.161	1900.0
CABA	0.042	0.170	1225.0	0.049	0.187	2088.0
NE	0.027	0.166	1387.0	0.054	0.171	2020.0
Diffuse	0.051	0.000	98.0	0.039	0.000	811.0
ζ_t^α	0.037			0.051		
Outcome Averages	0.051			0.039		
Sample Size	98			811		

Notes: This table reports the results of the debt refinancing experiment using our multi- prior adaptive algorithm. In column 2-4, we report outcomes for the monthly payment experimental arm, and in columns 5-7 we report the outcomes for the interest rate experimental arm. For the stopping rule, we set $B = 200$ and a probability of making a mistake below 1%. We used an ϵ-greedy algorithm that set $\epsilon = 0.20$.

Online Appendix

A Notation and some definitions

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

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

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

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

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

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

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

B Proof of Proposition 2.1

The proof of the proposition uses the following lemma, whose proof is relegated to the end of the section.

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

$$\alpha_t^o(d,x) = \frac{\phi(m_t(d,x);\zeta_0^o(d,x),(N_t(d,x)+\nu_0^o(d,x))/(N_t(d,x)\nu_0^o(d,x)))}{\sum_{o=0}^{L}\phi(m_t(d,x);\zeta_0^o(d,x),(N_t(d,x)+\nu_0^o(d,x))/(N_t(d,x)\nu_0^o(d,x)))},$$

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

Proof of Proposition 2.1. (1) To show this part we note that $\nu_0^o(d,x) > 0$ and thus $\mathbb{E}\mathbb{V}_{d,x}(o) \to -\infty$ iff the bias diverges to plus infinity. From the characterization In Lemma B.1 and the fact that $\phi(m_t(d,x);\zeta_0^o(d,x),(N_t(d,x)+\nu_0^o(d,x))/(N_t(d,x)\nu_0^o(d,x))) = \phi(m_t(d,x)-\theta(d,x);\zeta_0^o(d,x)-\theta(d,x),(N_t(d,x)+\nu_0^o(d,x))/(N_t(d,x)\nu_0^o(d,x)))$ will vanish if the bias diverges to plus infinity, the desired result is obtained.

(2) Observe that

$$\begin{aligned}&\log\phi(m_t(d,x);\zeta_0^o(d,x),(N_t(d,x)+\nu_0^o(d,x))/(N_t(d,x)\nu_0^o(d,x)))\\&= C-0.5\log\left((N_t(d,x)+\nu_0^o(d,x))/(N_t(d,x)\nu_0^o(d,x))\right)-0.5\frac{\left(m_t(d,x)-\zeta_0^o(d,x)\right)^2}{(N_t(d,x)+\nu_0^o(d,x))/(N_t(d,x)\nu_0^o(d,x))}\end{aligned}$$

where C is some universal constant.

If $\inf_t t^{-1}\sum_{s=1}^t \delta_s(d|x) > 0$, by LLN $N_t(d,x)$ will diverge with probability approaching 1 and $m_t(d,x) = \theta(d,x)+o_{\mathbf{P}}(1)$. Therefore,

$$\begin{aligned}&\log\phi(m_t(d,x);\zeta_0^o(d,x),(N_t(d,x)+\nu_0^o(d,x))/(N_t(d,x)\nu_0^o(d,x)))\\=&C-0.5\log\left(1/\nu_0^o(d,x))\right)-0.5\left(\theta(d,x)-\zeta_0^o(d,x)\right)^2\nu_0^o(d,x))+o_{\mathbf{P}}(1)\\=&C+0.5\mathbb{EV}_{(d,x)}(o)+o_{\mathbf{P}}(1)\end{aligned}$$

where the last line follows from the expression in Definition 1. Hence, for any $o\in\{1,\dots,L\}$,

$$\alpha_t^o(d,x) = \frac{e^{0.5\mathbb{EV}_{(d,x)}(o)}}{\sum_{o'=0}^L e^{0.5\mathbb{EV}_{(d,x)}(o')}}+o_{\mathbf{P}}(1).$$

Moreover, this expression implies

$$\frac{\alpha_t^{o'}(d,x)}{\alpha_t^o(d,x)} = e^{-0.5(\mathbb{EV}_{(d,x)}(o)-\mathbb{EV}_{(d,x)}(o'))}+o_{\mathbf{P}}(1).$$

and thus, given definition 2, if o is more externally valid than o', $\mathbb{EV}_{(d,x)}(o)-\mathbb{EV}_{(d,x)}(o') > 0$ so that $\frac{\alpha_t^{o'}(d,x)}{\alpha_t^o(d,x)} < 1+o_{\mathbf{P}}(1)$. If o is the (only) externally valid source, then according to our definition 2, $\mathbb{EV}_{(d,x)}(o) = \infty$ and with the convention that $c/\infty = 0$ for any number c, the result follows.

□

B.1 Proofs of Supplemental Lemmas

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

$$\begin{aligned}
&\int \prod_{s=1}^{t} (p_\theta(Y_s))^{1\{D_s(x)=d\}} \mu_0^o(d,x)(d\theta) \\
=&\int (2\pi)^{-0.5\sum_{s=1}^{t} 1\{D_s(x)=d\}} \exp\left\{-\frac{1}{2}\sum_{s=1}^{t} 1\{D_s(x)=d\}\left(Y_s(d,x)-m_t(d,x)\right)^2\right\} \\
&\times \exp\left\{-\frac{1}{2}\sum_{s=1}^{t} 1\{D_s(x)=d\}\left(m_t(d,x)-\theta\right)^2\right\} \\
&\times \exp\left\{-\sum_{s=1}^{t} 1\{D_s(x)=d\}\left(Y_s(d,x)-m_t(d,x)\right)\left(m_t(d,x)-\theta\right)\right\} \phi(\theta;\zeta_0^o(d,x),1/\nu_0^o(d,x))d\theta.
\end{aligned}$$

Observe that $\sum_{s=1}^{t} 1\{D_s(x)=d\}\left(Y_s(d,x)-m_t(d,x)\right)=0$, so, letting $N_t(d,x) := \sum_{s=1}^{t} 1\{D_s(x)=d\}$ it follows that

$$\begin{aligned}
\int \prod_{s=1}^{t} (p_\theta(Y_s))^{1\{D_s(x)=d\}} \mu_0^o(d,x)(d\theta) =& \frac{\exp\left\{-\frac{1}{2}\sum_{s=1}^{t} 1\{D_s(x)=d\}\left(Y_s(d,x)-m_t(d,x)\right)^2\right\}}{(2\pi)^{0.5\sum_{s=1}^{t} 1\{D_s(x)=d\}+0.5} N_t(d,x)^{0.5}} \\
&\times \int (2\pi/N_t(d,x))^{-1/2} \exp\left\{-\frac{1}{2}\left(m_t(d,x)-\theta\right)^2 N_t(d,x)\right\} \\
&\phi(\theta;\zeta_0^o(d,x),1/\nu_0^o(d,x))d\theta.
\end{aligned}$$

The expression of the integral can be viewed as a convolution between to Gaussian PDFs one indexed by $(0,1/N_t(d,x))$ and $(\zeta_0^o(d,x),1/\nu_0^o(d,x))$ resp, which in turn is equivalent to PDF of the sum of the corresponding random variables evaluated at $m_t(d,x)$. Therefore,

$$\int \prod_{s=1}^{t} (p_\theta(Y_s))^{1\{D_s(x)=d\}} \mu_0^o(d,x)(d\theta) = C\phi(m_t(d,x);\zeta_0^o(d,x),(N_t(d,x)+\nu_0^o(d,x))/(N_t(d,x)\nu_0^o(d,x)))$$

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

Hence

$$\alpha_t^o(d,x) = \frac{\phi(m_t(d,x);\zeta_0^o(d,x),(N_t(d,x)+\nu_0^o(d,x))/(N_t(d,x)\nu_0^o(d,x)))}{\sum_{o=0}^{L}\phi(m_t(d,x);\zeta_0^o(d,x),(N_t(d,x)+\nu_0^o(d,x))/(N_t(d,x)\nu_0^o(d,x)))}$$

and the desired result follows. $\square$

C Non-stochastic Bounds

C.1 Non-stochastic Bounds for α_t

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

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

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

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

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

Lemma C.1. *For any $o \in \{0,...,L\}$, any $t \in \{1,...,T\}$, any $(d,x) \in \mathbb{D} \times \mathbb{X}$ and any $\eta \in (0, h_t(d,x))$ and $\delta > 0$ such that $f_t(d,x) - h_t(d,x) \geq -\eta$ and $|J_t(d,x) - f_t(d,x)\theta(d,x)| \leq \delta$, it follows that*

$$\underline{\alpha}_t^o(\delta, h_t(d,x) - \eta, |\bar{\zeta}_0(d,x)|, \nu_0(d,x)) \leq \alpha_t^o(d,x) \leq \overline{\alpha}_t^o(\delta, h_t(d,x) - \eta, |\bar{\zeta}_0(d,x)|, \nu_0(d,x)).$$

Proof of Lemma C.1 . By Lemma B.1 it suffices to bound $\phi(J_t(d,x)/f_t(d,x);\zeta_0^o(d,x),(N_t(d,x)+\nu_0^o(d,x))/(N_t(d,x)\nu_0^o(d,x$
To do this, note that

$$\begin{aligned}
&\phi(J_t(d,x)/f_t(d,x);\zeta_0^o(d,x),(N_t(d,x)+\nu_0^o(d,x))/(N_t(d,x)\nu_0^o(d,x))) \\
&= \frac{\sqrt{(N_t(d,x)\nu_0^o(d,x))}}{2\pi\sqrt{(N_t(d,x)+\nu_0^o(d,x))}} \exp\left\{-0.5\frac{\left(J_t(d,x) - f_t(d,x)\zeta_0^o(d,x)\right)^2}{f_t(d,x)} \frac{N_t(d,x)\nu_0^o(d,x)}{f_t(d,x)(N_t(d,x)+\nu_0^o(d,x))}\right\} \\
&= \frac{\sqrt{f_t(d,x)\nu_0^o(d,x)}}{2\pi\sqrt{(f_t(d,x)+\nu_0^o(d,x)/t)}} \exp\left\{-0.5\frac{\left(\bar{J}_t(d,x) - f_t(d,x)\bar{\zeta}_0^o(d,x)\right)^2}{f_t(d,x)} \frac{\nu_0^o(d,x)}{(f_t(d,x)+\nu_0^o(d,x)/t)}\right\}
\end{aligned}$$

where $\bar{\cdot}$ indicates centered at $\theta(d,x)$. Henceforth, let $\sigma_t^2 := (f_t(d,x) + \nu_0^o(d,x)/t)/(f_t(d,x)\nu_0^o(d,x))$.

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

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

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

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

Also, under these conditions,

$$\begin{aligned}\sigma_t^2 \geq&(1+\nu_0^o(d,x)/t)/(\nu_0^o(d,x)) \geq (1+\nu_0^o(d,x)/T)/(\nu_0^o(d,x)) =: \underline{\sigma}_t^2\\ \leq&(h_t(d,x)-\eta+\nu_0^o(d,x)/t)/((h_t(d,x)-\eta)\nu_0^o(d,x)) =: \overline{\sigma}_t^2.\end{aligned}$$

Therefore,

$$\begin{aligned}\log\phi(\bar{m}_t(d,x) - \bar{\zeta}_0^o(d,x);0,\sigma_t^2(d,x)) \geq& -\log\sigma_t - 0.5\frac{(\delta+|\bar{\zeta}_0^o(d,x)|)^2}{f_t(d,x)^2\sigma_t^2} + Cte \geq -\log\overline{\sigma}_t - 0.5\frac{(\delta+|\bar{\zeta}_0^o(d,x)|)^2}{(h_t(d,x)-\eta)^2\underline{\sigma}_t^2} + Cte\\ =:&\underline{\ell}_t(\delta, h_t(d,x)-\eta, |\bar{\zeta}_0^o(d,x)|, \nu_0^o(d,x)) + Cte\end{aligned}$$

and

$$\begin{aligned}\log\phi(\bar{m}_t(d,x) - \bar{\zeta}_0^o(d,x);0,\sigma_t^2(d,x)) \leq& -\log\sigma_t - 0.5\frac{\max\{(h_t(d,x)-\eta))^2(\bar{\zeta}_0^o(d,x))^2 - 2\delta|\bar{\zeta}_0^o(d,x)|, 0\}}{f_t(d,x)^2\sigma_t^2} + Cte\\ \leq& -\log\underline{\sigma}_t - 0.5\frac{\max\{(h_t(d,x)-\eta))^2(\bar{\zeta}_0^o(d,x))^2 - 2\delta|\bar{\zeta}_0^o(d,x)|, 0\}}{\overline{\sigma}_t^2} + Cte\\ =:&\overline{\ell}_t(\delta, h_t(d,x)-\eta, |\bar{\zeta}_0^o(d,x)|, \nu_0^o(d,x)).\end{aligned}$$

□

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

1. $\delta \mapsto \underline{\ell}_t(\delta, g, |\bar{\zeta}_0^o(d,x)|, \nu_0^o(d,x))$ *is decreasing and* $\delta \mapsto \overline{\ell}_t(\delta, g, |\bar{\zeta}_0^o(d,x)|, \nu_0^o(d,x))$ *is non-decreasing.*
2. $g \mapsto \underline{\ell}_t(\delta, g, |\bar{\zeta}_0^o(d,x)|, \nu_0^o(d,x))$ *is increasing and* $g \mapsto \overline{\ell}_t(\delta, g, |\bar{\zeta}_0^o(d,x)|, \nu_0^o(d,x))$ *is decreasing.*

3. $\delta \mapsto \overline{\alpha}_t^o(\delta, g, |\zeta_0(d,x)|, \nu_0(d,x))$ *is increasing and* $\delta \mapsto \overline{\alpha}_t^o(\delta, g, |\zeta_0(d,x)|, \nu_0(d,x))$ *is decreasing.*

4. $g \mapsto \overline{\alpha}_t^o(\delta, g, |\zeta_0(d,x)|, \nu_0(d,x))$ *is decreasing and* $g \mapsto \underline{\alpha}_t^o(\delta, g, |\zeta_0(d,x)|, \nu_0(d,x))$ *is increasing.*

5. $t \mapsto \overline{\alpha}_t^o(\delta, g, |\zeta_0(d,x)|, \nu_0(d,x))$ *is increasing and* $t \mapsto \underline{\alpha}_t^o(\delta, g, |\zeta_0(d,x)|, \nu_0(d,x))$ *is decreasing.*

Proof of Lemma C.2. (1) It is easy to see that $\underline{\ell}_t(\delta, g, |\bar{\zeta}_0^o(d,x)|, \nu_0^o(d,x), h_t(d,x))$ is decreasing in δ and $\overline{\ell}_t(\delta, g, |\bar{\zeta}_0^o(d,x)|, \nu_0^o(d,x))$ is non-decreasing in δ.

(2) We first observe that $\underline{\sigma}_t^2$ is constant as a function of $g = h_t(d,x) - \eta$ and $\overline{\sigma}_t^2$ is an decreasing function of $g := h_t(d,x) - \eta$. Also, note that $\frac{d\underline{\ell}_t(\delta,g,|\bar{\zeta}_0^o(d,x)|,\nu_0^o(d,x))}{dg} = -\frac{1}{\overline{\sigma}}\frac{d\overline{\sigma}_t}{dg} + \frac{(\delta+|\bar{\zeta}_0^o(d,x)|)^2}{(g)^3\underline{\sigma}_t^2}$, thus, since $g \geq 0$, $\underline{\ell}_t(\delta, g, |\bar{\zeta}_0^o(d,x)|, \nu_0^o(d,x))$ is increasing as a function of g. Similarly, $\overline{\ell}_t(\delta, g, |\bar{\zeta}_0^o(d,x)|, \nu_0^o(d,x))$ is increasing as a function of $\overline{\sigma}_t^2$ and decreasing as a direct function of g, thus by computing the derivative it can be shown that it is decreasing in g.

(3-4) By parts (1), it readily follows that $\overline{\alpha}_t^o(\delta, g, |\zeta_0(d,x)|, \nu_0(d,x))$ is increasing in δ and $\underline{\alpha}_t^o(\delta, g, |\zeta_0(d,x)|, \nu_0(d,x))$ is decreasing in δ. And by part (2) $\overline{\alpha}_t^o(\delta, g, |\zeta_0(d,x)|, \nu_0(d,x))$ is decreasing in g and $\underline{\alpha}_t^o(\delta, g, |\zeta_0(d,x)|, \nu_0(d,x))$ is increasing in g.

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

These results imply that $t \mapsto \overline{\alpha}_t^o(\delta, g, |\zeta_0(d,x)|, \nu_0(d,x))$ is increasing and $t \mapsto \underline{\alpha}_t^o(\delta, g, |\zeta_0(d,x)|, \nu_0(d,x))$ is decreasing. ◻

C.2 Non-stochastic Bounds for ζ_t^o

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

$$\Omega(\gamma, g, \zeta_0^o(d,x), \nu_0^o(d,x)) := \frac{\gamma}{g + \nu_0^o(d,x)/t} + \Omega_0(g, \zeta_0^o(d,x), \nu_0^o(d,x))$$

and

$$\Omega_0(g, \zeta_0^o(d,x), \nu_0^o(d,x)) := \nu_0^o(d,x)\left(\frac{(\zeta_0^o(d,x))_+/t}{g + \nu_0^o(d,x)/t} + \frac{(\zeta_0^o(d,x))_-/T}{1 + \nu_0^o(d,x)/t}\right)$$

for any $(\gamma, g, \zeta_0^o(d,x), \nu_0^o(d,x)) \in D(\Omega)$, where for any real number a, $a_+ := \max\{a, 0\}$ and $a_- := \min\{a, 0\}$.

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

1. $\gamma \mapsto \Omega(\gamma, g, \zeta_0^o(d,x), \nu_0^o(d,x))$ *is increasing.*

2. $g \mapsto \Omega(\gamma, g, \zeta_0^o(d,x), \nu_0^o(d,x))$ *is decreasing and* $g \mapsto \Omega_0(g, \zeta_0^o(d,x), \nu_0^o(d,x))$ *is non-increasing.*

3. *If* $\zeta_0(d,x) \leq 0$, $\nu_0(d,x) \mapsto \Omega(\gamma, g, \zeta_0^o(d,x), \nu_0^o(d,x))$ *is decreasing; and if* $\zeta_0(d,x) \leq (\geq) 0$ $\nu_0(d,x) \mapsto \Omega_0(g, \zeta_0^o(d,x), \nu_0^o(d,x))$ *is non-increasing (non-decreasing).*

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

(2) By inspection, $g \mapsto \Omega_0(g, \zeta_0^o(d,x), \nu_0^o(d,x))$ is non-increasing. In addition $g \mapsto \gamma/(g + \nu_0^o(d,x)/t)$ is decreasing.

(3) Trivial. $\square$

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

1. $|\zeta_t^o(d,x) - \theta(d,x)| \leq \Omega(\gamma, h_t(d,x) - \eta, |\zeta_0^o(d,x) - \theta(d,x)|, \nu_0^o(d,x))$.

2. $\zeta_t^o(d,x) - \theta(d,x) \leq \Omega(\gamma, h_t(d,x) - \eta, \zeta_0^o(d,x) - \theta(d,x), \nu_0^o(d,x))$.

3. $-(\zeta_t^o(d,x) - \theta(d,x)) \leq \Omega(\gamma, h_t(d,x) - \eta, -(\zeta_0^o(d,x) - \theta(d,x)), \nu_0^o(d,x))$.

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

$$
\begin{aligned}
|\zeta_t^o(d,x) - \theta(d,x)| \leq \frac{\gamma + |\zeta_0^o(d,x) - \theta(d,x)| \nu_0^o(d,x)/t}{\iota_t(d,x) - \eta + \nu_0^o(d,x)/t} \leq & \frac{\gamma + |\zeta_0^o(d,x) - \theta(d,x)| \nu_0^o(d,x)/t}{h_t(d,x) - \eta + \nu_0^o(d,x)/t} \\
= & \Omega(\gamma, h_t(d,x) - \eta, |\zeta_0^o(d,x) - \theta(d,x)|, \nu_0^o(d,x)).
\end{aligned}
$$

(2-3) The proof is analogous and thus omitted. $\square$

C.3 Non-stochastic Bounds for ζ_t^α

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

$$
\begin{aligned}
\Gamma_0(\gamma, g, \zeta_0(d,x), \nu_0(d,x)) := & \sum_{o=0}^{L} \overline{\alpha}_t^o(\gamma, g, \zeta_0(d,x), \nu_0(d,x)) \Omega_0^+(g, \zeta_0^o(d,x), \nu_0^o(d,x)) \\
& + \sum_{o=0}^{L} \underline{\alpha}_t^o(\gamma, g, \zeta_0(d,x), \nu_0(d,x)) \Omega_0^-(g, \zeta_0^o(d,x), \nu_0^o(d,x))
\end{aligned}
$$

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

$$
\Gamma(\gamma, g, \zeta_0(d,x), \nu_0(d,x)) := \gamma \sum_{o=0}^{L} \frac{\overline{\alpha}_t^o(\gamma, g, \zeta_0(d,x), \nu_0(d,x))}{g + \nu_0^o(d,x)/t} + \Gamma_0(\gamma, g, \zeta_0(d,x), \nu_0(d,x))
$$

for any $(\gamma, g, \zeta_0(d,x), \nu_0(d,x)) \in D(\Gamma)$.

Remark C.1. *Another possible formulation for $\Gamma(\gamma, g, \zeta_0(d,x), \nu_0(d,x))$ is $\max_{o\in\{0,\ldots,L\}} \Omega(\gamma, g, \zeta_0^o(d,x), \nu_0^o(d,x))$, so one can define Γ as the minimum of this expression and the one above, and depending on the context one can use one bound or the other. The same applies to Γ_0 and Ω_0. For the sake of the exposition, however, we do not make this bound explicit.* $\triangle$

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

1. *$\delta \mapsto \Gamma(\delta, g, \zeta_0(d,x), \nu_0(d,x))$ is increasing and $\delta \mapsto \Gamma(\delta, g, \zeta_0(d,x), \nu_0(d,x))$ is non-decreasing.*
2. *$g \mapsto \Gamma(\delta, g, \zeta_0(d,x), \nu_0(d,x))$ and $g \mapsto \Gamma_0(\delta, g, \zeta_0(d,x), \nu_0(d,x))$ are decreasing.*
3. *For any positive sequences $(\delta_t, g_t)_t$, $\Gamma(\delta_t, g_t, \zeta_0(d,x), \nu_0(d,x)) = O\left(\frac{\delta_t + t^{-1}}{g_t + t^{-1}}\right)$ and $\Gamma_0(\delta_t, g_t, \zeta_0(d,x), \nu_0(d,x)) = O\left(\frac{t^{-1}}{g_t + t^{-1}}\right)$.*

Proof of Lemma C.5. (1) (we only establish the results for Γ as for Γ_0 is analogous) $\Gamma(\delta, g, \zeta_0(d,x), \nu_0(d,x))$ is the sum of products

$$\overline{\alpha}_t^o(\delta, g, |\zeta_0(d,x)|, \nu_0(d,x))\Omega^+(\delta, g, \zeta_0^o(d,x), \nu_0^o(d,x)) + \underline{\alpha}^o(\delta, g, |\zeta_0(d,x)|, \nu_0(d,x))\Omega^-(\delta, g, \zeta_0^o(d,x), \nu_0^o(d,x)).$$

Observe that, by Lemma C.2(3), $\overline{\alpha}_t^o(\delta, g, |\zeta_0(d,x)|, \nu_0(d,x))$ is increasing as a function of δ and $\Omega^+(\delta, g, \zeta_0^o(d,x), \nu_0^o(d,x))$ is non-decreasing as a function of δ by Lemma C.3(1). Since both quantities are positive, it follows that $\overline{\alpha}_t^o(\delta, g, |\zeta_0(d,x)|, \nu_0(d,x))\Omega^+(\delta, g, \zeta_0^o(d,x), \nu_0^o(d,x))$ is non-decreasing (increasing if $\Omega^+(\delta, g, \zeta_0^o(d,x), \nu_0^o(d,x)) > 0$). Similarly, by Lemmas C.2(3) and C.3(1), $\underline{\alpha}_t^o$ is decreasing and Ω^- is non-positive and non-decreasing as a function of δ, so the product is is non-decreasing (increasing if $\Omega^- < 0$). Thus, $\Gamma(\delta, g, \zeta_0(d,x), \nu_0(d,x))$ is increasing as a function of δ.

(2) (we only establish the results for Γ as for Γ_0 is analogous) By a similar argument, Lemma C.3(2) and Lemma C.2(4) it follows that $\Gamma(\delta, g, \zeta_0(d,x), \nu_0(d,x))$ is decreasing as a function of g.

(3) Clearly, $\max_{o\in\{0,\ldots,L\}} \Omega(\delta_t, g_t, \zeta_0(d,x), \nu_0(d,x)) = O\left(\frac{\delta_t + t^{-1}}{g_t + t^{-1}}\right)$. Thus $\Gamma(\delta_t, g_t, \zeta_0(d,x), \nu_0(d,x))$ inherits the same rate. Similarly, $\max_{o\in\{0,\ldots,L\}} \Omega_0(g_t, \zeta_0(d,x), \nu_0(d,x)) = O\left(\frac{t^{-1}}{g_t + t^{-1}}\right)$ and $\Gamma_0(\delta_t, g_t, \zeta_0(d,x), \nu_0(d,x))$ inherits the same rate. $\square$

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

1. *$|\zeta_t^\alpha(d,x) - \theta(d,x)| \leq \Gamma(\gamma, h_t(d,x) - \eta, \zeta_0(d,x) - \theta(d,x), \nu_0(d,x))$.*

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

C.3.1 Relationship between Γ and Ω

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

Lemma C.7. *For each $o \in \{0,...,L\}$, $(d,x) \in \mathbb{D}\times\mathbb{X}$, $\gamma, g \geq 0$ and $\nu_0(d,x)$,*

$$\lim_{\bar{\zeta}_0^{-o}(d,x)\to\infty} \left|\Gamma(\gamma,g,|\zeta_0(d,x)|,\nu_0(d,x)) - \Omega(\gamma,g,|\zeta_0^o(d,x)|,\nu_0^o(d,x))\right| = 0$$

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

$$\lim_{\bar{\zeta}_0^{o'}(d,x)\to\infty} \underline{\ell}_t(\gamma,g,|\zeta_0^{o'}(d,x)|,\nu_0^{o'}(d,x)) = \lim_{\bar{\zeta}_0^{o'}(d,x)\to\infty} \overline{\ell}_t(\gamma,g,|\zeta_0^{o'}(d,x)|,\nu_0^{o'}(d,x)) = -\infty$$

Moreover, in both cases the rate is $O(-|\zeta_0^{o'}(d,x)|^2)$ (observe that the Oh depends on $(\gamma,g,\nu_0^{o'}(d,x))$). Hence, for any $o' \neq o$,

$$\underline{\alpha}_t^{o'}(\gamma,g,|\zeta_0(d,x)|,\nu_0(d,x)) = O(e^{-|\zeta_0^{o'}(d,x)|^2}) \text{ and } \overline{\alpha}_t^{o'}(\gamma,g,|\zeta_0(d,x)|,\nu_0(d,x)) = O(e^{-|\zeta_0^{o'}(d,x)|^2}).$$

That is, they converge to 0 at exponential rate.

On the other hand, $\Omega(\gamma,g,|\zeta_0^{o'}(d,x)|,\nu_0^{o'}(d,x)) = O(|\zeta_0^{o'}(d,x)|)$, hence

$$\begin{aligned}&\overline{\alpha}_t^{o'}(\gamma,g,|\zeta_0(d,x)|,\nu_0(d,x))\Omega^+(\gamma,g,|\zeta_0^{o'}(d,x)|,\nu_0^{o'}(d,x))\\&+\underline{\alpha}_t^{o'}(\gamma,g,|\zeta_0(d,x)|,\nu_0(d,x))\Omega^-(\gamma,g,|\zeta_0^{o'}(d,x)|,\nu_0^{o'}(d,x)) = O(e^{-|\zeta_0^{o'}(d,x)|^2}|\zeta_0^{o'}(d,x)|)\end{aligned}$$

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

On the other hand, these results imply that

$$\lim_{\bar{\zeta}_0^{-o}(d,x)\to\infty} \overline{\alpha}_t^o(d,x) = \min\left\{1, \frac{e^{\overline{\ell}_t \gamma,g,|\zeta_0^o(d,x)|,\nu_0^o(d,x))}}{e^{\underline{\ell}_t \gamma,g,|\zeta_0^o(d,x)|,\nu_0^o(d,x))}}\right\} = 1$$

where the last equality follows because $\overline{\ell}_t(\gamma,g,|\zeta_0^o(d,x)|,\nu_0^o(d,x)) \geq \underline{\ell}_t(\gamma,g,|\zeta_0^o(d,x)|,\nu_0^o(d,x))$.

Therefore $\lim_{\bar{\zeta}_0^{-o}(d,x)\to\infty} \Gamma(\gamma,g,|\zeta_0(d,x)|,\nu_0(d,x)) = \Omega(\gamma,g,|\zeta_0^o(d,x)|,\nu_0^o(d,x))$, as desired. □

D Concentration Inequalities

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

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

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

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

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

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

and

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

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

Remark D.1 (Remarks on Lemma D.1)**.** *We use Assumption 2(i) in the first part of the lemma. in particular, it is used in order to get an upper bound with exponential decay. The assumption, however, could be replaced by sub-exponential or any other type of control on the MGF of $Y(d)$, e.g., $E[e^{\lambda(Y(d)-\theta(d))}] \leq e^{\kappa(\lambda)}$ for some decreasing function $\lambda \mapsto \kappa(\lambda)$. This change, however, will affect the upper bound obtained in the lemma; it will decay slower than the current one. In fact, up to constant, the result in the lemma will change to*

$$\mathbf{P}\left(\left|t^{-1}\sum_{s=1}^{t}(Y_s(d) - \theta(d))1\{D_s = d\}\right| \geq a\right) \leq 2e^{-t\max_{\lambda\geq 0}\{a\lambda - \kappa(\lambda)\}}.$$

△

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

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

Observe that

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

where $E_t[.]$ denotes the conditional expectation under $\mathbf{P}$ given $(Y_s)_{s=1}^{t-1}$ and $(D_s)_{s=1}^{t}$ (but not $Y_t(d)$). Observe that $Y_t(d)$ is independent of past Y's, given D_t. This observation and the fact that $Y_t(d)$ is sub-gaussian (Assumption 2) imply

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

Iterating in this fashion,

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

Therefore, for any $\lambda > 0$

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

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

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

By analogous calculations, it is easy to show that

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

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

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

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

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

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

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

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

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

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

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

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

E Appendix for Section 3.1

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

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

We now prove Proposition 3.1.

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

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

Conditional on these sets, by Lemma C.6,

$$|\zeta_t^{\alpha}(d) - \theta(d)| \leq \Gamma(\gamma, h_t(d) - \eta, \bar{\zeta}_0(d), \nu_0(d))$$

Therefore, for any $a > 0$,

$$\mathbf{P}\left(|\zeta_t^\alpha(d) - \theta(d)| > a\right) \le 1\{\Gamma(\gamma, h_t(d) - \eta, \bar{\zeta}_0(d), \nu_0(d)) > a\} + \mathbf{P}\left(S(t,\gamma)^C\right) + \mathbf{P}\left(R(t,\eta)^C\right) + \mathbf{P}\left(U(t)^C\right).$$

By Lemma D.1 with $T = t$ and the definition of exploration structure, it follows that

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

We now choose γ, η and a for any $\varepsilon > 0$. Let $\eta = \eta_t = h_t(d)\sqrt{0.25t^{-1}\varepsilon}$, $\gamma = \gamma_t = \sqrt{2t^{-1}\upsilon\varepsilon}\sigma(d)$ (which satisfies $\eta \le h_t(d)$). By Lemma C.5(1), $g \mapsto \Gamma(\gamma_t, g, \bar{\zeta}_0(d), \nu_0(d))$ is decreasing and since $h_t(d)(1 - \sqrt{0.25t^{-1}\varepsilon}) \ge 0.5h_t(d) \iff t0.5^2 \ge 0.25\varepsilon \iff t \ge \varepsilon$, then $\Gamma(\gamma_t, h_t(d)(1 - \sqrt{0.25t^{-1}\varepsilon}), \bar{\zeta}_0(d), \nu_0(d)) \le \Gamma(\gamma_t, 0.5h_t(d), \bar{\zeta}_0(d), \nu_0(d)) =: a_t = a$. With these choices,

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

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

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

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

We now prove Corollary 3.1.

Proof of Corollary 3.1. We can prove the result using limits. For any given $o \neq 0$, let $|\bar{s}_0^o(d)| := \sqrt{\nu_0^o(d)}|\bar{\zeta}_0^o(d)|$ and let $|\bar{s}_0^{-0}(d)|$ be the $L \times 1$ vector, excluding $|\bar{s}_0^0(d)|$. We consider the limit of this quantity going to ∞.

By Lemma C.7 applied to $o = 0$, for each $\gamma \ge 0$ and $\eta \le h_t(d)$,

$$\lim_{|\bar{s}_0^{-0}(d)| \to \infty} |\Gamma\left(\gamma, h_t(d) - \eta, |\bar{\zeta}_0(d)|, \nu_0(d)\right) - \Omega(\gamma, h_t(d) - \eta, |\bar{\zeta}_0^o(d)|, \nu_0^o(d))| = 0.$$

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

$$\Omega(\gamma, h_t(d) - \eta, |\bar{\zeta}_0^o(d)|, \nu_0^o(d)) \ge \Gamma\left(\gamma, h_t(d) - \eta, |\bar{\zeta}_0(d)|, \nu_0(d)\right) - \delta$$

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

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

F Appendix for Section 3.2

Proposition 3.2 follows from this more general lemma that allows for biased sources. To state this lemma we define, for each $d \in \mathbb{D}$, $\eta^*_d : \mathbb{N} \times [0,1] \times \mathbb{R}_+ \to \mathbb{R}_+ \cup \{+\infty\}$ as follows: For any $(t, h_t(d), \Delta) \in \mathbb{N} \times [0,1] \times \mathbb{R}_+$, if $\Gamma_0(\gamma_t, h_t(d) - \eta, (-1)^{1\{d=M\}} \zeta_0(d), \nu_0(d)) < 0.5\Delta$ for all η, then we choose $\eta^*_d(t, h_t(d), \Delta) = +\infty$; otherwise,

$$\eta^*_d(t, h_t(d), \Delta) := \max\left\{\eta \colon \Gamma_0(\gamma_t, h_t(d) - \eta, (-1)^{1\{d=M\}} \zeta_0(d), \nu_0(d)) \leq 0.5\Delta \textit{ and } \eta \leq h_t(d)\right\}$$

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

The quantity $\eta^*_d(t, h_t(d), \Delta)$ defines the concentration rate of $f_t(d)$. Intuitively, this quantity is the highest level of feasible *constant* experimentation, i.e., a constant that is less or equal than $h_t(d)$, such that the incidence of the priors on the aggregated posterior mean — this incidence is given by the function Γ_0 — is small relative to the true discrepancy of the ATEs, given by Δ.

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

$$\mathbf{P}\left(\max_{d \neq M}\{\zeta^\alpha_\tau(d) - \zeta^\alpha_\tau(M)\} > 0 \cap \{\tau = t\}\right) \leq 2\sum_{d=0}^{M}\left(e^{-0.5\frac{\gamma_t(d)^2}{\upsilon\sigma(d)^2}} + e^{-4t\eta^*_d(t,h_t(d),\Delta)^2}\right) \tag{F.1}$$

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

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

This lemma shows that the quantity $\eta^*_d(t, h_t(d), \Delta)$ is key for understanding how the primitives of our setup — i.e. the exploration structure and Δ — affect the upper bound for the probability of a mistake. The upper bound for the probability of a mistake decays exponentially with t and is non-increasing in $h_t(d)$ and Δ. Intuitively, as the degree of exploration increases, the data become less dependent on the past and thus more informative, resulting in a tighter bound. Also, as Δ becomes more positive, so does the difference between the PM's posteriors, which also decreases the probability of making a mistake.

Proof of Lemma F.1 . We divide the proof into several steps. Throughout the proof, we use the following definitions. For any $t \in \mathbb{N}$,

$$\mathcal{J}_t(\gamma, d) := \{|\sum_{s=1}^{t} \bar{Y}_s(d)|/\sqrt{t} \leq \gamma\}, \ \forall \gamma > 0,$$

$$\mathcal{V}(t, d) := \{\iota_t(d) \geq h_t(d)\},$$

$$\mathcal{E}_t(\eta, d) := \{|f_t(d) - \iota_t(d)| \leq \eta\}, \ \forall \eta > 0.$$

STEP 1 In this step we show that

$$\{\max_{d\neq M}\{\zeta_t^\alpha(d)-\zeta_t^\alpha(M)\}>0\}\cap\{\tau=t\}\subseteq\{\max_{d\neq M}\left\{\bar{\zeta}_t^\alpha(d)-\bar{\zeta}_t^\alpha(M)-c_t(\gamma_t,d,M)\right\}>\Delta\}\cap\{\tau=t\}$$

and that

$$\begin{aligned}\{\max_{d\neq M}\left\{\bar{\zeta}_t^\alpha(d)-\bar{\zeta}_t^\alpha(M)-c_t(\gamma_t,d,M)\right\}>\Delta\}\subseteq\cup_d\left\{\sum_{o=0}^{L}\alpha_t^o(d)\frac{(-1)^{1\{d=M\}}\bar{\zeta}_0^o(d)\nu_0^o(d)/t}{f_t(d)+\nu_0^o(d)/t}>0.5\Delta\right\}\cap\mathcal{J}_t(\gamma_t,d)\\ \cup\mathcal{J}_t(\gamma_t,d)^C.\end{aligned}$$

Since

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

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

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

$$\{\max_{d\neq M}\{\zeta_t^\alpha(d)-\zeta_t^\alpha(M)\}>0\cap\tau=t\}\subseteq\{\max_{d\neq M}\left\{\bar{\zeta}_t^\alpha(d)-\bar{\zeta}_t^\alpha(M)-c_t(\gamma_t,d,M)\right\}>\Delta\}.$$

Observe that

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

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

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

$$\begin{aligned}&\cup_{d\neq M}\{\left\{\bar{\zeta}_t^\alpha(d)-\bar{\zeta}_t^\alpha(M)-c_t(\gamma_t,d,M)\right\}>\Delta\}\cap\{\bar{\zeta}_t^\alpha(M)+c_t(\gamma_t,M)\geq-0.5\Delta\}\\ &\cup\{\bar{\zeta}_t^\alpha(M)+c_t(\gamma_t,M)<-0.5\Delta\}\\ =&\cup_{d\neq M}\{\bar{\zeta}_t^\alpha(d)>c_t(\gamma_t,d)+0.5\Delta\}\cup\{\bar{\zeta}_t^\alpha(M)<-(c_t(\gamma_t,M)+0.5\Delta)\}.\end{aligned}$$

We now bound each of the sets in the RHS.

Observe that $\bar{\zeta}_t^\alpha(d)=\left(\sum_{o=1}^{L}\frac{\alpha_t^o(d)}{N_t(d)+\nu_0^o(d)}\right)\left(\sum_{s=1}^{t}1\{D_s=d\}\bar{Y}_s(d)\right)+\sum_{o=1}^{L}\alpha_t^o(d)\frac{\nu_0^o(d)\bar{\zeta}_0^o(d)}{N_t(d)+\nu_0^o(d)}$. Conditional on the set $\mathcal{J}_t(\gamma,d)$, it follows that

$$\bar{\zeta}_t^{\alpha}(d) \le \gamma\sqrt{t}\left(\sum_{o=1}^{L}\frac{\alpha_t^o(d)}{N_t(d)+\nu_0^o(d)}\right)+\sum_{o=1}^{L}\alpha_t^o(d)\frac{\bar{\zeta}_0^o(d)\nu_0^o(d)/t}{f_t(d)+\nu_0^o(d)/t}$$
$$\text{and } \bar{\zeta}_t^{\alpha}(d) \ge -\gamma\sqrt{t}\left(\sum_{o=1}^{L}\frac{\alpha_t^o(d)}{N_t(d)+\nu_0^o(d)}\right)+\sum_{o=1}^{L}\alpha_t^o(d)\frac{\bar{\zeta}_0^o(d)\nu_0^o(d)/t}{f_t(d)+\nu_0^o(d)/t}$$

By the definition of c_t and choosing $\gamma=\gamma_t(d)$, it follows that for any $d\in\{0,\dots,M-1\}$,

$$\begin{aligned}\{\bar{\zeta}_t^{\alpha}(d) > c_t(\gamma_t,d)+0.5\Delta\} \subseteq &\{\bar{\zeta}_t^{\alpha}(d) > c_t(\gamma_t,d)+0.5\Delta\}\cap\mathcal{J}_t(\gamma_t,d)\cup\mathcal{J}_t(\gamma_t,d)^C\\ \subseteq &\left\{\sum_{o=0}^{L}\alpha_t^o(d)\frac{\bar{\zeta}_0^o(d)\nu_0^o(d)/t}{f_t(d)+\nu_0^o(d)/t} > 0.5\Delta\right\}\cap\mathcal{J}_t(\gamma_t,d)\cup\mathcal{J}_t(\gamma_t,d)^C.\end{aligned}$$

Similarly,

$$\{\bar{\zeta}_t^{\alpha}(M) > -(c_t(\gamma_t,M)+0.5\Delta)\} \subseteq \left\{\sum_{o=0}^{L}\alpha_t^o(M)\frac{(-\bar{\zeta}_0^o(M))\nu_0^o(M)/t}{f_t(M)+\nu_0^o(M)/t} > 0.5\Delta\right\}\cap\mathcal{J}_t(\gamma_t,M)\cup\mathcal{J}_t(\gamma_t,M)^C.$$

Hence,

$$\begin{aligned}\{\max_{d\neq M}\left\{\bar{\zeta}_t^{\alpha}(d)-\bar{\zeta}_t^{\alpha}(M)-c_t(\gamma_t,d,M)\right\} > \Delta\} \subseteq &\cup_d\left\{\sum_{o=0}^{L}\alpha_t^o(d)\frac{(-1)^{1\{d=M\}}\bar{\zeta}_0^o(d)\nu_0^o(d)/t}{f_t(d)+\nu_0^o(d)/t} > 0.5\Delta\right\}\cap\mathcal{J}_t(\gamma_t,d)\\ &\cup\mathcal{J}_t(\gamma_t,d)^C.\end{aligned}$$

STEP 2. We now bound $\mathbf{P}\left(\max_{d\neq M}\{\zeta_\tau^{\alpha}(d)-\zeta_\tau^{\alpha}(M)\} > 0\cap\{\tau=t\}\right)$ when $\forall d\colon (-1)^{1\{d=M\}}\bar{\zeta}_0(d)\le 0$. By Step 1,

$$\begin{aligned}\mathbf{P}\left(\max_{d\neq M}\{\zeta_\tau^{\alpha}(d)-\zeta_\tau^{\alpha}(M)\} > 0\cap\{\tau=t\}\right) \le &\mathbf{P}\left(\max_{d\neq M}\{\bar{\zeta}_t^{\alpha}(d)-\bar{\zeta}_t^{\alpha}(M)\} > \Delta\right)\\ \le &\mathbf{P}\left(\cup_d\left\{\sum_{o=0}^{L}\alpha_t^o(d)\frac{(-1)^{1\{d=M\}}\bar{\zeta}_0^o(d)\nu_0^o(d)/t}{f_t(d)+\nu_0^o(d)/t} > 0.5\Delta\right\}\cap\mathcal{J}_t(\gamma_t,d)\right)\\ &+\mathbf{P}(\mathcal{J}_t(\gamma_t,d)^C).\end{aligned}$$

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

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

Observe that

$$\mathbf{P}\left(\max_{d\neq M}\{\zeta^{\alpha}_{\tau}(d)-\zeta^{\alpha}_{\tau}(M)\}>0\cap\{\tau=t\}\right)\leq\mathbf{P}\left(\max_{d\neq M}\{\zeta^{\alpha}_{\tau}(d)-\zeta^{\alpha}_{\tau}(M)\}>0\cap\{\tau=t\}\cap\mathcal{V}(t,d)\right)+\mathbf{P}\left(\mathcal{V}(t,d)^{C}\right)$$
$$\leq\mathbf{P}\left(\max_{d\neq M}\{\zeta^{\alpha}_{\tau}(d)-\zeta^{\alpha}_{\tau}(M)\}>0\cap\{\tau=t\}\cap\mathcal{V}(t,d)\right)+\omega_t(d).$$

We now bound $\mathbf{P}\left(\max_{d\neq M}\{\zeta^{\alpha}_{\tau}(d)-\zeta^{\alpha}_{\tau}(M)\}>0\cap\{\tau=t\}\cap\mathcal{V}(t,d)\right)$. By Step 1 and the union bound, this probability can be bounded by

$$\sum_d\mathbf{P}\left(\left\{\sum_{o=0}^{L}\alpha^o_t(d)\frac{(-1)^{1\{d=M\}}\bar{\zeta}^o_0(d)\nu^o_0(d)/t}{f_t(d)+\nu^o_0(d)/t}>0.5\Delta\right\}\cap\mathcal{J}_t(\gamma_t,d)\cap\mathcal{E}_t(\eta,d)\right)$$
$$+\sum_d\left(\mathbf{P}\left(\mathcal{J}_t(\gamma_t,d)^C\right)+\mathbf{P}\left(\mathcal{E}_t(\eta,d)^C\right)\right).$$

for some $\eta>0$ to be chosen below.

We now bound the first term in the display above. To do this, observe that $\mathcal{E}_t(\eta,d)$ and $\mathcal{V}(t,d)$, imply $\{f_t(d)\geq h_t(d)-\eta\}$. This fact and Lemmas C.4(2) and C.1, imply that under $\mathcal{J}_t(\gamma_t,d)\cap\{f_t(d)\geq h_t(d)-\eta\}$, it follows that for any $d\in\mathbb{D}$,

$$\alpha^o_t(d)\frac{\bar{\zeta}^o_0(d)\nu^o_0(d)/t}{f_t(d)+\nu^o_0(d)/t}\leq\alpha^o_t(d)\Omega_0(h_t(d)-\eta,(-1)^{d=M}\bar{\zeta}^o_0(d),\nu^o_0(d))$$
$$\leq\overline{\alpha}^o_t(\gamma_t,h_t(d)-\eta,|\zeta_0(d)|,\nu_0(d))\Omega^+_0(h_t(d)-\eta,(-1)^{d=M}\bar{\zeta}^o_0(d),\nu^o_0(d))$$
$$+\underline{\alpha}^o_t(\gamma_t,h_t(d)-\eta,|\zeta_0(d)|,\nu_0(d))\Omega^-_0(h_t(d)-\eta,(-1)^{d=M}\bar{\zeta}^o_0(d),\nu^o_0(d)).$$

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

$$\left\{\sum_{o=0}^{L}\alpha^o_t(d)\frac{(-1)^{d=M}\bar{\zeta}^o_0(d)\nu^o_0(d)/t}{f_t(d)+\nu^o_0(d)/t}>0.5\Delta\right\}\cap\mathcal{J}_t(\gamma_t,d)\cap\mathcal{E}_t(\eta,d)$$
$$\subseteq\left\{\Gamma_0(\gamma_t,h_t(d)-\eta,(-1)^{1\{d=M\}}\bar{\zeta}_0(d),\nu_0(d))>0.5\Delta\right\}\cap\mathcal{J}_t(\gamma_t,d)\cap\mathcal{E}_t(\eta,d),$$

which in turn implies

$$\mathbf{P}\left(\max_{d\neq M}\{\zeta^{\alpha}_{\tau}(d)-\zeta^{\alpha}_{\tau}(M)\}>0\cap\{\tau=t\}\cap\mathcal{V}(t,d)\right)\leq\sum_d 1\{\mathcal{U}_d(\gamma_t,h_t(d)-\eta,\Delta)\}$$
$$+\sum_d\left(\mathbf{P}\left(\mathcal{J}_t(\gamma_t,d)^C\right)+\mathbf{P}\left(\mathcal{E}_t(\eta,d)^C\right)\right)\tag{F.3}$$

where $\mathcal{U}_d(\gamma_t,h_t(d)-\eta,\Delta):=\left\{\Gamma_0(\gamma_t,h_t(d)-\eta,(-1)^{1\{d=M\}}\bar{\zeta}_0(d),\nu_0(d))>0.5\Delta\right\}$.

We now choose η so that the first term in the RHS is naught. If $\Gamma_0(\gamma_t, h_t(d)-\eta, (-1)^{1\{d=M\}}\bar{\zeta}_0(d), \nu_0(d)) < 0.5\Delta$ for all η, then we choose $\eta^*_d(t, h_t(d), \Delta) = +\infty$; otherwise,

$$\eta^*_d(t, h_t(d), \Delta) := \max\left\{\eta\colon \Gamma_0(\gamma_t, h_t(d)-\eta, (-1)^{1\{d=M\}}\bar{\zeta}_0(d), \nu_0(d)) \leq 0.5\Delta \textit{ and } \eta \leq h_t(d)\right\}$$

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

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

$$\mathbf{P}\left(\max_{d\neq M}\{\zeta^{\alpha}_{\tau}(d) - \zeta^{\alpha}_{\tau}(M)\} > 0 \cap \{\tau = t\} \cap \mathcal{V}(t,d)\right) \leq \sum_d \left(\mathbf{P}\left(\mathcal{J}_t(\gamma_t, d)^C\right) + \mathbf{P}\left(\mathcal{E}_t(\eta^*_d(t, h_t(d), \Delta), d)^C\right)\right).$$

Thus, by Lemma D.1,

$$\mathbf{P}\left(\max_{d\neq M}\{\zeta^{\alpha}_{\tau}(d) - \zeta^{\alpha}_{\tau}(M)\} > 0 \cap \{\tau = t\}\right) \leq 2\sum_{d=0}^{M}\left(e^{-0.5\frac{\gamma_t^2}{\upsilon\sigma(d)^2}} + e^{-4t\eta^*_d(t,h_t(d),\Delta)^2} + \omega_t(d)\right)$$

STEP 4. We conclude the proof by showing some properties of η^*_d. First, $t \mapsto \eta^*_d(t, h_t(d), \Delta)$ is non-decreasing. To show this, first note that $\Gamma_0(\gamma_t, h_t(d)-\eta, (-1)^{1\{d=M\}}\bar{\zeta}_0(d), \nu_0(d))$ is (implicitly) a function of t and thus it suffices to show it is non-increasing (for a fixed $h_t(d)$) and $\eta \mapsto \Gamma_0(\gamma_t, h_t(d)-\eta, (-1)^{1\{d=M\}}\bar{\zeta}_0(d), \nu_0(d))$ is non-decreasing. This follows from Lemma C.5(2) and the fact that g (in that lemma) equals $h_t(d)-\eta$. ;
we now show the former.

By construction of Γ_0 it suffices to show that $t \mapsto K_t(\gamma_t, h_t(d)-\eta) := \overline{\alpha}^o_t(\gamma_t, h_t(d)-\eta, |\bar{\zeta}_0(d)|, \nu_0(d))\Omega^+_0(h_t(d)-\eta, (-1)^{d=M}\bar{\zeta}^o_0(d), \nu^o_0(d)) + \underline{\alpha}^o_t(\gamma_t, h_t(d)-\eta, |\bar{\zeta}_0(d)|, \nu_0(d))\Omega^-_0(h_t(d)-\eta, (-1)^{d=M}\bar{\zeta}^o_0(d), \nu^o_0(d))$ is non-increasing for each o. If $(-1)^{d=M}\bar{\zeta}^o_0(d) \geq 0$ then Ω_0 is positive and decreasing as a function of t (see its definition). In addition, $\underline{\sigma}^o_t$ and $\overline{\sigma}^o_t$ are non-increasing and decreasing in t resp. Hence $\overline{\alpha}^o_t$ is decreasing in t and positive. Thus, $K_t(\gamma, h_t(d)-\eta) = \overline{\alpha}^o_t(\gamma, h_t(d)-\eta, |\bar{\zeta}_0(d)|, \nu_0(d))\Omega_0(h_t(d)-\eta, (-1)^{d=M}\bar{\zeta}^o_0(d), \nu^o_0(d))$ is decreasing. In addition, by the proof of Lemma C.5(1) it follows that $\gamma \mapsto K_t(\gamma, h_t(d)-\eta)$ is increasing and since $t \mapsto \gamma_t$ is non-increasing it follows that $t \mapsto K_t(\gamma_t, h_t(d)-\eta)$ is non-increasing for this case. If $(-1)^{d=M}\bar{\zeta}^o_0(d) \leq 0$ then Ω_0 is negative and decreasing as a function of t (see its definition). Also, $\underline{\alpha}^o_t$ is increasing in t (see Lemma C.2(5)) and positive. Thus, $t \mapsto K_t(\gamma, h_t(d)-\eta) = \underline{\alpha}^o_t(\gamma_t, h_t(d)-\eta, |\bar{\zeta}_0(d)|, \nu_0(d))\Omega^-_0(h_t(d)-\eta, (-1)^{d=M}\bar{\zeta}^o_0(d), \nu^o_0(d))$ is decreasing in this case. We thus showed that $t \mapsto \Gamma_0(\gamma_t, h_t(d)-\eta, (-1)^{1\{d=M\}}\bar{\zeta}_0(d), \nu_0(d))$ is non-increasing.

Second, $\Delta \mapsto \eta^*_d(t, h_t(d), \Delta)$ is non-decreasing. To show this is sufficient to show that $\eta \mapsto \Gamma_0(\gamma_t, h_t(d)-\eta, (-1)^{1\{d=M\}}\bar{\zeta}_0(d), \nu_0(d))$ is non-decreasing. This follows from Lemma C.5(2) and the fact that g (in that lemma) equals $h_t(d)-\eta$.

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

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

Lemma F.2. *Suppose all the conditions of Lemma F.1 hold and* $\frac{|\zeta_0^0(d)-\theta(d)|\nu_0^0(d)/t}{h_t(d)+\nu_0^0(d)/t} \leq 0.5\Delta$.[19] *Then, for any* $\varepsilon > 0$*, there exists a C such that for all* $\min_{o\neq 0}|\zeta_0^o(d)-\theta(d)| \geq C$*, it follows that*

$$\mathbf{P}\left(\max_{d\neq M}\{\zeta_\tau^\alpha(d)-\zeta_\tau^\alpha(M)\} > 0 \cap \{\tau = t\}\right) \leq \sum_{d=0}^{M}(2e^{-\frac{0.5(\gamma_t(d))^2}{\upsilon\sigma(d)^2}} + e^{-4t(\eta_d^{oracle}(t,h_t(d),(\Delta-\varepsilon)/(1+\varepsilon)))^2})$$
$$+1\{\forall d\colon (-1)^{1\{d=M\}}\bar{\zeta}_0(d) \leq 0\}\omega_t(d).$$

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

$$\max\left\{\eta\colon \frac{|\zeta_0^0(d)-\theta(d)|\nu_0^0(d)/t}{h_t(d)-\eta+\nu_0^0(d)/t} \leq 0.50\frac{\Delta-\varepsilon}{1+\varepsilon} \text{ and } \eta \leq \epsilon\right\}.$$

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

The behavior of η_d^* determines whether the upper bound embodies an oracle property similar to the one we demonstrated for the concentration rates. Given the properties of the weights illustrated in Proposition 2.1 and Lemma B.1, it is easy to show that if sources other than $o = 0$ are sufficiently stubborn, then η_d^* becomes arbitrary close to η_d^{oracle}, where η_d^{oracle} is defined as the largest η such that $\frac{|\zeta_0^0(d)-\theta(d)|\nu_0^0(d)/t}{h_t(d)-\eta+\nu_0^0(d)/t} \leq 0.5\Delta$, which is the relevant quantity determining the probability of mistake for the *least stubborn* source. It then follows that the bound obtained in Proposition 3.2 would be arbitrary close to the oracle one; the corollary below formalizes this discussion.

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

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

$$\begin{aligned}
&\{\bar{\zeta}_t^\alpha(d) > (-1)^{1\{d=M\}}(c_t(\gamma_t, d)+0.5\Delta)\} \cap \mathcal{V}(t,d) \\
&\subseteq \left\{\sum_{o=0}^{L}\alpha_t^o(d)\frac{(-1)^{1\{d=M\}}\bar{\zeta}_0^o(d)\nu_0^o(d)/t}{f_t(d)+\nu_0^o(d)/t} > 0.5\Delta\right\} \cap \mathcal{J}_t(\gamma_t, d) \cap \mathcal{E}_t(\eta, d) \cap \mathcal{V}(t,d) \\
&\cup \mathcal{J}_t(\gamma_t, d)^C \cup \mathcal{E}_t(\eta, d)^C.
\end{aligned}$$

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

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

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

$$\begin{aligned}&\{\bar{\zeta}_t^\alpha(d) > (-1)^{1\{d=M\}}(c_t(\gamma_t,d)+0.5\Delta)\}\\ \subseteq&\left\{(1+\varepsilon)\frac{|\bar{\zeta}_0^0(d)|\nu_0^o(d)/t}{f_t(d)+\nu_0^o(d)/t}+\varepsilon>0.5\Delta\right\}\cap\mathcal{J}_t(\gamma_t,d)\cap\mathcal{E}_t(\eta,d)\cup\mathcal{J}_t(\gamma_t,d)^C\cup\mathcal{E}_t(\eta,d)^C.\end{aligned}$$

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

$$\eta_d^{oracle}(t,h_t(d),(\Delta-\varepsilon)/(1+\varepsilon)):=\max\left\{\eta\colon\ \frac{|\zeta_0^0(d)-\theta(d)|\nu_0^0(d)/t}{h_t(d)-\eta+\nu_0^0(d)/t}\leq 0.50\frac{\Delta-\varepsilon}{1+\varepsilon}\ \textit{and}\ \eta\leq\epsilon\right\}.$$

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

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

□

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

Lemma F.3. *Suppose all the conditions of Proposition 3.2 hold, and, for any t, $\gamma_t(d)\geq 2\sqrt{\upsilon}\sigma(d)A$ for all $d\in\mathbb{D}$ and $\frac{4\min_{d\in\mathbb{D}}(\eta_d^*(t,h_t(d),\Delta))^2}{A'}\geq\frac{1}{t}$ with A and A' such that*

$$A\geq-\log\frac{\beta}{M+1}\ \textit{and}\ A'\geq-\log\frac{\beta}{M+1}\tag{F.4}$$

Then

$$\max_{t\in\{B,\dots,T\}}\mathbf{P}\left(\max_{d\neq M}\{\zeta_\tau^\alpha(d)-\zeta_\tau^\alpha(M)\}>0\cap\{\tau=t\}\right)\leq\beta.$$

Observe that the condition $\frac{4\min_{d\in\mathbb{D}}(\eta_d^*(t,h_t(d),\Delta))^2}{A'}\geq\frac{1}{t}$ is trivially satisfied in the setup of Corollary 3.3 since there $\eta_d^* = \infty$.

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

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

By our choice of γ_t and the condition on A', the first term in the RHS is less or equal than $2(M+1)e^{-A}$ whereas the second is equal to $(M+1)e^{-A'}$. Hence, the desired result immediately follows.

□

G Appendix for Section 3.3

To show Proposition 3.3 we use the following lemmas, whose proofs are relegated to the end of the section.

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

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

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

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

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

where

$$\Lambda_t(\Delta):=4\sum_{d=0}^{M}t^{-1}\sum_{s=1}^{t}\left(e^{-0.5(s-1)\frac{\gamma^2}{\upsilon\sigma(d)^2}}+e^{-4(s-1)\eta_d^*(s-1,\bar{\Xi}_t/(M+1),\Delta)^2}\right).$$

Proof of Proposition 3.3. Since sources are assumed to be unbiased, $\eta^*=\infty$. Therefore, by Lemma G.2,

$$\mathbf{P}\left(\max_d\theta(d)-t^{-1}\sum_{s=1}^{t}Y_s>\sqrt{\frac{\gamma}{t}}\left(\sqrt{2\upsilon}\sigma(d)+\frac{||\theta||_1}{2}\right)-||\theta||_1\left(\sqrt{1-\bar{\Xi}_t}\sqrt{e^\gamma 4\sum_{d=0}^{M}t^{-1}\sum_{s=1}^{t}e^{-0.5(s-1)\frac{\gamma^2}{\upsilon\sigma(d)^2}}}+\frac{\bar{\Xi}_t}{M+1}\right)\right)\leq 5e^{-\gamma}.$$

Observe that $t^{-1}\sum_{s=1}^{t} e^{-0.5(s-1)\frac{\gamma^2}{\upsilon\sigma(d)^2}} = t^{-1}\int_0^{t-1} e^{-0.5x\frac{\gamma^2}{\upsilon\sigma(d)^2}}dx = \frac{\upsilon\sigma(d)^2}{2t\gamma^2}(1-e^{-0.5(t-1)\frac{\gamma^2}{\upsilon\sigma(d)^2}})$ which can be lower bounded by $\frac{\upsilon\sigma(d)^2}{2t\gamma^2}$. Thus, the desired result follows.

□

G.1 Proofs of Lemmas

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

$$t^{-1}\sum_{s=1}^{t} Y_s - \max_d \theta(d) = \left(\sum_{d=0}^{M} t^{-1}\sum_{s=1}^{t} 1\{D_s = d\}(Y_s(d)-\theta(d))\right) + \left(\sum_{d=0}^{M}\theta(d)(f_t(d)-\iota_t(d))\right) + \sum_{d=0}^{M}\theta(d)\iota_t(d) - \max_d \theta(d)$$
$$=: Term1 + Term2 + Term3.$$

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

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

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

□

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

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

Thus,

$$\begin{aligned}
&\mathbf{P}\left(\max_d \theta(d) - t^{-1}\sum_{s=1}^t Y_s > -\sqrt{\frac{\gamma}{t}}\left(\sqrt{2\upsilon}\sigma(d) + \frac{||\theta||_1}{2}\right) - ||\theta||_1\left(\sqrt{t^{-1}\sum_{s=1}^t (1-\Xi_s)^2}\sqrt{e^{\gamma}\Lambda_t(\Delta)} + \frac{\bar{\Xi}_t}{M+1}\right)\right)\\
\leq &\mathbf{P}\left(\max_d \theta(d) - t^{-1}\sum_{s=1}^t Y_s > -\sqrt{\frac{\gamma}{t}}\left(\sqrt{2\upsilon}\sigma(d) + \frac{||\theta||_1}{2}\right) + \max_d \theta(d) - \sum_{d=0}^M \theta(d)\iota_t(d)\right)\\
&+\mathbf{P}\left(\max_d \theta(d) - \sum_{d=0}^M \theta(d)\iota_t(d) \leq -||\theta||_1\left(\sqrt{t^{-1}\sum_{s=1}^t (1-\Xi_s)^2}\sqrt{e^{\gamma}\Lambda_t(\Delta)} + \frac{\bar{\Xi}_t}{M+1}\right)\right)\\
\leq &4e^{-\gamma} + \mathbf{P}\left(\max_d \theta(d) - \sum_{d=0}^M \theta(d)\iota_t(d) \leq -||\theta||_1\left(\sqrt{t^{-1}\sum_{s=1}^t (1-\Xi_s)^2}\sqrt{e^{\gamma}\Lambda_t(\Delta)} + \frac{\bar{\Xi}_t}{M+1}\right)\right)
\end{aligned}$$

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

$$\begin{aligned}
&\sum_{d=0}^M \theta(d) t^{-1}\sum_{s=1}^t 1\{d = \arg\max_a \theta(a)\} - \sum_{d=0}^M \theta(d)\iota_t(d)\\
=&t^{-1}\sum_{s=1}^t (1-\Xi_s)\sum_{d=0}^M \theta(d)(1\{d = \arg\max_a \theta(a)\} - 1\{d = \arg\max_a \zeta^{\alpha}_{s-1}(d)\}) + \bar{\Xi}_t\left(\max_d \theta(d) - \sum_{d=0}^M \frac{\theta(d)}{M+1}\right).
\end{aligned}$$

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

$$(1-\Xi_s)\sum_{d=0}^M \theta(d)(1\{d = \arg\max_a \theta(a)\} - 1\{d = \arg\max_a \zeta^{\alpha}_{s-1}(d)\}) \leq (1-\bar{\Xi}_t)1\{\max_{a\neq M}\{\zeta^{\alpha}_{s-1}(a) - \zeta^{\alpha}_{s-1}(M)\} > 0\}||\theta||_1.$$

Thus,

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

Hence, by the Cauchy-Swarchz inequality,

$$\begin{aligned}
&\mathbf{P}\left(\max_d \theta(d) - \sum_{d=0}^M \theta(d)\iota_t(d) \leq -||\theta||_1\left(\sqrt{t^{-1}\sum_{s=1}^t (1-\Xi_s)^2}\sqrt{e^{\gamma}\Lambda_t(\Delta)} + \frac{\bar{\Xi}_t}{M+1}\right)\right)\\
\leq &\mathbf{P}\left(t^{-1}\sum_{s=1}^t 1\{\max_{d\neq M}\{\zeta^{\alpha}_{s-1}(d) - \zeta^{\alpha}_{s-1}(M)\} > 0\} \geq e^{\gamma}\Lambda_t(\Delta)\right),
\end{aligned}$$

Thus, by the Markov inequality,

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

By Lemma F.1 and the fact that by our particular choice of stopping rule $h_t(d)=\bar{\Xi}_t/(M+1)$ and $\omega_t(d)=0$, it follows that

$$\begin{aligned}&\mathbf{P}\left(t^{-1}\sum_{s=1}^{t}1\{\max_{d\neq M}\{\zeta_{s-1}^{\alpha}(d)-\zeta_{s-1}^{\alpha}(M)\}>0\}\geq e^{\gamma}\Lambda_t(\Delta)\right)\\&\leq e^{-\gamma}(\Lambda_t(\Delta))^{-1}4\sum_{d=0}^{M}t^{-1}\sum_{s=1}^{t}\left(e^{-0.5(s-1)\frac{\gamma^2}{\upsilon\sigma(d)^2}}+e^{-4(s-1)\eta_d^*(s-1,\bar{\Xi}_t/(M+1),\Delta)^2}\right)\end{aligned}$$

□

H Relationship to ambiguity aversion and Empirical Hierarchical Bayes

In this section, we discuss alternative interpretations of and potential extensions to our learning model with multiple sources.

H.1 Extensions to an ambiguity aversion PM

Our interpretation of the problem is one where, at the beginning of the experiment, the PM is confronted with difference sources of information which she is either unwilling or unable to discern which one — or even what combinations of them — present the best description of nature. Using the terminology from the decision theory literature, we formalize this feature as the PM facing ambiguity, and thus we depart from the standard Bayesian updating model and use a “multi prior" Bayesian updating problem, wherein each prior represents a source (see Epstein and Schneider (2003) and references therein). It is important to note, however, that while we borrow the conceptual framework of the ambiguity aversion literature, our goal is very different. In particular, we are not concerned with dynamic optimality and thus consistency concerns do not apply.

We now present some remarks regarding the PM’s attitude towards this ambiguity and discuss some extensions. For each $(d,x)\in\mathbb{D}\times\mathbb{X}$, the object of interest is the average effect of each treatment, and, at each instance t, the PM will estimate it using

$$\zeta_t^{\alpha}(d,x) := \int y \int_{\Theta} p_{\theta}(y)\mu_t^{\alpha}(d,x)(d\theta)dy =: \sum_{o=0}^{L} \alpha_t^o(d,x)\zeta_t^o(d,x). \tag{H.1}$$

Thus, at each instance t, one can think of the PM solving this estimation problem

$$\max_{\zeta\in\mathbb{R}} \sum_{o=0}^{L} \alpha_t^o(d,x) \int (y-\zeta)^2 \int_{\Theta} p_{\theta}(y)\mu_t^o(d,x)(d\theta)dy.$$

Once we cast the problem in this form, we can see that by taken a (weighted) average over sources, we are postulating that the PM is not averse to the uncertainty over the sources; i.e., is not averse to the ambiguity generated by the sources. A possible extension would be one where the aforementioned optimization problem is replaced by

$$\max_{\zeta\in\mathbb{R}} \phi^{-1}\left(\sum_{o=0}^{L} \alpha_t^o(d,x)\phi\left(\int (y-\zeta)^2 \int_{\Theta} p_{\theta}(y)\mu_t^o(d,x)(d\theta)dy\right)\right) \tag{H.2}$$

where ϕ is a concave function. The form of ϕ will dictate how averse the PM is to having uncertainty over sources. This modeling choice is analogous to the smooth ambiguity model put forward by Klibanoff et al. (2009). For instance, by suitably chosing ϕ, the PM will use "the worst source" to construct an estimator of the average treatment effect.

Learning vs. max-min Considering extensions of the type presented in expression H.2 is outside the scope of the current paper, here we simply point out what we think should be a desirable property of this potential extension. Consider a very simple case where there are 3 sources. Initially, the PM was not able to discard any of these sources, but after enough instances, evidence suggests that one of this source is not externally valid (in the sense defined above) while the other two seem roughly equal. It seems overly pessimistic for the PM to guard herself against *all* three source — e.g. to apply a max-min criteria over all three — as the data already discarded one. That is, we believe that the ambiguity averse criteria and the aggregation method should take into account the accumulation of new evidence through learning. To our knowledge, there is no agreed upon way of doing this in the decision theoretic literature.

H.2 Alternative interpretation of our aggregation method.

We conclude this section by providing an alternative interpretation — based on an Empirical Hierarchical Bayes model — to the multi prior one, *for the "ambguity neutral" PM case* i.e., expression H.1.

Consider a Hierarchical Bayes model (HBM) where, for each $(d,x) \in \mathbb{D}\times\mathbb{X}$, $Y(d,x)$ is thought to be drawn from a Gaussian PDF with mean θ and variance 1. In turn, θ is thought to be drawn from $\phi(\cdot; a_0(d,x), b_0(d,x))$. The hierarchical aspect of this Bayes model is that parameters $(a_0(d,x), b_0(d,x))$ are thought to be random, coming from a distribution $P_0(\cdot|d,x)$. A particular case for $P_0(\cdot|d,x)$ is one where $(a_0(d,x), b_0(d,x))$ can only take finitely many values given by $(\zeta_0^o(d,x), 1/\nu_0^o(d,x))$ with $o = \{0,...,L\}$ and $P_0(\cdot|d,x)$ assigns probability $\pi_0^o(d,x)$ to each. At each instance t, it can be shown that the posterior of θ, given the observed data (for (d,x)) and a particular value of $(a_0(d,x), b_0(d,x)) = (\zeta_0^o(d,x), 1/\nu_0^o(d,x))$ is Gaussian with its posterior mean coinciding with expression for ζ_t^o. Moreover, the (subjective) mean of $Y(d,x)$ at instance t is given by $\sum_{o=0}^{L} \pi_0^o(d,x)\zeta_t^o(d,x)$, which is analogous to expression H.1 but with $\pi_0^o(d,x)$ instead of $\alpha_t^o(d,x)$.

We now show that by choosing $\pi_0^o(d,x)$ according to the empirical Bayes methodology (see Robbins (1992) and references therein) we recover $\alpha_t^o(d,x)$. That is, in this model, for each $(d,x) \in \mathbb{D} \times \mathbb{X}$, the likelihood over the observed outcome $Y^t := (Y_1(D_1,x),...,Y_t(D_t,x))$ given $D^t := (D_1,...,D_t) = d^t$ is indexed solely by the prior distribution $P_0(.|d,x)$ — in particular by $(\pi_0^o(d,x))_{o=0}^L$. By following the empirical Bayes methodology, one can "choose" $(\pi_0^o(d,x))_{o=0}^L$ to maximize such likelihood. The proposition below shows that such choice coincides with $(\alpha_t^o(d,x))_{o=0}^L$. Henceforth, we omit the dependence on x to simplify the notation.

Proposition H.1. *Let $f(Y^t \mid D^t; P_0)$ the likelihood over the observed outcome $Y^t := (Y_1(D_1),...,Y_t(D_t))$ given $D^t := (D_1,...,D_t)$ and prior distribution P_0. Then, for any $d \in \mathbb{D}$,*

$$(\alpha_t^o(d))_{o=0}^L = \arg\max_{P_0 \in \mathcal{P}(d)} \log f(Y^t \mid D^t = d^t; P_0),$$

where $\mathcal{P}(d)$ only considers probabilities with support on the points $(\zeta_0^o(d), 1/\nu_0^o(d))_{o=0}^L$.

Proof. Observe that the HBM gives a likelihood over the observed outcome $Y^t := (Y_1(D_1),...,Y_t(D_t))$ and $D^t := (D_1,...,D_t)$ is given by

$$\begin{aligned} f(Y^t \mid D^t) &= \int_\Theta f(Y^t \mid D^t, \theta)\,\Pr(d\theta) = \int_\Theta \prod_{d^t} p(Y^t(d^t) \mid d^t, \theta)^{1\{D^t = d^t\}}\,\Pr(d\theta) \\ &= \int_\Theta \prod_{s=1}^{t} \prod_{d^t} p(Y_s(d_s) \mid d_s, \theta)^{1\{D_s = d_s\}}\,\Pr(d\theta) \\ &= \int_\Theta \prod_{s=1}^{t} \prod_{d^t} p(Y_s(d_s) \mid \theta)^{1\{D_s = d_s\}}\,\Pr(d\theta), \end{aligned}$$

where the second line follows because the agent assumes that $Y_t(d)$ is independent of D_t; the third from the assumption that $Y(d)$ is viewed to be Gaussian with mean θ and variance 1 (it doesn't depend on d); and Pr is given by $\int \phi(\cdot; a, b) P_0(da, db|d)$. Thus, the likelihood depends on only one parameter, $P_0(.|d)$. To make it explicit, we use $f(.|.; P_0(.|d))$.

Now consider the following estimation problem for treatment $\mathbf{d} \in \mathbb{D}$,

$$\arg \max_{P_0 \in \mathcal{P}(\mathbf{d})} \log f(Y^t \mid D^t = \mathbf{d}^t; P_0),$$

which is equivalent to

$$\arg \max_{P_0(\cdot|\mathbf{d}) \in \mathcal{P}(\mathbf{d})} \log \int \left(\int_\Theta \prod_{s=1}^{t} p(Y_s(\mathbf{d}) \mid \theta)^{1\{D_s = \mathbf{d}\}} \phi(\theta; a, b) d\theta \right) P_0(da, db|\mathbf{d}). \tag{H.3}$$

Now consider the particular case where $\mathcal{P}(\mathbf{d})$ only considers probabilities with support on the points $(\zeta_0^o(\mathbf{d}), 1/\nu_0^o(\mathbf{d}))_{o=0}^L$. Then, it is clear that

$$\arg \max_{\alpha^0, \alpha^1, \ldots, \alpha^L \in \Delta_L} \log \sum_{o=0}^{L} \left(\int_\Theta \prod_{s=1}^{t} p(Y_s(\mathbf{d}) \mid \theta)^{1\{D_s = \mathbf{d}\}} \mu_0^o(d)(d\theta) \right) \alpha^o.$$

and the optimal choice is given by $(\alpha_t^o(\mathbf{d}))_{o=0}^L$. □

Hence, our model admits an alternative interpretation to our preferred one, based on an empirical HBM wherein the prior distribution P_0 is estimated *in each instance t*. This result is akin to certainty equivalence type results where a risk neutral agents acts as if there is no stochasticity in the underlying data.

We conclude by pointing out that the equivalence between these two interpretations breaks down if we consider a PM with a (strictly) concave ϕ in expression H.2. In the same way that certainty equivalence results break down if the agent is risk averse.

Online Supplemental Material

I General Learning Model

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

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

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

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

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

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

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

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

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

for any Borel set $A \subseteq \boldsymbol{\Theta}$, where

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

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

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

$$\int f(\boldsymbol{\theta})\mu_1^o(x)(d\boldsymbol{\theta}) = \int f(\boldsymbol{\theta}_0,...,\boldsymbol{\theta}_M) \frac{p_{\theta,d}(Y_1(x))\mu_0^o(d,x)(d\theta)\prod_{d'\neq d}\mu_0^o(d',x)(d\theta)}{\int_\Theta p_{\theta,d}(Y_1(x))\mu_0^o(d,x)(d\theta)}$$

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

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

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

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

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

A more succinct notation that captures these nuances is given by

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

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

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

for any $t \geq 1$.

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

J Examples of policy rules and their corresponding exploration structure

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

J.1 Examples of Policy Rules.

We present a series of examples of policy rules — and their associated exploration structure — in the context of the Gaussian learning framework.

Example 2 (Generalized ϵ-Greedy Policy Rule)**.** *A commonly-used policy function that is admissible in our framework is the so-called Epsilon-Greedy policy rule, given by*

$$\delta_t(y^{t-1}, d^{t-1})(d|x) = (M+1)\,\epsilon\,\frac{1}{M+1} + (1-(M+1)\,\epsilon)1\{d = \arg\max_a \zeta^{\alpha}_{t-1}(a,x)\},\ \forall t. \tag{J.1}$$

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

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

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

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

Assumption 3. *There exists an $\epsilon \in (0, 1/(M+1))$ such that $\delta(\cdot)(\cdot|x) \geq \epsilon$ for all $x \in \mathbb{X}$.*

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

It is straightforward to show that a structure of exploration for this class of policy rules is given by $h_t(d|x) = \epsilon$ and $\omega_t(d,x) = 0$. $\triangle$

Example 3 (Optimal policy function)**.** *The optimal policy function of this problem solves the Bellman equation problem with a per-period payoff given by the $\sum_{x\in\mathbb{X}} \zeta^{\alpha}(d,x)$ (or some other aggregator for x). Our framework does allow for such policy function but there are no guarantees that it will have a non-trivial* *exploration structure. Instead, one can consider a "perturbed" version of the form:*[20]

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

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

Example 4 (Thompson Sampling & refinements)**.** *Sampling schemes like Thompson's (Thompson (1933)) and others can be viewed as $\delta_t(d|x)=\pi_t(d|x)$ where $\pi_t(d|x)$ is a probability that treatment d yields the highest expected outcome and it is associated with the beliefs of the PM at time t, $(\zeta_{t-1},\nu_{t-1},\alpha_{t-1})$. For instance, in Thompson sampling $\pi_t(d|x)$ is constructed using the posterior beliefs $\mu^{\alpha}_{t-1}(d,x)$.*

For Thompson sampling, it is easy to show that Assumption 3 holds within the Markov Gaussian model and with $Y(d,x)$ having bounded support. Thus the exploration structure is such that $\epsilon := \inf_t h_t(d|x)>0$ and $\omega_t(d,x)=0$. In the other cases, Assumption 3 may not hold if $\pi_t(d|x)$ fails to be uniformly bounded from below, but a non-trivial exploration structure can still be obtained exploiting the fact that the subjective probability has full support and that $Y(d,x)$ is bounded with high probability. The next proposition present a structure of exploration for this such case.[21]

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

$$\begin{aligned}h_t(d|x) =& t^{-1}\sum_{s=1}^{t}(1-\max_{o\in\{0,\ldots,L\}}\Phi(a'_s+b_s;0,1/(s+\nu_0^o(d,x))))\prod_{l\neq d}\prod_{o=0}^{L}\int_{-b_s+a'_s}^{b_s-a'_s}\phi(y;0,1/\nu_0^o(l,x))dy,\\ \omega_t(d,x) =& 1-e^{\sum_{s=1}^{t-1}\log(\min_{d,x}P(-a'_s\leq Y(d,x)\leq a'_s|d,x))},\end{aligned}$$

is an exploration structure for the Thompson sampling policy rule.

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

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

with $e_s(d|x) := (1-\max_{o\in\{0,\ldots,L\}}\Phi(a'_s+b_s;0,1/(s+\nu_0^o(d,x))))\prod_{l\neq d}\prod_{o=0}^{L}\int_{-b_s+a'_s}^{b_s-a'_s}\phi(y;0,1/\nu_0^o(l,x))dy$ as this implies that $h_t(d|x)=t^{-1}\sum_{s=1}^{t}e_s(d|x)$ is an exploration index.

To do this, let, for each $t\in\{1,\ldots,T\}$ and $\mathbf{a}' := (a'_s)_{s\leq t}>0$, $S_t(\mathbf{a}') := \{(Y_s(.,.))_{s=1}^{t-1}:\ \forall s\leq t-1,\ |Y_s(.,.)|\leq a'_s\}$.

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

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

Observe that under this set $\max_{d,x,o} |\zeta_s^o(d,x)| \le \max\{a'_s, |\zeta_0^o(d,x)|\} =: a_s$. Thus, it suffices to show that

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

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

$$\pi_s(d|x) = \Pr(\hat{\zeta}_s^\alpha(d,x) \ge \max_{a \ne d} \hat{\zeta}_s^\alpha(a,x)) = \int \Pr(\hat{\zeta}_s^\alpha(d,x) \ge z \mid z = \max_{u \ne d} \hat{\zeta}_s^\alpha(u,x)) \Pr(dz)$$

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

Thus,

$$\begin{aligned}
\pi_s(d|x) =& \Pr(\hat{\zeta}_s^\alpha(d,x) \ge \max_{u \ne d} \hat{\zeta}_t^\alpha(u,x)) = \int (1 - \sum_{o=0}^{L} \alpha_{t-1}^o(d,x)\Phi(z; \zeta_{s-1}^o(d,x), 1/\nu_{s-1}^o(d,x))) \Pr(dz) \\
\ge& \int_K (1 - \sum_{o=0}^{L} \alpha_{s-1}^o(d,x)\Phi(z; \zeta_{s-1}^o(d,x), 1/\nu_{s-1}^o(d,x))) \Pr(dz) \\
\ge& (1 - \max_o \max_{z \in K} \Phi(z; \zeta_{s-1}^o(d,x), 1/\nu_{s-1}^o(d,x))) \Pr(K)
\end{aligned}$$

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

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

$$\begin{aligned}
\pi_s(d|x) \ge& (1 - \max_o \Phi(b; \zeta_{s-1}^o(d,x), 1/\nu_{s-1}^o(d,x))) \Pr(K) \\
\ge& (1 - \max_o \Phi(b; -a, 1/\nu_{s-1}^o(d,x))) \Pr(K) \\
=& (1 - \max_o \Phi(a+b; 0, 1/\nu_{s-1}^o(d,x))) \Pr(K) \\
\ge& (1 - \max_o \Phi(a+b; 0, 1/(s + \nu_0^o(d,x)))) \Pr(K)
\end{aligned}$$

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

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

$$\Pr(|\max_{l \ne d} \hat{\zeta}_s^\alpha(l,x)| \le b) \ge \Pr(\max_{l \ne d} \max_o |\hat{\zeta}_s^o(l,x)| \le b) = \prod_{l \ne d} \prod_{o=0}^{L} \Pr(|\hat{\zeta}_s^o(l,x)| \le b).$$

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

$$\begin{aligned}
\Pr(|\hat{\zeta}_s^o(l,x)| \le b) =&(\Phi(b;\zeta_{s-1}^o(l,x),1/\nu_{s-1}^o(l,x)) - \Phi(-b;\zeta_{s-1}^o(l,x),1/\nu_{s-1}^o(l,x)))\\
=&(\Phi(b-\zeta_{s-1}^o(l,x);0,1/\nu_{s-1}^o(l,x)) - \Phi(-b-\zeta_{s-1}^o(l,x);0,1/\nu_{s-1}^o(l,x)))\\
\ge&(\Phi(b-a;0,1/\nu_{s-1}^o(l,x)) - \Phi(-b+a;0,1/\nu_{s-1}^o(l,x)))\\
\ge&(\Phi(b-a;0,1/\nu_0^o(l,x)) - \Phi(-b+a;0,1/\nu_0^o(l,x)))
\end{aligned}$$

where the third line follows because Φ is increasing in its first argument and the fourth line follows because $\nu_{s-1}^o \ge \nu_0^o$ and $b-a>0$.

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

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

$\triangle$