

# Adaptivity and Confounding in Multi-Armed Bandit Experiments

Chao Qin and Daniel Russo

Columbia University

June 14, 2022

## Abstract

We explore a new model of bandit experiments where a potentially nonstationary sequence of contexts influences arms’ performance. Context-unaware algorithms risk confounding while those that perform correct inference face information delays. Our main insight is that an algorithm we call deconfounded Thompson sampling strikes a delicate balance between adaptivity and robustness. Its adaptivity leads to optimal efficiency properties in easy stationary instances, but it displays surprising resilience in hard nonstationary ones which cause other adaptive algorithms to fail.

## 1 Introduction

Multi-armed (MAB) bandit algorithms are designed to adapt their experimentation rapidly as evidence is gathered. By quickly shifting measurements away from less promising actions or ‘arms’, they focus measurement effort where it is most useful. The efficiency benefits of adaptive allocation are highlighted in the foundational theory of [Lai and Robbins \[1985\]](#). They show that asymptotically optimal procedures sample each suboptimal arm with frequency inversely proportional to its squared distance from optimality<sup>1</sup>. Upper confidence bound (UCB) algorithms and Thompson sampling [[Lattimore and Szepesvári, 2020a](#)] are simple adaptive algorithms known to match this long-run behavior [[Lai and Robbins, 1985](#), [Cappé et al., 2013](#), [Kaufmann et al., 2012](#), [Agrawal and Goyal, 2013a](#)]. These flexible algorithms have then become the dominant approach to efficient experimentation in more complicated sequential decision-making problems, including reinforcement learning (RL). In RL, it is well understood that non-adaptive experimentation can be hopelessly inefficient<sup>2</sup>.

The adaptivity of MAB algorithms makes them fragile. As described in [Section 5](#), if nonstationary confounding factors influence arms’ performance, this can cause a UCB algorithm to mistakenly discard good arms or can introduce information delays that UCB algorithms struggle with. Classical randomized controlled trials (RCTs) are designed to mitigate this risk. They sample arms according to fixed proportions across time, giving up on adaptivity completely.

We explore this tension through a model of bandit experiments where a sequence of observed contexts influences arms’ performance. Special cases of the model include ‘easy’ stationary instances — where highly adaptive algorithms offer efficiency benefits — and ‘hard’ nonstationary ones — where a traditional RCT is difficult to beat. Our main insight is that an algorithm we call *deconfounded Thompson sampling* strikes a

---

<sup>1</sup>More generally, if rewards noise is non-Gaussian, sampling frequencies are inversely proportional to a certain Kullback-Leibler divergence.

<sup>2</sup>The insight dates back at least to [Kearns and Singh \[2002\]](#). See for instance [[Osband et al., 2019](#), Section 4] for a detailed explanation of the inefficiency of various simple exploration policies. To converge on optimal behavior, these can require a number of measurements that scales exponentially in the number of states, which is prohibitive. Successful algorithms need to efficiently navigate to states of the Markov decision process in order to learn about them, which requires adjusting how actions are selected based on early feedback about the true transition probabilities.

delicate balance between adaptivity and robustness. It has optimal efficiency properties in the easy stationary instances, but displays surprising resilience in hard nonstationary ones.

## 1.1 Our contributions

This paper makes several contributions. First, we introduce a new model of adaptive experiments. This model combines the information structure of a Bayesian linear contextual bandit problem [Li et al., 2010] with the population level decision objective of Bubeck et al. [2009] or Kitagawa and Tetenov [2018]. We highlight in Section 5 that substantial challenges arise when the context sequence may follow a nonstationary pattern. We describe two examples of this form. Example 2 describes a weeklong experiment where the rewards an arm generates depend on the day. In this case, the source of nonstationarity is an observed exogenous factor. In Example 1, we show that latent (i.e. unobserved) confounders can be modeled by taking the observed context to be an indicator of the time at which a decision was made. An exogenous Gaussian process then influences arms’ rewards and we give examples of how specifying a covariance structure allows a modeler to specify the severity of plausible nonstationarity. Section 5 shows that natural MAB algorithms fail in such problems and gives insight into why.

A key feature of the model is that the experimenter’s ultimate goal is to reach a decision involving some degree of standardization. That is, after the experiment they do not wish employ a policy which makes different decisions in every context. We discuss reasons that many real world experiments have this feature, including operational, social, consistency, and fairness benefits of standardization, and incentive-compatibility constraints that restrict which mechanisms can be implemented post-experiment. This form of decision-objective was popularized in the economics literature by Kitagawa and Tetenov [2018].

Second, in Section 6 we propose deconfounded Thompson sampling (DTS), a simple and flexible approach that can address such problems. Thompson sampling (TS) is one of the leading approaches to efficient exploration in both academia and industry. See Chappelle and Li [2011], Scott [2010], Russo et al. [2018] for background. DTS makes two critical modifications to TS as it would usually be applied in a contextual bandit problem. The first modification makes the algorithm suitable for learning about a decision which will be implemented across many future contexts. In the setting of Example 2, rather than sample an arm according to the posterior probability it is the best action for the current day — as would standard TS — DTS samples an arm according to the posterior probability it offers the best average performance throughout the week. The second modification makes the algorithm focus on post-experiment performance rather than solely on the reward generating during the experiment itself. For this, we adopt the top-two sampling strategy of Russo [2020]. This modification explores suboptimal arms more aggressively by running Thompson sampling until two distinct actions are drawn and then randomly picking among those “top-two”.

Third, in Section 7 we prove that DTS offers robust performance even in settings with severe risk of delay and confounding. We study what Athey and Wager [2021] call the *utilitarian regret*, by evaluating its expected value conditioned on an arbitrary sequence of contexts. We provide a finite-time bound that depends only on the information contained in the contexts and is completely independent of the order in which they arrive, demonstrating robustness to nonstationary confounders that are modeled by the algorithm. As discussed above, any algorithm must grapple with delayed information revelation due to the context sequence. This result also allows for an arbitrary delay in observing reward realizations. The analysis in this section seems to offer substantial innovation. Past proofs have analyzed TS as method that implicitly behaves optimistically in the face of uncertainty [Agrawal and Goyal, 2013b, Russo and Van Roy, 2014], but natural optimistic algorithms fail to address Example 2 so an alternative theory is required. Our proofs use inverse propensity weights implicitly in the analysis of the posterior distribution, highlighting the importance of randomization and offering an interesting connection to the causal inference literature.

Fourth, Section 8 proves a sharp asymptotic optimality result in settings with i.i.d contexts. The key to this result is characterizing the proportion of measurements DTS dedicates to each arm asymptotically. Because DTS adjusts how future measurements are collected as it learns about each arm, these limiting ratios depend on the true (and initially unknown) parameter vectors. The way DTS adapts its measurement effort is shown to maximize the asymptotic speed of learning: in a sense made formal later, the posterior

probability assigned to the true best arm converges to one at an exponential rate and with the best possible exponent. DTS also minimizes the average sample size required to attain a vanishing simple regret. That result is a fully “frequentist” guarantee. The results in this section build on recent analyses of top-two sampling algorithms due to Russo [2020], Qin et al. [2017] and Shang et al. [2020] as well as lower bound techniques of Chernoff [1959] and Garivier and Kaufmann [2016].

Informally, these theoretical results suggest that DTS strikes a delicate balance between ensuring robustness in the presence of confounders and maximizing efficiency by adjusting measurement effort in response to feedback. One can interpret our robustness guarantee, and its supporting proof, as reflecting an important *lack of adaptivity* of DTS when faced with certain nonstationary context sequences. Robust performance in a highly nonstationary problem like Example 2 reflects that the algorithm may continue sampling each arm with roughly equal probability, even if some arms do not perform well on early days of the week. The asymptotic optimality result instead stresses the *useful adaptivity* of DTS in the face of a more benign context sequence. The algorithm itself seems likely to balance between these two goals in a range of examples that lie in between the extremes studied in our theoretical results.

The paper focuses on a particular decision objective. The experimenter’s goal is to reach a completely standardized decision — a single decision that is then implemented across the future population — and we measure the regret incurred after the experiment but not during it. The tension between adaptivity and robustness is much broader, and algorithms based on Thompson sampling seem flexible enough to treat many other objectives. We discuss extensions in Appendix A.

## 2 Related literature

Some of the paper’s main contributions appear to be completely new to the literature. Examples 1 and 2, and the challenges they raise appear to be new. The DTS algorithm is new. The proof of its robustness in problems like Example 2 (see Prop 1), appears to use techniques that have no precedent in the literature. The paper also touches on many issues that have been previously studied in the literature, however. We try to give a full account of related ideas below.

**Adversarial nonstationarity and seeking the best-of-both-worlds.** Nonstochastic bandit models [Auer et al., 2002] allow the sequence of rewards for each arm to be generated by an adversary. See Lattimore and Szepesvári [2020a] for an overview. The decision-maker’s goal is to select a sequence of actions that perform nearly as well in as any single fixed action would have on the reward sequence.

Robustness comes at a cost, and so Bubeck and Slivkins [2012] design an algorithm that attains the “best-of-both worlds”. Their procedure behaves like an algorithm designed to work with i.i.d observations, and switches to behave like nonstochastic bandit algorithm if it detects that nature is not i.i.d. They prove both robustness in the nonstochastic world and a result about efficient adaptivity in the i.i.d world.

That work focuses on reward earned during the experiment. Jamieson and Talwalkar [2016] formulates a nonstochastic best-arm identification problem where the focus is on post-experiment performance. In this model, reward sequences can be arbitrary, as long as they have a cesaro limit. The goal is to identify the arm with the best limiting value. Abbasi-Yadkori et al. [2018] and Shen [2019] provide best-arm identification algorithms with guarantees in the nonstochastic case that still offer competitive performance in the stochastic case relative to algorithms designed for that setting. Especially related to us is the work of Abbasi-Yadkori et al. [2018]. They show there is an inherent price to being robust against adversarial reward sequences. They provide a robust algorithm whose instance-dependent guarantees in problems with i.i.d rewards are nearly optimal among robust algorithms, but algorithms designed only for the i.i.d setting offer provably better performance in such problems.

The algorithm design and modeling in this paper are considerably different from the literature above. At a high level, however, this paper also seeks a “best-of-both worlds” result. We strike a somewhat different balance than Abbasi-Yadkori et al. [2018]. Section 8 provides sharp asymptotic optimality guarantees in i.i.d settings which would not be matched by an algorithm like that of Abbasi-Yadkori et al. [2018] — both due to the inherent price of robustness and because they do not use contextual observations to reduce

the variance of their estimators. In a setting like Example 1, where the structured prior restricts which nonstationary reward patterns are plausible, we expect DTS to adapt much more quickly than the algorithm of Abbasi-Yadkori et al. [2018]. This comes at a cost, however, as DTS with that prior is unlikely to be robust if there is an adversarial pattern to rewards.

This literature has used nonstationary Bayesian models like Example 1 in some recent proofs. In particular, Bubeck and Eldan [2016], Lattimore and Szepesvári [2020b], Lattimore and Gyorgy [2021] have analyzed the value of adversarial bandit problems by using the minimax theorem to convert it into a Bayesian problem and then applying information theoretic proofs in the style of Russo and Van Roy [2016, 2018]. To our knowledge, these information theoretic proofs cannot be used to prove Proposition 1.

**Delayed observations in MABs.** A number of MAB papers establish theoretical bounds on regret when observations are subject to delay. See for instance Dudík et al. [2011a], Joulani et al. [2013], Zhou et al. [2019] and references therein. As noted in Section 5.2, information delays arise organically in our model as a consequence of particular nonstationary context sequence. In dealing with this, we developed analysis techniques that also apply when there reward observations are subject to delay. In particular, Proposition 1 shows formally that robust results for DTS are possible even when reward signals are observed only after some delay of *arbitrary length*. Since at least Chapelle and Li [2011], it has been recognized informally that Thompson sampling has some resilience to delayed feedback since it is a randomized algorithm. To our knowledge, the most closely related theoretical result to ours is one by Kandasamy et al. [2018] on parallelized Thompson sampling. They provide a bound on Bayesian simple regret that degrades if there is a very high degree of parallelization. By contrast, our bound allows for delay in observing rewards that is as long as the experiment itself, which is equivalent to a problem where all arm pulls must be decided in parallel at the start of the experiment. Notice that even without delayed reward observations, this magnitude of delayed learning arises implicitly in Example 2, since uncertainty does not resolve until the end of the experiment.

A related preprint by Wu and Wager [2022] was posted online contemporaneously with our paper. This work includes empirical studies showing that vanilla Thompson sampling outperforms many algorithms designed specifically to work with delayed rewards. They also establish problem-dependent regret bounds which have an additive dependence on the magnitude of delay.

**Asymptotic efficiency in MABs.** Sharp results regarding the asymptotic limits of adaptive experiments were worked out decades ago. The asymptotic efficiency limits in unstructured bandit problems were worked out by Lai and Robbins [1985], and extended soon thereafter to structured parametric bandits and control of structured Markov chains by Agrawal et al. [1989b,a], Graves and Lai [1997]. The aforementioned papers focus on the reward earned during the experiment. Much earlier work by Chernoff [1959], Albert [1961] works out closely related asymptotic limits for adaptive experiments where the emphasis is on the quality of decision post-experiment. These papers characterize complexity through a game between an player and nature. This game has a special structure in the problem of selecting the best arm, and specialized results in that case are given in Jennison et al. [1982], Chan and Lai [2006] and revisited more recently in Garivier and Kaufmann [2016]. See Lattimore and Szepesvári [2017], Johari et al. [2017], Massoulié and Xu [2018], Combes et al. [2017], Degenne et al. [2020b], Van Parys and Golrezaei [2020], Degenne et al. [2020a], Wang et al. [2021], Jourdan et al. [2021] or Kirschner et al. [2021] for recent work along these lines .

Two very recent papers by Kato and Ariu [2021] and Russac et al. [2021] provide contextual extensions of Chan and Lai [2006], Garivier and Kaufmann [2016]. As in our work, their goal is to reach an effective population-level decision at the end of the experiment. Their results are asymptotic and assume contexts are i.i.d, making them comparable to our study of efficient adaptivity in Section 8 but not our study of robustness to nonstationarity in Section 7. Their models can be viewed as a special case of a linear contextual model, like ours, but where the context vectors are standard basis vectors. Both papers characterize asymptotic limits through a two-player game of the style studied in Chernoff [1959], Garivier and Kaufmann [2016] and other papers above.

Section 8 makes two contributions beyond the aforementioned literature. First, we show that the

equilibrium of Chernoff’s two player game has a simple structure in our problem. This result allows us to restrict to algorithms we call *context independent*. [Russac et al. \[2021\]](#) writes down the min-max optimization problems defining sample complexity under both context-dependent and context-independent sampling rules, but does not observe that these are equal. DTS satisfies this context independence property, which greatly simplifies our analysis of it. In particular, we can mostly follow [Russo \[2016\]](#), [Qin et al. \[2017\]](#) and [Shang et al. \[2020\]](#), who showed that top-two sampling algorithms attain lower bounds in the style of [Chernoff \[1959\]](#) in problems without contexts.

**Exploration sampling.** Building on top-two Thompson sampling [[Russo, 2016](#)], [Kasy and Sautmann \[2021\]](#) proposed an algorithm called exploration sampling. It follows from Theorem 1, part 1, of that paper that exploration sampling behaves identically to top-two TS once the posterior has concentrated. Since Section 8 focuses on such an asymptotic regime, we expect a deconfounded variant of exploration sampling has a nearly identical guarantee and analysis. It is unclear whether the finite time results in Section 5 apply to exploration sampling. The main reason we use a top-two sampling strategy is because it can be implemented without explicitly computing the posterior distribution of the optimal action. See Section 6.

**Dynamic programming based approaches.** The celebrated Gittins index theorem characterizes the solution to discounted Bayesian MAB problems [[Gittins and Jones, 1974](#), [Gittins, 1979](#)]. Most changes to this model render the associated dynamic program (DP) intractable, including introducing correlated beliefs, contexts, or a focus on post-experiment performance. A number papers propose heuristic but effective approximations to DP solutions [[Chick et al., 2010](#), [Chick and Inoue, 2001](#), [Frazier et al., 2008, 2009](#), [Chick and Frazier, 2012](#), [Ryzhov et al., 2012](#), [Chick et al., 2021](#)]. This paper shares the modeling perspective of those above, but takes a very different algorithmic approach and differs in its emphasis on resilience to confounding.

**Causal inference in MABs and adaptive allocation for treatment effect estimation.** Somewhat orthogonal to our paper, but also featuring words like “confounding”, there is a substantial and growing literature on policy evaluation from logged contextual bandit data [see e.g. [Dudík et al., 2011b, 2014](#), [Swaminathan and Joachims, 2015](#), [Li et al., 2015](#), [Wang et al., 2017](#), [Dimakopoulou et al., 2017](#), [Kallus and Zhou, 2018](#), [Farajtabar et al., 2018](#), [Zhou et al., 2018](#), [Dimakopoulou et al., 2019](#), [Athey and Wager, 2021](#)]. The work of [Bareinboim et al. \[2015\]](#) studies a setting where logged data, which might be subject to unobserved confounding, is used to warm-start Thompson sampling. Unlike this line of work, we consider the design of an adaptive experimentation algorithm, so the way decisions are made is changing but under our control. The connection to issues of confounding is made formal in Section 5.4.

Another related literature studies adaptive experimentation with a focus on treatment effect estimation. A landmark work in this area is the biased coin design of [Efron \[1971\]](#). We refer the reader to [Bugni et al. \[2018\]](#), [Kato et al. \[2020\]](#), [Bhat et al. \[2020\]](#) for recent reviews in the literature, each offering a distinct perspective. These papers focus on minimizing the variance of estimators. Our focus is on the quality of the final decision, allowing our methods to gather few measurements of some inferior arms.

[Farias et al. \[2022\]](#) recently posted a preprint that is related to our work. Their model is most similar to Example 1, where an exogenous time trend shifts the rewards an arm generates. They assume access to observations of non-experimental units and use synthetic control techniques to estimate and control for the exogenous trend. One technical difference is that our result in Proposition 1 has no dependence on the dimension of the context space, suggesting that our algorithms pay a surprisingly minimal price by using contexts to deconfound. It is an open question whether such guarantees are possible in the setting of [Farias et al. \[2022\]](#).

### 3 Motivating discussion of model features



The model we study closely resembles standard models in the literature, but has some salient differences. To motivate the role of standardization and of contexts in our model, we begin with a prototypical product testing problem.

### 3.1 Product testing example

Figure 1 shows the shortcuts displayed at the top of the Spotify home page. This feature makes it easy for users to find their favorite or recently played content. A personalized machine learning algorithm is trained to predict some success label, like the user streams given an impression. Below we discuss how to model a hypothetical experiment that is conducted to optimize this portion of the page. We describe a specific example for purposes of illustration only and the discussion does not necessarily reflect the specifics of that product or the available data.

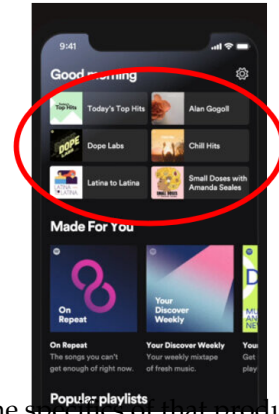


Figure 1: Shortcuts

- *Treatment arms:* An arm specifies a choice of the the user interface (UI), like whether to employ four, six, or eight items at at time, as well as which candidate machine learning algorithms is used to select and rank the displayed items.
- *Objective of the experiment:* The goal is to select a single arm that is later deployed across the whole population of users. Arms are evaluated based on the average reward they generate. The goal is *not* to learn a rule that changes the UI and the ML algorithm on the basis of a user’s recent behavior, the time of day, etc.
- *Reward measure:* A possible reward measure in this case might be the percentage of a user’s streams on the treatment day which are initiated from the home screen (rather than a user relying on search). This measure mirrors the product goal of making retrieval easy.
- *Context:* When a user is exposed to a treatment arm, variables other than the resulting rewards are logged. Many apps track measures like the user’s device, product plan, and age. Measures of a user’s long run and recent usage of the app prior to treatment are logged. Context could indicate whether the user is navigating via voice commands, which might make them more likely to use search. Contexts could also indicate whether there is a special promotion running, like the “Wrapped Campaign”, which changes the interface and diverts traffic. Finally, the timing of the treatment exposure is logged.

From this example, let us highlight two main insights:

1. The objective is to reach a *standardized* UI and choice of ML ranking algorithm.
2. The contextual factors may explain more of the variability in rewards than the treatment arms. These are observed passively without a need for experimentation.

We comment on the reasons for standardization in the next subsection. Now, let us comment on the role of context. In the example above, expected reward may vary substantially with the timing of the user’s exposure, whether the user is using voice commands, and whether they previously found a lot of their content from search. The impact of changing the architecture of a neural network used for item selection may be small compared to this exogenous variation. A data scientist could measure the frequency of contexts before the experiment. For instance, they could measure the fraction of users active on a given day who use ios, android, or a desktop and which fraction occur on weekdays vs weekends. During the time of the experiment itself, context is logged also for users who are held back from the experiment. Experiments are run to compare treatment decisions, not to learn about these contexts.

## 3.2 Reasons for standardization

Despite substantial possible benefits of personalization, public policies, operations processes, medical procedures, products, and prices are often relatively standardized. The reasons for this are varied and may be difficult to incorporate into a reward measure associated with an individual’s response to the treatment decision:

- *Operational benefits:* In the example given above, selecting a single UI and ML algorithm allows product designers and engineers to maintain and iterate on a standard product. Standardization is ubiquitous in mass-manufactured physical goods or in repeated operations involving humans because of efficiency benefits.
- *Fairness, ethical, or legal constraints:* In the year 2000, Amazon tested strategies which charged customers different prices for the same good<sup>3</sup>. They faced backlash from customers who believed the practice to be unfair. They appear not to have engaged in the practice since. Many forms of unequal treatment are not only perceived to be unfair, but are illegal in many countries.
- *Incentive compatibility constraints:* Consider an experiment designed to learn how to price. If the experiment selects a policy or pricing mechanism that charges different prices based on timing or past customer behavior, this mechanism may not be incentive compatible. Customers may respond optimally by modifying behavior to avoid price increases.
- *Social benefits:* On a social media platform, a dating app, or a two sided marketplace, standardizing the product for those who are posting content may improve the experience for those who consume that content. Digital education opens up the possibility of personalizing course content. However, a hidden cost of this is that students would not be able to easily discuss with each other.
- *Consistency benefits:* Users may expect a consistent and familiar experience. In the product testing example above, changing the UI based on the user’s last ten minutes of usage, or whether it is currently morning or evening, might create an erratic and frustrating experience.
- *Sample complexity benefits:* Much less data may be required to select a single arm than to identify a more complex policy. Our theory makes this formal.

For most of the paper, we focus on extreme case of standardization where the goal of the experiment is to commit to a single population-level decision. Generalizations of DTS that treat learning of a policy within a restricted class are briefly discussed in Section A.

## 4 Problem formulation

After running an experiment, a decision-maker must select among  $k$  arms. The performance of an arm depends on the context in which it is employed. Each context is represented by a  $d$  dimensional feature vector and the set of possible contexts is denoted by  $\mathcal{X}$ . For each arm  $i \in [k] := \{1, \dots, k\}$ , there is an uncertain arm-specific parameter  $\theta^{(i)}$ , which we model as a draw  $\theta^{(i)} \sim N(\mu_{1,i}, \Sigma_{1,i})$  from a multivariate Gaussian prior. We let  $\theta = (\theta^{(1)}, \dots, \theta^{(k)})$  denote the concatenation of the vectors. A linear function  $\mu(\theta, i, x) = \langle \theta^{(i)}, x \rangle$  determines the performance of arm  $i \in [k]$  in context  $x \in \mathcal{X}$ .

We assume the decision-maker has access to a probability distribution  $w$  over contexts that encodes the frequency with which they expect contexts to occur in the future. We call this either the *target distribution* or *population distribution*, where the latter suggests that  $w$  denotes the characteristics of a population of

---

<sup>3</sup><https://www.computerworld.com/article/2588337/amazon-apologizes-for-price-testing-program-that-angered-customers.html>

individuals. See Remark 1 for further discussion. If employed across a large number of future periods, arm  $i$  would generate average reward

$$\mu(\theta, i, w) := \sum_{x \in \mathcal{X}} w(x) \cdot \mu(\theta, i, x) = \langle \theta^{(i)}, X_{\text{pop}} \rangle \quad \text{where} \quad X_{\text{pop}} := \sum_{x \in \mathcal{X}} w(x)x. \quad (1)$$

If  $\mathcal{X}$  is uncountably infinite, the sum above is replaced with an integral. If the decision-maker knew  $\theta$ , the optimal arm to employ in the future would be  $I^* = I^*(\theta) \in \arg \max_{i \in [k]} \mu(\theta, i, w)$ . Many of the modeling choices made here were discussed at length in the previous section.

For technical or notational convenience, we make several additional assumptions. First, in both of our main results we will assume that the context vectors have uniformly bounded norm. Second, we assume that the arm-specific parameters  $\theta^{(i)}$  are drawn independently across arms, allowing us to track beliefs separately across arms in the analysis. Assume also that the prior covariance matrix  $\Sigma_{1,i}$  is the same for each arm  $i$  and is positive definite. Write  $\Sigma_1 \triangleq \Sigma_{1,1} = \dots = \Sigma_{1,k}$ .

**Sequential learning.** The decision-maker can reduce uncertainty about  $\theta$  through experimentation. In each period,  $t \in \mathbb{N} := \{1, 2, \dots\}$ , they select an arm  $I_t \in [k]$  in some context  $X_t \in \mathcal{X}$  and observe a real valued reward signal  $R_t = \langle \theta^{(I_t)}, X_t \rangle + W_t$ , where  $W_t \mid \theta, X_t \sim N(0, \sigma^2)$  is Gaussian noise drawn independently across time. Rewards are observed after a fixed<sup>4</sup> lag of  $L \geq 1$  periods. The information available when choosing  $I_t$  is the history  $H_t = (X_{1:t}, I_{1:t-1}, R_{1:t-L})$ . Formally, the action  $I_t$  must be chosen as a function of  $H_t$  and some random seed  $\zeta_t$  that is independent of all else. We assume the context sequence  $(X_t)_{t \in \mathbb{N}}$  is independent of  $\theta$ , so that the decision-maker cannot passively learn the impact of their actions by observing the contexts.

The distribution of  $\theta^{(i)}$  conditioned on  $H_t$  is multivariate Gaussian with covariance and mean given by  $\Sigma_{t,i} = \Sigma_{1,i}$  and  $\mu_{t,i} = \mu_{1,i}$  for  $t \leq L$  and

$$\Sigma_{t,i} = \left( \Sigma_1^{-1} + \sigma^{-2} \sum_{\ell=1}^{t-L} \mathbb{1}(I_\ell = i) X_\ell X_\ell^\top \right)^{-1} \quad \text{and} \quad \mu_{t,i} = \Sigma_{t,i} \left( \Sigma_1^{-1} \mu_{1,i} + \sigma^{-2} \sum_{\ell=1}^{t-L} \mathbb{1}(I_\ell = i) X_\ell R_\ell \right) \quad (2)$$

for  $t > L$ . Posterior beliefs about  $\theta$  induce posterior beliefs about  $I^*$ . We set  $\alpha_{t,i} = \mathbb{P}(I^* = i \mid H_t)$  for any period  $t \in \mathbb{N}$  and arm  $i \in [k]$ . Since  $\mu(\theta, i, w)$  is a linear function of  $\theta^{(i)}$ , it also has a Gaussian posterior. We write  $\mu(\theta, i, w) \mid H_t \sim N(m_{t,i}, s_{t,i}^2)$  where

$$s_{t,i}^2 = X_{\text{pop}}^\top \Sigma_{t,i} X_{\text{pop}} = \|X_{\text{pop}}\|_{\Sigma_{t,i}}^2 \quad \text{and} \quad m_{t,i} = \langle X_{\text{pop}}, \mu_{t,i} \rangle. \quad (3)$$

Notice that the Latin alphabet is used for the posterior parameters of the scalar quantity  $\mu(\theta, i, w)$  and the Greek alphabet is used for the posterior parameters of the vector  $\theta^{(i)}$ .

**Performance measures.** Let  $H_T^+ = (X_{1:T}, I_{1:T}, R_{1:T})$  denote all information generated by a  $T$ -period experiment, including the delayed reward outcomes. On the basis of the information in  $H_T^+$ , the decision-maker selects some arm  $\hat{I}_T^+$  (which implicitly we imagine being deployed throughout many future periods). For most of the paper, we focus on the *Bayes optimal selection rule*,

$$\hat{I}_T^+ \in \arg \max_{i \in [k]} \mathbb{E} [\mu(\theta, i, w) \mid H_T^+] \quad (4)$$

To show one (“frequentist”) impossibility result, we let  $\hat{I}_T^+$  be chosen as any function of  $H_T^+$  instead of specifying the Bayes selection is used. Notice that the decision  $\hat{I}_T^+$  can be made using the full results of the experiment  $H_T^+$  while a measurement decision  $I_t$  must be made in real-time based on partial information  $H_t$ .

<sup>4</sup>one could easily extend all results and algorithms to an alternate model where the delay at time  $t$  when arm  $i$  is chosen is  $L_{t,i} \leq L$ . The term  $L$  could be as long as the horizon of the problem and Proposition 1 has no dependence on  $L$ .



The non-negative random variable

$$\Delta_T = \mu(\theta, I^*, w) - \mu(\theta, \hat{I}_T^+, w)$$

measures the shortfall in future performance caused by selecting an arm  $I_T^+$  with only the incomplete information about  $\theta$  accrued after  $T$  measurements. We call  $\Delta_T$  the *simple regret* at time  $T$ , after [Bubeck et al. \[2009\]](#). Having in mind policy decisions where  $\mu(\theta, i, x)$  denotes the utility generated for an individual with feature  $x$ , [Athey and Wager \[2021\]](#) call this the *utilitarian regret*. The goal in the problem, informally, is to experiment intelligently so that simple regret is small after using as few measurements as possible. This objective can be formalized in two different ways in [Propositions 1 and 3](#), allowing us to give insight into different aspects of the performance of DTS.

**Remark 1** (On the population distribution). *In many experiments, one has access to information about the characteristics of a population, but does not know the performance of different treatment arms. If the contexts were i.i.d and representative of the target population, one could estimate  $w$  and  $X_{\text{pop}}$  during the early part of the experiment. Treating  $w$  as being derived from external data allows us to handle experiments where the contexts are not representative.*

*We typically have in mind the following choice for the population-weights  $X_{\text{pop}}$ . A tech company is running a very short experiment, perhaps lasting only a few days, to optimize some feature of the UI. The population weights are picked by setting  $X_{\text{pop}}$  to be the average among context vector among visitors to that page over the past three months. This is likely a decent reflection of visits to the page over the next three months. But the contexts distribution over, for instance, the first quarter of a short experiment is not necessarily representative of the population distribution. Some clear examples of this are discussed in the next section. There features like time-of-day or day-of-week are explicitly included as part of the context.*

## 5 Modeling of nonstationarity and the failure of alternative algorithms

### 5.1 Modeling of latent non-stationary confounders

The problem formulation seems, at first glance, to only capture observed contextual factors. The next example shows how latent confounders can be modeled as well. It is helpful to compare this example to the randomization based inference framework as applied to RCTs, which dates back to [Fisher \[1925\]](#). In that framework, one would not view the reward an arm generates as being a sample from some fixed population. Instead, potential outcomes (i.e. the samples one would observe if an arm were pulled that period) can follow some arbitrary pattern, and the objective is to measure an arm’s average performance in hindsight throughout the time of the experiment. The experimenter’s randomization serves as the “reasoned basis” for inference. This is similar to nonstochastic MAB models [[Auer et al., 2002](#)], where nature selects rewards adversarially and performance is assessed relative to the best fixed-arm in hindsight. The model below might be viewed as a Bayesian analogue where nature’s behavior is stochastic but nonstationary. The practitioner’s choice of a prior then governs the kinds of patterns the algorithm guards against.

**Example 1** (A Bayesian model of latent confounders). *Suppose  $X_t = e_t$  almost surely for each  $t \in [T]$  where  $e_t \in \mathbb{R}^T$  denotes the  $t$ -th standard basis vector. Let the target distribution  $w$  be uniform over  $\{e_1, \dots, e_T\}$ . The performance of arm  $i$  at time  $t$  is the  $t$ -th component of  $\theta^{(i)}$ , i.e.,  $\theta_t^{(i)} = \langle \theta^{(i)}, e_t \rangle$ , and the loss*

$$\Delta_T = \max_{i \in [k]} \left( \frac{\theta_1^{(i)} + \dots + \theta_T^{(i)}}{T} \right) - \left( \frac{\theta_1^{(\hat{I}_T^+)} + \dots + \theta_T^{(\hat{I}_T^+)}}{T} \right) \quad (5)$$

*measures the shortfall in average performance of the chosen arm  $\hat{I}_T^+$  across the contexts that occurred during the experiment. The vector  $\theta$  is drawn from a multivariate Gaussian distribution.*

*Imposing structure on this prior restricts the kinds of patterns that are likely. For instance, suppose  $\theta_t^{(i)} = \theta_0^{(i)} + \epsilon_t$  where the correlation  $\text{corr}(\epsilon_t, \epsilon_{\tilde{t}}) = \exp\{-|t - \tilde{t}|/\kappa\}$ . Then an exogenous process  $(\epsilon_t : t \in [T])$  shifts arms’ rewards*

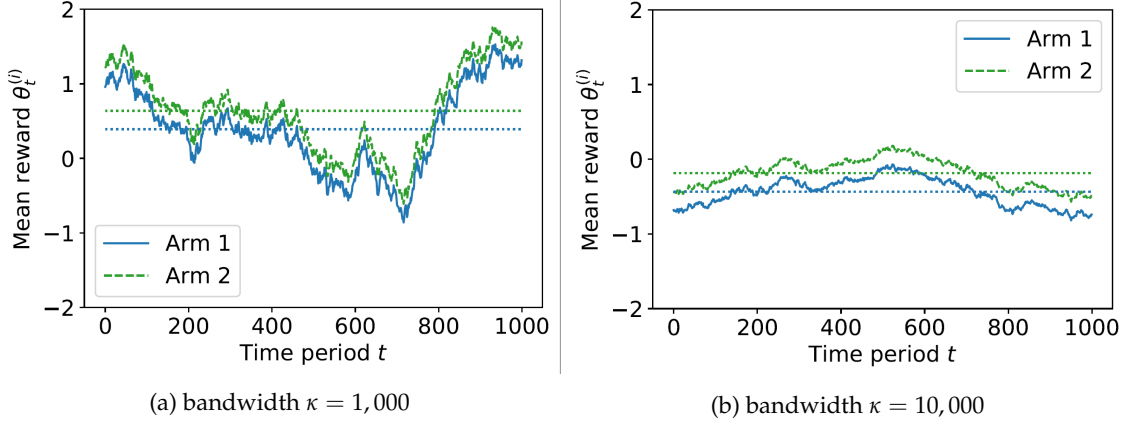


Figure 2: Two draws of  $\{\theta_t^{(i)}\}_{t \in [1000], i \in [2]}$  from the structured prior described in Example 1. The horizontal lines denote time averages and the vertical distance between them is  $\Delta_{1000}$ .

equally, as depicted in Figures 2a,2b. One might imagine an experiment running in December, where proximity to Christmas has a larger impact on behavior than the choice of arm. Take  $R_{t,i} = \theta_0^{(i)} + \epsilon_t + W_t$  to denote the potential reward outcome of arm  $i$ . When the time difference  $t - \bar{t}$  is small relative to the bandwidth parameter  $\kappa$ , the difference in rewards

$$R_{t,i} - R_{\bar{t},j} = \theta_0^{(i)} - \theta_0^{(j)} + (\epsilon_t - \epsilon_{\bar{t}}) + (W_t - W_{\bar{t}}) \approx \theta_0^{(i)} - \theta_0^{(j)} + (W_t - W_{\bar{t}})$$

behaves like a noisy sample of the difference in true qualities  $\theta_0^{(i)} - \theta_0^{(j)}$ . On the other hand, if one arm were measured only early in the experiment and another were measured only at the end, then the decision-maker might be unable to compare their performance.

## 5.2 Day-of-week effects

The next example illustrates the challenges that arise in special cases of the model where the context sequence has a nonstationary pattern. It describes a week-long experiment where observations are influenced by day-of-week effects, a standard concern in A/B testing practice [Kohavi et al., 2020].

**Example 2** (Day-of-week effects). Consider an online retailer who runs a weeklong experiment to find the profit maximizing price to charge for a product across future weeks. The distribution of demand in each period assumed to be normal, so a reward indicating the profit earned also follows a normal distribution. Different days of the week have different demand functions. This scenario can be modeled as a special case of the model above, where each context  $X_t \in \{e_1, \dots, e_7\} \subset \mathbb{R}^7$  is one of the standard basis vectors. Suppose  $T = 7m$  and the context at time  $t$  is  $X_t = e_{\lceil t/m \rceil}$ , meaning the first  $m$  periods are Sunday, the next  $m$  are Monday, and so on, with the final  $m$  being Saturday. The price is adjusted in each period and offered to the next customer or a small batch of customers, generating reward  $R_t = \langle \theta^{(I_t)}, X_t \rangle + W_t$ . Let the target distribution  $w$  be uniform over  $\{e_1, \dots, e_7\}$ . The performance of arm  $i$  on day  $x$  is the  $x$ -th component of the vector  $\theta^{(i)}$ , i.e.,  $\theta_x^{(i)} = \langle \theta^{(i)}, e_x \rangle$ . After the end of the experiment, the decision-maker picks a single price  $\hat{I}_T^+$  to employ across future weeks, and the loss

$$\Delta_T = \max_{i \in [k]} \left( \frac{\theta_1^{(i)} + \dots + \theta_7^{(i)}}{7} \right) - \left( \frac{\theta_1^{(\hat{I}_T^+)} + \dots + \theta_7^{(\hat{I}_T^+)}}{7} \right) \quad (6)$$

measures the shortfall in future profit due to making this choice with an incomplete resolution of uncertainty about average demand. Fairness and incentive-compatibility reasons to learn a single price were discussed earlier.

The decision-maker begins with prior belief under which  $\theta \sim N(\mu, \Sigma)$ . This might, for instance, arise from a latent

variable model where  $\theta_x^{(i)} = \theta_{i,x}^{\text{idio}} + \theta_i^{\text{arm}} + \theta_x^{\text{day}}$  is determined by an effect  $\theta_{i,x}^{\text{idio}}$  that is idiosyncratic to a specific arm and day, an effect  $\theta_i^{\text{arm}}$  associated with the chosen arm, and a shared day-of-week effect  $\theta_x^{\text{day}}$ . Placing an independent normal prior on the idiosyncratic, arm-specific, and day-specific effects induces a structured covariance matrix  $\Sigma$ . When the idiosyncratic terms have large variance, the decision-maker must guard against almost arbitrary nonstationary patterns. If these are believed to have smaller magnitude, the decision-maker may be able rule out some very poor arms early in the experiment.

Two notable challenges arise organically in this example:

1. *Distribution shift*: Observations early in the experiment reflect demand patterns early in the week. An algorithm that ignores day-of-week effects when performing inference risks confounding. It can incorrectly shift measurement effort away from an arm whose performance is subpar on a specific day but not average across all days.
2. *Information delay*: Observations early in the experiment reflect demand patterns early in the week. Under an algorithm that correctly models day-of-week effects when performing inference, playing an arm repeatedly on Monday will not resolve uncertainty about its performance. Regardless of how many customers arrive per day, uncertainty only fully resolves at the end of the week-long experiment.

Unfortunately, these features, make it difficult to apply common algorithms from the MAB literature to this problem.

### 5.3 Failure of context-unaware algorithms due to distribution shift

Building on the discussion in the previous subsection, this subsection highlights the risk of confounding for an algorithm that does not model day-of-week effects when performing inference.

We set

$$\tilde{s}_{t,i}^2 = \left( \tilde{s}_{1,i}^{-2} + \sigma^{-2} \sum_{\ell=1}^{t-1} \mathbb{1}(I_\ell = i) \right)^{-1} \quad \text{and} \quad \tilde{m}_{t,i} = \tilde{s}_{t,i}^2 \left( \sigma^{-2} \sum_{\ell=1}^{t-1} \mathbb{1}(I_\ell = i) R_\ell \right).$$

These are the posterior updating equations if  $\theta_1^{(i)} \sim N(0, \tilde{s}_{1,i}^2)$  and the algorithm (incorrectly) ignores day-of-week effects and assumes  $\theta_1^{(i)} = \theta_2^{(i)} = \dots = \theta_7^{(i)}$  almost surely. We define these expressions for  $\sigma^2 = 0$  by taking the limit as  $\sigma^2 \downarrow 0$ . In particular, we set  $\tilde{s}_{t,i}^2 = 0$  if arm  $i$  has been played previously and  $\tilde{m}_{t,i}$  to be 0 if arm  $i$  was never played previously and to be the empirical average reward otherwise.

Based on this, define context-unaware Thompson sampling. It chooses an arm at time  $t$  according to

$$I_t \in \arg \max_{i \in [k]} v_{t,i} \quad \text{where} \quad v_{t,i} \mid H_t \sim N(\tilde{m}_{t,i}, \tilde{s}_{t,i}^2). \quad (7)$$

In the above equation  $v_{t,1}, v_{t,2}, \dots, v_{t,k}$  are sampled independently. The next lemma formalizes that this algorithm risks confounding. The same result applies to a context-unaware UCB algorithm, which forms UCBs based on  $\tilde{m}_{t,i}$  and  $\tilde{s}_{t,i}^2$ . At a high-level, these algorithms fail because the way they perform inference does not reflect the problem's true information structure.

**Lemma 1** (Failure of context-unaware TS). *Consider Example 3. Suppose the components of the vector  $\theta = (\theta_x^{(i)})_{i \in [2], x \in [2]}$  are independent with  $\theta_x^{(1)} \sim N(0, 1)$  and  $\theta_x^{(2)} \sim N(0, 2)$  for  $x \in [2]$ . Suppose the noise variance  $\sigma^2 = 0$  and observations are not subject to delay (i.e.  $L = 1$ ). If (7) holds, there exists an absolute numerical constant  $c > 0$  such that for all  $T \in \mathbb{N}$ ,  $\mathbb{E}[\Delta_T] \geq c$ .*

Example 3 is a simplification of Example 2.

**Example 3** (Simplified day-of-week effects). *Consider a two-day experiment with  $k = 2$  arms and context set  $\mathcal{X} = \{e_1, e_2\} \subset \mathbb{R}^2$ . The context sequence is non-random, with  $X_t = e_1$  for  $t \leq \lfloor \frac{T}{2} \rfloor$ ,  $X_t = e_2$  for  $t > \lfloor \frac{T}{2} \rfloor$ , and the goal is to identify the best arm under equal context weights  $w$ . The components of vector  $\theta = (\theta_x^{(i)})_{i \in [2], x \in [2]}$  are*

independent and Gaussian with  $\theta_x^{(i)} = \langle \theta^{(i)}, e_x \rangle$  being the performance of arm  $i$  on day  $x$ . The reward at time  $t$  is  $R_t = \langle \theta^{(I_t)}, X_t \rangle + W_t$  where  $W_t \sim N(0, \sigma^2)$  is i.i.d Gaussian noise.

## 5.4 Formal discussion of confounding

In Subsection 5.3, we say informally that context-unaware UCB and TS risk confounding. Let us describe, in precise mathematical terms, what is meant by this.

Let  $R_t^{(i)} = \langle \theta^{(i)}, X_t \rangle + W_t$  denote the potential reward of arm  $i$  at time  $t$ . For arbitrary  $T \in \mathbb{N}$ , let  $\tau$  denote a time drawn uniformly at random from  $\{1, \dots, T\}$ , independent of all else. Then  $\mathbb{E}[R_{\tau, i}]$  denotes the average treatment effect of arm  $i$  across the contexts in the experiment. This is a primary quantity of interest in settings like Example 3 where the empirical distribution  $w_T(x) = \frac{1}{T} \sum_{t=1}^T \mathbb{1}(X_t = x)$  is close, in an appropriate sense, to the population distribution  $w$ .

By the construction of the algorithm, the following conditional unconfoundedness property holds:  $I_\tau \perp (R_\tau^{(i)})_{i \in [k]} \mid X_\tau$ . That is, the selected arm is independent of potential outcomes conditioned on context. But context-unaware algorithms like those in Subsection 5.3 do not store a record of context. We say their inferences are subject to confounding, since in general  $I_\tau \not\perp (R_\tau^{(i)})_{i \in [k]}$ . In particular  $\mathbb{E}[R_\tau^{(I_\tau)} \mid I_\tau = i] \neq \mathbb{E}[R_\tau^{(i)}]$ . The reason for this is adaptive selection together with nonstationarity of the context sequence. This allows for a pattern in arm selection to coincide with a pattern in the realized context sequence. The later could generate a pattern across time in the potential outcomes.

Our introduction of the random index  $\tau$  is nonstandard. One can interpret this as testing whether unconfoundedness holds after randomly permuting the dataset  $(X_t, I_t, R_t)_{t=1, \dots, T}$ . This is important because both the correct Bayesian inference (Sec 4) and the context-unaware procedure describe above update parameters in a manner that is invariant to the order in which past data was observed.

## 5.5 Failure of deconfounded UCB due to information delays

This section elaborates on the challenge of information delays mentioned in Subsection 5.2. Consider a UCB algorithm designed to correctly reflect the experimenter's goal and the problem's information structure. Reflecting that the true goal in Example 3 is to select an arm with strong performance throughout the week, not on a specific day, it plays the arm with the highest UCB on its average performance throughout the week:

$$I_t \in \arg \max_{i \in [k]} m_{t,i} + z \cdot s_{t,i} \quad \text{for all } t \in \mathbb{N}. \quad (8)$$

where  $m_{t,i}$  and  $s_{t,i}$  are defined in Equation (3) and  $z > 0$  is a tuning parameter. When  $z = 1.645$ , the term  $m_{t,i} + z \cdot s_{t,i}$  measures the 95% quantile of the posterior distribution. Note that (8) is equivalent to an algorithm whose optimism results from maximizing over parameters in an ellipsoidal confidence set<sup>5</sup>. This can be seen to be the UCB analogue of our proposed DTS algorithm, so we call it *deconfounded* UCB. It still selects the arm with the highest upside but accounts for observed contexts when performing inference.

The next result shows formally that deconfounded UCB fails to collect adequate data, regardless of the length of the time horizon. The issue is that the UCB in (8) is sometimes higher for action 2 for each of the first  $T/2$  periods. Action 1 is then never sampled in context 1, so learning is incomplete. This holds true regardless of how  $z$  is set and holds for time dependent tuning parameters. The issue is that, unlike common bandit settings, UCBs do not diminish when actions are repeatedly sampled in a single context. The main limitation of deconfounded UCB is its inability to cope with information delays.

**Lemma 2** (Failure of deconfounded UCB). *Consider Example 3. Suppose that the components of the vector  $\theta = (\theta_x^{(i)})_{i \in [2], x \in [2]}$  are independent with  $\theta_x^{(1)} \sim N(0, 1)$  and  $\theta_x^{(2)} \sim N(0, 3)$  for  $x \in [2]$ . Suppose the noise variance*

<sup>5</sup>In particular, consider the confidence set  $\Theta_{t,i} = \{\hat{\theta}^{(i)} : \|\hat{\theta}^{(i)} - \mu_{t,i}\|_{\Sigma_{t,i}^{-1}} \leq z\}$  derived from the level sets of the posterior. Then  $\max_{\hat{\theta}^{(i)} \in \Theta_{t,i}} \langle X_{\text{pop}}, \hat{\theta}^{(i)} \rangle = \langle X_{\text{pop}}, \mu_{t,i} \rangle + z \|X_{\text{pop}}\|_{\Sigma_{t,i}} = m_{t,i} + z \cdot s_{t,i}$ . So one can rewrite (8) as  $\arg \max_{i \in [k]} \max_{\hat{\theta}^{(i)} \in \Theta_{t,i}} \langle X_{\text{pop}}, \hat{\theta}^{(i)} \rangle$ .

$\sigma^2 = 0$  and observations are not subject to delay (i.e.  $L = 1$ ). If (8) holds, there is an absolute numerical constant  $c > 0$  such that for all  $T \in \mathbb{N}$ ,  $\mathbb{E} [\Delta_T] \geq c$ .

## 6 Deconfounded Thompson sampling

Now that we have had a thorough discussion of a new model of MAB bandit experiments and the challenges that arise, the rest of the paper focuses on understanding how deconfounded Thompson sampling overcomes these challenges.

DTS can be defined succinctly. At each time period  $t \in \mathbb{N}$ , it selects an arm to measure through the following procedure:

*Continue sampling from  $\alpha_t$  until two distinct arms are chosen.  
Flip a (biased) coin to select among these two.*

Recall that  $\alpha_t \in \mathbb{R}^k$  is defined by  $\alpha_{t,j} = \mathbb{P}(I^* = j \mid H_t)$ . We explain below how to efficiently sample from this distribution. Throughout the paper, we take  $\beta_t \in (0, 1]$  to be the probability the first sample from  $\alpha_t$  is played. By default, we recommend an unbiased coin ( $\beta_t = 1/2$ ) but this is discussed further below.

DTS can be understood as making two modifications to Thompson sampling in contextual bandits:

1. *Changing the learning target:* Thompson sampling for contextual bandits usually samples an action according to the probability it maximizes the mean reward in the current context. In particular, one sets  $\mathbb{P}(I_t = i \mid H_t) = \mathbb{P}(i = \arg \max_{j \in [k]} \mu(\theta, j, X_t) \mid H_t)$ . DTS is instead based on sampling from the posterior distribution of the arm  $I^*$ , which is the arm that maximizes the average reward in the target population rather than in the current context. Sampling from  $\alpha_t$  controls for confounders while directing exploration toward learning about the target arm of interest. Appendix B shows that the algorithm may fail without this change.
2. *Resampling:* Consider a problem without contexts. Then standard TS draws  $I_t$  from  $\alpha_t$ , without resampling. This algorithm is designed to maximize the reward earned throughout the experiment, implicitly imagining that the experimentation process never ends. But it performs poorly if there is an interest also in being able to rapidly stop and commit confidently to a decision. To understand the issue, imagine that  $\alpha_{t,1} = .95$ , so the algorithm believes there is a 95% chance that arm 1 is optimal. Then TS plays arm 1 in roughly 19/20 periods, making it very slow to gather information about alternatives. TS would be very slow to reach 99% confidence as a result and this is exacerbated if even higher confidence is desired.

To overcome this issue, [Russo \[2020\]](#) suggests a “top-two sampling” version of TS, which continues drawing arms from TS until two distinct options are drawn and then flips a biased coin to select among these two. To understand the resampling step, imagine that  $\alpha_{t,1} \rightarrow 1$  as  $t \rightarrow \infty$ . In this limit, the first sample from  $\alpha_t$  is nearly always arm 1 and this is played with probability  $\beta_t$ . Otherwise, an arm is chosen by resampling, and the chance of picking arm  $j \neq 1$  is roughly  $\mathbb{P}(I_t = j \mid H_t, I_t \neq 1) \sim \frac{\alpha_{t,j}}{1 - \alpha_{t,1}} = \mathbb{P}(I^* = j \mid H_t, I^* \neq 1)$ . Resampling shifts  $1 - \beta_t$  fraction of measurement effort away from arm  $I^*$  and assigns it to the strongest challengers. In particular, a challenger is sampled according to its conditional probability of being optimal.

It is worth noting that Proposition 1, which is about robustness to delay and non-stationary contexts, holds even if  $\beta_t = 1$  and there is no resampling. The resampling is only needed for the faster asymptotic rates in Section 8.

By default in this paper, we have in mind that DTS is implemented with a fair coin ( $\beta_t = 1/2$ ). Fixing a higher bias might be helpful to a practitioner. This would focus more measurement effort on the most promising arm, providing more confidence about the rewards it generates and reducing the expected regret incurred during the experiment. On the other hand, a longer experiment might be required to reach confidence about the best arm if a high bias is used. We discuss in Subsection 8.3 how the bias might be tuned adaptively



as data is observed to maximize certain asymptotic performance measures. Section 9 sketches a formal theory of how the choice of  $\beta$  trades off between the regret incurred in the experiment and the length of the experiment.

### Notable features of DTS.

Before proceeding, it is worth highlighting a few important features of DTS. First, let us draw a contrast with another popular strategy: UCB algorithms. These are based on the principle of *optimism in the face of uncertainty*. The decision-maker responds to uncertainty by playing whichever action is best in the best plausible model given current information. Notice that DTS, by default, *randomizes in the face of uncertainty*. Indeed, with a symmetric prior, one would have  $\alpha_{1,1} = \dots = \alpha_{1,k} = 1/k$  and so the initial arm  $I_1$  is sampled uniformly at random. As information is gathered, beliefs are updated and the decision-maker becomes less likely to sample inferior arms. In Examples 1 and 2, uncertainty is slow to resolve so  $\alpha_t$  may stay close to uniform for many periods. The algorithm’s randomization gives it a chance of sampling all plausibly optimal arms in all contexts. This appears to be critical to some of its robustness properties.

Another striking feature of the algorithm is that decisions at time  $t$  do not depend on the context at time  $t$ . That decisions are *context independent* in this way could offer substantial practical benefits. Even if contexts are logged, enormous engineering resources might be required to develop a system that observes contexts and responds in real time. For instance, assessing  $X_t$  could easily require querying several different datasets containing the current user’s interaction history and then applying a trained machine learning algorithm that generates a compact feature vector from this history. With a context independent algorithm, this could be done without substantial latency requirements.

### Efficient computation.

Following conventional implementation of Thompson sampling, a generic approach to sampling from  $\alpha_t$  is to sample a parameter vector  $\tilde{\theta}$  from the posterior distribution of  $\theta$  and then to find the arm  $\arg \max_{i \in [k]} \mu(\tilde{\theta}, i, w)$  that is best under this sample. The structure of Gaussian linear belief models allows for an even cleaner implementation of DTS. Because the population average reward of arm  $i$ ,  $\mu(\theta, i, w)$ , has a Gaussian posterior with posterior parameters given in (3), one can directly perform inference on the population average rewards.

The pseudocode below almost perfectly mirrors top-two TS in problems without contexts, except that the posterior parameters  $(m_{t+1,i}, s_{t+1,i}^2)$  are updated in a manner that controls for observed confounders, reflects the target population of contexts, and may be affected by delayed observations. By default, we imagine  $\beta_t = 1/2$ , but the pseudocode allows for adaptive tuning of the coin’s bias.

A possible concern is that it might take an enormous number of samples until the top-two arms differ (i.e. until  $I_t^{(1)} \neq I_t^{(2)}$ ). However, each fresh sample has chance  $1 - \alpha_{t,I_t^{(1)}}$  of generating a different arm, so this while-loop is expected to require many iterations only if the the posterior has already concentrated on a single action. In that case, it makes sense to terminate the experiment. When the posterior concentrates, there are also a variety of asymptotic approximations that could be used to calculate selection probabilities

and avoid repeated sampling<sup>6</sup>.

---

**Algorithm 1:** DTS allocation rule in Gaussian best-arm learning

---

Input prior parameters  $(\mu_{1,i}, \Sigma_{1,i})_{i \in [k]}$ , population weights  $X_{\text{pop}}$  and noise variance  $\sigma^2$ .

**for**  $t = 1, 2, \dots$  **do**

Sample  $v_i \sim N(m_{t,i}, s_{t,i}^2)$  for  $i \in [k]$  and set  $I_t^{(1)} = \arg \max_{i \in [k]} v_i$ ;

**do**

Sample  $v_i \sim N(m_{t,i}, s_{t,i}^2)$  for  $i \in [k]$  and set  $I_t^{(2)} = \arg \max_{i \in [k]} v_i$ ;

**while**  $I_t^{(1)} = I_t^{(2)}$ ;

Flip coin  $C_t \in \{0, 1\}$  with bias  $\mathbb{P}(C_t = 1) = \beta_t$ ;

Play arm  $I_t = I_t^{(1)}C_t + I_t^{(2)}(1 - C_t)$ ;

Gather delayed observation  $o = (I_{t-L+1}, X_{t-L+1}, R_{t-L+1})$ ;

Calculate posterior parameters  $m_{t+1,i}, s_{t+1,i}^2$  for  $i \in [k]$  according to (3) to reflect  $o$ ;

Calculate new tuning parameter  $\beta_{t+1}$  if using adaptive tuning;

**end**

---

## 7 Result 1: robustness to delay and confounding

Here we provide the first of two guarantees for DTS. The focus here is on assurances of robustness. We do this by establishing generic bounds on simple regret that essentially mirror regret guarantees satisfied when actions are selected uniformly at random. The challenge is to show that the adaptivity of DTS does not make the algorithm's performance brittle, in contrast to the algorithms described in Section 5. In the next section, we complement this study of robustness with a study of the efficiency benefits of DTS's adaptivity.

The next subsection describes a bound on simple regret. The following subsection then explains the novel analysis technique used to establish this bound, which implicitly uses inverse propensity weights [Horvitz and Thompson, 1952] to show the posterior of the best arm concentrates once enough contexts are observed. A challenge in the proof is that DTS violates the so-called overlap condition, which is core to most understanding of propensity score methods [Angrist and Pischke, 2008]. The randomized nature of deconfounded TS underlies our analysis technique, and given the failure of deconfounded UCB this appears to be more than an artifact of the chosen proof technique.

### 7.1 Performance guarantee

Because we do not require contexts to be i.i.d, there is no guarantee that the observed context sequence provides the information required to select the best-arm. We measure the information-content of the contexts through the quantity

$$V(X_{1:T}) = X_{\text{pop}}^\top \left( \Sigma_1^{-1} + \sigma^{-2} \sum_{t=1}^T X_t X_t^\top \right)^{-1} X_{\text{pop}}. \quad (9)$$

The matrix  $\left( \Sigma_1^{-1} + \sigma^{-2} \sum_{t=1}^T X_t X_t^\top \right)^{-1}$  appearing in (9) would be the posterior covariance matrix of  $\theta^{(i)}$  at the end of the experimentation horizon if that arm were played in every period. We similarly think of  $V(X_{1:T})$  as the posterior variance  $\text{Var}(\mu(\theta, i, w) \mid H_T^+)$  of the population effect of arm  $i$  if we observed the reward it generated in every period of the experiment. Notice that what makes the day-of-week effects in Example 3 challenging is *the order* in which contexts arrive. One can show that in that example  $V(X_{1:T}) \leq \sigma^2/T$ , reflecting that observing a single arm throughout the entire experiment would be informative.

---

<sup>6</sup>For instance, in the do-while loop we could select arm  $I_t^{(2)}$  by sampling with  $\mathbb{P}(I_t^{(2)} = i \mid H_t, I_t^{(1)}) \propto \Phi(-Z_{t,i^{(1)},i})$  where  $\Phi(\cdot)$  is the CDF of standard normal distribution and the z-values are defined in (13). The z-values underlie our asymptotic analysis, and we conjecture this version of DTS would attain the same asymptotic guarantees.

If arms were selected uniformly at random, we might expect the posterior variance of each one to scale roughly as  $k \cdot V(X_{1:T})$ , reflecting that information is divided equally across the arms. The next result establishes a simple regret bound for DTS that scales as  $\sqrt{k \cdot V(X_{1:T})}$ . One can think of this result as indicating a robustness property: the algorithm can cope with arbitrary context order and delayed reward observations, offering a guarantee matching what we would attain under a uniform allocation even when the context order and delay are severe. This holds even in settings like Example 1, where nature might pick complex nonstationary patterns. Of course, DTS is actually a highly adaptive algorithm, so it is subtle to show it satisfies this kind of robustness property and avoids the pitfalls described in Section 5.

For random variables  $X$  and  $Y$ , let  $\mathbb{H}(X)$  and  $\mathbb{H}(X|Y)$  denote the Shannon entropy and conditional Shannon entropy of  $X$ .

**Proposition 1.** *Suppose that  $\|X_t\|_2 \leq 1$  almost surely for  $t \in \mathbb{N}$ . If DTS is applied with tuning parameters satisfying  $\inf_{t \in \mathbb{N}} \beta_t \geq \frac{1}{2}$  almost surely and with the Bayes optimal selection rule in (4), then for any  $T \in \mathbb{N}$ ,*

$$\mathbb{E} [\Delta_T \mid X_{1:T}] \leq 2\sqrt{\iota \cdot k \cdot \mathbb{H}(I^* \mid H_T^+) \cdot V(X_{1:T})}$$

where  $\iota = \max \left\{ 9 \log \left( d \lambda_{\max}(\Sigma_1) \left[ \lambda_{\max}(\Sigma_1^{-1}) + \sigma^{-2} T \right] \right) \cdot \lambda_{\max}(\Sigma_1), 9 \right\}$ .

Notice that this result makes no assumptions about the delay  $L$ . The delay could be large enough ( $L \geq T$ ) to completely inhibit adaptivity during experimentation or short enough ( $L = 1$ ) that there are no explicit delays in reward observations. The term  $\iota$  in the bound above comes from applying a matrix concentration inequality to control the impact of the algorithm's action randomization. The bound applied there is crude and that term can almost certainly be tightened. The conditional entropy  $\mathbb{H}(I^* \mid H_T^+)$  is always smaller than the entropy term  $\mathbb{H}(I^*)$  that appears in Russo and Van Roy [2016]. The conditional entropy is upper bounded by  $\log(k)$ , but it can be much smaller if the decision-maker is likely to be confident about the identity of the best arm at the end of the problem. The requirement that  $\beta_t \geq \frac{1}{2}$  is satisfied by the default recommendation of using an unbiased coin ( $\beta_t = \frac{1}{2}$ ). If this is relaxed, the result would instead depend on  $(\inf_{t \in \mathbb{N}} \beta_t)^{-1}$ .

Under a natural condition that ensures the context sequence contains sufficient information about the population distribution, the next corollary of Proposition 1 gives a bound on simple regret that scales as  $\tilde{O}(\sqrt{k/T})$ . This result is nearly independent of the dimension of the linear model  $d$ , reflecting the sample complexity benefits of standardization described in Section 3.2. To understand the condition we place on the context sequence, imagine that  $X_{1:T} = x_{1:T}$  and, temporarily, define  $\bar{x} = \frac{1}{T} \sum_{t=1}^T x_t$  to be the empirical mean and  $\text{Cov}_x = \frac{1}{T} \sum_{t=1}^T (x_t - \bar{x})(x_t - \bar{x})^\top$ . Then  $\frac{1}{T} \sum_{t=1}^T x_t x_t^\top = \bar{x} \bar{x}^\top + \text{Cov}_x \succeq \bar{x} \bar{x}^\top$ . Our condition on the contexts always holds if  $\bar{x} = X_{\text{pop}}$ . If context vectors also have high variance in every direction, the bound in this corollary is conservative.

**Corollary 1.** *Under the conditions of Proposition 1, for any sequence  $x_{1:T} \in \mathcal{X}^T$  with  $\frac{1}{T} \sum_{t=1}^T x_t x_t^\top \succeq X_{\text{pop}} X_{\text{pop}}^\top$ ,*

$$\mathbb{E} [\Delta_T \mid X_{1:T} = x_{1:T}] \leq 2\sigma \sqrt{\frac{\iota \cdot k \cdot \mathbb{H}(I^* \mid H_T^+)}{T}} \leq 2\sigma \sqrt{\frac{\iota \cdot k \cdot \log(k)}{T}}$$

where  $\iota$  is given in Proposition 1.

Let us compare this result to what we would attain if there were no contexts (i.e.  $X_t = 1$  almost surely). Using Proposition 8 of Russo and Van Roy [2018], we can immediately convert expected regret bounds for Thompson sampling of Russo and Van Roy [2016] into a result of the form  $\mathbb{E}[\Delta_T] = O\left(\sigma \sqrt{\frac{k \log(k)}{T}}\right)$ . Here we get the same result, despite needing to control for a sequence of contexts.

The  $\tilde{O}(\sqrt{k/T})$  scaling here is reminiscent of ‘‘gap independent’’ regret lower-bounds in the multi-armed bandit literature. See Exercise 33.1 of Lattimore and Szepesvari [2020a] for a precise statement and their proof of Theorem 15.2 for a lower-bound instance which solves that exercise. Such lower bounds come from

constructing, for each fixed  $T$ , a prior distribution over instances under which any algorithm is unlikely to identify the best arm within  $T$  periods<sup>7</sup>. While suggestive, these lower bound instances use non-Gaussian priors and therefore do not apply in our setting. Extending our results to non-Gaussian priors is an open direction.

## 7.2 Proof outline

The proof of Proposition 1 is given in Appendix C. Here we give an outline of the main steps. The argument uses inverse propensity weighting implicitly in the analysis of the posterior distribution. In this way, the randomized nature of DTS is critical to the proof. This analysis appears to be quite different from others in the bandit literature and is one of the key innovations in this work.

Throughout the proof outline, assume  $X_{1:T}$  is nonrandom but arbitrary. Otherwise, one can condition on  $X_{1:T}$  in all expectations.

Set  $S_{t,i} = \text{Cov}(\theta^{(i)} | H_t^+) = \Sigma_{t+L,i}$  to be the posterior variance of  $\theta^{(i)}$  after observing reward generated by all  $t$  actions. This is a natural measure, because the final selection in period  $T$  is made based on the full history  $H_T^+$  (see Equation (4)). We define also *de-randomized* and *full-information* matrices by

$$\tilde{S}_{t,i} = \left( \Sigma_1^{-1} + \sigma^{-2} \sum_{\ell=1}^t \mathbb{P}(I_\ell = i | H_\ell) X_\ell X_\ell^\top \right)^{-1} \quad \text{and} \quad S_{\text{full}} = \left( \Sigma_1^{-1} + \sigma^{-2} \sum_{t=1}^T X_t X_t^\top \right)^{-1}.$$

The de-randomized matrix  $\tilde{S}_{t,i}$  is like  $S_{t,i}$ , but replaces  $\mathbb{1}(I_t = i)$  with the conditional probability  $\mathbb{P}(I_t = i | H_t)$ . The full information matrix imagines we were able to gather a measurement from every arm in every context.

Through two lemmas we establish the bounds on simple regret,

$$\mathbb{E} [\Delta_T] \leq \sqrt{2\mathbb{H}(I^* | H_T^+) X_{\text{pop}}^\top \mathbb{E} [S_{T,I^*}] X_{\text{pop}}} \leq \sqrt{2\iota \mathbb{H}(I^* | H_T^+) X_{\text{pop}}^\top \mathbb{E} [\tilde{S}_{T,I^*}] X_{\text{pop}}}, \quad (10)$$

where  $\iota$  is as in Proposition 1. The first inequality is Lemma 8 and it follows mainly from Russo and Zou [2019]. The second inequality is Lemma 10. That result builds on the proof of a matrix generalization of Friedman's inequality by Tropp [2011]. It requires careful technical analysis, however, which is given in Appendix C.5. The challenge is that we need a relation between matrices like  $S_{T,i} \preceq \iota \tilde{S}_{T,i}$  to hold with high probability, whereas Tropp [2011] bounds  $\lambda_{\max}(S_{T,i}^{-1} - \tilde{S}_{T,i}^{-1})$  in terms of  $\lambda_{\max}(\tilde{S}_{T,i}^{-1})$ .

Equation (10) bounds expected regret in terms of the expected (de-randomized) posterior variance at the end of the experiment. Crucially, the result only depends on the covariance matrix evaluated at the optimal arm  $I^*$ . This allows us to provide a bound for an algorithm like DTS, which may sample some arms very infrequently.

The next result uses an operator Jensen inequality to relate the (de-randomized) posterior covariance  $\tilde{S}_{T,I^*}$  to the sampling variance of an inverse-propensity weighted least squares estimator. In particular, consider

$$\hat{\theta}_T^{(i)} = \arg \min_{\hat{\theta} \in \mathbb{R}^d} \sum_{t=1}^T (X_t^\top \hat{\theta} - \hat{R}_{t,i})^2 + \sigma^2 \|\hat{\theta}\|_{\Sigma_1^{-1}}^2 \quad \text{where} \quad \hat{R}_{t,i} = \frac{R_t \mathbb{1}(I_t = i)}{\mathbb{P}(I_t = i | H_t)},$$

which is ridge regression applied to rewards that are de-biased by inverse propensity weighting. If  $\mathbb{P}(I_t = i | H_t)$  and the  $X_t$  were fixed and non-random, then the sampling variance of  $\hat{\theta}_T^{(i)}$  could be shown to equal

$$\text{Cov}(\theta_T^{(i)} | \theta, X_{1:T}) = S_{\text{full}} \mathbb{E} \left[ \Sigma_1^{-1} + \sigma^{-2} \sum_{t=1}^T \frac{X_t X_t^\top (1 + c_t)}{\mathbb{P}(I_t = i | H_t)} \mid \theta, X_{1:T} \right] S_{\text{full}}$$

<sup>7</sup>This differs from Bayesian prior-specific Lai [1987], Komiyama et al. [2021] asymptotics, where first the prior is fixed and then the time horizon is scaled to infinity.

where  $c_t = \sigma^{-2} (\mu(\theta, i, X_t))^2 (1 - \mathbb{P}(I_t = i | H_t)) \geq 0$  and this notation is used only temporarily to simplify this formula. Lemma 3 can therefore be viewed as giving a bound on the posterior covariance by a comparison to inverse propensity weighting. (This connection is useful for understanding, but not necessary for completing the proof.)

**Lemma 3** (Propensity matching type variance bound). *For any  $i \in [k]$ , with probability one,*

$$\tilde{S}_{T,i} \preceq S_{\text{full}} \left( \Sigma_1^{-1} + \sigma^{-2} \sum_{t=1}^T \frac{X_t X_t^\top}{\mathbb{P}(I_t = i | H_t)} \right) S_{\text{full}}.$$

Let  $\psi_{t,i} \triangleq \mathbb{P}(I_t = i | H_t)$  be the propensity assigned to arm  $i$  at time  $t$ . Under DTS, the propensities could be extremely small for some arms. Thankfully, we only need to study  $\tilde{S}_{T,I^*}$ , and we expect DTS will not neglect the optimal arm, i.e.  $\psi_{t,I^*}^{-1}$  will not be too large. In fact, we have  $\psi_{t,i} \geq \beta_t \mathbb{P}(I^* = i | H_t)$  where  $\beta_t \geq \frac{1}{2}$  by assumption, and so Lemma 11 in the appendix shows

$$\mathbb{E} \left[ \frac{1}{\psi_{t,I^*}} \right] = \mathbb{E} \left[ \sum_{i=1}^k \frac{\mathbb{1}(I^* = i)}{\psi_{t,i}} \right] = \mathbb{E} \left[ \sum_{i=1}^k \mathbb{E} \left[ \frac{\mathbb{1}(I^* = i)}{\psi_{t,i}} \mid H_t \right] \right] = \mathbb{E} \left[ \sum_{i=1}^k \frac{\mathbb{P}(I^* = i | H_t)}{\psi_{t,i}} \right] \leq 2k. \quad (11)$$

Notice that this property relies in a somewhat delicate way on the fact that propensities  $(\psi_{t,1}, \dots, \psi_{t,k})$  are “matched” to posterior beliefs  $(\alpha_{t,1}, \dots, \alpha_{t,k})$ . With (11), we are able to show  $\mathbb{E} [\tilde{S}_{T,I^*}] \leq 2k \cdot S_{\text{full}}$  and plugging this into (10) completes the proof.

The intuition behind this proof, following (11), is that DTS provides at least as much information about the true optimal arm as if fraction  $1/k$  of measurement effort were assigned to that arm in each context. In this way, DTS can never be too much worse than non-adaptive sampling. As described in Section 5, the closely related deconfounded UCB might, by contrast, fail to gather any information about the optimal arm in particular contexts.

## 8 Result 2: adaptivity and asymptotic optimality

Like most popular multi-armed bandit algorithms, DTS allocates measurement effort adaptively. As time proceeds, it learns about the quality of each arms and shifts measurement effort away from those that are clearly inferior. A higher fraction of samples are allocated to high quality arms, since the algorithm needs a more precise measurement of their quality in order to confidently distinguish them from the best arm. One can interpret Proposition 1 as reflecting a *lack of adaptivity* of DTS when faced with certain nonstationary context sequences. Robust performance in the day of week problem of Example 2, reflects that the algorithm may continue sampling each arm with roughly equal probability, even if some arms do not perform well on early days of the week. This section instead stresses the *useful adaptivity* of DTS in the face of a more benign context sequence. As a consequence of this adaptivity, we are able formalize a strong asymptotic optimality property DTS in settings where context vectors are drawn i.i.d.

In problems without contexts, the properties we highlight are largely known in the literature. They follow from analyses of top-two sampling algorithms initiated by Russo [2020] and extended in conference papers of Qin et al. [2017] and Shang et al. [2020] as well as lower bound techniques of Chernoff [1959] and Garivier and Kaufmann [2016]. Our work extends these properties to address contextual problems. Because DTS sampling probabilities are context-independent, as highlighted in Section 6, extending its asymptotic analysis to contextual case is technical demanding, but uses ideas similar to those appearing in past work. The main new insight is in the lower bound, where we need to show that context-independent sampling is optimal. This means that an algorithm that observes contexts after choosing an arm can perform as well asymptotically as one which observed contexts prior to arm selection. The key step there is in Proposition 5 in the appendix.

The results in this section have a very different style from the results in Section 5. It may be helpful to draw an analogy to the two styles of regret bounds that dominate the multi-armed bandit literature. One



style of result is “gap-independent” and tend to scale with the square-root of the time horizon. Matching lower bounds are constructed by, for a fixed time horizon, picking a hard instance  $\theta$  where it is difficult for any algorithm to identify the best arm within that horizon. At a high level, Corollary 1, which depends on the inverse square root of the time horizon, is an analogous bound for Bayesian best-arm identification. The second style of result, introduced by Lai and Robbins [1985] is a gap-dependent regret bound. One first fixes the parameter  $\theta$  determining all arm means and then scales the time horizon. Optimal algorithms depend just logarithmically on the time horizon—an exponential improvement—but the constant term is problem dependent and explodes if arm-means are not well separated. Algorithms which attain the optimal problem dependent constant must play sub-optimal arms with frequency exactly proportional to the inverse square of their optimality gap. Popular algorithms, like Thompson sampling and UCB, adapt their arm measurements in response to reward observations in a manner that naturally matches with these asymptotic limits. Our results in this section are analogous to these gap-dependent regret bounds, but where the performance measure stresses the regret incurred when implementing a decision post-experiment rather than the regret within the experiment. (See Appendix A for conjectures around the relations between these goals). Similar to the divide between polynomial and logarithmic cumulative regret, we move from the polynomial regret bounds in Corollary 1 to an exponential rate of convergence in Propositions 2 and 3. Our focus is essentially on characterizing the problem-dependent exponent governing the rate of convergence under DTS.

## 8.1 Preliminaries: assumptions and notation

We assume that contexts are drawn i.i.d from some distribution that is informative, in the sense that the second moment matrix is positive definite. For simplicity, we assume the context distribution is bounded. Notice that we still allow for distribution shift, since the average context vector  $\mathbb{E}[X_1]$  during the experiment may not equal the average in the target population,  $X_{\text{pop}}$ .

**Assumption 1.** Assume  $(X_1, X_2, \dots)$  are independent and identically distributed with bounded support and  $\Lambda \triangleq \mathbb{E}[X_1 X_1^\top] \succ 0$ .

In our general formulation, the experimenter may need to select arm  $I_t$  to measure using only the rewards  $R_1, \dots, R_{t-L}$  realized at least  $L$  periods ago. Here we assume  $L = 1$ . Our motivation for this assumption is to simplify the somewhat intricate proof of the upper bound portion of Proposition 3. We don’t believe it is necessary, and indeed, the proof sketches we give for Lemma 4 and Proposition 2, as well as the lower bound in Proposition 3 do not change if  $L$  exceeds one.

**Assumption 2** (No delay). Assume  $L = 1$ .

Our main results here are restricted to parameter vectors in a set  $\Theta$  under which each arm generates a distinct population average reward. Under the prior,  $\theta \in \Theta$  almost surely. Formally, define

$$\Theta \triangleq \left\{ \theta \in \mathbb{R}^{dk} : i \neq j \implies \mu(\theta, i, w) \neq \mu(\theta, j, w) \right\}. \quad (12)$$

In much of the analysis in this section, we consider the z-score for the difference in means,

$$Z_{t,i,j} \triangleq \frac{m_{t,i} - m_{t,j}}{\sqrt{s_{t,i}^2 + s_{t,j}^2}}. \quad (13)$$

This is a measure of the strength of evidence that one arm outperforms another in the population. Recall the definitions of  $m_{t,i} = \mathbb{E}[\mu(\theta, i, w) | H_t]$  and  $s_{t,i}^2 = \text{Var}(\mu(\theta, i, w) | H_t)$  given in Equation (3).

Finally, let  $\mathcal{S} = \{p \in \mathbb{R}_+^k : p_1 + \dots + p_k = 1\}$  denote the  $k - 1$  dimensional probability simplex.

## 8.2 Characterizing the speed of learning under context-independent sampling

The next two subsections are meant to be expository. Labeled as “proof sketches”, we offer short arguments that contain most of the key intuition behind the asymptotic behavior of DTS. These include (1) the proportion

of measurements DTS allocates to each arm asymptotically, (2) the speed at which the posterior converges, and (3) some reason to conjecture that no algorithm could outperform DTS. Our formal optimality guarantees are deferred to Subsection 8.4. The proofs of the main results largely supersede the arguments in the next two subsections, but those proofs are highly technical and a concise treatment may help the reader.

The main limitation of the next two subsections is that we restrict to analyzing algorithms under which sampling is *context independent* and *ergodic* in the sense below. Section 6 highlighted that DTS obeys the context independence property, but (15) would to be established. Formal lower bounds will need to show that algorithms not obeying these properties cannot offer better performance. Both issues are treated later. Notice that in this definition, the limiting proportions are allowed to depend on the draw of  $\theta$ .

**Definition 1.** A sampling rule is said to be context independent if

$$I_t \perp X_t \mid I_{1:(t-1)}, X_{1:(t-1)}, R_{1:(t-L)}. \quad (14)$$

It is said to be ergodic if there exists a strictly positive vector  $p = p(\theta) \in \mathcal{S}$  such that for each  $i \in [k]$ , with probability one,

$$\lim_{t \rightarrow \infty} p_{t,i} = p_i \quad \text{where} \quad p_{t,i} \triangleq \frac{1}{t} \sum_{\ell=1}^t \mathbb{1}(I_\ell = i). \quad (15)$$

In the next lemma, we study the speed at which the posterior beliefs convergence to the truth when samples are gathered in a manner that is context independent and ergodic. It is shown that  $\alpha_{t,I^*} \rightarrow 1$  at an exponential rate as  $t \rightarrow \infty$  and the precise exponent is characterized in terms of the limiting sampling proportions in (15). It worth mentioning that, despite the use of context-independent sampling, observing contexts allows for faster posterior convergence since realized contexts “explain” some of the variance in realized rewards.

**Lemma 4.** Suppose Assumptions 1 and 2 hold and the sampling rule satisfies (14) and (15). Then, with probability one

$$\lim_{t \rightarrow \infty} \frac{1}{t} \log(1 - \alpha_{t,I^*}) = - \min_{j \neq I^*} \frac{(\mu(\theta, I^*, w) - \mu(\theta, j, w))^2}{2\sigma^2 \|X_{\text{pop}}\|_{\Lambda^{-1}}^2 (p_{I^*}^{-1} + p_j^{-1})}. \quad (16)$$

*Proof.* Proof sketch of Lemma 4. It is easiest to understand this analysis as showing that (16) holds almost surely conditioned on the draw of state of nature  $\theta \in \Theta$ . (Recall that  $\theta \in \Theta$  with probability 1). Functions of  $\theta$  like  $I^*$ ,  $p = p(\theta)$ , or  $\mu(\theta, i, w)$  are nonrandom conditioned on the state of nature. A quantity like  $\alpha_{t,i}$  is a function of the observed data and therefore is a random variable.

Using an appropriate law of large numbers and Assumption 1, one can show that,

$$\frac{1}{t} \Sigma_{t,i}^{-1} = \frac{1}{t} \left( \Sigma_1^{-1} + \sigma^{-2} \sum_{\ell=1}^{t-L} \mathbb{1}\{I_\ell = i\} X_\ell X_\ell^\top \right) \rightarrow p_i \sigma^{-2} \Lambda. \quad (17)$$

Recall  $\Lambda = \mathbb{E}[X_1 X_1^\top]$ . An immediate consequence is that  $t s_{t,i}^2 = X_{\text{pop}}^\top (t \Sigma_{t,i}) X_{\text{pop}} \rightarrow p_i^{-1} \sigma^2 \|X_{\text{pop}}\|_{\Lambda^{-1}}^2$ . Similarly, one can show that  $m_{t,i} \rightarrow \mu(\theta, i, w)$ . Using this, and the definition of the z-values in (13), we have

$$\lim_{t \rightarrow \infty} \frac{Z_{t,i,j}^2}{t} = \frac{(\mu(\theta, i, w) - \mu(\theta, j, w))^2}{\sigma^2 \|X_{\text{pop}}\|_{\Lambda^{-1}}^2 (p_i^{-1} + p_j^{-1})}. \quad (18)$$

Now consider  $\alpha_{t,i}$ , the chance arm  $i$  is optimal under the Gaussian posterior distribution maintained by the algorithm. Each  $Z_{t,i,j}$  follows a normal distribution with unit variance under the posterior. The probability an arm is the best under the posterior can be bounded in terms of such pairwise comparisons. We have  $\max_{j \neq i} \Phi(-Z_{t,i,j}) \leq 1 - \alpha_{t,i} \leq k \max_{j \neq i} \Phi(-Z_{t,i,j})$ , where  $\Phi(\cdot)$  denotes the CDF of the standard normal distribution (In words, the chance arm  $i$  is suboptimal under the posterior is at least as large as the chance any particular arm  $j$  is better and no larger than the sum of this probability across arms).

Using this, we find,

$$\lim_{t \rightarrow \infty} \frac{1}{t} \log(1 - \alpha_{t, I^*}) = \lim_{t \rightarrow \infty} \max_{j \neq I^*} \frac{1}{t} \log \Phi(-Z_{t, I^*, j}) = \lim_{t \rightarrow \infty} \max_{j \neq I^*} \left( -\frac{1}{2t} Z_{t, I^*, j}^2 \right) = -\min_{j \neq I^*} \frac{(\mu(\theta, I^*, w) - \mu(\theta, j, w))^2}{2\sigma^2 \|X_{\text{pop}}\|_{\Lambda^{-1}}^2 (p_{I^*}^{-1} + p_j^{-1})},$$

where the second equality uses that  $\log \Phi(-x) \sim -x^2/2$  as  $x \rightarrow \infty$  [Dembo and Zeitouni, 2009, Equation 1.1.3] (Recall that  $\Phi(-x) \approx e^{-x^2/2}$  for large  $x$ ).  $\square$

It is clear from this result that adjusting the long-run sampling proportions based on the state of nature  $\theta$  can result in faster learning. Define the optimal exponent and optimal long-run sampling ratios as

$$\Gamma_{\theta}^{-1} \triangleq \max_{p \in \mathcal{S}} \min_{j \neq I^*} \frac{(\mu(\theta, I^*, w) - \mu(\theta, j, w))^2}{2\sigma^2 \|X_{\text{pop}}\|_{\Lambda^{-1}}^2 (p_{I^*}^{-1} + p_j^{-1})} \quad \text{and} \quad p^*(\theta) = \arg \max_{p \in \mathcal{S}} \min_{j \neq I^*} \frac{(\mu(\theta, I^*, w) - \mu(\theta, j, w))^2}{2\sigma^2 \|X_{\text{pop}}\|_{\Lambda^{-1}}^2 (p_{I^*}^{-1} + p_j^{-1})}. \quad (19)$$

Critically, these long-run sampling ratios depend on the true state of nature  $\theta$ , which is unknown at the start of the problem. Only an adaptive algorithm, which adjusts how often an arm is measured as it learns about their quality, could match the convergence rate under  $p^*(\theta)$ .

The optimization problem defining  $p^*(\theta)$  has a well-known<sup>8</sup> solution from the literature on best-arm identification without contexts. See Equation (13) in Glynn and Juneja [2004]. The solution  $p^*$  is the unique vector satisfying:

$$\frac{\mu(\theta, I^*, w) - \mu(\theta, i, w)}{\sqrt{(p_{I^*}^*)^{-1} + (p_i^*)^{-1}}} = \frac{\mu(\theta, I^*, w) - \mu(\theta, j, w)}{\sqrt{(p_{I^*}^*)^{-1} + (p_j^*)^{-1}}} \quad \forall i, j \neq I^* \quad (20)$$

$$p_{I^*}^* = \sqrt{\sum_{i \neq I^*} (p_i^*)^2}. \quad (21)$$

We call (20) an *information balance* property. From the left-hand-side of (18), one can see that this property ensures  $Z_{t, I^*, j}$  grows at the same rate for every  $j \neq I^*$ .

### 8.3 The asymptotic sampling proportions under DTS maximize the speed of learning

This subsection continues providing “proof sketches” that illuminate the asymptotic behavior of DTS. Our analysis suggests DTS satisfies an information balance property like (20) automatically. However, one would need to adjust the  $\beta_t$  parameter carefully in order to align with (21). The following algorithm gives a simple method which adjusts  $\beta_t$  by pretending as if the posterior mean  $\mu_t = \mathbb{E}[\theta | H_t]$  is the truth. We discuss the advantages and drawbacks of tuning further in the conclusion.

---

#### Algorithm 2: Plug-in Optimal $\beta_t$

---

```

if  $\arg \max_{i \in [k]} m_{t,i}$  is not unique then
  | Set  $\beta_t = \beta_{t-1}$ , with initial default  $\beta_1 = 1/2$ .
end
else
  | Solve for  $\hat{p} = p^*(\mu_t)$  where  $p^*(\cdot)$  is defined in (19). Set  $\beta_t = \hat{p}_{\hat{I}_t}$  where  $\hat{I}_t = \arg \max_i m_{t,i}$ .
end

```

---

It turns out that  $\beta_t$  can be calculated by solving a one dimensional fixed-point equation. This could be implemented, for example, by using bisection search or by Newton’s method. If computation became a

<sup>8</sup>The optimization problem (19) appears in Glynn and Juneja [2004]. This is partly a coincidence. Since observation noise is Gaussian, a certain KL divergence is symmetric in its arguments, and our complexity measure aligns with Glynn and Juneja [2004]. More generally, the allocations appearing in Chernoff [1959], Kaufmann et al. [2016], Russo [2020] differ from that in Glynn and Juneja [2004].

concern there are several paths to making this step even more efficient<sup>9</sup>.

**Lemma 5.** *Suppose  $\arg \max_{i \in [k]} m_{t,i}$  is unique. Then*

$$\beta_t = \frac{1}{1 + \sum_{i \neq \hat{I}_t} \left[ (m_{t,\hat{I}_t} - m_{t,i})^2 y_t - 1 \right]^{-1}} \quad \text{where } y_t \text{ solves } \sum_{i \neq \hat{I}_t} \left[ (m_{t,\hat{I}_t} - m_{t,i})^2 y - 1 \right]^{-2} = 1.$$

The next proposition suggests that as DTS gathers information about  $\theta$ , it adjusts its measurement gathering to align with  $p^*(\theta)$ . This adaptivity of DTS was highlighted in the paper's motivation. Notice that to use this result, we would need to separately argue that DTS is ergodic, in the sense of (15). A much longer argument removes this condition. See Subsection G.

**Proposition 2.** *Suppose Assumptions 1, 2 hold and DTS is applied with  $\beta_t$  set by Algorithm 2. Then,*

$$\lim_{t \rightarrow \infty} p_{t,i} = p_i^*(\theta) \quad \forall i \in [k] \quad \text{and} \quad \lim_{t \rightarrow \infty} \frac{1}{t} \log(1 - \alpha_{t,I^*}) = -\Gamma_\theta^{-1}$$

both hold on any sample path where  $\lim_{t \rightarrow \infty} p_{t,i}$  exists and is strictly positive for each  $i \in [k]$ .

*Proof.* Proof sketch of Proposition 2. As in the proof of Lemma 4, it may be easiest to understand this analysis as a limit that holds almost surely conditioned on the draw of state of nature  $\theta \in \Theta$ . See the comment in that proof.

If  $\lim_{t \rightarrow \infty} p_{t,i} = p_i^*(\theta) \quad \forall i \in [k]$ , then the convergence of  $\alpha_{t,I^*}$  follows from Lemma 4.

Now, suppose that there exists some probability vector  $p = p(\theta)$  with  $\lim_{t \rightarrow \infty} p_{t,i} = p_i > 0$  for each  $i \in [k]$ . It follows as in the proof of Lemma 4 that the posterior mean converges, i.e.  $\mu_{t,i} \rightarrow \theta^{(i)}$  and  $m_{t,i} \rightarrow \mu(\theta, w, i)$ . From the conclusion of Lemma 4 we know that  $\alpha_{t,I^*} \rightarrow 1$ .

Using the definition of DTS in Section 6, one can show that the the probability  $\psi_{t,i} \triangleq \mathbb{P}(I_t = i \mid H_t)$  assigned to sampling arm  $i$  in period  $t$  has the explicit form

$$\psi_{t,i} = \alpha_{t,i} \left[ \beta_t + (1 - \beta_t) \sum_{j \neq i} \frac{\alpha_{t,j}}{1 - \alpha_{t,i}} \right]. \quad (22)$$

We first establish the information balance property (20). Suppose, by contradiction, that it does not hold and there exists some arm  $i \neq I^*$  with

$$\epsilon = \frac{(\mu(\theta, I^*, w) - \mu(\theta, i, w))^2}{2\sigma^2 \|X_{\text{pop}}\|_{\Lambda^{-1}}^2 (p_{I^*}^{-1} + p_i^{-1})} - \min_{j \neq I^*} \frac{(\mu(\theta, I^*, w) - \mu(\theta, j, w))^2}{2\sigma^2 \|X_{\text{pop}}\|_{\Lambda^{-1}}^2 (p_{I^*}^{-1} + p_j^{-1})} > 0. \quad (23)$$

In this case,  $p_i$  is too large for information balance to hold. We show however that (23) implies that  $\psi_{t,i} \rightarrow 0$  exponentially fast, implying  $p_{t,i} \rightarrow 0$  and yielding contradiction. To show  $\psi_{t,i}$  diminishes at an exponential rate, observe that for  $t$  large enough such that  $I^* = \arg \max_j \alpha_{t,j}$ , we have

$$\psi_{t,i} = \alpha_{t,i} \left[ \beta_t + (1 - \beta_t) \sum_{j \neq i} \frac{\alpha_{t,j}}{1 - \alpha_{t,j}} \right] \leq \alpha_{t,i} \beta_t + \alpha_{t,i} (1 - \beta_t) \frac{1}{1 - \alpha_{t,I^*}} \leq \frac{\alpha_{t,i}}{1 - \alpha_{t,I^*}}.$$

<sup>9</sup>Our results would not change if  $\beta_t$  were updated only in select periods. For instance, we could update  $\beta_t$  only in periods  $t = 10, 100, 1000, 10000 \dots$ . Another option is to use an incremental algorithm to track the solution to the fixed point equation. One can view the solution of the fixed point equation,  $y_t = f(m_t)$ , as smooth function of the posterior means  $m_t = (m_{t,i} : i \in [k])$ . Then, one has  $y_{t+1} = y_t + \langle \nabla f(m_t), m_{t+1} - m_t \rangle + O(\|m_{t+1} - m_t\|^2)$ , so incremental updates to  $y_t$  could be used if enough samples have been gathered from each arm and the posterior mean changes only slightly between periods.

We find,

$$\begin{aligned}
\limsup_{t \rightarrow \infty} \frac{1}{t} \log(\psi_{t,i}) &\leq \limsup_{t \rightarrow \infty} \frac{1}{t} [\log(\alpha_{t,i}) - \log(1 - \alpha_{t,I^*})] \\
&\leq \limsup_{t \rightarrow \infty} \frac{1}{t} \log(\Phi(-Z_{t,I^*,i})) + \min_{j \neq I^*} \frac{(\mu(\theta, I^*, w) - \mu(\theta, j, w))^2}{2\sigma^2 \|X_{\text{pop}}\|_{\Lambda^{-1}}^2 (p_{I^*}^{-1} + p_j^{-1})} \\
&= -\frac{(\mu(\theta, I^*, w) - \mu(\theta, i, w))^2}{2\sigma^2 \|X_{\text{pop}}\|_{\Lambda^{-1}}^2 (p_{I^*}^{-1} + p_i^{-1})} + \min_{j \neq I^*} \frac{(\mu(\theta, I^*, w) - \mu(\theta, j, w))^2}{2\sigma^2 \|X_{\text{pop}}\|_{\Lambda^{-1}}^2 (p_{I^*}^{-1} + p_j^{-1})} = -\epsilon,
\end{aligned}$$

where the second inequality uses Lemma 4 and the equality follows as in the proof of Lemma 4. In particular, use (18) together with the fact that  $\log \Phi(-x) \sim -x^2/2$  as  $x \rightarrow \infty$ . This concludes the proof by contradiction.

Condition (21) follows by construction of the tuning algorithm. Since  $\alpha_{t,i} \rightarrow 0$  for  $i \neq I^*$ , it is immediate from (22) that  $\psi_{t,I^*} - \beta_t \rightarrow 0$ . One can show that  $p^*(\cdot)$  is continuous at  $\theta$ , and so  $\lim_{t \rightarrow \infty} p^*(\mu_t) = p^*(\theta)$ . It follows that  $\beta_t \rightarrow p_{I^*}^*(\theta)$ . This gives us that the probability assigned sampling arm  $I^*$  in each period converges as  $\psi_{t,I^*} \rightarrow p_{I^*}^*(\theta)$ . As a result  $p_{t,I^*} \rightarrow p_{I^*}^*(\theta)$ .  $\square$

## 8.4 Main result 2: formal asymptotic optimality guarantees for DTS

The results above give sharp insight into the asymptotic behavior of DTS, but they don't provide a fully coherent performance guarantee. Indeed, for a "Bayesian" who views  $\theta$  as a random variable, the sample-path specific limit in Lemma 4 is peculiar. Integrating over states of nature  $\theta$  would be more conventional, but our analytical tools are not refined enough for this. For a "frequentist" who prefers to study performance conditioned on the state of nature, it would be desirable to have performance measure that does not involve posterior beliefs. Here we provide one such guarantee.

The next result studies an objective used in the classic work of Chernoff [1959]. We discuss connections to popular "fixed-budget" and "fixed confidence" problem formulations in Appendix E. We evaluate performance at a random time  $\tau$  at which the decision-maker decides to stop collecting measurements. We take the total expected cost incurred by running an adaptive experiment, when the true state of nature is  $\theta_0$ , to be

$$\mathbb{E}[c\tau + \Delta_\tau \mid \theta = \theta_0], \quad (24)$$

where  $\tau$  denotes the chosen stopping time,  $c > 0$  is a cost per-period of experimentation, and  $\Delta_\tau$  is the simple regret of the final decision. Sharp results can be established through asymptotic analysis as  $c$  tends to zero. This is a regime where the cost of gathering one more observation is negligible relative to the cost of committing to a sub-optimal final decision. It arises if one imagines the final decision will later be implemented for a very large number of periods<sup>10</sup>.

A natural procedure stops when the z-values from comparing against the estimated best-arm are uniformly large. Specifically, consider the stopping rule

$$\tau = \inf \left\{ t \in \mathbb{N} : \min_{j \neq \hat{I}_t} Z_{t, \hat{I}_t, j} \geq \gamma_t + \sqrt{\gamma_t} \right\} \quad \text{where} \quad \gamma_t = \sqrt{2 \log \left( \frac{t^3}{\delta} \right)} \quad \text{and} \quad \hat{I}_t \in \arg \max_{i \in [k]} m_{t,i}. \quad (25)$$

Here  $\delta$  is a parameter that controls the desired level of confidence.

The next result shows that the expected cost incurred by DTS when the state of nature is  $\theta = \theta_0$  is no greater than  $c [\Gamma_{\theta_0} \log(1/c)]$  as  $c \rightarrow 0$ . This result is possible only by using fewer than  $\Gamma_{\theta_0} \log(1/c)$  samples on average while ensuring simple regret at the stopping time is  $\tilde{O}(c)$ . The logarithmic dependence of the sample size on  $c$  mimics the exponential rate of convergence of the posterior we showed in the previous section. The rest of the proposition shows the upper bound cannot be improved upon. The notation

<sup>10</sup>Notice that if  $c = c'/n$  then (24) is equivalent to  $\mathbb{E}[c'\tau + n\Delta_\tau \mid \theta = \theta_0]$ , where  $c'$  is a fixed experimentation cost and  $n$  can be thought of as measuring the number of individuals in the population who will receive whatever treatment arm is chosen.



$f(c, \theta_0) = o_{\theta_0}(1)$  means that  $\lim_{c \rightarrow 0} f(c, \theta_0) = 0$ . The inclusion of a subscript is meant to highlight that  $\theta$  remains fixed as  $c$  varies.

To avoid any possible ambiguity, let us be precise about the set of procedures to which the lower bound applies. An admissible algorithm for this problem consists of a stopping rule, a sampling rule, and a selection rule. A stopping rule  $\tau$ , like (25), is a random time where  $\mathbb{1}(\tau \leq t)$  is a function of  $(X_{1:(t-1)}, I_{1:(t-1)}, R_{1:(t-1)})$ . A sampling rule, like DTS, defines a sequence of random variables  $(I_t : t \in \mathbb{N})$ . The arm  $I_t$  is a function of  $(X_{1:t}, I_{1:(t-1)}, R_{1:(t-1)}, \zeta_t)$  for some random seed  $\zeta_t$  that is independent of all else. A selection rule, like the Bayes selection in (4), defines  $(\hat{I}_t^+ : t \in \mathbb{N})$  where  $\hat{I}_t^+$  is a function of  $(X_{1:t}, I_{1:t}, R_{1:t})$ . The stopping rule must choose whether to stop in period  $t - 1$  before seeing the context, arm selection, and reward in round  $t$ . The sampling rule may pick arm  $I_t$  on the basis of the current context, but it cannot “peak” at the reward realization  $R_t$ . However, the predicted best arm  $\hat{I}_t^+$  is chosen *after* observing  $R_t$ .

**Proposition 3.** *Suppose Assumptions 1 and 2 hold. If DTS is applied with  $\beta_t$  set by Algorithm 2 and stopping time  $\tau$  defined in (25) with parameter  $\delta = c$ , and the Bayes optimal selection rule in (4), then as  $c \rightarrow 0$ ,*

$$\mathbb{E}[c\tau + \Delta_\tau \mid \theta = \theta_0] \leq [1 + o_{\theta_0}(1)]\Gamma_{\theta_0}c \log(1/c) \quad \text{for all } \theta_0 \in \Theta.$$

*Under any admissible sampling rule, selection rule, and stopping rule  $\tau = \tau(c)$ , if as  $c \rightarrow 0$ ,*

$$\mathbb{E}[c\tau + \Delta_\tau \mid \theta = \theta_0] < [1 + o_{\theta_0}(1)]\Gamma_{\theta_0}c \log(1/c) \quad \text{for some } \theta_0 \in \Theta,$$

*then*

$$\lim_{c \rightarrow \infty} \frac{\mathbb{E}[c\tau + \Delta_\tau \mid \theta = \theta_1]}{c \log(1/c)} = \infty \quad \text{for some } \theta_1 \in \Theta. \quad (26)$$

Recall that the proof sketches in the previous two subsections restricted to sampling rules which are context-independent, in the sense of (14), and ergodic, in the sense of (15). The lower bound here makes no such restrictions. This shows that context-independent sampling is asymptotically optimal, constituting one of the main insights of this section. The proofs are established in Appendix F, G. Part of the upper bound argument needs to make formal our proof sketches that  $p_{t,i} \rightarrow p_i^*(\theta)$  under DTS, with an appropriate notion of convergence. That result is given in Proposition 7. Its proof is omitted, since it mimics steps in Russo [2020], Qin et al. [2017], and Shang et al. [2020], but for completeness it is provided in a separate technical report [Qin and Russo, 2022].

## 9 Conclusion

This paper proposes a new model of bandit experiments. To optimally treat this range of examples it covers, an algorithm must strike a delicate balance between efficient adaptivity — the main focus of the MAB literature — and robustness to nonstationarity and confounding — the main focus of classical randomized controlled trials. The paper provides evidence that careful modifications to Thompson sampling, one of the most widely used MAB algorithms, allows an experimenter to strike such a balance.

Robustness to nonstationarity has been a focus of a segment of the MAB literature that models rewards a choice picked by an adversary [Auer et al., 2002]. We take a less conservative approach, allowing the decision-maker to express which nonstationary patterns are plausible, as in Example 1, or to view observable contexts as the only possible source of nonstationarity, as in Example 2. This kind of modeling flexibility is most interesting and valuable when motivated concretely by real world problems. One of the most exciting possibilities for future work is to test this approach with data arising from real experiments. In depth simulation experiments could also provide insight into sensitivity to model or prior mis-specification.

## References

Y. Abbasi-Yadkori, P. Bartlett, V. Gabillon, A. Malek, and M. Valko. Best of both worlds: Stochastic & adversarial best-arm identification. In *Conference on Learning Theory*, pages 918–949. PMLR, 2018.

- M. Abeille and A. Lazaric. Linear Thompson sampling revisited. *Electronic Journal of Statistics*, 11(2):5165–5197, 2017.
- R. Agrawal, D. Teneketzis, and V. Anantharam. Asymptotically efficient adaptive allocation schemes for controlled markov chains: finite parameter space. *IEEE Transactions on Automatic Control*, 34(12):1249–1259, 1989a.
- R. Agrawal, D. Teneketzis, and V. Anantharam. Asymptotically efficient adaptive allocation schemes for controlled i.i.d. processes: finite parameter space. *IEEE Transactions on Automatic Control*, 34(3):258–267, 1989b.
- S. Agrawal and N. Goyal. Further optimal regret bounds for Thompson sampling. In *Artificial intelligence and statistics*, pages 99–107. PMLR, 2013a.
- S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135, 2013b.
- A. E. Albert. The sequential design of experiments for infinitely many states of nature. *The Annals of Mathematical Statistics*, 32(3):774–799, 1961.
- J. D. Angrist and J.-S. Pischke. *Mostly harmless econometrics*. Princeton university press, 2008.
- K. Ariu, M. Kato, J. Komiyama, K. McAlinn, and C. Qin. Policy choice and best arm identification: Asymptotic analysis of exploration sampling. *arXiv preprint arXiv:2109.08229*, 2021.
- S. Athey and S. Wager. Policy learning with observational data. *Econometrica*, 89(1):133–161, 2021.
- P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.
- E. M. Azevedo, A. Deng, J. L. Montiel Olea, and E. G. Weyl. Empirical Bayes estimation of treatment effects with many a/b tests: An overview. In *AEA Papers and Proceedings*, volume 109, pages 43–47, 2019.
- E. Bareinboim, A. Forney, and J. Pearl. Bandits with unobserved confounders: A causal approach. *Advances in Neural Information Processing Systems*, 28:1342–1350, 2015.
- S. Basu, B. Kveton, M. Zaheer, and C. Szepesvári. No regrets for learning the prior in bandits. *Advances in Neural Information Processing Systems*, 34, 2021.
- N. Bhat, V. F. Farias, C. C. Moallemi, and D. Sinha. Near-optimal a-b testing. *Management Science*, 66(10):4477–4495, 2020.
- S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.
- S. Bubeck and R. Eldan. Multi-scale exploration of convex functions and bandit convex optimization. In *Conference on Learning Theory*, pages 583–589. PMLR, 2016.
- S. Bubeck and A. Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *Conference on Learning Theory*, pages 42–1. JMLR Workshop and Conference Proceedings, 2012.
- S. Bubeck, R. Munos, and G. Stoltz. Pure exploration in multi-armed bandits problems. In *International conference on Algorithmic learning theory*, pages 23–37. Springer, 2009.
- F. A. Bugni, I. A. Canay, and A. M. Shaikh. Inference under covariate-adaptive randomization. *Journal of the American Statistical Association*, 113(524):1784–1796, 2018.
- O. Cappé, A. Garivier, O.-A. Maillard, R. Munos, and G. Stoltz. Kullback-leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, pages 1516–1541, 2013.
- A. Carpentier and A. Locatelli. Tight (lower) bounds for the fixed budget best arm identification bandit problem. In *Conference on Learning Theory*, pages 590–604. PMLR, 2016.
- H. P. Chan and T. L. Lai. Sequential generalized likelihood ratios and adaptive treatment allocation for optimal sequential selection. *Sequential Analysis*, 25(2):179–201, 2006.
- O. Chapelle and L. Li. An empirical evaluation of Thompson sampling. *Advances in neural information processing systems*, 24:2249–2257, 2011.
- C.-H. Chen, J. Lin, E. Yücesan, and S. E. Chick. Simulation budget allocation for further enhancing the efficiency of ordinal optimization. *Discrete Event Dynamic Systems*, 10(3):251–270, 2000.
- H. Chen, W. Lu, and R. Song. Statistical inference for online decision making: In a contextual bandit setting. *Journal of the American Statistical Association*, pages 1–16, 2020.
- H. Chernoff. Sequential design of experiments. *Annals of Mathematical Statistics*, 30(3):755–770, 1959.
- S. E. Chick and P. Frazier. Sequential sampling with economics of selection procedures. *Management Science*, 58(3):550–569, 2012.
- S. E. Chick and K. Inoue. New two-stage and sequential procedures for selecting the best simulated system. *Operations Research*, 49(5):732–743, 2001.
- S. E. Chick, J. Branke, and C. Schmidt. Sequential sampling to myopically maximize the expected value of information. *INFORMS Journal on Computing*, 22(1):71–80, 2010.
- S. E. Chick, N. Gans, and Ö. Yapar. Bayesian sequential learning for clinical trials of multiple correlated medical interventions. *Management Science*, 2021.
- R. Combes, S. Magureanu, and A. Proutiere. Minimal exploration in structured stochastic bandits. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 1761–1769, 2017.
- R. Degenne, P. Ménard, X. Shang, and M. Valko. Gamification of pure exploration for linear bandits. In *International Conference on Machine Learning*, pages 2432–2442. PMLR, 2020a.

- R. Degenne, H. Shao, and W. Koolen. Structure adaptive algorithms for stochastic bandits. In *International Conference on Machine Learning*, pages 2443–2452. PMLR, 2020b.
- A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*, volume 38. Springer Science & Business Media, 2009.
- Y. Deshpande, L. Mackey, V. Syrgkanis, and M. Taddy. Accurate inference for adaptive linear models. In *International Conference on Machine Learning*, pages 1194–1203. PMLR, 2018.
- M. Dimakopoulou, Z. Zhou, S. Athey, and G. Imbens. Estimation considerations in contextual bandits. *arXiv preprint arXiv:1711.07077*, 2017.
- M. Dimakopoulou, Z. Zhou, S. Athey, and G. Imbens. Balanced linear contextual bandits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3445–3453, 2019.
- M. Dimakopoulou, Z. Ren, and Z. Zhou. Online multi-armed bandits with adaptive inference. *Advances in Neural Information Processing Systems*, 34, 2021.
- D. Dimmery, E. Bakshy, and J. Sekhon. Shrinkage estimators in online experiments. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2914–2922, 2019.
- M. Dudík, D. Hsu, S. Kale, N. Karampatziakis, J. Langford, L. Reyzin, and T. Zhang. Efficient optimal learning for contextual bandits. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 169–178, 2011a.
- M. Dudík, J. Langford, and L. Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 1097–1104, 2011b.
- M. Dudík, D. Erhan, J. Langford, and L. Li. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511, 2014.
- D. Eckles and M. Kaptein. Bootstrap Thompson sampling and sequential decision problems in the behavioral sciences. *Sage Open*, 9(2): 2158244019851675, 2019.
- B. Efron. Forcing a sequential experiment to be balanced. *Biometrika*, 58(3):403–417, 1971.
- M. Farajtabar, Y. Chow, and M. Ghavamzadeh. More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*, pages 1447–1456. PMLR, 2018.
- V. Farias, C. Moallemi, T. Peng, and A. Zheng. Synthetically controlled bandits. *arXiv preprint arXiv:2202.07079*, 2022.
- F. Finan and D. Pouzo. Reinforcing rcts with multiple priors while learning about external validity. *arXiv preprint arXiv:2112.09170*, 2021.
- R. A. Fisher. The design of experiments. *The Design of Experiments*, 1925.
- P. Frazier, W. Powell, and S. Dayanik. The knowledge-gradient policy for correlated normal beliefs. *INFORMS journal on Computing*, 21(4):599–613, 2009.
- P. I. Frazier, W. B. Powell, and S. Dayanik. A knowledge-gradient policy for sequential information collection. *SIAM Journal on Control and Optimization*, 47(5):2410–2439, 2008.
- A. Garivier and E. Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027. PMLR, 2016.
- J. C. Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2): 148–164, 1979.
- J. C. Gittins and D. M. Jones. A dynamic allocation index for the sequential design of experiments. In J. Gani, editor, *Progress in Statistics*, pages 241–266. North-Holland, Amsterdam, NL, 1974.
- P. Glynn and S. Juneja. A large deviations perspective on ordinal optimization. In *Proceedings of the 2004 Winter Simulation Conference, 2004.*, volume 1. IEEE, 2004.
- T. L. Graves and T. L. Lai. Asymptotically efficient adaptive choice of control laws in controlled markov chains. *SIAM journal on control and optimization*, 35(3):715–743, 1997.
- R. M. Gray. *Entropy and information theory*. Springer Science & Business Media, 2011.
- V. Hadad, D. A. Hirshberg, R. Zhan, S. Wager, and S. Athey. Confidence intervals for policy evaluation in adaptive experiments. *Proceedings of the National Academy of Sciences*, 118(15), 2021.
- J. Hong, B. Kveton, M. Zaheer, and M. Ghavamzadeh. Hierarchical Bayesian bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 7724–7741. PMLR, 2022.
- D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.
- C.-W. Hsu, B. Kveton, O. Meshi, M. Mladenov, and C. Szepesvári. Empirical Bayes regret minimization. *arXiv preprint arXiv:1904.02664*, 2019.
- K. Jamieson and A. Talwalkar. Non-stochastic best arm identification and hyperparameter optimization. In *Artificial Intelligence and Statistics*, pages 240–248. PMLR, 2016.
- C. Jennison, I. M. Johnstone, and B. W. Turnbull. Asymptotically optimal procedures for sequential adaptive selection of the best of several normal means. In *Statistical decision theory and related topics III*, pages 55–86. Elsevier, 1982.
- R. Johari, V. Kamble, and Y. Kanoria. Matching while learning. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 119–119, 2017.
- P. Joulani, A. Gyorgy, and C. Szepesvári. Online learning under delayed feedback. In *International Conference on Machine Learning*, pages 1453–1461. PMLR, 2013.

- M. Jourdan, M. Mutný, J. Kirschner, and A. Krause. Efficient pure exploration for combinatorial bandits with semi-bandit feedback. In *Algorithmic Learning Theory*, pages 805–849. PMLR, 2021.
- N. Kallus and A. Zhou. Confounding-robust policy improvement. *Advances in neural information processing systems*, 31, 2018.
- K. Kandasamy, A. Krishnamurthy, J. Schneider, and B. Póczos. Parallelised Bayesian optimisation via Thompson sampling. In *International Conference on Artificial Intelligence and Statistics*, pages 133–142. PMLR, 2018.
- M. Kasy and A. Sautmann. Adaptive treatment assignment in experiments for policy choice. *Econometrica*, 89(1):113–132, 2021.
- M. Kato and K. Ariu. The role of contextual information in best arm identification. *arXiv preprint arXiv:2106.14077*, 2021.
- M. Kato, T. Ishihara, J. Honda, and Y. Narita. Efficient adaptive experimental design for average treatment effect estimation. *arXiv preprint arXiv:2002.05308*, 2020.
- E. Kaufmann, N. Korda, and R. Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *International conference on algorithmic learning theory*, pages 199–213. Springer, 2012.
- E. Kaufmann, O. Cappé, and A. Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42, 2016.
- M. Kearns and S. Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2):209–232, 2002.
- J. Kirschner, T. Lattimore, C. Vernade, and C. Szepesvári. Asymptotically optimal information-directed sampling. In *Conference on Learning Theory*, pages 2777–2821. PMLR, 2021.
- T. Kitagawa and A. Tetenov. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018.
- R. Kohavi, D. Tang, and Y. Xu. *Trustworthy online controlled experiments: A practical guide to A/B testing*. Cambridge University Press, 2020.
- J. Komiyama, K. Ariu, M. Kato, and C. Qin. Optimal simple regret in Bayesian best arm identification. *arXiv preprint arXiv:2111.09885*, 2021.
- T. L. Lai. Adaptive treatment allocation and the multi-armed bandit problem. *The Annals of Statistics*, pages 1091–1114, 1987.
- T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.
- T. L. Lai and C. Z. Wei. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, 10(1):154–166, 1982.
- T. Lattimore and A. Gyorgy. Mirror descent and the information ratio. In *Conference on Learning Theory*, pages 2965–2992. PMLR, 2021.
- T. Lattimore and C. Szepesvári. The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *Artificial Intelligence and Statistics*, pages 728–737. PMLR, 2017.
- T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020a.
- T. Lattimore and C. Szepesvári. Exploration by optimisation in partial monitoring. In *Conference on Learning Theory*, pages 2488–2515. PMLR, 2020b.
- L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.
- L. Li, R. Munos, and C. Szepesvári. Toward minimax off-policy value estimation. In *Artificial Intelligence and Statistics*, pages 608–616. PMLR, 2015.
- E. H. Lieb. Convex trace functions and the wigner-yanase-dyson conjecture. *Advances in Mathematics*, 11(3):267–288, 1973.
- L. Massoulié and K. Xu. On the capacity of information processing systems. *Operations Research*, 66(2):568–586, 2018.
- I. Osband, C. Blundell, A. Pritzel, and B. Van Roy. Deep exploration via bootstrapped DQN. *Advances in neural information processing systems*, 29:4026–4034, 2016.
- I. Osband, B. Van Roy, D. J. Russo, and Z. Wen. Deep exploration via randomized value functions. *Journal of Machine Learning Research*, 20:1–62, 2019.
- A. Peleg, N. Pearl, and R. Meir. Metalearning linear bandits by prior update. In *International Conference on Artificial Intelligence and Statistics*, pages 2885–2926. PMLR, 2022.
- C. Qin and D. Russo. Electronic companion to adaptivity and confounding in multi-armed bandit experiments. *Available at SSRN 4115833*, 2022.
- C. Qin, D. Klabjan, and D. Russo. Improving the expected improvement algorithm. *Advances in Neural Information Processing Systems*, 2017:5382–5392, 2017.
- C. Riquelme, G. Tucker, and J. Snoek. Deep Bayesian bandits showdown: An empirical comparison of Bayesian deep networks for Thompson sampling. In *International Conference on Learning Representations*, 2018.
- Y. Russac, C. Katsimerou, D. Bohle, O. Cappé, A. Garivier, and W. M. Koolen. A/B/n testing with control in the presence of subpopulations. In *Advances in Neural Information Processing Systems*, 2021.
- D. Russo. Simple Bayesian algorithms for best arm identification. In *Conference on Learning Theory*, pages 1417–1418. PMLR, 2016.
- D. Russo. Simple Bayesian algorithms for best-arm identification. *Operations Research*, 68(6):1625–1647, 2020.
- D. Russo and B. Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.
- D. Russo and B. Van Roy. An information-theoretic analysis of Thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471, 2016.
- D. Russo and B. Van Roy. Learning to optimize via information-directed sampling. *Operations Research*, 66(1):230–252, 2018.

- D. Russo and B. Van Roy. Satisficing in time-sensitive bandit learning. *Mathematics of Operations Research*, 2022.
- D. Russo and J. Zou. How much does your data exploration overfit? controlling bias via information usage. *IEEE Transactions on Information Theory*, 66(1):302–323, 2019.
- D. J. Russo, B. Van Roy, A. Kazerouni, I. Osband, and Z. Wen. A tutorial on Thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.
- I. O. Ryzhov, W. B. Powell, and P. I. Frazier. The knowledge gradient algorithm for a general class of online learning problems. *Operations Research*, 60(1):180–195, 2012.
- S. L. Scott. A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26(6):639–658, 2010.
- X. Shang, R. Heide, P. Menard, E. Kaufmann, and M. Valko. Fixed-confidence guarantees for Bayesian best-arm identification. In *International Conference on Artificial Intelligence and Statistics*, pages 1823–1832. PMLR, 2020.
- C. Shen. Universal best arm identification. *IEEE Transactions on Signal Processing*, 67(17):4464–4478, 2019.
- M. Simchowitz, C. Tosh, A. Krishnamurthy, D. Hsu, T. Lykouris, M. Dudík, and R. E. Schapire. Bayesian decision-making under misspecified priors with applications to meta-learning. *Advances in Neural Information Processing Systems*, 34, 2021.
- A. Swaminathan and T. Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *The Journal of Machine Learning Research*, 16(1):1731–1755, 2015.
- J. A. Tropp. Freedman’s inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270, 2011.
- J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.
- J. A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.
- B. Van Parys and N. Golrezaei. Optimal learning for structured bandits. *Available at SSRN 3651397*, 2020.
- P.-A. Wang, R.-C. Tzeng, and A. Proutiere. Fast pure exploration via frank-wolfe. *Advances in Neural Information Processing Systems*, 34, 2021.
- Y.-X. Wang, A. Agarwal, and M. Dudík. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning*, pages 3589–3597. PMLR, 2017.
- H. Wu and S. Wager. Thompson sampling with unrestricted delays. *arXiv preprint arXiv:2202.12431*, 2022.
- K. Zhang, L. Janson, and S. Murphy. Inference for batched bandits. *Advances in Neural Information Processing Systems*, 33:9818–9829, 2020.
- Z. Zhou, S. Athey, and S. Wager. Offline multi-action policy learning: Generalization and optimization. *arXiv preprint arXiv:1810.04778*, 2018.
- Z. Zhou, R. Xu, and J. Blanchet. Learning in generalized linear contextual bandits with stochastic delays. *Advances in Neural Information Processing Systems*, 32:5197–5208, 2019.

## A Extensions

We describe a number of possible extensions. In some cases, we know how to complete the extensions that are mentioned, but doing so would lengthen or complicate the paper. For others, substantial further research is required.

### Policy learning.

One direction for future work is to generalize our results about DTS. As one natural generalization, suppose the goal is to identify the best policy from a pre-specified class  $\Pi$ . Each element  $\pi \in \Pi$  is a mapping from  $\mathcal{X}$  to  $[k]$ . Natural justifications for constraining  $\Pi$  were discussed at length in Section 3.2. Overloading notation, take  $\mu(\theta, \pi, w) = \sum_x w(x)\mu(\theta, \pi(x), x)$  to be the average reward if  $\pi$  is employed throughout the population. We can naturally generalize DTS by taking  $\pi^* = \pi^*(\theta) \in \arg \max_{\pi} \mu(\theta, \pi, w)$  to be the optimal policy and  $\alpha_{t,\pi} = \mathbb{P}(\pi^* = \pi \mid H_t)$  to be the posterior measure of  $\pi^*$ . So far we have studied the coarsest possible segmentation, where  $\Pi$  consists of  $k$  policies that prescribe a fixed arm in every context. The typical definition of TS in contextual bandits takes  $\Pi$  to be all possible decision-rules. There is it is critical that sampling from  $\alpha_t$  can be done efficiently, without actually enumerating the space of possible policies. When contexts are i.i.d, we know how to analyze DTS for any policy class by using the information theoretic analysis of Russo and Van Roy [2016, 2022]. We leave the writing of this generalization to future work. More broadly, TS has been studied in a range of complex problems, including reinforcement learning, and it would be interesting to extend the ideas in this paper to that setting.



### Tuning $\beta$ and in-experiment vs post-experiment performance.

We view the information balance equation (20) as capturing a property of any reasonable sampling rule. An algorithm that violates this property will gather an order of magnitude more evidence against the optimality of some arms than others — which is wasteful. Efficient information gathering requires shifting effort away from arms that are overwhelming likely to be suboptimal in favor of those whose status is more uncertain. DTS does this automatically.

The best value of  $\beta_t$  to use depends greatly on the precise objective, however. In general, when  $\beta_t$  is large, DTS behaves more like a typical bandit algorithm, allocating a higher fraction of effort to the arm top-one TS would select. To explore this, let  $\Delta_i = \mu(\theta, I^*, w) - \mu(\theta, i, w)$  denote the optimality gap of arm  $i$  and consider a problem with no contexts, so this quantity also denotes the regret incurred each time arm  $i$  is chosen. As in the proof sketch of Proposition 2, the precise fraction of samples DTS allocates to an arm is determined by the information balance condition (20). When applied with fixed parameter  $\beta$ , this becomes

$$\frac{\Delta_i}{\sqrt{\beta^{-1} + (p_i^*)^{-1}}} = \frac{\Delta_j}{\sqrt{\beta^{-1} + (p_j^*)^{-1}}} \quad \forall i, j \neq I^*.$$

Write the solution of this balance equation as  $(p_1^{(\beta)}, \dots, p_k^{(\beta)})$ . As  $\beta \rightarrow 1$ ,  $(1 - \beta)^{-1} p_i^\beta \rightarrow \Delta_i^{-2} / \sum_{j \neq I^*} \Delta_j^{-2}$ , which, following [Lai and Robbins \[1985\]](#), is the usual recommendation when minimizing regret within the experiment. A careful reading of our proof of Proposition 3, applied with general  $\beta$ , appears to yield the result

$$\frac{\mathbb{E} \left[ \sum_{t=1}^{\tau_c} \Delta_{\hat{I}_t} \mid \theta \right]}{\mathbb{E}[\tau_c \mid \theta]} = \sum_{i=1}^k \frac{\mathbb{E} \left[ \sum_{t=1}^{\tau_c} \mathbb{1}(I_t = i) \mid \theta \right] \Delta_i}{\mathbb{E}[\tau_c \mid \theta]} = \sum_{i \neq I^*} p_i^{(\beta)} \Delta_i + o(1).$$

Taking  $\beta \rightarrow 1$  minimizes the per-period regret measure on the right hand side, but this requires running a longer experiment, which can be inherently costly due to operational considerations. How much is the speed of learning reduced by picking  $\beta$  bounded away from 1? In a related model of best-arm identification, Theorem 1 of [Russo \[2020\]](#) provides results for arbitrary fixed  $\beta$ : in that case,  $\frac{1}{t} \log(1 - \alpha_{t, I^*}) \rightarrow -\Gamma_{\theta, \beta}^{-1}$  where bounds like  $\Gamma_{\theta, 0.5}^{-1} \geq (1/2)\Gamma_{\theta}^{-1}$  control the worst-case impact of setting  $\beta$  that large. It seems the same analysis could be carried out in the setting of this paper. We leave a more formal investigation of the tradeoff between in-experiment and post-experiment regret for future work.

### Modified inference or learned priors.

Our paper focuses on sequential decision-making under uncertainty rather than uncertainty quantification (e.g. the construction of confidence intervals). Proposition 1 relies on the accuracy of Bayesian inference. Proposition 3 is “frequentist”, but asymptotic, and implicitly relies on the fact that the impact of the prior vanishes asymptotically.

One perspective in the literature is that Thompson sampling, and its variants, describe a way of making decisions given uncertainty and that the use of a posterior for uncertainty quantification can be swapped with some alternative. One option is to use bootstrap samples rather than posterior samples [[Eckles and Kaptein, 2019](#), [Russo et al., 2018](#)]. This has been used most substantially in reinforcement learning [[Riquelme et al., 2018](#), [Osband et al., 2016](#)]. Others update the parameters of a parametric posterior distribution using some alternative logic. [Agrawal and Goyal \[2013b\]](#) and [Abeille and Lazaric \[2017\]](#) inflate posterior variances in order to establish worst-case guarantees for a TS-style algorithm. [Dimakopoulou et al. \[2021\]](#) recently proposed a parameter updating rule motivated by doubly robust estimation techniques. It is an open question whether these techniques work effectively in the setting of this paper.

An alternative perspective in the literature is that the use of a prior is essential for incorporating knowledge that is available before the experiment begins. This view is advocated in [Scott \[2010\]](#) and [Russo et al. \[2018\]](#), for example, as well as in the discussion of Examples 1 and 2 above. Taking this viewpoint, many recent papers have formally studied the learning of prior parameters from other available data or across

many experiments; see [Dimmery et al. \[2019\]](#), [Hsu et al. \[2019\]](#), [Azevedo et al. \[2019\]](#), [Hong et al. \[2022\]](#), [Basu et al. \[2021\]](#), [Simchowitx et al. \[2021\]](#), [Peleg et al. \[2022\]](#). A recent preprint by [Finan and Pouzo \[2021\]](#) studies combining many prior sources of information in adaptive experiments.

### Confidence intervals post-experiment

Results of [Lai and Wei \[1982\]](#) can sometimes be used to establish consistency and asymptotic normality for least-squares regression when the covariates are adaptively chosen. Unfortunately, their conditions are often violated when data is collected by bandit algorithms, as highlighted by [Deshpande et al. \[2018\]](#). An emerging stream of literature develops nonstandard inferential procedures to quantify uncertainty after an adaptive experiment [[Hadad et al., 2021](#), [Chen et al., 2020](#), [Zhang et al., 2020](#)].

Thankfully, this is not necessary in our setting. DTS has special structures which mean that the conditions of [Lai and Wei \[1982\]](#) apply to without modification. Formally, under Assumption 1, for any unit vector  $u$  and scalar  $z$ ,

$$\mathbb{P} \left\{ \left\langle u, \Sigma_{t,i}^{-1/2} (\theta^{(i)} - \mu_{t,i}) \right\rangle \leq z \mid \theta \right\} \rightarrow \Phi(z) \quad \text{as } t \rightarrow \infty.$$

Because we use  $\Sigma_{t,i}^{-1/2}$  and  $\mu_{t,i}$ , this result technically refers to ridge regression. The same result applies to ordinary least squares, but this requires new notation. The key fact is that  $\frac{1}{t} \sum_{\ell=1}^t \mathbb{1}(I_\ell = i) X_\ell X_\ell^\top \rightarrow p_i^*(\theta) \mathbb{E}[X_1 X_1^\top] \succ 0$ . This stability property allows one<sup>11</sup> to apply theorem 3 of [Lai and Wei \[1982\]](#). Something like this key property was already noted in (17). What drives this stability? It seems DTS has two key special properties. One is the context independence property in (14). Another is the convergence to non-vanishing asymptotic sampling proportions, with  $\frac{1}{t} \sum_{\ell=1}^t \mathbb{1}(I_\ell = i) \rightarrow p_i^*(\theta) > 0$ . This relies on the use of resampling ( $\beta_t < 1$ ) and would not apply to standard Thompson sampling ( $\beta_t = 1$ ).

### Treating correlated beliefs across arms.

We have presented the case where the parameter vectors  $\theta^{(i)}$  and  $\theta^{(j)}$  are uncorrelated for different arms  $i \neq j$ . This greatly simplified some of the presentation. But, as reflected in Examples 1 and 2, the possibility of such correlation allows for much richer modeling. Here we aim to (A) explain how to update beliefs and implement the algorithm in a case with correlation and (B) explain how, with minimal modifications, most theoretical bounds still hold in this case.

We can write our model in the general form

$$R_t = \langle \theta, \phi(X_t, I_t) \rangle + W_t \quad \text{where} \quad \theta = \begin{bmatrix} \theta^{(1)} \\ \vdots \\ \theta^{(k)} \end{bmatrix} \quad \text{and} \quad \phi(x, i) = [0 \cdots 0 \ x_1 \cdots x_d \ 0 \cdots 0]^\top.$$

The posterior mean and covariance can be written in the form

$$\mu_t = \begin{bmatrix} \mu_{t,1} \\ \vdots \\ \mu_{t,k} \end{bmatrix} \in \mathbb{R}^{dk} \quad \text{and} \quad \Sigma_t = \begin{bmatrix} \Sigma_{t,11} & \cdots & \Sigma_{t,1k} \\ \vdots & \ddots & \vdots \\ \Sigma_{t,k1} & \cdots & \Sigma_{t,kk} \end{bmatrix} \in \mathbb{R}^{dk \times dk}$$

where  $\mu_{t,i} = \mathbb{E}[\theta^{(i)} \mid H_t]$  and  $\Sigma_{t,ij} = \mathbb{E}[(\theta^{(i)} - \mu_{t,i})(\theta^{(j)} - \mu_{t,j})^\top]$ . These can be updated according to the equations  $\Sigma_t = \left( \Sigma_1^{-1} + \sigma^{-2} \sum_{\ell=1}^{t-1} \phi(X_\ell, I_\ell) \phi(X_\ell, I_\ell)^\top \right)^{-1}$  and  $\mu_t = \Sigma_t \left( \Sigma_1^{-1} \mu_1 + \sigma^{-2} \sum_{\ell=1}^{t-1} \phi(X_\ell, I_\ell) R_\ell \right)$ . The population average reward  $\mu(\theta, i, w)$  also has a Gaussian posterior,  $[\mu(\theta, 1, w) \cdots \mu(\theta, k, w)]^\top \mid H_t \sim$

<sup>11</sup>To translate between their notation and ours, note that what they call  $X_t$  should be viewed as the censored vector  $\mathbb{1}(I_t = i) X_t$  in our problem. The adaptivity in this vector comes from the choice of  $I_t$ . Their equation 4.2 follows by taking  $B_t = \sqrt{t \cdot p_i^*(\theta) \mathbb{E}[X_1 X_1^\top]}^{1/2}$ . Their equation 4.3 follows easily since contexts are assumed to be bounded.

$N(m_t, S_t)$ , with mean and covariance

$$m_t = \begin{bmatrix} \langle X_{\text{pop}}, \mu_{t,1} \rangle \\ \vdots \\ \langle X_{\text{pop}}, \mu_{t,k} \rangle \end{bmatrix} \in \mathbb{R}^k \quad \text{and} \quad S_t = \begin{bmatrix} X_{\text{pop}}^\top \Sigma_{t,11} X_{\text{pop}} & \cdots & X_{\text{pop}}^\top \Sigma_{t,1k} X_{\text{pop}} \\ \vdots & \ddots & \vdots \\ X_{\text{pop}}^\top \Sigma_{t,k1} X_{\text{pop}} & \cdots & X_{\text{pop}}^\top \Sigma_{t,kk} X_{\text{pop}} \end{bmatrix} \in \mathbb{R}^{k \times k}.$$

The implementation of DTS extends gracefully. The definition of  $I^* = \arg \max_{i \in [k]} \langle X_{\text{pop}}, \theta^{(i)} \rangle$  and  $\alpha_{t,i} = \mathbb{P}(I^* = i \mid H_t)$  are unchanged. When the tuning parameter is  $\beta_t = 1$ , DTS picks an arm to measure by sampling from  $\alpha_{t,i}$ . That is,  $\mathbb{P}(I_t = i \mid H_t) = \alpha_{t,i}$ . A sample from  $\alpha_t$  can be generated by sampling  $v \sim N(m_t, S_t) \in \mathbb{R}^k$  and picking  $I_t = \arg \max_i v_i$ .

It appears that all of the guarantees for DTS extend easily to the case of a correlated prior across arms. The key insight is that the posterior covariance matrix in the correlated case is both lower and upper bounded by the covariance which would arise in the uncorrelated case. Take  $\underline{\Sigma}_1 = \lambda_{\min}(\Sigma_1)I \in \mathbb{R}^{dk \times dk}$  and  $\bar{\Sigma}_1 = \lambda_{\max}(\Sigma_1)I \in \mathbb{R}^{dk \times dk}$ . These dominate the prior covariance in the sense that  $\underline{\Sigma}_1 \preceq \Sigma_1 \preceq \bar{\Sigma}_1$ . This then implies

$$\underline{\Sigma}_t \triangleq \left( \underline{\Sigma}_1^{-1} + \sum_{\ell=1}^{t-L} \phi(X_\ell, I_\ell) \phi(X_\ell, I_\ell)^\top \right)^{-1} \preceq \Sigma_t \preceq \left( \bar{\Sigma}_1^{-1} + \sum_{\ell=1}^{t-L} \phi(X_\ell, I_\ell) \phi(X_\ell, I_\ell)^\top \right)^{-1} \triangleq \bar{\Sigma}_t. \quad (27)$$

These expressions upper and lower bound the true covariance by fictitious ones are updated as if  $\theta^{(i)}$  and  $\theta^{(j)}$  are uncorrelated. These fictitious matrices follow the same formulas used throughout this paper; for instance,  $\bar{\Sigma}_{t,ii} = \left( I / \lambda_{\max}(\Sigma_1) + \sum_{\ell=1}^{t-L} \mathbb{1}(I_\ell = i) X_\ell X_\ell^\top \right)^{-1}$ . Since the proof of Proposition 1 boils down to upper bounding the posterior covariance, the bounds above let us extend Proposition 1's conclusions to the correlated case with almost no modifications<sup>12</sup>. The proof sketches of Lemma 4 and Proposition 2 go through identically. The key is that the impact of correlation vanishes asymptotically. Formally,  $\frac{1}{t} \Sigma_t$  converges asymptotically to a block diagonal matrix with  $\frac{1}{t} \Sigma_{t,ii} \rightarrow p_i \sigma^2 \Lambda^{-1}$ , recovering (17). The lower bound portion of Proposition 3 does not depend on the prior. We conjecture that the upper bound analysis underlying Proposition 3 can be easily extended to account for correlation, using bounds like (27), but we have not verified each line.

## B Proofs from Section 5: analysis of counterexamples

### B.1 Proof of Lemma 1

We first recall the statement of the result.

**Lemma 1** (Failure of context-unaware TS). *Consider Example 3. Suppose the components of the vector  $\theta = (\theta_x^{(i)})_{i \in [2], x \in [2]}$  are independent with  $\theta_x^{(1)} \sim N(0, 1)$  and  $\theta_x^{(2)} \sim N(0, 2)$  for  $x \in [2]$ . Suppose the noise variance  $\sigma^2 = 0$  and observations are not subject to delay (i.e.  $L = 1$ ). If (7) holds, there exists an absolute numerical constant  $c > 0$  such that for all  $T \in \mathbb{N}$ ,  $\mathbb{E}[\Delta_T] \geq c$ .*

*Proof.* Proof of Lemma 1. It is not hard to show that  $\mathbb{E}[\Delta_t] > 0$  for any fixed  $t$ . To show the result, then, it is without loss of generality to assume  $T \geq 4$ . Let  $\Theta'$  denote the set of parameter vectors satisfying the following properties:

1.  $\frac{\theta_1^{(2)} + \theta_2^{(2)}}{2} > \frac{\theta_1^{(1)} + \theta_2^{(1)}}{2}$ : This implies that the optimal arm is  $I^* = 2$

<sup>12</sup>To produce a valid proof, it seems we need to make only a single change. Instead of defining  $S_{t,i} = \text{Cov}(\theta^{(i)} | H_t^+) = \left( \bar{\Sigma}_{1,i}^{-1} + \sigma^{-2} \sum_{\ell=1}^t \mathbb{1}\{I_\ell = i\} X_\ell X_\ell^\top \right)^{-1}$  and assuming independence, define  $S_{t,i} = \left( I / \lambda + \sigma^{-2} \sum_{\ell=1}^t \mathbb{1}\{I_\ell = i\} X_\ell X_\ell^\top \right)^{-1} \preceq \text{Cov}(\theta^{(i)} | H_t^+)$ , where the inequality holds as long as  $\lambda$  exceeds the maximum eigenvalue of prior covariance matrix. Lemma 8 still holds. Anywhere  $\Sigma_1$  appears in the proof or result, replace it with  $\lambda I$ .

2.  $\min \left\{ \theta_1^{(1)}, \theta_2^{(1)} \right\} > \theta_1^{(2)}$ : This implies arm 1 appears to be the best if arm 2 is only measured in the context  $e_1$ .
3.  $\theta_1^{(1)} < 0$ : This implies that if arm 1 is sampled in the first period, arm 2 has at least a  $\frac{1}{2}$  chance of being sampled in the second period.

In the first period, let  $\nu_{1,1} \sim N(\tilde{m}_{1,1}, \tilde{s}_{1,1}^2)$  and  $\nu_{1,2} \sim N(\tilde{m}_{1,2}, \tilde{s}_{1,2}^2)$  denote the sampled parameters, and we denote the probability of playing arm 1 by

$$c_0 \triangleq \mathbb{P}(\nu_{1,1} > \nu_{1,2}) > 0.$$

Conditioned on the event that  $\theta \in \Theta'$  and  $I_1 = 1$ , we have  $(\tilde{m}_{2,1}, \tilde{s}_{2,1}^2) = (\theta_1^{(1)}, 0)$  and the probability of playing arm 2 in the second period is

$$\mathbb{P}(\nu_{2,2} > \theta_1^{(1)} \mid \theta \in \Theta', I_1 = 1) \geq \frac{1}{2}$$

where the inequality holds due to Condition 3 above. Conditioned on the event that  $\theta \in \Theta'$ ,  $I_1 = 1$  and  $I_2 = 2$ , we have  $(\tilde{m}_{3,1}, \tilde{s}_{3,1}^2) = (\theta_1^{(1)}, 0)$  and  $(\tilde{m}_{3,2}, \tilde{s}_{3,2}^2) = (\theta_1^{(2)}, 0)$ . Due to Condition 2, TS will always measure arm 1 afterwards. Hence,

$$\begin{aligned} \mathbb{E}[\Delta_T] &\geq \mathbb{E} \left[ \left( \frac{\theta_1^{(2)} + \theta_2^{(2)}}{2} - \frac{\theta_1^{(1)} + \theta_2^{(1)}}{2} \right) \mathbb{1}(\theta \in \Theta') \mathbb{1}(I_1 = 1, I_2 = 2) \right] \\ &\geq \frac{c_0}{2} \mathbb{E} \left[ \left( \frac{\theta_1^{(2)} + \theta_2^{(2)}}{2} - \frac{\theta_1^{(1)} + \theta_2^{(1)}}{2} \right) \mathbb{1}(\theta \in \Theta') \right] > 0. \end{aligned}$$

This completes the proof.  $\square$

## B.2 Proof of Lemma 2

We first recall the statement of the result.

**Lemma 2** (Failure of deconfounded UCB). *Consider Example 3. Suppose that the components of the vector  $\theta = (\theta_x^{(i)})_{i \in [2], x \in [2]}$  are independent with  $\theta_x^{(1)} \sim N(0, 1)$  and  $\theta_x^{(2)} \sim N(0, 3)$  for  $x \in [2]$ . Suppose the noise variance  $\sigma^2 = 0$  and observations are not subject to delay (i.e.  $L = 1$ ). If (8) holds, there is an absolute numerical constant  $c > 0$  such that for all  $T \in \mathbb{N}$ ,  $\mathbb{E}[\Delta_T] \geq c$ .*

*Proof.* Proof of Lemma 2. Let  $U_{t,i} = m_{t,i} + z \cdot s_{t,i}$  denote the UCB of arm  $i$ . Since  $U_{1,2} > U_{1,1}$ , the initial arm selection is  $I_1 = 2$ . When  $\theta_1^{(2)} > 0$ , the posterior mean and standard deviation satisfy  $m_{2,2} = \frac{\theta_1^{(2)}}{2} > 0 = m_{2,1}$  and  $s_{2,2} = \frac{1}{2} \sqrt{\text{Var}(\theta_1^{(2)} \mid H_2) + \text{Var}(\theta_2^{(2)} \mid H_2)} = \sqrt{\frac{3}{2}} > \sqrt{\frac{1}{2}} = s_{2,1}$ . This implies  $U_{2,2} > U_{2,1}$  and arm  $I_2 = 2$  is again selected. This process repeats, showing that if  $\theta_1^{(2)} > 0$ , then arm 2 is chosen for each of the first  $\lfloor \frac{T}{2} \rfloor$  periods. We can lower bound simple regret by imagining that the decision-maker has perfect knowledge of  $\theta_1^{(2)}, \theta_2^{(2)}$  and  $\theta_2^{(1)}$  when selecting  $\hat{f}_T \in \arg \max_{i \in [2]} \mathbb{E} \left[ \frac{\theta_1^{(i)} + \theta_2^{(i)}}{2} \mid H_T^+ \right]$ , resulting in:

$$\mathbb{E}[\Delta_T] \geq \mathbb{E} \left[ \left( \max \left\{ \frac{\theta_1^{(1)} + \theta_2^{(1)}}{2}, \frac{\theta_1^{(2)} + \theta_2^{(2)}}{2} \right\} - \max \left\{ \frac{\theta_2^{(1)}}{2}, \frac{\theta_1^{(2)} + \theta_2^{(2)}}{2} \right\} \right) \mathbb{1}(\theta_1^{(2)} > 0) \right] > 0.$$

The strict inequality is due to the gap in Jensen's inequality, reflecting the value of having perfect information about  $\theta_1^{(1)}$  when making a decision.  $\square$

### B.3 Failure of contextual bandit algorithms

The goal in our formulation is to select among a very restricted set of decision-rules: those that choose a common action, irrespective of context. Experimentation should be tailored to this objective. Here, we give insight into potential failures when an exploration algorithm is designed with a different learning target in mind. Consider the following example. There are three actions, and the decision-maker would like to identify the best action to employ on average, across all contexts. Imagine that the context set describes two customer segments. Action 1 appeals to one segment, but is highly unappealing to the other. For action 2, the situation is reversed. Action 3 is not ideal for either segment, but is also not disliked by either. When personalization is inappropriate or costly, action 3 may be the preferred communal option.

The next example does not align with our formulation, because we take the prior distribution to be non-Gaussian. Similar issues can arise with a Gaussian prior, but its unbounded nature always allows for a nonzero – even if very small – chance that the mainstream action is better even for a specific segment. We omit analytical calculations of this more intricate case, since Example 4 seems already to capture the main intuition.

**Example 4** (A mainstream action). *Consider a problem with  $k = 3$  arms and context set  $\mathcal{X} = \{e_1, e_2\} \subset \mathbb{R}^2$ . The population distribution  $w$  is uniform over  $\mathcal{X}$  and  $(X_t)_{t \in \mathbb{N}}$  are drawn i.i.d from  $w$ . The components of the parameter vector  $\theta = (\theta_1, \theta_2, \theta_3)$  are drawn independently with  $\theta_x \sim \text{Uniform}(\{1, 2\})$  for  $x \in [2]$  and  $\theta_3 \sim \text{Uniform}([0, 1])$ . Rewards are noiseless, with  $R_t = \mu(\theta, I_t, X_t)$ . Observations are not subject to delay (i.e.  $L = 1$ ). Action 3's performance is insensitive to the context, and it always generates mean-reward  $\mu(\theta, 3, e_x) = \theta_3$  for  $x \in [2]$ . Actions 1 and 2 generate mean rewards in context  $e_x \in \mathcal{X}$  given by*

$$\mu(\theta, 1, e_x) = \frac{1}{2} + \frac{1}{2} \mathbb{1}(\theta_x = 1) \quad \text{and} \quad \mu(\theta, 2, e_x) = \frac{1}{2} + \frac{1}{2} \mathbb{1}(\theta_x = 2).$$

The next lemma formalizes that contextual Thompson sampling, which selects an action according to the posterior probability it is the optimal action for the current context, has simple regret that does not vanish even as the horizon grows. The same result applies to appropriate contextual versions of UCB. The simple reason is that action 3 is never sampled, because it does not maximize the reward in either context. This means no information about  $\theta_0$  is gathered and the decision-maker cannot determine whether action 3 is the best arm to select. If the goal is to identify the best policy within a restricted class, the exploration algorithm needs to be designed so that it gathers the right information for this task. The proof follows from this argument and is omitted for brevity. At a high level, contextual TS fails here because it does not reflect the true decision-objective.

**Lemma 6** (Failure of contextual TS). *Consider Example 4. Contextual TS at time  $t$  chooses an arm  $I_t$  such that for each  $i \in [k]$ ,  $\mathbb{P}(I_t = i \mid H_t) = \mathbb{P}(\arg \max_{j \in [k]} \mu(\theta, j, X_t) = i \mid H_t)$ . There is an absolute numerical constant  $c > 0$  such that for all  $T \in \mathbb{N}$ ,  $\mathbb{E}[\Delta_T] \geq c$ .*

## C Proof of Proposition 1

### Matrix convex combinations.

Let  $\mathcal{S}^n$  denote the set of symmetric  $n \times n$  square matrices, and let  $\mathcal{S}_+^n \subset \mathcal{S}^n$  denote the set of symmetric positive semidefinite matrices. A scalar function  $f : \mathbb{R} \rightarrow \mathbb{R}$  can be extended to a function on symmetric matrices as follows. For any symmetric matrix  $A \in \mathcal{S}_n$ , one can write  $A = \sum_{i=1}^n \lambda_i u_i u_i^\top$  where each  $\lambda_i$  is a real eigenvalue and  $u_i$  is the corresponding eigenvector. By defining  $f(A) = \sum_{i=1}^n f(\lambda_i) u_i u_i^\top$ , we have extended  $f$  to a function mapping from  $\mathcal{S}_n$  to  $\mathcal{S}_n$ . A function  $f$  is said to be monotone increasing on the space of positive semidefinite matrices if for  $A, B \in \mathcal{S}_+^n$ ,  $A \preceq B$  implies  $f(A) \preceq f(B)$  and monotone decreasing if this implies  $f(A) \succeq f(B)$ . A function  $f$  is said to be operator convex on the space of positive definite matrices if for any  $A, B \in \mathcal{S}_+^n$  and scalar  $\lambda \in [0, 1]$ ,  $f(\gamma A + (1 - \gamma)B) \preceq \gamma f(A) + (1 - \gamma)f(B)$ . For our purposes, a key fact is that the inverse function  $f(A) = A^{-1}$  is convex and monotone decreasing.

To prove Proposition 1, we need to leverage a generalization of Jensen’s inequality that applies to matrix convex combinations. The following definitions and result can be found in Tropp [2015]. It provides a deep generalization of Jensen’s inequality for operator convex functions, extending to a situation where the weights are matrices rather than scalars.

**Definition 2.** Let  $B_1, B_2$  be Hermitian matrices. If  $A_1^\top A_1 + A_2^\top A_2 = I$ , then the Hermitian matrix  $A_1^\top B_1 A_1 + A_2^\top B_2 A_2$  is called a matrix convex combination of  $B_1$  and  $B_2$ .

The next result in Theorem 8.5.2 in Tropp [2015] and a self-contained proof is given there.

**Lemma 7 (Operator Jensen Inequality).** Let  $f$  be an operator convex on the space of symmetric positive semidefinite matrices  $\mathbf{S}_+^n$ . Let  $B_1, B_2 \in \mathbf{S}_+^n$ . If  $A_1^\top A_1 + A_2^\top A_2 = I$  then,

$$f\left(A_1^\top B_1 A_1 + A_2^\top B_2 A_2\right) \preceq A_1^\top f(B_1) A_1 + A_2^\top f(B_2) A_2.$$

## C.1 Relating simple regret to posterior variance

Define the full information covariance matrix

$$S_{t,i} = \text{Cov}\left(\theta^{(i)} | H_t^+\right) = \left(\Sigma_1^{-1} + \sigma^{-2} \sum_{\ell=1}^t \mathbb{1}\{I_\ell = i\} X_\ell X_\ell^\top\right)^{-1}.$$

This is posterior variance of  $\theta^{(i)}$  after observing the rewards generated by all  $t$  actions. See Section 4 for a discussion of the delayed history  $H_t$  and the full information history  $H_t^+$ . One can also see that  $S_{t,i} = \Sigma_{t+L,i}$ .

The first key to our result reduces the problem of controlling the simple regret to the problem of controlling the expected posterior variance of the optimal arm.

**Lemma 8.** Under the conditions of Proposition 1,

$$\mathbb{E}[\Delta_T] \leq \sqrt{2\mathbb{H}(I^* | H_T^+) X_{\text{pop}}^\top \mathbb{E}[S_{T,I^*}] X_{\text{pop}}}$$

*Proof.* Proof of Lemma 8. Define  $Z_i = \mu(\theta, i, w)$  to be the uncertain population performance of arm  $i$ ,  $M_{T,i} = \mathbb{E}[Z_i | H_T^+] = \langle X_{\text{pop}}, \mathbb{E}[\theta^{(i)} | H_T^+] \rangle$  to be its posterior mean, and  $\sigma_{T,i}^2 = \text{Var}(\mu(\theta, i, w) | H_T^+) = X_{\text{pop}}^\top S_{T,i} X_{\text{pop}}$  to be its posterior variance. This notation is used only in this proof. Note that  $Z_i | H_T^+ \sim N(M_{T,i}, \sigma_{T,i}^2)$ . Take  $Z = (Z_1, \dots, Z_k)$  to be the vector. Let  $\tilde{I}$  be a sample from the posterior distribution of  $I^*$ , i.e.  $\mathbb{P}(\tilde{I} = i | H_T^+) = \mathbb{P}(I^* = i | H_T^+)$  but  $\tilde{I}$  is independent of  $(Z_1, \dots, Z_k)$  conditioned on  $H_T^+$ . We let  $\mathbb{I}_{H_T^+}(Y_1; Y_2)$  denote the mutual information between random variables  $Y_1$  and  $Y_2$  under the distribution  $\mathbb{P}((Y_1, Y_2) \in \cdot | H_T^+)$ . This is random, due to its dependence on the history. Taking expectations yields the usual definition of mutual information, with  $\mathbb{E}[\mathbb{I}_{H_T^+}(Y_1; Y_2)] = \mathbb{I}(Y_1; Y_2 | H_T^+)$ . This notation is used in this proof alone.



Recall that  $\hat{I}_T \in \arg \max_{i \in [k]} \mathbb{E} [\mu(\theta, i, w) \mid H_T^+]$  is the Bayes selection. Then

$$\begin{aligned}
\mathbb{E} [\Delta_T] &= \mathbb{E} [Z_{I^*} - Z_{\hat{I}_T}] = \mathbb{E} [Z_{I^*} - \mathbb{E} [Z_{\hat{I}_T} \mid H_T^+]] \leq \mathbb{E} [Z_{I^*} - \mathbb{E} [Z_{\tilde{I}} \mid H_T^+]] \\
&= \mathbb{E} [Z_{I^*} - \mathbb{E} [M_{T, \tilde{I}} \mid H_T^+]] \\
&\stackrel{(a)}{=} \mathbb{E} [Z_{I^*} - \mathbb{E} [M_{T, I^*} \mid H_T^+]] \\
&= \mathbb{E} [\mathbb{E} [Z_{I^*} - M_{T, I^*} \mid H_T^+]] \\
&\stackrel{(b)}{\leq} \mathbb{E} \left[ \sqrt{\mathbb{E} [\sigma_{T, I^*}^2 \mid H_T^+]} \sqrt{2\mathbb{I}_{H_T^+} (I^*; Z)} \right] \\
&\stackrel{(c)}{\leq} \sqrt{\mathbb{E} [\mathbb{E} [\sigma_{T, I^*}^2 \mid H_T^+]]} \sqrt{2\mathbb{E} [\mathbb{I}_{H_T^+} (I^*; Z)]} \\
&\stackrel{(d)}{=} \sqrt{\mathbb{E} [\sigma_{T, I^*}^2]} \sqrt{2\mathbb{I} (I^*; Z \mid H_T^+)} \\
&\leq \sqrt{\mathbb{E} [\sigma_{T, I^*}^2]} \sqrt{2\mathbb{H} (I^* \mid H_T^+)} \\
&= \sqrt{\mathbb{E} [X_{\text{pop}}^\top S_{T, I^*} X_{\text{pop}}]} \sqrt{2\mathbb{H} (I^* \mid H_T^+)}
\end{aligned}$$

Early steps of the proof use the tower property of conditional expectation. Step (a) is crucial and uses that fact that  $\tilde{I}$  and  $I^*$  have the same distribution conditions on  $H_T^+$  and the vector  $(M_{T, 1}, \dots, M_{T, k})$  is nonrandom conditioned on  $H_T^+$  (formally is measurable with respect to the sigma-algebra  $H_T^+$  generates). Step (b) applies Proposition 8 of [Russo and Zou \[2019\]](#), which is stated below. Step (c) applies the Hölder inequality, step (d) applies the tower property of conditional expectation, and the final step uses that entropy bounds mutual information.  $\square$

**Lemma 9** (Proposition 8 of [Russo and Zou \[2019\]](#)). *Consider a random vector  $Z \in \mathbb{R}^n$  and a random index  $I \in [n]$ . Suppose that for each  $i \in [n]$ ,  $Z_i$  has mean  $\mu_i$  and the distribution of  $Z_i - \mu_i$  is sub-Gaussian with variance proxy  $\sigma_i^2$ . Then*

$$|\mathbb{E} [Z_I - \mu_I]| \leq \sqrt{\mathbb{E} [\sigma_I^2]} \sqrt{2\mathbb{I} (Z; I)}.$$

In the setting of the above lemma, a standard sub-Gaussian maximal inequality would bound the largest deviation of  $Z_i$  from its mean as  $\mathbb{E} [\max_{i \in [k]} |Z_i - \mu_i|] \leq (\max_{i \in [k]} \sigma_i) \sqrt{2 \log(n)}$ . For our purposes, the lemma offers a critical improvement because it depends only on the variance at the likely realizations of  $I$ . A second improvement, which is the focus of the discussion in [Russo and Zou \[2019\]](#), is that the mutual information term  $\mathbb{I} (Z; I)$  could be much smaller than  $\log(n)$ .

## C.2 De-randomization

Even if we condition on  $X_{1:T}$  as we do in Proposition 1, the posterior covariance is random because the actions  $I_1, \dots, I_T$  are random. We consider the de-randomized analogue of the posterior covariance matrix,

$$\tilde{S}_{t,i} = \left( \Sigma_1^{-1} + \sigma^{-2} \sum_{\ell=1}^t \mathbb{P}(I_\ell = i \mid H_\ell) X_\ell X_\ell^\top \right)^{-1}, \quad (28)$$

which effectively replaces  $\mathbb{1}(I_\ell = i)$  with its conditional expectation,  $\alpha_{\ell,i} = \mathbb{P}(I_\ell = i \mid H_\ell)$ . This matrix measures the extent to which the algorithm's allocation decisions  $\alpha_1, \dots, \alpha_t$  explore in each direction. This might be thought of as the posterior variance under Thompson sampling in a smoothed problem that allows for fractional arm measurement decisions. The next lemma controls the impact of randomization, ensuring the posterior variance cannot be more than  $\iota$  times larger in *any* direction than the idealized posterior

covariance in (28). Using this, we focus on the idealized posterior covariance for the remainder of the proof of Proposition 1.

**Lemma 10** (Controlling the impact of action randomization). *For any  $T \in \mathbb{N}$ ,*

$$\mathbb{E} [S_{T,I^*}] \preceq \iota \cdot \mathbb{E} [\tilde{S}_{T,I^*}]$$

where  $\iota$  is defined as in Proposition 1.

The proof of Lemma 10, uses concentration techniques for matrix valued martingales that were popularized by Tropp [2011, 2012]. By specializing a result like the matrix Friedman inequality of Tropp [2011], we would be able to establish an inequality like

$$\lambda_{\max} \left( S_{T,i}^{-1} - \tilde{S}_{T,i}^{-1} \right) \leq c \lambda_{\max} \left( \tilde{S}_{T,i}^{-1} \right) + \log \left( \frac{d}{\delta} \right).$$

We modify the proofs of Tropp [2011] to prove an inequality of the form

$$S_{T,i}^{-1} - \tilde{S}_{T,i}^{-1} \preceq c \tilde{S}_{T,i}^{-1} + \log \left( \frac{d}{\delta} \right) I \quad (29)$$

also holds with high probability. For our purposes (29) offers a critical improvement. The issue is that the decision-maker might gather very little information about some context directions and a lot of information about others. In that case,  $\tilde{S}_{T,i}^{-1}$  would have some eigenvalues that are very small and others that are very large. To establish a result like Lemma 10, we need a result that bounds the scale of deviation along each particular direction in a more refined way. From (29), a direct (though messy) argument yields Lemma 10. This foray into matrix concentration is somewhat tangential, so we defer this proof until Subsection C.5.

### C.3 Using implicit propensity scores to complete the proof

Now, define

$$S_{\text{full}} = \left( \Sigma_1^{-1} + \sigma^{-2} \sum_{t=1}^T X_t X_t^\top \right)^{-1},$$

which is the posterior covariance of  $\theta^{(i)}$  if the reward of arm  $i$  were observed in every time period. The next lemma is the most essential one in our analysis. It upper bounds the idealized posterior covariance matrix by an expression that relates to the sampling variance of the inverse-propensity weighted least squares estimator. In particular, consider

$$\hat{\theta}_T^{(i)} = \arg \min_{\hat{\theta} \in \mathbb{R}^d} \sum_{t=1}^T \left( X_t^\top \hat{\theta} - \hat{R}_{t,i} \right)^2 + \sigma^2 \|\hat{\theta}\|_{\Sigma_1^{-1}}^2 \quad \text{where} \quad \hat{R}_{t,i} = \frac{R_t \mathbb{1}(I_t = i)}{\mathbb{P}(I_t = i | H_t)},$$

which is ridge regression applied to rewards that are de-biased by inverse propensity weighting. If  $\mathbb{P}(I_t = i | H_t)$  and  $X_t$  were fixed and non-random, then the sampling variance of  $\hat{\theta}_T^{(i)}$  could be shown to equal

$$\text{Cov} \left( \hat{\theta}_T^{(i)} \mid \theta, X_{1:T} \right) = S_{\text{full}} \mathbb{E} \left[ \Sigma_1^{-1} + \sigma^{-2} \sum_{t=1}^T \frac{X_t X_t^\top (1 + c_t)}{\mathbb{P}(I_t = i | H_t)} \mid \theta, X_{1:T} \right] S_{\text{full}}$$

where  $c_t = \sigma^{-2} (\mu(\theta, i, X_t))^2 (1 - \mathbb{P}(I_t = i | H_t)) \geq 0$  and this notation is used only temporarily to simplify this formula. This formula mirrors the behavior of the bound in Lemma 3. Lemma 3 can therefore be viewed as giving a bound on the posterior covariance by implicit propensity weighting.

**Lemma 3** (Propensity matching type variance bound). *For any  $i \in [k]$ , with probability one,*

$$\tilde{S}_{T,i} \preceq S_{\text{full}} \left( \Sigma_1^{-1} + \sigma^{-2} \sum_{t=1}^T \frac{X_t X_t^\top}{\mathbb{P}(I_t = i | H_t)} \right) S_{\text{full}}.$$

*Proof.* Proof of Lemma 3. We use specialized notation in the proof. Since  $i$  is fixed, drop it from notation and write  $p_t = \mathbb{P}(I_t = i | H_t)$ . Set  $B_t = \sigma^{-2} X_t X_t^\top$ . For notational convenience, set  $B_0 = \Sigma_1^{-1}$  and  $p_0 = 1$ . We then can write  $\tilde{S}_{T,i}^{-1} = \sum_{t=0}^T p_t B_t$ . Define the full precision matrix  $\Lambda \triangleq S_{\text{full}}^{-1} = \sum_{t=0}^T B_t$ . We have,

$$\begin{aligned} \tilde{S}_{T,i} &= \left( \sum_{t=0}^T p_t B_t \right)^{-1} = \Lambda^{-1/2} \left[ \Lambda^{-1/2} \left( \sum_{t=0}^T p_t B_t \right) \Lambda^{-1/2} \right]^{-1} \Lambda^{-1/2} \\ &= \Lambda^{-1/2} \left[ \Lambda^{-1/2} \left( \sum_{t=0}^T B_t^{1/2} (p_t I) B_t^{1/2} \right) \Lambda^{-1/2} \right]^{-1} \Lambda^{-1/2} \\ &= \Lambda^{-1/2} \left[ \sum_{t=0}^T \left( \Lambda^{-1/2} B_t^{1/2} \right) (p_t I) \left( B_t^{1/2} \Lambda^{-1/2} \right) \right]^{-1} \Lambda^{-1/2} \\ &\preceq \Lambda^{-1/2} \left[ \sum_{t=0}^T \left( \Lambda^{-1/2} B_t^{1/2} \right) (p_t I)^{-1} \left( B_t^{1/2} \Lambda^{-1/2} \right) \right] \Lambda^{-1/2} \\ &= \Lambda^{-1} \left( \sum_{t=0}^T \frac{B_t}{p_t} \right) \Lambda^{-1}. \end{aligned}$$

The inequality applies the operator Jensen inequality in Lemma 7, using that

$$\sum_{t=0}^T \left( \Lambda^{-1/2} B_t^{1/2} \right) \left( \Lambda^{-1/2} B_t^{1/2} \right)^\top = \Lambda^{-1/2} \left( \sum_{t=0}^T B_t \right) \Lambda^{-1/2} = \Lambda^{-1/2} \Lambda \Lambda^{-1/2} = I.$$

□

To simplify the discussion, consider the case where DTS is applied with  $\beta_t = 1$ . Then, conditional probability some action  $i \in [k]$  is selected is the posterior probability it is optimal, i.e.  $\mathbb{P}(I_t = i | H_t) = \mathbb{P}(I^* = i | H_t)$ . The challenge with using the posterior probabilities as inverse propensity weights is that they often decay to zero at an exponential rate for some actions. The key to our proof is that we only need to control the propensity assigned to the optimal action and this does not shrink, due to the next lemma. Recall  $\alpha_{t,i} = \mathbb{P}(I^* = i | H_t)$  is best-arm learning problems.

**Lemma 11** (Inverse propensity of the optimal action). *For any  $t \in \mathbb{N}$ ,  $\mathbb{E} \left[ \frac{1}{\alpha_{t,I^*}} \right] = k$ .*

*Proof.* Proof of Lemma 11. By the tower property of conditional expectation, and the fact that  $\alpha_{t,i}$  is  $H_t$  measurable,

$$\mathbb{E} \left[ \frac{1}{\alpha_{t,I^*}} \right] = \sum_{i=1}^k \mathbb{E} \left[ \frac{\mathbb{1}(I^* = i)}{\alpha_{t,i}} \right] = \sum_{i=1}^k \mathbb{E} \left[ \mathbb{E} \left[ \frac{\mathbb{1}(I^* = i)}{\alpha_{t,i}} \mid H_t \right] \right] = \sum_{i=1}^k \mathbb{E} \left[ \frac{\alpha_{t,i}}{\alpha_{t,i}} \right] = k.$$

□

Below, we use this to bounding the  $\mathbb{E} [S_{T,I^*}]$  term in Lemma 8, which, crucially, down-weights the importance of having low posterior variance at arms that are unlikely to be optimal. Combining the next lemma, Lemma 8 and Lemma 10 completes the proof of Proposition 1.

**Lemma 12** (Posterior variance bound). *Under the conditions of Proposition 1,*

$$\mathbb{E} [\tilde{S}_{T,I^*}] \preceq 2k \cdot S_{\text{full}}.$$

*Proof.* Proof of Lemma 12. Under DTS with  $\beta_t \geq \frac{1}{2}$ , we have  $\mathbb{P}(I_t = i \mid H_t) \geq \frac{\alpha_{t,i}}{2}$ . By Lemma 3,

$$\tilde{S}_{T,i} \preceq S_{\text{full}} \left( \Sigma_1^{-1} + \sigma^{-2} \sum_{t=1}^T \frac{X_t X_t^\top}{\mathbb{P}(I_t = i \mid H_t)} \right) S_{\text{full}} \preceq S_{\text{full}} \left( \Sigma_1^{-1} + \sigma^{-2} \sum_{t=1}^T \frac{2X_t X_t^\top}{\alpha_{t,i}} \right) S_{\text{full}}.$$

Applying this with Lemma 11 gives

$$\begin{aligned} \mathbb{E} [\tilde{S}_{T,I^*}] &\preceq S_{\text{full}} \left( \Sigma_1^{-1} + 2\sigma^{-2} \sum_{t=1}^T X_t X_t^\top \mathbb{E} \left[ \frac{1}{\alpha_{t,I^*}} \right] \right) S_{\text{full}} \preceq S_{\text{full}} \left( \Sigma_1^{-1} + 2k\sigma^{-2} \sum_{t=1}^T X_t X_t^\top \right) S_{\text{full}} \\ &\preceq S_{\text{full}} \left( 2k\Sigma_1^{-1} + 2k\sigma^{-2} \sum_{t=1}^T X_t X_t^\top \right) S_{\text{full}} \\ &= 2k \cdot S_{\text{full}}. \end{aligned}$$

□

## C.4 Proof of Corollary 1

*Proof.* Proof of Corollary 1. Our goal is to show that  $V(x_{1:T}) \preceq \sigma^2/T$  where

$$V(x_{1:T}) = X_{\text{pop}}^\top \left( \Sigma_1^{-1} + \sigma^{-2} \sum_{t=1}^T x_t x_t^\top \right)^{-1} X_{\text{pop}} \preceq X_{\text{pop}}^\top \left( \underbrace{\lambda_{\min}(\Sigma_1^{-1}) I + \sigma^{-2} T X_{\text{pop}} X_{\text{pop}}^\top}_{:=\Lambda} \right)^{-1} X_{\text{pop}}.$$

The matrix  $\Lambda$  has an eigenvector  $v = X_{\text{pop}}$  with eigenvalue  $\lambda = \lambda_{\min}(\Sigma_1^{-1}) + \sigma^{-2} T \|X_{\text{pop}}\|_2^2$ . Then  $v$  is also an eigenvalue of  $\Lambda^{-1}$  with corresponding eigenvalue  $1/\lambda$ . Therefore  $v^\top \Lambda^{-1} v = v^\top [v/\lambda] = \|v\|_2^2/\lambda \leq \sigma^2/T$ . □

## C.5 Returning to prove Lemma 10

**Notation.**

For most of the remainder of this proof, we fix an arm  $i$ . It is cleaner to omit  $i$  from the notation for the time being. We then face the following problem. There is a random process  $(Z_t)_{t \in \mathbb{N}}$  where  $Z_t \in \{0, 1\}$  ( $\dots$  think of  $Z_t$  as  $\mathbb{1}(I_t = i)$ ) and a deterministic sequence of vectors  $(x_t)_{t \in \mathbb{N}}$  where  $x_t \in \mathbb{R}^d$  satisfies  $\|x_t\|_2 \leq 1$ . Suppose that  $(Z_t)_{t \in \mathbb{N}}$  is adapted to some filtration  $(\mathcal{F}_t)_{t \in \mathbb{N}}$ . We take  $\mathbb{P}_t(\cdot) = \mathbb{P}(\cdot \mid \mathcal{F}_t)$  and  $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$ .

We are interested in the random matrices

$$S_T^{-1} = \Sigma_1^{-1} + \sum_{t=1}^T Z_t x_t x_t^\top \quad \text{and} \quad \tilde{S}_T^{-1} = \Sigma_1^{-1} + \sum_{t=1}^T p_t x_t x_t^\top$$

where  $p_t = \mathbb{P}_{t-1}(Z_t = 1)$ .

Define  $D_t = (p_t - Z_t)x_t x_t^\top$ . We study

$$A_0 := 0 \quad \text{and} \quad A_t := \tilde{S}_t^{-1} - S_t^{-1} = \sum_{\ell=1}^t D_\ell,$$

which in the sum of matrix martingale differences. We will follow Tropp [2011] fairly closely. Define

$\phi_t(\gamma) := \log \mathbb{E}_{t-1} [e^{\gamma D_t}]$ . Then set

$$\Phi_0 := 0 \quad \text{and} \quad \Phi_t(\gamma) := \sum_{\ell=1}^t \phi_\ell(\gamma).$$

Recognize that both  $A_0$  and  $\Phi_0$  are matrices where all elements equal zero. Here  $\Phi_T(\gamma)$  measures the total variability of the process. Our aim is to show that  $A_T$  can only be large if  $\Phi_T(\gamma)$  is large.

### A Bernstein-type bound on the cumulants.

We first recall a random matrix analogue of Bernstein's bound on the moment generating function of bounded random variables.

**Lemma 13** (Lemma 6.7 of [Tropp \[2012\]](#)). *Suppose  $D$  is a random self-adjoint matrix that satisfies*

$$\mathbb{E}D = 0 \quad \text{and} \quad \mathbb{P}(\lambda_{\max}(D) \leq 1) = 1.$$

Then

$$\mathbb{E}e^{\gamma D} \preceq \exp\left\{(e^\gamma - \gamma - 1)\mathbb{E}[D^2]\right\} \quad \forall \gamma > 0.$$

As a consequence of this, we can bound the sum of cumulants  $\Phi_t(\gamma)$  by a simpler quantity that closely mimics  $\tilde{S}_T$ .

**Lemma 14.** *For any  $\gamma > 0$ ,*

$$\Phi_T(\gamma) \preceq (e^\gamma - \gamma - 1) \sum_{t=1}^T p_t x_t x_t^\top.$$

*Proof.* Proof of Lemma 14. For notional convenience define  $f(\gamma) = e^\gamma - \gamma - 1$ . We have  $\lambda_{\max}(D_t) \leq |\alpha_t - Z_t| \text{tr}(x_t x_t^\top) \leq \|x_t\|_2^2 \leq 1$  where the last inequality follows from the assumption and  $\text{tr}$  denotes the trace operator. This allow us to apply the matrix Bernstein inequality above. By Lemma 13, we have immediately that  $\phi_t(\gamma) \preceq f(\gamma)\mathbb{E}_{t-1}[D_t^2]$ . Using this gives,

$$\begin{aligned} \Phi_T(\gamma) &\preceq f(\gamma) \sum_{t=1}^T \mathbb{E}_{t-1}[D_t^2] = f(\gamma) \sum_{t=1}^T \mathbb{E}_{t-1}[(p_t - Z_t)^2] (x_t x_t^\top) (x_t x_t^\top) \\ &= f(\gamma) \sum_{t=1}^T p_t(1 - p_t) \|x_t\|_2^2 x_t x_t^\top \\ &\preceq f(\gamma) \sum_{t=1}^T p_t x_t x_t^\top. \end{aligned}$$

□

### An exponential super-martingale.

Again, our goal is to show that  $A_t$  can only be large if  $\Phi_t(\gamma)$  is large. To this end, define

$$M_t(\gamma) = \text{tr} \exp\{\gamma A_t - \Phi_t(\gamma)\} \quad \forall t \in \{0, 1, \dots\}$$

where  $\text{tr}$  denotes the trace operator. We now show that  $M_t(\gamma)$  is a super-martingale, following the proof of Lemma 2.1 of [Tropp \[2011\]](#). We first state a powerful result of [Lieb \[1973\]](#) and then recall a simple corollary that is stated also in [Tropp \[2011\]](#).

**Theorem 1** (Lieb, 1973). *Fix a self-adjoint matrix  $H$ . The function  $A \mapsto \text{tr} \exp(H + \log(A))$  is concave on the positive-definite cone.*

**Corollary 2.** Fix a self-adjoint matrix  $H$ . For a random self-adjoint matrix  $X$ ,

$$\mathbb{E} [\operatorname{tr} \exp(H + X)] \leq \operatorname{tr} \exp\{H + \log(\mathbb{E}e^X)\}.$$

We now conclude that  $M_t(\gamma)$  is a super-martingale.

**Corollary 3.** For each  $\gamma > 0$ ,  $\{M_t(\gamma) : t = 0, 1, \dots\}$  is a super-martingale with initial value  $M_0(\gamma) = d$ .

*Proof.* Proof of Corollary 3. By definition,

$$M_0(\gamma) = \operatorname{tr} \exp\{\gamma A_0 - \Phi_0(\gamma)\} = \operatorname{tr} \exp\{0\} = \operatorname{tr} I = d.$$

For  $t > 0$ , taking conditional expectations gives

$$\begin{aligned} \mathbb{E}_{t-1} [M_t(\gamma)] &= \mathbb{E}_{t-1} [\operatorname{tr} \exp\{\gamma A_{t-1} - \Phi_t(\gamma) + \gamma D_t\}] \\ &\leq \operatorname{tr} \exp\{\gamma A_{t-1} - \Phi_t(\gamma) + \phi_t(\gamma)\} \\ &= \operatorname{tr} \exp\{\gamma A_{t-1} - \Phi_{t-1}(\gamma)\} = M_{t-1}(\gamma), \end{aligned}$$

where the inequality follows by Corollary 2. □

### Boundary crossing probabilities.

Here is where we begin to deviate from Tropp [2011]. The next result gives a boundary that  $A_t$  is unlikely to ever cross. The proof applies the same stopping time argument as the proof of one of Doob's martingale inequalities.

**Lemma 15.** For any fixed  $\delta > 0$  and  $\gamma > 0$ , with probability exceeding  $1 - \delta$ ,

$$A_t \preceq \frac{1}{\gamma} \left[ \Phi_t(\gamma) + \log \left( \frac{d}{\delta} \right) I \right]$$

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

*Proof.* Proof of Lemma 15. Fix  $\gamma > 0$  throughout. We have

$$\begin{aligned} \mathbb{P}(\lambda_{\max}(\gamma A_t - \Phi_t(\gamma)) \geq y) &= \mathbb{P}(e^{\lambda_{\max}(\gamma A_t - \Phi_t(\gamma))} \geq e^y) \leq \mathbb{P}(\operatorname{tr} e^{(\gamma A_t - \Phi_t(\gamma))} \geq e^y) \\ &\leq e^{-y} \mathbb{E} [\operatorname{tr} e^{(\gamma A_t - \Phi_t(\gamma))}] \\ &= e^{-y} \mathbb{E} [M_t(\gamma)]. \end{aligned}$$

The same inequalities hold for any bounded stopping time  $\tau$ , yielding

$$\mathbb{P}(\lambda_{\max}(\gamma A_\tau - \Phi_\tau(\gamma)) \geq y) \leq e^{-y} \mathbb{E} [M_\tau(\gamma)].$$

Take  $\tau = \inf\{t \in \mathbb{N} : \lambda_{\max}(\gamma A_t - \Phi_t(\gamma)) \geq y\}$ , with the convention that  $\tau = \infty$  if  $\lambda_{\max}(\gamma A_t - \Phi_t(\gamma)) < y$  for every  $t \in \mathbb{N}$ . Then,

$$\mathbb{P}(\exists t \leq N : \lambda_{\max}(\gamma A_t - \Phi_t(\gamma)) \geq y) = \mathbb{P}(\lambda_{\max}(\gamma A_{\tau \wedge N} - \Phi_{\tau \wedge N}(\gamma)) \geq y) \leq \mathbb{E} [M_{\tau \wedge N}(\gamma)] e^{-y} \leq d e^{-y}.$$

That  $\mathbb{E} [M_{\tau \wedge N}(\gamma)] \leq d$  uses Corollary 3 and Doob's optional sampling theorem. Taking  $N \rightarrow \infty$  and applying the monotone convergence theorem gives,

$$\mathbb{P}(\exists t \in \mathbb{N} : \lambda_{\max}(\gamma A_t - \Phi_t(\gamma)) \geq y) \leq d e^{-y}.$$



Choosing  $y = \log(d/\delta)$ , we have that with probability at least  $1 - \delta$ , the inequality

$$\lambda_{\max}(\gamma A_t - \Phi_t(\gamma)) \leq \log\left(\frac{d}{\delta}\right)$$

holds for every  $t \in \mathbb{N}$ . □

The next result bounds the deviation  $S_t^{-1} - \tilde{S}_t^{-1}$  in terms of  $\tilde{S}_t^{-1}$ . Notice that  $2 - e < 0$ , so the right hand side if this inequality is a negative definite matrix. But we also get the inequality  $S_t^{-1} \succeq (3 - e)\tilde{S}_t^{-1} - \log(d/\delta)I$  and here it is important that  $3 - e > 0$ .

**Lemma 16.** *For any given  $\delta > 0$ , with probability at least  $1 - \delta$ , the inequality*

$$S_t^{-1} - \tilde{S}_t^{-1} \succeq (2 - e)\tilde{S}_t^{-1} - \log(d/\delta)I$$

holds for every  $t \in \mathbb{N}$ .

*Proof.* Proof of Lemma 16. Define  $f(\gamma) = e^\gamma - \gamma - 1$ . Recall  $A_t := \tilde{S}_t^{-1} - S_t^{-1}$ . By applying Lemmas 15 and 14, we get that with probability at least  $1 - \delta$ ,

$$S_t^{-1} - \tilde{S}_t^{-1} = -A_t \succeq -\frac{1}{\gamma} [\Phi_t(\gamma) + \log(d/\delta)I] \succeq -\frac{1}{\gamma} \left[ f(\gamma) \sum_{\ell=1}^t p_\ell x_\ell x_\ell^\top + \log(d/\delta)I \right].$$

Picking  $\gamma = 1$ , we have

$$S_t^{-1} - \tilde{S}_t^{-1} \succeq (2 - e) \sum_{\ell=1}^t p_\ell x_\ell x_\ell^\top - \log(d/\delta)I = (2 - e)\tilde{S}_t^{-1} - \log(d/\delta)I.$$

□

### Multiplicative ratio deviation inequality.

We reformulate the results to provide a ‘‘multiplicative’’ bound.

**Lemma 17.** *For any  $\delta > 0$ , with probability at least  $1 - \delta$ , the following inequality holds simultaneously for every  $t \in \mathbb{N}$ :*

$$S_t \preceq c_\delta \tilde{S}_t \quad \text{where} \quad c_\delta = \max\{8 \log(d/\delta) \cdot \lambda_{\max}(\Sigma_1), 8\}$$

*Proof.* Proof of Lemma 17. From Lemma 16, we have that  $S_t^{-1} \succeq (3 - e)\tilde{S}_t^{-1} - \log(d/\delta)I$ . We also know that  $S_t^{-1} \succeq \Sigma_1^{-1}$ . Combining these implies that for any arbitrary unit vector  $u$ ,

$$u^\top S_t^{-1} u \geq \max\left\{ \lambda_{\min}(\Sigma_1^{-1}), (3 - e) u^\top \tilde{S}_t^{-1} u - \log(d/\delta) \right\}.$$

If  $(3 - e) u^\top \tilde{S}_t^{-1} u \geq 2 \log(d/\delta)$ , we have  $u^\top S_t^{-1} u \geq \frac{3-e}{2} u^\top \tilde{S}_t^{-1} u$ . On the other hand, if  $(3 - e) u^\top \tilde{S}_t^{-1} u \leq 2 \log(d/\delta)$ , we have

$$u^\top S_t^{-1} u \geq \lambda_{\min}(\Sigma_1^{-1}) = \frac{\lambda_{\min}(\Sigma_1^{-1})}{2 \log(d/\delta)} \cdot 2 \log(d/\delta) \geq \frac{\lambda_{\min}(\Sigma_1^{-1})}{\log(d/\delta)} \cdot \frac{3 - e}{2} \cdot u^\top \tilde{S}_t^{-1} u$$

In either case we have that for an arbitrary unit vector  $u$ ,

$$u^\top S_t^{-1} u \geq c_1 u^\top \tilde{S}_t^{-1} u \quad \text{where} \quad c_1 = \min\left\{ \frac{\lambda_{\min}(\Sigma_1^{-1})}{\log(d/\delta)}, 1 \right\} \cdot \frac{3 - e}{2}$$

Viewing this as a relation of the form  $S_t^{-1} \succeq c_1 \tilde{S}_t^{-1}$  and taking the inverse gives the result. We also simplify the expression using that  $\frac{2}{3-e} < 8$ . □

### Completing the proof of Lemma 10.

The remainder of the proof follows by a messy calculation.

*Proof.* Proof of Lemma 10. By Lemma 17, for any  $\delta > 0$ , with probability exceeding  $1 - \delta$ , the following inequality holds simultaneously for every  $T \in \mathbb{N}$ :

$$S_{T,i} \preceq c_\delta \tilde{S}_{T,i} \quad \text{where} \quad c_\delta = \max \{8 \log(d/\delta) \cdot \lambda_{\max}(\Sigma_1), 8\}.$$

Combining this with the fact that  $S_{T,i} \preceq \Sigma_1$  almost surely, we have, that for every  $\delta > 0$

$$\mathbb{E} [S_{T,i}] \preceq c_\delta \mathbb{E} [\tilde{S}_{T,i}] + \delta \Sigma_1.$$

Using also that  $\tilde{S}_{T,i} \succeq S_{\text{full}}$  almost surely and choosing  $\delta^* = \lambda_{\min}(S_{\text{full}}) / \lambda_{\max}(\Sigma_1)$  gives,

$$\mathbb{E} [S_{T,i}] \preceq c_{\delta^*} \mathbb{E} [\tilde{S}_{T,i}] + \delta^* \Sigma_1 \preceq c_{\delta^*} \mathbb{E} [\tilde{S}_{T,i}] + \lambda_{\min}(S_{\text{full}}) I \preceq (c_{\delta^*} + 1) \mathbb{E} [\tilde{S}_{T,i}].$$

Now,

$$\begin{aligned} c_{\delta^*} + 1 &= \max \{8 \log(d/\delta^*) \cdot \lambda_{\max}(\Sigma_1), 8\} + 1 \\ &\leq \max \{9 \log(d/\delta^*) \cdot \lambda_{\max}(\Sigma_1), 9\} \\ &= \max \left\{ 9 \log \left( \frac{d \lambda_{\max}(\Sigma_1)}{\lambda_{\min}(S_{\text{full}})} \right) \cdot \lambda_{\max}(\Sigma_1), 9 \right\} \\ &= \max \left\{ 9 \log \left( d \lambda_{\max}(\Sigma_1) \lambda_{\max}(S_{\text{full}}^{-1}) \right) \cdot \lambda_{\max}(\Sigma_1), 9 \right\} \\ &\leq \max \left\{ 9 \log \left( d \lambda_{\max}(\Sigma_1) \left[ \lambda_{\max}(\Sigma_1^{-1}) + \sigma^{-2} T \right] \right) \cdot \lambda_{\max}(\Sigma_1), 9 \right\}. \end{aligned}$$

The last step uses that  $S_{\text{full}}^{-1} = \Sigma_1^{-1} + \sum_{t=1}^T x_t x_t^\top \preceq \left( \lambda_{\max}(\Sigma_1^{-1}) + \sigma^{-2} T \right) I$ . □

## D Efficient Implementation of Algorithm 2 and Proof of Lemma 5

In this section, we show that the tuning parameter  $\beta_t$  in Algorithm 2 can be calculated by solving a one-dimensional fixed-point equation. This could be implemented efficiently by using bisection search or by Newton's method.

If  $\arg \max_{i \in [k]} m_{t,i}$  is not unique, Algorithm 2 simply sets  $\beta_t = \beta_{t-1}$ . Now suppose  $\arg \max_{i \in [k]} m_{t,i}$  is unique. Recall  $\hat{I}_t = \arg \max_{i \in [k]} m_{t,i}$ , and define  $\Delta_{t,i} \triangleq m_{t,\hat{I}_t} - m_{t,i}$  for each  $i \in [k]$ . Equation (20) implies there exists  $y > 0$  such that

$$\frac{1 + \hat{p}_{\hat{I}_t} \hat{p}_i^{-1}}{\Delta_{t,i}^2} = y, \quad \forall i \neq \hat{I}_t.$$

Clearly  $y > \max_{i \neq \hat{I}_t} \Delta_{t,i}^{-2}$  and

$$\frac{\hat{p}_{\hat{I}_t}}{\hat{p}_i} = \Delta_{t,i}^2 y - 1, \quad \forall i \neq \hat{I}_t. \tag{30}$$

Equations (30) and (21) imply

$$\sum_{i \neq \hat{I}_t} \left( \Delta_{t,i}^2 y - 1 \right)^{-2} = 1. \tag{31}$$

Let  $y_t$  solves Equation (31). We can solve this fixed-point equation for  $y$  using, for example, bisection search or Newton's method. Notice that if Newton's method is used, one may wish to save the value of  $y_{t-1}$  solved in the previous time period, which provides an effective initial point for finding an updated value  $y_t$ . Finally

$\sum_{i \in [k]} \hat{p}_i = 1$  and Equation (30) imply

$$\beta_t = \hat{p}_{\hat{I}_t} = \frac{1}{1 + \sum_{i \neq \hat{I}_t} (\Delta_{t,i}^2 y_t - 1)^{-1}}.$$

This completes the proof of Lemma 5

In this section, we also include the following uniform bounds on the tuning parameters produced by Algorithm 2.

**Lemma 18.** *Under Algorithm 2, for any  $t \in \mathbb{N}$ ,*

$$\frac{1}{1 + \sqrt{k-1}} \leq \beta_t \leq 1/2. \quad (32)$$

*Proof.* Proof of Lemma 18. We prove this result by induction. Under Algorithm 2,  $\beta_1 = 1/2$  satisfies Equation (32), and suppose  $\beta_{t-1}$  satisfies Equation (32). If  $\arg \max_{i \in [k]} m_{t,i}$  is not unique, we are done since  $\beta_t = \beta_{t-1}$ . Now suppose  $\hat{I}_t = \arg \max_{i \in [k]} m_{t,i}$  is unique, and we are going to show  $\beta_t = \hat{p}_{\hat{I}_t}$  satisfies Equation (32) where  $\hat{p} = p^*(\mu_t)$  is the solution to Equations (20) and (21). By Equation (21),

$$\hat{p}_{\hat{I}_t} = \sqrt{\sum_{i \neq \hat{I}_t} \hat{p}_i^2} \leq \sum_{i \neq \hat{I}_t} \hat{p}_i = 1 - \hat{p}_{\hat{I}_t},$$

which gives  $\hat{p}_{\hat{I}_t} \leq 1/2$ . On the other hand, by using the Cauchy-Schwarz inequality, we have

$$\hat{p}_{\hat{I}_t} = \sqrt{\sum_{i \neq \hat{I}_t} \hat{p}_i^2} \geq \frac{\sum_{i \neq \hat{I}_t} \hat{p}_i}{\sqrt{k-1}} = \frac{1 - \hat{p}_{\hat{I}_t}}{\sqrt{k-1}},$$

which gives  $\hat{p}_{\hat{I}_t} \geq \frac{1}{1 + \sqrt{k-1}}$ . This completes the proof.  $\square$

## E Relation to fixed-budget and fixed-confidence objectives in best arm identification

It is natural to wonder why Subsection 8.4, somewhat abruptly, introduces a stopping time  $\tau$ . Another popular formulation, called the “fixed-budget” setting, fixes the length of the experiment in advance. Some papers carry out large-deviations analysis of the probability of incorrect selection as the budget grows [Glynn and Juneja, 2004, Chen et al., 2000]. When reward distributions are Gaussian, those papers suggest the allocation  $p^*(\theta)$  and complexity term  $\Gamma_\theta$ . In other cases the terms do not quite match. A lower bound of Carpentier and Locatelli [2016] implies that these large deviations rates are not attainable by an adaptive algorithm. See the discussion in Ariu et al. [2021]. These open theoretical questions are not our primary focus, and we sidestep them by allowing for adaptive stopping.

Another popular problem formulation in the best-arm identification literature is called the “fixed confidence” setting. This formulation places a uniform constraint on the probability of incorrect selection, restricting to algorithms satisfying  $\sup_{\theta_0} \mathbb{P}(\hat{I}_\tau^+ \neq I^*(\theta_0) \mid \theta = \theta_0) \leq \delta$ . For problems without contexts, Chan and Lai [2006] and Kaufmann et al. [2016] characterize the minimal expected sample size  $\mathbb{E}[\tau \mid \theta = \theta_0]$  attainable by algorithms satisfying this constraint in the limit where  $\delta \rightarrow 0$ . In a special case of our formulation without contexts, the complexity measure  $\Gamma_\theta$  and optimal sampling proportions  $p^*(\theta)$  match those in Chan and Lai [2006], Kaufmann et al. [2016] and also the results regarding posterior convergence in Russo [2020].

We prefer the objective in (24) as it reflects both the costs and benefits of sampling in a single measure. This choice has some impact on the proofs of our technical results. Upper bounds are easier to derive as we no longer require a stopping rule that guarantees a very precise bound like  $\sup_{\theta_0} \mathbb{P}(\hat{I}_\tau^+ \neq I^*(\theta_0) \mid \theta = \theta_0) \leq \delta$ .

The cost of this is that the lower bound is more difficult to derive. In Proposition 4 in the appendix, we lower bound the average sample size among any algorithm satisfying  $\limsup_{c \rightarrow 0} c^{-1} \mathbb{E}[\Delta_\tau \mid \theta = \theta_0] < \infty$  for each fixed  $\theta_0$ . Hence, the lower bound applies to procedures that suffer much larger simple regret in some states of nature than others, as long as the scaling in  $c$  is preserved. Our understanding is that existing proofs would not immediately yield a lower bound on (24) even in problems without contexts. Remark 2 in the proof discusses the technical challenges.

## F Proof of the Lower Bound in Proposition 3

### F.1 Outline of the proof

Recall that  $\Gamma_\theta^{-1}$  was defined in (19) through an optimization problem over sampling proportions  $(p_1, \dots, p_k)$ . A measure like this already restricts, implicitly, to context-independent sampling rules whose behavior is characterized in this way. More generally, in a problem with Gaussian linear structure, one would at least need to reason about a design matrix like  $M_i = \frac{1}{i} \mathbb{E} [\sum_{\ell=1}^i \mathbb{1}(I_\ell = i) X_\ell X_\ell^\top]$ .

Using change of measure arguments that date back at least to Chernoff [1959], we reduce the lower bound studying the value of a zero-sum game between an experimenter and a skeptic. This game generalizes a two player game that appears in Chernoff [1959], Russo [2020], and Garivier and Kaufmann [2016] to treat problems with contexts. In this game, both players suspect the true state is  $\theta = \theta_0$  and so the best arm is  $I^*(\theta_0)$ . Trying to collect the evidence to certify this, the experimenter picks a collection of positive semi-definite matrices  $M = (M_1, \dots, M_k)$  which obey the constraint  $M_1 + \dots + M_k \preceq \mathbb{E}[X_1 X_1^\top]$ . The skeptic responds by picking an alternative state of nature  $\theta_1$  that would induce a different optimal arm. The skeptic's goal is to challenge the experimenter's claim by picking  $\theta_1$  that makes a certain KL divergence small.

Our key result is Proposition 5, given in the appendix, which solves for an equilibrium of this game. In equilibrium, the experimenter picks  $M_i^* = p_i^* \mathbb{E} [X_1 X_1^\top]$ , which reflects context-independent sampling with the long-run proportions in (19). The value of the game is shown to equal  $\Gamma_{\theta_0}^{-1}$ . These steps allow us to establish that more complex strategies cannot outperform context-independent sampling asymptotically.

#### A restatement of the result.

The lower bound in Proposition 3 follows from the next proposition, which characterizes the sample size required to ensure simple regret vanishes at a desired rate. To see this, suppose that (26) does not hold. Since  $c\tau \geq 0$ , this implies  $\mathbb{E}[\Delta_\tau \mid \theta = \theta_1] = O_{\theta_1}(c \log(1/c))$  for every  $\theta_1$ . Then (34) implies

$$\mathbb{E}[c\tau + \Delta_\tau \mid \theta = \theta_0] \geq c \mathbb{E}[\tau \mid \theta = \theta_0] \geq c [1 + o_{\theta_0}(1)] \Gamma_{\theta_0} \log(1/[c \log(1/c)]) = c [1 + o_{\theta_0}(1)] \Gamma_{\theta_0} \log(1/c).$$

The final equality uses that  $\log(1/[c \log(1/c)]) = \log(1/c)[1 - \frac{\log \log(1/c)}{\log(1/c)}] = \log(1/c)[1 + o(1)]$ .

**Proposition 4.** *Suppose Assumption 1 holds. Under any admissible sampling rule, stopping rule  $\tau = \tau(\delta)$ , and selection rule, if as  $\delta \rightarrow 0$ ,*

$$\mathbb{E}[\Delta_\tau \mid \theta = \theta_0] = O_{\theta_0}(\delta) \quad \text{for all } \theta_0 \in \Theta, \quad (33)$$

then

$$\mathbb{E}[\tau \mid \theta = \theta_0] \geq [1 + o_{\theta_0}(1)] \Gamma_{\theta_0} \log(1/\delta) \quad \text{for all } \theta_0 \in \Theta. \quad (34)$$

In this section we describe the sequence of results that lead to the lower bound in Proposition 4. It is derived formally as the solution to a two-player zero-sum game between an experimenter and nature.

#### A general change of measure argument.

Consider two alternative states of nature  $\theta_0, \theta_1 \in \Theta$  under which the optimal arm differs (i.e.  $I^*(\theta_0) \neq I^*(\theta_1)$ ). The decision-maker would like to reach the correct decision in each case. That is, if the decision-maker stops

at time  $\tau$  and picks arm  $\hat{I}_\tau^+$ , they would like  $\mathbb{P}(\hat{I}_\tau^+ = I^*(\theta_0) \mid \theta = \theta_0)$  and  $\mathbb{P}(\hat{I}_\tau^+ = I^*(\theta_1) \mid \theta = \theta_1)$  to each be close to one. Stated differently, they would like there to be a large divergence between the distribution of  $\hat{I}_\tau^+$  under  $\theta_0$  and  $\theta_1$ . The next result makes this formal.

For the next two lemmas, we need to extend the definition of the variables  $\hat{I}_\tau$  and  $H_\tau$  on sample paths where  $\tau = \infty$ . Define  $\hat{I}_\tau = \emptyset \notin \{1, \dots, k\}$  if  $\tau = \infty$ , indicating that  $\hat{I}_\tau$  cannot correctly identify the best arm. When  $\tau = \infty$ , take  $H_\tau = ((X_\ell, I_\ell, R_\ell) : \ell \in \mathbb{N})$  to be the infinite history generated by the allocation rule.

**Lemma 19** (Bretagnolle–Huber inequality). *Fix any  $\theta_0, \theta_1 \in \Theta$ .*

$$\mathbb{P}(\hat{I}_\tau^+ \neq I^*(\theta_0) \mid \theta = \theta_0) + \mathbb{P}(\hat{I}_\tau^+ = I^*(\theta_0) \mid \theta = \theta_1) \geq \frac{1}{2} \exp \left\{ -D_{\text{KL}}(\mathbb{P}(H_\tau^+ \in \cdot \mid \theta = \theta_0) \parallel \mathbb{P}(H_\tau^+ \in \cdot \mid \theta = \theta_1)) \right\}.$$

*Proof.* Proof of Lemma 19. See [Lattimore and Szepesvári, 2020a, Theorem 14.2].  $\square$

If (33) holds as  $\delta \rightarrow 0$ , the left-hand-side in the lemma above is  $O(\delta)$ . Taking the logarithm of both sides and dividing each side by  $\log(1/\delta)$  yields the next corollary.

**Corollary 4.** *Suppose (33) holds as  $\delta \rightarrow 0$ . Fix any  $\theta_1 \in \Theta$  such that  $I^*(\theta_1) \neq I^*(\theta_0)$ . Then,*

$$\limsup_{\delta \rightarrow 0} \frac{D_{\text{KL}}(\mathbb{P}(H_\tau^+ \in \cdot \mid \theta = \theta_0) \parallel \mathbb{P}(H_\tau^+ \in \cdot \mid \theta = \theta_1))}{\log(1/\delta)} \geq 1.$$

If the decision-maker wants to reach a correct decision with high probability, they need to gather a collection of observations (or history) that discriminates between  $\theta_0$  and  $\theta_1$  in the sense of making the divergence above large. Thankfully, this divergence has an explicit expression, which is given in the lemma below.

**Lemma 20.** *Fix any  $\theta_0, \theta_1 \in \Theta$ . If  $\mathbb{P}(\tau < \infty \mid \theta = \theta_0) = 1$ , then*

$$D_{\text{KL}}(\mathbb{P}(H_\tau^+ \in \cdot \mid \theta = \theta_0) \parallel \mathbb{P}(H_\tau^+ \in \cdot \mid \theta = \theta_1)) = \frac{1}{2\sigma^2} \sum_{i \in [k]} (\theta_0^{(i)} - \theta_1^{(i)})^\top \mathbb{E} \left[ \sum_{t=1}^{\tau} \mathbb{1}(I_t = i) X_t X_t^\top \mid \theta = \theta_0 \right] (\theta_0^{(i)} - \theta_1^{(i)}).$$

*Proof.* Proof of Lemma 20. Recall that  $H_t^+ = ((X_\ell, I_\ell, R_\ell) : \ell = 1, \dots, t)$ . Let  $\mathbb{P}_i = \mathbb{P}(\cdot \mid \theta = \theta_i)$  and  $\mathbb{E}_i = \mathbb{E}[\cdot \mid \theta = \theta_i]$  for  $i = 0, 1$ . The chain rule for stopping times in Lemma 21 implies

$$\begin{aligned} D_{\text{KL}}(\mathbb{P}_0(H_\tau^+ \in \cdot) \parallel \mathbb{P}_1(H_\tau^+ \in \cdot)) &= \mathbb{E}_0 \left[ \sum_{t=1}^{\tau} D_{\text{KL}}(\mathbb{P}_0((X_t, I_t, R_t) \in \cdot \mid H_{t-1}^+) \parallel \mathbb{P}_1((X_t, I_t, R_t) \in \cdot \mid H_{t-1}^+)) \right] \\ &= \mathbb{E}_0 \left[ \sum_{t=1}^{\tau} D_{\text{KL}}(\mathbb{P}_0((X_t, I_t) \in \cdot \mid H_{t-1}^+) \parallel \mathbb{P}_1((X_t, I_t) \in \cdot \mid H_{t-1}^+)) \right] \\ &\quad + \mathbb{E}_0 \left[ \sum_{t=1}^{\tau} D_{\text{KL}}(\mathbb{P}_0(R_t \in \cdot \mid H_{t-1}^+, X_t, I_t) \parallel \mathbb{P}_1(R_t \in \cdot \mid H_{t-1}^+, X_t, I_t)) \right] \\ &= \mathbb{E}_0 \left[ \sum_{t=1}^{\tau} \frac{1}{2\sigma^2} \left( X_t^\top (\theta_0^{(I_t)} - \theta_1^{(I_t)}) \right)^2 \right]. \end{aligned}$$

The result then follows from a simple calculation. The second equality above applies the chain rule. The final equality uses that, conditioned on  $H_{t-1}^+$ ,  $(X_t, I_t)$  have the same distribution under each  $\mathbb{P}_i$  and that under  $\mathbb{P}_i$ ,  $R_t \mid H_{t-1}^+, X_t, I_t \sim N(X_t^\top \theta_i^{(I_t)}, \sigma^2)$ .  $\square$

**Lemma 21** (Chain rule with stopping times). *Consider two probability spaces  $(\Omega, \mathcal{F}, \mathbb{P})$  and  $(\Omega, \mathcal{F}, \mathbb{Q})$  where  $\mathbb{Q}$  is absolutely continuous with respect to  $\mathbb{P}$ . Take  $\tau$  to be a  $\mathbb{P}$ -almost-surely finite stopping time adapted to  $(Z_t)_{t \in \mathbb{N}}$ .*

Then,

$$D_{\text{KL}}(\mathbb{P}(Z_{1:\tau} \in \cdot) \parallel \mathbb{Q}(Z_{1:\tau} \in \cdot)) = \mathbb{E}_{\mathbb{P}} \left[ \sum_{t=1}^{\tau} D_{\text{KL}} \left( \mathbb{P}(Z_t \in \cdot \mid Z_{1:(t-1)}) \parallel \mathbb{Q}(Z_t \in \cdot \mid Z_{1:(t-1)}) \right) \right]$$

*Proof.* Proof of Lemma 21. The result follows by applying the usual chain rule to the censored random variables  $\tilde{Z}_t = Z_t \mathbb{1}(\tau \geq t)$ . It is clear that the sequence  $Z_{1:\tau} = (Z_1, \dots, Z_{\tau})$  contains the same information as  $\tilde{Z}_{1:\infty} = (\tilde{Z}_1, \tilde{Z}_2, \dots)$ . Formally, there exists a function  $f$  with  $\tilde{Z}_{1:\infty} = f(Z_{1:\tau})$  and  $Z_{1:\tau} = f^{-1}(\tilde{Z}_{1:\infty})$ . It follows by the data-processing inequality that

$$D_{\text{KL}}(\mathbb{P}(Z_{1:\tau} \in \cdot) \parallel \mathbb{Q}(Z_{1:\tau} \in \cdot)) = D_{\text{KL}}(\mathbb{P}(\tilde{Z}_{1:\infty} \in \cdot) \parallel \mathbb{Q}(\tilde{Z}_{1:\infty} \in \cdot)).$$

Now,

$$\begin{aligned} D_{\text{KL}}(\mathbb{P}(\tilde{Z}_{1:\infty} \in \cdot) \parallel \mathbb{Q}(\tilde{Z}_{1:\infty} \in \cdot)) &= \lim_{T \rightarrow \infty} D_{\text{KL}}(\mathbb{P}(\tilde{Z}_{1:T} \in \cdot) \parallel \mathbb{Q}(\tilde{Z}_{1:T} \in \cdot)) \\ &= \lim_{T \rightarrow \infty} \mathbb{E}_{\mathbb{P}} \left[ \sum_{t=1}^T D_{\text{KL}} \left( \mathbb{P}(\tilde{Z}_t \in \cdot \mid \tilde{Z}_{1:(t-1)}) \parallel \mathbb{Q}(\tilde{Z}_t \in \cdot \mid \tilde{Z}_{1:(t-1)}) \right) \right] \\ &= \lim_{T \rightarrow \infty} \mathbb{E}_{\mathbb{P}} \left[ \sum_{t=1}^{\tau \wedge T} D_{\text{KL}} \left( \mathbb{P}(\tilde{Z}_t \in \cdot \mid \tilde{Z}_{1:(t-1)}) \parallel \mathbb{Q}(\tilde{Z}_t \in \cdot \mid \tilde{Z}_{1:(t-1)}) \right) \right] \\ &= \mathbb{E}_{\mathbb{P}} \left[ \sum_{t=1}^{\tau} D_{\text{KL}} \left( \mathbb{P}(\tilde{Z}_t \in \cdot \mid \tilde{Z}_{1:(t-1)}) \parallel \mathbb{Q}(\tilde{Z}_t \in \cdot \mid \tilde{Z}_{1:(t-1)}) \right) \right] \\ &= \mathbb{E}_{\mathbb{P}} \left[ \sum_{t=1}^{\tau} D_{\text{KL}} \left( \mathbb{P}(Z_t \in \cdot \mid Z_{1:(t-1)}) \parallel \mathbb{Q}(Z_t \in \cdot \mid Z_{1:(t-1)}) \right) \right]. \end{aligned}$$

The first equality is Corollary 5.2.5 in Gray [2011]. The second equality is the usual chain rule for KL-divergences. The third equality recognizes that conditioned on  $\tau < t$ ,  $\tilde{Z}_t = 0$  almost surely under both  $\mathbb{P}(\cdot)$  and  $\mathbb{Q}(\cdot)$  and so  $D_{\text{KL}}(\mathbb{P}(\tilde{Z}_t \in \cdot \mid \tilde{Z}_{1:(t-1)}) \parallel \mathbb{Q}(\tilde{Z}_t \in \cdot \mid \tilde{Z}_{1:(t-1)})) = 0$ . The fourth equality uses that KL-divergences are non-negative and applies the monotone convergence theorem to interchange the limit and expectation.  $\square$

### Two player game.

Based on the expression above, we define a two player zero-sum game between the experimenter, who tries to gather convincing evidence, and nature, who picks an alternative  $\theta_a$  for which the experimenter's evidence least convincing. The value of this game will determine our lower bound. Mimicking the expression for the KL divergence, define for a sequence of positive semidefinite matrices  $M = (M_1, \dots, M_k)$ ,

$$\Gamma_{\theta}^{-1}(M, \theta_a) \triangleq \frac{1}{2\sigma^2} \sum_{i \in [k]} \left( \theta^{(i)} - \theta_a^{(i)} \right)^{\top} M_i \left( \theta^{(i)} - \theta_a^{(i)} \right). \quad (35)$$

The KL divergence appearing in Lemma 20 is equal to  $\mathbb{E}[\tau \mid \theta = \theta_0] \Gamma_{\theta_0}^{-1}(M, \theta_a)$  if one were to plug-in a value  $M_i = \mathbb{E}[\sum_{t=1}^{\tau} \mathbb{1}(I_t = i) X_t X_t^{\top} \mid \theta_0] / \mathbb{E}[\tau \mid \theta = \theta_0]$ . Notice then that there is a natural constraint on how large the components of  $M$  can be, since

$$\sum_{i \in [k]} M_i = \mathbb{E}[X_1 X_1^{\top}] = \Lambda.$$



To bound the possible accumulation of information, we consider an experimenter who is free to select matrices from the family

$$\mathbb{M} = \left\{ (M_1, \dots, M_k) : \sum_{i \in [k]} M_i = \Lambda, M_i \succeq 0, M_i = M_i^\top \forall i \in [k] \right\}.$$

Nature will respond to the player by picking some alternative state of nature  $\theta_a$  from the following set that the design  $M$  does a poor job of ruling out,

$$\text{Alt}(\theta_0) \triangleq \left\{ \theta' \in \mathbb{R}^{dk} : I^*(\theta') \neq I^*(\theta_0) \right\}.$$

The equilibrium value of this game is  $\sup_{M \in \mathbb{M}} \inf_{\theta_a \in \text{Alt}(\theta)} \Gamma_\theta^{-1}(M, \theta_a)$ .

The next proposition describes an equilibrium of this game. The value of the game is shown to equal the simplified complexity term  $\Gamma_\theta^{-1}$  defined in (19). The experimenter's equilibrium strategy mimics context-independent sampling with proportions given by the optimal proportions in (20) and (21). Nature plays a distribution over just  $k-1$  vectors, defined as follows: for each arm  $i \neq I^*$  define  $\hat{\theta}_i = (\hat{\theta}_i^{(1)}, \dots, \hat{\theta}_i^{(k)})$  by

$$\hat{\theta}_i^{(i)} = \theta^{(i)} + \frac{\eta}{p_i^*} \Lambda^{-1} X_{\text{pop}} \quad \hat{\theta}_i^{(I^*)} = \theta^{(I^*)} - \frac{\eta}{p_{I^*}^*} \Lambda^{-1} X_{\text{pop}}, \quad \hat{\theta}_i^{(j)} = \theta^{(j)} \quad \forall j \notin \{i, I^*\} \quad (36)$$

where

$$\eta = \frac{(\mu(\theta, I^*, w) - \mu(\theta, i, w))^2}{2 \|X_{\text{pop}}\|_{\Lambda^{-1}}^2 [(p_i^*)^{-1} + (p_{I^*}^*)^{-1}]} \quad \forall i \neq I^*.$$

That  $\eta$  does not depend on  $i$  follows from (20). Let  $\mathcal{D}(\text{Alt}(\theta_0))$  denote the set of distributions over alternative states of nature and overload notation to write  $\Gamma_\theta^{-1}(M, q) = \mathbb{E}_{\theta_a \sim q} [\Gamma_\theta^{-1}(M, \theta_a)]$  for  $q \in \mathcal{D}(\text{Alt}(\theta))$ .

**Proposition 5.** *The complexity term  $\Gamma_\theta^{-1}$  defined in (19) satisfies*

$$\Gamma_\theta^{-1} = \sup_{M \in \mathbb{M}} \inf_{\theta_a \in \text{Alt}(\theta)} \Gamma_\theta^{-1}(M, \theta_a). \quad (37)$$

Let  $M_i^* = p_i^* \Lambda$  where  $(p_1^*, \dots, p_k^*)$  is defined by (20) and (21). Let  $q^*$  be a distribution supported on  $\{\hat{\theta}_i : i \neq I^*\}$ , defined in (36), where  $q^*(\hat{\theta}_i) = (p_i^*/p_{I^*}^*)^2$ . Then,

$$\sup_{M \in \mathbb{M}} \Gamma_\theta^{-1}(M, q^*) = \Gamma_\theta^{-1}(M^*, q^*) = \inf_{q \in \mathcal{D}(\text{Alt}(\theta))} \Gamma_\theta^{-1}(M^*, q). \quad (38)$$

*Proof.* Proof of Proposition 5. To start, we consider the best-response of nature,  $\min_{\theta_a \in \text{Alt}(\theta_0)} \Gamma_\theta^{-1}(M^*, \theta_a)$ . Throughout the proof use  $I^*$  to denote  $I^*(\theta)$ . It is immediate that

$$\min_{\theta_a \in \text{Alt}(\theta)} \frac{1}{2\sigma^2} \sum_{i \in [k]} \|\theta^{(i)} - \theta_a^{(i)}\|_{M_i^*}^2 = \frac{1}{2\sigma^2} \min_{i \neq I^*} \min_{X_{\text{pop}}^\top (\theta_a^{(i)} - \theta_a^{(I^*)}) \geq 0} \|\theta^{(i)} - \theta_a^{(i)}\|_{M_i^*}^2 + \|\theta^{(I^*)} - \theta_a^{(I^*)}\|_{M_{I^*}^*}^2.$$

Apply the transformation of variables  $Z_j = \theta^{(j)} - \theta_a^{(j)}$ . Then for each alternative arm  $i$  we have the optimization problem

$$\begin{aligned} \min \quad & \frac{1}{2\sigma^2} \left( \|Z_i\|_{M_i^*}^2 + \|Z_{I^*}\|_{M_{I^*}^*}^2 \right) \\ \text{subject to} \quad & (Z_i - Z_{I^*})^\top X_{\text{pop}} \geq \Delta_i \end{aligned} \quad (39)$$

where  $\Delta_i = \mu(\theta, I^*, w) - \mu(\theta, i, w) = X_{\text{pop}}^\top (\theta^{(I^*)} - \theta^{(i)})$ .

The KKT conditions for the problem (39) are

$$\begin{aligned}
M_i^* Z_i &= \eta X_{\text{pop}} \\
M_{I^*}^* Z_{I^*} &= -\eta X_{\text{pop}} \\
\eta \left[ (Z_i - Z_{I^*})^\top X_{\text{pop}} - \Delta_i \right] &= 0 \\
\eta &\geq 0.
\end{aligned}$$

A solution is given by

$$Z_i = \eta (M_i^*)^{-1} X_{\text{pop}} = \frac{\eta \Lambda^{-1} X_{\text{pop}}}{p_i^*} \quad \text{and} \quad Z_{I^*} = -\eta (M_{I^*}^*)^{-1} X_{\text{pop}} = -\frac{\eta \Lambda^{-1} X_{\text{pop}}}{p_{I^*}^*}$$

where

$$\eta = \frac{\Delta_i}{2X_{\text{pop}}^\top [(M_i^*)^{-1} + (M_{I^*}^*)^{-1}] X_{\text{pop}}} = \frac{(\mu(\theta, I^*, w) - \mu(\theta, i, w))^2}{2\|X_{\text{pop}}\|_{\Lambda^{-1}}^2 [(p_i^*)^{-1} + (p_{I^*}^*)^{-1}]}$$

Since the optimization problem is convex, satisfying the KKT conditions is sufficient to imply optimality. Recalling the transformation of variables  $\theta_a^{(i)} = \theta^{(i)} + Z_i$ , we find the optimal solution to (39) is exactly  $\hat{\theta}_i$  in (36). Plugging in shows that the optimal objective value of (39) is equal to  $\eta/\sigma^2$ . By the information balance property (20),  $\eta$  does not depend on  $i$ . In summary, we have shown that for each arm  $i \neq I^*$

$$\min_{\theta_a \in \text{Alt}(\theta)} \Gamma_\theta^{-1}(M^*, \theta_a) = \Gamma_\theta^{-1}(M^*, \hat{\theta}_i) = \frac{\eta}{\sigma^2} = \Gamma_\theta^{-1}.$$

The last equality recalls the definition of  $\Gamma_\theta$  in (19). Since we have just shown that each  $\hat{\theta}_i$  is a best response, so is the randomized strategy  $q^*$ .

Now, consider experiment's choice of  $M \in \mathbb{M}$  in response to nature's choice of  $q^*$ . We denote the standard

inner product on matrices by  $\langle A, B \rangle = \text{Trace}(A^\top B)$ . For any  $M \in \mathbb{M}$ ,

$$\begin{aligned}
\Gamma_\theta^{-1}(M, q^*) &= \sum_{j \neq I^*} q^*(\hat{\theta}_j) \left\{ \frac{1}{2\sigma^2} \sum_{i \in [k]} (\theta^{(i)} - \hat{\theta}_j^{(i)})^\top M_i (\theta^{(i)} - \hat{\theta}_j^{(i)}) \right\} \\
&= \frac{1}{2\sigma^2} \sum_{j \neq I^*} q^*(\hat{\theta}_j) \left[ \left\| \frac{\eta}{p_j^*} \Lambda^{-1} X_{\text{pop}} \right\|_{M_j}^2 + \left\| \frac{\eta}{p_{I^*}^*} \Lambda^{-1} X_{\text{pop}} \right\|_{M_{I^*}}^2 \right] \\
&= \frac{\eta^2}{2\sigma^2} \sum_{j \neq I^*} q^*(\hat{\theta}_j) \left[ \frac{1}{(p_j^*)^2} \left\| \Lambda^{-1} X_{\text{pop}} \right\|_{M_j}^2 + \frac{1}{(p_{I^*}^*)^2} \left\| \Lambda^{-1} X_{\text{pop}} \right\|_{M_{I^*}}^2 \right] \\
&= \frac{\eta^2}{2\sigma^2} \sum_{j \neq I^*} q^*(\hat{\theta}_j) \left[ \frac{\langle B, M_j \rangle}{(p_j^*)^2} + \frac{\langle B, M_{I^*} \rangle}{(p_{I^*}^*)^2} \right] \quad \text{where } B = (\Lambda^{-1} X_{\text{pop}})(\Lambda^{-1} X_{\text{pop}})^\top \\
&= \frac{\eta^2}{2\sigma^2} \sum_{j \neq I^*} \left[ \frac{q^*(\hat{\theta}_j)}{(p_j^*)^2} \right] \langle B, M_j \rangle + \left[ \frac{1}{(p_{I^*}^*)^2} \right] \langle B, M_{I^*} \rangle \\
&= \frac{\eta^2}{2\sigma^2} \sum_{j \neq I^*} \left[ \frac{1}{(p_{I^*}^*)^2} \right] \langle B, M_j \rangle + \left[ \frac{1}{(p_{I^*}^*)^2} \right] \langle B, M_{I^*} \rangle \quad \left( \text{since } q^*(\hat{\theta}_j) = \left( \frac{p_j^*}{p_{I^*}^*} \right)^2 \right) \\
&= \frac{\eta^2}{2\sigma^2 (p_{I^*}^*)^2} \langle B, \sum_{i=1}^k M_i \rangle \\
&= \frac{\eta^2}{2\sigma^2 (p_{I^*}^*)^2} \langle B, \Lambda \rangle.
\end{aligned}$$

When nature selects  $q^*$ , the experimenter is indifferent between all  $M = (M_1, \dots, M_k) \in \mathbb{M}$ . This implies in particular that  $M^* \in \arg \max_{M \in \mathbb{M}} \Gamma^{-1}(M, q^*)$ .

We have established (38). In particular,  $M^*$  and  $q^*$  form an equilibrium. We have already shown  $\Gamma_\theta^{-1} = \Gamma_\theta^{-1}(M^*, q^*)$ , so the simplified complexity  $\Gamma_\theta^{-1}$  is the equilibrium value of the game. We have to show (37). Since  $M^*$  and  $q^*$  form an equilibrium,

$$\inf_{q \in \mathcal{D}(\text{Alt}(\theta))} \sup_{M \in \mathbb{M}} \Gamma_\theta^{-1}(M, q) \leq \sup_{M \in \mathbb{M}} \Gamma_\theta^{-1}(M, q^*) = \Gamma_\theta^{-1}(M^*, q^*) = \inf_{q \in \mathcal{D}(\text{Alt}(\theta))} \Gamma_\theta^{-1}(M^*, q) \leq \sup_{M \in \mathbb{M}} \inf_{q \in \mathcal{D}(\text{Alt}(\theta))} \Gamma_\theta^{-1}(M, q).$$

The reverse inequality,

$$\sup_{M \in \mathbb{M}} \inf_{q \in \mathcal{D}(\text{Alt}(\theta))} \Gamma_\theta^{-1}(M, q) \leq \inf_{q \in \mathcal{D}(\text{Alt}(\theta))} \sup_{M \in \mathbb{M}} \Gamma_\theta^{-1}(M, q), \quad (40)$$

also holds, and so the inequalities above must hold with equality. Inequality (40) is a general result known as the max-min inequality ([Boyd and Vandenberghe, 2004, Chapter 5]). □

### Completing the proof.

A slight subtlety arises in completing the proof. The property (33), which underlies Corollary 4, is assumed to only hold for parameters in  $\Theta$ . This is the set of parameters with distinct arm means. The parameters  $\hat{\theta}_j$  played by nature in equilibrium are in the closure of  $\Theta$ , but not in  $\Theta$  itself. To complete the proof, we fix  $k-1$  parameters  $(\theta_j : j \neq I^*)$  where each  $\theta_j \in \Theta$  and then later take a limit as  $\theta_j \rightarrow \hat{\theta}_j$ . Combining the results

described above, we find

$$\begin{aligned}
& \log(1/\delta)[1 + o_\theta(1)] \\
& \leq \sum_{j \neq I^*(\theta_0)} q^*(\hat{\theta}_j) D_{\text{KL}}(\mathbb{P}(H_\tau^+ \in \cdot \mid \theta) \parallel \mathbb{P}(H_\tau^+ \in \cdot \mid \theta = \theta_j)) \\
& = \mathbb{E}[\tau \mid \theta] \sum_{j \neq I^*(\theta_0)} q^*(\hat{\theta}_j) \frac{1}{2\sigma^2} \sum_{i \in [k]} (\theta^{(i)} - \theta_j^{(i)})^\top \frac{\mathbb{E}[\sum_{t=1}^\tau \mathbb{1}(I_t = i) X_t X_t^\top \mid \theta]}{\mathbb{E}[\tau \mid \theta]} (\theta^{(i)} - \theta_j^{(i)}) \\
& \leq \mathbb{E}[\tau \mid \theta] \sup_{M \in \mathbb{M}} \sum_{j \neq I^*(\theta)} q^*(\hat{\theta}_j) \Gamma_\theta^{-1}(M, \theta_j).
\end{aligned}$$

The first inequality uses Corollary 4. Notice that it is critical that nature's equilibrium strategy randomizes over a *finite* set of vectors as this allows us to pass a limit through the sum<sup>13</sup>. The second inequality uses that matrices of the form  $M_i = \frac{\mathbb{E}[\sum_{t=1}^\tau \mathbb{1}(I_t = i) X_t X_t^\top \mid \theta = \theta_0]}{\mathbb{E}[\tau \mid \theta = \theta_0]}$  are a feasible choice for the experimenter.

Dividing each side by  $\log(1/\delta)$ , taking  $\delta \rightarrow 0$  and then taking  $\theta_j \rightarrow \hat{\theta}_j$  gives

$$\begin{aligned}
\liminf_{\delta \rightarrow 0} \frac{\mathbb{E}[\tau \mid \theta]}{\log(1/\delta)} & \geq \sup_{M \in \mathbb{M}} \sum_{j \neq I^*(\theta)} q^*(\hat{\theta}_j) \Gamma_\theta^{-1}(M, \theta_j) \longrightarrow \sup_{M \in \mathbb{M}} \sum_{j \neq I^*(\theta)} q^*(\hat{\theta}_j) \Gamma_\theta^{-1}(M, \hat{\theta}_j) \\
& = \sup_{M \in \mathbb{M}} \Gamma_\theta^{-1}(M, q^*) = \Gamma_\theta^{-1}.
\end{aligned}$$

**Remark 2** (Novelty in the lower bound proof). *There are two main points of novelty in this argument. The first is in the characterization of the experimenter's equilibrium choice of  $M$  in Proposition 5. This shows that context-independent sampling is asymptotically efficient. The second point of novelty is that Proposition 4 places a relatively weak constraint on the algorithm. The constraint (33) holds for each given  $\theta_0$ , but not uniformly.*

*The challenges this causes are subtle. Arguments like those in Garivier and Kaufmann [2016] seem to rely on a uniform constraint like  $\sup_{\theta_1 \in \Theta} \mathbb{P}(\hat{I}_\tau^+ \neq I^*(\theta_1) \mid \theta = \theta_1) \leq \delta$ . Each choice of  $\theta_1$  leads to a bound that is roughly of the form  $\mathbb{E}[\tau \mid \theta = \theta_0] \geq \Gamma_{\theta_0}^{-1}(M, \theta_1) \log(1/\delta)$ . Optimal results are derived by first maximizing over  $\theta_1$  and then minimizing over  $M$ . This argument breaks in our case, because (33) does not restrict  $\sup_{\theta_1 \in \Theta} \mathbb{E}[\Delta_\tau \mid \theta = \theta_1]$ , since the set  $\Theta$  is infinite. We avoid this problem by showing nature's worst-case behavior in Proposition 5 randomizes only over a finite number of states of nature. This is used at the very end of the proof.*

## G Proof of the Upper Bound in Proposition 3

### G.0.1 Outline of the proof.

The upper bound consists of two parts. A relatively straightforward part is given in Proposition 6, which shows that under stopping rule (25) with  $\delta = c$ ,  $\mathbb{E}[\Delta_\tau \mid \theta = \theta_0] \leq O_{\theta_0}(c) = o_{\theta_0}(c \log(1/c))$ . Establishing the upper bound then requires bounding the average sample size. Past work shows us how to bound the average sample size in problems without contexts. Arguments in Russo [2020] allow one to show formally that  $p_{t,i} \rightarrow p_i^*$  almost surely for each  $i \in [k]$ . If one makes the asymptotic approximation

$$Z_{t, I^*, j}^2 \approx t \frac{(\mu(\theta, I^*, w) - \mu(\theta, j, w))^2}{\sigma^2 \|X_{\text{pop}}\|_{\Lambda^{-1}}^2 \left( (p_{I^*}^*)^{-1} + (p_j^*)^{-1} \right)} = 2t \Gamma_\theta^{-1} \quad (41)$$

as in (18), then one can immediately bound  $\tau$ , the time required for all z-scores to cross a threshold that scales with  $\sqrt{2 \log(1/c)}$ , by a term of order  $\Gamma_\theta \log(1/c)$ . Almost sure convergence of  $p_{t,i}$  ensures (41) is eventually accurate for each sample path. As observed in Qin et al. [2017], however, to bound the *expected value* of a

<sup>13</sup>If there is a family of function  $f_i(\delta)$ , indexed by  $i \in I$ , with  $f_i = o(1)$  for each fixed  $i$ . This means  $\lim_{\delta \rightarrow 0} f_i(\delta) = 0$  for each given  $i \in I$ . When  $I$  is finite, it follows that  $\sum_{i \in I} f_i(\delta) = o(1)$  but this may not hold if  $I$  is infinite.

stopping time, we need to make sure the *expected time* elapsed before the approximation (41) is accurate is itself finite. For a different top-two sampling algorithm, Qin et al. [2017] does this by showing a slightly stronger notion of convergence: for each  $\epsilon > 0$  there is a (sample path dependent) time  $T$  with  $|p_{t,i} - p_i^*| \leq \epsilon$  for  $t \geq T$  and moreover  $\mathbb{E}[T] < \infty$ . Shang et al. [2020] extends this argument to apply to top-two Thompson sampling, i.e. to DTS in the setting without contexts.

To rigorously complete the proofs, we need to modify the argument in Russo [2020], Qin et al. [2017], and Shang et al. [2020]. The main difference is that we need to control for the randomness in the realized contexts. For instance, concentration inequalities are used to control for the finite sample error in an expression like equation (17). This introduces extra terms into nearly every expression in past proofs. To keep this paper concise and focused while ensuring all claims are reproducible, we have made the complete argument available in an separate technical report [Qin and Russo, 2022].

## G.1 Validity of the stopping time.

We begin with a result bounding the simple regret incurred when the stopping rule in (25) is used. In the statement,  $f(\theta, \delta) = \mathbb{E}[\Delta_\tau \mathbb{1}(\tau < \infty) \mid \theta]$  is viewed as some function of the state of nature and of the parameter  $\delta$  of the stopping time. The notation  $f(\theta, \delta) = O_\theta(\delta)$  means  $\limsup_{\delta \rightarrow 0} \frac{f(\theta, \delta)}{\delta} < \infty$ .

**Proposition 6.** *Let  $\tau$  denote the stopping time in (25). Then,*

$$\mathbb{E}[\Delta_\tau \mathbb{1}(\tau < \infty) \mid \theta] \leq O_\theta(\delta).$$

The proof is developed over the next few subsections.

### G.1.1 Latent reward and context tables

Because DTS is arm sampling under DTS is context independent (see (14)), we can analyze its behavior as if contexts are observed only *after* an arm is sampled. With this view, we have written down a model where, DTS selects an action  $I_t$  and then observes the context  $X_t$  and reward  $R_t = \langle X_t, \theta^{(I_t)} \rangle + W_t$ . An alternative description of the probability space would generate all the randomness upfront, using what [Lattimore and Szepesvári, 2020a, Section 4.6] call a *reward table* and a *context table*. Precisely, let  $\tilde{X}_{n,i} \sim p_X$  and  $\tilde{W}_{n,i} \sim N(0, \sigma^2)$  denote virtual contexts and rewards drawn i.i.d across  $n \in \mathbb{N}$  and  $i \in [k]$ . Assume  $(\tilde{X}_{n,i})_{n \in \mathbb{N}, i \in [k]}$ ,  $(\tilde{R}_{n,i})_{n \in \mathbb{N}, i \in [k]}$  and the algorithm’s internal randomness (used for arm sampling) are jointly independent. Let  $\tilde{R}_{n,i} = \langle \theta^{(i)}, \tilde{X}_{n,i} \rangle + \tilde{W}_{n,i}$  denote the reward corresponding to the  $n$ th play of arm  $i$ . Then for  $t \in \mathbb{N}$

$$(I_t, X_t, R_t) \stackrel{D}{=} (I_t, \tilde{X}_{N_{t,I_t}+1, I_t}, \tilde{R}_{N_{t,I_t}+1, I_t})$$

where  $\stackrel{D}{=}$  denotes equality in distribution (or “in law”).

We define  $\mathbb{N}_0 \triangleq \mathbb{N} \cup \{0\}$ . For  $(n, i) \in \mathbb{N}_0 \times [k]$ , the distribution of  $\theta^{(i)}$  conditioned on  $(\tilde{X}_{\ell,i})_{\ell \in [n]}$  is multivariate Gaussian with covariance and mean given by

$$\tilde{\Sigma}_{n,i} = \left( \tilde{\Sigma}_{0,i}^{-1} + \sigma^{-2} \sum_{\ell=1}^n \tilde{X}_{\ell,i} \tilde{X}_{\ell,i}^\top \right)^{-1} \quad \text{and} \quad \tilde{\mu}_{n,i} = \tilde{\Sigma}_{n,i} \left( \tilde{\Sigma}_{0,i}^{-1} \tilde{\mu}_{0,i} + \sigma^{-2} \sum_{\ell=1}^n \tilde{X}_{\ell,i} \tilde{R}_{\ell,i} \right) \quad (42)$$

where  $(\tilde{\mu}_{0,i}, \tilde{\Sigma}_{0,i})$  equals the prior mean and variance  $(\mu_{1,i}, \Sigma_{1,i})$  used by the algorithm.<sup>14</sup> Equation (42) mimics the updating rule in Equation (2), but is based on latent reward and context tables. Another difference is that  $(\tilde{\mu}_{n,i}, \tilde{\Sigma}_{n,i})$  in Equation (42) starts from  $n = 0$  (when the arm has not been sampled, the corresponding number of observations is 0), while  $(\mu_{t,i}, \Sigma_{t,i})$  in Equation (2) starts from time  $t = 1$  where  $(\mu_{1,i}, \Sigma_{1,i})$  is the prior mean and variance (when there is no observation seen so far).

<sup>14</sup>A side note is that here we overload the notations. In the definition of a context unaware algorithm in Section 5, we use  $\tilde{\mu}_{t,i}$  and  $\tilde{\Sigma}_{t,i}$  to denote an (incorrect) posterior mean and standard deviation at time  $t$ , while here  $\tilde{\mu}_{n,i}$  and  $\tilde{\Sigma}_{n,i}$  are the true posterior mean and standard deviation after having  $n$  samples based on the latent reward and context tables.

Since  $\mu(\theta, i, w)$  is a linear in  $\theta^{(i)}$ , it also has a Gaussian posterior given the latent reward and context tables. We write

$$\mu(\theta, i, w) \mid (\tilde{X}_{\ell,i}, \tilde{R}_{\ell,i})_{\ell \in [n]} \sim N(\tilde{m}_{n,i}, \tilde{s}_{n,i}^2)$$

where similar to those in Equation (3), we let

$$\tilde{s}_{n,i}^2 = X_{\text{pop}}^\top \tilde{\Sigma}_{n,i} X_{\text{pop}} \quad \text{and} \quad \tilde{m}_{n,i} = \langle X_{\text{pop}}, \tilde{\mu}_{n,i} \rangle. \quad (43)$$

### G.1.2 Frequentist distribution of the posterior mean

When conditioning on  $\theta$ , the posterior mean can be viewed as some function of the observed data. We characterize the sampling distribution of the error in the posterior mean. It is normally distributed. The variance of the sampling distribution is close to the posterior variance term  $\tilde{s}_{n,i}^2$  tracked by the algorithm: it is never larger (.. since the posterior calculation is factoring in believed variance in  $\theta$ , which is not present here) and the difference is a negligible fraction of  $\tilde{s}_{n,i}^2$  as  $n$  grows. The average error is nearly zero, but there is bias due to regularization toward the prior mean. The magnitude of bias is on the order of the  $\tilde{s}_{n,i}$ , implying it shrinks as evidence is gathered. We emphasize again that all analysis in this section is implicitly conditioned on  $\theta$ .

**Lemma 22.** For any  $n \in \mathbb{N}_0$  and  $i \in [k]$ ,

$$\tilde{m}_{n,i} - \mu(\theta, i, w) \mid \theta, \{\tilde{X}_{n',i'}\}_{(n',i') \in \mathbb{N} \times [k]} \sim N\left(\tilde{B}_{n,i}, \tilde{s}_{n,i}^2 - \|\tilde{\Sigma}_{n,i} X_{\text{pop}}\|_{\tilde{\Sigma}_{0,i}^{-1}}^2\right)$$

where the bias is given by

$$\tilde{B}_{n,i} \triangleq \left\langle X_{\text{pop}}, \tilde{\Sigma}_{n,i} \tilde{\Sigma}_{0,i}^{-1} \left(\tilde{\mu}_{0,i} - \theta^{(i)}\right) \right\rangle.$$

In addition, the following bounds hold:

$$\|\tilde{B}_{n,i}\| \leq \tilde{s}_{n,i} \|\tilde{\mu}_{0,i} - \theta^{(i)}\|_{\tilde{\Sigma}_{0,i}^{-1}} \quad \text{and} \quad \tilde{s}_{n,i}^2 \left[1 - \lambda_{\max}\left(\tilde{\Sigma}_{n,i}^{1/2} \tilde{\Sigma}_{0,i}^{-1} \tilde{\Sigma}_{n,i}^{1/2}\right)\right] \leq \tilde{s}_{n,i}^2 - \|\tilde{\Sigma}_{n,i} X_{\text{pop}}\|_{\tilde{\Sigma}_{0,i}^{-1}}^2 \leq \tilde{s}_{n,i}^2$$

*Proof.* Proof of Lemma 22. Fix  $(n, i) \in \mathbb{N}_0 \times [k]$ . We can write

$$\tilde{m}_{n,i} - \mu(\theta, i, w) - \tilde{B}_{n,i} = \left\langle X_{\text{pop}}, \left(\tilde{\mu}_{n,i} - \theta^{(i)}\right) - \tilde{\Sigma}_{n,i} \tilde{\Sigma}_{0,i}^{-1} \left(\tilde{\mu}_{0,i} - \theta^{(i)}\right) \right\rangle \triangleq P_{n,i}.$$

A little linear algebra shows that

$$\begin{aligned} \tilde{\Sigma}_{n,i}^{-1} \tilde{\mu}_{n,i} - \tilde{\Sigma}_{0,i}^{-1} \tilde{\mu}_{0,i} &= \sigma^{-2} \sum_{\ell=1}^n \tilde{X}_{\ell,i} \tilde{R}_{\ell,i} \\ &= \left( \sigma^{-2} \sum_{\ell=1}^n \tilde{X}_{\ell,i} \tilde{X}_{\ell,i}^\top \right) \theta^{(i)} + \sigma^{-2} \sum_{\ell=1}^n \tilde{X}_{\ell,i} \tilde{W}_{\ell,i} \\ &= \left( \tilde{\Sigma}_{n,i}^{-1} - \tilde{\Sigma}_{0,i}^{-1} \right) \theta^{(i)} + \sigma^{-2} \sum_{\ell=1}^n \tilde{X}_{\ell,i} \tilde{W}_{\ell,i}, \end{aligned}$$

which gives

$$\tilde{\Sigma}_{n,i}^{-1} \left(\tilde{\mu}_{n,i} - \theta^{(i)}\right) - \tilde{\Sigma}_{0,i}^{-1} \left(\tilde{\mu}_{0,i} - \theta^{(i)}\right) = \sigma^{-2} \sum_{\ell=1}^n \tilde{X}_{\ell,i} \tilde{W}_{\ell,i},$$

and thus

$$P_{n,i} = \sigma^{-2} \sum_{\ell=1}^n \left\langle X_{\text{pop}}, \tilde{\Sigma}_{n,i} \tilde{X}_{\ell,i} \right\rangle \tilde{W}_{\ell,i}.$$



Notice that

$$\begin{aligned}
\sum_{\ell=1}^n \langle X_{\text{pop}}, \tilde{\Sigma}_{n,i} \tilde{X}_{\ell,i} \rangle^2 &= X_{\text{pop}}^\top \tilde{\Sigma}_{n,i} \left( \sum_{\ell=1}^n \tilde{X}_{\ell,i} \tilde{X}_{\ell,i}^\top \right) \tilde{\Sigma}_{n,i} X_{\text{pop}} \\
&= \sigma^2 X_{\text{pop}}^\top \tilde{\Sigma}_{n,i} \left( \tilde{\Sigma}_{n,i}^{-1} - \tilde{\Sigma}_{0,i}^{-1} \right) \tilde{\Sigma}_{n,i} X_{\text{pop}} \\
&= \sigma^2 X_{\text{pop}}^\top \tilde{\Sigma}_{n,i} X_{\text{pop}} - \sigma^2 X_{\text{pop}}^\top \tilde{\Sigma}_{n,i} \left( \tilde{\Sigma}_{0,i}^{-1} \right) \tilde{\Sigma}_{n,i} X_{\text{pop}} \\
&= \sigma^2 \tilde{\varsigma}_{n,i}^2 - \sigma^2 \|\tilde{\Sigma}_{n,i} X_{\text{pop}}\|_{\tilde{\Sigma}_{0,i}^{-1}}^2.
\end{aligned}$$

which implies the first claim.

To bound the bias, we use

$$\begin{aligned}
\tilde{B}_{n,i} &= \left\langle X_{\text{pop}}, \tilde{\Sigma}_{n,i} \tilde{\Sigma}_{0,i}^{-1} \left( \tilde{\mu}_{0,i} - \theta^{(i)} \right) \right\rangle = \left\langle X_{\text{pop}}, \tilde{\Sigma}_{0,i}^{-1} \left( \tilde{\mu}_{0,i} - \theta^{(i)} \right) \right\rangle_{\tilde{\Sigma}_{n,i}} \\
&\leq \|X_{\text{pop}}\|_{\tilde{\Sigma}_{n,i}} \|\tilde{\Sigma}_{0,i}^{-1} \left( \tilde{\mu}_{0,i} - \theta^{(i)} \right)\|_{\tilde{\Sigma}_{0,i}} \\
&= \|X_{\text{pop}}\|_{\tilde{\Sigma}_{n,i}} \|\tilde{\mu}_{0,i} - \theta^{(i)}\|_{\tilde{\Sigma}_{0,i}^{-1}},
\end{aligned}$$

where we use the notation  $\langle x, y \rangle_A = x^\top A y$  for a positive definite matrix  $A$ , the first inequality is Cauchy-Schwartz, and the final equality can be seen by directly writing out the expressions for the two norms.

Finally, the offset term for the variance can be bounded as,

$$\|\tilde{\Sigma}_{n,i} X_{\text{pop}}\|_{\tilde{\Sigma}_{0,i}^{-1}}^2 = X_{\text{pop}}^\top \tilde{\Sigma}_{n,i} \tilde{\Sigma}_{0,i}^{-1} \tilde{\Sigma}_{n,i} X_{\text{pop}} = X_{\text{pop}}^\top \tilde{\Sigma}_{n,i}^{1/2} \left( \tilde{\Sigma}_{n,i}^{1/2} \tilde{\Sigma}_{0,i}^{-1} \tilde{\Sigma}_{n,i}^{1/2} \right) \tilde{\Sigma}_{n,i}^{1/2} X_{\text{pop}} \leq \lambda_{\max} \left( \tilde{\Sigma}_{n,i}^{1/2} \tilde{\Sigma}_{0,i}^{-1} \tilde{\Sigma}_{n,i}^{1/2} \right) \tilde{\varsigma}_{n,i}^2.$$

□

### G.1.3 Proof of Proposition 6

We now use Lemma 22 to prove Proposition 6.

*Proof.* Proof of Proposition 6. Let  $X = \{\tilde{X}_{n',i'}\}_{(n',i') \in \mathbb{N} \times [k]}$ . Our goal is to show that  $\mathbb{E} [\Delta_{\tau_\delta} \mathbb{1}(\tau_\delta < \infty) \mid \theta, X] = O_\theta(\delta)$ . This big-O notation is equivalent to the statement

$$\limsup_{\delta \rightarrow 0} \frac{\mathbb{E} [\Delta_{\tau_\delta} \mathbb{1}(\tau_\delta < \infty) \mid \theta, X]}{\delta} < \infty. \tag{44}$$

Following Subsubsection G.1.1, for each  $i, j \in [k]$  and  $n_i, n_j \in \mathbb{N}_0$ , we define the latent variable

$$\tilde{Z}_{n_i, n_j, i, j} = \frac{\tilde{m}_{n_i, i} - \tilde{m}_{n_j, j}}{\sqrt{\tilde{\varsigma}_{n_i, i}^2 + \tilde{\varsigma}_{n_j, j}^2}}.$$

This definition aligns with the z-scores used in the stopping rule (25). If arms  $i$  and  $j$  have been played  $n_i$

and  $n_j$  times by time  $t$ , then  $Z_{t,i,j} = \tilde{Z}_{n_i, n_j, i, j}$ . Then we have

$$\begin{aligned}
& \mathbb{E} [\Delta_{\tau_\delta} \mathbb{1}(\tau_\delta < \infty) \mid \theta, X] \\
&= \mathbb{E} [(\mu(\theta, I^*, w) - \mu(\theta, \hat{I}_{\tau_\delta}, w)) \mathbb{1}(\tau_\delta < \infty) \mid \theta, X] \\
&= \sum_{i \neq I^*} (\mu(\theta, I^*, w) - \mu(\theta, i, w)) \mathbb{P}(\tau_\delta < \infty, \hat{I}_{\tau_\delta} = i \mid \theta, X) \\
&\leq \sum_{i \neq I^*} (\mu(\theta, I^*, w) - \mu(\theta, i, w)) \mathbb{P}(\exists t \in \mathbb{N} : Z_{t,i,I^*} \geq \gamma_{t,\delta} + \sqrt{\gamma_{t,\delta}} \mid \theta, X) \\
&\leq \sum_{i \neq I^*} (\mu(\theta, I^*, w) - \mu(\theta, i, w)) \mathbb{P}(\exists t \in \mathbb{N} : Z_{t,i,I^*} \geq \gamma_{t,\delta} + \sqrt{\gamma_{t,\delta}} \mid \theta, X) \\
&\leq \sum_{i \neq I^*} (\mu(\theta, I^*, w) - \mu(\theta, i, w)) \mathbb{P}(\exists n_i, n_{I^*} \in \mathbb{N}_0 : \tilde{Z}_{n_i, n_{I^*}, i, I^*} \geq \gamma_{n_i+n_{I^*}+1,\delta} + \sqrt{\gamma_{n_i+n_{I^*}+1,\delta}} \mid \theta, X)
\end{aligned}$$

where in the final inequality, since  $n_i + n_{I^*} \leq t - 1$ , by monotonicity,  $\gamma_{n_i+n_{I^*}+1,\delta} \leq \gamma_{t,\delta}$ .

By Lemma 22 and the bounds therein, conditioned on  $\theta$  and  $X$ ,  $\tilde{Z}_{n_i, n_{I^*}, i, I^*}$  follows Gaussian distribution with  $\text{Var} [\tilde{Z}_{n_i, n_{I^*}, i, I^*} \mid \theta, X] \leq 1$  and

$$\begin{aligned}
\mathbb{E} [\tilde{Z}_{n_i, n_{I^*}, i, I^*} \mid \theta, X] &= \frac{\mu(\theta, i, w) - \mu(\theta, I^*, w)}{\sqrt{\tilde{s}_{n,i}^2 + \tilde{s}_{n,I^*}^2}} + \frac{B_{n,i} - B_{n,I^*}}{\sqrt{\tilde{s}_{n,i}^2 + \tilde{s}_{n,I^*}^2}} \\
&\leq \frac{\mu(\theta, i, w) - \mu(\theta, I^*, w)}{\sqrt{\tilde{s}_{n,i}^2 + \tilde{s}_{n,I^*}^2}} + \frac{|B_{n,i}| + |B_{n,I^*}|}{\sqrt{\tilde{s}_{n,i}^2 + \tilde{s}_{n,I^*}^2}} \\
&\leq \frac{\mu(\theta, i, w) - \mu(\theta, I^*, w)}{\sqrt{\tilde{s}_{n,i}^2 + \tilde{s}_{n,I^*}^2}} + \frac{\tilde{s}_{n,i} \|\tilde{\mu}_{0,i} - \theta^{(i)}\|_{\tilde{\Sigma}_{0,i}^{-1}} + \tilde{s}_{n,I^*} \|\tilde{\mu}_{0,I^*} - \theta^{(I^*)}\|_{\tilde{\Sigma}_{0,I^*}^{-1}}}{\sqrt{\tilde{s}_{n,i}^2 + \tilde{s}_{n,I^*}^2}} \\
&\leq \frac{\mu(\theta, i, w) - \mu(\theta, I^*, w)}{\sqrt{\tilde{s}_{n,I^*}^2 + \tilde{s}_{n,i}^2}} + \epsilon(\theta) \\
&\leq \epsilon(\theta),
\end{aligned}$$

where  $\epsilon(\theta) \triangleq 2 \max_{j \in [k]} \|\tilde{\mu}_{0,j} - \theta^{(j)}\|_{\tilde{\Sigma}_{0,j}^{-1}}$  in the second-to-last inequality, and we use the Cauchy-Schwarz inequality.

For any  $\theta$ , there is a  $\delta_\theta$  such that for every  $\delta \leq \delta_\theta$ , we have  $\epsilon(\theta) \leq \sqrt{\gamma_{1,\delta}}$ . Then, for  $\delta \leq \delta_\theta$  we have

$$\begin{aligned}
\mathbb{P}(\tilde{Z}_{n_i, n_{I^*}, i, I^*} \geq \gamma_{n_i+n_{I^*}+1,\delta} + \sqrt{\gamma_{n_i+n_{I^*}+1,\delta}} \mid \theta, X) &\leq \Phi(-\gamma_{n_i+n_{I^*}+1,\delta} - \sqrt{\gamma_{n_i+n_{I^*}+1,\delta}} + \epsilon(\theta)) \\
&\leq \Phi(-\gamma_{n_i+n_{I^*}+1,\delta} - \sqrt{\gamma_{1,\delta}} + \epsilon(\theta)) \\
&\leq \Phi(-\gamma_{n_i+n_{I^*}+1,\delta}) \\
&\leq \exp(-\gamma_{n_i+n_{I^*}+1,\delta}^2/2) \\
&= \frac{\delta}{(1+n_i+n_{I^*})^3}
\end{aligned}$$

where the first inequality follows from the monotonicity of  $\gamma_{t,\delta}$  in  $t$ , and the last inequality applies a common upper bound on the tail of the Gaussian CDF.

This shows that for  $\delta \leq \delta_\theta$ ,

$$\begin{aligned} \mathbb{E} [\Delta_{\tau_\delta} \mathbb{1}(\tau_\delta < \infty) \mid \theta] &= \mathbb{E} [\mathbb{E} [\Delta_{\tau_\delta} \mathbb{1}(\tau_\delta < \infty) \mid \theta, X] \mid \theta] \\ &\leq \sum_{i \neq I^*} (\mu(\theta, I^*, w) - \mu(\theta, i, w)) \left( \sum_{n_i=0}^{\infty} \sum_{n_{I^*}=0}^{\infty} \frac{\delta}{(1+n_i+n_{I^*})^3} \right), \end{aligned}$$

which implies (44). □

## G.2 Bounding the average sample size

Condition on  $\theta = \theta_0$  for some  $\theta_0 \in \Theta$ . Consider any sampling rule that is “ergodic” in the sense of (15). That is, for each  $i \in [k]$  there exists  $p_i = p_i(\theta_0) > 0$ , such that  $\lim_{t \rightarrow \infty} p_{t,i} = p_i$ . Equation (18) implies

$$\lim_{t \rightarrow \infty} \frac{Z_{t,i,j}^2}{t} = \frac{(\mu(\theta_0, w, i) - \mu(\theta_0, w, j))^2}{\sigma^2 \|X_{\text{pop}}\|_{\Lambda^{-1}}^2 (p_i^{-1} + p_j^{-1})},$$

for any  $i, j \in [k]$ . Recall that  $\hat{I}_t \in \arg \max_{i \in [k]} m_{t,i}$  is the empirical best-arm at time  $t$ . With probability 1, for sufficiently large  $t$  (i.e. for  $t$  larger than some sample path dependent time),  $\hat{I}_t = I^*(\theta_0)$  is the true optimal arm. So

$$\lim_{t \rightarrow \infty} \frac{Z_{t,\hat{I}_t,j}^2}{t} = \frac{(\mu(\theta_0, w, I^*(\theta_0)) - \mu(\theta_0, w, j))^2}{\sigma^2 \|X_{\text{pop}}\|_{\Lambda^{-1}}^2 (p_{I^*}^{-1} + p_j^{-1})}.$$

We gave proof sketches in Section 8 which suggest that under DTS,  $\lim_{t \rightarrow \infty} p_{t,i} = p_i^*(\theta_0)$  with probability 1. Assuming this holds, we have that for each  $j \in [k]$ , with probability 1,

$$\lim_{t \rightarrow \infty} \frac{Z_{t,\hat{I}_t,j}^2}{t} = \frac{(\mu(\theta_0, w, I^*(\theta_0)) - \mu(\theta_0, w, j))^2}{\sigma^2 \|X_{\text{pop}}\|_{\Lambda^{-1}}^2 [(p_{I^*}^*)^{-1} + (p_j^*)^{-1}]} = 2\Gamma_{\theta_0}^{-1},$$

where the second equality is the definition of the optimal allocation  $p^*(\theta_0)$  and optimal exponent  $\Gamma_{\theta_0}$  in (19).

Writing out the definition almost sure convergence explicitly, this means the following: for every  $\epsilon > 0$  there exists a random time  $T$  with  $\mathbb{P}(T < \infty \mid \theta = \theta_0) = 1$ , such that for every  $t \geq T$

$$\left| \frac{Z_{t,\hat{I}_t,j}^2}{t} - 2\Gamma_{\theta_0}^{-1} \right| \leq \epsilon \quad \text{for all } j \in [k]. \quad (45)$$

The following proposition strengthens this notion of convergence.

**Proposition 7.** *Suppose DTS is applied with  $\beta_t$  set by Algorithm 2 and condition on the event that  $\theta = \theta_0$  for some  $\theta_0 \in \Theta$ . For every  $\epsilon > 0$ , there exists a random time  $T$  with  $\mathbb{E}[T \mid \theta = \theta_0] < \infty$  such that (45) holds for each  $t \geq T$ .*

*Proof.* Notes on the proof of Proposition 7. First, suppose that there are no contexts. That is, assume  $X_t \in \mathbb{R}$  and  $X_t = 1$  for each  $t$  almost surely. In that case DTS is just the top-two Thompson sampling algorithm proposed in Russo [2020]. Argument like the ones in Russo [2020] imply almost sure convergence to the optimal proportions, i.e.  $\lim_{t \rightarrow \infty} p_{t,i} = p_i^*(\theta_0)$  with probability 1. This means there exists  $T$  with  $\mathbb{P}(T < \infty \mid \theta = \theta_0) = 1$  such that (45) holds for  $t \geq T$ . Qin et al. [2017] studies a different top-two sampling algorithm and shows one can ensure  $\mathbb{E}[T] < \infty$ . Shang et al. [2020] show this kind of result for top-two Thompson sampling.

The result above requires extending to the the case with contexts. We have already sketched in Proposition 2 why  $p_{t,i} \rightarrow p_i^*$  for each  $i \in [k]$ , if they converge at all. Because arm-sampling under DTS is context independent, the presence of contexts did not substantially influence the argument there. Indeed, the context

independence property in (14) means one can analyze DTS as if an arm is sampled and then the algorithm observes  $(X_t, R_t)$ . So a contextual extension requires controlling for the noise in contexts and not just noise in rewards.

Unfortunately, the the proofs of [Qin et al. \[2017\]](#) and [Shang et al. \[2020\]](#) are very long and detailed. Rewriting each line here, while adding in factors that deal with the noise in realized contexts, would add 25 pages to this paper and detract from our focus. For reproducibility, we have written a complete argument and make that available as a supplementary report.  $\square$

Given the convergence above, it is not hard to bound the average sample size under DTS. Notice that the condition that  $\mathbb{E}[T \mid \theta] < \infty$ , rather than usual condition that  $\mathbb{P}(T < \infty \mid \theta) = 1$ , plays a critical role in the proof.

**Proposition 8.** *Suppose DTS is applied with  $\beta_t$  set by 2. Let  $\tau$  denote the stopping time defined in (25). Then for any  $\theta_0 \in \Theta$ , as  $\delta \rightarrow 0$ ,*

$$\mathbb{E}[\tau \mid \theta = \theta_0] \leq [1 + o_{\theta_0}(1)] \Gamma_{\theta_0} \log(1/\delta).$$

*Proof.* Proof of Proposition 8. Fix  $\epsilon > 0$ . There exists  $T$  with  $\mathbb{E}[T \mid \theta = \theta_0] < \infty$  such that for  $t \geq T$ ,

$$\min_{i \neq \hat{I}_t} Z_{t, \hat{I}_t, i} \geq \sqrt{2t \left( \Gamma_{\theta_0}^{-1} - \epsilon \right)}.$$

The algorithm stops as soon as  $\min_{i \neq \hat{I}_t} Z_{t, \hat{I}_t, i} \geq \gamma_t$  and thus

$$\tau \leq \max\{T, t_\delta\} \leq T + t_\delta$$

where  $t_\delta$  is a deterministic time defined as

$$\begin{aligned} t_\delta &\triangleq \inf \left\{ t \in \mathbb{N} : \sqrt{2t \left( \Gamma_{\theta_0}^{-1} - \epsilon \right)} \geq \gamma_{t, \delta} + \sqrt{\gamma_{t, \delta}} \right\} \\ &= \inf \left\{ t \in \mathbb{N} : t \left( \Gamma_{\theta_0}^{-1} - \epsilon \right) \geq \log(1/\delta) + 3 \log(t) + \sqrt{\log(1/\delta) + 3 \log(t)} \right\}. \end{aligned}$$

The second equality just recalls the definition of  $\gamma_t$ . Taking expectations and dividing by  $\log(1/\delta)$  gives

$$\limsup_{\delta \rightarrow 0} \frac{\mathbb{E}[\tau \mid \theta = \theta_0]}{\log(1/\delta)} \leq \limsup_{\delta \rightarrow 0} \frac{\mathbb{E}[T \mid \theta = \theta_0]}{\log(1/\delta)} + \limsup_{\delta \rightarrow 0} \frac{t_\delta}{\log(1/\delta)} = \limsup_{\delta \rightarrow 0} \frac{t_\delta}{\log(1/\delta)} = \frac{1}{\Gamma_{\theta_0}^{-1} - \epsilon}.$$

Since the bound holds for arbitrary  $\epsilon$ , taking  $\epsilon \downarrow 0$  gives the result.  $\square$

### G.3 Completing the proof of the upper bound in Proposition 3

We note that the upper bound in Proposition 3 follows immediately from the results we have just established. In particular, we know that for  $\delta = c$ ,

$$\mathbb{E}[\tau \mid \theta = \theta_0] \leq [1 + o_{\theta_0}(1)] \Gamma_{\theta_0} \log(1/c).$$

In addition, when  $\delta = c$ ,

$$\mathbb{E}[\Delta_\tau \mid \theta = \theta_0] = \mathbb{E}[\Delta_\tau \mathbb{1}(\tau < \infty) \mid \theta = \theta_0] \leq O_{\theta_0}(c) = o_{\theta_0}(c \log(1/c)),$$

where the first equality holds since  $\mathbb{E}[\tau \mid \theta = \theta_0] < \infty$  ensures  $\tau$  is finite almost surely. Putting the two together gives the desired result:

$$\mathbb{E}[c\tau + \Delta_\tau \mid \theta = \theta_0] \leq [1 + o_{\theta_0}(1)] \Gamma_{\theta_0} c \log(1/c).$$