Adaptive Experimentation in the Presence of Exogenous Nonstationary Variation

Chao Qin and Daniel Russo
Columbia University
August 29, 2023

Abstract

We investigate experiments that are designed to select a treatment arm for population deployment. Multi-armed bandit algorithms can enhance efficiency by dynamically allocating measurement effort towards higher performing arms based on observed feedback. However, such dynamics can result in brittle behavior in the face of nonstationary exogenous factors influencing arms’ performance during the experiment. To counter this, we propose deconfounded Thompson sampling (DTS), a more robust variant of the prominent Thompson sampling algorithm. As observations accumulate, DTS projects the population-level performance of an arm while controlling for the context within which observed treatment decisions were made. Contexts here might capture a comprehensible source of variation, such as the country of a treated individual, or simply record the time of treatment. We provide bounds on both within-experiment and post-experiment regret of DTS, illustrating its resilience to exogenous variation and the delicate balance it strikes between exploration and exploitation. Our proofs leverage inverse propensity weights to analyze the evolution of the posterior distribution, a departure from established methods in the literature. Hinting that new understanding is indeed necessary, we show that a deconfounded variant of the popular upper confidence bound algorithm can fail completely.

1 Introduction

Multi-armed bandit (MAB) algorithms are crafted to enhance efficiency beyond what classical randomized controlled trials (RCTs) offer. In contrast to RCTs which maintain a fixed probability for assigning treatment arms throughout an experiment, MAB algorithms dynamically redistribute measurement effort towards higher performing arms based on observed feedback. Such strategies not only reduce experimental cost — since inferior arms are played less frequently – but variants of MAB algorithms can increase statistical power in identifying the most effective arm [Bubeck and Slivkins, 2012, Kaufmann et al., 2016, Russo, 2020]. These efficiency advantages have motivated widespread adoption of MAB algorithms for selecting and personalizing digital content.

Traditional application of MAB algorithms, as commonly advised in academic texts [Lattimore and Szepesvári, 2020] and industry-centric blogs [Amadio, 2020], presupposes that rewards linked with a particular arm selection are independently and identically distributed (i.i.d.) over time. However, this assumption often fails in real-world scenarios.

To elucidate, consider a hypothetical scenario. In this scenario, the audio streaming platform Spotify aims to optimize the shortcuts displayed in Figure 1. Imagine an experiment lasting one week. Each time period in the MAB model might correspond to a particular user who just opened the app, treatment arms might be slight alterations in the user interface of the shortcuts, and a positive ‘reward’ could signify a user

1Shortcuts are a real product feature that enables users to conveniently access their favorite or recently played content. The discussion, however, does not necessarily mirror the specifics of the product or the available data. This example is presented solely for illustrative purposes.
locating an item to listen to without navigating away from the home page. The i.i.d. assumption implies that a random sample of users who open the app on Monday morning will exhibit behavior similar to another random sample of users who open the app later in the week, such as on Friday evening. Yet, substantial variation in user behavior can occur over time.

RCTs are designed to be robust to exogenous variation like this. Since the probability of an arm being selected remains constant throughout the experiment, averaging the reward produced by an arm provides an unbiased estimate of the performance it would have yielded if it were deployed consistently to all users throughout the time period of the experiment. By varying arm selection probabilities over time, MAB algorithms lose this inherent resilience to nonstationary patterns. Of course, by abandoning adaptive arm selection, RCTs lose the efficiency advantages offered by MAB algorithms.

1.1 An overview of the paper

We propose a new model in which nonstationary exogenous factors influence treatment arms’ performance during an experiment. Adapting the prominent Thompson sampling algorithm [Thompson, 1933] to this model yields a new, more robust, variant which we call deconfounded Thompson sampling. We illustrate the algorithm’s performance through simulations and conduct substantive theoretical analysis. To help the reader digest the full paper, this section provides an abbreviated overview of our model, proposed algorithm, and results.

1.1.1 A new model of bandit experiments

Among a set of \( k \) predefined treatment arms, indexed as \( [k] \triangleq \{1, \ldots, k\} \), a decision-maker (DM) aims to select an arm \( I_{\text{post}} \in [k] \) to deploy to the population at the end of the experiment. The experiment proceeds sequentially across \( T \) rounds, thought of as representing interactions with distinct individuals or ‘users.’ In each round \( t \in [T] \), the DM selects a treatment arm \( I_t \in [k] \) and observes a noisy reward \( R_{t,I_t} \in \mathbb{R} \) signaling the quality of the outcome. The DM also observes a vector of exogenous factors \( X_t \in \mathbb{R}^d \) which influence rewards; these might encode things like features of the user, the weather, or timing of the interaction. Temporal patterns in factors \( X_1, \ldots, X_T \) drive temporal patterns in rewards.

In keeping with the tradition of the literature, we call these exogenous factors “contexts”. However, unlike contextual bandit models [Li et al., 2010] in which contexts are used to segment or personalize decision-rules, here they are used to control for exogenous sources of variation in experiments that seek to deploy a single treatment arm to the population. This reflects common experimental practice. Consider the representative example displayed in Figure 1, where the goal is to establish a consistent user interface rather than one that undergoes erratic changes as a user’s context (e.g. the time of day, their recent interactions) changes. Refer to Appendix D for a more substantive discussion and a generalization of our formulation that accommodates personalization.

A Bayesian linear model allows the DM to draw inferences about the population-level reward an arm generates as observations are gathered. The mean reward signal of arm \( i \in [k] \) in context \( x \in \mathbb{R}^d \) is \( r_{\theta}(i, x) = \langle \theta(i), x \rangle \), where the parameter vector \( \theta = (\theta(1), \ldots, \theta(k)) \in \mathbb{R}^{k \cdot d} \) is drawn from a multi-variate Gaussian prior, denoted \( \theta \sim N(\mu_1, \Sigma_1) \) where \( \mu_1 \in \mathbb{R}^{k \cdot d} \) and \( \Sigma_1 \in \mathbb{R}^{kd \times kd} \). The reward realized at time \( t \) is

\[
R_{t,I_t} = r_{\theta}(I_t, X_t) + W_{t,I_t},
\]

where \( W_{t,I_t} \sim N(0, \sigma_t^2) \) is independent Gaussian noise. In modeling reward noise as independent, we are implicitly assuming that any exogenous nonstationarity is “explained” by the contexts. The quality of the
deployment arm $l_{\text{post}}$ is assessed through its population-level reward $r_\theta(l_{\text{post}})$. We model the population-level reward of an arm,
\[
r_\theta(i) = \mathbb{E}_{x \sim D_{\text{pop}}} [r_\theta(i, x)] = \langle \theta(i), x_{\text{pop}} \rangle \quad \text{where} \quad x_{\text{pop}} = \mathbb{E}_{x \sim D_{\text{pop}}} [x],
\]
as the average reward over contexts drawn from a pre-defined population distribution $D_{\text{pop}}$. Because the decision-maker knows $x_{\text{pop}}$ and the contexts are observable, standard calculations allow one to compute the (multi-variate Gaussian) posterior distribution of the population-level rewards $(r_\theta(1), \ldots, r_\theta(k))$ as observations accumulate.

Why model $x_{\text{pop}}$ as known to the DM? Continuing the example in Figure 1, we imagine the company might form an empirical population distribution by subsampling from the features of users who visit the home page over e.g. the month prior to the experiment. Controlled experiments are usually conducted to evaluate differences in how treatment arms perform; using them to estimate passively observable quantities is wasteful.

1.1.2 Models of contextual variation subsume other models of nonstationarity

This turns out to be a surprisingly rich modeling framework. The term ‘context’ evokes a comprehensible source of exogenous variation. However, as illustrated in the next example, one can also model bandit experiments with nonstationary rewards whose pattern, seemingly, cannot be explained by any observable factor. We treat this as a special case of our formulation by taking the time period at which an arm was selected to be an observable context.

**Example 1** (Modeling latent exogenous variation with contexts). Take $d = T$ and assume $X_{1:T}$ is deterministic with the $t$th context equal to the $t$th standard basis vector: $X_t = e_t \in \mathbb{R}^T$. Let $D_{\text{pop}}$ be the uniform distribution over $\{e_1, \ldots, e_T\}$. In this setting, the reward at time $t$, $R_{t, I_t} = \theta_1^{(i_t)} + W_{t, I_t}$ is a noisy sample of $\theta_1^{(i_t)}$ and the experimenter’s goal is to select the arm
\[
I^* \in \arg \max_{i \in [k]} r_\theta(i) \quad \text{where} \quad r_\theta(i) = \frac{1}{T} \sum_{t=1}^{T} \theta_1^{(i, t)}, \quad (2)
\]
which has highest average reward throughout the experiment.$^2$

The prior $\theta \sim N(\mu_1, \Sigma_1)$ allows the decision-maker to draw inferences based on observations so far. Note that a vanilla bandit experiment is an extreme, degenerate, special case, where the rank of $\Sigma_1$ is $k$ and $\theta_1^{(i)} = \cdots = \theta_T^{(i)}$ almost surely. Figure 2 represents a structured prior on $\theta$ that allows the decision-maker to guard against certain nonstationarity patterns while still allowing them to use past observations to forecast arms’ relative performance.

Example 4, presented in the appendix, illustrates a setting in which contexts represent more comprehensible sources of variation. There, a context indicates a user’s country of some app. We assume this is an observable user feature, so the platform can calculate population average weights $x_{\text{pop}}$ by looking up the mix of countries among users who opened the app over a long period prior to the start of the experiment. Notice that, due to timezone differences, the mix of countries among users arriving during a particular hour within the experiment may not reflect the population proportions. In our model, the DM can ‘control for’ this source of exogenous variation, which might otherwise confound their inferences.

1.1.3 A new algorithm: deconfounded Thompson sampling

We propose deconfounded Thompson sampling (DTS). It is a more robust variant of the Thompson sampling (TS) algorithm, which is popular in both academic and industrial contexts [Chapelle and Li, 2011, Scott, 2010, \footnote{This objective is implicit in the way that average treatment effects are estimated in A/B tests, and we choose to mimic this standard practice in Example 1. The rationale behind this practice is subtle, however. What does it mean to optimize a backward-looking objective when nonstationarity is a concern? Our partial answer is that the objective in (2) reflects a belief that an arm that outperformed others over a substantial time-span, like a couple of weeks, is likely to continue its strong performance. This belief is consistent with concerns about nonstationarity in other forms, like exogenous time trends that shift all arm’s mean rewards (see Figure 2) or more cyclic patterns, where the performance of an arm depends on the time of day.}].

\[\]
Figure 2: Two draws of \( \{\theta_t^{(i)}\}_{t\in[1000],i\in[2]} \) in a special case of Example 1 in which parameters follow the latent variable model \( \theta_t^{(i)} = \theta_0^{(i)} + \epsilon_t \). The exogenous process \( \{\epsilon_i\} \) has correlation \( \text{corr}(\epsilon_t, \epsilon_{t'}) = \exp\{-|t - t'|/\kappa\} \). The horizontal lines denote time averages and the vertical distance between them is \( |r_0(1) - r_0(2)| \).

As observations are gathered, it projects the population-level performance of an arm while controlling for the contexts in which past decisions are made. The probability it selects an arm in a given period during the experiment corresponds to the posterior probability of that arm being optimal for population deployment.

To define DTS precisely, observe that the optimal deployment arm \( I^* = \arg\max_{i\in[k]} \theta_0(i) \) is a random variable, due to its dependence on the uncertain parameter \( \theta \). At any time period \( t \) within the experiment, DTS randomly samples an arm \( I_t \) to measure with sampling probabilities

\[
P(I_t = i \mid H_t) = P(I^* = i \mid H_t),
\]

where \( H_t \) is the full history of rewards and contexts observable so far. Sampling probabilities do not depend on the current context. As with standard TS, there is a very simple way to implement this sampling step; Algorithm 1, presented in Section 3, calculates the posterior mean \( \mu_t \) and covariance \( \Sigma_t \) of \( \theta \), samples \( \hat{\theta} \sim N(\mu_t, \Sigma_t) \) and picks \( I_t \in \arg\max_{i\in[k]} r_{\hat{\theta}}(i) \). At the end of the experiment, DTS selects the arm \( I_{\text{post}} \in \arg\max_{i\in[k]} E[r_{\hat{\theta}}(i) \mid H_{\text{post}}] \), where \( H_{\text{post}} \) consists of all observations available at the end of the experiment.

1.1.4 Numerical illustration: a teaser

Figure 3 is a teaser of a numerical illustration of DTS we provide in Section 5. It simulates bandit algorithms applied to a hypothetical week-long experiment, conducted to select an arm to deploy across future weeks. Day-of-week effects influence reward observations during the experiment. The experiment involves 700 time periods (representing distinct users); the first 100 time periods occur during the context ‘Monday’, the next 100 occur during ‘Tuesday’ and so on. Focusing solely on DTS, Figure 3 captures a delicate balance it strikes between exploration and exploitation. By the end of the week, it has explored enough to deploy a near-optimal arm, reflected in its low post-experiment regret \( E[r_{\hat{\theta}}(I^*) - r_{\hat{\theta}}(I_{\text{post}})] \). But it reduces the cost of experimentation by redistributing measurement effort to higher performing arms during the experiment, reflected in its low cumulative within-experiment regret \( E\left[ \sum_{t=1}^{\ell} (r_{\hat{\theta}}(I^*) - r_{\hat{\theta}}(I_t)) \right] \). Section 5 also plots two related performance metrics.

The algorithm labeled ‘round-robin’, is an algorithm that operates like an RCT, sampling arms uniformly throughout the experiment. The algorithm labeled ‘sequential elimination’ brings round-robin closer to DTS. It removes arms from consideration if their posterior probability of being optimal drops below a small threshold. Round robin, sequential elimination, and DTS all appear to be effective at deploying a (nearly) optimal treatment arm at the end of the experiment. The primary advantage of DTS is its ability to reduce
regret incurred during the experiment.

Other natural bandit algorithms fare poorly in the experiment, failing to gather the information required to select a good arm by the end of the experiment. One of these is ‘context unaware TS’ — a standard implementation of TS which (incorrectly) assumes rewards are i.i.d. Another is deconfounded UCB, a variant of the upper confidence bound algorithms which dominate much of the literature on exploration in multi-armed bandit problems and reinforcement learning [Auer et al., 2002a, 2008, Rusmevichientong and Tsitsiklis, 2010]. Whereas DTS chooses an arm by maximizing a posterior sample from arms’s population-level variance, deconfounded UCB maximizes and upper confidence bound on the same quantity. Appendix E provides formal counterexamples for these two algorithms and also for a third — a variant of Thompson sampling used for contextual linear bandit problems.

1.1.5 Theoretical analysis

DTS is a relatively straightforward adaptation of Thompson sampling to our model. Perhaps surprisingly, rigorously understanding its performance required us to develop a completely original approach to analyzing bandit algorithms. Our proofs, distinct from others in the literature, use inverse propensity weights to analyze the evolution of the posterior distribution. The theorem statement below is also distinctive — depending on what we call “attainable precision” rather the length of time horizon. At a high-level, the challenge is that learning dynamics in our model markedly deviate from those in i.i.d. bandit models. Unlike i.i.d. environments, where the DM can choose to rapidly resolve uncertainty through exploration, our model introduces an unavoidable delay in this process as the DM awaits the occurrence of relevant contexts.

When specialized to models with i.i.d. rewards (i.e. no contextual variation), DTS is just standard TS and the theorem below provides per-period regret bounds on the order of $\sigma\sqrt{k/t}$, recovering standard results in the literature. More generally, the bound depends on what we call attainable precision — the inverse posterior variance of an arm’s population level variance assuming the DM chose to exclusively measure that arm in all contexts that have occurred so far. Precision measures how much uncertainty is resolvable if the DM explored as aggressively as possible. A subtle element of this result is that DTS does not explore as aggressively as possible: instead the regret bound in (3) and its performance in Figure 3 suggest it aggressively ‘exploits’ past observations to select good arms.
Theorem (Informal version of our main result when there is no observation delay). Define

\[ \text{Precision}(X_{1:t}) \triangleq \min_{i \in [k]} \frac{1}{\text{Var}(r_{\theta}(i) \mid (X_{1}, R_{1}, \ldots, X_{t}, R_{t})))}. \]

Fix any context sequence \( x_{1:T} = (x_{1}, \ldots, x_{T}) \) with \( \|x_{t}\|_2 \leq 1 \). Under DTS, within-experiment regret at any time \( t \in [T] \) is bounded as

\[ \mathbb{E}[r_{\theta}(I^*) - r_{\theta}(I_t) \mid X_{1:t} = x_{1:T}] \leq \tilde{O}\left( \sqrt{\frac{k}{\text{Precision}(x_{1:(t-1)})}} \right) \]

and post-experiment regret is bounded as

\[ \mathbb{E}[r_{\theta}(I^*) - r_{\theta}(I_{\text{post}}) \mid X_{1:T} = x_{1:T}] \leq \tilde{O}\left( \frac{k}{\text{Precision}(x_{1:T})} \right). \]

Our full theoretical results extend the above theorem in substantial ways. First, they accommodate settings in which the DM only observes rewards within the experiment after some delay. Second, the appendix provides additional results that are more similar to past literature: Proposition 3 bounds what we term the total “within-experiment contextual regret” of DTS and Proposition 4 extends the bound to a extension of DTS that aims to learn personalized policies.

1.2 Connections to the literature

Learning with resilience to exogenous nonstationarity. Two approaches, Thompson sampling and upper confidence bound algorithms, dominate much of the literature on multi-armed bandit algorithms. However, we are not aware of any previous papers examining their ability to identify an effective treatment arm despite exogenous nonstationary variation.

A large literature on nonstochastic bandit problems considers a related goal: they aim to design procedures that earn rewards within the experiment which are competitive with that of the best stable decision, even when reward sequences are not i.i.d. This literature was launched by Auer et al. [2002b] and is reviewed in Lattimore and Szepesvári [2020]. Our work is a substantial departure, making precise comparisons difficult. Our Bayesian model emphasizes the role of contextual variation as a driver of nonstationarity and prioritizes the quality of post-experiment decision-making. The nonstochastic MAB literature instead assumes rewards are picked by an intelligent adaptive adversary and aims to attain low within-experiment regret despite this fact. A few papers [Abbasi-Yadkori et al., 2018, Jamieson and Talwalkar, 2016] which study the problem of nonstochastic best-arm selection are more similar in (implicitly) considering post-experiment performance, but still differ in how nonstationarity is modeled. Algorithm design in the adversarial bandit literature is usually tightly coupled to worst-case theoretical bounds, typically resulting in algorithms which are much more conservative than Thompson sampling. Appendix C provides a more precise discussion of nonstochastic bandit models.

A related paper by Farias et al. [2022] was posted online concurrently with our paper. Their model is most similar to Example 1 and the reward-model in Figure 2, in that an exogenous time trend additively shifts all arm’s rewards. 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 results 1 has (essentially) no dependence on the dimension of the context space, suggesting that our algorithms use contexts to deconfound with minimal cost. It is an open question whether such guarantees are possible in the setting of Farias et al. [2022].

Adapting decisions to respond to exogenous variation. Our focus on reaching a stable decision despite non i.i.d. exogenous variation distinguishes this work from most of the literature on decision-making in nonstationary environments. Works like Mellor and Shapiro [2013], Besbes et al. [2015], Cheung et al. [2019],
Trovo et al. [2020], Abbasi-Yadkori et al. [2022] and Suk and Kpotufe [2022] focus on adapting decision-making rules as the environment evolves. These papers may provide a natural model for a recommendation system where items, which represent the arms, may lose relevance over time, requiring an adaptable system. Our model is particularly well-suited to scenarios such as the A/B testing problem described previously.

Similarly, the focus on reaching a stable decision distinguishes our work from a large literature which emphasizes how decision-making can respond to evolving context. For instance, in standard linear contextual bandit models [Li et al., 2010], the DM aims to converge on a decision-rule mapping contexts to actions that maximizes expected reward accrued in each specific context. In our model, we use contextual observations to draw reliable inferences from past reward observations, rather than as input for a context-reactive decision rule. Although our model reflects a common experimental practice, it is a departure from much of the bandit literature. As a result, we provide a thorough discussion in Appendix D. That section includes a generalization of DTS for learning personalized decision-rules. Appendix E establishes that, without modification, contextual bandit algorithms can fail for our objective.

**Within-experiment and post-experiment decision quality.** Our paper is somewhat atypical in considering both within-experiment and post-experiment decision quality. In one of the the most classical formulations of a multi-armed bandit, due to Lai and Robbins [1985] one aims to minimize exploration costs while, in the long-run, almost always choosing an optimal action. There is no notion of post-experiment decisions, and the sole performance measure is what we call within-experiment regret. Another segment of the literature, focuses solely on post-experiment performance. Papers in this literature go by a variety of names, including pure-exploration in MABs [Bubeck et al., 2009], best-arm identification [Kaufmann et al., 2016], or ranking and selection [Kim and Nelson, 2006].

In models with i.i.d. reward observations, what we call post-experiment regret is widely studied. It is often called “simple regret” [Bubeck et al., 2009] or “expected opportunity cost” [Frazier et al., 2008]. These differ from another common performance metric, which considers only the probability a suboptimal arm is selected, because it more severely penalizes selection of very low quality arms. Studying combined objectives is quite natural. See the rich decision-theoretic model of clinical trials in Chick et al. [2021], for example. Rather than combine within-experiment and post-experiment regret into a single coherent objective function, we treat DTS as a heuristic that does not perfectly optimize any goal. We study its performance according to both regret measures. Other papers that study both within- and post-experiment decision quality include Degenne et al. [2019], Caria et al. [2020], Athey et al. [2022], Krishnamurthy et al. [2023] and Zhong et al. [2023].

**Learning with resilience to delayed reward observations.** In cases with no contextual variation, DTS corresponds to standard Thompson sampling. Even then, Theorem 1 is notable in providing guarantees when reward observations are subject to delay. Our bound on post-experiment regret in the second part of Theorem 1 permits delay in observing rewards as long as the experiment itself, paralleling a situation where all arm pulls must be pre-determined at the experiment’s outset. Previous work by Kandasamy et al. [2018] provided guarantees for a Thompson sampling which allocates a batch of arm selections at once; however, their performance guarantees degrade with increasing batch size. A related preprint by Wu and Wager [2022] was posted online concurrently with our paper, showing that vanilla Thompson sampling outperforms many algorithms designed specifically to address problems with delayed rewards. The first part of Theorem 1, which bounds within-experiment regret, is different from and complementary to their theoretical bounds. Beyond results on Thompson sampling, a number of MAB papers establish theoretical bounds on regret when observations are subject to delay. See for instance Dudík et al. [2011], Joulani et al. [2013], Zhou et al. [2019] and references therein.
2 Formal problem formulation allowing for observation delay

We provide a complete problem formulation that is more formal than the one contained in Subsection 1.1.1. One substantive generalization is that we allow for a reward observation delay of \( L \geq 1 \) periods. This means that the arm selection at time \( t \) must be based on rewards associated with arms played more than \( L \) periods earlier, which constrains adaptivity within an experiment. Nevertheless, the post-experiment arm deployment decision can still incorporate the full experiment results \((I_1, R_{1,i_1}, \ldots, I_T, R_{T,i_f})\). That is, we imagine that the DM waits for reward realizations before population deployment.

Discussion of modeling choices. The presentation here is mathematically precise but omits discussion of subtle modeling choices. Some of these modeling choices were already discussed briefly in Subsection 1.1.1. The first sections of the appendix provide more detailed comparisons to the literature; see Appendix D for a discussion of connections to contextual bandit models and Appendix C for a discussion of adversarial nonstationary bandit models. The reader may choose to skip to those section after reading the formulation, or may proceed directly to the main results. To understand the flexibility of this abstract modeling framework, one might look to various examples we present; see Example 1 in Section 1, Example 2 in Section 5 and Examples 4 and 5 in Appendix A.

Mathematical notation. For an integer \( k \), we write \([k] = \{1, \ldots, k\}\). For a sequence \( x_1, x_2, \ldots \), we use the “Matlab style” indexing notation \( x_{n:} = (x_n, \ldots, x_H)\) to refer to sub-sequences. All vectors in this paper are viewed as column vectors. We use \( \langle x, y \rangle = x^\top y \) to denote the standard inner product between two vectors. For three random variables \( X, Y \) and \( Z \), the notation \( X \perp Y | Z \) means that \( X \) and \( Y \) are independent and \( X \perp Y | Z \) means they are independent conditioned on \( Z \).

Our model. The DM would like to deploy the utilitarian optimal arm \( I^* = \arg \max_{i \in [k]} r_\theta(i) \) in the population, where \( r_\theta(i) \) denotes the population average reward of arm \( i \). We model the population average reward as the average over heterogeneous conditional average rewards among contexts drawn from some population distribution:

\[
    r_\theta(i) = \mathbb{E}_{x \sim \mathcal{D}_{\text{pop}}} [r_\theta(i, x)] = \langle \theta(i), x_{\text{pop}} \rangle, \tag{4}
\]

where \( \mathcal{D}_{\text{pop}} \) is a distribution over \( d \) dimensional context vectors, \( x_{\text{pop}} = \mathbb{E}_{x \sim \mathcal{D}_{\text{pop}}} [x] \in \mathbb{R}^d \) is the mean context vector, \( r_\theta(i, x) = \langle \theta(i), x \rangle \) is a linear model governing how mean-rewards vary across contexts. As discussed in the introduction, we assume that the DM knows \( x_{\text{pop}} \). But the DM is uncertain about the parameter \( \theta = (\theta(1), \ldots, \theta(k)) \in \mathbb{R}^{kd} \), and knows only that it is drawn from a multivariate Gaussian prior, denoted \( \theta \sim N(\mu_1, \Sigma_1) \) where \( \mu_1 \in \mathbb{R}^{kd} \) and \( \Sigma_1 \in \mathbb{R}^{kd \times kd} \).

To resolve uncertainty, the DM conducts a \( T \) period experiment. In any period \( t \in [T] \) during the experiment, the DM observes a context \( X_t \in \mathbb{R}^d \) and chooses an arm \( I_t \in [k] \). After a delay of \( L \geq 1 \) periods, they observe a reward \( R_{t,i} \) associated with the selected arm. Formally, the potential reward of arm \( i \) at time \( t \) is

\[
    R_{t,i} = r_\theta(i, X_t) + W_{t,i}, \tag{5}
\]

where \( W_{t,i} \sim N(0, \sigma^2) \) is i.i.d. Gaussian noise that is assumed to be jointly independent of \( \theta \), the contexts \( X_{1:t} \) and the decisions \( I_{1:t} \).

The sequence of contexts \( X_{1:T} \) within the experiment is drawn from a distribution \( \mathcal{D}_{\text{exp}} \) over \( \mathcal{X}^T \), where \( \mathcal{X} \subset \{ x \in \mathbb{R}^d : \| x \|_2 \leq 1 \} \) is a subset of context vectors with bounded norm; an important special case is where \( \mathcal{D}_{\text{exp}} \) is a point mass on a particular sequence \( x_{1:T} \). The algorithms that we study do not require prior knowledge of \( \mathcal{D}_{\text{exp}} \). We assume the draw of \( X_{1:T} \) is independent of \( \theta \), so that the DM cannot resolve their uncertainty by passively observing contexts, and assume that \( X_{(t+1):T} \perp I_{1:t} | X_{1:t} \), so that the DM cannot purposefully influence future contexts through their arm selection.

The DM employs a policy \( \pi = (\pi_1, \ldots, \pi_T, \pi_{\text{post}}) \). To treat randomized policies, we will allow the policy to take as input random seeds \((\xi_1, \ldots, \xi_T, \xi_{\text{post}})\), which are drawn i.i.d., and are independent from the context.
sequence, potential rewards, and \( \theta \). For a period \( t \in [T] \) within the experiment, \( \pi_t \) determines an arm to sample as
\[
I_t = \pi_t( H_t, X_t, \xi_t ) \quad \text{where} \quad H_t \triangleq \left( X_{1:(t-1)}, I_{1:(t-1)}, R_{1:(t-1)} \right);
\]
for notational convenience, we write \( R_\ell \triangleq R_\ell, I_\ell \) for \( \ell \in [T] \) and define \( R_{1:(t-1)} = \emptyset \) for any \( t \in [L] \). At the end of the experiment, the post-experiment decision-rule \( \pi_{\text{post}} \) selects an to deploy in the population as
\[
I_{\text{post}} = \pi_{\text{post}}( H_{\text{post}}, \xi_{\text{post}} ) \quad \text{where} \quad H_{\text{post}} \triangleq ( X_{1:T}, I_{1:T}, R_{1:T} ).
\]

We consider two measures of the performance of an algorithm \( \pi_t \):

**Expected post-experiment (utilitarian) regret:** \( \mathbb{E}[\Delta_{\text{post}}] \) where \( \Delta_{\text{post}} \triangleq r_\theta(I^*) - r_\theta(I_{\text{post}}) \).

**Expected within-experiment (utilitarian) regret:** \( \mathbb{E}[\Delta_t] \) where \( \Delta_t \triangleq r_\theta(I^*) - r_\theta(I_t) \).

The term ‘utilitarian’ is taken from Athey and Wager [2021] and reflects that an arm’s performance is measured in terms of its average reward or ‘utility’ it generates within a population. Appendix B studies a measure of regret on the contexts encountered within the experiment.

Post-experiment regret measures whether the policy is able to choose an arm with near-optimal population average reward at the end of the experiment. Within-experiment (utilitarian) regret captures whether a decision made within the experiment has near-optimal population average reward. Attaining low within-experiment regret indicates that the DM was able to select arms of similar quality within the experiment to the arm they hoped to employ post-experiment (with ‘similarity’ assessed through \( r_\theta(\cdot) \)); This can be thought of as an indicating a reduced cost of experimentation.

We make a couple of extra assumptions to simplify the presentation. First, we assume \( x_{\text{pop}} \neq 0 \). Otherwise, it is known at the beginning of the experiment that each arm’s population average reward is zero. Next, we assume that \( \Sigma_1 \) has full rank (although some eigenvalues could be arbitrarily small). This allows us to write some expressions in terms of matrix inverses. Together, the two assumptions imply that, with probability 1, there is a unique solution to the maximization problem defining \( I^* \) and we do not need to discuss tie-breaking rules. For similar reasons, assume \( \sigma > 0 \).

### 2.1 Remarks on interpretation

**Remark 1** (A unified objective function). Roughly speaking, we interpret attaining low post-experiment regret as a constraint on the policies a decision-maker could employ. In practice, an experimenter is unlikely to knowingly choose a policy that is incapable of deploying an effective arm to the population. Subject to this (loosely defined) constraint, we seek an policy that reduces experimentation costs by minimizing within-experiment regret. This kind of evaluation is clearest in the the interpretation of the experiment results in Section 5. There, some algorithms are effectively disqualified due to suffering high post-experiment regret. Several algorithms attain comparable post-experiment regret, but nevertheless differ substantially in the regret they incur within the experiment.

**Remark 2** (Connection to i.i.d. bandits). Classical bandit models with i.i.d. reward observations are a special case of the model in which there is no variation in contextual observations. Specifically, take contexts to be one dimensional \( (d = 1) \), and assume that \( X_1 = X_2 = \cdots = X_T = x_{\text{pop}} \). Then potential reward observations \( (r_1, \ldots, r_T) \) are i.i.d. samples with mean \( r_\theta(i) \). In this special case, post-experiment regret is often called “simple regret” [Bubeck et al., 2009]. We specialize our results to this case in Corollary 1.

**Remark 3** (Interpretation of post-experiment regret). Our discussion implicitly imagines that treatment decisions continue after the end of the experiment. Here, we make that explicit. Extend the time horizon by \( N \) periods; The DM
chooses arm $I_t = I_{\text{post}}$ for all $t \in \{T + 1, \ldots, T + N\}$. Then, the reward earned post experiment is:

$$
\sum_{t=T+1}^{T+N} r_{\theta}(I_t, X_t) = \sum_{t=T+1}^{T+N} r_{\theta}(I_{\text{post}}, X_t) = \sum_{t=T+1}^{T+N} \langle \theta(I_{\text{post}}), X_t \rangle = N \cdot r_{\theta}(I_{\text{post}}) \approx N \cdot r_{\theta}(I_{\text{post}})
$$

where $\hat{X}_{\text{pop}} = \frac{1}{N} \sum_{t=T+1}^{T+N} X_t$ is the average post-experiment context and the final approximate equality holds when $\hat{X}_{\text{pop}} \approx x_{\text{pop}}$. The approximate equality is exact if $\hat{X}_{\text{pop}} = x_{\text{pop}}$. What we call post-experiment utilitarian regret is the per-period expected regret of the post-experiment decision under a post-experiment context distribution whose mean matches the DM’s target context weights $x_{\text{pop}}$.

**Remark 4** (Comparison to the probability of incorrect selection). One might also be interested in comparing post-experiment regret to the probability of incorrect selection, $P(I^* \neq I_{\text{post}})$, a metric that is widely studied in the literature. We can write

$$
P(I_{\text{post}} \neq I^*) = \mathbb{E} \left[ \sum_{i \neq I^*} 1(I_{\text{post}} = i) \right] \quad \& \quad \mathbb{E}[\Delta_{\text{post}}] = \mathbb{E} \left[ \sum_{i \neq I^*} 1(I_{\text{post}} = i) \left( r_{\theta}(I^*) - r_{\theta}(i) \right) \right],
$$

revealing that post-experiment regret is similar to the probability of incorrect selection, except it is more forgiving of instances where “incorrect” but very nearly optimal arms are deployed post-experiment.

### 3 Deconfounded Thompson sampling

We propose deconfounded Thompson sampling (DTS). It is the natural way of applying Thompson sampling to our problem. At each time period $t \in [T]$, it selects an arm to measure randomly by sampling from the posterior distribution of the optimal arm:

$$
P(I_t = i \mid H_t, X_t) = P(I^* = i \mid H_t, X_t), \quad \forall i \in [k]. \quad (6)
$$

At the end of the experiment, DTS chooses the arm with highest expected reward in the population under posterior beliefs:

$$
I_{\text{post}} \in \arg \max_{i \in [k]} \mathbb{E} \left[ r_{\theta}(i) \mid H_{\text{post}} \right]. \quad (7)
$$

These definitions make no explicit reference to contextual observations. But implicitly, through proper Bayesian inference, DTS is using contextual observations to ‘deconfound’ its reward observations. Full pseudocode is given below.
To build intuition, we first consider a special case of the result that applies to vanilla bandit problems. In this case, DTS is just standard TS and the results we provide here are (essentially) known. By presenting them in a style that mirrors our main theorem, we hope to make it easier to digest the main theorem itself.

4 Main result

4.1 Warmup: bound in vanilla bandit environments

To build intuition, we first consider a special case of the our result that applies to vanilla bandit problems. In this case, DTS is just standard TS and the results we provide here are (essentially) known. By presenting them in a style that mirrors our main theorem, we hope to make it easier to digest the main theorem itself.

A striking feature of DTS is that the decision at time \( t \) does not depend on the context at time \( t \) — or even contexts in the past \( L - 1 \) periods. That is, in (6),

\[
P(I^* = i \mid H_t, X_t) = P \left( I^* = i \mid X_{1:1}, I_{1}, \ldots, I_{t-L}, R_{1}, \ldots, R_{t-L} \right).
\]

This equation uses that, conditioned on the observations \( X_{1:1}, I_{1}, \ldots, I_{t-L}, R_{1}, \ldots, R_{t-L} \), the latent variable \( \theta \) is independent of the additional arm selections and observed contexts. 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.
Under the next assumption, potential arm rewards \( (R_{t,i})_{t \in [T]} \) are i.i.d. with mean \( r_\theta(i) \) and rewards are observed immediately after an arm is played.

**Assumption 1** (Vanilla bandit problem). Suppose that \( L = 1 \) (no delay), the context dimension is \( d = 1 \), and with probability 1, \( X_1 = X_2 = \cdots = X_T = x_{\text{pop}} \).

Under this assumption, DTS is just standard TS followed by selecting the arm (7) at the end of the experiment. Summing the bound in (9) over \( t \in [T] \), yields familiar \( \tilde{O}(\sqrt{kT}) \) cumulative regret bounds for Thompson sampling [See e.g. Russo and Van Roy, 2016]. The form in (9) is stronger, since it bounds performance loss in every period, rather than on average. The bound in (10) ensures that TS gathers the variation. The main intellectual challenge is that learning dynamics in our model markedly deviate from those in i.i.d. bandit models. In i.i.d. environments, the DM can choose to quickly resolve uncertainty about an arm’s population-level performance through exploration. By contrast, our model introduces an unavoidable delay in this process as the DM awaits the occurrence of relevant contexts. A bound like (9), which maximizes posterior expected reward given all the information acquired throughout the experiment. This result is not commonly stated in the literature, but it is implied by the algorithm’s cumulative regret bounds; See [Russo and Van Roy, 2018, Proposition 8].

Recall that \( \Delta_t = r_\theta(I^*) - r_\theta(I_t) \) is the regret of the exploratory actions picked by DTS within the experiment. The post-experiment regret \( \Delta_{\text{post}} = r_\theta(I^*) - r_\theta(I_{\text{post}}) \) is the regret of the arm \( I_{\text{post}} \in \arg \max_{i \in [k]} E[r_\theta(i) \mid H_{\text{post}}} \) which maximizes posterior expected reward given all the information acquired throughout the experiment. It is possible to show that, in general, \( E[\Delta_{\text{post}}] \leq E[\Delta_t] \) for any \( t \), since \( I_{\text{post}} \) is selected based on more information and does not involve exploration. In this sense, (9) is the stronger and more surprising property.

**Corollary 1.** If Assumption 1 holds, then under DTS, within-experiment regret is bounded as

\[
E[\Delta_t] \leq \sigma \sqrt{\frac{2 \cdot t \cdot k \cdot \log(k)}{t - 1}}, \quad \forall t \in \{2, \ldots, T\},
\]

where \( i \) is defined in (15). Post-experiment regret is bounded as

\[
E[\Delta_{\text{post}}] \leq \sigma \sqrt{\frac{2 \cdot t \cdot k \cdot \log(k)}{T}}.
\]

Our use the term \( i \) to capture a messy factor which comes from the application of a concentration inequality. In our main regime of interest, \( i \) is a numerical constant, so we defer discussion until after our main theorem.

**4.2 General result**

We seek a generalization of Corollary 1 that holds throughout the scope of our problem formulation, removing the need for Assumption 1 and establishing the robustness of DTS to exogenous nonstationary variation. The main intellectual challenge is that learning dynamics in our model markedly deviate from those in i.i.d. bandit models. In i.i.d. environments, the DM can choose to quickly resolve uncertainty about an arm’s population-level performance through exploration. By contrast, our model introduces an unavoidable delay in this process as the DM awaits the occurrence of relevant contexts. A bound like (9), which says that DTS makes near optimal decisions as soon \( t \) is large, may not be possible under some context sequences.

Instead of depending explicitly on the number of arm pulls \( t \), our bound depends on a what we call **attainable precision**, defined as

\[
\text{Precision}(X_{1:t}) \triangleq \min_{i \in [k]} \frac{1}{\text{Var}(r_\theta(i) \mid (X_1, R_{1,i}, \ldots, X_t, R_{t,i}))}
\]

\[
= \min_{i \in [k]} \left( x_{\text{pop}}^\top \left[ \text{Cov}(\theta(i))^{-1} + \sigma^{-2} \sum_{t=1}^T X_t X_t^\top \right]^{-1} x_{\text{pop}} \right)^{-1},
\]

\[\text{Technical (9) can only be summed over } t \geq 2. \text{ It is easy to provide separate bounds when } t = 1.\]
To treat cases with no reward observations, define $\text{Precision}(X_{1:t-L}) \triangleq \min_{i \in [k]} \frac{1}{\text{Var}(y_{t-i})}$ when $t \leq L$. Precision is the inverse posterior variance of the arm’s population average reward if the potential reward outcomes $(R_{i,t})_{t=1,\ldots,T}$ from measuring the arm in contexts $X_1, \ldots, X_t$ were observable. The formula in (12) uses standard rules for computing Gaussian posterior distributions. Under Assumption 1, $\text{Precision}(X_{1:t}) \geq \sigma^{-2} \cdot t$ and Theorem 1 implies the corollary stated above. More generally, if the contexts so far are reflective of the population distribution (e.g. they are drawn i.i.d.), then precision scales as $\sigma^{-2} \cdot t$, with no or minimal dependence on context dimension; See Lemma 1. But precision can behave quite differently if contexts have a strong non-stationary pattern; see Figure 4 in Section 5.

Attainable precision measures whether the decision-maker could have precisely estimated an arm’s population average reward by playing it in each context observed so far in the experiment. The next theorem formalizes a striking result about DTS: once high precision is attainable, the expected regret of each subsequent decision made by DTS is low. The result generalizes Corollary 1 to problems with exogenous nonstationary variation and delayed reward observations. Full discussion is deferred until Subsection 4.4.

**Theorem 1** (Bound on within- and post-experiment utilitarian regret). Fix any sequence $x_{1:T} \in \mathcal{X}^T$. Under DTS, within-experiment regret is bounded as

$$
\mathbb{E}[\Delta_t \mid X_{1:T} = x_{1:T}] \leq \sqrt{2 \cdot t \cdot k \cdot \log(k)} \cdot \text{Precision}(x_{1:t-L}), \quad \forall t \in [T],
$$

where $t$ is defined in Equation (15). Post-experiment regret is bounded as

$$
\mathbb{E}[\Delta_{\text{post}} \mid X_{1:T} = x_{1:T}] \leq \sqrt{2 \cdot t \cdot k \cdot \log(k)} \cdot \text{Precision}(x_{1:T}).
$$

We define

$$
\iota \triangleq \max \left\{ 8 \left( \sigma^{-2} / \lambda_{\min}(\Sigma_1^{-1}) \right) \cdot \log \left( dk \lambda_{\max}(\Sigma_1) \left[ \lambda_{\max}(\Sigma_1^{-1}) + \sigma^{-2} T \right] \right), 1, 9 \right\}
$$

$$
= \tilde{O} \left( \max \left\{ \lambda_{\max}(\Sigma_1) / \sigma^2, 1 \right\} \right),
$$

where $\tilde{O}$ hides logarithmic factors. This term comes from applying a concentration inequality to control for the impact of randomness in action selection; see inequality (a) in the proof sketch in Section 4.5. We are interested in problems with a low signal-to-noise ratio — where a single user interaction does not resolve much uncertainty — in which case $\iota$ is a constant.

**Remark 5** (Treating $\iota$ as a constant). In choosing to downplay the importance of $\iota$, we are implicitly assuming that the signal-to-noise ratio is $\tilde{O}(1)$, i.e. we are in a regime where observing a single reward realization does not resolve most prior uncertainty. Indeed, many A/B tests involve just a few treatment arms, but still require (many) millions of users to attain statistical power. A line of the literature formally studies such a regime by taking a diffusion limit of bandit problems [Kuang and Wager, 2023, Fan and Glynn, 2021, Araman and Caldentey, 2022, Adusumilli, 2023] which is similar to letting $\lambda_{\max}(\Sigma_1) / \sigma^2 \to 0$ but taking the time horizon $T \to \infty$ at a comparable rate. In such a limit, $\iota = 9$, a numerical constant that comes from crude application of concentration inequalities.

Notice that our overall regret bounds do not degrade as the noise variance $\sigma^2$ tends to zero. In that case, two factors of $\sigma$, one in $\iota$ and the other in the right-hand-side of (9) or (10), will cancel. However, our analysis is not well suited to tightly bounding the regret incurred when $\sigma \approx 0$. 
4.3 Growth rate of attainable precision

In benign settings, where observed contexts are generally reflective of the population distribution, precision in period $t$ scales with $\sigma^{-2} \cdot t$ and does not depend on the context dimension. In such cases, the bounds in Theorem 1 are roughly on the order of $\sigma \sqrt{k/(t-L)}$ or $\sigma \sqrt{k/T}$.

The next lemma provides four results in such settings. The first result is a generic bound from which other bounds follow. The second considers standard $k$-armed bandit problem, viewed as a special case of our formulation. The third generalizes the second, allowing for arbitrary context order while requiring that the empirical mean of the contexts matches the population mean. The fourth result integrates the first result with concentration inequalities applied to sample covariance matrices.

**Lemma 1** (Bound on attainable precision). Fix any sequence $x_{1:T} \in X^T$ and $t \in [T]$.

1. (Generic bound) Let $S_x \triangleq \frac{1}{t} \sum_{\ell=1}^t x_\ell x_\ell^\top$ denote the empirical second moment matrix and $\hat{S}_x \triangleq S_x + \frac{c^2 \lambda_{\min}(\Sigma_1^{-1})}{t} I$ (where $I \in \mathbb{R}^{d \times d}$ is an identity matrix). Then

   $$\text{Precision}(x_{1:t}) \geq \sigma^{-2} t \cdot \left( x_{\text{pop}}^\top \hat{S}_x^{-1} x_{\text{pop}} \right)^{-1}.$$  

2. (Vanilla bandit) Suppose $d = 1$ and $x_\ell = 1 = x_{\text{pop}}$ for each $\ell \in [t]$. Then

   $$\text{Precision}(x_{1:t}) = \min_{i \in [k]} \Sigma_{1,ii}^{-1} \Sigma_{1,ii}^{-1} + \sigma^{-2} t \geq \lambda_{\min} \left( \Sigma_1^{-1} \right) + \sigma^{-2} t,$$

   where $\Sigma_{1,ii}$ is the $(i,i)$-th element of the prior covariance matrix $\Sigma_1$.

3. (No empirical distribution shift) Suppose $\frac{1}{t} \sum_{\ell=1}^t x_\ell = x_{\text{pop}}$. Then

   $$\text{Precision}(x_{1:t}) \geq \lambda_{\min} \left( \Sigma_1^{-1} \right) \|x_{\text{pop}}\|^2 + \sigma^{-2} t.$$  

4. (i.i.d. contexts) Suppose $X_1, \ldots, X_t$ are drawn i.i.d. from a distribution satisfying that $\mathbb{E}[X_1 X_1^\top] \succeq c \cdot x_{\text{pop}} x_{\text{pop}}^\top$ for some $c > 0$. Then for any $\delta > 0$, with probability greater than $1 - \delta$,

   $$\text{Precision}(x_{1:t}) \geq \lambda_{\min} \left( \Sigma_1^{-1} \right) \|x_{\text{pop}}\|^2 + c \cdot \sigma^{-2} t - 4c^2 \cdot \|x_{\text{pop}}\|^2 \cdot 2t \log \left( \frac{d}{\delta} \right).$$

   and

   $$t \geq \frac{128\|x_{\text{pop}}\|^4 \log \left( \frac{d}{\delta} \right)}{c^2} \implies \text{Precision}(x_{1:t}) \geq \lambda_{\min} \left( \Sigma_1^{-1} \right) \|x_{\text{pop}}\|^2 + \frac{c}{2} \cdot \sigma^{-2} t.$$  

Beyond the settings considered in this lemma, attainable precision can depend in an interesting way on the context sequence and the prior distribution. Figure 4 in Section 5.2 provides an illustration.

4.4 Discussion of the main result

Theorem 1 has several striking implications about the performance of DTS.

A delicate balance between exploration and exploitation. The attainable precision in estimating an arm’s performance, defined in (11), imagines that the potential reward outcomes of that arm were observed in every period. An adaptive algorithm can try to emulate this by selecting arms uniformly at random, roughly leading to the same bound on post-experiment performance as in (14). But doing so would forego the possibility of having low regret within the experiment as shown for DTS in (13). Attaining these two guarantees requires striking a delicate balance between exploring arms to gather all attainable
information that is useful, and also aggressively exploiting this information by shifting measurement effort away from bad arms.

**Robustness to context order.** Theorem 1 highlights DTS’s robustness when faced with a challenging context order. For instance, Example 2, presented in the next section, studies a weeklong experiment in which the first $T/7$ contexts are Monday, then next $T/7$ are Tuesday, and so on, until the last $T/7$ contexts are Sunday. The bound in (14) implies that the DTS still gathers adequate information by the end of the experiment. Sections 5 and 6 explain why this context order can create challenges in the design and analysis of bandit algorithms.

**Robustness to delayed observations.** Recall that rewards are observed only after some delay of $L \geq 1$ periods. When $L$ is very large, DTS is not able to get feedback on the decisions during the experimentation phase. For that reason, the bound in (13) measures precision offered by the context sequence up to $L$ periods ago. This mild dependence on $L$ provides assurances of robustness. According to our formulation, the decision-maker can wait for all rewards observations to realize before implementing a decision in the population, which is why it is possible for (14) to have no dependence on $L$. That bound suggests that, even in the face of extreme delay, DTS’s arm selections will provide adequate information if one waits for the rewards to realize.

**Low price of using contexts to deconfound.** The result highlights the low price of using rich contextual information to deconfound. Unlike contextual bandit results, under which regret generally scales polynomially in the context dimension $d$ [Agrawal and Goyal, 2013], our bound has at most a logarithmic dependence on $d$ when the contexts satisfy the conditions of Lemma 1. This also mimics the bound in Proposition 3 in Appendix B, which is completely independent of the dimension of context vectors. Of course, bounds that are nearly independent of the context dimension offer a stronger guarantee. More importantly, they offer a different conceptual guidance to a practitioner: when using contexts to ‘deconfound’ inferences, but not to personalize decisions, it is better to use very rich features.

### 4.5 Analysis

The analysis leading to Theorem 1 may be of independent interest. We outline ideas underlying the proof of (13), which is the more delicate part, with (14) following as a corollary of the analysis. A key quantity in the analysis is the posterior standard deviation of population average reward: for any $t \in [T]$ and $i \in [k],$

$$s_{t,i} \equiv \sqrt{\text{Var}(r_\theta(i) \mid H_t)}.$$

We also define the propensity (also called “propensity score”) assigned to arm $i$ at time $t$ by

$$p_{t,i} \equiv \Pr(I_t = i \mid H_t, X_t).$$

Learning about the population reward of an arm has limited value if that arm is believed to be very unlikely to be optimal. The term

$$\mathbb{E} \left[ s_{t,I_t}^2 \mid H_t, X_{1:T} \right] = \sum_{i=1}^{k} \Pr(I_t = i \mid H_t) \cdot s_{t,i}^2 = \sum_{i=1}^{k} p_{t,i} \cdot s_{t,i}^2,$$

assesses remaining uncertainty about the performance of arms while giving low weight to arms that are unlikely to be optimal under the posterior. (That $X_{1:T}$ does not appear on the right-hand-side of (16) follows from logic similar to Equation (8)).

The proof highlights two key properties of DTS.

**DTS exploits what is known.** The next result shows DTS has small expected regret in any period if the posterior uncertainty in (16) is small. A relatively short proof is given in Appendix F.2.
Proposition 1 (Reduction to estimation). Under DTS, for any $t \in [T]$,
\[
E \left[ \Delta_t \mid H_t, X_{1:T} \right] \leqslant \sqrt{2 \log(k)E \left[ s_{1,t}^2 \mid H_t, X_{1:T} \right]} \quad \text{and} \quad E \left[ \Delta_t \mid X_{1:T} \right] \leqslant \sqrt{2 \log(k)E \left[ s_{1,t}^2 \mid X_{1:T} \right]}.
\]

**DTS explores the optimal arm.** The next proposition formalizes that, regardless of the context sequence and delay $L$, DTS is expected to assign high propensity to the optimal arm — in the sense that the expected inverse propensity is uniformly bounded. Although the proof is very short, we call this a proposition to reflect the critical role it plays in our analysis.

Proposition 2. Under DTS, for any $t \in [T]$,
\[
E \left[ \frac{1}{p_{1,t^*}} \mid X_{1:T} \right] = k.
\]

**Proof.** By the tower property,
\[
E \left[ \frac{1}{p_{1,t^*}} \mid X_{1:T} \right] = E \left[ E \left[ \frac{1}{p_{1,t^*}} \mid H_t, X_{1:T} \right] \mid X_{1:T} \right] = E \left[ \frac{P(I^* = i \mid H_t, X_{1:T})}{p_{1,i}} \mid X_{1:T} \right] = E \left[ \frac{P(I^* = i \mid H_t, X_{1})}{p_{1,i}} \mid X_{1:T} \right] = 1.
\]

The penultimate equality uses that $X_{1:(t-1)}$ is already contained in the history $H_t$ and that $X_{(t+1):T}$ is independent of $\theta$ conditioned on $H_t$. The last equality uses the definition of DTS in (6). We conclude,
\[
E \left[ \frac{1}{p_{1,t^*}} \mid X_{1:T} \right] = \frac{1}{p_{1,t^*}} = \frac{\sum_{i=1}^k \frac{1}{p_{1,i}}}{p_{1,t^*}} = \sum_{i=1}^k \frac{1}{p_{1,i}}.
\]

A delicate balance between exploration and exploitation. To get some intuition for these results, let’s compare them to what could be attained under alternative algorithms. First, consider an RCT which sets $p_{1,i} = 1/k$ for each period $t$ and arm $i$. This algorithm explores aggressively, if naively. Assuming all arms equally likely to be optimal (i.e. $P(I^* = i) = 1/k$), then this method would attain the same bounds in Proposition 2, but it would not attain the low-regret property in Proposition 1. Next, consider a greedy algorithm, which selects the arm $\arg \max_{i \in [k]} E[\theta(i) \mid H_t]$ in each time period. That algorithm “exploits what is known” and attains the bound in Proposition 1, but it may neglect to explore the optimal arm and does not satisfy a bound like Proposition 2.

That DTS attains both properties reflects a delicate balance it strikes between exploration and exploitation. It is aggressive in shifting measurement effort away from poor arms, leading to Proposition 1, but it is still assured to explore all arms which might be optimal, leading to Proposition 2.

**Completing the proof.** To complete the proof, we show that sufficient exploration of the optimal arm, in the sense of Proposition 2, controls the expected posterior variance of the optimal arm (i.e. $E[s_{1,t^*}^2]$) which appears in Proposition 1. The full analysis is quite subtle, but it is possible to give a thorough proof sketch in a special case.

**Proof sketch in the orthogonal case.** Consider a special case of our formulation. To avoid writing conditional expectations, assume $X_{1:T} = x_{1:T} \in \mathcal{X}^T$ with probability 1 for some arbitrary sequence $x_{1:T}$. Now, Assume $\Sigma^{-1} = \lambda I$ (representing independent beliefs) and that $\|x_t\| = 1$ for each $t$ (so pairs of context vectors are either orthogonal or aligned). In this special case, the posterior covariance matrix $\Sigma \in \mathbb{R}^{dk \times dk}$ is diagonal.
with entries
\[ \sigma_{i,j}^2 \equiv \text{Var} \left( \theta^{(i)} \mid H_t \right) = \left( \lambda + \sigma^{-2} \sum_{\ell=1}^{t-L} \mathbb{1}(I_\ell = i)x_{t,\ell,j}^2 \right)^{-1}, \quad \forall i \in [k], j \in [d], \]
along the diagonal. Moreover, since \( r_\theta(i) = x_{\text{pop}}^* \theta^{(i)} \), the posterior variance of population average reward can be written as \( s_{i,j}^2 = \sum_{j=1}^d x_{\text{pop},j}^2 \sigma_{i,j}^2 \). We then bound this as
\[
E \left[ s_{i,t}^2 \right] = \sum_{j=1}^d x_{\text{pop},j}^2 \cdot E \left[ \left( \lambda + \sigma^{-2} \sum_{\ell=1}^{t-L} \mathbb{1}(I_\ell = I^*)x_{t,\ell,j}^2 \right)^{-1} \right]
\leq t \sum_{j=1}^d x_{\text{pop},j}^2 \cdot E \left[ \left( \lambda + \sigma^{-2} \sum_{\ell=1}^{t-L} p_{\ell,t} x_{t,\ell,j}^2 \right)^{-2} \right] \cdot \left( \lambda + \sigma^{-2} \sum_{\ell=1}^{t-L} x_{t,\ell,j}^2 \right) \cdot k
= \frac{t \cdot k}{\text{Precision} \left( x_{1:(t-L)} \right)}.
\]
Inequality (c) applies Proposition 2. Inequality (b) uses Jensen’s inequality.

Inequality (a) requires a detailed proof, but we can provide semi-rigorous intuition. To study both sides of the inequality (a), fix any arm \( i \) and define \( D_{t,j} = \|I_\ell = i \| - p_{t,j} x_{t,j}^2 \). This has zero conditional mean (i.e. \( E[D_{t,j} \mid D_{t,j}, \ldots, D_{t-1,j}] = 0 \)) and conditional variance \( v_{t,j} = E[D_{t,j}^2 \mid D_{t,j}, \ldots, D_{t-1,j}] = p_{t,j} (1 - p_{t,j}) x_{t,j}^4 \leq p_{t,j} x_{t,j}^2 \) (by the assumption that \( \|x_\ell\|_2 \leq 1 \)). Then,
\[
\sum_{\ell=1}^{t-L} \mathbb{1}(I_\ell = I^*) x_{t,\ell,j}^2 = \sum_{\ell=1}^{t-L} p_{t,j} x_{t,\ell,j}^2 + \sum_{\ell=1}^{t-L} D_{t,j} \approx \sum_{\ell=1}^{t-L} p_{t,j} x_{t,\ell,j}^2 + O \left( \left\| D_{t,j} \right\|_2 \right) = \sum_{\ell=1}^{t-L} p_{t,j} x_{t,\ell,j}^2 + O \left( \sum_{\ell=1}^{t-L} v_{t,j} \right).
\]
The approximate equality (marked \( \approx \)) can be loosely justified through the martingale central limit theorem. The rigorous proof, given in Appendix F.4, instead relies on a non-asymptotic martingale concentration inequalities.

This proof technique generalizes to problems with non-orthogonal context vectors, but it requires careful matrix-valued generalizations of all key inequalities. A generalization of inequality (a) is given in Lemma 6, in Appendix F.4. In proving this, we developed a new concentration inequality for matrix-valued martingales (i.e. Proposition 5), which may be of independent interest. In Appendix F.5, Lemma 8 presents a matrix-valued generalization of inequality (b). Its proof relies on a remarkable generalization of Jensen’s inequality to operator convex functions, which we restate as Lemma 7.
5 Numerical illustration

We provide numerical experiments that motivate our theory and help the reader build intuition. Specifically, these illustrations provide a glimpse of the challenges outlined in Section 6 and of DTS’s intricate balance of exploration and exploitation, which we formalized in Theorem 1. While we compare DTS with alternative algorithms, our intent is not to conduct extensive competitive benchmarking.

5.1 An example with day of week effects

Our simulations center around Example 2, which demonstrates the challenges faced when the context sequence exhibits nonstationary pattern. The example models a week-long experiment where observations are influenced by day-of-week effects, a routine concern in A/B testing [Kohavi et al., 2020].

Example 2 (Day-of-week effects). Consider an online retailer conducting a weeklong experiment to find the price that maximizes profit from selling a product in subsequent weeks. Demand is assumed to follow a normal distribution, implying that profit also follows a normal distribution. Demand varies according to the day of the week. This scenario can be mapped to a special case of the model in Section 2, where each context $X_t \in \{e_1, \ldots, e_7\} \subset \mathbb{R}^7$ is one of the standard basis vectors. Suppose $T = 7m$ and the context at time $t$ is $X_t = e_{[t/m]}$, signifying that first $m$ periods are Sunday, the next $m$ are Monday, and so on, with the final $m$ being Saturday. The price $I_t$ is adjusted in each period and offered to the next customer (a time period could also represent a small batch of customers), generating reward $R_{t,i} \equiv \langle \theta^{(i)}, X_t \rangle + W_{t,i}$ representing the profit earned. There is no delay in observing rewards (i.e. $L = 1$). Let the population distribution $D_{\text{pop}}$ be uniform over $\{e_1, \ldots, e_7\}$. The performance of arm $i$ on day $x$ is the $x$-th component of the vector $\theta^{(i)}$, i.e., $\theta^{(i)}_x = \langle \theta^{(i)}, e_x \rangle$. At the end of the experiment, the decision-maker picks a single price $I_{\text{post}}$ to employ across future weeks. The loss incurred due to the decision made under incomplete resolution of uncertainty about average demand is measured by

$$\Delta_{\text{post}} = \max_{i \in \{1, \ldots, 7\}} \left( \frac{\theta^{(i)}_1 + \cdots + \theta^{(i)}_7}{7} \right) - \left( \frac{\theta^{(I_{\text{post}})}_1 + \cdots + \theta^{(I_{\text{post}})}_7}{7} \right).$$

The reasons for learning a single price, pertaining to fairness and incentive-compatibility, are discussed in Appendix D.3.

The decision-maker begins with prior belief that $\theta \sim N(\mu, \Sigma)$. We consider a structured prior induced from a latent variable model where $\theta^{(i)}_x = \theta^{(i)}_{\text{arm}} + \theta^{(i)}_{\text{day}}$ is determined by an effect $\theta^{(i)}_{\text{arm}}$ that is idiosyncratic to a specific arm and day, an effect $\theta^{(i)}_{\text{day}}$ associated with the chosen arm, and a shared day-of-week effect $\theta^{(i)}_{\text{day}}$. 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 be cautious of almost arbitrary nonstationary patterns. If these are believed to have smaller magnitude, the decision-maker may be able to rule out some very poor arms early in the experiment.

5.2 Attainable precision and delayed learning due to context order

Figure 4 plots attainable precision in a special case of Example 2. Recall this is defined as

$$\text{Precision}(X_{1:t}) \triangleq \min_{i \in \{1, \ldots, 7\}} \frac{1}{\text{Var}(\theta(i) \mid (X_{1}, R_{1,i}, \ldots, X_{t}, R_{t,i}))}$$

and assesses the remaining uncertainty a decision-maker would have about an arm’s population-level performance assuming they chose to measure that arm exclusively. Arms are a priori symmetric in our example, so the minimum and maximum above are redundant. We plot this in two cases.
Sequential context order: Context 1 occurs for the first 100 periods (‘Monday’), context 2 occurs for the next 100 periods (‘Tuesday’), and so on.

Shuffled context order: Contexts are drawn i.i.d. across periods with uniform probabilities.

In both cases, attainable precision at the end of the experiment is \( \text{Precision}(X_{1:T}) = \sigma^{-2} \cdot T = 700 \). If we interpret this as a ‘large’ value, then the bound in equation (14) of Theorem 1 suggests that DTS will attain low post-experiment regret in either case. However, the evolution of attainable precision within the experiment looks very different depending on the context order.

When contexts are shuffled, precision displays linear growth with \( \text{Precision}(X_{1:t}) \approx \sigma^{-2} \cdot t \), mirroring the bounds in Subsection 4.3. The posterior variance, which is the inverse of precision, undergoes a rapid decrease following the onset of the experiment. This indicates that if the DM chose to explore an arm \( i \) aggressively at the beginning of the experiment, they could resolve uncertainty about its population-level performance \( \theta_i \).

The behavior of attainable precision changes substantially under a sequential context order. It grows slowly at the beginning of the experiment, reflecting that resolving uncertainty about an arm’s population level performance requires waiting for certain contexts to become observable. In fact, the figure displays fairly sharp jumps in attainable precision when new contexts become observable — marked in Figure 4 by alternating grey and white shaded columns.

5.3 Algorithms compared

5.3.1 Methods for selecting arms within an experiment

Our numerical experiments compare the following procedures for selecting arms \( I_1, \ldots, I_T \) within-the-experiment.

Deconfounded Thompson sampling: Implements Algorithm 1.
Deconfounded UCB: The UCB analogue of deconfounded TS. This algorithm defines an upper confidence bound $U_{t,i} = \mathbb{E}[r_\theta(i) \mid H_t] + z\sqrt{\text{Var}(r_\theta(i) \mid H_t)}$ on the population-average reward $r_\theta(i)$ of arm $i$, then it selects the arm $I_t = \arg \max_{i \in [k]} U_{t,i}$. In this study, we use $z = 3$, but alternative choices produce qualitatively similar outcomes.

Context-unaware Thompson sampling: A version of TS that acts as if rewards were i.i.d. and there were no contextual observations. It imagines each sample of arm $i$ is a draw with mean $\mathbb{E}[R_{t,i} \mid \theta] = r_\theta(i)$ and noise variance $\text{Var}(R_{t,i} \mid \theta) = \sigma^2 + \max_x \text{Var}(\theta_x^{\text{day}})$; the noise variance is inflated since the algorithm is not accounting for variance driven by context.

Round-robin sampling: The algorithm samples arm 1 when $t = 1$, arm 2 when $t = 2, \ldots$, arm $k$ when $t = k$, and then starts the cycle again, sampling arm 1 when $t = k + 1$ and so on.

Sequential elimination: The algorithm maintains a set of contending arms, which contains all arms at initialization. At the start of any period, an arm whose posterior probability of being optimal, $\mathbb{P}(I^* = i \mid H_t)$, falls below some threshold $\delta/k$ is removed from the set of contending arms. We set $\delta = 0.05$, reflecting a goal of having less than a 5% chance of eliminating the best arm. A suitable variant of round-robin sampling is used to select an arm to sample in each period from the arms still in contention.

### 5.3.2 Methods for selecting an arm to deploy post-experiment

Every procedure we evaluate selects an arm post-experiment in a Bayes optimal manner.

Minimizing regret: Set $I_{\text{post}} \in \arg \max_{i \in [k]} \mathbb{E}[r_\theta(i) \mid H_{\text{post}}]$ to be the Bayes optimal arm for a decision-maker who wishes to maximize population-level reward. To visualize decision-quality if the experiment we stopped early, we set $\hat{I}_t \in \arg \max_{i \in [k]} \mathbb{E}[r_\theta(i) \mid H_t]$ and evaluate the regret $\mathbb{E}[r_\theta(I^*) - r_\theta(\hat{I}_t)]$. Figure 5 presents this as “future regret if experiment were stopped.”

Minimizing the probability of incorrect selection: Set $I_{\text{post}} \in \arg \max_{i \in [k]} \mathbb{P}(I^* = i \mid H_{\text{post}})$ the Bayes optimal arm for a decision-maker who wishes to maximize the probability of correct selection. To visualize decision-quality if the experiment we stopped early, we set $\hat{I}_t \in \arg \max_{i \in [k]} \mathbb{P}(I^* = i \mid H_t)$ and evaluate the probability of correct selection $\mathbb{P}(\hat{I}_t = I^*)$. Figure 5 presents this as “confidence in identity of the best arm.”

Because these rules are Bayes optimal, an algorithm which suffers high post-experiment regret, or attains low probability of correct selection, does so because of inadequate information gathering within the experiment; it is not possible to improve performance by changing how decisions are made post-experiment\(^4\).

### 5.4 Discussion of experiment results

We simulate algorithms applied to Example 2. Our simulations use noise variance $\sigma^2 = 1$. The latent variables $\theta_x^{\text{arm}}$ and $\theta_x^{\text{day}}$, $\theta_{ix}^{\text{di}}$ have mean zero and prior standard deviation 0.5, 1.0, and 0.8 respectively. We make a number of observations:

Results with shuffled context order. With shuffled context order, all algorithms succeed in confidently identifying the best arm and have low post-experiment regret. Bandit algorithms like TS and UCB shift sampling effort away from clearly bad actions within the experiment and this reduces the regret they incur. Context unaware TS succeeds when contexts are shuffled by treating (unmodeled) contexts as if

\[^4\text{Consider any other rule } \pi_{\text{post}} \text{ that selects an arm } \hat{I}_{\text{post}} = f(H_{\text{post}}). \text{ Then}
\]

\[
\mathbb{E}[r_\theta(I_{\text{post}})] = \mathbb{E}[\mathbb{E}[r_\theta(I_{\text{post}}) \mid H_{\text{post}}]] = \mathbb{E}\left[\mathbb{E}\left[\max_{i \in [k]} r_\theta(i) \mid H_{\text{post}}\right]\right] \geq \mathbb{E}\left[\mathbb{E}\left[\max_{i \in [k]} r_\theta(I_{\text{post}}) \mid H_{\text{post}}\right]\right] = \mathbb{E}[r_\theta(I_{\text{post}})].
\]

A procedure that selects the arm with highest posterior mean at the end of the experiment yields greater expected reward post-experiment than any alternative, regardless of which procedure (e.g. DTS or deconfounded UCB) is used to sample arms during the experiment.
they were i.i.d. observation noise. Even when contexts are i.i.d., this is not statistically efficient since ‘controlling for’ observed contexts would reduce variance. This is reflected in the fact that the regret of context unaware TS is larger than that of DTS in Figure 5a (though, this is not a huge issue for our particular experiment parameters).

Delayed learning due to context order. For concreteness, let’s focus on round-robin sampling. In the experiment with shuffled context order, round-robin sampling quickly found a near optimal arm to deploy in the population. Due to low reward noise (i.e. small $\sigma^2$), uncertainty resolves rapidly. With sequential context order, despite low reward noise, uncertainty about an arm’s performance on Sunday only resolves at the end of the experiment. Hence, uncertainty about an arm’s average performance throughout the week only resolves at the end of the experiment. The top-left of Figure 5b shows that uncertainty about the identity of the optimal arm resolves in sharp jumps at the start of each day — a behavior that is quite different from what is depicted in Figure 5a. At least qualitatively, this finding parallels the behavior of attainable precision in Figure 4.

Robustness to sequential context order. DTS, round-robin sampling, and sequential elimination demonstrate robustness to sequential context order, while deconfounded UCB and context unaware TS appear brittle. Notably, DTS, round-robin sampling, and sequential elimination suffer tiny post-experiment regret once all contexts have been observed. In contrast, even after all days of the week have been observed, context-unaware TS and deconfounded UCB cannot identify an optimal arm to deploy post-experiment. Since all algorithms were evaluated assuming that correct posterior inferences were used for post-experiment arm selection, the failure of these algorithms indicates an inadequacy in the information they gather.

The performance differences between deconfounded TS and a deconfounded (Bayesian) UCB in Figure 5b are quite striking, given that the literature has often emphasized the similarities between these algorithms. A closer look at the experiment results reveals that deconfounded UCB often plays only a single arm on certain days of the week, completely failing to gather information about some arms on some days of the week. See the next section for further discussion.

Aggressive exploitation. DTS incurs lower regret within the experiment than both round-robin sampling and sequential elimination. This is attributable to its aggressive approach in shifting effort away from arms that have a low posterior probability of being optimal given current evidence. By comparison, sequential elimination incurs greater regret within the experiment as it cannot respond to weak initial evidence of an arm’s poor performance; sequential elimination treats all arms equally unless it is highly confident that a particular arm can be ruled out.

Contextual regret. In addition to our main regret measure, we compare algorithms in terms of what we term their cumulative “within-experiment contextual regret”: $E \left[ \sum_{t=1}^{T} (R_{t,I}^* - R_{t,I}) \right]$. DTS seems to perform well according to this metric as well. Appendix B confirms that this is always true by bounding the cumulative contextual regret of DTS. One should not focus on the fact that deconfounded UCB attains negative contextual regret in this particular experiment. This is not a general phenomenon, and it is possible to construct examples, along the lines of Example 3, in which it incurs large contextual regret.

6 Challenges of our model: the unexpected failure of deconfounded UCB

As expected, our numerical experiments show that context-unaware algorithms can falter. Controlling for exogenous variation is critical to drawing accurate inferences about arms’ performance.

The numerical experiments, however, indicate that our model hosts additional surprises. While controlling for sources of exogenous variation is crucial, it can introduce unavoidable delays in the resolution of
Figure 5: Algorithm performance in Example 2 with $m = 100$ periods per context and noise variance $\sigma^2 = 1$. The latent variables $\theta^\text{arm}_i$ and $\theta^\text{day}_x$, $\theta^\text{idio}_x$ have mean zero and prior standard deviation 0.5, 1.0, and 0.8 respectively.

uncertainty as the DM anticipates relevant contexts that have yet to occur. To illustrate this point, we present a simplified variant of Example 2.

Example 3 (Simplified day-of-week effects). Consider a two-day experiment with $k = 2$ arms and context set $X = \{e_1, e_2\} \subset \mathbb{R}^2$. The context sequence is deterministic, with $X_t = e_1$ for $t \leq \lfloor T/2 \rfloor$, $X_t = e_2$ for $t > \lfloor T/2 \rfloor$. The goal is to identify the best arm under equal context weights $x_{\text{pop}} = (0.5, 0.5)$. The components of vector $\theta = (\theta^{(i)}_x)_{i \in [2], x \in [2]}$ are independent with $\theta^{(i)}_x = (\theta^{(i)}, e_x)$ being the performance of arm $i$ on day $x$. The reward at time $t$ is $R_t, l_x = \langle \theta^{(i)}, X_t \rangle$ (i.e. $\sigma = 0$ so there is no reward noise\footnote{Technically, we assumed $\sigma > 0$ at the end of the problem formulation, writing that this allowed us to write expressions like $1/\sigma^2$, which appear often in the analysis. One could take $\sigma$ to be extremely close to 0 in this example, but the presentation is much cleaner if it equals zero exactly.}). Reward observations are not subject to delay (i.e. $L = 1$).

It is straightforward to design a learning procedure for this example. With no observation noise, the DM merely needs to play both arms once in each of the two contexts. However, unlike in an i.i.d. bandit model, the DM cannot opt for aggressive exploration to rapidly resolve uncertainty. Understanding an arm’s population-level performance requires waiting until the second half of the experiment when the second context becomes observable. Before that, the DM remains uncertain.

Algorithms are differentiated by how they explore when faced with this uncertainty about population-level performance that they cannot rapidly resolve. The following lemma shows that deconfounded UCB continues to sample one arm repeatedly during the first half of the experiment. Because of this failure of information gathering, it can’t evaluate one arm’s population-level performance even at the end of the experiment.

Lemma 2 (Failure of deconfounded UCB). Consider Example 3. Suppose that $\theta^{(1)}_x \sim N(0, 1)$ and $\theta^{(2)}_x \sim N(0, 3)$ for $x \in [2]$. If, for any fixed $\varepsilon > 0$, $I_t \in \arg\max_{i \in [k]} E[r_\theta(i) \mid H_t] + \varepsilon \sqrt{\text{Var}(r_\theta(i) \mid H_t)}$ holds for every $t$, then there is an absolute numerical constant $c > 0$ such that for all $T \in \mathbb{N}$, $E[\Delta_{\text{post}}] \geq c$.

Proof sketch. During the first half the experiment, when the context is $e_1$, deconfounded UCB plays only action 1. The UCB for action 1 exceeds that of action 2, and this UCB stays very large until after the second
context is observed. Since the reward of arm 2 in context $e_1$ is never observed, the DM may fail to deploy an optimal arm in the population. A complete proof is provided in Appendix E.

Intuitively, it seems that DTS might avoid this information-gathering failure. During the first half of the experiment, DTS would continue (randomly) sampling both arms, only shifting measurement effort away from an under-performing arm once its posterior probability of being optimal is low and further information gathering is not useful. Our theory confirms this intuition, demonstrating that, despite its aggressive exploration, DTS gathers enough information to ensure low post-experiment regret across a broad class of problems.

It is likely possible to modify deconfounded UCB so that it performs well in this straightforward example.\textsuperscript{6} We leave this to future work, and instead focus on showing that DTS explores efficiently without any such modifications.

7 Conclusion

7.1 Closing thoughts

This paper proposes a new way to model adaptive experiments conducted in the presence of nonstationary variation. Out of this model comes a more robust variant of the prominent Thompson sampling algorithm. We provide several theoretical results that provide assurances of its robustness. At a casual glance, one might expect developing this theory to require a routine – if intricate – exercise in adapting widely used arguments in the literature. Perhaps surprisingly, this problem class raises many new subtleties, as is reflected in the failure of deconfounded UCB, the departure of learning dynamics in Section 5 from those in i.i.d. bandit problems, and the original theorem statement and proof in Section 4.

Our model is quite flexible. Special cases of it, like Example 1 in the introduction or Example 4 in the appendix, differ significantly. The extensions covered below provide even more flexibility. On the positive side, this flexibility expands the scope of problems to which DTS and our theory can be applied. Unfortunately, it also leaves a practitioner with many subtle modeling choices. A nice complement to this paper would be one focuses on a narrow real-world use case and carefully documents many of the modeling choices involved.

7.2 Extensions

We close by mentioning two extensions that broaden the applicability of DTS.

Policy learning. Thompson sampling can be readily applied to contextual bandit problems where the goal is to learn an optimal policy that segments or personalized its decisions on the basis of observed contexts. In proposing DTS, we have shown how to adapt Thompson sampling so as to control for exogenous sources of variation while learning a stable decision-rule: one which does not react to evolving context. Appendix D provides a full discussion of and motivation for this difference. In that section, we also extend DTS to learn policies that are reactive to some parts of the context but not others. We explain how to provide a more conventional regret bound for that algorithm, but are not certain how to extend the proof of Theorem 1 to treat this generalization.

Top-two sampling and a prioritization of within-experiment regret. We have evaluated DTS in terms of two broad performance criteria: the regret incurred (or reward accrued) during the experiment and the regret incurred (or reward accrued) post-experiment. For those who wish to prioritize attaining very low post-experiment experiment regret, it may be helpful to consider Top-two sampling [Russo, 2020] variants of

\textsuperscript{6}One can define an algorithm that plays arms randomly with probabilities that depend on upper confidence bounds. One can also force the algorithm to continue sampling all arms with high UCBs, eliminating arms once it is clear that they underperform. These make the decision-making logic similar to Thompson sampling or sequential elimination, respectively.
DTS that explore more aggressively. Top-two 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 the probability mass function $P(I^* = \cdot | H_t)$ until two distinct arms are chosen. Flip a (biased) coin to select one among these two.

This procedure bootstraps standard randomized arm selection by DTS, defining a new way of sampling arms by running it as a subroutine. We denote the first arm sampled by top-two DTS by $\hat{I}_t$ and call this the “leader”. Denote the second arm sampled by $\hat{J}_t$ and call this the challenger. The overall sampling probabilities obey the formula

$$P(I^* = i | H_t) = \beta P(I^* = i | H_t) + (1 - \beta) \sum_{j \neq i} P(I^* = j)P(I^* = i | I^* \neq j, H_t).$$

To understand the intuition behind this modification, consider a scenario in which the DM is 95% confident that in the identify of the optimal arm; For instance, $P(I^* = 1 | H_t) = 0.95$. In such scenarios, standard DTS plays arm 1 95% of the time, rarely gathering information about other arms. The top-two modification encourages the algorithm to more aggressively explore the most promising challengers to arm 1. This change can reduce the length of experiment (i.e. $T$ in our formulation) required to reach very high confidence.

A burgeoning body of theory establishes senses in which this kind of procedure is asymptotically optimal [Russo, 2020, Qin et al., 2017, Shang et al., 2020, Jourdan et al., 2022]. Most of that theory involves problems without contexts, but a companion to this paper studies asymptotic efficiency of top-two DTS in problems with contextual variation.

**Beyond Gaussian noise.** Our results require a Gaussian prior and noise. This case is especially tractable analytically, allowing for an especially efficient implementation of DTS that avoids the need for approximate posterior sampling. However, we conjecture that an analogue of our theoretical results should hold more generally. An analogue of Proposition 3, in the appendix, holds when reward noise is sub-Gaussian and the norm of $\theta$ is bounded almost surely. But the proof of Theorem 1 relies on the analytical form of the Guassian posterior and we do not know how to generalize it.

**Choosing a prior.** The choice of a bandwidth parameter in the prior displayed in Figure 1, for instance, is a delicate choice. Yet, most choices are likely to offer more robustness than applying vanilla Thompson sampling, an extreme special case of that prior under which is there no nonstationarity in rewards.

One possibility is to set prior parameters using data from past experiments. An online retailer who regularly conducts pricing experiments can use data from these past experiments to calibrate hyper-parameters governing the structure and severity of plausible nonstationarity. For more insights into this ‘empirical Bayesian’ perspective, refer to Azevedo et al. [2019], Dimmery et al. [2019], Bastani et al. [2022], and McDonald et al. [2023].

**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.

Y. Abbasi-Yadkori, A. Gyorgy, and N. Lazic. A new look at dynamic regret for non-stationary stochastic bandits. *arXiv preprint arXiv:2201.06532*, 2022.

K. Adusumilli. Risk and optimal policies in bandit experiments, 2023.

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

B. Amadio. Multi-armed bandits and the stitch fix experimentation platform, 2020. URL https://multithreaded.stitchfix.com/blog/2020/08/05/bandits/. Accessed: May 31, 2023.

V. F. Araman and R. A. Caldentey. Diffusion approximations for a class of sequential experimentation problems. *Management Science*, 68 (8):5998–5997, 2022.
We illustrate two very different models of exogenous variation that can be viewed as special cases of our general problem formulation. The first example considers a bandit experiment where a single observable factor — a user’s country — explains the non-stationary pattern of rewards. Of course, this is a simplified example. One may include many other observable features and also create more intricate models that combine observable factors with the latent ones modeled in Example 1.

### Example 4 (Comprehensible observed contexts)

A video streaming website is testing a small change to the layout of its homepage. The platform operates in $d$ different countries, and the context $X_t \in \{e_1, \ldots, e_d\}$ is a standard basis vector that encodes a user’s country. We assume this is a recorded feature.

The target population context vector $x_{\text{pop}} = (x_{\text{pop},1}, \ldots, x_{\text{pop},d}) \in \mathbb{R}^d$ measures the long-term fraction of user visits among those who hail from each country. The platform estimates this by querying a database that records all user visits over the past several months. (Implicit in this approach is an assumption that $x_{\text{pop}}$ is a reasonable reflection of the users who will visit over the next few months.)

Individuals tend to visit this video streaming website between 7pm-11pm in their local timezone. Due to timezone differences, the mix of countries among users arriving during a particular hour within the experiment may not reflect...
As the volume of data grows, the prior washes away and country/arm-specific means are estimated through an empirical arm in the first place.

I results from the necessity to experiment in order to learn I earned under the utilitarian optimal arm I of the chosen arm

Define the contextual regret of an algorithm

$$
\Delta_t(X_t) = r_\theta(I^*, X_t) - r_\theta(I_t, X_t)
$$

where

$$
\Delta_t = \frac{1}{T} \sum_{i=1}^{d} x_{pop,i} \mathbb{E}\left[\theta_{\ell}^{(i)} | H_t\right]
$$

As the volume of data grows, the prior washes away and country/arm-specific means are estimated through an empirical averaging. The population average reward is estimated in (18). This is a Bayesian analogue of a very common technique known as post-stratification.

The next example illustrates that it is possible to combine the modeling approaches taken in Examples 4 and 1.

Example 5 (A mixture of latent and observed factors). The context at time t is a tuple $X_t = (1, Z_t, e_t) \in \mathbb{R}^{1+p+T}$, where the vector $Z_t \in \mathbb{R}^p$ encodes other observable user features, like the country in Example 4, and $e_t \in \mathbb{R}^T$ is the $t^{'th}$ standard basis vector and indicates the current time period. Take $x_{pop} = (1, z_{pop}, \frac{1}{T} \sum_{t=1}^{T} e_1)$, where $z_{pop}$ is a population effect. Rather than specify a prior mean and covariance over the latent parameters $\theta = (\theta^{(1)}, \ldots, \theta^{(p)})$, it is more interpretable to write $\theta^{(i)} = (\alpha^{(i)}, \gamma^{(i)} + \beta, \epsilon)$ and specify a prior mean and covariance for jointly Gaussian latent parameters $(\alpha^{(i)})_{i \in [k]} \in \mathbb{R}^k, (\gamma^{(i)})_{i \in [k]} \in \mathbb{R}^{kp}, \beta \in \mathbb{R}^p$, and $\epsilon = (e_t)_{t \in [T]} \in \mathbb{R}^T$. Under this model, potential arm reward

$$
R_{i,t} = \alpha^{(i)} + \gamma^{(i)} Z_t + \beta Z_t + \epsilon_t + W_{i,t}
$$

is determined by an arm-specific effect, an interaction effect $\gamma^{(i)} Z_t$ between an arm-specific parameter and the observable user features, and arm-shared effects explained by observable user features or a latent time trend. A prior $\gamma^{(i)} \sim N(0, \lambda^2 I)$ where $\lambda$ is a small scalar causes the decision-maker to shrink the posterior mean of arm-specific parameters toward zero. The population mean reward of an arm,

$$
r_{\theta}(i) = \alpha^{(i)} + \gamma^{(i)} z_{pop} + \beta z_{pop} + \frac{1}{T} \sum_{t=1}^{T} e_t
$$

measures how arm i would have performed in hindsight over the past T periods among a cohort of users whose average observable features match $z_{pop}$.

B Additional theoretical theoretical guarantee: a bound on contextual regret

Define the contextual regret of an algorithm $\Delta_t(X_t) = r_\theta(I^*, X_t) - r_\theta(I_t, X_t)$ to be the shortfall in performance of the chosen arm $I_t$ in some context within the experiment, relative to the reward that would have been earned under the utilitarian optimal arm $I^*$. In fact, $\mathbb{E}[\Delta_t(X_t)] = \mathbb{E}[R_{I^*} - R_{I_t}]$.

Bounds on cumulative contextual regret can be interpreted as a limit on the decrease in reward that results from the necessity to experiment in order to learn $I^*$. Remark 6 below highlights that caution is needed when comparing algorithms in terms of their contextual regret, as it is possible to attain negative contextual regret by systematically violating the reasons the experimenter aimed to deploy a stable treatment arm in the first place.
Somewhat remarkably, the next result bounds the shortfall in reward accrued under DTS in terms of the number of actions, placing no conditions at all on the dimension of the context space or the pattern of nonstationarity in rewards that the context sequence may induce.

**Proposition 3 (Bound on within-experiment contextual regret).** Fix any context sequence \( x_{1:T} \in \mathcal{X}^T \). If \( L = 1 \) (no observation delay), then under DTS,

\[
\mathbb{E} \left[ \frac{\sum_{t=1}^T \Delta_t(X_t) | X_{1:T} = x_{1:T}}{T} \right] \leq c_R \sqrt{\frac{2k \log(k)}{T}},
\]

where \( c_R^2 = \max_{t \in [T], j \in [k]} \text{Var}(R_{t,j} | X_t = x_t) \).

*Proof.* The proof follows the information-theoretic analysis of Russo and Van Roy [2016]. While that paper studies vanilla Thompson sampling in i.i.d. environments, the proof applies without substantial changes to DTS in nonstationary environments.

We use \( \mathbb{H}(Z) \) to denote the entropy of a random variable \( Z \) and \( \mathbb{I}(Z_1, Z_2) \) to denote mutual information between \( Z_1 \) and \( Z_2 \). Let \( G_t = \mathbb{I}_{\mathbb{P}(\cdot | H_t, X_t)}(I^*; (I_t, R_{t,j})) \) be the mutual information (or ‘information gain’) between \( I^* \) and the observation \( (I_t, R_{t,j}) \) under the conditional probability measure \( \mathbb{P}(\cdot | H_t, X_t) \). This is a random variable due to the randomness in \( \mathbb{P}(\cdot | H_t, X_t) \). The convention in information theory is to integrate over that randomness, with conditional mutual information defined as \( \mathbb{I}(I^*; (I_t, R_{t,j}) | H_t, X_t) = \mathbb{E}[G_t] \).

Following Proposition 3, and Corollary 1, of Russo and Van Roy [2016], the probability matching property of DTS, \( \mathbb{P}(I_t = i | H_t, X_t) = \mathbb{P}(I^* = i | H_t, X_t) \) implies the following bound on the so-called ‘information ratio’:

\[
\Gamma_t = \frac{(\mathbb{E}[R_{t,I^*} - R_{t,I} | H_t, X_t])^2}{\mathbb{I}_t(I^*; (I_t, R_{t,j}))} \leq 2c_R^2 k \triangleq \Gamma.
\]

Re-arranging this expression summing over \( t \)

\[
\mathbb{E} \left[ \sum_{t=1}^T \Delta_t(X_t) \right] = \mathbb{E} \left[ \sum_{t=1}^T R_{t,I^*} - R_{t,I} \right] = \mathbb{E} \left[ \sum_{t=1}^T R_{t,I^*} - R_{t,I} | H_t, X_t \right] = \mathbb{E} \left[ \sum_{t=1}^T \sqrt{\Gamma_t} G_t \right] \leq \sqrt{T \cdot \Gamma \cdot \mathbb{E} \left[ \sum_{t=1}^T G_t \right]}.
\]

Now we show that expected cumulative information gain is bounded by prior entropy. We have,

\[
\mathbb{E}[G_t] = \mathbb{I}(I^*; (I_t, R_{t,j}) | H_t, X_t) = \mathbb{I}(I^*; (X_t, I_t, R_{t,j}) | H_t) - \mathbb{I}(I^*; X_t | H_t) = \mathbb{I}(I^*; (X_t, I_t, R_{t,j}) | H_t),
\]

where the first equality uses the chain rule and the second uses that \( X_t \) is independent of \( \theta \). Now, since \( H_t = (X_t, I_t, R_{t,j}) \), the chain rule and non-negativity of conditional entropy imply,

\[
\mathbb{E} \left[ \sum_{t=1}^T G_t \right] = \sum_{t=1}^T \mathbb{I}(I^*; (X_t, I_t, R_{t,j}) | H_t) = \mathbb{I}(I^*; H_{T+1}) = \mathbb{H}(I^*) - \mathbb{H}(I^* | H_{T+1}) \leq \mathbb{H}(I^*).
\]

The final claim follows from using the coarse upper bound \( \mathbb{H}(I^*) \leq \log(k) \) and dividing by \( T \). \( \square \)

**Remark 6 (Care is needed when interpreting contextual regret).** Imagine treatment arms represent possible prices, rewards reflect revenue earned by displaying a price to a customer, and context observations are features of the customer. Suppose those customer features are predictive of the customer’s race. It is plausible that pricing based on race would increase revenue, but the company understands that this to be illegal, unethical, and reputationally damaging. For that reason, they seek to deploy a fixed price, \( \text{I_{pos}} \), after the experiment. In this setting, a bandit algorithm could attain low—even negative — within-experiment contextual regret by targeting its prices based on customer features. But then the algorithm’s decision-making within the experiment clearly goes against the way the company hopes to make decisions post-experiment. More examples like this are discussed in Appendix D.3.
C Discussion of adversarial nonstationary bandits

A special case of our formulation produces a Bayesian analogue of common adversarial bandit models [Auer et al., 2002b, Lattimore and Szepesvári, 2020]. Assume the context at time $t$ is the $t$th standard basis vector: $X_t = e_t \in \mathbb{R}^T$. The reward at time $t$ is then

$$R_{t,i} = \theta_t^{(i)} + W_{t,i},$$

If one chooses $x_{\text{pop}} = (1/T, \ldots, 1/T)$, then

$$I^* = \arg \max_{i \in [k]} \frac{1}{T} \sum_{t=1}^{T} \theta_t^{(i)}$$

is the best-arm in hindsight over the course of the experiment and per-period within-experiment contextual regret (See Appendix B),

$$\frac{1}{T} \sum_{t=1}^{T} \Delta_t(X_t) = \max_{i \in [k]} \frac{1}{T} \sum_{t=1}^{T} \theta_t^{(i)} - \frac{1}{T} \sum_{t=1}^{T} \theta_t^{(I^*)}$$

benchmarks the performance of selected arms within the experiment against the best fixed arm. This matches the performance measure in the adversarial bandit literature. Post-experiment utilitarian regret assesses whether the algorithm can select an arm $I_{\text{post}}$ at the end of the experiment whose hindsight performance is competitive with that of the hindsight-optimal arm $I^*$.

In this special case, our bound on contextual regret in Proposition 3 is then reminiscent of results in the adversarial bandit literature. Indeed that case, $O\left(\sqrt{k \log(k)/T}\right)$ regret bounds are well known, even when rewards are picked by an adversary [Auer et al., 2002b]. What distinguishes Proposition 3 is that it applies to a very different algorithm.

Our model and algorithm deviates from the adversarial bandit literature in two substantive ways. Both may allow the DM to write off arms with poor population-level performance earlier in the experiment than would be possible in a typical adversarial model:

1. A structured prior distribution over $\theta$ may restrict the form of nonstationarity that is plausible. Classical i.i.d. bandits are an extreme special case in which the covariance structure over $\theta_1, \ldots, \theta_T$ is degenerate. Other structured priors, like Example 1, would guide an algorithm like DTS to guard against particular forms of nonstationarity.

2. Our formulation accommodates rich contextual observations that capture features beyond the current time period. Example 4, presented in Appendix A, provides a simple illustration. In that example, it may be possible to infer an arm’s population-level performance early in the experiment.

It is also worth emphasizing that a bound on post-experiment regret, like Theorem 1, cannot be deduced from bonds on cumulative within-experiment contextual regret, like Proposition 3. So-called “online-to-batch” conversions do not work when the rewards sequence is nonstationary.

D Comparison to contextual bandits and extension to policy learning problems

D.1 Comparison to linear contextual bandit models

The information structure of our problem corresponds to that in a classical linear contextual bandit problem [Li et al., 2010, Agrawal and Goyal, 2013], but the learning objective differs. In a typical linear contextual bandit problem, the DM wishes to learn an optimal treatment-rule $\pi^*: \mathcal{X} \rightarrow [k]$ satisfying $\pi^*(x) \in \arg \max_{i \in [k]} \mathbb{E}[R_{t,i} | \theta, X_t = x]$ for each $x \in \mathcal{X}$. TS for contextual bandit problem selects an arm $I_t$ at time $t$
randomly with sampling probabilities,
\[ P(I_t = i \mid H_t, X_t) = P(\pi^*(X_t) = i \mid H_t, X_t), \]
which are matched to the posterior distribution of the reward maximizing arm \textit{in the current context}.

Practitioners are, inevitably, faced with question of which features to include in the context vector \( x \in \mathcal{X} \). If one is using contextual TS, including a feature has two implications:

**Inference:** Including a feature in the context vectors directs the algorithm to ‘control’ for past variation in this feature when making inferences about the reward an arm will generate in the future.

**Reactivity:** Including a feature in the context vectors directs the algorithm to segment decisions it makes on the basis of this feature. In practice, this could mean that individuals who are different along this dimension receive different treatments or that, across several interactions, an individual receives different treatments as this feature changes.

Under our model, these two issues are decoupled. Deconfounded Thompson sampling is designed to account for contextual variation when performing inference while still learning a population level decision-rule that is not reactive to context. Appendix E presents an example in which contextual TS fails to gather the information necessary to select a good population-level arm \( I_{post} \); simply put, its exploration is directed toward a different goal.

Why would an experimenter aim to learn a policy that does not react to (some components of) an observed context? One reason, which cuts across applications, is that this can vastly reduce data requirements. Beyond this, we provide a substantive discussion in Section D.3. We now extend DTS to react to \textit{particular} components of the context.

### D.2 Generalization of DTS

We sketch a generalization of DTS which aims, 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 \([k]\). Overloading notation, take

\[ r_\theta(\pi) = \int_{\mathcal{X}} r_\theta(\pi(x), x) dD_{pop}(x) \]  \hspace{1cm} (21)

to be the average reward accrued by \( \pi \) under the population, generalizing (4). Define \( \pi^* \in \arg \max_{\pi \in \Pi} r_\theta(\pi) \) to the policy within the policy class which maximizes average reward under the true parameter \( \theta \). To simplify the presentation, assume that this maximum exists and is unique almost surely.

This objective can interpolate between two extremes:

1. **Complete standardization:** The policy class \( \Pi = \{\pi^{(1)}, \ldots, \pi^{(k)}\} \) has just \( k \) elements. Each \( \pi^{(i)} \) maps any \( x \in \mathcal{X} \) to \( \pi(x) = i \), corresponding to a decision-rule that does not segment its decisions on the basis of context. This models the hypothetical A/B test considered in the introduction and recovers our formulation in Section 2.

2. **Complete personalization:** This policy class \( \Pi \) contains all possible functions mapping \( \mathcal{X} \) to \([k]\). This is a common formulation in contextual bandit models [Li et al., 2010]. Given perfect knowledge of \( \theta \), the optimal policy plays \( \pi^*(x) \in \arg \max_{i \in [k]} r_\theta(i, x) \). In this sense, optimal decision-making completely decouples across contexts.

The next two examples illustrate policy classes in between these two extremes.

**Example 6 (Segmentation).** The context space is divided into \( m \) disjoints segments as \( \mathcal{X} = \mathcal{X}_1 \cup \cdots \cup \mathcal{X}_m \). Segments may, for instance, represent distinct geographical regions. The policy class \( \Pi \) consists of all rules \( \pi : \mathcal{X} \to [k] \) obeying for each segment \( j \) the constraint \( \pi(x) = \pi(x') \) for all \( x, x' \in \mathcal{X}_j \). That is, the policy class consists of rules that associate each segment with an action.
Example 7 (Protected features). A context \( x = x_{1:d} = (x_1, \ldots, x_d) \) is divided into two parts. The policy can be react to the first \( d_0 \) features when selecting actions but features \( x_{d_0+1:x} \) are protected attributes that may only be used to deconfound inferences when looking at past data. Formally, the policy class

\[
\Pi = \left\{ \pi : \mathcal{X} \to [k] \mid x_{1:d_0} = x'_{1:d_0} \implies \pi(x) = \pi(x') \right\}
\]

consists of all decision-rules whose output is invariant to the protected attributes.

Natural justifications for constraining \( \Pi \) are discussed at length in Section D.3.

Let us generalize DTS to treat such problems. We view DTS as a rule for selecting a sequence of policies \( (\pi_1, \ldots, \pi_T, \pi_{\text{post}}) \). Within the experiment, the arm selected at time \( t \) is determined as \( I_t = \pi_t(X_t) \); one could equivalently view DTS as a rule for selecting these arms. In constructing the reward measure \( r_\theta(\pi_{\text{post}}) \), we implicitly assume post-experiment decisions are by applying \( \pi_{\text{post}} \) to the observed context. The model defining reward realizations within the experiment is the same as before. Building on the definitions of DTS in (6) and (7), generalized DTS randomly samples a policy at time \( t \) according to

\[
\mathbb{P}(\pi_I = \pi \mid H_I) = \mathbb{P}(\pi^* = \pi \mid H_I) \quad \forall \pi \in \Pi
\]

and selects a policy to deploy in the population according to

\[
\pi_{\text{post}} \in \arg \max_{\pi \in \Pi} \mathbb{E}[r_\theta(\pi) \mid H_{\text{post}}].
\]

With a completely standardized policy class, this algorithm is DTS. With a completely personalized policy class, it is the standard definition of Thompson sampling in contextual bandits. This is a purely intellectual definition of the algorithm and whether it can be implemented efficiently depends on the structure of the policy class and reward model.

The next result generalizes Proposition 3, which bounds the the within-experiment contextual regret of DTS. It depends on the entropy of the optimal policy, which is always bounded as \( H(\pi^*) \leq \log(\lvert \Pi \rvert) \). Under complete standardization, \( H(\pi^*) \leq \log(k) \), recovering Proposition 3. Under complete personalization, entropy scales with the dimension of the feature vectors, and this proposition roughly yields a bound on the order of \( \sigma \sqrt{kd/T} \). In between these extremes, the entropy term reflects the complexity of the policy class. Similar results that depend on the logarithm of the size of the policy class, rather than entropy, are known in the non-stochastic bandit literature [Beygelzimer et al., 2011]. The novelty in this result is in providing a similar guarantee for a very different type of algorithm, using a different (information-theoretic) proof technique.

Proposition 4 (Generalized within-experiment contextual regret). Assume \( L = 1 \) (no observation delay). Furthermore, assume the policy class \( \Pi \) is finite. Define the within-experiment contextual regret by \( \Delta_t(X_t) = r_\theta(\pi^*(X_t), X_t) - r_\theta(\pi_t(X_t), X_t) \). Then,

\[
\mathbb{E} \left[ \frac{\sum_{t=1}^T \Delta_t(X_t) \mid X_{1:T} = x_{1:T}}{T} \right] \leq \sigma_R \sqrt{\frac{2 \cdot k \cdot H(\pi^*)}{T}}.
\]

where \( \sigma_R^2 = \sup_{t \in [T], i \in [k]} \text{Var}(R_{t,i} \mid X_t = x_i) \).

This result offers some assurances, but, unfortunately, we do not know how to extend our analysis of utilitarian regret in Theorem 1 to analyze this generalized form of DTS. Here is a short proof sketch; the same kind of argument was used to prove Lemma 4.12 of Min and Russo [2023].

Proof. The proof is the same as that of Proposition 3. We detail only the changes and do not rewrite the proof. Define \( I^*_t = \pi^*(X_t) \). The probability matching property with respect to policies in (22) implies that \( \mathbb{P}(I_t = i \mid X_t, H_t) = \mathbb{P}(I^*_t = i \mid X_t, H_t) \). Using this, we have the following bound on the so-called
‘information ratio’:
\[
\Gamma_t \triangleq \left( \frac{E[R_{t,I^*_t} - R_{t,I_t} | H_t, X_t]}{I_t(\pi^*; (I_t, R_{t,I_t}))} \right)^2 \leq \left( \frac{E[R_{t,I^*_t} - R_{t,I_t} | H_t, X_t]}{I_t(I_t^*; (I_t, R_{t,I_t}))} \right)^2 \leq 2\sigma^2_R k \triangleq \Gamma,
\]
where the first step is the data processing inequality. The second step is the same as in (20) in the proof of Proposition 3 and follows using the argument as in Proposition 3 and Corollary 1, of Russo and Van Roy [2016]. From here we use the same argument as in the proof of Proposition 3, but now we define the information gain \( G_t = I_t(\pi^*; (I_t, R_{t,I_t})) \) as being relative the optimal policy \( \pi^* \) rather than the optimal arm \( I_t^* \).

D.3 Reasons for standardization

Continuing the discussion above, we might say that DTS implements a standardized decision at the end of an experiment, since the arm \( I_{\text{post}} \) is applied across all future contexts. 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 described in Figure 1, 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.\(^7\) 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 in Figure 1, 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.

Most of these considerations cannot be captured through a policy-level reward function in the form (21). Rather than modify the objective function, we have incorporated them via constraints on the policy class.

\(^7\)https://www.computerworld.com/article/2588337/amazon-apologizes-for-price-testing-program-that-angered-customers.html
E Failure of alternative algorithms

E.1 Failure of deconfounded UCB: Proof of Lemma 2

We begin by restating the claim in Section 6.

**Lemma 2 (Failure of deconfounded UCB).** Consider Example 3. Suppose that \( \theta^{(1)}_x \sim N(0, 1) \) and \( \theta^{(2)}_x \sim N(0, 3) \) for \( x \in [2] \). If, for any fixed \( z > 0 \), \( I_t \in \arg \max_{i \in [k]} \mathbb{E}[r_{\theta}(i) \mid H_t] + z \sqrt{\text{Var}(r_{\theta}(i) \mid H_t)} \) holds for every \( t \), there is an absolute numerical constant \( c > 0 \) such that for all \( T \in \mathbb{N} \), \( \mathbb{E} \left[ \Delta_{\text{post}} \right] \geq c \).

**Proof.** Let \( U_{t,i} = \mathbb{E}[r_{\theta}(i) \mid H_t] + z \cdot \sqrt{\text{Var}(r_{\theta}(i) \mid H_t)} \) denote the UCB of arm \( i \). Since \( U_{t,2} > U_{t,1} \), the initial arm selection is \( I_1 = 2 \). When \( \theta^{(2)}_1 > 0 \), the posterior mean and standard deviation satisfy \( m_{2,2} = \frac{\theta^{(2)}_2}{2} > 0 = m_{2,1} \) and \( s_{2,2} = \frac{1}{2} \sqrt{\text{Var} \left( \theta^{(2)}_1 \mid H_2 \right) + \text{Var} \left( \theta^{(2)}_2 \mid H_2 \right)} = \sqrt{\frac{3}{2}} > \sqrt{\frac{1}{2}} = s_{2,1} \). This implies \( U_{2,2} > U_{2,1} \) and arm \( I_2 = 2 \) is again selected. This process repeats, showing that if \( \theta^{(2)}_1 > 0 \), then arm 2 is chosen for each of the first \( \left[ \frac{T}{2} \right] \) periods. We can lower bound simple regret by imagining that the decision-maker has perfect knowledge of \( \theta^{(2)}_1, \theta^{(2)}_2 \) and \( \theta^{(1)}_2 \) when selecting \( I_{\text{post}} \in \arg \max_{i \in [2]} \mathbb{E} \left[ \frac{\theta^{(i)} + \theta^{(i)}_1}{2} \mid H_{\text{post}} \right] \), resulting in:

\[
\mathbb{E} \left[ \Delta_{\text{post}} \right] \geq \mathbb{E} \left[ \max \left( \frac{\theta^{(1)}_1 + \theta^{(2)}_1}{2}, \frac{\theta^{(2)}_1 + \theta^{(2)}_2}{2} \right) - \max \left( \frac{\theta^{(2)}_1}{2}, \frac{\theta^{(2)}_1 + \theta^{(2)}_2}{2} \right) \right] > 0.
\]

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. \( \square \)

E.2 Failure of context-unaware algorithms

Section 5 showed that a context-unaware version of Thompson sampling can fail. Here, we make that observation formal, just as we have for deconfounded UCB.

We define context-unaware Thompson sampling to be an arm algorithm that chooses arm at time \( t \) according to

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

where \( \tilde{m}_{t,i} \) and \( \tilde{s}_{t,i}^2 \) are parameters of a pseudo-posterior, defined below. In (24), \( v_{t,1}, v_{t,2}, \ldots, v_{t,k} \) are sampled independently.

The pseudo-posterior is updated as if observations were i.i.d. From the algebra of Bayes rule for Gaussian, when \( \sigma^2 > 0 \), we define this as

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

where \( \tilde{s}_{i}^{-2} > 0 \) is some initial value. The natural definition when there is no observation noise (i.e. \( \sigma^2 = 0 \)) is derived 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.

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. The proof is provided at the end of this subsection.
\textbf{Lemma 3} (Failure of context-unaware TS). Consider Example 3, presented in Section 6. 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 $x \in [2]$. Let $L = 1$ (no delay). If (24) holds, there exists an absolute numerical constant $c > 0$ such that for all $T \in \mathbb{N}$, $\mathbb{E}[\Delta_{\text{post}}] \geq c$.

The next remark interprets the failure of context-unaware TS in terms of confounding, using the potential outcomes formalism of Rubin [1979].

\textbf{Remark 7} (Interpretation as confounding). One can view the failure of context-unaware TS as being driven by ‘confounding’ due to omitted contextual variables. Let $\tau$ denote the event that the inequality holds due to Condition 3 above. Conditioned on the event that $\tau$, we have

$$I_{\tau} \perp (R_{\tau,i} : i \in [k]) \mid X_{\tau}.$$  

Conditioned on the context, the chosen arm is independent of potential reward outcomes. But context-unaware TS performs inferences without conditioning on contexts, and due to the co-occurring patterns in the contexts sequence and the sequence of chosen arms

$$I_{\tau} \not\perp (R_{\tau,i} : i \in [k]).$$

We now prove Lemma 3.

\textbf{Proof of Lemma 3}. It is not hard to show that $\mathbb{E}[\Delta_{\text{post}}] > 0$ 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 $l^* = 2$

2. $\min \{\theta_1^{(1)}, \theta_2^{(1)}\} > \theta_1^{(2)}$: 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 $v_{1,1} \sim N\left(\bar{m}_{1,1}, \bar{s}_{1,1}^2\right)$ and $v_{1,2} \sim N\left(\bar{m}_{1,2}, \bar{s}_{1,2}^2\right)$ denote the sampled parameters, and we denote the probability of playing arm 1 by

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

Conditioned on the event that $\theta \in \Theta'$ and $l_1 = 1$, we have $(\bar{m}_{2,1}, \bar{s}_{2,1}^2) = (\theta_1^{(1)}, 0)$ and the probability of playing arm 2 in the second period is

$$\mathbb{P}\left(v_{2,2} > \theta_1^{(1)} \mid \theta \in \Theta', l_1 = 1\right) \geq \frac{1}{2}$$

where the inequality holds due to Condition 3 above. Conditioned on the event that $\theta \in \Theta'$, $l_1 = 1$ and $l_2 = 2$, we have $(\bar{m}_{3,1}, \bar{s}_{3,1}^2) = (\theta_1^{(1)}, 0)$ and $(\bar{m}_{3,2}, \bar{s}_{3,2}^2) = (\theta_2^{(2)}, 0)$. Due to Condition 2, TS will always measure arm 1 afterwards. Hence,

$$\mathbb{E}[\Delta_{\text{post}}] \geq \mathbb{E}\left[\left(\frac{\theta_1^{(2)} + \theta_2^{(2)}}{2} - \frac{\theta_1^{(1)} + \theta_2^{(1)}}{2}\right) \mathbb{1}(\theta \in \Theta') \mathbb{1}(l_1 = 1, l_2 = 2)\right]$$

$$\geq \frac{c_0}{2} \mathbb{E}\left[\left(\frac{\theta_1^{(2)} + \theta_2^{(2)}}{2} - \frac{\theta_1^{(1)} + \theta_2^{(1)}}{2}\right) \mathbb{1}(\theta \in \Theta')\right] > 0.$$  

This completes the proof. \qed
E.3 Failure of contextual bandit algorithms

The goal in our formulation is to select one 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 8 seems already to capture the main intuition.

Example 8 (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 \( D_{\text{pop}} \) is uniform over \( \mathcal{X} \) and \( (X_t)_{t \in \mathbb{N}} \) are drawn i.i.d. from \( D_{\text{pop}} \). For the first two arms \( (i \in [2]) \) and \( x \in [2] \), it holds almost surely that

\[
\theta_x^{(i)} = 1(x = i) - 1(x \neq i).
\]

The third arm \( (i = 3) \), is insensitive to context, with \( \theta_1^{(3)} = \theta_2^{(3)} = U \) where \( U \sim \text{Uniform}[0,1] \). Rewards are noiseless, with \( R_{t,i} = r_{\theta}(I_t, X_t) = \langle \theta^{(i)}, X_t \rangle \). Hence, if \( X_t = e_1 \), then \( R_{t,i} = \theta_1^{(i)} \); otherwise \( R_{t,i} = \theta_2^{(i)} \). Observations are not subject to delay (i.e. \( L = 1 \)).

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 \( \theta^{(3)} \) 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 (i.e. those that select the same arm, irrespective of context), 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 4 (Failure of contextual TS). Consider Example 8. Contextual TS at time \( t \) chooses an arm \( I_t \) such that for each \( i \in [k] \), \( P(I_t = i \mid H_t, X_t) = P \left( \arg \max_{j \in [k]} r_{\theta}(j, X_t) = i \mid H_t, X_t \right) \). There is an absolute numerical constant \( c > 0 \) such that for all \( T \in \mathbb{N} \), \( E \left[ \Delta_{\text{post}} \right] \geq c \).

F Proof of Theorem 1

We begin by restating the theorem.

Theorem 1 (Bound on within- and post-experiment utilitarian regret). Fix any sequence \( x_{1:T} \in \mathcal{X}^T \). Under DTS, within-experiment regret is bounded as

\[
E \left[ \Delta_t \mid X_{1:T} = x_{1:T} \right] \leq \frac{2 \cdot t \cdot k \cdot \log(k)}{\text{Precision} \left( x_{1:(t-L)} \right)}, \quad \forall t \in [T],
\]

\( (13) \)
where \( i \) is defined in Equation (15). Post-experiment regret is bounded as

\[
\mathbb{E} \left[ \Delta_{\text{post}} \mid X_{1:T} = x_{1:T} \right] \leq \sqrt{\frac{2 \cdot \epsilon \cdot k \cdot \log(k)}{\text{Precision}(x_{1:T})}},
\]

(14)

The proof is broken into several parts.

**F.1 Proof of (14)**

The more delicate result is (13), with (14) following essentially as a corollary. Notice that the right hand side of (14) matches the right hand side of (13) if we could set \( t = T + L \).

**Argument deriving (14) from (13).** Recall that DTS’s decision at time \( t \) does not depend on the context at time \( t \) or even contexts in the past \( L - 1 \) periods (see (8)). Let \( \tilde{H}_t \triangleq (X_{1:(t-L)}, I_1:(t-L), R_1:(t-L)) \) be the effective history used by DTS at time \( t \in [T] \) (where \( \tilde{H}_t = \emptyset \) for \( t \in [L] \)). With some abuse of notation, extend the definition of \( \tilde{H}_t \) for \( t \in \{T+1, \ldots, T+L\} \) as \( \tilde{H}_t = (X_{1:(t-L)}, I_1:(t-L), R_1:(t-L)) \). Recall that for all \( t \in [T] \), the definition of DTS is that \( \mathbb{P} (I_t = i \mid \tilde{H}_t) = \mathbb{P} (I^* = i \mid \tilde{H}_t) \). Extend this definition for \( t \in \{T+1, \ldots, T+L\} \).

Define the greedy decision based on \( \tilde{H}_t \) at time \( t \in [T+L] \) by

\[
I^*_t \in \arg \max_{i \in [k]} \mathbb{E} \left[ r_\theta(i) \mid \tilde{H}_t \right].
\]

Recall that \( I_{\text{post}} \) is the arm chosen by DTS at the end of the experiment. Since \( H_{\text{post}} = (X_{1:T}, I_1:T, R_1:T) = \tilde{H}_{T+L} \), we have \( I_{\text{post}} = I^*_{T+L} \).

Now, for \( t \in [T+L] \), define the two performance measures

\[
\Delta_t^{\text{explo}} = r_\theta(I^*) - r_\theta(I_t) \quad \text{and} \quad \Delta_t^{\text{greedy}} = r_\theta(I^*) - r_\theta(I^*_t).
\]

The expected regret of the greedy decision is always smaller:

\[
\mathbb{E} \left[ \Delta_t^{\text{greedy}} \right] = \mathbb{E} \left[ \mathbb{E} \left[ r_\theta(I^*) - r_\theta(I^*_t) \mid \tilde{H}_t \right] \right] \leq \mathbb{E} \left[ \mathbb{E} \left[ r_\theta(I^*) - r_\theta(I_t) \mid \tilde{H}_t \right] \right] = \mathbb{E} \left[ \Delta_t^{\text{explo}} \right].
\]

(25)

Here to simplify this argument, assume \( X_{1:T} \) is an arbitrary deterministic sequence, equal to some fixed \( x_{1:T} \) almost surely. Since it is deterministic, we do not need to condition on it in expectations.

A careful reading of the proof\(^8\) of (13) reveals that it applies to bound

\[
\mathbb{E} \left[ \Delta_t^{\text{explo}} \right] \leq \sqrt{\frac{2 \cdot \epsilon \cdot k \cdot \log(k)}{\text{Precision}(x_{1:(t-L)})}},
\]

even for \( t \in \{T+1, \ldots, T+L\} \). Note that that expected within-experiment regret is \( \mathbb{E}[\Delta_t] = \mathbb{E} \left[ \Delta_t^{\text{explo}} \right] \) and expected post-experiment regret is \( \mathbb{E}[\Delta_{\text{post}}] = \mathbb{E} \left[ \Delta_t^{\text{greedy}} \right] \). Then picking \( t = T + L \) and combining this with (25) yields (14).

\( \square \)

**F.2 Proof of Proposition 1**

The first key to establishing Theorem 1 is Proposition 1, restated below. This reduces the problem of controlling the utilitarian regret to the problem of controlling the expected posterior variance of the optimal arm. Recall that \( s_{t,i} = \sqrt{\text{Var}(r_\theta(i) \mid \tilde{H}_t)}. \)

\(^8\)It is not proper style to cite a proof rather than a result. In this case, the modification is quite simple, though. Follow the exact same steps, but interpret \( t \) as possibly falling in the range \( \{T+1, \ldots, T+L\} \).
Proposition 1 (Reduction to estimation). Under DTS, for any $t \in [T]$, 

$$
\mathbb{E} [\Delta_t \mid H_t, X_{1:T}] \leq \sqrt{2 \log(k) \mathbb{E} \left[ s^2_{l_t^*} \mid H_t, X_{1:T} \right]} \quad \text{and} \quad \mathbb{E} [\Delta_t \mid X_{1:T}] \leq \sqrt{2 \log(k) \mathbb{E} \left[ s^2_{l_t^*} \mid X_{1:T} \right]}.
$$

Proof. In this proof, we avoid writing conditional expectations by letting $X_{1:T} = x_{1:T} \in \mathcal{X}^T$ with probability 1 for some arbitrary sequence $x_{1:T}$.

We focus on proving the bound on $\mathbb{E} [\Delta_t]$. Define $Z_i = r_\theta(i)$ to be the uncertain population performance of arm $i$, $m_{t,i} = \mathbb{E} [Z_i \mid H_t]$ to be its posterior mean, and $s^2_{t,i} = \text{Var} (Z_i \mid H_t)$ to be its posterior variance. The notation $Z_i$ and $m_{t,i}$ is used only in this proof. Note that $Z_i \mid H_t \sim N \left( m_{t,i}, s^2_{t,i} \right)$. Take $Z = (Z_1, \ldots, Z_k)$ to be the vector. Under DTS, $I_t$ is a sample from the posterior, i.e. $\mathbb{P} (I_t = i \mid H_t) = \mathbb{P} (I^* = i \mid H_t)$ but $I_t$ is independent of $(Z_1, \ldots, Z_k)$ 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 $\mathbb{P} ( (Y_1, Y_2) \in \cdot \mid H_t )$. This is random, due to its dependence on the history. Taking expectations yields the usual definition of mutual information, with $\mathbb{E} [\mathbb{I}_{H_t} (Y_1; Y_2)] = \mathbb{I} (Y_1; Y_2 \mid H_t)$. This notation is used in this proof alone.

We have,

$$
\mathbb{E} [\Delta_t] = \mathbb{E} [Z_{l^*} - Z_{I_t^*}^*] = \mathbb{E} [Z_{l^*} - \mathbb{E} [Z_{I_t^*} \mid H_t]] \\
= \mathbb{E} [Z_{l^*} - \mathbb{E} [m_{t,I_t^*} \mid H_t]] \\
\overset{(a)}{=} \mathbb{E} [Z_{l^*} - \mathbb{E} [m_{t,I_t^*} \mid H_t]] \\
= \mathbb{E} [\mathbb{E} [Z_{l^*} - m_{t,I_t^*} \mid H_t]] \\
\overset{(b)}{\leq} \mathbb{E} \left[ \sqrt{\mathbb{E} [s^2_{l_t^*} \mid H_t]} \sqrt{2 \mathbb{I}_{H_t} (I^*; Z)} \right] \\
\overset{(c)}{\leq} \sqrt{\mathbb{E} \left[ \mathbb{E} [s^2_{l_t^*} \mid H_t] \right]} \sqrt{2 \mathbb{E} [\mathbb{I}_{H_t} (I^*; Z)]} \\
\overset{(d)}{= \frac{\mathbb{E} [s^2_{l_t^*}]}{\sqrt{2 \mathbb{I} (I^* \mid H_t)}}} \\
\leq \sqrt{\mathbb{E} [s^2_{l_t^*}] \sqrt{2 \mathbb{I} (I^* \mid H_t)}}.
$$

Early steps of the proof use the tower property of conditional expectation. Step (a) is crucial and uses that fact that $I_t$ and $I^*$ have the same distribution conditioned on $H_t$ and that the vector $(m_{t,1}, \ldots, m_{t,k})$ 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. The proposition uses the coarse upper bound $\mathbb{I} (I^* \mid H_t) \leq \log(k)$ to simplify the presentation. The bound on $\mathbb{E} [\Delta_t \mid H_t]$ follows from (c) using that $\mathbb{I}_{H_t} (I^*; Z) \leq \log(k)$. \qed

Lemma 5 (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 $\sigma^2_i$. Then

$$
\mathbb{E} [Z_i - \mu_i] \leq \sqrt{\mathbb{E} [\sigma^2_i]} \sqrt{2 \mathbb{I} (Z; I)}.
$$

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 $\mathbb{E} \left[ \max_{i \in [k]} |Z_i - \mu_i| \right] \leq (\max_{i \in [k]} \sigma_i) \sqrt{2 \log(n)}$. 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 $\log(n)$. 

37
F.3 Optionally sampled matrix-valued processes

One part of our proof (namely, Lemma 6) relies on a new result on optionally sampled matrix-valued processes. Stated in the abstract form below, one can view the positive definite matrix $V_t$ as generalized ‘reward’ or ‘value’ and $Z_t$ as a (randomized) decision of whether to collect that value. The result bounds the impact of randomization on the reward accrued.

Let $S^d$ denote the set of symmetric $d \times d$ square matrices and $S^d_+ \subset S^d$ denote the set of symmetric positive semidefinite matrices.

**Proposition 5** (Optionally sampled matrix-valued process). Consider a deterministic sequence of positive semidefinite matrices $V_1, V_2, \ldots \in S^d$ satisfying $\sup_{t \in \mathbb{N}} \lambda_{\text{max}}(V_t) \leq 1$, and a random process $(Z_t)_{t \in \mathbb{N}}$ taking values in $\{0, 1\}$ that is adapted to some filtration $(\mathcal{F}_t)_{t \in \mathbb{N}_0}$. For $n \in \mathbb{N}$, define

$$S_n = \sum_{t=1}^{n} Z_t V_t \quad \text{and} \quad \tilde{S}_n = \sum_{t=1}^{n} \mathbb{P}(Z_t = 1 \mid \mathcal{F}_{t-1}) V_t.$$

Then, for any $\delta > 0$, with probability exceeding $1 - \delta$,

$$S_n \geq (3 - e) \tilde{S}_n - \log \left( \frac{d}{\delta} \right) I, \quad \forall n \in \mathbb{N}.$$

**Proof.** See Section G for a complete proof. The analysis builds on a beautiful theory of the concentration of matrix-valued martingales by Tropp [2011].

F.4 Introducing a smoothed observation model

Our goal is to establish a regret bound by bounding $\mathbb{E} \left[ s_{i,t}^2 \right]$. As a first step toward this, we introduce a ‘smoothed’ observation model as a device in the analysis. In this model, arms can be played fractionally; when the algorithm picks an effort allocation $(p_{t,1}, \ldots, p_{t,k})$, it observes in response noisy reward signals $(\tilde{R}_{t,1}, \ldots, \tilde{R}_{t,k})$ where the standard deviation of $\tilde{R}_{t,j}$ is $\frac{c}{p_{t,j}}$. The notation $\tilde{S}_t, \tilde{\Sigma}_{t,j}$ and $\tilde{s}_{t,j}^2$ is used to denote posterior (co)variances under this smoothed model.

**Definition 1** (Smoothed observation model). Define $\tilde{R}_{t,j} = r_\Theta(i, X_t) + \frac{W_{i,t}}{p_{t,j}}$ so that $\tilde{R}_{t,j} \mid (H_t, p_{t,j}, \theta, X_t) \sim N \left( r_\Theta(i, X_t), \frac{\Sigma_{t,j}}{p_{t,j}} \right)$.

Set

$$\Sigma_t = \text{Cov} \left[ \theta \mid \left( X_{t, \ell}, (p_{\ell,j}, \tilde{R}_{\ell,j})_{j \in [k]} \right)_{\ell \in [t-L]} \right],$$

$$\Sigma_{t,j} = \text{Cov} \left[ \theta^{(i)} \mid \left( X_{t, \ell}, (p_{\ell,j}, \tilde{R}_{\ell,j})_{j \in [k]} \right)_{\ell \in [t-L]} \right],$$

$$\tilde{s}_{t,j}^2 = \text{Var} \left[ \langle \theta^{(i)}, x_{\text{pop}} \rangle \mid \left( X_{t, \ell}, (p_{\ell,j}, \tilde{R}_{\ell,j})_{j \in [k]} \right)_{\ell \in [t-L]} \right].$$

(As a warning, the notation $\tilde{s}_{t,j}$ means something different in Subsection E.2, where it is used to define a heuristic algorithm.) Posterior variances under the smoothed model are known functions of the chosen arm propensities and the context sequence. The next fact illustrates this for $\tilde{\Sigma}_t$. Define $\phi(x, i) = (0, \ldots, 0, x_1, \ldots, x_d, 0, \ldots, 0) \in \mathbb{R}^{dh}$ to be the concatenation of $k$ subvectors of size $d$, where the $i$th subvector is $x$. Other quantities of interest can be derived from $\tilde{\Sigma}_t$. For instance, $\tilde{s}_{t,j}^2 = \phi(x_{\text{pop}}, i) \top \tilde{\Sigma}_t \phi(x_{\text{pop}}, i)$.

---

9Direct application of that paper establishes a scalar inequality of the form $\lambda_{\text{max}}(\tilde{S}_n - S_n) \leq c \lambda_{\text{max}}(S_n) + \log \left( \frac{1}{\delta} \right)$. For our purposes Proposition 5 offers a critical improvement. It is able to provide a meaningful bound on $u \top \tilde{S}_n u$ even for directions $u \in \mathbb{R}^d$ for which $u \top \tilde{S}_n u$ is extremely small.
Fact 1. \( \Sigma_t \) obeys the formula
\[
\Sigma_t = \left( \Sigma_t^{-1} + \sigma^{-2} \sum_{\ell=1}^{t-L} \sum_{i=1}^k p_{\ell,i} \phi(X_{\ell,i}) \phi(X_{\ell,i})^\top \right)^{-1}
\]
whereas
\[
\Sigma_t = \left( \Sigma_t^{-1} + \sigma^{-2} \sum_{\ell=1}^{t-L} \sum_{i=1}^k \mathbb{1}(I_{\ell} = i) \phi(X_{\ell,i}) \phi(X_{\ell,i})^\top \right)^{-1}.
\]

The next result allows us to rigorously use the evolution of the posterior variance in the smoothed model to study the evolution of posterior covariance in the true model. It follows by applying Proposition 5 to our problem.

Lemma 6. For any \( \delta > 0 \),
\[
\mathbb{P} \left( \Sigma_t^{-1} \succeq (3 - \epsilon) \Sigma_t^{-1} - \sigma^{-2} \log \left( \frac{dk}{\delta} \right) I \quad \forall t \geq L, X_{1:T} \right) \geq 1 - \delta.
\]

Proof of Lemma 6. Most of the analysis uses precision matrices, rather than covariance matrices. Write the posterior precision matrix \( \Sigma_t^{-1} \) in the form.
\[
\Sigma_t^{-1} = \Sigma_t^{-1} + \sigma^{-2} \begin{pmatrix} \sum_{\ell=1}^{t-L} \mathbb{1}(I_{\ell} = 1) X_{\ell} X_{\ell}^\top & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sum_{\ell=1}^{t-L} \mathbb{1}(I_{\ell} = k) X_{\ell} X_{\ell}^\top \end{pmatrix} \triangleq \Sigma_t^{-1} + \sigma^{-2} S_t.
\]
The posterior precision matrix in the smoothed observation model \( \tilde{\Sigma}_t^{-1} \in \mathbb{R}^{dk \times dk} \) is defined by
\[
\tilde{\Sigma}_t^{-1} = \Sigma_t^{-1} + \sigma^{-2} \begin{pmatrix} \sum_{\ell=1}^{t-L} p_{\ell,1} X_{\ell} X_{\ell}^\top & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sum_{\ell=1}^{t-L} p_{\ell,k} X_{\ell} X_{\ell}^\top \end{pmatrix} \triangleq \Sigma_t^{-1} + \sigma^{-2} \tilde{S}_t.
\]
For a fixed \( i \in [k] \), we apply Proposition 5 with \( V_\ell = X_\ell X_\ell^\top \) and \( Z_\ell = \mathbb{1}(I_{\ell} = i) \), and \( \mathcal{F}_{\ell-1} = \sigma(H_\ell) \) taken to be the sigma algebra generated by the history. Proposition 5 applies, since \( \lambda_{\max}(V_\ell) = \| X_\ell \|_2^2 \leq 1 \) where the first equality is Fact 2 and the norm bound is an assumption in the problem formulation. Observe that \( p_{\ell,i} = \mathbb{P}(Z_\ell = i \mid \mathcal{F}_{\ell-1}) \). Hence, for any \( \delta' > 0 \), with probability exceeding \( 1 - \delta' \),
\[
\sum_{\ell=1}^{t-L} \mathbb{1}(I_{\ell} = i) X_{\ell} X_{\ell}^\top \succeq (3 - \epsilon) \left( \sum_{\ell=1}^{t-L} p_{\ell,i} X_{\ell} X_{\ell}^\top \right) - \log \left( \frac{d}{\delta'} \right) I, \quad \forall t \geq L.
\]
Taking \( \delta = \frac{\delta'}{k} \) and applying a union bound, we have that with probability exceeding \( 1 - \delta \),
\[
S_t \succeq (3 - \epsilon) \tilde{S}_t - \log \left( \frac{dk}{\delta} \right) I, \quad \forall t \geq L,
\]

39
where $S_t$ and $\delta_t$ are defined earlier in this proof. Then on the event that $5_t \geq (3 - e)\delta_t - \log \left( \frac{dK}{\delta} \right) I$, we have

$$
\Sigma_t^{-1} = \Sigma_t^{-1} + \sigma^{-2} S_t \succeq \Sigma_t^{-1} + (3 - e)\sigma^{-2} \delta_t - \sigma^{-2} \log \left( \frac{dK}{\delta} \right) I
$$

$$
= \Sigma_t^{-1} + (3 - e) \left( \Sigma_t^{-1} - \Sigma_t^{-1} \right) - \sigma^{-2} \log \left( \frac{dK}{\delta} \right) I
$$

$$
= (e - 2) \Sigma_t^{-1} + (3 - e) \Sigma_t^{-1} - \sigma^{-2} \log \left( \frac{dK}{\delta} \right) I
$$

$$
\geq (3 - e) \Sigma_t^{-1} - \sigma^{-2} \log \left( \frac{dK}{\delta} \right) I.
$$

This completes the proof. \hfill \Box

An adaptation of the high probability bound above a bound in expectation, here stated in terms of the scalar quantities $s_{t,i}^2$ and $\hat{s}_{t,i}^2$ that are needed in the analysis. The proof follows by a messy calculation.

**Corollary 2.** For any $t \geq L$,

$$
\mathbb{E} \left[ s_{t,i}^2 \mid \theta, X_{1:T} \right] \leq t \cdot \mathbb{E} \left[ \hat{s}_{t,i}^2 \mid \theta, X_{1:T} \right], \quad \forall i \in [k].
$$

In particular,

$$
\mathbb{E} \left[ \hat{s}_{t,i}^2 \mid \theta, X_{1:T} \right] \leq t \cdot \mathbb{E} \left[ \hat{s}_{t,i}^2 \mid \theta, X_{1:T} \right].
$$

**Proof.** The first step is to prove an inequality of the form $\Sigma_t \preceq c_\delta \Sigma_t$, which holds with high probability. In particular we prove that for any $\delta > 0$, conditioned on $X_{1:T}$ and $\theta$, with probability exceeding $1 - \delta$, the following inequality holds simultaneously for every $t \geq L$:

$$
\Sigma_t \preceq c_\delta \Sigma_t \quad \text{where} \quad c_\delta = 8 \cdot \max \left\{ \frac{-\sigma^2}{2}, \frac{\lambda_{\max}(\Sigma_t) \cdot \log \left( \frac{dK}{\delta} \right)}{1} \right\}.
$$

(26)

We know that $\Sigma_t^{-1} \succeq \Sigma_t^{-1}$. Combining this with Lemma 6 implies that for any arbitrary unit vector $u$,

$$
u^T \Sigma_t^{-1} u \geq \max \left\{ \lambda_{\min} \left( \Sigma_t^{-1} \right), (3 - e) u^T \Sigma_t^{-1} u - \sigma^{-2} \log \left( \frac{dK}{\delta} \right) \right\}.
$$

If $(3 - e) u^T \Sigma_t^{-1} u \geq 2 \sigma^{-2} \log \left( \frac{dK}{\delta} \right)$, we have $u^T \Sigma_t^{-1} u \geq \frac{3 - e}{2} u^T \Sigma_t^{-1} u$. On the other hand, if $(3 - e) u^T \Sigma_t^{-1} u \leq 2 \sigma^{-2} \log \left( \frac{dK}{\delta} \right)$, we have

$$
u^T \Sigma_t^{-1} u \geq \lambda_{\min} \left( \Sigma_t^{-1} \right) = \frac{\lambda_{\min} \left( \Sigma_t^{-1} \right)}{\sigma^{-2} \log \left( \frac{dK}{\delta} \right)} \cdot \sigma^{-2} \log \left( \frac{dK}{\delta} \right) \geq \frac{\lambda_{\min} \left( \Sigma_t^{-1} \right)}{\sigma^{-2} \log \left( \frac{dK}{\delta} \right)} \cdot \frac{3 - e}{2} \cdot u^T \Sigma_t^{-1} u.
$$

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

$$
u^T \Sigma_t^{-1} u \geq c_1 u^T \Sigma_t^{-1} u \quad \text{where} \quad c_1 = \min \left\{ \frac{\lambda_{\min} \left( \Sigma_t^{-1} \right)}{\sigma^{-2} \log \left( \frac{dK}{\delta} \right)}, 1 \right\} \cdot \frac{3 - e}{2}.
$$

We can simplify the expression using that $\frac{2}{\log e} < 8$. Viewing this as a relation of the form $\Sigma_t^{-1} \preceq \frac{1}{c_1} \Sigma_t^{-1} \preceq c_1 \Sigma_t^{-1} \preceq \Sigma_t^{-1} + \sigma^{-2} (t - L) I$, (which holds since $\lambda_{\max}(X_t X_t^T) = \|X_t\|^2 \leq 1$ by Fact 2). We have that for every
\(\delta > 0\)

\[
\mathbb{E} [\Sigma_t \mid \theta, X_{1:T}] = c_\delta \mathbb{E} [\Sigma_t \mid \theta, X_{1:T}] + \mathbb{E} [\Sigma_t (1 - \chi_d) \mid \theta, X_{1:T}]
\]

\[
\leq c_\delta \mathbb{E} [\Sigma_t \mid \theta, X_{1:T}] + \mathbb{E} [\Sigma_t (1 - \chi_d) \mid \theta, X_{1:T}]
\]

\[
\leq c_\delta \mathbb{E} [\Sigma_t \mid \theta, X_{1:T}] + \delta \Sigma_t
\]

\[
\leq \mathbb{E} [\Sigma_t \mid \theta, X_{1:T}] \left( c_\delta + \delta \frac{\lambda_{\text{max}} (\Sigma_t)}{\lambda_{\text{min}} (\mathbb{E} [\Sigma_t \mid \theta, X_{1:T}])} \right)
\]

\[
= \mathbb{E} [\Sigma_t \mid \theta, X_{1:T}] \left( c_\delta + \delta \lambda_{\text{max}} (\Sigma_t) \lambda_{\text{max}} \left( (\mathbb{E} [\Sigma_t \mid \theta, X_{1:T}])^{-1} \right) \right)
\]

\[
\leq \mathbb{E} [\Sigma_t \mid \theta, X_{1:T}] \left( c_\delta + \delta \lambda_{\text{max}} (\Sigma_t) \lambda_{\text{max}} \left( \Sigma_t^{-1} + \sigma^{-2} (t - L) I \right) \right)
\]

\[
= \mathbb{E} [\Sigma_t \mid \theta, X_{1:T}] \left( c_\delta + \delta \lambda_{\text{max}} (\Sigma_t) \left[ \lambda_{\text{max}} \left( \Sigma_t^{-1} \right) + \sigma^{-2} (t - L) \right] \right).
\]

Hence,

\[
\mathbb{E} [\Sigma_t \mid \theta, X_{1:T}] \leq (c_\delta + 1) \mathbb{E} [\Sigma_t \mid \theta, X_{1:T}] \quad \text{where} \quad \delta^* = \left( \lambda_{\text{max}} (\Sigma_t) \left[ \lambda_{\text{max}} (\Sigma_t^{-1}) + \sigma^{-2} (t - L) \right] \right)^{-1}.
\]

Now,

\[
c_\delta + 1 = 8 \cdot \max \left\{ \sigma^{-2} \cdot \lambda_{\text{max}} (\Sigma_t) \cdot \log \left( \frac{dk}{\delta^*} \right), 1 \right\} + 1
\]

\[
= \max \left\{ 8 \sigma^{-2} \cdot \lambda_{\text{max}} (\Sigma_t) \cdot \log \left( \frac{dk \lambda_{\text{max}} (\Sigma_t)}{\lambda_{\text{max}} (\Sigma_t^{-1}) + \sigma^{-2} (t - L)} \right) + 1, 9 \right\}
\]

\[
\leq \max \left\{ 8 \sigma^{-2} \cdot \lambda_{\text{max}} (\Sigma_t) \cdot \log \left( \frac{dk \lambda_{\text{max}} (\Sigma_t)}{\lambda_{\text{max}} (\Sigma_t^{-1}) + \sigma^{-2} T} \right) + 1, 9 \right\}
\]

\[
\triangleq I_t.
\]

Fix \(i \in [k]\) and use again the notation \(\phi(x, i) = (0, \ldots, 0, x, 0, \ldots, 0) \in \mathbb{R}^{dk}\) to be the concatenation of \(k\) subvectors of size \(d\), where the \(i\)-th subvector is \(x\). Then \(s^2_{t,i} = \phi(x_{\text{pop}}, i)^\top \Sigma_t \phi(x_{\text{pop}}, i)\) and \(s^2_{t,i} = \phi(x_{\text{pop}}, i)^\top \Sigma_t \phi(x_{\text{pop}}, i)\). We have

\[
\mathbb{E} [s^2_{t,i} \mid \theta, X_{1:T}] = \mathbb{E} \left[ \phi(x_{\text{pop}}, i)^\top \Sigma_t \phi(x_{\text{pop}}, i) \mid \theta, X_{1:T} \right]
\]

\[
= \phi(x_{\text{pop}}, i)^\top \mathbb{E} [\Sigma_t \mid \theta, X_{1:T}] \phi(x_{\text{pop}}, i)
\]

\[
\leq I_t \cdot \phi(x_{\text{pop}}, i)^\top \mathbb{E} [\Sigma_t \mid \theta, X_{1:T}] \phi(x_{\text{pop}}, i)
\]

\[
= I_t \cdot \mathbb{E} \left[ \phi(x_{\text{pop}}, i)^\top \Sigma_t \phi(x_{\text{pop}}, i) \mid \theta, X_{1:T} \right]
\]

\[
= I_t \cdot \mathbb{E} [s^2_{t,i} \mid \theta, X_{1:T}].
\]

Since \(I_t\) is non-random conditioned on \(\theta\), we also have

\[
\mathbb{E} [s^2_{t,i} \mid \theta, X_{1:T}] \leq I_t \cdot \mathbb{E} [s^2_{t,i} \mid \theta, X_{1:T}].
\]

\(\square\)

### F.5 Bounding the posterior precision by attainable precision

We prove the following result, which applies to DTS, since the condition below holds under DTS, as shown in Proposition 2.
Proposition 6. If $\mathbb{E} \left[ \frac{1}{p_{L,t}} | X_{1:T} \right] \leq 1$ for each $t \in [T]$ and $i \in [k]$, then for any $t \geq L$, 

$$\mathbb{E} \left[ s_{L,t}^2 | X_{1:T} \right] \leq t \cdot \frac{k}{\text{Precision} \left( X_{1:(t-L)} \right)}.$$ 

Taking expectations of the inequality for $I^*$ in Corollary 2 and using the tower property of conditional expectations yields, $\mathbb{E} \left[ s_{L,t}^2 | X_{1:T} \right] \leq t \cdot \mathbb{E} \left[ s_{L,t}^2 | X_{1:T} \right]$. Hence it suffices to prove

$$\mathbb{E} \left[ s_{L,t}^2 | X_{1:T} \right] \leq \frac{k}{\text{Precision} \left( X_{1:(t-L)} \right)}.$$ (27)

Our main goal in this section is to prove (27) holds when arms are sampled according to DTS.

Preliminaries: matrix convex combinations.

Let $S^n$ denote the set of symmetric $n \times n$ square matrices, and let $S^n_+ \subset S^n$ 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 $A \in S_n$, 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 $S_n$ to $S_n$. A function $f$ is said to be monotone increasing on the space of positive semidefinite matrices if for $A, B \in S^n_+$, $A \succeq B$ implies $f(A) \leq f(B)$ 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 $A, B \in S^n_+$ and scalar $\lambda \in [0, 1]$, $f(\gamma A + (1-\gamma)B) \leq \gamma f(A) + (1-\gamma)f(B)$. For our purposes, a key fact is that the inverse function $f(A) = A^{-1}$ is convex and monotone decreasing.

To prove Proposition 6, we need to leverage a generalization of Jensen’s inequality that applies to matrix convex combinations. The following definitions can be found in Tropp [2015].

Definition 2 (Definition 8.5.1 in Tropp [2015] – Matrix Convex Combination). Let $B_1, B_2$ be Hermitian matrices (i.e. self-adjoint 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. 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.

Lemma 7 (Theorem 8.5.2 in Tropp [2015] – Operator Jensen Inequality). Let $f$ be an operator convex on the space of symmetric positive semidefinite matrices $S^n_+$. Let $B_1, B_2 \in S^n_+$. If $A_1^\top A_1 + A_2^\top A_2 = I$ then,

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

By induction, the lemma can be generalized to situations with more than two pairs of matrices.

Using inverse propensity weights to analyze the evolution of posterior.

The notation $V$ in Lemma 8 is used only to simplify this lemma statement and is not used again in this paper. Observe that if the action selection is not randomized, and satisfies $p_{\ell,i_t} = 1$ for each $\ell \in [t-L]$, then $\Sigma_t = \text{Cov}(\theta | R_{1,i_t}, \ldots, R_{t-L,i_{t-L}}, X_{1}, \ldots, X_{t-L})$ and the bound in Lemma 8 holds with equality.

Lemma 8 (Inverse-propensity weighted posterior evolution). Fix any sequence of arms $i_1, \ldots, i_{t-L}$. Then, with probability 1,

$$\Sigma_t \preceq V \left( \Sigma_{i_{t-L}}^{-1} + \sigma^2 \sum_{\ell=1}^{t-L} \frac{\phi(X_{\ell,i_{\ell}}) \phi(X_{\ell,i_{\ell}})^\top}{p_{\ell,i_{\ell}}} \right) V,$$

where $\phi(x)$ is the probability density function of the normal distribution with mean $x$ and variance $\sigma^2$.
where
\[
V = \text{Cov}(\theta \mid R_{1,i_1}, \ldots, R_{1-L,i_{1-L}}, X_{1}, \ldots, X_{1-L}) = \left(\Sigma_1^{-1} + \sigma^{-2} \sum_{\ell=1}^{1-L} \phi(X_{\ell,i_{\ell}})\phi(X_{\ell,i_{\ell}})^\top\right)^{-1}.
\]

**Proof.** First observe that
\[
\hat{\Sigma}_t = \left(\Sigma_1^{-1} + \sigma^{-2} \sum_{\ell=1}^{1-L} \sum_{i=1}^k p_{\ell,i}\phi(X_{\ell,i})\phi(X_{\ell,i})^\top\right)^{-1} \leq \left(\Sigma_1^{-1} + \sigma^{-2} \sum_{\ell=1}^{1-L} \sum_{i=1}^k p_{\ell,i}\phi(X_{\ell,i})\phi(X_{\ell,i})^\top\right)^{-1}.
\]

Since \(i_1, \ldots, i_{1-L}\) are fixed, drop them from notation and write \(p_\ell = p_{\ell,i_\ell} = \mathbb{P}(I_\ell = i_\ell \mid H_\ell)\). Set \(B_\ell = \sigma^{-2}\phi(X_{\ell,i_\ell})\phi(X_{\ell,i_\ell})^\top\). For notational convenience, set \(B_0 = \Sigma_1^{-1}\) and \(p_0 = 1\). Then \(V = \left(\sum_{\ell=0}^{1-L} B_\ell\right)^{-1}\), and the above inequality becomes
\[
\hat{\Sigma}_t \geq \left(\sum_{\ell=0}^{1-L} p_\ell B_\ell\right)^{-1}
\]
\[
= V^{1/2} \left[V^{1/2} \left(\sum_{\ell=0}^{1-L} p_\ell B_\ell\right) V^{1/2}\right]^{-1} V^{1/2}
\]
\[
= V^{1/2} \left[V^{1/2} \left(\sum_{\ell=0}^{1-L} B_\ell^{1/2} (p_\ell I) B_\ell^{1/2}\right) V^{1/2}\right]^{-1} V^{1/2}
\]
\[
= V^{1/2} \left[\sum_{\ell=0}^{1-L} \left(V^{1/2} B_\ell^{1/2}\right) (p_\ell I) \left(B_\ell^{1/2} V^{1/2}\right)\right]^{-1} V^{1/2}
\]
\[
\leq V^{1/2} \left[\sum_{\ell=0}^{1-L} \left(V^{1/2} B_\ell^{1/2}\right) (p_\ell I)^{-1} \left(B_\ell^{1/2} V^{1/2}\right)\right] V^{1/2}
\]
\[
= V \left(\sum_{\ell=0}^{1-L} B_\ell\right) V,\]
where the last inequality applies the operator Jensen inequality in Lemma 7, using that
\[
\sum_{\ell=0}^{1-L} \left(V^{1/2} B_\ell^{1/2}\right) \left(B_\ell^{1/2} V^{1/2}\right) = V^{1/2} \left(\sum_{\ell=0}^{1-L} B_\ell\right) V^{1/2} = V^{1/2} V^{-1} V^{1/2} = I.
\]

\[\square\]

**Completing the proof Proposition 6.**

Now we specialize this result to proof Proposition 6.

**Proof.** To start, we have
\[
\mathbb{E}\left[s_{T,I}^2 \mid X_{1:T}\right] = \mathbb{E}\left[\left.\mathbb{E}\left[s_{T,I}^2 \mid \theta, X_{1:T}\right]\right| X_{1:T}\right] \leq \mathbb{E}\left[\left.\mathbb{E}\left[s_{T,I}^2 \mid \theta, X_{1:T}\right]\right| X_{1:T}\right]
\]
\[
= \mathbb{E}\left[\left.\mathbb{E}\left[s_{T,I}^2 \mid X_{1:T}\right]\right| X_{1:T}\right],
\]
where the inequality applies Corollary 2, using that \(I^*\) is non-random conditioned on \(\theta\). The remainder of the proof bounds \(s_{T,I}^2\).
For $i \in [k]$, take
\[
\phi_i = \phi(x_{\text{pop}, i}) = (0, \ldots, 0, x_{\text{pop}, i+1}, \ldots, x_{\text{pop}, d}, 0, \ldots, 0) \in \mathbb{R}^{k_d}
\]
to be a vector whose $i$-th subvector is $x_{\text{pop}}$, and then $r_\theta(i) = \phi_i^\top \theta$. We can write
\[
s_{t,i}^2 = \text{Var} \left[ r_\theta(i) \mid (X_{\ell, (p_{\ell,j}, R_{\ell,j})} \in [k])_{\ell \in [t-L]} \right]
= \text{Var} \left[ \phi_i^\top \theta \mid (X_{\ell, (p_{\ell,j}, R_{\ell,j})} \in [k])_{\ell \in [t-L]} \right]
= \phi_i^\top \text{Cov} \left[ \theta \mid (X_{\ell, (p_{\ell,j}, R_{\ell,j})} \in [k])_{\ell \in [t-L]} \right] \phi_i
= \phi_i^\top \Sigma_i \phi_i.
\]
Now set
\[
V_i = \text{Cov}(\theta \mid R_{1,i}, \ldots, R_{t-L,i}, X_1, \ldots, X_{t-L}) = \left( \Sigma_1^{-1} + \sigma^{-2} \sum_{\ell=1}^{t-L} \phi(X_{\ell,i})\phi(X_{\ell,i})^\top \right)^{-1}.
\]
Applying Lemma 8 with $i_\ell = i$ for each $\ell \in [t-L]$ gives,
\[
s_{t,i}^2 \leq \phi_i^\top V_i \left( \Sigma_1^{-1} + \sigma^{-2} \sum_{\ell=1}^{t-L} \frac{\phi(X_{\ell,i})\phi(X_{\ell,i})^\top}{p_{\ell,i}} \right) V_i \phi_i.
\]
Next,
\[
s_{t,i}^2 = \sum_{i=1}^k \mathbb{1}(I^* = i)s_{t,i}^2 \leq \sum_{i=1}^k \mathbb{1}(I^* = i)\phi_i^\top V_i \left( \Sigma_1^{-1} + \sigma^{-2} \sum_{\ell=1}^{t-L} \frac{\phi(X_{\ell,i})\phi(X_{\ell,i})^\top}{p_{\ell,i}} \right) V_i \phi_i
\leq \sum_{i=1}^k \phi_i^\top V_i \left( \Sigma_1^{-1} + \sigma^{-2} \sum_{\ell=1}^{t-L} \frac{\mathbb{1}(I^* = i)}{p_{\ell,i}} \phi(X_{i})\phi(X_{i})^\top \right) V_i \phi_i.
\]
Since $\mathbb{E} \left[ \frac{\mathbb{1}(I^* = i)}{p_{\ell,i}} \mid X_{1:T} \right] \leq 1$ for any $\ell$ and $i$, we have
\[
\mathbb{E} \left[ s_{t,i}^2 \mid X_{1:T} \right] \leq \sum_{i=1}^k \phi_i^\top V_i \left( \Sigma_1^{-1} + \sigma^{-2} \sum_{\ell=1}^{t-L} \phi(X_{\ell,i})\phi(X_{\ell,i})^\top \right) V_i \phi_i
= \sum_{i=1}^k \phi_i^\top V_i V_i^{-1} V_i \phi_i
= \sum_{i=1}^k \phi_i^\top V_i \phi_i.
\]
Recalling that \( V_t = \text{Cov}(\theta \mid R_{1,j}, \ldots, R_{t-L,j}, X_1, \ldots, X_{t-L}) \) and that \( r_{\theta}(i) = \phi_i^T \theta \), gives

\[
\sum_{i=1}^{k} \phi_i^T V_i \phi_i = \sum_{i=1}^{k} \phi_i^T \text{Cov}(\theta \mid R_{1,j}, \ldots, R_{t-L,j}, X_1, \ldots, X_{t-L}) \phi_i \\
= \sum_{i=1}^{k} \text{Var} \left( r_{\theta}(i) \mid R_{1,j}, \ldots, R_{t-L,j}, X_1, \ldots, X_{t-L} \right) \\
\leq k \cdot \max_{i \in [k]} \text{Var} \left( r_{\theta}(i) \mid R_{1,j}, \ldots, R_{t-L,j}, X_1, \ldots, X_{t-L} \right) \\
= \frac{k}{\text{Precision} \left( X_{t,(t-L)} \right)},
\]

completing the proof.

\[\square\]

## G Matrix-valued Martingales and the proof of Proposition 5

We begin by restating the result.

**Proposition 5** (Optionally sampled matrix-valued process). Consider a deterministic sequence of positive semidefinite matrices \( V_1, V_2, \ldots \in \mathbb{S}^d \) satisfying \( \sup_{t \in \mathbb{N}} \lambda_{\text{max}}(V_t) \leq 1 \), and a random process \((Z_t)_{t \in \mathbb{N}}\) taking values in \( \{0,1\} \) that is adapted to some filtration \((\mathcal{F}_t)_{t \in \mathbb{N}_0}\). For \( n \in \mathbb{N} \), define

\[
S_n = \sum_{t=1}^{n} Z_t V_t \quad \text{and} \quad \hat{S}_n = \sum_{t=1}^{n} \mathbb{P}(Z_t = 1 \mid \mathcal{F}_{t-1}) V_t.
\]

Then, for any \( \delta > 0 \), with probability exceeding \( 1 - \delta \),

\[
S_n \geq (3 - \epsilon) \hat{S}_n - \log \left( \frac{d}{\delta} \right) I, \quad \forall n \in \mathbb{N}.
\]

We let \( \mathbb{P}_t(\cdot) = \mathbb{P}(\cdot \mid \mathcal{F}_t) \) and \( \mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_t] \). Set \( p_t = \mathbb{E}_{t-1}[Z_t] = \mathbb{P}_{t-1}(Z_t = 1) \).

Throughout this proof, we use some specialized notation. Define \( D_t = (p_t - Z_t) V_t \). We study

\[
A_0 \triangleq 0 \in \mathbb{R}^{d \times d} \quad \text{and} \quad A_n \triangleq \hat{S}_n - S_n = \sum_{t=1}^{n} D_t,
\]

which is the sum of matrix martingale differences. We will follow Tropp [2011] fairly closely. Define \( \Phi_t(\gamma) \triangleq \log \left( \mathbb{E}_{t-1} \left[ e^{\gamma D_t} \right] \right) \). Then set

\[
\Phi_0 \triangleq 0 \in \mathbb{R}^{d \times d} \quad \text{and} \quad \Phi_n(\gamma) \triangleq \sum_{t=1}^{n} \Phi_t(\gamma).
\]

Recognize that both \( A_0 \) and \( \Phi_0 \) are matrices where all elements equal zero. Here \( \Phi_n(\gamma) \) measures the total variability of the process. Our aim is to show that \( A_n \) can only be large if \( \Phi_n(\gamma) \) 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 9** (Lemma 6.7 in Tropp [2012]). Suppose \( D \) is a random self-adjoint matrix that satisfies

\[
\mathbb{E}[D] = 0 \quad \text{and} \quad \mathbb{P} \left( \lambda_{\text{max}}(D) \leq 1 \right) = 1.
\]
Then
\[ \mathbb{E} \left[ e^{\gamma D} \right] \leq \exp \left\{ (e^\gamma - \gamma - 1) \mathbb{E} \left[ D^2 \right] \right\}, \quad \forall \gamma > 0. \]

As a consequence of this, we can bound the sum of cumulants \( \Phi_n(\gamma) \) by a simpler quantity that closely mimics \( S_n \).

**Lemma 10.** For any \( \gamma > 0 \) and \( n \in \mathbb{N} \cup \{0\} \),
\[
\Phi_n(\gamma) \leq (e^\gamma - \gamma - 1) \sum_{i=1}^{n} p_i V_i = (e^\gamma - \gamma - 1) S_n.
\]

**Proof of Lemma 10.** We have \( \lambda_{\max}(D_t) \leq |p_t - Z_t| \lambda_{\max}(V_t) \leq 1 \) where the first inequality holds since \( V_t \) is positive semidefinite and the last inequality follows from an assumption on the maximum eigenvalue of \( V_t \). This allows us to apply the matrix Bernstein inequality above. For notional convenience define \( f(\gamma) = e^\gamma - \gamma - 1 \). By Lemma 9, we have that
\[
\phi_t(\gamma) = \log \mathbb{E}_{t-1} [e^{\gamma D_t}] \leq f(\gamma) \mathbb{E}_{t-1} [D_t^2].
\]
Using this gives,
\[
\Phi_n(\gamma) \leq f(\gamma) \sum_{i=1}^{n} \mathbb{E}_{t-1} [D_t^2] = f(\gamma) \sum_{i=1}^{n} \mathbb{E}_{t-1} [(p_t - Z_t)^2] V_i^2
\leq f(\gamma) \sum_{i=1}^{n} p_t (1 - p_t) V_i
\leq f(\gamma) \sum_{i=1}^{n} p_t V_t,
\]

as desired. To prove the first inequality, recall that all eigenvalues of \( V_t \in \mathbb{S}^d_+ \) are non-negative and smaller than 1. Write \( V_t = \sum_{i=1}^{d} \lambda_i u_i u_i^\top \) in terms of its eigenvalues \( \lambda_i \in [0, 1] \) and eigenvectors \( u_i \). Then \( V_t^2 u_i = V_t(\lambda_i u_i) = \lambda_i^2 u_i \). The matrix \( V_t^2 \) also has \( (u_1, \ldots, u_d) \) as eigenvectors, but with corresponding smaller corresponding eigenvalues \( (\lambda_1^2, \ldots, \lambda_d^2) \).

**An exponential super-martingale.**

Again, our goal is to show that \( A_n \) can only be large if \( \Phi_n(\gamma) \) is large. To this end, define
\[
M_n(\gamma) = \text{tr} \exp(\gamma A_n - \Phi_n(\gamma)), \quad \forall n \in \{0, 1, \ldots\},
\]
where \( \text{tr} \) denotes the trace operator. We now show that \( M_n(\gamma) \) 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 2** (Theorem 6 in 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 3** (Corollary 1.5 in Tropp [2011]). Fix a self-adjoint matrix \( H \). For a random self-adjoint matrix \( X \),
\[
\mathbb{E} [ \text{tr} \exp(H + X)] \leq \text{tr} \exp \left( H + \log \left( \mathbb{E} e^X \right) \right).
\]

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

**Corollary 4.** For each \( \gamma > 0 \), \( \{M_n(\gamma) : n = 0, 1, \ldots\} \) is a super-martingale with initial value \( M_0(\gamma) = d \).
Proof of Corollary 4. By definition,

\[ M_0(\gamma) = \text{tr} \exp(\gamma A_0 - \Phi_0(\gamma)) = \text{tr} \exp(0) = \text{tr}I = d. \]

For \( n > 0 \), taking conditional expectations gives

\[
\mathbb{E}_{n-1} [M_n(\gamma)] = \mathbb{E}_{n-1} [\text{tr} \exp(\gamma A_{n-1} - \Phi_n(\gamma) + \gamma D_n)] \\
\leq \text{tr} \exp(\gamma A_{n-1} - \Phi_n(\gamma) + \Phi_n(\gamma)) \\
= \text{tr} \exp(\gamma A_{n-1} - \Phi_n(\gamma)) = M_{n-1}(\gamma),
\]

where the inequality follows by Corollary 3, using that \( \log \). \( \square \)

Boundary crossing probabilities.

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

Lemma 11. For any fixed \( \delta > 0 \) and \( \gamma > 0 \), with probability exceeding \( 1 - \delta \),

\[ A_n \leq \frac{1}{\gamma} \left[ \Phi_n(\gamma) + \log \left( \frac{d}{\delta} \right) I \right], \quad \forall n \in \mathbb{N}. \]

Proof of Lemma 11. Fix \( \gamma > 0 \) throughout. Let \( y \in \mathbb{R} \). We have

\[
\mathbb{P} (\lambda_{\text{max}} (\gamma A_n - \Phi_n(\gamma)) \geq y) = \mathbb{P} \left( e^{\lambda_{\text{max}}(\gamma A_n - \Phi_n(\gamma))} \geq e^y \right) \leq \mathbb{P} \left( \text{tr} e^{\gamma A_n - \Phi_n(\gamma)} \geq e^y \right) \\
\leq e^{-y} \mathbb{E} \left[ \text{tr} e^{\gamma A_n - \Phi_n(\gamma)} \right] \\
= e^{-y} \mathbb{E} [M_n(\gamma)].
\]

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

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

Take \( \tau = \inf \{ n \in \mathbb{N} : \lambda_{\text{max}} (\gamma A_n - \Phi_n(\gamma)) \geq y \} \), with the convention that \( \tau = \infty \) if \( \lambda_{\text{max}} (\gamma A_n - \Phi_n(\gamma)) < y \) for every \( n \in \mathbb{N} \). Then,

\[
\mathbb{P} (\exists n \leq N : \lambda_{\text{max}} (\gamma A_n - \Phi_n(\gamma)) \geq y) = \mathbb{P} (\lambda_{\text{max}} (\gamma A_{\tau \wedge N} - \Phi_{\tau \wedge N}(\gamma)) \geq y) \\
\leq e^{-y} \mathbb{E} [M_{\tau \wedge N}(\gamma)] \\
\leq e^{-y} d.
\]

That \( \mathbb{E} [M_{\tau \wedge N}(\gamma)] \leq d \) uses Corollary 4 and Doob’s optional sampling theorem. Taking \( N \to \infty \) and applying the monotone convergence theorem gives,

\[
\mathbb{P} (\exists n \in \mathbb{N} : \lambda_{\text{max}} (\gamma A_n - \Phi_n(\gamma)) \geq y) \leq e^{-y} d.
\]

For any \( \delta > 0 \), we choose \( y = \log(d/\delta) \), and then with probability at least \( 1 - \delta \),

\[
\lambda_{\text{max}} (\gamma A_n - \Phi_n(\gamma)) \leq \log \left( \frac{d}{\delta} \right), \quad \forall n \in \mathbb{N}.
\]

\( \square \)

Combining our results completes the proof of Proposition 5.
Proof of Proposition 5. Define \( f(\gamma) = e^\gamma - \gamma - 1 \). Recall \( A_n = S_n - \bar{S}_n \). By applying Lemmas 11 and 10, we get that with probability at least \( 1 - \delta \),

\[
S_n - \bar{S}_n = -A_n \geq -\frac{1}{\gamma} \left[ \Phi_n(\gamma) + \log \left( \frac{d}{\delta} \right) I \right] \geq -\frac{1}{\gamma} \left[ f(\gamma) \sum_{t=1}^{n} p_t V_t + \log \left( \frac{d}{\delta} \right) I \right].
\]

Picking \( \gamma = 1 \), we have

\[
S_n - \bar{S}_n \geq (2 - e) \sum_{t=1}^{n} p_t V_t - \log \left( \frac{d}{\delta} \right) I = (2 - e) \bar{S}_n - \log \left( \frac{d}{\delta} \right) I.
\]

Adding \( \bar{S}_n \) to both sides yields the result. \( \square \)

**H Bounds on attainable precision: proof of Lemma 1**

In this section, we use \( \| \cdot \| \) to denote the spectral norm. First, we restate the claim.

**Lemma 1 (Bound on attainable precision).** Fix any sequence \( x_{1:T} \in \mathcal{X}^T \) and \( t \in [T] \).

1. (Generic bound) Let \( S_x \triangleq \frac{1}{t} \sum_{\ell=1}^{t} x_{\ell} x_{\ell}^\top \) denote the empirical second moment matrix and \( \bar{S}_x \triangleq S_x + \sigma^2 \lambda_{\min}(\Sigma_1^{-1}) I \) (where \( I \in \mathbb{R}^{d \times d} \) is an identify matrix). Then

\[
\text{Precision}(x_{1:t}) \geq \sigma^{-2} t \cdot \left( x_{\text{pop}}^\top \bar{S}_x^{-1} x_{\text{pop}} \right)^{-1}.
\]

2. (Vanilla bandit) Suppose \( d = 1 \) and \( x_{\ell} = 1 = x_{\text{pop}} \) for each \( \ell \in [t] \). Then

\[
\text{Precision}(x_{1:t}) = \min_{i \in [k]} \Sigma_{1,ii}^{-1} + \sigma^{-2} t \geq \lambda_{\min} \left( \Sigma_1^{-1} \right) + \sigma^{-2} t,
\]

where \( \Sigma_{1,ii} \) is the \((i,i)\)-th element of the prior covariance matrix \( \Sigma_1 \).

3. (No empirical distribution shift) Suppose \( \frac{1}{t} \sum_{\ell=1}^{t} x_{\ell} = x_{\text{pop}} \). Then

\[
\text{Precision}(x_{1:t}) \geq \lambda_{\min} \left( \Sigma_1^{-1} \right) \| x_{\text{pop}} \|_2^{-2} + \sigma^{-2} t.
\]

4. (I.i.d. contexts) Suppose \( X_1, \ldots, X_t \) are drawn i.i.d. from a distribution satisfying that \( \mathbb{E}[X_1 X_1^\top] \succeq c \cdot x_{\text{pop}} x_{\text{pop}}^\top \) for some \( c \geq 0 \). Then for any \( \delta > 0 \), with probability greater than \( 1 - \delta \),

\[
\text{Precision}(X_{1:t}) \geq \lambda_{\min} \left( \Sigma_1^{-1} \right) \| x_{\text{pop}} \|_2^{-2} + c \cdot \sigma^{-2} t - 4 \sigma^{-2} \| x_{\text{pop}} \|_2^{-2} \sqrt{2t \log \left( \frac{d}{\delta} \right)}. \]

and

\[
t \geq \frac{128 \| x_{\text{pop}} \|_2^{-4} \log \left( \frac{1}{\delta} \right)}{c^2} \implies \text{Precision}(X_{1:t}) \geq \lambda_{\min} \left( \Sigma_1^{-1} \right) \| x_{\text{pop}} \|_2^{-2} + \frac{c}{2} \cdot \sigma^{-2} t.
\]
Proof of Lemma 1. The definition of precision in (12) gives

\[
\frac{1}{\text{Precision}(x_{1:t})} = \max_{i \in [k]} x_{\text{pop}}^\top \left[ \text{Cov} \left( \theta^{(i)} \right)^{-1} + \sigma^{-2} \sum_{\ell = 1}^t x_{\ell} x_{\ell}^\top \right]^{-1} x_{\text{pop}} \\
\leq x_{\text{pop}}^\top \left[ \lambda_{\min} \left( \Sigma^{-1}_1 \right) I + \sigma^{-2} \sum_{\ell = 1}^t x_{\ell} x_{\ell}^\top \right]^{-1} x_{\text{pop}} \\
= x_{\text{pop}}^\top \left( \sigma^{-2} t S_{x} \right)^{-1} x_{\text{pop}},
\]

where the inequality uses Lemma 12. Taking the inverse on both sides yields the generic bound on precision. For vanilla bandit (with \( d = 1 \)), since \( \theta^{(i)} \in \mathbb{R} \) for each \( i \in [k] \) and \( x_{\ell} = x_{\text{pop}} \) for each \( \ell \in [t] \), the definition of precision in (12) becomes

\[
\text{Precision}(x_{1:t}) = \min_{i \in [k]} \text{Var} \left( \theta^{(i)} \right)^{-1} + \sigma^{-2} t = \min_{i \in [k]} \Sigma_{1,ii}^{-1} + \sigma^{-2} t,
\]

and the above generic bound gives

\[
\text{Precision}(x_{1:t}) \geq \lambda_{\min} \left( \Sigma^{-1}_1 \right) + \sigma^{-2} t.
\]

Next we analyze the setting without empirical distribution shift. By (28) and Lemma 13,

\[
\frac{1}{\text{Precision}(x_{1:t})} \leq x_{\text{pop}}^\top \left[ \lambda_{\min} \left( \Sigma^{-1}_1 \right) \cdot I + \sigma^{-2} \sum_{\ell = 1}^t x_{\ell} x_{\ell}^\top \right]^{-1} x_{\text{pop}} \\
\leq x_{\text{pop}}^\top \left[ \lambda_{\min} \left( \Sigma^{-1}_1 \right) \cdot I + \sigma^{-2} t \cdot x_{\text{pop}} x_{\text{pop}}^\top \right]^{-1} x_{\text{pop}}.
\]

Fact 2 implies that \( x_{\text{pop}} x_{\text{pop}}^\top \) only has one non-zero eigenvalue \( \|x_{\text{pop}}\|_2^2 \) with a corresponding eigenvector \( \frac{x_{\text{pop}}}{\|x_{\text{pop}}\|_2} \) (recall that \( \|x_{\text{pop}}\|_2 \neq 0 \)), so the eigendecomposition of \( x_{\text{pop}} x_{\text{pop}}^\top \) can be written as

\[
x_{\text{pop}} x_{\text{pop}}^\top = QAQ^\top
\]

where \( \Lambda = \text{diag} \left( \|x_{\text{pop}}\|_2^2, 0, \ldots, 0 \right) \in \mathbb{R}^{d \times d} \) is the diagonal matrix whose diagonal elements are the eigenvalues of \( x_{\text{pop}} x_{\text{pop}}^\top \), and \( Q \in \mathbb{R}^{d \times d} \) is a corresponding orthogonal matrix with the first column being \( \frac{x_{\text{pop}}}{\|x_{\text{pop}}\|_2} \). Then we have

\[
\lambda_{\min} \left( \Sigma^{-1}_1 \right) \cdot I + \sigma^{-2} t \cdot x_{\text{pop}} x_{\text{pop}}^\top \\
= Q \cdot \left[ \lambda_{\min} \left( \Sigma^{-1}_1 \right) \cdot I \right] \cdot Q^\top + Q \cdot \left[ \sigma^{-2} t \cdot \Lambda \right] \cdot Q^\top \\
= Q \cdot \text{diag} \left( \lambda_{\min} \left( \Sigma^{-1}_1 \right) + \sigma^{-2} t \|x_{\text{pop}}\|_2^2, \lambda_{\min} \left( \Sigma^{-1}_1 \right), \ldots, \lambda_{\min} \left( \Sigma^{-1}_1 \right) \right) \cdot Q^\top.
\]

Hence,

\[
\left[ \lambda_{\min} \left( \Sigma^{-1}_1 \right) \cdot I + \sigma^{-2} t \cdot x_{\text{pop}} x_{\text{pop}}^\top \right]^{-1} \\
= Q \cdot \text{diag} \left( \frac{1}{\lambda_{\min} \left( \Sigma^{-1}_1 \right) + \sigma^{-2} t \|x_{\text{pop}}\|_2^2}, \frac{1}{\lambda_{\min} \left( \Sigma^{-1}_1 \right)}, \ldots, \frac{1}{\lambda_{\min} \left( \Sigma^{-1}_1 \right)} \right) \cdot Q^\top,
\]

49
where the first and second inequalities apply the triangle inequality and Jensen’s inequality, respectively; the penultimate equality follows from that $Q \in \mathbb{R}^{d \times d}$ is the orthogonal matrix with the first column being $\frac{x_{\text{pop}}}{\|x_{\text{pop}}\|^2}$. Taking the inverse on both sides gives the lower bound on precision when there is no empirical distribution shift.

Lastly we study the setting with i.i.d. contexts. Let $\Omega = \mathbb{E}[X_1 X_1^\top]$ and $Z_\ell = X_\ell X_\ell^\top - \Omega$ for $\ell \in [t]$, and we have $\Omega = \mathbb{E}[X_1 X_1^\top] \succeq c x_{\text{pop}} x_{\text{pop}}^\top$. Then by (28),

\[
\frac{1}{\text{Precision}(X_{1:t})} \leq x_{\text{pop}}^\top \left[ \lambda_{\min} \left( \Sigma_1^{-1} \right) \cdot I + \sigma^{-2} \sum_{\ell = 1}^t X_\ell X_\ell^\top \right]^{-1} x_{\text{pop}}
\]

\[
= x_{\text{pop}}^\top \left[ \lambda_{\min} \left( \Sigma_1^{-1} \right) \cdot I + \sigma^{-2} \sum_{\ell = 1}^t Z_\ell \right]^{-1} x_{\text{pop}}
\]

\[
\leq x_{\text{pop}}^\top \left[ \lambda_{\min} \left( \Sigma_1^{-1} \right) \cdot I + \sigma^{-2} \sum_{\ell = 1}^t Z_\ell + c \cdot \sigma^{-2} \cdot x_{\text{pop}} x_{\text{pop}}^\top \right]^{-1} x_{\text{pop}}.
\]

Note that the spectral norm of $Z_\ell$ can be bounded as follows, by the triangle inequality,

\[
\|Z_\ell\| \leq \|X_\ell X_\ell^\top\| + \|\mathbb{E}[X_1 X_1^\top]\| \leq \|X_1 X_1^\top\| + \mathbb{E}[\|X_1 X_1^\top\|] = \|X_1\|_2^2 + \mathbb{E}\left[\|X_1\|_2^2\] \leq 2,
\]

where the first and second inequalities apply the triangle inequality and Jensen’s inequality, respectively; the next equality uses Fact 2; the last inequality follows from an assumption on the maximum $\ell_2$ norm of context vectors. This implies $Z_\ell^2 \leq 4I$. By the matrix Hoeffding inequality in Lemma 14, for $x \geq 0$, with probability at least $1 - d \exp(-x^2/(32t))$

\[
\lambda_{\min} \left( \sum_{\ell = 1}^t Z_\ell \right) > -x, \quad \text{and thus} \quad \sum_{\ell = 1}^t Z_\ell \succeq -xI.
\]
Hence, for $x \geq 0$, with probability at least $1 - d \exp(-x^2/(32t))$,

$$\frac{1}{\text{Precision}(X_{1:t})} \leq x_{\text{pop}}^\top \left[ \lambda_{\text{min}} \left( \Sigma_1^{-1} \right) \cdot I + \sigma^{-2} \sum_{\ell = 1}^t Z_{\ell} + c \cdot \sigma^{-2} t \cdot x_{\text{pop}}^\top x_{\text{pop}} \right]^{-1} x_{\text{pop}}$$

$$\leq x_{\text{pop}}^\top \left[ \left( \lambda_{\text{min}} \left( \Sigma_1^{-1} \right) - \sigma^{-2} x \right) I + c \cdot \sigma^{-2} t \cdot x_{\text{pop}}^\top x_{\text{pop}} \right]^{-1} x_{\text{pop}}$$

$$= x_{\text{pop}}^\top Q \cdot \left( \begin{array}{cccc} \frac{1}{\lambda_{\text{min}} \left( \Sigma_1^{-1} \right) - \sigma^{-2} x + c \cdot \sigma^{-2} t \| x_{\text{pop}} \|_2^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\lambda_{\text{min}} \left( \Sigma_1^{-1} \right) - \sigma^{-2} x} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\lambda_{\text{min}} \left( \Sigma_1^{-1} \right) - \sigma^{-2} x} \end{array} \right) \cdot Q^\top x_{\text{pop}}$$

$$= \left( \| x_{\text{pop}} \|_2, 0, \ldots, 0 \right) \left( \begin{array}{cccc} \frac{1}{\lambda_{\text{min}} \left( \Sigma_1^{-1} \right) - \sigma^{-2} x + c \cdot \sigma^{-2} t \| x_{\text{pop}} \|_2^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\lambda_{\text{min}} \left( \Sigma_1^{-1} \right) - \sigma^{-2} x} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\lambda_{\text{min}} \left( \Sigma_1^{-1} \right) - \sigma^{-2} x} \end{array} \right) \left( \begin{array}{c} \| x_{\text{pop}} \|_2 \\ 0 \\ \vdots \\ 0 \end{array} \right)$$

$$= \frac{\| x_{\text{pop}} \|_2^2}{\lambda_{\text{min}} \left( \Sigma_1^{-1} \right) - \sigma^{-2} x + c \cdot \sigma^{-2} t \| x_{\text{pop}} \|_2^2}$$

where the equalities above follow the same analysis for the setting with no empirical distribution shift. Equivalently, for $\delta > 0$, with probability at least $1 - \delta$,

$$\frac{1}{\text{Precision}(X_{1:t})} \leq \frac{\| x_{\text{pop}} \|_2^2}{\lambda_{\text{min}} \left( \Sigma_1^{-1} \right) - 4\sigma^{-2} \sqrt{2t \log \frac{d}{\delta}} + c \cdot \sigma^{-2} t \| x_{\text{pop}} \|_2^2}$$

and taking the inverse on both sides gives

$$\text{Precision}(X_{1:t}) \geq \lambda_{\text{min}} \left( \Sigma_1^{-1} \right) \| x_{\text{pop}} \|_2^2 + c \cdot \sigma^{-2} t - 4\sigma^{-2} \| x_{\text{pop}} \|_2^2 \sqrt{2t \log \frac{d}{\delta}}.$$  

\[ \square \]

**Supporting results.** We introduce several supporting results for the proof of Lemma 1.

**Lemma 12.** For $i \in [k]$, $\lambda_{\text{min}} \left( \text{Cov} \left( \hat{\theta}^{(i)} \right) \right) \geq \lambda_{\text{min}} \left( \Sigma_1^{-1} \right)$.

**Proof.** Fix $i \in [K]$. We prove an equivalent statement: $\lambda_{\text{max}} \left( \text{Cov} \left( \hat{\theta}^{(i)} \right) \right) \leq \lambda_{\text{max}} \left( \Sigma_1 \right)$. The $\ell_2$ induced norm (i.e. spectral norm) of a positive semidefinite (and symmetric) matrix equals its largest eigenvalue, so we
\[
\lambda_{\text{max}} \left( \text{Cov} \left( \theta^{(i)} \right) \right) = \max_{x = (x_1, \ldots, x_d) \in \mathbb{R}^d : \|x\|_2 = 1} x^\top \text{Cov} \left( \theta^{(i)} \right) x
\]
\[
= \max_{x \in \mathbb{R}^d : \|x\|_2 = 1} \phi(x, i)^\top \Sigma_1 \phi(x, i)
\]
\[
\leq \max_{z \in \mathbb{R}^{d \times k}} z^\top \Sigma_1 z
\]
\[
= \lambda_{\text{max}} (\Sigma_1).
\]

The second equality above holds because \( \phi(x, i) = (0, \ldots, 0, x_1, \ldots, x_d, 0, \ldots, 0)^\top \in \mathbb{R}^{kd} \) has non-zero entries only in the \( i \)-th subvector, and then we only need to consider the corresponding submatrix of \( \Sigma_1 \) when calculating the quadratic term. This completes the proof.

\[\square\]

**Lemma 13.** For any \( t \in \mathbb{N} \),
\[
t \sum_{\ell=1}^t x_\ell x_\ell^\top \succeq \left( \sum_{\ell=1}^t x_\ell \right) \left( \sum_{\ell=1}^t x_\ell \right)^\top.
\]

*Proof.* Let \( \bar{x} = \frac{1}{t} \sum_{\ell=1}^t x_\ell \). Then the statement follows from
\[
\sum_{\ell=1}^t x_\ell x_\ell^\top = t \bar{x} \bar{x}^\top + \sum_{\ell=1}^t (x_\ell - \bar{x})(x_\ell - \bar{x})^\top \succeq t \bar{x} \bar{x}^\top.
\]

\[\square\]

**Fact 2.** Let \( x \in \mathbb{R}^d \). The matrix \( xx^\top \in \mathbb{R}^{d \times d} \) has only one potentially non-zero eigenvalue \( \|x\|_2^2 \) with a corresponding eigenvector \( x \). The spectral norm of \( xx^\top \), denoted by \( \|xx^\top\| \), equals \( \|x\|_2^2 \).

**Lemma 14** (Matrix Hoeffding – Theorem 1.3 in [Tropp, 2012]). Consider a finite sequence \( \{X_n\} \) of independent, random, self-adjoint matrices with dimension \( d \) and a sequence \( \{Y_n\} \) of fixed self-adjoint matrices. Assume that each random matrix satisfies
\[
\mathbb{E}[X_n] = 0 \quad \text{and} \quad X_n^2 \preceq Y_n^2 \quad \text{almost surely}.
\]
Then, for all \( x \geq 0 \),
\[
P \left( \lambda_{\text{max}} \left( \sum_n X_n \right) \geq x \right) \leq d \cdot \exp \left( \frac{-x^2}{8 \|\sum_n Y_n^2\|} \right)
\]
and\(^{10}\)
\[
P \left( \lambda_{\text{min}} \left( \sum_n X_n \right) \leq -x \right) \leq d \cdot \exp \left( \frac{-x^2}{8 \|\sum_n Y_n^2\|} \right).
\]

\(^{10}\)The inequality below follows from applying the inequality above to \( \{-X_n\} \) and \( \{Y_n\} \) and using \( P (\lambda_{\text{min}} \left( \sum_n X_n \right) \leq -x) = P (\lambda_{\text{max}} \left( \sum_n -X_n \right) \geq x) \). See Remark 3.10 (Minimum Eigenvalue) in [Tropp, 2012].